



Alicja Nowakowska

Politechnika Wrocławska

Raport R - Pakiety statystyczne

Wrocław, 19 stycznia 2020

Spis treści

1	Wstęp teoretyczny	2
2	Realizacja zadania	3
2.1	Treść zadania	3
2.2	Implementacja w R	3
2.3	Wyniki	4
2.3.1	Zadanie 1.- $B(7,p)$	4
2.3.2	Zadanie 2.- $B(35,p)$	7
2.3.3	Zadanie 3.- $B(250,p)$	10
3	Wnioski	11

1 Wstęp teoretyczny

Testowanie, które przeprowadzamy służy nam do sprawdzenia hipotezy dotyczących pewnych danych. Testowaniem statystycznym nazywamy regułę postępowania, która każdej możliwej próbie przyporządkowuje decyzję przyjęcia lub odrzucenia hipotezy. Z kolei moc testu to prawdopodobieństwo niepopelnienia błędu drugiego rodzaju czyli nieodrzućenia hipotezy zerowej, gdy w rzeczywistości jest ona fałszywa. Mówmy, że test jest δ^* jest jednostajnie najmocniejszy na przedziale na poziomie istotności α , jeśli:

1. δ^* jest testem na poziomie istotności α ,
2. dla każdego testu δ na poziomie istotności α , mamy:

$$P_{\theta}(\delta^*(X) = 1) \geq P_{\theta}(\delta(X) = 1), \quad (1)$$

dla każdego $\theta \in \Theta_1$.

Przedziały ufności jakie stosujemy w naszym raporcie są przedziałami ufności stosowanymi dla rozkładu normalnego, przedział ufności Wilsona (1927r.), który zadany jest wzorem:

$$\left(p + \frac{z_{1-\alpha/2}^2}{2n} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}} \right) / \left(1 + \frac{z_{1-\alpha/2}^2}{n} \right) \quad (2)$$

Szczególnie ma on zastosowanie dla nawet małej wielkości prób i skrajnego prawdopodobieństwa. Wiemy, że standardowy przedział ufności nie jest zbyt wiarygodny, gdy mamy małą próbkę lub prawdopodobieństwo bliskie 0 lub 1. Intuicyjnie możemy przyjąć, że wartość środkowa tego przedziału jest średnią ważoną.

Następnym przedziałem ufności, który użyliśmy w raporcie to przedział Cloppera-Pearsona, który jest zadany wzorem:

$$\beta\left(\frac{\alpha}{2}; x, n - x + 1\right) < p < \beta\left(1 - \frac{\alpha}{2}; x + 1, n - x\right), \quad (3)$$

gdzie β pochodzi z rozkładu Beta, α to parametr kształtu, n to ilość prób i x to ilość sukcesów. Ten przedział ufności jest wczesną i bardzo powszechną metodą obliczania dwumianowych przedziałów ufności. Jest on często nazywany metodą „dokładną” dlatego, że opiera się na

skumulowanych prawdopodobieństwach rozkładu dwumianowego. Wadą tej metody jest to, że przedziały nie są dokładne w sposób, który można by przypuszczać: nieciągły charakter rozkładu dwumianowego wyklucza każdy przedział z dokładnym pokryciem dla wszystkich proporcji populacji.

Ostatnim przedziałem na jakim się skupiliśmy, był przedział Jeffreysa, który wyraża się wzorem i tworzą go kwantyle:

$$(q_{\alpha/2}, q_{1-\alpha/2}) \quad (4)$$

Jego właściwości są podobne do właściwości Wilsona.

2 Realizacja zadania

2.1 Treść zadania

Nasze zadanie polegało na obserwacji zmiennych losowych z rozkładu Bernoulliego o parametrach $n \in \{7, 35, 250\}$ i $p \in (0, 1)$. Hipoteza zerowa i alternatywa miały następującą postać:

$$H_0 : p = 0.5,$$

$$H_1 : p \neq 0.5.$$

Korzystając z symulacji Monte Carlo miałyśmy narysować wykres funkcji mocy testu na poziomie istotności $\alpha = 0.05$ dla testu opartego o przedział Wilsona, Cloppera-Pearsona i Jeffreysa. Ponadto miałyśmy stwierdzić, czy istnieje test jednostajnie najmocniejszy z nich.

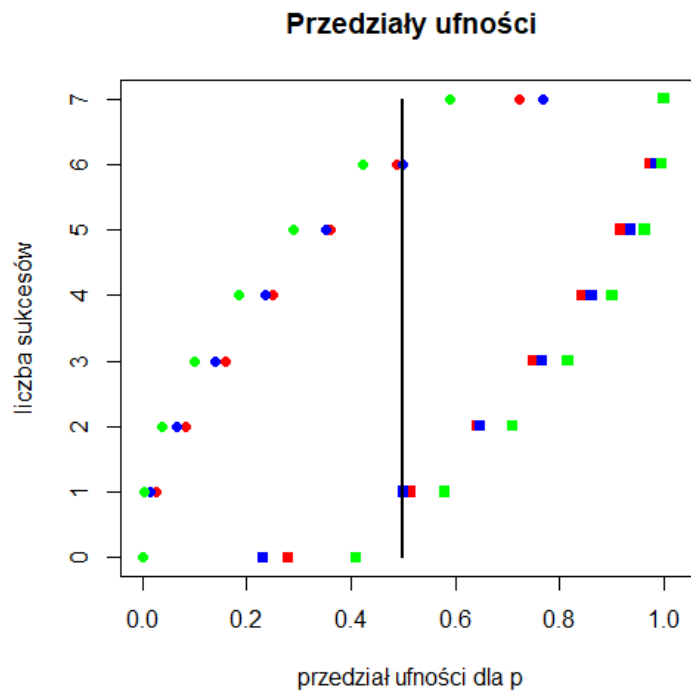
2.2 Implementacja w R

Do wykonania zadania użyliśmy biblioteki „prevalence”, która posiada funkcję *propCI* obliczającą przedział ufności w oparciu o metodę Wilsona („wilson”), Clopper-Pearsona („exact”) i Jeffreysa („jeffreys”). Istotną była także funkcja *rbinom* zwracająca zmienne losowe z rozkładu $B(n, p)$. Do narysowania wykresów użyliśmy funkcji *plot*, *lines* i *points*.

2.3 Wyniki

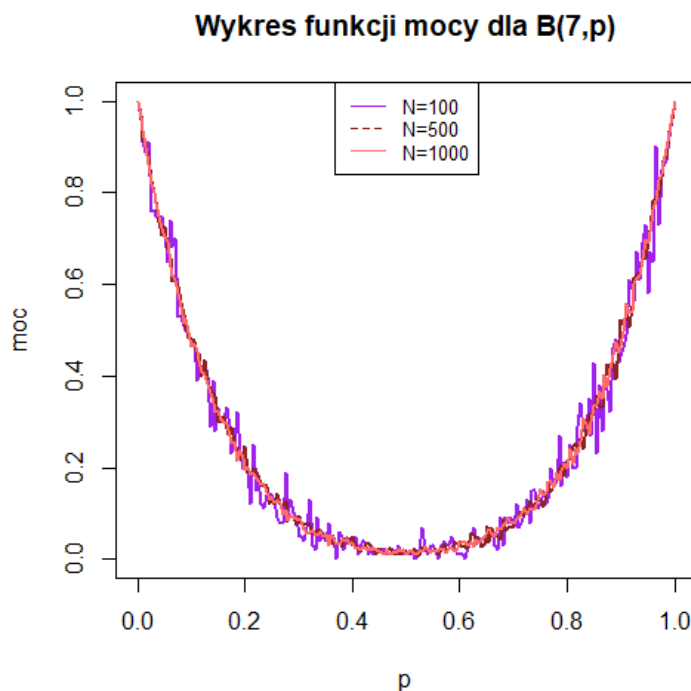
2.3.1 Zadanie 1.- $B(7,p)$

Na początek stworzyliśmy wykres przedstawiający otrzymane przedziały ufności dla p w zależności od możliwej wartości zmiennej losowej z rozkładu $B(7,p)$ (równej liczbie sukcesów w 7 niezależnych próbach z prawdopodobieństwem sukcesu p). Różnymi kolorami oznaczono przedziały ufności dla poszczególnych metod. Dla liczby sukcesów równej 0 dolna granica przedziału jest taka sama dla wszystkich metod, a dla 7 górna.



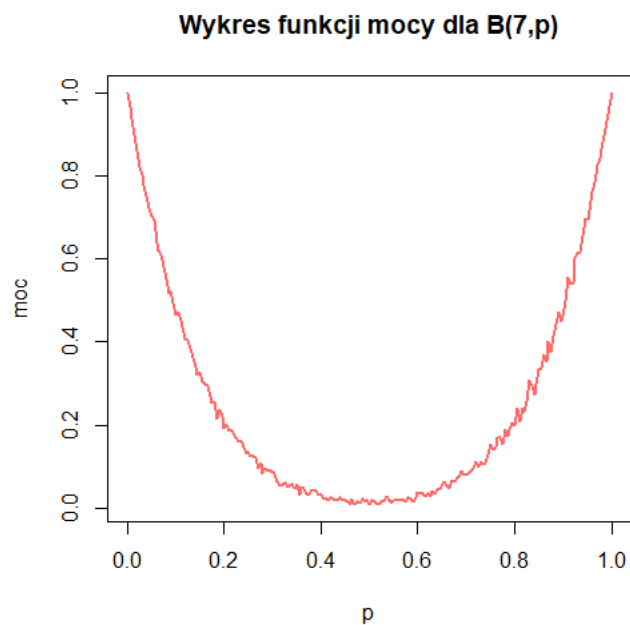
Rysunek 1: $B(7,p)$

Analizując wizualnie Rysunek 1. możemy zauważyć, że tylko dla liczby sukcesów równej 1 i 6 możliwe, że zachodzi sytuacja, w której $p = 0.5$ nie należy do jednego z przedziałów, podczas gdy do innych, owszem. Jednakże, dokładne obliczenia pokazują, że górna granica przedziału niebieskiego dla liczby sukcesów równej 1 ma wartość 0.5007975, zatem $p = 0.5$ zawiera się we wszystkich 3 przedziałach. Dla liczby sukcesów równej 6 dolna granica niebieskiego przedziału to 0.49920254, więc zachodzi ta sama sytuacja, co dla liczby sukcesów równej 1. Możemy, więc zauważyć, że $p = 0.5$ zawsze należy lub nie do wszystkich 3 przedziałów. Ta obserwacja prowadzi nas do wniosku, że wykres funkcji mocy dla wszystkich 3 testów w tym przypadku będzie taki sam.



Rysunek 2: Krzywe funkcji mocy w zależności od liczby powtórzeń Monte Carlo

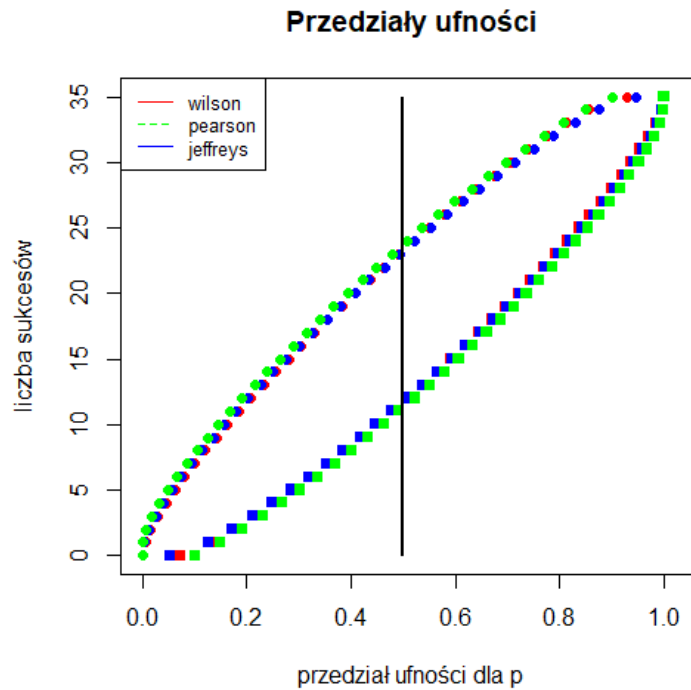
Sporządziliśmy wykresy funkcji mocy w zależności od liczby powtórzeń Monte Carlo. Wraz ze wzrostem liczby powtórzeń krzywa jest dokładniejsza. Ponadto wykres funkcji mocy wskazuje na fakt, że im bliższe jest p do 0.5 tym test słabszy. Szczególnie duży spadek mocy obserwujemy dla p z przedziału od 0.4 do 0.6. Zadowalające wyniki otrzymujemy dla p bliskich 0 i 1.



Rysunek 3: Wykres funkcji mocy dla $B(7,p)$

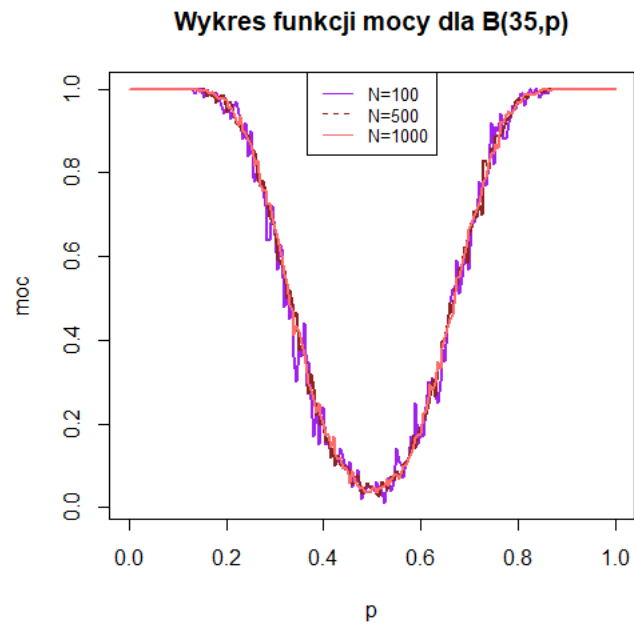
2.3.2 Zadanie 2.- $B(35,p)$

Analogicznie do zadania 1 na początek zbadaliśmy przedziały ufności dla poszczególnych metod. Podobnie jak w zadaniu 1. nie ma miejsca sytuacja, w której 0.5 należy do przedziału otrzymanego jedną z metod, podczas gdy do innych nie. W związku z tym wykres funkcji mocy będzie taki sam dla wszystkich 3 metod.

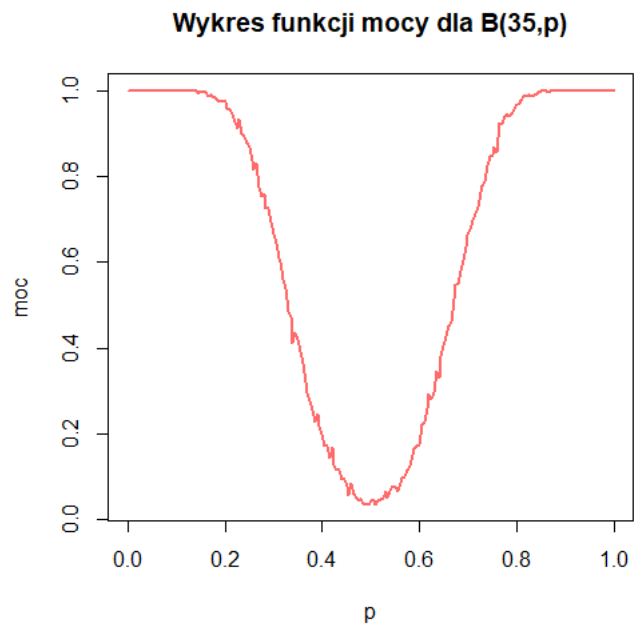


Rysunek 4: $B(35,p)$

Wraz ze wzrostem powtórzeń Monte Carlo otrzymujemy dokładniejszy obraz funkcji mocy. Zauważamy, że dla zmiennych z rozkładu $B(35, p)$ moc testu jest znacząco lepsza w porównaniu do $B(7, p)$. Test dla $p \in (0, 0.2) \cup (0.8, 1)$ prawie zawsze pozwala na odrzucenie fałszywej hipotezy zerowej. Obserwujemy też gwałtowniejszy spadek mocy niż w poprzednim zadaniu. Przedział, dla którego moc testu jest bardzo słaba, wynosząca prawie 0, jest znacząco mniejszy niż ten dla $B(7, p)$.

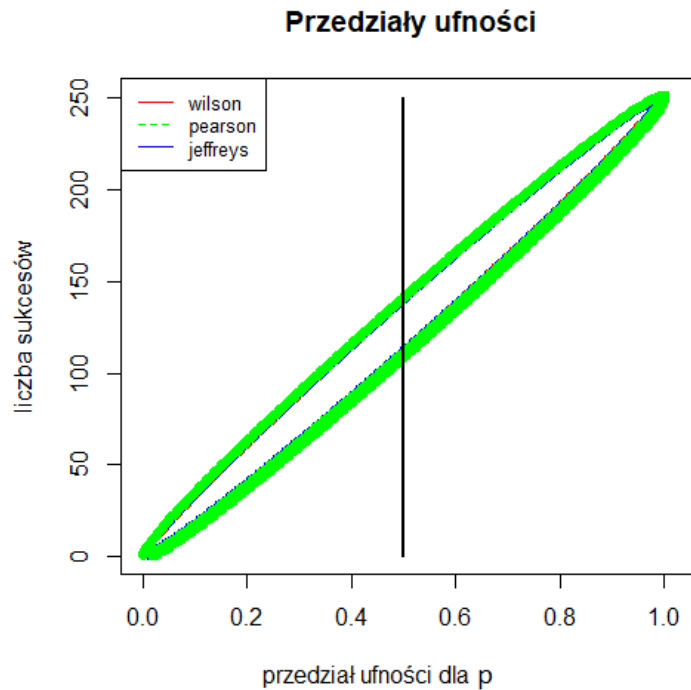


Rysunek 5: Krzywe funkcji mocy w zależności od liczby powtórzeń Monte Carlo



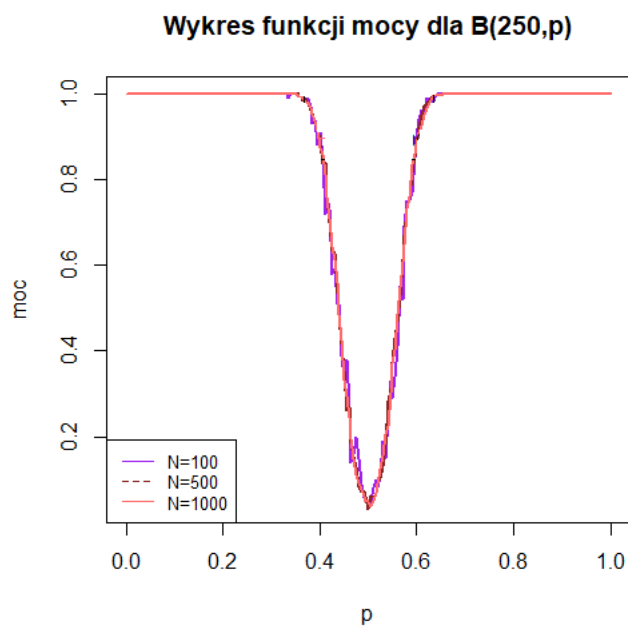
Rysunek 6: Wykres funkcji mocy dla $B(35, p)$

2.3.3 Zadanie 3.- $B(250, p)$

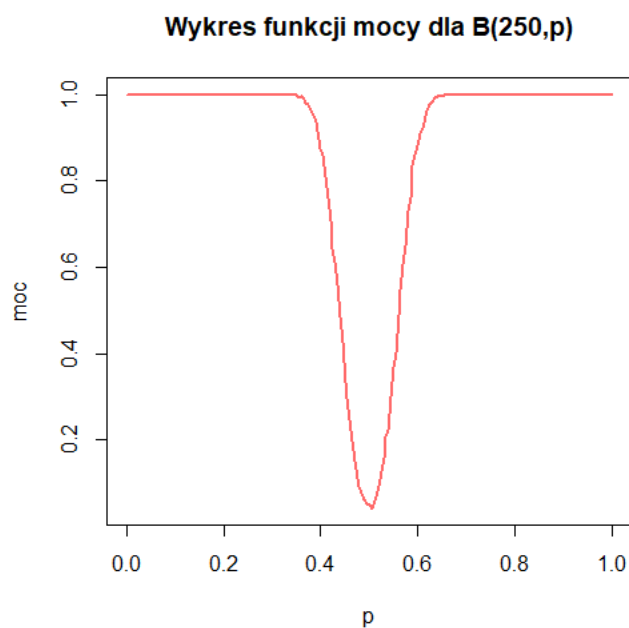


Rysunek 7: Przedziały ufności dla $B(250, p)$

Także i w tym zadaniu $p = 0.5$ zawsze należy lub nie do każdego z przedziałów ufności. Możemy też zauważyć, że prawdopodobnie największe przedziały otrzymuje się dla tych otrzymanych metodą Cloppera-Pearsona, choć różnice są niewielkie. Otrzymany wykres funkcji mocy metodą Monte Carlo jest identyczny dla wszystkich trzech testów. Prawdopodobieństwo odrzucenia hipotezy zerowej gdy $p \in (0, 0.4) \cup (0.6, 1)$ jest bliskie 1. Obserwujemy też jeszcze gwałtowniejszy spadek mocy niż ten w zadaniu 2.. Przedział, dla którego moc testu jest bardzo słaba jest w przybliżeniu dwukrotnie mniejszy, a ten, dla którego moc testu jest bardzo dobra jest dwukrotnie większy.



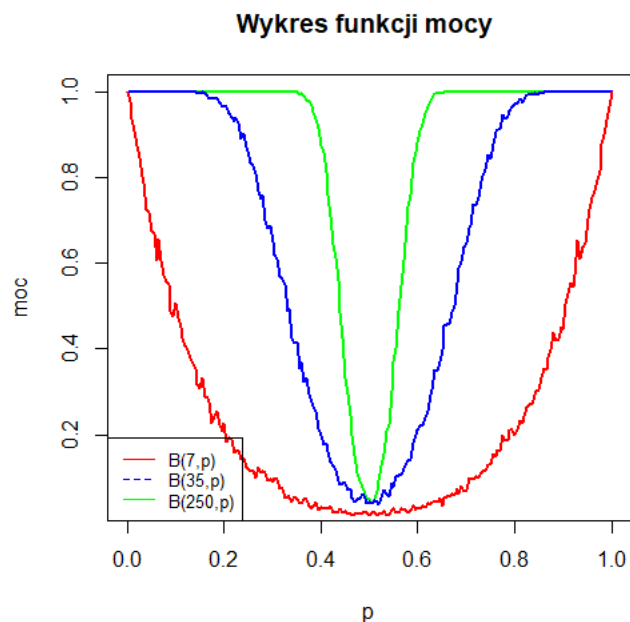
Rysunek 8: Krzywe funkcji mocy w zależności od liczby powtórzeń Monte Carlo



Rysunek 9: Wykres funkcji mocy dla $B(250, p)$

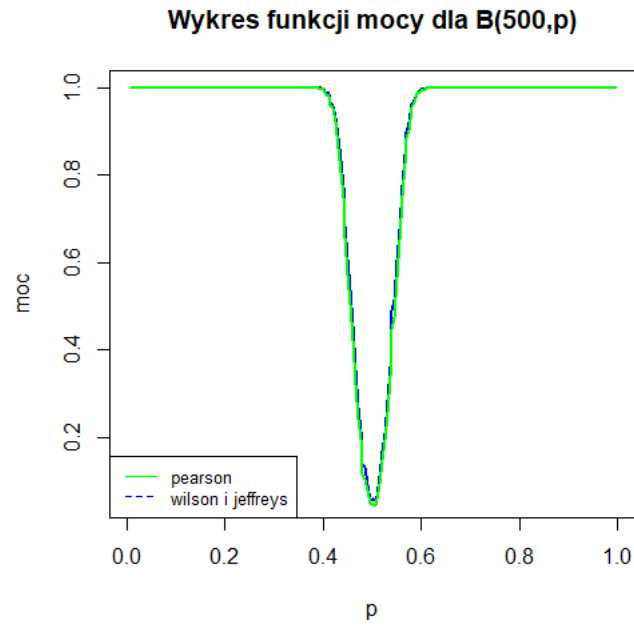
3 Wnioski

Pytanie o test jednostajnie najmocniejszy, a więc ten, który posiada największą moc jest przy dotychczasowej analizie jest nierozstrzygalne. Ponieważ, $p = 0.5$ dla naszych n zawsze



Rysunek 10: Porównanie funkcji mocy w zależności od zadanych w zadaniu n

należy lub nie należy jednocześnie do wszystkich przedziałów ufności, nawet większa liczba powtórzeń Monte Carlo nie przybliży nas do rozróżnienia mocy testu dla tych n . Dlatego też sprawdziliśmy jak zachowuje się funkcja mocy dla większych n takich jak 500 i 1000. Na podstawie sporządzonych wykresów stwierdzamy, że dla testów opartych o metodę Wilsona i Jeffreysa wyniki są wciąż identyczne, jednakże test oparty o metodę Pearsona dla p bliskiego 0.5 ma nieznacznie słabszą moc niż dwa pozostałe. Prawdopodobnie wynika to z nieznacznie dłuższych przedziałów ufności dla metody Pearsona w porównaniu do dwóch pozostałych metod. Z tego powodu uważamy, że test Wilsona lub Jeffreysa jest lepszym wyborem, jednakże różnice są bardzo małe. Ponadto obserwując wykresy funkcji mocy dla różnych metod widzimy, że im większe n tym lepsza moc testu. Wynika to z faktu, że mamy możliwość większej liczby niezależnych prób. Jednakże różnice w mocy testu między $n = 7$ i $n = 250$ są dużo większe (możemy zaobserwować znacząco lepsze wyniki przy zwiększeniu n) niż te między $n = 500$ i $n = 1000$.

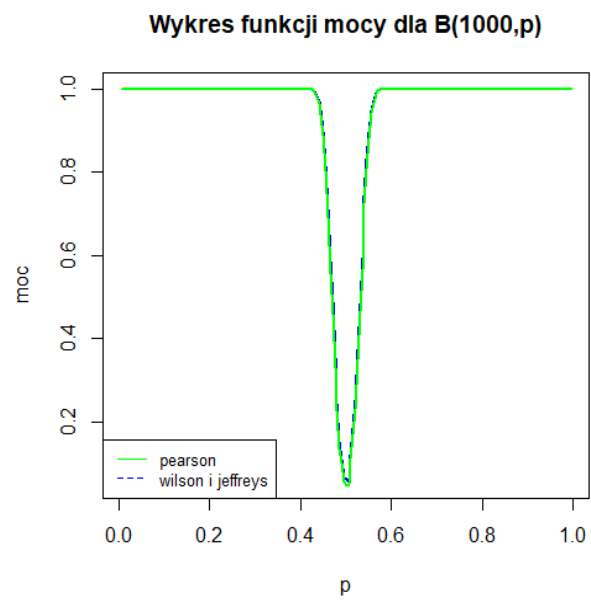


Rysunek 11: Wykres funkcji mocy dla $B(500, p)$

Wilson i Jeffreys i tak dają te same rezultaty, ale Pearson jest nieznacznie gorszy.

Literatura

- [1] https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval
- [2] <https://corplingstats.wordpress.com/2014/02/05/binomial-normal-wilson/>
- [3] <https://www.r-project.org>



Rysunek 12: Wykres funkcji mocy dla $B(1000, p)$