Laboratorium: Lista 3
Testy rangowe

Statystyka w zastosowaniach

Sylwia Majchrowska
Matematyka

2 kwietnia 2016

Spis treści

1 Wstęp

Rozważany jest problem testowania dla dwóch prób: $X_1, ..., X_{n_1} \sim F(x)$, $Y_1, ..., Y_{n_2} \sim G(x)$, gdzie $H_0 : F(x) = G(x)$ vs $H_A : F(x) \neq G(x)$.

2 Zadanie pierwsze

Zadanie polegało na wyznaczeniu prawdopodobieństwa popełnienia błędu pierwszego rodzaju gdy $n_1 = 5$ i $n_2 = 10$, $n_1 = 10$ i $n_2 = 20$ oraz $n_1 = 20$ i $n_2 = 40$ dla $\alpha = 0.05$. Rozkład, z którego losowałam dane to standardowy rozkład normalny (oraz rozkład wykładniczy z parametrem $\lambda = 1$). Dodatkowo należało wyznaczyć standardowy błąd estymatora (przedział ufności). P-wartości zostały wyliczone w oparciu o:

a) dokładny test Wilcoxona

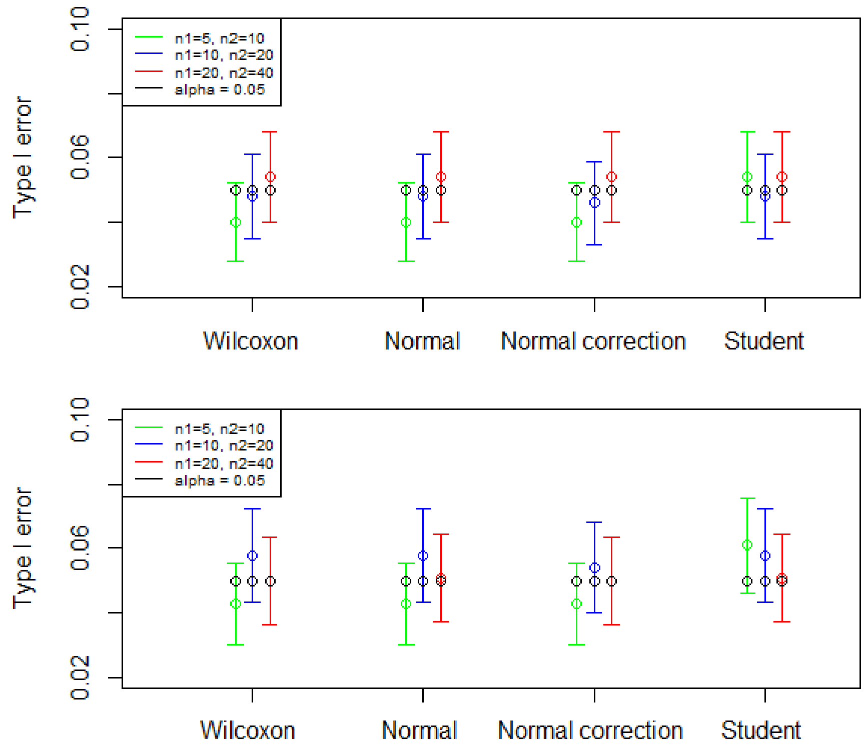
b) przybliżenie rozkładem normalnym (z poprawką na ciągłość i bez tej poprawki)

c) test Studenta, w którym wartości obserwacji są zastąpione rangami (liczonymi po połączeniu obu prób)

Estymatory wyznaczono w oparciu o 1000 powtórzeń eksperymentu.

Wyestymowane prawdopodobieństwa popełnienia błędu I rodzaju:

Test	$n_1 = 5$ $n_2 = 10$
dokładny test Wilcoxona	0.040 ± 0.013
przybliżenie rozkładem normalnym	0.040 ± 0.013
przybliżenie z poprawką na ciągłość	0.040 ± 0.013
test Studenta	0.054 ± 0.015
Test	$n_1 = 10$ $n_2 = 20$
dokładny test Wilcoxona	0.048 ± 0.014
przybliżenie rozkładem normalnym	0.048 ± 0.014
przybliżenie z poprawką na ciągłość	0.046 ± 0.013
test Studenta	0.048 ± 0.014
Test	$n_1 = 20$ $n_2 = 40$
dokładny test Wilcoxona	0.054 ± 0.015
przybliżenie rozkładem normalnym	0.054 ± 0.015
przybliżenie z poprawką na ciągłość	0.054 ± 0.015
test Studenta	0.054 ± 0.015



Rysunek 1: Wyestymowanie prawdopodobieństwa dla danych pochodzących z rozkładu normalnego (górny wykres) i wykładniczego (dolny wykres) oraz jego faktyczna wartość (czarne punkty).

Wyestymowane prawdopodobieństwo popełnienia błędu I rodzaju dla wszystkich rozważanych przypadków jest bliskie jego rzeczywistej wartości. Dodatkowo $\alpha = 0.5$ w każdym wariancie znalazła się w wyznaczonym dla wyestymowanych prawdopodobieństw przedziale. W przypadku danych pochodzących z rozkładu wykładniczego (dolny wykres z rys. nr 1) wraz ze wzrostem liczebności próby wyestymowana $\hat{\alpha}$ jest bliższa rzeczywistej wartości α, a dla przybliżeń p-wartości opartych na dokładnym teście Wilcoxona oraz zastosowania w nim przybliżenia rozkładem normalnym z poprawką Yetsa na ciągłość jest jej równa. Używając testu Studenta opartego na rangach widać, że prawdopodobieństwo popełnienia błędu I rodzaju jest trochę większe niż w pozostałych przypadkach, jednak wciąż bliskie 0.05, co jest związane z tym, że zastąpienie wartości obserwacji przez rangi sprawia, że otrzymana statystyka zbiega do statystyki dla testu Wilcoxona. Podobne wyniki dla dokładnego testu Wilcoxona i zastosowania przybliżenia rozkładem normalnym sugeruje, że rozkład normalny dobrze aproksymuje rozkład statystyki testu Wilcoxona-Manna-Whitneya.

Kod źródłowy zadania 1 w języku R:

```
set.seed(100)
alpha = 0.05
k = 1000 #liczba powtorzen eksperymentu

zad1 <- function(alfa,n1,n2,symNum){
  Xe=matrix(rnorm(n1*symNum),n1,symNum)
  Ye=matrix(rnorm(n2*symNum),n2,symNum)
  # a) test Wilcoxona
  pwarA = sapply(1:symNum, function(i){wilcox.test(Xe[,i
      ], Ye[,i], exact = TRUE)$p.value})
  WilcoxA = sum(pwarA<alfa)/symNum

  # b1) przyblizenie rozkladem normalnym bez poprawki
  pwarnbpB = sapply(1:symNum, function(i){wilcox.test(Xe
      [,i], Ye[,i], exact = FALSE, correct = FALSE)$p.
      value})
  WilcoxnbpB = sum(pwarnbpB<alfa)/symNum

  # b2) przyblizenie rozkladem normalnym z poprawka
  pwarnzpB = sapply(1:symNum, function(i){wilcox.test(Xe
      [,i], Ye[,i], exact = FALSE, correct = TRUE)$p.value
      })
  WilcoxnzpB = sum(pwarnzpB<alfa)/symNum

  # c) test Studenta dla rang
  Z = rbind(Xe,Ye)
  Z = apply(Z,2,rank)
  X = Z[1:n1,]
  Y = Z[n1+1:n2,]
  pwarC = sapply(1:symNum, function(i){t.test(X[,i], Y[,i
      ], var.equal = TRUE)$p.value})
  Student = sum(pwarC<alfa)/symNum

  wynik = c(WilcoxA, WilcoxnbpB, WilcoxnzpB, Student)
  return (wynik)
}
I = zad1(alpha,5,10,k)
II = zad1(alpha,10,20,k)
III = zad1(alpha,20,40,k)

bladI = qnorm(1-alpha/2)*sqrt(I*(1-I)/k)
bladII = qnorm(1-alpha/2)*sqrt(II*(1-II)/k)
bladIII = qnorm(1-alpha/2)*sqrt(III*(1-III)/k)
```

3 Zadanie drugie

Tym razem należało się skupić na porównaniu mocy testu Wilcoxona i testu Studenta w modelu z parametrem przesunięcia $G(x) = F(x - \Delta)$. Rozważano sytuacje gdy $\Delta = 1$ oraz funkcja $F(x)$ była dystrybuantą rozkładu:

a) $N(0,1)$,

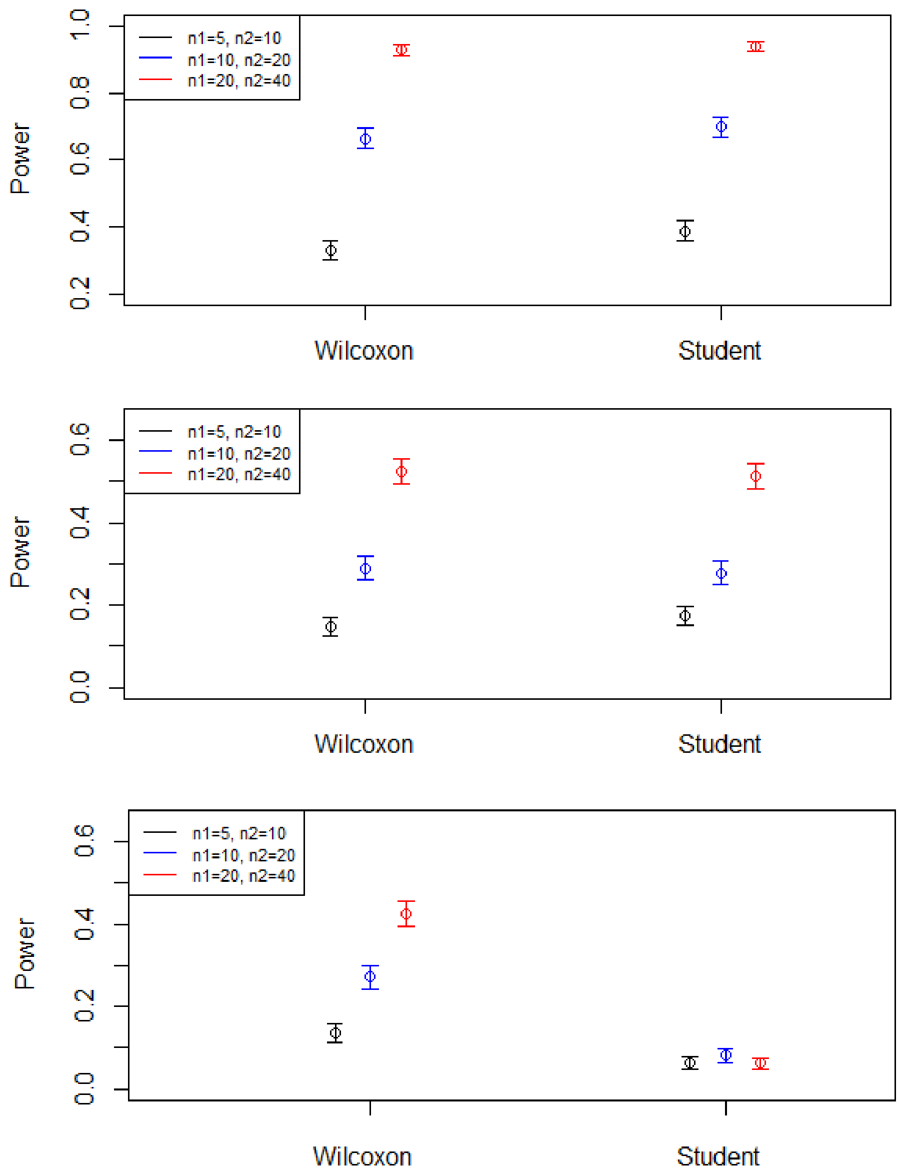
b) logistycznego(0,1),

c) Cauchy'ego.

Wyestymowana moc testu:

$F(x)$	Test Wilcoxona $n_1 = 5$ $n_2 = 10$	Test Studenta $n_1 = 5$ $n_2 = 10$
$N(0,1)$	0.330 ± 0.030	0.387 ± 0.031
logistyczny (0,1)	0.148 ± 0.023	0.173 ± 0.024
Cauchy'ego	0.136 ± 0.022	0.064 ± 0.016
$F(x)$	Test Wilcoxona $n_1 = 10$ $n_2 = 20$	Test Studenta $n_1 = 10$ $n_2 = 20$
$N(0,1)$	0.664 ± 0.030	0.698 ± 0.029
logistyczny (0,1)	0.298 ± 0.029	0.278 ± 0.028
Cauchy'ego	0.272 ± 0.028	0.082 ± 0.018
$F(x)$	Test Wilcoxona $n_1 = 20$ $n_2 = 40$	Test Studenta $n_1 = 20$ $n_2 = 40$
$N(0,1)$	0.928 ± 0.017	0.938 ± 0.015
logistyczny (0,1)	0.525 ± 0.031	0.514 ± 0.031
Cauchy'ego	0.426 ± 0.031	0.062 ± 0.015

Na podstawie otrzymanych wyników można stwierdzić, że test Wilcoxona w modelu z przesunięciem dla danych pochodzących z rozkładu normalnego radzi sobie gorzej niż test Studenta, natomiast dla danych pochodzących z rozkładu Cauchu'ego i logistycznego lepiej. W ogólności moce testów rosną wraz ze wzrostem liczebności próby (dla rozkładu Cauchy'ego utrzymują się na podobnym poziomie przy zastosowaniu testu Studenta). Warto zaznaczy, że dopiero wykorzystanie prób o liczebnościach $n_1 = 20$ i $n_2 = 40$ sprawia, iż skuteczność testów jest bliska bądź większa od 50% (biorąc pod uwagę test Studenta wyjątek stanowi rozkład Cauchy'ego - co jest spowodowane obserwacjami odstającymi - oraz rozkład normalny, przy którym zostały spełnione założenia dla testu Studenta(dane pochodzą z rozkładów normalnych o równych wariancjach)).



Rysunek 2: Wyestymowanie moce dla dystrybuanty rozkładu normalnego (górny wykres), logistycznego (środek) i Cauchy'ego (dolny wykres).

Kod źródłowy zadania 2 w języku R:

```
set.seed(100)
alpha = 0.05
k = 1000 #liczba powtorzen eksperymentu

zad2 <- function(alfa,n1,n2,symNum){

  Xe=matrix(rnorm(n1*symNum),n1,symNum)
  Ye=matrix(rnorm(n2*symNum),n2,symNum) - 1
  #Xe=matrix(rlogis(n1*symNum),n1,symNum)
  #Ye=matrix(rlogis(n2*symNum),n2,symNum) - 1
  #Xe=matrix(rcauchy(n1*symNum),n1,symNum)
  #Ye=matrix(rcauchy(n2*symNum),n2,symNum) - 1

  # test Wilcoxona
  pwarA = sapply(1:symNum, function(i){wilcox.test(Xe[,i
      ], Ye[,i], exact = TRUE)$p.value})
  Wilcox = sum(pwarA<alfa)/symNum

  # test Studenta
  pwarC = sapply(1:symNum, function(i){t.test(Xe[,i], Ye
      [,i], var.equal = TRUE)$p.value})
  Student = sum(pwarC<alfa)/symNum

  wynik = c(Wilcox, Student)
  return (wynik)
}
```

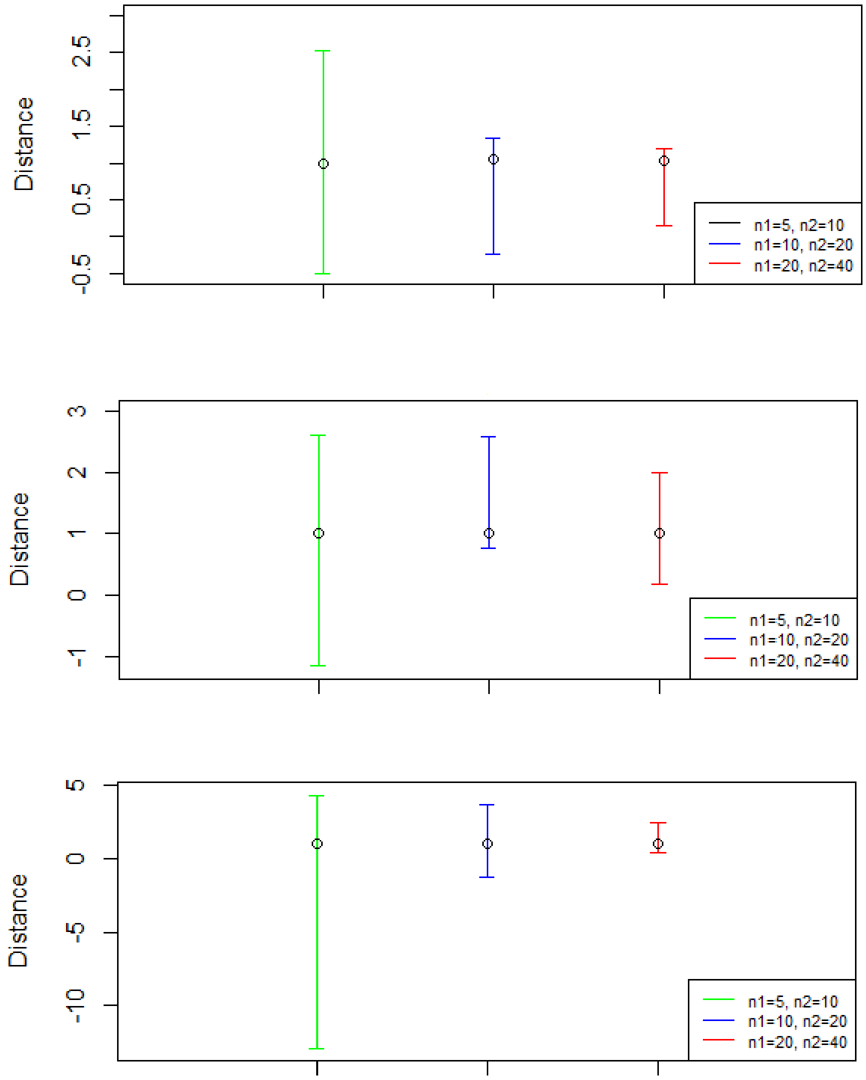
4 Zadanie trzecie

W oparciu o symulacje komputerowe należało oszacować jakość produkowanego przez funkcję wilcox.test parametru Δ. Rozmiary prób i rozkłady były takie same jak w zadaniu 2.

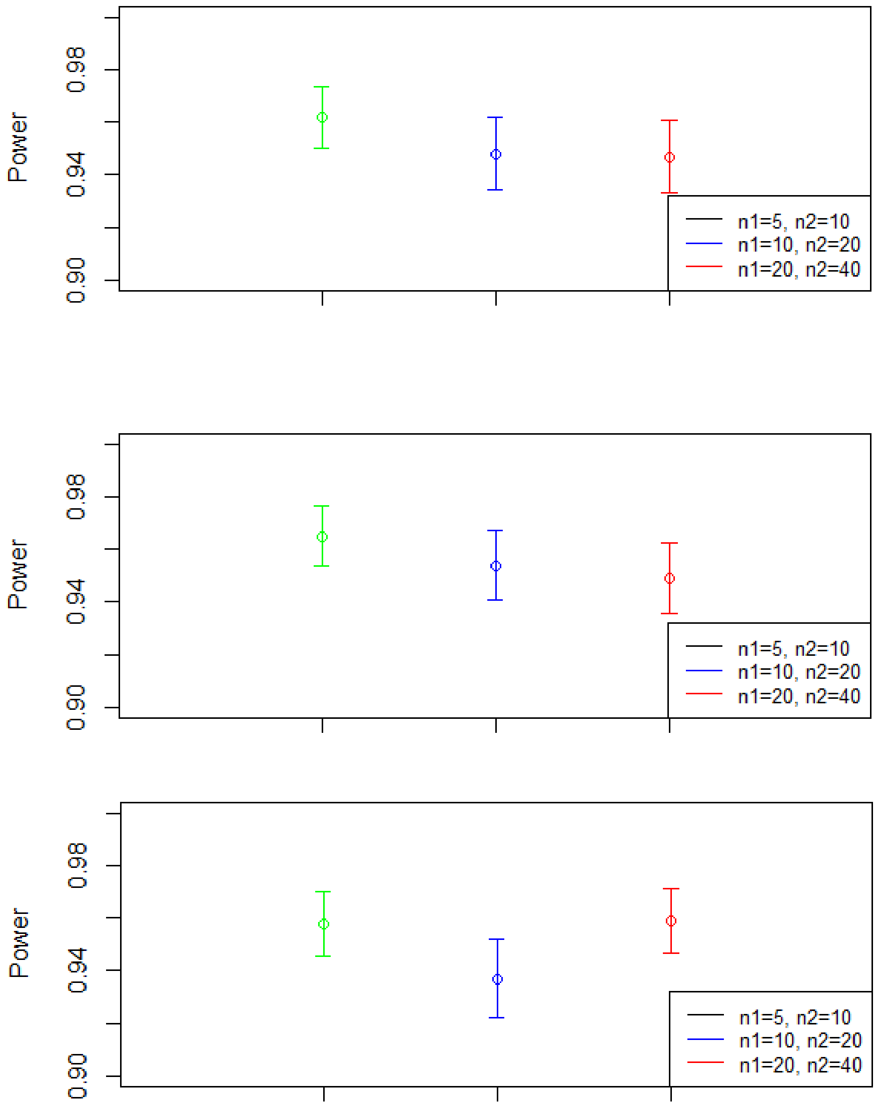
Dobrany parametr Δ oraz jego przedział ufności, produkowany przez funkcję wilcox.test dla pojedynczego eksperymentu:

Rozkład	$n_1 = 5$ $n_2 = 10$	Przedział ufności
$N(0,1)$	1	[-0.4982454, 2.5346679]
logistyczny (0,1)	1	[-1.1496410, 2.6152300]
Cauchy'ego	1	[-28.8493763, 4.7625887]
Rozkład	$n_1 = 10$ $n_2 = 20$	Przedział ufności
$N(0,1)$	1	[-0.2362745, 1.3322221]
logistyczny (0,1)	1	[0.7741937, 2.5919675]
Cauchy'ego	1	[-0.0560786, 1.5428109]
Rozkład	$n_1 = 20$ $n_2 = 40$	Przedział ufności
$N(0,1)$	1	[0.1445036, 1.1911338]
logistyczny (0,1)	1	[0.1827939, 1.9852246]
Cauchy'ego	1	[-0.8450537, 1.5862728]

Uśredniona pseudomediana jest bliska rzeczywistej wartości parametru Δ. Wyjątkiem są dane otrzymane dla rozkładu Cauchy'ego dla najmniejszej liczebności prób (rozkład ciężkoogonowy - brak EX, Var(X)). Dla przykładowego losowania (powyższa tabela oraz wykresy nr 3) rzeczywista wartość parametru przesunięcia mieści się w wyznaczonych przedziałach ufności. Jeżeli przyjrzymy się wykresom nr 4 to zauważymy, że prawdopodobieństwo tego zdarzenia jest dość duże - nieznacznie gorsze w przypadku rozkładu Cauchy'ego w porównaniu do reszty podpunktów (dla dystrybuant pochodzących z rozkładu normalnego oraz logistycznego). Warto zaznaczyć, że długość produkowanych przedziałów maleje wraz ze wzrostem wielkości prób. Generalnie produkowany przez funkcję *wilcox.test* przedział ufności nie jest zbyt dobry (bardzo szeroki) dla małej próby, a dużo lepszy dla większej ilości danych.



Rysunek 3: Wyprodukowane przez funkcję *wilcox.test* przedziały ufności dla dystrybuanty rozkładu normalnego (górny wykres), logistycznego (środek) i Cauchy'ego (dolny wykres) wraz z zaznaczoną rzeczywistą wartością $\Delta = 1$ (czarne punkty).



Rysunek 4: Wyznaczone prawdopodobieństwa faktu, że prawdziwa $\Delta = 1$ wpada do przedziału ufności dla dystrybuanty rozkładu normalnego (górny wykres), logistycznego (środek) i Cauchy'ego (dolny wykres).

```
alpha = 0.05
k = 1000 #liczba powtorzen eksperymentu

los <- function(n1,n2){

  Xe=rnorm(n1)
  Ye=rnorm(n2) - 1
  #Xe=rlogis(n1)
  #Ye=rlogis(n2) - 1
  #Xe=rcauchy(n1)
  #Ye=rcauchy(n2) - 1

  delta = wilcox.test(Xe, Ye, exact = TRUE, conf.int =
     TRUE)$estimate
  prze = wilcox.test(Xe, Ye, exact = TRUE, conf.int =
     TRUE)$conf.int
  dlu = prze[2]-prze[1]

  return (c(prze[1],prze[2],dlu, delta))
}

zad3 <- function(n1,n2,symNum){

  Xe=matrix(rnorm(n1*symNum),n1,symNum)
  Ye=matrix(rnorm(n2*symNum),n2,symNum) - 1
  #Xe=matrix(rlogis(n1*symNum),n1,symNum)
  #Ye=matrix(rlogis(n2*symNum),n2,symNum) - 1
  #Xe=matrix(rcauchy(n1*symNum),n1,symNum)
  #Ye=matrix(rcauchy(n2*symNum),n2,symNum) - 1

  delta = sapply(1:symNum, function(i){wilcox.test(Xe[,i
     ], Ye[,i], exact = TRUE, conf.int = TRUE)$estimate})
  prze = sapply(1:symNum, function(i){wilcox.test(Xe[,i],
      Ye[,i], exact = TRUE, conf.int = TRUE)$conf.int})
  zlicz = 0
  for (i in 1:symNum){
    if (1 > prze[1,i] && 1 < prze[2,i])
      zlicz = zlicz + 1
  }

  return (c(zlicz/symNum, delta))
}
```