Statystyka opisowa

Statystyka zajmuje się zasadami i metodami uogólniania wyników otrzymanych z próby losowej na całą populację (czyli zbiorowość, z której została pobrana próba). Takie postępowanie nazywamy **wnioskowaniem statystycznym**.

Zbiór wartości dla interesującej badacza cechy (lub cech) u wszystkich jednostek populacji fizycznej tworzy tzw. **populację generalną**.

Jeżeli zbiór elementów populacji generalnej jest skończony, to będziemy ją określać jako populację skończoną. Przykładem może być np. zbiór drzew wiśni w pewnym sadzie, zbiór wyprodukowanych produktów danego dnia.

W przypadku, gdy zbiór elementów populacji jest nieskończony, to populację określamy jako **nieskończoną.** Przykładem niech będzie zbiór ocen ogólnych

(konsumenckich) pewnego produktu (np. kompotu z wiśni).

W populacji mogą nas interesować cechy ilościowe, które będziemy nazywać **mierzalnymi** jak i cechy jakościowe, czyli **niemierzalne**.

Formalnie, populację generalną będziemy traktować jako zbiór niezależnych realizacji pewnej zmiennej losowej jedno lub wielowymiarowej (wiele cech badanych jednocześnie).

Celem badania statystycznego może być poznanie rozkładu danej cechy jak i oszacowanie charakterystyk tego rozkładu.

Jeżeli zmienna losowa X jest modelem probabilistycznym dla pewnej cechy w populacji generalnej (np. rozkład N(172,200) dla wzrostu dorosłego Polaka), to **rozkład prawdopodobieństwa** zmiennej modelowej opisuje **rozkład częstości** występowania różnych wartości tej cechy

(prawdopodobieństwo $P(160<X<190)$ opisuje prawdopodobieństwo wzrostu między 160 a 190cm), a parametry rozkładu tej zmiennej ($m=172$, $\sigma=200$) są jednocześnie **parametrami populacji**.

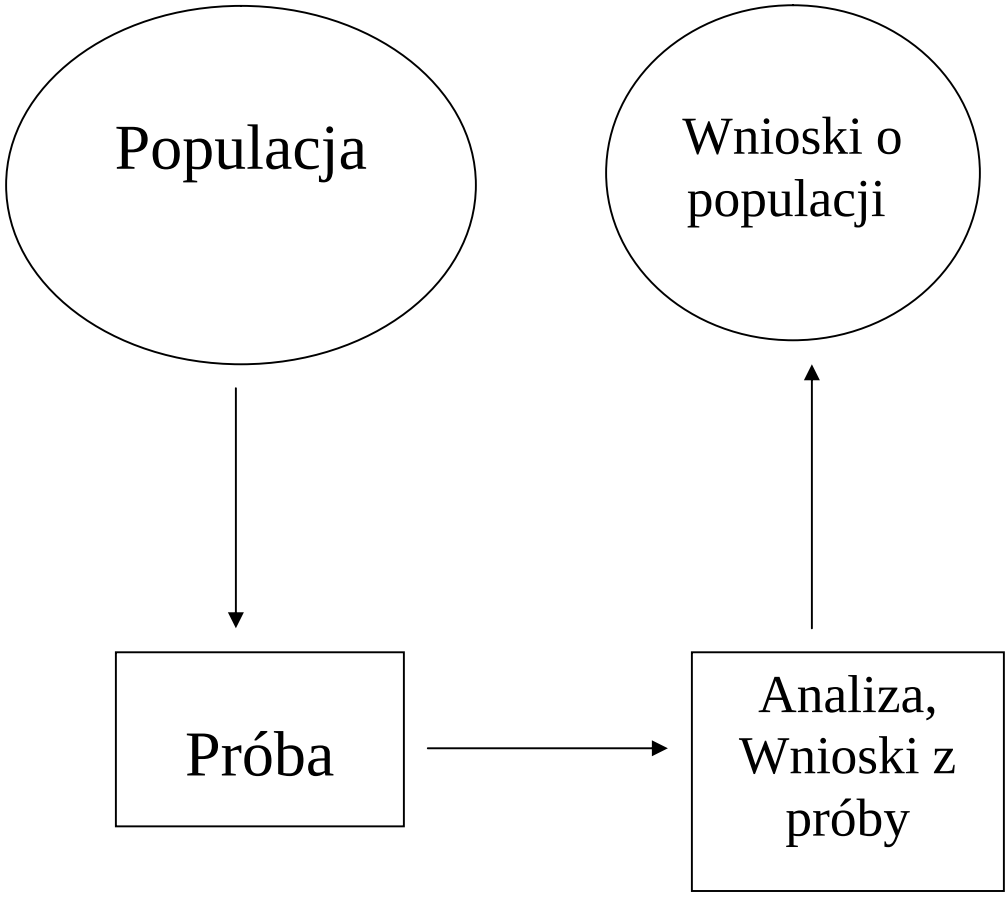
Badanie statystyczne może być badaniem

- **pełnym -** jeżeli obejmuje wszystkie elementy populacji generalnej;
- **częściowym -** jeżeli ograniczone jest do pewnej części populacji generalnej.

Tę część populacji generalnej, na której wykonywane jest badanie statystyczne nazywamy **populacją próbną** lub **próbą**.

Statystyka matematyczna zajmuje się tylko badaniami częściowymi, przy czym muszą być jeszcze spełnione określone warunki doboru próby.

Podstawowym warunkiem, jaki musi być spełniony w badaniach częściowych jest losowy dobór próby. Tak otrzymaną próbę nazywamy **próbą losową**.



Próba prosta (prosta próba losowa)
Jeżeli elementy próby zostały pobrane w taki sposób, aby:

- każdy element populacji generalnej miał tę samą szansę znalezienia się w próbie,
- losowanie elementów próby było niezależne,

- próba była dostatecznie liczna,

to możemy oczekiwać, że prawidłowości występujące w populacji znajdą swoje odbicie w próbie (jeśli tak jest, to taką próbę nazywamy próbą **reprezentatywną**).

Jak wcześniej powiedzieliśmy, próba ma dostarczyć informacji o analizowanej zmiennej w populacji, między innymi na podstawie elementów próby będziemy szacować (oceniać, **estymować**) nieznane parametry populacji.

Przykład:
Fabryka produkuje **kubeczki jogurtu** (ok. 200g). **Populacją** są produkowane kubeczki jogurtu w danym roku, a **cechą** może być masa netto. Aby próba była **próbą prostą**, należy tak ustalić **sposób losowania**, aby:

§ każdy wyprodukowany kubeczek miał jednakową szansę należeć do próby (te produkowane w styczniu i te w grudniu, produkowane rano i po południu,...),

- pobranie do próby było niezależne (pobieramy kubeczek do próby niezależnie od tego, czy dzisiaj pobraliśmy już 1,5 czy 0 kubeczków)
- pobierzemy wystarczającą ich ilość (np 50 przy określeniu masy, albo 200 dla określenia odchylenia standardowego)

Przykłady innych metod doboru próby niż próba prosta

Próba celowa (dobór ekspercki) – dobór nieprobabilistyczny. Staramy się wybrać jednostki tak, aby próba była reprezentatywna. Stosowanie – gdy mamy dość dobre informacje o obiektach a możliwe jest przebadanie tylko małej ich liczby.

Próba warstwowa – obiekty dzielone są na grupy. Z każdej grupy osobno obiekty losowane są jak w próbie prostej. Liczebność prób pobranych z grup nie musi być proporcjonalna do proporcji tych grup (np. badamy 100 osób palących i 100

niepalących). Większa efektywność badań niż próba prosta. Stosowana np. przez GUS.

Niech cecha (zmienna losowa) X ma pewien rozkład normalny X~N(m, σ^2), oraz niech x_i ($i = 1, 2, ..., n$) oznacza n-elementową próbę losową z populacji nieskończonej.

Ocenami niebciążonymi średniej i wariancji w populacji generalnej są odpowiednio:

- **średnia z próby**:

$$\hat{m} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **wariancja** i **odchylenie standardowe:**

$$\hat{s}^2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)$$

$$\hat{s} = s = \sqrt{s^2}$$

Te wartości wyznaczane na podstawie próby są najbardziej prawdopodobnymi wartościami odpowiednich parametrów populacji.

Średnia arytmetyczna, stosowana bardzo często, jest wskazana wtedy, gdy sumowanie wyników ma sens. Jeśli nie, można posłużyć się średnią geometryczną (np. gdy notujemy kolejne zmiany procentowe pewnej wielkości) lub harmoniczną (poszukujemy średniego współczynnika/ilorazu, np. gdy uśredniamy prędkość przejazdu danego odcinka).

Kolejny wskaźnik z grupy parametrów zmienności to **współczynnik zmienności v:**

$$v = \frac{s}{\bar{x}} \cdot 100\%$$

Wyraża zmienność cechy w stosunku do jej wartości średniej. Jest jednostką niemianowaną, więc może służyć do porównania zmienności cech wyrażanych w

różnych jednostkach (np. zmienność masy w kg i zmienność wysokości w cm).

Mediana w próbie (**m_e**) jest wartością środkową lub średnią dwu wartości środkowych spośród uporządkowanych obserwacji. (tyle samo obserwacji powyżej co poniżej mediany)

Powyższe charakterystyki to podstawowe parametry próby wykorzystywane w praktyce.

Wartość średnia i mediana są miarami położenia, a wariancja, odchylenie standardowe i współczynnik zmienności miarami rozproszenia badanej właściwości.

Dla konkretnych prób tej samej (badanej) populacji określone parametry przyjmują nieco różne wartości.

Estymatorom nieznanych parametrów populacji stawiamy wiele wymogów, które

mają zapewnić dobre oszacowanie nieznanych charakterystyk.

Między innymi oczekujemy by estymator był **nieobciążony**, czyli przeciętnie trafnie oceniał nieznany parametr populacyjny (nie może średnio zawyżać. nie może średnio zaniżać).

Oczekujemy też by estymator był **efektywny**, to znaczy charakteryzował się małą wariancją (dość dokładnie przybliżał wartość parametru populacyjnego).

Średnia z próby $\bar{x}$ jest najefektywniejszym nieobciążonym estymatorem wartości oczekiwanej ***m*** rozkładu Normalnego.

Przykład

Badano plon pewnej odmiany pszenicy-X, pobrano małą **próbę prostą**, n=5,

$$x_i: 35, 37, 40, 38, 40$$

Wartość średnia jest równa

$$\hat{m} = \bar{x} = \frac{35+37+40+38+40}{5} = 38$$

Mediana ma wartość 38.

Wariancja, odchylenie standardowe i współczynnik zmienności w próbie, liczone według podanych formuł, są równe:

$$\hat{s}^2 = s^2 = \frac{18}{5-1} = 4{,}5$$

$$\hat{s} = s = \sqrt{4{,}5} = 2{,}121$$

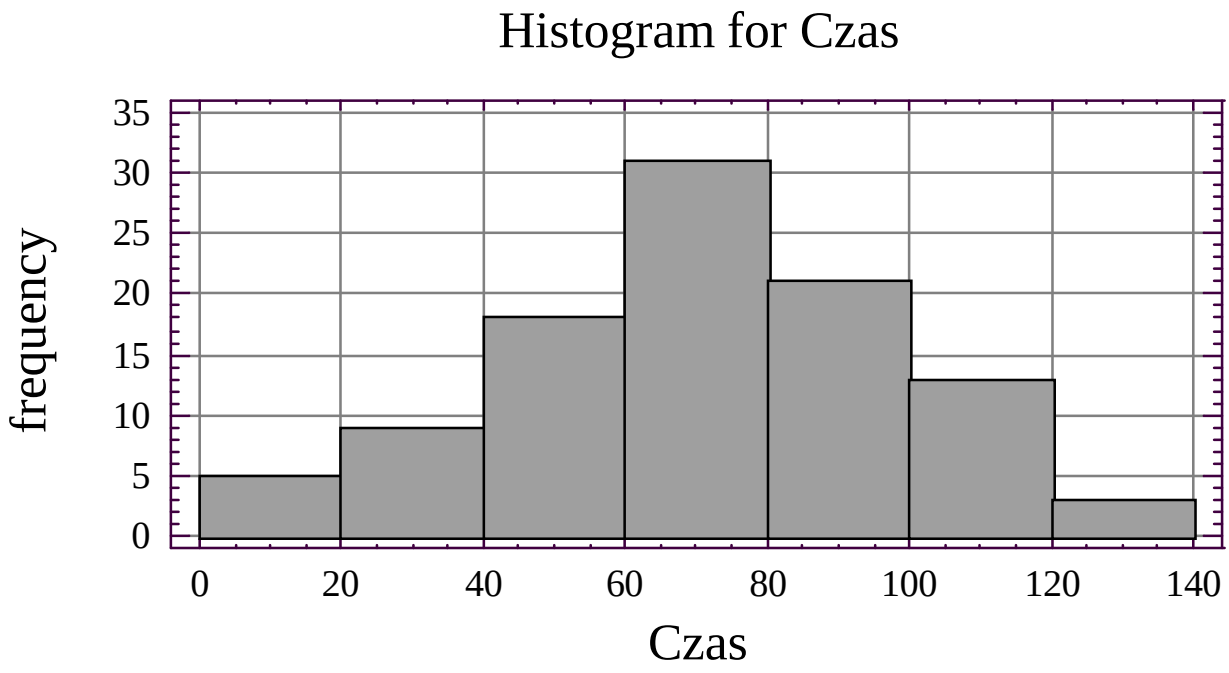
$$V = 5{,}582\ \%$$

Dla licznej próby konstruujemy **szereg rozdzielczy** – zestawienie wskazujące na rozkład wartości w próbie (być może w populacji), na przykład:

Przykład: Badając czas obsługi przy kasie sklepowej 100 losowo wybranych klientów uzyskano następujące wyniki (w sekundach):

X – czas obsługi przy kasie	**n_i**
X<20	5
20<X<40	9
40<X<60	18
60<X<80	31
80<X<100	21
100<X<120	13
X>120	3

Wyniki można przedstawić graficznie w postaci histogramu.



Parametry próby dużej, sklasyfikowanej w powyższy sposób, można policzyć jak dla próby prostej lub według formuł dla szeregu rozdzielczego.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{1}{100} 7100 = 71$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 n_i}{N-1} = \frac{1}{N-1}\left(\sum_{i=1}^{n} x_i^2 n_i - \frac{(\sum_{i=1}^{n} x_i n_i)^2}{N} \right)$$

$$= \frac{1}{99}\left(583600 - \frac{7010^2}{100} \right) \approx 803 \Rightarrow S \approx 28{,}3$$

Tabelkę z szeregiem rozdzielczym rozszerzono o odpowiednie kolumny

X	$\mathbf{n_i}$	$\mathbf{x_i}$	$\mathbf{n_i * x_i}$	$\mathbf{x_i^2}$	$\mathbf{n_i * x_i^2}$
X<20	5	10	50	100	500
20<X<40	9	30	270	900	8100
40<X<60	18	50	900	2500	45000
60<X<80	31	70	2170	4900	151900
80<X<100	21	90	1890	8100	170100
100<X<120	13	110	1430	12100	157300
X>120	3	130	390	16900	50700
suma	**100**		**7100**		**583600**

Konstrukcja szeregu rozdzielczego.

- Granice przedziałów w szeregu rozdzielczym powinny być “okrągłe”
- liczba przedziałów powinna być w okolicy 5log(N) lub 1+3,322log(N)

W omawianym przypadku granice przedziałów były równe (20,40, ..., 120) i liczba przedziałów (7) była w okolicy wartości 10 lub 7,6

Dla szeregu rozdzielczego można określić również funkcję dystrybuanty empirycznej.