# 9 Signifikanztests Teil III

**Zoltán Zomotor**

**Versionsstand: 12. April 2016, 11:29**

**Die nummerierten Felder bitte während der Vorlesung ausfüllen.**



Bitte hier notieren, was beim Bearbeiten unklar geblieben ist:

## Inhaltsverzeichnis



## 1 Kontingenztest[1]

- Gegeben: Zwei verbundene einfache Stichproben

1 $X_1, \dots, X_n$ und $Y_1, \dots, Y_n$

- Hypothesenpaar:

$H_0$: Die Merkmale $X$ und $Y$ sind 2 in $G$ unabhängig

[1] Bedeutung von „Kontingenz“ (duden.de): Häufigkeit bzw. Grad der Wahrscheinlichkeit des gemeinsamen Auftretens zweier Sachverhalte, Merkmale usw.

$H_1$: Die Merkmale $X$ und $Y$ sind [3] in $G$ abhängig

- Vorgehensweise entspricht dem [4] $\chi^2$-Anpassungstest
- basiert aber auf [5] einer Kontingenztabelle
- Vorgehen:

Schritt 1: Ein Signifikanzniveau [6] $\alpha$ festlegen.

Schritt 2: Den Testfunktionswert folgendermaßen ermitteln:

2.1 Die $x$-Achse in $k \geq 2$ und die $y$-Achse in $\ell \geq 2$ disjunkte, aneinander angrenzende Intervalle [7] $A_1, \dots, A_k$ bzw. $B_1, \dots, B_\ell$ unterteilen.

2.2 Eine Kontingenztabelle mit Randhäufigkeiten erstellen:

| | | $Y$ | | | | |
|---|---|---|---|---|---|---|
| | | $B_1$ | $B_2$ | $\cdots$ | $B_\ell$ | |
| $X$ | $A_1$ | $h_{11}$ | $h_{12}$ | $\cdots$ | $h_{1\ell}$ | $h_{A_1}$ |
| | $A_2$ | $h_{21}$ | $h_{22}$ | $\cdots$ | $h_{2\ell}$ | $h_{A_2}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| | $A_k$ | $h_{k1}$ | $h_{k2}$ | $\cdots$ | $h_{k\ell}$ | $h_{A_k}$ |
| | | $h_{B_1}$ | $h_{B_1}$ | $\cdots$ | $h_{B_\ell}$ | [8] $n$ |

Dabei bezeichnen [9] $h_{ij}$, $h_{A_i}$ und $h_{B_j}$ die Anzahl der beobachteten Paare [10] $(x, y)$ in $A_i \times B_j$, $A_i$ und $B_j$ .

2.3 Zu jeder Kombination aus $i = 1, \dots k$ und $j = 1, \dots, \ell$ die Größe [11] $\tilde{h}_{ij} = h_{A_i} \cdot h_{B_j}/n$ berechnen.

2.4 Den Testfunktionswert $v$ folgendermaßen berechnen:

[12]

$$v = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{h_{ij}^2}{\tilde{h}_{ij}} - n$$

Schritt 3: Mit dem Fraktilswert $x_{1-\alpha}$ der 13 $\chi^2((k-1)\cdot(\ell-1))$ -Verteilung den Verwerfungsbereich 14 $B = (x_{1-\alpha}, \infty)$ festlegen.

Schritt 4: $H_0$ genau dann ablehnen, wenn 15 $v \in B$ gilt.

- Bemerkungen:
  - (i) Test ist nur anwendbar, wenn
    16 $h_{ij} \geq 5$ bzw. $\tilde{h}_{ij} \geq 5$ für alle $i, j$

    Falls nicht erfüllt $\Rightarrow$ 17 Intervalle zusammenlegen!
  - (ii) Falls diskrete Verteilungen:
    - pro Ausprägung ein Intervall (falls (i) erfüllt)
    - Schritt 2.1 entfällt
  - (iii) Falls $k = \ell = 2$:
    - Schritt 2.3: kann entfallen
    - Schritt 2.4: Berechne Testfunktionswert gemäß

    18
    $$v = \frac{n(h_{11}h_{22} - h_{12}h_{21})^2}{h_{A_1}h_{A_2}h_{B_1}h_{B_2}}$$
  - (iv) Falls Randwahrscheinlichkeiten $p_i = P(X \in A_i), q_i = P(Y \in B_j)$ bekannt:
    - Schritt 2.2: Randhäufigkeiten entfallen
    - Schritt 2.3: Ersetze 19 $\tilde{h}_{ij}$ durch $n \cdot p_i \cdot q_j$
    - Schritt 3: Ersetze 20 $\chi^2((k-1)\cdot()\ell-1)$ durch $\chi^2(k\cdot\ell-1)$

- Beispiel: Eine regionale Tageszeitung lässt per Telefon ihre Leser befragen, wie gut sie sich durch die Zeitung informiert fühlen. Zur Auswahl stehen die Antworten gut, mäßig, schlecht. Außerdem wurde nach dem Alter des Urteilenden gefragt. Hier das Ergebnis:

| | | Bewertung | | |
|---|---|---|---|---|
| | | gut | mäßig | schlecht |
| Alter | unter 20 | 24 | 11 | 5 |
| | 20 bis 40 | 32 | 17 | 11 |
| | über 40 | 64 | 22 | 14 |

Ist die Meinung der Leser über ihre Zeitung mit dem Signifikanzniveau $\alpha = 0.01$ signifikant vom Alter abhängig? $\Rightarrow$ $H_0$: „Alter und Meinung sind voneinander unabhängig“

21

1. $\alpha = 0.01$

2. 2.1 Diskrete Verteilung, $A_j, B_j$ gegeben

   2.2

| $h_{ij}$ | | | | |
|---|---|---|---|---|
| | 24 | 11 | 5 | 22 40 |
| | 32 | 17 | 11 | 60 |
| | 64 | 22 | 14 | 100 |
| | 23 120 | 50 | 30 | 24 200 |

   2.3

$\tilde{h}_{ij}$

25

| | | | |
|---|---|---|---|
| 24 | 10 | 6 | 40 |
| 36 | 15 | 9 | 60 |
| 60 | 25 | 15 | 100 |
| 120 | 50 | 30 | 200 |

   2.4

26

$$v = \frac{24^2}{24} + \frac{11^2}{10} + \frac{5^2}{6}$$
$$\frac{32^2}{36} + \frac{17^2}{15} + \frac{11^2}{9}$$
$$\frac{64^2}{60} + \frac{22^2}{25} + \frac{14^2}{15} - 200 = 2.1$$

27

3. $\chi^2((3-1)\cdot(3-1)) = \chi^2(4)$: $x_{1-\alpha} = x_{0.99} = 13.28 \Rightarrow B = (9.2, \infty)$

28

4. $v \notin B \Rightarrow H_0$ beibehalten.

Die Meinung hängt somit 29 nicht signifikant vom Alter ab.

## 2 Korrelationstest

- Gegeben: Zwei verbundene einfache Stichproben

30 $X_1, \dots, X_n$ und $Y_1, \dots, Y_n$

mit

31 $X_i \sim N(\mu_1, \sigma_1)$ und $Y_i \sim N(\mu_2, \sigma_2)$

- Hypothesenpaare:

32

a) $H_0\colon \varrho = 0$ , $H_1\colon \varrho \neq 0$

b) $H_0\colon \varrho = 0$ (oder $\varrho \geq 0$), $H_1\colon \varrho < 0$

c) $H_0\colon \varrho = 0$ (oder $\varrho \leq 0$), $H_1\colon \varrho > 0$

- Vorgehen:

Schritt 1: Ein Signifikanzniveau 33 $\alpha$ festlegen.

Schritt 2: Mit Hilfe des empirischen oder Stichproben-Korrelationskoeffizienten

34 $r = \dfrac{s_{xy}}{s_x \cdot s_y}$ mit der Stichproben-Kovarianz

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n-1} \left( \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right)$$

der Beobachtungspaare $(x_i, y_i)$, $i = 1, \dots, n$ den Testfunktionswert $v$ berechnen:

35 $$v = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

Schritt 3: Den Verwerfungsbereich festlegen:

36

| | |
|---|---|
| $B = (-\infty, -X_{1-\alpha/2}) \cup (x_{1-\alpha/2}, \infty)$ | Lücke 32 Fall a) |
| $B = (-\infty, -X_{1-\alpha})$ | Lücke 32 Fall b) |
| $B = (x_{1-\alpha}, \infty)$ | Lücke 32 Fall c) |

Fraktilswerte $x_{1-\alpha/2}$ bzw. $x_{1-\alpha}$ der 37 $t(n-2)$ -Verteilung entnehmen.

Schritt 4: $H_0$ genau dann verwerfen, wenn 38 $v \in B$ gilt

- Beispiel ([BV12, Aufgabe 3.3.11]): In einer Behörde wurde an 25 zufällig ausgewählten Arbeitstagen des Jahres 2009 jeweils die Gesamtlänge $Y$ aller abgehenden Telefonate ermitlelt (in Minuten). Um die Abhängigkeit der „Telefonierfreudigkeit“ vom Wetter zu untersuchen, wurde an diesen Tagen zusätzlich um 13 Uhr der Luftdruck $X$ registriert (in Hektopascal). Aus den $n = 25$ Beobachtungspaaren sind dann
  - die empirischen Varianzen $s_x^2 = 94.09$ bzw. $s_y^2 = 1\,263\,376$ und
  - die empirische Kovarianz $s_{xy} = 3925$

  berechnet worden.

  a) Unterstellen Sie normalverteilte Beobachtungen und prüfen Sie, ob die Gesprächsdauer signifikant ($\alpha = 0.05$) vom Luftdruck abhängt.

  b) Was bedeutet der empirische Korrelationskoeffizient in der vorliegenden Situation?

- Hypothesenpaar: 39 $H_0\colon \varrho = 0,\ H_1\colon \varrho \neq 0$

1. 40 $\alpha = 0.05$

2. 41
$$r = \frac{s_{xy}}{s_x s_y} = \frac{3925}{\sqrt{94.09} \cdot \sqrt{1\,263\,376}} = 0.36$$
$$v = \sqrt{n-1}\frac{r}{1-r^2} = \sqrt{23} \cdot \frac{0.36}{\sqrt{1-0.36^2}} \approx 1.85$$

3. 42
$$t(n-2) = t(23){:}x_{1-\alpha/2} = x_{0.975} = 2.07$$
$$\Rightarrow B = (-\infty, -2.07) \cup (2.07, \infty)$$

4. 43 $v \notin B \Rightarrow H_0$ beibehalten.

Es lässt sich somit 44 nicht nachweisen , dass die Gesprächsdauer signifikant vom Luftdruck abhängt.

Bedeutung von $r$:

Leichte Abhängigkeit zwischen $X$ und $Y$.
$r > 0 \Leftrightarrow$ Je höher der Luftdruck, um so länger wird telefoniert.

## Literatur

[BV12] Udo Bankhofer und Jürgen Vogel. *Übungsbuch Datenanalyse und Statistik*. Springer Gabler, 2012.