Wykład nr 2

Statystyka opisowa część 2

Plan wykładu

1. Uwagi wstępne

2. Miary tendencji centralnej

2.1. Wartości średnie

2.2. Miary pozycyjne

2.3. Dominanta

3. Miary rozproszenia

4. Miary asymetrii

5. Miary koncentracji

6. Podsumowanie

Statystyka opisowa

Cel – zwięzłe przedstawienie ogólnej charakterystyki istotnych właściwości badanej zbiorowości.

Podstawowe zadania:

1. Określenie przeciętnego wielkości i rozmieszczenia wartości zmiennej – miary położenia / tendencji centralnej.
2. Określenie granic zmienności wartości zmiennej – miary rozproszenia (zmienności, dyspersji).
3. Określenie parametrów rozkładu wartości zmiennej – miary asymetrii i koncentracji rozkładu oraz podobieństwa struktury.
4. Ocena zmienności zjawisk w czasie.
5. Określenie współzmienności – miary współzmienności.

Statystyki opisowe w odniesieniu do skal pomiarowych

Miary położenia

Miary położenia (przeciętne) wskazują miejsce, w którym leży wartość najlepiej reprezentująca wielkości wchodzące w skład szeregu statystycznego.

Miary położenia informują o przeciętnym (średnim, typowym) poziomie wartości rozważanej cechy.

Typowe miary położenia:

Średnia arytmetyczna

Stosowane dla skal metrycznych

Średnia arytmetyczna

Definiowana jako

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

gdzie x_i wartość i-tego pomiaru, a n – liczebność populacji.

Przykład:

W pewnym doświadczeniu medycznym bada się czas snu (w minutach) pacjentów. Zmierzono u $n = 12$ losowo wybranych pacjentów czas: 435, 389, 533, 324, 561, 395, 416, 500, 499, 397, 356 i 398.

Średni czas $\bar{x} = \frac{435 + 389 + \ldots + 398}{12} = ?$

Własności średniej arytmetycznej:

$$\sum_{i=1}^{n} (x_i - \bar{x})$$

inne (mniej dogodne później)

Inne rodzaje średnich arytmetycznych

Średnia ważona (w_i to tzw. wagi):

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Przykład testu psychologicznego:

A co z szeregami rozdzielczymi?

Średnia szeregu punktowego:

Jeśli wartości wyników, jakie przybiera zmienna w próbie $x_1, x_2, \ldots, x_k$ występują z licznością $n_1, n_2, \ldots, n_k$, to średnia arytmetyczna jest zdefiniowana jako:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \cdot x_i}{n} ,$$

gdzie k jest liczbą różnych wartości zmiennej, a $n = \sum_{i=1}^{k} n_i$.

Przykład:

W przedsiębiorstwie zajmującą się produkcją pewnych narzędzi badano wydajność pracy (ilość sztuk wyprodukowanych w ciągu dnia) 60 pracowników. Ustalmy, jaka jest średnia wydajność pracy przypadająca na jedną osobę.

Dostępne jest zestawienie:

Średnia arytmetyczna dla szeregu rozdzielczego przedziałowego

zdefiniowana jako:

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i \cdot \dot{x}_i}{n},$$

gdzie k jest liczbą różnych wartości zmiennej,

n_i – liczność i-tego przedziału klasowego,

$\dot{x}_i$ – wartość średnia i-tego przedziału klasowego.

Komentarz – jest to przybliżony sposób obliczeń.

Przykład 4.

Czy dotychczasowe sposoby wyznaczania średniej są zawsze właściwe?

Ograniczenia średniej arytmetycznej:

- wartości skrajne mają silny wpływ na jej wartość,
- średniej arytmetycznej nie można policzyć, gdy skrajne przedziały szeregu są rozwarte,
- traci swoją wartość poznawczą w przypadku rozkładów silnie asymetrycznych i wielomodowych,
- nieadekwatna dla niemetrycznych skal, prób małolicznych.

Przykład obliczenia średniej z szeregu…

Rozważmy rozkład miesięcznych zasadniczych wynagrodzeń pracowników zatrudnionych w pewnej firmie. (…) Oblicz średnie wynagrodzenie pracownika z wyższym wykształceniem.

Inne rodzaje średnich

Średnia ucinana (z parametrem k) :

$$\bar{x}_{tk} = \frac{1}{n - 2 \cdot k} \sum_{i=k+1}^{n-k} x_i$$

Średnie dla analizy dynamik zjawisk i stosowane dla skal ilorazowych

Średnia geometryczna –stosowana dla oceny średniego tempa zmian zjawiska w czasie (oraz gdy w szeregu występują znaczne różnice między obserwacjami; mniej wrażliwa na krańcowe obserwacje „odstające").

Zdefiniowana jako:

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \ldots x_n}$$

Średnie harmoniczne:

Używana, gdy wartości zmiennej podane są w jednostkach względnych, np. przeciętna szybkość (w km/godz.), przeciętna cena towarów (wyrażona w liczbie jednostek za jednostkę pieniężną), gęstość zaludnienia (os/km^2).

Średnia harmoniczna prosta – zdefiniowana jako

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

Miary pozycyjne

Określają pozycję pewnego (typowego) przypadku w stosunku do innych przypadków (ze względu na ich położenie w zbiorowości).

Mediana (wartość środkowa) próby

Jest to wartość, dla której dokładnie połowa wyników w próbie jest od niej mniejsza lub równa, a druga połowa jest od niej większa lub równa.

Jak wyznaczać medianę?

$$x_{med} = \begin{cases} x_{((n+1)/2)} & \text{gdy n jest nieparzyste} \\ 0{,}5 \cdot (x_{(n/2)} + x_{(n/2+1)}) & \text{gdy n jest parzyste} \end{cases}$$

Przykład 9.

Zapytano 7 osób o wiek i otrzymano następujące odpowiedzi:

18, 21, 43, 27, 51, 35, 29 lat. Wyznacz medianę.

Wyznaczenie mediany z szeregu rozdzielczego:

$$x_{med} = x_0 + \frac{h_0}{n_0} \cdot (\frac{n}{2} - F_{-1})$$

gdzie:

x_0 – dolna granica przedziału klasowego, który zawiera pierwszych 50% skumulowanych częstości.

h_0 – rozpiętość przedziału klasowego zawierającego medianę.

n_0 – częstość odpowiadająca przedziałowi klasowemu zawierającego medianę.

n – ogólna liczna obserwacji.

F_{-1} – częstość skumulowana przedziału poprzedzającego przedział klasowego, który zawiera pierwszych 50% skumulowanych częstości.

Przykład obliczeń:

Kwantale (decyle, kwartyle, percentyle)

Wartości cechy (mierzonej na skali, co najmniej porządkowej), które dzielą próbę na określone części pod względem liczby obserwacji.

Najczęściej stosowane:

– kwartale (podział na 4 części),

– decyle (podział na 10 części),

– percentyle (podział na 100 części),

Ilustracja graficzna:

Dominanta (wartość modalna, moda)

Definicja: jest to ta kategoria zmiennej nominalnej (lub porządkowej czy wartość liczbowa), która występuje najczęściej.

Przykład.

Zbadano marki komputerów, które używa 10 osób. Są one następujące:

HP, Del, Cm, HP, IBM, Cm, Cm, No, Cm, IBM.

Wyznaczyć wartość modalną.

Wyniki pomiarów powinny być wcześniej pogrupowane w odpowiednie szeregi.

Wyznaczanie dominanty jest dopuszczalne, gdy rozkład zmiennej jest jednomodalny a jego asymetria jest umiarkowana.

Jeśli zbiorowość jest niejednorodna, a rozkład zmiennej ma 2,3 lub więcej szczytów to mówimy o rozkładach binominalnych, trimodalnych, itd.

Wyznaczenie dominanty z szeregu rozdzielczego:

$$x_{\text{mod}} = x_0 + \frac{n_0 - n_{-1}}{(n_0 - n_{-1}) + (n_0 - n_{+1})} \cdot h_0$$

gdzie:

x_0 – dolna granica przedziału klasowego z największą częstością.

h_0 – rozpiętość przedziału klasowego zawierającego dominantę

n_0 – częstość odpowiadająca przedziałowi klasowemu z największą częstością.

n_{-1} – częstość odpowiadająca przedziałowi poprzedzającemu.

n_{+1} – częstość odpowiadająca przedziałowi następnemu.

Przykład:

Miary rozproszenia

Nazywane także miarami zmienności lub dyspersji.

Służą do oceny czy wartości cechy są bardzo rozproszone lub skoncentrowane wokół wartości przeciętnej.

Miary rozproszeni są zależne od skali pomiarowej.

Skala przedziałowa

Rozstęp (różnica między pomiarem najwyższym i najniższym).

$$R = \mathrm{x}_{max} - x_{min}$$

Odchylenie średnie

$$D_m = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Wariancja i *odchylenie standardowe*:

wariancja w próbie

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

odchylenie standardowe s – pierwiastek z wariancji

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

Różnice w obliczaniu dla próby i populacji

Przykład:

W dwóch grupach chorych zmierzono ciśnienie skurczowe krwi. Otrzymano następujące wyniki: (Grupa-1 145, 125, 130, 155, 140, 150, 135) (Grupa-2 115, 150, 100, 180, 140, 165, 130). Oblicz podstawowe miary tendencji centralnej i rozproszenia.

Własności odchylenia standardowego:

- jest tym większe, im większy jest rozrzut wokół średniej,
- gdyby wszystkie pomiary były sobie równe, odchylenie standardowe również byłoby równe zero,
- w przypadku rozkładu normalnego – obowiązuje tzw. reguła „trzech sigm".

Inne komentarze:

- żaden wskaźnik rozproszenia nie powinien zmieniać swej wartości, gdy do wszystkich elementów próby zostanie dodana ta sama liczba (dodatnią lub ujemną),
- pomnożenie każdego elementu próby przez tę samą liczbę powinno prowadzić do pomnożenia wskaźnika przez wartość bezwzględną tej liczby,
- wskaźniki rozproszenia nie są odporne na wartości odstające w próbie.

Wyznaczenie wariancji z szeregu rozdzielczego:

$$s^2 = \frac{\sum_{i=1}^{k} n_i \cdot (\dot{x}_i - \bar{x})^2}{n-1}$$

Przykład:

Inne miary rozproszenia

Skala nominalna – *miary informacji* (np. entropia rozkładu danych).

Skala porządkowa – *rozstęp międzykwartylowy* - ćwiartkowy (przedział między kwartylem pierwszym a trzecim).

$$IQR = Q_3 - Q_1$$

Skala ilorazowa – tzw. *względne miary zróżnicowania*

Współczynnik zmienności:

$$V_s = \frac{s}{\overline{x}}$$

Stosowany w porównywania siły dyspersji:

– kilku zbiorowości pod względem tej samej zmiennej,

– jednej zbiorowości, ale ze względu na kilka różnych zmiennych.

Przykład:

Miary asymetrii

Po co są potrzebne?

Przykład ilustracyjny

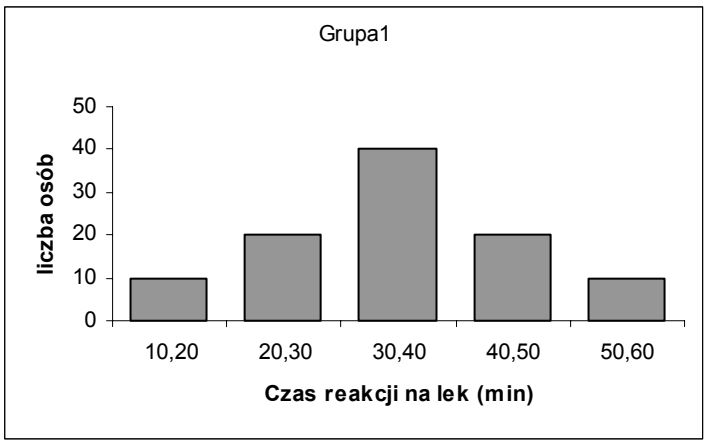
Badano czas reakcji (w minutach) na lek w trzech grupach 100-osobowych. Dane przedstawiono w poniższej tabeli

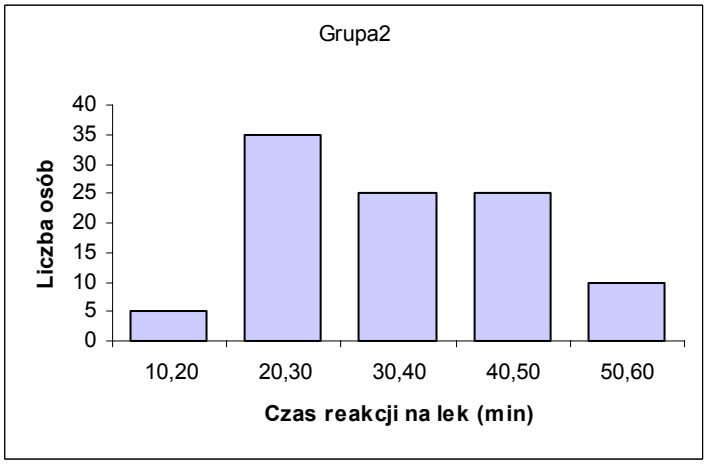
Czas reakcji	Grupa 1	Grupa 2	Grupa 3
10,20	10	5	10
20,30	20	35	25
30,40	40	25	25
40,50	20	25	35
50,60	10	10	5

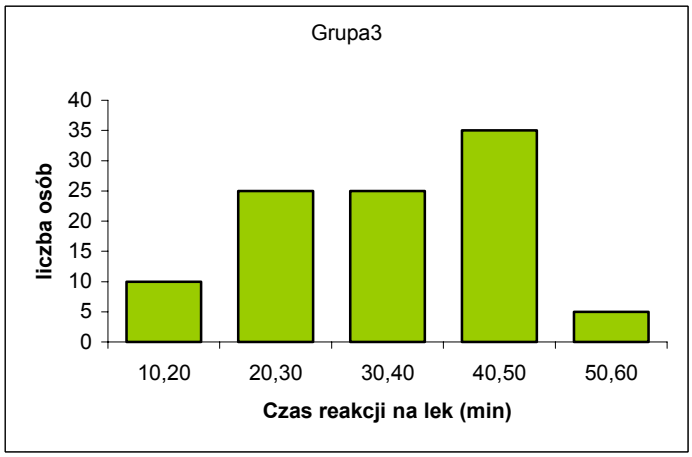
Średnia arytmetyczna i wariancja są jednakowe dla wszystkich grup i wynoszą $\overline{x}=35$ oraz $s^2=120$.

Ale czy grupy są identyczne? A co z wizualizacją rozkładów?

Wykonajmy histogramy!







Dokonaj interpretacji!

Miary koncentracji

Podsumowanie

Omówiono:

Więcej o danych:

późniejsze wykłady + literatura.

Warto **rozszerzyć wiedzę**, np.

- zapoznać się z zagadnieniami,
- ,
- …