Statistische Methoden der Datenanalyse

Markus Schumacher

Übung XI

Matthew Beckingham und Markus Warsinsky

15.7.2009

Anwesenheitsaufgaben

Aufgabe 46 *Konfidenzintervalle für eine Exponentialverteilung*

Betrachtet werden soll eine Stichprobe mit Umfang n einer nach der Exponentialverteilung mit Mittelwert τ

$$f(x;\tau) = \frac{1}{\tau}\exp\left(-\frac{x}{\tau}\right) \tag{1}$$

verteilten Zufallsvariablen x. Der Maximum Likelihood Schätzer des Mittelwertes ist dann gegeben durch:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2}$$

Die Wahrscheinlichkeitsdichtefunktion für diese Zufallsvariable ist dann, wie in der Vorlesung gezeigt, gegeben durch:

$$g(\hat{\tau};\tau) = \frac{n^n}{(n-1)!}\frac{\hat{\tau}^{n-1}}{\tau^n}\exp\left(-n\frac{\hat{\tau}}{\tau}\right). \tag{3}$$

Im folgenden sollen mit verschiedenen Methoden Aussagen über den wahren Wert τ getroffen werden.

(i) Wie groß ist die Standardabweichung auf den Schätzer $\hat{\tau}$. Nehmen Sie an, es wurde in einer Stichprobe vom Umfang $n = 5$ der Schätzwert $\hat{\tau} = 1$ gemessen. Geben Sie für diesen Fall das Intervall entsprechend plus und minus einer Standardabweichung um den Schätzer an.

(ii) Im folgenden soll nun mittels der Neyman-Konstruktion das Vertrauensintervall für den wahren Wert konstruiert werden. Zeigen Sie dazu zunächst, dass die beiden den Konfidenzbereich („confidence belt“) definierenden Kurven $u_\alpha(\tau)$ und $l_\beta(\tau)$ gegeben sind durch:

$$u_\alpha(\tau) = \frac{\tau}{2n}F^{-1}_{\chi^2}(1-\alpha;2n), \tag{4}$$

$$l_\beta(\tau) = \frac{\tau}{2n}F^{-1}_{\chi^2}(\beta;2n), \tag{5}$$

wobei $0 \leq \alpha,\beta \leq 1$ die Wahrscheinlichkeiten dafür sind, ein $\hat{\tau}$ zu erhalten, welches größer bzw. kleiner gleich u_α bzw. l_β ist. $F^{-1}_{\chi^2}$ sind die Quantile der χ^2-Verteilung. Benutzen Sie zur Herleitung folgende Beziehung zwischen der kumulativen χ^2-Verteilung und den inkompletten Gamma-Funktionen $P(n,x)$:

$$F_{\chi^2}(2x;2n) = P(n,x) \equiv \frac{1}{(n-1)!}\int_0^x \exp(-t)t^{n-1}dt. \tag{6}$$

(iii) Zeichnen Sie u_α und l_β für $\alpha = \beta = 0.159$ und $n = 5$. (<u>Hinweis:</u> $F^{-1}_{\chi^2}(0.159;10) = 5.685$, $F^{-1}_{\chi^2}(1-0.159;10) = 14.32$)

(iv) Ermitteln Sie das Vertrauensintervall $[a,b]$ als Funktion des Schätzwertes $\hat{\tau}$, des Stichprobenumfangs n und der Vertrauensniveaus α und β. Zeichnen Sie das Vertrauensintervall in das erstellte Diagramm ein. Ermitteln Sie das Vertrauensintervall für $\alpha = \beta = 0.159$ und $n = 5$. Vergleichen Sie dieses Resultat mit dem aus Unteraufgabe (i).

(v) Eine weitere Möglichkeit, das Vertrauensintervall abzuschätzen ist über die sogenannte Bartlett-Funktion. Diese lässt sich aus der maximierten Likelihoodfunktion L berechnen zu:

$$S(\tau) = \frac{\frac{\partial \ln L}{\partial \tau}}{E\left[-\frac{\partial^2 \ln L}{\partial \tau^2}\right]^{\frac{1}{2}}}. \tag{7}$$

Für grosse n geht die Bartlett-Funtion in eine Gaussverteilung mit Mittelwert Null und Standardabweichung eins über.

a) Berechnen Sie die Bartlett-Funktion für das Beispiel einer zugrundeliegenden Exponentialverteilung.

b) Benutzen Sie die Bartlett-Funktion, um ein Vertrauensintervall zu konstruieren. Benutzen Sie dabei die Annahme einer Normalverteilung.

c) Berechnen Sie das Vertrauensintervall explizit für $\alpha = \beta = 0.159$ und $n = 5$. Vergleichen Sie das Ergebnis mit der exakten Rechnung aus (iv).

Aufgabe 47 *Vertrauensintervalle für den Korrelationskoeffizienten*

In vielen Fällen sind nicht Mittelwerte oder Breiten von Verteilungen von Interesse, sondern insbesondere auch die Korrelation von zwei Größen. Oftmals versucht man aus statistischen Daten festzustellen, ob zwei Meßgrößen innerhalb der vorhandenen Statistik unabhängig voneinander sind, oder nicht.

Gegeben seien zwei Zufallsvariablen x und y, die nach einer zweidimensionalen Gaussverteilung mit Korrelationskoeffizient ρ verteilt sind. Unter Verwendung einer Stichprobe vom Umfang n (bei jeder Messung wurde sowohl ein x als auch ein y Wert gemessen), soll der Korrelationskoeffizient ρ durch einen Schätzer r abgeschätzt und Vertrauensintervalle für ρ angegeben werden. Als Schätzer kann beispielsweise

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{k=1}^{n}(y_k - \bar{y})^2\right)^{\frac{1}{2}}} \tag{8}$$

benutzt werden. Dessen Erwartungswert und Varianz ergeben sich dann zu

$$E[r] = \rho - \frac{\rho(1-\rho^2)}{2n} + \mathcal{O}(n^{-2}), \tag{9}$$

$$V[r] = \frac{1}{n}(1-\rho^2)^2 + \mathcal{O}(n^{-2}). \tag{10}$$

Die Wahrscheinlichkeitsdichtefunktion $g(r;\rho,n)$ geht nun allerdings nur sehr langsam mit grossen n gegen eine Gaussverteilung ($n \geq 500$ wird empfohlen). Um nun ein Vertrauensintervall zu erhalten, müßte man die Neyman-Konstruktion durchführen, was in diesem Fall sehr kompliziert werden würde.

Stattdessen bietet sich an, eine transformierte Variable

$$z = \tanh^{-1} r = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{11}$$

zu benutzen, mit der man einen Schätzer auf die Größe

$$\zeta = \tanh^{-1} \rho = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \tag{12}$$

erhält. Man kann zeigen, dass rwartungswertswert und Varianz von z sich näherungsweise ergeben zu:

$$E[z] = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}, \tag{13}$$

$$V[z] = \frac{1}{n-3}. \tag{14}$$

Es kann nun gezeigt werden, dass die so definierte Zufallsvariable z wesentlich schneller, d.h. für kleinere n nach einer Gaussfunktion verteilt ist.

(i) Benutzen Sie diese Eigenschaft, um ein Vertrauensintervalle $[a,b]$ für ζ und $[A,B]$ für ρ zu erhalten.

(ii) Wie würde das Vertrauensintervall von ρ aussehen, wenn man die Transformation nicht benutzen würde und direkt annehmen würde, dass r gaussverteilt ist?

(iii) Betrachten Sie nun eine Messung mit $n = 20$ und einem erhaltenen r von 0.5. Geben Sie Konfidenzintervalle zum Vertrauensniveau von 68.3% und 99% für diese beiden Methoden an. Inwiefern unterscheiden sich die erhaltenen Aussagen?

(iv) Mit welchem Vertrauensniveau würde man nach den beiden Methoden eine Korrelation von Null ausschliessen.