10.1 Metoda maximální věrohodnosti

▷ **Příklad 10.1.** Predikce cen na burze:
Když se predikují ceny nějaké komodity na burze, postupuje se následovně: Vytvoří se model pro vývoj ceny dané komodity a na základě již známých dat se odhadnou parametry v tomto modelu. Tyto parametry se pak použijí pro odhad ceny v budoucnu. Jednou z možností jak parametry odhadnout je použití metody maximální věrohodnosti, která je funkcí parametrů modelu při daných fixních datech (cen komodity).
Nech X je náhodná veličina značící cenu jedné určité komodity na burze. Předpokládjme, že máme k dispozici n hodnot $\mathbf{x} = (x_1, \ldots, x_n)$ v n různých časech. Náš model bude velmi jednoduchý. Cena v čase t bude záviset na ceně v čase $t-1$ plus konstanta a tzv. bílý šum, tedy:

$$x_t = \mu(1-\rho) + \rho x_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \quad X_1 \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

Model obsahuje tři paremtry μ, ρ, σ, označme vektor těchto parametrů jako $\theta = (\mu, \rho, \sigma)$.
Potom věrohodnostní funkce pro daná data $\mathbf{x} = (x_2, \cdots, x_n)$ a libovolné parametry θ vypadá následovně:

$$L(\theta|\mathbf{x}) = f_\theta(\mathbf{x}) = f_\theta(x_2|x_1) \cdot f_\theta(x_3|x_2), \ldots, f_\theta(x_n|x_{n-1})$$

Podmíněná hustota $f_\theta(X_t|X_{t-1} = x_{t-1})$ patří do třídy normálních distribucí s parametry:

$$E[X_t|X_{t-1} = x_{t-1}] = \mu(1-\rho) + \rho x_{t-1} + E[\epsilon_t] = \mu(1-\rho) + \rho x_{t-1}$$

$$var(X_t|X_{t-1} = x_{t-1}) = var(\epsilon_t) = \sigma^2$$

Užitím předchozího dostáváme:

$$\begin{aligned}
L(\theta|\mathbf{x}) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x_2 - \mu(1-\rho) - \rho x_1)^2) \cdot, \ldots, \\
&\cdot \ldots, \cdot \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x_n - \mu(1-\rho) - \rho x_{n-1})^2) = \\
&= \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x_i - \mu(1-\rho) - \rho x_{i-1})^2) \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^{n-1} \cdot exp(\sum_{i=2}^{n} -\frac{1}{2\sigma^2}(x_i - \mu(1-\rho) - \rho x_{i-1})^2)
\end{aligned}$$

Z výpočetních důvodů použijeme logaritmus věrohodnostní funkce, což můžeme, protože logaritmus nějaké funkce má maximum ve stejném bodě jako funkce samotná.

$$\begin{aligned}
\ln(L(\theta|\mathbf{x})) &= \ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^{n-1} + \ln\Big(\exp\Big(\sum_{i=2}^{n} -\frac{1}{2\sigma^2}(x_i - \mu(1-\rho) - \rho x_{i-1})^2\Big)\Big) = \\
&= (n-1)\ln\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big) + \Big(\sum_{i=2}^{n} -\frac{1}{2\sigma^2}(x_i - \mu(1-\rho) - \rho x_{i-1})^2\Big)
\end{aligned}$$

Příslušný kód pak v R může vypadat následovně:

```
likelihood=function(mi,ro,sigma,x){
   n=length(x)
   xi=x[2:n]
   xim=x[1:(n-1)]
   si=sum((-1/(2*sigma^2))*((xi-mi*(1-ro)-ro*xim)^2))
   return(n*log(1/(sqrt(2*pi)*sigma))+si)
}
```

Jak již bylo zmíněno, věrohodností funkce je funkcí parametrů, nikoliv dat.
Zvolte si paramtery μ, ρ, σ a vygenerujte dle modelu v ronici 1 data (zvolte alespoň $n = 1000$). Následně ověřte, že maximalizací funkce likelihood dostanete dobrý odhad zvolených parametrů.

► **Příklad 10.1.** Víte, že výběr 12, 11.2, 13.5,12.3,13.8,11.9 pochází z rozdělení s hustotou

$$f(x;\phi) = \begin{cases} \frac{\phi}{x^{\phi+1}} & x > 1 \\ 0 & \text{jinak} \end{cases}$$

Metodou maximální věrohodnosti stanovte parametr ϕ. [0.397]

10.2 Bodové odhady a intervaly spolehlivosti

► **Příklad 10.2.** Uvažujme náhodnou veličinu X, která má binomické rozdělení s parametry n, p. Ukažte, že $\widehat{p} = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$ je tzv. asymptoticky nestranný odhad parametru p, tj. pro $n \to \infty$ je $\widehat{p}$ nestraným odhadem p.

► **Příklad 10.3.** Nechť 3.9, 2.8, 5.5, 2.4, 8.6 je výběr z Gamma rozdělení se známým parametrem $\alpha = 4$. Užitím metody momentů určete bodový odhad pro neznámý parametr θ. [1.16]

▷ **Příklad 10.2.** Délka životnosti žárovek vyráběných určitou firmou má normální rozdělení, se směrodatnou dochylkou $\sigma = 40\,h$. Náhodně vybereme 30 žárovek a určime jejich živostnost $x_1, \ldots, x_{30}$. Průměrná životnost v tomto výběru byla $\overline{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = 780$ hodin. Určete 95 % interval spolehlivosti pro střední hodnotu životnosti žárovek vyráběných touto firmou.
Protože víme, že náhodná veličina $W = \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim$ má normální rozdělení $\mathcal{N}(0,1)$, můžeme psát:

$$P\Big(\Big\{ - z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \leq z_{\alpha/2}\Big\}\Big) = 1-\alpha$$

Po úpravě dostáváme:

$$P\Big(\Big\{\overline{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\Big\}\Big) = 1-\alpha$$

V našem případě je $\alpha = 0.05$, horní 2.5 % percentil $z_{\alpha/2}$ lze najít ve statistických tabulkách, nebo pomocí kvantilové funkce v R, kde parametr `lower.tail` nastavíme na `FALSE`. Můžeme samozřejmě využít vlastnosti distribuční funkce $\Phi(z_p) = 1 - p$.

```
> qnorm(0.025,lower.tail=FALSE)
[1] 1.959964
> qnorm(1-0.025)
[1] 1.959964
```

Po dosazení dostaneme 95 % interval spolehlivosti pro průměrnou životnost všech žárovek vyráběných danou firmu jako interval $(765.69, 794.31)$.

Jak velký musí být výběr, chceme-li, aby 95 % interval spolehlivosti byl $(770, 790)$?
Abychom interval spolehlivost zúžili, je potřeba zvětšit výběr, tj. vybrat více žárovek a prozkoumat jejich životnost. Předpokládejme, že jsme vybrali n žárovek a že průměrná životnost tohoto výběru je stejná $\overline{x} = 780$. Chceme nyní určit n tak aby $P(\{\overline{X} - 10 \leq \mu \leq \overline{X} + 10\}) = 0.95$, tedy: $z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} = 10$. Po úpravě a dosazení dostáváme $n = 61.46$, či-li museli bychom vybrat 62 žárovek.

► **Příklad 10.4.** Vraťme se k předchozímu příkladu a předpokládejme nyní, že směrodatná odchylka pro životnost žárovek není známá. Ovšem z výběru 30 žárovek jsme spočítali výběrovou směrodatnou odchylku jako $S = 40$ hodin. Určete nyní 95 % interval spolehlivosti pro střední hodnotu μ pomocí náhodné veličiny mající t-rozdělení. [(765.06,794.94)]

► **Příklad 10.5.** Na řece se měří koncentrace zinku. Náhodně se vybralo 36 lokací, průměrná koncentrace byla 2.6 gramu na mililitr vody. Předpokládejte, že standardní odchylka koncentrace zinku v řece je $\sigma = 0.3$. Určete 95 % a 99 % interval spolehlivosti pro střední hodnotu koncentrace zinku v řece. $[(2.50, 2.70); (2.47, 2.73)]$

► **Příklad 10.6.** Dvě aerolinky A,B odlétají ve stejný čas ze stejného letiště do stejné destinace a soupeří tak o pasažéry. Předpokládejte, že za 1 den obě společnosti dohromady přepraví 1000 pasažérů a že pasažéři si přepravní společnost vybírají nezávisle jeden na druhém, každý s pravděpodobností $p = 1/2$. Aerolinka A chce mít jistotu, že přepraví všechny pasažéry, kteří o její služby projeví zájem a nepřenechá tak nic konkurenci. To by ovšem znamenalo nasadit na trasu letadlo s 1000 místy, což je zbytečně nákladné a pravděpodobnost, že si všichni pasažéři v jeden den zvolí jenom aerolinku A je téměř nulová (kolik přesně?). Co byste manažerům z aerolinek A poradili? Kolika místné letadlo mají nasadit?

► **Příklad 10.7.** Chceme určit procentuální zastoupení kuřáků v České Republice. Náhodně vybereme n lidí. Jak velý musí být výběr, abychom p určili s přesností na jedno desetinné místo? Zvolte hladinu spolehlivosti 95 %. [385]

► **Příklad 10.8.** V nemocnici mají dva nové druhy antibiotik. První druh je vyzkoušen na 14 pacientech, jejich průměrná doba léčení byla 17 dnů. Druhý druh byl podán 16ti pacientům, jejich průměrná doba léčerní byla 19 dnů. Pro jednoduchost předpokládejme, že $\sigma_1 = 1.5$ a $\sigma_2 = 1.8$ jsou známé (ve skutečnosti bychom museli spočítat výběrové směrodatné odchylky). Určete 95 % interval spolehlivosti pro rozdíl středních hodnot, za předpokladu, že doba léčení u obou antibiotik má normální rozdělení.

Reference

[1] Kerns G. J.: Elementary Probability on Finite Sample Spaces, 2009, reference manual package prob, available from: http://CRAN.R-project.org/package=prob

[2] Kerns G. J: Introduction to Probability and Statistics Using R, First Edition http://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf

[3] Kenneth Baclawski: Introduction to Probability with R *Chapman and Hall/CRC*, ISBN 978-1420065213.

[4] Walpole R. E, Myers R., Myers S, Ye K. : Probability & Statistics for Engineers & Scientists *Prentice Hall*, ISBN:0-13-098469-8