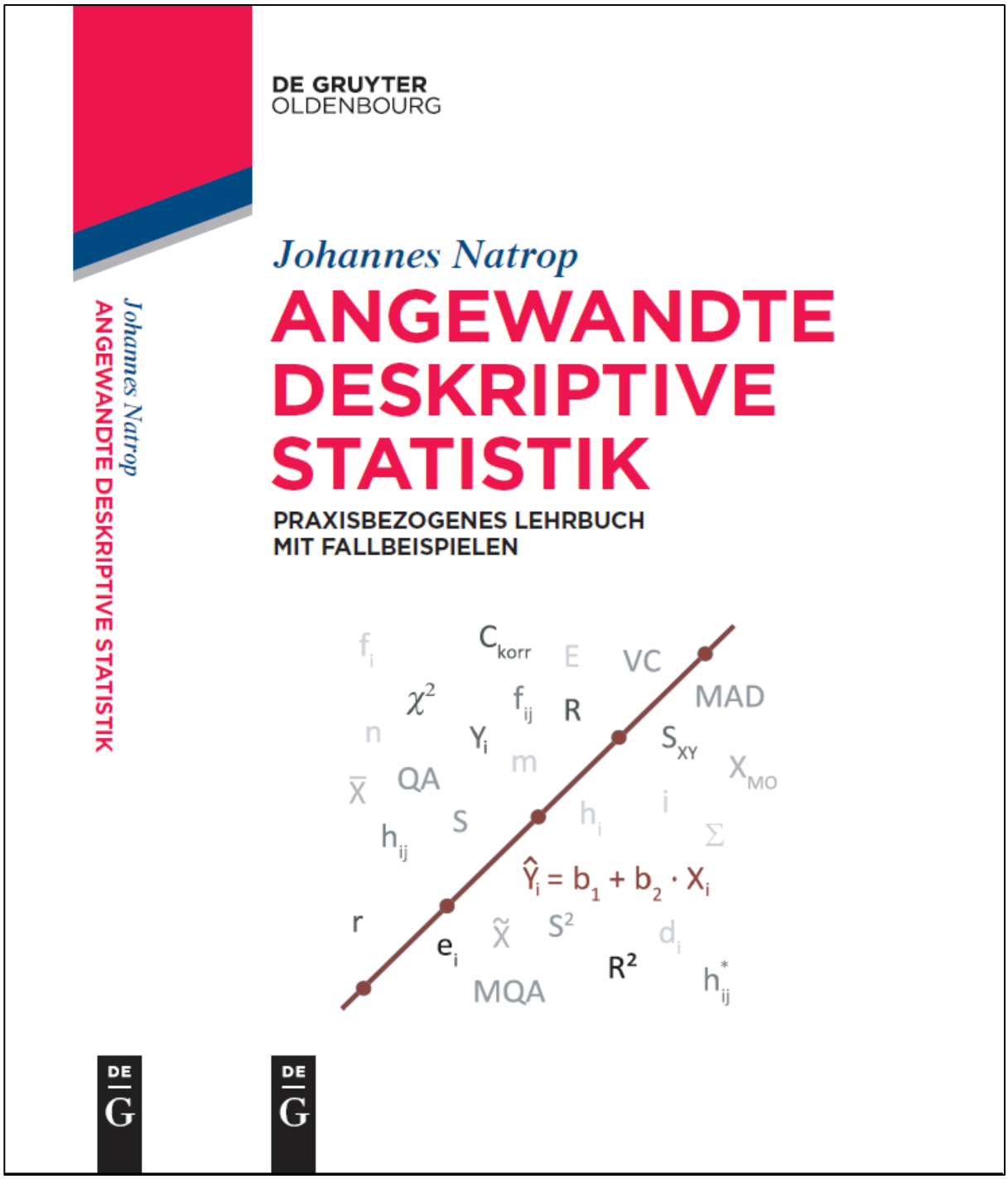
Prof. Dr. Johannes Natrop, Hochschule Bonn-Rhein-Sieg
E-Mail: Johannes.Natrop@H-BRS.de

Lösungen zu den Aufgaben (1 – 37) und den beiden Musterklausuren des nachfolgenden Lehrbuchs[1]

Stand: 01.11.2016

http://fb01.h-brs.de/wirtschaftsanktaugustinmedia/Angewandte_Deskriptive_Statistik_Loesungen.pdf



[1] Besonderer Dank gilt Herrn Thomas Neifer für die tatkräftige Unterstützung bei der Erstellung der Lösungen.

Lösungen zu den Aufgaben 1 – 37

Aufgabe 1: Photovoltaikanlage (S. 54)	
Merkmalsträger	Dünnschichtmodule
Merkmal	kWh je qm Modulfläche
Merkmalsausprägung	Konkret gemessene kWh je qm Modulfläche (Beobachtungswert)
Sachliche Abgrenzung i. e. S.	Zum Merkmalsträger: Wann liegt ein Dünnschichtmodul vor? Zum Merkmal: Wie ist die Modulfläche definiert? Weitere Aspekte der sachlichen Abgrenzung: exakt südliche Ausrichtung; beliebige Bauart zugelassen; Neigung der Solarmodule von 37
Räumliche Abgrenzung	Gebiet der Landeshauptstadt München
Zeitliche Abgrenzung	am 01.07.2014 zwischen 12.00 und 13.00 Uhr
Statistische Masse	alle betrachteten Merkmalsträger
Beobachtungswert	Merkmalswerte, beobachtete Merkmalsausprägungen

Aufgabe 2: Vielfalt des Weines (S. 55)

Geben Sie die Skalierung folgender Merkmale an und begründen Sie Ihre Wahl:

Merkmal	**Skalierung**	**Begründung**
Alkoholgehalt im Wein	Verhältnisskala	Definiert sind: Rangfolge („mehr – weniger“), Abstände und mathematischer Nullpunkt. Somit sind folgende Aussagen möglich: Ein Wein mit z.B. 10 % Alkoholgehalt enhält doppelt so viel Alkohol, wie ein Wein mit 5 % Alkohol (diese Aussage ist wegen der Existenz des absoluten Nullpunktes möglich). Ein Wein mit z. B. 10 % Alkohol enthält 5 %-Punkte mehr Alkohol als ein Wein mit 5 % Alkohol (diese Aussage ist möglich, da Abstände definiert sind).
Rebsorten	Nominalskala	Keine Rangfolge gegeben; Namen der Rebsorten wie Beaujolais, Chardonnay, Gewürztraminer, Grauburgunder, Müller-Thurgau, Riesling, Silvaner, Spätburgunder, Weißburgunder etc. stehen gleichberechtigt nebeneinander. Vergleiche im Sinne von „besser bzw. schlechter“ oder andere Vergleiche sind nicht möglich. Rechenoperationen sind ebenfalls nicht möglich.
Präferenz für Weine	Ordinalskala	Rangfolge gegeben, d. h. Aussagen im Sinne von „besser – schlechter“ sind möglich. Aber es sind keine Abstände definiert, d. h. es kann nicht gesagt werden, um wie viel besser oder schlechter die verschiedenen Weine sind.
Weinanbaugebiete	Nominalskala	Keine Rangfolge gegeben; Namen der Anbaugebiete wie „Ahr“, „Baden“, „Franken“, „Mosel“, „Rheingau“, „Rheinhessen“, „Sachsen“ etc. dienen nur der Identifizierung; keine Aussage im Sinne von „besser – schlechter“ gegeben.
Temperatur des Weines (°C)	Intervallskala	Hier sind Rangfolge und Abstände gegeben, aber ein mathematischer Nullpunkt ist nicht vorhanden (0 °C sind willkürlich gewählt). Damit lässt sich nur sagen, dass ein Wein, der eine Temperatur von 15 °C aufweist, 10 ° C wärmer ist als ein Wein, der eine Temperatur von 5 °C aufweist. Achtung: Relationen im Sinne von doppelt so warm etc. sind nicht möglich, da kein mathematischer Nullpunkt gegeben ist.

Aufgabe 3: Waldbrandschaden durch Selbstentzündung (S. 56)	
Merkmalsträger	Die betrachteten 11 Jahre (nicht die 5 Bundesländer)
Merkmal	**Schäden in Mio. €**
Merkmalsausprägung	**konkreter Schaden in Mio. €**
Sachliche Abgrenzung	• durch Selbstentzündung entstandene Waldbrandschäden • Wie ist „Selbstentzündung“ definiert? • Was ist ein „Waldbrand“? • Welcher Schaden ist entstanden?
Räumliche Abgrenzung	• die betrachteten 5 Bundesländer: • Wie sind die Grenzen der betrachteten 5 Bundesländer definiert?
Skalierung	Verhältnisskala; Nullpunkt gegeben: Ein Schaden von z.B. 88 Mio. € ist doppelt so viel wie ein Schaden von 44 Mio. €; zudem: Abstände gegeben

Aufgabe 4: Skalierung von Merkmalen (S. 56)		
Merkmal	Skala	Begründung
a) **Merkmal X**: Einkommen der Beschäftigten eines Unternehmens	Verhältnisskala	Gegeben sind: Rangfolge im Einkommen („mehr bzw. weniger“), Abstände und mathematischer Nullpunkt. Es lässt sich sagen, dass ein Beschäftigter mit 4 000 € Monatsgehalt doppelt so viel verdient, wie ein Beschäftigter mit 2 000 € Monatsgehalt (Aussage ist wegen der Existenz des absoluten Nullpunktes möglich). Ein Beschäftigter mit 4 000 € Monatsgehalt verdient 1 000 € mehr als ein Beschäftigter mit 3 000 € Monatsgehalt (Aussage ist möglich, da Abstände definiert sind).
Merkmal Y: Alter der Beschäftigten eines Unternehmens	Verhältnisskala	Gegeben sind: Rangfolge im Alter („älter bzw. jünger“), Abstände und mathematischer Nullpunkt. Es lässt sich sagen, dass ein Beschäftigter mit 40 Jahren doppelt so alt ist wie ein Beschäftigter mit 20 Jahren (Aussage ist wegen der Existenz des absoluten Nullpunktes, d. h. des Geburtszeitpunktes möglich). Ein Beschäftigter mit 40 Jahren ist 20 Jahre älter als ein Beschäftigter mit 20 Jahren (Aussage ist möglich, da Abstände definiert sind).
b) **Merkmal X**: Verschiedene Güteklassen eines Konsumgutes	Ordinalskala	Rangfolge der Güteklassen („besser bzw. schlechter“) gegeben, aber keine Abstände und kein mathem. Nullpunkt quantifizierbar. Mögliche Aussage: „Ein Konsumgut der Güterklasse (AA+) ist besser als ein Konsumgut der Güterklasse (A).“ Vergleichende Aussagen, wie z. B. „das eine Konsumgut ist um (…) besser als das andere Konsumgut“ oder „ das eine Konsumgut ist doppelt so gut wie das andere Konsumgut“, sind nicht möglich.
Merkmal Y: Preis des Konsumgutes	Verhältnisskala	Gegeben sind: Rangfolge der Preise („mehr bzw. weniger“), Abstände und mathematischer Nullpunkt (vgl. Ausführungen zu Einkommen in Beispiel a).
c) **Merkmal X**: Studiendauer von Hochschulabsolventen der BWL	Verhältnisskala	Gegeben sind: Rangfolge der Studiendauer („länger bzw. kürzer“), Abstände und mathematischer Nullpunkt (vgl. Ausführungen zum Einkommen in Beispiel a).
Merkmal Y: Einkünfte der Studierenden (z. B. BAFÖG, Erwerbstätigkeit, Unterstützung durch Angehörige)	Nominalskala	Keine Rangfolge der Einkunftsarten der Studierenden; die Einkunftsarten stehen gleichberechtigt nebeneinander; Bezeichnungen der Einkunftsarten dienen nur der Identifizierung; keine Aussage im Sinne von „besser bzw. schlechter“ möglich; Abstände und Nullpunkt sind damit auch nicht gegeben.

Aufgabe 5: Insolvenzstatistik 2005 (S. 70)		
a) Begriffe „Merkmalsträger, Merkmal, Merkmalsausprägung“		
Merkmals-träger	**insolvente** Unternehmen (Hinweis: Wichtig ist der Zusatz, dass es sich um insolvente Unternehmen handelt).	**Begründung**: n = 39 213 insolvente Unternehmen werden im Hinblick auf ihre Eigenschaft, d. h. im Hinblick auf das Merkmal „Rechtsform“ untersucht;
Merkmal	Rechtsform	
Merkmals-ausprägung	**konkret vorliegende** Rechtsform (EUN, PG, GmbH, AG, SR)	

b) Häufigkeitstabelle			
i	**X_i**	**h_i**	**f_i**
1	EUN	16 299	0,4156
2	PG	3 071	0,0783
3	GmbH	18 938	0,4830
4	AG	415	0,0106
5	SR	490	0,0125
Σ		**39 213**	**1,0000**

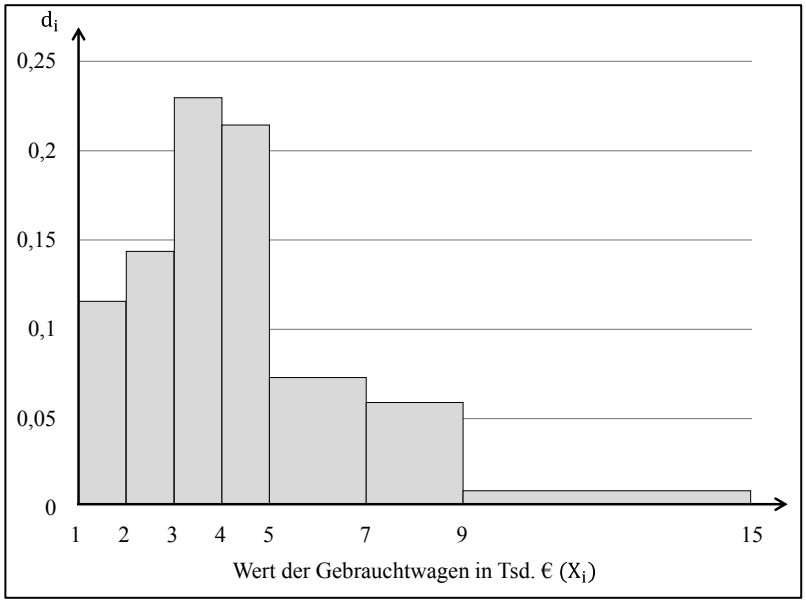
c) Sachliche Abgrenzung der Begriffe und Skalierung	
Die sachliche Abgrenzung i. e. S. dient der Abgrenzung der Merkmalsträger und der Merkmalsausprägungen sowie weiterer Tatbestände der Erhebung. Die Abgrenzung soll die Merkmalsträger im Hinblick auf das Untersuchungsmerkmal eindeutig identifizieren und sicherstellen, dass die statistische Erhebung und Auswertung der Merkmalsträger der Abgrenzung des Fragestellers entspricht.	
Sachl. Abgrenzung der Begriffe	**Abgrenzungsfrage**
Merkmalsträger	Was ist ein Unternehmen? Wann liegt ein insolventes Unternehmen vor?
Merkmalsausprägung	Wann sind die Rechtsformen „EUN“, „PG“, … etc. gegeben?
Skalierung Merkmalsträger	
Es liegt eine Nominalskala vor. Alle Merkmalsausprägungen (hier: konkret vorliegende Rechtsformen) stehen gleichberechtigt nebeneinander. Z. B. ist die Rechtsform „EUN“ nicht besser oder schlechter als die Rechtsform „PG“. Damit weisen die Rechtsformen keine Rangfolge auf. Ist diese Rangfolge nicht definiert, lassen sich auch keine Abstände und kein mathem. Nullpunkt bestimmen.	

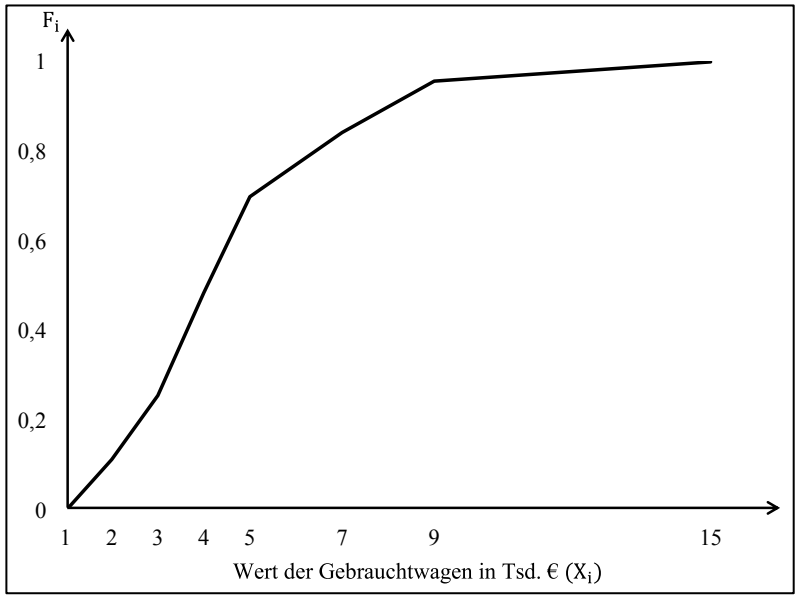
Aufgabe 6: Häufigkeitstabelle zum Wertbestand eines Gebrauchtwagenlagers (S. 96)						
i	**X_i**	**ΔX_i**	**h_i**	**f_i**	**F_i**	**d_i *)**
1	1 bis unter 2	1	8	0,1143	0,1143	0,1143
2	2 bis unter 3	1	10	0,1428	0,2571	0,1428
3	3 bis unter 4	1	16	0,2286	0,4857	0,2286
4	4 bis unter 5	1	15	0,2143	0,7000	0,2143
5	5 bis unter 7	2	10	0,1428	0,8428	0,0714
6	7 bis unter 9	2	8	0,1143	0,9571	0,0572
7	9 bis unter 15	6	3	0,0429	1,0000	0,0072
Σ			**70**	**1,0000**		
***)** Als Normklassenbreite wurde $\Delta X^n = 1$ gewählt, da $\mathbf{\Delta X_i = 1}$ am häufigsten vorkommt.						

Hinweise zur Häufigkeitstabelle der Aufgabe 6:

$$\Delta X^n = 1; \quad d_5 = \frac{f_5}{\Delta X_5} \cdot \Delta X^n = \frac{0{,}1428}{2} \cdot 1 = 0{,}0714; \quad d_6 = \frac{f_6}{\Delta X_6} \cdot \Delta X^n = \frac{0{,}1143}{2} \cdot 1 = 0{,}0572$$

$$d_7 = \frac{f_7}{\Delta X_7} \cdot \Delta X^n = \frac{0{,}0429}{6} \cdot 1 = 0{,}0072$$

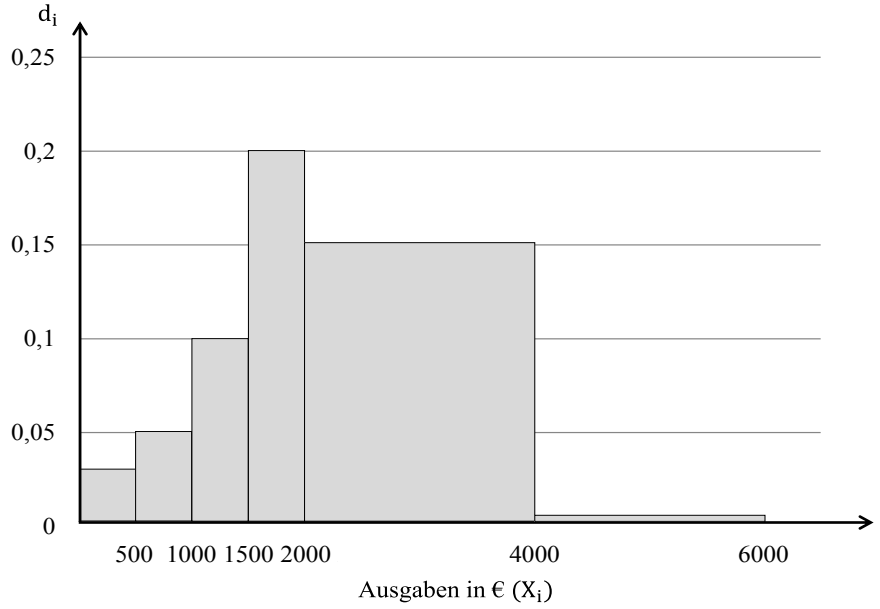




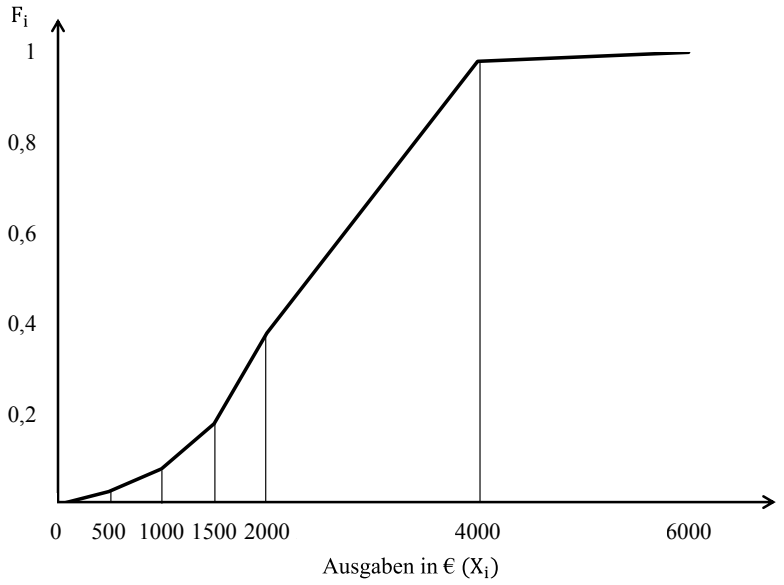
Aufgabe 7a: Häufigkeitsverteilung der PKW-Reparaturausgaben (S. 96/97)						
i	**X_i**	**h_i**	**f_i**	**F_i**	**ΔX_i**	**d_i**
1	0 bis unter 500	30	0,03	0,03	500	0,030
2	500 bis unter 1 000	50	0,05	0,08	500	0,050
3	1 000 bis unter 1 500	100	0,10	0,18	500	0,100
4	1 500 bis unter 2 000	200	0,20	0,38	500	0,200
5	2 000 bis unter 4 000	600	0,60	0,98	2 000	0,150
6	4 000 bis unter 6 000	20	0,02	1,00	2 000	0,005
	Σ	**1 000**	**1,00**			

$$\Delta X^n = 500; \quad d_5 = \frac{f_5}{\Delta X_5} \cdot \Delta X^n = \frac{0{,}60}{2\,000} \cdot 500 = 0{,}15; \quad d_6 = \frac{f_6}{\Delta X_6} \cdot \Delta X^n = \frac{0{,}02}{2\,000} \cdot 500 = 0{,}005$$

Histogramm der relativen Häufigkeitsverteilung der Reparaturausgaben



Verteilungsfunktion der Reparaturausgaben



Aufgabe 7b)

$$F(X \leq 3\,000) = 0{,}38 + 0{,}60 \cdot \frac{3\,000 - 2\,000}{2\,000} = 0{,}68$$

68 % der PKW-Besitzer tätigen Ausgaben für Reparaturen/Inspektionen in Höhe von 3 000 € oder weniger (alternativ: von höchstens 3 000 €).

$$F(X \leq 1\,750) = 0{,}18 + 0{,}20 \cdot \frac{1\,750 - 1\,500}{500} = 0{,}28; \quad F(X > 1\,750) = 1 - 0{,}28 = 0{,}72$$

72 % der PKW-Besitzer geben mehr als 1 750 € für Reparaturen/Inspektionen aus.

Natrop, Johannes: Angewandte Deskriptive Statistik. Praxisbezogenes Lehrbuch mit Fallbeispielen, Berlin u.a., 2015.

Aufgabe 7c)

Bei metrisch skalierten Merkmalen ist die Anzahl der vorliegenden Merkmalsausprägungen (insb. bei stetigen Merkmalen) häufig so groß, dass die Häufigkeitsverteilung sehr unübersichtlich ausfällt oder aufgrund der Vielzahl der Merkmalsausprägungen (MA) nicht sinnvoll gebildet werden kann (ggfs. kommt jede MA nur einmal vor, d. h. weist eine absolute Häufigkeit von „eins“ auf). Aus diesem Grund müssen benachbarte Merkmalsausprägungen zu Klassen zusammengefasst werden. Die Zusammenfassung hat so zu erfolgen, dass eine gegebene Übersichtlichkeit mit möglichst wenig Informationsverlust verbunden ist: Denn durch die Klassifizierung lässt sich nicht mehr erkennen, wie sich die Häufigkeiten auf die Merkmalswerte der Klassen verteilen. Damit sich dieser Informationsverlust in Grenzen hält, sind u. a. drei „Regeln“ zur optimalen Klassenbildung zu berücksichtigen (zu näheren Einzelheiten s. Ausführungen auf S. 80 ff im Buch):

- Es sollten nach Möglichkeit nur homogen besetzte Intervalle, d. h. Merkmalswerte mit ähnlich großer Häufigkeit zusammengefasst werden.
- Der häufigste Wert sollte nach Möglichkeit in der Klassenmitte liegen (Klassenmitte als Repräsentant der Klasse).
- Wenn möglich sollte eine einheitliche Klassenbreite gewählt werden. Dies ist häufig aber nicht möglich, da die Informationsdichte bei den verschiedenen Merkmalswerten unterschiedlich ausfällt. Liegt eine hohe Informationsdichte vor, d. h. konzentrieren sich die Merkmalswerte auf bestimmte Merkmalsintervalle und weisen diese zudem stärker schwankende Häufigkeiten auf (inhomogene Häufigkeitsverteilung), so ist für diese Bereiche eine kleine Klassenbreite zu wählen, um die vielen unterschiedlichen Informationen auch möglichst genau zu erfassen. Demgegenüber sind in Bereichen mit wenigen Merkmalswerten, die eventuell auch noch mit einheitlicher Häufigkeit auftreten, nur wenige Klassen mit größerer Klassenbreite zu berücksichtigen. Streuen die Merkmalswerte sehr stark in einem großen Intervall und liegen hier nur wenige Merkmalswerte vor, so sind zudem offene Randklassen zu wählen (vgl. hierzu S. 87 im Buch).

Aufgabe 8: Median der Personenzahl in Privathaushalten 2013 (S. 123)

Zum Modus:
Der Modus ist bei nicht klassifizierten Daten derjenige Merkmalswert, der mit der größten absoluten oder relativen Häufigkeit vorkommt. Die Merkmalsausprägung ($X_1 = 1$) weist mit einer absoluten Häufigkeit von $h_1 = 16{,}176$ Mio. Haushalten die größte Häufigkeit auf;
somit gilt: $\mathbf{X_{Mo} = 1}$ (da häufigster Wert).

Hinweis 1: Der Modus ist exakt und eindeutig anzugeben und darf nicht mit der Häufigkeit verwechselt werden. **Fehlerhaft** wäre eine Angabe, die auf die Häufigkeit als Modus abstellt (häufiges „Fettnäpfchen“ = Modalwert der Fettnäpfchen in Klausuren). Der Modus ist ein Mittelwert, so dass es sich bei ihm nur um einen Merkmalswert, nicht aber um eine Häufigkeit handeln kann.
Hinweis 2: Liegen klassifizierte Daten mit unterschiedlichen Klassenbreiten vor, kann der Modus nur über die Dichte anstelle der Häufigkeit gebildet werden. Zum Begriff der Dichte s. die Ausführungen im Buch auf S. 88 ff.

Zum Median:
Der Median beschreibt die Merkmalsausprägung des mittleren Merkmalsträgers der geordneten Urliste. In diesem Bsp. liegt eine H.V. mit einer sehr hohen Beobachtungszahl (n) vor. Daher kann der Median über $F_i = 0{,}5$ bzw. $H_i = \frac{n}{2}$ berechnet werden. Für das vorliegende Beispiel gilt:
$H_i = \frac{n}{2} = \frac{39\,933}{2} = 19\,966{,}5$ für $X = 2$. Daraus folgt: $X_{Me} = 2$ Pers. je Haushalt, da bei der 2. Merkmalsausprägung $X = 2$ die Hälfte der Merkmalsträger erfasst ist und damit $F_i = 0{,}5$ erreicht wird. 50 % der Personen in Privathaushalten leben in Haushalten mit 2 und weniger Personen und 50 % der Privatpersonen leben in Haushalten mit 2 und mehr Personen.

Aufgabe 9: PKW-Autovermietung (S. 130)

Können die Merkmalswerte des Merkmals Y als lineare Funktion der Merkmalswerte des Merkmals X mit den Parametern (a) und (b) dargestellt werden, gilt also ($Y = a + b \cdot X$), so lässt sich über das arithmetische Mittel von X direkt auch das arithmetische Mittel für Y ermitteln.

Wird nämlich das arithmetische Mittel des Merkmals X in die oben angeführte lineare Funktion eingesetzt, so ergibt sich daraus das arithmetische Mittel des Merkmals Y. Somit:
$\overline{Y} = a + b \cdot \overline{X}$

Diese Eigenschaft des arithmetischen Mittels einer linear transformierten Größe wird in der Statistik zur Ableitung verschiedener Formeln benötigt. (Hinweis: So leitet sich hieraus z. B. die Eigenschaft ab, dass die Regressionsfunktion durch den Schwerpunkt der Punktwolke verläuft; vgl. hierzu die Ausführungen im Kapitel 6.4 des Buches; S. 287 f. Analog lässt sich später in modifizierter Form diese Eigenschaft des arithmetischen Mittels auch auf die Eigenschaft der Varianz und der Standardabweichung einer aus der Variablen (Merkmal) X linear hervorgegangenen Größe Y übertragen (vgl. S. 182 f); das Verständnis der intuitiv einsichtigen Eigenschaft des arithmetischen Mittels erleichtert das Verständnis der Eigenschaft der Varianz bei linear transformierten Werten).

In diesem Beispiel lautet die lineare Funktion: $Y = 20 + 0{,}20 \cdot X$
Diese Funktion ermittelt die Einnahmen der Autovermietung (Merkmal Y) in Abhängigkeit von den zurückgelegten Kilometern (Merkmal X). Die PKW-Mieter fahren durchschnittlich 200 km ($\overline{X} = 200$); somit kann aufgrund der linearen Beziehung von X und Y aus dem arithmetischen Mittel der zurückgelegten PKW-Strecke auf das arithmetische Mittel der PKW-Mieteinnahmen geschlossen werden, denn es gilt:
$\overline{Y} = 20 + 0{,}20 \cdot \overline{X} = 20 + 0{,}20 \cdot 200 = 60\ €$
Die Autovermietung würde durchschnittlich Einnahmen $\overline{Y}$ in Höhe von $\overline{Y} = 60\ €$ je PKW erzielen.

Aufgabe 10: Umsatzrenditen von zwei Unternehmen (S. 134)

Die beiden UN stehen gleichberechtigt nebeneinander (2 Teilgesamtheiten), d. h. es liegt hier eine additive Verknüpfung der betrachteten Merkmalswerte vor. Daher ist zur Ermittlung der durchschnittlichen Umsatzrendite des Gesamtunternehmens das arithmetische Mittel heranzuziehen. Dabei erfolgt die Gewichtung mit den relativen Anteilen des Nenners der betrachteten Größe. Hier wird die Umsatzrendite betrachtet. Diese lautet:

$$\text{Umsatzrendite} = \frac{\text{Gewinn}}{\text{Umsatz}} \cdot 100$$

Der Nenner der Umsatzrendite enthält den Umsatz. Somit ist Umsatzrendite der beiden Unternehmen mit den Umsatzanteilen der jeweiligen Unternehmen zu gewichten. Die Unternehmung U_1 hat einen Umsatzanteil von 75 %, die Unternehmung U_2 hat einen Umsatzanteil von 25 %. Damit ergibt sich für das gewogene arithmetische Mittel der Umsatzrenditen beider Unternehmen:
$\overline{X} = 1{,}1\ \% \cdot 0{,}75 + 6{,}2\ \% \cdot 0{,}25 = 2{,}375\ \%$

Ergebnis: Im Jahr 2010 betrug die durchschnittliche Umsatzrendite des Gesamtunternehmens 2,375 %.
(Hinweis: Die Umsatzrendite des Gesamtunternehmens lässt sich auch ermitteln, indem die Absolutwerte für die Gewinne und die Umsätze beider Unternehmen jeweils addiert und dann gemäß der Definition der Umsatzrendite dividiert werden. Soll die durchschnittliche Umsatzrendite des fusionierten Unternehmens jedoch „auf die Schnelle“ in einer „Überschlagsrechnung“ anhand der Renditen der Einzelunternehmen beurteilt werden, müssen die heranzuziehenden Gewichte bekannt sein. Da hier mit den Umsatzanteilen gewichtet wird und die Unternehmung 1 das wesentlich höhere Umsatzgewicht aufweist, muss die gesuchte Umsatzrendite des Gesamtunternehmens eher in der Nähe der Rendite des Unternehmens 1, als in der Nähe der Rendite des Unternehmens 2 liegen).

Aufgabe 11: Durchschnittspreis für Obst (S. 143)

Im Folgenden soll die Einkaufsmenge an Äpfeln und Birnen durch die Größen X_1 bzw. X_2 erfasst werden. Die Einkaufsmenge an Obst wird durch die Größe X dargestellt.

- Die Äpfel- bzw. Birnenpreise werden durch die Symbole $P(X_1)$ und $P(X_2)$ abgebildet.
- Im vorliegenden Beispiel betragen der Apfelpreis $P(X_1) = 2\ €/kg$ und der Birnenpreis $P(X_2) = 1\ €/kg$.
- Der durchschnittliche Preis für das Obst soll durch das Symbol $\overline{P(X)}$ bei Verwendung des arithmetischen Mittels bzw. $\overline{P(X)}_H$ bei Verwendung des harmonischen Mittels dargestellt werden.

A. Ermittlung des durchschnittlichen Obstpreises unter Verwendung des arithmetischen Mittels:

Der Durchschnittspreis für Obst $\overline{P(X)}$ lässt sich über das gewogene arithmetische Mittel der Äpfel- und Birnenpreise ermitteln. Da der Obstpreis die Definition (€ je kg) aufweist, wären zur Bestimmung des Durchschnittspreises des eingekauften Obstes die Preise von Äpfeln und Birnen mit den Anteilen (f_i) des Nenners, d. h. mit den Gewichtsanteilen (in kg) der eingekauften Äpfel und Birnen zu gewichten (Gewichtung mit den relativen Anteilen des Nenners; hier: kg).

Im vorliegenden Beispiel wurden Äpfel und Birnen für einen Ausgabenbetrag von jeweils 10 € eingekauft. Da sich die Einkaufsmenge jeweils als Ausgabe/Einkaufspreis errechnet, ergibt sich für die Einkaufsmenge der Äpfel (X_1): $X_1 = 10\ [€]\ /\ 2\ [€/kg] = 5$ kg Äpfel. Analog errechnet sich für die Einkaufsmenge der Birnen (X_2) eine Einkaufsmenge von $X_1 = 10\ [€]\ /\ 1\ [€/kg] = 10$ kg Birnen. Insgesamt wurden somit 15 kg Obst (X) eingekauft. Die Gewichtsanteile der Äpfel betragen ($f_1 = 5/15 = 1/3$) und die Gewichtsanteile der Birnen betragen ($f_2 = 10/15 = 2/3$).

Unter Verwendung dieser Mengengewichte errechnet sich der Durchschnittspreis des Obstes als gewogenes arithmetisches Mittel:

Durchschnittlich (Ø) gezahlter Obstpreis

$$\overline{P(X)} = P(X_1) \cdot f_1 + P(X_2) \cdot f_2 = 2\ € \cdot \frac{1}{3} + 1\ € \cdot \frac{2}{3} = 1{,}34\ €/kg.$$

B. Ermittlung des durchschnittlichen Obstpreises unter Verwendung des harmonischen Mittels:

Allerdings sind in dieser Aufgabe nur die Anteile des Zählers, d. h. die Ausgabenanteile bekannt, und die Anteile des Nenners mussten zuvor ermittelt werden. Sollen stattdessen die bekannten Anteile des Zählers (Ausgabenanteile) als Gewichte Verwendung finden, so ist nicht das arithmetische Mittel, sondern das **harmonische Mittel** zur Berechnung des Durchschnittspreises zu verwenden. (Gewichtung mit den relativen Anteilen (f_i) des Zählers, d. h. den Ausgabeanteilen in €).

Da jeweils für 10 € Äpfel und Birnen eingekauft wurden, betragen die Ausgabenanteile für Äpfel bzw. Birnen jeweils $10/20 = \frac{1}{2}$. Werden diese Ausgabeanteile in der Formel für das harmonische Mittel als Gewichte verwendet, ergibt sich für den durchschnittlich gezahlten Obstpreis:

$$\overline{P(X)}_H = 1/(\frac{1}{P(X_1)} \cdot f_1 + \frac{1}{P(X_2)} \cdot f_2) = 1/(\frac{1}{2} \cdot \frac{10}{20} + \frac{1}{1} \cdot \frac{10}{20}) = 1{,}34\ €/kg$$

Der Durchschnittspreis des Obstes beträgt somit auch bei Verwendung des harmonischen Mittels 1,34 €/kg.

Aufgabe 12: Mittelwerte im Vergleich (S. 158 bzw. S. 96/97)

Zum Modus:

In dieser Aufgabe liegen klassifizierte Daten mit unterschiedlichen Klassenbreiten vor: Daher kann der Modus nur unter Verwendung der Dichte bestimmt werden. Die größte Dichte liegt in Klasse $(i = 4)$ mit $(d_4 = 0{,}2)$. Die Klassenmitte dieser Klasse mit der größten Dichte stellt den Modus dar. Somit gilt:

$X_{Mo} = 1\,750\ €$

Zum Median:

„Feinberechnung" des Median bei klassifizierten Daten:

$$X_{Me} = 2\,000 + 2\,000 \cdot \frac{0{,}5 - 0{,}38}{0{,}98 - 0{,}38} = 2\,400\ €$$

Zum arithmetischen Mittel:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m} X'_i \cdot h_i \quad \text{mit:} \sum_{i=1}^{m} X'_i \cdot h_i = 2\,420\,000 \text{ (s. Vorgabe)}$$

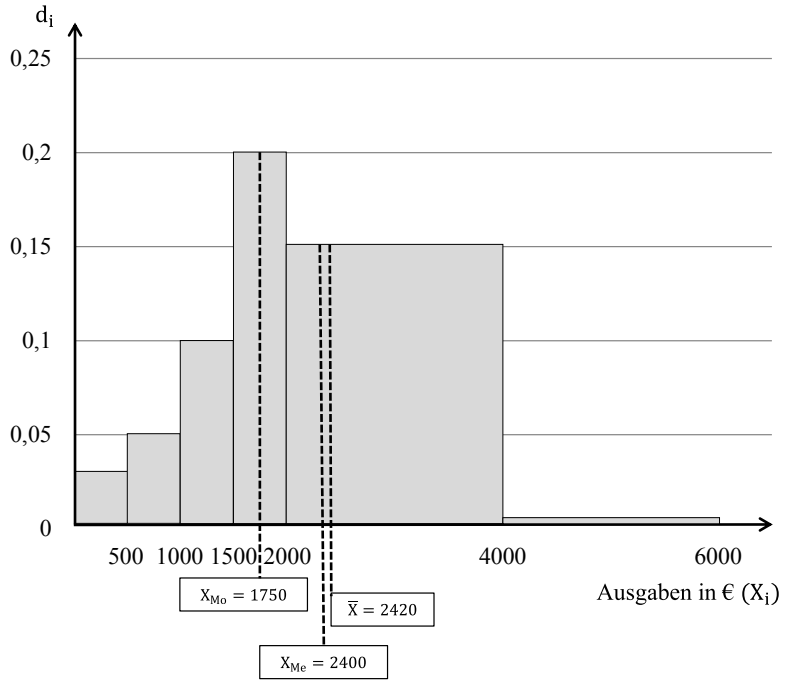
$$\overline{X} = \frac{1}{1\,000} \cdot 2\,420\,000 = 2\,420\ €$$

Vergleich der Mittelwerte (Fechnersche Lageregel):

Mithilfe der Fechnerschen Lageregel lässt sich über die Schiefe der H.V. folgende Aussage treffen: $(X_{Mo} = 1\,750\ €) < (X_{Me} = 2\,400\ €) < (\overline{X} = 2\,420\ €)$

Somit liegt eine **linkssteile** und **rechtsschiefe** Häufigkeitsverteilung vor (siehe hierzu auch das Histogramm der Aufgabe 7, das nachfolgend unter Einbeziehung der Mittelwerte nochmals dargestellt wird).

Histogramm der relativen Häufigkeitsverteilung der Reparaturausgaben



Aufgabe 13: Durchschnittliche Gewinnentwicklung im Zeitablauf (S. 158)

Aufgabe 13a)

Die durchschnittliche prozentuale Gewinnentwicklung der Jahre 2009 – 2013 ist multiplikativ über Wachstumsfaktoren verknüpft. Daher ist zur Berechnung der durchschnittlichen prozentualen Gewinnentwicklung $\overline{W_G}$ im Gesamtzeitraum 2009 – 2013 das geometrische Mittel heranzuziehen. In diesem Beispiel sind die absoluten Gewinne der einzelnen Jahre bekannt (Absolutwerte gegeben). Daher lässt sich die durchschnittliche prozentuale Wachstumsrate der Gewinnentwicklung $\overline{W_G}$ ermitteln als:

$$\overline{W_G} = \left[\left(\frac{\text{Endwert}}{\text{Anfangswert}}\right)^{\frac{1}{n}} - 1\right] \cdot 100$$

Die Größe (n) gibt hierbei die Anzahl der Wachstumsfaktoren an. Hier liegen n = 4 Wachstumsfaktoren vor; somit ergibt sich:

$$\overline{W_G} = \left[\left(\frac{35}{12}\right)^{\frac{1}{4}} - 1\right] \cdot 100 = 30{,}68\ \%$$

Ergebnis: Der Gewinn ist durchschnittlich um 30,68 % p. a. gestiegen.

Aufgabe 13 b)

Der Gewinnzuwachs ist auch in diesem Beispiel multiplikativ über die Wachstumsfaktoren verknüpft, wobei nun Wachstumsraten und nicht Absolutwerte der Gewinnentwicklung gegeben sind. Die Wachstumsraten sind zunächst in vier Wachstumsfaktoren WF_i für die $i = 1, \ldots, 4$ betrachteten Jahre umzuwandeln. Somit ergibt sich:

$WF_1 = 1{,}5;\ WF_2 = 1{,}5;\ WF_3 = 1{,}5;\ WF_4 = 0{,}9$

Hieraus ermittelt sich für den Gesamtzeitraum der folgende durchschnittliche (Ø) gesamte Wachstumsfaktor $\overline{X}_G$: $\overline{X}_G = \sqrt[4]{1{,}5 \cdot 1{,}5 \cdot 1{,}5 \cdot 0{,}9} = 1{,}3202$

Der Ø Wachstumsfaktor beträgt 1,3202. Wird dieser Wachstumsfaktor wieder in die Ø Wachstumsrate $\overline{W_G}$ umgerechnet, ergibt sich: $\overline{W_G} = (1{,}3202 - 1) \cdot 100 = 32{,}02\ \%$

Ein Vergleich der Ergebnisse von Unternehmen C und Unternehmen C-Ultra zeigt, dass das Unternehmen C-Ultra im gesamten Zeitraum mit einer durchschnittlichen jährlichen Wachstumsrate von $\overline{W_G} = 32{,}02\ \%$ p. a. ein höheres durchschnittliches Gewinnwachstum p. a. erzielen konnte als Unternehmen C mit $\overline{W_G} = 30{,}68\ \%$.

Aufgabe 14: Durchschnittliche Wachstumsrate des Umsatzes im Zeitablauf (S. 158)

Die Wachstumsfaktoren der Umsatzentwicklung bauen multiplikativ aufeinander auf, so dass zur Ermittlung der Ø jährlichen Wachstumsrate das geometrische Mittel zu verwenden ist.

Für **UN A** sind Absolutwerte der Umsatzentwicklung vorgegeben; somit ermittelt sich die durchschnittliche jährliche Wachstumsrate der Umsatzentwicklung wie folgt:

$$\overline{W_G} = \left[\left(\frac{\text{Endwert}}{\text{Anfangswert}}\right)^{\frac{1}{n}} - 1\right] \cdot 100 = \left[\left(\frac{329}{200}\right)^{\frac{1}{3}} - 1\right] \cdot 100 = 18{,}05\ \%$$

mit: n = Anzahl der Wachstumsfaktoren = 3

Für **UN B** sind für n = 3 Jahre die Wachstumsraten der jährlichen Umsatzentwicklung vorgegeben. Diese Wachstumsraten sind zunächst in drei Wachstumsfaktoren WF_i für die $(i = 1, \ldots, 3)$ betrachteten Jahre umzuwandeln. Somit ergibt sich:

$WF_1 = 1{,}2;\ WF_2 = 0{,}95;\ WF_3 = 1{,}516;$

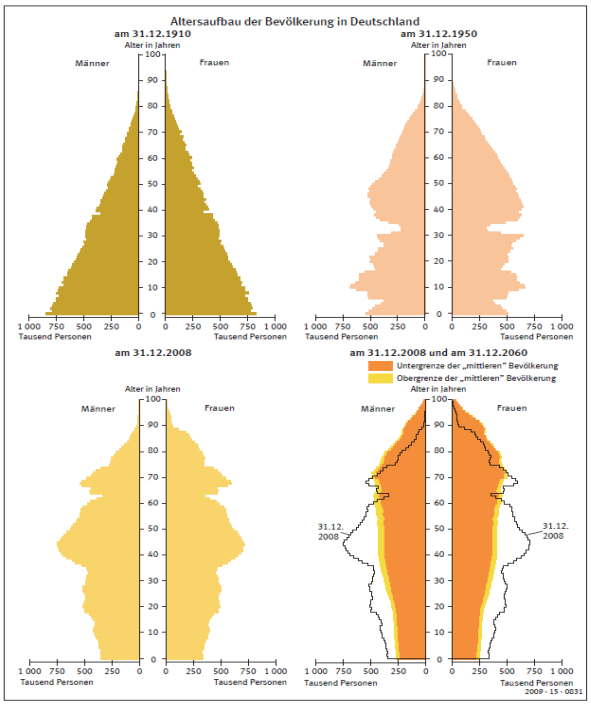
Hieraus ermittelt sich für den Gesamtzeitraum der folgende durchschnittliche gesamte Wachstumsfaktor $\overline{X}_G$: $\overline{X}_G = \sqrt[3]{1{,}2 \cdot 0{,}95 \cdot 1{,}516} = 1{,}2001$

Fortsetzung Aufgabe 14: Der durchschnittliche Wachstumsfaktor 1,2001 muss noch in die durchschnittliche Wachstumsrate $\overline{W_G}$ zurückgerechnet werden.
Es ergibt sich: $\overline{W_G} = (1{,}2001 - 1) \cdot 100 = 20{,}01\ \%$
Unternehmen B hat mit 20,01 % p. a. eine höhere durchschnittliche Wachstumsrate des Umsatzes erzielt als Unternehmen A mit nur 18,05 % p. a.

Aufgabe 15: Demographische Alterung im Blickpunkt (S. 159)

In Abb. II-3-9 werden das **Medianalter** und das **Durchschnittsalter** von Frauen und Männern im Zeitablauf dargestellt. Es zeigt sich, dass bei beiden Geschlechtern in der Vergangenheit das Medianalter stets niedriger ausfiel als das Durchschnittsalter. Zu Beginn dieses Jahrtausends hat sich die relative Höhe beider Altersmittelwerte umgedreht und zwar zunächst bei den Männern (im Jahr 2002) und dann bei den Frauen (im Jahr 2008). Wenn die Altersverteilung der Männer und Frauen jeweils einer unimodalen Häufigkeitsverteilung unterliegt, bedeutet der Wechsel der Höhe beider Mittelwerte, dass gemäß der Fechnerschen Lageregel die Altersverteilung von einer linkssteilen (rechtsschiefen) H.V. (mit: $\overline{X} > X_{Me} > X_{Mo}$) in eine rechtssteile (linksschiefe) H.V. umschwenkt (mit: $\overline{X} < X_{Me} < X_{Mo}$). Bis zur Jahrtausendwende lag somit eine linkssteile Altersverteilung vor, bei der das Modalalter geringer ausfiel als das Medianalter und dieses wiederum kleiner war als das Durchschnittsalter. Dies bedeutet, dass die meisten Menschen (Modus) ein jüngeres Alter aufwiesen. Seit der Jahrtausendwende hat sich diese Reihenfolge der Höhe der Mittelwerte durch die demographische Alterung verändert. Denn rückläufige Geburtenraten haben allmählich bei gleichzeitig ansteigender Lebenserwartung die Altersverteilung der Bevölkerung von einer links- in eine rechtssteile H.V. umgewandelt. Nun stellt ein höheres Alter die häufigste Merkmalsausprägung (Modus) dar, so dass das Modalalter das Medianalter übertrifft und dieses wieder höher ausfällt als das Durchschnittsalter.

Dieser Wechsel der Altersstruktur wird durch die nachfolgende Bevölkerungspyramide nochmals deutlich. Allerdings verlief in der Vergangenheit die Häufigkeitsverteilung des Alters nur annähernd unimodal, so dass die Rangfolge der Höhe der Mittelwerte nur eingeschränkt auf die Schiefe der Altersstruktur schließen lässt. Für die prognostizierten Zahlen der Zukunft (12. koordinierte Bevölkerungsmodellrechnung der Bundesregierung) liegt allerdings ein weitgehend (geglätteter) unimodaler Verlauf vor mit einem Modalwert von gut 70 Jahren (!):



Quelle: Stat. Bundesamt: Bevölkerung Deutschlands bis 2060, 12. Koordinierte Bevölkerungsvorausberechnung, Wiesbaden 2009, Schaubild 3, S. 15.

Hinweis zur Aufgabe 15:

- Das Medianalter teilt die Bevölkerung von der Altersstruktur in zwei gleich große Teile. 50 % der Bevölkerung sind älter als das Medianalter und 50 % sind jünger. In Folge des demographischen Wandels nimmt das Medianalter durch das abnehmende Geburtenniveau i. V. m. der durch verbesserte Lebensbedingungen ansteigenden Lebenserwartung stetig zu.

- Das Durchschnittsalter ermittelt sich als gewogenes arithmetisches Mittel, wobei ab der Jahrtausendwende die wenigen, sehr jungen Altersgruppen dazu beitragen, dass das Durchschnittsalter geringer ausfällt als das Median- und das Modalalter.

Aufgabe 16: Verwendung des arithmetischen Mittels (S. 159)

Diese Aussage kann so nicht getroffen werden. Die Verwendung der jeweiligen Mittelwerte wird durch die Verknüpfung der Merkmalswerte bestimmt und unterliegt keiner Gestaltungsmöglichkeit. Das arithmetische Mittel ist bei additiver Verknüpfung der Merkmalswerte zu verwenden, während das geometrische Mittel bei multiplikativer Verknüpfung der Merkmalswerte zum Einsatz kommt.

Aufgabe 17: Fragen Sie Ihren Statistiker oder fahren Sie in die Werkstatt (S. 159)

Der Spritverbrauch je km wird durch folgende Größe definiert:

$$\text{Spritverbrauch je km} = \frac{\text{verbrauchte Liter}}{\text{zurückgelegte Kilometer}}$$

- Der durchschnittliche Spritverbrauch einer Gesamtstrecke wird als gewogenes arithmetisches Mittel berechnet. Dabei erfolgt die Gewichtung der Spritverbräuche der Teilstrecken mit den Anteilen des Nenners der Definition, d. h. mit den jeweils zurückgelegten Kilometeranteilen.
- Schwankt der durchschnittliche Verbrauch zu Beginn der Fahrt noch sehr stark, so stabilisiert sich dieser Ø Spritverbrauch mit der zurückgelegten Kilometerzahl. Dies ist darauf zurückzuführen, dass mit zunehmender Kilometerzahl das hohe Gewicht der bereits zurückgelegten Strecke in den Spritverbrauch einfließt, so dass der aktuelle Spritverbrauch wegen des niedrigen Gewichts der aktuell zurückgelegten Strecke immer weniger ins Gewicht fällt und den Durchschnittswert immer weniger prägt.
- Dies soll im Folgenden mit zwei konkreten Zahlenbeispielen verdeutlicht werden: Es sei zunächst angenommen, dass ein PKW nur eine relativ kurze Strecke von insgesamt **5 km** zurücklegt, wobei auf den ersten **4 km** der Gesamtstrecke konstant **niedrige** Verbrauchswerte auftreten und auf dem restlichen **1 km** der Strecke der Verbrauchswert **außergewöhnlich hoch** ist. Bei der Berechnung des Durchschnittsverbrauchs für die zurückgelegten 5 km fließen nun die extrem hohen Verbrauchswerte des letzten Kilometers mit einem Gewicht von 1/5 in den Gesamtverbrauch ein. Daher wird für diesen zuletzt zurückgelegten Kilometer der angenommene hohe Verbrauch sich stark auf den Durchschnittsverbrauch auswirken und diesen ansteigen lassen.
- Anders verhält es sich, wenn die Analyse für eine kurze Fahrtstrecke nach einer bereits länger zurückgelegten Fahrtstrecke von **500 km** erfolgt: Für dieses Szenario sei angenommen, dass für eine zuerst zurückgelegte Strecke von beispielsweise **499 km** konstant **niedrige** Verbrauchswerte aufgetreten sind, bevor auf dem **letzten Kilometer** der momentane Spritverbrauch verkehrsbedingt **stark ansteigt**. Der sehr hohe Spritverbrauch des letzten Kilometers würde nun aber nur mit einem sehr kleinen Gewicht von $(1/500)$ in den Durchschnittsverbrauch einfließen. Dies hat zur Folge, dass trotz des zuletzt hohen Spritverbrauchs der durchschnittliche Verbrauch so gut wie keine Veränderung erfährt.

Aufgabe 18: Armutsschwelle und ihre Veränderung (S. 159)

Die Armutsschwelle definiert das Einkommen, ab dem eine Person als arm bezeichnet wird. Es handelt sich hierbei um eine relative und nicht um eine absolute Armutsschwelle. Als relativ arm wird eine Person bezeichnet, wenn sie weniger als eine bestimmte Prozentzahl (50 % oder 60 %) vom mittleren Einkommen aufweist. Als mittleres Einkommen werden das Durchschnittseinkommen oder das Medianeinkommen herangezogen.

- Wird die Armutsschwelle am **arithmetischen Mittel** ausgerichtet, indem z. B. alle Personen als arm gelten, die weniger als 50 % des Durchschnittseinkommens aufweisen, so sinkt die Armutsschwelle, sobald eine Person mit hohem Einkommen weniger verdient. Denn durch das geringere Einkommen der Einkommensstarken nimmt das Durchschnittseinkommen ab, so dass bei unverändertem Einkommen der Armen der Anteil derjenigen Personen abnimmt, die weniger als 50 % des Durchschnittseinkommens aufweisen. Daher sinkt die Armutsquote aufgrund der Einkommensrückgänge der gut Verdienenden, auch wenn bei den Armen keine Veränderungen eintreten. Dies hat zur Folge, dass weniger Arme unter die Armutsschwelle fallen, obwohl sich deren Einkommenssituation nicht verändert hat.
- Wird die Armutsschwelle hingegen am **Medianeinkommen** ausgerichtet, so ändert sich das Medianeinkommen und damit die Armutsquote solange nicht, wie das Einkommen der einkommensstärksten 50 % nicht unter das Medianeinkommen fällt. Kommt es also zu Einkommensschwankungen in der oberen Hälfte der Einkommensbezieher, bleibt das Medianeinkommen unverändert, so dass bei unverändertem Einkommen der armen Bevölkerung (untere Hälfte der Einkommensbezieher) auch die Armutsquote sich nicht verändert.
  Damit die Armutsquote unter Verwendung des Median sinkt, muss somit ein sehr hoher Einkommensbezieher so starke Einkommensverluste aufweisen, dass sein Einkommen in die untere Hälfte der Einkommensbezieher fällt.

Diese „Tücken" der Armutsquote entstehen, weil diese als **relative** Armut und nicht als **absolute** Armut definiert wird.

Aufgabe 19: MAD im Einzelhandel (S. 172)

Aufgabe 19a)

Der Median ermittelt sich als Merkmalsausprägung des mittleren Merkmalsträgers der geordneten Urliste. Daher müssen die Merkmalswerte nach ihrer Größe geordnet werden, bevor der Median bestimmt werden kann. Es ergibt sich folgende Reihenfolge der geordneten 11 Merkmalswerte:

2; 3; 4; 5; 6; 7; 8; 9; 9; 10; 11

Da in diesem Beispiel ein kleines und ungerades (n) vorliegt, wird der Median über folgende Formel gebildet:

$$X_{Me} = X_{[(n+1)/2]} = X_{[(11+1)/2]} = X_{[6]} = 7$$

(Hinweis: die 6. Merkmalsausprägung der geordneten Reihe weist den Merkmalswert 7 auf).

Somit: $X_{Me} = 7$

Aufgabe 19b)

$$\mathrm{MAD}(X_{Me}) = \frac{1}{11} \cdot (|2-7| + |3-7| + |4-7| + |5-7| + |6-7| + |7-7| + |8-7| + |9-7| + |9-7| + |10-7| + |11-7|) = 2{,}45 \text{ (in Tsd.) €}$$

Interpretation: Im Durchschnitt weichen die 11 Merkmalswerte der Werbeausgaben um 2 450 € vom Medianwert der Werbeausgaben von 7 000 € nach unten bzw. oben ab.

Bei dieser Ermittlung der MAD handelt es sich um die ausführliche Version der Berechnung der MAD für Einzelwerte. Die Berechnung lässt sich auch über eine „Kurzversion" darstellen. Es gilt:

$$\begin{aligned}\mathrm{MAD}(X_{Me}) &= \frac{1}{11} \cdot (|2-7| + |3-7| + |4-7| + |5-7| + |6-7| + |7-7| + |8-7| + |9-7| + |9-7| + |10-7| + |11-7|) \\ &= \frac{1}{11} \cdot [|2+3+4+5+6-5\cdot 7|] + [|8+9+9+10+11-5\cdot 7|] \\ &= \tfrac{1}{11} \cdot [|-15| + |12|] = 2{,}45 \text{ €}\end{aligned}$$

Somit gilt für die Kurzversion der Berechnung (Angaben in 1 000 €):

$$\mathrm{MAD}(X_{Me}) = \frac{1}{11} \cdot [|2+3+4+5+6-5\cdot 7| + |8+9+9+10+11-5\cdot 7|] = 2{,}45 \text{ €}$$

Aufgabe 20: Varianz und Standardabweichung (S. 180, S. 172, S. 158, S. 96/97)

$$S^2 = \frac{1}{n}\sum_{i=1}^{m} X_i'^2 \cdot h_i - \overline{X}^2 \quad \text{mit:} \sum_{i=1}^{m} X_i'^2 \cdot h_i = 6\,698\,750\,000$$

$$S^2 = \frac{1}{1\,000} \cdot 6\,698\,750\,000 - 2\,420^2 = 842\,350 \text{ €}^2$$

$$S = \sqrt{842\,350} = 917{,}7963 \text{ €}$$

Aufgabe 21: Varianz und Standardabweichung (S. 187)

Zur Lösung s. Tabelle III-A-2 im Anhang auf S. 341 im Buch.

Ergebnis:

- Strategie 2 weist zwar einen leicht höheren Erwartungswert des Ertrages von 680 € anstelle von 660 € aus, d. h. einen um (+ 20 €) höheren Erwartungswert.
- Dafür fällt das Risiko der Anlagestrategie 2 aber bei einer Varianz von 225 600 €^{2} deutlich höher aus als bei Anlagestrategie 1 mit einer Varianz von 36 400 €^{2}.
- Welche Strategie daher zu realisieren ist, hängt von der Risikoneigung ab. Ein risikoscheuer Anleger wird Strategie 1 wählen, ein risikofreudiger Anleger wird sich für Strategie 2 entscheiden.

Aufgabe 22: Varianz und Standardabweichung (S. 187)

i	X_i	h_i	f_i	F_i	$X_i \cdot h_i$	$\lvert X_i - X_{Me}\rvert \cdot h_i$	$\lvert X_i - \bar{X}\rvert \cdot h_i$	${X_i}^2 \cdot h_i$
1	0	9	0,09	0,09	0	13,5	14,4	0
2	1	41	0,41	0,50	41	20,5	24,6	41
3	2	35	0,35	0,85	70	17,5	14	140
4	3	11	0,11	0,96	33	16,5	15,4	99
5	4	4	0,04	1,00	16	10	9,6	64
	Σ	**100**	**1,00**		**160**	**78**	**78**	**344**

Modus X_{Mo}:

Hier liegen nicht klassifizierte Daten vor. Der Modus stellt daher die häufigste Merkmalsausprägung dar. Mit einer relativen Häufigkeit von 41 % kommt die Ausprägung $X_2 = 1$ am häufigsten vor. Somit gilt: $\mathbf{X_{Mo} = 1}$ **Statistikbuch** (häufigster Merkmalswert mit $f_2 = 41\ \%$)

Median X_{Me}:

Hier liegt eine gerade Beobachtungszahl nicht klassifizierter Merkmalswerte vor. Daher wird der Median ermittelt als:

$$X_{Me} = 0{,}5 \cdot \left(X_{[100/2]} + X_{[(100/2)+1]}\right) = 0{,}5 \cdot \left(X_{[50]} + X_{[51]}\right) = 0{,}5 \cdot (1 + 2) = 1{,}5$$

Somit: $\mathbf{X_{Me} = 1{,}5}$ **Statistikbücher**

Anhand dieses Beispiels lässt sich erkennen, dass der Median bei einer kleinen Beobachtungszahl (n) nicht exakt über die Definition ($F_i = 0{,}5$) ermittelt werden kann. Dies würde in diesem Beispiel zu einem ungenauen Ergebnis ($X_{Me} = 1$) führen, da der Median in der Mitte der H.V. empfindlich auf Schwankungen reagiert. Die Mitte wird definiert durch den 50. und 51. Merkmalswert. Der 50. Merkmalswert beträgt „1"; der 51. Merkmalswert beträgt „2". Die Ermittlung des Medians über ($F_i = 0{,}5$) kann daher nur dann Verwendung finden, wenn aufgrund einer großen Beobachtungszahl (n) die Merkmalswerte in der Mitte der H.V. nicht variieren. (Wichtiger Hinweis: Die Ermittlung des Medians über die Formel $X_{Me} = 0{,}5 \cdot \left(X_{[n/2]} + X_{[(n/2)+1]}\right)$, d. h. über die Verwendung des arithmetischen Mittels (bei gerader Anzahl n) ist nur dann möglich, wenn es sich beim Merkmal X mindestens um ein intervallskaliertes und nicht nur um ein ordinalskaliertes Merkmal handelt).

Arihmetisches Mittel $\bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{m} X_i \cdot h_i = \frac{1}{100} \cdot 160 = \mathbf{1{,}6}\ \textbf{Statistikbücher}$$

Berechnung der MAD für eine H.V.:

$$\mathrm{MAD}(X_{Me}) = \frac{1}{n}\sum_{i=1}^{m} \left|X_i - \tilde{X}\right| \cdot h_i = \frac{1}{100} \cdot 78 = \mathbf{0{,}78}\ \textbf{Statistikbücher}$$

$$\mathrm{MAD}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{m} |X_i - \bar{X}| \cdot h_i = \frac{1}{100} \cdot 78 = \mathbf{0{,}78}\ \textbf{Statistikbücher}$$

Die durchschnittliche Abweichung der Merkmalswerte vom Median ($X_{Me} = 1{,}5$) bzw. arithmetischen Mittel ($\bar{X} = 1{,}6$) beträgt (übereinstimmend) 0,78 Statistikbücher.

Hinweis:
Die MAD stimmt in diesem Beispiel für beide Mittelwerte (X_{Me}) und ($\bar{X}$) überein. Dies stellt eine Ausnahme bei diskreten Merkmalswerten dar und ist darauf zurückzuführen, dass sowohl beim Median als auch beim arithmetischen Mittel 50 % der Merkmalsträger den jeweiligen Mittelwert oder einen kleineren Wert aufweisen. Ebenso weisen 50 % den jeweiligen Mittelwert oder einen höheren Wert auf. Somit bilden sowohl der Median als auch das arithmetische Mittel übereinstimmend den Merkmalswert in der Mitte der Häufigkeitsverteilung ab (bei stetigen Merkmalswerten kann diese Situation für verschiedene Mittelwerte nicht eintreten).

Fortsetzung Hinweis von Aufgabe 22:

Veränderungen des Mittelwertes haben solange keinen Einfluss auf die MAD, wie sich dieser Mittelwert in der mittleren Position der H. V. befindet und gleichviele Merkmalsträger einen kleineren oder größeren Merkmalswert als den Mittelwert aufweisen. Dass dieser Sachverhalt so eintritt, kann anhand der vereinfachten Formel für die Berechnung der MAD aufgezeigt werden: So zeigt ein Blick auf die letzte Zeile der Aufgabenlösung zu Aufgabe 19b, dass der Mittelwert keinen Einfluss auf die MAD hat, wenn eine gleich hohe Anzahl von Merkmalsträgern einen Merkmalswert aufweist, der kleiner als der Mittelwert bzw. größer als der Mittelwert ist (in dem Beispiel wird jeweils 5 mal der Medianwert „7“ von den Merkmalswerten abgezogen, die größer bzw. kleiner als der Median sind). Zudem ist Folgendes zu beachten: Da der Median sich jeweils in der Mitte der H.V. befindet, haben unterschiedliche (d. h. verzerrte Medianwerte), die auf unterschiedliche bzw. ungenaue Berechnungsweisen zurückgehen, keinen Einfluss auf die MAD. (Würde beispielsweise anstelle des Medians 1,5 ein Medianwert von 1,0 verwendet, so ergäbe sich für die MAD ebenfalls ein unveränderter Wert). Diese besondere Situation, dass verschiedene Werte für die Mittelwerte keinen Einfluss auf die MAD ausüben, kann eher bei diskreten Merkmalswerten, kaum aber bei stetigen Merkmalswerten eintreten. Bei stetigen Merkmalen haben unterschiedliche Mittelwerte zur Folge, dass die Anzahl der Merkmalswerte vor und nach den verschiedenen Mittelwerten nicht übereinstimmen wird und damit die MAD zu unterschiedlichen Ergebnissen gelangt, wie die vereinfachte Formel für die MAD erkennen lässt.)

Ergebnisse für die Varianz und die Standardabweichung.

$$S^2 = \frac{1}{n}\sum_{i=1}^{m} X_i^2 \cdot h_i - \overline{X}^2 = \frac{1}{100}\sum_{i=1}^{4} X_i^2 \cdot h_i - \overline{X}^2 = \frac{1}{100} \cdot 344 - 1{,}6^2 = \mathbf{0,88}\ (\textbf{Statistikbücher})^2$$

$S = \sqrt{0{,}88} = 0{,}9381$ Statistikbücher. Die Standardabweichung beträgt **0,9381 Statistikbücher**.

Standardabweichung und Varianz lassen sich inhaltlich nicht interpretieren. Die Standardabweichung stellt nicht die durchschnittliche Abweichung dar (wie die MAD), sondern es ist nur eine technische Interpretation im Sinne der Rechenformel möglich.

Aufgabe 23: Z-Transformation (S. 188)

Die Größe Z sei wie folgt definiert: $Z = \frac{X - \overline{X}}{S_X}$

Die Größe Z (Variable Z, Merkmalswert von Z) wird als Z-Standardisierung bezeichnet. Sie geht über eine lineare Transformation aus der Größe X (Variable X, Merkmalswert von X) hervor. Es lässt sich zeigen, dass Z immer ein arithmetisches Mittel von „null“ und eine Varianz/Standardabweichung von „eins“ aufweist, d. h. $\overline{Z} = 0$; $S_Z = 1$.

Hinweis: Diese beiden Eigenschaften sind von zentraler Bedeutung in vielen statistischen Anwendungen, z. B. bei der Überführung normalverteilter Merkmalswerte in sogenannte standardnormalverteilte Merkmalswerte in der Schließenden Statistik. Aber auch bei den multivariaten Verfahren werden wegen dieser beiden Eigenschaften ($\overline{Z}$= 0; $S_Z = 1$) immer wieder standardisierte Werte verwendet. Formaler Beweis, dass $\overline{Z} = 0$:

$$Z = \frac{X - \overline{X}}{S_X} = \frac{X}{S_X} - \frac{\overline{X}}{S_X} = \frac{1}{S_X} \cdot (X - \overline{X})$$

Aufgrund der Eigenschaft des arithmetischen Mittels von linear transformierten Werten gilt für das arithmetische Mittel $\overline{Z}$:

$\overline{Z} = \frac{1}{S_X} \cdot (\overline{X} - \overline{X}) = 0$ (wird in die lineare Transformationsbeziehung Z = F(X) für X das arithm. Mittel eingesetzt, so ergibt sich für Z ebenfalls das arithmetische Mittel, das zugleich „null“ beträgt).

Fortsetzung Aufgabe 23:

Formaler Beweis, dass $S_Z = 1$:

Aufgrund der Z-Standardisierung lässt sich Z auch formulieren als:

$$Z = \frac{X - \overline{X}}{S_X} = -\frac{\overline{X}}{S_X} + \frac{1}{S_X} \cdot X$$

Hinweis: Das arithmetische Mittel $\overline{X}$ und die Standardabweichung S_X stellen fest vorgegebene Größen (sogenannte Konstanten) dar.

Wird die Konstante $-\frac{\overline{X}}{S_X}$ als "a" und die Konstante $\left(\frac{1}{S_X}\right)$ als "b" bezeichnet, so gilt:

$Z = a + b \cdot X$

Stellt Z eine lineare Transformation der Größe X dar, so gilt für die Berechnung der Varianz der Größe Z:

$S_Z^2 = b^2 \cdot S_X^2$ (zur Varianz von linear transformierten Größen s. die Ausführungen auf S. 182 ff)

Damit ergibt sich unter Verwendung von $b = \frac{1}{S_X}$

$$S_Z^2 = \left(\frac{1}{S_X}\right)^2 \cdot S_X^2 = \frac{1}{S_X^2} \cdot S_X^2 = 1$$

Damit besitzt Z eine Varianz $S_Z^2 = 1$ und eine Standardabweichung von $S_Z = 1$

Aufgabe 24: „1 – 20 – 2 im Skatspiel“ (S. 229)

Herz-Ass wird durch die Merkmalskombination (X_2, Y_8) dargestellt.

Für die relative Häufigkeit von „Herz-Ass“ gilt: $f_{28} = \frac{1}{32}$

Für die relative Randhäufigkeit von „Herz“ gilt: $f_{2.} = \frac{1}{4}$

Die theoretisch erwartete relative Häufigkeit von „Herz-Ass“ kann auch über den Laplace-Wahrscheinlichkeitsbegriff abgebildet werden:

$$\text{Wahrscheinlichkeit (W)} = \frac{\text{Anzahl der günstigen Elementarereignisse}}{\text{Anzahl der möglichen Elementarereignisse}}$$

Somit lautet die theoretisch erwartete relative Häufigkeit bzw. die Wahrscheinlichkeit für „Herz-Ass “:

$$W(\text{Herz} - \text{Ass}) = \frac{\text{Anzahl der günstigen Elementarereignisse}}{\text{Anzahl der möglichen Elementarereignisse}} = \frac{1}{32}$$

Die theoretisch erwartete relative Häufigkeit von „Herz-Ass“ wird mittels des Multiplikationssatzes der Wahrscheinlichkeit bei Unabhängigkeit bestimmt, indem die relativen Randhäufigkeiten von „Herz“ bzw. „Ass“ miteinander multipliziert werden:

Somit: $W(\text{Herz} - \text{Ass}) = f_{28}^* = f_{2.} \cdot f_{.8} = \frac{1}{4} \cdot \frac{1}{8} = \frac{1}{32}$

Bedingte relative Häufigkeit von „Herz-Ass“, wenn bekannt ist, dass eine „Herz-Karte“ gezogen wurde:

$$f(Y_8/X_2) = \frac{h_{28}}{h_{2.}} = \frac{1}{8} = \frac{f_{28}}{f_{2.}} = \frac{1/32}{1/4} = \frac{1}{8}$$

Aufgabe 25: Verkehrsunfälle und Alkohol im Straßenverkehr (S. 230)

Aufgabe 25a)

Gemeinsame absolute Häufigkeiten (X = Uhrzeit; Y = Alkoholstatus); Angaben in 1 000			
Alkohol (Y) / Uhrzeit (X)	Nein	Ja	Gesamt
18 Uhr abends bis 4 Uhr morgens	58 197,72	8 509,58	66 707,30
ab 4 Uhr morgens bis 18 Uhr abends	218 934,24	5 463,46	224 397,70
Gesamt	277 131,96	13 973,04	291 105

Gemeinsame relative Häufigkeiten (X = Uhrzeit; Y = Alkoholstatus); Angaben in %			
Alkohol (Y) / Zeit (X)	Nein	Ja	Gesamt
18 Uhr abends bis 4 Uhr morgens	19,99	2,92	22,91
ab 4 Uhr morgens bis 18 Uhr abends	75,21	1,88	77,09
Gesamt	95,20	4,80	100

Aufgabe 25 b)

$$f(Y_2/X_2) = \frac{h_{22}}{h_{2.}} = \frac{5\,463{,}46}{224\,397{,}70} = 0{,}0243$$

Die bedingte relative Häufigkeit für einen Unfall unter Alkoholeinfluss unter der Bedingung, dass dieser zwischen 4 Uhr morgens und 18 Uhr abends geschieht, beträgt 2,43 %.

$$f(X_2/Y_2) = \frac{h_{22}}{h_{.2}} = \frac{5\,463{,}46}{13\,973{,}04} = 0{,}391$$

Die bedingte relative Häufigkeit für einen Unfall zwischen 4 Uhr morgens und 18 Uhr abends unter der Bedingung, dass dieser unter Alkoholeinfluss stattfand, beträgt 39,10 %.

Fortsetzung Aufgabe 25: Verkehrsunfälle und Alkohol im Straßenverkehr (S. 230)

Die Daten der Aufgabe 25 lassen sich in eine „Gesamtbetrachtung von absoluten, relativen und bedingten relativen Häufigkeiten der Merkmale X und Y einbinden. Das gesamte Zahlenwerk stellt sich wie nachfolgend dar:

Gemeinsame absolute Häufigkeiten in 2013 (X = Uhrzeit, Y = Alkoholstatus); Angaben in 1 000			
Alkohol (Y) / Zeit (X)	Nein	Ja	Gesamt
18 Uhr abends bis 4 Uhr morgens	58,198	8,510	66,708
ab 4 Uhr morgens bis 18 Uhr abends	218,934	5,463	224,397
Gesamt	277,132	13,973	291,105

1) 13,973 = 4,8 % von 291,105
2) Ohne Alkohol = 291,105 – 13,973 = 277,132
3) 21,0 % von 277,132 (ohne Alkohol) = 58,198 entfielen auf „18 Uhr abends – 4 Uhr morgens“
4) 60,9 % von 13,973 (mit Alkohol) = 8,510 entfielen auf „18 Uhr abends – 4 Uhr morgens“

Relative Häufigkeit der Unfälle in 2013 (in %)		
Alkohol (Y) / Zeit (X)	Nein	Ja
18 Uhr abends bis 4 Uhr morgens	19,992	2,923
ab 4 Uhr morgens bis 18 Uhr abends	75,208	1,877
Gesamt	95,200	4,800

Bedingte relative Häufigkeit für (Uhrzeit/Alkoholstatus) in %		
Alkohol (Y) / Zeit (X)	Nein	Ja
18 Uhr abends bis 4 Uhr morgens	21,00	60,90
ab 4 Uhr morgens bis 18 Uhr abends	79,00	39,10
Gesamt	100,00	100,00

Bedingte relative Häufigkeit für (Alkoholstatus/Uhrzeit) in %			
Alkohol (Y) / Zeit (X)	Nein	Ja	Gesamt
18 Uhr abends bis 4 Uhr morgens	87,24	12,76	100,00
ab 4 Uhr morgens bis 18 Uhr abends	97,57	2,43	100,00

Aufgabe 26: Varianz und Standardabweichung (S. 232/ S. 233)

Aufgabe 26 a)

$f_{22} = 0{,}10$	$f_{2.} = 0{,}30$
$f(Y_2/X_2) = \frac{h_{22}}{h_{2.}} = \frac{10}{30} = \frac{1}{3}$	$f(X_2/Y_2) = \frac{h_{22}}{h_{.2}} = \frac{10}{30} = \frac{1}{3}$
$h_{22}^{*} = \frac{h_{2.} \cdot h_{.2}}{n} = \frac{30 \cdot 30}{100} = 9$	$f_{22}^{*} = f_{2.} \cdot f_{.2} = 0{,}30 \cdot 0{,}30 = 0{,}09$
$\overline{X} = 0 \cdot 0{,}46 + 1 \cdot 0{,}30 + 2 \cdot 0{,}24 = 0{,}78$	$\overline{Y} = 0 \cdot 0{,}60 + 1 \cdot 0{,}30 + 2 \cdot 0{,}10 = 0{,}50$

$S_X^2 = (0^{\,2} \cdot 0{,}46 + 1^2 \cdot 0{,}30 + 2^2 \cdot 0{,}24) - 0{,}78^2 = 0{,}6516$

$S_X = \sqrt{0{,}6516} = 0{,}8072$

$S_Y^2 = (0^{\,2} \cdot 0{,}60 + 1^2 \cdot 0{,}30 + 2^2 \cdot 0{,}10) - 0{,}50^2 = 0{,}45$

$S_Y = \sqrt{0{,}45} = 0{,}6708$

Aufgabe 26 b)

Bei Unabhängigkeit der Ausfallhäufigkeit der Maschinen X und Y müssen die bedingten relativen Häufigkeiten des Merkmals X unter der Bedingung Y, also $f(X_i/Y_j)$ für $i = 1, \dots, 3$ und $j = 1, \dots, 3$ mit der relativen Randhäufigkeit $f(X_i)$ übereinstimmen. Analog müssen die bedingten relativen Häufigkeiten des Merkmals Y unter der Bedingung X, also $f(Y_j/X_i)$ für $i = 1, \dots, 3$ und $j = 1, \dots, 3$ mit der relativen Randhäufigkeit $f(Y_j)$ übereinstimmen.

So muss z. B. gelten, dass die bedingten relativen Häufigkeiten der ersten Ausprägung des Merkmals X (also X_1 = "Kein Ausfall") unter der Bedingung der verschiedenen Merkmalsausprägungen des Merkmals Y übereinstimmen. Zudem müssen diese bedingten Häufigkeiten mit der Randhäufigkeit der ersten Merkmalsausprägung des Merkmals X (also $f(X_1)$) übereinstimmen. (Hinweis: diese Bedingungen müssen für alle Ausprägungen des Merkmals X gelten. Zudem müssen sie für alle bedingten Häufigkeiten des Merkmals Y unter der Bedingung X gelten. Im Folgenden werden nur die bedingten Häufigkeiten $f(X_1/Y_r)$ (für $r = 1, \dots 3$) dargestellt.)

Somit muss u. a. bei Unabhängigkeit gelten: $f(X_1/Y_1) = f(X_1/Y_2) = f(X_1/Y_3) = f(X_1)$

Die bedingten relativen Häufigkeiten des Merkmals X unter der Bedingung der verschiedenen Ausprägungen des Merkmals Y betragen:

$$f(X_1/Y_1) = \frac{h_{11}}{h_{.1}} = \frac{30}{60} = 0{,}50;\ f(X_1/Y_2) = \frac{h_{12}}{h_{.2}} = \frac{14}{30} = 0{,}467;\ f(X_1/Y_3) = \frac{h_{13}}{h_{.3}} = \frac{2}{10} = 0{,}20$$

Die **<u>nicht</u> bedingte relative Häufigkeit, d.h. die relative Häufigkeit des Merkmals X** beträgt:

$f(X_1) = 0{,}46$

Die **bedingten relativen Häufigkeiten der ersten Ausprägung X_1 des Merkmals X unter der Bedingung der verschiedenen Ausprägungen des Merkmals Y** weichen voneinander ab und stimmen auch nicht mit der relativen Häufigkeit der ersten Ausprägung des Merkmals X überein:

Somit: $[f(X_1/Y_1) = 0{,}50] \neq [f(X_1/Y_2) = 0{,}467] \neq [f(X_1/Y_3) = 0{,}20] \neq [f(X_1) = 0{,}46]$
Damit ist die Bedingung für Unabhängigkeit der Merkmale X und Y nicht erfüllt.

Dass die Bedingung für Unabhängigkeit nicht erfüllt ist, kann auch daraus ersehen werden, dass die theoretische erwarteten und die empirisch beobachteten absoluten und relativen Häufigkeit in den einzelnen Merkmalskombinationen nicht übereinstimmen.

Aufgabe 27: Kovarianz der Ausfallhäufigkeit zweier Maschinen (S. 249, S. 232/233)

Auf Basis der Ergebnisse der Aufgabe 26 ermittelt sich folgende Kovarianz:

$$S_{XY} = (0 \cdot 0 \cdot 0{,}30 + 0 \cdot 1 \cdot 0{,}14 + 0 \cdot 2 \cdot 0{,}02 + 1 \cdot 0 \cdot 0{,}18 + 1 \cdot 1 \cdot 0{,}10 + 1 \cdot 2 \cdot 0{,}02 + 2 \cdot 0 \cdot 0{,}12 + 2 \cdot 1 \cdot 0{,}06 + 2 \cdot 2 \cdot 0{,}06) - (0{,}78 \cdot 0{,}50) = 0{,}11$$

- Das positive Vorzeichen der Kovarianz impliziert, dass eine positive lineare Abhängigkeit zwischen den Ausfallhäufigkeiten beider Maschinen bestehen könnte. Die Kovarianz weist folgende Dimension auf: (Ausfälle/Tag)².
- Allerdings kann von der Höhe der Kovarianz nicht auf die Stärke des Zusammenhangs der metrischen Merkmale X und Y geschlossen werden. Eine nähere Interpretation der positiven Kovarianz ist somit nicht gegeben. Es lässt sich mittels des Ergebnisses der Kovarianz von $S_{XY} = 0{,}11$ lediglich schließen, dass eine positive lineare Abhängigkeit ($S_{XY} > 0$) vorliegt. Die Stärke der positiven Abhängigkeit kann mit der mit den Standardabweichungen von X und Y normierten Kovarianz, d. h. mit dem Korrelationskoeffizienten nach Bravais-Pearson ermittelt werden. Gleichwohl ist auch dieses Ergebnis mit Vorsicht zu betrachten, da der Zufall diese positive Kovarianz bzw. Korrelation hervorgerufen haben kann. Die Auswirkungen des Zufalls auf das Ergebnis lassen sich letztlich nur mithilfe der schließenden Statistik beurteilen.

Aufgabe 28: Korrelationskoeffizient und graphische Darstellung (S. 259/ S. 260)

Aufgabe 28 a)

Es sei von zwei metrisch skalierten Merkmalen X und Y ausgegangen, für die die lineare Abhängigkeit über den Bravais-Pearson-Korrelationskoeffizienten beschrieben werden kann.

Situation I:

- Einem weitgehend unveränderten X-Wert können mehrere, stark schwankende Y-Werte zugeordnet werden. Es herrscht Unabhängigkeit, da sich die Merkmalswerte von Y auch dann verändern, wenn der Merkmalswert von X sich nicht verändert. In dieser Situation führt die lineare Unabhängigkeit zu einer Kovarianz von $S_{XY} = 0$ und damit auch zu einem Bravais-Pearson Korrelationskoeffizienten von r = 0.

Situation II:

- Bei steigendem X steigt auch Y, so dass zwischen beiden Merkmalswerten eine positive Abhängigkeit besteht und die Kovarianz ein positives Vorzeichen aufweist. Über den positiven Wert der Kovarianz ist keine Aussage möglich, da die Kovarianz keinen Maximalwert aufweist.
- Wird die Kovarianz in den Bravais-Pearson Korrelationskoeffizienten (r) überführt, so wird (r) einen positiven Wert im Intervall ($0 < r < 1$) aufweisen. Ein Korrelationskoeffizient r = 1 (perfekte positive lineare Abhängigkeit) ist nicht möglich, da nicht alle Punkte auf einer Geraden liegen. Erklärung: Lägen alle Punkte auf einer Geraden, könnte sich beispielsweise ein Y Wert nur dann verändern, wenn sich auch der X-Wert verändert. Weichen die Punkte von einer Geraden ab, hat dies zur Folge, dass bei Veränderungen des X-Wertes der Y-Wert sich unterproportional verändert; X und Y stehen somit in keiner perfekten linearen Beziehung zueinander.

Situation III

- Bei steigendem X sinkt Y. Somit liegt eine negative Abhängigkeit (Beziehung) zwischen X und Y vor. Kovarianz und Korrelationskoeffizient sind negativ. Im Beispiel liegt eine perfekte **nichtlineare** Beziehung zwischen X und Y vor: Die Größe Y verändert sich gemäß dieser nichtlinearen Abhängigkeit nur dann, wenn sich auch die Größe X verändert.
- Allerdings liegt **keine perfekte lineare** Beziehung vor. Würde durch die Punktwolke der Merkmalskombinationen eine Gerade gelegt, so würden die Merkmalskombinationen von X und Y nicht alle auf dieser Geraden liegen. Damit kann (r) nicht den Wert (r = –1) annehmen. Der Wert des Bravais-Pearson Korrelationskoeffizienten (r) würde im Intervall ($-1 < r < 0$) liegen.

Fortsetzung Aufgabe 28:

Aufgabe 28b:

Situation IV:

Für die Merkmale X und Y liegen zwei Teilgesamten vor. In jeder Teilgesamtheit sind jedem gegebenen Y-Wert mehrere X-Werte zugeordnet. Daher besteht in jeder Teilgesamtheit jeweils Unabhängigkeit zwischen den beiden Merkmalen X und Y. In diesem Beispiel soll das Merkmal X das Alter und das Merkmal Y das Körpergewicht wiedergeben. Dabei stellen die Männer die Teilgesamtheit mit den höheren Merkmalswerten X und Y dar:

Diese beschriebene Situation konnte z. B. häufiger in den letzten Jahrzehnten beobachtet werden, als Männer aufgrund des Wehrdienstes oder des Zivildienstes häufig das Studium später begannen als die Frauen. Wird zudem angenommen, dass Männer tendenziell ein höheres Gewicht als Frauen aufweisen, dann stellt sich die in Situation IV beschriebene Datenkonstellation ein: Die Männer (Teilgesamtheit mit den höheren Werten der Merkmale X und Y) waren zu Studienbeginn im Vergleich zu den Frauen tendenziell älter und wiesen zugleich ein etwas höheres Körpergewicht auf. Würde nun für beide Teilgesamtheiten (d. h. ohne Differenzierung nach dem Geschlecht) der Zusammenhang zwischen dem Alter und dem Körpergewicht untersucht, so könnte zwischen den Merkmalen X und Y eine positive Beziehung festgestellt werden (positive Kovarianz und positiver Korrelationskoeffizient), obwohl für jede Teilgesamtheit Unabhängigkeit zwischen den Merkmalen besteht. Der Statistiker spricht in diesem Zusammenhang von **Scheinkorrelation**. Da in der Analyse nicht nach dem Geschlecht differenziert wird, entsteht der scheinbare Eindruck eines Zusammenhangs zwischen dem Alter und der Körpergröße, obwohl in Wirklichkeit eine dritte, sogenannte latente (verborgene) Einflussgröße Merkmal Z (z. B: Geschlecht) für die vorgetäuschte Beziehung zwischen dem Alter und der Körpergröße verantwortlich ist (zum Begriff der Scheinkorrelation vgl. auch die Ausführungen in Kap. 6.5, S. 301)

Aufgabe 28c)

Diese Aussage ist nicht korrekt, da die Anwendung des Bravais-Pearson Korrelationskoeffizienten (r) metrisch skalierte Merkmale voraussetzt und r zudem nur die Stärke linearer Abhängigkeiten zum Ausdruck bringen kann.

Aufgabe 29: Varianz und Standardabweichung (S. 260, S. 249, S. 232/233)

Unter Einbeziehung der Ergebnisse der Aufgabe 26 gilt:

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{0{,}11}{0{,}8072 \cdot 0{,}6708} = 0{,}2032$$

Die Merkmale X und Y weisen eine positive Beziehung auf. Der Bravais-Pearson Korrelationskoeffizienten (r) bewegt sich bei positiver linearer Abhängigkeit von X und Y grundsätzlich im Intervall $(0 < r \leq 1)$. Im gegebenen Beispiel liegt der Wert von (r) im unteren Bereich des Intervalls, so dass eine schwache positive, lineare Korrelation zwischen den Ausfällen/Tag der Maschine X und den Ausfällen/Tag der Maschine Y besteht.

Aufgabe 30: Varianz und Standardabweichung (S. 265)

Aufgabe 30a)

- Merkmal X (Note) ist ordinalskaliert. Es ist eine Rangfolge der Schwimmbäder gegeben. Ein Schwimmbad mit der Note „2“ ist besser als ein Schwimmbad mit der Note „3“. Es sind aber keine Abstände definiert. Es kann z. B. nicht gesagt werden, dass Schwimmbäder mit den Noten „1“ bzw. „2“ sich in ihrer Qualität genauso unterscheiden wie z. B. Schwimmbäder mit den Noten „4“ bzw. „5“. Da kein math. Nullpunkt besteht, lassen sich auch keine Relationen bilden. Es kann z. B. nicht gesagt werden, dass ein Schwimmbad mit der Note „2“ doppelt so gut ist wie ein Schwimmbad mit der Note „4“.
- Merkmal Y ist verhältnisskaliert. Es existieren Rangfolge, Abstand und ein natürlicher Nullpunkt. Das dreißigjährige Schwimmbad ist 18 Jahre älter als das zwölfjährige Schwimmbad. Wegen des natürlichen Nullpunktes lässt sich auch sagen: „Das 30-jährige Schwimmbad ist doppelt so alt wie das 15-jährige Schwimmbad.

Die schwächste Skala bestimmt das Verfahren zur Ermittlung der Stärke des Zusammenhangs. Hier stellt die Ordinalskala die schwächste Skala dar. Daher ist ein Verfahren für ordinalskalierte Skalen zu wählen. Eine Möglichkeit besteht in der Verwendung des Rangkorrelationskoeffizienten nach Spearman (zu weiteren möglichen Verfahren bei einer Ordinalskala vgl. die Ausführungen auf S. 263, Fußnote 188 im Buch).

Hinweis: Grundsätzlich lassen sich auch Zusammenhangsmaße für nominalskalierte Verfahren wie der „korrigierte Kontingenzkoeffizient nach Pearson“ verwenden. Dabei ist allerdings zu beachten, dass ein Informationsverlust gegenüber denjenigen Verfahren stattfindet, die auf ordinalskalierte Merkmalswerte angewendet werden können (wie z. B. dem Rangkorrelationskoeffizienten nach Spearman, vgl. hierzu S. 265).

Aufgabe 30b)

Die nachfolgende Tabelle A zeigt, dass sich unter Verwendung der Formel für den Rangkorrelationskoeffizient dasselbe Ergebnis für R ermittelt, wie bei Berechnung des Rangkorrelationskoeffizienten R über den Bravais-Person Korrelationskoeffizient (letzterer verwendet die Ränge der Merkmalswerte anstelle der Merkmalswerte selbst). Der Korrelationskoeffizient R beträgt R = 0,8857. Zwischen den Rängen der Merkmale X und Y besteht eine starke **positive** lineare Korrelation. Dies ist gleichbedeutend damit, dass die älteren Schwimmbäder auch eine schlechtere Bewertung erfahren haben. Die Korrelation ist stark ausgeprägt, da R sich bei einer positiven Abhängigkeit grundsätzlich im Intervall $(0 < R \leq 1)$ bewegt und hier R im oberen Bereich liegt.

Bei der Tabelle A wurden die Ränge gleichläufig gestaltet. Dies wird dadurch erreicht, dass bei Merkmal X eine schlechtere Note auch mit einer großen Rangzahl einhergeht (Note 6 erhält den schlechtesten Rang, hier also Rang 6). Bei Merkmal Y geht ebenfalls ein höheres Alter mit einer großen Rangzahl einher (das älteste Schwimmbad erhält Rang 6, das jüngste Schwimmbad erhält Rang 1).

Anstelle des Gleichlaufs der Ränge können diese auch gegenläufig verlaufen, wenn z. B. das älteste Schwimmbad nicht den Rang 6, sondern den Rang 1 erhält und das jüngste Schwimmbad den Rangwert „6“ anstelle einer „1“ zugewiesen bekommt. Aus Tabelle B ist die gegenläufige Rangfolge ersichtlich. Bei ihr errechnet sich der gleiche Korrelationskoeffizient wie bei der gleichläufigen Anordnung der Ränge, nur mit einem negativen Vorzeichen, also R = –0,8857. Wird dieses Ergebnis interpretiert, bedeutet es weiterhin, dass in der Qualität und im Alter der Schwimmbäder eine positive Korrelation besteht, d. h. eine schlechte Bewertung mit einem hohen Alter verbunden ist. Lediglich die Rangzahlen sind negativ korreliert, da ein gutes Schwimmbad (kleine Rangzahlen) mit einem niedrigen Alter (große Rangzahlen) einhergeht.

Tabelle A zur Aufgabe 30b: Berechnung des Rangkorrelationskoeffizienten nach Spearman (X und Y verlaufen gleichläufig in den Rängen)

i	Schwimmbad*) X_i	Rang von X_i: $Rg(X_i)$	Alter Y_i in Jahren	Rang von Y_i: $Rg(Y_i)$	D_i = $Rg(X_i)$ minus $Rg(Y_i)$	D_i^2	$[Rg(X_i)]^2$	$[Rg(Y_i)]^2$	$Rg(X_i) \cdot Rg(Y_i)$
1	1	1	5	2	-1	1	1	4	2
2	3	3	9	3	0	0	9	9	9
3	6	6	30	6	0	0	36	36	36
4	2	2	1	1	1	1	4	1	2
5	4	4	15	5	-1	1	16	25	20
6	5	5	12	4	1	1	25	16	20
Σ		21	72	21	0	4	91	91	89

***) Schlüssel für Schwimmbadbewertung:**
1 = sehr gut; 2 = gut; 3 = befriedigend; 4 = ausreichend; 5 = mangelhaft; 6 = ungenügend

$$R = 1 - \frac{6 \cdot \sum D_i^2}{n \cdot (n^2 - 1)} = 1 - [(6 \cdot 4)/(6 \cdot 35)] = +0{,}885714$$

$S_{XY} = 1/6 \cdot 89 - (21/6) \cdot (21/6) = +2{,}583333$
$S_X = [1/6 \cdot 91 - (21/6) \cdot (21/6)]^{0,5} = +1{,}707825$
$S_Y = [1/6 \cdot 91 - (21/6) \cdot (21/6)]^{0,5} = +1{,}707825$
$r_{XY} = S_{XY}/(S_X \cdot S_Y) = 2{,}583333/(1{,}707825)^2 = 0{,}885714$

Tabelle B zur Aufgabe 30b: Berechnung des Rangkorrelationskoeffizienten nach Spearman (X und Y verlaufen entgegengesetzt in den Rängen)

i	Schwimmbad*) X_i	Rang von X_i: $Rg(X_i)$	Alter Y_i in Jahren	Rang von Y_i: $Rg(Y_i)$	D_i = $Rg(X_i)$ minus $Rg(Y_i)$	D_i^2	$[Rg(X_i)]^2$	$[Rg(Y_i)]^2$	$Rg(X_i) \cdot Rg(Y_i)$
1	1	1	5	5	-4	16	1	25	5
2	3	3	9	4	-1	1	9	16	12
3	6	6	30	1	5	25	36	1	6
4	2	2	1	6	-4	16	4	36	12
5	4	4	15	2	2	4	16	4	8
6	5	5	12	3	2	4	25	9	15
Σ		21	72	21	0	66	91	91	58

***) Schlüssel für Schwimmbadbewertung:**
1 = sehr gut; 2 = gut; 3 = befriedigend; 4 = ausreichend; 5 = mangelhaft; 6 = ungenügend

$$R = 1 - \frac{6 \cdot \sum D_i^2}{n \cdot (n^2 - 1)} = 1 - [(6 \cdot 66)/(6 \cdot 35)] = -0{,}885714$$

$S_{XY} = 1/6 \cdot 58 - (21/6) \cdot (21/6) = -2{,}583333$
$S_X = [1/6 \cdot 91 - (21/6) \cdot (21/6)]^{0,5} = +1{,}707825$
$S_Y = [1/6 \cdot 91 - (21/6) \cdot (21/6)]^{0,5} = +1{,}707825$
$r_{XY} = S_{XY}/(S_X \cdot S_Y) = -2{,}583333/(1{,}707825)^2 = -0{,}885714$

Aufgabe 31: Verschiedene Zusammenhangsmaße (S. 271, S. 56)

Die Skalierung der Merkmale wurde bereits in Aufgabe 4 beschrieben; daher sei hier nur das Ergebnis ohne Begründung wiederholt, um auf dieser Basis das geeignete Zusammenhangsmaß zu beschreiben.
Im Folgenden ist zu beachten, dass die schwächste Skala jeweils das Verfahren bestimmt, das zur Beurteilung der Stärke des Zusammenhangs anzuwenden ist.

Skalierung	Schwächste Skala	Geeignetes Verfahren (z. B.)
Merkmal X: Einkommen der Beschäftigten eines Unternehmens = Verhältnisskala Merkmal Y: Alter der Beschäftigten = Verhältnisskala	Verhältnisskala	Bravais-Pearson Korrelationskoeffizient
Merkmal X: verschiedene Güteklassen eines Konsumgutes = Ordinalskala Merkmal Y: Preis des Konsumgutes = Verhältnisskala	Ordinalskala	Rangkorrelationskoeffizient nach Spearman
Merkmal X: Studiendauer von Hochschulabsolventen der BWL = Verhältnisskala Merkmal Y: Einkommensarten der Studierenden = Nominalskala	Nominalskala	Korrigierter Kontingenzkoeffizient nach Pearson

Grundsätzlich kann ein Verfahren, was auf eine höherwertige Skala anzuwenden ist, nicht auf eine einfache Skala angewandt werden, wohl aber umgekehrt. Wird ein Verfahren für eine einfache Skala auch bei einer höherwertigen Skala eingesetzt, findet ein Informationsverlust statt, da das einfache Verfahren nicht alle Informationen der komplexeren Skala verarbeiten kann (vgl. S. 265). Das Verfahren ist quasi nicht „sensibel“ genug, um die vielen Informationen voll zu erfassen und auszuwerten (vgl. hierzu den Vergleich mit der Erdbebenmessung über alternative Verfahren; s. hierzu S. 252).
Bezogen auf die hier vorliegenden drei Beispiele der Merkmale X und Y bedeutet dies konkret: Im ersten Fallbeispiel der Verhältnisskala können auch Verfahren für ordinal- und nominalskalierte Merkmalswerte zur Messung der Stärke des Zusammenhangs zum Einsatz kommen (allerdings mit Informationsverlusten). Im zweiten Fall der Ordinalskala könnte auch ein Verfahren für nominalskalierte Merkmalswerte Verwendung finden (ebenfalls mit Informationsverlusten). Im dritten Fall der Nominalskala können nur die verschiedenen Verfahren für nominalskalierte Merkmale zur Messung der Stärke des Zusammenhangs herangezogen werden (wie z. B. der korrigierte Kontingenzkoeffizient nach Pearson oder Cramers V, vgl. S. 270).

Aufgabe 32: Welcher Bootsanleger darf es denn sein? (S. 271)

Gemeinsame absolute Häufigkeiten von Merkmal X und Y

Anleger (Y) / Farbe der Boote (X)	Anleger 1	Anleger 2	Gesamt
Rot	22	30	52
Blau	34	43	77
Rest	64	27	91
Gesamt	120	100	220

Aufgabe 32a)

Gemeinsame relative Häufigkeiten für Merkmal X und Y

Anleger (Y) / Farbe der Boote (X)	Anleger 1	Anleger 2	Gesamt
Rot	0,100	0,137	0,237
Blau	0,154	0,195	0,349
Rest	0,291	0,123	0,414
Gesamt	0,5455	0,455	1,000

Aufgabe 32b)

Die bedingten relativen Häufigkeiten für Anleger 1 bzw. Anleger 2 unter der Bedingung, dass ein rotes Boot gewünscht wird, stellen sich wie folgt dar:

$$f(Y_1/X_1) = \frac{h_{11}}{h_{1.}} = \frac{22}{52} = 0{,}4231$$

$$f(Y_2/X_1) = \frac{h_{12}}{h_{1.}} = \frac{30}{52} = 0{,}5769$$

Der Passagier sollte an Anleger 2 warten, da dort die bedingte relative Häufigkeit, ein rotes Boot zu erhalten, höher ausfällt als für Anleger 1.

Aufgabe 32c)

Unter der Annahme, dass Unabhängigkeit von Bootsfarbe und Anleger besteht, bestimmen sich die theoretischen absoluten Häufigkeiten wie folgt über die Randhäufigkeiten:

$$h_{11}^* = \frac{h_{1.} \cdot h_{.1}}{n} = \frac{52 \cdot 120}{220} = 28{,}3636\,; \quad h_{12}^* = \frac{h_{1.} \cdot h_{.2}}{n} = \frac{52 \cdot 100}{220} = 23{,}6363$$

$$h_{21}^* = \frac{h_{2.} \cdot h_{.1}}{n} = \frac{77 \cdot 120}{220} = 42; \quad h_{22}^* = \frac{h_{2.} \cdot h_{.2}}{n} = \frac{77 \cdot 100}{220} = 35$$

$$h_{31}^* = \frac{h_{3.} \cdot h_{.1}}{n} = \frac{91 \cdot 120}{220} = 49{,}6363; \quad h_{32}^* = \frac{h_{3.} \cdot h_{.1}}{n} = \frac{91 \cdot 100}{220} = 41{,}3636$$

Aufgabe 32d)

Merkmal X und Y weisen jeweils eine Nominalskala auf. Die Farben und die Anleger stehen gleichrangig nebeneinander und eine Rangfolge kann nicht gebildet werden. Es lässt sich z. B. nicht sagen, dass die Farbe „Rot“ besser oder schlechter als die Farbe „Blau“ oder Anleger 1 besser oder schlechter als Anleger 2 ist. Da nominalskalierte Merkmale vorliegen, kann der Zusammenhang nur über ein Verfahren für nominalskalierte Merkmale wie z. B. über den korrigierten Kontingenzkoeffizienten ermittelt werden. Hierzu wird zunächst die Größe χ^2gebildet, die dann in den korrigierten Kontingenzkoeffizienten einfließt:

$$\chi^2 = \frac{(22-28{,}3636)^2}{28{,}3636} + \frac{(30-23{,}6363)^2}{23{,}6363} + \frac{(34-42)^2}{42} +$$

$$+\frac{(43-35)^2}{35} + \frac{(64-49{,}6363)^2}{49{,}6363} + \frac{(27-41{,}3636)^2}{41{,}3636} = 15{,}6378$$

$$C_{korr} = \sqrt{\frac{15{,}6378}{15{,}6378+220} \cdot \frac{2}{1}} = 0{,}3643$$

Da C_{korr} sich hier mit einem Wert von C_{korr} = 0,3643 im unteren Bereich des möglichen Intervalls ($0 \leq C_{korr} \leq 1$) befindet, besteht eine schwache Abhängigkeit zwischen der Farbe der Boote und dem Anleger. Diese Abhängigkeit könnte aber auch nur zufällig zustande gekommen sein. Ob der Zufall für diese Abweichungen verantwortlich ist, lässt sich nur mit einem Test der Schließenden Statistik wie z. B. dem χ^2-Unabhängigkeitstest beantworten (vgl. hierzu Teil D des Anhangs auf S. 368).

Aufgabe 32e)

Liegt Unabhängigkeit der Merkmale X und Y vor, so muss u. a. gelten:

$f(X_1/Y_1) = f(X_1/Y_2) = f(X_1)$

Diese Bedingung ist hier nicht erfüllt:

$$[f(X_1/Y_1) = \frac{22}{120}] \neq [f(X_1/Y_2) = \frac{30}{100}] \neq [f(X_1) = \frac{52}{220}]$$

Es könnte somit eine Abhängigkeit von Anleger und Farbe vorliegen, wenn die Abweichungen nicht auf den Zufall zurückzuführen sind (vgl. Ausführungen zur Aufgabe 32e).

Aufgabe 32f)

Die Größe χ^2 wird gemäß folgender Formel gebildet.

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{r} \frac{\left(h_{ij} - h_{ij}^*\right)^2}{h_{ij}^*} \quad \text{mit: } h_{ij}^* = \frac{\left(h_{i.} \cdot h_{.j}\right)}{n} \quad \text{für } i = 1, \dots, m; j = 1, \dots, r$$

Liegt Unabhängigkeit der Merkmale X und Y vor, so stimmen – bis auf zufällige Abweichungen – die empirischen und die theoretisch erwarteten Häufigkeiten überein und χ^2 nimmt den Wert „0“ an. Allerdings weist die Größe χ^2 – ähnlich wie die Kovarianz S_{XY} – den Nachteil auf, dass sie selbst bei starker Abhängigkeit nicht gegen einen konstanten Maximalwert konvergiert, sondern dieser von der Zahl der Beobachtungswerte (n) und der Zahl der Merkmalsausprägungen abhängig ist. Die Abhängigkeit von der Beobachtungszahl (n) und der Zahl der Merkmalsausprägungen kann vermieden werden, indem χ^2 in den „korrigierten (oder normierten) Kontingenzkoeffizient (C_{korr})“ überführt wird, der maximal den Wert 1 annimmt und wie folgt definiert ist:

$$C_{korr} = \sqrt{\frac{\chi^2}{\chi^2 + n} \cdot \frac{C^*}{C^* - 1}} \quad \text{mit: } C^* = \text{Min}(m, r)$$

Je nach der Stärke der Abhängigkeit bewegt sich C_{korr} im möglichen Intervall ($0 \leq C_{korr} \leq 1$). Bei Unabhängigkeit beträgt der Wert von $\chi^2 = 0$, bei perfekter Abhängigkeit gilt $C_{korr} = 1$.

Aufgabe 33: Stellung im Beruf nach Geschlecht (S. 272)

Beide Merkmale X und Y weisen eine Nominalskala auf: Die Merkmale stehen gleichberechtigt nebeneinander (zu einer ausführlichen Begründung der Skala vergleiche die Ausführungen zu Aufgabe 2 und 4, die sich analog auf dieses Beispiel übertragen lassen).

Die schwächste Skala ist hier die Nominalskala, so dass z. B. der korrigierte Kontingenzkoeffizient nach Pearson zur Anwendung kommen kann, um die Abhängigkeit zwischen den Merkmalen X und Y zu beschreiben; hierzu sind zunächst die theoretisch erwarteten absoluten Häufigkeiten zu ermitteln:

$$h_{11}^{*} = \frac{h_{1.} \cdot h_{.1}}{n} = \frac{4{,}64 \cdot 21{,}67}{40{,}16} = 2{,}5037 \qquad h_{12}^{*} = \frac{h_{1.} \cdot h_{.2}}{n} = \frac{4{,}64 \cdot 18{,}49}{40{,}16} = 2{,}1363$$

$$h_{21}^{*} = \frac{h_{2.} \cdot h_{.1}}{n} = \frac{35{,}52 \cdot 21{,}67}{40{,}16} = 19{,}1663 \qquad h_{22}^{*} = \frac{h_{2.} \cdot h_{.2}}{n} = \frac{35{,}52 \cdot 18{,}49}{40{,}16} = 16{,}3537$$

Auf dieser Basis ermittelt sich χ^2 wie folgt:

$$\chi^2 = \frac{(3{,}10 - 2{,}5037)^2}{2{,}5037} + \frac{(1{,}54 - 2{,}1363)^2}{2{,}1363} + \frac{(18{,}57 - 19{,}1663)^2}{19{,}1663} + \frac{(16{,}95 - 16{,}3537)^2}{16{,}3537} = 0{,}3488$$

Wird der Wert für χ^2 in die Formel für C_{korr} eingesetzt, ergibt sich:

$$C_{korr} = \sqrt{\frac{0{,}3488}{0{,}3488 + 40{,}16} \cdot \frac{2}{1}} = 0{,}1312$$

Es herrscht ein sehr schwacher Zusammenhang zwischen Stellung im Beruf und Geschlecht, da der Wert $C_{korr} = 0{,}1312$ im unteren Bereich des möglichen Intervalls ($0 \leq C_{korr} \leq 1$) liegt.

Aufgabe 34: Waldschadensbericht Nordrhein-Westfalen 2014 (S. 272)

Aufgabe 34a)

- Merkmalsträger = Waldbäume des Landes NRW, da sie im Hinblick auf das Merkmal „Schadenszustand“ untersucht werden;
- Merkmal = Schadenszustand;
- Merkmalsausprägung = konkrete Schadstufe, z. B. „ohne Kronenverlichtung (0 – 10 % Verlichtung)“, „schwache Kronenverlichtung (11 – 25 % Verlichtung)“, etc.

Aufgabe 34b)

Zur sachlichen Abgrenzung des Merkmalsträgers:

- Was ist als Baum eines Waldes anzusehen? (Auch die Bäume im Vorgarten der Einwohner?)
- Was ist unter „Wald“ zu verstehen?

Zur sachlichen Abgrenzung der Merkmalsausprägungen:

- Wie ist eine Schadstufe definiert?
- Was bedeutet „ohne Kronenverlichtung“, „schwache Kronenverlichtung“ etc.?

Fortsetzung Aufgabe 34:

Aufgabe 34c)

Hier liegt eine Ordinalskala vor, denn es ist eine Rangfolge gegeben, aber es sind keine Abstände der Merkmalsausprägungen quantifizierbar. Die Schadstufen lassen sich nach der Stärke der Schädigung im Sinne von „schwächere bzw. stärkere Beschädigung“ unterscheiden. Eine Schädigung der Stufe „schwache Kronenverlichtung“ ist geringer als eine Schädigung der Stufe „mittelstarke Kronenverlichtung“. Jedoch lässt sich für die verschiedenen Ausprägungen nicht aussagen, wie weit die Schädigung sich unterscheidet. Auch besteht kein absoluter Nullpunkt in der Erfassung der Merkmalsausprägungen. Folglich lassen sich auch keine Relationen bilden und es ist z. B. die Aussage nicht möglich, dass die Schädigung der Stufe „mittelstarke Kronenverlichtung“ ein Vielfaches einer geringeren Schädigungsstufe beträgt.

Aufgabe 34d)

Aufgrund der Ordinalskala können lediglich Modus und Median berechnet werden: Der Modus lässt sich bei allen Skalen ermitteln; der Median setzt eine Rangfolge der Merkmalsausprägungen voraus, die mit der Ordinalskala gegeben ist. Ein arithmetisches Mittel kann nicht bestimmt werden, da in dem vorliegenden Fallbeispiel Abstände der Merkmalsausprägungen nicht definiert sind (s. o.), d. h. ein arithmetisches Mittel kann erst aber einer Intervallskala ermittelt werden.

Modus X_{Mo}:

Der Modus ist definiert als „die häufigste Merkmalsausprägung“; in diesem Beispiel kommt die Merkmalsausprägung "schwache Kronenverlichtung" mit einer relativen Häufigkeit von 41 % am häufigsten vor, so dass sie den Modus darstellt. Somit gilt: X_{Mo} = "schwache Kronenverlichtung" (Hinweis: eine Antwort: „Modus = 41 %“ kann als Fettnäpfchen der Statistik angesehen werden!)
Bei der Bestimmung des häufigsten Wertes ist zu beachten, dass die Schadstufen (2 – 4) zusammengefasst sind und damit die Häufigkeiten der verschiedenen Schadstufen nicht direkt verglichen werden können. Werden die Schadstufen (2 – 4) als Zusammenfassung anderer Schadstufen verstanden und nicht als eine eigene definierte Stufe, müssten für eine unverzerrte Bestimmung des Modus die relativen Häufigkeiten der jeweiligen Schadstufen 2, 3 und 4 bekannt sein. Da im vorliegenden Beispiel aber selbst für die zusammengefassten Schadstufen (2 – 4) die Häufigkeit kleiner ausfällt als bei der Schadstufe (X_1 = "schwache Kronenverlichtung"), kann der Modus $\mathbf{X_{Mo}}$ = **"schwache Kronenverlichtung"** als unverzerrter Modus angesehen werden.

Median X_{Me}:

Hier liegt eine große Beobachtungszahl (n) vor (viele Bäume), so dass der Median über die Verteilungsfunktion (relative Summenhäufigkeit F_i) bestimmt werden kann. Bei derjenigen Merkmalsausprägung, bei der F_i den Wert $F_i = 0{,}5$ erreicht, liegt der Median. Im vorliegenden Beispiel wird $F_i = 0{,}5$ bei der 2. Ausprägung X_2 = Schadstufe 1 = "schwache Kronenverlichtung erreicht. Mit $F_2 = 64\,\%$ weisen 64 % der Bäume eine Schädigung höchstens der Schadstufe 2 = "schwache Kronenverlichtung" auf. Der Median liegt somit in der zweiten Schadstufe, da 50 % der Bäume eine Schädigung der Schadstufe 2 oder weniger aufweisen und 50 % der Bäume eine Schädigung der Schadstufe 2 oder mehr besitzen. Somit gilt: X_{Me} = "schwache Kronenverlichtung".

Aufgabe 34d)

Soll der Zusammenhang zwischen der Schadstufe (Merkmal X) und der Baumart (Merkmal Y) untersucht werden, bestimmt die schwächste Skala das Verfahren. In diesem Beispiel weist das Merkmal Y „Baumart“ eine Nominalskala und damit die schwächste Skala auf. Die Baumart ist nominalskaliert, da die verschiedenen Baumarten wie z. B. Kiefer, Buche etc. gleichberechtigt nebeneinander stehen und keine Rangfolge in den Baumarten besteht. Es kann nicht gesagt werden, dass z. B. die Baumart „Buche“ besser oder schlechter als eine andere Baumart (z. B. als „Kiefer“) ist. Auch sind die Abstände und der mathematische Nullpunkt nicht definiert. Da die Nominalskala die schwächste Skala beider Merkmale ist, kommt ein Verfahren für nominalskalierte Merkmale wie z. B. der korrigierte Kontingenzkoeffizient nach Pearson zur Anwendung, um die Stärke der Abhängigkeit zu bestimmen.

Aufgabe 35: Prof. Emsig und die Regressionsanalyse (S. 303)

Hilfsangaben: $\overline{X} = 80\ \%$ der Vorlesungen: $\overline{Y} = 70\ \%$ der Punkte;

$b_2 = 0{,}6$ [% der Punkte je %-Punkt der besuchten Vorlesung] (ergibt sich aus dem Text der Aufgabe)

Aufgabe 35a)

$\widehat{Y}_i = b_1 + 0{,}6 \cdot X_i$

$b_1 = \overline{Y} - b_2 \cdot \overline{X}$

$b_1 = 70 - 0{,}6 \cdot 80 = 22$

Daraus folgt: $\widehat{Y}_i = 22 + 0{,}6 \cdot X_i$

Aufgabe 35b)

$\widehat{Y}_i = 22 + 0{,}6 \cdot 0\ \%$

$\widehat{Y}_i = 22$

Die Punktzahl bei einer Teilnahmequote von 0 % beträgt 22 % der Punkte, d. h. 22 Punkte.

Aufgabe 35c)

Hier liegt eine (lineare) Einfachregression vor; daher lässt sich das Bestimmtheitsmaß R^2 durch die Quadrierung des Bravais-Pearson Korrelationskoeffizienten (r) bilden.

$R^2 = r^2 = 0{,}8^2 = 0{,}64$

Interpretation: 64 % der Schwankungen der erzielten Punktzahl (Varianz von Y) werden über die Schwankungen der Teilnahmequote an der Vorlesung (Varianz von X) erklärt. 36 % der Schwankungen der Punktzahl werden nicht erklärt und sind zufallsbedingt. (Hinweis: die nicht erklärten Schwankungen können nicht auf eine andere Einflussgröße zurückgeführt werden, da die Regressionsfunktion alle erklärenden Variablen erfassen muss, damit keine Fehlspezifikation der Regressionsfunktion vorliegt).

Aufgabe 36: Regressionsanalyse auf dem Wohnungsmarkt (S. 336)

Aufgabe 36a)

Hier liegen Einzelwerte vor. Der Median wird als Merkmalsausprägung des mittleren Merkmalsträgers der geordneten Urliste ermittelt. Dazu sind die Einzelwerte zunächst in eine Reihenfolge zu bringen:

Geordnete Merkmalswerte X_i										
i	1	2	3	4	5	6	7	8	9	10
$\mathbf{X_i}$	25	35	45	55	65	75	95	105	120	150

Hier liegt mit (n = 10) eine gerade Beobachtungszahl (n) vor; daher ergibt sich der Median als:

$X_{Me} = 0{,}5 \cdot \left(X_{[10/2]} + X_{[(10/2)+1]}\right)$

$X_{Me} = 0{,}5 \cdot \left(X_{[5]} + X_{[6]}\right)$

$X_{Me} = 0{,}5 \cdot (65 + 75)$

$X_{Me} = 70\ \text{qm}$

Hinweis auf Fettnäpfchen in den Klausuren:

1) Ordnen der Zahlen nicht vergessen!
2) Korrekte Übersetzung von der Position der mittleren Merkmalswerte auf die Merkmalswerte selbst beachten.

Aufgabe 36b)

$b_2 = \frac{S_{XY}}{S_X^2}$; $S_X^2 = \frac{1}{10} \cdot 73\,700 - \left(\frac{1}{10} \cdot 770\right)^2 = 1\,441\ \text{qm}^2$; $b_2 = \frac{7\,492{,}7380}{1\,441} = 5{,}1997$

Steigt die Wohnfläche um 1 qm, so erhöht sich die Nettokaltmiete um 5,1997 €.

Aufgabe 36c)

Unter einem Residuum ist die Abweichung des empirischen Wertes Y_i von dem über die Regressionsfunktion geschätzten Wert $\widehat{Y}_i$ zu verstehen, also: $e_i = Y_i - \widehat{Y}_i$

Zur Ermittlung des Residuums ist zunächst die Regressionsfunktion zu bestimmen. Da b_2 bereits bekannt ist (s. Aufgabe 36b), lässt sich b_1 u. a. über folgende Beziehung ermitteln:

$$b_1 = \overline{Y} - b_2 \cdot \overline{X} = \left(\frac{1}{10} \cdot 4\,533{,}81\right) - 5{,}1997 \cdot 77 = 53{,}0041$$

Somit gilt für die Regressionsfunktion: $\widehat{Y}_i = 53{,}0041 + 5{,}1997 \cdot X_i$

Damit ergibt sich für die 7. Wohnung mit einer Wohnfläche von 65 qm ein Residuum i. H. v.:

$e_i = Y_i - \widehat{Y}_i = 296{,}84 - (53{,}0041 + 5{,}1997 \cdot X_i)$

$e_i = 296{,}84 - 53{,}0041 - 5{,}1997 \cdot 65 = -94{,}1446$

Aufgabe 36d)

Die Regressionsfunktion wird beim Kleinste-Quadrate-Verfahren (K-Q-V) derart durch die Regressionsfunktion gelegt, dass die Summe der Residuen immer „null“ ergibt, also: $\sum e_i = 0$. Dieses Ergebnis resultiert aus der Minimierung der Summe der quadrierten Abweichungen der Residuen (SAQ), also $SAQ = \sum e_i^2 = \text{Minimum}!$ Hierzu werden die partiellen Ableitungen von SAQ nach b_1 und b_2 gebildet und gleich „null“ gesetzt. Aus der partiellen Ableitung nach b_1 ergibt sich die 1. Eigenschaft der linearen Einfachregression beim KQV:

$SAQ'(b_1) = -2 \cdot \sum(Y_i - b_1 - b_2 \cdot X_i) = 0$ (1. Normalgleichung)

Wegen $(Y_i - b_1 - b_2 \cdot X_i) = e_i$ folgt: $SAQ'(b_1) = \sum e_i = 0$

Die Regressionsfunktion wird somit auf eine Weise durch die Punktwolke gelegt, dass die Summe der Residuen null wird ($\sum e_i = 0$).

Aufgabe 36e)

Für die lineare Regressionsfunktion ($\widehat{Y}_i = 53{,}0041 + 5{,}1997 \cdot X_i$) ist R^2 zu bestimmen, mit:

$$R^2 = \frac{SQE}{SQT} = \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT}$$

mit: $SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} Y_i^2 - b_1 \cdot \sum_{i=1}^{n} Y_i - b_2 \cdot \sum_{i=1}^{n} X_i \cdot Y_i$; wobei: $\sum_{i=1}^{n} Y_i^2 = 2\,479\,949{,}3569$

Somit: $SQR = 2\,479\,949{,}3569 - 53{,}0041 \cdot 4\,533{,}81 - 5{,}1997 \cdot 424\,030{,}75 = 34\,806{,}1475$

$$SQT = n \cdot S_Y^2 = n \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\overline{Y}\right)^2 = 10 \cdot \left(\frac{1}{10} \cdot 2\,479\,949{,}3569 - 453{,}381^2\right) = 424\,406{,}0453$$

$$R^2 = 1 - \frac{34\,806{,}1475}{424\,406{,}0453} = 0{,}917989$$

91,8 % der Gesamtstreuung der Nettokaltmieten (Varianz von Y_i) wird über die Streuung der Wohnflächen (X_i), d. h. über die Regressionsfunktion erklärt. 8,2 % der Streuung der Nettokaltmieten wird nicht erklärt (SQE) und ist somit zufallsbedingt.

Aufgabe 37: Regressionsanalyse im Gesundheitsbereich (S. 337)

Aufgabe 37a)

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - (\overline{X})^2 = \frac{1}{22}\cdot 259{,}8396 - \left(\frac{1}{22}\cdot 67{,}76\right)^2 = 2{,}3245$$

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}X_i \cdot Y_i - \overline{X}\cdot\overline{Y} = \frac{1}{22}\cdot 661{,}495 - \left(\frac{1}{22}\cdot 67{,}76\right)\cdot\left(\frac{1}{22}\cdot 195{,}10\right) = 2{,}7540$$

Aufgabe 37b)

$$b_2 = \frac{S_{XY}}{S_X^2} = \frac{2{,}7540}{2{,}3245} = 1{,}1848$$

Steigt das monatliche Pro-Kopf-Einkommen um 1 000 $ an, so verändert sich der prozentuale Anteil der Gesundheitsausgaben am BIP um 1,1848 Prozentpunkte.

Aufgabe 37c)

Gesucht ist der Niveauparameter (absolute Glied) b_1:

$$b_1 = \overline{Y} - b_2\cdot\overline{X} = 8{,}8381 - 1{,}1848\cdot 3{,}08 = 5{,}1890.$$

Der Schnittpunkt mit der Y-Achse liegt an der Stelle Y = 5,1890 % (Gesundheitsausgaben/BIP).

Aufgabe 37d)

Hier liegt eine lineare Einfachregression vor. Daher lässt sich das Bestimmtheitsmaß vereinfacht über den quadrierten Bravais-Pearson Korrelationskoeffizienten ermitteln:

$$r = \frac{S_{XY}}{S_X\cdot S_Y} = \frac{2{,}7540}{\sqrt{2{,}3245}\cdot 2{,}1714} = 0{,}8319; \quad \text{somit: } R^2 = r^2 = 0{,}8319^2 = 0{,}6921$$

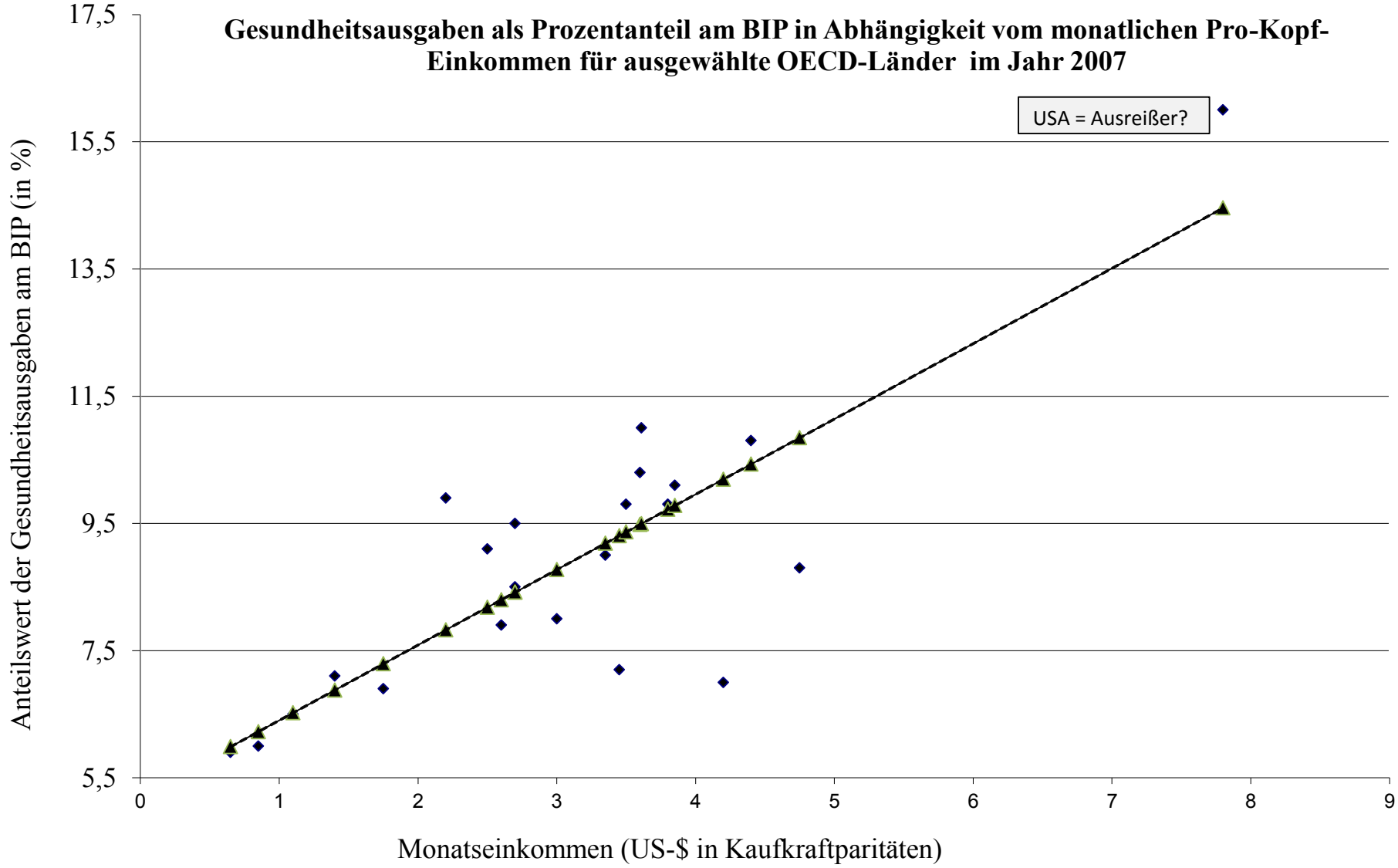
Damit lassen sich 69,21 % der Gesamtstreuung der anteiligen Gesundheitsausgaben am BIP (Varianz von Y_i) über die Streuung der monatlichen Pro-Kopf-Einkommen (X_i), d. h. über die Regressionsfunktion erklären. 30,79 % der Streuung der anteiligen Gesundheitsausgaben am BIP werden nicht erklärt (SQE) und sind somit zufallsbedingt.

Aufgabe 37e) Gemäß der Abbildung der Aufgabe 37 weichen die Daten für die USA von den Daten der anderen Staaten ab, so dass die USA als „Ausreißer" betrachtet werden könnte. Ausreißer in der Regressionsanalyse können eine Scheinkorrelation auslösen oder die Ergebnisse verzerren (zur Scheinkorrelation vgl. die Ausführungen auf S. 301, insbesondere S. 302). Dies würde im vorliegenden Beispiel bedeuten, dass eine Beziehung zwischen dem Anteil der Gesundheitsausgaben und dem Einkommen unterstellt wird, die de facto nicht oder nicht so stark besteht. Die nachfolgende Graphik der Gesundheitsausgaben als Anteil am BIP zeigt, dass die Merkmalskombination der USA leicht nach oben vom Trend abweicht, wobei der Trend, d. h. die Steigung der Regressionsfunktion durch die Merkmalskombination der USA selbst nach oben gedreht wird. Dies zeigt eine Analyse ohne die USA-Daten für die restlichen 21 Länder. In diesem Fall ergeben sich folgende Werte (Werte unter Einbeziehung der USA zum Vergleich in Klammern):

$b_1 = 5{,}93\ (5{,}20);\ b_2 = 0{,}91\ (1{,}1848);\ R^2 = 0{,}457(0{,}69)$

Dieser Trend würde noch weiter abgeschwächt und das Bestimmtheitsmaß weiter gesenkt, wenn auch die beiden Merkmalswerte für die Türkei und Mexiko aus der Analyse herausgenommen würden. Dies zeigt, dass eine Datenbereinigung um „Ausreißer" sehr subjektiv ist und das Ergebnis stark verändern kann. Daher ist bei der Anwendung und Interpretation der Regressionsanalyse in diesen Situationen vermeintlicher Ausreißer Vorsicht geboten. Solange aber die statistischen T-Werte für die Regressionsparameter auch unter Ausschluss der vermeintlichen Ausreißer mit höherer Signifikanz gesicherte T-Werte für die Regressionsparameter ausweist, kann davon ausgegangen werden, dass eine Abhängigkeit zwischen X und Y zwar besteht (also keine Scheinkorrelation vorliegt), diese Abhängigkeit aber ggfs. durch vermeintliche Ausreißer verzerrt wird (Zu den T-Werten der Regressionsparameter vgl. z. B. die Anmerkungen in Tab. 6-2-1 (S. 283) sowie S. 345 ff.

Fortsetzung von Aufgabe 37e)



Natrop, Johannes: Angewandte Deskriptive Statistik. Praxisbezogenes Lehrbuch mit Fallbeispielen, Berlin u.a., 2015.

Musterklausur I, S. 370 ff. (Hinweis zur Bearbeitung der Aufgaben: Die Antworten sind im Sinne der Wiederholung des Stoffes sehr umfassend formuliert und lassen sich z. T. auch kürzer fassen. Erforderlich ist jeweils eine ausreichende Begründung bzw. ein nachvollziehbarer Rechengang. Eine Darstellung der Formeln ist bei Rechengängen nicht erforderlich, sofern nicht explizit danach gefragt wird. Bitte darauf achten, dass bei allen Berechnungen ersichtlich ist, was formal ermittelt wurde. Eine bloße Auflistung von Zahlen ohne formalen Hinweis darauf, was berechnet wird, ist nicht ausreichend und führt zu Punktabzügen in der Bewertung!)

Aufgabe 1: Richtig oder Falsch (Hinweis: Die Seitenangaben beziehen sich auf das Buch).

Aufgabe 1a: Aussage zur relativen Summenhäufigkeit (S. 93)

Diese Aussage ist falsch. Die relativen Summenhäufigkeiten (F_i) werden bei klassifizierten Daten erst an der Klassenobergrenze erreicht. Daher werden die relativen Summenhäufigkeiten (F_i) den Klassenobergrenzen und nicht den Klassenmitten zugeordnet.

Aufgabe 1b: Aussage zu den Randhäufigkeiten einer zweidimensionalen H.V. (S. 205 – 208)

Werden die gemeinsamen absoluten Häufigkeiten $h(X_i/Y_j)$ über alle Ausprägungen ($i = 1, \dots, m$) des Merkmals X summiert, also für jeder Spalte über alle Zeilen addiert, so ergibt sich die absolute Randhäufigkeit $h(Y_j)$ oder $h_{.j}$ des Merkmals Y (und nicht des Merkmals X). Die Aussage ist somit falsch.

Aufgabe 1c: Aussage zur Kovarianz und zur Unabhängigkeit (S. 233)

Diese Aussage ist nicht korrekt. Von einer Kovarianz $S_{XY} = 0$ kann nicht stets auf Unabhängigkeit geschlossen werden. Es kann nur auf lineare Unabhängigkeit geschlossen werden. Gleichwohl kann bei einer Kovarianz von $S_{XY} = 0$ eine nichtlineare oder sogar perfekte nichtlineare Abhängigkeit bestehen (vgl. S. 248 im Buch). Dieser Sachverhalt ist auf die Problematik zurückzuführen, dass die Kovarianz nur lineare Abhängigkeiten abzubilden vermag. Dieses Problem kann auch durch Überführung der Kovarianz in den Korrelationskoeffizienten nicht gelöst werden.

Aufgabe 1d: Korrelationskoeffizient (S. 252) und Regressionskoeffizient (S. 276)

Der Korrelationskoeffizient $r = \frac{S_{XY}}{S_X \cdot S_Y}$ ist **eine dimensionslose** Größe. Dadurch, dass im Zähler die Kovarianz mit der multiplikativen Verknüpfung der Dimensionen beider Merkmale erfasst wird und im Nenner die Standardabweichungen beider Merkmale (ebenfalls multiplikativ verknüpft) aufgeführt sind, kürzen sich die Dimensionen jeweils weg, so dass r dimensionslos ist.

Die Regressionskoeffizienten (b_1) (Niveauparameter) und (b_2) (Steigungsparameter) stellen hingegen keine dimensionslosen Größen dar. Der Niveauparameter weist die Dimension der Y-Werte auf. Der Steigungsparameter überführt die Dimension der exogenen Variablen (X-Wert) in die Dimension der endogenen Variablen (Y-Wert). Wird z. B. in einer Regressionsfunktion die Beziehung zwischen der Miete in € (endogene Variable) und der Wohnfläche in qm (exogene Variable) beschrieben, so zeigt der Steigungsparameter b_2 auf, wie die Miete in € ansteigt, wenn die Wohnfläche um einen qm zunimmt. Damit weist b_2 die Dimension [€ je qm] auf. Der Niveauparameter (b_1) gibt an, welchen Wert Y für (X = 0) aufweist (Schnittpunkt der Regressionsfunktion mit der Y-Achse). Daher besitzt der Parameter b_1 die Dimension der Y-Werte, d.h. im vorliegenden Fall die Dimension „€“.

Aufgabe 2: Mikrozensus 2013

Aufgabe 2a: Berechnung von h_5

$$f_1 = \frac{h_1}{n}; \quad \text{somit: } n = \frac{h_1}{f_1} = \frac{5\,518}{0{,}148549} = 37\,146; \qquad h_5 = f_5 \cdot n = 0{,}190815 \cdot 37\,146 = 7\,088$$

(Hinweis: eine Lösung ist auch etwas schneller möglich über:

$$n = \frac{h_1}{f_1} = \frac{h_5}{f_5}; \quad \text{somit: } h_5 = f_5 \cdot \frac{h_1}{f_1} = 0{,}190815 \cdot \frac{5\,518}{0{,}148549} = 7\,088$$

Aufgabe 2b: Mittelwerte

Zum Modus:

Hier liegen klassifizierte Daten mit unterschiedlicher Klassenbreite vor; daher ist der Modus als Klassenmitte der dichtesten Klasse definiert. Aus den Daten ist ersichtlich (da $f_1 = d_1$), dass die berechnete Dichte mit der normierten Klassenbreite $\Delta X^n = 700$ errechnet wurde. Damit ermittelt sich die Dichte d_3 der 3. Klasse wie folgt:

$$d_3 = \frac{f_3}{\Delta X_3} \cdot \Delta X^n = \frac{0{,}198137}{400} \cdot 700 = 0{,}346740$$

Die dichteste Klasse stellt somit $d_3 = 0{,}346740$ in der 3. Klasse dar.

Die Klassenmitte der 3. Klasse stellt den Modus dar. Dies ist somit:

$$X_{Mo} = \frac{1100 + 1500}{2} = 1\,300$$

Zum Median:

Bei klassifizierten Daten erfolgt eine „Feinberechnung" des Medians ($F_i = 0{,}5$):

$$X_{Me} = X_3^u + \Delta X_i \cdot \frac{0{,}5 - F(X_3^u)}{F(X_3^o) - F(X_3^u)} = 1\,100 + 400 \cdot \frac{0{,}5 - 0{,}3047}{0{,}5029 - 0{,}3047} = 1\,494{,}1473$$

Zum arithmetischen Mittel:

$$\overline{X} = \sum_{i=1}^{m} X_i' \cdot f_i = 1\,643{,}66 \quad \text{(siehe Hilfsangabe)}$$

Somit gilt: $\overline{X} = 1\,643{,}66$

Aufgabe 2c: Anteilswerte

Zum Anteilswert der Erwerbstätigen, die mehr als das Medianeinkommen zur Verfügung haben:

Der Median zerlegt die H.V. in zwei gleich große Hälften ($F_i = 0{,}5$). Somit beträgt der Anteilswert der Erwerbstätigen, die mehr als das Medianeinkommen zur Verfügung haben, 50 %.

Zum Anteilswert der Erwerbstätigen, die mehr als das arithmetische Mittel $\overline{X} = 1\,643{,}66$ zur Verfügung haben:

Zunächst wird der Anteil der Erwerbstätigen berechnet, die ein Einkommen **unterhalb** des arithmetischen Mittels $\overline{X} = 1\,643{,}66$ aufweisen.

$$F(X \leq 1\,643{,}66) = F(X_4^u) + f_4 \cdot \frac{1\,643{,}66 - X_4^u}{\Delta X_4} = 0{,}5029 + 0{,}208878 \cdot \frac{1\,643{,}66 - 1\,500}{500} = 0{,}5629$$

Damit ermittelt sich der Anteilswert der Erwerbstätigen, die **mehr** als das arithmetische Mittel $\overline{X} = 1\,643{,}66$ aufweisen als: $F(X > 1\,643{,}66) = 1 - 0{,}5629 = 0{,}4371$

43,71 % der Erwerbstätigen haben mehr als das Durchschnittsnettoeinkommen zur Verfügung.

Hinweis auf eine weitere mögliche Frage: Wäre zusätzlich nach dem Anteilswert der Erwerbstätigen gefragt, die mehr als das Medianeinkommen und weniger als das arithmetische Mittel aufweisen, so ergäbe sich: $F(X_{Me} \leq X \leq \overline{X}) = 0{,}5629 - 0{,}5 = 0{,}0629$ (also 6,29 %).

Aufgabe 2d: MAD

$$MAD(X_{Me}) = \sum_{i=1}^{6} |X_i' - X_{Me}| \cdot f_i = \sum_{i=1}^{5} |X_i' - X_{Me}| \cdot f_i + |X_6' - X_{Me}| \cdot f_6$$

$$\text{mit: } \sum_{i=1}^{5} |X_i' - X_{Me}| \cdot f_i = 537{,}0689 \quad \text{(siehe Hilfsangaben)}$$

Somit: $MAD(X_{Me}) = 537{,}0689 + |3\,700 - 1\,494{,}1473| \cdot 0{,}097426(X_{Me}) = 751{,}9763$

Die durchschnittliche (absolute) Abweichung der Merkmalswerte vom Median beträgt 751,9763 €.

Aufgabe 2e: Schiefe einer H.V.

Mithilfe der Fechnerschen Lageregel lässt sich anhand der Mittelwerte folgende Schiefe der vorliegenden H.V. ermitteln:

$(X_{Mo} = 1\,300) < (X_{Me} = 1\,494{,}1473) < (\overline{X} = 1\,643{,}66)$

Damit liegt eine linkssteile oder rechtsschiefe H.V. vor.

Aufgabe 2f: Geometrisches Mittel

Die Veränderungsraten der Nettoeinkommen bauen in den Jahren 2012 bis 2014 aufeinander auf, d. h. die Wachstumsfaktoren sind multiplikativ verknüpft. Daher ist hier das geometrische Mittel anzuwenden.

Die Wachstumsfaktoren ermitteln sich als:

$WF_1 = 1{,}20 \quad WF_2 = 0{,}90 \quad WF_3 = 0{,}90$ mit: $n = 3$ Wachstumsfaktoren

Für den gesamten Wachstumsfaktor ergibt sich:

$$\overline{X}_G = \sqrt[3]{1{,}20 \cdot 0{,}90 \cdot 0{,}90} = 0{,}9906$$

Der durchschnittliche Wachstumsfaktor beträgt 0,9906.

Eine Rückumwandlung in die durchschnittliche Wachstumsrate ergibt:

$$\overline{W}_G = (0{,}9906 - 1) \cdot 100 = -0{,}94$$

Das Nettoeinkommen von Frau Müller hat sich in den Jahren 2012 – 2014 um durchschnittlich (–0,94 %) pro Jahr (p. a.) verringert.

Aufgabe 2g: Arithmetisches Mittel von Teilgesamtheiten

Die Nettoeinkommenssteigerungen der Eheleute finden parallel zueinander im Jahr 2012 statt, d. h. es liegt eine additiv (und keine multiplikative) Verknüpfung vor. Daher ist das gewogene arithmetische Mittel heranzuziehen, wobei die Wachstumsraten mit den Anteilen des Nenners der Wachstumsrate gewichtet werden. Eine Wachstumsrate weist im Nenner die Ausgangswerte, hier die Werte des Ausgangsjahres aus. Daher sind die Wachstumsraten mit den Einkommensanteilen im Ausgangsjahr zu gewichten, die 2/3 (Frau Müller) und 1/3 (Herr Müller) betragen. Somit ergibt sich für das gewogene arithmetische Mittel:

$$\overline{X} = 20\,\% \cdot \frac{2}{3} + 40\,\% \cdot \frac{1}{3} = 26{,}67\,\%$$

Die durchschnittliche Steigerung des Nettoeinkommens der Eheleute betrug 26,67 % im Jahr 2012.

Aufgabe 3: Mietwohnungen in Deutschland 2006

Aufgabe 3a: Arithmetisches Mittel der Miete

$$\overline{Y} == \frac{1}{n}\sum_{i=1}^{r} Y_j \cdot h_{.j} = \frac{1}{2\,191} \cdot (4{,}25 \cdot 240 + 4{,}75 \cdot 312 + 6 \cdot 1\,639) = 5{,}6303\ €$$

Aufgabe 3b: Standardabweichung der Miete

$$S_Y^2 = \frac{1}{n}\sum_{j=1}^{r} Y'^2_j \cdot h_{.j} - \overline{Y}^2 \quad \text{mit:} \quad \sum_{j=1}^{r} Y'^2_j \cdot h_{.j} = 70\,378{,}5 \quad \text{(siehe Hilfsangaben)}$$

$$S_Y^2 = \frac{1}{2\,191} \cdot 70\,378{,}5 - 5{,}6303^2 = 0{,}4255 \quad \text{somit: } S_Y = \sqrt{0{,}4255} = 0{,}6523$$

Aufgabe 3c: Bedingte relative Häufigkeit

$f(Y_1/X_1) = \frac{h_{11}}{h_{1.}} = \frac{33}{427} = 0{,}0773$ (bitte immer formal angeben, was berechnet wurde; hier: $f(Y_1/X_1)$)

Die (bedingte) relative Häufigkeit für eine Miete von „4 bis unter 4,5 €“ unter der Bedingung, dass die Mietwohnungen ein Alter von „5 bis unter 9 Jahren“ aufweisen, beträgt 7,73 %.

Aufgabe 3d: Theoretisch erwartete gemeinsame relative Häufigkeit

Unter der Annahme, dass zwischen dem Alter der Mietwohnungen und der Miete Unabhängigkeit besteht, lassen sich die erwarteten gemeinsamen relativen Häufigkeiten über die Randhäufigkeiten ermitteln. Für die Merkmalskombination (X_1, Y_1) ergibt sich für f_{11}^*:

$$f_{11}^* = f_{1.} \cdot f_{.1} = \frac{427}{2\,191} \cdot \frac{240}{2\,191} = 0{,}02135$$

Die theoretisch erwartete relative Häufigkeit beträgt 2,135 %.

Aufgabe 3e: Kovarianz

Die Kovarianz bestimmt sich formal wie folgt:

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{r} X_i \cdot Y_j \cdot h(X_i, Y_j) - \overline{X} \cdot \overline{Y}$$

Der Ansatz zur Berechnung der Kovarianz lautet in diesem Beispiel unter Beachtung der vorgegebenen arithmetischen Mittel für X und Y:

$$S_{XY} = \frac{1}{2\,191} \cdot [7 \cdot 4{,}25 \cdot 33 + 7 \cdot 4{,}75 \cdot 48 + 7 \cdot 6 \cdot 346 + 11{,}5 \cdot 4{,}25 \cdot 35 + 11{,}5 \cdot 4{,}75 \cdot 45$$
$$+ 11{,}5 \cdot 6 \cdot 311 + 20 \cdot 4{,}25 \cdot 172 + 20 \cdot 4{,}75 \cdot 219 + 20 \cdot 6 \cdot 982] - 15{,}9496 \cdot 5{,}6303$$

Aufgabe 3f: Kovarianz

In diesem Beispiel beträgt die Kovarianz $S_{XY} = -0{,}3429$ (siehe Vorgabe). Die Kovarianz erfasst die gemeinsame Streuung der Merkmale X und Y. Sie gibt somit Auskunft darüber, wie sich die Merkmalswerte von Y verändern, wenn die Merkmalswerte von X variieren und umgekehrt. Eine Kovarianz von –0,3429 impliziert, dass im vorliegenden Fall eine negative lineare Abhängigkeit gegeben ist. Dies bedeutet, dass mit zunehmendem Alter der Wohnungen die Miethöhe tendenziell abnimmt. Wie stark diese negative lineare Abhängigkeit ausgeprägt ist, lässt sich aus den Werten der Kovarianz nicht ablesen, da die Kovarianz keinen Maximalwert hat. Anders verhält es sich beim Bravais-Pearson Korrelationskoeffizienten: Er wird gebildet, indem die Kovarianz durch die Standardabweichungen der Merkmalswerte dividiert wird. Sein Vorteil besteht im Vergleich zur Kovarianz darin, dass er bei negativer Korrelation einen Maximalwert von r = –1 (bzw. bei positiver Korrelation einen Maximalwert r = +1) aufweist. Insoweit kann aus der Höhe von r auf die Stärke des linearen Zusammenhangs geschlossen werden. Weitere Einschränkungen der Aussagekraft der Kovarianz im Vergleich zum Korrelationskoeffizienten finden sich auf S. 244 ff im Buch.

Aufgabe 3g: Korrelationskoeffizient

$r = \frac{S_{XY}}{S_X \cdot S_Y}$ (mit: $S_X = 5{,}4244$; siehe Vorgabe; $S_Y = 0{,}6523$; siehe Aufgabe 3b)

$r = \frac{-0{,}3429}{5{,}4244 \cdot 0{,}6523} = -0{,}0969$

Da r sich bei einer negativen Korrelation grundsätzlich im Intervall $(-1 < r < 0)$ bewegt und hier ein Wert nahe „null" realisiert wird, besteht eine sehr schwache, negative, lineare Korrelation zwischen dem Alter der Mietwohnungen und der Miete.

Aufgabe 3h: Zusammenhangsmaß

Merkmal Z ist nominalskaliert, da beide Merkmalsausprägungen Z = 0 und Z = 1 gleichberechtigt nebeneinander stehen. Es kann lediglich gesagt werden, ob eine Verkehrsanbindung besteht oder nicht. Es ist keine Rangfolge gegeben, d. h. es lässt sich grundsätzlich nicht sagen, ob eine Anbindung besser oder schlechter ist als keine Anbindung (Hinweis: es wird hier nicht nach den Präferenzen gefragt, so dass eine Rangfolge nicht gegeben ist. Zudem kann eine aus verkehrspolitischer Sicht günstige Anbindung aus anderen Gründen nachteilig sein, so dass es hier um eine neutrale Beurteilung der Situation geht). Auch sind Abstände und ein mathematischer Nullpunkt nicht gegeben. Merkmal Y ist verhältnisskaliert. Da die schwächste Skalierung (hier: Nominalskala) das Zusammenhangsmaß bestimmt, ist unabhängig von der Skalierung von Y ein Zusammenhangsmaß für nominalskalierte Merkmalswerte anzuwenden. Dies kann z. B. der korrigierte Kontingenzkoeffizient nach Pearson sein.

Aufgabe 4: Regressionsanalyse am Beispiel eines Verlagsunternehmens

Aufgabe 4a: Steigungsparameter b_2

Hier ist lediglich der Steigungsparameter b_2 gesucht, der beschreibt, wie sich der Umsatz im Durchschnitt der Filialen entwickelt, wenn die Werbeausgaben um 1 Einheit (hier: 1 000 €) ansteigen.

$b_2 = \frac{S_{XY}}{S_X^2}$ ($S_{XY} = 0{,}1221$; siehe Vorgabe)

mit: $S_X^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$ mit: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Hieraus folgt: $b_2 = \frac{0{,}1221}{\frac{1}{11} \cdot 13{,}62 - \left(\frac{1}{11} \cdot 11\right)^2} = 0{,}5126$

Der Umsatz verändert sich um 0,51260 Mio. €, wenn die Werbeausgaben um 1 000 € ansteigen.

Aufgabe 4b: Kleinstes-Quadrate-Verfahren (K-Q-V)

Diese Aussage ist nicht korrekt. Die Voraussetzung $\sum e_i = 0$ ist nicht eindeutig, da es viele Regressionsfunktionen gibt, die diese Bedingung erfüllen. Dies ist dadurch bedingt, dass sich bei alternativer Wahl der Regressionsfunktion jeweils Residuen errechnen, deren unterschiedliche positiven und negativen Werte sich immer zu „null" aufaddieren (sogenanntes „Plus-Minus-Problem", das häufiger in der Statistik auftritt). Daher kommt beim Kleinste-Quadrate-Verfahren folgendes Kriterium für die Auswahl der Regressionsfunktion zu Anwendung: Die Summe der Residuen wird quadriert und anschließend minimiert ($\sum e_i^2 = \text{Minimum}$).

Aufgabe 4c: Bravais-Pearson Korrelationskoeffizient

Steigen die Werbeausgaben (X), so steigt auch der Umsatz (Y). Infolge dessen liegt eine positive Korrelation vor, und der Bravais-Pearson Korrelationskoeffizient besitzt ein positives Vorzeichen.

Es herrscht **keine perfekte** positive Korrelation, da nicht alle Punkte auf einer Geraden liegen. Somit wird r nicht den Maximalwert „r = +1" erreichen, sondern im Intervall $(0 < r < 1)$ liegen.

Musterklausur II, S. 374 (Zur Bearbeitung siehe die analogen Hinweise der Musterklausur I)

Aufgabe 1: Statistische Begriffe und Formeln im Lichte der Fußball-WM 2014

Aufgabe 1a: Merkmalsträger, Merkmal, Merkmalsausprägung

Merkmalsträger	Personenkraftwagen, die im Hinblick auf das Merkmal „Fußball-Fanartikel“ untersucht werden.
Merkmal	PKW-Fanartikel
Merkmalsausprägung	konkreter PKW-Fanartikel, z. B. „nur Spiegel Cover, 2er-Set für Deutschland“ oder „nur Autoflagge, 1er-Bestückung für Hitzfelds Schweizer Mannschaft“

Aufgabe 1b: Aussage zur Skalierung richtig oder falsch?

Diese Aussage ist fehlerhaft.

Es handelt sich hierbei um ein nominalskaliertes Merkmal, da die verschiedenen Merkmalsausprägungen gleichberechtigt nebeneinander stehen. Es kann für einen neutralen Beobachter (auch diese soll es bei einer Fußball-WM geben) nicht gesagt werden, dass zum Beispiel die Ausprägung. „nur Spiegel Cover, 2er-Set für Deutschland“ besser oder schlechter ist als z. B. die Merkmalsausprägung „nur Autoflagge, 1er-Bestückung für Hitzfelds Schweizer Mannschaft“. Damit besteht weder eine Rangfolge, noch sind Abstände gegeben. Auch existiert kein mathematischer Nullpunkt.

Die Behauptung, dass die Merkmalsausprägung „mehrere Fanartikel gleichzeitig“ doppelt so häufig vorkommt, wie „nur Spiegel Cover“ ist korrekt. Hierbei handelt es sich aber um die relativen Häufigkeiten, die bei jeder Skala, d.h. auch bei einer Nominalskala gebildet werden können. Für die Frage der Skalierung ist aber nicht die relative Häufigkeit, sondern die Merkmalsausprägung selbst entscheidend. Die Merkmalsausprägung lässt nur eine Aussage im Sinne von „eine bestimmte Merkmalsausprägung ist vorhanden oder nicht“ zu. Damit liegt hier eine Nominalskala vor.

Aufgabe 1c: Aussage zum Modus richtig oder falsch?

Diese Aussage ist falsch. Der Modus beschreibt die am häufigsten vorkommende **Merkmalsausprägung**. Hier wurde jedoch die relative Häufigkeit selbst als Modus angeführt. Ein Modus ist ein Mittelwert und dieser kann nur eine Merkmalsausprägung und nicht eine Häufigkeit darstellen!

Die richtige Lösung lautet: $X_{Mo} =$ "mehrere Fanartikel gleichzeitig"

(Hinweis: Bitte in dieser Form stets eindeutig anführen; unzureichend ist eine oberflächige, nicht genau zuzuordnende Antwort gemäß folgender Aussage: „Der Modus ist der häufigste Wert, hier 50 %“; gemeint bei dieser Antwort ist: „Der Modus lautet "mehrere Fanartikel gleichzeitig", da diese Merkmalsausprägung am häufigsten vorkommt, nämlich in 50 % aller Ausprägungen“

Aufgabe 1d: 1. Eigenschaft des arithmetischen Mittels

Diese Formel besagt, dass die aufsummierten einfachen Abweichungen der Merkmalswerte von ihrem arithmetischen Mittel immer „null“ ergeben (vgl. auch S. 128 f im Buch, wo sich auch ein Zahlenbeispiel findet).

$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$. Es gilt:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - n \cdot \overline{X} = \sum_{i=1}^{n} X_i - n \cdot \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i = 0$$

Diese Aussage besagt, dass die Summe der Abweichungen der Merkmalswerte vom arithmetischen Mittel (Bezugsgröße) immer „null“ ergibt.

Hinweis: Diese Eigenschaft ist deshalb von zentraler Bedeutung, da die Abweichungen der Merkmalswerte vom Zentrum (arithm. Mittel) als Streuungsmaß genutzt werden. Damit die Summe dieser Abweichungen wegen der „Plus-Minus“ Problematik nicht immer „null“ ergibt, beseitigen die verschiedenen Streuungsmaße das Vorzeichen durch Absolutbeträge oder Quadrierungen.

Aufgabe 2: Mikrozensus 2013

Aufgabe 2a: Berechnung der Mittelwerte

Für die Berechnung der Mittelwerte werden folgende Größen benötigt:

Zum Modus:

normierte Klassenbreite; hier. $\Delta X^n = 500$, da bei einer Klassenbreite von 500 gilt: $f_4 = d_4$

$$d_3 = \frac{f_3}{\Delta X_3} \cdot \Delta X^n = \frac{0{,}0777}{200} \cdot 500 = 0{,}19425$$

Die größte Dichte liegt mit d_3 in der 3. Klasse vor. Die Klassenmitte (X'_3) dieser Klasse bildet den Modus, somit:

$$X_{Mo} = \frac{1300 + 1500}{2} = 1400$$

Zum Median:

$F(X_4^0) = F_5 - f_5 = 0{,}6617 - 0{,}1529 = 0{,}5088$

$F(X_3^0) = F_4 - f_4 = 0{,}5088 - 0{,}1679 = 0{,}3409$

Bei klassifizierten Daten erfolgt eine „Feinberechnung" des Median für $(F_i = 0{,}5)$; der Median liegt somit in der 4. Klasse (etwa knapp unter der Obergrenze der Klasse, da $F(X_4^o) = 0{,}5088$)

$$X_{Me} = X_4^u + \Delta X_4 \cdot \frac{0{,}5 - [F(X_4^u) = F(X_3^0)]}{f(X_4)} = 1\,500 + 500 \cdot \frac{0{,}5 - 0{,}3409}{0{,}1679} = 1\,973{,}7939$$

Zum arithmetischen Mittel:

$\overline{X} = \sum_{i=1}^{m} X'_i \cdot f_i = 2\,701{,}75$ (siehe Hilfsangabe). Somit gilt: $\overline{X} = 2\,701{,}75$

Aufgabe 2b: Graphische Darstellung der H.V.

In dieser Aufgabe liegen unterschiedlichen Klassenbreiten vor. Die graphische Darstellung der Häufigkeitsverteilung (Histogramm) muss flächenproportional erfolgen. Eine flächenproportionale Darstellung ist bei unterschiedlichen Klassenbreiten nur dann gegeben, wenn das Histogramm mittels der Dichte (hier relative Dichte) dargestellt wird.

Begründung: Die Häufigkeiten steigen in den Klassen überproportional an, in denen eine im Vergleich zu den anderen Klassen überproportional große Klassenbreite vorliegt. Umgekehrt fallen die relativen Häufigkeiten in den Klassen unterproportional aus, in denen unterproportional kleine Klassenbreiten vorliegen. Werden die Häufigkeiten und nicht die Dichten höhenproportional dargestellt, steigen auch die Säulen des Histogramms in den Klassen mit größerer Klassenbreite überproportional an und umgekehrt. Dies ist bei einer Darstellung des Histogramms über die Dichte nicht der Fall, da die Dichte eine künstlich berechnete Häufigkeit je einheitlicher Klassenbreite darstellt (siehe Beispiel vom Schmetterlingskescher im Buch, S. 88 ff, insbesondere S. 91).

Aufgabe 2c: MAD

$$MAD(\overline{X}) = \frac{1}{n}\sum_{i=1}^{8} |X'_i - \overline{X}| \cdot h_i = \frac{1}{n}(|X'_1 - \overline{X}| \cdot h_1 + \sum_{i=2}^{8} |X'_i - \overline{X}| \cdot h_i)$$

mit: $X'_1 = 550$ (Berechnung); $\sum_{i=2}^{8} |X'_i - \overline{X}| \cdot h_i = 54\,209{,}0351$ (siehe Hilfsangaben)

$$MAD(\overline{X}) = \frac{1}{39{,}9}|550 - 2701{,}75| \cdot 4{,}9 + 54\,209{,}0351 = \frac{1}{39{,}9}(10543{,}575 + 54\,209{,}0351) = 1622{,}8724$$

Die durchschnittliche (absolute) Abweichung der Haushaltsnettoeinkommen der Erwerbstätigen vom arithmetischen Mittel 2 701,75 € (Durchschnittseinkommen) beträgt 1 622,8724 €.

Aufgabe 2d: Schiefe der Häufigkeitsverteilung

Mithilfe der Fechnerschen Lageregel lässt sich anhand der Mittelwerte folgende Schiefe der hier vorliegenden H.V. ermitteln:

$(X_{Mo} = 1\,400) < (X_{Me} = 1\,973{,}7939) < (\overline{X} = 2\,701{,}75)$

Damit liegt eine linkssteile oder rechtsschiefe H.V. vor.

(**Hinweis**: Weitere Fragestellungen könnten lauten: „Stellen Sie die vorliegende H.V. als Histogramm graphisch dar. Eine grobe graphische Skizze mit den markanten Punkten ist ausreichend. Allerdings sollte Ihre Skizze die Schiefe der H.V. und die Mittelwerte deutlich erkennen lassen. Die Achsenbeschriftungen sollten vollständig erfasst sein. Bedenken Sie bei der Darstellung des Histogramms auch, dass unterschiedliche Klassenbreiten vorliegen".
Hinweis zur Antwort: Es ist hier die Dichte zu verwenden, wenn unterschiedliche Klassenbreiten dargestellt werden. Zur Darstellung eines linkssteilen Histogramms bei einheitlicher Klassenbreite und unter Verwendung von relativen Häufigkeiten vgl. die Abb. II-3-12a auf S. 154 im Buch.)

Aufgabe 2e: Feinberechnung der Anteilswerte

Zum Anteilswert der Erwerbstätigen, die weniger als das Medianeinkommen zur Verfügung haben: Der Median zerlegt die H.V. in zwei gleich große Hälften ($F_i = 0{,}5$). Somit beträgt der Anteilswert der Erwerbstätigen, die weniger als das Medianeinkommen zur Verfügung haben, 50 %.

Zum Anteilswert der Erwerbstätigen, die mehr als das arithmetische Mittel $\overline{X} = 2\,701{,}75$ zur Verfügung haben: Zunächst wird der Anteil der Erwerbstätigen berechnet, die ein Einkommen **unterhalb** des arithmetischen Mittels $\overline{X} = 2\,701{,}75$ aufweisen.

$$F(X \leq 2\,701{,}75) = F(X_6^u) + f_6 \cdot \frac{2\,701{,}75 - X_6^u}{\Delta X_6} = 0{,}6617 + 0{,}1654 \cdot \frac{2\,701{,}75 - 2600}{1000} = 0{,}6785$$

Damit ermittelt sich Anteilswert der Erwerbstätigen, die **mehr** als das arithmetische Mittel $\overline{X} = 2\,701{,}75$ aufweisen als: $F(X > 2\,701{,}75) = 1 - 0{,}6785 = 0{,}3215$

Somit gilt: $F(X_{Me} \geq X \geq \overline{X}) = 0{,}5 + 0{,}3215 = 0{,}8215$ (also 82,15 %).

82,15 % der Erwerbstätigen verdienen im Hinblick auf das Haushaltsnettoeinkommen weniger als das Medianeinkommen und mehr als das Durchschnittsnettoeinkommen.

Aufgabe 3: Fußballweltmeisterschaft 2014 in Brasilien

Aufgabe 3a: Skalierungen von Merkmal X und Merkmal Y

Merkmal X: Herkunft der Mannschaften

Das Merkmal X ist nominalskaliert, da die verschiedenen Herkunftsländer der Fußballmannschaften gleichberechtigt nebeneinander stehen. Es kann z. B. nur gesagt werden, ob eine Mannschaft aus einem bestimmten Herkunftsland (Kontinent) wie z. B. Afrika oder aus einem anderen Herkunftsland (Kontinent) kommt. Eine Rangfolge der Herkunftsländer (Kontinente) beispielsweise dergestalt, dass „ eine Mannschaft aus Afrika besser oder schlechter ist als eine Mannschaft aus Süd/Osteuropa" kann aufgrund der Angaben des Herkunftslandes nicht getroffen werden. Auch lassen sich zu den Namen der Herkunftsländer keine Abstände oder einen mathematischen Nullpunkt definieren.

Merkmal Y: Platzierung der Mannschaften am Ende der Vorrunde

Hier liegt eine Ordinalskala vor, denn es ist eine Rangfolge („höher bzw. niedriger platziert") gegeben; Platz 1 ist beispielsweise besser als Platz 2 und nachfolgende Rangplätze. Allerdings lassen sich aus den Rängen keine Rückschlüsse daraus gewinnen, wie sehr die Mannschaften mit unterschiedlichen Rängen sich in ihren Leistungsfähigkeiten unterscheiden, d.h. Abstände sind nicht gegeben (eine solche Angabe ist lediglich anhand der Torbilanz bzw. Punktebilanz möglich). Zudem lässt die Angabe zur Rangfolge auch keine Aussage zum mathematischen Nullpunkt zu. Somit sind auch keine Aussagen zur relativen Leistungsfähigkeit in dem Sinne möglich, dass beispielsweise eine Mannschaft auf Rang 4 nur halb so gut abgeschnitten hat wie eine Mannschaft auf Rang 2.

Aufgabe 3b: Relative Häufigkeit und bedingte relative Häufigkeit

Relative Häufigkeit: $f(X_2, Y_1) = \frac{h(X_2, Y_1)}{n} = \frac{h_{21}}{n} = \frac{4}{32} = 0{,}125\ (12{,}5\ \%)$

(Interpretation als zusätzlicher Hinweis: Relative empirische Häufigkeit für eine Mannschaften aus Mittel-/Südamerika, in der Vorrunde den ersten Platz zu erlangen. 12,5 % der Mannschaften, die in der Vorrunde den ersten Rang erreichten, kamen aus Mittel-/Südamerika)

Bedingte relative Häufigkeit:

$f(Y_1/X_2) = \frac{h(X_2, Y_1)}{h(X_2)} = \frac{h_{21}}{h_{2.}} = \frac{4}{9} = 0{,}445\ (44{,}5\ \%)$

(Interpretation als zusätzlicher Hinweis: Die relative Häufigkeit, in der Vorrunde den ersten Platz zu erlangen, wenn es sich um eine Mannschaft aus Mittel-/Südamerika handelte, betrug 44,5 % (im Vergleich zur relativen Häufigkeit, d. h. im Vergleich zur nicht bedingten relativen Häufigkeit von 12,5 %)

(Hinweis zu einer modifizierten Frage: Bei Frage 3a wurde hier der gesuchte Wert formal angegeben. Die Frage könnte aber auch verbal auf Basis der Interpretation des gesuchten Wertes formuliert sein, also z. B. „Wie groß ist die relative Häufigkeit, in der Vorrunde der Fußballweltmeisterschaft den ersten Platz zu erlangen, unter der Bedingung, dass es sich um eine Mannschaft aus Mittel-/Südamerika handelt." Bei der Antwort ist formal auch anzugeben, welche statistische Größe gesucht wird, also: $f(Y_1/X_2)$; es reicht nicht aus, lediglich den Rechengang für die gesuchte Größe darzustellen.)

Aufgabe 3c: Bedingte relative Häufigkeiten und Randhäufigkeiten bei Unabhängigkeit

$f(X_1/Y_1) = f(X_1/Y_2) = f(X_1/Y_3) = f(X_1/Y_3) = f(X_1/Y_4) = f(X_1)$

(Hinweis 1: Eine zusätzliche Frage nach der Interpretation der Gleichung kann wie folgt beantwortet werden: „Sind die Merkmale X und Y unabhängig voneinander, dann ist die bedingte relative Häufigkeit für die erste Ausprägung des Merkmals X (also: X_1 = Mannschaft aus Afrika) unabhängig von der Bedingung, ob die Mannschaft den ersten, zweiten oder weiteren Rang (also Y_1, Y_2, …) erzielt hat. Die bedingte relative Häufigkeit für die erste Ausprägung des Merkmals X (also: X_1 = Mannschaft aus Afrika) wird nur durch die relative Häufigkeit der Mannschaft aus Afrika, also $f(X_1)$ geprägt und nicht durch den erzielten Rang in der Vorrunde).

(Hinweis 2: Ist die Unabhängigkeitsbedingung für $[f(X_1/Y_j)\ (j = 1, \ldots, 4)]$ erfüllt, dann gelten auch für andere X_i ($i = 2, \ldots 5$) die Unabhängigkeitsbedingungen. Außerdem sind spiegelbildlich auch für $f(Y_j/X_i)$ analoge Bedingungen erfüllt.)

Aufgabe 3d: Theoretisch erwartete gemeinsame absolute Häufigkeit

Besteht Unabhängigkeit zwischen den beiden Merkmalen X und Y, so lässt sich die theoretisch erwartete gemeinsame absolute Häufigkeit (h_{22}^*) für den 2. Rang (Y_2) der Mannschaft aus Mittel-/Südamerika (X_2) über die jeweiligen absoluten Randhäufigkeiten und die Zahl der Merkmalsträger (n) wie folgt ermitteln:

$h_{22}^* = \frac{h_{2.} \cdot h_{.2}}{n} = \frac{9 \cdot 8}{32} = 2{,}25$

(Zusätzlicher Hinweis: Hier weicht die theoretisch erwartete Häufigkeit ($h_{22}^* = 2{,}25$) von der empirisch beobachteten gemeinsamen Häufigkeit ($h_{22} = 3$) ab. Dies deutet darauf hin, dass – von zufälligen Abweichungen abgesehen – die Merkmale X und Y eine gewisse Abhängigkeit aufweisen, d. h. Rangplatz und Herkunftsland (hier: für Mittel-/Südamerika (X_2)) nicht unabhängig voneinander sind. Ein geeignetes Maß zur Messung der Stärke des Zusammenhangs zwischen den beiden Merkmalen stellt der korrigierte Kontingenzkoeffizient nach Pearson dar. Inwieweit der Zufall für die Abweichungen zwischen den empirischen und den theoretisch erwarteten Häufigkeiten verantwortlich ist, lässt sich nur mit einem Test der Schließenden Statistik wie z.B. dem $\chi^2 - Unabhängigkeitstest$ beantworten; vgl. hierzu Teil D des Anhangs auf S. 368.)

Aufgabe 3e: Korrigierter Kontingenzkoeffizient und Fußballweltmeisterschaft 2014

Die Berechnung des korrigierten Kontingenzkoeffizienten C_{korr} erfolgt über die Größe χ^2, welche die quadrierten Abweichungen der empirischen von ihren theoretisch erwarteten Häufigkeiten aufsummiert und durch die jeweilige theoretisch erwartete Häufigkeit dividiert. Die wegen identischer Randhäufigkeiten z. T. identischen Ergebnisse für die theoretisch erwarteten Häufigkeiten lauten:

(Hinweis: Da sich die Rechengänge stark wiederholen, reicht es in der Klausur aus, wenn der Rechengang nur für unterschiedliche Randhäufigkeiten explizit angegeben wird und ansonsten nur das Ergebnis für h_{ij}^* dargestellt wird

z. B. $h_{11}^* = \frac{h_{1.} \cdot h_{.1}}{n} = \frac{5 \cdot 8}{32} = 1{,}25$; ebenso gilt für alle weiteren Kombinationen h_{12}^* bis $h_{14}^* = 1{,}25$,

da die Randhäufigkeiten jeweils identisch sind) . Im Folgenden seien dennoch die vollständigen Rechengänge wiedergegeben)

$$h_{11}^* = \frac{h_{1.} \cdot h_{.1}}{n} = \frac{5 \cdot 8}{32} = 1{,}25 \quad h_{12}^* = \frac{h_{1.} \cdot h_{.2}}{n} = \frac{5 \cdot 8}{32} = 1{,}25 \quad h_{13}^* = \frac{h_{1.} \cdot h_{.3}}{n} = \frac{5 \cdot 8}{32} = 1{,}25$$

$$h_{14}^* = \frac{h_{1.} \cdot h_{.4}}{n} = \frac{5 \cdot 8}{32} = 1{,}25$$

$$h_{21}^* = \frac{h_{2.} \cdot h_{.1}}{n} = \frac{9 \cdot 8}{32} = 2{,}25 \quad h_{22}^* = \frac{h_{2.} \cdot h_{.2}}{n} = \frac{9 \cdot 8}{32} = 2{,}25$$

$$h_{23}^* = \frac{h_{2.} \cdot h_{.3}}{n} = \frac{9 \cdot 8}{32} = 2{,}25 \quad h_{24}^* = \frac{h_{2.} \cdot h_{.4}}{n} = \frac{9 \cdot 8}{32} = 2{,}25$$

$$h_{31}^* = \frac{h_{3.} \cdot h_{.1}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{32}^* = \frac{h_{3.} \cdot h_{.2}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{33}^* = \frac{h_{3.} \cdot h_{.3}}{n} = \frac{6 \cdot 8}{32} = 1{,}5$$

$$h_{34}^* = \frac{h_{3.} \cdot h_{.4}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{41}^* = \frac{h_{4.} \cdot h_{.1}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{42}^* = \frac{h_{4.} \cdot h_{.2}}{n} = \frac{6 \cdot 8}{32} = 1{,}5$$

$$h_{43}^* = \frac{h_{4.} \cdot h_{.3}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{44}^* = \frac{h_{4.} \cdot h_{.4}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{51}^* = \frac{h_{5.} \cdot h_{.1}}{n} = \frac{6 \cdot 8}{32} = 1{,}5$$

$$h_{52}^* = \frac{h_{5.} \cdot h_{.2}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{53}^* = \frac{h_{5.} \cdot h_{.3}}{n} = \frac{6 \cdot 8}{32} = 1{,}5 \quad h_{54}^* = \frac{h_{5.} \cdot h_{.4}}{n} = \frac{6 \cdot 8}{32} = 1{,}5$$

Anschließend ist Chi-Quadrat zu bilden: $\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{r} \frac{\left(h_{ij} - h_{ij}^*\right)^2}{h_{ij}^*}$

$$\chi^2 = \frac{(0-1{,}25)^2}{1{,}25} + \frac{(2-1{,}25)^2}{1{,}25} + \frac{(1-1{,}25)^2}{1{,}25} + \frac{(2-1{,}25)^2}{1{,}25} + \frac{(4-2{,}25)^2}{2{,}25} + \frac{(3-2{,}25)^2}{2{,}25}$$
$$+ \frac{(1-2{,}25)^2}{2{,}25} + \frac{(1-2{,}25)^2}{2{,}25} + \frac{(0-1{,}5)^2}{1{,}5} + \frac{(1-1{,}5)^2}{1{,}5} + \frac{(5-1{,}5)^2}{1{,}5} + \frac{(0-1{,}5)^2}{1{,}5}$$
$$+ \frac{(4-1{,}5)^2}{1{,}5} + \frac{(1-1{,}5)^2}{1{,}5} + \frac{(0-1{,}5)^2}{1{,}5} + \frac{(1-1{,}5)^2}{1{,}5} + \frac{(4-1{,}5)^2}{1{,}5} + \frac{(0-1{,}5)^2}{1{,}5} + \frac{(1-1{,}5)^2}{1{,}5}$$
$$+ \frac{(1-1{,}5)^2}{1{,}5} + \frac{(4-1{,}5)^2}{1{,}5}$$

$$\chi^2 = 1{,}25 + 0{,}45 + 0{,}05 + 0{,}45 + 1{,}361 + 0{,}25 + 0{,}694 + 0{,}694 + 1{,}5 + 0{,}17 + 8{,}17 + 1{,}5$$
$$+4{,}17 + 0{,}17 + 1{,}5 + 0{,}17 + 1{,}5 + 0{,}17 + 0{,}17 + 4{,}17 = 28{,}559$$

Mittels χ^2 lässt sich nun der korrigierte Kontingenzkoeffizient bilden (gerundete Werte):

$$C_{korr} = \sqrt{\frac{\chi^2}{\chi^2 + n} \cdot \frac{C^*}{C^* - 1}} \quad \text{mit: } C^* = \text{Min}(m, r) = 4;\ C_{korr} = \sqrt{\frac{28{,}559}{28{,}559 + 32} \cdot \frac{4}{3}} = 0{,}7930$$

Damit liegt ein starker Zusammenhang zwischen der Herkunft der Mannschaft und ihrer Platzierung vor, da der Wert $C_{korr.} = 0{,}7930$ im oberen Bereich des möglichen Intervalls ($0 \leq C_{korr} \leq 1$) liegt.

Aufgabe 3f: Mittelwerte der Rangplätze in den Vorrundenspielen der Fußball-WM 2014

Sollen die Mittelwerte für das ordinalskalierte Merkmal Y (Rangplatz) für die Länder aus Mittel /Südamerika bestimmt werden, so lassen sich nur Modus und Median bilden. Der Modus lässt sich für alle Skalen berechnen; der Median setzt eine Ordinalskala voraus, da Rangfolgen ermittelt werden müssen. Das arithmetische Mittel kann erst ab einer Intervallskalierung gebildet werden, da hierzu Abstände zwischen den Merkmalsausprägungen definiert sein müssen. Daher lässt sich bei der vorliegenden Ordinalskala kein arithmetisches Mittel bilden (eine Aussage im Sinne „der durchschnittliche Rang betrug…" ist hier also nicht möglich; Fettnäpfchen der Statistik beim „Volksmund"!)

Zum Modus:

Hier liegen nichtklassifizierte Daten vor; daher wird der Modus durch die häufigste Merkmalsausprägung bestimmt; der Modus lautet: $\mathbf{X_{Mo} = Platz\ 1}$ (da die häufigste Merkmalsausprägung für die Länder Mittel-/Südamerika (X_2) die absolute Häufigkeit ($h_{21} = 4$) beträgt.)

Zum Median:

Hier liegt eine H.V. mit einer **ungeraden** Beobachtungszahl (n = 9) vor, daher lässt sich der Median ermitteln über:

$X_{Me} = X_{\left[\frac{n+1}{2}\right]} = X_{\left[\frac{9+1}{2}\right]} = X_{[5]}$ mit: $X_{[5]} =$ Platz 2

Der Median stellt also die Merkmalsausprägung des fünften Merkmalsträgers der geordneten Urliste dar, folglich: $\mathbf{X_{Me} = Platz\ 2}$

Aufgabe 4: Umsätze in Abhängigkeit von den Marketingaufwendungen

Aufgabe 4a: Durchschnittlicher Umsatz

Bevor der durchschnittliche Umsatz ($\overline{Y}$) ermittelt werden kann, müssen die Randhäufigkeiten des Merkmals Y gebildet werden: $h_{.1} = 6 \quad h_{.2} = 12 \quad h_{.3} = 12$

$$\overline{Y} = \frac{1}{30} \cdot (100 \cdot 6 + 200 \cdot 12 + 300 \cdot 12) = 220$$

Der durchschnittliche Umsatz des Unternehmens beträgt 220 Mio. €.

Aufgabe 4b: Bravais-Pearson Korrelationskoeffizient

$r = \frac{S_{XY}}{S_X \cdot S_Y}$ (siehe Hilfsangaben)

Hier sind Kovarianz und die Standardabweichung der Marketingaufwendungen bereits gegeben; somit muss die Standardabweichung des Umsatzes noch berechnet werden.

$S_Y = \sqrt{\frac{1}{30} \cdot 1\,620\,000 - 220^2} = 74{,}8331$ Daraus folgt für r:

$$r = \frac{81{,}3333}{1{,}6111 \cdot 74{,}8331} = 0{,}6746$$

Der Bravais-Pearson-Korrelationskoeffizienten (r) nimmt bei positiver Abhängigkeit von X und Y einen Wert zwischen ($0 < r \leq 1$). Im vorliegenden Beispiel bewegt sich (r) im mittleren Bereich des Intervalls, so dass eine mittlere, positive lineare Korrelation zwischen den Marketingaufwendungen und dem Umsatz besteht.

Aufgabe 4c: Richtig oder falsche Aussagen zur Kovarianz?

Alle drei Aussagen sind falsch:

- Die Kovarianz stellt keine dimensionslose Kennzahl dar, da die beiden Dimensionen der Merkmale multiplikativ verknüpft sind (hier z. B. [$€ \cdot € = €^2$]).

(Noch Aufgabe 4c):

- Weiterhin kann die Kovarianz nur für metrisch skalierte Merkmale ermittelt werden, damit die erforderlichen Rechenoperationen (Differenzen; Produkte) für die Merkmalswerte X und Y möglich sind. Lediglich Zusammenhangsmaße, die für nominalskalierte Merkmale geeignet sind (wie z. B. der korrigierte Kontingenzkoeffizient nach Pearson) lassen sich auf alle Skalen anwenden.
- Auch kann sich die Kovarianz im negativen Bereich bewegen, wenn zwischen den Merkmalswerten X und Y eine negative Abhängigkeit vorliegt. Anders als bei der Varianz, bei der die negativen Abweichungen der Merkmalswerte durch die Quadrierung in positive Abweichungen umgewandelt werden, bleiben bei der Kovarianz bei einer negativen Abhängigkeit die negativen Vorzeichen erhalten (positive oder negative Abweichungen der Merkmalswerte X bzw. Y von ihren jeweiligen Durchschnitten werden multipliziert, aber nicht quadriert).

Aufgabe 4d: Kleinste Quadrate Verfahren (K-Q-V)

Hier ist lediglich der Steigungsparameter b_2 zu ermitteln, da dieser die Umsatzentwicklung in Abhängigkeit von der Veränderung der Marketingaufwendungen beschreibt.

$$b_2 = \frac{S_{XY}}{S_X^2} \quad \text{(siehe Hilfsangaben; } S_{XY} = 81{,}3333;\ S_X = 1{,}61112)$$

$$b_2 = \frac{81{,}3333}{1{,}6111^2} = 31{,}3345$$

Erhöhen sich die Marketingaufwendungen um 1 Einheit, d. h. um 1 Mio. €, so steigt der Umsatz um 31,3345 Mio. € an.

(Hinweis: Die Regressionsanalyse wird für Einzelwerte berechnet; dabei muss eine ausreichende Zahl an Merkmalskombinationen gegeben sein, die auch eine gewisse Streuung aufweisen. Je größer die Anzahl der Merkmalswerte, desto weniger zufallsbedingt sind die Ergebnisse. Die Güte der Regressionsfunktion steigt c. p., je mehr die Merkmalswerte streuen (vgl. die Formel für Bestimmtheitsmaß im Buch auf S. 296 ff, insbesondere S. 297). Liegen die Merkmalswerte als H.V. vor, d. h. kommen die Merkmalskombinationen häufiger vor (d.h. sind die h_{ij} hoch) und werden diese Häufigkeiten in Einzelwerte übersetzt, so kommen bestimmte Merkmalswerte mehrfach vor. Dies bedeutet gleichzeitig, dass die Streuung der Merkmalswerte abnimmt. Dadurch verschlechtert sich der „fit" der Regressionsfunktion und bewirkt c. p., dass der Erklärungswert (Bestimmtheitsmaß) der Regressionsfunktion abnimmt.

Aufgabe 4e: Niveauparameter der Regressionsfunktion

Bei Marketingaufwendungen von „0 €" würde das absolute Glied b_1 die Höhe des Umsatzes wiedergeben: $b_1 = \overline{Y} - b_2 \cdot \overline{X} = 220 - 31{,}3345 \cdot 3{,}2667 = 117{,}6396$. Bei Marketingaufwendungen von „0" € würde sich für das Unternehmen ein Umsatz von 117,6396 Mio. € ergeben.

Aufgabe 4f: Kovarianz und Korrelationskoeffizient

Die Kovarianz würde hier einen negativen Wert von $S_{XY} < 0$ annehmen. Eine negative Kovarianz bedeutet für das konkrete Beispiel, dass mit steigendem Preis des Gutes der Absatz des Gutes abnimmt. Wie stark diese negative lineare Abhängigkeit ausgeprägt ist, lässt sich aus den Werten der Kovarianz nicht ablesen, da die Kovarianz keinen Maximalwert hat. Anders verhält es sich beim Bravais-Pearson Korrelationskoeffizienten: Er wird gebildet, indem die Kovarianz durch die Standardabweichungen der Merkmalswerte dividiert wird. Sein Vorteil besteht im Vergleich zur Kovarianz darin, dass er bei negativer Korrelation einen Maximalwert von $r = -1$ (bzw. bei positiver Korrelation einen Maximalwert $r = +1$) aufweisen kann. Im vorliegenden Beispiel liegt allerdings eine **perfekte <u>nicht</u>lineare** Abhängigkeit (und damit **keine perfekt lineare** Abhängigkeit) vor. Wird durch die Punktwolke eine Gerade gelegt, so werden die Merkmalswerte nicht alle auf einer Geraden liegen. Dies hat zur Folge, dass der Korrelationskoeffizient nicht den Maximalwert ($r = -1$) erreicht, d. h. er wird im Intervall ($-1 < r < 0$) liegen (nahe, aber nicht gleich $r = -1$).

Der Bravais-Pearson Korrelationskoeffizient hat analog zur Kovarianz ein negatives Vorzeichen.