Pakiety statystyczne raport 2

Adam Wrzesiński, Joanna Kołaczek

18.07.2022

Spis treści

1 Wstęp

Niniejszy raport powstał na potrzeby realizacji laboratorium z Pakietów Statystycznych, prowadzonych przez dr inż. XXXX, do wykładu dr inż. Andrzeja Giniewicza.

Będziemy testować hipotezy statystyczne dla wartości średniej i wariancji w rodzinie rozkładów normalnych. Zobrazujemy także obszary krytyczne, wyznaczymy p-wartości oraz prawdopodobieństwo wystąpienia błędów I i II rodzaju. Życzymy miłej lektury.

2 Transformata Sinh-arcsinh

Przy generowaniu rozkładu normalnego, musimy ustalić parametr położenia μ (dla rozkładu normalnego jest to średnia) oraz parametr skali σ (w tym przypadku będzie to odchylenie standardowe). Transformata sinh-arcsinh rozkładu normalnego wprowadza nowe parametry, które kontrolują asymetrię ν i ciężkość ogonów τ. Dane cztery parametry definiują rozkład normalny Sinh-arcsinh jako:

$$X = \mu + \sigma \cdot \sinh\left[\frac{\sinh^{-1}(Z) + \nu}{\tau}\right],$$

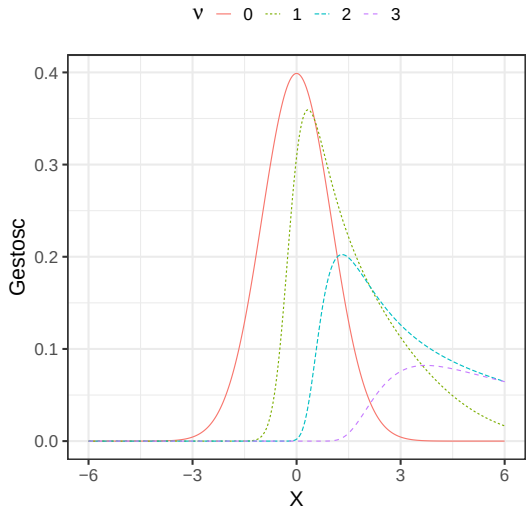
gdzie Z jest zmienną losową z standardowego rozkładu normalnego oraz:

$$\sinh(x) = \frac{\mathrm{e}^{x} + \mathrm{e}^{-x}}{2}, \quad \sinh^{-1}(x) = \log(x + \sqrt{1 + x^2}).$$

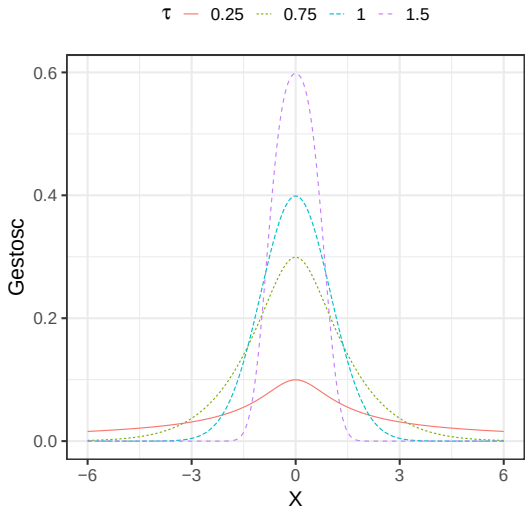
Dystrybuanta rozkładu Sinh-arcinh zastosowanego na rozkład normalny wygląda następująco:

$$F(x; \mu, \sigma, \nu, \tau) = \phi\left(\sinh\left(\tau \sinh^{-1}\left(\frac{x - \mu}{\sigma}\right) - \nu\right)\right).$$

gdzie ϕ do dystrybuanta standardowego rozkładu normalnego $\mathcal{N}(0,1)$



Rysunek 1: Transformata Sinh-arcsinh z modyfikacją ν



Rysunek 2: Transformata Sinh-arcsinh z modyfikacja τ

Na rysunkach 1, 2 widzimy przykładowe rozkłady normalne po transformacji. Jak widać wartości $\nu > 0$ oznaczają rozkłady prawoskośne (analogicznie $\nu < 0$ będą lewoskośne), $\tau > 1$ oznacza rozkłady o chudszych ogonach niż normalny, natomiast $\tau < 1$ grubszych. Warto jednak podkreślić, że ν i τ nie są skośnością oraz kurtozą, jedynie parametrami kontrolującymi je nie wprost. W dalszej części raportu, sprawdzimy jak zmiany skośności oraz kurtozy wpływają na moc omawianych testów.

3 Wybrane testy statystyczne

Testy statystyczne mają zadanie oszacować prawdopodobieństwo spełnienia pewnej hipotezy statystycznej w populacji na podstawie próby losowej z tej populacji. W raporcie przyjrzymy się jak zmienia się moc wybranych testów normalności w obliczu transformaty sinh-arcsinh. Hipoteza zerowa H_0 którą przyjmujemy w każdym z nich brzmi: *Próba pochodzi z rozkładu normalnego* przeciwko hipotezie alternatywnej H_1: *Próba nie pochodzi z rozkładu normalnego.*

Test Shapiro-Wilka

Opiera się on na statystyce W, wyliczanej na podstawie danych z próbki, która będzie tym bliższa 1 im bardziej prawdopodobne jest to, że dane pochodzą z rozkładu normalnego.

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

gdzie $x_{(i)}$ z nawiasami w indeksie dolnym to i-ta najmniejsza liczba w próbie (nie mylić z x_i) a $\bar{x}$ oznacza średnią próbkową. Współczynniki a_i dane są jako:

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C},$$

gdzie C jest wektorem norm:

$$C = \|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}},$$

a wektor m:

$$m = (m_1, \ldots, m_n)^T$$

składa się z wartości oczekiwanych statystyki porządkowej (ang. *order statistics*) niezależnych zmiennych losowych próbkowanych ze standardowego rozkładu normalnego, natomiast V jest macierzą kowariancji tejże statystyki.

Test Kołmogorowa-Lillieforsa

Statystyka tego testu, to maksymalna bezwzględna różnica między empiryczną a hipotetyczną funkcją rozkładu skumulowanego. Można obliczyć ją jako $D = D^+, D^-$, gdzie

$$D^+ = \max_{i=1,\ldots,n} (i/n - p_{(i)}), \quad D^- = \max_{i=1,\ldots,n} (p_{(i)} - (i-1)/n),$$

gdzie:

$$p_{(i)} = \Phi([x_{(i)} - \overline{x}]/s).$$

Tutaj Φ jest funkcją skumulowanego standardowego rozkładu normalnego, $\bar{x}$ i s są średnią i odchyleniem standardowym wartości danych.

Test Jarque-Bera

Jest to test dopasowania (ang. *goodness of fit*), określający czy dane z próbki mają skośność i kurtozę odpowiadającą rozkładowi normalnemu. Statystyka JB testu jest zawsze większa bądź równa zero. Jeżeli jest daleka od zera, sygnalizuje, że dane nie mają rozkładu normalnego.

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right),$$

gdzie n oznacza ilość obserwacji, natomiast S oznacza skośność a K kurtozę:

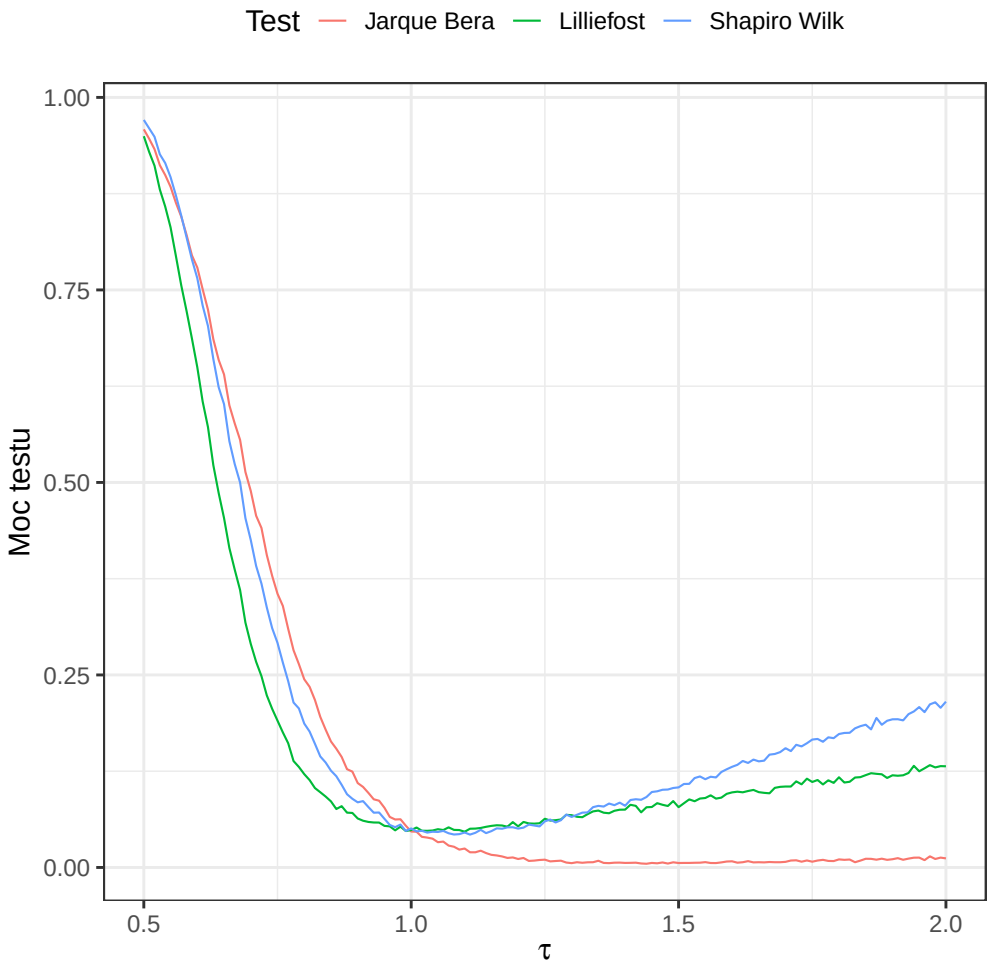
$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^{3/2}}, \qquad K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)^2},$$

Jeśli dane pochodzą z rozkładu normalnego, statystyka JB zbiega asymptotycznie do rozkładu chi kwadrat z dwoma stopniami swobody.

4 Zadania

Zadanie 1

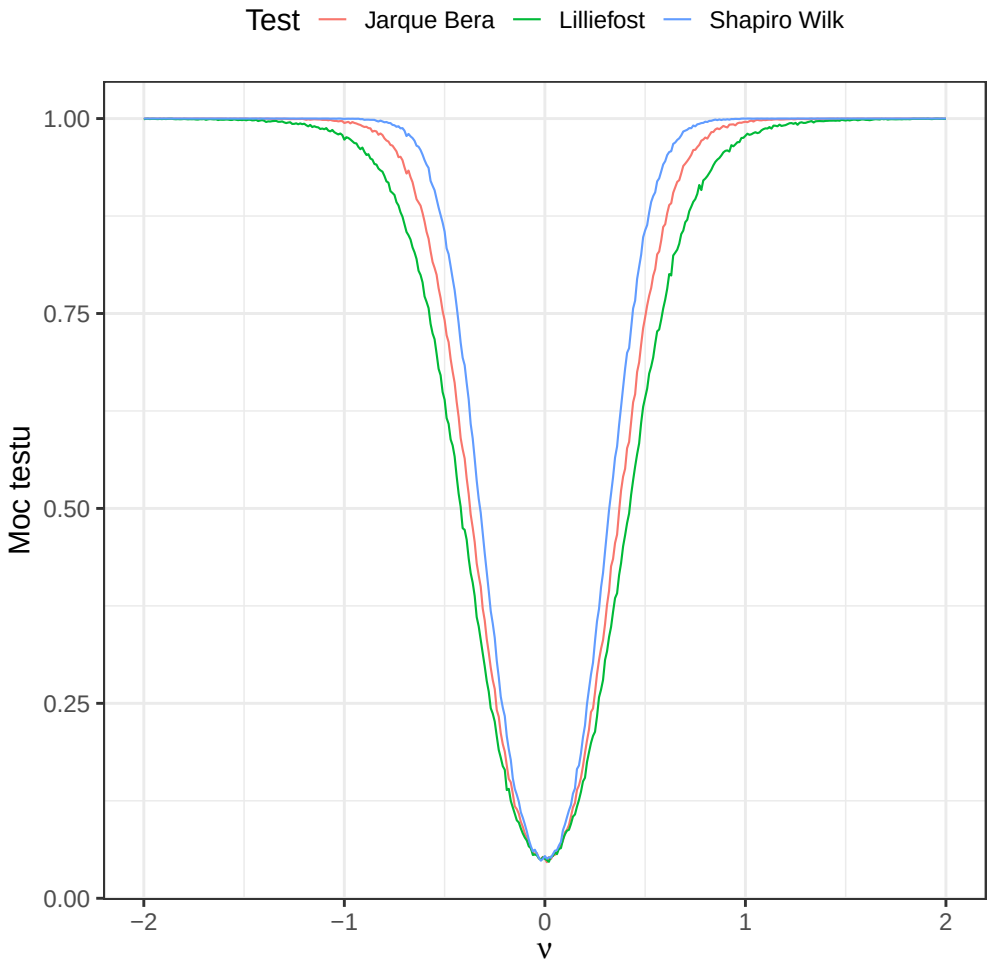
Dysponujemy próbą o rozmiarze 100 z rozkładu normalnego $\mathcal{N}(-1,3)$ przekształconego przez transformatę Sinh-arcsinh z $\nu = 0$. Na wykresie 3 sprawdzamy jak zmiana parametru τ na przedziale (0.5,2), wpływa na moc poszczególnych testów. Widzimy, że dla małych τ wszystkie testy radzą sobie dobrze. Niestety moc testu jest niska dla rozkładów leptokurtycznych i wzrasta powoli dla testów Kołmogorowa-Lillieforsa oraz Shapiro-Wilka. W przypadku testu Jarque-Bera moc testu pozostaje w przybliżeniu stała. Na zadanym przedziale nie ma testu jednostajnie najmocniejszego.



Rysunek 3: Moc testu dla rozkładu Sinh-arcsinh wobec zmienności τ

Zadanie 2

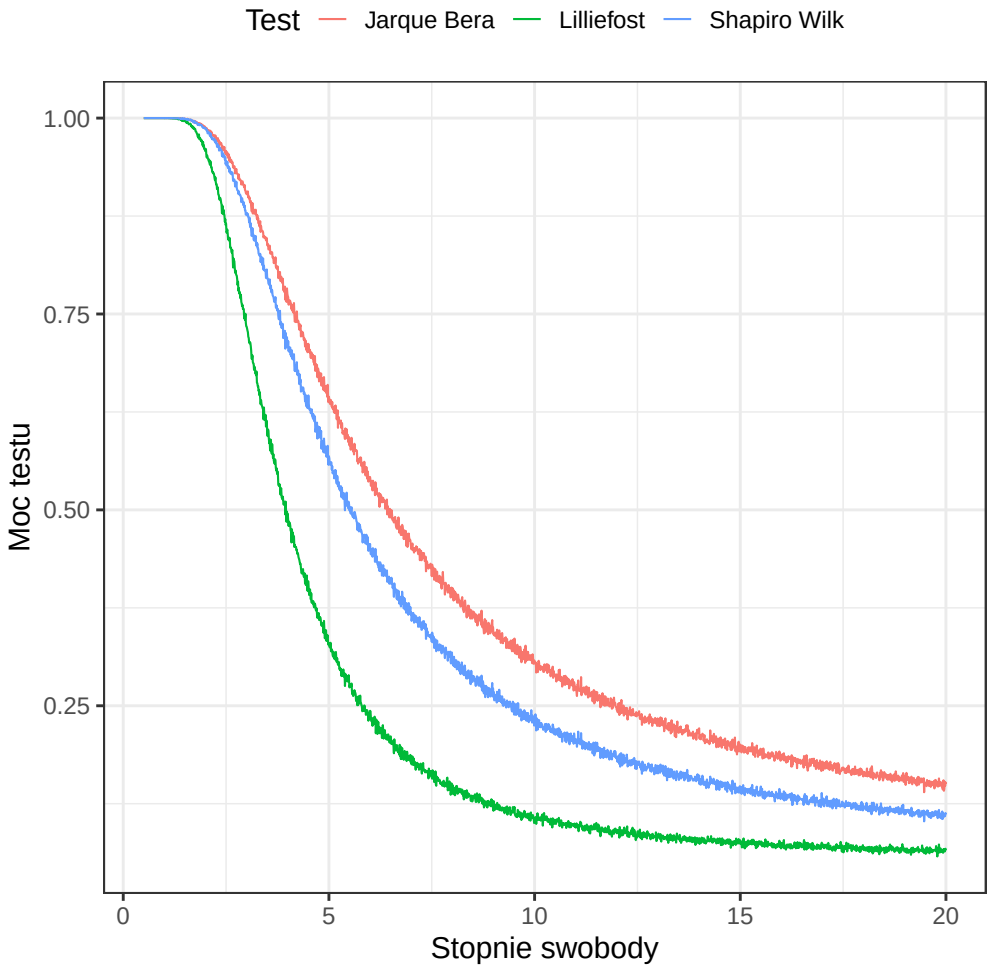
Dysponujemy próbą o rozmiarze 100 z rozkładu normalnego $\mathcal{N}(-1,3)$ przekształconego przez transformatę Sinh-arcsinh z $\tau = 1$. Na wykresie 4 sprawdzamy jak zmiana parametru ν na przedziale $(-2,2)$, wpływa na moc poszczególnych testów. Zauważamy symetrię względem prostej $\nu = 0$. Wynika to z faktu, że dla przeciwnych ν skośność jest również przeciwna. Tym razem możemy wyróżnić jednostajnie najmocniejszy test - Shapiro-Wilka. Niemniej jednak na wyborze innych testów wiele nie stracimy.



Rysunek 4: Moc testu dla rozkładu Sinh-arcsinh wobec zmienności ν

Zadanie 3

Mamy próbę $(X_1, \ldots, X_{100})$ taką, że zmienne losowe $Y_i = \frac{X_i - 1}{3}$ są z rozkładu t-Studenta $\mathcal{T}(\nu)$. Ponieważ X_t jest liniową funkcją zmiennej z rozkładu t-Studenta oczekujemy, że wraz ze wzrostem liczby stopni swobody, jego rozkład będzie zbiegał do rozkłądu normalnego. Rzeczywiście, wykres 5 potwierdza nasze przypuszczenia - moce testów zbiegają asymptotycznie do zera. Jednostajnie najmocniejszym testem jest test Jarque-Bera. Najgorzej spisuje się zaś test Kołmogorowa-Lillieforsa.



Rysunek 5: Moc testu dla rozkładu t-Studenta wobec zmienności stopni swobody

5 Podsumowanie

Rozwiązane zadania pokazały, że nie ma jednego uniwersalnego jednostajnie najmocniejszego testu. Wszystkie rozważane przez nas testy mają trudności w rozpoznaniu rozkładów podobnych do rozkładu normalnego. Przy testowaniu hipotez warto zatem wykonać kilka niezależnych testów oraz podchodzić do każdego przypadku indywidualnie.

6 Źródłą