|  |
| --- |
| Laboratorium semestr zimowy 21/22 |
| Sprawozdanie |
| Metody probabilistyczne w informatyce |

|  |
| --- |
| Bartłomiej Błaszczyk  236382 3i3 NS |

Spis treści

[Wstęp 2](#_Toc95361668)

[Metoda Histogramu 3](#_Toc95361669)

[Przebieg badania 3](#_Toc95361670)

[Test serii 4](#_Toc95361671)

[Przebieg badania 4](#_Toc95361672)

[Przygotowanie danych do analizy statystycznej - eliminacja błędów grubych 6](#_Toc95361673)

[Przebieg badania 6](#_Toc95361674)

[Rozkład normalny 8](#_Toc95361675)

[Przebieg badania 9](#_Toc95361676)

[Rozkład logarytmo-normalny 12](#_Toc95361677)

[Przebieg badania 12](#_Toc95361678)

[Równanie 1 4](#_Toc95361693)

[Równanie 2 4](#_Toc95361694)

[Równanie 3 6](#_Toc95361695)

[Równanie 4 7](#_Toc95361696)

[Równanie 5 8](#_Toc95361697)

[Tabela 1 3](#_Toc95361706)

[Tabela 2 3](#_Toc95361707)

[Tabela 3 5](#_Toc95361708)

[Tabela 4 6](#_Toc95361709)

[Tabela 5 6](#_Toc95361710)

[Tabela 6 7](#_Toc95361711)

[Tabela 7 9](#_Toc95361712)

[Tabela 8 11](#_Toc95361713)

[Wykres 1 3](#_Toc95361725)

[Wykres 2 5](#_Toc95361726)

[Wykres 3 7](#_Toc95361727)

[Wykres 4 8](#_Toc95361728)

[Wykres 5 10](#_Toc95361729)

[Wykres 6 11](#_Toc95361730)

[Wykres 7 12](#_Toc95361731)

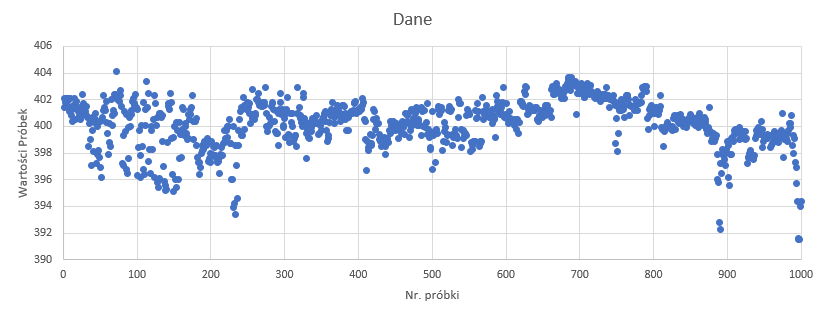
[Wykres 8 13](#_Toc95361732)

[Wykres 9 14](#_Toc95361733)

# Wstęp

Przedmiotem sprawozdania jest badanie statystyczne zestawu danych dostarczonych przez prowadzącego na zajęciach z przedmiotu metody probabilistyczne w informatyce.

Dane opracowane na laboratorium pochodzą z pliku: dane2.txt.



Dane

W sprawozdaniu znajduje się opracowanie metod badawczych używanych do statystycznej analizy danych z wykorzystaniem różnych metod, jak i opis niezbędnych kroków potrzebnych do obróbki badanego zbioru, aby uzyskać miarodajne informacje.

# Metoda Histogramu

Metoda analizy statystycznej, która opiera się na szeregu rozdzielczym.

Metoda ta daje dość atrakcyjne graficznie wyniki, jednakże wymaga dodatkowych założeń co do podziału zakresu zmiennej losowej na klasy i co do liczności realizacji zmiennej losowej w poszczególnych klasach. Zarówno przy wyborze granic klas jak i przy wyborze liczności w klasach występuje dość duża niejednoznaczność kryteriów, mogąca dawać spore różnice wyników.

## Przebieg badania

Zaczynamy od ponumerowania próbek i określenia podstawowych parametrów badanego wektora.



Tabela

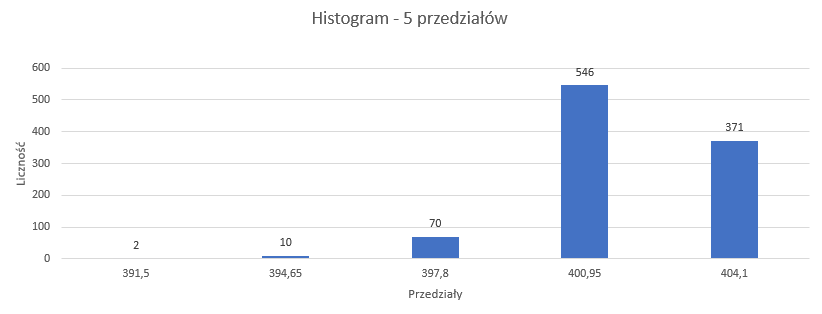
Następnie dzielimy zakres wartości wektora na określoną ilość przedziałów i zliczamy liczność wystąpień próbek w danym przedziale.

Obraz zawierający stół

Opis wygenerowany automatycznie

Tabela

Po takim zabiegu jesteśmy wstanie przygotować wykres dla pięciu przedziałów.



Wykres

Badanie zostało przeprowadzone jeszcze dla dziesięciu i piętnastu przedziałów.

# Test serii

Wykonując badania statystyczne zwykle obserwuje się dwa rodzaje zdarzeń: coś się wydarzyło lub coś się nie wydarzyło albo coś jest czerwone lub nie jest czerwone. Są to zatem przypadki rozkładu zero-jedynkowego. We wszystkich takich przypadkach można utworzyć ciąg elementów dwóch rodzajów.

Test serii do oceny losowości można stosować nie tylko wówczas, gdy zmienna losowa przybiera wartości 0 lub 1, czyli gdy podlega rozkładowi dwumianowemu, lecz również przy badaniach wartości zmiennej losowej ciągłej.

## Przebieg badania

Badanie zaczynamy od obliczenia wartości średniej i mediany, oraz wygenerowania nowej zmiennej losowej zgodnej ze wzorami:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Równanie

W kolejnym kroku wylicza się wartości krytyczne za pomocą podanych narzędzi:

Obraz zawierający stół

Opis wygenerowany automatycznie

Równanie

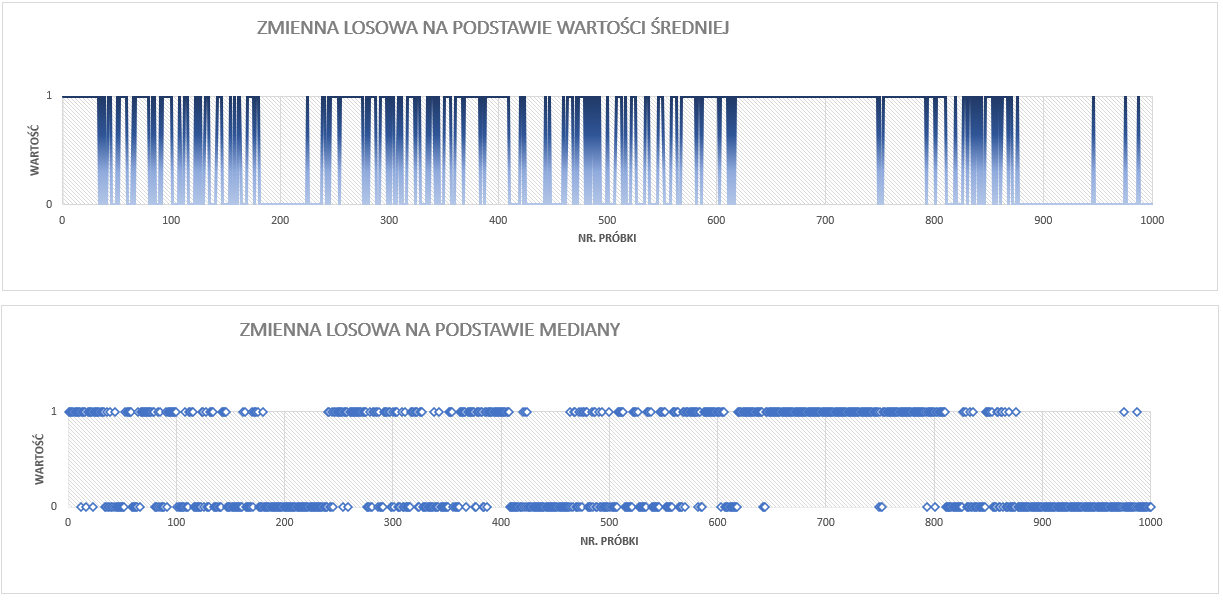
|  |  |
| --- | --- |
| WARTOŚĆ | obliczona dla danej populacji |
| n1 | Liczba zer dla nowo wygenerowanej zmiennej losowej |
| n2 | Liczba jedynek dla nowej zmiennej |
| n | Rozmiar populacji |
| µ | Środek rozrzutu, który dla rozkładu normalnego pokrywa się z wartością oczekiwaną, medianą i modą |
| σ | odchylenie standardowe |
| k1 | Pierwsza wartość krytyczna |
| k2 | Druga wartość krytyczna |
| LICZBA SERII | Zliczenie kolejnych par powtarzających się po sobie wystąpień nowej zmiennej losowej |

Obraz zawierający stół

Opis wygenerowany automatycznie

Tabela

Po skończonych obliczeniach można naszkicować wykresy pokazujące rozkład nowej zmiennej losowej w podziale na zbiory.



Wykres

I określić czy badany zbiór spełnia warunek losowości, czyli czy znajduje się pomiędzy wartościami krytycznymi.



Rysunek

# Przygotowanie danych do analizy statystycznej - eliminacja błędów grubych

Czasem w badanej populacji spotkamy się z anomaliami w danych. Należy wtedy się zastanowić nad sensem występowania wartości odstających od reszty. Czasem jednak ciężko jest ocenić jak bardzo odstaje dana próbka; należy wtedy skorzystać z testu błędów grubych opartego na statystykach.

## Przebieg badania

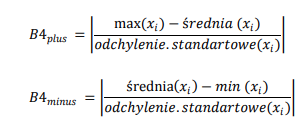
Analizę zaczęliśmy od posortowania danych i wyznaczenia podstawowych parametrów wektora.

Obraz zawierający stół

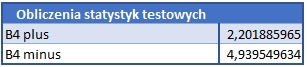
Opis wygenerowany automatycznie

Tabela

Odchylenie standardowe jest nam potrzebne do określenia wartości, które opisują jak bardzo nasze dane są rozrzucane od górnej granicy zakresu wartości i dolnej.



Równanie



Tabela

Teraz należy wyliczyć wartość b4, którą porównamy z wartościami Bi, które pozwolą nam sprawdzić, czy przy założonym poziomie istotności, istnieją próbki obarczone błędem grubym.

|  |  |
| --- | --- |
| α | Poziom istotności |
| n | Liczność zbioru |
| β | 1- (α/n) |
| y | kwantyl |
| b4 | Wartość krytyczna |

Obraz zawierający stół

Opis wygenerowany automatycznie

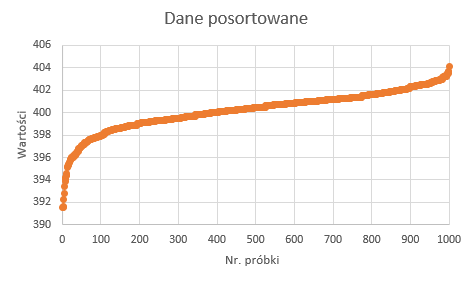
Równanie

Obraz zawierający stół

Opis wygenerowany automatycznie

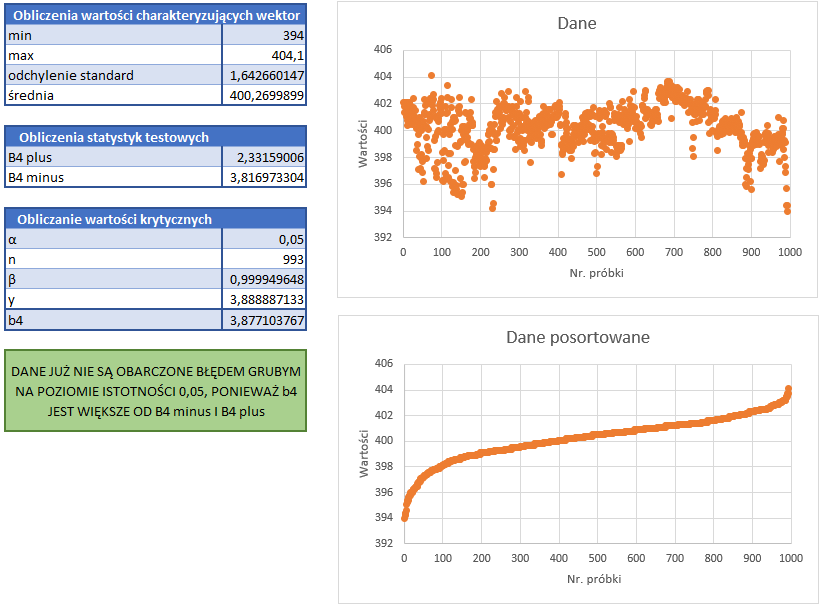
Tabela

Po posortowaniu danych rzeczywiście jesteśmy wstanie zauważyć, że od dolnej granicy przedziału wartości jest dużo większy rozrzut, niż chociażby, przy górnej granicy.



Wykres

Należy zatem wyeliminować kolejne próbki sprawdzając za każdym razem jak zmieniają się nasze wartości krytyczne.



Wykres

# Rozkład normalny

Rozkład normalny jest najstarszym, najlepiej zbadanym i bardzo istotnym dla praktyki inżynierskiej rozkładem prawdopodobieństwa zmiennej losowej ciągłej X. Bazuje on na centralnym twierdzeniu granicznym, które brzmi następująco: Jeśli Xi są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, takiej samej wartości oczekiwanej µ oraz dodatniej i skończonej wariancji σ2 to zmienna losowa o postaci…

Obraz zawierający tekst, zegar, zegarek

Opis wygenerowany automatycznie

Równanie

… zbiega według rozkładu do standardowego rozkładu normalnego, gdy n rośnie do nieskończoności.

## Przebieg badania

Stworzenie szeregu kumulacyjnego (posortowanie danych w kolejności rosnącej.

Obliczenie podstawowych parametrów wektora jak: średnia, mediana, odchylenie standardowe, wartość maksymalna i minimalna.

Obraz zawierający stół

Opis wygenerowany automatycznie

Tabela

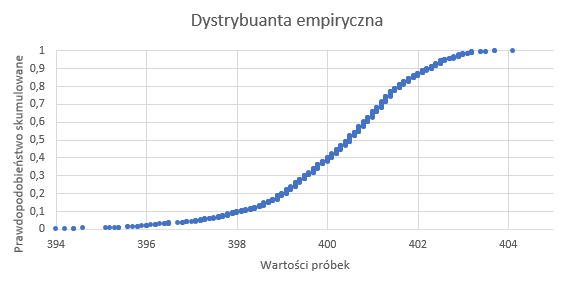
Z wykorzystaniem wzoru 𝑝𝑖 = 𝑖𝑙𝑜 𝑠𝑐\_𝑝𝑟𝑜𝑏𝑒𝑘 𝑖 𝑛+1 należy stworzyć wektor reprezentujący prawdopodobieństwo skumulowane – dla przypomnienia wartości w tym wektorze zależą wyłącznie od liczności wektora n.

Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek

Przedstawienie na wykresie danych o współrzędnych (x,p), gdzie x to dane posortowane a p to odpowiadające im prawdopodobieństwa. Wykres ma mieć charakter punktowy.



Wykres

Wyznaczenie wartości standaryzowanych y dla wartości wektora p zgodnie z procedurą przedstawioną na wykładzie.

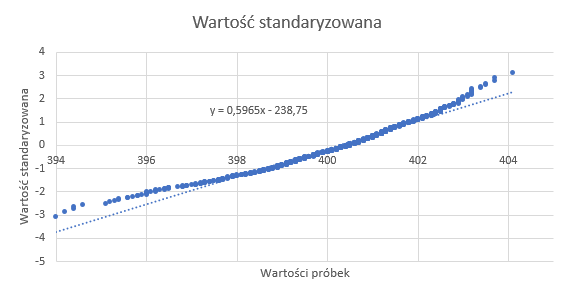
Obraz zawierający stół

Opis wygenerowany automatycznie

Rysunek

yi =ROZKŁ.NORMALNY.ODWR(pi , 0, 1)

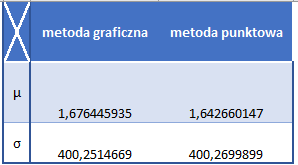
Przedstawienie na wykresie danych o współrzędnych (x,y), gdzie x to dane posortowane a y to odpowiadające im wartości standaryzowane. Punkty powinny mieć charakter linii prostej



Wykres

Dobieramy prostą regresji dla w/w zestawu punktów. UWAGA: Proszę pamiętać, że na wykładzie prosta ta jest definiowana jako y = a + bx, zaś wzór prostej regresji w Excelu może mieć postać   
y = ax +b. W związku z tym należy odpowiednio zrewidować wzory na parametry rozkładu Gaussa. Oczywiście prosta ta musi znaleźć się na wykresie razem z punktami.

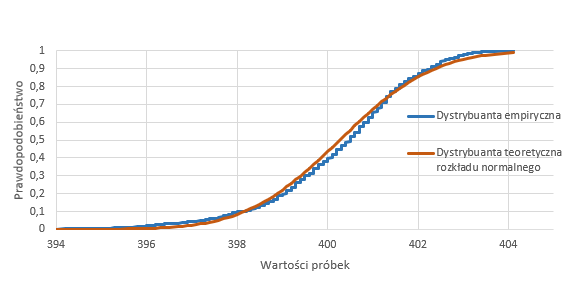
Obliczamy parametry μ i σ z otrzymanych wartości a i b



Tabela

a to współczynnik kierunkowy, a b to wyraz wolny równania prostej linii trendu.

Przedstawiamy na wykresie dystrybuantę rozkładu Gaussa (zgodnie ze stosowanym wzorem) o parametrach μ i σ nanosząc na wykres także punkty o współrzędnych (x,p) jak na pierwszym wykresie



Wykres

# Rozkład logarytmo-normalny

Jeżeli w centralnym twierdzeniu granicznym zamiast o sumie niezależnych czynników losowych mówić o ich iloczynie to zamiast rozkładu normalnego mamy do czynienia z rozkładem logarytmo-normalnym

## Przebieg badania

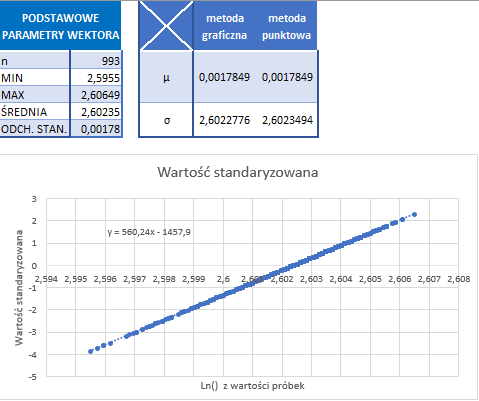
po punkcie 2 należy wektor danych uporządkowanych zlogarytmować logarytmem naturalnym, i taki wektor dalej analizować jak w przypadku rozkładu Gaussa.

Obraz zawierający stół

Opis wygenerowany automatycznie

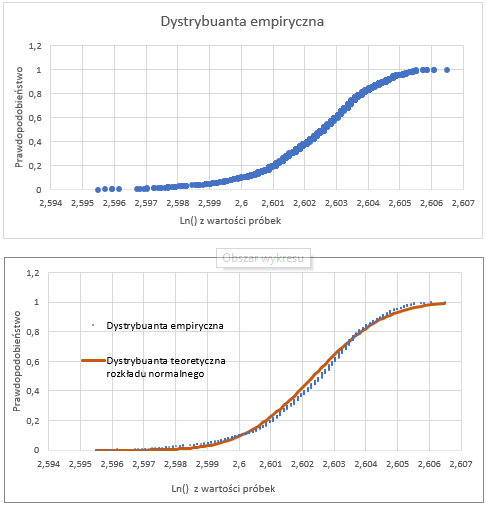
Rysunek

przy wykreślaniu rozkładu teoretycznego należy zastosować wzór na dystrybuantę rozkładu logarytmo-normalnego a parametry rozkładu μ i σ będą z dziedziny logarytmów



Wykres

- na końcu ze wzorów z wykładu policzyć parametry powiązane z rozkładem (wartość oczekiwana, mediana, moda i odchylenie standardowe jako pierwiastek z wariancji)



Wykres