IV. ZMIENNE LOSOWE DWUWYMIAROWE

4.1. Rozkład zmiennej losowej dwuwymiarowej

Definicja 4.1. Uporządkowaną parę (X, Y) nazywamy *zmienną losową dwuwymiarową*, jeśli każda ze zmiennych X i Y jest zmienną losową.

Definicja 4.2. Funkcję rzeczywistą $F(x, y)$ zmiennych rzeczywistych x i y określoną na całej płaszczyźnie Oxy jako prawdopodobieństwo, że zmienna losowa X przyjmie wartość mniejszą od x oraz zmienna losowa Y przyjmie wartość mniejszą od y nazywamy *dystrybuantą zmiennej losowej dwuwymiarowej*, czyli

$$F(x, y) = P(X < x, Y < y).$$

Dystrybuanta $F(x, y)$ jest względem każdego z argumentów x i y funkcją:

a) niemalejącą,
b) co najmniej lewostronnie ciągłą,
c) spełniającą warunki graniczne

$$\lim_{\substack{x \to \infty \\ y \to \infty}} F(x, y) = 1, \quad \lim_{x \to -\infty} F(x, y) = 0, \quad \lim_{y \to -\infty} F(x, y) = 0,$$

d) o własności

$$F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0,$$

gdzie $x_1 < x_2$ i $y_1 < y_2$ oznaczają liczby rzeczywiste.

Powyższe warunki są zarazem konieczne i wystarczające na to, by funkcja $F(x, y)$ była dystrybuantą dwuwymiarowej zmiennej losowej (X, Y).

Definicja 4.3. Mówimy, że zmienna losowa dwuwymiarowa (X, Y) jest *typu skokowego*, jeżeli dla każdej pary wskaźników i oraz k jest określona funkcja

$$P(X = x_i, Y = y_k) = p_{ik} > 0$$

spełniająca warunek

$$F(x, y) = \sum_{\substack{x_i < x \\ y_k < y}} p_{ik}$$

dla każdej pary wartości rzeczywistych x i y.

Funkcję tę nazywa się *funkcją prawdopodobieństwa*, punkty (x_i, y_k) – *punktami skokowymi*, a prawdopodobieństwa p_{ik} – *skokami*. Bezpośrednio z podanej definicji wynika, że

$$\sum_{i,k} p_{ik} = 1.$$

Definicja 4.4. Mówimy, że zmienna losowa (X, Y) jest *typu ciągłego*, gdy istnieje taka nieujemna i całkowalna na całej płaszczyźnie Oxy funkcja $f(x, y)$, że

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, ds dt$$

dla każdej pary wartości rzeczywistych x i y.

Funkcję $f(x,y)$ nazywa się *gęstością prawdopodobieństwa* lub krótko *gęstością* zmiennej losowej ciągłej (X, Y). Z definicji tej wynika, że

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1.$$

Mówimy, że jest dany rozkład prawdopodobieństwa zmiennej losowej (X, Y), gdy jest znana dystrybuanta albo też gdy jest znana funkcja prawdopodobieństwa dla zmiennej losowej skokowej lub gęstość prawdopodobieństwa dla zmiennej losowej ciągłej.

Przykład 4.1. Zbadać, czy funkcja

$$f(x, y) = \begin{cases} \frac{1}{8}(x^2 - y^2)e^{-x} & \text{dla } |y| \le x, \\ 0 \text{ dla innych } (x, y) \end{cases}$$

jest gęstością dwuwymiarowej zmiennej losowej (X, Y).
Podana funkcja jest nieujemna, gdyż dla $|y| \le x$ mamy $x^2 - y^2 \ge 0$. Wystarczy zatem sprawdzić, że

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1.$$

Mamy

$$\int_{-\infty}^{\infty} f(x, y)\, dy = \frac{1}{8} \int_{-x}^{x} (x^2 - y^2) e^{-x} dy = \frac{1}{8} \left(x^2 e^{-x} \int_{-x}^{x} dy - e^{-x} \int_{-x}^{x} y^2 dy \right)$$

$$= \frac{1}{8} \left(x^2 e^{-x} y \Big|_{-x}^{x} - e^{-x} \frac{y^3}{3} \Big|_{-x}^{x} \right) = \frac{1}{8} \left(2x^3 e^{-x} - \frac{2}{3} x^3 e^{-x} \right) = \frac{1}{6} x^3 e^{-x}.$$

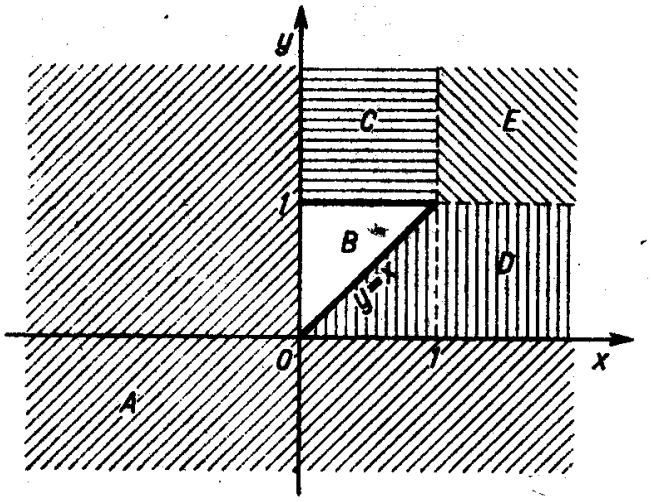
Stąd

$$\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(x,y)\,dy\right]dx = \frac{1}{6}\int_{-\infty}^{\infty} x^3 e^{-x}dx = \frac{1}{6}\cdot\Gamma(4) = \frac{1}{6}\cdot 3! = 1.$$

Przykład 4.2. Zmienna losowa (X, Y) podlega rozkładowi o gęstości

$$f(x,y) = \begin{cases} \dfrac{1}{2\sqrt{xy}} & \text{dla } 0 < x \le y \le 1, \\ 0 & \text{dla innych } (x,y). \end{cases}$$

Znaleźć dystrybuantę zmiennej losowej (X, Y).
Dystrybuantę wyznaczymy kolejno dla obszarów przedstawionych na poniższym rysunku.



W obszarze A mamy $x \le 0$ lub $y \le 0$. Z określenia gęstości $f(x, y)$ wynika, że w tym obszarze mamy $F(x, y) = 0$.
Obszar B jest określony nierównościami $0 < x \le 1$ oraz $x \le y \le 1$. Dla punktów (x, y) położonych w tym obszarze mamy (zob. rys. a))

$$F(x,y) = \int_0^x\left(\int_x^y \frac{1}{2\sqrt{xy}}\,dy\right)dx = \int_0^x\left(\frac{1}{2\sqrt{x}}\int_x^y \frac{1}{\sqrt{y}}\,dy\right)dx = \int_0^x\left(\frac{1}{2\sqrt{x}}\cdot 2\sqrt{y}\,\Big|_x^y\right)dx$$

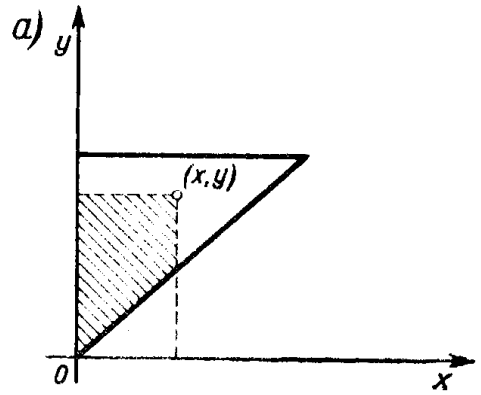
$$= \int_0^x \frac{1}{\sqrt{x}}\left(\sqrt{y}-\sqrt{x}\right)dx = \int_0^x\left(\frac{\sqrt{y}}{\sqrt{x}} - 1\right)dx = \sqrt{y}\int_0^x \frac{1}{\sqrt{x}}\,dx - \int_0^x dx$$

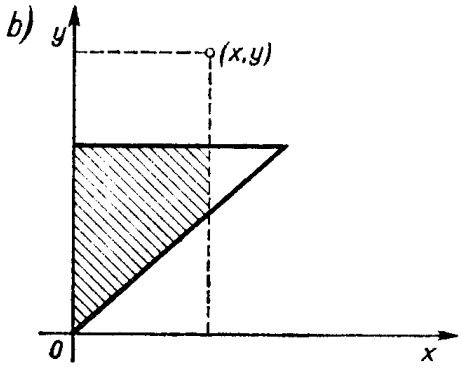
$$= \sqrt{y}\cdot 2\sqrt{x}\,\Big|_0^x - x\,\Big|_0^x = 2\sqrt{xy} - x.$$

W obszarze C są spełnione nierówności $0 < x \le 1$ oraz $1 < y < \infty$. W obszarze tym mamy (zob. rys. b))

$$F(x,y) = \int_0^x\left(\int_x^1 \frac{1}{2\sqrt{xy}}\,dy\right)dx = F(x,1) = 2\sqrt{x} - x,$$

przy czym wykorzystaliśmy tu wynik otrzymany dla obszaru B.



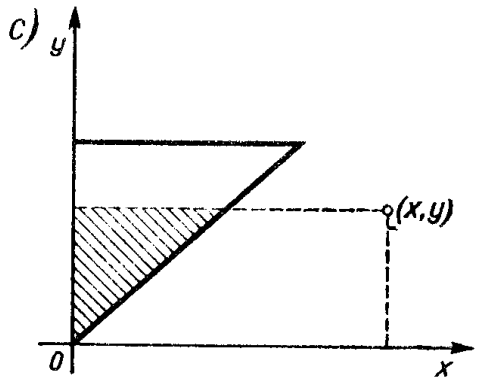


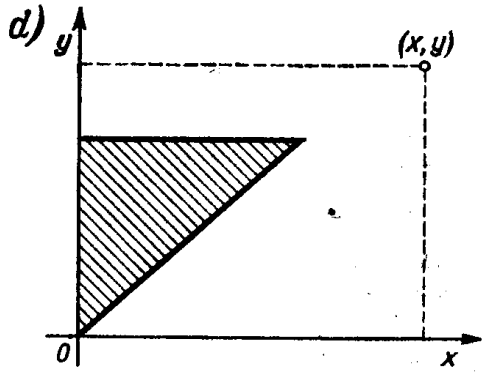
W obszarze D mamy $0 < x < \infty$ oraz $0 < y \leq 1$ i $y < x$, a więc (zob. rys. C)

$$F(x, y) = \int_0^y \left(\int_x^y \frac{1}{2\sqrt{xy}} dy \right) dx = F(y, y) = y,$$

przy czym ponownie wykorzystaliśmy wynik otrzymany dla obszaru B. Wreszcie, dla obszaru E spełnione są nierówności $1 < x < \infty$ oraz $1 < y < \infty$ i otrzymujemy (zob. rys. d))

$$F(x, y) = \int_0^1 \left(\int_x^1 f(x, y) dy \right) dx = F(1, 1) = 1.$$





4.2. Rozkłady brzegowe

Niech prawdopodobieństwo p_{ik} będzie określone wzorem

$$p_{ik} = P(X = x_i, Y = y_k) > 0.$$

Oznaczmy

$$p_{i\cdot} = P(X = x_i, Y = y_1 \text{ lub } X = x_i, Y = y_2 \text{ lub } \ldots)$$

$$= P(X = x_i, Y = y_1) + P(X = x_i, Y = y_2) + \ldots$$
$$= \sum_k P(X = x_i, Y = y_k) = \sum_k p_{ik}$$

i podobnie

$$p_{\cdot k} = \sum_i P(X = x_i, Y = y_k) = \sum_i p_{ik}.$$

Oznacza to, że wartość $p_{i\cdot}$ jest prawdopodobieństwem, że zmienna losowa X przyjmie wartość x_i, gdy zmienna losowa Y przyjmie którąkolwiek z możliwych wartości, a wartość $p_{\cdot k}$ jest prawdopodobieństwem, że zmienna losowa Y przyjmie wartość y_k, gdy zmienna losowa X przyjmie którąkolwiek z możliwych wartości. Mamy przy tym

$$\sum_i p_{i\cdot} = \sum_i \sum_k p_{ik} = 1 \text{ i } \sum_k p_{\cdot k} = \sum_k \sum_i p_{ik} = 1.$$

Definicja 4.5. Rozkład prawdopodobieństwa wyznaczony przez liczby $p_{i\cdot}$ lub $p_{\cdot k}$ nazywamy *rozkładem brzegowym zmiennej losowej skokowej X* lub *Y* w dwuwymiarowym rozkładzie zmiennej losowej (*X*, *Y*), a wyrażenie $p_{i\cdot}$ lub $p_{\cdot k}$ nazywamy *funkcją prawdopodobieństwa* tego rozkładu.

Funkcję prawdopodobieństwa można podać w postaci wzoru lub tzw. tablicy dwuwejściowej (gdy zmienna losowa przyjmuje skończoną liczbę wartości).

Dystrybuantę $F_1(x)$ określa wzór

$$F_1(x) = \sum_{x_i < x} p_{i\cdot} = \sum_{x_i < x} \sum_k p_{ik},$$

co oznacza, że sumowanie dotyczy wszystkich wartości k oraz tych wartości wskaźnika i, dla których jest spełniona nierówność $x_i < x$. Podobnie jest określona dystrybuanta $F_2(y)$:

$$F_2(y) = \sum_{y_k < y} p_{\cdot k} = \sum_{y_k < y} \sum_i p_{ik}.$$

Podobnie, jak dla zmiennej losowej skokowej, wprowadza się pojęcie rozkładów brzegowych dla zmiennej losowej ciągłej. Niech $f(x, y)$ oznacza gęstość dwuwymiarowej zmiennej losowej (*X*, *Y*). Wprowadźmy oznaczenia

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

Funkcje te są całkowalne na całej osi, nieujemne oraz spełniają warunki

$$\int_{-\infty}^{\infty} f_1(x)\, dx = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y)\, dy \right) dx = 1, \quad \int_{-\infty}^{\infty} f_2(y)\, dy = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y)\, dx \right) dy = 1.$$

Definicja 4.6. Rozkład wyznaczony przez funkcje $f_1(x)$ lub $f_2(y)$ nazywamy *rozkładem brzegowym* zmiennej losowej ciągłej *X* lub *Y* w dwuwymiarowym rozkładzie zmiennej

losowej (X, Y), a funkcje $f_1(x)$ lub $f_2(y)$ nazywamy *gęstościami prawdopodobieństwa* tych rozkładów.

Dystrybuanty są określone następującymi wzorami:

$$F_1(x) = \int_{-\infty}^{x} f_1(x)\,dx = \int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f(x,y)\,dy\right)dx,$$
$$F_2(y) = \int_{-\infty}^{y} f_2(y)\,dy = \int_{-\infty}^{y}\left(\int_{-\infty}^{\infty} f(x,y)\,dx\right)dy.$$

Przykład 4.3. Dana jest funkcja

$$f(x,y) = \begin{cases} Cxy & \text{dla } 1 \le x \le 2,\ 2 \le y \le 4, \\ 0 & \text{dla innych } x \text{ i } y. \end{cases}$$

A. Wyznaczyć stałą C tak, aby podana funkcja określała rozkład.
B. Podać rozkłady brzegowe.
C. Określić dystrybuantę.
Ad A.
Stałą wyznaczamy z warunku

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$$

Z uwagi na definicję funkcji $f(x, y)$ mamy

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_{2}^{4}\int_{1}^{2} Cxy\,dx\,dy = C\int_{2}^{4}\left(y\int_{1}^{2} x\,dx\right)dy$$
$$= C\int_{2}^{4}\left[y\left(\frac{x^2}{2}\right)\Big|_{1}^{2}\right]dy = C\int_{2}^{4} y\left(2-\frac{1}{2}\right)dy$$
$$= \frac{3}{2}C\int_{2}^{4} y\,dy = \frac{3}{2}C\left(\frac{y^2}{2}\right)\Big|_{2}^{4} = \frac{3}{2}C(8-2) = 9C.$$

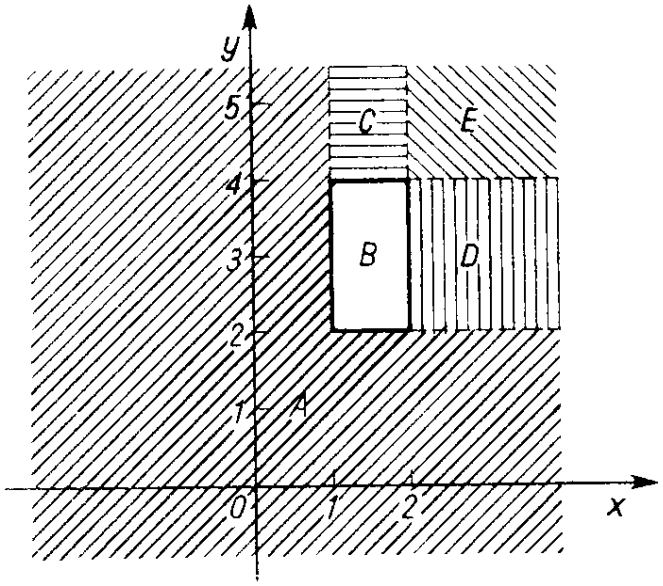
Zatem $9C = 1$, skąd $C = 1/9$.
Ad B.
Mamy

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \frac{1}{9}x\int_{2}^{4} y\,dy = \frac{1}{9}x\left(\frac{y^2}{2}\right)\Big|_{2}^{4} = \frac{1}{9}x(8-2) = \frac{2}{3}x \text{ dla } 1 \le x \le 2,$$
$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \frac{1}{9}y\int_{1}^{2} x\,dx = \frac{1}{9}y\left(\frac{x^2}{2}\right)\Big|_{1}^{2} = \frac{1}{9}y\left(2-\frac{1}{2}\right) = \frac{1}{6}y \text{ dla } 2 \le y \le 4.$$

Możemy zatem napisać

$$f_1(x)=\begin{cases}\frac{2}{3}x & \text{dla } 1\le x\le 2,\\ 0 & \text{dla innych } x,\end{cases}\qquad f_2(y)=\begin{cases}\frac{1}{6}y & \text{dla } 2\le y\le 4,\\ 0 & \text{dla innych } y.\end{cases}$$

Ad C.
Dystrybuantę wyznaczymy kolejno dla obszarów pokazanych na poniższym rysunku.



Dla obszaru A mamy $x \le 1$ lub $y \le 2$. Ponieważ funkcja gęstości $f(x, y)$ jest w tym obszarze równa 0, więc

$$F(x, y) = 0.$$

W obszarze B spełnione są nierówności $1 \le x \le 2$ oraz $2 \le y \le 4$ i mamy

$$F(x,y)=\frac{1}{9}\int_1^x\int_2^y xy\,dy\,dx=\frac{1}{9}\int_1^x x\left(\int_2^y y\,dy\right)dx=\frac{1}{9}\int_1^x x\left(\frac{y^2}{2}\right)\Bigg|_2^x dx$$
$$=\frac{1}{9}\int_1^x x\left(\frac{y^2}{2}-2\right)dx=\frac{y^2-4}{18}\int_1^x x\,dx=\frac{1}{18}(y^2-4)\left(\frac{x^2}{2}\right)\Bigg|_1^x$$
$$=\frac{1}{18}(y^2-4)\left(\frac{x^2}{2}-\frac{1}{2}\right)=\frac{1}{36}(y^2-4)(x^2-1)=\frac{1}{36}(x^2y^2-4x^2-y^2+4).$$

Dla obszaru C mamy $1 \le x \le 2$ oraz $4 \le y < \infty$, więc

$$F(x,y)=\frac{1}{9}\int_1^x\int_2^4 xy\,dy\,dx=F(x,4)=\frac{1}{36}(16x^2-4x^2-16+4)=\frac{1}{3}(x^2-1)=F_1(x),$$

przy czym skorzystaliśmy tu z wyniku uzyskanego dla obszaru B.

W obszarze D zachodzą nierówności $2 < x < \infty$ oraz $2 \le y \le 4$ i ponownie korzystając z wyniku dla obszaru B otrzymujemy

$$F(x,y) = \frac{1}{9}\int_1^2\int_2^y xy\,dy\,dx = F(2,y) = \frac{1}{36}(4y^2 - 16 - y^2 + 4) = \frac{1}{12}(y^2 - 4) = F_2(y).$$

Wreszcie, dla obszary E, w którym zachodzą nierówności $2 < x < \infty$ i $4 < y < \infty$, mamy

$$F(x,y) = \frac{1}{9}\int_1^2\int_2^4 xy\,dy\,dx = F(2,4) = \frac{1}{36}(64 - 16 - 16 + 4) = 1.$$

Zadania

1. Podać gęstości rozkładów brzegowych zmiennych X i Y w dwuwymiarowym rozkładzie zmiennej losowej (X, Y) o gęstości danej wzorem

$$f(x,y) = \begin{cases} \frac{1}{8}(x^2 - y^2)\mathrm{e}^{-x} & \text{dla } |y| \le x, \\ 0 \text{ dla innych } (x,y). \end{cases}$$

2. Podać rozkłady brzegowe zmiennych X i Y rozkładu podanego w przykładzie 4.2.

3. Znaleźć gęstości i dystrybuanty brzegowe zmiennych X i Y rozkładu w dwuwymiarowym rozkładzie zmiennej losowej (X, Y) o gęstości

$$f(x,y) = \begin{cases} \mathrm{e}^{-y} & \text{dla } 0 \le x \le \infty,\ x \le y < \infty, \\ 0 \text{ dla innych } (x,y). \end{cases}$$

4. Wyznaczyć dystrybuantę $F(x,y)$, gęstości brzegowe $f_1(x), f_2(y)$, dystrybuanty brzegowe $F_1(x)$, $F_2(y)$ rozkładu prawdopodobieństwa dwuwymiarowej zmiennej losowej (X, Y), jeśli gęstość jest dana wzorem

$$f(x,y) = \begin{cases} 1 & \text{dla } 0 \le x \le 1,\ x \le y \le 2 - x, \\ 0 & \text{dla innych } (x,y). \end{cases}$$

V. ELEMENTY STATYSTYKI MATEMATYCZNEJ

5.1. Przedmiot i zadania statystyki matematycznej

Statystyka jest nauką zajmującą się liczbowym opisywaniem masowych zjawisk i procesów. Najpowszechniej znana statystyka zajmuje się rejestrowaniem danych potrzebnych ludziom zajmującym się polityką i gospodarką. Dane te dotyczą zwykle ludności, jej podziału na różne zawody, wielkości produkcji w poszczególnych działach gospodarki narodowej, wielkości obrotów handlowych, wielkości spożycia różnych artykułów itp.

Badany zbiór nazywa się *populacją* (niekoniecznie musi to być zbiór ludzi, np. populacja przedsiębiorstw, populacja gospodarstw rolnych). Dla danej populacji wyróżnia się pewną *cechę* (lub cechy) i każdemu elementowi populacji przyporządkowuje się pewną *wartość cechy*. Zbiór par

(wartość cechy, liczba jednostek w populacji o danej wartości cechy)

nazywa się *rozkładem cechy w populacji*.

Często pojawia się konieczność uzyskania danych w inny sposób niż za pomocą spisywania wszystkich jednostek populacji. Ten inny sposób to tzw. *badanie reprezentacyjne* lub badanie wyrywkowe. Polega ono na tym, że z populacji wybiera się tylko pewną liczbę jednostek i każdej z wybranych jednostek, w wyniku odpowiedniego badania, przyporządkowuje się wartość badanej cechy. Zbiór badanych jednostek nazywa się *próbką*.

Powstaje problem uogólnienia wyników badania wyrywkowego na całą populację. Pojawiają się natychmiast następujące pytania:

- czy takie uogólnienie jest możliwe,
- jeżeli tak, to jak je przeprowadzić,
- jak jest wiarygodność tak uogólnionych danych?

Zagadnienie opisania populacji przez podanie rozkładu interesującej nas cechy sprowadza się do zadania podania rozkładu pewnej zmiennej losowej. Często jest wystarczające oszacowanie tylko pewnych syntetycznych wskaźników dotyczących tej cechy. Takie zadania sprowadzają się do szacowanie niektórych parametrów zmiennych losowych lub prawdopodobieństw pewnych zdarzeń losowych.

Zadania szacowania rozkładu zmiennych losowych, parametrów tych rozkładów (np. wartości oczekiwanej) lub prawdopodobieństw różnych zdarzeń nazywają się *zadaniami estymacji*.

Często w praktyce zdarza się, że rozważając pewne zagadnienie mamy gotową hipotezę dotyczącą tego zagadnienia i zadanie polega na tym, żeby hipotezę tę sprawdzić. *Hipotezą statystyczną* nazywa się każde przypuszczenie o rozkładzie zmiennej losowej, jego parametrach lub prawdopodobieństwach pewnych zdarzeń. Część statystyki, która zajmuje się metodami sprawdzania takich hipotez nazywa się *teorią weryfikacji hipotez statystycznych* lub teorią testów statystycznych.

Przykład 5.1. Wyjmijmy z portmonetki dowolną monetę. Sprawdzenie, czy jest ona symetryczna sprowadza się do zweryfikowania hipotezy, że zmienna losowa przyjmująca wartość 1, gdy wynikiem rzutu monetą jest orzeł i wartość 0 w przeciwnym przypadku, ma wartość oczekiwaną równą 1/2. Hipoteza ta może być równoważnie sformułowana następująco: zdarzenie polegające na wyrzuceniu orła jest równe 1/2.

5.2. Pojęcie próbki

Rozważmy przykład 5.1. W celu sprawdzenia, czy moneta jest symetryczna, wykonujemy serię niezależnych rzutów tą monetą. Niech X oznacza zmienną losową przyjmują wartość 1, gdy wynikiem rzutu jest orzeł i wartość 0 w przeciwnym przypadku. Niech X_i oznacza zmienną losową – wynik i-tego rzutu. Liczba wszystkich rzutów niech wynosi n. Każda ze zmiennych losowych X_i ma jednakowy rozkład, taki jak zmienna losowa X i zmienne losowe X_i są niezależne. Zadanie polega na sformułowaniu pewnych wniosków o rozkładzie zmiennej losowej X na podstawie wyników $X_1, X_2, \ldots, X_n$ przeprowadzonych rzutów. Zmienne losowe $X_1, X_2, \ldots, X_n$ stanowią próbkę w naszym zagadnieniu.

Definicja 5.1. Niech X oznacza zmienną losową. *Próbką n-elementową* nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$ o jednakowym rozkładzie, takim jak rozkład zmiennej losowej X.

Wnioskowanie statystyczne polega na formułowaniu różnych twierdzeń o rozkładzie zmiennej losowej X na podstawie próbki $X_1, X_2, \ldots, X_n$.

5.3. Zagadnienia estymacji

Rozpatrzmy problem ogólny. Niech X oznacza zmienną losową o nieznanej dystrybuancie $F(X)$, a $X_1, X_2, \ldots, X_n$ – próbkę dla zmiennej losowej X. Przypuśćmy, że zadanie polega na oszacowaniu pewnego parametru liczbowego ϑ w tym rozkładzie (np. wartości oczekiwanej, wariancji, prawdopodobieństwa pewnego ustalonego zdarzenia, np. $P(X > a)$, gdzie a oznacza pewną stałą).

Za oszacowanie parametru ϑ przyjmiemy zaobserwowaną wartość pewnej funkcji określonej na zbiorze wszystkich wartości z próbki, tzn. funkcji, której argumentami są zmienne losowe X_1, $X_2, \ldots, X_n$. Oznaczmy tę funkcję następująco:

$$\hat{\vartheta}_n = \hat{\vartheta}_n(X_1, X_2, \ldots, X_n).$$

Funkcję $\hat{\vartheta}_n$ nazywamy *estymatorem parametru* ϑ.

Określając funkcję $\hat{\vartheta}_n$ dla każdej wartości n, otrzymujemy ciąg funkcji $\hat{\vartheta}_n$ ($n = 1, 2, \ldots, n$). Aby estymator $\hat{\vartheta}_n$ mógł być praktycznie wykorzystany, muszą być spełnione pewne dodatkowe warunki, które powinny zagwarantować, że obserwowane w próbce wartości $\hat{\vartheta}_n$ nie będą bardzo odbiegały od prawdziwych wartości parametru ϑ. Ograniczymy się do dwóch warunków.

Definicja 5.2. Jeżeli dla każdej dodatniej liczby ε

$$\lim_{n\to\infty} P\left(\left|\hat{\vartheta}_n - \vartheta\right| > \varepsilon\right) = 0,$$

to estymator $\hat{\vartheta}_n$ nazywamy *estymatorem zgodnym*.

Definicja 5.3. Jeśli

$$E\left(\hat{\vartheta}_n\right) = \vartheta,$$

to estymator $\hat{\vartheta}_n$ nazywamy *estymatorem nieobciążonym*.

Przykład 5.2. Niech dana będzie zmienna losowa X i niech zadanie polega na oszacowaniu prawdopodobieństwa $p = P(X > a)$, gdzie a oznacza pewną stałą. Niech X_1, X_2, ... , X_n oznacza próbkę i zdefiniujmy nowe zmienne losowe

$$Y_i = \begin{cases} 1, \text{ gdy } X_i > a, \\ 0, \text{ gdy } X_i \le a. \end{cases}$$

Określmy następnie estymator $\hat{p}_n$ parametru p:

$$\hat{p}_n = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n).$$

Pokazać, że estymator $\hat{p}_n$ jest zgodnym i nieobciążonym estymatorem parametru p.

Z określenia zmiennych losowych Y_i wynika, że zdarzenie, iż zmienna losowa Y_i jest równa 1 jest równoważne zdarzeniu, że zmienna losowa X_i przyjmie wartość większą od a. Ponieważ zmienne losowe X_i są niezależne, więc zdarzenia $X_i > a$ są też niezależne. Suma $Y_1 + Y_2 + ... + Y_n$ może być więc interpretowana jako liczba zajść pewnego zdarzenia w schemacie Bernoulliego. W takim przypadku jest spełnione tzw. *prawo wielkich liczb Bernolliego*:

Twierdzenie 5.1. *Jeżeli $\{X_i\}$ oznacza ciąg zmiennych losowych o wspólnym rozkładzie zero-jedynkowym, tj.*

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p$$

oraz

$$Y_n = \sum_{i=1}^{n} X_i,$$

to ciąg

$$\left\{U_n = \frac{Y_n}{n}\right\}$$

jest stochastycznie zbieżny do p.

Dowód pomijamy. ■

Innymi słowy: ponieważ zmienna Y_n jest zmienną losową o rozkładzie dwumianowym, która przyjmuje wartości k ($k = 0, 1, ... , n$), a ułamek k/n oznacza częs-

tość względną pojawienia się określonego zdarzenia w n doświadczeniach przeprowadzonych według schematu Bernoulliego, to możemy powiedzieć, że

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| \geq \varepsilon\right) = 0.$$

Stąd wynika, że estymator $\hat{p}_n$ jest estymatorem zgodnym.
Ponadto mamy

$$E(Y_i) = 1 \cdot P(X_i > a) + 0 \cdot P(X_i \leq a) = p.$$

Ale wartość oczekiwana sumy zmiennych losowych jest równa sumie wartości oczekiwanych tych zmiennych, więc

$$E(Y_1 + Y_2 + \ldots + Y_n) = \sum_{i=1}^{n} E(Y_i) = np.$$

Korzystając z twierdzenia 3.7 mamy

$$E\left(\hat{p}_n\right) = E\left(\frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n)\right) = \frac{1}{n} E(Y_1 + Y_2 + \ldots + Y_n) = p,$$

a to oznacza, że estymator $\hat{p}_n$ jest estymatorem nieobciążonym parametru p.

Wynik uzyskany dla parametru p z powyższego przykładu może być uogólniony.

Twierdzenie 5.2. *Średnia z próbki jest zgodnym i nieobciążonym estymatorem wartości oczekiwanej, jeśli tylko ta wartość oczekiwana istnieje.*

Dowód. Oznaczmy średnią z próbki przez $\overline{X}_n$. Mamy

$$E(\overline{X}_n) = E\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) = \frac{1}{n} E(X_1 + X_2 + \ldots + X_n)$$
$$= \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)) = \frac{1}{n} \cdot n \cdot E(X) = E(X),$$

gdzie X oznacza zmienną losową o takim samym rozkładzie, jak $X_1, X_2, \ldots, X_n$, czyli jak próbka. Dowód zgodności estymatora $\overline{X}_n$ wynika z tzw. *prawa wielkich liczb Chinczyna*, którego dowód pomijamy:

Twierdzenie 5.3. *Niech $X_1, X_2, \ldots, X_n$ oznacza ciąg niezależnych zmiennych losowych o jednakowym rozkładzie. Jeżeli wartość oczekiwana $m = E(X)$ istnieje, to dla każdej liczby $\varepsilon > 0$*

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - m\right| \geq \varepsilon\right) = 0. \qquad \blacksquare$$