

**Klausur zur Vorlesung Statistik I und II,  
WS 2008/09  
für Diplom-Studenten  
Lösungsvorschlag**

**Aufgabe 1**

**(40 Punkte)**

*Um die wesentlichen Beeinflussungsfaktoren für die Passagier- und Güterverkehrsleistung deutscher Flughäfen zu bestimmen, wird eine Bestandsaufnahme relevanter Eigenschaften der verschiedenen Flughäfen aufgrund der letzten verfügbaren Statistiken erstellt. Insbesondere Zahl der Start- und Landebahnen und ihre Länge, ob es ein Nachtflugverbot gibt, die Wirtschaftskraft und Zahl der Hotelbetten im Einzugsgebiet (Region im 100 km bzw 1:30 Kfz-h-Umkreis), Größe und Entfernung des nächsten Konkurrenz-Flughafens, abgefertigte Passagierzahlen, abgefertigte Gütermenge, ob sich eine Hauptstadt im Einzugsgebiet befindet, ob eine Fluggesellschaft den Flughafen als Heimatflughafen benutzt und wie hoch die Start/Landegebühren sind.*

- (a) *Geben Sie die statistische Gesamtheit und den Merkmalsträger an. Grenzen Sie die statistische Gesamtheit ab.*

Merkmalsträger: Ein Verkehrsflughafen; Stat. Gesamtheit: Alle Flughäfen

- in Deutschland (räumlich)
- zur Zeit der “letzten verfügbaren Statistiken” (zeitlich)
- auf denen es Passagier- oder Güterverkehr gibt, also Verkehrsflughäfen, nicht Sportflughäfen etc. (sachlich).

- (b) *Geben Sie von allen oben genannten Merkmalen die Skalierung an.*

- Zahl der Start- und Landebahnen: Kardinal-absolut
- Länge: Kardinal
- Nachtflugverbot: Nominal. Außerdem dichotom (binär).
- Wirtschaftskraft, Hotelbetten: Kardinal
- Größe und Entfernung des nächsten Konkurrenz-Flughafens, abgefertigte Passagierzahlen, abgefertigte Gütermenge: Kardinal
- Hauptstadt im Einzugsgebiet: Nominal-dichotom
- Heimatflughafen einer Fluggesellschaft: Nominal-dichotom
- Start-,Landegebühren: Kardinal

Mit kardinal-verhältnisskaliert, diskret bzw. (quasi-)stetig etc kann man Pluspunkte gewinnen.

- (c) *Die Analyse der Beeinflussungsfaktoren soll mit einer multivariaten Regressionsanalyse durchgeführt werden. Welche zwei abhängigen Variablen sind dabei sinnvoll?*  
abgefertigte Passagierzahlen und Gütermenge

- (d) Stellen Sie nun für den Passagierverkehr eine multivariate lineare Regressionsfunktion auf, welche mindestens fünf unabhängige Variable enthält.

$$\hat{y}(x) = \sum_j \beta_j x_j$$

mit z.B.  $x_1$ =Zahl der Start- und Landebahnen,  $x_2$ =Länge (m) der längsten Bahn,  $x_3 \in \{0, 1\}$  mit 0=Nachtflugverbot, 1=kein Nachtflugverbot  $x_4$ =Wirtschaftskraft (€/Jahr) der Stadt,  $x_5$ =Zahl der Hotelbetten ...

- (e) *Wie könnte man nominalskalierte Variable wie die Existenz eines Nachtflugverbots in die Schätzfunktion der linearen Regression einbringen?*

Durch Abbildung auf eine Pseudovariablen mit den Werten 0 und 1, welche z.B. "Nachtflugverbot" bzw. "kein Nachtflugverbot" zugeordnet werden.

## Aufgabe 2

(60 Punkte)

Auf einem mit Tempolimit 120 versehenen Autobahnabschnitt ergaben Induktionsschleifendetektoren während eines Tages die in der Aufgabenstellung angegebenen Zählergebnisse.

- (a) *Wieviel Fahrzeuge wurden insgesamt erfasst? Welchem Verkehrsfluss (Kfz/h) entspricht das im Mittel?*

$$n = \sum_{k=1}^9 h_k = 800 + 3\,300 + \dots = \underline{\underline{18\,350 \text{ Kfz}}}.$$

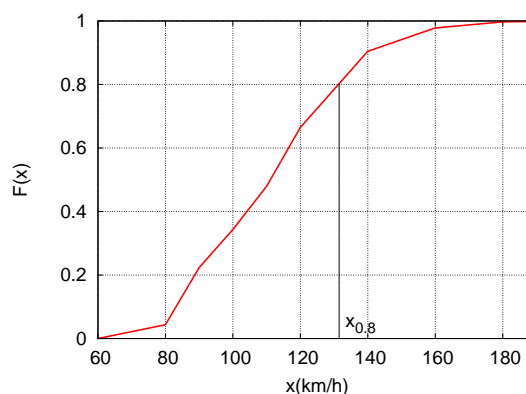
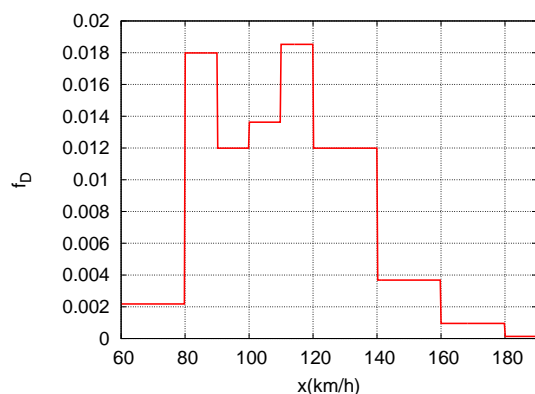
Mittlerer Fluss in Kfz/h:

$$Q = \frac{n}{24 \text{ h}} = \underline{\underline{765 \text{ Kfz/h}}}.$$

- (b) *Bestimmen Sie die relativen Häufigkeiten, Summenhäufigkeiten und die empirische Dichtefunktion für jede Klasse.*

Klasse (km/h)	$f_k$	$F_k$	$f_k^D$
1: 60-80	0.0436	0.0436	0.00218
2: 80-90	0.180	0.223	0.0180
3: 90-100	0.120	0.343	0.0120
4: 100-110	0.136	0.480	0.0136
5: 110-120	0.185	0.665	0.0185
6: 120-140	0.240	0.905	0.0120
7: 140-160	0.0736	0.978	0.0037
8: 160-180	0.0191	0.997	0.0010
9: 180-200	0.00272	1.0	0.0001

- (c) *Zeichnen Sie die Dichte- und Verteilungsfunktion in die Diagramme ein. Ist die Verteilung bimodal? Wenn ja, was könnte der Grund sein?*



Verteilung ist bimodal; PKW- und "LKW-Peak".

- (d) *Bestimmen Sie das arithmetische Mittel, den Median und den bzw. die Modi (mit Feinberechnung). Zeichnen Sie in die Grafik das 80. Perzentil ein.*

$$\bar{x} = \sum_{k=1}^9 f_k x_k^* = \underline{\underline{111.3 \text{ km/h}}}.$$

Median in Klasse 5. In km/h:

$$x_{0.5} = x_5^u + \Delta x \frac{0.5 - f_4}{f_5} = \underline{\underline{111.1}}$$

Es gibt zwei Modi, einen in Klasse 2 und einen in Klasse 5. Feinberechnung:

$$x_{\text{mod},1} = x_2^u + \Delta x_2 \frac{f_2^D - f_1^D}{2f_2^D - f_1^D - f_3^D} = \underline{\underline{97.3}},$$

$$x_{\text{mod},2} = x_5^u + \Delta x_5 \frac{f_5^D - f_4^D}{2f_5^D - f_4^D - f_6^D} = \underline{\underline{114.3}}$$

- (e) Bestimmen Sie Varianz und die Standardabweichung und zeichnen Sie einen Boxplot dieser Verteilung.

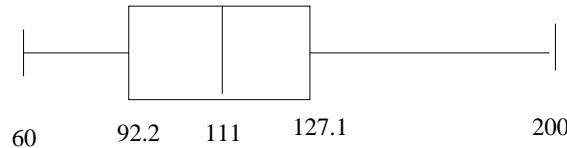
$$V(x) = s_x^2 = \sum_{k=1}^9 f_k (x_k^* - \bar{x})^2 = \underline{\underline{515}}, \quad s_x = \sqrt{s_x^2} = \underline{\underline{22.7}}$$

Boxplot:

$$x_{\min} = 60, \quad x_{\max} = 200, \quad x_{0.5} = 111.1,$$

sowie

$$x_{0.25} = x_3^u + \Delta x_3 \frac{0.25 - f_2}{f_3} = 92.2, \quad x_{0.75} = x_6^u + \Delta x_6 \frac{0.75 - f_5}{f_6} = 127.1$$



- (f) Am Messquerschnitt soll nun ein "Blitzer" installiert werden. Mit wieviel Einnahmen kann man (zumindest anfangs) pro Tag rechnen, wenn nach folgender (vereinfachter aber aktueller) Gebührentabelle "abgerechnet" wird und jeweils 3 km/h Toleranz gegeben wird? Warum werden durch die Gleichverteilungsannahme innerhalb der Klassen die Einnahmen überschätzt?

Überschreitung um (km/h)	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	> 70
€	30	80	120	160	240	440	600

Gegeben sind Strafkostenklassen  $j$  mit Strafkosten  $S_j$  innerhalb der dazugehörigen Geschwindigkeitsbereiche  $[\tilde{x}_j^u, \tilde{x}_j^o]$ , also in Euro bzw. km/h und mit Berücksichtigung der Toleranz:

$$\begin{aligned} S_1 &= 30, & \tilde{x}_1^u &= 133, & \tilde{x}_1^o &= 143 \\ S_2 &= 80, & \tilde{x}_2^u &= 143, & \tilde{x}_2^o &= 153 \\ &\dots & \dots & \dots & & \\ S_7 &= 600, & \tilde{x}_7^u &= 193, & \tilde{x}_7^o &= 200 \text{ (bzw. } \infty) \end{aligned}$$

Der Erwartungswert der Einnahmen pro Fahrzeug ist also gegeben durch

$$E(S) = \sum_{j=1}^7 S_j \left( F(\tilde{x}_j^o) - F(\tilde{x}_j^u) \right)$$

und mit  $F(\tilde{x}_1^o) - F(\tilde{x}_1^u) = 7f_6^D + 3f_7^D$ ,  $F(\tilde{x}_2^o) - F(\tilde{x}_2^u) = 10f_7^D$ , ..., schließlich

$$E(S) = S_1 \left( 7f_6^D + 3f_7^D \right) + S_2 \left( 10f_7^D \right) + S_3 \left( (7f_7^D + 3f_8^D) \right) + \dots + S_7 \left( 7f_9^D \right) = \underline{\underline{13.6}}.$$

Die täglichen Einnahmen (welche sehr schnell abnehmen!) sind also

$$nE(S) = \underline{\underline{250\,000\text{ €}}}.$$

**Aufgabe 3****(30 Punkte)**

Zwischen Deutschland und China soll die Kaufkraftparität (Yuán pro Euro) anhand folgender deutscher und chinesischer monatlicher Warenkörbe ermittelt werden:

Gut	Sack Reis	Big Mac	Miete	Laptop	Bahnfahrt
Preis D	8 €	3.50 €	400 €/Mon.	1 000 €	0.20 €/km
Menge D	0.1	5	1 Monat	0.05	400 km
Preis China	15 Yuán	25 Yuán	700 Yuán/Mon.	5 000 Yuán	0.20 Yuán/km
Menge China	1	1	1 Monat	0.02	1 000 km

- Geben Sie die Kaufkraftparität bezüglich des deutschen Warenkorbs an und vergleichen Sie mit dem offiziellen Wechselkurs (5.5 Yuán/€)

Preis des deutschen Warenkorbes in Deutschland:

$$P(W_D, D) = \sum_{i=1}^5 p_i(D)q_i(D) = 8 \text{ €} * 0.1 + \dots = 548.3 \text{ €}$$

Preis des deutschen Warenkorbes in China:

$$P(W_D, C) = \sum_{i=1}^5 p_i(C)q_i(D) = 15 \text{ Yuan} * 0.1 + \dots = 1156.5 \text{ Yuan}$$

Preis des chinesischen Warenkorbes in Deutschland:

$$P(W_C, D) = \sum_{i=1}^5 p_i(D)q_i(C) = 8 \text{ €} * 1 + \dots = 631.5 \text{ €}$$

Preis des chinesischen Warenkorbes in China:

$$P(W_C, C) = \sum_{i=1}^5 p_i(C)q_i(C) = 15 \text{ Yuan} * 1 + \dots = 1040 \text{ Yuan}$$

Damit Kaufkraftparität bezüglich des deutschen Warenkorbes:

$$K_D = \frac{P(W_D, C)}{P(W_D, D)} = \underline{\underline{2.11 \text{ Yuán/€}}}$$

- Ermitteln Sie nun die Kaufkraftparität bezüglich des chinesischen Warenkorbs.

Kaufkraftparität bezüglich des chinesischen Warenkorbes:

$$K_C = \frac{P(W_C, C)}{P(W_C, D)} = \underline{\underline{1.65 \text{ Yuán/€}}} \text{ bzw. } \underline{\underline{0.61 \text{ €/Yuán}}}$$

- Nach einer These des Wirtschaftswissenschaftlers Balassa haben in Entwicklungs- und Schwellenländern nicht-handelbare Güter einen niedrigeren Preis als es dem Wechselkurs entspricht, während er bei ideal (ohne Kosten) handelbaren Gütern in etwa gleich dem Wechselkurs ist. Trifft diese These hier zu?

In etwa schon: Nicht handelbar sind Miete und Bahnfahrt, da sie auf lokalen Immobilien bzw. Infrastrukturen beruhen. Bei ihnen ist der Preis in China, bezogen

auf den Wechselkurs und die Preise in Deutschland, günstig. Mit geringen Kosten handelbar sind Laptop und Big-Mac: Preis ungefähr konsistent mit Wechselkurs. Der Sack Reis ist handelbar, aber (da schwer und billig) mit vergleichsweise hohen Transportkosten, so dass sein in China bezogen auf den Kurs günstiger Preis nicht der Theorie widerspricht.

- *Um die Wechselkurse aller Länder beurteilen zu können, hat der "New Economist" den Big-Mac-Index eingeführt, bei dem der Warenkorb zur Berechnung der Kaufkraftparität einheitlich aus einem (überall erhältlichen) Big Mac besteht. Nennen Sie einen Vor- und einen Nachteil dieses Index!*

Vorteil: Kaufkraftparitäten bezüglich beliebiger Währungskombinationen berechenbar.

Nachteil: Sehr einseitig und verfälschend. In einigen Ländern ist der Big-Mac ein bedeutender Teil des Warenkorbes, in anderen nicht. Außerdem von einem Konzern abhängig. Die damit verbundenen undurchschaubaren Mischkalkulationen und mögliche Quer- oder Anschub-Subventionen verfälschen das Ergebnis.

## Aufgabe 4

(50 Punkte)

In dieser Aufgabe werden Fahrzeuge betrachtet, welche an einen festen Streckenquerschnitt vorbeifahren und dort durch Induktionsschleifen detektiert werden.

- (a) Bei geringem Verkehrsaufkommen ist die Zeitlücke zwischen zwei Fahrzeugen annähernd exponentialverteilt. Bestimmen Sie den Parameter dieser Verteilung als Funktion des Verkehrsflusses  $Q$  (Fahrzeuge pro Zeiteinheit).

Die Exponentialverteilung mit der Dichte

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

hat einen einzigen Parameter  $\lambda$ . 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

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

- (b) Gegeben ist nun ein Fluss von 1 Fahrzeug pro Minute. Mit welcher Wahrscheinlichkeit kommen in den nächsten 5 Minuten höchstens zwei Fahrzeuge vorbei? Wie hoch ist der Median des zeitlichen Abstandes zwischen zwei Fahrzeugen?

Da der zeitliche Abstand zweier Fahrzeuge nach Aufgabenstellung exponentialverteilt ist, ist die Fahrzeugzahl  $Y$  selbst poissonverteilt,  $Y \sim \text{Po}(\mu)$ . 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

$$\begin{aligned} P(Y \leq 2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= e^{-\mu} \left( 1 + \mu + \frac{\mu^2}{2!} \right) = \underline{\underline{12.5\%}}. \end{aligned}$$

Median:

$$\begin{aligned} F(x) = 1 - e^{-\lambda x} &\stackrel{!}{=} \frac{1}{2} \\ \frac{1}{2} &= e^{-\lambda x} \\ \ln 2 &= \lambda x \end{aligned}$$

also

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

- (c) Nehmen Sie nun an, dass die Geschwindigkeit aller Fahrzeuge genau  $72 \text{ km/h}$  bzw.  $20 \text{ m/s}$  beträgt. Wie lautet die Verteilungsfunktion der räumlichen Fahrzeugabstände, wenn das Verkehrsaufkommen nach wie vor 1 Fahrzeug pro Minute beträgt?

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

$$Z = aX = 20X$$

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

$$E(Z) = E(aX) = aE(X) = \frac{a}{\lambda} \stackrel{!}{=} \frac{1}{\lambda_z} \quad \Rightarrow \quad \lambda_z = \frac{\lambda}{a} = \frac{1}{60 \text{ s} * 20 \text{ m/s}} = \underline{\underline{\frac{1}{1200 \text{ m}}}}$$

- (d) Für höhere Verkehrsaufkommen (dichter Verkehr) und Richtungsfahrbahnen mit nur einem Fahrstreifen ist die Exponentialverteilung zur Beschreibung der Abstände nicht geeignet. Ermitteln Sie dazu den aus ihr folgenden wahrscheinlichsten (zeitlichen) Abstand und erläutern Sie den Sachverhalt.

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) Um einen zeitlichen Mindestabstand  $x_{\min}$  zu berücksichtigen, wird die Verteilung modifiziert. Die Dichtefunktion der neuen Verteilung ist gegeben durch

$$f(x) = \begin{cases} 0 & x < x_{\min} \\ \tilde{\lambda} e^{-\tilde{\lambda}(x-x_{\min})} & x \geq x_{\min} \end{cases} .$$

Ermitteln Sie  $\tilde{\lambda}$  in Abhängigkeit des Verkehrsaufkommens  $Q$  so, dass der mittlere zeitliche Abstand  $E(X)$  nach wie vor durch den Kehrwert  $1/Q$  des Verkehrsflusses gegeben ist.

Ausrechnen von  $E(X)$ :

$$\begin{aligned} E(X) &= \int_0^{\infty} x f(x) dx \\ &= \int_{x_{\min}}^{\infty} \tilde{\lambda} x e^{-\tilde{\lambda}(x-x_{\min})} dx \\ &\stackrel{x=y+x_{\min}}{=} \int_0^{\infty} (x_{\min} + y) \tilde{\lambda} e^{-\tilde{\lambda}y} dy \\ &= x_{\min} + \int_0^{\infty} \tilde{\lambda} y e^{-\tilde{\lambda}y} dy \\ &= x_{\min} + \frac{1}{\tilde{\lambda}} \end{aligned}$$

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

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

(f) Beachten Sie nun die beiden Werte  $Q_1 = 1$  Kfz/min und  $Q_2 = 30$  Kfz/min des Verkehrsaufkommens (geringer bzw. starker Verkehr). Für den Parameter  $\tilde{\lambda}_1$  der Verteilung von Aufgabenteil (e) gilt dann  $\tilde{\lambda}_1 = 1.01695$  (geringer Verkehr) bzw.  $\tilde{\lambda}_2 = 60$  (starker Verkehr). Beurteilen Sie die Auswirkung der Modellierung des Mindestabstandes für beide Fälle, indem Sie jeweils

- (i) den Wert des Parameters  $\tilde{\lambda}$  mit dem entsprechenden Wert aus der Exponentialverteilung vergleichen,
- (ii) den Median der beiden Verteilungen vergleichen.

Für welches der beiden Verkehrsaufkommen  $Q_1$  und  $Q_2$  fällt der Unterschied zwischen der ursprünglichen und der modifizierten Exponentialverteilung geringer aus?

- (i) Schwacher Verkehr:  $\lambda_1 = Q_1 = 1$ ,  $\tilde{\lambda}_1 = 1.01695$ : Kaum ein Unterschied.  
Starker Verkehr:  $\lambda_2 = Q_2 = 30$ ,  $\tilde{\lambda}_2 = 60$ : Unterschied um Faktor zwei!
- (ii) Median der ursprünglichen Verteilung (vgl. Teil (b)):

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

Median der modifizierten Verteilung:

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

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

$$\begin{aligned} Q = Q_1 = 1/(60 \text{ s}) &\Rightarrow x_{0.5} = 41.6 \text{ s} = 0.693 \text{ min}, & \tilde{x}_{0.5} = 41.9 \text{ s} = 0.698 \text{ min} \\ Q = Q_2 = 1/(2 \text{ s}) &\Rightarrow x_{0.5} = 1.39 \text{ s} = 0.0231 \text{ min}, & \tilde{x}_{0.5} = 1.69 \text{ s} = 0.028 \text{ min} \end{aligned}$$

also ebenfalls nur beim hohen Fluss ein deutlicher Unterschied.

## Aufgabe 5

(30 Punkte)

Bei einer Erhebung zum Mobilitätsverhalten soll auch das Einkommen als relevante Einflussgröße ermittelt werden. Da es heikel ist, danach direkt zu fragen, wird das Einkommen indirekt durch Erhebung von mit dem Einkommen korrelierten, aber unproblematischeren Merkmalen abgeschätzt. Konkret wird nach der Zahl der im Haushalt verfügbaren Autos gefragt. Aus Volkszählendaten ist folgendes bekannt:

- In der Einkommensklasse 1 (unter 30 000 €/Jahr) besitzen 40% kein Kfz, 55% ein Kfz und nur 5% mehr als 1 Kfz. In der Einkommensklasse 2 (30 000 bis 50 000 €/Jahr) ist die Aufteilung 20% (kein Kfz), 60% (ein Kfz) und 20% (mehrere), während in der Einkommensklasse 3 (>50 000 €/Jahr) 5% (kein Kfz), 45% (ein Kfz) und 50% (mehrere) gilt.
- 20% der Haushalte gehören zu Einkommensklasse 1, 60% zu Klasse 2 und der Rest zu Klasse 3.

Wie hoch sind die Wahrscheinlichkeiten dafür, dass (i) Haushalte ohne Kfz, (ii) Haushalte mit einem Kfz, (iii) Haushalte mit mehreren Kfz jeweils zu den Einkommensklassen 1, 2 oder 3 gehören?

Wir kategorisieren die Einkommensklassen mit  $A_i$  und die Zahl der Kfz mit  $B_j$ :

- $A_i \Leftrightarrow$  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

$$P(A_1) = 0.2, \quad P(A_2) = 0.6, \quad P(A_3) = 0.2.$$

- Bedingte Wahrscheinlichkeiten

$$\begin{aligned} P(B_1|A_1) &= 0.40 & P(B_2|A_1) &= 0.55, & P(B_3|A_1) &= 0.05, \\ P(B_1|A_2) &= 0.20 & P(B_2|A_2) &= 0.60, & P(B_3|A_2) &= 0.20, \\ P(B_1|A_3) &= 0.05 & P(B_2|A_3) &= 0.45, & P(B_3|A_3) &= 0.50. \end{aligned}$$

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

$$\begin{aligned} P(B_1) &= P(B_1|A_1)P(A_1) + P(B_1|A_2)P(A_2) + P(B_1|A_3)P(A_3) = \underline{0.21}, \\ P(B_2) &= P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2) + P(B_2|A_3)P(A_3) = \underline{0.56}, \\ P(B_3) &= P(B_3|A_1)P(A_1) + P(B_3|A_2)P(A_2) + P(B_3|A_3)P(A_3) = \underline{0.23}. \end{aligned}$$

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

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

Die Zahlenwerte eingesetzt:

$$\begin{aligned}P(A_1|B_1) &= 0.381 & P(A_2|B_1) &= 0.571, & P(A_3|B_1) &= 0.048, \\P(A_1|B_2) &= 0.196 & P(A_2|B_2) &= 0.643, & P(A_3|B_2) &= 0.161, \\P(A_1|B_3) &= 0.043 & P(A_2|B_3) &= 0.522, & P(A_3|B_3) &= 0.435.\end{aligned}$$

Das Ergebnis zeigt die erwartete Tendenz, dass Haushalten mit mehreren Kfz mit höherer Wahrscheinlichkeit eine obere Einkommensklasse besitzen als die mit keinem Fahrzeug.  
*Hinweis:* Mit dem Wahrscheinlichkeitsbaum (3 Äste, je 3 Zweige) ist die Lösung ebenfalls im Standardverfahren erhältlich.

**Aufgabe 6****(40 Punkte)**

Auf einem Motorprüfstand wird die Leistung vier gleichartiger Motoren gemessen:

Motoren-Nr	1	2	3	4
Leistung (kW)	98.5	99.0	95.5	98.0

- (a) Berechnen Sie zu einer Fehlerwahrscheinlichkeit von 5% das Konfidenzintervall der mittleren Leistung

Erwartungswert-Schätzung bei unbekannter Varianz  $\Rightarrow$  t-Statistik.

Mittelwertschätzer:

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

Varianzschätzer:

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

Quantil der Student-Statistik aus der beigelegten Tabelle:

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

Halbe Breite des Konfidenzintervalls:

$$\Delta x_\alpha = \frac{s t_{1-\alpha/2}^{(n-1)}}{\sqrt{n}} = \underline{\underline{2.473}}$$

Konfidenzintervall:

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

- (b) Der Fahrzeughersteller gibt die mittlere Leistung mit "mindestens 100 kW" an. Kann man ihn bei einer Fehlerwahrscheinlichkeit von 5% der Falschaussage bezichtigen?

Statistischer Test auf Ungleichheit.

1. Nullhypothese  $H_0: \mu \geq \mu_0 = 100$  ("Im Zweifel für den Angeklagten")
2. Test-Statistik:  $T = \sqrt{n} \frac{\bar{X} - \mu_0}{s} \sim T(n-1) = T(3)$
3. Realisierung aus der Stichprobe:  $t = 2 \frac{97.75 - 100}{1.554} = \underline{\underline{-2.895}}$
4. 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 in (a) errechnete Konfidenzintervall enthält den Wert der Nullhypothese des Aufgabenteils (b), dennoch kann man in (b) die Nullhypothese verwerfen. Erläutern Sie diese scheinbar widersprüchlichen Ergebnisse.

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) Zu welcher Fehlerwahrscheinlichkeit kann man die Nullhypothese "Leistung mindestens 100 kW" gerade noch verwerfen? Lesen Sie das Ergebnis von der Abbildung am Ende dieser Aufgabe ab! Falls Sie (b) nicht gerechnet haben, nehmen Sie einen Stichprobenwert der Test-Statistik von  $-2.8$  an.

Hier muss das Kriterium der Ablehnung grenzwertig erfüllt sein, also

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

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

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

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

$$1 - F^{(3)}(-t) = 1 - F^{(3)}(2.8) = 1 - 0.97 = 0.03 \quad \Rightarrow \quad \alpha = \underline{\underline{3\%}}$$

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

- (e) Wie groß ist die Wahrscheinlichkeit für einen "Fehler zweiter Art", wenn der tatsächliche Mittelwert der (gaußverteilten) Motorleistung  $\mu = 98$  kW beträgt und der Test wie bei (b) durchgeführt wird? Hinweis: Sie müssen hier anhand der Test-Statistik bezüglich der ursprünglichen Nullhypothese  $\mu_0 = 100$  kW den Wert  $\bar{x}_c$  von  $\bar{X}$  finden, oberhalb dem der Test fälschlicherweise angenommen wird. Tatsächlich gehorcht aber die "wahre" Statistik  $\sqrt{n}(\bar{X} - \mu)/S$  und nicht  $\sqrt{n}(\bar{X} - \mu_0)/S$  der Studentverteilung. Setzen Sie nun  $\bar{x}_c$  in die wahre Statistik ein. Verwenden Sie wieder die Grafik

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

$$\bar{x}_c = \mu_0 - t_{0.95}^{(3)} \frac{s}{\sqrt{n}} = \underline{\underline{98.2}}$$

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

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S} > t_c = \sqrt{n} \frac{\bar{x}_c - \mu}{s} = \underline{\underline{0.22}}$$

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

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

Also beträgt der Fehler zweiter Art in diesem konkreten Beispiel etwa 40%. (Im Extremfall, falls  $\mu$  nur minimal kleiner als  $\mu_0$  ist, kann der Fehler zweiter Art bis zum Wert  $1 - \alpha = 95\%$  steigen!)