Laboratorium semestr zimowy 21/22

Sprawozdanie

Metody probabilistyczne w informatyce

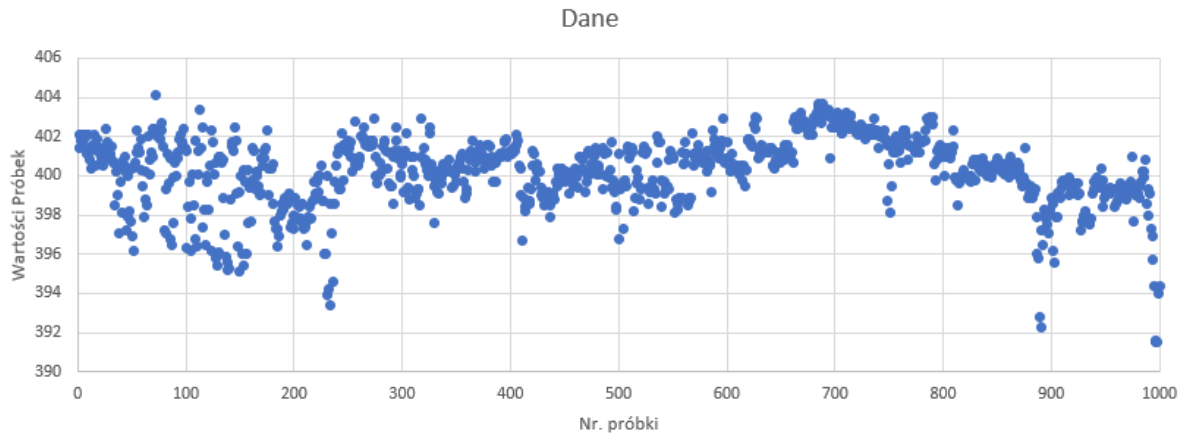
Bartłomiej Błaszczyk
236382 3i3 NS

Spis treści

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.

Przedziały	
min.	391,5
maks.	404,1
Przedział wartości	12,6

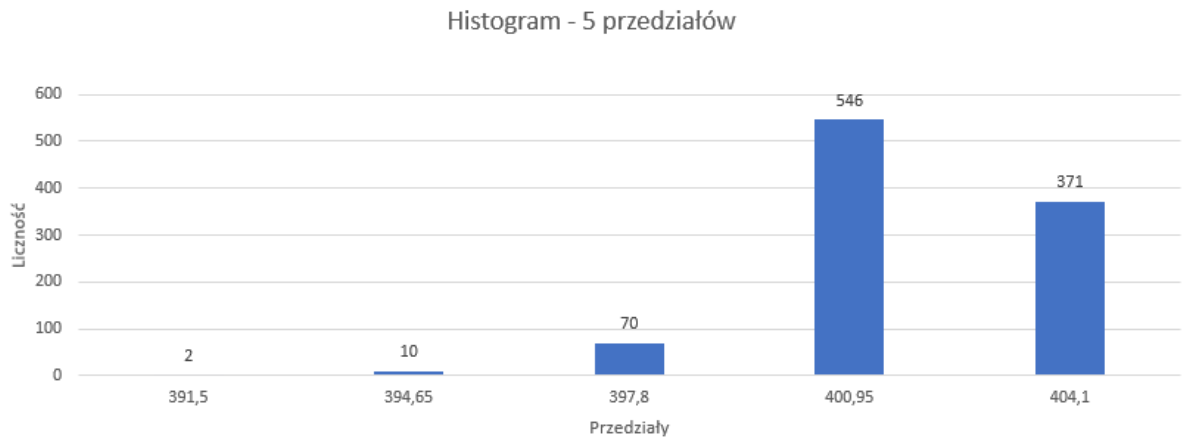
Tabela 1

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

5 przedziałów	
Wielkość przedziału	3,15
Przedziały	Liczność
391,5	2
394,65	10
397,8	70
400,95	546
404,1	371

Tabela 2

Po takim zabiegu jesteśmy wstanie przygotować wykres dla pięciu przedziałów.



Wykres 1

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:

$x > x_{śr}$ daje 1 $\quad$ $x > x_{mediana}$ daje 1

$x < x_{śr}$ daje 0 $\quad$ $x < x_{mediana}$ daje 0

Równanie 1

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

$$\mu = \frac{2 \cdot n_1 \cdot n_2}{n} + 1$$

$$\sigma = \sqrt{\frac{2 \cdot n_1 \cdot n_2 . \cdot (2 \cdot n_1 \cdot n_2 - n)}{(n-1) \cdot n^2}}$$

Równanie 2

WARTOŚĆ	obliczona dla danej populacji
n_1	Liczba zer dla nowo wygenerowanej zmiennej losowej
n_2	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

	OBLICZENIA DLA WARTOŚCI ŚREDNIEJ	OBLICZENIA DLA MEDIANY
WARTOŚĆ	400,2151	400,5
n_1	449	525
n_2	551	475
n	1000	1000
μ	495,798	499,75
σ	15,63888863	15,76392444
k1	ROZKŁ.NORMALNY.ODWR(0,05; K6;K7)	473,8206517
k2	ROZKŁ.NORMALNY.ODWR(0,95; K6;K7)	525,6793483
LICZBA SERII	SUMA(E2:E1001)	826

Tabela 3

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

Wykres 2

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

DLA WARTOŚCI ŚREDNIEJ WARUNEK LOSOWOŚCI NIE ZOSTAŁ SPEŁNIONY 470,1 < 813,0 < 521,5
DLA MEDIANY WARUNEK LOSOWOŚCI NIE ZOSTAŁ SPEŁNIONY 473,8 < 826,0 < 525,7

Rysunek 1

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.

Obliczenia wartości charakteryzujących wektor	
min	391,5
max	404,1
odchylenie standard	1,764351134
średnia	400,2151

Tabela 4

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

$$B4_{plus} = \left| \frac{\max(x_i) - średnia\ (x_i)}{odchylenie.standartowe(x_i)} \right|$$

$$B4_{minus} = \left| \frac{średnia(x_i) - min\ (x_i)}{odchylenie.standartowe(x_i)} \right|$$

Równanie 3

Obliczenia statystyk testowych	
B4 plus	2,201885965
B4 minus	4,939549634

Tabela 5

Teraz należy wyliczyć wartość b4, którą porównamy z wartościami B_i, 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

$$b4 := y \cdot \sqrt{\left[\frac{2 \cdot (n-1)}{2 \cdot n - 5 + y^2 + \left(3 + y^2 + 2 \cdot y^4\right) \cdot \frac{1}{6 \cdot (2 \cdot n - 5)}}\right]}$$

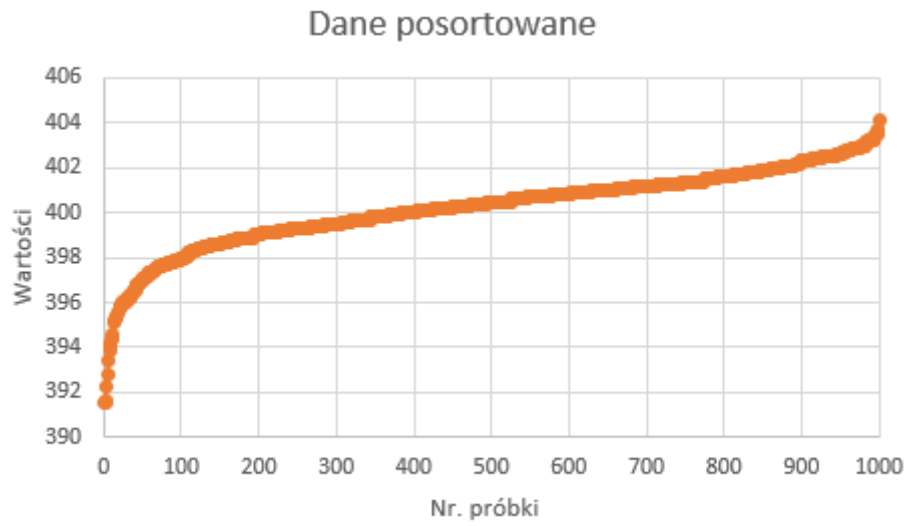
Równanie 4

Obliczanie wartości krytycznych	
α	0,05
n	1000
β	0,99995
y	3,890591886
b4	3,878790494

DANE SĄ OBARCZONE BŁĘDEM GRUBYM NA POZIOMIE ISTOTNOŚCI 0,05, PONIEWAŻ b4 JEST MNIEJSZE OD B4 minus

Tabela 6

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



Wykres 3

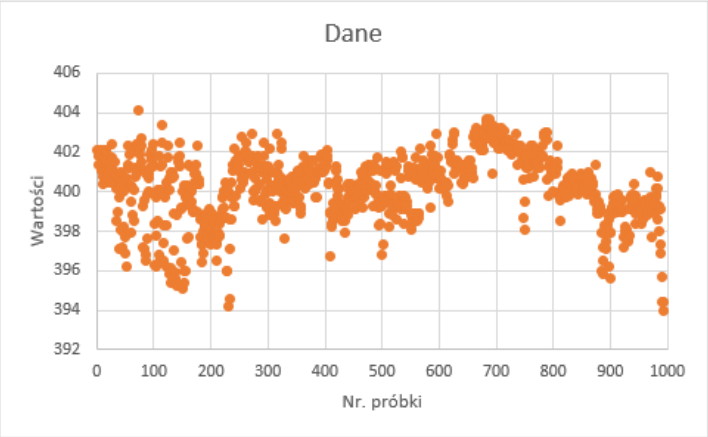
Należy zatem wyeliminować kolejne próbki sprawdzając za każdym razem jak zmieniają się nasze wartości krytyczne.

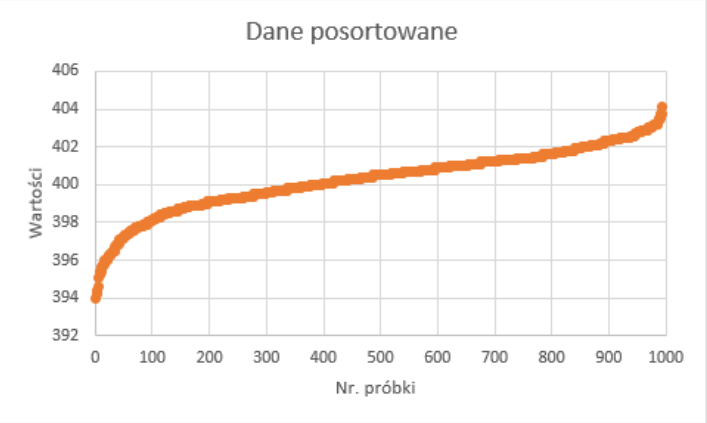
Obliczenia wartości charakteryzujących wektor	
min	394
max	404,1
odchylenie standard	1,642660147
średnia	400,2699899

Obliczenia statystyk testowych	
B4 plus	2,33159006
B4 minus	3,816973304

Obliczanie wartości krytycznych	
α	0,05
n	993
β	0,999949648
y	3,888887133
b4	3,877103767

DANE JUŻ NIE SĄ OBARCZONE BŁĘDEM GRUBYM NA POZIOMIE ISTOTNOŚCI 0,05, PONIEWAŻ b4 JEST WIĘKSZE OD B4 minus I B4 plus





Wykres 4

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…

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Równanie 5

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

Przebieg badania

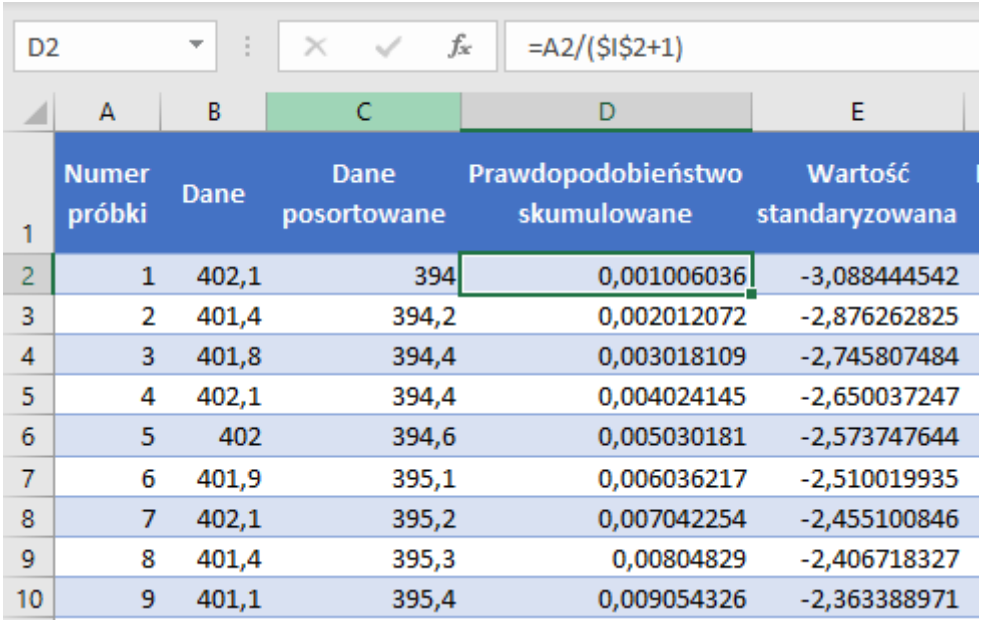
Stworzenie szeregu kumulacyjnego (posortowanie danych w kolejności rosnącej.

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

PODSTAWOWE PARAMETRY WEKTORA	
n	993
MIN	394
MAX	404,1
ŚREDNIA	400,2699899
ODCH. STAN.	1,642660147

Tabela 7

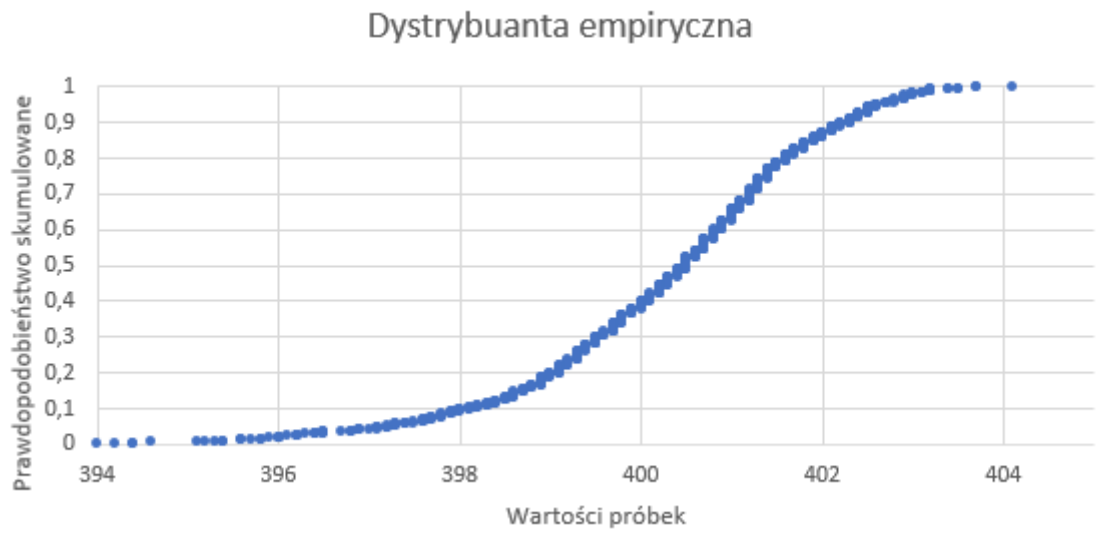
Z wykorzystaniem wzoru $pi = ilo\ sc\_probek\ i\ n$+1 należy stworzyć wektor reprezentujący prawdopodobieństwo skumulowane – dla przypomnienia wartości w tym wektorze zależą wyłącznie od liczności wektora n.



	A	B	C	D	E
1	Numer próbki	Dane	Dane posortowane	Prawdopodobieństwo skumulowane	Wartość standaryzowana
2	1	402,1	394	0,001006036	-3,088444542
3	2	401,4	394,2	0,002012072	-2,876262825
4	3	401,8	394,4	0,003018109	-2,745807484
5	4	402,1	394,4	0,004024145	-2,650037247
6	5	402	394,6	0,005030181	-2,573747644
7	6	401,9	395,1	0,006036217	-2,510019935
8	7	402,1	395,2	0,007042254	-2,455100846
9	8	401,4	395,3	0,00804829	-2,406718327
10	9	401,1	395,4	0,009054326	-2,363388971

Rysunek 2

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 5

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

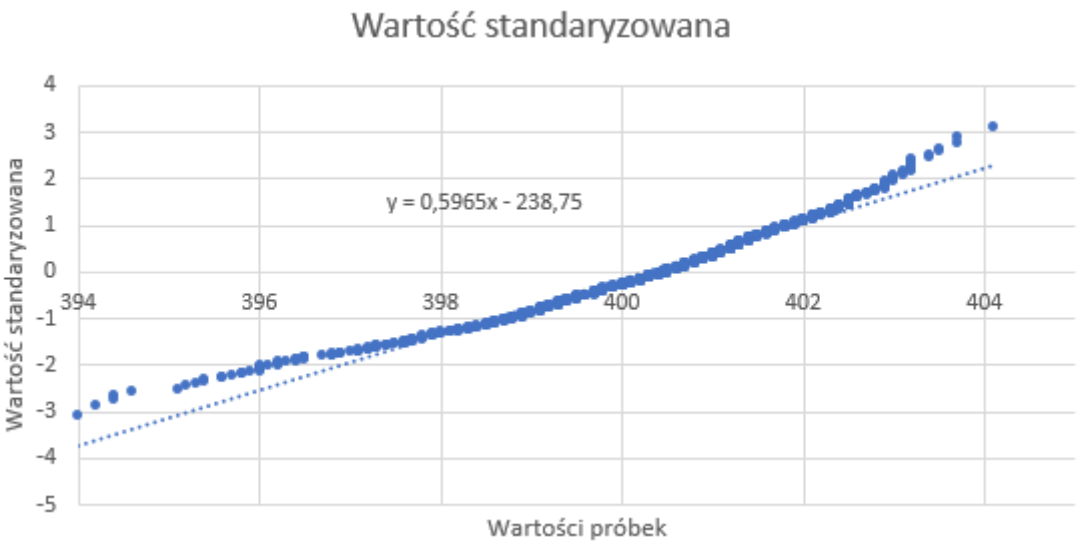
E2 =ROZKŁ.NORMALNY.ODWR(D2;0;1)

	A	B	C	D	E
1	Numer próbki	Dane	Dane posortowane	Prawdopodobieństwo skumulowane	Wartość standaryzowana
2	1	402,1	394	0,001006036	-3,088444542
3	2	401,4	394,2	0,002012072	-2,876262825
4	3	401,8	394,4	0,003018109	-2,745807484
5	4	402,1	394,4	0,004024145	-2,650037247
6	5	402	394,6	0,005030181	-2,573747644
7	6	401,9	395,1	0,006036217	-2,510019935
8	7	402,1	395,2	0,007042254	-2,455100846
9	8	401,4	395,3	0,00804829	-2,406718327
10	9	401,1	395,4	0,009054326	-2,363388971

Rysunek 3

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 6

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

	metoda graficzna	metoda punktowa
μ	1,676445935	1,642660147
σ	400,2514669	400,2699899

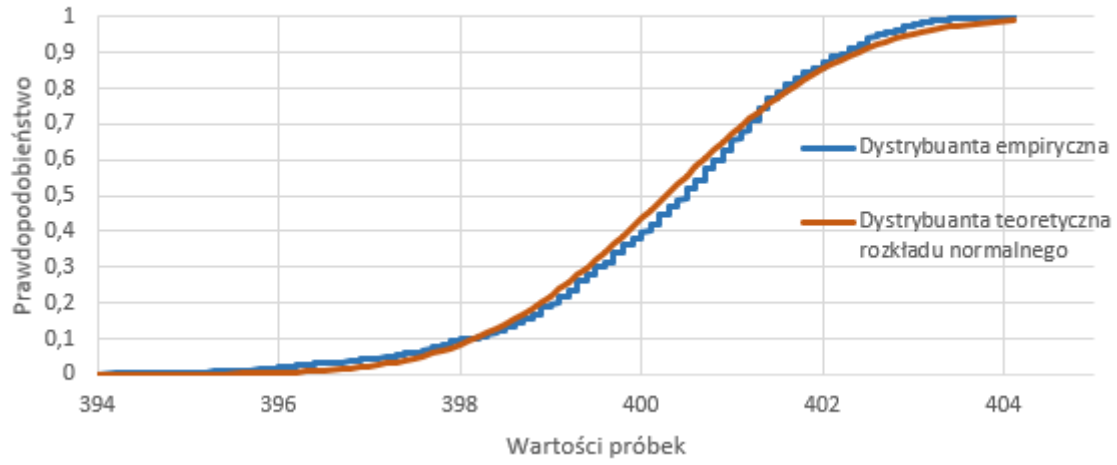
Tabela 8

$$\sigma := \frac{1}{b}$$

$$\mu := -\left(\frac{a}{b}\right)$$

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 7

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.

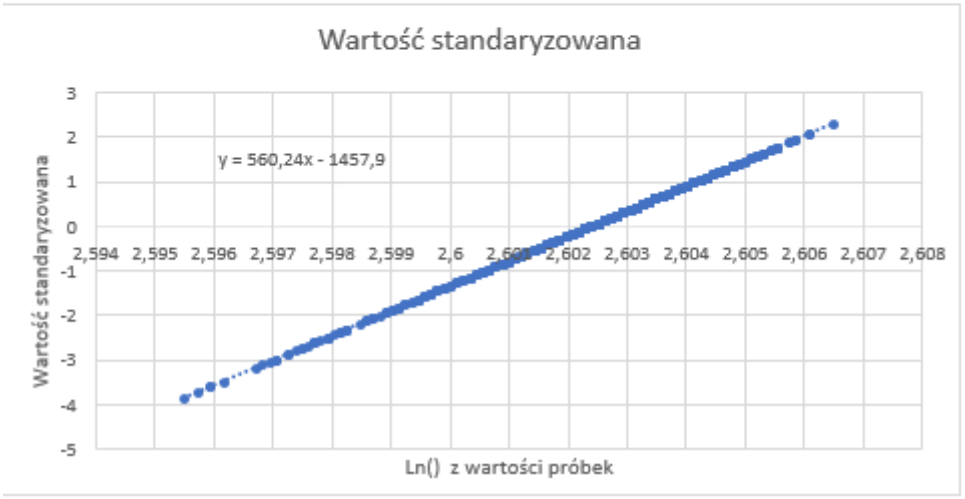
	A	B	C	D	E	F	G
1	Numer próbki	Dane	Dane posortowane	Logarytm naturalny	Prawdopodobieństwo skumulowane	Wartość standaryzowana	Dystrybuanta teoretyczna rozkładu normalnego
2	1	402,1	394	LOG(C2)	A2/(J2+1)	(D2-J5)/J6	ROZKŁ.NORMALNY.S(F2;PRA
3	2	401,4	394,2	2,59571662	0,002012072	-3,745897453	8,9875E-05
4	3	401,8	394,4	2,595936906	0,003018109	-3,621618465	0,000146383
5	4	402,1	394,4	2,595936906	0,004024145	-3,621618465	0,000146383
6	5	402	394,6	2,596157081	0,005030181	-3,497402482	0,000234906
7	6	401,9	395,1	2,59670703	0,006036217	-3,187137756	0,000718442
8	7	402,1	395,2	2,596816936	0,007042254	-3,12513193	0,000888627
9	8	401,4	395,3	2,596926814	0,00804829	-3,063141791	0,001095131
10	9	401,1	395,4	2,597036665	0,009054326	-3,001167333	0,001344734

Rysunek 4

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

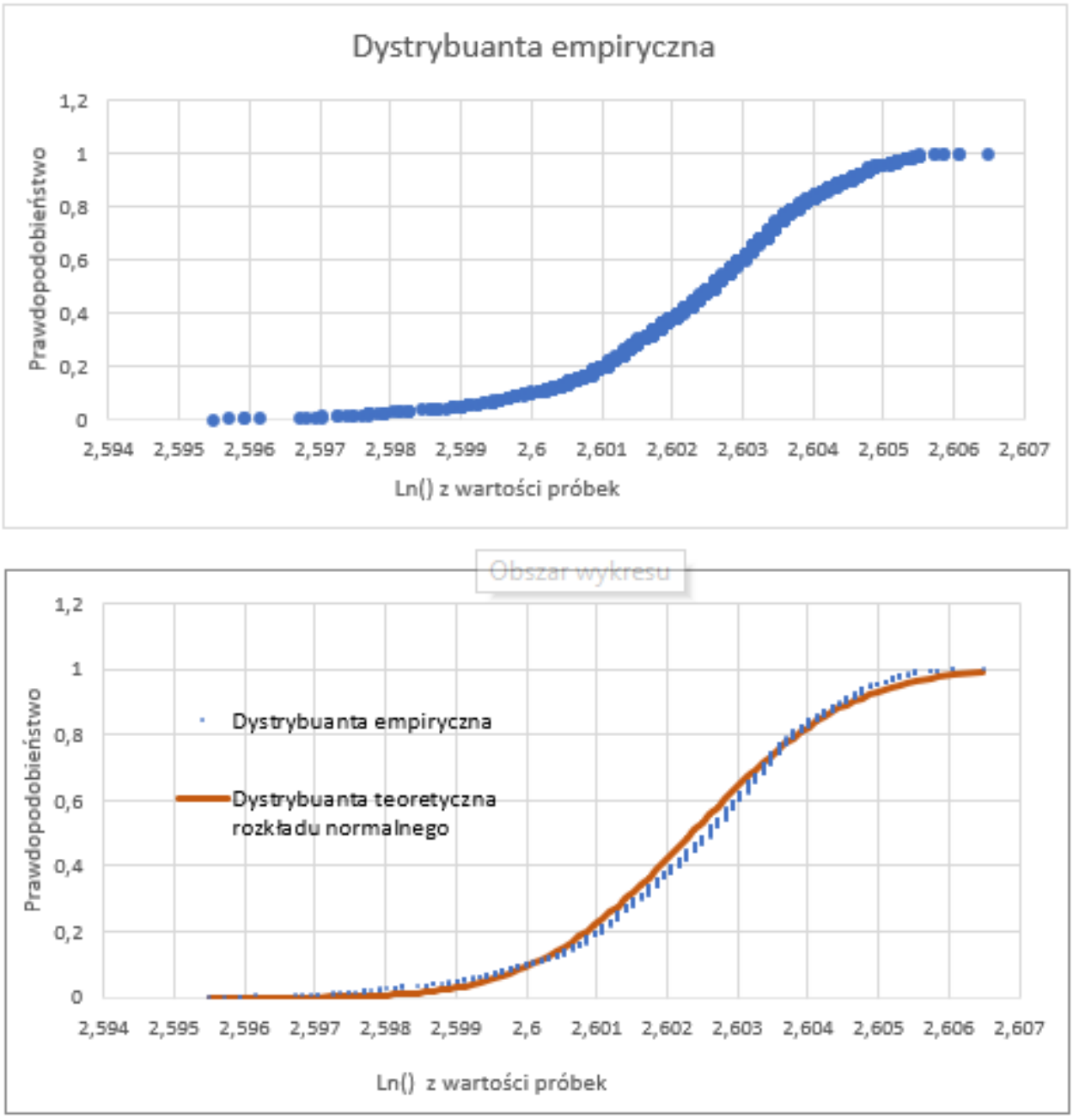
PODSTAWOWE PARAMETRY WEKTORA	
n	993
MIN	2,5955
MAX	2,60649
ŚREDNIA	2,60235
ODCH. STAN.	0,00178

	metoda graficzna	metoda punktowa
μ	0,0017849	0,0017849
σ	2,6022776	2,6023494



Wykres 8

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


Dystrybuanta empiryczna
Prawdopodobieństwo
1,2
1
0,8
0,6
0,4
0,2
0
2,594 2,595 2,596 2,597 2,598 2,599 2,6 2,601 2,602 2,603 2,604 2,605 2,606 2,607
Ln() z wartości próbek
Obszar wykresu
Dystrybuanta empiryczna
Dystrybuanta teoretyczna rozkładu normalnego
Prawdopodobieństwo
Ln() z wartości próbek

Wykres 9