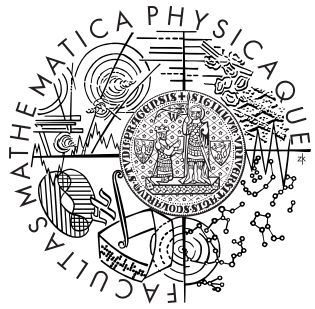
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Agáta Eštóková

Postačující statistiky a exponenciální rozdělení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Shokirov Bobosharif
Studijní program: Matematika, Finanční a pojistná matematika

2010

Především bych se chtěla poděkovat mému vedoucímu Mgr. Shokirovi Bobosharifovi za cenné rady a připomínky při vedení této práce. Velké poděkování patří mé spolubydlíci, Vere Djordjilovičové. V neposledí řadě chci poděkovat Petrovi Huszárovi a Szabolcsovi Cséfalvaymu za jazykovou korekturu a technické připomínky. Na závěr nesmím zpomenout ani na své rodiče, jím jsem vděčná za umožnění studia.

Prohlašuji, že jsem svou bakalářskou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne Agáta Eštóková

Obsah

Úvod **6**

1 Základní definice a věty **8**
1.1 Základní definice a pojmy 8
1.2 Halmosova a Savageova faktorizační věta 9

2 Exponenciální rodina rozdělení **15**
2.1 Jednorozměrné rozdělení 15
2.2 Exponenciální rodina rozdělení 20
2.2.1 Exponenciální rodina s parametrem polohy $F(x-\theta)$ 23
2.2.2 Exponenciální rodina s parametrem měřítka $F(\frac{x}{\sigma})$ . . 25

3 Příklady **27**

Závěr **33**

Literatura **34**

Název práce: Postačující statistiky a exponencialní rozdelení
Autor: Agáta Eštóková
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Shokirov Bobosharif
e-mail vedoucího: bobosari@karlin.mff.cuni.cz

Abstrakt: Jedna ze základných problému matematické statistiky představuje určení neznámého rozdělení nebo odhad neznámého parametru z předepsané rodiny rozdělení s využitím výsledku nezávislých pozorování. Je známé, že v mnoha případech řešení tohoto problému nevyžaduje znalost všech výsledků, nýbrž pouze jejich určitou funkci. Způsob konstrukce takových funkcí, které poskytují postačující materiál pro řešení tohoto problému, přinesl ideu postačujících statistik. V uvedené práci definujeme postačující statistiky a prostudujeme kritérium postačitelnosti, známé jako Fisherovo-Neymannovo faktorizční kritérium zobecněné Halmosem a Savagem. Prozkoumáme rodinu jednorozměrných rozdělení, jejichž určité mocniny připouštějí postačující statistiky. Dále se podíváme na minimální a netriviální postačující statistiky a ukážeme, že mezi regulárními rodinami rozdělení pouze exponenciální rodina hustot má tu vlastnot, že její určitá mocnina vyhovuje faktorizačnímu kritériu. Nezbytná část práce je věnována výpočtu postačujících statistik pro některé rodiny rozdělení. Na závěr přikládáme vhodné příklady.

Klíčová slova: statistika, postačující statistika, minimální postačující statistika, faktorizační věta, rodina exponenciálních rozdělení.

Title: Sufficient statistics and exponential family of distributions
Author: Agáta Eštóková
Department: Department of Probability and Mathematical Statistics
Supervisor: Shokirov Bobosharif
Supervisor's e-mail address: bobosari@karlin.mff.cuni.cz

Abstract: One of the fundamental problems of mathematical statistics is the determination of an unknown probability distribution or estimation of the unknown parameter(s) from some prescribed family of distributions by using the results of independent observations. It is known that in many cases, solving this problem does not require necessarily the complete knowledge of the results themselves but only certain functions of them.

The way of constructing such functions, which provide enough material for solving these problems, brought the idea of sufficient statistics. In the presented work, we give the definition of sufficient statistics and study the criteria of sufficiency, known as the Fisher-Neymann factorization theorem generalized by Halmos and Savage. We study the family of one-dimensional distributions, some power of which allows the sufficient statistics. Further, we look at minimal and nontrivial sufficient statistics and show that among the regular family of distributions only exponential type of densities have the property that some power of them satisfies factorization criteria. Essential part of the work is devoted to the calculation of sufficient statistics for some family of distributions. At the end, appropriate examples are given.

Keywords: statistics, sufficient statistics, minimal sufficient statistics, factorization theorem, exponential family.

Úvod

Postačující statistiky mají důležitý význam v matematické statistice, zejména v oblasti teorií odhadu. Uvažujme takovou statistiku, která obsahuje všechny informace, které lze z daného náhodného výběru $X_1, ..., X_n$ vzhledem k odhadovanému parametru získat. Statistika s touto vlastností se nazývá postačující.

Idea postačující statistiky vznikla jako důsledek základního problému matematické statistiky, při němž chceme určit neznámé rozdělení a jeho parametry ze znalosti realizací n nezávislých pokusů. Již dříve bylo objeveno, že řešení tohoto problému nevyžaduje nutně znalost výsledku jednotlivých pokusů, pouze jejich určité funkce.

Je žrejmé, že samotný náhodný výběr $X_1, ..., X_n$ je postačující statistikou. Taková postačující statistika se nazývá triviální. V reálných případech se ovšem zajímáme o postačující statistiky, které jsou dány menším počtem složek, o tzv. netriviální postačující statistiky. Pokud lze dané rozdělení napsat jako součin dvou funkcí, kde jedna závisí pouze na náhodném výběru, a druhá na určitém parametru (popřípadě na náhodném výběru také), bude mít toto rozdělení vzhledem k tomuto parametru netriviální postačující statistiku. Ukážeme, že takové rozdělení musí patřit do tzv. rodiny exponenciálních rozdělení. Mezi různými rodinami rozdělení pouze tato rodina má tu vlastnost, že k ní existuje postačující statistika, počet jejíž složek zůstává omezený při rostoucím rozsahu výběru.

V první kapitole uvádíme základnou definici statistiky, postačující a minimální postačující statistiky, které jsou nezbytné ke konstrukci dalších tvrzení. Zároveň dokážeme jednoduché kritérium pro nalezení postačující statisiky.

Druhá kapitola se zaměřuje na jednorozměrné rodiny rozdělení. Ukážeme v ní, že mezi rodinami rozdělení, jejichž nosič nezávisí na parametru, pouze v exponenciální rodině rozdělení existuje postačující statistika, jejíž rozměr s rostoucím rozsahem výběru zůstane omezený. Podmínka postačitelnosti

ostře omezuje možné formy rozdělení. Ve druhé kapitole tedy popisujeme dvě významné typy exponenciálních rodin: $F(x-\theta)$ s parametrem polohy a $1/\sigma F(1/\sigma)$ s parametrem měřítka. V těchto speciálních případech upřesníme tvar exponenciální rodiny.

Příklady aplikace výše uvedené teorie obsahuje kapitola třetí.

Kapitola 1

Základní definice a věty

V této kapitole uvedeme základní definice a věty, které potřebujeme ke konstrukci hlubších tvrzení. Pokud to není uvedeno jinak, definice a tvrzení jsou z [2].

Nechť $(\Omega, \mathcal{A}, P_\theta)$ je pravděpodobnostní prostor, kde Ω je neprázdná množina, $\mathcal{A}$ je σ-algebra podmnožin Ω a P_θ je pravděpodobnostní míra, která patří do rodiny $\{P_\theta, \theta \in \Theta\}$, kde Θ je parametrický prostor. Nechť $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{S})$ je náhodná veličina, kde $\mathcal{S}$ je σ-algebra podmnožin $\mathcal{X}$. Nechť $\sigma(X) = \sigma([X \in B], B \in \mathcal{S})$ je σ-algebra generována náhodní veličinou X. Potom výběrový prostor $\mathcal{X}$ je množina hodnot, kterých může nabývat náhodná veličina X. Nechť $n \in N$ a označme $\mathcal{X}^n = \times_{k=1}^n \mathcal{X}_k$, kde $\mathcal{X}_k := \mathcal{X}$ kartézsky součin a $\mathcal{S}^n = \otimes_{k=1}^n \mathcal{S}_k$, kde $\mathcal{S}_k := \mathcal{S}$ součinovou σ-algebru. Vektor $\mathbb{X} = (X_1, \ldots, X_n) : (\Omega, \mathcal{A}) \to (\mathcal{X}^n, \mathcal{S}^n)$ se nazývá náhodný vektor.

1.1 Základní definice a pojmy

Definice 1.1.1. Měřitelné zobrazení $T : (\mathcal{X}^n, \mathcal{S}^n) \to (R^m, \mathcal{B}^m)$, kde $m \leq n$ a $\mathcal{B}^m$ je množina borelovských podmnožin R^m, se nazýva statistika.

Definice 1.1.2. Statistika $S : (\mathcal{X}^n, \mathcal{S}^n) \to (R^m, \mathcal{B}^m)$ je postačující pro rodinu $\{P_\theta, \theta \in \Theta\}$, pokud pro každé $A \in \mathcal{S}^n$ existuje funkce $\psi_A = \psi_A[S(\mathbb{X})]$ taková, že pro každé $\theta \in \Theta$ platí

$$P_\theta(A|S) = \psi_A, \quad P_\theta\text{-s.j.} \tag{1.1.1}$$

Podmínka (1.1.1) říká, že podmíněná pravděpodobnost libovolného jevu za podmínky S je nezávislá na parametru θ. Pro rodinu $\{P_\theta, \theta \in \Theta\}$ mohou existovat různé postačující statistiky. My se snažíme najít postačující

statistiku s co nejmenším počtem složek tak, aby se přitom neztratila žádná informace o parametru θ. K tomu účelu definujeme minimální postačující statistiku.

Definice 1.1.3. Postačující statistika $T : (\mathcal{X}^n, \mathcal{S}^n) \to \left(R^l, \mathcal{B}^l\right)$ se nazývá minimální pro rodinu $\{P_\theta, \theta \in \Theta\}$, pokud T je funkcí kterékoliv jiné postačující statistiky.

To znamená, že statistika T je minimální pro $\{P_\theta, \theta \in \Theta\}$, pokud pro libovolnou postačující statistiku $S : (\mathcal{X}^n, \mathcal{S}^n) \to (R^m, \mathcal{B}^m)$ platí, že

$$T^{-1}(\mathcal{B}^l) \subset S^{-1}(\mathcal{B}^m),$$

kde $T^{-1}(\mathcal{B}^l)$ a $S^{-1}(\mathcal{B}^m)$ jsou σ-algebry generované náhodnými veličinami S a T.

Pokud T je minimální postačující statistika a pro libovolnou postačující statistiku S platí, $S^{-1}(\mathcal{B}^m) \subset T^{-1}(\mathcal{B}^l)$, potom $S^{-1}(\mathcal{B}^m) = T^{-1}(\mathcal{B}^l)$. To znamená, že statistika, která závisí na minimální postačující statistice je sama minimální postačující statistikou. Dvě statistky, které jsou současně minimální a postačující, jsou ekvivalentní.

1.2 Halmosova a Savageova faktorizační věta

Definice 1.2.1. Nechť $(\mathcal{X}, \mathcal{S})$ je měřitelný prostor, kde $\mathcal{S}$ je σ-algebra a μ a ν jsou míry na $\mathcal{S}$. Řekneme, že ν je absolutně spojitá vzhledem k μ, jestliže pro každou $E \in \mathcal{S}$ platí

$$\mu(E) = 0 \Rightarrow \nu(E) = 0.$$

Následující věta popisuje rodinu rozdělení pro kterou je daná statistika S postačující. Poznamenejme, že každé rozdělení $\{P_\theta, \theta \in \Theta\}$ je absolutně spojité vzhledem k nějaké míře μ. V takovém případě říkáme, že rodina $\{P_\theta\}$ je dominována mírou μ. Následující Věta a Lemma se najde v knize [2].

Věta 1.2.2. *Statistika $S : (\mathcal{X}^n, \mathcal{S}^n) \to (R^m, \mathcal{B}^m)$ je postačující pro rodinu $\{P_\theta, \theta \in \Theta\}$, která je dominována mírou μ právě tehdy, když hustota $\frac{dP_\theta}{d\mu} = p(\mathbf{x}; \theta)$ připouští faktorizaci:*

$$p(\mathbf{x}; \theta) = \mathbf{R}[S(\mathbf{x}); \theta]\mathbf{r}(\mathbf{x}) \quad \text{s. j.} \quad \mu, \theta \in \Theta, \tag{1.2.1}$$

kde $\mathbf{R}(\cdot; \theta)$ je nezáporná $\mathcal{B}^m$-měřitelná funkce a $\mathbf{r}(\mathbf{x})$ je nezáporná $\mathcal{S}^n$-měřitelná funkce.

Věta 1.2.2 v této formulaci byla dokázána Halmosem a Savagem [2]. Dřív a v méně zobecněné formě tuto větu dokázali Fisher a Neymann. Proto se v literatuře většinou nazývají Fisherovou-Neymanovou faktorizační větou nebo kritériem.

K důkazu Věty 1.2.2 potřebujeme následující Lemma.

Lemma 1.2.3. *Pokud rodina rozdělení* $\{P_\theta, \theta \in \Theta\}$ *je dominována mírou* μ*, potom existuje její spočetná podmnožina* $\{P_{\theta_1}, P_{\theta_2}, \ldots\}$ *taková, že platí následující implikace*

$$P_{\theta_i}(A) = 0 \quad pro \quad i = 1, 2, \ldots \quad \Rightarrow \quad P_\theta(A) = 0, \quad pro\ každé \quad \theta \in \Theta.$$

Důkaz lemmy 1.2.3 je možné najít v literatuře, například v [3].

Uvažujme pravděpodobnostní míru

$$\lambda = \sum_i c_i P_{\theta_i}, \tag{1.2.2}$$

kde $c_i > 0$ a $\sum_i c_i = 1$. Tato míra má takovou vlastnost, že pokud $\lambda(A) = 0$ potom $P_\theta(A) \equiv 0, \theta \in \Theta$ a naopak, když $P_\theta(A) \equiv 0$ potom $\lambda(A) = 0$.

Důkaz. **Halmosova a Savageova faktorizační věta.**
Nutná podmínka. Nechť je statistika $S(\mathbf{x})$ postačující pro rodinu P_θ. Potom z Lemmy 1.2.3 vyplýva, že pokud $P_\theta(A|S) = \psi_A$, potom i $\lambda(A|S) = \psi_A$. Vskutku, pro $\forall B \in S^{-1}(\mathcal{B}^m)$ dostaneme

$$\int_B \psi_A d\lambda = \int_B \psi_A \sum_i c_i dP_{\theta_i} = \sum_i c_i \int_B \psi_A dP_{\theta_i} =$$

$$= \sum_i c_i \int_{B\cap A} dP_{\theta_i} = \sum_i c_i P_{\theta_i}(B \cap A) = \lambda(B \cap A).$$

Položme $\frac{dP_\theta}{d\lambda} = f_\theta$ a označme χ_A jako indikátor množiny A:

$$\chi_A = \begin{cases} 1 & \text{pro } \mathbf{x} \in A \\ 0 & \text{pro } \mathbf{x} \notin A. \end{cases}$$

Vzhledem k tomu, že $E_\lambda(f_\theta|S) = f_\theta$ a $E[(P_\theta(A|S)] = P_\theta(A)$, dostaneme pro každé $A \in \mathcal{S}^n$

$$E_\lambda(\chi_A f_\theta) = P_\theta(A) = E_\theta[P_\theta(A|S)] = E_\lambda[f_\theta P_\theta(A|S)] =$$

$$= E_\lambda[f_\theta \lambda(A|S)] = E_\lambda[E_\lambda(f_\theta|S)\lambda(A|S)] =$$
$$= E_\lambda[E_\lambda(\chi_A E_\lambda(f_\theta|S)|S)] = E_\lambda[E_\lambda(\chi_A E_\lambda(f_\theta|S))].$$

Z toho, $f_\theta = E_\lambda(f_\theta|S)$ λ-s.v. a teda hustota f_θ je $\mathcal{B}^m$-měřitelná. Dále,

$$p(x;\theta) = \frac{dP_\theta}{d\mu} = \frac{dP_\theta}{d\lambda} \cdot \frac{d\lambda}{d\mu} = f_\theta \cdot r$$

a pokud $f_\theta = E_\lambda(f_\theta|S)$, dostaneme (1.2.1).
Postačující podmínka.
Z (1.2.1) plyne, že

$$\frac{d\lambda}{d\mu} = r(\mathbf{x}) \sum_i R[S(\mathbf{x}); \theta_i] = r(\mathbf{x}) \cdot G[S(\mathbf{x})].$$

Teď položme

$$\widetilde{f}_\theta = \begin{cases} \frac{R[S(\mathbf{x});\theta]}{G[S(\mathbf{x})]}, & \text{pokud} \quad G[S(\mathbf{x})] > 0; \\ 0, & \text{pokud} \quad G[S(\mathbf{x})] = 0, \end{cases}$$

kde $G[S(\mathbf{x})] = \sum_i R[S(\mathbf{x}); \theta_i]$. Je jednoduché ověřit, že $\mathcal{B}^m$-měřitelná funkce $\widetilde{f}$ je jedna z derivací $\frac{dP_\theta}{d\lambda}$, z čehož $\widetilde{f}_\theta = f_\theta$ λ-s.v. Pro $\forall A \in \mathcal{S}^n$ máme:

$$E_\theta[P_\theta(A|S)] = P_\theta(A) = E_\lambda(f_\theta \chi_A) = E_\lambda[E_\lambda(f_\theta \chi_A|S)] =$$
$$= E_\lambda[f_\theta E_\lambda(\chi_A|S)] = E_\theta[E_\lambda(\chi_A|S)] = E_\theta[\lambda(A|S)].$$

To znamená, že $E_\theta[P_\theta(A|S)] = E_\theta[\lambda(A|S)]$. Pokud poslední výraz platí pro každé $A \in \mathcal{A}$, platí také pro jev $A \cap B$, kde $B \in S^{-1}(\mathcal{B}^m)$. Pro takové případy nabývá tento výraz tvar:

$$E_\theta[\chi_B P_\theta(A|S)] = E_\theta[\chi_B \lambda(A|S)],$$

z čehož plyne $P_\theta(A|S) = \lambda(A|S)$ P_θ-s.j. To znamená, že statistika $S(\mathbf{x})$ je postačující. □

Lemma 1.2.4. *Nechť S je postačující statistika pro rodinu $\{P_\theta, \theta \in \Theta\}$, která je dominovánaFf mírou μ. Potom pro libovolnou měřitelnou funkci $\phi(x)$, pro kterou platí $E_\theta|\phi| < \infty$ pro každé $\theta \in \Theta$, existuje funkce $\widetilde{\phi}$ taková, že*

$$E_\theta(\phi|S) = \widetilde{\phi}, \quad P_\theta\text{-s.j.} \quad \theta \in \Theta. \tag{1.2.3}$$

Důkaz. Položme

$$\Theta_N = \{\theta \in \Theta : N < E_\theta|\phi| \leq N+1\}.$$

Z Lemmy 1.2.3 vyplývá, že $\bigcup_{N=0}^{\infty} \Theta_N = \Theta$. Pro každou Θ_N zkonstruujeme $\{\lambda_N\}$ podle (1.2.2) a položíme

$$\nu = \sum_{N=0}^{\infty} \frac{\lambda_N}{2^{N+1}}.$$

Rozdíl mezi mírou ν a λ spočívá v tom, že $E_\theta|\phi| < \infty$, ale $E_\lambda|\phi|$ může nabývat i ∞. Pokud postupujeme podle důkazu Věty 1.2.2 dostaneme, že

$$E_\theta(\phi|S) = E_\nu(\phi|S), \quad \text{s. v.}[P_\theta], \quad \theta \in \Theta.$$

□

Tutíž (1.2.3) se dá považovat za definicí postačující statistiky. Vlastnost (1.2.3) postačující statistiky se používá v teorii odhadu.

Struktura minimální postačující statistiky se dá popsat pomocí faktorizačního kritéria pokud předpokládáme, že $p(\mathbf{x}, \theta) > 0$ pro každé $\mathbf{x} \in \mathcal{X}^n$ a $\theta \in \Theta$. Vezměme θ_0 z množiny Θ pevné. Uvažujme statistiku

$$T : \mathbf{x} \to g_{\mathbf{x}}(\theta) = \ln \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta_0)} \tag{1.2.4}$$

(značení $g_{\mathbf{x}}(\theta)$ znamená, že g je funkcí parametru). Statistika $T(\mathbf{x})$ zobrazuje prostor $(\mathcal{X}^n, \mathcal{S}^n)$ do prostoru funkcí $\mathcal{T} = \{g_{\mathbf{x}}(\theta); \mathbf{x} \in \mathcal{X}^n\}$ definovaných na Θ. V prostoru $\mathcal{T}$ uvažujme σ-algebru nad konečně-rozměrnými válci, tedy nad množinami tvaru:

$$\{\gamma \in \mathcal{T} : (\gamma(\theta_1), \ldots, \gamma(\theta_s)) \in \mathbf{B}_s\},$$

kde $\theta_1, \ldots, \theta_s$ jsou body z množiny Θ a $\mathbf{B}_s$ je s-dimenzionální Borelovská množina. Potom statistika (1.2.4) představuje měřitelné zobrazení z $(\mathcal{X}^n, \mathcal{S}^n)$ do $(R^l, \mathcal{B}^l)$.
Předpokládejme, že existuje (konečný nebo nekonečný) interval I z R takový, že funkce

$$\begin{aligned} p(x;\theta) &> 0 \qquad x \in I, \quad \theta \in \Theta, \\ p(x;\theta) &= 0 \qquad x \notin I, \quad \theta \in \Theta \end{aligned}$$

je spojitá a diferencovatelná vzhledem k $x \in I$, a $\theta \in \Theta$. Pokud jsou tyto podmínky splněny, říkáme, že rodina hustot $\{p(\mathbf{x};\theta), \theta \in \Theta\}$ je regulární.

Následující dvě věty a její důkazy se najdou v článku [1].

Věta 1.2.5. *Minimální postačující statistika pro regulární rodinu* $\{P_\theta,\ \theta \in \Theta\}$ *je funkce*

$$g_x(\theta) = \ln p(x,\theta) - \ln p(x,\theta_0). \tag{1.2.5}$$

Důkaz. (a) Nejprve ukážeme, že statistika $g_x(\theta)$ je postačující.
Z rovnosti (1.2.5) dostaneme $p(x,\theta) = \mathrm{e}^{g_x(\theta)} p(x,\theta_0)$, což nám dává tvar (1.2.1), neboť $\mathrm{e}^{g_x(\theta)}$ závisí na parametru, ale $p(x,\theta_0)$ nikoliv.

(b) Zbývá ukázat, že statistika $g_x(\theta)$ je zároveň také minimální postačující. Nechť $\mathbf{S}(x)$ je postačující statistika, potom z (1.2.5) máme:

$$g_x(\theta) = \ln p(S(x),\theta) - \ln p(S(x),\theta_0),$$

kde $g_x(\theta)$ závisí na $S(x)$ což je ekvivalentní s definicí minimální postačující statistiky. □

Poznámka 1.2.6. Pokud $\Theta \subset R^k$ a $p(x,\theta)$ je spojitá a diferencovatelná v $\theta, \theta \in \Theta$ funkce, potom další minimální postačující statistika pro rodinu $\{P_\theta, \theta \in \Theta\}$ je

$$f_x(\theta) = grad_\theta \ln p(x,\theta) = \frac{1}{p(x,\theta)} grad_\theta p(x,\theta).$$

Věta 1.2.7. *Mějme náhodný výběr o rozsahu* n*. Potom minimální postačující statistika pro regulární rodinu* $\{P_\theta,\ \theta \in \Theta\}$ *je:*

$$g_{x_1,\ldots,x_n}(\theta) = g_{x_1}(\theta) + g_{x_2}(\theta) + \cdots + g_{x_n}(\theta).$$

Důkaz. Z nezávislosti náhodných veličin $X_1, \ldots, X_n$ platí

$$p(\mathbf{x},\theta) = p(x_1,\theta)p(x_2,\theta)\cdots p(x_n,\theta).$$

Potom podle Věty 1.2.5:

$$g_{x_i}(\theta) = \ln p(x_i,\theta) - \ln p(x_i,\theta_0), \quad i = 1,\ldots,n$$

$$
\begin{aligned}
g_{x_1,\ldots,x_n}(\theta) &= [\ln p(x_1,\theta) + \cdots + \ln p(x_n,\theta)] - [\ln p(x_1,\theta_0) + \cdots + \ln p(x_n,\theta_0)] \\
&= \sum_{i=1}^{n} \ln p(x_i,\theta) - \sum_{i=1}^{n} \ln p(x_i,\theta_0).
\end{aligned}
$$

Z toho dostáváme

$$
\begin{aligned}
\sum_{i=1}^{n} \ln p(x_i,\theta) &= g_{x_1,\ldots,x_n}(\theta) + \sum_{i=1}^{n} \ln p(x_i,\theta_0) \\
\prod_{i=1}^{n} p(x_i,\theta) &= \mathrm{e}^{g_{x_1,\ldots,x_n}(\theta)} \prod_{i=1}^{n} p(x_i,\theta_0)
\end{aligned}
$$

což je ekvivalentní s (1.2.1) pro

$$
\begin{aligned}
\mathbf{R}[S(\mathbf{x});\theta] &= \mathrm{e}^{g_{x_1,\ldots,x_n}(\theta)} \\
\mathbf{r}(\mathbf{x}) &= \prod_{i=1}^{n} p(x_i,\theta_0).
\end{aligned}
$$

□

Kapitola 2

Exponenciální rodina rozdělení

V této kapitole popisujeme rodinu jednorozměrních rozdělení, které mají takovou vlastnost, že některé její mocniny umožňují netriviální postačující statistiky. Zdá se, že rodina rozdělení s touto vlastností je exponenciálního typu.

2.1 Jednorozměrné rozdělení

Postupujeme podle knihy [2].

V této kapitole učiňíme předpoklad, že rozdělení $\{P_\theta, \theta \in \Theta\}$ na prostoru $\mathbf{R}^{mn}$ se σ-algebrou borelovských množin je dané vztahem

$$P_\theta(A) = \int \cdots \int I_A(\mathbf{x}_1, \ldots, \mathbf{x}_n) dF(\mathbf{x_1}, \theta) \cdots dF(\mathbf{x_n}, \theta), \qquad (2.1.1)$$

kde $F(\mathbf{x}, \theta)$ je distribuční funkce na $\mathbf{R}^m$ a $I_A(x_1, \ldots, x_n)$ je indikátor množiny A,

$$\begin{aligned} I_A(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= 1 \qquad & \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in A, \\ I_A(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= 0 \qquad & \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \notin A. \end{aligned}$$

V takovém případe říkáme, že rozdělení P_θ je n-tou mocninou
$F : dP_\theta = dF(\mathbf{x_1}, \theta) \cdots dF(\mathbf{x_n}, \theta)$. Ukážeme jaké podmínky musí $F(\mathbf{x}, \theta)$ splňovat aby pro rodinu $\{P_\theta, \theta \in \Theta\}$ existovala netriviální postačující statistika. Omezíme se na případ $m = 1$.
Pokud je distribuční funkce $F(x; \theta)$ dána hustotou $f(x; \theta)$ vzhledem k Lebesgueove míře, potom rodina $\{P_\theta, \theta \in \Theta\}$ je dominována Lebesgueovou mírou

μ na $\mathbf{R}^n$. Potom podle Halmosovy a Savagovy věty statistika S je postačující pro rodinu (2.1.1) a z toho vyplývá, že

$$\frac{dP_\theta}{d\mu} = p(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \mathbf{R}[S(\mathbf{x}); \theta]\mathbf{r}(\mathbf{x}).$$

To znamená, že když $S(\mathbf{x})$ je postačující statistika pro rodinu (2.1.1) potom platí rovnost

$$\prod_{i=1}^{n} f(x_i, \theta) = \mathbf{R}[S(\mathbf{x}); \theta]\mathbf{r}(\mathbf{x}), \quad \mathbf{x} = (x_1, \ldots, x_n). \tag{2.1.2}$$

Problém nalezení jednorozměrných rozdělení, která připouštějí netriviální statistiky, se omezí na popis všech funkcí $f(\mathbf{x}; \theta)$, které splňujou (2.1.2) pro některé $\mathbf{R}$ a $\mathbf{r}$.

Definice 2.1.1. Statistika $S(\mathbf{x}) : (R^n, \mathcal{B}^n) \to (R, \mathcal{B})$ v bodě $\mathbf{x}$ se nazývá triviální, jestliže na nějakém okolí U tohoto bodu platí:

$$S(\mathbf{x}') = S(\mathbf{x}'') \Rightarrow \mathbf{x}' = \mathbf{x}'', \quad \mathbf{x}', \mathbf{x}'' \in U. \tag{2.1.3}$$

Definice triviální statisiky se dá zformulovat ekvivalentním způsobem: Statistika $S(\mathbf{x})$, je triviální v bodě $\mathbf{x} = (x_1, \ldots, x_n)$ pokud existuje okolí U tohoto bodu takové, že

$$S^{-1}(\mathcal{B}) \cap U = \mathcal{B}^n \cap U,$$

kde průnik σ-algeber na okolí U je tvořen průnikem původních σ-algeber na tomto okolí. Statistika, která v žádném bodě není triviální, se nazývá netriviální statistika.

Označme L minimální lineární prostor funkcí definovaných na intervalu I, který obsahuje konstantní funkci $g_x(\theta) = 1$ a každou funkci

$$g_x(\theta) = \ln\left[\frac{f(x; \theta)}{f(x, \theta_0)}\right], \qquad \theta \in \Theta,$$

kde θ_0 je pevný parametr z množiny Θ a θ probíhá celým parametrickým prostorem Θ. Nechť je dimenze $L = r + 1$ ($r = \infty$ se nevylučuje).

Věta 2.1.2. *Nechť rodina hustot* (2.1.2) *je regulární. Potom*
a) pro $n \leq r$ *tato rodina rozdělení nemá netriviální postačující statistiku.*
b) pokud $n \geq r$ *a funkce* $1, h_1(x), ..., h_r(x)$ *tvoří bázi* L*, pak systém funkcí*

$$S_i(x_1, ..., x_n) = h_i(x_1) + ... + h_i(x_n), \quad pro \quad i = 1, ..., r \tag{2.1.4}$$

je funkcionálně nezávislý a $S(\mathbf{x}) = \{S_1(\mathbf{x}), ..., S_r(\mathbf{x})\}$ *tvoří minimální postačující statistiku pro rodinu rozdělení* (2.1.2).

Důkaz. Nechť $n \geq r + 1$ a funkce $1, h_1(x), ..., h_r(x)$ tvoří bázi prostoru L. Podle předpokladu, h_i jsou lineárními kombinacemi 1 a $g_\theta(x)$ a jsou spojitě diferencovatelné. Kdyby $S_1(\mathbf{x}), ..., S_r(\mathbf{x})$ byly funkcionálně závislé, potom pro pevné $x_{r+1}, ..., x_n$ by v bodech $x_1, ..., x_r$ mělo identicky platit, že Jacobian $\frac{\partial(S_1,...,S_r)}{\partial(x_1,...,x_r)} = 0$, tj.:

$$\frac{\partial(S_1, ..., S_r)}{\partial(x_1, ..., x_r)} = \begin{vmatrix} h_1'(x_1) & h_1'(x_2) & \cdots & h_1'(x_r) \\ h_2'(x_1) & h_2'(x_2) & \cdots & h_2'(x_r) \\ \vdots & \vdots & \ddots & \vdots \\ h_r'(x_1) & h_r'(x_2) & \cdots & h_r'(x_r) \end{vmatrix} = 0 \tag{2.1.5}$$

Ukážeme, že z (2.1.5) plyne lineární závislost funkcí $1, h_1(x), ..., h_r(x)$.
Pro $r = 1$: $h_1'(x_1) = 0$ právě tehdy, když $h_1(x_1)$ je konstanta.
Předpokládejme, že předchozí tvrzení platí pro $r - 1$. Potom když

$$\frac{\partial(S_1, ..., S_{r-1})}{\partial(x_1, ..., x_{r-1})} \equiv 0,$$

pak funkce $1, h_1(x), ..., h_{r-1}(x)$ jsou lineárně závislé. To znamená, že existuje netriviální lineární kombinace prvků $1, h_1(x), ..., h_{r-1}(x)$,

$$\sum_{i=0}^{r-1} a_i h_i = 0,$$

kde $h_0 = 1$ a $a_i, i = 0, \ldots, r-1$ jsou konstanty a alespoň jeden z a_i je nenulový. Potom funkce $1, h_1(x), ..., h_r(x)$ jsou také lineárně závislé. Protože víme, že $\sum_{i=0}^{r-1} a_i h_i = 0$ a alespoň jedna z $a_i, i = 0, \ldots, r-1$ je nenulová, potom a_r může nabývat hodnotu 0. Předpokládáme, že existuje bod $(x_1^0, \ldots, x_{r-1}^0) \in R$ v kterém platí

$$\frac{\partial(S_1, ..., S_{r-1})}{\partial(x_1^0, \ldots, x_{r-1}^0)} \neq 0.$$

Rozšířením determinantu (2.1.5) v bodě $x_1^0, \ldots, x_{r-1}^0, x$ v rámci prvků posledního sloupce dostáváme

$$A_1 h_1{}'(x) + A_2 h_2{}'(x) + \cdots + A_r h_r{}'(x) \equiv 0, \qquad x \in I, \tag{2.1.6}$$

kde $A_1, \ldots, A_r$ jsou konstanty v následujícím tvaru

$$A_1 = (-1)^{1+r} \begin{vmatrix} h_2'(x_2^0) & h_2'(x_3^0) & \cdots & h_2'(x_{r-1}^0) \\ h_3'(x_2^0) & h_3'(x_3^0) & \cdots & h_3'(x_{r-1}^0) \\ \vdots & \vdots & \ddots & \vdots \\ h_r'(x_2^0) & h_r'(x_3^0) & \cdots & h_r'(x_{r-1}^0) \end{vmatrix},$$

$$A_2 = (-1)^{2+r} \begin{vmatrix} h_1'(x_1^0) & h_1'(x_3^0) & \cdots & h_1'(x_{r-1}^0) \\ h_3'(x_1^0) & h_3'(x_3^0) & \cdots & h_3'(x_{r-1}^0) \\ \vdots & \vdots & \ddots & \vdots \\ h_r'(x_1^0) & h_r'(x_3^0) & \cdots & h_r'(x_{r-1}^0) \end{vmatrix},$$

$$\vdots$$

$$A_r = (-1)^{r+r} \begin{vmatrix} h_1'(x_1^0) & h_1'(x_2^0) & \cdots & h_1'(x_{r-1}^0) \\ h_2'(x_1^0) & h_2'(x_2^0) & \cdots & h_2'(x_{r-1}^0) \\ \vdots & \vdots & \ddots & \vdots \\ h_{r-1}'(x_1^0) & h_{r-1}'(x_2^0) & \cdots & h_{r-1}'(x_{r-1}^0) \end{vmatrix} \neq 0.$$

Z (2.1.6) plyne, že funkce $h(x) = A_1 h_1{}'(x) + A_2 h_2{}'(x) + \cdots + A_r h_r{}'(x)$ je konstantní na celém I a proto funkce $1, h_1(x), ..., h_r(x)$ jsou lineárně závislé na I. Toto je ve sporu s předpokaledem, že $1, h_1(x), ..., h_r(x)$ tvoří bází prostoru L. Tím jsme dokázali funkcionální nezávislost statistiky $S_1(\mathbf{x}), ..., S_r(\mathbf{x})$. Dále podle Věty 1.2.5,

$$\begin{aligned} \ln \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta_0)} &= \sum_{i=1}^{n} g_{x_i}(\theta) = \sum_{i=1}^{n} \ln \frac{f(\mathbf{x_i}, \theta)}{f(\mathbf{x_i}, \theta_0)} \\ &= \sum_{i=1}^{n} \left[\sum_{k=1}^{r} c_k(\theta) h_k(x_i) + c_0(\theta) \right] \\ &= \sum_{k=1}^{r} c_k(\theta) S_k(\mathbf{x}) + n c_0(\theta). \end{aligned}$$

Z toho:

$$\frac{p(\mathbf{x},\theta)}{p(\mathbf{x},\theta_0)} = \exp\left\{\sum_{k=1}^{r} c_k(\theta)S_k(\mathbf{x}) + nc_0(\theta)\right\}.$$

Tudíž

$$p(\mathbf{x},\theta) = \mathbf{R}[S(\mathbf{x});\theta]\mathbf{r}(\mathbf{x}),$$

kde

$$\begin{aligned}\mathbf{R}[S(\mathbf{x});\theta] &= \exp\left[\sum_{k=1}^{n} c_k(\theta)S_k(\mathbf{x}) + nc_0(\theta)\right],\\ \mathbf{r}(\mathbf{x}) &= p(\mathbf{x},\theta_0).\end{aligned}$$

Potom statistika $S(\mathbf{x})$ je postačující pro rodinu (2.1.2). Z definice prostoru L plyne

$$h_k(x) = \sum_s c_{ks} g_x(\theta_s) + c_{k0}.$$

Odtud

$$\begin{aligned}S_k(\mathbf{x}) &= \sum_{i=1}^{n} h_k(x_i) = \sum_{i=1}^{n}\sum_s c_{ks} g_{x_i}(\theta_s) + nc_{k0}\\ &= \sum_s c_{ks}\sum_{i=1}^{n} g_{x_i}(\theta_s) + nc_{k0}.\end{aligned}$$

a víme, že statistika

$$T : x \to g_x(\theta) = \ln\frac{p(x,\theta)}{p(x,\theta_0)} = \sum_{i=1}^{n} g_{x_i}(\theta) \tag{2.1.7}$$

je minimální postačující statistika. Ale z předcházejícího vztahu plyne, že $S_k(\mathbf{x}) = \tilde{S}_k[T(\mathbf{x})]$, kde $T(\mathbf{x}) = \sum_{i=1}^{n} g_{x_i}(\theta_s)$ a proto

$$S(\mathbf{x}) = \{S_1(\mathbf{x}), ..., S_r(\mathbf{x})\} : (\mathbf{R}^n, \mathcal{B}^n) \to (\mathbf{R}^r, \mathcal{B}^r)$$

je minimální postačující statistika pro rodinu (2.1.2).

V následujícím kroku dokážeme, že pro $n \leq r$ rodina (2.1.2) nepřipouští netriviální postačující statistiky.
V prostoru L zvolíme funkce $1, h_1(x), ..., h_r(x)$ tak, aby byly lineárně nezávislé. Potom podle první části důkazu je Jacobiho determinant nenulový

$$\frac{\partial(S_1, ..., S_n)}{\partial(x_1, ..., x_n)} \neq 0$$

alespoň v bodě $({x_1}^0, ..., {x_n}^0)$. A podle věty o inverzní funkci:

$$x_i = x_i(S_1, ..., S_n), \qquad i = 1, ..., n$$

tedy máme

$$\mathcal{R}^n \cap U \subset S^{-1}(\mathcal{R}^n) \cap U \subset T^{-1}(\mathcal{B}^l) \cap U \tag{2.1.8}$$

kde T je minimální postačující statistika. Podle (2.1.8) je minimální postačující statistika triviální dokonce v bodě $\mathbf{x}^0 = ({x_1}^0, ..., {x_n}^0)'$. Z toho vyplývá, že rodina (2.1.2) nemá netriviální postačující statistiku. □

2.2 Exponenciální rodina rozdělení

Významná rodina rozdělení, u kterých lze určit postačující statistiky je exponencialná rodina. Do této rodiny rozdělení patří například: binomické, Poissonovo, normální, exponenciální a gamma rozdělení. V následující větě podle [2] popíšeme všechny regulární hustoty, které splňují (2.1.2).

Věta 2.2.1. *Pokud regulární hustota $f(x, \theta)$, $\theta \in \Theta$ splňuje* (2.1.2) *s netriviální statistikou $S(\mathbf{x})$, potom pro nějaké $r \leq n - 1$*

$$f(x; \theta) = \exp\left\{\sum_{i=1}^{r} c_i(\theta) h_i(x) + c_0(\theta) + h_0(x)\right\} \tag{2.2.1}$$

kde funkce $h_1(x), ..., h_r(x)$ jsou spojitě diferencovatelné a systém funkcí $\{1, h_1, ..., h_r\}$ je lineárně nezávislý.

To znamená, že mezi regulárními rodinami pouze exponenciální rodina hustot (2.2.1) má takové vlastnosti, že některé její mocniny splňují (2.1.2) s netriviální statistikou $S(\mathbf{x})$.

Důkaz. 1) Z předpokladu věty a z netriviality statistiky $S(\mathbf{x})$ plyne, že dimenze $L = r + 1 \leq n$. Nechť funkce $h_1(x), ..., h_r(x)$ tvoří báze prostoru L a $\{1, c_1, ..., c_r\}$ jsou konstanty z L, pak statistiku $g_x(\theta)$ lze zapsat ve tvaru

$$g_x(\theta) = \ln \frac{f(x,\theta)}{f(x,\theta_0)} = \sum_{i=1}^{r} c_i(\theta) h_i(x) + c_0(\theta), \qquad \theta \in \Theta,$$

neboli

$$\ln f(x;\theta) - \ln f(x,\theta_0) \;=\; \sum_{i=1}^{r} c_i(\theta) h_i(x) + c_0(\theta). \tag{2.2.2}$$

Označme $h_0(x) = f(x,\theta_0)$. Vidíme, že (2.2.2) je ekvivalentní s (2.2.1):

$$\begin{aligned} \ln f(x;\theta) &= \sum_{i=1}^{r} c_i(\theta) h_i(x) + c_0(\theta) + h_0(x), \\ f(x;\theta) &= \exp\left\{ \sum_{i=1}^{r} c_i(\theta) h_i(x) + c_0(\theta) + h_0(x) \right\}. \end{aligned}$$

A obráceně, pokud $f(x;\theta)$ zapíšeme ve tvaru (2.2.1), potom

$$\begin{aligned} p(\mathbf{x},\theta) &= \prod_{i=1}^{n} f(x_i,\theta) \\ &= \exp\left\{ \sum_{i=1}^{r} c_i(\theta) \sum_{j=1}^{n} h_i(x_j) + n c_0(\theta) + \sum_{j=1}^{n} h_0(x_j) \right\} \\ &= \mathbf{R}(S_1, ..., S_r; \theta) \mathbf{r}(\mathbf{x}), \end{aligned}$$

a tedy

$$S_i \;=\; \sum_{j=1}^{n} h_i(x_j).$$

Z toho vidíme, že $S(\mathbf{x}) = \{S_1(\mathbf{x}), ..., S_r(\mathbf{x})\}$ je postačující pro rodinu (2.1.2). Vzhledem k tomu, že $r \leq n-1$ a funkce $h_1(x), ..., h_r(x)$ jsou spojitě diferencovatelné, je postačující statistika v tomto případě netriviální. □

Podle [1] zformulujeme následující definici a větu s důkazem.

Definice 2.2.2. Hodností systému distribučních funkcí $\{P_\theta, \theta \in \Theta\}$ na prostoru $\mathcal{X}$ rozumíme největší takové r, pro které platí, že pro žádné konečné $n \leq r$ rodina $\{P_\theta, \theta \in \Theta\}$ nemá netriviální postačující statistiky pro výběry o rozsahu n na prostoru $\mathcal{X}$.

Věta 2.2.3. *Nechť*

$$f(x,\theta) = \exp\left\{\sum_{i=1}^{r} h_i(x)c_i(\theta) + c_0(\theta) + h_0(x)\right\}.$$

Pak hodnost systému distribučních funkcí $\{P_\theta, \theta \in \Theta\}$ není větší něž r.
Pokud $(1, h_1, ..., h_r)$ a $(1, c_1, ..., c_r)$ jsou lineárně nezávislé systémy funkcí, pak hodnost $\{P_\theta, \theta \in \Theta\}$ je r a pro každé $n \geq r$ jsou funkce v systému funkcí

$$S_i(x_1, ..., x_n) = \sum_{j=1}^{n} h_i(x_j), \qquad (i = 1, ..., r)$$

funkcionálně nezávislé a tvoří n-rozměrnou minimální postačující statistiku pro rodinu (2.1.2).

Důkaz. Máme:

$$\begin{aligned} g_x(\theta) &= \ln f(x,\theta) - \ln f(x,\theta_0) \\ &= \sum_{i=1}^{r} (c_i(\theta) - c_i)h_i(x) + c_0(\theta) - c_0, \end{aligned} \tag{2.2.3}$$

kde $c_i = c_i(\theta_0)$. To znamená, že dimenze $L \leq r + 1$ a hodnost $\{P_\theta, \theta \in \Theta\}$ se nerovná n. Dále, pokud $(1, c_1(\theta), ..., c_r(\theta))$ jsou lineárně nezávislé, $(c_1(\theta) - c_1(\theta_0), \ldots, c_r(\theta) - c_r(\theta_0))$ jsou také lineárně nezávislé. Z toho plyne, že $\theta_1, ..., \theta_r \in \Theta$ můžeme zvolit tak, že se determinant $\|c_i(\theta_j) - c_i(\theta_0)\| \neq 0$.
Pokud v (2.2.3) nahradíme θ $(\theta_1, ..., \theta_r) \in \Theta$ a vyřešíme vzniklou soustavu rovnic, dostaneme:

$$h_i(x) = \sum_{j=1}^{r} b_i{}^j g_{\theta_j}(x) + b_i{}^0, \qquad i = 1, ..., r \tag{2.2.4}$$

kde $b_i{}^j$ jsou konstanty. Z rovnosti (2.2.4) vidíme, že $(1, h_1, ..., h_r)$ patří do L a generují L.
Pokud funkce $(1, h_1, ..., h_r)$ jsou lineárně nezávislé, potom tyto funkce tvoří bázi L a důkaz plyne z Věty 2.1.2. □

2.2.1 Exponenciální rodina s parametrem polohy $F(x-\theta)$

Uvažujeme exponenciální rodinu rozdělení, která podle (2.2.1) má tvar

$$f(x;\theta) = \exp\left\{\sum_{i=1}^{r} c_i(\theta)h_i(x) + c_0(\theta) + h_0(x)\right\}.$$

Táto rodina závisí na parametru $\theta \in \Theta$. V speciálním případě, kdy θ je parametrem polohy, se dá exponenciální rodinu definovat důkladněji.

Následující věta a její důkaz je k nalezení v [2].

Věta 2.2.4. *Nechť se regulární hustota $f(x-\theta)$, která závisí na parametru polohy $\theta \in R$ dá napsat ve formě* (2.2.1). *Potom pro libovolné $s \in N$*

$$f(x) = \exp\left[\sum_{i=1}^{s} c_i x^{n_i} \mathrm{e}^{\mu_i x}\right], \quad x \in R, \tag{2.2.5}$$

kde n_i jsou nezáporná, celá čísla a μ_i, c_i jsou komplexní čísla.

Důkaz. Z předpokladu regularity $f(x-\theta)$ plyne, že $f(x-\theta) > 0$ pro $\forall x \in R^1$ a $\forall \theta \in R^1$. Nechť $H(x) = \ln f(x)$. Pro $\theta \in R^1$ prostor funkcí $H(x-\theta)$ má konečnou dimenzi a z toho plyne, že

$$H(x-\theta) = \sum_{j=1}^{k} a_j(\theta)H_j(x)$$

kde $H_j(x) = H(x-\theta_j)$. Z důkazu Věty 2.1.2 plyne, že existují $x_1, ..., x_k$ takové, že

$$\begin{vmatrix} H_1(x_1) & H_2(x_1) & \cdots & H_k(x_1) \\ H_1(x_2) & H_2(x_2) & \cdots & H_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ H_1(x_k) & H_2(x_k) & \cdots & H_k(x_k) \end{vmatrix} \neq 0.$$

Potom v systému

$$H(x_i-\theta) = \sum_{j=1}^{k} a_j(\theta)H_j(x_i), \quad i = 1, ..., k,$$

se funkce $a_j(\theta)$ dají vyjádřit lineárně ve smyslu $H(x_i - \theta)$ a protože

$$\begin{aligned}
H_1(x_1)a_1(\theta) + H_2(x_1)a_2(\theta) + \cdots + H_k(x_1)a_k(\theta) &= H(x_1 - \theta) \\
H_1(x_2)a_1(\theta) + H_2(x_2)a_2(\theta) + \cdots + H_k(x_2)a_k(\theta) &= H(x_2 - \theta) \\
&\cdots \\
H_1(x_k)a_1(\theta) + H_2(x_k)a_2(\theta) + \cdots + H_k(x_k)a_k(\theta) &= H(x_k - \theta)
\end{aligned}$$

nebo v matici formě

$$\begin{aligned}
H(\mathbf{x})A(\theta) &= H(\mathbf{x} - \theta), \\
A(\theta) &= H^{-1}H(\mathbf{x} - \theta).
\end{aligned}$$

kde

$$H(\mathbf{x}) = \begin{pmatrix} H_1(x_1) & H_2(x_1) & \cdots & H_k(x_1) \\ H_1(x_2) & H_2(x_2) & \cdots & H_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ H_1(x_k) & H_2(x_k) & \cdots & H_k(x_k) \end{pmatrix},$$

$$A(\theta) = (a_1(\theta), a_2(\theta), \cdots, a_k(\theta))'$$

a

$$H(\mathbf{x} - \theta) = (H(x_1 - \theta), H(x_2 - \theta), \cdots, H(x_k - \theta))'.$$

Funkce $H(x_j - \theta), j = 1, \ldots, k$ jsou diferencovatelné podle θ. Z toho plyne, že $a_i(\theta), i = 1, \ldots, k$ jsou také diferencovatelné podle θ. A platí

$$\frac{\partial H(x - \theta)}{\partial \theta} = H_j'(x - \theta) = \sum_{i=1}^{k} a_j'(\theta) H_j(x); \quad i = 1, \ldots, k, \tag{2.2.6}$$

Položme v (2.2.6) $\theta = \theta_1, \ldots, \theta_k$ a dostaneme

$$H_j'(x) = \sum_{i=1}^{k} A_{ij}(\theta_i) H_j(x); \quad i = 1, \ldots, k, \tag{2.2.7}$$

kde $A_{ij} = a_j'(\theta_i)$. Rovnost (2.2.7) je systém lineárních diferenciálních rovnic s konstantními koeficienty. Z teorie diferenciálních rovnic víme, že její zobecnění řešení má tvar

$$H_j(x) = \left[\sum_{i=1}^{k} c_{ij} x^{n_i} \mathrm{e}^{\mu_i x} \right]. \tag{2.2.8}$$

Dále víme, že $H_j(x) = \ln f_j(x)$, a z podmínky (2.2.8) plyne tvrzení věty. □

2.2.2 Exponenciální rodina s parametrem měřítka $F(\frac{x}{\sigma})$

V tomto odstavci postupujeme podle [2].

Dalším speciálním parametrem pro exponenciální rodinu (2.2.1) je parametr měřítka, $\sigma \in R_+^1$. Pokud $f(x/\theta)$ je regulární, potom

$$\begin{array}{lcll} f(x) & > & 0 \quad \text{pro} \quad x \in R & \text{nebo} \\ f(x) & > & 0 \quad \text{pro} \quad x \in R_+^1 & \\ f(x) & = & 0 \quad \text{pro} \quad x \in R_-^1 & \text{nebo} \\ f(x) & > & 0 \quad \text{pro} \quad x \in R_-^1 & \\ f(x) & = & 0 \quad \text{pro} \quad x \in R_+^1 & \end{array}$$

Všechny případy můžeme uvažovat stejným způsobem. My se zde omezíme na druhý případ, když $f(x) > 0$ pro $x \in R_+^1$ a $f(x) = 0$ pro $x \in R_-^1$.

Věta 2.2.5. *Nechť regulární hustota $(1/\sigma)f(x/\sigma)$, kde $f(x) > 0$ pro $x \in R_+^1$ a $f(x) \equiv 0$ pro $x \in R_-^1$, která závisí na parametru měřítka $\sigma \in R_+^1$, patří do exponenciální třídy hustot. Potom pro libovolné $s \in N$*

$$f(x) = \exp\left[\sum_{i=1}^{s} c_i (\ln x)^{n_i} x^{\mu_i}\right], \quad x \in R_+^1, \tag{2.2.9}$$

kde n_i jsou nezáporná, celá čísla a μ_i, c_i jsou komplexní čísla.

Důkaz. Položme $x = \mathrm{e}^y, \sigma = \mathrm{e}^\rho$. Nechť $f(x) > 0$ pro $x \in R_+^1$ a $f(x) = 0$ pro R_-^1. Potom nová hustota má tvar

$$h(y, \rho) = \mathrm{e}^{y-\rho} f(\mathrm{e}^{y-\rho}) = h(y - \rho).$$

Podrobněji, máme

$$F\left(\frac{x}{\sigma}\right) \;=\; \frac{1}{\sigma}\int_0^x f\left(\frac{t}{\sigma}\right) \mathrm{d}t.$$

Použijeme substituce $t = \mathrm{e}^y, \mathrm{d}t = \mathrm{e}^y \mathrm{d}y, \sigma = \mathrm{e}^\rho$,

$$\begin{aligned} \frac{1}{\sigma}\int_0^x f\left(\frac{t}{\sigma}\right) \mathrm{d}t &= \frac{1}{\sigma}\int_{-\infty}^{\ln(x)} f\left(\frac{\mathrm{e}^y}{\sigma}\right) \mathrm{e}^y \,\mathrm{d}y = \mathrm{e}^{-\rho}\int_{-\infty}^{\ln(x)} f\left(\mathrm{e}^{y-\rho}\right) \mathrm{e}^y \,\mathrm{d}y \\ &= \int_{-\infty}^{\ln(x)} f\left(\mathrm{e}^{y-\rho}\right) \mathrm{e}^{y-\rho} \,\mathrm{d}y. \end{aligned}$$

Z posledního vyjádření vyplývá, že hustota distribuční funkce $F(\frac{x}{\sigma})$ je $h(y-\rho) = h(y,\rho) = \mathrm{e}^{y-\rho} f(\mathrm{e}^{y-\rho})$. Potom podle Věty 2.2.4 dostaneme

$$h(y) = \exp\left[\sum_{i=1}^{s} c_i y^{n_i} \mathrm{e}^{\mu_i x}\right]$$

z čeho vyplývá hledaný tvar (2.2.9) pro původní rodinu hustot. □

Kapitola 3

Příklady

V této kapitole uvádíme několik příkladů výpočtu postačující statistiky převzaté z [1] a [4].

V následujícím příkladě ukážeme postup jak spočítat postačující statistiku pro parametr binomického rozdělení podle Definice 1.1.2.

Příklad 1. Nechť $X_1, ..., X_n$ jsou nezávislé stejně rozdělené nahodné veličiny s binomickým rozdělením s parametry m a p,
tj. $\mathsf{Bi}(m, p)$.

$$\mathsf{P}(X_i = k) = \binom{m}{k} p^k (1-p)^{m-k}, \quad k = 0, \ldots, m \quad i = 1, \ldots, n$$

Potom sdružené rozdělení $X_1, ..., X_n$ je binomické s parametry mn a p, tj. $\mathbb{X} \sim \mathsf{Bi}(mn, p)$:

$$\begin{aligned} \mathsf{P}(X_1 = x_1, ..., X_n = x_n) &= \prod_{i=1}^{n} \mathsf{P}(X_i = x_i) = \\ &= \prod_{i=1}^{n} \binom{m}{x_i} p^{\sum_{i=1}^n x_i} (1-p)^{mn - \sum_{i=1}^n x_i} \end{aligned}$$

Označme $s = \sum_{i=1}^n x_i$, potom $S = \sum_{i=1}^n X_i$ je postačující statistika pro

parametr p, jelikož podmíněné rozdělení nezávisí na parametru p:

$$\begin{aligned}
\mathsf{P}(X_1 = x_1, ..., X_n = x_n | \sum_{i=1}^{n} x_i = s) &= \frac{\mathsf{P}(X_1 = x_1, ..., X_n = x_n, \sum_{i=1}^{n} x_i = s)}{\mathsf{P}(\sum_{i=1}^{n} x_i = s)} \\
&= \frac{\mathsf{P}(X_1 = x_1, ..., X_n = x_n)}{\mathsf{P}(\sum_{i=1}^{n} x_i = s}) = \\
&= \frac{\prod_{i=1}^{n} \binom{m}{x_i} p^s (1-p)^{mn-s}}{\binom{mn}{s} p^s (1-p)^{mn-s}} = \\
&= \frac{\prod_{i=1}^{n} \binom{m}{x_i}}{\binom{mn}{s}}.
\end{aligned}$$

Výše uvedená rovnost platí pro $s = \sum_{i=1}^{n} x_i$. Je-li $s \neq \sum_{i=1}^{n} x_i$, pak tato podmíněná pravděpodobnost se rovná nule. Dále ukážeme, jak lze pomocí faktorizačního kritéria určit postačující statistiky pro dané rozdělení.

$$\mathbf{R}[S(\mathbf{x}); p] = p^{\sum_{i=1}^{n} x_i} (1-p)^{mn - \sum_{i=1}^{n} x_i}$$

a

$$\mathbf{r}(\mathbf{x}) = \prod_{i=1}^{n} \binom{m}{x_i}.$$

Z toho plyne, že $S = \sum_{i=1}^{n} X_i$ je postačující statistikou pro parametr p.

Příklad 2. Nechť $X_1, ..., X_n$ jsou nezávislé stejně rozdělené náhodné veličiny. X_i má normální rozdělení s parametry $\mu \in R$ a $\sigma^2 > 0$, $\theta = (\mu, \sigma^2)$ tj. $X_i \sim N(\mu, \sigma^2)$ s hustotou

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x_i - \mu)^2}{2\sigma^2}}, \quad x_i \in R, \quad i = 1, \ldots, n.$$

Potom hustota sdruženého rozdělení má za podmínky nezávislosti náhodných veličin tvar:

$$\begin{aligned}
f(x_1, ..., x_n; \mu, \sigma^2) &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left\{ -\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \right\} \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left\{ -\frac{\sum_{i=1}^{n} {x_i}^2 + 2\mu \sum_{i=1}^{n} x_i - n\mu^2}{2\sigma^2} \right\} \\
&= \frac{1}{\sqrt{(2\pi)^n}} \frac{1}{\sigma^n} \exp\left\{ -\frac{\sum_{i=1}^{n} {x_i}^2 + 2\mu \sum_{i=1}^{n} x_i - n\mu^2}{2\sigma^2} \right\},
\end{aligned}$$

což odpovídá tvaru (1.2.1) pro

$$\mathbf{R}[S(\mathbf{x});\theta] \;=\; \frac{1}{\sigma^n}\exp\left\{-\frac{\sum_{i=1}^n {x_i}^2 + 2\mu\sum_{i=1}^n x_i - n\mu^2}{2\sigma^2}\right\}$$

a

$$\mathbf{r}(\mathbf{x}) \;=\; \frac{1}{\sqrt{(2\pi)^n}}.$$

Postačující statistika pro parametr $\theta = (\mu, \sigma^2)$ je $\mathbf{S} = (\sum_{i=1}^n x_i, \sum_{i=1}^n {x_i}^2)$. Jako pokračování příkladu odvodíme postačující statistiky normálního rozdělení pro parametry μ a σ^2 aplikací (1.2.1). Spojitá funcke $f(x;\mu,\sigma) > 0$ je diferencovatelná na celém $\mathbb{R}$ a pro $\theta = (\mu, \sigma^2)$ podle Věty 1.2.5 máme: Pro $X \sim N(\mu, \sigma^2)$

$$\begin{aligned}
g_x(\mu,\sigma) &= \ln p(x,\mu,\sigma) - \ln p(x,\mu,\sigma),\\
g_x(\mu,\sigma) &= \ln\frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2} - \ln\frac{1}{\sigma_0\sqrt{2\pi}} + \frac{(x-\mu_0)^2}{2{\sigma_0}^2}\\
&= \ln\frac{\sigma_0\sqrt{2\pi}}{\sigma\sqrt{2\pi}} - \frac{x^2-2\mu x+\mu^2}{2\sigma^2} + \frac{x^2-2\mu_0 x+{\mu_0}^2}{2{\sigma_0}^2}\\
&= -\frac{x^2}{2}\left(\frac{1}{\sigma^2} - \frac{1}{{\sigma_0}^2}\right) + x\left(\frac{\mu}{\sigma^2} - \frac{\mu}{{\sigma_0}^2}\right) - \left(\frac{\mu^2}{2\sigma^2} - \frac{{\mu_0}^2}{2{\sigma_0}^2} + \ln\frac{\sigma}{\sigma_0}\right).
\end{aligned}$$

$$g_{x_1,\ldots,x_n}(\mu,\sigma^2) = \ln p\left(\sum_{i=1}^n x_i,\mu,\sigma\right) - \ln p\left(\sum_{i=1}^n x_i,\mu,\sigma_0\right)$$

$$g_{x_1,\ldots,x_n}(\mu,\sigma^2) = g_{x_1}(\mu,\sigma^2) + g_{x_2}(\mu,\sigma^2) + \ldots + g_{x_n}(\mu,\sigma^2) = \sum_{i=1}^n g_{x_i}(\mu,\sigma^2) =$$

$$= -\frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{{\sigma_0}^2}\right)\sum_{i=1}^n {x_i}^2 + \left(\frac{\mu}{\sigma^2} - \frac{\mu}{{\sigma_0}^2}\right)\sum_{i=1}^n x_i - \left(\frac{\mu^2}{2\sigma^2} - \frac{{\mu_0}^2}{2{\sigma_0}^2} + \ln\frac{\sigma}{\sigma_0}\right).$$

Podle Věty 1.2.7 je funkce $g_{x_1,\ldots,x_n}(\mu,\sigma^2)$ minimální postačující statistikou pro μ a σ^2 a je ekvivalentní statistice

$$S(x_1,\ldots,x_n) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n {x_i}^2\right).$$

Následující tři příklady se dají nalézt v článku [1].

Příklad 3. Nechť pravděpodobnostní hustota je definována jako:

$$p(x,\beta,\lambda,\mu) = \begin{cases} 0, & x \le 0 \\ Cx^{\beta}\mathrm{e}^{-\lambda x^{\mu}}, & x > 0. \end{cases} \tag{3.0.1}$$

Hodnotu C se dá určit z podmínky $\int_{\infty}^{-\infty} p(x,\beta,\lambda,\mu) = 1$. Nechť λ a β jsou pevné, $\mu \in (0,+\infty)$, potom rodina rozdělení:

$$\mathcal{F} = \{p_{\mu}(x,\beta,\lambda); \mu \in (0,+\infty); \quad \lambda,\beta \quad \text{pevné}\}$$

je regulární na $(0,+\infty)$, podle (1.2.5):

$$\begin{aligned} g_x(\mu) &= \ln p(x,\beta,\lambda,\mu) - \ln p(x,\beta_0,\lambda_0,\mu_0) \\ &= \ln(Cx^{\beta}\mathrm{e}^{-\lambda x^{\mu}}) - \ln(Cx^{\beta_0}\mathrm{e}^{-\lambda_0 x^{\mu_0}}) \\ &= \ln(Cx^{\beta}) - \lambda x^{\mu} - \ln(C_0 x^{\beta_0}) + \lambda_0 x_0^{\mu} \\ &= \lambda x^{\mu} + (\ln C + \ln C_0 + \lambda_0 x^{\mu_0}), \quad x > 0 \end{aligned}$$

Pokud μ je neznámé, potom neexistuje netriviální postačující statistika pro hustotu (3.0.1). V tomto případě dimenze $L = \infty$ a to se nezmění, když se α a β nepovažují za konstanty. Pokud μ je známé, potom hustota (3.0.1) se dá napsat ve tvaru:

$$\begin{aligned} p(x,\beta,\lambda,\mu) &= Cx^{\beta}\mathrm{e}^{-\lambda x^{\mu}} \\ &= \exp\left\{-\lambda x^{\mu} + \beta \ln x + \ln C\right\}, \qquad x > 0 \end{aligned}$$

Sdružené rozdělení má tvar

$$\begin{aligned} \prod_{i=1}^{n} p(x_i,\beta,\lambda,\mu) &= \prod_{i=1}^{n} \exp\{-\lambda {x_i}^{\mu} + \beta \ln x_i + \ln C\} \\ &= \exp\left\{-\lambda \sum_{i=1}^{n} {x_i}^{\mu} + n \ln C + \beta \sum_{i=1}^{n} \ln x_i\right\}. \end{aligned}$$

Když obě parametry β a λ jsou neznámé a μ je známé pak $\sum_{i=1}^{n} {x_i}^{\mu}$ a $\sum_{i=1}^{n} \ln x_i$ tvoří minimální postačující statistiku pro β a λ.
Když β je nezmámý parametr (λ známé) a μ je známé, potom $\sum_{i=1}^{n} \ln x_i$ tvoří minimální postačující statistiku pro β.
Když λ je nezmámý parametr (β známé) a μ je známé, potom $\sum_{i=1}^{n} {x_i}^{\mu}$ tvoří minimální postačující statistiku pro λ.

Příklad 4. Mějme hustotu definovanou jako

$$p(x,\alpha,\sigma)=\frac{1}{2\sigma}\mathrm{e}^{-\frac{|x-\alpha|}{\sigma}},\quad \alpha\in\mathbb{R},\quad \sigma^2>0,\quad \infty<x<-\infty.$$

.

Sdružené rozdělení má tvar

$$\prod_{i=1}^{n}p(x_i,\alpha,\sigma)=\frac{1}{(2\sigma)^n}\mathrm{e}^{-\frac{\sum_{i=1}^{n}|x_i-\alpha|}{\sigma}}.$$

Podle Věty 1.2.7:

$$\begin{aligned} g_{x_i}(\theta) &= \ln p(x_i,\alpha,\sigma)-\ln p(x_i,\alpha_0,\sigma), \qquad i=1,...,n\\ \sum_{i=1}^{n}g_{x_i}(\theta) &= \sum_{i=1}^{n}\frac{|x_i-\alpha|-|x_i-\alpha_0|}{\sigma}\\ &= \frac{1}{\sigma}\sum_{i=1}^{n}|x_i-\alpha|-\frac{1}{\sigma}\sum_{i=1}^{n}|x_i-\alpha_0| \end{aligned}$$

z toho vidíme, že pro známé α minimální postačující statistika pro σ je $\sum_{i=1}^{n}|x_i-\alpha|$.
Když α je neznámé potom rodina distribučních funkcí má nekonečnou hodnost na prostoru L podle věty 1.2.2 a obsahuje nekonečně mnoho nezávislých funkcí $|x-\alpha|$ a tedy minimální postačující statistika neexistuje.

Na následujícím příkladu ukážeme, jaký tvar má postačující statistika pro hustotu z exponenciální rodiny rozdělení s parametrem polohy.

Příklad 5. Hustota z exponenciální rodiny je definovaná následujícím způsobem:

$$p(x)=\mathrm{e}^{-x+\mathrm{e}^{-x})}\qquad -\infty<x<\infty. \tag{3.0.2}$$

Potom hustota

$$p(x-a)=\mathrm{e}^{-[(x-a)+\mathrm{e}^{-(x-a)}]}\qquad x,a\in R \tag{3.0.3}$$

je regulární na R a závisí na parametru polohy a. Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené náhodné veličiny, kde X_i má hustotu ve tvaru (3.0.3). Potom sdružená hustota má tvar

$$\prod_{i=1}^{n} p(x_i - a) = \mathrm{e}^{-\sum_{i=1}^{n}(x_i - a)} \mathrm{e}^{-\sum_{i=1}^{n} \mathrm{e}^{(x_i - a)}}.$$

To znamená, že postačující statistika pro parametr polohy a je $\sum_{i=1}^{n} \mathrm{e}^{x_i}$.

Následující příklad je převzatý z [4].

Příklad 6. Nechť $D \subset R^n$ je možina na Euklidovském prostoru, $x \in R^n$ a $\theta \in \Theta$ kde Θ je množina parametru. Pro $\forall \theta \in \Theta$ je hustota $p(x, \theta)$ spojitá a $p(x, \theta) > 0$ pro $x \in D$. Uvažujme zobrazení

$$x \to g_x(\theta) = \ln p(x; \theta) - \ln p(x; \theta_0).$$

Ke každému $x \in D$ zkonstruujme funkci $g_x(\theta)$, kde θ probíhá celý parametrický prostor Θ a θ_0 je pevné. Toto zobrazení je postačující statistika

$$p(x; \theta) = \exp[g_x(\theta) \cdot g_x(\theta_0)].$$

Závěr

Předkládaná práce se zabývala významnou problematikou matematické statistiky, která se snaží vyčerpat maximální informaci z určitého náhodného výběru. Zavedli jsme pojem postačující statistiky a rozlišili jsme triviální, netriviální a minimální postačující statistiky. Pomocí Halmasovi a Savagovi faktorizační věty jsme ukázali jakým způsobem pro daný parametr spočítat postačující statistiku.

Poté jsme se zaměřili na jednorozměrné rozdělení, jejichž určité mocniny připouštějí netriviální postačující statistiky. Dospěli jsme k důležitému výsledku, podle kterého exponenciální rodina hustot má takovou vlastnot, že připouští omezení pomocí postačující statistiky. Pomocí speciálních parametrů se dalo zformulovat exponenciální rodinu rozdělení přesněji a to v případech s parametrem polohy a s parametrem měřítka. V rámci poslední kapitoly jsme uvedli několik příkladu jak spočítat postačující a minimální postačící statistiky pro dané rozdělení.

Literatura

[1] Dynkin, E. B.: *Necessary and sufficient statistics for a family of probability distributions*, Addison-Wesley, Reading, 1985.

[2] Kagan, A. M., Linnik, Yu. V., Rao, C. P.: *Characterization problems in mathematical statistics*, Moscow, 1972.

[3] Lehmann, E. L.: *Testing Statistical Hypotheses*, New York, 1997.

[4] Linnik, Yu, V.: *Statistical problems with nuisance parameters*, Amer. Math. Soc., 1968, (Translated from Russian)