Statystyka Matematyczna

Anna Janicka

Wykład III, 07.03.2016

MODEL STATYSTYCZNY – CD.

MODEL NORMALNY

Plan na dzisiaj

1. Model statystyczny – cd.
   - pojęcie statystyki
2. Model normalny. Rozkłady statystyk w modelu normalnym
   - Chi-kwadrat
   - t-Studenta
   - F (Fishera-Snedecora)
3. Wstęp do estymacji

Model Statystyczny – przypomnienie

w RP było: $(\Omega, \mathrm{F}, \mathrm{P})$

Model statystyczny: $(\mathrm{X}, \mathrm{F}_{\mathrm{X}}, \mathrm{P})$

gdzie:

X – przestrzeń wartości obserwowanej zmiennej losowej *X* (często *n*-wymiarowa, jeśli mamy *n*-wymiarową próbkę zmiennych $X_1, \ldots, X_n$)

F_{X} – σ-ciało na X

P – rodzina rozkładów prawdopodobieństw P_θ, indeksowana parametrem $\theta \in \Theta$

W mniej formalnym opisie zwykle podaje się: X, P, Θ

Model statystyczny – przykład

Sformułowanie: Kontroler przebadał partię 50 sztuk towaru. Wyniki są następujące (1– towar wadliwy, 0 – bez wad):

0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1

$X = \{0,1\}^n$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

dla $\theta \in [0,1]$

(u nas n=50 oraz $X_2 = X_{10} = X_{15} = X_{32} = X_{42} = X_{50} = 1$, pozostałe $X_i = 0$)

Model statystyczny – przykład cd.

Alternatywne sformułowanie (jeśli notujemy tylko *liczbę* wadliwych elementów w próbie):

X = {0,1, 2, ..., n} – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

dla $\theta \in [0,1]$

(u nas n=50 oraz X=6)

Model statystyczny – przykład cd. (2): pytania

Mamy konkretne dane (próbkę):

- Jaka jest wartość parametru θ?
  - interesuje nas konkretna wartość
  - interesuje nas przedział (*ufności*)

  → *zagadnienie estymacji*

- Weryfikacja hipotezy, że $\theta = 0{,}1$

  → *testowanie hipotez statystycznych*

- → *ew. predykcje*

Model Statystyczny: Przykład 2

Wzrosty na giełdzie. Analityk bada długość *okresów wzrostowych* na giełdzie. Interesuje go czas wzrostu kursu (do pierwszego spadku), w dniach. Załóżmy, że czasy wzrostu $X_1, X_2, \ldots, X_n$ są próbką z rozkładu wykładniczego Exp(λ).

λ – nieznany parametr

X =$(0,\infty)^n$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_\lambda(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = \prod_{i=1}^{n}(1 - e^{-\lambda x_i})$$

$$f_\lambda(x_1, x_2, \ldots, x_n) = \lambda^n e^{-\lambda \Sigma x_i}$$

dla $\lambda > 0$

Model Statystyczny: Przykład 3

Pomiar z błędem losowym: powtarzamy pomiar wielkości μ, wyniki poszczególnych pomiarów są niezależnymi zmiennymi los. $X_1, X_2, \ldots, X_n$, bo maszyna do pomiaru niedoskonała. Każdy z pomiarów ma jednakowy rozkład normalny $N(\mu, \sigma^2)$.

μ, σ^2 – nieznane parametry (a więc $\theta = (\mu, \sigma)$)

$X = \mathrm{R}^n$ – przestrzeń próbkowa

Łączny rozkład prawdopodobieństwa:

$$P_{\mu,\sigma}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) = \prod_{i=1}^{n} \Phi\left(\frac{x_i - \mu}{\sigma}\right) \text{ lub}$$

$$f_{\mu,\sigma}(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

dla $\mu \in \mathrm{R}$, $\sigma > 0$

Statystyki

Estymację parametrów (punktową, przedziałową) czy testowanie hipotez statystycznych przeprowadza się na podstawie tzw. *statystyk*:

Statystyka = dowolna funkcja obserwacji, czyli zmienna losowa postaci

$$T = T(X_1, X_2, ..., X_n)$$

Statystyka nie może zależeć od parametru θ, np. $X_1 + X_2 - \theta$

Statystyki – przykład

$$T_1 = \sum_{i=1}^{n} X_i, \quad T_2 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad T_3 = \frac{1}{n}\sum_{i=1}^{n} X_i - 0{,}1$$

są statystykami dla pierwszego sformułowania;

$$T_1 = X, \quad T_2 = \frac{X}{n}, \quad T_3 = \frac{X}{n} - 0{,}1$$

są statystykami dla drugiego sformułowania

Wybór statystyki zależy od pytania, na które mamy odpowiedzieć.

Rozkład statystyki

Przy odpowiedzi na pytania z wykorzystaniem statystyk będziemy musieli znać **rozkład statystyki**. Mimo że sama statystyka nie zależy od nieznanych parametrów, jej rozkład – owszem.

Ważne rozkłady statystyk: w modelu normalnym

Model normalny

Najbardziej powszechne założenie stosowane przy badaniach statystycznych:

$X_1, X_2, \ldots, X_n$ są próbką z rozkładu normalnego N(μ, σ^2).

Ważne statystyki w tym modelu:

średnia $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

odchylenie std: $S = \sqrt{S^2}$

jakie są ich rozkłady?

Model normalny – cd.

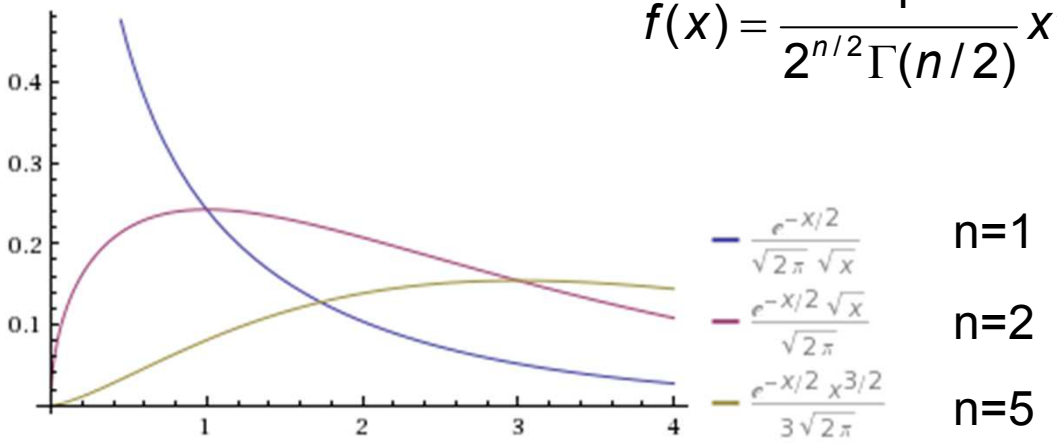
Rozkład $\overline{X}$: jako przeskalowana suma (niezależnych) zmiennych z rozkładu normalnego

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

Rozkład S^2?

Rozkład chi-kwadrat $\chi^2(n)$

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \quad \text{dla } x > 0$$



n=1

n=2

n=5

Suma kwadratów *n* niezależnych zmiennych losowych o rozkładach N(0,1) ma rozkład chi kwadrat z *n* stopniami swobody, $\chi^2(n)$ ew. χ^2_n

$\chi^2(n)$ jest szczególnym przypadkiem rozkładu gamma: Γ(n/2, 1/2)

EX = n, Var$X$ = 2n

Model normalny – cd. (1)

Tw. W modelu normalnym, statystyki $\overline{X}$ i S^2 są niezależnymi zmiennymi losowymi, t. że

$$\overline{X} \sim N(\mu, \sigma^2/n) \qquad \frac{(\overline{X}-\mu)}{\sigma}\sqrt{n} \sim N(0,1)$$

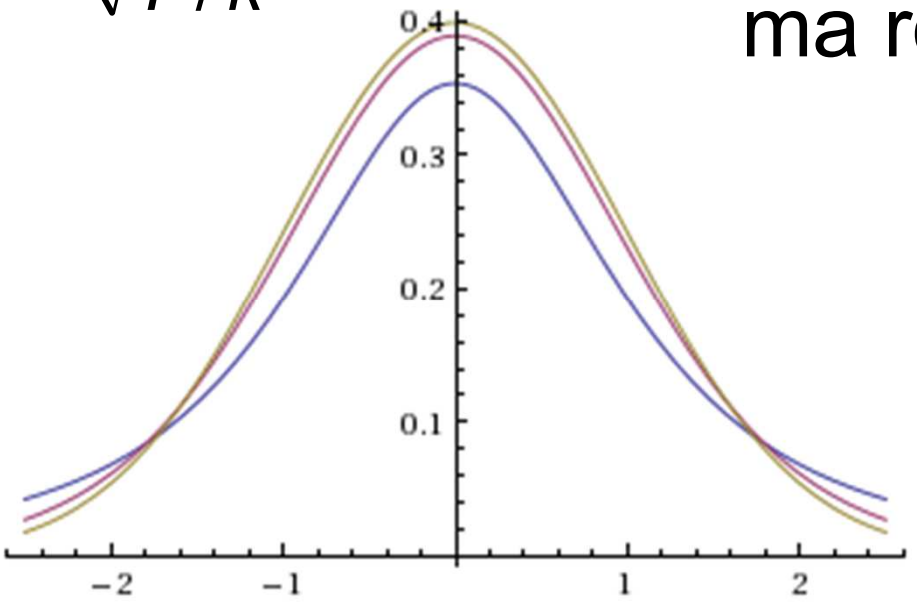
$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

w szczególności:

$$E_{\mu,\sigma} S^2 = \sigma^2, \text{ oraz } \mathrm{Var} S^2 = 2\sigma^4/(n-1)$$

Rozkład t-Studenta *t*(*k*), *k* = 1,2,…

$T = \dfrac{X}{\sqrt{Y/k}}$ dla X i Y niezależnych $X \sim N(0,1)$, $Y \sim \chi^2(k)$ ma rozkład t-Studenta z k stopniami swobody



$\frac{1}{(x^2+2)^{3/2}}$ k=2

$\frac{984\,375}{8(x^2+10)^{11/2}}$ k=10

$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ N(0,1)

$$g_n(x) = \frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma(\frac{1}{2}(n+1))}{\Gamma(\frac{1}{2}n)} \left(1 + \frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)}, \quad n = 1, 2, \ldots.$$

EX = 0 dla k>1, Var$X$ = $k/(k-2)$ dla $k>2$

Model normalny – cd. (2)

Tw. W modelu normalnym, *zmienna*

$$T = \sqrt{n}(\overline{X} - \mu) \Big/ S$$

ma rozkład t-Studenta z n -1 stopniami swobody, $T \sim t(n-1)$

Rozkład F-Snedecora $F(d_1,d_2)$, $d_1,d_2 = 1, 2,\ldots$

F ma rozkład $F(d_1,d_2)$, jeśli $F = \dfrac{Y_1/d_1}{Y_2/d_2}$,

gdzie Y_i są niezależnymi zmiennymi losowymi o rozkładzie $\chi^2(d_i)$

$$g_{d_1,d_2}(x) = \frac{\left(\frac{d_1\,x}{d_1\,x+d_2}\right)^{d_1/2}\left(1-\frac{d_1\,x}{d_1\,x+d_2}\right)^{d_2/2}}{x\,\mathrm{B}(d_1/2,d_2/2)}1_{(0,\infty)}(x)$$

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},\quad a,\,b>0$$

Model normalny – cd. (3)

Jeśli mamy więcej niż jedną (sub)populację:

Tw. Jeśli $X_1, X_2, \ldots, X_n$ są próbką z rozkładu normalnego $N(\mu_X, \sigma_X^2)$, zaś $Y_1, Y_2, \ldots, Y_m$ są (niezależną) próbką z rozkładu normalnego $N(\mu_Y, \sigma_Y^2)$, to *zmienna*

to nie jest statystyka!

$$F = \frac{S_X^2 \sigma_Y^2}{S_Y^2 \sigma_X^2} \sim F(n-1, m-1)$$

przy dodatkowym założeniu $\sigma_X^2 = \sigma_Y^2$, $F = \frac{S_X^2}{S_Y^2} \sim F(n-1, m-1)$ już jest statystyką

Estymacja punktowa

- Wybór, na podstawie danych, *najlepszego* parametru θ spośród parametrów, jakie mogą opisywać rozkład P_θ
- **Esytmator** parametru θ to <u>dowolna</u> statystyka $T = T(X_1, X_2, ..., X_n)$

  o wartościach w zbiorze Θ (interpretujemy ją jako przybliżenie θ). Zwykle zapisywany jako $\hat{\theta}$
- Czasem estymowane nie θ, a $g(\theta)$.

Przykład estymacji – częstość empiryczna

Kontrola jakości:

0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1

- Model stat. np. X = {0,1, 2, ..., n} (tu n=50),

$$P_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{dla } \theta \in [0,1]$$

- parametr θ: p-stwo sztuki wadliwej
- oczywisty estymator: $\hat{\theta} = {}^{X}\!/_{n} = {}^{6}\!/_{50}$

  n – liczebność próby

  X – liczebność wadliwych sztuk

przy innym sformułowaniu modelu,
ten estymator to średnia

Kłopoty z częstością (i nie tylko)...

Przykład: trzy genotypy w populacji, występują w proporcjach $\theta^2 : 2\theta(1-\theta) : (1-\theta)^2$

W populacji *n* osób zaobserwowano odpowiednio N_1, N_2, N_3 osobników poszczególnych genotypów.

Czy powinniśmy wziąć $\hat{\theta} = \sqrt{N_1/n}$, czy raczej $\hat{\theta} = 1 - \sqrt{N_3/n}$, a może $\hat{\theta} = N_1/n + \frac{1}{2} N_2/n$, a może jeszcze jakiś inny estymator?

→ Jak wybrać najlepszy?

Estymacja – statystyki próbkowe

Charakterystyki próbkowe:

estymatory tworzone w oparciu o rozkład empiryczny (dystrybuantę empiryczną)

Dystrybuanta empiryczna

- Niech $X_1, X_2, \ldots, X_n$ – próbka z rozkładu o dystrybuancie F (model z rodziną $\{P_F\}$)

  (n-ta) **dystrybuanta empiryczna**

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,t]}(X_i) = \frac{\text{liczba obserwacji } X_i : X_i \le t}{n}$$

- Dla ustalonych realizacji X_i jest to dystrybuanta rozkładu empirycznego (równomiernego na punktach $x_1, x_2, \ldots, x_n$). Dla ustalonego t jest to statystyka dla próby losowej: zmienna losowa o rozkładzie

$$P(\hat{F}(t) = \tfrac{k}{n}) = \binom{n}{k} F(t)^k (1-F(t))^{n-k}, \quad k = 0,1,\ldots,n$$

Dystrybuanta empiryczna: własności

1. $E_F \hat{F}_n(t) = F(t)$
2. $\mathrm{Var}\hat{F}_n(t) = \frac{1}{n} F(t)(1-F(t))$
3. z CTG: $\frac{\hat{F}_n(t) - F(t)}{\sqrt{F(t)(1-F(t))}} \sqrt{n} \xrightarrow[n\to\infty]{} N(0,1)$

   czyli dla dowolnego *z:* $P\left(\frac{\hat{F}_n(t) - F(t)}{\sqrt{F(t)(1-F(t))}} \sqrt{n} \le z \right) \to \Phi(z)$

4. Tw. Gliwenki-Cantelliego

   (podstawowe tw. statystyki)

   $\sup_{t \in \mathrm{R}} |\hat{F}_n(t) - F(t)| \xrightarrow{p.n.} 0$

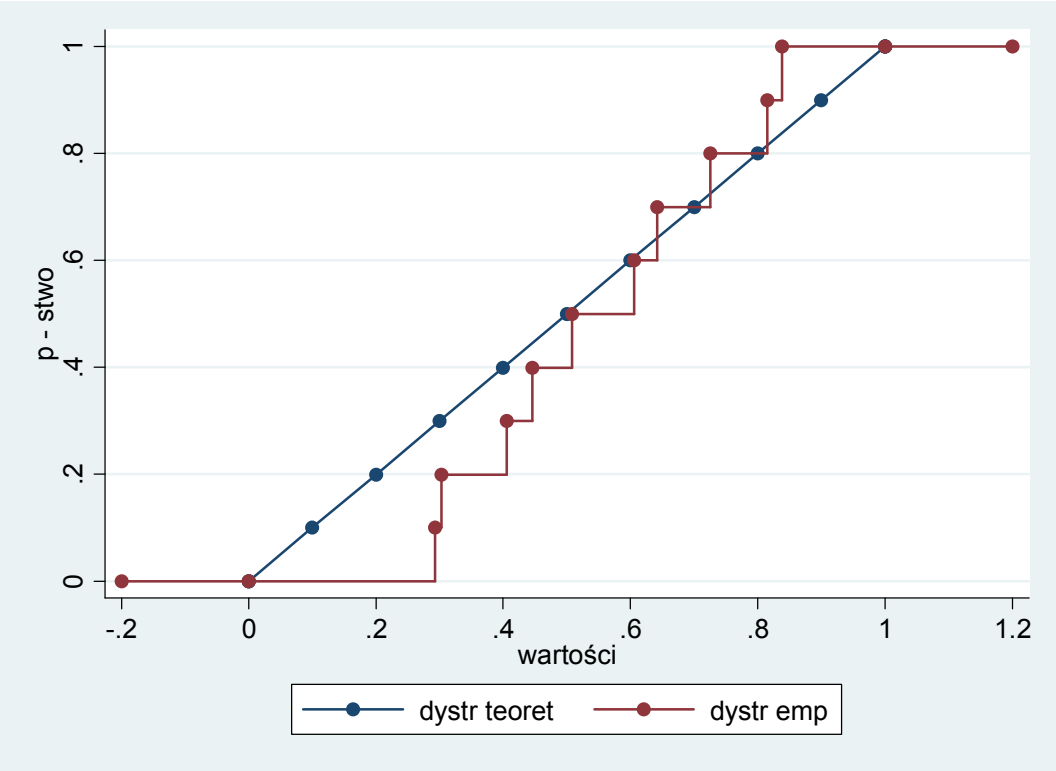
   dla $n \to \infty$

jeśli liczebność próby wzrasta, to możemy poznać nieznany rozkład prawdopodobieństwa z dowolną dokładnością

Dystrybuanta empiryczna – przykład

Dane z rozkładu U[0,1], *n*=10

0,29 0,30 0,40 0,44 0,50 0,60 0,64 0,72 0,81 0,83



Statystyki pozycyjne

- Niech $X_1, X_2, \ldots, X_n$ – próbka z rozkładu o dystrybuancie F. Porządkujemy je w kolejności rosnącej, i oznaczamy

  $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ ← **statystyki pozycyjne**

  (w szczególności $X_{1:n}$ = min, $X_{n:n}$ = max)
- Dystrybuanta empiryczna jest funkcją schodkową, stałą na przedziałach

  $[X_{i:n}, X_{i+1:n})$

Rozkłady statystyk pozycyjnych

- Niech $X_1, X_2, \ldots, X_n$ – niezależne zmienne losowe o dystrybuancie F. Wówczas $X_{k:n}$ ma rozkład o dystrybuancie

$$F_{k:n}(x) = P(X_{k:n} \le x) = \sum_{i=k}^{n} \binom{n}{i} \big(F(x)\big)^i \big(1 - F(x)\big)^{n-i}$$

- jeśli dodatkowo rozkład jest ciągły o gstości f, to $X_{k:n}$ ma rozkład o gęstości

$$f_{k:n}(x) = n \binom{n-1}{k-1} f(x) \big(F(x)\big)^{k-1} \big(1 - F(x)\big)^{n-k}$$


Warsaw University
Faculty of Economic Sciences