## 6.4 Podstawowe metody statystyczne

Spóbujemy teraz w dopuszczalnym uproszczeniu przedstawić istotę *analizy statystycznej*. W szczególności udzielimy odpowiedzi na postawione wcześniej pytania (patrz przykład 5.2.1).

Z metodologicznego punktu widzenia analiza statystyczna cechy $\mathbb{X}$ populacji generalnej sprowadza się do:

1. pobrania próby prostej,
2. określenia z dużym prawdopodobieństwem (na ogół 0,95) typu rozkładu cechy $\mathbb{X}$,
3. oszacowania z dużym prawdopodobieństwem (j.w.) parametrów, od których ten rozkład zależy,
4. weryfikacji z dużym prawdopodobieństwem (j.w.) wyników z punktów (2) i (3).

Na temat (1) już wypowiadaliśmy się. Zauważyliśmy, że z jednej strony nie będziemy tej kwestii dyskutowali, z drugiej strony umiejętność próbkowania stanowi podstawę całego rozumowania. Mówiąc inaczej, jeśli w wyniku zakończonej analizy statystycznej pojawią się znaczące różnice pomiędzy uzyskanymi wynikami a dalszą obserwacją cechy populacji generalnej, to na ogół źródłem tych rozbieżności jest niewłaściwe pobranie materiału statystycznego.

W rozdziale 3 wprowadziliśmy podstawowe pojęcie teorii prawdopodobieństwa-rozkład prawdopodobieństwa. Zauważyliśmy, że można wprowadzić kategorię *typu* rozkładu. Dokładniej, możemy mówić o: rozkładach dyskretnych, całkowicie niedyskretnych, w szczególności o rozkładach ciągłych. W ramach tych typów możemy myśleć np. o rozkładzie Bernoulliego, Poissona, wykładniczym, jednostajnym czy normalnym. Istotą metody (2) jest określenie na podstawie pobranej próby prostej takiego właśnie typu rozkładu cechy populacji generalnej. Podstawową metodą jest tutaj metoda bazująca na pojęciu *dystrybuanty empirycznej* i *histogramu* (patrz [4], [10]). Na podstawie próby prostej wykreśla się dystrybuantę schodkową i próbuje się ją zidentyfikować przeglądając katalog znanych rozkładów. Ponieważ ciąg takich dystrybuant empirycznych w określony sposób zbieżny jest do dystrybuanty (w tym przypadku *teoretycznej*) cechy $\mathbb{X}$, więc przy dostatecznie liczebnych próbach prostych metoda ta bywa skuteczna. Tym bardziej, jeśli szukany rozkład cechy pochodzi z tego katalogu.

Niestety, nawet po pomyślnej identyfikacji typu rozkładu nadal pozostaje nierozwiązany problem. Typowe rozkłady zależą bowiem od parametrów, np.

$$\mathbb{X} \in B(n,p),\ \mathbb{X} \in \mathcal{P}(\lambda),\ \mathbb{X} \in \mathcal{J}([a,b]),\ \mathbb{X} \in \mathcal{N}(m,\sigma)\ \ \text{itd.}$$

Pozostaje problem wyznaczenie tych parametrów, o czym mówi metoda (3). Jest to tzw. *teoria estymacji*. Ograniczymy się tylko do metody *estymacji przedziałowej*. O innych rodzajach estymacji można dowiedzieć się np. z [4].

Weźmy próbę prostą cechy $\mathbb{X}$ populacji generalnej $(x_1, x_2, \ldots, x_n)$. Wiemy, że istnieje wtedy przestrzeń probabilistyczna $(\Omega, \Sigma, P)$ oraz ciąg niezależnych zmiennych losowych $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n$ o jednakowym rozkładzie co cecha $\mathbb{X}$, że

$$(x_1, x_2, \ldots, x_n) = \Big(\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n\Big)(\omega_o),$$

dla pewnego zdarzenia elementarnego $\omega_o$.

Definiujemy funkcję:

$$\mathbf{R} \ni x \longrightarrow S(x, \omega_o) \in [0, 1],$$

1. niech $(x_{k_1}, x_{k_2}, \ldots, x_{k_n})$ oznacza niemalejące uporządkowanie próby prostej $(x_1, x_2, \ldots, x_n)$,

2. kładziemy
$$S(x, \omega_o) = \frac{1}{n}|\{j\colon\ x_{k_j} < x\}|,\ x \in \mathbf{R}.$$

Zauważmy, że dla każdego $x \leqslant x_{k_1}$, $S(x, \omega_o) = 0$ oraz dla $x > x_{k_n}$, $S(x, \omega_o) = 1$. Ponadto dla $x$ należących do przedziału $(x_{k_j}, x_{k_{j+1}})$, $S(x, \omega_o) = \frac{j}{n}$, a więc w każdej sytuacji $S(x, \omega_o) \in [0, 1]$. Wprost z definicji wynika, że:

1. $S$ jest funkcją niemalejącą,

2.
$$\lim_{x \to -\infty} S(x, \omega_o) = 0,\ \lim_{x \to +\infty} S(x, \omega_o) = 1.$$

Wreszcie można pokazać, że $S$ jest lewostronnie ciągła. Ponieważ funkcja $S$ wzięła się z materiału statystycznego i ma wszystkie własności dystrybuanty, nazywa się *dystrybuantą empiryczną* cechy populacji generalnej.

Uwolnimy teraz wartość $\omega_o$, a więc potraktujemy ją jako drugi argument dystrybuanty empirycznej $S$. Wtedy dla każdej ustalonej wartości $x = x_o$ dostaniemy odwzorowanie

$$\Omega \ni \omega \longrightarrow S(x_o, \omega) \in \mathbf{R},$$

które jest zmienną losową. Znajdziemy jej rozkład.

Oznaczmy przez $F$ dystrybuantę cechy $\mathbb{X}$. Wtedy dla każdego $1 \leqslant j \leqslant n$

$$P(\{\omega \in \Omega\colon\ \mathbb{X}_j(\omega) < x_o\}) = F(x_o).$$

Z definicji dystrybuanty empirycznej, dla każdego $\omega \in \Omega$, $S(x_o, \omega)$ rejestruje częstość pojawienia się pewnego zdarzenia $A$ w $n$ niezależnych próbach, bowiem

$$S(x_o, \omega) = \frac{1}{n}|\{j\colon\ \mathbb{X}_j(\omega) < x_o\}|$$

i zmienne losowe $\mathbb{X}_j$ są niezależne. Oznacza to, że

$$S(x_o, \omega) \in B(n, p),$$

gdzie $p = F(x_o)$ i dlatego

$$P(\{\omega \in \Omega\colon\ S(x_o, \omega) = \frac{j}{n}\}) = \binom{n}{j}(F(x_o))^j(1 - F(x_o))^{n-j}.$$

Znaczenie dystrybuanty empirycznej dla metod statystycznych wyjaśnia następujące twierdzenie Gliwienki (patrz [4])

**Twierdzenie 6.4.1** *Niech $(S_n)$ oznacza ciąg dystrybuant empirycznych odpowiadający wektorom losowym $(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ niezależnych zmiennych losowych o tym samym rozkładzie $F$ co cecha $\mathbb{X}$ populacji generalnej. Wtedy*

$$\forall_{x \in \mathbf{R}}\ S_n(x, \omega) \xrightarrow{p.w.} F(x),$$

*a więc*

$$P(\{\omega \in \Omega\colon\ \lim_{n \to +\infty} S_n(x, \omega) = F(x)\}) = 1.$$

W takim razie, jeśli dysponujemy odpowiednio liczebną próbą prostą, to z prawdopodobieństwem jeden $S_n(x, \omega) \cong F(x)$. Zatem wielce prawdopodobne jest, że dla zdarzenia elementarnego $\omega_o$ prawdziwe jest przybliżenie

$$S_n(x, \omega_o) \cong F(x).$$

Spróbujemy wyjaśnić to jeszcze raz na przykładzie.

**Przykład 6.4.1** *Zbadamy metodą dystrybuanty empirycznej naturę zjawiska polegającego na wielokrotnym rzucaniu kostką do gry. W omawianym przypadku podstawowe pytanie sprowadza się do rozstrzygnięcia kwestii czy kostka jest symetryczna. Wyobraźmy sobie, że rzucaliśmy kostką 120 razy i odnotowaliśmy następujące wyniki:*

| *liczba oczek* | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *liczba obserwacji* | 23 | 27 | 18 | 22 | 10 | 20 |

*Gdybyśmy mieli pewność, że próba jest dostatecznie liczebna, to z twierdzenia Gliwienki wnioskowalibyśmy, że* $S_{120}(x,\omega_o) \cong F(x)$*, czyli z pewną dokładnością poznalibyśmy rozkład teoretyczny cechy naszej populacji, która opisuje naturę kostki. Z definicji dystrybuanty empirycznej wynika, że*

$$S_{120}(x,\omega_o) = \begin{cases} 0, & x \leqslant 1 \\ \frac{23}{120} \cong 0,19, & 1 < x \leqslant 2 \\ \frac{50}{120} \cong 0,42, & 2 < x \leqslant 3 \\ \frac{68}{120} \cong 0,57, & 3 < x \leqslant 4 \\ \frac{90}{120} \cong 0,75, & 4 < x \leqslant 5 \\ \frac{100}{120} \cong 0,83, & 5 < x \leqslant 6 \\ 1, & 6 < x. \end{cases}$$

*Ponieważ* $\frac{1}{6} \cong 0,17$*, więc postać dystrybuanty empirycznej raczej wyklucza symetrię w obserwowanym zjawisku.*

*Powyższą sytuację możemy zilustrować inaczej. W tym celu podzielmy prostą rzeczywistą następującymi przedziałami:*

$$I_1 = (-\infty, 1),\ I_2 = [1,2),\ I_3 = [2,3),\ I_4 = [3,4),$$

$$I_5 = [4,5),\ I_6 = [5,6),\ I_7 = [6,7),\ I_8 = [7,+\infty).$$

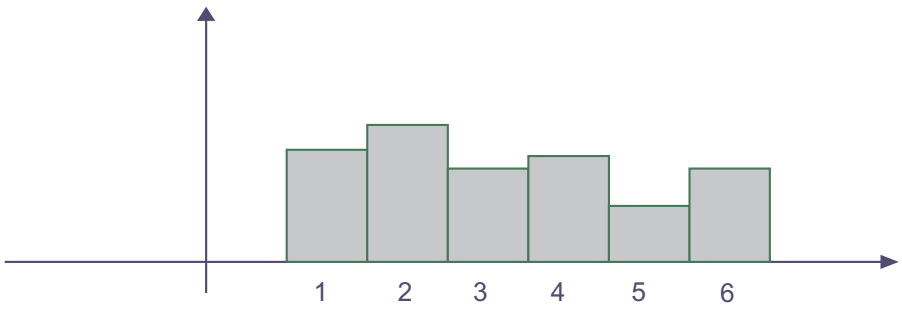
*Definiujemy funkcję* $h\colon \mathbf{R} \to \mathbf{R}$ *wzorem*

$$h(x) = \begin{cases} 0, & x \in I_1 \cup I_8 \\ \frac{w(I_j)}{120 \cdot |I_j|}, x \in I_j,\ 2 \leqslant j \leqslant 7, & \end{cases}$$

*gdzie* $w(I_j)$ *oznacza liczbę z tabeli odpowiadającą wartości* $j$*,* $|I_j|$ *jest długością przedziału* $I_j$ *u nas równą jeden. Podstawiając dane liczbowe otrzymamy*

$$h(x) = \begin{cases} 0, & x \in I_1 \cup I_8 \\ \frac{23}{120} \cong 0,19, & x \in I_2 \\ \frac{27}{120} \cong 0,23, & x \in I_3 \\ \frac{18}{120} \cong 0,15, & x \in I_4 \\ \frac{22}{120} \cong 0,18, & x \in I_5 \\ \frac{10}{120} \cong 0,08, & x \in I_6 \\ \frac{20}{120} \cong 0,17, & x \in I_7. \end{cases}$$

*Poniżej przedstawiliśmy tzw. wykres słupkowy tej funkcji. Zauważmy, że funkcja ta zawsze przyjmuje wartości nieujemne a ponadto suma pól wszystkich słupków jest równa jeden. Tak powstały wykres słupkowy nazywamy histogramem z próby prostej populacji generalnej.*




Rysunek 6.1: wykres histogramu dla próby prostej

I ogólnie, niech

$$(x_1, x_2, \ldots, x_n) = (\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n)(\omega_o)$$

będzie próbą prostą. Podamy zasadę konstrukcji histogramu dla tej próby.

1. Zaznaczamy na osi liczbowj kolejne elementy próby prostej. W ten sposób otrzymamy przedział $[a, b]$, gdzie odpowiednio $a$ jest elementem najmniejszym, $b$ największym w tej próbie.

2. Przedział ten dzielimy na parami rozłączne podprzedziały, których ilość $k$ ustalmy według zasady:

$$k = \sqrt{n} \text{ lub } k \cong 1 + 3,322 \cdot \log n.$$

   Otrzymujemy w ten sposób przedziały

$$[a, a_2), [a_2, a_3), \ldots, [a_{k-1}, b), [b, a_{k+1}).$$

3. Jeśli $n_i$ oznacza liczbę elementów próby prostej w przedziale $[a_i, a_{i+1}]$, $i = 1, 2, \ldots, k$, to przyjmujemy

$$h(x) = \begin{cases} 0, & x < a \text{ lub } x \geqslant a_{k+1} \\ \frac{n_i}{n(a_{i+1} - a_i)}, & x \in [a_i, a_{i+1}]. \end{cases}$$

4. Na tej podstawie konstruujemy wykres słupkowy histogramu. Kolejny $i$–ty słupek w podstawie ma odcinek $[a_i, a_{i+1}]$ oraz wysokość określoną wartością funkcji $h$ na tym przedziale. Wtedy pole takiego słupka jest równe $\frac{n_i}{n}$. Dlatego suma wszystkich pól wynosi jeden. Jeżeli wiemy, że cecha $\mathbb{X}$ obserwowanej populacji generalnej jest typu ciągłego, to wtedy histogram ten stanowi przybliżenie funkcji gęstości tego rozkładu.