Materiały do wykładu PMPiSt 10.12.2020

Temat. Wprowadzenie do STATYSTYKI MATEMATYCZNEJ.

Wstęp. W Teorii Prawdopodobieństwa ilekroć wspominamy o rozkładzie prawdopodobieństwa, to zakładamy, że

1°. Istnieje co najmniej jedno ZL

2°. jego model - MDK (Ω, Σ, P)

3) zmienna losowa X, i

jej rozkład, czyli funkcja

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

Znana.

Dalej T.P. pokazuje jak obliczyć $m_X = EX$, $\sigma_X^2 = Var(X)$, jak już znamy m_X, σ_X^2 w procesie obserwacji ZL, oraz dla $A \in \Sigma$, potrafi obliczyć $P(A)$.

Oczywiście T.P. dostarcza innych faktów, np. podaje przykłady rozkładów prawdopodobieństwa, bada ich własności asymptotyczne (tw. graniczne).

Wszystko co zostało ustalone w ramach T.P. jest do dyspozycji S.M. W takim razie, co jest głównym zadaniem SM?

Odpowiedź brzmi: ustalenie postaci rozkładu obserwowanego ZL, czyli na wyjściu badania statystycznego pytamy o $F_X = ?$

Na wejściu tego badania mamy to co pojawia się na wyjściu T.P., czyli znamy F_X i od tego momentu inicjatywę przejmuje T.P.

S.M. na tyle usamodzielniła się, że wprowadziła „własne" pojęcia i metody, chociaż większość z nich za podstawy przyjmuje fundamentalne twierdzenia T.P. (np. tw. graniczne).

ZL w S.M. postrzegane jest następująco:

jest to mnogość, gdzie o przynależności do niej decyduje własność, która spełniona jest dla wszystkich jej elementów.

Tę mnogość nazywamy POPULACJĄ GENERALNĄ, i oznaczamy przez $\mathcal{P}$.

Każdy jej element $p \in \mathcal{P}$ nazwiemy JEDNOSTKĄ STATYSTYCZNĄ, gdzie

$p \in \mathcal{P} \iff p$ ma własność, którą oznaczamy przez X i nazywamy CECHĄ $\mathcal{P}$.

W takim razie możemy:

a) dla danej cechy X dobrać odpowiadającą jej populację $\mathcal{P}$

b) dla danej populacji $\mathcal{P}$, określić cechy X.

<u>Uwaga.</u> Na ogół dla danej populacji $\mathcal{P}$ mamy kilka cech: X, Y, Z itd.

<u>W tym kursie będą rozważali tylko przypadek pojedynczej cechy.</u>

Przykład 1

Populacja P	cecha X	Obszar
1) WYBORCY	PREFERENCJE WYBORCZE	np. POLITOLOGIA
2) Chory / potencjalnie chory	zespół chorobowy	ZDROWIE " "
3) Wolumen Produkcji	wadliwość	działalność gospodarcza (Controling)
4) RYNEK PAPIERÓW WARTOŚCIOWYCH	ZMIANA WARTOŚCI JEDNOSTKI	RYNEK FINANSOWY (np. GIEŁDA)
5) Obszar terenu i czas	czynnik klimatyczny (np. temp., ciśnienie, siła i kierunek wiatru, rodzaj opadu i jego intensywność)	Prognozowanie pogody
6) Towar (jako podaż na RYNKU)	zmiana wartości jego jednostki	RYNEK OBROTU, np. MANDEL

Model teoretyczny populacji i jej cechy

Treści PA podporządkowane, i.e.:

a) Populacja jest zjawiskiem MASOWYM, a nie zbiorem jest „duży” ze względu na swoją miarę liczebności = MOC

b) Postrzegamy populacji poprzez FILTR w postaci cechy X, jeśli ta cecha ~~...~~ zostanie „dobrze” ~~zdia~~ zidentyfikowana pozwoli zrozumieć zachowanie się populacji, w szczególności pozwoli je sterować, wywierać na nią wpływ

Ze względu na masowy charakter P.G. – $\mathcal{P}$, cechy X opisuje się w kategoriach T.P.

Dlatego: (•) $\mathcal{P}$ przedstawia się jako M.P.K. (Ω, Σ, P), czyli $p \in \mathcal{P}$ oznacza, że $\omega_p \in \Omega$

$$(p \longleftrightarrow \omega_p)$$

(••) X oznacza rozkład prawdopodobieństwa, czyli F_X, którego na tym etapie nie znamy!

– J –

(...) Kluczowym problemem jest wyznaczenie rozkładu F_X

Opis metodologii umożliwiającej rozwiązanie problemu (...)

I) Określenie ~~...~~ P.G. $\mathbb{P}$ oraz jej cechy X

(p. np. Przykład 1)

II) Pobranie MATERIAŁU STATYSTYCZNEGO:

W tym celu wybiera się z $\mathbb{P}$ podzbiór $\mathbb{P}_0$, (czyli dla modelu z Ω, $\Omega_0 = \{\omega_1, \omega_2, \dots, \omega_n\}$), gdzie $\mathbb{P}_0$ „właściwie reprezentuje" $\mathbb{P}$ (analiza P.1) i „zbadanie" tego podzbioru (np. metodą ankietowania).

W modelu oznacza to, że

$$\Omega_0 = \{\omega_1, \omega_2, \dots, \omega_n\} \longrightarrow (X(\omega_1), X(\omega_2), \dots, X(\omega_n))$$

$$= (x_1, x_2, \dots, x_n), \quad x_k \in \mathbb{R}, \quad k = 1, 2, \dots, n,$$

czyli dla $1 \le j \le n$

$$\omega_j \longrightarrow X(\omega_j) = x_j \in \mathbb{R} \text{ - liczbowy}$$

wynik ankietowania ω_j.

gdzie n „odpowiednio duża” dobrana liczba.
TAK POBRANY MATERIAŁ STATYSTYCZNY NAZYWAMY PRÓBĄ PROSTĄ (PP), a proces jej pozyskiwania nazywamy PRÓBKOWANIEM.

Z teoretycznego punktu widzenia konstrukcję P.P. można przeprowadzić inaczej.

a) dla danej cechy X i $n \geq 2$ bierzemy n-niezależnych kopii $X_1, X_2, \ldots, X_n$, czyli $d(X_i) = d(X)$.

b) bierzemy obiekt, który w T.P. nazywa się WEKTOREM LOSOWYM

$$(X_1, X_2, \ldots, X_n)$$

Można wykazać, że jeśli P.P. ma postać

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

to istnieje $\omega_0 \in \Omega$, że

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

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

Dalej będą założenia, że P.P. $(x_1, x_2, \dots, x_n)$ dana jest jak w $(\#)$.

Uwaga

1) Jak zobaczymy dalej kluczową sprawę jest długości PP, czyli $n \geq 2$.

2) $(\#)$ możemy zinterpretować następująco: zamiast ankietować n jednostek statystycznych z P_0 (czyli Ω_0), wybrano pojedynczą ($\omega_0 \in \Omega$) i w sposób niezależny (stochastyczne) zbadano ją pod kątem cechy X.

3) Przypuśćmy na chwilę, że dla P.G. $\mathcal{P}$ z cechą X pobraliśmy P.O.

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

gdzie $\{X_1, \ldots, X_n\}$ stochastycznie niezależne i $d(X_j) = d(X)$, $j = 1, 2, \ldots n$.

<u>Co nam to daje?</u>

Na pewno mamy wszystko na temat $\mathcal{P}_0$. Ale $\mathcal{P}_0$ to nie $\mathcal{P}$! Zatem P.P. nie jest rozwiązanym problemem.

W takim razie <u>potrzebna jest metoda</u> (metody), która pozwala <u>ekstrapolować</u> wyniki uzyskane <u>dla $\mathcal{P}_0$</u> na <u>całe</u> P.G. $\mathcal{P}$.

Idea tej ekstrapolacji jest następująca:

III. Ustala się najpierw TYP rozkładu cechy X, czyli np. – czy jest dyskretny, czy ciągły
– jeśli dyskretny, to czy np. dwupunktowy, B – Bernoulliego, P – Poissona
– jeśli ciągły, to czy np. J – jednostajny, W – wykładniczy, N – normalny,

bez wnikania w szczegóły – wartości parametrów, np. jako P(λ), to na tym etapie nie ustala się λ, J[⟨a,b⟩], a, b itd.

Robi to się za pomocą pojęcia DYSTRYBUANTY EMPIRYCZNEJ, co pokażemy dalej.

Po identyfikacji typu rozkładu cechy X P.G. P na podstawie P.P., tej samej próby używa się w kolejnym kroku metodologii.

IV Ustalenie wartości parametrów rozkładu, czyli mamy $F_X(\theta)$, nie znamy $\theta = ?$

Należą tu metody ESTYMACJI, o czym pomówimy później.

Zatem ten etap daje $F_X(\theta_0)$, gdzie θ_0 ustalono metodą IV, F_X metodą III z tej samej P.P.

Tak uzyskane wyniki na każdym etapie metodologii poddaje się na koniec weryfikacji (rodzaj AUDYTU).

(V) Należą się tu metody TESTOWANIA HIPOTEZ STATYSTYCZNYCH.

Jeśli faza (V) wypadnie POZYTYWNIE, uzyskane wcześniej wyniki z fazy I–IV mają swoje „potwierdzenie" — stają się „wiarygodne".

Wtedy ustalony rozkład cechy X p.G. P, który modelu T.P. pozwala mierzyć zachowania się P, np.

P.2. Weźmy sytuację (1) z P.1. oraz załóżmy, iż $X(\omega) = \{1, 2, \dots, k\}$, gdzie

$j \in X(\omega)$ — numer listy zgłoszonej w procesie wyborczym.

Jeśli znamy rozkład X, czyli

1	2	3	...	k
p_1	p_2	p_3	...	p_k

, do naszej przewidzieć zwycięzcę, biorąc $\max_{1 \le j \le k} \{p_j\}$.

Na kolejnych wykładach omówimy szczegółowo fazy metodologii III – V.