Politechnika Wrocławska

Niezawodność i diagnostyka układów cyfrowych 2

Projekt – etap 2

Analiza danych

Szymon Datko

szymon.datko@pwr.edu.pl

Wydział Elektroniki,

Politechnika Wrocławska

semestr letni 2020/2021

HR EXCELLENCE IN RESEARCH

W skrócie

Wymagania wstępne:

- przygotowanie narzędzia, symulującego działanie określonego systemu,
- eksport mierzalnych parametrów systemu w wygodnym formacie,
- przykład – plik w wygodnym do przetwarzania formacie CSV:
  - `9978, 17, 4, 1  # wynik jednego uruchomienia programu`
  - `9979, 18, 3, 0  # wynik drugiego uruchomienia (próby)`
  - powyższe numery mogą tu oznaczać liczby pakietów przesłanych:
    - poprawnie (bez błędów),
    - z błędami naprawialnymi,
    - z błędami naprawialnymi,
    - z błędami niewykrytymi.

Cel etapu:

- określić spodziewane wartości parametrów wyjściowych modelu,
- wykorzystując przy tym różne narzędzia statystyczne.

---

1) Aby określenie było rzetelne, musimy najpierw wykonać dostatecznie dużo prób (więcej: Metoda Monte Carlo).
2) Wszystkie próby w tym etapie dotyczą jednej ustalonej konfiguracji parametrów wejściowych systemu, czyli na przykład wielkości pakietu, rozmiaru danych do przesłania, stosowanego algorytm kodowania, itp.

Politechnika Wrocławska

Część I

Metody analiz statystycznych

Średnia arytmetyczna

Klasycznie rozumiana, intuicyjna wartość – suma przez liczbę elementów.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N}$$

Odchylenie standardowe z próby:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N}}$$

Razem $\bar{x}$ i s dostarczają podstawowych informacji na temat rozkładu wartości.

Niektórzy twierdzą, iż lepiej jest zastosować $N - 1$ w mianowniku wzoru na wariancję s.
Hasło dla zainteresowanych: obciążenie estymatora wariancji.
W przypadku naszych zajęć nie ma to dużego znaczenia.

Statystyka pięciopunktowa

Częściej spotykana pod angielskim sformułowaniem ***five-number summary***.

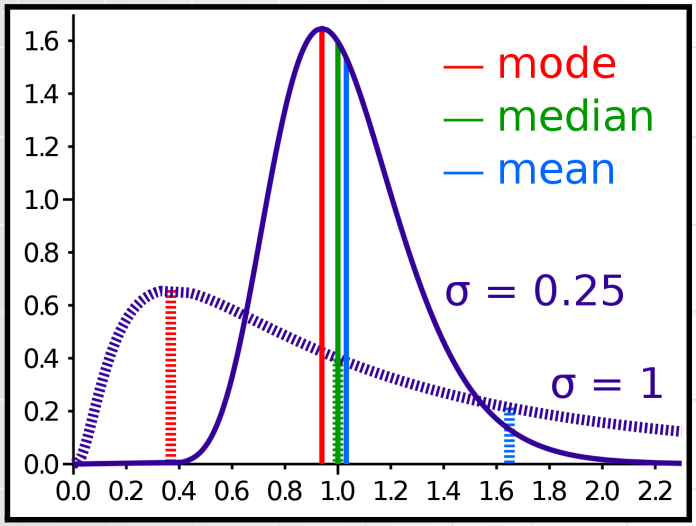
Zapewnia więcej szczegółów na temat rozkładu danych w zbiorze.

Wyznacza się pięć liczb – tak zwanych **kwartyli** rozkładu:

- Q_0 – wartość minimalna,
- Q_1 – wartość większa niż 25% danych w zbiorze,
- Q_2 – mediana (wartość większa niż 50% danych w zbiorze),
- Q_3 – wartość większa niż 75% danych w zbiorze,
- Q_4 – wartość maksymalna.

Zazwyczaj oblicza się także przedział międzykwartylowy $IQR = Q_3 - Q_1$.

Przykład: określenie skośności rozkładu.



Źródło obrazka: https://en.wikipedia.org/wiki/Average

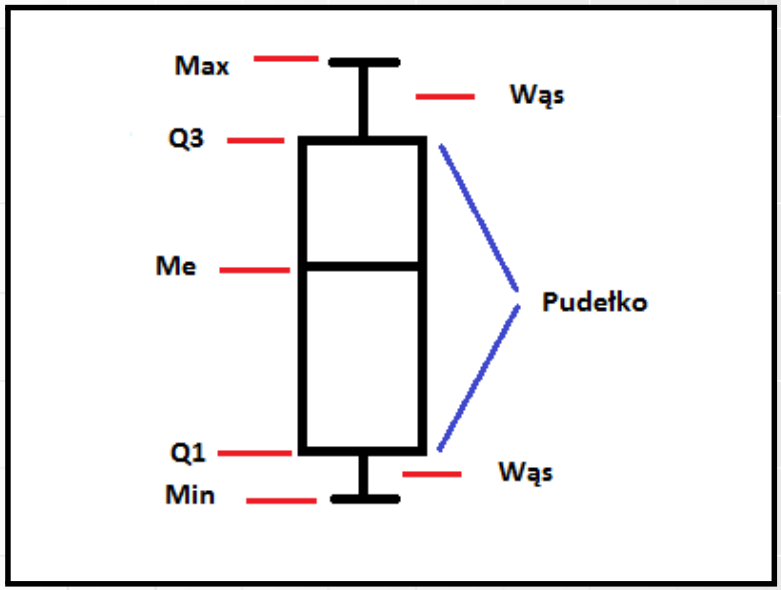
Elementy Q_0 i Q_4 nie stanowią typowej notacji. Zapis ten jest inspirowany implementacjami informatycznymi.

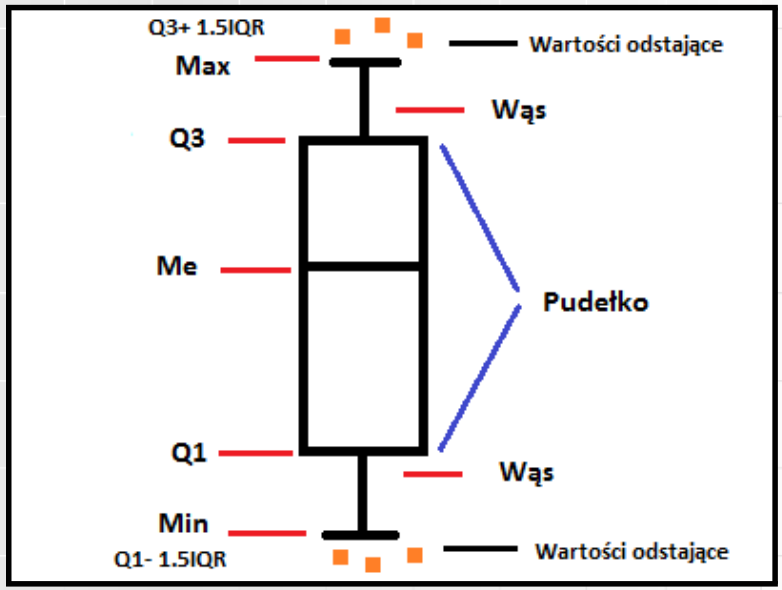
Wykres pudełkowy

Sposób graficznej wizualizacji rozkładu wartości w zbiorze danych.

Opiera się na kwartylach rozkładu/statystyce pięciopunktowej:

- W wariancie podstawowym stosuje się wartości Q_0, Q_1, Q_2, Q_3 i Q_4.
- W wariancie złożonym stosuje się tylko wartości Q_1, Q_2 i Q_3:
  - definiuje się dodatkowo $min = Q_1 - 1.5 \cdot IQR$ oraz $max = Q_3 + 1.5 \cdot IQR$,
  - dane mniejsze od *min* i większe od *max* określa się jako odstające.





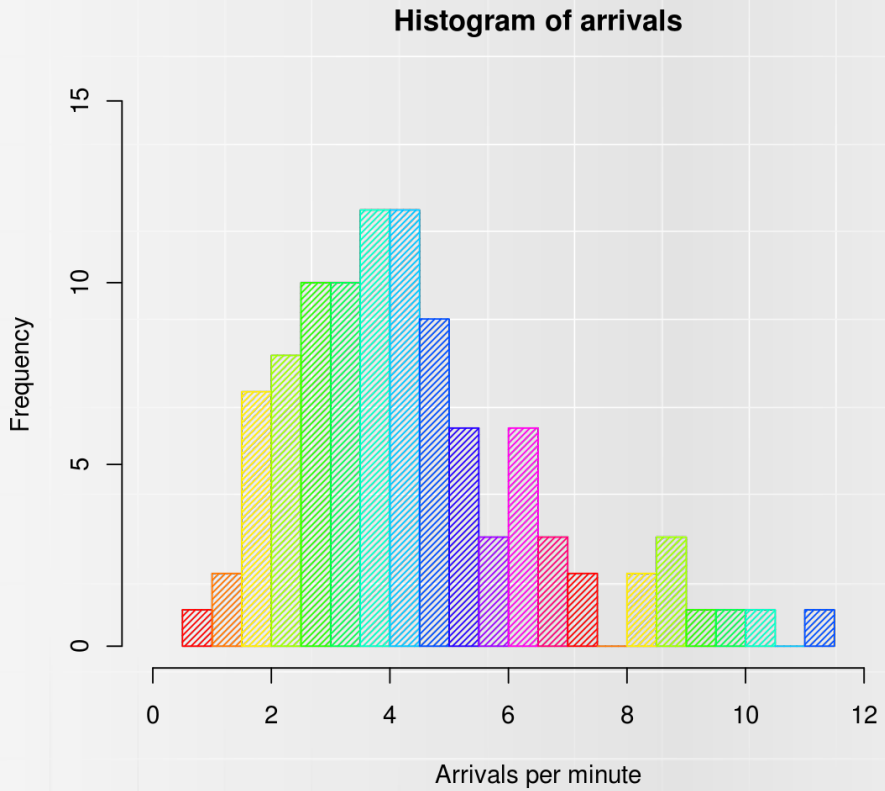
Źródło i więcej informacji: https://mfiles.pl/pl/index.php/Wykres_pudełkowy (ang. *boxplot*)

Wybór wariantu nie ma znaczenia w projekcie, choć niektórzy statystycy nie uznają istnienia wariantu podstawowego.

Histogram

Wykres przedstawiający zakres wartości oraz częstość lub liczbę ich wystąpień.

- Graficzna wizualizacja rozkładu uzyskanych wartości.
- W przybliżeniu odpowiada funkcji gęstości prawdopodobieństwa.



Pytanie otwarte – jak dobrać szerokość przedziałów klasowych dla zakresu wartości?*

Źródło obrazka: https://en.wikipedia.org/wiki/Histogram

* – Więcej informacji: https://www.statystyczny.pl/pora-na-rysunki-histogram/

Dopasowanie danych do modelu

Zakładamy, że dane pochodzą z określonego rozkładu/modelu parametrycznego.

- Poszukujemy najlepszych wartości parametrów tego modelu dla danych.
- Intuicyjnie: wartości obliczone z modelu powinny odpowiadać tym z próby.

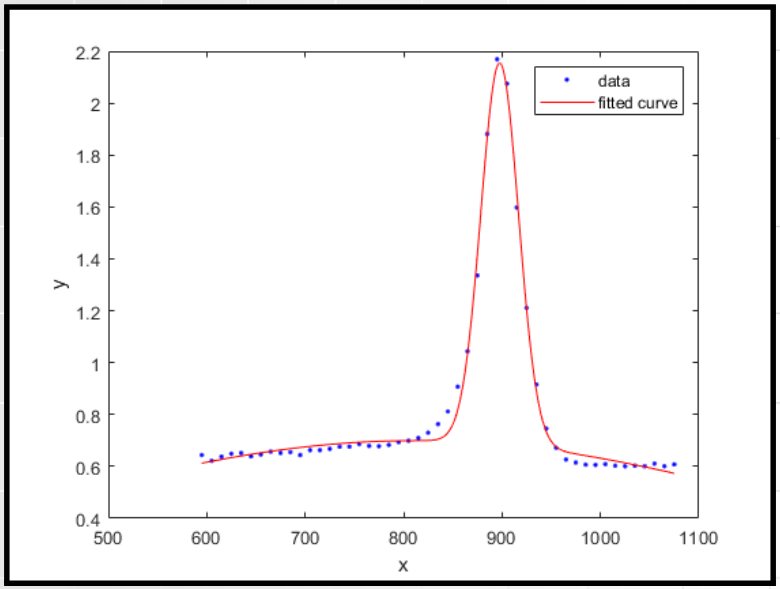
Na potrzeby projektu załóżmy, że histogram wartości ma cechy rozkładu Gaussa.

Model będzie stanowiła funkcja postaci

$$f_{\mathcal{N}}(A, \mu, \sigma) = A \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Wynik działania procedury to 3 wartości:

- A – amplituda (tutaj nieistotna),
- μ – wartość średnia,
- σ – odchylenie standardowe.



W przypadku danych z rozkładu Gaussa: $\mu = \bar{x} = Q_2$ oraz $\sigma = s = \frac{IQR}{1.3490}$.

Źródło obrazu: https://www.mathworks.com/help/curvefit/gaussian.html

Więcej informacji: https://en.wikipedia.org/wiki/Curve_fitting (ang. *Curve Fitting*)