Schriftliche Prüfung im Grundwissen

Angewandte Stochastik

Klausur mit Lösungen

gemäß Prüfungsordnung 5
der Deutschen Aktuarvereinigung e. V.

am 17. Mai 2024

Hinweise:

- Als Hilfsmittel für die Klausur sind Literatur und Unterlagen, die inhaltlich (ggf. teilweise) den Lernzielkatalog behandeln, sowie frühere Klausuren mit Lösungsvorschlägen jeweils in Papierform und 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 besteht aus 28 Seiten.
- Mit Ausnahme der Multiple-Choice-Fragen sind alle Antworten zu begründen und bei Rechenaufgaben muss der Lösungsweg ersichtlich sein.
- 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]*

Es wird ein Bestand von Rentenversicherungen untersucht, in dem alle Zugänge bis zum Alter 70 erfolgen. Ein Storno der Rentenversicherung ist nicht möglich. Ausgewertet werden die Alter 71 bis 75 bezüglich der Sterblichkeit. Die Lebensdauern in dem untersuchten Bestand von zehn versicherten Personen (VP) werden wie in der folgenden Tabelle aufbereitet. Hierbei bedeutet

1 die Person hat das jeweilige Alter überlebt

X die Person ist in dem Alter verstorben.

	Lebensdauer bzw. Alter				
Person	71	72	73	74	75
P_1	1	X			
P_2	1	1	1	X	
P_3	X				
P_4	1	1	1	1	1
P_5	1	X			
P_6	1	1	X		
P_7	1	1	1	1	1
P_8	1	1	1	1	X
P_9	1	1	1	1	1
P_{10}	1	1	1	1	1

(a) *[8 Punkte]* Ermitteln Sie mit der Geburtsjahrmethode die Sterblichkeiten für jedes der Alter 71 bis 75.

(b) *[22 Punkte]*

(i) *[6 Punkte]* Geben Sie allgemein den Nelson-Aalen-Schätzer für die Hazardfunktion an. Passen Sie die Formulierung an die vorliegende Aufgabe an.

Leiten Sie aus der obigen Tabelle die in dem Nelson-Aalen-Schätzer verwendeten Realisationen d_j und n_j der Zufallsvariablen D_j und N_j ab.

(ii) *[6 Punkte]* Begründen Sie, dass für $N_j = n_j$ die Zufallsvariable $(n_j - D_j)$ binomialverteilt ist und geben Sie die Parameter der Binomialverteilung an.

(iii) *[6 Punkte]* Zeigen Sie, dass für $j = 71, \ldots, 75$ für den Erwartungswert und die Varianz von $1 - \frac{D_j}{n_j}$ gilt

$$E\left(1 - \frac{D_j}{n_j}\right) = 1 - q_j = p_j$$
$$Var\left(1 - \frac{D_j}{n_j}\right) = \frac{p_j(1-p_j)}{n_j}.$$

(iv) *[4 Punkte]* Ermitteln Sie für $j = 71, \ldots, 75$ Schätzwerte der Hazardfunktion mit Hilfe des Nelson-Aalen-Schätzers aus (i).

Lösung

(a) Bei der Geburtsjahrmethode werden nur die Geburtsjahrgänge betrachtet, deren Todesfälle im Alter x ausschließlich in den Beobachtungszeitraum fallen können. Geburtsjahrgänge, deren Todesfälle im Alter x auch vor oder nach dem Beobachtungszeitraum auftreten können, bleiben dabei völlig unberücksichtigt.

Bezeichne

N_j die unter Risiko stehende Personengesamtheit des Alters j,

D_j Tote aus N_j

B = Beobachtungszeitraum

G = Geburtsjahr

$N_j(B, G)$ die Personen aus der Personengesamtheit, die im Beobachtungszeitraum B das Alter j erreichen und das $(j+1)$-te Lebensjahr vollenden (könnten) und die im Geburtsjahr G geboren wurden,

$D_j(B, G)$ die Personen aus $N_j(B, G)$, die im Beobachtungszeitraum sterben.

$N_j(B) := \sum_G N_j(B, G)$ Summe der $N_j(B, G)$ über alle Geburtsjahre

$D_j(B) := \sum_G D_j(B, G)$ Summe der $D_j(B, G)$ über alle Geburtsjahre

Die rohe Sterbewahrscheinlichkeit nach der Geburtsjahrmethode ohne Unterscheidung der Geburtsjahre ist dann definiert durch:

$$q_j = \frac{\left|D_j(B)\right|}{\left|N_j(B)\right|}$$

Anzahl der Toten, Anzahl der unter Risiko stehenden Personen und Sterblichkeiten:

Alter j	71	72	73	74	75
Todesfälle D_j	1	2	1	1	1
Anzahl unter Risiko N_j	10	9	7	6	5
Sterblichkeit q_j	0,100	0,222	0,143	0,167	0,200

(b) (i) Der Nelson-Aalen-Schätzer für die kumulierte Hazard-Funktion lautet

$$\hat{\Lambda}(t) = \begin{cases} 0 & t < t_{(1)} \\ \sum_{j|t_{(j)} \leq t} \frac{d_j}{n_j} & \text{sonst.} \end{cases}$$

Da sich die Aufgabenstellung auf Alter bezieht, handelt es sich um diskrete Zeitpunkte im Abstand von einem Jahr. Demzufolge gilt $t(1) = 71, \ldots, t(5) = 75$ und der Nelson-Aalen-Schätzer lautet für $t \in \mathbb{N}$

$$\hat{\Lambda}(t) = \begin{cases} 0 & t < 71 \\ \sum_{j=71}^{t} \frac{d_j}{n_j} & \text{sonst.} \end{cases} \tag{1}$$

Für $x = j$ gilt $L_x(B, G) = n_j$ und $T_x(B, G) = d_j$

(ii) Am Beginn des Alters j stehen n_j Personen unter Risiko. Im Alter j versterben $D_j = d_j$ versicherte Personen. Dann ist $n_j - d_j = n_{j+1}$ die Anzahl der Personen unter Risiko im Alter $j + 1$. Dann ist

$$p_j = P(T > t_j | T > t_{j-1}), j = 71, \ldots, 75$$

die Überlebenswahrscheinlichkeit im Alter j und es gilt $(n_j - D_j)$ ist binomialverteilt $B(n_j, p_j)$.

(iii) Für den Erwartungswert und die Varianz gilt dann

$$\begin{aligned} E(n_j - D_j) &= n_j - E(D_j) \\ &= n_j - n_j q_j \\ &= n_j - n_j(1 - p_j) \\ &= n_j p_j \\ Var(n_j - D_j) &= n_j p_j (1 - p_j) \end{aligned}$$

Daraus ergibt sich für den Erwartungswert und die Varianz von $(1 - \frac{D_j}{n_j})$

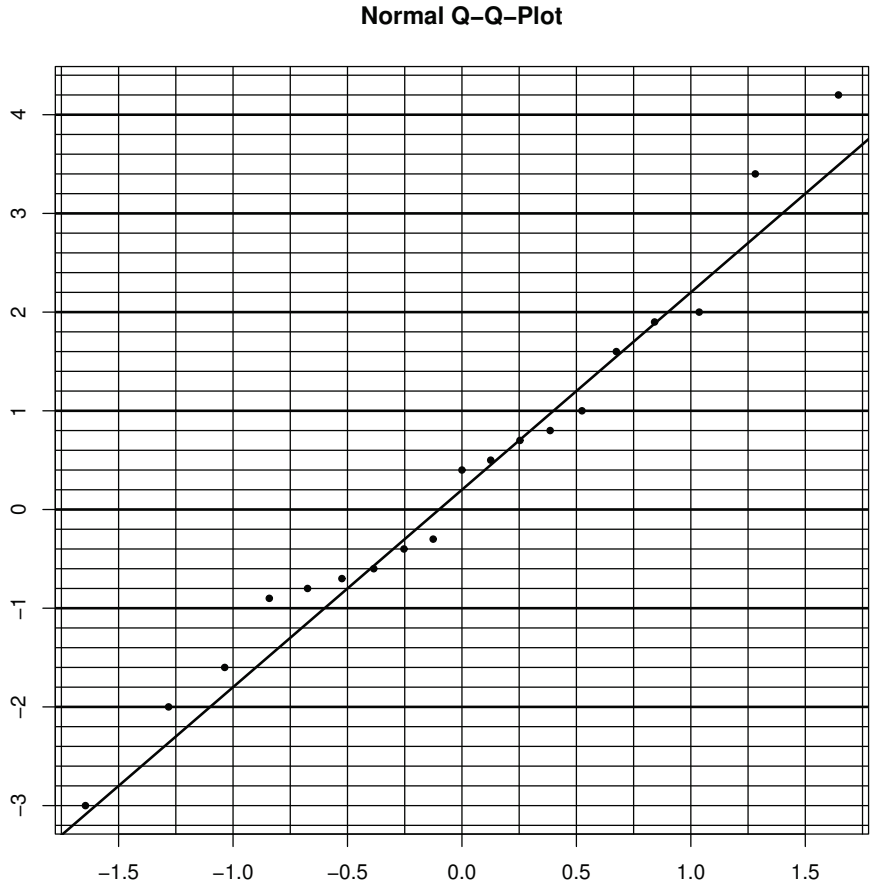
$$\begin{aligned} E\left(1 - \frac{D_j}{n_j}\right) &= \frac{1}{n_j} E(n_j - D_j) = p_j \\ Var\left(1 - \frac{D_j}{n_j}\right) &= \frac{1}{{n_j}^2} Var(n_j - D_j) \\ &= \frac{1}{{n_j}^2} n_j p_j (1 - p_j) \\ &= \frac{p_j(1 - p_j)}{n_j} \end{aligned}$$

(iv) Schätzwerte der Hazardfunktion (vgl. (i))

Alter t bzw. j	71	72	73	74	75
Todesfälle d_j	1	2	1	1	1
Anzahl unter Risiko n_j	10	9	7	6	5
Sterblichkeit $\frac{d_j}{n_j}$	0,100	0,222	0,143	0,167	0,200
Hazardfunktion $\sum_{j=71}^{t} \frac{d_j}{n_j}$	0,100	0,322	0,465	0,632	0,832

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

(a) *[22 Punkte]* Gegeben seien Versicherungsdaten $x_1, \ldots, x_{19}$. Die nachfolgende Grafik zeigt den zu den Daten zugehörigen Normal Q-Q-Plot und die Anpassungsgerade.



(i) *[3 Punkte]* Erläutern Sie die Grafik. Erklären Sie insbesondere die Größen, die auf x- und y-Achse aufgetragen sind. Welche Verteilung liegt dem Q-Q-Plot zugrunde?

(ii) *[11 Punkte]* Erstellen Sie einen Boxplot zu den Daten $x_1, \ldots, x_{19}$. Begründen Sie die gewählten bzw. die ausgerechneten Werte. Gibt es Ausreißer?

(iii) *[2 Punkte]* Ist die Verteilungsannahme plausibel? Begründen Sie Ihre Entscheidung, nennen Sie zwei Gründe.

(iv) *[3 Punkte]* Bestimmen Sie aus der Grafik plausible Schätzer für die Parameter der zugrundeliegenden Verteilung.

(v) *[3 Punkte]* In der Modellierung wird oft die Lognormal-Verteilung verwendet. Kommt diese Verteilung hier in Frage? Begründen Sie Ihre Antwort.

(b) *[8 Punkte]* Bei den folgenden Multiple Choice-Fragen entscheiden Sie für jede Aussage, ob diese allgemein richtig oder falsch ist. Eine falsche Antwort ergibt 0 Punkte, und für jede richtige Antwort gibt es 2 Punkte; Minuspunkte werden nicht vergeben. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben. Eine Begründung ist nicht erforderlich.

(i) Aus dem Q-Q-Plot für eine Stichprobe kann man die empirische Verteilungsfunktion erstellen.

(ii) Aus einem Histogramm für eine Stichprobe kann man einen Q-Q-Plot erstellen.

(iii) Seien $(x_1, y_1)^\top, \ldots, (x_n, y_n)^\top$ unabhängige Realisierungen des Zufallsvektors $(X, Y)^\top$ wobei die Zufallsvariablen X, Y nicht unabhängig seien. Dann liegen die Punkte (x_i, y_i), $i = 1, \ldots, n$ näherungsweise auf einer Geraden.

(iv) Gegeben seien zwei Stichproben $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ mit paarweise verschiedenen x_i und y_i, $i = 1, \ldots, n$. Seien $x_{(i)}$, $y_{(i)}$ die Ordnungsstatistiken. Dann ist Kendalls τ von $(x_{(i)}, y_{(i)})$ gleich eins.

Lösung

(a) *[22 Punkte]*

(i) *[3 Punkte]* Auf der x-Achse sind die Quantile der Standardnormalverteilung $u_{k/20}$, $k = 1, \ldots, 19$ abgetragen, auf der y-Achse die $x_{(k)}$, $k = 1, \ldots, 19$, wobei $x_{(k)}$ die aufsteigend geordnete Stichprobe bezeichnet. Dem Q-Q-Plot liegt demnach die Standardnormalverteilung zugrunde.

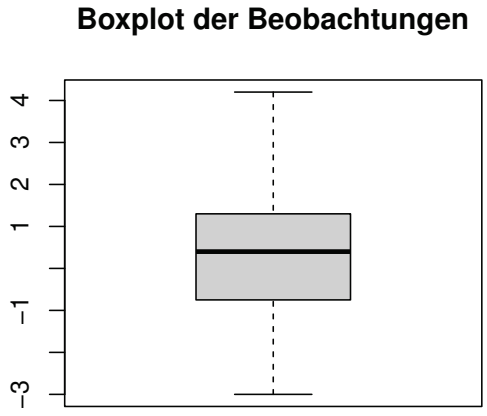
(ii) *[11 Punkte]*

$$\begin{aligned} x_{25\%} = x_{(5)} &\approx -0,8, && \text{(oberes Quartil)} \\ x_{50\%} = x_{(10)} &\approx 0,4, && \text{(Median)} \\ x_{75\%} = x_{(15)} &\approx 1,6 && \text{(unteres Quartil)} \\ x_{(19)} &\approx 4,2 && \text{(Maximum)} \\ x_{(1)} &\approx -3 && \text{(Minimum)} \end{aligned}$$

Aus den Quartilen ergibt sich der Interquartilsabstand $IQD = x_{(15)} - x_{(5)} = 2,4$. Ferner gilt

$$\begin{aligned} x_{25\%} - 1,5IQD &\approx -0,8 - 3,6 < x_{(1)} \\ x_{75\%} + 1,5IQD &\approx 1,6 + 3,6 > x_{(19)}. \end{aligned}$$

Somit gibt es keine Aureißer.



(iii) *[2 Punkte]* Die Annahme der Normalverteilung ist plausibel. Die Anpassungsgerade passt die Werte gut an. Außerdem ist der Boxplot symmetrisch, so wie es bei der Normalverteilung sein sollte.

(iv) *[3 Punkte]* Man liest bei der eingezeichneten Ausgleichsgeraden den Achsenabschnitt $\hat{\mu} = 0,2$ und die Steigung $\hat{\sigma} = 2$ ab, also $\hat{\sigma}^2 = 4$.

(v) *[3 Punkte]* Die Lognormalverteilung kommt nicht in Frage, da negative Realisierungen vorliegen.

(b) *[8 Punkte = je 2]* Begründungen sind nicht erforderlich.

(i) Richtig

(ii) Falsch

(iii) Falsch

(iv) Richtig

Aufgabe 3. *[36 Punkte]* Induktive Statistik

Eine Autoversicherung zählt die Schäden pro Versicherungsrisiko i, $i = 1, \ldots, n$ (Zielvariable: *Anzahl.Schäden*). Weiter stehen das metrische Merkmal *Fahrzeugwert* (in Einheiten von 10000 Euro) und das Merkmal *Region* (Ausprägungen: `A, B, C, D`) zur Verfügung. Sie möchten herausfinden, ob und wie die Anzahl der Schäden von diesen beiden Merkmalen abhängt.

(a) *[6 Punkte]* Fragen zum Modell

(i) *[2 Punkte]*

i. Welche Annahme treffen Sie für die Verteilung der Zielvariable (Begründung)? Welches Regressionsmodell verwenden Sie (nur verbal)?

ii. Geben Sie die Likelihood der n Beobachtungen als Funktion der Erwartungswerte μ_i der Zielvariable an.

(ii) *[4 Punkte]* Geben Sie die kanonische Link–Funktion und die Response-Funktion an, die Sie verwenden. Formulieren Sie mathematisch beides als Funktion der Erwartungswerte der Zielvariable. (*Verwenden Sie dazu ggf. in allgemeiner Form den Vektor der Kovariablen $\mathbf{x}_i$ und den Parametervektor $\boldsymbol{\beta}$*)

(b) *[24 Punkte]* Die Analyse eines Modells mit kanonischer Link–Funktion liefert folgende Ausgabe (R Output):

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.71669    0.03533 -76.894  < 2e-16 ***
Fahrzeugwert  0.04999    0.01065   4.696 2.66e-06 ***
RegionB       0.02008    0.03449   0.582   0.5605
RegionC      -0.12594    0.05249      ?    0.0164 *
RegionD       0.13359    0.06452   2.070   0.0384 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 26768  on 67855  degrees of freedom
Residual deviance: 26732  on 67851  degrees of freedom
AIC: 36176
```

(i) *[1 Punkt]* Wie viele Versicherungsrisiken n sind in der Analyse berücksichtigt? *(Kurze Begründung!)*

(ii) *[5 Punkte]*

i. *[2 Punkte]* Wie lautet die Modellgleichung des entsprechenden Modells? Berücksichtigen Sie dabei, dass *Region* eine kategoriale Variable ist.

ii. *[3 Punkte]* Welche Kodierung wird für *Region* verwendet? Welche Kategorie ist die Referenzkategorie? Geben Sie die Kodierungsvektoren für die *Region* `A` und die *Region* `D` an!

(iii) *[2 Punkte]* Berechnen Sie den fehlenden Wert der Ausgabe gekennzeichnet durch '?'.

(iv) *[4 Punkte]* Wie testen Sie, ob der *Fahrzeugwert* einen signifikanten Einfluss auf die Anzahl der Schäden hat? ($\alpha = 0.05$). Geben Sie die Hypothesen H_0 und H_1 an. Geben Sie an, wie die Teststatistik berechnet wird und welcher Verteilung diese asymptotisch folgt. Wie lässt sich die asymptotische Verteilung begründen?

(v) *[4 Punkte]* Wie testen Sie, ob die *Region* **insgesamt** einen signifikanten Einfluss auf die Anzahl der Schäden hat? ($\alpha = 0.05$). Geben Sie die Hypothesen H_0 und H_1 an. Können Sie diese Hypothesen ohne weitere Informationen mit Hilfe des R Outputs überprüfen?

(vi) *[6 Punkte]* Sie erhalten die folgende Ausgabe:

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

Response: Anzahl.Schäden
             LR Chisq Df Pr(>Chisq)
Fahrzeugwert   20.315  1  6.567e-06 ***
Region         14.900  3   0.001904 **
```

Wie sind die Spalten zu interpretieren? Gehen Sie dabei auch auf die folgenden Fragestellungen ein: Wie wird der Wert in der Spalte `LR Chisq` berechnet? Was lässt sich über dessen Verteilung sagen? Wie kann der Wert in Spalte `Pr(>Chisq)` zur Testentscheidung verwendet werden?

(vii) *[2 Punkte]* Berechnen Sie die geschätzte Anzahl der Schäden für einen Vertrag eines 30000 Euro teuren Fahrzeugs in *Region* `B`.

(c) *[6 Punkte]*

Ein weiteres Modell mit einer im Vergleich zum Modell in (b) *zusätzlichen* (metrischen) Kovariable im Modell liefert folgende Ausgabe:

```
    Null deviance: 26768  on 67855  degrees of freedom
Residual deviance: 25435  on 67850  degrees of freedom
AIC: 34881
```

(i) *[2 Punkte]* Begründen Sie, ob die Variable in das Modell aufgenommen werden sollte oder nicht.

(ii) *[4 Punkte]* Berechnen Sie die entsprechende Statistik des Likelihoodquotienten-Tests.

Lösung Aufgabe 3

(a) *[6 Punkte]* Fragen zum Modell

(i) *[2 Punkte]*

i. Poisson-Verteilung, da die Zielvariable eine Anzahl ist, Poisson-GLM/Poisson-Regressionsmodell (*1P*)

ii. `Anzahl.Schäden` y_i:

$$L = \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} \exp(-\mu_i)$$

(*1P*)

(ii) *[4 Punkte]* Kanonischer Link: log-Link, $\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$, wobei $\mu_i = E(Y_i|\mathbf{x}_i)$ die Erwartungswerte sind ($i = 1, \ldots, n$), also die erwartete Anzahl an Schäden pro Versicherungsrisiko i (*2P*)

Response-Funktion: Exponentialfunktion, $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ (*2P*)

(b) *[24 Punkte]*

(i) *[1 Punkt]* Dies lässt sich aus den Angaben der Freiheitsgrade ('degrees of freedom') zur 'deviance' (Devianz) ablesen: $n = 67855 + 1 = 67851 + 5 =$ **67856**. Null deviance: Modell mit Intercept, d.h. degrees of freedom +1

Residual deviance: Modell mit 5 Parametern, d.h. degrees of freedom +5

(ii) *[5 Punkte]*

i. Modellgleichung:

$$\log(\mu) = \beta_0 + \beta_F FW + \beta_{B,Reg} Reg_B + \beta_{C,Reg} Reg_C + \beta_{D,Reg} Reg_D$$

(*2P*)

Abkürzungen: FW (*Fahrzeugwert*), Reg (*Region*). μ: erwartete Anzahl Schäden

ii. Dummykodierung. Dabei sind Reg_B, Reg_C, Reg_D entsprechende Dummy-Variablen, die 1 sind, wenn der Versicherungsnehmer die entsprechende Region als Merkmalsausprägung aufweist (und 0 sonst, d.h. Reg_A ist die Referenzregion). (*1P*)

Die Kodierungsvektoren sind $(0, 0, 0)$ für Reg_A und $(0, 0, 1)$ für Reg_D. (*2P*)

(iii) *[2 Punkte]* Fehlender Wert der Ausgabe ('?'):

$$'?' = -0.12594/0.05249 = -2.399$$

(iv) *[4 Punkte]* Da *Fahrzeugwert* metrisch ist, läßt sich hier der übliche Test basierend auf der Normalverteilung (zum Niveau $\alpha = 0.05$) verwenden (1P).

$$H_0 : \beta_F = 0 \quad \text{versus} \quad H_1 : \beta_F \neq 0$$

(1P) H_0 wird abgelehnt (d.h. *Fahrzeugwert* hat einen signifikanten Einfluss), wenn

$$|z| = |\hat{\beta}_F / \hat{\sigma}_{\beta_F}| > z_{0.975} \ ,$$

d.h., wenn der z-Wert betragsmäßig größer ist als das 97.5%-Quantil der Standardnormalverteilung (*1P*).

D.h., die Annahme ist, dass z asymptotisch (Standard-)normal verteilt ist (folgt aus Eigenschaft der Maximum-Likelihood-Schätzung) (*1P*).

(v) *[4 Punkte] Region* ist kategorial mit 4 Kategorien, daher lauten die Hypothesen:

$$H_0 : \beta_{B,Region} = \beta_{C,Region} = \beta_{D,Region} = 0$$

versus

$$H_1 : \text{mindestens ein Koeffizient } \beta_{j,Region} \text{ von Region ist ungleich 0}$$

(*2P*) Nein, der R Output genügt nicht, da hier nur die einzelnen Koeffizienten zu den Dummyvariablen getestet werden können, nicht aber alle simultan. *(2P)*

(vi) *[6 Punkte]* 1. Spalte: Kovariablen

2. Spalte: χ^2-Statistik des Likelihoodquotienten-Test, wenn jeweils die eine Kovariable weggelassen wird (*1P*)

3. Spalte: Freiheitsgrade der χ^2-Statistik (*1P*)

4. Spalte: p-Wert (*1P*).

Spalte LR `Chisq`: Teststatistik des Likelihood-Quotienten-Tests (LQT), asymptotisch χ^2-verteilt mit $df = 3$ Freiheitsgraden (*1P*)

p-Wert: Wahrscheinlichkeit, unter H_0 eine entsprechende χ^2-Teststatistik zu beobachten, die größer ist als die beobachtete (z.B. 20.315) (*1P*)

H_0 ablehnen, wenn p-Wert kleiner als das Signifikanzniveau α ist (*1P*)

(vii) *[2 Punkte]* Vorhergesagter Wert für *Anzahl.Schäden*: 0.078

Berechnung des linearer Prädiktors:

$$-2.71669 + 3 * 0.04999 + 0.02008 = -2.54664$$

(*1P*)

Anwendung der Response-Funktion:

$$\exp(-2.54664) = \mathbf{0.078}$$

(*1P*)

(c) *[6 Punkte]*

(i) *[2 Punkte]* Vergleich der AIC Werte des Modells in (b) mit dem neuen Modell:

AIC in (b): 36176

AIC mit neuer Kovariable: 34881

(*1P*)

'Kleiner ist besser', d.h. neue Variable sollte im Modell aufgenommen werden. (*1P*)

(ii) *[4 Punkte]* Berechnen der Statistik des Likelihoodquotienten-Tests:

$$AIC = -2L + 2p \quad \text{d.h.} \quad -2L = AIC - 2p$$

mit L log-Likelihood, p: Anzahl Parameter (*1P*)

$$\begin{aligned} \chi^2 &= -2(L_1 - L_2) = -2L_1 + 2L_2 \\ &= AIC_1 - 2p_1 - AIC_2 + 2p_2 \\ &= 36176 - 10 - 34881 + 12 \\ &= \mathbf{1297} \end{aligned}$$

Dabei ist L_1 die log Likelihood des Modells *ohne* die neue Variable (entsprechend AIC_1) und L_2 das Modell mit neuer Variable (entsprechend AIC_2). Weiterhin ist $p_1 = 5$, $p_2 = 6$. (*3P*)

Aufgabe 4. Zeitreihenanalyse *[24 Punkte]*

(a) *[3 Punkte]* Nennen Sie drei Annahmen für schwach stationäre Prozesse.

(b) *[6 Punkte]* Bei den folgenden Multiple Choice-Fragen entscheiden Sie für jede Aussage, ob diese allgemein richtig oder falsch ist. Eine falsche Antwort ergibt 0 Punkte, und für jede richtige Antwort gibt es 2 Punkte; Minuspunkte werden nicht vergeben. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben. Eine Begründung ist nicht erforderlich.

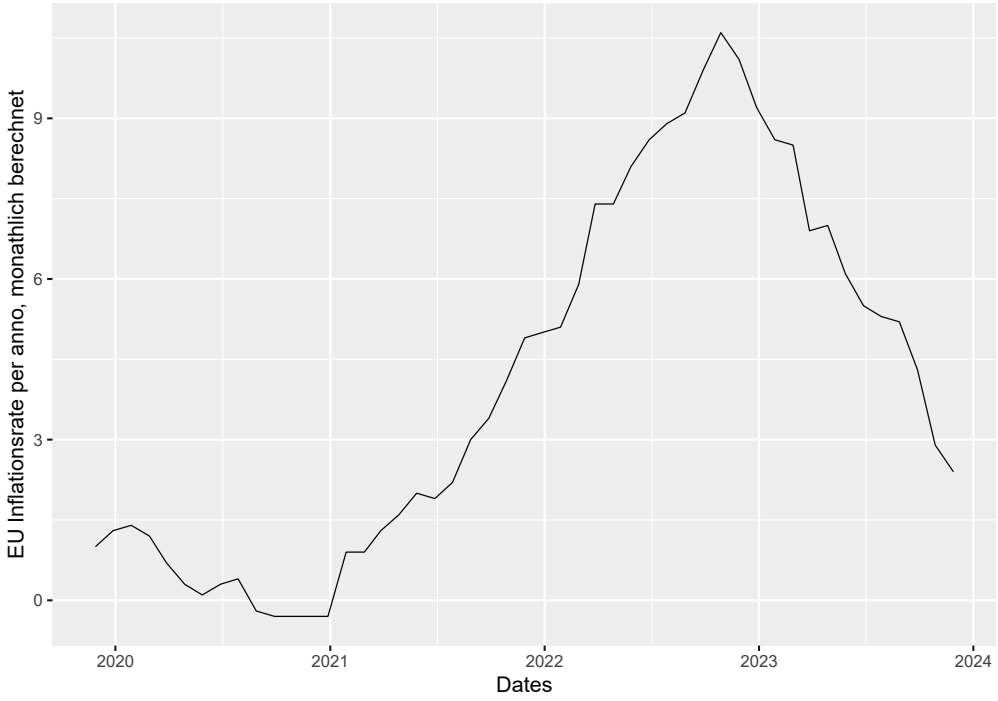
(i) Bei AR(p) Modellen kann man zukünftige Werte schätzen, nicht aber deren Unsicherheit.

(ii) ARCH Modelle werden verwendet um die bedingte Varianz zu modellieren, wohingegen ARMA Modelle geeignet sind den bedingten Erwartungswert zu modellieren.

(iii) Möchte man eine Zeitreihe um den linearen Trend bereinigen, kann man dies mittels Differenzen zweiter Ordnung erreichen.

(c) *[15 Punkte]*

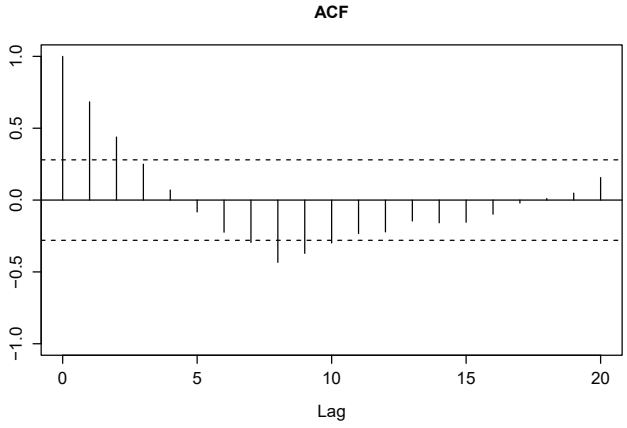
(i) *[3 Punkte]* Die folgende Grafik enthält die Inflationsrate in der EU (in Prozent) per anno, monatlich berechnet. Nennen Sie die drei Annahmen für Zeitreihen (-modelle) und geben Sie kurz an, ob diese Annahmen in diesem Fall erfüllt sind (Abschätzung anhand der Grafik).

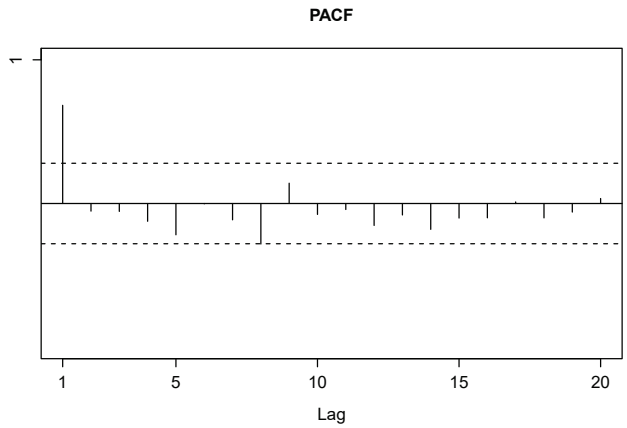


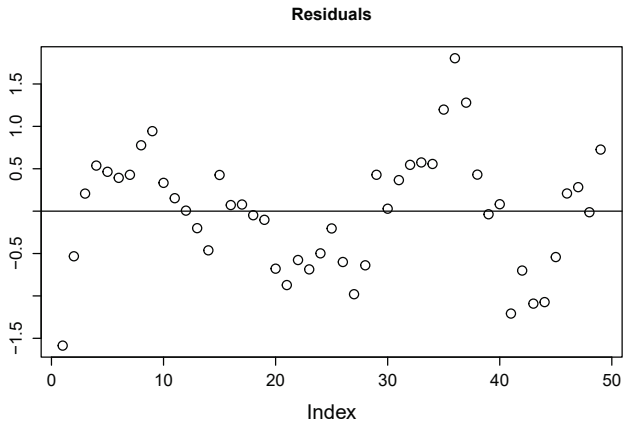
(ii) *[1 Punkt]* Nennen Sie einen Grund warum die Zeitreihe nicht stationär ist.

(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). Bitte beachten Sie, dass die PACF erst ab *lag=1* definiert ist.







(v) *[2 Punkte]* Sie sollen für die bereinigte Zeitreihe ein Modell schätzen. Ma-

chen Sie einen Vorschlag und begründen ihn kurz!

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

Lösung

(a) *[3 Punkte]*

(i) Konstanter Ewartungswert.

(ii) Konstante Varianz.

(iii) Die Autokorrelationsfunktion hängt nur von dem Lag ab, nicht aber von den jeweiligen Zeitpunkten.

(b) *[6 Punkte]*

(i) falsch

(ii) richtig

(iii) falsch

(c) (i) *[3 Punkte]*

i. Die Werte sind zu äquidistanten - gleichabständigen Zeitpunkten vorhanden.

ii. Die Werte sind Realisierungen einer metrisch skalierten Variable.

iii. Die Werte sind korreliert (abhängig würde auch akzeptiert werden).

Die Bedingungen sind erfüllt.

(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 fallen bis *Lag=8* auf ca -0.5, danach steigen sie wieder.

ii. Partielle Autokorrelationsfunktion: Abgesehen von *Lag=1* streuen die Werte nahe dem Wert Null.

iii. Residuen-Plot: Die Residuen streuen um den Wert Null; sie weisen einen zyklisch-artigen Verlauf auf

(v) *[2 Punkte]* Die Werte der partiellen Autokorrelationsfunktion suggerieren ein Modell mit autoregressiver Komponente (z.B. AR(1), und die Werte und Struktur der Autokorrelationen zumindest bis *Lag=8* legen zudem nahe eine moving average Komponente mitzuverwenden. Dies wäre dann ein ARMA Modell. Ein SARIMA Modell wird auch akzeptiert, wenn jemand argumentiert, dass die Residuen zyklische Schwankungen haben.

(vi) *[2 Punkte]* Aufgrund der Struktur in der ACF, PACF und Residuen sollte die Verwendung eines Zeitreihenmodelles hilfreich sein für die Modellierung des bedingten Erwartungswertes bzw. der Abhängigkeitsstruktur und daher die kurzfristige Prognose verbessern.

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

(a) *[4 Punkte]* Bei den folgenden Multiple Choice-Fragen entscheiden Sie für jede Aussage, ob diese allgemein richtig oder falsch ist. Eine falsche Antwort ergibt 0 Punkte, und für jede richtige Antwort gibt es 2 Punkte; Minuspunkte werden nicht vergeben. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben. Eine Begründung ist nicht erforderlich.

(i) Im Bayes'schen Credibility-Modell ermittelt man die a-posteriori-Verteilung der Schadenhöhen bei gegebenen Beobachtungen des Strukturparameters.

(ii) Für die Berechnung der linearisierten Credibility-Prämie $H^{**} = z_n \cdot \bar{X} + (1 - z_n) \cdot E(X)$ verwendet man einen Credibility-Faktor z_n der gegen 1 strebt, wenn die Beobachtungsanzahl n unendlich groß wird.

(b) *[6 Punkte]* An 6 Objekten wurden in den vergangenen 5 Jahren folgende Schadenbeobachtungen gemacht:

	Jahr 1	Jahr 2	Jahr 3	Jahr 4	Jahr 5	Mittelwert pro Zeile	Varianz pro Zeile
Objekt 1	0,00	0,00	0,00	0,07	0,00	0,0140	0,0010
Objekt 2	0,00	0,77	0,66	0,61	0,00	0,4080	0,1421
Objekt 3	0,38	0,00	0,02	0,49	0,00	0,1780	0,0566
Objekt 4	0,54	0,64	0,00	0,36	0,04	0,3160	0,0833
Objekt 5	0,00	0,93	0,49	0,15	0,00	0,3140	0,1586
Objekt 6	0,45	0,00	0,00	0,00	0,00	0,0900	0,0405
				Mittelwert pro Spalte		0,2200	0,0803
				Varianz pro Spalte		0,0229	

Berechnen Sie unter den Annahmen des Bühlmann-Modells die Credibility-Prämie von Objekt 1.

(c) *[20 Punkte]* Die Schadenhöhen X in einem Bayes'schen Credibility-Modell folgen einer Verteilung mit Dichte

$$f_{X|\Lambda=\lambda}(x) = (\lambda + 1) \cdot \exp\{-(\lambda + 1) \cdot x\}$$

für $x > 0$. Der Parameter λ sei dabei die Realisierung eines zufälligen Strukturparameters Λ, welcher einer Verteilung mit Dichte

$$f_\Lambda(\lambda) = (\lambda + 1)^{-2}$$

für $\lambda > 0$ folgt. Folgende zehn Beobachtungen x_i von X liegen vor:

i	1	2	3	4	5	6	7	8	9	10
x_i	0,1	1,5	0,4	0,2	1	0,5	0,3	0,1	0,7	0,2

Berechnen Sie den Wert der zugehörigen allgemeinen Credibility-Prämie. Gehen Sie dazu in folgenden Teilschritten vor:

(i) *[2 Punkte]* Geben Sie die Funktion $H(\lambda) = E(X|\Lambda = \lambda)$ an. Dabei können Sie verwenden, dass $f_{X|\Lambda=\lambda}$ die Dichte einer Exponentialverteilung mit Parameter $\lambda + 1$ ist.

(ii) *[5 Punkte]* Berechnen Sie die gemeinsame Dichte g des Strukturparameters und der Schadenbeobachtungen, und setzen Sie in diese die in der Tabelle gegebenen Werte $x_1, \dots, x_{10}$ ein.

Kontrollergebnis: $g(\lambda, x_1, \dots, x_{10}) = (\lambda + 1)^8 \cdot \exp\{-5 \cdot (\lambda + 1)\}$.

(iii) *[4 Punkte]* Rechnen Sie nach, dass die Dichte der a-posteriori-Verteilung des Strukturparameters unter den gegebenen Schadenbeobachtungen

$$f_{\Lambda|X_1=x_1,\dots,X_{10}=x_{10}}(\lambda) = \frac{5^9}{8! \cdot (1 - F_{(9;5)}(1))} \cdot (\lambda + 1)^8 \cdot \exp\{-5 \cdot (\lambda + 1)\}$$

ist.

Lösungshinweis: Ohne Beweis können Sie verwenden, dass

$$\int_0^\infty (\lambda + 1)^m \cdot \exp\{-a \cdot (\lambda + 1)\}\, d\lambda = \frac{m!}{a^{m+1}} \left(1 - F_{(1+m;\, a)}(1)\right)$$

mit $F_{(1+m;a)}$ *= Verteilungsfunktion der Gamma-Verteilung* $\Gamma(1 + m; a)$.

(iv) *[9 Punkte]* Berechnen Sie den Wert der allgemeinen Credibility-Prämie H^*.

Lösungshinweis: Nutzen Sie nochmals das Integral aus dem Lösungshinweis zu (iii). Ein Statistikprogramm oder Excel liefert für die Gammaverteilung $F_{(8;\,5)}(1) = 0{,}1334$ *und* $F_{(9;\,5)}(1) = 0{,}0681$.

Lösung:

(a)

(i) falsch: Man ermittelt die a-posteriori-Verteilung des Strukturparameters bei gegebenen Schadenbeobachtungen.

(ii) richtig

(b) Der Credibility-Faktor im Bühlmann-Modell berechnet sich gemäß

$$\hat{z} = 1 - \frac{\hat{E}\left(\sigma^2(\Theta)\right)}{n \cdot \widehat{Var}(\bar{X})} = 1 - \frac{0{,}0803}{5 \cdot 0{,}0229} = 0{,}2987.$$

Für die Credibility-Prämie des ersten Objekts gilt

$$H_1^{**} = \hat{z} \cdot \overline{X_1} + (1 - \hat{z}) \cdot \hat{E}\left(\mu(\Theta)\right)$$
$$= 0{,}2987 \cdot 0{,}0140 + (1 - 0{,}2987) \cdot 0{,}2200 = 0{,}1585.$$

(c)

(i) Da die Schadenhöhen exponentialverteilt sind, gilt $H(\lambda) = E(X|\Lambda = \lambda) = \frac{1}{\lambda+1}$.

(ii) Für die gemeinsame Dichte gilt

$$g(\lambda, x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X|\Lambda=\lambda}(x_i) \cdot f_\Lambda(\lambda) = (\lambda+1)^n \prod_{i=1}^{n} \exp\{-(\lambda+1) \cdot x_i\} \cdot (\lambda+1)^{-2}$$
$$= (\lambda+1)^{n-2} \cdot \exp\left\{-(\lambda+1) \cdot \sum_{i=1}^{n} x_i\right\}$$

Im vorliegenden Fall ist $n = 10$ und $\sum_{i=1}^{n} x_i = 5$, so dass

$$g(\lambda, x_1, \dots, x_{10}) = (\lambda+1)^8 \cdot \exp\{-5 \cdot (\lambda+1)\}.$$

(iii) Die a-posteriori-Dichte ist

$$f_{\Lambda|X_1=x_1,\dots,X_{10}=x_{10}}(\lambda) = \frac{1}{c} \cdot\ g(\lambda, x_1, \dots, x_{10})$$

mit Normierungskonstante (gemäß Hinweis)

$$c = \int_0^\infty g(\lambda, x_1, \dots, x_{10})\, d\lambda = \int_0^\infty (\lambda+1)^8 \cdot \exp\{-5 \cdot (\lambda+1)\}\, d\lambda = \frac{8!}{5^9}\left(1 - F_{(9;\,5)}(1)\right).$$

Somit ergibt sich

$$f_{\Lambda|X_1=x_1,\dots,X_{10}=x_{10}}(\lambda) = \frac{5^9}{8! \cdot (1 - F_{(9;5)}(1))} \cdot (\lambda+1)^8 \cdot \exp\{-5 \cdot (\lambda+1)\}$$

(iv) Für die allgemeine Credibility-Prämie ergibt sich mit dem Hinweis

$$\begin{aligned} H^* &= \int_0^\infty H(\lambda) \cdot f_{\Lambda|X_1=x_1,\dots,X_{10}=x_{10}}(\lambda)\; d\lambda \\ &= \frac{5^9}{8! \cdot \left(1 - F_{(9;5)}(1)\right)} \cdot \int_0^\infty \frac{1}{\lambda+1} \cdot (\lambda+1)^8 \cdot \exp\{-5 \cdot (\lambda+1)\}\; d\lambda \\ &= \frac{5^9}{8! \cdot \left(1 - F_{(9;5)}(1)\right)} \cdot \int_0^\infty (\lambda+1)^7 \cdot \exp\{-5 \cdot (\lambda+1)\}\; d\lambda \\ &= \frac{5^9}{8! \cdot \left(1 - F_{(9;5)}(1)\right)} \cdot \frac{7! \cdot \left(1 - F_{(8;\,5)}(1)\right)}{5^8} \\ &= \frac{5 \cdot \left(1 - F_{(8;\,5)}(1)\right)}{8 \cdot \left(1 - F_{(9;5)}(1)\right)} = \frac{5 \cdot (1 - 0{,}1334)}{8 \cdot (1 - 0{,}0681)} = 0{,}5812. \end{aligned}$$

Aufgabe 6. *[Monte-Carlo Methoden] [30 Punkte]*

(a) *[24 Punkte]* Gegeben sei eine Zufallsvariable X mit Dichte

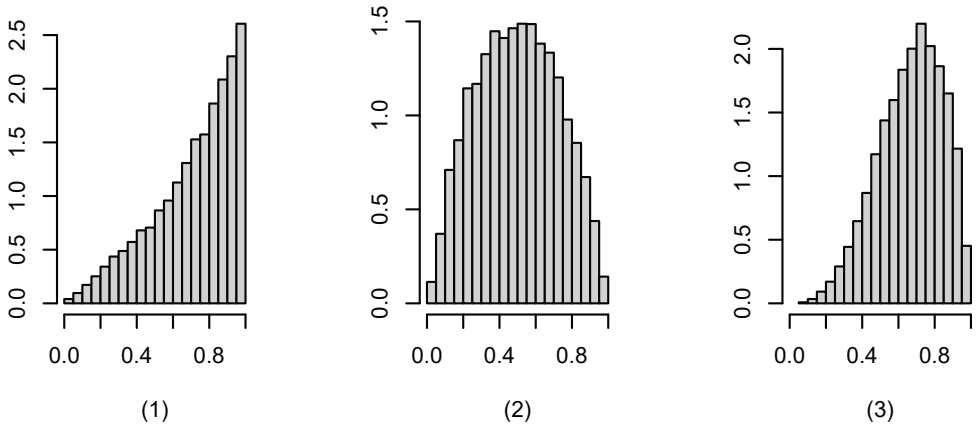
$$f(x) = \frac{4}{3}x(x^2+1) \cdot \mathbb{1}_{(0,1)}(x).$$

(i) *[10 Punkte]* Zeigen Sie, dass folgender Algorithmus Zufallszahlen aus f erzeugt:

- Erzeuge $u \sim \mathrm{U}(0, 1)$
- Berechne

$$x = \sqrt{\sqrt{3u+1}-1}$$

(ii) *[2 Punkte]* Welches der drei folgenden Histogramme (1)–(3) zeigt das Ergebnis eines Durchlaufs des Algorithmus aus (i) (je 10.000 Durchläufe)? Begründen Sie Ihre Auswahl.



(iii) *[2 Punkte]* Es soll $E(X^2)$ mit Hilfe einer Monte-Carlo Simulation geschätzt werden. Auf Basis des Algorithmus aus (i) wurden dafür N Zufallszahlen $x_1, \ldots, x_N$ generiert. Wie wird daraus der Monte-Carlo Schätzer S_N für $E(X^2)$ berechnet? Geben Sie eine Formel an (ohne Begründung).

(iv) *[10 Punkte]* In Fortführung von Teil (iii) sei nun $n = 2.000.000$ und $S_n = 0,7108215$. Zudem seien die Werte $Var(X^2) = 0,0802$ und $\Phi^{-1}(0,975) = 1,959964$ bekannt. Bestimmen Sie daraus ein Konfidenzintervall der Schätzung. Was kann man damit über die Anzahl der korrekten Nachkommastellen von S_n sagen?

(b) *[6 Punkte]* Bei den folgenden Multiple-Choice-Fragen entscheiden Sie für jede Aussage, ob diese allgemein richtig oder falsch ist. Eine falsche Antwort ergibt

0 Punkte, und für jede richtige Antwort gibt es 2 Punkte; Minuspunkte werden nicht vergeben. Schreiben Sie Ihre Antworten auf die Lösungsblätter, die Sie abgeben. Eine Begründung ist nicht erforderlich.

(i) Bei der Box-Muller-Methode werden aus zwei unabhängigen gleichverteilten Zufallszahlen zwei perfekt korrelierte standardnormalverteilte Zufallszahlen erzeugt.

(ii) Verdoppelt man die Anzahl der Simulationen zur Monte Carlo Schätzung einer stochastischen Kenngröße, dann halbiert sich die Länge des zugehörigen Konfidenzintervalls.

(iii) Bei der Simulation eines Pfades einer geometrischen Brownschen Bewegung mit Hilfe der Euler-Methode entsteht kein Diskretisierungsfehler.

Lösungsvorschlag

(a)

(i) Es handelt sich um eine Anwendung der Inversionsmethode: Die Verteilungsfunktion lautet

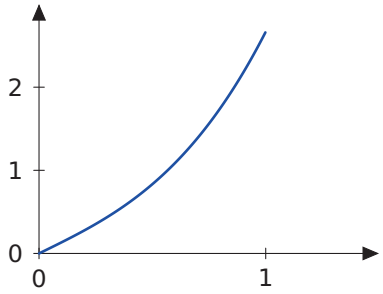
$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \left[\frac{1}{3}t^4 + \frac{2}{3}t^2\right]_0^x = \frac{1}{3}x^4 + \frac{2}{3}x^2$$

für $0 < x < 1$. Damit ist die Gleichung $F(x) = u$ nach x aufzulösen:

$$\begin{aligned}
& \tfrac{1}{3}x^4 + \tfrac{2}{3}x^2 &&= u \\
\Leftrightarrow \quad & x^4 + 2x^2 &&= 3u \\
\overset{t:=x^2}{\Leftrightarrow} \quad & t^2 + 2t &&= 3u \\
\Leftrightarrow \quad & t &&= \sqrt{3u+1} - 1 \\
\Leftrightarrow \quad & x &&= \sqrt{\sqrt{3u+1} - 1} = F^{-1}(u).
\end{aligned}$$

Im letzten Schritt wird die positive (äußere) Wurzel genommen, da x positiv ist.

(ii) Das Histogramm ist eine diskretisierte Variante des Graphen der Dichtefunktion. Auch wenn man keine Möglichkeit hat den Plot von f zu erstellen, sieht man, dass die Histogramme in (2) und (3) von Dichten stammen, die im Punkt $x = 1$ sehr kleine Werte haben. Dagegen hat f in $x = 1$ den Wert 8/3. Daher kommt nur das Histogramm (1) in Frage.
Zum Vergleich ein Plot von f über $[0, 1]$:



(iii) Der Schätzer lautet

$$S_n = \frac{1}{n}\sum_{k=1}^{n} x_k^2.$$

(iv) Das Konfidenzintervall zum Niveau α des Monte-Carlo Schätzers S_N des Erwartungswertes einer Zufallsvariablen Y lautet

$$\left[S_n - \frac{\sigma(Y)}{\sqrt{n}}\,\Phi^{-1}\left(\frac{\alpha+1}{2}\right),\, S_n + \frac{\sigma(Y)}{\sqrt{n}}\,\Phi^{-1}\left(\frac{\alpha+1}{2}\right)\right].$$

Da $\Phi^{-1}(0,975)$ gegeben ist, kann man mit den gegebenen Werten das 95%-Konfidenzintervall

$$[0,710429;\ 0,711214]$$

für den Erwartungswert von $Y = X^2$ berechnen. Das bedeutet: Die Wahrscheinlichkeit, dass das auf Basis der Zufallszahlen erzeugte Intervall den wahren Wert $E(X^2)$ enthält, beträgt 95%. Mit einer entsprechenden Irrtumswahrscheinlichkeit von 5% hat S_n daher zwei korrekte Nachkommastellen (0,71).

(b)

(i) Falsch, es sind unabhängige standardnormalverteilte Zufallszahlen.

(ii) Falsch, es reduziert sich auf $1/\sqrt{2}$ mal die ursprüngliche Länge.

(iii) Falsch, die Euler-Methode resultiert immer in normalverteilten Simulationen, aber die geometrische Brownsche Bewegung hat eine Log-Normalverteilung.