Statystyka
w pracy badawczej nauczyciela

Wykład 3: Analiza struktury zbiorowości statystycznej

dr inż. Walery Susłow
walery.suslow@ie.tu.koszalin.pl

Zadania analityczne (1)

- Analiza przewiduje badanie podobieństw i różnic, jakie występują między elementami danej zbiorowości ze względu na wyróżnioną cechę statystyczną.
- Podobieństwa badamy określając *przeciętny poziom* (tendencję centralną) cechy w danej zbiorowości.
- Różnice badamy, określając *dyspersję*, *asymetrię* i *koncentrację* wariantów cechy.

Zadania analityczne (2)

- *Dyspersja* (zmienność, rozproszenie) – zróżnicowanie wariantów cechy w stosunku do przeciętnego poziomu tej cechy.
- *Asymetria* (skośność) – wewnętrzne rozmieszczenie jednostek w ramach danej zbiorowości z punktem widzenia cechy badanej.
- *Koncentracja* – stopień nierównomierności rozkładu wariantów cechy pomiędzy poszczególne jednostki danej zbiorowości oraz stopień skupienia wariantów cechy wokół przeciętnego poziomu tej cechy.

Parametry opisujące zbiorowość

Klasyczne	Pozycyjne
Przeciętny poziom	
x – średnia arytmetyczna	**D** – dominanta, **Me** – mediana
G – średnia geometryczna	Q_1, Q_3 – kwartyle
H – średnia harmoniczna	**Dci** – decyle
Ocena dyspersji	
d – odchylenie przeciętne	O_2 – obszar zmienności
s – odchylenie standardowe, s^2 - wariancja	**Q** – odchylenie ćwiartkowe
V_x - współczynnik zmienności	V_{Me} – pozycyjny współczynnik zmienności
Ocena asymetrii	
W_S – współczynnik asymetrii	A_S – współczynnik asymetrii
α_3 – moment 3 centralny, standaryzowany	
Ocena skupienia	
α_4 – moment 4 centralny, standaryzowany	

Średnia arytmetyczna (SA)

- Jest to suma wartości cechy wszystkich jednostek badanej zbiorowości podzieloną przez liczbę tych jednostek.
- Wadą SA jest wrażliwość na przypadki odstające.
- Wartości badanej cechy oznaczymy przez: x_1, x_2, x_3, … , x_n, SA szeregu statystycznego wyznaczymy z poniższego wzoru:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Średnia arytmetyczna — przykład

- Dwóch pedagogów bada uczniów. Przeprowadzono obserwację czasu trwania tych badań w minutach i zanotowano następujące wyniki:
- Dla pedagoga A: 12, 15, 15, 18, 20
- Dla pedagoga B: 10, 10, 12, 12, 15, 15, 18, 20, 21, 21
- Korzystając ze wzoru uzyskujemy:

$$\bar{x}_A = \frac{12+15+15+18+20}{5} = \frac{80}{5} = 16 \text{ min}$$

$$\bar{x}_B = \frac{10+10+12+12+15+15+18+20+21+21}{10} = \frac{154}{10} = 15{,}4 \text{ min}$$

Średnia geometryczna (SG)

- Jest pierwiastkiem *n*-tego stopnia z iloczynu *n* wartości zmiennej.
- Jest właściwą charakterystyką centralnego skupienia:
  - gdy wartości jednostek zbiorowości statystycznej mają charakter miar względnych.
  - gdy zjawisko wykazuje wyraźną asymetrię,
  - gdy brak ważkich argumentów dla pominięcia wartości ekstremalnych.
- Jest wykorzystywana do opisywania średniego tempa zmian zjawisk.

$$\bar{x}_G = \sqrt[n]{x_1 \, x_2 \ldots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Średnia geometryczna — przykład

- W ciągu trzech kolejnych lat liczba dzieci przyjętych do przedszkoli wynosiła odpowiednio: 500, 750, 825. Jaki był średni względny przyrost liczby nowych wychowanków?
- Uwaga: wartości cechy statystycznej w tym zadaniu to przyrosty liczby wychowanków w kolejnych latach, tzn.:

$$x_1 = \frac{750}{500} = 1{,}5$$

$$x_2 = \frac{825}{750} = 1{,}1$$

$$\bar{x}_G = \sqrt{1{,}5 \cdot 1{,}1} = 1{,}28$$

Gdyby w tym przykładzie zastosować średnią arytmetyczną (1,5+1,1)÷2=1,3, to wynikałoby z tego, że w 3-cim roku, powinno być 500×1,3×1,3=845 nowych wychowanków.

Średnia harmoniczna (SH)

- Jest odwrotnością średniej arytmetycznej z odwrotności wartości zmiennych.
- Jest stosowana zdecydowanie rzadziej niż arytmetyczna. Konieczność jej użycia zachodzi, gdy wartości cechy statystycznej podawane są w przeliczeniu na stałą jednostkę innej zmiennej, np. prędkość w km/h, gęstość zaludnienia w osobach/km^2, spożycie w kg/osobę, itp.
- W przypadku szeregów szczegółowych SH obliczamy według wzoru:

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

Średnia harmoniczna — przykład

- W ciągu 8 godzin pracy w zakładzie poprawczym obserwowano pracę trzech wychowawców. Na wykonanie obowiązków związanych z jednym wychowankiem wychowawca A potrzebował 4 min, wychowawca B – 6 min, a wychowawca C – 12 min. Jaki jest średni czas zużywany na jednego wychowanka?

$$\bar{x}_H = \frac{3}{\frac{1}{4}+\frac{1}{6}+\frac{1}{12}} = 6\,\text{min/osobę}$$

Średnia arytmetyczna wynosi w tym przypadku 7.3 min/osobę, przy takim tempie pracy, trzech wychowawców w ciągu zmiany (8 godzin) zbadałoby 3×480÷7,333=196 osób. W rzeczywistości wychowawca A mógłby zbadać 480÷4=120 wychowanków, wychowawca B – 480÷6=80, a wychowawca C – 480÷12=40, co daje łącznie 120+80+40=240 wychowanków.

Wartość modalna (dominanta)

- Jest to wartość cechy, która najczęściej (najliczniej) występuje w badanej zbiorowości statystycznej.
- Można, stwierdzić, że jest to wartość typowa dla tej zbiorowości.
- Wartość modalną przedstawiać będziemy następująco: $D = x_d$, gdzie x_d wartość *i-tej* cechy, dla której $n_i = max$.
- Wartość modalna, jako miara pozycyjna, jest odporna na występowanie przypadków odstających.

Mediana albo wartość środkowa (1)

- Jest to wartość cechy, która rozdziela zbiorowość na dwie równe części, zajmując środkową pozycję w szeregu statystycznym.
- Sposób wyznaczania wartości mediany M_e uzależniony jest od wielu czynników, a do najważniejszych należy zaliczyć:
  - liczbę elementów szeregu (parzysta czy nieparzysta),
  - typ szeregu (szereg szczegółowy czy rozdzielczy).

Mediana albo wartość środkowa (2)

Gdy *n* jest nieparzyste

$$M_e = x_{\frac{n+1}{2}}$$

$$P_{Me} = \frac{n+1}{2}$$

Gdy *n* jest parzyste

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

$$P_{Me} = \frac{n}{2}$$

Kwantyle

- Mediana informuje, poniżej i powyżej jakiej wartości cechy znajduje się 50% zbiorowości.
- Można podzielić zbiorowość na większą liczbę części.
- Wartości ustalające podział nazywamy kwantylami.
- Najczęściej stosowane są:
  - *kwartyle* – dzielą szereg statystyczny na 4 części,
  - *decyle* – dzielą szereg statystyczny na 10 części,
  - *centyle* – dzielą szereg statystyczny na 100 części.

Kwartyle

- *Kwartyl pierwszy (Q1)* dzieli zbiorowość na dwie części tak, że 25% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi pierwszemu, a 75% równe bądź wyższe.
- *Kwartyl drugi* jest to mediana (Me).
- *Kwartyl trzeci (Q3)* dzieli zbiorowość na dwie części tak, że 75% jednostek zbiorowości ma wartości cechy niższe bądź równe kwartylowi trzeciemu, a 25% równe bądź wyższe.
- Kwartyle wyznacza się w sposób analogiczny do mediany.

Decyle

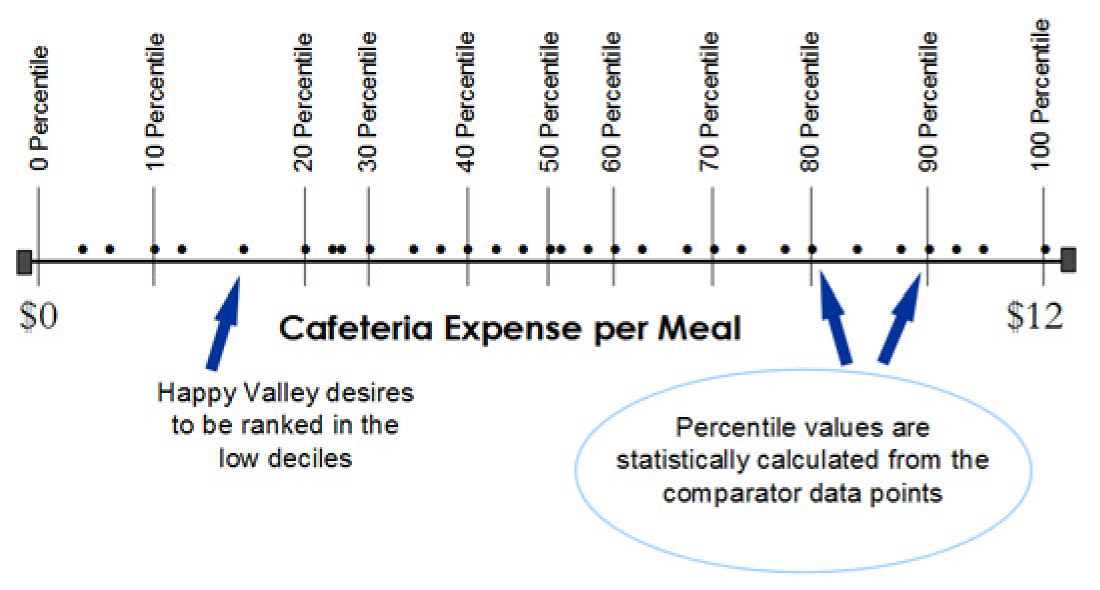


Table I. Response Decile Analysis

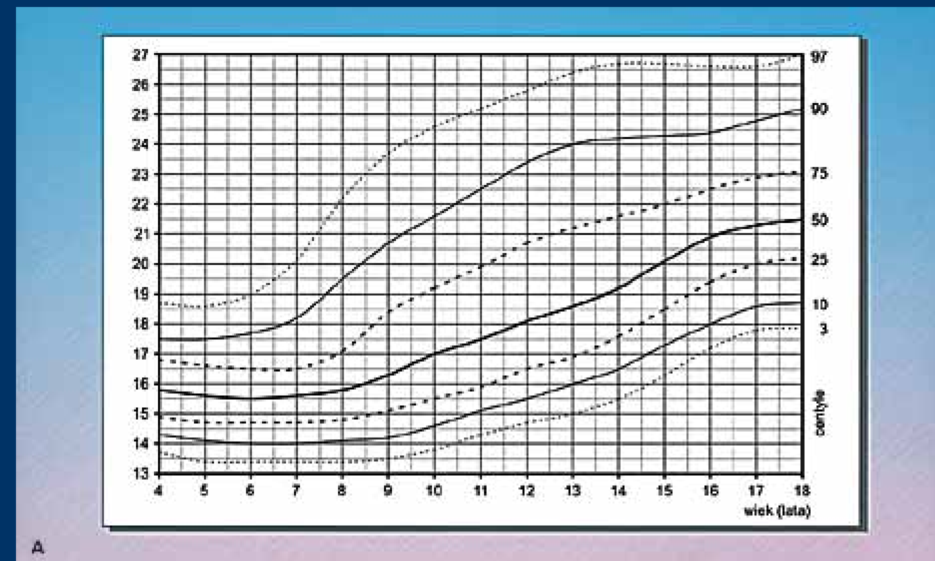
Decile	Number of Individuals	Number of Responders	Decile Response Rate	Cumulative Response Rate	Cum Lift
top	7,410	911	12.30%	12.30%	294
2	7,410	544	7.30%	9.80%	235
3	7,410	437	5.90%	8.50%	203
4	7,410	322	4.30%	7.50%	178
5	7,410	258	3.50%	6.70%	159
6	7,410	188	2.50%	6.00%	143
7	7,410	130	1.80%	5.40%	129
8	7,410	163	2.20%	5.00%	119
9	7,410	124	1.70%	4.60%	110
bottom	7,410	24	0.30%	4.20%	100
Total	74,100	3,101	4.20%		

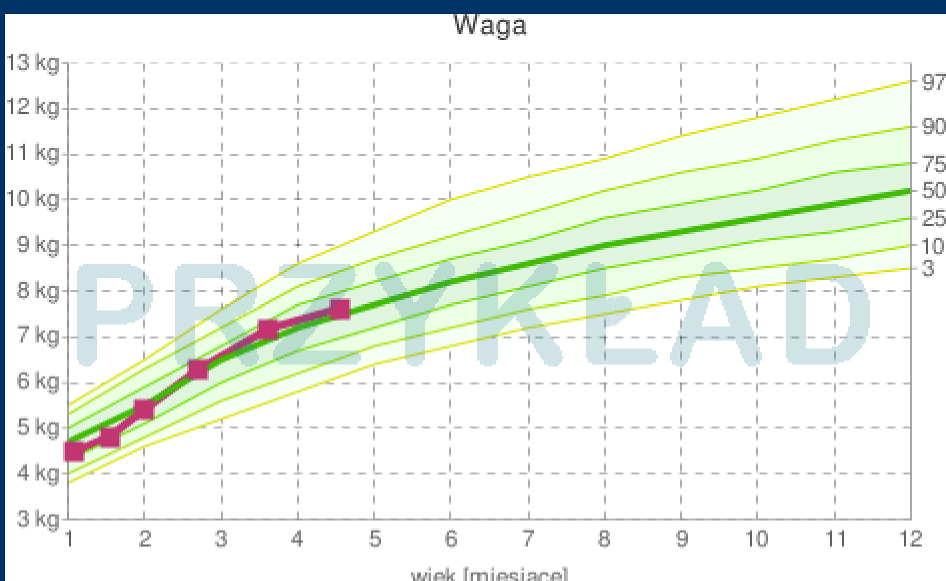
Centyle

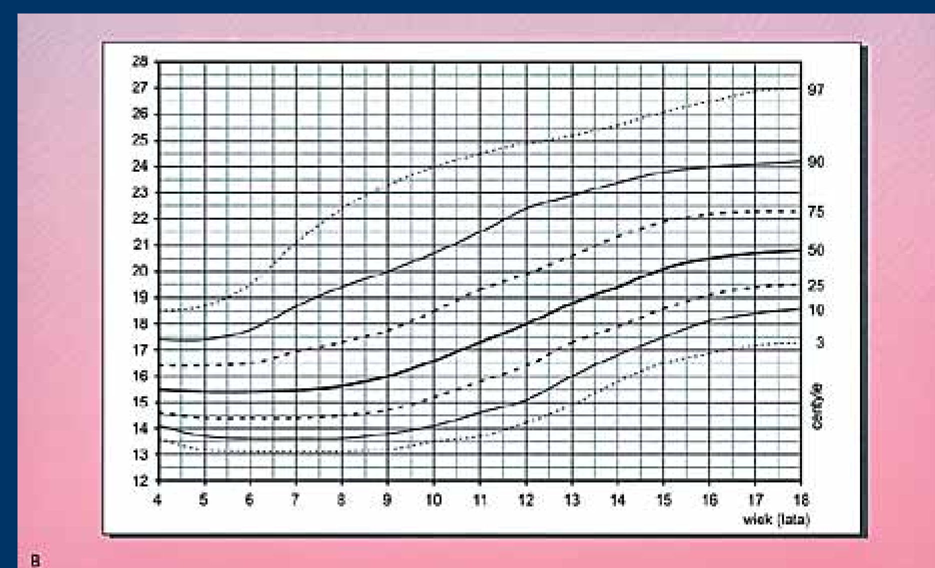
- Centyle stosowane są dla prób o dużej liczebności. Centyl 50 odpowiada medianie, a centyle 25 i 75 to odpowiednio kwartyle Q1 i Q3.
- Centyle są często stosowane do odnoszenie różnych pomiarów antropometrycznych u badanego dziecka do ogółu populacji dzieci.
- Siatki centylowe są to wykresy na których zaznaczono kilka wybranych centyli (zwykle 3, 10, 25, 50, 75, 90 i 97) w zależności od wieku dla wybranego parametru antropometrycznego.

Siatki centylowe (1)

Siatki centylowe wskaźnika względnej masy ciała (BMI) chłopców i dziewcząt warszawskich (opr. I. Palczewska, Z. Niedźwiecka, 1999 r.)



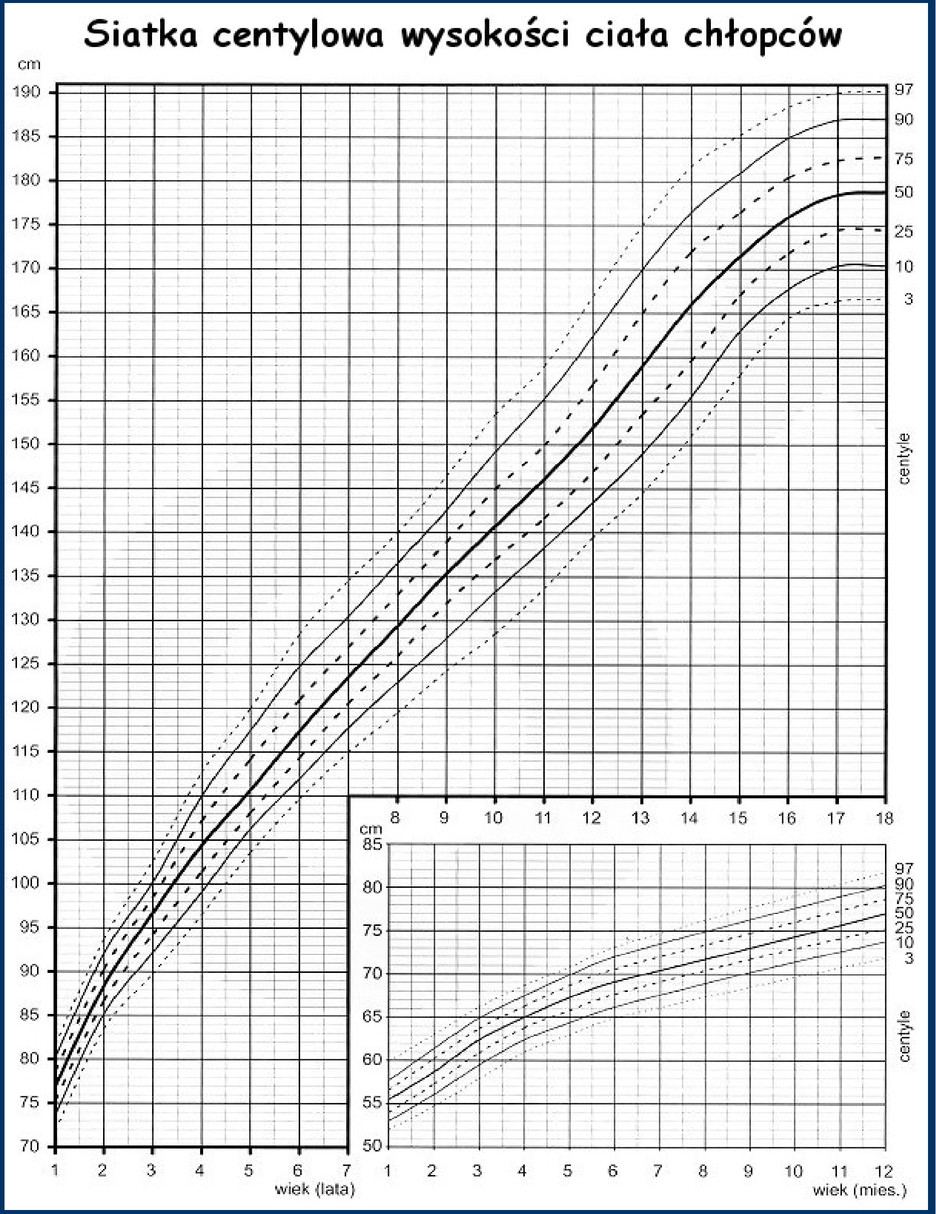




Siatki centylowe (2)

Siatka centylowa wzrostu u chłopców

Przykładowe zadanie:
Ocenić wzrost 10 letniego chłopca, mierzącego 130 cm.



Ocena rozproszenia populacji

- Rozrzut wartości cechy statystycznej wokół wartości przeciętnej opisują miary zmienności.
- *Miary zmienności* charakteryzują stopień zróżnicowania elementów zbiorowości pod względem badanej cechy:
  - Miary *klasyczne* – ich wartość jest wyznaczana w oparciu o wszystkie obserwacje – są to *wariancja, odchylenie standardowe, odchylenie przeciętne, współczynnik zmienności*.
  - Miary *pozycyjne* – na ich wartość wpływają tylko wybrane obserwacje z próby uporządkowanej – są to *rozstęp, odchylenie ćwiartkowe, współczynniki zmienności*.

Wariancja

- Jest to średnia arytmetyczna kwadratów odchyleń poszczególnych obserwacji od średniej arytmetycznej zbiorowości.
- Nie jest wyrażona w jednostkach cechy, ale w jednostkach podniesionych do kwadratu.
- Jest wskaźnikiem "rozrzutu" wartości zmiennej losowej wokół jej wartości oczekiwanej – średniej.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Odchylenie standardowe

- Jest pierwiastkiem kwadratowym wariancji.
- Jest miarą zmienności o jednostce zgodnej z mianem badanej cechy statystycznej.
- Określa przeciętne zróżnicowanie poszczególnych wartości cechy od średniej arytmetycznej.

$$s=\sqrt{\frac{\left(x_1-\bar{x}\right)^2+\left(x_2-\bar{x}\right)^2+\ldots+\left(x_n-\bar{x}\right)^2}{n}}=\sqrt{\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{n}}$$

Typowy obszar zmienności populacji

- Określamy, znając średnią arytmetyczną i odchylenie standardowe próby:

$$(\bar{x}-s)<x_{typowe}<(\bar{x}+s)$$

- W tym zakresie mieszczą się około 2/3 (dokładniej 68%) wszystkich elementów badanej zbiorowości statystycznej:
  - Poza przedział ± jednego odchylenia standardowego od średniej wykracza około 32% obserwacji.
  - Poza przedział ± dwóch odchyleń standardowych – tylko około 5% obserwacji.
  - Prawdopodobieństwo, wystąpienia obserwacji spoza przedziału ± trzech odchyleń standardowych od średniej jest uważane za znikome, bo wynosi około 0,3%.

Odchylenie przeciętne

- Jest to średnia arytmetyczna bezwzględnych odchyleń wartości cechy od jej średniej arytmetycznej.
- Jest miarą rzadziej stosowaną w analizach statystycznych niż odchylenie standardowe.
- Odchylenie przeciętne jest zawsze mniejsze od odchylenia standardowego, policzonego dla tego samego szeregu.

$$d=\frac{|x_1-\bar{x}|+|x_2-\bar{x}|+\ldots+|x_n-\bar{x}|}{n}=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$$

Cd. przykładu ze slajdu 6

- Wariancja czasu badania ucznia przez pedagogów A i B wynosi:

$$s_A^2=\frac{(12-16)^2+(15-16)^2+(15-16)^2+(18-16)^2+(20-16)^2}{5}=$$

$$\frac{16+1+1+4+16}{5}=\frac{38}{5}=7{,}6\text{min}^2\Rightarrow s_A=\sqrt{7{,}6}=2{,}76\text{min}$$

$$s_B^2=17{,}24\text{min}^2\Rightarrow s_B=\sqrt{17{,}24}=4{,}15\text{min}$$

- Więc, typowy obszar zmienności wynosi:
  - dla pedagoga A: 16±2,76 min (13,24 min; 18,76 min)
  - dla pedagoga B: 15,4±4,15 min (11,25 min; 19,55 min)

Rozstęp

- Jest to różnica między najwyższą i najniższą zaobserwowaną wartością cechy statystycznej: $R = x_{max} - x_{min}$.
- Jest najprostszą miarą zmienności.
- Opisuje empiryczny obszar zmienności badanej cechy, nie dając informacji o zróżnicowaniu poszczególnych wartości cechy w całej zbiorowości.
- Zależy od jednostkowych przypadków ekstremalnych, a nie od zróżnicowania typowych obserwacji przeważających w zbiorowości.

Odchylenie ćwiartkowe

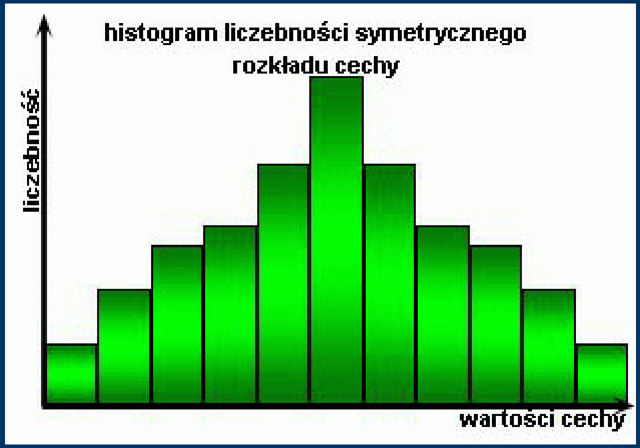
- Jest to połową różnicy między trzecim, a pierwszym kwartylem: $$Q = \frac{(Q_3 - Q_1)}{2}$$
- Jest miarą zmienności najczęściej używaną w parze z medianą.
- Mierzy poziom zróżnicowania 50% elementów, pozostałych po odrzuceniu wartości najmniejszych (25%) i największych (25%).
- Typowy obszar zmienności przez kwartyli wyrażamy tak: $(Q_1 < x_{typ} < Q_3)$

Współczynniki zmienności

- Jest to iloraz bezwzględnej miary zmienności cechy i przeciętnej wartości tej cechy (V_x i V_{Me}).
- Stosuje się do oceny zróżnicowania kilku zbiorowości pod względem tej samej cechy, ewentualnie tej samej zbiorowości pod względem kilku cech.
- Jest wielkością niemianowaną. Wartości współczynników podaje się z reguły w procentach.
- Jest przyjęte, że *zbiorowość* uważa się za *niejednorodną*, gdy współczynnik zmienności jest większy niż *10%*.

Miary asymetrii zbiorowości

- Porównanie średniego poziomu cechy i jej rozproszenia nie zawsze wykazuje różnic między badanymi zbiorowościami.
- Rozkłady mogą być symetryczne lub asymetryczne (prawostronnie lub lewostronnie) przy tym samym rozproszeniu.



Miary asymetrii zbiorowości

- Asymetrię można ocenić porównując różnice pomiędzy średnią arytmetyczną, a medianą lub modalną.
  - Asymetria prawostronna: $\bar{x} > Me > Mo$
  - Asymetria lewostronna: $\bar{x} < Me < Mo$

- Współczynnik asymetrii (wskaźnik skośności) przyjmuje wartość 0 przy braku asymetrii, wartości dodatnie przy asymetrii prawostronnej, a wartości ujemne przy asymetrii lewostronnej:

$$W_s = (\bar{x} - Mo) \qquad A_s = (Q_3 - Me) - (Me - Q_1)$$

Klasyczny współczynnik asymetrii

$$a_3 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3 \cdot n} = \frac{(x_1 - \bar{x})^3 + (x_2 - \bar{x})^3 + \ldots + (x_n - \bar{x})^3}{s^3 \cdot n}$$