Werkzeuge der empirischen Forschung
R-Version

Wolfgang Kössler
(R-Übersetzung: Frank Fuhlbrück)

Institut für Informatik, Humboldt-Universität zu Berlin

Sommersemester 2012 17. August 2012

Inhalt (1)

Inhalt (2)

Inhalt (3)

1. Einleitung

Statistik und Wahrscheinlichkeitsrechnung

Stochastik

- befasst sich mit zufälligen Erscheinungen
  Häufigkeit, Wahrscheinlichkeit und Zufall
  grch: Kunst des geschickten Vermutens
- Teilgebiete
  - Wahrscheinlichkeitsrechnung
  - Statistik

Wahrscheinlichkeitsrechnung

gegebene Grundgesamtheit (Verteilung) $\rightarrow$ Aussagen über Realisierungen einer Zufallsvariablen treffen.

Einleitung

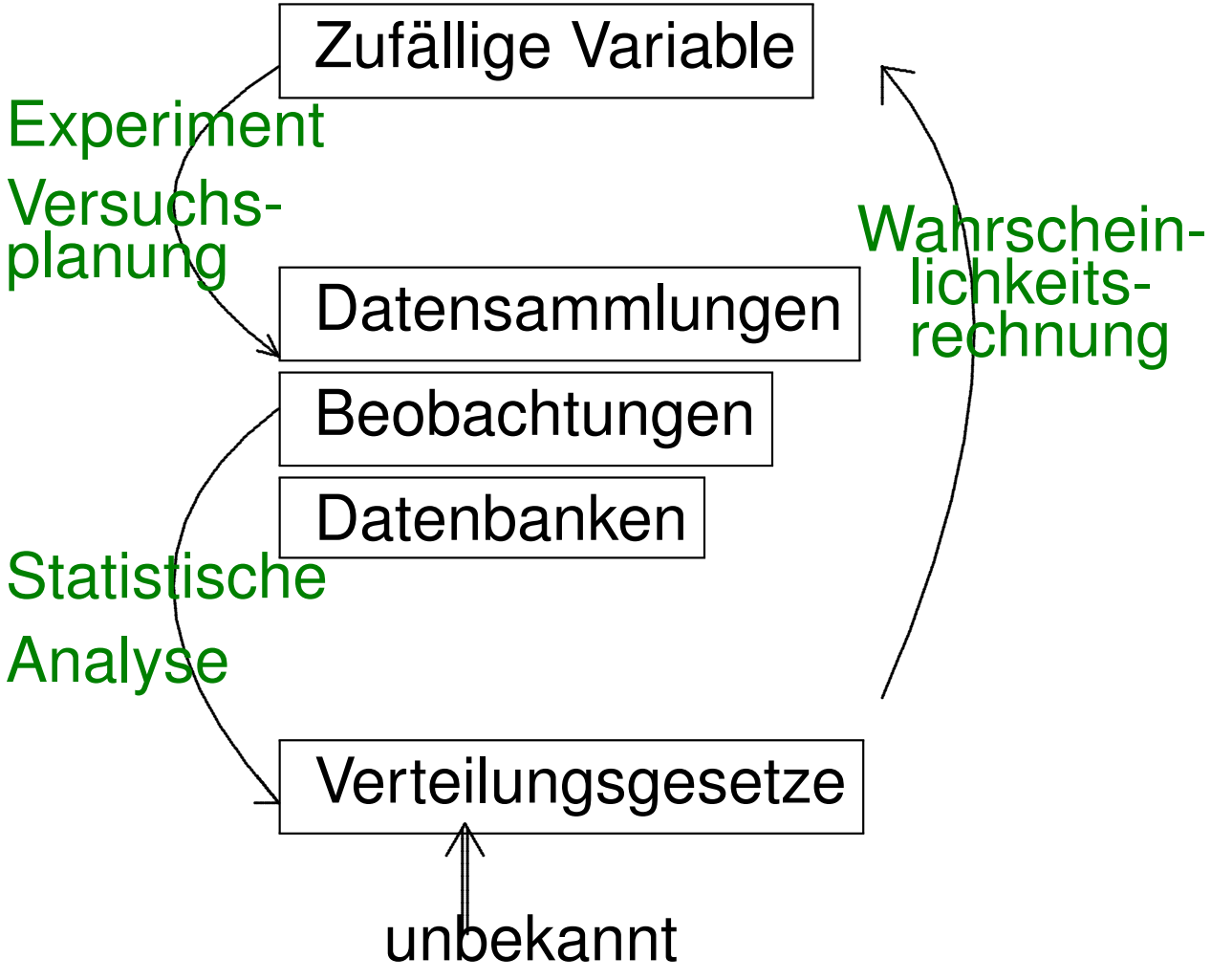
Statistik

Statistik

- Gesamtheit aller Methoden zur Analyse zufallsbehafteter Datenmengen
- Gegeben: (Besondere) zufallsbehaftete Datenmengen
- Gesucht: (Allgemeine) Aussagen über die zugrundeliegende Grundgesamtheit
- Teilgebiete:
  - Beschreibende oder Deskriptive Statistik
  - Induktive Statistik
  - Explorative oder Hyothesen-generierende Statistik (data mining)

Einleitung

Überblick: Statistik



Einleitung

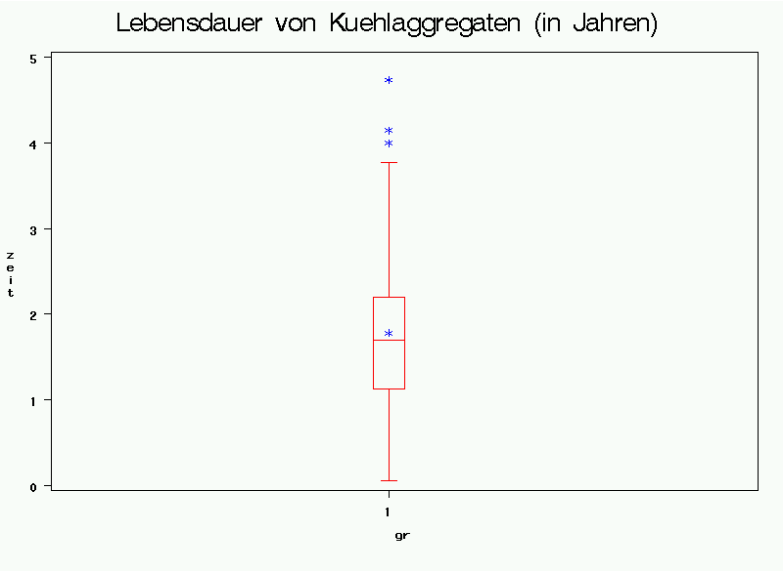
Beschreibene Statistik

Beschreibene Statistik

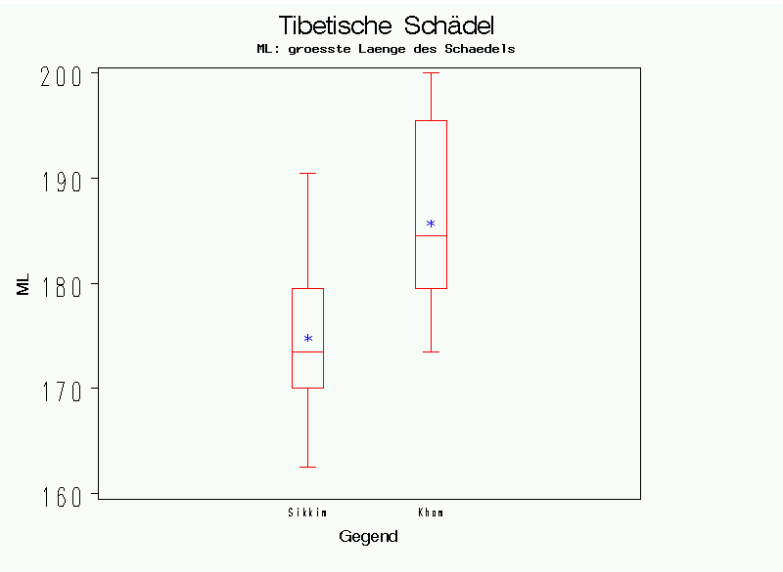
- statistische Maßzahlen: Mittelwerte, Streuungen, Quantile, ...
- Box-Blots
- Q-Q Plots
- Balkendiagramme
- Zusammenhangsmaße
- Punktediagramme (Scatterplots)

Boxplots - Beispiele

Lebensdauern von 100 Kühlaggregaten

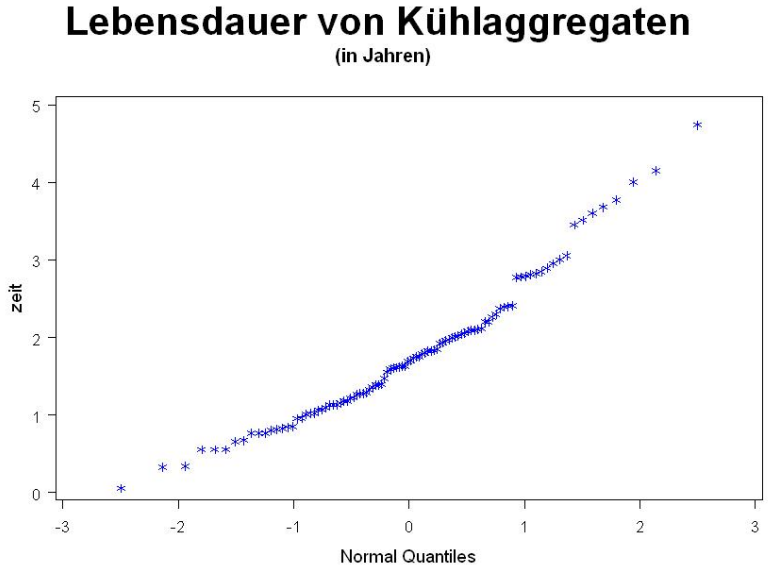


Schädelmaße in zwei Regionen Tibets

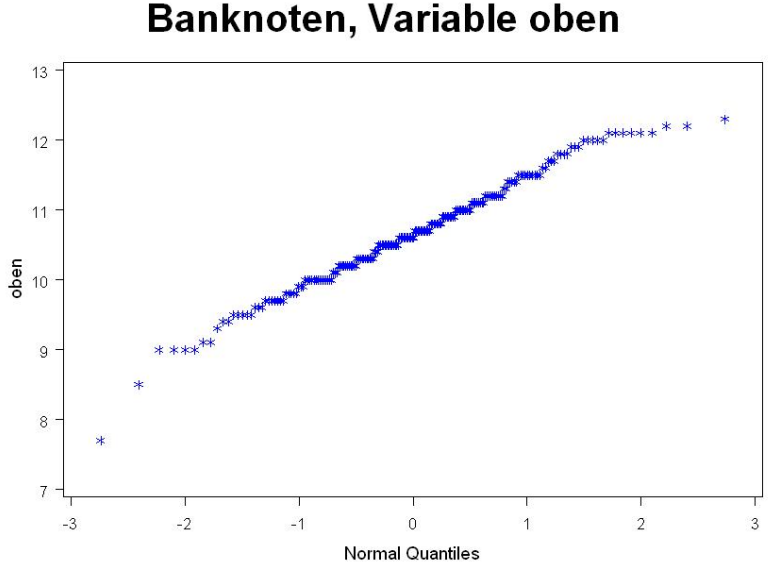


Q-Q Plots - Beispiele (1/2)

Lebensdauern von 100 Kühlaggregaten

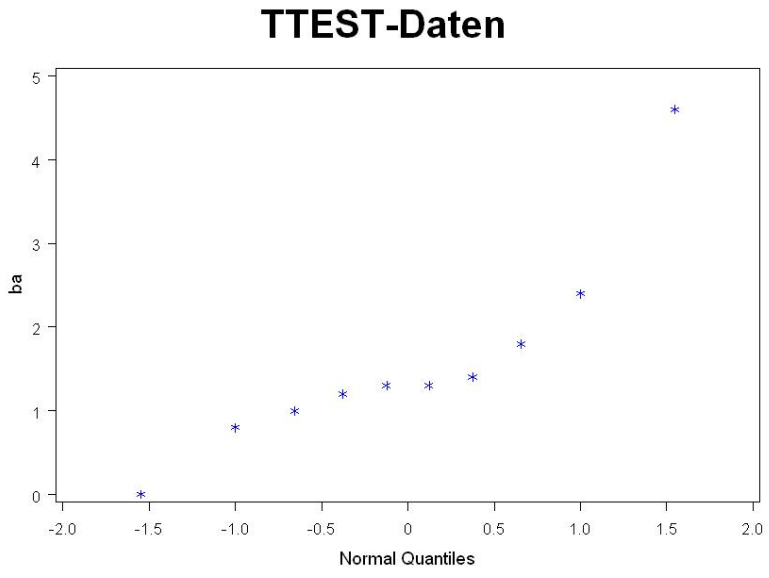


Abmessungen von Banknoten



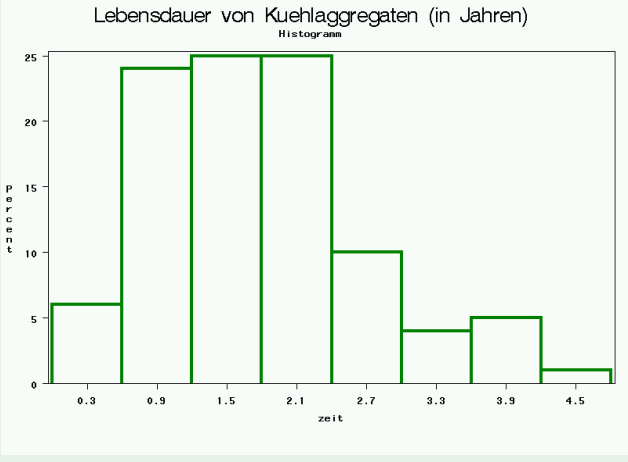
Q-Q Plots - Beispiele (2/2)

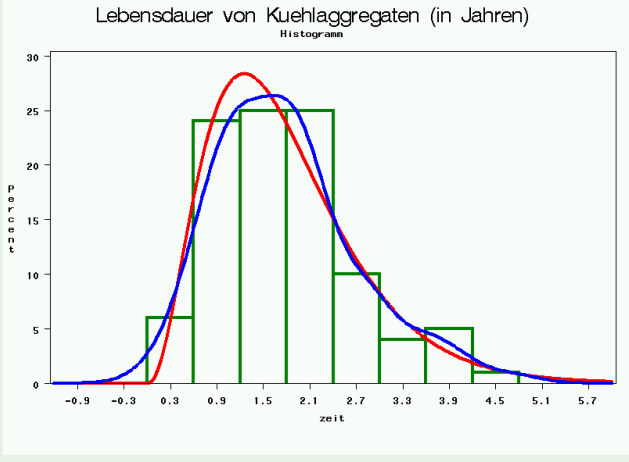
Verlängerung der Schlafdauer



Dichteschätzung, Beispiel

Kühlaggregate



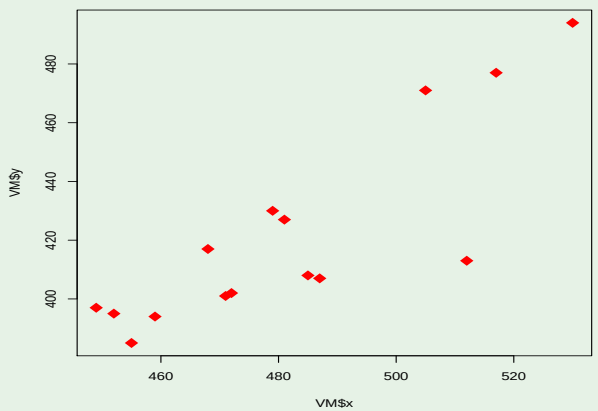


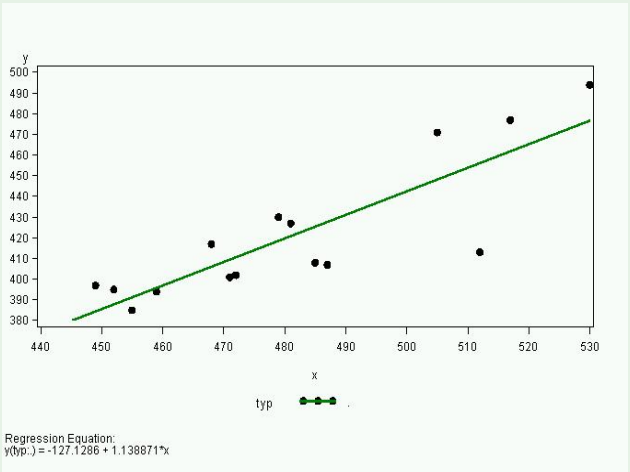
Histogramm

Parametrische Dichteschätzung (Gamma)

Nichtparametrische Dichteschätzung

Länge und Breite von Venusmuscheln





Einleitung

Schließende Statistik

Schließende Statistik

- Vergleich von Behandlungen, Grundgesamtheiten, Effekten
  $\rightarrow$ t-Test, Wilcoxon-Test, ANOVA, Kruskal-Wallis-Test, Friedman-Test
- Ursache-Wirkungsanalysen, Vorhersagen, Bestimmen funktionaler Beziehungen, Trendbestimmungen
  $\rightarrow$ lineare, nichtlineare Regression
  $\rightarrow$ Kurvenschätzung
  $\rightarrow$ logistische Regression
  $\rightarrow$ Korrelation und Unabhängigkeit

Einleitung

Schließende Statistik

Schließende Statistik

- Klassifikation
  - → Clusteranalyse
  - → Hauptkomponentenanalyse
  - → Faktorenanalyse
  - → Diskriminanzanalyse
- weitere Verfahren
  - → Lebensdaueranalyse (Zuverlässigkeit)
  - → Qualitätskontrolle
  - → Zeitreihenanalyse

Einleitung

Vergleich von Behandlungen, Grundgesamtheiten, Effekten

Vergleich von Behandlungen, Grundgesamtheiten, Effekten

- Einstichprobenproblem
  Messungen sollen mit einem vorgegebenen Wert verglichen werden
- Zweistichprobenproblem
  - Vergleich zweier unabhängiger Stichproben
  - Vergleich zweier abhängiger Stichproben
- Vergleich mehrerer unabhängiger Stichproben
- Vergleich mehrerer abhängiger Stichproben

Einleitung

Ein- und Zweistichprobenproblem

Eine Stichprobe

- Banknoten: vorgegebene Länge eingehalten?

$\rightarrow$ Einstichproben t-Test, Signed-Wilcoxon-Test

Abhängige und Unabhängige Stichproben

- Vergleich zweier unabhängiger Stichproben
  - echte - gefälschte Banknoten
  - Schädel aus verschiedenen Gegenden Tibets

  $\rightarrow$ t-Test, Wilcoxon-Test
- Vergleich zweier abhängiger Stichproben
  Länge des Scheines oben und unten
  $\rightarrow$ Einstichproben t-Test, Vorzeichen-Wilcoxon-Test

Einleitung

Vergleich von Behandlungen, Grundgesamtheiten, Effekten

Abhängige und Unabhängige Stichproben

- Vergleich mehrerer unabhängiger Stichproben: Ägypt. Schädel: mehrere Grundgesamtheiten, Epochen
  $\rightarrow$ ANOVA, Kruskal-Wallis-Test
- Vergleich mehrerer abhängiger Stichproben Blutdruck von Patienten an mehreren aufeinanderfolgenden Tagen, (Faktoren: Patient, Tag)
  Preisrichter beim Synchronschwimmen
  $\rightarrow$ 2 fakt. Varianzanalyse, Friedman-Test

Einleitung

Ursache - Wirkungsanalysen

Ursache - Wirkungsanalysen

- Ursache - Wirkungsanalysen
  - Zusammenhangsanalyse
  - Bestimmen funktionaler Beziehungen
  - Trends, Vorhersagen
- Beispiele:
  - Bluthochdruck - Rauchgewohnheiten
  - Blutdruck - Proteinuria
  - Größe - Gewicht
  - Sterblichkeit - Wasserhärte

$\rightarrow$ Lineare, Nichtlineare und Nichtparametrische Regression

$\rightarrow$ Korrelation

Einleitung

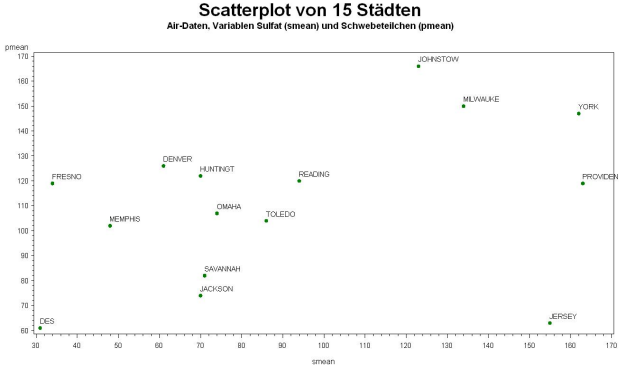
Klassifikation

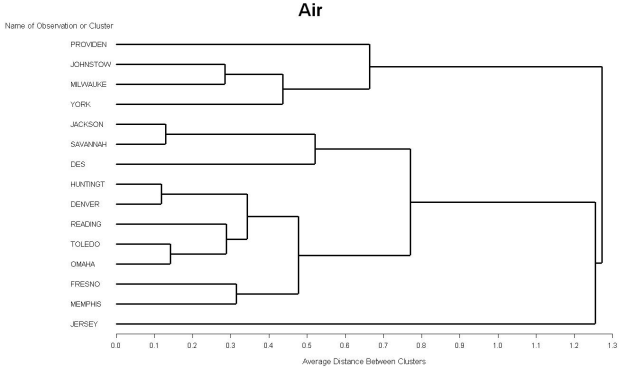
Klassifikation

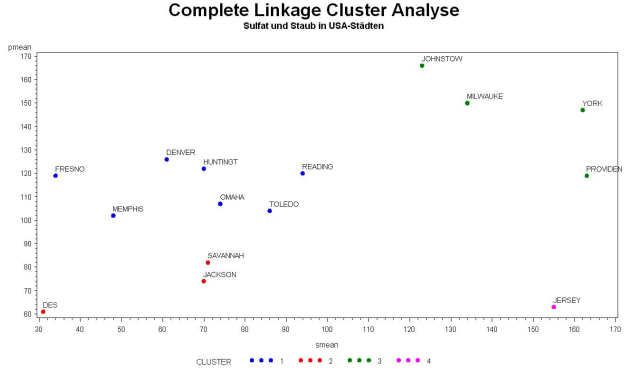
- Auffinden von Gruppen in Daten
  → Clusteranalyse
- Individuen sollen einer von vorgegebenen Klassen zugeordnet werden
  → Diskriminanzanalyse
  → Logistische Regression
- Datensatz hat Variablen, die mehr oder weniger voneinander abhängen.
  Welche Struktur besteht zwischen den Variablen?
  → Hauptkomponentenanalyse
  → Faktorenanalyse

Hierarchische Clusteranalyse

Beispiel: Luftverschmutzung in USA-Städten

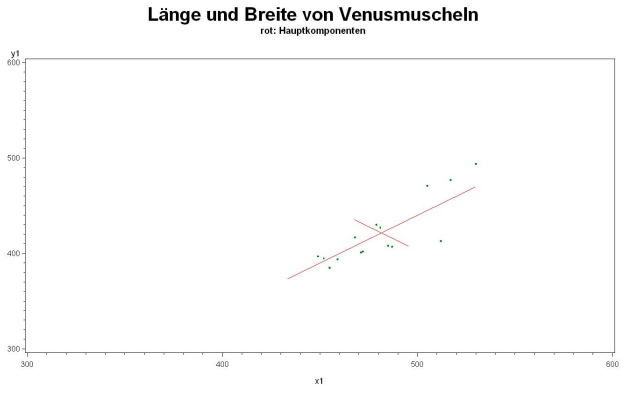


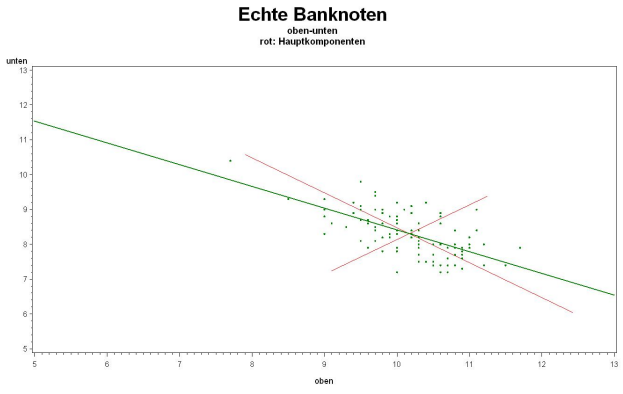




Hauptkomponentenanalyse

Beispiele





Frage: Wie kann man diese ausgezeichnete Richtung erfassen?

Einleitung

Literatur

Literatur (1)

Dufner, Jensen, Schumacher (1992). Statistik mit SAS, Teubner.

Falk, Becker, Marohn (1995). Angewandte Statistik mit SAS, Springer.

Graf, Ortseifen (1995). Statistische und grafische Datenanalyse mit SAS, Spektrum akademischer Verlag Heidelberg.

Krämer, Schoffer, Tschiersch (2004). Datenanalyse mit SAS, Springer.

SAS-Online Dokumentation, SAS-Handbücher

Einleitung

Literatur (2)

Hartung (1993). Statistik, Lehr- und Handbuch, Oldenbourg.

Sachs (1999). Angewandte Statistik, Springer.

Handl, A. (2002). Multivariate Analysemethoden, Springer.

Schlittgen, R. (2008). Einführung in die Statistik, Oldenbourg.

Backhaus, Erichsen, Plinke, Weiber (2010). Multivariate Analysemethoden, Springer.

Büning, Trenkler (1994). Nichtparametrische Statistische Methoden, DeGruyter Berlin.

Bortz, J. (1999). Statistik für Sozialwissenschaftler, Springer.

Einleitung

Statistik Software

Statistik-Software

SAS	-	sehr umfangreich, universell
	-	weit verbreitet
SPSS	-	umfangreich
	-	Anwendung vor allem in Biowiss., Medizin, Sozialwiss.
SYSTAT	-	ähnlich wie SPSS
	-	sehr gut
BMDP	-	umfangreich
S, S^+, **R**	-	funktionale Sprachen
	-	R: frei verfügbar

STATA, STATGRAPHICS, XPLORE, MATHEMATICA, MATLAB ..

Einleitung

Statistik Software (2)

	SAS	R
Umfang	+	+
Verfügbarkeit	+	++
Preis	(-)	++
Validierung	+	-
Dokumentation	+	-
Große Datensätze	+	-
User Community	+	+
Graphik		+
Kontinuität	+	Kern gut Zusatzpakete ?
Haftung	?	?
Erlernbarkeit	+	+

Mitschriften nach R. Vonk: KSFE 2010.

Einleitung

R auf den Informatikservern

R auf den Informatikservern

- Starten von R
  1. beim Sun-Server rabe (oder star ...) einloggen:
     `ssh -XC rabe` die Argumente bedeuten:
     -X: X-Weiterleitung (nur bei Plots nötig)
     -C: Kompression benutzen
  2. Start von R: R (interaktiv)
     oder `R --vanilla --slave < Quelltext.R`
  3. Laden von R-Quelltext: `source("Quelltext.R")`
- Beenden der Sitzung
  `q() oder CTRL-D`

Einleitung

R auf dem eigenen Rechner

R auf dem eigenen Rechner

- Linux
  - debian-basierte (u.a. Ubuntu): Paket r-base
  - Suse: R-patched, Fedora: R, Arch: r
  - teilweise neuere unter
    `http://cran.r-project.org/bin/`
- OS X:
  - R-....pkg unter
    `http://cran.r-project.org/bin/macosx/`
  - oder über Macports R, Fink: r-base, Homebrew: r
- Windows:
  `http://cran.r-project.org/bin/windows/base/`

Einleitung

R-Hilfe

R-Hilfe

- Hilfe zu Funktion/Paket: `?Name` oder `?"Name"`
- Suche in der gesamten Hilfe `??Begriff` oder `??"Begriff"`
- Suche im Hilfeartikel unter Unix mit / (falls Hilfe nicht im HTML-Modus)
- Autovervollständigung: (vermuteten) Namen beginnen, dann TAB drücken (hilfreich z.B. bei Verteilungen)
- Modus: `getOption("help_type")`, setzen: `options(help_type = "html")` (oder `"text"`)

Einleitung

Aufbau eines R-Programms

Grundlegende Syntax von R

- Zuweisung: `a = 10.2` oder gleichwertig `a <- 10.2`
- Vektorbildung: `c(c(1,2),c(1,2))` bildet Vektor $(1, 2, 1, 2)$
- arithmetische Op.: `+`,`*`,`^`,`%%`(modulo) etc. wirken bei Vektoren komponentenweise: `c(2,3) * c(2,2)` ergibt `c(4,6)`
- if (auch mit Ausdrücken!): `if(bed) ausd1 else ausd2` z.B. `if(5) 10 else 11` ergibt `10`
- for: `for(var in seq) ausd`
  Der Ausdruck sollte eine Anweisung sein (`print(var)`)

Einleitung

Aufbau eines R-Programms (2)

Grundlegende Syntax von R

- Eigene Funktion definieren: `function(arglist) ausd`
  Beispiel: `nachf = function(i) i+1`
- längere Funktionen mit {}:
  Beispiel: `{nachff = function(i) i+100;i+2}`
  Der letzte Ausdruck wird zurückgegeben: `nachff(2)` ist `4`.
- explizite Rückgabe durch `return(wert)`
- Funktionen aufrufen:
  Parameter werden durch Position oder Name festgelegt:
  `nachff(2)` oder `nachff(i=2)`

Einleitung

Aufbau eines R-Programms (3)

Grundlegende Syntax von R

- Kommentare: Zeilen mit # am Anfang
- Befehlsende: Newline oder ;
- Variablennamen: Umlaute etc. erlaubt
  Groß- und Kleinschreibung wird unterschieden!

Einleitung

Daten

Daten

Ausgangspunkt sind die Daten, die für die Analyse relevant sind.
Die Struktur der Daten hat die folgende allgemeine Form: x_{ij}

Objekte	Merkmale						
	1	2	3	..	j	..	p
1							
2							
3							
..							
i					x_{ij}		
..							
N							

Wert oder Ausprägung des Merkmals j am Objekt i

Einleitung

Daten (2)

Daten

p: Anzahl der Merkmale

N: Gesamtanzahl der einbezogenen Objekte (Individuen)

Objekte	Merkmale						
	1	2	3	..	j	..	p
1							
2							
3							
..							
i					x_{ij}		
..							
N							

Qualität des Datenmaterials wird im Wesentlichen durch die Auswahl der Objekte aus einer größeren Grundgesamtheit bestimmt.

Einleitung

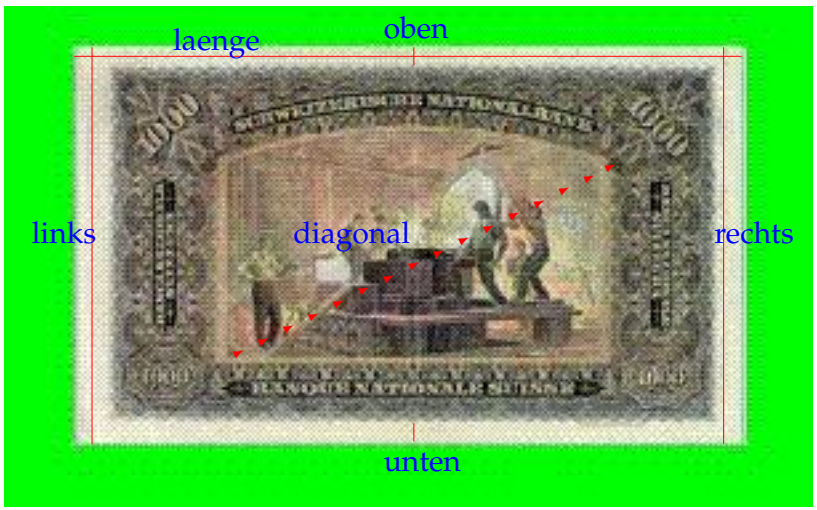
Daten (3)

Beispiele

- Objekte: Patienten einer Klinik
  Merkmale: Alter, Geschlecht, Krankheiten
- Objekte: Bäckereien in einer bestimmten Region
  Merkmale: Anzahl der Beschäftigten, Geräteausstattung, Umsatz, Produktpalette
- Objekte: Banknoten
  Merkmale: Längenparameter

Ein 1000-Franken Schein

1



Einleitung

Daten (4)

Datenmatrix

- Zeilen: Individuen, Objekte, Beobachtungen
- Spalten: Merkmalsausprägungen, -werte, -realisierungen

Banknote	Merkmale						
	laenge	oben	unten	..	j	..	gr
1							
2							
3							
..							
i					x_{ij}		
..							
N							

Einleitung

Daten (5)

Merkmale

- Definition: **Merkmale** sind Zufallsvariablen, die für jedes Individuum (Objekt) eine bestimmte Realisierung (Merkmalsausprägung) haben.
- Stetige Merkmale: laenge, oben
- Diskrete Merkmale: gr (Gruppe)

Banknote	Merkmale						
	laenge	oben	unten	..	j	..	gr
1							
2							
..							

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

2. Dateneingabe und Transformation

2.0 Datentypen in R

Vektor	alle Werte von einem Typ (u.a. double, integer,logical, character)	`c(...)`
Array	Vektor mit mehreren Dimensionen	`array(...)`
Matrix	zweidimensionales Array	`matrix(...)`
Liste	Werte können verschiedene Typen haben	`list(...)`
data.frame	Liste von Listen mit Spalten als Variablen und eindeutig benannten Zeilen (z.B. Probanden), wichtigste Datenstruktur für Statistik in R	`data.frame()` oder `as.data.frame()`

Inhalt

Dateneingabe und Transformation

2.1 Eingabe innerhalb des Programms

```
dfr = as.data.frame(rbind(
c(X=1,Y=2,Z=3),
c(11,2,3),
c(2,5,NA)
))
```

damit hat der data.frame `dfr` folgende Gestalt:

```
   X Y  Z
1  1 2  3
2 11 2  3
3  2 5 NA
```

`rbind` verknüpft zeilenweise (r: row) Vektoren zu einer Matrix
`NA` (not available) muss auch am Ende angegeben werden, sonst wird zyklisch aufgefüllt

Dateneingabe und Transformation

Eingabe innerhalb des Programms - alternativ

```
dfr =
   read.table(stdin(),col.names=c("X","Y","Z"))
1 2 3
11 2 3
2 5

#ab hier ggf. weiterer Quelltext
```

Achtung:

Funktioniert nur bei Einlesen über `R < Datei.R`, nicht über `source()`
Grund: keine Umleitung der Standardeingabe, `source` parsed komplette Datei vor Ausführung

Dateneingabe und Transformation

Eingabe innerhalb des Programms - alternativ

```
dfr =
   read.table(col.names=c("X","Y","Z"),text="
1 2 3
11 2 3
2 5
")
```

Funktioniert auch per `source()`.

Inhalt

Dateneingabe und Transformation

2.2 Direkte Eingabe durch Benuzter

Eingabe per Fenster oder Konsole

- `edit(data.frame(matrix(ncol=5)))` öffnet Fenster zum editieren
- `edit(Objekt)` öffnet Texteditor falls Objekt weder Matrix noch data.frame
- `read.table(file=stdin())` liest data.frame über Konsole ein
- `scan()` liest Vektor über Konsole ein (file=stdin() ist hier Standard)
- `scan` und `read.table` lesen über Konsole nur bis zur ersten Leerzeile $\Rightarrow$ dadurch auch Eingaben zwischen Quelltext möglich

Inhalt

Dateneingabe und Transformation

2.3 Zugriff auf einzelne Daten

Zugriff auf einzelne Daten

`dfr` beinhaltet 3 Variablen (X,Y,Z) mit je max. drei Beobachtungen (Individuen o.ä., noch unbenannt)

- Zeilen benennen: `rownames(dfr) = c("P1","P2","P3")`
- Zeilen/Spalten über Namen auswählen: `dfr["P1",]`,`dfr[,"Z"]`, `dfr["P1","Z"]`
- Zeilen/Spalten über Indizes auswählen (ab 1): `dfr[1,]`,`dfr[,3]`, `dfr[1,3]`
- Zeilen mit bestimmter Eigenschaft wählen (hier X < 10): `dfr[dfr[,"X"] < 10,]` oder `subset(dfr,X < 10)`

Inhalt

2.4 Eingabe durch externes File (plain text)

Der einfache Fall: read.table

`read.table(file, header, sep, quote, row.names, col.names, colClasses, nrows, skip, blank.lines.skip, stringsAsFactors, fileEncoding)` (und weitere Parameter)

- file : absoluter oder relativer Dateiname **oder ganze URL**
- header: Spaltennamen aus erster Zeile lesen?
- sep: Trennzeichen (Standard sind alle white spaces)
- quote: Anführungszeichen
- dec: Dezimaltrennzeichen (Standsrd ist .)
- col.names / row.names: Namen der Spalten und Zeilen
- colClasses: Vektor aus ïnteger", "numeric", "character", . . .

2.4 Eingabe durch externes File (plain text)

Der einfache Fall: read.table (Fortsetzung)

`read.table(file, header, sep, quote, row.names, col.names, colClasses, nrows, skip, blank.lines.skip, stringsAsFactors, fileEncoding)` (und weitere Parameter)

- nrows: Anzahl der zu lesenden Zeilen
- skip : Anzahl der am Anfang auszulassenden Zeilen
- blank.lines.skip: Leere Zeilen auslassen?
- stringsAsFactors: Strings werden als Faktoren codiert (effiziente Speicherung u.a. für Varianzanalyse, Strings lassen sich aber nicht mehr als solche verwenden)
- fileEncoding: Latin1, UTF-8 etc.

2.4 Eingabe durch externes File (plain text)

Der einfache Fall: read.table – ein Beispiel

```
banknote = read.table(file =
"http://www2.informatik.hu-berlin.de/
~koessler/SAS_Kurs/SAS_Vorlesung_Beispiele/
Vorles_Bsp/BANKNOTE.DAT",
colClasses=c("integer", "numeric", "numeric",
   "numeric", "numeric", "numeric", "numeric"),
col.names=c("nummer","laenge", "links",
   "rechts", "unten", "oben", "diagonal"),
row.names = 1)
```

2.4 Eingabe durch externes File (plain text)

Der einfache Fall: read.table – ein Beispiel (Fortsetzung)

	laenge	links	rechts	unten	oben	diagonal
1	214.8	131.0	131.1	9.0	9.7	141.0
2	214.6	129.7	129.7	8.1	9.5	141.7
3	214.8	129.7	129.7	8.7	9.6	142.2
4	214.8	129.7	129.6	7.5	10.4	142.0
5	215.0	129.6	129.7	10.4	7.7	141.8

⋮

`row.names = 1` nutzt die Spalte 1 (nummer) als Zeilenbennennung

`colClasses` ist hier nicht nötig, die automatische Bestimmung der Typen liefert das richtige Ergebnis

2.4 Eingabe durch externes File (plain text)

flexibler, aber unhandlicher: scan

```
as.data.frame(scan(file, what, ...))
```

- scan gibt nicht direkt einen data.frame zurück
- aber es kann **mehr als einen Datensatz je Zeile** lesen
- what : Liste aus Typen, wird zyklisch wiederholt, falls eine Zeile länger ist
  Achtung: `col.Classes=c("integer","logical")`
  entspricht `what=list(integer(),logical())`

Eingabe durch externes File (Fremdformate und Datenbanken)

Fremdformate und Datenbanken

- EXCEL,OpenDocumentSpreadsheet: verschiedene Pakete (speedR,gnumeric, RODBC(EXCEL unter Windows) . . .)
- Pakete für verschiedene Datenbanken: RODBC, RMySql, RSQlite . . .
- SAS/SPSS: foreign (meist instaliert): `read.xport(...)` bzw. `read.spss(...)`
- weitere Pakete (u.a. für neuere EXCEL-Dateien) sind nicht im CRAN, sondern über externe Projekte verfügbar (s. nächste Folie)

Einschub: Pakete und das CRAN

Pakete und das CRAN

- Viele Funktionen in Pakete ausgelagert, laden mit: `library(Paketname)`
- die meisten sind über das CRAN (Comprehensive R Archive Network, vgl. CTAN für TEX) zu finden: `http://cran.r-project.org`
- Installation mit `install.packages(Paketname)`
- Parameter repos für andere Quellen als CRAN, z.B. für das Omegaprojekt `install.packages(Paketname, repos="http://www.omegahat.org/R")`

Inhalt

2.5 Speichern,Laden,Löschen von Objekten

Speichern in .Rdata-Dateien

- `save(Objekt1,Objekt2,...,Objektn, file="Pfad.Rdata")` sichert die Objekte in einem für R schnell zu verarbeitenden Format
- Alternativ: `save(list=Namensliste, ...)` (Namensliste: `list("Objekt1",...)`) oder `save.image(file)` (sichert alle Objekte, wie ein `y` beim Beenden)
- `load(file,envir)` läd die Datei und fügt Objekte der Umgebung envir hinzu (normalerweise die globale Umgebung)

2.5 Speichern,Laden,Löschen von Objekten

Auflisten und Löschen von Objekten

- `ls()` gibt Liste aller Objekte der aktuellen Umgebung zurück
- `ls(.GlobalEnv)` gibt Liste aller Objekte der gloablen Umgebung zurück (nützlich in Fuktionen)
- `rm(Objekt1,...,Objektn,envir)` löscht die Objekte aus der (aktuellen) Umgebung, verhält sich mit Liste wie `save`

Inhalt

2.6 Zusammenfügen von data.frames

Der einfache Fall: rbind / cbind

- `rbind(...)` verknüpft Zeilenweise (s.o.)
- `cbind(...)` verknüpft Spaltenweise
- beide ex. auch für Vektoren, Listen und Matrizen $\Rightarrow$ Ergebnis ist nur data.frame, falls mind. ein Argument data.frame ist
- Zeilen- resp. Spaltennamen müssen/sollten verschieden sein
  Beispiel: Hat `d1` eine Spalte ID und `d2` ebenfalls, so hat `cbind(d1,d2)` zwei solche Spalten.

2.6 Zusammenfügen von data.frames

Der allgemeine Fall: merge

- `merge(x,y,by,by.x,by.y,sort,...)` verküpft **zwei** data.frames ähnlich einem join bei Datenbanken, d.h. über gemeinsame Schlüsselspalten
- by: Name der Spalten, falls in x und y identisch
- by.x,by.y: Namen der Spalten, falls verschieden z.B.: `by.x = "IdentNr",by.y="ID"`
- sort: Nach der by-Spalte sortieren?

2.6 Zusammenfügen von data.frames

Beispiel: cbind vs. merge

```
d1 = data.frame(cbind(X=c(2,5,4,1,3),
   Xsq=c(2,5,4,1,3)^2))
d2 = data.frame(cbind(Zahl=1:5,
   HochDrei=(1:5)^3))
```

```
  X Xsq
1 2   4
2 5  25
3 4  16
4 1   1
5 3   9
```

```
  Zahl HochDrei
1    1        1
2    2        8
3    3       27
4    4       64
5    5      125
```

2.6 Zusammenfügen von data.frames

Beispiel: cbind vs. merge (Fortsetzung)

```
dcbind = cbind(d1,d2)
dmerge = merge(d1,d2,by.x="X",by.y="Zahl")
```

dcbind

	X	Xsq	Zahl	HochDrei
1	2	4	1	1
2	5	25	2	8
3	4	16	3	27
4	1	1	4	64
5	3	9	5	125

dmerge

	X	Xsq	HochDrei
1	1	1	1
2	2	4	8
3	3	9	27
4	4	16	64
5	5	25	125

2.6 Zusammenfügen von data.frames

Sortieren: sort und order

- `sort`: Sortieren von Vektoren
  `sort(c(2,3,1))` ergibt Vektor `c(1,2,3)`
- `order`: Elementreihenfolge von Vekoren und Listen:
  `order(c(2,3,1))` ergibt Vektor `c(3,1,2)`,
  d.h. die **Permutation**, die `c(2,3,1)` in `c(1,2,3)` **überführt**.
- Sortieren eines data.frames `dfr` nach Spalte V:
  `dfr[order(dfr[,"V"]),]`
  Lies: Wähle in der Reihenfolge die Zeilen aus `dfr`,
  die eine Rangfolge der Spalte V aus `dfr` ist.

2.6 Zusammenfügen von data.frames

Beispiel: Einsatz von order

```
dnoso =merge(d1,d2,by.x="X",by.y="Zahl",sort=F)
dmerge = dmnoso[order(dnoso[,"X"]),]
```

dnoso

	X	Xsq	HochDrei
1	2	4	8
2	5	25	125
3	4	16	64
4	1	1	1
5	3	9	27

order(dnoso [,"X"]):

dmerge

	X	Xsq	HochDrei
1	1	1	1
2	2	4	8
3	3	9	27
4	4	16	64
5	5	25	125

```
[1] 4 1 5 3 2
```

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

3. Wahrscheinlichkeitsrechnung

3.1 Grundbegriffe

Eine Grundgesamtheit (oder Population)

ist eine Menge von Objekten, die gewissen Kriterien genügen. Die einzelnen Objekte heißen Individuen.

- Menge aller Haushalte
- Menge aller Studenten
- Menge aller Studenten der HUB
- Menge aller Einwohner von GB
- Menge aller Heroin-Abhängigen
- Menge aller Bewohner Tibets
- Menge aller verschiedenen Computer
- Menge aller Schweizer Franken
- Menge aller Wettkämpfer

Grundbegriffe

Zufällige Stichprobe

Die gesamte Population zu erfassen und zu untersuchen ist meist zu aufwendig, deshalb beschränkt man sich auf zufällige Stichproben.

Zufällige Stichprobe

Eine zufällige Stichprobe ist eine zufällige Teilmenge der Grundgesamtheit, wobei jede Stichprobe gleichen Umfangs gleichwahrscheinlich ist.
(oder: bei der jedes Element mit ‘der gleichen Wahrscheinlichkeit’ ausgewählt wird).

Bemerkung: Ein (auszuwertender) Datensatz ist (i.d.R.) eine Stichprobe.

Grundbegriffe

Klassifikation von Merkmalen

Nominale Merkmale

Die Ausprägungen sind lediglich Bezeichnungen für Zustände oder Sachverhalte.
Sie können auch durch Zahlen kodiert sein!

Bsp: Familienstand, Nationalität, Beruf

Dichotome Merkmale

Hat das (nominale) Merkmal nur 2 Ausprägungen, so heißt es auch binär oder dichotom.

gut - schlecht
männlich - weiblich
wahr - falsch

Klassifikation von Merkmalen

Ordinale und metrische Merkmale

Ordinale Merkmale (Rangskala)

Die Menge der Merkmalsausprägungen besitzt eine Rangordnung!

Rangzahlen einer Rangliste (z.B. beim Sport)
Härtegrade
Schulzensuren

Metrische Merkmale (kardinale/quantitative M.)

Werte können auf der Zahlengeraden aufgetragen werden (metrische Skala)

Messwerte, Längen, Größen, Gewichte, Alter

Klassifikation von Merkmalen

Metrische Merkmale

Metrische Merkmale werden unterschieden nach:

Diskrete Merkmale

nehmen höchstens abzählbar viele Werte an.

Alter, Länge einer Warteschlange

Stetige Merkmale

können Werte in jedem Punkt eines Intervalls annehmen, z.B. $x \in [a, b]$, $x \in (-\infty, \infty)$.

Metrische Merkmale sind immer auch ordinal.

Grundbegriffe

Stichprobenraum

Der Stichprobenraum Ω eines zufälligen Experiments

ist die Menge aller möglichen Versuchsausgänge
Die Elemente ω des Stichprobenraums Ω heißen Elementarereignisse.

- Münzwurf $\Omega = \{Z, B\}$
- Würfel $\Omega = \{1, ..., 6\}$
- Qualitätskontrolle $\Omega = \{$gut, schlecht$\}$
- Lebensdauer einer Glühlampe $\Omega = [0, \infty)$
- 100m - Zeit $\Omega =$[9.81, 20)
- Blutdruck, Herzfrequenz
- Länge einer Warteschlange $\Omega = \{0, 1, 2, ...\}$
- Anzahl der radioaktiven Teilchen beim Zerfall
- Wasserstand eines Flusses $\Omega = [0, \ldots)$

Grundbegriffe

Ein Ereignis ist eine Teilmenge $A, A \subseteq \Omega$

Lebensdauer ≤ 10 min.
Augensumme gerade.
Warteschlange hat Länge von ≤ 10 Personen.

Realisierungen sind die Ergebnisse des Experiments

(die realisierten Elemente von Ω)

Verknüpfungen von Ereignissen werden durch entsprechende Mengenverknüpfungen beschrieben

$A \cup B$	A oder B tritt ein
$A \cap B$	A und B tritt ein
$\overline{A} = \Omega \setminus A$	A tritt nicht ein.

Grundbegriffe

Ereignisfeld

Forderung (damit die Verknüpfungen auch immer ausgeführt werden können):
Die Ereignisse liegen in einem Ereignisfeld (σ-Algebra) $\mathfrak{E}$.

Ereignisfeld

Das Mengensystem $\mathfrak{E} \subseteq \mathfrak{P}(\Omega)$ heißt Ereignisfeld, falls gilt:

1. $\Omega \in \mathfrak{E}$
2. $A \in \mathfrak{E} \Longrightarrow \overline{A} \in \mathfrak{E}$
3. $A_i \in \mathfrak{E}, i = 1, 2, \ldots \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathfrak{E}$.

Inhalt

3.2 Wahrscheinlichkeit

Das Axiomsystem von Kolmogorov

Sei $\mathfrak{E}$ ein Ereignisfeld. Die Abbildung

$$P : \mathfrak{E} \longrightarrow \mathbb{R}$$

heißt Wahrscheinlichkeit, falls sie folgende Eigenschaften hat:

1. Für alle $A \in \mathfrak{E}$ gilt: $0 \leq P(A) \leq 1$.
2. $P(\Omega) = 1$.
3. Sei A_i eine Folge von Ereignissen, $A_i \in \mathfrak{E}$,

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i),$$

falls $\quad A_i \cap A_j = \emptyset \quad \forall i, i \neq j$

Wahrscheinlichkeit

Eigenschaften (1)

$$P(\overline{A}) = 1 - P(A).$$

Beweis:

$$\begin{aligned} 1 &= P(\Omega) && \text{Axiom 2} \\ &= P(A \cup \overline{A}) \\ &= P(A) + P(\overline{A}) && \text{Axiom 3} \end{aligned}$$

Wahrscheinlichkeit

Eigenschaften (2)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Beweis:

$$\begin{aligned}
P(A \cup B) &= P((A \cap B) \cup (A \cap \overline{B}) \cup (B \cap \overline{A})) \\
&= \underbrace{P(A \cap B) + P(A \cap \overline{B})}_{} \\
&\quad + P(B \cap \overline{A}) \quad \text{Axiom 3} \\
&= P(A) + \underbrace{P(B \cap \overline{A}) + P(A \cap B)}_{} \\
&\quad - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}$$

Inhalt

3.3 Zufallsvariablen

Eine (messbare) Abbildung heißt Zufallsvariable.

$$\begin{aligned} X: \quad \Omega &\longrightarrow \mathbb{R} \\ \omega &\longmapsto r \end{aligned}$$

Diskrete Zufallsvariable

Die Zufallsvariable X heißt diskret, wenn X nur endlich viele oder abzählbar unendlich viele Werte x_i annehmen kann. Jeder dieser Werte kann mit einer gewissen Wkt. $p_i = P(X = x_i)$ auftreten. $(p_i > 0)$

- geografische Lage (N,O,S,W)
- Länge einer Warteschlange
- Anzahl der erreichten Punkte in der Klausur.

Stetige Zufallsvariable

Stetige Zufallsvariable

Die Zufallsvariable X heißt stetig, falls X beliebige Werte in einem Intervall (a, b), $[a, b]$, $\overline{(a, b]}$, $(a, b]$, $(-\infty, a)$, (b, ∞), $(-\infty, a]$, $[b, \infty)$, $(-\infty, \infty)$ annehmen kann.

- Wassergehalt von Butter
- Messgrößen (z.B. bei der Banknote)
- Lebensdauer von Kühlschränken

Verteilungsfunktion

Diskrete Zufallsvariable

$$F_X(x) := P(X \leq x) = \sum_{i:i \leq x} p_i = \sum_{i=0}^{x} p_i$$

heißt Verteilungsfunktion der diskreten zufälligen Variable X

Manchmal wird die Verteilungsfunktion auch durch $P(X < x)$ definiert.

Stetige Zufallsvariable

Die Zufallsvariable X wird mit Hilfe der sogen. Dichtefunktion f beschrieben,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

Inhalt

3.4 Diskrete Zufallsvariablen

Bezeichnung

$$X \in \{x_1, x_2, x_3, \ldots\}$$

$$X : \begin{pmatrix} x_1 & x_2 & x_3 & \cdots x_n & \cdots \\ p_1 & p_2 & p_3 & \cdots p_n & \cdots \end{pmatrix}$$

$$p_i = P(X = x_i) > 0, \quad i = 1, 2, 3, \ldots$$

$$\sum_{i=1}^{\infty} p_i = 1$$

Diskrete Zufallsvariablen

Beispiele

Zweimaliges Werfen einer Münze

$\Omega = \{ZZ, ZB, BZ, BB\}$, $X :=$ Anzahl von Blatt

$$X : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Erfolge bei n Versuchen

X: Anzahl der "Erfolge" bei n Versuchen, wobei jeder der n Versuche eine Erfolgswahrscheinlichkeit p hat.

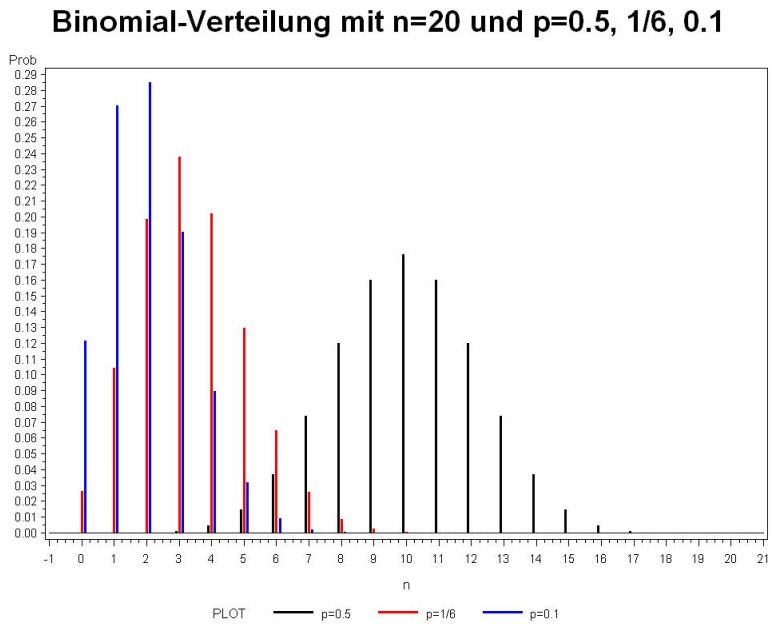
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{Binomialwkt.}$$

$$F_X(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \quad \text{Vf.}$$

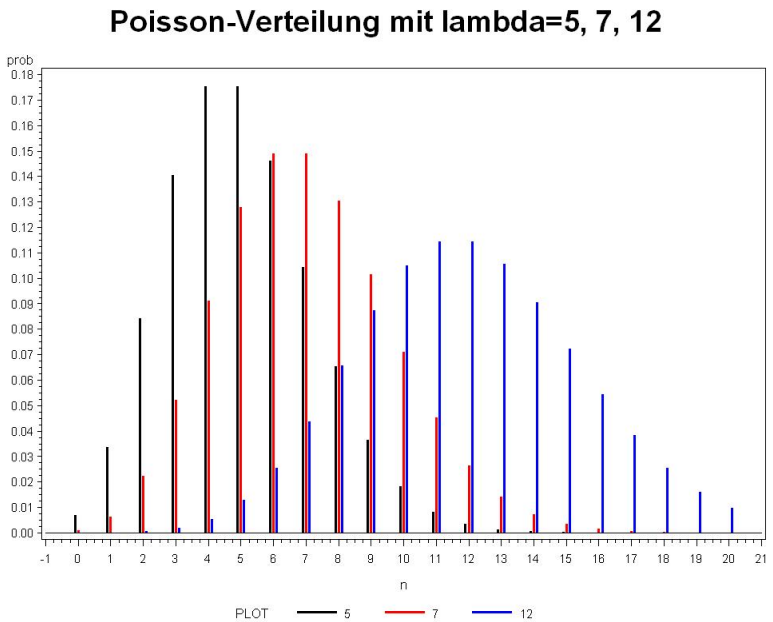
Diskrete Zufallsvariablen

Wahrscheinlichkeitsfunktionen

Binomial



Poisson



Diskrete Zufallsvariablen

Übungsaufgabe

Würfeln 20 mal. Wkt. für mindestens 4 Sechsen?

X: Anzahl der Sechsen.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F_X(3) = 1 - \sum_{i=0}^{3} P(X = i)$$

$$= 1 - \left(\frac{5}{6}\right)^{20} - 20\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{19} - \frac{20 \cdot 19}{2}\left(\frac{1}{6}\right)^{2}\left(\frac{5}{6}\right)^{18} -$$

$$- \frac{20 \cdot 19 \cdot 18}{6}\left(\frac{1}{6}\right)^{3}\left(\frac{5}{6}\right)^{17}$$

= `1 - pbinom(3,20,1/6)`

≈ 0.43.

Diskrete Zufallsvariablen

Poisson (1)

X: Anzahl der Anrufe pro Zeiteinheit

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \cdots \\ p_0 & p_1 & p_2 & p_3 & \cdots \end{pmatrix}$$

$$p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \quad \lambda > 0$$

$$\sum_{i=0}^{\infty} p_i = \underbrace{\sum_{i=0}^{\infty} \frac{\lambda^i}{i!}}_{e^{\lambda}} e^{-\lambda} = 1.$$

Bez.: $X \sim Poi(\lambda)$, wobei λ ein noch unbestimmter Parameter ist. Er kann als mittlere Rate aufgefaßt werden.

Diskrete Zufallsvariablen

Poisson (2), Motivation

Sei $\{N_t\}_{t \in T}$ eine Menge von Zufallsvariablen (ein stochastischer Prozess) mit den Eigenschaften:

- V1: Zuwächse sind unabhängig, dh. die Zufallsvariablen $N_{t+h} - N_t$ und $N_t - N_{t-h}$ sind unabhängig
- V2: es ist egal wo wir das Zeitintervall betrachten, dh. N_{t+h} und N_t haben dieselbe Verteilung
- V3: Wkt., daß mindestens ein Ereignis in der Zeit h eintritt, z.B. ein Kunde ankommt.
  $p(h) = a \cdot h + o(h), \quad a > 0, h \to 0$
- V4: Wkt. für $k \geq 2$ Ereignisse in der Zeit h: $o(h)$

Diskrete Zufallsvariablen

Poisson (3)

Frage: Wkt. bis zum Zeitpunkt t genau i Ereignisse?
(eingetroffene Kunden, zerfallene Teilchen)

$$P_k(t) := P(N_t = k), \qquad P_k(t) = 0 \quad \text{für} \quad k < 0$$

$$P_k(t) = \frac{a^k t^k}{k!} e^{-at}, \qquad k \geq 0$$

Poisson-Verteilung mit Parameter $\lambda = at$.

Beweis: Stochastik-Vorlesung.

Diskrete Zufallsvariablen

Poisson (4)

Binomial und Poisson

Seien $X_n \sim Bi(n,p)$ $\quad Y \sim Poi(\lambda)$

Für $n \cdot p = \lambda$ gilt: $\quad P(X_n = k) \longrightarrow_{n\to\infty} P(Y = k).$

Beweis:

$$\begin{aligned} P(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} (\frac{\lambda}{n})^k (1-\frac{\lambda}{n})^{n-k} \\ &= \frac{1}{k!} \underbrace{\frac{n(n-1)\cdots(n-k+1)}{(n-\lambda)^k}}_{\to 1} \lambda^k \underbrace{(1-\frac{\lambda}{n})^n}_{\to e^{-\lambda}} \end{aligned}$$

Diskrete Zufallsvariablen

Geometrische Verteilung

Münzwurf solange bis B(Blatt) kommt

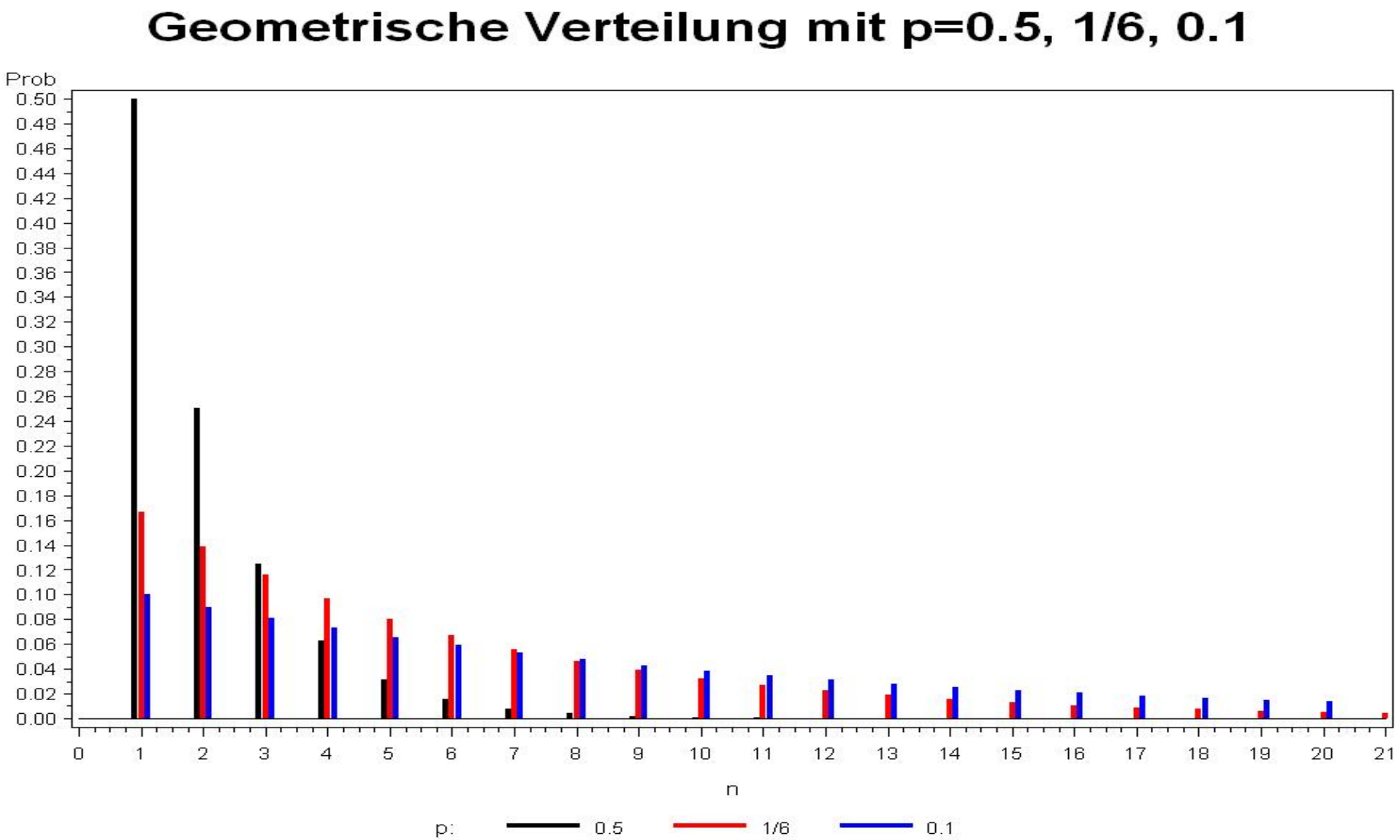
Ω = {B, ZB, ZZB, ...}
X := Anzahl der Würfe bis zum ersten Blatt.

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n & \cdots \\ (1/2) & (1/2)^2 & (1/2)^3 & (1/2)^4 & \cdots & (1/2)^n & \cdots \end{pmatrix}$$

$$\sum_{i=1}^{\infty} p_i = \sum_{i=1}^{\infty} (1/2)^i = \frac{1}{1-\frac{1}{2}} - 1 = 1$$ geometrische Reihe

geometrische Verteilung mit p=1/2, $p_i = (1/2)^i$.
allgemeiner: $p_i = p^{i-1}(1-p)$.

Geometrische Verteilung



Diskrete Zufallsvariablen

Hypergeometrische Verteilung (1)

Qualitätskontrolle

Warenlieferung mit N Stücken, davon genau n schlecht. Frage: Wkt., in einer Stichprobe vom Umfang m sind höchstens k Stück schlecht?

X: Anzahl der schlechten Stücke in der Stichprobe.

$$P(X = k) = \frac{\binom{n}{k} \cdot \binom{N-n}{m-k}}{\binom{N}{m}}$$

$\binom{N}{m}$: # möglichen Stichproben.

$\binom{n}{k}$: # Möglichkeiten, aus n schlechten Stücken in der Population k schlechte Stücke zu ziehen.

$\binom{N-n}{m-k}$: # Möglichkeiten, aus $N-n$ guten Stücken in der Population $m-k$ gute Stücke zu ziehen.

Diskrete Zufallsvariablen

Hypergeometrische Verteilung (2)

Offenbar: $0 \leq x \leq \min(n, m), \quad m - x \leq N - n.$

Eine Zufallsvariable mit der Verteilungsfunktion

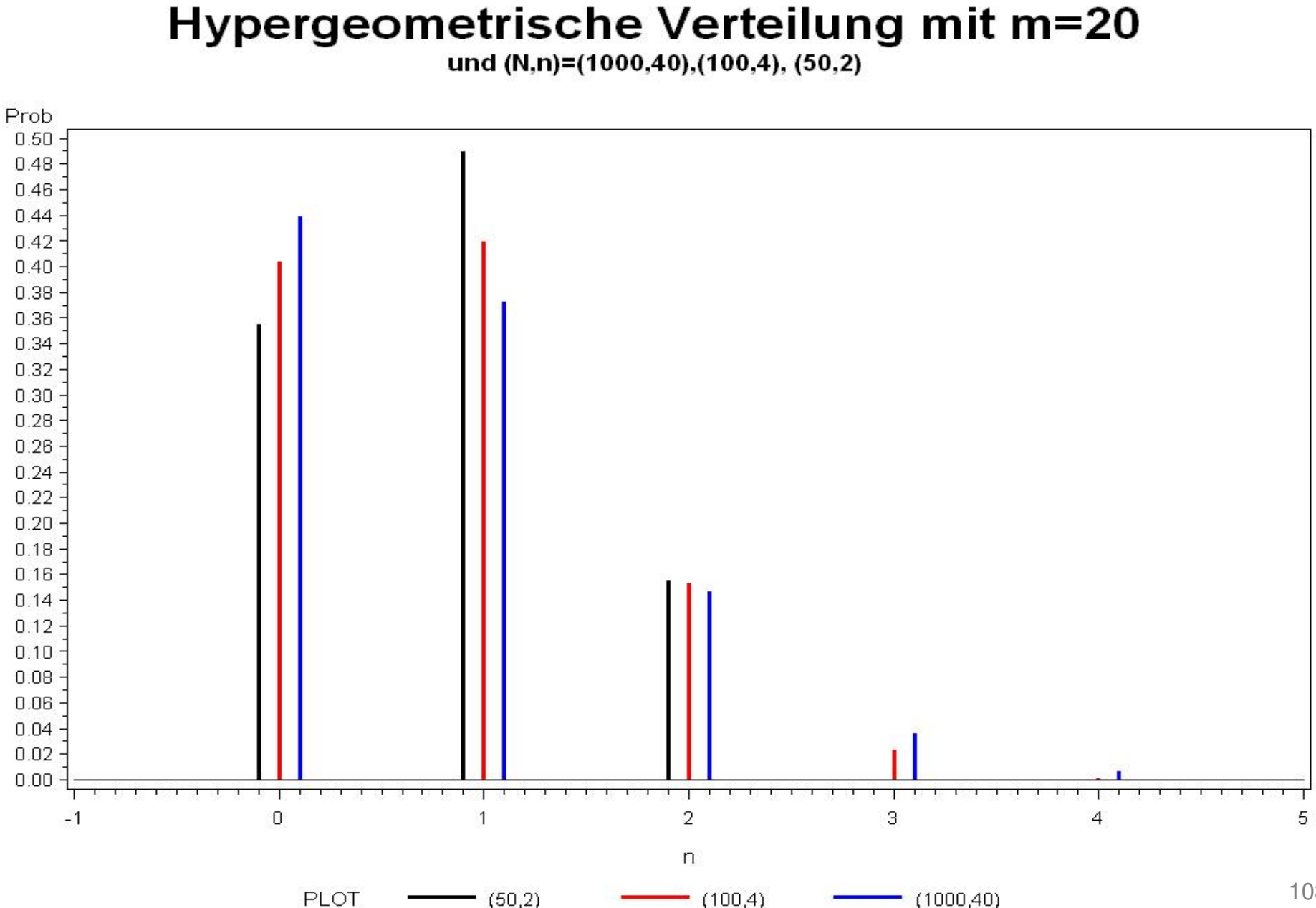
$$F(k|H_{N,n,m}) = \sum_{x=0}^{k} \frac{\binom{n}{x} \cdot \binom{N-n}{m-x}}{\binom{N}{m}}$$

heißt hypergeometrisch verteilt.

Bemerkung: Für $N \to \infty$, $n \to \infty$, $\frac{n}{N} \to p$ gilt:

$$f(x|H_{N,n,m}) \to \binom{m}{x} p^x (1-p)^{m-x} = f(x|Bi(m,p))$$

Hypergeometrische Verteilung



R-Anweisungen

Verteilungen	Dichten
`pbinom`(m,n,p)	`dbinom`(m,n,p)
`ppois`(m,λ)	`dpois`(m,λ)
`pgeom`(i,p)	`dgeom`(i,p)
`phyper`$(k,n,N-n,m)$	`dhyper`$(k,n,N-n,m)$

In den Wahrscheinlichkeiten können Parameter auftreten, die in der Regel unbekannt sind.

Die Parameter sind anhand der Beobachtungen (der Daten) zu bestimmen/zu schätzen!
$\longrightarrow$ Aufgabe der Statistik

Inhalt

3.5 Stetige Zufallsvariablen

Sei X stetig auf (a,b), wobei a, b unendlich sein können,
$a \leq x_0 < x_1 \leq b$
$P(X = x_0) = 0, \quad P(x_0 < X < x_1) > 0$ (wenn $f > 0$).

Die Funktion f heißt Dichtefunktion (von X) falls:

1. $f(x) \geq 0, \quad a < x < b.$
2. $\int_a^b f(x)\,dx = 1.$

Die stetige Zufallsvariable X wird also durch seine Dichtefunktion beschrieben.

$$P(c < X < d) = \int_c^d f(x)\,dx.$$

Die Dichtefunktion hängt i.A. von unbekannten Parametern ab, die geschätzt werden müssen.

Beispiele

Gleich- und Exponentialverteilung

Gleichverteilung auf [a,b], $X \sim R(a,b)$, $a < b$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{falls } a \leq x \leq b\ , \\ 0 & \text{sonst.} \end{cases}$$

- Referenzverteilung - Zufallszahlen

Exponentialverteilung, $X \sim Exp(\lambda)$, $(\lambda > 0)$

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & \text{falls } x \geq 0, \\ 0 & \text{sonst.} \end{cases} \qquad F(x) = \begin{cases} 0 & \text{falls } x \leq 0 \\ 1 - e^{-\frac{x}{\lambda}} & \text{falls } x > 0. \end{cases}$$

- Lebensdauer - Zeitdauer zwischen Ankünften

Beispiele

Exponentialverteilung (2)

Gedächtnislosigkeit

Eine Verteilung P (mit Verteilungsfunktion F) heißt gedächtnislos, wenn für alle $s, t \geq 0$, gilt:

$$P(X \geq s + t | X \geq t) = P(X \geq s).$$

Es gilt (Definition der bedingten Wahrscheinlichkeit)

$$\begin{aligned} P(X \geq s + t | X \geq t) &= \frac{P(\{X \geq s + t\} \cap \{X \geq t\})}{P(X \geq t)} \\ &= \frac{P(X \geq s + t)}{P(X \geq t)}. \end{aligned}$$

Gedächtnislosigkeit

Cauchy-Funtionalgleichung

Eine Verteilung ist also gedächtnislos, gdw.

$$\frac{P(X \geq s+t)}{P(X \geq t)} = P(X \geq s) \quad \text{gdw.} \quad \frac{1-F(s+t)}{1-F(t)} = 1 - F(s).$$

Überlebensfunktion (oder Zuverlässigkeitsfunktion)

$$G(t) = 1 - F(t)$$

Die Verteilungsfunktion F (mit der Überlebensfunktion G) ist also gedächtnislos gdw.

$$G(s+t) = G(s) \cdot G(t) \quad \text{für alle} \quad s, t \geq 0$$

Cauchy-Funktionalgleichung

Eine Lösung

Satz: Die Exponentialverteilung ist gedächtnislos.

Beweis: Die Verteilungsfunktion ist (sei $\lambda' := \frac{1}{\lambda}$)

$$F(t) = P(X < t) = \begin{cases} 1 - e^{-\lambda' t} & \text{falls} \quad t \geq 0 \\ 0 & \text{sonst,} \end{cases}$$

und die Überlebensfunktion

$$G(t) = 1 - F(t) = 1 - (1 - e^{-\lambda' t}) = e^{-\lambda' t}.$$

Folglich erhalten wir

$$G(s+t) = e^{-\lambda'(s+t)} = e^{-\lambda' s} e^{-\lambda' t} = G(s) \cdot G(t).$$

Cauchy-Funktionalgleichung

Die einzige Lösung

Satz:

Sei F eine stetige Verteilungsfunktion mit
$F(0) = 0$ und $G(t) = 1 - F(t)$.
Es gelte die Cauchy-Funktionalgleichung

$$G(s+t) = G(s) \cdot G(t) \quad \text{für alle} \quad s, t \geq 0.$$

Dann gilt für alle t, $t > 0$,

$$F(t) = 1 - e^{-\lambda t},$$

wobei $\lambda > 0$. D.h. F ist Exponential-Verteilungsfunktion.

Beweis: Stochastik-Vorlesung.

Beispiele

Normalverteilung (NV)

Dichtefunktion und Verteilungsfunktion

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)} \tag{1}$$

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \tag{2}$$

$(-\infty < x < \infty), \quad -\infty < \mu < \infty, \sigma^2 > 0.$

Bez.: $X \sim \mathcal{N}(\mu, \sigma^2)$, μ: Lageparameter, σ: Skalenparameter
Normalverteilung: wichtigste Verteilung in der Statistik
warum? $\longrightarrow$ später.

R-Funktionen

`dexp(x,` $\frac{1}{\lambda}$`)`	Dichtefunktion
`pexp(x,` $\frac{1}{\lambda}$`)`	Verteilungsfkt.
`dnorm(x,`μ,σ`)`	Dichtefunktion
`pnorm(x,`μ,σ`)`	Verteilungsfkt.
`qnorm(u,`μ,σ`)`	Quantilfunktion

Stetige Zufallsvariablen

Weitere wichtige Verteilungen

Weibull-Verteilung	`pweibull(x,a,`λ`)`
Gamma-Verteilung	`pgamma(x,a,`λ`)`
χ^2-Verteilung	`pchisq(`λ, ν`)`
t-Verteilung	`pt(x,`δ, ν`)`
F-Verteilung	`pf(x,`δ, ν_1, ν_2`)`

Die drei letzten Verteilungen werden vor allem bei statistischen Tests benötigt (später).

Descr_Weibull
Descr_Gamma

Wahrscheinlichkeitsverteilungen in R

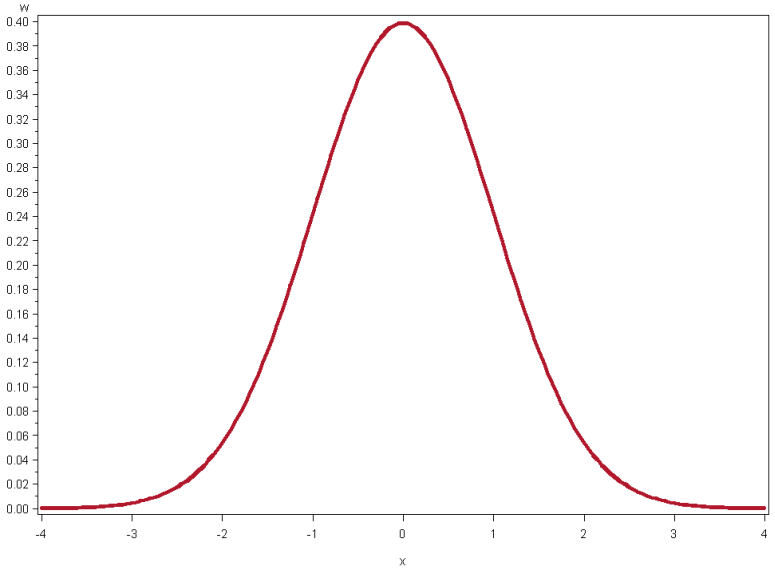
pVERT(x,Parameterliste)	Verteilungsfkt.
dVERT(x,Parameterliste)	Dichtefkt. (Wkt.fkt.)
qVERT(u,Parameterliste)	Quantilfkt.
rVERT(n, Parameterliste)	generiert pseudozuf. VERT-verteilten Vektor mit n Elementen

Autovervollständigung zum Finden nutzen!

Inhalt

3.6 Normalverteilung (1)

Dichtefunktion der Standard-Normalverteilung



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)}$$

Gauß

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Eine Zufallsvariable mit dieser Dichte $f(x)$ heißt normalverteilt mit Parametern μ und σ^2.

Normalverteilung (2)

Satz: f auf der letzten Folie ist Dichte.

Beweis: 1. $f(x) \geq 0 \; \forall x \in \mathbf{R}$ und $\sigma > 0$.

2. bleibt z.z.

$$\lim_{x \to \infty} F(x) = \int_{-\infty}^{\infty} f(t)\, dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt = 1.$$

Wir bezeichnen

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx =: I.$$

Normalverteilung (3)

Wir betrachten zunächst:

$$
\begin{aligned}
I^2 &= \left(\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right)^2 \\
&= \frac{1}{2\pi\sigma^2} \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right) \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \right) \\
&= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right) e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\
&= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dx\, dy
\end{aligned}
$$

Normalverteilung (4)

Substitution:

$$s := \frac{x-\mu}{\sigma} \qquad t := \frac{y-\mu}{\sigma}.$$

$$dx = \sigma\, ds \qquad dy = \sigma\, dt.$$

Wir erhalten damit:

$$\begin{aligned} I^2 &= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}s^2} e^{-\frac{1}{2}t^2} \sigma^2\, ds\, dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(s^2+t^2)}\, ds\, dt \end{aligned}$$

Normalverteilung (5)

Weitere Substitution (Polarkoordinaten):

$$s = r\cos\varphi \qquad t = r\sin\varphi.$$

Dann gilt allgemein nach der Substitutionsregel:

$$\int\int g(s,t)\,ds\,dt = \int\int g(r,\varphi)\det J\,dr\,d\varphi,$$

wobei hier:

$$\begin{aligned}
\det J = |J| &= \begin{vmatrix} \frac{\partial s}{\partial r} & \frac{\partial s}{\partial \varphi} \\ \frac{\partial t}{\partial r} & \frac{\partial t}{\partial \varphi} \end{vmatrix} \\
&= \begin{vmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{vmatrix} \\
&= r\cos^2\varphi + r\sin^2\varphi \\
&= r(\cos^2\varphi + \sin^2\varphi) = r
\end{aligned}$$

Normalverteilung (6)

$$
\begin{aligned}
I^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}(r^2 \cos^2 \varphi + r^2 \sin^2 \varphi)} r \, dr \, d\varphi \\
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2} r^2} r \, dr \, d\varphi \\
&= \frac{1}{2\pi} \int_0^{2\pi} \left[-e^{-\frac{r^2}{2}} \right]_0^{\infty} d\varphi \\
&= \frac{1}{2\pi} \int_0^{2\pi} d\varphi = \frac{1}{2\pi} 2\pi = 1
\end{aligned}
$$

Normalverteilung

Standard-Normalverteilung

$\mu = 0, \quad \sigma^2 = 1$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \qquad \text{Dichte}$$

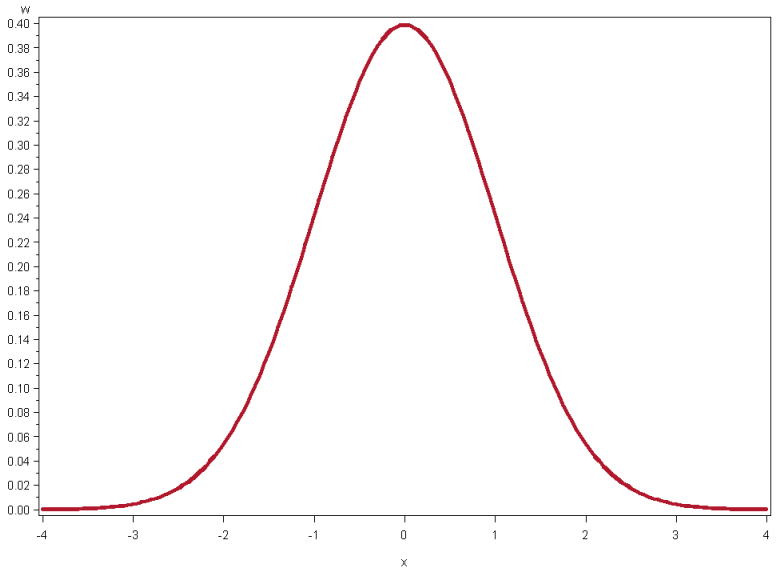
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \qquad \text{Verteilungsfunktion}$$

$\varphi(x), \Phi(x)$ sind tabelliert.

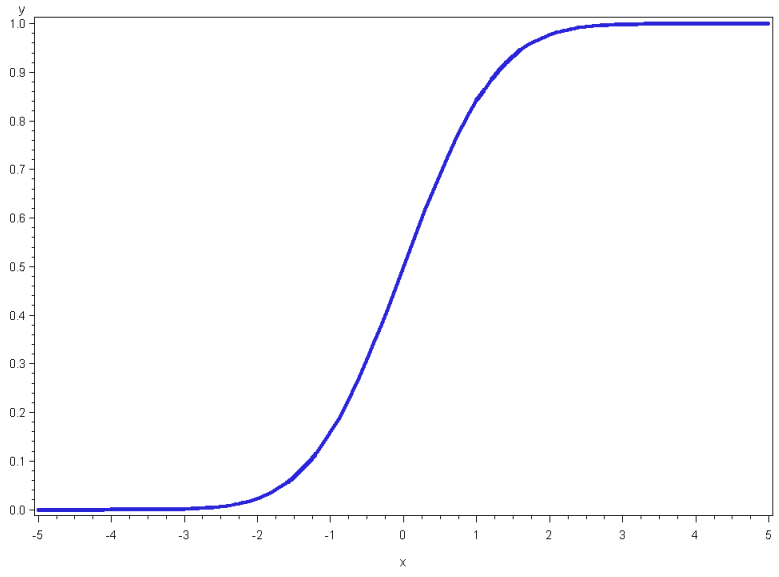
Es geht auch einfacher mit `pnorm` und `dnorm`.

Standardnormalverteilung (1)

Dichtefunktion der Standard-Normalverteilung



Verteilungsfunktion der Standard-Normalverteilung



$$\varphi(x) = \varphi(-x) \qquad \Phi(x) = 1 - \Phi(-x)$$

$$P(a < X < b) = \Phi(b) - \Phi(a)$$

Descr_normal.R

Standardnormalverteilung (2)

Frage: Für welches x gilt: $\Phi(x) = \alpha$?

$x = \Phi^{-1}(\alpha)$ α-Quantil.
$\Phi^{-1}(\alpha)$ als Funktion: Quantilfunktion

R: `qnorm(`α`)`

Normalverteilung

Beziehung zur Standard-Normalverteilung

Sei $X \sim \mathcal{N}(0,1)$. Dann $P(a < X < b) = \Phi(b) - \Phi(a)$.

Satz. Es gilt:

$$
\begin{aligned}
X \sim \mathcal{N}(0,1) &\iff \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2) \\
X \sim \mathcal{N}(\mu, \sigma^2) &\iff \alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2) \\
X \sim \mathcal{N}(\mu, \sigma^2) &\iff \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)
\end{aligned}
$$

Beweis: Wir zeigen nur 1. ($\rightarrow$). Sei $X \sim \mathcal{N}(0,1)$.

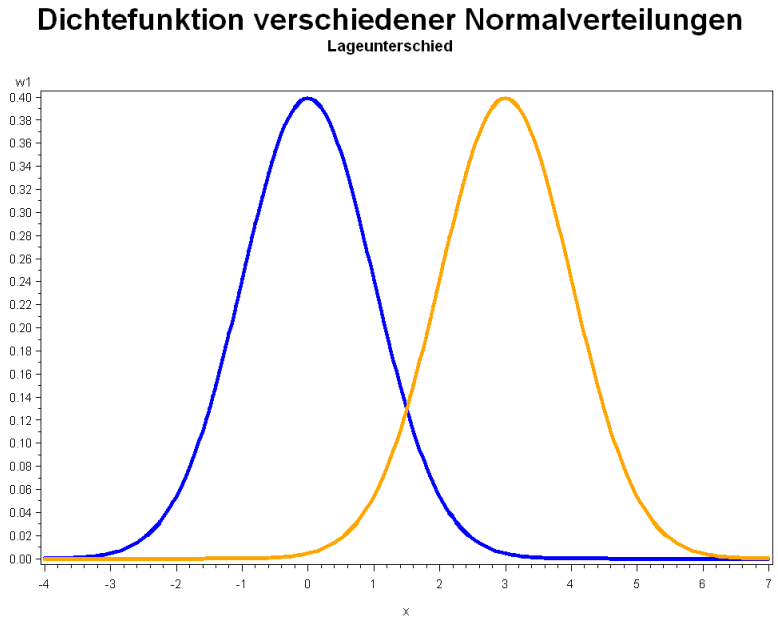
$$
\begin{aligned}
P(\sigma X + \mu \leq x) &= P(X \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}) = \\
= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(u-\mu)^2/(2\sigma^2)}\, du
\end{aligned}
$$

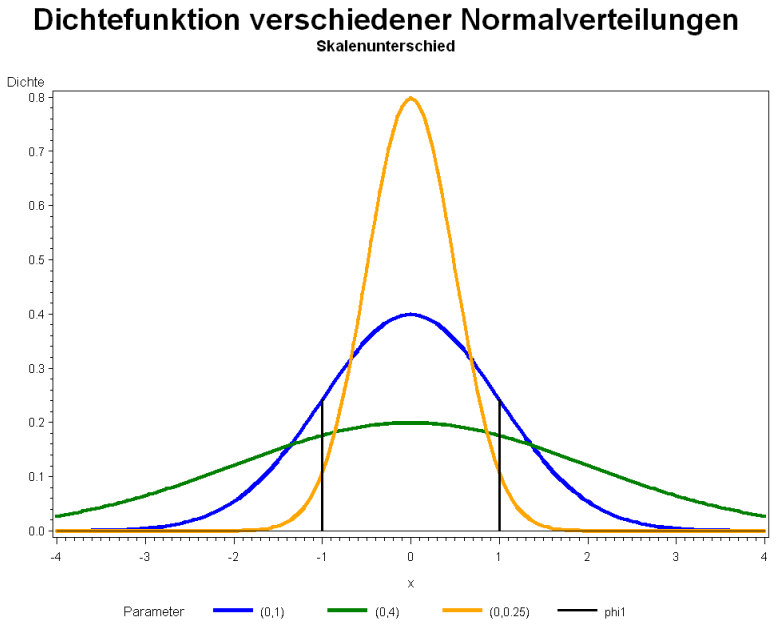
Normalverteilung

Unterschiedliche Parameter (1)

Vergleichen Sie

a) σ^2 fest, μ verschieden
b) μ fest, σ^2 verschieden





Descr_Normal_1.R

Normalverteilung

Unterschiedliche Parameter (2)

Satz: Seien $X_1 \sim \mathcal{N}(\mu, \sigma_1^2), X_2 \sim \mathcal{N}(\mu, \sigma_2^2)$,

$\sigma_1^2 < \sigma_2^2$ und $a > 0$. Dann gilt:

$$P(\mu - a < X_1 < \mu + a) > P(\mu - a < X_2 < \mu + a).$$

Beweis:

$$\begin{aligned} P(\mu - a < X_1 < \mu + a) &= P(\frac{-a}{\sigma_1} < \frac{X_1 - \mu}{\sigma_1} < \frac{a}{\sigma_1}) \\ &= \Phi(\frac{a}{\sigma_1}) - \Phi(-\frac{a}{\sigma_1}) \\ &> \Phi(\frac{a}{\sigma_2}) - \Phi(-\frac{a}{\sigma_2}) \\ &= P(\mu - a < X_2 < \mu + a). \end{aligned}$$

Normalverteilung

Beispiel: $X_1 \sim \mathcal{N}(10,4), X_2 \sim \mathcal{N}(10,9), a = 1.$

$$
\begin{aligned}
P(9 < X_1 < 11) &= \Phi(\frac{11-10}{2}) - \Phi(\frac{9-10}{2}) \\
&= \Phi(\frac{1}{2}) - \Phi(-\frac{1}{2}) = 2 \cdot \Phi(\frac{1}{2}) - 1 \\
&= 2 \cdot 0.6915 - 1 = 0.383.
\end{aligned}
$$

$$
\begin{aligned}
P(9 < X_2 < 11) &= \Phi(\frac{11-10}{3}) - \Phi(\frac{9-10}{3}) \\
&= \Phi(\frac{1}{3}) - \Phi(-\frac{1}{3}) = 2 \cdot \Phi(\frac{1}{3}) - 1 \\
&= 2 \cdot 0.6306 - 1 = 0.26112.
\end{aligned}
$$

Descr_Normal_3.R

Wahrscheinlichkeitsverteilungen

Zusammenfassung (1)

Diskrete Verteilungen

Binomial $X \sim B(n,p)$

X : Anzahl von “Erfolgen”, n Versuche, Erfolgswkt. p.

Poisson $X \sim Poi(\lambda)$

X : Anzahl von “Erfolgen”, n Versuche, Erfolgswkt. p,
n groß und p klein, $n \cdot p = \lambda$.
X : $\#$ Ankünfte in einem Zeitintervall.

Geometrisch, $X \sim Geo(p)$

X :: Zahl der Versuche bis zum ersten “Erfolg”.

Wahrscheinlichkeitsverteilungen

Zusammenfassung (2)

Stetige Verteilungen

Gleichverteilung $X \sim R(a, b)$

Zufallszahlen

Exponential $X \sim Exp(\lambda)$

“gedächtnislose” stetige Verteilung.

Normal $X \sim \mathcal{N}(\mu, \sigma^2)$

Zentraler Grenzwertsatz
Fehlergesetz (viele kleine unabhängige Fehler)

Inhalt

3.7 Erwartungswert

Einleitende Motivation

Eine Münze wird 3 mal geworfen.
Wie oft können wir erwarten, daß Blatt oben liegt?
Wie oft wird im Mittel Blatt oben liegen?

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1/8 & 3/8 & 3/8 & 1/8 \end{pmatrix}$$

Erwartungswert:
$0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1.5$
D.h. bei 10maliger Durchführung des Experiments können wir im Mittel mit 15mal Blatt rechnen!

Erwartungswert

Diskrete Zufallsvariable

Sei X diskrete Zufallsvariable

$$X : \begin{pmatrix} x_1 & \dots & x_n & \dots \\ p_1 & \dots & p_n & \dots \end{pmatrix}$$

$$\mathbf{E}X = \sum_{i=1}^{\infty} p_i x_i = \sum_{i=1}^{\infty} x_i p_i$$

heißt Erwartungswert von X.

Erwartungswert

$X \sim \text{Poisson}(\lambda)$

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & \dots \\ p_0 & p_1 & p_2 & p_3 & \dots \end{pmatrix} \qquad p_i = \frac{\lambda^i}{i!} e^{-\lambda}$$

$$\begin{aligned} \mathbf{E}X &= \sum_{i=0}^{\infty} p_i i \\ &= \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} e^{-\lambda} \cdot i \\ &= \lambda \underbrace{\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}}_{e^{\lambda}} e^{-\lambda} = \lambda. \end{aligned}$$

Interpretation: z.B. mittlere Ankunftsrate.

Erwartungswert

$X \sim Bi(n,p)$

$$
\begin{aligned}
\mathbf{E}X &= \sum_{k=0}^{n} k \binom{n}{k} p^k \cdot (1-p)^{n-k} \\
&= p \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= p \cdot n \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= p \cdot n \underbrace{\sum_{i=0}^{n-1} \binom{n-1}{i} p^i (1-p)^{n-1-i}}_{=1}, \qquad k = i+1 \\
&= n \cdot p.
\end{aligned}
$$

Erwartungswert

Stetige Verteilung

Sei X stetig mit Dichte f. Die Größe

$$\mathbf{E}X = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

heißt Erwartungswert von X.

$X \sim Exp(\lambda), \qquad \lambda > 0$

$$\mathbf{E}X = \int_{0}^{\infty} x \cdot \frac{1}{\lambda} \cdot e^{-\frac{x}{\lambda}} dx = \lambda$$

Erwartungswert

Normalverteilung

$X \sim \mathcal{N}(\mu, \sigma^2)$

$$
\begin{aligned}
\mathbf{E}X &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-(\frac{x-\mu}{\sigma})^2/2} dx \\
&= \int_{-\infty}^{\infty} (\sigma t + \mu) \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt \qquad \frac{x-\mu}{\sigma} = t, \quad dx = \sigma dt \\
&= \mu + \frac{1}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} \sigma \cdot t \cdot e^{\frac{-t^2}{2}} dt}_{=0} = \mu.
\end{aligned}
$$

Erwartungswert

Gleichverteilung

$X \sim R(a,b)$, gleichverteilt auf dem Intervall (a,b)

$$\begin{aligned} \mathbf{E}X &= \frac{1}{b-a}\int_a^b x dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

Erwartungswert

Eigenschaften des Erwartungswertes

$\mathbf{E}$ ist Linearer Operator

$$\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y.$$

Beweis: folgt aus Eigenschaften von Reihen und Integralen. ∎

Regel des Faulen Statistikers

Sei X Zufallsvariable, $g\colon \mathbb{R} \longrightarrow \mathbb{R}$ (rechtsseitig) stetig $\Rightarrow$

$$\mathbf{E}(g(X)) = \begin{cases} \sum_{i=0}^{\infty} g(x_i)p_i & \text{, falls X diskret} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{, falls X stetig,} \end{cases}$$

vorausgesetzt die Erwartungswerte existieren.

Beweis: Transformationsformel (s. Stochastik)

Inhalt

3.8 Die Varianz (Streuung)

Definition

Ang., die betrachteten Erwartungswerte existieren.

$$var(X) = \mathbf{E}(X - \mathbf{E}X)^2$$

heißt Varianz der Zufallsvariable X.

$$\sigma = \sqrt{Var(X)}$$

heißt Standardabweichung der Zufallsvariablen X.

Bez.: $var(X), Var(X), varX, \sigma^2, \sigma_X^2, \sigma, \sigma_X$.

Sei $\mu := \mathbf{E}X$.

Die Varianz

Stetige und diskrete Zufallsvariablen

Wenn X diskret, so gilt:

$$var(X) = \sum_{i=0}^{\infty} (x_i - \mu)^2 p_i$$

Wenn X stetig, so gilt:

$$var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx,$$

wobei f die Dichte von X ist.

$var(X)$: mittlere quadratische Abweichung von X und $\mathbf{E}X$.

Die Varianz

Eigenschaften der Varianz

$$
\begin{aligned}
var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}(X - \mu)^2 \\
&= \mathbf{E}(X^2 - 2\mu X + \mu^2) \\
&= \mathbf{E}X^2 - \mu^2 \\
var(aX + b) &= a^2 var(X), \quad a, b \in \mathbb{R}. \\
var(X) &= 0 \Longleftrightarrow \exists c : \quad P(X = c) = 1.
\end{aligned}
$$

Unabhängigkeit

Unabhängigkeit von Zufallsvariablen

Zwei Zufallsvariablen X und Y heißen unabhängig, falls

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$$

für alle $x, y \in \mathbb{R}$.

Zwei Ereignisse A und B heißen unabhängig, falls

$$P(A, B) = P(A) \cdot P(B)$$

X und Y sind also unabhängig gdw. die Ereignisse $X \leq x$ und $Y \leq y$ unabhängig sind für alle $x, y \in \mathbb{R}$.

Erwartungswert und Varianz

Eigenschaften

Seien X und Y stochastisch unabhängig. Dann

$$\mathbf{E}(X \cdot Y) = \mathbf{E}X \cdot \mathbf{E}Y.$$

Beweis: Übung

Seien X und Y unabhängig. Dann gilt

$$var(X + Y) = var(X) + var(Y).$$

Beweis: Übung

Die Varianz

Poisson-Verteilung

Wahrscheinlichkeitsfunktion

$$P(X=i)=\frac{\lambda^i}{i!}e^{-\lambda}, \quad i=0,1,2,\ldots \qquad \mathbf{E}(X)=\lambda$$

$$\begin{aligned} var(X) &= \mathbf{E}(X-\mathbf{E}X)^2=\sum_{i=0}^{\infty}(i-\lambda)^2 p_i \\ &= \sum_{i=2}^{\infty} i\cdot(i-1)p_i+\sum_{i=0}^{\infty} ip_i-2\lambda\sum_{i=0}^{\infty} ip_i+\lambda^2\sum_{i=0}^{\infty} p_i \\ &= e^{-\lambda}\lambda^2\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!}+\lambda-2\lambda^2+\lambda^2=\lambda. \end{aligned}$$

Die Varianz

Binomialverteilung, $X \sim B(n,p)$

Wahrscheinlichkeitsfunktion

$$P(X=k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

$$var(X) = np(1-p).$$

(ohne Beweis, ÜA)

Die Varianz

Gleichverteilung auf (a,b)

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a,b) \\ 0 & \text{sonst.} \end{cases} \qquad \mathbf{E}X = \frac{a+b}{2}.$$

$$\begin{aligned} \mathbf{E}X^2 &= \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{1}{3}x^3\Big|_a^b \cdot \frac{1}{b-a} \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}. \\ var(X) &= \mathbf{E}X^2 - (\mathbf{E}X)^2 = \frac{1}{12}(4a^2+4ab+4b^2-3a^2-6ab-3b^2) \\ &= \frac{1}{12}(a^2-2ab+b^2) = \frac{(b-a)^2}{12}. \end{aligned}$$

Die Varianz

Exponentialverteilung

Dichte

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & \text{falls } x \geq 0, \\ 0 & \text{sonst.} \end{cases}$$

$$\begin{aligned} \mathbf{E}X &= \lambda. \\ \mathbf{E}X^2 &= \int_0^\infty x^2 \frac{1}{\lambda} e^{-\frac{x}{\lambda}} \, dx = 2 \cdot \lambda^2 \quad \text{(ÜA)}. \\ var(X) &= \lambda^2. \end{aligned}$$

Die Varianz

Normalverteilung: $var(X) = \sigma^2$

$$
\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx \\
\mathbf{E}(X-\mu)^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\, dx \\
&= \sigma^2 \int_{-\infty}^{\infty} t^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = \sigma^2 \int_{-\infty}^{\infty} (-t)(-t\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}})\, dt \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(-te^{-t^2/2}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1) e^{-\frac{t^2}{2}}\, dt\right) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\, dt = \sigma^2.
\end{aligned}
$$

Bei Normalverteilung sind also die Parameter μ und σ^2 Erwartungswert und Varianz.

Inhalt

Formmaße (1)

(Theoretische) Schiefe

$$\beta_1 = \mathbf{E}\left(\frac{X - \mathbf{E}X}{\sqrt{var(X)}}\right)^3$$

β_1	=	0	falls F symmetrisch
β_1	$<$	0	falls F linksschief
β_1	$>$	0	falls F rechtsschief

ÜA: Berechnen Sie die (theoretische) Schiefe von

$$X : \begin{pmatrix} \frac{1}{2}(-4-\sqrt{6}) & -1 & \frac{1}{2}(-4+\sqrt{6}) & 2 & 3 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

und von

$$Y : \begin{pmatrix} -9 & -7 & 2 & 4 & 10 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

Formmaße (2)

(Theoretische) Wölbung, Kurtosis

$$\beta_2 = \mathbf{E}\left(\frac{X - \mathbf{E}X}{\sqrt{var(X)}}\right)^4 - 3$$

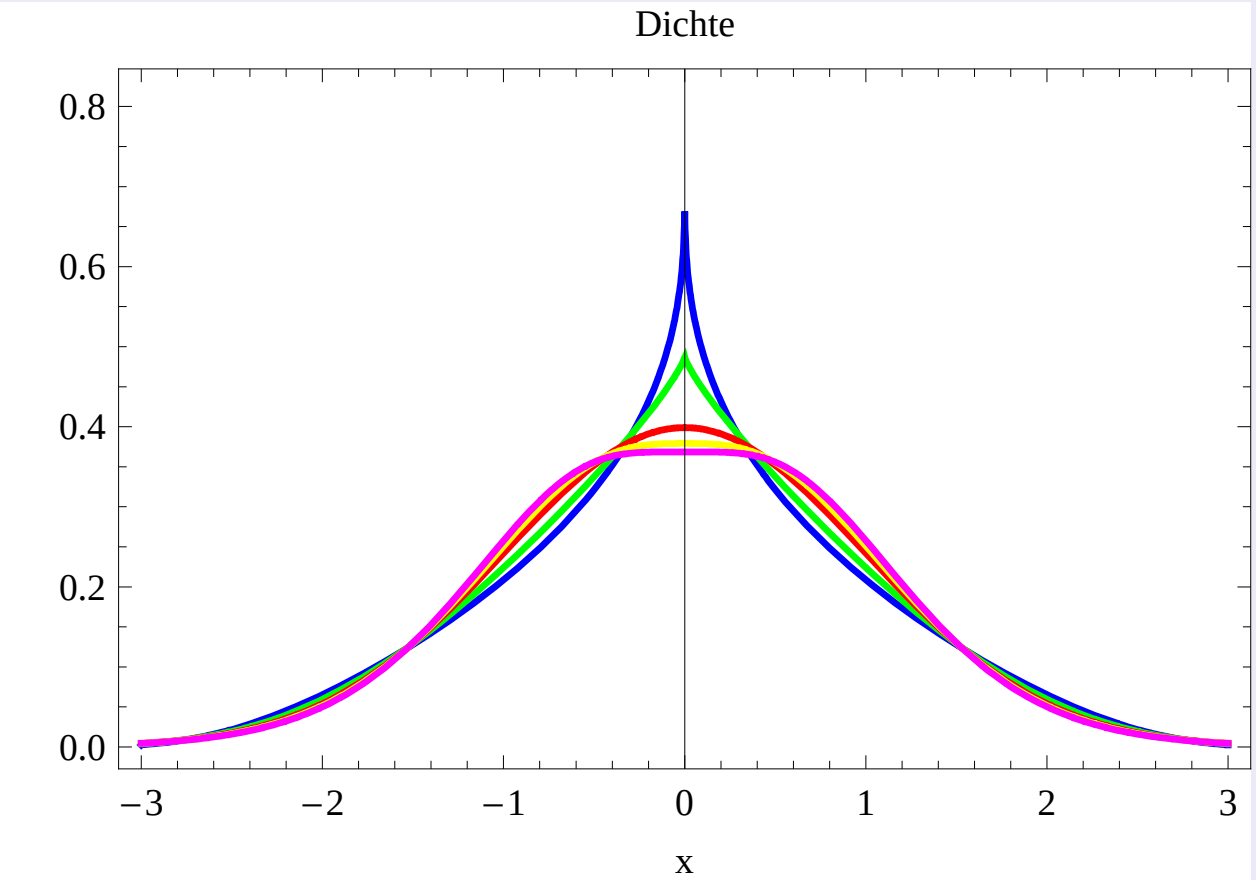
$\beta_2 = 0$ bei Normalverteilung
$\beta_2 > 0$ Tails “dicker, länger, stärker” als bei NV (?)
$\beta_2 < 0$ Tails “dünner, kürzer, schwächer” als bei NV (?)

$\beta_2 = 0$ heißt nicht notwendig: $F \sim$ Normal.

Formmaße (3)

Kurtosis

Dichten mit $\mathbf{E}(X) = 0, var(X) = 1, \beta_1 = 0, \beta_2 = 0$



Formmaße (4)

Theoretische Schiefe und Kurtosis verschiedener Verteilungen

Verteilung	Schiefe	Kurtosis
normal	0	0
gleich	0	-1.2
Doppelexp	0	3
Exponential	2	6
Bi(n,p)	$\frac{1-2p}{\sqrt{np(1-p)}}$	$-\frac{6}{n}+\frac{1}{np(1-p)}$
Poi(λ)	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$
Geo(p)	$\frac{2-p}{\sqrt{1-p}}$	$6+\frac{p^2}{1-p}$

Inhalt

3.9 Normalverteilung (2)

Besondere Eigenschaften

(schwaches) Gesetz der Großen Zahlen

Seien X_i unabhängig, identisch verteilt, $EX_i = \mu$

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to_p \mathbf{E}X$$

Zentraler Grenzwertsatz

Seien X_i unabhängig, identisch verteilt,
$EX_i = \mu, varX_i = \sigma^2$.

$$Z_n := \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \to Z, \quad Z \sim \mathcal{N}(0,1).$$

Descr_Binomial_2.R Descr_Exp.R

Normalverteilung

Fehlertheorie

Fehler sind unter folgenden Annahmen (asymptotisch) normalverteilt:

- Jeder Fehler ist Summe einer sehr großen Anzahl sehr kleiner, gleich großer Fehler, die verschiedene Ursachen haben.
- Die verschiedenen Fehlerkomponenten sind unabhängig.
- Jede Fehlerkomponente ist mit Wkt. 0.5 positiv und mit Wkt. 0.5 negativ.

Normalverteilung

Maximale Entropie (zur Information)

gegeben: Erwartungswert μ und Varianz σ^2

gesucht: Wahrscheinlichkeitsdichte f auf $(-\infty, \infty)$ mit

$$\int x f(x)\, dx = \mu, \quad \int (x-\mu)^2 f(x)\, dx = \sigma^2$$

und maximimaler Entropie:

$$H(f) := -\int f(x) \log f(x)\, dx$$

$\Longrightarrow$ f =Normaldichte.

Literatur: Rao: Lineare Statistische Methoden, 3.a.1.

Normalverteilung

Die Summe normalverteilter Zufallsvariablen

Die Summe normalverteilter Zufallsvariablen ist normalverteilt.

Seien $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Dann

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2).$$

(ρ: Korrelationskoeffizient zwischen X_1 und X_2, s.u.)

Beweis: über charakteristische Funktionen (Fouriertransformationen der Dichte) oder über die Faltungsformel (Stochastik-Vorlesung) oder über eine Verallg. des Satzes der Totalen Wkt.

Inhalt (1)

Inhalt (2)

Inhalt (3)

4. Statistische Maßzahlen für quantitative Merkmale

Inhalt

Lagemaße (Lokationsparameter)

Das arithmetische Mittel

Die angegebenen Maßzahlen sind empirisch, d.h. sie sind Schätzungen für die wahre (i.A. unbekannte) Lage.

Mittelwert (`mean`)

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

$\overline{X}_n \to_{n\to\infty} \mathbf{E}X$ Gesetz der Großen Zahlen.

Voraussetzungen:

a) X_i i.i.d., $\mathbf{E}X_i < \infty$ (Chintchin) oder

b) X_i beliebig, $\mathbf{E}X_i^2 < \infty$ (Tschebychev)

Lagemaße (2)

Quantile

Die Beobachtungen $x_1, \ldots, x_n$ werden der Größe nach geordnet:
$x_{(1)} \leq \ldots \leq x_{(n)}$.
Sei $0 \leq \alpha \leq 1, \quad \alpha \cdot n = \lfloor \alpha \cdot n \rfloor + r =: j + r.$

Quantile (Perzentile)

$$x_\alpha = \begin{cases} x_{(j+1)} \text{ für } r > 0 \\ 1/2(x_{(j)} + x_{(j+1)}) \text{ für } r = 0 \end{cases}$$

(empirisches) α-Quantil bzw. $\alpha \cdot 100\%$ Perzentil

mindestens $\lfloor \alpha \cdot n \rfloor$ der Werte $(x_1, \ldots, x_n)$ sind $\leq x_\alpha$
mindestens $\lfloor (1-\alpha)n \rfloor$ sind $\geq x_\alpha$

Vereinbarung: $x_0 := x_{(1)} \quad x_1 := x_{(n)}$

Bem.: x_α ist Schätzung von $F^{-1}(\alpha)$

Quantile

Beispiel

$$\begin{array}{ccccccccc} x_{(1)} & < & x_{(2)} & < & x_{(3)} & < & x_{(4)} & < & x_{(5)} \\ 1.5 & < & 2.7 & < & 2.8 & < & 3.0 & < & 3.1 \end{array}$$

$\alpha = 0.25$:

$\alpha \cdot n = 0.25 \cdot 5 = 1.25 = 1 + 0.25$

$\rightarrow x_\alpha = x_{0.25} = x_{(2)} = 2.7$

$\alpha = 0.75$:

$\alpha \cdot n = 0.75 \cdot 5 = 3.75 = 3 + 0.75$

$\rightarrow x_\alpha = x_{0.75} = x_{(4)} = 3.0$

$\alpha = 0.5$:

$\alpha \cdot n = 0.5 \cdot 5 = 2.5 = 2 + 0.5$

$\rightarrow x_\alpha = x_{0.5} = x_{(3)} = 2.8$

Lagemaße (3)

Median

ist das 0.5-Quantil $x_{0.5}$.

Quartile

heißen die 0.25- und 0.75-Quantile $x_{0.25}$ und $x_{0.75}$.

Modalwert

häufigster Wert

theoretischer Modalwert:
diskrete Merkmale: der wahrscheinlichste Wert
stetige Merkmale: Wert mit der größten Dichte

Lagemaße (4)

- Der Mittelwert ist in vielen Fällen eine 'gute' Lageschätzung, aber nicht robust (gegen Ausreißer).
- Der Median ist robust, aber meist nicht so 'gut'.

getrimmte Mittel, (α-)getrimmtes Mittel

$$\overline{X}_\alpha := \frac{x_{(\lfloor n\cdot\alpha\rfloor+1)} + \ldots + x_{(n-\lfloor n\cdot\alpha\rfloor)}}{n - 2\lfloor n\cdot\alpha\rfloor}, \qquad \alpha \in [0, \frac{1}{2})$$

Die $\lfloor n\cdot\alpha\rfloor$ kleinsten und $\lfloor n\cdot\alpha\rfloor$ größten Werte werden weggelassen und dann das arithmetische Mittel gebildet.

$\overline{X}_\alpha$ ist robuster als $\overline{X}$ und effizienter als $x_{0.5}$.

Lagemaße (5)

winsorisiertes Mittel, (α-)winsorisiertes Mittel

Sei $\alpha \in [0, \frac{1}{2})$ und jetzt $n_1 := \lfloor n \cdot \alpha \rfloor + 1$.

$$\overline{X}_{\alpha,w} := \frac{n_1 x_{(n_1)} + x_{(n_1+1)} + \ldots + x_{(n-n_1)} + n_1 x_{(n-n_1+1)}}{n}$$

Die $\lfloor n \cdot \alpha \rfloor$ kleinsten und $\lfloor n \cdot \alpha \rfloor$ größten Werte werden "herangeschoben" und dann das arithmetische Mittel gebildet.

- winsorisiertes Mittel ist robuster als $\overline{X}$ und effizienter als $x_{0.5}$.

Empfehlung für $\overline{X}_\alpha, \overline{X}_{\alpha,w}$: $\alpha : 0.1 \quad \ldots \quad 0.2$.

Lageschätzungen mit R

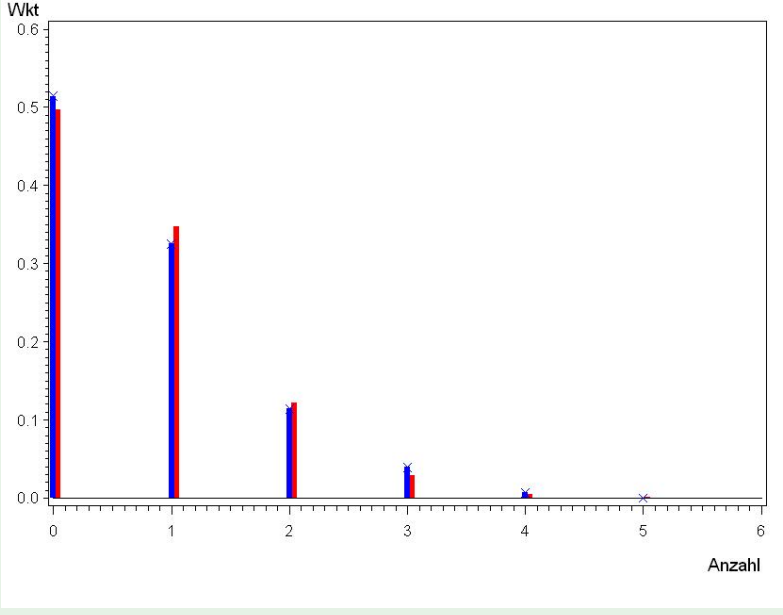
Mittelwert:	`mean()`
Median:	`median()`
getrimmte Mittel:	`mean(obj,trim=Anteil)`
abs. Anzahl Werte trimmen:	`mean(obj,trim=Anz/length(obj))`
winsorisierte Mittel:	`winsor.mean(obj,trim=Anteil)` aus Pakte `"psych"` verwenden?
Modalwert:	einige unsch. Mglkt.
Quartile:	`quantile()`;
bel. Quantile:	`quantile(obj,probs=c(0.33,0.9))` gibt 0.33 und 0.9-Quantile
Mittelw,Quartile und Median:	`summary(obj)`

Beispiele (1)

Tödliche Unfälle durch Pferdetritte

14 Corps, 20 Jahre, insges. 280 Einheiten. Erfasst wurde für jede Einheit die Anzahl der tödlichen Unfälle durch Pferdetritte.

Anzahl	Häufigkeit
0	144
1	91
2	32
3	11
4	2
5	0



Poisson-Verteilung geeignet (?)

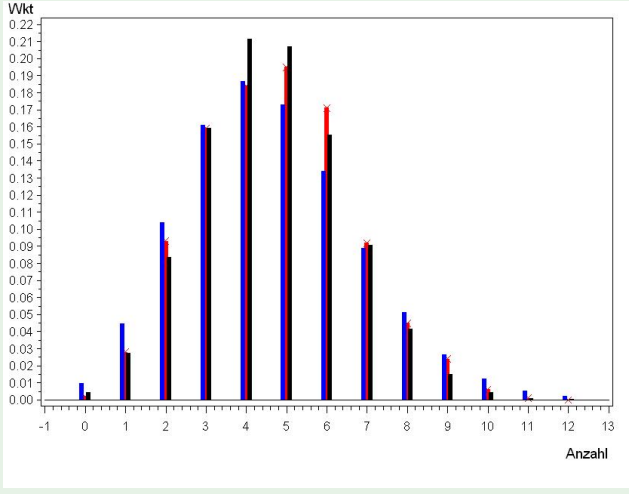
Schätzung von λ durch $\overline{X}$.

Beispiele (2)

Anzahl von schwarzen Feldern

Ein Zufallszahlengenerator soll zufällige Bildpunkte erzeugen, weiß mit Wkt. 0.71 und schwarz mit Wkt. 0.29.

Dazu wurde ein großes Quadrat in 1000 Teilquadrate mit je 16 Bildpunkten zerlegt. Gezählt wurde jeweils die Anzahl der schwarzen Bildpunkte.



n	0	1	2	3	4	5	6	7	8	9	10	11	12
h	2	28	93	159	184	195	171	92	45	24	6	1	0

Binomial-Verteilung (schwarz) geeignet (?)

Ang. p unbekannt. Schätzung von np durch $\overline{X}$.

Inhalt

Eigenschaften von Schätzungen (1)

Sei $\hat{\theta}_n$ eine Schätzung von θ, die auf n Beobachtungen beruht.

Konsistenz (Minimalforderung)

$\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta$

Erwartungstreue, Asymptotische Erwartungstreue

$\mathbf{E}\hat{\theta}_n = \theta$
$\mathbf{E}\hat{\theta}_n \to_{n\to\infty} \theta$

“gute”, “effiziente” Schätzung

var $\hat{\theta}_n$ möglichst klein

Eigenschaften von Schätzungen (2)

optimale Schätzung

wenn var $\hat{\theta}_n$ den kleinstmöglichen Wert annimmt für alle erwartungstreuen (e-treuen) Schätzungen.

Mean Square Error (MSE)

$$\begin{aligned} \text{MSE} &= \text{var}\ \hat{\theta}_n + \text{bias}^2\ \hat{\theta}_n \\ &= \text{var}\ \hat{\theta}_n + (E\hat{\theta}_n - \theta)^2 \end{aligned}$$

soll minimal oder möglichst klein sein.

robuste Schätzung

Eigenschaften sollten “möglichst” auch bei (kleinen) Abweichungen von der (Normal-) Verteilungsannahme gelten

Eigenschaften von Schätzungen (3)

Cramer-Rao Ungleichung

θ: zu schätzender Parameter einer Population (Dichte f).
$\hat{\theta} = \hat{\theta}_n$: eine erwartungstreue Schätzung von θ.

Cramer-Rao-Ungleichung

$$var(\hat{\theta}) \geq \frac{1}{n \cdot I(f,\theta)},$$

Fisher-Information

$$I(f,\theta) \quad = \quad \mathbf{E}\left(\frac{\partial \ln f(X,\theta)}{\partial \theta}\right)^2 = \int \left(\frac{\partial \ln f(x,\theta)}{\partial \theta}\right)^2 f(x,\theta)\, dx$$

Die Varianz einer Schätzung kann, bei gegebenem Stochprobenumfang, nicht beliebig klein werden.

Eigenschaften von Schätzungen (4)

Beispiele

f normal

$$f(x,\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned}
\ln f(x,\mu) &= -\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2} \\
\frac{\partial \ln f(x,\mu)}{\partial \mu} &= \frac{x-\mu}{\sigma} \cdot \frac{1}{\sigma} \\
I(f,\mu) &= \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^2 \cdot f(x,\mu)\, dx = \frac{1}{\sigma^2}.
\end{aligned}$$

Eigenschaften von Schätzungen (5)

Beispiele (2)

Nach der Cramer-Rao-Ungleichung gilt also für jede Lageschätzung

$$var(\hat{\theta}) \geq \frac{1}{nI(f,\theta)} = \frac{\sigma^2}{n},$$

insbesondere

$$var\overline{X} \geq \frac{\sigma^2}{n}.$$

Vergleichen Sie das mit:

$$var\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} varX_i = \frac{\sigma^2}{n}.$$

Bei Normalverteilung ist also $\overline{X}$ Lageschätzung mit minimaler Varianz.

Eigenschaften von Schätzungen (6)

Beispiele (3)

f exponential

$$f(x,\lambda) = \begin{cases} \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} & \text{falls} \quad x \geq 0 \\ 0 & \text{sonst.} \end{cases} \qquad I(f,\lambda) = \frac{1}{\lambda^2} \qquad \text{(ÜA)}$$

Die Cramer-Rao-Schranke ist also: $\dfrac{1}{nI(\lambda)} = \dfrac{\lambda^2}{n}$.

Vergleichen Sie mit: $\operatorname{var}\overline{X} = \dfrac{\lambda^2}{n}$.

Bei Exponentialverteilung ist also $\overline{X}$ Parameterschätzung mit minimaler Varianz.

Eigenschaften von Schätzungen (7)

Beispiele (4)

f Doppelexponential (=Laplace)

$$f(x,\lambda,\mu) = \frac{1}{2}\begin{cases} \frac{1}{\lambda}e^{-\frac{1}{\lambda}(x-\mu)} & \text{falls} \quad x \geq \mu \\ \frac{1}{\lambda}e^{\frac{1}{\lambda}(x-\mu)} & \text{falls} \quad x < \mu \end{cases}$$

Der hier interessierende (Lage-) Paramter ist μ.

$$I(f,\mu) = \frac{1}{\lambda^2}. \quad \text{(ÜA)} \qquad var(\overline{X}) = \frac{2\lambda^2}{n}. \quad \text{(ÜA)}$$

Für den Median $x_{0.5}$ gilt:

$$\text{var}(x_{0.5}) \sim \frac{\lambda^2}{n}. \quad \text{(ÜA}^*\text{)}$$

Inhalt

Schätzmethoden

Momentenmethode

Man drückt den zu schätzenden Parameter durch die Momente, z.B. $\mathbf{E}(X)$, aus.
Dann werden die Momente durch die entsprechenden *empirischen* Momente,
z.B. der Erwartungswert durch $\overline{X}$, ersetzt.

Maximum-Likelihood-Schätzung (ML-Schätzung)

Es wird der Schätzwert für den unbekannten Parameter ermittelt, der anhand der vorliegenden Daten, am meisten für diesen Paramter spricht (most likely).

Schätzmethoden

Kleinste-Quadrat-Schätzung (KQS)

Sei θ der zu schätzende Parameter. Man geht aus von einem Modell, z.B.

$$Y_i = g(\theta, X_i) + \epsilon_i$$

Dannn versucht man die Summe der Fehlerquadrate

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - g(\theta, X_i))^2.$$

zu minimieren (Kleinste Quadrate).

Momentenschätzung

Momentenschätzung bei Normalverteilung

Seien $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mu = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{\mu} = \overline{X}$$

$$\sigma^2 = \mathbf{E}(X - \mathbf{E}X)^2 \;\Rightarrow\; \hat{\sigma}^2 = \overline{(X_i - \overline{X})^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Momentenschätzung bei Exponentialverteilung

Seien $X_1, \ldots, X_n \sim Exp(\lambda)$.

$$\lambda = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{\lambda} = \overline{X}$$

Momentenschätzung

Momentenschätzung bei Binomialverteilung

Seien $X_1, \ldots, X_n \sim Bi(1, p)$.

$$p = \mathbf{E}X_i \quad \Longrightarrow \quad \hat{p} = \overline{X}$$

der relative Anteil der Realisierungen $x_i = 1$.

Maximum-Likelihood-Schätzung

ML-Schätzung bei Binomialverteilung

Beobachten n=1000 Jugendliche. Stichprobe $(X_1, \ldots, X_n)$

$X_i = 1$ falls Übergewicht festgestellt

$X_i = 0$ sonst.

Die Wkt., daß die beobachtete Stichprobe auftritt, wenn der Parameter p vorliegt ist

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$= \quad p^k(1-p)^{n-k}, \quad \text{wobei} \quad k = \sum_{i=1}^{n} x_i.$$

Maximum-Likelihood-Schätzung

Binomialverteilung

Der ML-Schätzer ist der Wert, der diese Funktion, $L_n(p)$, Likelihood-Funktion genannt, bzgl. p maximiert.
Maximieren statt $L_n(p)$: $\ln L_n(p)$ (Arg.Max. ist dasselbe).

$$\begin{aligned} \ln L_n(p) &= \ln\left(p^k(1-p)^{n-k}\right) \\ &= k \ln p + (n-k)\ln(1-p). \end{aligned}$$

Ableiten nach p und Nullsetzen liefert:

$$\frac{k}{p} - \frac{n-k}{1-p} = 0$$

Maximum-Likelihood-Schätzung

Binomialverteilung

Die einzige Lösung ist:

$$\hat{p} = \frac{k}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Für ein relatives Extremum in (0,1) kommt nur dieser Wert in Betracht.
Müssen aber noch die Likelihood-Funktion an den Rändern betrachten:
Für $p = 0$ und $p = 1$ wird $\ln L(p) = -\infty$. Also:

$$\hat{p}_{ML} = \frac{k}{n}.$$

Maximum-Likelihood-Schätzung

Normalverteilung, μ unbekannt, σ^2 bekannt

ML-Schätzung bei Normalverteilung

Likelihood: $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$, die gemeinsame Dichtefunktion der X_i.

Seien $X_1,\ldots,X_n$ unabhängig, $X_i \sim \mathcal{N}(\mu,1)$.

Likelihood:

$$\begin{aligned} L_n(\mu) &= \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \text{(Unabhängigkeit)} \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} \end{aligned}$$

Maximum-Likelihood-Schätzung

Normalverteilung, 2

$$\begin{aligned} \ln L_n(\mu) &= -n \ln(\sqrt{2\pi}) + \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2}\right) \\ \frac{\partial L_n(\mu)}{\partial \mu} &= \sum_{i=1}^{n} (x_i - \mu) \end{aligned}$$

Nullsetzen liefert die Maximum-Likelihood-Schätzung

$$\hat{\mu} = \overline{X}.$$

Maximum-Likelihood-Schätzung

Normalverteilung, μ und σ^2 unbekannt

$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $x_1, \ldots, x_n$: Beobachtungen

$$
\begin{aligned}
L_n(\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x_i - \mu)^2\Big) \\
&= \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\Big) \\
&= \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\Big(-\frac{nS^2}{2\sigma^2}\Big) \exp\Big(-\frac{n(\overline{X} - \mu)^2}{2\sigma^2}\Big)
\end{aligned}
$$

wobei $S^2 = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.
Die letzte Gleichung folgt aus:
$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \overline{X} + \overline{X} - \mu)^2 = nS^2 + n(\overline{X} - \mu)^2$

Maximum-Likelihood-Schätzung

Normalverteilung, Fortsetzung

Log-Likelihood:

$$\ln L(\mu,\sigma) = -\ln\sqrt{2\pi} - n\ln\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\sigma^2}$$

Lösen des Gleichungssystems

$$\begin{aligned} 0 &= \frac{\partial \ln L(\mu,\sigma)}{\partial\mu} = \frac{\overline{X}-\mu}{\sigma^2} \\ 0 &= \frac{\partial \ln L(\mu,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{nS^2}{\sigma^3} + \frac{n(\overline{X}-\mu)^2}{\sigma^3} \end{aligned}$$

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}^2 = S^2$$

Maximum-Likelihood-Schätzung

Gleichverteilung

ML-Schätzung bei Gleichverteilung auf $(0, \theta)$

Likelihood: $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$,
die gemeinsame Dichtefunktion der X_i.
Seien $X_1, \ldots, X_n$ unabhängig, $X_i \sim R(0, \theta)$, d.h.

$$f_{X_i}(x_i) = \begin{cases} \frac{1}{\theta} & \text{falls} \quad 0 \leq x_i \leq \theta \\ 0 & \text{sonst} \end{cases}$$

Maximum-Likelihood-Schätzung

Gleichverteilung, 2

Likelihood:

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \text{(Unabhängigkeit)} \\ &= \begin{cases} \frac{1}{\theta^n} & \text{falls} \quad 0 \leq x_i \leq \theta \quad \forall x_i \\ 0 & \text{sonst} \end{cases} \end{aligned}$$

Maximal, wenn $\theta \geq x_1, \ldots, x_n$, und wenn θ möglichst klein, also

$$\hat{\theta} = \max(x_1, \ldots, x_n).$$

Maximum-Likelihood-Schätzung

Gemischte Normalverteilung

Dichte ($\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, p)$):

$$f(x; \boldsymbol{\theta}) = (1-p)\phi\Big(\frac{x-\mu_1}{\sigma_1}\Big) + p\phi\Big(\frac{x-\mu_2}{\sigma_2}\Big)$$

$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ mit Wkt. $(1-p)$ und $X_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$ mit Wkt. $(1-p)$, aber p ist nicht bekannt.
Likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\Big((1-p)\phi(\frac{x_i-\mu_1}{\sigma_1}) + p\phi(\frac{x_i-\mu_2}{\sigma_2})\Big)$$

Maximieren des (log-)Likelihood $\rightarrow$ Newton-Raphson o. EM-Algorithmus (Stochastik-Vorlesung)

Eigenschaften von ML-Schätzern

Unter Regularitätsannahmen gilt

- ML-Schätzungen sind konsistent.
- sie sind (asymptotisch) effizient, d.h. sie haben minimale Varianz.
  Die Varianz ist durch die Cramér-Rao Ungleichung gegeben.
- sie sind asymptotisch normal verteilt (wichtig für die Konstruktion von Konfidenzintervallen, s.u.)
- Nachteil: ML-Schätzungen beruhen auf Verteilungsannahmen.

Kleinste Quadrat Schätzung

KQS des Lageparameters

Modell:

$$Y_i = \mu + \epsilon_i$$

Die Summe der Fehlerquadrate

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \mu)^2.$$

minimieren: Differenzieren und Nullsetzen liefert:

$$\hat{\mu}_{KQS} = \overline{Y}.$$

Kleinste Quadrat-Schätzung

KQS im einfachen linearen Regressionsmodell

$$Y_i = \theta_2 + \theta_1 X_i + \epsilon_i$$

$$f(X, \theta_1, \theta_2) = \theta_1 X + \theta_2$$

$$\frac{\partial f}{\partial \theta_1} = X \qquad \frac{\partial f}{\partial \theta_2} = 1$$

Minimieren von $\sum (Y_i - f(X_i, \theta_1, \theta_2))^2$ liefert:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (Y_i - (\theta_1 X_i + \theta_2)) \cdot X_i &= 0 \\ \frac{1}{n} \sum_{i=1}^{n} (Y_i - (\theta_1 X_i + \theta_2)) \cdot 1 &= 0 \end{aligned}$$

Kleinste Quadrat-Schätzung

$\Rightarrow$

$$\begin{aligned} \sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \theta_2 \sum_i X_i &= 0 \\ \sum_i Y_i - \theta_1 \sum_i X_i - \theta_2 \cdot n &= 0 \end{aligned}$$

Die zweite Gleichung nach θ_2 auflösen:

$$\theta_2 = \frac{1}{n} \sum_i Y_i - \theta_1 \frac{1}{n} \sum_i X_i$$

und in die erste einsetzen:

Kleinste Quadrat-Schätzung

$$\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \frac{1}{n} \sum_i Y_i \sum_i X_i + \theta_1 \frac{1}{n} \sum_i X_i \sum_i X_i = 0$$

$$\sum_i X_i Y_i - \frac{1}{n} \sum_i Y_i \sum_i X_i - \theta_1 \Big((\sum_i X_i^2 - \frac{1}{n} \sum_i X_i \sum_i X_i \Big) = 0$$

$\Rightarrow$

$$\begin{aligned} \hat{\theta}_1 &= \frac{\sum_i X_i Y_i - \frac{1}{n} \sum_i X_i \sum_i Y_i}{\sum_i X_i^2 - \frac{1}{n} (\sum_i X_i)^2} = \frac{S_{XY}}{S_X^2} \\ \hat{\theta}_2 &= \frac{1}{n} \Big(\sum_i Y_i - \hat{\theta}_1 \sum_i X_i \Big) \end{aligned}$$

Einschub: Die Funktion plot & Co

(vgl. ÜA 9)

Darstellung von diskreten Verteilungen

`plot(x,y,type,main,xlab,ylab,col,axes)`

- x: Vektor aus Ordinaten (Abzissenwerte: 1 bs `length(x)`),
  x: Koordinaten der Punkte oder
  x: Abzissenwerte und y: Ordinatenwerte
- type: Nadelplot: ”h”, Punkteplot: ”p”,Linien(verb.Punkte): ”l”
- xlab, ylab: Beschriftung der Achsen
- col: Farbe der Punkte, Linien oder Nadeln
- main, sub: Haupt- und Untertitel
- axes: Achsen zeichen? (nachträglich mit `axes()`)

Einschub: Die Funktion plot & Co

(vgl. ÜA 9)

Hinzufügen zum Plot

`lines(x,y,type,main,xlab,ylab,col,axes)`

- `plot` erstellt immer neue (Teil-)Grafik
- zum Hinzufügen in bestehende: `lines` oder `points`
- einziger Unterschied: Standard für type: `"p"` bei `points`, `"l"` bei `lines`

Einschub: Die Funktion plot & Co

(vgl. ÜA 9)

Darstellung von Funktionen und Dichten

`curve(expr, from,to,n, add,...)`

- expr: Funktionsname oder Ausdruck, in dem `x` vorkommt z.B. `curve(x*sin(x))`
- from, to: Intervall auf der Abszisse für das gezeichnet wird (alternativ Parameter `xlim=c(from,to)`)
- n: Anzahl der Stützstellen
- add: Hinzufügen zu bestehendem Plot? (sonst neuer)

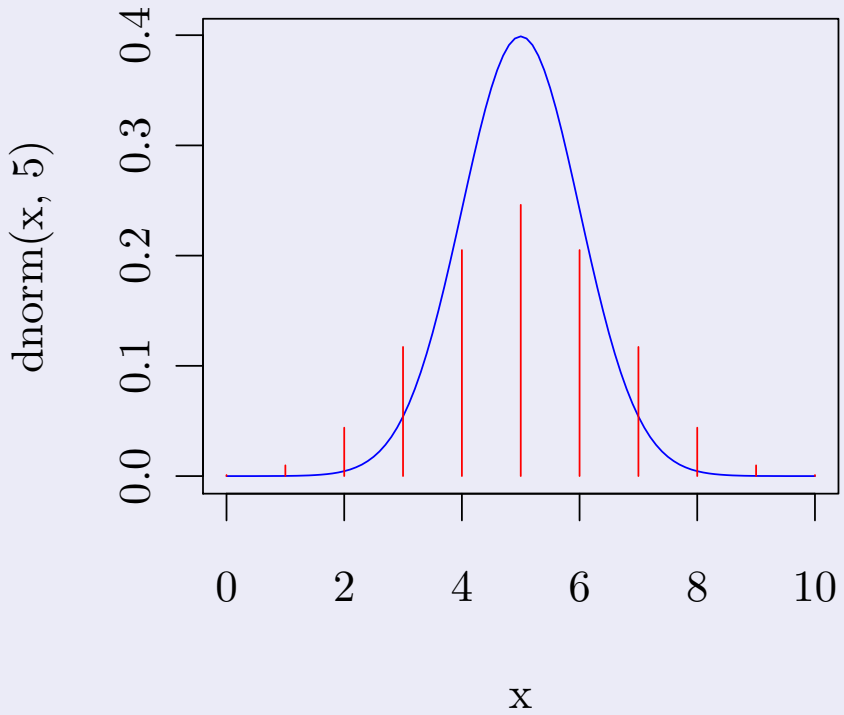
`plot(Funktionsname,from,to)` ist weniger flexibel

Einschub: Die Funktion plot & Co

(vgl. ÜA 9)

Beispiel

```
curve(dnorm(x,5),xlim=c(0,10),col="blue")
lines(0:10,dbinom(0:10,10,1/2),
   type="h",col="red")
```



Inhalt

Streuungsmaße

Die angegebenen Maßzahlen sind empirisch, d.h. sie sind Schätzungen für die wahre Varianz

(empirische) Varianz (Streuung)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2$$

$$s^2 \rightarrow_{n \rightarrow \infty} var(X)$$

Warum Division durch $(n-1)$: Erwartungstreue (ÜA)

Standardabweichung

$$s = \sqrt{s^2}$$

Streuungsmaße (2)

Spannweite (Range)

$$x_{(n)} - x_{(1)}$$

(Inter-)Quartilsabstand, IR

$$IR = x_{0.75} - x_{0.25}$$

Wenn $X \sim \mathcal{N}$ so $\mathbf{E}(IR/1.34898) = \sigma$.

Mittlere absolute Abweichung vom Median

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{0.5}|$$

Streuungsmaße (3)

Median absolute deviation, MAD

$$MAD = med(|x_i - x_{0.5}|)$$

Wenn $X \sim \mathcal{N}$ so $\mathbf{E}(1.4826 \cdot MAD) = \sigma$

Variationskoeffizient

$$CV = \frac{s \cdot 100}{\overline{X}}$$

Gini's Mean Difference

$$G = \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j| \qquad X \sim \mathcal{N} \Rightarrow \mathbf{E}\left(\frac{\sqrt{\pi}}{2} G\right) = \sigma$$

Streuungsmaße (4)

S_n und Q_n (Croux, Rousseuw 1992, 1993)

$$
\begin{aligned}
S_n &= 1.1926 \cdot med_i(med_j|x_i - x_j|) \\
Q_n &= 2.219 \cdot \{|x_i - x_j|, i < j\}_{(k)}
\end{aligned}
$$

$k = \binom{h}{2}, h = \lfloor \frac{n}{2} \rfloor + 1$

$\{...\}_{(k)}$ bezeichnet das k-größte Element der Menge.

- Die konstanten Faktoren sichern Erwartungstreue bei Normalverteilung, $X \sim \mathcal{N}$:
  $\Rightarrow \mathbf{E}(S_n) = \mathbf{E}(Q_n) = \sigma$

Streuungsmaße (5)

Eigenschaften:

- Varianz und Standardabweichung und Spannweite sind nicht “robust”.
- IR und MAD sind robust.
  (MAD etwas besser da höherer “Bruchpunkt”)
- G ist bedingt robust, effizient bei F normal.
- IR und MAD sind wenig effizient.
  (0.37 bei Normal)
- S_n oder Q_n sind geeignetste Schätzungen.

Streuungsmaße (6)

Nicht-Robuste Skalenschätzungen

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \\
\text{Range} &= x_{(n)} - x_{(1)} \\
CV &= \frac{s \cdot 100}{\overline{X}}
\end{aligned}
$$

Streuungsmaße (7)

Robuste Skalenschätzungen

$$
\begin{aligned}
IR &= x_{0.75} - x_{0.25} \\
MAD &= med(|x_i - x_{0.5}|) \\
G &= \frac{1}{\binom{n}{2}} \sum_{i<j} |x_i - x_j| \\
S_n &= 1.1926 \cdot med_i(med_j|x_i - x_j|) \\
Q_n &= 2.219 \cdot \{|x_i - x_j|, i < j\}_{(k)} \\
& \quad k = \binom{h}{2}, h = \lfloor \frac{n}{2} \rfloor + 1
\end{aligned}
$$

Streuungsmaße mit R

emp. Standardabw.:	`sd(obj)`
Range:	`diff(range(obj))`
CV:	`100*sd(obj)/mean(obj)`
IR:	`IQR(obj)`
	`c(0.25,0.75), names=F))`
MAD:	`mad(obj)`
S_n,Q_n:	`Sn(obj)`,`Qn(obj)`
	aus Paket `"robustbase"`
	verwenden?
G:	`gini.mean.diff()`;
	aus Paket `"lmomco"`
	verwenden?

Inhalt

Formmaße (1)

(Theoretische) Schiefe

$$\beta_1 = \mathbf{E}\left(\frac{X - \mathbf{E}X}{\sqrt{var(X)}}\right)^3$$

(Empirische) Schiefe

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{s}\right)^3$$

R : `beta1 = mean(((X-mean(X)/sd(X))^3)`

Formmaße (2)

(Theoretische) Wölbung, Kurtosis

$$\beta_2 = \mathbf{E}\left(\frac{X - \mathbf{E}X}{\sqrt{var(X)}}\right)^4 - 3$$

(Empirische) Wölbung, Kurtosis

$$\hat{\beta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}}{s}\right)^4 - 3$$

R : `beta2 = mean(((X-mean(X)/sd(X))^4) - 3`

Formmaße (3)

Exzeß

$\beta_2 + 3$ $\quad$ $\hat{\beta}_2 + 3$

$\beta_2 = 0$	bei Normalverteilung
$\beta_2 > 0$	Tails “dicker, länger, stärker” als bei NV
$\beta_2 < 0$	Tails “dünner, kürzer, schwächer” als bei NV

Erinnerung:

$\beta_2 = 0$ heißt nicht notwendig: $F \sim$ Normal.

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

5.1 Box-Plots

Ziel: übersichtliche Darstellung der Daten.

Box-Plots

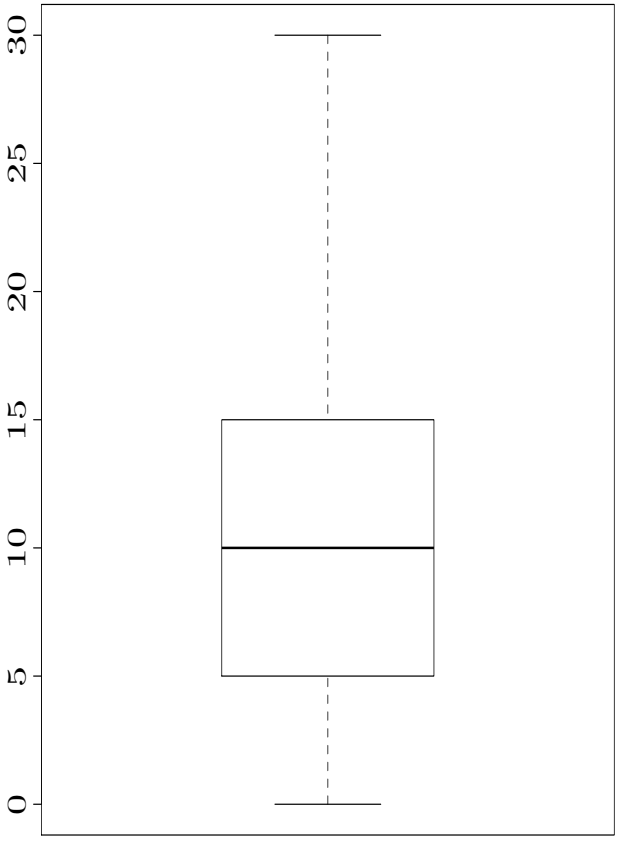
Funktion: `boxplot(x,range,...)`
zeichnet Box mit Linie beim Median und Rahmen bei Quartilen.

Parameter range

bestimmt die Länge der Whiskers (engl.: Schnurrhaare):
Whiskers bis max./min. Wert im Intervall
$[x_{0.25} - range \cdot IR, x_{0.75} + range \cdot IR]$
Falls `range = 0` $\Rightarrow$ Whiskers bis Extremwerte (egal wie groß)
Standard: `range = 1.5`

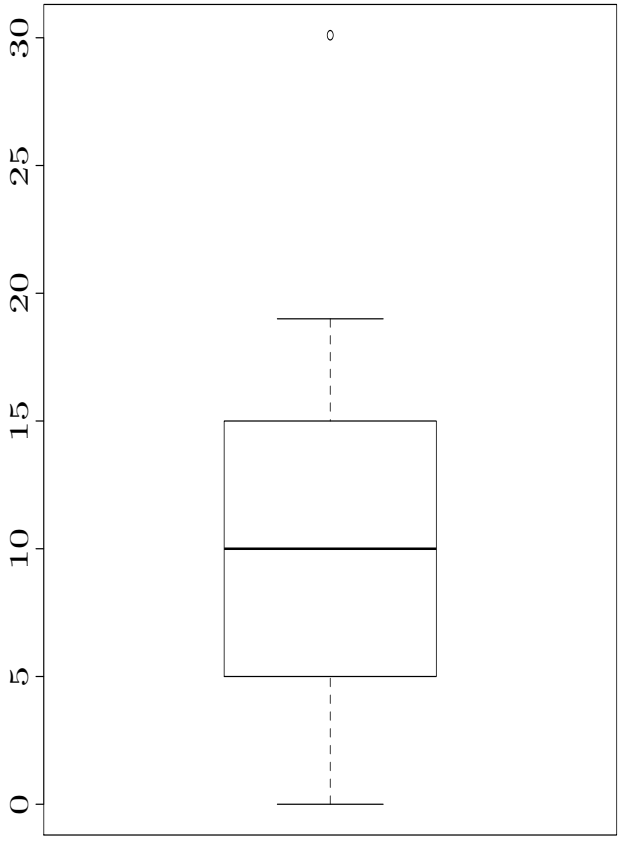
5.1 Box-Plots

```
boxplot(c(0:19,30))
```

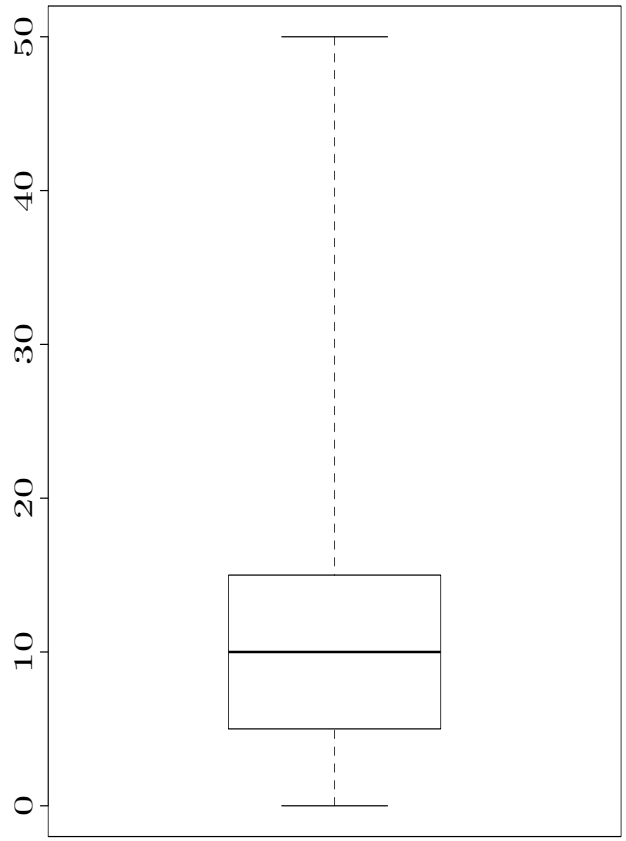


$IR = 10, x_{75} = 15$

```
boxplot(c(0:19,30.1))
```




```
boxplot(c(0:19,50),range=0)
```



Erläuterung zum Wert range=1.5

$X \sim \mathcal{N}(\mu, \sigma^2)$

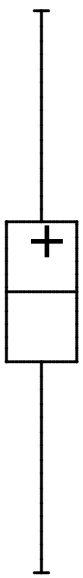
etwa 99% der Daten liegen zwischen den “fences” (den ...).

$$
\begin{aligned}
0.99 &= 0.995 - 0.005 \\
&= \Phi(2.575) - \Phi(-2.575) \\
&= P(\mu - 2.575\sigma < X < \mu + 2.575\sigma) \\
&\approx P(x_{0.5} - 2.575 \cdot \underbrace{0.7434 \cdot IR}_{} < X < \\
&\qquad x_{0.5} + 2.575 \cdot \underbrace{0.7434 \cdot IR}_{}) \\
&= P(x_{0.5} - 1.914 \cdot IR < X < x_{0.5} + 1.914 \cdot IR) \\
&\approx P(x_{0.5} - 2 \cdot IR < X < x_{0.5} + 2 \cdot IR) \\
&= P(x_{0.25} - 1.5 \cdot IR < X < x_{0.75} + 1.5 \cdot IR)
\end{aligned}
$$

5.1 Box-Plots

* Ausreißer ??

$x_{0.75} + 3 \cdot IR$



$x_{0.75} + 1.5 \cdot IR$ oder Max.

$x_{0.75}$ $\overline{X}$

$x_{0.5}$

$x_{0.25}$

$x_{0.25} - 1.5 \cdot IR$ oder Min.

$x_{0.25} - 3 \cdot IR$

komplexere Box-Plots in R

Geg.: data.frame `dfr` mit Merkmalen m1,m2 und Gruppierungsmerkmale gr1,gr2

Ein Merkmal, mehrere Gruppen: Formeln

```
boxplot(m1 ~ gr1,data=dfr)
boxplot(m1 ~ gr1*gr2,data=dfr)
```

`m1 ~ gr` ist eine Formel (lies Merkmal m1 in Abh. von Gruppe(n) aus gr1)

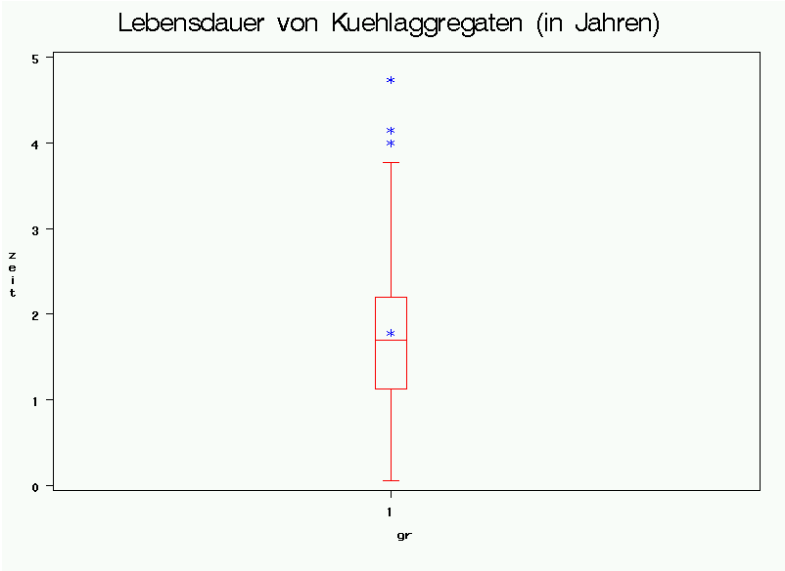
Mehrere Merkmale, eine Gruppe

`boxplot(dfr[1,2])` bzw.

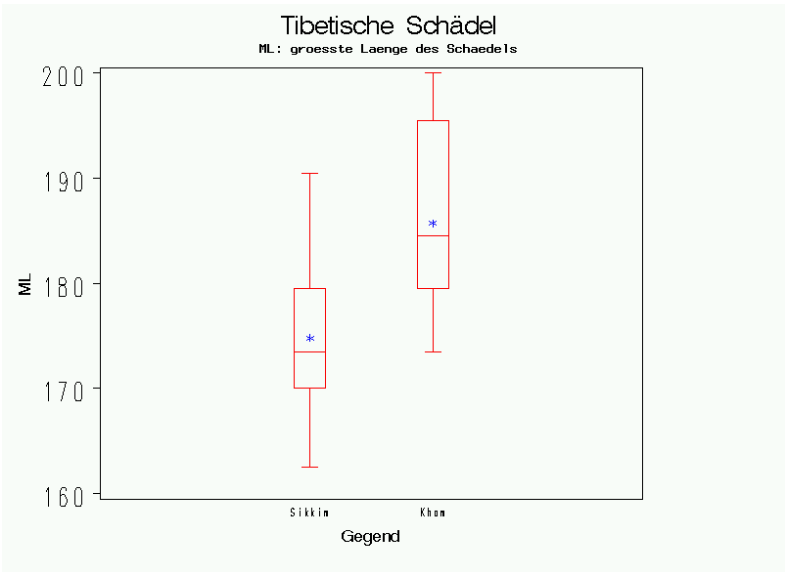
```
boxplot(dfr[c("m1","m2")])
```

Boxplots - Beispiele

Lebensdauern von 100 Kühlaggregaten



Schädelmaße in zwei Regionen Tibets



Inhalt

5.2 Probability Plots

Erinnerung: Normalverteilung

(i) Dichte der Standard-Normalverteilung

$$\phi(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty$$

(ii) Verteilungsfunktion der Standard-Normal

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{t^2}{2}} dt, \qquad -\infty < x < \infty$$

(iii) Dichte der Normalverteilung

$$\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{\sigma^2}},$$

mit Erwartungswert μ und Varianz σ^2.

Probability Plots

Erinnerung: Normalverteilung, Quantile

Der Wert $\Phi^{-1}(u)$ heißt u-Quantil

der Standard-Normalverteilung.

Die Funktion $\Phi^{-1}(u), u \in (0,1)$, heißt Quantilfunktion

der Standard-Normalverteilung.

$\alpha = 0.05$

$$u_\alpha = \Phi^{-1}(1-\alpha) = \Phi^{-1}(0.95) = 1.645$$
$$\Phi^{-1}(1-\frac{\alpha}{2}) = \Phi^{-1}(0.975) = 1.96$$

$\Phi^{-1}(\alpha)$: α-Quantil, theoretisch

$\hat{x}_\alpha = x_{(\lfloor \alpha n \rfloor)}$: α-Quantil, empirisch

Q-Q-Plot

$$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

$$\frac{x_\alpha - \mu}{\sigma} = u_\alpha = \Phi^{-1}(\alpha) \quad \text{gdw.} \quad x_\alpha = \sigma\Phi^{-1}(\alpha) + \mu$$

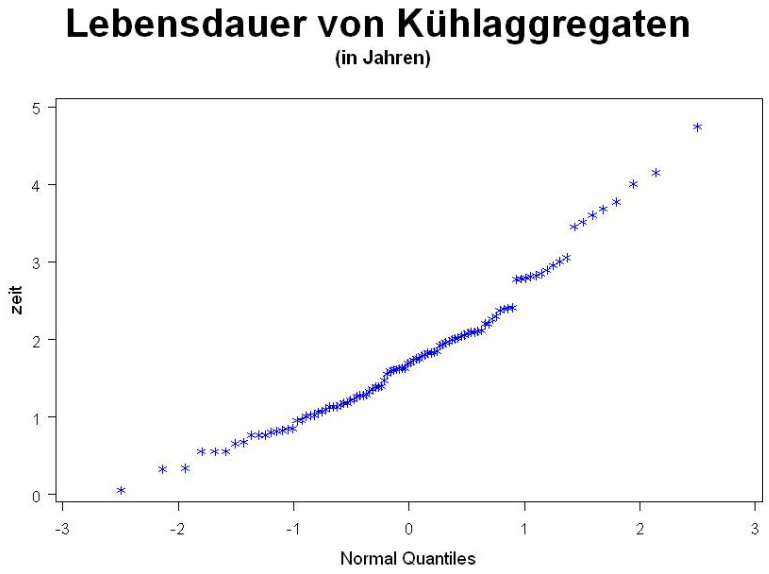
Wenn Normalverteilung zutrifft, so müssen die Punkte $(\Phi^{-1}(\alpha), \hat{x}_\alpha)$ etwa auf einer Geraden liegen,

$$\Phi^{-1}(\alpha) \approx \frac{\hat{x}_\alpha - \mu}{\sigma} = \frac{x_{(\lfloor \alpha n \rfloor)} - \mu}{\sigma}$$

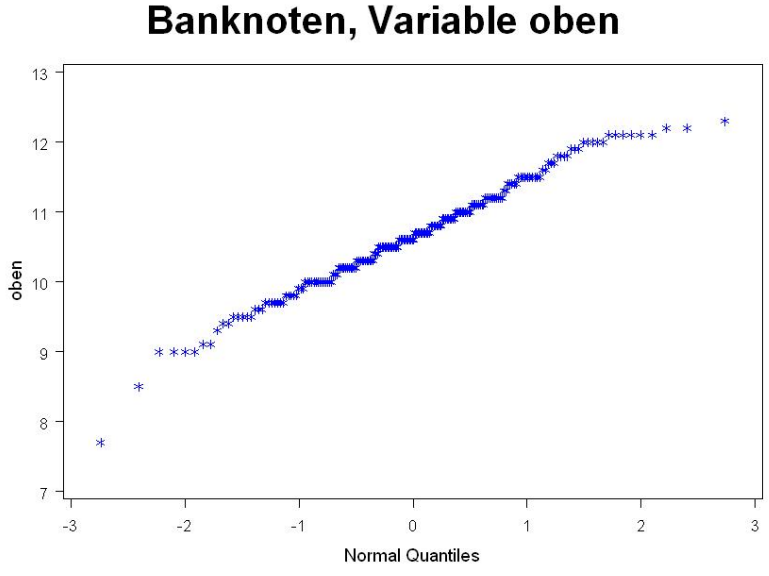
`qqnorm(obj);qqline(obj)`
`qqline` plottet theoretische Werte als Vergleichsgerade
Je näher die Punkte an der Gerade liegen, desto näher sind wir an der NV.

Q-Q Plots - Beispiele (1/2)

Lebensdauern von 100 Kühlaggregaten

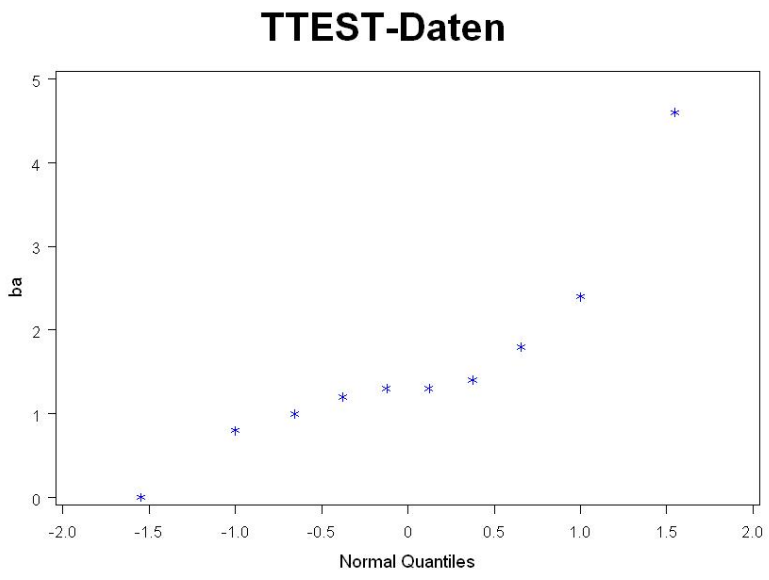


Abmessungen von Banknoten



Q-Q Plots - Beispiele (2/2)

Verlängerung der Schlafdauer



Probability Plot

```
qqnorm(obj,xaxt="n",xlab="Theoretical
   Probabilities")
qqline(obj)
axis(1,at=seq(-3,3,0.5),
   labels=c(round(pnorm(seq(-3,3,0.5)),3)))
```

wie oben, x-Achse hat die selbe Skala, aber eine andere Beschriftung, statt $\Phi^{-1}(u)$ steht u, also werden die Punkte $\left(\alpha, x_{(\lfloor \alpha n \rfloor)}\right)$ geplottet.

Q-Q Plot

Übersicht

Eigenschaften der QQ-Kurve	Interpretation
wenige Punkte weg von der Geraden	Ausreißer
linkes Ende unter der Linie rechtes Ende über der Linie	lange Tails
linkes Ende über der Linie rechtes Ende unter der Linie	kurze Tails
gebogene Kurve, steigender Anstieg	rechtsschief
gebogene Kurve, fallender Anstieg	linksschief
Plateaus und Sprünge	diskrete Daten gerundete Dat.
Gerade $y = x$	empirische $\approx$ theoretische Verteil.
Gerade $y = ax + b$	nur Lage- oder Skalenunterschied

Inhalt

5.3 Häufigkeitsdiagramme: hist & Co

hist

```
hist(obj,breaks,freq,border,col,density,angle)
```

- breaks: Einteilung der Klassen: Names eine Algor. (Standard: `"Sturges"`), Anzahl der Klassen, Vektor mit den Klassengrenzen (breaks) oder Funktion zum Berechnen der Grenzen
- freq: absolute Anzahlen (frequencies)? (sonst Anteile)
- border: Farbe der Rahmen
- col: Farbe der Füllung/Schraffur
- density: Dichte der Schraffur (Standard: voll ausgefüllt)
- angle: Winkel der Schraffur (math. Drehsinn)

5.3 Häufigkeitsdiagramme: hist & Co

(echte) Histogramme

`hist(obj,breaks,freq,border,col,density,angle,...)`

- breaks mit Vektor aus Grenzen (muss Min. und Max. abdecken!) $\Rightarrow$
  Histogramm mit Eigenschaft

$$\sum_{\text{Blöcke } b} \text{Intervallbreite}(b) \cdot \text{Anteil}(b) = 1$$

  wird gezeichnet
- `truehist` aus dem Paket `MASS` erhält diese Eigenschaft immer.

5.3 Häufigkeitsdiagramme: hist & Co

alternative Funktionen

Zunächst mit `hist(...,plot=FALSE)$counts` oder `table` Häufigkeiten ermitteln, dann

- `barplot(...,horiz=TRUE)`: horizontaler Plot
- `plot(...,type="h")`: Nadelplot
- `pie()`: Tortendiagramm

Parametrische Dichteschätzung

Vorgabe: Modell, z.B. Normalverteilung oder Gammaverteilung
Lediglich die Parameter werden geschätzt (hier über Momente):

```
curve(dnorm(x=x,mean(obj),sd(obj)))
curve(dgamma(x=x,shape=(mean(obj)/sd(obj))^2,
rate=mean(obj)/(sd(obj)^2)))
```

Frage: Wie wird geschätzt?

bei Normalverteilung ist das klar: $\overline{X}$ und s^2 sind optimale Schätzungen für μ und σ^2.

Wie findet man (gute) Schätzungen bei anderen Verteilungen?
$\rightarrow$ Abschnitt Schätzmethoden.

Nichtparametrsiche Dichteschätzung

Überlagerung der Daten mit einer (Dichte-) Funktion

$K(t)$ eine Kernfunktion,

$$\int K(t)\,dt = 1, \quad \int tK(t)\,dt = 0,$$
$$\int t^2K(t)\,dt = 1, \quad \int K^2(t)\,dt < \infty$$

Dichteschätzung oder Dichtefunktionsschätzung.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-x_i}{h})$$

$x_1, \ldots, x_n$: die Beobachtungen.
h: ein sogenannter Glättungsparameter.

Nichtparametrsiche Dichteschätzung

Nichtparametrsiche Dichteschätzung in R

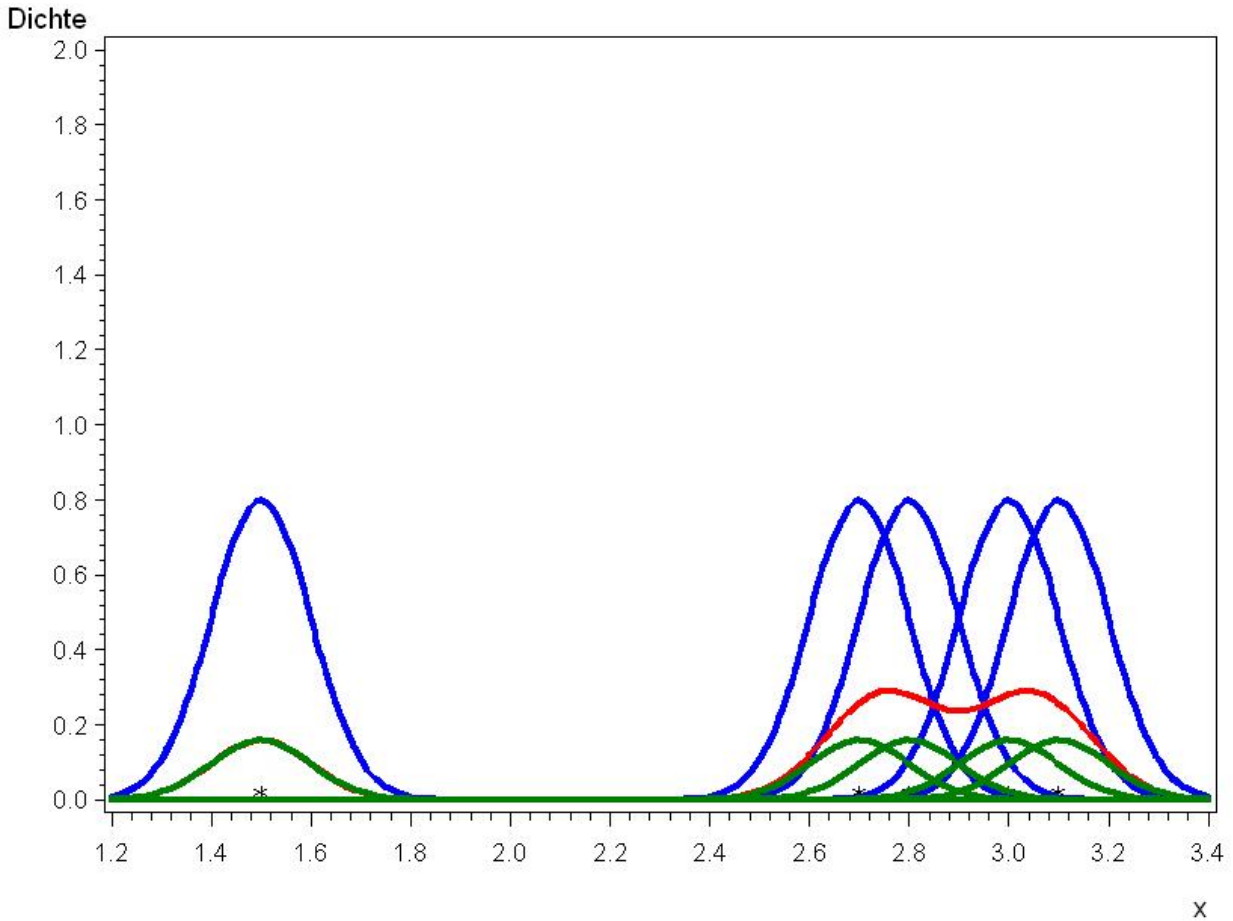
Funktion density

```
density(obj,kernel,from,to,n)
```

- kernel: Name einer Kernfunktion (Standard: `"gaussian"`)
- form,to: Intervall für Schätzung
- Anzahl der Stellen (Standard: 512, Zweierpotenz angebracht)

Dichteschätzung

Motivation Kern-Dichteschätzung



Beispiel

Histogramm und Dichteschätzung in R

Kühlaggregate

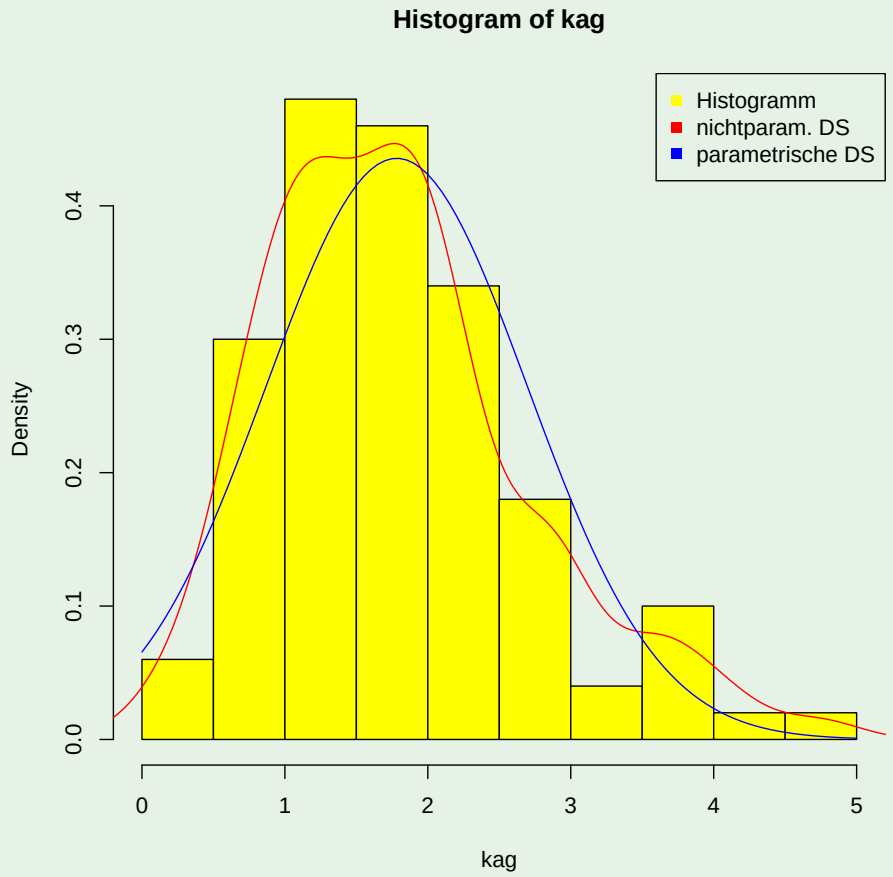
```
kag = scan()
1.29 1.38 2.89 ... 1.55 0.55 1.26 1.18

hist(kag,col="yellow",freq=F)
lines(density(kag),col="red")
curve(dnorm(x,mean=mean(kag),
   sd=sd(kag)),add=T,col="blue")
legend("topright",pch=15,
   col=c("yellow","red","blue"),
   legend=c("Histogramm",
"nichtparam. DS", "parametrische DS"))
```

Beispiel

Histogramm und Dichteschätzung in R

Kühlaggregate



Einschub: Graphikparameter in R

Die Funktion par

Details der Darstellung

`par(adj,bg,cex,din,family,fig,font,lab,lty, mfcol,new,pch,xlog,...)` (viele weitere, s. Hilfe)
Plotfunktionen greifen auf Werte von `par` zurück → vor dem Plotten setzen!
Einige (z.B. col) auch direkt beim Aufrufen der Funktionen setzbar(s. `plot`,`boxplot` etc.)

- adj: Ausrichtung des Texts von 0 linksb. bis 1 rechtsbündig
- bg: Hintergrundfarbe (u.a. als `"Farbe"` oder `"#RRBBGG"`), fg existiert auch
- cex: Vervielfachungsfaktor der Standardschriftgröße

Achtung: bg und cex sind auch direkte Parameter versch. Fkt.

Einschub: Graphikparameter in R

Die Funktion par(2)

Details der Darstellung

`par(adj,bg,cex,din,family,fig,font,lab,lty, mfcol,new,pch,xlog,...)` (viele weitere, s. Hilfe)

- din: `c(Breite,Höhe)` des Plotbereichs (d.h. des Fensters, falls nicht in Datei geplottet wird) in Zoll (inch)
- family: Schriftartenfamilie (z.B. `"serif"`), Optionen variieren je nach OS und Fenster vs. Datei.
- fig: Vektor `(x,y,h,b)` Position und Größe der eigentlichen Figur
- font: 1 Standard, 2 fett, 3 kursiv und 4 fett und kursiv
- lab: `c(x,y,nutzlos)`: x: Anzahl der Striche an der x-Achse

Einschub: Graphikparameter in R

Die Funktion par(3)

Details der Darstellung

`par(adj,bg,cex,din,family,fig,font,lab,lty, mfcol,new,pch,xlog,...)` (viele weitere, s. Hilfe)

- lty: Linientyp: 1 durchg. 2 gestrichelt 3 gepunktet 4 Strichpunktlinie 5 lange Striche 6 Doppelstriche
- **mfcol,mfrow**: `c(Zeilen,Spalten)` mehrere Plots in eine Graphik (Gitter). mfcol zeichnet spaltenweise, mfrow zeilenweise
- new: nächsten Plot **hinzufügen** (!)
  new heißt: Schon für neuen Plot vorbereitet, Löschen des Inhalts nicht nötig
  wird nach jedem Plot auf `FALSE` gesetzt

Einschub: Graphikparameter in R

Die Funktion par(4)

Details der Darstellung

`par(adj,bg,cex,din,family,fig,font,lab,lty, mfcol,new,pch,xlog,...)` (viele weitere, s. Hilfe)

- pch: (plot character): Zeichen für Punkte im Plot, Zahl (Bedeutung unter `?points`,s. `legend` im letzten Beispiel) oder einzelnes Zeichen
- xlog: logarithmische Skale nutzen?

`par` gibt alte Werte zurück $\Rightarrow$ speichern und zurücksezten möglich

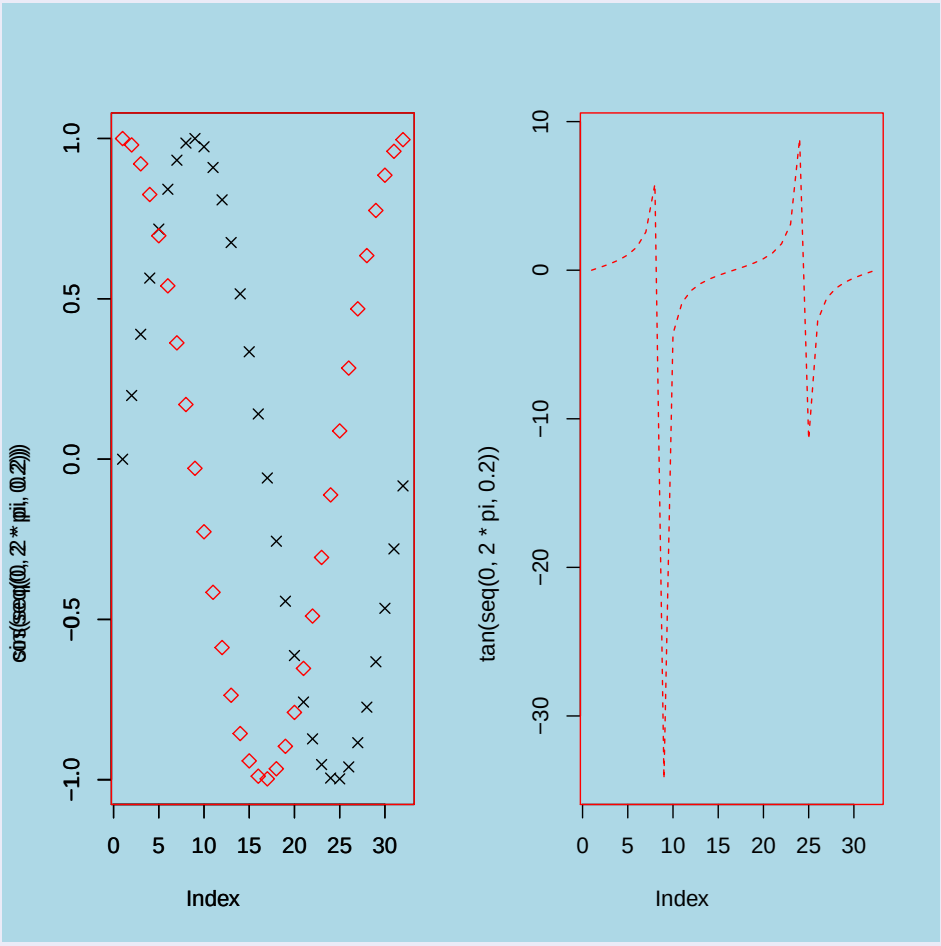
Einschub: Graphikparameter in R

Beispiel

```
oldpar = par(pch=4,
   mfrow=c(1,2),bg="lightblue", adj=0.3)
plot(sin(seq(0,2*pi,0.2)))
par(new=TRUE,pch=5,col="red")
plot(cos(seq(0,2*pi,0.2)))
plot(tan(seq(0,2*pi,0.2)),type="l",lty=2)
par(oldpar)
```

Einschub: Graphikparameter in R

Beispiel



`new=TRUE` sorgt für Doppelplot im linken Feld.

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

6.1 Häufigkeitstabellen

Die Prozedur FREQ

Ein-, zwei- und höherdimensionale Häufigkeiten

Eindimensionale Zufallsvariablen

$$X: \quad \begin{pmatrix} x_0 & x_1 & \cdots & x_n & \cdots \\ p_0 & p_1 & \cdots & p_n & \cdots \end{pmatrix}$$

Die p_i sind zu schätzen:

$$\hat{p}_i = \frac{n_i}{N}$$

N: Stichprobenumfang n_i: relative Häufigkeiten

`prop.table(table(x))`

`table`: absol. Tabelle, `prop.table`: abs. $\rightarrow$ relativ

DescrFreqBanknote.R DescrFreq.R

Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel

3maliges Werfen einer Münze

X: Anzahl von Blatt nach 3 Würfen
Y: Anzahl von Blatt nach 2 Würfen

Element von Ω	X	Y
BBB	3	2
BBZ	2	2
BZB	2	1
BZZ	1	1
ZBB	2	1
ZBZ	1	1
ZZB	1	0
ZZZ	0	0

Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel (Fortsetzung)

Besetzungswahrscheinlichkeiten

$X \mid Y$	0	1	2	
0	$\frac{1}{8}$	0	0	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{4}$	0	$\frac{3}{8}$
2	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$
3	0	0	$\frac{1}{8}$	$\frac{1}{8}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix} \qquad Y : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Tabelle der zweidimensionalen Wahrscheinlichkeiten

$X \mid Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_N	
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1N}	$p_{1.}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2N}	$p_{2.}$
$\cdots$							
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iN}	$p_{i.}$
$\cdots$							
x_M	p_{M1}	p_{M2}	$\cdots$	p_{Mj}	$\cdots$	p_{MN}	$p_{M.}$
	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.N}$	1

Zweidimensionale diskrete Zufallsgrößen

Zweidimensionale Zufallsvariable

Seien X, Y Zufallsgrößen. Das Paar (X, Y) heißt zweidimensionale Zufallsvariable.

Seien X und Y diskret und (x_i, y_j) die möglichen Ergebnisse von (X, Y), $\quad i = 1, \ldots, M, j = 1, \ldots, N$.

gemeinsame Wahrscheinlichkeitsfunktion von (X, Y)

$$p_{ij} = P(X = x_i, Y = y_j),$$

$$\begin{aligned} p_{ij} &\geq 0 \\ \sum_{i,j} p_{ij} &= 1 \end{aligned}$$

$$p_{i.} := \sum_{j=1}^{N} p_{ij}$$

$$p_{.j} := \sum_{i=1}^{M} p_{ij}$$

Zweidimensionale diskrete Zufallsgrößen

X und Y heißen unabhängig, wenn

$$p_{ij} = P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j) = p_{i.} \cdot p_{.j}$$

$p_{i.}$ und $p_{.j}$ heißen Randwahrscheinlichkeiten.

Zweidimensionale diskrete Zufallsgrößen

Beispiel

Treiben Sie Sport?

X: 0 - nein 1 - ja

Y: 0 - weiblich 1 - männlich

X | Y	0	1	
0	p_{00}	p_{01}	$p_{0.}$
1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	

p_{ij}: unbekannt!

Frage: Ist das Sportverhalten von Männern und Frauen unterschiedlich? Hängt das Sportverhalten vom Geschlecht ab?

Zweidimensionale diskrete Zufallsgrößen

Kontingenztafel

Befragung liefert Häufigkeiten für die einzelnen Felder. Anhand dieser Häufigkeiten werden die Wahrscheinlichkeiten geschätzt!

Die Tabelle der Häufigkeiten heißt Kontingenztafel

X | Y	0	1	# der beobachteten	
0	n_{00}	n_{01}	$n_{0.}$	Nichtsportler
1	n_{10}	n_{11}	$n_{1.}$	Sportler
	$n_{.0}$	$n_{.1}$		
	# der befragten Frauen	# der befragten Männer		

Mehrdimensionale diskrete Zufallsgrößen

Häufigkeitstabellen in R

Geg.: data.frame `dfr` mit Spalten `X`,`Y` und `Z`

2 bzw. 3 Dimensionen

`table(dfr$X,dfr$Y)` bzw. `table(dfr$X,dfr$Y,dfr$Z)`
oder
`ftable(X ~ Y,data=dfr)` bzw.
`ftable(X ~ Y+Z,data=dfr)`

alle Dimensionen

`table(dfr)` oder `ftable(dfr)`

Parameter exclude

Werte ausschließen: z.B. `NA`, `NaN` oder `Inf`

Mehrdimensionale diskrete Zufallsgrößen

Häufigkeitstabellen in R (2)

Geg.: data.frame `dfr` mit Spalten `X`,`Y` und `Z`

Funktion margin.table

`margin.table(table(dfr),dim)` gibt Randtabelle für Dimensionen dim zurück.,d.h. `dim =c (2,3)` für `Y`×`Z`

as.data.frame und xtabs

`as.data.frame(table(...))` macht aus Zeilen-/Spaltennamen Variablen, `xtabs()` umgekehrt

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \underset{\leftarrow}{\overset{\texttt{as.data.frame}}{\rightarrow}} \underset{\texttt{xtabs}}{} \begin{pmatrix} 0 & 0 & a \\ 0 & 1 & b \\ 1 & 0 & c \\ 1 & 1 & d \end{pmatrix}$$

Assoziationsmaße

nur für mehrdim. Tabellen

χ^2-Statistik

$$\sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

Φ-Koeffizient für 2x2 Tafeln

$$\Phi^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1.}p_{2.}p_{.1}p_{.2}}$$

Odds Ratio für 2x2 Tafeln

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Schätzung: Ersetzen der Wahrscheinlichkeiten durch die jeweiligen relativen Häufigkeiten.

Assoziationsmaße, Beispiel

Mendelsche Kreuzungsversuche

```
erbsen=read.table(stdin(),
col.names = c(
"rund","gruen","Anzahl"))
0 0 101
0 1  32
1 0 315
1 1 108

erbstab = xtabs(
Anzahl ~ rund+gruen,
data=erbsen)
```

```
chisq.test(erbstab,
   correct=F)
phi(erbstab,
   digits=4)
#phi aus Paket
   "psych"
OR =
   (erbstab[1,1]*
   erbstab[2,2])/
 (erbstab[1,2]*
   erbstab[2,1])
```

$\chi^2 = 0.1163$ Φ-Koeffizient$= 0.0145$ $OR = 1.0821$

Inhalt

6.2 Zusammenhangsmaße

zwischen Zufallsvariablen X, Y

Erinnerung: Varianz der Zufallsvariablen X

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 \\ &= \mathbf{E}[(X - \mathbf{E}X)(X - \mathbf{E}X)] \end{aligned}$$

Kovarianz der Zufallsvariablen X und Y

$$\begin{aligned} Cov(X, Y) &= \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y) \\ &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \end{aligned}$$

Korrelation der Zufallsvariablen X und Y

$$Corr(X, Y) = \frac{\mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]}{\sqrt{var(X) \cdot var(Y)}}$$

Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})(x_i - \overline{X})$$

empirische Kovarianz

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})$$

empirische Korrelation, Pearson-Korrelationskoeffizient

$$r_{XY} := \frac{s_{XY}}{s_X s_Y}$$

Pearson-Korrelationskoeffizient

Eigenschaften

- Es gilt stets:
$$-1 \leq r_{XY} \leq 1.$$

- Der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen
$$x \longrightarrow a + bx$$

- $|r_{XY}| = 1$ gdw. alle Punkte auf einer Geraden liegen,
$y = mx + b, m \neq 0$
$r_{XY} = \quad 1 \rightarrow$ Anstieg > 0
$r_{XY} = -1 \rightarrow$ Anstieg < 0

Pearson-Korrelationskoeffizient

- Der Pearson-Korrelationskoeffizient ist also ein Maß für die lineare Abhängigkeit von X und Y.
- $r_{XY} \approx 0 \longrightarrow$ keine lineare Beziehung zwischen X und Y erkennbar, aber es sind durchaus andere Abhängigkeiten möglich!
- Der Pearson-Korrelationskoeffizient ist nicht robust gegen Ausreißer (siehe Übung)

Realisierung in R:
`cor(x,y,method="pearson")` berechnet Koeffizient
`cor.test(x,y,method="pearson")` berechnet Koeffizient + Signifikanztest (später)
`method="pearson"` ist Standard und kann entfallen.

Spearman-Korrelationskoeffizient

Spearman-Rangkorrelationskoeffizient

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}}$$

R_i: Rang von X_i in der geordneten Stichprobe $X_{(1)} \leq \ldots \leq X_{(n)}$
S_i: Rang von Y_i in der geordneten Stichprobe $Y_{(1)} \leq \ldots \leq Y_{(n)}$

`cor(x,y,method="spearman")` bzw.
`cor.test(x,y,method="spearman")`

Spearman-Korrelationskoeffizient

$$
\begin{aligned}
r_S &= \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \\
&= \frac{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \\
&= 1 - \frac{6 \cdot \sum_{i=1}^{n}(R_i - S_i)^2}{n \cdot (n^2 - 1)}
\end{aligned}
$$

$$
-1 \leq r_S \leq +1
$$

$|r_S| = 1$ gdw. X_i, Y_i in gleicher oder entgegengesetzter Weise geordnet sind!

Spearman-Korrelationskoeffizient

Beweis der letzten Formel (1)

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})^2}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})}}$$

Nenner:

$$\sum_{i=1}^{n}(R_i - \overline{R})^2 = \sum_{i=1}^{n}(S_i - \overline{S})^2 = \sum_{i=1}^{n}(i - \frac{n+1}{2})^2$$

$$\begin{aligned}
&= \sum i^2 - 2 \cdot \frac{n+1}{2} \sum i + n \cdot (\frac{n+1}{2})^2 \\
&= \frac{n \cdot (n+1) \cdot (2n+1)}{6} - \frac{n \cdot (n+1)^2}{2} + \frac{n \cdot (n+1)^2}{4} \\
&= \frac{n \cdot (n+1)}{12} \cdot [2 \cdot (2n+1) - 3 \cdot (n+1)] \\
&= \frac{(n-1) \cdot n \cdot (n+1)}{12} = \frac{n \cdot (n^2-1)}{12}
\end{aligned}$$

Spearman-Korrelationskoeffizient

Beweis der letzten Formel (2)

Zähler:

$$\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S}) = \sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})$$

$$= \sum_{i=1}^{n} R_i S_i - 2 \cdot \frac{n+1}{2} \sum_{i=1}^{n} R_i + n \cdot (\frac{n+1}{2})^2$$

$$= \sum_{i=1}^{n} R_i S_i - \frac{n \cdot (n+1)^2}{4}$$

Damit erhalten wir eine weitere Darstellung für r_S:

$$r_S = 12 \cdot \frac{\sum_{i=1}^{n} R_i S_i - \frac{n \cdot (n+1)^2}{4}}{(n-1) \cdot n \cdot (n+1)}$$

Spearman-Korrelationskoeffizient

Andere Darstellung für den Zähler

Setzen: $d_i := R_i - S_i = (R_i - \frac{n+1}{2}) + (\frac{n+1}{2} - S_i)$

$$\begin{aligned}
\sum d_i^2 &= \sum (R_i - \frac{n+1}{2})^2 + \sum (S_i - \frac{n+1}{2})^2 \\
&\quad -2 \sum (R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2}) \\
&= \frac{(n-1)n(n+1)}{12} + \frac{(n-1)n(n+1)}{12} \\
&\quad -2 \cdot r_S \cdot \frac{(n-1)n(n+1)}{12} \\
&= \frac{(n-1)n(n+1)}{6}(1 - r_S) \\
r_S &= 1 - \frac{6 \sum d_i^2}{(n-1)n(n+1)}
\end{aligned}$$

Spearman-Korrelationskoeffizient

Drei Darstellungen

$$
\begin{aligned}
r_S &= \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}} \\
&= 12 \cdot \frac{\sum_{i=1}^{n} R_i S_i - \frac{n\cdot(n+1)^2}{4}}{(n-1)n(n+1)} \\
&= 1 - \frac{6\sum (R_i - S_i)^2}{(n-1)n(n+1)}
\end{aligned}
$$

Bem.: Es gilt:

a) $-1 \leq r_S \leq 1$

b) $r_S = 1 \Leftrightarrow R_i = S_i \quad \forall i = 1, \ldots, n$

c) $r_S = -1 \Leftrightarrow R_i = n + 1 - S_i \quad \forall i = 1, \ldots, n$

Vergleich der Korrelationskoeffizienten

Pearson - Spearman

Vorteile Spearman

- es genügt ordinales Meßniveau
- leicht zu berechnen
- r_S ist invariant gegenüber monotonen Transformationen
- gute Interpretation, wenn $r_S \approx -1, 0, 1$ (wie bei Pearson)
- eignet sich als Teststatistik für einen Test auf Unabhängigkeit
- ist robust gegen Abweichungen von der NV.

Vergleich der Korrelationskoeffizienten

Pearson - Spearman

Nachteile Spearman

- wenn kardinales (stetiges) Meßniveau $\longrightarrow$ evtl. Informationsverlust
- schwierige Interpretation, wenn r_S nicht nahe 0, 1, oder -1

(gilt eingeschränkt auch für Pearson)

Kendalls τ (Konkordanzkoeffizient)

$(X_i, Y_i), i = 1, ..., n$

$$a_{ij} = \begin{cases} 1, \text{ falls } x_i < x_j \wedge y_i < y_j \text{ oder} \\ \qquad\qquad x_i > x_j \wedge y_i > y_j \\ -1, \text{ falls } x_i < x_j \wedge y_i > y_j \text{ oder} \\ \qquad\qquad x_i > x_j \wedge y_i < y_j \\ 0, \text{ sonst} \end{cases}$$

$$= sgn[(x_i - x_j)(y_i - y_j)]$$

Falls $a_{ij} = 1$ so heißen die Paare konkordant
Falls $a_{ij} = -1$ " diskordant
Falls $a_{ij} = 0$ " gebunden

Kendalls τ (Konkordanzkoeffizient)

$$\begin{aligned}\tau &= \frac{2 \cdot \sum_{i<j} a_{ij}}{N \cdot (N-1)} = \frac{1}{\binom{N}{2}} \cdot \sum_{i<j} a_{ij} \\ &= \frac{\sharp \text{ konkordanter Paare - } \sharp \text{ diskordanter Paare}}{\binom{N}{2}}\end{aligned}$$

Bem.: einfache Berechnung, wenn neue Paare hinzukommen

Bem.: meist gilt: $|\tau| < |r_S|$. Approximation von τ:

$$\tau_{appr.} = \frac{2}{3}\frac{N+1}{N} r_S$$

```
cor(x,y,method="kendall")
```

Inhalt

6.3 Das Regressionsproblem

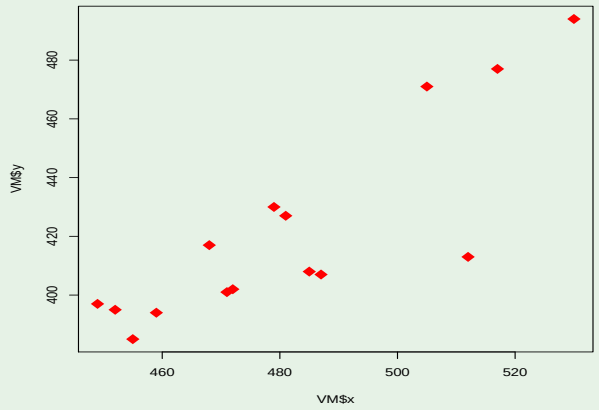
Scatterplots

Scatterplot

Zweidimensionale Stichproben können als Punkte in der Ebene dargestellt werden

Länge und Breite von Venusmuscheln

```
plot(VM$x,VM$y,
col="red",pch=18,
cex=2)
```



Descr_Scatter.R Descr_Scatter1.R

Das Regressionsproblem

X, Y: Zufallsvariablen (auch mehrdimensional)

Modell:

$$Y = f(X, \underbrace{\theta_1, \ldots, \theta_p}_{Parameter}) + \underbrace{\epsilon}_{zuf.Fehler}, \qquad \epsilon \sim (0, \sigma^2).$$

f linear, bekannt bis auf Parameter:
lineare Regression

f nichtlinear, bekannt bis auf Parameter:
nichtlineare Regression

f unbekannt: nichtparametrische Regression

Regression

f bekannt (bis auf Parameter)

Aufgabe:

$$min_{\theta_1,\ldots,\theta_p}\mathbf{E}(Y - f(\mathbf{X}, \theta_1, \ldots, \theta_p))^2$$

$\theta_1, \ldots, \theta_p$ unbekannt.

Beobachtungen: $(Y_i, \mathbf{X}_i)$.

Erwartungswert durch arithmetisches Mittel ersetzen

$$min_{\theta_1,\ldots,\theta_p}\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i, \theta_1, \ldots, \theta_p))^2$$

Kleinste Quadrat-Schätzung für $\theta_1, \ldots, \theta_p$ (KQS)
Least-Squares-Estimation (LSE)

Regression

f bekannt (bis auf Parameter)

Lösung des Minimum-Problems

$$min_{\theta_1,\ldots,\theta_p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i, \theta_1, \ldots, \theta_p))^2$$

zu minimierende Funktion nach den Parametern differenzieren und Null setzen:

$$\frac{2}{n} \cdot \sum_{i=1}^{n} (Y_i - f(\mathbf{X}_i, \theta_1, \ldots, \theta_p)) \cdot \frac{\partial f(\mathbf{X}_i, \theta_1, \ldots, \theta_p)}{\partial \theta_j} = 0$$

$j = 1, \ldots, p, \Rightarrow$ Gleichungssystem mit p Gleichungen.

Regression

f linear: lineares Gleichungssystem (1)

$$f(X, \theta_1, \theta_2) = \theta_1 X + \theta_2$$

$$\frac{\partial f}{\partial \theta_1} = X \qquad \frac{\partial f}{\partial \theta_2} = 1$$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(Y_i - (\theta_1 X_i + \theta_2)) \cdot X_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n}(Y_i - (\theta_1 X_i + \theta_2)) \cdot 1 &= 0 \\
\hline
\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \theta_2 \sum_i X_i &= 0 \\
\sum_i Y_i - \theta_1 \sum_i X_i - \theta_2 \cdot n &= 0
\end{aligned}$$

Regression

f linear: lineares Gleichungssystem (2)

Die zweite Gleichung nach θ_2 auflösen:

$$\theta_2 = \frac{1}{n}\sum_i Y_i - \theta_1 \frac{1}{n}\sum_i X_i$$

und in die erste einsetzen:

$$\sum_i X_i Y_i - \theta_1 \sum_i X_i^2 - \frac{1}{n}\sum_i Y_i \sum_i X_i + \theta_1 \frac{1}{n}\sum_i X_i \sum_i X_i = 0$$

$$\sum_i X_i Y_i - \frac{1}{n}\sum_i Y_i \sum_i X_i - \theta_1 \Big(\big(\sum_i X_i^2 - \frac{1}{n}\sum_i X_i \sum_i X_i\big) = 0$$

$\Rightarrow$

$$\hat{\theta}_1 = \frac{\sum_i X_i Y_i - \frac{1}{n}\sum_i X_i \sum_i Y_i}{\sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2} = \frac{S_{XY}}{S_X^2}, \hat{\theta}_2 = \frac{1}{n}\big(\sum_i Y_i - \hat{\theta}_1 \sum_i X_i\big)$$

Regression

Zähler und Nenner in $\hat{\theta}_1$

$$
\begin{aligned}
S_{XY} &= \frac{1}{n-1}\sum_i (X_i - \overline{X})(Y_i - \overline{Y}) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - \overline{X}\sum_i Y_i - \overline{Y}\sum_i X_i + n\overline{XY}\Big) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - n\overline{XY} - n\overline{XY} + n\overline{XY}\Big) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - n\overline{XY}\Big) \\
&= \frac{1}{n-1}\Big(\sum_i X_i Y_i - \frac{1}{n}\sum_i X_i \sum_i Y_i\Big) \\
S_{X^2} &= \frac{1}{n-1}\Big(\sum_i X_i X_i - \frac{1}{n}\sum_i X_i \sum_i X_i\Big)
\end{aligned}
$$

Spezialfall $f(X, \theta) = \theta$ (konstant)

$$Y_i = \theta + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2)$$

Minimierungsaufgabe:

$$\min_{\theta} \left(\sum_{i=1}^{n} (Y_i - \theta)^2 \right)$$

Lösung:

$$2 \sum_{i=1}^{n} (Y_i - \theta) = 0 \qquad \sum_{i=1}^{n} Y_i - n\theta = 0$$

$$\hat{\theta} = \frac{1}{n} \sum Y_i = \overline{Y}$$

D.h. $\overline{Y}$ ist auch KQS.

Spezialfall $f(X, \theta) = \theta$

Schätzung des Schätzfehlers

$$\sigma^2_{Y_i} = \sigma^2_{\theta+\epsilon_i} = \sigma^2_{\epsilon_i} = \sigma^2.$$

Schätzfehler:

$$\begin{aligned}
\sigma^2_{\hat{\theta}} &= var(\hat{\theta}) = var(\frac{1}{n} \cdot \sum Y_i) = \frac{1}{n^2} \cdot n \cdot varY_i \\
&= \frac{1}{n} \cdot \sigma^2 \qquad \rightarrow_{n\rightarrow\infty} \quad 0 \\
\hat{\sigma}^2_{\hat{\theta}} &= \frac{\hat{\sigma}^2}{n}
\end{aligned}$$

Lineare und Nichtlineare Regression

f: linear, $f(X, \theta_1, \theta_2) = \theta_1 X + \theta_2$

θ_1 und θ_2 werden geschätzt.

Descr_Scatter_1.R Descr_Scatter_Heroin.R

f: nichtlinear, z.B. $f(X, \theta_1, \theta_2) = ln(\theta_1 X + \theta_2)$

a) Lösung des nichtlinearen Gleichungssystems

b) wird auf den linearen Fall zurückgeführt

$$\begin{aligned} Y &= \ln(\theta_1 X + \theta_2) + \epsilon \\ e^Y &= \theta_1 X + \theta_2 + \tilde{\epsilon} \end{aligned}$$

Modelle sind aber i.A. nicht äquivalent!

Weitere nichtlineare Regressionsfunktionen

$$
\begin{aligned}
f(t) &= a + bt + ct^2 && \text{Parabel} \\
f(t) &= at^b && \text{Potenzfunktion} \\
f(t) &= ae^t && \text{Exponentialfunktion} \\
f(t) &= k - ae^{-t} && \\
f(t) &= \frac{k}{1 + be^{-ct}} && \text{logistische Funktion} \\
\ln f(t) &= k - \frac{a}{b + t} && \text{Johnson-Funktion} \\
\ln f(t) &= k - \lambda e^{-t} && \text{Gompertz-Funktion}
\end{aligned}
$$

Parametrische Regression in R

`lm` (lineare Modelle)

`lm(formula,data)`, Formeln haben die Form `Y ~ f`, wobei `f` Variablen und ihre Beziehungen enthält:

- Summe: `A + B`
- Interaktion: `A:B` (s. Varianzanalyse)
- Abkürzungen: `A * B` = `A + B + A:B` und $\hat{Ak}$ =`A * ...* A`
- Funktionen: `log(A)`
- Arithmetische Operationen: `I(A*A)`

Die zu schätzenden Paramter werden nicht aufgeführt:
`Y ~ A + B` bedeutet:
Modell ist $Y = c_A A + c_B B + c$ und c_a,c_b und c (Intercept) sind zu schätzen.

Parametrische Regression in R

Beispiel Venusmuscheln

```
venusm = scan(what=list(integer(),integer()))
530 494 517 477 505 471 512 413 487 407
481 427 485 408 479 430 452 395 468 417
459 394 449 397 472 402 471 401 455 385
names(venusm) = c("x","y")
lm(y~x,venusm); plot(lm(y~x,venusm))
lm(y~I(x^2)+x,venusm); plot(lm(y~x,venusm))
```

Die Breite y (in mm) von Venusmuscheln wird in Abh. von Ihrer Länge x betrachtet. Das erste Modell ist linear, das zweite quadratisch. `plot` gibt mehrere Plots aus.

Nichtparametrische Regression

f unbekannt, aber "glatt"

Sei f 2x stetig differenzierbar, $f \in C_2$, $\lambda \geq 0$

Ziel: $$\min_{f \in C_2} \left(\sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \cdot \int \left(f''(x)\right)^2 dx \right)$$

Lösung: Glättender Kubischer Spline.

Descr_Scatter.R

`smooth.spline(x,y,spar,all.knots)`

spar:	Glättungsparameter (meist aus $(0, 1]$)
	spar=0+ε: Interpolierender Spline (0 Orig.)
	spar=1: Gerade
all.knots:	alle Punkte mit paarw. versch. Abszissenwerten verwenden?

Nichtparametrische Regression

Kernschätzung, Motivation

geg.: Kernfunktion K, standardisierte Dichte, z.B. Normaldichte, Epanechnikov-Kern.

Regressionsmodell:

$$
\begin{aligned}
Y &= f(X) + \epsilon, \quad \epsilon \sim (0, \sigma^2) \quad \text{also} \\
\mathbf{E}(Y|X = x) &= f(x) \\
f(x) &= \mathbf{E}(Y|X = x) = \int y f_{Y|X}(y|x)\, dy \\
&= \int y \frac{g(x,y)}{f_0(x)}\, dy = \frac{\int y g(x,y) dy}{f_0(x)}
\end{aligned}
$$

Regression

Kernschätzung

$$f(x) = \frac{\int y g(x,y) dy}{f_0(x)}$$

$g(x,y)$: gemeinsame Dichte von (X,Y)
$f_0(x)$: Randdichte von X
$f_{Y|X}$: bedingte Dichte von Y

Der Nenner wird geschätzt durch

$$\hat{f}_0(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \cdot K(\frac{x - x_i}{h})$$

und der Zähler durch

$$\frac{1}{n} \sum_{i=1}^{n} y_i \hat{g}(x_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} y_i \cdot \frac{1}{h} \cdot K(\frac{x - x_i}{h})$$

Regression

Kernschätzung

Beide zusammen ergeben die

Kernschätzung

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i \cdot \frac{1}{h} \cdot K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} \frac{1}{h} \cdot K\left(\frac{x-x_i}{h}\right)}$$

K: Kernfunktion
h: Glättungsparameter

Beschreibende Statistik

Zusammenfassung (1)

Verteilungsfunktion

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

diskrete Verteilung

$$F(x) = \sum_{i:i \leq x} p_i \qquad p_i = P(X = x_i)$$

stetige Verteilung

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad f(t): \text{ Dichte.}$$

Bsp: diskrete Verteilung: Binomial, Poisson
stetige Verteilung: Normal,Gleich,Exp

Beschreibende Statistik

Zusammenfassung (2)

Erwartungswert

$$\mathbf{E}(X) = \begin{cases} \sum x_i p_i & X \text{ diskret} \\ \int x f(x) dx & X \text{ stetig} \end{cases}$$

Varianz

$$var(X) = \mathbf{E}(X - \mathbf{E}X)^2$$

Normalverteilung, Dichte

$$f(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{x^2}{2}} \quad \text{Standard}$$

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Beschreibende Statistik

Zusammenfassung (3)

Gesetz der Großen Zahlen ($\mathbf{E}(X) < \infty$)

$$\overline{X} \longrightarrow \mathbf{E}X, \qquad \overline{X} = \frac{1}{n}\sum X_i$$

Zentraler Grenzwertsatz (X_i iid)

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \longrightarrow Z \sim \mathcal{N}(0,1)$$

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{s} \longrightarrow Z \sim \mathcal{N}(0,1)$$

$$s^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2 \to \sigma^2$$

Beschreibende Statistik

Zusammenfassung (4)

Statistische Maßzahlen

Lagemaße: $\overline{X}, x_{0.5}, x_\alpha, x_{0.25}, x_{0.75}, \overline{x}_\alpha, \overline{x}_{\alpha,w}$

Skalenmaße: s^2, s, R, IR, MAD, Gini, S_n, Q_n

Formmaße: β_1, β_2

`mean`, `median`, `quantile`, `winsor.mean`, `summary`
`sd`, `diff(range(obj))`, `mad`, `Sn`, `Qn` (Pkt.: robustbase)

Beschreibende Statistik

Zusammenfassung (5)

Boxplots	`boxplot`
Häufigkeitsdiagramme	`hist`
	`plot(table(...))`
	`barplot(table(...))`
Häufigkeitstabellen:	`table` (abs.)
	`prop.table(table(...))` (rel.)
Zusammenhangsmaße:	`cor`, `cor.test`
Pearson, Spearman,	
Kendall-Korrelationskoeff.	
Scatterplots	`plot`
Regression:	`lm`, `plot(lm(...))`

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

7. Statistische Tests

7.1 Einführung und Übersicht

Sei X ein Merkmal (eine Zufallsvariable),
$F_X(x) = P(X \leq x) = P_\theta(X \leq x) = F_{X,\theta}(x)$ θ: Parametervektor

Beispiel: $\theta = (\mu, \sigma^2)$

μ: Erwartungswert von X
σ^2: Varianz von X

$X_1, X_2, \ldots, X_n$ Beobachtungen von X

$\mu \approx \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$
$\sigma^2 \approx \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = s^2$

D.h. die unbekannten Parameter werden geschätzt.

Statistische Tests: Einführung

Problem

Schätzungen können sehr schlecht ausfallen!

I.a. vertritt der Fachexperte gewisse Hypothesen bzgl. der (unbekannten) Parameterwerte!

Diese Hypothesen werden verworfen, wenn die erhaltenen Schätzwerte (z.B. $\overline{X}, s^2$) mit ihnen nicht in Einklang stehen.

Statistische Tests: Einführung

Eine verwandte Problemstellung

Elektronischer Großhandel: TV-Geräte

Händler sagt: Ausschußquote $p \leq 1\%$ $(p = 0.01)$
Käufer wäre einverstanden, prüft aber N Geräte!
Davon: N_f fehlerhaft, N_f - Teststatistik

$$\frac{N_f}{N} \cdot 100\% \gg 1\% \Rightarrow \text{ Ablehnung}$$

Zwei Fehler möglich

a) Zufällig N_f zu groß! $\quad p < 0.01$
$\Rightarrow$ Käufer lehnt ab

b) Zufällig N_f zu klein! $\quad p$ groß, $p \gg 0.01$
$\Rightarrow$ Käufer kauft

Statistische Tests: Einführung

Risiken - Fehler

Risiko des Händlers

Käufer lehnt gute Ware ab (weil N_f zufällig zu groß)

Risiko des Käufers

Käufer kauft schlechte Ware (weil N_f zufällig zu klein)

Risiken sollen quantifiziert werden:

a) $P(\text{ Nicht kaufen } | \, p \leq 1\%)$
b) $P(\text{ Kaufen } | \, p > 1\%)$

Beide Risiken nicht gleichzeitig zu minimieren.

Lösung:

$P(\text{ Nicht kaufen } | \, p \leq 1\%) = \alpha$ vorgeben
$P(\text{ Kaufen } | \, p > 1\%)$ minimieren (oder es versuchen)

Hypothesentest

Beispiel: Einstichproben-Lagetest

Sei μ ein Lageparameter, z.B. der Erwartungswert.
Sei μ_0 ein vorgegebener Wert.

Nullhypothese und Alternativhypothese

a) $H_0: \quad \mu \leq \mu_0 \qquad H_A: \mu > \mu_0$
b) $H_0: \quad \mu \geq \mu_0 \qquad H_A: \mu < \mu_0$
c) $H_0: \quad \mu = \mu_0 \qquad H_A: \mu \neq \mu_0$

Teststatistik

$$T(X_1, ..., X_n) = \frac{\overline{X} - \mu_0}{s} \cdot \sqrt{n}$$

T heißt auch Testgröße, Prüfgröße, Stichprobenfunktion.

Hypothesentest

Allgemein

Die Entscheidung für H_A oder für H_0 wird anhand einer Teststatistik

$$T = T(x_1, \ldots, x_n)$$

gefällt.
Liegt der Wert von T in einem vorher bestimmten Bereich K, dem sogen. Ablehnungsbereich oder kritischen Bereich, dann wird H_0 abgelehnt, anderenfalls wird H_0 nicht abgelehnt.

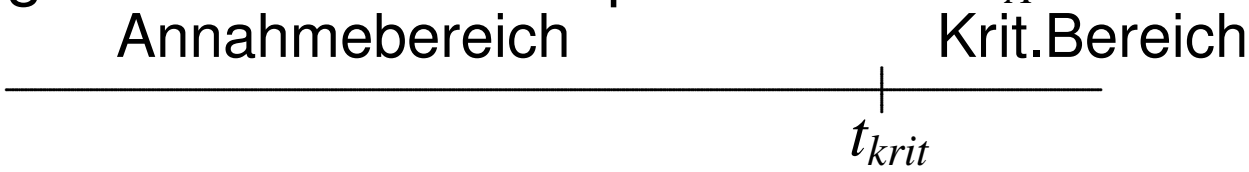
$T \in K \Rightarrow H_0$ ablehnen, Entscheidung für H_A
$T \notin K \Rightarrow H_0$ nicht ablehnen, Entscheidung für H_0.

Hypothesentest

Annahme- und Ablehnungsbereich

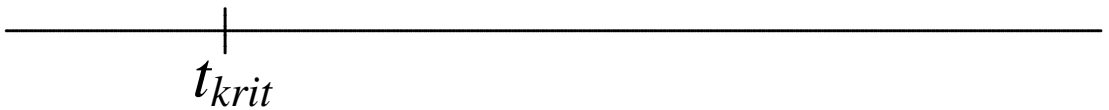
a) $H_0: \quad \mu \leq \mu_0 \qquad H_A: \mu > \mu_0$

große Werte von T sprechen für H_A.



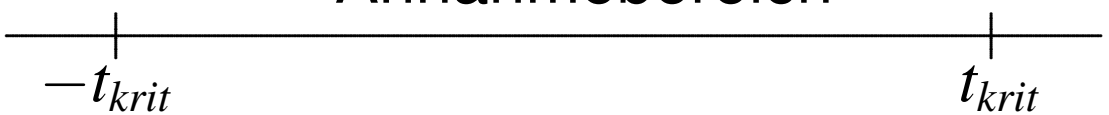
b) $H_0 \quad \mu \geq \mu_0 \qquad H_A: \quad \mu < \mu_0$

kleine Werte von T sprechen für H_A.



c) $H_0: \quad \mu = \mu_0 \qquad H_A: \quad \mu \neq \mu_0$

große Werte von $|T|$ sprechen für H_A.



Hypothesentest

Fehler 1. Art, Fehler 2. Art

Fehler 1.Art

Entscheidung für H_A obwohl H_0 richtig ist.

Fehler 2.Art

Entscheidung für H_0 obwohl H_A richtig ist

	Entscheidung für H_0	Entscheidung für H_A
H_0 richtig	richtig, Sicherheitswkt. $1-\alpha$	Fehler 1. Art Fehlerwkt. α.
H_A richtig	Fehler 2.Art Fehlerwkt. 1-β	richtig, Güte β

Entscheidung für H_0 heißt nicht notwendig, dass H_0 richtig ist.

Hypothesentest

Fehler 1. Art, Fehler 2. Art

α und $(1-\beta)$ können nicht gleichzeitig minimiert werden.

$\Rightarrow$ Man gibt α vor (z.B. $\alpha = 0.05$), d.h. man behält α unter Kontrolle und versucht die Teststatistik so zu definieren, daß β maximal wird.

β (und manchmal auch α) hängen von wahren (i.A. unbekannten) Parametern ab.

Signifikanzniveau

$$\alpha = \sup_{\theta \in \mathbf{\Theta}_0} \beta(\theta).$$

$\mathbf{\Theta}_0$: Nullhypothesenraum, also z.B. die Menge $\{\mu : \mu \geq \mu_0\}$ oder $\{\mu : \mu = \mu_0\}$.

Gütefunktion

Gütefunktion

$$\beta = \beta(\theta) = \beta(\mu) = P_\mu(T \in K)$$

K heißt Ablehnungsbereich oder Kritischer Bereich.

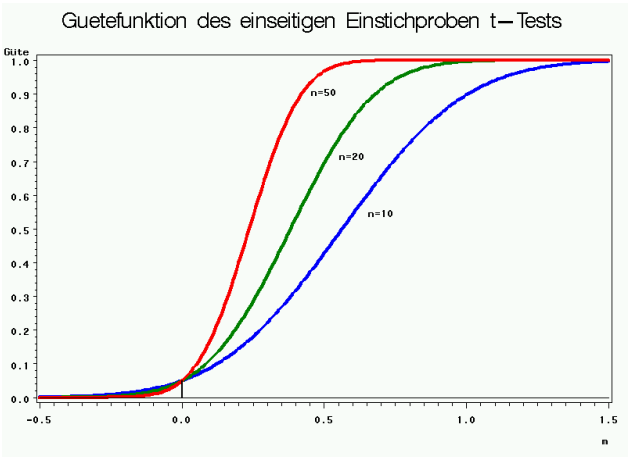
Beispiel: t-Test

$$\begin{aligned}\beta(\mu) &= P(T \in K) \qquad K\text{: kritischer Bereich} \\ &= P(T > t_{1-\alpha,n-1} | \mu, \sigma^2) \\ &= 1 - pt(t_{1-\alpha,n-1}, n-1, nc)\end{aligned}$$

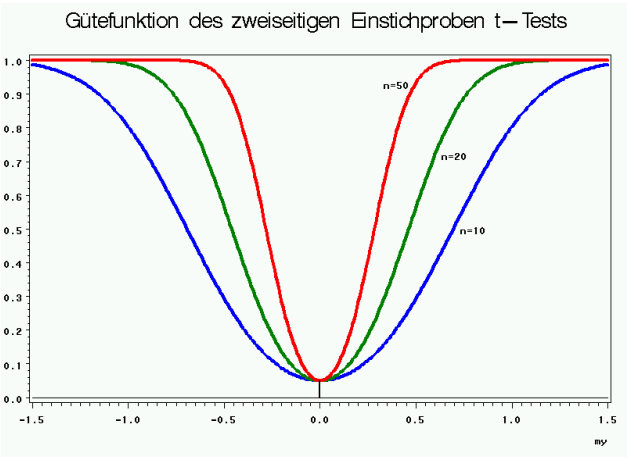
$nc = \sqrt{n}\frac{\mu-\mu_0}{\sigma}$:	Nichtzentralitätsparameter
$t_{1-\alpha,n-1}$:	kritischer Wert
$K = [t_{1-\alpha,n-1}, \infty)$:	kritischer Bereich.

Gütefunktion

Einseitiger Test



Zweiseitiger Test



Test_Guete_t.R

Test_Guete_t2.R

Gütefunktion

Ideal:

Unter H_0: Güte 0 (d.h. Fehler 1. Art =0)
Unter H_A: Güte 1 (d.h. Fehler 2. Art =0)

Das ist aber nicht möglich!

Ziel:

Test mit möglichst großer Gütefunktion (unter H_A).

Wir schlagen natürlich nur solche “sinnvollen” Tests vor.

Lagetests

(bei Normalverteilungsannahme, 1)

Einstichprobenproblem

$H_0 : \mu \leq \mu_0 \quad H_A : \mu > \mu_0$

$H_0 : \mu \geq \mu_0 \quad H_A : \mu < \mu_0$

$H_0 : \mu = \mu_0 \quad H_A : \mu \neq \mu_0$

Einstichproben t-Test

`t.test(x,mu,alternative)`

alt.: `"two.sided"`, `"less"` oder `"greater"`

Zweistichprobenproblem

$H_0 : \mu_1 \leq \mu_2 \quad H_A : \mu_1 > \mu_2$

$H_0 : \mu_1 \geq \mu_2 \quad H_A : \mu_1 < \mu_2$

$H_0 : \mu_1 = \mu_2 \quad H_A : \mu_1 \neq \mu_2$

Einstichproben t-Test (verbundene Stichproben)

t-Test (unverb. Stichproben)

`t.test(x,y,mu,alternative, paired)`

`paired=TRUE` verbunden

`mu` vermutete Differenz

Lage- und Skalentests

(bei Normalverteilungsannahme, 2)

c-Stichprobenproblem

$H_0 : \mu_1 = ... = \mu_c$ $H_A : \exists(i,j) : \mu_i \neq \mu_j$

einfache Varianzanalyse

`aov`,`lm`,`anova`

Andere Alternativen sind z.B.: $\mu_1 \leq ... \leq \mu_c$ $\mu_1 \geq ... \geq \mu_c$

Skalentest

Zwei unverbundene Stichproben

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_A : \sigma_1^2 \neq \sigma_2^2$$

`var.test` (nur bei Normalverteilung)

`ansari.test`,`levene.test` (Pkt. lawstat)

p-Werte

bisher: “H_0 abgelehnt” oder “H_0 beibehalten”
$\Rightarrow$ wenig informativ.

Wir könnten uns auch bei jedem α fragen, ob H_0 abgelehnt wird oder nicht.

Wenn der Test bei Signifikanzniveau α ablehnt, wird er das auch für $\alpha' > \alpha$ tun.

Es gibt also ein kleinstes α, bei dem der Test H_0 ablehnt.

Der p-Wert

ist das kleinste α, bei dem wir H_0 ablehnen können.

Test_t_p_value

p-Wert

T: (zufällige) Teststatistik, t: beobachtete Teststatistik

Nullhypopthese:

$$H_0 : \mu = \mu_0$$

Zweiseitige Alternative: $\mu \neq \mu_0$
p-Wert $= P_0(|T| > |t|)$

Einseitige Alternative: $\mu < \mu_0$
p-Wert $= P_0(T < t)$

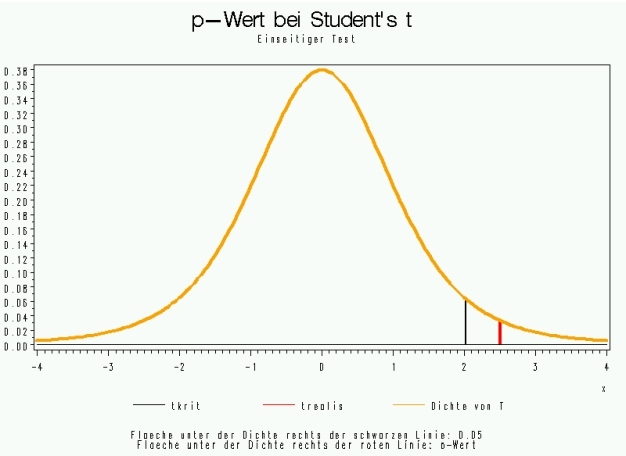
Einseitige Alternative: $\mu > \mu_0$
p-Wert $= P_0(T > t)$

Der p-Wert heißt auch Überschreitungswahrscheinlichkeit.

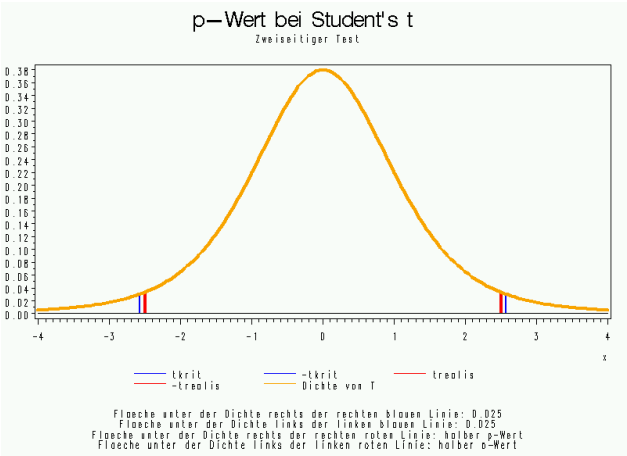
p-Wert

Illustration

Einseitiger Test



Zweiseitiger Test



Fäche unter der Dichte rechts der schwarzen Linie:

0.05 0.025

Fäche unter der Dichte rechts der roten Linie:

p-Wert halber p-Wert

links entsprechend.

Bewertung von p-Werten

Der p-Wert ist also, grob, ein Maß für den Grad dafür, dass die Nullhypothese nicht zutrifft.

(vorsichtige) Interpretation

p-Wert	Grad des Nicht-Zutreffens von H_0
< 0.01	sehr streng gegen H_0
$0.01 \ldots 0.05$	streng gegen H_0
$0.05 \ldots 0.1$	schwach gegen H_0
> 0.1	wenig oder gar nichts gegen H_0

Warnung:

Ein großer p-Wert heisst noch lange nicht, dass H_0 zutrifft. H_0 kann zutreffen,
Der große p-Wert kann aber auch daran liegen, dass der Test niedrige Güte hat!

p-Wert und kritischer Wert

Einseitige Alternative, $t_{krit} = t_{1-\alpha}$

$t \leq t_{krit} \Leftrightarrow$ p-Wert $\geq \alpha \Longrightarrow H_0$ angenommen,
$t > t_{krit} \Leftrightarrow$ p-Wert $< \alpha \Longrightarrow H_0$ abgelehnt.

Zweiseitige Alternative, $t_{krit} = t_{1-\alpha/2}$

$|t| \leq t_{krit} \Leftrightarrow$ p-Wert $\geq \alpha \Longrightarrow H_0$ angenommen,
$|t| > t_{krit} \Leftrightarrow$ p-Wert $< \alpha \Longrightarrow H_0$ abgelehnt.

Ausgabe bei R entpricht Wert von `alternative`

Der p-Wert ist **nicht** die Wahrscheinlichkeit, dass H_0 zurifft

$P(H_0|\text{Daten}) \neq$ p-Wert.

Inhalt

7.2 Einstichprobenproblem

	Nulhypothese	Alternative
a)	$H_0: \quad \mu \leq \mu_0$	$H_A: \mu > \mu_0$
b)	$H_0: \quad \mu \geq \mu_0$	$H_A: \mu < \mu_0$
c)	$H_0: \quad \mu = \mu_0$	$H_A: \mu \neq \mu_0$

Teststatistik

$$T(X_1, ..., X_n) = \frac{\overline{X} - \mu_0}{s} \cdot \sqrt{n}$$


'Student' in 1908

'Student'

Durchführung des Tests mit

`t.test(data,mu=`μ_0`)`

Einstichprobenproblem

Beispiel: Banknoten

Test_t1_Banknote.R

μ_0	gr	p-Wert Pr$>\|t\|$		
215	1	0.4258	$> \alpha = 0.05$	nosign
	2	< 0.0001	$< \alpha = 0.05$	sign.
214.9	1	0.0784	$> \alpha = 0.05$	nosign.
	2	0.03	$< \alpha = 0.05$	sign.

Das sind also zweiseitige p-Werte (Alternative c)).
Was machen wir bei Alternative a) oder b)? $\rightarrow$ s.u.

vorgegeben: Fehler 1.Art α (Signifikanzniveau)
(üblich ist $\alpha = 0.05$ oder $\alpha = 0.01$)
d.h. $P_{\mu_0}(|T| > t_{krit}) = \alpha$.

Verteilung der Teststatistik T

Nehmen wir in unserem Beispiel an, die Beobachtungen

$$X_i \sim \mathcal{N}(\mu_0, \sigma^2), \quad , i = 1, \ldots, n$$

sind normal und unabhängig, dann hat die (zufällige) Teststatistik T eine t-Verteilung (Student's t),

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{s} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} =: t_{n-1}$$

(t-Verteilung mit $n-1$ Freiheitsgraden) und

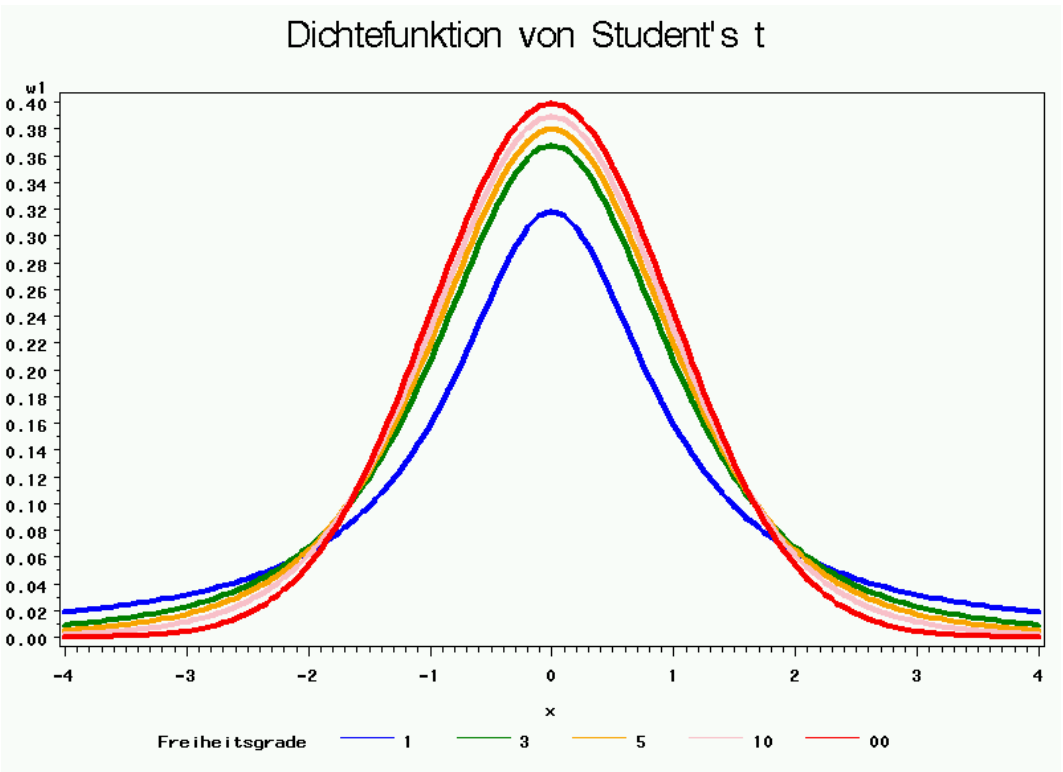
$$t_{krit} = t_{1-\frac{\alpha}{2},n-1}$$

ist $(1-\frac{\alpha}{2})$ - Quantil einer t-Verteilung mit $n-1$ Freiheitsgraden.

Dichtefunktion einer t-Verteilung

mit $\nu(= n - 1)$ Freiheitsgraden (FG)

$$f_{t_\nu}(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu \cdot \pi} \cdot \Gamma(\frac{\nu}{2})} \cdot (1 + \frac{x^2}{\nu})^{-\frac{\nu+1}{2}} \qquad dt(x, \nu)$$



Test_t_Dichte.R

Einstichprobenproblem

t-Test

a) $H_0 : \mu \leq \mu_0 \qquad H_A : \mu > \mu_0$
$\Rightarrow$ große Werte von

$$T = \frac{\overline{X} - \mu_0}{s} \cdot \sqrt{n}$$

indizieren Gültigkeit von H_A.

b) $H_0 : \mu \geq \mu_0 \qquad H_A : \mu < \mu_0$
$\Rightarrow$ kleine Werte von T indizieren H_A

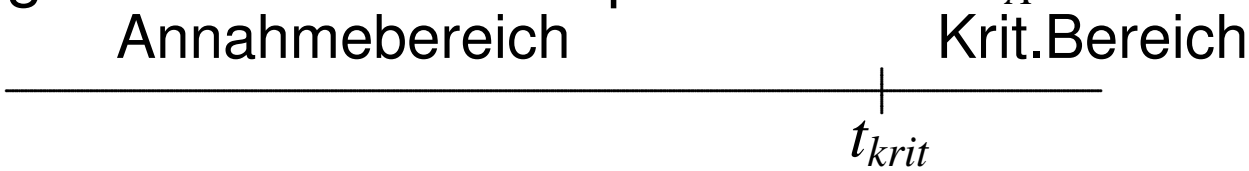
c) $H_0 : \mu = \mu_0 \qquad H_A : \mu \neq \mu_0$
$\Rightarrow |T|$ groß indiziert Gültigkeit von H_A.

Hypothesentest

Annahme- und Ablehnungsbereich

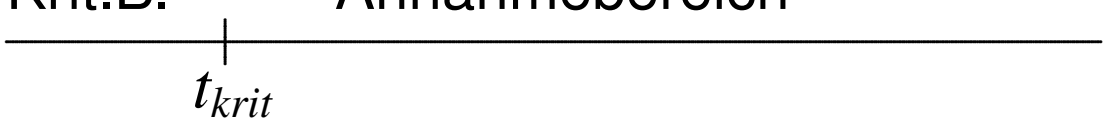
a) $H_0: \quad \mu \leq \mu_0 \qquad H_A: \mu > \mu_0$

große Werte von T sprechen für H_A.



b) $H_0 \quad \mu \geq \mu_0 \qquad H_A: \quad \mu < \mu_0$

kleine Werte von T sprechen für H_A.



c) $H_0: \quad \mu = \mu_0 \qquad H_A: \quad \mu \neq \mu_0$

große Werte von $|T|$ sprechen für H_A.

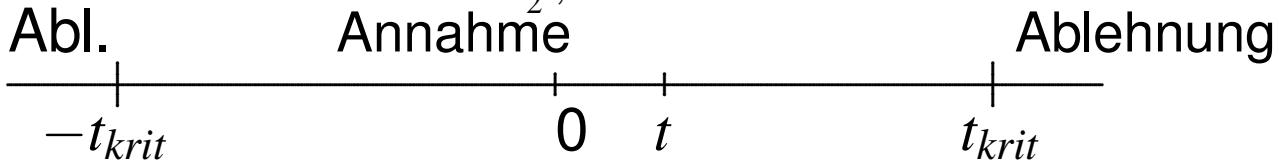
Hypothesentest

Sei jetzt t eine Realisierung von T.

Zweiseitige Alternative $H_A : \mu \neq \mu_0$

Wenn $|t| > t_{krit} = t_{1-\frac{\alpha}{2},n-1}$ so H_0 abgelehnt.

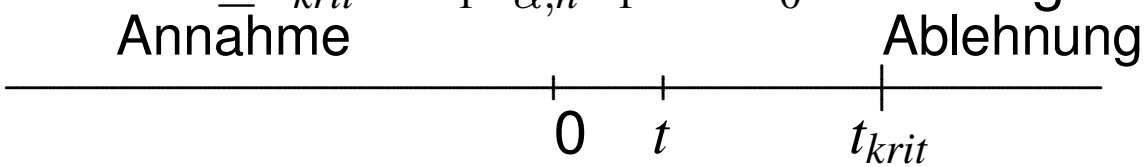
Wenn $|t| \leq t_{krit} = t_{1-\frac{\alpha}{2},n-1}$ so H_0 nicht abgel.



Einseitige Alternative $H_A : \mu > \mu_0$

Wenn $t > t_{krit} = t_{1-\alpha,n-1}$ so H_0 abgelehnt.

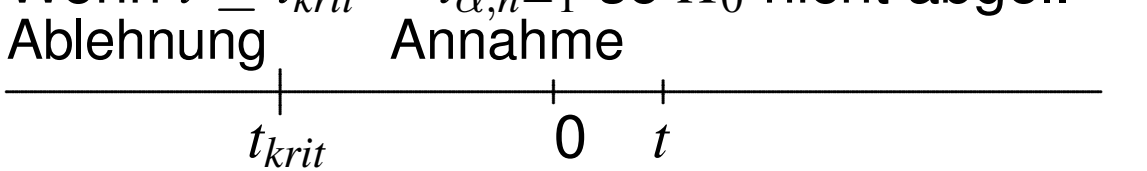
Wenn $t \leq t_{krit} = t_{1-\alpha,n-1}$ so H_0 nicht abgel.



Einseitige Alternative: $H_A : \mu < \mu_0$

Wenn $t < t_{krit} = t_{\alpha,n-1}$ so H_0 abgelehnt.

Wenn $t \geq t_{krit} = t_{\alpha,n-1}$ so H_0 nicht abgel.



p-Werte bei einseitigen Alternativen (1)

Erinnerung: Der zweiseitige p-Wert ist: $P(|T| > t)$.

$$\begin{aligned}
P(|T| > t) &= P((T > t) \vee (-T > t)) \\
&= P((T > t) \vee (T < -t)) \\
&= 2 \cdot P(T > t), \qquad t > 0 \\
P(T > t) &= P(T < -t) \\
&= 1 - P(T \geq -t) \\
&= 1 - \frac{1}{2} P(|T| > -t), \qquad t \leq 0
\end{aligned}$$

(Die Verteilung von T ist stetig und symmetrisch.)

p-Werte bei einseitigen Alternativen (2)

Fall a) $H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$

$$\text{p-Wert} = P(T > t) = \begin{cases} \frac{1}{2}P(|T| > t), & \text{falls } t > 0 \\ 1 - \frac{1}{2}P(|T| > -t), & \text{falls } t \leq 0 \end{cases}$$

Ablehnung von H_0 falls $P(T > t) < \alpha$.

```
t.test(data,mu=μ0,alternative="greater")
```

Fall b) $H_0 :$ $\mu \geq \mu_0$ $H_a : \mu < \mu_0$

$$\text{p-Wert} = P(T < t) = \begin{cases} \frac{1}{2}P(|T| > |t|), & \text{falls } t \leq 0 \\ 1 - \frac{1}{2}P(|T| > -t), & \text{falls } t > 0 \end{cases}$$

Ablehnung von H_0 falls $P(T < t) < \alpha$

```
t.test(data,mu=μ0,alternative="less")
```

Zusammenfassung Einstichprobenproblem (1)

Teststatistik

$$T = \sqrt{n} \cdot \frac{\overline{X} - \mu_0}{S} \qquad \text{Realisierung } t$$

$$\overline{X} = \frac{1}{n} \sum_i X_i, \qquad S^2 = \frac{1}{n-1} \sum_i (X_i - \overline{X})^2$$

Zweiseitige Alternative, $H_0 : \mu = \mu_0 \qquad H_A : \mu \neq \mu_0$

$\lvert t \rvert > t_{krit}$	$\Leftrightarrow$	H_0 ablehnen
p-value $< \alpha$	$\Leftrightarrow$	H_0 ablehnen
"Pr $> \lvert t \rvert$" $< \alpha$	$\Rightarrow$	H_0 ablehnen

Zusammenfassung Einstichprobenproblem (2)

Einseitige Alternative, $H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$

$t > 0$ und $\frac{\text{p-value}}{2} < \alpha \Leftrightarrow H_0$ ablehnen

Einseitige Alternative, $H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$

$t < 0$ und $\frac{\text{p-value}}{2} < \alpha \Leftrightarrow H_0$ ablehnen

Konfidenzbereiche (1)

am Beispiel des t-Tests

$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \sqrt{n} \cdot \frac{\overline{X}-\mu}{S} \sim t_{n-1}$ wenn μ der wahre (Lokations-) Parameter ist. $\Rightarrow$

$$P(\underbrace{-t_{1-\frac{\alpha}{2},n-1} \leq \sqrt{n} \cdot \frac{\overline{X}-\mu}{s} \leq t_{1-\frac{\alpha}{2},n-1}}_{(*)}) = 1-\alpha$$

Die Ungleichungen sind äquivalent zu

$$\begin{aligned}
(*) \quad &\Leftrightarrow \quad -\frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \leq \overline{X} - \mu \leq \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \\
&\Leftrightarrow \quad -\overline{X} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \leq -\mu \leq -\overline{X} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \\
&\Leftrightarrow \quad \overline{X} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \geq \mu \geq \overline{X} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \\
&\Leftrightarrow \quad \overline{X} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1} \leq \mu \leq \overline{X} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2},n-1}
\end{aligned}$$

Konfidenzbereiche (2)

$(1-\alpha)$ Konfidenzintervall für den (unbekannten) Parameter μ

$$\left[\overline{X} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2},n-1}, \overline{X} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2},n-1}\right]$$

`t.test(..., conf.level=`$1-\alpha$`)`

Konfidenzbereiche (3)

$(1-\alpha)$ Konfidenzintervall für den (unbekannten) Median

$$[x(k), x(n-k+1)] \text{ mit}$$
$$P(Y_n < k) \leq \frac{\alpha}{2} \text{ und } Y_n \sim B(n, 0.5)$$

```
n = length(x)
k = qbinom(α/2, n, 0.5)
sorted = sort(x)
confint= c(sorted[k],sorted[n-k+1])
```

Konfidenzbereiche (4)

Beispiel

Test_t1_Banknote

$(1-\alpha)$-Konfidenzintervalle für den Lageparameter $\mu = \mathbf{E}$'laenge':

	echt		gefälscht	
$\alpha = 0.01$	214.87	215.07	214.73	214.92
$\alpha = 0.05$	214.89	215.05	214.75	214.89
$\alpha = 0.05$	214.9	215.1	214.7	214.9
verteilungsfr. KI (für Median)				

`t.test(...,conf.level =` $1-\alpha$`)`

verteilungsfrei: `confint` der vorherigen Folie

Einseitige Konfidenzintervalle mit

`t.test(...,alternative="less")` bzw. `"greater"`

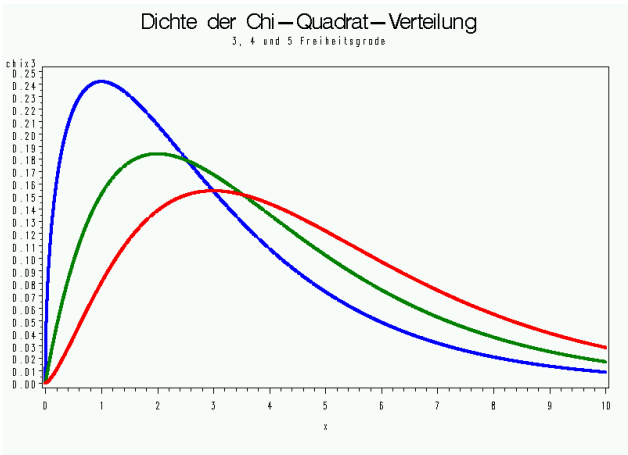
Konfidenzintervalle für σ^2

bei Normalverteilung

$$X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2), \text{unabhängig} \quad \Rightarrow \quad (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Dichte einer χ^2_ν-Verteilung

$$f_{\chi^2_\nu}(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} e^{-x/2} x^{\nu/2-1} & \text{falls} \quad x \geq 0 \\ 0 & \text{sonst.} \end{cases}$$



Test_Chi2_Dichte

Konfidenzintervall für σ^2 (2)

bei Normalverteilung

$$P\left(\chi^2_{\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}\right) = 1-\alpha$$

auflösen nach σ^2:

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}\right) \\
&= P\left(\frac{1}{\chi^2_{1-\alpha/2,n-1}} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi^2_{\alpha/2,n-1}}\right) \\
&= P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right)
\end{aligned}$$

Konfidenzintervall für σ^2 (3)

nur bei Normalverteilung!

Konfidenzintervall

(Vertrauensintervall) für den (unbekannten) Parameter σ^2

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$$

```
alpha = α
n = length(x)
confint.var = (n-1)*var(x)/
qchisq(c(1-alpha/2,alpha/2),n-1)
```

Inhalt

7.3 Vergleich zweier abhängiger Gruppen (verbundene Stichproben)

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$
$$H_0 : \mu_1 \geq \mu_2 \quad H_1 : \mu_1 < \mu_2$$
$$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$$

- Gewicht einer Person zu den Zeitpunkten t_1, t_2.
- Banknoten (oben- unten, links - rechts)
- Patient nimmt Medikament 1 und 2
- Kreuz- und selbstbefruchtete Pflanzen

Test_t2_Banknote Test_t2_Darwin

Vergleich zweier abhängiger Gruppen

Folgende Möglichkeiten:

a) Transformation $Z := X_1 - X_2$ und testen auf $\mu = 0$

```
t.test(x1-x2)
```

b) Mit der 2 Argumenten und `paired`:

```
t.test(x1,x2,paired=TRUE)
```

Inhalt

7.4 Vergleich zweier unabhängiger Gruppen (unverbundene Stichproben)

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$
$$H_0 : \mu_1 < \mu_2 \qquad H_1 : \mu_1 \geq \mu_2$$
$$H_0 : \mu_1 > \mu_2 \qquad H_1 : \mu_1 \leq \mu_2$$

- Tibetische Schädel (Sikkim - Kham)
- Wasserhärte (Nord - Süd)
- Klinikaufenthalt (Klinik1 - Klinik2)
- Banknoten (echt - gefälscht)

Test_t2_Tibetan Test_t2_Heroin Test_t2_Banknote

Vergleich zweier unabhängiger Gruppen (2)

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Fall 1: Varianzen σ_1^2, σ_2^2 sind gleich
Fall 2: Varianzen σ_1^2, σ_2^2 sind verschieden

Fall 1:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{}$$

Vergleich zweier unabhängiger Gruppen (3)

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

Fall 1: Varianzen σ_1^2, σ_2^2 sind gleich
Fall 2: Varianzen σ_1^2, σ_2^2 sind verschieden

Fall 1:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}}$$

n, m: Umfänge Stichprobe 1 und 2

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{1i} - \overline{X}_1)^2, \qquad S_2^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_{2i} - \overline{X}_2)^2$$

Erläuterung des Quotienten T

$X_1 \sim \mathcal{N}(\mu_1, \sigma^2), X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$

$$\overline{X}_1 \sim \mathcal{N}(\mu_1, \sigma^2 \cdot \frac{1}{n}), \quad \overline{X}_2 \sim \mathcal{N}(\mu_2, \sigma^2 \cdot \frac{1}{m})$$

$$\frac{(n-1)}{\sigma^2} \cdot S_1^2 \sim \chi^2_{n-1}, \quad \frac{(n-1)}{\sigma^2} \cdot S_2^2 \sim \chi^2_{m-1}$$

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}(\mu_1 - \mu_2, \sigma^2 \cdot (\frac{1}{n} + \frac{1}{m}))$$

$$\frac{1}{\sigma^2}\left((n-1) \cdot S_1^2 + (m-1) \cdot S_2^2\right) \sim \chi^2_{n+m-2}$$

$$T \sim t_{n+m-2}$$

unter H_0 ($\mu_1 = \mu_2$).

Vergleich zweier unabhängiger Gruppen (4)

T ist eine Zufallsgröße!

Werte von T werden mit gewissen Wahrscheinlichkeiten angenommen!

Die Wahrscheinlichkeit dafür, daß T sehr große Werte annimmt (wenn H_0 richtig ist) ist also sehr klein.

Sei jetzt t eine Realisierung von T (also der Wert, der bei Ausrechnen anhand der gegebenen Daten entsteht).

Wenn jetzt t sehr groß, $|t| \in K$ (krit. Bereich)
(aber die Wkt. dafür ist sehr klein, wenn H_0 richtig ist)
$\Rightarrow H_0$ ablehnen.

Vergleich zweier unabhängiger Gruppen (ungleiche Varianzen)

Fall 2: Varianzen ungleich

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$$

$T \sim t_\nu$ approximativ. Die Zahl ν der Freiheitsgrade wird auch approximativ berechnet. (Welch-Test, 1937)

R bietet Tests für beide Fälle (gleiche, ungleiche Varianzen) an. Satterthwaite-Approximation (1946).

`t.test(X1,X2,var.equal=TRUE)` bei gleichen Varianzen, `t.test(X1,X2)` sonst (`var.equal=FALSE` ist Standard).

Vergleich zweier unabhängiger Gruppen

Welchen Test soll man nehmen?

- Aus Vorinformation ist vielleicht bekannt, ob man gleiche Varianzen annehmen kann.
- Man könnte einen Test auf gleiche Varianzen vorschalten

Problem: 2 stufiger Test

Wird das Signifikanzniveau eingehalten??

Vorschlag

gleich den t-Test für ungleiche Varianzen nehmen
ist einigermaßen robust gegen Abweichungen von der Normalverteilung, aber nicht gegen Ausreißer

Inhalt

7.5 Test auf Gleichheit der Varianzen

Voraussetzung: Normalverteilung

$H_0 : \sigma_1^2 = \sigma_2^2$ $\qquad$ $H_1 : \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{S_1^2}{S_2^2} \sim F_{n-1,m-1}$$

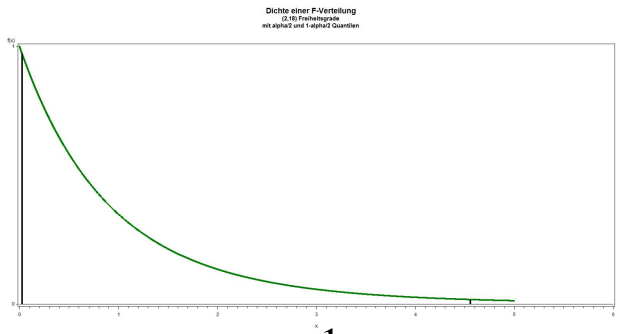
(Fisher-) F- Verteilung mit $(n-1, m-1)$ Freiheitsgraden.

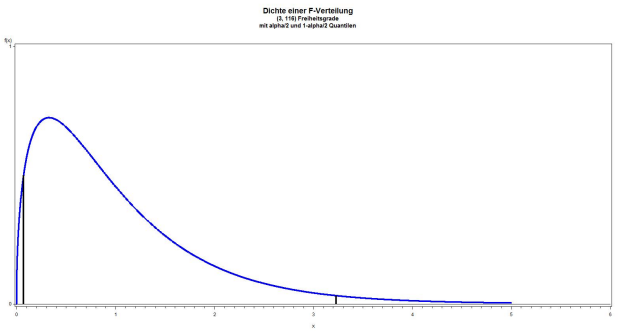
F ist Quotient zweier unabhängiger χ^2-verteilter Zufallsgrößen.
H_0 ablehnen, falls

$$\frac{s_1^2}{s_2^2} < F_{\frac{\alpha}{2},n-1,m-1} \quad \text{oder} \quad \frac{s_1^2}{s_2^2} > F_{1-\frac{\alpha}{2},n-1,m-1}$$

Test auf Gleichheit der Varianzen

F-Test





$$F_{\frac{\alpha}{2},n-1,m-1} = \frac{1}{F_{1-\frac{\alpha}{2},m-1,n-1}}$$

(beachten: Freiheitsgrade vertauschen sich)

$\Rightarrow H_0$ ablehnen, falls

$$\frac{s_1^2}{s_2^2} < \frac{1}{F_{1-\frac{\alpha}{2},m-1,n-1}} \quad \text{oder} \quad \frac{s_1^2}{s_2^2} > F_{1-\frac{\alpha}{2},n-1,m-1} \quad \Leftrightarrow$$

$$\frac{s_2^2}{s_1^2} > F_{1-\frac{\alpha}{2},m-1,n-1} \quad \text{oder} \quad \frac{s_1^2}{s_2^2} > F_{1-\frac{\alpha}{2},n-1,m-1}$$

Test auf Gleichheit der Varianzen

F-Test, prakt. Durchführung

$s_M^2 := max(s_1^2, s_2^2)$ $\quad$ $s_m^2 := min(s_1^2, s_2^2)$

n_M, n_m: die entsprechenden Stichprobenumfänge

$$\Rightarrow H_0 \text{ ablehnen, falls } \frac{s_M^2}{s_m^2} > F_{1-\frac{\alpha}{2}, n_M-1, n_m-1}.$$

Formulierung mit p-Werten

$$\Rightarrow H_0 \text{ ablehnen, falls p-Wert} = P(F > \frac{s_M^2}{s_m^2}) < \frac{\alpha}{2}$$

$F \sim F_{n_M-1, n_m-1}$

`var.test(X1,X2,ratio)` (`ratio=1` ist Standard)

Test_F_Dichte

Inhalt

Ein- und Zweistichprobenproblem

Anmerkungen (1)

- Der F-Test (zum Skalenvergleich) ist sehr empfindlich gegenüber Abweichungen von der Normalverteilungsannahme

$\Rightarrow$ mit größter Vorsicht genießen.

- Der Einstichproben- t-Test ist nicht robust!
- Der Zweistichproben t-Test ist etwas robuster als der t-Test im Einstichprobenproblem
- Ausreißer können extremen Einfluss haben (ÜA).
- Wenn Gleichheit der Varianzen unklar $\Rightarrow$

t-Test mit ungleichen Varianzen nehmen.

(ist bei gleichen Varianzen nur ganz wenig weniger effizient)

Ein- und Zweistichprobenproblem

Anmerkungen (2)

- Besser nicht auf das Ergebnis des F-Tests verlassen. (Problematik: 2-Stufentest, Nicht-Robustheit).
- Es gibt robustere Skalentests

$\Rightarrow$ Levene Test und Brown-Forsythe Test.

Inhalt

Test auf Gleichheit der Varianzen

Levene-Test

Bilden die Werte

$$\begin{aligned} X_j^* &:= |X_j - \overline{X}| \\ Y_j^* &:= |Y_j - \overline{Y}| \end{aligned}$$

Skalenunterschiede in (X, Y) spiegeln sich jetzt in Lageunterschieden in (X^*, Y^*) wieder.
Mit den “neuen Beobachtungen” wird jetzt ein t-Test durchgeführt.
Die t-Verteilung der entsprechenden Teststatistik gilt nur approximativ.

Test auf Gleichheit der Varianzen

Brown-Forsythe Test

Analog zum Levene-Test, nur hier bilden wir die Werte

$$\begin{aligned} X_j^* &:= |X_j - \text{med}_i X_i| \\ Y_j^* &:= |Y_j - \text{med}_i Y_i| \end{aligned}$$

Beide Tests, Levene und Brown-Forsythe, sind (einigermaßen) robust gegen Abweichungen von der Normalverteilung.

Test auf Gleichheit der Varianzen

Syntax

`levene.test(y,group,...)` erwartet eine Variable und einen gleichlangen Gruppierungsvektor. Für Stichproben als separate Vektoren ist daher eine Umformung nötig:

```
require(lawstat)
xf = data.frame(
rbind(cbind(val=x1,fact=1),cbind(x2,2))
#mean = Levene
levene.test(xf$val,xf$fact,location="mean")
#median = Brown-Forsythe
levene.test(xf$val,xf$fact,location="median")
```

oder (hässlich, aber kurz):

```
levene.test(c(x1,x2),c(x1^0,x1^0+1))
```

Test_t2_Banknote

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

8. Varianzanalyse

8.1 Vergleich von k unabhängigen Gruppen

einfaktorielle, einfache Varianzanalyse

A: Faktor (Gruppenvariable) mit k Stufen (Faktorstufen)

Modell

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1...k, j = 1...n_i$$

μ:	Gesamterwartungswert
α_i:	Effekt der i-ten Stufe von A
ϵ_{ij}:	Fehler, $\epsilon_{ij} \sim (0, \sigma^2)$
Y_{ij}:	j-te Beobachtung der i-ten Faktorstufe
$\sum_{i=1}^{k} \alpha_i = 0$	Parametrisierungsbedingung

Einfache Varianzanalyse

$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k$
$H_1 : \alpha_i \neq \alpha_l$ (für ein $i \neq l$)

Im Fall $k = 2$ führt dieses Testproblem auf das Zweistichprobenproblem ($\rightarrow$ t-Test).

Output der Maschinen gleich?
Klausurergebnisse unterschiedlich?
Mageninhalt der Eidechsen gleich?
Cortisolgehalt unterschiedlich?

ANOVA_Maschinen
Varianzanalyse_Modelle\PI12erg
GLM_Eidechsen GLM_Cortisol

Varianzanalyse

Varianzanalyse macht eine Streuungszerlegung:

Gesamt-varianz	=	Varianz zwischen den Faktorstufen	+	Varianz innerhalb der Faktorstufen	
SST (Total)	=	SSB (Between)	+	SSW (Within)	(SSE) (Error)

$$N = \sum_{i=1}^{k} n_i$$

$$\overline{Y}_i = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} Y_{ij}, \qquad \overline{Y} = \frac{1}{N} \sum_{i,j} Y_{i,j}$$

Einfache Varianzanalyse

Satz: Es gilt

$$SSB + SSW = SST$$

wobei

$$SSB = \sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2 \qquad \text{(Between)}$$

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2 \qquad \text{(Within)}$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2. \qquad \text{(Total)}$$

Einfache Varianzanalyse

Satz: $SSB + SSW = SST$

Beweis:

$$SSB = \sum_i n_i \overline{Y}_i^2 - 2 \cdot N \cdot \overline{Y}^2 + \overline{Y}^2 \cdot N$$

$$SSW = \sum_{i,j} Y_{ij}^2 - 2 \cdot \sum_i n_i \overline{Y}_i^2 + \sum_i n_i \overline{Y}_i^2$$

$$SSB + SSW =$$

$$= \sum_{i,j} Y_{ij}^2 + \sum_i n_i \overline{Y}_i^2 - N \cdot \overline{Y}^2 - \sum_i n_i \overline{Y}_i^2$$

$$= \sum_{i,j} Y_{ij}^2 - N \cdot \overline{Y}^2 = \sum_j \sum_i (Y_{ij} - \overline{Y})^2 = SST \quad \blacksquare$$

Varianzanalyse in R

anova

```
anova(model)
```

berechnet Varianzanalyse von Modellen (s. `lm`).

aov

```
aov(formula,data)
```

vereint beide Schritte (Modell und ANOVA), mit `summary` wird auch ein p-Wert ausgegeben.

```
anova(lm(v ~ fac,dat))
summary(aov(v ~ fac,dat))
```

wobei `dat` die Variable `v` und den Faktor `fac` enthält.

Einschub: Faktoren in R

- Faktoren sind Daten mit sog. nominalem Niveau, d.h. sie können lediglich nach identisch/nicht identisch unterschieden werden
- Beispiel Lage: Nord/Süd, Geschlecht
- R speichert Faktoren intern als Integer
- Die verschiedenen möglichen Werte heißen Level
- `factor(c(1,1,2))` und `c(1,1,2)` sehen gleich aus, werden aber von `lm` anders behandelt (ÜA?)

Einfache Varianzanalyse (1)

Response: v

	Df	Sum Sq	MeanSq	F value	$Pr(>F)$
fac	k-1	SSB(M)	MSB	$\frac{MSB}{MSE}$	p-Wert
Residuals	N-k	SSW(E)	MSE		

$$MSB = \frac{SSB}{k-1}, \qquad MSE = \frac{SSW}{N-k}$$

$$H_0: \quad \alpha_1 = \cdots = \alpha_k$$
$$H_1: \quad \exists (i,j) : \alpha_i \neq \alpha_j$$

Einfache Varianzanalyse (2)

H_0 wird getestet mit

$$F = \frac{MSB}{MSE} = \frac{\text{Mittlere Var. zwischen d. Gruppen}}{\text{Mittlere Var. innerhalb d. Gruppen}}$$

$$= \frac{N-k}{k-1} \qquad \frac{SSB}{SSW} = \frac{N-k}{k-1} \qquad \frac{SST - SSW}{SSW}$$

F groß, $F > F_{1-\alpha,k-1,N-k} \Leftrightarrow H_0$ abgelehnt

Bestimmtheitsmaß

$$R^2 := \frac{SSB}{SST} = \frac{SST - SSW}{SST} = 1 - \frac{SSW}{SST}$$

Der Anteil der Varianz, der durch das Modell bestimmt wird, heißt Bestimmtheitsmaß

Einfache Varianzanalyse (3)

Offenbar: $0 \leq R^2 \leq 1$.

$$F = \frac{MSB}{MSE} = \frac{N-k}{k-1} \cdot \frac{SSB}{SST} \cdot \frac{SST}{SSW} = \frac{N-k}{k-1} \cdot \frac{R^2}{1-R^2}$$

$$R^2 \to 0 \quad \Longrightarrow \quad F \to 0$$
$$R^2 \to 1 \quad \Longrightarrow \quad F \to \infty.$$

Schätzung der Modellstandardabweichung σ

$$RootMSE = \sqrt{MSE} = \sqrt{\tfrac{1}{N-k} SSE}$$

Variationskoeffizient

$$CV = \frac{100 \cdot RootMSE}{\overline{Y}}$$

Einfache Varianzanalyse

Anmerkungen

- Der F-Test in der Varianzanalyse ist (einigermaßen) robust gegenüber Abweichungen von der Normalverteilungsannahme
- Die Funktion `lm` liefert sehr viele Ausgaben, die sich mit `plot` abbilden und mit weiteren Funktionen auswerten lassen.
  `residuals` gibt die Residuen eines Modells zurück.
- F-Test verlangt auch Varianzhomogenität
  Daten balanciert (gleiche Stichprobenumfänge)
  $\rightarrow$ Abweichungen nicht so schwerwiegend.

Einfache Varianzanalyse

Test auf Varianzhomogenität

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$
$$H_1 : \exists(i,l) : \sigma_i^2 \neq \sigma_l^2$$

Levene Test (1960)

```
levene.test(...,location="mean")
```

$$Z_{ij}^* = |Y_{ij} - \overline{Y}_{i.}|$$

Brown-Forsythe-Test (1974)

```
levene.test(...,location="median")
```

$$Z_{ij}^* = |Y_{ij} - medY_i|$$

Einfache Varianzanalyse

Test auf Varianzhomogenität (2)

Mit diesen neuen ZV wird eine Varianzanalyse durchgeführt.

$$W = \frac{\frac{1}{k-1}\sum n_i(\overline{Z}^*_{i.} - \overline{Z}^*)^2}{\frac{1}{N-k}\sum_{i,j}(Z^*_{ij} - \overline{Z}^*_{i.})^2} \sim F_{k-1,N-k}.$$

GLM_Cortisol

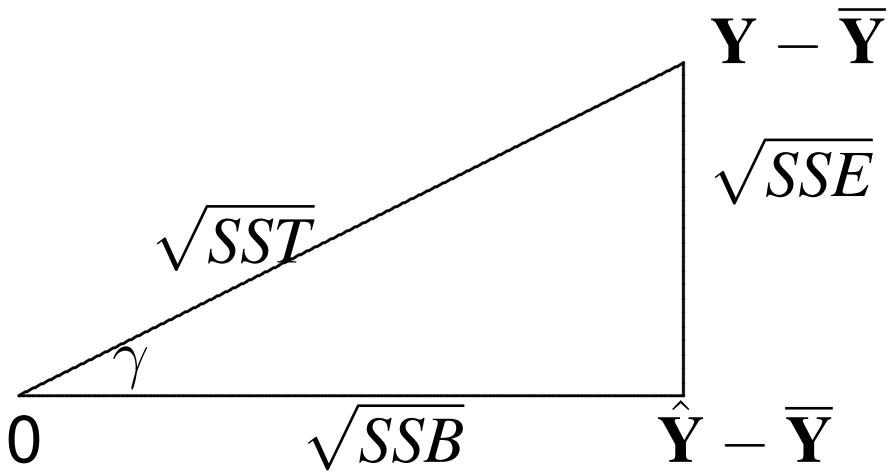
Geometrische Veranschaulichung

zur Varianzanalyse

$$\mathbf{Y} = (Y_{11}, \ldots, Y_{kn_k}) \quad \text{Dimension } N$$

$$\hat{\mathbf{Y}} = (\underbrace{\overline{Y}_1, \ldots, \overline{Y}_1}_{n_1\text{mal}}, \ldots, \underbrace{\overline{Y}_k, \ldots, \overline{Y}_k}_{n_2\text{mal}})$$

$$\overline{\mathbf{Y}} = (\underbrace{\overline{Y}, \ldots, \overline{Y}}_{N\text{mal}}), \quad \overline{Y} = \frac{1}{N}\sum_{i,j} Y_{ij}$$



$$SSB + SSW = SST \qquad R^2 = cos^2\gamma$$

$$||\hat{\mathbf{Y}} - \overline{Y}||^2 + ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \overline{Y}||^2$$

Inhalt

8.2 Multiple Vergleiche

Problemstellung: H_0 abgelehnt, aber zwischen welchen Faktorstufen liegt der Unterschied?

- Idee: Alle Paarvergleiche machen.
- Problem: Wenn wir das Signifikanzniveau $\alpha(= 0.05)$ so lassen, wird das Testniveau nicht eingehalten!
- Veranschaulichung: Bei 20 gleichzeitigen Tests können wir $20 \cdot \alpha = 1$ Ablehnung erwarten, auch wenn H_0 richtig ist.

Multiple Vergleiche

Lösungsmöglichkeiten (1)

Bonferroni-Korrektur

Signifikanzniveau für die gleichzeitigen Tests herabsetzen auf $\frac{\alpha_{nom}}{\binom{k}{2}}$,
bei $k = 4$ wäre das etwa $\frac{\alpha_{nom}}{\binom{4}{2}} = \frac{0.05}{6}$.
Begründung: Bonferroni-Ungleichung.

A_i: Ereignis, H_{0i} (i-ter Paarvergleich) abgelehnt.

$$\underbrace{P_0\left(\bigcup A_i\right)}_{\text{Wkt,}H_{0i}\geq 1\text{mal abgelehnt}} \leq \sum_{i=1}^{M} P(A_i) \leq M \cdot \frac{\alpha}{M} = \alpha$$

M: Anzahl der Paarvergleiche.

Multiple Vergleiche

Lösungsmöglichkeiten (1)

Bonferroni-Korrektur in R

`pairwise.t.test(x, g, p.adjust.method,pool.sd)` mit `p.adjust.method = "bonferroni"` führt für jedes Paar von Faktorstufen aus `g` einen t-Test aus und passt die p-Werte an.
`pool.sd` (Standard: `TRUE`) legt fest, ob die Varianzen gemeinsam oder separat geschätzt werden sollen.

Bem.: Es gibt eine Fülle weiterer Methoden (s. `?p.adjust`).

Multiple Vergleiche

Lösungsmöglichkeiten (2)

Tukeys „Honest Significant Difference“-Methode

Bilden die $\overline{Y}_j$ und die Spannweite dazu $w = max_{i,j}|\overline{Y}_i - \overline{Y}_j|$.
Dazu kommt noch die empirische Standardabweichung s.

$$t\text{max} = \frac{w}{s}$$

die sogenannnte *studentisierte Spannweite*.
Diese hat (wenn die $Y_i \sim \mathcal{N}$) eine (dem R-Programmierer) wohlbekannte Verteilung, und entsprechende Quantile und kritische Werte.
Damit erhalten wir simultane Konfidenzintervalle für alle Paardifferenzen $\mu_i - \mu_j$. Liegt 0 nicht darin, so wird $H_{0,ij}: \ \mu_i = \mu_j$ abgelehnt zugunsten von $H_{A,ij}: \ \mu_i \neq \mu_j$.

Multiple Vergleiche

Lösungsmöglichkeiten (2)

Tukeys „Honest Significant Difference“-Methode in R

`TukeyHSD(aov(v~fac,dat))`

`TukeyHSD` wird auf das Ergebnis von `aov` angewendet (funktioniert nicht mit `lm` oder `anova(lm())`).

Inhalt

8.3 Vergleich von k abhängigen Gruppen

(2-faktorielle Varianzanalyse)

Modell:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim (0, \sigma^2)$$

$i = 1, \ldots, a, j = 1, \ldots, b.$
(eine Beobachtung je Zelle)

Das Modell ist überparametrisiert, deswegen Bedingung:
$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0.$
Folgene Hypothesen sind zu testen:

$H_{0a}: \quad \alpha_1 = \cdots = \alpha_a = 0$ gegen $H_{1a}: \quad \exists (i_1, i_2) : \alpha_{i_1} \neq \alpha_{i_2}$

$H_{0b}: \quad \beta_1 = \cdots = \beta_b = 0$ gegen $H_{1a}: \quad \exists (j_1, j_2) : \beta_{j_1} \neq \beta_{j_2}$

GLM_Synchro GLM_Cache

2-faktorielle Varianzanalyse

$$\overline{Y}_{..} = \frac{1}{a \cdot b} \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{ij} \quad \text{arith. Mittel aller Beob.}$$

$$\overline{Y}_{i.} = \frac{1}{b} \sum_{j=1}^{b} Y_{ij} \quad \text{Mittel der i-ten Stufe von A}$$

$$\overline{Y}_{.j} = \frac{1}{a} \sum_{i=1}^{a} Y_{ij} \quad \text{Mittel der j-ten Stufe von B}$$

$$SSA := b \sum_{i=1}^{a} (\overline{Y}_{i.} - \overline{Y}_{..})^2 \qquad SSB := a \sum_{j=1}^{b} (\overline{Y}_{.j} - \overline{Y}_{..})^2$$

$$SSE := \sum_{i=1}^{a} \sum_{j=1}^{b} (Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2$$

$$SST := \sum_{i=1}^{a} \sum_{j=1}^{b} (Y_{ij} - \overline{Y}_{..})^2$$

2-faktorielle Varianzanalyse

Quadratsummenzerlegung

Dependent Variable: Y

	Df	Sum Sq.	Mean Sq.	F-value	$Pr(>F)$	
A	a-1	SSA	MSA	$\frac{MSA}{MSE}$	H_{1a}	
B	b-1	SSB	MSB	$\frac{MSB}{MSE}$	H_{1b}	
Model	a+b-2	SSM	MSM	$\frac{MSM}{MSE}$	H_1	nicht enthalten
Error	(a-1)(b-1)	SSE	MSE			
Total	a b - 1	SST				

$$SSM = SSA + SSB \qquad SST = SSA + SSB + SSE$$

$$MSA = \frac{SSA}{(a-1)} \qquad MSB = \frac{SSB}{(b-1)}$$

$$MSM = \frac{SSA + SSB}{a+b-2} \qquad MSE = \frac{SSE}{(a-1)(b-1)}$$

2-faktorielle Varianzanalyse

Tests (1)

H_{0a} gegen H_{1a}:

$$F_1 = \frac{MSA}{MSE} = \frac{\text{mittl. Var. zwischen Stufen von A}}{\text{mittl. Var. innerhalb d. Gruppen}}$$

$$F_1 \quad \sim \quad F_{a-1,(a-1)(b-1)}$$

H_{0b} gegen H_{1b}:

$$F_2 = \frac{MSB}{MSE} = \frac{\text{mittl. Var. zwischen Stufen von B}}{\text{mittl. Var. innerhalb d. Gruppen}}$$

$$F_2 \quad \sim \quad F_{b-1,(a-1)(b-1)}$$

große Werte von F führen zur Ablehnung!

$$F_1 \quad > \quad F_{1-\alpha,a-1,(a-1)(b-1)} \rightarrow \text{Ablehnung von } H_{0a}$$
$$F_2 \quad > \quad F_{1-\alpha,b-1,(a-1)(b-1)} \rightarrow \text{Ablehnung von } H_{0b}$$

2-faktorielle Varianzanalyse

Tests (2)

H_0: $\alpha_1 = \cdots \alpha_a = 0$ und $\beta_1 = \cdots \beta_a = 0$ gegen
H_1: $\exists (i_1, i_2)$: $\alpha_{i_1} \neq \alpha_{i_2} \vee \exists (j_1, j_2)$: $\beta_{j_1} \neq \beta_{j_2}$.

$$F = \frac{MSModell}{MSE} = \frac{SSA + SSB}{SSE} \cdot \frac{(a-1)(b-1)}{a+b-2}$$

$$\begin{aligned} MSModell &= \frac{SSModell}{a+b-2} \\ SSModell &= SSA + SSB. \end{aligned}$$

H_0 ablehnen, falls

$$F > F_{1-\alpha, a+b-2, (a-1)(b-1)}.$$

Zweifaktorielle Varianzanalyse

Programm

```
#falls A, B noch keine Faktoren:
X$A = factor(X$A)
X$B = factor(X$B)
#eigentliche Analyse
anova(lm(Y~A+B,X))
#F-Statistik und p-Wert des ges. Modells
summary(lm(Y~A+B,X))
```

Achtung: `anova` berechnet nur sog. TypI-Summen(s. nächster Abschnitt)! Hier (balancierte Stichprobe) gilt aber $SSM = SSA + SSB$, dadurch macht das keinen Unterschied.

Inhalt

8.4 Weitere Varianzanalyse-Modelle

8.4.1 Mehrere Beobachtungen pro Kombination der Faktoren A und B

a) balancierter Fall $\rightarrow$ eindeutig
b) unbalancierter Fall $\rightarrow$
Es gibt verschiedene Möglichkeiten die Fehlerquadratsummen zu zerlegen.
`anova` beherrscht nur Typ-I-Summen.
besser: `Anova(lm(Y A+B,X),type=3)` Paket car
Typ-III-Summen hängen nicht von Reihenfolge ab (`A+B` vs. `B+A`).

3 Forscher graben eine Reihe von Schädeln in 3 verschiedenen Schichten aus.

Gemessen wird die Nasenlänge.
? Forschereffekt, Schichteneffekt

Weitere Varianzanalyse-Modelle

Mehrere Beobachtungen pro Kombination der Faktoren A und B (2)

Klinische Untersuchung in mehreren Zentren

Ein Medikament zur Gewichtsreduktion soll getestet werden.

1: Medikament
0: Placebo
1-6: Zentren

Modell:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

Es interessiert nur das Medikament, nicht das Zentrum:

$$H_0: \quad \alpha_0 = \alpha_1 \qquad H_1: \alpha_0 < \alpha_1$$

Weitere Varianzanalyse-Modelle

```
anova(lm(Y~Medik+Zentrum))
#oder
Anova(lm(Y~Medik+Zentrum,type=3))
```

GLM_Drugeffect

Zum Output: wie bisher.

Balanzierter Fall: Variante I und III identisch.

Unbalancierter Fall: Typ III-Summen zu bevorzugen, da der entsprechende Test unabhängig von den Stichprobenumfängen ist.

Weitere Varianzanalyse-Modelle

8.4.2 Wechselwirkungen ins Modell mit aufnehmen

$$Y_{ijk} = \alpha + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

(+Reparametrisierungsbedingungen)
geht nur, wenn für jede Faktorstufenkombination mehrere Beobachtungen vorliegen.

```
anova(lm(Y~A+B+A:B))
#oder kurz (siehe Folie zu lm)
anova(lm(Y~A*B))
```

GLM_Insekten

Weitere Varianzanalyse-Modelle

Modell mit Wechselwirkungen

Folgene Hypothesen sind zu testen:

$$H_{0a}: \quad \alpha_1 = \cdots = \alpha_a = 0 \quad \text{gegen}$$
$$H_{1a}: \quad \exists (i_1, i_2) : \alpha_{i_1} \neq \alpha_{i_2}$$

$$H_{0b}: \quad \beta_1 = \cdots = \beta_b = 0 \quad \text{gegen}$$
$$H_{1a}: \quad \exists (j_1, j_2) : \beta_{j_1} \neq \beta_{j_2}$$

$$H_{0c}: \quad \gamma_{11} = \cdots = \gamma_{a*b} = 0 \quad \text{gegen}$$
$$H_{1c}: \quad \exists (j_1, j_2) : \gamma_{j_1, j_2} \neq 0$$

Weitere Varianzanalyse-Modelle

8.4.3 Faktoren (Effekte, Faktorstufen) sind zufällig

hier ist Schätzung der Varianzkomponenten interessant und evtl. ein Hypothesentest

Preisrichter seien zufällig ausgewählt.

Die Frage ist, ob die Variabilität in den Scores an den Preisrichtern liegt?

$$Y_{ij} = \mu + \underbrace{A_i}_{\text{zufällig}} + b_j + \epsilon_{ij}$$

$$A_i \sim (0, \sigma_P^2)$$

$$\epsilon_{ij} \sim (0, \sigma^2)$$

Varianzkomponentenschätzung

```
varcomp(lme(Score~1,
  random=1|Preisrichter/Wettkaempfer))
#varcomp aus Paket ape
#lme aus Pkat nlme
```

GLM_syncro_zufaelligeEffekte

Annahme: A_i, B_j und ϵ_{ij} unabhängig.

$$var(Y_{ij}) = var(A_i) + var(B_j) + var(\epsilon_{ij})$$

Output: Schätzungen für die Varianzkomponenten.

Weitere Varianzanalyse-Modelle

8.3.4 Mehr als 2 Faktoren

- höherfaktorielle VA

Frequenzspektren

Gemessen wird die Amplitude bei 35 verschiedenen Frequenzen, 4 Füllungen, 3 Richtungen,
jede Messung wird 5 mal wiederholt.
? Füllungs-, Richtungseffekt, Wiederholungseffekt? Frequenzeffekt?
$\rightarrow$ 4 Faktoren.

```
Anova(lm(Y~A+B+C+D),data,type=3)
```

/Beratung/Vogt/Glaeser1

Weitere Varianzanalyse-Modelle

8.3.5 Hierarchische Modelle

Die Faktoren liegen in hierarch. Ordnung vor.

						A						
	A1			A2			A3			A4		
B11	B12	B13	B21	B22	B23	B31	B32	B33	B41	B42	B43	

(mit zufäll. Effekten)

Kalzium-Gehalt verschiedener Pflanzen und von verschiedenen Blättern
4 Pflanzen werden zufällig ausgewählt
3 Blätter davon
2 Stichproben zu 100mg von jedem Blatt
Frage: Gibt es zwischen Pflanzen oder zwischen Blättern unterschiedliche CA-Konzentrationen?

Weitere Varianzanalyse-Modelle

Hierarchische Modelle (2)

Modell

$$Y_{ijk} = \mu + A_i + B_{ij} + \epsilon_{ijk}$$

$$A_i \sim \mathcal{N}(0, \sigma_a^2) \quad B_{ij} \sim \mathcal{N}(0, \sigma_b^2) \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

hier: n=2 a=4 b=3

$$\begin{aligned} \text{var} Y_{ijk} &= \text{var} A_i + \text{var} B_{ij} + \text{var} \epsilon_{ijk} \\ &= \sigma_a^2 + \sigma_b^2 + \sigma^2 \end{aligned}$$

$$H_{0a} : \sigma_a^2 = 0 \qquad H_{0b} : \sigma_b^2 = 0$$

GLM_hierarch

Weitere Varianzanalyse-Modelle

Hierarchische Modelle (3)

```
#lmer in Paket lme4
lmer(Y~(1|A)+(1|B/A))
```

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

9. Anpassungstests

Anpassungstests

9. 1 Einführung

Problem

Klassische Test- und Schätzverfahren sind oft konzipiert unter der Normalverteilungsannahme.

Frage

Gilt sie überhaupt?

Gilt die Normalverteilung? (1)

Hampel, 1980, Biometr. Journal

Eine Zeitlang glaubte (fast) jeder an das ’normale Fehlergesetz’,

die Mathematiker, weil sie es für ein empirisches Faktum hielten,

und die Anwender, weil sie es für ein mathematisches Gesetz hielten.

Gilt die Normalverteilung? (2)

Geary 1947, Biometrika

Normality is a myth;
there never was,
and never will be,
a normal distribution.

Anpassungstests

$(X_1, \ldots, X_n)$ iid., $X_i \sim F$, F unbekannt.

Anpassungstest auf eine spezifizierte Verteilung:

$$H_0: \quad F = F_0 \qquad \text{gegen} \qquad H_1: \quad F \neq F_0.$$

I.A. hängt F von unbekannten Parametern ab.

Anpassungstest auf eine Normalverteilung:

$$H_0: F(x) = \Phi\Big(\frac{x-\mu}{\sigma}\Big) \quad (\mu, \sigma \text{ unbekannt})$$

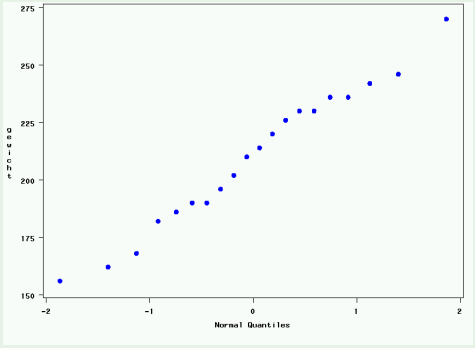
$$H_1: F(x) \neq \Phi\Big(\frac{x-\mu}{\sigma}\Big) \quad \forall \mu, \sigma, \sigma > 0$$

(Φ: Verteilungsfunktion der Standardnormal.).

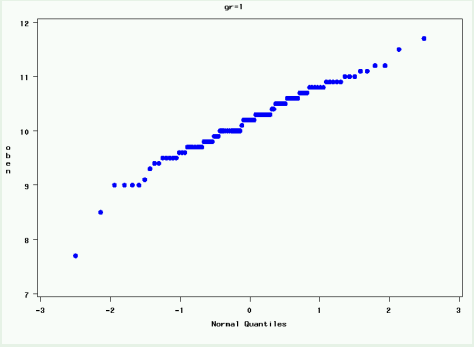
Anpassungstests

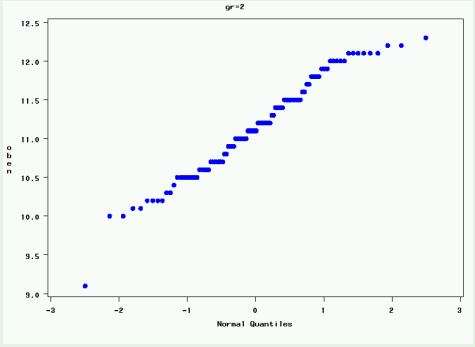
Gewicht von Hühnern

156	162	168	182	186
190	190	196	202	210
214	220	226	230	230
236	236	242	246	270



Abmessungen von Banknoten, oben (echt, falsch)





Inhalt

9.2 Auf der empirischen Verteilungsfunktion beruhende Tests

Empirische Verteilungsfunktion

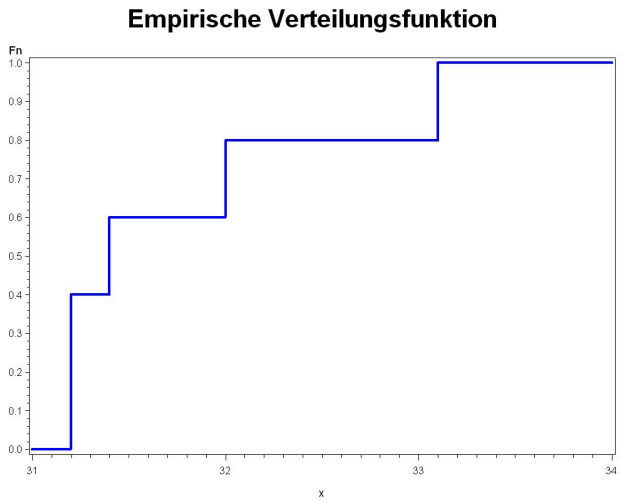
Seien $X_1, ..., X_n$ unabh. Beobachtungen,
$X_{(1)} \leq ... \leq X_{(n)}$ die geordneten Beob.
Die Funktion

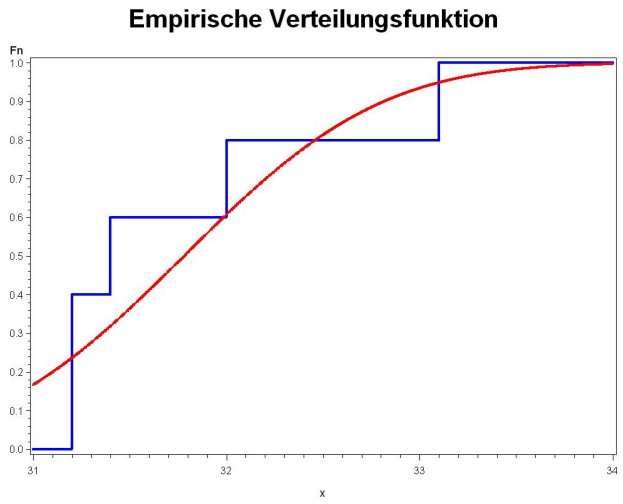
$$F_n(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x < X_{(i+1)} \\ 1 & X_{(n)} \leq x \end{cases} \qquad i = 1...n$$

heißt empirische Verteilungsfunktion.

Satz v. Glivento-Cantelli: $F_n(x) \to F(x)$.
(Hauptsatz der math. Statistik genannt)

Die empirische Verteilungsfunktion





Anpassungstests

Auf der empirischen Verteilungsfunktion beruhende Tests

Kolmogorov-Smirnov-Test

$$D = \sqrt{n} \sup_x |F_n(x) - F_0(x)|$$

Cramér-von Mises-Test

$$\text{W-sq} = n \int_{-\infty}^{\infty} \big(F_n(x) - F_0(x)\big)^2 dF_0(x)$$

Anderson-Darling-Test

$$\text{A-sq} = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} dF_0(x)$$

Anpassungstests auf Normalverteilung

Auf der empirischen Verteilungsfunktion beruhende Tests

hier:

$$\begin{aligned} F_0(x) &= \Phi\Big(\frac{x-\hat{\mu}}{\hat{\sigma}}\Big), \\ \hat{\mu} = \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\ \hat{\sigma}^2 = s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

$D \sim D_n$ (Kolmogorov-Verteilung) approx.

$$\lim_{n\to\infty} P_0\Big(D < \frac{x}{\sqrt{n}}\Big) = 1 - 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}$$

(Kolmogorov, 1933).

Anpassungstests auf Normalverteilung

Auf der empirischen Verteilungsfunktion beruhende Tests (2)

Modifikationen für endliche Stichproben (zur Info.)

$$\mathbf{D}: \quad D \cdot (\sqrt{n} - 0.01 + 0.85/\sqrt{n})/\sqrt{n}$$

$$\mathbf{A-sq}: \quad A\text{-sq} \cdot (1.0 + 0.75/n + 2.25/n^2)$$

$$\mathbf{W}\text{-sq}: \quad W\text{-sq} \cdot (1.0 + 0.5/n)$$

Große Werte von D, A-sq und W-sq führen jeweils zur Ablehnung von H_0.
p-Werte werden vom Programm berechnet.

Test_GoF_Banknote.R
Test_GoFDarwin.R

Anpassungstests auf Normalverteilung

Auf der empirischen Verteilungsfunktion beruhende Tests in R

Kolmogorov-Smirnov-Test: `ks.test`

`ks.test(x, y,alternative, exact = NULL)`
`x` ist eine Stichprobe, `y` Stichprobe oder Name einer Verteilung oder Verteilungsfunktion (`pnorm` für Normalverteilung).

Cramér-von Mises-Test: `cvm.test` (Paket nortest)

`cvm.test(x)` test, ob `x` normalverteilt ist.

Anderson-Darling-Test: `ad.test` (Paket nortest)

`ad.test(x)` test, ob `x` normalverteilt ist.

Inhalt

Anpassungstests

9.3 Shapiro-Wilk-Test (1)

Vorbemerkungen:

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \qquad Y_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

$i = 1, \ldots, n$.
Geordnete Beobachtungen:

$$X_{(1)} \leq \ldots \leq X_{(n)} \qquad Y_{(1)} \leq \ldots \leq Y_{(n)}.$$

Die Erwartungswerte

$$\begin{aligned} m_i &:= E(Y_{(i)}) \\ &= \frac{n!}{(i-1)!(n-i)!} \cdot \int_{-\infty}^{\infty} t\, \Phi^{i-1}(t)(1-\Phi(t))^{n-i}\phi(t)dt \end{aligned}$$

sind bekannt (und vertafelt).

Shapiro-Wilk-Test (2)

Approximation (Blom, 1958)

$$m_i \approx \tilde{m}_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

$$\begin{aligned} \mathbf{E}X_{(i)} &= \mu + \sigma m_i \\ X_{(i)} &= \mu + \sigma m_i + \epsilon_i \end{aligned}$$

einfaches lineares Regressionsmodell mit Parametern μ, σ.
$\mathbf{E}\epsilon_i = 0$, aber die ϵ_i sind nicht unabhängig.

$$\mathbf{V} := \mathrm{cov}(Y_{(i)}, Y_{(j)}), \qquad \mathbf{m}' := (m_1, \ldots, m_n)$$

$$\mathbf{X}' := (X_{(1)}, \ldots, X_{(n)}).$$

Shapiro-Wilk-Test (3)

Verallgemeinerter Kleinster Quadrat-Schätzer von σ:

$$\hat{\sigma} = \frac{\mathbf{m}'\mathbf{V}^{-1}\mathbf{X}}{\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}}$$

wird verglichen mit der gewöhnlichen Standardabweichung s

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Bem.: Der gewöhnliche Kleinster Quadrat-Schätzer von σ ist:

$$\hat{\sigma}_{KQS} = \frac{\mathbf{m}'\mathbf{X}}{\mathbf{m}'\mathbf{m}}.$$

Shapiro-Wilk Test (4)

Shapiro-Wilk-Statistik

$$W = \frac{\hat{\sigma}^2}{s^2(n-1)} \cdot \frac{(\mathbf{m}'\mathbf{V}^{-1}\mathbf{m})^2}{\mathbf{m}'\mathbf{V}^{-2}\mathbf{m}} = \frac{(\mathbf{h}'\mathbf{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \mathbf{h}'\mathbf{h}}$$

wobei $\mathbf{h}' = \mathbf{m}'\mathbf{V}^{-1}$ (bekannt, vertafelt).

Wegen $\sum h_i = 0$ folgt:
W ist Quadrat des (empir.) Korrelationskoeffizienten von $\mathbf{h}$ und $\mathbf{X}$:

$$W = \frac{\left(\sum_{i=1}^{n}(X_i - \overline{X})(h_i - \overline{h})\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(h_i - \overline{h})^2},$$

Shapiro-Wilk Test (5)

$$W = \frac{\left(\sum_{i=1}^{n}(X_i - \overline{X})(h_i - \overline{h})\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(h_i - \overline{h})^2},$$

- Offenbar: $0 \leq W \leq 1$.
- $W \approx 1$ indiziert, dass $\mathbf{h}' = \mathbf{m}'\mathbf{V}^{-1} (\approx 2\mathbf{m}')$: ein Vielfaches von $\mathbf{X}$ ist.
  D.h. die Punkte $(m_i, X_{(i)})$ liegen etwa auf einer Geraden, was Normalverteilung indiziert.
- H_0 wird ablehnt, falls $W < W_\alpha(n)$. R verwendet dabei noch eine Transformation von W

Shapiro-Wilk Test (6)

Scores der 1. Wettkämpferinnen (5 Preisrichter)

$X = (31.2, 31.2, 31.4, 32.0, 33.1)$
Mit der Funktion `sd` erhalten wir $s = 0.80747$,
weiter ist $h \approx (-2.88, -0.99, 0, 0.99, 2.88) \approx 2\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right)$
(ausser h_1 und h_5, siehe R-Code)
Für die Shapiro-Wilk Statistik bekommen wir

$$W = \mathrm{cor}(X, h)^2 \approx 0.81121$$

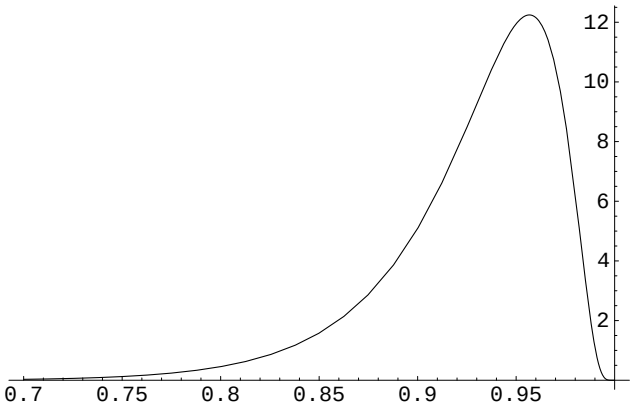
ShapiroWilk_Synchro.R

Shapiro-Wilk Test (7)

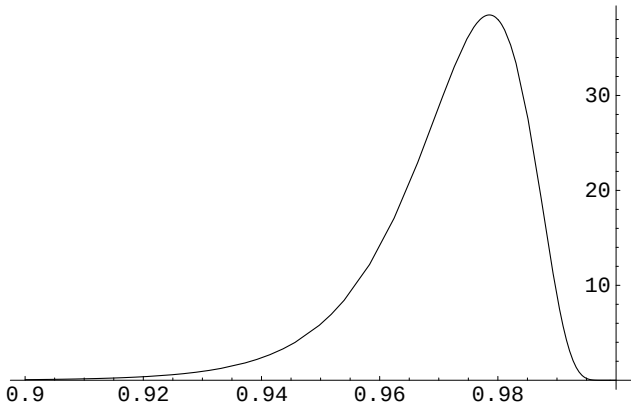
Approximative Dichtefunktion von W (unter H_0)

$$W = \frac{\left(\sum_{i=1}^{n}(X_i - \overline{X})(h_i - \overline{h})\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(h_i - \overline{h})^2},$$

$n = 10$



$n = 50$



Anpassungstests

- R verwendet eine Approximation der Verteilung von W ab Stichprobengröße ≥ 4 .
- Der Shapiro-Wilk-Test erweist sich für kleinere, mittlere und größere Stichprobenumfänge als geeignetster Test (höchste Güte).
- Früher wurde meist der sogen. χ^2-Anpassungstest verwendet. Dieser hat jedoch geringe Güte.
- W ist etwas besser als A-sq, besser als W-sq, und viel besser als D und χ^2.
- D ist nur für sehr große Stichprobenumfänge zu empfehlen ($n \geq 2000$).

Anpassungstests

- Man sollte beim Test auf Normalverteilung das Signifikanzniveau auf $\alpha = 0.1$ hochsetzen, insbesondere wenn wenig robuste Tests (die NV verlangen) angewendet werden sollen.
- Robuste Tests haben meist geringen Effizienzverlust bei Vorliegen von Normalverteilung.

Anpassungstests

Durchführung des Shapiro-Wilk-Tests in R

`shapiro.test`

`shapiro.test(x)` teste `x` auf Normalverteilung ($\hat{\mu}, \hat{\sigma}$ werden genutzt).

Inhalt

Anpassungstests

9.4 Anpassungstests auf weitere Verteilungen

χ^2-Anpassungstest (Pearson, 1900)

Prinzip: Daten werden in p Klassen eingeteilt.
Klassenhäufigkeiten: N_i
theoretische Klassenhäufigkeiten: np_i

$$X^2 = \sum_{i=1}^{p} \frac{(N_i - np_i)^2}{np_i}$$

$X^2 \sim \chi^2_{p-1}$ asymptotisch (bei bekannten μ, σ^2) (Fisher, 1922)

$X^2 \sim \chi^2_{p-3}$ approx. (bei 2 zu schätzenden Parametern, ML-Schätzung mit gruppierten Daten oder Minimum-χ^2-Schätzung).

Anpassungstests

χ^2-Anpassungstest

Nachteile des χ^2-Anpassungstests

- Wert von X^2 abhängig von Klasseneinteilung.
- χ^2- Anpassungstest auf Normalverteilung hat geringe Güte.

Diskrete Verteilungen

Hier kann der χ^2-Anpassungstest genommen werden (natürliche Klasseneinteilung)
Funktion `chisq.test(x,p)`

Anpassungstests

χ^2-Anpassungstest

Diskrete Gleichverteilung

```
chisq.test(x)
```

Sonstige diskrete Verteilungen

wie oben, zusätzlich sind die Einzelwktn. explizit zu formulieren,

```
chisq.test(x,p=c(p1,p2,...))
```

Achtung: `p=` nutzen und nicht einfach 2.Argument setzen, dieses ist `y` und führt zu einem χ^2-Unabhängigkeitstest (später).

Test_GoF_Poisson PoissonHorsekicks

Anpassungstests

EDF-Tests

Stetige Verteilungen

`ks.test(x,y)` mit `y=` `pweibull`, `pexp`, `pgamma`, `pchisq`, `pt`, `plnorm` ...

Descr_Plot_Kuehl.R
Test_GoF_Darwin_1.R

Inhalt (1)

Inhalt (2)

Inhalt (3)

Inhalt

10. Nichtparametrische Tests

Übersicht

Es werden die wichtigsten Rang-Analoga zu den Tests in 7.2.-7.4., 8.1,8.3 behandelt.

Nichtparametrische Tests

10.1 Einführung

Was tun wenn Normalverteilung nicht vorliegt?

Nichtparametrische Tests

- sie verwenden keine Parameterschätzung (wie $\overline{X}, s$)
- sie halten das Signifikanzniveau (α) für jede stetige Verteilung (approx.) ein. α hängt also nicht von der zugrundeliegenden Verteilungsfunktion ab.
- sie sind relativ effizient. Der Effizienzverlust bei Normalvert. ist in vielen Fällen gering!

Annahme: Verteilungsfunktion ist stetig (wenn nicht anders vermerkt)

Inhalt

Nichtparametrische Tests

10.2 Einstichprobenproblem

	Nulhypothese	Alternative
a)	$H_0: \quad \mu \leq \mu_0$	$H_A: \quad \mu > \mu_0$
b)	$H_0: \quad \mu \geq \mu_0$	$H_A: \quad \mu < \mu_0$
c)	$H_0: \quad \mu = \mu_0$	$H_A: \quad \mu \neq \mu_0$

Vorzeichentest

Wie bisher werden die Differenzen $X_i - \mu_0$ gebildet.

$$V_i := \begin{cases} 1 & \text{falls } X_i - \mu_0 > 0 \\ 0 & \text{falls } X_i - \mu_0 < 0 \end{cases} \qquad V^+ = \sum_{i=1}^{n} V_i$$

$V^+ =$ # Differenzen mit positivem Vorzeichen

Nichtparametrische Tests

Vorzeichentest (2)

Der Fall $X_i - \mu_0 = 0$ tritt wegen der Stetigkeit

der Vf. nur mit Wkt. 0 auf. Sollte der Fall trotzdem eintreten (Meßungenauigkeit) so wird die entspr. Beobachtung weggelassen und der Stichprobenumfang entsprechend verringert.
(Nachteil: Es werden gerade Beobachtungen weggelassen, die für die Nullhypothese sprechen!)

Es gilt: $V^+ \sim Bi(n, \frac{1}{2})$

(V^+ = # “Erfolge” bei n Versuchen mit Wkt. je $\frac{1}{2}$).

$\Rightarrow$ krit. Werte könnten leicht selbst bestimmt werden:

$\texttt{qbinom}(1-\alpha, n, \frac{1}{2})$

Nichtparametrische Tests

Vorzeichentest (3)

Teststatistik

$$\underline{M} = V^+ - \frac{n}{2} \qquad (= \frac{V^+ - V^-}{2}) \quad (zentrierteStatistik)$$

n^+: Realisierung von V^+
n^-: Realisierung von V^-

Zweiseitiger p-Wert:
$P(|\underline{M}| \geq |n^+ - \frac{n}{2}|) = P(|\underline{M}| \geq \max(n^+, n^-) - \frac{n}{2})$=(*)

$$\text{denn } |n^+ - \frac{n}{2}| = \begin{cases} n^+ - \frac{n}{2} & n^+ > \frac{n}{2} \\ \frac{n}{2} - n^+ & n^+ < \frac{n}{2} \\ = n^- - \frac{n}{2} & \end{cases}$$

Nichtparametrische Tests

Vorzeichentest (4)

Der p-Wert ist gleich

$$
\begin{aligned}
(*) &= P\Big(V^+ - \frac{n}{2} \geq \max(n^+, n^-) - \frac{n}{2}\Big) + P\Big(\frac{n}{2} - V^+ \geq \max(n^+, n^-) - \frac{n}{2}\Big) \\
&= P\Big(V^+ \geq \max(n^+, n^-)\Big) + P\Big(n - V^+ \geq \max(n^+, n^-)\Big) \\
&= 2 \sum_{j=\max(n^+,n^-)}^{n} \binom{n}{j} (\frac{1}{2})^j (\frac{1}{2})^{n-j} = (\frac{1}{2})^{n-1} \sum_{j=\max(n^+,n^-)}^{n} \binom{n}{j} \\
&= (\frac{1}{2})^{n-1} \sum_{j=0}^{\min(n^+,n^-)} \binom{n}{j}.
\end{aligned}
$$

Nichtparametrische Tests

Vorzeichentest (5)

Die Verteilung von V^+ ist diskret, d.h. es gibt nicht zu jedem α einen entsprechenden kritischen Wert.

Aber: p-Werte gibt es immer, d.h.:

$$
\begin{array}{lcl}
p < \alpha & \Rightarrow & H_0 \text{ (c) ablehnen} \\
M > 0 \wedge \frac{p}{2} < \alpha & \Rightarrow & H_0 \text{ (b) ablehnen} \\
M < 0 \wedge \frac{p}{2} < \alpha & \Rightarrow & H_0 \text{ (a) ablehnen}
\end{array}
$$

Der Vorzeichentest ist meist nicht sehr effizient (Ausnahme: Verteilung=Doppelexponential)
besser ist der Wilcoxon-Vorzeichen-Rangtest

Nichtparametrische Tests

Vorzeichentest in R

```
sign.test = function(data,mu=0,...){
  sig = sign(data-mu);
  vplus = length(sig[sig == 1]);
  ties = length(sig[sig == 0]);
  n = length(data) - ties;
  binom.test(vplus,n,p=0.5,...);
}
```

Bem.: `...` kopiert die Argumente von `sign.test` zu `binom.test`. Dieses testet ein Ergebnis eines Bernoulliexperiments unter der H_0 das die Wkt. pro Experiment `p` ist.

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest

Wilcoxon-Vorzeichen-Rangtest

Bilden zu den "Beobachtungen" $D_i = |X_i - \mu_0|$ die Rangzahlen, d.h. den Rang (den Platz) in der geordneten Stichprobe

$$\underbrace{D_{(1)}}_{\text{Rang 1}} \leq \ldots \qquad \ldots \leq \underbrace{D_{(n)}}_{\text{Rang n}}$$

Sei R_i^+ der Rang von D_i.

$$W_n^+ = \sum_{i=1}^{n} R_i^+ \cdot V_i$$

Summe der Ränge von D_i für die $X_i - \mu_0 > 0$.

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest (2)

Erwartungswert und Varianz von W_n^+

$$\mathbf{E}_0 W_n^+ = \frac{1}{2}\sum_{i=1}^{n} R_i^+ = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n\cdot(n+1)}{4} \qquad \mathbf{E}V_i = \frac{1}{2}$$

$$\text{var } W_n^+ = \mathbf{E}(W_n^+ - \mathbf{E}W_n^+)^2 = \frac{n\cdot(n+1)(2n+1)}{24} \quad (\text{ÜA})$$

Die Berechnung der exakten Verteilung von W_n^+ kann durch Auszählen aller Permutationen erfolgen
($\rightarrow$ schon für kleinere n größere Rechenzeit!)
Deshalb verwendet man (für mittlere und große n) die asymptotische Verteilung.

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest (3)

Asymtotische Verteilung

$$W_n^+ \sim \mathcal{N}(EW_n^+, \text{var}W_n^+) \qquad \text{asymptotisch}$$

Große Werte von

$$\frac{|W_n^+ - EW_n^+|}{\sqrt{\text{var } W_n^+}}$$

führen zur Ablehnung von H_0.

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest (4)

R-Implementation (Wilcoxon-Vorzeichen-Test)

$$S = W_n^+ - EW_n^+ = \sum_{i=1}^{n} R_i^+ V_i - \frac{n(n+1)}{4}$$

R_i^+ Rang von $|X_i - \mu_0|$,
Summe nur über positive $X_i - \mu_0$

$n \leq 20$: p-Werte aus der exakten Verteilung von S.

$n > 20$: Es wird auch eine t-Approximation angeboten:

$$t = \frac{S \cdot \sqrt{n-1}}{\sqrt{n \text{ Var } (S) - S^2}} \sim t_{n-1}$$

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest (5)

Bindungen (= Meßwertwiederholungen): Ränge werden gemittelt.

Sei t_i: # Bindungen in der i-ten Gruppe.
Korrektur in Var(S):

$$\text{var}(S) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{2}\sum t_i(t_i+1)(t_i-1)$$

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest (6)

IQ-Werte von Studenten (Wiwi)

$H_0 : \mu = \mu_0 = 110$ $\quad$ $H_1 : \mu > \mu_0$

x_i=IQ	d_i	$\lvert d_i \rvert$	r_i^+	V_i
99	-11	11	5	0
131	21	21	8	1
118	8	8	3	1
112	2	2	1	1
128	18	18	7	1
136	26	26	10	1
120	10	10	4	1
107	-3	3	2	0
134	24	24	9	1
122	12	12	6	1

$d_i = x_i - 110$

Vorzeichentest:

$M = 8 - \frac{10}{2}$

p-Wert(exakt) = 0.1094

Wilcoxon-signed

$W^+ - \mathbf{E}(W^+) = 48 - \frac{10 \cdot 11}{4} = 20.5.$

p-Wert=0.0371.

Test_IQ_Daten

Nichtparametrische Tests

Wilcoxon-Vorzeichen-Rangtest (7)

- Im Gegensatz zum Vorzeichentest ist der Vorzeichen-Wilcoxon-Test (= signed rank test) sehr effizient, bei NV nur wenig schlechter, bei den meisten Vf. besser als der t-Test.
  $\Rightarrow$ Wenn NV nicht gesichert ist Vorzeichen-Wilcoxon-Test nehmen!
- Der Vorzeichentest und der Wilcoxon-Test sind sogen. Rangtests, da sie nur auf den Rangzahlen der Beobachtungen beruhen.
  Es gibt weitere Rangtests.
- Durchführung des Wilcoxon-Vorzeichen-Rangtest:
  `wilcox.test(x,alternative,mu,exact,...)`

Nichtparametrische Tests

Zwei verbundene Stichproben

Bilden $Z := X - Y$ und testen wie beim Einstichprobenproblem, z.B.

$$H_0 : \mu_Z = 0 \qquad H_1 : \mu_Z \neq 0$$

Banknoten: oben-unten, links-rechts
Darwin: kreuz-selbstbefruchtete Pflanzen

```
sign.test(x-y)
wilcox.test(x,y,paired=TRUE)
```

Npar_1_Banknote Npar_1_Darwin

Nichtparametrische Tests

Weitere Problemstellungen im Einstichprobenfall (1)

Binärvariablen

Sei X eine 0-1 Variable, d.h.

$$P(X = 0) = p, \quad P(X = 1) = 1 - p$$

$H_0 : p = p_0$ T: Anzahl der Beobachtungen in Klasse 0.

H_{1a} $p < p_0$: p-Wert $= P(T \leq t)$ = `pbinom(`t, n, p_o`)`

H_{1b} $p > p_0$: p-Wert $= P(T \geq t)$

H_{1c} $p \neq p_0$: p-Wert $= P(T \leq t$ oder $T \geq n - t + 1)$

Binomialtest

`binom.test(x=`t`,n,p)`

Nichtparametrische Tests

Binomialtest

```
binom.test(sum(var > 0),length(var),0.8)
```

Binomialtest_toxaemia.R

Warenliefung, ÜA

Der Hersteller behauptet, höchstens 5% sind schlecht.
Sie haben $n = 20$ Stücke geprüft, und $X = 3$ schlechte Stücke gefunden. Hat der Hersteller recht?
Betrachten Sie sowohl die exakte als auch die asymptotische Version.

Konfidenzintervalle:

a) Normalapproximation
b) exakt: Binomialverteilung (`pbinom`)

Nichtparametrische Tests

Weitere Problemstellungen im Einstichprobenfall (4)

Zum Vergleich, zur Erinnerung und Ergänzung

Diskrete Gleichverteilung

```
chisq.test(x)
```

Anpassungstest auf vorgegebene diskrete Verteilung

wie oben, zusätzlich sind die Einzelwktn. explizit zu formulieren,
`chisq.test(x,p=c(p1,p2,...))`
Achtung: `p=` nutzen und nicht einfach 2.Argument setzen, dieses ist `y` und führt zu einem χ^2-Unabhängigkeitstest (später).

Nichtparametrische Konfidenzintervalle

$(1-\alpha)$-Konfidenzintervall für p-Quantil, d.h. für x_p

Die Verteilung der j-ten Ordnungsstatistik $X_{(j)}$:

$$P(X_{(j)} < x) = \sum_{i=j}^{n} \binom{n}{i} F(x)^i (1-F(x))^{n-i}$$

‘Erfolg” gdw. $X_i < x$, “Erfolgswkt.” $F(x)$.
Insbesondere, für $x = x_p$ (das wahre p-Quantil)

$$\begin{aligned} P(X_{(j)} < x_p) &= \sum_{i=j}^{n} \binom{n}{i} F(x_p)^i (1-F(x_p))^{n-i} \\ &= \sum_{i=j}^{n} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Nichtparametrische Konfidenzintervalle

$$P(X_{(j)} < x_p) = \sum_{i=j+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

Untere und obere Konfidengrenzen $X_{(l)}$ und $X_{(u)}$ für x_p werden so bestimmt, dass l und u (möglichst) symmetrisch um $\lfloor np \rfloor + 1$ und so dass

$$P(X_{(l)} \leq x_p < X_{(u)}) = \sum_{i=l}^{u-1} \binom{n}{i} p^i (1-p)^{n-i} \geq 1 - \alpha$$

($X_{(\lfloor np \rfloor)}$ ist Schätzung für x_p.)

Nichtparametrische Konfidenzintervalle

$(1-\alpha)$ Konfidenzintervall für x_p

```
n = length(x);  npf = floor(n*p); alpha= α
ci.ind = c(l=npf,u=npf+1)
while(diff(pbinom(ci.ind,n,p)) < (1-alpha)){
  ci.ind["u"] = ci.ind["u"]+1
  if(diff(pbinom(ci.ind,n,p) >= (1-alpha)))
    break
  ci.ind["l"] = ci.ind["l"]-1
}
conf.int = sort(x)[ci.ind]
```

Inhalt

Nichtparametrische Tests

10.3 Zwei unverbundene Stichproben: Wilcoxon Test

Wir setzen keine Normalverteilung voraus, aber den gleichen Verteilungstyp, insbesondere gleiche Varianzen

$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$H_0 : \mu_1 \geq \mu_2 \quad H_1 : \mu_1 < \mu_2$

$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$

Wir fassen die Beobachtungen

$X_{11}, ..., X_{1n}, X_{21}, ..., X_{2m}$

zu einer Stichprobe zusammen und bilden die Rangzahlen R_{ij}, $i = 1, 2, j = 1 \ldots n, m$

$$\underbrace{z_{(1)}}_{\text{Rang 1}} \leq \ldots \leq \underbrace{z_{(n+m)}}_{\text{Rang n+m}}$$

Nichtparametrische Tests

Wilcoxon-Test

Summe der Ränge zur 1. bzw. 2. Stichprobe

$$S_1 = \sum_{j=1}^{n} R_{1j} \qquad S_2 = \sum_{j=1}^{m} R_{2j}$$

Die Erwartungswerte (unter H_0) sind

$$\mathbf{E}_0 S_1 = \frac{n(n+m+1)}{2} \quad \text{und} \quad \mathbf{E}_0 S_2 = \frac{m(n+m+1)}{2}$$

und die Varianzen

$$\text{var} S_1 = \text{var}\, S_2 = \frac{n \cdot m(n+m+1)}{12}.$$

Nichtparametrische Tests

Wilcoxon-Test (2)

Sei S die Statistik S_1 oder S_2, die zur kleineren Stichprobe gehört.

Die Teststatistik des Wilcoxon-Tests ist

$$Z = \frac{S - E(S)}{\sqrt{varS}}$$

$$Z \sim \mathcal{N}(0,1) \qquad \text{approximativ}$$

(0.5 = Stetigkeitskorrektur)
bei Bindungen: korrigierte (kleinere) Varianz

Npar1way_Carnitinfraktion.R
Npar1way_Banknote.R
Npar1way_Heroin.R
Npar1way_Tibetan.R

Nichtparametrische Tests

Wilcoxon-Test (3)

- R gibt die Teststatistik (Z) und den p-Wert je nach Wahl von `alternative` an.

a) $H_0 : \mu_1 = \mu_2$ $\quad H_1 : \mu_1 \neq \mu_2$

$\Rightarrow$ two-sided $Pr > |Z| = P(|Z| > Z)$

b) $H_0 : \mu_1 \leq \mu_2$ $\quad H_1 : \mu_1 > \mu_2$

$\Rightarrow$ one-sided $z > 0$

$\rightarrow P(Z > z) = Pr > Z$

c) $H_0 : \mu_1 \geq \mu_2$ $\quad H_1 : \mu_1 < \mu_2$

$\Rightarrow$ one-sided $z < 0$

$\rightarrow P(Z < z) = Pr < Z$

- R bietet die Normalapproximation oder exakte p-Werte an.

```
wilcox.test(x,y,alternative,exact)
```

(nötige Option `paired=FALSE` ist Standard)

Nichtparametrische Tests

Zwei unverbundene Stichproben: Kolmogorov-Smirnov Test

Wir machen gar keine Verteilungsannahmen.

$$
\begin{array}{llll}
H_0: & F_1 = F_2 & H_1: & F_1 \neq F_2 \\
H_0: & F_1 \leq F_2 & H_1: & F_1 > F_2 \\
H_0: & F_1 \geq F_2 & H_1: & F_1 < F_2
\end{array}
$$

Kolmogorov-Smirnov Test

$$
\begin{aligned}
D &= \max_i |F_1(x) - F_2(x)| && \text{(zweiseitig, EDF)} \\
D^+ &= \max_i (F_1(x) - F_2(x)) && \text{(einseitig, D)} \\
D^- &= \max_i (F_2(x) - F_1(x)) && \text{(einseitig, D)}
\end{aligned}
$$

```
ks.texst(x,y,alternative,exact)
```

Zweistichprobenproblem

Allgemeine Empfehlungen (1)

- Wenn Normalverteilung, gleiche Varianzen und keine Ausreißer: t-Test
- Wenn Normalverteilung, ungleiche oder unbekannte Varianzen und keine Ausreißer: Welch-Test (t-Test, unpooled, Satterthwaite)
- Wenn "sehr nahe" an Normalverteilung und keine Ausreißer: wie bei Normalverteilung
- keine Normalverteilung oder unbekannte Verteilung, gleiche Varianzen, und etwa gleicher Verteilungstyp (Ausreißer in begrenztem Maße erlaubt): Wilcoxon Test

  oder: Adaptiver Test (z.B. Paket: adaptTest)

Zweistichprobenproblem

Allgemeine Empfehlungen (2)

- keine Normalverteilung oder unbekannte Verteilung, ungleiche Varianzen, und etwa gleicher Verteilungstyp (Ausreißer in begrenztem Maße erlaubt)
  $n_1 \approx n_2$ oder $(n_1 > n_2, \sigma_1 < \sigma_2)$: Wilcoxon Test
- keine Normalverteilung, Verteilungstypen verschieden, ungleiche Varianzen (kleine Varianz zu kleinem Stichprobenumfang): K-S Test

  oder: Brunner-Munzel Test (Paket lawstat)

Inhalt

Nichtparametrische Tests

10.4 Mehrere unverbundene Stichproben

Modell:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim (0, \sigma^2), \quad j = 1, ..., n_i, i = 1, .., k$$

$$H_0 : \mu_1 = ... = \mu_k \qquad H_1 : \exists(\mu_{i_1}, \mu_{i_2}) \ \mu_{i_1} \neq \mu_{i_2}$$

Wir fassen alle Beobachtungen
$X_{11}, ..., X_{1n_1}, ..., X_{k1}, ..., X_{kn_k}$
zusammen und bilden die Rangzahlen R_{ij}, $i = 1...k, j = 1...n_i$.

Mit den Rangzahlen führen wir eine
einfaktorielle Varianzanalyse durch
= Kruskal-Wallis Test

Nichtparametrische Tests

Mehrere unverbundene Stichproben

Kruskal-Wallis Test

$$KW = \frac{\sum_{i=1}^{k}(T_i - E_0(T_i))^2 \cdot n_i}{S^2}, \quad \text{wobei}$$

$$T_i = \frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij} \quad \text{mittl. Rangsumme der i-ten Gruppe}$$

Kruskal-Wallis	Varianzanalyse
T_i	$\overline{Y}_{i.}$
$E_0 T_i = \frac{N+1}{2}$	$\overline{Y}_{..} = \overline{Y}$
Zähler	SSB
$S^2 = \frac{(N-1)N(N+1)}{12}$	SST
$= \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2)$	
$N = \sum_{i=1}^{k} n_i$	Gesamtstichprobenumfang

Nichtparametrische Tests

Kruskal-Wallis-Test (2)

$$
\begin{aligned}
S^2 &= \sum_i \sum_j (R_{ij} - \frac{N+1}{2})^2 = \quad = \sum_{k=1}^{N} (k - \frac{N+1}{2})^2 \\
&= \sum_k k^2 - (N+1) \sum_k k + \frac{(N+1)^2}{4} \cdot N \\
&= \frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{2} + \frac{(N+1)^2}{4} \cdot N \\
&= \frac{(N+1) \cdot N}{12} \left(4N + 2 - 6N - 6 + 3N + 3\right) \\
&= \frac{N(N+1)}{12} \cdot (N-1) = \frac{(N-1) \cdot N \cdot (N+1)}{12}.
\end{aligned}
$$

Nichtparametrische Tests

Kruskal-Wallis-Test (3)

Vorteil: S^2 ist nicht zufällig, hängt nur vom Stichprobenumfang ab.

$KW \sim \chi^2_{k-1}$ (asymptotisch)

H_0 ablehnen, falls p-value $< \alpha$

R: Funktion und Output

`kruskal.test(x,g,...)`

chi-squared: realisierte KW

df=$k-1$: Freiheitsgrade.

Npar1way_Maschinen.R

~_Varianzanalyse_Modelle_PI12erg.R

Nichtparametrische Tests

Kruskal-Wallis-Test (4)

- Bei Bindungen erfolgt eine Korrektur der Statistik
- KW-Test ist relativ effizient bei NV. Bei Nicht-NV meist besser als der Varianzanalyse-F-Test.
- KW-Test hält (wie alle nichtparametrischen Tests) asymptotisch das Signifikanzniveau ein.
- kleine Stichproben ($N \leq 20$): exakte p-Werte möglich mit der Funktion `wilcox_test` aus dem Paket coin (`_` statt `.`).

Inhalt

Nichtparametrische Tests

10.5 Mehrere verbundene Stichproben: Friedman Test

Modell, wie bei der 2-faktoriellen Varianzanalyse

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad \epsilon_{ij} \sim (0, \sigma^2), \qquad j = 1, \ldots, k, i = 1, \ldots, n$$

$$H_0 : \beta_1 = \ldots = \beta_k (= 0) \qquad H_1 : \exists (j_1, j_2) : \beta_{j1} \neq \beta_{j2}$$

Ränge werden zeilenweise gebildet, $Y_{1(1)} \leq \ldots \leq Y_{1(k)}$
R_{ij} der Rang von Y_{ij} in der i-ten Zeile.

Nichtparametrische Tests

Friedman Test (2)

	Behandlung				
Block	1	2	...	k	Zeilensummme
1	R_{11}	R_{12}	...	R_{1k}	$\frac{k(k+1)}{2}$
.					
.					
n	R_{n1}	R_{n2}	...	R_{nk}	$\frac{k(k+1)}{2}$
	$R_{.1}$	$R_{.2}$	...	$R_{.k}$	$\frac{nk(k+1)}{2}$
	$n\overline{R}_{.1}$	$n\overline{R}_{.2}$	...	$n\overline{R}_{.k}$	

$$F_k = \frac{n^2 \sum_{j=1}^{k} (\overline{R}_{.j} - E(\overline{R}_{.j}))^2}{n \cdot k(k+1)/12}$$

Nichtparametrische Tests

Friedman Test (3)

$$F_k = \frac{n^2 \sum_{j=1}^{k} (\overline{R}_{.j} - E(\overline{R}_{.j}))^2}{n \cdot k(k+1)/12}$$

$\overline{R}_{.j} = \frac{1}{n} \sum_{i=1}^{n} R_{ij}$ Spaltenmittel der j-ten Spalte (Vergleiche mit $\overline{Y}_{.j}$)

$E\overline{R}_{.j} = \frac{1}{n} \cdot \frac{n(k+1)}{2} = \frac{k+1}{2}$ (Vergleiche mit $\overline{Y}_{..}$)

Unter H_0 : $F_k \sim \chi^2_{k-1}$ (asympt.)

H_0 ablehnen, falls $F_k > \chi^2_{1-\alpha,k-1}$
oder falls p-value $< \alpha$.

Nichtparametrische Tests

Friedman-Test (4)

- Bei Bindungen Korrektur des Nenners.
- Für kleinere n ist Friedman-Test (asy.) meist etwas konservativ (d.h. der wahre Fehler 1. Art ist kleiner als z.B. 0.05).
- Für größere k (etwa $k \geq 5$) ist der Friedman-Test (bei NV) einigermaßen effizient.
- Für $k = 2$ ist der Friedman-Test zum Vorzeichentest äquivalent (also nicht besonders effizient).

Friedman-Test (5)

Durchführung des Friedman-Tests

Daten als Vektor oder Matrix

```
friedman.test(y, groups, blocks, ...)
```

Daten und Faktoren als `data.frame`

```
friedman.test(formula, data,...)
```

Test_Friedman_Hypnose.R Test_Friedman_Synchro.R
Hier ist nur die folgende Zeile interessant:
Row Mean Scores Differ

Inhalt (1)

Inhalt (2)

Inhalt (3)

11. Korrelation und Regression

Übersicht

Inhalt

11.1 Korrelation und Unabhängigkeit

Unabhängigkeit und Unkorreliertheit, Wdh.

Die Zufallsvariablen $X_1, \ldots, X_N$ heißen unabhängig, falls für alle $x_1, \ldots, x_N \in \mathbb{R}$

$$P(X_1 < x_1, \ldots, X_N < x_N) = P(X_1 < x_1) \cdots P(X_N < x_N)$$

Die Zufallsvariablen $X_1, \ldots, X_N$ heißen unkorreliert, falls

$$\mathbf{E}(X_1 \cdots X_N) = \mathbf{E}(X_1) \cdots \mathbf{E}(X_N).$$

Unabhängigkeit $\Rightarrow$ Unkorreliertheit

$\not\Leftarrow$

Unabhängigkeit $\Leftrightarrow$ Unkorreliertheit falls $X_i \sim \mathcal{N}$

Korrelation und Unabhängigkeit

Fall a) Stetige (metrische) Merkmale

Seien (X_i, Y_i), $i = 1, \ldots, N$ unabhängige bivariate Zufallsvariablen. Wir testen

H_0 : X und Y sind unabhängig (unkorreliert) gegen

H_1 : X und Y sind linear abhängig (korreliert)

Pearson-Korrelation

$$r_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

$$T = \sqrt{N-2} \cdot \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sim t_{N-2}$$

wird in R zur Berechnung der p-Werte verwendet.

Korrelation und Unabhängigkeit

Fall a) Stetige (metrische) Merkmale (3)

H_0 : X und Y sind unabhängig (unkorreliert) gegen

H_1 : X und Y sind monoton abhängig

Spearman-Rangkorrelationskoeffizient

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}}$$

Weitere Korrelationskoeffizienten: Kendall.
Wenn keine Normalverteilung vorliegt, so Spearman oder Kendall nehmen!

Korrelation und Unabhängigkeit

a) Metrisch skalierte Merkmale

`cor.test(x,y,method,conf.level,...)`

b) Ordinal oder nominal skalierte Merkmale

`chisq.test(x,y)` (beide abhängige Stichproben) oder
`chisq.test(x)` (Kontingenztafel)
`chisq.test(x,y)` = `chisq.test(table(x,y))`

Descr_Scatter.R Descr_Scatter_Heroin.R

Korrelation und Unabhängigkeit

Ordinal oder nominal skalierte Merkmale

Frage: Bestehen Abhängigkeiten?

Geschlecht - Studienfach Studiengang - Note
Geburtsmonat - IQ

Antwort: χ^2 - Unabhängigkeitstest (Pearson, 1908)

Annahme:
X hat Ausprägungen $a_1, \ldots, a_m$
Y hat Ausprägungen $b_1, \ldots, b_l$

(sind die Daten metrisch, so wird automatisch eine Klasseneinteilung vorgenommen.)

$$P(X = a_i) = p_{i.} \qquad P(Y = b_j) = p_{.j}$$

$$P(X = a_i, Y = b_j) = p_{ij}$$

Unabhängigkeitstests

Häufigkeitstabelle (= Kontingenztafel)

$X\|Y$	b_1	b_2	$\cdots$	b_j	$\cdots$	b_l	
a_1	h_{11}	h_{12}	$\cdots$	h_{1j}	$\cdots$	h_{1l}	$h_{1.}$
a_2	h_{21}	h_{22}	$\cdots$	h_{2j}	$\cdots$	h_{2l}	$h_{2.}$
$\cdots$							
a_i	h_{i1}	h_{i2}	$\cdots$	h_{ij}	$\cdots$	h_{iN}	$h_{i.}$
$\cdots$							
a_m	h_{m1}	h_{m2}	$\cdots$	h_{mj}	$\cdots$	h_{ml}	$h_{m.}$
	$h_{.1}$	$h_{.2}$	$\cdots$	$h_{.j}$	$\cdots$	$h_{.l}$	$h_{..}$=N

h_{ij}: Häufigkeiten

Unabhängigkeitstests

Die Häufigkeiten h_{ij} werden verglichen mit den theoretischen Häufigkeiten np_{ij}.

$$\begin{aligned} H_0: \quad & p_{ij} = p_{i.} \cdot p_{.j}, \quad && i = 1, ..., m, j = 1, ...l \\ H_1: \quad & p_{ij} \neq p_{i.} \cdot p_{.j}, \quad && \text{für ein Paar} (i, j) \end{aligned}$$

H_0: X und Y sind unabhängig.
H_1: X und Y sind abhängig.

Betrachten zunächst die Stichprobenfunktion

$$\widetilde{T} = \sum_i \sum_j \frac{(h_{ij} - np_{ij})^2}{np_{ij}}$$

Unabhängigkeitstests

Konstruktion der Teststatistik

Problem: $p_{i.}$ und $p_{.j}$ sind unbekannt. Sie müssen also geschätzt werden,
das sind $m + l - 2$ Parameter ($\sum p_{i.} = \sum p_{.j} = 1$)

$$\hat{p}_{i.} = \frac{h_{i.}}{N} \qquad \hat{p}_{.j} = \frac{h_{.j}}{N}$$

$$h_{i.} = \sum_{j=1}^{l} h_{ij} \qquad h_{.j} = \sum_{i=1}^{m} h_{ij}$$

Unabhängigkeitstests

Einsetzen der Schätzungen in $\widetilde{T}$ (unter H_0)

$$\begin{aligned} Q_P &= \sum_i \sum_j \frac{(h_{ij} - n\hat{p}_{i.}\hat{p}_{.j})^2}{n\hat{p}_{i.}\hat{p}_{.j}} \\ &= n \sum_i \sum_j \frac{(h_{ij} - \frac{h_{i.}h_{.j}}{n})^2}{h_{i.}h_{.j}} \\ &\sim \chi^2_{(m-1)(l-1)} \quad \text{approx. unter} H_0 \end{aligned}$$

Die Anzahl der Freiheitsgrade ergibt sich aus:

$$m \cdot l - 1 - \underbrace{(m + l - 2)}_{\#\text{geschätzte Werte}}$$

H_0 ablehnen, falls

$$Q_P > \chi^2_{(m-1)(l-1)}, \quad \text{bzw. falls p-Wert} < \alpha$$

Korrelation und Unabhängigkeit

Faustregel für die Anwendung des χ^2-Unabhängigkeitstests:

- alle $h_{ij} > 0$.
- $h_{ij} \geq 5$ für mindestens 80% der Zellen,
sonst Klassen zusammenfassen.

Descr_Freq_Heroin_Unabhaengigkeitstest

Korrelation und Unabhängigkeit

Weitere Unabhängigkeitstests (1)

- LQ-χ^2- Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{h_{ij}}{h_{i.} h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Continuity Adjusted χ^2 (bei R nur: 2x2-Tafel, dann Standard)

$$Q_c = N \sum_i \sum_j \frac{\max(0, |h_{ij} - \frac{h_{i.} h_{.j}}{N}| - 0.5)^2}{h_{i.} h_{.j}} \sim \chi^2_{(m-1)(l-1)}$$

- Mantel-Haenszel (`mantelhaen.test`, r_{XY}: Pearson-Korr.)

$$Q_{MH} = (N-1) r^2_{XY} \sim \chi^2_1$$

- Phi-Koeffizient

$$\Phi = \begin{cases} \frac{h_{11} h_{22} - h_{12} h_{21}}{\sqrt{h_{1.} h_{2.} h_{.1} h_{.2}}} & m = l = 2 \\ \sqrt{Q_p / n} & sonst \end{cases}$$

Weitere Unabhängigkeitstests (2)

- Kontingenzkoeffizient

$$P = \sqrt{\frac{Q_P}{Q_P + N}}$$

- Fishers Exact Test (`fisher.test`, bei 2x2-Tafeln)

durch Auszählen aller Tafel-Möglichkeiten bei gegebenen Rändern.
(gilt als etwas konservativ.)

- Cramers V

$$V = \begin{cases} \Phi & \text{falls} \quad \text{2x2 Tafel} \\ \sqrt{\frac{Q_P/N}{\min(m-1, l-1)}} & \text{sonst} \end{cases}$$

Weitere Unabhängigkeitstests (3)

Anmerkungen

- Mantel- Haenszel Test verlangt ordinale Skalierung, vgl. $(N-1)r_{XY}^2$

‘gut’ gegen lineare Abhängigkeit.

- Der χ^2 Unabhängigkeitstest testet gegen allgemeine Abhängigkeit.
- Der LQ-Test G^2 ist plausibel und geeignet.
- Der LQ-Test G^2 und der χ^2 Unabhängigkeitstest sind asymptotisch äquivalent.

Unabhängigkeitstests

Φ-Koeffizient (2x2 Tafel)

Y \ X	Sportler	Nichtsportler	Summe
w	p_{11}	p_{12}	$p_{1.}$
m	p_{21}	p_{22}	$p_{2.}$
Summe	$p_{.1}$	$p_{.2}$	1

$$X \sim Bi(1, p_{.2}) \qquad Y \sim Bi(1, p_{2.})$$

$$\begin{aligned} \mathbf{E}(X) &= p_{.2} \quad var(X) = p_{.2}(1 - p_{.2}) = p_{.2}p_{.1} \\ \mathbf{E}(Y) &= p_{2.} \quad var(Y) = p_{2.}(1 - p_{2.}) = p_{2.}p_{1.} \end{aligned}$$

$$cov(X, Y) = \mathbf{E}(X \cdot Y) - \mathbf{E}(X)\mathbf{E}(Y) = p_{22} - p_{.2}p_{2.}$$

Unabhängigkeitstests

Korrelationskoeffizient in einer 2x2 Tafel

$$\rho = \frac{p_{22} - p_{.2}p_{2.}}{\sqrt{p_{.2}p_{1.}p_{2.}p_{.1}}} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{.2}p_{2.}p_{1.}p_{.1}}}$$

$$\begin{aligned} p_{22} - p_{2.}p_{.2} &= p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) \\ &= p_{22} - (p_{21}p_{12} + p_{22}p_{12} + p_{21}p_{22} + p_{22}^2) \\ &= p_{22}(1 - p_{12} - p_{21} - p_{22}) - p_{21}p_{12} \\ &= p_{22}p_{11} - p_{21}p_{12} \end{aligned}$$

Für $m = l = 2$ ist der Phi-Koeffizient eine Schätzung des Korrelationskoeffizienten.

Inhalt

11.2 Lineare Regression

Einfache lineare Regression (vgl. Kap. 6.3)

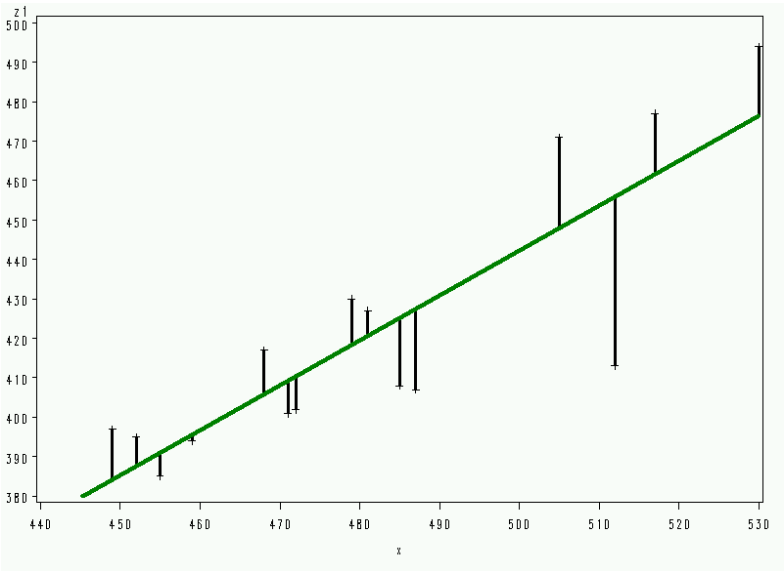
$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i \qquad \epsilon_i \sim (0, \sigma^2)$$

$$\begin{aligned} \hat{\theta}_1 &= \frac{S_{XY}}{S_X^2} \\ \hat{\theta}_0 &= \frac{1}{n}\left(\sum Y_i - \hat{\theta}_1 \sum X_i\right) = \overline{Y} - \hat{\theta}_1 \overline{X} \end{aligned}$$

als Lösung der Minimumaufgabe

$$\sum_{i=1}^{n} (Y_i - \theta_1 X_1 - \theta_0)^2 \rightarrow \text{ min.}$$

Lineare Regression (2)



Die Summe der Quadrate der Länge der Streckenabschnitte soll minimal werden.

$$S_{XY} = \frac{1}{n-1}\sum_i (X_i - \overline{X})(Y_i - \overline{Y})$$

$$S_X^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

Regression_Venusmuscheln
Regression_Plot

Lineare Regression (3)

Zur Erinnerung:

```
lm(y ~ x,data)
```

Lineare Regression

Multiple lineare Regression

Modell

$$\begin{aligned} Y_i &= \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \ldots + \theta_m x_{mi} + \epsilon_i \\ Y_i &= \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \ldots + \theta_m X_{mi} + \epsilon_i \end{aligned}$$

Y_i, ϵ_i Zufallsvariablen, unabh., $\epsilon_i \sim (0, \sigma^2)$, $i = 1 \ldots n$
$\theta_0 \ldots \theta_m, \sigma$: Modellparameter $\Rightarrow$ zu schätzen

Man unterscheidet Fälle:
$x_i = (x_{1i}, \ldots, x_{mi})$ fest, und $X_i = (X_{1i}, \ldots, X_{mi})$ zufällig
oder auch gemischt.
Matrix-Schreibweise:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

Lineare Regression

Multiple lineare Regression (2)

Modell

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1m} \\ . & . & \dots & . \\ 1 & X_{n1} & \dots & X_{nm} \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \dots \\ \theta_m \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \dots \\ \epsilon_n \end{pmatrix}$$

Methode der kleinsten Quadrate: Bestimme $\hat{\theta}$ so daß

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})$$

Lineare Regression

Multiple lineare Regression (3)

Kleinste Quadrat-Schätzung

Vor.: $\text{rg}(\mathbf{X}'\mathbf{X}) = m$ (voll)

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

```
theta = solve(t(X) %*% X) %*% t(X)%*%Y
```

wenn $(\mathbf{X}'\mathbf{X})$ nicht regulär: verallg. Inverse (Moore-Penrose)

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$$

```
theta = ginv(t(X) %*% X) %*% t(X)%*%Y
```

Lineare Regression

Multiple lineare Regression (4)

Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (1)

$$
\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1} &= \left[\begin{pmatrix} 1 & 1 & \dots & 1 \\ X_{11} & . & \dots & X_{n1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ \dots & \dots \\ 1 & X_{n1} \end{pmatrix} \right]^{-1} \\
&= \begin{pmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{pmatrix}^{-1} \qquad (X_i = X_{1i}) \\
&= \frac{1}{n \sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}
\end{aligned}
$$

Lineare Regression

Multiple lineare Regression (5)

Kleinste Quadrat-Schätzung, Spezialfall $m = 1$ (2)

$$
\begin{aligned}
\mathbf{X}'\mathbf{Y} &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ X_1 & . & \dots & X_n \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix} \\
\hat{\theta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i \\ -\sum X_i \sum Y_i + n \sum X_i Y_i \end{pmatrix}
\end{aligned}
$$

Lineare Regression

Multiple lineare Regression (6)

Schätzung für $\mathbf{Y}$: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$

Vergleiche mit $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$

Einsetzen von $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$:

$$\begin{aligned} \hat{\mathbf{Y}} &= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y} \end{aligned}$$

$\mathbf{H}$: Hat-Matrix

Aus dem Beobachtungsvektor $\mathbf{Y}$ wird der geschätzte Beobachtungsvektor $\hat{\mathbf{Y}}$.

Lineare Regression

Multiple Lineare Regression (7)

Quadratsummenaufspaltung:

$$\underbrace{\sum (Y_i - \overline{Y})^2}_{SST} = \underbrace{\sum (\hat{Y}_i - \overline{Y})^2}_{SSM} + \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{SSE}$$

$MST = \frac{1}{n-1} SST$: Schätzung für die Gesamtvarianz.
$MSE = \frac{1}{n-m-1} SSE = \hat{\sigma}^2$. (erwartungstreu)
$MSM = \frac{1}{m} SSM$ ($m+1$ Einflussvariablen)

Bestimmtheitsmaß (wie bei der Varianzanalyse)

$$R^2 = \frac{SSM}{SST}.$$

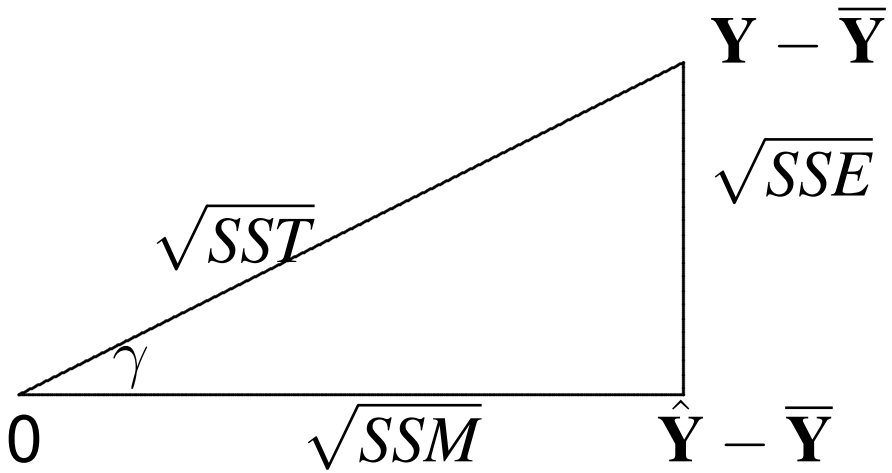
Geometrische Veranschaulichung

zur Multiplen Linearen Regression

$\mathbf{Y} = (Y_{11}, \ldots, Y_{kn_k})$ Dimension N

$\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_{kn_k})$

$\overline{\mathbf{Y}} = (\underbrace{\overline{Y}, \ldots, \overline{Y}}_{\text{n mal}})$, $\overline{Y} = \frac{1}{N} \sum_{i,j} Y_{ij}$



$$SSM + SSE = SST \qquad R^2 = cos^2\gamma$$

$$||\hat{\mathbf{Y}} - \overline{Y}||^2 + ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \overline{Y}||^2$$

Lineare Regression

Multiple Lineare Regression (8)

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_m = 0 \quad H_1 :\sim H_0$$

Unter der Annahme $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ gilt:

$$F = \frac{SSM}{SSE} \cdot \frac{n-m-1}{m} \sim F_{m,n-m-1}$$

```
md = lm(y~ x1+x2+x3,data) md.sum =
summary(lm(y~ x1+x2+x3,data))
```

Regression_Tibetan Regression_Phosphor

Lineare Regression

Multiple Lineare Regression (9)

Zusätzliche Hypothesen, z.B.

$$H_{0a} : \theta_1 = 0 \,, \qquad H_{1a} : \theta_1 \neq 0$$
$$H_{0b} : \theta_k = 0 \,, \qquad H_{1b} : \theta_k \neq 0$$
$$H_{0c} : \theta_1 = \theta_2 = 0 \,, \qquad H_{1c} : \theta_1 \neq 0 \vee \theta_2 \neq 0$$

Lineare Regression

Multiple Lineare Regression (10)

R^2-adjustiert für Anzahl p der Parameter im Modell

$$Adj\_R^2 = 1 - \frac{n-i}{n-p}(1-R^2)$$

$$\begin{cases} i = 0 & \text{ohne intercept} \\ i = 1 & \text{mit intercept} \end{cases}$$

Dependent Mean: Mittelwert der abhängigen Variable (Y)

StdError MeanPredict: Standardfehler für vorhergesagten Erwartungswert

Lineare Regression

Multiple Lineare Regression (11)

```
md = lm(y~x1+x2+x3,dat); md.sum = summary(md)
```

Rückgaben von `lm`/`summary(lm())`

`md$model$y`	Y_i
`md$fitted.values`	$\hat{Y}_i = \hat{\boldsymbol{\theta}}\mathbf{X}$
`md.sum$sigma`	$\hat{\sigma}_{\hat{Y}_i}$
`confint(md)`	KI für θ
`mod$residuals`	$e_i = Y_i - \hat{Y}_i$
`rstandard(md)`	StdErrorResidual : $s\sqrt{1-h_{ii}}$
`md.sum$r.squared`	R^2
`md.sum$adj.r.squared`	Adj_R^2
`hatvalues(md)`	Diagonale von $\mathbf{H}$

Lineare Regression

Multiple Lineare Regression (12)

Konfidenzintervalle für allg. Parameter ϑ_i:

$$\frac{\hat{\vartheta}_i - \vartheta_i}{S_{\hat{\vartheta}_i}} \sim t_{n-1} \qquad \underline{\text{Vor.}}\ \epsilon_j \sim \mathcal{N}(0, \sigma^2) \quad \text{u.a.}$$

KI: $[\hat{\vartheta}_i - t_{1-\frac{\alpha}{2},n-1} \cdot S_{\hat{\vartheta}_i}, \hat{\vartheta}_i + t_{1-\frac{\alpha}{2},n-1} \cdot S_{\hat{\vartheta}_i}]$

95% Konfidenzintervall für $\underline{E(Y_i)}$

($\vartheta_i = \mathbf{E}(Y_i)$, `predict(lm(...),interval="confidence")`)
Nur die Variabilität in der Parameterschätzung wird berücksichtigt.

Lineare Regression

Multiple Lineare Regression (13)

95% Konfidenzintervall für Vorhersagen $\underline{Y}_i$

$(\vartheta_i = Y_i)$
Die Variabilität im Fehlerterm wird mit berücksichtigt.

95% Konfidenzintervall für θ

$(\vartheta_i = \theta_j,$ `confint(md)` $)$

Multiple Lineare Regression

Residualanalyse (1)

Studentisierte Residuen (`rstudent(lm(...))`)

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

$e_i = y_i - \hat{y}_i$ (Residuen) sind korreliert,
var $e_i = \sigma^2(1-h_{ii})$ $s = \hat{\sigma}$

Cook's D_i (`cooks.distance(lm(...))`)

$$D_i = \frac{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})}{(m+1)S^2}, \qquad i = 1...n$$

beschreibt den Einfluß der i-ten Beobachtung auf die Parameterschätzung
$\hat{\boldsymbol{\theta}}_{(i)}$: KQS von $\boldsymbol{\theta}$ ohne Beobachtung i.
Faustregel: $D_i > 1 \rightarrow$ 'starker' Einfluß

Multiple Lineare Regression

Residualanalyse (2)

Predicted Residual SS (`PRESS`,u.a. Paket qpcR)

$$\sum (y_i - \hat{y}_{i(i)})^2$$

$\hat{y}_{i(i)}$: i-te Beobachtung weggelassen.

"Test" auf Autokorrelation: Durbin-Watson-Test (`dwt(lm(...))`, Paket car)

$$DW = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

DW=2: Unkorreliertheit der Residuen

Multiple Lineare Regression

Residualanalyse (3)

Weitere Bewertung der Residuen

```
mod = lm(y~x1+x2+x3,data)
plot(residuals(mod))
shapiro.test(residuals(mod))
points(rstudent(mod))
shapiro.test(rstudent(mod))
point(PRESS(mod)$residuals)
shapiro.test(PRESS(mod)$residuals)
```

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Modellwahl in der linearen Regression

Akaikes „an information criterion“:

`step(model,scope,direction)`

`scope=c(lower,upper)` oder `scope=upper`

`direction=` `"forward"`,`"backward"` oder `"both"`

`backward`: Alle Variablen in `upper`, die mit größten p-Wert werden nacheinander herausgenommen, bis nur noch Var aus `lower`

`forward`: Start mit Variablen aus `lower`, die Var. mit kleinstem p-Wert kommt hinzu bis max. alle aus `upper` enthalten sind.

`both`: Schritte in beide Richtungen möglich.

Lineare Regression

Modellwahl in der linearen Regression (2)

Einzelschritte

`add1(model,scope)` und
`drop1(model)`
Eine Variable (aus Formel `scope`) wird bei `add1` hinzugefügt, bei `drop1` eine aus dem bisherigen Modell entfernt.

alle Teilmodelle testen

```
leap(x=data[c("x1","x2")],y=data["y"],
method="Cp")
```
testet alle Modelle mit Variablen aus den Spalten von `x` und abh. Variable `y`.
Berechnet jeweils das Kriterium `method`. Zu Mallows Cp s.u. (`ädjr2"` und `"r2"` auh möglich)

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

a) Teste auf $rg(\mathbf{X}'\mathbf{X})$ nicht voll ($< m+1$)

```
rankMatrix(data[c("x1","x2")])
```

b) Condition number

$\sqrt{\frac{\lambda max}{\lambda min}}$ $\lambda_{max}, \lambda_{min}$ größter u. kleinster Eigenwert von $\mathbf{X}'\mathbf{X}$ (ohne 1-Spalte).

```
rankMatrix(as.matrix(data[c("x1","x2")]))
```

gr. Konditionszahl (etwa > 30): schlechte Kond. ($\approx$ lin. Abh.)

c) C(p): Mallows (1973) Kriterium für die Modellwahl

$$C(p) = \frac{SSE_p}{MSE} - n + 2p$$

SSE_p: SSE im Modell mit p Parametern

Multiple Lineare Regression (Forts.)

Modellwahl in der linearen Regression

$$R^2 = \frac{SSM}{SST}.$$

$$C(p) = \frac{SSE_p}{MSE} - n + 2p$$

SSE_p: SSE im Modell mit p Parametern
Ziel: R^2 groß, $C(p)$ nahe p

Idee von $C(p)$: Wenn die Wahl von p Parametern gut, dann

$$MSE \approx MSE_p = \frac{SSE_p}{n-p} \quad \Rightarrow \quad C(p) \approx n - p - n + 2p = p$$

Regression_Tibetan_Modellwahl

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

Einfache Varianzanalyse: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ \dots \\ Y_{n_1 1} \\ Y_{12} \\ \dots \\ Y_{n_2 2} \\ \dots \\ \dots \\ Y_{1k} \\ \dots \\ Y_{n_k k} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ . & . & . & & . \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ . & . & . & & . \\ 1 & . & 1 & \dots & 0 \\ . & . & 0 & \dots & 0 \\ . & . & . & & . \\ 1 & . & . & 0 & 1 \\ . & . & . & & \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_k \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \epsilon_{n_k k} \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

Lineare Regression

Multiple Lineare Regression (Fortsetzung)

$$\begin{pmatrix} Y_1 \\ \dots \\ \dots \\ Y_N \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ . & \dots & & \dots \\ . & \dots & & \dots \\ 1 & X_{N1} & \dots & X_{Np} \end{pmatrix} \begin{pmatrix} \mu \\ \theta_1 \\ \dots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \dots \\ \dots \\ \epsilon_N \end{pmatrix} \Leftrightarrow$$

$$\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

Inhalt

11.3 Robuste lineare Regression

Mögliche Probleme bei der linearen Regression

Probleme	Lösungsansätze
• Ausreißer	Robuste Lineare Regression
• keine Normalverteilung	Datentransformation, (L_1-Regression)
• kein linearer Zusammenhang	Nichtlineare Regression Nichtparametrische Regression
• Zielvariable nicht stetig	Logistische Regression

Robuste Lineare Regression (Skizze)

Ausreißer können auftreten in
- Y-Richtung
- X-Richtung(en) (Leverage points)
- Y- und X- Richtungen

Fall: Ausreißer(verdacht) in Y-Richtung:
es werden nicht die Abstandsquadrate minimiert, sondern (z.B.) die Gewichtsfunktion (Bisquare Biweight, Huber)

$$W(x,c) = \begin{cases} 1 - \left(\frac{x}{c}\right)^2 & \text{falls} \quad |x| < c \\ 0 & \text{sonst.} \end{cases}$$

verwendet.

Robuste Lineare Regression (2)

Außerdem wird der Skalenparameter σ nicht durch s sondern durch den MAD geschätzt.

```
#rlm aus Paket MASS
rlm(formula,data,scale.est="MAD",
  psi=psi.bisquare)
#oder psi.huber, psi.hampel
```

Regression_Phosphor

Robuste Lineare Regression (3)

Diagnosestatistiken

Ausreißer: standardis. robust residual $>$ cutoff (outlier)

Leverage Point: robuste MCD-Distanz $>$ cutoff (Leverage)

Mahalanobis-Distanz

$\approx$ mit Kovarianzmatrix gewichteter mittlerer quadratischer Abstand von $\overline{X}$.

Robust MCD Distance:

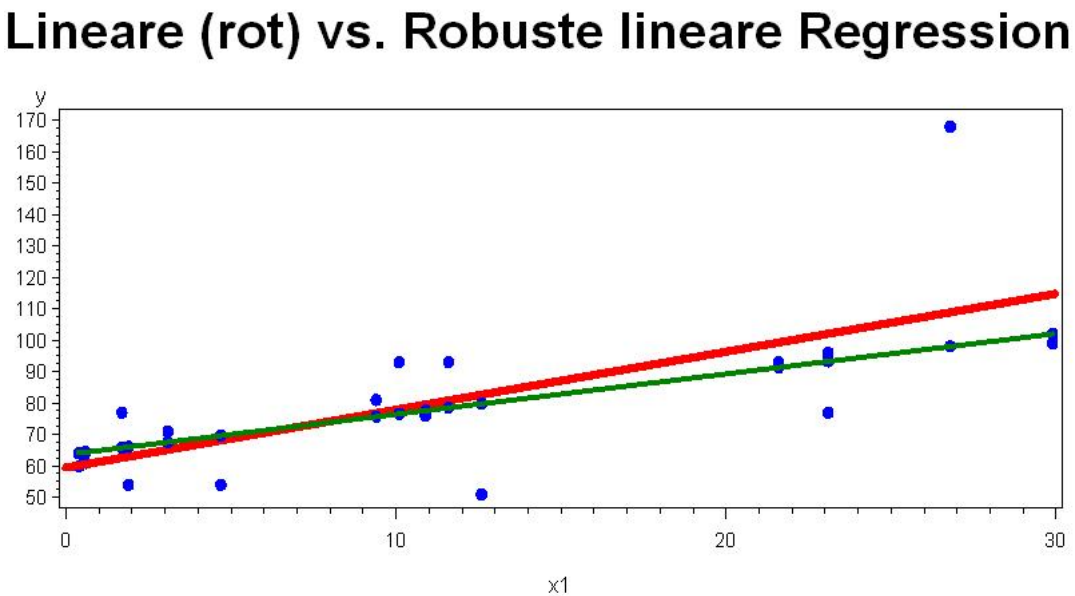
anstelle von $\overline{X}$: robuste multivariate Lokationsschätzung (MCD)

Goodness of fit: zum Modellvergleich

je größer R^2, je kleiner AICR, BICR desto besser.

Robuste Lineare Regression (3)

Beispiel: Phosphorfraktionen



Inhalt

11.4 Nichtlineare Regression

Quasilineare Regression

z.B. Polynomregression

$Y_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \epsilon_i$

wird auf lineare Regression zurückgeführt

$x_{ij} := x_i^j$

Echt nichtlineare Regression, z.B. Wachstumskurven

$$y = \alpha + \frac{\gamma}{1 + exp(-\beta(x - \mu))} \quad \text{logistische Fkt.}$$

$$y = \alpha + \gamma exp(-exp(-\beta(x - \mu))) \quad \text{Gompertzfkt.}$$

Modell, f wird als bekannt angenommen

$$\begin{aligned} Y &= f(x,\theta) + \epsilon \qquad \epsilon \sim (0,\sigma^2) \\ \mathbf{Y} &= \mathbf{F}(\mathbf{X},\boldsymbol{\theta}) + \boldsymbol{\epsilon} \end{aligned}$$

$$L(\boldsymbol{\theta}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sum_i (Y_i - \mathbf{F}(\mathbf{X}_i,\boldsymbol{\theta}))^2 \longrightarrow min_\theta$$

Dazu werden Iterationsverfahren verwendet.

```
f = function(x) ...
nlm(f,p=Anfangswertswerte,
    print.level=2)
```

Nichtlineare Regression (2)

Ausgabe

`minimum`	Zielwert
`gradient`	Ableitung
`code`	Abbruchgrund (s.Hilfe)
`iterations`	Anzahl Schritte

Nlin1_usapop.R
Nlin1_usapop_est.R
Nlin2_wind.R

Inhalt

11.5 Nichtparametrische Regression

Modell: f unbekannt, aber "glatt"

$$Y_i = f(\mathbf{x}_i) + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$ ($\mathbf{x}_i$ fest oder zufällig)

$$\min_{f \in C^2} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$$

- $\int (f'')^2$: Strafterm
- λ: Glättungsparameter
  - $\lambda \to 0$: Interpolierender Spline
  - $\lambda \to \infty$: lineare Regression

Lösung der Minimumaufgabe: natürlicher kubischer Spline

Nichtparametrische Regression (2)

Thin Plate Splines aus Paket fields:

```
Tps(x,y,lambda)
```

Wahl der Glättungsparameter

Kreuzvalidierung (Standard)

vorgeben: `lambda=Wert`

Nichtparametrische Regression (3)

Ausgabe

$Log_{10}(n * \hat{\lambda})$
Strafterm $\int (f'')^2(t)\, dt$
Residual Sum of Squares
Schätzung für σ, $\hat{\sigma}^2 = \frac{RSS}{sp(I-A)}$, A: entspricht der Hat-Matrix.

Npar_USApop.R
Npar_Banknote.R

Visualisierung

```
#vier Diagramme, also 2x2
par(mfrow=c(2,2))
plot(Tps(x,y,lambda))
```

Inhalt

11.6 Logistische Regression

Y: Binäre Zielgröße, $P(Y = 1) = p, P(Y = 0) = 1 - p$,
$Y \sim B(1, p)$

Wenn wir lineare Regression machen würden:

$$\begin{aligned} Y_i &= \alpha + \beta x_i + \epsilon_i \\ \mathbf{E}Y_i &= \alpha + \beta x_i, \quad \mathbf{E}\epsilon_i = 0 \\ p_i &= \alpha + \beta x_i \end{aligned}$$

Problem: Wahrscheinlichkeiten sind beschränkt, lineare Funktionen aber nicht.

Ausweg: Odds ratio $OR := \frac{p}{1-p}$

nach oben unbeschränkt, aber nicht nach unten

Logistische Regression (2)

Logit

$$Logit(p) := \ln\left(\frac{p}{1-p}\right)$$

ist auch nach unten unbeschränkt.

Modell

$$\begin{aligned} Logit(p_i) &= \ln\left(\frac{p_i}{1-p_i}\right) \\ &= \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \boldsymbol{\beta}' \mathbf{x}_i, \end{aligned}$$

$i = 1, \ldots, n, p_i = P(Y_i = 1)$.
$\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ik}), \boldsymbol{\beta}' = (\alpha, \beta_1, \ldots, \beta_k)$.

Umstellen der letzten Gleichung liefert

Logistische Regression (3)

$$p_i = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}} = 1 - \frac{1}{1 + e^{\beta' \mathbf{x}_i}}.$$

Gegeben sind Beobachtungen: $(y_i, \mathbf{x}_i)$.
Unbekannt sind p_i.

Frage: Wie schätzen wir β ?

Methode: Maximum-Likelihood

```
glm(y~x,data,family=binomial(link="logit"))
```

Logistic_banknote
Logistic_tibetan
Logistic_water

Logistische Regression (4)

Maximum-Likelihood Schätzung der Parameter

Idee: Eine Schätzung ist “gut”, wenn sie für die beobachteten Daten die “plausibelste” ist, wenn sie eine hohe Wkt. produziert.

Ziel: maximiere (die Bebachtungen sind unabhängig)

$$L = P(y_1) \cdot P(y_2) \cdots P(y_n) = \prod_{i=1}^{n} P(y_i).$$

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{mit Wkt. } p_i \\ 0 & \text{mit Wkt. } 1 - p_i \end{cases} \\ P(y_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \end{aligned}$$

$$\begin{aligned} P(0) &= p_i^0(1-p_i)^{1-0} = 1 - p_i \\ P(1) &= p_i^1(1-p_i)^{1-1} = p_i \end{aligned}$$

hier: x_i bekannt (Beobachtungen) — p_i zu schätzen

Logistische Regression (5)

Maximum-Likelihood Schätzung der Parameter (2)

Einsetzen

$$
\begin{aligned}
L &= \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i} (1 - p_i) \\
\ln L &= \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1 - p_i}\right) + \sum_{i=1}^{n} \ln(1 - p_i) \\
&= \sum_{i=1}^{n} \boldsymbol{\beta}' \mathbf{x}_i y_i - \sum_{i=1}^{n} \ln\left(1 + e^{\boldsymbol{\beta}' \mathbf{x}_i}\right)
\end{aligned}
$$

Da der Logarithmus monoton wachsend ist, genügt es $\ln L$ zu maximieren.

Logistische Regression (6)

$$\begin{aligned}
\frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \left(1 + e^{\boldsymbol{\beta}' \mathbf{x}_i}\right)^{-1} e^{\boldsymbol{\beta}' \mathbf{x}_i} \\
&= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \left(1 + e^{-\boldsymbol{\beta}' \mathbf{x}_i}\right)^{-1} \\
&= \sum_{i=1}^{n} \mathbf{x}_i y_i - \sum_{i=1}^{n} \mathbf{x}_i \hat{y}_i,
\end{aligned}$$

wobei

$$\hat{y}_i = \frac{1}{1 + e^{-\boldsymbol{\beta}' \mathbf{x}_i}}$$

die Vorhersagewkt. für y_i bei gegebenen $\mathbf{x}_i$.

Logistische Regression (7)

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = 0$$

ist Nichtlineares Gleichungssystem
$\rightarrow$ numerische Lösung, z.B. Newton-Raphson Methode
hier: = Fisher Scoring
$\mathbf{U}(\boldsymbol{\beta})$: Vektor der ersten Ableitungen von $\ln L$
$\mathbf{I}(\boldsymbol{\beta})$: Matrix der zweiten Ableitungen von $\ln L$
Iteration

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}(\boldsymbol{\beta}_j)$$

Konvergenz? hoffentlich.

Vergleiche: Newton-Verfahren ($k = 1$) zur Lösung von $g(x) = 0$.

Logistische Regression (8)

Output von `print(summary(glm(...)))`

Aufruf von `glm`
Modellanpassungsstatistiken (Deviance)
geshätzte Parameter
AIC
Anzahl der Fisher-Scoring-Schritte
ggf. Warnungen bei Nichtkonvergenz oder
angepassten Wahrscheinlichkeiten von 0 oder 1

wie bei `lm` enthaltn die zurückgegebenen Objekte weitere Informationen. (`md = glm(...);md$...;summary(md)$dots`)

Logistische Regression (9)

Modellanpassungsstatistiken

zum Vergleich verschiedener Modelle
je größer AIC, SC, desto besser
je kleiner Deviance $-2 \ln L$ desto besser
$-2 \ln L$: Abweichung vom saturierten Modell,
d.h. vom anhand der Daten (bei perfekter Anpassung) möglichen Modell
Hosmer-Lemeshov Anpassungstest (Option LACKFIT)

Logistische Regression (10)

Vorhersagefähigkeit des Modells

(Association of Predicted probabilities)
alle möglichen Paare (y_i, y_j) werden verglichen bzgl. ihres Vorhersagewertes $(\hat{y}_i, \hat{y}_j)$
Anteil der konkordanten Paare C
Kendall-Konkordanzkoeffizient Tau-a
Somer's D, Gamma, c hängen mit C zusammen.
Tau-a kann als Bestimmtheitsmaß interpretiert werden.

Inhalt

Regressionsverfahren

Kurze Übersicht (1)

a) Lineare Regression

Modell:

$$Y_i = \theta_0 + \sum_{j=1}^{m} \theta_j X_{ij} + \epsilon_i$$

$\epsilon_i \sim (0, \sigma^2)$, $i = 1, \ldots, n$
Y_i, ϵ_i zufällig
X_i zufällig oder fest
$\theta_0 \ldots \theta_m; \sigma$: Modellparameter

`lm`

```
lm(y~ x1+x2+x3,data)
```

Regressionsverfahren

Kurze Übersicht (2)

b) Robuste Lineare Regression

Modell wie bei der linearen Regression

$$Y_i = \theta_0 + \sum_{j=1}^{m} \theta_j X_{ij} + \epsilon_i$$

robuste Abstandsfunktion
MAD statt s als Skalenschätzung.

`rlm` aus Paket MASS

```
rlm(formula,data,scale.est="MAD",
psi=psi.bisquare)
```

Regressionsverfahren

Kurze Übersicht (3)

c) Nichtlineare Regression

Modell:

$$Y_i = f(X_{1i}, \ldots, X_{mi}, \theta_1, \ldots, \theta_p) + \epsilon_i$$

f: bekannt (i.A. nichtlinear)

`nlm`

```
f = function(x) ... nlm(f,p=Anfangswertswerte,
print.level=2)
```

Regressionsverfahren

Kurze Übersicht (4)

d) Nichtparametrische Regression

Modell:

$$Y_i = f(X_{1i}, \ldots, X_{mi}) + \epsilon_i$$

f unbekannt, aber "glatt", z.B. $f \in C^2$.

`Tps` aus Paket fields

```
Tps(x,y,lambda)
```

Regression_Phosphor_Uebersicht.R

Regressionsverfahren

Kurze Übersicht (5)

e) Logistische Regression

Y: binäre Zielgröße

$$p_i = P(Y_i = 1) = \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}}.$$

Parameter: β.
Odds ratio: $\frac{p_i}{1-p_i}$

`glm`

```
glm(y x,data,family=binomial(link="logit"))
```

Inhalt (1)

Inhalt (2)

Inhalt (3)

12. Zufallszahlen

- werden nach einem determinist. Algorithmus erzeugt $\Rightarrow$ Pseudozufallszahlen

- wirken wie zufäll. Zahlen (sollen sie jedenfalls)

Algorithmus:

Startwert x_0, $\quad x_{n+1} = f(x_n)$ (z.B. Kongruenzen)

Der alte Generator von SAS

$$x_{n+1} = \underbrace{397204094}_{2\cdot 7\cdot 7\cdot 4053103} x_n \mod(2^{31}-1) \qquad u_n = \frac{x_n}{2^{31}-1}$$

liefert gleichverteilte Zufallszahlen $u_n \in (0,1)$.

Zufallszahlen (2)

Der aktuelle Standard-Generator von R und SAS

Mersenne Twister

Der Algorithmus ist schwieriger (s. z.B. Wikipedia)

Algorithmus abfragen oder ändern

`RNGkind()` bzw. `RNGkind(kind="neuer Algo", normal.kind="Algo für NV")`
auch eigene Generatoren möglich (s. `?Random.user`)

zufälliger Startwert

`set.seed(`x_1`)`
Der interne Startwert wird dann durch x_1 ersetzt

Zufallszahlen (3)

auf $(a,b)^k$ gleichverteilter Zufallsvektor

`x=runif(`k`,min=`a`,max=`b`)` $(0,1)^k$ ist Standard

Normalverteilte Zufallszahlen

`x=rnorm(`k`,mu=`μ`,sd=`σ`)` erzeugt Zufallsvektor mit $\mathcal{N}(\mu,\sigma)$-verteilten Komponenten. $\mu = 0$ und $\sigma = 1$ sind Standard.

andere Verteilungen

Zu jeder Verteilung `p...` existiert i.d.R. neben Dichte `d...` und Quantilfunktion `q...` auch ein Zufallsgenerator `r...`.

Zufallszahlen (4)

vorgegebene stetige Verteilung

wird z.B. aus gleichverteilter Zufallsvariable U_i mittels Quantilfunktion ($F^{-1}(U_i)$) gewonnen.

diskrete Verteilungen

werden erzeugt durch Klasseneinteilung des Intervalls $(0, 1)$ entsprechend der vorgegebenen Wahrscheinlichkeiten p_i, also

$$(0, p_1], (p_1, p_1 + p_2], (p_1 + p_2, p_1 + p_2 + p_3],$$

$$\ldots, (p_1 + \cdots + p_{k-1}, 1)$$

Zufallszahlen (5)

Wünschenswerte Eigenschaften

- Einfacher Algorithmus, wenig Rechenzeit.
- möglichst viele verschieden Zufallszahlen sollen erzeugbar sein

$\Rightarrow$ lange Periode.

- k-Tupel $(U_1, \ldots, U_k) \sim R(0,1)^k$, $k \leq 10$

$\Rightarrow$ Test auf Gleichverteilung.

- "Unabhängigkeit"

Test auf Autokorrelation (z.B. Durbin-Watson Test, vgl. Regression)
Plot der Punkte (U_i, U_{i+k}), $k = 1, 2...$
es sollten keine Muster zu erkennen sein.

Zufallszahlen_test.R Zufallszahlen_Dichte.R

Inhalt (1)

Inhalt (2)

Inhalt (3)

13. Clusteranalyse

Ziel: Zusammenfassung von
- “ähnlichen” Objekten zu Gruppen (Clustern),
- unähnliche Objekte in verschiedene Cluster.

Cluster sind vorher nicht bekannt.

20 Patienten, Blutanalyse

Merkmale: Eisengehalt X_1, alkalische Phosphate X_2

Umweltverschmutzung in verschiedenen Städten

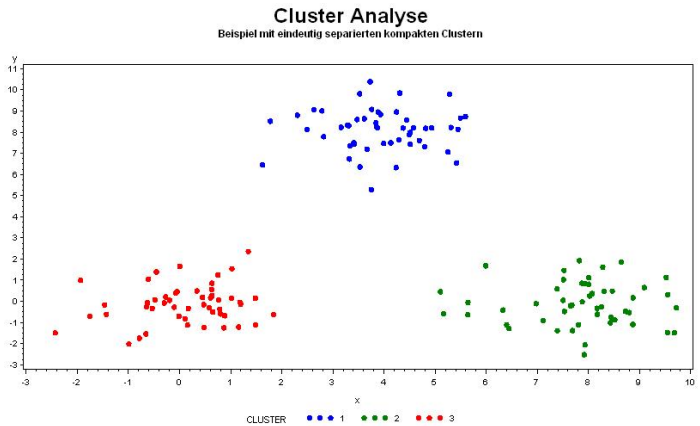
Merkmale: Schwebeteilchen, Schwefeldioxid

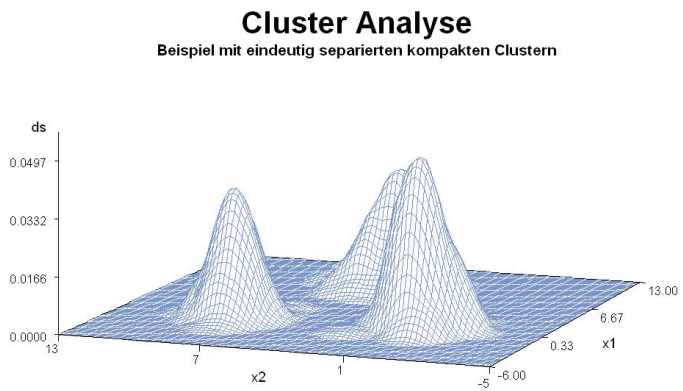
Byzantinische Münzen

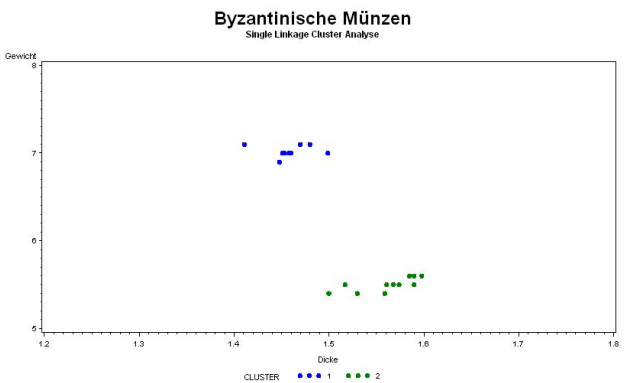
Lassen sich gesammelte Münzen verschiedenen Epochen zuordnen?

Clusteranalyse

Beispiel







Clusteranalyse

Wir unterscheiden:

partitionierende Clusteranalyse

Zahl der Cluster ist vorgegeben

`kmeans(x,centers,iter.max,algorithm)`

`centers` kann Anzahl k sein oder Liste von k Zentren.

hierarchische Clusteranalyse

`hclust(dist(...))`

ggf. `plot(...)`

Fuzzy Clusteranalyse

`fanny` (Paket cluster)

Clusteranalyse

Abstandsdefinitionen (p: # Merkmale)

Euklidischer Abstand (das ist Standard)

$$d_E^2(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2$$

City-Block Abstand (Manhattan-Abstand)

$$d_C(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

Tschebyschev-Abstand

$$d_T(x,y) = \max_i |x_i - y_i|$$

Clusteranalyse

Anmerkungen zu den Abständen

- Die Variablen sollten i.A. vor der Analyse standardisiert werden (`sapply(data,scale)`), da Variablen mit großer Varianz sonst großen Einfluß haben.
  davor: Ausreißer beseitigen.

Hierarchische Clusteranalyse

Methoden (1)

Die Methoden unterscheiden sich durch die Definition der Abstände $D(C_i, C_j)$ zwischen Clustern C_i und C_j.

Single Linkage

$$D_S(C_i, C_j) = \min\{d(k,l), k \in C_i, l \in C_j)\}$$

Complete Linkage

$$D_C(C_i, C_j) = \max\{d(k,l), k \in C_i, l \in C_j)\}$$

Centroid

$D_{CE}(C_i, C_j) = d(\overline{X}_i, \overline{X}_j)$ Abstände der Schwerpunkte

Hierarchische Clusteranalyse

Methoden (2)

Average Linkage

$D_A(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k \in C_i, j \in C_j} d(k, l)$

Ward

ANOVA-Abstände innerhalb der Cluster minimieren, außerhalb maximieren. Nach Umrechnen erhält man
$D_W(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} D_{CE}(C_i, C_j)$.

Density Linkage

beruht auf nichtparametrischer Dichteschätzung (DENSITY, TWOSTAGE)

Hierarchische Clusteranalyse

Tendenzen

WARD:	Cluster mit etwa gleicher Anzahl von Objekten
AVERAGE:	ballförmige Cluster
SINGLE:	große Cluster, “Ketteneffekt”, langgestreckte Cluster
COMPLETE:	kompakte, kleine Cluster

Im Mittel erweisen sich **Average Linkage** und **Ward** sowie die nichtparametrischen Methoden als die geeignetsten Methoden.

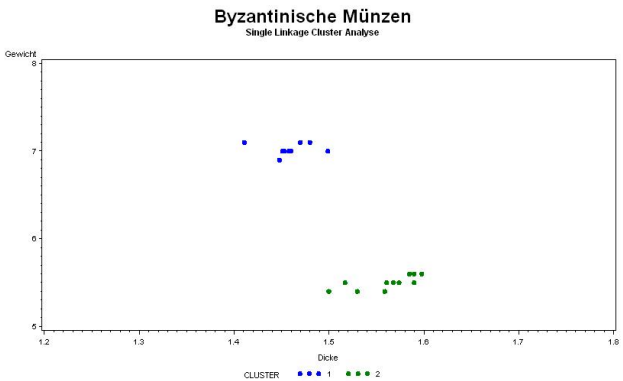
Hierarchische Clusteranalyse

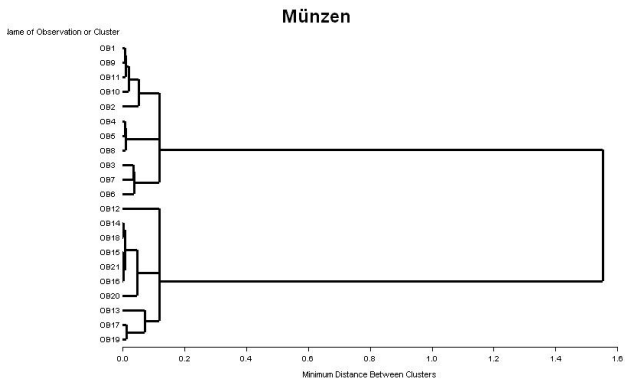
Agglomerative Verfahren

1. Beginne mit der totalen Zerlegung, d.h.
   $Z = \{C_1, ..., C_n\}, C_i \cap C_j = \varnothing \quad C_i = \{O_i\}$
2. Suche C_r, C_l : $\quad d(C_r, C_l) = \min_{i \neq j} d(C_i, C_j)$
3. Fusioniere C_r, C_l zu einem neuen Cluster:
   $$C_r^{new} = C_r \cup C_l$$
4. Ändere die r-te Zeile und Spalte der Distanzmatrix durch Berechnung der Abstände von C_r^{new} zu den anderen Clustern!
   Streiche die l-te Zeile und Spalte!
5. Beende nach n-1 Schritten, ansonsten fahre bei 2. mit geänderter Distanzmatrix fort!

Clusteranalyse

Beispiel





Hierarchische Clusteranalyse

Anmerkungen

- `hclust`s Methoden sind agglomerativ. Im Paket cluster sind die Methoden `agnes` und `diana` enthalten, letztere bietet divisive Methoden.

Hierarchische Clusteranalyse

zu WARD:

ANOVA Abstände innerhalb eines Clusters i

$$D_i = \frac{1}{n_i} \sum_{l \in C_i} d^2(O_l, \overline{X}_i)$$

Fusioniere die Cluster C_i und C_j, wenn

$$D_{CE}(C_i, C_j) - D_i - D_j \longrightarrow \min_{i,j}$$

Clusteranalyse

Durchführung

`x.dist = dist(x,method)`
`method` ist die zu verw. Norm
Falls gewünschte Norm mit `dist` nicht möglich oder Distanzmatrix aus anderer Quelle als normiertem Raum:
`x.dist = as.dist(Distanzmatrix)`

`x.clust = hclust(x.dist,method)`
`method` kann `"ward"`,`"single"`, `"complete"`, `"average"`, `"mcquitty"`, `"median"` oder `"centroid"` sein

`plot(x.clust)` Dendrogramm
`cutree(x.clust,k` oder `h)`
Cluster der Elemente nach Höhe `h` oder Clusteranzahl `k`.

Hierarchische Clusteranalyse

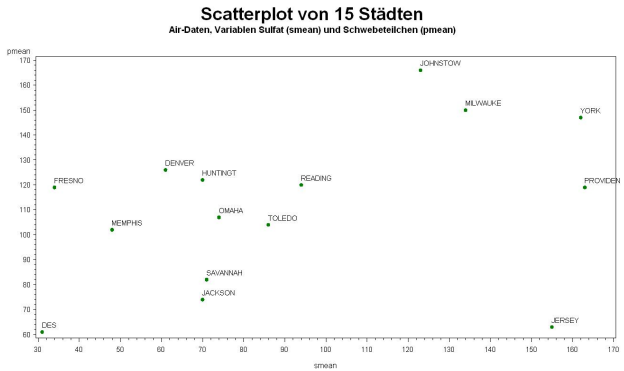
Das Objekt `x.clust=hclust(...)`

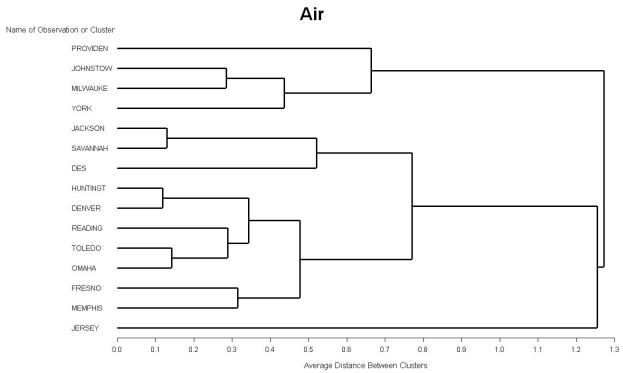
`x.clust$height[i]`	Höhe im Baum von `x[i]`
`x.clust$merge`	Reihenfolge der Aggl. (siehe Hilfe)
`x.clust$order`	Permutation von `x`, sodass Dendrogramm ohne Überschneidungen plottbar.

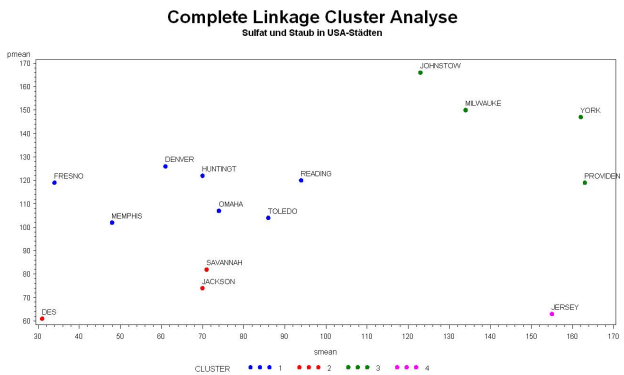
Cluster_Air.R
Cluster.R
Cluster_Banknoten.R
Cluster_Muenzen.R

Hierarchische Clusteranalyse

Beispiel: Luftverschmutzung in USA-Städten







3D-Darstellung von Datenpunkten

`cloud(z~x+y,data)` aus Paket lattice
`scatterplot3d(dfr)` aus gleichnamigem Paket
`plot3d(dfr)` aus Paket rgl (braucht OpenGL, nicht für Export geeignet, per Maus drehbar)
alle ähnlich zu normalem `plot` aufrufbar

3D-Darstellung von Flächen, Kontur-Plot

`persp` und `persp3d`

`persp(x,y,z)` plottet beschr. Fläche, wobei `z` `length(x)` Zeilen und `length(y)` Spalten hat.
D.h. für alle Koordinatenpaare aus `x` und `y` ist ein Wert `z` vorhanden.
`perp3d` nutzt OpenGL und ist mit `plot3d` kombinierbar!

`contour`

`contur` benötigt dieselbe Eingabe wie `persp` zeichnet aber einen 2D-Konturplot (d.h. Höhenlinien).

`image`

`image` benötigt dieselbe Eingabe wie `persp` zeichnet aber einen 2D-Plot mit Farbe als 3.Dimension.

Glatte 3D-Darstellung

Beispiel mit Tps und persp

```
#berechne Thin plate spline
x.tps = Tps(banknoteecht[c("oben","unten")],
  banknoteecht["laenge"])
#isoliere und ordne Koordinaten
ob = sort(unique(banknoteecht$oben))
ut = sort(unique(banknoteecht$unten))
#sage je Paar aus (ob × ut) die laenge vor.
x.pred =
   predict(x.tps,expand.grid(oben=ob,unten=ut))
#zeiche die Vorhersage perspektivisch
persp(ob, ut, x.pred)
```

Siehe auch Programm Npar_Banknote.R

Inhalt (1)

Inhalt (2)

Inhalt (3)

14. Hauptkomponentenanalyse

Problemstellung

- viele (hoch) korrelierte Variablen

$\rightarrow$ diese sollen ersetzt werden, durch neue, unkorrelierte Variablen, durch eine lineare Transformation

- **Ziel**: wenig neue Variablen,

die aber möglichst viel Information aus den Daten erhalten.

Daten: Punkte im p-dimensionalen Raum
Ziel: Projektion in einen p'-dimensionalen
(p' $\leq$ p) Teilraum mit möglichst viel erhaltener Information.

Hauptkomponenten_Venusmuscheln.R $(p = 2)$

Hauptkomponentenanalyse (2)

Annahmen

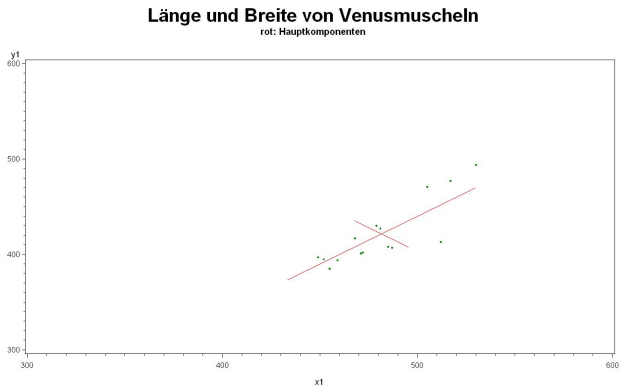
Daten sind Realisierungen eines p-variaten zufälligen Vektors $\mathbf{X} := (X_1, ..., X_p)$ mit
$E\mathbf{X} = 0$ und $\operatorname{var} \mathbf{X} = \mathbf{\Sigma} > 0$
(Kovarianzmatrix, positiv definit)

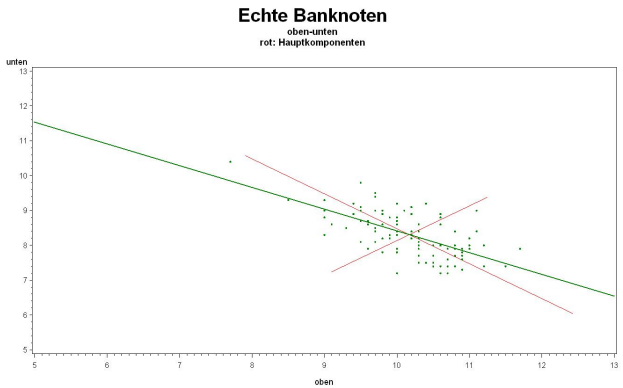
Bem: Die erste Bedingung erreicht man durch zentrieren um die Mittelwerte $\overline{X}_{.j}$, $j = 1, \ldots, p$

Wenn zwischen einzelnen Komponenten des zufälligen Vektors ein gewisser (etwa ein linearer) Zusammenhang besteht, so ist eine Dimensionsreduzierung möglich.
Der Zusammenhang wird durch Gerade dargestellt (ausgezeichnete Richtung in der Ebene).

Hauptkomponentenanalyse

Beispiele





Frage: Wie kann man diese ausgezeichnete Richtung erfassen?

Hauptkomponentenanalyse (3)

1. Hauptkomponente. Die Linearkombination

$$Y_1 = \sum_{j=1}^{p} b_{1j} X_j$$

ist so zu bestimmen, dass var $Y_1 \to$ max.
unter Normierungsbedingung $(\sum_j b_{1j}^2 = 1)$
(Die Variablen werden zentriert, $X_j' = X_j - \overline{X}_{.j}$)

2. Hauptkomponente. Die Linearkombination

$$Y_2 = \sum_{j=1}^{p} b_{2j} X_j$$

ist so zu bestimmen, dass var $Y_2 \to$ max,
unter Normierungsbedingung $(\sum_j b_{2j}^2 = 1)$
und unter der Bedingung $\text{cov}(Y_1, Y_2) = 0$

Hauptkomponentenanalyse (4)

Die Bedingung cov $(Y_1, Y_2) = 0$ sichert Unkorreliertheit der Hauptkomponenten.

Hauptkomponenten sind durch die Korrelationsmatrix eindeutig bestimmt.

Hauptachsentransformation: $\mathbf{\Sigma = U' \Lambda U}$

$\mathbf{\Sigma}$: (empir.) Korrelationsmatrix (bekannt)

$\mathbf{U}$: Orthogonalmatrix

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}$$

λ_i : Eigenwerte, sei $\lambda_1 \geq \dots \geq \lambda_p \geq 0$

Hauptkomponentenanalyse (5)

Hauptkomponenten

$\mathbf{Y} = \mathbf{U} \cdot \mathbf{X}$

Mahalanobis-Distanz eines Datenpunktes $\mathbf{X} = (X_1, \ldots, X_p)$ zum Ursprung:

$$\begin{aligned} \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} &= \mathbf{X}'\mathbf{U}'\boldsymbol{\Lambda}^{-1}\mathbf{U}\mathbf{X} = \mathbf{Y}'\boldsymbol{\Lambda}^{-1}\mathbf{Y} \\ &= \sum_{i=1}^{p} \frac{Y_i^2}{\lambda_i}. \end{aligned}$$

Die Konturen sind Ellipsoide.

Hauptkomponentenanalyse (6)

Hauptkomponentenanalyse in R

`prcomp`

`prcomp(data,tol)`
`data` ist Matrix oder data.frame.
Nur Hauptkomponenten deren Standardabweichung größer als `tol`$\sqrt{\text{var}Y_1}$ ist werden hinzugefügt.

Inhalt (1)

Inhalt (2)

Inhalt (3)

Zusammenfassung (1)

Basiswissen

- Klassifikation von Merkmalen
- Wahrscheinlichkeit
- Zufallsvariable
- Diskrete Zufallsvariablen (insbes. Binomial)
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert, Varianz
- Gesetz der großen Zahlen,
  Zentraler Grenzwertsatz

Zusammenfassung (2)

Beschreibende Statistik

(Robuste) Lage- und Skalenschätzungen

`summary, mean, median, winsor.mean, quantile, sd, IQR, mad, Sn, Qn` (u.a. Pakete psych und robustbase)

Boxplots

einfach: `boxplot(x)`
Formeln: `boxplot(m1 ~ gr1,data=dfr)`

Häufigkeitsdiagramme:

`hist(obj,breaks,freq,...)`

Scatterplots, Regressionsgerade:

`plot(x,y);abline(lm(x y))`

Zusammenfassung (3)

Statistische Tests

Testproblem: Nullhypothese - Alternative, z.B.

$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$

Entscheidung für H_0/gegen H_0: anhand einer

Teststatistik, z.B.

$$T = \frac{\overline{X} - \mu_0}{S} \cdot \sqrt{n}$$

Entscheidung

$|t| > t_{krit} \Rightarrow H_0$ ablehnen, $\quad P(|T| > t_{krit}) = \alpha$

α : Fehler 1. Art, Signifikanzniveau (in der Regel vorgegeben)

Zusammenfassung (4)

Statistische Tests (2)

p-Wert (zweiseitg)

$P(|T| > t)$, wobei t: Realisierung von T

p-Wert $< \alpha \Rightarrow H_0$ ablehnen

p-Wert $\geq \alpha \Rightarrow H_0$ nicht ablehnen

Gütefunktion

P(H_0 abgelehnt$|\mu$ richtig) = $\beta(\mu)$
Fehler 2.Art: $1 - \beta(\mu)$

Wir betrachten Tests mit einer vergleichsweise hohen Gütefunktion.

Zusammenfassung (5)

Einseitige Tests

Alternative geht in eine Richtung, (aus sachlichen Gründen kann es nur eine Richtung geben)

z.B. $\mu > \mu_0$

Zweiseitige Tests

Alternative geht in alle Richtungen,

z.B. $\mu \neq \mu_0$

Zusammenfassung (6)

Übersicht über Mittelwertvergleiche

k	unverbunden	verbunden
1	Einstichproben t-Test, Vorzeichen-Wilcoxon-Test	
	`t.test(x,mu), wilcox.test(x,mu)`	
2	t-Test	t-Test
	`t.test(x,y)`	`t.test(x,y,paired=TRUE)`
	Wilcoxon-Test	Vorzeichen-Wilcoxon-Test
	`wilcox.test(x,y)`	`wilcox.test(x,y,paired=T)`
$>$ 2	einfache Varianzana. = einfaktorielle VA	einfaches Blockexperiment = zweifaktorielle VA
	`anova(lm(x~y))`	`anova(lm(x~y+z))`
	Kruskal-Wallis-Test	Friedman-Test
	`kruskal.test(a~gr)`	`friedman.test(a~gr | bl)`

Zusammenfassung (7)

Anpassungstest auf Normalverteilung:

`shapiro.test(x)` oder `ad.test(x)` (Paket nortest)
Shapiro-Wilk-Test oder Anderson-Darling-Test

Anpassungstest auf Verteilung mit begrenzter Anzahl von Ausprägungen

`chisq.test(x,p)`
(`p` $= p_1, \ldots, p_k$ ggf. vorher ausrechnen)

Zusammenfassung (8)

Test auf Korrelation (metrisch oder ordinal skalierte Merkmale)

`cor.test(x,y,type="pearson")` bzw. `"spearmen"`/`"kendall"`

Test auf Unabhängigkeit (beliebig skalierte Merkmale):

`chisq.test(x,y)` = `chisq.test(table(x,y))`

Zusammenfassung (9)

Lineare Regression (1)

Parameterschätzung und Test

```
mod=lm(Y~ Var1+Var2+Var3...)
mod.sum = summary(mod)
```

Modellwahl

```
step(mod,direction)
leap(x,y,method)
```

Zusammenfassung (10)

Lineare Regression (2)

Residualanalyse

Plotten und Test auf Nomalverteilung:

```
plot(residuals(mod))
shapiro.test(residuals(mod))
points(rstudent(mod))
shapiro.test(rstudent(mod))
```

Zusammenfassung (11)

Sonstige Regressionsverfahren, nur Übersicht

Robuste Lineare Regression
Nichtlineare Regression
Nichtparametrische Regression
Logistische Regression

Zusammenfassung (12)

Hierarchische Clusteranalyse:

Standardisieren und Distanzmatrix:

`x.dist = dist(scale(x))`

`x.clust = hclust(x.dist,method)`

(`method="ward"`,`"single"`, `"complete"`,...)

`plot(x.clust)` Dendrogramm plotten

`cutree(x.clust,k` oder `h)`

Cluster der Elemente nach Höhe `h` oder Clusteranzahl `k`.

Zusammenfassung (13)

Konfidenzbereiche

für Parameter im Regressionsmodell

```
prd=predict(mod,interval="confidence")
confint(mod)
```

Grafische Darstellung von Konfidenzbereichen bei der Regression

```
plot(y)
```

Plotte untere und obere Grenzen:

```
lines(prd[,2],col="red")
lines(prd[,3],col="blue")
```

Zusammenfassung (14)

Wichtige Sprachelemente

Normalverteilte Zufallsvariable

mit festem Startwert `set.seed(`x_1`)`

`rnorm(k)`

k-Vektor, Komp. univariat normalverteilt

Gleichverteilte Zufallsvariable

`runif(k)`

sonstige Zufallsvariable

`r`Name der Verteilung

Zusammenfassung (15)

Wahrscheinlichkeitsverteilungen:

Verteilungsfunktion (Parameter)

`p`Verteilung`(q,Parameterliste)`

Dichte oder Wahrscheinlichkeitsfunktion (Parameter)

`d`Verteilung`(x,Parameterliste)`
z.B. `dnorm(x,0,1)`
`dbinom(x,n,p)`

Quantile

Standardnormal: `qnorm(u)` $u \in (0,1)$.
`q`Verteilung`(n,Parameterliste)`

Übungen (1)

1. Folgen und Reihen, Potenzreihen
2. Differential- und Integralrechnung, Normalverteilung
3. Integrralrechnung, Rechnen mit Erwartungswerten
4. Berechnen von Erwartungswerten, Berechnen von robusten Lage- und Skalenschätzungen
5. Berechnen von Korrelationen
6. Korrelationen, Einfluss von Ausreißern, Minima von Funktionen zweier Veränderlicher
7. Aufgabenblatt 7, Regressionsmodel, Berechnen von t-Teststatistiken
8. Aufgabenblatt 8, t-Test und Varianzanalyse

Übungen (2)

9. Aufgabenblatt 9,
   Produkt von Matrizen, Eigenwerte, Eigenvektoren
10. Aufgabenblatt 10,
    Lineare Algebra, Matrizenrechnung, χ^2-Verteilung
11. Aufgabenblatt 11
12. Aufgabenblatt 12

Übungsaufgaben

7,8,9 Wahrscheinlichkeitsverteilungen

10,11 Statist. Maßzahlen, Boxplots

11 Histogramme, Dichteschätzung

14,15,26,30,33,34,35 Korrelation, Unabhängigkeit, Lineare Regression

16-18,20-22,23-25 Lagetests, Anpassungstests

20,23 Varianzanalyse

27-29,31-32 Nichtparametrische Tests

36,37 Zufallszahlen

37 Clusteranalyse