Využitie funkcie vierohodnosti na konštrukciu približných a exaktných konfidenčných oblastí[1]

Viktor WITKOVSKÝ

*Ústav merania, Slovenská akadémia vied, Bratislava
witkovsky@savba.sk

ROBUST 2010

Hora Matky Boží – Klášter Králíky

31.1.-5.2.2010

[1] Práca vznikla vďaka podpore grantov VEGA 1/0077/09, VEGA 2/0019/10 a APVV LPP-0388-09.

Abstrakt

Konštrukcia konfidenčných oblastí pre parametre modelu založených na funkcii vierohodnosti, resp. funkcii podielu vierohodností, je alternatívou k viac tradičnej metóde konštrukcie konfidenčných oblastí tzv. waldovského typu, ktorá vychádza z asymptotickej normality rozdelenia odhadov metódou maximálnej vierohodnosti (MLE).

Na rozdiel od MLE, práve postup konštrukcie vierohodnostných oblastí parametrov, založených na pozorovanej funkcii vierohodnosti, je v súlade s Fisherovým chápaním riešenia problému štatistického odhadovania, pozri Uusipaikka (2008).

Cieľom príspevku je ilustrovať konštrukciu približných a exaktných konfidenčných oblastí pre parametre modelu založených na funkcii vierohodnosti v niektorých špeciálnych prípadoch, s poukázaním na možné využitie zovšeobecnených pivotov, resp. fiduciálneho rozdelenia parametrov:

- pre parametre lineárneho regresného modelu,
- pre parametre jednoduchého modelu merania založeného na digitalizovaných meraniach,
- pre variančné komponenty v zmiešanom lineárnom modeli s dvomi variančnými komponentami.

Obsah

- **Základné pojmy:**
  - Koncept funkcie vierohodnosti
  - Interpretácia Fisherovho chápania teórie odhadovania
- **Vierohodnostná oblasť pre parametre lineárneho regresného modelu:**
  - Exaktný LR test
  - Exaktná konfidenčná oblasť pre parametre modelu
- **Zovšeobecnená inferencia:**
  - Zovšeobecnené pivoty (GPQ)
  - Fiduciálne rozdelenie
- **Vierohodnostná oblasť založená na fiduciálnom rozdelení resp. GPQ:**
  1. Vierohodnostná oblasť pre parametre lineárneho regresného modelu
  2. Vierohodnostná oblasť pre parametre modelu digitalizovaných meraní
  3. Vierohodnostná oblasť pre varianční komponenty v lineárnom zmiešanom modeli s dvomi variančnými komponentami

Označme

- napozorovanú odozvu: y, resp. y_{obs}, $y \in \mathcal{Y}$.
  y je realizácia náhodnej premennej Y, $\mathcal{Y}$ označuje výberový priestor,
- štatistický model: $\mathcal{M}$.
  V prípade parametrických štatistických modelov a za predpokladu existencie združenej hustoty pravdepodobnostného rozdelenia pre $y \in \mathcal{Y}$ — teda $f(y, \theta)$, kde θ označuje vektor parametrov, $\theta \in \Theta$, kde Θ je parametrický priestor — je uvažovaným modelom parametrizovaná trieda pravdepodobnostných rozdelení:

$$\mathcal{M} = \{f(y \mid \theta) : y \in \mathcal{Y}, \theta \in \Theta\}.$$

Koncept funkcie vierohodnosti bol vypracovaný R.A. Fisherom v období 1912-1921:

Označme $A \equiv A(Y)$ náhodnú udalosť vzťahujúcu sa k odozve Y, nech $A_{obs} \equiv A(y_{obs}) \subset \mathcal{Y}$. Funkcia vierohodnosti je definovaná ako:

$$L_{\mathcal{M}}(\theta \,|\, A = A_{obs}) = \Pr(A = A_{obs} \,|\, \theta).$$

- Ak model $\mathcal{M}$ pozostáva z triedy diskrétnych pravdepodobnostných rozdelení a $A_{obs} = \{y_{obs}\}$, potom

  $$L_{\mathcal{M}}(\theta \,|\, Y = y_{obs}) = f(Y = y_{obs} \,|\, \theta),$$

  kde $f(\cdot \,|\, \theta)$ označuje pravdepodobnostnú funkciu (pmf) diskrétnej náhodnej premennej Y.
- Ak $\mathcal{M}$ pozostáva z absolútne spojitých pravdepodobnostných rozdelení a $A(Y) = \{\tilde{y} \in \mathcal{Y} : Y - \delta < \tilde{y} < Y + \delta\}$ (jednorozmerný prípad), potom

  $$L_{\mathcal{M}}(\theta \,|\, Y = y_{obs}) = \Pr(A(Y) = A(y_{obs})) \approx f(y_{obs} \,|\, \theta) 2\delta,$$

  kde $f(\cdot \,|\, \theta)$ označuje hustotu náhodnej premennej Y (pdf).

Uvažujme transformáciu odozvy jednorozmernej Y pomocou hladkej invertovateľnej funkcie $h(\cdot)$: $Z = h(Y)$. Nech $g(\cdot)$ označuje inverznú funkciu k $h(\cdot)$.

- Ak $A(Z) = \{\tilde{z} \in \mathcal{Z} : Z - \delta < \tilde{z} < Z - \delta\}$ označuje náhodnú udalosť vzťahujúcu sa k odozve $Z = h(Y)$, potom

$$
\begin{aligned}
L_{\mathcal{M}}(\theta \,|\, Z = z_{obs}) &= \Pr(A(Z) = A(z_{obs})) \\
&= \Pr(z_{obs} - \delta < Z < z_{obs} + \delta) \\
&= \Pr(g(z_{obs} - \delta) < Y < g(z_{obs} + \delta)) \\
&\approx f(y_{obs} \,|\, \theta) \times |g'(h(y_{obs}))| \, 2\delta \\
&= c(y_{obs}) \times f(y_{obs} \,|\, \theta).
\end{aligned}
$$

  kde $c(y_{obs})$ nezávisí od parametra θ.

- Funkcia vierohodnosti je (aspoň približne) proporcionálna hodnote funkcie hustoty $f(y_{obs} \,|\, \cdot)$, ako funkcie parametra θ, teda $L_{\mathcal{M}}(\theta \,|\, Y = y_{obs}) \propto f(y_{obs} \,|\, \theta)$.
- Uvedená aproximácia môže zlyhať v špecifických situáciách vo vyššiedimenzionálnom parametrickom priestore, preto je v takýchto situáciach rozumné využívať exaktnú funkciu vierohodnosti $L_{\mathcal{M}}(\theta \,|\, A = A_{obs}) = \Pr(A = A_{obs} \,|\, \theta)$.

Fisher chápal funkciu podielu vierohodností (likelihood ratio), založenú na štatistickej evidencii $(y_{obs}, \mathcal{M})$,

$$LR(\theta_1, \theta_2 \,|\, y_{obs}) = \frac{L(\theta_1 \,|\, y_{obs})}{L(\theta_2 \,|\, y_{obs})},$$

ako relatívnu mieru toho ako pozorovaná odozva y_{obs} podporuje preferenciu hodnoty parametra θ_1 vzhľadom k hodnote θ_2 v rámci uvažovaného modelu $\mathcal{M}$.

- Funkcia podielu vierohodností nezávisí od funkcie $c(y_{obs})$ — preto namiesto funkcie vierohodnosti $L(\theta|y_{obs})$ možno pracovať s relatívnou funkciou vierohodnosti, ktorá nadobúda hodnoty v intervale $[0, 1]$:

  $$R(\theta \,|\, y_{obs}) = \frac{L(\theta \,|\, y_{obs})}{L(\hat{\theta} \,|\, y_{obs})},$$

  kde $\hat{\theta} \in \Theta$ je taká hodnota parametra, ktorá maximalizuje funkciu vierohodnosti. Keďže $\hat{\theta}$ nemusí existovať, presnejšie by bolo uvádzať v menovateli hodnotu $\sup_{\theta \in \Theta} L(\theta \,|\, y_{obs})$.

Fisher rozpracoval svoju teóriu odhadovania postupne v rokoch 1912-1956. Základné princípy tejto teórie možno zhrnuť v niekoľkých bodoch:

- **Odmietnutie apriórneho rozdelenia pre parametre:** Štatistická inferencia o neznámych parametroch by nemala závisieť od akéhokoľvek apriórneho rozdelenia parametrov, pokiaľ takéto apriórne rozdelenie nie je založené na reálnej (fyzikálnej) znalosti danej situácie.
- **Inariantnosť:** Štatistická inferencia by mala byť invariantná vzhľadom k jedno-jednoznačným transformáciam parametrov.
- **Efektívnosť:** Štatistická inferencia by mala byť efektívna v tom zmysle, že využíva optimálne všetku informáciu obsiahnutú v pozorovaniach.
- **Malé rozsahy výberov:** Štatistická inferencia by mala byť presne aplikovateľná pre malé (konečné) rozsahy pozorovaní.

Základné pojmy

Interpretácia Fisherovej teórie odhadu

Fisher považoval za riešenie problému odhadovania súbor oblastí vierohodnosti odvodených od napozorovanej (relatívnej) funkcie vierohodnosti.

- Vierohodnostná oblasť (*likelihood-based region*) pre parameter θ je

$$\mathcal{R}(y_{obs}) = \{\theta : \theta \in \Theta, R(\theta \mid y_{obs}) \geq c(y_{obs})\}$$

  pre vhodne zvolenú konštantu (resp. funkciu) $c(y_{obs}) \in (0, 1)$.
- Fisherov výrok štatistickej inferencie (riešenie problému odhadovania) znie: Neznámy parameter (vektor) θ patrí do stanoveného intervalu (oblasti) s napozorovanou relatívnou vierohodnosťou väčšou ako $c(y_{obs})$.
- Problémom zostáva optimálne určenie konštanty c resp. funkcie $c(y_{obs})$ (c môže funkčne závisieť od y_{obs}), tak aby stanovená vierohodnostná oblasť pokrývala neznámy parameter s vopred určenou pravdepodobnosťou, vzhľadom na variabilitu možných výsledkov štatistického experimentu $y \in \mathcal{Y}$.
- Fisher teda hodnotu $\hat{\theta}$, t.j. bodový odhad MLE (odhad metódou maximálnej vierohodnosi), nepovažoval za riešenie problému odhadovania. Navyše, konfidnčné oblasti waldovského typu, odvodené na základe asymptotického rozdelenia odhadu MLE (*MLE-based region*), nemožno považovať za riešenie v súlade s Fisherovým chápaním riešenia problému odhadovania.

Nech $\psi = g(\theta)$ je q-rozmerná reálna funkcia parametrov – *parameter of interest*. Konštrukciu vierohodnostných oblastí pre ψ možno korektne definovať pomocou napozorovanej profilovej funkcie vierohodnosti pre funkciu $g(\theta) = \psi$:

- Profilová funkcia vierohodnosti pre funkciu parametrov $g(\cdot)$ je

$$L_g(\psi \,|\, y_{obs}) = \sup_{\theta \in \Theta : g(\theta) = \psi} L(\theta \,|\, y_{obs}) = L(\hat{\theta}_\psi \,|\, y_{obs}),$$

  kde $\hat{\theta}_\psi = \arg\max_{\theta \in \Theta : g(\theta) = \psi} L(\theta \,|\, y_{obs})$. Funkcia

$$\ell_g(\psi \,|\, y_{obs}) = \log(L_g(\psi \,|\, y_{obs}))$$

  je profilová *log-likelihood* funkcia.

- Relatívna profilová vierohodnostná funkcia a relatívna profilová *log-likelihood* funkcia sú definované adekvátne:

$$R_g(\psi \,|\, y_{obs}) = \frac{L_g(\psi \,|\, y_{obs})}{L_g(\hat{\psi} \,|\, y_{obs})} = \frac{L_g(\psi \,|\, y_{obs})}{L(\hat{\theta} \,|\, y_{obs})}$$

$$r_g(\psi | y_{obs}) = \ell_g(\psi \,|\, y_{obs}) - \ell_g(\hat{\psi} \,|\, y_{obs}) = \ell_g(\psi \,|\, y_{obs}) - \ell(\hat{\theta} \,|\, y_{obs}).$$

Základné pojmy

Formulácia problému odhadovania pre funkciu parametrov $g(\theta)$

- Vierohodnostná oblasť pre $\psi = g(\theta)$ je definovaná ako množina

$$\mathcal{R}_g(y_{obs}) = \{\psi : R_g(\psi \,|\, y_{obs}) \geq c(y_{obs})\} = \{\psi : r_g(\psi \,|\, y_{obs}) \geq \log(c(y_{obs}))\}.$$

- Predpokladajme platnosť štatistického modelu $\mathcal{M} = \{f(y \,|\, \theta) : y \in \mathcal{Y}, \theta \in \Theta\}$. Ak možno určiť konštantu (funckiu) $c(\cdot)$ tak, že pre každdú (skutočnú) hodnotu parametra $\theta^\star \in \Theta$ platí

$$\Pr(\mathcal{R}_g(Y) \ni \theta^\star) = \Pr(R_g(g(\theta^\star) \,|\, Y) \geq c(Y)) = 1 - \alpha,$$

  kde $(1 - \alpha)$, $\alpha \in (0, 1)$, je vopred pevne stanovená pravdepodobnostná hladina (*confidence level*), potom vierohodnostná oblasť $\mathcal{R}_g(Y) \equiv \mathcal{R}_{g,1-\alpha}(Y)$ je presnou $(1 - \alpha)$-konfidenčnou oblasťou pre $g(\theta)$.

- Pokiaľ uvedený vzťah paltí približne pre každdú (skutočnú) hodnotu parametra $\theta^\star \in \Theta$, teda

$$\Pr(\mathcal{R}_g(Y) \ni \theta^\star) = \Pr(R_g(g(\theta^\star) \,|\, Y) \geq c(Y)) \approx 1 - \alpha,$$

  hovoríme o približnej $(1 - \alpha)$-konfidenčnej oblasti pre $g(\theta)$.

- Vierohodnostná oblasť $\mathcal{R}_{g,1-\alpha}(y_{obs})$, určená na základe štatistickej evidencie $(y_{obs}, \mathcal{M})$, je realizáciou $(1 - \alpha)$-konfidenčnej oblasti $\mathcal{R}_{g,1-\alpha}(Y)$ pre $\psi = g(\theta)$.

Vo všeobecnosti, približné $(1-\alpha)$-konfidenčné oblasti založené na funkcii vierohodnosti možno určiť na základe výsledku o asymptotickom rozdelení štatistiky logaritmu podielu vierohodností $r_g(\psi \mid Y)$:

- Za pomerne všeobecných predpokladov o funkcii $g(\cdot)$ a o štatistickom modeli $\mathcal{M}$ platí tvrdenie, pozri napr. Fan et al. (JASA 2000):

  $$-2r_g(\psi \mid Y) \overset{n\to\infty}{\sim} \chi^2_q,$$

  kde χ^2_q označuje chi-kvadrát rozdelenie s q-stupňami voľnosti, kde q je dimenzia parametra $\psi = g(\theta)$.

  Pokiaľ $g(\theta) = \theta$, potom $q = d$, kde d je dimenzia parametra $\theta \in \Theta \subseteq R^d$.

- Realizácia približnej $(1-\alpha)$-konfidenčnej oblasti pre $g(\theta)$ je potom určená ako

  $$\mathcal{R}_{g,1-\alpha}(y_{obs}) = \{\psi : r_g(\psi|y_{obs}) \geq -\chi^2_{q,1-\alpha}/2\} = \{\psi : -2r_g(\psi \mid y_{obs}) \leq \chi^2_{q,1-\alpha}\},$$

  kde $\chi^2_{q,1-\alpha}$ je $(1-\alpha)$-kvantil chi-kvadrát rozdelenie s q-stupňami voľnosti.

Uvažujme lineárny regresný model s normálne rozdelenými chybami:

$$Y = \boldsymbol{X}\beta + \sigma Z,$$

kde

- Y je n-rozmerný náhodný vektor pozorovaní, y označuje realizáciu Y,
- $\boldsymbol{X}$ je $n \times k$ nestochastická matica plnej hodnosti,
- β je k-rozmerný vektor regresných parametrov,
- σ je štandardná odchýlka chýb, $\sigma > 0$,
- Z je n-rozmerný štandardizovaný náhodný vektor chýb, $Z \sim N(0, I_n)$.

Najskôr, budeme uvažovať problém testovania hypotézy

$$H_0 : (\beta, \sigma) = (\beta_0, \sigma_0) \text{ oproti alternatíve } H_1 : (\beta, \sigma) \neq (\beta_0, \sigma_0).$$

pričom logaritmus vierohodnostnej funkcie (*log-likelihood*) je

$$\ell_{\mathcal{M}}((\beta, \sigma) \,|\, y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - \boldsymbol{X}\beta)'(y - \boldsymbol{X}\beta).$$

Nech

$$\lambda((\beta,\sigma)\,|\,y) = -2r_{\mathcal{M}}((\beta,\sigma)\,|\,y),$$

$$r_{\mathcal{M}}((\beta,\sigma)\,|\,y) = \ell_{\mathcal{M}}((\beta,\sigma)\,|\,y) - \ell_{\mathcal{M}}((\hat\beta,\hat\sigma)\,|\,y),$$

kde

- $\hat\beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$ — MLE odhad parametra β,
- $\hat\sigma$ — MLE odhad parametra σ, t.j. $\hat\sigma = \sqrt{(y-\boldsymbol{X}\hat\beta)'(y-\boldsymbol{X}\hat\beta)/n}$.

Odtiaľ teda

$$\lambda((\beta,\sigma)\,|\,y) = \frac{1}{\sigma^2}(y-\boldsymbol{X}\beta)'(y-\boldsymbol{X}\beta) - n\log\left(\frac{\hat\sigma^2}{\sigma^2}\right) - n,$$

Uvažovaná trieda rozdelení spĺňa podmienky regularity. Za platnosti nulovej hypotézy $H_0 : (\beta,\sigma) = (\beta_0,\sigma_0)$ dostávame $\lambda((\beta_0,\sigma_0)\,|\,Y) \overset{n\to\infty}{\sim} \chi^2_{k+1}$.
Realizácia približnej $(1-\alpha)$-konfidenčnej oblasti pre (β,σ) je potom určená ako

$$\mathcal{R}_{1-\alpha}(y) = \{(\beta,\sigma) : \lambda((\beta,\sigma)\,|\,y) \le \chi^2_{k+1,1-\alpha}\},$$

kde $\chi^2_{k+1,1-\alpha}$ je $(1-\alpha)$-kvantil chi-kvadrát rozdelenie s $(k+1)$-stupňami voľnosti.

Za daných predpokladov a za platnosti nulovej hypotézy H_0, možno jednoducho odvodiť exaktné rozdelenie tesovacej štatistiky $\lambda((\beta_0,\sigma_0)\,|\,Y)$:

$$\begin{aligned}
\lambda((\beta_0,\sigma_0)\,|\,Y) &\sim \frac{1}{\sigma_0^2}(Y-\boldsymbol{X}\beta_0)'(Y-\boldsymbol{X}\beta_0) - n\log\left(\frac{(Y-\boldsymbol{X}\beta_0)'M_{\boldsymbol{X}}(Y-\boldsymbol{X}\beta_0)}{n\sigma_0^2}\right) - n \\
&\sim \frac{1}{\sigma_0^2}(\sigma_0^2 Z'Z) - n\log\left(\frac{\sigma_0^2 Z'M_{\boldsymbol{X}}Z}{n\sigma_0^2}\right) - n \\
&\sim Z'Z - n\log\left(Z'M_{\boldsymbol{X}}Z\right) + n\left(\log(n)-1\right) \\
&\sim Z'(P_{\boldsymbol{X}}+M_{\boldsymbol{X}})Z - n\log\left(Z'M_{\boldsymbol{X}}Z\right) + n\left(\log(n)-1\right) \\
&\sim W_k + W_{n-k} - n\log\left(W_{n-k}\right) + n\left(\log(n)-1\right),
\end{aligned}$$

kde

- $P_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$,
- $M_{\boldsymbol{X}} = I_n - P_{\boldsymbol{X}}$,
- $Z \sim N(0, I_n)$,
- $W_k \sim \chi^2_k$ a $W_{n-k} \sim \chi^2_{n-k}$ sú navzájom nezávislé náhodné premenné.

Rozdelenie testovacej štatistiky $\lambda((\beta_0, \sigma_0) \,|\, Y)$ závisí len od rozsahu výberu n a hodnosti matice k (nezávisí od konkrétnej hodnoty matice $\boldsymbol{X}$).

LR test zamietá nulovú hypotézu H_0 pre veľké hodnoty $\lambda((\beta_0, \sigma_0) \,|\, y)$, teda pre

$$\lambda((\beta_0, \sigma_0) \,|\, y) > \lambda_{1-\alpha},$$

kde $\lambda_{1-\alpha}$ označuje $(1-\alpha)$-kvantil rozdelenia náhodnej premennej $\lambda((\beta_0, \sigma_0) \,|\, Y)$. Tabuľky kritických hodnôt možno nájsť v práci Chvosteková-Witkovský (MSR 2009).

Pre veľké rozsahy výberov platí:

$$\lambda_{1-\alpha} \overset{n\to\infty}{\longrightarrow} \chi^2_{k+1,1-\alpha},$$

kde $\chi^2_{k+1,1-\alpha}$ je $(1-\alpha)$-kvantil chi-kvadrát rozdelenie s $(k+1)$-stupňami voľnosti.

Invertovaním LR testu dostávame realizáciu exaktnej $(1-\alpha)$-konfidenčnej oblasti pre (β, σ):

$$\mathcal{R}_{1-\alpha}(y) = \{(\beta, \sigma) : \lambda((\beta, \sigma) \,|\, y) \leq \lambda_{1-\alpha}\}.$$

- Metódy zovšeobecnenej inferencie boli navrhnuté ako metódy približnej štatistickej inferencie (o špecifických parametroch záujmu) v situáciách, keď sa v modeli vyskytujú rušivé parametre a klasické metódy štatistickej inferencie v týchto situáciach neexistujú, alebo sú nevhodné.
- Pojem *zovšeobecnenej testovacej premennej* a *zovšeobecnenej p-hodnoty* zaviedli Tsui a Weerahandi (JASA, 1989).
- Hoci invertovaním zovšeobecnených testov možno konštruovať aj (zovšeobecnené) konfidenčné intervaly (oblasti), Weerahandi (JASA, 1993) definoval pre takúto situáciu vhodnejší pojem *zovšeobecnený pivot* (GPQ — Generalized Pivotal Quantity).
- Hannig et al. (JASA, 2006) zaviedli pojem *fiduciálny zovšeobecnený pivot* a odhalili spojitosť zovšeobecnenej inferencie a fiduciálnym argumetom, ktorý zaviedol Fisher (1935).

Nech Y je pozorovateľný náhodný vektor, ktorého rozdelenie závisí od (vektorového) parametra θ. Nech $\psi = g(\theta)$ je parameter záujmu.

Nech Y^* je náhodný vektor s rovnakým rozdelením ako Y, stochasticky nezávislý od Y.

Nech $GPQ(Y, Y^*, \theta)$ označuje náhodnú premennú, ktorá je funckiou Y, Y^* a θ.

Požiadavky na vlastnosti funkcie $GPQ(Y, Y^*, \theta)$:

1. Podmienené rozdelenie náhodnej premennej $GPQ(Y, Y^*, \theta \mid Y = y)$ nezávisí od θ.
2. Pre každú realizáciu y, $GPQ(y, y, \theta)$ závisí (funkčne) od θ len cez $\psi = g(\theta)$.
3. Pre každú realizáciu y, $GPQ(y, y, \theta) = \psi = g(\theta)$.

- Pokiaľ $GPQ(Y, Y^*, \theta)$ spĺňa požiadavky 1 a 2, potom $GPQ(Y, Y^*, \theta)$ nazývame zovšeobecneným pivotom pre parameter $\psi = g(\theta)$.
- Pokiaľ $GPQ(Y, Y^*, \theta)$ spĺňa požiadavky 1 a 3, potom $GPQ(Y, Y^*, \theta)$ nazývame fiduciálnym zovšeobecneným pivotom pre parameter $\psi = g(\theta)$.

Príklad: Lineárny regresný model
Nech $S^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n-k) \Rightarrow (n-k)S^2/\sigma^2 \sim \chi^2_{n-k}$.

- Fiduciálny zovšeobecnený pivot pre parameter σ:

$$GPQ(S^2, S^{2*}, (\beta, \sigma)) = \sqrt{\frac{(n-k)S^2}{(n-k)S^{2*}/\sigma^2}}$$

$$GPQ(s^2_{obs}, S^{2*}, (\beta, \sigma)) \sim \sqrt{\frac{(n-k)s^2_{obs}}{W^\star_{n-k}}}, \text{ kde } W^\star_{n-k} \sim \chi^2_{n-k}$$

$$GPQ(s^2_{obs}, s^2_{obs}, (\beta, \sigma)) = \sqrt{\frac{(n-k)s^2_{obs}}{(n-k)s^2_{obs}/\sigma^2}} = \sigma,$$

- Rozdelenie náhodnej premennej $\tilde{\sigma} = GPQ(s^2_{obs}, S^{2*}, (\beta, \sigma)) \sim \sqrt{\frac{(n-k)s^2_{obs}}{W^\star_{n-k}}}$ je v tomto prípade totožné s marginálnym fiduciálnym rozdelením pre parameter σ.
- Podobne možno určit GPQ a náhodnú premennú $\tilde{\beta} = \hat{\beta}_{obs} - \tilde{\sigma}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Z^*$ s marginálnym fiduciálnym rozdelením pre parameter β, kde $Z^* \sim N(0, I_n)$ je náhodný vektor nezávislý od $W^\star_{n-k} \sim \chi^2_{n-k}$.

Fiduciálny argument je založený na výmene úlohy parametrov a dát. Uvádzame tu stručný popis konštrukcie fiduciálnej distribúcie podľa návrhu z prác J. Hanniga (2006, 2009):

- Nech
$$Y = G(\theta, Z)$$
je známy systém štrukturálnych rovníc (dáta generujúci mechanizmus), ktorý spája náhodný vektor pozorovaní Y s vektorom parametrov $\theta \in \Theta$ a náhodným vektorom Z (chybový - šumový - proces), so známou distribučnou funkciou, ktorá nezávisí od parametrov.
- Pre ľubovoľnú (pozorovanú) hodnotu y náhodného vektora Y a pre realizáciu z vektora Z definujeme množinovo-hodnotovú funkciu
$$Q(y, z) = \{\theta : y = G(\theta, z)\}.$$
Funkcia $Q(Y, Z)$ reprezentuje 'inverznú funkciu' ku G. Pre ľubovoľnú pevnú hodnotu y a z je výsledkom $Q(y, z)$ buď prázdna množina, jednoprvková množina obsahujúca jednoznačný parameter θ, alebo zložitejšia mnohoprvková množina možných parametrov θ, ktorá závisia od vzťahov medzi y, z a G.

- Nech $V(\cdot)$ definuje akýkoľvek náhodný mechanizmus na merateľných množinách. Pre ľubovoľnú množinu S, $V(S)$ náhodne generuje prvok z uzáveru množiny S.

- Potom, špeciálny výber (v závislosti od G, Q, a V) zovšeobecnenej fiduciálnej distribúcie parametra θ, označme ju $F_\theta(\theta^\star)$ pre pevné $\theta^\star$, je definovaný ako akákoľvek verzia podmienenej distribúcie náhodnej premennej $V(Q(y, Z^\star))$ za podmienky $Q(y, Z^\star) \neq \emptyset$, teda

$$F_\theta(\theta^\star) = \Pr\left(V(Q(y, Z^\star)) \leq \theta^\star | Q(y, Z^\star) \neq \emptyset\right),$$

  kde y napozorovaná hodnota náhodného vektora Y a $Z^\star$ je nezávislá náhodná premenná s rovnakým rozdelením ako Z.

- Úloha nájsť explicitný tvar fiduciálnej distribúcie môže byť veľmi zložitá. Avšak, uvedený postup dáva priamy návod na generovanie náhodného výberu parametrov z danej fiduciálnej distribúcie.

- Uvažujme špeciálny prípad lineárneho regresného modelu — náhodný výber $Y_1, Y_2, \ldots, Y_n$, kde
$$Y_i = \mu + \sigma Z_i,\ \text{pričom } Z_i \sim N(0,1).$$

- Pre postačujúce štatistiky $\bar{Y}$ a S_Y^2 platí $\bar{Y} = \mu + (\sigma/\sqrt{n})Z$, kde $Z \sim N(0,1)$ a $S_Y^2 = \sigma^2 W/(n-1)$, kde $W \sim \chi^2_{(n-1)}$, pričom sú navzájom nezávislé.

- Štrukturálne rovnice definujeme v tvare $\left(\bar{Y}, S_Y^2\right) = G((\mu,\sigma),(Z,W))$, teda
$$G((\mu,\sigma),(Z,W)) = \left(\mu + (\sigma/\sqrt{n})Z, \sigma^2 W/(n-1)\right).$$

- Inverznú funkciu $Q(\cdot,\cdot)$ definujeme ako

$$Q\left((\bar{y}, s_Y^2), (z, w)\right) = \left\{\left(\bar{y} - \sqrt{\frac{(n-1)s_Y^2}{nw}}\, z, \sqrt{\frac{(n-1)s_Y^2}{w}}\right)\right\},$$

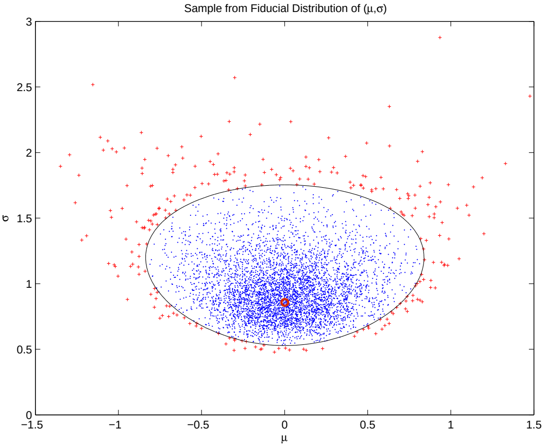
kde $\bar{y}$ je pozorovaná hodnota (realizácia) $\bar{Y}$ a s_Y^2 je realizácia S_Y^2. Tu $Q(\cdot,\cdot)$ vždy definuje neprázdnu množinu, ktorá obsahuje jedinú hodnotu z parametrického priestoru.

Teda fiduciálna distribúcia nezávisí od žiadneho náhodného mechanizmu $V(\cdot)$:

$$F_{(\mu,\sigma)}((\mu^\star, \sigma^\star)) = \Pr\left(\bar{y} - \sqrt{\frac{(n-1)s_Y^2}{nW^\star}}\, Z^\star \leq \mu^\star, \sqrt{\frac{(n-1)s_Y^2}{W^\star}} \leq \sigma^\star\right).$$

Toto rozdelenie je totožné s Fisherovým fiduciálnym rozdelením, ktoré odvodil v práci Fisher (1935).

Príklad: Náhodný výber rozsahu $N = 5000$ *zo združenej fiduciálnej distribúcie parametrov* μ *a* σ *založený na* $n = 10$ *nezávislých meraniach* $Y_i \sim N(0,1)$, $i = 1,\dots,10$, *pričom* $\bar{y} = 0.0013$ *a* $s_Y^2 = 0.9034^2$.



Vierohodnostná oblasť založená na fiduciálnom rozdelení resp. GPQ

Nech $\lambda(\theta \mid y_{obs}) = -2r_{\mathcal{M}}(\theta \mid y_{obs})$ označuje $(-2)\times$ relatívnu logaritmickú funkciu vierohodnosti — *log-likelihood funkciu* pre napozorovanú odozvu $y_{obs} \in \mathcal{Y}$, teda realizáciu náhodného vektora Y.

- Náhodnú premennú $\lambda(\theta \mid Y)$ budeme nazývať *výberová log-likelihood funkcia* pre θ.
- Nech $\tilde{\theta}$ označuje náhodnú premennú s fiduciálnym rozdelením pre parameter θ, teda $\tilde{\theta} \sim F_{\theta}$, resp. náhodný vektor s rozdelením odvodeným z podmieneného rozdelenia príslušných zovšeobecnených pivotov (GPQ).
- Náhodnú premennú $\lambda(\tilde{\theta} \mid y_{obs})$ budeme nazývať *zovšeobecnená (fiduciálna) log-likelihood funkcia* pre θ.
- Nech $\lambda_{1-\alpha}(y_{obs})$ označuje $(1-\alpha)$-kvantil rozdelenia náhodnej premennej $\lambda(\tilde{\theta} \mid y_{obs})$.
- Potom vierohodnostnú oblasť založenú na fiduciálnom rozdelení (resp. GPQ) pre θ, teda množinu

  $$\mathcal{R}_{1-\alpha}(y_{obs}) = \{\theta : \lambda(\theta \mid y_{obs}) \leq \lambda_{1-\alpha}(y_{obs})\},$$

  budeme považovať za realizáciu približnej $(1-\alpha)$-konfidenčnej oblasti pre θ.

Nech $y_{obs} \in \mathcal{Y}$ je napozorovaná odozva a nech $(\tilde\beta, \tilde\sigma)$ označuje náhodný vektor s fiduciálnym (GPQ) rozdelením pre (β, σ). Teda,

$$\tilde\sigma^2 \sim \frac{(y_{obs} - \boldsymbol{X}\hat\beta_{obs})'(y_{obs} - \boldsymbol{X}\hat\beta_{obs})}{W^\star_{n-k}} \sim \frac{(n-k)s^2_{obs}}{W^\star_{n-k}} \sim \frac{n\hat\sigma^2_{obs}}{W^\star_{n-k}},$$

$$\tilde\beta \sim \hat\beta_{obs} - \tilde\sigma(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Z^\star,$$

pričom $W^\star_{n-k} \sim \chi^2_{n-k}$ a $Z^\star \sim N(0, I_n)$ sú stochasticky nezávislé.
Potom rozdelenie *zovšeobecnenej (fiduciálnej) log-likelihood funkcie* $\lambda((\tilde\beta, \tilde\sigma) \,|\, y_{obs})$ je

$$\begin{aligned}
\lambda((\tilde\beta, \tilde\sigma) \,|\, y_{obs}) &\sim \frac{(y_{obs} - \boldsymbol{X}\tilde\beta)'(y_{obs} - \boldsymbol{X}\tilde\beta)}{\tilde\sigma^2} - n\log\left(\frac{\hat\sigma^2_{obs}}{\tilde\sigma^2}\right) - n \\
&\sim \frac{(y_{obs} - \boldsymbol{X}\hat\beta_{obs})'(y_{obs} - \boldsymbol{X}\hat\beta_{obs})}{\tilde\sigma^2} + \frac{\tilde\sigma^2 Z^{\star\prime}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Z^\star}{\tilde\sigma^2} - n\log\left(\frac{\hat\sigma^2_{obs}}{\tilde\sigma^2}\right) - n \\
&\sim W^\star_{n-k} + W^\star_k - n\log\left(W^\star_{n-k}\right) + n(\log(n) - 1),
\end{aligned}$$

kde $W^\star_k \sim \chi^2_k$ a $W^\star_{n-k} \sim \chi^2_{n-k}$ sú navzájom nezávislé náhodné premenné.

- Rozdelenie náhodnej premennej $\lambda((\tilde\beta, \tilde\sigma) \,|\, y_{obs})$ je totožné s rozdelením $\lambda((\beta, \sigma) \,|\, Y)$, teda oba prístupy vedú k rovnakým (exaktným) konfidenčným oblastiam pre (β, σ).

Uvažujme jednoduchý model nezávislých digitalizovaných meraní so známym parametrom rozlíšenia $\delta > 0$, neznámymi parametrami $\mu \in R^1$ (parameter polohy) a $\sigma > 0$ (štandardná odchýlka), pričom výsledkami meraní sú realizácie nezávislých náhodných veličín $Y = (Y_1, Y_2, \ldots, Y_n)$, kde

$$Y_i = \lfloor (\mu + \sigma Z_i)/\delta + 0.5 \rfloor \delta, \text{ kde } Z_i \sim N(0,1).$$

Rozdelenie pravdepodobnosti náhodnej premennej Y_i je určené pravdepodobnostnou funkciou $\{P_{k\delta}\}$, $k = 0, \pm 1, \pm 2, \ldots$, pričom $P_{k\delta} = \Pr(Y_i = k\delta)$ je

$$P_{k\delta} = \Phi\left([(k + 0.5)\delta - \mu]/\sigma\right) - \Phi\left([(k - 0.5)\delta - \mu]/\sigma\right),$$

kde $\Phi(\cdot)$ je distribučná funkcia rozdelenia $N(0,1)$.
Logaritmus vierohodnostnej funkcie pre (μ, σ) je

$$\ell((\mu,\sigma) \,|\, y_{obs}) = \sum_{i=1}^{n} \ln(P_{y_i\delta}) = \sum_{\{k_i : k_i = y_i/\delta\}} \ln(P_{k_i\delta}).$$

Pre tento model je $(-2\times)$ relatívna log-likelihood funkcia daná ako

$$\lambda((\mu,\sigma) \,|\, y_{obs}) = -2 \left(\ell((\mu,\sigma) \,|\, y_{obs}) - \sup_{(\mu,\sigma)} \ell((\mu,\sigma) \,|\, y_{obs}) \right),$$

pričom MLE odhad parametra (μ, σ) nemusí existovať.

V danom modeli nie je známe exaktné rozdelenie *výberovej log-likelihood funkcie* $\lambda((\mu,\sigma)\,|\,Y)$.

Vierohodnostnú oblasť pre parametre modelu digitalizovaných meraní možno preto konštruovať ako približnú $(1-\alpha)$-konfidenčnú oblasť pre (μ,σ):

1. Na základe predpokladu platnosti asymptotického rozdelenia, teda ako množinu

$$\mathcal{R}_{1-\alpha}(y_{obs}) = \{(\mu,\sigma) : \lambda((\mu,\sigma)\,|\,y_{obs}) \leq \chi^2_{2,1-\alpha}\},$$

   kde $\chi^2_{2,1-\alpha}$ je $(1-\alpha)$-kvantil chi-kvadát rozdelenia χ^2_2.

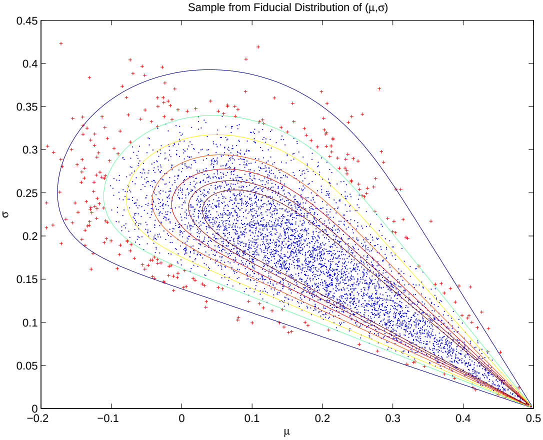
2. Na základe generovania realizácii náhodných premenných $(\tilde{\mu},\tilde{\sigma})$ z fiduciálneho rozdelenia pre (μ,σ), teda ako množinu

$$\mathcal{R}_{1-\alpha}(y_{obs}) = \{(\mu,\sigma) : \lambda((\mu,\sigma)\,|\,y_{obs}) \leq \hat{\lambda}_{1-\alpha}(y_{obs})\},$$

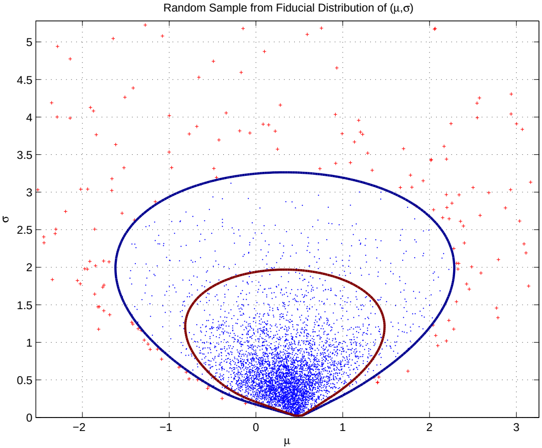
   kde $\hat{\lambda}_{1-\alpha}(y_{obs})$ je $(1-\alpha)$-kvantil odhadnutý z empirickej distribučnej funkcie *zovšeobecnenej fiduciálnej log-likelihood funkcie* $\lambda((\tilde{\mu},\tilde{\sigma})\,|\,y_{obs})$.

Metódy a algoritmy na generovanie realizácii $(\tilde{\mu},\tilde{\sigma})$ z fiduciálneho rozdelenia publikovali Hannig et al (2007), Wimmer Jr. (2009), Witkovský a Wimmer (2009).

Náhodný výber rozsahu N $= 5000$ *zo združenej fiduciálnej distribúcie parametrov* μ *a* σ *založený na* n $= [29, 1]$ *nezávislých pozorovaniach hodnôt* 0 *a* 1 *na prístroji s rozlíšením* $\delta = 1$.



Náhodný výber rozsahu N $= 5000$ *zo združenej fiduciálnej distribúcie parametrov* μ *a* σ *založený na n* $= [4, 1]$ *nezávislých pozorovaniach hodnôt* 0 *a* 1 *na prístroji s rozlíšením* $\delta = 1$*. Porovnanie empirickej a asymptotickej približnej* 95%*-konfidenčnej oblasti pre* (μ, σ)*.*



Uvažujme lineárny zmiešaný model

$$Y = \boldsymbol{X}\beta + \boldsymbol{Z}\eta + \varepsilon,$$

- Y — n-rozmerný vektor pozorovaní,
- $\boldsymbol{X}$ — $(n \times k)$ matica plánu,
- β — k-vektor regresných parametrov (pevné efekty),
- $\boldsymbol{Z}$ — $(n \times q)$ matica plánu,
- η — k-vektor náhodných efektov, $\eta \sim N(0, \sigma_A^2 I_q)$,
- ε — n-vektor náhodných chýb, $\varepsilon \sim N(0, \sigma^2 I_n)$.

Teda,

$$Y \sim N(\boldsymbol{X}\beta, \sigma_A^2 \boldsymbol{V} + \sigma^2 I_n), \text{ kde } \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{Z}'.$$

V prípade, že máme záujem len o variančné komponenty (σ_A^2, σ^2), inferencia môže byť založená na maximálnom invariante $Y^\star$:

$$Y^\star = \boldsymbol{B}Y, \text{ kde } \boldsymbol{B}'\boldsymbol{B} = M_{\boldsymbol{X}} \quad \& \quad \boldsymbol{B}\boldsymbol{B}' = I_\nu, \quad \nu = rank(\boldsymbol{X}).$$

Potom

$$Y^\star \sim N(0, \boldsymbol{\Sigma}),$$

- $\boldsymbol{\Sigma} = \sigma_A^2 \boldsymbol{W} + \sigma^2 I_\nu = \sum_{i=1}^r (\sigma_A^2 \lambda_i + \sigma^2) \boldsymbol{E}_i$,
- $\boldsymbol{W} = \boldsymbol{BVB}' = \sum_{i=1}^r \lambda_i \boldsymbol{E}_i$, je spektrálny rozklad, pričom λ_i sú vlatné čísla s násobnosťou ν_i, $\boldsymbol{E}_i$ sú navzájom ortogonálne matice, $i = 1, \ldots, r$.

Označme

$$U_i = Y^{\star\prime} \boldsymbol{E}_i Y^\star,$$

Potom

- $U = (U_1, \ldots, U_r)$ je (minimálna) postačujúca štatistika,
- $U_i/(\sigma_A^2 \lambda_i + \sigma^2) = W_i \sim \chi^2_{\nu_i}$ sú navzájom nezávislé,
- $Y^{\star\prime} \boldsymbol{\Sigma}^{-1} Y^\star = \sum_{i=1}^r W_i$,
- $\det(\boldsymbol{\Sigma}) = \Pi_{i=1}^r (\sigma_A^2 \lambda_i + \sigma^2)^{\nu_i}$.

Odtiaľ dostávame logaritmus funkcie vierohodnosti

$$\ell((\sigma_A, \sigma) \,|\, u) = -\frac{1}{2} \left[\nu \log(2\pi) + \sum_{i=1}^r \nu_i \log(\sigma_A^2 \lambda_i + \sigma^2) + \sum_{i=1}^r u_i/(\sigma_A^2 \lambda_i + \sigma^2) \right].$$

Teda

$$\ell((\sigma_A, \sigma) \,|\, u) = \ell(\theta \,|\, u) = -\frac{1}{2}\left[\nu \log(2\pi) + \sum_{i=1}^{r} \nu_i \log(\theta_i) + \sum_{i=1}^{r} u_i/\theta_i\right].$$

kde $\theta = (\theta_1, \ldots, \theta_r)$ a $\theta_i = (\sigma_A^2 \lambda_i + \sigma^2)$.

Minimálna postačujúca štatistika, vo všeobecnosti pre $r > 2$, nie je úplná pre (σ_A^2, σ^2). Problém konštrukcie konfidenčných intervalov resp. oblastí pre σ_A^2, resp. (σ_A^2, σ^2) je stále predmetom výskumu, pozri napr. Arendacká (2008).

Keďže $U_i/\theta_i = W_i \sim \chi^2_{\nu_i}$, GPQ pre θ_i je

$$\tilde{\theta}_i = U_i/W_i^\star, \quad \text{kde} \quad W^\star \sim \chi^2_{\nu_i}.$$

Pre $U_i = u_{i,obs}$ je podmienené rozdelenie $\tilde{\theta}_i \sim u_{i,obs}/W_i^\star$.

Označme

$$\lambda(\theta \,|\, u) = -2\left[\ell(\theta \,|\, u) - \ell(\hat{\theta} \,|\, u)\right] - 2\left[\ell((\sigma_A, \sigma) \,|\, u) - \ell((\hat{\sigma_A}, \hat{\sigma}) \,|\, u)\right], \text{ kde } \hat{\theta} = \hat{\theta}_{MLE}.$$

Potom $\lambda(\tilde{\theta} \,|\, u_{obs})$ je **zovšeobecnená log-likelihood funkcia**.
Vierohodnostnú oblasť pre variančné komponenty možno preto konštruovať ako približnú $(1-\alpha)$-konfidenčnú oblasť pre (σ_A, σ):

1. Na základe predpokladu platnosti asymptotického rozdelenia, teda ako množinu

$$\mathcal{R}_{1-\alpha}(u_{obs}) = \{(\sigma_A, \sigma) : \lambda((\sigma_A, \sigma) \,|\, u_{obs}) \leq \chi^2_{2,1-\alpha}\},$$

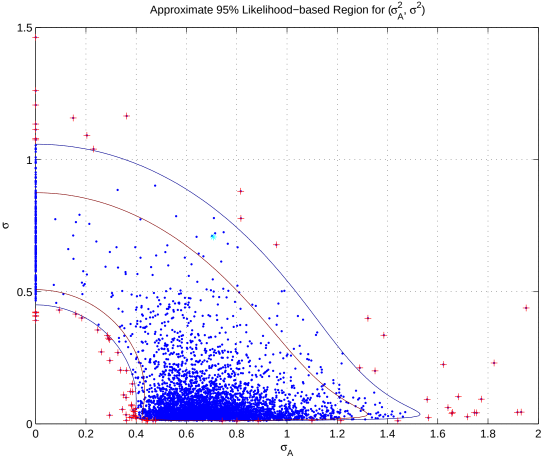
kde $\chi^2_{2,1-\alpha}$ je $(1-\alpha)$-kvantil chi-kvadát rozdelenia χ^2_2.

2. Na základe generovania realizácii náhodných premenných $\tilde{\theta}_i$ z GPQ pre $\theta_i = (\sigma_A^2 \lambda_i + \sigma^2)$, teda ako množinu

$$\mathcal{R}_{1-\alpha}(u_{obs}) = \{(\sigma_A, \sigma) : \lambda((\sigma_A, \sigma) \,|\, u_{obs}) \leq \hat{\lambda}_{1-\alpha}(u_{obs})\},$$

kde $\hat{\lambda}_{1-\alpha}(u_{obs})$ je $(1-\alpha)$-kvantil odhadnutý z empirickej distribučnej funkcie $\lambda(\tilde{\theta} \,|\, u_{obs})$.

Náhodný výber rozsahu N $= 5000$ *z distribúcie zovšeobecnených pivotov pre parametre* σ_A *a* σ.
Porovnanie empirickej a asymptotickej približnej 95%*-konfidenčnej oblasti pre* (μ, σ).



Odkazy I

Arendacká, B.:
Interval Estimators for a Variance Component in Mixed Linear Models with Two Variance Components.
Dizertačná práca na získanie vedecko-akademickej hodnosti philosophiae doctor (PhD) vo vednom odbore 11-06-9 Pravdepodobnosť a matematická štatistika. Ústav merania SAV, Bratislava, 2008.

Chvosteková, M. - Witkovský, V.:
Exact likelihood ratio test for the parameters of the linear regression model with normal errors.
Measurement Science Review, 9 (1), 2009, 1-8.

Fan, J. - Hung, H.-N. - Wong, W.-H.:
Geometric understanding of likelihood ratio statistics.
Journal of the American Statistical Association, 95(451), 2000, 836-841.

Fisher, R.A.:
The fiducial argument in statistical inference.
Annals of Eugenics, 6, 1935, 391-398.

Odkazy II

Hannig, J. - Iyer, H. - Patterson, P.:
Fiducial generalized confidence intervals.
Journal of the American Statistical Association, 101, 2006, 254,269.

Hannig, J. - Iyer, H. - Wang, C.M.:
Fiducial approach to uncertainty assessment accounting for error due to instrument resolution.
Metrologia, 44, 2007, 476–483.

Hannig, J.:
On generalized fiducial inference.
Statistica Sinica, 19, 2009, 491-544.

Tsui, K.W. - Weerahandi, S.:
Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters.
Journal of the American Statistical Association, 84, 1989, 602,607.

Odkazy III

Uusipaikka E.:
What did Fisher mean by an estimate?
Electronic Journal of Statistics, `arXiv:0807.3397v1 [math.ST]`, 2008. Submitted to the Annals of Applied Probability.

Weerahandi, S.:
Generalized confidence intervals.
Journal of the American Statistical Association, 88, 1993, 899,905.

Wimmer Jr., G.:
Algoritmus výpočtu približných konfidenčných intervalov parametra polohy z digitalizovaných meraní.
In: J. Antoch, G. Dohnal (Eds.), *ROBUST 2008: Sborník prací 15. letní školy JČMF*. Praha, 2009, 513-532.

Witkovský, V. - Wimmer, G.:
Konfidenčné intervaly založené na digitalizovaných meraniach.
In: J. Antoch, G. Dohnal (Eds.), *ROBUST 2008: Sborník prací 15. letní školy JČMF*. Praha, 2009, 513-532.

Odkazy IV

Witkovský, V. - Wimmer, G.:
Interval estimation of the mean of a normal distribution based on quantized observations.
Mathematica Slovaca, 59(5), 2009, 627-645.