Statistische Datenanalyse

Eine Einführung

Prof. em. Werner A. Stahel, ETH Zürich

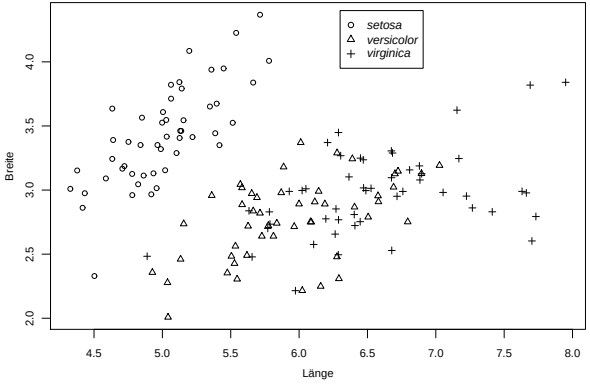
Kurs der DECHEMA, 5.-6. Nov. 2015

Folien

1. Beispiele, Fragestellungen

a Iris-Daten Der bekannteste Datensatz der (Multivariaten) Statistik.

!!! bild


setosa
versicolor
virginica
Breite
Länge
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

1

b Fragestellung: Kann man die Arten

auf Grund der Blüten- und Petalbrätter unterscheiden?

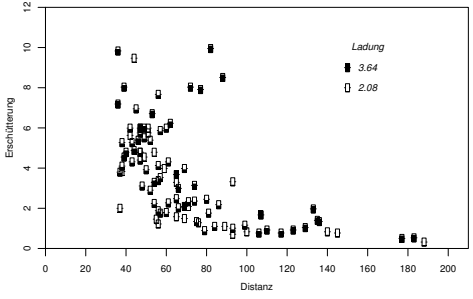
1

c Beispiel Sprengungen

Sprengungen für einen Autobahn-Tunnel unter Schaffhausen.

Erschütterung, gemessen in exponierten Häusern:

Abhängigkeit der Erschütterung von Distanz und Ladung.



1

d Fragestellung:

Wie hängt die Erschütterung von der Distanz und der Ladung ab?

$\longrightarrow$ Ladung anpassen, damit die Erschütterung nicht zu gross wird.

1

d Fragestellung:

Wie hängt die Erschütterung von der Distanz und der Ladung ab?

$\longrightarrow$ Ladung anpassen, damit die Erschütterung nicht zu gross wird.

e Regression ist die zentrale Methodik der Statistik!

2. Übersicht

Ziel: Methoden der statistischen Datenanalyse kennenlernen

... mit Hilfe der Statistik-Software R.

Nach Möglichkeit Anwendung auf eigene Daten. Wer hat solche?

Unterschiedliche Kenntnisse in Statistik und R

3. R-Einführung

a

- Weshalb R?
- Befehle statt Menue-Klicks
- R-Studio
- Wie R lernen?

3

b Statistische Daten

Beobachtungen, Variable $\longrightarrow$ **data.frame**

Import: **read.table("http://stat.ethz.ch/Teaching/Datasets/ WBL/spreng.dat"**

Besser: R-Studio starten, “Import Dataset”, Dateinamen eingeben.

Im linken oberen Feld “R Script” einstellen (Button linke obere Ecke)

Hier tippen Sie Befehle ein.

```
head(spreng)
```

liefert die ersten Zeilen des Datensatzes

c Elemente der Sprache R (ursprünglich S)

- "Vektoren", Skalare (Vektoren der Länge 1), Matrizen, Arrays
- Modes: **numeric, character, logical, complex, ...**
- Listen
- Funktionen
- **class** $\longrightarrow$ "objektorientierte" Sprache
- **data.frame** = Liste der Klasse **"data.frame"** ,
  geeignet für statistische Datensätze:
  Nominale (kategorielle) Variable als **factor** s gespeichert

d Grundoperationen

- Elemente auswählen

  spreng[3,1] 3. Zeile, 1. Spalte

  spreng[,"dist"] Spalte (Variable) “dist”

  (Variable haben Namen, Zeilen auch,

  Elemente von Vektoren oder Listen können Namen haben)

- Häufigste Aktion: Zuweisung

  d <- spreng[,"dist"]

  re <- range(spreng[,"ersch"])

  rd <- range(d)

- “Alles” ist Funktion!

  Hier: **range** liefert Minimum und Maximum.

- Ohne Zuweisung wird **print** ausgeführt

  = Resultat auf dem Schirm angezeigt.

  range(d) oder **print(range(d))**

  wählt “die” geeignete Form der Ausgabe

  (reagiert auf **class** dessen, was ausgegeben werden soll,

  “generische Funktion”.)

e Funktionen

- Einfache statistische und mathematische Funktionen: **mean, median, min, max, range, log, sqrt, ...**
- Kompliziertere Funktionen produzieren Listen:

  dc <- hist(spreng[,"ersch"], plot=FALSE) klassiert Daten

  str(dc) Was für ein Objekt ist das?

```
> dc <- hist(spreng[,"ersch"], plot=FALSE)
> str(dc)
List of 6
 $ breaks  : num [1:8] 0 2 4 6 8 10 12 14
 $ counts  : int [1:7] 32 25 26 6 6 0 1
 $ density : num [1:7] 0.1667 0.1302 0.1354 0.0312 0.0312 .
 $ mids    : num [1:7] 1 3 5 7 9 11 13
 $ xname   : chr "spreng[, \"ersch\"]"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

f Listen fassen beliebige Objekte zusammen.

“Komponenten” auswählen: $ oder **dc[["counts"]]**

```
> dc$counts
[1] 32 25 26  6  6  0  1
```

Mehrere Komponenten $\longrightarrow$ (Teil-) Liste: **dc[1:3]**

3

g Generische Funktionen

print, summary, plot reagieren auf die **class**

des 1. Arguments $\longrightarrow$ rufen die entsprechende Methode auf.

$\longrightarrow$ Flexibilität des objektorientierten Programmierens,

manchmal schwierig zu durchschauen

z.B. liefert **print** nicht immer alle Komponenten einer Liste.

h Grafik 1

- Histogramm: **hist(spreng[,"ersch"])**
- **plot(ersch ∼ dist, data=spreng)**

  Funktion **plot** , erstes Argument enthält ∼

  $\longrightarrow$ vom Typ **formula** ,

  später gebraucht für (Regressions-) Modell.

  (**plot(spreng[,"dist"],spreng[,"ersch"])** geht auch.)

i Funktions-Aufrufe

- **hist** hat viele Argumente, die das Aussehen steuern

  Diese werden mit **Argumentname = Wert** gesetzt:

  hist(spreng[,"dist"], nclass=15, main="Distanz")

- Namen der Argumente und default-Werte erhält man über

  ?hist oder **help(hist)**

  – oder sie werden beim Tippen des Namens angezeigt

  (R-Studio u.a.)