Estymacja parametrów populacji

Estymacja parametrów populacji

Estymacja polega na szacowaniu wartości parametrów rozkładu lub postaci samego rozkładu zmiennej losowej, na podstawie próby statystycznej.

Estymacje można podzielić na:

- ***estymację punktową*** - wyznaczanie na podstawie próby statystycznej konkretnych wartości parametrów dla całej zbiorowości generalnej (mogą to być takie parametry jak wartość oczekiwana, mediana, czy odchylenie standardowe),
- ***estymację przedziałową*** polegającą na konstruowaniu przedziału liczbowego, który z ustalonym z góry, wysokim prawdopodobieństwem pokrywa nieznaną wartość szacowanego parametru. Przedział taki nazywa się przedziałem ufności, a prawdopodobieństwo, z jakim pokrywa on szacowany parametr - współczynnikiem ufności.

Przedział ufności

Granice przedziału ufności są **losowe**, a więc dla konkretnych prób będziemy uzyskiwać różne wartości. Uzyskany konkretny przedział będziemy interpretować następująco: **w** ***1-α*** **procentach przypadków przedział (*a*, *b*) pokrywa nieznaną wartość parametru.**

Oznacza to jednocześnie, że średnio w α procentach przypadków wyznaczony przedział **nie pokrywa** szacowanego parametru.

Przedział ufności (c.d.)

Dokładność estymacji parametru określa rozpiętość przedziału ufności będąca różnicą między jego górną i dolną granicą: $d = b - a$.

Rozpiętość przedziału ufności zależy między innymi od przyjętego poziomu ufności $1-\alpha$: im to prawdopodobieństwo jest bliższe jedności, tym rozpiętość przedziału jest większa (a precyzja oceny mniejsza).

W zastosowaniach praktycznych najczęściej stosujemy poziomy ufności rzędu 0.90, 0.95 czy 0.99 (α odpowiednio 0.10, 0.05 czy 0.01)

Przedział ufności dla wartości oczekiwanej – znane odchylenie standardowe

Jeśli znamy odchylenie standardowe zbiorowości, to wartość szacowanej średniej, z prawdopodobieństwem równym 1-α, znajduje się w przedziale danym wzorem:

$$P\left(\bar{x}-u_\alpha \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x}+u_\alpha \frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

gdzie:

$\bar{x}$ -średnia arytmetyczna,

u_α -wartość odczytana z tablicy rozkładu *t*-Studenta dla liczby stopni swobody r=∞,

σ - odchylenie standardowe,

n - liczebność próby,

m - wartość oczekiwana.

Przykład

Zakładając, że ceny jednostkowe lokali mieszkalnych z danego przykładu w miejscowości A w pierwszym kwartale 2019 r. mają rozkład zbliżony do rozkładu normalnego *N(6460,241)* oraz, że znane jest odchylenie standardowe zbiorowości $\sigma = 241$ oraz liczebność próby n = 30 oszacować przedział ufności dla nieznanej wartości średniej zbiorowości. Przyjmijmy współczynnik ufności 1-α = 0,95

Rozwiązanie

Podstawiając powyższe dane do danego wzoru oraz odczytując z tablicy zmiennej losowej *t*-Studenta wartość krytyczną, dla liczby stopni swobody *r*=∞ (albowiem odchylenie standardowe zbiorowości jest znane) i α = 0,05 otrzymujemy:

$$P\left(6460 - 1{,}96\frac{241}{\sqrt{30}} \leq m \leq 6460 + 1{,}96\frac{241}{\sqrt{30}}\right)=0{,}95$$

Oznacza to, że przedział liczbowy $\mathrm{P}(6373{,}8 \leq m \leq 6546{,}2) = 0{,}95$ z prawdopodobieństwem 1-α = 0,95 pokrywa nieznaną wartość m.

Przedział ufności dla wartości oczekiwanej – odchylenie standardowe nie jest znane

Jeżeli odchylenie standardowe zbiorowości nie jest znane, to przedział ufności dla wartości oczekiwanej m należy skonstruować w oparciu o rozkład t-Studenta:

$$P\left(\bar{x} - t_{\alpha;r}\frac{s}{\sqrt{n-1}} \le m \le \bar{x} + t_{\alpha;r}\frac{s}{\sqrt{n-1}}\right) = 1 - \alpha$$

gdzie:

$r = n$ -1 stopni swobody,

s - odchylenie standardowe.

Przykład

Zakładając, że ceny jednostkowe lokali mieszkalnych z danego przykładu w miejscowości A w pierwszym kwartale 2019 r. mają rozkład zbliżony do rozkładu normalnego $N(6460, \sigma)$, *s = 241 oraz* liczebność próby n = 29 oszacować przedział ufności dla nieznanej wartości średniej. Przyjmijmy współczynnik ufności 1-α = 0,95

Rozwiązanie

Podstawiając powyższe dane do wzoru oraz odczytując z tablicy zmiennej losowej *t*-Studenta wartość, dla liczby stopni swobody *r* = *n*-1=28 i α = 0,05 otrzymujemy:

$$P\left(6460 - 2{,}048\frac{241}{\sqrt{28}} \leq m \leq 6460 + 2{,}048\frac{241}{\sqrt{28}}\right) = 0{,}95$$

$$P(6366{,}7 \leq m \leq 6553{,}3) = 0{,}95$$

Można zatem stwierdzić, że z prawdopodobieństwem 95% średnia cena jednostkowa lokali mieszkalnych w miejscowości A zawiera się w przedziale liczbowym o końcach 6366,7 zł/m^2 i 6553,3 zł/m^2.

Jeżeli liczba obserwacji n dąży do nieskończoności, to różnica między wyżej podanymi przedziałami jest bardzo mała. Dzieje się tak dlatego, że rozkład t-Studenta jest zbieżny do rozkładu normalnego. Występuje to wtedy, gdy liczba stopni swobody (n-1) wzrasta nieograniczenie. Począwszy od n=30 różnicę między tymi przedziałami można praktycznie zaniedbać.

Przedział ufności dla wariancji σ^2 w populacji normalnej

Niech zmienna losowa $X \sim N(m,\sigma)$ oraz niech x_i ($i = 1, 2, ..., n$) oznacza n-elementową próbę losową.

Statystyka
$$\chi^2 = \frac{ns^2}{\sigma^2}$$

ma rozkład χ^2 z liczbą stopni swobody $v = n - 1$.

Dla ustalonego α można określić takie dwie wartości $\chi^2_{\frac{\alpha}{2},n-1}$ i $\chi^2_{1-\frac{\alpha}{2},n-1}$, dla których spełnione są równości:

$$P(\chi^2 \geq \chi^2_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2} \qquad P(\chi^2 \geq \chi^2_{1-\frac{\alpha}{2},n-1}) = 1 - \frac{\alpha}{2}$$

Przedział ufności dla wariancji σ^2 w populacji normalnej (c.d.)

Z obu wzorów wynika, że

$$P(\chi^2_{1-\frac{\alpha}{2},n-1} < \chi^2 < \chi^2_{\frac{\alpha}{2},n-1}) = 1-\alpha$$

Po odpowiednich przekształceniach otrzymujemy przedział ufności dla wariancji:

$$P\left(\frac{ns^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{ns^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

Przedział ufności dla odchylenia standardowego σ w populacji normalnej.

Pierwiastkując krańce przedziału ufności dla wariancji otrzymujemy poszukiwany przedział dla odchylenia standardowego:

$$P\left(\sqrt{\frac{ns^2}{\chi^2_{\frac{\alpha}{2},n-1}}} < \sigma < \sqrt{\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}}\right) = 1 - \alpha$$

Przedział ufności dla odchylenia standardowego w populacji normalnej – dla dużej próby

$$\mathrm{P}\left(S - u_{\alpha}\frac{S}{\sqrt{2n}} < \sigma < S + u_{\alpha}\frac{S}{\sqrt{2n}}\right)=1-\alpha$$

Gdzie:

u_{α} - wartość odczytana z tablicy rozkładu t-Studenta dla liczby stopni swobody $r=\infty$,

Przedział ufności dla prawdopodobieństwa w populacji normalnej

$$P\left(\frac{m}{n} - u_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}} < p < \frac{m}{n} + u_\alpha\sqrt{\frac{\frac{m}{n}\left(1-\frac{m}{n}\right)}{n}}\right) = 1-\alpha$$

Gdzie:
m - liczba jednostek w próbie mających wyróżnioną cechę,
n - liczebność próby,

$W = \frac{m}{n}$ - wskaźnik struktury w próbie, który jest estymatorem

prawdopodobieństwa p w populacji generalnej

u_α - wartość odczytana z tablicy rozkładu t-Studenta dla liczby stopni swobody $r = \infty$,

Uzasadnienie wielkości próby

$$n = \frac{u_\alpha^2 \sigma^2}{d^2}$$

(wariancja jest znana)

$$n = \frac{t_\alpha^2 \widehat{s}^2}{d^2}$$

(wariancja jest nieznana)

gdzie:

$$\widehat{s}^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (x_i - \overline{x})^2$$

$$n = \frac{u_\alpha^2 pq}{d^2}$$

Zbiorowość generalna ma rozkład dwupunktowy z parametrem p (p jest frakcją jedynek lub elementów wyróżnionych w zbiorowości.)

Jeżeli nie znamy rzędu wielkości szacowanego wskaźnika struktury p, to przyjmując za iloczyn pq jego największą wartość ¼ , otrzymujemy poniższy wzór:

$$n = \frac{u_\alpha^2}{4d^2}$$

gdzie:

d - dopuszczalny, ustalony z góry maksymalny błąd szacunku wartości m.