Statystyka Opisowa

cz. 1

Bartłomiej Karaban

Program

- Idea statystyki. Podstawowe pojęcia
- Statystyka opisowa i matematyczna. Przykłady
- Skale pomiarowe
- Miary położenia

„Lepiej znać prawdę niedokładnie niż dokładnie się mylić.”

J. M. Keynes

Idea statystyki. Podstawowe pojęcia.

Statystyka - nauka o zbieraniu, przetwarzaniu i analizie danych o zjawiskach masowych.

Statystyka opisowa- dziedzina statystyki zajmująca się analizą danych bez użycia rachunku prawdopodobieństwa. Przedstawia wyniki w sposób uporządkowany, jasny i prosty.

Statystyka matematyczna (wnioskowanie statystyczne) - dziedzina statystyki wykorzystująca rachunek prawdopodobieństwa do wnioskowania na podstawie zredukowanej liczby danych (próby). Pozwala określić jaki błąd popełniamy uogólniając wyniki z próby na populację.

Idea statystyki. Podstawowe pojęcia.

Populacja generalna- Zbiór wszystkich jednostek podlegających badaniu

Próba- Podzbiór jednostek z populacji generalnej. Najczęściej wybierany w sposób losowy.

Statystyka opisowa. Przykład.

German Credit Data

	Age	Sex	Job	Housing	Saving.accounts	Checking.account	Credit.amount	Duration	Purpose	Risk
844	50	male	2	own	little	NA	1559	24	business	good
13	22	female	2	own	little	moderate	1567	12	radio/TV	good
104	35	male	2	rent	little	moderate	1919	9	furniture/equipment	good
51	29	male	1	own	NA	moderate	2333	24	furniture/equipment	good
922	37	male	3	own	quite rich	NA	12749	48	radio/TV	good
721	34	male	3	own	little	rich	1337	9	radio/TV	bad
836	48	male	2	own	little	little	1082	12	car	bad
228	53	male	3	free	little	little	7865	12	furniture/equipment	bad
538	37	female	2	own	little	moderate	3612	18	furniture/equipment	good
527	31	female	2	own	moderate	NA	1532	15	education	good

https://raw.githubusercontent.com/STWUR/eRementarz-06-04-2019/master/dane/german_credit_data.csv

Populacja: 1000 wniosków kredytowych (każdemu wierszowi tabeli odpowiada 1 udzielony kredyt)

Próba: 10 Wybranych losowo z populacji wniosków kredytowych.

Cecha: każdy atrybut opisujący udzielony kredyt (każdej kolumnie odpowiada 1 cecha)

Statystyka opisowa. Przykład.

German Credit Data

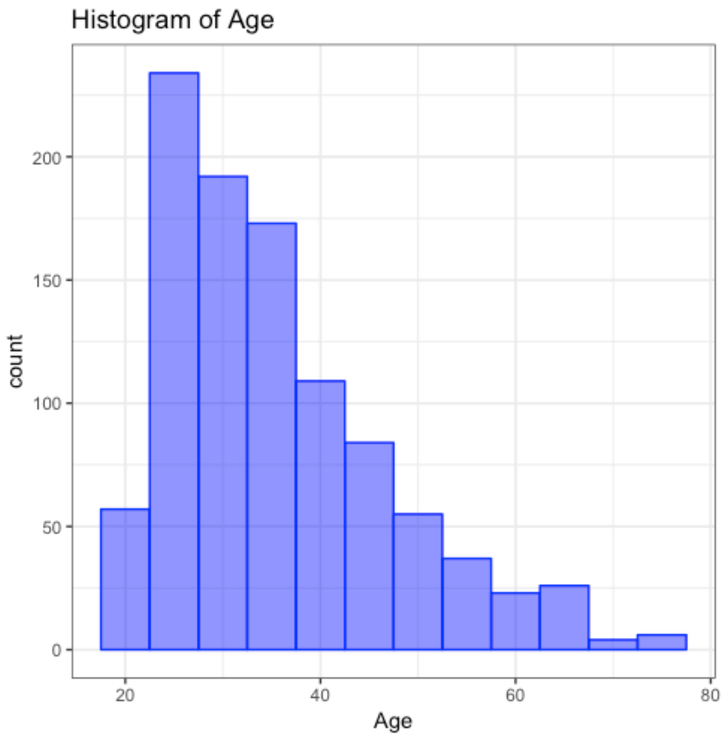
	Age	Sex	Job	Housing	Saving.accounts	Checking.account	Credit.amount	Duration	Purpose	Risk
844	50	male	2	own	little	NA	1559	24	business	good
13	22	female	2	own	little	moderate	1567	12	radio/TV	good
104	35	male	2	rent	little	moderate	1919	9	furniture/equipment	good
51	29	male	1	own	NA	moderate	2333	24	furniture/equipment	good
922	37	male	3	own	quite rich	NA	12749	48	radio/TV	good
721	34	male	3	own	little	rich	1337	9	radio/TV	bad
836	48	male	2	own	little	little	1082	12	car	bad
228	53	male	3	free	little	little	7865	12	furniture/equipment	bad
538	37	female	2	own	little	moderate	3612	18	furniture/equipment	good
527	31	female	2	own	moderate	NA	1532	15	education	good

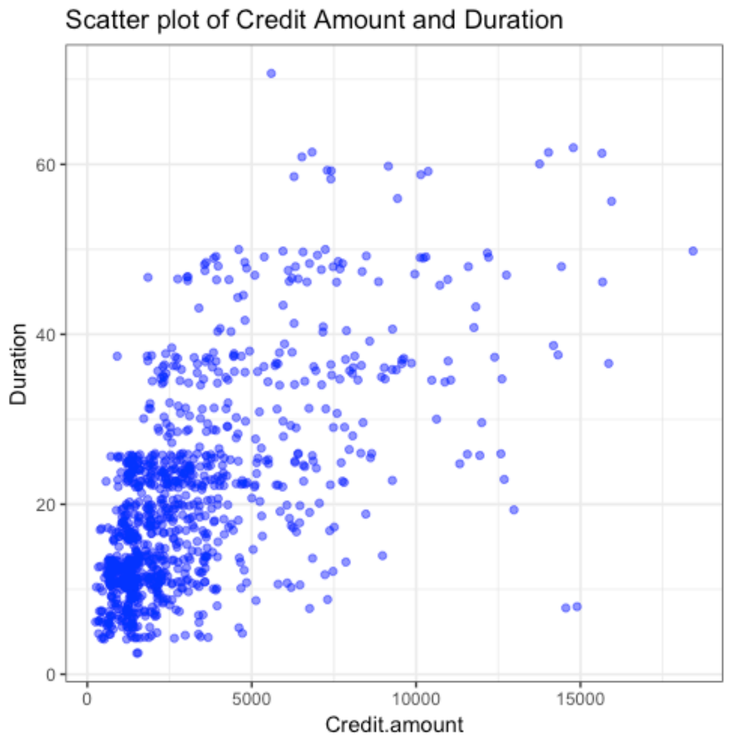
Miary pozycyjne cechy Age (populacja)

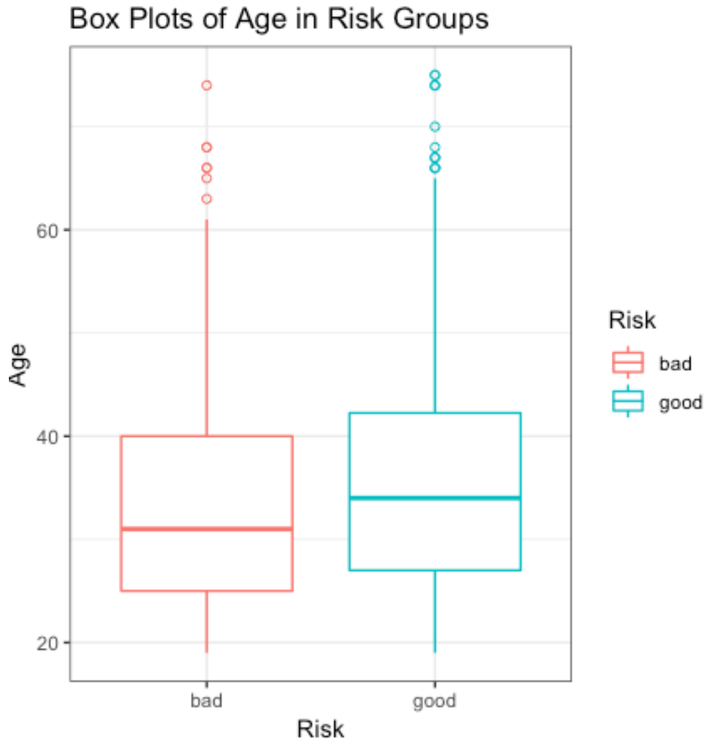
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.00	27.00	33.00	35.55	42.00	75.00

Bartłomiej Karaban

Statystyka opisowa. Przykład.

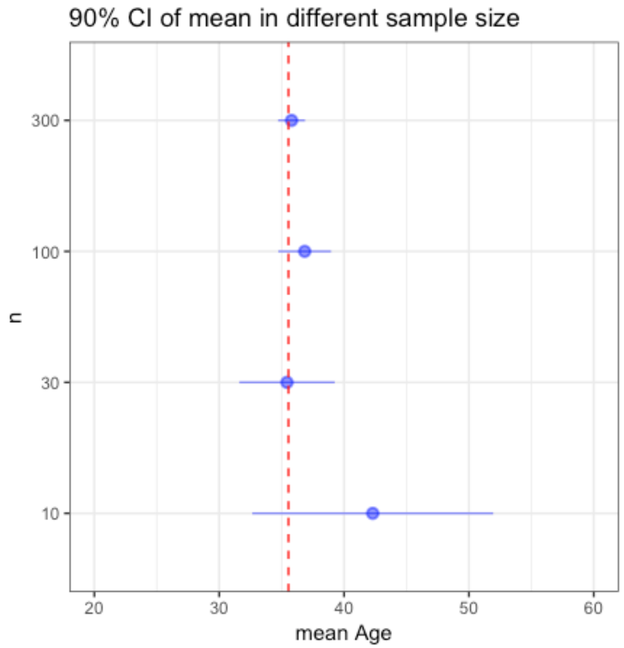


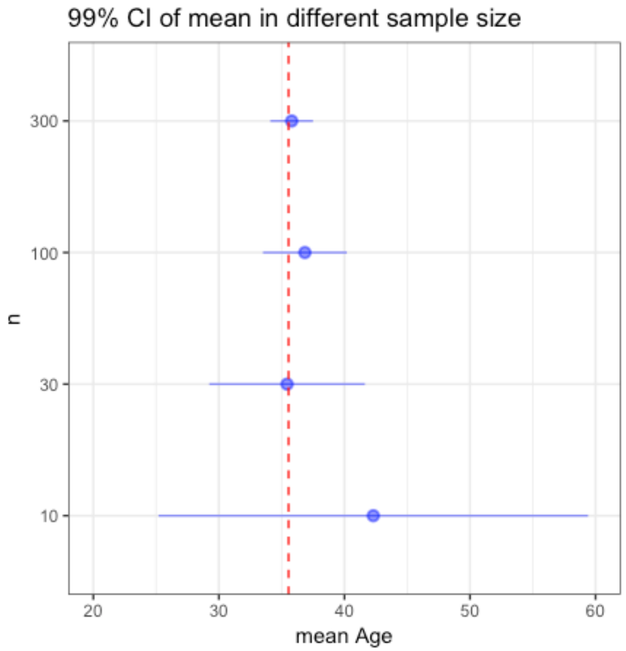




Statystyka matematyczna. Przykład.

- Wnioskujemy o parametrze populacji na podstawie wylosowanej próbki.
- Interesuje jaki jest średni wiek w całej populacji na podstawie wylosowanych przypadków.
- Z góry zakładamy na ile chcemy być pewni naszego oszacowania.





Statystyka matematyczna. Dyskusja.

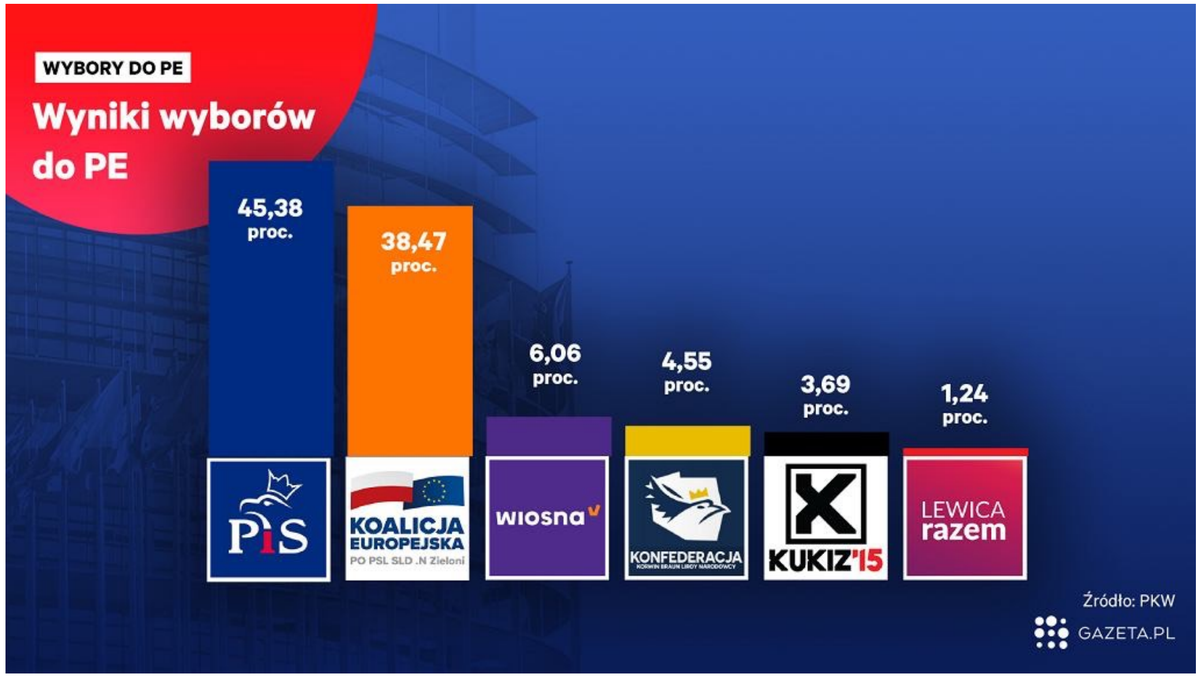
- Wyniki wyborów Exit Poll.
- Próba około 1000 respondentów.
- Szacowany błąd wynosi 4 % przy zadanym poziomie ufności.

1. KW PRAWO I SPRAWIEDLIWOŚĆ - 42.4%
2. KKW KOALICJA EUROPEJSKA PO PSL SLD .N ZIELONI - 39.1%
3. KW WIOSNA ROBERTA BIEDRONIA - 6.6%
4. KWW KONFEDERACJA KORWIN BRAUN LIROY NARODOWCY - 6.1%
5. KWW KUKIZ'15 - 4.1%
6. KW LEWICA RAZEM - 1,3%
7. INNE - 0,4%

- Patrząc na powyższe wyniki oszacuj zakres błędu dla ugrupowań od 1 do 5 miejsca.
- Co należało by zrobić żeby zmniejszyć zakres błędu przy tym samym poziomie ufności?
- Jak zmieniłby się zakres błędu badania przy tej samej liczności próby lecz mniejszym poziomie ufności?

Statystyka matematyczna. Dyskusja.

1. KW PRAWO I SPRAWIEDLIWOŚĆ - 42.4%
2. KKW KOALICJA EUROPEJSKA PO PSL SLD .N ZIELONI - 39.1%
3. KW WIOSNA ROBERTA BIEDRONIA - 6.6%
4. KWW KONFEDERACJA KORWIN BRAUN LIROY NARODOWCY - 6.1%
5. KWW KUKIZ'15 - 4.1%
6. KW LEWICA RAZEM - 1,3%
7. INNE - 0,4%



Bartłomiej Karaban

Skale pomiarowe

1. Skala nominalna
2. Skala porządkowa
3. Skala przedziałowa
4. Skala ilorazowa

Skale pomiarowe. Skala nominalna

W skali nominalnej obiekty są opisane za pomocą klas, przy czym klasom nie przypisuje się rangi. Np kolor oczu (zielony, piwny, niebieski). Jeżeli w tej skali występują cyfry to są traktowane jako etykiety dla klas i tylko zastępują nazwę klasy (np. 1 -zielony, 2 - piwny, 3 - niebieski)

	Age	Sex	Job	Housing	Saving.accounts	Checking.account	Credit.amount	Duration	Purpose	Risk
844	50	male	2	own	little	NA	1559	24	business	good
13	22	female	2	own	little	moderate	1567	12	radio/TV	good
104	35	male	2	rent	little	moderate	1919	9	furniture/equipment	good
51	29	male	1	own	NA	moderate	2333	24	furniture/equipment	good
922	37	male	3	own	quite rich	NA	12749	48	radio/TV	good
721	34	male	3	own	little	rich	1337	9	radio/TV	bad
836	48	male	2	own	little	little	1082	12	car	bad
228	53	male	3	free	little	little	7865	12	furniture/equipment	bad
538	37	female	2	own	little	moderate	3612	18	furniture/equipment	good
527	31	female	2	own	moderate	NA	1532	15	education	good

Zadanie: wskaż miary w powyższej tabeli, które nie zostały wskazane jako nominalne a nimi są.

Bartłomiej Karaban

Skale pomiarowe. Skala porządkowa

W skali porządkowej obiekty także są opisane za pomocą klas, jednak klasom przypisuje się rangi. Np. ryzyko - niskie, średnie wysokie . Tak jak w skali nominalnej klasy można kodować za pomocą cyfr, lecz tutaj liczba oznacza rangę. W tej skali nie dowiadujemy się o ile jeden obiekt jest lepszy od drugiego, lecz jedynie tyle, że jest lepszy.

	Age	Sex	Job	Housing	Saving.accounts	Checking.account	Credit.amount	Duration	Purpose	Risk
844	50	male	2	own	little	NA	1559	24	business	good
13	22	female	2	own	little	moderate	1567	12	radio/TV	good
104	35	male	2	rent	little	moderate	1919	9	furniture/equipment	good
51	29	male	1	own	NA	moderate	2333	24	furniture/equipment	good
922	37	male	3	own	quite rich	NA	12749	48	radio/TV	good
721	34	male	3	own	little	rich	1337	9	radio/TV	bad
836	48	male	2	own	little	little	1082	12	car	bad
228	53	male	3	free	little	little	7865	12	furniture/equipment	bad
538	37	female	2	own	little	moderate	3612	18	furniture/equipment	good
527	31	female	2	own	moderate	NA	1532	15	education	good

Zadanie: wskaż miary w powyższej tabeli, które nie zostały wskazane jako nominalne a nimi są.

Bartłomiej Karaban

Skale pomiarowe. Skala przedziałowa i ilorazowa

W skali przedziałowej i ilorazowej obiekty są opisane za pomocą cyfr (cechy ilościowe), które reprezentują wyniki pomiaru i można wykonywać na nich operacje matematyczne. Różnica pomiędzy skalami tkwi w zakresie operacji matematycznych jaki można na skalach wykonywać.

	Age	Sex	Job	Housing	Saving.accounts	Checking.account	Credit.amount	Duration	Purpose	Risk
844	50	male	2	own	little	NA	1559	24	business	good
13	22	female	2	own	little	moderate	1567	12	radio/TV	good
104	35	male	2	rent	little	moderate	1919	9	furniture/equipment	good
51	29	male	1	own	NA	moderate	2333	24	furniture/equipment	good
922	37	male	3	own	quite rich	NA	12749	48	radio/TV	good
721	34	male	3	own	little	rich	1337	9	radio/TV	bad
836	48	male	2	own	little	little	1082	12	car	bad
228	53	male	3	free	little	little	7865	12	furniture/equipment	bad
538	37	female	2	own	little	moderate	3612	18	furniture/equipment	good
527	31	female	2	own	moderate	NA	1532	15	education	good

Zadanie: Zaproponuj przekształcenie dowolnej cechy ilościowej na skalę porządkową.

Case Study: Przedziały score'owe w analizie ryzyka kredytowego.

Bartłomiej Karaban

Miary położenia

1. Średnia arytmetyczna.
2. Dominanta (moda).
3. Kwartyle i percentyle

Miary położenia. Średnia arytmetyczna

Średnią zbioru wyników obserwacji, zwaną także **przeciętną**, jest suma wartości wszystkich wyników obserwacji przez liczbę elementów tego zbioru.

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

Gdzie $\sum$ jest znakiem sumowania. Sumowanie rozciąga się wszystkie wyniki obserwacji.

Zadanie:

1. oblicz średnią arytmetyczną dla cechy Credit.amount z próbki zbioru German Credit Data (poprzednie slajdy).
2. Dodaj do próbki kolejną wartość: 100 000. Jak zmieni się wartość średniej?

Dyskusja: ostrożnie ze średnią

Miary położenia. Dominanta

Dominantą (modą) w zbiorze danych jest ta wartość, która w tym zbiorze występuje najczęściej.

wartości	6	9	10	12	13	14	15	16	17	18	19	20	21	22	24
częstości	1	1	1	1	1	2	1	3	2	2	2	1	1	1	1

Zadania:

1. Na podstawie danych z poprzedniego zadania zastanów się, na których cechach jest sens liczyć dominantę

2.* Zaimplementuj w R funkcję **get_mode(x)**, która jako argument przyjmie wektor wartości i zwróci dominantę z tego wektora. Podpowiedź: do rozwiązania użyteczne będą funkcje: **table()** i **which.max()**. Przykład oczekiwanego wywołania i wyniku:

```
> get_mode(credits$Job)
  2
630
```

Bartłomiej Karaban

Miary położenia. Percentyle

P - tym **percentylem** w zbiorze liczb (uporządkowanych według wielkości) jest taka wartość, poniżej której znajduje się *P%* liczb z tego zbioru. Miejsce *P*-tego percentyla określa wzór :

$$(n+1)P/100$$

Przykład 1:

W danym dniu zebrano dane sprzedażowe o wartości sprzedaży dla każdego z 20 sprzedawców:

6, 9, 10, 12, 13, 14, 14, 15, 16, 16, 16, 17, 17, 18, 18, 19, 20, 21, 22, 24

Dane zostały już uporządkowane.

Znajdźmy 50-ty percentyl:

1. Korzystając ze wzoru musimy znaleźć wynik obserwacji na miejscu:

(20 + 1)*(50/100) = 21*0.5 = 10.5

2. Licząc dane od najmniejszej do największej okazuje się, że 10-ta obserwacja przyjmuje wartość 16 i że taka sama jest wartość obserwacji 11.

Dlatego wynikiem obserwacji który zajmowałby w kolejności miejsce 10.5 jest liczba 16 (połowa „drogi” pomiędzy obserwacją 10 i 11)

3. 50 percentyl jest równy 16

Miary położenia. Percentyle

P - tym **percentylem** w zbiorze liczb (uporządkowanych według wielkości) jest taka wartość, poniżej której znajduje się *P%* liczb z tego zbioru. Miejsce *P*-tego percentyla określa wzór :

$$(n+1)P/100$$

Przykład 2:

W danym dniu zebrano dane sprzedażowe o wartości sprzedaży dla każdego z 20 sprzedawców:

6, 9, 10, 12, 13, 14, 14, 15, 16, 16, 16, 17, 17, 18, 18, 19, 20, 21, 22, 24

Dane zostały już uporządkowane.

Znajdźmy 80-ty percentyl:

1. Korzystając ze wzoru musimy znaleźć wynik obserwacji na miejscu:

(20 + 1)*(80/100) = 21*0.8 = 16.8

2. Licząc dane od najmniejszej do największej okazuje się, że 16-ta obserwacja przyjmuje wartość 19 i a 17-ta jest równa 20.

Dlatego wynikiem obserwacji który zajmowałby w kolejności miejsce 16.8 jest liczba 19.8 (0.8 „drogi” pomiędzy obserwacją 16 i 17)

3. 80 percentyl jest równy 19.8

Miary położenia. Kwartyle - szczególne percetyle

Pierwszy kwartyl to 25-ty percentyl, czyli wartość poniżej której znajduje się jedna czwarta wyników obserwacji. Często oznaczany symbolem Q1

Drugi kwartyl to 50-ty percentyl - jest to najważniejszy kwartyl mający specjalną nazwę: **mediania.**

Trzeci kwartyl to 75-ty percentyl - czyli wartość poniżej której znajduje się trzy czwarte obserwacji. Często oznaczany symbolem Q3.

Rozstęp miedzykwartylowy (IQR) - *to różnica pomiędzy wartością Q3 i Q1. Jest to jedna z miar rozproszenia obserwacji*

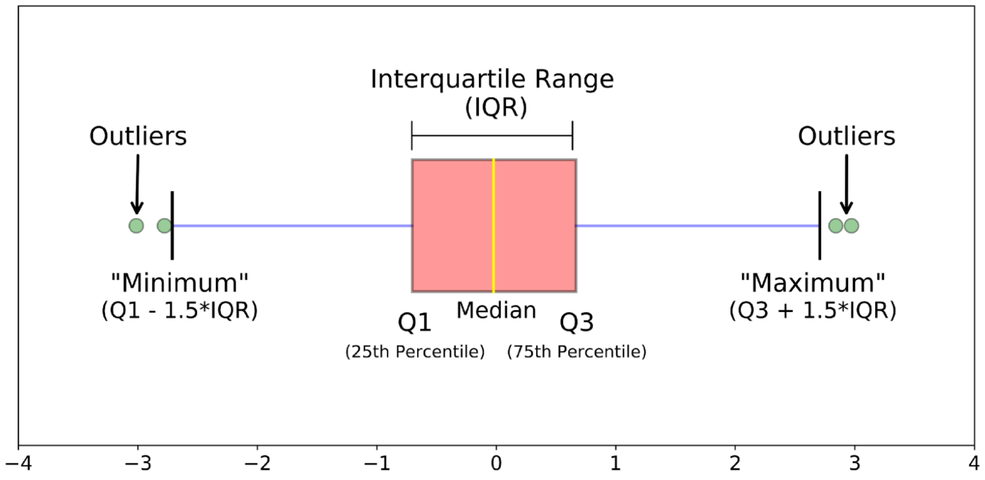
Zadanie:

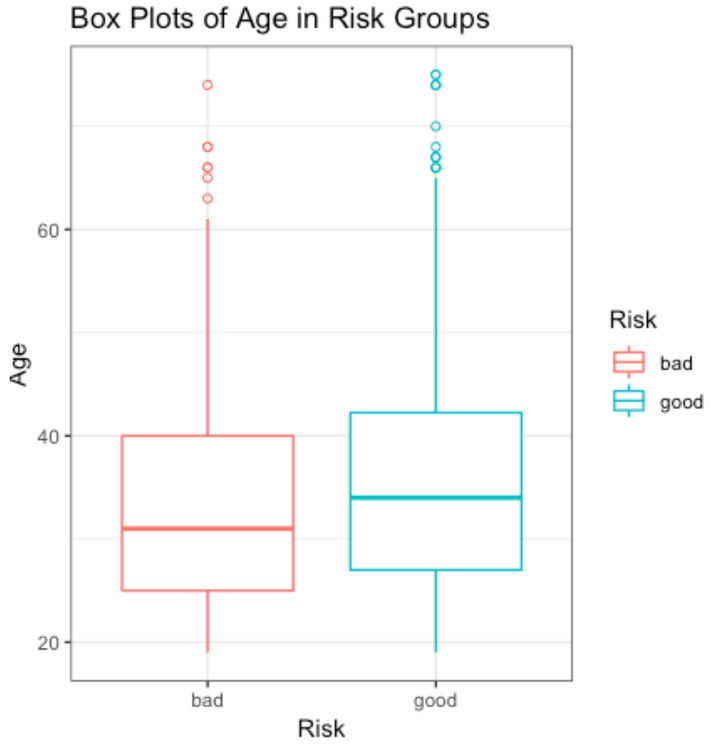
1. oblicz Q1, medianę, Q2 dla cechy Credit.amount z próbki zbioru German Credit Data (poprzednie slajdy).
2. Dodaj do próbki kolejną wartość: 100 000. Jak zmieni się wartość obliczonych kwartyli?

Wykres ramka-wąsy (boxplot)

Wykres-pudełko (ramka-wąsy, boxplot) jest obrazem następujących charakterystyk danych:

- mediana,
- dolny kwartyl (Q1),
- górny kwartyl (Q3),
- najmniejszy wynik obserwacji,
- największy wynik obserwacji.





Bartłomiej Karaban

Literatura

1. Aczel A., *Statystyka w zarządzaniu*, PWN, Warszawa 2000
2. Starzyńska W., *Statystyka praktyczna*, PWN, Warszawa 2005
3. Wawrzynek J., *Metody opisu i wnioskowania statystycznego*, Wydawnictwo AE we Wrocławiu, Wrocław 2007