Schriftliche Prüfung im Grundwissen

Angewandte Stochastik

Klausur mit Lösungen

gemäß Prüfungsordnung 5
der Deutschen Aktuarvereinigung e. V.

am 19. Mai 2023

Hinweise:

- Als Hilfsmittel sind Seminarunterlagen und Aufgaben in Papierform, handschriftliche Notizen im Rahmen der normalen Schulung sowie ein nicht programmierbarer Taschenrechner zugelassen.
- Die Gesamtpunktzahl beträgt 180 Punkte. Die Klausur ist bestanden, wenn mindestens 90 Punkte erreicht werden.
- Bitte prüfen Sie die Ihnen vorliegende Prüfungsklausur auf Vollständigkeit. Die Klausur mit Lösungen besteht aus 26 Seiten.
- Mit Ausnahme der MC-Fragen sind alle Antworten zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.
- Bei MC-Fragen entscheiden Sie für jede Aussage, ob diese wahr oder falsch ist. Eine falsche Antwort gibt 0 Punkte; Minuspunkte werden nicht vergeben. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben. Eine Begründung ist nicht erforderlich.
- Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet.

Mitglieder der Prüfungskommission:

Prof. Torsten Becker, Dr. Richard Herrmann,
Prof. Christian Heumann, Dr. Stefan Pilz,
Prof. Viktor Sandor, Dr. Dominik Schäfer

Aufgabe 1. *[Lebensdauermodelle] [30 Punkte]*

Betrachtet wird ein Bestand von Risikoversicherungen für den Todesfall mit vierjähriger Laufzeit. Untersucht wird der Teilbestand von 1000 Versicherten, die bei Beginn der Versicherung 40 Jahre alt sind.

(a) [22 Punkte] Nehmen Sie für diese Teilaufgabe an, dass Storno oder ein anderer Abgang außer Tod während der Vertragslaufzeit vor Ablauf nicht vorliegt. Es wurden folgende Anzahlen an Todesfällen beobachtet:

Alter	40	41	42	43
Todesfälle	5	6	7	8

(i) [4 Punkte] Bestimmen Sie die Anzahl der im jeweiligen Alter unter Risiko stehenden Personen.

(ii) [6 Punkte] Formulieren und erläutern Sie allgemein die Survivalfunktion für den Fall, dass die Zufallsvariable T nur die diskreten Werte $t_0 = 39$, $t_1 = 40$, $t_2 = 41$, $t_3 = 42$, $t_4 = 43$ annehmen kann.

(iii) [6 Punkte] Berechnen Sie die Survivalfunktion mit Hilfe des Kaplan-Meier-Schätzers.

(iv) [6 Punkte] Liegt eine Zensierung der Beobachtungen vor? Wenn ja, in welchen Altern? Begründen Sie Ihre Antwort.

(b) [8 Punkte] Nehmen Sie für diese Teilaufgabe an, dass Storno als weiterer Abgang während der Vertragslaufzeit vorliegt. Es wurden folgende Anzahlen an Todes- und Stornofällen beobachtet:

Alter	40	41	42	43
Todesfälle	5	6	7	8
Storno	0	50	50	0

(i) [2 Punkte] Bestimmen Sie die Anzahl der im jeweiligen Alter unter Risiko stehenden Personen.

(ii) [4 Punkte] Liegt eine Zensierung der Beobachtungen vor? Wenn ja, in welchen Altern? Begründen Sie Ihre Antwort.

(iii) [2 Punkte] Berechnen Sie die Survivalfunktion mit Hilfe des Kaplan-Meier-Schätzers.

Lösung

(a) (i) [4 Punkte]

Anzahl der unter Risiko stehenden Personen

Alter j	40	41	42	43	44
Todesfälle d_j	5	6	7	8	0
Anzahl unter Risiko n_j	1000	995	989	982	0

(ii) [6 Punkte]

Sei $t > 0$ und $0 =: t_0 < t_1 < \cdots < t_{n-1} < t_n := t$.

Die Survivalfunktion lautet allgemein

$$S(t) = \prod_{i=1}^{n} P(T > t_i | T > t_{i-1}). \tag{1}$$

Im vorliegenden diskreten Fall sind die Realisationen t_i der Zufallsvariable T die jeweiligen Alter in Jahren. Der Ausdruck $P(T > t_i | T > t_{i-1}) =: p_i$ bezeichnet die Wahrscheinlichkeit, dass die Person das Alter t_i überlebt unter der Voraussetzung, dass die Person das Alter t_{i-1} überlebt hat. Für die t_i gilt $t_0 = 39$, $t_1 = 40$, $t_2 = 41$, $t_3 = 42$, $t_4 = 43$.

Für die Aufgabenstellung lautet damit die diskrete Survivalfunktion für $k = 1, \ldots, 4$

$$S(t_k) = \prod_{i=1}^{k} P(T > t_i | T > t_{i-1}) = \prod_{j=40}^{40+k-1} P(T > j | T > j-1) = \prod_{j=40}^{40+k-1} p_j. \tag{2}$$

Für $k = 1$ gilt

$$P(T > t_1 | T > t_0) = P(T > t_1) \tag{3}$$

bzw.

$$P(T > 40 | T > 39) = P(T > 40). \tag{4}$$

(iii) [6 Punkte]

Bezeichne d_j die Anzahl der Abgänge wegen Tod im Alter j und n_j die Anzahl der Versicherten unter Risiko im Alter j. Dann wird die Überlebenswahrscheinlichkeit

$$P(T > j | T > j-1) = p_j \tag{5}$$

geschätzt durch

$$\hat{p}_j := 1 - \frac{d_j}{n_j}. \tag{6}$$

Der Kaplan-Meier-Schätzer für die Survivalfunktion für $S(t)$ lautet

$$\begin{aligned} \hat{S}(t) &= \begin{cases} 1 & \text{falls } t < t_{(1)} \\ \prod_{j|t_{(j)} \le t} \hat{p}_j & \text{sonst} \end{cases} \\ &= \begin{cases} 1 & \text{falls } t < t_{(1)} \\ \prod_{j|t_{(j)} \le t} (1 - \frac{d_j}{n_j}) & \text{sonst} \end{cases} \end{aligned}$$

Die Schätzer $\hat{p}_j$ lauten

$$\hat{p}_{40} = 1 - \frac{d_{40}}{n_{40}} = 1 - \frac{5}{1000} = 0,9950$$
$$\hat{p}_{41} = 1 - \frac{d_{41}}{n_{41}} = 1 - \frac{6}{995} = 0,9940$$
$$\hat{p}_{42} = 1 - \frac{d_{42}}{n_{42}} = 1 - \frac{7}{989} = 0,9929$$
$$\hat{p}_{43} = 1 - \frac{d_{43}}{n_{43}} = 1 - \frac{8}{982} = 0,9919$$

Die geschätzten Werte $\hat{S}$ der Survivalfunktion lauten dann

$$\hat{S}(40) = 0,9950$$
$$\hat{S}(41) = 0,9950 \cdot 0,9940 = 0,9890$$
$$\hat{S}(42) = 0,9950 \cdot 0,9940 \cdot 0,9929 = 0,9820$$
$$\hat{S}(43) = 0,9950 \cdot 0,9940 \cdot 0,9929 \cdot 0,9919 = 0,9740$$

(iv) [6 Punkte]

Eine Zensierung der Beobachtungen liegt vor, wenn die Beobachtung beendet wird, bevor das Ereignis (hier Tod) eintritt.

Es werden die Lebensdauern $T_1, \ldots, T_n$ von n Individuen beobachtet und $C_1, \ldots, C_n$ seien die Zensierungszeiten der Individuen. Die Zufallsvariablen T_i, C_i, $i = 1, \ldots, n$ seien unabhängig. Beobachtet werden $T_i^* := \min(T_i, C_i)$.

Da bei dieser Teilaufgabe angenommen wird, dass ein Abgang aus dem Bestand nur bei dem zu beobachtenden Ereignis Tod stattfindet, liegt in den Altern 40 bis 43 keine Zensierung der Beobachtungen vor.

Da im Alter 44 die Vertragslaufzeit beendet ist, scheiden alle Versicherten mit Erreichen des Alters 44 aus der Beobachtung aus, d.h. im Alter 44 liegt eine Zensierung der Beobachtung für alle Versicherten des Bestandes vor.

(b) (i) [2 Punkte]

Anzahl der unter Risiko stehenden Personen

Alter j	40	41	42	43	44
Todesfälle d_j	5	6	7	8	0
Storno	0	50	50	0	0
Anzahl unter Risiko n_j	1000	995	939	882	0

(ii) [4 Punkte]

Da bei dieser Teilaufgabe angenommen wird, dass ein Abgang aus dem Bestand auch wegen Storno eintreten kann, ist eine Zensierung auch in den Altern 40 bis 43 möglich. Gemäß der Aufgabenstellung findet eine Zensierung wegen Storno in den Altern 41 und 42 sowie mit Erreichen des Alters 44 aufgrund des Endes der Vertragslaufzeit für alle Versicherten statt.

(iii) [2 Punkte]

Die Schätzer $\hat{p}_j$ lauten

$$\hat{p}_{40} = 1 - \frac{d_{40}}{n_{40}} = 1 - \frac{5}{1000} = 0,9950$$

$$\hat{p}_{41} = 1 - \frac{d_{41}}{n_{41}} = 1 - \frac{6}{995} = 0,9940$$

$$\hat{p}_{42} = 1 - \frac{d_{42}}{n_{42}} = 1 - \frac{7}{939} = 0,9925$$

$$\hat{p}_{43} = 1 - \frac{d_{43}}{n_{43}} = 1 - \frac{8}{882} = 0,9909$$

Die geschätzten Werte $\hat{S}$ der Survivalfunktion lauten dann

$$\hat{S}(40) = 0,9950$$
$$\hat{S}(41) = 0,9950 \cdot 0,9940 = 0,9890$$
$$\hat{S}(42) = 0,9950 \cdot 0,9940 \cdot 0,9925 = 0,9816$$
$$\hat{S}(43) = 0,9950 \cdot 0,9940 \cdot 0,9925 \cdot 0,9909 = 0,9727$$

Aufgabe 2. *[Deskriptive Statistik, Abhängigkeitsmaße] [30 Punkte]*

Für (a)-(e) ist eine bivariate Stichprobe

$$(x_1, y_1), \ldots (x_n, y_n)$$

gegeben, wobei x_i eine Temperatur und y_i der Gasverbrauch ist. Die Werte x_i und y_i sind paarweise verschieden.

In Abbildung 1 links ist das Streudiagramm der Daten, rechts sind die die jeweiligen Ränge der Daten gegeneinander aufgetragen.

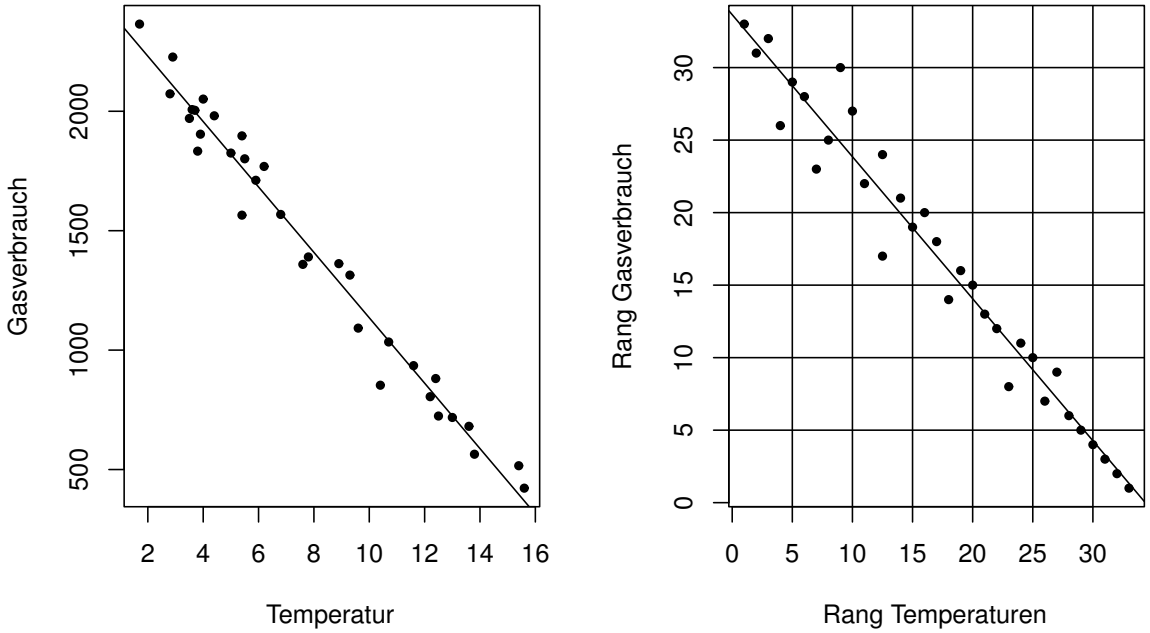


Abbildung 1: Streudiagramm und Rangplot

(a) *[4 Punkte]* Welche Größenordnung und Vorzeichen erwarten Sie für die Korrelationskoeffizienten nach Pearson bzw. Spearman. Begründen Sie Ihre Antwort.

(b) *[4 Punkte]* In Abbildung 1 rechts beträgt die Anzahl der konkordanten Punktepaare 26. Bestimmen Sie die Anzahl der diskordanten Punktepaare und die Rangkorrelation r_τ.

(c) *[3 Punkte]* Es wird ein univariates lineares Modell angepasst für die Zielvariable Gasverbrauch und die Kovariable Temperatur. Für die standardisierten Residuen erhalten wir den Normal Q-Q-Plot in Abbildung 2. Erläutern Sie die Werte auf den Achsen. Welche Verteilung liegt dem Plot zugrunde?

(d) *[11 Punkte]*

(i) *[8 Punkte]* Erstellen Sie einen einfachen Boxplot für die standardisierten Residuen. Für die benötigten Quantile verwenden Sie die Abbildung 2. Geben Sie sie mit einer Genauigkeit von 0,05 an.

(ii) *[3 Punkte]* Gibt es in diesem Fall einen Unterschied zwischen dem einfachen und dem modifizierten Box-Plot? Begründen Sie Ihre Antwort.

(e) *[2 Punkte]* Interpretieren Sie hinsichtlich der Verteilungsannahme den Normal-Q-Q Plot in Abbildung 2 und den Boxplot. Nennen Sie jeweils ein Argument für und eins gegen die Verteilungsannahme.

(f) *[6 Punkte] Bei den folgenden MC-Fragen entscheiden Sie für jede Aussage, ob diese allgemein wahr oder falsch ist. Eine falsche Antwort ergibt 0 Punkte; Minuspunkte werden nicht vergeben. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben. Eine Begründung ist nicht erforderlich.*

A Der Zweistichproben Q-Q Plot ist monoton wachsend.

B Aus dem Zweistichproben Q-Q-Plot kann man erkennen ob Abhängigkeit vorliegt.

C Seien x_i, y_i, $i = 1, \ldots, n$ Realisierungen identisch unabhängig verteilter Zufallsvariablen. Dann liegen die Punkte (x_i, y_i) näherungsweise auf einer Geraden.

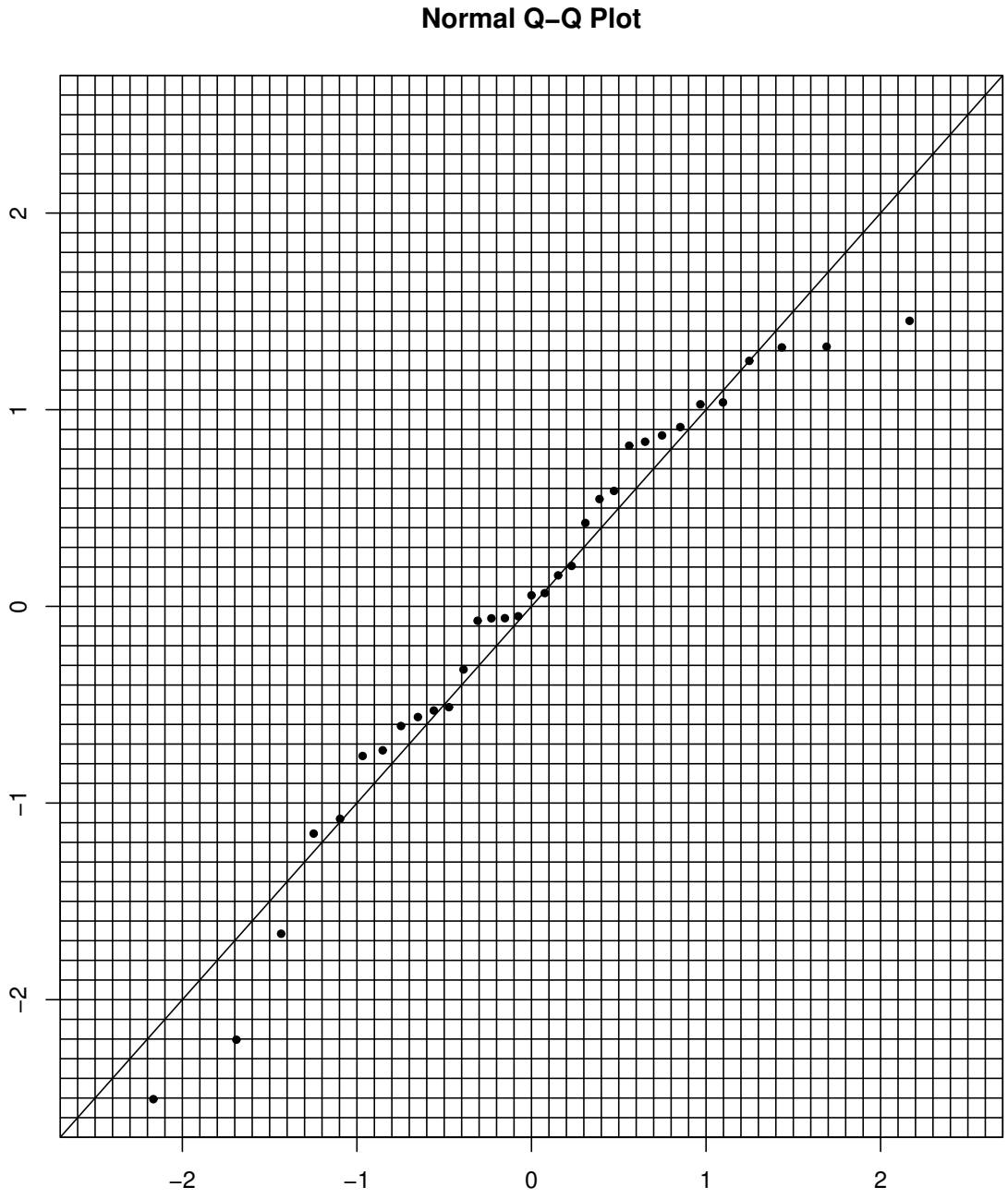


Abbildung 2: Q-Q-Plot der standardisierten Residuen

Lösung

(a) *[4 Punkte = je 2 für die Beschreibung und 2 für die richtige Zuordnung]* Die Korrelation nach Pearson müsste negativ sein und nahe -1, da die Punkte nahe der Ausgleichsgeraden liegen. Dir Rangkorrelation nach Spearman ist nach Definition die Pearson-Korrelation der Ränge $(R(x_i), R(y_i))$. In Abbildung 1 rechts sieht man analog zu links einen negativen Zusammenhang, die Punkte liegen nahe an der Ausgleichsgeraden.

(b) *[4 Punkte = je 2]* Aus der Abbildung 1 rechts liest man $n = 33$ ab. Es gibt damit insgesamt $\binom{33}{2} = 528$ Punktepaare. Da die x_i und y_i paarweise verschieden sind, sind ist jedes Punktepaar konkordant oder diskordant. Damit sind 502 Punktepaare diskordant und

$$r_\tau = 1 - \frac{4 \cdot 502}{33 \cdot 32} = -\frac{119}{132} \approx -0,902$$

(c) *[3 Punkte]* Auf der x-Achse sind die Quantile der Standardnormalverteilung $u_{\frac{i}{34}}$, $i = 1, \ldots, 33$ abgetragen, auf der y-Achse die geordenten standardisierten Residuen. Es liegt die Standardnormalverteilung zugrunde, die eingezeichnete Gerade ist die Identitätsgerade.

(d) *[(i) 8 Punkte=4 für die Werte, 4 für den Plot, (ii) 3 Punkte]*

(i) Bei $n = 33$ erhalten wir wegen $\lfloor 33 \cdot 0,25 \rfloor + 1 = 9$, $\lfloor 33 \cdot 0,5 \rfloor + 1 = 17$, $\lfloor 33 \cdot 0,75 \rfloor + 1 = 25$

$$\begin{aligned}
x_{25\%} &= x_{(9)} \approx -0,55, && \text{(oberes Quartil)}\\
x_{50\%} &= x_{(17)} \approx 0,05, && \text{(Median)}\\
x_{75\%} &= x_{(25)} \approx 0,85, && \text{(unteres Quartil)}\\
x_{(33)} &\approx 1,55 && \text{(Maximum)}\\
x_{(1)} &\approx -2,5 && \text{(Minimum)}
\end{aligned}$$

Boxplot, vgl. Abbildung 3

(ii) Aus den Quartilen ergibt sich der Interquartilsabstand $IQD = 0,85 - (-0,55) \approx 1,4$. Der modifizierte Boxplot sieht genauso aus wie der einfache Boxplot, da es keine Ausreißer gibt:

$$\begin{aligned}
x_{25\%} - 1,5IQD &\approx -0,55 - 2,1 < x_{(1)}\\
x_{75\%} + 1,5IQD &\approx 0,85 + 2,1 > x_{(33)}.
\end{aligned}$$

(e) *[2 Punkte]* Für die Normalverteilung spricht, dass der Q-Q-Plot gut von der Identitäsgeraden angenähert wird. Dagegen sprechen die Punkte an den Rändern, die systematisch von der Geraden abweichen. Dagegen spricht auch, dass der Box-Plot in Abbildung 3 nicht symmetrisch ist.

(f) *[6 Punkte = je 2]* Begründungen sind nicht erforderlich.

A: Richtig, es werden die geordneten Stichproben gegeneinader gezeichnet.
B: Falsch, wegen A.
C: Falsch, das Streudiagramm sollte regellos sein.

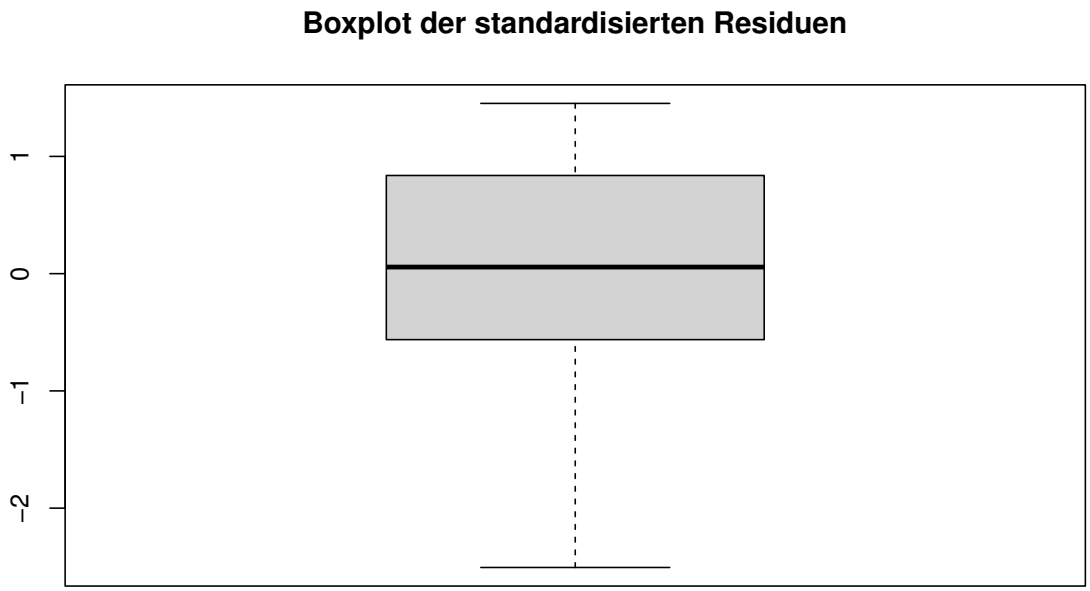


Abbildung 3: Boxplot der standardisierten Residuen

Aufgabe 3. *[36 Punkte]* Induktive Statistik

Gegeben ist die Storno-Statistik von Verträgen nach 3 Tarif-Segmenten und 2 Vertriebswegen. Es sind jeweils die Jahresanfangs-Bestände und die Storni aus diesen Beständen im Laufe des Jahres angegeben.

Jahresanfangs-Bestände			Storni		
Vertriebsweg	1	2	Vertriebsweg	1	2
Segment 1	50000	100000	Segment 1	2000	6000
Segment 2	80000	20000	Segment 2	2000	400
Segment 3	5000	300000	Segment 3	100	15000

(a) *[2 Punkte]* Berechnen Sie die *relativen* Storno-Häufigkeiten für die Kombinationen aus Segment 2 und Vertriebsweg 1 sowie Segment 2 und Vertriebsweg 2.

(b) *[4 Punkte]* Welche diskreten Verteilungen sind zur Modellierung der Storno–Wahrscheinlichkeiten naheliegend? Unterscheiden Sie dabei den Fall, dass ein einzelner Vertrag betrachtet wird und den Fall, dass pro Kombination alle Verträge betrachtet werden, wenn die Unabhängigkeit der Verträge vorausgesetzt werden kann.

(c) *[3 Punkte]* Geben Sie die kanonische Link–Funktion und die Response-Funktion der in b) genannten Verteilung für einzelne Verträge an (verbal und als mathematische Funktion).

(d) *[2 Punkte]* Wie lautet die Modellgleichung für ein GLM–Modell mit Wechselwirkung?

(e) *[17 Punkte]* Die Ausgabe eines GLM mit kanonischer Link–Funktion liefert (Segment entspricht der Variablen `segment`, Vertriebsweg entspricht der Variablen `vert.weg`)

```
Coefficients:
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -3.17805    0.02282 -139.255  < 2e-16 ***
segment2           -0.48551    0.03215  -15.101  < 2e-16 ***
segment3           -0.71377    0.10356   -6.892 5.49e-12 ***
vert.weg2           0.42652    0.02642   16.142  < 2e-16 ***
segment2:vert.weg2 -0.65478    0.06133  -10.675  < 2e-16 ***
segment3:vert.weg2  0.52086    0.10475    4.972 6.61e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

AIC: 66.924
```

(i) *[7 Punkte]* Welche Kodierung wird verwendet? Welche Kategorien sind die Referenzkategorien? Wie sieht die Designmatrix für die 6 verschiedenen Kombinationen aus Segment und Vertriebsweg aus (inklusive Intercept)?

(ii) *[6 Punkte]* Wie würden Sie testen, ob die Wechselwirkung signfikant ist ($\alpha = 0.05$)? Geben Sie die Hypothesen an. Interpretieren Sie die Spalten der folgenden Ausgabe hinsichtlich der zu testenden Wechselwirkung:

```
Analysis of Deviance Table (Type II tests)

                 LR Chisq Df Pr(>Chisq)
segment            750.68  2  < 2.2e-16 ***
vert.weg           241.48  1  < 2.2e-16 ***
segment:vert.weg   172.95  2  < 2.2e-16 ***
```

(iii) *[4 Punkte]* Berechnen Sie die geschätzte Storno–Wahrscheinlichkeit für einen Vertrag in Segment 2 und Vertriebsweg 2. Stimmt diese mit der in (a) berechneten Wahrscheinlichkeit überein? Berechnen sie auf Basis aller Verträge mit dieser Kombination ein approximatives 95%-Konfidenzintervall für die Storno-Wahrscheinlichkeit ($z_{0.975} = 1.96$).

(f) *[8 Punkte]* Eine weitere Analyse liefert folgende Ausgabe:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.10938    0.01961 -158.55   <2e-16 ***
segment2    -0.67169    0.02581  -26.03   <2e-16 ***
segment3    -0.17419    0.01528  -11.40   <2e-16 ***
vert.weg2    0.33345    0.02177     ?     <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

AIC: 235.87
```

(i) *[1 Punkt]* Um welches Modell handelt es sich?

(ii) *[2 Punkte]* Berechnen Sie den Wert für das Fragezeichen.

(iii) *[5 Punkte]* Können Sie auf der Basis dieser Ausgabe und des Modells in e) entscheiden, welches Modell (besser) geeignet ist? Berechnen Sie damit die entsprechende Teststatistik aus (e) (ii).

Lösung Aufgabe 3

(a) *[2 Punkte]* Relative Storno–Häufigkeiten für jede der 6 Zellen (nur 2.Zeile ist gefragt!).

	Vert-Weg 1	Vert-Weg 2
Segment 1	0.04	0.06
Segment 2	**0.025**	**0.02**
Segment 3	0.02	0.05

(b) *[4 Punkte]* Binomialverteilung, da für jeden Vertrag nur 0 (kein Storno) oder 1 (Storno) möglich. Für jeden einzelnen Vertrag erhält man eine Bernoulli-Verteilung. Bei Annahme der Unabhängigkeit der Verträge kann pro Zelle j (”gleiche Kovariablen Segment und Vert-Weg”) eine Binomialverteilung mit n_j, π_j angenommen werden.

(c) *[3 Punkte]* Der kanonische Link ist der Logit–Link.

$$\text{Linkfunktion} : \log(\pi_j/(1-\pi_j)) = x_j'\beta$$

$$\text{Responsefunktion} : \pi_j = \exp(x_j'\beta)/(1+\exp(x_j'\beta))$$

(d) *[2 Punkte]* Die Modellgleichung lautet

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \text{Segment} + \text{Vert-Weg} + \text{Segment*Vert-Weg}$$

Formulierung der Gleichung mit den Dummy-Variablen wird auch gewertet.

(e) *[17 Punkte]*

(i) *[7 Punkte]* Dummy–Kodierung. Kategorie 1 von Segment und Kategorie 1 von Vert-Weg sind die Referenzkategorien.

```
Zelle  (Intercept) segment2 segment3 vert.weg2 segment2:vert.weg2 segment3:vert.weg2
1           1        0        0         0                 0                 0
2           1        0        0         1                 0                 0
3           1        1        0         0                 0                 0
4           1        1        0         1                 1                 0
5           1        0        1         0                 0                 0
6           1        0        1         1                 0                 1
```

(ii) *[6 Punkte]* Die Interaktion hat zwei Parameter. Deshalb kann man nicht einfach die p–Werte der einzelnen Koeffizienten betrachten. Möglich sind prinzipiell Likelihood–Quotienten–Test (LQT), Wald–Test, oder Score–Test.

$$H_0: \quad \beta_{segment2:vert.weg2} = \beta_{segment3:vert.weg2} = 0$$
$$H_1: \quad \text{mindestens einer der beiden Parameter ist nicht 0}$$

Der LQT liefert eine signifikante Wechselwirkung zum Signifikanzniveau $\alpha = 0.05$:

```
Analysis of Deviance Table (Type II tests)

                LR Chisq Df Pr(>Chisq)
segment           750.68  2  < 2.2e-16 ***
vert.weg          241.48  1  < 2.2e-16 ***
segment:vert.weg  172.95  2  < 2.2e-16 ***
```

Die letzte Zeile liefert die Teststatistik (172.95), Freiheitsgrade df=2 und p-Wert (kleiner $\alpha = 0.05$, daher Ablehnung von H_0).

(iii) *[4 Punkte]* Linearer Prädiktor:

$$\eta = -3.17805 - 0.48551 + 0.42652 - 0.65478 = -3.89182$$

Wahrscheinlichkeit:

$$\begin{aligned} \pi_{\text{Segment 2, Vert-Weg 2}} &= \frac{\exp(-3.89182)}{1+\exp(-3.89182)} \\ &= \frac{1}{1+\exp(-(-3.89182)} \\ &= \mathbf{0.02}. \end{aligned}$$

Gleiches Ergebnis wie unter a), da saturiertes Modell. Ein approximatives Konfidenzintervall ist ($n_j = 20000$):

$$0.02 \pm 1.96\sqrt{\frac{0.02 \cdot 0.98}{20000}} = [0.0180597; 0.0219403] \approx [0.018; 0.022].$$

(f) *[8 Punkte]*

(i) *[1 Punkt]* Es handelt sich um das Modell ohne Interaktion.

(ii) *[2 Punkte]* Der z–Wert ist $z = 0.33345/0.02177 = 15.31695$.

(iii) *[5 Punkte]* Es gilt: AIC des Modells ohne Interaktion ist 235.87, das AIC des Modells mit Interaktion aus e) ist 66.924. Damit folgt: Modell mit Interaktion ist besser, da es kleineres AIC hat.

$$\text{LR Chisq} = 235.87 - 66.294 + (12 - 8) = 172.946 \approx 172.95$$

.

Aufgabe 4. *[24 Punkte]* Zeitreihenanalyse

(a) *[3 Punkte]* Nennen Sie drei Ziele der Zeitreihenanalyse.

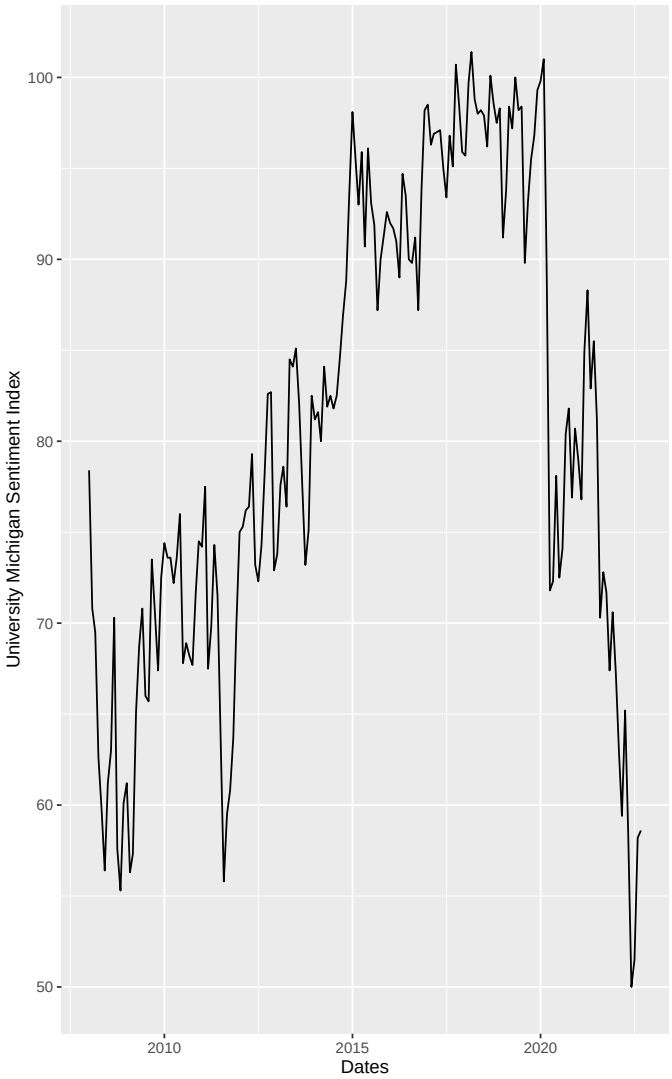
(b) *[6 Punkte]* Entscheiden Sie, ob die folgenden Aussagen richtig oder falsch sind. *Für jede richtige [falsche] Antwort gibt es zwei [Null] Punkte*.

A) Der AR(1) Prozess $y_t = \alpha_1 * y_{t-1} + u_t$, u_t iid N(0,1), ist immer stationär.

B) Bei AR(1) Modellen nimmt die Autokorrelation als Funktion des lags einen monotonen Verlauf an.

C) Das AIC Kriterium kann zur Bestimmung des Parameters p eines AR(p) Prozesses verwendet werden.

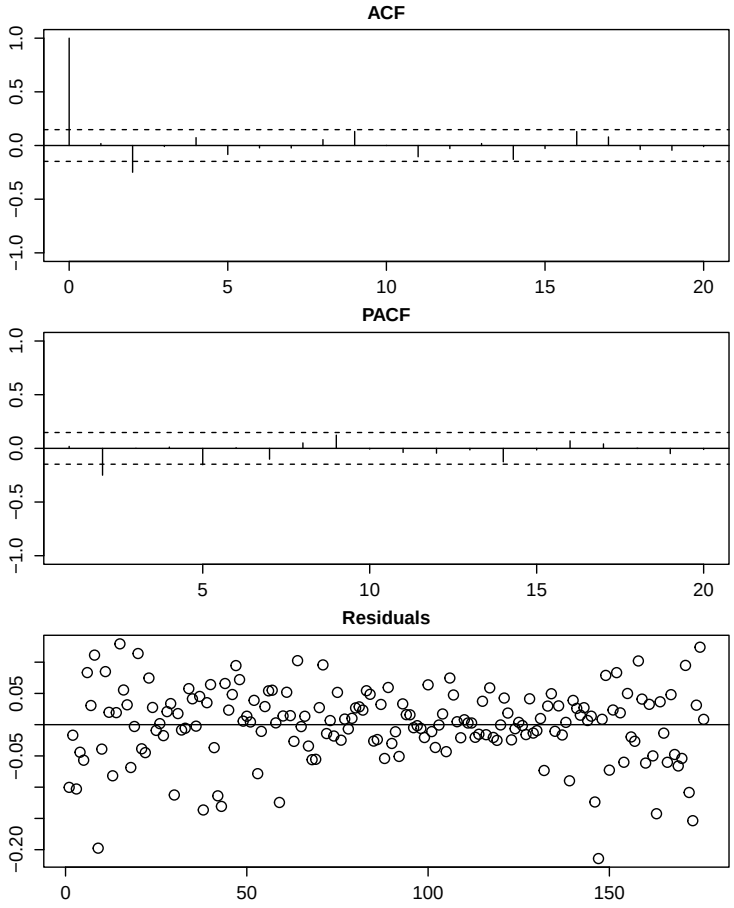
(c) (i) *[3 Punkte]* Die folgende Grafik enthält den monatlichen Sentiment Index der Universität Michigan (Januar 2008 - September 2022), der die Konsumneigung misst. Nennen Sie drei Charakteristika, die geeignet sind den Verlauf der Zeitreihe zu beschreiben!



(ii) *[1 Punkt]* Woran erkennen Sie, dass die Zeitreihe nicht stationär ist?

(iii) *[4 Punkte]* Welche Transformationen bzw. Bereinigungen schlagen Sie vor um eine stationäre Zeitreihe zu erreichen? (zwei Vorschläge mit kurzer Begründung)

(iv) *[3 Punkte]* Für die bereinigte Zeitreihe wurde die Autokorrelationsfunktion bzw. die partielle Autokorrelationsfunktion bzw. die Residuen berechnet und visualisiert. Interpretieren Sie die Grafiken (jeweils 1 Punkt).



(v) *[2 Punkte]* Sie sollen für die bereinigte Zeitreihe ein Modell schätzen. Machen Sie einen Vorschlag und begründen ihn kurz!

(vi) *[2 Punkte]* Wie beurteilen Sie den Einsatz von diesem Zeitreihenmodell (d.h. auf Basis der bereinigten Zeitreihe) für die kurzfristige Prognose (mit kurzer Begründung)?

Lösung

(a) *[3 Punkte]*

(i) Modellierung der Zeitreihe durch (meist stationäre) stochastische Prozesse.

(ii) Prognose zukünftiger Werte, sowie Quantifizierung deren Unsicherheit durch Prognoseintervalle.

(iii) Beschreibung der Zeitreihe grafisch und mit einfachen statistischen Verfahren, z.B. Hervorheben von Trends durch Glättung mittels linearer KQ–Schätzung, gleitender Durchschnitte.

(b) *[6 Punkte]*

A) falsch, z.B. falls $\alpha_1 > 1$ ist.
B) falsch; Gegenbeispiel, negativer Parameter α_1 führt zu Oszillation.
C) richtig.

(c) (i) *[3 Punkte]* Die Werte der Zeitreihe sind positiv. Es gibt Trendperioden (2008-2020, 2020-2022). Lokal monotone Pattern. Die Aussage, dass die Zeitreihe keine saisonalen Komponenten hat wird auch akzeptiert.

(ii) *[1 Punkte]* Die Zeitreihe ist nicht stationär; Gründe: Trend, lokale Pattern (ein Grund genügt).

(iii) *[4 Punkte]* Transformationen

i. Trendbereinigung durch Differenzen erster Ordnung.

ii. Gleitende Durchschnitte.

(iv) *[3 Punkte]* Interpretation

i. Autokorrelationsfunktion: Die Werte sind für alle Lags > 0 nahe dem Wert Null.

ii. Partielle Autokorrelationsfunktion: Die Werte sind für alle Lags nahe dem Wert Null.

iii. Residuen-Plot: Die Residuen streuen um den Wert Null; es gibt zwei Werte mit ca. -0.2

(v) *[2 Punkte]* Aufgrund der Werte der (partiellen) Autokorrelationsfunktion empfiehlt sich ein weißes Rauschen Modell.

(vi) *[2 Punkte]* Da es keine nennenswerten (partiellen) Autokorrelationen gibt, sind die Zeitreihenmodelle nicht hilfreich für die kurzfristige Prognose.

Aufgabe 5. *[Credibility-Theorie, 30 Punkte]*

(a) *[4 Punkte]* Ergänzen Sie im folgenden Lückentext die vier fehlenden spezifischen Fachbegriffe [A] bis [D]. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben.

Im Bayes'schen Credibility-Modell wird die sogenannte [A] *-Verteilung des Strukturparameters* Θ *mittels Beobachtungen der Schadenhöhe* X *in eine angepasste, sogenannte* [B] *-Verteilung umgewandelt. Um daraus die Credibility-Prämie* H^* *zu erhalten, berechnet man mit dieser angepassten Verteilung den* [C] *von* $H(\Theta) = E[X|\Theta]$. *Wenn sich die Credibility-Prämie in der* Form $H^{**} = z\bar{X} + (1-z)E(X)$ *darstellen lässt, spricht man von einer linearisierten Credibility-Prämie.* z *bezeichnet man darin als* [D].

(b) *[26 Punkte]* Die jährliche Schadenanzahl X in einem Kollektiv von Haftpflichtversicherungen wird durch eine Poissonverteilung mit Erwartungswert ϑ beschrieben.

Der Parameter ϑ der Poissonverteilung wird als Strukturparameter Θ in einem Bayes'schen Credibility-Modell betrachtet. Als a-priori-Verteilung von Θ wird eine Gammaverteilung $\Gamma(\alpha, \lambda)$ mit den Parametern $\alpha, \lambda > 0$ angesetzt.

Hinweis: Ohne Beweis können Sie in <u>allen</u> folgenden Teilaufgaben benutzen, dass für eine $\Gamma(\alpha, \lambda)$–Verteilung der Erwartungswert $\frac{\alpha}{\lambda}$ beträgt und die Varianz $\frac{\alpha}{\lambda^2}$.

(i) *[2 Punkte]* Berechnen Sie das Verhältnis $\sqrt{Var(\Theta)}/E(\Theta)$ von Standardabweichung und Varianz der a-priori-Verteilung von Θ in Abhängigkeit des Parameters α.

Welcher Wert für α ergibt sich, wenn das Verhältnis 20% beträgt? *(Kontrollergebnis* $\alpha = 25$*).*

(ii) *[4 Punkte]* Berechnen Sie die erwartete Schadenanzahl $E(X)$ in Abhängigkeit der Parameter α und λ.

Welchen Wert für λ setzt man an, wenn die erwartete Schadenzahl 20 beträgt und $\alpha = 25$ ist? *(Kontrollergebnis* $\lambda = 1{,}25$*).*

(iii) *[6 Punkte]* In $n = 2$ Beobachtungsjahren wurde die mittlere Schadenanzahl $\bar{X} = 15$ beobachtet. Berechnen Sie mit den Angaben aus (b2)

den Wert des Credibility-Schätzers $H^* = E[H(\Theta)|\mathbf{X}]$, wobei $H(\Theta) = E[X|\Theta]$ und $\mathbf{X}$ die beobachteten Schadenanzahlen bezeichnet. Begründen Sie die einzelnen Teilschritte Ihrer Rechnung.

(Hinweis: Ohne Beweis können Sie davon ausgehen, dass die a-posteriori-Verteilung des Strukturparameters Θ unter den gegebenen Beobachtungen eine $\Gamma(\alpha + n\,\bar{X}, \lambda + n)$–Verteilung ist.)

(iv) *[14 Punkte]* Berechnen Sie den Wert des linearisierten Credibility-Schätzers $H^{**} = z\,\bar{X} + (1 - z)\,E(X)$, der sich ergibt wenn man für die Verteilung der Schadenanzahl X anstelle der Poissonverteilung eine Binomialverteilung $X \sim Bin(40, \Theta)$ mit Strukturparameter Θ verwendet.

Als a-priori-Verteilung von Θ wird dabei eine Verteilung mit $E(\Theta) = 0{,}5$ sowie $Var(\Theta) = 0{,}01$ und $E(\Theta^2) = 0{,}26$ angesetzt. Es liegen wieder $n = 2$ Beobachtungsjahre mit $\bar{X} = 15$ vor.

Lösung:

(a) A = a-priori

B = a-posteriori

C = Erwartungswert

D = Credibility-Faktor

(b)

(i) Unter den Voraussetzungen der Aufgabe gilt

$$\frac{\sqrt{Var(\Theta)}}{E(\Theta)} = \frac{\sqrt{\alpha}/\lambda}{\alpha/\lambda} = \frac{1}{\sqrt{\alpha}}.$$

Aus $\frac{1}{\sqrt{\alpha}} = 20\%$ folgt $\alpha = 25$.

(ii) Da X bei gegebenem Θ einer Poissonverteilung mit Erwartungswert Θ folgt, ergibt sich unter Nutzung des Hinweises

$$E(X) = E(E[X|\Theta]) = E(\Theta) = \frac{\alpha}{\lambda}.$$

Man fordert somit $20 = \frac{\alpha}{\lambda} = \frac{25}{\lambda}$, also $\lambda = \frac{25}{20} = 1{,}25$.

(iii) Da X bei gegebenem Θ einer Poissonverteilung mit Parameter Θ folgt, gilt $H(\Theta) = E[X|\Theta] = \Theta$, so dass

$$H^* = E[H(\Theta)|\mathbf{X}] = E[\Theta|\bar{X}] = \frac{\alpha + n\,\bar{X}}{\lambda + n},$$

wobei sich die letzte Gleichheit aus dem Erwartungswert der angegebenen a-posteriori-Verteilung ergibt. Für den konkreten Wert von H^* ergibt sich damit

$$H^* = \frac{25 + 2 \cdot 15}{1{,}25 + 2} = 16{,}92.$$

(iv) Aufgrund von $X \sim Bin(40, \Theta)$ ergibt sich $H(\Theta) = E[X|\Theta] = 40 \cdot \Theta$ und $Var[X|\Theta] = 40 \cdot \Theta(1 - \Theta)$.

Für die Berechnung des Credibility-Faktors z benötigt man

$$Var\big(H(\Theta)\big) = Var(40 \cdot \Theta) = 1600 \cdot Var(\Theta) = 1600 \cdot 0{,}01 = 16$$

und

$$E(Var[X|\Theta]) = 40 \cdot E\big(\Theta(1-\Theta)\big) = 40 \cdot \{E(\Theta) - E(\Theta^2)\} = 40 \cdot \{0{,}5 - 0{,}26\} = 9{,}6.$$

Daraus folgt der Credibility-Faktor

$$z = \frac{Var\big(H(\Theta)\big)}{\frac{1}{n}E(Var[X|\Theta]) + Var\big(H(\Theta)\big)} = \frac{16}{\frac{9{,}6}{2} + 16} = 0{,}7692.$$

Mit

$$E(X) = E\big(H(\Theta)\big) = 40 \cdot E(\Theta) = 40 \cdot 0{,}5 = 20$$

ergibt sich schließlich

$$H^{**} = z\,\bar{X} + (1-z)\,E(X) = 0{,}7692 \cdot 15 + (1 - 0{,}7692) \cdot 20 = 16{,}15.$$

Aufgabe 6. *[Stochastische Prozesse und deren Simulation] [30 Punkte]*
Zur Modellierung des Ausfallrisikos verwendet ein Kreditversicherer drei Bonitätsklassen A, B, C sowie die Ausfallklasse D. Die Zugehörigkeit eines Kreditnehmers zu einer der Klassen zum Zeitpunkt $k \in \mathbb{N}_0$ sei mit X_k bezeichnet und hängt nur von der Ausfallklasse im Zeitpunkt $k-1$ ab. Findet ein Ausfall statt, ist der Kreditnehmer zum nächsten ganzzahligen Zeitpunkt in Klasse D und verbleibt dort.

(a) Gegeben seien die folgenden Wahrscheinlichkeiten:

$$\begin{array}{lll} P(X_{k+1}=A\,|\,X_k=A)=0,8 & P(X_{k+1}=B\,|\,X_k=A)=0,15 & P(X_{k+1}=D\,|\,X_k=A)=0 \\ P(X_{k+1}=B\,|\,X_k=B)=0,65 & P(X_{k+1}=C\,|\,X_k=B)=0,15 & P(X_{k+1}=D\,|\,X_k=B)=0,1 \\ P(X_{k+1}=A\,|\,X_k=C)=0,05 & P(X_{k+1}=B\,|\,X_k=C)=0,45 & P(X_{k+1}=C\,|\,X_k=C)=0,25 \end{array}$$

(i) *[8 Punkte]* Geben Sie die fehlenden Wahrscheinlichkeiten an und begründen Sie, warum es sich bei $(X_k)_{k\in\mathbb{N}}$ um eine homogene Markov-Kette mit endlichem Zustandsraum handelt.

(ii) *[2 Punkte]* Stellen Sie die zugehörige Übergangsmatrix Π auf.

(iii) *[4 Punkte]* Eine Berechnung der Eigenwerte und -vektoren von Π in R ergibt folgenden Output (`$values` sind die Eigenwerte, `$vectors` die Eigenvektoren):

```
> eigen(Pi)
eigen() decomposition
$values
[1] 1.0000000 0.9223298 0.6552204 0.1224498

$vectors
     [,1]      [,2]       [,3]       [,4]
[1,]  0.5 0.7726441  0.7038410 -0.0108969
[2,]  0.5 0.4995178 -0.5169395 -0.2715648
[3,]  0.5 0.3917947 -0.4872181  0.9623585
[4,]  0.5 0.0000000  0.0000000  0.0000000
```

Leiten Sie damit die Existenz einer asymptotischen Verteilung der Markov-Kette her.

(iv) *[4 Punkte]* Geben Sie die asymptotische Verteilung an. Begründen Sie Ihre Antwort.

(b) Die Anzahl N_t der Ausfallereignisse bis zum Zeitpunkt t im Portfolio des Kreditversicherers sei als homogener Poisson-Prozess mit Intensität $\lambda = 4$ modelliert. Dieser Prozess soll mit der Monte-Carlo-Methode simuliert werden. Dazu werden die Zwischenankunftszeiten zwischen aufeinander folgenden Ausfallereignissen betrachtet.

(i) *[4 Punkte]* Welcher Verteilung genügen diese Zwischenankunftszeiten? Geben Sie einen Simulationsalgorithmus für diese Verteilung an, der auf einer direkten Verwendung gleichverteilter Zufallszahlen beruht. Wie heißt die von Ihnen genannte Methode?

(ii) *[8 Punkte]* Mit dem Befehl `runif(3)` wurden in R die folgenden Zahlen generiert:

$$u_1 = 0,2244 \qquad u_2 = 0,8072 \qquad u_3 = 0,2167$$

Verwenden Sie diese Zahlen und den Algorithmus aus (e), um $N_{0,5}$ zu simulieren.

Lösungen

(ai) Die fehlenden Wahrscheinlichkeiten ergeben sich dadurch, dass die Summe der Wahrscheinlichkeiten bei gegebener Bedingung eins sein muss:

$$\begin{aligned} P(X_{k+1} = C \,|\, X_k = A) &= 1 - (0,8 + 0,15 + 0) = 0,05 \\ P(X_{k+1} = A \,|\, X_k = B) &= 1 - (0,65 + 0,15 + 0,1) = 0,1 \\ P(X_{k+1} = D \,|\, X_k = C) &= 1 - (0,05 + 0,45 + 0,25) = 0,25. \end{aligned}$$

Für die Bedingung $X_k = D$ ist zu beachten, dass nach einem Ausfall kein Weg zurückführt (ein sog. absorbierender Zustand), so dass

$$P(X_{k+1} = A \,|\, X_k = D) = P(X_{k+1} = B \,|\, X_k = D) = P(X_{k+1} = C \,|\, X_k = D) = 0$$

sowie

$$P(X_{k+1} = D \,|\, X_k = D) = 1.$$

Man erkennt, dass die Kette den endlichen Zustandsraum $S = \{A, B, C, D\}$ besitzt und die Eintrittswahrscheinlichkeiten für X_{k+1} nach Aufgabenstellung nur vom Wert von X_k abhängt, und kein X_i für $i < k$ eine Rolle spielt. Damit ist die Markoveigenschaft erfüllt. Da die gegebenen Wahrscheinlichkeiten ferner nicht vom Wert von k abhängen, ist die Kette auch homogen.

(aii) Die gegebenen und berechneten Werte werden als Matrix zusammengestellt:

$$\Pi = \begin{pmatrix} 0,8 & 0,15 & 0,05 & 0 \\ 0,1 & 0,65 & 0,15 & 0,1 \\ 0,05 & 0,45 & 0,25 & 0,25 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(aiii) Die Eigenwerte lauten

$$\lambda_1 = 1, \quad \lambda_2 = 0,9223298, \quad \lambda_3 = 0,6552204, \quad \lambda_4 = 0,1224498.$$

Der Konvergenzsatz für endliche homogene Markov-Ketten ist anwendbar: Der Eigenwert λ_1 ist der einzige vom Betrag eins; zudem sind alle Eigenwerte verschieden, so dass jeder (und damit auch λ_1) die algebraische Vielfachheit eins besitzt. Damit ist die Existenz einer asymptotischen Verteilung – unabhängig vom Startzustand – gesichert.

(aiv) Die asymptotische Verteilung ist die (eindeutige) normierte Lösung der Gleichung $\boldsymbol{p} = \boldsymbol{p} \cdot \Pi$. Man erkennt sofort, dass $\boldsymbol{p} = (0, 0, 0, 1)$ eine solche Lösung und damit die asymptotische Verteilung ist.

(bi) Die Zwischenankunftszeiten eines Poisson-Prozesses mit Intensität λ sind exponentialverteilt mit Parameter λ. Eine solche Verteilung kann auf Basis einer $U(0,1)$-Zufallszahl u simuliert werden über

$$x = -\frac{1}{\lambda} \ln(u).$$

(Alternativ auch $x = -\frac{1}{\lambda} \ln(1-u)$.) Dies ist die Inversionsmethode.

(bii) Man erhält auf Basis der gegebenen Zahlen (welche U(0, 1)-Zufallszahlen sind) und $\lambda = 4$ die Zwischenankunftszeiten

	Mit $\ln(u)$	Mit $\ln(1-u)$
u_1	$0,373581276$	$0,063529589$
u_2	$0,053545952$	$0,411525474$
u_3	$0,382310343$	$0,061059879$

Es gilt somit für die Zeitpunkte der Ausfälle (Summe der Zwischenankunftszeiten)

	Mit $\ln(u)$	Mit $\ln(1-u)$
1. Ausfall	$0,373581276$	$0,063529589$
2. Ausfall	$0,427127229$	$0,475055063$
3. Ausfall	$0,809437571$	$0,536114942$

In beiden Fällen ist der simulierte Wert für $N_{0,5}$ demnach 2.