

III. Schließende Statistik 2

1. Korrelation und Unabhängigkeit
2. Lineare Regression
3. Nichtlineare Regression
4. Nichtparametrische Regression
5. Logistische Regression
6. Zufallszahlen
7. Clusteranalyse
8. Hauptkomponentenanalyse
9. Faktorenanalyse
10. Diskriminanzanalyse

III. Schließende Statistik II

Anmerkung: In der Vorlesung werden viele Details weggelassen, diese sind dann auch nicht prüfungsrelevant.

1. Korrelation und Unabhängigkeit

Def.: Die Zufallsvariablen X_1, \dots, X_N heißen unabhängig, falls für alle $x_1, \dots, x_N \in \mathbb{R}$

$$P(X_1 < x_1, \dots, X_N < x_N) = P(X_1 < x_1) \cdots P(X_N < x_N)$$

Def.: Die Zufallsvariablen X_1, \dots, X_N heißen unkorreliert, falls

$$\mathbf{E}(X_1 \cdots X_N) = \mathbf{E}(X_1) \cdots \mathbf{E}(X_N).$$

Bem.:

Unabhängigkeit \Rightarrow Unkorreliertheit

\nLeftarrow

$X_i \sim \text{Normal} \Rightarrow \text{Unabh.} = \text{Unkorr.}$

Fall a) Stetige (metrische) Merkmale

Seien (X_i, Y_i) , $i = 1, \dots, N$ unabhängige bivariate Zufallsvariablen.

Pearson-Korrelation

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Es gilt:

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

wird in SAS zur Berechnung der p-Werte verwendet.

Weitere Korrelationskoeffizienten:

Spearman, Kendall

wenn keine NV so diese nehmen!

`Descr_Scatter.sas`

`Descr_Scatter_Heroin.sas`

a) Metrisch skalierte merkmale

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

b) Ordinal oder nominal skalierte Merkmale

```
PROC FREQ;  
  TABLES var1*var2 / CHISQ;  
RUN;
```

Fall b) Ordinal oder nominal skalierte Merkmale

Bsp. Geschlecht - Studienfach

Studiengang - Note

Geburtsmonat - IQ

Frage: Bestehen Abhängigkeiten?

Antwort: χ^2 - Unabhängigkeitstest (Pearson, 1908)

Annahme:

X hat Ausprägungen a_1, \dots, a_m

Y hat Ausprägungen b_1, \dots, b_l

(sind die Daten metrisch, so wird automatisch eine Klasseneinteilung vorgenommen.)

$$P(X = a_i) = p_i. \quad P(Y = b_j) = p_{.j}$$

$$P(X = a_i, Y = b_j) = p_{ij}$$

Häufigkeitstabelle (= Kontingenztafel)

$X Y$	b_1	b_2	\dots	b_j	\dots	b_l	
a_1	h_{11}	h_{12}	\dots	h_{1j}	\dots	h_{1l}	$h_{1.}$
a_2	h_{21}	h_{22}	\dots	h_{2j}	\dots	h_{2l}	$h_{2.}$
\dots							
a_i	h_{i1}	h_{i2}	\dots	h_{ij}	\dots	h_{iN}	$h_{i.}$
\dots							
a_m	h_{m1}	h_{m2}	\dots	h_{mj}	\dots	h_{ml}	$h_{m.}$
	$h_{.1}$	$h_{.2}$	\dots	$h_{.j}$	\dots	$h_{.l}$	$h_{..}=N$

h_{ij} : Häufigkeiten,

werden verglichen mit den exakten Häufigkeiten np_{ij} .

$$H_0 : p_{ij} = p_{i.} \cdot p_{.j}, \quad i = 1, \dots, m, j = 1, \dots, l$$

$$H_1 : p_{ij} \neq p_{i.} \cdot p_{.j}, \quad \text{für ein Paar}(i, j)$$

H_0 : X und Y sind unabhängig.

Stichprobenfunktion

$$\tilde{T} = \sum_i \sum_j \frac{(h_{ij} - Np_{ij})^2}{Np_{ij}}$$

Unter H_0 : $p_{ij} = p_{i.} \cdot p_{.j}$,

aber $p_{i.}$ und $p_{.j}$ sind unbekannt. Sie müssen also geschätzt werden,

das sind $m + l - 2$ Parameter ($\sum p_{i.} = \sum p_{.j} = 1$)

$$\hat{p}_{i.} = \frac{h_{i.}}{N} \quad \hat{p}_{.j} = \frac{h_{.j}}{N}$$

$$h_{i.} = \sum_{j=1}^l h_{ij} \quad h_{.j} = \sum_{i=1}^m h_{ij}$$

Einsetzen der Schätzungen in \tilde{T} (unter H_0)

$$\begin{aligned}
 Q_P &= \sum_i \sum_j \frac{(h_{ij} - N\hat{p}_{i.}\hat{p}_{.j})^2}{N\hat{p}_{i.}\hat{p}_{.j}} \\
 &= N \sum_i \sum_j \frac{(h_{ij} - \frac{h_{i.}h_{.j}}{N})^2}{h_{i.}h_{.j}} \\
 &\sim \chi_{(m-1)(l-1)}^2 \quad \text{approx. unter } H_0
 \end{aligned}$$

Die Anzahl der Freiheitsgrade ergibt sich aus:

$$m \cdot l - 1 - \underbrace{(m + l - 2)}_{\text{\#geschätzte Werte}}$$

H_0 ablehnen, falls

$$Q_P > \chi_{(m-1)(l-1)}^2,$$

bzw. falls p-Wert $< \alpha$.

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

Weitere Unabhängigkeitstests

- LQ- χ^2 - Unabhängigkeitstest

$$G^2 = 2 \sum_i \sum_j h_{ij} \ln \frac{h_{ij}}{h_{i.}h_{.j}} \sim \chi_{(m-1)(l-1)}^2$$

- Continuity Adjusted χ^2 (bei SAS nur: 2x2-Tafel)

$$Q_c = N \sum_i \sum_j \frac{\max(0, |h_{ij} - \frac{h_{i.}h_{.j}}{N}| - 0.5)^2}{h_{i.}h_{.j}} \sim \chi_{(m-1)(l-1)}^2$$

- Mantel-Haenszel (r_{XY} : Pearson-Korrelation)

$$Q_{MH} = (N - 1)r_{XY}^2 \sim \chi_1^2$$

- Phi-Koeffizient

$$\Phi = \begin{cases} \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1.}h_{2.}h_{.1}h_{.2}}} & m = l = 2 \\ \sqrt{Q_p/N} & \text{sonst} \end{cases}$$

- Kontingenzkoeffizient

$$P = \sqrt{\frac{Q_P}{Q_P + N}}$$

- Fishers Exact Test (bei 2x2-Tafeln)
 durch Auszählen aller Tafel-Möglichkeiten bei
gegebenen Rändern.
 (gilt als etwas konservativ.)

- Cramers V

$$V = \begin{cases} \Phi & \text{falls } 2 \times 2 \text{ Tafel} \\ \sqrt{\frac{Q_P/N}{\min(m-1, l-1)}} & \text{sonst} \end{cases}$$

Bem.

- Mantel- Haenszel Test verlangt ordinale Skalierung,
 vgl. $(N - 1)r_{XY}^2$
 ‘gut’ gegen lineare Abhängigkeit.
- Der χ^2 Unabhängigkeitstest testet gegen allg. Unabhängigkeit.
- Der LQ-Test G^2 ist plausibel und geeignet.
- Der LQ-Test G^2 und der χ^2 Unabhängigkeitstest sind asymptotisch äquivalent.

Φ-Koeffizient (2x2 Tafel)

Y \ X	Raucher	Nichtraucher	Summe
w	p_{11}	p_{12}	$p_{1.}$
m	p_{21}	p_{22}	$p_{2.}$
Summe	$p_{.1}$	$p_{.2}$	1

$$X \sim Bi(1, p_{.2}) \quad Y \sim Bi(1, p_{2.})$$

$$\mathbf{E}(X) = p_{.2} \quad \text{var}(X) = p_{.2}(1 - p_{.2}) = p_{.2}p_{.1}$$

$$\mathbf{E}(Y) = p_{2.} \quad \text{var}(Y) = p_{2.}(1 - p_{2.}) = p_{2.}p_{1.}$$

$$\text{cov}(X, Y) = \mathbf{E}(X \cdot Y) - \mathbf{E}(X)\mathbf{E}(Y) = p_{22} - p_{.2}p_{2.}$$

Korrelationskoeffizient:

$$\rho = \frac{p_{22} - p_{.2}p_{2.}}{\sqrt{p_{.2}p_{.1}p_{2.}p_{1.}}} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{.2}p_{2.}p_{.1}p_{1.}}}$$

$$\begin{aligned} p_{22} - p_{.2}p_{2.} &= p_{22} - (p_{21} + p_{22})(p_{12} + p_{22}) \\ &= p_{22} - (p_{21}p_{12} + p_{22}p_{12} + p_{21}p_{22} + p_{22}^2) \\ &= p_{22}(1 - p_{12} - p_{21} - p_{22}) - p_{21}p_{12} \\ &= p_{22}p_{11} - p_{21}p_{12} \end{aligned}$$

Für $m = l = 2$ ist der Phi-Koeffizient eine Schätzung des Korrelationskoeffizienten.

Der Run-Test

Es seien nun X_1, \dots, X_n identisch verteilte Zufallsgrößen.

H_0 : X_1, \dots, X_n sind unabhängig

H_1 : X_1, \dots, X_n sind abhängig

Def. Jeder Teilabschnitt einer Folge unabhängiger, identisch verteilter Zufallsgrößen, in dem die Zufallsgrößen in aufsteigend geordnet sind, heißt Run. Bsp.:

Wir teilen eine Folge in Runs ein:

Folge von Zufallsgrößen	2 1 2 3 2 4 1 7 8 9 0									
Run	I.	II.	III.	IV.	V.					
Länge des Runs	1	3	2	4	1					

Satz Es sei X_1, \dots, X_n eine Folge identisch verteilter unabhängiger Zufallsgrößen mit Dann gilt für die zufällige Länge \underline{R} eines Runs:

$$P(\underline{R} = r) = \frac{r}{(r + 1)!}.$$

$$\underline{R} : \begin{pmatrix} 1 & 2 & \dots & r & \dots \\ \frac{1}{2} & \frac{1}{3} & \dots & \frac{r}{(r+1)!} & \dots \end{pmatrix}.$$

Bem. 2 *Es gilt:*

$$\begin{aligned} \sum_{i=1}^{\infty} P(\underline{R} = i) &= \sum_{i=1}^{\infty} \frac{i}{(i+1)!} \\ &= \sum_{i=1}^{\infty} \left(\frac{1}{i!} - \frac{1}{(i+1)!} \right) \\ &= \sum_{i=1}^{\infty} \frac{1}{i!} - \sum_{i=1}^{\infty} \frac{1}{(i+1)!} \\ &= \left(\sum_{i=0}^{\infty} \frac{1}{i!} - 1 \right) - \left(\sum_{i=0}^{\infty} \frac{1}{(i+1)!} - 1 \right) \\ &= (e - 1) - (e - 1 - 1) = 1. \end{aligned}$$

Beobachtungen: R_1, \dots, R_m sei die Folge der Längen der auftretenden Runs.

Run-Test: χ^2 -Anpassungstest auf diese Verteilung

SAS bietet den Test in dieser Form nicht an, er ist jedoch selbst zu programmieren:

```
data Wkt/*Berechnung der Wktn. und der Runs*/
  DO r=1 to 10;
    p_r=Gamma(r+2); /*Berechnung der Wktn.*/
  END;
run;
data rundata;
  set SAS-datei;
  /*Berechnung der Runs R: UEA 10 P.*/
run;
proc freq data=rundata;
  tables R /chisq testp=(p_1 p_2 ... p_r);
  /*p_r einsetzen*/
run;
```