

Klausur zur Vorlesung Statistik I und II, WS 2007/08

Lösungsvorschlag

Aufgabe 1

(40 Punkte)

- (a) – Merkmalsträger: Ein Wintersportort
– Statistische Gesamtheit: Alle Wintersportorte in Bayern
– Merkmale und beispielhafte Werte:
- * nominalskalierte: Zugehörigkeit zu Alpen oder Mittelgebirge (“Alpen”, “Mittelgebirge”), Existenz von Beschneiungsanlagen (“ja”, “nein”)
 - * ordinalskalierte: Schneesicherheit (“ziemlich schneesicher”), Attraktivität (“eher unattraktiv”)
 - * verhältnisskalierte: Gesamtlänge an Skipisten (72 km) und Loipen (91 km)
 - * absolutskalierte: Zahlen der Hotel- und Pensionsbetten, Zahl der Lifte (42)
- (b) Die Fahrleistung F aller 65-85-jährigen Senioren ist gegeben durch das Produkt aus der Gesamtzahl $n = pN$ an Senioren (N =deutsche Gesamtbevölkerung, p =Anteil der Altersschicht), den Führerscheinanteil α und die mittlere jährliche Fahrleistung f der Führerscheinbesitzer unter den Senioren. Dies wird noch aufgespalten nach Geschlecht $g \in \{\text{♀}, \text{male}\}$, so dass sich für die Gegenwart $t = t_0$ ergibt:

$$\begin{aligned} F(t_0) &= \sum_{g=\text{♂}, \text{♀}} N(t_0)p_g(t_0)\alpha_g(t_0)f \\ &= N(t_0)(0.11 * 0.5 + 0.09 * 0.8) f(t_0) \end{aligned}$$

Modellgestützte Prognose 20 Jahre in die Zukunft (Zeit t):

- Prognostizierte Zahl der Senioren: $n_g(t) = N(t_0)p_g^{45-65}w_g$ mit dem gegenwärtigen Bevölkerungsanteil p_g^{45-65} der 45-65-jährigen und der Überlebenswahrscheinlichkeit w_g (2/3 für $g = \text{♂}$, 87.5% für $g = \text{♀}$).
- Prognostizierter Anteil der Führerscheinbesitzer: Der Anteil verschiebt sich einfach von den 45-65-Jährigen auf die 65-85-Jährigen: $\alpha_g(t) = \alpha_g^{45-65}(t_0)$.
- Prognostizierte Fahrleistung pro Führerscheinbesitzer und Senior: Bleibt nach Aufgabenstellung konstant, $f(t) = f(t_0)$.

Damit

$$F(t) = N(t_0)(0.16 * 0.8 * 0.875 + 0.15 * 0.9 * 2/3)f(t_0)$$

und damit der prognostizierte Wachstumsfaktor ($N(t_0)$ und $f(t_0)$ kürzen sich raus):

$$\frac{F(t)}{F(t_0)} = \frac{16 * 80 * 0.875 + 15 * 90 * 2/3}{11 * 50 + 9 * 80} = \frac{202}{127} = \underline{\underline{1.591}}.$$

Die prognostizierte Wachstumsrate (relativer Anstieg) der Fahrleistung der Senioren beträgt damit 59.1%

Der prognostizierte Wachstumsfaktor der Zahl der Senioren ist analog gegeben durch

$$\frac{n(t)}{n(t_0)} = \frac{0.16 * 0.875 + 0.15 * \frac{2}{3}}{0.11 + 0.09} = \underline{\underline{1.2}},$$

also ein prognostizierter Anstieg um 20%.

Aufgabe 2

(60 Punkte)

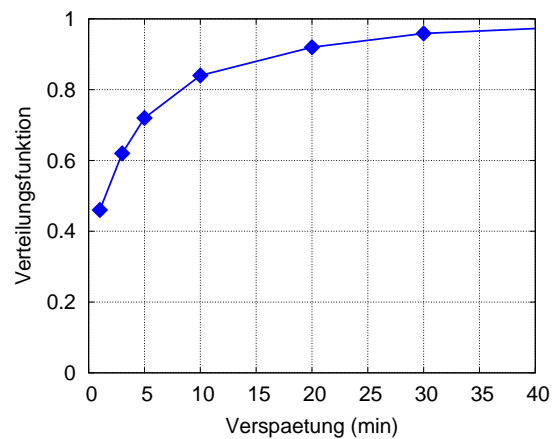
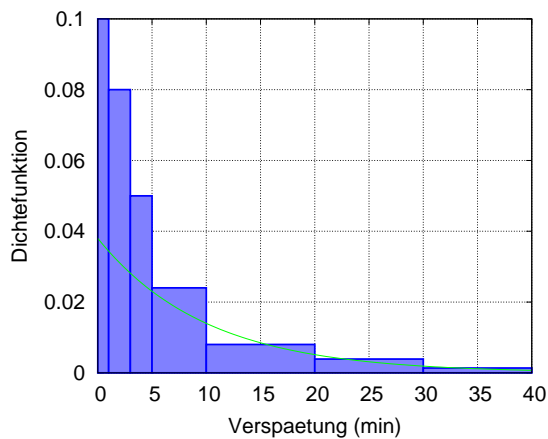
In allen Teilaufgaben sei h_k die Zahl der erfassten Züge in Klasse k und x die Verspätung in Minuten.

(a) Gesamtzahl $n = \sum_{k=1}^7 h_k = \underline{\underline{23261}}$.

(b),(c) Siehe Arbeitstabelle:

Verspätungsklasse (Minuten)	h_k	f_k	F_k	f_k^D
0-1	10699	0.460	0.460	0.460
1-3	3723	0.160	0.620	0.080
3-5	2326	0.100	0.720	0.050
5-10	2791	0.120	0.840	0.024
10-20	1861	0.080	0.920	0.008
20-30	905	0.039	0.959	0.004
30-60	956	0.041	1.000	0.001

(d) Plots der Verteilungs- und Dichtefunktion:



(e) Lagemaße:

$$\text{Arithmetisches Mittel } \bar{x} = \sum_{k=1}^7 f_k x_k^* = \underline{\underline{5.872}}$$

$$\text{Der Median ist in Klasse 2: } x_{0.5} = x_2^u + \frac{\Delta x_2(0.5 - F_1)}{f_2} = \underline{\underline{1.50}}$$

Der Modalwert ist in Klasse 1: Sowohl die Klassenmitte ($x_D = x_1^* = 0.5$) als auch die untere Grenze $x_D = 0$ (ist wohl realistischer) sind gültige Antworten.

Die Verteilung ist linkssteil bzw. rechtsschief, die Schiefe ist also größer Null.

(g) Streumaße:

$$\text{Varianz: } s_x^2 = \sum_{k=1}^7 f_k (x_k^* - \bar{x})^2 = \underline{\underline{100.2}}$$

$$\text{Standardabweichung: } s = \sqrt{s^2} = \underline{\underline{10.01}}$$

$$\text{Mittlere Absolute Abweichung: } s_{\text{MAD}} = \sum_{k=1}^7 f_k |x_k^* - \bar{x}| = \underline{\underline{5.292}}$$

(h) Wahrscheinlichkeiten und Quantile:

$$- P(\text{pünktlich}) = P(X \leq 3) = F_2 = \underline{\underline{62\%}}$$

- Anteil der Züge mit Verspätung mehr als 15 Minuten: Diese Verspätung ist in Klasse 5 zu finden:

$$P(X > 15) = 1 - P(X \leq 15) = 1 - F(15) = 1 - \left(F_4 + \frac{1}{2}f_5\right) = \underline{\underline{12\%}}$$

- 90% Quantil (90. Perzentil) ist ebenfalls in Klasse 5:

$$x_{0.9} = x_5^u + \frac{\Delta x_5(0.9 - F_4)}{f_5} = \underline{\underline{17.5}}$$

Aufgabe 3**(20 Punkte)**

- (a) Index der Verkehrstopfer in Westdeutschland zur Basis 1960:

$$I_t^W = \frac{x_t}{x_{1960}} = 1,182/140,135/140,8/14$$

(siehe Tabelle)

- (b) Die Anstiegsfaktoren der Verkehrstopfer werden in Westdeutschland und der ehemaligen DDR gleich angenommen. Damit ist der Index bis 1990 derselbe und damit auch derselbe in Gesamtdeutschland (Superskript D):

$$I_t^W = I_t^{\text{DDR}} = I_t^D, \quad t \leq 1990.$$

Für die Jahre nach 1990 wird eine Verknüpfung vorgenommen:

$$I_t^D = I_{1990}^W \frac{x_t^D}{x_{1990}^D}.$$

(Zahlenwerte siehe Tabelle)

- (c) Hier muss man die Verkehrstopfer durch die Verkehrsleistung
- V
- teilen, um die Risikokennziffer
- $y = x/V$
- zu erhalten. Da wieder Indizes mit der Basis 1960 berechnet werden sollen, kommt es auch hier nicht auf die Absolutwerte an und man erhält einfach

$$(I_t)_{\text{Verkehrsleistung}} = \frac{I_t^D V_{1960}}{V_t} = 1,182/140 * 207/430, \dots, 8/14 * 51/118 * 207/907.$$

(Zahlenwerte siehe Tabelle)

Zu (a)-(c)

Jahr	Index Verkehrstopfer Westdeutschland	Index Verkehrstopfer Gesamtdeutschland	Index Risikokennziffer Gesamtdeutschland
1960	1	1	1
1970	1.3	1.3	0.626
1980	0.964	0.964	0.347
1990	0.571	0.571	0.165
2000	-	0.373	0.091
2006	-	0.247	0.056

Das kilometerbezogene Risiko zu verunglücken ist also gegenüber 1960 um den Faktor $1/0.056 \approx 18$ zurückgegangen!

Aufgabe 4**(20 Punkte)**

(a) Übersicht in Tabelle:

Abgrenzungskriterium	Grundgesamtheit	Stichprobe
räumlich	Deutschland	10 Städte
zeitlich	“aktuell” bzw. 2007	23.9.-31.10.2007, abzüglich 5 Streiktage
sachlich	Fern- und Regionalverkehr	Fern- und Regionalverkehr

(b) Statistische Einheiten: (i) ein Zug des Fernverkehrs, (ii) ein Zug des Regionalverkehrs
 Statistische Gesamtheiten: Alle Züge, welche räumlich, zeitlich und sachlich gemäß Teil (a) abgegrenzt sind.

Statistisches Merkmal: Die Verspätung in Minuten.

(c) Repräsentativität bedeutet, dass die Verteilung der Merkmalsausprägungen in der Stichprobe, zumindest im Mittel, gleich der Verteilung in der Grundgesamtheit ist. Dies bedeutet insbesondere, dass die Erwartungswerte des Stichprobenmittels und der Stichprobenvarianz gleich den entsprechenden Werten in der Grundgesamtheit sind.

Berücksichtigung der Streiktage würde zu einer Verzerrung des Mittelwertes nach oben führen.

(d) Andere mögliche verzerrende Einflüsse sind z.B. die Auswahl der Bahnhöfe (nur große Städte) und die Jahreszeit (weder Urlaubsverkehr noch witterungsbedingte Beeinträchtigungen im Erhebungszeitraum).

Aufgabe 5**(40 Punkte)**(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 $\lambda = 1/10$: $P(X < x | X > 0) = F(x | X > 0) = 1 - e^{-\lambda x}$. Damit gilt für die unbedingte Verteilungsfunktion $P(X < x) = F(x) = p_0 + (1 - p_0) (1 - e^{-\lambda x}) = 1 - (1 - p_0)e^{-\lambda x}$.

Damit insgesamt

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

Die Wahrscheinlichkeit, den Anschluss zu verpassen, ist also gleich

$$p_v = P(X > 7) = 1 - P(x \leq 7) = (1 - p_0)e^{-0.7} = \underline{\underline{0.189}}.$$

(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 $y_{01} = 261$: Die Zufallsvariable $Y_1 = y_{01} + X$ hat die Verteilungsfunktion

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

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 $Y_2 = 739$ (die Differenz zwischen 8:15 am Folgetag und 18:55) mit der Verpass-Wahrscheinlichkeit $p_v = 0.189$. Insgesamt also

$$F_2(y) = \begin{cases} 0 & y < 279 \\ 1 - p_v & 279 \leq y < 739 \\ 1 & y \geq 739. \end{cases}$$

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

$$F_{x_1}(x) = F_{x_2}(x) = F(x) = \begin{cases} 0 & x < 0 \\ 1 - (1 - p_0)e^{-\lambda x} & x \geq 0 \end{cases}$$

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

$$p_a = P(Z \leq 7) = P(X_1 - X_2 \leq 7)$$

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 $f_3(x) = f_2(-x)$:

$$f_Z(z) = \int_{-\infty}^{\infty} f_1(x)f_3(z - x) dx = \int_{-\infty}^{\infty} f_1(x)f_2(x - z) dx$$

und damit

$$p_a = P(Z \leq 7) = \int_{-\infty}^7 f_Z(z) dz.$$

Ausrechnen mit dieser Formel würde Differenzieren sowie zweimal Integrieren erfordern, ist aber nicht verlangt. (Für Interessierte aber auf S. 9 gezeigt).

Aufgabe 6**(40 Punkte)**

(a) Tabelle:

Schätzer	erwartungstreu	konsistent	effizient
$\hat{\mu}_1$	Y	Y	Y
$\hat{\mu}_2$	N	Y	N
$\hat{\mu}_3$	Y	Y	N
$\hat{\mu}_4$	Y	N	N

Zum Berechnen der Erwartungswerte $E(\cdot)$ und Varianzen $V(\cdot)$ werden die Beziehungen

$$E(aX + bY) = aE(x) + bE(y), \quad V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab \operatorname{Cov}(X, Y)$$

verwendet, wobei die Kovarianz $\operatorname{Cov}(X, Y) = 0$ bei Zufallsstichproben. Also

$$\begin{aligned} E(\hat{\mu}_1) &= \mu, & V(\hat{\mu}_1) &= \frac{\sigma^2}{n} \\ E(\hat{\mu}_2) &= \frac{n}{n-1}\mu, & V(\hat{\mu}_2) &= \frac{\sigma^2}{n-1} \\ E(\hat{\mu}_3) &= \mu, & V(\hat{\mu}_3) &= \frac{2\sigma^2}{n} \\ E(\hat{\mu}_4) &= \mu, & V(\hat{\mu}_4) &= \frac{\sigma^2(1+4+9+16)}{100} = \frac{3\sigma^2}{10} \end{aligned}$$

(b) Tabelle:

Schätzer	erwartungstreu	konsistent	effizient
$\hat{\sigma}_1^2$	N	Y	N
$\hat{\sigma}_2^2$	Y	Y	Y
$\hat{\sigma}_3^2$	Y	Y	N

(Nicht verlangt, zur Info:)

$$E(\hat{\sigma}_1^2) = \frac{n}{n-1}\sigma^2, \quad E(\hat{\sigma}_2^2) = \sigma^2, \quad E(\hat{\sigma}_3^2) = \sigma^2.$$

Aufgabe 7

(20 Punkte)

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

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

Test-Statistik:

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

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

$$\underline{\underline{q = 5.03.}}$$

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

Da $q_{0.975}^{(1)} = \underline{\underline{5.024 \approx q}}$, ist die Grenz-Fehlerwahrscheinlichkeit, bei der man die Nullhypothese gerade noch ablehnen kann, durch $\alpha - 1 - 9 = 2.5\%$ gegeben.

Zu Aufgabe 5(c): Verpass-Wahrscheinlichkeit für den Fall, dass auch der Anschlusszug verspätung haben kann

Seien X und Y zwei unabhängige Zufallsvariable. Dann folgt direkt aus der Definition der Verteilungsfunktion ein "Faltungssatz" für die Verteilungsfunktion der Summe $Z = X + Y$:

$$\begin{aligned} F(z) &= P(Z \leq z) = P(X + Y \leq z) \\ &= \int dx f_x(x) P(x + Y \leq z) \\ &= \int dx f_x(x) P(Y \leq z - x) \\ &= \underline{\underline{\int dx f_x(x) F_y(z - x)}} \end{aligned}$$

Hier ist X die Verspätung des ersten Zuges und Y die *negative* Verspätung des zweiten Zuges. Die Verspätungen beider Züge haben die Verteilungsfunktion $F(x)$ und die Dichte $f(x)$ und damit $-Y$ die Dichte $f(-x)$ und die Verteilungsfunktion $1 - F(-x)$, also

$$F_z(z) = \int dx f(x)(1 - F(x - z)) = 1 - \int dx f(x)F(x - z).$$

Hier wurde $\int dx f(x) = 1$ ausgenutzt. Nun ist $F(x) = f(x) = 0$ für $x < 0$ und damit kann die untere Integrationsgrenze auf das Maximum von 0 und z gesetzt werden:

$$F_z(z) = 1 - \int_{\max(0,z)}^{\infty} dx f(x)F(x - z)$$

Für $z > 0$ wird $f(x)$ nur für $x > 0$ benötigt:

$$f(x) = F'(x) = (1 - p_0)\lambda e^{-\lambda x} \quad \text{falls } x > 0$$

Zusammen mit $F(x) = 1 - (1 - p_0)e^{-\lambda x}$ erhält man

$$\begin{aligned} F_z(z) &= 1 - \int_z^{\infty} dx \lambda (1 - p_0) e^{-\lambda x} (1 - (1 - p_0) e^{-\lambda x}) \\ &= 1 - \lambda (1 - p_0) \left[-\frac{1}{\lambda} e^{-\lambda x} \right]_z^{\infty} + \lambda (1 - p_0)^2 \left[-\frac{1}{2\lambda} e^{-2\lambda x} \right]_z^{\infty} \\ &= \underline{\underline{1 - \frac{1}{2}(1 - p_0^2)e^{-\lambda z}}}. \end{aligned}$$

Für 7 Minuten geht die Verpasswahrscheinlichkeit $1 - F_z(7)$ von 18.9 Minuten auf 15.3 Minuten herunter.

Sonderfälle:

- Für $p_0 = 0$ ist $F_z(z) = \frac{1}{2}e^{-\lambda z}$. Die Wahrscheinlichkeit, dass der zweite Zug um mehr als z Minuten pünktlicher als der erste ist, ist $1/2$ (der Zug ist überhaupt pünktlicher) multipliziert mit $e^{-\lambda z}$ (der erste Zug ist mehr als z Minuten verspäteter als der zweite).

- Für $z \rightarrow 0^+$ gilt $F_z(0^+) = 1 - \frac{1}{2}(1 - p_0^2)$ bzw. $1 - F_z(0^+) = \frac{1}{2}(1 - p_0^2)$. Anschaulich: Die Wahrscheinlichkeit $1 - F_z(0^+)$ dafür, dass der erste Zug unpünktlicher als der zweite ist, ist die Summe aus der Wahrscheinlichkeit $(1 - p_0)p_0$ dafür, dass der erste Zug unpünktlich und der zweite Zug pünktlich ist, zuzüglich der Wahrscheinlichkeit $(1 - p_0)^2/2$ dafür, dass beide Züge unpünktlich sind, und der erste außerdem unpünktlicher als der zweite ist. Zusammen ergibt sich $\frac{1}{2}(1 - p_0^2)$.

