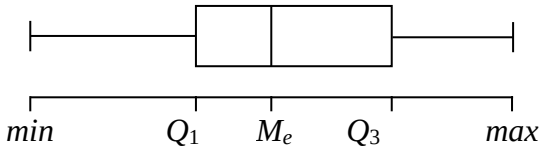
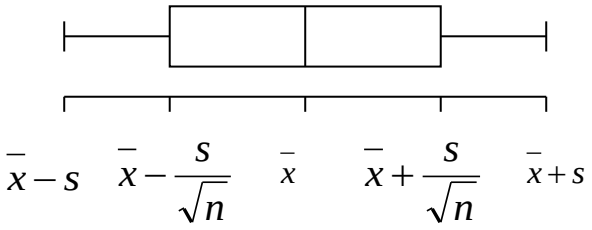
2. GRAFICZNA PREZENTACJA DANYCH

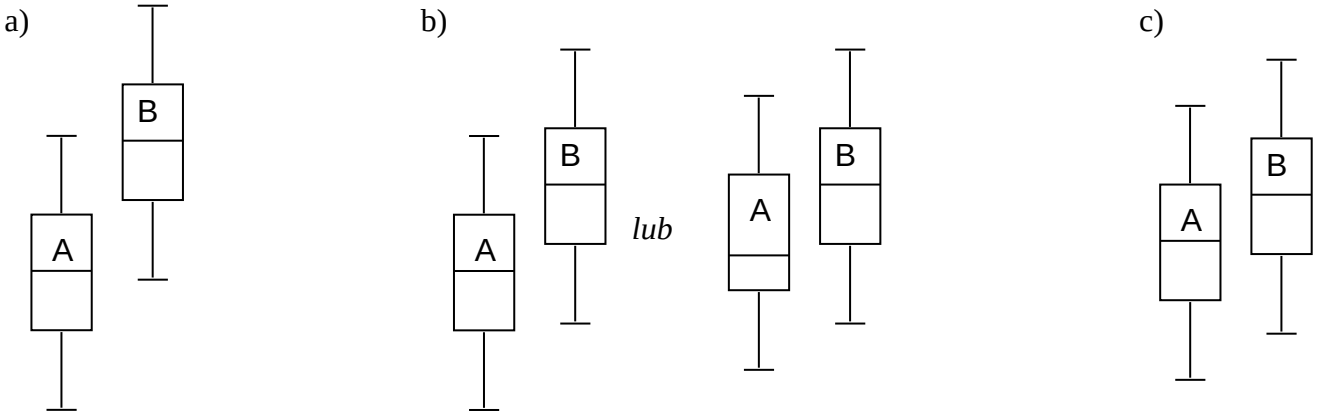
2.1. Wykres pudełkowy

Wykres pudełkowy – pozwala na ilustrację miar rozkładu jednowymiarowej zmiennej losowej. Jest rysowany w kilku odmianach:

Ilustrowane miary	***Wygląd***
min, max Q_1, Me, Q_3	
$\bar{x}$, s	

Wykresy pudełkowe są często rysowane dla wielu grup danych w celu ich porównania. Wyniki porównania można uogólnić na całe populacje jeśli rozmiary próbek na podstawie których narysowane były wykresy nie są mniejsze od 30.

Poniżej rozważone zostały typowe układy dwóch wykresów pudełkowych.



Rys. 1. Typowe układy dwóch wykresów pudełkowych

W przypadku:

a) prostokąty wykresów nie nachodzą na siebie – w tej sytuacji uznaje się że ***B jest większe od A***,

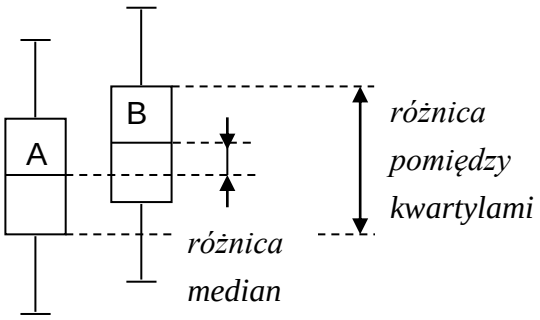
b) prostokąty nachodzą na siebie ale jedna z median jest poza prostokątem drugiego wykresu – w tej sytuacji wyciągany jest wniosek, że ***B jest prawdopodobnie większe od A***,

c) prostokąty nachodzą na siebie, obydwie mediany są wewnątrz prostokątów sąsiedniego wykresu – w tej ***sytuacji nie można stwierdzić, że istnieje różnica pomiędzy A i B***.

W przypadku c) można wyznaczyć iloraz różnicy median do różnicy pomiędzy dalej położonym dolnym i górnym kwartylem. Jeśli wielkość tak wyznaczonego wskaźnika przekracza pewną wartość graniczną uznaje się że jest ***tendencja do różnicy pomiędzy A i B***.

Za wartość graniczną uznaje się wielkość:

- 0,33 dla próbek o rozmiarze 30,
- 0,2 dla próbek o rozmiarze 100,
- 0,1 dla próbek o rozmiarze 1000.



Rys. 2. Istnienie „tendencji do różnicy”

2.2. Histogram

Histogram to wykres słupkowy ilustrujący rozkład prawdopodobieństwa określonej cechy, jest konstruowany na podstawie danych szeregu rozdzielczego. Podstawę słupków stanowią przedziały klasowe szeregu, przyjmuje się że liczba tych przedziałów powinna być równa pierwiastkowi z liczby pomiarów, zwykle rysowane są histogramy o 6 do 12 przedziałach. Wysokość słupków jest ustalana dla określonej wartości lub przedziału na podstawie:

- *liczebności* (histogram liczebności),
- *częstości* (histogram częstości),
- *skumulowanych liczebności* (histogram liczebności skumulowanych),
- *skumulowanych częstości* (histogram częstości skumulowanych).

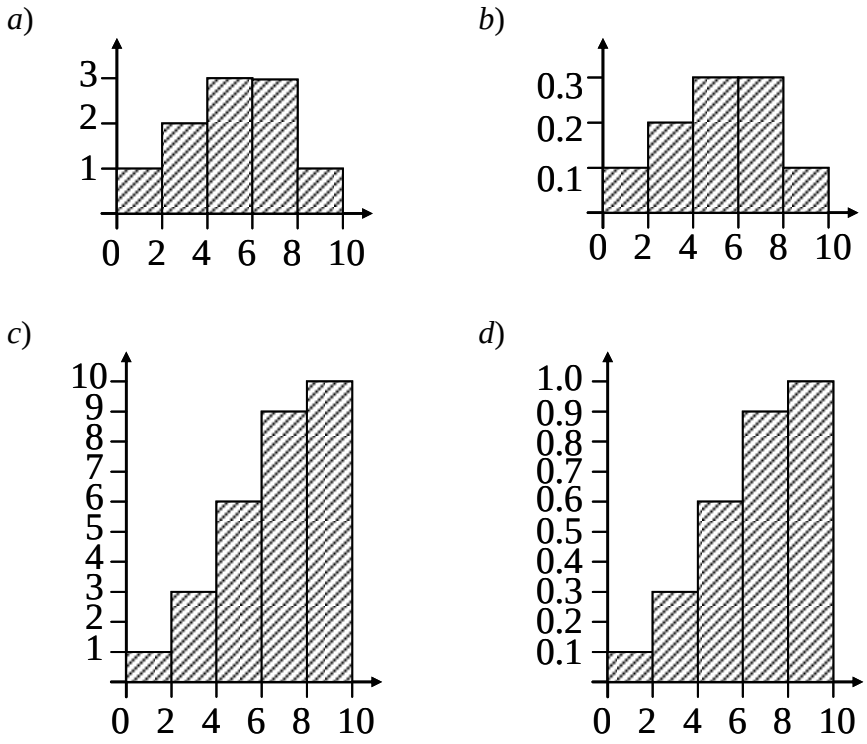
Przykład 1.

Na rysunkach przedstawione zostały histogramy przedstawiające rozkład wyników pomiarów długości pewnego detalu.

19.5	20	20.9	21.2	21.6	21.6	22	22	22.6	24

Histogramy zostały skonstruowane po utworzeniu poniższego szeregu rozdzielczego.

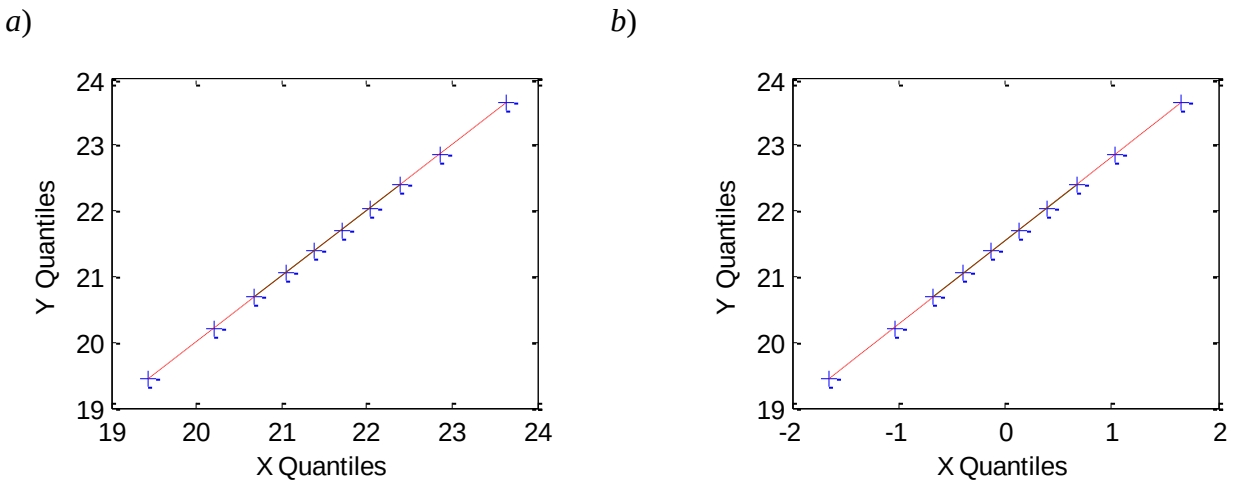
	[19,20)	[20,21)	[21,22)	[22,23)	[23,24]
liczebności	1	2	3	3	1
częstości	0.1	0.2	0.3	0.3	0.1
l. skumul.	1	3	6	9	10
c. skumul.	0.1	0.3	0.6	0.9	1.0



Rys.3. Histogramy a)liczebności, b) częstości, c) liczebności skumulowanych d) częstości skumulowanych.

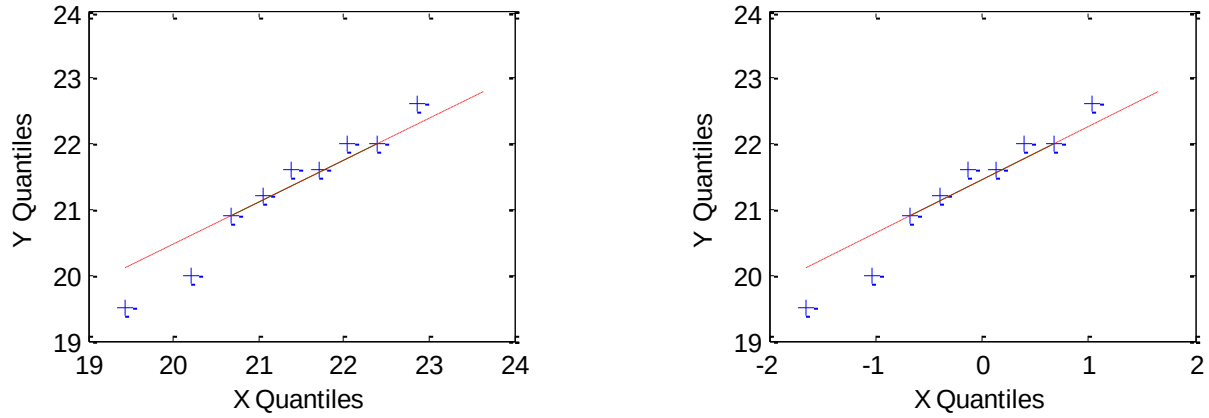
2.3. Wykres kwanty–kwantyl (wykres K-K lub Q-Q)

Wykres *Q–Q* umożliwia porównanie kwantyli dwóch zmiennych, jest często wykorzystywany do porównania rozkładu zmiennej empirycznej z teoretycznym rozkładem tej zmiennej. Na osi poziomej odkładane są kwantyle wyznaczone z rozkładu pierwszej zmiennej a na osi pionowej drugiej zmiennej (lub kwantyle wynikające z teoretycznego rozkładu badanej zmiennej).



Rys. 4. Wykres Q–Q porównujący a) rozkład kwantyli teoretycznych tego samego rozkładu normalnego, b) rozkład kwantyli teoretycznych dwóch różnych rozkładów normalnych

a) *b*)



Rys. 5. Wykres Q–Q porównujący a) rozkład kwantyli z próby z kwantylami teoretycznymi rozkładu o średniej i odchyleniu standardowym wyznaczonym na podstawie próby, b) rozkład kwantyli z próby z kwantylami teoretycznymi rozkładu normalnego standaryzowanego.

Jeżeli kwantyle obydwu rozkładów układają się na wykresie tworząc linię prostą to obydwa rozkłady są idealnie do siebie dopasowane. Idealne dopasowanie można uzyskać np. porównując ze sobą kwantyle tego samego rozkładu teoretycznego lub kwantyle teoretyczne dwóch różnych rozkładów normalnych.
W przypadku porównywania kwantyli z próby (empirycznych) z kwantylami teoretycznymi punkty na wykresie tym bardziej dopasowane są do linii prostej im lepiej rozkład empiryczny jest dopasowany do rozkładu teoretycznego. Wykres *Q–Q* jest bardzo często uzupełniany linią prostą pokazują liniowy trend danych.

Wyznaczanie kwantyla z próby

Kwantyl rzędu p *(0<p<1)*

wartość cechy x_p, która dzieli szereg na dwie części w taki sposób, że:

- w pierwszej części znajduje się 100*p* [%] elementów próbki (wartości tych elementów są mniejsze lub równe *kwantylowi* x_p),
- w drugiej części znajduje się 100(1–*p*) [%] elementów (wartości tych elementów są większe bądź równe *kwantylowi* x_p).

Powyższa definicja pozwala na pewną dowolność przy wyznaczaniu kwantyli. Załóżmy, że dany jest szereg prosty:

1	2	3	4	5

Definicja jednoznacznie określa kwantyl rzędu 0.5 (medianę) – jest on równy 3, wyznaczenie kwantyla rzędu np. 0.1 nie jest już jednoznaczne i stosowane mogą być różne algorytmy na wyznaczenie jego wartości.
Metoda Clevelanda polega na przyporządkowaniu kolejnym wartościom uporządkowanego rosnąco szeregu prostego prawdopodobieństw:

$$p_i = \frac{i - 0.5}{n},$$

gdzie: i – numer wartości w szeregu, n – wielkość próbki
Wartości o rzędach różnych od wyznaczonych są interpolowane.

Dla przykładowego szeregu prawdopodobieństwa, zgodnie z powyższym wzorem wynosiłyby: 0.1, 0.3, 0.5, 0.7 i 0.9 więc kwantyl rzędu 0.1 byłby równy 1.

Wyznaczanie kwantyla rozkładu teoretycznego

Kwantyle teoretyczne są zdefiniowane jednoznacznie, kwantyl rzędu p wyznacza się z zależności: $F^{-1}(p)$ – wyznaczana jest odwrotność dystrybuanty rozkładu teoretycznego dla prawdopodobieństwa równego rzędowi kwantyla.

2.4. Wykres prawdopodobieństwo-prawdopodobieństwo (wykres P–P)

Wykres *P–P* umożliwia porównanie skumulowanego prawdopodobieństwa rozkładu dwóch zmiennych, jest często wykorzystywany, podobnie jak wykres *Q–Q* do porównania rozkładu zmiennej empirycznej z teoretycznym rozkładem tej zmiennej. Na osi poziomej odkładane są skumulowane prawdopodobieństwa wyznaczone z próby, prawdopodobieństwa te wyznacza się po uporządkowaniu (rosnąco) danych z próby i wyznaczeniu dla każdej z wartości dystrybuanty rozkładu teoretycznego o parametrach wyznaczonych na podstawie próby. Na osi pionowej odkładane są skumulowane prawdopodobieństwa rozkładu teoretycznego, prawdopodobieństwa te wyznaczane są np. jako: $p_i = (i - 0.5)/n$. Podobnie jak w przypadku wykresu *Q–Q* punkty na wykresie *P–P* są tym bardziej dopasowane są do linii prostej im bardziej rozkłady są zgodne, wykres *P–P* jest często uzupełniany liniowym trendem ułatwiającym sprawdzenie zgodności rozkładów.

2.5. Wykresy normalności

Wykres porównujące rozkład danych empirycznych z teoretycznym rozkładem normalnym nazywane są *wykresami normalności*, jeżeli na wykresach *Q–Q* i *P–P* porównywany jest rozkład z próby z rozkładem normalnym wykresy te nazywane są *wykresami normalności Q–Q* i *wykresami normalności P–P*. Na *wykresach normalności Q–Q* oś pionowa, na której odkładane są kwantyle wynikające z teoretycznego rozkładu normalnego badanej zmiennej, może być również opisywane rzędami kwantyli.