Metody nieparametryczne w statystyce

Tomasz Wójtowicz

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

Metoda Monte Carlo

Załóżmy, że:

- interesuje nas pewien parametr θ rozkładu zmiennej losowej X i że θ można przedstawić w postaci:

$$\theta = E\big(h(X)\big),$$

  gdzie h jest funkcją, której postać znamy.
- jesteśmy w stanie wygenerować próbę pseudolosową z rozkładu X.

Na podstawie wyników tej próby oszacujemy wartość parametru θ.

Metoda Monte Carlo

Definicja

Niech dla pewnej wartości n $X_1, \dots, X_n$ będzie próbą pseudolosową z rozkładu zmiennej losowej X. Średnią

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

nazywamy estymatorem parametru $\theta = E\big(h(X)\big)$ otrzymanym metodą Monte Carlo.

Metropolis N., Ulam S., (1949) The Monte Carlo Method, „Journal of the American Statistical Association”, Vol. 44, No. 247.

Prawa wielkich liczb

Podstawą teoretyczną Metody Monte Carlo są prawa wielkich liczb.

Prawo wielkich liczb Bernoulliego

Niech S_n będzie liczbą sukcesów w schemacie Bernoulliego z parametrami n i p.

Wtedy

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = 1$$

Nierówność Czebyszewa:

Niech X_k będzie ciągiem zmiennych losowych o skończonej wariancji σ^2, to

$$P(|\bar{X}_n - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

Prawa wielkich liczb

Wnioski:

Niech X_k będzie ciągiem zmiennych losowych o skończonych wartościach oczekiwanych i skończonych wariancjach. Niech X_k będą parami nieskorelowane.

$$\lim_{n\to\infty} P(|\bar{X}_n - E\bar{X}_n| < \varepsilon) = 1$$

Jeżeli $EX_k = \mu$ i $D^2X_k = \sigma^2$, to

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$$

Mocne prawo wielkich liczb:

$$P\left[\lim_{n\to\infty}(\bar{X}_n - EX_n) = 0\right] = 1$$

Prawa wielkich liczb

Wniosek:

Jeżeli X_n jest ciągiem niezależnych zmiennych losowych o rozkładzie jednostajnym na przedziale $[a, b]$,a h pewną funkcją określoną na $[a, b]$, to w szczególności

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \to \frac{1}{b-a}\int_a^b h(x)dx$$

Metoda Monte Carlo

Definicja

Niech dla pewnej wartości n $X_1, \dots, X_n$ będzie próbą pseudolosową z rozkładu zmiennej losowej X. Średnią

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

nazywamy estymatorem parametru $\theta = E\big(h(X)\big)$ otrzymanym metodą Monte Carlo.

Metoda Monte Carlo

Przykłady:

- gdy θ jest odchyleniem standardowym X, to np. $\theta = \sqrt{EX^2 - \mu^2}$, czyli
$$h(x) = \sqrt{x^2 - \mu^2}$$
- gdy θ jest wartością dystrybuanty zmiennej X dla pewnego t_0, tzn.
$$\theta = F(t_0) = P(X \leq t_0),$$
wtedy możemy zapisać, że $\theta = E(h(X))$, gdzie
$$h(x) = \begin{cases} 1 & x \leq t_0 \\ 0 & x > t_0 \end{cases}$$
- niech X będzie zmienną dyskretną o wartościach $x_1, x_2, \ldots$ z prawdopodobieństwem $p_k = P(X = x_k)$ i $\theta = p_{k_0}$ dla pewnego k_0. Wtedy
$$h(x) = \begin{cases} 1 & x \leq x_{k_0} \\ 0 & x \neq x_{k_0} \end{cases}$$

Metoda Monte Carlo

Parametr θ nie musi mieć interpretacji probabilistycznej.

Przykład

Może nas interesować oszacowanie całki

$$\theta = \int_0^1 h(t)dt$$

W tym celu wystarczy zauważyć, że jest to wartość oczekiwana zmiennej losowej $h(X)$, gdzie X jest zmienną losową o rozkładzie jednostajnym na $[0,1]$.

W analogiczny sposób możemy szacować wartości dowolnych całek oznaczonych i pól pod wykresami funkcji.

Błąd standardowy

Błąd estymatora $\bar{h}$, tzn. różnica $\bar{h} - \theta$ jest losowy. Możemy jednak ocenić jego zmienność:

$$D^2\left(\bar{h} - \theta\right) = D^2\left(\bar{h}\right) = D^2\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right) = \frac{1}{n}D^2(X)$$

czyli **błąd standardowy estymatora** $\bar{h}$ ma postać:

$$S_{\bar{h}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(h(X_i) - \bar{h}\right)^2}$$

Wniosek:

Na tej podstawie możemy dobrać liczność próby pseudolosowej tak, by błąd standardowy był z dużym prawdopodobieństwem mały.

Przedział ufności

Dla dużych n

$$\frac{\bar{h} - \theta}{S_{\bar{h}}}$$

ma w przybliżeniu standardowy rozkład normalny $N(0,1)$.

To pozwala na konstrukcję przybliżonego przedziału ufności, bo

$$P\left(\bar{h} - u_\alpha S_{\bar{h}} < \theta < \bar{h} + u_\alpha S_{\bar{h}}\right) \approx 1 - \alpha$$

czyli przedział ufności dla θ, dla współczynnika ufności $1 - \alpha$ jest w przybliżeniu równy:

$$\left(\bar{h} - u_\alpha S_{\bar{h}}, \bar{h} + u_\alpha S_{\bar{h}}\right).$$

W tym przypadku można też tak dobrać liczbę danych by przedział ufności z dużym prawdopodobieństwem nie przekraczał zadanej szerokości.

Przedział ufności

Niekiedy możemy z góry oszacować szerokość przedziału ufności niezależnie od wylosowanej próby.

Przykład

Gdy $\theta = P(X \leq t_0)$. Wtedy

$$\bar{h} = \hat{F}(t_0) = \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \frac{liczba\ obserwacji \leq t_0}{n}$$

Wtedy $nF(t_0)$ ma rozkład dwumianowy z parametrami n i $p = F(t_0)$, czyli

$$D^2\left(\bar{h}\right) = \frac{F(t_0)\left(1 - F(t_0)\right)}{n} \leq \frac{1}{4n}$$

bo $p(1-p) \leq 1/4$.

Przedział ufności

Za pomocą metody Monte Carlo możemy też wyznaczyć przedział ufności dla dowolnej zmiennej losowej W.

Dla zadanego współczynnika ufności $1-\alpha$ szukamy takiego przedziału $\left[q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}\right]$, że

$$P\left(q_{\frac{\alpha}{2}} \leq W \leq q_{1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

W tym celu wystarczy wyznaczyć kwantyle $q_{\frac{\alpha}{2}}$ i $q_{1-\frac{\alpha}{2}}$ rzędu $\frac{\alpha}{2}$ i $1-\frac{\alpha}{2}$ rozkładu W.

Na podstawie wylosowanej próby o długości n wyznaczamy statystyki porządkowe rzędu $[n\alpha/2]$ i $[n(1-\alpha/2)]$ i oznaczamy je przez $\widetilde{w}_{n,\alpha/2}$ i $\widetilde{w}_{n,1-\alpha/2}$.

Wtedy:

$$P\left(\widetilde{w}_{n,\alpha/2} \leq W \leq \widetilde{w}_{n,1-\alpha/2}\right) \approx 1-\alpha.$$

p-value i moc testu

Jeżeli potrafimy wyznaczyć kwantyle dowolnego rozkładu, to możemy oszacować p-value i moc testu.

Aby zweryfikować hipotezę, że w partii 2000 elementów występuje mniej niż 10% braków wylosowano z niej próbę o liczebności 100 elementów. 14 spośród nich było wybrakowanych. Czy są podstawy do odrzucenia hipotezy głównej na poziomie istotności $\alpha = 0{,}05$?

$$H_0: p = 0{,}1$$

$$H_1: p < 0{,}1$$

Losujemy n prób 100-elementowych i szacujemy prawdopodobieństwo $P(X \geq 14)$. W ten sposób uzyskujemy wartość p-value tego testu.

Modelowanie eksperymentów

Metodę Monte Carlo można wykorzystać również przy analizie skomplikowanych eksperymentów losowych. Zwłaszcza takich, że:

- ze względu na duży stopień złożoności trudno jest obliczyć wartości interesujących nas parametrów,
- trudno jest stworzyć adekwatny model analityczny pozwalający opisać dany problem.

Przykład:

Usługa składa się z dwóch etapów: A i B. Czas realizacji (w godzinach) etapu A jest losowy i ma rozkład wykładniczy z $\lambda = 2$. Czas realizacji etapu B ma rozkład jednostajny na przedziale $(1/6, 1/3)$. Są 2 niezależne stanowiska do realizacji etapu A i jedno do etapu B/ Klienci zgłaszają się zgodnie z procesem Poissona z $\lambda = 1$, tzn. przybywają w losowych momentach czasu. Jeżeli jest wolne stanowisko etapu A to są obsługiwani, jeżeli nie to rezygnują z usługi. Po zakończeniu etapu A klienci czekają w kolejności na realizację etapu B. Interesuje nas czas W obsługi klienta.