Koordinator
Konstantin Rusch

Wahrscheinlichkeit & Statistik

Lösungen 12

1. Die Vereinigten Staaten wollen das derzeitige Niveau ihrer N strategischen Ölreserven erkennen. Sie modellieren das Niveau einer einzelnen Ölreserve mit der Zufallsvariable X. Sie benutzen für X die Dichte

$$f_{\alpha,\beta}(x) = cx^{-\alpha}1_{\{x\geq\beta\}}\,,$$

wobei $\beta > 0$ and $\alpha > 1$ Parameter sind.

a) Berechnen Sie die Konstante c. Zeigen Sie, dass $E(X^s) = \frac{(\alpha-1)\beta^s}{\alpha-1-s}$, wenn $\alpha > s+1$. Berechnen Sie den Erwartungswert $E(X)$ und die Varianz $\mathrm{Var}(X)$.

Nehmen Sie an, dass der Parameter β bekannt ist. Sie wollen den Parameter α aus einer Stichprobe mit n Beobachtungen schätzen.

b) Bestimmen Sie den Momentenschätzer und den Maximum-Likelihood-Schätzer für α. Schreiben Sie die allgemeinen Formeln für n Beobachtungen hin.

Für $i = 1\ldots N$ sei X_i jene Zufallsvariable, welche das Niveau der i-ten Ölreserve repräsentiert. Sei $S_N = \sum_{i=1}^N X_i$ die Zufallsvariable, die das Gesamtniveau der N Ölreserven darstellt.

c) Wie gross ist etwa die Wahrscheinlichkeit, dass das Gesamtniveau der N strategischen Ölreserven eine Mindestgrenze δ unterschreitet? Treffen Sie dabei geeignete Annahmen, und beschreiben Sie diese. Drücken Sie diese Wahrscheinlichkeit mit α, β, N, δ und der kumulativen Verteilungsfunktion Φ der Standardnormalverteilung aus.

Lösung:

a) Damit $f_{\alpha,\beta}$ eine Dichte ist, muss gelten $\int_{\mathbb{R}} f_{\alpha,\beta}(x)dx = 1$. Also berechnen wir $\int_{\mathbb{R}} f_{\alpha,\beta}(x)dx = \int_\beta^\infty cx^{-\alpha}dx = c\left[\frac{x^{-\alpha+1}}{-\alpha+1}\right]_\beta^\infty = c\frac{\beta^{-\alpha+1}}{\alpha-1}$. Somit erhalten wir $c = (\alpha-1)\beta^{\alpha-1}$. Daraus folgt

$$\begin{aligned} E(X^s) = \int_{\mathbb{R}} x^s f_{\alpha,\beta}(x)dx &= \int_\beta^\infty (\alpha-1)\beta^{\alpha-1}x^{s-\alpha}dx \\ &= (\alpha-1)\beta^{\alpha-1}\left[\frac{x^{s-\alpha+1}}{s-\alpha+1}\right]_\beta^\infty = \frac{(\alpha-1)\beta^s}{\alpha-1-s}\,, \end{aligned}$$

für $\alpha > s+1$. Insbesondere ist $E(X) = \frac{(\alpha-1)\beta}{\alpha-2}$, $E(X^2) = \frac{(\alpha-1)\beta^2}{\alpha-3}$, $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{(\alpha-1)\beta^2}{(\alpha-3)(\alpha-2)^2}$, für $\alpha > 1$.

b) Wir wählen α so, dass das s-te theoretische Moment von X, definiert als $\mu_s = E(X^s)$, und das s-te empirische Moment von $x_1, \ldots, x_n$, definiert als $m_s = \frac{1}{n}\sum_{i=1}^n x_i^s$, übereinstimmen. Also haben wir für $s = 1$

$$\mu_1(\alpha) = m_1 \Leftrightarrow \frac{(\alpha-1)\beta}{\alpha-2} = m_1 \Leftrightarrow \alpha = \frac{\beta - 2m_1}{\beta - m_1} = 1 - \frac{m_1}{\beta - m_1} \Rightarrow \alpha = 1 - \frac{\frac{1}{n}\sum_{i=1}^n x_i}{\beta - \frac{1}{n}\sum_{i=1}^n x_i}\,.$$

Dies führt zum Momentenschätzer

$$\hat{\alpha}_{MoM} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n X_i}{\beta - \frac{1}{n}\sum_{i=1}^n X_i} = 1 - \frac{\bar{X}_n}{\beta - \bar{X}_n}.$$

Da die Beobachtungen unabhängig sind, können wir die Likelihoodfunktion schreiben als $L(x_1, \ldots, x_n, \alpha, \beta) = \prod_{i=1}^n f_{\alpha,\beta}(x_i)$. Die log-Likelihoodfunktion ist

$$\begin{aligned} l(x_1, \ldots, x_n, \alpha, \beta) = \log\left(\prod_{i=1}^n f_{\alpha,\beta}(x_i)\right) &= \sum_{i=1}^n \log f_{\alpha,\beta}(x_i) \\ &= \sum_{i=1}^n \log\left((\alpha-1)\beta^{\alpha-1}x_i^{-\alpha}\right) \\ &= n\log(\alpha-1) + n(\alpha-1)\log\beta - \alpha\sum_{i=1}^n \log x_i\,. \end{aligned}$$

Wir berechnen die Ableitung der log-Likelihoodfunktion als

$$\frac{\partial l(x_1, \ldots, x_n, \alpha, \beta)}{\partial\alpha} = \frac{n}{\alpha-1} + n\log\beta - \sum_{i=1}^n \log x_i = \frac{n}{\alpha-1} + \sum_{i=1}^n \log\frac{\beta}{x_i}\,.$$

Aus $\frac{\partial l(x_1,\ldots,x_n,\alpha,\beta)}{\partial\alpha} = 0$ folgt also $\alpha = 1 - \frac{n}{\sum_{i=1}^n \log\frac{\beta}{x_i}}$. Da die zweite Abteilung

$$\frac{\partial l^2(x_1, \ldots, x_n, \alpha, \beta)}{\partial\alpha^2} = -\frac{n}{(\alpha-1)^2}$$

negativ ist, schliessen wir, dass α ein Maximum ist. Der gesuchte Maximum-Likelihood-Schätzer ist also

$$\hat{\alpha}_{MLE} = 1 - \frac{n}{\sum_{i=1}^n \log\frac{\beta}{X_i}} = 1 + \frac{1}{\overline{(\log X)}_n - \log\beta}\,,$$

wobei $\overline{(\log X)}_n = \frac{1}{n}\sum_{i=1}^n \log X_i$.

c) Wir nehmen an, dass die N strategischen Ölreserven unabhängig voneinander sind. Eine zweite Annahme ist, dass N gross genug ist, damit die Normalverteilung via den zentralen Grenzwertsatz eine gute Approximation liefert. Laut dem zentralen

Grenzwertsatz ist S_N approximativ $\mathcal{N}(N\mu_{X_1}, N\sigma^2_{X_1})$-verteilt mit $\mu_{X_1} = E(X) = \frac{(\alpha-1)\beta}{\alpha-2}$ und $\sigma^2_{X_1} = \mathrm{Var}(X) = \frac{(\alpha-1)\beta^2}{(\alpha-3)(\alpha-2)^2}$.

Die Wahrscheinlichkeit, dass das Gesamtniveau der N strategischen Ölreserven die Mindestgrenze δ unterschreitet, ist approximativ gegeben durch

$$\begin{aligned} P(S_N \leq \delta) &= P\left(\frac{S_N - \mu_{S_N}}{\sigma_{S_N}} \leq \frac{\delta - \mu_{S_N}}{\sigma_{S_N}}\right) \approx P\left(Z \leq \frac{\delta - \mu_{S_N}}{\sigma_{S_N}}\right) \\ &= \Phi\left(\frac{\delta - \mu_{S_N}}{\sigma_{S_N}}\right) = \Phi\left(\frac{\delta - N\mu_{X_1}}{\sqrt{N}\sigma_{X_1}}\right) = \Phi\left(\frac{\delta - N\frac{(\alpha-1)\beta}{\alpha-2}}{\sqrt{N\frac{(\alpha-1)\beta^2}{(\alpha-3)(\alpha-2)^2}}}\right). \end{aligned}$$

2. In einem Sägewerk wird das Schnittholz einer bestimmten Sortierklasse einer Qualitätskontrolle unterzogen. Pro Produktionstag wird eine Stichprobe mit zehn Brettern entnommen und jedes Brett auf seine Steifigkeit getestet. Basierend auf Erfahrung kann angenommen werden, dass die Steifigkeit eines Brettes normalverteilt ist mit bekannter Standardabweichung $\sigma = 1430$ MPa [1].

a) Leiten Sie die Formel des 95%-Konfidenzintervalls für μ nach 15 Produktionstagen her.

b) Berechnen Sie aus a) das realisierte Konfidenzintervall für einen beobachteten Stichprobenmittelwert von $\overline{x} = 11'000$ MPa (nach 15 Produktionstagen).

c) Wieviele Stichproben wären nötig, damit die Breite des Konfidenzintervalls kleiner als 200 MPa ist?

Lösung:

a) Wir wissen, dass

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n) \ .$$

Also gilt

$$P\left(-z_{1-\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha,$$

wobei $z_{1-\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ das $(1 - \frac{\alpha}{2}) \times 100\%$-Quantil der Standardnormalverteilung ist. Durch Umformen erhalten wir

$$P\left(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Also hat das Vertrauensintervall die Form

$$\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

[1] Das Pascal ist eine abgeleitete SI-Einheit des Drucks sowie der mechanischen Spannung. Sie ist nach Blaise Pascal benannt und folgendermassen definiert: 1 Pa= $1\text{kg}\times\text{m}^{-1}\times\text{s}^{-2} = 1\text{N}\times\text{m}^{-2}$. Ein Pascal ist also der Druck, den eine Kraft von einem Newton auf eine Fläche von einem Quadratmeter ausübt. 1 MPa $= 10^6$ Pa $= 10$ bar.

mit $\alpha = 0.05$. Das Vertrauensintervall nach 15 Produktionstagen (mit Stichproben von je 10 Brettern) ist

$$\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{150}}.$$

b) Aus a) wissen wir, dass das Vertrauensintervall die Form $\overline{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ hat. Durch Einsetzen erhalten wir $(10771.15, 11228.85)$, wobei wir verwendet haben, dass $\Phi^{-1}(1-\frac{\alpha}{2}) = 1.96$ für $\alpha = 0.05$.

c) Die Breite des Vertrauensintervalls ist gegeben durch $2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. Also folgt, dass

$$\begin{aligned} 2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} &\leq 200\,, \\ \frac{z_{1-\alpha/2}\sigma}{100} &\leq \sqrt{n}\,, \\ 785.57 &\leq n. \end{aligned}$$

Das heisst, es muss $n \geq 786$ gelten. Dies bedeutet, dass wir entweder mindestens 79 Tage lang Stichproben von 10 Brettern sammeln oder alternativ während 15 Produktionstagen täglich mindestens 53 Bretter testen müssten.

3. Sei X eine Zufallsvariable. Wir haben 10 Beobachtungen $(x_1, \ldots, x_{10})$. Wir nehmen an, dass X $\mathcal{N}(\mu, \sigma)$-verteilt ist mit $\mu = 7$. Das arithmetische Mittel ist $\overline{x} = 8$ und die empirische Standardabweichung ist $s = 2.97$.

a) Führen Sie einen geeigneten Test auf dem 5%-Niveau durch, um zu beurteilen, ob $\mu > 7$ ist.

b) Führen Sie den gleichen Test auf dem 20%-Niveau durch.

c) Welches ist das kleinste Niveau, auf dem der Test die Nullhypothese noch verwirft (d.h. der p-Wert)?

d) Jetzt haben wir 90 Beobachtungen, wobei $\overline{x} = 8$ und $s = 2.98$. Führen Sie den gleichen Test auf dem 5%-Niveau durch.

e) Wie gross ist der p-Wert?

Lösung:

a) Da die Daten normalverteilt sind und σ unbekannt ist, führen wir einen t-Test zum Niveau $\alpha = 5\%$ durch. Wir wollen testen, ob $\mu > 7$. Daraus ergibt sich $\mu_0 := 7$ und ein einseitiger Test mit den Hypothesen

$$H_0 : \mu = \mu_0 = 7 \qquad\qquad H_A : \mu > \mu_0.$$

Mit $t_{n-1,1-\alpha} = t_{9,0.95} = 1.833$ ist der Verwerfungsbereich gegeben durch $K = [1.833; \infty)$. Die Realisierung t der Teststatistik T errechnet sich zu

$$t = \sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma} = \sqrt{10}\frac{8-7}{2.97} = 1.065 \notin K,$$

weshalb die Nullhypothese nicht verworfen wird.

b) Mit $t_{n-1,1-\alpha} = t_{9,0.80} = 0.883$ ist der Verwerfungsbereich $K = [0.883; \infty)$. Die Realisierung der Teststatistik T ist immer $t = 1.065$. Die Nullhypothese wird diesmal verworfen.

c) Der p-Wert für die gegebene Realisierung der Teststatistik errechnet sich zu $p = P(T \geq 1.065) = 1 - P(T \leq 1.065) = 1 - 0.836 = 0.164$ (lineare Interpolation zwischen $t_{9,0.80} = 0.883$ und $t_{9,0.90} = 1.383$). Wenn wir einen Test zum Niveau $\alpha \geq 16.4\%$ durchführen, wird die Nullhypothese verworfen. Wenn $\alpha < 16.4\%$, wird die Nullhypothese nicht verworfen.

d) Der Verwerfungsbereich ist $K = [t_{n-1,1-\alpha}; \infty) = [t_{89,0.95}; \infty) \approx [t_{90,0.95}; \infty) = [1.662; \infty)$. Die Realisierung t der Teststatistik T errechnet sich zu

$$t = \sqrt{n}\frac{\overline{x}_n - \mu_0}{\sigma} = \sqrt{90}\frac{8-7}{2.98} = 3.185 \in K,$$

weshalb die Nullhypothese (ziemlich klar) verworfen wird.

e) Der p-Wert für die gegebene Realisierung der Teststatistik errechnet sich zu $p = P(T \geq 3.185) = 1 - P(T \leq 3.185)$. In der Tabelle ist $t_{90,0.995} = 2.632$, also $p < 0.005 = 0.5\%$. Man bekommt $p = 0.1\%$ mit Excel. Wir könnten also einen Test zum Niveau $\alpha = 0.1\%$ durchführen und die Nullhypothese würde immer noch verworfen.

4. Welche der untenstehenden Aussagen ist **richtig**?

a) Sei X eine exponentialverteilte Zufallsvariable mit Parameter λ, d.h. $X \sim Exp(\lambda)$. Berechnen Sie die momenterzeugende Funktion $M_X(t)$.

1. $M_X(t) = \frac{\lambda}{\lambda - t}$ für $t < \lambda$.
2. $M_X(t) = \frac{\lambda}{t - \lambda}$ für $t > \lambda$.
3. $M_X(t) = \lambda$ für $t \in \mathbb{R}$.

b) Seien $X_1, \ldots, X_n$ unter P_ϑ i.i.d. $\sim Be(\vartheta)$ mit $\vartheta \in [0,1]$. Wir betrachten die Schätzer $T_1 = \sum_{i=1}^n X_i$ und $T_2 = X_2$. Welche der folgenden Aussagen trifft zu?

1. T_1 ist erwartungstreu und T_2 ist nicht erwartungstreu.
2. T_1 und T_2 sind beide erwartungstreu.
3. T_1 ist nicht erwartungstreu und T_2 ist erwartungstreu.

c) Seien $X_1, \ldots, X_n$ unter P_ϑ i.i.d. mit endlichem Erwartungswert. Ferner sei $Y = \sum_{i=1}^n \lambda_i X_i$ für beliebige $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ mit $\lambda_1 + \cdots + \lambda_n = 1$. Ist dann Y ein erwartungstreuer Schätzer für $E_\vartheta[X]$?

1. Diese Aussage ist richtig.
2. Diese Aussage ist falsch.
3. Es sind zu wenig Informationen vorhanden, um eine Aussage zu machen.

d) Seien $X_1, \ldots, X_n$ unabhängig und binomialverteilt. Ferner sei $S_n = \sum_{i=1}^n X_i$ und $\delta > 0$. Dann gilt

$$P\Big[S_n \geq (1+\delta)E[S_n]\Big] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{E[S_n]}.$$

1. Diese Aussage ist falsch.
2. Diese Aussage ist richtig.
3. Wir müssen die Parameter der Binomialverteilung kennen, um eine Aussage machen zu können.

e) Aus einer Urne mit 40 roten und 80 blauen Kugeln wird 27 Mal eine Kugel gezogen und wieder zurückgelegt. Sei A das Ereignis, dass total mindestens 18 rote Kugeln gezogen werden. Welche Aussage ist richtig?

1. $P[A] = \sum_{k=18}^{27} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{27-k}$.
2. Wegen der Chebyshev-Ungleichung gilt $P[A] > \frac{27 \cdot \frac{1}{3} \cdot \frac{2}{3}}{9^2}$.
3. Wegen der Chernoff-Schranke gilt $P[A] \leq \left(\frac{e^1}{2^2}\right)^9$.

Lösung:

a) 1. ist richtig. Man rechnet

$$\begin{aligned} M_X(t) = E[e^{tX}] &= \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \lambda \int_0^\infty e^{-(\lambda - t)x} dx = -\frac{\lambda}{\lambda - t} e^{-(\lambda-t)x}\Big|_0^\infty \\ &= \begin{cases} \frac{\lambda}{\lambda - t} & \text{falls } t < \lambda, \\ \infty & \text{falls } t \geq \lambda. \end{cases} \end{aligned}$$

Alternativ: Da X positiv ist, ist $M_X(t)$ monoton wachsend. Somit kann man 2. ausschliessen. Ausserdem kann 3. i.A. nicht stimmen, da $M_X(0) = 1$ (für alle λ).

b) 3. ist richtig. T_1 ist nicht erwartungstreu, da

$$E_\vartheta[T_1] = \sum_{i=1}^n E_\vartheta[X_i] = n\vartheta \neq \vartheta \qquad \text{für } n > 1,$$

und T_2 ist erwartungstreu, da

$$E_\vartheta[T_2] = E_\vartheta[X_2] = \vartheta \qquad \text{für alle } \vartheta.$$

c) 1. ist die richtige Anwort. Wegen der Linearität und da $X_1, \ldots, X_n$ unter P_ϑ identisch verteilt sind, gilt

$$E_\vartheta[Y] = \sum_{i=1}^n \lambda_i E_\vartheta[X_i] \overset{\text{i.d.}}{=} E_\vartheta[X_1] \sum_{i=1}^n \lambda_i = E_\vartheta[X_1].$$

d) 2. ist richtig. Da die $X_1, \ldots, X_n$ unabhängig und binomialverteilt sind, gibt es $m_1, \ldots, m_n \in \mathbb{N}$, $p_1, \ldots, p_n \in [0,1]$ und unabhängige Bernoulli-Zufallsvariablen $Y_{i,k}$, so dass $X_i = \sum_{k=1}^{m_i} Y_{i,k}$ und $Y_{i,k} \sim Be(p_i)$. Schreiben wir $S_n = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \sum_{k=1}^{m_i} Y_{i,k}$, so sind die Voraussetzungen von Satz 5.7 erfüllt und die angegebene Ungleichung gilt.

e) 3. ist richtig. Die Aussage 1. ist falsch, da der Binomialkoeffizient fehlt. Die Aussage 2. ist nicht richtig, da das Ungleichungszeichen umgekehrt sein sollte. Um 3. zu zeigen, bezeichnen wir mit X_i, $i = 1, \ldots 27$, unabhängige Bernoulli-verteilte Zufallsvariablen, und definieren $X_i = 1$, falls eine rote Kugel gezogen wurde und $X_i = 0$, falls eine schwarze Kugel gezogen wurde. Dann ist $X_i \sim Be(\frac{1}{3})$. Sei $S_n = \sum_{i=1}^{n} X_i$. Somit folgt, dass $E[S_{27}] = 27 \cdot \frac{1}{3} = 9$ ist. Also erhalten wir mit Satz 5.7 (mit $\delta = 1$) aus der Vorlesung die Ungleichung

$$P[A] = P[S_{27} \geq 18] = P[S_{27} \geq (1+1) \cdot 9] \leq \left(\frac{e^1}{2^2}\right)^9.$$