Statistické toleranční a predikční oblasti, spojitá rozdělení, Wilksovy toleranční meze.

Definice – toleranční interval (opět)

Mějme náhodný výběr $\{x_1, x_2,..., x_n\}$ (iid) z náhodné proměnné ξ s distribuční funkcí $F_\xi(x)$ a dvě statistiky $L(x_1, x_2,..., x_n)$a $U(x_1, x_2,..., x_n)$pod **tolerančním (statistickým) intervalem** (β, γ) budeme rozumět obě statistiky, pokud platí:

- $L(x_1, x_2,..., x_n) < U(x_1, x_2,..., x_n)$
- $P_{x_1,x_2,...x_n}\left(P_\xi\{L(x_1, x_2,..., x_n) \le \xi \le U(x_1, x_2,..., x_n)\} \ge \beta\right) \ge \gamma$

Volněji: pravděpodobnost toho, že interval $\langle L,U \rangle$ „pokryje“ více než $100\,\beta\,\%$ pravděpodobnosti výskytu náhodné proměnné ξ je alespoň $100\,\gamma\,\%$.

Definice - predikční interval (oblast)

Za výše uvedených podmínek budeme pod β predikčním intervalem rozumět obě statistiky, pokud platí

- $E_{x_1,x_2,...x_n}\left[P_\xi\{L(x_1, x_2,..., x_n) \le \xi \le U(x_1, x_2,..., x_n)\}\right] \ge \beta$

Volněji:

U tolerančního intervalu je požadováno to, aby pravděpodobnost jevu, že náhodná proměnná bude ležet ve „spočtených-odhadnutých“ mezích s pravděpodobností β bude alespoň γ.

U predikčního intervalu je alternativně požadováno to, aby střední hodnota pravděpodobnosti jevu, že náhodná proměnná bude ležet ve „spočtených-odhadnutých“ mezích byla alespoň β.

Filosofie Wilksových mezí a některé potřebné vztahy

V některých případech není přijatelný vhodný předpoklad o rozdělení, ze kterého pochází pozorování analyzované náhodné proměnné. Pak je možné využít pořadových statistik:

Mějme (iid) náhodný výběr $\{x_1, x_2,..., x_n\}$ rozsahu n náhodné proměnné ξ s distribuční funkcí $F(x)$ a hustotou $f(x)$, $F(x)$ spojitá, rostoucí, 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 $x_{(1)} < x_{(2)} < ... < x_{(n)}$.

Námět: Nestačí pro podmínku $x_{(1)} < x_{(2)} < ... < x_{(n)}$ předpoklad $F(x)$ spojitá?

Pro libovolné $i < j \le n$ budeme nazývat $F_\xi(x_{(j)}) - F_\xi(x_{(i)})$ pravděpodobnostním pokrytím intervalu $\langle x_{(i)}, x_{(j)} \rangle$. Uspořádaný náhodný výběr $(x_{(1)}, x_{(2)},..., x_{(n)})$ dělí R_1 na $n+1$ intervalů

$I_i = \langle x_{(i-1)}, x_{(i)} \rangle$, kde $x_{(0)} = -\infty$ a $x_{(n+1)} = +\infty$. Budeme označovat $c_i = P(\xi \in I_i)$ jako pravděpodobnostní pokrytí intervalu $I_i = \langle x_{(i-1)}, x_{(i)} \rangle$. Tato pokrytí jsou náhodné proměnné spojené vztahem $\sum_{i=1}^{n+1} c_i = 1$.

Součet *W* **libovolných** $r \le n$ po dvou různých náhodných proměnných c_i má pak hustotu

$$f_W(x) = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)} x^{r-1}(1-x)^{n-r} = \frac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r} = \frac{r}{r}\frac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r} =$$

$$= r\binom{n}{r} x^{r-1}(1-x)^{n-r}.$$ Tedy Beta rozdělení.

Dále: Součet *W* libovolných $r \le n$ po dvou různých náhodných proměnných c_i a *V* libovolných $s \le n; r+s \le n$ po dvou různých náhodných proměnných c_i. Množiny sčítanců vstupujících do *W* a do *V* jsou disjunktní. Potom dvourozměrná hustota

$$f_{W,V}(u,v) = \frac{\Gamma(n+1)}{\Gamma(r)\Gamma(s)\Gamma(n-r-s+1)} u^{r-1}v^{s-1}(1-u-v)^{n-r-s} =$$

$$= \frac{n!}{(r-1)!(s-1)!(n-r-s)!} u^{r-1}v^{s-1}(1-u-v)^{n-r-s} =$$

$$= \frac{(r+s)!}{(r-1)!(s-1)!}\frac{n!}{(r+s)!(n-r-s)!} u^{r-1}v^{s-1}(1-u-v)^{n-r-s} =$$

$$= \frac{(r+s)!}{(r-1)!(s-1)!}\binom{n}{r+s} u^{r-1}v^{s-1}(1-u-v)^{n-r-s} = rs\binom{r+s}{s}\binom{n}{r+s} u^{r-1}v^{s-1}(1-u-v)^{n-r-s}.$$

Důkazy a odvození výše uvedených vztahů viz: Wilks S. S.: Order statistics. Bull. Amer. Math. Soc. Volume 54, Number 1, Part 1 (1948), 6 50. Motivace a některé dílčí vztahy v přílohách.

Námět: spočtěte k uvedeným hustotám jejich distribuční funkce.

Příklad využití:

Toleranční (statistický) limit pro jednorozměrnou populaci.

Hledáme toleranční meze ve tvaru $L(x_1, x_2, ..., x_n) = x_{(k)}$ a $U(x_1, x_2, ..., x_n) = x_{(n-k+1)}$. Potom:

$$\gamma = P_{x_1,x_2,...x_n}\left[P_\xi\{L(x_1, x_2, ..., x_n) \le \xi \le U(x_1, x_2, ..., x_n)\} \ge \beta\right] = P_{x_1,x_2,...x_n}\left[F_\xi(U) - F_\xi(L) \ge \beta\right] =$$

$$= P_{x_1,x_2,...x_n}\left[F_\xi(x_{(n-k+1)}) - F_\xi(x_{(k)}) \ge \beta\right] = P\left[\sum_{j=1}^{n-k+1} c_j - \sum_{j=1}^{k} c_j \ge \beta\right] = P\left[\sum_{j=k+1}^{n-k+1} c_j \ge \beta\right] = P[S \ge \beta],$$ kde

$S = \sum_{j=k+1}^{n-k+1} c_j$, tj. součet $n - 2k + 1$ elementárních pravděpodobnostních pokrytí.

Shrnuto: $\gamma = P[S \geq \beta] = 1 - P(S < \beta) = 1 - F_S(\beta)$. Ale náhodná proměnná S má hustotu $f_S(x) = \frac{\Gamma(n+1)}{\Gamma(n-2k+1)\Gamma(n-(n-2k+1)+1)} x^{n-2k+1-1}(1-x)^{n-(n-2k+1)}$ (viz výše) a ta po úpravě dá $f_S(x) = \frac{\Gamma(n+1)}{\Gamma(n-2k+1)\Gamma(2k)} x^{n-2k}(1-x)^{2k-1}$.

Odtud: $1 - F_S(\beta) = 1 - \int_0^\beta \frac{\Gamma(n+1)}{\Gamma(n-2k+1)\Gamma(2k)} x^{n-2k}(1-x)^{2k-1} dx = 1 - I(\beta; n-2k+1, 2k)$.

Kde $I(\beta; n-2k+1, 2k)$ je distribuční funkce beta rozdělení (viz dále) s parametry $a = n-2k+1, b = 2k$.

Opět shrnuto $\gamma = 1 - I(\beta; n-2k+1, 2k) \Leftrightarrow 1-\gamma = I(\beta; n-2k+1, 2k)$. Parametrickým řešením $\beta = I^{-1}(1-\gamma; n-2k+1, 2k)$ této rovnice nalezneme vztah mezi n, k pro dané β. Poznámka: pro dostatečně velké γ (a vhodně zvolené) bude k malé.

Jednostranné toleranční intervaly

Pravostranný $P_{x_1,x_2,\ldots,x_n}\left[P_\xi\{\xi \leq U(x_1,x_2,\ldots,x_n)\} \geq \beta\right] = P_{x_1,x_2,\ldots,x_n}\left[P_\xi\{\xi < U(x_1,x_2,\ldots,x_n)\} \geq \beta\right] \geq \gamma$

Levostranný $P_{x_1,x_2,\ldots,x_n}\left[P_\xi\{L(x_1,x_2,\ldots,x_n) \leq \xi\} \geq \beta\right] \geq \gamma$

Pravostranný Wilksův toleranční interval

Hledáme takové *j*, aby platilo: $P_{x_1,x_2,\ldots,x_n}\left[P_\xi\{\xi < x_{(j)}\} \geq \beta\right] \geq \gamma$. Ale $P_\xi\{\xi < x_{(j)}\} = F(x_{(j)})$, proto $P_{x_1,x_2,\ldots,x_n}\left[F(x_{(j)}) \geq \beta\right] \geq \gamma \Leftrightarrow P_{x_1,x_2,\ldots,x_n}\left[y_{(j)} \geq \beta\right] \geq \gamma$. $y_{(j)}$ je j-tá pořádková statistika výběru z rovnoměrného rozdělení na intervalu $(0,1)$, pak $P(y_j < x) = F_{y_j}(x) = x \Leftrightarrow 0 \leq x \leq 1; j = 1,\ldots,n$ a

$$P(y_{(j)} < x) = F_{(j)}(x) = \sum_{i=j}^{n} \binom{n}{i} \left(F_{y_j}(x)\right)^i \left(1 - F_{y_j}(x)\right)^{n-i} = \sum_{i=j}^{n} \binom{n}{i} x^i (1-x)^{n-i}.$$

Podmínku $P_{x_1,x_2,\ldots,x_n}\left[F(x_{(j)}) \geq \beta\right] \geq \gamma$ odtud přepíšeme na $1 - \sum_{i=j}^{n} \binom{n}{i} \beta^i (1-\beta)^{n-i} \geq \gamma$. To je ekvivalentní $\sum_{i=0}^{j-1} \binom{n}{i} \beta^i (1-\beta)^{n-i} \geq \gamma$. Tu řešíme vůči j. To je poměrně snadné numericky. Jinou, obdobnou, možnost řešení lze získat pomocí následující rovnosti $\sum_{i=0}^{j-1} \binom{n}{i} \beta^i (1-\beta)^{n-i} = F_{F_{(2(n-j+1),2j)}}\left(\frac{j}{n-j+1}\frac{1-\beta}{\beta}\right)$, kde $F_{F_{(2(n-j+1),2j)}}(x)$ je distribuční funkce F-rozdělení s $\nu_1 = 2(n-j+1)$ a $\nu_2 = 2j$ stupni volnosti (Jaroslav Hátle, Jiří Likeš: Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974, str. 141). Srovnatelnou možností je i využití $\sum_{i=0}^{j-1} \binom{n}{i} \beta^i (1-\beta)^{n-i} = \frac{1}{B(n-j+1, j)} \int_0^{1-\beta} x^{n-j}(1-x)^{j-1} dx$, tedy pomocí distribuční funkce Beta (Jaroslav Hátle, Jiří Likeš: Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974, str. 133).

Levostranný Wilksův toleranční interval

Hledáme takové i, aby platilo: $P_{x_1,x_2,\ldots,x_n}\left[P_\xi\left\{x_{(i)} \geq \xi\right\} \geq \beta\right] \geq \gamma$. Ale $P_\xi\left\{\xi \geq x_{(i)}\right\} = 1 - F\left(x_{(i)}\right)$, proto $P_{x_1,x_2,\ldots,x_n}\left[1 - F\left(x_{(i)}\right) \geq \beta\right] \geq \gamma \Leftrightarrow P_{x_1,x_2,\ldots,x_n}\left[1 - y_{(i)} \geq \beta\right] \geq \gamma \Leftrightarrow P_{x_1,x_2,\ldots,x_n}\left[y_{(i)} < 1-\beta\right] \geq \gamma$. $y_{(i)}$ je i-tá pořádková statistika výběru z rovnoměrného rozdělení na intervalu $(0,1)$, pak $P_{x_1,x_2,\ldots,x_n}\left[y_{(i)} < 1-\beta\right] = \sum_{j=i}^{n}\binom{n}{j}(1-\beta)^j \beta^{n-j} = 1 - \sum_{j=0}^{i-1}\binom{n}{j}(1-\beta)^j \beta^{n-j}$. Nalezení levostranného tolerančního intervalu je nalezením nejmenšího i splňujícího nerovnost $1 - \sum_{j=0}^{i-1}\binom{n}{j}(1-\beta)^j \beta^{n-j} \geq \gamma$. Opět se jako nejvhodnější jeví numerické řešení nebo s využitím vztahů uvedených u pravostranného intervalu.

Pro numerické řešení výše uvedených úloh se jeví jako výhodný následující vztah:

$\sum_{i=o}^{j}\binom{n}{i}x^i(1-x)^{n-i} = \sum_{i=o}^{j} a_i$, kde $a_i = \binom{n}{i}x^i(1-x)^{n-i}$, $n\lg(1-x) = \lg a_0$ a $a_{i+1} = \frac{x}{1-x}\frac{n-i}{i+1}a_i$.

Námět: Wilksových mezí lze použít pro testování extrémních pozorování. Formulujte takovou úlohu. Extrémní pozorování je natolik velké (nebo malé), že lze vyloučit hypotézu, že se takové pozorování řídí stejným rozdělením jako ostatní pozorování. Nalezněte řešení takové úlohy pomocí Wilksových mezí.

Námět: Lze nalézt pro libovolné (β, γ) a pevně daný rozsah výběru řešení problému tolerančních mezí Wilksovými mezemi?

Námět: Pokud bude odpověď na předchozí otázku negativní, nalezněte nejmenší rozsah náhodného výběru *n* tak aby byl problém řešitelný pomocí Wilksových mezí.

Doporučená a zdrojová literatura:

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
Machek J.	Teorie odhadu, SPN Praha 1974, skripta MFFUK
Miloš Jílek	Statistické toleranční meze. SNTL Praha 1988.
Wilks S. S.	Determination of sample size for getting tolerance limits. AMS, 12, 1941. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177731788
Wilks S. S.	Order statistics. Bull. Amer. Math. Soc. Volume 54, Number 1, Part 1 (1948), 6-50. http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.bams/1183511502

Příloha: Některé užitečné pojmy a vztahy:

Beta funkce: $B(a,b)=\int_0^1 x^{a-1}(1-x)^{b-1}dx \qquad a,b>0$

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)};\quad B(a,b)=B(b,a);\quad B(a,b)=\frac{a-1}{a+b-1}B(a-1,b);\quad B(a,b)=\frac{b-1}{a+b-1}B(a,b-1);$$

$$\sum_{i=k}^{n}\binom{n}{i}p^i(1-p)^{n-i}=\frac{n!}{(k-1)!\,(n-k)!}\int_0^p x^{k-1}(1-x)^{n-k}dx$$

Beta rozdělení: Jeho distribuční funkce

$I(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x z^{a-1}(1-z)^{b-1}dz \qquad a,b>0;\quad 0\le x\le 1$ a jeho hustota:

$i(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \quad 0\le x\le 1$

Námět: Určete momenty beta rozdělení (zvl. pro a,b přirozená).

Příloha:

Odvození a důkazy některých potřebných vztahů.

Mějme (iid) náhodný výběr $\{x_1,x_2,...,x_n\}$ rozsahu n náhodné proměnné ξ s distribuční funkcí $F(x)$ a hustotou $f(x)$, $F(x)$ spojitá, rostoucí, 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 $x_{(1)}<x_{(2)}<...<x_{(n)}$.

Sdružená hustota dvou pořádkových statistik $1\le i<j\le n$; $(x_{(i)},x_{(j)})$; $x_{(i)}<x_{(j)}$

$$f_{(i,j)}(x,y)=\begin{cases}\dfrac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\left(F(x)\right)^{i-1}\left(F(y)-F(x)\right)^{j-i-1}\left(1-F(y)\right)^{n-j}f(x)f(y) & \Leftrightarrow -\infty<x<y<+\infty\\ 0 \quad \text{jinak}\end{cases}$$

Např.: Jaroslav Hátle, Jiří Likeš: Základy počtu pravděpodobnosti a matematické statistiky. SNTL. Praha 1974. Str. 157-161 (je zde však jen odkaz na další práce).

Pro $1=i<j=n$ $\quad f_{(1,n)}(x,y)=\begin{cases}n(n-1)\left(F(y)-F(x)\right)^{n-2}f(x)f(y) & \Leftrightarrow -\infty<x<y<+\infty\\ 0 \quad \text{jinak}\end{cases}$, tj.

sdružená hustota minima a maxima.

Pro $1 \le i < i+1 \le n$

$$f_{(i,i+1)}(x,y) = \begin{cases} \dfrac{n!}{(i-1)!(n-i-1)!}\left(F(x)\right)^{i-1}\left(1-F(y)\right)^{n-i-1} f(x) f(y) \Leftrightarrow -\infty < x < y < +\infty \\ 0 \quad \text{jinak} \end{cases}$$

, tj. sdružená hustota po sobě následujících pořádkových statistik.

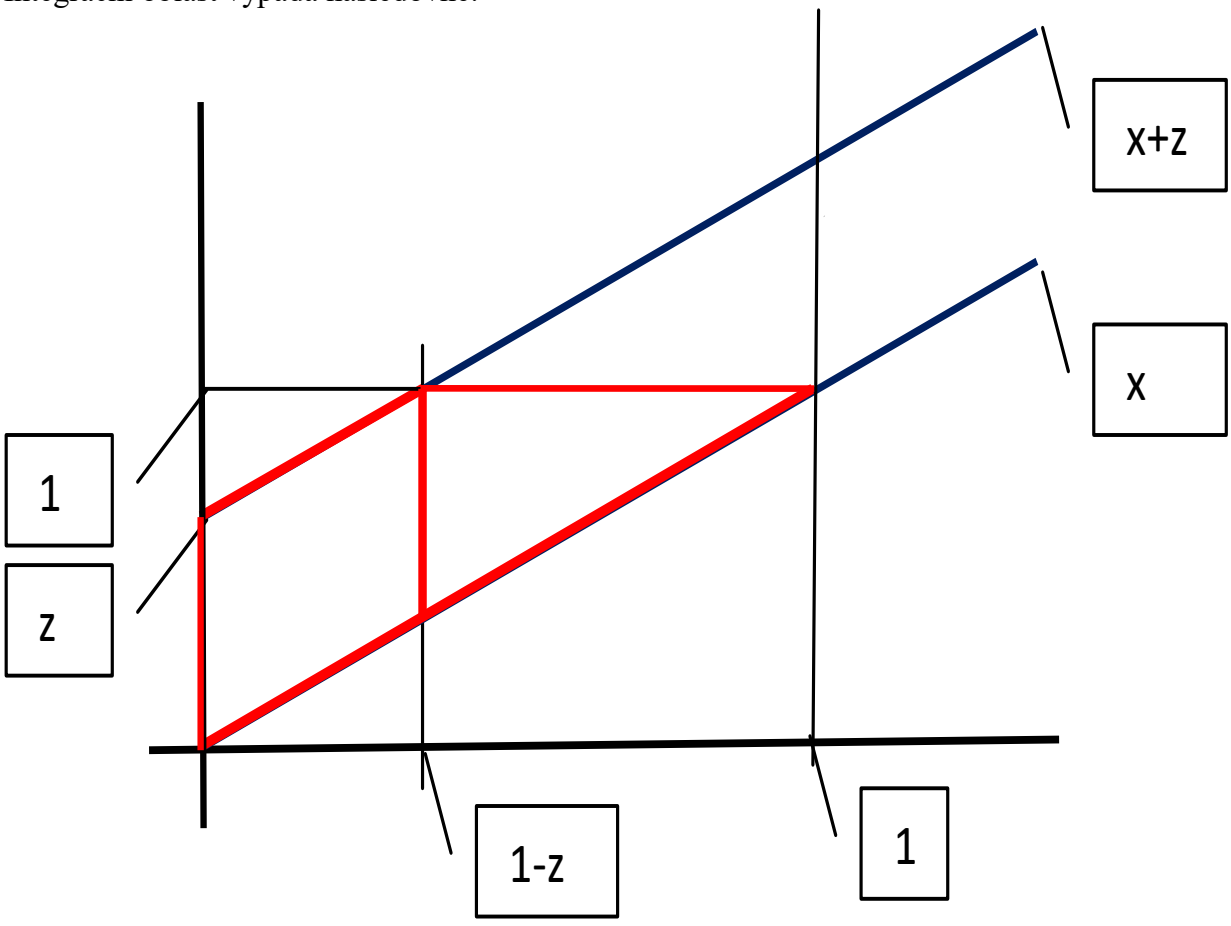
Tvrzení: Jestliže $\{x_1, x_2, ..., x_n\}$ je náhodný výběr pak $\{y_1, y_2, ..., y_n\}$, kde $y_i = F(x_i)$ je náhodný výběr transformovaných náhodných proměnných a $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$, kde $x_{(1)} < x_{(2)} < ... < x_{(n)}$ je uspořádaný náhodný výběr netransformovaných náhodných proměnných a $\{y_{(1)}, y_{(2)}, ..., y_{(n)}\}$, kde $y_{(1)} < y_{(2)} < ... < y_{(n)}$ a $y_{(i)} = F(x_{(i)})$. Tj. uvedená transformace, při definovaných podmínkách na $F(x)$ dodržuje pořadí uspořádaného náhodného výběru.

Důkaz: $F(x)$ je spojitá, rostoucí a dodržuje tedy ostré nerovnosti.

Potom: $P(x_{(i)} < x < x_{(j)}) = F(x_{(j)}) - F(x_{(i)}) = y_{(j)} - y_{(i)} = d_{i,j};\ 1 \le i < j \le n$. Tato pravděpodobnost je tedy náhodnou veličinou s distribucí:

$$G(z) = P(y_{(j)} - y_{(i)} < z) = \iint\limits_{\substack{0<y-x<y \\ 0<x<1 \\ 0<y<1 \\ 0<z<1}} f_{(i,j)}(x,y)\,dxdy = (A)$$

Integrační oblast vypadá následovně:



Proto: $(A)=\int_0^{1-z}\int_x^{x+z} f_{(i,j)}(x,y)\,dy\,dx+\int_{1-z}^{1}\int_x^{1} f_{(i,j)}(x,y)\,dy\,dx$. Shrnuto:

$G(z)=\int_0^{1-z}\int_x^{x+z} f_{(i,j)}(x,y)\,dy\,dx+\int_{1-z}^{1}\int_x^{1} f_{(i,j)}(x,y)\,dy\,dx$ a odtud hustota

$$g(z)=\frac{d}{dz}G(z)=-\int_{1-z}^{1} f_{(i,j)}(x,y)\,dy+\int_0^{1-z} f_{(i,j)}(x,x+z)\,dx-\left(-\int_{1-z}^{1} f_{(i,j)}(x,y)\,dy\right)=\int_0^{1-z} f_{(i,j)}(x,x+z)\,dx$$

Opět shrnuto: $g(z)=\int_0^{1-z} f_{(i,j)}(x,x+z)\,dx$. Po dosazení za

$$f_{(i,j)}(x,y)=\frac{n!}{(i-1)!(j-i-1)!(n-j)!}x^{i-1}(y-x)^{j-i-1}(1-y)^{n-j}$$, dostaneme:

$$g(z)=\frac{n!}{(i-1)!(j-i-1)!(n-j)!}\int_0^{1-z}x^{i-1}(z)^{j-i-1}(1-x-z)^{n-j}\,dx=\frac{z^{j-i-1}n!}{(i-1)!(j-i-1)!(n-j)!}\int_0^{1-z}x^{i-1}(1-x-z)^{n-j}\,dx$$

Po výpočtu integrálu: $\int_0^{1-z}x^{i-1}(1-x-z)^{n-j}\,dx=\frac{(i-1)!(n-j)!}{(n-j+i)!}(1-z)^{n-j+i}$ (viz příloha 1.),

dostaneme:

$$g(z)=\frac{n!}{(i-1)!(j-i-1)!(n-j)!}z^{j-i-1}(1-z)^{n-j+i}\frac{(i-1)!(n-j)!}{(n-j+i)!}=\frac{n!}{(j-i-1)!(n-j+i)!}z^{j-i-1}(1-z)^{n-j+i}$$

Závěr: $g(z)=\frac{n!}{(j-i-1)!(n-j+i)!}z^{j-i-1}(1-z)^{n-j+i}$. Náhodná veličina

$P(x_{(i)}<x<x_{(j)})=F(x_{(j)})-F(x_{(i)})=y_{(j)}-y_{(i)}=d_{i,j};\ 1\le i<j\le n$ má BETA rozdělení s parametry $\alpha=j-i;\beta=n-j+i+1$, tedy $G_{d_{i,j}}(z)=\frac{n!}{(j-i-1)!(n-j+i)!}\int_0^{z}x^{j-i-1}(1-x)^{n-j+i}\,dx$.

Příloha 1.: Výpočet integrálu: $\int_0^{1-z}x^{i-1}(1-x-z)^{n-j}\,dx$

$\int_0^{1-z}x^{i-1}(1-x-z)^{n-j}\,dx=(A)$ $\quad v=\frac{x}{1-z}$

$x=v(1-z)$

$dx=dv(1-z)$

$$(A)=\int_0^{1}(v(1-z))^{i-1}(1-v(1-z)-z)^{n-j}(1-z)\,dv=$$

$$=(1-z)^{i}\int_0^{1}v^{i-1}(1-v+vz-z)^{n-j}\,dv=$$

$$=(1-z)^{i}\int_0^{1}v^{i-1}((1-v)-z(1-v))^{n-j}\,dv=$$

$$=(1-z)^{i}(1-z)^{n-i}\int_0^{1}v^{i-1}(1-v)^{n-j}\,dv=$$

$$=(1-z)^{n-j+i}\cdot B(i,n-j+1)=$$

$$=(1-z)^{n-j+i}\frac{(i-1)!(n-j)!}{(n-j+i)!}$$

$\alpha=i$

$\beta=n-j+1$

$\alpha+\beta=n-j+i+1$