Sprawozdanie 5

Alicja Wiączkowska

2023-12-19

Contents

Wstęp

Dane są dwie grupy próbek:

- $X = (X_1, X_2, ..., X_n)$ iid. z rozkładu o ciągłej dystrybuancie F
- $Y = (Y_1, Y_2, ..., Y_m)$ iid. z rozkładu o ciągłej dystrybuancie G

Celem sprawodzania jest porównanie metod testowania hipotezy $H_0 : F = G$ (próbki pochodzą z tego samego rozkładu) przeciwko alternatywie $H_1 : F \neq G$ (próbkipochodzą z dwóch różnych rozkładów).

Analizie poddane zostały funkcje mocy testów:

- Wilcoxona oparty na statystyce $W = (T_{\varphi_1})^2$
- Ansari-Bradley'a oparty na statystyce $AB = (T_{\varphi_2})^2$

- Lepage'a oparty na statystyce $L = W + AB$
- Kołmogorowa-Smirnowa oparty na statystyce KS.

Postać klasycznej liniowej statystyki rangowej T_φ zależy od funkcji φ. Jeśli funkcja ta spełnia warunki:

- $\int_0^1 \varphi(u)du = 0$
- $\int_0^1 \varphi^2(u)du = 1$,

to przy prawdziwoścci hipotezy zerowej statystyka T_φ ma asymptotyczny rozkład standardowy normalny niezależnie od rozkładu wejściowych wektorów X i Y. Kwadrat tej statystyki zatem również nie zależy od rozkładów wejściowych wektorów.

Można sprawdzić, że funkcje $\varphi_1(u) = \sqrt{3}(2u-1)$ i $\varphi_2 = \sqrt{48}(\frac{1}{4} - |u - \frac{1}{2}|)$ wykorzystywane do konstrukcji statystyk testowych W i AB spełnają powyższe warunki. Statystyka testowa L jest sumą statystyk $W + AB$, zatem jej rozkład rówież nie zależy od rozkładu, z którego pochodzą wektory X i Y. Z tego powodu wartość krytyczną w teście Wilcoxona, Ansari-Bradley'a oraz Lepage'a można wyznaczyć globalnie, tzn. niezależnie od rozkładów, z których będą pochodzić wektory X i Y.

Z kolei statystyka KS zależy od maksymalnej odległości między dystrybuantami empirycznymi wyznaczonymi na podstawie X i Y. Jeśli oba wektory pochodzą z tego samego rozkładu, statystyka testowa będzie przyjjmować małe wartości, niezależnie od kształtu funkcji gęstości na podstawie której losowano wektory.

W symulacjach przyjęto poziom istotności $\alpha = 0.05$.

Zad 1 Empiryczne wyznaczanie wartości krytycznej dla statystyki testowej

Wszytkie badane statystyki testowe przyjmują tylko wartości nieujemne.

Wartość krytyczna, to liczba c, która dla danej statystyki testowej T ($T \in \{W, AB, L, KS\}$) i ustalonego poziomu istotności α spełnia

$$\mathbb{P}(T > c \mid H_0) = \alpha$$

Symuacja

W celu wyznaczenia wartości krytycznych dla poszczególnych testów wykonano 10000-krotnie symulację polegającą na wylosowaniu n obserwacji X i $m = n$ obserwacji Y z rozkładu $\mathcal{N}(0, 1)$, a następnie na ich podstawie obliczono statystyki testowe. W skutku dla każdej ze statystyk otrzymano 10000-elementowy wektor i na jego podstawie wyznaczono wartość c korzystając z kwantyli empirycznych. Ponieważ próbki X i Y były losowane z tego samego rozkładu, hipoteza zerowa jest prawdziwa. Wówczas można wyznaczyć $c = Q_{1-\alpha}$, gdzie $Q_{1-\alpha}$ jest kwantylem empirycznym statystyki T rzędu $1 - \alpha$.

Symulacje wykonano dla $n = 20$ oraz $n = 50$.

	W	AB	L	KS
n=m=20	3.780750	3.888000	5.898750	1.264911
n=m=50	3.790128	3.871488	6.007985	1.300000

W przypadkach $n = m = 20$ oraz $n = m = 50$ otrzymano zbliżone wyniki.

Uzasadnienie poprawności sposobu generowania wartości krytycznych

Jeśli rozkład statystyki testowej T jest znany i zadany dystrybuantą $\mathcal{F}$, wartość krytyczna c spełnia:

$$\alpha = \mathbb{P}(T > c \mid H_0) = 1 - \mathbb{P}(T \leq c \mid H_0) = \mathcal{F}(c)$$

Zatem skoro $\alpha = \mathcal{F}(c)$, to $c = \mathcal{F}^{-1}(c)$, gdzie $\mathcal{F}^{-1}(c)$ oznacza kwantyl rzędu c.

W przypadku nieznajomości rozkładu statystyki testowej T, jej dystrybuntę $\mathcal{F}$ należy przybliżyć dystrybuantą empiryczną $\mathcal{F}_M$. Aby ją wyznaczyć potrzebny jest wektor $(Z_1, Z_2, ..., Z_M)$ iid. próbek zmiennej losowej o rozkładzie zadanym szukaną dystrybuantą $\mathcal{F}$. na podstawie wektora $(Z_1, Z_2, ..., Z_M)$ można wyznaczyć dystrybuantę empiryczną $\mathcal{F}_M$, która przy $M \to \infty$ zbiega do rzeczywistej dystrybuanty $\mathcal{F}$.

Każdy z elementów wektora $(Z_1, Z_2, ..., Z_M)$ wyznaczono obliczając wartość danej statstyki T dla każdorazowo niezależnie wylosowanych prób X i Y, tak jak opisano w symulacji. Przyjęto $M = 10000$. Ponieważ wektor X oraz Y był wylosowany niezależnie z zadanego rozkładu do obliczenia wartości statystyki testowej Z_i, można przyjąć, że $(Z_1, Z_2, ..., Z_M)$ jest iid. próbą z rozkładu zadanego dystrybuantą $\mathcal{F}$. W takim razie wartość dystrybuanty empirycznej $\mathcal{F}_M(\alpha)$ dobrze przybliża rzeczywistą wartość $c = \mathcal{F}(\alpha)$.

Rozkład statystyki testowej nie zależy od rozkładu, z którego generowano wektory X i Y, dlatego wartości krytyczne c wyznaczone przy generowaniu próbek X, Y z $\mathcal{N}(0, 1)$ stosują się także, w sytuacji generowania próbek X, Y z innych rozkładów.

Zad 2 Wyznaczanie funkcji mocy testów statystycznych w zależności od parametru położenia rozkładu wektora Y

Moc testu γ to prawdopodobieństwo odrzucenia hipotezy zerowej H_0 przy prawdziwości alternatywy H_1 w rozpatrywanych przypadkach wyraża się wzorem:

$$\gamma = \mathbb{P}(T > c \mid H_1)$$

Symulacja

W celu wyznaczenia mocy każdego z badanych testów wykonano 10000-krotnie symulację polegającą na wylosowaniu n obserwacji X i $m = n$ obserwacji Y z zadanego rozkładu o różnych parametrach dla każdego z wektorów. Następnie na ich podstawie obliczono statystyki testowe. W skutku dla każdej ze statystyk otrzymano 10000-elementowy wektor $Z = (Z_1, ..., Z_{10^4})$ i na jego podstawie wyznaczono moc testu.

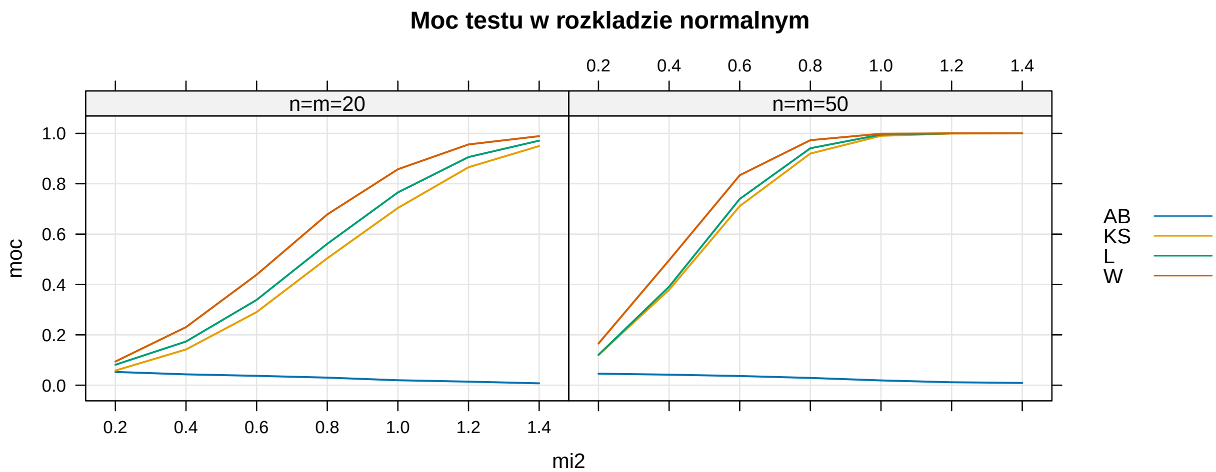
Ponieważ parametry rozkładu dla wektorów X i Y były różne, hipoteza alternatywna H_1 jest prawdziwa, a przybliżoną moc testu γ można empirycznie uzyskać ze wzoru:

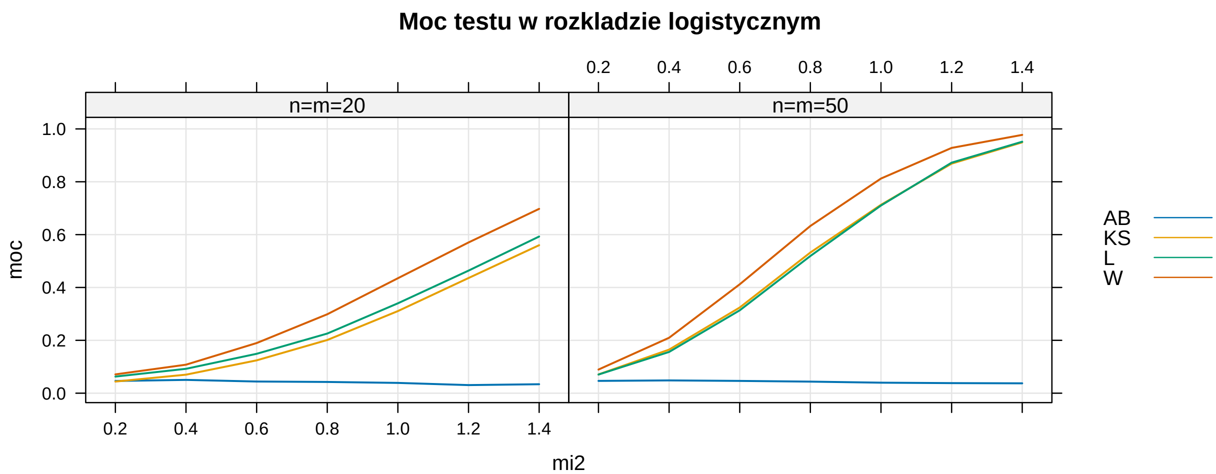
$$\gamma = \mathbb{P}(T > c \mid H_1) \approx \frac{|\{i : Z_i > c\}|}{10^4},$$

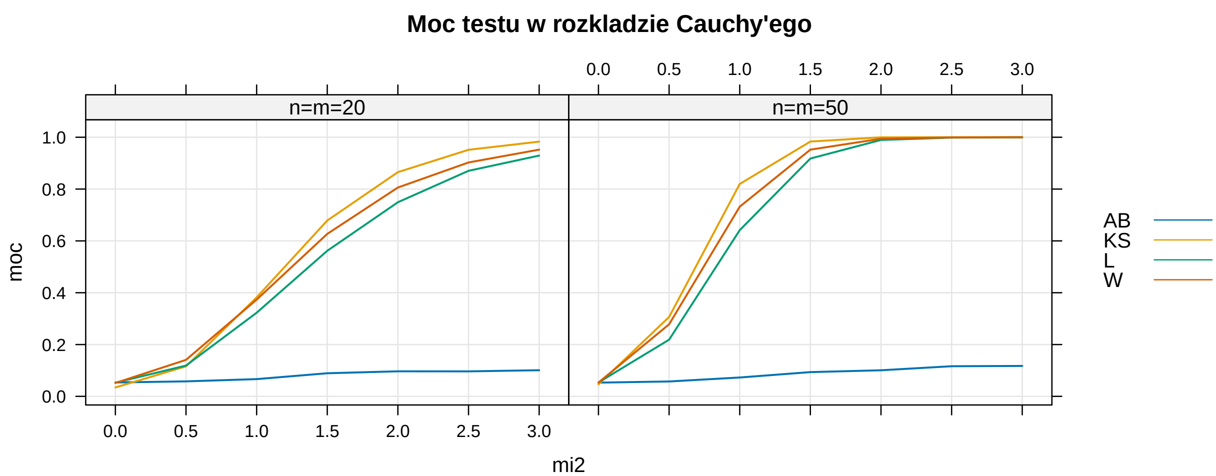
gdzie c jest wyznaczoną wcześniej wartością krytyczną.

Za parametry rozkładu dla wektora X przyjęto: $\mu_1 = 0$ i $\sigma_1 = 1$. Natomiast dla rozkłady wektora Y: parametr położenia μ_2 podlegał zmianom, a za parametr skali przyjęto $\sigma_2 = 1$.

Na wykresach poniżej przedstawiono moc testu w zależności od wartości parametru μ_2 wyznaczoną przy długościach wektorów X i Y ustalonych jako $n = m = 20$ oraz $n = m = 50$.







Dla wszystkich testów, z wyjątkiem Ansari-Bradley'a moc γ rośnie wraz ze wzrostem różnicy pomiędzy parametrami położenia w rozkładach, z których pochodzą wektory X i Y. Moc wszyskich testów jest na niskim poziomie, gdy μ_1 i μ_2 są do siebie zbliżone, co jest spowodowane występującym wówczas znacznym podobieństwem dystrybuant F i G. Ponadto moc tych testów Wilcoxona, Lepage'a oraz Kołmogorowa-Smirnowa jest tym lepsza, im więcej obserwacji znajdowało się w wektorach X i Y.

Najlepsze rezultaty dla rozkładu normalnego i logistycznego uzyskiwane są w teście Wilcoxona, choć testy

Lepage'a oraz Kołmogorowa-Smirnowa uzyskują porównywalną moc. Natomiast w przypadku Rozkładu Cauchy'ego test Kołmogorowa-Smirnowa sprawdza się najlepiej.

We wszystkich trzech przypadkach jednak moc testu Ansari-Bradley'a nie przekracza 0.2. Co więcej, w przypadku generowania wektorów X i Y z rozkładów normalnego lub logistycznego wartość γ zdaje się maleć wraz ze wzrostem odległości między μ_1 a μ_2 zarówno w przypadku $n = m = 20$ jak i dla $n = m = 50$. Z tego powodu należy uznać, że wyniki tego testu nie są dostatecznie wiarygodne, by stwierdzić czy zachodzi równość dystrybuant $F = G$ w przypadku różnicy parametrów położenia.

Zad 3 Wyznaczanie funkcji mocy testów statystycznych w zależności od parametru skali rozkładu wektora Y

Moc testu γ to prawdopodobieństwo odrzucenia hipotezy zerowej H_0 przy prawdziwości alternatywy H_1 w rozpatrywanych przypadkach wyraża się wzorem:

$$\gamma = \mathbb{P}(T > c \mid H_1)$$

Symulacja

W celu wyznaczenia mocy każdego z badanych testów wykonano 10000-krotnie symulację polegającą na wylosowaniu n obserwacji X i $m = n$ obserwacji Y z zadanego rozkładu o różnych parametrach dla każdego z wektorów. Następnie na ich podstawie obliczono statystyki testowe. W skutku dla każdej ze statystyk otrzymano 10000-elementowy wektor $Z = (Z_1, ..., Z_{10^4})$ i na jego podstawie wyznaczono moc testu.

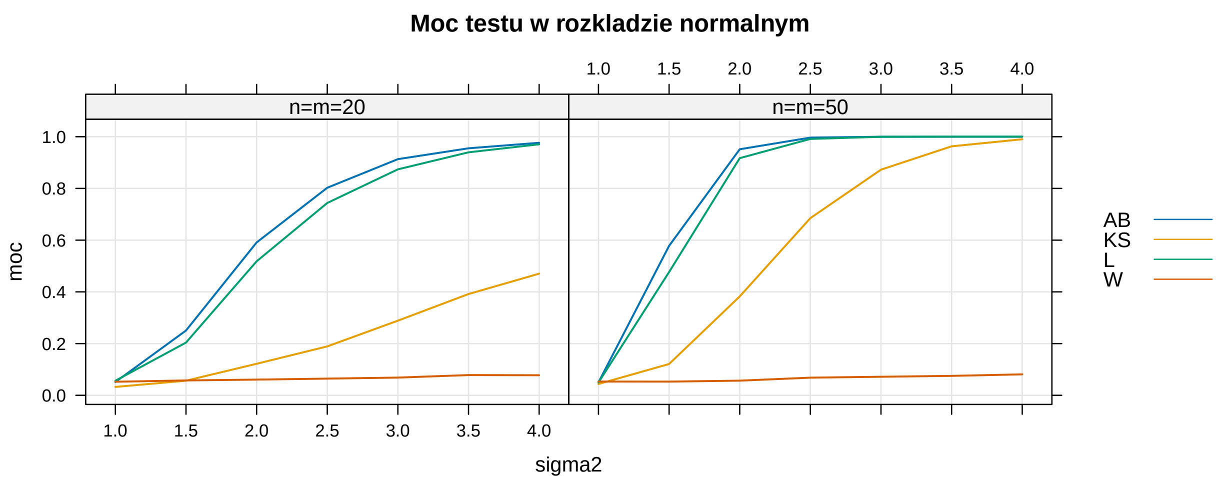
Ponieważ parametry rozkładu dla wektorów X i Y były różne, hipoteza alternatywna H_1 jest prawdziwa, a przybliżoną moc testu γ można empirycznie uzyskać ze wzoru:

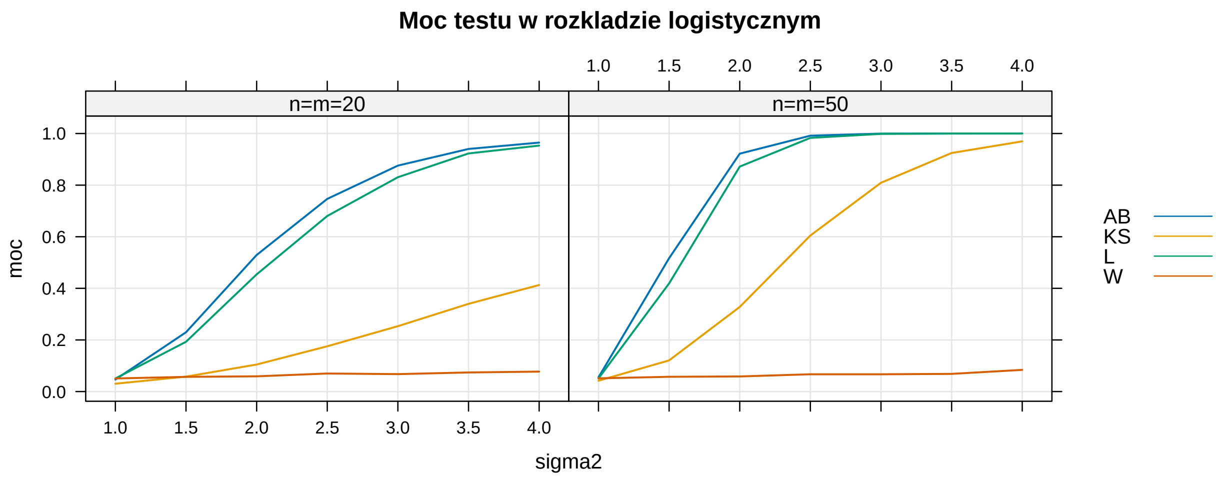
$$\gamma = \mathbb{P}(T > c \mid H_1) \approx \frac{|\{i : Z_i > c\}|}{10^4},$$

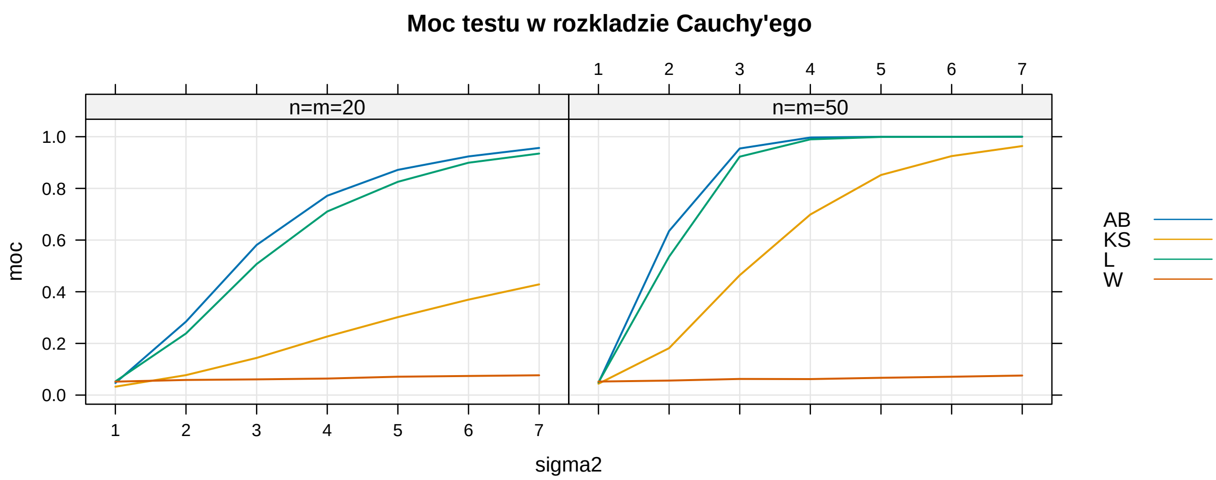
gdzie c jest wyznaczoną na początku wartością krytyczną.

Za parametry rozkładu dla wektora X przyjęto: $\mu_1 = 0$ i $\sigma_1 = 1$. Natomiast dla rozkłady wektora Y: parametr położenia $\mu_2 = 0$, a parametr skali σ_2 podlegał zmianom.

Na wykresach poniżej przedstawiono moc testu w zależności od wartości parametru σ_2 wyznaczoną przy długościach wektorów X i Y ustalonych jako $n = m = 20$ oraz $n = m = 50$.







W przypadku wszystkich rozkładów największą moc uzyskano dla testu Ansari-Bradley'a, test Lepage'a wypada porównywalnie do niego.

Test Kołmogorowa-Smirnowa osiąga mniej satysfakcjonującą moc, choć nadal może być użyteczny, w przeciwieństwie do testu Wilcoxona, dla którego γ utrzymuje się poniżej 0.1 dla wszystkich badanych rozkładów, niezlależnie od ilości próbek w wektorach X i Y.

Test Kołmogorowa-Smirnowa wypada gorzej w przypadku różnych parametrów skali, w stosunku do przypadku różnicy między paametrami położenia, ponieważ parametry te inaczej wpływają na zmianę dystrybuanty: μ wpływa na translację, a σ zmienia kształt krzywej analogicznie do przekształcenia liniowego. odległości między wartościami na przesuniętych dystrybuantach w badanych rozkładach są większe niż w przypadku przekształcenia liniowego krzywej dystrybuanty.

Zad 4 Wyznaczanie funkcji mocy testów statystycznych w zależności od parametru położenia i skali rozkładu wektora Y

Moc testu γ to prawdopodobieństwo odrzucenia hipotezy zerowej H_0 przy prawdziwości alternatywy H_1 w rozpatrywanych przypadkach wyraża się wzorem:

$$\gamma = \mathbb{P}(T > c \mid H_1)$$

Symulacja

W celu wyznaczenia mocy każdego z badanych testów wykonano 10000-krotnie symulację polegającą na wylosowaniu n obserwacji X i $m = n$ obserwacji Y z zadanego rozkładu o różnych parametrach dla każdego z wektorów. Następnie na ich podstawie obliczono statystyki testowe. W skutku dla każdej ze statystyk otrzymano 10000-elementowy wektor $Z = (Z_1, ..., Z_{10^4})$ i na jego podstawie wyznaczono moc testu.

Ponieważ parametry rozkładu dla wektorów X i Y były różne, hipoteza alternatywna H_1 jest prawdziwa, a przybliżoną moc testu γ można empirycznie uzyskać ze wzoru:

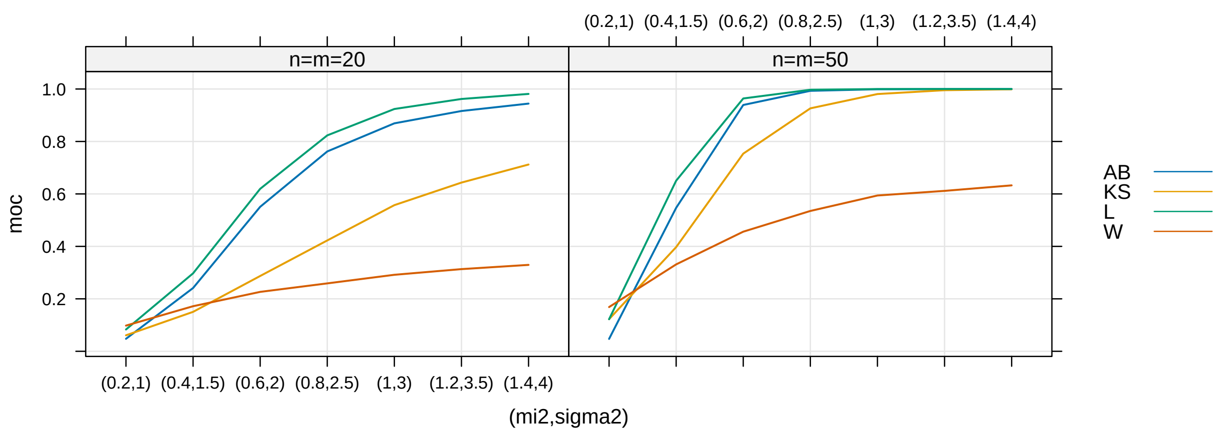
$$\gamma = \mathbb{P}(T > c \mid H_1) \approx \frac{|\{i : Z_i > c\}|}{10^4},$$

gdzie c jest wyznaczoną na początku wartością krytyczną.

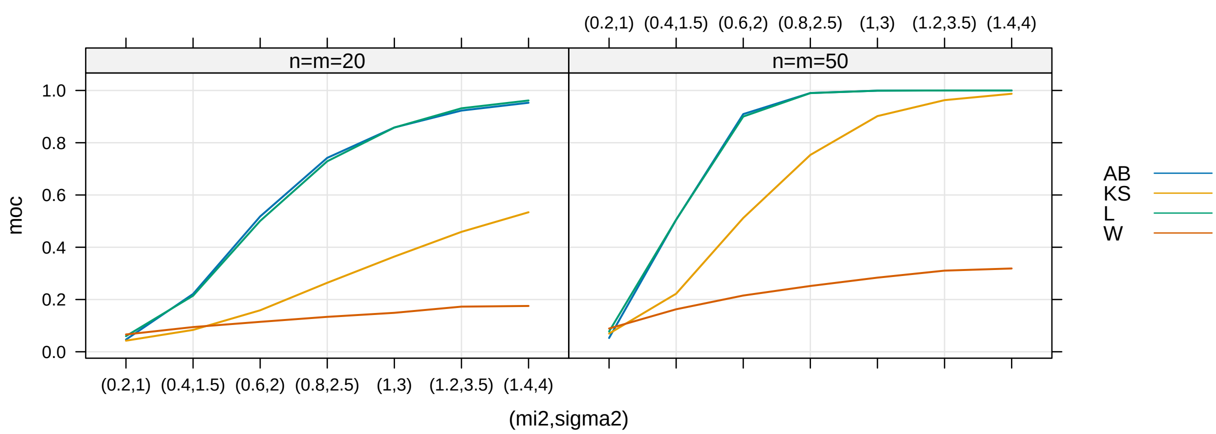
Za parametry rozkładu dla wektora X przyjęto: $\mu_1 = 0$ i $\sigma_1 = 1$. Natomiast dla rozkłady wektora Y: parametr położenia μ_2 oraz skali σ_2 podlegał zmianom.

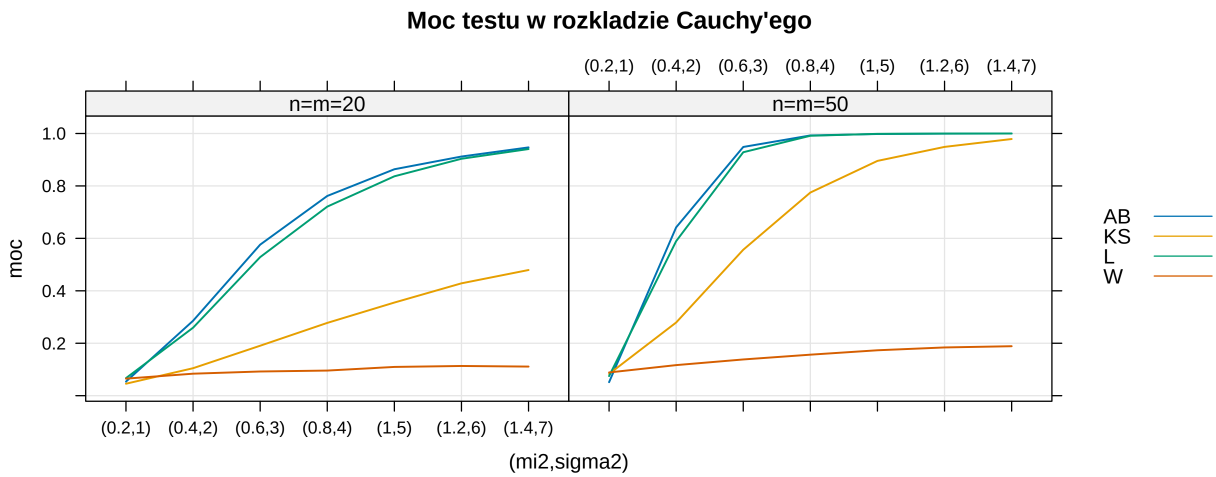
Na wykresach poniżej przedstawiono moc testu w zależności od wartości parametrów μ_2 i σ_2 wyznaczoną przy długościach wektorów X i Y ustalonych jako $n = m = 20$ oraz $n = m = 50$.

Moc testu w rozkladzie normalnym



Moc testu w rozkladzie logistycznym





Powyższe wykresy są pewną wypadkową wyników wyznaczonych we wcześniejszych rozważaniach. Test Lepege'a sprawdzał się bardzo dobrze zarówno przy znieniającym się parametrze przesunięcia μ_2 jak i skali σ_2. Dobrze sprawdził się również w połączeniu tych przypadków, gdy manipulowano oboma parametrami rozkładu zadającego dystrybuantę G.

Test Ansari-Bradley'a, który źle sprawdzał się w przypadku zmian parametru przesunięcia μ_2, ale usyskiwał najwyższą moc, gdy manipulowano parametrem skaliσ_2, w tej części uzyskał wartości γ porównywalne z testem Lepege'a.

Test Kołmogorowa-Smirnowa uzyskał nieco lepsze wyniki, niż w przypadku zmian jedynie parametru skali, jednak jego moc rośnie Mackenzie wolniej niż w teście Lepege'a i Ansari-Bradley'a.

Moc testu Wilcoxona rośnie bardzo powoli, jest lepsze niż w przypadku manipulacji samym parametrem skali, jednak *gamma* w tym przypadku jest niesatysfakcjonująca.

Oczywiście moc testu jest tym większa im liczebniejsze były wektory X i Y.

Podsumowując, można stwierdzić, że przy braku informacji, którymi parametrami różnią się rozkłady, najoptymalniej jest użyć testu Lepege'a do zbadania czy $F = G$.