Rozdział 6

Wstęp do statystyki matematycznej

6.1 Cecha populacji generalnej

W rozdziale tym zaprezentujemy metodę probabilistycznego opisu zaobserwowanego zjawiska. W takim razie (patrz rozdział 2.4) zjawisko to będziemy nazywali zjawiskiem losowym. Będziemy zakładali, że mamy do czynienia z pewną mnogością charakteryzującą się tym, że jej wszystkie elementy posiadają tę samą interesującą nas własność, którą chcemy poznać.

Zbiorowość tę dalej będziemy nazywali *populacją generalną*, a badaną własność tej populacji– *cechą populacji generalnej* i będziemy ją oznaczali przez $\mathbb{X}$. U podstaw metodologii poznania własności cechy $\mathbb{X}$ danej populacji generalnej leży *proces próbkowania*. Chodzi o to, że na ogół populacja generalna jest bardzo liczna i nie ma możliwości, aby własności cechy $\mathbb{X}$ tej populacji można by było poznać, obserwując całą populację. Wybiera się wtedy jej reprezentację. W wyniku obserwacji elementów tej reprezentacji pozyskujemy tzw. *materiał statystyczny*. Proces ten dalej krótko będziemy nazywali *próbkowaniem*, a reprezentację populacji generalnej, na której przeprowadzamy próbkowanie, *jej próbą*. Jeśli przyjmiemy, że próba jest n–elementowym podzbiorem populacji generalnej, to uzyskany w wyniku obserwacji tej próby materiał statystyczny możemy opisać w postacji ciągu $(x_1 \dots x_n)$, gdzie kolejne wyrazy x_j opisują tę samą własność, a więc własność cechy $\mathbb{X}$. W dalszym ciągu będziemy zakładali, że wartości x_j są wielkościami wymiernymi, a więc liczbami rzeczywistymi. Zatem materiał statystyczny będzie miał postać ciągu liczbowego skończonego. Będziemy go nazywali *próbą cechy* $\mathbb{X}$ z populacji generalnej.

Mając już pobrany materiał statystyczny skonstruujemy odpowiadający mu model probabilistyczny, a więc własności cechy $\mathbb{X}$ opiszemy na tej podstawie w kategoriach teorii prawdopodobieństwa. Na gruncie tej teorii zostanie on zbadany. Pokażemy jak otrzymane wyniki odnoszące się tylko do pewnej podzbio-

rowości całej populacji generalnej (jej próby) będzie można przenieść na cała populację i jak takie uogólnienie należy rozumieć.

Prześledźmy wprowadzone wyżej pojęcia na następującym przykładzie:

Przykład 6.1.1 *Próbujemy określić preferencje wyborcze obywateli danego państwa. Przynależność do populacji wyborców określa wtedy obowiązująca ordynacja wyborcza. Przykład naszego państwa pokazuje, że liczebność tej populacji może być całkiem spora (przeszło 20 mln.). Ale nie tylko to stanowi problem. Zbiorowość ta zawsze jest bardzo zróżnicowana, co związane jest z miejscem zamieszkania (jaki region, miasto, wieś itp.), wiekiem, wykonywanym zawodem, płcią itd. Aby materiał statystyczny był wiarygodny, należy ten podział i liczebność populacji generalnej uwzględnić, wybierając jej reprezentację. Powiedzmy wyraźnie, że to nie będzie przedmiotem tego opracowania, aczkolwiek ta faza analizy bywa kluczowa. Z technicznego punktu widzenia, po wybraniu reprezentacji, materiał statystyczny pozyskuje się techniką ankietowania. Również sama konstrukcja takiej ankiety nie będzie przedmiotem naszej analizy. Wspominamy o tym tylko dlatego, aby Czytelnik lepiej mógł sobie wyobrazić proces, który próbujemy tutaj opisać.*

W efekcie otrzymamy próbę $(x_1 \dots x_n)$. *Okazuje się, że w efekcie zamodelowania tego procesu i dodatkowych założeń na tę próbę, stosując metody statystyki matematycznej, będzie można odpowiedzieć np. na następujące pytania:*

1. *najprawdopodobnie na kogo zagłosują mieszkańcy danego województwa;*
2. *najprawdopodobnie na kogo zagłosują ludzie z wyższym wykształceniem;*
3. *najprawdopodobnie na kogo zagłosują ludzie powyżej 55 roku życia;*

i na wiele innych podobnych.

6.2 Model probabilistyczny próby prostej

Zaczniemy od następującego przykładu:

Przykład 6.2.1 *Wyobraźmy sobie, że pewien zakład produkuje w określonym cyklu czasowym (np. dziennym) pewną partię tego samego produktu. Cała ta partia z założenia ma trafić do dystrybucji. Aby ustrzec się przed wpadką, firma zmuszona jest do przeprowadzania kontroli bieżącej produkcji. Z powodu liczebności wyprodukowanej partii i braku czasu nie można tą kontrolą objąć całej produkcji.*

W tym przypadku stosuje się również metody probabilistyczne. Z całej partii, kierując się wieloma względami (np. zmianowym trybem pracy, godzinami pracy itd.), po skontrolowaniu tylko wybranych produktów, a więc elementów próby, pozyskuje się materiał statystyczny. Jego analiza na przykładzie skonstruowanego modelu probabilistycznego pozwala na uzyskanie np. odpowiedzi na następujące pytanie:

jak często w całej populacji pojawiają się detale wadliwe,

o ile ocena przydatności produktu odbywała się według najprostszego kryterium polegającego na ocenie w kategoriach dobry–zły.

W przykładzie 5.1.1 wspomnieliśmy, że materiał statystyczny musi spełniać jeszcze jeden bardzo ważny wymóg, aby analiza statystyczna takiej próby mogła spełnić swoje zadanie. Poniżej zajmiemy się między innymi i tym zagadnieniem. W pierwszej kolejności pokażemy, jak można zbudować model probabilistyczny pozwalający na podstawie materiału statystycznego opisać badaną cechę danej populacji generalnej. Zrobimy to w dwóch etapach. W etapie pierwszym przyjmiemy, że:

1. cecha $\mathbb{X}$ populacji generalnej jest zmienną losową o rozkładzie F, który jest nieznany i próbujemy go poznać,

2. zmienna ta określona jest na przestrzeni zdarzeń Ω_o, która reprezentuje populację generalną związaną z tą cechą,

3. σ-ciało zdarzeń Σ_o jest generowane przez rodzinę

$$\{\omega \in \Omega_o : \mathbb{X}(\omega) < t,\ t \in \mathbf{R}\},$$

4. funkcja prawdopodobieństwa P_o jest taka, że

$$F(x) = P_o(\{\omega \in \Omega_o : \mathbb{X}(\omega) < x\}),\ \text{dla każdego}\ \ x \in \mathbf{R}.$$

Wtedy Ω_o jako zbiór wszystkich zdarzeń elementarnych przedstawia populację generalną, a przyporządkowanie

$$\Omega_o \ni \omega \to \mathbb{X}(\omega) \in \mathbf{R}$$

opisuje proces obserwacji własności cechy $\mathbb{X}$ dla danego elementu ω wybranego z tej populacji. Natomiast n–elementową reprezentację tej populacji-jej próbę możemy traktować jako n–elementowy podzbiór całej populacji

$$\Omega_n = \{\omega_1^o \dots \omega_n^o\}.$$

Materiał statystyczny-próba z populacji generalnej, będzie miał postać

$$\Big(\mathbb{X}(\omega_1^o) \dots \mathbb{X}(\omega_n^o)\Big).$$

W fazie drugiej konstrukcji modelu probabilistycznego odpowiadajęcego zjawisku obserwacji populacji generalnej na podstawie wyboru próby skorzystamy z wyników podrozdziału 2.3, gdzie była mowa o prawdopodobieństwie produktowym. Powtarzając tamtą konstrukcję $n \geq 2$–razy otrzymamy:

$$\Omega = \underbrace{\Omega_o \otimes \dots \Omega_o}_{n-razy},$$

z σ ciałem produktowym Σ i z prawdopodobieństwem produktowym P.

Na tak skonstruowanej przestrzeni probabilistycznej zdefiniujemy następujący ciąg zmiennych losowych:

$$\mathbb{X}_j(\omega) = \mathbb{X}_j(\omega_1, \dots \omega_j, \dots \omega_n) = \mathbb{X}(\omega_j)$$

dla każdego $j \in \{1 \dots n\}$.

Zachodzi następujące twierdzenie (patrz też Dodatek)

Twierdzenie 6.2.1 *Niech zmienne $\mathbb{X}_j$ będą określone na przestrzeni probabilistycznej (Ω, Σ, P) jak wyżej. Wtedy:*

1.

zmienne losowe $\mathbb{X}_j$ mają jednakowe rozkłady jak rozklad cechy $\mathbb{X}$,

2.

zmienne te są niezależne,

3.

$$(\mathbb{X}_1, \ldots \mathbb{X}_n)(\omega^o) = (\mathbb{X}(\omega_1^o), \ldots \mathbb{X}(\omega_n^o)) = (x_1, \ldots x_n),$$

gdzie $\omega^o = (\omega_1^o, \ldots \omega_n^o)$.

Możemy wreszcie doprecyzować pojęcie próby, o czym wspominaliśmy wcześniej.

Definicja 6.2.1 *Niech dany będzie ciąg liczb* $(x_1 \ldots x_n)$ *będący efektem obserwacji cechy* $\mathbb{X}$ *na przykładzie wybranej próby populacji generalnej.*

Powiemy, że ciąg ten jest próbą prostą, jeśli istnieje przestrzeń probabilistyczna i n zmiennych losowych niezależnych o tym samym rozkładzie co badana cecha $\mathbb{X}$*, że zachodzi wzór (3) powyższego twierdzenia.*

Spróbujmy przybliżyć lepiej znaczenie tej definicji. Do tej pory własności tej samej cechy obserwowaliśmy na różnych elementach populacji generalnej wybierając z niej próbę. Tak naprawdę chodziło nam o coś więcej, aby ta obserwacja poszczególnych elementów próby przebiegała w sposób *niezależny*. Na etapie pierwszym opisu tego modelu przetłumaczenie empirycznie rozumianej niezależności sprawia problemy. Stąd w kroku drugim, biorąc za wzór model produktowy, tę niezależność dostaliśmy niejako za darmo. Zmieniła się jednak interpretacja całego procesu pozyskiwania materiału statystycznego. Bowiem fakt, że zaczęliśmy operować ciągiem zmiennych losowych niezależnych o rozkładzie tym samym co cecha $\mathbb{X}$ oznacza, że n–krotnie powtarzaliśmy w sposób niezależny od siebie to samo doświadczenie (patrz też przykład 2.4.2).

A zatem zwrot:

> z populacji generalnej w wyniku obserwacji jej cechy $\mathbb{X}$ pobrano materiał statystyczny w postaci próby prostej $(x_1 \ldots x_n)$,

w myśl powyższych ustaleń oznacza, że każda liczba x_j jest zaobserwowaną wartością zmiennej losowej $\mathbb{X}_j$, dla pewnego zdarzenia elementarnego ω_o, gdzie zmienne te mają ten sam rozkład co cecha $\mathbb{X}$ i są parami niezależne. Z drugiej strony możemy mówić o odwzorowaniu, które było przedmiotem naszych rozważań w rozdziale 4, czyli *wektora losowego*

$$\Omega \ni \omega \to (\mathbb{X}_1, \ldots \mathbb{X}_n)(\omega),$$

które w statystyce odgrywa kluczową rolę. W takim razie przy tej interpretacji próba prosta jest wartością wektora losowego o składowych będących niezależnymi zmiennymi losowymi o tym samym rozkładzie co cecha $\mathbb{X}$.

Przykład 6.2.2 *Należy ocenić partie produktu finalnego pod kątem jego wadliwości. Z teoretycznego punktu widzenia cecha $\mathbb{X}$ ma rozkład dwupunktowy. Zakładając, że obserwacja dobrego produktu zwraca wartość 1, a wadliwego 0, dostaniemy*

$$P(\{\omega \in \Omega : \mathbb{X}(\omega) = 1\}) = p, \ P(\{\omega \in \Omega : \mathbb{X}(\omega) = 0\}) = 1 - p.$$

Zatem ocenę można sprowadzić do prostego pytania: jaka jest wartość liczbowa parametru p?

Przypuśćmy, że dysponujemy n–elementową próbą prostą

$$(x_1 \dots x_n), \ \text{gdzie } x_j \in \{0, 1\}.$$

Istnieje więc wektor losowy $(\mathbb{X}_1 \dots \mathbb{X}_n)$, taki że

1.
$$d(\mathbb{X}_j) = d(\mathbb{X}),$$

2.
$$\mathbb{X}_j \text{ są niezależne,}$$

3.
$$\exists_{\omega_o \in \Omega} \ \ (x_1 \dots x_n) = (\mathbb{X}_1, \dots \mathbb{X}_n)(\omega_o).$$

Z drugiej strony, z Mocnego Prawa Wielkich Liczb wiadomo, że jeśli weźmiemy ciąg zmiennych losowych

$$\overline{\mathbb{Y}_k} = \frac{1}{k}(\mathbb{Y}_1 + \dots \mathbb{Y}_k),$$

gdzie $\mathbb{Y}_k$ mają jednakowe rozkłady, są niezależne i mają drugie momenty, to

$$\overline{\mathbb{Y}_k}(\omega) \longrightarrow p, \ \text{dla} \ \ \omega \in \Omega_1 \ \ i \ \ P(\Omega_1) = 1.$$

W naszym przypadku, gdybyśmy wiedzieli, czy:

1. *zdarzenie elementarne ω_o określające naszą próbę prostą jest elementem zdarzenia Ω_1,*
2. *liczebność próby prostej jest dostatecznie duża,*

to moglibyśmy ustalić na tej podstawie następujące przybliżenie

$$p \simeq \frac{1}{n}(x_1 + \dots x_n).$$

Dalej spróbujemy udzielić odpowiedzi na postawione wyżej pytania.

6.3 Pojęcie statystyki

Zmienne losowe $\overline{\mathbb{Y}_k}$ użyte w przykładzie 5.2.2 stanowią przykład tzw. *statystyki.*

Definicja 6.3.1 *Każdą zmienną losową $\mathbb{Z}$, która powstaje poprzez złożenie wektora losowego*

$$(\mathbb{X}_1 \dots \mathbb{X}_k), \text{ gdzie } \mathbb{X}_j \text{ są o tym samym rozkładzie co cecha } \mathbb{X}$$

z rzeczywistą funkcją ciągłą k–zmiennych f (patrz Dodatek), będziemy nazywali statystyką.

Mamy więc

$$\mathbb{Z}(\omega) = f((\mathbb{X}_1(\omega) \dots \mathbb{X}_k(\omega)) \text{ dla } \omega \in \Omega.$$

W przykładzie 5.2.2 funkcja f określona jest wzorem

$$f(u) = \frac{1}{k}(u_1 + \dots u_k), \text{ gdzie } u = (u_1, \dots u_k).$$

Przyjmijmy następujące oznaczenia:

$$\text{dla próby prostej } (x_1 \dots x_n) = (\mathbb{X}_1 \dots \mathbb{X}_n)(\omega^o)$$

symbolem

$$\overline{\mathbb{X}}_n = \frac{1}{n}(\mathbb{X}_1 + \dots \mathbb{X}_n)$$

oznaczymy statystykę zwaną *średnią z próby.*

Natomiast jej wartość dla zdarzenia elementarnego ω_o

$$\overline{x_n} = \overline{\mathbb{X}}_n(\omega_o)$$

będziemy nazywali *średnią empiryczną z próby.*

Wśród ważniejszych statystyk należy wymienić następujące:

Definicja 6.3.2 *Niech $(x_1 \dots x_n) = (\mathbb{X}_1 \dots \mathbb{X}_n)(\omega_o)$ będzie próbą prostą.*

1. *Momentem rzędu $k \geq 1$ z próby będziemy nazywali statystykę*

$$\mathbb{M}_k = \frac{1}{n}(\mathbb{X}_1^k + \dots \mathbb{X}_n^k).$$

Z momentem tym związany jest moment empiryczny rzędu k

$$m_k = \frac{1}{n}(x_1^k + \dots x_n^k).$$

2. Momentem centralnym rzędu $k \geq 1$ z próby będziemy nazywali statystykę

$$\mathbb{C}_k = \frac{1}{n}\sum_{j=1}^{n}(\mathbb{X}_j - \mathbb{M}_1)^k.$$

Podobnie jak wyżej, dla momentu empirycznego mamy

$$c_k = \frac{1}{n}\sum_{j=1}^{n}(x_j - m_1)^k,$$

a liczbę c_k nazywamy empirycznym momentem centralnym rzędu k.

Uwaga 6.3.1 *Zauważmy, że* $\mathbb{M}_1 = \overline{\mathbb{X}}_n$.

W dalszym ciągu moment centralny rzędu 2 będziemy nazywali *wariancją z próby* i będziemy ją oznaczali przez $\mathbb{S}^2$. Z pewnego względu będziemy też używali modyfikacji wariancji, a mianowicie statystyki

$$\widehat{\mathbb{S}}^2 = \frac{n}{n-1}\mathbb{S}^2.$$

Zauważmy, że:

Fakt 6.3.1 *Niech cecha $\mathbb{X}$ ma wartość oczekiwaną m i wariancję σ^2. Wtedy*

$$E\mathbb{X}_n = m, \ E\widehat{\mathbb{S}}^2 = \sigma^2.$$

Na zakończenie tego podrozdziału wprowadzimy jeszcze dwie statystyki. Statystyki te związane są z dwoma bardzo ważnymi rozkładami w teorii prawdopodobieństwa, o których do tej pory nie wspominaliśmy. Stało się tak dlatego, że złożoność definicji tych rozkładów wykracza poza przewidziany zakres tego opracowania. Z drugiej strony, ponieważ rozkłady te są stablicowane, możemy pozwolić sobie na te uproszczenia bez uszczerbku dla zrozumienia roli, jaką odgrywają.

Zachodzi następujące twierdzenie

Twierdzenie 6.3.1 *Niech cecha $\mathbb{X}$ ma rozkład $\mathcal{N}(m, \sigma^2)$, $(\mathbb{X}_1 \ldots \mathbb{X}_n)$ będzie wektorem losowym złożonym z niezależnych zmiennych losowych o rozkładzie równym $\mathbb{X}$.*

Wtedy:

1.

$$\textit{statystyka} \quad \chi^2_{n-1} = \frac{n\mathbb{S}^2}{\sigma^2},$$

zwana statystyką chi-kwadrat Pearsona ma rozkład chi-kwadrat o n-1 stopniach swobody,

2.

$$\textit{statystyka} \quad t_{n-1} = \frac{\overline{\mathbb{X}}_n}{\mathbb{S}}\sqrt{n-1}$$

nazywana jest statystyką Goseta i ma rozkład t-Studenta o n-1 stopniach swobody,

3.

$$\textit{statystyka} \quad \mathbb{N} = \frac{\overline{\mathbb{X}}_n - m}{\sigma}\sqrt{n}$$

ma rozkład typu $\mathcal{N}(0,1)$ (patrz Twierdzenie 3.4.8).

Uwaga 6.3.2 *Dowolną zmienną losową o rozkładzie chi-kwadrat o k stopniach swobody będziemy oznaczali przez χ^2_k. Podobnie będzie ze zmienną losową o rozkładzie t-Studenta, którą oznaczymy przez t_k.*

Oba powyższe rozkłady są stablicowane. W tablicach rozkładów mamy podane tzw. ich wartości krytyczne:

1. *dla zmiennej χ^2_k*
$$P(\{\omega \in \Omega : \chi^2_k > \chi^2_\alpha\}) = \alpha,$$
2. *dla zmiennej t_k*
$$P(\{\omega \in \Omega : \mid t_k \mid > t_\alpha\}) = \alpha,$$

w obu przypadkach dla $1 \leq k \leq 30$.

Uwaga 6.3.3 *Dla $k > 30$, z CTG wynika, że rozkłady obu statystyk: Pearsona i t-Studenda dobrze przybliża standardowy rozkład normalny.*

Zatem

$$P(\{\omega \in \Omega : \chi^2_k < \chi^2_\alpha\}) \simeq \Phi(\chi^2_\alpha)$$
$$P(\{\omega \in \Omega : t_k < t\}) \simeq \Phi(t), \ t > 0.$$