

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

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

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

dla pewnego zdarzenia elementarnego ω_o .

Definiujemy funkcję:

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

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

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

Zauważmy, że dla każdego $x \leq x_{k_1}$, $S(x, \omega_o) = 0$ 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.

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

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

Uwolnimy teraz wartość ω_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 dystrybantę cechy \mathbb{X} . Wtedy dla każdego $1 \leq j \leq n$

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

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

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

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

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

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 $(\mathbb{X}_1, \dots, \mathbb{X}_n)$ niezależnych zmiennych losowych o tym samym rozkładzie F co cecha \mathbb{X} populacji generalnej. Wtedy*

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

a więc

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

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 ω_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

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

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:

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

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

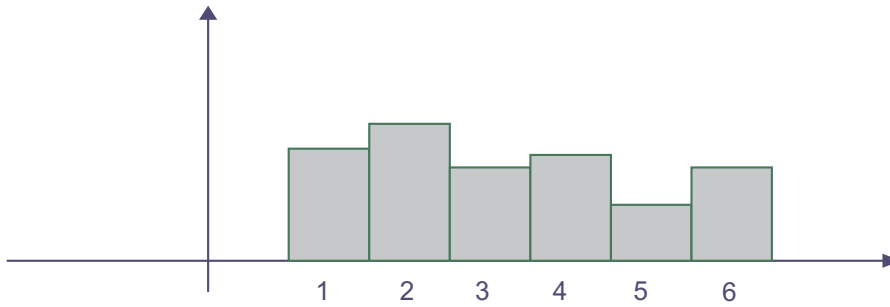
Definiujemy funkcję $h: \mathbf{R} \rightarrow \mathbf{R}$ wzorem

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

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

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

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

1. Zaznaczamy na osi liczbowej 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

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

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

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

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.