Statystyczna Analiza Danych

Analiza mocy testów normalności & ANOVA

Wprowadzenie:

Cel projektu:

Celem niniejszego projekt jest analiza mocy statystycznej testów normalności oraz analiza wariancji (ANOVA). Projekt ten ma na celu zbadanie, jak różne testy normalności radzą sobie z różnymi typami rozkładów danych, a także ocenić ich skuteczność i niezawodność w praktycznych zastosowaniach. Sprawdzimy takie testy jak: test Shapiro-Wilka, Andersona-Darlinga, Jarque-Bera, Lilieforsa.

Oprócz tego chcemy sprawdzić czy kolor i marka telefonów może być powodem różnic cenowych. Zbadamy marki: Apple, Huawei, Samsung i Xiaomi. Do tego celu zastosujemy analizę wariancji ANOVA.

Dane:

Dane techniczne zostały pozyskane ze strony mgsm.pl, a pojedyncze wartości zaciągnięte ze strony producenta lub ofert sprzedaży. Ceny zostały pozyskane ze stron ceneo.pl i allegro.pl. Dotyczą wyłącznie nieużywanych telefonów i są możliwie najniższe.

Modele telefonów były wybierane w sposób losowy. Następnie, były wybierane różne wariancje tego samego modelu telefonu, zarówno względem parametrów technicznych, jak i dostępnych kolorów. Najmniej wariancji miały telefony marki Huawei, z tego też powodu ich różnych modeli jest najwięcej.

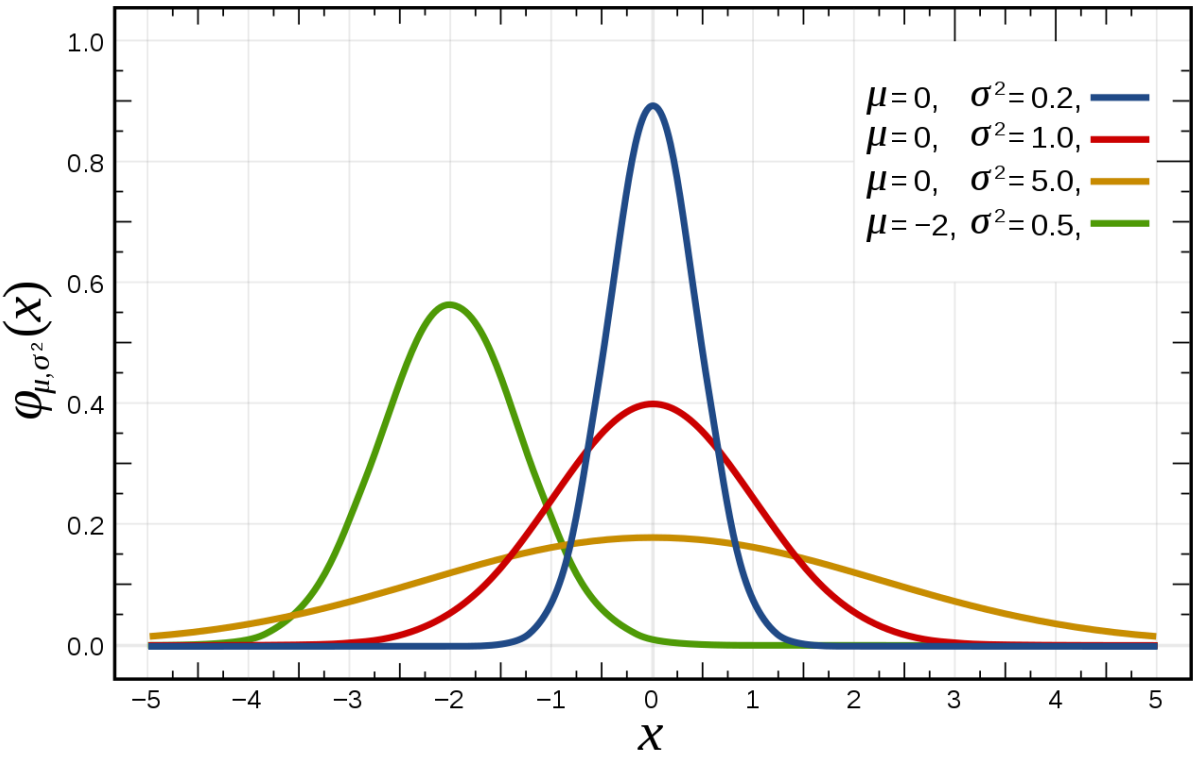
Dla każdej marki telefonu jest około 50 obserwacji.

Analiza mocy testów normalności:

Moc testu jest to prawdopodobieństwo uniknięcia błędu drugiego rodzaju, czyli przyjęcia hipotezy zerowej, gdy w rzeczywistości jest ona fałszywa.

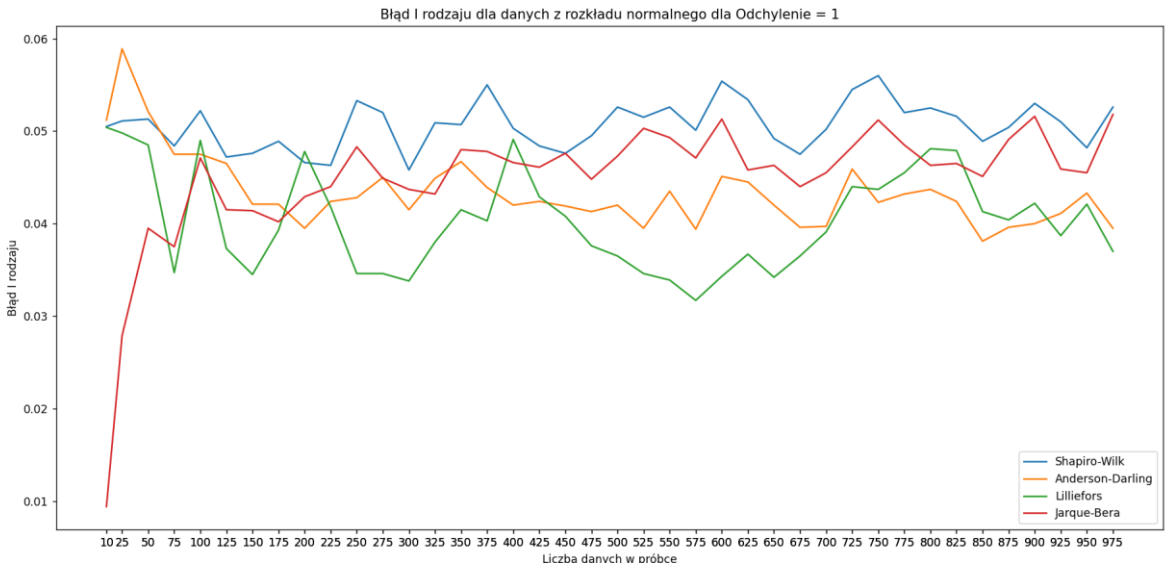
W naszej pracy zbadamy moc testów na podstawie następujących rozkładów – (rozkład normalny, rozkład t-studenta, rozkład lognormalny, rozkład gamma). W zależności od danego rozkładu będziemy brać różne parametry (odchylenie standardowe, liczba stopni swobody, skala, wielkość próby). Dla każdej kombinacji parametrów sprawdzimy normalność 10.000 razy. Moc testu będzie liczona w następujący sposób l/10000 gdzie l to liczba odrzuconych hipotez zerowych.

Rozkład normalny:



$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

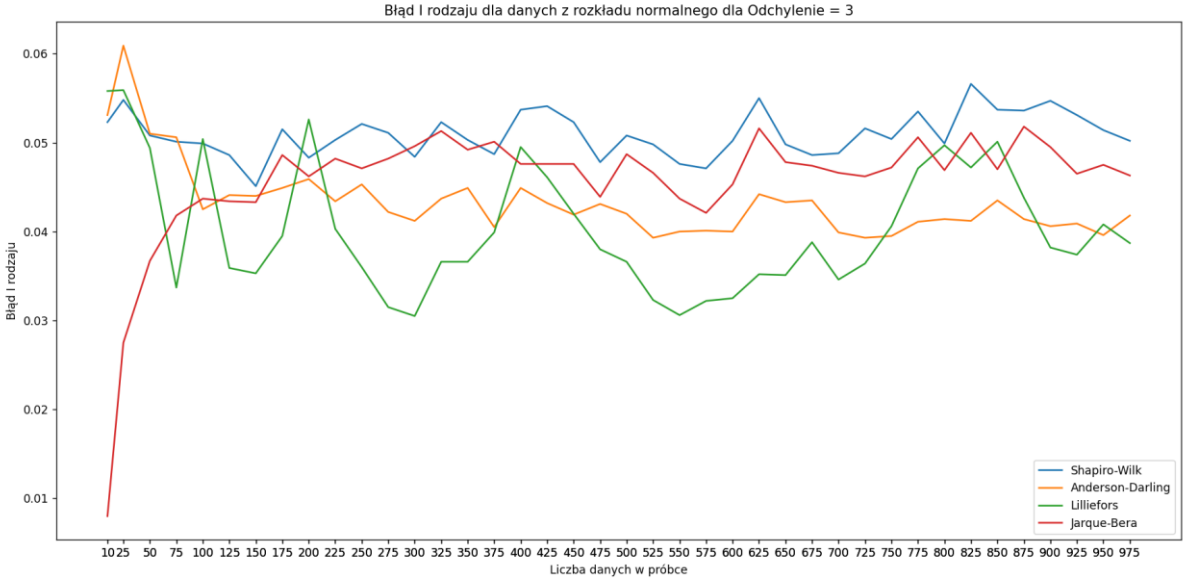
Błąd I rodzaju jest to odrzucenie hipotezy, która jest prawdziwa.



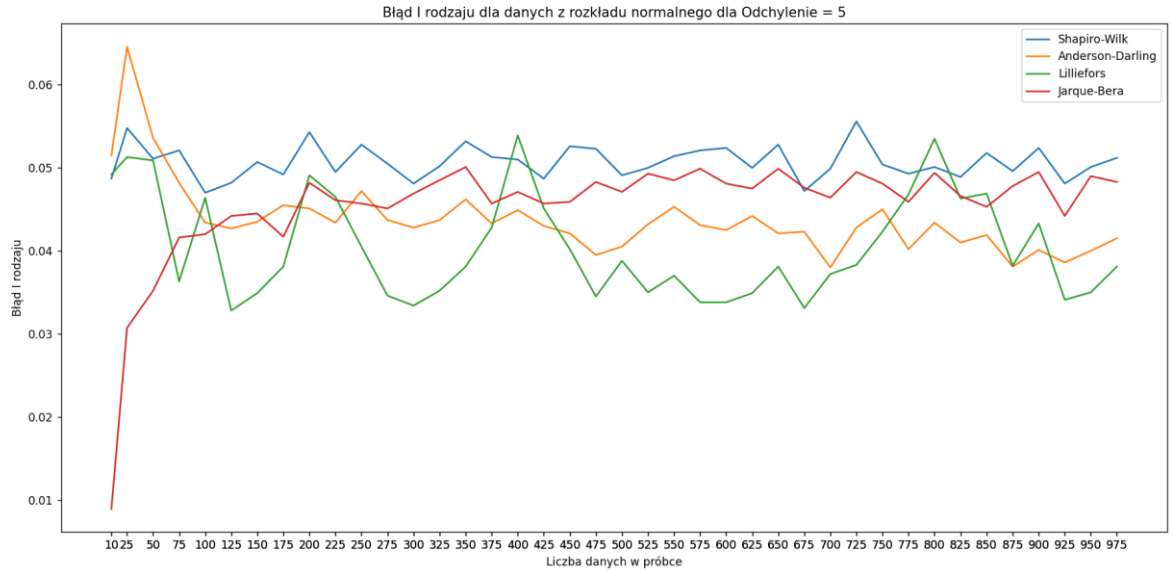
Przeprowadzone badanie ujawnia parę ciekawych aspektów. Po pierwsze, od około 50 danych w próbce, najwięcej błędu I rodzaju popełnia test Shapiro-Wilka. Trochę lepszy jest w tym przedziale test Jarque-Bera. Test Andersona-Darlinga jest porównywalny do testu Shapiro-Wilka do około 350

danych w próbce, a następnie staje się od niego lepszy. Dla mniejszej ilości danych w próbce (do około 50) test Andersona-Darlinga jest najgorszy ze wszystkich dostępnych, a najmniej błędu I rodzaju popełnia test Jarque-Bera.

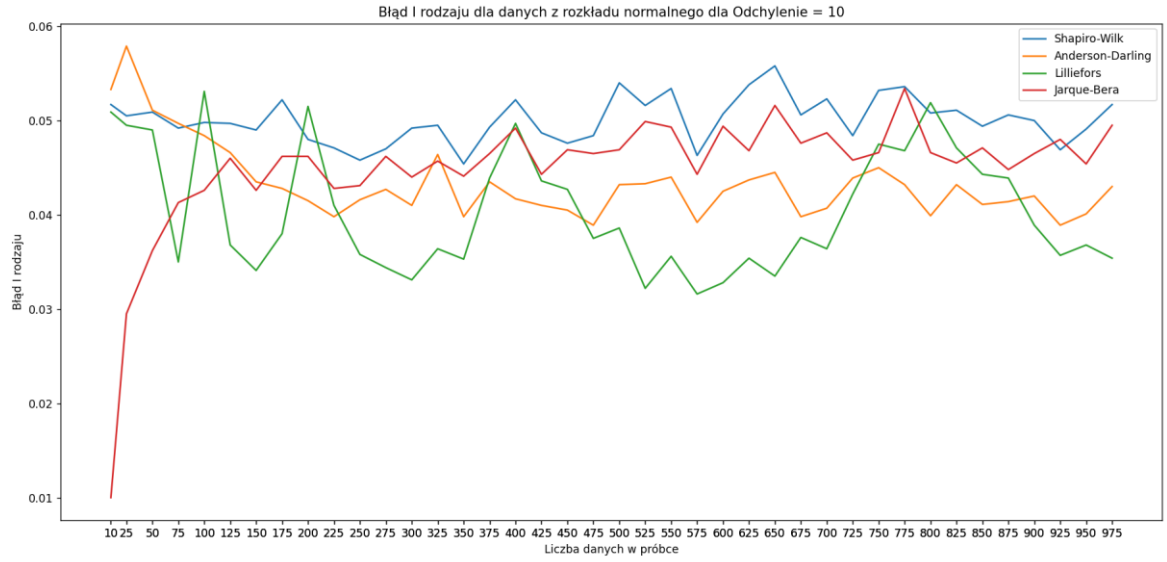
Test lilliforsa jest najbardziej nieprzewidywalnym testem ze wszystkich badanych. Potrafi przez dodanie o 25 danych do próby zmienić procent popełnionego błędu o około 2%. Należy do przedziałów, gdzie jest zarówno najlepszy, jak i najgorszy.



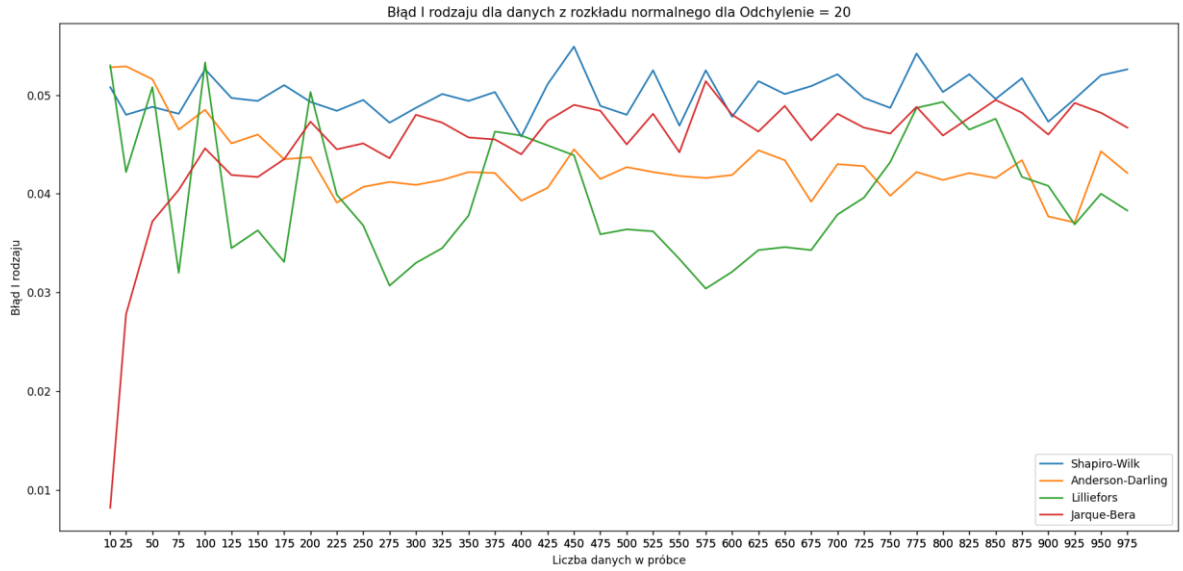
Przedstawione badanie można opisać identycznie jak dla sprawdzania odchylenia standardowego = 1.



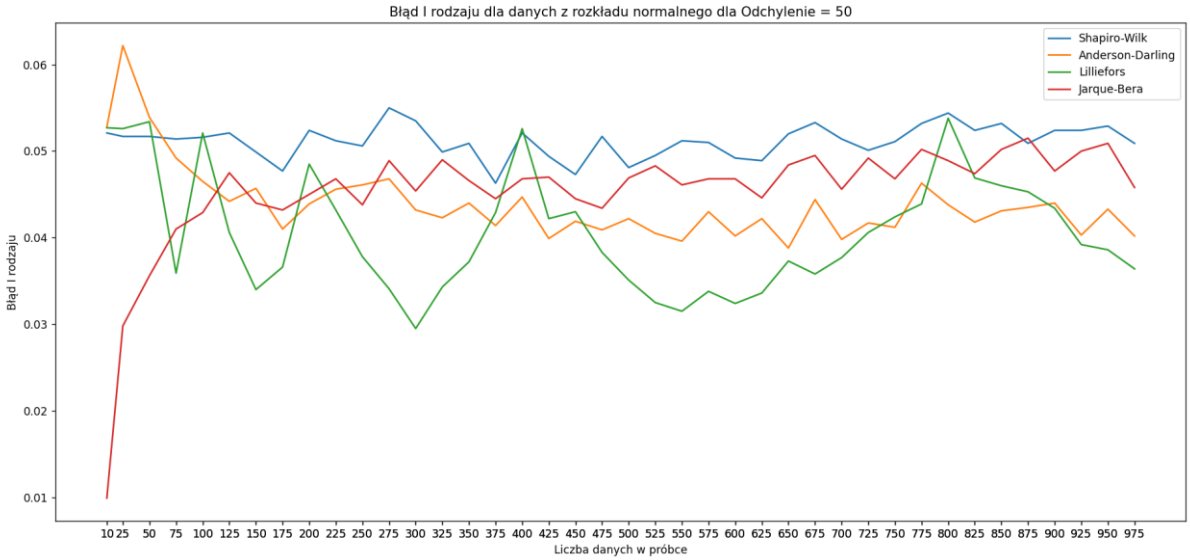
Przedstawione badanie można opisać identycznie jak dla sprawdzania odchylenia standardowego = 1.



Przedstawione badanie można opisać identycznie jak dla sprawdzania odchylenia standardowego = 1.



Przedstawione badanie można opisać identycznie jak dla sprawdzania odchylenia standardowego = 1.

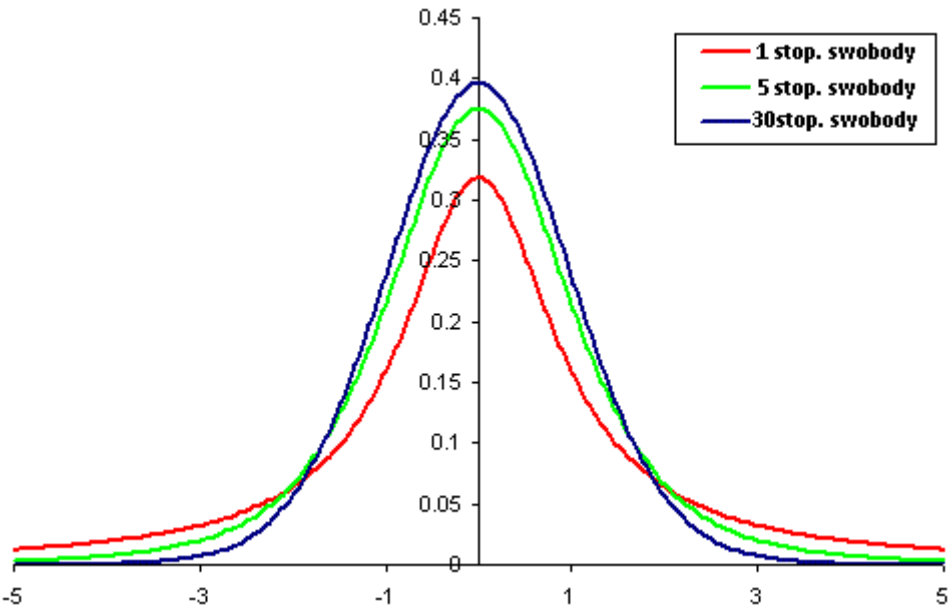


Przedstawione badanie można opisać identycznie jak dla sprawdzania odchylenia standardowego = 1.

Podsumowanie rozkładu normalnego

Różne wartości odchylenia standardowego nie mają żadnego istotnego wpływu na pojawianie się błędu pierwszego rodzaju. Pojawiające się błędy dla Shapiro-Wilka wahają się w okolicy 5%, czyli zastosowanego poziomu istotności. Od około 150 danych w próbie, test Jarque-Bera także oscyluje w okolicy 5%.

Rozkład t-studenta

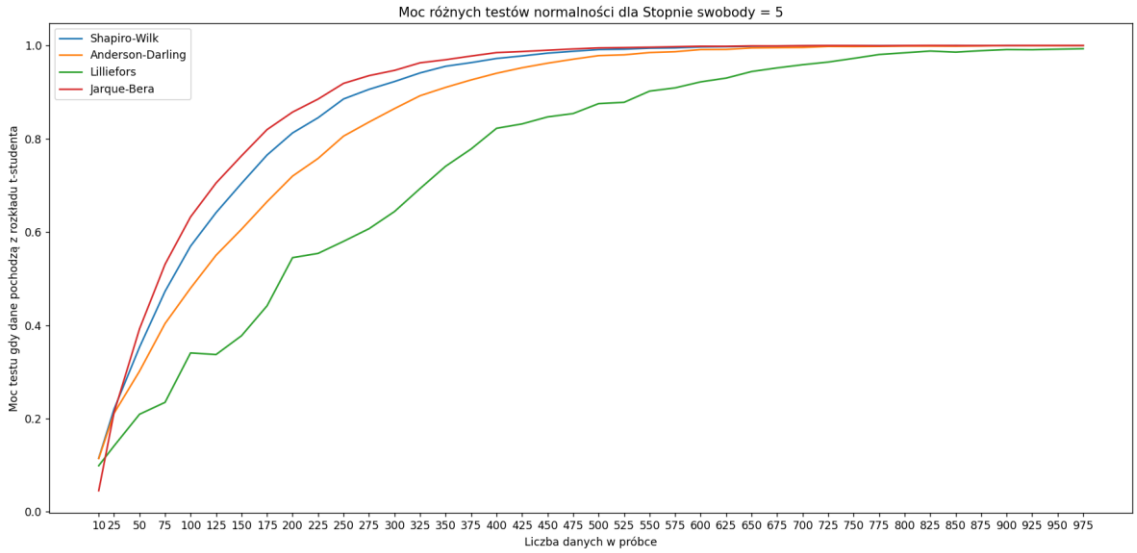


Rozkład Studenta z n stopniami swobody jest rozkładem zmiennej losowej T postaci:

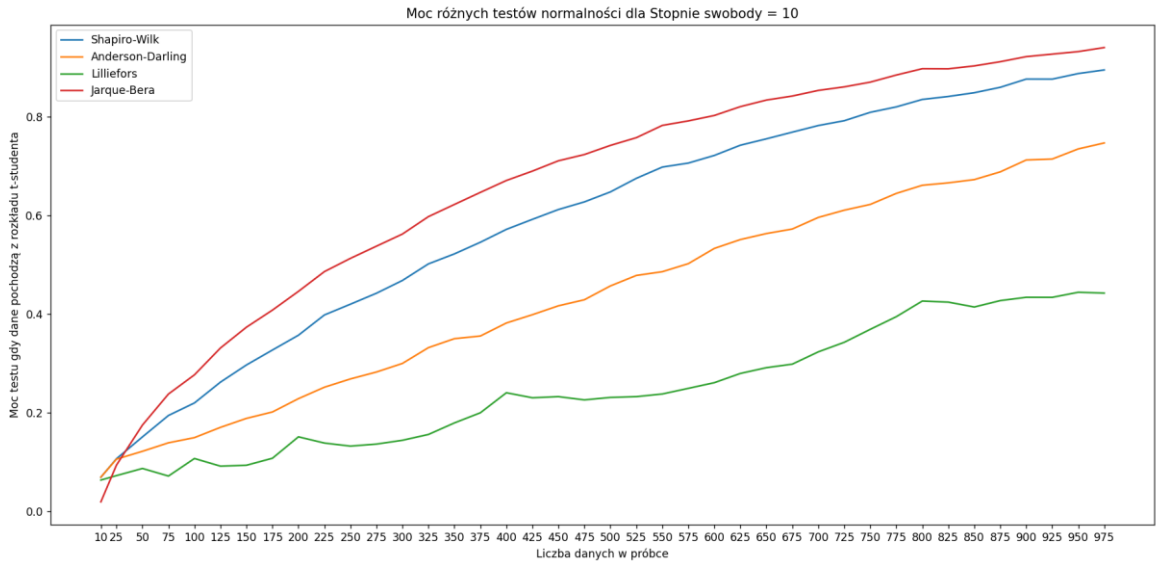
$$T = \frac{U\sqrt{n}}{\sqrt{Z}}$$

Gdzie,
U jest zmienną losową mającą standardowy rozkład normalny N (0 , 1)
Z jest zmienną losową o rozkładzie chi kwadrat o n stopniach swobody
U i Z są niezależne.

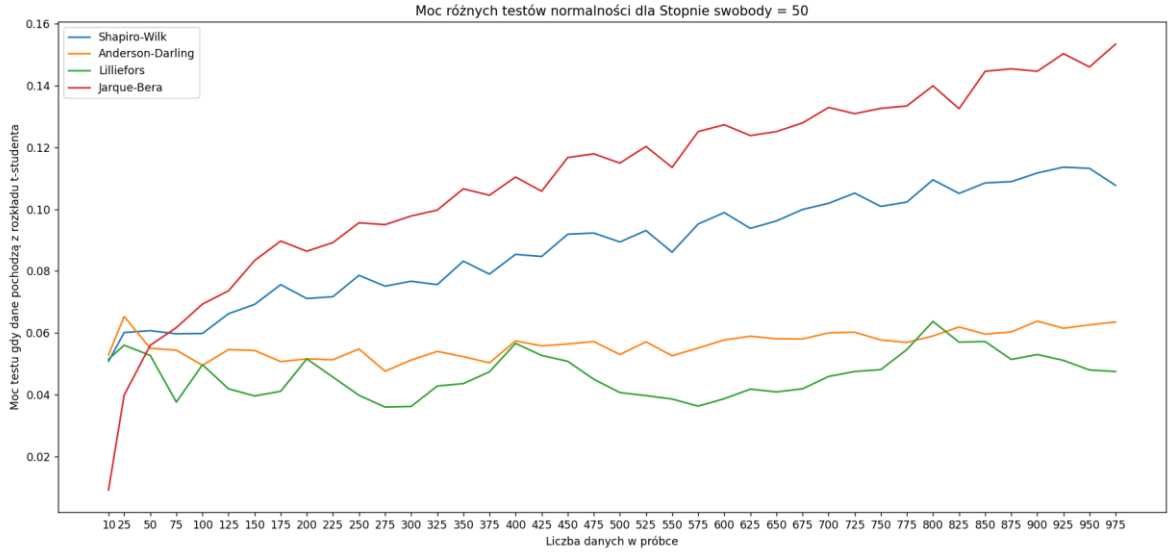
Wszystkie testy szybko zbiegają do 1. Rozkład t-studenta dla 2 stopni swobody ewidentnie jest różny od rozkładu normalnego, co wykryły wszystkie testy. Zgodnie z oczekiwaniami większa liczba danych zwiększa moc testu, gdyż większa próbka w lepszy sposób odwzorowuje populacje.



Wzrost stopni swobody na 5, istotnie zwiększył problemy w rozpoznawaniu rozkładu, jednakże przy 250 danych w próbce, ten problem w większości testów nie jest istotny. Jedynie test Lillieforsa jest mocny od około 500 danych w próbie. Co ciekawe, test Lillieforsa jest silniejszy od Jarque-Bera i prawdopodobnie najsilniejszy ze wszystkich przy bardzo małej próbie (mniejszej od 10). Pomimo to, w praktycznie całym badaniu test Jarque-Bera jest najmocniejszym testem i porównywalny z testem Shapiro-Wilka.

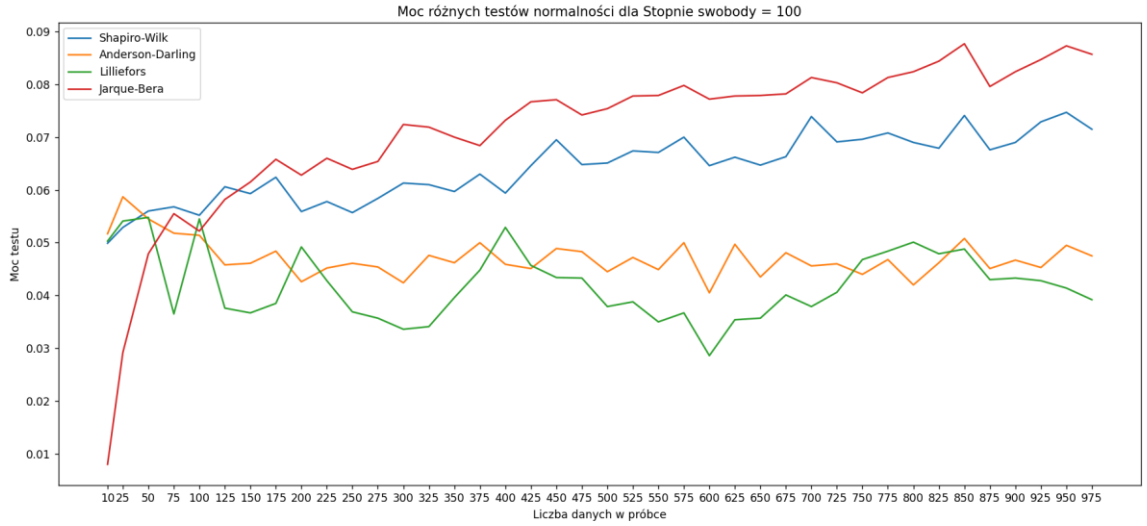


Podwojenie liczby stopni swobody na tyle przybliżyło rozkład t-studenta do rozkładu normalnego, że żaden test nie osiągnął mocy równej 1. W tym przypadku, widać ewidentne rozwarstwienie się mocy testów, gdzie na prowadzenie wyszedł test Jarque-Bera. Dosyć podobnie wypada test Shapiro-Wilka, jednakże ten, dla wzrostu ilości danych w próbie, ale jednocześnie przy małej ilości tych danych oddala się od testu Jaque-Bera, a następnie przy większej ilości danych, dla wzrostu ilości danych w próbie, zbliża się do tego testu. Funkcja mocy tych dwóch testów od liczebności próby w badanym przedziale mają postać zbliżoną wyglądem do funkcji logarytmicznej. W przypadku testu Andersona-Darlinga moc zwiększa się liniowo, jednakże tempo tego wzrostu jest istotnie mniejsze od dwóch wcześniej omawianych. Test Lilliefors jest podobny do testu Andersona-Darlinga z taką różnicą, że moc tego testu rośnie jeszcze wolniej, oraz dodatkowo występują lokalne gwałtowne zmiany mocy.

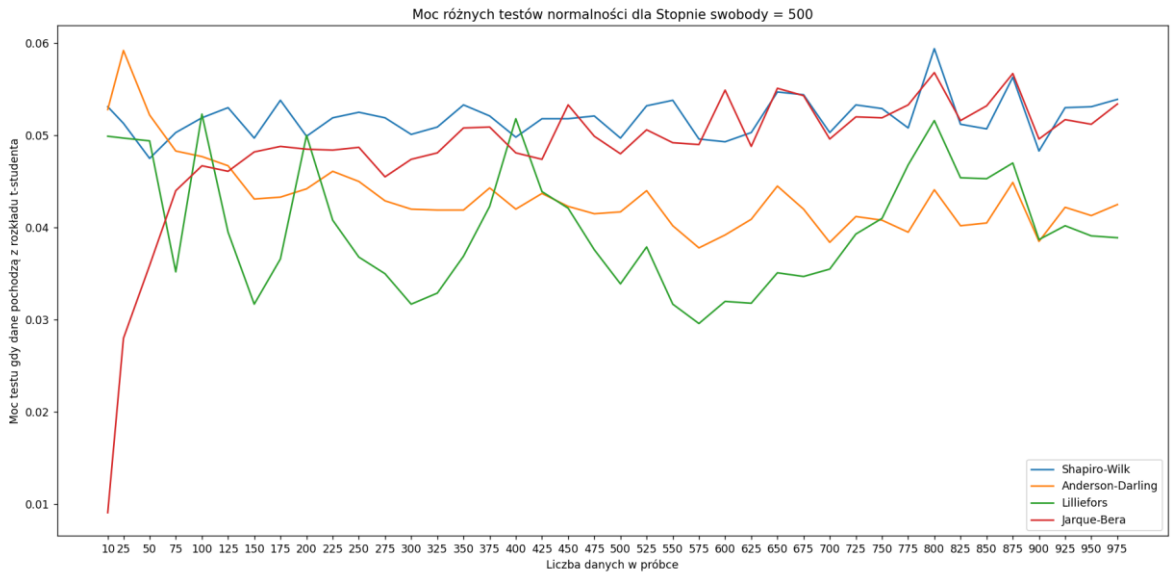


Rozkład t-studenta o 50 stopniach swobody jest bardzo trudny do rozróżnienia od rozkładu normalnego, zarówno metodą na oko, jak i testami. Żaden test w badanym przedziale danych w próbce nie osiągnął progu 20% mocy testu. Każdy test w większości przypadków uznawał badany rozkład za normalny. Dla bardzo małej ilości danych, ciężko jest wyróżnić najlepszy test, za to ewidentnie najgorszym jest test Jarque-Bera. Ten test bardzo szybko zwiększa moc

wychodząc na prowadzenie. Dosyć porównywalnym testem dla nieco większej próbki danych jest test Shapiro-Wilka. Oba te testy, wraz ze wzrostem liczebności próbki, zwiększają moc. Dwa pozostałe testy nie wykazują jednoznacznej tendencji wzrostowej. Test Andersona-Darlinga stabilnie oscyluje w okolicy 5,5% mocy testu. Test Lilliforsa jest dużo bardziej niestabilny. Występują u niego liczne nagłe zmiany mocy, pomimo których to widać, że ten test jest trochę słabszy od testu Andersona-Darlinga.

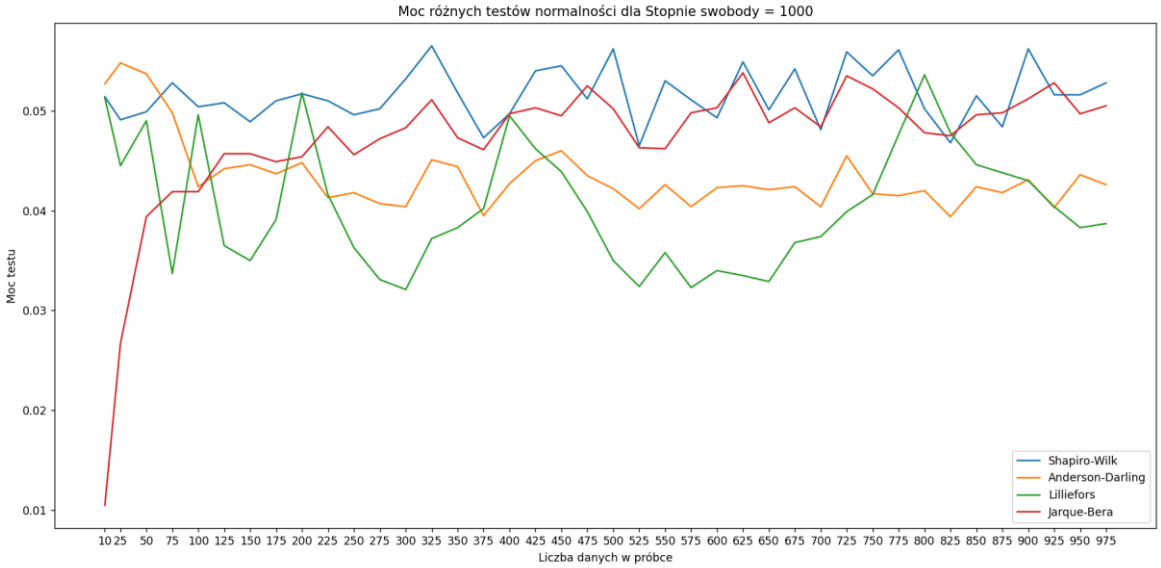


Dalsze zwiększanie stopni swobody w teście t-studenta coraz bardziej upodabnia ten rozkład do rozkładu normalnego. Moc testów stała się bardziej porównywalna, a szczególnie dwa najgorsze – Lillieforsa i Andersona-Darlinga. Do około 150 danych w próbce nie ma jednoznacznie najmocniejszego testu, a większość jest na porównywalnym poziomie. Dla większej ilości danych w próbce uwidacznia się najmocniejszy test – Jarque-Bera.

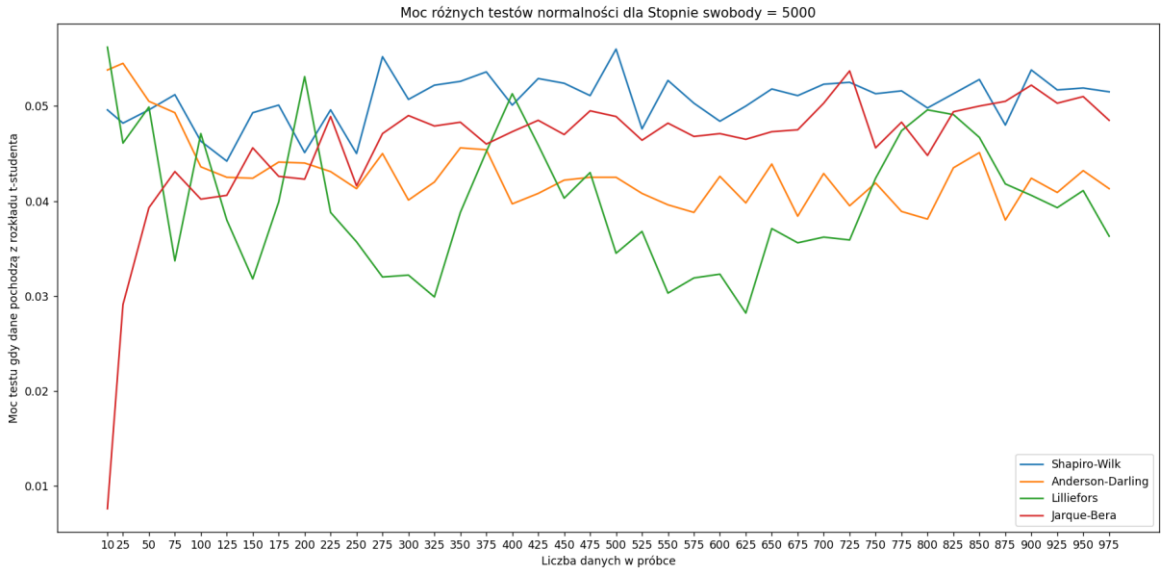


Rozkład t-studenta o 500 stopniach swobody uniemożliwia testom skutecznego rozpoznawania rozkładu. Każdy z badanych testów wykazuje podobną moc i bywa dla pewnej liczby danych w próbce najlepszym ze wszystkich. Test Lillieforsa jeszcze bardziej uwydatnia swoje nagłe zmiany

mocy. Dodatkowo, test Andersona-Darlinga wydaje się, jakby wraz ze wzrostem liczby danych w próbce, zmniejszał swoją moc.



Przeprowadzone badanie dla rozkładu t-studenta o 1000 stopniach swobody posiada bardzo podobne wyniki względem rozkładu t-studenta o 500 stopniach swobody.



Przeprowadzone badanie dla rozkładu t-studenta o 5000 stopniach swobody posiada bardzo podobne wyniki względem rozkładu t-studenta o 500 stopniach swobody.

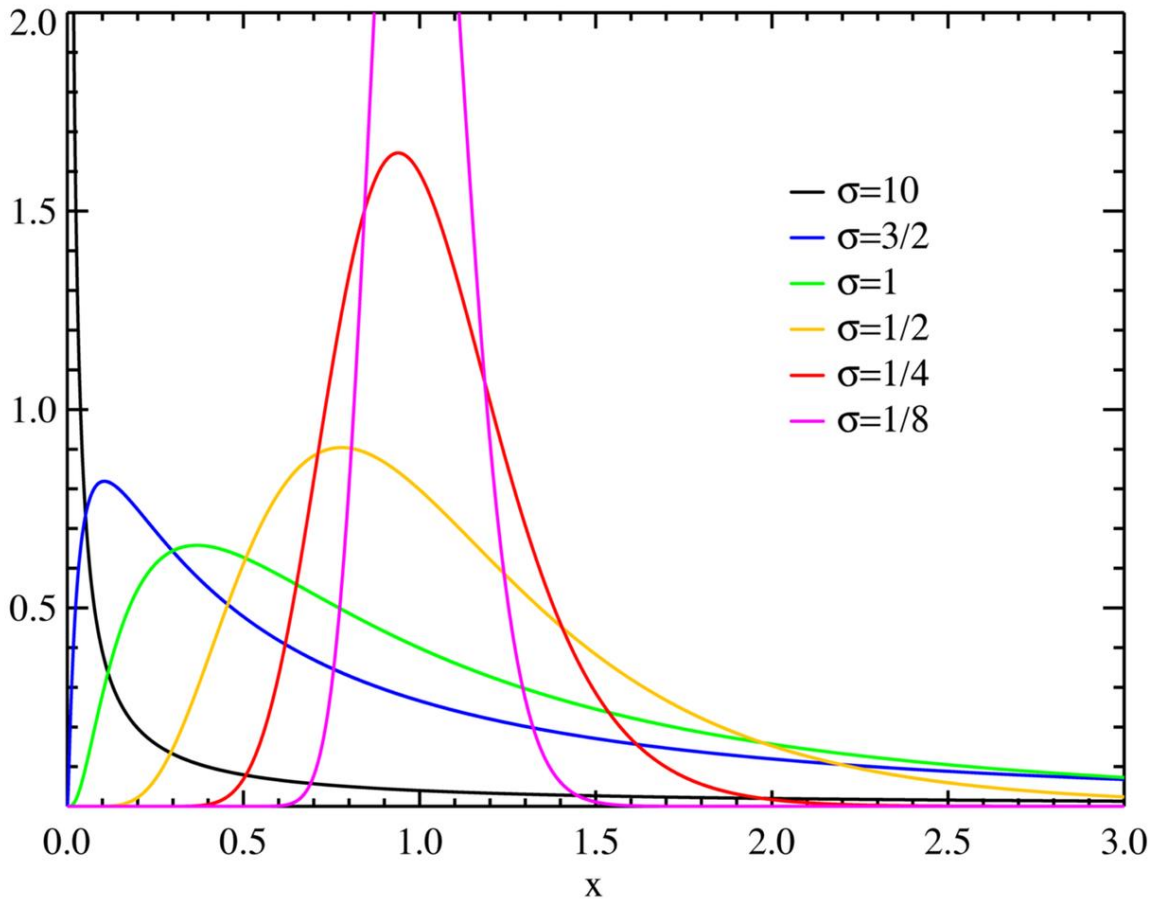
Podsumowanie rozkładu t-studenta

Przeprowadzone badanie ukazuje, że liczba stopni swobody w rozkładzie t-studenta istotnie wpływa na popełnianie błędów II rodzaju przebadanych testów. Wraz ze wzrostem liczby stopni swobody, testy normalności częściej uznają rozkład t-studenta za rozkład normalny. Dodatkowo, wzrost liczby st. swobody wizualnie wypłaszcza funkcje mocy, zależne od liczby danych w próbie, a test Lillieforsa jeszcze bardziej uwydatnia swoje nagłe zmiany mocy. To ostatnie spostrzeżenie może być spowodowane tym, że lewa oś zmienia się w zależności od przedziału uzyskiwanych wartości przez funkcję mocy i nie zawsze posiada te same przedziały. Tą tezę

umacnia fakt, że wahania tego testu wynoszą zazwyczaj około 2% mocy. W takim wypadku, nie można określić, że test Lillieforsa zwiększa przedział wahań, ale jego wahania wraz ze wzrostem liczby stopni swobody, stają się istotniejszą częścią generowanych wyników.

Rozkład lognormalny:

Rozkład lognormalny dla średniej 0 i różnym odchyleniu



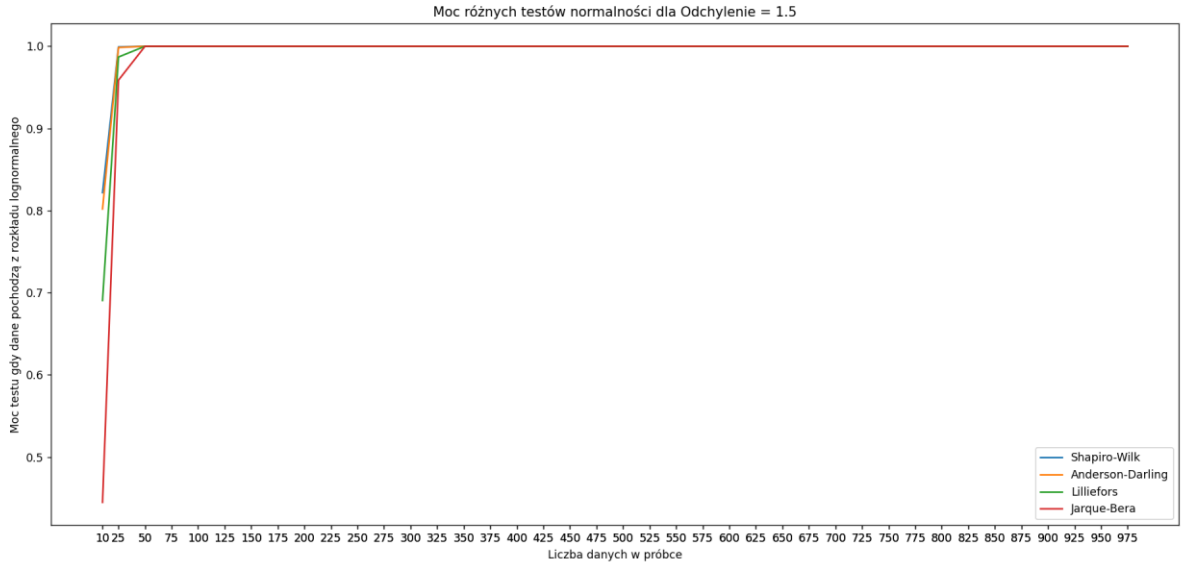
Niech X będzie zmienną losową przyjmująca wartości dodatnie. Zmienna ta ma rozkład lognormalny z parametrami μ (wartość oczekiwana) i σ^2 (odchylenie standardowe), gdy zmienna Y = lnX ma rozkład normalny z parametrami μ i σ^2.

Gęstość rozkładu lognormalnego

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), & x > 0 \\ 0, & x \leqslant 0 \end{cases}$$

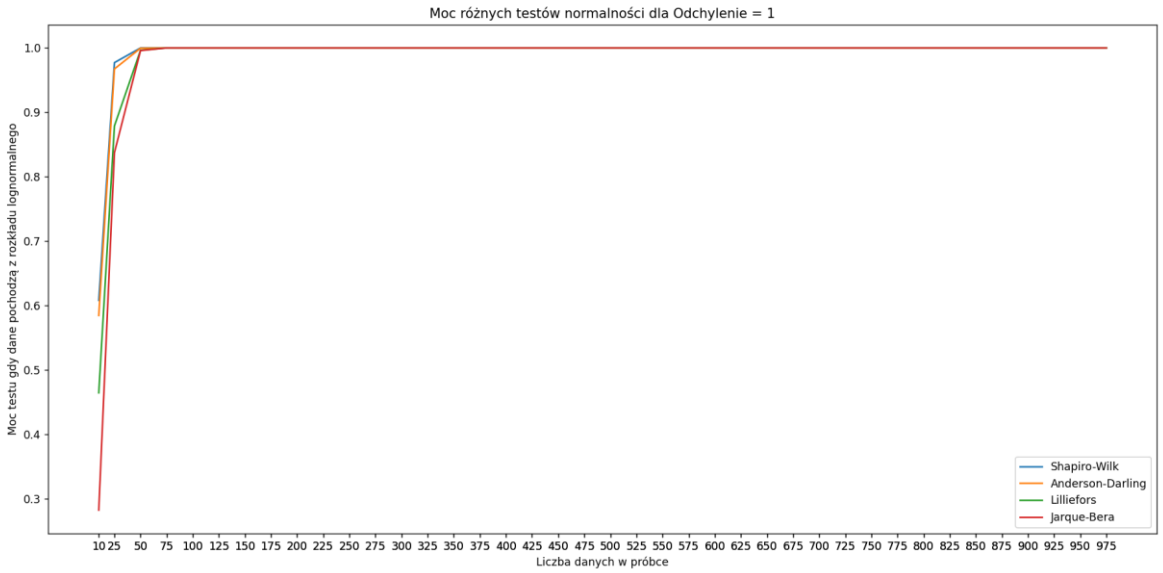
Do naszego badania będziemy brać średnią 0, a odchylenie standardowe będzie zmieniane. Odchylenie będzie z listy [3/2,1,1/2,1/4,1/8,1/16,1/32]. W miarę zmniejszania odchylenia rozkład lognormalny powinien przypominać rozkład normalny.

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,1.5)



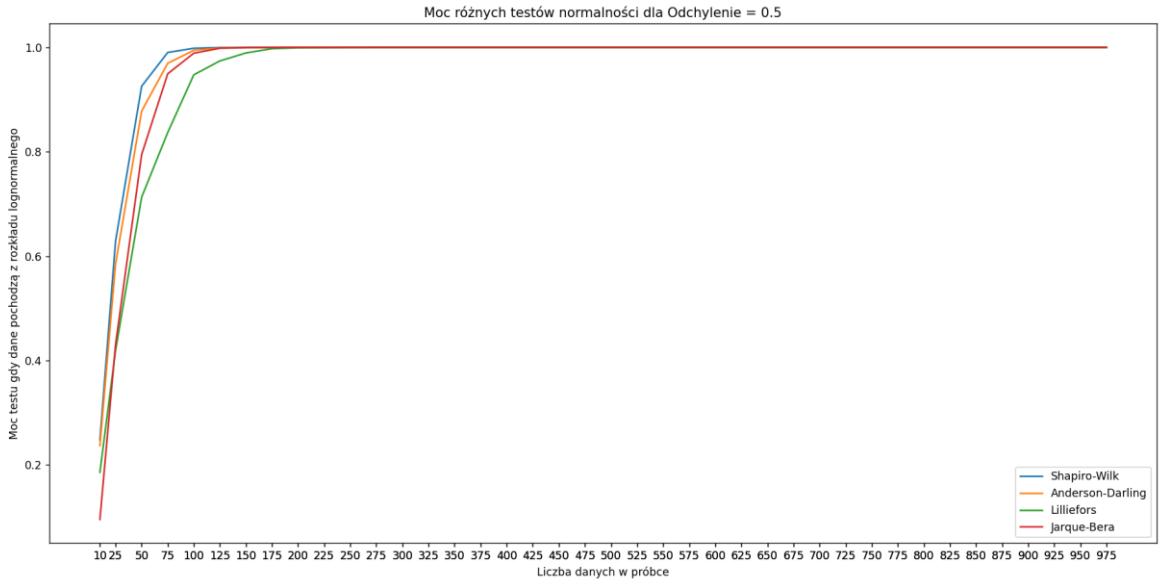
Jak widać moc testu szybko zbiega do 1. Rozkład lognormalny(0,3/2) jest mocno prawoskośny wszystkie testy dobrze zinterpretowały dane wejściowe.

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,1)



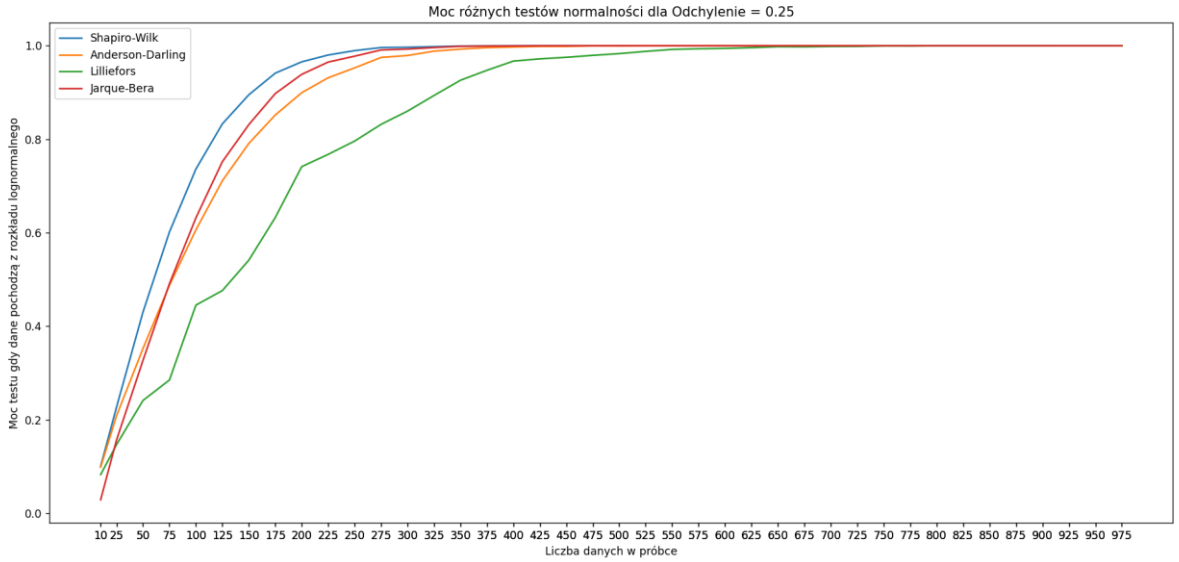
Podobnie jak w poprzednim przypadku widać moc testu szybko zbiega do 1. Rozkład lognormalny(0,1) jest prawoskośny. Testy dobrze zinterpretowały dane wejściowe. Wszystkie wypadły podobnie.

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,0.5)



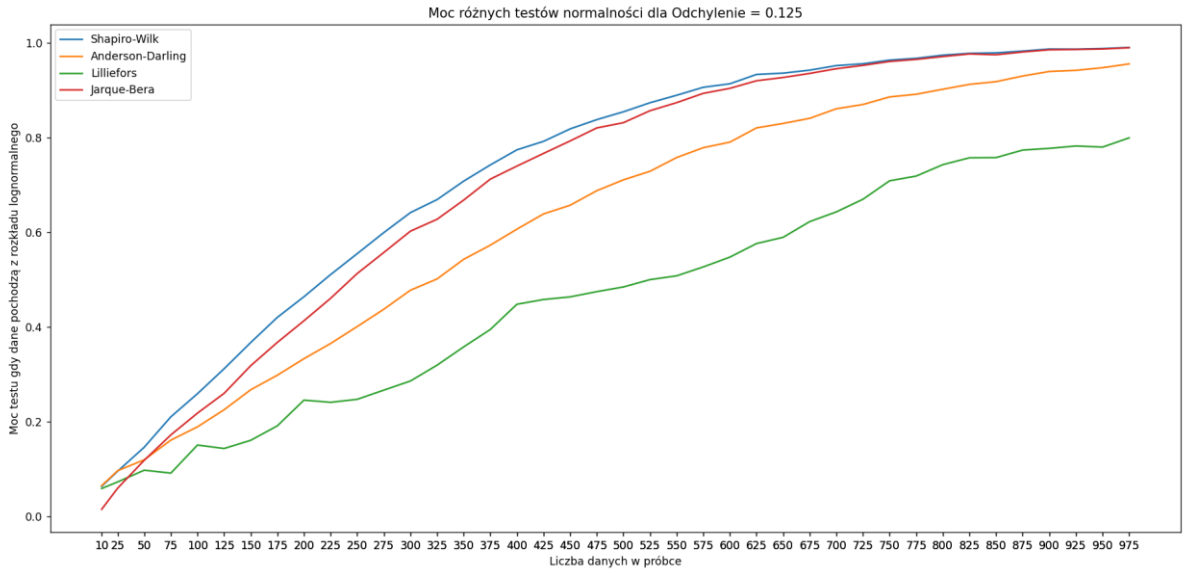
Wynik zdaję się być zaskakujący gdyż rozkład lognormalny z odchyleniem ½ jest stosunkowo podobny do rozkładu normalnego, jest odrobine prawoskośny. Testy natomiast dosyć szybko zaczęły odrzucać h0 o normalności danych. Najgorzej wypadł test Lillieforsa, który potrzebował 175 danych aby za każdym razem odrzucić h0. Wraz ze wzrostem liczby obserwacji moc testu się poprawia, a większa próba powinna lepiej odwzorowywać przybliżony rozkład.

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,1/4)



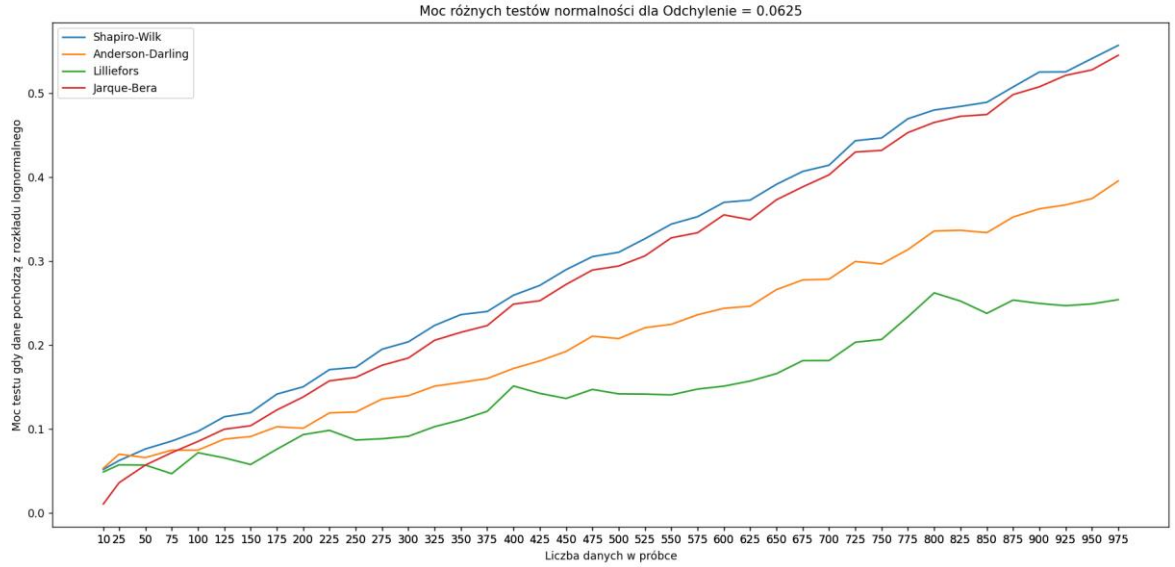
Zauważalne jest istotne rozwarstwienie funkcji mocy osiąganych testów, przy czym ponownie najgorzej wypadł test Lillieforsa. Pozostałe testy wyszły porównywalnie. Widać, że testy potrzebują większej ilości danych w próbce, aby móc odrzucić hipotezę zerową względem poprzednich badań. Pomimo to, każdy wynik funkcji mocy testu zbiega do 100% (czyli wartości maksymalnej).

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,1/8)



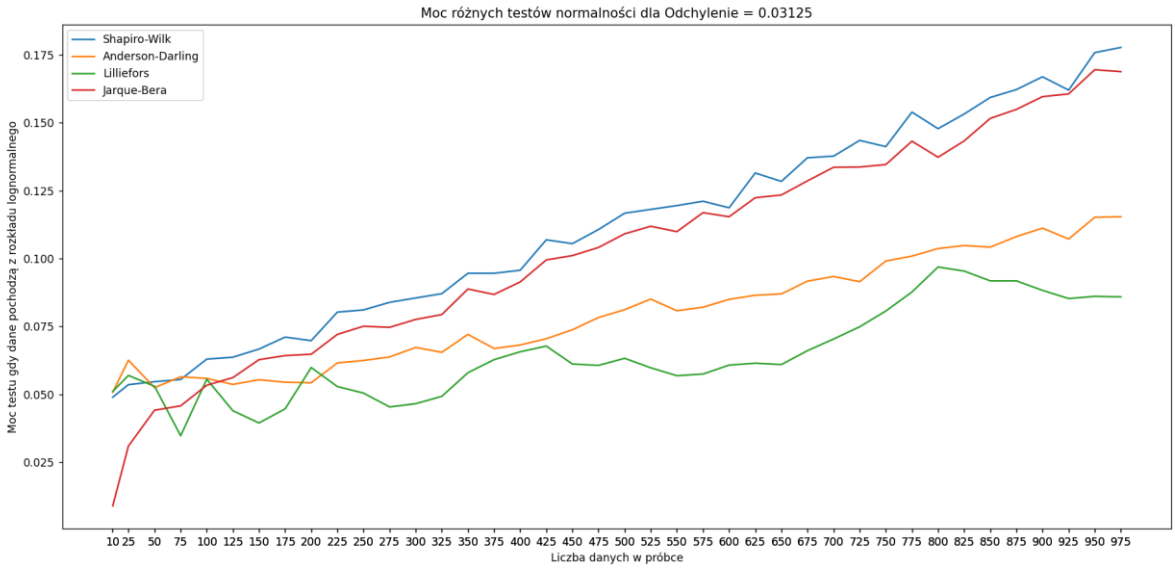
Jest to pierwsza sytuacja, gdzie moc testu nie zawsze zbiega do jedynki. Widać, że testy zaczynają mieć problem z rozróżnianiem tego, czy dane pochodzą z rozkładu normalnego. Tak niskie odchylenie powoduje większą koncentracje (kurtoza >0) wokół średniej, co widocznie powoduje błędną interpretację przez testy. Ponownie najgorzej wypadł test Lillieforsa, jednakże najlepiej wypadły testy Jarque-Bera i Shapiro-Wilka.

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,1/16).



Moc testów, dla tak niskiego odchylenia standardowego, powoduje, że przeprowadzone testy, są coraz bardziej skłonne zaklasyfikować, że dane pochodzą z rozkładu normalnego. Przy tak niskim odchyleniu standardowym, rozkład lognormalny, przypomina rozkład normalny o wysokiej kurtozie, stąd zauważalne są niskie poziomy mocy testów. Zdecydowanie górują testy Shapiro-Wilka i Jarque -Bera, a ponownie najgorzej wypadł test Lillieforsa. Dla każdego testu, zauważalny jest trend wzrostowy mocy testu.

Moc testów normalności dla danych z rozkładu lognormalnego(0 ,1/32)



Zauważalny jest fakt, iż dla przeprowadzonego badania, zastosowane testy nie są w stanie rozróżnić rozkładu lognormalnego od rozkładu normalnego. Ponownie ranking testów układa się w tej samej kolejności – Shapiro-Wilk, Jarque-Bery, Anderson-Darling, Lilliefors. Zauważalna jest także zależność, że wraz ze wzrostem liczby danych w próbce, moc testów rośnie.

Podsumowanie rozkładu lognormalnego:

Im mniejsze odchylenie w rozkładzie lognormalnym tym rozkład ten bardziej przypomina rozkład normalny co widać było na powyższych wykresach. Najgorszym testem okazał się być test Lillieforsa, dosyć kiepsko wyszedł też Anderson Darling. Zdecydowanie najlepszymi testami okazały się być test Shapiro-Wilka i Jarque-Bera. Im większa była liczebność próby tym większa była moc testu, co nie jest specjalnie zaskakujące gdyż, im większa próbka tym rozkład staje się bardziej podobny do teorytycznego. Oprócz tego dla większej ilości danych małe różnice stają się bardziej istotne statystycznie.

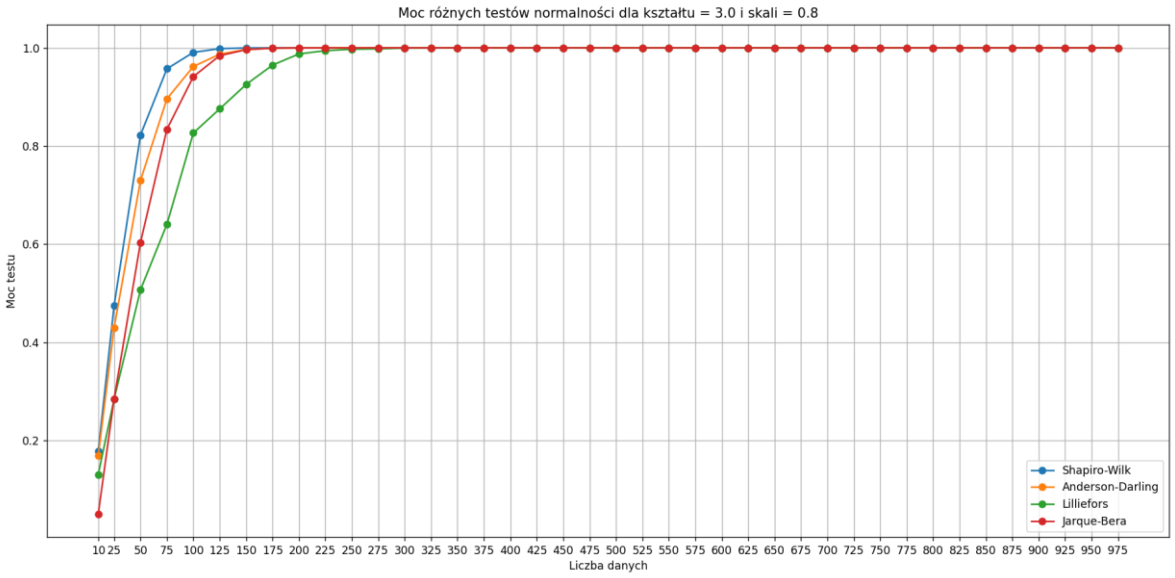
Rozkład Gamma

W tych badaniach, będziemy obserwowali, jak zmiana parametru kształtu i skali, zmieni moc wybranych testów. Aby przybliżyć rozkładem gamma rozkład normalny N(2,1) należy wziąć parametry k=4, θ=0.5.

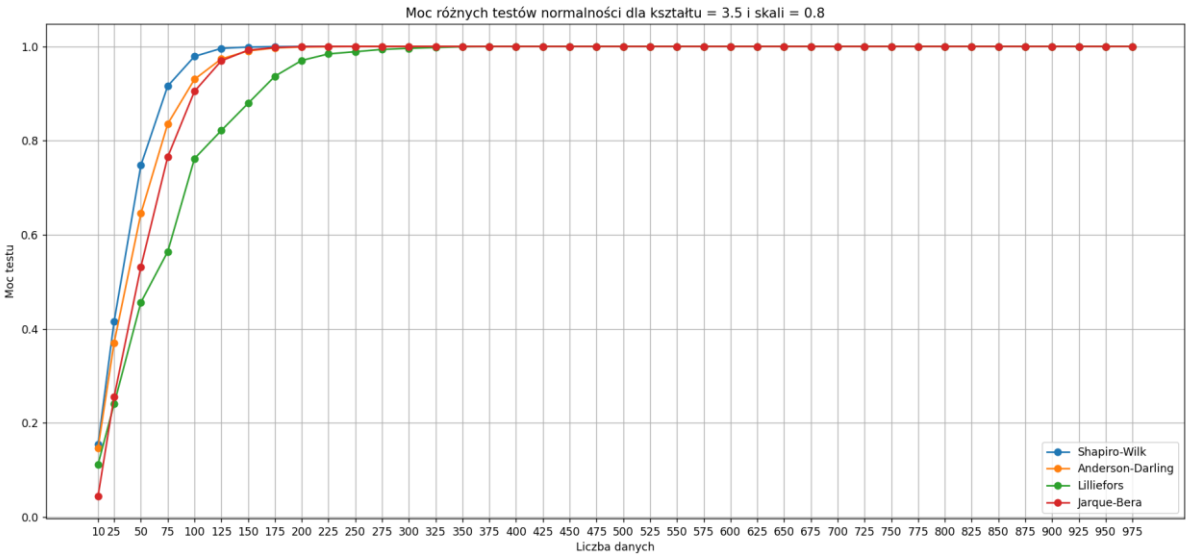
Rozkład gamma z parametrami kształtu k i skali θ ma średnią (wartość oczekiwaną) równą k*θ. Dlatego aby przyrównać do wartości oczekiwanej w rozkładzie normalnym równej u przyrównujemy k*θ = u. Wariancja rozkładu gamma jest równa $k*\theta^2$. Aby wariancja rozkładu gamma była równa wariancji rozkładu normalnego σ^2, ustawiamy: $k\theta^2=\sigma^2$. Podstawiając za u wartość 2, a za σ^2 wartość 1 otrzymujemy układ równań.
k*θ = 2
$k\theta^2$= 1
Górne równanie możemy pomnożyć przez θ otrzymując $k\theta^2 = 2\theta \Rightarrow 2\theta = 1 \Rightarrow \theta = 0.5$
Podstawiając pod pierwsze równanie otrzymamy k = 4.

Do naszej analizy weźmiemy wartości z listy k = [3,3.5,4,4.5] i θ = [0.35,0.5, 0.65, 0.8]

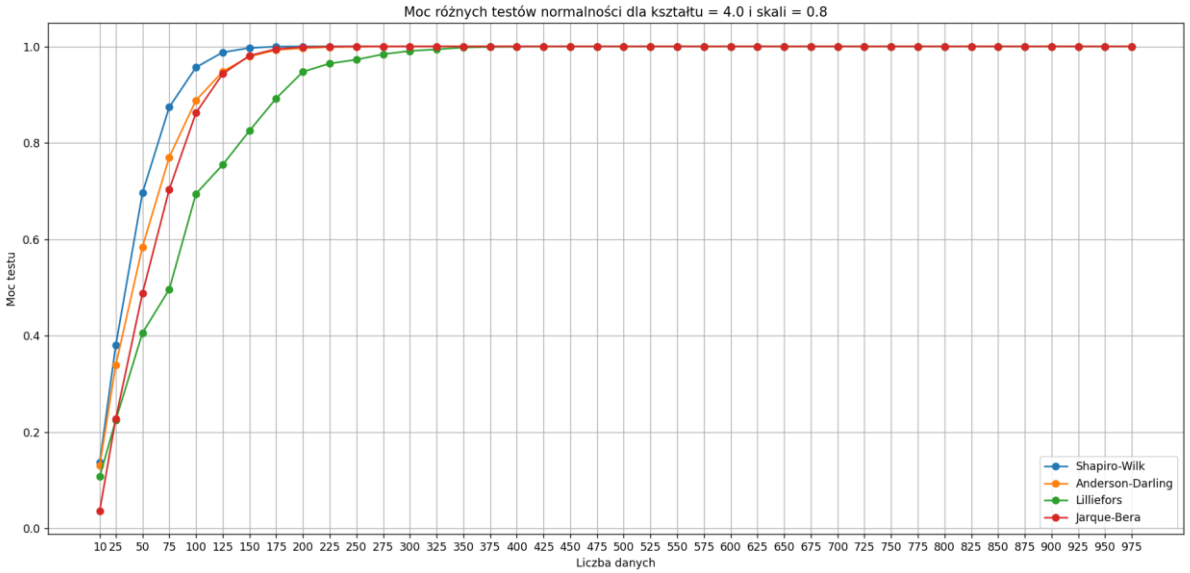
Moc testów w zależności od liczebności próbki i parametrów kształtu i skali, gdy dane pochodzą z rozkładu gamma



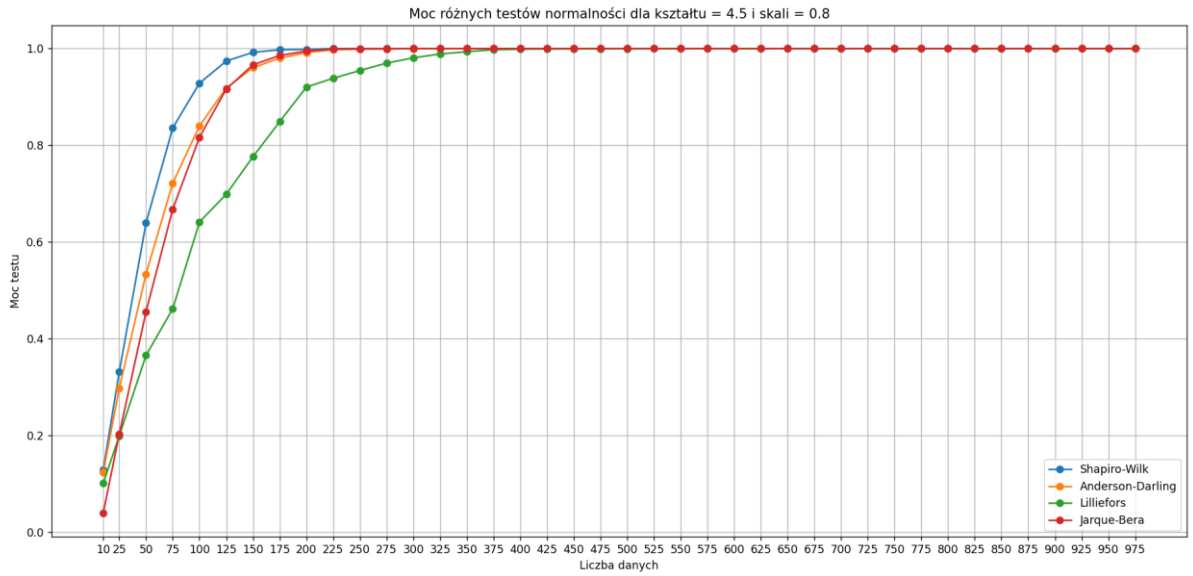
Jak można zauważyć, ponownie najgorzej wypada test Lillieforsa. Poza tym, wszystkie testy szybko zbiegają do 1. Najszybciej zbiega Shapiro-Wilk. Dla małej próbki testy nie wypadają najlepiej natomiast już dla 100 danych moc testu jest wysoka dla każdego testu.



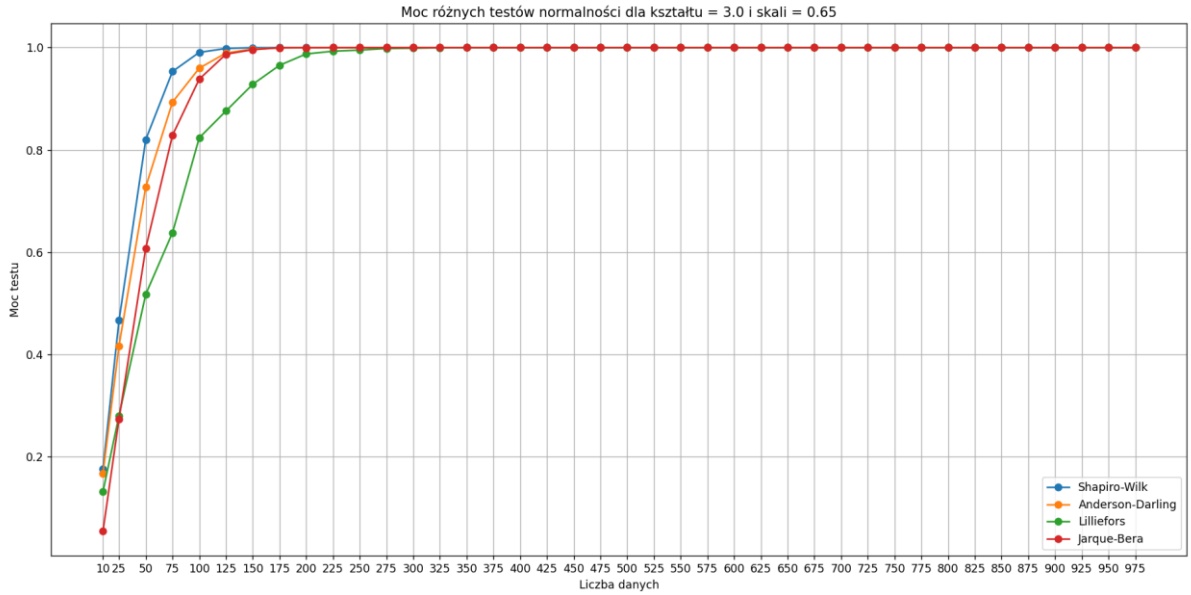
Moce testów wyglądają bardzo podobnie co w przykładzie gdy kształt wynosi 3 i skala 0.8. Natomiast nieco później Lilliefors zbiega do mocy testu równego 1.



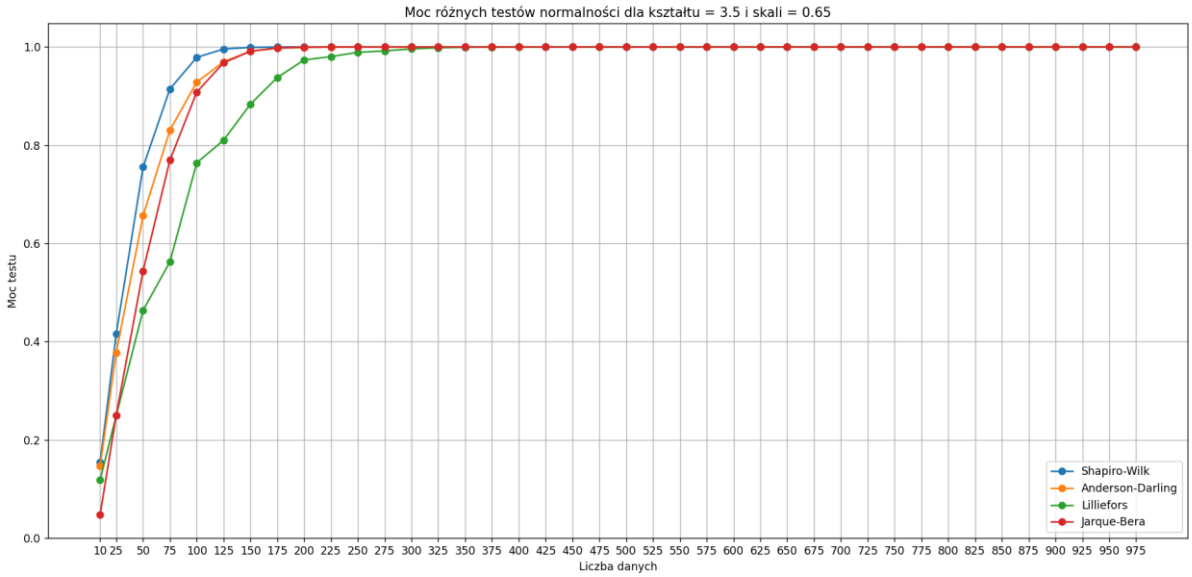
Wykres wygląda identycznie, natomiast test Anderon-Darling nieco później osiągnął moc testu większą od 0.8.



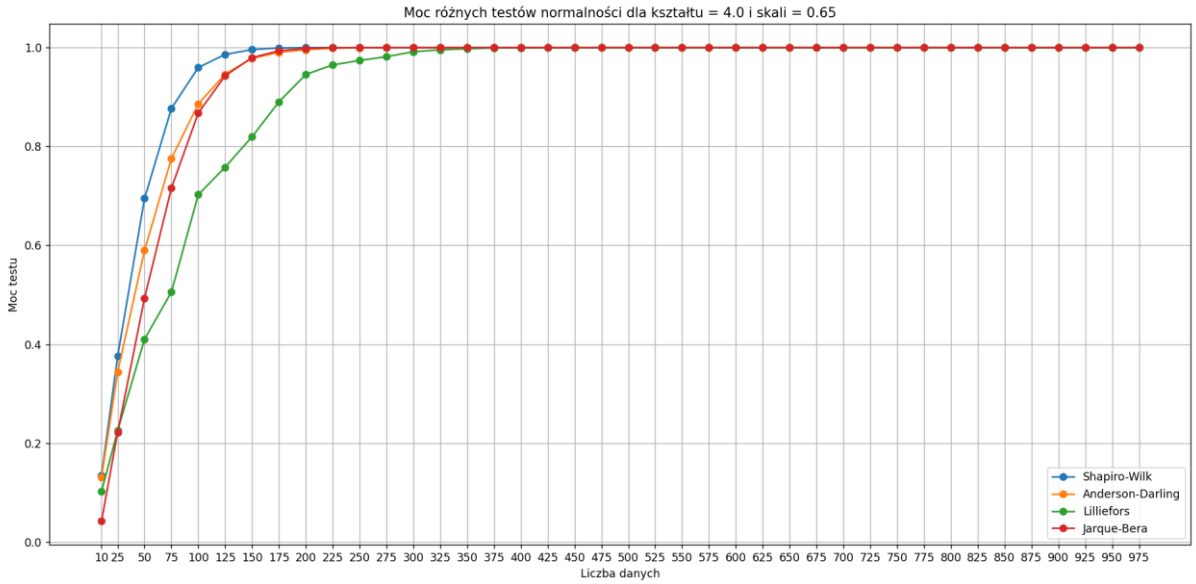
Moce testu wyglądają identycznie jak w poprzednim przypadku



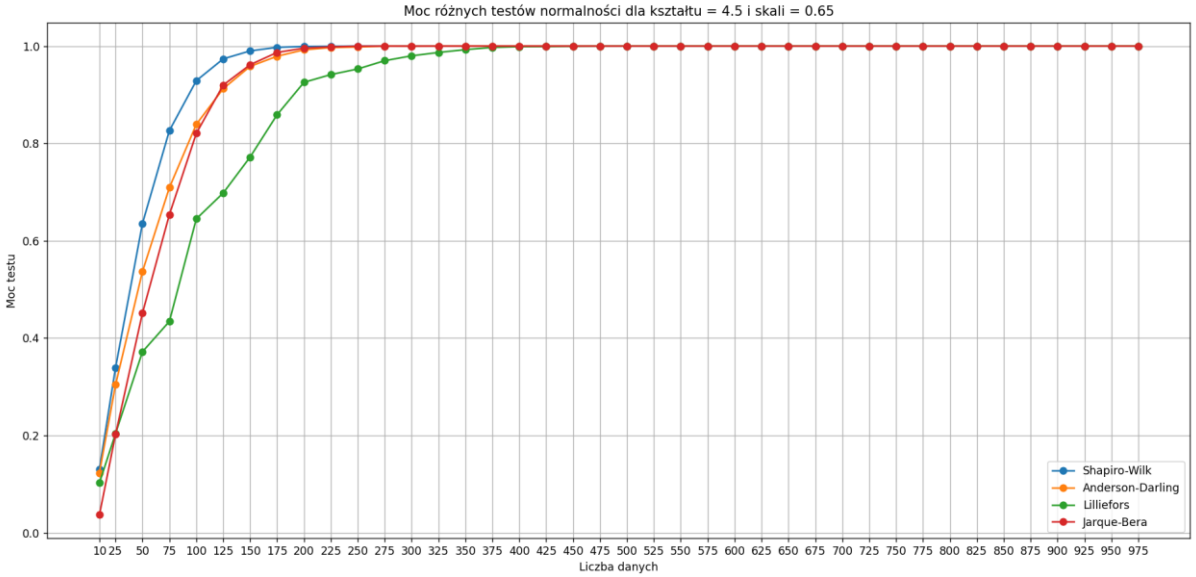
Testy bardzo szybko zbiegają do 1. Najgorzej wypada znowu Lilliefors, natomiast nie różni się specjalnie od innych.



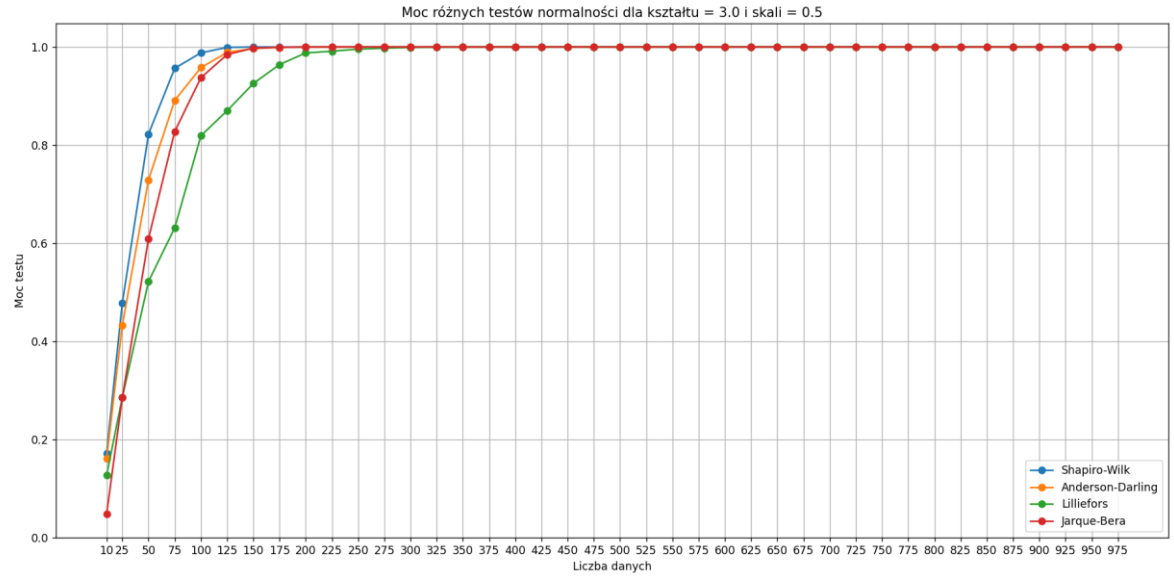
Shapiro-Wilk nieco szybciej zbiega do 1. Wykresy wyglądają identycznie.



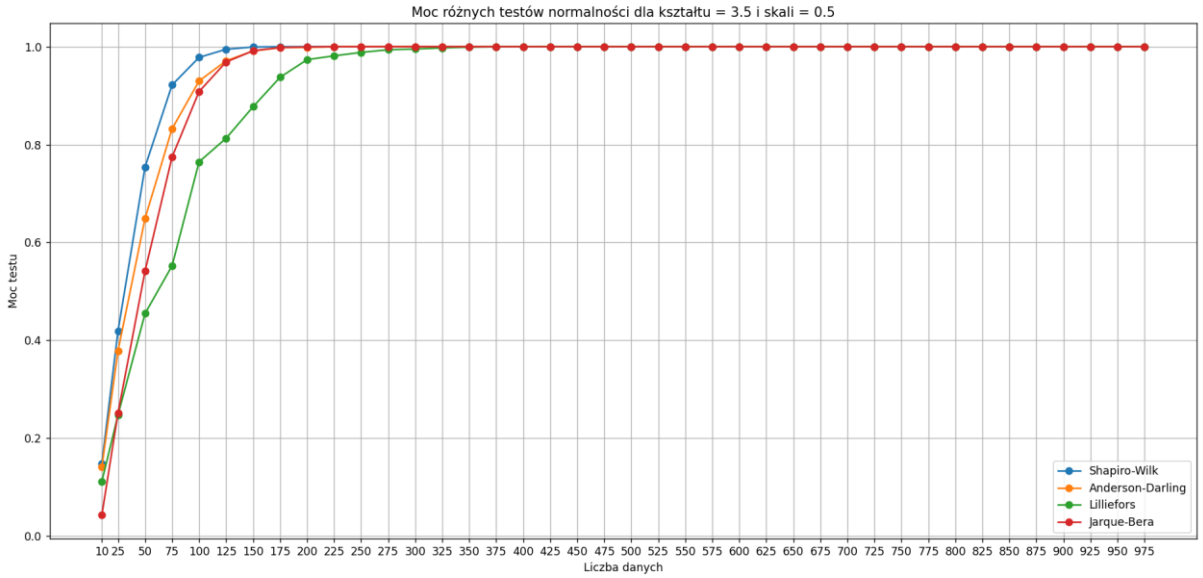
Wykres wygląda identycznie co poprzednio.



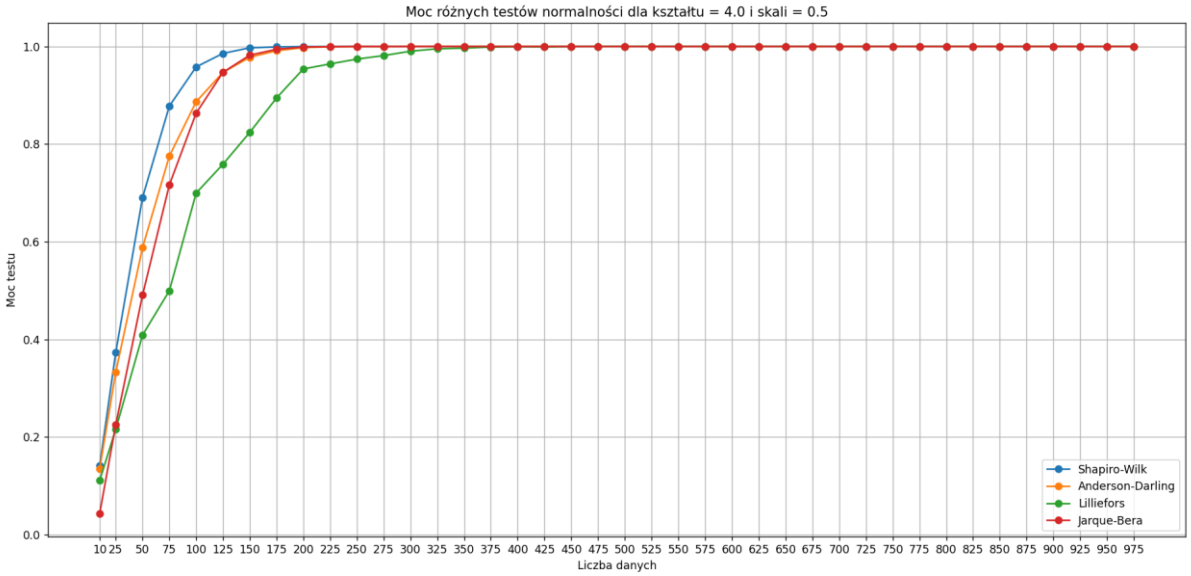
Wykres wygląda identycznie jak poprzedni.



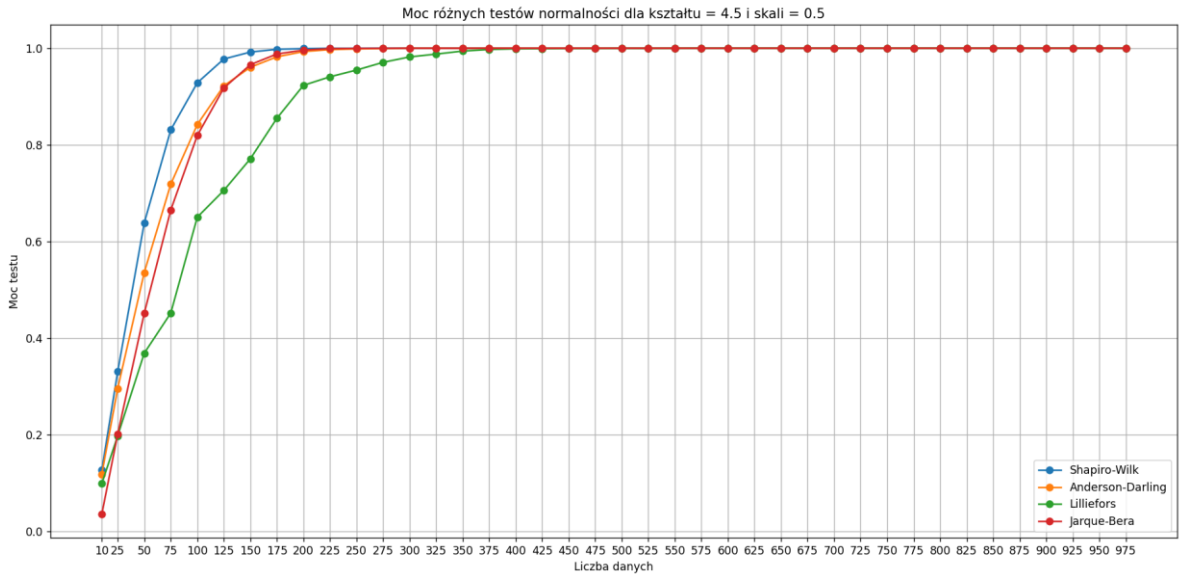
Dla skali 0.5 nieco szybciej zbiegają się testy.



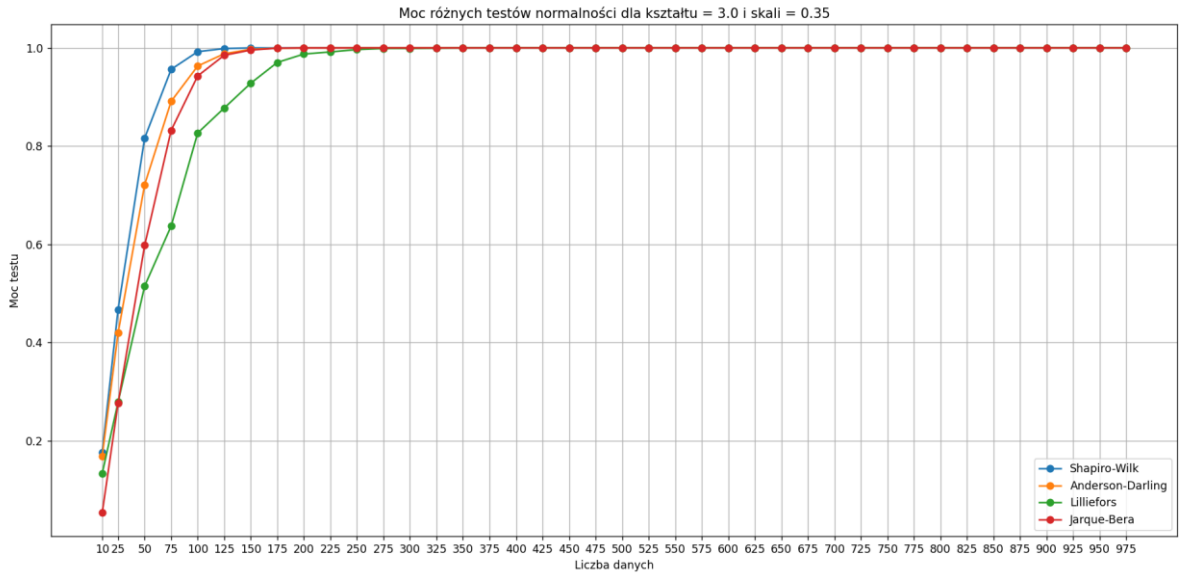
Lilliefors nieco później zbiegł do 1 od poprzedniego



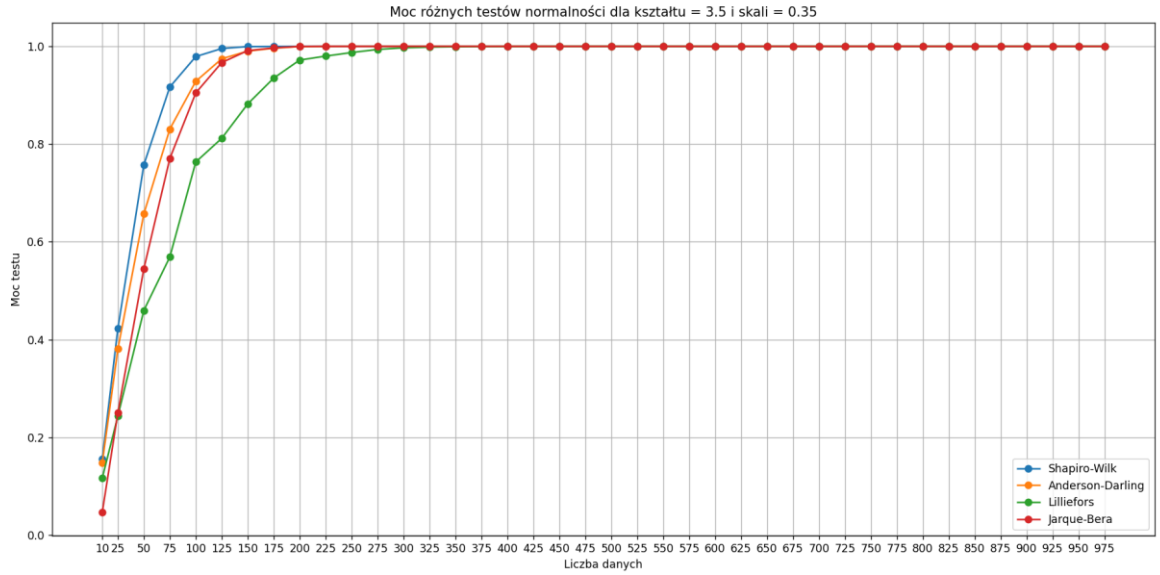
Wszystkie testy zbiegają się dosyć szybko do 1, mimo tego, że rozkład gamma dla kształtu 4 i skali 0.5 przypomina rozkład normalny. Choć można zauważyć, że Lilliefors dopiero dla 350 danych zbiegł do 1.



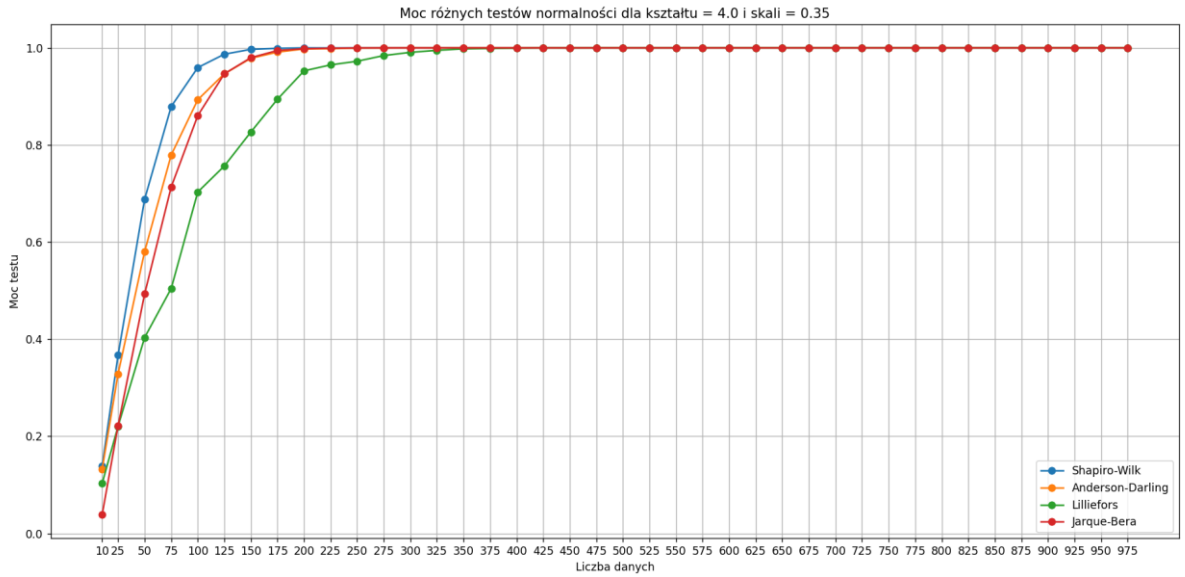
Wykres dosyć podobny.



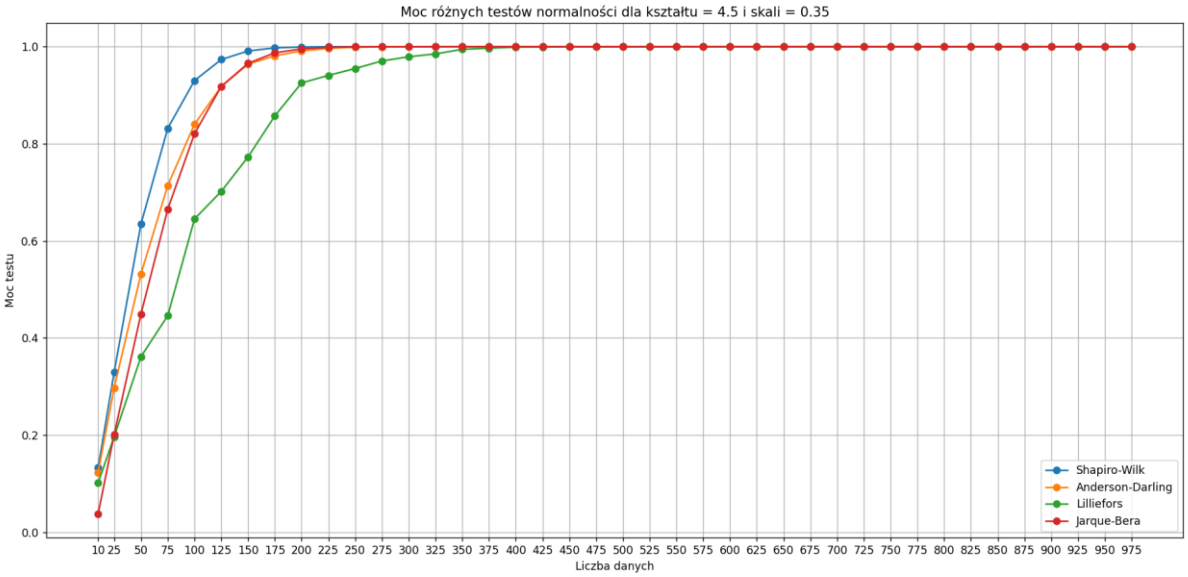
Zdaje się, że testy w tym wypadku najszybciej zbiegły do 1. Można to stwierdzić porównując aktualne badanie z badaniem stosującym kształt równy 4 i skalę równą 0.5.



Lilliefors później zbiega się do 1 gdyż rozkład ten bardziej przypomina rozkład normalny.



Lilliefors później zbiega się do 1 gdyż rozkład ten bardziej przypomina rozkład normalny, kształt wynosi 4.



Wykres dość podobny do poprzedniego.

Podsumowanie rozkładu gamma

Testy nieco szybciej zbiegają do 1 niż początkowo zakładaliśmy. Wszystkie testy trafnie oceniły, że dane nie pochodzą z rozkładu normalnego. Najlepiej wypadł test Shapiro-Wilka, następnie porównywalnie Anderson-Darling i Jarque-Bera, zdecydowanie najgorzej poradził sobie Lillieofors. Dla mniejszej liczebności próbki, testy miały dosyć małą moc testu, natomiast wszystkie testy miały widoczny trend rosnący, zatem szybko zbiegały do 1. Kształty i skale, jakie dobraliśmy, nie miały większego wpływu na wynik, natomiast im kształt był bliższy do 4 i skala była bliższa 0.5, tym testy nieco później zbiegały do 1, jednakże różnica była nieznaczna.

ANOVA

Do analizy wariancji ANOVA wykorzystamy zbiór telefonów.

Kolor, a cena

Naszym celem jest sprawdzenie, czy kolor obudowy wpływa na cenę. Wybrane telefony są stosunkowo nowe na rynku. Wybraliśmy telefony z 4 marek: Apple, Samsung, Huawei i Xiaomi. Wyodrębniliśmy 7 różnych kolorów (biały, czarny, czerwony, niebieski, srebrny, zielony, żółty)

Liczność grupy

Kolor	Liczebność
Biały	27
Czarny	40
Czerwony	23
Niebieski	36
Srebrny	24
Zielony	30
Żółty	22

Maksymalna liczebność w grupie wynosi 40, natomiast minimalna wynosi 22, co jest zauważalną różnicą w liczebności pomiędzy grupami. Jednakże, przy tej liczebności grup, nie powinno to być istotnym problemem, szczególnie, że większość grup oscyluje licznością w okolicy 30. Właśnie z tego powodu, uznaliśmy, że takie dysproporcje nie wpłyną istotnie na wynik, a błędy obliczeniowe, które mogą się pojawić, będą bliskie granicy błędu.

Normalność Danych

Zbadamy normalność danych w grupach, która jest jednym z założeń ANOVY. Dokonamy tego testem Shapiro-Wilka, ponieważ jest to mocny test sprawdzający normalność danych.

Kolor	P-value
Biały	0.381
Czarny	0.047
Czerwony	0.858
Niebieski	0.144
Srebrny	0.204
Zielony	0.962
Żółty	0.265

Nie ma podstaw do odrzucenia hipotezy 0 (na poziomie istotności $\alpha=0.05$) oprócz koloru czarnego. Jednakże wartość p_value jest dosyć blisko poziomu istotności, a dodatkowo przy większej liczebności, normalność nie jest aż tak dużym problemem w ANOVIE.

Jednorodność wariancji w grupach

Test Bartleta służy do sprawdzenia jednorodności wariancji w grupach, która to jednorodność jest kluczowym założeniem ANOVY. W naszym wypadku p_value wyniosło w przybliżeniu 0.5, zatem nie ma podstaw do odrzucenia h0 o jednorodności wariancji. Także założenie zostało spełnione

ANOVA rezultat

	df	sum_sq	mean_sq	F	PR(>F)
Kolor	6.0	4.269482e+07	7.115804e+06	1.236871	0.289029
Residual	195.0	1.121848e+09	5.753069e+06	NaN	NaN

P_value wynosi 0.28, czyli nie ma podstaw do odrzucenia h0 o równości wszystkich średnich w grupie, zatem nie ma wystarczających dowodów statystycznych na to, że kolor wpływa na cenę.

ANOVA dla marki

Liczność grupy

Marka	Liczebność
Apple	52
Samsung	55
Huawei	45
Xiaomi	50

W tym wypadku grupy są równoliczne, także jedno z założeń ANOVY zostało spełnione

Normalność w grupach

Marka	P_value
Apple	0.009
Samsung	0.135
Huawei	0.0003
Xiaomi	0.054

Dane z Huawei nie pochodzą z rozkładu normalnego, zatem nie spełniają założeń ANOVY. Gdybyśmy pozbyli się tych danych, to pozostałe grupy nie spełniałyby założenia jednorodności wariancji (p_value 0.0009 w teście Bartleta), zatem zastosujemy test Kruskalla-Wallisa.

Test Kruskalla-Wallisa

P_value wynosi 6.25e-16 zatem istnieją statystycznie istotne różnice w cenach pomiędzy badanymi grupami.

Autorzy:

Tomasz Zapart,

Radosław Mocarski