

Klausur zur Vorlesung Statistik I für Bachelors und Statistik I und II für Diplomstudenten, SS 2008

Lösungsvorschlag

Aufgabe 1

(20 Punkte)

Die deutsche Unfallstatistik listet für das Jahr 2007 alle Unfälle, welche der Polizei gemeldet wurden. Diese sollen bezüglich der an einem Unfall beteiligten Fahrzeuge und Personen, der jeweiligen Zahl der Todesopfer, Schwer- und Leichtverletzten sowie des entstandenen Sachschadens untersucht werden. Dabei wird auch nach Unfallursache (z.B. zu hohe Geschwindigkeit oder Vorfahrtsverletzung), nach Alter und Geschlecht des Haupt-Unfallverursachers, des Verkehrsmittels des Hauptverursachers und der Verletzten und Getöteten (Kfz, Rad, zu Fuß oder ÖPNV) sowie nach Stadtstraßen-, Landstraßen- und Autobahnunfällen unterschieden.

- (a) Geben Sie die statistische Gesamtheit und den Merkmalsträger an. Grenzen Sie die statistische Gesamtheit räumlich, zeitlich und sachlich ab.

Lösung

- Stat. Einheit bzw. Merkmalsträger: Unfälle
- Stat. Gesamtheit: Alle Unfälle mit folgenden Abgrenzungen:
 - * räumlich: Deutschland
 - * Zeitlich: Jahr 2007
 - * Sachlich: Alle der Polizei gemeldeten Verkehrsunfälle, also auch solche mit Fußgängern und Radfahrern, aber keine, die der Polizei nicht gemeldet wurden.

- (b) Geben Sie die in der Aufgabenstellung erwähnten Merkmale und ihre Skalierung an. Unterscheiden Sie dabei auch die Verhältnisskalierung von der Absolutskalierung.

Lösung (A=Absolutskalierung, V=Verhältnisskala, K=Kardinalskalierung, N=Nominalskalierung, Ord=Ordinalskalierung):

Merkmal	Skalierung
Zahl der Personen	K bzw. A
Art der Fahrzeuge	Nom
Zahl der Fahrzeuge	K bzw. A
Zahl der Todesopfer, Schwer- und Leichtverletzten	jeweils A
Sachschaden	K bzw. V
Unfallursache	Nom
Alter Hauptverursacher	K bzw. V
Geschlecht Hauptverursacher	Nom (dichotom)
Verkehrsmittel des Hauptverursacher	Nom
Verkehrsmittel der anderen Beteiligten	jeweils Nom
Straßenkategorie	Ord

Hinweis: Bei den beteiligten Personen hat man eine Hierarchie von Merkmalen: Einerseits ist deren Anzahl ein Merkmal der stat. Einheit "Unfall", gleichzeitig spielen sie aber auch die Rolle eines Merkmalsträgers mit Merkmalen wie Alter, Geschlecht, Rolle beim Unfall usw.

- (c) *Es wird ein Gesetzesvorschlag eingebracht, nach dem ältere Personen einen regelmäßigen Fahrtauglichkeitstest machen müssen, um ihren Führerschein zu behalten. Dies soll mit Hilfe der Unfallstatistik begründet werden. Welche Kombination der obigen Merkmale würde man dabei zugrunde legen? Welche wichtigen zusätzlichen Angaben fehlen für eine seriöse Begründung?*

Lösung:

- Relevante angegebene Merkmale: Alter, Geschlecht und Verkehrsmittel des Haupt-Unfallverursachers;
- Zusätzlich noch wichtig: Die Fahrleistungen (z.B. Kilometer pro Tag) als Fahrer von Kfz und Fahrrad bzw. als Fußgänger. Nur dann kann man abschätzen, inwieweit "die Alten" (wie übrigens auch die "Jungen") ein Risiko darstellen.

Aufgabe 2

(60 Punkte)

Einer Verkehrsstatistik entnimmt man folgende Aufgliederung der transportierten Güter (in Millionen Tonnen) nach verschiedenen Verkehrsmitteln und Entfernungsstufen:

Entfernung (km)	Gütermenge LKW (in Mio t)	Gütermenge Bahn (in Mio t)	Gütermenge Schiff (in Mio t)
0 – 50	1 020	9	0
50 – 100	660	39	10
100 – 200	340	84	95
200 – 500	150	55	40
500 – 1 000	60	21	7

- (a) *Wieviel Tonnen Güter wurden mit den drei Verkehrsarten jeweils transportiert?*

Lösung: Es sei y_{ik} die in der Entfernungsklasse k mit Verkehrsmittel i (1: LKW; 2: Bahn; 3: Schiff) transportierte Menge (in Mio t) an Gütern. Dann gilt

$$\text{LKW : } M_1 = \sum_{k=1}^5 y_{1k} = \underline{\underline{2230}},$$

$$\text{Bahn : } M_2 = \sum_{k=1}^5 y_{2k} = \underline{\underline{208}},$$

$$\text{Schiff : } M_3 = \sum_{k=1}^5 y_{3k} = \underline{\underline{152}}.$$

- (b) *Wie ist die Aufteilung der Transportmenge auf die drei Verkehrsarten?*

Lösung: Es sei $M = M_1 + M_2 + M_3 = \underline{2590}$ die insgesamt transportierte Menge in Mio t. Dann gilt

$$\text{Anteil LKW : } A_1 = \frac{M_1}{M} = \underline{0.861},$$

$$\text{Anteil Bahn : } A_2 = \frac{M_2}{M} = \underline{0.080},$$

$$\text{Anteil Schiff : } A_3 = \frac{M_3}{M} = \underline{0.059}.$$

- (c) Berechnen Sie nun für die drei Verkehrsarten jeweils die Transportleistung in Milliarden tkm (Tonnen-Kilometer). Nehmen Sie dabei an, dass innerhalb einer Entfernungsklasse die Transportweiten gleichverteilt sind. Berechnen Sie auch die Anteile der drei Verkehrsarten an der Leistung. Erstellen Sie Tortendiagramme für die transportierte Menge und die Transportleistung.

Lösung: Die Transportleistung (in Mrd tkm) in jeder Klasse ist die transportierte Menge y_{ik} multipliziert mit der Klassenmitte x_k^* der jeweiligen Transportweiten:

$$L_{ik} = x_k^* y_{ik}.$$

Also

$$\text{Leistung LKW : } L_1 = \sum_{k=1}^5 x_k^* y_{1k} = \underline{223.5},$$

$$\text{Leistung Bahn : } L_2 = \sum_{k=1}^5 x_k^* y_{2k} = \underline{50.8},$$

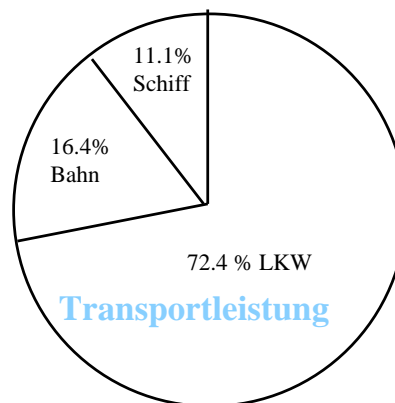
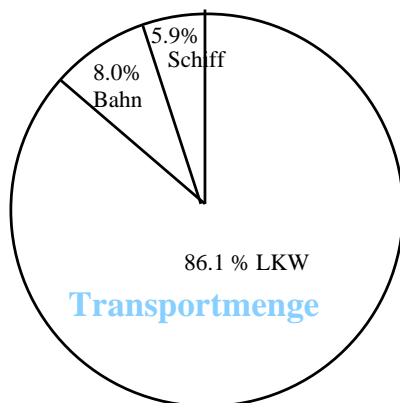
$$\text{Leistung Schiff : } L_3 = \sum_{k=1}^5 x_k^* y_{3k} = \underline{34.3}.$$

Es sei $L = L_1 + L_2 + L_3 = \underline{308.5}$ die gesamte Transportleistung in Mrd tkm. Dann gilt

$$\text{Leistungsanteil LKW : } A_1^{(L)} = \frac{L_1}{L} = \underline{0.724},$$

$$\text{Leistungsanteil Bahn : } A_2^{(L)} = \frac{L_2}{L} = \underline{0.165},$$

$$\text{Leistungsanteil Schiff : } A_3^{(L)} = \frac{L_3}{L} = \underline{0.111}.$$



- (d) Berechnen Sie für die drei Verkehrsmittel jeweils die mittlere Transportweite.

Lösung:

$$\text{Mittl. Transportweite LKW : } \bar{x}_1 = \sum_{k=1}^5 x_k^* y_{1k} = \underline{\underline{100.2 \text{ km}}},$$

$$\text{Mittl. Transportweite Bahn : } \bar{x}_2 = \sum_{k=1}^5 x_k^* y_{2k} = \underline{\underline{244.0 \text{ km}}},$$

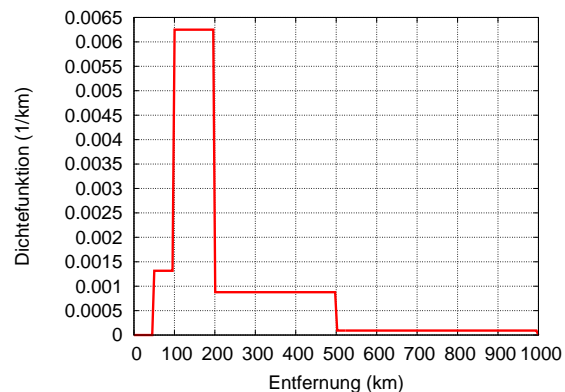
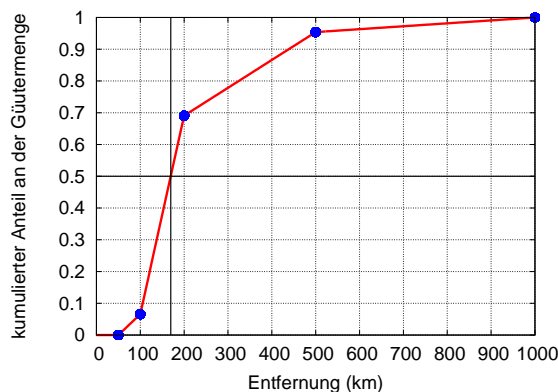
$$\text{Mittl. Transportweite Schiff : } \bar{x}_3 = \sum_{k=1}^5 x_k^* y_{3k} = \underline{\underline{225.3 \text{ km}}}.$$

- (e) *Konzentrieren Sie sich nun auf den **Schiffstransport**: Berechnen Sie die Verteilungs- und Dichtefunktionen der transportierten Mengen bezüglich der Transportweite und zeichnen Sie das Ergebnis in die beiden Diagramme ein.*

Lösung:

Wertetabelle (enthält auch Lösung zu (f)):

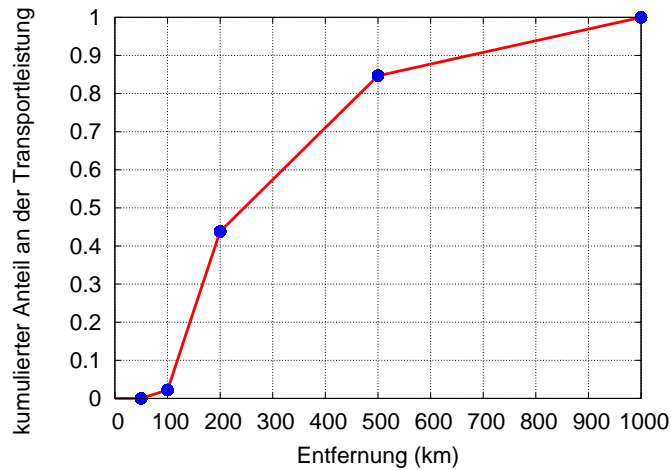
Klasse	x_k^*	f_k	F_k	$f_k^{(L)}$	$F_k^{(L)}$
0-50 km	25 km	0	0	0	0
50-100 km	75 km	0.0658	0.0658	0.0218	0.0218
100-200 km	150 km	0.625	0.691	0.416	0.438
200-500 km	350 km	0.263	0.954	0.409	0.847
500-1000 km	750 km	0.046	1	0.153	1



- (f) *Berechnen Sie nun die Verteilungsfunktion der Transportleistung im Güterschiffsverkehr bezüglich der Entfernung und tragen Sie das Ergebnis in das folgende Diagramm ein*

Lösung:

Relative Leistungsanteile $f_k^{(L)}$ und kumulierte Leistungsanteile $F_k^{(L)}$ zum Plotten der Verteilungsfunktion siehe Teil (e). Plot selbst:



- (g) *Schiffe werden stichprobenartig auf geschmuggelte bzw. verbotene Ladung untersucht. Wie hoch ist die Wahrscheinlichkeit, dass die überprüften Güter zwischen 500 km und 1000 km weit transportiert werden, wenn die Stichproben (i) am Start- oder Zielpunkt stattfinden, (ii) die Kontrollen irgendwo unterwegs durchgeführt werden?*

Lösung:

- (i) Stichproben am Start- oder Zielpunkt: Dann ist (unter Annahme, dass die Größe der Schiffe nicht mit der Transportweite korreliert ist) die Zahl der aus- bzw. einlaufenden Schiffe proportional zur transportierten Menge. Damit ist der Anteil durch $f_5 = \underline{4.6\%}$
- (ii) Kontrollen irgendwo unterwegs: Dieser Anteil wächst offensichtlich mit den insgesamt zurückgelegten Kilometern in der jeweiligen ENtfernungsklasse, ist also proportional zum Leistungsanteil $f_5^{(L)} = \underline{15.3\%}$.

Aufgabe 3

(40 Punkte)

Gegeben ist folgende Statistik der Übernachtungspreise, Auslastungen und Sternenzahlen einiger Hotels in Dresden:

Übernachtungspreis (in €)	72	67	78	40	34	50	98	116	82	31	58	15
Auslastung (%)	77	42	90	58	56	40	90	75	85	51	76	49
Sternenzahl	3	3	3	1	2	2	4	4	4	1	2	1

Ihre Statistik-Software erzeugt aus den Daten folgende Streudiagramme [hier nicht gezeigt] und folgende Lagemaße, Varianzen und Kovarianzen bezüglich des Preises x (in €) und der Auslastung y (in %) aus:

$$\bar{x} = 61.75, \quad \bar{y} = 65.75, \quad s_x^2 = 799.2, \quad s_y^2 = 312.0, \quad s_{xy} = 341.3.$$

- (a) Führen Sie eine lineare Regression der Auslastung als Funktion des Preises durch und zeichnen Sie den Graph der Regressionsfunktion in das linke Streudiagramm ein.

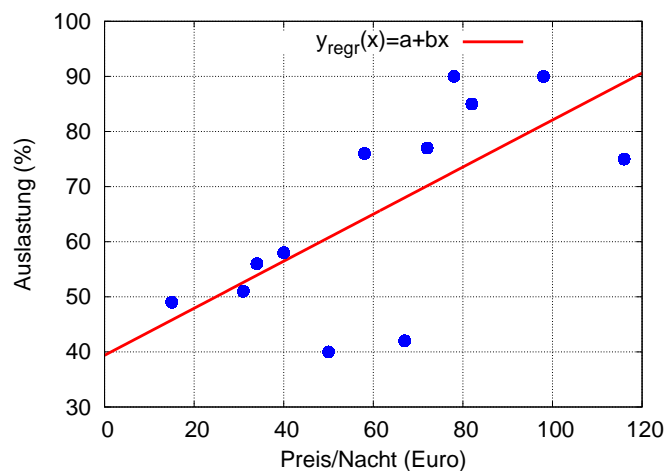
Lösung: Univariate lineare Regression der Abhängigkeit von y von x :

$$\hat{y}(x) = a + bx$$

wobei die für die zur Berechnung der Koeffizienten nötigen Mittelwerte, Varianzen und Kovarianzen bereits in der Aufgabenstellung gegeben sind:

$$b = \frac{s_{xy}}{s_x^2} = \underline{\underline{0.428\%/\text{€}}}, \quad a = \bar{y} - b\bar{x} = \underline{\underline{39.4}}.$$

Graph:



- (b) Was sind für diesen Sachverhalt die offensichtlichen Schwächen der linearen Regression? Zeigen Sie anhand eines frei wählbaren Zahlenbeispiels, dass diese Regressionsfunktion zu unsinnigen Aussagen führen kann.

Lösung: z.B. gibt es selbst bei kostenlosen Hotels ($x = 0$) nur eine endliche Auslastung von $a = 39.4\%$. Dies kann ja (bei sehr geringer Nachfrage und schlechter Hotelqualität) theoretisch noch sein. Hingegen steigt bei Preisen oberhalb von $x_{100} = (100 - a)/b = 142 \text{ €}$ die Auslastung auf Werte oberhalb von 100%, was nicht sein kann.

- (c) Sie erhalten in Aufgabenteil (a) das Ergebnis, dass die Auslastung mit dem Übernachtungspreis tendenziell steigt. Ist dies unsinnig oder plausibel? Begründen Sie ihre Antwort mit dem rechten Streudiagramm und der vom Statistikprogramm gelieferten Korrelation $r_{xz} = 0.929$ zwischen Übernachtungspreis und Sternezahl z .

Lösung: Es ist plausibel, da die Korrelation $r_{xz} = 0.929$ besagt, dass mit dem Preis tendenziell auch die Sternezahl und damit die Qualität steigt. Wenn nun die Qualität mit dem Preis überproportional steigt, dann finden viele Touristen die teureren Hotels attraktiver.

- (d) Geben Sie für die Abhängigkeit der Auslastung vom Hotelpreis die erklärte Varianz, die nicht erklärte Varianz (Residualvarianz), die Gesamtvarianz und das Bestimmtheitsmaß an.

Lösung:

Da es sich um eine lineare Regression handelt, gilt nach Formelsammlung oder Unterlagen:

$$B = B_{xy} = r_{xy}^2 = \underline{0.468}.$$

Ebenfalls für lineare Regression gilt das Additionsgesetz

$$\sigma_y^2 = \sigma_E^2 + \sigma_R^2$$

mit der erklärten Varianz σ_E^2 und der Residualvarianz σ_R^2 . Außerdem gilt ja nach Definition des Bestimmtheitsmaßes

$$U = 1 - B = \frac{\sigma_R^2}{\sigma_y^2}$$

Mit der in der Aufgabenstellung angegebenen Gesamtvarianz

$$\sigma^2 = \sigma_y^2 = \underline{312}$$

gilt somit

$$\sigma_R^2 = (1 - B)\sigma_y^2 = \underline{166.3}, \quad \sigma_E^2 = \sigma_y^2 - \sigma_R^2 = \underline{166.3}.$$

Aufgabe 4**(30 Punkte)**

Bei der Korrektur eines Dokuments auf Rechtschreibfehler entdeckte der erste Überprüfer 5 Fehler und der zweite 4 Fehler. Darunter waren zwei Fehler, welche von beiden entdeckt wurden. Man kann davon ausgehen, dass die beiden Personen unabhängig voneinander korrigiert haben und die Wahrscheinlichkeit für das Aufspüren eines Fehlers bei allen Fehlern dieselbe ist.

- (a) Nutzen Sie die Regeln für bedingte Wahrscheinlichkeiten und Wahrscheinlichkeiten unabhängiger Ereignisse, um die Gesamtzahl der Fehler abzuschätzen. Wieviele unentdeckte Fehler sind danach zu erwarten?

Lösung: Mit der Fehlerzahl n , der Entdeckungswahrscheinlichkeiten p_A und p_B gilt anhand der Aufgabenstellung und bei unabhängig die Fehler entdeckenden Korrektoren A und B:

$$\begin{aligned} 5 &= np_A, \\ 4 &= np_B, \\ 2 &= np_{APB}. \end{aligned}$$

Aus der zweiten und dritten Gl. ergibt sich $p_A = \underline{0.5}$, aus der ersten und dritten $p_B = \underline{0.4}$ und damit

$$n = \frac{2}{p_{APB}} = \underline{10}.$$

- (b) Könnten es auch mehr oder weniger sein? Nehmen Sie dazu als Nullhypothese H_0 an, dass sich tatsächlich 10 Fehler eingeschlichen haben und die Wahrscheinlichkeit, einen Fehler aufzuspüren, bei der ersten Person bei $1/2$ und bei der zweiten bei $2/5$ liegt. Berechnen Sie unter dieser Nullhypothese die Verteilung der von der ersten

und zweiten Person entdeckten Fehlerzahl sowie die Verteilung der Gesamtzahl der entdeckten Fehler.

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

$$n_A \sim B(10, 0.5), \quad n_B \sim B(10, 0.4)$$

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

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

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

Aufgabe 5

(25 Punkte)

Für neue Fahrerassistenzsysteme (z.B. ein Kreuzungsassistent oder eine verbesserte Navigation) ist es wichtig, zu bestimmen, wie stark sich die Fahrtrichtung innerhalb eines bestimmten Zeitraums ändert. Diese Änderung kann prinzipiell auf dreierlei Arten gemessen werden:

1. Über den aufgezeichneten Lenkwinkel und die Geschwindigkeit,
2. über einen "Gierdatensensor", welcher direkt die Drehung des Fahrzeuges misst,
3. und über die Auswertung der vom Navigationssystem gelieferten Positionen.

Für eine tatsächliche Änderung der Fahrtrichtung von 30 Grad liefern alle Methoden im Mittel das richtige Ergebnis, aber mit Unsicherheiten (einfachen Standardabweichungen) von $\sigma_1 = 2$ Grad, $\sigma_2 = 2$ Grad und $\sigma_3 = 4$ Grad. Kann man durch Kombinationen der einzelnen Messungen den Fehler verringern?

- (a) Berechnen Sie zunächst die Standardabweichung der Richtungsbestimmung, falls dazu das arithmetische Mittel der drei Messungen verwendet wird.

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

$$\sigma_1^2 = \sigma_2^2 = 4, \quad \sigma_3^2 = 16.$$

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

$$\bar{X} = \frac{1}{3} \sum_{i=1}^3 X_i, \quad V(\bar{X}) = \frac{1}{9} \sum_{i=1}^3 X_i^2 = \frac{24}{9} = \frac{8}{3} \Rightarrow \sigma_{\bar{x}} = \sqrt{V(\bar{X})} = \underline{\underline{1.63}}.$$

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

- (b) Kann man die Genauigkeit optimieren, indem man die "schlechtere" Messmethode 3 ignoriert oder weniger gewichtet? Bestimmen Sie dazu die Varianz eines gewichteten arithmetischen Mittels der drei Messungen, wobei Sie die ersten beiden Messungen gleich gewichten. Minimieren Sie die Varianz durch Variation der Gewichtungen. Mit welchem Anteil trägt die Messmethode 3 im optimalen Fall noch bei? Wie hoch ist die resultierende Varianz?

Lösung: 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

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

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 berechnet sich wieder mit den Rechenregeln für unabhängige Zufallsgrößen aus der Formelsammlung:

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

Nun Minimieren bezüglich w durch Ableiten und Nullsetzen:

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

Damit

$$(w_1)_{\text{opt}} = (w_2)_{\text{opt}} = \underline{\underline{\frac{4}{9}}}, \quad (w_3)_{\text{opt}} = \underline{\underline{\frac{1}{9}}}.$$

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 $\bar{X}(w_{\text{opt}})$ ist gegeben durch

$$V(w_{\text{opt}}) = \frac{16}{81} * 8 + \frac{1}{81} * 16 = \frac{16}{9}$$

und damit

$$\sigma(w_{\text{opt}}) = \underline{\underline{\frac{4}{3}}} = 1.333.$$

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

Aufgabe 6

(15 Punkte)

Eine bestimmte Bank benötigt für ein Jahr Geld und kann sich dieses nur (von anderen Banken oder Investoren) ausleihen, wenn sie 6% Zinsen bietet. Ausfallsichere Bundesanleihen bieten nur 4% Zinsen. Wie hoch ist die von den Marktteilnehmern angenommene subjektive Wahrscheinlichkeit dafür, dass diese Bank innerhalb des Jahres Pleite geht (und damit das ausgeliehene Geld samt Zinsen verloren ist) ?

Lösung 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

$$E(X) = (1 - p)(1 + r) + p * 0$$

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

Durch Gleichsetzen mit dem Auszahlungsbetrag $1 + r_0$ der Bundesanleihe erhält man

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

Aufgabe 7

(55 Punkte)

Bei der Neueinschreibung an eine Universität bewerben sich die Studenten im allgemeinen an mehreren Universitäten gleichzeitig. Die Universitäten nehmen die geeigneten Kandidaten an. Allerdings erhalten viele Studenten von mehreren Universitäten Einladungen und wählen nun ihrerseits aus. Um die Studienplätze eines Studiengangs möglichst optimal zu besetzen, werden daher mehr Studenten akzeptiert als Studienplätze zur Verfügung stehen. In einem bestimmten Studiengang mit 200 Plätzen ergaben sich in den letzten Jahren bei den ersten 200 Einladungen folgende Absagen von Seiten der Studenten:

Jahr	2005	2006	2007	2008
Zahl der Absagen	83	103	63	111

- (a) Geben Sie Schätzer für den Erwartungswert und die Varianz der Zahl der Absagen bzw. der Absagerquote an.

Lösung: Schätzung von Erwartungswert μ und Varianz σ^2 aus der Stichprobe:

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

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

- (b) Mit welcher Wahrscheinlichkeit ergibt sich höchstens 80 Absagen, wenn Sie eine Gaußverteilung der Absagenquoten mit einem Erwartungswert von 45% und einer Standardabweichung von 10% annehmen?

Bei 200 Einladungen ergeben sich mit den Angaben der Aufgabenstellung für den wahren Mittelwert und die wahre Standardabweichung der Zahl der Absagen

$$\mu = 200 * 0.45 = \underline{\underline{90}}, \quad \sigma = 200 * 0.10 = \underline{\underline{20}}.$$

Damit ist die Größe

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

standardnormalverteilt. Damit

$$\begin{aligned} P(X < 80) &= P\left(Z < \frac{80 - 90}{20}\right) \\ &= P\left(Z < -\frac{1}{2}\right) \\ &= \Phi\left(-\frac{1}{2}\right) = 1 - \Phi\left(\frac{1}{2}\right) = \underline{\underline{0.309}}. \end{aligned}$$

- (c) Berechnen Sie das Konfidenzintervall der Absagenquote zur Fehlerwahrscheinlichkeit von 5%.

Lösung: 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:

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

Mit

$$t_{0.975}^{(3)} \hat{\sigma} / \sqrt{n} = 3.182 * \sqrt{462.7} / 2 = \underline{\underline{34.2}}$$

ergibt sich das Konfidenzintervall

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

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

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

- (d) Es gibt zwei mögliche Ursachen für die Schwankungen: (i) Schwankungen durch die endliche Zahl der Studenten, (ii) Schwankungen der wahrgenommenen Attraktivität des Studiengangs von seiten der Studenten von einem Jahr zum nächsten. Welche Schwankungsursache ist hier größer? Nehmen Sie dazu Unabhängigkeit der beiden Ursachen an und vergleichen Sie die in (a) geschätzte Gesamtvarianz mit der erwarteten Varianz bei konstanter Attraktivität. In letzterem Fall gilt für jeden Studenten die gleiche Absagewahrscheinlichkeit, welche Sie wieder mit 45% annehmen können.

Lösung

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

$$\sigma_{(i)}^2 = 200\theta(1 - \theta) = \underline{\underline{49.5}}$$

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) Um den Studiengang trotz der Absagen zu füllen, werden an die Studenten 350 Einladungen für die 200 Plätze verschickt. Wie groß ist das Risiko, dass sich mehr als 200 Studenten einschreiben und damit der Studiengang überfüllt ist? Nehmen Sie nun an, dass die Schwankungen von Jahr zu Jahr überwiegen, so dass der erwartete Anteilswert der Absagen, nicht aber der Erwartungswert der Anzahl der Absagen unabhängig von der Zahl der angenommenen Studenten ist.

Lösung Bei 350 Einladungen und 200 Plätzen droht Überfüllung, falls die Ablehnquote unterhalb von

$$\theta_c = \frac{150}{350} = \frac{3}{7} = \underline{\underline{42.9\%}}$$

liegt.

Nach Aufgabenstellung bleibt der Erwartungswert und die Varianz der Ablehnquote 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 bekannte Erwartungswert von $\theta_0 = \frac{\mu_0}{200} = 0.45$ und eine bekannte Varianz $\sigma_\theta^2 = 0.01$. Damit ist

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

standardnormalverteilt und es gilt

$$\begin{aligned} P(\theta < \theta_c) &= P\left(Z > \frac{\theta_c - \theta_0}{\sigma_\theta}\right) \\ &= P(Z < 10(0.42 - 0.45)) = P(Z < -0.3) = 1 - \Phi(0.3) = \underline{\underline{0.382}}. \end{aligned}$$

- (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 $P(Z > (\theta_c - \theta_0)/0.1)$ hat man die Wahrscheinlichkeit für Überbelegung gegeben durch

$$\begin{aligned} P\left(T_3 < \frac{\theta_c - \hat{\mu}/200}{\hat{\sigma}/200}\right) &= P(T_3 < (0.42 - 0.45)/0.1075) \\ &= P(T_3 < -0.278) = 1 - P(T_3 < 0.278). \end{aligned}$$

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

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