Statystyka opisowa

Cz. 2

Bartłomiej Karaban

Program

- Miary zmienności
- Miary asymetrii
- Miary koncentracji
- 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

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

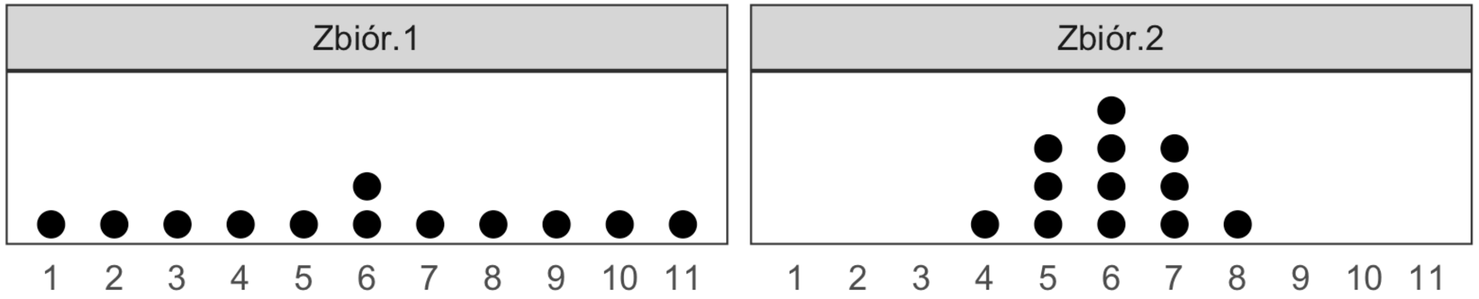
Me = 6, D = 6, średnia =6

Bartłomiej Karaban

Miary zmienności

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.

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

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.

$$IQR = Q_3 - Q_1$$

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.

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

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`.

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 %.

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

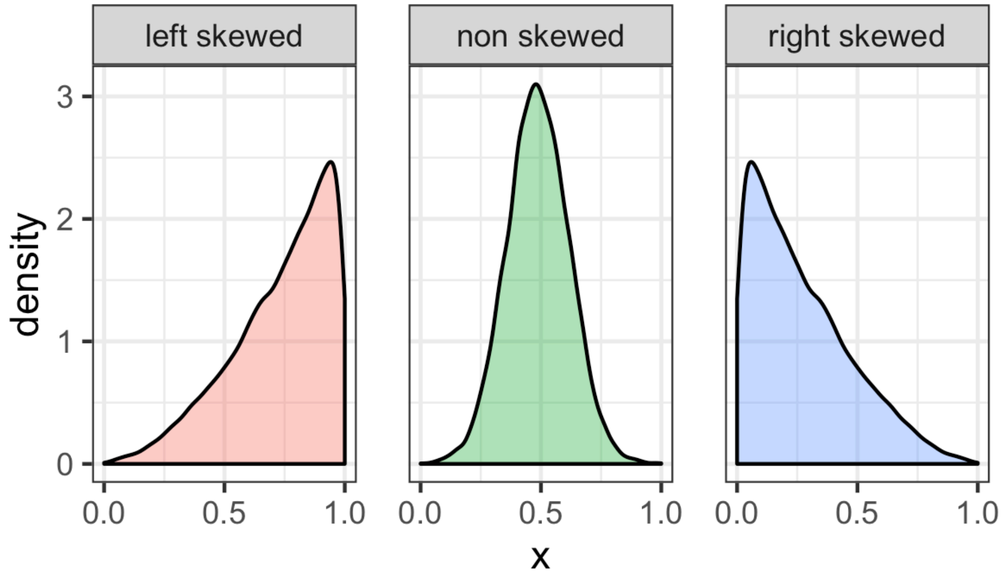
Zadanie:
Dla dwóch zbiorów z poprzedniego przykładu oblicz współczynnik zmienności.
Zaimplementuj jako funkcję w R i udostępnij na githubie.

Miary asymetrii

Skośność

***Skośność* (*skewness*)** jest miarą asymetrii rozkładu częstości

Jednym ze sposobów sprawdzenia asymetrii jest ocena wykresu rozkładu. Kierunek opadania „dłuższego ramienia” wskazuje czy rozkład jest prawo czy lewoskośny.

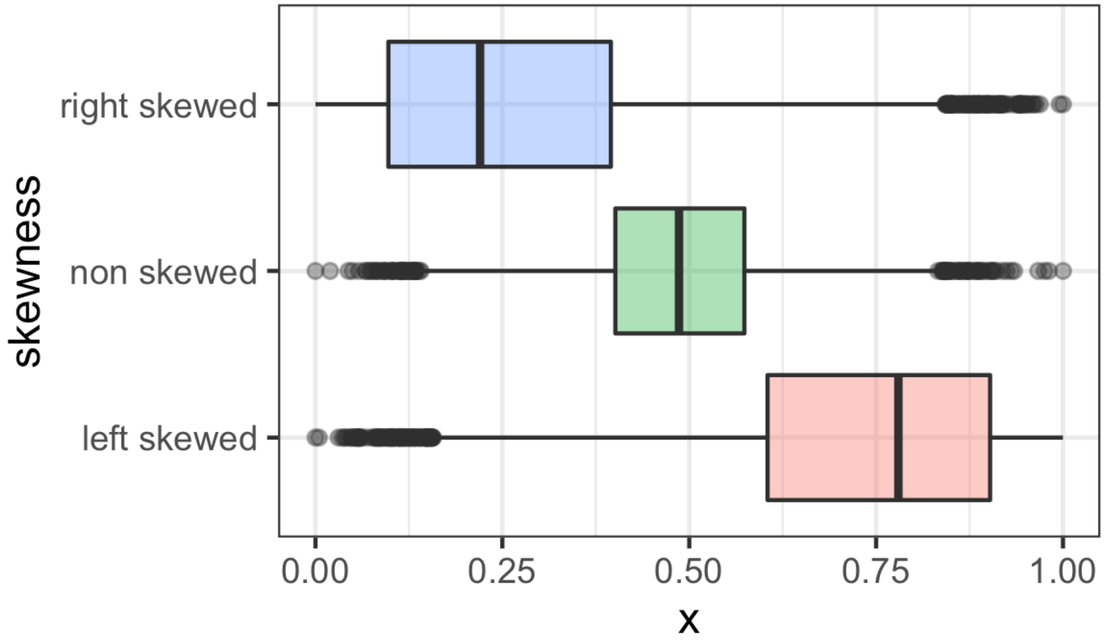


Bartłomiej Karaban

Miary asymetrii

Skośność

Kolejnym rodzajem wykresu, który pozwala na ocenę skośności jest wykres ramka-wąsy (*box-plot*).
Strona, po której znajduje się dłuższa część pudełka od mediany wyznacza kierunek skośności.



Bartłomiej Karaban

Miary asymetrii

Skośność

Określanie kierunku asymetrii na podstawie relacji pomiędzy miarami położenia:

$\mu = Me = D$ Rozkład symetryczny

$\mu \geq Me \geq D$ Rozkład asymetryczny, prawoskośny

$\mu \leq Me \leq D$ Rozkład asymetryczny, lewoskośny

Bartłomiej Karaban

Miary asymetrii

współczynnik asymetrii

Określenie kierunku asymetrii na podstawie współczynników asymetrii:

I współczynnik asymetrii Pearsona: $$A = \frac{\mu - D}{\sigma}$$

II współczynnik asymetrii Pearsona: $$A = 3\frac{\mu - Me}{\sigma}$$

pozycyjny współczynnik asymetrii: $$A = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

Interpretacja:

$A \approx 0$	Rozkład symetryczny
$A \geq 0$	Rozkład asymetryczny, prawoskośny
$A \leq 0$	Rozkład asymetryczny, lewoskośny

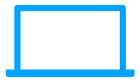
Wyniki z powyższych wzorów mogą zwracać wynik z różnym znakiem dla takich samych danych.

Bartłomiej Karaban

Miary asymetrii

współczynnik asymetrii

Czy poznane współczynniki asymetrii mogą być używane do porównania między sobą skośności rozkładów, które reprezentują pomiary cech w różnych jednostkach lub rzędach wielkości? (np.porównanie skośności pomiarów IQ z pomiarami wagi)



- W repozytorium ds_workshop w istniejącym pliku `excercises/014_StatystykaOpisowaSkosnosc.R` zaimplementuj funkcje obliczające II współczynnik asymetrii Pearsona i pozycyjny współczynnik asymetrii.
- Załaduj ramkę danych `skewed_data` z pliku `data/012_skewed_data.Rdata` używając funkcji `load()` i zapoznaj się z danymi.
- Sprawdź czy w przypadku obydwu współczynników wynik ma tez sam znak.
- Sprawdź czy zmiana rzędu wielkości zmiennej `skewed_data$x` wpłynie w jakikolwiek sposób na wynik.

Miary asymetrii

Współczynnik skośności

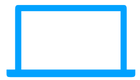
Kolejna miara asymetrii, ta oparta jest o tzw. trzeci moment centralny:

$$Skewness = \frac{M_3}{\sigma^3}$$

$$Skewness = \frac{\frac{\sum_{i=1}^{n}(x_i - \mu)^3}{N}}{\sigma^3}$$

Interpretacja:

$Skewness \approx 0$	Rozkład symetryczny
$Skewness \geq 0$	Rozkład asymetryczny, prawoskośny
$Skewness \leq 0$	Rozkład asymetryczny, lewoskośny



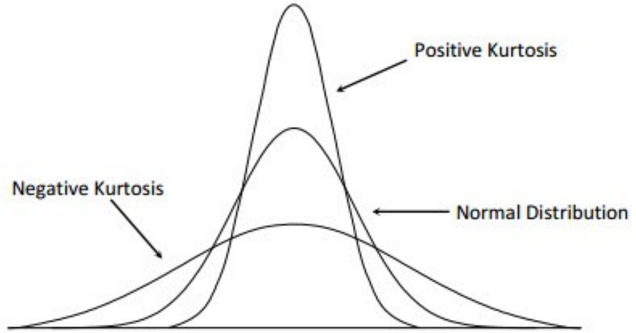
- W repozytorium ds_workshop w istniejącym pliku `excercises/014_StatystykaOpisowaSkosnosc.R` zaimplementuj funkcje współczynnika skośności i porównaj wynik z wynikiem funkcji `moments::skewness()`. Wykonaj benchmark Twojej funkcji i funkcji `skewness` na 10 000 powtórzeniach.

Bartłomiej Karaban

Miary koncentracji

Kurtoza

Kurtoza (kurthosis) opisuje jak bardzo dane skoncentrowane wokół średniej w porównaniu z rozkładem normalnym



$$K = \frac{M_4}{\sigma^4}$$

$$K = \frac{\frac{\sum_{i=1}^{n} (x_i - \mu)^4}{N}}{\sigma^4}$$

Interpretacja:

$K = 3$ Rozkład mezokurtyczny, normalny

$K \geq 3$ Rozkład leptokurtyczny, „szpiczasty”

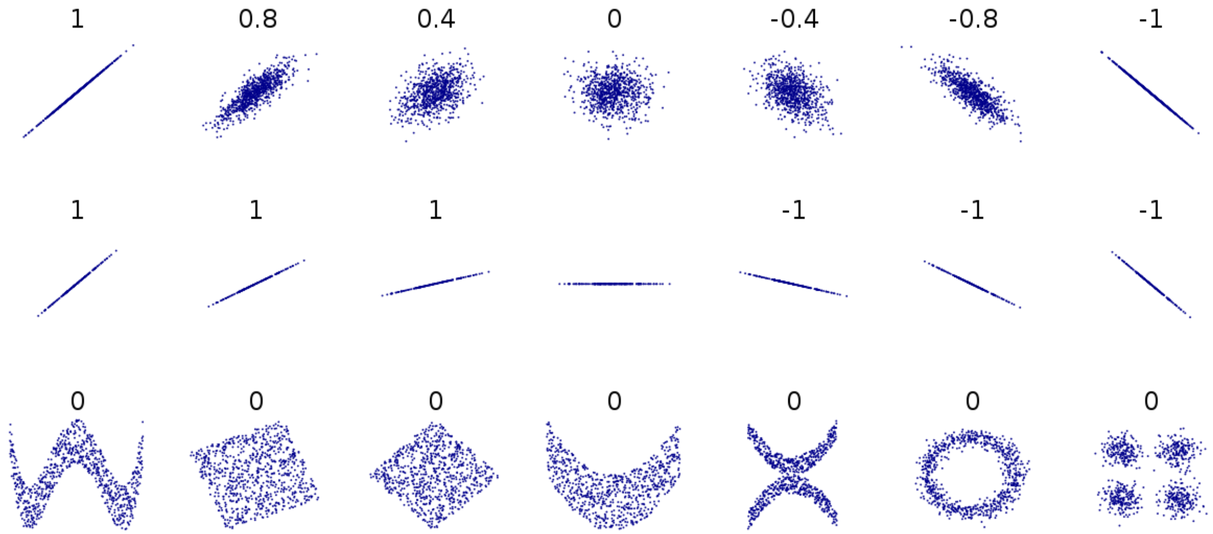
$K \leq 3$ Rozkład platokurtyczny, „płaski”

Język R posiada bibliotekę `moments`, gdzie za pomocą funkcji `kurthosis` można obliczyć kurtozę.

Bartłomiej Karaban

Korelacja

Korelacja (*corrrelation*) między dwoma losowymi zmiennymi X i Y jest miarą siły liniowego związku między tymi zmiennymi.



$\rho = 0$ Brak liniowego związku

$\rho = 1$ Ścisły dodatni związek, wraz ze wzrostem cechy X wrzasta cecha Y

$\rho = -1$ Ścisły ujemny związek, wraz ze wzrostem cechy X spada cecha Y

Korelacja nie jest tożsama z przyczynowością, tzn. Nie możemy interpretować, że wzrost lub spadek jednej cechy powoduje określone zachowanie drugiej.

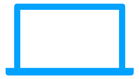
Bartłomiej Karaban

Korelacja

Korelacja (*corrrelation*) między dwoma losowymi zmiennymi X i Y jest miarą siły liniowego związku między tymi zmiennymi.

$$cov(X, Y) = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right) \left(y_i - \bar{y}\right)}{N}$$

$$\rho = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$



- W repozytorium ds_workshop utwórz plik `excercises/014_StatystykaOpisowaKorelacja.R` i za pomocą funkcji bazowej zbadaj korelację dowolnych dwóch zmiennych ze zbioru `iris`.
- Zbadaj korelację zmiennych „każda z każdą”. Wynik przedstaw w postaci macierzy korelacji.
- Zwizualizuj uzyskaną macierz korelacji za pomocą pakietu `corrplot`.

Dziękuję za uwagę