Symulacyjna analiza mocy wybranych testów statystycznych

Marcin Miśkiewicz, Adrian Sobczak

23 stycznia 2022

1 Wprowadzenie

Celem niniejszego sprawozdania jest symulacyjna analiza mocy trzech wybranych testów statycznych dla danych pochodzących z trzech różnych rozkładów. W każdym przypadku sprawdzimy czy wśród rozpatrywanych testów istnieje test jednoznacznie najmocniejszy, a także zwrócimy uwagę na to jak testy działają dla danych niespełniających potrzebnych założeń.

1.1 Opis analizy

Badanymi testami będą: test z (przy założeniu $\sigma = 2$), test t oraz test rang znakowanych Wilcoxona. Rozpatrzymy trzy przypadki: dane z rozkładu $\mathcal{N}(\mu, 2^2)$, dane z rozkładu $\mathcal{N}(\mu, 4^2)$ oraz dane z rozkładu $\mathcal{E}(\frac{1}{\mu})$. W każdym przypadku przyjmiemy pewną wartość μ, wygenerujemy 100 obserwacji z danego rozkładu i będziemy testować $H_0 : \mu = 1$ przeciwko $H_1 : \mu \neq 1$, na poziomie istotności $\alpha = 0.05$. Korzystając z metody Monte Carlo, procedurę tę powtórzymy 100 razy i za moc przyjmiemy ile procent przeprowadzonych testów prowadziło do odrzucenia hipotezy zerowej. Wszystkie omówione kroki wykonamy dla wielu μ z przedziału zawierającego wartości z hipotezy zerowej oraz alternatywnej.

1.2 Opis rozpatrywanych testów

Dla dwustronnego testu z zakładamy, że $X_1, \ldots, X_n$ są niezależnymi zmiennymi losowymi z rozkładu $\mathcal{N}(\mu, \sigma^2)$, gdzie μ jest nieznane, a σ jest znaną wariancją. Statystyką testową jest $\bar{X}$, czyli średnia z próby. Dla testu nieobciążonego, funkcja mocy wyraża się wzorem

$$\beta(\mu) = \Phi\left(\sqrt{n}\frac{\mu_0 - \mu}{\sigma} - z_{1-\frac{\alpha}{2}}\right) + 1 - \Phi\left(\sqrt{n}\frac{\mu_0 - \mu}{\sigma} + z_{1-\frac{\alpha}{2}}\right),$$

gdzie Φ to dystrybuanta, a $z_{1-\frac{\alpha}{2}}$ to kwantyl rzędu $1 - \frac{\alpha}{2}$ standardowego rozkładu normalnego.

Dla dwustronnego testu t zakładamy, że $X_1, \ldots, X_n$ są niezależnymi zmiennymi losowymi z rozkładu $\mathcal{N}(\mu, \sigma^2)$, gdzie μ i σ są nieznane. Statystyką testową również jest $\bar{X}$, a sam test oparty jest na fakcie

$$t = \sqrt{n}\frac{\bar{X} - \mu_0}{S} \sim \mathcal{T}(n-1),$$

gdzie S to nieobciążony estymator wariancji, a $\mathcal{T}(n-1)$ to rozkład t-Studenta z $n-1$ stopniami swobody. Testu t używa się gdy nie znamy wartości σ, jednak warto zauważyć, że ze względu na nieobciążoność estymatora S, dla dużych prób wyniki uzyskane przy pomocy testu t są bardzo podobne do tych z testu z.

Test rang znakowanych Wilcoxona jest często używanym nieparametrycznym odpowiednikiem testu t-Studenta, jednak nie wymaga on założenia o normalności badanej próby. Kosztem tej uniwersalności jest mniejsza moc gdy dane faktycznie pochodzą z rozkładu normalnego – wtedy test t to lepszy wybór. W teście rang znakowanych Wilcoxona hipoteza zerowa jest postaci

$$H_0 : X - \mu_0 \stackrel{st}{=} \mu_0 - X,$$

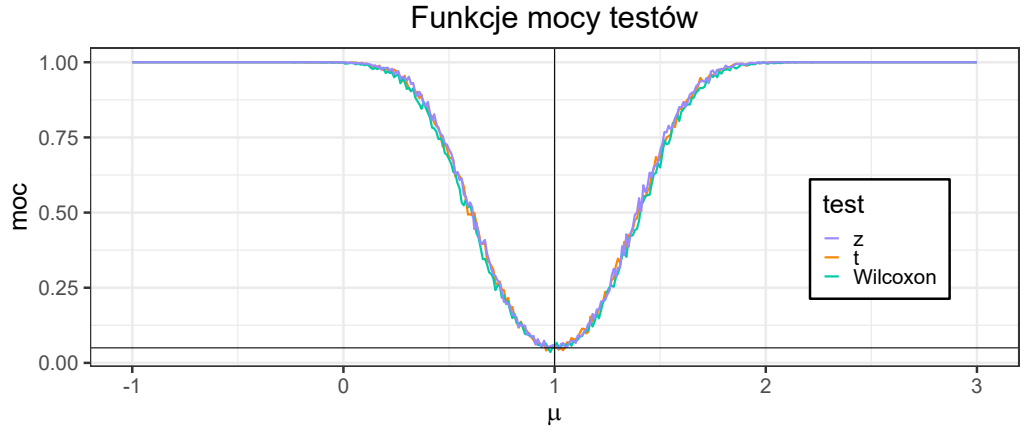
gdzie zapisana równość to równość stochastyczna. W naszym przypadku będziemy rozpatrywać X z rozkładu normalnego i z rozkładu wykładniczego. Rozkład normalny jest symetryczny, dlatego wynik będziemy mogli zinterpretować jako test mediany (i równoważnie średniej). Rozkład wykładniczy nie jest symetryczny, ale ze względu na ciągłość, będziemy mogli zinterpretować wynik jako test jednoznacznie wyznaczonej pseudomediany.

2 Analiza

Analizę podzielimy na trzy opisane wcześniej przypadki.

2.1 Przypadek 1.

Rozpatrujemy próbę $X_1, \ldots, X_{100}$ z rozkładu $\mathcal{N}(\mu, 2^2)$. Weźmy μ z przedziału $(-1, 3)$, z krokiem $h = 0.01$. Na wykresie (1) linia pionowa została poprowadzona w punkcie $\mu = 1$ z hipotezy zerowej, a linia pozioma wskazuje na



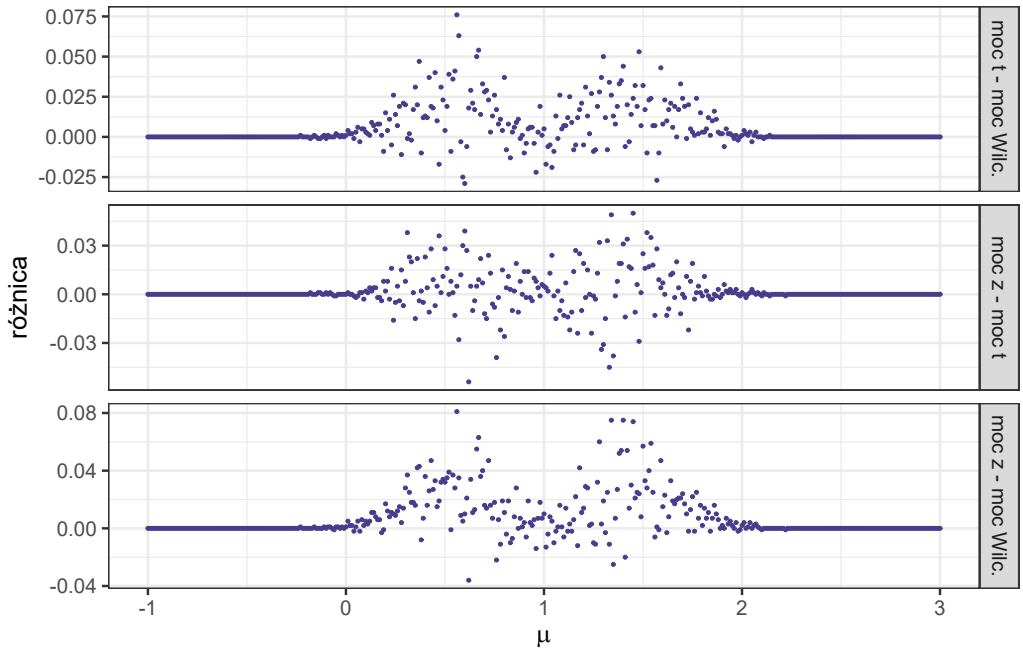
Rysunek 1: Funkcje mocy przeprowadzonych testów dla $X_i \sim \mathcal{N}(\mu, 2^2)$.

przyjęty poziom istotności $\alpha = 0.05$. Możemy zauważyć, że symulacyjnie wyznaczone funkcje mocy są do siebie bardzo zbliżone. Wszystkie testy zdają się być nieobciążone, a prawdopodobieństwo odrzucenia prawdziwej hipotezy zerowej przez każdy z nich jest, zgodnie z oczekiwaniami, w okolicach 0.05. Na wykresie (2) zaznaczono różnicę mocy dla każdej pary testów. Pomijając te wartości μ dla których testy mają jednakową moc, okazuje się, że test z jest mocniejszy od testu t w 61.17% przypadków, a od testu Wilcoxona w 81.78% przypadków. Test t jest mocniejszy od testu Wilcoxona w 78.08% przypadków. Z wykresu (2) możemy odczytać, że różnice w mocy są rzędu kilku setnych, a dla $\mu = 1$ testy działają bardzo podobnie.

Tak więc, pomimo tego, że na pierwszy rzut oka funkcje mocy są zbliżone, dla danych z rozkładu $\mathcal{N}(0, 2^2)$ najlepszym wyborem zdaje się być test z (oczywiście cały czas mówimy o teście przy założeniu $\sigma = 2$). Niemniej jednak, na podstawie przeprowadzonych symulacji, nie jesteśmy w stanie wskazać testu jednostajnie najmocniejszego.

2.2 Przypadek 2.

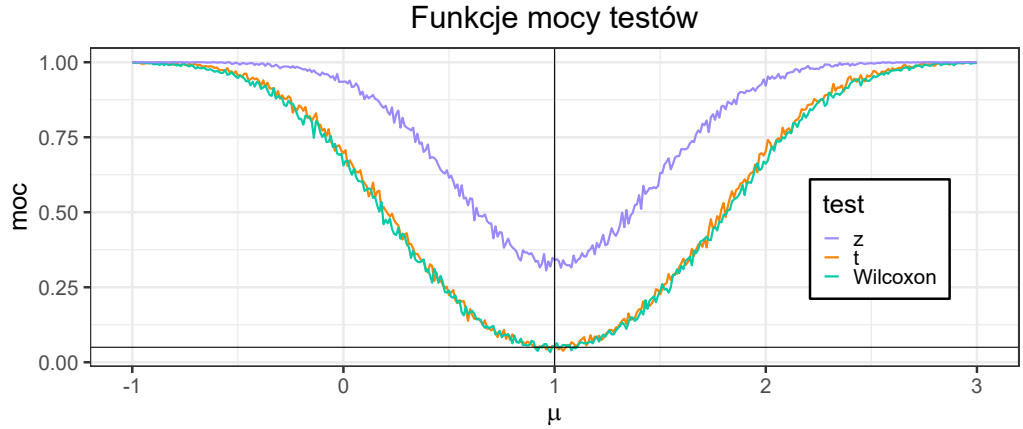
Tym razem rozpatrujemy próbę $X_1, \ldots, X_{100}$ z rozkładu $\mathcal{N}(\mu, 4^2)$. Weźmy μ z przedziału $(-1, 3)$, z krokiem $h = 0.01$. Podobnie jak w poprzednim przykładzie, na wykresie (3) linia pionowa została poprowadzona w punkcie $\mu = 1$ z hipotezy zerowej, a linia pozioma odpowiada przyjętemu poziomowi istotności $\alpha = 0.05$. Analizując wykres (3), od razu możemy zauważyć, że test z nie jest na zadanym poziomie istotności – hipotezę zerową odrzuca z prawdopodobieństwem bliskim 0.35. Domyślamy się, że jest to spowodowane niewłaściwą wartością wariancji przyjętej w teście. Cały czas zakładamy,



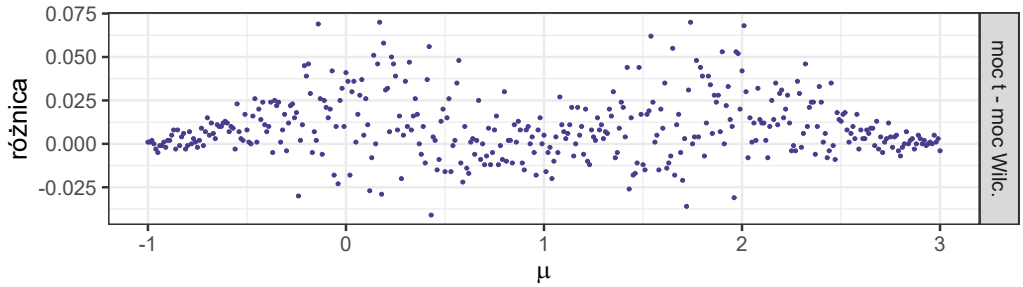
Rysunek 2: Różnice mocy przeprowadzonych testów dla $X_i \sim \mathcal{N}(\mu, 2^2)$.

że $\sigma^2 = 4$, gdzie w rzeczywistości $\sigma^2 = 16$, zatem test ten został niewłaściwie użyty. Problem ten nie dotyczy zaś testu t, gdyż przypomnijmy, że sam test opiera się na nieobciążonym estymatorze wariancji z badanej próby, a nie na teoretycznej wartości σ^2. Podobnie dobrze radzi sobie test rang znakowanych Wilcoxona – wygenerowana symulacyjnie funkcja mocy tego testu jest bardzo podobna do tej dla testu t. Wykres (4) przedstawia różnicę mocy tych dwóch testów dla przyjętych wartości μ. Okazuje się, że pomijając te wartości μ dla których testy mają jednakową moc, test t jest mocniejszy od testu rang znakowanych Wilcoxona w 73.7% przypadków. Podobnie jak w przypadku (1), różnice w mocy są rzędu kilku setnych, a dla $\mu = 1$ oba testy działają bardzo podobnie.

Na podstawie przeprowadzonych symulacji nie jesteśmy w stanie wskazać testu jednoznacznie najmocniejszego. Chociaż test z zdaje się być mocniejszy od pozostałych w każdym punkcie alternatywy, to porównywanie go do testów o innym poziomie istotności nie ma zbyt wiele sensu. Nie zważając na poziom istotności zawsze jesteśmy w stanie podać test mocniejszy lub niegorszy od pozostałych, chociażby jednostajnie najmocniejszy test $\varphi(X) = 1$. Jeśli nie mamy pewności co do wartości wariancji rozkładu z którego pochodzi badana próba, powinniśmy skorzystać z testu t (oczywiście przy założeniu normalności próby) lub z testu rang znakowanych Wilcoxona. Spodziewali-



Rysunek 3: Funkcje mocy przeprowadzonych testów dla $X_i \sim \mathcal{N}(\mu, 4^2)$.



Rysunek 4: Różnice mocy przeprowadzonych testów dla $X_i \sim \mathcal{N}(\mu, 4^2)$.

śmy się, że test rang znakowanych Wilcoxona wypadnie nieco gorzej na tle testu t, jednak dla rozpatrywanej próby o rozmiarze 100, różnice w mocy były bardzo niewielkie. Ich wartości mogą wynikać również z błędu samej metody Monte Carlo.

2.3 Przypadek 3.

W ostatnim badanym przez nas przypadku rozpatrujemy próbę $X_1, \ldots, X_{100}$ z rozkładu $\mathcal{E}(\frac{1}{\mu})$. Weźmy μ z przedziału $(0.01, 4)$, z krokiem $h = 0.01$. Podobnie jak wcześniej, na wykresie (5) linia pionowa została poprowadzona w punkcie $\mu = 1$ z hipotezy zerowej, a linia pozioma odpowiada przyjętemu poziomowi istotności $\alpha = 0.05$. Na wykresie (5) wyraźnie widać, że test z nie działa dobrze dla danych pochodzących z rozkładu wykładniczego. Test zdecydowanie nie jest na przyjętym poziomie istotności i w wielu punktach z hipotezy alternatywnej jego moc jest bliska 0. Zaskakujące może być to, jak

dobrze w tym przypadku wypada test t, którego jednym z założeń jest normalność badanej próby. Zauważmy jednak, że w teście t statystyką testową jest $\bar{X}$. Z centralnego twierdzenia granicznego wiemy, że dla dużych rozmiarów próby, rozkład $\bar{X}$ jest zbliżony do rozkładu normalnego, zatem możemy przypuszczać, że dla badanych 100 obserwacji, test t zadziała przyzwoicie. Aby przekonać się, czy faktycznie średnia ze stu niezależnych zmiennych losowych z rozkładu wykładniczego ma rozkład zbliżony do rozkładu normalnego, wykonamy symulację. Wygenerujemy wektor stu realizacji zmiennej losowej z rozkładu wykładniczego z przykładowym parametrem $\lambda = 1$ i obliczymy średnią próbkową. Procedurę tę powtórzymy sto razy i na wektorze stu takich średnich przeprowadzimy test Shapiro-Wilka na normalność. Otrzymujemy

$$\text{statystyka testowa W} = 0.99, \text{ p-wartość} = 0.71,$$

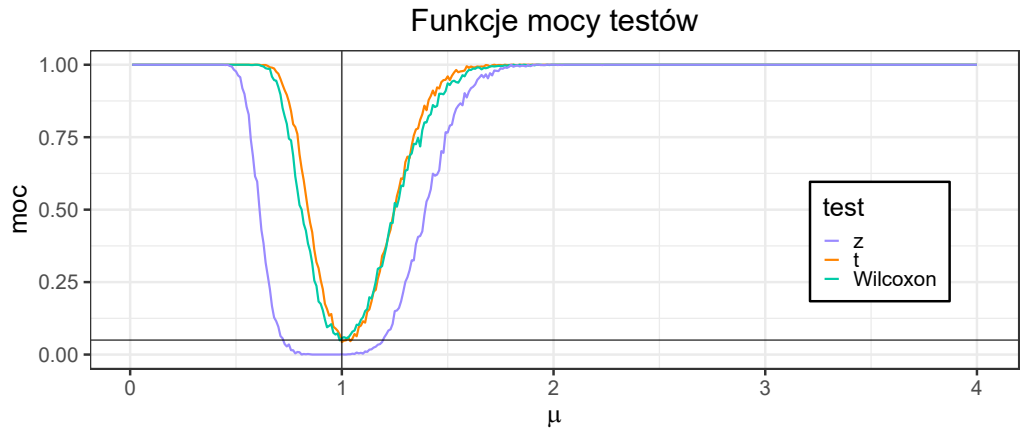
czyli nie mamy podstaw by odrzucić hipotezę zerową świadczącą o normalności badanej próby. Domyślamy się, że dla innych wartości λ wyniki byłyby podobne. Warto również wspomnieć o tym, że dla próby niepochodzącej z rozkładu normalnego, rozkład statystyki $(n-1)S^2/\sigma^2$ może znacząco odbiegać od rozkładu χ^2, jednak z twierdzenia Slutsky'ego wnioskujemy, że dla dużej próby nie będzie to miało większego wpływu na rozkład statystyki t. Zatem możemy stierdzić, że test t będzie działał dobrze nawet dla próby o rozkładzie innym niż normalny, pod warunkiem, że próba ta będzie duża.

Test rang znakowanych Wilcoxona, będący testem nieparametrycznym, również działa dobrze dla badanej próby. Sam test traktujemy jako test pseudomediany, czyli mediany średniej zmiennej i jej niezależnej kopii. Dla rozkładu $X_1, X_2 \sim \mathcal{E}(\lambda)$ mamy

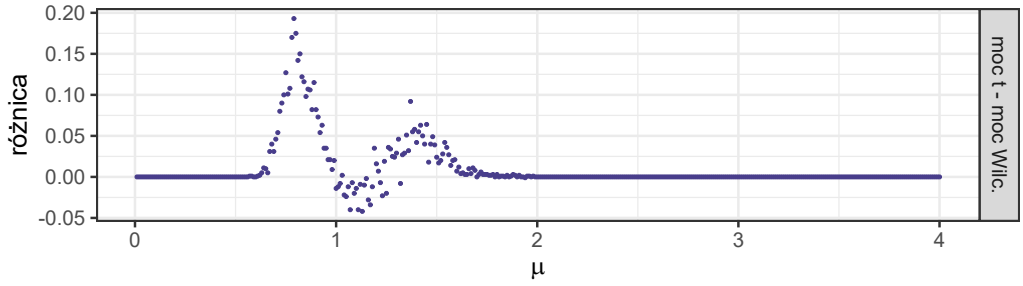
$$\frac{X_1 + X_2}{2} \sim \mathcal{G}(2, 2\lambda)$$

$$q_{2,2\lambda}(0.5) = \frac{q_{2,2}(0.5)}{\lambda} \approx \frac{0.8391735}{\lambda} = 0.8391735\mu.$$

Na wykresie (6) możemy zobaczyć, że dla $\mu < \mu_0$ test rang znakowanych Wilcoxona wypada zdecydowanie gorzej niż test t – różnica mocy sięga nawet 0.2. Okazuje się, że pomijając te wartości μ dla których testy mają jednakową moc, test t jest mocniejszy od testu rang znakowanych Wilcoxona w 82.44% przypadków. Nie możemy zatem stwierdzić, że test t jest jednostajnie najmocniejszy, gdyż nadal mamy przypadki $\mu \in H_1$, z którymi test rang znakowanych Wilcoxona poradził sobie lepiej.



Rysunek 5: Funkcje mocy przeprowadzonych testów dla $X_i \sim \mathcal{E}(\frac{1}{\mu})$.



Rysunek 6: Różnice mocy przeprowadzonych testów dla $X_i \sim \mathcal{E}(\frac{1}{\mu})$.

3 Podsumowanie

Podsumujmy obserwacje i wnioski uzyskane z analizowanych przypadków. Dla próby $X_1, \ldots, X_{100}$ z rozkładu $\mathcal{N}(\mu, 2^2)$ każde założenie rozpatrywanych testów jest spełnione i wszystkie testy mają podobną moc dla μ z przedziału $(-1, 3)$. Minimalnie najlepszy zdaje się być test z, jednak nie możemy jednoznacznie stwierdzić, by był to test jednostajnie najmocniejszy.

Dla próby $X_1, \ldots, X_{100}$ z rozkładu $\mathcal{N}(\mu, 4^2)$ test z (przy założeniu $\sigma = 2$) nie jest odpowiednim wyborem. Jego moc w punkcie μ_0 nie odpowiada zadanemu poziomowi istotności i porównania do pozostałych testów nie niosą sensownych informacji. Test t i test rang znakowanych Wilcoxona są odporne na zmianę wartości wariancji rozkładu. Wszystkie założenia tych testów zostały spełnione, a uzyskane wyniki są bardzo podobne. Nie możemy wyróżnić testu jednostajnie najmocniejszego.

Dla próby $X_1, \ldots, X_{100}$ z rozkładu $\mathcal{E}(\frac{1}{\mu})$ test z działa najgorzej. Nie mamy tu spełnionego założenia o normalności próby, czego skutkiem jest fatalna moc testu. Pomimo że założenie to dotyczy również testu t, test ten wypada bardzo dobrze. Jest to efekt działania centralnego twierdzenia granicznego dla tak licznej próby. Okazuje się, że gdy dysponujemy wieloma danymi, test t może być dobrym wyborem pomimo niespełnionych założeń. Ze względu na swoją uniwersalność, test rang znakowanych Wilcoxona również działa na zadowalającym poziomie. W przypadku rozkładu wykładniczego, test ten traktujemy jako test pseudomediany. Chociaż w większości przypadków test t działał najlepiej, nie możemy jednoznacznie stwierdzić by był to test jednostajnie najmocniejszy.