

Rachunek prawdopodobieństwa

MWS, wykład 1

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 16 marca 2021

Przestrzeń probabilistyczna

- ▶ $\Omega = \{\omega_1, \omega_2, \dots\}$ przestrzeń zdarzeń elementarnych,
- ▶ ω_i zdarzenia elementarne
- ▶ $A \subset \Omega$ zdarzenie (losowe) - interpretacja: dowolny podzbiór Ω

Przykład 1: Eksperyment polegający na dwukrotnym rzucie monetą

- ▶ $\Omega = \{oo, or, ro, rr\}$
- ▶ zdarzenie "pierwszy wypadnie orzeł": $A = \{oo, or\}$

Przykład 2: Wybieramy się na pocztę. Interesuje nas czas oczekiwania w kolejce:

- ▶ $\Omega = \{t : t \geq 0\}$
- ▶ zdarzenie "nie dłużej niż 5 minut": $A = \{t : 0 \leq t \leq 5\text{min}\}$

Definicja

Prawdopodobieństwo, to funkcja \mathbb{P} przyporządkowująca zdarzeniom liczby rzeczywiste w taki sposób, że:

1. $\mathbb{P}(A) \geq 0$ dla każdego zdarzenia A ,
2. $\mathbb{P}(\Omega) = 1$,
3. Dla każdych rozłącznych zdarzeń A_1, A_2, A_3, \dots

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Przykład: Niech $\mathbb{P}(\{oo\}) = \mathbb{P}(\{or\}) = \mathbb{P}(\{ro\}) = \mathbb{P}(\{rr\}) = \frac{1}{4}$

ad 2 $\mathbb{P}(\{oo, or, ro, rr\}) = 1$

ad 3 np. $\mathbb{P}(\{oo, or\}) = \mathbb{P}(\{oo\}) + \mathbb{P}(\{or\}) = \frac{1}{2}$

- ▶ $\mathbb{P}(\emptyset) = 0$,
- ▶ $A \subset B \rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$,
- ▶ $0 \leq \mathbb{P}(A) \leq 1$,
- ▶ $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$, (A^C oznacza dopełnienie)
- ▶ $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Przykład cd.

ad 2 np. $A = \{oo\}$, $B = \{oo, ro\}$

ad 3 np. $A = \{oo\} \Rightarrow A^C = \{ro, or, rr\}$

Definicja

Zdarzenia A i B są **niezależne**, jeśli

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Definicja

Zdarzenia A_i , $i \in I$ są niezależne, jeśli dla każdego skończonego podzbioru $J \subset I$:

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i).$$

$N = 100$ żołnierzy strzela do samolotu (każdy z nich trafia z p-stwem $p = \frac{1}{100}$). Jakie jest p-stwo, że co najmniej jeden trafi?

$$\mathbb{P}(\text{któryś trafi}) = 1 - \mathbb{P}(\text{żaden nie trafi}) = 1 - (1 - p)^N = 0.63$$

Zmienne losowe i ich dystrybuanty

Zmienna losowa to funkcja

$$X: \Omega \rightarrow E \subset \mathbb{R}.$$

Jej dystrybuenta:

$$F_X: \mathbb{R} \rightarrow [0, 1], \quad F_X(x) = P(X \leq x).$$

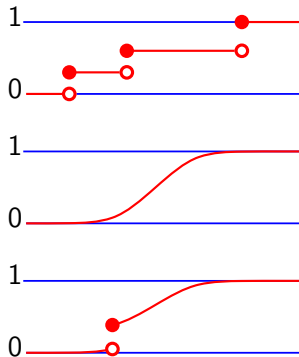
Duże litery — zmienne losowe, małe litery — ich wartości.

Przykład: Zdefiniujmy zmienną losową: $X(\{oo\}) = 0$, $X(\{or\}) = 1$,
 $X(\{ro\}) = 2$, $X(\{rr\}) = 3$

wtedy dystrybuenta spełnia: $F_X(0) = \frac{1}{4}$, $F_X(1) = \frac{1}{2}$, $F_X(2) = \frac{3}{4}$,
 $F_X(3) = 1$

Każda dystrybuanta $F(x)$ jest funkcją

- ▶ niemalejącą,
- ▶ dążącą do 0 dla $x \rightarrow -\infty$,
- ▶ dążącą do 1 dla $x \rightarrow +\infty$,
- ▶ prawostronnie ciągłą,
- ▶ posiadającą lewostronne granice,
- ▶ różniczkowalną prawie wszędzie.



Definicja

$$X: \Omega \rightarrow E \subset \mathbb{R}.$$

X jest **dyskretną zmienną losową**, jeśli zbiór E jest co najwyżej przeliczalny ($E = \{x_1, x_2, \dots, x_N\}$ lub $E = \{x_1, x_2, \dots\}$).

Funkcja prawdopodobieństwa zmiennej losowej X :

$$f_X: E \rightarrow \mathbb{R}, \quad f_X(x) = P(X = x).$$

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i).$$

Przykłady dyskretnych zmiennych losowych

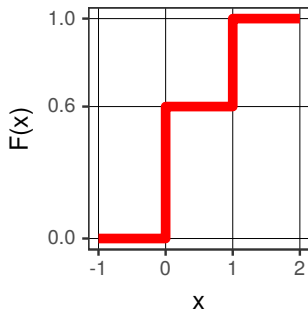
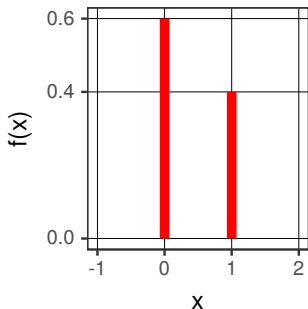
Rozkład Bernoulliego

X ma **rozkład Bernoulliego** (ewent. dwupunktowy) z parametrem p ,
 $X \sim \text{Bern}(p)$, jeśli

$$X: \Omega \rightarrow \{0, 1\}$$

tzn. zmienna losowa może przyjąć tylko dwie wartości, oraz

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p.$$



Przykłady dyskretnych zmiennych losowych

Rozkład dwumianowym

X ma **rozkład dwumianowy** (*ang. binomial*) z parametrami n, p ,
 $X \sim \text{Binom}(n, p)$, jeśli

$$X: \Omega \rightarrow \{0, 1, \dots, n\}$$

oraz

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Suma n niezależnych zmiennych o rozkładzie $\text{Bern}(p)$ ma rozkład $\text{Binom}(n, p)$.

$\text{Binom}(n, p)$ opisuje liczbę sukcesów w tzw. schemacie Bernoulliego (n niezależnych prób; w każdej próbie tylko dwie możliwości - sukces /porażka; ich prawdopodobieństwa takie same w każdej próbie).

Przykład

Rzucamy 10 razy kostką.

Jakie jest prawdopodobieństwo, że jedynka wypadnie dokładnie 4 razy?

Choć kostka ma 6 ścian, można ograniczyć się tylko do dwóch możliwości:

sukces wypadła jedynka ($p = 1/6$)

porażka wypadła inna liczba oczek ($p = 5/6$)

Jest to więc schemat Bernoulliego.

$$\mathbb{P}(X = 4) = \binom{10}{4} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 = 0.054$$

Przykłady dyskretnych zmiennych losowych

Rozkład Poissona

X ma **rozkład Poissona** z parametrem λ , $X \sim \text{Pois}(\lambda)$, jeśli

$$X: \Omega \rightarrow \{0, 1, 2, \dots\}$$

oraz

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Rozkład Poissona z parametrem λ można otrzymać

- ▶ „przechodząc do granicy” ($n \rightarrow \infty$) z rozkładem dwumianowym $\text{Binom}(n, p)$ w taki sposób, by $np = \lambda$.
- ▶ analizując liczbę niezależnych zdarzeń w danym przedziale czasu.
 $\mathbb{P}(X=k)$ jest prawdopodobieństwem tego, że liczba ta wynosi k jeśli oczekiwana liczba zdarzeń w tym przedziale jest równa λ .

Definicja

Zmienna losowa $X: \Omega \rightarrow E \subset \mathbb{R}$ jest **zmienną ciągłą** jeśli istnieje **funkcja gęstości prawdopodobieństwa** $f_X: E \rightarrow \mathbb{R}$ taka, że

- ▶ $f_X(x) \geq 0$,
- ▶ $\int_{-\infty}^{+\infty} f_X(x) dx = 1$,
- ▶ $\mathbb{P}(a < X < b) = \int_a^b f_X(x) dx, \quad \forall a \leq b$.

$$F_X(x) = \int_{-\infty}^x f_X(x_1) dx_1.$$

Dla wszystkich punktów $x \in \mathbb{R}$, w których F_X jest różniczkowalna zachodzi równość

$$f_X(x) = F'_X(x).$$

Przykłady ciągłych zmiennych losowych

Rozkład jednostajny

X ma **rozkład jednostajny** (*ang. uniform*) na przedziale $[a, b]$,
 $X \sim \text{Unif}([a, b])$, jeśli

$$X: \Omega \rightarrow [a, b]$$

oraz

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{jeśli } x \in [a, b], \\ 0, & \text{jeśli } x \notin [a, b]. \end{cases}$$

Przykłady ciągłych zmiennych losowych

Rozkład wykładniczy

X ma **rozkład wykładniczy** (*ang. exponential*) z parametrem λ ,
 $X \sim \text{Exp}(\lambda)$, jeśli

$$X: \Omega \rightarrow \mathbb{R}$$

oraz

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Przykłady ciągłych zmiennych losowych

Rozkład normalny (Gausa)

X ma **rozkład normalny** z parametrami μ , σ^2 , $X \sim N(\mu, \sigma^2)$, jeśli

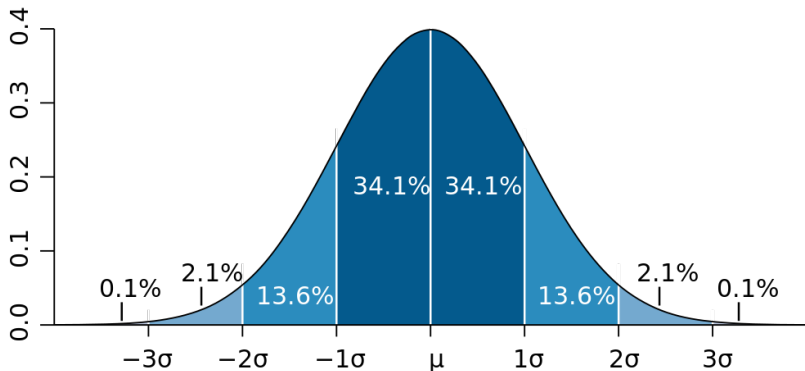
$$X: \Omega \rightarrow \mathbb{R}$$

oraz

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Funkcję gęstości oraz dystrybuantę zmiennej o rozkładzie $N(0, 1)$ oznacza się, odpowiednio, literami ϕ oraz Φ , a kwantyl rzędu p — z_p .

Reguła trzech sigm (reguła 68–95–99.7)



Odwrotna dystrybuanta (funkcja kwantylowa)

Jeśli X jest zmienną losową o dystrybuancie F , to **odwrotną dystrybuantą** tej zmiennej nazywamy funkcję

$$F^{-1}: [0, 1] \rightarrow \mathbb{R}, \quad F^{-1}(q) = \inf\{x: F(x) \geq q\}.$$

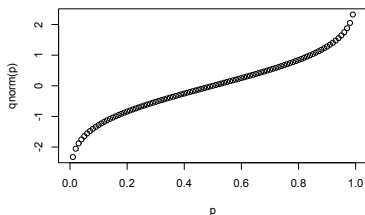
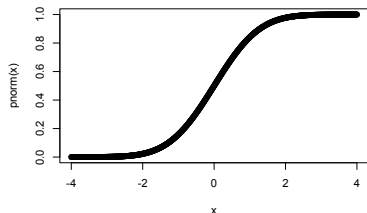
Innymi słowy

$$F^{-1}(p) \leq x \quad \text{wtedy i tylko wtedy gdy} \quad p \leq F(x).$$

$F^{-1}(\frac{1}{4})$ pierwszy (dolny) kwartyl rozkładu,

$F^{-1}(\frac{1}{2})$ mediana rozkładu,

$F^{-1}(\frac{3}{4})$ trzeci (górny) kwartyl.



Zmienna wielowymiarowa, to zmienna postaci:

$$X = (X_1, \dots, X_k): \Omega \rightarrow E \subset \mathbb{R}^k, \quad k > 1.$$

Podobnie jak w przypadku jednowymiarowym definiujemy funkcje prawdopodobieństwa (dla zm. dyskretnych) i funkcje gęstości prawdopodobieństwa (dla zm. ciągłych).

Rozkłady łączne i brzegowe

Zmienne dyskretne

Jeśli $X = (X_1, \dots, X_N)$ jest N -wymiarową dyskretną zmienną losową, to **łączną funkcję prawdopodobieństwa** nazywamy funkcję:

$$\begin{aligned} f_X: (x_1, \dots, x_N) &\mapsto \mathbb{P}(X = (x_1, \dots, x_N)) = \\ &= \mathbb{P}(X_1 = x_1, \dots, X_N = x_N). \end{aligned}$$

Brzegowe funkcje prawdopodobieństwa to funkcje prawdopodobieństwa określone zależnościami:

$$\begin{aligned} f_{X_i}(x_i) &= \mathbb{P}(X_i = x_i) = \\ &= \sum_{y_1} \cdots \sum_{y_{i-1}} \sum_{y_{i+1}} \cdots \sum_{y_N} f_X(y_1, \dots, y_{i-1}, x_i, y_{i+1}, \dots, y_N). \end{aligned}$$

Podobnie definiuje się dystrybuanty łączne i brzegowe.

Przykłady rozkładów wielowymiarowych

Rozkład normalny

$X = (X_1, \dots, X_n)$ ma **rozkład normalny** z parametrami

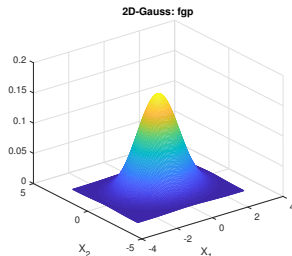
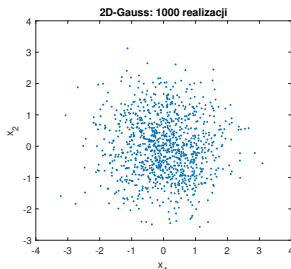
$$\mu = (\mu_1, \dots, \mu_n), \quad \Sigma = (\sigma_{ij})_{i,j=1,\dots,n}$$

(Σ macierz nieujemnie określona), $X \sim N(\mu, \Sigma)$, jeśli

$$X: \Omega \rightarrow \mathbb{R}^n$$

oraz dla $x = (x_1, \dots, x_n)$

$$f_X(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T}.$$



Niezależność zmiennych losowych

Niech X_1, \dots, X_N zmienne losowe z łączną funkcją prawdopodobieństwa (lub gęstości prawdopodobieństwa) f . Wówczas zmienne te są niezależne, jeśli

$$f(x_1, \dots, x_N) = \prod_{i=1}^N f_{X_i}(x_i), \quad \forall x_1, \dots, x_N.$$

Równanie powyższe pozwala znacznie uprościć obliczenia w przypadku niezależnych zmiennych losowych.

Wielowymiarowy rozkład normalny dla niezależnych zmiennych:

$$f(x_1, \dots, x_N) = \prod_{i=1}^N f_{\text{norm}}(x_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2}$$

Definicja

Jeśli X_1, \dots, X_N są niezależnymi zmiennymi losowymi o tym samym rozkładzie, to wektor (X_1, \dots, X_N) nazywamy **próbą losową** rozmiaru N z tego rozkładu.

Jeśli X_1, \dots, X_N próba losowa i X_i zadane dystrybuantą F lub funkcją (gęstości) prawdopodobieństwa f , to piszemy także

$$X_1, \dots, X_N \sim F \text{ lub } X_1, \dots, X_N \sim f$$

(Próba losowa z rozkładu F).

Przekształcenia zmiennych losowych

Zmienne dyskretne

- ▶ X dyskretna zmienna losowa,
- ▶ $h: \mathbb{R} \rightarrow \mathbb{R}$.
- ▶ Jaki rozkład ma zmienna losowa $Y = h(X)$?

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(X \in h^{-1}(y)).$$

Przykład: $Y = e^X$

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(e^X = y) = \mathbb{P}(X = \ln(y)) = f_X(\ln(y))$$

(dla $y > 0$).

Przekształcenia zmiennych losowych

Zmienne ciągłe

X ciągła zmienna losowa, $h: \mathbb{R} \rightarrow \mathbb{R}$. Jaki rozkład ma zmienna losowa $Y = h(X)$?

Zaczynamy od dystrybuanty!

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \in \{x: h(x) \leq y\}) = \int_{\{x: h(x) \leq y\}} f_X(x) dx,$$

Potem, w razie potrzeby, wyznaczamy funkcję gęstości prawd.

$$f_Y(y) = F'_Y(y).$$

Przekształcenia zmiennych losowych

Zmienne ciągłe — przykład

$$X \sim N(\mu, \sigma^2), \quad Y = e^X.$$

$$f_Y(y) \stackrel{!}{=} \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}}.$$

Rozkład zadany powyższą funkcją gęstości nazywamy **rozkładem logarytmicznie normalnym** z parametrami μ, σ^2 ($Y \sim \ln N(\mu, \sigma^2)$).

Dwa ciekawe i użyteczne fakty

Fakt

X ciągła zmienna losowa o dystrybuancie F. Wówczas

$$Z = F(X) \sim \text{Unif}([0, 1]).$$

Fakt

$U \sim \text{Unif}([0, 1])$ i $F: \mathbb{R} \rightarrow [0, 1]$ prawostronnie ciągła funkcja niemalejąca spełniająca warunki

$$\lim_{t \rightarrow +\infty} F(t) = 1, \quad \lim_{t \rightarrow -\infty} F(t) = 0,$$

to zmienna $X = F^{-1}(U)$ opisana jest dystrybuantą F.

Wartość oczekiwana zmiennej losowej X , to liczba

$$\mu_X = \mathbb{E}(X) = \begin{cases} \sum_x xf(x), & \text{jeśli } X \text{ dyskr.} \\ \int xf(x)dx, & \text{jeśli } X \text{ ciągła} \end{cases}$$

Wartość oczekiwana istnieje wtw, gdy (odpowiednio dla zm. dyskretnej i ciągłej)

$$\sum_x |x|f(x) < \infty, \quad \int |x|f(x)dx < \infty.$$

$$X \sim \text{Bern}(p) \quad \Rightarrow \quad \mathbb{E}(X) = p.$$

$$(\text{rozkład Cauchy'ego}): f_X(x) = \frac{1}{\pi(1+x^2)} \Rightarrow \mathbb{E}(X) \text{ nie istnieje!}$$

$$Y = h(X), \quad \mathbb{E}(Y) = ?$$

$$\mathbb{E}(Y) = \int h(x) f_X(x) dx,$$

(podobnie w przypadku dyskretnym).

Przykład:

$$\mathbb{E}(X^2) = \int x^2 f_X(x) dx,$$

X_1, \dots, X_n zmienne los. (**niekoniecznie niezależne**)

$$\mathbb{E}(a_1X_1 + \dots + a_nX_n) = a_1\mathbb{E}(X_1) + \dots + a_n\mathbb{E}(X_n).$$

Jeśli X_1, \dots, X_n **niezależne**, to

$$\mathbb{E}(X_1X_2 \dots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \dots \mathbb{E}(X_n).$$

Wariancję zmiennej losowej X nazywamy liczbę

$$\mathbb{V}X = \sigma_X^2 = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X - \mu_X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

Odchyleniem standardowym zmiennej losowej X nazywamy liczbę

$$\sigma_X = \sqrt{\mathbb{V}X}.$$

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}X, \quad \forall a, b \in \mathbb{R}.$$

Jeśli X_1, \dots, X_n **niezależne** zm. los. oraz a_1, \dots, a_n pewne stałe, to

$$\mathbb{V}(a_1X_1 + \dots + a_nX_n) = a_1^2 \mathbb{V}X_1 + \dots + a_n^2 \mathbb{V}X_n.$$

Wariancja i średnia z próby

Średnią z próby losowej X_1, \dots, X_n nazywamy zmienną losową

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

Wariancją z próby losowej X_1, \dots, X_n nazywamy zmienną losową

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

Jeśli próba jest z rozkładu o wartości oczekiwanej μ i wariancji σ^2 ($E(X_i) = \mu$ i $VX_i = \sigma^2$), to

$$E\bar{X}_n = \mu, \quad V\bar{X}_n = \frac{\sigma^2}{n}, \quad ES_n^2 = \sigma^2.$$

Kowariancją zmiennych losowych X i Y nazywamy liczbę

$$\mathbb{C}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

Współczynnikiem korelacji tych zmiennych nazywamy liczbę

$$\rho_{X,Y} = \rho(X, Y) = \frac{\mathbb{C}(X, Y)}{\sigma_X \sigma_Y}, \quad -1 \leq \rho_{X,Y} \leq 1.$$

$$\mathbb{C}(X, X) = \mathbb{V}X, \left(\text{pencil icon}\right)$$

$$\mathbb{V}(X + Y) = \mathbb{V}X + \mathbb{V}Y + 2\mathbb{C}(X, Y),$$

$$\mathbb{V}(X - Y) = \mathbb{V}X + \mathbb{V}Y - 2\mathbb{C}(X, Y),$$

$$\mathbb{V}\left(\sum_i (a_i X_i)\right) = \sum_i a_i^2 \mathbb{V}X_i + 2 \sum_{i < j} a_i a_j \mathbb{C}(X_i, X_j).$$

Jeśli X, Y niezależne, to $\mathbb{C}(X, Y) = 0$.

Jeśli $Y = aX + b$, to $\rho_{X,Y} = \operatorname{sgn} a$.

Co można powiedzieć o granicznym zachowaniu ciągu zmiennych losowych

$$X_1, X_2, \dots?$$

Rodzaje zbieżności

X, X_1, X_2, \dots zm. los. o dystrybuantach F, F_1, F_2, \dots

Definicja

Ciąg X_n **zbiega do X według prawdopodobieństwa**, $X_n \xrightarrow{\mathbb{P}} X$, jeśli

$$\mathbb{P}(|X_n - X| > \epsilon) \xrightarrow{n \rightarrow \infty} 0 \quad \forall \epsilon > 0.$$

Ciąg X_n **zbiega do X według rozkładu**, $X_n \xrightarrow{d} X$, jeśli

$$\lim_{n \rightarrow \infty} F_n(t) = F(t)$$

dla każdego punktu t , w którym F jest ciągłe.

Jeśli $X_n \xrightarrow{\mathbb{P}} X$, to również $X_n \xrightarrow{d} X$! (implikacja w drugą stronę nie zachodzi).

Twierdzenie ((słabe) Prawo Wielkich Liczb (PWL))

Jeśli X_1, X_2, \dots są niezależnymi zmiennymi losowymi o tym samym rozkładzie i skończonej wartości oczekiwanej ($|\mathbb{E}X_1| < \infty$), to

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}X_1.$$

Uwaga: zmienna losowa zbiega do wartości deterministycznej.

Twierdzenie (Centralne twierdzenie graniczne (CTG))

Jeśli X_1, X_2, \dots są niezależnymi zmiennymi losowymi o tym samym rozkładzie, $\mathbb{E}X_1 = \mu$, $\mathbb{V}X_1 = \sigma^2$, to

$$\sqrt{n} (\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Równoważnie

$$\frac{\sqrt{n} (\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1),$$

$$\frac{\bar{X}_n - \mathbb{E}\bar{X}_n}{\sqrt{\mathbb{V}\bar{X}_n}} \xrightarrow{d} N(0, 1),$$

Przy założeniach CTG także

$$\frac{\sqrt{n} (\bar{X}_n - \mu)}{S_n} \xrightarrow{d} N(0, 1),$$

Metoda momentów

MWS, wykład 2

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 16 marca 2021

Model parametryczny

Przykład

x_i to wynik i -tego pomiaru pewnej nieznannej wielkości x (pomiaru obarczone są pewnym losowym błędem).

- ▶ Jak oszacować wartość x ?
- ▶ Jak oszacować parametry błędu pomiaru (np. wariancję)?
- ▶ Jak oszacować prawdopodobieństwo, że następny pomiar da wynik większy niż dotychczasowe?
- ▶ itd.

Potrzebny model,

np. $X_i \sim N(x, \sigma^2)$ o pewnych parametrach x i σ^2 ,

- ▶ X_i to niezależne zmienne losowe,
- ▶ x_i to ich realizacje, czyli
- ▶ x_1, \dots, x_n to **próba losowa** z rozkładu $N(x, \sigma^2)$.

Model parametryczny

Parametryczna rodzina rozkładów

W modelu parametrycznym mamy do czynienia z rodziną rozkładów

$$\mathcal{F} = \{F_{\theta}(x) : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^k,$$

gdzie

$$\theta = (\theta_1, \dots, \theta_k)$$

to parametr (rozkładu).

Zamiast dystrybuant F_{θ} będziemy też rozważać funkcje prawdopodobieństwa lub funkcje gęstości prawdopodobieństwa f_{θ} .

Czasami będziemy stosowali oznaczenia typu:

$$F(x; \theta), f(x; \theta), \dots$$

Będziemy dalej zakładać, że mamy do dyspozycji próbę losową

$$X_1, \dots, X_n,$$

(ciąg **niezależnych zmiennych o tym samym rozkładzie**) z pewnego rozkładu F_θ należącego do rodziny parametrycznej

$$\mathcal{F} = \{F_\theta(x) : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^k.$$

Uwagi:

- ▶ wartość parametru θ nie jest znana, ale
- ▶ zmienne X_1, \dots, X_n odpowiadają tej samej wartości θ .

Model parametryczny

Typowe zagadnienia

Typowymi zagadnieniami rozważanymi w kontekście modeli parametrycznych są:

- ▶ wyznaczenie, na podstawie obserwacji (próby): X_1, \dots, X_n , nieznanego parametru θ , tzn. wyznaczenie takiej funkcji (estymatora)

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n),$$

która w jakimś sensie (w jakim?) przybliża (jak dobrze?) θ (teoria estymacji).

- ▶ sprawdzenie, na podstawie obserwacji: X_1, \dots, X_n , czy θ spełnia (z jakim prawdopodobieństwem?) pewne warunki, np.

$$\theta = \theta_1, \quad \theta > \theta_1, \quad \theta \neq \theta_1$$

(testowanie hipotez statystycznych, teoria detekcji).

Model parametryczny — przykłady

Rodzina rozkładów Bernoulliego

$$\mathcal{F} = \{f_p(x) : p \in [0, 1]\} \quad (\theta = p).$$

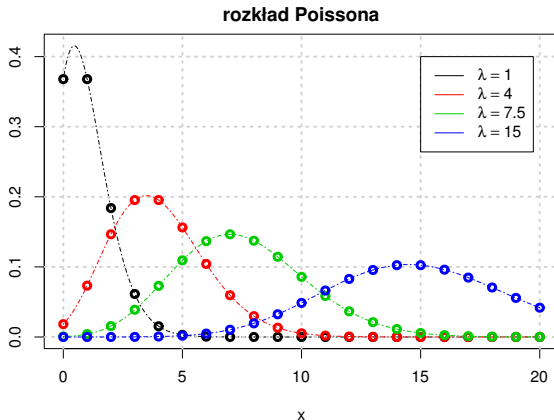
$$f_p(x) = \begin{cases} p & x = 1, \\ 1 - p & x = 0. \end{cases}$$

Próba losowa X_1, \dots, X_N z rozkładu z rodziny \mathcal{F} może modelować N niezależnych rzutów tą samą monetą, o której nie wiemy jak bardzo jest niesymetryczna.

Model parametryczny — przykłady

Rodzina rozkładów Poissona

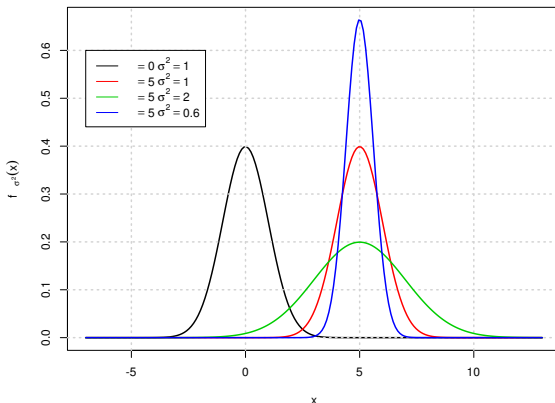
$$\mathcal{F} = \{f_\lambda(x) : \lambda > 0\} \quad (\theta = \lambda).$$
$$f_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x \in \mathbb{N} = \{0, 1, 2, \dots\}.$$



Rodzina rozkładów normalnych

$$\mathcal{F} = \{f_{(\mu, \sigma^2)}(x) : \mu \in \mathbb{R}, \sigma^2 > 0\} \quad \theta = (\mu, \sigma^2).$$

$$f_{(\mu, \sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



$$f_{\alpha,\beta}(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0, \alpha, \beta > 0,$$

gdzie

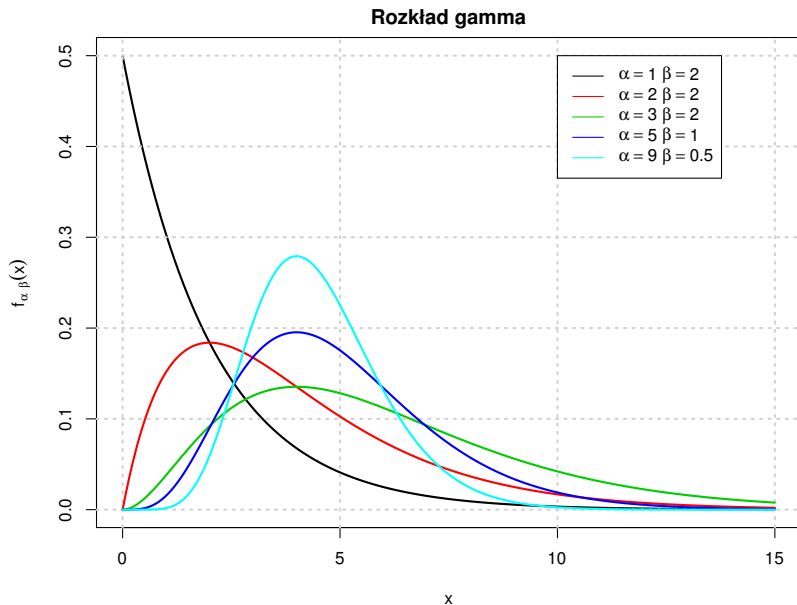
$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du, \quad \alpha > 0.$$

$$\Gamma(x+1) = x\Gamma(x) \quad x > 0, \quad \Gamma(1) = 1, \quad \Gamma(n+1) = n! \quad n \in \mathbb{N}.$$

Dla $\alpha = 1$ otrzymujemy rozkład $\text{Exp}(\frac{1}{\beta})$.

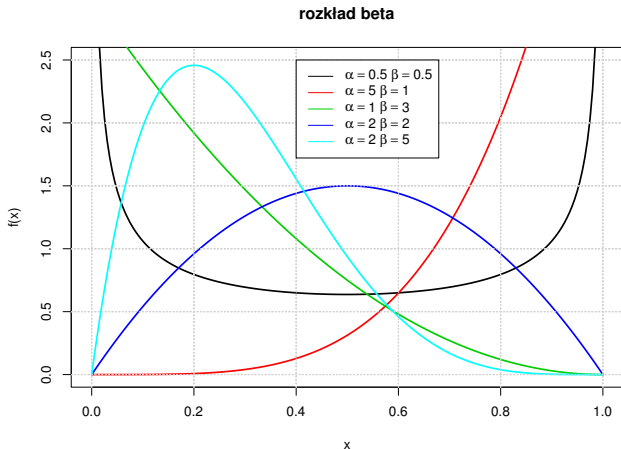
Uwaga! Czasami zamiast **parametru skali** β używa się parametru $\frac{1}{\beta}$.

Rozkład Gamma — przykłady



Rozkład beta

$$f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad x \in [0, 1], \alpha, \beta > 0$$



Model parametryczny

Niedokładność wyznaczenia θ

Źródła błędu:

1. typowy błąd estymacji wynikający z losowego charakteru modelu;
2. błąd określenia modelu (przyjęty model nie pasuje idealnie do opisywanej rzeczywistości)

Wniosek: Jest dopuszczalne, aby model nie odzwierciedlał „idealnie” badanego zjawiska.

Np. wzrost ludzi możemy modelować rozkładem normalnym z parametrami $\theta = (\mu, \sigma)$.

Czasami zamiast parametru

$$\theta = (\theta_1, \dots, \theta_k)$$

zainteresowani jesteśmy pewną funkcją $T(\theta)$ tego parametru.

Przykład: założmy, że czas życia X osobników pewnej populacji (ludzi, zwierząt, czy urządzeń elektronicznych) modelujemy rozkładem $\text{Gamma}(\alpha, \beta)$, gdzie

$$\theta = (\alpha, \beta) \in \Theta \subset \mathbb{R}^2.$$

Możemy być zainteresowani średnią długością życia.

Ta średnia długość życia jest równa:

$$\mathbb{E}X \stackrel{\text{pencil}}{=} \alpha\beta = T(\alpha, \beta) = T(\theta).$$

Metoda momentów

Przykład wprowadzający

Rozważmy rodzinę

$$\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}, \quad \theta = (\mu, \sigma^2).$$

oraz n -elementową próbę $X_1, \dots, X_n \sim N(\mu, \sigma^2)$.

Chcemy na podstawie tej próby wyznaczyć nieznane μ, σ .

Zgodnie z prawem wielkich liczb mamy

$$1. \quad \bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}X_1 = \mu$$

$$2. \quad \overline{(X^2)}_n = \frac{X_1^2 + \dots + X_n^2}{n} \xrightarrow{\mathbb{P}} \mathbb{E}X_1^2 = \mathbb{V}(X_1) + (\mathbb{E}X_1)^2 = \sigma^2 + \mu^2.$$

Za estymatory parametrów μ, σ możemy zatem przyjąć, odpowiednio,

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma} = \sqrt{\overline{(X^2)}_n - (\bar{X}_n)^2}.$$

Definicja

Momentem k -tego rzędu zmiennej losowej X nazywamy liczbę

$$\mu_k = \mathbb{E}X^k.$$

Definicja

Momentem k -tego rzędu z próby X_1, \dots, X_n nazywamy zmienną losową

$$m_k = \frac{1}{n} \sum_{i=1}^n X_i^k = \overline{(X^k)}_n.$$

Definicja

Estymator metody momentów dla parametru

$\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$ i próby losowej X_1, \dots, X_n , to taka statystyka (funkcja próby, a więc zmienna losowa!)

$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, która „podstawiona w miejsce parametru θ ” daje wybrane momenty rozkładu równe odpowiednim momentom z próby (np. $\mu_1 = m_1$, $\mu_2 = m_2$ i $\mu_4 = m_4$).

Najczęściej w powyższej definicji przez „wybrane” rozumie się pierwsze kolejne k momenty pozwalające wyznaczyć estymator parametru.

Próba losowa X_1, \dots, X_n — estymator $\hat{\theta}(X_1, \dots, X_n)$ (**zmienna losowa**).

Wartości próby x_1, \dots, x_n — wartość estymatora $\hat{\theta}(x_1, \dots, x_n)$ (**liczba, wektor liczbowy**).

Algorytm metody momentów

krok 1 Przedstaw momenty niskich rzędów jako funkcje poszukiwanych parametrów $\theta_1, \dots, \theta_k$

$$(\mu_1, \dots, \mu_k) = g(\theta_1, \dots, \theta_k)$$

krok 2 Odwróć funkcje z poprzedniego kroku, aby wyrazić poszukiwane parametry $\theta_1, \dots, \theta_k$ jako funkcje momentów.

$$(\theta_1, \dots, \theta_k) = g^{-1}(\mu_1, \dots, \mu_k)$$

krok 3 W wyrażeniach z poprzedniego punktu w miejsce momentów wstaw momenty z próby

$$(\hat{\theta}_1, \dots, \hat{\theta}_k) = g^{-1}(m_1, \dots, m_k)$$

Metoda momentów

Przykład 1

$$X_1, \dots, X_n \sim \text{Bern}(p).$$

Jaką postać ma estymator m. m. parametru p ?

(krok 1 i 2) Wartość oczekiwana:

$$\mu_1 = \mathbb{E}X = p$$

Wartość oczekiwana z próby:

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i.$$

(krok 3) Zatem estymatorem m. m. parametru p jest

$$\hat{p} = m_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n.$$

$$X_1, \dots, X_n \sim \text{Pois}(\lambda).$$

Jaką postać ma estymator m. m. parametru λ ?

$$\lambda = \mu_1.$$

Zatem estymatorem m. m. parametru λ jest średnia z próby:

$$\hat{\lambda} = m_1.$$

Metoda momentów

Przykład 2a (Bortkiewicz, 1898)

Analizowana była liczba zgonów od kopnięcia konia dla 10 korpusów pruskiej kawalerii w przeciągu 20-letniego okresu (mamy 200 „korpuso-lat”).

liczba zgonów/rok	liczba „korpuso-lat”
0	109
1	65
2	22
3	3
4	1

Próbujemy „dopasować” rozkład $\text{Pois}(\lambda)$.

$$\hat{\lambda} = (109 \times 0 + 65 \times 1 + 22 \times 2 + 3 \times 3 + 1 \times 4) / 200 = 0.61$$

Metoda momentów

Przykład 2a (Bortkiewicz, 1898) — podsumowanie

zgonów/rok (k)	„korpuso-lat”	częstość	$P_{\hat{\lambda}}(k)$	$P_{\hat{\lambda}}(k) \cdot 200$
0	109	0.545	0.543	109
1	65	0.325	0.331	66
2	22	0.110	0.101	20
3	3	0.015	0.021	4
4	1	0.005	0.003	1

Funkcja generująca momenty

Definicja

Funkcją generującą momenty dla zmiennej X o pewnym rozkładzie lub po prostu funkcją generującą momenty dla tego rozkładu nazywamy funkcję określoną zależnością

$$M(t) = \mathbb{E}e^{tX}, \quad t \in \mathbb{R}.$$

Twierdzenie

Jeśli funkcja generująca momenty $M(t)$ dla zmiennej X jest dobrze określona w pewnym otoczeniu zera, to

- ▶ *M jednoznacznie określa rozkład zmiennej X ,*
- ▶ *$\mu_k = M^{(k)}(0)$.*

Funkcja generująca momenty

Rozkład normalny

$$X \sim N(\mu, \sigma^2).$$

$$M(t) = \mathbb{E}e^{tX} = \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = e^{t\mu + \sigma^2 t^2/2}.$$

$$\mathbb{E}X = \mu, \mathbb{E}X^2 = \mu^2 + \sigma^2, \mathbb{E}X^3 \stackrel{\text{pencil}}{=} \mu^3 + 3\mu\sigma^2, \mathbb{E}X^4 \stackrel{\text{pencil}}{=} \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4.$$

Funkcja generująca momenty

Rozkład gamma

$$X \sim \text{Gamma}(\alpha, \beta).$$

$$f_{\alpha, \beta}(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0, \alpha, \beta > 0.$$

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du, \quad \alpha > 0.$$

$$M(t) = (1 - \beta t)^{-\alpha}, \quad t < \frac{1}{\beta}.$$

$$\mu_1 \stackrel{\text{pencil}}{=} \alpha\beta, \quad \mu_2 \stackrel{\text{pencil}}{=} \alpha(\alpha + 1)\beta^2.$$

$$X_1, \dots, X_n \sim \text{Gamma}(\alpha, \beta).$$

$$\mu_1 = \alpha\beta, \quad \mu_2 = \alpha(\alpha + 1)\beta^2.$$

Wyznaczamy parametry w funkcji momentów:

$$\mu_2 = \mu_1^2 + \mu_1\beta \implies \beta = \frac{\mu_2 - \mu_1^2}{\mu_1}, \quad \alpha = \frac{\mu_1}{\beta} = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

Estymatorami m. m. parametrów α , β są

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}, \quad \hat{\beta} = \frac{m_2 - m_1^2}{m_1}.$$

Zmieniając licznosc próby $n = 1, 2, \dots$ dostajemy ciąg estymatorów:

$$\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \dots, \hat{\theta}_{(n)}, \dots$$

Twierdzenie

Estymator m. m. dla parametru θ jest zgodny, tzn.

$$\hat{\theta}_{(n)} \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta.$$

Właściwości estymatorów metody momentów

Asymptotyczna normalność

Twierdzenie

Estymator m. m. dla parametru θ jest *asymptotycznie normalny*.
W przypadku skalarne parametru ($\theta = \theta_1$) oznacza to

$$\sqrt{n} \left(\hat{\theta}_{(n)} - \theta \right) \xrightarrow[n \rightarrow \infty]{d} N(0, \sigma^2),$$

gdzie

$$\sigma^2 = \frac{g_2(\theta) - g_1^2(\theta)}{(g_1'(\theta))^2} = \frac{\mu_2(\theta) - \mu_1^2(\theta)}{\left(\frac{\partial \mu_1(\theta)}{\partial \theta} \right)^2}.$$

Uwaga: wyrażenie w liczniku opisuje wariancję rozkładu X wyrażoną przez θ .

Przykład

Rozkład wykładniczy $\text{Exp}(\lambda)$ ma fgp

$$f(x) = \lambda e^{-\lambda x}$$

Przyjmijmy, że mamy próbę losową z tego rozkładu:

$$X_1, \dots, X_n \sim \text{Exp}(\lambda).$$

- ▶ Wartość oczekiwana jest równa

$$\mathbb{E}(x) = 1/\lambda,$$

- ▶ więc estymator parametru λ wg. metody momentów to

$$\hat{\lambda} = \frac{1}{m_1} = \frac{1}{\frac{1}{n} \sum_{i=1}^n X_i}$$

- ▶ parametr σ^2 z twierdzenia o asymptotycznej zbieżności do rozkładu normalnego jest równy:

$$\sigma^2 = \frac{\lambda^{-2}}{(-\lambda^{-2})^2} = \lambda^2 (\text{pencil icon})$$

Co oznacza zbieżność $\hat{\lambda}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \lambda$?

$$\mathcal{F} = \{\text{Exp}(\lambda) : \lambda > 0\}$$

$$\hat{\lambda}_n = 1/\bar{X}_n$$

np. $\lambda = 2$, $\forall_\epsilon \mathbb{P}(|\hat{\lambda}_n - 2| > \epsilon) \xrightarrow{n \rightarrow \infty} 0$, weźmy np. $\epsilon_0 = 0.01$

$$\mathbb{P}(|\hat{\lambda}_{10} - 2| > \epsilon_0) \approx 0.99,$$

$$\mathbb{P}(|\hat{\lambda}_{100} - 2| > \epsilon_0) \approx 0.96,$$

$$\mathbb{P}(|\hat{\lambda}_{1000} - 2| > \epsilon_0) \approx 0.87,$$

$$\mathbb{P}(|\hat{\lambda}_{10000} - 2| > \epsilon_0) \approx 0.62,$$

$$\mathbb{P}(|\hat{\lambda}_{100000} - 2| > \epsilon_0) \approx 0.11,$$

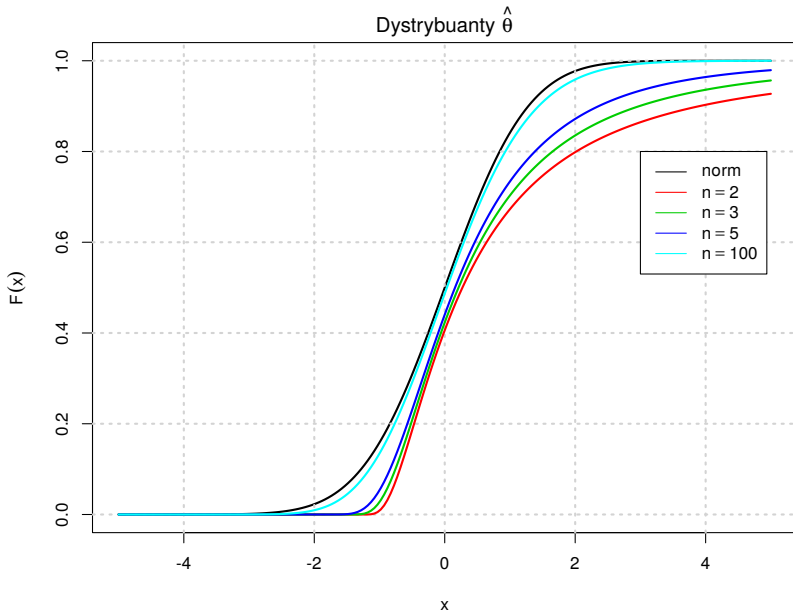
$$\mathbb{P}(|\hat{\lambda}_{200000} - 2| > \epsilon_0) \approx 0.03,$$

Co oznacza zbieżność $\sqrt{n} \left(\hat{\lambda}_n - \lambda \right) \xrightarrow[n \rightarrow \infty]{d} N(0, \lambda^2)$?

Oznacza punktową zbieżność dystrybuant F_n zmiennych $Y_n = \sqrt{n} \left(\hat{\lambda}_n - \lambda \right)$ do dystrybuanty F rozkładu $N(0, \lambda^2)$, t.j.

$$\forall t \in \mathbb{R} \quad F_n(t) \xrightarrow[n \rightarrow \infty]{} F(t).$$

Co oznacza zbieżność $\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow[n \rightarrow \infty]{d} N(0, \lambda^2)$?



Estymatory największej wiarygodności

MWS, wykład 3

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 23 marca 2021

Definicja

Model parametryczny:

$$\mathcal{F} = \{f_{\theta}(x) : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^k,$$

gdzie $\theta = (\theta_1, \dots, \theta_k)$ to parametr (rozkładu). Zamiast funkcji gęstości można rozważać dystrybuanty lub funkcje prawdopodobieństwa.

Funkcja wiarygodności

$$\mathcal{F} = \{f_{\theta}(x) : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^k,$$

X_1, \dots, X_n próba losowa (niezależne zmienne losowe o tym samym rozkładzie) z rozkładu odpowiadającemu pewnej **nieznanej** wartości θ .

Definicja (Funkcja wiarygodności)

Funkcja wiarygodności \mathcal{L}_n to funkcja określona formułą

$$\mathcal{L}_n(x_1, \dots, x_n; \theta) = f_{\theta}(x_1) \times \dots \times f_{\theta}(x_n).$$

- ▶ Ogólniej: łączna funkcja gęstości prawdopodobieństwa
- ▶ Skrócone oznaczenie: $\mathcal{L}_n(\theta)$
- ▶ Logarytmiczna funkcja wiarygodności ℓ_n to

$$\ell_n(\theta) = \ell_n(x_1, \dots, x_n; \theta) = \ln(\mathcal{L}_n(\theta)).$$

Definicja

Estymator największej wiarygodności $\hat{\theta}_n$ parametru θ to taka wartość parametru θ , dla której funkcja wiarygodności przyjmuje maksimum, czyli

$$\hat{\theta}_n = \arg \max_{\theta} \mathcal{L}_n(X_1, \dots, X_n; \theta).$$

- ▶ Interpretacja: taki parametr θ , który prowadzi do największego p-stwa otrzymanych wyników pomiaru.
- ▶ Estymator NW jest zmienną losową (jak każdy estymator)
- ▶ Równoważnie:

$$\hat{\theta}_n = \arg \max_{\theta} \ell_n(X_1, \dots, X_n; \theta).$$

Przykład: estymator n. w. dla rozkładu Bern(p)

Zadanie: Rzucono n razy monetą. Znaleźć estymator N.W. p -stwa wypadnięcia orła.

- ▶ Tworzymy zmienną losową X . Zdarzeniu „orzeł” przypisujemy 1, a zdarzeniu „reszka” przypisujemy 0. Zmienna ta ma rozkład Bernoulliego.
- ▶ Przypomnienie: X ma rozkład Bernoulliego z parametrem p , $X \sim \text{Bern}(p)$, jeśli zmienna losowa może przyjąć tylko dwie wartości, oraz

$$f_p(1) = p, \quad f_p(0) = 1 - p.$$

- ▶ Rodzina rozkładów Bernoulliego z nieznanym parametrem

$$\mathcal{F} = \{f_p(x) : p \in [0, 1]\}$$

Przykład: estymator n. w. dla rozkładu Bern(p)

Wyznaczenie estymatora:

1. Wyznaczamy funkcję wiarygodności (nieznany parametr to p)

$$\mathcal{L}_n(p) = f_p(X_1) \times \cdots \times f_p(X_n) = p^{L_1}(1-p)^{L_0}$$

gdzie L_1 to liczba 1, a L_0 to liczba 0, $L_1 + L_0 = n$.

2. Szukamy maksimum tej funkcji ze względu na p

$$\ell_n(p) = L_1 \ln p + L_0 \ln(1-p).$$

$$0 = \frac{d\ell_n(p)}{dp} = \frac{L_1}{p} - \frac{L_0}{1-p}.$$

3. Wynik to:

$$\hat{p}_n \stackrel{\text{pencil}}{=} \frac{L_1}{L_0 + L_1} = \frac{X_1 + \cdots + X_n}{n} = \bar{X}_n = m_1.$$

Przykład: Estymator n. w. dla rozkładu $N(\mu, \sigma^2)$

$$\mathcal{F} = \{f_{\mu, \sigma^2}(x) : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\},$$

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Logarytmiczna funkcja wiarygodności

$$\ell_n(\mu, \sigma^2) = -n \ln \sqrt{2\pi} - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

$$\frac{\partial \ell_n(\mu, \sigma^2)}{\partial \mu} = 0 \xRightarrow{\text{pencil}} \hat{\mu} = \bar{X}_n.$$

$$\frac{\partial \ell_n(\mu, \sigma^2)}{\partial \sigma^2} = 0 \xRightarrow{\text{pencil}} \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Estymator n. w. dla rozkładu $\text{Unif}([0, \theta])$

$$\mathcal{F} = \{f_\theta(x) : \theta > 0\}, \quad f_\theta(x) = \begin{cases} \frac{1}{\theta} & \text{dla } x \in [0, \theta] \\ 0 & \text{w p.p.} \end{cases}$$

Funkcja wiarygodności (funkcja $\mathbb{1}$ to indykator zbioru)

$$\mathcal{L}_n(\theta) = \frac{1}{\theta^n} \mathbb{1}_{[0, \theta]} \left(\max(X_1, \dots, X_n) \right).$$

$$\hat{\theta}_n = \max(X_1, \dots, X_n)$$

Uwaga: Metoda momentów daje w tym wypadku

$$\hat{\theta}_n = 2\bar{X}_n.$$

Estymator n. w. dla rozkładu Γ

$$\mathcal{F} = \{f_{\alpha,\beta}(x) : (\alpha, \beta) \in (0, \infty) \times (0, \infty)\},$$

$$f_{\alpha,\beta}(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},$$

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du, \quad \alpha > 0.$$

Logarytmiczna funkcja wiarygodności:

$$\ell_n(\alpha, \beta) = (\alpha - 1) \sum_{k=1}^n \ln X_k - \sum_{k=1}^n X_k / \beta - n\alpha \ln \beta - n \ln \Gamma(\alpha).$$

$$\frac{\partial \ell_n}{\partial \beta} = 0 \implies \hat{\beta} = \frac{\bar{X}_n}{\hat{\alpha}}.$$

$$\frac{\partial \ell_n}{\partial \alpha} = 0 \implies \overline{(\ln X)_n} - \ln \bar{X}_n = \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \ln \hat{\alpha}.$$

Powyższe równanie rozwiązuje się numerycznie ...

Definicja

Obciążenie estymatora $\hat{\theta}_n$ parametru θ to wielkość

$$\mathbb{E}\hat{\theta} - \theta.$$

- ▶ Estymator $\hat{\theta}_n$ parametru θ nazywany jest **estymatorem nieobciążonym**, jeśli jego obciążenie jest zerowe, czyli

$$\mathbb{E}\hat{\theta} = \theta.$$

Przykład: $\mathcal{F} = \{\text{Unif}([0, \theta]): \theta > 0\}$

- ▶ Estymator (parametru θ) m. m. jest nieobciążony:

$$\mathbb{E}\hat{\theta}_n = \mathbb{E}(2\bar{X}_n) = \frac{2}{n}\mathbb{E}(X_1 + \dots + X_n) = \frac{2}{n}n\frac{\theta}{2} = \theta.$$

- ▶ Estymator (parametru θ) n.w. jest obciążony:

$$\mathbb{E}\hat{\theta}_n = \mathbb{E}(\max(X_1, \dots, X_n)) < \theta.$$

- ▶ Czy to znaczy, że estymator m. m. jest w tym wypadku lepszy?

Własności estymatorów n. w.

Twierdzenie

Estymator n. w. dla parametru θ (pod pewnymi założeniami co do regularności modelu) jest

- ▶ *zgodny*, tzn.

$$\hat{\theta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \theta,$$

- ▶ *asymptotycznie normalny*, tzn. (przypadek skalarne parametru)

$$\sqrt{n} \left(\hat{\theta}_n - \theta \right) \xrightarrow[n \rightarrow \infty]{d} N(0, \sigma_{MLE}^2),$$

- ▶ *ekwiwariantny*, tzn. jeśli $\hat{\theta}_n$ jest estymatorem n. w. dla θ , to $g(\hat{\theta}_n)$ jest estymatorem n. w. dla $g(\theta)$.
- ▶ *asymptotycznie optymalny* (asymptotycznie efektywny)

Zajmiemy się teraz wyznaczaniem σ_{MLE}^2 .

Funkcja informacji Fishera

$$\mathcal{F} = \{f_{\theta}(x) : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^k,$$

X_{θ} zmienna losowa o rozkładzie zadany fun. gęst. $f_{\theta}(x)$.

Definicja (Funkcja informacji Fishera)

Funkcją informacji Fishera (informacją Fishera) dla rodziny \mathcal{F} ze skalarnym ($k = 1$) parametrem θ nazywamy odwzorowanie

$$\Theta \ni \theta \mapsto I(\theta) = \mathbb{E} (\ell'(\theta))^2 = \mathbb{E} \left(\frac{\partial \ln f_{\theta}(X_{\theta})}{\partial \theta} \right)^2 = \mathbb{E} \left(\frac{f'_{\theta}(X_{\theta})}{f_{\theta}(X_{\theta})} \right)^2.$$

Uwaga:

W przypadku wektorowego parametru θ ($k > 1$) $I(\theta)$ jest macierzą o i, j -tym elemencie będącym funkcją określonym przez

$$\mathbb{E} \left(\frac{\partial \ell(\theta)}{\partial \theta_i} \frac{\partial \ell(\theta)}{\partial \theta_j} \right).$$

Fakt

Przy pewnych założeniach

$$\mathbb{E} (\ell''(\theta)) = \mathbb{E} \left(\frac{\partial^2 \ln f_{\theta}(X_{\theta})}{\partial \theta^2} \right) = -I(\theta).$$

Dowód $\mathbb{E} (\ell''(\theta)) = -I(\theta)$.

$$\ell'(\theta) = \frac{f'_{\theta}(X_{\theta})}{f_{\theta}(X_{\theta})}, \quad \ell''(\theta) = \frac{f''_{\theta}(X_{\theta})}{f_{\theta}(X_{\theta})} - \frac{(f'_{\theta}(X_{\theta}))^2}{f_{\theta}^2(X_{\theta})}, \quad ' = \frac{\partial}{\partial \theta}.$$

$$\int f_{\theta}(x) dx = 1 \implies \int \frac{\partial^k f_{\theta}(x)}{\partial \theta^k} dx = 0 \implies \int f''_{\theta}(x) dx = 0.$$

$$\begin{aligned} \mathbb{E} (\ell''(\theta)) &= \int \left(\frac{f''_{\theta}(x)}{f_{\theta}(x)} - \frac{(f'_{\theta}(x))^2}{f_{\theta}^2(x)} \right) f_{\theta}(x) dx \\ &= 0 - \mathbb{E} (\ell'(\theta)^2) = -I(\theta). \end{aligned}$$



Twierdzenie o zbieżności estymatorów n. w.

Twierdzenie (Asymptotyczna normalność est. n. w.)

Jeśli $\hat{\theta}_n$ jest estymatorem n. w., to

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} N \left(0, \frac{1}{I(\theta)} \right).$$

Równoważnie:

$$\sqrt{n I(\theta)} (\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, 1).$$

Co więcej

$$\sqrt{n I(\hat{\theta})} (\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, 1).$$

Więc σ_{MLE}^2 to odwrotność informacji Fishera.

Przykład: Rozkład Bern(p)

Funkcja gęstości prawdopodobieństwa

$$f_p(x) = p\mathbb{1}_{\{1\}}(x) + (1-p)\mathbb{1}_{\{0\}}(x) = p^x(1-p)^{1-x}.$$

$$\ell'(p) = (x \ln p + (1-x) \ln(1-p))' = \frac{x}{p} - \frac{1-x}{1-p}.$$

$$\ell''(p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

$$I(p) = -\mathbb{E}\ell''(p) = \frac{\mathbb{E}X}{p^2} + \frac{1-\mathbb{E}X}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

Przykład: Rozkład $\text{Exp}(\lambda)$

Funkcja gęstości prawdopodobieństwa

$$f_{\lambda}(x) = \lambda e^{-\lambda x} \quad \text{dla } x \geq 0.$$

$$\ell(\lambda) = \ln \lambda - \lambda x, \quad \ell''(\lambda) = -\frac{1}{\lambda^2}.$$

$$I(\lambda) = -\mathbb{E} \left(-\frac{1}{\lambda^2} \right) = \frac{1}{\lambda^2}.$$

Przykład: Rozkład Pois(λ)

Funkcja gęstości prawdopodobieństwa

$$f_{\lambda}(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad x = 0, 1, \dots$$

$$\ell(\lambda) = x \ln \lambda - \lambda - \ln x!, \quad \ell''(\lambda) = -\frac{x}{\lambda^2}.$$

$$I(\lambda) = -\mathbb{E} \left(-\frac{X}{\lambda^2} \right) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Nierówność Crámera-Rao

Niech X_1, \dots, X_n próba losowa (zmienne niezależne o tym samym rozkładzie odpowiadającym nieznannej wartości parametru θ),
 $S = S(X_1, \dots, X_n)$ statystyka oraz

$$m(\theta) = \mathbb{E}S(X_1, \dots, X_n).$$

Twierdzenie (Crámera-Rao)

$$\mathbb{V}(S) \geq \frac{(m'(\theta))^2}{nI(\theta)},$$

przy czym równość wtw, gdy

$$S(X_1, \dots, X_n) = m(\theta) + t(\theta)\ell'_n(X_1, \dots, X_n; \theta),$$

dla pewnej funkcji $t(\theta)$.

Przypomnienie: wariancja S to $\mathbb{V}(S) = \mathbb{E}(S - m(\theta))^2$.

Definicja

Niech $S = S(X_1, \dots, X_n)$ będzie statystyką oraz $m(\theta) = \mathbb{E}S$.
 S nazywamy **efektywnym estymatorem** wielkości $m(\theta)$, jeśli

$$\mathbb{E}(S - m(\theta))^2 = \frac{(m'(\theta))^2}{nI(\theta)},$$

tzn. w nierówności Crámera-Rao zachodzi równość.

Pokazaliśmy, że S jest estymatorem efektywnym $m(\theta)$ wtw, gdy istnieje funkcja $t(\theta)$ taka, że

$$S(X_1, \dots, X_n) = m(\theta) + t(\theta)\ell'_n(X_1, \dots, X_n; \theta).$$

Estymator efektywny wielkości $m(\theta)$ nie musi istnieć!

Asymptotyczna efektywność estymatorów n. w.

Niech $\hat{\theta}$ będzie nieobciążonym ($E\hat{\theta} = \theta$) estymatorem największej wiarygodności.

- ▶ Nierówność Crámera-Rao ma w tym przypadku postać:

$$V(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

- ▶ Estymatora efektywny to taki, który spełnia:

$$V(\hat{\theta}) = \frac{1}{nI(\theta)}$$

- ▶ Wiemy również, że

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n \rightarrow \infty]{d} N\left(0, \frac{1}{I(\theta)}\right).$$

Wariancja lewej strony to $V(\sqrt{n}(\hat{\theta} - \theta)) = nV(\hat{\theta})$ więc

$$nV(\hat{\theta}) \rightarrow \frac{1}{I(\theta)}.$$

Ta własność estym. n.w. to **asymptotyczna efektywność**.

Definicja

Rodzina rozkładów $\mathcal{F} = \{f_{\theta}(x) : \theta \in \Theta \subset \mathbb{R}\}$ nazywana jest **rodziną wykładniczą**, jeśli funkcja gęstości prawdopodobieństwa (funkcja prawdopodobieństwa w przypadku dyskretnym) daje się przedstawić w postaci

$$f_{\theta}(x) = a(\theta)b(x)e^{c(\theta)d(x)}.$$

Wykładnicze rodziny rozkładów tworzą np. rozkłady: Gaussa, Poissona, Bernoulliego, dwumianowy, wielomianowy, wykładniczy, gamma,...

$$f_{\theta}(x) = a(\theta)b(x)e^{c(\theta)d(x)}.$$

Fakt

Dla wykładniczej rodziny rozkładów statystyka

$$S = \frac{1}{n} \sum_{i=1}^n d(X_i)$$

jest efektywnym estymatorem wielkości

$$m(\theta) = \mathbb{E}S = -\frac{a'(\theta)}{a(\theta)c'(\theta)}.$$

$f_{\theta}(x) = a(\theta)b(x)e^{c(\theta)d(x)}$. Chcemy pokazać, że $S = \frac{1}{n} \sum_{i=1}^n d(X_i)$ jest efektywnym estymatorem $m(\theta) = \mathbb{E}S = -\frac{a'(\theta)}{a(\theta)c'(\theta)}$.

Musimy pokazać:

$$\mathbb{E}S \stackrel{(\star)}{=} -\frac{a'(\theta)}{a(\theta)c'(\theta)} = m(\theta) \text{ oraz } S = m(\theta) + t(\theta)\ell'_n(\theta).$$

$$\ell'(x; \theta) = \frac{\partial}{\partial \theta} (\ln a(\theta) + \ln b(x) + c(\theta)d(x)) = \frac{a'(\theta)}{a(\theta)} + c'(\theta)d(x).$$

Stąd $\ell'_n(\theta) = n \frac{a'(\theta)}{a(\theta)} + c'(\theta) \sum_{i=1}^n d(X_i)$, czyli

$$S(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n d(X_i) = \frac{1}{nc'(\theta)} \ell'_n(X_1, \dots, X_n; \theta) - \frac{a'(\theta)}{a(\theta)c'(\theta)}.$$

(\star) dostajemy dzięki $\mathbb{E}\ell'_n = 0$

Rodzina

$$\mathcal{F} = \{f_{\theta}(x) : \theta \in \Theta \subset \mathbb{R}^k\}$$

$$\theta = (\theta_1, \dots, \theta_k)$$

jest nazywana wykładniczą rodziną rozkładów jeśli

$$f_{\theta}(x) = a(\theta)b(x)e^{\sum_{i=1}^k c_i(\theta_i)d_i(x)}.$$

W tym kontekście dostajemy zestaw estymatorów efektywnych

$$S_i = \frac{1}{n} \sum_{i=1}^k d_i(X), \quad \mathbb{E}S_i = -\frac{\frac{\partial a}{\partial \theta_i}}{a(\theta) \frac{\partial c_i}{\partial \theta_i}}.$$

Przykład efektywnego estymatora

Rozważmy rodzinę rozkładów Poissona:

$$f_{\lambda}(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots$$

$$f_{\lambda}(x) = \underbrace{e^{-\lambda}}_{a(\lambda)} \underbrace{\frac{1}{x!}}_{b(x)} \exp(\underbrace{\ln \lambda}_{c(\lambda)} \underbrace{x}_{d(x)}).$$

$$S = \frac{1}{n} \sum_{i=1}^n d(X_i) = \bar{X}_n$$

jest zatem estymatorem efektywnym wielkości

$$m(\lambda) = \mathbb{E}S = -\frac{a'(\lambda)}{a(\lambda)c'(\lambda)} = \frac{-(-e^{-\lambda})}{e^{-\lambda}(1/\lambda)} = \lambda.$$

Modele i wnioskowanie statystyczne

Wprowadzenie do R

Konrad Jędrzejewski

ISE PW

24 października 2012

Informacje ogólne

- R jest pakietem (środowiskiem) przeznaczonym do zaawansowanych obliczeń statystycznych.
- Licencja GNU GPL – całkowicie bezpłatny.
- Źródła: <http://www.r-project.org>.
- CRAN (*Comprehensive R Archive Network*).
- Platformy: Windows, Linux, Unix, MacOS.
- Język R jest językiem interpretowanym, a nie kompilowanym.
- R Commander.

Środowisko

- Konsola
 - >
 - +
- Rozróżnialność małych wielkich liter
- Mechanizm „strzałek” – poprzednie komendy
- Pomoc
 - **help**(nazwa),
 - ~~?nazwa~~
 - **apropos**(nazwa)
 - ??nazwa

Środowisko

- Funkcje – wywołanie.
 - nazwa(arg1, arg2, arg3 = wartość)
 - `x = 1:100; y = rnorm(100); plot(x, y, type = "l")`
 - **args**(legend)
 - **example**(legend)
- # komentarz
- Przydatne funkcje
 - **ls()**, **objects()**
 - **rm()**
 - **print**("napis")
 - **print**(dane)

Środowisko - pakiety

- Pakiety
 - **library()** – lista pakietów zainstalowanych
 - **search()** – lista pakietów załadowanych
- Załadowanie pakietu
 - **library(nazwa_pakietu)**
 - **require(nazwa_pakietu)**
- Usunięcie pakietu (z pamięci)
 - **detach(package:nazwa_pakietu)**

Własne skrypty, funkcje

- Katalog roboczy
 - **getwd()**, **setwd("nazwa_katalogu")**
 - **dir()**, **list.files()**
- Uruchamianie własnych skryptów
 - **source("nazwa_skryptu")**
- Dołączanie własnych funkcji
 - **source(„nazwa_pliku")**
 - **ls()**
 - „funkcja1_z_nazwa_pliku” „funkcja2_z_nazwa_pliku”
 - **rm("funkcja1_z_nazwa_pliku")**

Wczytywanie i zapisywanie danych

- Pobieranie z plików:
 - dane = **scan**('c:/plik.txt')
 - dane = **read.table**('plik.txt', **header** = T)
 - dane = **read.csv**('Zeszyt1.csv', **sep** = ";" **header** = T, **dec** = ',')
 - **names**(dane), **rownames**(dane), **dimnames**(dane)
 - **write**(x, 'plik.txt')
 - **write.table**(dane, **file** = 'plik.txt'), **write.csv**()
- Edycja/zmiana danych
 - **edit**(dane), **fix**(dane)
- Pobieranie z pakietów
 - **data**(uspop, **package** = 'datasets')

Typy danych

- Numeryczny
 - 2.345
 - 3.5e-15
- Znakowy
 - 'a', "abc", \n, \t
- Zespólny
 - $x = 3 + 4i$ **Mod(x), Arg(x), Re(x), Im(x)**
- Logiczny
 - TRUE\T, FALSE\F

Struktury danych

- Wektor
- Tablica/macierz
- Faktor (factor)
- Lista
- Ramka (dataframe)

Wektor

- Tworzenie
 - `x = c(1,2,3,4)`, `x <- c(1,2,3,4)`
 - `x = c("bdb","db","dst","bdb")`
 - `x = c(TRUE,FALSE,TRUE,FALSE)` ; `y = c(T,F,T,F)`
- Indeksowanie wektorów
 - `x[3]`
 - `x[2:4]`
 - `x[c(2,5,8)]`
- Operacje arytmetyczne

Generowanie wektorów

- `:`
 - `x = 1:100`
 - `x = 100:1`
- `seq()`
 - `x = seq(0, 5, by = 0.25)`
 - `x = seq(0, 5, length = 10)`
- `rep()`
 - `x = rep(c(1,2,3), 4)`
 - `x = rep(c(1,2,3), each = 4)`

Tablica

- Tablica (array) jest wektorem zawierającym dodatkowe dane określające uporządkowanie elementów w tablicy.
- **dim()**
 - `x = 1:20`
 - `dim(x) = c(4,5)`
 - `attributes(x)`
 - `dimnames(x) = list(letters[1:4],LETTERS[1:5])`

Tablica

- **matrix()**
 - **matrix**(1:20, 4, 5)
- **array()**
 - **array**(1:20,c(4,5))
- **rbind(), cbind()**
 - **x = rbind**(1:3,4:6); **y = cbind**(1:3, 4:6)
- Mnożenie macierzy
 - **z = x%*%y**
- Indeksowanie
 - **x[2,3], x[1:3,1:2], x[2,], x[,3]**

Faktor (factor)

- Faktor (factor) jest strukturą przechowującą oprócz szeregu danych informacje o powtórzeniach takich samych wartości oraz zbiorze unikalnych wartości.
- **factor()**
 - faktor = **factor**(c(2,3,4), levels=1:5)
 - punkty = **factor**(c(95,56,74,80,52,99,35,74))
 - oceny = **factor**(c("bdb","dst","db","dst","bdb","ndst","db"))
- **levels()**
 - **levels**(oceny)
- **table()**
 - **table**(oceny)

Lista

- Lista (list) jest uporządkowanym zbiorem elementów różnego typu.
 - Lista = **list**("Jan","Kowalski",1990,"Warszawa","TRUE")
 - Lista = **list**(imie="Jan",nazwisko="Kowalski",rok_ur=1990,zam="Warszawa",stud="TRUE")
- Wybór z listy
 - Lista\$nazwisko
 - Lista[2][1]
- Dodawanie
 - Lista\$imie[2] = „Jakub”; Lista\$nazwisko[2] = "Nowak"; ...
 - Lista2 = **list**(imie=c("Jan","Piotr"),nazwisko=c("Kowalski","Nowak"),rok_ur=c(1991,1995),zam=c("Warszawa","Poznan"),stud=c(T,F))

Ramka

- Ramka (dataframe) jest macierzą, w której poszczególne kolumny mogą zawierać wartości różnego typu.
- Tworzenie
 - `ramka = data.frame(LETTERS[1:6], seq(10,60, by = 10), seq(10,60, by = 10) > 35)`
 - `names(ramka) = c("Litera", "Punkty", " Punkty > 35")`
- Wybór z ramki
 - `ramka[3,] ; ramka[,2]`
 - `ramka$Punkty`
 - `ramka$Litera[2]`
 - `ramka$"Punkty > 35"`

Generowanie liczb losowych

- **sample()**
 - **sample(1:6, 4, replace = T)**
- Rozkłady zmiennych – binom, geom, hyper, pois, norm, unif, exp, chisq, f, t, beta, gamma
- Przedrostki: r, d, p, q
 - **x = rnorm(100, mean = 2, sd = 3)**
 - **dnorm(0, mean = 0, sd = 1) ; dunif(0.4, min=0, max=2)**
 - **pnorm(0, mean = 0, sd = 1) ; punif(0.4, min=0, max=2)**
 - **qnorm(0.95, mean = 0, sd = 1); qunif(0.4, min=0, max=2)**
- Podstawowe funkcje
 - **mean(), sd(), var(), median(), quantile()**

Wykresy

- Podstawowe funkcje
 - **plot()**,
 - **plot(x,y, xlab = "opis x", ylab = "opis y", main = "tytul")**
 - kolejne dane na wykresie - **points()**, **lines()**
 - **hist(rnorm(1000))**
 - **pie(1:6, labels = LETTERS[1:6])**
 - **grid()**, **title()**, **legend()**
- Zarządzanie oknami
 - **dev.new()**, **x11()**, **dev.off()**, **dev.cur()**, **dev.set(nr_device)**

Wykresy

- Zapisywanie
 - **> jpeg('rys.jpg')**
 - > plot(x,y)**
 - > dev.off()**
 - **> plot(x,y)**
 - > dev.copy(png, 'rys.png')**
 - > dev.off()**
 - **> dev.print(pdf, 'rys.pdf')**

Programowanie w R

- Instrukcja warunkowa
- Pętle
- Skrypty
- Funkcje

Instrukcja warunkowa

- **if**(warunek) wyrażenie
- **if**(warunek) wyrażenie1 **else** wyrażenie2
- **ifelse**(warunek, a, b)
- **switch**(zmienna, wartosc1 = akcja1, wartosc2 = akcja2, ...)
- Operatory logiczne: **&**, **|**, **!**, **xor(x,y)**, **==**, **!=**, **<**, **>**, **<=**, **<=**, **isTRUE(x)**, **&&**, **||**.

Pętle

- **for**(licznik **in** start:koniec) wyrażenie
- **while**(warunek) wyrażenie
- **repeat** wyrażenie
- **break**
- **next**
- Przykłady

Skrypty

- **source("nazwa_skryptu.R")**

- Przykład

Skrypt1.R

```
x = 1:100
```

```
y = x^2
```

```
plot(x, y, type="l")
```

```
grid()
```

```
> source("Skrypt1.R")
```


Funkcje

- nazwa_funkcji = **function**(arg1, arg2, arg3 = wartość)
{ciało funkcji}
- Zwracane wartości - ostatnia linia
- **return()**
- Przeciążanie funkcji
- **stop()**
- **warning()**

Przedziały ufności

MWS, wykład 5

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 13 kwietnia 2021

Przedziały ufności

Przedziały ufności

X_1, \dots, X_n próba losowa odp. pewnej (nieznanej) wartości parametru $\theta \in \Theta$.

Definicja

Przedziałem ufności na poziomie ufności γ dla parametru θ nazywamy przedział $\left[A(X_1, \dots, X_n), B(X_1, \dots, X_n) \right]$, dla którego zachodzi nierówność

$$\mathbb{P} \left(\theta \in \left[A(X_1, \dots, X_n), B(X_1, \dots, X_n) \right] \right) \geq \gamma.$$


Uwagi:

- ▶ w praktyce zazwyczaj chodzi o równość
- ▶ przedział ten jest losowy (tak samo jak estymator punktowy)
- ▶ poziom ufności wyraża się często przez $1 - \alpha$ zamiast γ

Sposoby wyznaczania przedziałów ufności

..czyli spis treści wykładu

Przedziały ufności dla estymatora można wyznaczać:

- 
- ▶ W sposób dokładny wykorzystując znajomość analitycznej postaci rozkładu estymatora,
 - ▶ trzeba ją najpierw wyznaczyć, a to bywa trudne
 - ▶ W sposób przybliżony wykorzystując graniczny (normalny) rozkład estymatora,
 - ▶ już łatwiej, choć nieco mniej dokładnie
 - ▶ W sposób przybliżony przez symulacje (tzw. bootstrap)
 - ▶ jeszcze łatwiej.

Przykład wykorzystania rozkładu granicznego


Liczba zgonów od kopnięcia konia dla 10 korpusów pruskiej kawalerii zebrana z 20-letniego okresu (mamy 200 „korpuso-lat”).

Wyznaczanie przedziału uf. z wykorzystaniem rozkładu granicznego

ilość zgonów/rok	0	1	2	3	4
liczba „korpuso-lat”	109	65	22	3	1

$$\hat{\lambda} = (109 \times 0 + 65 \times 1 + 22 \times 2 + 3 \times 3 + 1 \times 4)/200 = 0.61$$

$$\sqrt{n}(\hat{\lambda} - \lambda) \approx N(0, 1/I(\hat{\lambda})) = N(0, \hat{\lambda}) = \sqrt{\hat{\lambda}}N(0, 1)$$

Przedział ufności dla λ na poziomie ufności $\gamma = 1 - \alpha$ można zatem przybliżyć jako 

$$\begin{aligned} & \left[\hat{\lambda} - z_{(1+\gamma)/2} \sqrt{\hat{\lambda}/n}, \hat{\lambda} + z_{(1+\gamma)/2} \sqrt{\hat{\lambda}/n} \right] \\ &= \left[\hat{\lambda} - z_{1-\frac{\alpha}{2}} \sqrt{\hat{\lambda}/n}, \hat{\lambda} + z_{1-\frac{\alpha}{2}} \sqrt{\hat{\lambda}/n} \right] \stackrel{\gamma=0.95}{\approx} [0.50, 0.72]. \end{aligned}$$

Bootstrap parametryczny

- ▶ Mamy do dyspozycji próbę, na podstawie której estymujemy parametr rozkładu $\hat{\theta}$. Chcemy zbadać jakie są własności (np. wariancja, kwantyle, rozkład) tego estymatora lub innej statystyki.
- ▶ Nie dysponując dodatkowymi danymi próbujemy wykorzystać to co mamy (bootstrapping).
- ▶ W przypadku modelu parametrycznego generujemy wiele losowych prób (liczności takiej jak oryginalna próba), dla każdej z nich wyznaczamy wartość badanej statystyki i badamy rozkład/wariancję/kwantyle/itp. tych wartości.
- ▶ Żeby wygenerować próby musimy znać parametr θ rozkładu!
- ▶ Zamiast θ wykorzystujemy (istota bootstrapu parametrycznego) estymatę $\hat{\theta}$ uzyskaną na podstawie oryginalnych danych.

Wyznaczania przedziału uf. metodą bootstrapu parametrycznego

ilość zgonów/rok	0	1	2	3	4
liczba „korpuso-lat”	109	65	22	3	1

$$\hat{\lambda} = 0.61$$

$$[y_1, \dots, y_n] = \text{sample}([x_1, \dots, x_n])$$

- ▶ Generujemy $N \gg 1$ zestawów po $n = 200$ liczb z rozkładu $\text{Pois}(\hat{\lambda})$.
- ▶ Dla każdego zestawu wyliczamy wartość estymatora.
- ▶ Wyznaczamy przedział ufności $[A, B]$ (na poziomie ufności γ) tak, aby obejmował $\gamma \times N$ spośród wyznaczonych N wartości estymatora.
- ▶ Dla $\gamma = 0.95$, $N = 1000$, $\hat{\lambda} = 0.61$ w wyniku symulacji otrzymujemy

$$[A, B] \approx [0.45, 0.77].$$

Bootstrap nieparametryczny

- ▶ Mamy do dyspozycji próbę, na podstawie której wyznaczamy pewną statystkę T . Chcemy zbadać jakie są własności tej statystyki.
- ▶ W przypadku **braku** modelu parametrycznego generujemy wiele losowych prób (liczności takiej jak oryginalna próba) „z tego samego rozkładu co wyjściowa próba”. Dla każdej wygenerowanej próby wyznaczamy wartość statystyki T , a następnie badamy rozkład/wariancję/kwantyle/itp. tych wartości.
- ▶ Żeby wygenerować próby musimy znać rozkład!
- ▶ Nie dysponując dodatkowymi danymi próbujemy wykorzystać to co mamy (bootstrapping).
- ▶ Generujemy kolejne realizacje próby poprzez przepróbkowanie (losowanie z powtórzeniami) posiadanej jednej próby.
- ▶ Odpowiada to przyjęciu za nieznany rozkład rozkładu o dystrybuancie równej dystrybuancie (empirycznej) uzyskanej z wyjściowej próby.

Przedziały ufności dla parametrów rozkładu Gaussa

Rozkład Normalny (Gaussa) — przypomnienie

X ma **rozkład normalny (Gaussa)** z parametrami μ , σ ,
 $X \sim N(\mu, \sigma^2)$, jeśli

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

- ▶ Funkcję gęstości oraz dystrybuantę zmiennej o rozkładzie $N(0, 1)$ oznacza się czasem odpowiednio, literami ϕ oraz Φ .
- ▶ Kwantyl rzędu α dla rozkładu $N(0, 1)$ oznacza się zazwyczaj przez z_α .

$$Y = \frac{X - \mu}{\sigma} \sim N(0, 1) \quad \text{Var}(aX) = a^2 \text{Var}(X)$$

Estymacja μ gdy σ^2 znane

$$X_k = \mu + \epsilon_k, \quad k = 1, \dots, n,$$

$\epsilon_1, \dots, \epsilon_n$ niezależne zmienne losowe $\sim N(0, \sigma^2)$.

- ▶ X_k mogą modelować pomiary badanej wielkości μ , gdzie błąd pomiaru (np. związany z przyrządem) ma rozkład $N(0, \sigma^2)$.
- ▶ Mamy do czynienia z rodziną (parametryczną):

$$\mathcal{F} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R} \right\}, \quad \sigma^2 \text{ ustalone.}$$

- ▶ Jest to rodzina wykładnicza

- ▶ \bar{X}_n jest efektywnym estymatorem μ .

- ▶ Jest on nieobciążony i ma rozkład normalny,
- ▶ jego wariancja, to

$$V(\bar{X}_n) = \frac{1}{nI(\mu)} = \frac{\sigma^2}{n}.$$

X_1, \dots, X_n σ^2
 $\frac{1}{n} \sum x_n$ σ^2/n

Przedziały ufności dla μ gdy σ^2 znane

- ▶ Estymator μ ma postać

$$\hat{\mu} = \bar{X}_n$$

- ▶ ma on rozkład

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

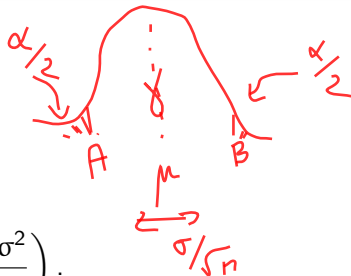
- ▶ Przedziałem ufności dla μ na poziomie ufności $\gamma = 1 - \alpha$ jest zatem (📎)

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \right].$$

- ▶ Uwaga: $1 - \alpha/2 = (1 + \gamma)/2$.

$$\bar{X} - \mu > z \cdot \frac{\sigma}{\sqrt{n}}$$

$$P(Y < z_{\alpha/2}) = \frac{\alpha}{2}$$




Estymacja σ^2 , znane μ

- ▶ Dysponujemy próbą losową

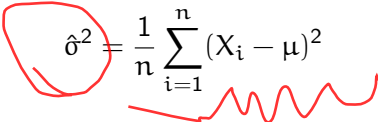
$$X_k = \mu + \epsilon_k, \quad k = 1, \dots, n, \quad \epsilon_1, \dots, \epsilon_n \text{ niezależne } \sim N(0, \sigma^2).$$

- ▶ np. X_k mogą modelować pomiary znanej wielkości (wzorca) μ , gdzie błąd pomiaru (np. związany z przyrządem) ma rozkład $N(0, \sigma^2)$.
- ▶ Mamy do czynienia z rodziną (parametryczną):

$$\mathcal{F} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \sigma^2 > 0 \right\}, \quad \mu \text{ znane.}$$

- ▶ Jest to rodzina wykładnicza .

- ▶ Efektywnym estymatorem parametru σ^2 jest


$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

- ▶ jego wariancja to

$$V(\hat{\sigma}^2) = 1/(nI(\sigma^2)) = 2\sigma^4/n.$$

Definicja

Niech X_1, \dots, X_n niezależne zmienne losowe o rozkładzie $N(0, 1)$.
Rozkładem χ^2 o n stopniach swobody (rozkładem χ_n^2) nazywamy
rozkład zmiennej losowej

$$X_1^2 + \dots + X_n^2.$$

Jaką funkcję gęstości ma rozkład χ_1^2 ?

$$\overbrace{X_1 \quad X_2} \\ f_{\text{gum}}(X_1 + X_2) = f_{\text{gum}}(x_1) \cdot f_{\text{gum}}(x_2)$$

Funkcja gęstości prawdopodobieństwa rozkładu χ_1^2

Niech $\underline{X^2} \sim \chi_1^2$.

$$X^2 \leq 4 \quad \underline{-\sqrt{4} \leq X \leq \sqrt{4}}$$

- ▶ Można obliczyć dystrybuantę tego rozkładu:

$$F_{X^2}(x) = P(X^2 \leq x) = \int_{-\sqrt{x}}^{+\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

$$P(-\sqrt{x} \leq X \leq \sqrt{x})$$

- ▶ FGP ma więc postać:

$$f_{X^2}(x) = \frac{d}{dx} F_{X^2}(x) = \dots = \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}}.$$

- ▶ Rozkład χ_1^2 jest zatem równy rozkładowi Gamma z parametrami $\alpha = \frac{1}{2}$, $\beta = 2$, czyli $\text{Gamma}(\frac{1}{2}, 2)$.
- ▶ Jaką funkcję gęstości ma rozkład χ_n^2 ?
 - ▶ Pokażemy, że

$$\chi_n^2 = \text{Gamma}\left(\frac{n}{2}, 2\right).$$

Suma niezależnych rozkładów Gamma(\cdot , β)

- ▶ Jeśli $Y \sim \text{Gamma}(\alpha, \beta)$, to

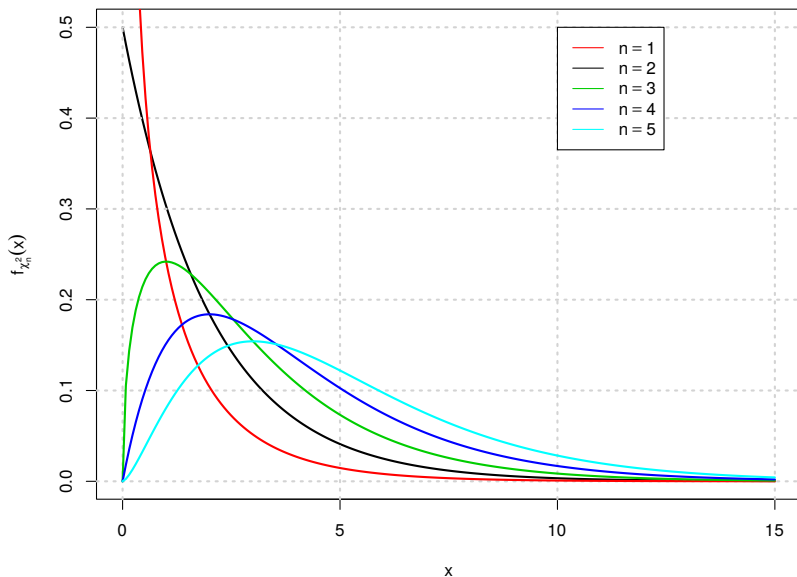
$$M_Y(t) = \mathbb{E}e^{tY} = \left(\frac{1}{1 - \beta t} \right)^\alpha.$$

- ▶ Dla niezależnych zmiennych losowych $Y_k \sim \text{Gamma}(\alpha_k, \beta)$ mamy:

$$\mathbb{E} \exp \left(t \sum_{i=1}^n Y_i \right) = \underbrace{\left(\frac{1}{1 - \beta t} \right)^{\sum_{i=1}^n \alpha_i}}.$$

- ▶ Stąd $Y_1 + \dots + Y_n \sim \text{Gamma}(\alpha_1 + \dots + \alpha_n, \beta)$.
- ▶ W szczególności $\chi_n^2 = \text{Gamma}(\frac{n}{2}, 2)$.

Rozkład χ^2 — wykresy funkcji gęstości



Przedziały ufności dla σ^2 (μ znane)

$X_k \sim N(\mu, \sigma^2)$ niezależne, μ znane.

- ▶ Estymator wariancji σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

- ▶ więc $\frac{n\hat{\sigma}^2}{\sigma^2}$ ma rozkład χ_n^2 (bo $\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^n (\frac{X_i - \mu}{\sigma})^2$.)
- ▶ Przedziałem ufności dla σ^2 na poziomie ufności γ jest zatem

$$\left[\frac{n\hat{\sigma}^2}{F_{\chi_n^2}^{-1}(1-b)}, \frac{n\hat{\sigma}^2}{F_{\chi_n^2}^{-1}(a)} \right], \quad a, b \geq 0, a + b = 1 - \gamma.$$

- ▶ Można wziąć np. $a = b = (1 - \gamma)/2$.

Estymacja σ^2 , nieznane μ i σ^2

$X_k = \mu + \epsilon_k$, $k = 1, \dots, n$, $\epsilon_1, \dots, \epsilon_n$ niezależne $\sim N(0, \sigma^2)$.

- ▶ np. X_k mogą modelować pomiary nieznanej wielkości (wzorca) μ , gdzie błąd pomiaru (np. związany z przyrządem) ma rozkład $N(0, \sigma^2)$ z nieznanymi odchyleniem standardowym σ .
- ▶ Mamy do czynienia z rodziną (parametryczną):

$$\mathcal{F} = \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$$

- ▶ Estymator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2$$

(\bar{X}_n zamiast μ) nie jest już estymatorem nieobciążonym.

- ▶ Nieobciążonym (patrz 1 wykład) i asymptotycznie efektywnym estymatorem parametru σ^2 jest za to

$$S_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2.$$

Rozkład estymatora S^2

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2, \quad X_k \sim N(\mu, \sigma^2) \text{ niezależne.}$$


- Pokażemy, że $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
 - W tym celu wprowadzimy nową zmienną W i wykorzystamy funkcję tworzącą momenty

$$W = \sum_{k=1}^n \left(\frac{X_k - \mu}{\sigma} \right)^2 = \underbrace{\frac{1}{\sigma^2} \sum_{k=1}^n (X_k - \bar{X}_n)^2}_U + \underbrace{\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right)^2}_V.$$

- Mamy $W \sim \chi^2_n$, $V \sim \chi^2_1$ oraz

$$M_U(t) = \frac{M_W(t)}{M_V(t)} = \left(\frac{1}{1-2t} \right)^{(n-1)/2}.$$

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2, \quad X_k \sim N(\mu, \sigma^2) \text{ niezależne.}$$

- ▶ Wiemy już, że $\frac{(n-1)S^2}{\sigma^2} \sim \text{Gamma}\left(\frac{n-1}{2}, 2\right) = \chi^2_{n-1}$.
- ▶ Przedziałem ufności dla σ^2 na poziomie ufności γ jest zatem 

$$\left[\frac{(n-1)S^2}{F_{\chi^2_{n-1}}^{-1}(1-b)}, \frac{(n-1)S^2}{F_{\chi^2_{n-1}}^{-1}(a)} \right], \quad a, b \geq 0, a + b = 1 - \gamma.$$

Można wziąć np. $a = b = (1 - \gamma)/2$.

Przedziały ufności dla μ , gdy σ^2 nieznane

$X_k \sim N(\mu, \sigma^2)$ niezależne.

- ▶ Estymator μ ma postać

$$\hat{\mu} = \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ czyli } \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{\sigma^2}} \sim N(0, 1).$$

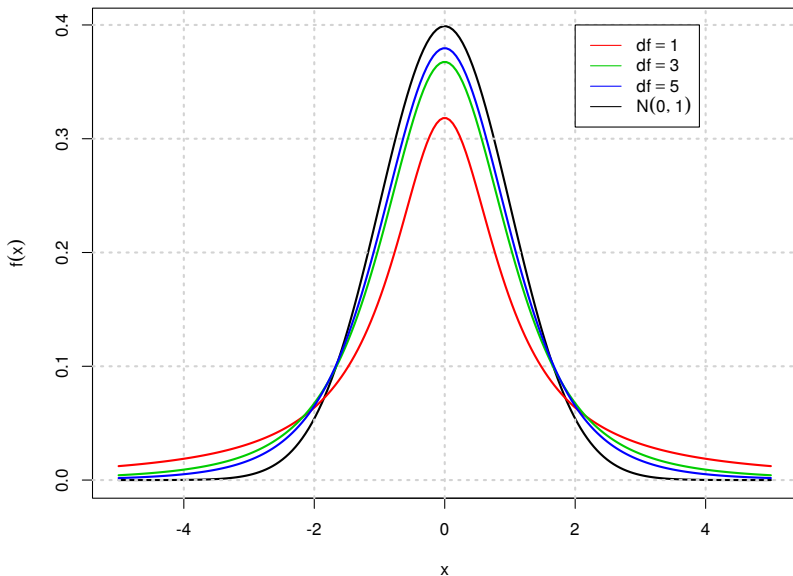
- ▶ Nie znamy σ^2 !
- ▶ Możemy za to rozważyć zmienną

$$t_{n-1} = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sqrt{S^2}}.$$

- ▶ Rozkład tej zmiennej nazywany jest rozkładem Studenta (t-Studenta) o $n-1$ stopniach swobody.
- ▶ Przedziałem ufności dla μ na poziomie ufności γ jest zatem (📎)

$$\left[\bar{X}_n - \frac{S}{\sqrt{n}} F_{t_{n-1}}^{-1} \left(\frac{1+\gamma}{2} \right), \bar{X}_n + \frac{S}{\sqrt{n}} F_{t_{n-1}}^{-1} \left(\frac{1+\gamma}{2} \right) \right].$$

Rozkład t-Studenta — wykresy funkcji gęstości



Estymacja Bayesowska

MWS, wykład 6

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 20 kwietnia 2021

Czym właściwie jest prawdopodobieństwo

- ▶ **prawdopodobieństwo obiektywne** nazywane także prawdopodobieństwem w sensie częstościowym.
 - ▶ Jeśli prawdopodobieństwo (częstościowe) wyrzucenia orła pewną monetą wynosi $\frac{1}{2}$, to

$$\lim_{n \rightarrow \infty} \frac{\text{liczba orłów w pierwszych } n \text{ rzutach}}{n} = \frac{1}{2}.$$

- ▶ **prawdopodobieństwo subiektywne** nazywane także prawdopodobieństwem w sensie bayesowskim.
 - ▶ W tym sensie prowadzący przedmiot może stwierdzić, że np. student A zaliczy przedmiot MWS z prawdopodobieństwem 80%, a student B — z prawdopodobieństwem 99%. Studenci A i B mogą oceniać te prawdopodobieństwa inaczej. Co więcej prawdopodobieństwa te mogą się zmieniać wraz z czasem (napływ nowych informacji).

Prawdopodobieństwo Bayesowskie spełnia wszystkie aksjomaty prawdopodobieństwa:

- ▶ $\mathbb{P}(A) \geq 0$ dla każdego zdarzenia A ,
- ▶ $\mathbb{P}(\Omega) = 1$,
- ▶ Dla każdych rozłącznych zdarzeń A_1, A_2, A_3, \dots

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Prawdopodobieństwo warunkowe

W podejściu Bayesowskim kluczową rolę gra prawdopodobieństwo warunkowe:

- ▶ $\mathbb{P}(A)$, to nasze przekonanie, że prawdziwe jest A .
W szczególnych przypadkach można interpretować $\mathbb{P}(A)$ jako prawdopodobieństwo a priori (przed wykonaniem eksperymentu/pomiaru).
- ▶ $\mathbb{P}(A|B)$, to nasze przekonanie, że prawdziwe jest A pod warunkiem, że zachodzi B . Można je interpretować jako prawdopodobieństwo a posteriori (po wykonaniu eksperymentu/pomiaru, którego wynik opisywany jest przez B).

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

$$P(A) P(B|A)$$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B) \mathbb{P}(B)}{\mathbb{P}(A)}$$

Prawdopodobieństwo warunkowe a intuicja

Przykład 1

Mamy do dyspozycji 3 monety:

1. monetę orzeł-reszka (O/R),
2. monetę orzeł-orzeł (O/O),
3. monetę reszka-reszka (R/R).

Losowo wybieramy monetę i rzucamy nią dostając orła. Jakie jest prawdopodobieństwo tego, że z drugiej strony jest też orzeł?

Prawdopodobieństwo warunkowe a intuicja

Przykład 1

Mamy do dyspozycji 3 monety:

1. monetę orzeł-reszka (O/R),
2. monetę orzeł-orzeł (O/O),
3. monetę reszka-reszka (R/R).

Losowo wybieramy monetę i rzucamy nią dostając orła. Jakie jest prawdopodobieństwo tego, że z drugiej strony jest też orzeł?

$$\frac{2}{3} \quad \left(\text{🖊} \right)$$

Prawdopodobieństwo warunkowe a intuicja

Przykład 2

W pewnej zabawie mamy do wyboru 3 drzwi. Za jednymi z nich stoi samochód, który wygramy wybierając te drzwi. Za pozostałymi stoją kozy. Wybieramy jedno z drzwi (nie zaglądamy za nie), a następnie prowadzący zabawę otwiera jedno z pozostałych drzwi ukazując kozę. Następnie zadaje pytanie, czy zmieniamy nasz wybór. Zmienić wybór/ nie zmienić / wszystko jedno?



Prawdopodobieństwo warunkowe a intuicja

Przykład 2

W pewnej zabawie mamy do wyboru 3 drzwi. Za jednymi z nich stoi samochód, który wygramy wybierając te drzwi. Za pozostałymi stoją kozy. Wybieramy jedno z drzwi (nie zaglądamy za nie), a następnie prowadzący zabawę otwiera jedno z pozostałych drzwi ukazując kozę. Następnie zadaje pytanie, czy zmieniamy nasz wybór. Zmienić wybór/ nie zmienić / wszystko jedno?



Zmienić wybór! (wygramy z prawd. $\frac{2}{3}$; gdy nie zmienimy mamy szanse $\frac{1}{3}$.).

Problemy z właściwym szacowaniem prawdopodobieństw warunkowych

Przykład (Eddy, 1982)

Przed 100 lekarzami postawiono następujący problem dotyczący analizowania wyników mammograficznych badań przesiewowych:

- ▶ W przypadku braku pewnych dodatkowych informacji, prawdopodobieństwo, że kobieta (w odp. wieku i kondycji) ma raka piersi wynosi 1%.
- ▶ Jeśli pacjentka ma raka piersi, to prawdopodobieństwo, że radiolog na podstawie badania postawi właściwą diagnozę wynosi 80%.
- ▶ Jeśli pacjentka ma zmiany nienowotworowe, to prawdopodobieństwo, że radiolog zdiagnozuje raka wynosi 10%.

Jakie jest prawdopodobieństwo, że pacjentka z pozytywnym wynikiem mammografii ma istotnie raka piersi?

Problemy z właściwym szacowaniem prawdopodobieństw warunkowych

Przykład (Eddy, 1982)

Przed 100 lekarzami postawiono następujący problem dotyczący analizowania wyników mammograficznych badań przesiewowych:

- ▶ W przypadku braku pewnych dodatkowych informacji, prawdopodobieństwo, że kobieta (w odp. wieku i kondycji) ma raka piersi wynosi 1%.
- ▶ Jeśli pacjentka ma raka piersi, to prawdopodobieństwo, że radiolog na podstawie badania postawi właściwą diagnozę wynosi 80%.
- ▶ Jeśli pacjentka ma zmiany nienowotworowe, to prawdopodobieństwo, że radiolog zdiagnozuje raka wynosi 10%.

Jakie jest prawdopodobieństwo, że pacjentka z pozytywnym wynikiem mammografii ma istotnie raka piersi?

- ▶ 95 na 100 lekarzy oszacowało to prawdopodobieństwo na około 75%.
- ▶ Wynosi ono jednak tylko 7.5%

Twierdzenie (Reguła Bayesa)

Jeśli A jest pewnym zdarzeniem oraz rozłączne zdarzenia B_1, \dots, B_n o niezerowym prawdopodobieństwie pokrywają całą przestrzeń probabilistyczną Ω ($\bigcup_{i=1}^n B_i = \Omega$), to

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

$P(A)$

wersja „ciągła”:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x)f_X(x)dx}.$$

- ▶ Parametryczny model:

$$\mathcal{F} = \{f_p(x) : p \in \Theta\}, \quad \Theta \subset \mathbb{R}^k,$$

z parametrem $p = (p_1, \dots, p_k)$

- ▶ Zamiast funkcji gęstości można rozważać funkcje prawdopodobieństwa lub dystrybuanty.
- ▶ W podejściu **bayesowskim** zakładamy, że parametr p jest **zmienną losową** o pewnym rozkładzie zadany np. funkcją prawdopodobieństwa:

- ▶ **a priori**

→ przed dyspensowaniem

$$f_{\text{prior}} = f_p : \Theta \rightarrow \mathbb{R}, \quad p \mapsto f(p)$$

wynikającą z naszej wiedzy i doświadczenia,

- ▶ **a posteriori**

$$f_{\text{post}} = f_{p|X} : \Theta \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad (p, x_1, \dots, x_n) \mapsto f_{\text{post}}(p, x_1, \dots, x_n)$$

wynikającą z rozkładu **a priori** oraz zaobserwowanej próby

$$X = (X_1, \dots, X_n).$$

Konstrukcja rozkładu a posteriori

- ▶ Model parametryczny $\mathcal{F} = \{f_p(x) : p \in \Theta\}$.

- ▶ Zgodnie z regułą Bayesa

$$\begin{aligned} \underline{f_{\text{post}}(p, x)} &= \underline{f_{p|x}(p, x)} = \frac{\overbrace{f_{x|p}(x, p)}^{\text{funkcja wiarygodności}} \overbrace{f_p(p)}^{\text{a priori}}}{\text{"}f_X(x)\text{"}} = \frac{f_p(x)f_p(p)}{\int_{\Theta} f_p(x)f_p(p)dp} \\ &= \frac{f_p(x)f_{\text{prior}}(p)}{\int_{\Theta} f_p(x)f_{\text{prior}}(p)dp}, \end{aligned}$$

gdzie

$$f_p(x) = f_p(x_1)f_p(x_2) \dots f_p(x_n) = \mathcal{L}_n(x_1, \dots, x_n; p)$$

- ▶ gęstość a posteriori \propto fun. wiarygodności \times gęst. a priori.

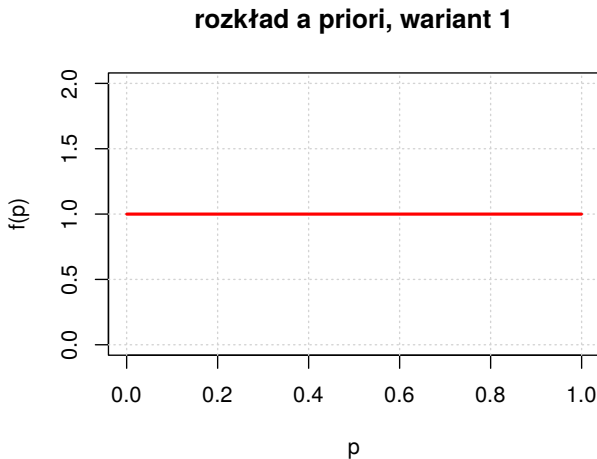
Przykład Bayesowskiego podejścia do problemu

Spotykamy znajomego, który proponuje nam zakład oparty na rzutach monetą.

- ▶ Prawdopodobieństwo p wypadnięcia orła to parametr modelu.
- ▶ Chcemy oszacować p przed zakładem.
- ▶ W zależności od sytuacji możemy zakładać różne rozkłady a priori.

Rozkład a priori — wariant 1

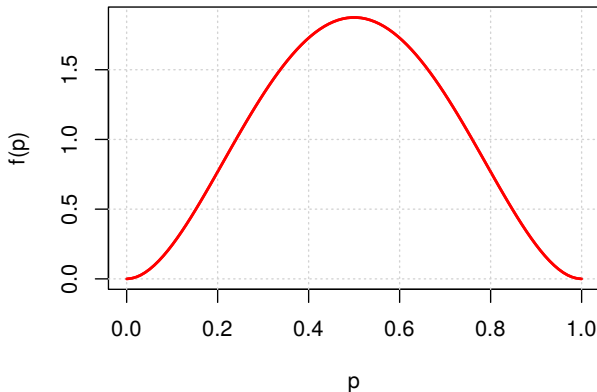
Jeśli chcemy być „obiektywni” możemy założyć rozkład jednostajny.



Rozkład a priori — wariant 2

Wiedząc, że monety są zazwyczaj symetryczne można zakładać rozkład skupiony wokół wartości $p = \frac{1}{2}$.

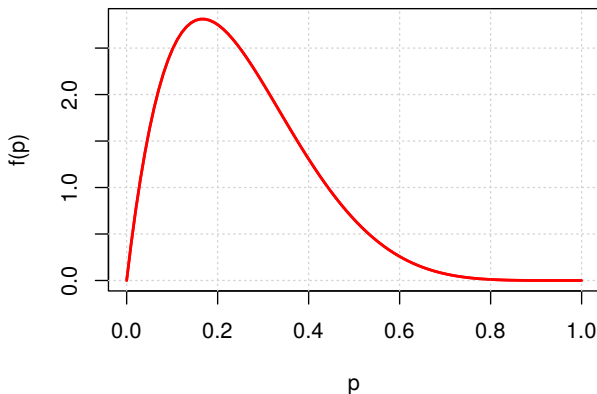
rozkład a priori, wariant 2



Rozkład a priori — wariant 3

Znając monetę (np. będąc świadkiem innych zakładów z wykorzystaniem tej samej monety), można zakładać rozkład:

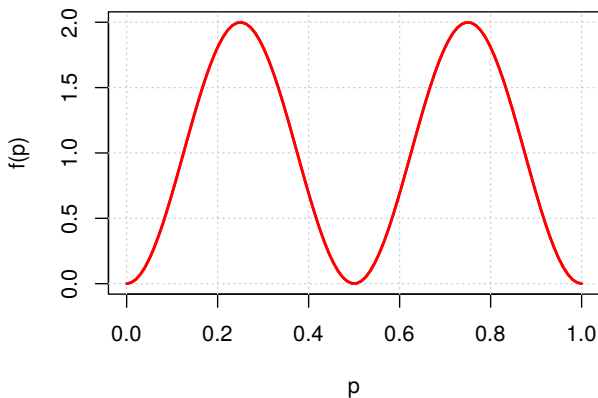
rozkład a priori, wariant 3



Rozkład a priori — wariant 4

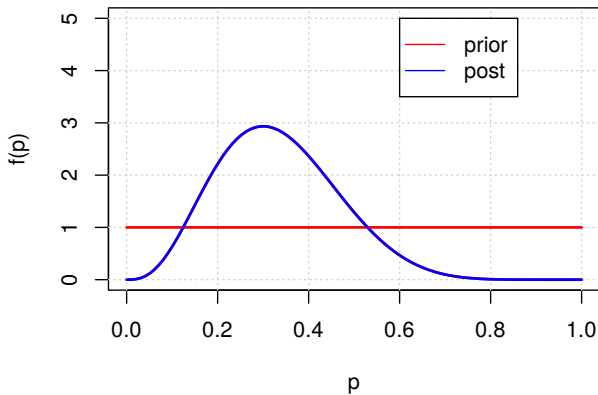
Znając „przewrotność” właściciela monety można też zakładać rozkład bimodalny.

rozkład a priori, wariant 4

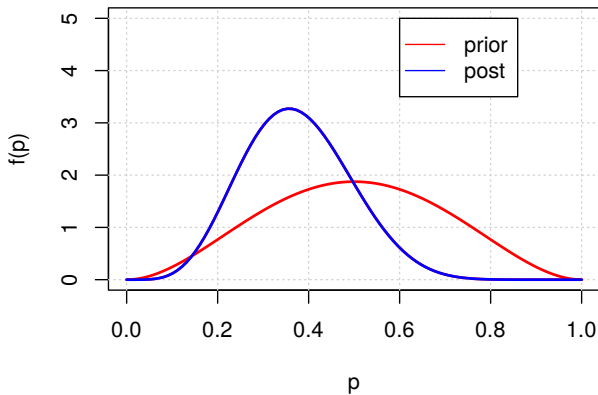


Rzucamy „próbnie” 10 razy monetą i notujemy 3 orły.

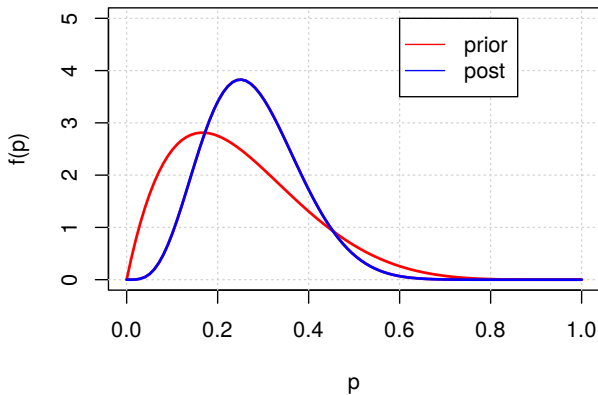
rozkład a posteriori, wariant 1



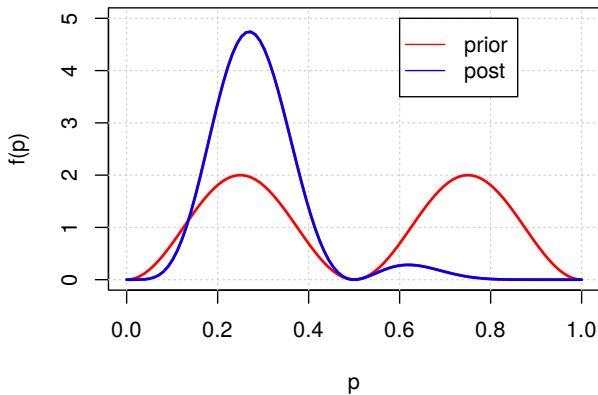
rozkład a posteriori, wariant 2

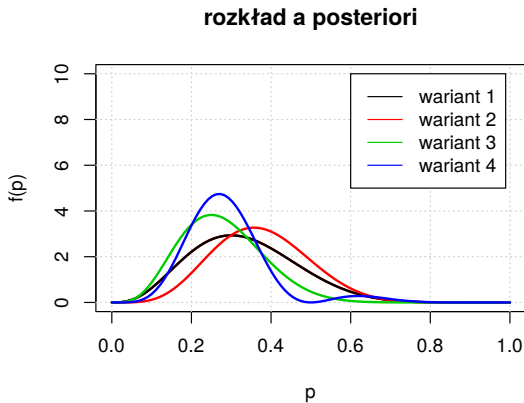


rozkład a posteriori, wariant 3



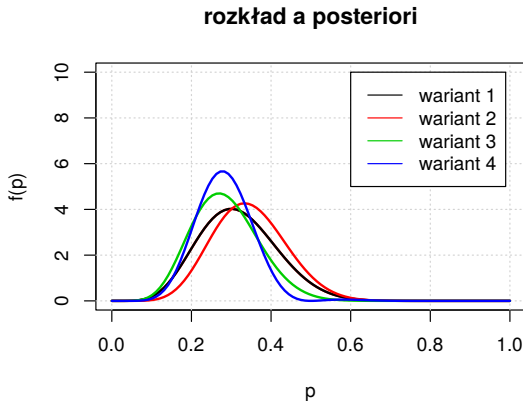
rozkład a posteriori, wariant 4





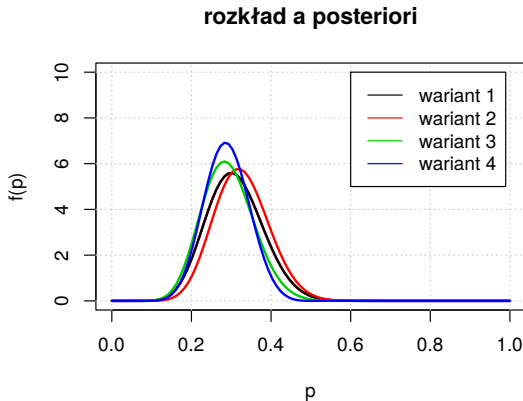
Rozkład a posteriori — podsumowanie (c.d.)

A gdybyśmy rzucili 20 razy uzyskując 6 orłów.



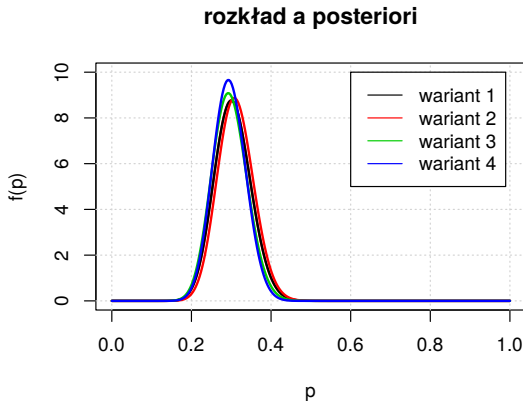
Rozkład a posteriori — podsumowanie (c.d.)

A gdybyśmy rzucili 40 razy uzyskując 12 orłów.



Rozkład a posteriori — podsumowanie (c.d.)

A gdybyśmy rzucili 100 razy uzyskując 30 orłów.



- ▶ Model parametryczny $\mathcal{F} = \{f_p(x) : p \in \Theta\}$.
- ▶ Mając rozkład a priori parametru p (funkcja gęstości $f_{\text{prior}}(p)$), parametryczny model oraz próbę $X = (X_1, \dots, X_n)$ możemy wyznaczyć rozkład a posteriori parametru (funkcję gęstości $f_{\text{post}}(p, x)$):

$$f_{\text{post}}(p, x) = \frac{\mathcal{L}_n(x; p) f_{\text{prior}}(p)}{\int_{\Theta} \mathcal{L}_n(x; p) f_{\text{prior}}(p) dp}.$$

Definicja

Bayesowskim estymatorem \hat{p} parametru p nazywamy wartość oczekiwaną tego parametru wyliczoną za pomocą rozkładu a posteriori:

$$\hat{p} = \int_{\Theta} p f_{\text{post}}(p) dp.$$

Czasami zamiast wartości oczekiwanej wg rozkładu a posteriori, definiuje się estymator bayesowski \hat{p} jako **dominantę** (inaczej wartość modalną lub modę) rozkładu a posteriori, czyli wartość parametru p , dla której funkcja gęstości f_{post} przyjmuje wartość maksymalną.

Fakt

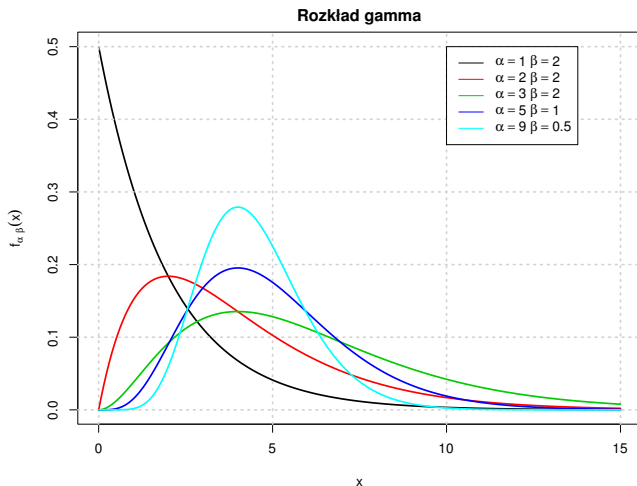
Niezależnie od wyboru (jednej z dwóch) definicji estymatora bayesowskiego, dla dużych prób (asymptotycznie) ma on te same własności co estymator największej wiarygodności.

Niezwykłe użyteczne rozkłady — przypomnienie

- ▶ rozkład normalny $N(\mu, \sigma^2)$ (nośnik \mathbb{R}),
- ▶ rozkład gamma $\text{Gamma}(\alpha, \beta)$ (nośnik $[0, +\infty)$),
- ▶ rozkład beta $\text{Beta}(\alpha, \beta)$ (nośnik $[0, 1]$).

Rozkład gamma — przypomnienie

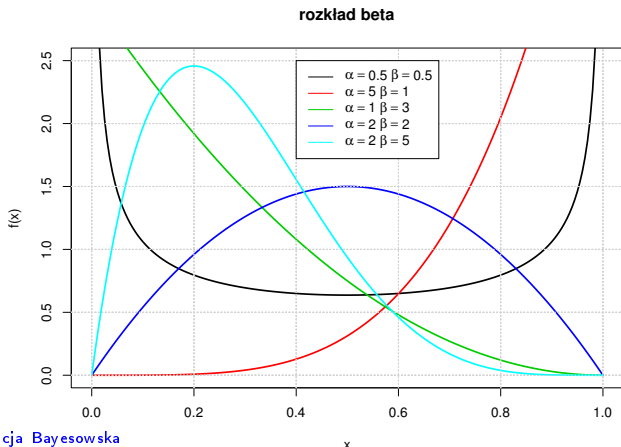
$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad \mathbb{E}X = \alpha\beta, \quad \mathbb{V}X = \alpha\beta^2$$



Rozkład beta — przypomnienie

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad \alpha, \beta > 0$$

$$\mathbb{E}X = \frac{\alpha}{\alpha + \beta}, \quad \mathbb{V}X = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$



Rozkład sprzężony dla modelu parametrycznego $\mathcal{F} = \{f_p(x) : p \in \Theta\}$, to taki rozkład (typ rozkładu) X , że przy rozkładzie a priori typu X dostaje się rozkład a posteriori również typu X (być może z innymi parametrami).
Przykłady:

model	par.	r. sprz.	par. a priori	parametry rozkł. a posteriori
Bern	p	Beta	α, β	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$
Pois	λ	Γ	α, β	$\alpha + \sum_{i=1}^n x_i, \frac{\beta}{n\beta+1}$
N	μ	N	μ_0, σ_0^2	$\frac{\mu_0 \sigma_0^{-2} + \sigma^{-2} \sum_{i=1}^n x_i}{\sigma_0^{-2} + n\sigma^{-2}}, (\sigma_0^{-2} + n\sigma^{-2})^{-1}$
N	σ^2	Γ^{-1}	α, β	$\alpha + n/2, \beta + \sum_{i=1}^n (x_i - \mu)^2/2$
Exp	λ	Γ	α, β	$\alpha + n, \frac{\beta}{1 + \beta \sum_{i=1}^n x_i}$

$$X \sim \Gamma$$

$$\frac{1}{X} \sim \Gamma^{-1}$$



Jak opisać wynik estymacji.

a) $\hat{\theta}$ ← estym. punktowa

b) $[\hat{A}_\theta, \dots, \hat{B}_\theta]$ ← estym. przedziałowa

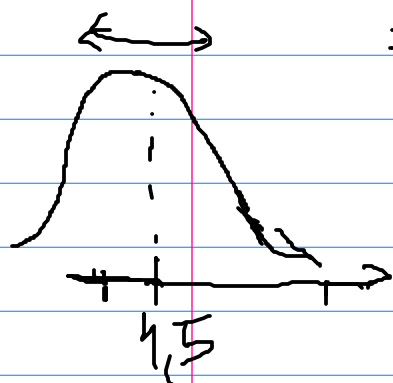
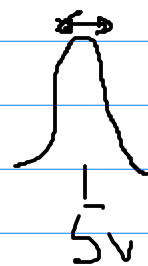
c)  fgp parametru θ

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \{ \underbrace{f(x_1, \dots, x_n; \theta)} \}$$

a) Jak „włożyć” dodatkowe wiedzy (sposób eksp. wzm.)

b) Maksym. jakiej mac tego

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} f(\theta | x_1, \dots, x_n) \\ &= \underset{\theta}{\operatorname{argmax}} \underbrace{f(x_1, \dots, x_n | \theta)}_{\text{funkcja wiarygodności}} \cdot \underbrace{f(\theta)}_{\text{f.p. a priori}} \end{aligned}$$



funkcja wiarygodności

wzrostad postwa θ
sprzed eksp. wzm.

f.p. a posteriori

f.p. a priori

Hipotezy statystyczne i ich weryfikacja

MWS, wykład 7

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 28 kwietnia 2021

Hipotezy statystyczne są hipotezami dotyczącymi rozkładów prawdopodobieństwa.

Wyróżniamy:

- ▶ **hipotezy proste** mówiące, że pewna zmienna ma określony (jednoznacznie) rozkład. Np. moneta jest symetryczna (zmienna opisująca ilość orłów w n rzutach ma rozkład $\text{Binom}(n, \frac{1}{2})$),
- ▶ **hipotezy złożone** nie określają jednoznacznie rozkładu prawdopodobieństwa (np. rozkład normalny z nieznaną wariancją) lub nie określają go w ogóle (np. jakiś rozkład nie będący $\text{Bern}(\frac{1}{2})$).

Mamy do czynienia z dwoma hipotezami:

H_0 hipotezą zerową, czyli hipotezą, którą w pewnym sensie poddajemy weryfikacji,

H_1 hipotezą alternatywną, czyli taką, o której zakładamy, że zachodzi, jeśli hipoteza zerowa okaże się nieprawdziwa.

Definicja

Odrzucenie hipotezy H_0 , gdy jest ona prawdziwa nazywamy **błędem I rodzaju**. Prawdopodobieństwo popełnienia takiego błędu na podstawie pewnego testu nazywane jest **poziomem istotności** tego testu i oznaczane najczęściej symbolem α .

Definicja

Przyjęcie hipotezy H_0 , gdy w rzeczywistości nie jest ona prawdziwa nazywamy **błędem II rodzaju**. Prawdopodobieństwo popełnienia takiego błędu na podstawie pewnego testu oznaczane jest najczęściej symbolem β . Wielkość $1 - \beta$ nazywana jest **mocą testu**.

Czym mniejszy poziom istotności testu (α) i czym większa moc tego testu ($1 - \beta$) tym lepiej.

Przykład

Obiekt obserwowany w systemie radarowym może podlegać testowi z hipotezami:

H_0 — obiekt jest pociskiem odrzutowym,

H_1 — obiekt jest pasażerskim odrzutowcem.

- ▶ Błąd I rodzaju polega na zignorowaniu zagrożenia (potraktowanie pocisku jako odrzutowca),
- ▶ Błąd II rodzaju polega na fałszywym alarmie (mogącym się zakończyć zestrzeleniem samolotu pasażerskiego).

Przykład

Pewien test medyczny ma określić czy pacjent cierpi na schorzenie X:

H_1 — tak, cierpi (pozytywny wynik testu),

H_0 — nie (negatywny wynik testu).

- ▶ Błąd I rodzaju polega na „fałszywym alarmie” (podjęcie niepotrzebnego leczenia mogącego mieć negatywne skutki uboczne),
- ▶ Błąd II rodzaju polega na zignorowaniu zagrożenia (nie podjęcie terapii gdy jest ona potrzebna).

Definicja

Statystyka decyzyjna (ew. testowa) to statystyka (funkcja próby), na podstawie której weryfikujemy hipotezę. **Obszar krytyczny testu**, to obszar wartości tej statystyki, który prowadzi do odrzucenia hipotezy zerowej.

Przykład 2 prostych hipotez

W pudełku znajdują się dwie monety: symetryczna oraz taka, że prawdopodobieństwo wypadnięcia orła wynosi 0.7. Wyciągamy losowo jedną monetę, rzucamy nią 10 razy, notujemy co „wypadło” i na tej podstawie chcemy weryfikować

H_0 : wylosowaliśmy monetę symetryczną wobec

H_1 : wylosowaliśmy tę drugą monetę.

Obie hipotezy są proste. Liczba orłów opisana jest rozkładem $\text{Binom}(10, 0.5)$ lub $\text{Binom}(10, 0.7)$.

	0	1	2	3	4	5	6	7	8	9	10
$p = .5$.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010
$p = .7$.0000	.0001	.0014	.0090	.0368	.1029	.2001	.2668	.2335	.1211	.0282

Kontynuacja przykładu

k	0	1	2	3	4	5	6	7	8	9	10
$P(X = k p = .5)$.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010
$P(X = k p = .7)$.0000	.0001	.0014	.0090	.0368	.1029	.2001	.2668	.2335	.1211	.0288

Za statystykę decyzyjną weźmy **iloraz wiarygodności**

$$R = \frac{P(X|H_1)}{P(X|H_0)}.$$

x	0	1	2	3	4	5	6	7	8	9	10
$R(x)$	0.006	0.014	0.033	0.077	0.18	0.42	0.98	2.3	5.3	12.4	28.9

Określamy obszar krytyczny jako $\{R: R > c\}$ (c to tzw. wartość krytyczna).

- ▶ Biorąc $c = 30$ nie popełnimy błędu I rodzaju.
- ▶ Biorąc $c = 0$ nie popełnimy błędu II rodzaju.
- ▶ Biorąc $c = 1$ (odrzucaamy H_0 jeśli wypadło > 6 orłów) popełnimy błąd I rodzaju z prawd. 0.17, a II rodzaju — z prawd. 0.35.
- ▶ Biorąc $c = 10$ (odrzucaamy H_0 jeśli > 8 orłów) popełnimy błąd I rodzaju z prawd. 0.01, a II rodzaju — z prawd. 0.85.


Lemat (Neymana-Pearsona)

Niech H_0 oraz H_1 będą hipotezami prostymi oraz niech dany będzie test oparty na ilorazie wiarygodności (tzn. test odrzucający hipotezę H_0 gdy iloraz wiarygodności jest większy niż pewna stała c) na pewnym poziomie istotności α . Wówczas jest to test o największej mocy spośród wszystkich testów na poziomach istotności nie przekraczających α .

Niech X_1, \dots, X_n będą niezależne o rozkładzie normalnym ze znaną wariancją σ^2 oraz

$$H_0: \mu = \mu_0$$

$$H_1: \mu = \mu_1.$$

Test oparty na ilorazie wiarygodności, to test oparty na średniej \bar{X}_n z próby. 

Kontynuacja przykładu z monetą

x	0	1	2	3	4	5	6	7	8	9	10
$R(x)$	0.006	0.014	0.033	0.077	0.18	0.42	0.98	2.3	5.3	12.4	28.9

- ▶ W związku z dyskretnością zbioru wartości ilorazu wiarygodności, zadając obszary krytyczne postaci $\{R > c\}$, można otrzymać dyskretną liczbę „osiągalnych” poziomów istotności.
- ▶ Np. test z obszarem krytycznym $\{R > 30\}$ będzie miał zerowy poziom istotności (taki sam poziom dla $\{R > \frac{P(10|H_1)}{P(10|H_0)}\}$);
- ▶ test z obszarem krytycznym $\{R > 28\}$ będzie miał poziom istotności $\alpha \approx 0.001$ (taki sam poziom dla $\{R > \frac{P(9|H_1)}{P(9|H_0)}\}$);
- ▶ test z obszarem krytycznym $\{R > 10\}$ będzie miał poziom istotności $\alpha \approx 0.011$ (taki sam poziom dla $\{R > \frac{P(8|H_1)}{P(8|H_0)}\}$);
- ▶ Czy można skonstruować test na poziomie istotności np. $\alpha = 0.005$?

Test randomizowany to test, który w zależności od wartości statystyki decyzyjnej:

- ▶ odrzuca hipotezę zerową (obszar krytyczny),
- ▶ nie odrzuca hipotezy zerowej (obszar akceptacji),
- ▶ odrzuca hipotezę zerową losowo z zadany
prawdopodobieństwem (na „granicy obszarów”).

Kontynuacja przykładu z monetą

x	0	1	2	3	4	5	6	7	8	9	10
p = .5	.0010	.0098	.0439	.1172	.2051	.2461	.2051	.1172	.0439	.0098	.0010
p = .7	.0000	.0001	.0014	.0090	.0368	.1029	.2001	.2668	.2335	.1211	.0282

x	0	1	2	3	4	5	6	7	8	9	10
R(x)	0.006	0.014	0.033	0.077	0.18	0.42	0.98	2.3	5.3	12.4	28.9

Skonstruujemy test (randomizowany) z poziomem istotności $\alpha \approx 0.005$.

$$\begin{cases} \text{odrzucaamy } H_0 & \text{jeśli } R > \frac{\mathbb{P}(9|H_1)}{\mathbb{P}(9|H_0)}, \\ \text{nie odrzucaamy } H_0 & \text{jeśli } R < \frac{\mathbb{P}(9|H_1)}{\mathbb{P}(9|H_0)}, \\ \text{odrzucaamy } H_0 \text{ z pr. } p_* & \text{jeśli } R = \frac{\mathbb{P}(9|H_1)}{\mathbb{P}(9|H_0)}. \end{cases}$$

$$0.005 = \alpha = \mathbb{P}(10|H_0) + \mathbb{P}(9|H_0)p_* \approx 0.001 + p_* 0.0098 \Rightarrow p_* \approx 0.4.$$


Lemat Neymana-Pearsona wymaga aby obie hipotezy H_0 i H_1 były proste.

Jeśli H_1 jest hipotezą złożoną (można ją traktować jako składającą się z wielu hipotez prostych) oraz pewien test jest najmocniejszy (spełnia założenia lematu N-P) dla dowolnej prostej hipotezy alternatywnej z H_1 (tzn. będącej składową oryginalnej hipotezy H_1), to taki test nazywamy **testem jednostajnie najmocniejszym**.

Niech X_1, \dots, X_n będą niezależne o rozkładzie normalnym ze znaną wariancją σ^2 oraz

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0 \quad \text{hipoteza jednostronna.}$$

Test oparty na ilorazie wiarygodności (obszar krytyczny postaci $\{\bar{X}_n > x_0\}$) jest testem jednostajnie najmocniejszym. 

Ale ten sam test w przypadku hipotezy (**dwustronnej**) $H_1: \mu \neq \mu_0$ testem jednostajnie najmocniejszym już nie jest.

Niech T będzie statystyką decyzyjną oraz niech obszar krytyczny testu na poziomie istotności α będzie postaci

$$\{T > t_0\},$$

gdzie t_0 dobrane tak by $\mathbb{P}(T > t_0 | H_0) = \alpha$.

Definicja

p-wartością dla zaobserwowanej próby nazywamy minimalną wartość poziomu istotności α , dla której hipoteza zerowa byłaby odrzucona.

p-wartość można interpretować jako prawdopodobieństwo, pod warunkiem H_0 , uzyskania wartość statystyki testowej tak samo lub bardziej „ekstremalnej” niż wartość wyznaczona dla zaobserwowanej próby.

Przykład: weryfikacja zdolności nadprzyrodzonych

Osoba twierdząca, że ma nadprzyrodzone zdolności proszona jest o rozpoznanie jednego z 4 kolorów 20 losowo (bez zwracania) wyciągniętych kart (z 52-kartowej talii). T — liczba poprawnie odgadniętych kart.

H_0 : osoba zgaduje,

H_1 : osoba ma szósty zmysł.

- ▶ H_0 jest prosta (T ma wówczas rozkład $\text{Binom}(20, \frac{1}{4})$), a
- ▶ H_1 złożona.
- ▶ Załóżmy, że osoba trafnie odgadła kolory 9 kart. Hipoteza zerowa zostałaby odrzucona np. przy poziomie istotności $\alpha = 0.05$, a nie zostałaby odrzucona dla $\alpha = 0.01$.
- ▶ p-wartością dla wyniku tego eksperymentu jest 0.041 ($\mathbb{P}(T \geq 9 | H_0) \approx 0.041$).
- ▶ (dla 10 odgadniętych kart p-wartość wynosiłaby 0.014.)

Uogólniony iloraz wiarygodności

Niech $H_0: \theta \in \Omega_0$, $H_1: \theta \in \Omega_1$ oraz $\Omega = \Omega_0 \cup \Omega_1$.

Uogólniony iloraz wiarygodności dany jest formułą

$$\Lambda^* = \frac{\max_{\theta \in \Omega_1} \mathcal{L}(\theta)}{\max_{\theta \in \Omega_0} \mathcal{L}(\theta)}.$$

Duże wartości Λ^* „dyskredytują” hipotezę H_0 .

Często wygodniej jest posługiwać się ilorazem

$$\Lambda = \frac{\max_{\theta \in \Omega} \mathcal{L}(\theta)}{\max_{\theta \in \Omega_0} \mathcal{L}(\theta)}.$$

$(\Lambda = \max(\Lambda^*, 1))$. Obszar krytyczny testu opartego o u.i.w. jest postaci $\{\Lambda > \lambda_0\}$, gdzie λ_0 dobierana tak, by zapewnić zadany poziom istotności.

Przykład

Rodzina rozkładów $N(\mu, \sigma^2)$ ze znaną σ^2 oraz hipotezy

$$H_0: \mu = \mu_0 \quad \text{wobec} \quad H_1: \mu \neq \mu_0.$$

$$\Omega_0 = \{\mu_0\}, \quad \Omega_1 = \{\mu: \mu \neq \mu_0\}, \quad \Omega = \mathbb{R}.$$

$$\Lambda \stackrel{\text{pencil}}{=} \exp \left(\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \bar{X})^2 \right) \right).$$

Duże wartości Λ odpowiadają dużym wartościom

$$2 \ln \Lambda \stackrel{\text{pencil}}{=} n(\bar{X} - \mu_0)^2 / \sigma^2.$$

Jeśli hipoteza H_0 jest prawdziwa, to rozkład $2 \ln \Lambda$ jest χ_1^2 .

Dla zadanego poziomu istotności α możemy zdefiniować obszar krytyczny:

$$(\bar{X} - \mu_0)^2 \stackrel{\text{pencil}}{>} \frac{\sigma^2}{n} F_{\chi_1^2}^{-1}(1 - \alpha).$$

Równoważnie

$$|\bar{X} - \mu_0| \stackrel{\text{pencil}}{>} \frac{\sigma}{\sqrt{n}} \Phi^{-1}(1 - \frac{\alpha}{2}).$$

Badanie zgodności rozkładów

MWS, wykład 8

Rafał Rytel-Andrianik
na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 18 maja 2021

Dysponujemy próbą X_1, \dots, X_n i chcemy sprawdzić czy pochodzi ona z danego rozkładu,

- ▶ inaczej: zbadać czy rozkład próby jest **zgodny** z danym rozkładem;
- ▶ ten hipotetyczny rozkład może być określony np. funkcją gęstości prawdopodobieństwa lub dystrybuantą.

Często o rozkładzie, z którym chcemy sprawdzić zgodność danych, wiemy tylko, że należy do pewnej rodziny (np. rozkładów normalnych), a nie znamy wartości parametrów. Wówczas:

- ▶ najpierw estymujemy parametry rozkładu (np. średnią i wariancję),
- ▶ następnie badamy zgodność danych (próby) z rozkładem o wyestymowanych parametrach.

Przypadek 1

Hipotetyczny rozkład jest dyskretny

Przykład wiodący

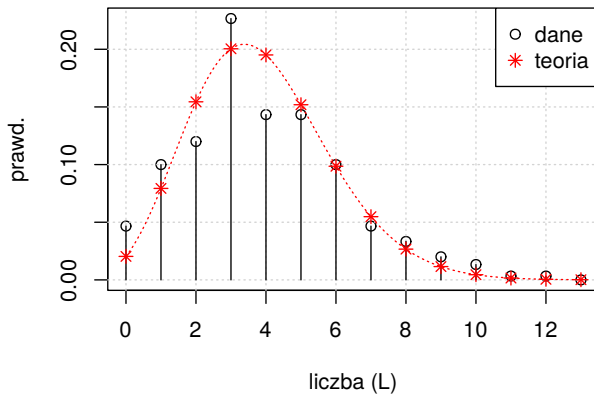
W poniższej tabeli przedstawione są liczby pojazdów skręcających na pewnym skrzyżowaniu w prawo w przeciągu 300 3-minutowych przedziałów czasu. Będziemy badali zgodność tych danych z rozkładem Poissona.

S	0	1	2	3	4	5	6
L	14	30	36	68	43	30	14
S	7	8	9	10	11	12	13+
L	14	10	6	4	1	1	0

- ▶ S to Liczba „skrętów” w przeciągu 3 minut.
- ▶ L to liczba 3-minutowych przedziałów z zadaną liczbą „skrętów”.

$$\hat{\lambda} = \frac{0 \times 14 + 1 \times 30 + \dots + 12 \times 1}{14 + 30 + \dots + 1} \approx 3.9$$

Wynik estymacji parametru λ



W kierunku testu zgodności χ^2

Rozważmy problem losowego rozmieszczenia n kul w r koszykach, przy założeniu, że prawdopodobieństwo umieszczenia kuli w i -tym koszyku jest równe p_i .

Niech X_k oznacza numer koszyka, do którego wpadła k -ta kula.

$$\mathbb{P}(X_k = i) = p_i, \quad p_1 + p_2 + \cdots + p_r = 1.$$

Niech v_i oznacza liczbę kul, które wpadły do i -tego koszyka.

$$\mathbb{E}v_i = np_i.$$

Chcemy zbadać jakiego odstępstwa zmiennych v_i od ich wartości średnich np_i możemy się spodziewać.

Twierdzenie Pearsona

n kul, r koszyków, $p_i = \mathbb{P}(X = i)$, v_i — liczba kul w i -tym koszyku.

Chcemy zbadać jakiego odstępstwa zmiennych v_i od ich wartości średnich np_i możemy się spodziewać.

Twierdzenie (Pearsona)

Rozkład zmiennej losowej

$$T = \sum_{i=1}^r \frac{(v_i - np_i)^2}{np_i}$$

zbiega do rozkładu χ^2 z $(r - 1)$ stopniami swobody (χ_{r-1}^2).

v_i (ustalone i) można potraktować jako sumę n niezależnych zmiennych o rozkładzie Bernoulliego z prawdopodobieństwem sukcesu p_i .

Zatem

$$\mathbb{V} v_i = n(p_i - p_i^2) = np_i(1 - p_i).$$

W szczególności na mocy centralnego twierdzenia granicznego

$$\frac{v_i - np_i}{\sqrt{np_i(1 - p_i)}} \rightarrow N(0, 1).$$

Innymi słowy

$$\frac{v_i - np_i}{\sqrt{np_i}} \rightarrow N(0, 1 - p_i).$$

Test zgodności rozkładu

Rozważamy próbę X_1, \dots, X_n niezależnych zmiennych losowych o tym samym, **dyskretnym** rozkładzie.

Oznaczmy przez B_1, \dots, B_r zbiór wartości jakie mogą przyjmować zmienne X oraz przez p_1, \dots, p_r prawdopodobieństwa przyjmowania poszczególnych wartości.

Chcemy zbadać, czy próba X_1, \dots, X_n odpowiada wartościom pewnych ustalonych prawdopodobieństw p_1^*, \dots, p_r^* , czyli czy zachodzi hipoteza:

$$H_0 : p_1 = p_1^*, \dots, p_r = p_r^*$$

wobec hipotezy alternatywnej

$$H_1 : \text{przynajmniej dla jednego } i \text{ jest } p_i \neq p_i^*.$$

$$H_0 : p_1 = p_1^*, \dots, p_r = p_r^*$$

H_1 : przynajmniej dla jednego i jest $p_i \neq p_i^*$.

Za statystykę decyzyjną przyjmujemy

$$T = \sum_{i=1}^r \frac{(v_i - np_i^*)^2}{np_i^*}.$$

Jeśli rzeczywiście $p_i = p_i^*$, $i = 1, \dots, r$, to na mocy tw. Pearsona

$$T \xrightarrow{d} \chi_{r-1}^2.$$

Gdyby natomiast pewne $p_i \neq p_i^*$, to

$$\frac{v_i - np_i^*}{\sqrt{np_i^*}} \stackrel{\text{pencil}}{=} \sqrt{\frac{p_i}{p_i^*}} \frac{v_i - np_i}{\sqrt{np_i}} + \sqrt{n} \frac{p_i - p_i^*}{\sqrt{p_i^*}}.$$

Zatem wystarczy aby jedno $p_i \neq p_i^*$ aby $T \xrightarrow{n \rightarrow \infty} \infty$.

Test zgodności rozkładu χ^2

$$H_0 : p_1 = p_1^*, \dots, p_r = p_r^*$$

H_1 : przynajmniej dla jednego i jest $p_i \neq p_i^*$.

Za statystykę decyzyjną przyjmujemy

$$T = \sum_{i=1}^r \frac{(v_i - np_i^*)^2}{np_i^*}.$$

Reguła decyzyjna w teście zgodności χ^2 :

$$\begin{cases} H_0 : T \leq c, \\ H_1 : T > c \end{cases}.$$

Stałą c dobieramy tak, by zapewnić określony poziom istotności α testu. Dla dużych prób można szacować:

$$c \approx F_{\chi_{r-1}^2}^{-1}(1 - \alpha).$$

Kontynuacja przykładu z pojazdami

Mamy $r = 14$ „koszyków” (0 skrętów, 1 skręt, ..., co najmniej 13 skrętów). Chcemy sprawdzić, czy prawdopodobieństwa „wpadnięcia” $n = 300$ „kul” do poszczególnych „koszyków” są równe:

$$p_0^* = \frac{3.9^0}{0!e^{3.9}}, p_1^* = \frac{3.9^1}{1!e^{3.9}}, \dots, p_{12}^* = \frac{3.9^{12}}{12!e^{3.9}}, p_{13+}^* = 1 - p_0 - \dots - p_{12}.$$

Wyznaczamy wartość krytyczną testu dla $\alpha = 0.05$:

$$T \sim \chi_{13}^2, \quad c = F_{\chi_{13}^2}^{-1}(1 - 0.05) \approx 22.4.$$

Wartość statystyki T dla naszych danych: $T \approx 32.6 > c$. Przy wybranym poziomie istotności hipotezę o tym, że rozkład jest $\text{Pois}(3.9)$ należy odrzucić!

p-wartość dla danego testu jest równa ≈ 0.002 !

Kontynuacja przykładu z pojazdami

Jaką hipotezę sprawdzaliśmy?

Taką, że zaobserwowane liczby skrętów odpowiadają rozkładowi $\text{Pois}(3.9)$.

Jak sprawdzić, czy te liczby odpowiadają rozkładowi $\text{Pois}(\lambda)$ dla jakiegokolwiek λ ?

Test zgodności χ^2 dla hipotezy złożonej

Fakt

Jeśli w wyjściowej sytuacji z kulami rozważymy hipotezę H_0 : każde p_i jest równe $p_i(\theta)$, dla wspólnego $\theta \in \Theta$ wobec hipotezy alternatywnej H_1 przeciwnej do H_0 , i jeżeli $\hat{\theta}$ jest estymatorem największej wiarygodności

$$\text{tzn. } \hat{\theta} = \arg \max_{\theta \in \Theta} p_1(\theta)^{v_1} \dots p_r(\theta)^{v_r},$$

to

$$T = \sum_{i=1}^r \frac{(v_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow[n \rightarrow \infty]{d} \chi_{r-s-1}^2,$$

gdzie s to wymiar przestrzeni parametrów Θ .

Kontynuacja przykładu z pojazdami

Test na poziomie istotności dla hipotezy zerowej: dane „układają się” według pewnego rozkładu Poissona wobec hipotezy alternatywnej: dane nie układają się wg żadnego rozkładu Poissona wygląda podobnie do poprzednio skonstruowanego, gdyż stosowaliśmy go do estymatora największej wiarygodności.

Różnice: $\lambda \in (0, \infty) = \Theta$, zatem $s = \dim \Theta = 1$, rozkład graniczny statystyki T jest zatem rozkładem χ^2 o 12 (a nie 13) stopniach swobody.

p-wartość dla nowego testu wynosi ≈ 0.001 !

Średnio raz na tysiąc (!) razy 300-elementowa próba z rozkładu Poissona będzie tak „słabo” lub jeszcze „gorzej” zgodna z rozkładem Poissona niż rozważane w przykładzie dane.

Przypadek 2

Hipotetyczny rozkład jest ciągły

Jak sprawdzać zgodność rozkładu dla ciągłych rozkładów

Rozważmy problem polegający na sprawdzeniu, czy dana próba losowa X_1, \dots, X_n „pochodzi” z rozkładu ciągłego (np. normalnego $N(\mu, \sigma^2)$ o zadanych parametrach) zadanego pewną dystrybucją F .

Rozwiązanie 1 - dyskretyzacja rozkładu

Podzielić zbiór wartości, które mogą przyjmować zmienne X , na skończoną liczbę przedziałów. Na podstawie danego (ciągłego) rozkładu wyznaczyć prawdopodobieństwa „wpadnięcia” do każdego z przedziałów. Policzyc ile ze zmiennych X wpada do każdego z przedziałów i przeprowadzić test zgodności χ^2 .

Rozwiązanie 2 - test Kołomorgowa-Smirnowa

Definicja

Dystrybuanta empiryczna dla próby X_1, \dots, X_n to dystrybuanta F_n określona wzorem:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i \leq t),$$

gdzie (funkcja charakterystyczna, indyktor zbioru)

$$\mathbb{1}(X_i \leq t) = \begin{cases} 1 & \text{jeśli } X_i \leq t, \\ 0 & \text{jeśli } X_i > t \end{cases}$$

X_1, \dots, X_n próba los. z rozkładu o dystrybuancie F .

Fakt

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[\mathbb{P}]{n \rightarrow \infty} 0.$$

Twierdzenie

Rozkład zmiennej losowej

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

nie zależy od dystrybuanty F (tzn. dla każdej dystrybuanty ciągłego rozkładu i niezależnej próby X_1, \dots, X_n pochodzącej z tego rozkładu, powyższa zmienna ma taki sam rozkład).

X_1, \dots, X_n próba los. z rozkładu o dystrybuancie F .

Twierdzenie

Niech $D_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$.

Wtedy

$$P(D_n \leq t) \xrightarrow{n \rightarrow \infty} H(t) = 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 t^2}.$$

($H(t)$ to dystrybuanta rozkładu Kołmogorowa-Smirnowa).

Test Kołmogorowa-Smirnowa

X_1, \dots, X_n próba losowa.

F pewna ustalona dystrybuanta rozkładu ciągłego.

H_0 : zmienne X_1, \dots, X_n pochodzą z rozkładu o dystryb. F ,

H_1 : zmienne X_1, \dots, X_n nie pochodzą z rozkładu o dystryb. F .

Statystyka decyzyjna: $D_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$.

Reguła decyzyjna:
$$\begin{cases} H_0: D_n \leq c, \\ H_1: D_n > c \end{cases}.$$

Stałą c dobiera się tak, by zapewnić określony poziom istotności α testu (na podstawie rozkładu D_n lub rozkładu granicznego Kołmogorowa-Smirnowa).

R: funkcja `ks.test`

Przykład

Rozkład t-Studenta dla n stopni swobody zbliża się do rozkładu $N(0,1)$ gdy liczba stopni swobody rośnie.

Czy `ks.test` da się „nabrać”, że dane z rozkładu Studenta są gaussowskie?

—

Najpierw dla liczby stopni swobody równej 4.

R

```
> x = rt(1000, 4)
> ks.test(x, 'pnorm', 0, sd(x))
      One-sample Kolmogorov-Smirnov test
data:  x
D = 0.062471, p-value = 0.0008152
alternative hypothesis: two-sided
```

Interpretacja p-wartości: Wybrać H_1 (x nie są gaussowskie) można nawet dla poziomu istotności (p-stwo błędnego wyboru H_1) 0.0008. Czyli można śmiało wybrać H_1 .

Teraz dla liczby stopni swobody równej 20.

R

```
> x = rt(1000, 20);  
> ks.test(x, 'pnorm', 0, sd(x))  
      One-sample Kolmogorov-Smirnov test  
data:  x  
D = 0.025455, p-value = 0.5361  
alternative hypothesis: two-sided
```

Tym razem:

- ▶ $D = 0.025$ - mniejsze, więc dystrybuanta empiryczna jest bliżej dystrybuanty rozkładu normalnego
- ▶ $p\text{-wartość} = 0.54$ - wskazuje na H_0 , tzn. dystrybuanta empiryczna x nie odbiega **istotnie** od dystrybuanty rozkładu normalnego.

Istnieje wiele testów służących badaniu normalności rozkładu (H_0):

- ▶ test D'Agostino na skośność (rozkład normalny ma zerową skośność; duża wartość skośności z próby świadczy na niekorzyść H_0),
- ▶ test Anscombe-Glynn-a na kurtozę (rozkład normalny ma kurtozę równą 3; duża odległość kurtozy z próby od tej wartości świadczy na niekorzyść H_0),
- ▶ test Jarque-Bera (kombinacja testów skośności i kurtozy),
- ▶ test Shapiro-Wilka (oparty na statystykach pozycyjnych).

W pakiecie R dostępne są funkcje: `agostino.test()`, `kurtosis.test()`, `jarque.test()` (biblioteka `moments`) oraz `shapiro.test()`.

Porównywanie prób

MWS, wykład 9

Rafał Rytel-Andrianik

na podstawie slajdów Marka Rupniewskiego

Instytut Systemów Elektronicznych
Politechnika Warszawska

wersja: 26 maja 2021

- ▶ Mamy do dyspozycji *dwie* próby X_1, \dots, X_n oraz Y_1, \dots, Y_m .
- ▶ Każda próba jest potencjalnie z innego rozkładu.
- ▶ Chcemy sprawdzić, czy wartości średnie tych dwóch rozkładów są równe.

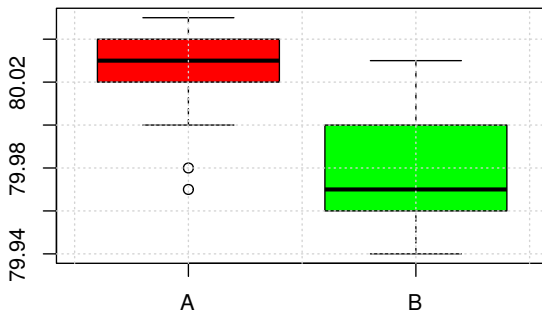
Przykład

Dwie metody A i B były użyte do wyznaczenia całkowitego ciepła potrzebnego do ogrzania i stopienia lodu od temperatury -72°C do wody o temperaturze 0°C .

Wyniki w $[\text{cal/g}]$

A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03	80.02	80.00	80.02
B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97					

wykres pudełkowy



Metody oparte na rozkładzie normalnym

Metody oparte na rozkładzie normalnym

Jeśli próba X_1, \dots, X_n jest z rozkładu $N(\mu_X, \sigma^2)$, a niezależna od niej próba Y_1, \dots, Y_m jest z rozkładu $N(\mu_Y, \sigma^2)$ (ta sama wariancja), to

$$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma^2(n^{-1} + m^{-1})).$$

Zazwyczaj wariancja nie jest dana i trzeba ją estymować z próby:

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}, \quad s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \quad s_Y^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m-1}.$$

Zmienna losowa

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s\sqrt{n^{-1} + m^{-1}}}$$

ma rozkład t-Studenta z $m + n - 2$ stopniami swobody.

Dwie niezależne próby X_1, \dots, X_n oraz Y_1, \dots, Y_m z rozkładów, odpowiednio, $N(\mu_X, \sigma^2)$ oraz $N(\mu_Y, \sigma^2)$ (μ_X, μ_Y, σ^2 nieznane). Wyznaczamy wartość statystyki decyzyjnej,

$$T = \frac{\bar{X} - \bar{Y}}{s\sqrt{n^{-1} + m^{-1}}}.$$

Dla testu dwustronnego ($H_0 : \mu_X = \mu_Y$ wobec $H_1 : \mu_X \neq \mu_Y$) i poziomu istotności α obszar krytyczny testu jest postaci

$$|T| > c, \quad c = F_{t_{m+n-2}}^{-1} \left(1 - \frac{\alpha}{2} \right).$$

Dla testu jednostronnego ($H_1 : \mu_X > \mu_Y$) obszar krytyczny:

$$T > c, \quad c = F_{t_{m+n-2}}^{-1} (1 - \alpha).$$

Przykład — kontynuacja

A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03	80.02	80.00	80.02
B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97					

$$\bar{X} = 80.02, \quad \bar{Y} = 79.98, \quad s_X = 0.024, \quad s_Y = 0.031$$

$$s = \sqrt{\frac{12s_X^2 + 7s_Y^2}{19}}, \quad s = 0.027.$$

$$T = 3.33$$

Dla poziomu istotności $\alpha = 0.01$:

- ▶ $c = 2.861$
- ▶ $|T| > c$ więc są podstawy aby przyjąć H_1 (tzn. średnie nie są równe).

(p-wartość dla dwustronnego testu mniejsza niż 0.01.)

Ale kto powiedział, że wariancje rozkładów X -ów i Y -ów są takie same?

Próby niezależne o nieznanym i być może różnych wariancjach

$$X_i \sim N(\mu_X, \sigma_X^2), i = 1, \dots, n, \quad Y_j \sim N(\mu_Y, \sigma_Y^2), j = 1, \dots, m.$$

$$T = \frac{\bar{X} - \bar{Y}}{s}, \quad \text{gdzie} \quad s^2 = \text{Var}(\bar{X} - \bar{Y}) = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

Statystyka decyzyjna T ma rozkład zbliżony do rozkładu t-studenta z liczbą stopni swobody

$$d \approx \frac{(s^2)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}$$

(po zaokrągleniu do liczby naturalnej).

Dla testu dwustronnego ($H_0 : \mu_X = \mu_Y$ wobec $H_1 : \mu_X \neq \mu_Y$) i poziomu istotności α obszar krytyczny testu jest postaci

$$|T| > c, \quad c = F_{t_d}^{-1} \left(1 - \frac{\alpha}{2} \right).$$

W pakiecie R do przeprowadzania testu równości średnich dla rozkładów normalnych służy funkcja `t.test()`.

Przykład — kontynuacja

A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03	80.02	80.00	80.02
B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97					

$$\bar{X} = 80.02, \quad \bar{Y} = 79.98, \quad s_X = 0.024, \quad s_Y = 0.031$$

$$s^2 = \frac{s_X^2}{13} + \frac{s_Y^2}{8} = 0.00017, \quad T = 3.25.$$

Liczba stopni swobody: $d = \text{round}(12.03) = 12$.

Dla poziomu istotności $\alpha = 0.01$:

- ▶ $c = 3.05$.
- ▶ Ponownie $|T| > c$, więc również bez zakładania równości wariancji są podstawy aby odrzucić H_0 (równość średnich).

Przykład obliczeń w R

Parametry funkcji `t.test`:

```
t.test(x, y = NULL,  
       alternative = c("two.sided", "less", "greater"),  
       mu = 0, paired = FALSE, var.equal = FALSE,  
       conf.level = 0.95, ...)
```

Przykład użycia:

```
> a=c(79.98,80.04,80.02,80.04,80.03,80.03,80.04,  
+ 79.97,80.05,80.03,80.02,80.00,80.02);  
> b=c(80.02,79.94,79.98,79.97,79.97,80.03,79.95,79.97);  
> t.test(a,b)
```

Welch Two Sample t-test

```
data:  a and b  
t = 3.2499, df = 12.027, p-value = 0.006939  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.01385526 0.07018320  
sample estimates:  
mean of x mean of y  
80.02077 79.97875
```

Uwagi o mocy testu porównującego średnie

Moc testu ($H_0 : \mu_X = \mu_Y$ wobec $H_1 : \mu_X \neq \mu_Y$) zależy od

- ▶ Różnicy między średnimi $\Delta = |\mu_X - \mu_Y|$ (większa różnica — większa moc),
- ▶ Poziomu istotności testu α (większy poziom istotności — większa moc),
- ▶ Wariancji prób σ^2 (mniejsza wariancja — większa moc),
- ▶ Rozmiarów prób n, m (większe rozmiary — większa moc).

Założmy, że σ , α oraz Δ są dane oraz, że $n = m$.

$$V(\bar{X} - \bar{Y}) \stackrel{\text{pencil}}{=} \frac{2\sigma^2}{n}.$$

Statystyka decyzyjna:

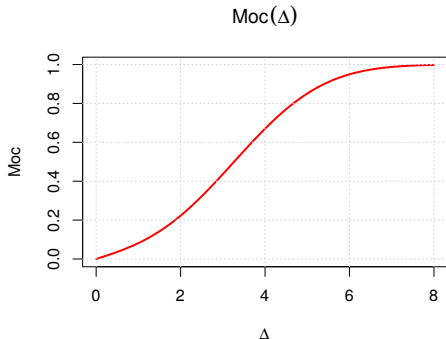
$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{2/n}}.$$

Obszar krytyczny testu: $|Z| > c = \Phi^{-1}(1 - \frac{\alpha}{2})$.

Moc testu w przypadku $\mu_X - \mu_Y = \Delta$:

$$\mathbb{P}\left(\frac{|\bar{X} - \bar{Y}|}{\sigma\sqrt{2/n}} > c_\alpha\right) \stackrel{\text{pencil}}{=} 1 - \Phi\left(c_\alpha - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right) + \Phi\left(-c_\alpha - \frac{\Delta}{\sigma}\sqrt{\frac{n}{2}}\right).$$

Przykład: $m = n = 18$, $\sigma = 5$, $\alpha = 0.05$



Podobnie, jeśli chcemy dla $\Delta = 1$ wykryć różnicę między wartościami średnimi z prawdopodobieństwem 0.9, to

$$\Phi\left(1.96 - \frac{\Delta}{\sigma} \sqrt{\frac{n}{2}}\right) \approx 0.1 \Rightarrow n \approx 525.$$

Będziemy rozważać próbę złożoną z par

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

gdzie niezależne są: X_i od X_j ; X_i od Y_j ; oraz Y_i od Y_j dla $i \neq j$.

Nie zakłada się natomiast niezależności X_i od Y_i .

Jaki zysk daje parowanie?

Niech

$$\mathbb{E}X_i = \mu_X, \mathbb{V}X_i = \sigma_X^2, \mathbb{E}Y_i = \mu_Y, \mathbb{V}Y_i = \sigma_Y^2, \mathbb{C}(X, Y) = \sigma_{XY}$$

oraz niech $D_i = X_i - Y_i$.

Wtedy $\bar{D} = \bar{X} - \bar{Y}$ więc

$$\mathbb{E}\bar{D} = \mu_X - \mu_Y, \quad \mathbb{V}\bar{D} = (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})/n.$$

Gdyby korelacji między X_i a Y_i nie było, to

$$\mathbb{V}\bar{D} = (\sigma_X^2 + \sigma_Y^2)/n.$$

Jeśli rozkład różnic jest rozkładem normalnym $N(\mu_D, \sigma_D^2)$ (μ_D oraz σ_D^2 nieznane), to zmienna losowa

$$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{n}}$$

ma rozkład t-Studenta z $n - 1$ stopniami swobody.

Test dla hipotezy alternatywnej dwustronnej $\mu_D \neq 0$ ma obszar krytyczny

$$|\bar{D}\sqrt{n}/s_D| > c,$$

gdzie

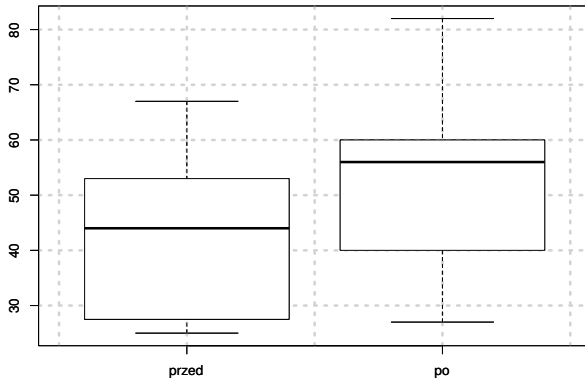
$$c = F_{t_{n-1}}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

W tabeli procentowy udział płytek krwi, które uległy złączeniu w odpowiedzi na odp. stymulację przed i po wypaleniu papierosa przez 11 osób.

przed	25	25	27	44	30	67	53	53	52	60	28
po	27	29	37	56	46	82	57	80	61	59	43

Czy udział płytek, które uległy agregacji się zwiększył?

Przykład — wykres pudełkowy



Przykład — test bez parowania

```
> t.test(po,przed)
```

Welch Two Sample t-test

data: po and przed

t = 1.4164, df = 19.516, p-value = 0.1724

alternative hypothesis: true difference in means is
not equal to 0

95 percent confidence interval:

-4.880458 25.425913

sample estimates:

mean of x mean of y

52.45455 42.18182

Przykład — parowanie

przed	25	25	27	44	30	67	53	53	52	60	28
po	27	29	37	56	46	82	57	80	61	59	43
różnica	2	4	10	12	16	15	4	27	9	-1	15

$$\bar{D} \approx 10.27, s_D \approx 4.27.$$

Dla $\alpha = 0.01$, wartość krytyczna testu $c = 3.17$.

Przykład — obliczenia w pakiecie R

Test z parowaniem

```
> t.test(po,przed,paired=TRUE)
```

Paired t-test

data: po and przed

t = 4.2716, df = 10, p-value = 0.001633

alternative hypothesis: true difference in means is
not equal to 0

95 percent confidence interval:

4.91431 15.63114

sample estimates:

mean of the differences

10.27273

Testy nieparametryczne

czyli bez zakładania konkretnego rozkładu elementów próby

Test Manna-Whitneya (test sumy rang Wilcoxona)

X_1, \dots, X_n próba z pewnego rozkładu F_X ,

Y_1, \dots, Y_m tzw. **próba kontrolna**, niezależna od powyższej,
z pewnego rozkładu F_Y .

Chcemy badać, czy (w odpowiednim sensie) wartości X i Y są na podobnym poziomie (z tego samego rozkładu).

Np. X_i poziom białych krwinek po kuracji badanym lekiem, a Y_j poziom tych krwinek w grupie kontrolnej (nie poddawanej kuracji).

Hipoteza zerowa:

$$H_0: F_X = F_Y.$$

Hipoteza alternatywna jednostronna:

$$H_1: \mathbb{P}(X > Y) > \mathbb{P}(X < Y)$$

lub dwustronna

$$H_1: \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

Idea testu Manna-Whitneya(-Wilcoxona)

1. szeregujemy elementy X_i, Y_j w kolejności rosnącej,
2. sumujemy rangi elementów Y_j (ich numery w uszeregowanym w poprzednim punkcie ciągu),
3. „zbyt mała” lub „zbyt duża” wartość powyższej sumy skłania do odrzucenia hipotezy zerowej (wobec hipotezy alternatywnej dwustronnej).

Przykład

Przykład: $X_1 = 1(1)$, $X_2 = 3(2)$, $Y_1 = 6(4)$, $Y_2 = 4(3)$
(w nawiasach rangi).

Suma rang próby kontrolnej: $R = 4 + 3 = 7$.

Czy to dostatecznie mało/dużo do odrzucenia hipotezy zerowej?

Gdy zachodzi H_0 to rangi przypisane próbie kontrolnej są z równym prawdopodobieństwem równe (u, v) dla każdego $1 \leq u < v \leq m + n$.

Rangi	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
R	3	4	5	5	6	7

$$\mathbb{P}(R \geq 7) = \frac{1}{6}.$$

Rozkład sumy rang (R) jest tablicowany dla wielu możliwych n i m .

W pakiecie R do przeprowadzania testu

Manna-Whitneya-Wilcoxa służy funkcja `wilcox.test()`.

```
> wilcox.test(a,b)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: a and b
```

```
W = 89, p-value = 0.007497
```

```
alternative hypothesis: true location shift is not equal to 0
```

```
Warning message:
```

```
In wilcox.test.default(a, b) : cannot compute exact p-value with ties
```

Test Wilcoxona dla par

W przypadku, gdy nie mamy podstaw do zakładania, że różnice między wartościami w każdej parze mają rozkład normalny możemy wykorzystać test Wilcoxona:

1. Sortujemy n par według rosnących **modułów** różnic (między wartościami pary),
2. Nadajemy każdej parze rangę równą pozycji modułu różnicy w uporządkowanym ciągu,
3. Parom o ujemnej różnicy zmieniamy rangi na przeciwne ($x \mapsto -x$),
4. Obliczamy statystykę W_+ równą sumie dodatnich rang,
5. „Zbyt małe” lub „zbyt duże” wartości W_+ świadczą na niekorzyść hipotezy zerowej ($F_X = F_Y$).

$$\mathbb{E}W_+ = \frac{n(n+1)}{4}, \quad \mathbb{V}W_+ = \frac{n(n+1)(2n+1)}{24}.$$

Przykład (koncentracja płytek krwi)

Obliczenia w R

```
> wilcox.test(przed,po,paired=TRUE)
```

Wilcoxon signed rank test with continuity correction

data: przed and po

V = 1, p-value = 0.005056

alternative hypothesis: true location shift is not equal to 0

Warning message:

In wilcox.test.default(przed, po, paired = TRUE) :

cannot compute exact p-value with ties