Prof. Vincent Tassion
Laurin Köhler-Schindler

Wahrscheinlichkeit & Statistik

Serie 12

Abgabe bis Mittwoch (25.05.2022) um 10:15 Uhr

Diese Serie beschäftigt sich mit approximativen Konfidenzintervallen. Ausserdem geht es um Tests und die grundlegenden Definitionen in diesem Kontext.

Weitere Informationen und Instruktionen zur Abgabe unter
https://metaphor.ethz.ch/x/2022/fs/401-0614-00L/

Aufgabe 12.1 [Approximatives Konfidenzintervall I: Geometrische Verteilung]
Sei $\Theta = [1/2, 1]$. Wir betrachten die Modellfamilie $(\mathbb{P}_\theta)_{\theta\in\Theta}$, wobei $X_1, \dots, X_n$ unter $\mathbb{P}_\theta$ unabhängig, identisch verteilt sind mit $X_1 \sim \mathrm{Geom}(\theta)$. In Quiz 10 haben wir bereits gesehen, dass der Maximum-Lieklihood Schätzer für θ gegeben ist durch

$$T_{ML} = \frac{n}{\sum_{i=1}^n X_i}.$$

Bestimme ein approximative Konfidenzintervall für θ mit Niveau 95%.
Hinweis: Für alle $\theta \in [1/2, 1]$ gilt $\frac{\sqrt{1-\theta}}{\theta} \leq \sqrt{2}$. Weiterhin wurde in den Serien 5 und 6 berechnent, dass $\mathbb{E}_\theta[X_1] = 1/\theta$ und $Var_\theta[X_1] = (1-\theta)/\theta^2$.

Aufgabe 12.2 [Approximatives Konfidenzintervall II: Binomialverteilung]
Um die Anzahl N der Forellen in einem See zu bestimmen, wird folgendermassen vorgegangen (Capture-Recapture-Methode): In einem ersten Schritt werden 500 Forellen gefangen, markiert und wieder ausgesetzt. In einem zweiten Schritt werden nochmals 200 Forellen gefangen und die Anzahl X der markierten Forellen bestimmt.

(a) Für X wird oft eine Binomialverteilung angenommen, $X \sim \mathrm{Bin}(n, \theta)$ (θ bezeichne die Wahrscheinlichkeit, dass ein im zweiten Schritt gefangener Fisch markiert ist).
Wie gross ist n? Wie gross ist der Parameter θ, wenn die Gesamtzahl der Forellen im See $N = 2000$ bzw. $N = 5000$ ist?

(b) Die tatsächliche Beobachtung für X ergibt den Wert 40. Gebe eine vernünftige Schätzung für den Parameter θ an, und leite daraus eine Schätzung für die Gesamtzahl N der Forellen im See ab.

(c) Bestimme ein approximatives Konfidenzintervall für θ mit Level 95% und daraus ein approximatives Konfidenzintervall für N mit Level 95%.

Hinweis: Benutze für den Schätzer von θ den zentralen Grenzwertsatz.

Aufgabe 12.3 [Test I: Gezinkter Würfel]
Bei einem Würfel mit sechs Seiten soll getestet werden, ob der Würfel gezinkt ist und eher auf der Sechs landet. Hierzu wird ein Experiment durchgeführt, bei dem zehn Mal gewürfelt und jeweils die beobachtete Augenzahl notiert wird. Wir gehen davon aus, dass alle Würfe unabhängig voneinander sind und die Wahrscheinlichkeit, eine 1, 2, 3, 4 oder 5 zu würfeln, gleich ist. Wir modellieren die Ausgänge der Würfe als eine Stichprobe $X_1, \ldots, X_{10}$, wobei $X_i = 1$ bedeutet, dass der i-te Wurf eine Sechs ist, und $X_i = 0$ sonst. Wir erhalten folgende Resultate:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0	0	1	0	1	1	0	1	0	0

(a) Bestimme ein geeignetes Modell $(\mathbb{P}_\theta)_{\theta\in\Theta}$, d.h. einen Parameterraum und die Verteilungen von $X_1, \ldots, X_{10}$ unter jedem $\mathbb{P}_\theta$.

(b) Bestimme eine geeignete Hypothese H_0 und Alternative H_A.

(c) Sei $T = \sum_{i=1}^{10} X_i$ die Teststatistik. Welcher Verteilung folgt T?

(d) Sei $K = (4, 10]$ der Verwerfungsbereich. Berechne die Wahrscheinlichkeit für einen Fehler 1. Art.

(e) Beschreibe den Testentscheid basierend auf den obigen Resultaten.

Aufgabe 12.4 [Test II: Normalverteilung]
Seien $X_1, \ldots, X_{12}$ unabhängig und je $\mathcal{N}(\mu, \sigma^2)$-verteilt unter $\mathbb{P}_\theta$, wobei $\theta = \mu$ ein unbekannter Parameter ist. Die Standardabweichung $\sigma = 0.0499$ ist bekannt. Wir haben folgende Daten für die Stichprobe gegeben:

x_1	x_2	x_3	x_4	x_5	x_6
1.00781	1.00646	1.00801	1.00833	1.00738	1.00687
x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
1.00783	1.00936	1.00564	1.00543	1.00794	1.01060

Wir testen die Hypothese $H_0\colon \mu = \mu_0 = 1.0085$ gegen die Alternative $H_1 : \mu \neq \mu_0$.

(a) Bestimme a und b so, dass die Teststatistik $T := \frac{\sum_{i=1}^{12} X_i + b}{a} \sim \mathcal{N}(0, 1)$-verteilt ist unter $\mathbb{P}_{\mu_0}$.

(b) Wähle $K := (-\infty, -c_{\neq}) \cup (c_{\neq}, \infty)$ als Verwerfungsbereich für ein zu bestimmendes $c_{\neq} \geq 0$. Teste H_0 gegen H_A für das Signifikantsniveau 5%.

(c) Berechne die Macht des oben durchgeführten Tests an der Stelle $\mu = 1.008$.

Lösung 12.1 Da $X_1 \sim \mathrm{Geom}(\theta)$ den Erwartungswert $1/\theta$ und die Varianz $(1-\theta)/\theta^2$ hat, ist

$$\frac{\sum_{i=1}^n X_i - n/\theta}{\sqrt{n(1-\theta)/\theta^2}}$$

approximativ $\mathcal{N}(0,1)$-verteilt für $n \to \infty$ nach dem Zentralen Grenzwertsatz. Nach dem Hinweis gilt $\sqrt{1-\theta}/\theta \leq \sqrt{2}$ für alle $\theta \in [1/2, 1]$ und somit gilt für alle $\theta \in [1/2, 1]$, dass

$$\mathbb{P}_\theta\left[-1.96 \leq \frac{\sum_{i=1}^n X_i - n/\theta}{\sqrt{2n}} \leq 1.96\right] \geq \mathbb{P}_\theta\left[-1.96 \leq \frac{\sum_{i=1}^n X_i - n/\theta}{\sqrt{n(1-\theta)/\theta^2}} \leq 1.96\right] \geq 0.95.$$

Da wir ein approximatives Konfidenzintervall suchen, also n gross annehmen, können wir auch annehmen, dass $\sqrt{n} \geq 1.96 \cdot \sqrt{2}$ und somit $\sum_{i=1}^n X_i \geq n \geq 1.96 \cdot \sqrt{2n}$, da X_i Werte in $\mathbb{N} \setminus \{0\}$ annimmt. Damit gilt

$$-1.96 \leq \frac{\sum_{i=1}^n X_i - n/\theta}{\sqrt{2n}} \leq 1.96 \iff \frac{1}{(T_{ML})^{-1} + \frac{1.96\sqrt{2}}{\sqrt{n}}} \leq \theta \leq \frac{1}{(T_{ML})^{-1} - \frac{1.96\sqrt{2}}{\sqrt{n}}},$$

also ergibt sich ein approximatives 95%-Vertrauensintervall für θ als

$$\left[\frac{1}{(T_{ML})^{-1} + \frac{1.96\sqrt{2}}{\sqrt{n}}}, \frac{1}{(T_{ML})^{-1} - \frac{1.96\sqrt{2}}{\sqrt{n}}}\right].$$

Lösung 12.2

(a) Weil wir im zweiten Schritt 200 Fische herausziehen, ist $n = 200$. Die Wahrscheinlichkeit, im zweiten Schritt einen markierten Fisch zu fangen, ist also

$$\theta = \frac{\text{Anzahl markierte Fische}}{\text{Totale Anzahl Fische im See}} = \frac{500}{N}. \tag{1}$$

Für $N = 2000$ ergibt dies $\theta = 1/4$, und für $N = 5000$ erhält man $\theta = 1/10$.

(b) Wir schätzen θ durch $T = X/n$. Der realisierte Schätzwert ist also

$$T^{(\theta)}(\omega) = X(\omega)/n = 40/200 = 1/5.$$

Wenn wir (1) nach N auflösen, ergibt sich für die Gesamtanzahl der Forellen im See die Schätzung

$$T^{(N)} = \frac{500}{T^{(\theta)}} = \frac{500n}{X},$$

mit realisiertem Wert

$$T^{(N)}(\omega) = \frac{500}{T^{(\theta)}(\omega)} = 2500.$$

(c) Da $X \sim \mathrm{Bin}(n, \theta)$ Erwartungswert $n\theta$ und Varianz $n\theta(1-\theta)$ hat, ist

$$\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

approximativ $\mathcal{N}(0,1)$-verteilt für $n \to \infty$ nach dem Zentralen Grenzwertsatz. Also gilt für alle $\theta \in [0,1]$, dass

$$\mathbb{P}_\theta\left[-1.96 \leq \frac{X - n\theta}{\sqrt{n/4}} \leq 1.96\right] \geq \mathbb{P}_\theta\left[-1.96 \leq \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \leq 1.96\right] \geq 0.95,$$

wobei wir wie in der Vorlesung verwendet haben, dass $\theta(1-\theta) \leq 1/4$ für alle $\theta \in [0,1]$. Somit ergibt sich ein approximatives 95%-Vertrauensintervall für θ als

$$\left[T^{(\theta)} - \frac{1.96}{2\sqrt{n}}, T^{(\theta)} + \frac{1.96}{2\sqrt{n}}\right].$$

Der realisierte Wert ergibt sich durch Einsetzen von $T^{(\theta)}(\omega) = 1/5$ als $[0.13, 0.27]$. Unter Verwendung von (1) ergibt sich daraus $[1851, 3846]$ als realisiertes 95%-Konfidenzintervall für N. Das ist Methode 1 im Skript. Mit Methode 2 erhalten wir das realisierte 95%-Konfidenzintervall $[1957, 3459]$ für N.

Lösung 12.3

(a) Aus den Annahmen folgt, dass $X_1, \ldots, X_{10}$ unabhängig und jeweils $\mathrm{Ber}(\theta)$-verteilt sind unter $\mathbb{P}_\theta$ mit unbekanntem Erfolgsparameter $\theta := P[X_1 = 1] \in \Theta = [0,1]$.

(b) Unsere Nullhypothese ist, dass der Würfel nicht gezinkt ist; also ist

$$H_0 : \theta = \frac{1}{6}, \quad \text{d.h. } \Theta_0 = \left\{\frac{1}{6}\right\}.$$

Die Alternativhypothese, dass der Würfel gezinkt ist, ist dann

$$H_A : \theta > 1/6, \quad \text{d.h. } \Theta_A = \left(\frac{1}{6}, 1\right].$$

(c) Da $X_1, \ldots, X_{10}$ unabhängig und $\mathrm{Ber}(\theta)$-verteilt sind unter $\mathbb{P}_\theta$, ist $T \sim \mathrm{Bin}(10, \theta)$ unter $\mathbb{P}_\theta$.

(d) Die Wahrscheinlichkeit für einen Fehler 1. Art ist

$$\begin{aligned}
\mathbb{P}_{\frac{1}{6}}[T \in K] &= \mathbb{P}_{\frac{1}{6}}[T \geq 5] \\
&= \sum_{k=5}^{10} \binom{10}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{10-k} \\
&= \frac{1}{6^{10}} \left(\binom{10}{5} 5^5 + \binom{10}{6} 5^4 + \binom{10}{7} 5^3 \right. \\
&\qquad \left. + \binom{10}{8} 5^2 + \binom{10}{9} 5^1 + \binom{10}{10} 5^0 \right) \\
&= 0.0155.
\end{aligned}$$

(e) Da $t(x_1, \ldots, x_{10}) = t(X_1(\omega), \ldots, X_{10}(\omega)) = T(\omega) = 4$ nicht im Verwerfungsbereich liegt, verwerfen wir die Nullhypothese nicht.

Lösung 12.4

(a) Wegen $\sum_{i=1}^{12} X_i \sim \mathcal{N}(12\mu_0, 12\sigma^2)$ unter $\mathbb{P}_{\mu_0}$ ist

$$T = \frac{\sum_{i=1}^{12} X_i - 12\mu_0}{\sigma\sqrt{12}} \sim \mathcal{N}(0,1)$$

unter $\mathbb{P}_{\mu_0}$. Wir wählen also $a = \sigma\sqrt{12}$ und $b = -12\mu_0$.

(b) Sei $\alpha = 0.05$. Wir führen einen Test mit der Teststatistik T und dem Verwerfungsbereich K durch, verwerfen also die Hypothese, falls $|T| > c_{\neq}$ für ein noch zu bestimmendes $c_{\neq}$. Die Definition des Niveaus ergibt

$$\begin{aligned}\alpha = \mathbb{P}_{\mu_0}[T \in K] = \mathbb{P}_{\mu}\big[T \notin [-c_{\neq}, c_{\neq}]\big] &= \mathbb{P}_{\mu_0}[T < -c_{\neq}] + \mathbb{P}_{\mu_0}[T > c_{\neq}] \\ &= \Phi(-c_{\neq}) + 1 - \Phi(c_{\neq}) = 2 - 2\Phi(c_{\neq}),\end{aligned}$$

da $T \sim \mathcal{N}(0,1)$-verteilt ist unter $\mathbb{P}_{\mu_0}$. Aus

$$\Phi(c_{\neq}) = 1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975$$

folgt, dass $c_{\neq} = \Phi^{-1}(0.975) = z_{0.975} = 1.96$. Der realisierte Schätzwert ist

$$T(\omega) = t(x_1, \ldots, x_{10}) = -0.00598;$$

also verwerfen wir die Hypothese nicht.

(c) Die Macht des Tests an der Stelle μ ist

$$\begin{aligned}\mathbb{P}_{\mu}[T \in K] &= \mathbb{P}_{\mu}\big[T \notin [-c_{\neq}, c_{\neq}]\big] \\ &= \mathbb{P}_{\mu}\left[\frac{\sum_{i=1}^{12} X_i - 12\mu}{\sigma\sqrt{12}} < -c_{\neq} + \frac{\sqrt{12}}{\sigma}(\mu_0 - \mu)\right] \\ &\qquad + \mathbb{P}_{\mu}\left[\frac{\sum_{i=1}^{12} X_i - 12\mu}{\sigma\sqrt{12}} > c_{\neq} + \frac{\sqrt{12}}{\sigma}(\mu_0 - \mu)\right] \\ &= \Phi\left(-c_{\neq} + \frac{\sqrt{12}}{\sigma}(\mu_0 - \mu)\right) + 1 - \Phi\left(c_{\neq} + \frac{\sqrt{12}}{\sigma}(\mu_0 - \mu)\right) \\ &= \Phi(-1.93) + 1 - \Phi(1.99) = 0.0501.\end{aligned}$$

Das ist ein niedriger Wert, und der Grund dafür ist, dass μ sehr nahe bei μ_0 liegt.