Opis statystyczny

Punktem wyjścia do wnioskowania statystycznego (uogólnianie wyników badania próby na populację generalną) jest odpowiednia analiza rozkładu badanej cechy w tej próbie. Metody służące do analizy rozkładu cechy w próbie są nazywane metodami *opisu statystycznego*.

Opis statystyczny sprowadza się do wyznaczenia pewnych liczbowych parametrów charakteryzujących badany rozkład. Opis statystyczny może być zamkniętym badaniem (w przypadku skończonej zbiorowości generalnej).

Stosowane w analizach parametry:

- **Miary położenia (przeciętne, średnie)**
- **Miary zmienności (zróżnicowania, dyspersji, rozproszenia)**
- **Miary asymetrii (skośności)**
- **Miary skupienia**

Miary położenia :

- ***klasyczne***
  - średnia arytmetyczna
  - średnia geometryczna
  - średnia harmoniczna
- ***pozycyjne***
  - dominanta
  - mediana
  - kwantyle

Oznaczenia:

$\overline{x}_A$ -średnia arytmetyczna

$\overline{x}_G$ -średnia geometryczna

$\overline{x}_H$ -średnia harmoniczna

$\hat{x}_i$ -środek i-tego przedziału klasowego

n_i -liczebność i-tego wariantu cechy

N -liczebność badanej zbiorowości

r -liczba wariantów cechy

Średnia arytmetyczna

1… $\bar{x}_A = \frac{1}{N}\sum_{i=1}^{N} x_i$ szereg szczegółowy

1a… $\bar{x}_A = \frac{1}{N}\sum_{i=1}^{r} x_i n_i = \sum_{i=1}^{r} x_i \omega_i$ szereg rozdzielczy punktowy

1b… $\bar{x}_A = \frac{1}{N}\sum_{i=1}^{r} \hat{x}_i n_i = \sum_{i=1}^{r} \hat{x}_i \omega_i$ szereg rozdzielczy przedziałowy o domkniętych przedziałach klasowych

gdzie $\sum_{i=1}^{r} n_i = N$

*Uwaga: D*la szeregów przedziałowych wyznacza się tzw. średnią ważoną (wagami są częstości ω_i)

UWAGi:

1. Środki przedziałów uznajemy za reprezentatywne, ale one tylko w przybliżeniu odzwierciedlają rzeczywiste wartości; stąd dla szeregów rozdzielczych przedziałowych wartości: średniej arytmetycznej wyznaczonej wg wzoru (1b) i średniej arytmetycznej wyznaczonej dla szeregu szczegółowego wg wzoru (1) na ogół będą się różnić.

2. Średnia arytmetyczna jest pewną abstrakcyjną wielkością (wypadkową wszystkich obserwacji) i nie musi należeć do zbioru wartości cechy.

3. Dla szeregów rozdzielczych przedziałowych o otwartych przedziałach klasowych przed obliczeniem średniej należy przedziały domknąć; przyjmuje się, że otwarte przedziały można domknąć, jeśli ich liczebność jest mniejsza niż 0,05N.

Własności średniej arytmetycznej

1. $N\overline{x}_A = \sum_{i=1}^{N} x_i$ $\quad \left(N\overline{x}_A = \sum_{i=1}^{r} x_i n_i ; \quad N\overline{x}_A = \sum_{i=1}^{r} \hat{x}_i n_i \right)$

2. $x_{min} \le \overline{x} \le x_{max}$

3. $\sum_{i=1}^{N} (x_i - \overline{x}) = 0$ $\quad \left(\sum_{i=1}^{r} (x_i - \overline{x}) n_i = 0 \ ; \quad \sum_{i=1}^{r} (\hat{x}_i - \overline{x}) n_i = 0 \right)$

4. $\min_{c \in R} \sum_{i=1}^{N} (x_i - c)^2 = \sum_{i=1}^{N} (x_i - \overline{x})^2$ $\quad \left(\min_{c \in R} \sum_{i=1}^{r} (x_i - c)^2 n_i = \sum_{i=1}^{r} (x_i - \overline{x})^2 n_i \right)$

Własności średniej arytmetycznej (c.d.)

5. Jeśli wszystkie wartości cechy powiększymy (pomniejszymy, pomnożymy, podzielimy) o/przez pewną stałą, to średnia arytmetyczna będzie równa sumie (różnicy, iloczynowi, ilorazowi) średniej arytmetycznej wyjściowej cechy i tej stałej.
6. Średnia arytmetyczna sumy (różnicy) cech równa się sumie (różnicy) ich średnich arytmetycznych.
7. Na poziom średniej arytmetycznej silny wpływ mają wartości ekstremalne (skrajne), przy czym wpływ jest silniejszy w przypadku wysokich wartości cechy.

UWAGA: Średnia arytmetyczna jest miarą prawidłową dla zbiorowości w których rozkład cechy jest jednomodalny, symetryczny lub o niewielkiej asymetrii. Jeśli tak nie jest, to dla scharakteryzowania średniego poziomu zjawiska należy wykorzystać przeciętne pozycyjne.

Załóżmy, że zbiorowość jest podzielona na m rozłącznych grup i znamy średnią arytmetyczną wartości cechy dla każdej z grup.

Niech

$\tilde{x}_j$ (j=1,2,...,m) oznacza średnią arytmetyczną obliczoną dla j-tej grupy,

n_j – liczebność j-tej grupy

N – ogólna liczebność próby

$\bar{x}_A$ – średnia arytmetyczna dla wszystkich grup łącznie

Wtedy

2... $\tilde{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i$ dla j=1,2,...,m oraz $\sum_{j=1}^{m} n_j = N$

3... $\bar{x}_A = \frac{1}{N}\sum_{j=1}^{m} \tilde{x}_j n_j$

Średnia geometryczna

4... $\overline{x}_G = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N} = \sqrt[N]{\prod_{i=1}^{N} x_i}$ dla szeregów szczegółowych

4a... $\overline{x}_G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_r^{n_r}} = \sqrt[N]{\prod_{i=1}^{r} x_i^{n_i}}$ dla szeregów rozdzielczych

Średnia geometryczna ma zastosowanie wtedy, gdy zjawiska ujmowane są dynamicznie, przy badaniu średniego tempa zmian zjawisk w czasie.

Średnia harmoniczna

5… $\overline{x}_H = \dfrac{N}{\sum_{i=1}^{N} \dfrac{1}{x_i}}$ dla szeregów wyliczających

6… $\overline{x}_H = \dfrac{N}{\sum_{i=1}^{r} \dfrac{1}{x_i} \cdot n_i}$ dla szeregów rozdzielczych punktowych

7… $\overline{x}_H = \dfrac{N}{\sum_{i=1}^{r} \dfrac{1}{\hat{x}_i} \cdot n_i}$ dla szeregów rozdzielczych przedziałowych

Średnią harmoniczną stosuje się, gdy wartości cechy podane są w jednostkach względnych (km/godz, kg/osobę).

UWAGA: Dla konkretnej cechy tylko jedna średnia klasyczna jest odpowiednia.

Dominanta

Dominanta (modalna, moda, wartość najczęstsza) w rozkładzie empirycznym Do – ta wartość cechy, której odpowiada największa liczebność (częstość).

- Dominanta nie zawsze istnieje.
- Na podstawie przedziałowego szeregu rozdzielczego dominantę można wyznaczyć jedynie wówczas, gdy przedziały klasowe w tym szeregu mają jednakową rozpiętość (wysoka liczebność mogłaby być spowodowana szerszą rozpiętością tego przedziału w stosunku do innych).
- Dla szeregów rozdzielczych przedziałowych można poprzestać na wskazaniu przedziału zawierającego dominantę.

Zwykle dla dokładniejszego wyznaczenia mody stosuje się wzór interpolacyjny (8), wyprowadzony przy założeniu, że wzrost liczebności w poszczególnych przedziałach klasowych jest proporcjonalny do wzrostu wartości cechy.

8... $$Do = x_{0d} + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \cdot h_d$$

gdzie

x_{0d} – dolna granica przedziału, w którym występuje dominanta

h_d – rozpiętość przedziału z dominantą

n_d, n_{d-1}, n_{d+1} – liczebności przedziału zawierającego dominantę, poprzedniego, następnego (odpowiednio)

Uwaga: We wzorze (8) liczebności można zastąpić częstościami.

Graficzne wyznaczanie dominanty

- Wyznaczyć histogram dla przedziału klasowego zawierającego dominantę, poprzedniego i następnego.
- Z górnych wierzchołków najwyższego prostokąta należy wykreślić dwa odcinki łączące po przekątnej bliższe górne wierzchołki sąsiednich prostokątów.
- Rzut punktu przecięcia tych odcinków na oś odciętych jest dominantą.

Uwagi:

1. Jeśli liczebności przedziałów sąsiadujących z przedziałem dominanty są jednakowe, to dominanta jest równa środkowi klasy dominującej.
2. Wyznaczanie dominanty jest uzasadnione wówczas, gdy rozkład empiryczny jest jednomodalny i asymetria jest umiarkowana.

Kwantyle

Kwantyl rzędu p w rozkładzie empirycznym – taka wartość cechy k_p, dla której jako pierwszej dystrybuanta empiryczna spełnia relację

9… $F(k_p) \geq p$ $0 < p < 1$

W statystyce opisowej wyróżnia się:

- ***kwartyle*** (kwantyle rzędu $\frac{k}{4}$ k = 1,2,3)
- ***decyle*** (kwantyle rzędu $\frac{k}{10}$ k = 1,2,3,…,9)
- ***centyle*** (kwantyle rzędu $\frac{k}{100}$ k = 1,2,3,…,99)

Kwartyle:

Q_1 - kwartyl pierwszy (rzędu $\frac{1}{4}$)

me - kwartyl drugi (rzędu $\frac{2}{4}$) - mediana

Q_3 - kwartyl trzeci (rzędu $\frac{3}{4}$)

Mediana jest tą wartością cechy, którą posiada środkowa jednostka w uporządkowanym (niemalejąco) ciągu wartości cechy; gdy tych jednostek jest więcej bierze się ich średnią arytmetyczną, tzn. dla szeregów wyliczających

10...

$$\text{me} = \begin{cases} x_{\frac{N+1}{2}} & \text{gdy N - nieparzyste} \\ \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}) & \text{gdy N - parzyste} \end{cases}$$

Własności mediany

1. $$\min_{c \in R} \sum_{i=1}^{N} | x_i - c | = \sum_{i=1}^{N} | x_i - me |$$

2. Mediana nie reaguje na zmiany wartości cech skrajnych jednostek (na obserwacje nietypowe).
3. Przy zmianie próby mediana ulega większym zmianom niż średnia arytmetyczna.

Uwaga: Mediana obok średniej arytmetycznej jest najczęściej stosowanym parametrem; może być obliczona, gdy nie można obliczyć średniej arytmetycznej (otwarte przedziały).

Do wyznaczenia kwartyli z szeregów rozdzielczych przedziałowych stosuje się wzór interpolacyjny (11), wyprowadzony przy założeniu, że wzrost liczebności w poszczególnych przedziałach klasowych jest proporcjonalny do wzrostu wartości cechy.

11... $$Q_{4p} = x_{0Q} + \frac{h_Q}{n_Q}[\, p \cdot N - n^{sk}_{Q-1}\,]$$

gdzie:

p – rząd kwartyla

x_{0Q} – dolna granica przedziału, w którym jest wartość kwartyla

h_Q – rozpiętość przedziału kwartyla

n_Q – liczebność przedziału kwartyla

n^{sk}_{Q-1} – liczebność skumulowana w przedziale poprzedzającym klasę kwartyla

N – ogólna liczebność zbiorowości.

Do graficznego wyznaczenia kwartyli wykorzystuje się wielobok skumulowany liczebności (częstości) - łamana łącząca punkty o współrzędnych: górna granica przedziału klasowego, odpowiadająca danej klasie liczebność (częstość) skumulowana.

Wartość kwartyla rzędu p stanowi odczytana na osi odciętych (wartości cechy) liczna odpowiadająca skumulowanej liczebności równej pN (skumulowanej częstości równej p); N jest ogólną liczebnością zbiorowości .

W przypadku rozkładu umiarkowanie asymetrycznego zachodzi wzór Pearsona

12... $\bar{x} - Do \approx 3 \cdot (\bar{x} - me)$

Miary zmienności:

- ***klasyczne***
  - wariancja
  - odchylenie standardowe
  - odchylenie przeciętne
  - współczynnik zmienności
- ***pozycyjne***
  - rozstęp
  - odchylenie ćwiartkowe
  - współczynnik zmienności

Wariancja

Wariancja to średnia arytmetyczna kwadratów odchyleń wartości cechy od średniej

13… $$S^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$ szereg szczegółowy

13a… $$S^2 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_i}{N} = \sum_{i=1}^{r} (x_i - \bar{x})^2 \omega_i$$ szereg rozdzielczy punktowy

13b… $$S^2 = \frac{\sum_{i=1}^{r} (\hat{x}_i - \bar{x})^2 n_i}{N} = \sum_{i=1}^{r} (\hat{x}_i - \bar{x})^2 \omega_i$$ szereg rozdzielczy przedziałowy

gdzie $$\sum_{i=1}^{r} n_i = N$$

Uwaga: Wariancja dla szeregów rozdzielczych przedziałowych jest zawyżona (bierzemy środki klas; liczba przedziałów jest odwrotnie proporcjonalna do ich rozpiętości, więc przeszacowanie jest tym większe im mniej jest klas). Zaleca się stosowanie poprawki Shepparda równej $\frac{h^2}{12}$

14... $$S^2 = \frac{\sum_{i=1}^{r} (\hat{x}_i - \bar{x})^2 n_i}{N} - \frac{h^2}{12} = \sum_{i=1}^{r} (\hat{x}_i - \bar{x})^2 \omega_i - \frac{h^2}{12}$$

gdzie h – rozpiętość przedziałów klasowych.

Dla wariancji zachodzi

15... $$S^2 = \overline{x^2} - \bar{x}^2$$

Jeśli zbiorowość jest podzielona na m rozłącznych grup, to wariancja dla całej zbiorowości, tzw. wariancja ogólna jest sumą dwóch składników:

- wariancji wewnątrzgrupowej (średnia arytmetyczna wariancji grup);
- wariancji międzygrupowej (wariancja średnich grupowych)

Wariancja ogólna wyraża się wzorem

16... $$S^2 = \overline{S_i^2} + S^2(\tilde{x}_i)$$

gdzie: $$\overline{S_i^2} = \frac{\sum_{i=1}^{m} S_i^2 n_i}{N} \qquad S^2(\tilde{x}_i) = \frac{\sum_{i=1}^{m} (\tilde{x}_i - \bar{x})^2 n_i}{N}$$

$\overline{S_i^2}$ - wariancja wewnątrzgrupowa

$S^2(\tilde{x}_i)$ - wariancja międzygrupowa

$\tilde{x}_i$ (i=1,2,...,m) oznacza średnią arytmetyczną obliczoną dla i-tej grupy

n_i - liczebność i-tej grupy

N - ogólna liczebność próby

$\bar{x}$ - średnia arytmetyczna dla wszystkich grup łącznie

Odchylenie standardowe

Odchylenie standardowe S jest to pierwiastek z wariancji.

Wyraża się w mianach takich jak badana cecha. Określa przeciętne zróżnicowanie poszczególnych wartości cechy w stosunku do średniej arytmetycznej.

Typowy obszar zmienności

17… $$\overline{x} - S \leq x_{typ} \leq \overline{x} + S$$

Na ogół w obszarze tym mieszczą się wartości cechy około 2/3 jednostek badanej zbiorowości

Uwagi:

1. Odchylenie standardowe jest najczęściej stosowanym parametrem statystycznym.
2. Obliczane jest na podstawie wszystkich obserwacji.
3. Im zbiorowość bardziej zróżnicowana, tym większa wariancja i odchylenie standardowe.
4. Na podstawie nierówności Czebyszewa, sformułowano tzw. regułę trzech sigm która mówi, że wystąpienie obserwacji o wartości cechy poza przedziałem $(\overline{x} - 3S; \overline{x} + 3S)$ jest mało prawdopodobne.
5. Dla rozkładów normalnych lub zbliżonych do normalnych: tylko 1/3 obserwacji wykracza poza typowy przedział obserwacji $(\overline{x} - S; \overline{x} + S)$

   tylko 5% obserwacji wykracza poza przedział $(\overline{x} - 2S; \overline{x} + 2S)$

   a około 0,3% obserwacji poza przedział $(\overline{x} - 3S; \overline{x} + 3S)$

Odchylenie przeciętne

Odchylenie przeciętne d wyraża się wzorem

18… $$d = \frac{\sum_{i=1}^{N} | x_i - \overline{x} |}{N}$$ szereg szczegółowy

18a… $$d = \frac{\sum_{i=1}^{r} | x_i - \overline{x} | n_i}{N} = \sum_{i=1}^{r} | x_i - \overline{x} | \omega_i$$ szereg rozdzielczy punktowy

18b… $$d = \frac{\sum_{i=1}^{r} | \hat{x}_i - \overline{x} | n_i}{N} = \sum_{i=1}^{r} | \hat{x}_i - \overline{x} | \omega_i$$ szereg rozdzielczy przedziałowy

gdzie $$\sum_{i=1}^{r} n_i = N$$

Rozstęp

Rozstęp R to bardzo ogólna miara zmienności

19... $R = x_{max} - x_{min}$

Odchylenie ćwiartkowe

Odchylenie ćwiartkowe Q mierzy poziom zróżnicowania tylko części jednostek (po odrzuceniu 25% o najmniejszej i 25% o największej wartości cechy)

20... $Q = \frac{(Q_3 - me) + (me - Q_1)}{2} = \frac{Q_3 - Q_1}{2}$

Zachodzi związek $Q < d < S$.

Typowy obszar zmienności cechy (w oparciu o parametry pozycyjne)

20.... $me - Q \leq x_{typ} \leq me + Q$

Współczynnik zmienności

Dotychczas omówione miary dyspersji są miarami bezwzględnymi (w takich jednostkach jak cecha); nie można więc ich wykorzystać do porównywania rozkładów cech w zbiorowościach. Dlatego w analizie dyspersji stosuje się względną miarę rozproszenia – współczynnik zmienności.

Współczynnik zmienności jest stosunkiem bezwzględnej miary zróżnicowania cechy do przeciętnej wartości cechy (jest miarą niemianowaną, najczęściej podawaną w procentach).

W zależności od przyjętych miar przeciętnych i dyspersji wyróżnia się współczynniki zmienności:

- Klasyczne

21... $$V_S = \frac{S}{\bar{x}}$$

22... $$V_d = \frac{d}{\bar{x}}$$

- Pozycyjne

23... $$V_Q = \frac{Q}{me}$$

24... $$V_{Q_1Q_3} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad \left(Q = \frac{Q_3 - Q_1}{2} \right)$$

Miary asymetrii

Rozkłady mogą różnić się kierunkami i siłą asymetrii.

W szeregach symetrycznych $\bar{x} = me = Do$ $(Q_3 - me) - (me - Q_1) = 0$

$\bar{x} - Do$ - wskaźnik skośności (określa kierunek asymetrii)

$(Q_3 - me) - (me - Q_1)$ - pozycyjny wskaźnik skośności

Asymetria lewostronna: $\bar{x} < me < Do$

$(Q_3 - me) - (me - Q_1) < 0$

Asymetria prawostronna:

$\bar{x} > me > Do$

$(Q_3 - me) - (me - Q_1) > 0$

Miary asymetrii (współczynniki skośności) określają kierunek i siłę asymetrii

Klasyczne współczynniki asymetrii:

25... $$A_S = \frac{\bar{x} - Do}{S}$$

26... $$A_d = \frac{\bar{x} - Do}{d}$$

27... $$A = \frac{m_3}{S^3}$$ gdzie $$m_3 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N}$$ $$\left(m_3 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^3 n_i}{N} \right)$$

Pozycyjny współczynnik asymetrii

28... $$A_Q = \frac{(Q_3 - me) - (me - Q_1)}{(Q_3 - me) + (me - Q_1)} = \frac{Q_3 + Q_1 - 2me}{2Q}$$

Uwaga: Im większa wartość bezwzględna współczynnika asymetrii, tym silniejsza asymetria

Miary koncentracji

Współczynnik skupienia (kurtoza) – miara skupienia obserwacji wokół średniej

29... $K = \frac{m_4}{S^4}$ gdzie $m_4 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N}$ $\left(m_4 = \frac{\sum_{i=1}^{r}(x_i - \bar{x})^4 n_i}{N}\right)$

Im wyższa wartość K, tym bardziej wysmukła krzywa liczebności, więc większa koncentracja wartości cechy wokół średniej.

Jeśli zbiorowość ma rozkład normalny, to K = 3.

K < 3 -rozkład ***platokurtyczny*** bardziej spłaszczony od normalnego

K > 3 - rozkład ***leptokurtyczny*** bardziej wysmukły od normalnego

30... $K' = \frac{m_4}{S^4} - 3$

$K' < 0$ - rozkład platokurtyczny

$K' > 0$ -rozkład leptokurtyczny