Wykład IV: Wnioskowanie statystyczne - podstawy

Alicja Janic

Politechnika Wrocławska *alicja.janic@pwr.edu.pl*

4 grudzień 2020

Prosta próba losowa

Definicja

Prostą próbą losową o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, ..., X_n$ określonych na przestrzeni Ω i takich, że każda z nich ma ten sam rozkład

Konkretny ciąg wartości $x_1, x_2, ..., x_n$ próby losowej $X_1, X_2, ..., X_n$ będziemy nazywać **realizacją próby losowej**.
Jak wybierać reprezentatywną grupę elementów populacji?
Wybór n elementów populacji powinien być dokonany w taki sposób, żeby każdy podzbiór populacji, składających się z n elementów miał taką samą szansę wybrania

Średnia w prostej próbie losowej

Definicja

Średnią w prostej próbie losowej $X_1, ..., X_n$ o liczności n nazywamy statystykę

$$\overline{X} = \frac{(X_1 + X_2 + ... + X_n)}{n}$$

Powyższa definicja jest szczególnym przypadkiem ogólnej definicji statystyki dla funkcji T zdefiniowanej w sposób następujący:
$T(X_1, X_2, ..., X_n) = \frac{X_1+X_2+...+X_n}{n}$
Jaki jest cel przeniesienia definicji średniej z poziomu próby wartości na poziom prostej próby losowej? Badając rozkład średniej $\overline{X}$, jesteśmy w stanie przeanalizować jej zachowanie dla potencjalnej próby w terminach rachunku prawdopodobieństwa

Prawo wielkich liczb

Twierdzenie

Niech X będzie zmienną losową o skończonej wariancji, $\sigma_X^2 < \infty$ i niech $X_1, ..., X_n$ będzie prostą próbą losową z rozkładu zmiennej losowej X. Wówczas dla dowolnie małej dodatniej liczby ϵ prawdopodobieństwo

$$P(\overline{X} \in [\mu_X - \epsilon, \mu_X + \epsilon])$$

jest bliskie 1 dla dużych liczności próby n

Prawo wielkich liczb uzasadnia słuszność użycia średniej $\overline{X}$ jako oszacowania wartości średniej μ_X

Wskaźniki rozkładu średniej w prostej próbie losowej

Stwierdzenie

Niech $\overline{X}$ będzie średnią w prostej próbie losowej o liczności n z rozkładu o wartości średniej m i wariancji σ^2. Wówczas

$$E\left(\overline{X}\right) = \mu_{\overline{X}} = m \qquad \text{oraz} \qquad Var\left(\overline{X}\right) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Jeżeli rozkład próby prostej jest normalny z parametrami m i σ to

$$\overline{X} \sim \mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

Przykłady

Przykład 1

Stwierdzono, że wzrost dorosłych Polaków jest cechą o rozkładzie normalnym o wartości średniej $m = 176cm$ i odchyleniu standardowym $\sigma = 6,5cm$. Obliczyć prawdopodobieństwo, że średnia $\overline{X}$ dla prostej próby losowej o liczności 100 różni od prawdziwej wartości m o więcej niż $1,5cm$

Centralne twierdzenie graniczne dla średniej

Twierdzenie

Niech $X_1, ..., X_n$ będzie prostą próbą losową z rozkładu o średniej m i skończonej wariancji σ^2. Wówczas dla dużych liczności próby n rozkład standaryzowanej średniej $\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ jest bliski standardowemu rozkładowi normalnemu $\mathcal{N}(0,1)$ Mianowicie, dla dowolnych liczb a i b, $a \leqslant b$, i zmiennej losowej Z o standardowym rozkładzie normalnym

$$P\left(a \leqslant \frac{\overline{X}-m}{\sigma/\sqrt{n}} \leqslant b\right) \longrightarrow P(a \leqslant Z \leqslant b) = \Phi(b) - \Phi(a)$$

dla n dążących do nieskończoności. Równoważnie, rozkład średniej $\overline{X}$ jest w przybliżeniu równy rozkładowi normalnemu $\mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$

Centralne twierdzenie graniczne dla sumy

Twierdzenie

Niech $X_1, ..., X_n$ *będzie prostą próbą losową z rozkładu o średniej* m *i skończonej wariancji* σ^2. *Wówczas dla dużych liczności próby* n *rozkład standaryzowanej sumy* $\frac{\sum_{i=1}^n X_i - nm}{\sigma\sqrt{n}}$ *jest bliski standardowemu rozkładowi normalnemu* $\mathcal{N}(0,1)$ *Mianowicie, dla dowolnych liczb* a *i* b, $a \leqslant b$, *i zmiennej losowej* Z *o standardowym rozkładzie normalnym*

$$P\left(a \leqslant \frac{\sum_{i=1}^n X_i - nm}{\sigma\sqrt{n}} \leqslant b\right) \longrightarrow P(a \leqslant Z \leqslant b) = \Phi(b) - \Phi(a)$$

dla n *dążących do nieskończoności. Równoważnie, rozkład sumy* $\sum_{i=1}^n X_i$ *jest w przybliżeniu równy rozkładowi normalnemu* $\mathcal{N}\left(nm, \sigma\sqrt{n}\right)$

Poprawka w przybliżeniu normalnym

Poprawka w przybliżeniu normalnym

w Centralnym Twierdzeniu Granicznym nie założyliśmy ciągłości rozkładu zmiennych X_i. W przypadku gdy zmienne X_i w prostej próbie losowej przyjmują wartości całkowite, dokonuje się poprawki w przybliżeniu normalnym, opartej na następującym rozumowaniu. Jeśli każda wartość w próbie jest liczbą całkowitą, to suma $\sum_{i=1}^{n} X_i$ przyjmuje tylko wartości bedące liczbami całkowitymi i dlatego dla dowolnych liczb całkowitych a i b możemy napisać

$$P(a \leqslant \sum_{i=1}^{n} X_i \leqslant b) = P(a - 0.5 \leqslant \sum_{i=1}^{n} X_i \leqslant b + 0.5)$$

Przykłady

Przykład 2

Komputer dodaje 1200 liczb rzeczywistych i każdą zaokrągla do najbliższej liczby całkowitej. Błędy zaokrągleń są niezależne i mają rozkład jednostajny na przedziale $[-0.5, 0.5]$. Obliczyć prawdopodobieństwo, że błąd w obliczeniu sumy przekroczy 10

Rozwiązanie

Wprowadzając oznaczenia: X_i - błąd zaokrąglenia i-tej liczby otrzymujemy:

$$E(X_i) = 0 = m, \quad Var(X_i) = \frac{1}{12} = \sigma^2 \quad \text{oraz}$$
$$Var\left(\sum_{i=1}^{1200} X_i\right) = 1200 \cdot \frac{1}{12} = 100 = n\sigma^2$$

Zatem

$$P\left(\sum_{i=1}^{1200} X_i > 10\right) = P\left(\frac{\sum_{i=1}^{1200} X_i - 0}{10} > \frac{10-0}{10}\right) \approx P(Z > 1)$$
$$= 1 - \Phi(1) = 0.1587$$

Przykłady

Przykład 3

Czas pracy diody ma rozkład wykładniczy o wartości oczekiwanej $m = 900h$. Zgromadzono zapas 100 diod. Obliczyć prawdopodobieństwo, że wystarczy ich na 100.000 godzin pracy, jeżeli każdą diodę włączamy natychmiast po uszkodzeniu poprzedniej. Ile diod trzeba mieć w zapasie, by z prawdopodobieństwem większym niż 0,99 wystarczyło ich na 99.000 godzin pracy urządzenia.

Rozwiązanie

Przykłady

Przykład 4

Rzucamy 30 razy kostką. Oszacujmy prawdopodobieństwo, że suma wyrzuconych oczek jest liczbą między 100 a 110

Rozwiązanie

Częstość

Częstość

Częstością w prostej próbie losowej nazywamy statystykę

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$$

gdzie $X_1, X_2, ..., X_n$ jest prostą próbą losową z rozkładu dwupunktowego o wartościach 0 i 1. Statystykę $\hat{p}$ obliczoną dla konkretnych wartości w próbie nazywamy wartością częstości

Stwierdzenie

Częstość $\hat{p}$ pomnożona przez liczność próby n ma rozkład dwumianowy $\mathcal{B}in(n, p)$. Ponadto

$$E(\hat{p}) = p \quad \text{oraz} \quad Var(\hat{p}) = \frac{p(1-p)}{n}$$

Twierdzenie Moivre'a-Laplace'a

Twierdzenie

Dla dowolnych rzeczywistych a i b

$$P\left(a \leqslant \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leqslant b\right) \longrightarrow \Phi(b)-\Phi(a)$$

gdy n dąży do nieskończoności
Analogicznie, jeśli $S_n = n\hat{p}$

$$P\left(a \leqslant \frac{S_n-np}{\sqrt{np(1-p)}} \leqslant b\right) \longrightarrow \Phi(b)-\Phi(a)$$

gdy n dąży do nieskończoności

Przybliżenie normalne uznajemy za zadowalające, gdy $np \geqslant 5$ i $n(1-p) \geqslant 5$

Przykłady

Przykład 5

W populacji dorosłych Polaków 39% ma kłopoty ze snem. Oszacujmy prawdopodobieństwo, że w prostej próbie losowej dorosłych Polaków o liczności 100, częstość osób mających kłopoty ze snem nie przekroczy 0,33

Rozwiązanie