SS07 WeF Zusammenfassung

Oswald Berthold

April 11, 2008

Contents

Index[1] - SS07_WeF[2]

1 Allgemeines

Dies ist meine prüfungsvorbereitende Zusammenfassung der Vorlesung "Werkzeuge der empirischen Forschung" von PD Dr. Wolfgang Kössler im SS07. Seitenzahlen beziehen sich auf die Folien zur Vorlesung WeF SS06 von W. Kössler, Institut f. Informatik der Humboldt Universität zu Berlin.

2 Basiskommandos in R

2.1 Dateien lesen / schreiben

```
## Datei ansehen
file.show('pfad/zur/datei.dat')
## datei als zeilenvektor einlesen, danach 'reformatieren'
scan('pfad/zur/datei.dat')
## falls es das format erlaubt gleich als datenframe einlesen
read.table('pfad/zur/datei')
```

2.2 Ausgabe

```
cat
print
dev.print(postscript, file='...')
```

2.3 Matrix und data.frame Manipulation

```
## basics
A <- matrix(seq(1, 9), 3, 3)
B <- matrix(seq(1, 9), 3, 3, byrow=T)
## dimensionen eines arrays / einer matrix
dim(A)
## indizierung
A[1]
A[1,]
A[,1]
## filtern
A[1,c(T, F, F)]
A[1,A[1,]>2]
## subset
subset(dataframe, select=c(spalte1,spalte2,...))
subset(dataframe, select=-c(spalte1,spalte2,...))
```

[1] index.pdf
[2] SS07_WeF.pdf

```
## operationen
apply(daten, spaltenweise(1)/zeilenweise(2), function)
colSums(A)
colMeans(A)
rowSums(B)
rowMeans(B)
## sortieren
order
sort
rank
[]
## zusammenfuegen
merge
cbind
rbind
## kontingenz-tafel / contingency oder frequency table
table(x)
## fuer stetige ZVs empfiehlt sich vorher ein
cut(x, breaks(anzahl der klasseneinteilungen))
## datenframe umbauen, auch neue spalten koennen angegeben werden
transform(df, spaltename=transformationsfunktion, spltn=tf, ...)
## faktor / gruppeneinteilung
f <- factor(x)
levels(f)
## generate levels
gl()
## differenz zweier zahlen
diff(c(a, b))
## stichprobenfunktion
sample(daten, laenge, replace=F)
```

2.4 Grafik

```
## scatterplot
plot(...)
## adding stuff
points
abline(from, to)
abline(v=0)
abline(h=0)
lines(x, y, ...)
spline(...)
smooth.spline(...)
legend(x, y, text, pch
text(x, y, ...
boxplot(vector1, vec2, vec3, ..., parms, names=c(...), boxwex
```

```
## boxplot formelinterface
boxplot(wert ~ gruppe, ...
## multiple plots in einem fenster
layout(matrix(1:6, 3, 2)) ## 6 felder, 3 zeilen, 2 spalten
par(mfrow=c(nr, nc)) ## aequivalent
## stamm-blatt diagramm
stem(x)
## histogram
hist(x, anzahl-bins, .., plot=T|F)
## qq-plot
qqnorm(x)
qqline()
```

2.5 Regression

```
lm(y ~ x, data=...)
glm
nls(y ~ a + b * x + c * x^2 ..., data=, start=list(a=1,b=1,...))
coef
formula()
residuals()
## um model fit zu plotten
predict(model)
## nur predict.lm() kann konfidenz-intervall und vorhersage-intervall berechnen
predict(y~x, interval='c')
```

2.6 Hypothesentests

```
## t-Test: 1-sample,
t.test()
library(BSDA); sign.test()
wilcox.test()
## Varianzvergleich
var.test()
levene.test()
## korrelation / unabhängigkeit
cor.test()
chisq.test()
summary(table(...)) ## fuehrt chi-quadrat test auf unabhaengigkeit aus
## mehrstichproben mittelwertsvergleich
aov(wert ~ gruppe, data=dataframe) ## mehrere stichproben in einem vektor,
                                   ## mit gruppenvariable gekennzeichnet
anova(lm(wert ~ gruppe, data=dataframe)) ## wie aov
## Anpassungstest
```

2.7 Diverses

```
ecdf(daten) ## empirsche verteilungsfunktion
plot(ecdf(daten), verticals=T, do.p=F) ## schoenere darstellung
```

3 Beschreibende Statistik

3.1 Schätzungen

3.1.1 Eigenschaften von Schätzungen $\hat{\theta}$

Ab S. 119
Sei $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ eine Schätzung des Parameters θ die auf n Beobachten beruht.

- $\hat{\theta}_n \rightarrow_{n\rightarrow\infty} \theta$, Konsistenz, Minimalforderung
- $E\hat{\theta}_n = \theta$, Erwartungstreue bzw. asymptotische Erwartungstreue
- $\mathrm{var}\hat{\theta}_n$ möglichst klein: "gute", "effiziente" Schätzung
- wenn $\mathrm{var}\hat{\theta}_n$ kleinstmöglich dann ist $\hat{\theta}_n$ "optimale" Schätzung
- MSE (mean squared error) soll minimal sein: $MSE = \mathrm{var}\hat{\theta}_n + bias^2\hat{\theta}_n = \mathrm{var}\hat{\theta}_n + (E\hat{\theta}_n - \theta)^2$
- Eigenschaften sollen auch bei Abweichungen von der Normalvtlg. gelten: robuste Schätzung.

Cramer-Rao Schranke, Fisher-Information S. 121–128
Sei $\hat{\theta} = \theta_n$ eine erwartungstreue Schätzung von θ.
Dann gilt die Cramer-Rao Ungleichung:

$$var(\hat{\theta}) \geq \frac{1}{nI(f,\theta)'}$$

mit

$$I(f,\theta) = E\left(\frac{\partial lnf(x,\theta)}{\partial\theta}\right)^2$$

die *Fisher-Information*.

3.1.2 Schätzmethoden:

- Momentemethode: wahre Momente werden durch empirische Momente ersetzt.
- Maximum-Likelihood Methode: Aufstellen der Likelihood-Funktion $L(X_1, ..., X_n, a, b, ...)$ als gemeinsame Dichte der Stichprobe $X = (X_1, ..., X_n)$. Dann wird $logL(...)$ maximiert.
- Kleinste-Quadrat-Schätzung: ...

3.1.3 Lageschätzungen

Ab S. 111

1. Mittelwert:
$$\overline{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$$
$$\overline{X} \to_{n\to\infty} EX \quad \text{Gesetz der grossen Zahlen}$$
Unter der Voraussetzung dass der Erwartungswert existiert. Der Mittelwert ist meist ein "gute" Schätzung aber nicht robust.

2. Quantile: α-Quantil x_α: mindestens $\alpha \cdot n$ der Werte $(x_1, ..., x_n)$ sind kleiner oder gleich x_α, mindestens $(1-\alpha) \cdot n$ der Werte $(x_1, ..., x_n)$ sind grösser oder gleich x_α.

3. Median: ist das 0.5-Quantil $x_{0.5}$. Der Median ist robust aber meist nicht so "gut".

4. Quartile: heissen die 0.25- und 0.75-Quantile $x_{0.25}$ und $x_{0.75}$

5. Modalwert: Häufigster Wert falls diskrete ZV, Wert mit grösster Dichte falls stetige ZV.

6. Getrimmtes Mittel: (Ausreisserschutz) Sei $\alpha \in [0, \frac{1}{2})$. $\overline{X}_\alpha := \frac{X_{(\lfloor n\cdot\alpha\rfloor+1)}+...+X_{(n-\lfloor n\cdot\alpha\rfloor)}}{n-2\lfloor n\cdot\alpha\rfloor}$

7. Winsorisiertes Mittel: Sei $\alpha \in [0, \frac{1}{2}$ und $n_1 := \lfloor n \cdot \alpha \rfloor + 1$. Dann heisst
$$\overline{X}_{\alpha,\omega} := \frac{n_1 X_{(n_1)} + X_{(n_1+1)} + ... + X_{(n-n_1)} + n_1 X_{(n-n_1+1)}}{n}$$
α-winsorisiertes Mittel. Die jeweils $\lfloor n\cdot\alpha\rfloor$ kleinsten und grössten Werte werden "herangeschoben" und dann das arithmetische Mittel gebildet. $\alpha : 0.1, ..., 0.2$.

3.1.4 Skalenschätzungen / Streuungsmasse

Ab S. 129

1. Varianz:
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$
$$s^2 \to_{n\to\infty} var(X). \text{ Division durch } n-1\text{: Erwartungstreue}$$

2. Standardabweichung:
$$s = \sqrt{s^2}$$

3. Spannweite / Range:
$$Range = X_{(n)} - X_{(1)}$$

4. (Inter-)Quartilsabstand:
$$IR = s_F = x_{0.75} - x_{0.25}$$

5. Mittlere absolute Abweichung vom Median:

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

6. Median aboslute Abweichung vom Median (MAD):

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

7. Variationskoeffizient:

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

8. Gini's Mean Difference:

$$G = \frac{1}{\binom{n}{2}}\sum_{i<j}|x_i - x_j|$$

   G ist mässig robust aber effizient.

9.

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

Bei $X \sim N \Rightarrow$ Skalierungsfaktoren fuer IR, MAD, G nach sigma.

- Varianz, Standardabweichung und Spannweite sind nicht "robust".
- Quartilsabstand und MAD sind robust, MAD etwas besser.
- G ist bedingt robust, effizient bei F normal.
- MAD ist wenig effizient
- S_n oder Q_n sind am geeignetsten.

3.1.5 Formmasse

- Schiefe (skewness): Theoretische Schiefe

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

  Empirische Schiefe

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

  β_1: gleich 0 F symmetrisch, kleiner 0 linksschief, grösser 0 rechtsschief

- Wölbung (kurtosis) Theoretische Wölbung

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

  Empirische Wölbung

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

3.2 Boxplots

Schematische und skeletale Boxplots. Von oben nach unten:
Schematisch:

- Ausreisser
- $x_{0.75} + 1.5 \cdot IR$
- $x_{0.75}$
- Empirisches Mittel (+)
- $x_{0.5}$
- $x_{0.25}$
- $x_{0.25} - 1.5 \cdot IR$

Skeletal:

- Max
- Wie schematisch
- Min

3.2.1 R

```
boxplot(x, y, args...)
```

3.3 Stamm- und Blatt Diagramme (Stemplot)

Ähnlich dem Histogramm

1. 1. Stelle in 10er Darstellung: Stamm
2. 2. Stelle in 10er Darstellung: Blätter, Ziffer wird explizit angegeben, Verfeinerung des Balkendiagramms

3.3.1 R

```
stem(runif(10))
```

3.4 Q-Q-Plot

Es werden die Quantile der Normalverteilung und die Quantile der empirischen Verteilung, also die Punkte $(\Phi^{-1}(\alpha), x_\alpha)$ gegeneinander geplottet, falls $F \sim N(\mu, \sigma$ sollten die Punkte in etwa auf einer Geraden liegen.

3.4.1 R

```
x <- rnorm(100)
qqnorm(x)
qqline(x)
```

3.5 Häufigkeitstabellen

Chart, eigentlich wie Histogramm. Vertikal und horizontal.

3.5.1 Eindimensionale Zufallsvariable

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

Die p_i sind zu schätzen:

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

mit N Stichprobenumfang, n_i: relative Häufigkeiten.

3.5.2 Zweidimensionale Zufallsvariablen

Das Paar (X, Y) heisst zweidimensionale ZV. Seien X und Y diskret und (x_i, y_i) die möglichen Ergebnisse von (X, Y).

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

$i = 1, ..., M, j = 1, ..., N$ heisst gemeinsame Wahrscheinlichkeitsfunktion von (X, Y).

Eigenschaften

$$p_{ij} \geq 0$$

$$\sum_{i,j} p_{ij} = 1$$

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

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

Z.B. Rauchverhalten 0,1 und Geschlecht m,w. Die Tabelle der Häufigkeiten heisst Kontingenztafel.
Def.: X und Y heissen unabhängig genau dann wenn

$$p_{ij} = p_{i.} \cdot p_{.j}$$

(strip.chart, barchart)

3.5.3 R

```
x <- rnorm(100)
hist(x)
x <- floor(runif(100, 0, 3))
y <- floor(runif(100, 0, 3))
table(cbind(x,y))
```

3.6 Histogramme und Dichteschätzung

Histogramm oder auch Zähldichte.
Überlagerung des Histogramms mit einer glatten Dichtefunktion: Dichteschätzung mittels Kernfunktion.
Seien $x_1, ..., x_n$ die Beobachtungen und sei $K(t)$ eine sogenannte Kernfunktion sowie

$$\begin{array}{ll} \int K(t)dt = 1 & \int tK(t)dt = 0 \\ \int t^2K(t)dt = 1 & \int K^2(t)dt < \infty \end{array}$$

und h ein sogen. Glättungsparameter, dann heisst

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

Dichteschätzung oder Dichtefunktionsschätzung.

3.6.1 R

```
x -> rnorm(100)
hist(x, freq=F)
lines(density(x))
```

3.7 Zusammenhangsmasse

3.7.1 Scatterplots

Zweidimensionale Stichproben können als Punkte in der Ebene (Punktwolke) dargestellt werden: Scatterplot.

3.7.2 Kovarianz, Korrelation, Korrelationskoeffizienten

Seinen X, Y Zufallsvariablen.
Varianz: $var(X) = E(X - EX)^2 = E[(X - EX)(X - EX)] = EX^2 - E^2X$
Definition: Kovarianz

$$Cov(X, Y) := E[(X - EX)(Y - EY)] = XXX$$

Definition: Korrelation

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

Empirische Varianz:

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

Empirische Kovarianz:

$$s_{XY}^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Normierung: s_X, s_Y: empirische Standardabw. von X, Y.

1. Pearsonscher (empirischer) Korrelationskoeffizient:

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

   - Es gilt: $-1 \leq r_{XY} \leq 1$
   - Der Korrelationskoeefizient ist invariant gegenüber linearen Transformationen: $x \rightarrow a + bx$
   - $|r_{XY}| = 1$ gdw. alle Punkte auf einer Geraden liegen, $y = mx + b, m \neq 0$
   - Korrelationskoeffizient ist ein Mass für die ineare Abhängigkeit von X und Y.
   - $r_{XY} = 0$ heisst: keine lineare Abhängigkeit, andere Abhängigkeiten sind aber durchaus möglich.

2. Spearman Rangkorrelationskoeffizient

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

   R_i =Rang von X_i in der geordneten Stichprobe

$$X_{(1)} \leq ... \leq X_{(n)}$$

   S_i =Rang von Y_i in der geordneten Stichprobe

$$Y_{(1)} \leq ... \leq Y_{(n)}$$

   r_S ist auch für ordinale Merkmale geeignet, die $X_1, ..., X_n; Y_1, ..., Y_n$ werden sersetzt durch Rangzahlen

$$X_i \rightarrow R_{X_i} = R_i$$

$$Y_i \to R_{Y_i} = S_i$$

Dann ist $R_{X_i} = 1$ falls $X_i = min(X_i, ..., X_n)$,

$$R_{X_i} = n \text{ falls } X_i = max(X_i, ..., X_n)$$

3. Kendalls Konkordanzkoeffizient $(X_i, Y_i), i = 1, ..., n$

$$a_{ij} = \begin{cases} 1 & \text{falls } x_i < x_j \wedge y_i < y_j \text{ oder } x_i > x_j \wedge y_i > y_j \\ -1 & \text{falls } x_i < x_j \wedge y_i > y_j \text{ oder } x_i > x_j \wedge y_i < y_j \\ 0 & \text{sonst.} \end{cases} = sgn[(X_i - X_j)(Y_i - Y_j)]$$

Falls

$$\begin{array}{cc} a_{ij} = 1 & \text{Paar heisst konkordant.} \\ a_{ij} = -1 & \text{Paar heisst diskordant.} \\ a_{ij} = 0 & \text{Paar heisst gebunden.} \end{array}$$

Kendalls Konkordanzkoeffizient τ:

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

Vergleich Pearson - Spearman
Vorteile Spearman

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

Nachteile Spearman

- wenn kardinales (stetiges) Messniveau: Informationsverlust
- schwierige Interpretation wenn r_S nicht nahe -1,0,1 (gilt eingeschränkt auch für Pearson).

3.8 Regressionsproblem

S. 200, ...
Seien X, Y Zufallsvariablen (entsprechend höherdimensionaler Fall).
Ein Modell ist:

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

Dabei gibt es folgenden Fälle für f:

- linear, bekannte Form, suchen nur Parameter
- nonlinear, bekannte Form, suchen nur Parameter
- unbekannt, nichtparametrische Regression

Für f bekannt: Minimieren den quadratischen Erwartungswert des Fehlers zwischen Y und der Vorhersage.

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

Die $\theta_1, ..., \theta_p$ sind unbekannt und werden anhand der Beobachten X_i, Y_i geschätzt mit LSE:

$$min_{\theta_1,...,\theta_p} \frac{1}{n} \sum_{i=1}^{n} (Y - f(X, \theta_1, ..., \theta_p))^2$$

Lösung des Minimu-Problems durch Ableiten und Nullsetzen des obigen Ausdrucks, führt auf GS mit p Gleichungen.
Lineare Regression führt auf Polynom, sonst auch nichtlineare Basisfunktionen (ln, Exponentialfkt., ...)
Für f unbekannt: z.B. kubischern Spline, Kernschätzung.

3.9 Zusammenfassung

Siehe Folien: S. 213 - 217

4 Schliessende Statistik

4.1 Hypothesentests

Es werden 2 Hypothesen aufgestellt bzgl. der Parameter eines Problems.

$$\begin{array}{rlll} \text{Einseitige Alternative:} & H_0: & \mu \leq \mu_0 & H_A \quad \mu > \mu_0 \\ \text{Einseitige Alternative:} & H_0: & \mu \geq \mu_0 & H_A \quad \mu < \mu_0 \\ \text{Zweiseitige Alternative:} & H_0: & \mu = \mu_0 & H_A \quad \mu \neq \mu_0 \end{array}$$

Teststatistik:

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

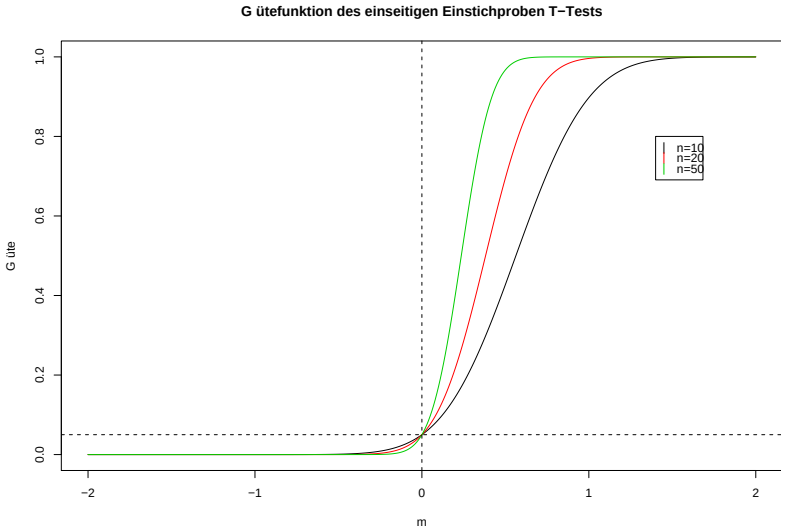
Die Teststatistik geht gegen null für n gegen unendlich wegen des Gesetzes der grossen Zahlen: $\overline{X} \longrightarrow_{n \to \infty} EX$ (mit n -> unendl. geht der empirische Mittelwert gegen den wahren Mittelwert).
Die Entscheidung für H_0 oder H_A wird anhand der Teststatistik gefällt. Zeigt der Wert von T in einen vorher bestimmten Bereich, den kritischen oder Ablehnungs-bereich wird H_0 abgelehnt. Sonst wird H_0 beibehalten. Die Testgrösse T ist t-verteilt mit $n-1$ Freiheitsgraden wobei die n die Stichprobengrösse ist.
Bei dieser Entscheidung kann man Fehlentscheidungen treffen:

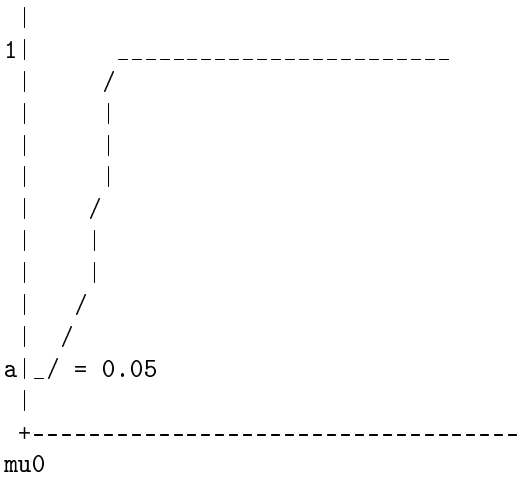
- Entscheidung für H_A obwohl H_0 richtig ist: Fehler 1. Art
- Entscheidung für H_0 obwohl H_A richtig ist: Fehler 2. Art

	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. $1-\beta$	richtig, Güte β

I.a. wird α festgelegt und β maximiert.
$\beta(\theta)$ heisst Gütefunktion

4.2 Gütefunktion





4.3 T-Test

4.3.1 Einstichprobenproblem

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

Teststatistik:

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

R

```
t.test(rnorm(10, 0, 1), mu=0, alternative='less')
```

Dabei ist meist der Fehler 1. Art α z.B. $\alpha = 0.05, \alpha = 0.01$ d.h. $P_{\mu_0}(|T| > t_{krit}) = \alpha$. α heisst Signifikanzniveau.
$\overline{T}$ ist eine Zufallsgrösse und besitzt eine best. Wahrscheinlichkeitsverteilung, in diesem Fall eine t-Verteilung (Student's t), genauer $T \sim t_{n-1}$.
p-Wert: Die Grösse $P(|T| > t)$ heisst p-Wert (p-value). Wenn also $p \geq \alpha$ so H_0 angenommen, sonst abgelehnt. Andere Interpretation: p-Wert ist Wahrscheinlichkeit der beobachteten Daten wenn H_0 richtig ist.
Dichtefunktion einer t-Verteilung mit $\nu = n - 1$ Freiheitsgraden:

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

Verteilungsfunktion:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Konfidenzintervalle Das Intervall

$$\overline{X} - \frac{s}{\sqrt{n}} \cdot t_{t-\frac{\alpha}{2}, n-1}, \overline{X} + \frac{s}{\sqrt{n}} \cdot t_{t-\frac{\alpha}{2}, n-1}$$

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

4.3.2 Zweistichprobenproblem

$$H_0: \quad \mu_1 = \mu_2 \quad H_1: \quad \begin{array}{l} \mu_1 \neq \mu_2 \\ \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \end{array}$$

Vergleich zweier abhängiger Gruppen Beispiele:

- Gewicht einer Person zu den Zeitpunkten t_1, t_2
- Banknoten: oben - unten, links - rechts

R

```
t.test(x, y, paired=T, alternative=c('two.sided', 'less', 'greater'))
```

Vergleich zweier unabhängiger Gruppen Seien $X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$ Es werden 2 Fälle unterschieden:

1. Die Varianzen sind gleich.
2. Die Varianzen sind verschieden oder unbekannt

R

```
t.test(x, y, var.equal=T)
```

4.3.3 Skalentests

F-Test zum Vergleich zweier Varianzen

$$F = \frac{S_1^2}{S_2^2} \sim F_{n-1,m-1}$$

F-Verteilung mit $(n-1, m-1)$ Freiheitsgraden (Fisher-Verteilung).
F ist der Quotient zweier unabhängiger χ^2-verteilter Grössen.

Robuste Skalentests Besser Skalentests: Levene-Test, Brown-Forsythe-Test: Bilden neue ZV durch Betrag der Differenz der ZV und dem Mittelwert bzw. dem Median. Diese neue Zufallsvariable wird t-Test unterzogen. Lässt sich erweitern auf den Vergleich der Varianzen von k Stichproben.

4.3.4 R

```
var.test(x,y)
library(car)
levene.test(x,y)
```

4.3.5 Voraussetzungen für die Anwendung des T-Tests

- Normalverteilung
- Varianzhomogenität

Ist das Verhältnis der Varianzen bekannt (gleich, ungleich)?
Es kann ein Test auf gleiche Varianzen vorgeschaltet werden: F-Test
Aber: -stufiger Test ist problematische bezüglich des Signifikanzniveaus.

- F-Test (zum Skalenvergleich) ist nicht robust.
- Einstichproben t-Test ist nicht robust.
- Zweistichproben t-Test etwas robuster.
- Ausreisserempfindlichkeit
- Wenn Varianz unklar /unbekannt: unterschiedliche Varianzen annehmen.

4.4 Varianzanalyse

4.4.1 Vergleich von k unabhängigen Gruppen

A: Faktor (Gruppenvariable) Y: anhängiges Merkmal /Responsevariable
Modell: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}\ i = 1...k, j = 1...n_i$

$$\begin{aligned} H_0: &\quad \alpha_1 = \alpha_2 = ... = \alpha_k \\ H_1: &\quad \alpha_i \neq \alpha_l (\text{ für ein } i \neq l) \end{aligned}$$

Streuungszerlegung: Gesamtvarianz (SST) ist Varianz zwischen den Gruppen (SSB) plus Varianz innerhalb der Faktorstufen (SSW) + Fehler (SSE).
$F = \frac{MSB}{MSE}$, also mittlere Varianz zw. den Gruppen durch mittl. Varianz innerhalb der Gruppen.
Bestimmtheitsmass: R^2 := SSB/SST

- Der F-Test in der Varianzanalyse ist robust.
- Verlangt aber auch Varianzhomogenität, jedoch sind Abweichungen nicht so schwerwiegend
- Bei versch. Varianzen: Welch-Modifikation
- Gibt auch hier Test auf Varianzhomogenität: Levene, BF

R XXX

```
anova.lm
anova.glm
```

In R gibt es diese HOV Tests

```
bartlett.test(stats)    Bartlett Test of Homogeneity of Variances
fligner.test(stats)     Fligner-Killeen Test of Homogeneity of
```

4.4.2 Vergleich von k verbundenen Stichproben

Varianz innerhalb der Gruppen, Varianz ziwschen den Gruppen: Quadratsummen Chisq-verteilt und unabhaengig
$F = c * SSB/SST$ (summed square between / summed square total)
$= \frac{\chi^2_{k-1}}{\chi^2_{N-k}}$, N ist gesamtstichprobenumfang
-> aov, anova

4.5 Anpassungstests

Ab S. 294
Klassische Test- und Schätzverfahren sind meist unter der Normalverteilungsannahme konzipiert. Gilt diese überhaupt?
Sei $(X_1, ..., X_n)$ eine unabhängige Stichprobe, $X_i \sim F, F$ unbekannt.

1. Anpassungstest auf eine spezifizierte Verteilung:

$H_0 : F = F_0 \quad$ gegen $H_1 : F \neq F_0$, wobei F i.A. von unbekannten Parametern abhängt.

2. Anpassungstest auf Normalverteilung:

$$\begin{array}{lrcl} H_0: & F(x) & = & \Phi(\frac{x-\mu}{\sigma})(\mu, \sigma \text{ unbekannt}) \\ H_1: & F(x) & \neq & \Phi(\frac{x-\mu}{\sigma})\forall \mu, \sigma, \sigma > 0 \end{array}$$

4.5.1 Auf der empirischen Verteilungsfunktion beruhende Tests

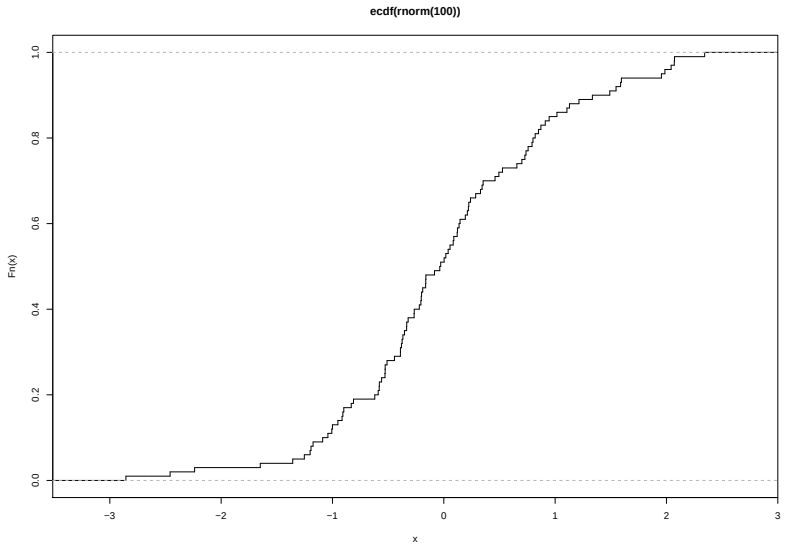
Seien $X_{(1)} \leq ... \leq X_{(n)}$ die geordneten Beobachtungen. Die Funktion

$$F_n(x) = \begin{cases} 0 & x, X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x < X_{(i+1)} \quad i = 1...n \\ 1 & X_{(n)} \leq x \end{cases}$$

heisst empirische Verteilungsfunktion.
Satz von Glivenko-Cantelli: $F_n(x) \to F(x)$, Hauptsatz der mathematischen Statistik.

R

```
e <- ecdf(rnorm(100))
plot(e, verticals=T, do.points=F)
```



1. **Kolmogorov-Smirnov-Test**

$$\begin{array}{rcl} D & = & \sup_x |F_n(x) - F_0(x)| \\ & = & \max\left(\max_i(\frac{i}{n} - U_{(i)}), \max_i(U_{(i)} - \frac{i-1}{n})\right) \end{array}$$

2. **Cramer-von-Mises-Test**

$$\begin{aligned} W - sq &= n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dF_0(x) \\ &= \frac{1}{12n} + \sum_{i=1}^{n} (U_{(i)} - \frac{2i-1}{2n})^2 \end{aligned}$$

3. **Anderson-Darling-Test**

$$\begin{aligned} A - sq &= n \int_{-\infty}^{\infty} \underbrace{\frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))}} dF_0(x) \\ &= -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)(\ln U_{(i)} + \ln(1 - U_{n+1-i})) \end{aligned}$$

$U_{(i)} = F_0(X_{(i)}), \qquad X_{(1)} \leq ... \leq X_{(n)}$
$D \sim D_n$ ist Kolmogorov-verteilt. D wird approximiert. Für endliche Stichproben werden Modifikationen verwendet (S. 303)

R

```
library(nortest)
x <- rnorm(100) ##
ks.test(x)      ## einstichproben test default, normalverteilung
ks.test(x, 'pnorm') ## einstichproben test mit gegebener verteilungsfunktion
ks.test(x, y)   ## zweistichproben test: entstammen beide stichproben einer
                ## population mit gemeinsamer Verteilung?
## siehe auch: lillie.test
###
cvm.test(x)     ## aus nortest
ad.test(x)      ## aus nortest
shapiro.test(x) ## aus stat
sf.test(x)      ## aus nortest
```

4.5.2 Shapiro-Wilk-Test

Ab. S.304, Sachs S.341
Kurze Version: Die Teststatistik $\hat{W}$ ist der Quotient aus zwei Schätzungen für σ^2: das Quadrat einer kleinsten Fehlerquadratschätzung für die Steigung einer Regressionsgeraden im QQ-Plot und die Stichprobenvarianz. Im Fall einer Normalverteilung sollte der Quotient nahe bei 1 liegen.

$$\hat{W} = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)} \right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$x_{(i)}$ sind die geordneten Beobachtungen, a_i sind konstante Werte (Tabelle).
Lange Version: XXX

- Shapiro-Wilk-Test hat (für kleine, mittlere und grössere Stichprobenumfänge) die höchste Güte der angeführten Tests.
- Früher meist verwendet: χ^2-Anpassungstest. Hat geringe Güte.
- W ist etwas besser als $A - sq$, besser als $W - sq$ und viel besser als D und χ^2
- D erst ab Stichprobenumfängen $n \geq 2000$ zu empfehlen.
- Signifikanzniveau sollte auf $\alpha = 0.01$ hochgesetzt werden, besonders bei weniger robusten Tests.

4.5.3 χ^2-Anpassungstest (Pearson)

Daten werden in p Klassen eingeteilt.

- Klassenhäufigkeiten: N_i
- theoretische Häufigkeiten: np_i

$$X^2 = \sum_{i=1}^{p} \frac{(N_i - np_i)^2}{np_i}$$

Dann ist

- $X^2 \sim \chi^2_{p-1}$ asymptotisch verteilt (bei bekannten μ, σ^2), (Fisher 1922)
- $X^2 \sim \chi^2_{p-3}$ approximativ (bei zu schätzenden Parametern, ML-Schäztung mit gruppierten Daten oder Minimum χ^2-Schätzung).

Nachteile:

- Wert von X^2 abhängig von Klasseinteilung
- Geringe Güte

R

```
x <- rnorm(100)
chisq.test(x)
## oder
erbsen <- c(rep(1, 315), rep(2, 108), rep(3, 101), rep(4, 32));
chisq.test(table(erbsen)[], p=c(9,3,3,1), rescale.p=T)
## ...
```

4.6 Nicht-parametrische Tests

Analoga zu bereits behandelten parametrischen Tests.

1. Einstichprobenproblem, verbundene Stichproben: Vorzeichen-Test (Sign-Test), Vorzeichen-Wilcoxon-Test (Signed-Rank-Test)
2. Zwei unverbundene Stichproben: Wilcoxon-Test
3. Mehrere unabhängige Stichproben: Kruskal-Wallis-Test
4. Mehrere verbundene Stichproben: Friedman-Test

- Wenn keine Normalverteilung vorliegt: Verwendung von nicht-parametrischen Tests.
- Diese verwenden keine Parameterschätzungen
- Halten das Signifikanzniveau (α) für jede stetige Verteilung ein, sind also unabhängig von der zugrundeliegenden Verteilung.
- relativ effizient
- Annahme: stetige Verteilungsfunktion

4.6.1 Einstichprobenproblem

Hypothesen wie bei t-Test.
Es werden die Differenzen $X_i - \mu_0$ gebildet.

$$V_i := \begin{cases} 1 & \text{falls} X_i - \mu_0 > 0 \\ 0 & \text{falls} X_i - \mu_0 < 0 \end{cases}$$

Vorzeichentest $V^+ = \sum_{i=1}^{n} V_i =$# der Differenzen mit positiven Vorzeichen.
Der Fall $X_i - \mu_0 = 0$ kommt wegen Stetigkeit der Vtlgs.fkt. nur mit Wkt. 0 vor, falls doch wird die Beobachtung als Messungenauigkeit verworfen. Nachteil: Gerade Beobachtungen die für H_0 sprechen werden nicht berücksichtigt.
Es gilt: $V^+ \sim Bi(n, \frac{1}{2})$
Kritische Werte können daher leicht bestimmt werden:

$$Bi(1-\alpha, n, 1/2)$$

R

```
library(BSDA)
x <- rnorm(100)
y <- rnorm(100, 1, 1)
sign.test(x)
sign.test(y, md=1)
```

Wilcoxon-Vorzeichen-Test Bilden neue Beobachtungen $D_i = |X_i - \mu_0|$, zu diesen dann die Rangzahlen, d.h. den Rang in der geordneten Stichprobe:

$$\begin{array}{ccccccc} D_{(1)} & \leq & ... & & ... & \leq & D_{(n)} \\ \downarrow & & & & & & \uparrow \\ \text{Rang } 1 & & & & & & \text{Rang } n \\ R_{(1)} = 1 & & & & & & R_{(n)} = n \end{array}$$

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 $X_i - \mu_0 > 0$
Berechnen $E_0 W_n^+$ und $var(W_n^+)$.
Es gilt: $W_n^+ \sim N(EW_n^+, var(W_n^+))$ asympt.
XXX: Siehe Test_IQ_Daten.sas

R

```
x <- rexp(100)
wilcox.test(x, mu=0)
```

4.6.2 Zweistichprobenproblem

Zwei verbundene Stichproben Bilden $Z := X - Y$ und testen wie gehabt z.B.
$H_0 : \mu_Z = 0$
$H_A : \mu_Z \neq 0$
Beispiele: Banknote, Darwin

R

```
x <- rexp(100)
y <- rexp(100)
wilcox.test(x, y, paired=T)
```

Zwei unverbundene Stichproben Hypothesen wie gehabt: $H_0 : \mu_1 = \mu_2$ resp. $\leq, \geq$
Die Beobachtungen $X_{11}, ..., X_{1n}, X_{21}, ..., X_{2n}$ werden zu einer Stichprobe zusammengefasst und den Elementen dieser Rangzahlen zugeordnet:

$$z_{(1)} \leq ... \leq z_{(n+m)}$$

Seien nun R_{ij} die Rangzahlen zu x_{ij}, wobei $i = 1, 2; j = 1, ...n$, dann ist

$$S = \sum_{j=1}^{n} R_{ij} = \text{Summe der Ränge die zur ersten Stichprobe gehören.}$$

Unter H_0 gilt:

$$Z = \frac{S - ES}{\sqrt{var(S)}} \text{ ist näherungsweise } N(0,1) \text{ verteilt.}$$

R

```
x <- rexp(100)
y <- rexp(100)
wilcox.test(x, y)
```

4.6.3 Mehrere Stichproben

Unverbunden Modell: $Y_{ij} = \mu_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$, $\quad j = 1, ..., n, \quad i = 1, ..., k$
$H_0 : \mu_1 = ... = \mu_k$
$H_A : \exists(\mu_{i_1}, \mu_{i_2}) \quad \mu_{i_1} \neq \mu_{i_2}$
Wir fassen alle Beobachtungen

$$X_{11}, ..., X_{1n_1}, ..., X_{k1}, ..., X_{kn_k}$$

zusammen und bilden die Rangzahlen $R_{ij}, \quad i = 1...k, j = 1, ..., n_i$
Mit den Rangzahlen führen wir eine einfaktorielle VA durch: Kruskal-Wallis-Test

$$KW = \frac{\sum_{i=1}^{k} (T_i - E_0(T_i))^2 \cdot n_i}{S^2}$$

mit

$$T_i = \frac{1}{n_i} \sum j = 1 n_i R_{ij}$$

die mittlere Rangsumme der i-ten Gruppe (vgl. $\overline{Y}_i$ aus der VA).
$E_0 T_i = \frac{N+1}{2}$
$N = \sum_{i=1}^{k} n_i$ Gesamtstichprobenumfang
$S^2 = \frac{N \cdot (N+1)}{12} \cdot (N-1)$
$KW \sim \chi^2_{k-1}$ (asympt.)
H_0 ablehnen falls p-Wert $< \alpha$.

R

```
x <- rexp(100)
y <- rexp(100)
z <- rexp(100)
kruskal.test(x, y, z)
```

- Bei Bindungen erfolgt Korrektur, Mittel der Rangzahlen
- relativ effizient

Verbundene Stichproben Friedman-Test: Modell: $Y_{ij} = \mu + \alpha_i + \beta_i + \epsilon_{ij}$
$\epsilon_{ij} \sim N(0, \sigma^2), \quad j = 1...k, \quad i = 1...n$
$H_0 : \beta_1 = ... = \beta_k (= 0)$
$H_A : \exists (j_1, j_2) : \beta_{j1} \neq \beta_{j2}$
Ränge werden zeilenweise gebildet:

$$Y_{1(1)} \leq ... \leq Y_{1(k)}, \quad R_{1(1)} = 1, ...$$

R_{ij} der Rang von Y_{ij} in der i-ten Zeile.
Tabelle: S.340

$$F_k = \frac{n^2 \sum_{j=1}^{k} (\overline{R}_{.j} - E(\overline{R}_{.j}))^2}{n \cdot k(k+1)/12}$$

$\overline{R}_{.j} = \frac{1}{n} \sum_{i=1}^{n} R_{ij}$ Spaltenmittel der j-ten Spalte
$E\overline{R}_{.j} = \frac{1}{n} \frac{n(k+1)}{2} = \frac{k+1}{2}$
Unter $H_0 : F_k \sim \chi^2_{k-1}$ (asympt.), H_0 ablehnen falls $F_k > \chi^2_{1-\alpha,k-1}$ bzw. p-value $< \alpha$.

R

```
x <- rexp(100)
y <- rexp(100)
z <- rexp(100)
friedman.test(x, y, z)
```

4.7 Korrelation und Unabhängigkeit

Die Zufallsvariablen $X_1, ..., X_N$ heissen unabhängig falls für alle $x_1, ..., x_N \in R$ gilt:

$$P(X_1 < x_1, ..., X_N < x_N) = P(X_1 < x_1) \cdot ... \cdot P(X_N < x_N)$$

Die Zufallsvariablen $X_1, ..., X_N$ heissen unkorreliert falls:

$$E(X_1 \cdot \cdot \cdot X_N) = E(X_1) \cdot cdotcdotE(X_N)$$

Aus Unabhängigkeit folgt Unkorreliertheit aber die Umkehrung gilt nicht.
Aus $X_i \sim N(\mu, \sigma^2)$ folgt Unabhängigkeit und Unkorreliertheit.

4.7.1 Korrelationstest

Zwei Fälle

1. Stetige (metrische) Merkmale Mit r_{XY} Pearsonscher Korrelationskoeffizient ist

$$T = \sqrt{N-2} \cdot \frac{r_{XY}}{\sqrt{1 - r^2_{XY}}} \sim t_{N-2}$$

   Also t-Test anwendbar.

2. Ordinal oder Nominal skalierte Merkmale Z.B. Länge - Breite, Geschlecht - Studienfach, Studiengang - Note, Geburtsmontat - IQ $H_0 : p_{ij} = p_{i.} \cdot p_{.j}, \quad i = 1, ..., m; j = 1, ..., l \; H_A : p_{ij} \neq p_{i.} \cdot p_{.j}$ für ein Paar (i, j). Also $H_0 : X, Y$ sind unabhängig. Berechnen einer Teststatistik Q_P die χ^2-verteilt ist mit $(m-1) \cdot (l-1)$ Freiheitsgraden. Das ist der χ^2-Unabhängigkeitstest.

rangtest: spearman, kendall
Autokorrelationstest Durbin-Watson

4.8 Test auf Unabhängigkeit

χ^2-Unabhängigkeitstest, siehe oben.
Phi-Koeffizient
Run-Test

4.9 Lineare Regression

4.9.1 Einfache Lineare Regression

$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i \qquad \epsilon_i \sim (0, \sigma^2)$, auch oft β statt θ.

$$\hat{\theta}_1 = \frac{S_{XY}}{S_X^2}$$

$$\hat{\theta}_0 = \frac{1}{n}(\sum Y_i - \hat{\theta}_1 \sum X_i)$$

Lösung einer Minimumsaufgabe (S. 358)

4.9.2 Multiple Lineare Regression

S. 359

Modell:

4.9.3 Residualanalyse

Residuen auf NV testen. S. 370

R

```
skull <- read.table('../daten/skull.dat');
names(skull) <- c('Group','MB','BH','BL','NH')
attach(skull)
## verwende multiple lineare regression, modellgleichung:
f <- MB ~ BL + BH + NH

model <- lm(f, subset=Group==1)

plot(residuals(model)); abline(0, 0, col=2, lty=2);
```

4.10 Zufallszahlen

Erzeugung von gleichverteilten Zufallsvariablen.
Beliebige Verteilungen:

- stetig: anwenden der Quantilfunktion $F^{-1}(U_I)$ auf eine gleichverteilte ZV.
- diskret: zerteilen des Intervalls $(0, 1)$ in entsprechende grosse Teile. Länge des Intervalls ist die Wahrscheinlichkeit. Je nach Wert von U_i entsprechend zugeordnet diskreter Wert.

Siehe S. 400 sowie [projects/uni[3]/informatik/SfI/Pruefung/pruefung.vorbereitung.tex]

4.11 Clusteranalyse

Ziel: Zusammenfassen von "ähnlichen" Objekten zu Gruppen (Clustern). Unähnliche Objekte sollen in verschiedene Cluster. Cluster sind vorher nicht bekannt.
Es gibt zu unterscheiden:

- partionierende Clusteranalyse: Zahl der Cluster ist vorgegeben
- hierarchische Clusteranalyse
- Fuzzy Clusteranalyse: Zugehörigkeit eines Datenpunktes zu einem Cluster als Fuzzy-Wert

Für "Ähnlichkeit" eintscheidend: Definition des Abstandsmasses, wobei p : # Merkmale

- Euklidischer Abstand
$$d_E^2(x, y) = \sum_{i=1}^{p} (x_i - y_i)^2$$
- City-Block / Manhattan-Abstand
$$d_C(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$
- Tschebyscheff-Abstand
$$d_T(x, y) = \frac{max}{i} |x_i - y_i|$$

Agglomerative Verfahren: Jede Beobachtung ist ein Cluster, dann immer die zwei ähnlichsten Cluster zusammenfassen, bis es nurmehr ein Cluster gibt.
Die Methoden unterscheiden sich durch die Definitionen der Abstände $D(C_i, C_j)$ zwischen Clustern C_i und C_j.

- Single Linkage
$$D_S(C_i, C_j) = min\{d(k, l), k \in C_i, l \in C_j\}$$

[3] index.pdf

- Complete Linkage
$$D_C(C_i, C_j) = max\{d(k,l), k \in C_i, l \in C_j\}$$
- Centroid

  $D_{CE}(C_i, C_j) = d(\overline{X_i}, \overline{Y_i})$, Abstände der Schwerpunkte
- Average Linkage
$$D_A(C_i, C_j) = \frac{1}{n_i n_j} \sum_{k \in C_i, l \in C_j} d(k,l)$$
- Ward: Anova Abstände innerhalb der CLuster minimieren, nach Umrechnen erhält man
$$D_W(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} D_{CE}(C_i, C_j)$$

4.11.1 R

```
df <- data.frame(X=c(1,2,3), Y=c(4,5,6), method='euclidean')
## dist
hc <- hclust(dist(df), method='complete')
memb <- cutree(hc, k=3)
df[memb=1:3]
### oder
agnes(df)
```

5 TODO

- ein- und zwei-faktorielle Varianzanalyse
- nicht-parametrische Tests
- Konfidenzbereiche
- ein paar praktische Experimente: check uebungen, check astrostat tutorials, check gulli etc data

6 Literatur

1. W. Kössler: Folien zur Vorlesung "Werkzeuge der empirischen Forschung" SS06, SS07, `http://www2.informatik.hu-berlin.de/~koessler/`
2. L. Sachs: Angewandte Statistik mit Beispiel in R, Springer 2006
3. `http://astrostatistics.psu.edu/su07/R/`
4. R Online Documentation