Metody nieparametryczne w statystyce

Tomasz Wójtowicz

Wydział Zarządzania
AGH Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Niech $X_1, \ldots, X_n$ będzie próbą losową prostą z pewnego rozkładu F.
Niech $\hat{\theta} = T(X_1, .., X_n)$ będzie pewną statystyką.

Stosując metodę **Monte Carlo** można oszacować rozkład statystyki $\hat{\theta}$ poprzez wygenerowanie odpowiednio dużej próby wartości $\hat{\theta}_1, \ldots, \hat{\theta}_k$.

Warunek:
znajomość rozkładu F próby losowej $X_1, .., X_n$.

Problem:

- często nie znamy tego rozkładu,
- dysponujemy jedynie wartościami $x_1, \ldots, x_n$.

Metoda bootstrap

Metoda bootstrap

Niech:

- $X_1, \dots, X_n$ będzie próbą losową prostą z pewnego (nieznanego) rozkładu F,
- $x_1, \dots, x_n$ będą wartościami próby losowej $X_1, \dots, X_n$,
- $\hat{F}$ będzie dystrybuantą empiryczną zdefiniowaną na podstawie próby $x_1, \dots, x_n$,

wtedy

dystrybuanta $\hat{F}$ **jest przybliżeniem** rozkładu F.

Możemy więc zastosować $\hat{F}$ zamiast F do estymacji rozkładu statystyki $\hat{\theta}$,

tzn. ocenić rozkład $\hat{\theta}$ na podstawie wielu prób wygenerowanych z rozkładu $\hat{F}$ zamiast z rozkładu F.

Metoda bootstrap

Definicja

Próbę losową prostą $\boldsymbol{X}^* = (X_1^*, \dots, X_n^*)$ z rozkładu $\hat{F}$ dla ustalonej realizacji $x_1, \dots, x_n$ nazywamy **próbą typu bootstrap** (**próbą bootstrap**).

pull oneself up by one's bootstraps

– wybobyć się z opresji używając własnych sił

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

Metoda bootstrap

W celu otrzymania realizacji próby bootstrap dokonuje się n-krotnego losowania **ze zwracaniem** spośród elementów oryginalnej próby $x_1, \dots, x_n$.

Uwagi:

- losowość związana jest tylko z wyborem elementu spośród $x_1, \dots, x_n$,
- próbkę $x_1, \dots, x_n$ traktujemy jako populację, z której czerpiemy próby losowe,
- próba bootsrtap składa się z elementów próby $x_1, \dots, x_n$, przy czym niektóre wartości mogą się powtarzać (a niektóre mogą nie występować),
- w próbie bootstrap elementy z reguły się powtarzają,
- prawdopodobieństwo, że każdy element $x_1, \dots, x_n$ wystąpi dokładnie jeden raz wynosi $n!/n^n$.

Metoda bootstrap

Uzasadnienie:

szacowany parametr θ jest pewną funkcją dystrybuanty F, a jego estymator $\hat{\theta}$ otrzymujemy podstawiając do tej funkcji $\hat{F}$ zamiast F.

Przykład:

Spróbujmy oszacować wartość dystrybuanty F w pewnym punkcie t_0.

W tym przypadku $\theta = F(t_0)$, dla pewnego ustalonego t_0.

Estymatorem $F(t_0)$ jest oczywiście wartość dystrybuanty empirycznej $\hat{F}(t_0)$, tzn. $\hat{\theta} = \hat{F}(t_0)$.

Metoda bootstrap

Zasada:

Rozkład statystyki $T(\boldsymbol{X}^*) - \hat{\theta}$ dla próby bootstrap przy ustalonych wartościach realizacji $x_1, \dots, x_n$ jest dla regularnych statystyk T bliski rozkładowi $T(\boldsymbol{X}) - \theta$.

To oznacza, że:

- dla ustalonych $x_1, \dots, x_n$ kształt $T(\boldsymbol{X}^*)$ jest bliski kształtowi $T(\boldsymbol{X})$,
- położenie rozkładu statystyki $T(\boldsymbol{X}^*)$ jest przesunięte względem położenia rozkładu statystyki $T(\boldsymbol{X})$ o wielkość $\hat{\theta} - \theta$.

Metoda bootstrap

Kolejne kroki:

- na podstawie realizacji $x_1, \dots, x_n$ obliczyć wartość $\hat{\theta}$,
- na podstawie realizacji $x_1, \dots, x_n$ wylosować niezależne próby bootstrap $\boldsymbol{X_1^*}, \dots, \boldsymbol{X_k^*}$,
- obliczyć wartości $\hat{\theta}_1^* = T(\boldsymbol{X_1^*}) - \hat{\theta}$, $\hat{\theta}_2^* = T(\boldsymbol{X_2^*}) - \hat{\theta}$,..., $\hat{\theta}_k^* = T(\boldsymbol{X_k^*}) - \hat{\theta}$,
- skonstruować histogram wartości $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$, który jest przybliżeniem rozkładu statystyki $\hat{\theta} - \theta$.

Uzyskany rozkład jest przybliżeniem rozkładu błędów estymacji parametru θ.

Nazywamy go estymatorem rozkładu $\hat{\theta}$ uzyskanym metodą bootstrap.

Metoda bootstrap

Mając rozkład $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$ możemy oszacować zmienność estymatora $\hat{\theta} = T(X)$, co nie jest możliwe przy wykorzystaniu tylko wartości $x_1, \ldots, x_n$.

Definicja

Błędem standardowym typu bootstrap estymatora $\hat{\theta}$ nazywamy:

$$S_{\hat{\theta}} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2},$$

gdzie

$$\bar{\theta}^* = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i^*$$

Metoda bootstrap

Uwaga

Metoda bootstrap nie zawsze daje dobre wyniki

Przykład:

Niech $x_1, . . , x_n$ będzie realizacją próby losowej $X_1, \ldots, X_n$ z rozkładu jednostajnego na przedziale $[0, \theta]$, gdzie θ jest nieznanym parametrem.

Naturalnym estymatorem θ jest $\hat{\theta} = X_{(n)}$ czyli największa wartość w próbie losowej.

Przedziały ufności

Przedział ufności oparty na przybliżeniu normalnym

Jeżeli możemy przyjąć, że rozkład $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$ jest w przybliżeniu normalny, to wtedy również rozkład $\hat{\theta}$ jest w przybliżeniu normalny. Wtedy, dla współczynnika ufności $1 - \alpha$, przedział ufności dla θ ma postać:

$$\left(\hat{\theta} - u_\alpha \sigma_{\hat{\theta}}, \hat{\theta} + u_\alpha \sigma_{\hat{\theta}}\right)$$

Zastępując nieznane odchylenie standardowe estymatora $\sigma_{\hat{\theta}}$ bliskim mu błędem standardowym $S_{\hat{\theta}}$ uzyskujemy przedział ufności

$$\left(\hat{\theta} - u_\alpha S_{\hat{\theta}}, \hat{\theta} + u_\alpha S_{\hat{\theta}}\right)$$

dla przybliżonego współczynnika ufności $1 - \alpha$.

Przedziały ufności

Percentylowy przedział ufności typu bootstrap

Jeżeli przez q_α^* oznaczymy kwantyl rzędu α z rozkładu $\hat{\theta}^* - \hat{\theta}$, to wtedy

$$P_{\hat{F}}\left(q_{\alpha/2}^* \leq \hat{\theta}^* - \hat{\theta} \leq q_{1-\alpha/2}^*\right) = 1 - \alpha$$

Na podstawie zasady bootstrap mamy więc:

$$P_F\left(q_{\alpha/2}^* \leq \hat{\theta} - \theta \leq q_{1-\alpha/2}^*\right) = 1 - \alpha$$

czyli:

$$P_F\left(\hat{\theta} - q_{1-\alpha/2}^* \leq \theta \leq \hat{\theta} - q_{\alpha/2}^*\right) = 1 - \alpha$$

czyli przedział ufności postaci:

$$\left(\hat{\theta} - q_{1-\alpha/2}^*, \hat{\theta} - q_{\alpha/2}^*\right)$$

Metoda bootstrap

Percentylowy przedział ufności typu bootstrap

basic bootstrap

$$\left(\hat{\theta} - q^*_{1-\alpha/2}, \hat{\theta} - q^*_{\alpha/2}\right)$$

gdzie q^*_α oznacza kwantyl rzędu α z rozkładu $\hat{\theta}^* - \hat{\theta}$,

percentile

$$\left(q^*_{\alpha/2}, \quad q^*_{1-\alpha/2}\right)$$

gdzie q^*_α oznacza kwantyl rzędu α z rozkładu $\hat{\theta}^*$,

Metoda bootstrap

Studentyzacja

Powyższy schemat można też zastosować do statystyk studentyzowanych, tzn. statystyk postaci:

$$\frac{\hat{\theta} - \theta}{S_{\hat{\theta}}}$$

gdzie $S_{\hat{\theta}}$ jest bootstrapowym odchyleniem stand. $\hat{\theta}$.

Takie statystyki maja w przybliżeniu rozkład $N(0,1)$.

Z zasady bootstrap wynika, że rozkłady zmiennych:

$$\frac{\hat{\theta} - \theta}{S_{\hat{\theta}}} \quad \text{i} \quad \frac{\hat{\theta}^* - \hat{\theta}}{S_{\hat{\theta}^*}}$$

są bliskie.

Metoda bootstrap

W celu wyznaczenia przedziału ufności obliczamy dla $i = 1, \dots k$:

$$t_i^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{S_{\hat{\theta}_i^*}}$$

Uwaga!

Obliczenie $S_{\hat{\theta}_i^*}$ wymaga zastosowania metody bootstrap do próby bootstrapowej.

Próba ta może mieć dużo mniejszą liczność.

Wyznaczamy kwantyle $q_{\alpha/2}^*$ i $q_{1-\alpha/2}^*$ rozkładu $t_1^*, \dots, t_k^*$.

Wtedy, przedział ufności ma postać:

$$\left(\hat{\theta} - q_{1-\alpha/2}^* S_{\hat{\theta}}, \hat{\theta} - q_{\alpha/2}^* S_{\hat{\theta}}\right)$$

W tym przypadku kwantyle rozkładu normalnego zostały zastąpione przez kwantyle empiryczne.

Metoda bootstrap

Przykład:

Na podstawie poniższych wyników egzaminu:

38.9, 55.6, 53.7, 48.1, 56.5, 44.4, 54.6, 53.7, 43.5, 56.5, 50.0, 1.9 ,31.5, 27.8

wyznaczyć przedział ufności dla wartości oczekiwanej. Przyjąć $1 - \alpha = 0{,}95$.

Rozwiązanie:

$$\bar{X} = 44{,}5$$

Przedziały ufności

normalny: $(36{,}4; 51{,}7)$

bootstrapowy: $(37{,}3; 51{,}5)$

percentylowy: $(36{,}6; 50{,}8)$

studentyzowany: $(38{,}2; 58{,}7$