Centralne twierdzenie graniczne

Dariusz Uciński

Instytut Sterowania i Systemów Informatycznych
Universytet Zielonogórski

Wykład 4

Ważne uzupełnienie

Dwuwymiarowy rozkład normalny $\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

Można udowodnić, że $X \sim \mathcal{N}(\mu_X, \sigma_X)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$ oraz $\rho_{XY} = \rho$.

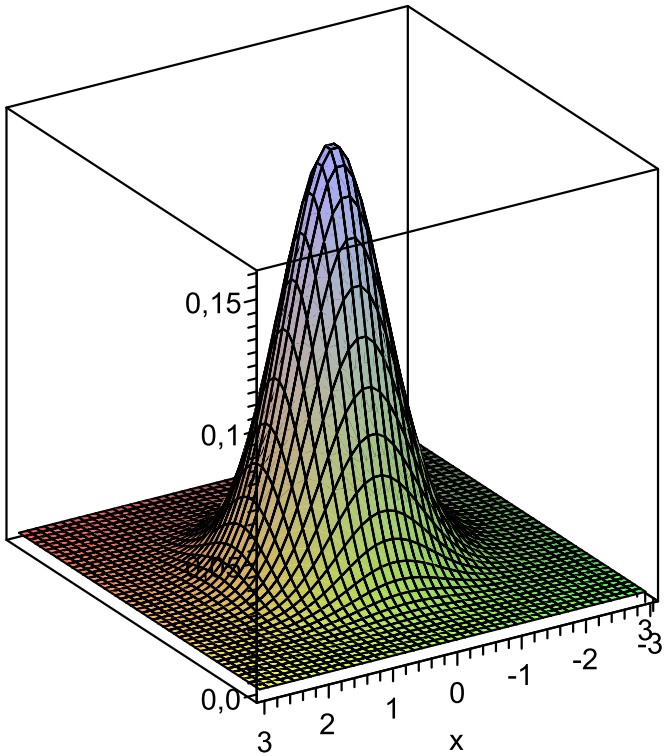
Ważne uzupełnienie

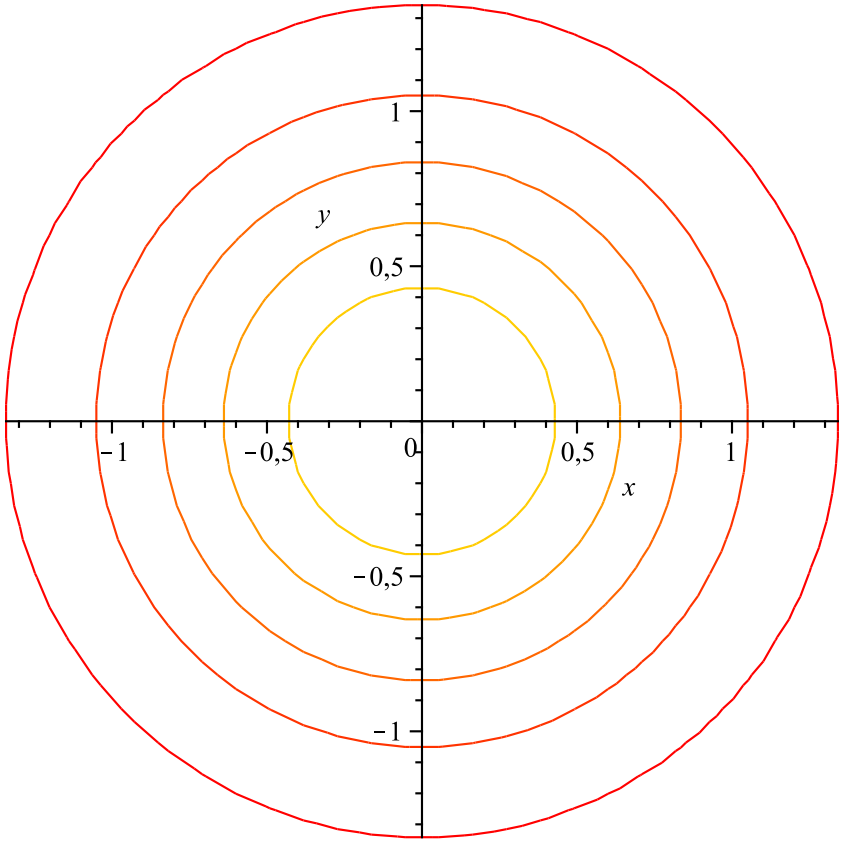
Dwuwymiarowy rozkład normalny $\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

Można udowodnić, że $X \sim \mathcal{N}(\mu_X, \sigma_X)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$ oraz $\rho_{XY} = \rho$.

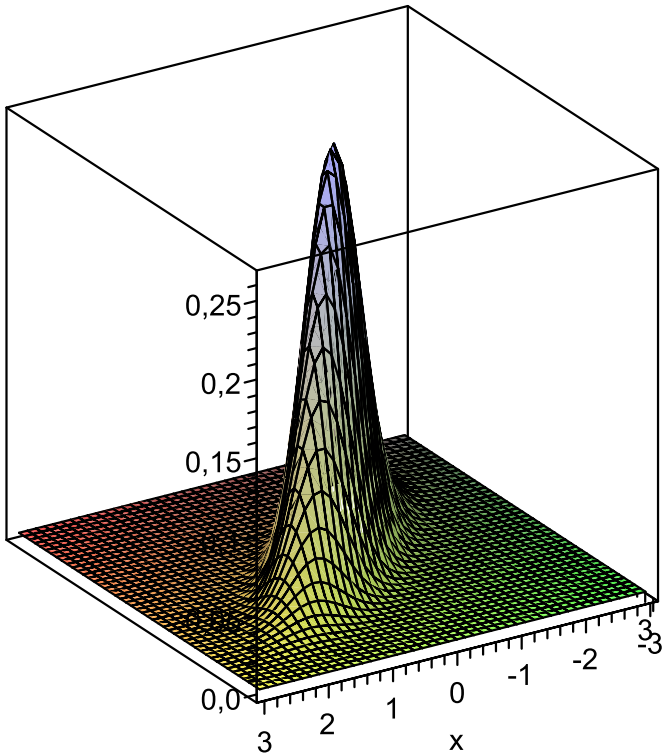
$\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$

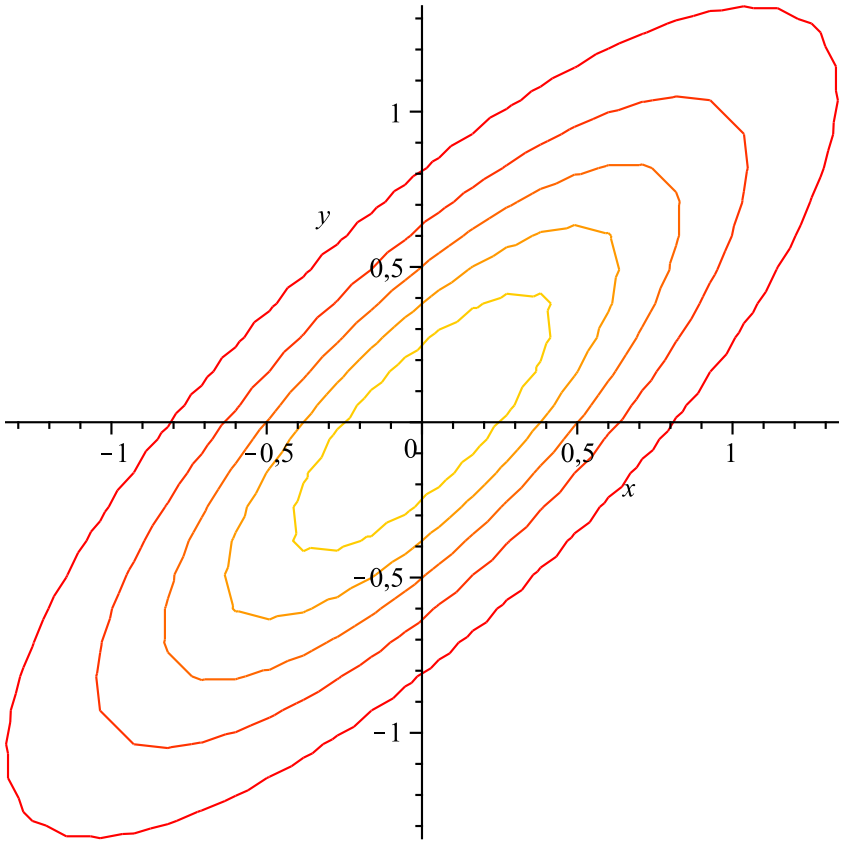




$$\sigma_X = \sigma_Y = 1,\ \rho = 0$$

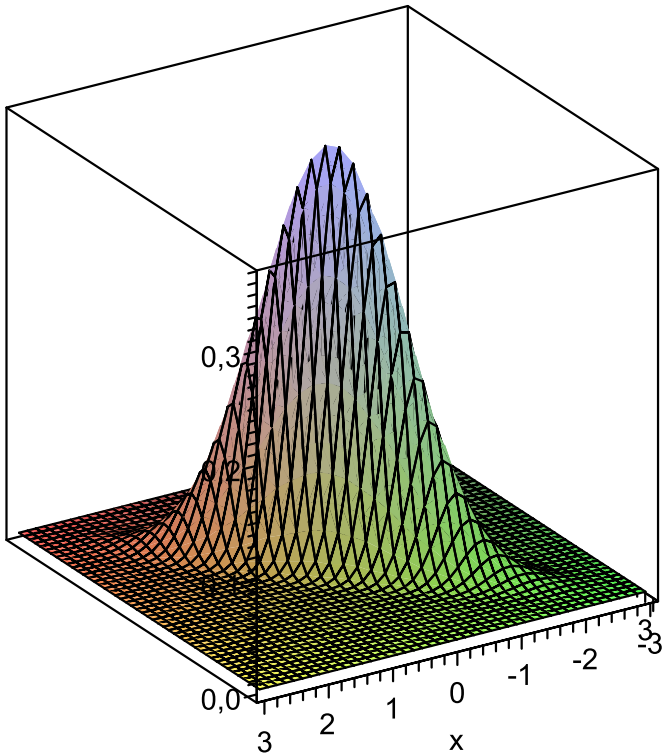
$\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$

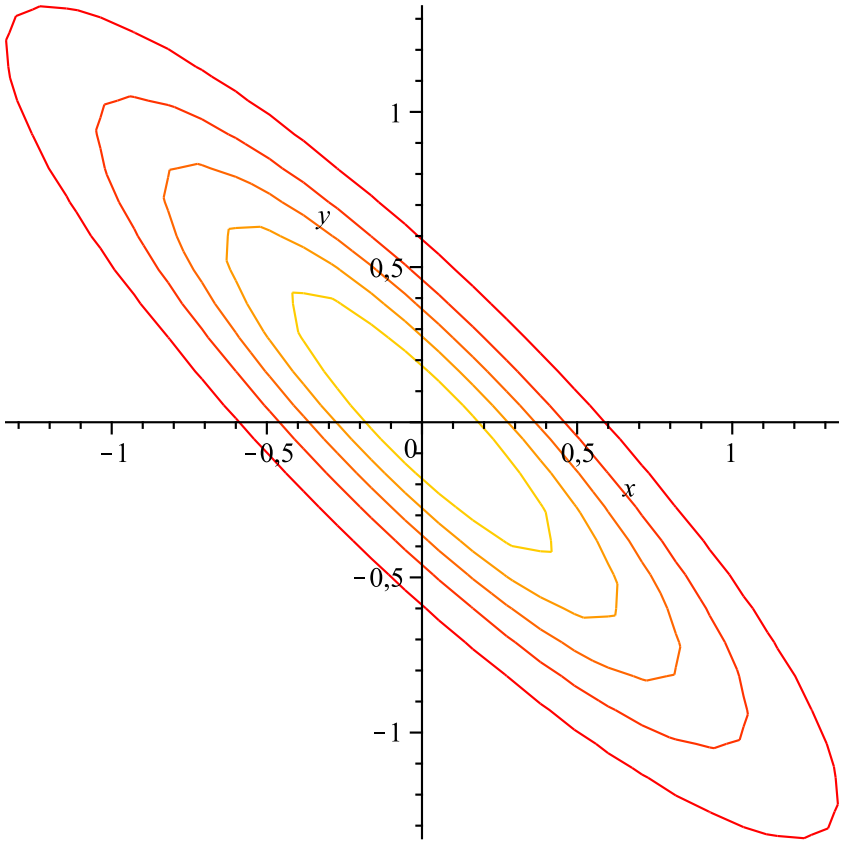




$$\sigma_X = \sigma_Y = 1,\ \rho = 0.8$$

$\mathcal{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$





$$\sigma_X = \sigma_Y = 1,\ \rho = -0.9$$

Dwuwymiarowy rozkład normalny

Twierdzenie

Dla X i Y mających dwuwumiarowy rozkład normalny zachodzi

$$X,\ Y \text{ — niezależne} \iff \rho = 0$$

Twierdzenie

Para zmiennych losowych X i Y ma dwuwymiarowy rozkład normalny wtedy i tylko wtedy, gdy każda kombinacja liniowa $aX + bY$ (a, b — dowolne stałe) ma rozkład normalny.

Dwuwymiarowy rozkład normalny

Twierdzenie

Dla X i Y mających dwuwumiarowy rozkład normalny zachodzi

$$X,\ Y \text{ — niezależne} \iff \rho = 0$$

Twierdzenie

Para zmiennych losowych X i Y ma dwuwymiarowy rozkład normalny wtedy i tylko wtedy, gdy każda kombinacja liniowa $aX + bY$ (a, b — dowolne stałe) ma rozkład normalny.

Wnioskowanie statystyczne — podstawy

Definicja

Prostą próbą losową o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ określonych na tej samej przestrzeni zdarzeń elementarnych $\mathcal{S}$ i takich, że każda ma ten sam rozkład.

Pytanie: Co to jest próba reprezentatywna i jak ją wybierać?

Statystyka — zmienna losowa $T(X_1, \ldots, X_n)$ będąca funkcją próby

Wnioskowanie statystyczne — podstawy

Definicja

Prostą próbą losową o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ określonych na tej samej przestrzeni zdarzeń elementarnych $\mathcal{S}$ i takich, że każda ma ten sam rozkład.

Pytanie: Co to jest próba reprezentatywna i jak ją wybierać?

Statystyka — zmienna losowa $T(X_1, \ldots, X_n)$ będąca funkcją próby

Wnioskowanie statystyczne — podstawy

Definicja

Prostą próbą losową o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ określonych na tej samej przestrzeni zdarzeń elementarnych $\mathcal{S}$ i takich, że każda ma ten sam rozkład.

Pytanie: Co to jest próba reprezentatywna i jak ją wybierać?

Statystyka — zmienna losowa $T(X_1, \ldots, X_n)$ będąca funkcją próby

Wnioskowanie statystyczne — podstawy

Definicja

Średnią w proste próbie losowej $X_1, \ldots, X_n$ o liczności n nazywamy statystykę

$$\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

Twierdzenie

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

$$P(\bar{X} \text{ należy do } [\mu_X - \varepsilon, \mu_X + \varepsilon])$$

jest bliskie 1 dla dużych liczności próby n.

Wnioskowanie statystyczne — podstawy

Definicja

Średnią w proste próbie losowej $X_1, \ldots, X_n$ o liczności n nazywamy statystykę

$$\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

Twierdzenie

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

$$P(\bar{X} \text{ należy do } [\mu_X - \varepsilon, \mu_X + \varepsilon])$$

jest bliskie 1 dla dużych liczności próby n.

Wnioskowanie statystyczne — podstawy

Stwierdzenie

Zachodzi

$$\mu_{\bar{X}} = \mu, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Przykład

Wzrost dorosłych Polaków ma rozkład $\mathcal{N}(176\text{ cm}, 6.5\text{ cm})$. Obliczyć prawdopodobieństwo, że średnia $\bar{X}$ dla prostej próby losowej o liczności 100 różni się od prawdziwej wartości $\mu = 176$ cm o więcej niż 1.5 cm.

Wnioskowanie statystyczne — podstawy

Stwierdzenie

Zachodzi

$$\mu_{\bar{X}} = \mu, \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Przykład

Wzrost dorosłych Polaków ma rozkład $\mathcal{N}(176\text{ cm}, 6.5\text{ cm})$. Obliczyć prawdopodobieństwo, że średnia $\bar{X}$ dla prostej próby losowej o liczności 100 różni się od prawdziwej wartości $\mu = 176$ cm o więcej niż 1.5 cm.

Przykład (c.d.)

Średnia $\bar{X}$ ma rozkład normalny jako kombinacja liniowa zmiennych $X_1, \ldots, X_n$ o rozkładzie normalnym. Co więcej, $\mu_{\bar{X}} = 176$ oraz $\sigma_{\bar{X}} = 6.5/\sqrt{100} = 0.65$.

$$\begin{aligned} P(|\bar{X} - \mu_{\bar{X}}| > 1.5) &= 1 - P(|\bar{X} - \mu_{\bar{X}}| \leqslant 1.5) \\ &= 1 - P\left(-\frac{1.5}{0.65} \leqslant \frac{\bar{X} - \mu_{\bar{X}}}{0.65} \leqslant \frac{1.5}{0.65}\right) \\ &= 1 - P(-2.31 \leqslant Z \leqslant 2.31) \\ &= 1 - [\Phi(2.31) - \Phi(-2.31)] = 0.0208 \end{aligned}$$

Pytanie: Jak postępować bez informacji o rozkładzie obserwacji?

Przykład (c.d.)

Średnia $\bar{X}$ ma rozkład normalny jako kombinacja liniowa zmiennych $X_1, \ldots, X_n$ o rozkładzie normalnym. Co więcej, $\mu_{\bar{X}} = 176$ oraz $\sigma_{\bar{X}} = 6.5/\sqrt{100} = 0.65$.

$$\begin{aligned} P(|\bar{X} - \mu_{\bar{X}}| > 1.5) &= 1 - P(|\bar{X} - \mu_{\bar{X}}| \leqslant 1.5) \\ &= 1 - P\left(-\frac{1.5}{0.65} \leqslant \frac{\bar{X} - \mu_{\bar{X}}}{0.65} \leqslant \frac{1.5}{0.65}\right) \\ &= 1 - P(-2.31 \leqslant Z \leqslant 2.31) \\ &= 1 - \left[\Phi(2.31) - \Phi(-2.31)\right] = 0.0208 \end{aligned}$$

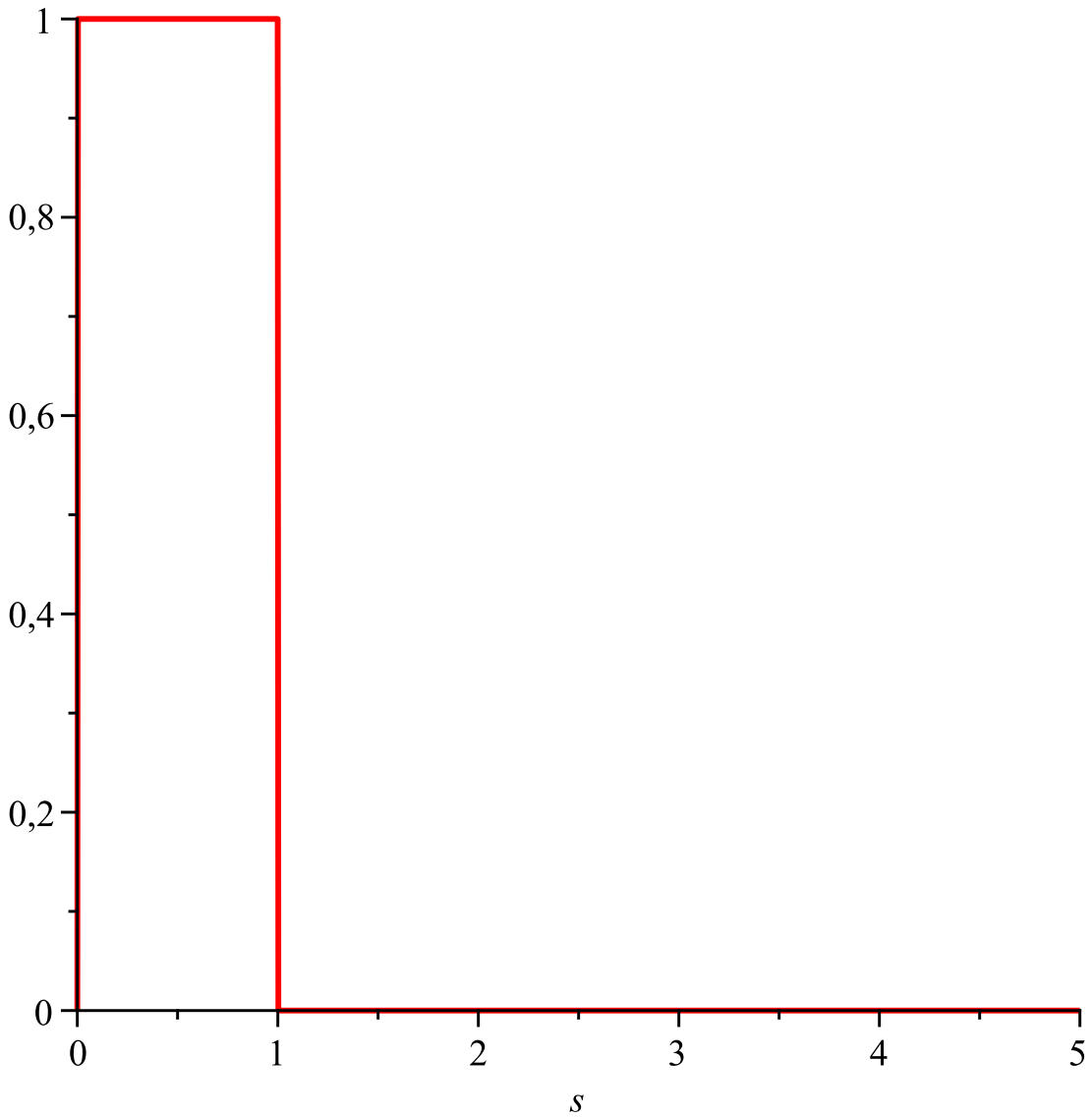
Pytanie: Jak postępować bez informacji o rozkładzie obserwacji?

Przykład (c.d.)

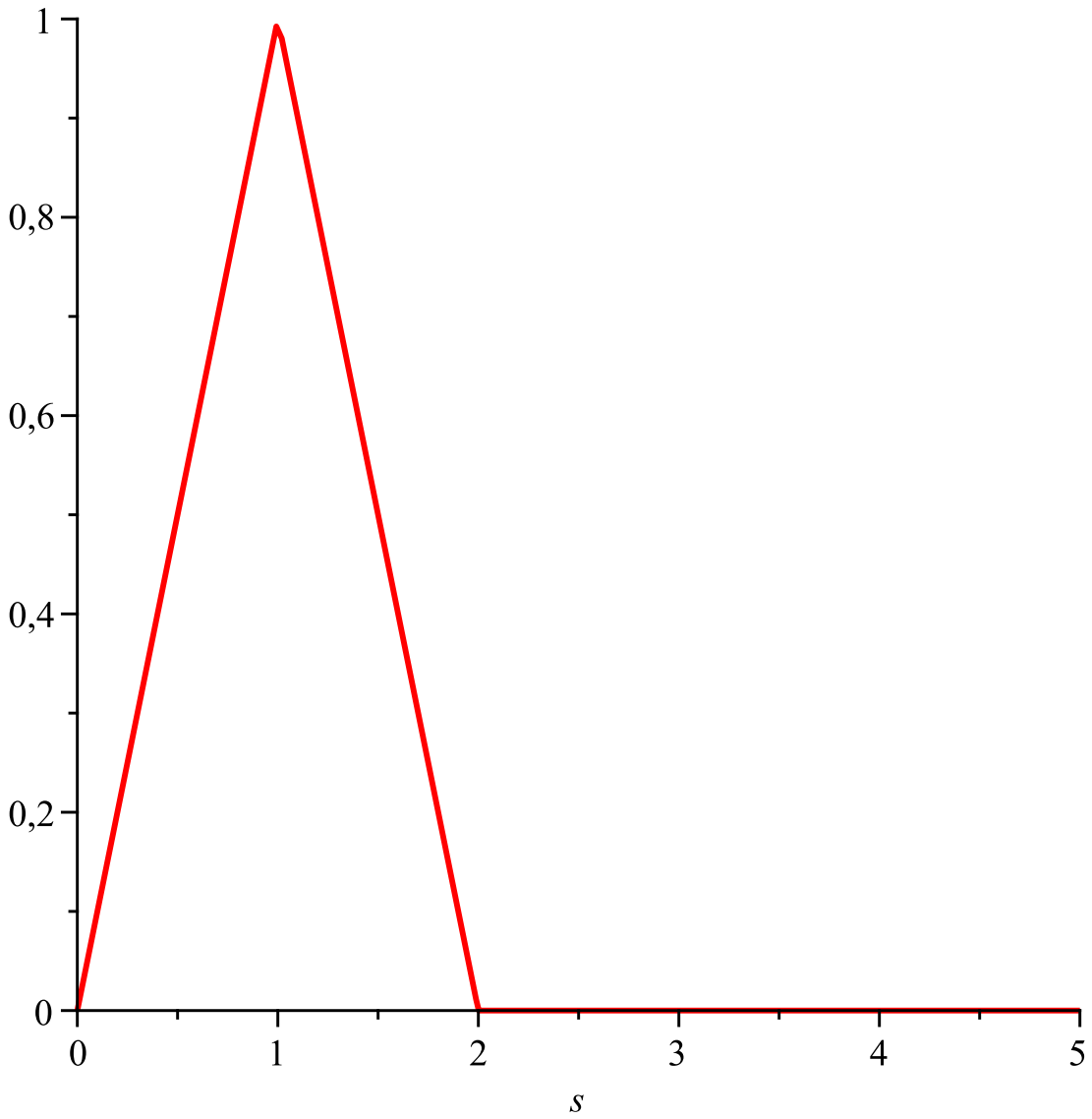
Średnia $\bar{X}$ ma rozkład normalny jako kombinacja liniowa zmiennych $X_1, \ldots, X_n$ o rozkładzie normalnym. Co więcej, $\mu_{\bar{X}} = 176$ oraz $\sigma_{\bar{X}} = 6.5/\sqrt{100} = 0.65$.

$$\begin{aligned} P(|\bar{X} - \mu_{\bar{X}}| > 1.5) &= 1 - P(|\bar{X} - \mu_{\bar{X}}| \leqslant 1.5) \\ &= 1 - P\left(-\frac{1.5}{0.65} \leqslant \frac{\bar{X} - \mu_{\bar{X}}}{0.65} \leqslant \frac{1.5}{0.65}\right) \\ &= 1 - P(-2.31 \leqslant Z \leqslant 2.31) \\ &= 1 - \left[\Phi(2.31) - \Phi(-2.31)\right] = 0.0208 \end{aligned}$$

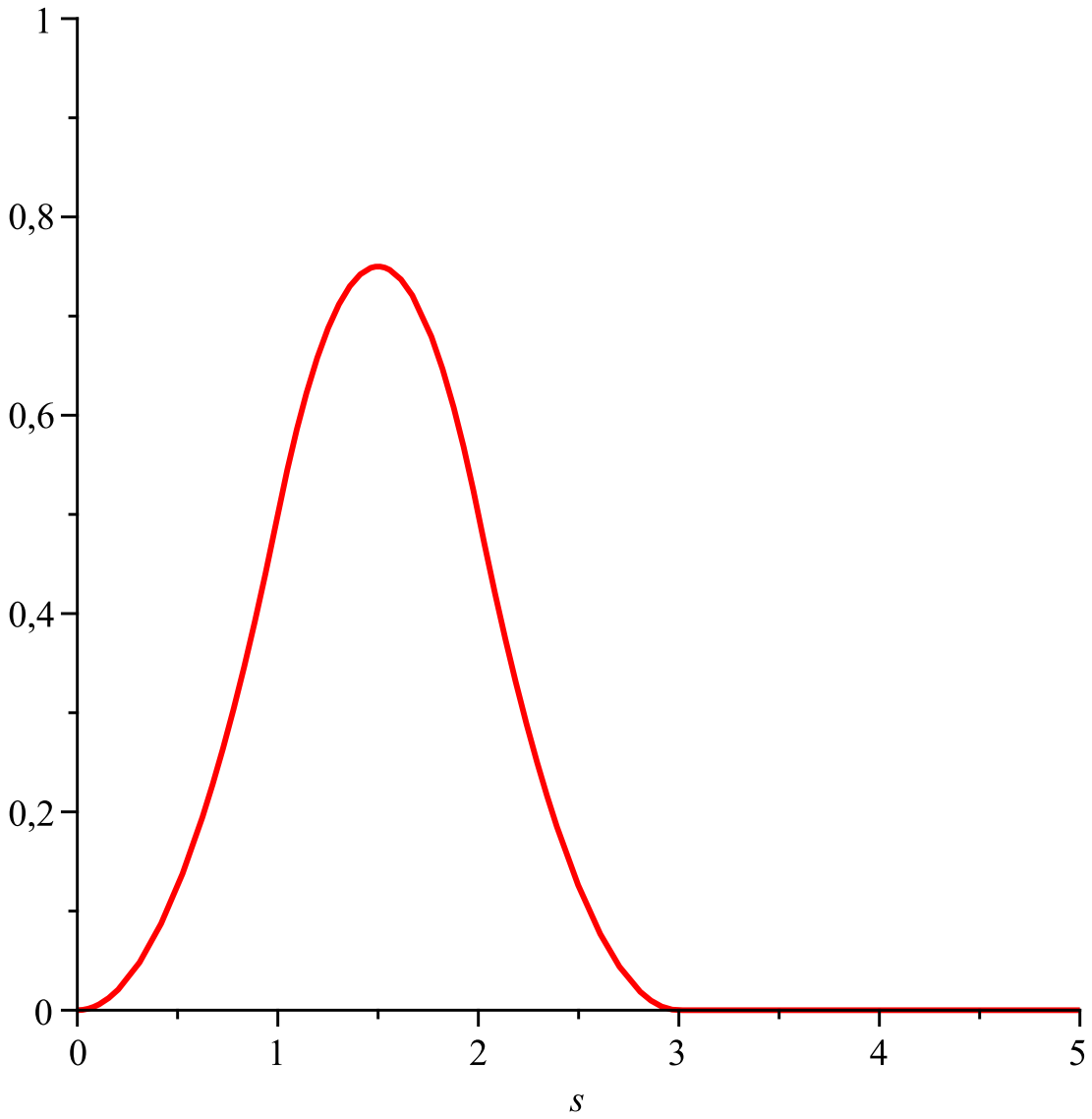
Pytanie: Jak postępować bez informacji o rozkładzie obserwacji?



Rozkład $S_1 = X_1$

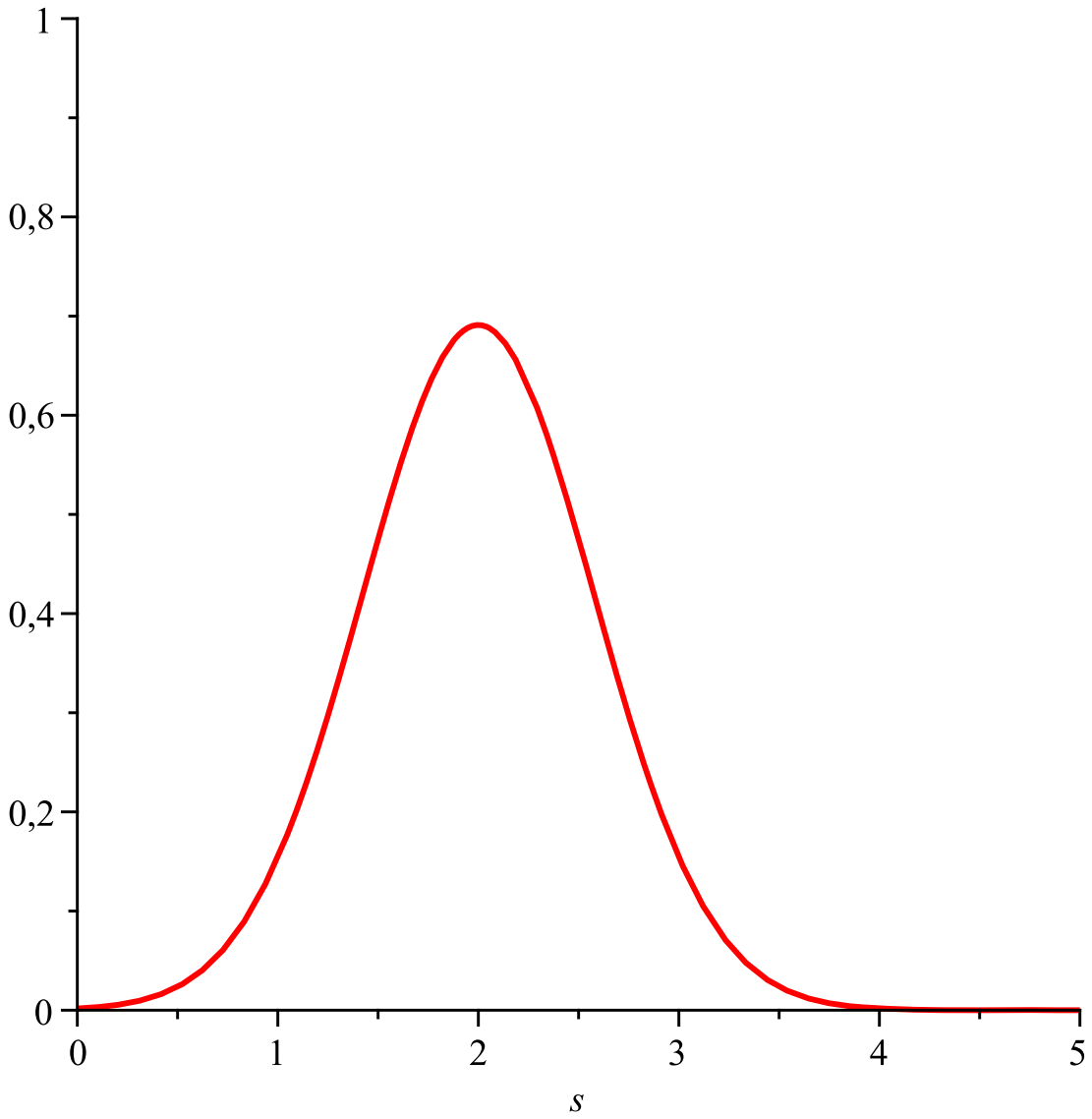


Rozkład $S_2 = X_1 + X_2$

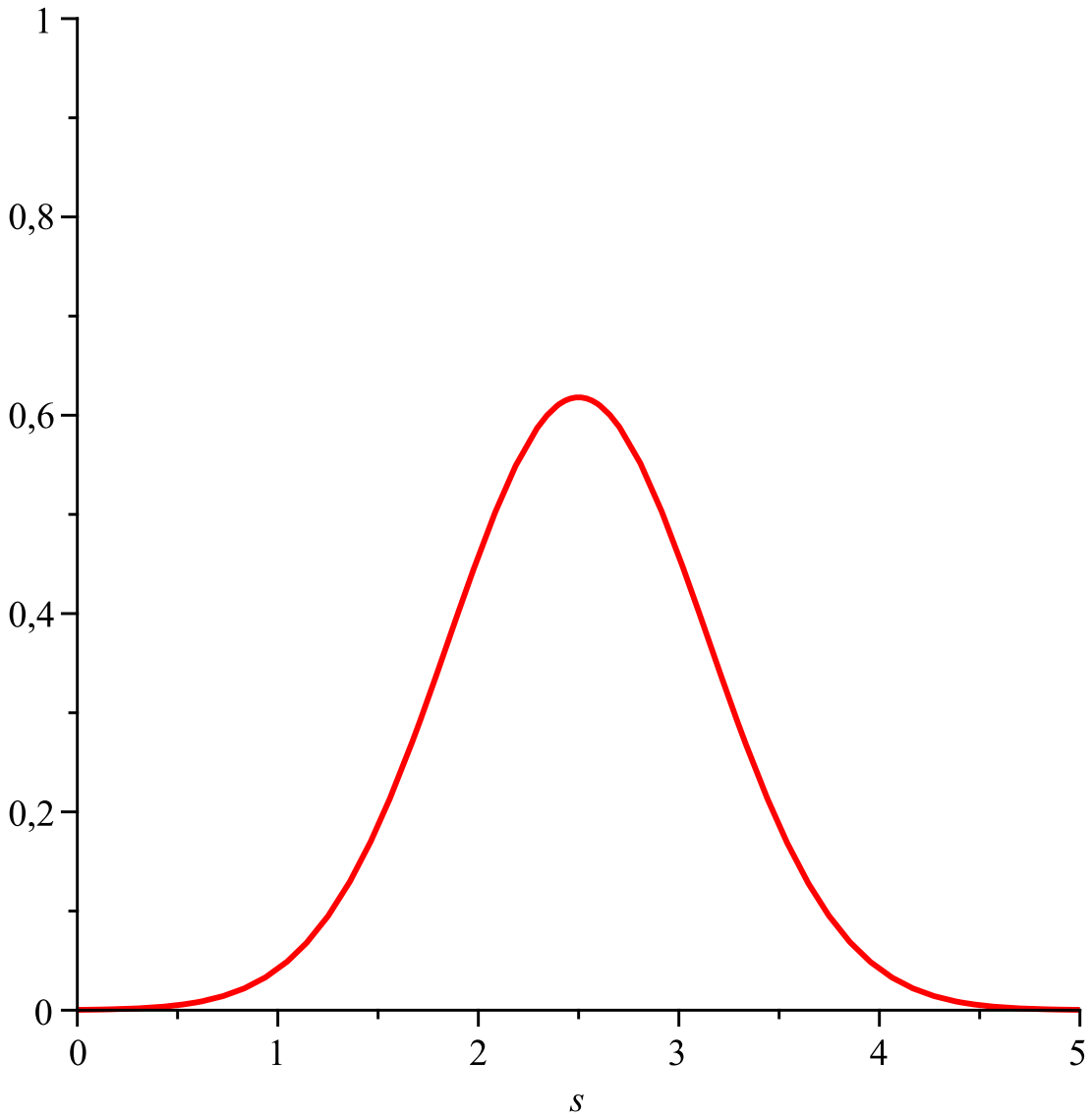


Rozkład $S_2 = X_1 + X_2 + X_3$

Wnioskowanie statystyczne — podstawy



Rozkład $S_2 = X_1 + X_2 + X_3 + X_4$



Rozkład $S_2 = X_1 + X_2 + X_3 + X_4 + X_5$

Twierdzenie

CENTRALNE TWIERDZENIE GRANICZNE. *Niech $X_1, \ldots, X_n$ będzie prostą próbą losową z rozkładu o średniej μ i skończonej wariancji σ^2. Wówczas dla dużych liczności próby n rozkład sumy $S_n := X_1 + \cdots + X_n$ jest bliski $\mathcal{N}(n\mu, \sqrt{n}\sigma)$.*

Oznacza to, że rozkład standaryzowanej średniej $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ jest bliski standaryzowanemu rozkładowi normalnemu $\mathcal{N}(0,1)$. Ściślej, dla dowolnych liczb a i b ($a \leqslant b$), i zmiennej Z o standardowym rozkładzie normalnym

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

dla $n \to \infty$. Równoważnie, rozkład średniej $\bar{X}$ jest w przybliżeniu równy rozkładowi $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

Twierdzenie

CENTRALNE TWIERDZENIE GRANICZNE. *Niech $X_1, \ldots, X_n$ będzie prostą próbą losową z rozkładu o średniej μ i skończonej wariancji σ^2. Wówczas dla dużych liczności próby n rozkład sumy $S_n := X_1 + \cdots + X_n$ jest bliski $\mathcal{N}(n\mu, \sqrt{n}\sigma)$. Oznacza to, że rozkład standaryzowanej średniej $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ jest bliski standaryzowanemu rozkładowi normalnemu $\mathcal{N}(0,1)$. Ściślej, dla dowolnych liczb a i b ($a \leqslant b$), i zmiennej Z o standardowym rozkładzie normalnym*

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

dla $n \to \infty$. Równoważnie, rozkład średniej $\bar{X}$ jest w przybliżeniu równy rozkładowi $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

Twierdzenie

CENTRALNE TWIERDZENIE GRANICZNE. *Niech $X_1, \ldots, X_n$ będzie prostą próbą losową z rozkładu o średniej μ i skończonej wariancji σ^2. Wówczas dla dużych liczności próby n rozkład sumy $S_n := X_1 + \cdots + X_n$ jest bliski $\mathcal{N}(n\mu, \sqrt{n}\sigma)$. Oznacza to, że rozkład standaryzowanej średniej $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ jest bliski standaryzowanemu rozkładowi normalnemu $\mathcal{N}(0,1)$. Ściślej, dla dowolnych liczb a i b ($a \leqslant b$), i zmiennej Z o standardowym rozkładzie normalnym*

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

dla $n \to \infty$. Równoważnie, rozkład średniej $\bar{X}$ jest w przybliżeniu równy rozkładowi $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

Twierdzenie

CENTRALNE TWIERDZENIE GRANICZNE. *Niech $X_1, \ldots, X_n$ będzie prostą próbą losową z rozkładu o średniej μ i skończonej wariancji σ^2. Wówczas dla dużych liczności próby n rozkład sumy $S_n := X_1 + \cdots + X_n$ jest bliski $\mathcal{N}(n\mu, \sqrt{n}\sigma)$. Oznacza to, że rozkład standaryzowanej średniej $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ jest bliski standaryzowanemu rozkładowi normalnemu $\mathcal{N}(0,1)$. Ściślej, dla dowolnych liczb a i b ($a \leqslant b$), i zmiennej Z o standardowym rozkładzie normalnym*

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

dla $n \to \infty$. Równoważnie, rozkład średniej $\bar{X}$ jest w przybliżeniu równy rozkładowi $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

Przykład

Przyjmijmy, że rozkład codziennego dojazdu do pracy jest w przybliżeniu rozkładem jednostajnym na odcinku [0.5 godz., 1 godz.] i że czasy dojazdów w różne dni są niezależne. Ile w przybliżeniu wynosi prawdopodobieństwo zdarzenia, że średni dzienny dojazd w ciągu 30 dni przekroczy 48 minut?

Niech X_i — czas dojazdu w i-tym dniu, $i = 1, \ldots, 30$. Mamy

$$\mu_{X_i} = \frac{0.5 + 1}{2} = \frac{3}{4}, \quad \sigma^2_{X_i} = \frac{(1 - 0.5)^2}{12} = \frac{1}{48}$$

Przykład

Przyjmijmy, że rozkład codziennego dojazdu do pracy jest w przybliżeniu rozkładem jednostajnym na odcinku [0.5 godz., 1 godz.] i że czasy dojazdów w różne dni są niezależne. Ile w przybliżeniu wynosi prawdopodobieństwo zdarzenia, że średni dzienny dojazd w ciągu 30 dni przekroczy 48 minut?

Niech X_i — czas dojazdu w i-tym dniu, $i = 1, \ldots, 30$. Mamy

$$\mu_{X_i} = \frac{0.5 + 1}{2} = \frac{3}{4}, \quad \sigma^2_{X_i} = \frac{(1 - 0.5)^2}{12} = \frac{1}{48}$$

Przykład (c.d.)

Interesuje nas prawdopodobieństwo $P(\bar{X} > 0.8)$ (48 minut to 0.8 godz.). Z CTG wiemy, że średnia $\bar{X}$ ma w przybliżeniu rozkład $\mathcal{N}\left(\frac{3}{4}, \sqrt{\frac{1}{30 \cdot 48}}\right)$.

W konsekwencji,

$$P(\bar{X} > 0.8) = P\left(\frac{\bar{X} - 3/4}{\sqrt{\frac{1}{30 \cdot 48}}} > \frac{0.8 - 3/4}{\sqrt{\frac{1}{30 \cdot 48}}}\right)$$
$$\approx P(Z > 1.89) = 0.03$$

Przykład (c.d.)

Interesuje nas prawdopodobieństwo $P(\bar{X} > 0.8)$ (48 minut to 0.8 godz.). Z CTG wiemy, że średnia $\bar{X}$ ma w przybliżeniu rozkład $\mathcal{N}\left(\frac{3}{4}, \sqrt{\frac{1}{30 \cdot 48}}\right)$.

W konsekwencji,

$$P(\bar{X} > 0.8) = P\left(\frac{\bar{X} - 3/4}{\sqrt{\frac{1}{30 \cdot 48}}} > \frac{0.8 - 3/4}{\sqrt{\frac{1}{30 \cdot 48}}}\right)$$
$$\approx P(Z > 1.89) = 0.03$$

Wnioskowanie statystyczne — podstawy

Rozważmy sytuację, gdy zmienna losowa X, z rozkładu której pochodzi rozważana próba losowa, przyjmuje tylko dwie wartości: 1 i 0. Oznaczmy

$$p = P(x = 1), \quad q = 1 - p = P(X = 0)$$

Interpretacja: Część jednostek w populacji ma pewną własność (dla nich $X = 1$), a częsć nie (dla nich $X = 0$). Liczbę p nazywa się proporcją.

Przypomnijmy, że $\mu_X = p$ oraz $\sigma_X^2 = pq = p(1 - p)$.

Wnioskowanie statystyczne — podstawy

Rozważmy sytuację, gdy zmienna losowa X, z rozkładu której pochodzi rozważana próba losowa, przyjmuje tylko dwie wartości: 1 i 0. Oznaczmy

$$p = P(x = 1), \quad q = 1 - p = P(X = 0)$$

Interpretacja: Część jednostek w populacji ma pewną własność (dla nich $X = 1$), a częsć nie (dla nich $X = 0$). Liczbę p nazywa się proporcją.

Przypomnijmy, że $\mu_X = p$ oraz $\sigma_X^2 = pq = p(1 - p)$.

Wnioskowanie statystyczne — podstawy

Rozważmy sytuację, gdy zmienna losowa X, z rozkładu której pochodzi rozważana próba losowa, przyjmuje tylko dwie wartości: 1 i 0. Oznaczmy

$$p = P(x = 1), \quad q = 1 - p = P(X = 0)$$

Interpretacja: Część jednostek w populacji ma pewną własność (dla nich $X = 1$), a częsć nie (dla nich $X = 0$). Liczbę p nazywa się proporcją.

Przypomnijmy, że $\mu_X = p$ oraz $\sigma_X^2 = pq = p(1 - p)$.

Wnioskowanie statystyczne — podstawy

Przykłady:

- populacja elementów wyprodukowanych w pewnej w danym miesiącu, z których pewna proporcja p jest wadliwa;
- populacja Polaków, z których pewna proporcja p jest leworęczna;
- populacja studentów szkół wyższych w Polsce, z których pewna proporcja p nie zdała przynajmniej jednego egzaminu w poprzednim semestrze.

Definicja

Częstością występowania w prostej próbie losowej nazywamy statystykę

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

gdzie: $X_1, \ldots, X_n$ — prosta próba losowa z rozkładu dwupunktowego o wartościach 0 i 1.

Wnioskowanie statystyczne — podstawy

Przykłady:

- populacja elementów wyprodukowanych w pewnej w danym miesiącu, z których pewna proporcja p jest wadliwa;
- populacja Polaków, z których pewna proporcja p jest leworęczna;
- populacja studentów szkół wyższych w Polsce, z których pewna proporcja p nie zdała przynajmniej jednego egzaminu w poprzednim semestrze.

Definicja

Częstością występowania w prostej próbie losowej nazywamy statystykę

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

gdzie: $X_1, \ldots, X_n$ — prosta próba losowa z rozkładu dwupunktowego o wartościach 0 i 1.

Twierdzenie

Dla dowolnych rzeczywistych a i b

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

gdy $n \to \infty$.

Do samodzielnego opracowania:

- nierówność Czebyszewa
- poprawka w przybliżeniu normalnym
- metody zbierania danych

Wnioskowanie statystyczne — podstawy

Twierdzenie

Dla dowolnych rzeczywistych a i b

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

gdy $n \to \infty$.

Do samodzielnego opracowania:

- nierówność Czebyszewa
- poprawka w przybliżeniu normalnym
- metody zbierania danych