Estymacja punktowa

Pojęcia wstępne

Jedną z form wnioskowania statystycznego jest *estymacja*, czyli szacowanie nieznanych wartości parametrów rozkładu pewnej cechy w populacji na podstawie próbki.

Niech zmienna losowa X będzie modelem badanej cechy populacji generalnej.

Będziemy dalej zakładać, że ciąg liczbowy $x_1, x_2, ..., x_n$ jest realizacją ciągu $X_1, X_2, ..., X_n$, gdzie X_i ,i=1, 2, ... , n, jest zmienną losową, której zbiorem możliwych wartości są wartości i-tego spośród n wylosowanych elementów. Zakładamy przy tym, że zmienne X_1 , X_2 , ... , X_n są niezależne i każda z nich ma rozkład taki sam, jak rozkład badanej cechy populacji. Ciąg takich zmiennych losowych będziemy nazywać *n-elementową próbą losową prostą*.

Ciąg liczb $x_1, x_2, \ldots x_n$ będziemy nazywać *zaobserwowaną próbą losową* bądź po prostu *próbką*.

Pojęcie estymatora

Niech $X_1, X_2, \ldots, X_n$ będzie próbą prostą z populacji, której cecha X, traktowana jako obserwowana zmienna losowa, ma rozkład zależny od nieznanej wartości parametru θ.

Każdą funkcję próby (statystykę)

$$\theta_n = \theta_n(X_1, X_2, \ldots, X_n)$$

której wartość $\theta_n(x_1, x_2, \ldots, x_n)$ dla aktualnie zaobserwowanej próbki $x_1, x_2, \ldots, x_n$, przyjmujemy jako ocenę (oszacowanie) nieznanej wartości parametru θ nazywamy *estymatorem parametru* θ:

$$\theta_n(x_1, x_2, \ldots, x_n) \approx \theta$$

Uwaga. Estymator – jako funkcja zmiennych losowych – sam jest zmienną losową. Zmienna ta przyjmuje wartości:

$$\theta_n^1, \theta_n^2, \ldots, \theta_n^k$$

Pożądane cechy estymatorów:

1) nieobciążoność estymatora

Estymator θ_n parametru θ nazywamy *nieobciążonym*, jeżeli dla dowolnej liczności próby n wartość oczekiwana estymatora jest równa szacowanemu parametrowi θ, czyli

$$E\theta_n = \theta .$$

2) efektywność estymatora

Estymator θ_n parametru θ nazywamy *najefektywniejszym*, jeżeli przy danej liczności próbki n ma najmniejszą wariancję $D^2\theta_n$.

3) zgodność estymatora

Estymator θ_n parametru θ nazywamy *zgodnym*, jeżeli wraz ze wzrostem liczności próbki wzrasta dokładność oszacowania parametru θ. Warunek ten możemy zapisać następująco

$$\bigwedge_{\varepsilon>0} \lim_{n\to\infty} P(|\theta_n - \theta| > \varepsilon) = 0.$$

Estymacja wartości oczekiwanej

Niech zmienna losowa X będzie modelem badanej cechy populacji generalnej. Chodzi nam o oszacowanie wartości oczekiwanej zmiennej losowej X, czyli $\theta = EX$.

Najczęściej używanym estymatorem wartości oczekiwanej jest średnia arytmetyczna z próby:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

gdzie X_1 , X_2 , ... , X_n jest próbą losową z populacji generalnej, w której obserwowana jest zmienna losowa X.

Jeżeli x_1 , x_2 , ... x_n jest aktualnie zaobserwowaną próbką, to

$$EX \approx \bar{x}\ , \quad \text{gdzie} \quad \bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n)$$

(czyli $\bar{x}$ jest zaobserwowaną wartością zmiennej los. $\overline{X}$).

Twierdzenie. Jeżeli cecha X elementów populacji generalnej ma dowolny rozkład i skończoną wartość oczekiwaną EX, to średnia arytmetyczna $\bar{X}$ n-elementowej próby z tej populacji jest nieobciążonym i zgodnym estymatorem wartości oczekiwanej. Jeżeli X ma rozkład normalny, to estymator ten jest najefektywniejszy.

Uwaga. W roli estymatora wartości oczekiwanej obserwowanej cechy X populacji generalnej używa się również mediany z próby X_{me} oraz średniej arytmetycznej wartości skrajnych

$$\frac{1}{2}(X_{\min} + X_{\max}).$$

Estymatory wariancji

Niech estymowanym parametrem obserwowanej zmiennej losowej X będzie wariancja, czyli $\theta = D^2X$. Najczęściej stosowanymi estymatorami wariancji są:

1) wariancja z próby

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

gdzie X_1, X_2, ..., X_n jest próbą losową z populacji generalnej, w której obserwowana jest zmienna losowa X, a $\bar{X}$ jest określoną wcześniej średnią arytmetyczną tej próby.

Uwaga. Estymator ten jest zgodny, ale nie jest nieobciążony.

2) skorygowana wariancja z próby

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

Uwaga. Estymator ten jest zgodny oraz nieobciążony.

3) jeżeli znana jest wartość oczekiwana EX, to można posłużyć się następującym estymatorem wariancji:

$$S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - EX)^2$$

Uwaga. Estymator ten jest zgodny oraz nieobciążony, a dla rozkładu normalnego – najefektywniejszy.

Estymatory odchylenia standardowego $\sigma = \sqrt{D^2 X}$ można utworzyć z powyższych estymatorów wariancji (przez pierwiastkowanie). Nie będą one jednak nieobciążone.

Estymacja przedziałowa

Pojęcia wstępne

Estymacja przedziałowa polega na szacowaniu nieznanej wartości parametru θ rozkładu cechy X w populacji za pomocą pewnego przedziału, który z założonym z góry prawdopodobieństwem pokrywa rzeczywistą wartość parametru θ.

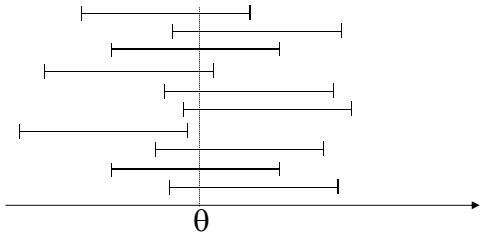
Przedziałem ufności dla parametru θ na *poziomie ufności* $1-\alpha$ $(0<\alpha<1)$ nazywamy przedział (θ_1 , θ_2) spełniający warunki:

- jego końce $\theta_1 = \theta_1(X_1, X_2, \ldots , X_n)$, $\theta_2 = \theta_2(X_1, X_2, \ldots , X_n)$ są funkcjami próby losowej i nie zależą od szacowanego parametru θ,
- prawdopodobieństwo pokrycia przez ten przedział nieznanej wartości parametru θ jest równe $1-\alpha$, tzn.

$$P(\theta_1 < \theta < \theta_2) = 1-\alpha .$$

Liczbę $1-\alpha$ nazywamy także *współczynnikiem ufności*. Najczęściej przyjmuje się $1-\alpha=0{,}99$, $1-\alpha=0{,}95$, $1-\alpha=0{,}90$.

Przedział liczbowy (g_1 , g_2), którego końce są zaobserwowanymi wartościami odpowiednio zmiennych losowych θ_1 , θ_2 dla aktualnie zaobserwowanej próbki $x_1, x_2, \ldots, x_n$ będziemy nazywać *realizacją przedziału ufności* (θ_1 , θ_2) lub *zaobserwowanym przedziałem ufności*. W konkretnych zagadnieniach praktycznych zawsze mamy do czynienia z wyznaczeniem realizacji przedziału ufności, a nie samego przedziału ufności.



Lewostronnym przedziałem ufności dla parametru θ ze współczynnikiem ufności $1-\alpha$ nazywamy przedział losowy $(\theta_1, +\infty)$, który z prawdopodobieństwem $1-\alpha$ pokrywa nieznaną wartość parametru θ, tj.

$$P(\theta_1 < \theta < +\infty) = 1-\alpha \ .$$

Podobnie definiujemy *prawostronny przedział ufności* $(-\infty, \theta_2)$

$$P(-\infty < \theta < \theta_2) = 1-\alpha \ .$$

Kwantylem rzędu p zmiennej losowej X typu ciągłego o dystrybuancie F i gęstości f nazywamy liczbę x_p spełniającą którykolwiek z następujących równoważnych warunków:

$$F(x_p) = p \ , \quad P(X < x_p) = p \ , \quad \int_{-\infty}^{x_p} f(x)dx = p \ .$$

Przedziały ufności dla wartości oczekiwanej

Model I. Założenia:

1) cecha X elementów populacji generalnej ma rozkład normalny $N(\mu, \sigma)$,
2) odchylenie standardowe σ jest znane przed pobraniem próbki,
3) n-liczność próbki dowolna.

Twierdzenie. Jeżeli $\overline{X}$ jest średnią arytmetyczną próby prostej z badanej populacji, daną równością

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

oraz liczby $u_{1-\frac{1}{2}\alpha}$, $u_{1-\alpha}$ oznaczają kwantyle odpowiednio rzędu $1-\frac{1}{2}\alpha$ oraz $1-\alpha$ zmiennej losowej U o rozkładzie normalnym $N(0, 1)$, to przedziały losowe

$$\left(\overline{X} - u_{1-\frac{1}{2}\alpha} \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + u_{1-\frac{1}{2}\alpha} \frac{\sigma}{\sqrt{n}} \right) ,$$

$$\left(\overline{X} - u_{1-\alpha} \frac{\sigma}{\sqrt{n}} \ , \ +\infty \right) ,$$

$$\left(-\infty \ , \ \overline{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right) ,$$

można uważać odpowiednio jako dwustronny, lewostronny, prawostronny przedział ufności ze współczynnikiem ufności $1-\alpha$ dla wartości oczekiwanej μ .

Przykład 1. Niech X oznacza zużycie przędzy (w gramach) na wyprodukowanie jednego metra bieżącego tkaniny płaszczowej. Dokonano $n = 9$ niezależnych, jednakowo dokładnych pomiarów i otrzymano następujące wyniki x_i:

473, 482, 489, 464, 476, 487, 468, 474, 480.

Zakładając, że rozważana tu cecha X ma rozkład normalny $N(\mu,8)$, wyznaczyć 95-procentową ($1-\alpha = 0{,}95$) realizację dwustronnego przedziału ufności dla przeciętnego zużycia przędzy na wyprodukowanie jednego metra tkaniny.

Model II. Założenia:

1) cecha X elementów populacji generalnej ma rozkład normalny $N(\mu,\sigma)$,

2) odchylenie standardowe σ nie jest znane,

3) n-liczność próbki dowolna.

Twierdzenie. Jeżeli $\overline{X}$ i S oznaczają odpowiednio średnią arytmetyczną z próby i odchylenie standardowe z próby w badanej populacji oraz liczby $t_{n-1,\,1-\frac{1}{2}\alpha}$ i $t_{n-1,1-\alpha}$ oznaczają kwantyle rzędu $1-\frac{1}{2}\alpha$ oraz $1-\alpha$ zmiennej losowej T o rozkładzie T-studenta z n-1 stopniami swobody, to przedziały losowe

$$\left(\overline{X}-t_{n-1,\,1-\frac{1}{2}\alpha}\frac{S}{\sqrt{n-1}}\;,\;\overline{X}+t_{n-1,\,1-\frac{1}{2}\alpha}\frac{S}{\sqrt{n-1}}\right),$$

$$\left(\overline{X}-t_{n-1,1-\alpha}\frac{S}{\sqrt{n-1}}\;,\;+\infty\right),$$

$$\left(-\infty\;,\;\overline{X}+t_{n-1,1-\alpha}\frac{S}{\sqrt{n-1}}\right),$$

można uważać odpowiednio jako dwustronny, lewostronny, prawostronny przedział ufności ze współczynnikiem ufności $1-\alpha$ dla wartości oczekiwanej μ cechy X elementów populacji generalnej.

Przykład 2. Treść jak w przykładzie 1, przy czym nie zakładamy znajomości odchylenia standardowego σ.
Wyznaczyć realizację 95-procentowego dwustronnego przedziału ufności dla przeciętnego zużycia przędzy na wyprodukowanie jednego metra tkaniny.

Model III. Założenia:

1) cecha X elementów populacji generalnej ma dowolny rozkład,
2) odchylenie standardowe σ cechy X nie jest znane,
3) n-liczność próbki duża $(n \geq 30)$.

Przedziały ufności (tak, jak w modelu I – w miejsce σ podstawiamy S lub $\widetilde{S}$):

$$\left(\overline{X} - u_{1-\frac{1}{2}\alpha} \frac{S}{\sqrt{n}} \ , \ \overline{X} + u_{1-\frac{1}{2}\alpha} \frac{S}{\sqrt{n}} \right) ,$$

$$\left(\overline{X} - u_{1-\alpha} \frac{S}{\sqrt{n}} \ , \ +\infty \right) , \qquad \left(-\infty \ , \ \overline{X} + u_{1-\alpha} \frac{S}{\sqrt{n}} \right) .$$

Przedziały ufności dla odchylenia standardowego

Model I. Założenia:

1) cecha X elementów populacji generalnej ma rozkład normalny $N(\mu,\sigma)$,
2) wartość oczekiwana μ i odchylenie standardowe σ nie są znane,
3) n-liczność próbki dowolna.

Przedziały ufności

$$\left(\sqrt{\frac{nS^2}{\chi^2_{n-1,\,1-\frac{1}{2}\alpha}}} \,,\, \sqrt{\frac{nS^2}{\chi^2_{n-1,\,\frac{1}{2}\alpha}}} \right),$$

$$\left(0 \,,\, \sqrt{\frac{nS^2}{\chi^2_{n-1,\,\alpha}}} \right), \qquad \left(\sqrt{\frac{nS^2}{\chi^2_{n-1,\,1-\alpha}}} \,,\, +\infty \right),$$

$\chi^2_{k,p}$ – kwantyl rzędu p zmiennej losowej o rozkładzie chi-kwadrat z k stopniami swobody (jego wartość odczytujemy z tablic).

Przykład 3. Treść jak w przykładzie 1, przy czym nie zakładamy znajomości odchylenia standardowego σ. Wyznaczyć realizację 90-procentowego dwustronnego przedziału ufności dla odchylenia standardowego.

Model II. Założenia:

1) cecha X elementów populacji generalnej ma rozkład normalny lub zbliżony do normalnego,
2) wartość oczekiwana μ i odchylenie standardowe σ nie są znane,
3) n-liczność próbki duża $(n \geq 30)$.

Przedziały ufności:

$$\left(\frac{S}{1 + \frac{u_{1-\frac{1}{2}\alpha}}{\sqrt{2n}}} \; , \; \frac{S}{1 - \frac{u_{1-\frac{1}{2}\alpha}}{\sqrt{2n}}} \right),$$

$$\left(0 \; , \; \frac{S}{1 - \frac{u_{1-\alpha}}{\sqrt{2n}}} \right), \qquad \left(\frac{S}{1 + \frac{u_{1-\alpha}}{\sqrt{2n}}} \; , \; +\infty \right).$$

u_p – kwantyl rzędu p standaryzowanego rozkładu normalnego

Przykład 4. W celu oszacowania czasu przebywania na zwolnieniach pracowników pewnego zakładu wybrano losowo grupę 100 pracowników i zanotowano liczby dni opuszczonych z powodu choroby w ciągu całego roku. Otrzymano następujące wyniki:

Liczba opuszczonych dni	Liczba pracowników
0-4	13
4-8	37
8-12	22
12-16	17
16-20	8
20-24	2
24-28	1

Na podstawie wyników tego badania znaleźć 95% realizację przedziału ufności dla średniego czasu niezdolności do pracy wszystkich pracowników tego zakładu.