

Przedział ufności dla mediany w przypadku populacji dyskretnej

Marta Sommer

8 kwietnia 2014

X_1, X_2, \dots, X_n – obserwacje

Hipoteza:

$$\begin{cases} H : M = M_0 \\ K : M \neq M_0 \end{cases}$$

Statystyka testowa:

$$T = \sum_{i=1}^n \mathbb{I}_{\{X_i > M_0\}}$$

Rozkład statystyki testowej:

$$T \sim \text{Bin}(n, \frac{1}{2})$$

Hipotezę H odrzucamy więc, gdy $T \geq k_{\frac{\alpha}{2}}$ lub $T \leq k'_{\frac{\alpha}{2}}$, gdzie $k_{\frac{\alpha}{2}}$ jest najmniejszą liczbą naturalną spełniającą nierówność:

$$\sum_{i=k_{\frac{\alpha}{2}}}^n \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2},$$

zaś $k'_{\frac{\alpha}{2}}$ jest największą liczbą naturalną spełniającą nierówność:

$$\sum_{i=0}^{k'_{\frac{\alpha}{2}}} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2}.$$

Hipotezy H nie odrzucamy natomiast w następującym przypadku:

$$k'_{\frac{\alpha}{2}} < T < k_{\frac{\alpha}{2}} \quad \equiv \quad k'_{\frac{\alpha}{2}} + 1 \leq T \leq k_{\frac{\alpha}{2}} - 1.$$

Lemat

Niech $r, s \in \{1, \dots, n\}$, $r < s$ oraz niech T będzie statystyką testu znaków. Wówczas prawdziwa jest teza:

$$\mathbb{P}(X_{r:n} \leq M \leq X_{s:n}) = \mathbb{P}(r \leq T(X_1, \dots, X_n) \leq s - 1).$$

$$k'_{\frac{\alpha}{2}} + 1 \leq T \leq k_{\frac{\alpha}{2}} - 1$$

Przyjmując $r := k'_{\frac{\alpha}{2}} + 1$ i $s := k_{\frac{\alpha}{2}}$, otrzymujemy przedział ufności

$$X_{(k'_{\frac{\alpha}{2}}+1):n} \leq M \leq X_{k_{\frac{\alpha}{2}}:n}$$

na poziomie $1 - \alpha$.

Uwaga

Wszystkie te rozważania przeprowadzane są przy założeniach ciągłości rozkładu oraz zerowego prawdopodobieństwa wystąpienia mediany!

Jak wygląda sytuacja dla rozkładu dyskretnego? Co dzieje się w przypadkach, gdy mediana występuje nawet kilkakrotnie?

Będziemy rozważać przedziały ufności oparte na statystyce porządkowej i mające następującą postać:

$$[X_{d:n}, X_{(n+1-d):n}]$$

Dla populacji ciągłej powyższy przedział wyznaczony jest na poziomie ufności równym

$$1 - \alpha = 1 - 2 \cdot \mathbb{P}(B \leq d - 1),$$

gdzie $B \sim \text{Bin}(n, \frac{1}{2})$.

Wynika to z dość prostego rachunku wykorzystującego to, że przy założeniu hipotezy zerowej $\mathbb{P}(B \geq t) = \mathbb{P}(B \leq n - t)$.

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\mathbb{P}(B \geq t) =$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\mathbb{P}(B \geq t) = \left| B \sim \text{Bin}(n, \frac{1}{2}) \right|$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\mathbb{P}(B \geq t) = \left| B \sim \text{Bin}(n, \frac{1}{2}) \right| \stackrel{H}{=} \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} =$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\begin{aligned} \mathbb{P}(B \geq t) &= \left| B \sim \text{Bin}(n, \frac{1}{2}) \right| \stackrel{H}{=} \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \\ &= \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^n \end{aligned}$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\begin{aligned}\mathbb{P}(B \geq t) &= \left| B \sim \text{Bin}(n, \frac{1}{2}) \right| \stackrel{H}{=} \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \\ &= \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n \sum_{i=t}^n \binom{n}{n-i}\end{aligned}$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\begin{aligned} \mathbb{P}(B \geq t) &= \left| B \sim \text{Bin}(n, \frac{1}{2}) \right| \stackrel{H}{=} \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \\ &= \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n \sum_{i=t}^n \binom{n}{n-i} = \\ &= \left(\frac{1}{2}\right)^n \sum_{i=0}^{n-t} \binom{n}{i} \end{aligned}$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\begin{aligned} \mathbb{P}(B \geq t) &= \left| B \sim \text{Bin}(n, \frac{1}{2}) \right| \stackrel{H}{=} \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \\ &= \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n \sum_{i=t}^n \binom{n}{n-i} = \\ &= \left(\frac{1}{2}\right)^n \sum_{i=0}^{n-t} \binom{n}{i} = \sum_{i=0}^{n-t} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} \end{aligned}$$

$$\mathbb{P}(B \geq t) \stackrel{?}{=} \mathbb{P}(B \leq n - t)$$

$$\begin{aligned} \mathbb{P}(B \geq t) &= \left| B \sim \text{Bin}(n, \frac{1}{2}) \right| \stackrel{H}{=} \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \\ &= \sum_{i=t}^n \binom{n}{i} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^n \sum_{i=t}^n \binom{n}{n-i} = \\ &= \left(\frac{1}{2}\right)^n \sum_{i=0}^{n-t} \binom{n}{i} = \sum_{i=0}^{n-t} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} = \\ &= \mathbb{P}(B \leq n - t) \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$1 - \alpha$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$1 - \alpha = \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}])$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$1 - \alpha = \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n})$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\ &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\ &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned} 1 \quad - \quad \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\ &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) = \\ &= \mathbb{P}(T \leq n - d) - \mathbb{P}(T < d) \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned}
 1 \quad - \quad \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\
 &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) = \\
 &= \mathbb{P}(T \leq n - d) - \mathbb{P}(T < d) = \mathbb{P}(T \leq n - d) - \mathbb{P}(T \leq d - 1)
 \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned}
 1 \quad - \quad \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\
 &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) = \\
 &= \mathbb{P}(T \leq n - d) - \mathbb{P}(T < d) = \mathbb{P}(T \leq n - d) - \mathbb{P}(T \leq d - 1) = \\
 &= \mathbb{P}(T \geq d) - \mathbb{P}(T \leq d - 1)
 \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned}
 1 \quad - \quad \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\
 &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) = \\
 &= \mathbb{P}(T \leq n - d) - \mathbb{P}(T < d) = \mathbb{P}(T \leq n - d) - \mathbb{P}(T \leq d - 1) = \\
 &= \mathbb{P}(T \geq d) - \mathbb{P}(T \leq d - 1) = 1 - \mathbb{P}(T < d) - \mathbb{P}(T \leq d - 1)
 \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned}
 1 - \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\
 &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) = \\
 &= \mathbb{P}(T \leq n - d) - \mathbb{P}(T < d) = \mathbb{P}(T \leq n - d) - \mathbb{P}(T \leq d - 1) = \\
 &= \mathbb{P}(T \geq d) - \mathbb{P}(T \leq d - 1) = 1 - \mathbb{P}(T < d) - \mathbb{P}(T \leq d - 1) = \\
 &= 1 - \mathbb{P}(T \leq d - 1) - \mathbb{P}(T \leq d - 1)
 \end{aligned}$$

$$1 - \alpha \stackrel{?}{=} 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

$$\begin{aligned}
 1 - \alpha &= \mathbb{P}(M \in [X_{d:n}, X_{(n+1-d):n}]) = \mathbb{P}(X_{d:n} \leq M \leq X_{(n+1-d):n}) = \\
 &= \mathbb{P}(d \leq T \leq n + 1 - d - 1) = \mathbb{P}(d \leq T \leq n - d) = \\
 &= \mathbb{P}(T \leq n - d) - \mathbb{P}(T < d) = \mathbb{P}(T \leq n - d) - \mathbb{P}(T \leq d - 1) = \\
 &= \mathbb{P}(T \geq d) - \mathbb{P}(T \leq d - 1) = 1 - \mathbb{P}(T < d) - \mathbb{P}(T \leq d - 1) = \\
 &= 1 - \mathbb{P}(T \leq d - 1) - \mathbb{P}(T \leq d - 1) = 1 - 2 \cdot \mathbb{P}(T \leq d - 1)
 \end{aligned}$$

Uwaga

Prawdziwym problemem jest określenie poziomu ufności dla znalezionego już przedziału ufności.

Ze względu na to, że dane są dyskretne, istnieje skończenie wiele (ewentualnie przeliczalnie wiele) możliwych przedziałów ufności, a tym samym – skończenie wiele poziomów ufności możliwych do osiągnięcia.

Poziom ufności wyznaczamy wprost ze wzoru

$$1 - \alpha = 1 - 2 \cdot \mathbb{P}(B \leq d - 1),$$

ignorując to, że dane są dyskretne.

Rozważamy poziomy ufności wyznaczone tym samym wzorem, co w metodzie 1, dla przedziału ufności $[X_{d:n}, X_{(n+1-d):n}]$:

$$1 - \alpha = 1 - 2 \cdot \mathbb{P}(B \leq d - 1)$$

oraz poziomy ufności wyznaczone dla nieznacznie niesymetrycznego przedziału ufności $[X_{(d+1):n}, X_{(n+1-d):n}]$:

$$1 - \alpha = 1 - \mathbb{P}(B \leq d - 1) - \mathbb{P}(B \leq d).$$

Wyznaczamy wszystkie możliwe przedziały i wybieramy ten, który ma najmniejszy poziom ufności większy lub równy 0,95.

Rozważamy symetryczny przedział ufności

$$[X_{d:n}, X_{(n+1-d):n}].$$

Stosując tę metodę, weźmiemy pod uwagę ewentualne obserwacje związane.

Niech r będzie najmniejszym indeksem, dla którego

$$X_{r:n} = X_{d:n}$$

oraz niech s będzie największym indeksem, dla którego

$$X_{s:n} = X_{(n+1-d):n}$$

Poziom ufności wyraża się wtedy wzorem:

$$1 - \alpha = 1 - \mathbb{P}(B \leq r - 1) - \mathbb{P}(B \leq n - s).$$

Rozważamy przedział ufności $[X_{d:n}, X_{(n+1-d):n}]$.

Metody te stosuje się, opierając się na odwróceniu dwustronnego testu znaków z obserwacjami związanymi.

Poziom ufności wyraża się następująco:

$$1 - \alpha = 1 - \max \{p.value(X_{d:n}-), p.value(X_{(n+1-d):n}+)\},$$

gdzie $p.value(c)$ to p-value dwustronnego testu o hipotezie:

$$\begin{cases} H : M = c \\ K : M \neq c \end{cases}$$

$X_{d:n}-$ oznacza pierwszą możliwą obserwację poniżej $X_{d:n}$, zaś

$X_{(n+1-d):n}+$ oznacza pierwszą możliwą obserwację powyżej

$X_{(n+1-d):n}$.

Korzystając z p-value wyznaczonego dla dwustronnego testu znaków przy założeniu ciągłości rozkładu, przyjmujemy:

$$p.value(c) = 2 \cdot \mathbb{P}(B \leq \min \{n_+^c, n_-^c\}),$$

gdzie n_+^c to liczba obserwacji większych niż c , zaś n_-^c to liczba obserwacji mniejszych niż c .

Rozważmy teraz sytuację podobną do tej z metody 4, biorąc jednak pod uwagę, że mamy do czynienia z rozkładem dyskretnym. W szczególności możemy przyjąć, że $X_{d:n-} = X_{d:n} - 1$, a $X_{(n+1-d):n+} = X_{(n+1-d):n} + 1$. Wtedy:

$$1 - \alpha = 1 - \max \{p.value(X_{d:n} - 1), p.value(X_{(n+1-d):n} + 1)\}.$$

Może się jednak zdarzyć, że obserwacja $X_{d:n} - 1$ nie wystąpi wcale lub wystąpi kilkakrotnie. Należy zatem policzyć p-value, biorąc pod uwagę obserwacje związane.

Wstęp do metod 5–7

Wprowadźmy następujące oznaczenia:

n_+^c – liczba obserwacji większych niż c ,

n_-^c – liczba obserwacji mniejszych niż c ,

n_0^c – liczba obserwacji równych c .

Rozważmy dwustronny test znaków, który odrzuca hipotezę dla dużych wartości

$$n_*^c = \max \{n_+^c, n_-^c\}.$$

P-value takiego testu będzie wyglądała następująco:

$$p.value(c) = \mathbb{P}(N_* \geq n_*^c \mid \tilde{p}_+, \tilde{p}_-, \tilde{p}_0),$$

gdzie $N_* = \max \{N_+, N_-\}$, a (N_+, N_0, N_-) ma rozkład wielomianowy z parametrami n i $(\tilde{p}_+, \tilde{p}_0, \tilde{p}_-)$, które spełniają warunki: $0 \leq \tilde{p}_+ \leq \frac{1}{2}$ oraz $0 \leq \tilde{p}_- \leq \frac{1}{2}$.

Ustaliwszy parametry $(\tilde{p}_+, \tilde{p}_0, \tilde{p}_-)$, wyznaczymy p-value, a tym samym poziom ufności.

Metody 5–7 pozwolą na różne sposoby przybliżać wartości $(\tilde{p}_+, \tilde{p}_0, \tilde{p}_-)$.

Metoda 5

Dla $n_*^c \leq \frac{n}{2}$:

$$\tilde{p}_{+mle} = \frac{n_+^c}{n}, \quad \tilde{p}_{0mle} = \frac{n_0^c}{n}, \quad \tilde{p}_{-mle} = \frac{n_-^c}{n}$$

Dla $n_+^c > \frac{n}{2}$:

$$\tilde{p}_{+mle} = \frac{1}{2}, \quad \tilde{p}_{0mle} = \frac{n_0^c}{2(n - n_+^c)}, \quad \tilde{p}_{-mle} = \frac{n_-^c}{2(n - n_+^c)}$$

Dla $n_+^c = n$:

$$\tilde{p}_{+mle} = \tilde{p}_{-mle} = \frac{1}{2}, \quad \tilde{p}_{0mle} = 0$$

Analogicznie dla $n_-^c > \frac{1}{2}$.

Metoda 6

Szukamy $(\tilde{p}_+, \tilde{p}_0, \tilde{p}_-)$, minimalizując wyrażenie:

$$\left(\frac{n_+^c}{n} - p_+\right)^2 + \left(\frac{n_0^c}{n} - p_0\right)^2 + \left(\frac{n_-^c}{n} - p_-\right)^2$$

przy warunkach $0 \leq \tilde{p}_+ \leq \frac{1}{2}$ oraz $0 \leq \tilde{p}_- \leq \frac{1}{2}$.

Rozwiązania tego problemu wyglądają następująco:

$$\tilde{p}_{+cql} = \frac{n_+^c}{n}, \quad \tilde{p}_{0cql} = \frac{n_0^c}{n}, \quad \tilde{p}_{-cql} = \frac{n_-^c}{n},$$

dla $n_*^c \leq \frac{n}{2}$.

$$\tilde{p}_{+cql} = \frac{1}{2}, \quad \tilde{p}_{0cql} = \frac{n_0^c}{n} + \frac{1}{2} \left(\frac{n_+^c}{n} - \frac{1}{2} \right), \quad \tilde{p}_{-cql} = \frac{n_-^c}{n} + \frac{1}{2} \left(\frac{n_+^c}{n} - \frac{1}{2} \right),$$

dla $n_+^c > \frac{n}{2}$. Analogicznie dla $n_-^c > \frac{1}{2}$.

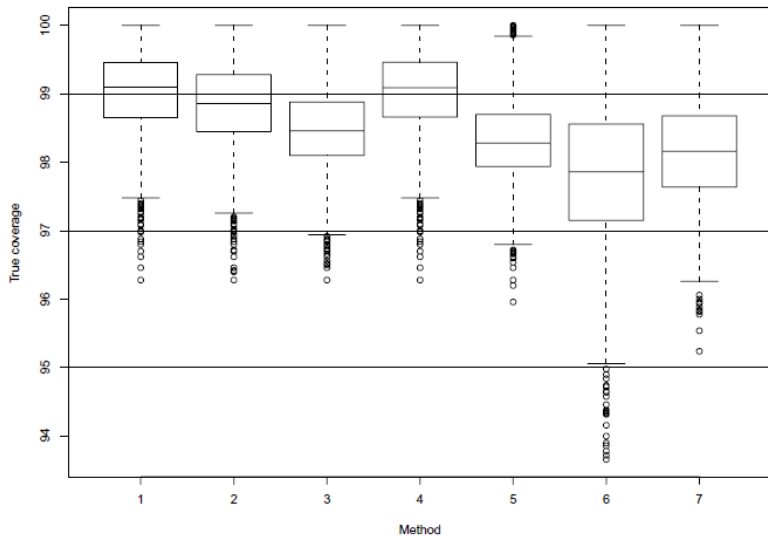
Jest nieznaczną modyfikacją metody 6.

Dla $n_0^c > 0$ zachodzi równość $(\tilde{p}_+, \tilde{p}_0, \tilde{p}_-) = (\tilde{p}_{+cql}, \tilde{p}_{0cql}, \tilde{p}_{-cql})$.

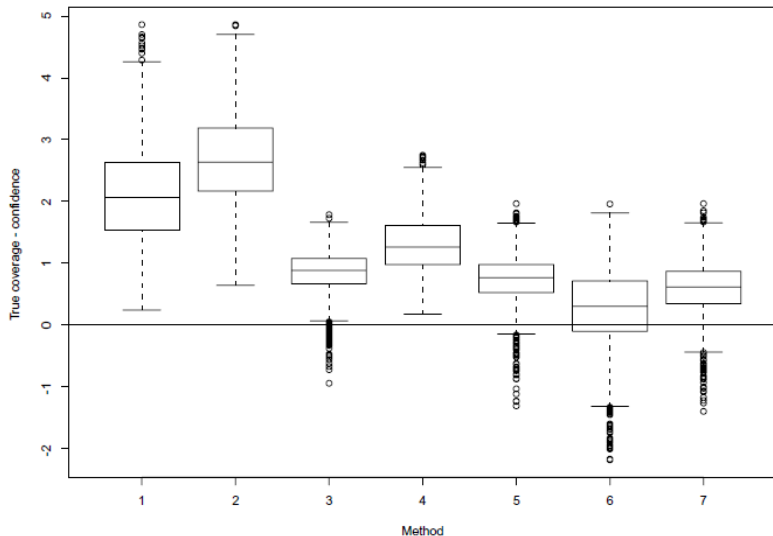
Zmiana następuje dla $n_0^c = 0$. Ma wtedy miejsce równość

$$(\tilde{p}_+, \tilde{p}_0, \tilde{p}_-) = \left(\frac{1}{2}, 0, \frac{1}{2} \right).$$

Porównanie metod



Porównanie metod, cd.



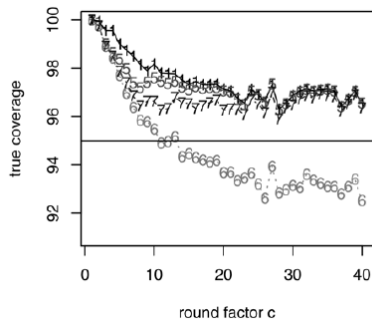
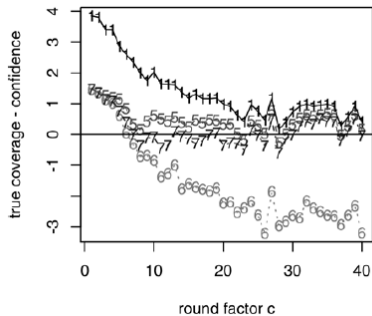
Intra-configuration rankings of
absolute difference between
reported confidence and true coverage

Method	Average rank
6	1.78
7	2.15
5	3.13
3	3.45
4	4.55
1	6.27
2	6.67

Intra-configuration rankings of
confidence interval length




Method	Average rank
6	1.56
7	2.61
5	3.00
3	3.94
2	5.11
4	5.88
1	5.88

Porównanie metod



Porównanie metod

Method	SIM scores ($n=84$) Sample median = 14			Ticks on sheep ($n=82$) Sample median = 5			Reading scores ($n=116$) Sample median = 14		
	lower	upper	confidence	lower	upper	confidence	lower	upper	confidence
1	10	18	96.25	4	6	96.48	11	16	96.77
2	10	18	96.25	4	6	95.25	11	16	95.85
3	10	18	98.84	4	5	96.02	12	16	95.14
4	10	18	98.84	4	6	98.02	11	16	98.01
5	10	18	99.41	4	5	96.70	12	16	96.05
6	11	16	96.63	4	5	96.99	12	16	96.13
7	10	18	99.42	4	5	96.99	12	16	96.13

-  Larocque, Denis, Randles, Ronald H. Confidence Intervals for a Discrete Population Median, *The American Statistician* 2008, nr 62, s. 32–39.
-  Emerson, J.D., Simon, G.A. Another Look at the Sign Test when Ties are Present: the Problem of Confidence Intervals, *The American Statistician* 1979, nr 33, s. 140–142.
-  Fong, D.Y.T., Kwan, C.W., Lam, K.F., Lam, K.S.L. Use of the Sign Test for the Median in the Presence of Ties, *The American Statistician* 2003, nr 58, s. 237–240.