

Základné pojmy matematickej štatistiky

Matematická štatistika sa zaoberá štúdiom prijatia rozhodnutí v tzv. podmienkach neurčitosti, t.j. vtedy, keď určité rozhodnutie závisí od náhodnej udalosti, pričom nie je známe, ktorý zo zákonov rozdelenia pravdepodobnosti pôsobí v danom prípade.

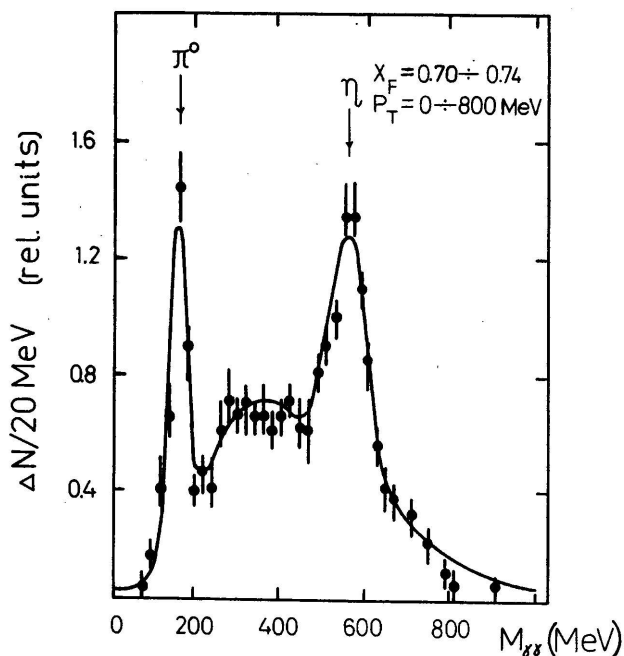
Uvažujme prípad, keď máme urobiť rozhodnutie závisiace od náhodnej udalosti, avšak nepoznáme ani súhrn podmienok γ určitého experimentu, ani to, ktorý zo zákonov pravdepodobnosti pôsobí v danom prípade. Vtedy je nutné realizovať experiment na základe meraní určitých veličín a za pomoci určitých metód urobiť príslušné závery.

Fyzikálny príklad

Uvažujme fyzikálny experiment zameraný na štúdium mezónov rozpadajúcich sa na 2 gama kvantá – napr.: $K^- p \rightarrow \gamma\gamma + X$.

- Invariantná hmotnosť $m_{\gamma\gamma} = (E_1 + E_2)^2 - (\vec{p}_1 + \vec{p}_2)^2$ bude predstavovať náhodnú veličinu charakterizujúcu náš experiment, pričom **jej odpovedajúci pravdepodobnostný zákon nepoznáme..**
- Realizácie tejto náhodnej veličiny sa získavajú opakovaním experimentu.
- Výsledkom experimentu je rozdelenie $m_{\gamma\gamma}$.

G.S. Bitsadze et al. / η -meson inclusive production



Obr. 1: spektrum invariantných hmotností $\gamma\gamma$ -parov charakterizujúce daný $\gamma\gamma$ -experiment

Predpoklad:

- Poznáme rozlíšenie nášho detektora, t.j. poznáme závislosť σ od hmotnosti M rozpadajúcej sa častice sa na $\gamma\gamma$.
- Poznáme hmotnosť prvej častice (π^0 -mezón), teda poznáme jemu zodpovedajúce rozdelenie hustoty pravdepodobnosti invariantnej hmotnosti $m_{\gamma\gamma}$.
- Nepoznáme hmotnosť druhej častice (η -mezón) a teda ani jemu zodpovedajúce rozdelenie hustoty pravdepodobnosti.

Možné úlohy, ktoré je treba riešiť metódami matematickej štatistiky:

1. Preveriť hypotézu, že prvou časticou je π^0 -mezón – teda pre známe rozdelenie hustoty pravdepodobnosti zistiť nakoľko je pravdepodobná hypotéza, že v experimente sa realizuje rozdelením $m_{\gamma\gamma}$ zodpovedajúce π^0 -mezónu.
2. Nájsť hustotu rozdelenia pravdepodobnosti pre druhú časticu, teda nájsť zákon pravdepodobnosti, ktorý popisuje jej prejavenie sa v experimente. Prakticky nájsť hmotnosť skúmanej častice, zodpovedajúcu σ rozdelenia $m_{\gamma\gamma}$, či sa jedná o jednu časticu a pod.

Poznámka: Vyššie spomínaný $\gamma\gamma$ -experiment prakticky delíme na 2 sub-experimenty: v rámci prvého ($m_{\gamma\gamma} < 300$ MeV) sa rieši **úloha 1** a v rámci druhého sub-experimentu ($m_{\gamma\gamma} > 300$ MeV) sa rieši **úloha 2**. Riešenie problému si vyžaduje prijať a testovať hypotézu o pozadí, t.j. o rozdelení $m_{\gamma\gamma}$ v prípade, že γ -kvantá nie sú geneticky spojené (nepochádzajú z rozpadu tej istej častice). Pozadie je možné preskúmať metódou MC, využívajúc pritom aj možnosť iteračného prístupu.

Niekoľko pojmov...

Náhodný výber

Predpokladajme, že určitý experiment sme n -krát opakovali za tých istých podmienok, a že sme získali n na sebe nezávislých pozorovaní $x_1, \dots, x_n$ určitej veličiny. Môžeme ich považovať za realizácie určitej náhodnej veličiny ξ charakterizujúcej náš prípad.

N -ticu $(x_1, \dots, x_n)$ môžeme považovať tiež za realizáciu vektorovej náhodnej premennej $\Xi = (\xi_1, \dots, \xi_n)$, ktorej zložky $\xi_i, i=1..n$ sú nezávislé a majú to isté rozdelenie pravdepodobnosti ako náhodná premenná ξ s distribučnou funkciou $F(x)$ charakterizujúca základný súbor.

Náhodný vektor $(\xi_1, \dots, \xi_n)$ sa nazýva **jednoduchým náhodným výberom** zo základného súboru. N -tica hodnôt $(x_1, \dots, x_n)$ sa nazýva realizáciou tohoto náhodného výberu.

Výberová funkcia.

Je to ľubovoľná reálna funkcia náhodného výberu Ξ teda $g(\xi_1, \dots, \xi_n)$ - je to tiež náhodná veličina.

Dôležité výberové funkcie. Nech náhodná premenná ξ má distribučnú funkciu $F(x)$ a nech $E(\xi) = a$ a $D(\xi) = \sigma^2$. Nech $\Xi = (\xi_1, \dots, \xi_n)$ je náhodný výber o rozsahu $n (>1)$ a nech výberová funkcia je:

$$a) \quad g(\xi_1, \dots, \xi_n) = \bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i, \text{ potom ju nazývame } \textbf{výberovým aritmetickým}$$

$$\textbf{priemerom}, \text{ pričom platí: } E(\bar{\xi}) = a, \quad D(\bar{\xi}) = \frac{\sigma^2}{n}$$

$$b) \quad h(\xi_1, \dots, \xi_n) = s^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2, \text{ potom ju nazývame } \textbf{výberovou disperziou},$$

$$\text{pričom platí: } E(s^2) = \frac{n-1}{n} \cdot \sigma^2.$$

Štatistický odhad

Výberová funkcia $g(\xi_1, \dots, \xi_n)$, ktorá má tú vlastnosť, že parametre jej rozdelenia súvisia s určitým parametrom θ základného súboru, nazývame **štatistickým odhadom** daného parametra.

Štatistický odhad, ktorého stredná hodnota sa rovná hodnote odhadovaného parametra sa nazýva **neodchýleným štatistickým odhadom**

Kvantil.

Nech $F(x)$ je distribučná funkcia spojitej náhodnej premennej ξ a nech je dané číslo $\alpha \in (0,1)$. Ak číslo x_α je riešením pravdepodobnostnej rovnice:

$$F(x_\alpha) = \alpha \quad \text{t.j.} \quad P(\xi < x_\alpha) = \alpha \quad (2.1),$$

potom x_α nazývame **$\alpha \times 100$ %-ným kvantilom**.

Testovanie hypotéz

Interval spoľahlivosti pre odhad strednej hodnoty

Predpoklad:

experiment je charakterizovaný náhodnou premennou ξ s normálnym rozdelením $N(x, a, \sigma^2)$. Pritom poznáme $\sigma^2 (= \sigma_0^2)$ avšak nepoznáme a .

Realizujeme n -krát experiment s výsledkami $x_1, x_2, \dots, x_n$, potom:

- $(x_1, x_2, \dots, x_n)$ predstavuje realizáciu náhodného výberu $\Xi = (\xi_1, \xi_2, \dots, \xi_n)$;
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ je realizácia výberového aritmetického priemeru $\bar{\xi}$.

Odhad strednej hodnoty a je možné urobiť na základe realizácie $\bar{\xi}$ riešením pravdepodobnostnej rovnice:

$$P(|a - \bar{\xi}| \geq k_\alpha) = \alpha \quad (2.2)$$

kde $\alpha \in (0,1)$ je dané číslo a hľadáme k_α .

Z (2) vyplýva:

$$P(a - k_\alpha < \bar{\xi} < a + k_\alpha) = 1 - \alpha \quad (2.2b)$$

$\Downarrow$

$$N\left(a + k_\alpha, a, \frac{\sigma^2}{n}\right) - N\left(a - k_\alpha, a, \frac{\sigma^2}{n}\right) = 1 - \alpha \quad (2.2c)$$

$\Downarrow$

$$N\left(a + k_\alpha, a, \frac{\sigma^2}{n}\right) = 1 - \frac{\alpha}{2} \quad (2.2d)$$

$$N\left(\frac{k_{\alpha}\sqrt{n}}{\sigma}, 0, 1\right) = 1 - \frac{\alpha}{2} \quad (2.3)$$

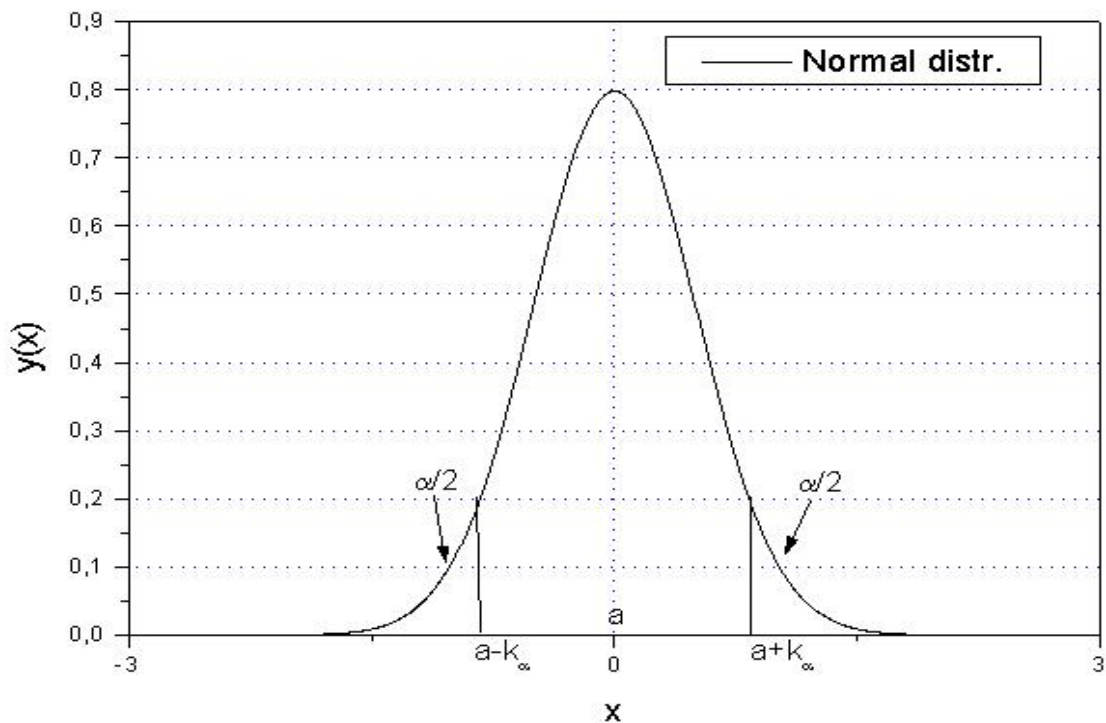
Rovnica (2.3) predstavuje rovnicu pre $(1-\alpha/2)\times 100\%$ kvantil.

Označme si:

$$u_{1-\frac{\alpha}{2}} = \frac{k_{\alpha}\sqrt{n}}{\sigma} \quad (2.4)$$

potom pre **interval spoľahlivosti pre parameter a** s koeficientom spoľahlivosti $1-\alpha$ dostávame:

$$\begin{aligned} |\bar{\xi} - a| &< k_{\alpha} = \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} \\ \bar{\xi} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} &< a < \bar{\xi} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha/2} \end{aligned} \quad (2.5)$$



Obr. 2: Interval spoľahlivosti pre odhad parametra a s koeficientom spoľahlivosti $1-\alpha$

Odhad disperzie a jej interval spoľahlivosti

Nech $\tilde{S}^2$ je neodchýlený odhad disperzie náhodnej premennej ξ charakterizujúcej náš experiment (základný súbor), potom náhodná premenná

$$\Psi = \frac{(N-1)\tilde{S}^2}{\sigma^2} = \sum_{i=1}^N \frac{(\xi_i - \bar{\xi})^2}{\sigma^2}$$

má rozdelenie χ^2 s $(N-1)$ stupňami voľnosti (je tam $N-1$ a nie N lebo premenné ξ_i sú viazané jednou podmienkou ($\bar{\xi} = \frac{1}{N} \sum_{i=1}^N \xi_i$)).

Interval spoľahlivosti pre σ^2 :

$$P\left\{a < \frac{(N-1)\tilde{S}^2}{\sigma^2} < b\right\} = \int_a^b P_{\eta}^{(N-1)}(x) dx = \gamma$$

Z čoho vyplýva:

$$P\left\{\frac{(N-1)\tilde{S}^2}{b} < \sigma^2 < \frac{(N-1)\tilde{S}^2}{a}\right\} = \gamma$$

Teda s pravdepodobnosťou γ sa disperzia σ^2 nachádza v intervale

$$\left(\frac{(N-1)\tilde{S}^2}{b}, \frac{(N-1)\tilde{S}^2}{a}\right).$$

Empirická distribučná funkcia

Nech ξ je náhodná premenná, $F_{\xi}(x)$ jej distribučná funkcia a $x_1, \dots, x_N$ sú jej nezávislé realizácie. Označme si $\nu_N(x)$ počet realizácií x_k , pre ktoré platí: $x_k < x$. Potom **empirická distribučná funkcia** je definovaná nasledovne:

$$F_N^*(x) = \frac{\nu_N(x)}{N}$$

Je zrejmé, že platí:

$$\lim_{N \rightarrow \infty} F_N^*(x) = F_{\xi}(x) \text{ dokonca } F_N^*(x) \Rightarrow F_{\xi}(x)$$

Teda empirická distribučná funkcia v limite $N \rightarrow \infty$ prechádza na $F_{\xi}(x)$ (dokonca rovnomerne k nej konverguje).

Kritérium blízkosti F_N^* a F_{ξ} . Ako kritérium blízkosti môžeme použiť rozdelenie χ^2 - dôvod je nasledovný:

$$\Psi = N \sum_{i=1}^L \frac{\left(\int_{\Delta_i} dF_N^*(x) - \int_{\Delta_i} dF_{\xi}(x) \right)^2}{\int_{\Delta_i} dF_{\xi}(x)}$$

Kde Δ_i $i=1, \dots, L$ sú intervaly, na ktoré sme rozbili obor hodnôt realizácií náhodnej premennej ξ . Pre $N \rightarrow \infty$ náhodná premenná Ψ predstavuje χ^2 -rozdelenie s $L-1$ stupňami voľnosti.

Metóda najmenších štvorcov

Pri analýze dát sa často stáva, že máme model, ale nepoznáme jeho všetky parametre, teda forma hustoty rozdelenia pravdepodobnosti (h.r.p.) je známe no nepoznáme všetky jej parametre. Napr. Skúmaná náhodná veličina má normálne rozdelenie, avšak nepoznáme jej strednú hodnotu resp. disperziu.

Predpokladajme, že v bodoch $x_1, \dots, x_n$ sme merali nezávislé náhodné veličiny $y_1, \dots, y_n$ s disperziami $\sigma_i^2, i=1..n$. Teda $\vec{x}$ je dané (nezávisle premenná) a $\vec{y}$ je náhodný výber, ktorý chceme použiť na odhad závislosti $E(y)=f(x)$. Nech teoretický model predpovedá závislosť $f(x; \vec{\theta})$ kde $\vec{\theta}$ je súbor L parametrov, ktoré chceme odhadnúť.

Parametre $\vec{\theta}$ sa nájdu minimalizáciou nasledovnej veľičiny:

$$L_{sq} = \sum_{i=1}^n \left(\frac{y_i - f(x_i, \vec{\theta})}{\sigma_i} \right)^2 = \min \quad (2.6)$$

Vo všeobecnosti merané veličiny môžu byť skorelované. V tomto prípade váhovou maticou (namiesto $1/\sigma_i^2$ v prípade nekorelovaných veličín) je inverzná kovariačná matica:

$$L_{sq} = \sum_{i,j=1}^n \left((y_i - f(x_i, \vec{\theta})) \cdot V_{i,j}^{-1} \cdot (y_j - f(x_j, \vec{\theta})) \right) \quad (2.7)$$

alebo vo vektorovom formalizme:

$$L_{sq} = (\vec{y} - \vec{f})^T \cdot V^{-1} \cdot (\vec{y} - \vec{f}) \quad (2.8)$$

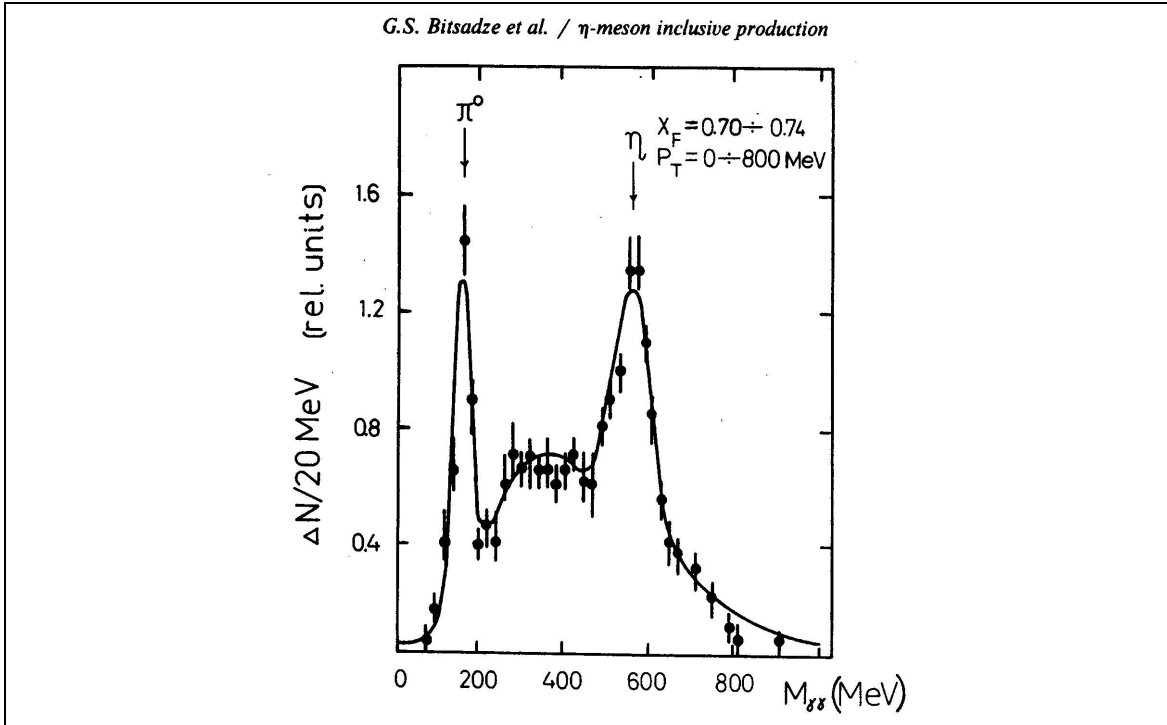
kde

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad \vec{f} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_n) \end{pmatrix} \quad V = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \dots & \sigma_{1,n}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 & \dots & \sigma_{2,n}^2 \\ \dots & \dots & \dots & \dots \\ \sigma_{n,1}^2 & \sigma_{n,2}^2 & \dots & \sigma_{n,n}^2 \end{bmatrix}$$

Podmienka minima, z ktorej vyplývajú rovnice pre hľadané parametre $\vec{\theta}$ sú:

$$\frac{\partial L_{sq}}{\partial \theta_1} = 0 \quad \frac{\partial L_{sq}}{\partial \theta_2} = 0 \quad \dots \quad \frac{\partial L_{sq}}{\partial \theta_L} = 0 \quad (2.9)$$

Príklad: Uvažujme vyššie spomínaný $\gamma\gamma$ -experiment. Invariantná hmotnosť $M_{\gamma\gamma}$



Obr. 3: Spektrum invariantných hmotností $m_{\gamma\gamma}$ daného $\gamma\gamma$ -experimentu obsahuje $\gamma\gamma$ -páry od rozpadu π^0 , η a neskorelované $\gamma\gamma$ -páry

geneticky spojených kvánt γ patrí buď π^0 -mezónu (normálne rozdelenie) alebo η -mezónu (normálne rozdelenie). Invariantná $M_{\gamma\gamma}$ geneticky nespojených kvánt γ patrí pozadiu (gamma funkcia). Náš model (hypotéza) vyzerá preto nasledovne:

$$f(x; \vec{\theta}) = c_1 \cdot G(x; m_\pi, \sigma_\pi) + c_2 \cdot G(x; m_\eta, \sigma_\eta) + (1 - c_1 - c_2) \cdot \Gamma(x; \alpha, \beta) \quad (2.10)$$

Vektor parametrov : $\vec{\theta} = (c_1, c_2, m_\pi, \sigma_\pi, m_\eta, \sigma_\eta, \alpha, \beta)$

Nezávisle premenná x_i : stredné hodnoty binov histogramu

Nezávisle náhodné premenné y_i : obsah jednotlivých binov histogramu.

Disperzie n.p. y_i : $\sigma_i^2 = y_i$

Príklad: Predpokladajme lineárny model ($f(x)=a+bx$) pre prípad súboru nekorelovaných bodov: $(x_i, y_i \pm \sigma_i) \quad i=1, \dots, n$
Potom

$$L_{SQ} = \sum_{i=1}^n \left(\frac{y_i - (a + bx_i)}{\sigma_i} \right)^2 \quad (2.11)$$

Z rovníc:

$$\frac{\partial L_{SQ}}{\partial a} = 0 \quad \text{a} \quad \frac{\partial L_{SQ}}{\partial b} = 0 \quad (2.12)$$

dostávame:

$$\begin{aligned} a &= \frac{S_y \cdot S_{xx} - S_x \cdot S_{xy}}{D} \\ b &= \frac{S_1 \cdot S_{xy} - S_x \cdot S_y}{D} \end{aligned} \quad (2.13)$$

kde

$$S_1 = \sum_i \frac{1}{\sigma_i^2} \quad S_x = \sum_i \frac{x_i}{\sigma_i^2} \quad S_{xx} = \sum_i \frac{x_i \cdot x_i}{\sigma_i^2} \quad (2.14)$$

$$S_y = \sum_i \frac{y_i}{\sigma_i^2} \quad S_{xy} = \sum_i \frac{x_i \cdot y_i}{\sigma_i^2} \quad D = S_1 \cdot S_{xx} - S_x^2$$

Disperziu (varianciu) parametrov parametrov $\vec{\theta}$ (v uvedenom prípade $\vec{\theta} = \begin{pmatrix} a \\ b \end{pmatrix}$)

získame nasledovne:

$$V(\vec{\theta}) = S \cdot V(\vec{y}) \cdot S^T \quad (2.15)$$

kde

$$S = \begin{pmatrix} \frac{\partial a}{\partial y_1} & \dots & \frac{\partial a}{\partial y_n} \\ \frac{\partial b}{\partial y_1} & \dots & \frac{\partial b}{\partial y_n} \end{pmatrix} \quad (2.16)$$

Konkrétne pre náš prípad:

$$V(\vec{\theta}) = \frac{1}{D} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \quad (2.17)$$

Metóda maximálnej pravdepodobnosti

Princíp: vyberáme takú hypotézu, ktorá dáva najväčšiu pravdepodobnosť pozorovaným udalostiam.

Uvažujme náhodnú premennú so spojitou h.r.p $p(x|\theta)$ kde θ je parameter ($x|\theta$ x dôrazňuje, že sa jedná o podmienenú pravdepodobnosť mať výsledok x pri hodnote parametra θ). Ak máme náhodný výber o rozsahu n , potom mu prislúcha h.r.p.:

$$P(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i | \theta) \quad (2.18)$$

Hľadaný parameter θ má pri daných realizáciách $x_1, \dots, x_n$ nasledovnú (podmienenú) pravdepodobnosť:

$$l(\theta | x) = \prod_{i=1}^n p(x_i | \theta) \quad (2.19)$$

$$\text{Log}(l) = L = \sum_{i=1}^n \text{Log}(p(x_i | \theta))$$

Hodnota parametra θ sa získava z podmienky maxima L :

$$\frac{\partial L}{\partial \theta} = 0 \quad (2.20)$$

Príklad:

Uskutočnili sme n pokusov s výsledkami $x_1, \dots, x_n$, pri čom náhodná premenná charakterizujúca náš experiment má normálne rozdelenie pravdepodobnosti, pričom nepoznáme jeho strednú hodnotu a disperziu:

$$p(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (2.21)$$

Pre logaritmicкую pravdepodobnosť (po vynechaní normovacích faktorov):

$$L = \log(l) = -\frac{n \log \sigma^2}{2} - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2} \quad (2.22)$$

Na základe:

$$\frac{\partial L}{\partial \mu} = 0 \quad \text{a} \quad \frac{\partial L}{\partial \sigma^2} = 0 \quad (2.23)$$

dostávame:

$$\hat{\mu} = \frac{\sum_i x_i}{n} \quad \hat{\sigma}^2 = \frac{\sum_i (x_i - \hat{\mu})^2}{n} \quad (2.24)$$