Materiały do wykładu PMPiSt, 17.12.2020

Temat. Wprowadzenie do wnioskowania statystycznego.

Wstęp. Będą zajmowali nas POPULACJA GENERALNA $\mathcal{P}$ z cechą X.
Jak wiemy, oznacza to, że X jest zmienną losową (o rozkładzie prawd.) określoną na modelu teoretycznym $\mathcal{P}$, czyli MPK (Ω, Σ, P).
Z praktycznego punktu widzenia $\mathcal{P}$ jest nam potrzebna do wyznaczenia rozkładu X. Robi się to w ramach tzw. WNIOSKOWANIA STATYSTYCZNEGO, korzystając z MATERIAŁU STATYSTYCZNEGO nazywanego PRÓBĄ PROSTĄ.
Z teoretycznego punktu widzenia P.P. ma postać

$$(*) \qquad (x_1, x_2, \dots, x_n) = (X_1, X_2, \dots, X_n)(\omega_0), \text{ dla p. } \omega_0 \in \Omega,$$

gdzie:
(i) $d(X_j) = d(X)$, $j = 1, 2, \dots, n$
(ii) $\{X_1, X_2, \dots, X_n\}$ są st. niez.
(iii) $x_j = X_j(\omega_0)$, $j = 1 \dots n$.

Wtedy $(X_1, X_2, \dots, X_n)$ nazywamy wektorem losowym odpowiadającym $\mathcal{P}$ i X.

Dalej zakładamy, że dla $\mathcal{P}$ i X dana będzie P.P. postaci $(*)$, oraz dla X istnieją $m = EX$ i $\sigma^2 = var(X)$.

Pojęcie statystyki. Ważniejsze przykłady i ich rozkłady.

Def 1.

Każdą zmienną losową Z, która ma postać

$$Z = f(X_1, X_2, .., X_n), \text{ czyli}$$

$$Z(\omega) = f(X_1(\omega), X_2(\omega), .., X_n(\omega)), \quad \omega \in \Omega,$$

gdzie $(X_1, X_2, .., X_n)$ jak w PP. nazywamy STATYSTYKĄ (f oznacza dowolną funkcję nieujemną ciągłą n-argumentów).

Przegląd statystyk

1) $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ – średnia z próby

Wtedy dla ω_0 jak w (*)

$$\bar{X}_n(\omega_0) = \frac{1}{n}(x_1 + x_2 + .. + x_n) \text{ nazywamy wartością}$$

zaobserwowaną $\bar{X}_n$.

Zauważmy, iż

$$E\bar{X}_n = EX = m, \quad var(\bar{X}_n) = \frac{1}{n^2} \cdot n \, var(X) = \frac{var(X)}{n} = \frac{\sigma^2}{n}$$

2) $M_2 = \frac{1}{n} \sum_{k=1}^{n} X_k^2$ – drugi moment z próby

$C_2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - M_2)^2$ – drugi moment centralny z próby

Dalej C_2 oznaczymy przez S^2.

$S^2(\omega_0)$ – wartość zaobserwowana drugiego momentu centralnego.

3) $\hat{S}^2 \stackrel{df}{=} \frac{n}{n-1} S^2 \quad (n \geq 2)$

$\hat{S}^2$ czytamy „S z daszkiem kwadrat".

Wtedy $E\hat{S}^2 = \sigma^2$

Do analizy statystycznej potrzebujemy jeszcze trzech kolejnych statystyk. Ich konstrukcje bazują na założeniu, że cecha X P.G ma rozkład $N(m, \sigma^2)$.

Wtedy dla wektora losowego $(X_1, X_2, \dots, X_n)$ jak w (*)

4) przez statystykę chi-kwadrat Pearsona o n-1 stopniach swobody rozumiemy

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

5) przez <u>statystykę Goseta</u> o $n-1$ st. swobody, albo <u>t-Studenta</u>

rozumiemy

$$t_{n-1} = \frac{\bar{X}_n}{S}\sqrt{n-1}$$

6) $N = \frac{\bar{X}_n - m}{\sigma}\sqrt{n} \in N(0,1)$

<u>Uwagi n/t rozkładów statystyk χ^2_{n-1} i t_{n-1}</u>

Niech χ^2_k oznacza statystykę chi-kwadrat o k st. swobody,

t_k - t-Studenta o k-stopniach swobody.

<u>Wtedy</u>:

1°. rozkłady χ^2_k i t_k <u>są stablicowane</u>

2°. W tablicach wykorzystuje się tzw. <u>wartości krytyczne</u> a nie dystrybuanty.

3°. dla χ^2_k mamy

$$P(\{\omega\in\Omega : \chi^2_k > \chi^2_\alpha\}) = \alpha, \text{ dla } \alpha\in(0,1)$$

Wtedy α nazywamy poziomem istotności,

χ^2_α - wartością krytyczną odpowiadającą α dla tej statystyki o k-st. swobody.

Przykład 1.

Weźmy $\alpha = 0,05$, $k = 10$. Wtedy

$$\chi^2_\alpha = \underline{18,307}$$

4°. dla t_k mamy

$$P(\{\omega \in \Omega : |t_k| > t_\alpha\}) = \alpha$$

$\downarrow$

wartość krytyczna

Przykład 2. Dla $\alpha = 0,1$, $k = 5$

$$\underline{t_\alpha = 2,015}.$$

Uwaga. Z tablic rozkładów χ^2_k i t_k korzystamy tylko dla $k \leq 30$.

Dla $k > 30$, z CTG wynika, iż mamy następujące przybliżenia:

$$P(\{\omega\in\Omega: \chi^2_k < \chi^2_\alpha\}) \cong \Phi(\chi^2_\alpha)$$

$$P(\{\omega\in\Omega: t_k < t_\alpha\}) \cong \Phi(t_\alpha), \quad t_\alpha > 0$$

WNIOSKOWANIE STATYSTYCZNE

Na poprzednim wykładzie przedstawiliśmy opisowo 5 etapów wnioskowania statystycznego.

Teraz przedstawimy metody, które pozwalają realizować etapy 3-5.

ETAP 3: określenie typu rozkładu cechy X - metoda DYSTRYBUANTY EMPIRYCZNEJ.

Niech $(x_1, x_2, \dots, x_n) = (X_1, X_2, \dots, X_n)(\omega_0)$ P.P. jak w (*).

Definiujemy funkcję

$$\mathbb{R} \ni t \longrightarrow S_n(t, \omega_0) \in [0,1],$$

gdzie

$$S_n(t, \omega_0) = \frac{1}{n} \# \{ j : x_j < t \}.$$

Wtedy: ① dla $t \le \min\{x_1, \dots, x_n\}$,

$$S_n(t, \omega_0) = 0, \text{ dlatego } \lim_{t \to -\infty} S_n(t, \omega_0) = 0$$

dla $t > \max\{x_1, x_2, \dots, x_n\}$,

$S_n(t, \omega_0) = 1$ i dlatego

$$\lim_{t \to +\infty} S_n(t, \omega_0) = 1$$

(2) $S_n(t, \omega_0)$ jest niemalejąca

(3) $S_n(t, \omega_0)$ jest lewostronnie ciągła.

Z (1) - (3) wynika, że $S_n(t, \omega_0)$ ma własności funkcji schodkowej - czyli dystrybuanty rozkł. dyskretnego.

Z tego powodu nazywamy ją dystrybuantą empiryczną.

Zadanie

Narysować D.E. dla P.P.

$$(0{,}97,\ 1{,}02,\ 1{,}27,\ 1{,}00,\ 1{,}1,\ 1{,}15,\ 1{,}71)$$

Aby zrozumieć dlaczego D.E. jest ważna w procesie wnioskowania statystycznego potrzebny szerszego spojrzenia na jej definicję.

Załóżmy na chwilę, że przy próbkowaniu przebiega nieskończenie wiele razy, a nie jak do tej pory tylko raz dla pewnego ω_0.

Zatem wygląda on następująco:

$$\Omega \ni \omega \longrightarrow (X_1, X_2, \ldots, X_n)(\omega) =$$

$$= \underbrace{(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))}_{\text{wynik dla } \omega}$$

Pozwala to nam zdefiniować funkcję

$$\Omega \times \mathbb{R} \ni (\omega, t) \longrightarrow S_n(\omega, t) \overset{df}{=} \frac{1}{n} \# \{ j : X_j(\omega) < t \}.$$

Zauważmy, że $S_n(\omega, t)\Big|_{\omega=\omega_0} = S_n(\omega_0, t)$.

<u>Tw. (GLIWIENKO)</u>

Przy założeniach jak wyżej,

$$P(\{\omega \in \Omega : S_n(\omega, t) \xrightarrow[n]{} F_X(t)\}) = 1.$$

Zatem dla odpowiednio dużych n (długości próby!) mamy aproksymację F_X za pomocą D.E, gdzie

$$\boxed{F_X(t) \cong S_n(w_0, t)}.$$

Ostatni wzór stanowi podstawę do określenia typu rozkładu X).

Zadanie 2.

P.6.4.1 – podręcznik

Zadanie 3.

Przeczytać materiał o histogramie (str. 155-160) podręcznika.

Dalej będziemy zakładali, że wynikiem tego etapu analizy statystycznej jest postać dystrybuanty cechy X,

$F_X(t, \theta)$, gdzie θ jest nieznaną wartością jej (jedynego!) parametru.

etap 4 . Metoda ESTYMACJI Przedziałowej.

Cel: określić wartości θ_0 nieznanej wartości parametru θ dystrybuanty cechy X P.G.

Metodologia metody

dla danej P.P. (*) (tej samej co użytej w kroku 3)
i ustalonej liczby $\alpha \in (0,1)$ należy skonstruować dwie statystyki Z_1, Z_2,

czyli $Z_1 = f_1(X_1, X_2 .., X_n)$,

$Z_2 = f_2(X_1, X_2 .., X_n)$,

gdzie:

(i) $Z_1(\omega) < Z_2(\omega)$ z prawd. 1

(ii) $P(\{\omega \in \Omega : \theta \in (Z_1(\omega), Z_2(\omega))\}) = \underbrace{1-\alpha}_{\text{poziom wiarygodności}}$

Wtedy (Z_1, Z_2) nazywamy <u>przedziałem losowym</u>

Biorąc teraz ω_0, gdzie

$$(X_1, X_2, \ldots, X_n)(\omega_0) = (x_1, x_2, \ldots, x_n) \quad (\text{P.P.})$$

otrzymujemy wynik metody:

$$\left[\ \theta \in (\underbrace{Z_1(\omega_0), Z_2(\omega_0)}_{\text{przedział liczby}}) \ \text{ z prawd. } 1-\alpha . \right]$$

Przegląd ważniejszych przypadków procedury ESTYMACJI

Omówmy teraz 3 standardowe sytuacje spotykane w stat. matematycznej, pokazując jak można skonstruować statystyki Z_1, Z_2, a więc i przedział losowy.

(I) Dla P i k cechy X mamy:

- PP. $(x_1, \dots, x_n) = (X_1, X_2, \dots, X_n)(\omega_0)$
- znamy rozkład: $X \in N(m, \sigma^2)$,
  gdzie σ^2 jest znane, $m = ?$.
- ustalmy poziom istotności $\alpha \in (0,1)$.

Definiujemy:

$$Z_1 = \bar{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}}, \quad Z_2 = \bar{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}},$$

gdzie liczbę u_α wyznaczamy z równości

$$P(\{\omega \in \Omega : Z_1(\omega) < m < Z_2(\omega)\}) = 1 - \alpha$$

Po podstawieniu wzorów na Z_1 i Z_2 dostajemy:

$$P(\{\omega \in \Omega : \bar{X}_n(\omega) - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \bar{X}_n(\omega) + u_\alpha \frac{\sigma}{\sqrt{n}}\}) = 1 - \alpha$$

i po przekształceniu

$$P(\{\omega \in \Omega : -u_\alpha < \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < u_\alpha\}) = 1 - \alpha$$

Z przeglądu statystyk, wiemy, że

$$\mathbb{N} = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \in N(0,1), \text{ zatem}$$

ostatnia równość oznacza, że

$$\Phi(u_\alpha) - \Phi(-u_\alpha) = 1-\alpha \equiv$$

$$2\Phi(u_\alpha) - 1 = 1-\alpha \equiv$$

$$\boxed{\Phi(u_\alpha) = 1-\frac{\alpha}{2}}.$$

Np. dla $\alpha = 0{,}05$, $\underline{u_\alpha = 1{,}96 \text{ (tablice!)}}$.

<u>Przykład</u>.

Niech $(2{,}23,\ 2{,}12,\ 1{,}97,\ 2{,}01)$ będzie P.P. cechy X P.G., gdzie metodą D.E. ustalono, że $X \in N(m, (0{,}25)^2)$. Dla $\alpha = 0{,}05$ estymować m.

<u>Uwaga</u>. Wszystkie omawiane, kiedy znane p prawdop. $P(\{\omega \in \Omega : \Theta \in (Z_1(\omega), Z_2(\omega))\})$

nie wymaga się, aby długość przedziału była duża liczba.

Z zadania

$$(2,23,\ 2,12,\ 1,97,\ 2,01) = (X_1, X_2, X_3, X_4)(\omega_0)$$

Dla $\alpha = 0,05$, $u_\alpha = 1,96$, bo $\Phi(u_\alpha) = 1 - \frac{\alpha}{2}$.

Dalej, ponieważ

$$\overline{X}_4(\omega_0) = 2,083,$$ mamy

$$Z_1(\omega_0) = \overline{X}_4(\omega_0) - u_\alpha \frac{\sigma}{\sqrt{4}} = 2,083 - 1,96\,\frac{0,25}{2}$$
$$= 1,838$$

$$Z_2(\omega_0) = \overline{X}_4(\omega_0) + u_\alpha \frac{\sigma}{\sqrt{4}} = 2,083 + 1,96\,\frac{0,25}{2}$$
$$= 2,328.$$

Zatem

$$m \in (1,838,\ 2,328) \text{ z prawd. } 0,95$$