SI materiały do wykładu 15.12.2020

Część I. I. Wprowadzenie do statystyki matemat.

1. Pojęcie Populacji Generalnej (PG), cechy P.G i zjawiska pr[z]ebadanego.

Postawienie problemu.

Będą obserwowali zjawisko o charakterze masowym – co będzie chodziło o mnogości znacznej z punktu widzenia jej liczebności.

Dokładniej, załóżmy że jeśli P oznacza tą mnogość, to

$p \in P \iff p$ ma własność X.

Tak zdefiniowaną mnogość P nazwiemy POPULACJĄ Generalną (P.G), a X jej CECHĄ.

Uwaga. Dana P.G. może mieć wiele cech. Ten kurs ograniczy tylko do pojedynczej cechy.

Przykłady P.G zostały przedstawione na wykładzie inauguracyjnym.

PROBLEM

W jaki sposób opisać P aby można było „zrozumieć" zachowanie każdego $p \in P$, które umożliwiały np. sterowanie układu P, podejmowanie decyzji itp.

Ponieważ, na podstawie def P, o przynależności p do P decyduje to, czy „p ma cechy X", jeśli opiszemy „właściwie" tę cechę, to uzyskamy rozwiązanie problemu.

Masowość zjawiska P wskazuje, iż opis powinien być stochastyczny, zatem X będzie oznaczać rozkład prawdop., czyli zmienną losową.

Wtedy $\forall p \in P$, $X(p)$ będzie miarą numeryczną tej cechy określoną na przypadkach p — jednostki statystyczne.

Podsumowując, rozwiązaniem problemu powinien być rozkład prawdopodobieństwa cechy X P.G.

Ustalenie tego rozkładu odbędzie się w drodze obserwacji P.G.

Ze względu na jej masowy charakter, proces obserwacji będzie zredukowany tylko do skończonego podzbioru $P_0 \subset P$, który będzie stanowić „reprezentację" P.G.

Taki sposób obserwacji P.G. nazwiemy PRÓBKOWANIEM.

2. Model teoretyczny P.G. i proces próbkowania.

Niech dana będzie P.G. $\mathcal{P}$ i jej cecha X.
Wtedy modelem teoretycznym $\mathcal{P}$ jest M.P.K (Ω, Σ, P),
czyli $p \in \mathcal{P} \equiv \omega = p \in \Omega$,
natomiast X jest zmienną losową określoną na Ω,
czyli $\Omega \ni \omega \longrightarrow X(\omega) \in \mathbb{R}$.

Wtedy funkcja

$$\mathbb{R} \ni t \longrightarrow F_X(t) = P(\{\omega \in \Omega : X(\omega) < t\})$$

określa jej rozkład prawdop.

Niech $\mathcal{P}_0 \subset \mathcal{P}$ będzie reprezentacją P.G., czyli
$\mathcal{P}_0 = \Omega_0 = \{\omega_1, \omega_2, \dots, \omega_n\} \subset \Omega$.

Wtedy ciąg liczbowy

$$(x_1, x_2, \dots, x_n) = (X(\omega_1), X(\omega_2), \dots, X(\omega_n))$$

jest wynikiem próbkowania. Aby taki ciąg mógł być wykorzystany w procesie tzw. WNIOSKOWANIA STATYST.

musi być wygenerowany we właściwy sposób. Decydują o tym dwie podstawowe kwestie:

(*)
- n – jego długość musi być „odpowiednio" duże
- liczby x_j muszą pochodzić z „odpowiednio" dobranych jednostek statystycznych.

Te dwie kwestie weryfikuje proces dobrania $P_0 = \Omega_0$.

W modelu teoretycznym przedstawia się to następująco.

Odpowiednio długi ciąg liczby $(x_1, x_2, \dots, x_n)$ spełnia wymagania (*) jeśli:

(i) $\exists\, \omega_0 \in \Omega_0 \; (= P_0)$

(ii) da się dobrać n – stochastycznie niezależnych kopii $X_1, X_2, \dots, X_n$ cechy X, że

$$(x_1, x_2, \dots, x_n) = (X_1, X_2, \dots, X_n)(\omega_0) =$$
$$= (X_1(\omega_0), X_2(\omega_0), \dots, X_n(\omega_0)).$$

Wtedy taki ciąg $(x_1, x_2 .., x_n)$ będziemy nazywali próbą prostą odpowiadającą P.G. P dla cechy X, natomiast obiekt $(X_1, X_2 .., X_n)$ wektorem losowym tej próby.

3. Metodologia analizy statystycznej – wnioskowania statystycznego.

Poniżej przedstawimy kolejne etapy A.S., które mają decydujący wpływ na rozw. problemu.

Etap 1. Identyfikacja P.G. P i jej cechy X, wybór jej reprezentacji P_0

Etap 2. Ustalenie modelu teoretycznego P.G : $(\Omega, \Sigma, P), X$

Etap 3. Przeprowadzenie procesu próbkowania:
- ustalenie $n \geq 2$
- wybór wektora losowego $(X_1 .., X_n)$
- wybór $\omega_0 \in \Omega$ takiego, że

$$(X(\omega_1), X(\omega_2) .., X(\omega_n)) = (X_1, X_2 .., X_n)(\omega_0)$$
$$\| \quad (x_1, x_2, ..., x_n)$$

czyli ustalić, czy pobrany z P w wyniku wyboru jej reprezentujący P_0 materiał statystyczny $(x_1, x_2.., x_n)$ jest próbą prostą, a nie spełnia wymogów

$$(**)\quad (x_1, x_2.., x_n) = (X_1, X_2..., X_n)(\omega_0).$$

Etap 1. Na podstawie P.P. $(**)$ ustalamy typ rozkładu $F_X(t,\theta)$, gdzie na tym etapie wartości parametru θ (może ich być więcej!) jest jeszcze nieustalona.

Przez określenie typu rozumiemy m.in. takie:

- ustalenie czy F_X jest dyskretny, ciągły
- czy np. $B(\underbrace{n,p}_{\text{parametry } \theta})$, $\underset{\theta}{\underset{\parallel}{P(\lambda)}}$, $J(\underbrace{[a,b]}_{\theta})$, $\underset{\theta}{\underset{\parallel}{W(\lambda)}}$, $N(\underbrace{m,\sigma^2}_{\theta})$ itp.

Podstawową metodą jest metoda DYSTRYBUANTY EMPIRYCZNEJ, którą zademonstrujemy dalej.

Etap 5.

Mając $F_X(t,\theta)$ jako wynik Etapu 4, w Etapie 5 określamy wartości θ.

Należy to do metody przybliżania, czyli ESTYMACJI.

Dalej poznamy jedną z metod ESTYMACJI — tzw. estymację przedziałową.

ETAP 6. Na tym etapie wnioskowania mamy już $F(t,\theta_0)$.

Zadaniem ETAPU 6 jest weryfikacja spójności pod kątem niesprzeczności ETAPÓW 1–5.

Do tego celu stosuje się metody HIPOTEZ STATYST. opartych na TESTACH STATYST.

Dalej poznamy je na odpowiednich przykładach.

4. Przedstawienie wybranych etapów oraz ilustracja projektu.

Etapy 1-3 nie są przedmiotem tego kursu.
Skoncentrujemy się na etapach 4-6.

Etap 4. Metoda dystrybuanty empirycznej.

Wprowadzenie.

Niech $(x_1, x_2, .., x_n) = (X_1, X_2 .., X_n)(\omega_0)$ będzie P.P. pobraną z P.G. P dla cechy X.

Definiujemy funkcję

$$\mathbb{R} \ni t \longrightarrow F_n(\omega_0, t) = \frac{|\{i : x_i < t\}|}{n}$$

Dalej nazywamy ją dystrybuantą empiryczną cechy X P.G.
Zaczniemy od wyjaśnienia roli tej funkcji w procesie wnioskowania statystycznego. Podamy na końcu projektu wyjaśnienie dlaczego F_n nazywamy dystrybuantą.

Twierdz o D.E.

Definy funkcję

$$\mathbb{R}\times\Omega \ni (t,\omega) \longrightarrow F_n(t,\omega) \overset{df}{=} \frac{1}{n} \left| \{ i : X_i(\omega) < t \} \right|,$$

(a nie dystrybuanta empiryczna to szczególny przypadek, kiedy $\omega = \omega_0$ i wtedy $(X_1, \dots, X_n)(\omega_0)$ jest P.P.), gdzie X cecha P.G. $\mathbb{P}$.

Wtedy

$$\mathbb{P}\big(\overbrace{\{\omega\in\Omega : F_n(t,\omega) \xrightarrow[n]{} F_X(t)\}}^{A}\big) = 1,$$

co oznacza, że dla odpowiednio dużych n mamy przybliżenie

$$F_X(t) \approx F_n(t,\omega).$$

Ponieważ zdanie A ma prawd. 1, mamy, że

$$F_X(t) \simeq F_n(t,\omega_0) \text{ z prawd. } 1,$$

dlatego ω_0 → możemy użyć P.P.

Przykład 1

Dla cechy X pewnej P.G. P pobrano P.P.

$(1{,}2,\ 1{,}0,\ 1{,}5,\ 0{,}98,\ 0{,}97)$.

Narysować dystrybuantę empir. odpowiadającą tej P.P.

Rozw.

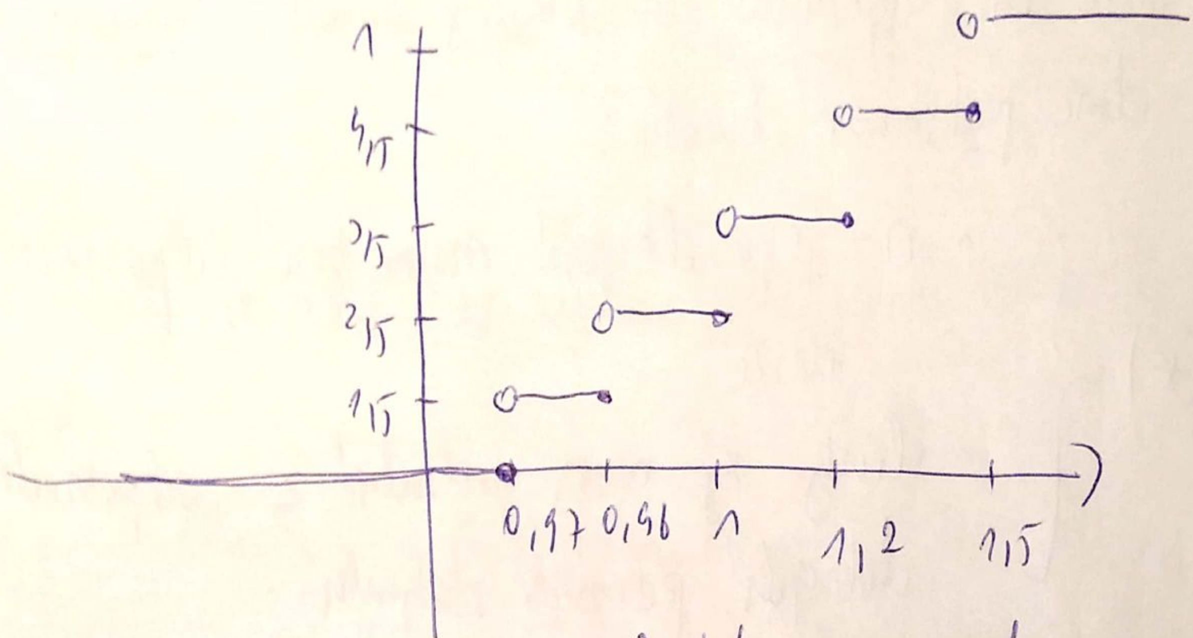
Z założenia mamy

$$(1{,}2,\ 1{,}0,\ 1{,}5,\ 0{,}98,\ 0{,}97)=(X_1, X_2, X_3, X_4, X_5)(\omega_0),$$

gdzie $X_1, X_2, \ldots, X_5$ stochast. niezależne o rozkł. X, dla p. $\omega_0 \in \Omega$.

Zgodnie z def. D.E. mamy

$$F_5(t,\omega_0)=\begin{cases} 0, & t \le 0{,}97 \\ 1/5 & 0{,}97 < t \le 0{,}98 \\ 2/5 & 0{,}98 < t \le 1{,}0 \\ 3/5 & 1{,}0 < t \le 1{,}2 \\ 4/5 & 1{,}2 < t \le 1{,}5 \\ 1 & 1{,}5 < t \end{cases}$$



Zatem $F(t,\omega)$ j. schodkową <u>dystrybuantą r. prawd.</u>

Na ~~[illegible]~~ podstawie <u>materiału historycznego</u>, metodą np. porównania ustala się typ $F_X(t,\theta)$ wykorzystując wykresy D.E. W ten sposób realizuje się <u>ETAP 4.</u>

<u>Uwaga.</u>

Istnieje metoda alternatywna — oparta na pojęciu HISTOGRAMU

<u>ZAD 1</u>. Przeczytać wyk. fragm. podręcznika o histogramie. Narysować histog. dla danych z P1.

ETAP 5 – wprowadzenie do t. estymacji przedziałowej.

Zadanie.

Dla cechy X P.G. w wyniku próbkowania pobrano P.P.:

$$(x_1, x_2, \ldots, x_n) = (X_1, X_2, \ldots, X_n)(w_0)$$

i metodą np. D.E. ustalono typ $F_X(t, \theta)$, gdzie $\theta = ?$.

(Np. ustalono, iż $X \sim P(\lambda)$, $\theta = \lambda = ?$; $X \sim N(m, \sigma^2)$, $\theta = m = ?$ $\vee$ $\theta = \sigma^2 = ?$)

Przypomnienie:

(i) dla wektora losowego $(X_1, X_2, \ldots, X_n)$ potrafimy skonstruować dwie zmienne losowe Z_1, Z_2:

$$Z_1(w) = f_1(X_1(w), X_2(w), \ldots, X_n(w)),$$

$$Z_2(w) = f_2(X_1(w), X_2(w), \ldots, X_n(w))$$

dla pewnych funkcji f_1, f_2, gdzie:

$Z_1(\omega) < Z_2(\omega) \quad \omega \in \Omega$

2) dla ustalonej liczby $\alpha \in (0,1)$ – tzw. POZIOM ISTOTNOŚCI (np. $\alpha = 0,05$; $\alpha = 0,1$)

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

POZIOM WIARYGODNOŚCI

Wtedy (Z_1, Z_2) nazywamy przedziałem losowym

Biorąc teraz ω_0 (z P.P.) mówimy, iż

$\theta \in (Z_1(\omega_0), Z_2(\omega_0))$ z prawd. $1-\alpha$

Z tego powodu przedział nazywamy estymacją przedziałową.

Dalej pokażemy dla wybranych rozkładów sposób implementacji estymacji przedziałowej.