Statystyka opisowa

Cz. 2

Bartłomiej Karaban

Program

- Miary zmienności
- Miary asymetrii
- Korelacja

Miary zmienności

Zbiór 1	1	2	3	4	5	6	6	7	8	9	10	11
Zbiór 2	4	5	5	5	6	6	6	6	7	7	8	8

Miary zmienności

Zbiór 1	1	2	3	4	5	6	6	7	8	9	10	11
Zbiór 2	4	5	5	5	6	6	6	6	7	7	8	8

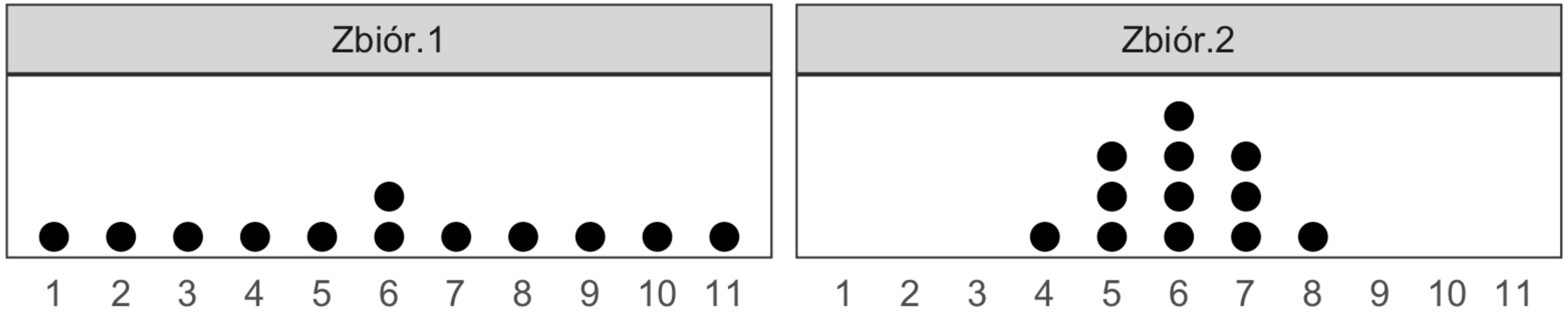
Me = 6, D = 6, średnia =6

Bartłomiej Karaban

Miary zmienności

Zbiór 1	1	2	3	4	5	6	6	7	8	9	10	11
Zbiór 2	4	5	5	5	6	6	6	6	7	7	8	8

Me = 6, D = 6, średnia =6



Bartłomiej Karaban

Miary zmienności

Rozstęp

Zbiór 1	1	2	3	4	5	6	6	7	8	9	10	11
Zbiór 2	4	5	5	5	6	6	6	6	7	7	8	8

Rozstęp (*range*) - różnica pomiędzy największą na najmniejszą zaobserwowaną wartością.

$$range = x_{max} - x_{min}$$

Zadanie:

1. Oblicz rozstępy dla powyższych danych. Co możesz powiedzieć o zmienności obu zbiorów względem siebie, zakładając, że dane przedstawiają pomiar tej samej cechy?

2. Utwórz wektory tych danych w R i oblicz rozstępy za jego pomocą.

3. Zdefiniuj własną funkcję `‚range’`

Bartłomiej Karaban

Miary zmienności

Rozstęp międzykwartylowy

Zbiór 1	1	2	3	4	5	6	6	7	8	9	10	11	Q1 = 3.75, Q3 =8.25
Zbiór 2	4	5	5	5	6	6	6	6	7	7	8	8	Q1 = 5, Q3 =7

Rozstęp międzykwartylowy (*Interquartile range, IQR)* - różnica pomiędzy wartością pierwszego i trzeciego kwartyla.

$$IQR \quad = \quad Q_3 - Q_1$$

Zadanie:

1. Oblicz *IQR* dla powyższych danych. Co możesz powiedzieć o zmienności obu zbiorów względem siebie, zakładając, że dane przedstawiają pomiar tej samej cechy?
2. Utwórz wektory tych danych w R i oblicz *IQR* za jego pomocą.

Bartłomiej Karaban

Miary zmienności

Wariancja

Wariancją ***(variance)*** w zbiorze wyników obserwacji nazywamy przeciętne kwadratowe odchylenie poszczególnych wyników obserwacji od ich średniej.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Bartłomiej Karaban

Miary zmienności

Wariancja

	$(x_i - \mu)$	$(x_i - \mu)^2$
1	1 - 6 = -5	25
2	2 - 6 = -4	16
3	3 - 6 = -3	9
4	4 - 6 = -2	4
5	5 - 6 = -1	1
6	6 - 6 = 0	0
6	6 - 6 = 0	0
7	7 - 6 = 1	1
8	8 - 6 = 2	4
9	9 - 6 = 3	9
10	10 - 6 = 4	16
11	11 - 6 = 5	25
suma	**0**	**110**

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{110}{12}$$

$$\sigma^2 \approx 9.17$$

Bartłomiej Karaban

Miary zmienności

Odchylenie standardowe

Odchyleniem standardowym (*standard deviation*) w zbiorze wyników nazywamy pierwiastek kwadratowy z wariancji.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Odchylenie standardowe dla poprzedniego przykładu wynosi zatem:

$\sigma \approx 3.03$

Bartłomiej Karaban

Miary zmienności

Wariancja, odchylenie standardowe

	$(x_i - \mu)$	$(x_i - \mu)^2$
4		
5		
5		
5		
6		
6		
6		
6		
7		
7		
8		
8		
suma		

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Zrób to sam:

- Oblicz wariancję i odchylenie standardowe
- Uzyskane wyniki porównaj z wynikami uzyskanymi w R za pomocą funkcji `var()` i `sd()`
- Napisz własne funkcje wariancji i odchylenia standardowego ze zmodyfikowanym licznikiem wariancji (n-1) i umieść je na githubie.
- Dokonaj porównania szybkości działania napisanych przez Ciebie funkcji z bazowymi funkcjami R i funkcjami napisanymi przez pozostałych uczestników warsztatu korzystając z pakietu `microbenchmark`.

Bartłomiej Karaban

Miary zmienności

Wariancja, odchylenie standardowe

Kilka uwag na temat wariancji i odchylenia standardowego:

- Obie te miary są tzw. bezwzględnymi miarami zmienności przez co można ich używać tylko i wyłącznie do porównywania zmienności pomiędzy cechami zmierzonymi w tych samych jednostkach.

  Przykład:

  Możemy porównać odchylenia standardowe zarobków brutto w Polsce dla kobiet i dla mężczyzn (jednostka PLN)

  Nie możemy porównać odchyleń standardowych zarobków brutto kobiet w Polsce i w Niemczech (PLN i EUR)

- Wariancja nie jest interpretowalna wprost. Tzn. Jeżeli obliczona warianacja zarobków wynosi 144, należy pamiętać, że jest to przeciętne kwadratowe odchylenie od średniej - ciężko jest mówić o PLN podniesionych do kwadratu, między innymi dlatego najczęściej używa się odchylenia standardowego.

- W przypadku gdy chcemy porównać zmienność pomiędzy cechami reprezentowanymi w różnych jednostkach właściwe są tzw. względne miary zmienności.

Miary zmienności

Współczynnik zmienności

Współczynnikiem zmienności (*coefficient of* variation) w zbiorze wyników obserwacji nazywamy stosunek ich odchylenia standardowego do średniej arytmetycznej. Wynik wyrażamy w %.

$$V_\sigma = \frac{\sigma}{\mu} \times 100$$

Zadanie:
Dla dwóch zbiorów z poprzedniego przykładu oblicz współczynnik zmienności.
Zaimplementuj jako funkcję w R i udostępnij na githubie.

Bartłomiej Karaban

CDN…