Aufgabe 1

Teil A

Im Rahmen einer Studienqualitätsumfrage wurden 200 Studierende nach dem Besuch eines Tutoriums ihrer Wahl befragt. Dabei wurden zwei statistische Größen erhoben:

X : Wie gut konnten die Studierenden dem Tutorium folgen?
mit den möglichen Ausprägungen: *schlecht* (a_1), *mittelmäßig* (a_2), *sehr gut* (a_3)

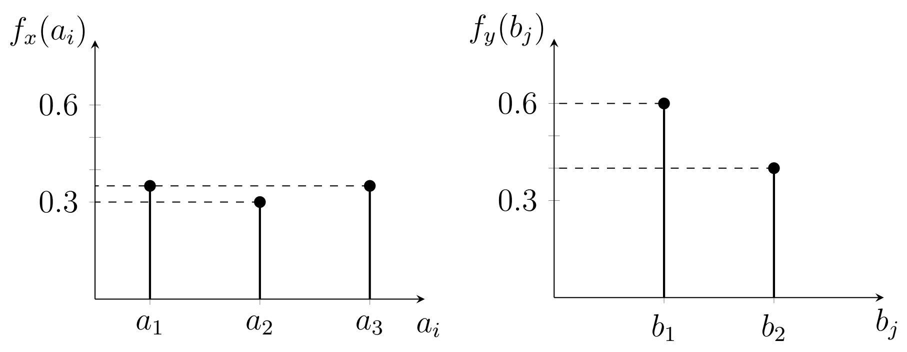
Y : Haben sich die Studierenden vor dem Tutorium mit dem Vorlesungsstoff befasst?
mit den möglichen Ausprägungen: *ja* (b_1), *nein* (b_2)

Aus der Umfrage ergab sich, dass **jeweils** 70 Studierende angaben, dem Tutorium schlecht bzw. sehr gut folgen zu können. Insgesamt gaben 120 Studierende an, sich vor dem Tutorium mit dem Vorlesungsstoff befasst zu haben.

a) Wie sind die statistische Merkmale skaliert?

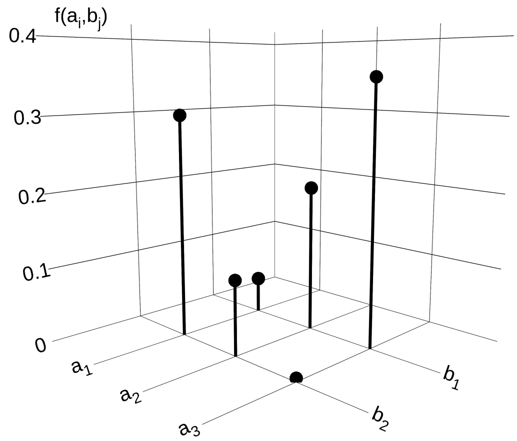
X ist ordinal (diskret) skaliert; Y ist nominal (diskret) skaliert.

b) Skizzieren Sie jeweils die relative Häufigkeitsfunktion für X und Y in einem Stabdiagramm.



Teil B

Nun wurden die Antworten beider Merkmale (X, Y) gemeinsam ausgewertet. Die zugehörige gemeinsame relative Häufigkeitsfunktion $f(a_i, b_j) := f_{ij}$ wurde grafisch dargestellt und in einer Kontingenztabelle zusammengefasst.



X \ Y		ja (b_1)	nein (b_2)	$\sum$
schlecht	(a_1)	0,05	0,30	0,35
mittelmäßig	(a_2)	0,20	0,10	0,30
sehr gut	(a_3)	0,35	0,00	0,35
$\sum$		0,60	0,40	1,00

c) Verbalisieren Sie f_{21}, $f_{\cdot 1}$, $f_{2\cdot}$, $f_x(a_1|b_1)$ und $f_y(b_2|a_1)$.

– $f_{21} = 0,20$,
d.h. 20% aller Studierenden geben an, dem Tutorium mittelmäßig folgen zu können *und* sich vor dem Tutorium mit dem Vorlesungsstoff befasst zu haben.
– $f_{\cdot 1} = 0,60$,
d.h. 60% aller Studierenden haben sich vor dem Tutorium mit dem Vorlesungsstoff befasst.
– $f_{2\cdot} = 0,3$,
d.h. 30% aller Studierenden können dem Tutorium mittelmäßig folgen.
– $f_x(a_1|b_1) = \frac{f_{11}}{f_{\cdot 1}} = \frac{0{,}05}{0{,}6} = \frac{1}{12}$,
d.h. ein Zwölftel der Studierenden, die sich vor dem Tutorium mit dem Vorlesungsstoff befasst haben, kann dem Tutorium schlecht folgen.
– $f_y(b_2|a_1) = \frac{f_{12}}{f_{1\cdot}} = \frac{0{,}3}{0{,}35} = \frac{6}{7}$,
d.h. sechs Siebtel der Studierenden, die dem Tutorium schlecht folgen können, befassten sich vor dem Tutorium nicht mit dem Vorlesungsstoff.

d) Erläutern Sie, weshalb die Gesamtheit aller
   i) gemeinsamen relativen Häufigkeiten f_{ij},
   ii) relativen Randhäufigkeiten $f_{i\cdot}$ bzw. $f_{\cdot j}$ und
   iii) bedingten relativen Häufigkeiten $f_x(a_i|b_j)$ bzw. $f_y(b_j|a_i)$

   jeweils eine diskrete relative Häufigkeits*verteilung* bildet.

Es gilt:

$$
\begin{aligned}
\textit{i)}\quad & \sum_{i=1}^{3}\sum_{j=1}^{2} f_{ij} &&= (0,05+0,30)+(0,20+0,10)+(0,35+0,00)=1\\
\textit{ii)}\quad & \sum_{i=1}^{3} f_{i\cdot} &&= 0,35+0,3+0,35=1 \quad \text{bzw.}\\
& \sum_{j=1}^{2} f_{\cdot j} &&= 0,60+0,4=1\\
\textit{iii)}\quad & \sum_{i=1}^{3} f_x(a_i|b_1) &&= f_x(a_1|b_1)+f_x(a_2|b_1)+f_x(a_3|b_1)\\
& &&= \frac{f(a_1,b_1)+f(a_2,b_1)+f(a_3,b_1)}{f_{\cdot 1}}\\
& &&= \frac{0,05+0,20+0,35}{0,60}=1 \quad (\text{analog für } f_x(a_i|b_2))\\
& \sum_{j=1}^{2} f_y(b_j|a_1) &&= f_y(b_1|a_1)+f_y(b_2|a_1)\\
& &&= \frac{f(b_1,a_1)+f(b_2,a_1)}{f_{1\cdot}}\\
& &&= \frac{0,05+0,30}{0,35}=1 \quad (\text{analog für } f_y(b_j|a_2) \text{ und } f_y(b_j|a_3))
\end{aligned}
$$

Die Gesamtheit aller relativen Häufigkeiten ergibt jeweils 1 und jede relative Häufigkeit nimmt einen Wert zwischen 0 und 1 an.

e) Skizzieren Sie die relativen Randhäufigkeiten von X und Y in Stabdiagrammen und erläutern Sie formal und grafisch den Zusammenhang zu den gemeinsamen relativen Häufigkeiten.

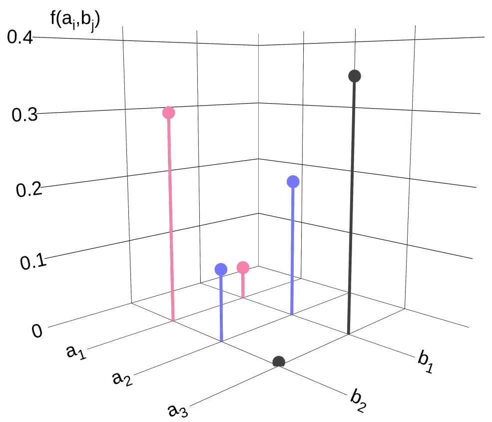
Für Skizze siehe Teilaufgabe b).

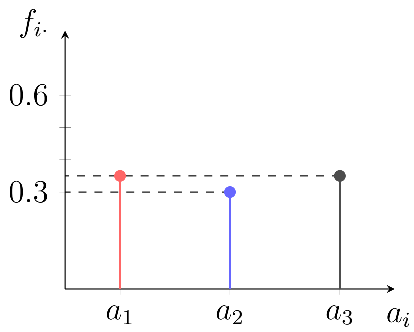
Formaler Zusammenhang:

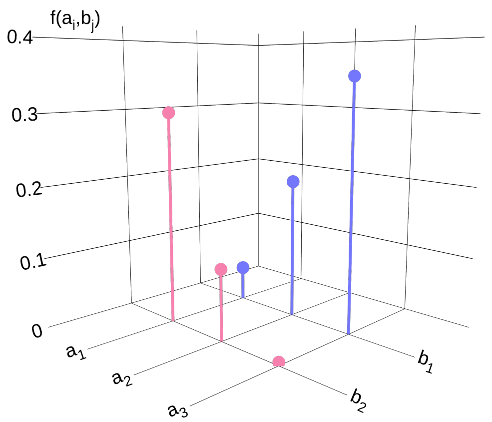
$$f_{i\cdot} = \sum_{j=1}^{2} f_{ij} \quad \text{für alle } i = 1,2,3 \qquad \text{bzw.} \qquad f_{\cdot j} = \sum_{i=1}^{3} f_{ij} \quad \text{für alle } j = 1,2$$

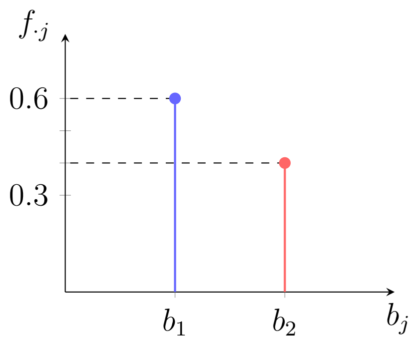
Grafischer Zusammenhang:

Die relativen Häufigkeiten der gemeinsamen relativen Häufigkeitsfunktion werden zum jeweiligen Rand hin aufsummiert. Daraus ergeben sich die relativen Randhäufigkeiten.







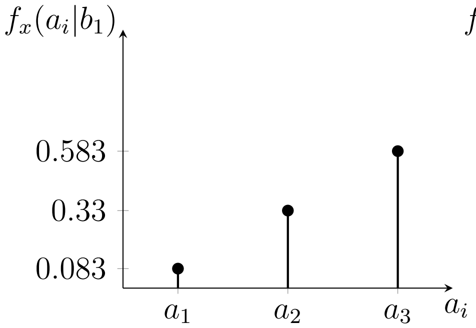


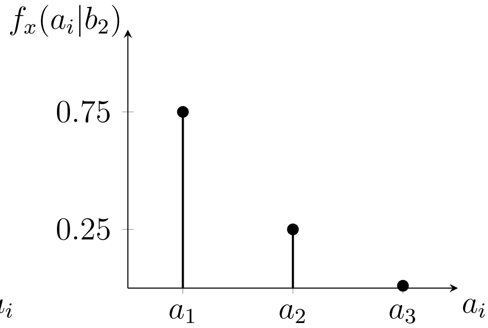
f) Skizzieren Sie die relative Häufigkeitsfunktion von X,

   i) bedingt auf die Studierenden, die sich vor dem Tutorium mit dem Vorlesungsstoff beschäftigt haben und

   ii) bedingt auf die Studierenden, die dies nicht getan haben

   in einem Stabdiagramm. Beurteilen Sie eine mögliche empirische Abhängigkeit zwischen den Merkmalen X und Y.



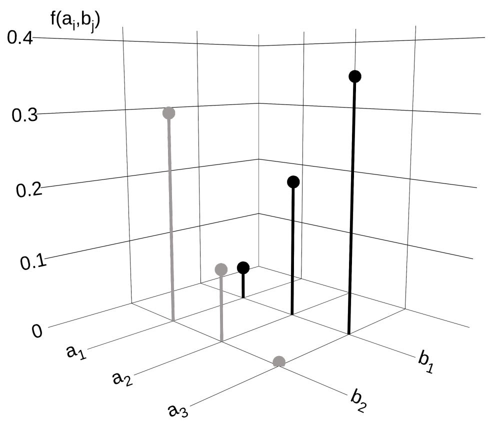


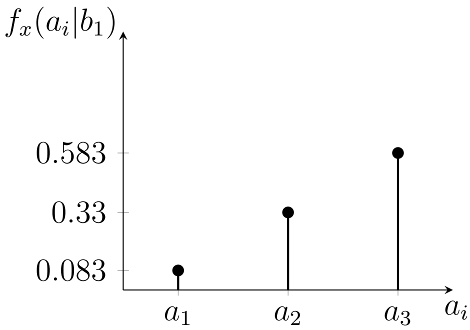
Die bedingten relativen Häufigkeitsverteilungen von X unterscheiden sich deutlich, je nachdem, ob $Y = b_1$ oder $Y = b_2$ gilt. Dies zeigt, dass X von der Ausprägung von Y abhängt, also X und Y *empirisch nicht unabhängig* sind.

g) Vergleichen Sie die in f) skizzierten bedingten relativen Häufigkeitsfunktionen mit der gemeinsamen relativen Häufigkeitsfunktion grafisch.

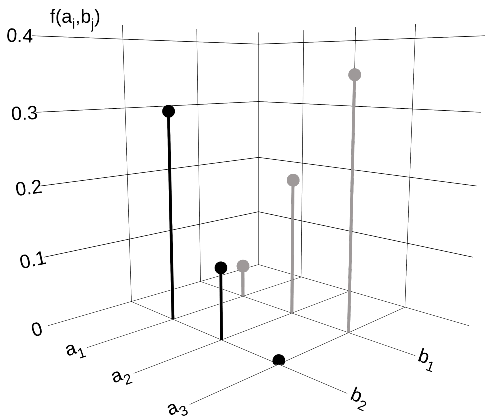
Unterschied: Die bedingten relativen Häufigkeiten bilden jeweils eine diskrete Häufigkeitsverteilung; f_{11}, f_{21}, f_{31} nicht (analog für f_{12}, f_{22}, f_{32}).

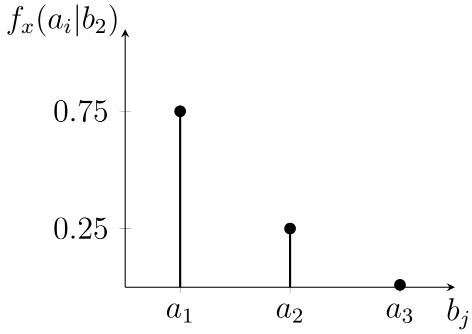
Gemeinsamkeit: Die Proportionalität zwischen den Ausprägungen bleibt erhalten. $\rightarrow$ Siehe Skizzen.





Fortsetzung zu g):





Teil C

h) Vervollständigen Sie die folgende Kontingenztabelle, indem Sie die (tatsächlichen) gemeinsamen absoluten Häufigkeiten $h(a_i, b_j) := h_{ij}$ eintragen. Berechnen Sie außerdem die (hypothetischen) absoluten Häufigkeiten $\tilde{h}_{ij}$, die Sie *unter der Annahme von empirischer Unabhängigkeit der Merkmale* X und Y erwarten würden.

X \ Y	ja (b_1)	nein (b_2)	$\sum$
schlecht (a_1)	10 / 42	60 / 28	70
mittelmäßig (a_2)	40 / 36	20 / 24	60
sehr gut (a_3)	70 / 42	0 / 28	70
$\sum$	120	80	200

i) Berechnen Sie den *Chi-Quadrat-Koeffizienten* χ^2. Geben Sie eine Intuition für die χ^2 Berechnungsformel.

Es gilt $\chi^2 := \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij}-\tilde{h}_{ij})^2}{\tilde{h}_{ij}}$, also

$$\begin{aligned}\chi^2 &= \frac{(10-42)^2}{42} + \frac{(60-28)^2}{28} + \frac{(40-36)^2}{36} + \frac{(20-24)^2}{24} + \frac{(70-42)^2}{42} + \frac{(0-28)^2}{28} \\ &= \frac{1024}{42} + \frac{1024}{28} + \frac{16}{36} + \frac{16}{24} + \frac{784}{42} + \frac{784}{28} \\ &\approx 108,73\end{aligned}$$

Intuition:
Die Abweichungen zwischen der beobachteten absoluten gemeinsamen Häufigkeit h_{ij} und der unter empirischer Unabhängigkeit erwarteten Häufigkeit $\tilde{h}_{ij}$ werden quadriert, damit sich positive und negative Abweichungen nicht gegenseitig aufheben. Anschließend werden die quadrierten Abweichungen $(h_{ij} - \tilde{h}_{ij})^2$ *relativ zur erwarteten Häufigkeit* $\tilde{h}_{ij}$ betrachtet. Diese Gewichtung ist notwendig, weil die Bedeutung einer Abweichung davon abhängt, wie groß die erwartete Häufigkeit unter Unabhängigkeit ist. Eine quadrierte Abweichung von $(h_{ij} - \tilde{h}_{ij})^2 = 5^2$ fällt beispielsweise *relativ stärker ins Gewicht*, wenn nur $\tilde{h}_{ij} = 10$ erwartet wird, als wenn die erwartete Häufigkeit $\tilde{h}_{ij} = 100$ beträgt. Mit anderen Worten: Dieselbe absolute Abweichung wirkt bei kleinen erwarteten Häufigkeiten stärker und bei großen erwarteten Häufigkeiten schwächer. Genau dieses Verhältnis wird in der Berechnung des χ^2-Werts berücksichtigt.

j) Berechnen Sie den *Kontingenzkoeffizienten* K und den *korrigierten Kontingenzkoeffizienten* K^* und interpretieren Sie Ihr Ergebnis im Kontext der Aufgabe.

$K := \sqrt{\frac{\chi^2}{\chi^2+n}}$, also $K := \sqrt{\frac{108,73}{108,73+200}} \approx 0,5935$

$K^* := \frac{K}{K_{max}}$ mit $K_{max} = \sqrt{\frac{m-1}{m}}$ und $m = min\{k, l\}$, also $K^* = \frac{0,5935}{\sqrt{\frac{2-1}{2}}} \approx 0.8393$

Der Wert von $K^* = 0.8393$ spricht für ein hohes Maß an empirischer Abhängigkeit zwischen den Merkmalen X und Y.