DHBW Mannheim
Informatik

Mannheim, 20.04.2023
Mathematik II/2

Lösungen zu Übungsblatt 2

Aufgabe 1. Für eine Untersuchung der Abhängigkeit des Einkommens vom Geschlecht werden die monatlichen Nettoeinkommen in drei Gruppen eingeteilt:

- g : geringes Einkommen (unter €1800).
- m : mittleres Einkommen (zwischen €1800 und €3500).
- h : hohes Einkommen (mehr als €3500).

Eine empirische Untersuchung ergibt folgende absolute Häufigkeiten:

	g	m	h
Frauen	140	100	60
Männer	90	240	140
divers	10	15	5

a) Erläutern Sie den Begriff *empirisch unabhängig* und überprüfen Sie die Merkmale *Geschlecht* und *Einkommen* in obiger Tabelle auf empirische Unabhängigkeit.

Lösung:

Die empirische Unabhängigkeit besagt, dass die bedingte Häufigkeit für das Auftreten einer bestimmten Einkommensklasse unter der Bedingung, dass ein bestimmtes Geschlecht vorliegt mit der relativen Häufigkeit für diese Einkommensklasse übereinstimmt, und entsprechend für das Geschlecht in Abhängigkeit vom Einkommen. Empirische Unabhängigkeit besagt also, dass die relative Häufigkeit dafür, in einer bestimmten Einkommensklasse zu liegen, nicht vom Geschlecht abhängt.

Wir ergänzen die Tabelle um die Randhäufigkeiten

	g	m	h	
Frauen	140	100	60	300
Männer	90	240	140	470
divers	10	15	5	30
	240	355	205	800

Der Stichprobenumfang ist $n = 800$. Es gilt

$$\frac{h_{1,\bullet} \cdot h_{\bullet,1}}{800} = \frac{300 \cdot 240}{800} = 90 \neq 140 = h_{1,1}$$

und damit sind die beiden Merkmale nicht empirisch unabhängig.

b) Ermittlen Sie den χ^2–Koeffizienten und den Kontingenzkoeffizienten für diese Daten und interpretieren Sie diese Koeffizienten.

Lösung:

Wir ermitteln die Häufigkeiten, die zu erwarten sind, wenn kein empirischer Zusammenhang vorliegt,

$$\widetilde{h_{i,j}} = \frac{h_{i,\bullet} \cdot h_{\bullet,j}}{800}$$

und erhalten folgende Tabelle

	g	m	h	
Frauen	90	$\frac{1065}{8}$	$\frac{615}{8}$	300
Männer	141	$\frac{3337}{16}$	$\frac{1927}{16}$	470
divers	9	$\frac{213}{16}$	$\frac{123}{16}$	30
	240	355	205	800

Damit ergibt sich χ^2 nach der Formel

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{\left(h_{i,j} - \widetilde{h_{i,j}}\right)^2}{\widetilde{h_{i,j}}}$$

also

$$\chi^2 = 67.352$$

Dieser Wert kann noch nicht direkt interpretiert werden, da er nicht normiert ist. Dazu benutzen wir den Kontingenzkoeffizienten

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}} = 0.2787$$

oder noch besser den korrigierten Kontingenzkoeffizienten

$$K^\star = \frac{K}{\sqrt{\frac{2}{3}}} = 0.3413$$

Die Kontingenzkoeffizienten zeigen, dass ein erkennbarer Zusammenhang zwischen Geschlecht und Zugehörigkeit zu einer Einkommensklasse vorliegt, dass allerdings das Geschlecht die Einkommensklasse noch nicht bestimmt (und auch nicht umgekehrt).

c) Angenommen, eine andere Untersuchung der Ihren Daten zugrundegelegten Grundgesamtheit ergibt ein Durchschnittseinkommen von € 3170 bei Frauen, € 3160 bei Männern und € 3180 bei Diversen und kommt zu dem Schluss, dass sich das Geschlecht nicht auf das durchschnittliche Einkommen auswirkt. Erläutern Sie, warum dieses Ergebnis Ihren Daten nicht widerspricht.

Lösung:

In das Durchschnittseinkommen gehen alle Werte, auch Extremwerte ein. Offensichtlich verfügen einige der Frauen und Diverse in der untersuchten Grundgesamtheit über ein relativ hohes Einkommen, was dazu führt, dass das Durchschnittseinkommen in dieser Gruppe relativ hoch ist. Das ändert jedoch nichts daran, dass die Frauen und Diverse in der oberen Einkommensgruppe deutlich unterproportional vertreten sind.

Aufgabe 2. Für eine Untersuchung der Abhängigkeit des Tätigkeitsfelds von Arbeitnehmern vom Ausbildungsstand werden die drei Ausbildungsniveaus *wrs* (maximal Werkrealschulabschluss), *abi* (mittlere Reife oder Abitur ohne Studium) und *stu* (abgeschlossenes Studium) festgelegt. Die Einsatzbereiche werden unterteilt in P (Produktion), V (Verwaltung) und M (Marketing).
Eine Erhebung der Arbeitnehmer in den verschiedenen Einsatzbereichen ergibt folgende absolute Häufigkeiten:

Bereich	P	V	M
wrs	180	90	30
abi	150	150	50
stu	70	110	170

a) Sind die Merkmale *Tätigkeitsfeld* und *Ausbildungsstand* empirisch unabhängig?

Lösung:

Zunächst ergänzen wir die Tabelle der Häufigkeiten um die Randhäufigkeiten:

Bereich	P	V	M	
wrs	180	90	30	300
abi	150	150	50	350
stu	70	110	170	350
	400	350	250	1000

Aus den Randhäufigkeiten ermitteln wir die Häufigkeiten, die zu erwarten sind, wenn kein Zusammenhang besteht,

$$\widetilde{h_{i,j}} = \frac{h_{i,\bullet} \cdot h_{\bullet,j}}{1000}$$

und erhalten daraus folgende erwartete Häufigkeitstabelle bei empirischer Unabhängigkeit

Bereich	P	V	M
wrs	120.00	105.00	75.00
abi	140.00	122.50	87.50
stu	140.00	122.50	87.50

Diese Tabelle weicht in allen Einträgen von der Tabelle der tatsächlichen Häufigkeiten ab, und zwar in den meisten Fällen ziemlich deutlich. Daher kann hier nicht mehr von eine empirischen Unabhängigkeit gesprochen werden.

b) Ermitteln Sie den χ^2–Koeffizienten, den Kontingenzkoeffizienten und den normierten Kontingenzkoeffizienten der beiden Merkmale und interpretieren Sie diese Zahlen.

Lösung:

Es gilt:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(h_{i,j}-\widetilde{h_{i,j}})^2}{\widetilde{h_{i,j}}}, \\ K &= \sqrt{\frac{\chi^2}{n+\chi^2}} \\ K^* &= \frac{K}{\sqrt{\frac{M-1}{M}}} \end{aligned}$$

wobei hier $n = 1000$ und $M = 3$, also

$$\begin{aligned} \chi^2 &= \frac{(180-120)^2}{120} + \frac{(90-105)^2}{105} + \frac{(30-75)^2}{75} \\ &\quad + \frac{(150-140)^2}{140} + \frac{(150-122.5)^2}{122.50} + \frac{(50-87.50)^2}{87.50} \\ &\quad + \frac{(70-140)^2}{140} + \frac{(110-122.5)^2}{122.50} + \frac{(170-87.50)^2}{87.50} \\ &= 196.2, \\ K &= 0.405, \\ K^* &= 0.496 \end{aligned}$$

Da K^* immer zwischen 0 und 1 liegt, zeigen die Kontingenzkoeffizienten eine deutliche empirische Abhängigkeit der Merkmale, wenn auch der Ausbildungsstand den Einsatzbereich noch nicht vollständig bestimmt (was bei $K^* = 1$ der Fall wäre).

Aufgabe 3. Wir betrachten die beiden Merkmale X: *Nettohaushaltseinkommen* (in 1000 €) im Jahr 2019 und Y: *Haushaltsausgaben* (in 1000 €) im Jahr 2019. Eine Untersuchung von 10 zufällig ausgewählten Haushalten lieferte folgendes Ergebnis:

k	1	2	3	4	5	6	7	8	9	10
X	44	37	33	42	58	24	63	51	40	39
Y	43	32	22	46	52	25	50	52	32	39

a) Ermitteln Sie den Bravais–Pearson–Korrelationkoeffizienten und den Spearman–Korrelationskoeffizienten und interpretieren Sie die Ergebnisse.

Lösung:

Aus den Formeln ergibt sich für die Merkmale $X =$ *Nettohaushaltseinkommen in 2019* und $Y =$ *Haushaltsausgaben in 2019* zunächst

$$\overline{x} = 43.1$$

und

$$\overline{y} = 39.3$$

und damit:

$$r_{X,Y} = \frac{\sum_{i=1}^{10}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{10}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{10}(y_i - \overline{y})^2}} = 0.876$$

Der Korrelationskoeffizient deutet also auf eine starke positive lineare Korrelation von Haushaltseinkommen und Haushaltsausgaben hin.

Für den Spearman–Korrelationskoeffizienten bestimmen wir zunächst die Ränge :

k	1	2	3	4	5	6	7	8	9	10
rg_X	4	8	9	5	2	10	1	3	6	7
rg_Y	5	7.5	10	4	1.5	9	3	1.5	7.5	6

Beachten Sie dabei, dass in der y–Zeile mehrere gleiche Werte sind, sodass an diesen Stellen die Ränge gemittelt werden müssen. Damit gilt $\overline{\mathrm{rg}_X} = 5.5$ und $\overline{\mathrm{rg}_5} = 5.5$ und eine Vorgehensweise identisch zu der bei Bravais–Pearson liefert

$$r_{Sp} = 0.9147$$

also auch eine sehr starke positive Rangkorrelation.

b) Bestimmen Sie die Regressionsgerade für das Merkmal $Y =$ *Ausgaben in 2019* in Abhängigkeit von $X =$ *Einkommen in 2019*.

Lösung:

Aus Teil a) wissen wir bereits

$$\overline{x} = 43.1$$

und

$$\overline{y} = 39.3$$

Die Koeffizienten $\widehat{a}$ und $\widehat{b}$ der Regressionsgerade

$$f(x) = \widehat{a} \cdot x + \widehat{b}$$

ermittelt sich laut Vorlesung nach der Formel

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.84$$

und

$$\widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x} = 3.24$$

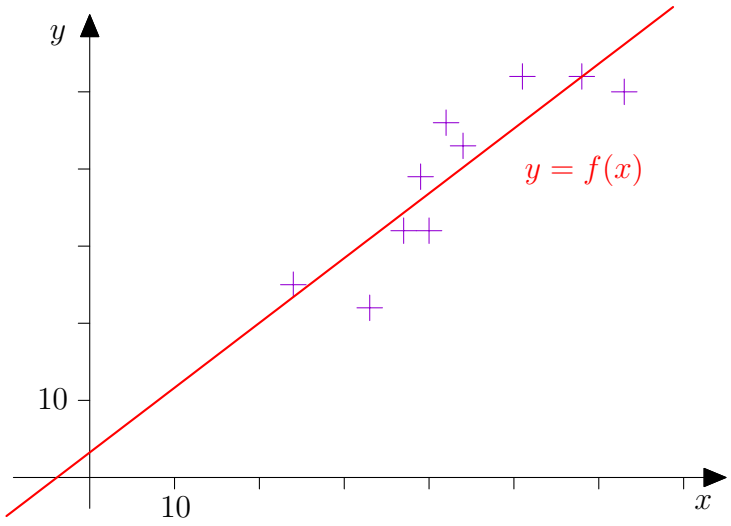
so dass also die Regressionsgerade die Form

$$f(x) = 0.84 \cdot x + 3.24$$

hat. Das Bestimmtheitsmaß der Regressionsgerade ist

$$R^2 = \frac{\sum_{i=1}^{n}(f(x_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 0.77$$

Das deutet auf einen sehr guten Erklärungsgrad der tatsächlichen Werte durch die Regressionsgerade hin, was man auch am Graphen sieht:



Aufgabe 4. Für sieben ausgewählte DAX–Unternehmen ergaben sich 2017 folgende Zahlen für Umsatz (in Mio. €) und Mitarbeiter (im Jahresmittel):

	1	2	3	4	5	6	7
Umsatz	230 682	164 330	126 149	98 678	83 049	74 942	60 444
Mitarbeiter	627 000	285 000	140 500	125 000	351 000	221 000	498 500

a) Ermitteln Sie den Bravais–Pearson–Korrelationkoeffizienten und den Spearman–Korrelationskoeffizienten und interpretieren Sie die Ergebnisse.

Lösung:

Aus den Formeln ergibt sich für die Merkmale $X =$ *Mitarbeiter in 2017* und $Y =$ *Umsatz in 2017* zunächst

$$\overline{x} = 321\,144$$

und

$$\overline{y} = 119\,753$$

und damit:

$$r_{X,Y} = \frac{\sum_{i=1}^{7}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{7}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{7}(y_i - \overline{y})^2}} = 0.4194$$

Für den Spearman–Korrelationskoeffizienten ist zunächst die Rangtabelle zu bestimmen:

	1	2	3	4	5	6	7
Rang Umsatz	1	2	3	4	5	6	7
Rang Mitarbeiter	1	4	6	7	3	5	2

Der Mittelwert der Ränge ist in beiden Fällen $\overline{rg} = \frac{7+1}{2} = 4$. Damit ergibt sich als Spearman–Korrelationskoeffizient nach der Berechnungsformel

$$r_{Sp} = \frac{\sum_{i=1}^{7}(rg(x_i) - 4) \cdot (rg(y_i) - 4)}{\sqrt{\sum_{i=1}^{7}(rg(x_i) - 4)^2 \cdot \sum_{i=1}^{7}(rg(y_i) - 4)^2}} = 0.0714$$

Der Bravais–Pearson–Korrelationkoeffizient deutet auf eine Korrelation zwischen Anzahl der Mitarbeiter und Umsatz hin, die zumindest mittelstark ausgeprägt ist. Der Spearman–Korrelationskoeffizient deutet dagegen eher auf eine schwache Korrelation der Ränge hin.

b) Bestimmen Sie die Regressionsgerade für das Merkmal $Y =$ *Umsatz in 2017* in Abhängigkeit von $X =$ *Mitarbeiter in 2017*.

Lösung:

Für die Bestimmung der Regressionsgerade benötigen wir die Mittelwerte. Hierzu gilt (siehe a))

$$\overline{x} = 321\,144$$

und

$$\overline{y} = 119\,753$$

Die Koeffizienten $\widehat{a}$ und $\widehat{b}$ der Regressionsgerade

$$f(x) = \widehat{a} \cdot x + \widehat{b}$$

ermittelt sich laut Vorlesung nach der Formel

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.135$$

und

$$\widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x} = 76\,333$$

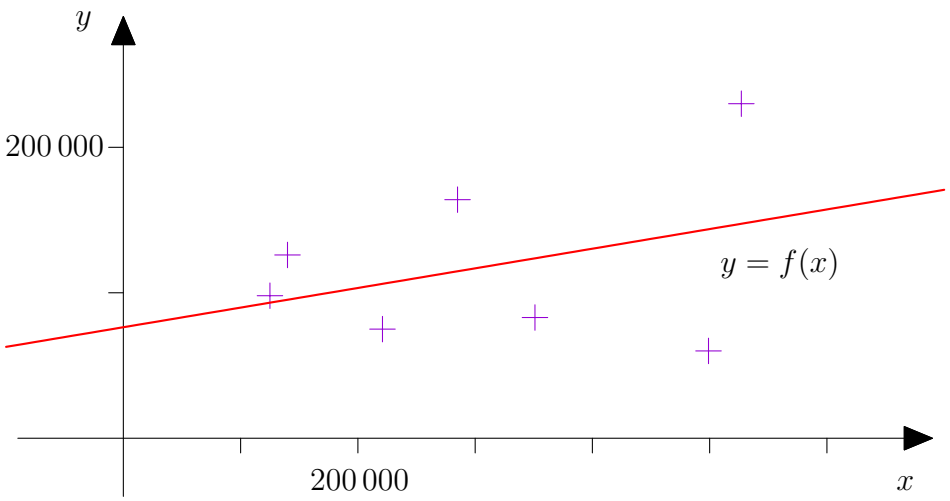
so dass also die Regressionsgerade die Form

$$f(x) = 0.135 \cdot x + 76\,333$$

hat. Das Bestimmtheitsmaß der Regressionsgerade ist

$$R^2 = \frac{\sum_{i=1}^{n}(f(x_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 0.176$$

Das deutet auf einen eher mäßigen Erklärungsgrad der tatsächlichen Werte durch die Regressionsgerade hin, man kann also sicherlich noch nicht von einer linearen Abhängigkeit des Umsatzes durch die Anzahl der Mitarbeiter sprechen. Das sieht man auch an der zugehörigen Graphik



Beachten Sie in dieser Aufgabe die Rolle von X und Y. Das Merkmal X ist hier in der zweiten Zeile aufgelistet. Falls Sie mit der umgekehrten Bezeichnung der Merkmale gearbeitet haben, sollten Sie das Ergebnis

$$\widehat{a'} = 1.30, \qquad \widehat{b'} = 165\,336$$

erhalten