Bootcamp Data Science

Statystyka

Przemysław Spurek

Estymatory

Definicja

Prostą próbą losową (lub krócej próbą losową) o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ określonych na przestrzeni zdarzeń elementarnych Ω i takich, że każda ze zmiennych ma taki sam rozkład.

Definicja

Statystyką nazywamy każdą zmienną losową będącą ustaloną funkcją próby losowej $X_1, X_2, \ldots, X_n$.

Definicja

Każdą statystykę, którą przyjmujemy jako oszacowanie (przybliżenie) nieznanego parametru rozkładu będziemy nazywać *estymatorem*.

Przedział ufności (confidence interval)

Metody estymacji, którymi zajmowaliśmy się dotychczas, pozwalały uzyskać oceny punktowe nieznanych parametrów rozkładu, przy czym nie potrafiliśmy dać odpowiedzi na pytanie, jaka jest dokładność uzyskanej oceny.

Przedział ufności (confidence interval)

Metody estymacji, którymi zajmowaliśmy się dotychczas, pozwalały uzyskać oceny punktowe nieznanych parametrów rozkładu, przy czym nie potrafiliśmy dać odpowiedzi na pytanie, jaka jest dokładność uzyskanej oceny.

Innym sposobem estymacji, dającym możliwość oceny tej dokładności, jest **metoda przedziałowa** polegająca na podaniu tzw. **przedziałów ufności** dla nieznanych parametrów danego rozkładu.

Przedział ufności (confidence interval)

Definicja

Przedziałem ufności dla parametru Θ na poziomie ufności $1-\alpha$ $(0 < \alpha < 1)$ nazywamy przedział (Θ_1, Θ_2) spełniający warunki:

1. jego końce są funkcjami $\Theta_1 = \Theta_1(X_1, \ldots, X_n)$, $\Theta_2 = \Theta_2(X_1, \ldots, X_n)$ i nie zależą od szacowanego parametru Θ
2. Prawdopodobieństwo pokrycia przez ten przedział nieznanego parametru Θ jest równe $1-\alpha$

$$P\big(\Theta_1(X_1, \ldots, X_n) < \Theta < \Theta_2(X_1, \ldots, X_n)\big) = 1 - \alpha$$

Przedział ufności (confidence interval)

Definicja

Przedziałem ufności dla parametru Θ na poziomie ufności $1-\alpha$ $(0 < \alpha < 1)$ nazywamy przedział (Θ_1, Θ_2) spełniający warunki:

1. jego końce są funkcjami $\Theta_1 = \Theta_1(X_1, \ldots, X_n)$, $\Theta_2 = \Theta_2(X_1, \ldots, X_n)$ i nie zależą od szacowanego parametru Θ
2. Prawdopodobieństwo pokrycia przez ten przedział nieznanego parametru Θ jest równe $1-\alpha$

$$P\big(\Theta_1(X_1, \ldots, X_n) < \Theta < \Theta_2(X_1, \ldots, X_n)\big) = 1 - \alpha$$

Definicja

Parametr $1-\alpha$ nazywa się współczynnikiem ufności.

Przedział ufności (confidence interval)

Jak widać z definicji końce przedziału ufności są zmiennymi losowymi. Nieznana wartość parametru θ może więc być pokryta przez ten losowy przedział bądź nie.

Przedział ufności (confidence interval)

Jak widać z definicji końce przedziału ufności są zmiennymi losowymi. Nieznana wartość parametru θ może więc być pokryta przez ten losowy przedział bądź nie.

Jeżeli jednak dla różnych zaobserwowanych próbek losowych $x_1, \ldots, x_n$ znajdziemy wiele realizacji przedziału ufności, to część z tych, które będą zawierać rzeczywistą wartość parametru θ w dużej liczbie tych realizacji, będzie w przybliżeniu równa $1 - \alpha$.

Przedział ufności (confidence interval)

`https://github.com/przem85/statistic_4/blob/master/D05_Z01_confidence_interval.ipynb`

Zadanie

Znaleźć przedział ufności dla nieznanej wartości średniej μ populacji, w której badana cecha ma rozkład $N(\mu, \sigma)$, w przypadku gdy σ jest znana, na podstawie n-elementowej próby prostej

$$X_1, \ldots, X_n.$$

Przedział ufności (confidence interval)

Uwaga

Widzimy, że nawet przy wykorzystaniu jednej statystyki U do wyznaczania szukanego **Przedziału ufności** w zależności od sposobu wyboru wartości α_1 i α_2 możemy utworzyć nieskończenie wiele przedziałów ufności.

Przedział ufności (confidence interval)

Uwaga

Widzimy, że nawet przy wykorzystaniu jednej statystyki U do wyznaczania szukanego **Przedziału ufności** w zależności od sposobu wyboru wartości α_1 i α_2 możemy utworzyć nieskończenie wiele przedziałów ufności.

Uwaga

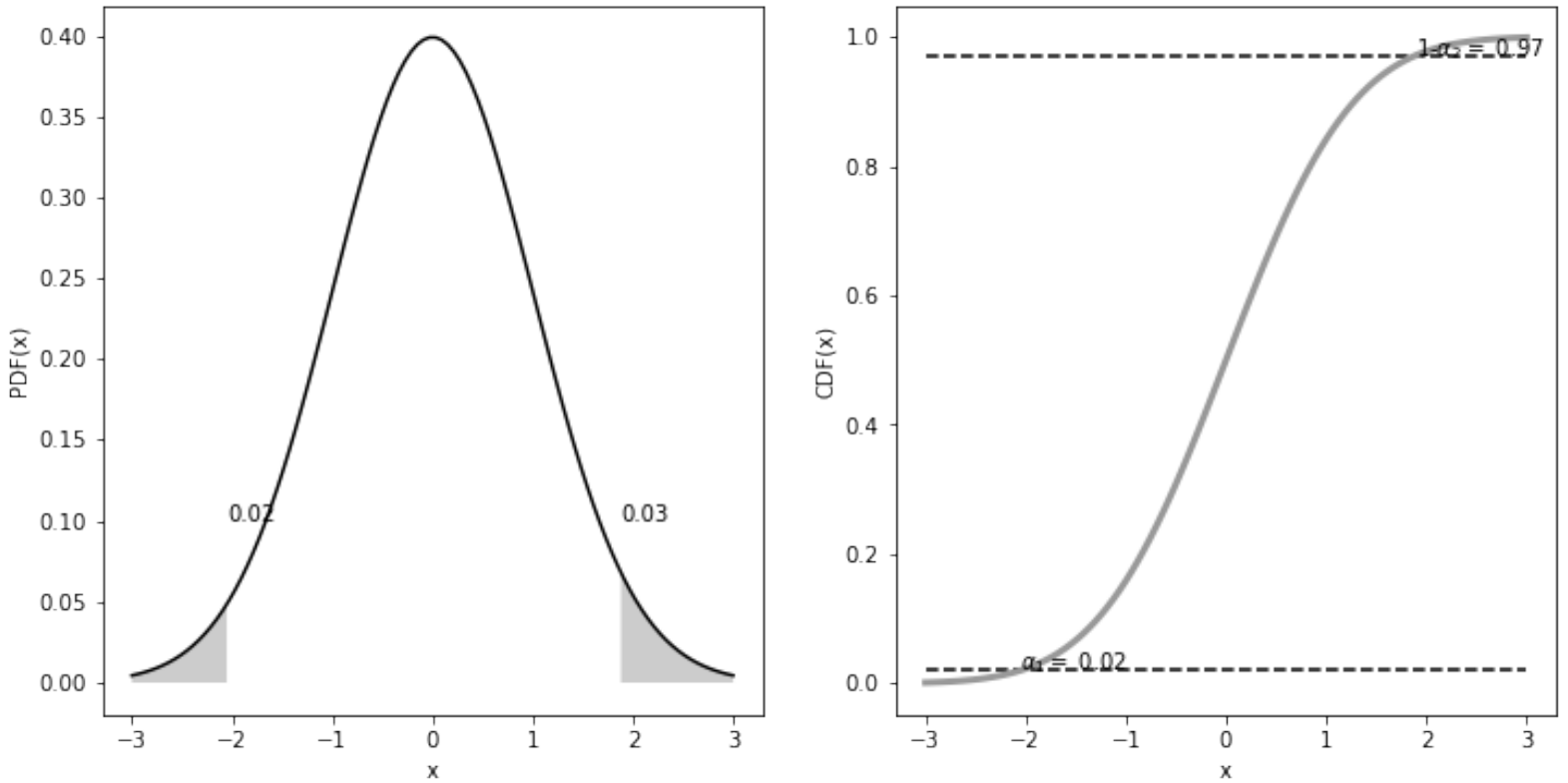
Gdy weźmiemy $\alpha_1 = 0$ to $\alpha_2 = \alpha$.
Wtedy

$$PPF(\alpha_1) = PPF(0) = -\infty, \qquad PPF(1-\alpha_2) = PPF(1-\alpha)$$

i uzyskujemy przedział

$$\left(\bar{X} - PPF(1-\alpha)\frac{\sigma}{\sqrt{n}}, \infty\right)$$

Przedział ufności (confidence interval)



Przedział ufności (confidence interval)

Uwaga

W praktyce najczęściej α_1 i α_1 wybieramy takie, aby

$$\alpha_1 = \alpha_2 = \frac{1}{2}\alpha,$$

otrzymując wówczas

$$\left(\bar{X} - PPF(1 - \frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}, \bar{X} - PPF(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right)$$

Przedział ufności (confidence interval)

Uwaga

W praktyce najczęściej α_1 i α_1 wybieramy takie, aby

$$\alpha_1 = \alpha_2 = \frac{1}{2}\alpha,$$

otrzymując wówczas

$$\left(\bar{X} - PPF(1 - \frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}, \bar{X} - PPF(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right)$$

Za względu na symetrię rozkładu normalnego można go zapisać jako

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

Przedział ufności (confidence interval)

Zadanie

Niech $\sigma = 2$. Załóżmy, że w naszym zadaniu w $n = 16$ próbach wypadło $\bar{x} = 34.1$.
Przyjmijmy współczynnik ufności 0.05 wyznacz przedział największej wiarygodności.

Przedział ufności – modele

Model 1.
Przedział ufności dla **nieznanej wartości przeciętnej** μ populacji, w której **badana cecha ma rozkład** $N(\mu, \sigma^2)$, w przypadku gdy σ **jest znana**, na podstawie n–elementowej próby $X_1, \ldots, X_n$ wynosi:

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

gdzie $u(\alpha)$ oznacza kwantyl rzędu α rozkładu normalnego $N(0, 1)$ oraz

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

Przedział ufności – modele

Zadanie (Już je zrobiliśmy)

Niech $\sigma = 2$. Na podstawie próbki o liczności $n = 16$ wyznaczono $\bar{X} = 34.1$. Przyjmując $\alpha = 0.05$ oblicz przedział ufności dla nieznanej wartości przeciętnej μ uzyskany dla danej próbki przy poziomie ufności $1 - \alpha = 0.95$.

Przedział ufności – modele

Model 2.
Przedział ufności dla **nieznanej wartości przeciętnej** μ populacji, w której **badana cecha ma rozkład** $N(\mu, \sigma^2)$, w przypadku gdy **zarówno** μ **jak** σ **są nieznane**, na podstawie n–elementowej próby $X_1, \ldots, X_n$ **(n<100)** wynosi

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

gdzie

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

oraz $t(\alpha, n)$ oznacza kwantyl rzędu α o n stopniach swobody rozkładu t–studenta.

Przedział ufności – modele

https://github.com/przem85/statistic_4/blob/master/D05_Z02_confidence_interval.ipynb

Zadanie

Zmierzono wytrzymałość 10 losowo wybranych gotowych elementów konstrukcji budowlanych i otrzymano następujące wyniki: 383, 284, 339, 340, 305, 386, 378, 335, 344, 346.
Zakładamy, że rozkład wytrzymałości tych elementów jest rozkładem normalnym $N(\mu, \sigma^2)$ o nieznanych parametrach. Wyznaczyć na podstawie tej próbki 95%–ową realizację przedziału ufności dla nieznanej wartości parametru μ badanej cechy populacji.

Przedział ufności – modele

https://github.com/przem85/statistic_4/blob/master/D05_Z03_confidence_interval.ipynb

Zadanie

W celu wyznaczenia ładunku elektrycznego wykonano 26 pomiarów tego ładunku metodą Millikana, otrzymując:

$$\bar{X} = 1.574 \cdot 10^{-19}, \; S = 0.043 \cdot 10^{-19}$$

Zakładając że pomiary pochodzą z rozkładu normalnego o nieznanym parametrze σ wyznaczyć na podstawie danych 99%-owy przedział ufności.

Model 3.
Przedział ufności dla **nieznanej wartości przeciętnej** μ populacji, w której **badana cecha ma rozkład** $N(\mu, \sigma^2)$, w przypadku gdy **zarówno** μ **jak** σ **są nieznane**, na podstawie n–elementowej próby $X_1, \ldots, X_n$ $(n \geq 100)$ wynosi

$$\left(\bar{X} - u\left(1 - \frac{\alpha}{2}\right)\frac{S^*}{\sqrt{n}}, \bar{X} + u\left(1 - \frac{\alpha}{2}\right)\frac{S^*}{\sqrt{n}}\right)$$

gdzie $u(\alpha)$ oznacza kwantyl rzędu α rozkładu normalnego $N(0,1)$ oraz

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

Przedział ufności – modele

https://github.com/przem85/statistic_4/blob/master/D05_Z04_confidence_interval.ipynb

Zadanie

Z populacji włókien bawełny pobrano 300–elementową próbkę włókien i zmierzono ich długości. Obliczono $\bar{X} = 27,43$ oraz $S^{*2} = 51,598$. Znaleźć 95% realizacje przedziału ufności dla nieznanej wartości.

Przedział ufności – modele

Model 4.
Przedział ufności dla **nieznanego odchylenia standardowego** σ populacji, w której **badana cecha ma rozkład** $N(\mu, \sigma^2)$, w przypadku gdy **zarówno** μ **jak** σ **są nieznane**, na podstawie n–elementowej próby $X_1, \ldots, X_n$ $(n < 50)$ wynosi

$$(g_1(\alpha, n-1) \cdot S^*, g_2(\alpha, n-1) \cdot S^*)$$

gdzie

$$g_1(\alpha, n-1) = \sqrt{\frac{n-1}{\chi^2(1-\frac{1}{2}\alpha, n-1)}} \text{ oraz } g_1(\alpha, n-1) = \sqrt{\frac{n-1}{\chi^2(\frac{1}{2}\alpha, n-1)}}$$

gdzie $\chi^2(\alpha, n)$ oznacza kwantyl rzędu α o n stopniach swobody rozkładu χ^2.

Przedział ufności – modele

https://github.com/przem85/statistic_4/blob/master/D05_Z05_confidence_interval.ipynb

Zadanie

Wykonano pomiar liczby skrętów dla losowo wybranych odcinków przędzy o długości 1 m, uzyskując wyniki: 87, 102, 119, 81, 97, 93, 100, 113, 99, 100, 112, 93, 95, 85, 123, 99.
Zakładając, że liczba skrętów odcinków przędzy ma rozkład normalny, znaleźć 90%-owe realizacje przedziałów ufności wariancji i odchylenia standardowego liczby skrętów całej partii przędzy.

Model 5.
Przedział ufności dla nieznanego **odchylenia standardowego** σ populacji, w której **badana cecha ma rozkład** $N(\mu, \sigma^2)$, w przypadku gdy **zarówno** μ **jak** σ **są nieznane**, na podstawie n–elementowej próby $X_1, \ldots, X_n$ $(n \geq 50)$ wynosi

$$\left(\frac{S\sqrt{2n}}{\sqrt{2n-3} + u(1-\frac{1}{2}\alpha)}, \frac{S\sqrt{2n}}{\sqrt{2n-3} - u(1-\frac{1}{2}\alpha)} \right)$$

gdzie $u(\alpha)$ oznacza kwantyl rzędu α rozkładu normalnego $N(0,1)$.

Przedział ufności – modele

https://github.com/przem85/statistic_4/blob/master/D05_Z06_confidence_interval.ipynb

Zadanie

W celu sprawdzenia dokładności skrawania za pomocą pewnego urządzenia, dokonano pomiarów wykonanych 50 części i otrzymano $S^2 = 0.00068$. Zakładając, że rozkład błędów wymiarów części jest normalny o nieznanym σ, na poziomie ufności 0.95 wyznaczyć na podstawie danych realizację przedziałów ufności dla odchylenia standardowego σ.

Przedział ufności – Zadania

https://github.com/przem85/statistic_4/blob/master/D05_Z07_confidence_interval.ipynb

Zadanie 1

W centrali telefonicznej dokonano 17 obserwacji długości losowo wybranych rozmów w ciągu jednego dnia i otrzymano (w min): $\bar{X} = 5.48$, $S = 1.2$; na tej podstawie – przy założeniu, że długość rozmów telefonicznych ma rozkład normalny – wyznaczyć 95%-ową realizację przedziału ufności dla wartości przeciętnej długości rozmowy telefonicznej przeprowadzonej za pośrednictwem tej centrali w danym dniu.

Przedział ufności – Zadania

https://github.com/przem85/statistic_4/blob/master/D05_Z08_confidence_interval.ipynb

Zadanie 2

Z grupy robotników pewnego zakładu wykonujących taką samą pracę wybrano w sposób losowy 13 pracowników i dokonano badania pod względem wydajności pracy (w szt./h) uzyskując dane: 21, 12, 11, 15, 9, 10, 17, 8, 16, 13, 12, 9, 18. Na tej podstawie zakładając, że badana cecha ma rozkład normalny, wyznaczyć 95%-ową realizację przedziału ufności dla nieznanej wartości przeciętnej wydajności pracy.

Przedział ufności – Zadania

https://github.com/przem85/statistic_4/blob/master/D05_Z09_confidence_interval.ipynb

Zadanie 3

W losowo wybranej grupie 10 samochodów osobowych przeprowadzono badanie zużycia benzyny na – tej samej dla wszystkich samochodów – trasie długości 100 km. Okazało się, że średnia zużycia benzyny (w l/100 km) dla tej grupy samochodów wynosi $\bar{X} = 8.1$; odchylenie standardowe $0,8$. Zakładając, że badana cecha ma rozkład normalny, wyznaczyć 99%-ową realizację przedziału ufności dla wartości przeciętnej zużycia benzyny przez samochody tej marki na rozpatrywanej trasie.

Przedział ufności – Zadania

https://github.com/przem85/statistic_4/blob/master/D05_Z10_confidence_interval.ipynb

Zadanie 4

Dwunastu tokarzy wykonuje takie same części. Ich średnie wydajności w sztukach na godzinę wynoszą odpowiednio: 4.6, 6.1, 10.3, 9.8, 6.7, 12.3, 14.5, 8.7, 9.0, 7.3, 8.8, 11.2. Znaleźć 98%–ową realizację przedziału ufności dla wariancji liczby sztuk wykonywanych w ciągu godziny przez jednego tokarza.