

STATYSTYKA

W rachunku prawdopodobieństwa mówiąc o rozkładzie zmiennej losowej zakłada się, że jest on znany i na tej podstawie wyznacza się prawdopodobieństwa różnych zdarzeń natomiast w statystyce obserwując wyniki doświadczeń losowych (czyli tzw. Realizacji zmiennej losowej) należy wyciągnąć wnioski o rozkładzie zmiennej losowej.

Założmy że mamy zmienną losową X , która jest określona na przestrzeni zdarzeń elementarnych Ω , nazywanej często w statystyce populacją. Zmienna losowa X jest natomiast interpretowana jako pewna cecha tej populacji, np. wzrost, waga, zarobki, preferencje wyborcze, parametry elektryczne produkowanych elementów itp.

Aby określić rozkład cechy X najlepiej byłoby zbadać całą populację. Jednak często zadanie takie jest niewykonalne lub zbyt kosztowne. Zamiast tego dokonuje się pomiaru dla pewnej, losowo wybranej części tej populacji (oznaczmy ją przez $x_1, \dots, x_n$). Część ta jest tak wybrana, aby otrzymać reprezentatywną próbę potencjalnych wartości zmiennej losowej X .

Wynik takiego badania próby też jest zmienną losową, rozkładzie takim samym jak zmienna X opisująca całą populację.

ESTYMATORY ZMIENNYCH LOSOWYCH

Definicja

Przez statystykę $T_n = T_n(X_1, \dots, X_n)$ będziemy rozumieli zmienną losową będącą funkcją próby losowej prostej $X_1, \dots, X_n$.

Statystykę służącą do oszacowania nieznanego parametru θ cechy X w populacji nazywamy **estymatorem**.

Pożądane własności estymatorów:

- nieobciążony
- zgodny
- efektywny

Definicja

Estymator T_n parametru θ jest nieobciążony, gdy jego wartość oczekiwana:

$$E(T_n) = \theta.$$

W przeciwnym przypadku obciążenie estymatora (ang.: *bias*) wynosi

$$b(T_n) = E(T_n) - \theta.$$

Estymator może być asymptotycznie nieobciążony, gdy

$$\lim_{n \rightarrow \infty} b(T_n) = 0.$$

Definicja

Estymator T_n parametru θ jest zgodny, jeżeli ciąg $\{T_n\}$ jest stochastycznie zbieżny do θ , tzn. dla każdego $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|T_n - \theta| > \varepsilon) = 0.$$

Oznacza to, że w dostatecznie dużej próbie ryzyko popełnienia dużego błędu estymacji jest niewielkie.

Estymator zgodny jest też asymptotycznie nieobciążony.

Definicja

Jeżeli dany jest zbiór wszystkich nieobciążonych estymatorów $T_n^1, T_n^2, \dots, T_n^k$ parametru θ , to estymator T_n^* nazywamy **estymatorem efektywnym**, jeżeli ma w tym zbiorze najmniejszą wariancję, tzn.

$$D^2(T_n^i) \leq D^2(T_n^*) \text{ dla każdego } i = 1 \dots k.$$

Innymi słowy, najlepszy estymator to taki, który ma najmniejszą wariancję. Wariancja estymatora zawsze jest większa od zera - nierówność Rao-Cramera:

$$D^2(T_n) = E(T_n - \theta)^2 \geq \frac{1}{nE\left(\left[\frac{\partial \log f(x; \theta)}{\partial \theta}\right]^2\right)}$$

METODY UZYSKIWANIA ESTYMATORÓW

Metoda momentów

Założmy, że cecha X elementów populacji generalnej ma rozkład normalny $N(m, \sigma)$.

Estymatorem wartości oczekiwanej m jest **średnia arytmetyczna**

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Jest to estymator nieobciążony, zgodny i efektywny.

Średnia arytmetyczna też ma rozkład normalny, ale o parametrach

$$N\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

Im większe n , tym większe prawdopodobieństwo że średnia arytmetyczna przyjmie wartość bliską „prawdziwej” wartości m .

Estymatorem wariancji σ^2 jest wariancja z próby:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

która ma rozkład χ^2 .

Niestety, jest to estymator **obciążony** (aczkolwiek asymptotycznie nieobciążony)

Nieobciążony estymator wariancji ma postać

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

W przypadku średniej arytmetycznej niezależnych zmiennych losowych o rozkładzie normalnym z nieznanym odchyleniem standardowym można skorzystać ze statystyki

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

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

METODY UZYSKIWANIA ESTYMATORÓW

Metoda największej wiarygodności

Metoda największej wiarygodności jest najbardziej uniwersalną metodą estymacji punktowej i jej estymatory asymptotycznie najlepsze własności.

Funkcją wiarygodności nazywamy funkcję postaci:

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \cdot \dots \cdot f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

gdzie zmienna losowa X ma rozkład prawdopodobieństwa f , $x_1, \dots, x_n$ to dane empiryczne, a θ jest szukanym parametrem (lub wektorem parametrów).

Estymacja parametru θ polega na znalezieniu maksimum funkcji wiarygodności.

W praktyce liczy się maksimum logarytmu funkcji wiarygodności: łatwiej jest policzyć pochodne sumy niż iloczynu.

Metoda największej wiarygodności

Przykład: za pomocą metody największej wiarygodności ustalić estymator wartości oczekiwanej dla rozkładu normalnego $N(m, \sigma)$

Funkcja wiarygodności ma w tym przypadku postać

$$L(m, \sigma^2, x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-m)^2}{2\sigma^2}} \text{ więc jej logarytm to}$$
$$l(m, \sigma^2) = \ln L(m, \sigma^2) = \sum_{i=1}^n \left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x_i-m)^2}{2\sigma^2} \right] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2 =$$
$$= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2.$$

Pochodne cząstkowe:

$$\frac{\partial l(m, \sigma^2)}{\partial m} = -\frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - m) \cdot (-1) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - m)$$
$$\frac{\partial l(m, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - m)^2.$$

Przyrównujemy do zera i wyliczamy m, σ - **dokończyć zadanie!**

Twierdzenie (prawo wielkich liczb Bernoulliego)

Niech $\{X_n\}$ ($n = 1, 2, \dots$) będzie ciągiem zmiennych losowych przyjmujących wartości

$$0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1$$

z prawdopodobieństwem

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

dla pewnego $0 < p < 1$ i $k = 0, 1, \dots, n$.

Wówczas ciąg zmiennych losowych $\{X_n\}$ jest stochastycznie zbieżny do p , tzn.

dla dowolnego $\varepsilon > 0$ jest

$$\lim_{n \rightarrow \infty} P(|X_n - p| > \varepsilon) = 0.$$

Twierdzenie to oznacza, że w schemacie Bernoulliego z prawdopodobieństwem sukcesu p wraz ze wzrostem liczby prób maleje do zera prawdopodobieństwo, że częstość sukcesów będzie znacząco różniła się od p . Innymi słowy: wzrasta do 1 prawdopodobieństwo, że częstość sukcesów będzie bliska p . Należy zauważyć, że tempo zmiernienia do 0 zależy zarówno od p jak i od ε :

