Aufgabe 1 (Biometrische Rechnungsgrundlagen) [30 Punkte]

Zwei Lebensversicherungsunternehmen A und B fusionieren. Im Rahmen der Fusion werden die Bestände der gemischten Versicherungen zu einer Risikogemeinschaft zusammengeführt, da sie (angeblich) gleiche Risiken darstellen. Als Aktuar werden Sie um Antworten zu folgenden Fragen gebeten:

a) [22 Punkte] Sind die biometrischen Risiken in beiden Beständen gleich?

Als Unterlagen haben Sie für beide Bestände und die letzten 5 Jahre (2013 - 2017) folgende Informationen zur Verfügung:

1. Versichertenbestand mit individuellen Angaben (Geschlecht, Geb.Dat., Datum des Versicherungsbeginns, Datum des Ablaufs der Versicherung, Vers.Summe, Deckungskapital) zum Bilanzstichtag (31.12.),
2. Liste der Abgänge im jeweiligen Geschäftsjahr mit den Angaben unter 1. und zusätzlich nur den Abgangsgrund Tod mit Todesdatum oder Ablauf der Versicherung (Storno wird vernachlässigt).

(i) [7 Punkte] Wählen und beschreiben Sie die aus Ihrer Sicht optimale Methode, mit der Sie die Unterschiede bei der Sterblichkeit der beiden Bestände analysieren, und begründen Sie Ihre Wahl. Verwenden Sie für das Alter das bürgerliche Alter (vollendete Jahre am letzten Geburtstag).

(ii) [7 Punkte] Welche Methoden kennen Sie, falls Ihnen die Informationen Ablauf der Versicherung und Todesdatum nicht zur Verfügung stehen. Stellen Sie eine dieser Methoden dar und vergleichen Sie es mit Ihrem Verfahren unter (i).

(iii) [8 Punkte] Welche Methoden kennen Sie, um Unterschiede zwischen den beiden Beständen zu erkennen. Wenden Sie eine dieser Methoden auf das folgende Beispiel an.

Alter x	q_x^A	q_x^B
41	1,0%	1,1%
42	1,2%	1,0%
43	0,7%	0,6%
44	1,1%	1,0%
45	1,4%	1,5%

b) [8 Punkte] Welche Einsparungen ergeben sich beim Sicherheitskapital?

Nehmen Sie an, dass die Schadenverteilungen beider Bestände stochastisch unabhängig sind und das Schwankungsrisiko in beiden Beständen durch ein entsprechendes Sicherheitskapital zum Sicherheitsniveau $\alpha=99\%$ berücksichtigt, das durch das 99%-Quantil der Schadenverteilung bestimmt wird.

Die 99%-Sicherheitskapitale zur Abdeckung des Schwankungsrisikos sind in beiden VU vorhanden und betragen für VU A: K^A = 2 Mio€ und für VU B: K^B = 5 Mio€.

Nehmen Sie weiter an, dass das Sicherheitskapital in beiden Beständen mit Hilfe der Normalverteilung errechnet wurde. Welches Sicherheitskapital ergibt sich dann nach dem Zusammenlegen der Bestände? Das 99%-Quantil der Standardnormalverteilung beträgt $u_{0,99}$ = 2,326.

Lösung

a)

(i) (7 Punkte)

Ermittlung der relativen Sterbehäufigkeiten aus den Beständen der letzten 5 Jahre. Da es sich um einen offenen Personenbestand handelt, ist nur die Verweildauermethode geeignet, um die Sterbehäufigkeiten exakt zu ermitteln. Die Verweildauermethode ist anwendbar, da in den Bestandsdaten der Todeszeitpunkt und das Ablaufdatum enthalten ist.

Bezeichne $d_{x,i}$ $(0 \le d_{x,i} \le 1)$ die Verweildauer der Person i im Alter x in der Personengesamtheit mit

$d_{x,i} < 1$ falls die Versicherung abläuft oder bei unterjährigem Versicherungsbeginn

$d_{x,i} = 1$ falls der Tod im Alter x eintritt oder die Versicherung im ganzen Jahr unter Risiko stand.

sowie

B	Beobachtungszeitraum [1.1.2013 , 1.1.2018)
G	Geburtsjahr
$L_x(B)$	die Personen aus der Personengesamtheit, die im Beobachtungszeitraum B das x. Lebensjahr vollenden (könnten),
$T_x(B)$	die Personen aus $L_x(B)$, die im Beobachtungszeitraum sterben.

$L_x^{'}(B) = L_x(B) \setminus T_x(B)$ Überlebende (Teilgesamtheit der Lebenden)

Die rohen Sterbewahrscheinlichkeiten werden definiert durch

$$q_x^v = \frac{\left|T_x(B)\right|}{\sum\limits_{i \in L_x(B)} d_{x,i}} = \frac{\left|T_x(B)\right|}{\left|T_x(B)\right| + \sum\limits_{i \in L_x^{'}(B)} d_{x,i}} .$$

(ii) (7 Punkte)

Methoden:

1. Geburtsjahrmethode
2. Sterbejahrmethode
3. Sterbeziffernverfahren

Darstellung und Begründung einer Methode (nur 1 Methode wird verlangt):

Geburtsjahrmethode

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

L	die unter Risiko stehende Personengesamtheit,
T	Tote aus L
$q = \frac{\lvert T\rvert}{\lvert L\rvert}$	relative Sterbehäufigkeit = rohe Sterbewahrscheinlichkeit
$L' = L \setminus T$	Überlebende (Teilgesamtheit der Lebenden)
B	Beobachtungszeitraum [1.1.2013 , 1.1.2018)
G	Geburtsjahr

$L_x(B, G)$ die Personen aus der Personengesamtheit, die im Beobachtungszeitraum B das x. Lebensjahr vollenden (könnten) und im Geburtsjahr G geboren wurden,

$T_x(B, G)$ die Personen aus $L_x(B, G)$, die im Beobachtungszeitraum sterben.

Die rohe Sterbewahrscheinlichkeit nach der Geburtsjahrmethode ist definiert durch:

$$q_x^G = \frac{\left|T_x\left(B,G\right)\right|}{\left|L_x\left(B,G\right)\right|}$$

Sterbejahrmethode

Gegenüber der Geburtsjahrmethode werden bei der Sterbejahrmethode sämtliche Todesfälle eines Alters x berücksichtigt. Die Sterbewahrscheinlichkeit kann jedoch nur näherungsweise und nicht exakt ermittelt werden.

Sei $B = \bigcup_{i=1}^n B_i$ der Beobachtungszeitraum und $G = \bigcup_{i=0}^n G_i$ die für das Alter x in Frage kommenden Geburtsjahrgänge. Deshalb wird die Annahme getroffen, dass von den „Rand“-Geburtsjahrgängen G_0 und G_n nur die Hälfte aller Todesfälle in B beobachtet werden. Die rohen Sterbewahrscheinlichkeiten werden daher ermittelt als

$$q_x^S = \frac{\left|T_x\left(B,G\right)\right|}{\frac{1}{2}\left|L_x\left(B,G_0\right)\right| + \left|\sum_{i=1}^{n-1} L_x\left(B,G_i\right)\right| + \frac{1}{2}\left|L_x\left(B,G_n\right)\right|}$$

Sterbeziffernverfahren

Im Gegensatz zur Geburtsjahr- und zur Sterbejahrmethode werden beim Sterbeziffernverfahren die Bestandsveränderungen innerhalb des Beobachtungszeitraums näherungsweise durch die Durchschnittsbildung der Personengesamtheit am Anfang und am Ende jedes Jahres der Beobachtungsperiode berücksichtigt. Zu- und Abgänge, die innerhalb eines Beobachtungsjahres stattfinden, können nicht miteinbezogen werden.

Sei $B = \bigcup_{i=1}^{n} B_i$ ein Beobachtungszeitraum von n Jahren und seien G die entsprechenden Geburtsjahrgänge.

Die Sterbeziffer lautet

$$\mathrm{k_x} = \frac{\left|\mathrm{T_x(B,G)}\right|}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{1}{2}\left(\left|\mathrm{L_x^A(B_i,G)}\right| + \left|\mathrm{L_x^E(B_i,G)}\right|\right)},$$

wobei L_x^A und L_x^E den Bestand der x-Jährigen am Anfang bzw. am Ende eines Jahres innerhalb des Beobachtungszeitraums bezeichnen.

k_x kann jedoch nicht als Wert für die relative Sterbehäufigkeit angesetzt werden. Stattdessen definiert man für die relative Sterbehäufigkeit

$$\mathrm{q_x^z} = \frac{2\mathrm{k_x}}{2+\mathrm{k_x}} \quad .$$

(iii) (8 Punkte)
Die Unterschiede zwischen den beiden Beständen können durch den Vergleich der relativen Sterbehäufigkeiten mithilfe eines Testverfahrens ermittelt werden.

Dazu wird die Nullhypothese $\mathrm{H_0}$ bzw. die Alternativhypothese $\mathrm{H_1}$ formuliert:

$\mathbf{H_0}$: Die Sterbewahrscheinlichkeiten stimmen in beiden Versichertenbeständen überein.

$\mathbf{H_1}$: Die Sterbewahrscheinlichkeiten in beiden Versichertenbeständen sind verschieden.

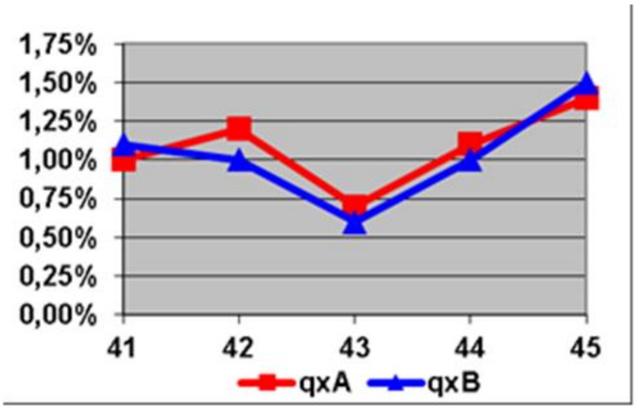
Mögliche Testverfahren:

- Vorzeichentest
- Iterationstest

(Es genügt, nur ein Testverfahren darzustellen)

Beispiel:

Alter x	q_x^A	q_x^B
41	1,0%	1,1%
42	1,2%	1,0%
43	0,7%	0,6%
44	1,1%	1,0%
45	1,4%	1,5%



Vorzeichentest

Geht man davon aus, dass in beiden Beständen die Sterbewahrscheinlichkeiten übereinstimmen, dann kann man annehmen, dass bei den Differenzen zwischen Sterbehäufigkeiten beider Bestände gleich viele positive wie negative Vorzeichen auftreten.

Teststatistik:

$$T = \sum_{x=41}^{45} 1_{\{q_x^A < q_x^B\}}$$

Für das Beispiel gilt: $T = 2$

D.h., man zählt die negativen Vorzeichen, die sich bei den Differenzen q_x^A - q_x^B ergeben.

Unter Gültigkeit der Nullhypothese H_0 ist T binomialverteilt mit Wahrscheinlichkeit ½ für jedes Vorzeichen, d.h. es gilt

$$T \sim B\left(5, \frac{1}{2}\right).$$

Zu einem vorgegebenen Signifikanzniveau α werden dann zwei Schwellenwerte n_α und $5 - n_\alpha$ bestimmt, so dass die Nullhypothese H_0 abgelehnt wird, wenn der Wert der Teststatistik n_α unterschreitet bzw. $5 - n_\alpha$ überschreitet.

Dabei wird n_α wegen der Symmetrie der Binomialverteilung bestimmt aus

$$2 \cdot \mathrm{P}\left(T < n_\alpha\right) = 2 \cdot \sum_{j=0}^{n_\alpha - 1} \binom{5}{j} \cdot \left(\frac{1}{2}\right)^{j} \cdot \left(\frac{1}{2}\right)^{5-j} \overset{!}{\leq} \alpha \, .$$

Iterationstest

Wenn man davon ausgeht, dass in beiden Beständen die Sterbewahrscheinlichkeiten übereinstimmen, dann kann man folgern, dass die beobachteten Wahrscheinlichkeiten des einen Bestandes „mal größer und mal kleiner“ als die des anderen sind, d.h. dass also viele Vorzeichenwechsel auftreten.

Teststatistik:

$$T = \sum_{x=42}^{45} 1_{\left\{Sign(q_x^A - q_x^B) \neq Sign(q_{x-1}^A - q_{x-1}^B)\right\}}$$

Für das Beispiel gilt: $T = 2$

Wie beim Vorzeichentest bildet man auch beim Iterationstest zunächst die Differenzen zwischen den Sterbehäufigkeiten desselben Alters. Im Anschluss zählt man die aufgetretenen Vorzeichenwechsel.

Unter der Gültigkeit der Nullhypothese H_0 ist T binomialverteilt mit Wahrscheinlichkeit ½ für jeden Vorzeichenwechsel, d.h. es gilt

$$T \sim B\left(4, \frac{1}{2}\right).$$

Zu vorgegebenem Signifikanzniveau α wird ein Schwellenwert $n_{\sim}$ so bestimmt, dass die Nullhypothese abgelehnt wird, wenn der Wert der Teststatistik $n_{\sim}$ unterschreitet.

Dabei wird $n_{\sim}$ bestimmt aus

$$\mathrm{P}(T < n_\alpha) = \sum_{j=0}^{n_\alpha - 1} \binom{4}{j} \cdot \left(\frac{1}{2}\right)^{j} \cdot \left(\frac{1}{2}\right)^{4-j} \overset{!}{\leq} \alpha \, .$$

b) (8 Punkte)

Bezeichne S^A den Schaden im Bestand A und S^B den im Bestand B.

Dann gilt näherungsweise $S^A \sim N(\mu_A, \sigma_A^2)$ und $S^B \sim N(\mu_B, \sigma_B^2)$

und für den Median gilt $\tilde{S}^A = «_A$ (analog für B).

Für den Value at Risk gilt dann mit Φ als Verteilungsfunktion der Standard-Normalverteilung

$$\begin{aligned} \mathrm{VaR}_{0,01}\left(S^A\right) &= \mu_A + \sigma_A \cdot \Phi^{-1}(1-0,01) \\ &= \mu_A + \sigma_A \cdot u_{0,99} \end{aligned}$$

und

$$K^A = \mathrm{VaR}_{0,01}(S^A) - \mu_A = \sigma_A \cdot u_{0,99} = 2 \text{ Mio€}$$

$$\Rightarrow \pm_A = \frac{2Mio€}{u_{0,99}}$$

sowie

$$\begin{aligned} \mathrm{VaR}_{0,01}\left(S^B\right) &= \mu_B + \sigma_B \cdot \Phi^{-1}(1-0,01) \\ &= \mu_B + \sigma_B \cdot u_{0,99} \end{aligned}$$

und

$$K^B = \mathrm{VaR}_{0,01}(S^B) - \mu_B = \sigma_B \cdot u_{0,99} = 5 \text{ Mio€}$$

$$\Rightarrow \pm_B = \frac{5Mio€}{u_{0,99}}$$

Da die Schäden in den beiden Versicherungsbeständen stochastisch unabhängig sind, gilt für die Summe S der zwei unabhängigen normalverteilten Zufallsvariablen S^A und S^B

$$S = S^A + S^B \sim N\left(\mu_A + \mu_B, \sigma_A^2 + \sigma_B^2\right)$$

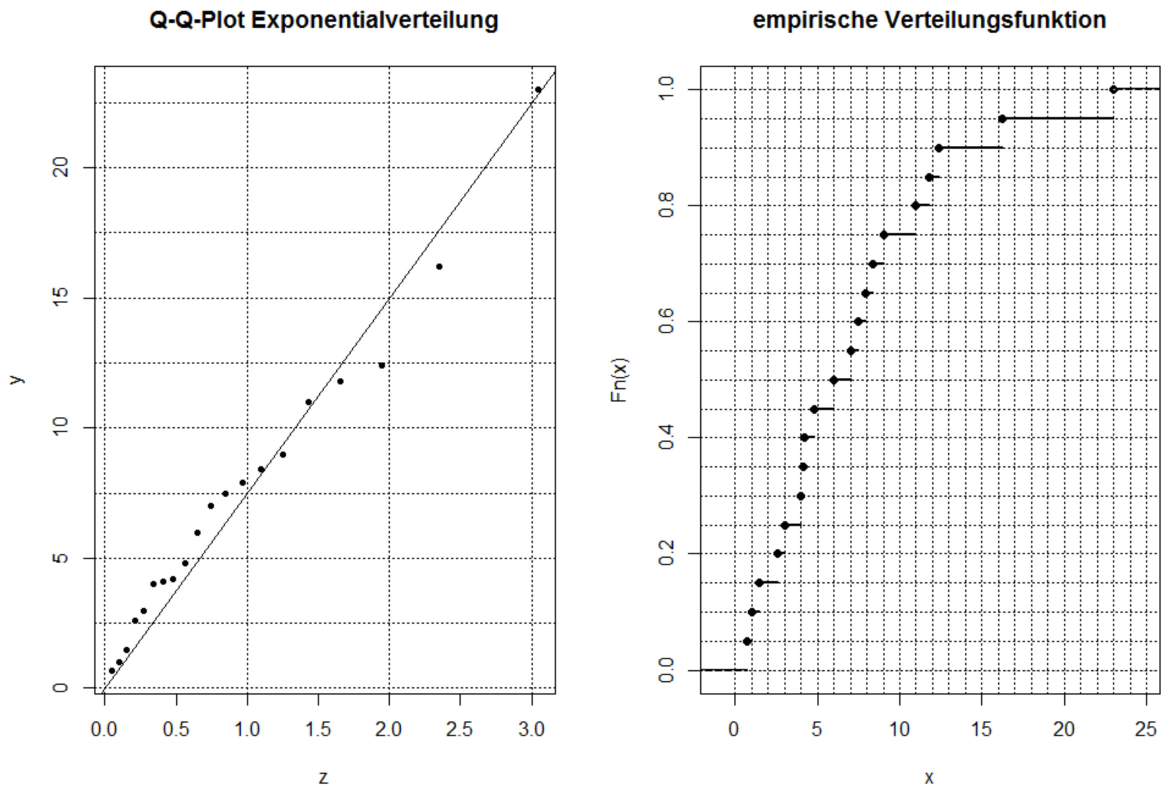
Für das erforderliche Sicherheitskapital K der fusionierten Bestände gilt dann:

$$
\begin{aligned}
K &= \mathrm{VaR}_{0,01}(S) - \mu_A - \mu_B \\
&= u_{0,99}\sqrt{\sigma_A^2 + \sigma_B^2} \\
&= u_{0,99}\frac{1}{u_{0,99}}\sqrt{2^2 + 5^2} \\
&= \sqrt{29}\,\text{Mio€} \\
&= 5,38\,\text{Mio€}
\end{aligned}
$$

Die Einsparung beim Sicherheitskapital beträgt 2+5-5,38 = 1,62 Mio€.

Aufgabe 2 (Deskriptive Statistik) [30 Punkte]

Der Aktuar eines Schadenversicherers untersucht, ob Schäden einer Exponentialverteilung Exp(λ) genügen. Es werden dazu Jahresgesamtschäden $x_1, \dots, x_n$ der Vergangenheit betrachtet. Es gilt $\sum_{i=1}^{n} x_i = 146{,}1$. In den folgenden Graphiken sind die empirische Verteilungsfunktion und ein geeigneter Q-Q-Plot mit der Ausgleichsgeraden gegeben.



a) *[7 Punkte]* Erstellen Sie ein Histogramm mit den Klassengrenzen 0, 5, 10, 25. Erläutern Sie genau ihr Vorgehen.
b) *[9 Punkte]* Erstellen Sie einen Boxplot zu den Daten. Verwenden Sie hierzu die empirische Verteilungsfunktion. Begründen Sie die gewählten bzw. die ausgerechneten Werte.
c) *[4 Punkte]* Erläutern Sie den Q-Q-Plot, insbesondere die auf der x-Achse und der y-Achse aufgetragenen Größen.
d) *[5 Punkte]* Warum stützt der Q-Q-Plot die Hypothese einer Exponentialverteilung? Welcher Schätzwert für λ ergibt sich daraus?

e) *[5 Punkte]* Für die Informationsmatrix $I(\lambda)$ von Exp(λ) gilt $I(\lambda) = \frac{1}{\lambda^2}$. Der ML-Schätzer von λ ist $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$ (beides müssen Sie **nicht** beweisen). Bestimmen Sie nun mit Hilfe der asymptotischen Normalverteilung von $\hat{\lambda}$ ein 95 % - Konfidenzintervall für λ aus den obigen Daten.

Folgende Quantile der Standardnormalverteilung sind gegeben:

$u_{90\%} = 1{,}28, u_{95\%} = 1{,}64, u_{97{,}5\%} = 1{,}96,\ u_{99\%} = 2{,}33$

Lösung Aufgabe 2

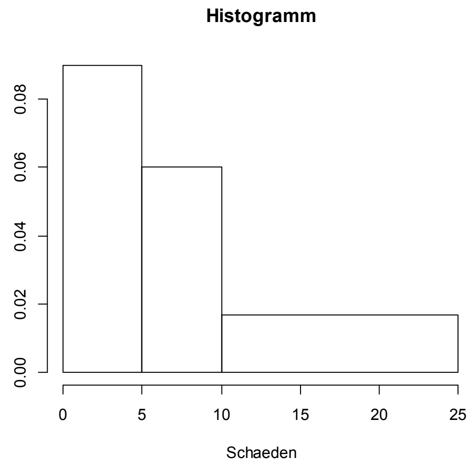
a) (7 Punkte)

Aus der empirischen Verteilungsfunktion erkennt man, dass 20 Daten vorliegen, 9 Schäden im Intervall $I_1 := [0,5]$, 6 Schäden im Intervall $I_2 := (5, 10]$ und 5 Schäden im Intervall $I_3 := (10,25]$. Für die Höhe der Rechtecke h_1, h_2 und h_3 folgt

$$5h_1 = \frac{9}{20} \Longrightarrow h_1 = 0{,}09$$

$$5h_2 = \frac{6}{20} \Longrightarrow h_2 = 0{,}06$$

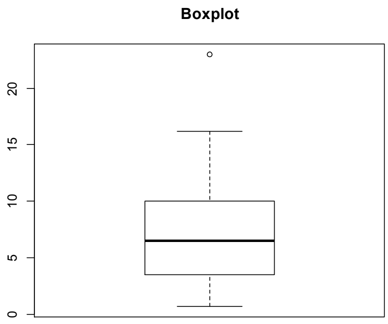
$$15h_3 = \frac{5}{20} \Longrightarrow h_3 = 0{,}017$$



b) (9 Punkte)

Die Sprungstellen der empirischen Verteilungsfunktion entsprechen der Ordnungsstatistik, man liest also ab $x_{(5)} = 3, x_{(6)} = 4, x_{(15)} = 9, x_{(16)} = 11$. Daraus ergeben sich die Quartile $x_{25\%} = 0{,}5(3+4) = 3{,}5, x_{75\%} = 0{,}5(9+11) = 10$.

Der Median ergibt sich aus $x_{50\%} = 0{,}5\left(x_{(10)} + x_{(11)}\right) = 0{,}5(6+7) = 6{,}5$. Der Interquartilsabstand ist 6,5. Die Box ist also im Bereich [3,5; 10], die Whisker bei 0 und 10+1,5*6,5=19,75 und der $x_{(20)} = 23$ als Ausreißer.



c) (4 Punkte)

Auf der x-Achse werden die Quantile der Exp(1) Verteilung aufgetragen, auf der y-Achse die Ordnungsstatistik. Für $u \in (0,1)$ gilt $u = 1 - e^{-x}$ erhält man durch Auflösen $x = -\ln(1-u)$ und somit sind im obigen Q-Q-Plot sind die Punkte $(-\ln(1-u_k), x_{(k)})$ k=1, …, 20 mit $u_k = \frac{k}{21}$ enthalten.

d) (5 Punkte)

Die Punkte des Q-Q-Plots liegen beim Vorliegen einer Exp((λ))-Verteilung auf einer Ursprungsgeraden durch den Nullpunkt mit Steigung $^1/_\lambda$, da es sich um eine reine Skalenfamilie handelt.

Da die Punkte nahezu in der Nähe der Ausgleichsgeraden liegen, spricht das für die Exponentialverteilung.

Die Ausgleichsgerade besitzt die Steigung 7,5 und somit wäre $^1/_{7,5}$ der sich aus dem Q-Q-Plot ergebende Wert.

e) (5 Punkte)

Der Schätzwert ist $\hat{\lambda} = \frac{20}{146,1} = 0,137$. Das Konfidenzintervall ergibt sich aus $\left(\hat{\lambda} - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{nI(\lambda)}}, \hat{\lambda} + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{nI(\lambda)}}\right) = \left(\hat{\lambda} - \frac{\lambda u_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \hat{\lambda} + \frac{\lambda u_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$ und somit mit $n = 20, u_{97,5\%} = 1,96$ und durch Einsetzen von $\hat{\lambda}$ anstelle von λ

$$\left(\hat{\lambda}\left(1 - \frac{1,96}{\sqrt{20}}\right), \hat{\lambda}\left(1 + \frac{1,96}{\sqrt{20}}\right)\right) = (0,077, \qquad 0,197)$$

Aufgabe 3. *[30 Punkte]* Induktive Statistik

Eine ausgewählte Teilmenge von $n = 10000$ Versicherten eines Versicherungsbestands wurde befragt, ob sie sich vorstellen könnten, ein bestimmtes neues Versicherungsprodukt zu erwerben. Es wurde nur eine Antwort mit "Ja" oder "Nein" zugelassen. Weiter sei π_i die Wahrscheinlichkeit, dass Versicherter i angibt, das neue Produkt zu erwerben. Die individuellen Wahrscheinlichkeiten sollen modelliert werden. Dazu stehen die Merkmale "Alter des Versicherungsnehmers" (metrisch, in Jahren) und "Familienstand des Versicherungsnehmers" (drei Kategorien: verheiratet, ledig, sonstige) zur Verfügung.

(a) *[4 Punkte]* Definieren Sie eine geeignete Zielvariable Y und geben Sie die Likelihood als Funktion der π_i, $i = 1, ..., n$ an.

(b) *[4 Punkte]* Stellen Sie ein Regressionsmodell mit kanonischer Linkfunktion auf, welches die beiden Merkmale als Haupteffekte enthält.

(c) *[12 Punkte]* Die Ausgabe des Regressionsmodells sieht folgendermaßen aus:

```
Koeffizienten:
                          Schätzung Standardfehler z-Wert  p-Wert
(Intercept)               -2.32105    0.24011      -9.667  < 2e-16
alter                      0.02097    0.00526          ?   6.73e-05
familienstand verheiratet  0.0 (!)
familienstand ledig       -0.11711    0.06091      -1.923   0.0545
familienstand sonstige    -1.20371    0.10582     -11.375  < 2e-16

AIC: 9174.7
(!): Parameterschätzung auf null gesetzt, da redundant.
```

(i) *[2 Punkte]* Welche Kodierung wird für das Merkmal "Familienstand" verwendet? Geben Sie den kodierten Vektor für eine 35–jährige ledige Person an.

(ii) *[2 Punkte]* Berechnen Sie den z-Wert für das Merkmal "Alter" (*2 Kommastellen*).

(iii) *[3 Punkte]* Interpretieren Sie den geschätzten Koeffizienten für das Merkmal "Alter" hinsichtlich seines quantitativen Einflusses auf die Zielvariable.

(iv) *[3 Punkte]* Welche Kaufwahrscheinlichkeit schätzt man für einen 40–jährigen, verheirateten Versicherungsnehmer (*2 Kommastellen*)?

(v) *[2 Punkte]* Das metrische Alter ist linear in den Prädiktor aufgenommen worden. Nennen Sie zwei weitere Möglichkeiten.

(d) *[10 Punkte]* Das Alter wird nun kategorisiert in zwei Kategorien (25–45 und 46–65) und es wird ein Modell mit Interaktion von Alter und Familienstand berechnet. Folgende Likelihoodquotienten–Statistik (LR) für die Interaktion ist gegeben:

```
                          LR     Df Pr(>Chisq)
alterkat:familienstand    1.526  ?  0.47
```

(i) *[2 Punkte]* Formulieren Sie die entsprechenden Hypothesen H_0 und H_1.

(ii) *[3 Punkte]* Geben Sie die fehlende Zahl der Freiheitsgrade (Df) an (*kurze Begründung*). Welche Verteilung hat damit die Teststatistik?

(iii) *[2 Punkte]* Geben Sie die Testentscheidung zum Niveau $\alpha = 0.05$ an (*kurze Begründung*).

(iv) *[3 Punkte]* Das Modell mit kategorialem Alter (ohne Interaktion) liefert ein AIC von 9177.9. Welches Modell würde man bevorzugen: das Modell mit metrischem Alter aus Teilaufgabe c) oder das Modell mit kategorialem Alter? (*Kurze Begründung*)

Lösung Aufgabe 3

(a) *[4 Punkte]* Eine geeignete Zielvariable Y ist binär: $Y_i = 1$, falls Versicherter i das Produkt erwirbt, $Y_i = 0$ sonst. Likelihood als Funktion der π_i:

$$L(\pi_1, \ldots, \pi_{10000}) = \prod_{i=1}^{10000} \pi_i^{y_i}(1-\pi_i)^{(1-y_i)} \ .$$

(b) *[4 Punkte]* Logistisches Regressionsmodell mit Alter und Familienstand als Haupteffekte. Familienstand hat drei Kategorien, deshalb müssen 2 Dummy–Variablen verwendet werden.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Alter} + \beta_2 \text{Familienstand}_1 + \beta_3 \text{Familienstand}_2 \ .$$

(c) *[12 Punkte]*

(i) *[2 Punkte]* Es wird eine Dummy–Kodierung mit "verheiratet" als Referenzkategorie verwendet. D.h. die Kodierung ist

Familienstand	Dummy 1 (Familienstand$_1$)	Dummy 2 (Familienstand$_2$)
verheiratet	0	0
ledig	1	0
sonstige	0	1

Damit hat eine 35–jährige versicherte Person den kodierten Vektor:

Intercept	Alter	Dummy 1 (Familienstand$_1$)	Dummy 2 (Familienstand$_2$)
1	35	1	0

oder kurz $x_i = (1, 35, 1, 0)$.

(ii) *[2 Punkte]* Der z-Wert ist Schätzwert dividiert durch geschätzten Standardfehler:

$$z_{\text{Alter}} = \frac{0.02097}{0.00526} = 3.99 \ .$$

(iii) *[3 Punkte]* 3 mögliche Antworten:

- Erhöht sich das Alter um 1 Jahr, so erhöht sich die logarithmierte Chance (für Kauf des Produkts vs. Nichtkauf),

$$\log\left(\frac{\pi}{1-\pi}\right) ,$$

  additiv um $\beta_1 = 0.02097$.

- Erhöht sich das Alter um 1 Jahr, erhöht sich die Chance (Odds, für Kauf des Produkts vs. Nichtkauf) (multiplikativ) um den Faktor $\exp(\beta_1) = 1.02$.

- Der Odds Ratio ist $\exp(\beta_1) = 1.02$, d.h. die Chance eines $(x+1)$–Jahre alten Versicherten (für Kauf des Produkts vs. Nichtkauf) ist um den Faktor $\exp(\beta_1) = 1.02$ höher als die Chance eines x–Jahre alten Versicherten.

(iv) *[3 Punkte]* Prädiktor $\eta_i = -2.32105 + 40 \cdot 0.02097 = -1.48$. Damit:

$$P(Y_i = 1) = \frac{1}{1 + \exp(1.48)} = \frac{\exp(-1.48)}{1 + \exp(-1.48)} = 0.19\ .$$

Die Kaufwahrscheinlichkeit wird auf etwa 0.19 geschätzt.

(v) *[2 Punkte]* Folgende Alternativen sind möglich: Aufnahme von Transformationen des Alters, also log(Alter) oder Alter^2 oder Generalisierte Additive Modelle (GAM) oder feinere Kategorisierung.

(d) *[10 Punkte]*

(i) *[2 Punkte]* H_0: Es besteht keine Interaktion zwischen Alter und Familienstand. H_1 : Es besteht eine Interaktion zwischen Alter und Familienstand.

(ii) *[3 Punkte]* Die Zahl der Freiheitsgrade ist 2. Begründung: Alter hat 2 und Familienstand 3 Kategorien. Damit ist die Zahl der benötigten Dummy-Variablen für die Interaktion $(2-1) \cdot (3-1) = 2$. Dies ist die Differenz der Anzahl der Parameter im Modell mit Interaktion im Vergleich zum Modell ohne Interaktion. Die Teststatistik ist asymptotisch χ^2-verteilt mit 2 Freiheitsgraden.

(iii) *[2 Punkte]* Die Testentscheidung ist: H_0 beibehalten (keine Interaktion), da der p-Wert mit 0.47 größer ist als das vorgegebene $\alpha = 0.05$.

(iv) *[3 Punkte]* Das Modell mit metrischem Alter wird bevorzugt, da das AIC kleiner (9174.7) ist als beim Modell mit kategorialem Alter (9177.9). Das AIC berücksichtigt die Zahl der Parameter im Modell und kann deshalb zum Vergleich nicht hierarchischer/genesteter Modelle verwendet werden. Es gilt: Modelle mit kleinerem AIC werden bevorzugt, da AIC $= -2$likelihood$+$ 2Anzahl Parameter.

Aufgabe 4. *[30 Punkte]* Credibility

Es soll die Wahrscheinlichkeit geschätzt werden, dass ein Industrieobjekt innerhalb eines Jahres von einem Betriebsunterbrechungsschaden betroffen ist. Bei dem betrachteten Objekt wurden in den letzten 5 Jahren insgesamt 3 Jahre mit Betriebsunterbrechungsschäden beobachtet.

In einem Credibility-Modell setzt der Aktuar dazu für die Schaden"höhen"verteilung eine Binomialverteilung an: $X \sim B(1, P)$, wobei P eine Zufallsvariable („Strukturparameter") mit Wertebereich $[0; 1]$ ist. Dabei repräsentiert $X = 1$ das Ereignis, dass innerhalb eines Jahres Betriebsunterbrechungsschäden auftreten. $X = 0$ wird gesetzt, wenn innerhalb eines Jahres keine Betriebsunterbrechungsschäden auftreten.

(a) *[10 Punkte]* Aus Portfolioinformationen sei bekannt, dass $E(P) = 30\%$ und $\sqrt{Var(P)} = 10\%$. Berechnen Sie auf Basis der oben genannten Verteilung von X die *linearisierte* Credibility-Prämie H^{**} für dieses Objekt (diese kann als Schätzer für die gesuchte Wahrscheinlichkeit interpretiert werden).

(b) *[2 Punkte]* Mit welcher a-priori-Verteilung von P würde man das Ergebnis aus (a) als *allgemeine* Credibility-Prämie erhalten (die Angabe des Namens des Verteilungstyps genügt)? Wie nennt man die Beziehung, in der a-priori- und Schadenhöhenverteilung in diesem Fall zueinander stehen?

(c) *[12 Punkte]* Der Aktuar ignoriert für den Moment die Einschränkung, dass P nur Werte in $[0; 1]$ annimmt. Er entscheidet sich, als a-priori-Verteilung von P eine Normalverteilung mit Dichte $f_P(p)$ zu verwenden, für die $E(P) = 30\%$ und $\sqrt{Var(P)} = 10\%$ gilt. Zeigen Sie, dass sich hieraus als a-posteriori-Dichte

$$\frac{f_P(p) \cdot (p^3 - 2p^4 + p^5)}{0{,}01398}$$

ergibt. Berechnen Sie hieraus die *allgemeine* Credibility-Prämie H^*.

Hinweis: Ohne Beweis können Sie folgende Momente von P verwenden:
$E(P^3) = 0{,}036$; $E(P^4) = 0{,}0138$; $E(P^5) = 0{,}00558$; $E(P^6) = 0{,}00236$.

(d) *[6 Punkte]* An 7 Objekten wurden in den vergangenen 5 Jahren folgende Beobachtungen von X gemacht:

	Jahr 1	Jahr 2	Jahr 3	Jahr 4	Jahr 5	Mittelwert pro Zeile	Varianz pro Zeile
Objekt 1	0	1	0	1	1	0,60	0,30
Objekt 2	1	0	0	0	0	0,20	0,20
Objekt 3	0	0	0	0	1	0,20	0,20
Objekt 4	0	1	1	1	1	0,80	0,20
Objekt 5	1	0	0	0	0	0,20	0,20
Objekt 6	0	0	0	0	0	0,00	0,00
Objekt 7	0	0	0	1	0	0,20	0,20
				Mittelwert pro Spalte		0,3143	0,1857
				Varianz pro Spalte		0,0781	

Berechnen Sie unter den Annahmen des *Bühlmann-Modells* einen Schätzer für den Credibility-Faktor sowie die Credibility-Prämie von Objekt 1.

Lösung:

(a) *[10 Punkte]* Die linearisierte Credibility-Prämie ist

$$H^{**} = z\overline{\mathbf{X}} + (1-z)E(X),$$ 1

wobei $\overline{\mathbf{X}} = 3/5$. Des Weiteren ist $H(p) = E[X|P=p] = p$, so dass $E(X) = E\big(H(P)\big) = E(P) = 30\%$ sowie $Var\big(H(P)\big) = Var(P) = 10\%^2 = 0{,}01$ und 6
$E\big(Var(X|P)\big) = E\big(P(1-P)\big) = E(P) - E(P^2) = E(P) - \big(E(P)\big)^2 - Var(P) = 30\% - 30\%^2 - 10\%^2 = 0{,}2$. Daraus ergibt sich der Credibility-Faktor für $n = 5$ als

$$z = \frac{Var(H(P))}{\frac{1}{n}E\big(Var(X|P)\big) + Var(H(P))} = \frac{0{,}01}{\frac{1}{5}\cdot 0{,}2 + 0{,}01} = 0{,}2$$ 2

und eine Credibility-Prämie von

$$H^{**} = 0{,}2 \cdot \tfrac{3}{5} + (1-0{,}2)\cdot 30\% = 36{,}0\%.$$ 1

(b) *[2 Punkte]* Die linearisierte Credibility-Prämie fällt im Fall konjugierter Verteilungen mit der allgemeinen Credibility-Prämie zusammen. Gemäß Lemma 7.4 des Skriptes wäre dies für eine Beta-Verteilung der Fall. 2

(c) *[12 Punkte]* Nach Lemma 7.1 berechnet sich die Dichte der a-posteriori-Verteilung durch

$$f_{P|\mathbf{X}=\mathbf{x}}(p) = \frac{f_P(p)\cdot\prod_{i=1}^n f_{X|P=p}(x_i)}{\int f_P(w)\cdot\prod_{i=1}^n f_{X|P=w}(x_i)\,dw}$$ 1

Der Zähler ist dabei

$$f_P(p)\cdot\prod_{i=1}^n f_{X|P=p}(x_i) = f_P(p)\cdot\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = f_P(p)\cdot p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

$$= f_P(p)\cdot p^3(1-p)^2 = f_P(p)\cdot(p^3 - 2p^4 + p^5).$$ 3

Der Nenner ergibt sich mit dem Hinweis aus

$$\int f_P(w)\cdot(w^3 - 2w^4 + w^5)\,dw = E(P^3 - 2P^4 + P^5)$$

$$= 0{,}036 - 2\cdot 0{,}0138 + 0{,}00558 = 0{,}01398.$$ 3

Für die allgemeine Credibility-Prämie folgt hieraus mit $H(p) = E[X|P=p] = p$ 1

$$H^* = \int p\, f_{P|\mathbf{X}=\mathbf{x}}(p)\, dp = \frac{1}{0{,}01389}\int p\, f_P(p)\cdot(p^3 - 2p^4 + p^5)\, dp = \frac{E(P^4 - 2P^5 + P^6)}{0{,}01398}$$
$$= \frac{0{,}0138 - 2\cdot 0{,}00558 + 0{,}00236}{0{,}01398} = \frac{0{,}005}{0{,}01398} = 35{,}8\%.$$ 4

(d) *[6 Punkte]* Im Bühlmann-Modell lautet der Schätzer für den Credibility-Faktor

$$\hat{z} = 1 - \frac{\hat{E}\big(\sigma^2(P)\big)}{n\cdot\widehat{Var}(\overline{\mathbf{X}})} = 1 - \frac{0{,}1857}{5\cdot 0{,}0781} = 0{,}5245.$$ 4

Für Objekt 1 erhält man als Credibility-Prämie

$$H^{**} = \hat{z}\,\overline{\mathbf{X}} + (1-\hat{z})\,\widehat{E}\big(\mu(P)\big) = 0{,}5245\cdot\tfrac{3}{5} + 0{,}4755\cdot 0{,}3143 = 46{,}4\%.$$ 2