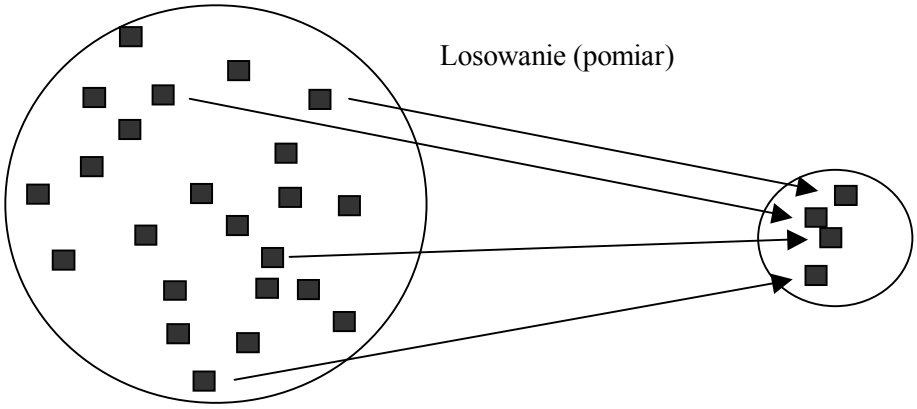
**STATYSTYKA OPISOWA**

Statystyka

Statystyka opisowa — Statystyka matematyczna




Populacja generalna (rezultaty potencjalnych pomiarów) — Próbka (rezultaty pomiarów)

Statystyka opisowa zajmuje się wstępnym opracowaniem wyników pomiarów (próbki) bez posługiwania się rachunkiem prawdopodobieństwa. Nie wyciągamy wniosków dotyczących populacji generalnej.
Niech $x_1, x_2, x_3,...x_n$ będzie próbką n-elementową. n – *liczność (liczebność)*. Parametry obliczone z próbki będą dalej nazywane statystykami.

## 1. Graficzne przedstawienie próbki: szereg rozdzielczy, histogram, łamana częstości

Rozstęp $R=x_{max}-x_{min}$

Klasy Dla próbek o dużej liczebności (n>30) elementy próbki grupuje się w klasach, tj. przedziałach o równej lub nierównej długości. Niech k oznacza ilość klas. Ile klas k przyjąć dla danej próbki? Można się kierować następującymi orientacyjnymi regułami:

$k \le 5 \lg(n)$ lub $k=1+3.32\ l8lg(n)$ lub $k=\sqrt{n}$

Zatem, gdy n=20, to k=4 ÷ 6, gdy n=40, to k=6 ÷ 8

Długość klasy $b \cong R/k$

Niech $n_i$ – liczność i-tej klasy, a $\bar{x}_i$ środek i-tej klasy. Wtedy pary liczb $(\bar{x}_i, n_i)$ nazywamy szeregiem rozdzielczym. Graficzne przedstawienie szeregu rozdzielczego nazywa się histogramem.
Na osi poziomej histogramu – środki klas lub granice poszczególnych klas, na osi pionowej histogramu – liczności klas, *częstości (frekwencje)* $w_i=n_i/n$, lub $v_i=w_i/b$. Łącząc punkty o współrzędnych $(\bar{x}_1 - b,0),(\bar{x}_i, v_i)$ dla i=1,...,k, $(\bar{x}_k + b,0)$ otrzymujemy tzw. łamaną częstości.

## 2. Statystyki lokacji rozkładu

Średnia arytmetyczna $\bar{x}$ liczb $x_1, x_2, x_3,...x_n$ określona jest wzorem

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

Charakterystyczna własność średniej arytmetycznej: suma wszystkich odchyleń jest równa zero; $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Średnia geometryczna $\overline{g}$ liczb dodatnich określona jest wzorem

$$\overline{g} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Średnia harmoniczna $\overline{h}$, różnych od zera liczb $x_1, x_2, x_3,...x_n$, nazywamy odwrotność średniej arytmetycznej odwrotności tych liczb

$$\overline{h} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

Mediana (wartość środkowa) $m_e$ – środkowa liczbę w *uporządkowanej* niemalejąco próbce (dla próbki o liczności nieparzystej) lub średnią arytmetyczną dwóch liczb środkowych (dla próbki o liczności parzystej).

Wartością modalną (modą, dominantą) $m_0$ próbki o powtarzających się wartościach nazywamy najczęściej powtarzającą się wartość, o ile istnieje, nie będącą $x_{min}$ ani $x_{max}$.

Jeżeli w szeregu rozdzielczym najliczniejsze są obie klasy skrajne, to szereg rozdzielczy nazywamy antymodalnym typu U, a środek najmniej licznej klasy antymodą. Gdy najliczniejsza jest jedna z klas skrajnych, to szereg rozdzielczy nazywamy antymodalnym typu J.

Rozkład dwumodalny – gdy występują dwie jednakowo liczne i najliczniejsze klasy nie będące skrajnymi.

Rozkład jednomodalny, dwuwierzchołkowy – występują dwie najliczniejsze klasy, ale nie są jednakowo liczne i nie są skrajnymi.

Kwantyl rzędu q ($0<q<1$) – taka wartość $x_q$, przed którą (tzn.dla $x \le x_q$) znajduje się 100q % elementów próbki. Gdy q=0.25, 0.5, 0.75, to takie kwantyle nazywamy kwartylami. Gdy q=0.25 mówimy o kwartylu dolnym, gdy q=0.75 mówimy o kwartylu górnym. Kwartyl q=0.5 jest medianą.

### 3. Statyki rozproszenia (rozrzutu, rozsiania) rozkładu

Rozstęp R;

Wariancja $s^2$ średnia arytmetyczna kwadratów odchyleń poszczególnych wartości $x_i$ od średniej arytmetycznej $\overline{x}$

$$s^2 = \frac{1}{n} \sum_{i-1}^{n} \left( x_i - \overline{x} \right)^2$$

Odchylenie standardowe $s = \sqrt{s^2}$

Odchylenie przeciętne $d_1$ od wartości średniej – średnia arytmetyczna wartości bezwzględnych odchyleń poszczególnych wartości $x_i$ od średniej arytmetycznej

$$d_1 = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{x} \right|$$

Odchylenie przeciętne $d_2$ od mediany – średnia arytmetyczna wartości bezwzględnych odchyleń poszczególnych wartości $x_i$ od mediany $m_e$

$$d_2 = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - m_e \right|$$

### 4. Statystyki kształtu rozkładu

Momentem zwykłym $m_l$ rzędu l próbki $x_1, x_2, x_3,...x_n$ nazywamy średnią arytmetyczną l-tych potęg wartości $x_i$

$$m_l = \frac{1}{n} \sum_{i=1}^{n} x_i^l$$

Zauważmy, że $m_1 = \overline{x}$

Momentem centralnym $M_l$ rzędu l próbki $x_1, x_2, x_3,...x_n$ nazywamy średnią arytmetyczną l-tych potęg odchyleń wartości $x_i$ od średniej arytmetycznej $\overline{x}$ próbki

$$M_l = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^l$$

Zauważmy, że $M_1=0$, $M_2=s^2$.

Współczynnik asymetrii (skośności) $g_1$

$$g_1 = \frac{M_3}{s^3}$$

gdzie s jest odchyleniem standardowym. Dla rozkładu normalnego $g_1=0$. Gdy rozkład ma długi „ogon" dla wartości większych od wartości średniej, to $g_1>0$, gdy „ogon" występuje po stronie wartości mniejszej niż średnia, to $g_1<0$.

Współczynnik koncentracji (skupienia), kurtoza K

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

gdzie s jest odchyleniem standardowym. Kurtoza ma wartość 3 dla rozkładu normalnego. Gdy K>3, to rozkład jest bardziej skupiony („szpiczasty") niż rozkład normalny, gdy K<3, to rozkład jest bardziej spłaszczony niż rozkład normalny.

Współczynnik spłaszczenia, eksces $g_2$

$$g_2=K-3$$

Dla rozkładu normalnego $g_2=0$.

Współczynnik zmienności ν

$$\nu = \frac{s}{\bar{x}}\cdot 100\%$$

gdzie s jest odchyleniem standardowym.

Współczynnik nierównomierności H

$$H = \frac{d_1}{\bar{x}}\cdot 100\%$$

gdzie $d_1$ jest odchyleniem przeciętnym od średniej arytmetycznej.

## 5. Graficzne przedstawienie próbki: prawdopodobieństwo skumulowane, wykres ramkowy

Zakładamy, że prawdopodobieństwo uzyskani każdego elementu próbki n elementowej jest równe 1/n. Uporządkujmy próbkę według wartości rosnących.

Prawdopodobieństwem skumulowanym (dystrybuantą empiryczną) p(x) dla danego x nazywamy prawdopodobieństwo otrzymania wartości mniejszej lub równej x: $p(x)=p(x_i \le x)$ w próbce uporządkowanej.

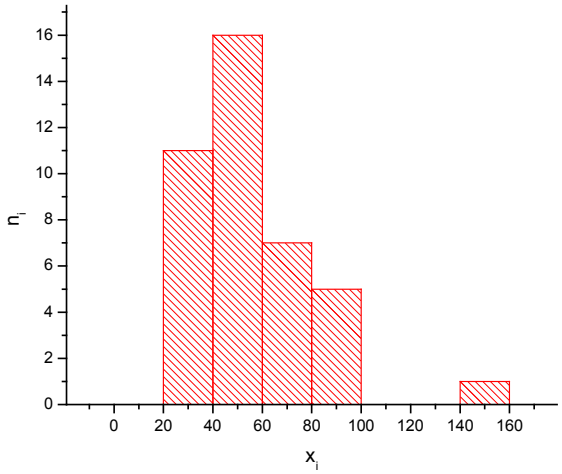
Jednym z wielu sposobów graficznej prezentacji próbki jest wykres ramkowy, potocznie nazywany 'pudełkiem z wąsami' (ang. box-and-whisker plot), zaproponowany w 1977 roku przez J.Tukey'a. Rysujemy najpierw prostokąt, którego dolny bok jest kwartylem dolnym, a górny bok kwartylem górnym. Pozioma linia dzieląca prostokąt to mediana. Wąsy powstają z połączenia powstałego pudełka z krótkimi liniami poziomymi, narysowanymi dla kwantyla q=0.95 (górny wąs) i kwantyla 0.05 (wąs dolny). Na rysunku zaznaczyć można także inne wartości kwantyli (np. 0.01 i 0.99), jak i inne statystyki próbki, np. wartość średnią, ekstremalne wartości w próbce, itp.

*PRZYKŁAD: Próbka 40. elementowa – utworzona za pomocą generatora liczb losowych, z rozkładu lognormalnego LND(4, 0.4) (Program MATHEMATICA)*

| | | | | |
|---|---|---|---|---|
| 48.4478 | 69.2368 | 21.6994 | 29.3819 | 65.3572 |
| 45.7823 | 55.4199 | 42.1859 | 47.8664 | 55.7535 |
| 87.1514 | 49.3306 | 37.5616 | 56.4771 | 26.8422 |
| 74.2661 | 51.3336 | 77.8302 | 40.1117 | 41.5877 |
| 55.8195 | 35.9834 | 67.6347 | 82.9544 | 42.1217 |
| 61.1744 | 35.7469 | 43.1695 | 48.9212 | 52.3768 |

```
63.7887     39.5142     153.613     98.6516     86.1010
30.4353     34.3459     39.4973     21.1369     91.6702
```

n=40, $x_{min}$=21.1369, $x_{max}$=153.613, R=132.476




Rys. 1. Histogram próbki. Zaznaczono granice klas (na osi x) i ilość elementów w klasie (na osi y)

Statystyki lokacji rozkładu:

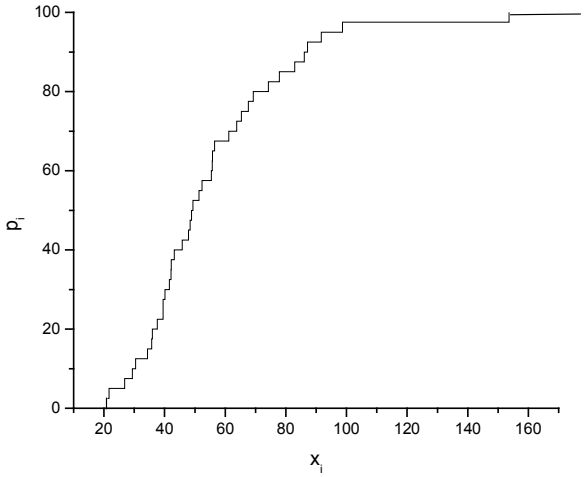
| | |
|---|---|
| średnia arytmetyczna | $\overline{x}$ =55.2071 |
| średnia geometryczna | $\overline{g}$ =50.5966 |
| średnia harmoniczna | $\overline{h}$ =46.5614 |
| mediana | $m_e$=49.1259 |
| modabrak | |

Statystyki rozproszenia:

| | |
|---|---|
| wariancja | $s^2$=615.69 |
| odchylenie standardowe | s=24.8131 |
| odchylenie przeciętne od $\overline{x}$ | $d_1$=18.2191 |
| odchylenie przeciętne od $m_e$ | $d_2$=12.5955 |

Statystyki kształtu:

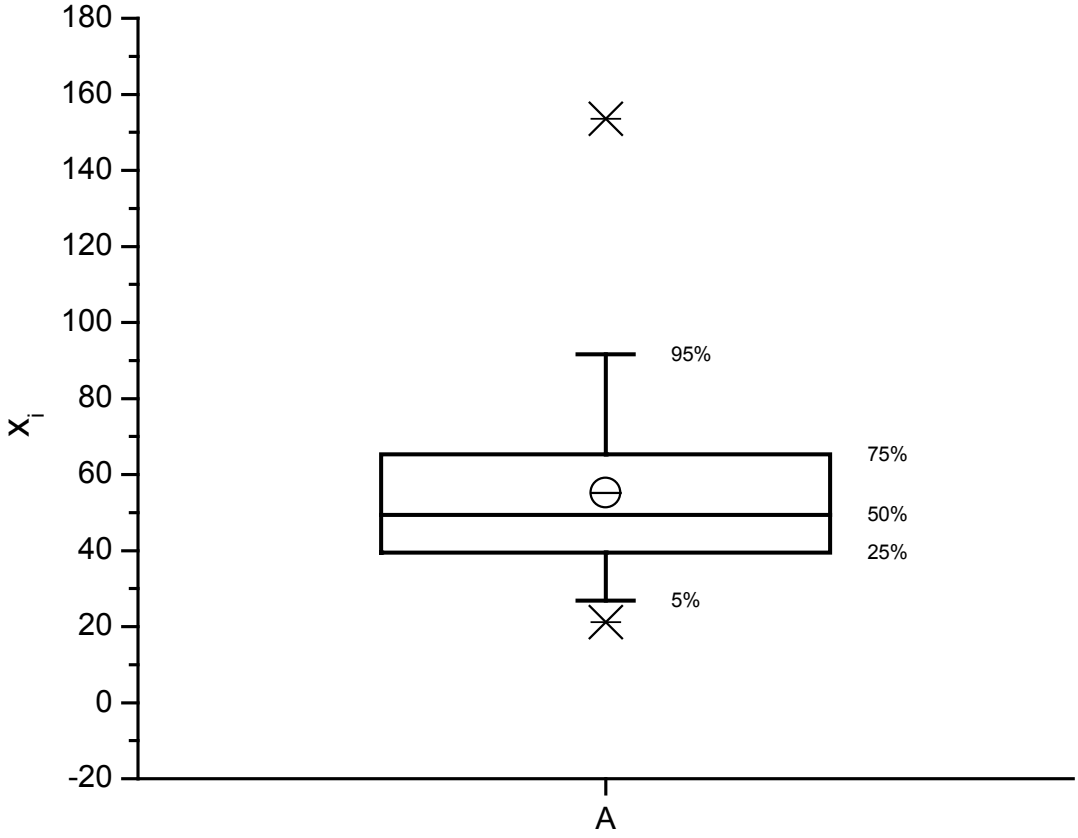
| | |
|---|---|
| moment centralny l=3 | $M_3$=25213 |
| moment centralny l=4 | $M_4$=2.67679·10$^6$ |
| współczynnik asymetrii | $g_1$=1.65037 |
| kurtoza | K=7.06139 |
| eksces | $g_2$=4.06139 |
| współczynnik zmienności | ν=44.94 % |
| współczynnik nierównomierności | H=33.00 % |




Rys. 2. Wykres skumulowanego prawdopodobieństwa $p_i$ ($x_i$) [*wyrażonego w %]* tego, że znajdziemy w próbce wartość ≤$x_i$

Kwantyle:

| | |
|---|---|
| kwantyl rzędu 0.01 | 21.1369 |
| kwantyl rzędu 0.05 | 21.6994 |
| kwantyl rzędu 0.25 | 39.4973 |
| kwantyl rzędu 0.50 | 48.9213 |
| kwantyl rzędu 0.75 | 65.3572 |
| kwantyl rzędu 0.95 | 91.6703 |
| kwantyl rzędu 0.99 | 153.614 |




Rys. 3. Wykres ramkowy: wartość średnia (kółko z poziomą kreską), wartości ekstremalne (poziome kreski), kwartyle (pudełko), kwantyle 0.05 i 0.95 (wąsy), kwantyle 0.01 i 0.99 (krzyżyki)

Literatura:

W.Krysicki i inni, *Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach, część II: Statystyka matematyczna*, PWN, Warszawa 1995

J.Tukey, *Explanatory Data Analysis. Reading*, MA:Addison-Wesley, 1977

Eric Weissteins's World of Mathematics, http://mathworld.wolfram.com/