TECHNISCHE
UNIVERSITÄT
DRESDEN

Fakultät Verkehrswissenschaften „Friedrich List“ Professur für Verkehrsökonometrie und -statistik

Statistik II für Verkehrswirtschaftler

Wintersemester 2011/12

Lösungen zu den Trainingsaufgaben

Inhaltsverzeichnis

1 Zeitreihenanalyse

Lösung T.1.1

(a) **Glatte Komponente:** Lässt sich mit dem symmetrischen gleitenden Mittel bestimmen. Die Periodendauer beträgt 1 Jahr bei Angabe in Quartalen ⇒ Periodenlänge $m = 4$. Damit gilt für die glatte Komponente:

$$G_{13} = \frac{1}{4}\left(\frac{x_1}{2} + x_2 + x_3 + x_4 + \frac{x_5}{2}\right) = 160, \tag{1}$$
$$G_{14} = \frac{1}{4}\left(\frac{x_2}{2} + x_3 + x_4 + x_5 + \frac{x_6}{2}\right) = 167.5, \tag{2}$$
$$\dots \qquad \dots \tag{3}$$
$$G_{32} = \frac{1}{4}\left(\frac{x_8}{2} + x_9 + x_{10} + x_{11} + \frac{x_{12}}{2}\right) = 210.625. \tag{4}$$

oder lässt sich einfacher mit einer Arbeitstabelle berechnen:

	Jahr	Quartal	x_{ij}	$\overline{G}_{ij}^{(4)}$
x_{11}	2007	Q_1	240	-
x_{12}	2007	Q_2	110	-
x_{13}	2007	Q_3	170	$(\frac{240}{2} + 110 + 170 + 95 + \frac{290}{2}) \cdot \frac{1}{4} = 160$
x_{14}	2007	Q_4	95	$(\frac{110}{2} + 170 + 95 + 290 + \frac{120}{2}) \cdot \frac{1}{4} = 167.5$
x_{21}	2008	Q_1	290	168.8
x_{22}	2008	Q_2	120	172.5
x_{23}	2008	Q_3	170	180.6
x_{24}	2008	Q_4	125	188.8
x_{31}	2009	Q_1	325	201.3
x_{32}	2009	Q_2	150	210.6
x_{33}	2009	Q_3	240	-
x_{34}	2009	Q_4	130	-

also insgesamt

	Q1	Q2	Q3	Q4
2007	-	-	160.0	167.5
2008	168.8	172.5	180.6	188.8
2009	201.3	210.6	-	-

(bis auf den ersten Wert ergeben sich durch Rundung die Werte des Aufgabenblattes) Die ersten und letzten beiden Werte der glatten Komponente lassen sich mit symmetrischen gleitenden Mitteln nicht berechnen!

(b) **Saisonkomponente:** Zunächst werden die vorläufigen Saisonanteile durch die Differenz der arithmetischen Mittel der Spalten und der Spaltenmittel der glatten Komponenten berechnet:

$$\tilde{S}_1 = \frac{x_{21} + x_{31} - G_{21} - G_{31}}{2} = \frac{290 + 325 - 168,8 - 201,3}{2} = 122.5, \quad (5)$$

$$\tilde{S}_2 = \frac{x_{22} + x_{32} - G_{22} - G_{32}}{2} = -57, \quad (6)$$

$$\tilde{S}_3 = \frac{x_{13} + x_{23} - G_{13} - G_{23}}{2} = \frac{170 + 170 - 160 - 180,6}{2} = -3, \quad (7)$$

$$\tilde{S}_4 = \frac{x_{14} + x_{24} - G_{14} - G_{24}}{2} = -68.15. \quad (8)$$

Es gilt $s_j = \tilde{S}_j - c$. Der Korrekturterm beträgt

$$c = \frac{1}{4}\sum_{j=1}^{4} \tilde{S}_j = -1.41, \quad (9)$$

und damit

$$S_1 = 122.5 - (-1.41) = 123.6, \quad S_2 = -55.9, \quad S_3 = -1.6, \quad S_4 = -69.6. \quad (10)$$

(c) **Zufallskomponmente** für das Jahr 2008: Allgemein gilt (in der Nomenklatur mit doppeltem Index) $U_{ij} = x_{ij} - G_{ij} - S_j$ also

$$U_{21} = 290 - 168.8 - 122.5 = -1.3, \quad U_{22} = 3.4, \quad U_{23} = -9, \quad U_{24} = 5.8. \quad (11)$$

(d) **Prognose der glatten Komponente für 2010:** Die letzte verfügbare glatte Komponente G_{32} ist für das 2. Quartal 2009. Für 2010 muss man also maximal sechs Quartale (Q_3 und Q_4 im Jahr 2009 sowie $Q_1 - Q_4$ im Jahr 2010) prognostizieren:

1. Quartal 2010: $G_{41} = G_{32} + 3(G_{32} - G_{31}) = 210.6 + 3 \cdot (210.6 - 201.3) = 238.5,$ (12)
2. Quartal 2010: $G_{42} = G_{32} + 4(G_{32} - G_{31}) = 247.8,$ (13)
3. Quartal 2010: $G_{43} = G_{32} + 5(G_{32} - G_{31}) = 257.1,$ (14)
4. Quartal 2010: $G_{44} = G_{32} + 6(G_{32} - G_{31}) = 266.4.$ (15)

(e) **Prognose der Übernachtungszahlen für 2010:** Hier addiert man alle regelmäßigen Komponenten, also G und S. Damit

1. Quartal 2010: $x_{41} = G_{41} + S_1 = 238.5 + 123.6 = 362.1,$ (16)
2. Quartal 2010: $x_{42} = G_{42} + S_2 = 191.9,$ (17)
3. Quartal 2010: $x_{43} = G_{43} + S_3 = 255.5,$ (18)
4. Quartal 2010: $x_{44} = G_{44} + S_4 = 196.8.$ (19)

(f) **Fehler bei direkter unbedingter Prognose:** Prognostiziert man die Übernachtungszahlen direkt durch die Fortschreibung der Zahlen selbst, erhält man starke Fehler dadurch, das der meiste Anteil des "Trends" in Wirklichkeit vom Auf- oder Abstieg des Saisomanteils kommt! Hier gilt[1]

$$1.\ \text{Quartal 2010: } h_{13} = 130 - (240 - 130) = 130 - 110 = 20, \tag{20}$$

$$2.\ \text{Quartal 2010: } h_{14} = 130 - 2 \cdot 110 = -90, \tag{21}$$

$$3.\ \text{Quartal 2010: } h_{15} = 130 - 3 \cdot 110 = -200, \tag{22}$$

$$4.\ \text{Quartal 2010: } h_{16} = 130 - 4 \cdot 110 = -310. \tag{23}$$

[1] Es gilt folgende Regel bei der Bedienung von Statistikprogrammen: “Junk in - Junk out”.

Lösung T.1.2

(a) Bei den Fortzugszahlen ist ein Schrumpfungsprozess festzustellen. Logarithmen von Wachstums- und Schrumpfungsprozesse sind i.d.R. additiv, also ist hier ein multiplikatives Modell geeignet, da dessen Logarithmen additiv sind.

(b) Seien x_{ij} die Fortzugszahlen im Quartal j des Jahres $2000 + i$

- Logarithmieren:

$$y_{ij} = \ln(x_{ij}), \quad y_{11} = 7.3778, \ \ldots, \ y_{44} = 7.272, \quad \bar{y} = 7.326. \tag{24}$$

- Aus der Formelsammlung Formel für additive Zerlegung mit $\tau = 4$ Quartalen:

$$\ln(\tilde{S}_j) = \frac{1}{n}\sum_i (y_{ij} - \bar{y}_{ij}^{(\tau)}) \tag{25}$$

  Die Summation geschieht über alle Jahre, zu denen für das jeweilige Quartal das arithmetische gleitende Mittel $\bar{y}_{ij}^{(\tau)}$ errechnet werden kann. Für die Quartale $j = 1$ und 2 sind dies die Jahresindices $i = 2, 3, 4$ und für die beiden anderen Quartale $i = 1, 2, 3$. Die Zahl der berücksichtigten Jahre ist also jeweils $n = 3$.

  Gleitende Mittel der Ordnung $\tau = 4$:

$$\bar{y}_{13}^{(4)} = \frac{y_{11}}{8} + \frac{y_{12} + y_{13} + y_{14}}{4} + \frac{y_{21}}{8} = 7.524, \quad \bar{y}_{14}^{(4)} = 7.504, \ldots, \bar{y}_{42}^{(4)} = 7.179. \tag{26}$$

  bzw. mit der Arbeitstabelle:

	x_{ij}	$y_{ij} = ln x_{ij}$	$\bar{y}_{ij}^{(4)}$
y_{11}	1600	7.378	-
y_{12}	1580	7.365	-
y_{13}	2420	7.792	$7.524 = \frac{7.378}{8} + \frac{7.365+7.792+7.591}{4} + \frac{7.320}{8}$
y_{14}	1980	7.591	7.504
y_{21}	1510	7.320	7.455
y_{22}	1420	7.258	7.386
y_{23}	1820	7.507	7.341
y_{24}	1520	7.326	7.307
y_{31}	1370	7.223	7.273
y_{32}	1200	7.090	7.249
y_{33}	1640	7.402	7.226
y_{34}	1390	7.237	7.198
y_{41}	1240	7.123	7.178
y_{42}	1060	6.966	7.179
y_{43}	1590	7.371	-
y_{44}	1440	7.272	-

  und damit

$$\ln(\tilde{S}_1) = \frac{7.320 - 7.455 + 7.223 - 7.273 + 7.123 - 7.178}{3} = -0.0803, \tag{27}$$

$$\ln(\tilde{S}_2) = -0.1664, \quad \ln(\tilde{S}_3) = 0.2034, \quad \ln(\tilde{S}_4) = 0.0486 \tag{28}$$

  und mit dem Korrekturfaktor $c = \frac{1}{4}\sum_{j=1}^{4} \ln(\tilde{S}_j) = \frac{-0.0803-0.1664+0.2034+0.0486}{4} = 0.0013$

und

$$\ln S_j = (\tilde{S_j}) - c \tag{29}$$

$$\ln(S_1) = -0.0803 - 0.0013 = -0.0816, \tag{30}$$
$$\ln(S_2) = -0.1678, \tag{31}$$
$$\ln(S_3) = 0.2021, \tag{32}$$
$$\ln(S_4) = 0.0473. \tag{33}$$

- Damit die Saisonindexziffern $S_j = e^{\ln S_j}$ der multiplikativen Zerlegung $X_{ij} = T_{ij}K_{ij}S_jU_{ij}$:

$$S_1 = 0.9216, \tag{34}$$
$$S_2 = 0.8456, \tag{35}$$
$$S_3 = 1.2240, \tag{36}$$
$$S_4 = 1.0484. \tag{37}$$

Es kann z.B. durch lineare Fortschreibung der Jahresmittel der logarithmierten Werte als Basis für die Saison-Multiplikatoren eine Prognose erstellt werden. (Eine konstante Fortschreibung, lineare Regression, lineare Fortschreibung der bei der Berechnung der $\ln(\tilde{S_1})$ benötigten gleitenden arithmetischen Mittel etc. ergeben alle volle Punktzahl, da die Methode für die Prognose in der Aufgabenstellung nichts spezifiziert wurde).

$$\text{Jahr 2003:} \quad \bar{y}_3 = 7.238, \quad \text{Jahr 2004:} \quad \bar{y}_4 = 7.183. \tag{38}$$

Für das Jahr 2005 erhält man durch lineare Fortschreibung

$$\bar{y}_5 \approx 2\bar{y}_4 - \bar{y}_3 = 7.128 \tag{39}$$

Hierauf wird die Saison-Charakteristik addiert, um die Prognose

$$y_{5j} = \bar{y}_5 + \ln(S_j) \tag{40}$$

zu erhalten mit den Werten

$$y_{51} = 7.128 + (-0.0816) = 7.04 \tag{41}$$
$$y_{52} = 6.96, \quad y_{53} = 7.33, \quad y_{54} = 7.18 \tag{42}$$

und für die Wegzugsraten selbst

$$x_{51} = e^{7.04} = 1141, \quad x_{52} = 1054, \tag{43}$$
$$x_{53} = 1525, \quad x_{54} = 1313. \tag{44}$$

Lösung T.1.3

(a) Dazu wertet man $S_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} - \bar{y}$ mit einer Arbeitstabelle aus:

	Q1	Q2	Q3	Q4	$\sum$
1999	200	260	210	150	
2000	220	270	220	140	
2001	210	280	200	160	
$\sum$	630	810	630	450	$2520 = 12\bar{y}$ $(\rightarrow \bar{y} = 210)$
$S_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij} - \bar{y}$	0	60	0	-60	0

In der letzten Summe hat man auch gleich das Verschwinden des Mittlewertes der Saisonfigur überprüft.

(b) Sie steigt. Da man bei stationären Zeitreihen keinen Trend hat, muss man einfach nur den Anstieg bzw. Abfall gegenüber dem selben Quartal des Vorjahres betrachten. Da nach Voraussetzung die Konjunkturkomponente =0 ist, geht der Anstieg auf die Restkomponente U_{ij} zurück.

Lösung T.1.4

Lösungsvorschlag folgt.

Lösung T.1.5

(a) Gleitendes Mittel der Ordnung $\tau = 2$. Mit der Formel

$$\bar{y}_t^{(\tau=2)} = \frac{1}{2}\left(\frac{1}{2}y_{t-1} + y_t + \frac{1}{2}y_{t+1}\right) \tag{45}$$
$$= \frac{1}{4}y_{t-1} + \frac{1}{2}y_t + \frac{1}{4}y_{t+1} \tag{46}$$

ergeben sich

t	y_t	$g_t := y_t^{(\tau=2)}$
W07	261	-
S07	142	199
W08	251	210
S08	196	236
W09	301	244
S09	178	237
W10	291	245
S10	220	-

(b) Phasendurchschnittsverfahren:

1. Ermittlung der Phasendurchschnitte

$$\tilde{S}_j = \frac{1}{p}\sum_{i=1}^{p}(y_{ij} - g_{ij}) \tag{47}$$

wobei $p = 3$ vollständige Perioden zur Verfügung stehen. Es ergeben sich

$$\tilde{S}_{\mathrm{W}} = \frac{1}{3}\big[(251 - 210) + (301 - 244) + (291 - 245)\big] \tag{48}$$
$$= \frac{1}{3}\big[41 + 57 + 46\big] \tag{49}$$
$$= 48 \tag{50}$$

und

$$\tilde{S}_{\mathrm{S}} = \frac{1}{3}\big[(142 - 199) + (196 - 236) + (178 - 237)\big] \tag{51}$$
$$= \frac{1}{3}\big[(-57) + (-40) + (-59)\big] \tag{52}$$
$$= -52 \tag{53}$$

2. Eliminieren des Mittelwertes

$$S_j = \tilde{S}_j - \frac{1}{\tau}\sum_{j'=1}^{\tau} \tilde{S}_{j'} \tag{54}$$
$$= \tilde{S}_j - \frac{1}{2}\left(\tilde{S}_\mathrm{W} + \tilde{S}_\mathrm{S}\right) \tag{55}$$
$$= \tilde{S}_j - \frac{1}{2}\left(48 + (-52)\right) \tag{56}$$
$$= \tilde{S}_j + 2 \tag{57}$$
$$\tag{58}$$

mit $\tau = 2$ Saisonen. Es ergeben sich die Saisonindexziffern

$$S_\mathrm{W} = \tilde{S}_\mathrm{W} + 2 \tag{59}$$
$$= 48 + 2 \tag{60}$$
$$= 50 \tag{61}$$

und

$$S_\mathrm{S} = \tilde{S}_\mathrm{S} + 2 \tag{62}$$
$$= -52 + 2 \tag{63}$$
$$= -50 \tag{64}$$

(c) Im Winter ist mit $S_\mathrm{W} = 50$ (Tausend) mehr Touristen zu rechnen als im **Jahresdurchschnitt**.

(d) Die letzte verfügbare Anstiegsrate ist

$$r = g_{W10} - g_{S09} \tag{65}$$
$$= 245 - 237 \tag{66}$$
$$= 8 \tag{67}$$

womit sich mit der Formel

$$\hat{g}_{t+1} = g_t + r \tag{68}$$

iterativ folgende Werte berechnen lassen:

t	g_t	$\hat{g}_t$
W07		
S07	199	
W08	210	
S08	236	
W09	244	
S09	237	
W10	245	
S10		253
W11		261
S11		269

(e) Mit dem Ansatz der additiven Verknüpfung berechnet sich ein Prognosewert mit der Formel

$$\hat{y}_{ij} = g_{ij} + S_j \tag{69}$$

womit sich

$$\begin{aligned} \hat{y}_{\mathrm{W}11} &= g_{\mathrm{W}11} + S_{\mathrm{W}} && (70) \\ &= 261 + 50 && (71) \\ &= 311 && (72) \end{aligned}$$

und

$$\begin{aligned} \hat{y}_{\mathrm{S}11} &= g_{\mathrm{S}11} + S_{\mathrm{S}} && (73) \\ &= 269 - 50 && (74) \\ &= 219 && (75) \end{aligned}$$

ergeben.

(f) Bestimmung der Restkomponente

$$u_{ij} = y_{ij} - g_{ij} - S_j \tag{76}$$

liefert

$$\begin{aligned} u_{\mathrm{W}10} &= y_{\mathrm{W}10} - g_{\mathrm{W}10} - S_{\mathrm{W}} && (77) \\ &= 291 - 245 - 50 && (78) \\ &= -4 && (79) \end{aligned}$$

und

$$\begin{aligned} u_{\mathrm{S}10} &= y_{\mathrm{S}10} - g_{\mathrm{S}10} - S_{\mathrm{S}} && (80) \\ &= 220 - 253 - (-50) && (81) \\ &= 17 && (82) \end{aligned}$$

Lösung T.1.6

(a) Das gleitende Mittel der Ordnung $\tau = 4$ lässt sich mit

$$t = 4(i-1) + j$$

wie folgt berechnen.

$$\bar{y}_t^{(4)} = \frac{y_{t-2}/2 + (y_{t-1} + y_t + y_{t+1}) + y_{t+2}/2}{4} \tag{83}$$

Beispielsweise ist

$$\bar{y}_{13}^{(4)} = \frac{9/2 + 65 + 89 + 25 + 9/2}{4} = \mathbf{47} \tag{84}$$

und damit die restlichen Werte

$\bar{y}_{ij}^{\tau}$	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	-	-	**47**	44
$i=2$	43	45	47	50
$i=3$	50	50	-	-

Die glatte Komponente für das erste und zweite Quartal des ersten Jahres und das dritte und vierte Quartal des letzten Jahres können nicht berechnet werden.

(b) Zuerst müssen die Differenzen zwischen der Anzahl der Fahrraddiebstähle y_{ij} und der glatten Komponente $\bar{y}_{ij}^{4}$ gebildet werden. Beispielsweise ergibt sich für

$$y_{13} - \bar{y}_{13}^{(4)} = 89 - 47 = \mathbf{42} \tag{85}$$

und für die restlichen Differenzen

$y_{ij} - \bar{y}_{ij}^{(\tau)}$	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	-	-	**42**	-19
$i=2$	-34	-4	58	-25
$i=3$	-25	-1	-	-

Die Mittelwerte der Differenzen ergeben die Phasendurchschnitte $\tilde{S}_j$ für jedes Quartal j. So ist

$$\tilde{S}_1 = \frac{(-34) + (-25)}{2} = \mathbf{-29.5} \tag{86}$$

und für alle Quartale

$\tilde{S}_j$	$j=1$	$j=2$	$j=3$	$j=4$
	-29.5	-2.5	50	-22

Die Phasendurchschnitte wiederum haben den Mittelwert

$$\frac{1}{4}\sum_{j=1}^{4} \tilde{S}_j = \frac{1}{4}(-29.5 - 2.5 + 50 - 22) \tag{87}$$

$$= -1 \tag{88}$$

Dieser muss nur noch abgezogen werden. Man erhält für die Saisonfigur S_j

S_j	$j=1$	$j=2$	$j=3$	$j=4$
	-28.5	-1.5	51	-21

(c) Im dritten Quartal gibt es durchschnittlich 51 Fahrraddiebstähle mehr als im Jahresmittel.

(d) Mit dem Ansatz

$$\hat{y}_{ij} = 50 + S_j \tag{89}$$

erhält man die Prognosewerte

$\hat{y}_{4j}$	$j=1$	$j=2$	$j=3$	$j=4$
	21.5	48.5	101	29

2 Wahrscheinlichkeitsrechnung

Lösung T.2.1

(a) Zwei Bedingungen sind nötig, um mit einem Einsatz von x Cents den Gewinn abzufsassen:

- Kein weiterer Bieter bietet x Cents, es gilt also $\overline{A_x}$
- x Cents sind der geringste einzelstehende Betrag, es gilt also A_n für alle natürlichzahligen Werte $n < x$.

Da nach Annahme alle Einsätze unabhängig sind, gilt das Multiplikationsgesetz der Wahrscheinlichkeiten:

$$P(G_x) = P\left(A_1 \cap A_2 \cap \ldots \cap A_{x-1} \cap \overline{A_x}\right) = (1 - P(A_x)) \prod_{n=1}^{x-1} P(A_n) \tag{90}$$

(b) Einfach durch Einsetzen von $x = 100\,000$ Cents:

$$P(A_{100\,000}) = e^{-100\,000\,\lambda} = e^{-1} = 0.368. \tag{91}$$

(c) Durch Logarithmieren werden Produkte zu Summen:

$$\ln P(G_x) = \ln(1 - P(A_x)) + \sum_{n=1}^{x-1} \ln P(A_n) \tag{92}$$

und durch Einsetzen $P(A_n) = e^{-\lambda n}$ mit der angegebenen Näherungsformel.

$$\ln P(G_x) = \ln(1 - e^{-\lambda x}) - \sum_{n=1}^{x-1} \lambda n = \ln(1 - e^{-\lambda x}) - \frac{\lambda x^2}{2}. \tag{93}$$

(d) Ableiten, Null setzen durch Anwendung der Kettenregel für Exponentialfunktionen. Es gilt: $f(x) = e^{u(x)} \Rightarrow f'(x) = u'(x) \cdot e^{u(x)}$ und damit ist die Ableitung von $-e^{-\lambda x} = \lambda e^{-\lambda x}$, also

$$\frac{d \ln P(G_x)}{dx} = \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda x}} - \lambda x \stackrel{!}{=} 0 \tag{94}$$

$$\frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda x}} = \lambda x \tag{95}$$

$$\lambda e^{-\lambda x} = \lambda x (1 - e^{-\lambda x}) \tag{96}$$

$$\lambda e^{-\lambda x} = \lambda x - e^{-\lambda x} \lambda x \tag{97}$$

$$\lambda e^{-\lambda x} = \lambda (x - e^{-\lambda x} x) \quad | \div \lambda e^{-\lambda x} \tag{98}$$

$$1 = \frac{x}{e^{-\lambda x}} - \frac{x e^{-\lambda x}}{e^{-\lambda x}} \tag{99}$$

$$\Rightarrow 1 = x \left(e^{\lambda x} - 1\right). \tag{100}$$

Warum ist es zur Bestimmung des optimalen Gebots gleichgültig, ob man die Wahrscheinlichkeit oder deren Logarithmus maximiert?

Der Logarithmus ist eine streng monotone Funktion. Deshalb ist das Argument des Maximums der nichtlogarithmierten und der logarithmierten Funktion identisch.

(e) **Optimales Gebot:**

$$x_{opt} = 1/\sqrt{\lambda} = 316.22 \;\Rightarrow\; 316 \text{ Cent.} \tag{101}$$

Gewinnwahrscheinlichkeit: (Hier darf man natürlich nicht den Logarithmus nehmen!)

$$p_{opt} = e^{-\frac{\lambda x_{opt}^2}{s}} \left(1 - e^{-\lambda x_{opt}}\right) = 1.91 \text{ Promille.} \tag{102}$$

Lösung T.2.2

Wir kategorisieren die Einkommensklassen mit A_i und die Zahl der Kfz mit B_i:

- $A_i \Leftrightarrow$ Einkommensklasse i
- $B_1 \Leftrightarrow$ kein Kfz, $B_2 \Leftrightarrow$ ein Kfz, $B_3 \Leftrightarrow$ mindestens zwei Kfz.

Aus dem Text entnimmt man dann folgenden Angaben:

- Unbedingte Wahrscheinlichkeiten

$$P(A_1) = 0.2, \quad P(A_2) = 0.6, \quad P(A_3) = 0.2. \tag{103}$$

- Bedingte Wahrscheinlichkeiten

$$P(B_1|A_1) = 0.40 \quad P(B_2|A_1) = 0.55, \quad P(B_3|A_1) = 0.05, \tag{104}$$
$$P(B_1|A_2) = 0.20 \quad P(B_2|A_2) = 0.60, \quad P(B_3|A_2) = 0.20, \tag{105}$$
$$P(B_1|A_3) = 0.05 \quad P(B_2|A_3) = 0.45, \quad P(B_3|A_3) = 0.50. \tag{106}$$

Mit dem Satz der totalen Wahrscheinlichkeit zunächst die unbedingten Wahrscheinlichkeiten der Ereignisse B_i:

$$P(B_1) = P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2) + P(B_1|A_3)P(A_3) = 0.21, \tag{107}$$
$$P(B_1) = 0.4 \cdot 0.2 + 0.2 \cdot 0.6 + 0.05 \cdot 0.2 = 0.21 \tag{108}$$
$$P(B_2) = P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2) + P(B_2|A_3)P(A_3) = 0.56, \tag{109}$$
$$P(B_3) = P(B_3|A_1)P(A_1) + P(B_3|A_2)P(A_2) + P(B_3|A_3)P(A_3) = 0.23. \tag{110}$$

Nun mit dem Satz von Bayes die gesuchten bedingten Wahrscheinlichkeiten $P(A_i|B_j)$:

$$P(A_i|B_j) = \frac{P(B_j|A_i)P(A_i)}{P(B_j)} \tag{111}$$

Die Zahlenwerte eingesetzt. Beispielsweise:

$$P(A_1|B_1) = \frac{0.4 \cdot 0.2}{0.21} = 0.381 \tag{112}$$
$$P(A_3|B_2) = \frac{0.45 \cdot 0.2}{0.56} = 0.161 \tag{113}$$

also

$$P(A_1|B_1) = 0.381 \quad P(A_2|B_1) = 0.571, \quad P(A_3|B_1) = 0.048, \tag{114}$$
$$P(A_1|B_2) = 0.196 \quad P(A_2|B_2) = 0.643, \quad P(A_3|B_2) = 0.161, \tag{115}$$
$$P(A_1|B_3) = 0.043 \quad P(A_2|B_3) = 0.522, \quad P(A_3|B_3) = 0.435. \tag{116}$$

Das Ergebnis zeigt die erwartete Tendenz, dass Haushalten mit mehreren Kfz mit höherer Wahrscheinlichkeit eine obere Einkommensklasse besitzten als die mit keinem Fahrzeug (Betrachtung der vertikalen Ergebnisse pro Einkommensklasse).

Hinweis: Mit dem Wahrscheinlichkeitsbaum (3 Äste, je 3 Zweige) ist die Lösung ebenfalls im Standardverfahren erhältlich.

Lösung T.2.3

(a) Mit der Fehlerzahl n, der Entdeckwahrscheinlichkeiten p_A und p_B gilt anhand der Aufgabenstellung und bei unabhängig die Fehler entdeckenden Korrektoren A und B:

$$5 = n \cdot p_A, \quad p_A = \frac{5}{n}, \tag{117}$$

$$4 = n \cdot p_B, \quad p_B = \frac{4}{n}, \tag{118}$$

$$2 = n \cdot p_A \cdot p_B, \quad 2 = n \cdot \frac{5}{n} \cdot \frac{4}{n}, \tag{119}$$

$$\Rightarrow n = 10, \quad p_A = 0.5, \quad p_B = 0.4 \tag{120}$$

bzw.

$$n = \frac{2}{p_A p_B} = 10. \tag{121}$$

(b) Nullhypothese H_0: $n = 10$ Fehler. Unter H_0 sind somit die von Person A oder B entdeckten Fehlerzahlen binomialverteilt:

$$n_A \sim B(10, 0.5), \quad n_B \sim B(10, 0.4) \tag{122}$$

Für die Wahrscheinlichkeiten $p_A(x)$ und $p_B(x)$ dafür, dass die Personen A oder B jeweils x Fehler entdecken, gilt also

$$p_A(x) = \binom{10}{x} \left(\frac{1}{2}\right)^{10}, \quad p_B(x) = \binom{10}{x} \left(\frac{2}{5}\right)^{x} \left(\frac{3}{5}\right)^{10-x}. \tag{123}$$

Somit können die personen durchaus mehr oder weniger als 5 bzw. 4 Fehler entdecken und damit auch die schlussgefolgerte (geschätzte) Gesamtfehlerzahl von der tatsächlichen Gesamtzahl abweichen.

Lösung T.2.4

Sei

- $r_0 = 0.04$ der (auf 1 Jahr bezogene, also einschl. Zinseszins) Zinssatz der sicheren Bundesanleihe,
- $r = 0.06$ der (auf 1 Jahr bezogene) Zinssatz der Bankanleihe
- p die Ausfallwahrscheinlichkeit der Bank in diesem Jahr.

Wir nehmen an, dass die erwartete Rendite bei den sicheren Papieren und der Bankanleihe dieselbe sein soll (in Wirklichkeit erwartet der Investor allerdings oft eine höhere erwartete Rendite als Risikozuschlag). Bei der Bankanleihe ist bei einer Ausfallwahrscheinlichkeit von p der Erwartungswert des Auszahlungsbetrags X pro eingezahlter Geldeinheit gegeben durch

$$E(X) = (1-p)(1+r) + p \cdot 0 \tag{124}$$

(mit der Wahrsch. p wird gar nichts zurückbezahlt, mit der Komplementärwahrscheinlichkeit der Einsatz Plus Zinsen).

Durch Gleichsetzten mit dem Auszahlbetrag $1 + r_0$ der Bundesanleihe erhält man

$$(1-p)(1+r) = 1 + r_0 \;\Rightarrow\; p = 1 - \frac{1+r_0}{1+r} = 1.89\% \tag{125}$$

Lösung T.2.5

(a) Die die bedingte Wahrscheinlichkeit kann man sich aus den Häufigkeiten herleiten. Korrektor A findet 75 Fehler, also $n_A = 75$. Von diesen Fehlern werden $n_C = 30$ auch von Korrektor B gefunden. Von allen $n_A = 75$ Fehlern, die Korrektor A findet (Bedingung A), wurden $n_C = 30$ auch vom anderen gefunden (betrachtetes Ereignis C). Entsprechend sind

$$P(C|A) = \frac{n_C}{n_A} = \frac{30}{75} = 40\% \tag{126}$$

$$P(C|B) = \frac{n_C}{n_B} = \frac{30}{60} = 50\% \tag{127}$$

(b) Es gilt nach Produktregel

$$P(C) = P(A \cap B) \tag{128}$$

$$= P(A) \cdot P(B|A) \tag{129}$$

$$= P(A) \cdot P(B) \quad \text{(da unabhängig)} \tag{130}$$

woraus nach Umformung folgt

$$P(B) = \frac{P(C)}{P(A)} = \frac{n_C}{n_A} = \frac{30}{75} = 40\% \tag{131}$$

$$P(A) = \frac{P(C)}{P(B)} = \frac{n_C}{n_B} = \frac{1}{2} = 50\%. \tag{132}$$

(c) Prüfer B entdeckt den Anteil $P(B) = 30/75$ aller Fehler und dieser Anteil entspricht $n_B = 60$ Fehlern. Damit ist die erwartete Gesamtfehlerzahl gegeben durch den Dreisatz

$$P(B) = \frac{n_B}{n_F} \tag{133}$$

$$n_F = \frac{n_B}{P(B)} \tag{134}$$

$$= \frac{60}{30/75} \tag{135}$$

$$= 150. \tag{136}$$

Es wurden insgesamt 105 Fehler entdeckt: 75 Fehler von Korrektor A und $(60 - 30)$ zusätzliche Fehler von B (welche A nicht entdeckt hat). Damit sind im Mittel noch 45 Fehler unentdeckt.

(d) **Gaußverteilung:** Ja, da bei unabhängiger Fehlerentstehung und 150 Fehlern insgesamt die Bedingungen für den Zentalen Grenzwertsatz erfüllt sind. Bei sehr wenigen Fehlern (sehr gute Schreiber oder sehr kurze Arbeiten) nicht anwendbar.

Exponentialverteilung: Nein. Zwar beschreibt diese die Abstände zwischen zwei Fehlern, nicht jedoch die Gesamtanzahl derselben.

Binomialverteilung: Theoretisch ja, aber extrem unpraktisch. Da man bei jedem Wort Schreibfehler machen kann, ist die Gesamtfehlerzahl eine Summe (nach Voraussetzung unabhängiger) binärwertiger Variabler (Wort ist fehlerhaft oder nicht) und damit binomialverteilt. Da aber pro Wort die Fehlerwahrscheinlichkeit klein ist, ist der Grenzfall „seltenes Ereignis“ relevant und damit die ...

Poissonverteilung: Dies ist (neben der Gaußverteilung) die praktikabelste Verteilung.

Hypergeometrische Verteilung: Diese behandelt eine Situation des Ziehens aus einem endlichen Reservoir „ohne Zurücklegen“. Die Fehler wären also voneinander abhängig und diese Verteilung daher nicht zutreffend.

Gleichverteilung: nicht zutreffend.

(e) Die Fehlerzahl in der Zusammenfassung ist poissonverteilt

$$X \sim \mathrm{Po}(\mu). \tag{137}$$

mit dem Parameter

$$\begin{aligned} \mu &= n \cdot \theta && (138)\\ &= 1.5 \cdot 3 && (139)\\ &= 4.5 && (140) \end{aligned}$$

Wahrscheinlichkeit der Fehlerfreiheit ist gegeben durch

$$\begin{aligned} P(X=0) &= \frac{\mu^0 \cdot e^{-\mu}}{0!} && (141)\\ &= \frac{1 \cdot e^{-4.5}}{1} && (142)\\ &= 1.11\% && (143) \end{aligned}$$

Die Wahrscheinlichkeit für mehr als drei Fehler ist

$$\begin{aligned} P(X>3) &= 1 - \sum_{x=0}^{3} P(X=x) && (144)\\ &= 1 - \sum_{x=0}^{3} \frac{\mu^x e^{-\mu}}{x!} && (145)\\ &= 1 - (0.0111 + 0.0500 + 0.1125 + 0.1687) && (146)\\ &= 1 - 0.342 && (147)\\ &= 65.8\% && (148) \end{aligned}$$

(f) Die Anzahl X der Fehler auf einer Seite ist poissonverteilt

$$X \sim \mathrm{Po}(\mu). \tag{149}$$

mit dem Parameter

$$\begin{aligned} \mu &= n \cdot \theta && (150)\\ &= 1 \cdot 3 && (151)\\ &= 3 && (152) \end{aligned}$$

Die Wahrscheinlichkeit, dass eine Seite fehlerfrei ist, ist

$$\begin{aligned} P(X=0) &= \frac{\mu^0 \cdot e^{-\mu}}{0!} && (153)\\ &= \frac{1 \cdot e^{-3}}{1} && (154)\\ &= 5\% && (155) \end{aligned}$$

Die Wahrscheinlichkeit, dass mindestens eine Seite einer $m = 50$-seitigen Arbeit fehlerfrei ist, beträgt

$$\begin{aligned} P &= 1 - \big(1 - P(X=0)\big)^m && (156)\\ &= 1 - \big(1 - 0.05\big)^{50} && (157)\\ &= 92.3\% && (158) \end{aligned}$$

Lösung T.2.6

(a) Nach den Angaben gilt fü die Ereignisse i =A, B oder U mit den dazugehörigen Quoten q_i, dass die Wahrscheinlichkeiten p_i proportional zu den inversen Quoten sind. Die Proportionalitätskonstante ergibt sich aus der Normierungsbedingung:

$$p_i = \frac{c}{q_i}, \quad p_1 + p_2 + p_3 = 1 \quad \Rightarrow \quad p_i = \frac{1/q_i}{1/q_1 + 1/q_2 + 1/q_3} \tag{159}$$

Mit $q_1 = q_A = 2.6$, $q_2 = q_B = 3.2$ und $q_3 = q_U = 2.4$ gilt

$$p_A = \frac{\frac{1}{2.6}}{\frac{1}{2.6} + \frac{1}{3.2} + \frac{1}{2.4}} = 0.345, \tag{160}$$

$$p_B = 0.281, \quad p_U = 0.374. \tag{161}$$

Hintergrund (nicht verlangt): Die subjektiven Definition kann überall dort zum Einsatz kommen, wo man mit richtigen Voraussagen Geld (oder Ehre) gewinnen kann, man mit seiner Entscheidung "den Markt" beeinflusst, so dass sichere Gewinne unmöglich werden und die Gewinnfunktion eine kardinalskalierte Größe ist. Hier kann man mit richtigen Tipps reales Geld (kardinalskaliert) gewinnen und beeinflusst mit seinen Entscheidungen den Markt, zumindest indirekt über den Einfluss auf die Quoten des Anbieters.

(b) Seien die Anteile der Wetteinsätze auf die Ereignisse i = A,B oder U durch θ_i gegeben. Dann ist der Erwartungswert des Gewinns des Wettanbieters in Einheiten des Gesamteinsatzes gegeben durch

$$G = 1 - \sum_{i=1}^{3} \theta_i p_i q_i = 1 - \sum_{i=1}^{3} \theta_i \frac{1}{\sum\limits_{j=1}^{3} \frac{1}{q_j}} = 1 - \frac{1}{\sum\limits_{j=1}^{3} \frac{1}{q_j}} \tag{162}$$

$$= 1 - \frac{2.6 \cdot 0.345 + 3.2 \cdot 0.281 + 2.4 \cdot 0.374}{3} = 0.102. \tag{163}$$

Hier wurde die Normierungsbedingung $\sum_i \theta_i = 1$ verwendet. Der Wettanbieter gewinnt also, *unabhängig davon, auf welches Ereignis die Spieler setzen*, im Mittel 10% der Einsätze. Wichtig ist dabei nur, dass die subjektive Wahrscheinlichkeit die tatsächliche widerspiegelt. Beeinflussen Wettbetrüger die tatsächlichen Ausgänge (z.B. durch Bestechung des Schiedsrichters und gleichzeitig hohe Einsätze auf den begünstigten Club), kann der Erwartungswert des Gewinns des Anbieters durchaus negativ werden!

(c) Statistische Wahrscheinlichkeiten = relative Häufigkeiten in der Zeitreihe der Vergangenheit:

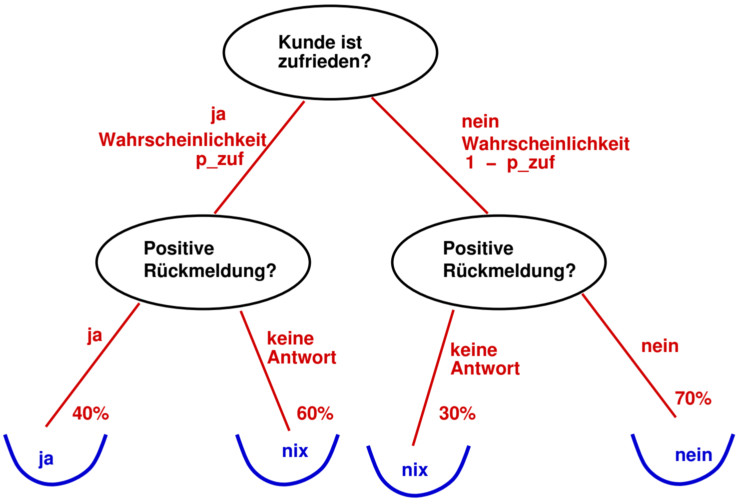
$$p_A^{stat} = \frac{3}{4}, \quad p_B = \frac{1}{4}, \quad p_U = 0. \tag{164}$$

Die klassische Laplace'sche Definition kann man nicht anwenden, da man die Grundgesamtheit, welche auch alle möglichen Spiele in der Zukunft mit einschließt, nicht abzählen kann. Außerdem sind die Elementarereignisse (z.B. 2:0, 2:1, 4:2 etc für das Ereignis A) nicht gleich wahrscheinlich.

Bemerkung am Rande: Am 17.2. hat Werder Bremen unerwartet gegen irgendein Bundesliga-Schlusslicht verloren. Daraufhin wurden auch die Quoten verändert, z.B. Stand 20.2.07, 20:30 h: $q_A = 2.4$, $q_B = 2.7$, $q_U = 3.2$. Die Gewinnwahrscheinlichkeit für Bremen beträgt damit nur noch $p_B = 0.337$ statt 0.374 (in Klausur der Wert von p_U, da dort p_A und p_U vertauscht wurde).

Lösung T.2.7

(a) Wahrscheinlichkeitsbaum:



(b) Es sei n_p die Zahl der positiven und n_n die Zahl der negativen Antworten. Anteilswert der positiven Antworten:

$$A_p = \frac{n_p}{n_p + n_n} = \frac{353}{353 + 402} = 0.4675. \tag{165}$$

Erwarteter Anteil der zufriedenen Kunden p_{zuf} unter der Berücksichtigung der unterschiedlichen Antwortraten: Es sei

- $p_{pz} = P(\text{positiv}|\text{zufrieden}) = 0.4$,
- $p_{p\bar{z}} = P(\text{positiv}|\overline{\text{zufrieden}}) = 0.0$,
- $p_{nz} = P(\text{negativ}|\text{zufrieden}) = 0.0$,
- $p_{n\bar{z}} = P(\text{negativ}|\overline{\text{zufrieden}}) = 0.7$

Dann ergibt sich aus dem Verhältnis der Antworten mit dem W-Baum:

$$q = \frac{n_p}{n_n} = \frac{353}{402} = 0.8781 = \frac{p_{zuf} p_{pz}}{(1 - p_{zuf}) p_{n\bar{z}}} \tag{166}$$

und nach Auflösung

$$q(1 - p_{zuf})p_{n\bar{z}} = p_{zuf} p_{pz} \tag{167}$$
$$q p_{n\bar{z}} - q p_{zuf} p_{n\bar{z}} = p_{zuf} p_{pz} \tag{168}$$
$$q p_{n\bar{z}} = p_{zuf}(p_{pz} + q p_{n\bar{z}}) \tag{169}$$
$$p_{zuf} = \frac{q p_{n\bar{z}}}{p_{pz} + q p_{n\bar{z}}} \tag{170}$$
$$= \frac{0.8781 \cdot 0.7}{0.4 + 0.8781 \cdot 0.7} = 0.6058. \tag{171}$$

(c) Nun gilt $p_{pz} = 0.4 - 0.1 \cdot 0.4 = 0.36$, $p_{nz} = 0.1 \cdot 0.4 = 0.04$, $p_{p\bar{z}} = 0.7 \cdot 0.1 = 0.07$ und $p_{n\bar{z}} = 0.7 - 0.1 \cdot 0.7 = 0.63$.

Damit ergibt sich für das bekannte Antwortverhältnis q nun

$$q = \frac{n_p}{n_n} = \frac{p_{zuf}p_{pz} + (1 - p_{zuf})p_{p\bar{z}}}{(1 - p_{zuf})p_{n\bar{z}} + p_{zuf}p_{nz}} \tag{172}$$

und nach Auflösung

$$q = \frac{0.36p_{zuf} + (1 - p_{zuf})0.07}{(1 - p_{zuf})0.63 + 0.04p_{zuf}} \tag{173}$$

$$q = \frac{0.29p_{zuf} + 0.07}{0.63 - 0.59p_{zuf}} \tag{174}$$

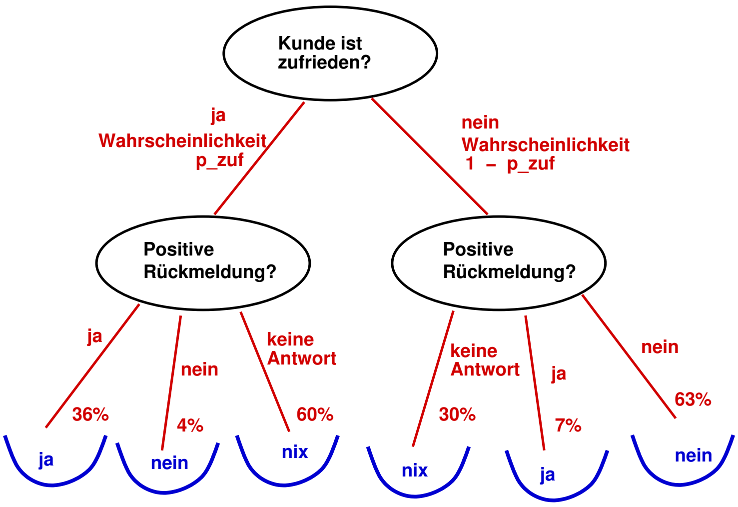
$$0.63q - 0.59qp_{zuf} = 0.29p_{zuf} + 0.07 \tag{175}$$

$$\frac{0.063q - 0.07}{0.29 + 0.59q} = p_{zuf} \tag{176}$$

mit $q = 0.8781$ ergibt sich

$$p_{zuf} = \frac{qp_{n\bar{z}} - p_{p\bar{z}}}{qp_{n\bar{z}} + p_{pz} - p_{p\bar{z}} - qp_{nz}} = 0.5980. \tag{177}$$

Siehe auch den W-Baum (nicht verlangt!) für diese Situation:



(d) Konfidenzintervall für Anteilswert für $\alpha = 4\%$:

$$\theta \in [A - \Delta A, A + \Delta A] \tag{178}$$

mit

$$A = \frac{353}{353 + 402} = 0.4675 \tag{179}$$

$$\Delta A = z_{1-\alpha/2}\sqrt{\frac{A(1-A)}{n}} = 2.054\sqrt{\frac{0.4675(1 - 0.4675)}{755}} = 0.00136 \tag{180}$$

und damit

$$\theta \in [0.466, 0.469] \tag{181}$$

Lösung T.2.8

Sei A_i das Ereignis "Ein Mitglied ist in Schadenfreiheitsklasse (SFK) i". Die Wahrscheinlichkeiten ergeben sich aus den relativen Häufigkeiten zu

$$P(A_1) = \frac{50000}{276667} = 0.1807, \quad P(A_2) = \frac{80000}{276667} = 0.2892, \tag{182}$$

$$P(A_3) = \frac{80000}{276667} = 0.2892, \quad P(A_4) = \frac{66667}{276667} = 0.2409. \tag{183}$$

Die bedingten Wahrscheinlichkeiten $P(A_i|S)$, dass ein Schadensfall (Ereignis S) von einem Mitglied der SFK i verursacht wird, lassen sich direkt aus der Tabelle ablesen:

$$P(A_1|S) = 0.3, \quad P(A_2|S) = 0.3, \quad P(A_3|S) = 0.24, \quad P(A_4|S) = 0.16. \tag{184}$$

Die gesuchten Wahrscheinlichkeiten $P(S|A_i)$, dass ein Mitglied der SFK i einen Schaden meldet, wird duch den Satz von Bayes berechnet:

$$P(S|A_i) = \frac{P(A_i|S)P(S)}{P(A_i)} \tag{185}$$

und damit insbesondere die Quotienten, bei denen sich die zwar berechenbare, aber irrelevante unbedingte Schadenswahrscheinlichkeit $P(S)$ herauskürzt:

$$\frac{P(S|A_i)}{P(S|A_1)} = \frac{P(A_i|S)P(A_1)}{P(A_1|S)P(A_i)} \tag{186}$$

Diese Bruchteile geben direkt den Rabatt in Form des Bruchteils des von der SFK1 zu zahlenden Versicherungsbetrages an:

$$\frac{P(S|A_2)}{P(S|A_1)} = \frac{0.3 \cdot 0.1807}{0.3 \cdot 0.2892} = 0.625, \quad \frac{P(S|A_3)}{P(S|A_1)} = 0.5, \quad \frac{P(S|A_4)}{P(S|A_1)} = 0.4. \tag{187}$$

Lösung T.2.9

(a) Die Wahrscheinlichkeit von 57% Staus bezieht sich auf den ganzen Nachmittag, während für den Pendler nur die Staus relevant sind, die ihn behindern, wenn er gerade vorbeifährt. Dies ist natürlich nur eine Untermenge aller nachmittäglichen Staus, daher die unterschiedlichen Wahrscheinlichkeiten.

(b) Sei

- A das Ereignis "Stau vorhanden",
- B das Ereignis "eine Staumeldung wurde durchgegeben".

Dann gilt aus der Fragestellung

$$P(A) = 0.2, \quad P(B) = \frac{0.2}{0.8} = 0.25, \quad P(B|A) = 0.8. \tag{188}$$

Aus dem Satz von Bayes ergibt sich damit

(i) Wahrscheinlichkeit für Stau, falls Staumeldung:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.2 \cdot 0.8}{0.25} = 0.64 \tag{189}$$

(ii) Wahrscheinlichkeit für Stau, falls keine Staumeldung:

$$P(A|\bar{B}) = \frac{P(A)P(\bar{B}|A)}{P(\bar{B})} = \frac{P(A)(1 - P(B|A))}{1 - P(B)} = \frac{0.2 \cdot 0.2}{0.75} = 0.053 \tag{190}$$

(c) Wahrscheinlichkeit von Falschmeldungen:

$$P(B|\bar{A}) = \frac{P(B)P(\bar{A}|B)}{P(\bar{A})} = \frac{P(B)(1 - P(A|B))}{1 - P(A)} = \frac{0.25 \cdot (1 - 0.64)}{0.8} = 0.1125 \tag{191}$$

(d) Es sei

- C das Ereignis "Stau auf Nebenstrecke"

mit

$$P(C|B) = 0.5, \quad P(C|\bar{B}) = 0.1 \tag{192}$$

(wobei $P(C|B) = 0.5$ entspricht Stau auf der Nebenstrecke, falls Staumeldung und $P(C|\bar{B})$ bedeutet Stau auf der Nebenstrecke, falls keine Durchsage)

(i) Hier ist es wichtig, weniger als 20 Minuten zu brauchen. Aus der Fahrtzeit - Entscheidungsmatrix sieht man, dass man dazu die Wahrscheinlichkeit, in einen Stau -egal, ob auf Haupt- oder Nebenstrecke- minimieren muss. Falls eine Stauwarnung gegeben wurde,trifft B zu und die Stauwahrscheinlichkeit $P(A|B) = 0.64$ auf der Hauptstrecke ist größer als die Stauwahrscheinlichkeit $P(C|B)$ auf der Nebenstrecke. Also ist eine Umfahrung sinnvoll.

(ii) Falls man die mittlere Reisezeit optimieren will, gilt es, den Erwartungswert $\langle \tau \rangle$ der Reisezeit zu minimieren. Sei

- D die Entscheidung "Ich nutze die Nebenstrecke"

Dann gilt (in Einheiten von Minuten):

$$\langle \tau \rangle \,|_D = 10 \text{ min } P(\bar{C}|B) + 40 \text{ min } P(C|B) = 10 \cdot (1 - 0.5) + 40 \cdot 0.5 = 25 \text{ min} \quad (193)$$

und

$$\langle \tau \rangle \,|_{\bar{D}} = 5 \text{ min } P(\bar{A}|B) + 30 \text{ min } P(A|B) = 5 \cdot (1 - 0.64) + 30 \cdot 0.64 = 21 \text{ min} \quad (194)$$

Also ist in diesem Falle die Entscheidung $\bar{D}$: “Ich bleibe auf der Autobahn” besser.

Lösung T.2.10

(a) Flugzeug: Die Ausfallwahrscheinlichkeit von Querruder, Triebwerke und Bremsklappen sind durch $p_Q = p_T = p_B = 10^{-3}$ gegeben. Die Wahrscheinlichkeit, dass während eines Fluges alle drei Systeme ausfallen und damit das Flugzeug lenkunfähig wird, ist durch

$$P(\text{Flugzeug wird lenkunfähig}) = p_Q p_T p_B = 10^{-9} \tag{195}$$

gegeben. Diese Wahrscheinlichkeit von $1 : 10^9$ muss übrigens beim derzeitigen Standard unterschritten werden, um ein Flugzeug zuzulassen.

(b) Die drei unabhängigen Einheiten C_i, $i = 1, 2, 3$, von Sensoren und Steuergeräten fallen je mit der Wahrscheinlichkeit $\theta = 10^{-4}$ pro Fahrt aus. Die Wahrscheinlichkeit, dass während einer Fahrt i dieser Systeme ausfallen, ist durch die Binomialverteilung

$$P(i \text{ Computer (“Steuergeräte”) fallen aus}) = p_B^{(3,\theta)}(i) = \binom{3}{i} \theta^i (1-\theta)^{3-i} \tag{196}$$

gegeben, also insbesondere die Ausfallwahrscheinlichkeiten für 2 oder 3 Steuergeräte durch

$$p_B^{(3,\theta)}(2) = 3\theta^2(1-\theta) = 3\ (10^{-8} + 10^{-12}), \quad p_B^{(3,\theta)}(3) = \theta^3 = 10^{-12}. \tag{197}$$

Damit ist die Wahrscheinlichkeit für den Ausfall von mindestens zwei Steuergeräten (bei zwei signifikanten Stellen) durch

$$p_C = p_B^{(3,\theta)}(2) + p_B^{(3,\theta)}(3) \approx 3 \cdot 10^{-8} \tag{198}$$

gegeben.

(c) Hier ist der Sachverhalt wie beim Flugzeug: Mindestens einer der Stellmotoren reicht aus. (Ein Ausfall eines Motors führt zum Stillstand dieses Motors, so dass, im Gegensatz zu den Steuergeräten für die Umsetzung der Lenkbewegung, auch ohne Voter eindeutig der defekte vom intakten Motor unterschieden werden kann). Damit ist bei der Ausfallwahrscheinlicheit $\hat{\theta} = 10^{-4}$ eines Motors die Versagenswahrscheinlichkeit

$$p_M = P(\text{alle beiden Stellmotoren fallen aus}) = (\hat{\theta})^2 = 10^{-8} \tag{199}$$

(d) Versagenswahrscheinlichkeit p_L des Gesamtsystems “Lenkung”:

$$p_L = P(\text{Steuergeräte oder Stellmotoren versagen}) = p_C + p_M - p_C p_M \approx p_C + p_M = 4 \cdot 10^{-8} \tag{200}$$

Dabei wurde die bei Unabhängigkeit von A und B gültige Beziehung

$$\begin{aligned} P(A \cup B) &= P(\overline{\bar{A} \cap \bar{B}}) = 1 - P(\bar{A} \cap \bar{B}) && (201)\\ &= 1 - P(\bar{A})\ P(\bar{B}|\bar{A}) = 1 - P(\bar{A})\ P(\bar{B}) && (202)\\ &= 1 - (1 - p_C)(1 - p_M) = p_C + p_M - p_C p_M && (203) \end{aligned}$$

angewandt und das Ergebnis auf zwei signifikante Stellen gerundet.

(e) Hier gilt im Prinzip dieselbe Formel, nur das für einen “Elektronikfehler” bereits das Versagen eines Steuergerätes oder eines Motors ausreicht. In diesem Falle würde sich nämlich auch bei noch funktionierender Lenkung die Versagenswahrscheinlichkeit drastisch erhöhen, so dass das Fahrzeug aus Sicherheitsgründen nicht mehr gestartet werden kann. Also

$$p_{Lenkelektronikfehler} = \tilde{p}_C + \tilde{p}_M - \tilde{p}_C\ \tilde{p}_M \approx \tilde{p}_C + \tilde{p}_M = 5 \cdot 10^{-4} \tag{204}$$

Dabei wurden die Wahrscheinlichkeiten für den Ausfall mindestens eines Steuergerätes oder Stellmotors durch die Binomialverteilung

$$\tilde{p}_C = 1 - p_B^{(3,\theta)}(0) \approx 3 \cdot 10^{-4}, \quad \tilde{p}_M = 1 - p_B^{(2,\hat{\theta})}(0) \approx 2 \cdot 10^{-4} \tag{205}$$

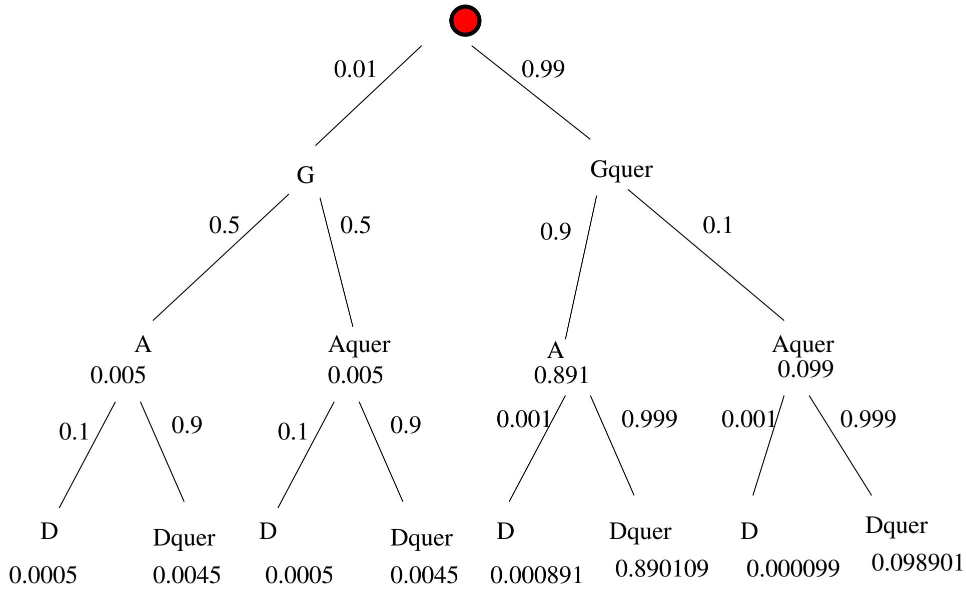
ermittelt.

Lösung T.2.11

Seien folgende Ereignisse definiert:

G	Gestohlenes Fahrzeug
A	Ausweis kann vorgezeigt werden
D	Drogen im Auto mitgeführt

Wahrscheinlichkeitsbaum:



(a) Ablesen der mittleren "Verzweigungsebene" des W-Baumes ergibt

(i) Unbedingte Wahrscheinlichkeit

$$P(A) = 0.005 + 0.891 = 89.6\% \tag{206}$$

(ii) Bedingte Wahrscheinlichkeit für gestohlenes Fahrzeug bei Vorweis eines Fahrzeugscheins:

$$P(G|A) = \frac{P(G \cap A)}{P(A)} = \frac{0.005}{0.005 + 0.891} = 0.558\% \tag{207}$$

(iii) Bedingte Wahrscheinlichkeit für gestohlenes Fahrzeug, falls kein Fahrzeugschein vorzeigbar:

$$P(G|\bar{A}) = \frac{P(G \cap \bar{A})}{P(\bar{A})} = \frac{0.005}{0.005 + 0.099} = 4.81\% \tag{208}$$

(b) Für die Wahrscheinlichkeiten des Drogentransports muss man auf die unterste "Verästelungsebene" des W-Baumes gehen:[2]

$$P(D \cap A) = 0.0005 + 0.000891 = 0.001391, \quad P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{0.001391}{0.896} = 0.155\% \tag{209}$$

[2] Man kann die mittlere und untere Ebene des W-Baumes auch vertauschen. Dann ergibt sich Aufgabenteil (b) aus der mittleren und (a) aus der unteren Ebene.

und

$$P(D \cap \bar{A}) = 0.0005 + 0.000099 = 0.000599, \quad P(D|\bar{A}) = \frac{P(D \cap \bar{A})}{P(\bar{A})} = \frac{0.000599}{1 - 0.896} = 0.576\%. \tag{210}$$

Diskussion (kein Bestandteil der eigentlichen Klausur)

Der Polizist erhält aus der W-Theorie folgende konkret relevante Informationen:

(i) Falls der Fahrzeugschein nicht vorgezeigt werden kann, wird durch diese Information die *a-priori-* Wahrscheinlichkeit für ein gestohlenes Fahrzeug von 1% um fast den Faktor 5 erhöht. Das Vorzeigen der Papiere erniedrigt hingegen reduziert diese Wahrscheinlichkeit um weniger als den Faktor 2.

(ii) Bei der Frage nach dem Drogentransport hilft die Information, ob ein Schein vorgezeigt werden kann, weniger deutlich: Die *a-priori*-Wahrscheinlichkeit für Drogentransport von etwa 2 Promille wird durch das Vorzeigen der Papiere nur marginal auf 1.55 Promille reduziert, während das Fehlen der Fahrzeugpapiere die Wahrscheinlichkeit immerhin auf 5.76 Promille erhöht.

(iii) Die ganze Problematik macht deutlich, wie die Wahrscheinlichkeit von der Information abhängt und sich bei Vorliegen neuer Information schlagartig ändern kann.

Lösung T.2.12

Lösungsvorschlag folgt.

Lösung T.2.13

(a) Es seien folgende Ereignisse definiert:

A: Inspektion wurde durchgeführt

B: Es gab mindestens eine Panne auf der Fahrt in die Türkei

C: Es gab mindestens eine Panne auf 1000 km Fahrt

Bekannt ist die unbedingte (totale) Wahrscheinlichkeit

$P(A) = 0.8,$

sowie die bedingten Wahrscheinlichkeiten

$P(C|A) = 0.005$ Pannenwahrscheinlichkeit pro 1000 km bei durchgeführter Inspektion

$P(C|\bar{A}) = 0.03$ Pannenwahrscheinlichkeit pro 1000 km, falls die Inspektion nicht durchgeführt wurde

Da die Pannenwahrscheinlichkeit nicht vom Streckenabschnitt bzw. von schon erlittenen Pannen abhängt, gilt für die Wahrsch., mit (A) oder ohne ($\bar{A}$) Inspektion *keine* Panne zu erleiden:

$$P(\bar{B}|A) = (1 - P(C|A))^4 = (1 - 0.005)^4 \tag{211}$$

bzw.

$$P(\bar{B}|\bar{A}) = (1 - P(C|\bar{A}))^4 = (1 - 0.03)^4 \tag{212}$$

Damit ergibt sich für die Wahrscheinlichkeiten, mindestens eine Panne zu erleiden:

$$P(B|A) = 1 - P(\bar{B}|A) = 1 - (1 - 0.005)^4 = 1.985\%, \tag{213}$$

$$P(B|\bar{A}) = 1 - P(\bar{B}|\bar{A}) = 1 - (1 - 0.03)^4 = 11.47\% \tag{214}$$

(b) Unbedingte (totale) Wahrscheinlichkeit $P(B)$:

$$P(B) = \sum_k P(B|A_k)P(A_k) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.0388 \tag{215}$$

Hier wurde $A_1 = A$ und $A_2 = \bar{A}$ gesetzt.

Die gesuchte Wahrscheinlichkeit $P(\bar{A}|B)$ ergibt sich mit dem Satz von Bayes (mit $A_k = A_2 = \bar{A}$):

$$P(\bar{A}|B) = \frac{P(B|\bar{A})P(\bar{A})}{P(B)} = \frac{0.1147 \cdot 0.2}{0.0388} = 0.59 \tag{216}$$

Während die A-Priori-Wahrscheinlichkeit, keine Inspektion durchgeführt zu haben, nur $P(\bar{A}) = 0.2$ beträgt, steigt sie auf $P(\bar{A}|B) = 59\%$, wenn man die zusätzliche Information hat, dass eine Panne vorliegt.

(c) Sei Ereignis D: "Sonsiger Grund für Verspätung". Dann gilt für Ereignis $E = D \cup B$: "Machinenteile kommen verspätet an" bei Unabhängigkeit der sonstigen Ursachen wie Zoll etc. von den Pannen:

$$P(E) = P(D \cup B) \stackrel{\text{DeMorgan}}{=} 1 - P(\bar{D} \cap \bar{B}) \stackrel{\text{Unabhh.}}{=} 1 - P(\bar{D})P(\bar{B}) = 1 - 0.99 \cdot 0.9612 = 4.84\% \tag{217}$$

Lösung T.2.14

Aus der Aufgabenstellung sind gegeben

$$P(A) = 50\% \qquad P(\overline{A}) = 50\% \tag{218}$$
$$P(B|A) = 40\% \qquad P(\overline{B}|A) = 60\% \tag{219}$$
$$P(B|\overline{A}) = 80\% \qquad P(\overline{B}|\overline{A}) = 20\% \tag{220}$$

(a) Totale Wahrscheinlichkeit

$$P(B) = \sum_{A_i=A,\overline{A}} P(B|A_i) \cdot P(A_i) \tag{221}$$
$$= P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A}) \tag{222}$$
$$= 40\% \cdot 50\% + 80\% \cdot 50\% \tag{223}$$
$$= 20\% + 40\% \tag{224}$$
$$= 60\% \tag{225}$$

(b) Satz von Bayes

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{226}$$
$$= \frac{40\% \cdot 50\%}{60\%} \tag{227}$$
$$= 33.3\% \tag{228}$$

(c) Die Zahl X der noch benötigten Stellplätzen ist binomialverteilt

$$X \sim B(n = 4, \theta = 66.7\%) \tag{229}$$

mit der Wahrscheinlichkeitsfunktion

$$P(X = x) = p_B^{(n,\theta)} = \binom{n}{x} \cdot \theta^x \cdot (1-\theta)^{n-x}. \tag{230}$$

So berechnet sich

$$P(X \leq 2) = 1 - P(X \geq 3) \tag{231}$$
$$= 1 - \big[P(X = 3) + P(X = 4)\big] \tag{232}$$
$$= 1 - \left[\binom{4}{3} \cdot 0.667^3 \cdot (1 - 0.667)^{4-3} + \binom{4}{4} \cdot 0.667^4 \cdot (1 - 0.667)^{4-4}\right] \tag{233}$$
$$= 1 - \big[4 \cdot 0.667^3 \cdot 0.333 + 1 \cdot 0.667^4 \cdot 1\big] \tag{234}$$
$$= 100\% - \big[39.53\% + 19.79\%\big] \tag{235}$$
$$= 40.68\% \tag{236}$$

(d) Ob ein Stellplatz i noch genutzt wird ($X_i = 1$) oder frei wird ($X_i = 0$) ist binomialverteilt

$$X_i \sim B(n = 1, \theta = 66.7\%) \tag{237}$$

mit dem Erwartungswert und der Varianz

$$\mu_i = \theta = 0.667 \tag{238}$$
$$\sigma_i^2 = \theta \cdot (1-\theta) = 0.2221 \tag{239}$$

Die Gesamtzahl X der benötigten Stellplätze ergibt sich als die Summe

$$X = \sum_{i=1}^{208} X_i \tag{240}$$

kann aufgrund des zentralen Grenzwertsatzes als normalverteilt

$$X \sim N(\mu, \sigma^2) \tag{241}$$

mit

$$\mu = \sum_{i=1}^{208} \mu_i = 208 \cdot 0.667 \qquad = 138.74 \tag{242}$$
$$\sigma^2 = \sum_{i=1}^{208} \sigma_i^2 = 208 \cdot 0.2221 \qquad = 46.197 \tag{243}$$

angenommen werden. Gesucht ist nun nach der Wahrscheinlichkeit

$$P(X \leq 150) = \Phi\left(z = \frac{150-\mu}{\sigma}\right) \tag{244}$$
$$= \Phi\left(z = \frac{150-138.74}{\sqrt{46.197}}\right) \tag{245}$$
$$= \Phi\left(z = 1.657\right) \tag{246}$$
$$\approx 95\% \tag{247}$$

Lösung T.2.15

(a) Direkt aus der Aufgabenstellung folgt

$$P(B|A) = 80\% \qquad P(\overline{B}|A) = 20\% \tag{248}$$
$$P(B|\overline{A}) = 30\% \qquad P(\overline{B}|\overline{A}) = 70\% \tag{249}$$

(b) Mit

$$P(A) = 40\% \tag{250}$$
$$P(\overline{A}) = 60\% \tag{251}$$

ergibt sich für die totale Wahrscheinlichkeit $P(B)$

$$P(B) = \sum_{A_i = A, \overline{A}} P(B|A_i) \cdot P(A_i) \tag{252}$$
$$= P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A}) \tag{253}$$
$$= 80\% \cdot 40\% + 30\% \cdot 60\% \tag{254}$$
$$= 32\% + 18\% \tag{255}$$
$$= 50\%. \tag{256}$$

(c) Mit dem Satz von Bayes lässt sich

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{257}$$
$$= \frac{80\% \cdot 40\%}{50\%} \tag{258}$$
$$= 64\% \tag{259}$$

ermitteln.

(d) Wieder mit dem Satz von Bayes:

$$P(\overline{A}|\overline{B}) = \frac{P(\overline{B}|\overline{A}) \cdot P(\overline{A})}{P(\overline{B})} \tag{260}$$
$$= \frac{70\% \cdot 60\%}{1 - 50\%} \tag{261}$$
$$= 84\% \tag{262}$$

(e) Mit gegebenem $P(B) = 55\%$ berechnet sich aus der Formel der totalen Wahrscheinlichkeit die gesuchte Wahrscheinlichkeit $P(A)$ durch Umstellen wie folgt

$$P(B) = \sum_{A_i = A, \overline{A}} P(B|A_i) \cdot P(A_i) \tag{263}$$

$$P(B) = P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A}) \qquad \Big| P(\overline{A}) = 1 - P(A) \tag{264}$$

$$P(B) = P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot \big[1 - P(A)\big] \qquad \Big| - P(B|\overline{A}) \tag{265}$$

$$P(B) - P(B|\overline{A}) = P(A) \cdot \big[P(B|A) - P(B|\overline{A})\big] \qquad \Big| : \big[P(B|A) - P(B|\overline{A})\big] \tag{266}$$

$$P(A) = \frac{P(B) - P(B|\overline{A})}{P(B|A) - P(B|\overline{A})} \tag{267}$$

$$= \frac{55\% - 30\%}{80\% - 30\%} \tag{268}$$

$$= 50\% \tag{269}$$

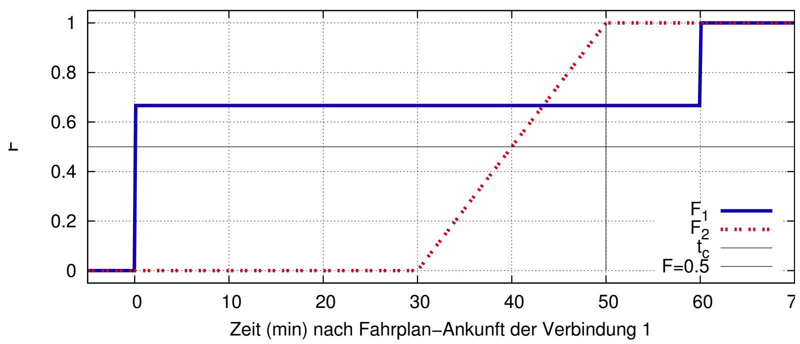
3 Wahrscheinlichkeitsverteilungen

Lösung T.3.1

(a) **Verteilungsfunktionen der Ankunftszeiten:** Die Verteilungsfunktionen der Verbindung 1 (mit Umsteigen) und Verbindung 2 (Direktverbindung) lauten [nicht verlangt, nur zeichnen]

$$F_1(T) = \begin{cases} 0 & T < 0 \\ 2/3 & 0 \le T \le 60 \\ 1 & \text{sonst.} \end{cases}, \quad F_2(T) = \begin{cases} 0 & T < 30 \\ \frac{T-30}{20} & 30 \le T \le 50 \\ 1 & \text{sonst.} \end{cases}$$

Der Wert 2/3 in der Verteilungsfunktion ist gleich der Wahrscheinlichkeit p_{happy}, den Anschlusszug *nicht* zu verpassen. Mit 20 min fahrplanmäßige Umsteigezeit, einer $G(0, 30)$-gleichverteilten Verspätung der Ankunft des ersten Zuges und pünktlichen Anschlusszug wird der Anschluss verpasst, wenn die Verspätung größer als 20 min ist, was mit der Wahrsch. $1 - p_{\text{happy}} = 1/3$ geschieht.



(b) **Median** (vgl. die obige Abbildung):

$$\text{Verbindung 1: } T_{0.5} = 0, \quad \text{Verbindung 2: } T_{0.5} = 30 + 0.5 \cdot 20 = 40 \tag{270}$$

Erwartungswert:

$$\text{Verbindung 1:} \qquad E(T) = 2/3 \cdot 0 + 1/3 \cdot 60 = 20, \tag{271}$$

$$\text{Verbindung 2:} \qquad E(T) = 30 + 0.5 \cdot 20 = 40. \tag{272}$$

(c) **Wahl der Verbindung**

(i) Falls das Ziel eine maximale Zeitersparnis im Mittel ist, sollte die Verbindung mit den kleineren Erwartungswert gewählt werden, also Verbindung 1.

(ii) Falls maximal 50 Minuten Verspätung erlaubt sind: Es sollte die Verbindung gewählt werden, für die dies mit höherer Wahrscheinlichkeit

$$P(T \le 50) = F(50) \tag{273}$$

erfüllt ist:

$$\text{Verbindung 1: } F(50) = 2/3, \quad \text{Verbindung 2: } F(50) = 1, \tag{274}$$

Also Verbindung 2.

(d) **Wahrscheinlichkeit und Streuparameter:**

- Standardabweichung:

$$\sigma_1^2 = 60^2 \left[\frac{2}{3} \left(\frac{1}{3} \right)^2 + \frac{1}{3} \left(\frac{2}{3} \right)^2 \right] = 60^2 \frac{6}{27}, \qquad \sigma = 20\sqrt{2} = 28.2, \tag{275}$$

$$\sigma_2^2 = \frac{20^2}{12}, \qquad \sigma_2 = \frac{20}{\sqrt{12}} = 5.8. \tag{276}$$

- Wahrscheinlichkeit $P(T < 45) = F(45)$:

$$\text{Verbindung 1: } F(45) = 2/3, \quad \text{Verbindung 2: } F(45) = 3/4. \tag{277}$$

- Spannweite:

$$\text{Verbindung 1: } R = 60, \quad \text{Verbindung 2: } R = 20. \tag{278}$$

Lösung T.3.2

(a) Die Zahl der ÖPNV-Entscheidungen kann nur dann mit der Binomialverteilung beschrieben werden, wenn die Entscheidungen der zehn Studenten bzw. Erwerbstätigen voneinander unabhängig sind. Dann hat man jeweils zehn unabhängige Wiederholungen eines Bernouulli-Experiments, die also durch die Binomialverteilung $B(10, \theta)$ beschrieben werden. Der Parameter $n = 10$ (in beiden Gruppen) ist also fest.

(b) Dazu nimmt man für die Studenten die Zeile mit $x = 7$ ÖV-Entscheidungen (für die Erwerbstätigen $x = 2$) und maximiert die Wahrscheinlichkeit bezüglich θ. Also $\theta = 0.7$ für Studenten (maximierte Wahrscheinlichkeit 0.27) und $\theta = 0.2$ für Erwerbstätige (maximierte Wahrscheinlichkeit 0.30)

(c) Binomialverteilung:

$$p_B^{(10,\theta)}(x) = \binom{10}{x} \theta^x (1-\theta)^{10-x} \tag{279}$$

Dies fasst man nun bei *festen* Wert von x als Funktion $L(\theta)$ des Parameters θ auf und maximiert bezüglich θ (unter Anwendung der Produkt und Kettenregel):

$$\frac{dL(\theta)}{d\theta} = \binom{10}{x} \left[x\theta^{x-1}(1-\theta)^{10-x} + \theta^x (10-x)(1-\theta)^{10-x-1}(-1) \right] \tag{280}$$

$$= \binom{10}{x} \theta^{x-1}(1-\theta)^{10-x-1} \left[x(1-\theta) - \theta(10-x) \right] \stackrel{!}{=} 0 \tag{281}$$

Also $x = \theta(10 - x + x)$ oder $\theta = \frac{x}{10}$. Die bei (b) abgelesenen Werte von $\theta = 0.7$ bzw. $\theta = 0.2$ treffen also exakt zu.

(d) Beschreibung mit Poissonverteilung ist möglich, falls ein Defekt keine weiteren Defekte nach sich zieht, also vollständig behoben wird (und es sich wirklich um Defekte, nicht um das Ersetzen von Verschleißteilen handelt, die danach natürlich nicht sofort wieder verschlissen sind).

Eine Binomialverteilung ist möglich, aber unpraktisch, da man ihren zweiten Parameter n ad-hoc definieren muss (z.B. $n = 100\,000$ Kilometerintervalle) und große Zahlen bekommt.

Da die Defekthäufigkeit eine niedrige ganze Zahl ist, ist eine Beschreibung mit der nur für stetige oder quasistetige Merkmale anwendbare Gaußverteilung nicht möglich.

(e)

$$\frac{dp_{Po}^{(\mu)}(x)}{d\mu} = \frac{d}{d\mu} \left(\mu^x e^{-\mu} / x! \right) \tag{282}$$

$$= \frac{\mu^{x-1} e^{-\mu}}{x!} (x - \mu) \stackrel{!}{=} 0 \tag{283}$$

Also ist der die Daten am besten beschreibende Poissonparameter μ gegeben durch $\mu = x$

Lösung T.3.3

(a) Die Exponentialverteilung $E(\lambda)$ mit dem Parameter λ hat die Dichtefunktion

$$f_E^{(\lambda)}(x) = \begin{cases} \frac{1}{\lambda}e^{-\lambda x} & x \geq 0 \\ 0 & \text{sonst.} \end{cases}$$

Der Modalwert (Maximum der Dichtefunktion) der Exponentialverteilung liegt also bei $x = 0$. Auf einer einspurigen Strecke mit dichtem Verkehr kommt aber wegen des notwendigen Sicherheitsabstandes zweier Fahrzeuge voneinander der Wert Null überhaupt nicht vor, also wäre eine Expoentialverteilung der Abstände dem Sachverhalt nicht angemessen.

Bei mehreren Spuren hingegen sind "longitudinale" Abstände (bezüglich der Straßenrichtung) nahe Null sehr wohl möglich, da sie von zwei Fahrzeugen auf verschiedenen Spuren verursacht werden können. Für den Fußgänger ist es ja wichtig, dass *alle* Spuren frei sind. Zusätzlich spielt bei geringem Verkehrsaufkommen selbst auf einer Spur der Anteil des "gebundenen" Verkehrs, bei dem der Abstand durch den Sicherheitsabstand begrenzt wird, nur eine untergeordnete Rolle, so dass die Fahrzeuge im Wesentlichen unabhängig voneinander fahren. Die Zahl der in einem festen Zeitabschnitt registrierten unabhängigen und "seltenen" Ereignisse, hier die Zahl der vorbeikommenden Fahrzeuge, ist Poissonverteilt und damit die zeitlichen Abstände zwischen den Fahrzeugen exponentialverteilt.

(b) Sei die Zufallsgröße T gleich dem zeitlichen Abstand zwischen zwei, auf einer beliebigen Spur vorbeikommenden Fahrzeuge. Dann gilt

$$T \sim E^{(\lambda)} \tag{284}$$

wobei der Parameter λ aus dem maximalen Verkehrsaufkommen bestimmt wird:

$$\lambda = \frac{1}{\langle T \rangle} = 600/h = \frac{600/h}{3600s/h} = 1/6s \tag{285}$$

Beim zweiten "="Zeichen wurde ausgenutzt, dass das Verkehrsaufkommen (Fahrzeuge pro Zeiteinheit) gleich dem Inversen der mittleren Zeitlücke ist.

Die Wahrscheinlichkeit für eine sofortige Überquerungsmöglichkeit ("freie Strecke") ist gegeben durch die Wahrscheinlichkeit, dass T größer als die Überquerungszeit $x = t = 12s$ ist:

$$P_f = P(T > t) = 1 - P(T \leq t) = 1 - F_E(t) = e^{-\lambda t} = 1 - (1 - e^{-\lambda t}) = e^{-2} = 0.135 \tag{286}$$

Da die Exponentialverteilung die Abstände voneinander unabhängiger Ereignisse beschreibt, ist es völlig unerheblich, ob beim Ankommen an die Straße gerade ein Auto vorbeigefahren ist oder ob es schon lange frei war.

(c) Für die mittlere Zahl der bis zu einer Überquerungsmöglichkeit vorbeifahrender Fahrzeuge berechnet man zuerst die Wahrscheinlichkeiten P_n, dass genau $n = 0, 1, 2, \ldots$ Fahrzeuge vorbeifahren, d.h. die ersten n Lücken sind zu klein und die $(n+1)$-te ist groß genug: Mit dem Multiplikationssatz für unabhängige Ereignisse erhält man

$$P(n) = (P(T \leq t))^n P(T > t) = (1 - P_f)^n P_f \tag{287}$$

(Probe, dass $P(n)$ eine Verteilungsfunktion der Variablen n ist:
$\sum_{n=0}^{\infty} P(n) = P_f \sum_{n=0}^{\infty}(1 - P_f)^n = P_f/(1 - (1 - P_f)) = 1.$)

Die gesuchte mittlere Fahrzeugzahl ergibt sich aus dem Erwartungswert:

$$\langle n \rangle = \sum_{n=0}^{\infty} n P_n = P_f \sum_{n=0}^{\infty} n (1 - P_f)^n. \tag{288}$$

Nutzt man nun die angegebene Formel mit $q = 1 - P_f$ aus, erhält man

$$\langle n \rangle = P_f \frac{1 - P_f}{P_f^2} = \frac{1 - P_f}{P_f} = 6.39 \text{ Fahrzeuge} \tag{289}$$

(d) Die mittlere Warte*zeit* ist gleich der mittleren Fahrzeugzahl $\langle n \rangle$, multipliziert mit der mittleren Länge $\bar{T}_t$ der Zeitlücke unter der Bedingung, dass die Zeitlücke T kleiner als t ist (andernfalls würde der Fußgänger ja nicht das nächste Fahrzeug abwarten). Mit der Abstands-Dichtefunktion

$$f_E(t) = \frac{1}{\lambda} e^{-\lambda t} \tag{290}$$

gilt

$$\bar{T}_t = \frac{\int_{t'=0}^{t} t' f_E(t') dt}{\int\limits_{t'=0}^{t} f_E(t') dt} = \frac{1}{\lambda} - \frac{t P_f}{1 - P_f} \tag{291}$$

und damit für die mittlere Wartezeit $\langle T_w \rangle$ letztendlich

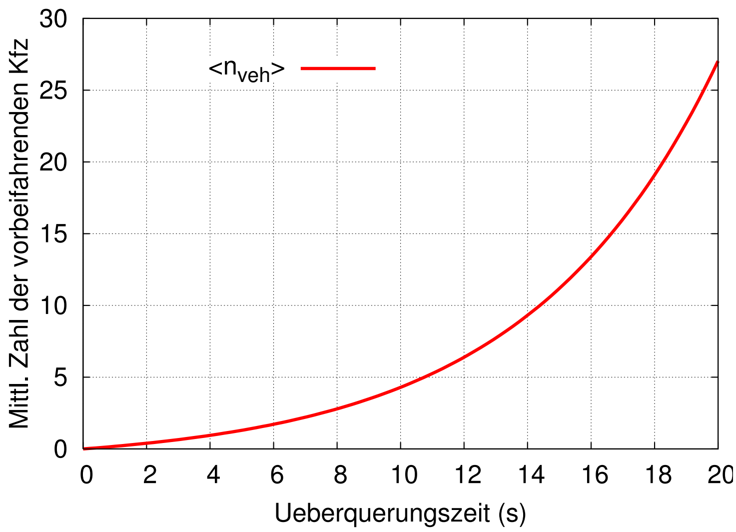
$$\langle T_w \rangle (t) = \langle n \rangle \bar{T}_t = \frac{1 - P_f}{P_f} \bar{T}_t = \frac{1 - P_f}{\lambda P_f} - t. \tag{292}$$

Zahlenwert mit $t = 12$ s, $1/\lambda = 6$ s und P_f von Aufgabenteil (b):

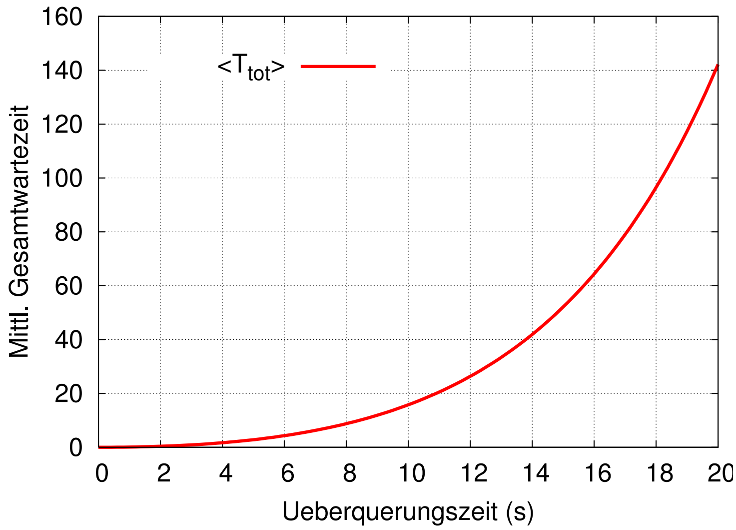
$$\langle T_w \rangle (12) = 26.3 \text{ s} \tag{293}$$

Da das Integral $\int x e^{-\lambda x} dx$ nicht angegeben war, gab es auch ohne Ausrechnen von $\bar{T}_t$, damit auch ohne Zahlenwerte, die volle Punktzahl!

Plots der Funktionen der mittleren Fahrzeugzahl und der mittleren Wartezeit als Funktion der Überquerungszeit (nicht verlangt, nur zur Veranschaulichung):



Mittl. Zahl der vorbeifahrenden Kfz
Ueberquerungszeit (s)
<n_veh>
0
5
10
15
20
25
30
0
2
4
6
8
10
12
14
16
18
20



Mittl. Gesamtwartezeit
Ueberquerungszeit (s)
<T_tot>
0
20
40
60
80
100
120
140
160
0
2
4
6
8
10
12
14
16
18
20

Lösung T.3.4

(a) Die Exponentialverteilung mit der Dichte

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{sonst} \end{cases} \tag{294}$$

hat einen einzigen Parameter λ. Es gilt $E(X) = \frac{1}{\lambda}$. Da nach Aufgabenstellung der Fluss Q durch "Fahrzeuge pro Zeiteinheit" definiert ist, ist $1/Q$ die Zeiteinheit pro Fahrzeug, also der mittlere zeitliche Abstand. Also

$$E(X) = \frac{1}{\lambda} = \frac{1}{Q} \quad \Rightarrow \quad \lambda = Q \tag{295}$$

(b) Da der zeitliche Abstand zweier Fahrzeuge nach Aufgabenstellung exponentialverteilt ist, ist die Fahrzeugzahl Y selbst poissonverteilt, $Y \sim \mathrm{Po}(\mu)$. Bei $Q = 1$ Fz/min. **Achtung!! fast alle haben in der Klausur mit einer Exponentialverteilung gerechnet!** Bei einem Intervall von 5 Minuten gilt also $E(Y) = \mu = 5$. Damit

$$P(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2) \tag{296}$$

$$= \frac{\mu^0 e^{-\mu}}{0!} + \frac{\mu^1 e^{-\mu}}{1!} + \frac{\mu^2 e^{-\mu}}{2!} = e^{-\mu}\left(1 + \mu + \frac{\mu^2}{2!}\right) = 12.5\%. \tag{297}$$

Median:

$$F(x) = 1 - e^{-\lambda x} \stackrel{!}{=} \frac{1}{2} \tag{298}$$

$$\frac{1}{2} = e^{-\lambda x} \tag{299}$$

$$\ln 2 = \lambda x \tag{300}$$

also

$$x_{0.5} = \frac{1}{\lambda} \ln 2 = \frac{1}{Q} \ln 2 = 0.693 \text{ min.} \tag{301}$$

(c) Es gilt für den zeitlichen Abstand $X \sim E(\lambda)$ mit $\lambda^{-1} = 1$ min $= 60$ s. Für den räumlichen Abstand$Z$ gilt bei konstanter Geschwindigkeit $V = 20$m/s und Rechnung in SI-Einheiten, also Metern und Sekunden:

$$Z = aX = 20X \tag{302}$$

Eine gleichförmige Skalierung um den Faktor a ändert die Verteilungsform (Exponentialverteilung) nicht, nur den Parameter, also $Z \sim E(\lambda_z)$ mit

$$E(Z) = E(aX) = aE(X) = \frac{a}{\lambda} \stackrel{!}{=} \frac{1}{\lambda_z} \quad \Rightarrow \quad \lambda_z = \frac{\lambda}{a} = \frac{1}{60 \text{ s} \cdot 20 \text{ m/s}} = \frac{1}{1\,200 \text{ m}} \tag{303}$$

(d) Mit der bereits oben erwähnten Dichte ist der wahrscheinlichste Abstand = Dichtemaximum = Null. Da auch die aggressivsten Raser nicht mit dem Abstand Null fahren, sondern einen minimalen "Sicherheits"-Abstand halten, gibt die Exponentialverteilung den Sachverhalt falsch wider.

(e) Ausrechnen von $E(X)$:

$$E(X) = \int_0^\infty x\, f(x) \mathrm{d}x \tag{304}$$

$$= \int_{x_{min}}^\infty \tilde{\lambda} x\, e^{-\tilde{\lambda}(x-x_{min})} \mathrm{d}x \tag{305}$$

$$\overset{x=y+x_{min}}{=} \int_0^\infty (x_{min}+y)\tilde{\lambda} e^{-\tilde{\lambda}(y+x_{min}-x_{min})} \mathrm{d}y \tag{306}$$

$$= x_{min} + \int_0^\infty \tilde{\lambda} y e^{-\tilde{\lambda} y} \mathrm{d}y \tag{307}$$

$$= x_{min} + \frac{1}{\tilde{\lambda}} \tag{308}$$

Das Integral in der vorletzten Zeile hat man durch Analogie gelöst, da es gleich dem Erwartungswert einer $E(\tilde{\lambda})$-Verteilung, also gleich $\frac{1}{\tilde{\lambda}}$ ist. Da nach wie vor $E(X) = \frac{1}{Q}$ erhält man durch Vergleich

$$\frac{1}{Q} = x_{min} + \frac{1}{\tilde{\lambda}} \quad \Rightarrow \quad \tilde{\lambda} = \frac{Q}{1+Qx_{min}}. \tag{309}$$

(f) (i) Schwacher Verkehr: $\lambda_1 = Q_1 = 1$, $\tilde{\lambda}_1 = 1.01695$: Kaum ein Unterschied.
Starker Verkehr: $\lambda_2 = Q_2 = 30$, $\tilde{\lambda}_2 = 60$: Unterwschied um Faktor zwei!

(ii) Median der ursprünglichen Verteilung (vgl. Teil (b)):

$$F(x) = 1 - e^{-\lambda x} = 0.5 \quad \Rightarrow \quad x_{0.5} = \frac{\ln 2}{\lambda} = \frac{\ln 2}{Q}. \tag{310}$$

Median der modifizierten Verteilung:

$$F(x) = 1 - e^{-\tilde{\lambda}(x-x_{min})} = 0.5 \quad \Rightarrow \quad \tilde{x}_{0.5} = x_{min} + \frac{\ln 2}{\tilde{\lambda}}. \tag{311}$$

Mit $x_{min} = 1$ s, Q_1 =1/(60 s), Q_2 =1/(2 s) ergibt sich

$$Q = Q_1 = 1/(60 \text{ s}) \Rightarrow \; x_{0.5} = \frac{\ln 2}{1/60s} = 41.6 \text{ s} = 0.693 \text{ min}, \quad \tilde{x}_{0.5} = 41.9 \text{ s} = 0.698 \text{ min} \tag{312}$$

$$Q = Q_2 = 1/(2 \text{ s}) \Rightarrow \; x_{0.5} = 1.39 \text{ s} = 0.0231 \text{ min}, \qquad \tilde{x}_{0.5} = 1.69 \text{ s} = 0.028 \text{ min} \tag{313}$$

also ebenfalls nur beim hohen Fluss ein deutlicher Unterschied.

Lösung T.3.5

(a) Sei X die Verspätung (in Minuten) des ersten Zuges der zweiten Verbindung.

- Die Züge fahren entweder pünktlich oder verspätet ab, nicht jedoch verfrüht, so dass $P(X < 0) = 0$, also F(x)=0 für $x < 0$.
- 62% aller Züge sind pünktlich: $p_0 = P(X = 0) = P(X \leq 0) = F(0) = 0.62$.
- Die Verspätungszeiten der restlichen $1 - p_0 = 38\%$ verspäteten Züge gehorchen einer Exponentialverteilung mit Parameter $\lambda = 1/10$ (da $E(x) = \mu = 1/\lambda \Rightarrow \lambda = 1/\mu$): $P(X < x|X > 0) = F(x|X > 0) = 1 - e^{-\lambda x}$. Damit gilt für die unbedingte Verteilungsfunktion $P(X < x) = F(x) = p_0 + (1 - p_0)\left(1 - e^{-\lambda x}\right) = p_0 + 1 - e^{-\lambda x} - p_0 + p_0 e^{-\lambda x} = 1 - (1 - p_0)e^{-\lambda x}$.

Damit insgesamt

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - (1 - p_0)e^{-\lambda x} & x \geq 0. \end{cases} \tag{314}$$

Die Wahrscheinlichkeit, den Anschluss zu verpassen, ist also gleich

$$p_v = P(X > 7) = 1 - P(x \leq 7) = (1 - p_0)e^{-7\cdot\frac{1}{10}} = 0.38 \cdot e^{-0.7} = 0.189. \tag{315}$$

(b) Die tatsächliche Reisezeit Y (in Minuten) setzt sich aus der Fahrplan-Reisezeit y_0 und der Verspätungszeit X zusammen (nur letztere ist eine Zufallsvariable!)

- Verbindung 1, Fahrplan-Reisezeit $y_{01} = 261$[3]: Die Zufallsvariable $Y_1 = y_{01} + X$ hat die Verteilungsfunktion

$$F_1(y) = F(y - y_{01}) = \begin{cases} 0 & y < y_{01} \\ 1 - (1 - p_0)e^{-\lambda(y - y_{01})} & y \geq y_{01}. \end{cases} \tag{316}$$

  mit $\lambda = 0.1$ und $p_0 = 62\%$.

- Verbindung 2: Da die Anschlusszüge nach Voraussetzung immer pünktlich abfahren, gibt es nur zwei mögliche Werte der Gesamtreisezeit Y_2: $Y_2 = y_{02} = 279$ mit der Wahrscheinlichkeit $1 - p_v = 0.811$ dafür, dass man den Anschluss erreicht, und $Y_2 = 800$ (die Differenz zwischen 8:15 am Folgetag und 18:55) mit der Verpass-Wahrscheinlichkeit $p_v = 0.189$. Insgesamt also

$$F_2(y) = \begin{cases} 0 & y < 279 \\ 1 - p_v & 279 \leq y < 800 \\ 1 & y \geq 800. \end{cases} \tag{317}$$

(c) Es sei nun X_1 die Verspätung des ersten Zuges und X_2 die Verspätung des Anschlusszuges, welche beide derselben Verteilung

$$F_{x1}(x) = F_{x2}(x) = F(x) = \begin{cases} 0 & x < 0 \\ 1 - (1 - p_0)e^{-\lambda x} & x \geq 0 \end{cases} \tag{318}$$

[3] $4h \cdot 60\ \mathrm{min}/h + 21\ \mathrm{min} = 261\ \mathrm{min}$

gehorchen. Der Anschluss wird erreicht, falls die Differenz $Z = X_1 - X_2$ kleiner oder gleich 7 (Minuten) ist. Dies ist mit der Wahrscheinlichkeit

$$p_a = P(Z \leq 7) = P(X_1 - X_2 \leq 7) \tag{319}$$

der Fall. In der Formelsammlung findet sich die Formel für die Dichtefunktion der *Summe* zweier unabhängiger Zufallsvariablen. Diese kann man anwenden, wenn man $Z = X_1 + (-X_2) = X_1 + X_3$ schreibt, wobei $f_3(x) = f_2(-x)$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_1(x) f_3(z - x) \mathrm{d}x = \int_{-\infty}^{\infty} f_1(x) f_2(x - z) \mathrm{d}x \tag{320}$$

und damit

$$p_a = P(Z \leq 7) = \int_{-\infty}^{7} f_Z(z) \mathrm{d}z. \tag{321}$$

Ausrechnen mit dieser Formel würde Differenzieren sowie zweimal Integrieren erfordern, ist aber nicht verlangt.

Lösung T.3.6

(a) Sei X die Schwangerschaftsdauer in Tagen. Da die Gaußverteilung $N(\mu, \sigma^2)$ symmetrisch ist, gilt für μ direkt die "Naegelesche Regel":

$$\mu = E(X) = 280. \tag{322}$$

Nach Angabe in der Aufgabenstellung gilt

$$P(279.5 < X \leq 280.5) = F(280.5) - F(279.5) = 0.04, \tag{323}$$

oder mit der Standardnormalverteilung $\Phi(z)$ mit $Z = (X - \mu)/\sigma$ und $F(-z) = 1 - F(z)$

$$0.04 = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1, \quad \text{bzw.} \quad \Phi(z) = 0.52. \tag{324}$$

Aus der Quantilstabelle ergibt dies $z_{0.52} = 0.0502$. Vergleich man dies mit der Definition $Z = (X - \mu)/\sigma$, ist also der Wert von σ gesucht, bei dem ein z-Wert von 0.0502 einen halben Tag Abweichung vom Mittelwert μ entspricht, also

$$\sigma = \frac{X - \mu}{z} = \frac{0.5}{0.0502} = 10.0. \tag{325}$$

(b) Mit $\sigma = 10$ (Tagen) entspricht einer 3 Wochen zu frühen Geburt einem z-Wert von $z_{Fg} = \frac{X-\mu}{\sigma} = \frac{(280-21)-280}{10} = -2.1$. Damit

$$P(\text{Frühgeburt}) = P(Z < z_{Fg}) = \Phi(-2.1) = 1 - \Phi(2.1) = 1 - 0.9821 = 1.79\%. \tag{326}$$

(c) Da das Kind zum errechneten Termin, entsprechend $Z = 0$, noch nicht da ist, muss offensichtlich $Z > 0$ sein. Da es nach Angabe keine Anzeichen einer unmittelbare Geburt gibt, bleibt die Form der Verteilung für $z > 0$ unverändert, aber natürlich können keine negativen Werte vorkommen, es gilt also

$$P(X \leq 0) := G(0) = 0. \tag{327}$$

Wegen der unveränderten Form wird daher die "linke" Seite der Dichtefunktion nach rechts "umgeklappt" und damit die Verteilungsfunktion $G(x)$ für X unter der Bedingung $X > 0$ gegeben durch

$$G(X) = 2\Phi(z) - 1 = 2\Phi\left(\frac{X - \mu}{\sigma}\right) - 1 \quad \text{für} \quad x \geq 0. \tag{328}$$

Zwei Wochen nach dem errechneten Termin, also für $z = (14 - 0)/10$, gilt daher

$$P(X \leq 280 + 14 | X \geq 280) = 2\Phi(1.4) - 1 = 2 \cdot 0.9192 - 1 = 83.8\%. \tag{329}$$

Das Kind wird also nur mit einer Wahrscheinlichkeit von 83.8 % in den nächsten zwei Wochen nach dem Naegeleschen Termin geboren, die Verwandten kommen also mit einer Wahrscheinlichkeit von 16.2% umsonst. Für die *unbedingte* Wahrscheinlichkeit gilt übrigens $P(X \leq 280 + 14) = \Phi(1.4) = 91.9\%$.

Lösung T.3.7

(a) Verteilungsfunktion für die Wartezeit x in Sekunden bei zufälliger Ankunft: Die Verteilungsfunktion $F(x)$ ist die Wahrscheinlichkeit dafür, dass die Wartezeit kleiner oder gleich x Sekunden beträgt.

- Offensichtlich kann x nur ≥ 0 sein. Deshalb $F(x) = 0$ für $x < 0$
- Mit der Wahrscheinlichkeit $p_0 = 30/90 = 1/3$ erreicht der Fußgänger den Übergang bei grüner Ampel, muss also nicht warten ($x = 0$). Also $F(0) = 1/3$.
- Die maximale Wartezeit beträgt 60 Sekunden, deshalb $F(x) = 1$ für $x > 60$.
- Falls der Fußgänger warten muss, ist wegen der zufälligen Ankunft die Wartezeit zwischen 0 und 60 Sekunden gleichverteilt mit der Dichte $f(x) = 2/3 \cdot 1/60$ (die 2/3 kommen daher, da der Fußgänger nur mit dieser Wahrscheinlichkeit überhaupt warten muss, also $F(60) - F(0) = 2/3$). Damit ist für $0 \leq x \leq 60$ die Verteilungsfunktion gegeben durch

$$F(x) = \frac{1}{3} + \frac{x}{90}. \tag{330}$$

(b) Zu betrachten: $2 \cdot 200 = 400$ Überquerungen. Bedingungen des Zentralen Grenzwertsatzes (ZGWS) (nicht verlangt!)

- Unabhängig: Da es sich um einen reinen Fußweg handelt, hängt die Ampelphase bei Erreichen nicht von irgendwelchen Synchronisationen ("Grüne Welle") ab. Abgesehen von dem Fall, dass der Fußgänger die Ampel an jedem Tag minutengenau zur selben Uhrzeit erreicht, ist auch eine mögliche Kopplung an die Uhrzeit (welche ihrerseits die Ampelphase bestimmen könnte) ausgeschlossen. Die 400 Zufallsereignisse "Wartezeit X_i bei Überquerung i" sind also unabhängig.
- Endliche Varianz. Da die Varianz nie größer als die Spannweite R zum Quadrat sein kann, ist mit $R = 60$ diese Bedingung erfüllt.
- Die größte Einzelvarianz darf höchstens 1/30 der Gesamtvarianz sein. Dies ist bei 400 identisch verteilten Summanden (i.i.d, independent, identically distributed variables) mit Beiträgen von jeweils 1/400 der Gesamtvarianz der Fall.

Der ZGWS sagt aus, dass die Gesamtwartezeit

$$Y = \sum_{i=1}^{400} X_i \sim N(\mu_y, \sigma_y^2) \tag{331}$$

mit

$$\mu = 400\mu_x, \quad \sigma_y^2 = 400\sigma_x^2. \tag{332}$$

Berechnung von Erwartungswert μ_x und Varianz σ_x^2 der Zufallsvariablen X_i: Hier muss man berücksichtigen, dass X sowohl diskrete als auch stetige Anteile hat:

- Mit der Wahrscheinlichkeit $p_0 = 1/3$ ist $X = 0$,
- Für $0 < x \leq 60$ ist X gleichverteilt mit der Dichte $f(x) = 1/90$.

Erwartungswert:

$$\mu_x = p_0 \cdot 0 + \int_0^{60} x f(x) \mathrm{d}x = \left[\frac{1}{2} \cdot \frac{1}{90} x^2\right]_0^{60} = \left[\frac{x^2}{180}\right]_0^{60} = \frac{1}{180} \cdot 60^2 - \frac{1}{180} \cdot 0^2 = 20. \tag{333}$$

Alternativ anschaulich: Mit Wahrscheinlichkeit $1 - p_0 = 2/3$ muss gewartet werden, und zwar im Mittel 30 Sekunden, also $\mu_x = \frac{2}{3} \cdot 30 = 20$.

Varianz:

$$\sigma_x^2 = \int\limits_0^{60} x^2 f(x)\mathrm{d}x - \mu_x^2 = \left[\frac{1}{3} \cdot \frac{1}{90} x^3\right]_0^{60} - 20^2 = \left[\frac{x^3}{270}\right]_0^{60} - 400 = \frac{21\,600}{270} - 400 = 400. \quad (334)$$

Damit ist die Gesamtwartezeit

$$Y \sim N(20 \cdot 400, 400 \cdot 400) \Rightarrow Y \sim N(8000, 400^2). \quad (335)$$

(c) Wahrscheinlichkeit für eine Wartezeit größer 2 h = $2h \cdot 3600$s/h $= 7200s$ ($Y > 7\,200$):

$$P(Y > 7\,200) = 1 - F(7\,200) = 1 - \Phi\left(\frac{7\,200 - 8\,000}{400}\right) \quad (336)$$

$$= 1 - \Phi(-2) = 1 - (1 - \Phi(2)) = \Phi(2) = 0.9772. \quad (337)$$

Wahrscheinlichkeit für eine Wartezeit kleiner 2 1/2 h ($Y < 9\,000$):

$$P(Y < 9\,000) = F(9\,000) = \Phi\left(\frac{9\,000 - 8\,000}{400}\right) = \Phi(2.5) = 0.9938. \quad (338)$$

Lösung T.3.8

(a)

$$P(X > 100) = 1 - P(X \leq 100) = 1 - F(100) = 100^{-\alpha} = 0.631\% \tag{339}$$

(b) Gesucht ist das 99.9% Quantil $x_{0.999}$. Die Gleichung für das Quantil x_q lautet:

$$P(X < x_q) = F(x_q) = 1 - x_q^{-\alpha} = q \quad \overset{q=0.999}{\Rightarrow} \quad x_q = (1-q)^{-1/\alpha} = 534. \tag{340}$$

(c) Mittlerer Umsatz pro Buch:

$$\bar{x} = \int_1^\infty x f(x)\mathrm{d}x = \int_1^\infty x F'(x)\mathrm{d}x \tag{341}$$

$$= \alpha \int_1^\infty x x^{-\alpha-1}\mathrm{d}x = \alpha \int_1^\infty x^{-\alpha}\mathrm{d}x \tag{342}$$

$$= \frac{\alpha}{1-\alpha}\left[x^{1-\alpha}\right]_1^\infty = \frac{\alpha}{\alpha-1} = 11. \tag{343}$$

(d) Das Integral ist wie das des Mittelwertes, nur mit variabler oberer Grenze:

$$P(x) = \left(\int_1^x x' f(x')\mathrm{d}x'\right) \Big/ \left(\int_1^\infty x' f(x')\mathrm{d}x'\right) \tag{344}$$

$$= \left[x'^{1-\alpha}\right]_1^x \Big/ \left[x'^{1-\alpha}\right]_1^\infty = 1 - x^{1-\alpha}. \tag{345}$$

$P(x)$ gibt den Anteil des Gesamtumsatzes an, welcher von Büchern mit Einzelumsatz kleiner oder gleich x (aber nach Voraussetzung ≥ 1) erzeugt wird.

(e) Da $P(x)$ den Umsatzanteil aller Bücher mit Einzelumsatz zwischen 1 und x angibt, ist der Anteil hier durch

$$A = P(100) = 1 - 100^{1-\alpha} = 1 - 100^{-0.1} = 36.9\%. \tag{346}$$

gegeben (Ohne die explizite untere Grenze könnte man auch $A = P(100) - P(1)$ berechnen, mit demselben Ergebnis)

Hinweis: Falls Sie Teil (d) nicht gelöst haben, verwenden Sie $P(x) = 1 - x^{1-\alpha}$

4 Induktive Statistik

Lösung T.4.1

(a) **Schätzer:** Es sei X die Geschwindigkeit in Metern pro Sekunde:

$$\hat{\mu} = \bar{x} = \frac{1}{16}\sum_{i=1}^{16} x_i = 1.526, \tag{347}$$

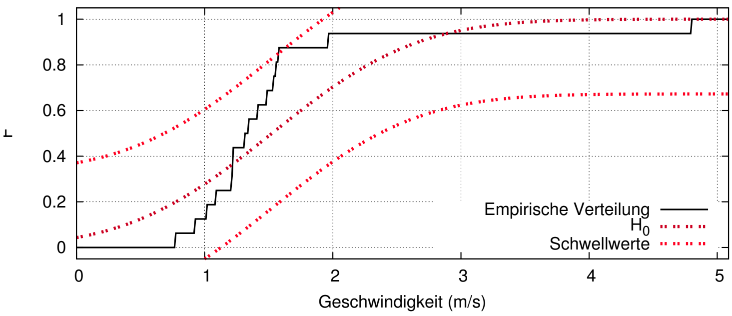
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{15}\sum_{i=1}^{16}(x_i-\bar{x})^2 = 0.846, \quad \hat{\sigma} = 0.920. \tag{348}$$

(b) **Kolmo-Test:** Die Realisierung d der Test-Statistik D des Kolmogorow-Smirnow-Tests ist durch den maximalen vertikalen Abstand $\max_x(|F(x) - F_0(x)|)$ der empirischen Verteilung bezüglich der H_0-Verteilung gegeben. Aus der Abbildung der Aufgabenstellung liest man ab: z.B. bei $F(x) = 0.9$ und $F_0(x) = 0.55$, damit ergibt sich $d = 0.35$ ($d = 0.4$ würde auch gehen, ebenfalls $d = 0.3$ mit falscher Schlussfolgerung).

Das $1-\alpha$ Quantil $d_{n,1-\alpha}$ der D-Statistik ist für $\alpha = 5\%$ und $n = 16$ nach der Vorlesung angenähert gegeben durch

$$d_{0.95,16} = \frac{c(0.05)}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}} = \frac{1.358}{\sqrt{16} + 0.12 + 0.11/\sqrt{16}} = 0.33. \tag{349}$$

Also kann H_0 abgelehnt werden, wenn auch knapp,[4] siehe Abbildung (die nicht verlangt war).

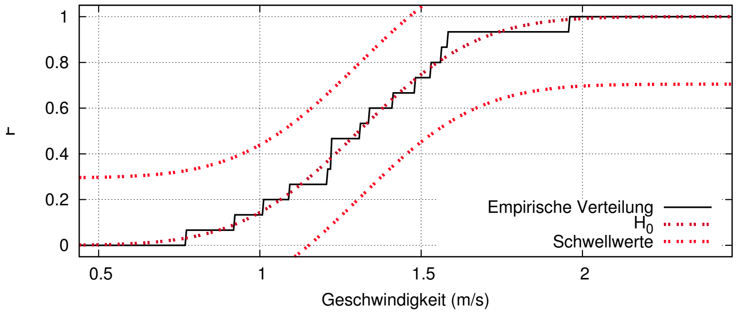


(c) **Ausreißer** Der Fußgänger mit $x = 4.8$ m/s (etwa 17 km/h) war ein Jogger (eher sogar ein veritabler Läufer). Da die Bemessung von Ampel-Freigabezeiten sich nach den langsamsten Fußgängern richten sollte, ist der Läufer offensichtlich irrelevant und sollte, ja muss aus den Daten eliminiert werden.[5]

Nach Eliminierung bekommt man folgenden Kolmo-Test (nicht verlangt):

[4] Falls man bei einer falschen Schätzung von d eine andere Schlussfolgerung bekommt, die Schlussfolgerung aber an sich logisch richtig ist, ergibt das volle Punktzahl.

[5] Ironischerweise würde die Berücksichtigung des Joggers sogar zu einer längeren, wenn nicht gar negativen Mindestfreigabezeit führen.



(d) **Perzentile der Geschwindigkeiten:** Hierzu benötigt man die Quantilstabelle der Standardnormalverteilung. Hier ist bei Gültigkeit von H_0 $Z = (X - \hat{\mu})/\hat{\sigma} \sim N(0,1)$ standardnormalverteilt, also

$$x_{0.9} = \hat{\mu} + \hat{\sigma} z_{0.9} \overset{z_{0.9}=1.2816}{=} 1.31\ \frac{m}{s} + 0.289\ \frac{m}{s} = 1.680, \tag{350}$$

$$x_{0.99} = \hat{\mu} + \hat{\sigma} z_{0.99} \overset{z_{0.9}=2.3263}{=} 1.982. \tag{351}$$

(e) **Perzentile der Überquerungszeiten:**2 Die Überquerungszeit $Y = b/X$ mit $b = 10$ m ist *micht* normalverteilt: Dies gilt zwar für Summen, Differenzen und Linearkombinationen, aber *nicht* für nichtlineare Funktionen (beispielsweise ist $Z^2 \sim \chi^2(1)$ verteilt). Für auf Quantile basierte Untersuchungen ist dies jedoch egal. Da die Funktion $Y(X)$ streng monoton fallend ist, kehren sich die Quantile allerdings um: Die größte Geschwindigkeit führt zur kleinsten Überquerungszeit, also

$$y_q = \frac{b}{x_{1-q}} = \frac{b}{\hat{\mu} + \hat{\sigma} z_{1-q}} = \frac{b}{\hat{\mu} - \hat{\sigma} z_q} \tag{352}$$

Die letzte Umformung ist nötig, da nur Quantile $q \geq 0.5$ in der Quantilstabelle aufgeführt sind. Also hier (Überquerungszeit-Quantile y_q in Sekunden):

$$y_{0.9} = \frac{b}{\hat{\mu} - \hat{\sigma} z_{0.9}} = \frac{10\ \text{m}}{1.31\ \frac{m}{s} - 0.289\ \frac{m}{s} \cdot 1.2816} = 10.64, \tag{353}$$

$$y_{0.99} = \frac{b}{\hat{\mu} - \hat{\sigma} z_{0.99}} = 15.68. \tag{354}$$

Lösung T.4.2

(a) **München – Hamburg**:

$$\hat{\mu} = \bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = 710.8, \quad \hat{\sigma} = s = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_i - \bar{x})^2} = 101.4 \tag{355}$$

Hof–Plauen:

$$\hat{\mu} = \bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = 220.8, \quad \hat{\sigma} = s = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_i - \bar{x})^2} = 46.6 \tag{356}$$

(b) Wegen des unbekannten wahren Mittelwerts muss mit den Quantilen der t-Verteilung anstelle der (implizit der Fahrgastzahl zugrundegelegten) Gaußverteilung gerechnet werden. Die Gesamtzahl m_{billig} der Billigtickets wird so bestimmt, dass beim 90. Perzentil $x_{0.9}$ der regulären Ticketzahl die Fahrgastzahl

$$Y = X + m_{billig} \tag{357}$$

gleich der Sitzplatzzahl m des Zuges ist, also

$$x_{0.9} + m_{billig} = m. \tag{358}$$

Das 90. Perzentil $x_{0.9}$ von x lässt sich durch das entsprechende Quantil $t_{0.9}^{(5)}$ der Student-t-Verteilung $(X - \bar{x})/S \sim T(5)$ ausdrücken:

$$\bar{x} + s\ t_{0.90}^{(5)} + m_{billig} = m \tag{359}$$

und damit schließlich die Zahl der anzubietenden Billig-Fahrkarten:

$$m_{billig} = m - \bar{x} - s\ t_{0.90}^{(5)} = \begin{cases} 950 - 710.8 - 101.4 \cdot 1.476 = 89 & \text{München – Hamburg} \\ 360 & \text{Hof – Plauen} \end{cases} \tag{360}$$

(Hier wurden zur Sicherheit die Billigticket-Zahlen immer auf ganze Werte *ab*gerundet.)

(c) Wegen der hinreichend großen Umfänge $n = 52 > 30$ und $m = 36 > 30$ der Stichproben vor und nach Einführung der Billigtickets kann hier mit der Gauß-Statistik gerechnet werden. Da es sich um nicht verbundene Stichproben handelt (unterschiedliche statistische Einheiten in den beiden Stichproben), kommt der entsprechende Test unverbundener Stichproben zum Einsatz. Die Test-Statistik lautet

$$Z = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \tag{361}$$

$$= \frac{\hat{\mu}_e - \hat{\mu}_a}{\sqrt{\frac{\hat{\sigma}_a^2}{n} + \frac{\hat{\sigma}_e^2}{m}}} \sim N(0,1). \tag{362}$$

Die Realisierungen für die beiden Verbindungen mit den in der Aufgabenstellung angegebenen Zahlenwerten für die beiden Schätzer von μ und σ^2 sind

$$z = \begin{cases} \frac{850-890}{\sqrt{\frac{110^2}{52}+\frac{70^2}{36}}} = -2.08 & \text{München – Hamburg} \\ 3.50 & \text{Hof – Plauen} \end{cases} \tag{363}$$

München – Hamburg: Die Abnahme ist zur Fehlerwahrscheinlichkeit α signifikant, falls $z \leq -z_{1-\alpha} = z_\alpha$ oder, unter Berücksichtigung, dass die Quantilsfunktion die Umkehrfunktion der Verteilungsfunktion ist:

$$\Phi(z) \leq \alpha \tag{364}$$

bzw. mit der Tabelle der Standardnormalverteilung:

$$\alpha \geq \Phi(z) = 1 - \Phi(-z) \approx 1 - \Phi(2.1) = 1 - 0.9821 = 0.018 = 1.8\% \tag{365}$$

Hof–Plauen: Die Zunahme ist zur Fehlerwahrscheinlichkeit α signifikant, falls $z \geq z_{1-\alpha}$, also

$$\Phi(z) \geq 1 - \alpha \tag{366}$$

bzw. mit der Tabelle der Standardnormalverteilung:

$$\alpha \geq 1 - \Phi(z) \approx 1 - \Phi(3.5) = 0.0002 = 0.02\% \tag{367}$$

Alternative Lösung:

Den Wert z der Gauß-Statistik direkt mit der Maschinerie der Tests auf Ungleichheit behandeln.

München – Hamburg: Man kann Hypothesen nur verwerfen, nicht aber annehmen, deshalb muss man das Gegenteil als Nullhypothese annehmen, also hier, wo man eigentlich wissen will, ob die Fahrgastzahl signifikant abgenommen hat ($\mu_e < \mu_a$), der Test auf "$\geq$":

- H_0: $\mu_e \geq \mu_a$
- Test-Statistik nach Gl. 361
- Realisierung aus den Stichproben: $z = -2.08$
- Entscheidung bei $\alpha = 1\%$: Nullhypothese ablehnbar, falls $z < z_\alpha = -z_{1-\alpha} = -2.33$. Dies ist nicht erfüllt, also ist H_0 *nicht* ablehnbar.

Hof – Plauen:

- H_0: $\mu_e \leq \mu_a$ (genau andersherum als bei München – Hamburg!)
- Test-Statistik nach Gl. 361
- Realisierung aus den Stichproben: $z = +3.50$
- Entscheidung bei $\alpha = 1\%$: Nullhypothese ablehnbar, falls $z > z_{1-\alpha} = 2.33$. Dies ist erfüllt, also ist H_0 ablehnbar und damit das Gegenteil davon, also eine Abnahme der Fahrgastzahlen, signifikant.

Lösung T.4.3

(a) Erwartungswert-Schätzung bei unbekannter Varianz ⇒ t-Statistik.

Mittelwertschätzer:

$$\bar{x} = \frac{1}{4}\sum_i x_i = 97.75 \tag{368}$$

Varianzschätzer:

$$s_x^2 = \frac{1}{4-1}\sum_i (x_i - \bar{x})^2 = 2.417. \tag{369}$$

Quantil der Student-Statistik aus der beigelegten Tabelle:

$$t_{1-\alpha/2}^{(n-1)} = t_{0.975}^{(3)} = 3.182 \tag{370}$$

Das Konfidenzintervall wird wie folgt berechnet: $\hat{\mu} \in \bar{x} \pm \frac{t_{1-\alpha/2}^{(n-1)} \cdot s}{\sqrt{n}}$ Halbe Breite des Konfidenzintervalls ist also:

$$\Delta x_\alpha = \frac{s\, t_{1-\alpha/2}^{(n-1)}}{\sqrt{n}} = \frac{3.182 \cdot \sqrt{2.417}}{\sqrt{4}} = 2.473 \tag{371}$$

Konfidenzintervall:

$$K_\alpha = [\bar{x} - \Delta x_\alpha, \bar{x} + \Delta x_\alpha] = [95.3, 100.2]. \tag{372}$$

(b) Statistischer Test auf Ungleichheit.

a) Nullhypothese H_0: $\mu \geq \mu_0 = 100$ ("Im Zweifel für den Angeklagten")

b) Test-Statistik: $T = \sqrt{n}\frac{\bar{X} - \mu_0}{s} \sim T(n-1) = T(3)$

c) Realisierung aus der Stichprobe: $t = 2\frac{97.75 - 100}{1.554} = -2.895$

d) Testentscheidung bei $\alpha = 5\%$: H_0 abgelehnt, falls $t < -t_{1-\alpha}^{(n-1)} = t_{0.95}^{(3)} = -2.353 \Rightarrow H_0$ abgelehnt.

(c) Das Konfidenzintervall ist nur zu einem symmetrischen Test äquivalent. Der asymmetrische Test auf Ungleichheit ist jedoch schärfer, da die Fehlerwahrscheinlichkeit ganz für eine Seite zur Verfügung steht, also kein Widerspruch. (Konsistent dazu würde man einen zweiseitigen Test nicht ablehnen können, da $|t| < t_{0.975}^{(3)}$)

(d) Hier muss das Kriterium der Ablehnung grenzwertig erfüllt sein, also

$$t = -t_{1-\alpha}^{(3)} \quad \text{bzw.} \quad -t = t_{1-\alpha}^{(3)} \tag{373}$$

Um daraus α zu erhalten, wird auf beiden Seiten die Umkehrfunktion der Quantilsfunktion, also die (kumulierte) Verteilungsfunktion $F^{(3)}$ der Student-3-Verteilung selbst, angewandt:

$$F^{(3)}(-t) = 1 - \alpha \quad \Rightarrow \alpha = 1 - F^{(3)}(-t) \tag{374}$$

Aus der Grafik der Aufgabenstellung ergibt sich für die Kurve "$n = 3$" mit dem Wert $t = -2.8$ der Aufgabenstellung der Wert

$$1 - F^{(3)}(-t) = 1 - F^{(3)}(2.8) = 1 - 0.97 = 0.03 \;\Rightarrow\; \alpha = 3\% \tag{375}$$

Für den wahren Wert $t = -2.89$ ist $1 - F^{(3)}(-t)$ und damit die Grenze für α etwas geringer, etwa 2.7% (jeder Wert zwischen 2% und 3.5% gibt volle Punktzahl)

(e) Grenzwert $\bar{x}_c$ für den empirischen Mittelwert, bei dem die Nullhypothese $\mu \geq \mu_0 = 100$ gerade noch bei $\alpha = 5\%$ angenommen wird:

$$\bar{x}_c = \mu_0 - t_{0.95}^{(3)} \frac{s}{\sqrt{n}} = 100 - \frac{2.353 \cdot \sqrt{2.417}}{\sqrt{4}} = 98.2 \tag{376}$$

Da der wahre Mittelwert $\mu = 98$, gehorcht in Wirklichkeit aber nicht $\sqrt{n}(\bar{X} - \mu_0)/S$ sondern $\sqrt{n}(\bar{X} - \mu)/S$ der Student-Verteilung mit drei Freiheitsgraden. Die Nullhypothese wird fälschlicherweise nicht abgelehnt, falls $\bar{X} > \bar{x}_c$, also

$$T = \sqrt{n}\frac{\bar{X} - \mu}{S} > t_c = \sqrt{n}\frac{\bar{x}_c - \mu}{s} = \sqrt{4}\frac{98.171 - 98}{\sqrt{2.417}} = 0.22 \tag{377}$$

Die Wahrscheinlichkeit dafür ist (Ablesung wieder aus der Grafik der Aufgabenstellung)

$$\beta = P(T > t_c) = 1 - P(T < t_c) = 1 - F^{(3)}(0.22) = 1 - 0.6 = 40\,\%. \tag{378}$$

Also beträgt der Fehler zweiter Art in diesem konkreten Beispiel etwa 40%. (Im Extremfall, falls μ nur minimal kleiner als μ_0 ist, kann der Fehler zweiter Art bis zum Wert $1 - \alpha = 95\%$ steigen!)

Lösung T.4.4

(a) Nimmt man als Einheit ein Winkelgrad, ergeben sich die Varianzen der Messgrößen X_1, X_2 und X_3 aus den Standardabweichungen:

$$\sigma_1^2 = \sigma_2^2 = 4, \quad \sigma_3^2 = 16. \tag{379}$$

Wie groß ist die Standardabweichung des arithmetischen Mittels? Berechnung unter Ausnutzung der Unabhängigkeit über die Varianz:

$$\bar{X} = \frac{1}{3}\sum_{i=1}^{3} X_i, \quad V(\bar{X}) = \frac{1}{9}\sum_{i=1}^{3} V(X_i) = \frac{24}{9} = \frac{8}{3} \Rightarrow \sigma_{\bar{x}} = \sqrt{V(\bar{X})} = 1.63. \tag{380}$$

Denn es gilt für $V(\bar{X})$, wenn für $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ eingesetzt wird

$$\begin{aligned}
V(\bar{X}) &= V(\frac{1}{n}\sum_i X_i) && (381)\\
&= \frac{1}{n^2} V(\sum_i X_i) && (382)\\
&= \frac{1}{n^2}\sum_i V(X_i) && (383)\\
&= \frac{1}{n^2}\sum_i \sigma^2 && (384)\\
&= \frac{1}{3^3}(4+4+16) && (385)\\
&= \frac{24}{9} && (386)
\end{aligned}$$

Dabei wurden die Rechenregel $V(aX) = a^2 V(X)$ und die für unabhängige X_i gültige Rechenregel $V(X_1 + X_2) = V(X_1) + V(X_2)$ aus der Formelsammlung angewandt.

(b) Hier gilt es, die Varianz eines gewichteten Mittels $w_1X_1 + w_2X_2 + w_3X_3$ der Sensordaten zu berechnen. Mit $w_1 + w_2 + w_3 = 1$ und $w_1 = w_2$ (Bedingungen aus der Aufgabenstellung) kann man mit $w_1 = w_2 = w$ und $w_3 = 1 - 2w$ schreiben

$$\bar{X}(w) = w(X_1) + X_2) + (1 - 2w)X_3. \tag{387}$$

Offensichtlich gilt für beliebige $w \in [0, 1]$, dass der Schätzer $\bar{X}(w)$ erwartungstreu bezüglich des wahren Winkels $E(X) = 30$ ist (nicht verlangt). Die Varianz verechnet sich wieder mit den Rechenregeln für unabhängige Zufallsgrößen aus der Formelsammlung:

$$V(\bar{X}(w)) = w^2(\sigma_1^2 + \sigma_2^2) + (1 - 2w)^2\sigma_3^2 = 8w^2 + 16(1 - 2w)^2. \tag{388}$$

Nun Minimieren bezüglich w durch Ableiten und Nullsetzen: *Achtung: Kettenregel!!*

$$V'(w) = 16w - 64(1 - 2w) \stackrel{!}{=} 0 \Rightarrow w_{opt} = \frac{4}{9}. \tag{389}$$

Damit

$$(w_1)_{opt} = (w_2)_{opt} = \frac{4}{9}, \quad (w_3)_{opt} = 1 - 2w = 1 - 2 \cdot \frac{4}{9} = \frac{1}{9}. \tag{390}$$

Also werden die beiden “besseren” Messergebnisse X_1 und X_2 vierfach bezüglich der “schlechteren” Messung 3 gewichtet.

Der Wert der Varianz beim effektiven (varianzminimalen) Schätzer $\bar{X}(w_{opt})$ ist gegeben durch

$$V(w_{opt}) = (\frac{4}{9})^2 \cdot 4 + (\frac{4}{9})^2 \cdot 4 + (\frac{1}{9})^2 \cdot 16 = \frac{16}{81} \cdot 8 + \frac{1}{81} \cdot 16 = \frac{16}{9} \tag{391}$$

und damit

$$\sigma(w_{opt}) = \sqrt{\frac{16}{9}} = \frac{4}{3} = 1.333. \tag{392}$$

Die Standardabweichung reduziert sich also bei optimaler Gewichtung der drei Messergebnisse von 1.63 auf 1.33 Winkelgrade.

Lösung T.4.5

(a) Schätzung von Erwartungswert μ und Varianz σ^2 aus der Stichprobe:

$$\hat{\mu} = \bar{X} = \frac{1}{4}\sum_{i=1}^{4} X_i = 90, \quad \hat{\sigma^2} = \frac{1}{3}\sum_{i=1}^{4}(X_i - \bar{X})^2 = 462.7. \tag{393}$$

Dabei ist bei der Stichprobenvarianz der Nenner $n-1=3$ wichtig!

(b) Bei 200 Einladungen ergeben sich mit den Angaben der Aufgabenstellung für den wahren Mittelwerte und die wahre Standardabweichung der Zahl der Absagen

$$\mu = 200 \cdot 0.45 = 90, \quad \sigma = 200 \cdot 0.10 = 20. \tag{394}$$

Damit ist die Größe

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1) \tag{395}$$

standardnormalverteilt. Damit

$$P(X < 80) = P\left(Z < \frac{80-90}{20}\right) \tag{396}$$

$$= P(Z < -\frac{1}{2}) \tag{397}$$

$$= \Phi(-\frac{1}{2} = 1 - \Phi(\frac{1}{2}) = 1 - 0.691 = 0.309. \tag{398}$$

(c) Im Gegensatz zu (b) werden hier keine bekannten Parameter μ und σ^2 angenommen, sondern diese müssen beide aus der Stichprobe geschätzt werden. Damit ist die t-Statistik und nicht die Gauß-Statistik für die Konfidenzintervalle $I_{0.05}$ der Absagen-Zahl zur Fehlerwahrscheinlichkeit $\alpha = 0.05$ relevant:

$$\bar{\mu} \in \bar{X} \pm \frac{t^{(3)}_{1-\alpha/2} \cdot \hat{\sigma}}{\sqrt{n}} \tag{399}$$

$$I_{0.05} = \left[\bar{X} - t^{(3)}_{0.975}\hat{\sigma}/\sqrt{n}, \bar{X} + t^{(3)}_{0.975}\hat{\sigma}/\sqrt{n}\right] \tag{400}$$

Mit

$$t^{(3)}_{0.975}\hat{\sigma}/\sqrt{n} = 3.182 \cdot \sqrt{462.7}/\sqrt{4} = 34.2 \tag{401}$$

ergibt sich das Konfidenzintervall

$$I_{0.05} = [55.8, 124.2]. \tag{402}$$

Das Konfidenzintervall für die Absagequote ergibt sich durch Division mit der Zahl $n = 200$ der Einladungen:

$$I_{\text{quote},0.05} = [0.279, 0.621]. \tag{403}$$

(d)
- Geschätzte Gesamtvarianz der Absagerzahl: $\hat{\sigma^2} = 463$.
- Varianz bei konstanter Attraktivität bzw. konstanter Ablehnquote $\theta = 45\%$:

$$\sigma^2_{(i)} = 200\theta(1-\theta) = 49.5 \tag{404}$$

Dies ist deutlich kleiner als die Gesamtvarianz, so dass der Großteil der Varianz nicht durch statistische Schwankungen bei konstanter mittlere Ablehnquote zustande kommt, sondern durch Schwankungen der Attraktivität, d.h. der wahren Ablehnquote, über die Jahre.

(e) Bei 350 Einladungen und 200 Plätzen droht Überfüllung, falls die Ablehnquote unterhalb von

$$\theta_c = \frac{350 - 200}{350} = \frac{150}{350} = \frac{3}{7} = 42.9\% \tag{405}$$

liegt.

Nach Aufgabenstellung bleibt der Erwartungswert und die Varianz der Ablehn*quote* und nicht der *Zahl* der Ablehnungen unabhängig von der Zahl der Einladungen. Es ist aus der Aufgabenstellung nicht klar, ob man für die Verteilung der Quoten (i) die Annahme bei (b) oder (ii) die Schätzer von (a) zugrundelegen soll. Deshalb werden sowohl die Variante (i) als auch (ii) voll gewertet.

(i) Mit den Annahmen von Aufgabenteil (b) ergibt sich für die Ablehnquoten θ ein bekannter Erwartungswert von $\theta_0 = \frac{\mu_0}{200} = 0.45$ und eine Standardabweichung von $\sigma_\theta = 0.1$ und damit bekannte Varianz $\sigma_\theta^2 = 0.01$. Damit ist

$$Z = \frac{\theta - \theta_0}{\sigma_\theta} = \frac{\theta - \theta_0}{0.1} \sim N(0,1) \tag{406}$$

standardnormalverteilt und es gilt

$$P(\theta < \theta_c) = P\left(Z > \frac{\theta_c - \theta_0}{\sigma_\theta}\right) \tag{407}$$

$$= P(Z < \frac{(0.42 - 0.45)}{0.1}) = P(Z < -0.3) = 1 - \Phi(0.3) = 1 - 0.6179 = 0.382. \tag{408}$$

(ii) Mittelwert und Varianz geschätzt wie in Teil (a). Dann wird Z durch die Zufallsvariable $T_3 \sim T(3)$ (t-Statistik mit 3 Freiheitsgraden) ersetzt und anstelle $P\left(Z > (\theta_c - \theta_0)/0.1\right)$ hat man die Wahrscheinlichkeit für Überbelegung gegeben durch

$$P\left(T_3 < \frac{\theta_c - \hat{\mu}/200}{\hat{\sigma}/200}\right) = P(T_3 < (0.42 - 0.45)/0.1075) \tag{409}$$

$$= P(T_3 < -0.278) = 1 - P(T_3 < 0.278). \tag{410}$$

Vergleich von $P(T_3 < 0.278)$ mit der Quantilstabelle für 3 Freiheitsgrade ergibt bei einem Quantil $q = 0.6$ einen Quantilswert, welcher nahezu gleich der Grenze 0.278 ist: $t_{0.6}^{(3)} = 0.277$. Da allgemein $P(T_3 < t_q^{(3)}) = q$, ist hier $P(T_3 < 0.277) = 0.6$. Damit ist die Wahrscheinlichkeit für Überbelegung gegeben durch

$$P(\text{Überbelegung}) = 1 - P(T_3 < 0.278) \approx 1 - P(T_3 < 0.277) = 0.4. \tag{411}$$

Lösung T.4.6

(a) Ein Schätzer ist

- erwatungstreu, wenn $E(\hat{\mu}) = \mu$
- konsistent, wenn $\lim_{n \to N} \hat{\mu} = \mu$, also je größer die Stichprobe, desto kleiner die Varianz
- effizient, wenn $V(\hat{\mu})$ minimal im Vergleich zu allen anderen Schätzern

Tabelle:

Schätzer	erwartungstreu	konsistent	effizient
$\hat{\mu}_1$	j	j	j
$\hat{\mu}_2$	n	j	n
$\hat{\mu}_3$	j	j	n
$\hat{\mu}_4$	j	n	n

Unter Ausnutzen der Beziehungen

$$E(aX + bY) = aE(X) + bE(Y) \quad \text{und} \tag{412}$$
$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\,\mathrm{Cov}(X, Y), \tag{413}$$

wobei die Kovarianz $\mathrm{Cov}(X, Y) = 0$ angenommen werden kann, erhält man:

$$E(\hat{\mu}_1) = \mu, \qquad V(\hat{\mu}_1) = \frac{\sigma^2}{n} \tag{414}$$
$$E(\hat{\mu}_2) = \frac{n}{n-1}\mu, \qquad V(\hat{\mu}_2) = \frac{n\sigma^2}{(n-1)^2} \tag{415}$$
$$E(\hat{\mu}_3) = \mu, \qquad V(\hat{\mu}_3) = \frac{2\sigma^2}{n} \tag{416}$$
$$E(\hat{\mu}_4) = \mu, \qquad V(\hat{\mu}_4) = \frac{\sigma^2(1+3+9+16)}{100} = \frac{3\sigma^2}{10} \tag{417}$$

Ausführlich:

(i) $\hat{\mu}_1 = \frac{1}{n}\sum_i x_i$

$$\begin{aligned} E(\hat{\mu}_1) &= E(\tfrac{1}{n}\textstyle\sum_i x_i) & V(\hat{\mu}_1) &= V(\tfrac{1}{n}\textstyle\sum_i x_i) \\ &= \tfrac{1}{n} \cdot E(\textstyle\sum_i x_i) & &= \tfrac{1}{n^2}V(\textstyle\sum_i x_i) \\ &= \tfrac{1}{n} \cdot \textstyle\sum_i E(x_i) & &= \tfrac{1}{n^2}\textstyle\sum_i \cdot V(x_i) \\ &= \tfrac{1}{n} \cdot \textstyle\sum_i \mu & &= \tfrac{1}{n^2}\textstyle\sum_i \sigma^2 \\ &= \tfrac{1}{n} \cdot n \cdot \mu & &= \tfrac{1}{n^2} \cdot n \cdot \sigma^2 \\ &= \mu & &= \tfrac{1}{n}\sigma^2 \end{aligned}$$

(ii) $\hat{\mu}_2 = \frac{1}{n-1}\sum_i x_i$

$$\begin{aligned} E(\hat\mu_2) &= E(\tfrac{1}{n-1}\sum_i x_i) & V(\hat\mu_2) &= V(\tfrac{1}{n-1}\sum_i x_i)\\ &= \tfrac{1}{n-1}\cdot E(\sum_i x_i) & &= \tfrac{1}{(n-1)^2}V(\sum_i x_i)\\ &= \tfrac{1}{n-1}\cdot \sum_i E(x_i) & &= \tfrac{1}{(n-1)^2}\sum_i \cdot V(x_i)\\ &= \tfrac{1}{n-1}\cdot \sum_i \mu & &= \tfrac{1}{(n-1)^2}\sum_i \sigma^2\\ &= \tfrac{n}{n-1}\cdot \mu & &= \tfrac{n}{(n-1)^2}\cdot \sigma^2 \end{aligned}$$

(iii) $\hat\mu_3 = \frac{2}{n}\sum_i^{\frac{n}{2}} x_i$

$$\begin{aligned} E(\hat\mu_3) &= E(\tfrac{2}{n}\sum_i^{\frac{n}{2}} x_i) & V(\hat\mu_3) &= V(\tfrac{2}{n}\sum_i^{\frac{n}{2}} x_i)\\ &= \tfrac{2}{n}\cdot E(\sum_i^{\frac{n}{2}} x_i) & &= \tfrac{4}{n^2}V(\sum_i^{\frac{n}{2}} x_i)\\ &= \tfrac{2}{n}\cdot \sum_i^{\frac{n}{2}} E(x_i) & &= \tfrac{4}{n^2}\sum_i^{\frac{n}{2}} \cdot V(x_i)\\ &= \tfrac{2}{n}\cdot \sum_i^{\frac{n}{2}} \mu & &= \tfrac{4}{n^2}\sum_i^{\frac{n}{2}} \sigma^2\\ &= \tfrac{2}{n}\cdot\tfrac{n}{2}\cdot \mu & &= \tfrac{4}{n^2}\cdot\tfrac{n}{2}\cdot \sigma^2\\ &= \mu & &= \tfrac{2}{n}\sigma^2 \end{aligned}$$

(iv) $\hat\mu_4 = \frac{1}{10}\sum_{i=1}^4 i\cdot x_i$

$$\begin{aligned} E(\hat\mu_4) &= E(\tfrac{1}{10}\sum_{i=1}^4 i\cdot x_i) & V(\hat\mu_4) &= V(\tfrac{1}{10}\sum_{i=1}^4 i\cdot x_i)\\ &= \tfrac{1}{10}\cdot E(\sum_{i=1}^4 i\cdot x_i) & &= \tfrac{1}{100}V(\sum_{i=1}^4 i\cdot x_i)\\ &= \tfrac{1}{10}\cdot \sum_{i=1}^4 iE(x_i) & &= \tfrac{1}{100}\sum_{i=1}^4 i^2\cdot V(x_i)\\ &= \tfrac{1}{10}\cdot \sum_{i=1}^4 i\cdot\mu & &= \tfrac{1}{100}\sum_{i=1}^4 i^2\cdot \sigma^2\\ &= \tfrac{1}{10}\cdot (1+2+3+4)\cdot\mu & &= \tfrac{1}{100}\cdot (1^2+2^2+3^2+4^2)\sigma^2\\ &= \mu & &= \tfrac{3}{10}\sigma^2 \end{aligned}$$

(b) Die Tabelle:

Schätzer	erwartungstreu	konsistent	effizient
$\hat\sigma_1^2$	n	j	n
$\hat\sigma_2^2$	j	j	j
$\hat\sigma_3^2$	j	j	n

(Nicht verlangt, zur Info:)

$$E\left(\hat\sigma_1^2\right) = \frac{n}{n-1}\sigma^2, \quad E\left(\hat\sigma_2^2\right) = \sigma^2, \quad E\left(\hat\sigma_3^2\right) = \sigma^2. \tag{418}$$

Lösung T.4.7

Nullhypothese: Anteil der Kfz-Besitzer in der Grundgesamtheit hängt nicht von den Studienfächern “Wirtschaft” bzw. “Biologie” ab.

Test-Statistik:

$$Q = \sum_{i=1}^{K_x}\sum_{j=1}^{K_y}(\frac{h_{ij}^2}{h_{ij}^e}) - n = n(-1 + \sum_{i=1}^{K_x}\sum_{j=1}^{K_y}\frac{h(x_i, y_j)^2}{h(x_i)h(y_j)}) \sim \chi^2(m) \quad (419)$$

Die Freiheitsgrade m lassen sich wie folgt berechnen: $m = (K_x - 1)(K_y - 1) = (2-1)(2-1) = 1$

Fach	Anzahl in Stichprobe	Anzahl Kfz-Besitzer	$\sum$
Biologie	234	87	321
Wirtschaft	789	401	1190
$\sum$	1023	488	1511

$$\begin{aligned} q &= 1511 \cdot (-1 + \frac{234^2}{1023 \cdot 321} + \frac{87^2}{488 \cdot 321} + \frac{789^2}{1023 \cdot 1190} + \frac{401^2}{488 \cdot 1190}) && (420)\\ &= 1511 \cdot (-1 + 1.0033) && (421)\\ &= 5.03 && (422) \end{aligned}$$

Alternativ

Vierfeldertest mit $a = 234$, $b = 87$, $c = 789$ und $d = 401$:

$$Q = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi^2(1). \quad (423)$$

Auswertung mit $n = a + b + c + d = 1511$:

$$q = 5.03. \quad (424)$$

Da das 95% Quantil $q_{0.95}^{(1)} = 3.84 < q$ ist, ist die Nullhypothese bei einer Fehlerwahrscheinlichkeit von 5% abzulehnen.

Da $q_{0.975}^{(1)} = 5.024 \approx q$, ist die Grenz-Fehlerwahrscheinlichkeit, bei der man die Nullhypothese gerade noch ablehnen kann, durch $\alpha = 2.5\%$ gegeben.

Lösung T.4.8

(a) Lineare Regression unter Verwendung der angegebenen Zwischengrößen $\bar{x} = 1996.5$, $\bar{y} = 0.6625$ und $s_x^2 = 20.25$ sowie der berechneten Summe $\sum_{i=1}^{4} x_i y_i = 5290.13$

$$b = \frac{s_{xy}}{s_x^2} \tag{425}$$
$$= \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}} \tag{426}$$
$$= \frac{5290.13 - 4 \cdot 1996.5 \cdot 0.6625}{15944130 - 4 \cdot 1996.5^2} \tag{427}$$
$$= \frac{-0.595}{81} \tag{428}$$
$$= -0.007346 \tag{429}$$

sowie

$$\overline{y} = a + b\overline{x} \tag{430}$$
$$a = \overline{y} - b\overline{x} \tag{431}$$
$$= 0.6625 - (-0.007346) \cdot 1996.5 \tag{432}$$
$$= 15.328 \tag{433}$$

und daraus die Regression:

$$\hat{y}(x) = 15.328 - 0.007346x. \tag{434}$$

Prognose für 2008:

$$\hat{y}(2008) = 15.328 - 0.007346 \cdot 2008 = 0.5772 \tag{435}$$

(b) Einseitiger T-Test mit $H_0 : \beta_0 \geq 0$, $n = 4$ und der Testvariable:

$$T_b = \frac{b - \beta_0}{\sqrt{\sigma_b^2}} \sim T_{1-\alpha}(n-m). \tag{436}$$

dabei sei $\beta_0 = 0$ und damit $T_b = \frac{b}{\sqrt{\sigma_b^2}}$. Des Weiteren ist $\sigma_b^2 = \frac{\sigma_R^2}{n \cdot s_x^2}$. Wird diese in die Gleichung für die Testvariable eingesetzt, dann erhalten wir

$$t_b = \frac{b}{\sqrt{\frac{\sigma_R^2}{n \cdot s_x^2}}} \tag{437}$$
$$= \frac{2 \cdot b \cdot s_x}{\sigma_R} \tag{438}$$

Die Residualvarianz berechnet sich

$$\sigma_R^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}(x_i))^2 \tag{439}$$
$$= \frac{1}{2} \cdot 0.0015043 \tag{440}$$
$$= 0.000752. \tag{441}$$

Damit ist

$$t_b = \frac{2 \cdot (-0.007346) \cdot \sqrt{20.25}}{\sqrt{0.000752}} = -2.411. \tag{442}$$

Die Nullhypothese “ÖV-Wegezahlen sinken nicht” kann nicht abgelehnt werden, falls

$$t_b > -t^{(2)}_{1-\alpha} = -t^{(2)}_{0.95} \tag{443}$$
$$-2.41 > -2.920 \tag{444}$$

Dies ist eine wahre Aussage $\Rightarrow$ man kann *nicht* auf eine sinkende Wegezahl schließen!

(c) Konfidenzintervall bei 10% Fehlerwahrscheinlichkeit für die Prognose 2008:

$$y(2008) \in [\hat{y}(2008) - \Delta(2008), \hat{y}(2008) + \Delta(2008)] \tag{445}$$

mit

$$\sigma_{\hat{y}}^2 = \frac{\sigma_R^2}{\sqrt{n}}\sqrt{1 + \frac{(x-\overline{x})^2}{s_x^2}} \tag{446}$$
$$= \frac{0.0007522}{2}\sqrt{1 + \frac{(2008 - 1996.5)^2}{20.25}} \tag{447}$$
$$= 0.001032. \tag{448}$$

Damit ist unter Verwendung von $t^{(2)}_{1-\alpha/2} = t^{(2)}_{0.95} = 2.920$

$$\Delta(2008) = \sqrt{0.001032} \cdot 2.920 = 0.0938. \tag{449}$$

und somit für das Konfidenzintervall

$$\hat{y}(2008) + \Delta(2008) = 0.5772 + 0.0938 = 0.6710 \tag{450}$$
$$\hat{y}(2008) - \Delta(2008) = 0.5772 - 0.0938 = 0.4834. \tag{451}$$

$$y(2008) \in [0.483, 0.671] \tag{452}$$

(d) Bei der längeren Zeitreihe erstrecken sich die Datenpunkte in die frühere DDR hinein. Durch die Umwälzungen zur Wendezeit ist sicher die Annahme der Homoskedastizität (die Varianz der Abweichungen der einzelnen Datenpunkte von der Regressionsfunktion ist konstant) und damit eine der Annahmen der Konfidenzintervall-Schätzung von Regressionsfunktionen nicht erfüllt. (die Erwähnung der Wende allein reicht zur Beantwortung dieses Teils).

Lösung T.4.9

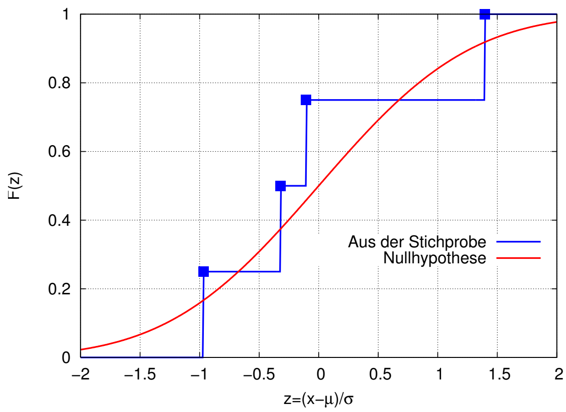
(a) Kolmogorow-Testvariable:

$$D = \max_x |F_{emp}(x) - F_0(x)| \tag{453}$$

Da bei Einzelwerten, im Ggs zu klassierten Daten, die Verteilungsfunktion $F_{emp}(x)$ eine Treppenfunktion ist (vgl. Plot), ergeben sich an jeder "Stufe" zwei Möglichkeiten für das Maximum: Am unteren und am oberen Rand der Stufe:

$$D = \max_{i=1...4, j=0,1} |F_{i-j} - F_0(x_i)| \tag{454}$$

Berechnen von $F(x_i)$ mittels der kummulierten relativen Häufigkeiten mit $\sum x_i = 4 \cdot \bar{x} = 5.98$ und damit z.B. $F(x_1) = f(x_1) = 1.45/5.98 \approx \mathbf{0.25}$ und $F(x_2) = f(x_1) + f(x_2) = 0.25 + 1.48/5.98 \approx 0.25 + 0.25 = \mathbf{0.50}$ usw. Soiehe dazu auch den Plot.



Um die Verteilungsfunktion der Nullhypothese zu ermitteln, verwendet man die Tabelle der Standardnormalverteilung $\Phi(z) \Rightarrow$ benötige die Werte $z_i = (x_i - \mu)/\sigma$ der Werte der standardnormalverteilten Zufallsvariablen Z an den Stellen x_i, so z.B. $(1.45-1.495)/\sqrt{0.002167} = \mathbf{-0.97}$.

Arbeitstabelle der nach steigenden Werten geordneten x_i:

i	x_i	$F(x_i)$	z_i	$F_0(x_i) = \Phi(z_i)$	$\|F_i - F_0(x_i)\|$	$\|F_{i-1} - F_0(x_i)\|$
1	1.45	**0.25**	**-0.97**	0.17	$0.08 = \|0.25 - 0.17\|$	$0.17 = \|0 - 0.17\|$*
2	1.48	**0.50**	-0.32	0.37	0.13	$0.12\| = 0.25 - 0.37\|$
3	1.49	0.75	-0.11	0.46	0.29	$0.04\| = 0.5 - 0.46\|$
4	1.56	1.00	1.40	0.92	0.08	$0.17\| = 0.75 - 0.92\|$

**unterer Rand der Stufe*

z_i lässt sich in der Tabelle für Standardnormalverteilungen nachschlagen und dann als $F_0(x_i) = \Phi(z_i)$ in der obigen Tabelle eintragen. Ist jedoch der Wert von z_i negativ, muss wie im 1. Fall $1 - \Phi(z_1) = 1 - 0.83 = 0.17$ berechnet werden. Also liegt das Maximum bei

$$D = 0.29 \tag{455}$$

Nullhypothese kann nur verworfen werden, falls

$$D > d_{0.05} = \frac{c_\alpha}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}} \stackrel{c_\alpha = 1.358}{=} 0.62. \tag{456}$$

Dies ist nicht der Fall, also H_0 nicht verwerfbar (und damit der nur bei Gaußverteilung gültige t-Test und χ^2-Varianztest zumindest plausibel).

(b)
- Nullhypothese H_0: $\mu > \mu_0 = 1.5$
- Testvariable: $T \sim T(n-1)$ (Student-Verteilung mit $n-1 = 3$ Freiheitsgraden)
- Realisierung:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \frac{1.495 - 1.5}{\sqrt{0.002167}}\sqrt{4} = -0.215. \quad (457)$$

- Quantil zu Fehlerwahrscheinlichkeit $\alpha = 5$ %: $t_{0.95}^{(3)} = 2.353$
- Testentscheidung: H_0 angenommen, falls

$$t > -t_{0.95}^{(3)} \quad \text{Wahre Aussage} \Rightarrow H_0 \text{ angenommen bzw. kann nicht abgelehnt werden.} \quad (458)$$

(c)
- Nullhypothese: $\sigma^2 < \sigma_0^2 = 0.05^2$
- Testvariable: $Q \sim \chi^2(n-1) = \chi^2(3)$
- Realisierung aus der Stichprobe: $q = \frac{(n-1)s^2}{\sigma_0^2} = \frac{3 \cdot 0.002167}{0.05^2} = 2.6$
- Entscheidung: H_0 angenommen bzw. nicht abgelehnt werden, falls $q < q_{0.95}^{(3)} = 7.8$. Dies ist der Fall!

(d) (i) Mittelwerttest mit komplementärer Nullhypothese $\overline{H_0} : \mu < 1.5$

Testvariable und Quantil wie oben, Testentscheidung: $\overline{H_0}$ angenommen, falls

$$t < +t_{0.95}^{(3)} \text{ also } 0.215 < 2.353 \quad \text{Wahre Aussage} \Rightarrow \overline{H_0} \text{ angenommen!} \quad (459)$$

(ii) Varianztest mit komplementärer Nullhypothese $\overline{H_0} : \sigma > \sigma_0 = 0.05^2$.

Testvariable wie oben, wegen der unsymmetrischen χ^2-Verteilung wird aber explizit das 5%-Quantil benötigt: $q_{0.05}^{(3)} = 0.35$

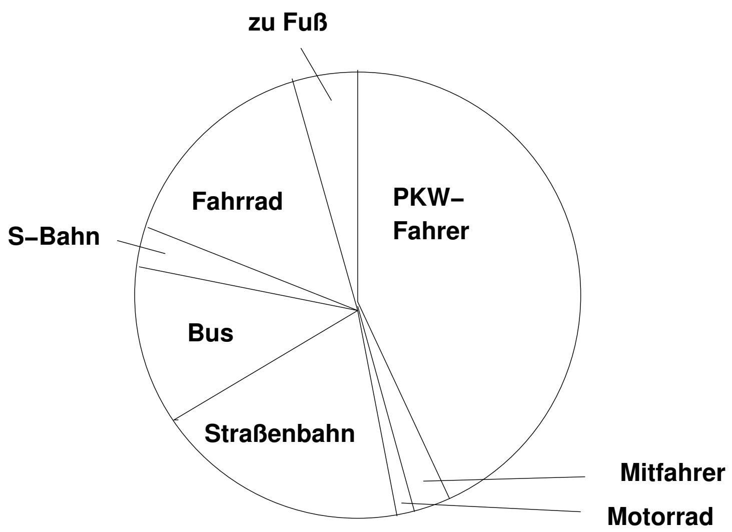
Testentscheidung: $\overline{H_0}$ angenommen, falls

$$q > +q_{0.05}^{(3)} \text{ also } 2.6 > 0.35 \quad \text{Wahre Aussage} \Rightarrow \overline{H_0} \text{ angenommen!} \quad (460)$$

Also werden auch die komplementären Nullyhpothesen nicht verworfen, d.h. der Mittelwert kann sehr wohl (mit einer Wahrscheinlichkeit weit oberhalb 5%) unter 1.5 und die Varianz oberhalb 0.05^2 liegen!

Lösung T.4.10

(a) Tortendiagramm (ein qualitatives Diagramm mit in etwa den richtigen Grißenverhältnissen ist OK)



(b) Da der Stichprobenfehler für Anteilswerte θ proportional zu $\theta(1-\theta)$ ist, ist der größte Fehler für den Anteil zu erwarten, welcher am nächsten an 50% liegt. Dies ist für den Anteil θ_1 an PKW-Fahrern mit dem Stichprobenergebnis $f_1 = 0.43$ der Fall. Bestimmung des Stichprobenumfangs aus der Konfidenzintervall-Formel für den Anteilswert (oder dem Test des Anteilswertes auf $\theta_0 = f_1$):

$$\theta \in f_1 \pm z_{1-\alpha/2}\sqrt{\frac{f_1(1-f_1)}{n}} \tag{461}$$

Nach Aufgabenstellung soll gelten: $|\theta - f_1| < 0.02$. Durch Umstellung erhält man für $\alpha = 5\%$:

$$n = \frac{f_1(1-f_1)z^2_{1-\alpha/2}}{(\theta - f_1)^2} = \frac{0.43 * 0.57 * 1.96^2}{0.02^2} = 2354. \tag{462}$$

(c) Zwei-Stichproben-Differenztest nach Skript 27.5. Es handelt sich hier um eine unverbundene Stichprobe, da es nicht notwendig ist, in den zwei Stichproben dieselben Personen zu befragen und auch die Umfänge unterschiedlich groß sein können:

$$Z = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \tag{463}$$

Hierbei ist

- Die Nullhypothese H_0: $\theta_0 = 11\% = 0.11$ bei beiden Stichproben,
- Die Zufallsvariablen X und Y sind zweiwertig mit den Werten 0 (“kein Fahrrad”) und 1 (“Fahrrad”) mit Wahrscheinlichkeiten (bei Zutreffen von H_0) von $(1-\theta_0)$ bzw. θ_0.
- Stichprobenmittel $\bar{X} = 11\%$ (erste Stichprobe) bzw. $\bar{Y} = 13\%$ (zweite Stichprobe),
- Die Varianzen bei Zutreffen von H_0 sind $s_x^2 = s_y^2 = \theta_0(1-\theta_0) = 0.0979$,
- Die Stichprobenumfänge sind $n = m = 2000$.

Alles eingesetzt ergibt die Realisierung z von Z:

$$Z = \frac{|0.11 - 0.13|}{\sqrt{\frac{2 \cdot 0.0979}{2000}}} = \frac{0.02\sqrt{2000}}{\sqrt{2 \cdot 0.0979}} = 2.02 \tag{464}$$

Das Quantil ist

$$z_{1-\alpha/2} = z_{0.995} = 2.576 \;\Rightarrow\; H_0 \text{ nicht widerlegbar.} \tag{465}$$

Die Annahme der Gleichheit der Fahrradanteile kann also durch die Stichproben nicht widerlegt werden.

Lösung T.4.11

(a) (i) Der Mittelungszeitraum τ muss deutlich größer als 1 Monat sein (September-Peak wird weitgehend weggemittelt wie in der in der Aufgabenstellung angegebenen Kurve), aber deutlich kleiner als 1 Jahr, damit die Jahresfigur nicht weggemittelt wird, z.B.

$$\tau = 3 \text{ Monate.} \tag{466}$$

(jeder Wert zwischen 2 und 11 Monaten ergibt volle Punktzahl).

(ii) Es handelt sich um eine periodische Zeitreihe, d.h. der 31. Dezember muss bei der Glättung stetig (und differenzierbar) mit dem 1. Januar verbunden werden. Die Zeitreihe ist quasi als "Ring" gechlossen und der 1. Januar ebenso Nachbar des 31. Dezember wie der 31. Dezember Nachbar des 30. Dezembers ist. Bei $\tau = 3$ Monaten würde man z.B. für den 31. Dezember das arithmetische Mittel der Geburten von 15. November bis 15. Februar nehmen.

(b) Realisierung d der Testgröße D: "maximale betragsmäßige Abweichung der empirischen Verteilungsfunktion von der Nullhypothese":

$$d = \frac{1}{n}\left(6 \text{ Kästchen} \cdot 30\text{Tage/Kästchenbreite} \cdot 50\text{/Kästchenhöhe}\right) \tag{467}$$

Die Einwohnerzahl Dresdens wird entweder zu $n = 481\ 429$ (von Aufgabe 1) oder aus dem in der Abbildung angegebenen Mittelwert berechnet:

$$n = 365 \text{ Tage} \cdot 1\ 310 \text{ Geburtstage/Tag} = 478\ 150 \tag{468}$$

und damit

$$d = \frac{9\ 000}{478\ 150} = 0.0188 \tag{469}$$

Das $(1-\alpha) = 99\%$ Quantil der D-Statistik aus der Formelsammlung:

$$d_{n,0.99} = \frac{1.628}{\sqrt{n} + 0.12 + 0.11/\sqrt{n}} = 0.00235 \tag{470}$$

Beim Vergleich zeigt sich:

$$0.00235 < 0.0188 \tag{471}$$

Also ist die Realisierung d weit oberhalb des Quantils und damit die Nullhypothese einer Gleichverteilung hochsignifikant abgelehnt.

(c) Auch die September-Spitze stellt eine kumulierte Abweichung gegenüber dem Mittelwert von etwa 2 Kästchen dar, also

$$d_{september} = \frac{1}{n}\left(2 \text{ Kästchen} \cdot 30\text{Tage/Kästchenbreite} \cdot 50\text{/Kästchenhöhe}\right) = 0.00627. \tag{472}$$

Der September-Peak als solcher stellt also ebenfalls eine signifikante Abweichung gegenüber der Geburten-Gleichverteilung dar. Dies kann viele Gründe haben, sowohl direkte ("erhöhte Fruchtbarkeit"), als auch indirekte (Weihnachten hat man eher Zeit etc). Man kann jedenfalls *nicht* von einer statistischen Korrelation auf die Ursache schließen, hat also *nicht* einen besonders "fruchtbaren" Empfängniszeitraum im Dezember nachgewiesen.

(d) Die "offizielle" vermutete Erklärung für diesen Sachverhalt ist die, dass bei Unklarheit des Geburtstags (Es werden ja alle Einwohner, auch sehr alte oder aus anderen Ländern eingewanderte erfasst!), der 1. Januar *ad hoc* angenommen wird. Dies kann auch von Seiten der Einwohnermelde-Software (Default ist der 1. Januar) etc geschehen. (Übrigens sind andere "runde" Geburtstage wie der 15. oder der Erste eines Monats auch signifikant höher vertreten. Auch in der Abbildung zu Aufgabe 4 ist eine erhöhte Geburtstagszahl am 1. Januar zu sehen).

Lösung T.4.12

(a) Bei Schätzung eines Mittelwertes bei unbekannter Varianz wird das Konfidenzintervall mit der t-Verteilung gebildet: Bei einer Fehlerwahrscheinlichkeit $\alpha = 5\%$ erhält man

$$\langle x \rangle \in [\bar{x} - t^{(n-1)}_{1-\alpha/2} \frac{s_x}{\sqrt{n}}, \bar{x} + t^{(n-1)}_{1-\alpha/2} \frac{s_x}{\sqrt{n}}] \tag{473}$$

Für die Wochen 18-24 gilt (mit der empirischen Varianz mit $(n-1)$ im Nenner):

$$n = 7, \bar{x} = \frac{12802}{7} = 1829, \quad t^{(6)}_{0.975} = 2.447, \quad s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = 247726, \quad s_x = 497.7 \tag{474}$$

und damit

$$\langle x \rangle \in 1829 \mp 2.447 \cdot \frac{497.7}{\sqrt{7}} \tag{475}$$

$$\langle x \rangle \in [1369, 2289] \tag{476}$$

(b) Analog für das zweite Intervall (Wochen 25-28):

$$n = 4, \bar{x} = \frac{10047}{4} = 2512, \quad t^{(3)}_{0.975} = 3.182, \quad s_x^2 = \frac{773279}{3} = 257760 \quad s_x = 507.7 \tag{477}$$

und damit

$$\langle x \rangle \in 2512 \mp 3.182 \cdot \frac{507.7}{\sqrt{4}} \tag{478}$$

$$\langle x \rangle \in [1704, 3320] \tag{479}$$

Die Klickraten sind also bei $\alpha = 5$ % nicht signifikant verschieden, da sich die Konfidenzintervalle aus a) und b) überlappen. Wir erhöhen die Fehlerwahrscheinlichkeit, um die Trennschärfe zu erhöhen: In der Tabelle liegen die Quantile der t-Verteilung für $q = 0.95, 0.9, 0.8, \ldots$ vor. Damit können die Konfidenzintervalle $[x_{1u}, x_{1o}]$ und $[x_{2u}, x_{2o}]$ der Klickraten für die Zeitintervalle 1 und 2 für $\alpha = 10\%$, $\alpha = 20\%$, $\alpha = 40\%$ ermittelt werden:

α	x_{1u}	x_{1o}	x_{2u}	x_{2o}
0.10	1463	2194	1914	3109
0.20	1557	2100	2096	2928
0.40	1658	1999	2264	2760

Erst bei einer Fehlerwahrscheinlichkeit von knapp über 20% ist eine hinreichende Trennschärfe für eine signifikante Unterscheidung gegeben, da sich die Konfidenzintervalle nicht mehr überlappen!

(c) In der Verteilung der täglichen Klickraten sieht man ein deutliches periodisches Muster mit einem Maximum an den Wochenenden, d.h. die Lancierung des potentiell klickratensteigernden Artikels fällt mit der Zeit von sowieso periodisch hohen Klickraten zusammen. Analysiert man die täglichen Daten ohne Vorverarbeitung einige Tage vor und nach dem Artikel, führt der Wochenend-Peak möglicherweise zu falschen Schlüssen. Lösung: Man *bereinigt* die Daten durch Herausrechnen der Saisonkomponente und analysiert die saisonbereinigten Daten.

Lösung T.4.13

(a) Die χ^2-verteilte Testvariable ist durch

$$Q = \sum_{i=1}^{5} \sum_{j=1}^{2} \left(\frac{h(x_i, y_j)^2}{h_{ij}^e} \right) - nQ = n(-1 + \sum_{i=1}^{5} \sum_{j=1}^{2} \left(\frac{h(x_i, y_j)^2}{h(x_i)h(y_i)} \right)) \tag{480}$$

mit der erwarteten absoluten Häufigkeit bei Zutreffen der Nullhypothese der homogenen Meinungsspektrums,

$$h_{ij}^e = \frac{h(x_i)h(y_j)}{n} \tag{481}$$

den Spaltensummen $h(y_j) = \sum_{i=1}^5 h(x_i, y_j)$, den Zeilensummen $h(x_i) = \sum_{j=1}^2 h(x_i, y_j)$ und $n = 200$. Dies führt zu einer Realisierung q der Testvariablen Q von

$$q = 200 \cdot (-1 + \frac{4^2}{14 \cdot 60} + \frac{10^2}{14 \cdot 140} + \frac{6^2}{34 \cdot 60} + \ldots + \frac{20^2}{40 \cdot 140}) \tag{482}$$

$$q = 10.69 \tag{483}$$

Die Nullhypothese ist bei einer Fehlerwahrscheinlichkeit von α angenommen bzw. kann nicht abgelehnt werden, falls die Ungleichung

$$q < q_{1-\alpha}^{(m)}, \quad m = (5-1)(2-1) = 4 \tag{484}$$

Da $q_{0.975}^{(4)} = 11.14$, ist dies bei $\alpha = 2.5\%$ der Fall. Ein homogenes Meinungsbild kann nicht abgelehnt werden.

(b) (i) Falsch. Aus der Tatsache, dass man H_0 bei einer Fehlerwahrscheinlichkeit von 2.5% nicht ablehnen kann folgt *nicht*, dass sie mit mindestens $(1-\alpha)$ wahr ist!

(ii) Richtig. Da $q_{0.975}^{(4)} = 9.49 < 10.69$, ist die obige Ungleichung nicht mehr erfüllt und H_0 kann abgelehnt werden.

(iii) Richtig. Sei T das Ereignis "Test angenommen", so könnte man die Wahrscheinlichkeit $P(H_0|T)$ dafür, dass bei Annahme des Tests die Nullhypothese auch richtig ist, prinzipiell mit dem Satz von Bayes berechnen, da aus der Definitionsgleichung der Fehlerwahrscheinlichkeit die Bedingung $P(T|H_0) = 1 - P(\bar{T}|H_0) \geq 1 - \alpha$ folgt und damit

$$P(H_0|T) = \frac{P(H_0)}{P(T)} P(T|H_0) \geq \frac{P(H_0)}{P(T)}(1-\alpha) \tag{485}$$

Da aber die a-Priori-Wahrscheinlichkeiten $P(H_0)$ und $P(T)$ i.A. nicht ermittelbar sind, scheitert dies meist. Man könnte hier sogar begründen, dass $P(H_0) = 0$, da H_0 eine exakte Gleichheit von reellen Zahlen (der wahren Meinungsanteile) impliziert, während $P(T)$ sicherlich ungleich 0 ist (sonst würde ja nie ein Test angeommen!) Damit erhält man $P(H_0|T) = 0$, d.h die Wahrscheinlichkeit dafür, dass H_0 falsch ist, kann sogar 100% betragen!

(iv) *Schwierig!* Falsch: Formuliert man einen Test, der die Alternativhypothese H_A bei einer Fehlerwahrscheinlichkeit von β testet und bezeichnet das Ereignis der Ablehnung von H_A durch diesen Test mit $\bar{U}$, ist der Fehler erster Art bei diesem Test durch $P(\bar{U}|H_A) < \beta$ definiert. Wieder erhält man die interessierende Größe $P(H_A|\bar{U})$ nur über den Satz von Bayes, also nur bei Kenntnis der A-Priori-Wahrscheinlichkeiten. Dies ist übrigens auch beim Test auf H_0 der Fall: Nicht einmal der Fehler erster Art erlaubt ohne weitere Informationen Aussagen über die Wahrscheinlichkeiten des Zutreffens von H_0 oder $\bar{H}_0$!

Lösung T.4.14

(a) Die Nullhypothese kann abgelehnt werden, falls die Realisierung $q_{geburten}$ der Testvariablen

$$Q = \sum_{k=1}^{K} \left(\frac{h_k^2}{h_k^e} \right) - n \tag{486}$$

größer als das Quantil

$$q_{0.975}^{(8-1)} = q_{0.975}^{(7)} = 16.14 \tag{487}$$

ist. Kein Parameter ist frei, deshalb ist die Zahl der Freiheitsgrade $m = K - 1 = 7$. Die konkrete Realisierung aus der Stichprobe ergibt mit $h_k^e = 8280/8 = 1035$ zu

$$q_{geburten} = 15.46 < q_{0.975}^{(7)}, \tag{488}$$

also kann die Annahme einer Gleichverteilung nicht abgelehnt werden.

Analog für die Zuzüge:

$$n_{zu} = 43\,100, \quad h_e^{zu} = 5\,387.5, \quad q_{zu} = 710 < q_{0.975}^{(7)} \tag{489}$$

Die Verteilung der Zuzüge auf die einzelnen Quartale weicht signifikant von einer Gleichverteilung ab.[6]

(b)

$$n_{zu,Q4} = 12\,600, \quad h_e^{zu} = 6\,300, \quad q_{zu,Q4} = 3.175 < q_{0.975}^{(2)} = 5.02 \tag{490}$$

Die Annahme einer gleichmäßigen Verteilung kann nicht widerlegt werden.

(c) Nein. Mit den hier behandelten statistischen Tests kann man den Fehler erster Art (Test lehnt Nullhypothese ab, obwohl sie zutrifft) klein halten bzw. kontrollieren. Über die Fehler zweiter Art: Annahme von H_0, obwohl sie nicht zutrifft, kann i.A. *nichts* ausgesagt werden. Insbesondere kann die Fehlerwahrscheinlichkeit zweiter Art nahe 1 oder sogar gleich 1 sein.

Zum zweiten Teil des Hinweises: Nicht relevant! Da die hypothetische Gesamtheit hier alle folgenden und vergangenen Jahre einschließt und damit unendlich ist, kennt man die konkreten Verhältnisse der Grundgesamtheit nicht; eine Aussage ist nicht möglich.

[6] Streng genommen müsste man berücksichtigen, dass die vier Quartale eines Jahres unterschiedlich viel Tage aufweisen! Dies fällt hier jedoch nicht ins Gewicht.

Lösung T.4.15

(a) Aus der Arbeitstabelle

i	x_i	$(x_i - \bar{x})^2$
1	3.05	0.0049
2	3.35	0.0529
3	3.20	0.0064
4	3.00	0.0144
5	3.00	0.0144
$\sum_i$	15.60	0.0930

folgen der Schätzer des Mittelwertes

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{491}$$
$$= \frac{15.6}{5} \tag{492}$$
$$= 3.12 \tag{493}$$

und der Varianz

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{494}$$
$$= \frac{0.0930}{4} \tag{495}$$
$$= 0.02325 \tag{496}$$

Das Konfidenzintervall des Mittelwertes bei unbekannter Varianz lautet

$$\mu \in \bar{x} \pm t_{1-\alpha/2}^{(n-1)} \cdot \sqrt{\frac{s^2}{n}} \tag{497}$$
$$\in 3.12 \pm t_{0.95}^{(4)} \cdot \sqrt{\frac{0.02325}{5}} \tag{498}$$
$$\in 3.12 \pm 2.132 \cdot 0.0682 \tag{499}$$
$$\in 3.12 \pm 0.1454 \tag{500}$$
$$\in [2.975, \quad 3.265] \tag{501}$$

(b) Test auf Mittelwert bei unbekannter Varianz

1. Nullhypothese:

$$H_0 : \mu < \mu_0 = 3 \tag{502}$$

2. Testvariable:

$$T = \frac{\bar{X} - \mu_0}{S} \cdot \sqrt{n} \qquad \sim T(n-1) \tag{503}$$

3. Realisierung:

$$t = \frac{\bar{x} - \mu_0}{s} \cdot \sqrt{n} \tag{504}$$
$$= \frac{3.12 - 3}{\sqrt{0.02325}} \cdot \sqrt{5} \tag{505}$$
$$= 1.7598 \tag{506}$$

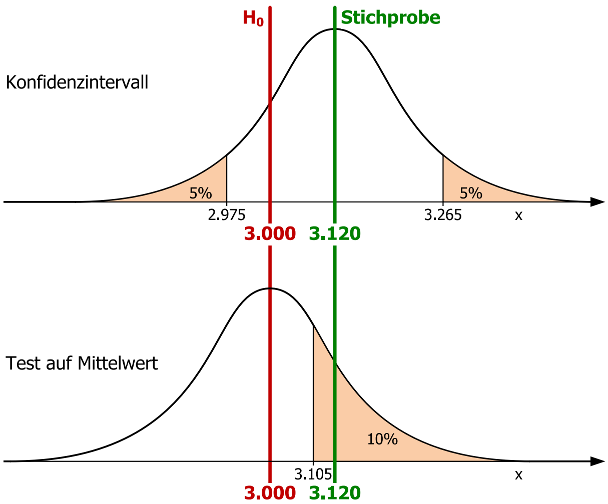
4. Entscheidung: H_0 ist abzulehnen, da

$$t > t_{1-\alpha}^{(n-1)} \tag{507}$$
$$1.7598 > t_{0.9}^{(4)} \tag{508}$$
$$1.7598 > 1.533 \tag{509}$$

Grenz-Fehlerwahrscheinlichkeit

$$t = t_{1-\alpha}^{(n-1)} \tag{510}$$
$$1.7598 = t_{1-\alpha}^{(4)} \tag{511}$$
$$\text{ist erfüllt für } \alpha = 7.66\% \tag{512}$$

(c) Einseitiger versus zweiseitiger Test.



(d) Mit dem Ansatz

$$\text{Reichweite} = \frac{\text{Tankvolumen}}{\text{Verbrauch}} \cdot 100 \tag{513}$$

folgt das Konfidenzintervall

$$\text{Reichweite} \in \left[\frac{25}{3.265} \cdot 100, \quad \frac{25}{2.975} \cdot 100\right] \tag{514}$$
$$\in [765.6, \quad 840.4] \tag{515}$$

(e) Test der Varianz

1. Nullhypothese:

$$H_0 : \sigma^2 > \sigma_0^2 = 0.4^2 = 0.16 \tag{516}$$

2. Testvariable:

$$Q = \frac{(n-1) \cdot S^2}{\sigma_0^2} \quad \sim \chi^2(n-1) \tag{517}$$

3. Realisierung:

$$q = \frac{(n-1) \cdot s^2}{\sigma_0^2} \tag{518}$$

$$= \frac{4 \cdot 0.02325}{0.16} \tag{519}$$

$$= 0.58125 \tag{520}$$

4. Entscheidung: H_0 kann abgelehnt werden, da

$$q < q_{\alpha}^{(n-1)} \tag{521}$$

$$0.58125 < q_{0.1}^{(4)} \tag{522}$$

$$0.58125 < 1.064 \tag{523}$$

erfüllt ist.

Lösung T.4.16

Lösungsvorschlag folgt.

Lösung T.4.17

(a) Konfidenzintervall für Anteilswert p bei bekannter relativer Häufigkeit f und $\alpha = 5\%$:

$$p \in \left[f - z_{0.975}\sqrt{\frac{f(1-f)}{n}}, f + z_{0.975}\sqrt{\frac{f(1-f)}{n}}\right] \tag{524}$$

So z.B. für die Fußgänger:

$$p \in [0.22 - 1.96 \cdot \sqrt{\frac{0.22(1-0.22)}{1000}}, 0.22 + 1.96 \cdot \sqrt{\frac{0.22(1-0.22)}{1000}}] \tag{525}$$

also:

- Fußgänger: $f \in [0.194324748102501, 0.245675251897499]$
- Rad: $f \in [0.0906068919458484, 0.129393108054152]$
- ÖPNV: $f \in [0.194324748102501, 0.245675251897499]$
- Kfz: $f \in [0.419165019863798, 0.480834980136202]$

(b) **Absoluter Fehler:** Für jedes einzelne Verkehrsmittel den entsprechenden Stichprobenumfang berechnen und das Maximum wählen. So ist nach dem Umstellen der Formel:

$$n = \frac{f(1-f)z^2_{1-\frac{\alpha}{2}}}{(f-p)^2} \tag{526}$$

Damit kann jetzt z.B. für die Kfz

$$n_{kfz} = \frac{0.45(1-0.45)1.96^2}{0.02^2} \tag{527}$$

$$n_{kfz} = 2377 \tag{528}$$

berechnet werden. (Der Stichprobenumfang n muss noch für zu Fuß, Bahn und Bus ermittelt werden.) Beim Vergleich erkennt man, dass die Kfz den höchsten Stichprobenumfang $\Rightarrow$ $n_{abs} = 2377$ verlangen.

Relativer Fehler: Sei

$$f - p = z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \tag{529}$$

und der relative Fehler

$$R = \frac{f-p}{f} = 0.05 \tag{530}$$

$$\tag{531}$$

$$0.05 = \frac{z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}}{f} \tag{532}$$

und nach dem Umstellen

$$n = \frac{f(1-f) \cdot z^2_{1-\alpha/2}}{(R \cdot f)^2} \tag{533}$$

Für beispielsweise Radfahrer:

$$n = \frac{0.11(1-0.11) \cdot 1.96^2}{(0.05 \cdot 0.11)^2} = 12433 \tag{534}$$

Die Radfahrer verlangen den höchsten Stichprobenumfang, wenn der relative Fehler unter 0.05 liegen soll. $\Rightarrow n_{rel} = 12433$

(c) Einseitiger Test mit $H_0 : \theta \geq 0.24$ bei relativer Häufigkeit $f = 0.22$ Realisation mit

$$z = \frac{(f-\theta)}{\sqrt{\theta(1-\theta)}} \cdot \sqrt{n} \tag{535}$$

$$= \frac{(0.22-0.24)}{\sqrt{0.24 \cdot (1-0.24)}} \cdot \sqrt{1000} \tag{536}$$

$$|z| = 1.48 \tag{537}$$

$\Rightarrow$ Wert der Z-Statistik ist mit $|z| = 1.48$ kleiner als $z_{0.95} = 1.645$ (aus der Quantil-Tabelle) $\Rightarrow$ H_0 angenommen bzw. nicht abgelehnt. Mit einer Irrtumswahrscheinlichkeit von 5% kann nicht abgelehnt werden, dass der Anteil der ÖPNV Nutzer größer ist als 24%.

(d) Z.B. testen, ob die Konfidenzintervalle davor und danach überschneiden. Falls nicht, signifikante Änderung. Oder die Differenz der Anteilswerte davor und danach auf die Hypothese H_0 :Differenz=0 testen. Falls H_0 abgelehnt $\Rightarrow$ signifikante Änderung.

Lösung T.4.18

Fachrichtung	zu Fuß	Rad	Bahn/Bus	Kfz	$\sum x_i$
Verkehrswirtschaft	18	30	30	22	100
Physik	8	18	10	14	50
BWL	34	50	50	66	200
$\sum y_j$	60	98	90	102	350

(a) Für beliebiges Element (ij) gilt hier $f(x_i, y_j) \neq f(x_i)f(y_j)$ z.B. x_1 =”Fachrichtung Verkehrswirtschaft”, y_1 =”benutzt Schusters Rappen”, $f(x_1, y_1) = 18/350$, $f(x_1) = 100/350$ (1. relative Häufigkeit der Zeilensumme), $f(y_1) = 60/350$ (relative Häufigkeit der 1. Spaltensumme). Daraus ergibt sich

$$\frac{18}{350} \neq \frac{100}{350} \cdot \frac{60}{350} \tag{538}$$

$$\frac{6300}{122500} \neq \frac{6000}{122500} \tag{539}$$

(b) a) H_0: Die Verkehrsmittelwahl hängt nicht vom Studiengang ab.

b) Testvariable

$$Q = \sum_{i=1}^{4} \sum_{j=1}^{3} \left(\frac{h(x_i, y_j)^2}{h_{ij}^e}\right) - n \quad \sim \chi^2(m) \tag{540}$$

$$h_{ij}^e = \frac{h(x_i)h(y_j)}{n} \tag{541}$$

c) Realisation

$$q = 350 \cdot \left(-1 + \frac{18^2}{60 \cdot 100} + \frac{8^2}{60 \cdot 50} + \ldots + \frac{66^2}{102 \cdot 200}\right) \tag{542}$$

$$= 6.195 \tag{543}$$

Die Anzahl der Freiheitsgerade berechnen sich mit $m = (4-1)(3-1) = 6$ und damit ist $q_{0.95}^{(6)} = 12.6 \Rightarrow$ Unabhängigkeit kann nicht verworfen werden, da $6.195 < 12.6$.

(c) Entsprechendes Quantil heraussuchen. Siehe Tabelle. Beispielsweise ist $q_{0.5}^{(6)} = 5.348$ und damit kleiner als das berechnete $q = 6.195 \Rightarrow$ Die Nullhypothese kann mit einer Irrtumswahrscheinlichkeit von 50% verworfen werden.

Lösung T.4.19

Nutzerklasse	Neues System ist besser	Neues System ist schlechter	keine Meinung	
Bahnbenutzer, vorwiegend Nahverkehr	20	270	10	300
Bahnbenutzer, vorwiegend Fernverkehr	200	60	40	300
Kfz-Fahrer und Sonstige	100	100	200	400
	320	430	250	1000

(a) $\frac{320}{1000} = 0.32$ finden das neue System besser, während $\frac{430}{1000} = 0.43$ das alte System bevorzugen.

(b) (i)

$$\theta \in f \mp z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \tag{544}$$

$$\theta \in 0.43 \mp 1.96\sqrt{\frac{0.43(1-0.43)}{1000}} \tag{545}$$

$$KI \in [0.3993, 0.4607] \tag{546}$$

(ii) $f = \frac{270}{300} = 0.9$

$$\theta \in f \mp z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \tag{547}$$

$$\theta \in 0.9 \mp 1.96\sqrt{\frac{0.9(1-0.9)}{1000}} \tag{548}$$

$$KI \in [0.8814, 0.9186] \tag{549}$$

(iii) $f = \frac{60}{300} = 0.2$

$$\theta \in f \mp z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \tag{550}$$

$$\theta \in 0.2 \mp 1.96\sqrt{\frac{0.2(1-0.2)}{1000}} \tag{551}$$

$$KI \in [0.1752, 0.2248] \tag{552}$$

(c) a) H_0: $\theta_0 \leq 0.4$

b) Testvariable $Z = \frac{f-\theta_0}{\sqrt{\theta_0(1-\theta_0)}}\sqrt{n}$

c) Realisation

$$Z = \frac{0.43-0.4}{\sqrt{0.4(1-0.4)}}\sqrt{1000} \tag{553}$$

$$Z = 1.9365 \tag{554}$$

Vergleiche mit $z_{1-\alpha} = z_{0.95} = 1.65$

d) Ergebnis: Die Nullhypothese kann verworfen werden, da $1.9365 > 1.65$

e) Zur Bestimmung der Grenzwahrscheinlichkeit wird ein neues α gesucht. $\alpha = 1 - \phi(1.94) = 1 - 0.9738 = 0.0262$ Also ab einer Fehlerwahrscheinlichkeit von 2.62% kann die Nullhypothese abgelehnt werden.

(d) Gegeben sei also $\alpha = 0.05$, $f = 0.4$ und ein Anteilsfehler von 2%.

$$0.02 = z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \tag{555}$$

$$n = \frac{z_{1-\alpha/2}^2 \cdot f(1-f)}{0.02^2} \tag{556}$$

$$n = \frac{1.96^2 \cdot 0.4 \cdot 0.6}{0.02^2} \tag{557}$$

$$n = 2305 \tag{558}$$

(e) Systematische Stichprobe oder Quotenverfahren o. ä.

Lösung T.4.20

(a) Nullhypothese H_0: "Motorleistung über 80 KW". Testvariable bei Gaußverteilung mit unbekannter Varianz:

$$T = \frac{\bar{x} - \mu_0}{S}\sqrt{10} \tag{559}$$

mit $\mu_0 = 80$ (KW), $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 78$, und $S^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{144}{9} = 16$. Also ist hier die Realisierung t von T:

$$t = \frac{78 - 80}{\sqrt{16}}\ \sqrt{10} = -1.581. \tag{560}$$

Die Nullhypothese kann nicht abgelehnt werden, falls

$$t \geq t_{\alpha}^{(n-1)} = -t_{1-\alpha}^{(n-1)} = -t_{0.95}^{(9)} \overset{\text{Tabelle}}{=} -1.833 \tag{561}$$

mit $t_{0.95}^{(9)}$ dem $(1-\alpha) = 95\%$-Quantil der t-Verteilung mit 9 Freiheitsgraden. $\Rightarrow$ Nullhypothese kann nicht abgelehnt werden!

(b) (i) $P(X < x) = F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{80-82}{\sqrt{16}}\right) = \Phi(-0.5)$

mit $\Phi(x)$ der tabellierten Standardnormalverteilung. Also

$$P(X < 80) = \Phi(-0.5) = 1 - \Phi(0.5) \overset{\text{Tabelle}}{=} 1 - 0.6915 = 0.3085 \tag{562}$$

(ii) $\Phi(\frac{86-82}{4}) = \Phi(1)$ und $\Phi(\frac{80-82}{4}) = \Phi(-0.5)$ und damit $P(80 \leq X \leq 86) = F(86) - F(80) = \Phi(1) - \Phi(-0.5)$

$= \Phi(1) - (1 - \Phi(0.5)) = 0.8413 - (1 - 0.6915) = 0.5328$

(c) Mit dem nichtparametrischen χ^2-Unabhängigkeitstest.

(d) Falls die Merkmale X (Motorleistung) und Y (Benzinverbrauch) unabhängig sind, ist die Wahrscheinlichkeit, dass gleichzeitig eine Bedingung A für Merkmal X und eine Bedingung B für Merkmal Y eintrifft, durch das Produkt gegeben:

$$p(A \cap B) = p(A)p(B) \tag{563}$$

(i) $p(A) = P(X \geq 80) = \Phi(0.5) = 0.6915$,
$p(B) = P(Y \leq 7) = \Phi(0) = 0.5 \Rightarrow \Phi(\frac{7-7}{0.5}) = \Phi(0)$,
$p(A \cap B) = p(A)p(B) = 0.3458$.

(ii) Der Motor erfüllt mindestens eine der Bedingungen A: Leistung $\geq$ 80 KW und B: Verbrauch $\leq$ 7l/(100 km), wenn nicht gleichzeitig die Leistung kleiner 80 KW UND der Verbrauch größer als 7 l/(100 km) ist:

$$P = 1 - P(C)P(D) \tag{564}$$

$P(C) = P(\overline{A}) = P(X < 80) = 1 - \Phi(0.5) = 1 - 0.6915 = 0.3085$,
$P(D) = P(\overline{B}) = P(Y > 7) = 0.5$,
$P = 1 - P(C)P(D) = 1 - 0.3085 \cdot 0.5 = 0.8458$.

Oder mit den De-Morgan'schen Sätzen ($\bar{A}$ ist die Negation von A):

$$P(A \cup B) = P(\overline{\overline{A} \cap \overline{B}}) = 1 - P(\overline{A} \cap \overline{B}) = 1 - p(\overline{A})p(\overline{B}). \tag{565}$$

Lösung T.4.21

Lösungsvorschlag folgt.

Lösung T.4.22

(a) Mit der Arbeitstabelle

i	x_i	$(x_i - \bar{x})^2$
1	50	100
2	70	100
3	60	0
4	70	100
5	70	100
6	50	100
7	50	100
$\sum_i$	420	600

ergeben sich als Punktschätzer für den Mittelwert

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{566}$$

$$= \frac{420}{7} \tag{567}$$

$$= 60 \tag{568}$$

und für die Varianz

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{569}$$

$$= \frac{600}{6} \tag{570}$$

$$= 100 \tag{571}$$

(b) Richtig ist $H_0 : \mu < 50$

Fehler 1. Art: Die Maßnahmen werden nicht implementiert, obwohl die Fahrtenzahl gering ist.

Fehler 2. Art: Es wird umstrukturiert und eingespart – allerdings umsonst, denn die Fahrtenzahl ist hoch.

Der Fehler 1. Art ist schwerwiegender, weil er das Taxiunternehmen in den Konkurs führt.

(c) Test des Mittelwertes bei unbekannter Varianz

1. Nullhypothese:

$$H_0 : \mu < \mu_0 = 50 \tag{572}$$

2. Testvariable:

$$T = \frac{\bar{X} - \mu_0}{s}\sqrt{n} \quad \sim T(n-1) \tag{573}$$

3. Realisierung:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} \tag{574}$$
$$= \frac{60 - 50}{\sqrt{100}}\sqrt{7} \tag{575}$$
$$= 2.646 \tag{576}$$

4. Entscheidung: H_0 ablehnen, da

$$t > t_{1-\alpha}^{(n-1)} \tag{577}$$
$$2.646 > t_{0.95}^{(6)} \tag{578}$$
$$2.646 > 1.943 \tag{579}$$

(d) Mögliche richtige Antworten sind:

- Falls die mittlere Fahrtenzahl tatsächlich kleiner als 50 ist, dann sind Fahrtenzahlen, wie sie vergangene Woche beobachtet wurden, äußerst unwahrscheinlich (kleiner als 5%).

- Nicht exakt, aber dennoch richtig: Man kann mit 95%iger Sicherheit davon ausgehen, dass die mittlere tägliche Fahrtenzahl größer (oder gleich) 50 ist.

(e) χ^2-Anpassungstest

1. Nullhypothese:

$$H_0 : \text{„Die Fahrten verteilen sich gleichmäßig auf die sieben Wochentage“} \tag{580}$$

2. Testvariable:

$$Q = \left(\sum_{k=1}^{K} \frac{h_k^2}{h_k^e}\right) - n \quad \sim \chi^2(m) \tag{581}$$

3. Realisierung: Mit

$$n = 420 \tag{582}$$
$$K = 7 \tag{583}$$
$$r = 0 \tag{584}$$
$$m = K - 1 - r = 6 \tag{585}$$
$$\alpha = 5\% \tag{586}$$

und der Arbeitstabelle

k	h_k	h_k^e	h_k^2/h_k^e
1	50	60	41.67
2	70	60	81.67
3	60	60	60.00
4	70	60	81.67
5	70	60	81.67
6	50	60	41.67
7	50	60	41.67
$\sum_k$	420		430.00

ergibt sich

$$q = \left(\sum_{k=1}^{K} \frac{h_k^2}{h_k^e} \right) - n \tag{587}$$
$$= 430 - 420 \tag{588}$$
$$= 10 \tag{589}$$

4. Entscheidung: H_0 ablehnen, falls

$$q > q_{1-\alpha}^{(m)} \tag{590}$$
$$10 > q_{0.95}^{(6)} \tag{591}$$
$$10 > 12.592 \tag{592}$$

was nicht der Fall ist. Also kann H_0 nicht abgelehnt werden.

(f) Mögliche richtige Antworten sind:

- Aus dem Nicht-Ablehnen von H_0 können keine haltbaren Schlüsse gezogen werden. Wie wahrscheinlich H_0 zutrifft (oder nicht) kann nicht eingegrenzt werden.
- Falls sich die Fahrten tatsächlich gleichmäßig auf die Wochentage verteilen, so liegt die Stichprobe im Bereich des „Normalen“.

Falsch wäre es, auf das Vorliegen einer anderen als der Gleichverteilung zu schließen.

Lösung T.4.23

(a) Der ÖV-Anteil der Befragten ist

$$f = h_2/n \tag{593}$$
$$= 54/200 \tag{594}$$
$$= 27\%. \tag{595}$$

(b) Das Konfidenzintervall des ÖV-Anteils lässt sich mit

$$\theta \in f \pm z_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}\sqrt{\frac{N-n}{N-1}} \qquad \Big| N \to \infty \tag{596}$$
$$\in f \pm z_{0.975}\sqrt{\frac{f(1-f)}{n}} \tag{597}$$
$$\in 27\% \pm 1.960 \cdot 0.0314 \tag{598}$$
$$\in 27\% \pm 6.2\% \tag{599}$$
$$\in \big[20.8\%, 33.2\%\big] \tag{600}$$

berechnen.

(c) Das ist kein Widerspruch, weil das Konfidenzintervall Abweichungen vom emprisch ermittelten Anteilswert symmetrisch zu beiden Seiten zulässt, der Test „auf kleiner" aber einseitig testet.

(d) χ^2-Unabhängigkeitstest

1. Nullhypothese

$$H_0 : \text{„Reiseweite und Verkehrsmittelwahl sind unabhängig."} \tag{601}$$

2. Testvariable

$$Q = \sum_{i=1}^{K_x}\sum_{j=1}^{K_y}\frac{h_{ij}^2}{h_{ij}^e} - n \sim \chi^2\big(m = (K_x-1)(K_y-1)\big) \tag{602}$$

3. Realisation: Mit

$$h_{ij}^e = \frac{h(x_i)h(y_j)}{n} \tag{603}$$

und den Arbeitstabellen

h_{ij}^e	MIV	ÖV	$\sum_j$
1	82.49	30.51	113
2	40.15	14.85	55
3	23.36	8.64	32
$\sum_i$	146	54	200

h_{ij}^2/h_{ij}^e	MIV	ÖV	$\sum_j$
1	77.59	35.69	113.28
2	34.10	21.82	55.92
3	36.00	1.04	37.04
$\sum_i$	147.68	58.55	206.24

erhält man

$$q = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{h_{ij}^2}{h_{ij}^e} - n \tag{604}$$
$$= 206.24 - 200 \tag{605}$$
$$= 6.24 \tag{606}$$
$$\tag{607}$$

4. Entscheidung: H_0 ablehnen, falls

$$q \geq q_{1-\alpha}\big(m = (3-1)(2-1)\big) \tag{608}$$
$$q \geq q_{0.95}(2) \tag{609}$$
$$6.24 \geq 5.991 \tag{610}$$

was hier der Fall ist.

(e) Mögliche richtige Antworten sind:

- Falls die Verkehrsmittelwahl tatsächlich von der Reiseweite unabhängig ist, wären die beobachteten Zahlen äußerst unwahrscheinlich (kleiner als 5%).
- Nicht exakt, aber dennoch richtig: Man kann mit 95%iger Sicherheit davon ausgehen, dass die Verkehrsmittelwahl von der Reiseweite abhängig ist.

(f) χ^2-Anpassungstest

1. Nullhypothese

$$H_0 : \text{„Die Reiseweite ist exponentialverteilt.“} \tag{611}$$

2. Testvariable

$$Q = \sum_{k=1}^{K} \frac{h_k^2}{h_k^e} - n \sim \chi^2(m = K - 1 - r) \tag{612}$$

3. Realisation. Mit

$$F_E(x) = 1 - e^{-\lambda x} \tag{613}$$

und

$$h_k^e = n \cdot \left[F_E(x_k^o) - F_E(x_k^u)\right] \tag{614}$$

erhält man über die Arbeitstabelle

k	x_k^u	x_k^o	h_k	$F_E(x_k^u)$	$F_E(x_k^o)$	h_k^e	h_k^2/h_k^e
1	0	50	113	0.000	0.565	113.00	113.00
2	50	100	55	0.565	0.811	49.20	61.48
3	100	∞	32	0.811	1.000	37.80	27.09
$\sum_k$			200			200.00	201.57

als Realisation

$$q = \sum_{k=1}^{K} \frac{h_k^2}{h_k^e} - n \tag{615}$$
$$= 201.57 - 200 \tag{616}$$
$$= 1.57 \tag{617}$$

4. Entscheidung: H_0 ablehnen, falls (mit $r = 0$ geschätzten Parametern)

$$q \geq q_{1-\alpha}\big(m = 3 - 1 - 0\big) \tag{618}$$
$$q \geq q_{0.95}(2) \tag{619}$$
$$1.57 \geq 5.991 \tag{620}$$

was hier nicht der Fall ist.

(g) Mögliche richtige Antworten sind:

- Aus dem Nicht-Ablehnen von H_0 können keine haltbaren Schlüsse gezogen werden. Wie wahrscheinlich H_0 zutrifft (oder nicht) kann nicht eingegrenzt werden.
- Falls die Reiseweiten tatsächlich exponentialverteilt sind, so liegt die Stichprobe im Bereich des „Normalen“.