Statystyka i Analiza Danych

W3: Statystyki i ich rozkłady

dr hab. inż. Katarzyna Filipiak, prof. PP

Instytut Matematyki
Politechnika Poznańska

2023/2024

Próba

Cel statystyki: wyciągnięcie wnioskow o populacji na podstawie zbioru obserwowanych danych (próby)

Próba – zbiór znanych, mierzalnych jednostek reprezentujących populację posiadającą badaną cechę

Powody próbkowania:

- ekonomiczne
- aktualizacja wiedzy
- duża populacja
- destrukcyjny charakter obserwacji
- niedostępność obserwacji

Próba prosta

Obserwację przed jej pobraniem modelujemy jako zmienną losową X o rozkładzie $f(x)$ - rozkładzie populacji.

Próba (losowa) prosta o liczebności n – zbiór n niezależnych zm. losowych $X_1, X_2, \ldots, X_n$ o takim samym rozkładzie $f(x)$ jak interesująca zm. losowa X w populacji.

$X_1, X_2, \ldots, X_n$ – zm. losowe reprezentujące nieznane pomiary, które w procesie losowania próby zamienią się w pierwszą, drugą, $\ldots, n$tą obserwację

$x_1, x_2, \ldots, x_n$ – obserwacje (realizacje zm. losowych $X_1, X_2, \ldots, X_n$)

Przykład 1

Badacz chce zweryfikować, ile rowerów przypada na rodzinę w pewnym mieście liczącym 50000 rodzin. Do badania wybranych zostało 100 rodzin, które następnie zapytano o liczbę posiadanych rowerów.
Niech X będzie zmienną losową zliczającą liczbę rowerów w rodzinie. Zdefiniuj populację, próbę, zmienną losową oraz podaj przykład obserwacji.

Populacja:

Próba:

Zmienna losowa:

Obserwacje (przykład):

Statystyki

Statystyka – dowolna funkcja zm. losowych $X_1, X_2, \ldots, X_n$ stanowiących próbę, nie zawierająca nieznanych parametrów, np.

- średnia próby,
- średnie odchylenie kwadratowe (wariancja) próby,
- odchylenie standardowe próby,
- wskaźnik struktury (proporcja, prawdopodobieństwo sukcesu).

Statystyka = funkcja zmiennych losowych = ZMIENNA LOSOWA!!!

Statystyka jest zm. losową, a więc posiada swój rozkład!

Przykład 2 (rozkład średniej z próby)

Niech dane w tabeli prezentują wyniki egzaminu z fizyki populacji studentów mechaniki (tak mała populacja w praktyce nie występuje, jednakże miniaturowa skala tego przykładu pozwoli zaprezentować rozkład średniej).

Imię	Daniel	Robert	Anna	Iwona	Jan
Ocena	3	2	3	4	2

Wybrana zostanie próba dwóch studentów z populacji. Wyznacz rozkład średniej z próby. X - ocena studenta, X_1, X_2 – oceny dwóch wybranych studentów

Populacja (zwykle nieznana):

Imię	D	R	A	I	J
Ocena	3	2	3	4	2

Przykład 2 – c.d.

Rozkład średniej z próby (1)

Obserwowana cecha w populacji – zmienna losowa $X \sim N(\mu, \sigma)$,
σ - znane

Próba: $X_1, X_2, \ldots, X_n, \qquad X_i \sim N(\mu, \sigma), \quad \sigma$ - znane

Średnia z próby:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ \sim \ N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Suma z próby:

$$T = X_1 + X_2 + \ldots + X_n \ \sim \ N(n \cdot \mu, \sqrt{n} \cdot \sigma)$$

Rozkład średniej z próby (2)

Obserwowana cecha w populacji – zmienna losowa X (dowolny rozkład)

Próba (duża): $X_1, X_2, \ldots, X_n \quad (n > 30)$

Centralne Twierdzenie Graniczne

W losowym próbkowaniu z dowolnej populacji o wartości oczekiwanej μ i odchyleniu standardowym σ rozkład $\overline{X}$ przy dużym n jest w przybliżeniu rozkładem normalnym z wartością oczekiwaną μ i odchyleniem standardowym $\sigma/\sqrt{n}$, tzn.

$$\overline{X} \underset{\text{app}}{\sim} N(\mu, \tfrac{\sigma}{\sqrt{n}}).$$

Suma z próby: $T = X_1 + X_2 + \ldots + X_n \underset{\text{app}}{\sim} N(n \cdot \mu, \sqrt{n} \cdot \sigma)$

Rozkład średniej z próby (3)

Obserwowana cecha w populacji – zmienna losowa $X \sim N(\mu, \sigma)$,
σ - nieznane

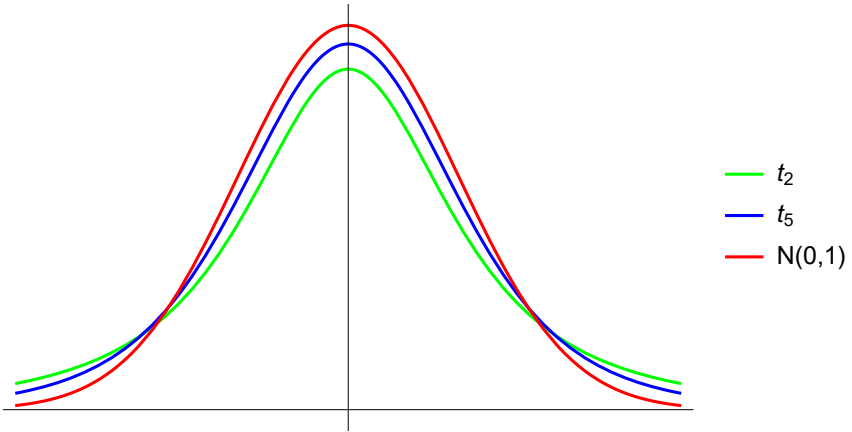
Próba: $X_1, X_2, \ldots, X_n$, $X_i \sim N(\mu, \sigma)$, σ - nieznane

(Standaryzowana) średnia z próby:

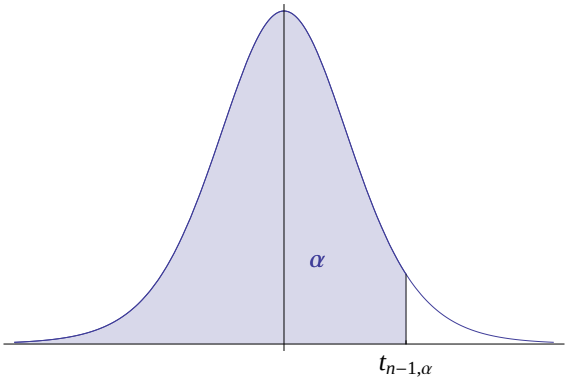
$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

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

Rozkład t-Studenta



Rozkład t-Studenta



$P(t < t_{n-1,\alpha}) = \alpha$

$t_{n-1,\alpha}$ – kwantyl rzędu α rozkładu t_{n-1}: $\texttt{qt}(\alpha,\, n-1)$

Przykład 3

Załóżmy, że waga dorosłych Polaków jest zmienną losową o rozkładzie normalnym ze średnią $\mu = 70$ i odchyleniem standardowym $\sigma = 10$.
9 studentów wsiadło do windy. Jakie jest prawdopodobieństwo, że ich całkowita waga przekroczyła dopuszczalną normę 650 kg?

Rozkład średniej z próby – podsumowanie

(1): $X_i \sim N(\mu, \sigma)$, μ, σ – znane:

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$
$$T \sim N(n\,\mu, \sqrt{n}\,\sigma)$$

(2): X_i ze znanymi μ, σ, rozkład X_i jest dowolny, duża próba:

$$\overline{X} \underset{\text{app}}{\sim} N(\mu, \frac{\sigma}{\sqrt{n}}) \quad \Rightarrow \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \underset{\text{app}}{\sim} N(0, 1)$$
$$T \underset{\text{app}}{\sim} N(n\,\mu, \sqrt{n}\,\sigma)$$

(3): $X_i \sim N(\mu, \sigma)$, μ – znane, σ – nieznane: $\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Wariancja z próby

Obserwowana cecha w populacji – zmienna losowa $X \sim N(\mu, \sigma)$,
σ - znane

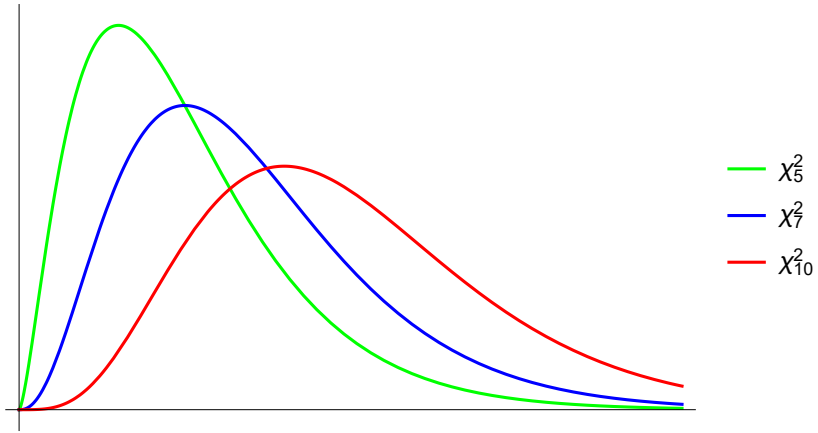
Próba: $X_1, X_2, \ldots, X_n$, $X_i \sim N(\mu, \sigma)$, σ - znane

Wariancja z próby: $$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$$

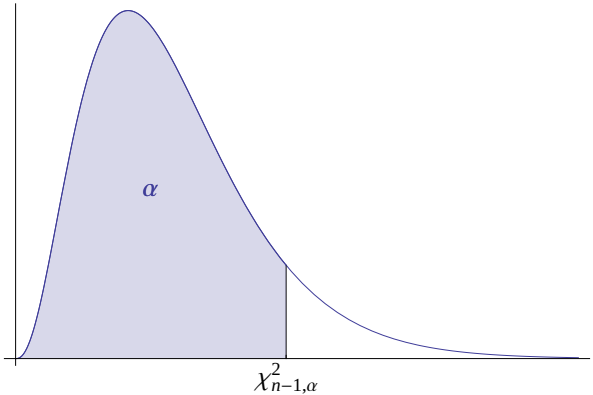
Rozkład wariancji z próby:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Rozkład χ^2 (chi-kwadrat)



Rozkład χ^2



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

$\chi^2_{n-1,\alpha}$ – kwantyl rzędu α rozkładu χ^2_{n-1}: $\texttt{qchisq}(\alpha, n-1)$

Proporcja populacyjna (prawdopodobieństwo sukcesu)

Obserwowana cecha w populacji – zmienna losowa $X \sim \text{bin}(1, p)$

Próba (duża): $X_1, X_2, \ldots, X_n \quad (n > 100), \quad X_i \sim \text{bin}(1, p)$

$$T = \sum_{i=1}^{n} X_i \text{ – liczba "sukcesów" w próbie}$$

Proporcja z próby: $\widehat{p} = \dfrac{T}{n}$

Rozkład proporcji populacyjnej:

$$\widehat{p} \underset{\text{app}}{\sim} N\left(p, \sqrt{pq/n}\right)$$

Przykład 5

Agencja reklamowa uruchomiła kampanię mającą wprowadzić na rynek nowy produkt. Na koniec kampanii przeprowadzono badanie na podstawie którego stwierdzono, że co najmniej 25% konsumentów kojarzy reklamowany produkt. Jeżeli 25% konsumentów rzeczywiście zna nowy produkt, to jakie jest prawdopodobieństwo, że nie więcej niż 232 losowo wybranych konsumentów spośród 1000 zna produkt?