Instytut Matematyczny
Polskiej Akademii Nauk

Paweł Marcin Kozyra

Optymalne oszacowania momentów kombinacji liniowych statystyk pozycyjnych i k-tych rekordów

streszczenie rozprawy doktorskiej

Promotor rozprawy
prof. dr hab. Tomasz Rychlik
Instytut Matematyczny
Polskiej Akademii Nauk

Warszawa, Styczeń 2017

Streszczenie

Dysertacja ta poświęcona jest wyznaczeniu optymalnych oszacowań wartości oczekiwanych i wariancji kombinacji liniowych statystyk pozycyjnych oraz k-tych rekordów skonstruowanych na bazie niezależnych zmiennych losowych o tym samym rozkładzie. Jest ona nowatorska w dwóch aspektach. Oszacowania wartości oczekiwanych zostały wyrażone w jednostkach średniej różnicy Giniego populacji, a oszacowania wariancji pojedynczych statystyk pozycyjnych i rekordów zostały uogólnione na przypadek nietrywialnych liniowych ich kombinacji. Zasadnicza idea naszego rozumowania polega na przedstawieniu wartości oczekiwanych, wariancji i kowariancji statystyk pozycyjnych i rekordowych w postaci całkowej w taki sposób, aby wyrażenie podcałkowe stanowiło złożenie pewnej funkcji (zwykle opisanej bardzo skomplikownym wzorem) z dystrybuantą rozważanych zmiennych losowych. Poniżej prezentujemy nasze ogólne rezultaty wraz z przykładowymi przypadkami szczególnymi.

Oszacowania wartości oczekiwanych L-statystyk. Przypuśćmy, że $X_1, \ldots, X_n$ są niezdegenerownymi, niezależnymi zmiennymi losowymi o tym samym rozkładzie ze skończoną średnią $\mu = \mathbb{E}X_1$. Niech $X_{1:n} \leqslant \ldots \leqslant X_{n:n}$ oznaczają ich statystyki pozycyjne. Najpierw koncentrujemy się na wyznaczeniu optymalnych oszacowań górnych i dolnych wartości oczekiwanych odpowiednio scentrowanych L-statystyk $\mathbb{E}\sum_{i=1}^n c_i(X_{i:n}-\mu)$, dla dowolnie ustalonych $c_1, \ldots, c_n \in \mathbb{R}$, oraz ich specjalnych przypadków. Są one wyrażone w jednostkach średniej różnicy Giniego $\Delta = \mathbb{E}|X_1 - X_2|$. Dla $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$ ze średnią arytmetyczną $\bar{c} = \frac{1}{n}\sum_{i=1}^n c_i$ definiujemy funkcję

$$\Xi_{\mathbf{c}}(u) = \sum_{i=0}^{n-2} \frac{n(n-1)}{2(i+1)(n-i-1)} \left[\sum_{k=1}^{i+1}(\bar{c} - c_k)\right] \binom{n-2}{i} u^i (1-u)^{n-2-i}$$

określoną na przedziale $[0, 1]$.

Twierdzenie (patrz Twierdzenie 2) *Przy powyższych założeniach i oznaczeniach, następujące oszacowania są optymalne*

$$\min_{0\leqslant u\leqslant 1} \Xi_{\mathbf{c}}(u) \leqslant \mathbb{E}\,\frac{\sum_{i=1}^n c_i(X_{i:n} - \mu)}{\Delta} \leqslant \max_{0\leqslant u\leqslant 1} \Xi_{\mathbf{c}}(u).$$

Jeśli $0 < u_0 < 1$ jest argumentem maksimum (odpowiednio minimum), to górne (dolne) oszacowanie jest osiągane wtedy i tylko wtedy, gdy dystrybuanta rozważanych zmiennych losowych ma postać

$$F(x) = \begin{cases} 0, & x < a, \\ u_0, & a \leqslant x < b, \\ 1, & x \geqslant b, \end{cases}$$

dla dowolnie ustalonych $a < b$.

Jeśli maksimum (minimum) wynosi albo $\Xi_{\mathbf{c}}(0)$ *albo* $\Xi_{\mathbf{c}}(1)$*, to górne (odpowiednio dolne) oszacowanie jest osiągane w granicy przez rozkłady dwu–punktowe takie, że prawdopodobieństwa mniejszego punktu zbiegają odpowiednio do* 0 *i* 1*.*

Jeżeli $\Xi_{\mathbf{c}}$ ma wielokrotne ekstrema, co zdarza się bardzo rzadko, to również inne rozkłady dyskretne osiągają te oszacowania. Nie omawiamy ich w tym streszczeniu. W celu skrócenia naszej prezentacji, będziemy też używać następującej konwencji. Pisząc, że dolne albo górne oszacowanie wynosi $\Xi_{\mathbf{c}}(u_0)$ dla pewnego $0 < u_0 < 1$, przyjmujemy milcząco, że oszacowanie to jest osiągane przez rozkłady dwu-punktowe opisane w powyższym twierdzeniu. Jeżeli oszacowanie jest równe $\Xi_{\mathbf{c}}(0)$ albo $\Xi_{\mathbf{c}}(1)$, to jest ono osiągane w granicy przez przez rozkłady dwu-punktowe o prawdopodobieństwie mniejszego punktu zbiegającym odpowiednio do 0 i 1.

Dla pojedynczej statystyki pozycyjnej $X_{r:n}$, $1 \leqslant r \leqslant n$, funkcja $\Xi_{\mathbf{c}}$ upraszcza się do

$$\Xi_{r:n}(u) = \sum_{i=0}^{r-2} \frac{n-1}{2(n-i-1)} \binom{n-2}{i} u^i (1-u)^{n-2-i} - \sum_{i=r-1}^{n-2} \frac{n-1}{2(i+1)} \binom{n-2}{i} u^i (1-u)^{n-2-i}.$$

Stwierdzenie (patrz Stwierdzenie 1) *(i) Dla skrajnych statystyk pozycyjnych mamy*

$$\begin{aligned} \Xi_{1:n}(0) = -\frac{n-1}{2} &\leqslant \mathbb{E}\frac{X_{1:n}-\mu}{\Delta} \leqslant \Xi_{1:n}(1) = -\frac{1}{2}, \\ \Xi_{n:n}(0) = \frac{1}{2} &\leqslant \mathbb{E}\frac{X_{n:n}-\mu}{\Delta} \leqslant \Xi_{n:n}(1) = \frac{n-1}{2}. \end{aligned}$$

(ii) Dla sąsiadujących ze skrajnymi statystyk pozycyjnych, pochodne $\Xi'_{r:n}(u)$*,* $r = 2, n-1$*, mają jedyne miejsca zerowe, odpowiednio* $v_1(2)$ *i* $u_1(n-1) = 1 - v_1(2)$*, i zachodzą nierówności*

$$\begin{aligned} \Xi_{2:n}(v_1(2)) &\leqslant \mathbb{E}\frac{X_{2:n}-\mu}{\Delta} \leqslant \Xi_{2:n}(0) = \frac{1}{2}, \\ \Xi_{n-1:n}(1) = -\frac{1}{2} &\leqslant \mathbb{E}\frac{X_{n-1:n}-\mu}{\Delta} \leqslant \Xi_{n-1:n}(u_1(n-1)). \end{aligned}$$

(iii) Dla $3 \leqslant r \leqslant n-2$*,* $\Xi'_{r:n}(u)$ *ma dwa miejsca zerowe* $u_1(r) < v_1(r)$ *w* $(0,1)$*, i wówczas*

$$\Xi_{r:n}(v_1(r)) \leqslant \mathbb{E}\frac{X_{r:n}-\mu}{\Delta} \leqslant \Xi_{r:n}(u_1(r)).$$

W podobny sposób szacujemy wartości oczekiwane różnic dwu statystyk pozycyjnych, uciętych i Winsoryzowanych średnich, ich różnic oraz średniego absolutnego odchylenia od mediany.

Oszacowania wariancji kombinacji liniowych statystyk pozycyjnych. Rozważamy niezależne zmienne losowe $X_1, \ldots, X_n$ o tym samym rozkładzie z niezerową i skończoną wariancją. Definiujemy

$$\begin{aligned}\Phi_{\mathbf{c}}(u,v) &= \left[\sum_{i=0}^{n-1} \frac{n}{i+1}\left(\sum_{k=1}^{i+1} c_k\right)\binom{n-1}{i} u^i (1-u)^{n-1-i}\right] \\ &\times \left[\sum_{j=0}^{n-1} \frac{n}{n-j}\left(\sum_{m=j+1}^{n} c_m\right)\binom{n-1}{j} v^j (1-v)^{n-1-j}\right] \\ &- \sum_{i=0}^{n-2}\sum_{j=i}^{n-2} \frac{n(n-1)}{(i+1)(n-1-j)}\left(\sum_{k=1}^{i+1} c_k\right)\left(\sum_{m=j+2}^{n} c_m\right) \\ &\times \frac{n!}{i!(j-i)!(n-j)!} u^i (v-u)^{j-i}(1-v)^{n-2-j}.\end{aligned}$$

dla $0 < u \leqslant v < 1$. Niech $\Psi_{\mathbf{c}}(u) = \Phi_{\mathbf{c}}(u,u)$.

Twierdzenie (patrz Twierdzenie 3) *Przy powyższych założeniach i oznaczeniach, zachodzi*

$$\frac{\mathbb{V}ar(\sum_{i=1}^n c_i X_{i:n})}{\mathbb{V}ar\, X_1} \leqslant \sup_{0<u\leqslant v<1} \Phi_{\mathbf{c}}(u,v).$$

Ponadto, jeżeli

$$\sup_{0<u\leqslant v<1} \Phi_{\mathbf{c}}(u,v) = \sup_{0<u<1} \Psi_{\mathbf{c}}(u),$$

to powyższe oszacowanie jest optymalne.

Pokazujemy także, że dolne oszacowanie ilorazu $\frac{\mathbb{V}ar(\sum_{i=1}^n c_i X_{i:n})}{\mathbb{V}ar\, X_1}$ jest równe 0 wtedy i tylko wtedy, gdy $c_1 c_n = 0$. W przypadku spacji $S_{i:n} = X_{i+1:n} - X_{i:n}$, $1 \leqslant i < n < \infty$, funkcja $\Psi_{\mathbf{c}}$ ma postać

$$\Psi_{i:n}(u) = \binom{n}{i} u^{i-1}(1-u)^{n-i-1}\left[1 - \binom{n}{i} u^i (1-u)^{n-i}\right].$$

Stwierdzenie (patrz Stwierdzenia 9 i 10) *Następujące oszacowania są optymalne.*
(i) Jeśli $n \geqslant 3$, to

$$\begin{aligned} 0 = \Psi_{1:n}(1) &\leqslant \frac{\mathbb{V}ar(X_{2:n} - X_{1:n})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{1:n}(0) = n, \\ 0 = \Psi_{n-1:n}(0) &\leqslant \frac{\mathbb{V}ar(X_{n:n} - X_{n-1:n})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{n-1:n}(1) = n. \end{aligned}$$

(ii) Jeśli $n \geqslant 4$ i $2 \leqslant i \leqslant n-2$, to pochodna $\Psi'_{i:n}$ ma albo jedno albo trzy miejsca zerowe, i wtedy

$$0 = \Psi_{i:n}(0) = \Psi_{i:n}(1) \leqslant \frac{Var(X_{i+1:n} - X_{i:n})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{i:n}(u_0),$$

gdzie u_0 jest albo pojedynczym miejscem zerowym pochodnej albo pierwszym bądź trzecim miejscem zerowym $\Psi'_{i:n}$. Spośród nich wybierany jest ten argument, dla którego wartość $\Psi_{i:n}$ jest większa.

Przypadek $n = 2$ oraz $i = 1$, dla którego dolne oszacowanie jest dodatnie, nie jest omawiany w tym streszczeniu.

Wartości oczekiwane kombinacji liniowych rekordów. Niech $X_1, X_2, \ldots$ będą niezależnymi zmiennymi losowymi o tym samym rozkładzie ze skończoną średnią μ. Ponadto, niech $R_{1,k}, R_{2,k}, \ldots$ oznaczają odpowiednie wartości k-tych rekordów. Załóżmy, że $\mathbb{E}R_{n,k} < \infty$ dla pewnych ustalonych n i k. Poniżej opisujemy optymalne dolne i górne oszacowania wartości oczekiwanych kombinacji liniowych k-tych rekordów $\mathbb{E}\left[\sum_{i=1}^n c_i(R_{i,k} - \mu)\right]$, scentrowanych względem średniej populacji, i wyrażonych w jednostkach średniej różnicy Giniego $\Delta = \mathbb{E}|X_1 - X_2|$. Używamy następujących oznaczeń

$$\begin{aligned}
\Xi_{n,k}(u) &= \frac{1}{2u}\left[(1-u)^{k-1}\sum_{i=0}^{n-1}\frac{\left[-k\ln(1-u)\right]^i}{i!} - 1\right], \\
\Xi_{\mathbf{c},k}(u) &= \sum_{i=1}^n c_i\Xi_{i,k}(u) = \frac{1}{2u}\left[(1-u)^{k-1}\sum_{i=0}^{n-1}\left(\sum_{j=i+1}^n c_j\right)\frac{[-k\ln(1-u)]^i}{i!} - \sum_{j=1}^n c_j\right],
\end{aligned}$$

gdzie $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$.

Twierdzenie (patrz Twierdzenie 5) *Przy powyższych założeniach i oznaczeniach następujące oszacowania*

$$\inf_{0<u<1}\Xi_{\underline{c},k}(u) \leqslant \frac{\mathbb{E}\left[\sum_{i=1}^n c_i(R_{i,k} - \mu)\right]}{\Delta} \leqslant \sup_{0<u<1}\Xi_{\underline{c},k}(u)$$

są optymalne.

Warunki osiągalności są podobne do tych z problemów oszacowywania wartości oczekiwanych L-statystyk. Jedyna różnica polega na tym, że dokładne wartości w punktach 0 i 1 są zastąpione przez odpowiednie granice, a dyskretne rozkłady dwu-punktowe są zastąpione przez ich ciągłe przybliżenia. Zgodnie z powyższym, wszystkie oszacowania dla kombinacji rekordów są osiągane w granicy. Dla pojedynczych rekordów $R_{n,k}$, zachodzi $\Xi_{\mathbf{c},k} = \Xi_{n,k}$.

Stwierdzenie (patrz Stwierdzenie 12) *Dla różnych liczb naturalych $n \geqslant 2$ i $k \geqslant 1$, mamy następujące optymalne oszacowania.*
(i) Dla $n \geqslant 2$ i $k = 1$, zachodzi

$$\frac{1}{2} = \Xi_{n,1}(0+) \leqslant \frac{\mathbb{E}(R_{n,1} - \mu)}{\Delta} \leqslant \Xi_{n,1}(1-) = \infty.$$

(ii) Jeżeli $n = k = 2$, *to*

$$-\frac{1}{2} = \Xi_{2,2}(1-) \leqslant \frac{\mathbb{E}(R_{2,2} - \mu)}{\Delta} \leqslant \Xi_{2,2}(0+) = \frac{1}{2}.$$

(iii) Dla $n \geqslant 3$ *i* $k = 2$

$$-\frac{1}{2} = \lim_{u \nearrow 1^-} \Xi_{n,2}(1-) \leqslant \frac{\mathbb{E}(R_{n,2} - \mu)}{\Delta} \leqslant \Xi_{n,2}(u_1) > \frac{1}{2},$$

gdzie u_1 *jest jedynym miejscem zerowym* $\Xi'_{n,2}$ *w przedziale* $(0, 1)$.
(iv) Dla $n = 2$ *oraz* $k \geqslant 3$

$$-\frac{1}{2} > \Xi_{2,k}(u_1) \leqslant \frac{\mathbb{E}(R_{2,k} - \mu)}{\Delta} \leqslant \Xi_{2,k}(0+) = \frac{1}{2},$$

gdzie $u_1 \in (0, 1)$ *jest jedynym miejscem zerowym* $\Xi'_{2,k}$ *in* $(0, 1)$.
(v) Dla $k \geqslant 3$ *i* $n \geqslant 3$, *mamy*

$$-\frac{1}{2} > \Xi_{n,k}(u_2) \leqslant \frac{\mathbb{E}(R_{n,k} - \mu)}{\Delta} \leqslant \Xi_{n,k}(u_1) > \frac{1}{2},$$

gdzie $0 < u_1 < u_2 < 1$ *są jedynymi rozwiązaniami równania* $\Xi'_{n,k}(u) = 0$.

Podobnie wyznaczamy oszacowania wartości oczekiwanych różnic wartości k-tych rekordów $\mathbb{E}(R_{n,k} - R_{m,k})$, $1 \leqslant m < n$.

Oszacowania wariancji kombinacji liniowych rekordów. Niech $X_1, X_2, \ldots$ będą niezależnymi zmiennymi losowymi o tym samym rozkładzie i skończonym drugim momencie. Załóżmy też, że $\mathbb{E}R_{n,k}^2 < \infty$. Dla danych liczb naturalnych n i k, oraz dla ustalonego niezerowego wektora $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$, definiujemy funkcję

$$\begin{aligned}
\Phi_{\mathbf{c},k}(u,v) &= \frac{(1-v)^{k-1}}{u} \left\{ \left[\sum_{j=1}^{n} c_j - (1-u)^k \sum_{i=0}^{n-1} \left(\sum_{j=i+1}^{n} c_j \right) \frac{[-k\ln(1-u)]^i}{i!} \right] \right. \\
&\times \sum_{i=0}^{n-1} \left(\sum_{j=i+1}^{n} c_j \right) \frac{[-k\ln(1-v)]^i}{i!} \\
&- \left. \sum_{1 \leqslant i < j \leqslant n} c_i c_j \sum_{p=0}^{j-i-1} \sum_{q=0}^{p} \frac{(-1)^q [-k\ln(1-u)]^{i+q} [-k\ln(1-v)]^{p-q}}{(i-1)!q!(p-q)!(p+i)} \right\}
\end{aligned}$$

określoną na trójkącie $0 < u \leqslant v < 1$, a także $\Psi_{\mathbf{c},k}(u) = \Phi_{\mathbf{c},k}(u,u)$, $0 < u < 1$.

Twierdzenie (patrz Twierdzenie 6) *Przy powyższych warunkach i oznaczeniach mamy*

$$\frac{\mathbb{V}ar(\sum_{i=1}^{n} c_i R_{i,k})}{\mathbb{V}ar\, X_1} \leqslant \sup_{0<u\leqslant v<1} \Phi_{\mathbf{c},k}(u,v).$$

Ponadto, jeśli

$$\sup_{0<u\leqslant v<1} \Phi_{\mathbf{c},k}(u,v) = \sup_{0<u<1} \Psi_{\mathbf{c},k}(u),$$

to oszacowanie jest optymalne.

Dowodzimy także, że dolne oszacowania wariancji kombinacji liniowych k-tych rekordów są równe 0, za wyjątkiem przypadku $k = 1$ i $c_1 \neq 0$. Dla spacji k-tych rekordów $R_{m+1,k} - R_{m,k}$, funkcja $\Psi_{\mathbf{c},k}$ ma prostszą postać

$$\Psi_{m,k}(u) = \frac{[-k\ln(1-u)]^m(1-u)^{k-1}}{um!}\left[1 - \frac{(1-u)^k[-k\ln(1-u)]^m}{m!}\right].$$

Stwierdzenie (patrz Stwierdzenie 14) *Następujące oszacowania są optymalne.*
(i) Jeśli $k = 1$ i $m \geqslant 1$, to

$$\frac{\mathbb{V}ar(R_{m+1,1} - R_{m,1})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{m,1}(1-) = +\infty.$$

(ii) Jeżeli $m = 1$ oraz $k \geqslant 2$, to

$$\frac{\mathbb{V}ar(R_{2,k} - R_{1,k})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{1,k}(0+) = k.$$

(iii) Jeśli albo $k = 2 \leqslant m$ albo $k \geqslant 3$ oraz $2 \leqslant m \leqslant \frac{2}{3}k$, to

$$\frac{\mathbb{V}ar(R_{m+1,k} - R_{m,k})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{m,k}(u_0),$$

gdzie $0 < u_0 < 1$ jest jedynym rozwiązaniem równania $\Psi'_{m,k}(u) = 0$.
(iv) Jeżeli wreszcie $k \geqslant 3$ oraz $m > \frac{2}{3}k$, to

$$\frac{\mathbb{V}ar(R_{m+1,k} - R_{m,k})}{\mathbb{V}ar\, X_1} \leqslant \Psi_{m,k}(u_0),$$

gdzie $0 < u_0 < 1$ argumentem, dla którego funkcja $\Psi_{m,k}$ osiąga swoje globalne maksimum na przedziale $(0, 1)$.

W ostatnim przypadku, formalnie możemy jedynie udowodnić, że $\Psi_{m,k}$ ma co najwyżej trzy lokalne maksima.