Metody bootstrapowe w statystyce

Jacek Gulgowski, Barbara Wolnik

Instytut Matematyki, Uniwersytet Gdański

2013-02-09

Szkic wystąpienia

1. Podstawowe zagadnienia i problemy statystyki
2. Opis metody bootstrap
3. Zastosowania metody bootstrap
4. Trochę uwag technicznych

Statystyka – pierwsze przybliżenie

- Populacja
- Badana cecha i jej rozkład w populacji
- Próba

Pytanie

Jak na podstawie zebranej próby oszacować rozkład badanej cechy w populacji?

Statystyka – spojrzenie matematyka

- Rozkład prawdopodobieństwa cechy w populacji – dany przez dystrybuantę F.
- Próba: ciąg realizacji niezależnych zmiennych losowych X_n o rozkładzie F.
- Estymacja pewnej charakterystyki liczbowej rozkładu badanej cechy w populacji: statystyka T jako funkcja liczona na podstawie próby.

Czy próba potrafi powiedzieć coś o rozkładzie, z którego pochodzi?
Dystrybuanta empiryczna z próby

$$F_n(x) = \frac{\#\{1 \leqslant j \leqslant n : \ X_j \leqslant x\}}{n}, \quad x \in \mathbb{R}$$

spełnia:

1) dla każdego $x \in \mathbb{R}$ zachodzi równość $E_F\left(F_n(x)\right) = F(x)$,
2) dla każdego $x \in \mathbb{R}$ mamy $P_F\left(\lim_{n\to\infty} F_n(x) = F(x)\right) = 1$,
3) dla każdego $x \in \mathbb{R}$ rozkład zmiennej losowej

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}},$$

przy $n \to \infty$ dąży do rozkładu normalnego $N(0,1)$.

Podstawowe twierdzenie statystyki matematycznej

Twierdzenie (Gliwienki-Cantelliego)

$$P_F\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|=0\right)=1$$

Jak estymować wartość badanego parametru

Jakie charakterystyki liczbowe rozkładu badanej cechy (badanych cech) w populacji próbujemy opisać?

- wartość średnią;
- wariancję (odchylenie standardowe);
- medianę;
- współczynnik korelacji zmiennych losowych;
- różnicę wartości średnich dwóch zmiennych losowych;
- współczynniki regresji liniowej;
- co tylko nam przyjdzie do głowy...

Jak estymować wartość badanego parametru

Zwykle nie trzeba zastanawiać się nad tym JAK estymować wartość danego parametru na podstawie zebranej próby $x_1, x_2, ..., x_n$, np. wiadomo, że estymatorem optymalnym dla wartości średniej jest

$$T(X_1, .., X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Jest to bowiem estymator nieobciążony, zgodny i najefektywniejszy. Zastosowanie innego estymatora (nieobciążonego) pogorszyłoby precyzję szacunku.

Co więc stanowi problem?

Z punktu widzenia matematyki:

- nie jest problemem zebranie próbki
- nie jest problemem policzenie oceny estymowanego parametru

Pytanie

Jaka jest wiarygodność uzyskanego wyniku?

Jak określić wiarygodność uzyskanego wyniku

Dwa równoważne podejścia:

- testowanie hipotez
- estymacja przedziałów ufności

Co musimy wiedzieć by ustalić wiarygodność uzyskanego wyniku?

Najlepiej, gdy znamy rozkład badanej statystyki $T(X_1, ..., X_n)$.
Czy w ogóle możemy taki rozkład poznać?
W wielu przypadkach tak!
Przypomnijmy najbardziej znane – i najczęściej wykorzystywane – twierdzenia statystyczne.

Rozkład prawdopodobieństwa średniej z próby dla rozkładu $N(m, \sigma)$

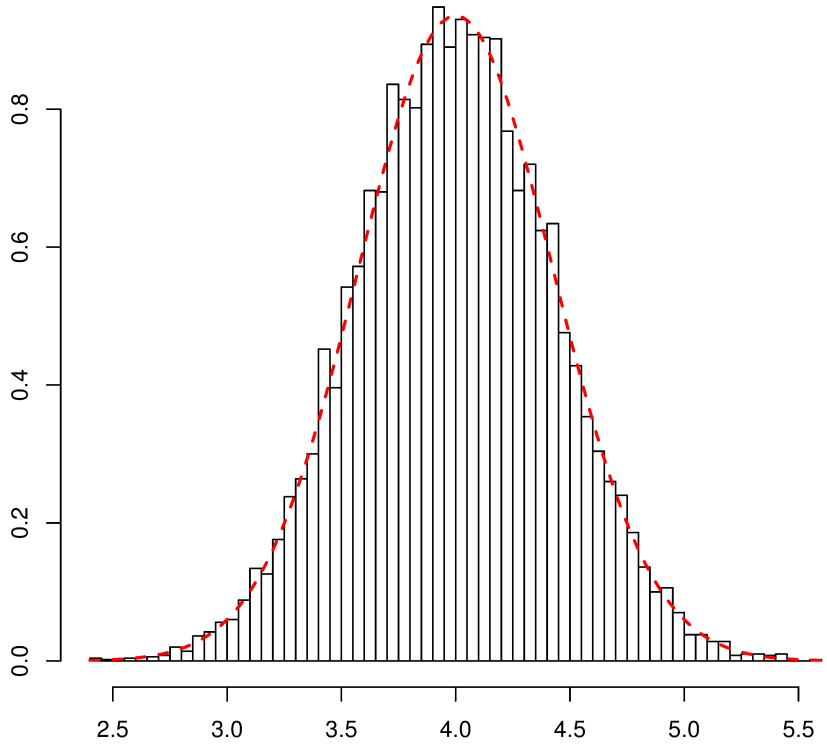
Twierdzenie

Jeżeli rozkład cechy w populacji jest rozkładem $N(m, \sigma)$, to rozkład statystyki

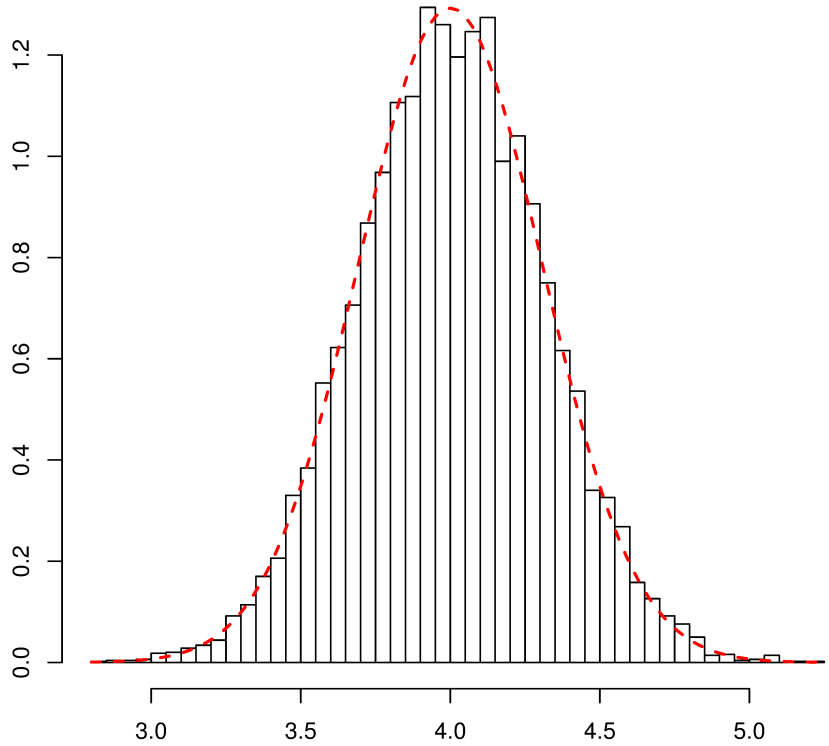
$$T(X_1, .., X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

jest rozkładem normalnym $N(m, \frac{\sigma}{\sqrt{n}})$.

Przykład rozkładu statystyki



Rysunek 1 : Empiryczny rozkład estymatora wartości średniej dla zmiennej losowej o rozkładzie $N(4, \sqrt{2})$. Próba o liczności $n = 11$ powtórzona $M = 9999$ razy



Rysunek 2 : Empiryczny rozkład estymatora wartości średniej dla zmiennej losowej o rozkładzie $N(4, \sqrt{2})$. Próba o liczności $n = 21$ powtórzona $M = 9999$ razy

Inne znane rozkłady statystyk.....

Dla $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ oraz $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ mamy

Twierdzenie

Jeżeli rozkład cechy w populacji jest rozkładem $N(m, \sigma)$, to statystyki

$$U^2 = \frac{nS^2}{\sigma^2}, \quad \hat{U}^2 = \frac{(n-1)\hat{S}^2}{\sigma^2}$$

mają rozkład χ^2 o $n-1$ stopniach swobody, natomiast statystyka

$$t = \frac{\bar{X} - \mu}{S}\sqrt{n-1}$$

ma rozkład t-Studenta o $n-1$ stopniach swobody.

Inne znane rozkłady statystyk.....

Twierdzenie

Jeżeli rozkład cechy w populacji jest rozkładem o ciągłej dystrybuancie F i gęstości f, to gęstość brzegowa dla każdej k-tej statystyki pozycyjnej $X_{(k)}$, $1 \leqslant k \leqslant n$ przyjmuje postać:

$$g(x) = \frac{n!}{(k-1)!(n-k)!} \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k} f(x).$$

Przykład do policzenia "na palcach" – rozkład zero-jedynkowy

W przypadku rozkładu zero-jedynkowego $P(X = 1) = p$, $P(X = 0) = 1 - p$, rozkład statystyki wartości średniej z próby (T) daje się policzyć bardzo prosto:

- dla $X_1, X_2, \ldots, X_n$ przestrzeń prób ma postać $\{0, 1\}^n$, zatem możliwe wartości statystyki T to $\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$,
- dla każdego $k \in \{0, 1, 2, \ldots, n\}$ równość $T = \frac{k}{n}$ zachodzi wtedy i tylko wtedy, gdy wśród wartości próby było k jedynek i $n - k$ zer, zatem

$$P\left(T = \frac{k}{n}\right) = \binom{n}{k} p^k (1 - p)^{n-k},$$

  tzn, statystyka nT ma rozkład Bernoulliego $B(n, p)$,
- z powyższego otrzymujemy natychmiast np. $E(T) = p$, $Var(T) = \frac{1}{n} p(1 - p)$.

Jak określić przedział ufności?

Jeżeli estymujemy parametr θ, to przedziałem ufności o współczynniku ufności $1-\alpha$ nazywamy taki przedział (θ_1, θ_2), że

$$P(\theta < \theta_1) = \frac{\alpha}{2}, \quad P(\theta > \theta_2) = \frac{\alpha}{2},$$

gdzie θ_1 i θ_2 są funkcjami wyznaczonymi na podstawie próby losowej.

Twierdzenie

Jeżeli znany jest rozkład nieobciążonego estymatora T parametru θ, to końce przedziału ufności możemy wyliczyć ze wzorów: $\theta_1 = t - t_{1-\frac{\alpha}{2}}$, $\theta_2 = t - t_{\frac{\alpha}{2}}$, *gdzie*

$$P(T - E(T) < t_{\frac{\alpha}{2}}) = \frac{\alpha}{2} = P(T - E(T) > t_{1-\frac{\alpha}{2}}).$$

Jak określić przedział ufności?

Na przykład

Twierdzenie

Jeżeli rozkład cechy w populacji jest rozkładem $N(m, \sigma)$, to przedział ufności dla parametru m jest następujący

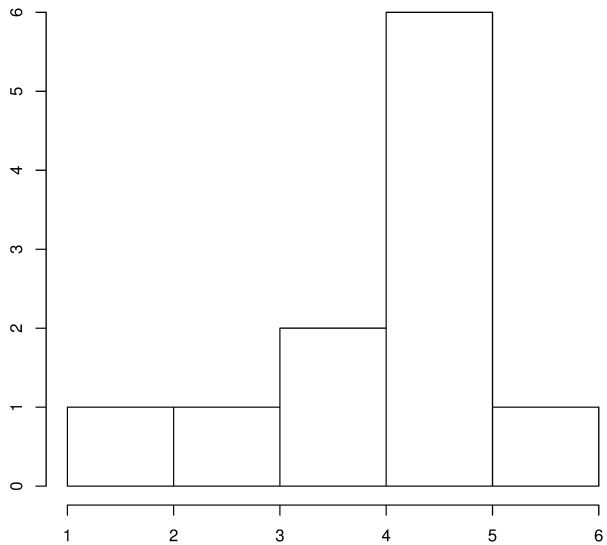
$$\left(\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right),$$

gdzie $z_{\frac{\alpha}{2}}$ jest wartością dytrybuanty standardowego rozkładu normalnego w punkcie $1 - \frac{\alpha}{2}$.

Przykład standardowego wnioskowania statystycznego

1	2	3	4	5	6	7	8	9	10	11
4.21	4.60	1.82	3.61	4.26	2.28	4.96	5.34	4.09	4.21	3.37

Tabela 1 : Próba losowa z rozkładu normalnego $N(4, \sqrt{2})$



Rysunek 3 : Wyniki losowania 11-elementowej próby z rozkładu normalnego $N(4, \sqrt{2})$

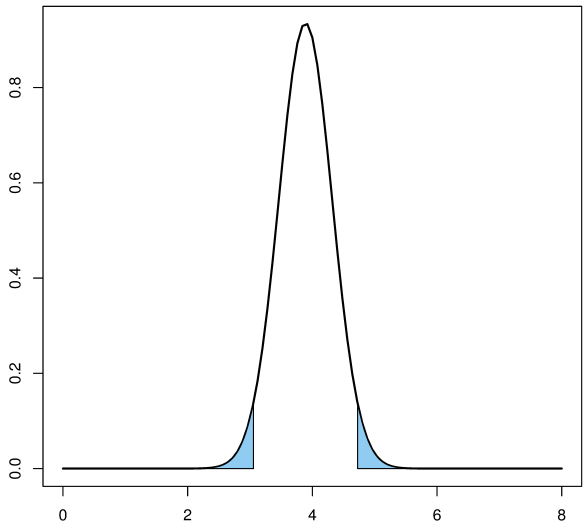
Przykład standardowego wnioskowania statystycznego – c.d.

Z wylosowanej próby możemy estymować wybrane parametry:

- wartość średnia z próby wynosi $3,89$;
- wariancja z próby wynosi 1.13;
- odchylenie standardowe z próby wynosi $1,06$;
- mediana z próby wynosi 4.21.

Przykład standardowego wnioskowania statystycznego – c.d.

Przedziały ufności dla wartości średniej przy założeniu, że znamy wariancję $\sigma^2 = 2$.
Zakładamy, że statystyka ma rozkład $N(3,89; \sqrt{2}/\sqrt{11})$. Przedział ufności na poziomie 95% dany jest przez odrzucenie "ogonów".



Rysunek 4 : Hipotetyczny rozkład statystyki T. Obszar niebieski to odrzucone "ogony", to co zostaje to przedział ufności

Przykład standardowego wnioskowania statystycznego – c.d.

Przedziały ufności dla wartości średniej:

- przy założeniu, że znamy wariancję $\sigma^2 = 2$;
- przy założeniu, że wariancję estymujemy z próby $s^2 = 1,13$;

.

z rozkładu normalnego	z rozkładu t-Studenta
$(3,05; 4,72)$	$(3,17; 4,60)$

Tabela 2 : 95% przedziały ufności wyliczone na podstawie wartości estymatorów dla próby z rozkładu normalnego

Rozkłady znane i nieznane

Pytanie

A co zrobić, gdy nie znamy rozkładu statystyki?

Pytanie

A co zrobić, gdy nie znamy w ogóle rozkładu zmiennej losowej opisującej cechę, którą badamy?

Metoda bootstrap

Punkt wyjścia: uzyskaliśmy próbę z badanej populacji. Pewnie niezbyt liczną – czy możemy coś więcej powiedzieć o rozkładzie badanej cechy niż to, co widzimy z histogramu uzyskanych danych? A gdyby ta nasza próba stała się dla nas populacją? Gdybyśmy zapytali się:

- Co stanie się z oceną parametru, gdy nasza próba będzie uboższa o jedno losowanie? Jaki wpływ ma każdy z uzyskanych wyników na wartość estymatora? Metoda "jackknife".
- Co stanie się, gdy z naszej n-elementowej próbki *wylosujemy* n-elementową próbę prostą i policzymy wartość estymatora na takiej "próbce z próbki"? Metoda "bootstrap".

Praca: Bradley Efron, "Bootstrap Methods: Another Look at the Jackknife", The Annals of Statistics, 1979, Vol. 7, No. 1, 1-26.

Opis metody bootstrap – wprowadzenie

Zakładamy, że mamy ciąg niezależnych rzeczywistych zmiennych losowych X_n o tym samy rozkładzie F. Rozkładu F nie znamy – chcemy jednak poznać pewne jego własności. W szczególności interesuje nas pewna charakterystyka liczbowa θ rozkładu F i wyliczająca ją statystyka t. Wtedy $\theta = t(F)$.
Dysponujemy jednak pewną informacją – zebraliśmy skończoną próbkę $x_1, x_2, ..., x_n$. Próbka ta wyznacza nam empiryczny rozkład prawdopodobieństwa $\hat{F}$, który możemy uznać za aproksymację rozkładu F. Również dla tego rozkładu $\hat{F}$ możemy obliczyć wartość interesującego nas parametru $t = t(\hat{F})$.

Opis metody bootstrap – rozwinięcie

Z rozkładem $\hat{F}$ robimy to samo, co przed chwilą zrobiliśmy z rozkładem F: budujemy ciąg niezależnych zmiennych losowych X_1^*, X_2^*, ..., X_n^* o rozkładzie $\hat{F}$. Czyli losujemy n-elementową próbkę, ale jedynie ze zbioru $\{x_1, x_2, ..., x_n\}$. Taka "próbka z próbki" nazywana jest próbką bootstrapową (*bootstrap sample*). Oznaczmy zmienną losową zwracającą próbę bootstrapową symbolem $\mathbf{X}^*$. Dla takiej próbki możemy policzyć wartość statystyki t i wprowadzić zmienną losową $T^* = t(\mathbf{X}^*)$.
Teraz możemy łatwo zrobić to, czego w zwykłej praktyce badawczej zwykle nie da się zrobić: powtarzamy proces losowania próbki bootstrapowej tak wiele razy jak chcemy. Otrzymujemy więc pewien ciąg próbek $\mathbf{X}_i^*$, $i = 1, 2, ..., R$ i odpowiadających jej wartości statystyki t, $t_i^* = t(\mathbf{X}_i^*)$.

Opis metody bootstrap – zakończenie

Założenie (Podstawowe założenie metody bootstrap)

Rozkład zmiennej losowej $T^* - t$ przypomina rozkład zmiennej losowej $T - \theta$.

Podsumowując – kluczowa dla wykorzystania metody bootstrap we wnioskowaniu statystycznym jest następująca analogia: próba bootstrapowa jest dla wylosowanej próbki tym, czym wylosowana próbka dla całej populacji.

Metoda bootstrap – podstawowe zastosowanie

Rozkład zmiennej T^* możemy wykorzystać do oszacowania podstawowych parametrów statystyki T

- obciążenia (bias)

$$b = \bar{t}^* - t,$$

gdzie $\bar{t}^*$ oznacza wartość średnią statystyki t wyliczaną ze wszystkich prób bootstrapowych (zakładamy, że mamy R prób bootstrapowych), tzn.

$$\bar{t}^* = \frac{1}{R}\sum_{i=1}^{R} t_i^*.$$

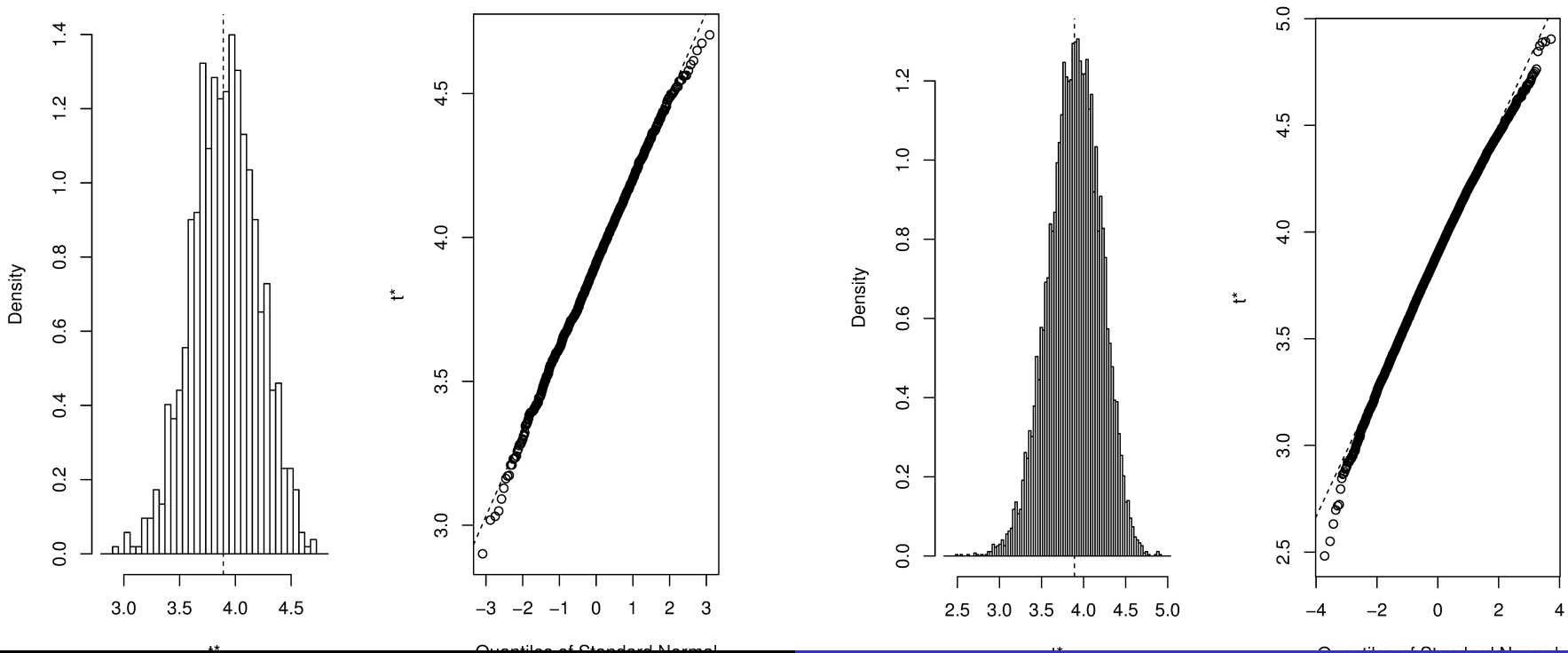
- wariancji

$$v^* = \frac{1}{R-1}\sum_{i=1}^{R}(t_i^* - \bar{t}^*)^2.$$

Metoda bootstrap – przykład zastosowania

Zbadajmy teraz metodą bootstrap estymator dla rozpatrywanej wcześniej próby z rozkładu normalnego $N(4, \sqrt{2})$

R (liczność próby)	t	obciążenie	odchylenie standardowe
1 000	3.8911398	0.0156	0.2918
10 000	3.8911398	-0.0002	0.3063



Metoda bootstrap dla rozkładu zero-jedynkowego – teoria

Widzieliśmy już, że w przypadku rozkładu zero-jedynkowego $P(X = 1) = p$, $P(X = 0) = 1 - p$, rozkład statystyki T wartości średniej z próby, która jest estymatorem parametru p, spełnia $nT \sim B(n, p)$, w szczególności $E(T) = p$ oraz $Var(T) = \frac{1}{n}p(1 - p)$.

Dla zaobserwowanych wartości próby $x_1, x_2, \ldots, x_n$ rozkład $\hat{F}$ jest także rozkładem zero-jedynkowym, ale o parametrze $\bar{x}$, zatem rozkład statystyki T^* spełnia $nT^* \sim B(n, \bar{x})$, dlatego $E(T^*) = \bar{x}$ i $Var(T^*) = \frac{1}{n}\bar{x}(1 - \bar{x})$.

Metoda bootstrapowa zastosowana w tym przypadku dałaby oszacowanie rozkładu $T - p$ poprzez rozkład $T^* - \bar{x}$ następująco

$$E(T - p) \approx E(T^* - \bar{x}) = 0$$

$$Var(T - p) \approx Var(T^* - \bar{x}) = Var(T^*) = \frac{1}{n}\bar{x}(1 - \bar{x}),$$

a jak wiemy, $\frac{1}{n}\bar{x}(1 - \bar{x}) \to \frac{1}{n}p(1 - p)$, gdy $n \to \infty$.

Metoda bootstrap dla rozkładu zero-jedynkowego – praktyka

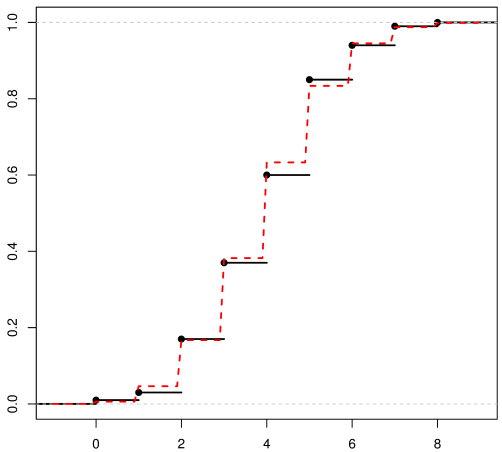
Zróbmy eksperyment z próbą o liczności $n = 10$ losowaną z rozkładu zero-jedynkowego o nieznanym parametrze $P(X = 1) = p$. Badana przez nas statystyka to wartość średnia

$$T(X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

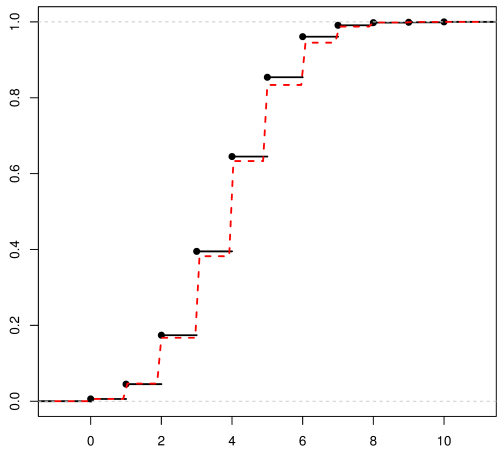
Wylosowana próba składa się z 4 jedynek i 6 zer. Daje nam to ocenę parametru p równą $t = 0,4$.

Metoda bootstrap dla rozkładu zero-jedynkowego – praktyka

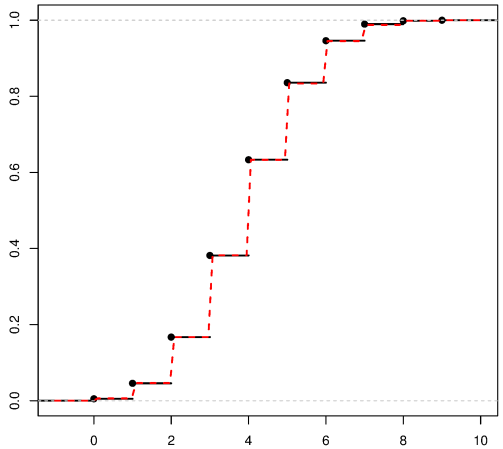
Popatrzmy na dystrybuantę rozkładu zmiennej T^* ($nT^* \sim B(n; 0,4)$) (czerwona linia przerywana) oraz empiryczną t^* wynikającą z prób bootstrapowych (czarne, ciągłe odcinki):



Rysunek 7 : $R = 100$



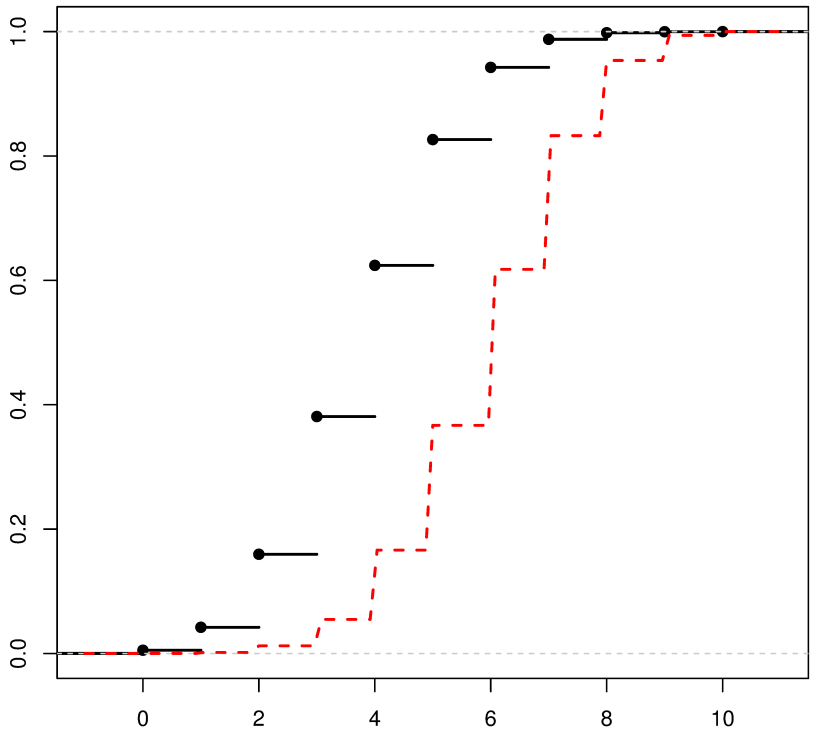
Rysunek 8 : $R = 1\ 000$



Rysunek 9 : $R = 10\ 000$

Metoda bootstrap dla rozkładu dwupunktowego – praktyka

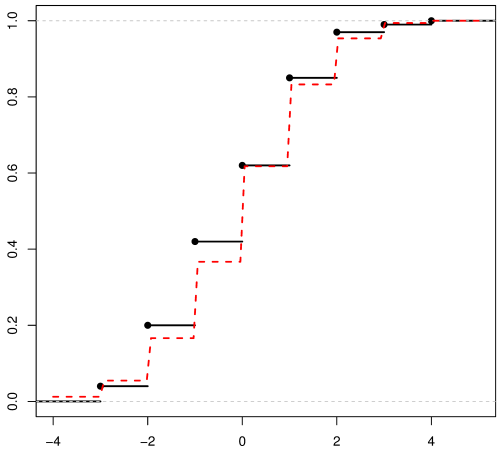
A co będzie, gdy rzeczywisty parametr $p = 0,6$? Czyli estymowana wartość jest wyraźnie inna niż w rzeczywistości. Nałóżmy na siebie dystrybuanty rozkładów T oraz T^*. Liczba prób wynosi 10 000.



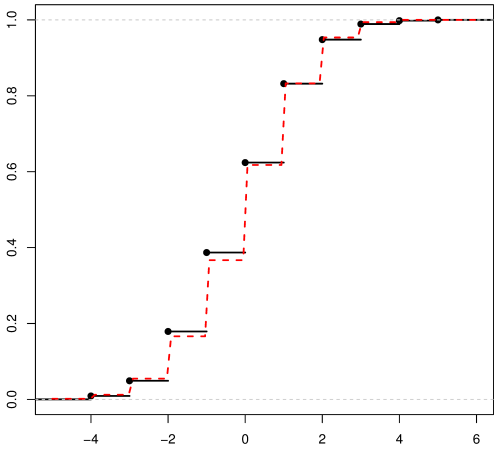
Rysunek 10 : $R = 10\ 000$

Metoda bootstrap dla rozkładu dwupunktowego – praktyka

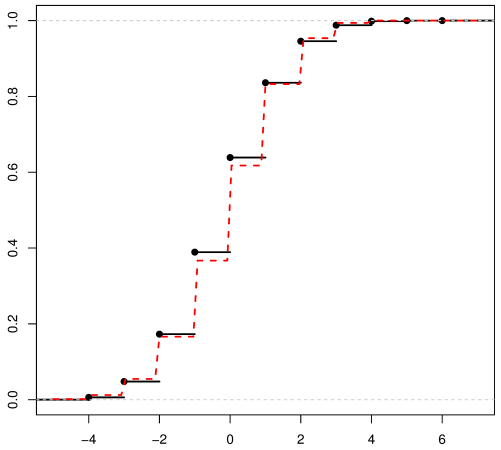
Popatrzmy na dystrybuantę teoretyczną zmiennej $T - p$ dla $p = 0,6$ (czerwona linia przerywana) oraz empiryczną $T^* - t$ wynikającą z prób bootstrapowych (czarne, ciągłe odcinki).



Rysunek 11 :
$R = 100$



Rysunek 12 :
$R = 1\ 000$



Rysunek 13 :
$R = 10\ 000$

Metoda bootstrap – przedziały ufności

Przedstawmy kilka metod pozwalających oszacować przedział, w którym z prawdopodobieństwem $1-\alpha$ znajduje się rzeczywista wartość estymowanego parametru. Każda z tych metod oparta jest na pewnym naturalnym pomyśle – wszystkie jednak dają inne wyniki i warto wiedzieć, jak te metody działają – choćby po to, by poznać ich dobre i złe strony.
W poniższych metodach zakładamy, że budowany jest równomierny przedział ufności na poziomie istotności $\alpha \in (0,1)$, tzn. taki, że prawdopodieństwo błędu zarówno prawo- jak i lewostronnego wynosi $\frac{\alpha}{2}$. Szukamy więc kwantyli $q_{\frac{\alpha}{2}}$ oraz $q_{1-\frac{\alpha}{2}}$, tj. wartości dla których

$$P(T-\theta \leqslant q_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \quad P(T-\theta \geqslant q_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

Normal interval

Ten sposób wyznaczania przedziału ufności oparty jest na założeniu, że estymator ma rozkład asymptotycznie normalny. Zakładamy, że zmienna losowa $T - \theta$ ma rozkład $N(0, \sqrt{v})$, co oznacza, że nasz przedział ufności dla poziomu α dany jest wzorem

$$(t - \sqrt{v} z_{1-\frac{\alpha}{2}}, t + \sqrt{v} z_{1-\frac{\alpha}{2}})$$

gdzie $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha)$, czyli kwantyl rzędu $1-\alpha$ standardowego rozkładu normalnego $N(0,1)$.
Za parametry rozkładu tego estymatora przyjmujemy bootstrapowe oszacowania dla t (czyli uwzględniamy obciążenie b!) oraz wariancji v.

Metoda kwantyli (percentile method)

Metoda polega na tym, że ustalamy kwantyle $\frac{\alpha}{2}$ oraz $1 - \frac{\alpha}{2}$ empirycznego rozkładu zmiennej t_n^* – kwantyle te wyznaczają poszukiwany przedział ufności.
Jak znaleźć kwantyle rzędu p? Jeżeli liczba $p(R+1)$ jest całkowita, to nie ma problemu – sortujemy wszystkie empirycznie uzyskane wartości t_k^*, gdzie $k = 1, 2, ..., R$ i wybieramy tę, która ma numer $p(R+1)$.

Metoda kwantyli (percentile method) – c.d.

Jeżeli jednak liczba $p(R+1)$ nie jest całkowitą, to odpowiednią wartość można interpolować jako liczbę gdzieś pomiędzy wartością k-tą a $k+1$-szą, gdzie $k = \lfloor p(R+1) \rfloor$. Zwykle nie stosuje się tu interpolacji liniowej, ale opartą na skali normalnej. Daje to wzór:

$$q^*_{p(R+1)} = q^*_k + \frac{\Phi^{-1}(p) - \Phi^{-1}(\frac{k}{R+1})}{\Phi^{-1}(\frac{k+1}{R+1}) - \Phi^{-1}(\frac{k}{R+1})}(q^*_{k+1} - q^*_k),$$

gdzie q^*_p oznacza kwantyl rzędu p ze statystyki bootstrapowej t^* (kwantyl empiryczny).
Gdy prób bootstrapowych jest mało i $p(R+1) < 1$ lub $p(R+1) > R$, to nie możemy wskazać odpowiedniej wartości. Rozsądnie jest wtedy przyjąć wartość kwantyla równą odpowiednio najmniejszej bądź największej uzyskanej wartości.

Podstawowa metoda bootstrapowa (basic bootstrap)

Popatrzmy na zależności:

$$P(T - \theta \leqslant q_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

$$P(T - \theta \geqslant q_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}.$$

Opisaliśmy prawdopodobieństwa, że wartość statystyki T leży w jednym z dwóch "ogonów" rozkładu prawdopodobieństwa statystyki T. Nas interesuje zdarzenie przeciwne, tzn. sytuacja kiedy wartość statystyki T leży w przedziale, który pokrywa $1 - \alpha$ "masy" rozkładu zmiennej losowej T. Stanie się tak, gdy Przepiszmy zdarzenia $T - \theta \geqslant q_{\frac{\alpha}{2}}$ oraz $T - \theta \leqslant q_{1-\frac{\alpha}{2}}$ jako $\theta \geqslant T - q_{1-\frac{\alpha}{2}}$ oraz $\theta \leqslant T - q_{\frac{\alpha}{2}}$.Wyznacza nam to przedział na θ dany jako

$$(t - q_{1-\frac{\alpha}{2}}, t - q_{\frac{\alpha}{2}}).$$

Podstawowa metoda bootstrapowa (basic bootstrap)

$$(t - q_{1-\frac{\alpha}{2}}, t - q_{\frac{\alpha}{2}}).$$

Tak zapisany przedział jest oczywiście mało praktyczny ponieważ nie znamy rozkładu zmiennej losowej $T - \theta$, nie znamy więc jego kwantyli. Chcemy jednak te kwantyle w jakiś sposób estymować. I tu przydaje się nasza analogia bootstrapowa: zakładamy, że dobrym przybliżeniem rozkładu $T - \theta$ jest rozkład $T^* - t$. I to kwantyle tego rozkładu chcemy znaleźć, i to te kwantyle $q^*_{\frac{\alpha}{2}}$, $q^*_{1-\frac{\alpha}{2}}$ wstawimy do wzoru zamiast $q_{\frac{\alpha}{2}}$ oraz $q_{1-\frac{\alpha}{2}}$.

Podstawowa metoda bootstrapowa (basic bootstrap)

Z ogólnych rozważań dotyczących estymacji statystyk pozycyjnych (kwantyli) dla ciągu niezależnych zmiennych losowych $X_1, ..., X_n$ o jednakowym rozkładzie (danym przez dystrybuantę K) wynika, że

$$EX_{(j)} = K^{-1}(\frac{j}{n+1}).$$

Oznacza to, że dobrym estymatorem kwantyla rzędu p rozkładu K (czyli wartości $K^{-1}(p)$) jest wartość zmiennej $X_{(p(n+1))}$, o ile $j = p(n+1)$ jest liczbą całkowitą. Zmienna losowa $X_{(p(n+1))}$ oznacza j-tą w kolejności od najmniejszej wylosowaną wartość zmiennej X_i w n próbach.
W naszej sytuacji nieznaną dystrybuantą K jest dystrybuanta rozkładu $T^* - t$, a jej kwantyl rzędu $\frac{\alpha}{2}$ możemy estymować wartością $t^*_{(\frac{\alpha}{2}(R+1))} - t$, co jest rozsądnym estymatorem kwantyla rzędu $\frac{\alpha}{2}$ zmiennej losowej $T^* - t$. Możemy więc powiedzieć, że

$$(t + t - t^*_{((1-\frac{\alpha}{2})(R+1))}, t + t - t^*_{(\frac{\alpha}{2}(R+1))}).$$

Metoda studentyzacji zmiennej losowej

W metodzie tej wychodzimy z założenia, że o ile nie znamy rozkładu statystyki $T - \theta$, to możemy przyjąć, że rozkład statystyki

$$Z = \frac{T - \theta}{\sqrt{V}}$$

może być bliski $N(0, 1)$. Wariancja V jest tu wariancją zmiennej losowej $T - \theta$ (czyli też zmiennej losowej T). To wariancja **estymatora** T, a nie wyjściowej zmiennej losowej!

Metoda studentyzacji zmiennej losowej

Metodą bootstrapową badamy teraz rozkład statystyki Z, czyli dla danej próby bootstrapowej liczymy

$$z^* = \frac{t^* - t}{\sqrt{v^*}},$$

gdzie v^* jest pewnym estymatorem wariancji liczonym dla próby bootstrapowej. Następnie do oszacowania przedziału ufności dla statystyki t dla poziomu istotności α liczymy kwantyle rzędu $\frac{\alpha}{2}$ oraz $1 - \frac{\alpha}{2}$ dla rozkładu z^*. Oznaczmy te kwantyle symbolem $z^*_{\frac{\alpha}{2}}$ oraz $z^*_{1-\frac{\alpha}{2}}$. Wtedy przedział ufności dany jest wzorem:

$$(t - \sqrt{v^*} z^*_{1-\frac{\alpha}{2}}, t - \sqrt{v^*} z^*_{\frac{\alpha}{2}}).$$

Kwantyle $z^*_{\frac{\alpha}{2}}$ oraz $z^*_{1-\frac{\alpha}{2}}$ oszacowane są z empirycznego rozkładu z^* podobnie jak w metodzie kwantyli.

Metoda studentyzacji zmiennej losowej

W metodzie tej kluczowe wydaje się estymowanie wariancji: i to zarówno z próbek bootstrapowych y_n^* jak i dla wyjściowej próby y. Oczywiście można tu wziąć wariancję estymowaną samą metodą bootstrapową, ale okazuje się, że lepsze własności daje estymacja wariancji oparta na bardziej zaawansowanych metodach.

Metoda studentyzacji zmiennej losowej

Jeśli chodzi o estymację wariancji T to przede wszystkim warto tu wymienić "non-parametric delta method", która dana jest wzorem:

$$v_L = \frac{1}{n^2} \sum_{i=1}^{n} l_i^2,$$

gdzie l_i jest tzw. "influence value". dane jako $l_i = l(y_i)$, gdzie

$$l(y) = L_t(y, \hat{F}) = \frac{\partial t[(1-\varepsilon)\hat{F} + \varepsilon H_y]}{\partial \varepsilon}|_{\varepsilon=0}$$

gdzie $\hat{F}$ jest empiryczną dystrybuantą zmiennej losowej (czyli dystrybuantą wynikająca z próbki), zaś H_y jest dystrybuantą rozkładu jednopunktowego skupionego w y.
W praktyce "influence value" może być wyliczane jako "empirical influence value", które może być wyliczane jako numeryczne przybliżenie pochodnej zdefiniowanej powyżej.

Przedziały ufności: inne metody

- metoda BC_a (bias corrected and accelerated)
- metoda ABC (approximate bootstrap condence intervals)

Przedziały ufności: przykłady

Estymacje przedziałów ufności dla poziomu 95% dla prób bootstrapowych różnej wielkości

- Dla 1 000 prób

Normal	Basic	Studentized	Percentile	BCa
(3.303, 4.448)	(3.304, 4.468)	(2.833, 4.438)	(3.315, 4.478)	(3.231, 4.408)

- Dla 10 000 prób

Normal	Basic	Studentized	Percentile	BCa
(3.291, 4.492)	(3.330, 4.512)	(2.870, 4.490)	(3.270, 4.452)	(3.186, 4.407)

Co jeszcze możemy zobaczyć

Dobrym nawykiem powinien być rzut oka na histogram prób bootstrapowych. Możemy potraktować to jako wskazówkę czy bootstrap generuje sensowne wyniki. W końcu oczekujemy na to, że rozkład prób boostrapowych przypominać będzie rozkład wartości statystyki w populacji. W szczególności pewnymi wskazówkami co do rozkładu statystyki może być np. asymetria rozkładu bootstrapowego. Z kolei rzut oka na wykres QQ porównujący rozkład T^* do rozkładu normalnego pozawala nam stwierdzić, że nie możemy dla statystyki T zakładać rozkładu normalnego.

A co jeśli rozkład T^* nie jest normalny

Można zmodyfikować statystykę przy pomocy pewnej rosnącej funkcji h tak, aby estymator $h(t(\cdot))$ miał już odpowiednie własności – np. by zmienna $h(T^*)$ miała już rozkład zbliżony do normalnego. Wtedy możemy przedział ufności zmiennej T^* dla poziomu α szacować przy pomocy wartości

$$h^{-1}(h(t) \pm z_\alpha \sqrt{v^*}),$$

gdzie z_α jest odpowiednim kwantylem rozkładu normalnego, zaś v^* jest estymowaną wariancją zmiennej $h(T^*)$.

Twierdzenie

Niech $\hat{\theta}_\alpha$ oznacza estymator kwantyla rzędu α uzyskany metodą studentized boostrap lub metodą BC_a dla próby liczności n. Wówczas:

$$P(\theta \leqslant \hat{\theta}_\alpha) = \alpha + O(\frac{1}{n}).$$

Twierdzenie

Niech $\hat{\theta}_\alpha$ oznacza estymator kwantyla rzędu α uzyskany jedną z metod normal interwal, kwantyli lub basic dla próby liczności n. Wówczas:

$$P(\theta \leqslant \hat{\theta}_\alpha) = \alpha + O(\frac{1}{\sqrt{n}}).$$

Regresja liniowa

Rozważmy model regresji postaci

$$y_i = f(x_i, \beta) + \varepsilon_i \quad i = 1, 2, \ldots, n,$$

gdzie y_i jest wartością zmiennej objaśnianej, x_i realizacją wektora zmiennych objaśniających, ε_i - wartością składnika losowego, natomiast β jest nieznanym wektorem parametrów, który chcemy estymować.
Zakładamy, że

$$\varepsilon_i \sim F \quad i = 1, 2, \ldots, n.$$

Przypuśćmy, że β estymujemy metodą najmniejszych kwadratów otrzymując jego ocenę $\hat{\beta}$.
Oszacowania wartości składników losowych są wówczas dane poprzez

$$\hat{\varepsilon}_i = y_i - f(x_i, \hat{\beta}) \quad i = 1, 2, \ldots, n.$$

Teraz definiujemy rozkład $\hat{F}$ poprzez

$$P(\hat{F} = \hat{\varepsilon}_i) = \frac{1}{n} \quad i = 1, 2, \ldots, n.$$

Następnie według tego rozkładu pobieramy próbkę bootstrapową $(\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*)$, dla której wyliczamy

$$y_i^* = f(x_i, \hat{\beta}) + \varepsilon_i^* \quad i = 1, 2, \ldots, n.$$

Otrzymane wartości y_i^* wykorzystujemy do ponownej estymacji metodą najmniejszych kwadratów, tym razem dla modelu

$$y_i^* = f(x_i, \beta) + \tilde{\varepsilon}_i \quad i = 1, 2, \ldots, n.$$

Jak zwykle w metodzie bootstrapowej, całą operację powtarzamy R razy i w wyniku otrzymujemy wartości $\hat{\beta}_1^*$, $\hat{\beta}_2^*$, ..., $\hat{\beta}_R^*$, które służą nam do oszacowania rozkładu $\hat{\beta}$.

Modyfikacje podstawowej metody

- symetryzacja wyników
- przydzielanie uzyskanym wynikom różnych wag
- balanced boostrap
- antithetic bootstrap

Modyfikacje podstawowej metody: różne wagi dla uzyskanych wyników

W stosowanych metodach nie musimy zakładać, że rozkład empiryczny $\hat{F}$ z próby $x_1, ..., x_n$ jest równomierny, tzn. że

$$P(\hat{F} = x_i) = \frac{1}{n}.$$

Różnym realizacjom zmiennej losowej można nadawać różne wagi – tym samym modyfikując rozkład $\hat{F}$. Nie zmienia to idei metody bootstrapowej – modyfikuje jedynie sposób losowania prób bootstrapowych.
Kiedy taka modyfikacja wag może być konieczna? Np. gdy zebrana próba jest niejednorodna pod względem wiarygodności (pewne dane są mniej wiarygodne niż inne), albo gdy posiadamy jakiekolwiek informacje wskazujące na charakter rozkładu.

Modyfikacje podstawowej metody: symetryzacja wyników

Jeżeli wiemy, że zmienna losowa ma rozkład symetryczny, to uzyskane wyniki można w sztuczny sposób "zsymetryzowac", tzn. dla każdej uzyskanej wartości zmiennej losowej dołożyć wartość symetryczną do niej względem np. mediany (która wydaje się tu lepszym estymatorem dla wartości względem której symetryczny jest rozkład niż np. wartość średnia).

Modyfikacje podstawowej metody: balanced bootstrap

Można zagwarantować sobie, że we wszystkich wygenerowanych R próbach bootstrapowych każda z wartości próbki $x_1, x_2, ..., x_n$ wystąpi tak samo często.
Jak można sobie to zagwarantować? Np. budując tablicę składającą się z R powtórzeń ciągu $(1, 2, 3..., n)$, a następnie losowo permutując tą tablicę. Tablica po permutacji znów dzielona jest na fragmenty długości n – i każdy taki fragment mówi nam, które elementy próby wyjściowej wybrać do próby bootstrapowej.

Modyfikacje podstawowej metody: antithetic boostrap

Jeżeli uporządkujemy uzyskane wartości próbki

$$x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \ldots \leqslant x_{(n)},$$

to dla każdej próby bootstrapowej możemy dołożyć "anty-próbkę" polegającą na tym, że jeśli w wyjściowej próbie pojawił się element pierwszy w kolejności, to w anty-próbie znajdzie się element pierwszy od końca. Jeżeli w wyjściowej próbie był element piaty, to w anty-próbie powinien znaleźć się element piąty od końca, itp. Daje to bardziej zbilansowany rozkład prób bootstrapowych – i lepsze własności w estymowaniu przedziałów ufności np. dla mediany.

Modyfikacje podstawowej metody: double boostrap

Każdą z uzyskanych prób boostrapowych można znów potraktować procedurą boostrapową – np. po to by estymować wariancję rozkładu T^* w metodzie studentyzacji zmiennej.
Z kolei wykonywanie prób boostrapowych na próbach bootstrapowych pozwala oszacować obciążenie estymatora statystyki $t(X_i^*)$, co pozwala estymować obciążenie obciążenia statystyki t, tym samym dając precyzyjniejszy (asymptotycznie) estymator obciążania.

Zastrzeżenia do metody bootstrap

Błędy w metodzie bootstrap:
(1) błąd próbkowania (próbka jako reprezentacja populacji)
(2) błąd próbkowania bootstrapowego (wynikający z tego, że nie wszystkie próbki bootstrapowe są wygenerowane) – ten błąd możemy kontrolować powtarzając próbkowanie wystarczającą liczbę razy
Podstawowe zastrzeżenie – metoda nie spełnia:

Reguła nr 1 (L. J. Gleser, The First Law of Applied Statistics)

Two individuals using the same statistical method on the same data should arrive at the same conclusion

Jak to w praktyce robić?

Symulacje wykonywane były w pakiecie R przy pomocy biblioteki `boot` oraz funkcji `boot(·)`, która wykonuje procedurę próbkowania z podanego wektora danych odpowiednią liczbę danych. Poniżej przedstawimy przykład wykorzystania tej procedury.
W procedurze tej możemy wykorzystać w zasadzie dowolną statystykę liczoną na próbie `x` dowolnej liczności. Funkcja zwraca pewien obiekt, który przechowuje wygenerowane dane. Poniżej przykład zastosowania:

```
samplemedian <- function(x, d) median(x[d])
b=boot(x,samplemedian,R=1000)
print(b)
plot(b)
boot.ci(b,conf=0.95)
```

Literatura

Efron B., *Bootstrap Methods: Another Look at the Jackknife*, The Annals of Statistics 7, 1979
Efron B., Tibshirani R.J., *An Introduction to the Bootstrap*, Chapman & Hall, 1993
DiCiccio T.J., Efron B., *Bootstrap Confidence Intervals*, Statistical Science 1996, Vol. 11, No. 3, 189-228
Davison A.C., Hinkley D.V., *Bootstrap Methods and their Application*, Cambridge University Presss, 1997
Domański Cz., Pruska K., *Nieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, 2000

I tu już koniec

Dziękujemy za uwagę!