Technische Universität Dortmund
Fakultät Statistik
Jun-Prof. Dr. Andreas Groll
M.Sc. Hendrik van der Wurp

Wintersemester 2018/2019
4. Dezember 2018

Übungen zur Vorlesung Wahrscheinlichkeitsrechnung und Mathematische Statistik (für Informatiker)

Blatt 4

Aufgabe 10: (per Hand)

Für die gewählten Biersorgen aus Aufgabe 6 stehe zusätzlich eine Variable Geschlecht zur Verfügung, die angibt, ob das Bier von einer weiblichen oder einer männlichen Person gewählt wurde. Dies führt auf die folgende Kontingenztafel K der beiden Variablen *Biersorte* und *Geschlecht*:

		Biersorte		
		B1	B2	B3
		Dortmunder Oniun	Dortmunder Kornen	Dortmunder Hasna
Geschlecht	weiblich	15	10	6
	männlich	25	50	14

(a) Berechnen Sie die (univariaten) relativen Randhäufigkeitsverteilungen und die bedingten Verteilungen, jeweils für die beiden Variablen Biersorte und Geschlecht.

(b) Berechnen Sie für die Kontingenztafel K den Wert der χ^2-Größe.

(c) Welche Einträge in der Kontingenztafel sind größer als der jeweils erwartete Wert unter Unabhängigkeit der beiden Variablen? Wie ist dies zu interpretieren?

Lösungsvorschlag:

(a) Die Spaltensummen sind 40, 60 und 20, die Zeilensummen 31 und 89. Mit der Gesamtzahl von $N = 120$ ergibt sich für die relativen Häufigkeiten mit $\frac{15}{120} = 0.125$, $\frac{10}{120} = 0.083$, $\frac{6}{120} = 0.05$ usw. die Kontingenztafel:

	B1	B2	B3	Σ
weiblich	0.125	0.083	0.05	0.258
männlich	0.208	0.417	0.117	0.742
Σ	0.333	0.5	0.167	1

Von hier aus (oder von der Kontingenztafel mit absoluten Zahlen direkt) lassen sich die bedingten Verteilungen berechnen. Zuerst bedingt auf die Biersorte, das heißt es wird auf die Anteile $\frac{40}{120} = 0.333$ für B1, $\frac{60}{120} = 0.5$ für B2 und $\frac{20}{120} = 0.167$ für B3 bedingt:

	B1	B2	B3
weiblich	$\frac{0.125}{0.333}$	$\frac{0.083}{0.5}$	$\frac{0.05}{0.167}$
männlich	$\frac{0.208}{0.333}$	$\frac{0.417}{0.5}$	$\frac{0.117}{0.167}$
Σ	1	1	1

$\approx$

	B1	B2	B3
weiblich	0.375	0.167	0.3
männlich	0.625	0.833	0.7
Σ	1	1	1

Problem: Aufsummierende Rundungsfehler. Z.B. $\frac{0.125}{0.333} \neq 0.375 = \frac{15}{40} = \frac{15/120}{40/120}$. Nach Möglichkeit immer mit exakten Werten rechnen, das heißt speziell Zwischenschritte mit Rundungen sollten vermieden werden.

Zuletzt bedingt auf Geschlecht:

	B1	B2	B3	Σ
weiblich	$\frac{15}{31}$	$\frac{10}{31}$	$\frac{6}{31}$	1
männlich	$\frac{25}{89}$	$\frac{50}{89}$	$\frac{14}{89}$	1

$\approx$

	B1	B2	B3	Σ
weiblich	0.484	0.323	0.194	1
männlich	0.281	0.562	0.157	1

(b) χ^2: Dazu wird zuerst die Kontingenztafel der bei Unabhängigkeit erwarteten Einträge benötigt. Diese berechnen sich über das Produkt der Randhäufigkeiten. Zum Beispiel für weiblich und B1: $\frac{N_{\cdot i} \cdot N_{j \cdot}}{N} = \frac{40 \cdot 31}{120} = 10.333$. So ergibt sich die Kontingenztafel der **erwarteten** Einträge:

	B1	B2	B3	Σ
weiblich	$\frac{40 \cdot 31}{120}$	$\frac{60 \cdot 31}{120}$	$\frac{20 \cdot 31}{120}$	31
männlich	$\frac{40 \cdot 89}{120}$	$\frac{60 \cdot 89}{120}$	$\frac{20 \cdot 89}{120}$	89
Σ	40	60	20	120

$\approx$

	B1	B2	B3	Σ
weiblich	10.333	15.5	5.167	31
männlich	29.667	44.5	14.833	89
Σ	40	60	20	120

Zusammen mit der ursprünglichen Kontingenztafel lässt sich dann χ^2 bestimmen als

$$\begin{aligned}
\chi^2 = & \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - v_{ij})^2}{v_{ij}} \\
= & \frac{(15 - 10.333)^2}{10.333} + \frac{(25 - 29.667)^2}{29.667} + \frac{(10 - 15.5)^2}{15.5} + \frac{(50 - 44.5)^2}{44.5} + \frac{(6 - 5.167)^2}{5.167} + \\
& \frac{(14 - 14.833)^2}{14.833} \\
= & 2.108 + 0.734 + 1.952 + 0.680 + 0.134 + 0.047 \\
= & 5.654
\end{aligned}$$

χ^2 liegt zwischen 0 und $N \cdot (\min\{I, J\} - 1) = 120$ und spricht hier daher für eine eher schwache Abhängigkeit. Der korrigierte Kontingenzkoeffizient nach Pearson bestätigt das:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N} \frac{\min\{I, J\}}{\min\{I, J\} - 1}} = \sqrt{\frac{5.654}{5.654 + 120} \frac{2}{1}} = 0.300$$

(c) Das Bier B1 trinken weniger weibliche Studierende als erwartet. Da über die Randsummen gerechnet wird, bedeutet das automatisch, dass es von mehr männlichen Studierenden als erwartet getrunken wird. Bier B2 ist bei den Frauen stärker als erwartet repräsentiert. Die Unterschiede sind jeweils ca. fünf Studierende bei B1 und B2 sowie beiden Geschlechtern. Bei B3 sind die Unterschiede zwischen realisierten und erwarteten Werten gering. Es existiert vermutlich ein (schwacher) Zusammenhang zwischen Geschlecht und Biersorte.

Aufgabe 12: (per Hand)

Gegeben seien vier Beobachtungen eines Datensatzes mit zwei Variablen X und Y:

$$x_1 = 0,\ x_2 = 3,\ x_3 = -3,\ x_4 = 2, \qquad y_1 = 5,\ y_2 = 4,\ y_3 = 3,\ y_4 = 9$$

(a) Berechnen Sie für die beiden Variablen Mittelwert, Varianz und Standardabweichung.

(b) Berechnen Sie für die beiden Variablen die Korrelationskoeffizienten nach Bravais-Pearson und nach Spearman.

(c) Berechnen Sie die Regressionsparameter des linearen Modells $y = c + dx$, bei dem also Y durch X vorhergesagt wird.

Lösungsvorschlag:

(a) Kennzahlen berechnen

$\rightarrow\ \bar{x} = \frac{0+3-3+2}{4} = \frac{1}{2}$

$\rightarrow\ s_x^2 = \frac{1}{3}((0-0.5)^2+(3-0.5)^2+(-3-0.5)^2+(2-0.5)^2) = \frac{1}{3}(0.25+6.25+12.25+2.25) = \frac{21}{3} = 7$

$\rightarrow\ s_x = \sqrt{7} \approx 2.646$

$\rightarrow\ \bar{y} = \frac{5+4+3+9}{4} = 5.25$

$\rightarrow\ s_y^2 = \frac{1}{3}((5-5.25)^2+(4-5.25)^2+(3-5.25)^2+(9-5.25)^2) = \frac{1}{3}(0.0625+1.5625+5.0625+14.0625) = \frac{20.75}{3} \approx 6.917$

$\rightarrow\ s_y \approx \sqrt{6.917}$

(b) Für den Korrelationskoeffizienten nach Bravais-Pearson wird noch die Kovarianz benötigt.

$$\begin{aligned} s_{xy} &= \frac{1}{3}\left((0-0.5)(5-5.25)+(3-0.5)(4-5.25)+(-3-0.5)(3-5.25)+(2-0.5)(9-5.25)\right) \\ &= \frac{1}{3}\cdot[-0.5\cdot(-0.25)+2.5\cdot(-1.25)+(-3.5)\cdot(-2.25)+1.5\cdot 3.75] \\ &= \frac{1}{3}\cdot(0.125-3.125+7.875+5.625) \\ &= \frac{1}{3}\cdot 10.5 \;=\; 3.5 \end{aligned}$$

Also lautet der Korrelationskoeffizient nach Bravais-Pearson:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{3.5}{\sqrt{7} * \sqrt{6.917}} \;\approx\; 0.503$$

Und für Spearman benötigt man noch die Ränge:
$R(x_1, x_2, x_3, x_4) = (2, 4, 1, 3)$ und
$R(y_1, y_2, y_3, y_4) = (3, 2, 1, 4)$. Damit gilt (Spearman ohne Bindungen):

$$\begin{aligned} r_{xy}^{\mathrm{Sp}} &= 1 - \frac{6}{N(N^2-1)}\sum_{n=1}^{N}\left(R(x_n)-R(y_n)\right)^2 \\ &= 1 - \frac{6}{4(16-1)}\left((2-3)^2+(4-2)^2+(1-1)^2+(3-4)^2\right) \\ &= 1 - \frac{6}{60}\cdot(1+4+1) \;=\; 1-\frac{36}{60} \;=\; 1-\frac{3}{5} \;=\; \frac{2}{5} \;=\; 0.4 \end{aligned}$$

Mit 0.503 und 0.4 deuten die Korrelationskoeffizienten von Bravais-Pearson und Spearman auf einen mittelstarken (linearen bzw. monotonen) Zusammenhang hin.

(c) Regression $y = c + dx$. Regressionsparameter c und d mit Formeln der Vorlesung:

$\rightarrow\ d = \frac{s_{xy}}{s_x^2} \ = \ \frac{3.5}{7} \ = \ 0.5$

$\rightarrow\ c = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x} \ = \ 5.25 - 0.5 \cdot 0.5 \ = \ 5$

$\rightarrow\ y = 5 + 0.5x$ ist die Regressionsgerade