Programowanie narzędzi analitycznych – Z07

1 Metoda Największej Wiarygodności

Zadanie 1
Z rozkładu $N(\mu, 1)$ wygenerowanych zostało 40 obserwacji:
-1.13,-0.62,0.06,-1.82,-0.27,-0.74,0.49,-0.35,-0.41,-0.05,-0.20,0.06,-1.06,0.58,1.20, 1.59,0.20,0.65,-0.53,-0.73,-1.16,1.34,1.77,0.70,-0.10,-0.30,0.60,0.49,-0.45,-1.15,-0.31,1.30,0.91,0.47,-0.44,1.23,-0.15,1.22,-1.26,-0.18
Oszacować parametr μ metodą największej wiarygodności. Przetestować statystyczną istotność parametru μ.

$$L(\mu; x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \tag{1}$$

$$\ln L(\mu; x) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2}$$

Zadanie 2
W bardzo wietrzny dzień w pewnej miejscowości w wielu miejscach rozpalano grilla za pomocą zapałek. Zanotowana została dla każdego przypadku liczba prób do osiągnięcia sukcesu:
0,0,3,1,3,11,0,7,2,3,4,2,0,1,2,1,3,2,4,2,1,4,6,4,0,6,0,0,4,2.
Sprawdzić istotność parametru p oraz przetestować hipotezę, że $p = 0.25$.

$$L(p; k) = \prod_{i=1}^{n}(1-p)^{k_i - 1}p = p^n(1-p)^{-n}\prod_{i=1}^{n}(1-p)^{k_i} \tag{3}$$

$$\ln L(p; k) = n\ln(p) - n\ln(1-p) + \ln(1-p)\sum_{i=1}^{n} k_i \tag{4}$$

Zadanie 3
Zbiór danych `ratings_Musical_Instruments.csv` zawiera informacje o liczbie recenzji różnych instrumentów muzycznych sprzedawanych w serwisie Amazon.
Dane pochodzą ze strony `http://jmcauley.ucsd.edu/data/amazon/`.

Załóżmy, że liczba recenzji instrumentów jest zmienną losową z rozkładu Poissona i poszczególne obserwacje są niezależne. Oczywiście w bazie danych nie ma obserwacji zerowych, więc należy zastosować rozkład Poissona obcięty w zerze. Wówczas funkcja wiarygodności ma postać:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i! * (1 - e^{-\lambda})}, \tag{5}$$

a jej logarytm:

$$\ln L(\lambda) = -n\lambda - n\ln\left(1 - e^{-\lambda}\right) + \ln(\lambda)\sum_{i=1}^{n} k_i - \sum_{i=1}^{n}\ln(k_i!). \tag{6}$$

Oszacować parametr λ MNW i przetestować hipotezę $\lambda = 6$.

Zadanie 4 (Zadanie 17 z [1] str. 162)
Zbiór danych `Quakes` z biblioteki `resampledata` zawiera informacje o czasie w dniach jaki upłynął między trzęsieniami ziemi o magnitudzie 6 lub większej. Zakładając, że czas między poszczególnymi trzęsieniami ziemi może być opisany rozkładem Weibulla oszacować parametry k oraz λ metodą największej wiarygodności. Zweryfikować hipotezę $H_0 : k = 1 \& \lambda = 20$ na poziomie istotności $\alpha = 0.05$.

$$L(k,\lambda;x) = \prod_{i=1}^{n} \frac{kx^{k-1}}{\lambda^k} \exp\left(-\frac{x_i}{\lambda}\right)^k = \frac{k^n}{\lambda^{kn}} \prod_{i=1}^{n} x_i^{k-1} \exp\left(-\sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right)^k\right) \tag{7}$$

$$\ln L(k,\lambda;x) = n\ln(k) - kn\ln(\lambda) + (k-1)\sum_{i=1}^{n}\ln(x_i) - \sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right)^k \tag{8}$$

Zadanie 5 (Na podstawie zadania 16 z [1] str. 162)
Nikt nie lubi stać w kolejkach, więc możliwość modelowania jest obiektem zainteresowania w badaniach dotyczących teorii kolejek. Zbiór danych `Service` z biblioteki `resampledata` zawiera informacje o czasie oczekiwania w minutach na obsłużenie dla 174 klientów. Zakładając niezależność obserwacji oraz to że pochodzą z rozkładu Gamma oszacować parametry k oraz θ metodą największej wiarygodności. Na poziomie istotności $\alpha = 0.05$ zweryfikować hipotezę $H_0 : k = 1 \ \& \ \theta = 2$.

$$\ln L(k,\theta;x) = (k-1)\sum_{i=1}^{n}\ln(x_i) - \sum_{i=1}^{n}\frac{x_i}{\theta} - nk\ln(\theta) - n\ln(\Gamma(k)) \tag{9}$$

Zadanie 6
Zbiór danych `contest_data.csv` pochodzi ze strony `https://www.kaggle.com/jaysobel/kcbs-bbq`.
This data set is the aggregate of 1,559 KCBS competitions from July 2013 through December 2016. The Kansas City Barbeque Society (KCBS) is "world's largest organization of barbeque and grilling enthusiasts with over 20,000 members worldwide." The data set was constructed by scraping the KCBS events page.

Wykorzystując niezerowe obserwacje dla zmiennej *prize* oszacować parametry rozkładu tej zmiennej - μ oraz σ metodą największej wiarygodności (zakładając rozkład normalny zmiennej *prize*).

$$\ln L(\mu,\sigma;x) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{10}$$

Zadanie 7
Z rozkładu $N(\mu, 2^2)$ wygenerowanych zostało 40 obserwacji:
0.62,-2.52,-0.31,-0.73,2.54,-1.52,-1.18,2.06,2.53,2.52,0.66,0.02,-0.93,-0.09,0.81 -1.60,2.70,0.52,1.75,-0.79, 3.66,-1.05,-1.32,-2.42,0.41,-2.09,2.67,1.36,0.94,0.58,-0.40,1.91,0.18,1.41,4.56,-0.01,-1.60,-0.07,1.79,2.23, 0.52,-2.81,-1.74,0.71,2.09,2.25,1.33,0.37,-2.04,2.29.
Oszacować parametr μ metodą największej wiarygodności. Przetestować statystyczną istotność parametru μ.

Zadanie 8
Rozważmy ponownie zbiór danych o nagrodach w konkursach barbecue. Nie ograniczając zbioru danych wyznaczyć parametry rozkładu normalnego nieprzyjmującego wartości ujemnych (w tym przypadku rozwiązania brzegowe).

$$X \sim N(\mu,\sigma^2) \Rightarrow f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\sum_{i=1}^{n}(x-\mu)^2}{2\sigma^2}\right)$$

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \mathbb{P}(X < x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad \Rightarrow \quad \mathbb{P}(X < 0) = \Phi\left(\frac{0-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu}{\sigma}\right)$$

Rozkład normalny nieprzyjmujący wartości ujemnych dla rozwiązań brzegowych ma funkcję gęstości:

$$f(x_i; \mu, \sigma) = \begin{cases} 1 - \Phi\left(\frac{\mu}{\sigma}\right) & \text{dla} \quad y_i = 0 \\ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) & \text{dla} \quad y_i > 0 \end{cases} \tag{11}$$

Funkcja $\Phi(.)$ jest dystrybuantą rozkładu standardowego normalnego i w programie R może być zapisana poleceniem `pnorm()`. W programowaniu funkcji wiarygodności:
`pnorm(mu/sigma, mean=0, sd=1)`.

Uwaga: Moglibyśmy rozważyć jeszcze inny rozkład: rozkład normalny obcięty w zerze, tj. taki, który określony jest tylko dla $x > 0$. Wówczas funkcja gęstości miałaby postać:

$$f(x; \mu, \sigma) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{1 - \Phi(0)} \tag{12}$$

2 Bibliografia

[1] Laura Chihara, Tim Hesterberg, *Mathematical Statistics with Resampling and R*, John Wiley&Sons, 2011.

