Metody probabilistyczne i statystyka - wykład dziewiąty[1]

dr Jarosław Kotowicz

Instytut Informatyki Uniwersytet w Białymstoku

wersja z roku ak. 2021/22

Spis treści

Wszechstronna analiza struktury

Na wszechstronną analizę struktury składają się cztery zagadnienia

1. analiza tendencji centralnej (średniego poziomu cechy),
2. analiza zróżnicowania,
3. analiza asymetrii,
4. analiza koncentracji.

Parametry statystyczne i ich podział. I

Analiza danych statystycznych prowadzi do przedstawienia wyników badań za pomocą charakterystyk liczbowych zwanych parametrami statystycznymi.

Podział parametrów statystycznych

1. miary położenia,
2. miary zmienności (rozproszenia, dyspersji, zróżnicowania),
3. miary asymetrii (skośności),
4. miary koncentracji (spłaszczenia).

Parametry statystyczne i ich podział. II

Uwaga 1

- *Miary położenia wyznaczają przeciętną wartość cechy statystycznej.*
- *Miary zmienności wyznaczają siłę zróżnicowania wartości cechy statystycznej. Pozwalają określić jakie jest zróżnicowanie wartości cechy statystycznej w zbiorze obserwacji (jak mocno „rozproszone" są poszczególne obserwacje).*
- *Miary asymetrii wyznaczają siłę skupienia wartości cechy statystycznej bliżej dolnej lub górnej granicy zbioru wartości.*
- *Miary koncentracji wyznaczają siłę skupienia wartości cechy statystycznej wokół wartości przeciętnej.*

Pozycyjne miary położenia. I

Definicja 1

Kwartyl i-ty ($i = 1, 2, 3$), jest to wartość cechy statystycznej, dzieląca zbiór obserwacji na dwie części w następujący sposób: w pierwszym zbiorze o liczebności równej co najmniej $\frac{i}{4}$ wszystkich obserwacji znajdują się obserwacje, których wartości nie przekraczają wartości i-tego kwartyla, a w drugim, o liczebności co najmniej $\frac{4-i}{4}$, mamy obserwacje, których wartości są co najmniej równe wartości i-tego kwartyla.

Uwaga 2

Kwartyl drugi, to mediana.

Możliwe są co najmniej dwa sposoby wyznaczania kwartyli dla szeregu szczegółowego

Pozycyjne miary położenia. II

1. Zbiorowość dzielimy na dwie części. W pierwszej są te jednostki których cechy przyjmują wartości mniejsze niż mediana, a w drugiej pozostałe i dla nich wyznaczamy mediany, które będą odpowiednio kwartylem pierwszym Q_1 i trzecim Q_3.
2. Stosujemy następujące wzory

Pozycyjne miary położenia. III

$$Q_1 := \begin{cases} \frac{1}{2}\left(x_{\frac{n}{4}} + x_{\frac{n}{4}+1}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m\} \\ x_{\frac{n+1}{4}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+1\} \\ x_{\frac{n}{4}+\frac{1}{2}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+2\} \\ \frac{1}{2}\left(x_{\frac{n+1}{4}-\frac{1}{2}} + x_{\frac{n+1}{4}+\frac{1}{2}}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m+3\} \end{cases},$$

$$Q_3 := \begin{cases} \frac{1}{2}\left(x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m\} \\ x_{\frac{3(n+1)}{4}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+1\} \\ x_{\frac{3n}{4}+\frac{1}{2}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+2\} \\ \frac{1}{2}\left(x_{\frac{3(n+1)}{4}-\frac{1}{2}} + x_{\frac{3(n+1)}{4}+\frac{1}{2}}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m+3\} \end{cases}.$$

Pozycyjne miary położenia. IV

Dla kwartyli pierwszego i trzeciego dla szeregu rozdzielczego stosujemy następujące wzór

$$Q_i := x_{0Q_i} + \frac{N_{Q_i} - \sum_{i=1}^{m-1} n_i}{n_m} h_m,$$

gdzie

m – numer klasy, w której znajduje się kwartyl Q_i,

x_{0Q_i} – granica dolna przedziału, w którym znajduje się kwartyl Q_i,

n_m – liczebność przedziału kwartyla Q_i,

$\sum_{i=1}^{m-1} n_i$ – liczebność skumulowana,

h_m – rozpiętość przedziału kwartyla Q_i,

Pozycyjne miary położenia. V

N_{Q_i} – pozycja kwartyla Q_i (przyjmujemy $N_{Q_1} = \frac{n}{4}$ i $N_{Q_3} = \frac{3n}{4}$).

Definicja 2

Niech $p \in]0,1[$. Kwantylem rzędu p jednowymiarowej zmiennej losowej X nazywamy liczbę κ_p taką, że

$$P\left(\{\omega : X(\omega) \leqslant \kappa_p\}\right) \geqslant p \wedge P\left(\{\omega : X(\omega) \geqslant \kappa_p\}\right) \geqslant 1 - p. \tag{1}$$

Będziemy korzystać z następujących wzorów interpolacyjnych dla kwantyli dla szeregu rozdzielczego przedziałowego

$$\kappa_p := x_{0p} + \frac{n(p) - n(x_{0p})}{n_p} h_p, \tag{2}$$

gdzie

Pozycyjne miary położenia. VI

x_{0p} – granica dolna przedziału, w którym znajduje się kwantyl rzędu p,

$n(p)$ – pozycja kwantyla rzędu p (zauważmy, że $n(p) = np$),

$n(x_{0p})$ – liczebność skumulowana dla przedziału poprzedzającego przedział kwantyla rzędu p,

h_p – rozpiętość przedziału, w którym znajduje się kwantyl rzędu p,

n_p – liczebność przedziału, w którym znajduje się kwantyl rzędu p.

Podział miary zmienności

1. Klasyczne
   1. wariancja,
   2. odchylenie standardowe,
   3. odchylenie przeciętne,
   4. współczynnik zmienności Pearsona,
   5. typowy obszar zmienności.
2. Pozycyjne
   1. rozstęp próby,
   2. odchylenie ćwiartkowe,
   3. współczynnik zmienności,
   4. typowy obszar zmienności.

Klasyczne miary zmienności. I

Definicja 3

Wariancją nazywamy liczbę równą

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \textit{(dla szeregu szczegółowego)},$$

$$s^2 := \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 n_i \quad \textit{(dla szeregu rozdzielczego)}.$$

Klasyczne miary zmienności. II

Uwaga 3

Wariancję można też zdefiniować następująco:

$$\tilde{s}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \textit{(dla szeregu szczegółowego)},$$

$$\tilde{s}^2 := \frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 n_i \quad \textit{(dla szeregu rozdzielczego)}.$$

- s^2 nazywana jest wariancją populacyjną, $\tilde{s}^2$ wariancją próbkową[2].
- Wariancja mierzy średni rozrzut wartości zmiennej losowej od jej wartości średniej.
- Intuicyjnie wariancja utożsamiana jest ze zróżnicowaniem zbiorowości.

Klasyczne miary zmienności. III

Uwaga 4

Odchylenie standardowe i przeciętne opisuje rozrzut wartości zmiennej losowej wokół średniej arytmetycznej.

Klasyczne miary zmienności. IV

Definicja 4

Odchylenie standardowe to pierwiastek kwadratowy z wariancji (oznaczamy je s lub $\tilde{s}$ w zależności od tego jak wyznaczamy wariancję).
Odchylenie przeciętne definiujemy następująco

$$d := \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}| \quad \textit{(dla szeregu szczegółowego)},$$

$$d := \frac{1}{n}\sum_{i=1}^{k} |\dot{x}_i - \overline{x}| n_i \quad \textit{(dla szeregu rozdzielczego)}.$$

Klasyczne miary zmienności. V

Definicja 5

Klasycznymi współczynnikami zmienności (Pearsona) (dla $\overline{x} \neq 0$) nazywamy liczbę równą odpowiednio

$$V_s := \frac{s}{\overline{x}}, \quad V_d := \frac{d}{\overline{x}}.$$

Uwaga 5

Typowy obszar zmienności ma największe zastosowanie w przypadku, gdy dane są wyraźnie skupione wokół wartości średniej.

Klasyczny typowy obszar zmienności wyznacza się z wykorzystaniem odchylenia standardowego i odchylenia przeciętnego.

Klasyczne miary zmienności. VI

Definicja 6

Typowy obszarem zmienności jest przedział określony warunkiem

- $T_s :=]\overline{x} - s, \overline{x} + s[$ *(odpowiednio* $]\overline{x} - \tilde{s}, \overline{x} + \tilde{s}[$*),*
- $T_d :=]\overline{x} - d, \overline{x} + d[$.

[2] Zobacz pomoc do MS Excela.

Pozycyjne miary zmienności. I

Definicja 7

Pozycyjnym rozstępem próby nazywamy liczbę

$$R := x_{\max} - x_{\min},$$

gdzie $x_{\max} := \max_{i \in \overline{1,n}} x_i$ *oraz* $x_{\min} := \min_{i \in \overline{1,n}} x_i$.

Definicja 8

Odchyleniem ćwiartkowym nazywamy liczbę

$$Q := \frac{Q_3 - Q_1}{2}.$$

Pozycyjne miary zmienności. II

Definicja 9

Pozycyjnym typowym obszarem zmienności nazywamy następujący przedział

$$T_Q :=]\,\mathrm{Me} - Q, \mathrm{Me} + Q[.$$

Definicja 10

Pozycyjnymi współczynnikami zmienności nazywamy liczby równe odpowiednio

$$V_Q := \frac{Q}{\mathrm{Me}},\ (\mathrm{Me} \neq 0) \quad \textit{oraz} \quad V_{Q_1,Q_3} := \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Uwaga 6

Współczynnik zmienności informuje o sile rozproszenia.

Szeregi symetryczne i o asymetrii. I

Asymetria rozkładu cechy statystycznej oznacza, że elementy zbiorowości statystycznej skupiają się bliżej dolnej albo bliżej górnej granicy tej zbiorowości.

- Jeśli jednostki zbiorowości skupiają się bliżej mniejszych wartości cechy, to mówimy, że asymetria jest prawostronna.
- Jeśli jednostki zbiorowości skupiają się bliżej większych wartości cechy, to mówimy, że asymetria jest lewostronna.

Szereg statystyczny nazywamy symetrycznym, gdy

$$\overline{x} = \mathrm{Me} = \mathrm{Mo}\,.$$

Mówimy, że szereg statystyczny ma asymetrię prawostronną, gdy

$$\overline{x} > \mathrm{Me} > \mathrm{Mo}\,.$$

Szeregi symetryczne i o asymetrii. II

Mówimy, że szereg statystyczny ma asymetrię lewostronną, gdy

$$\overline{x} < \mathrm{Me} < \mathrm{Mo}\,.$$

Podział miar asymetrii

1. Klasyczne
   - współczynnik asymetrii,
2. Pozycyjne
   - wskaźnik asymetrii,
   - współczynnik asymetrii.
3. Mieszane
   - wskaźnik asymetrii,
   - pierwszy współczynnik asymetrii Pearsona,
   - drugi współczynnik asymetrii Pearsona,

Klasyczne miary asymetrii

Trzeci moment centralny

$$\mu_3 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 \quad \text{(dla szeregu szczegółowego)},$$

$$\mu_3 := \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^3 n_i \quad \text{(dla szeregu rozdzielczego)}.$$

Definicja 11

Klasycznym współczynnikiem asymetrii nazywamy liczbę

$$A := \frac{\mu_3}{s^3}.$$

Pozycyjny miary asymetrii

Definicja 12

Pozycyjnym wskaźnikiem asymetrii nazywamy liczbę

$$W_s^Q := (Q_3 - \mathrm{Me}) - (\mathrm{Me} - Q_1).$$

Pozycyjnym współczynnikiem asymetrii nazywamy liczbę

$$A_Q := \frac{(Q_3 - \mathrm{Me}) - (\mathrm{Me} - Q_1)}{2Q}.$$

Uwaga 7

Pozycyjny współczynnik asymetrii określa kierunek i siłę asymetrii jednostek znajdujących się miedzy pierwszym z trzecim kwartylem.

Mieszane miary asymetrii. I

Definicja 13

Wskaźnikiem asymetrii nazywamy liczbę

$$W_s := \overline{x} - \mathrm{Mo}\,.$$

Mówimy o asymetrii lewostronne (odpowiednio prawostronnej) wtedy, gdy $W_s < 0$ *(odpowiednio* $W_s > 0$*).*

Wskaźnik asymetrii określa kierunek asymetrii i nie określa siły.

Mieszane miary asymetrii. II

Definicja 14

Pierwszym współczynnikiem asymetrii Pearsona nazywamy liczbę

$$A_s := \frac{\overline{x} - \mathrm{Mo}}{s} \quad \textit{oraz} \quad A_d := \frac{\overline{x} - \mathrm{Mo}}{d}.$$

Drugi współczynnikiem asymetrii Pearsona nazywamy liczbę

$$W_{s,2} := \frac{\overline{x} - \mathrm{Me}}{s} \quad \textit{oraz} \quad W_{d,2} := \frac{\overline{x} - \mathrm{Me}}{d}.$$

Miary koncentracji

1. Krzywa Lorenza.
2. Klasyczne miary koncentracji
   1. współczynnik koncentracji (kurtoza),
   2. współczynnik ekscesu,
   3. współczynnik Giniego.
3. Pozycyjne miary koncentracji
   1. współczynnik koncentracji.

Krzywa Lorenza. I

Niech obserwacje z_i, gdzie $i \in \overline{1, n}$ spełniają warunki

$$0 \leqslant z_1 \leqslant z_2 \leqslant \ldots \leqslant z_n,$$
$$\sum_{i=1}^{n} z_i > 0.$$

Krzywa Lorenza. II

Definicja 15

Krzywą Lorenza odpowiadającą obserwacjom z_i nazywamy łamaną łączącą kolejne punkty (x_i, y_i) $(i \in \overline{0, n})$, gdzie

$$x_0 = y_0 = 0,$$

$$x_k := \frac{k}{n},\ y_k := \frac{\sum_{i=1}^{k} z_i}{\sum_{i=1}^{n} z_i} \quad k \in \overline{1, n}.$$

Uwaga 8

Krzywa Lorenza jest zawarta w kwadracie jednostkowym, a ponadto łączy dolny lewy wierzchołek kwadratu z górnym prawym.

Klasyczne miary koncentracji. I

Czwarty moment centralny

$$\mu_4 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 \quad \text{(dla szeregu szczegółowego)},$$

$$\mu_4 := \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^4 n_i \quad \text{(dla szeregu rozdzielczego)}.$$

Definicja 16

Współczynnikiem koncentracji (kurtozą) nazywamy liczbę

$$\gamma_4 := \frac{\mu_4}{(s^2)^2}.$$

Klasyczne miary koncentracji. II

- Przyjmuje się, że zbiorowość statystyczna ma rozkład normalny wtedy, gdy $\gamma_4 = 3$.
- Czwarty moment centralny można też zdefiniować wzorem[3]

$$\mu_4' := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^4 \quad \text{(dla szeregu szczegółowego)},$$

$$\mu_4' := \frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^4 n_i \quad \text{(dla szeregu rozdzielczego)}.$$

I przyjąć następującą definicję współczynnika koncentracji

$$\gamma_4 := \frac{\mu_4'}{(\tilde{s}^2)^2}.$$

Klasyczne miary koncentracji. III

Definicja 17

Mówimy, że rozkład cechy w populacji jest leptokurtyczny (wysmukły) wtedy, gdy $\gamma_4 > 3$. *Natomiast, gdy* $\gamma_4 < 3$, *to o rozkładzie cechy mówimy, że jest platokurtyczny (spłaszczony).*

Definicja 18

Liczbę $K := \gamma_4 - 3$ *nazywamy współczynnikiem ekscesu.*

Uwaga 9

W literaturze czasami współczynnik ekscesu nazywany też jest kurtozą.

Klasyczne miary koncentracji. IV

Definicja 19

Współczynnik Giniego dla obserwacji $x_1, \ldots, x_n$ definiujemy następująco

$$G := \frac{\sum_{i,j=1}^{n} |x_i - x_j|}{n^2 \overline{x}}. \tag{3}$$

Jeżeli obserwacje x_i, dla $i \in \overline{1,n}$, uporządkowane są rosnąco, to współczynnik Giniego wyraża się wzorem

$$G := \frac{\sum_{i=1}^{n} (2i - n - 1) x_i}{n^2 \overline{x}}.$$

Klasyczne miary koncentracji. V

Uwaga 10

Współczynnik Giniego jest to pole zwarte pomiędzy krzywą Lorenza, a przekątna kwadratu jednostkowego.

[3] Zobacz [1]

Pozycyjny miary koncentracji

Definicja 20

Pozycyjnym współczynnikiem koncentracji nazywamy liczbę

$$W_s := \frac{D_9 - D_1}{Q_3 - Q_1},$$

gdzie D_i jest i-tym decylem oraz $D_1 = x_{\frac{n}{10}}, D_9 = x_{\frac{9n}{10}}$ dla szeregu szczegółowego i

$D_i = x_{0D_i} + \frac{N_{D_i} - \sum_{i=1}^{m-1} n_i}{n_m} h_m$ *dla szeregu rozdzielczego, a pozostałe oznaczenia są analogiczne.*

Przykład wyznaczania dla szeregu rozdzielczego przedziałowego jego parametrów

Czas obsługi	Liczebność	Liczebność skumulowana	Częstotliwość
0 - 20	3	3	0, 12
20 - 40	9	12	0, 36
40 - 60	6	18	0, 24
60 - 80	5	23	0, 20
80 - 100	2	25	0, 08
$\sum$	25		1, 00

Dla szeregu rozdzielczego zadanego tablicą mamy

$$\mathrm{Me} = \ldots$$
$$\mathrm{Me} = 40 + (12,5 - 12)\frac{20}{6} = 41,67$$
$$Q_1 = \ldots$$
$$Q_1 = 20 + (0,25 - 0,12)\frac{20}{0,36} = 27,2$$
$$Q_2 = \ldots$$
$$Q_3 = 60 + (0,75 - 0,72)\frac{20}{0,20} = 63$$
$$\mathrm{Mo} = \ldots$$
$$\mathrm{Mo} = 20 + \frac{9-3}{(9-3)+(9-6)}20 = 43,3$$

Spis treści

Próba losowa i realizacja

Definicja 21

Próbą losową prostą (n - elementową) nazywamy ciąg $(X_1, \ldots, X_n)$ niezależnych zmiennych losowych o jednakowych rozkładach identycznych takich, jak rozkład cechy statystycznej X w populacji generalnej.
Piszemy wtedy $X_1, \ldots, X_n \sim_{iid} F_X$, gdzie F_X jest dystrybuantą cechy X.

Uwaga 11

Realizację zmiennych losowych oznaczamy $(x_1, \ldots, x_n)$.

Definicja 22

Przestrzenią próby zmiennej losowej $(X_1, \ldots, X_n)$ nazywamy zbiór wszystkich możliwych realizacji $(x_1, \ldots, x_n)$.

Statystyka z próby i rozkład z próby

Definicja 23

Niech $X_1, \ldots, X_n \sim_{iid} F_X$ będzie próba losową prostą.
Statystyką z próby nazywamy zmienną losową Z_n będącą pewną funkcją zmiennych losowych $X_1, \ldots, X_n$ stanowiących próbę losową tzn. $Z_n = f(X_1, \ldots, X_n)$.
Niech $Z_n = f(X_1, \ldots, X_n)$ będzie statystyką z próby.
Rozkładem z próby nazywamy rozkład statystyki z próby Z_n.

Przykłady statystyk z próby. I

Niech $X_1, \ldots, X_n \sim_{iid} F_X$ będzie próba losową prostą.

Definicja 24

Średnią z próby definiujemy jako

$$\overline{X}_{(n)} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Przykłady statystyk z próby. II

Uwaga 12

1. *Jeżeli wiadomo ile wynosi liczebność próby indeks liczebności prób w statystyka będziemy opuszczać.*
2. *W dalszych rozważaniach, o ile nie będzie to powiedziane inaczej, rozważamy próbę losową prostą n elementową.*
3. *Przy powyższych ustaleniach średnią z próby oznaczać będziemy* $\overline{X}$.

Przykłady statystyk z próby. III

Definicja 25

Wariancję z próby definiujemy jednym ze wzorów

$$S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{4}$$

$$\tilde{S}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{5}$$

$$\dot{S}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2, \quad \textit{gdy znane jest } m. \tag{6}$$

Przykłady statystyk z próby. IV

Stwierdzenie 1

Zauważmy, że zachodzi zależność

$$S^2 = \frac{n-1}{n}\tilde{S}^2.$$

Uwaga 13

W przypadku małych prób przy ustalonym n dla statystyki z próby otrzymujemy tzw. rozkład dokładny statystyki.
W przypadku dużych prób dla statystyki z próby otrzymujemy rozkład graniczny statystyki, o ile oczywiście istnieje.

Wprowadzenie do rozkładów z próby.

W trzech pierwszych sekcjach będziemy rozpatrywać populację generalną lub dwie populacje, w której cecha statystyczna ma rozkład normalny, a wnioskować będziemy o klasycznych statystykach z próby – średniej i wariancji.

Ponadto będziemy w nich rozpatrywać rozkłady dokładne statystyk z próby.

Rozkład średniej dla populacji normalnej ze znanym σ. I

Niech cecha w populacji generalnej ma rozkład normalny $\mathcal{N}(m, \sigma)$. Załóżmy, że danymi wartością oczekiwana i odchyleniem standardowym. Pobieramy losowo próbę prostą $X_1, \ldots, X_n$.

Lemat 1

Dla średniej arytmetycznej z próby $\overline{X}$ przy znanym odchyleniu standardowym mamy $\mathbb{E}(\overline{X}) = m$ oraz $\mathbb{D}^2(\overline{X}) = \frac{\sigma^2}{n}$.

Twierdzenie 1

Średnia arytmetyczna z próby ma rozkład normalny $\mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$.

Rozkład średniej dla populacji normalnej ze znanym σ. II

Przykład 1 ([1, Przkład 8.2])

Przypuśćmy, że rozkład pewnej cech w populacji jest rozkładem $\mathcal{N}(50,3)$. *Rozważymy próbę* 9*-cio elementową. Wyznaczymy*

1. *rozkład średniej z próby,*
2. *prawdopodobieństwo zdarzenia polegającego na tym, że średnia w tej próbie będzie większa niż* 52.

Rozkład średniej dla populacji normalnej ze znanym σ. III

Mamy $\overline{X} \sim \mathcal{N}(50, 1)$, gdyż w tym wypadku $\frac{\sigma^2}{n} = \frac{9}{9} = 1$ oraz

$$\begin{aligned} P(\{\overline{X} > 52\}) &= P\left(\left\{\frac{\overline{X} - 50}{1} > \frac{52 - 50}{1}\right\}\right) \\ &= P\left(\left\{\frac{\overline{X} - 50}{1} > 2\right\}\right) \\ &= 1 - \Phi(2) = 1 - 0,97725 = 0,02275. \end{aligned}$$

Można ostatni wynik interpretować, że na 100000 przypadków w 2275 przypadkach średnia będzie przekraczać 52.

Rozkład średniej dla populacji normalnej z nieznanymi m i σ. I

Niech cecha w populacji generalnej ma rozkład normalny $\mathcal{N}(m, \sigma)$. Załóżmy, że nieznane są wartość oczekiwana i odchylenie standardowe. Pobieramy losowo próbę prostą $X_1, \ldots, X_n \sim_{iid} \mathcal{N}(m, \sigma)$.

Uwaga 14

Jeżeli nie znamy parametrów tj. m i σ, to nie możemy wyznaczyć parametrów rozkładów średniej z próby $\overline{X}$.

Rozkład średniej dla populacji normalnej z nieznanymi m i σ. II

Definicja 26

Statystykę z próby dla rozkładu średniej z populacji normalnej z nieznanymi parametrami średnią i odchyleniem standardowym definiujemy wzorem

$$T := \frac{\overline{X} - m}{S}\sqrt{n-1},$$

gdzie $S = \sqrt{S^2}$, *a* S^2 *jest wariancją z próby.*

Lemat 2

Statystka z próby może też być zapisana w postaci

$$T := \frac{\overline{X} - m}{\tilde{S}}\sqrt{n}.$$

Rozkład średniej dla populacji normalnej z nieznanymi m i σ. III

Twierdzenie 2

Mamy $T \sim t(n-1)$ i rozkład T jest niezależny od σ.

1. W tablicach statystycznych dla rozkładu t-Studenta podawane jest liczba $t_{\alpha,\nu}$, gdzie α jest ustalona liczbą z odcinka $]0,1[$, a ν jest liczbą stopni swobody. Liczba $t_{\alpha,\nu}$ wyznaczana jest z zależności
$$P(\{|T| \geqslant t_{\alpha,\nu}\}) = \alpha.$$
Ponadto zachodzi wtedy zależność
$$P(\{T \geqslant t_{\alpha,\nu}\}) = P(\{-t_{\alpha,\nu} \leqslant T\}) = \frac{\alpha}{2}.$$
2. Gdy liczba stopni swobody zbiega do nieskończoności, to rozkład t-Studenta zbiega do rozkładu normalnego standardowego.

Rozkład średniej dla populacji normalnej z nieznanymi m i σ. IV

3. Na ćwiczeniach rachunkowych przyjmujemy, że gdy liczba stopni swobody jest większa niż 30, to rozkład t-Studenta zastępujemy standardowym rozkładem normalnym.

Rozkład średniej dla populacji normalnej z nieznanym σ, ale znaną wartością oczekiwaną. I

W tym przypadku wykorzystujemy statystykę

$$T := \frac{\overline{X} - m}{\dot{S}}\sqrt{n}.$$

Twierdzenie 3

Rozkład T jest rozkładem t-Studenta o n stopniach swobody.

Przykład 2 ([2, Przykład 7.3])

W populacji o rozkładzie normalnym $\mathcal{N}(12, \sigma)$ wylosowano pobrano próbkę 9-cio elementową. Wyznaczymy prawdopodobieństwo, że średnia będzie większa od 11,5, jeżeli wiadomo, że odchylenie standardowe w próbie wynosi 1,5?

Rozkład średniej dla populacji normalnej z nieznanym σ, ale znaną wartością oczekiwaną. II

$$\begin{aligned} P(\{\overline{X} > 11{,}5\}) &= P\left(\left\{\frac{\overline{X}-12}{1{,}5}\cdot 3 > \frac{11{,}5-12}{1{,}5}\cdot 3\right\}\right) \\ &= P\left(\left\{\frac{\overline{X}-12}{1{,}5}\cdot 3 > -1\right\}\right) \\ &= P(\{T > -1\}) = 1 - P(\{T < 1\}) \\ &\approx 1 - P(\{T < 1{,}1\}) = 1 - \frac{0{,}3}{2} = 0{,}85. \end{aligned}$$

Wynik dokładny to 0,8282818.

Motywacja

Uwaga 15 (Przyczyna liczenia rozkładów różnic średnich)

Rozkład różnicy średnich stosuje się w praktyce w celu porównania średnich arytmetycznych obliczonych na podstawie dwóch niezależnych prób pochodzących z dwóch różnych populacji.

Rozkład różnicy średnich dla dwóch populacji normalnych ze znanymi odchyleniami standardowymi. I

Dane są dwie populacje generalne mające rozkłady normalny $\mathcal{N}(m_1, \sigma_1)$ i $\mathcal{N}(m_2, \sigma_2)$. Pobieramy losowe próby proste n_1 elementową i n_2 elementową.

Wtedy za statystykę z próby przyjmujemy różnice średnich arytmetycznych z próby

$$\overline{X}^{(1)} - \overline{X}^{(2)}.$$

Twierdzenie 4

Rozkład różnice średnich arytmetycznych z próby jest rozkładem normalny z parametrami $m_1 - m_2$ *i* $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ *tj.* $\mathcal{N}\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.

Rozkład różnicy średnich dla dwóch populacji normalnych ze znanymi odchyleniami standardowymi. II

Przykład 3 ([1, Przykład 8.3])

Przypuśćmy, że mamy dwie populacje o rozkładach $\mathcal{N}(170, 5)$ i $\mathcal{N}(166, 4)$. Pobieramy losowo i niezależnie próby ośmioelementową (pierwsza populacja) i dzisięcio elementową (druga populacja). Jak należy interpretować prawdopodobieństwo zdarzenia polegającego na tym, że średnia wyników dla próby 10-cio elementowej jest większa od średniej wyników dla próby 8-mio elementowej.

Przyjmując standardowe oznaczenia $\overline{X}^{(1)}$ i $\overline{X}^{(2)}$ mamy $\overline{X}^{(1)} - \overline{X}^{(2)}$ ma rozkład normalny

$$\mathcal{N}\left(170 - 166, \sqrt{\frac{25}{8} + \frac{16}{10}}\right) = \mathcal{N}\left(4, \frac{3}{2}\sqrt{\frac{21}{10}}\right).$$

Rozkład różnicy średnich dla dwóch populacji normalnych ze znanymi odchyleniami standardowymi. III

Licząc prawdopodobieństwo mamy

$$\begin{aligned} P(\{\overline{X}^{(2)} > \overline{X}^{(1)}\}) &= P\left(\left\{\frac{\overline{X}^{(1)} - \overline{X}^{(2)} - 4}{\frac{3}{2}\sqrt{\frac{21}{10}}} < \frac{0-4}{\frac{3}{2}\sqrt{\frac{21}{10}}}\right\}\right) \\ &= \Phi\left(-\frac{8\sqrt{10}}{3\sqrt{21}}\right) = 1 - \Phi\left(\frac{8\sqrt{10}}{3\sqrt{21}}\right) \\ &= 1 - \Phi(1,84) = 1 - 0,96712 = 0,03288. \end{aligned}$$

Ostatni wynik interpretujemy, że na 100000 przypadków w 3288 przypadkach średnia z 10-cio elementowej próby z drugiej populacji jest większa niż średnia z 8-mio elementowej z pierwszej populacji.

Rozkład różnicy średnich dla dwóch populacji normalnych z nieznanymi, ale jednakowymi odchyleniami standardowymi. I

Dane są dwie populacje generalne mające rozkłady normalny $\mathcal{N}(m_1, \sigma)$ i $\mathcal{N}(m_2, \sigma)$.

Pobieramy losowo próby proste n_1 elementową z pierwszej populacji i n_2 elementową z drugiej. Wtedy za statystykę z próby przyjmujemy

$$T := \frac{\overline{X}^{(1)} - \overline{X}^{(2)} - (m_1 - m_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

gdzie $S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$, a S_i^2 jest wariancją z próby dla $i = 1, 2$.

Twierdzenie 5

Rozkład ten jest rozkładem t-Studenta o $n_1 + n_2 - 2$ stopniach swobody.

Rozkład różnicy średnich dla dwóch populacji normalnych z nieznanymi, ale jednakowymi odchyleniami standardowymi. II

Uwaga 16

Zauważmy, że wariancja połączonej $S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$ *można wyrazić wzorem*

$$S_p^2 = \frac{(n_1 - 1)\tilde{S}_1^2 + (n_2 - 1)\tilde{S}_2^2}{n_1 + n_2 - 2}.$$

Rozkład wariancji w populacji normalnej. I

Dana jest populacja o rozkładzie cechy $\mathcal{N}(m, \sigma)$. Pobieramy losową n elementową próbę $(X_1, \ldots, X_n)$.

O wariancji σ^2 populacji generalnej wnioskujemy w oparciu o statystykę próbkową

$$\xi_n := \frac{nS^2}{\sigma^2}.$$

Twierdzenie 6

Mamy $\xi_n \sim \chi^2(n-1)$.

Rozkład wariancji w populacji normalnej. II

Wniosek 1

Prawdziwe są następujące równości:

1. $\xi_n = \frac{(n-1)\tilde{S}^2}{\sigma^2}$,
2. $\mathbb{E}(\xi_n) = n-1$ *oraz* $\mathbb{D}^2(\xi_n) = 2(n-1)$,
3. $\mathbb{E}(\tilde{S}^2) = \sigma^2$ *oraz* $\mathbb{D}^2(\tilde{S}^2) = \frac{2\sigma^4}{n-1}$,
4. $\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2$ *oraz* $\mathbb{D}^2(S^2) = \frac{2(n-1)\sigma^4}{n^2}$.

Rozkład wariancji w populacji normalnej ze znana wartością oczekiwaną

O wariancji σ^2 populacji generalnej wnioskujemy w oparciu o statystykę

$$\xi_n := \frac{n\dot{S}^2}{\sigma^2}.$$

Twierdzenie 7

Statystyka ξ_n ma rozkład chi-kwadrat o n stopniach swobody.

Rozkład ilorazu wariancji. I

Dane są dwie niezależne populacje o normalnym rozkładzie cechy statystycznej o dowolnych wartościach oczekiwanych oraz wariancjach równych odpowiednio σ_1^2 i σ_2^2. Pobieramy niezależnie z każdej populacji próby o liczebnościach równych n_1 i n_2 i budujemy statystyki S_1^2 oraz S_2^2.

O ilorazie wariancji będziemy wnioskować na podstawie statystyki

$$F := \frac{\frac{n_1 S_1^2}{\sigma_1^2 (n_1 - 1)}}{\frac{n_2 S_2^2}{\sigma_2^2 (n_2 - 1)}} \equiv \frac{\frac{\tilde{S}_1^2}{\sigma_1^2}}{\frac{\tilde{S}_2^2}{\sigma_2^2}}.$$

Twierdzenie 8

Statystyka F ma rozkład Fishera-Snedecora z liczbą stopni swobody licznika $n_1 - 1$ i liczbą stopni swobody mianownika $n_2 - 1$.

Rozkład ilorazu wariancji. II

Wniosek 2

Wartość oczekiwana statystyki F istniej dla $n_2 > 3$ i wynosi $\mathbb{E}(F) = \frac{n_2-1}{n_2-3}$, natomiast wariancja istniej dla $n_2 > 5$ i wynosi $\mathbb{D}^2(F) = \frac{2(n_2-1)^2(n_1+n_2-4)}{(n_1-1)(n_2-3)^2(n_2-5)}$.

O tym co jest w rzeczywistości

Rzeczywistość – populacja nie ma rozkładu normalnego, bądź nie znana jest postać rozkładu.

Praktyka – stosujemy graniczne rozkłady statystyk.

Warunek – duża liczebność próby.

Graniczny rozkład częstości (frakcji). I

X ma rozkład Bernoulliego z parametrami n i p tzn. $X \sim Bin(n, p)$.

Stosujemy statystykę (częstotliwościową)

$$\Omega := \frac{X}{n}.$$

Uwaga 17

1. *Zmienną losową Ω nazywamy częstością względną.*
2. *Ponieważ $P\left(\{\omega : \Omega(\omega) = \frac{k}{n}\}\right) = P\left(\{\omega : X(\omega) = k\}\right)$ dla $k \in \overline{0, n}$, więc Ω ma rozkład dwumianowy liczbach $\frac{k}{n}$ dla $k \in \overline{0, n}$.*

Graniczny rozkład częstości (frakcji). II

Twierdzenie 9

Mamy $\mathbb{E}(\Omega) = p$ *oraz* $\mathbb{D}^2(\Omega) = \frac{p(1-p)}{n}$.

Twierdzenie 10 (Wniosek z twierdzenia Moivre'a-Laplace'a)

Rozkładem granicznym Ω *jest rozkład normalny* $\mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Graniczny rozkład różnicy frakcji.

Zmienne losowe X_1 i X_2 mają rozkład Bernoulliego z parametrami równymi odpowiednio n_1, p_1 i n_2, p_2. Rozważmy częstości względne Ω_1 i Ω_2 i ich różnice (analizujemy rozkład różnicy frakcji)

$$\Omega_1 - \Omega_2.$$

Twierdzenie 11

Rozkładem granicznym $\Omega_1 - \Omega_2$ *jest rozkład normalny* $\mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$.

Graniczny rozkład średniej i różnicy średnich z próby

Zmienna losowa X ma nieznany rozkład z danymi parametrami: średnią m i odchyleniem standardowym σ.

Twierdzenie 12

Rozkład średniej z n elementowej próby $\overline{X}_{(n)}$ zbiega według rozkładu do rozkładu normalnego $\mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

Zmienne losowe X_1 i X_2 mają dowolne rozkłady z parametrami m_1, σ_1 i m_2, σ_2.

Twierdzenie 13

Rozkład różnicy średnich z próby $\overline{X}^{(1)} - \overline{X}^{(2)}$ zbiega do rozkładu normalnego $\mathcal{N}\left(m_1 - m_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$.

Bibliografia

[1] J. Jóźwiak i J. Podgórski. *Statystyka od podstaw*. Wyd. 5 zmienione. Warszawa: PWE, 2000.

[2] St. Ostasiewicz, Z. Rusnak i U. Siedlecka. *Statystyka. Elementy teorii i Zadania*. Wyd. 5 poprawione. Wrocław: Wyd. Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, 2003.