INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Analýza modelových úloh z vybraných aplikačních oblastí pomocí počítače

1 Formulace úlohy

Cílem této studie je posoudit funkci čističky odpadních vod s ohledem na přítomnost syntetických musk sloučenin ve vodě, která prošla touto čističkou. Proto bylo v řece Svratce poblíž čističky odpadních vod uloveno celkem 60 ryb - jelců (Leuciscus cephalus), přibližně stejného stáří.Z toho bylo 30 uloveno před čističkou (skupina 1) odpadních vod a 30 za čističkou odpadních vod (skupina 2). Potom bylo zjišťováno, jaké množství syntetických musk sloučenin (musk ambrette (AMB), musk tibetene (TIB)), phantolide (PH) a traseolide (TR)) bylo v jednotkovém množství svalů těchto ryb přítomno. Měřící přístroj měl dvě omezení mez kvantifikace a mez detekce (limit of detection LOD =x_D) a mez kvantifikace (limit of quantification LOQ=x_Q), $x_D < x_Q$. Takže data byla zleva dvojnásobně cenzorovaná časem (meze detekce a mez kvantifikace jsou v roli časového cenzoru) a pro každou skupinu byly znám počet cenzorovaných pozorování, tedy počet pozorovánín_D před mezi detekce, počet pozorování n_Qv intervalu $\langle x_D, x_Q\rangle$ a počet necenzorovaných pozorování n_0 a jejich konkrétní hodnoty. Získaná data lze najít v článku Fusek, Michálek, Zouhar, Vávrová: Statistical Analysis of Musk Compounds Concentrations in Fish Tissue Based on Doubly Left-Censored Samples. DEED 2013.

2 Model

Předpokládáme, že $X_1, \dots, X_n$ je náhodný výběr z exponenciálního rozdělení s parametrem λ a s hustotou $f(x,\lambda)$ a distribuční funkcí $F(x,\lambda)$. $X_{(1)}, \dots, X_{(n)}$ značí uspořádaný výběr.
Dvojité cenzorování s cenzory x_D a x_Q:
n_D je četnost cenzorovaných v $\langle 0, x_D\rangle$,
n_Q je četnost cenzorovaných v $(x_D, x_Q\rangle$,
n_O je četnost necenzorovaných, tedy větších než x_Q, $X_{(n-n_O+1)}, \dots, X_{(n)}$.
Vektor $(n_D, n_Q, n_O) \sim \text{Mu}_3(n, \theta_D, \theta_Q, \theta_O)$, $n = n_D + n_Q + n_O$ (multinomické rozdělení), a jednotlivá marginální rozdělení jsou tvaru
$n_D \sim \text{Bi}(n, \theta_D)$,
$n_Q \sim \text{Bi}(n, \theta_Q)$,
$n_O \sim \text{Bi}(n, \theta_O)$,
kde
$\theta_D = F(x_D, \lambda)$,
$\theta_Q = F(x_Q, \lambda) - F(x_D, \lambda)$,
$\theta_O = 1 - F(x_Q, \lambda)$.

3 Věrohodnostní rovnice a empirická Fisherova míra informace

Věrohodnostní funkce je tvaru

$$L(\lambda, n_D, n_Q, x_{(n-n_O+1)}, \dots, x_{(n)}) = \frac{n!}{n_D! n_Q!} \left[F(x_D,\lambda)\right]^{n_D} \left[F(x_Q,\lambda) - F(x_D,\lambda)\right]^{n_Q} \prod_{j=n-n_O+1}^{n} f(x_j).$$

Operační program Vzdělávání pro konkurenceschopnost
Název projektu: Inovace magisterského studijního programu Fakulty ekonomiky a managementu
Registrační číslo projektu: CZ.1.07/2.2.00/28.0326

PROJEKT JE SPOLUFINANCOVÁN EVROPSKÝM SOCIÁLNÍM FONDEM A STÁTNÍM ROZPOČTEM ČESKÉ REPUBLIKY.

Logaritmická věrohodnostní funkce je tvaru

$$l(\lambda, n_D, n_Q, x_{(n-n_O+1)}, \ldots, x_{(n)}) = \log\left(\frac{n!}{n_D! n_Q!}\right) + n_D \log\left[F(x_D, \lambda)\right] + n_Q \log\left[F(x_Q, \lambda) - F(x_D, \lambda)\right] + \sum_{j=n-n_O+1}^{n} \log\left[f(x_{(j)})\right].$$

$$\frac{\partial l}{\partial \lambda} = n_D \underbrace{\frac{F'_\lambda(x_D, \lambda)}{F(x_D, \lambda)}}_{H_D(x_D,\lambda)} + n_Q \underbrace{\frac{F'_\lambda(x_Q, \lambda) - F'_\lambda(x_D, \lambda)}{F(x_Q, \lambda) - F(x_D, \lambda)}}_{H_Q(x_Q,\lambda)} + \sum_{j=n-n_O+1}^{n} \frac{f'_\lambda(x_{(j)})}{f(x_{(j)})}$$

$$\frac{\partial^2 l}{\partial \lambda^2} = n_D \frac{\partial H_D(x_D, \lambda)}{\partial \lambda} + n_Q \frac{\partial H_Q(x_Q, \lambda)}{\partial \lambda} + \frac{\partial}{\partial \lambda} \sum_{j=n-n_O+1}^{n} \frac{f'_\lambda(x_{(j)})}{f(x_{(j)})}$$

4 Maximálně věrohodný odhad parametru ze předpokladu, že náhodný výběr je z $Ex(\lambda)$

Exponenciální rozdělení s distribuční funkcí $F(x) = 1 - \exp(-\lambda x)$ a hustotou $f(x) = \lambda \exp(-\lambda x)$. Logaritmická věrohodnostní funkce je tvaru

$$l(\lambda, n_D, n_Q, x_{(n-n_O+1)}, \ldots, x_{(n)}) = \log\left(\frac{n!}{n_D! n_Q!}\right) + n_D \log\left[1 - \exp(-\lambda x_D)\right] + n_Q \log\left[\exp(-\lambda x_D) - \exp(-\lambda x_Q)\right] + n_O \log(\lambda) - \lambda \sum_{i=n-n_O+1}^{n} x_{(i)}$$

$$\frac{\partial l}{\partial \lambda} = n_D \frac{x_D \exp(-\lambda x_D)}{1 - \exp(-\lambda x_D)} + n_Q \frac{x_Q \exp(-\lambda x_Q) - x_D \exp(-\lambda x_D)}{\exp(-\lambda x_D) - \exp(-\lambda x_Q)} + \frac{n_O}{\lambda} - \sum_{i=n-n_O+1}^{n} x_{(i)}$$

$$\frac{\partial^2 l}{\partial \lambda^2} = n_D \left\{ -\frac{x_D^2 \exp(-\lambda x_D)}{1 - \exp(-\lambda x_D)} - \frac{x_D^2 \exp(-2\lambda x_D)}{\left[1 - \exp(-\lambda x_D)\right]^2} \right\} + n_Q \left\{ \frac{x_D^2 \exp(-\lambda x_D) - x_Q^2 \exp(-\lambda x_Q)}{\exp(-\lambda x_D) - \exp(-\lambda x_Q)} - \frac{\left[x_Q \exp(-\lambda x_Q) - x_D \exp(-\lambda x_D)\right]^2}{\left[\exp(-\lambda x_D) - \exp(-\lambda x_Q)\right]^2} \right\} - \frac{n_O}{\lambda^2}$$

Empirická Fisherova informace je pak tvaru

$$J\text{empir} = -\frac{\partial^2 l}{\partial \lambda^2} = n_D \left\{ \frac{x_D^2 \exp(-\lambda x_D)}{1 - \exp(-\lambda x_D)} + \frac{x_D^2 \exp(-2\lambda x_D)}{\left[1 - \exp(-\lambda x_D)\right]^2} \right\} - n_Q \left\{ \frac{x_D^2 \exp(-\lambda x_D) - x_Q^2 \exp(-\lambda x_Q)}{\exp(-\lambda x_D) - \exp(-\lambda x_Q)} - \frac{\left[x_Q \exp(-\lambda x_Q) - x_D \exp(-\lambda x_D)\right]^2}{\left[\exp(-\lambda x_D) - \exp(-\lambda x_Q)\right]^2} \right\} + \frac{n_O}{\lambda^2}.$$

evropský sociální fond v ČR EVROPSKÁ UNIE MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost UNIVERZITA OBRANY
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Opět platí, že

$$\begin{aligned} \mathrm{E}(n_D) &= n\theta_D = n\left[1-\exp\left(-\lambda x_D\right)\right], \\ \mathrm{E}(n_Q) &= n\theta_Q = n\left[\exp\left(-\lambda x_D\right)-\exp\left(-\lambda x_Q\right)\right], \\ \mathrm{E}(n_O) &= n\theta_O = n\exp\left(-\lambda x_Q\right). \end{aligned}$$

Tedy teoretická Fisherova informace je tvaru

$$\begin{aligned} J\text{teor} = -\mathrm{E}\frac{\partial^2 l}{\partial\lambda^2} = n\frac{x_D^2\exp\left(-2\lambda x_D\right)}{\left[1-\exp\left(-\lambda x_D\right)\right]} + nx_Q^2\exp\left(-\lambda x_Q\right) \\ + n\frac{\left[x_Q\exp\left(-\lambda x_Q\right)-x_D\exp\left(-\lambda x_D\right)\right]^2}{\left[\exp\left(-\lambda x_D\right)-\exp\left(-\lambda x_Q\right)\right]} + \frac{n}{\lambda^2}\exp\left(-\lambda x_Q\right). \end{aligned}$$

5 Statistický test pro srovnání dvou cenzorovaných exponenciálních rozdělení

V předchozím odstavci byl nalezen maximálně věrohodný odhad parametru exponenciálního rozdělení $Ex(\lambda)$ a odhad jeho rozptylu. Protože maximálně věrohodné odhady mají asymptoticky normální rozdělení, lze statistický test pro posouzení funkce čističky založit na testování hypotézy, že parametry exponenciálního rozdělení, která popisují rozdělení dat před a za čističkou jsou stejné.Vyjdeme proto z nulové hypotézy H_0 : $\lambda_1 - \lambda_2 = 0$ proti alternativě $\mathrm{H}_1 : \lambda_1 - \lambda_2 \neq 0$, kde λ_1 aλ_2 jsou parametry exponenciálního rozdělení popisující data před a za čističkou. Testovací statistika Tje tvaru

$$\mathrm{T} = \frac{\widehat{\lambda}_1 - \widehat{\lambda}_2}{\sqrt{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}}, \tag{1}$$

kde $\widehat{\sigma}_1^2$ a $\widehat{\sigma}_2^2$ jsou odhadnuté rozptyly odhadů $\widehat{\lambda}_1$ a $\widehat{\lambda}_2$. Asymptotický rozptyl σ_k^2 je roven

$$\sigma_k^2 = J_k^{-1}, k = 1, 2, \tag{2}$$

kde J_k je Fisherova informační míra a lze ji stanovit pomocí druhé derivace logaritmické věrohodnostní funkce.

6 Výsledky statistického testu

Na základě použitého testu na hladině významnosti $\alpha = 0.05$ bylo možné rozhodnout, že není statistický významný rozdíl mezi skupinou skupinu 1 a skupinou 2 v musk sloučeninách Phantolide, Traseolide a Musk Ambrette. Statiticky významný rozdíl byl prokázán pro sloučeninu Musk Tibeten. Výsledky jsou v tabulce níže.

Sloučenina	H	p-hodnota
PH	0	0,47
TR	0	0,58
AMB	0	1,00
TIB	1	0,00

Tab. 1: Srovnání středních hodnot koncentrací musk sloučenin mezi skupinou 1 a 2. $\mathrm{H} = 0$ ($\mathrm{H} = 1$) značí, že nulová hypotéza nebyla zamítnuta (byla zamítnuta) na hladině významnosti 0,05.

Příklady k procvičení

1. Detailně projděte jednotlivé kroky při odvozování testovací statistiky T pro testování shody dvou zleva časem cenzorovaných výběrů. Pomocí software R nebo MATLAB si osvojte testy pro srovnání cenzorovaných výběrů z exponenciálního a Weibullova rozdělení na datech podle vaší volby.

2. Více než polovina zemětřesení v USA se vyskytuje na Aljašce. Konkrétně zemětřesení s magnitudou větší než 8 stupňů Richterovy stupnice se objevují v průměru každých 13 let. Předpokládejme, že tyto výskyty lze modelovat pomocí Poissonova procesu. Určete pravděpodobnost, že

   a) v deseti letech nebudou žádná zemětřesení,

   b) dvě po sobě jsoucí zemětřesení budou od sebe vzdáleny alespoň 5 let,

   c) během 13 let se vyskytne právě jedno zemětřesení,

   d) ve třech po sobě jsoucích dekádách bude právě jedno zemětřesení v každé dekádě.

3. Pojišťovna rozděluje své pojištěnce aut do tří kategorii podle rizika: vysoké střední a nízké. V daném roce nemá pojištěnec nehodu s pravděpodobností 0,6, jednu nehodu s pravděpodobností 0,2, dvě nehody s pravděpodobností 0,1 a více než dvě nehody s pravděpodobností 0,1. V případě, že pojištěnec nemá nehodu, posune se o jednu kategorii rizika níž, má-li jednu nehodu, zůstane ve stejné kategorii, pokud má dvě nehody, posune se o kategorii výš a jestliže má více než dvě nehody, přesune se vždy do kategorie s nejvyšším rizikem.

   a) Popište posloupnost pohybů mezi jednotlivými kategoriemi pojištěnců pomocí markovského řetězce.

   b) Jestliže začnete jako pojištěnec s nejnižším rizikem, jak dlouho lze očekávat, že v této kategorii zůstanete?

   c) Kolik let v průměru uplyne mezi dvěma po sobě jdoucími „návštěvami“ kategorie s nejvyšším rizikem?