

Zusammenfassung

Basiswissen

- Klassifikation von Merkmalen
- Wahrscheinlichkeit
- Zufallsvariable
- Diskrete Zufallsvariablen (insbes. Binomial)
- Stetige Zufallsvariablen
- Normalverteilung
- Erwartungswert, Varianz
- Gesetz der großen Zahlen,
Zentraler Grenzwertsatz

Beschreibende Statistik

(Robuste) Lage- und Skalenschätzungen

```
PROC UNIVARIATE TRIMMED=Zahl  
ROBUSTSCALE; RUN;
```

Boxplots

```
PROC BOXPLOT; PLOT Variable*Faktor  
/BOXSTYLE=SCHEMATIC; RUN;
```

Häufigkeitsdiagramme:

```
PATTERN1 ...;  
PROC GCHART; VBAR Variable; RUN;
```

Scatterplots, Regressionsgerade:

```
SYMBOL1 ...;  
PROC GPLOT; PLOT y*x=1 / REGEQN; RUN;
```

Zusammenfassung Statistische Tests und Multivariate Verfahren

Testproblem: Nullhypothese - Alternative, z.B.

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

Entscheidung für H_0 /gegen H_0 : anhand einer

Teststatistik, z.B.

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n}$$

Entscheidung

$$|t| > t_{krit} \Rightarrow H_0 \text{ ablehnen, } P(|T| > t_{krit}) = \alpha$$

α : Fehler 1. Art, Signifikanzniveau (in der Regel vorgegeben)

Zusammenfassung Statistische Tests (2)

p-Wert (zweiseitig)

$P(|T| > t)$, wobei t : Realisierung von T

p-Wert $< \alpha \Rightarrow H_0$ ablehnen

p-Wert $\geq \alpha \Rightarrow H_0$ nicht ablehnen

Gütefunktion

$P(H_0 \text{ abgelehnt} | \mu \text{ richtig}) = \beta(\mu)$

Fehler 2. Art: $1 - \beta(\mu)$

Wir betrachten Tests mit einer vergleichsweise hohen Gütefunktion.

Zusammenfassung Statistische Tests (3)

Einseitige Tests

Alternative geht in eine Richtung, (aus sachlichen Gründen kann es nur eine Richtung geben)

$$\text{z.B. } \mu > \mu_0$$

Zweiseitige Tests

Alternative geht in alle Richtungen,

$$\text{z.B. } \mu \neq \mu_0$$

Übersicht über Mittelwertvergleiche (1)

k	unverbunden	verbunden
1	Einstichproben t-Test, Vorzeichen-Wilcoxon-Test	
	PROC UNIVARIATE; o. PROC TTEST H0=Wert; VAR Variable; RUN	
2	t-Test	t-Test
	PROC TTEST; CLASS=Faktor; VAR Variable; RUN;	PROC TTEST; PAIRED Var1*Var2; RUN;
	Wilcoxon-Test	Vorzeichen-Wilcoxon-Test
	PROC NPAR1WAY WILCOXON; CLASS=Faktor; VAR Variable; RUN;	diff=a-b; PROC UNIVARIATE; VAR diff; RUN;

Übersicht über Mittelwertvergleiche (2)

einfache Varianzana. = einfaktorielle VA	einfaches Blockexperiment = zweifaktorielle VA
PROC ANOVA; CLASS Faktor; MODEL Y=Faktor; RUN; (PROC GLM)	PROC GLM; CLASS FaktorA FaktorB; MODEL Y=FaktorA FaktorB; RUN;
Kruskal-Wallis-Test	Friedman-Test
PROC NPAR1WAY Wilcoxon; CLASS Faktor; VAR var; RUN;	PROC FREQ; TABLES FaktorA*FaktorB*Y / CMH2 SCORES=RANK NOPRINT; RUN;

Anpassungstest auf Normalverteilung:

```
PROC UNIVARIATE NORMAL; VAR var; RUN;
```

Shapiro-Wilk-Test oder Anderson-Darling-Test

Anpassungstest auf Verteilung mit begrenzter Anzahl von Ausprägungen

```
PROC FREQ; TABLES Var1 /CHISQ NOPRINT  
          TESTP=(p1,p2,...pk); RUN;
```

(p_1, \dots, p_k) vorher ausrechnen)

Test auf Korrelation (metrisch oder ordinal skalierte Merkmale)

```
PROC CORR PEARSON SPEARMAN KENDALL; RUN;
```

Test auf Unabhängigkeit (beliebig skalierte Merkmale):

```
PROC FREQ;
```

```
TABLES Var1*Var2 /CHISQ NOPRINT; RUN;
```

Lineare Regression (1)

Parameterschätzung und Test

```
PROC REG;  
  MODEL Y=Var1 Var2 ... Varn / CLI CLM R;  
  TEST Var1=0 Var2=0; /*Zusaetzl.Hypothesen */  
RUN;
```

Modellwahl

```
PROC REG;  
  MODEL Y=Var1 Var2 ... Varn /  
  SELECTION=backward; RUN;
```

Lineare Regression (2)

Residualanalyse

```
PROC REG;  
  MODEL Y=Var1 Var2 ... Varn / R;  
  PLOT rstudent.*obs.; /*und/oder*/  
  PLOT residual.*y; residual.*predicted.;  
RUN;
```

und evtl. Test auf Normalverteilung.

Sonstige Regressionsverfahren, nur Übersicht

Robuste Lineare Regression

Nichtlineare Regression

Nichtparametrische Regression

Logistische Regression

Hierarchische Clusteranalyse:

```
PROC CLUSTER
  METHOD=Average
      (oder: CENTROID oder WARD)
  OUTTREE=baum; VAR Variablen; RUN;
PROC TREE DATA=baum
  NCLUSTERS = Anzahl der Cluster
      fuer GPLOT;
  OUT=Eingabedatei fuer Proc GPLOT;
RUN;
PROC GPLOT; PLOT VarA*VarB=cluster; RUN;
```

Konfidenzbereiche

für Parameter im Regressionsmodell

```
PROC REG;  
    MODEL Y=var1...varn/ CLI CLM;  
RUN;
```

Grafische Darstellung von Konfidenzbereichen bei der Regression

```
SYMBOL1 I=RLCLI95;  
PROC GPLOT; PLOT y*x=1; RUN;
```

Wichtige Sprachelemente

Normalverteilte Zufallsvariable

mit zufälligem Startwert: `seed=-1; RANNOR(seed);`

Gleichverteilte Zufallsvariable

mit zufälligem Startwert: `seed=-1; RANUNI(seed);`

Wahrscheinlichkeitsverteilungen:

Verteilungsfunktion (Parameter)

`CDF('Verteilung', z, Parameterliste)`

Dichte oder Wahrscheinlichkeitsfunktion (Parameter)

`PDF('Verteilung', z, Parameterliste)`

z.B.: ('normal', z, 0, 1)

('binomial', z, n, p)

Quantile

Standardnormal: `PROBIT(u)`, $u \in (0, 1)$.

`Quantile('Verteilung', z, Parameterliste)`

Übungen (1)

1. Folgen und Reihen, Potenzreihen
2. Differential- und Integralrechnung, Normalverteilung
3. Integralrechnung, 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, Regressionsmodell, 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
- 11h-14,29,32,33,34 Korrelation, Unabhängigkeit, Lineare Regression
- 15-21,23-25 Lagetests, Anpassungstests
- 19,22 Varianzanalyse
- 26-28,30-31 Nichtparametrische Tests
- 35,36 Zufallszahlen
 - 36 Clusteranalyse