Bodové odhady – některá užití pořádkových statistik

Mějme (iid) náhodný výběr $\{x_1, x_2, ..., x_n\}$ rozsahu n náhodné proměnné ξ s distribuční funkcí $F(x)$ a hustotou $f(x)$, pak vzestupně seřazená pozorování $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$, kde $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ budeme nazývat pořádkovými statistikami nad výběrem $\{x_1, x_2, ..., x_n\}$ a $x_{(i)}$ i-tou pořádkovou statistikou. Budeme dále předpokládat, že jednotlivá pozorování jsou po dvou různá, proto $x_{(1)} < x_{(2)} < ... < x_{(n)}$. Pak je

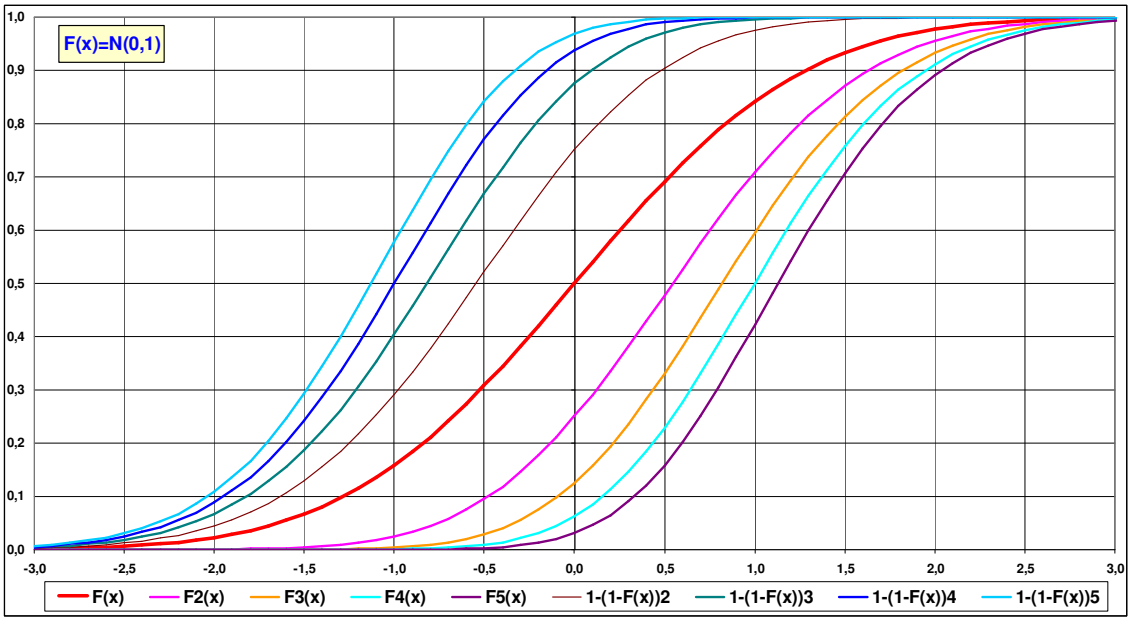
$$x_{(1)} = min\{x_1, x_2, ..., x_n\} \text{ a } x_{(n)} = max\{x_1, x_2, ..., x_n\}$$

Distribuční funkce i-té pořádkové statistiky: $F_{(i)}(x) = P\{x_{(i)} < x\}$ to je pravděpodobnost toho, že se v náhodném výběru nalezne alespoň i pozorování menších než x. Pravděpodobnost toho, že se v náhodném výběru nalezne právě i pozorování menších než x a právě n-i pozorování větších nebo rovno x je $P_i(x) = \binom{n}{i} (F(x))^i (1 - F(x))^{n-i}$. $F_{(i)}(x) = P\{x_{(i)} < x\}$ je pak součet předchozích pravděpodobností od i až do n (jedná se o disjunktní jevy). $F_{(i)}(x) = \sum_{j=i}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j}$. Speciálně je

$$F_{(n)}(x) = \binom{n}{n} (F(x))^n (1 - F(x))^{n-n} = (F(x))^n \text{ a}$$

$$F_{(1)}(x) = \sum_{j=1}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j} = 1 - \binom{n}{0} (F(x))^0 (1 - F(x))^{n-0} = 1 - (1 - F(x))^n$$

Což je, po řadě, distribuční funkce maxima a minima náhodného výběru.



Odtud můžeme dostat i výraz pro hustotu *i*-té pořádkové statistiky:

$$f_{(i)}(x) = \frac{d}{dx} F_{(i)}(x) = \frac{d}{dx} \sum_{j=i}^{n} \binom{n}{j} (F(x))^j (1-F(x))^{n-j} = \sum_{j=i}^{n} \binom{n}{j} j (F(x))^{j-1} (1-F(x))^{n-j} f(x) -$$

$$- \sum_{j=i}^{n} \binom{n}{j} (n-j) (F(x))^j (1-F(x))^{n-j-1} f(x) = A$$

Protože:

$$\binom{n}{j} j = j \frac{n!}{j!(n-j)!} = n \frac{(n-1)!}{(j-1)!(n-j)!} = n \binom{n-1}{j-1} \text{ a}$$

$$\binom{n}{j} (n-j) = (n-j) \frac{n!}{j!(n-j)!} = n \frac{(n-1)!}{j!(n-j-1)!} = n \binom{n-1}{j}$$

je

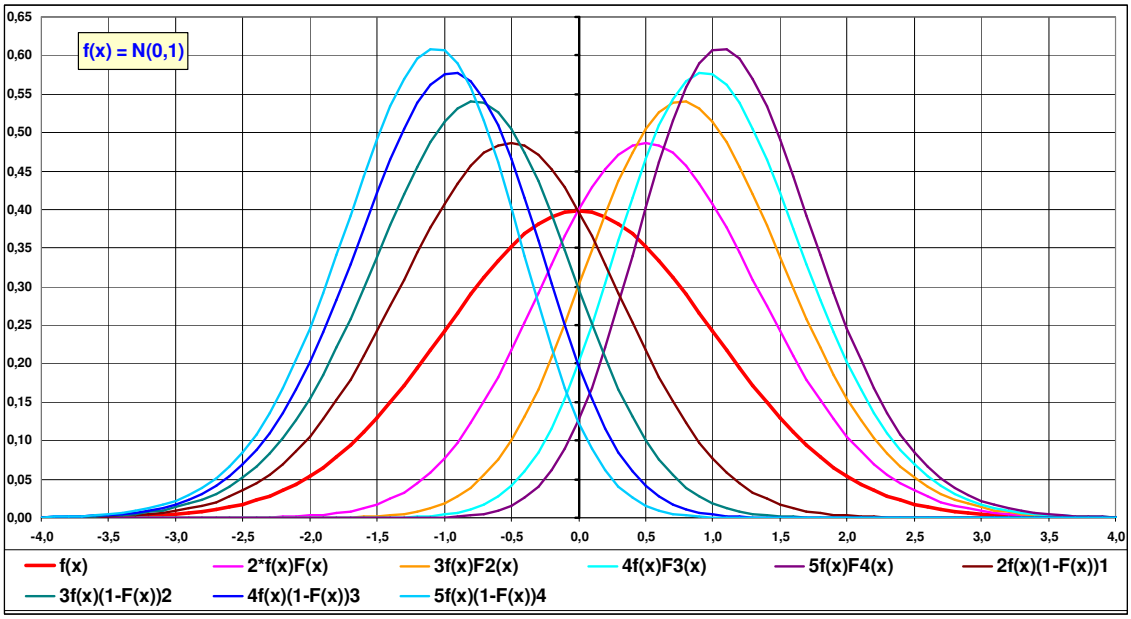
$$A = \sum_{j=i}^{n} n \binom{n-1}{j-1} (F(x))^{j-1} (1-F(x))^{n-j} f(x) - \sum_{j=i}^{n} n \binom{n-1}{j} (F(x))^j (1-F(x))^{n-j-1} f(x) =$$

$$= n \binom{n-1}{i-1} (F(x))^{i-1} (1-F(x))^{n-j} f(x) \text{ a tedy:}$$

$$f_{(i)}(x) = n \binom{n-1}{i-1} (F(x))^{i-1} (1-F(x))^{n-i} f(x)$$

Opět speciálně pro minimum a maximum:

$$f_{(1)}(x) = n f(x) (1-F(x))^{n-1} \quad \text{a} \quad f_{(n)}(x) = n f(x) F^{n-1}(x)$$



Symetrická rozdělení

Náhodná veličina má symetrické rozdělení podle střední hodnoty (obecně jakéhokoliv čísla) μ, jestliže pro její hustotu pravděpodobnosti platí $f(x) = f(2\mu - x)$. Pro distribuční funkci takového rozdělení platí:

$$F(x) = \int_{-\infty}^{x} f(s)ds = 1 - \int_{x}^{+\infty} f(s)ds = 1 - \int_{x}^{+\infty} f(2\mu - s)ds = 1 - \int_{-\infty}^{2\mu - x} f(s)ds = 1 - F(2\mu - x)$$

Medián a kvantil podruhé

p-kvantil je číslo x_p pro které platí $P\{x < x_p\} = F(x_p) = p$. Medián je 0,5-kvantil. Pro symetrické rozdělení pak $p = F(x_p) = 1 - F(2\mu - x_p)$ a pro medián je $0{,}5 = F(x_{0,5}) = 1 - F(2\mu - x_{0,5}) \Rightarrow 0{,}5 = F(x_{0,5})$ a $0{,}5 = F(2\mu - x_{0,5})$

proto $F(2\mu - x_{0,5}) = F(x_{0,5})$

Pokud uvažujeme navíc spojitou a rostoucí distribuční funkci, pak platí

$$2\mu - x_{0,5} = x_{0,5} \Leftrightarrow \mu = x_{0,5}.$$

Střední hodnota symetrického rozdělení:

$$E\{x\} = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{+\infty} xf(2\mu - x)dx = \int_{-\infty}^{+\infty} (2\mu - z)f(z)dz = 2\mu - E\{x\} \Rightarrow E\{x\} = \mu$$

Centrální momenty symetrického rozdělení:

$m_r\{x\} = \int_{-\infty}^{+\infty} (x-\mu)^r f(x)dx = \int_{-\infty}^{+\infty} (x-\mu)^r f(2\mu - x)dx = \int_{-\infty}^{+\infty} (\mu - z)^r f(z)dz$, proto pro *r lichá* je $\int_{-\infty}^{+\infty} (\mu - z)^r f(z)dz = -\int_{-\infty}^{+\infty} (z-\mu)^r f(z)dz$ a platí $m_r\{x\} = 0$ (pro r lichá). Symetrická rozdělení mají všechny liché centrální momenty nulové.

Výběrový medián

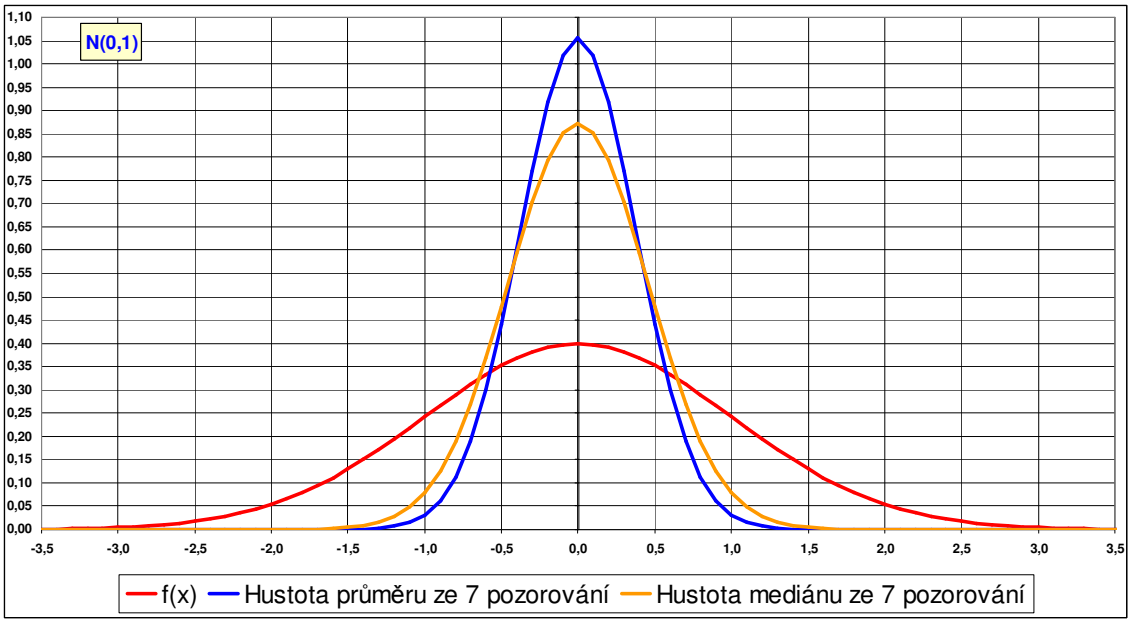
Pro n liché je $me = x_{\left(\frac{n+1}{2}\right)}$ a pro n sudé $me = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right)$. Ten pak má distribuční funkci a hustotu při výběru z rozdělení $F(x), f(x)$:

$$F_{(\frac{n+1}{2})}(x) = \sum_{j=\frac{n+1}{2}}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j} \text{ a}$$

$$f_{(\frac{n+1}{2})}(x) = n\binom{n-1}{\frac{n-1}{2}} (F(x))^{\frac{n-1}{2}} (1 - F(x))^{\frac{n-1}{2}} f(x) = n\binom{n-1}{\frac{n-1}{2}} (F(x)(1 - F(x)))^{\frac{n-1}{2}} f(x),$$

pro n-liché.

Srovnání hustoty průměru a výběrového mediánu



Z tohoto obrázku je vidět, že průměr má menší variabilitu. To je ovšem za předpokladu, že se jedná o výběr stejně rozdělených pozorování. Pokud se do výběru „namixují“ i pozorování s jiným rozdělením pravděpodobnosti, situace se mění, zvl. pro odlehlé (extrémní) hodnoty. V případě působení odlehlých pozorování z jiného rozdělení pravděpodobnosti bývá medián spolehlivější.

Sdružené rozdělení výběrového maxima a minima

$$F(y_1, y_n) = P\{x_{(1)} < y_1; x_{(n)} < y_n\} = [F(y_n)]^n - [F(y_n) - F(y_1)]^n$$

Poznámka: První člen je distribuce maxima a druhý člen je pravděpodobnost toho, že všechna pozorování budou mezi x_1 a x_n. (Rao C.R.: Lineární metody statistické indukce a jejich aplikace, ACADEMIA, Praha 1978, str.249).

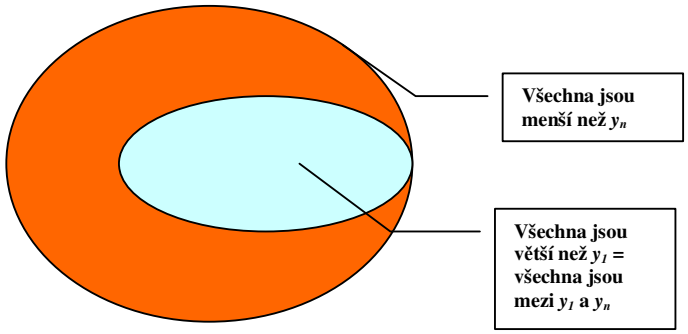
Odvození:

1. $y_1 < y_n \Rightarrow F(y_1, y_n) = P\{\min < y_1; \max < y_n\} = P\left\{\begin{matrix}[\text{nektera jsou mensi nez } y_1] \wedge \\ [\text{vsechna jsou mensi nez } y_n]\end{matrix}\right\} =$

$$= P\left\{\begin{matrix}\overline{[\text{vsechna jsou vetsi nez } y_1]} \wedge \\ [\text{vsechna jsou mensi nez } y_n]\end{matrix}\right\} = P\left\{\begin{matrix}[\text{vsechna jsou mezi } y_1 \text{ a } y_n] \wedge \\ [\text{vsechna jsou mensi nez } y_n]\end{matrix}\right\}$$

2. $y_1 \geq y_n \Rightarrow F(y_1, y_n) = P\{\min < y_1; \max < y_n\} = P\{[\text{vsechna jsou mensi nez } y_n]\}$
3. Je zřejmé, že jev $[\text{vsechna jsou mensi nez } y_n]$ je obsažen v jevu $\left\{\begin{matrix}[\text{nektera jsou mensi nez } y_1] \wedge \\ [\text{vsechna jsou mensi nez } y_n]\end{matrix}\right\}$ a proto platí :

$$F(y_1, y_n) = P\{\min < y_1; \max < y_n\} = [F(y_n)]^n - [F(y_n) - F(y_1)]^n$$



Tomu odpovídá hustota: $f(y_1, y_n) = n(n-1)[F(y_n) - F(y_1)]^{n-2} f(y_1) f(y_n)$. Odtud lze odvodit rozdělení náhodné veličiny $R = y_{(n)} - y_{(1)}$ - výběrového rozpětí:

Hustota výběrového rozpětí je pak:

$$f(R) = n(n-1)\int_{-\infty}^{+\infty} [F(x+R) - F(x)]^{n-2} f(x) f(x+R) dx \Leftrightarrow R \geq 0; = 0 \Leftrightarrow R < 0$$

Blíže, viz: Doc. Ing. Dagmar Blatná, CSc.: Neparametrické metody. Testy založené na pořádkových a pořadových statistikách. Skripta VŠE, Praha 1996.

Bodové odhady využívající pořádkových statistik

Výběry (uspořádané) z rovnoměrného rozdělení na intervalu (a,b).

$f(x) = \frac{1}{b-a} \Leftrightarrow a < x < b; f(x) = 0$ v ostatních případech.

$F(x) = 0 \Leftrightarrow x \leq a$,

$F(x) = \frac{x-a}{b-a} \Leftrightarrow a \leq x \leq b$,

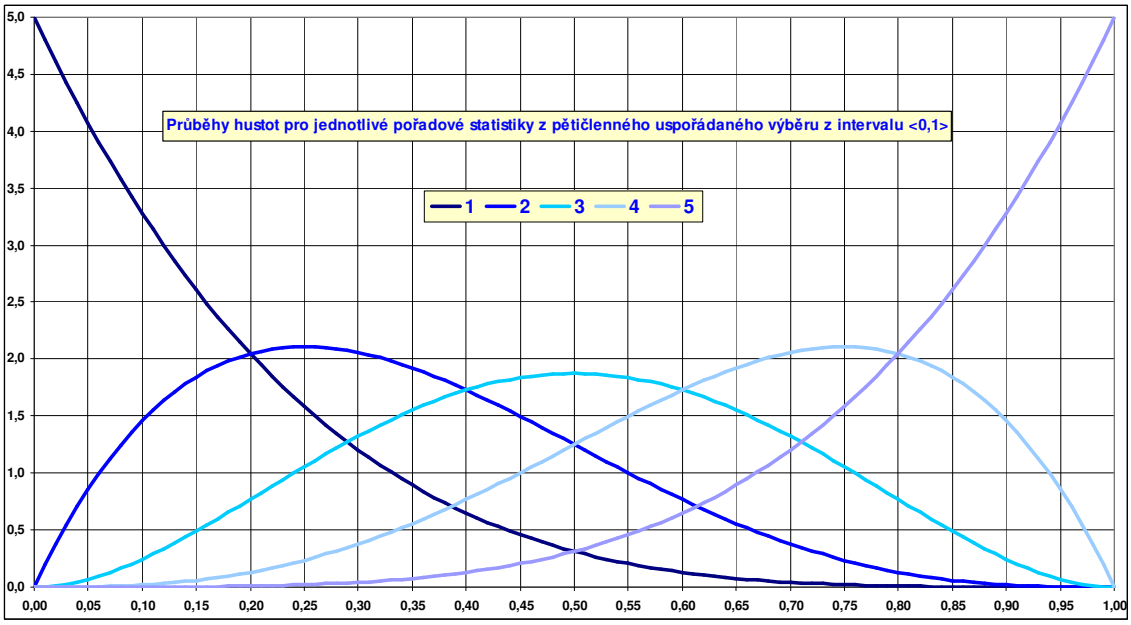
$F(x) = 1 \Leftrightarrow x \geq b$.

Protože:

$f_{(i)}(x) = n\binom{n-1}{i-1}(F(x))^{i-1}(1-F(x))^{n-i} f(x)$ je

$f_{(i)}(x) = n\binom{n-1}{i-1}\left(\frac{x-a}{b-a}\right)^{i-1}\left(\frac{b-x}{b-a}\right)^{n-i}\frac{1}{b-a} \Leftrightarrow a < x < b$ jinak $f_{(i)}(x) = 0$.

$f_{(i)}(x) = n(b-a)^{-n}\binom{n-1}{i-1}(x-a)^{i-1}(b-x)^{n-i} \Leftrightarrow a < x < b$ jinak $f_{(i)}(x) = 0$.



Potom:

$$f_{(1)}(x)=\frac{n}{(b-a)^n}(b-x)^{n-1},$$

$$E\{x_{(1)}\}=b-\frac{n}{n+1}(b-a),$$

$$f_{(n)}(x)=\frac{n}{(b-a)^n}(x-a)^{n-1}\ \text{a}$$

$$E\{x_{(n)}\}=\frac{n}{n+1}(b-a)+a.$$

Potom

$$E\{x_{(n)}-x_{(1)}\}=\frac{n-1}{n+1}(b-a)\ \text{a odtud}\ E\left\{(x_{(n)}-x_{(1)})\frac{n+1}{n-1}\right\}=(b-a).\ \text{Tedy:}$$

$(x_{(n)}-x_{(1)})\frac{n+1}{n-1}$ je nestranným odhadem variačního rozpětí z rovnoměrného rozdělení na intervalu (a,b).

Konstrukce odhadů parametrů a a b.

$$E\{x_{(1)}\} = b - \frac{n}{n+1}(b-a) = b - \frac{n}{n+1}E\left\{(x_{(n)} - x_{(1)})\frac{n+1}{n-1}\right\} = b - \frac{n}{n-1}E\{(x_{(n)} - x_{(1)})\} \Rightarrow$$

$$b = E\left\{x_{(1)} + \frac{n}{n-1}(x_{(n)} - x_{(1)})\right\} = E\left\{\frac{nx_{(n)} - x_{(1)}}{n-1}\right\} = E\left\{x_{(n)} + \frac{x_{(n)} - x_{(1)}}{n-1}\right\}, \text{ tedy:}$$

$x_{(n)} + \frac{x_{(n)} - x_{(1)}}{n-1}$ je nestranným odhadem parametru b.

$$E\{x_{(n)}\} = \frac{n}{n+1}(b-a) + a = \frac{n}{n+1}E\left\{(x_{(n)} - x_{(1)})\frac{n+1}{n-1}\right\} + a \Rightarrow$$

$$a = E\left\{x_{(n)} - \frac{n}{n-1}(x_{(n)} - x_{(1)})\right\} = E\left\{\frac{nx_{(1)} - x_{(n)}}{n-1}\right\} = E\left\{x_{(1)} - \frac{x_{(n)} - x_{(1)}}{n-1})\right\}, \text{ tedy:}$$

$x_{(1)} - \frac{x_{(n)} - x_{(1)}}{n-1})$ je nestranným odhadem parametru a.

Protože:

$\hat{b} = x_{(n)} + \frac{x_{(n)} - x_{(1)}}{n-1}; \hat{a} = x_{(1)} - \frac{x_{(n)} - x_{(1)}}{n-1}$ je $\hat{b} - \hat{a} = x_{(n)} + \frac{x_{(n)} - x_{(1)}}{n-1} - x_{(1)} + \frac{x_{(n)} - x_{(1)}}{n-1} =$

$= x_{(n)} - x_{(1)} + 2\frac{x_{(n)} - x_{(1)}}{n-1} = (x_{(n)} - x_{(1)})\left(1 + \frac{2}{n-1}\right) = (x_{(n)} - x_{(1)})\left(\frac{n+1}{n-1}\right)$. Tj. odhady $\hat{b}; \hat{a}; \hat{b} - \hat{a}; \hat{b} - \hat{a}$ jsou konzistentní (zde slovo konzistence znamená, že s nimi lze provádět uvedené operace, tj. platí: odhad rozdílu se v tomto případě rovná rozdílu odhadů).

Námět: Zkuste pro tento případ spočítat hustotu variačního rozpětí $x_{(n)} - x_{(1)}$.

Výběry (uspořádané) z posunutého exponenciálního rozdělení

$$E(A,\delta) \equiv f(x) = \frac{1}{\delta}e^{-\frac{1}{\delta}(x-A)} \Leftrightarrow x > A; f(x) = 0 \Leftrightarrow x \le A.$$

Pro posunuté exponenciální rozdělení

$E(A,\delta) \equiv f(x) = \frac{1}{\delta}e^{-\frac{1}{\delta}(x-A)} \Leftrightarrow x > A; f(x) = 0 \Leftrightarrow x \le A$ je nestranným odhadem parametrické funkce $A + \delta$ statistika $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, zatímco statistika:

$x_{(p)} - \frac{1}{q-p}\left(\sum_{i=1}^{p}\frac{1}{n-i+1}\right)\left[\sum_{i=p+1}^{q}(x_{(i)} - x_{(p)}) + (n-q)(x_{(q)} - x_{(p)})\right]$ *pro* $1 \le p < q \le n$ je nestranným odhadem parametru A.

Důkaz: $f_{(i)}(x) = \frac{n!}{(i-1)!(n-i)!}\frac{1}{\delta}\left[1-e^{-n\frac{a-A}{\delta}}\right]^{i-1} e^{-\frac{(n-i+1)(x-A)}{\delta}}$ pro uvedené exponenciální rozdělení.

Zdroj: Hátle, J., Likeš, J.: Základy počtu pravděpodobnosti a matematické statistiky, SNTL/ALFA, Praha 1974, str. 165 – 188.

Příklad: Nestranným odhadem pro parametr δ z posunutého exponenciálního rozdělení

$$\frac{1}{q-p}\left[\sum_{i=p+1}^{q}(x_{(i)}-x_{(p)})+(n-q)(x_{(q)}-x_{(p)})\right] \quad pro \quad 1 \le p < q \le n .$$

Námět: Specifikujte uvedené statistiky pro výběr z posunutého exponenciálního rozdělení pro $1 = p; q = n$.

Příklad: Spočtěte $E\{x_{(1)}\}, E\{x_{(n)}\}$ a $E\{x_{(n)} - x_{(1)}\}$ u výběru s posunutého exponenciálního rozdělení. Na základě těchto výpočtů se pokuste odvodit statistiku pro nestranný odhad δ. Jsou veličiny $x_{(n)}, x_{(1)}$ nezávislé?

Doporučená a zdrojová literatura:

Jiří Reif	Metody matematické statistiky, ZČU v Plzni 2004
Jaroslav Hátle, Jiří Likeš	Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974.
Alfréd Rényi	Teorie pravděpodobnosti, ACADEMIA, Praha 1972
C. Radhakrishna Rao	Lineární metody statistické indukce a jejich aplikace, ACADEMIA, Praha 1978
Dagmar Blatná	Neparametrické metody. Testy založené na pořádkových a pořadových statistikách. Skripta VŠE, Praha 1996.