6. Multivariate Verfahren
Übersicht

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.

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 SAS zur Berechnung der p-Werte verwendet.

Weitere Korrelationskoeffizienten:
Spearman, Kendall
wenn keine NV vorliegt, so diese nehmen!

Korrelation und Unabhängigkeit

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;
```

`Descr_Scatter.sas`
`Descr_Scatter_Heroin.sas`

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)

$X \mid Y$	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 SAS nur: 2x2-Tafel)

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

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

$$Q_{MH} = (N-1) r^2_{XY} \sim \chi^2_1$$

- 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 (bei 2x2-Tafeln)

durch Auszählen aller Tafel-Möglichkeiten bei gegebenen Rändern.
(gilt als etwas konservativ.)

- Cramers V

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

Weitere Unabhängigkeitstests (2)

Anmerkungen

- 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.

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.