Zusatzaufgaben I – Lösungen

Aufgabe 1)

Bei einem Wettbewerb erreichen die Teilnehmer folgenden Punktzahlen:
8; 8; 9; 13; 14; 16; 17; 20; 21; 22

Bestimmen bzw. berechnen Sie:
a) Modalwert
b) Spannweite
c) arithmetisches Mittel
d) Median
e) Standardabweichung
f) Zeichnen Sie ein Box-Plot

Lösungen:

a) $h = 8$

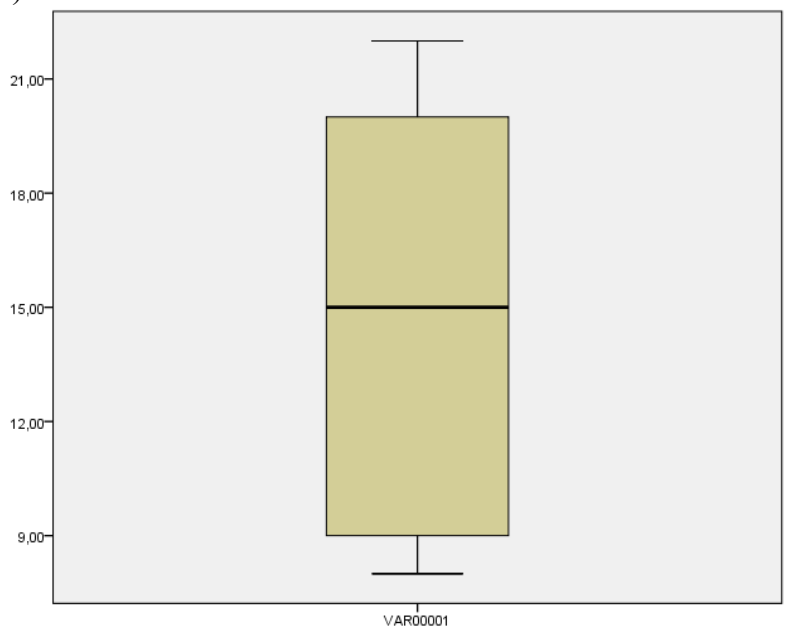
b) $R = 14$

c) $\bar{x} = 14{,}8$

d) $\tilde{x} = 15$

e) $s_X = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2} = 5{,}04$

f)



Aufgabe 2)

In Anschluss einer Gesundheitsstudie wurde mit Hilfe eines Regressionsmodells der Effekt des Einkommens auf den durchschnittlichen Zigarettenkonsum überprüft. Das Ergebnis der Schätzgleichung lautet: $\hat{y} = 14{,}452 - 0{,}002x$

a)
Interpretieren Sie das Ergebnis!

b)
Welchen durchschnittlichen Zigarettenkonsum sagt das Modell für eine Person mit einem Einkommen von 7000€ voraus?

Lösung:

a)
Die Regressionskonstante der Schätzgleichung verdeutlicht, dass bei einem (hypothetischen) Einkommen von 0 Euro der durchschnittliche Zigarettenkonsum bei etwa 14,5 Zigaretten am Tag liegt.
Der Regressionskoeffizient weist daraufhin, dass der Zigarettenkonsum mit zunehmendem Einkommen sinkt – und zwar um 0,002 Zigaretten am Tag je Euro mehr.

b)
Laut Modell konsumiert eine Person mit einem Einkommen von 7000€ durchschnittlich 0,452 Zigaretten am Tag.
14,452 – 0,002*7000 = 0,452

Aufgabe 3)

Zehn Personen unterschiedlichen Alters haben an einem Sprintwettbewerb teilgenommen. Obwohl die Zeiten nicht genau gemessen werden konnten, war es dennoch möglich zu bestimmen wer der Schnellste, Zweitschnellste usw. war. Neben den Rangplätzen sind in der folgenden Tabelle noch die Altersangaben notiert:

Person	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Rangplatz (X)	7	3	9	10	1	5	4	6	2	8
Alter (Y)	45	31	33	51	29	27	43	39	22	37

Überprüfen Sie mit Hilfe eines geeigneten Koeffizienten, ob ein Zusammenhang zwischen dem Alter und der Platzierung beim Wettbewerb besteht!

Lösungen:

Überprüfung mit Hilfe des Rangkorrelationskoeffizienten nach Spearman
(Transformation des Alters in Rangplätze)

Person	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Rangplatz (X)	7	3	9	10	1	5	4	6	2	8
Alter	9	4	5	10	3	2	8	7	1	6
$\mathbf{d_i}$	-2	-1	4	0	-2	3	-4	-1	1	2
$\mathbf{d_i^2}$	4	1	16	0	4	9	16	1	1	4

$$r_s = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n * (n^2 - 1)} = 1 - \frac{6 * 56}{10 * 99} = 0{,}661$$

Es besteht ein starker positiver Zusammenhang zwischen den beiden Variablen. Je älter die Person, desto schlechter die Platzierung im Wettbewerb (und umgekehrt)

Aufgabe 4)

Bestimmen Sie welches Skalenniveau bei den folgenden Variablen vorliegt!
Begründen Sie gegebenenfalls kurz! (10)

a) Filmgenre (nominal)

b) Inflationsrate (rational)

c) Rang beim Militär (ordinal)

d) Matrikelnummer (nominal)

e) Benzinverbrauch (rational)

f) Geburtsjahr (intervall)

g) Alkoholgehalt von Getränken (rational)

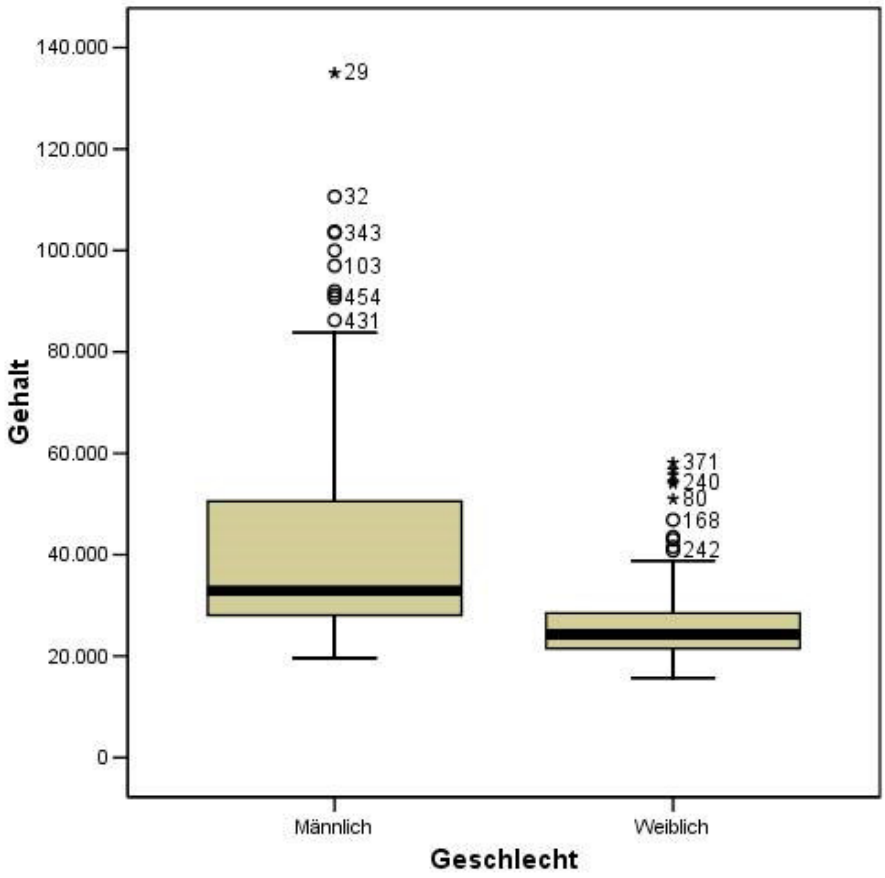
h) Klassifikation von Hotels nach Sternen (ordinal)

i) Preise für Lebensmittel (rational)

j) Konfession (nominal)

Aufgabe 5)

Bitte interpretieren Sie die folgende Grafik!



Lösungen:

In der Grafik ist für die Abbildung der weiblichen und der männlichen (jährlichen) Einkommensverteilung (Gehalt) jeweils ein Box-Plot zu sehen.
Verdeutlicht wird dadurch ein starker Unterschied der beiden jeweils rechtsschiefen Verteilungen. Das Einkommen der Männer ist im Vergleich mit dem der Frauen nicht nur deutlich höher, sondern auch deutlich heterogener.
Genauer: Bei den Männer liegt der Mindestwert bei 20000€, der Median bei ca. 33000 und der Maximalwert bei ca. 135000€ Bei den Frauen sind die Werte jeweils geringer ca. (17000, 23000, 60000). Die mittleren 50% der Männer (Quartilabstand) verdienen zwischen 28000 und 50000 Euro, mehrere Ausreißer sind ab ca. 82000 Euro abgebildet. Bei den Frauen verdienen die mittleren 50% zwischen 22000 und 28000 Euro (die Box ist somit viel schmaler) und die Werte der Ausreißer beginnen bei 38000 Euro.

Aufgabe 6)

Im Rahmen einer kleinen Umfrage sollten die Teilnehmer unter anderem angeben, ob Sie schon einmal absichtlich „schwarz“ gefahren sind. Im Folgenden sind die Ergebnisse aufgegliedert nach Geschlecht dargestellt:

	mind. einmal „schwarz“ gefahren		
	ja	nein	Gesamt
Geschlecht weiblich	8	25	33
männlich	17	12	29
Gesamt	25	37	62

a)

Berechnen Sie ein geeignetes Maß, um die Stärke des Zusammenhangs angeben zu können!

b)

Geben Sie darüber hinaus mit einem geeigneten Kennwert den Effekt des Geschlechts auf das Schwarzfahren an!

Lösungen:

a)

$$\phi = \frac{ad - bc}{\sqrt{(a+b) * (c+d) * (a+c) * (b+d)}} = \frac{8 * 12 - 25 * 17}{\sqrt{33 * 29 * 25 * 37}} = -0{,}35$$

Es besteht ein mittelstarker Zusammenhang zwischen den beiden Variablen (das Vorzeichen sollte nicht interpretiert werden).

b)
Berechnung der Odds oder der Prozentsatzdifferenz (hier Odds)

$$OR = \frac{n_{11}}{n_{12}} \,/\, \frac{n_{21}}{n_{22}} = \frac{8}{25} / \frac{17}{12} = 0{,}226$$

Laut Umfrage fahren Frauen ca. nur ein Viertel so häufig „schwarz“ im Vergleich zu den Männern.

Aufgabe 7)

Ein junger Forscher entschließt sich bei seiner Datenerhebung nicht direkt nach dem Einkommen der Probanden zu fragen, sondern lässt diese sich jeweils in Einkommensgruppen einordnen. Auf diese Weise erhofft er sich die Ausfallrate möglichst gering zu halten. Er erhält folgendes Ergebnis:

Alter	Freq.	Percent	Valid Percent	Cum. Percent
0- 499 Euro	145			
500- 999 Euro	308			
1000-1499 Euro	490			
1500-1999 Euro	616			
2000-2499 Euro	413			
2500-2999 Euro	237			
3000-3999 Euro	168			
4000-4999 Euro	41			
5000-8000 Euro	30			
keine Angabe	212			
Total	2,660			

a) Bitte vervollständigen Sie die Häufigkeitstabelle der gruppierten Einkommensangaben!

b) Bitte zeichnen Sie anschließend die empirische Verteilungsfunktion!

Lösungen:

a)

Alter	Freq.	Percent	Valid Percent	Cum. Percent
0- 499 Euro	145	5,45	5,92	5,92
500- 999 Euro	308	11,58	12,58	18,50
1000-1499 Euro	490	18,42	20,02	38,52
1500-1999 Euro	616	23,16	25,16	63,68
2000-2499 Euro	413	15,53	16,87	80,55
2500-2999 Euro	237	8,91	9,68	90,23
3000-3999 Euro	168	6,32	6,86	97,09
4000-4999 Euro	41	1,54	1,68	98,77
5000-8000 Euro	30	1,13	1,23	100,00
keine Angabe	212	7,97		
Total	2,660	2,448		

Aufgabe 8)

Im Folgenden sind die beiden Variablen „Gesundheitszustand“ und „subjektive Schichteinstufung“ kreuztabelliert. Überprüfen Sie mit einem geeigneten Maß, ob die Schichteinstufung einen Effekt auf den Gesundheitszustand hat!

Gesundheits-zustand	**Subjektive Schichteinstufung**			**Total**
	Unterschicht	**Mittelschicht**	**Oberschicht**	
eher schlecht	134	91	23	248
teils/teils	167	311	157	635
eher gut	78	183	139	400
Total	379	585	319	1283

a) Welches Maß ist geeignet, um überprüfen zu können, ob die Schichteinstufung einen Effekt auf den Gesundheitszustand hat?
b) Berechnen Sie dieses Maß und interpretieren Sie ihr Ergebnis!

Lösung:

a)
Asymmetrisches Maß für zwei ordinalskalierte Variablen → Somers d

b)

$$d_{YX} = \frac{C - D}{C + D + T_Y}$$

$C = 229799$

$D = 104527$

$T_Y = 194905$

$$d_{YX} = \frac{229799 - 104527}{229799 + 104527 + 194905} = 0{,}237$$

c)
Es besteht ein mäßiger/mittelstarker positiver Effekt. Je höher die Schichteinstufung, desto besser der Gesundheitszustand!

Aufgabe 9)

516 Leser einer Zeitschrift wurden unter anderem befragt für welchen Themenbereich sie sich am meisten interessieren.
Folgende Verteilung der Antworten konnte ermittelt werden:

Politik	204
Wirtschaft	46
Buntes	31
Kultur	28
Sport	188
Wetter	19

Bestimmen Sie die Streuung mit einem geeigneten Kennwert und interpretieren Sie diese!

Lösung:

Themenbereich	**Absolute Häufigkeit**	**relative Häufigkeit**	**quadrierte rel. H.**
Politik	204	0,395	0,156
Wirtschaft	46	0,089	0,008
Buntes	31	0,060	0,003
Kultur	28	0,054	0,003
Sport	188	0,364	0,133
Wetter	19	0,037	0,001
Gesamt	**516**	**1**	**0,304**

$$P = \frac{m}{m-1} * (1 - \sum_{i=1}^{m} p_i^2) = \frac{6}{5} * (1 - 0{,}304) = 0{,}835$$

Bei dieser Verteilung liegt eine mittlere Streuung vor. Es besteht weder Ähnlichkeit zu einer Einpunkt-Verteilung, noch zu einer Gleichverteilung.

Aufgabe 10)

Im Rahmen der sogenannten Drittvariablenkontrolle können u.a. eine Scheinkausalität oder ein Moderatoreffekt aufgedeckt werden.
Worin besteht der Unterschied?

Lösung:

Wenn ein im Rahmen einer bivariaten Analyse bestehender Zusammenhang zwischen den Variablen X und Y durch die Hinzunahme einer dritten Variablen nicht mehr gemessen werden kann, dann haben X und Y kausal nichts miteinander zu tun. Die statistische Korrelation wird nur deshalb hervorgerufen, weil sowohl X als auch Y von der dritten Variablen Z beeinflusst werden.

Im Unterschied zu dieser Scheinkausalität verschwindet bei einem Moderatoreffekt der bivariate Zusammenhang zwischen X und Y nicht durch die Hinzunahme einer dritten Variable Z in die Analyse. Vielmehr verändert sich die statistische Korrelation zwischen X und Y in ihrer Stärke, je nachdem welche Ausprägung Z annimmt. Die Variable Z moderiert demnach die Beziehung zwischen X und Y.

Aufgabe 11)

Welche der folgenden Aussagen sind zutreffend. Kreuzen Sie an.

	stimmt	stimmt nicht
1. Der Modalwert ist der Wert einer Verteilung, der die größte Auftretenswahrscheinlichkeit hat.	x	
2. Die Standardabweichung einer Verteilung ist immer größer als die Varianz.		x
3. Der Mittelwert $\bar{x}$ ist formal definiert als die an der Anzahl der Messwerte relativierte Summe der Einzelwerte.	x	
4. Bestehen die Ausprägungen einer Variablen aus Rangplätzen (bspw. Chartplatzierungen), so kann von metrischem Skalenniveau ausgegangen werden.		x
5. Die Bestimmung des arithmetischen Mittels aus einer nominalskalierten Variablen ist inhaltlich sinnvoll.		x
6. Wird in einer Untersuchung die soziale Schicht der Versuchsperson erfasst, so kann es sich nur um eine nominalskalierte Größe handeln.		x
7. Die Varianz ist ein Dispersionsmaß, das bei intervallskalierten Messwerten bestimmt werden kann.	x	
8. Die Varianz ist ein Maß für den Unterschied zwischen den Extremwerten einer Verteilung.		x
9. Eine Korrelation zeigt immer Ursache und Stärke des Zusammenhangs zweier Variablen an.		x
10. Bei einer exakt symmetrischen Verteilung reicht die Berechnung des arithmetischen Mittels, da keine neuen Erkenntnisse durch Modus und Median zu erwarten sind.	x	
11. Es kann nur einen Modus geben.		x
12. Die Regressionskonstante einer Regressionsgleichung kann nicht negativ sein.		x
13. Bei metrischen Skalen ist das arithmetische Mittel das geeignetste Maß der zentralen Tendenz.	x	
14. Jeder Regressionskoeffizient kann genutzt werden um die Stärke des Zusammenhangs zu interpretieren.		x
15. Eine andere Bezeichnung für die Varianz ist Streuung.	x	

16. Kumulierte Anteilswerte in einer Häufigkeitstabelle sind bei jedem Skalenniveau sinnvoll interpretierbar.		x
17. Cramers V hat einen Wertebereich von -1 bis 1.		x
18. Bei einer Ursache-Wirkungs-Beziehung muss auch immer eine Korrelation vorliegen.	x	
19. Beim Vorliegen einer Vierfeldertafel erhält man für die Berechnung von Phi und Cramers V das gleiche Ergebnis.	x	
20. Ein gruppiertes Balkendiagramm ist eine geeignete Methode für die grafische Darstellung von Beziehungen zwischen metrisch skalierten Variablen.		x
21. Ist das arithmetische Mittel einer Verteilung größer als der Median und der Modalwert, und der Median größer als der Modalwert, kann von einer rechtsschiefen Verteilung ausgegangen werden.	x	
22. Gamma, tau-b und der Kontingenzkoeffizient C sind Koeffizienten für ordinal skalierte Variablen.		x

Aufgabe 12)

Die deutschen Fernbuslinien kamen im Jahr 2014 auf 20,4 Millionen Fahrgäste, die sich wie folgt aufgliederte:

Mein Fernbus : 8,63 Mio.
Flixbus : 5,92 Mio.
Postbus : 2,61 Mio.
Berlinlinienbus : 0,89 Mio.
Eurolines : 0,48 Mio.
DeinBus : 0,27 Mio.
IC Bus (DB) : 0,24 Mio.
Sonstige : 1,36 Mio.

Bestimmen Sie die Konzentrationsrate für die beiden Marktführer!

Lösung:

$$C_r = \frac{\sum_{i=1}^{r} x_i}{\sum_{i=1}^{n} x_i} = C_2 = \frac{14{,}55}{20{,}4} = 0{,}713 \cong 71{,}3\%$$

Die beiden Marktführer vereinen über 70% der Fahrgäste unter sich.

Aufgabe 13)

In einer kleinen Studie sollten Probanden in einer vorgegebenen Zeit verschiedene Aufgaben lösen. Anschließend wurde das Alter, das Geschlecht (0 = weiblich; 1 = männlich) und der Intelligenzquotient erhoben. Es soll mit Hilfe einer linearen Regression untersucht werden, inwiefern diese drei unabhängigen Variablen einen Effekt auf die Zahl der gelösten Aufgaben haben.
Nachfolgend steht der entsprechende SPSS-Ausdruck:

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	.738[a]	.545	.460	2.64901

a. Einflußvariablen : (Konstante), Geschlecht, Intelligenzquotient, Alter

ANOVA[a]

Modell		Quadratsumme	df	Mittel der Quadrate	F	Sig.
1	Regression	134.674	3	44.891	6.397	.005[b]
	Nicht standardisierte Residuen	112.276	16	7.017		
	Gesamt	246.950	19			

a. Abhängige Variable: Gelöste Aufgaben

b. Einflußvariablen : (Konstante), Geschlecht, Intelligenzquotient, Alter

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Sig.
		RegressionskoeffizientB	Standardfehler	Beta		
1	(Konstante)	-20.961	7.111		-2.948	.009
	Intelligenzquotient	.111	.097	.286	1.135	.273
	Alter	.775	.404	.489	1.919	.073
	Geschlecht	-1.653	1.216	-.235	-1.359	.193

a. Abhängige Variable: Gelöste Aufgaben

a)
Interpretieren Sie die Erklärungskraft des Modells!

b)
Interpretieren Sie die Regressionskoeffizienten und die Konstante!

c)
Welche Variable hat den stärksten Einfluss?

Lösung:

a)
Das R-Quadrat verdeutlicht die (recht gute) Erklärungskraft des Modells. Der Wert von 0,545 besagt, dass 54,5% der Streuung der abhängigen Variablen „Gelöste Aufgaben“ durch die Streuung der unabhängigen Variablen erklärt wird.

b)
Der Regressionskoeffizient für den Intelligenzquotienten beträgt 0,111. Pro IQ-Punkt mehr, werden im Durchschnitt 0,111 Aufgaben mehr gelöst.

Der Regressionskoeffizient für das Alter beträgt 0,775. Pro Lebensjahr mehr, werden im Durchschnitt 0,775 Aufgaben mehr gelöst.

Der Regressionskoeffizient für das Geschlecht beträgt -1,653. Mit Blick auf die Kodierung bedeutet dies, dass Männer im Durchschnitt 1,653 Aufgaben weniger lösen als Frauen.

c)
Durch die standardisierten Koeffizienten wird deutlich, dass das Alter mit 0,489 einen mittelstarken/starken Einfluss hat und innerhalb des Modells die relevanteste Variable ist.