Wahrscheinlichkeit und Statistik

Musterlösung Serie 13

1. Wir wollen den Effekt von Ausreisser auf Vertrauensintervalle untersuchen. Seien $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ bei unbekanntem σ.

   a) Geben Sie das zweiseitige Vertrauensintervall für das unbekannte μ zum Niveau α an.

   b) Wie verhält sich das Vertrauensintervall für $x_1 \to \infty$ bei festen $x_2, \dots, x_n$?
   Hinweis: Für jedes $c \in \mathbb{R}$ gilt $\sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(c - \bar{x}_n)^2$.

Lösung:

a) $\left[\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}, \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right]$.

b) Der rechte Eckpunkt geht klarerweise gegen ∞, weil es mindestens $\bar{x}_n$ ist.
Für den linken Eckpunkt setzen wir $a \coloneqq \frac{1}{n} \sum_{i=2}^n x_i$. Mit dem Hinweis erhalten wir

$$\begin{aligned}
s_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - a)^2 - \underbrace{\frac{n}{n-1}(a - \bar{x}_n)^2}_{= \frac{1}{n(n-1)} x_1^2} \\
&= \underbrace{\frac{1}{n-1} \sum_{i=2} (x_i - a)^2}_{=:b} + \frac{1}{n-1}(x_1 - a)^2 - \frac{1}{n(n-1)} x_1^2 \\
&= b + \frac{1}{n-1}(x_1^2 - 2x_1 a + a^2) - \frac{1}{n(n-1)} x_1^2 \\
&= x_1^2 \left(\frac{1}{n-1} - \frac{1}{n(n-1)} - \frac{2a}{x_1} \frac{1}{n-1} + \frac{1}{n-1} \frac{a^2}{x_1^2} + \frac{b}{x_1^2} \right) =: x_1^2 f(x_1),
\end{aligned}$$

und $f(x_1) \to \frac{1}{n-1} - \frac{1}{n(n-1)} = \frac{1}{n}$, wenn $x_1 \to \infty$. Ausserdem ist $\bar{x}_n = \frac{1}{n} x_1 + a = x_1 \left(\frac{1}{n} + \frac{a}{x_1} \right) =: x_1 g(x_1)$, wobei $g(x_1) \to \frac{1}{n}$, wenn $x_1 \to \infty$. Insgesamt ergibt sich

$$\begin{aligned}
\bar{x}_n - \frac{s_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} &= \bar{x}_n \left(1 - \frac{s_n}{\bar{x}\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \right) = \bar{x}_n \left(1 - \frac{x_1 \sqrt{f(x_1)}}{x_1 g(x_1) \sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \right) \\
&= \bar{x}_n \underbrace{\left(1 - \frac{\sqrt{f(x_1)}}{g(x_1)\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \right)}_{\xrightarrow[x_1 \to \infty]{} 1 - t_{n-1,1-\frac{\alpha}{2}}}.
\end{aligned}$$

Für alle in der Praxis verwendeten Niveaus ist das t-Quantil > 1, d.h. der linke Endpunkt geht gegen $-\infty$, das Vertrauensintervall liefert überhaupt keine brauchbare Information mehr, jeder Wert ist plausibel für $\mu = \mathbb{E}[X]$.

Beweis vom Hinweis:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-c)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}_n+\bar{x}_n-c)^2\\ &= \sum_{i=1}^{n}(x_i-\bar{x}_n)^2+2(\bar{x}_n-c)\sum_{i=1}^{n}(x_i-\bar{x}_n)+n(\bar{x}_n-c)^2\\ &= \sum_{i=1}^{n}(x_i-\bar{x}_n)^2+n(\bar{x}_n-c)^2,\end{aligned}$$

wie behauptet.

2. Betrachten Sie die Nullhypothese $X \sim f(x)dx$ und die Alternative $X \sim f(x-1)dx$ in den beiden Fällen

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad \text{bzw.} \qquad f(x) = \frac{1}{\pi(1+x^2)},$$

und bestimmen die qualitative Form des Verwerfungsbereichs beim Likelihood-Quotienten-Test (Lemma von Neyman-Pearson) in den beiden Fällen. Kommentieren Sie den Unterschied.

Lösung: Beim Neyman-Pearson-Test der Hypothese $H_0 : X \sim f(x)\,dx$ gegen die Alternative $H_A : X \sim f(x-1)\,dx$ ist der Likelihood-Quotient gegeben durch

$$R(x) \coloneqq \frac{f_1(x)}{f_0(x)} = \frac{f(x-1)}{f(x)}.$$

a) Im Fall der Normalverteilung ist $R(x) = \frac{f(x-1)}{f(x)} = e^{x-\frac{1}{2}}$, und wir verwerfen H_0 wenn $R(x) > c = c(\alpha)$, d.h. wenn $x > \ln(c) + \frac{1}{2}$. Der Verwerfungsbereich ist von der Form (a, ∞) für ein $a \in \mathbb{R}$.

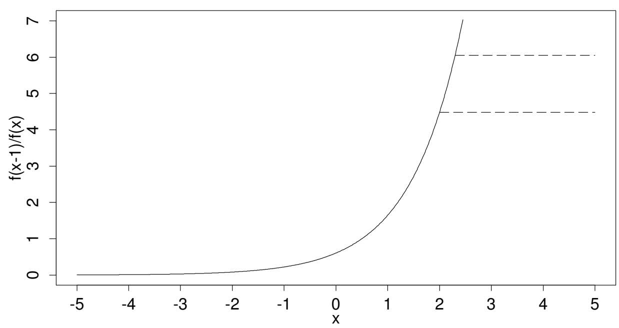


Abbildung 1: Verwerfungsbereiche für den Fall der Normalverteilung.

b) Im Fall der Cauchy-Verteilung ist $R(x) = \frac{f(x-1)}{f(x)} = \frac{x^2+1}{x^2-2x+2}$, und wir haben ein interessantes Verhalten, wie man in Abbildung 2 sehen kann. Wenn wir $c = 1$ setzen, haben wir ein unbeschränktes Intervall; wenn wir aber $c > 1$ setzen, erhalten wir ein beschränktes Intervall. (Natürlich hängt die Wahl von c vom Niveau α des Tests ab. Z.B. entspricht $c = 1$ einem Verwerfungsbereich von $x > 1/2$, was $\alpha \approx 0.3524$ entspricht.)

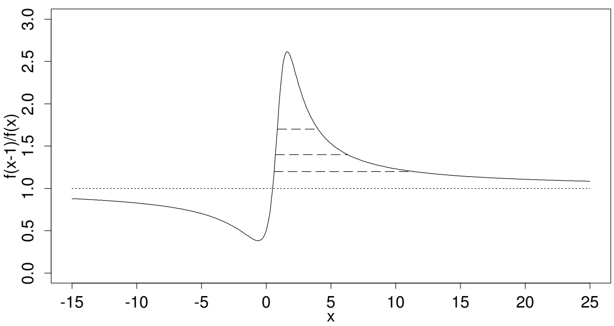
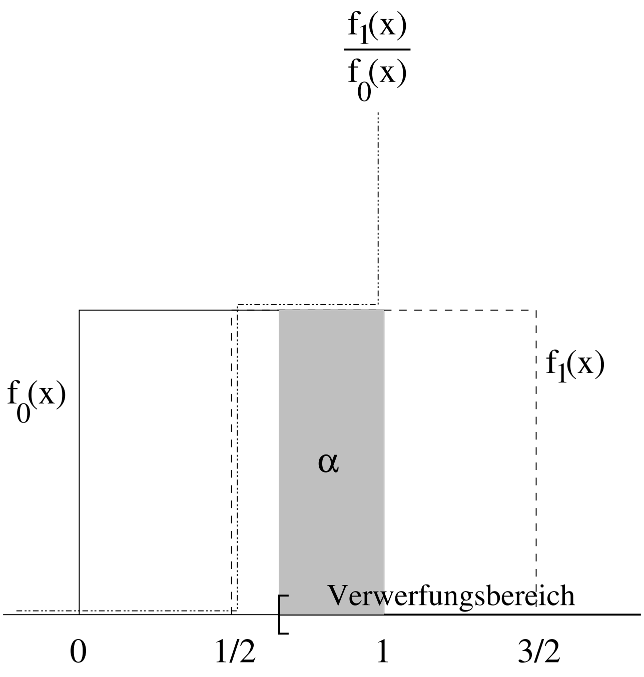


Abbildung 2: Verwerfungsbereiche für den Fall der Cauchy-Verteilung.

Die Erklärung ist, dass bei der Cauchyverteilung sehr grosse Werte sowohl unter H_0 als auch unter H_A etwa gleich unwahrscheinlich sind. Deshalb ist es für nicht allzu grosse Werte von α besser, bei grosser Beobachtung X die Nullhypothese nicht zu verwerfen. Bei der Normalverteilung tritt dieses Phänomen nicht auf, da grosse X unter der Alternativhypothese sehr viel weniger unwahrscheinlich sind als unter H_0.

3. Wir betrachten die beiden Verteilungen $\mu_0 = \text{Uniform}(0,1)$ und $\mu_1 = \text{Uniform}(\frac{1}{2}, \frac{3}{2})$.
   a) Geben Sie für jedes $\alpha \in (0,1)$ einen mächtigsten Test zum Niveau α an, der (150) und (151) in den Slides erfüllt. Ist ein solcher Test eindeutig?
   b) Zeigen Sie, dass man für gewisse α's einen mächtigsten Test φ' konstruieren kann, der (150) nicht erfüllt.

Lösung: Betrachten Sie die folgende Figur:



a) Für $\alpha \in (0, \frac{1}{2}]$ wähle $K' \subseteq (\frac{1}{2}, 1)$ mit $\mu_L(K') = \alpha$ (ist die Situation in der Figur) und als

Verwerfungsbereich $K = K' \cup [1, \infty)$. Dieser Test ist offensichtlich ein mächtigster Test zum Niveau α, und es gilt $\varphi = 1$, falls $f_1(x) > f_0(x)$, und $\varphi = 0$, falls $f_1(x) < f_0(x)$ (also $c = 1$).
Für $\alpha \in [\frac{1}{2}, 1)$ wähle $K' \subseteq (0, \frac{1}{2})$ mit $\mu_L(K') = \alpha - \frac{1}{2}$ und den Verwerfungsbereich $K = K' \cup (\frac{1}{2}, \infty)$. Auch dieser Test ist mächtigst zum Niveau α, und er erfüllt die Bedingungen vom Skript mit $c = 0$.
Solche Tests sind nicht eindeutig, da K' eine beliebige Teilmenge von $(\frac{1}{2}, 1)$ bzw. von $(0, \frac{1}{2})$ sein kann.

b) Für $\alpha \in (\frac{1}{2}, 1)$ (zum Beispiel für $\alpha = \frac{3}{4}$) wähle $K = (\frac{1}{2}, \infty)$. Dieser Test verwirft immer, wenn die Alternative wahr ist. Er hat also maximale Macht $\mathbb{E}_1[\varphi'] = 1$, obwohl er das zur Verfügung stehende Niveau nicht einmal ganz ausnützt: $\mathbb{E}_0[\varphi'] = \frac{1}{2} < \frac{3}{4}$.

4. Seien $X_1, \ldots, X_n$ unabhängig und identisch $\sim F$-verteilte Zufallsvariablen wobei F absolutstetig ist. Der Vorzeichentest ist ein Test der die Nullhypothese, dass der Median von F gleich einem vorgegebenen Wert m ist, d.h.

$$F^{-1}(\tfrac{1}{2}) = m.$$

Benutzen Sie den Dualitätssatz (S. 408 in den Slides), um ein Vertrauensintervall für den Median von F zum Niveau 95% zu konstruieren.

Lösung: Wir sollen einen Test zwischen den Hypothesen konstruieren:

$$H_0\colon F^{-1}(0.5) = m,$$

$$H_A\colon F^{-1}(0.5) \neq m.$$

Beim Vorzeichentest benutzen wir die Teststatistik $T_{n,m} = \sum_{i=1}^n \mathbf{1}_{\{X_i > m\}}$ und der Test ist gegeben durch

$$\varphi(\boldsymbol{x}) = 1 \qquad \Longleftrightarrow \qquad \left| T_{n,m} - \tfrac{n}{2}, \right| > c(n, \alpha),$$

wobei $\boldsymbol{x} = (x_1, \ldots, x_n)$ die Beobachtungen, n die Anzahl der Beobachtungen und α das Niveau des Tests bezeichnet. Weil F als stetig angenommen wurde, ist $T_{n,m}$ unter der Nullhypothese $\text{Bin}(n, 0.5)$-verteilt. Also ist $k = \frac{n}{2} - c(n, \alpha)$ bestimmt durch

$$\mathbb{P}_{H_0}(T_{n,m} < k) \sum_{j=0}^{k-1} \binom{n}{j} 0.5^n \leq \tfrac{\alpha}{2} < \sum_{j=0}^{k} \binom{n}{j} 0.5^n = \mathbb{P}_{H_0}(T_{n,m} \leq k),$$

und $\frac{n}{2} + c(n, \alpha) = n - k$. Wenn wir $A \subseteq \mathbb{R}^{n+1}$ definieren durch

$$A = \{(\boldsymbol{x}, m) \in \mathbb{R}^n \times \mathbb{R}\colon k \leq \sum_{i=1}^n \mathbf{1}_{\{X_i > m\}} \leq n - k\},$$

dann ist der Schnitt $K(m) = \{\boldsymbol{x} \in \mathbb{R}^n\colon (\boldsymbol{x}, m) \in A\}$ der Annahmebereich des Vorzeichentests zum Niveau α. Mit dem Dualitätssatz bildet der Schnitt $C(\boldsymbol{x}) = \{m \in \mathbb{R}\colon (\boldsymbol{x}, m) \in A\}$ daher einen Vertrauensbereich zum Niveau $1 - \alpha$. Wenn $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ die der Grösse nach geordneten Beobachtungen bezeichnet, dann gilt

$$m \in [x_{(j)}, x_{(j+1)}) \Leftrightarrow \sum_{i=1}^n \mathbf{1}_{\{X_i > m\}} = n - j.$$

Also ist $m \in C(\boldsymbol{x})$ genau dann, wenn $x_{(k)} \leq m \leq x_{(n-k)}$, d.h. $[X_{(k)}, X_{(n-k)}]$ ist ein Vertrauensintervall für den Median zum Niveau $1-\alpha$ (weil F als stetig angenommen wurde, spielt es keine Rolle, ob man das offene oder abgeschlossene Intervall nimmt). Wegen des zentralen Grenzwertsatzes gilt

$$\begin{aligned}
\mathbb{P}_m\left(k \leq \sum_{i=1}^n \mathbf{1}_{\{X_i>m\}} \leq n-k\right) &= \mathbb{P}_m\left(\frac{2}{\sqrt{n}}\left(k-\frac{n}{2}\right) \leq \frac{2}{\sqrt{n}}\left(\sum_{i=1}^n \mathbf{1}_{\{X_i>m\}} - \frac{n}{2}\right) \leq \frac{2}{\sqrt{n}}\left(\frac{n}{2}-k\right)\right) \\
&\approx \Phi\left(\frac{2}{\sqrt{n}}\left(\frac{n}{2}-k\right)\right) - \Phi\left(\frac{2}{\sqrt{n}}\left(k-\frac{n}{2}\right)\right) \\
&= 2\Phi\left(\frac{2}{\sqrt{n}}\left(\frac{n}{2}-k\right)\right) - 1 \geq 0.95,
\end{aligned}$$

Also folgt: $k \approx \frac{n}{2} - \frac{1.96}{2}\sqrt{n} \approx \lfloor \frac{n}{2} - \sqrt{n} \rfloor$. Dann ist

$$C((X_1, \ldots, X_n)) \approx [X_{(\lfloor n/2-\sqrt{n}\rfloor)}, X_{(\lfloor n/2+\sqrt{n}\rfloor)}]$$

ein Konfidenzintervall zum Niveau 95%.