Príklady k postačujúcej štatistike

Príklad 1.
Uvažujme náhodný vektor $(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)^T$, kde $\tilde{y}_1, \tilde{y}_2, \tilde{y}_3 \sim Geo(\theta)$ a $\tilde{y}_1, \tilde{y}_2, \tilde{y}_3$ sú nezávislé, t.j. každé z nich má rozdelenie $f(y_i|\theta) = (1-\theta)^{y_i}\theta$, $y_i \in Y = \{0, 1, 2, \ldots\}$. Rozdelenie náhodného vektora $(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)^T$ je teda (tým, že sú nezávislé)

$$f(y_1, y_2, y_3|\theta) = (1-\theta)^{y_1}\theta(1-\theta)^{y_2}\theta(1-\theta)^{y_3}\theta = (1-\theta)^{y_1+y_2+y_3}\theta^3,$$

$(y_1, y_2, y_3)^T \in \{0, 1, 2, \ldots\}^3$ (všetky možné trojice nezáporných celých čísel). To môžeme zapísať aj ako

$$f(y_1, y_2, y_3|\theta) = (1-\theta)^{\tau(y_1,y_2,y_3)}\theta^3 = h(y_1, y_2, y_3)q(\tau(y_1, y_2, y_3), \theta),$$

kde $\tau(y_1, y_2, y_3) = y_1 + y_2 + y_3$, $h(y_1, y_2, y_3) = 1$ a $q(t, \theta) = (1-\theta)^t\theta^3$, čiže z faktorizačnej vety platí, že $\tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3) = \tilde{y}_1 + \tilde{y}_2 + \tilde{y}_3$ je postačujúca štatistika pre θ.

Poďme teraz nájsť rozdelenie postačujúcej štatistiky $\tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)$ (budeme ho označovať $f^\star(\tau(y_1, y_2, y_3)|\theta)$):

- $\tau(y_1, y_2, y_3)$ dostaneme ako súčet troch nezáporných celých čísel, čo môže byť opäť iba nezáporné celé číslo - zaujímajú nás teda iba hodnoty $0, 1, 2, \ldots$.

- $f^\star(0|\theta) = P(\tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3) = 0|\theta) = P(\tilde{y}_1 + \tilde{y}_2 + \tilde{y}_3 = 0|\theta) =$
  $= P(\tilde{y}_1 = 0, \tilde{y}_2 = 0, \tilde{y}_3 = 0|\theta) = f(0, 0, 0|\theta) = \theta^3,$

- $f^\star(1|\theta) = P(\tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3) = 1|\theta) = P(\tilde{y}_1 + \tilde{y}_2 + \tilde{y}_3 = 1|\theta) =$
  $= f(1, 0, 0|\theta) + f(0, 1, 0|\theta) + f(0, 0, 1|\theta) = 3(1-\theta)\theta^3,$

- $f^\star(2|\theta) = f(2, 0, 0|\theta) + f(0, 2, 0|\theta) + f(0, 0, 2|\theta)$
  $+ f(1, 1, 0|\theta) + f(1, 0, 1|\theta) + f(0, 1, 1|\theta) = 6(1-\theta)^2\theta^3.$

  (všimnime si, že tento dlhý súčet vieme zapísať aj ako

  $$f^\star(2|\theta) = \sum_{(y_1,y_2,y_3):\tau(y_1,y_2,y_3)=2} f(y_1, y_2, y_3|\theta)$$

  - porovnaj s dôkazom z prednášky, prípadne z učebnice),

- analogicky pre ľubovoľné $t \in \{0, 1, 2, \ldots\}$

$$f^\star(t|\theta) = \sum_{(y_1,y_2,y_3):\tau(y_1,y_2,y_3)=t} f(y_1, y_2, y_3|\theta) = \binom{t+2}{t}(1-\theta)^t\theta^3$$

(toto vo všeobecnom prípade nemusí byť ľahko vypočítateľné - v tomto konkrétnom sa to dá kombinatoricky, resp. s využitím toho, že súčet nezávislých geometrických rozdelení má negatívne binomické rozdelenie; v dôkaze z prednášký nám ale stačí iba suma

$$f^\star(t|\theta) = \sum_{(y_1,y_2,y_3):\tau(y_1,y_2,y_3)=t} f(y_1,y_2,y_3|\theta)$$

a jej následný rozklad na

$$\begin{aligned} f^\star(t|\theta) &= \sum_{(y_1,y_2,y_3):\tau(y_1,y_2,y_3)=t} h(y_1,y_2,y_3)q(\tau(y_1,y_2,y_3),\theta) = \\ &= q(t,\theta) \sum_{(y_1,y_2,y_3):\tau(y_1,y_2,y_3)=t} h(y_1,y_2,y_3). \end{aligned}$$

a nepotrebujeme konkrétne ,,vyčíslenie").

Predpokladajme teraz, že θ má nejaké rozdelenie $\pi(\theta)$. Potom ak pozorujeme celú trojicu (y_1,y_2,y_3), tak dostaneme

$$\pi(\theta|y_1,y_2,y_3) = \frac{\pi(\theta)f(y_1,y_2,y_3)}{\int_\Theta \pi(\theta)f(y_1,y_2,y_3)\mathrm{d}\theta} = \frac{\pi(\theta)(1-\theta)^{y_1+y_2+y_3}\theta^3}{\int_\Theta \pi(\theta)(1-\theta)^{y_1+y_2+y_3}\theta^3\mathrm{d}\theta}.$$

Ak by sme pozorovali iba $\tau(y_1,y_2,y_3) = y_1+y_2+y_3$, tak dostaneme

$$\begin{aligned} \pi^\star(\theta|\tau(y_1,y_2,y_3)) &= \frac{\pi(\theta)f^\star(\tau(y_1,y_2,y_3))}{\int_\Theta \pi(\theta)f^\star(\tau(y_1,y_2,y_3))\mathrm{d}\theta} = \\ &= \frac{\pi(\theta)\binom{\tau(y_1,y_2,y_3)+2}{\tau(y_1,y_2,y_3)}(1-\theta)^{\tau(y_1,y_2,y_3)}\theta^3}{\int_\Theta \pi(\theta)\binom{\tau(y_1,y_2,y_3)+2}{\tau(y_1,y_2,y_3)}(1-\theta)^{\tau(y_1,y_2,y_3)}\theta^3\mathrm{d}\theta} = \\ &= \frac{\pi(\theta)(1-\theta)^{y_1+y_2+y_3}\theta^3}{\int_\Theta \pi(\theta)(1-\theta)^{y_1+y_2+y_3}\theta^3\mathrm{d}\theta}, \end{aligned}$$

teda rovnakú ,,novú informáciu o θ" ako keď sme museli pozorovať celú trojicu (y_1,y_2,y_3).

Poznámka. Na tomto príklade možno ilustrovať aj definíciu postačujúcej štatistiky, t.j. že rozdelenie $(\tilde{y}_1,\tilde{y}_2,\tilde{y}_3)|(\theta,\tau(\tilde{y}_1,\tilde{y}_2,\tilde{y}_3)=t)$ nezávisí od θ. Pre ľubovoľné zmysluplné k_1,k_2,k_3 a t počítajme
$P(\tilde{y}_1=k_1,\tilde{y}_2=k_2,\tilde{y}_3=k_3|\theta,\tau(\tilde{y}_1,\tilde{y}_2,\tilde{y}_3)=t)$

$$= \frac{P(\tilde{y}_1=k_1,\tilde{y}_2=k_2,\tilde{y}_3=k_3,\tau(\tilde{y}_1,\tilde{y}_2,\tilde{y}_3)=t|\theta)}{P(\tau(\tilde{y}_1,\tilde{y}_2,\tilde{y}_3)=t)|\theta)}.$$

Výraz v čitateli je nenulový iba ak $k_1 + k_2 + k_3 = t$ - v takom prípade môžeme časť $\tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3) = t$ vyhodiť, pretože vyplýva z predchádzajúcich troch ($\tilde{y}_1 = k_1, \tilde{y}_2 = k_2, \tilde{y}_3 = k_3 \Rightarrow \tilde{y}_1 + \tilde{y}_2 + \tilde{y}_3 = t$). Máme teda

$$\frac{P(\tilde{y}_1 = k_1, \tilde{y}_2 = k_2, \tilde{y}_3 = k_3, \tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3) = t|\theta)}{P(\tau(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3) = t)|\theta} =$$

$$= \frac{f(k_1, k_2, k_3|\theta)}{f^\star(t|\theta)} = \frac{\theta^{k_1+k_2+k_3}(1-\theta)^3}{\binom{t+2}{2}\theta^t(1-\theta)^3} = \frac{1}{\binom{t+2}{t}},$$

čo nezávisí od θ.

Príklad 2. (tu nie sú ovlnovkované náhodné premenné, ale malo by byť pomerne jasné, čo je čo).
Uvažujme náhodný vektor $(y_1, y_2)^T$, kde $y_1, y_2 \sim N(\theta, \sigma_y^2)$ a sú nezávislé, t.j

$$(y_1, y_2)^T \sim N\left(\begin{pmatrix}\theta\\ \theta\end{pmatrix}, \begin{pmatrix}\sigma_y^2 & 0\\ 0 & \sigma_y^2\end{pmatrix}\right)$$

O θ budeme predpokladať, že $\theta \sim N(\mu, \sigma^2)$, kde μ, σ^2 a podobne aj σ_y^2 sú známe.

Bude nás zaujímať aposteriórne rozdelenie θ pri pozorovanom náhodnom vektore $(y_1, y_2)^T$ ($\theta|(y_1, y_2) \sim$?). Keďže poznáme aj hustotu $\pi(\theta)$ a aj hustotu $f(y_1, y_2|\theta)$, vedeli by sme aposteriórne rozdelenie vypočítať priamo pomocou Bayesovho vzorca, avšak to by bolo zdĺhavé preto miesto toho využijeme iný postup (ktorý bude v konečnom dôsledku zrejme dlhší, ale zato kreatívnejší). Konkrétne použijeme

1. pravidlo reťazenia (to môžeme, lebo y_1 a y_2 sú nezávislé), t.j najprv zistíme rozdelenie $\theta|y_1$ a to použijeme ako nový prior (označíme ho θ_{y_1}); následne vypočítame posterior $\theta_{y_1}|y_2$, ktorý bude mať podľa pravidla reťazenia rovnaké rozdelenie ako $\theta|(y_1, y_2)$;

2. výsledok Príkladu 1.1 (robili sme aj na hodine), ktorý hovorí, že ak $\theta \sim N(\mu, \sigma^2)$ a $y_1|\theta \sim N(\theta, \sigma_y^2)$, potom $\theta_{y_1} = \theta|y_1 \sim N(\mu_{nove}, \sigma_{nove}^2)$, kde

$$\begin{aligned}\mu_{nove} &= \frac{y_1\sigma^2 + \mu\sigma_y^2}{\sigma^2 + \sigma_y^2}\\ \sigma_{nove}^2 &= \frac{\sigma^2\sigma_y^2}{\sigma^2 + \sigma_y^2}\end{aligned}$$

(pozor, je tu teraz trochu iné označenie ako v učebnici/na hodine - kvôli tomu, aby nevznikol chaos v indexoch - pointa je ale stále tá istá).

Máme teda $\theta_{y_1} \sim N(\mu_{nove}, \sigma^2_{nove})$ a opäť s využitím Príkladu 1.1 dostaneme, že

$$\theta_{y_1}|y_2 \sim N\left(\frac{y_2\sigma^2_{nove} + \mu_{nove}\sigma^2_y}{\sigma^2_{nove} + \sigma^2_y}, \frac{\sigma^2_{nove}\sigma^2_y}{\sigma^2_{nove} + \sigma^2_y}\right).$$

Pre prehľadnosť vypočítajme jednotlivo čitatele a menovatele:

- čitateľ v strednej hodnote:

$$\begin{aligned} y_2\sigma^2_{nove} + \mu_{nove}\sigma^2_y &= y_2\frac{\sigma^2\sigma^2_y}{\sigma^2 + \sigma^2_y} + \frac{y_1\sigma^2 + \mu\sigma^2_y}{\sigma^2 + \sigma^2_y}\sigma^2_y = \\ &= \frac{y_2\sigma^2\sigma^2_y + y_1\sigma^2\sigma^2_y + \mu\sigma^2_y\sigma^2_y}{\sigma^2 + \sigma^2_y} = \\ &= \frac{(y_1 + y_2)\sigma^2\sigma^2_y + \mu\sigma^2_y\sigma^2_y}{\sigma^2 + \sigma^2_y}, \end{aligned}$$

- čitateľ v disperzii:

$$\sigma^2_{nove}\sigma^2_y = \frac{\sigma^2\sigma^2_y}{\sigma^2 + \sigma^2_y}\sigma^2_y = \frac{\sigma^2\sigma^2_y\sigma^2_y}{\sigma^2 + \sigma^2_y},$$

- menovateľ (v obidvoch rovnaký):

$$\sigma^2_{nove} + \sigma^2_y = \frac{\sigma^2\sigma^2_y}{\sigma^2 + \sigma^2_y} + \sigma^2_y = \frac{\sigma^2\sigma^2_y}{\sigma^2 + \sigma^2_y} + \frac{\sigma^2_y\sigma^2 + \sigma^2_y\sigma^2_y}{\sigma^2 + \sigma^2_y} = \frac{2\sigma^2\sigma^2_y + \sigma^2_y\sigma^2_y}{\sigma^2 + \sigma^2_y}.$$

Po dosadení dostaneme, že stredná hodnota $\theta_{y_1}|y_2$ je

$$\frac{\frac{(y_1+y_2)\sigma^2\sigma^2_y + \mu\sigma^2_y\sigma^2_y}{\sigma^2+\sigma^2_y}}{\frac{2\sigma^2\sigma^2_y + \sigma^2_y\sigma^2_y}{\sigma^2+\sigma^2_y}} = \frac{(y_1 + y_2)\sigma^2\sigma^2_y + \mu\sigma^2_y\sigma^2_y}{2\sigma^2\sigma^2_y + \sigma^2_y\sigma^2_y} = \frac{(y_1 + y_2)\sigma^2 + \mu\sigma^2_y}{2\sigma^2 + \sigma^2_y}$$

a disperzia je

$$\frac{\frac{\sigma^2\sigma^2_y\sigma^2_y}{\sigma^2+\sigma^2_y}}{\frac{2\sigma^2\sigma^2_y + \sigma^2_y\sigma^2_y}{\sigma^2+\sigma^2_y}} = \frac{\sigma^2\sigma^2_y\sigma^2_y}{2\sigma^2\sigma^2_y + \sigma^2_y\sigma^2_y} = \frac{\sigma^2\sigma^2_y}{2\sigma^2 + \sigma^2_y}.$$

Finálny výsledok je teda

$$\theta|(y_1, y_2) \sim N\left(\frac{(y_1 + y_2)\sigma^2 + \mu\sigma^2_y}{2\sigma^2 + \sigma^2_y}, \frac{\sigma^2\sigma^2_y}{2\sigma^2 + \sigma^2_y}\right).$$

V praxi je ale na takúto situáciu jednoduchšie využiť vetu o postačujúcej štatistike. Z faktorizačnej vety nie je ťažké sa presvedčiť, že ak

$$(y_1, y_2)^T \sim N\left(\begin{pmatrix}\theta\\ \theta\end{pmatrix}, \begin{pmatrix}\sigma_y^2 & 0\\ 0 & \sigma_y^2\end{pmatrix}\right),$$

tak $\tau(y_1, y_2) = \frac{y_1+y_2}{2}$ je postačujúca štatistika pre parameter θ. Rovnako ľahko (ba dokonca o dosť ľahšie) vidno, že $\tau(y_1, y_2)|\theta \sim N(\theta, \frac{\sigma_y^2}{2})$ (lebo je to priemer nezávislých normálnych rozdelení). Veta o postačujúcej štatistike mi hovorí, že rozdelenie $\theta|(y_1, y_2)$ je rovnaké ako rozdelenie $\theta|\tau(y_1, y_2)$ (teda nepotrebujeme dve čísla (y_1, y_2), ale stačí nám jedno $(\frac{y_1+y_2}{2})$). Aplikáciou Príkladu 1.1 dostaneme, že

$$\theta|\tau(y_1, y_2) \sim N\left(\frac{\tau(y_1, y_2)\sigma^2 + \mu\frac{\sigma_y^2}{2}}{\sigma^2 + \frac{\sigma_y^2}{2}}, \frac{\sigma^2\frac{\sigma_y^2}{2}}{\sigma^2 + \frac{\sigma_y^2}{2}}\right),$$

z čoho po jednoduchej úprave (vynásobením oboch zlomkov jednotkou v tvare $\frac{2}{2}$) dostaneme

$$\theta|\tau(y_1, y_2) \sim N\left(\frac{(y_1 + y_2)\sigma^2 + \mu\sigma_y^2}{2\sigma^2 + \sigma_y^2}, \frac{\sigma^2\sigma_y^2}{2\sigma^2 + \sigma_y^2}\right),$$

čo vidíme, že je rovnaké rozdelenie ako $\theta|(y_1, y_2)$. Teda ak ovládame vetu o postačujúcej štatistike, tak toto je určite najrýchlejší postup, navyše ľahko zovšeobecniteľný na prípad $(y_1, \ldots, y_n)^T$.