Statistik 1 für WInf, WI(MB), WI(ET), WI(BI)
Übung 2, Lösungsvorschlag

Gruppenübung

G 3 Kontingenztabelle

In einem Experiment zur Wirkung von Alkohol auf die Reaktionszeit wurden insgesamt 400 Versuchspersonen zufällig in zwei Gruppen aufgeteilt. Eine der beiden Gruppen erhielt dabei eine standardisierte Menge Alkohol. Abschließend ergab sich die folgende Kontingenztabelle:

	Reaktion			
	gut	mittel	stark verzögert	
ohne Alkohol	120	60	20	
mit Alkohol	60	100	40	

a) Bestimmen Sie die Randhäufigkeiten dieser Kontingenztabelle, und interpretieren Sie diese, soweit dies sinnvoll ist.

b) Bestimmen Sie diejenige bedingte relative Häufigkeitsverteilung, die sinnvoll interpretierbar ist.

c) Bestimmen Sie den χ^2- und den Kontingenzkoeffizienten.

a) Man erhält die folgende Kontingenztafel inklusive Randhäufigkeiten; bei den Werten in Klammern handelt es sich um die absoluten Häufigkeiten, wenn Unabhängigkeit vorliegt. Diese werden in der Lösung von Teilaufgabe c) benötigt.

	Reaktion			
	gut	*mittel*	*stark verzögert*	
ohne Alkohol	*120*	*60*	*20*	*200*
	(90)	*(80)*	*(30)*	
mit Alkohol	*60*	*100*	*40*	*200*
	(90)	*(80)*	*(30)*	
	180	*160*	*60*	*400*

Die 400 Personen wurden jeweils zu gleichen Anteilen in die beiden Gruppen mit und ohne Alkohol eingeteilt. Insgesamt zeigten die allermeisten Versuchspersonen eine gute oder mittlere Reaktionszeit. Lediglich 60 (15 Prozent) zeigten eine stark verzögerte Reaktionszeit.

b) Als bedingte relative Häufigkeitsverteilung, gegeben die Person war alkoholisiert, ergibt sich:

	Reaktion			
	gut	*mittel*	*stark verzögert*	
mit Alkohol	*0.3*	*0.5*	*0.2*	*1*

Entsprechend ermittelt man als bedingte relative Häufigkeitsverteilung, gegeben die Person war nicht alkoholisiert:

	Reaktion			
	gut	*mittel*	*stark verzögert*	
ohne Alkohol	*0.6*	*0.3*	*0.1*	*1*

Ein Vergleich der beiden relativen Häufigkeitsverteilungen zeigt, dass die Reaktionszeiten bei alkoholisierten Personen insgesamt schlechter sind als in der Gruppe ohne Alkohol. Während in der Gruppe der nicht alkoholisierten Personen insgesamt 60 Prozent eine gute Reaktionszeit aufweisen, sind dies in der Gruppe der alkoholisierten Gruppe lediglich 30 Prozent.

c) Die Assoziationsmaße berechnen sich als

$$\begin{aligned}
\chi^2 &= \frac{(120-90)^2}{90} + \frac{(60-80)^2}{80} + \frac{(20-30)^2}{30} + \\
&\quad \frac{(60-90)^2}{90} + \frac{(100-80)^2}{80} + \frac{(40-30)^2}{30} = 36.67 \\
K &= \sqrt{\frac{36.67}{400+36.67}} = 0.29 \\
K_* &= \frac{0.29}{\sqrt{\frac{1}{2}}} = 0.41
\end{aligned}$$

(Beachte dabei: Aus $M = \min\{k,l\} = \min\{2,2\} = 2$ folgt $K_{max} = \sqrt{\frac{1}{2}}$.)

G 4 Lineare Regression

Gegeben sei die zweidimensionale Messreihe $(0,c),(1,1),(2,0)$ mit $c \in \mathbb{R}$. Bestimmen Sie in Abhängigkeit von c die Regressionsgerade.
Für welches $c \in \mathbb{R}$ liegt der erste Punkt auf der Regressionsgeraden? Fertigen Sie für diesen Fall und für den Fall $c = -1$ jeweils eine Skizze an.

Es gelten

$$\begin{aligned}
\bar{x} &= \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{3}{3} = 1 \\
\bar{y} &= \frac{1}{3}\sum_{i=1}^{3} y_i = \frac{c+1}{3} \\
s_X^2 &= \frac{1}{3}\sum_{i=1}^{3} (x_i - \bar{x})^2 = \frac{1+1}{3} = \frac{2}{3} \\
s_{XY} &= \frac{1}{3}\sum_{i=1}^{3} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{3}\left(\sum_{i=1}^{3} x_i y_i - 3\bar{x}\bar{y}\right) \\
&= \frac{1}{3}\left(1 - 3 \cdot 1 \cdot \frac{c+1}{3}\right) = -\frac{c}{3}
\end{aligned}$$

Daraus erhält man

$$a = \frac{s_{XY}}{s_X^2} = -\frac{c}{2}$$
$$b = \bar{y} - a\bar{x} = \frac{c+1}{3} + \frac{c}{2}\cdot 1 = \frac{2c+2+3c}{6} = \frac{5c+2}{6}$$

Also ist wird die Regressionsgerade durch die Gleichung

$$y = -\frac{c}{2}x + \frac{5c+2}{6}$$

beschrieben.

Für welches $c \in \mathbb{R}$ liegt $(0, c)$ auf der Regressionsgeraden?
Punkt in Gerade einsetzen liefert

$$c = -\frac{c}{2}\cdot 0 + \frac{5c+2}{6} \quad \Leftrightarrow \quad 6c = 5c+2 \quad \Leftrightarrow \quad c = 2.$$

Im Fall $c = 2$ liegen alle drei Punkte auf der Regressionsgeraden $y = -x + 2$. Falls $c = -1$ liegt keiner der Punkte auf der Regressionsgeraden $y = \frac{1}{2}x - \frac{1}{2}$.

G 5 Laplace-Wahrscheinlichkeit, Kombinatorik

a) In einer Urne befinden sich 15 Kugeln – 4 weiße, 5 schwarze und 6 rote. Nach gründlichem Mischen werden 5 Kugeln ohne Zurücklegen gezogen. Berechnen Sie die Wahrscheinlichkeit dafür, dass sich unter den gezogenen Kugeln

- keine weiße Kugel befindet,

  $P($ *"keine weiße Kugel"* $) = \frac{\binom{11}{5}}{\binom{15}{5}} = \frac{11\cdot 10\cdot 9\cdot 8\cdot 7}{15\cdot 14\cdot 13\cdot 12\cdot 11} = 0.1538$
- genau zwei schwarze Kugeln befinden,

  $P($ *"genau zwei schwarze Kugeln"* $) = \frac{\binom{5}{2}\binom{10}{3}}{\binom{15}{5}} = \frac{10\cdot 120}{3003} = 0.3996$
- ebensoviele schwarze wie rote Kugeln befinden.

  $P($ *"ebensoviele schwarze wie rote K."* $) = \frac{\binom{6}{1}\binom{5}{1}\binom{4}{3} + \binom{6}{2}\binom{5}{2}\binom{4}{1}}{\binom{15}{5}} = \frac{720}{3003} = 0.2398$

b) Eine Gruppe von 5 Studenten trifft in der Mensa auf einen Öffentlichkeitsarbeiter. Dieser verteilt unter ihnen 5 (ununterscheidbare) Werbegeschenke. Da einige der Studenten schubsen und drängeln, kann es passieren, dass manche von ihnen mehrere bzw. andere kein Werbegeschenk erhalten. Wieviele Möglichkeiten gibt es, die Werbegeschenke auf die 5 Studenten zu verteilen.

„Ziehen" von $k = 5$ aus einer Gruppe von $n = 5$, mit Wiederholung: $\binom{n+k-1}{k} = \binom{9}{5} = 126$

Hausübung

H 5 Kontingenztabelle

100 weibliche Patienten sind mit einer konventionellen Therapie behandelt worden. 85 Patientinnen wurden geheilt, 15 sind gestorben. Von 81 Patientinnen, die mit einer neuen Therapie behandelt wurden, konnten 77 geheilt entlassen werden, 4 sind gestorben.

a) Stellen Sie aus den genannten Häufigkeiten eine 2x2-Kontingenztabelle auf.

b) Ist die Heilungschance der Patientin unabhängig von der angewandten Therapie?

c) Berechnen Sie den Chi-Quadrat- und den (normierten) Kontingenzkoeffizienten. Unter welcher Bedingung erreicht der Kontingenzkoeffizient seinen größten Wert?

a) 2x2-Kontingenztafel (hier mit Randhäufigkeiten):

	geheilt	*gestorben*	*Randhäufigkeiten*
konventionelle Therapie	*85*	*15*	*100*
neue Therapie	*77*	*4*	*81*
Randhäufigkeiten	*162*	*19*	*181*

b) Wir beantworten diese Frage mit Hilfe der bedingten relativen Häufigkeiten für das Eintreten einer Heilung:

$$f(geheilt|konventionelleTherapie) = \frac{85}{100} = 0.85$$
$$f(geheilt|neueTherapie) = \frac{77}{81} \approx 0.95$$

bzw.

$$f(gestorben|konventionelleTherapie) = \frac{15}{100} = 0.15$$
$$f(gestorben|neueTherapie) = \frac{4}{81} \approx 0.05$$

Die Merkmale sind offensichtlich abhängig, da die relativen Häufigkeiten jeweils nicht überein stimmen.

c) $\chi^2 \approx 4.82$
Kontingenzkoeffizient $K \approx 0,1611$
korrigierter Kontingenzkoeffizient $K_* \approx 0,2278$
Der maximale (unkorrigierte) Kontigenzkoeffizient K_{max} für die Vierfeldertafel beträgt 0,7071; er tritt stets bei vollkommener Kontingenz auf, also wenn die entgegengesetzten Komplementärereignisse unbesetzt bleiben. In unserem

Beispiel müßten dabei bei der konventionellen Therapie alle Patienten sterben und bei der neuen Therapie alle Patienten geheilt werden (oder umgekehrt). Der maximale korrigierte Kontingenzkoeffizient ist gleich 1.

H 6 Korrelationskoeffizient

Zeigen Sie, dass der empirische Korrelationskoeffizient bis auf Vorzeichenwechsel invariant gegenüber linearen Transformationen der Daten ist, d.h.:
Sind $(x_1, y_1), \ldots, (x_n, y_n)$ und $(u_1, v_1), \ldots, (u_n, v_n)$ zweidimensionale Messreihen mit

$$u_i = a \cdot x_i + b, \quad v_i = c \cdot y_i + d, \quad i = 1, \ldots, n,$$

für Zahlen $a, b, c, d \in \mathbb{R}$ mit $a \cdot c \neq 0$, so gilt für die jeweiligen empirischen Korrelationskoeffizienten

$$r_{uv} = \begin{cases} r_{xy} & \text{falls } a \cdot c > 0, \\ -r_{xy} & \text{falls } a \cdot c < 0. \end{cases}$$

Es gelten

$$\begin{aligned} \bar{u} &= a\bar{x} + b, \quad s_u^2 = a^2 s_x^2, \\ \bar{v} &= c\bar{y} + d, \quad s_v^2 = c^2 s_y^2, \end{aligned}$$

so dass

$$\begin{aligned} s_{uv} &= \frac{1}{n} \sum_i (ax_i + b - a\bar{x} - b)(cy_i + d - c\bar{y} - d) \\ &= \frac{ac}{n} \sum_i (x_i + \bar{x})(y_i + d\bar{y}) \\ &= ac \cdot s_{xy} \end{aligned}$$

und

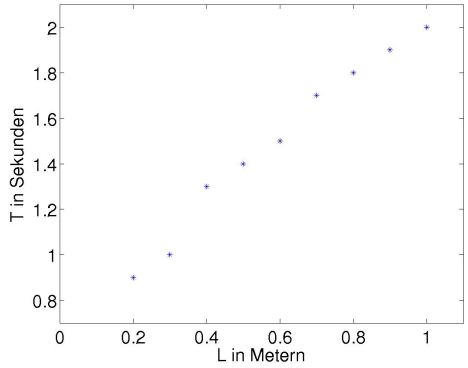
$$r_{uv} = \frac{s_{uv}}{s_u} s_v = \frac{ac}{|ac|} \cdot \frac{s_{xy}}{s_x s_y} = \begin{cases} r_{xy} & \text{für } ac > 0, \\ -r_{xy} & \text{für } ac < 0. \end{cases}$$

H 7 Nichtlineare Regression

Eine Messung der Periodendauer T der Schwingung eines Fadenpendels in Abhängigkeit von dessen Länge L ergab folgende Messwerte:

L in Metern	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
T in Sekunden	0.9	1	1.3	1.4	1.5	1.7	1.8	1.9	2

Stellen Sie die Daten als Punktwolke dar! Nehmen Sie einen Zusammenhang der Form $T = a \cdot L^b$ an und bestimmen Sie diejenigen Parameter a und b, für welche sich der zugehörige Funktionsgraph am besten der gegebenen Punktwolke anpasst, indem Sie die Daten geeignet transformieren und das Problem auf eine lineare Regression zurückführen.



Physik: $T = 2\pi \cdot \sqrt{\frac{L}{g}} \approx 2.0061 \cdot \sqrt{L}$ *mit Erdbeschleunigung* $g \approx 9.81 \frac{m}{s^2}$
Aus $T = a \cdot L^b$ *folgt* $\log T = \log a + b \log L$, *zwischen* $y := \log T$ *und* $x := \log L$ *bestünde also der lineare Zusammenhang* $y = \tilde{a} + \tilde{b}x$, *wobei* $\tilde{a} = \log a$ *und* $\tilde{b} = b$.
Umrechnen der Werte mit dem Logarithmus z.B. zur Basis 2 liefert:

$\log_2 L$	*-2.322*	*-1.737*	*-1.322*	*-1.000*	*-0.737*	*-0.515*	*-0.322*	*-0.152*	*0.000*
$\log_2 T$	*-0.152*	*0.000*	*0.379*	*0.485*	*0.585*	*0.766*	*0.848*	*0.926*	*1.000*

Wir erhalten schließlich $\bar{x} = -0.9007$ *und* $\bar{y} = 0.5374$, *weiterhin* $\sum x_i^2 = 12.0905$, $\sum y_i^2 = 3.9068$ *und* $\sum x_i y_i = -1.3943$. *Daraus ermitteln wir*

$$\begin{aligned} s_{xy} &= \frac{1}{9}\left(\sum x_i y_i - 9 \cdot \bar{x}\bar{y}\right) = 0.3291 \\ s_x^2 &= \frac{1}{9}\left(\sum x_i^2 - 9\bar{x}^2\right) = 0.5321 \end{aligned}$$

und

$$\begin{aligned} b = \tilde{b} &= \frac{s_{xy}}{s_x^2} = \frac{0.3291}{0.5321} = 0.6185 \\ \log_2 a = \tilde{a} &= \bar{y} - b \cdot \bar{x} = 0.5374 - 0.6185 \cdot (-0.9007) = 1.0944 \end{aligned}$$

Die Regressionsgerade für die transformierten Daten lautet also:

$$y = 1.0944 + 0.6185x$$

Rücktransformation liefert: $a = 2^{\tilde{a}} = 2^{1.0944} = 2.1353$ *und*

$$T = 2.1353 \cdot L^{0.6185}$$

H 8 Laplace-Wahrscheinlichkeit

Ein Skatspiel besteht aus 32 Karten, vier davon heißen Buben. Nach dem Mischen der Karten erhalten die drei Spieler jeweils 10 Karten, die übrigen zwei Karten bilden den Skat. Andreas, Bettina und Claudia spielen Skat.

Berechnen Sie die Wahrscheinlichkeiten folgender Ereignisse:

a) Es liegen genau zwei Buben im Skat.

Unter der Laplace-Annahme gilt: $P(D_2) = \frac{\binom{4}{2}}{\binom{32}{2}} = 0.012\,.$

b) Claudia hat genau einen Buben.

Jeder Spieler hat 10 Karten, also gibt es $\binom{4}{1}\binom{28}{9}$ *Möglichkeiten, genau einen Buben auf der Hand zu haben, d. h., unter der Laplace-Annahme gilt:*

$$P(C_1) = \frac{\binom{4}{1}\binom{28}{9}}{\binom{32}{10}} = 0.42825\,.$$

c) Claudia hat mindestens zwei Buben

Es gilt für das Ereignis C "Claudia hat mindestens 2 Buben":

$$\begin{aligned} P(C) &= 1 - (P(C_0) + P(C_1)) \\ &= 1 - \left(\frac{\binom{4}{0}\binom{28}{10}}{\binom{32}{10}} + P(C_1) \right) \\ &= 1 - (0.20342 + 0.42825) = 0.36833 \end{aligned}$$

d) Man nehme nun an, dass Claudias Karten schon ausgeteilt wurden und dass darunter genau ein Bube war. Wie groß ist jetzt die Wahrscheinlichkeit, dass Andreas genau einen Buben erhält?

Wenn Claudia genau einen Buben hat, sind noch 22 Karten (3 Buben und 19 übrige Karten) für Andreas möglich. Dabei gibt es $\binom{3}{1}\binom{19}{9}$ *Möglichkeiten dafür, dass Andreas genau einen Buben erhält. Da es insgesamt noch* $\binom{22}{10}$ *Möglichkeiten gibt, aus 22 Karten 10 auszuwählen, ist die Wahrscheinlichkeit, dass Andreas genau einen Buben erhält gleich*

$$\frac{\binom{3}{1}\binom{19}{9}}{\binom{22}{10}} = 0.42857\,.$$

e) Wie hoch ist die Wahrscheinlichkeit dafür, dass jeder Spieler genau einen Buben hat?

Sei D das Ereignis "jeder Spieler hat genau einen Buben", dann gilt:

$$P(D) = P(A_1 \cap B_1 \cap C_1).$$

Es gibt

$$J := \binom{4}{1}\binom{28}{9} \cdot \binom{3}{1}\binom{19}{9} \cdot \binom{2}{1}\binom{10}{9} \cdot \binom{1}{1}\binom{1}{1} \approx 1.531 \cdot 10^{14}$$

Möglichkeiten dafür, dass jeder Spieler genau einen Buben hat.
Da es insgesamt $I := \binom{32}{10}\binom{22}{10}\binom{12}{10}\binom{2}{2} \approx 2.7533 \cdot 10^{15}$ *verschiedene Möglichkeiten gibt, die 32 Karten auf Andreas, Bettina, Claudia und den Skat aufzuteilen, gilt*

$$P(D) = \frac{J}{I} \approx 0.05562\,.$$