Statystyka - raport 6

Łukasz Rębisz

29.01.2023

Implementacja wzorów teoretycznych

Niech $X_1, ..., X_n$ będą niezależnymi zmiennymi losowymi z rozkładu o ciągłej dystrybuancie F. Rozważmy problem testowania hipotezy

$$H_0 : F = F_0 \quad \text{vs} \quad H_1 : F \neq F_0,$$

gdzie F_0 jest znaną dystrybuantą.

Definiujemy nowe zmienne $U_1 = F_0(X_1), ..., U_n = F_0(X_n)$. Wtedy problem testowania (H_0, H_1) sprowadza się do weryfikowania

$$H_0 : U_k \sim U(0,1) \quad \text{vs} \quad H_1 : U_k \nsim U(0,1),$$

gdzie $U(0,1)$ oznacza rozkład jednostajny na odcinku $(0,1)$.

Poniższe testy badają pochodzenie danej próby z rozkładu $U(0,1)$.

Test chi-kwadrat Pearsona

Niech $A_1, ..., A_k$ będzie partycja odcinka (0,1). Niech $N_j = \#\{U_i \in A_j : i = 1, ..., n\}$, a $p_j = P_0(U_1 \in A_j), j = 1, ..., k$.

Klasyczny test chi-kwadrat Pearsona oparty jest na statystyce

$$P_k = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}.$$

Przy prawdziwości hipotezy zerowej statystyka P_k ma asymptotyczny rozkład $\chi^2(k-1)$. Hipotezę H_0 odrzucamy dla dużych wartości statystyki P_k.

```
P_k <- function(X, k){
  n <- length(X)

  vec_partition_left <- seq(0, 1, 1/k)
  vec_partition_right <- seq(0+1/k, 1, 1/k)

  sum_partition <- function(x){
  return(vec_partition_left < x &  x <= vec_partition_right)
  }
  vec_sum_partition <- Vectorize(sum_partition)

  vec_N <- as.vector(apply(vec_sum_partition(X), 1, sum))[1:k]

  P <- sum((vec_N - n/k)^2/(n/k))
  return(P)
}
```

Gładki test Neymana

Niech $\{b_j\}_{j\in\mathbb{N}}$ będzie układem ortonormalnym wielomianów Legendre'a w $L^2((0,1),du)$. Gładki test Neymana z k komponentami oparty jest na statystyce

$$N_k = \sum_{j=1}^{k} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_j(U_i) \right)^2 .$$

Przy prawdziwości hipotezy zerowej statystyka N_k ma asymptotyczny rozkład $\chi^2(k)$. Hipotezę H_0 odrzucamy dla dużych wartości statystyki N_k.

Zaimplementujumy unormowane wielomiany Legendre'a na (0, 1) (wielomiany stopni 1-8).

```
slegendre.polynomials(8, TRUE)
```

```
legendre_1 <- function(x){return(-1.732051 + 3.464102*x)}
legendre_2 <- function(x){return(2.236068 - 13.41641*x + 13.41641*x^2)}
legendre_3 <- function(x){return(-2.645751 + 31.74902*x - 79.37254*x^2 +
                                   52.91503*x^3)}
legendre_4 <- function(x){return(3 - 60*x + 270*x^2 - 420*x^3 + 210*x^4)}
legendre_5 <- function(x){return(-3.316625 + 99.49874*x - 696.4912*x^2
                                 + 1857.31*x^3 - 2089.474*x^4 + 835.7894*x^5)}
legendre_6 <- function(x){return(3.605551 - 151.4332*x + 1514.332*x^2
                                 - 6057.326*x^3 + 11357.49*x^4 - 9994.588*x^5
                                 + 3331.529*x^6)}
legendre_7 <- function(x){return(-3.872983 + 216.8871*x - 2927.975*x^2
                                 + 16266.53*x^3 - 44732.96*x^4 + 64415.46*x^5
                                 - 46522.28*x^6 + 13292.08*x^7)}
legendre_8 <- function(x){return(4.123106 - 296.8636*x + 5195.113*x^2
                                 - 38097.5*x^3 + 142865.6*x^4 - 297160.5*x^5
                                 + 346687.2*x^6 - 212257.5*x^7 + 53064.37*x^8)}

legendre <- function(x){
  return(c(legendre_1(x),legendre_2(x),legendre_3(x),legendre_4(x),
           legendre_5(x),legendre_6(x),legendre_7(x),legendre_8(x)))
}

vec_legendre <- Vectorize(legendre)
```

```
N_k <- function(X, k){
  n <- length(X)
  matrix_legendre <- as.matrix(vec_legendre(X)[1:k,])
  if(k==1){matrix_legendre <- t(matrix_legendre)}
  N <- sum((apply(matrix_legendre, 1, sum)/sqrt(n))^2)
  return(N)
}
```

Test Kołmogorowa-Smirnowa

Test oparty jest na statystyce

$$KS = \sqrt{n} \sup_{u\in(0,1)} |G_n(u) - u|,$$

gdzie G_n jest dystrybuantą empiryczną w próbie $U_1, ..., U_n$. Przy prawdziwości hipotezy zerowej statystyka KS ma asymptotyczny rozkład Kołmogorowa. Hipotezę H_0 odrzucamy dla dużych wartości statystyki KS.

```r
distribution1 <- function(a, X){
  return(mean(X <= a))
}

distribution2 <- function(a, X){
  return(mean(X < a))
}

KS <- function(X){
  f1 <- function(a){return(distribution1(a, X))}
  f2 <- function(a){return(distribution2(a, X))}

  n <- length(X)

  KS <- sqrt(n)*max(sapply(X, function(x) max(abs(f1(x) - x), abs(f2(x) - x))))
  return(KS)
}
```

Celem raportu jest zbadanie własności danych rozwiązań problemu testowego

$$H_0 : F = F_0 \quad \text{vs} \quad H_1 : F \neq F_0.$$

Przeanalizujemy moce testów (na poziomie istotności $\alpha = 0.05$):

i) test chi-kwadrat Pearsona oparty na statystyce P_4 oraz P_8 z równomierną partycją,

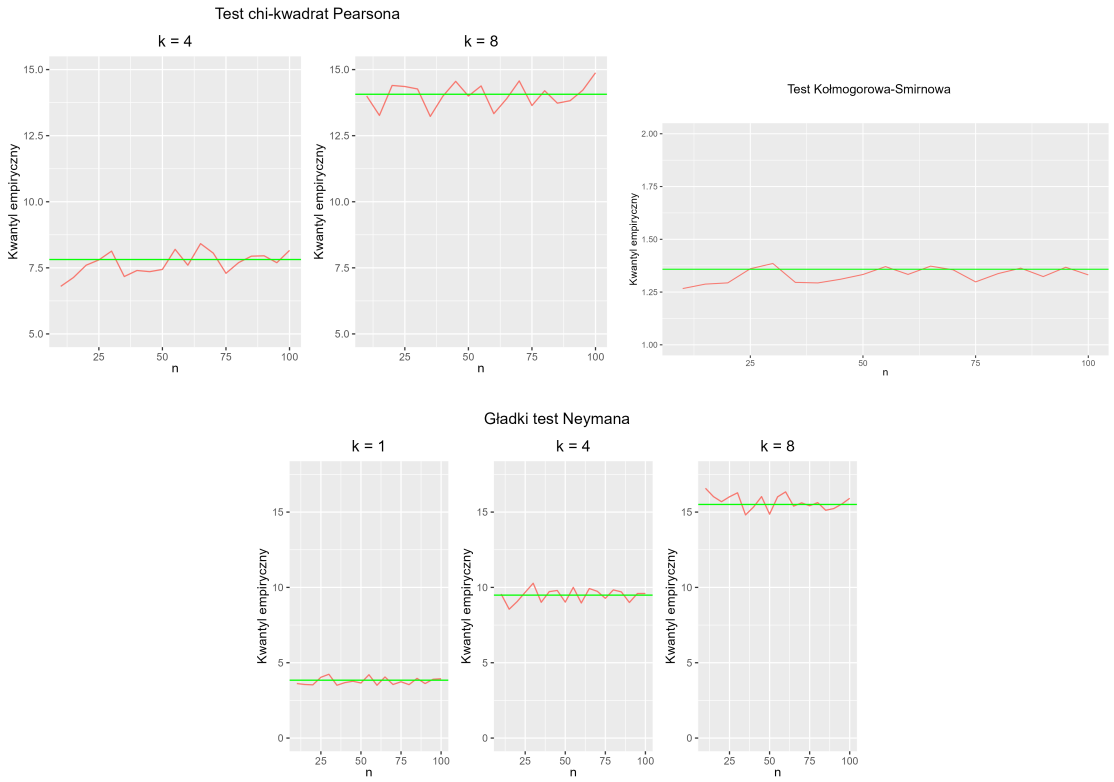
ii) gładki test Neymana z 1, 4 i 8 składowymi,

iii) test Kołmogorowa-Smirnowa oparty na statystyce KS.

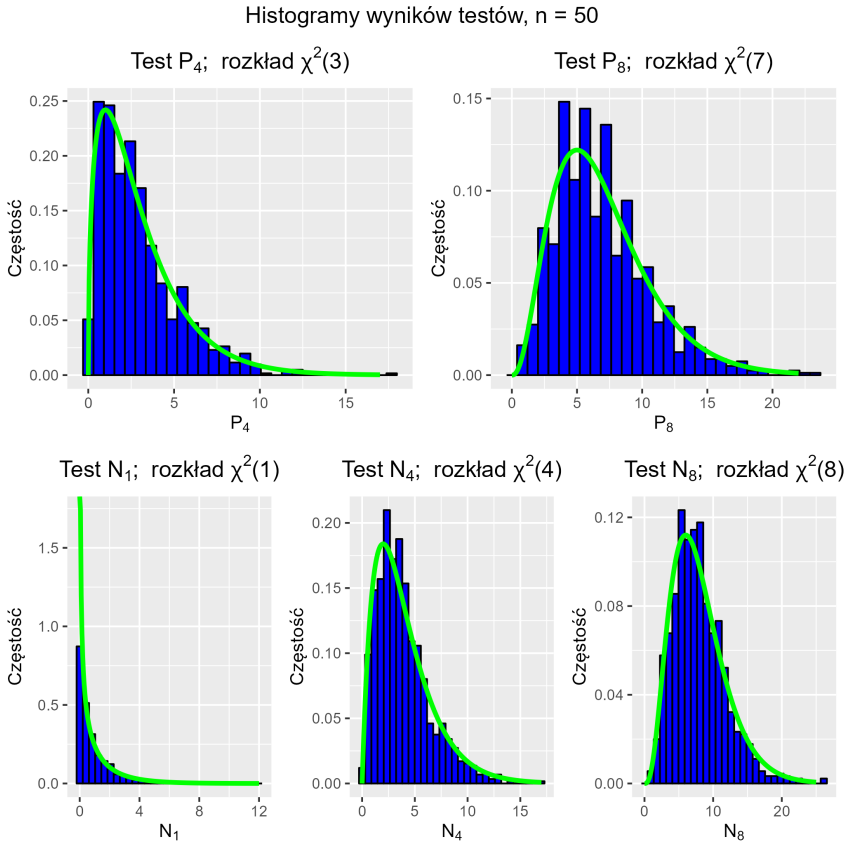
Zadanie 1

Wyznaczmy kwantyle empiryczne dla danych testów dla prób rozmiarów $n = 10, 20, \ldots, 100$ pochodzących z rozkładu $U(0, 1)$. Porównajmy otrzymane wyniki z kwantylami asymptotycznymi (teoretycznymi). Liczba powtórzeń eksperymentu $M = 1000$.

```r
ex_1 <- function(n){
  X <- runif(n)
  P_4 <- P_k(X, 4)
  P_8 <- P_k(X, 8)
  N_1 <- N_k(X, 1)
  N_4 <- N_k(X, 4)
  N_8 <- N_k(X, 8)
  KS <- KS(X)
  return(c(P_4, P_8, N_1, N_4, N_8, KS))
}
vec_ex_1 <- Vectorize(ex_1)
alpha <- 0.05
# Kwantyle teoretyczne:
vec_q <- c(qchisq(1-alpha, 3), qchisq(1-alpha, 7), qchisq(1-alpha, 1),
           qchisq(1-alpha, 4), qchisq(1-alpha, 8), sqrt(-log(alpha/2)/2))
```



Analiza powyższych wykresów pokazuje, że dla każdego z analizowanych testów **kwantyle empiryczne** oscylują wokół wartości danego **kwantylu asymptotycznego**. Nawet dla prób małych rozmiarów wartości kwantyli empirycznych są zbliżone do wartości teoretycznych. Zatem przyjęcie w każdym z testów wartości asymptotycznej za wartość kwantylu jest uzasadnione. Zwłaszcza w przypadku testu Kołmogorowa-Smirnowa wartości empiryczne są bardzo dobrym przybliżeniem wartości asymptotycznej. Sprawdźmy, jak blisko danego rozkładu asymptotycznego jest rozkład wyników pozostałych testów, ustalmy rozmiar prób $n = 50$.



Analiza powyższych histogramów pokazuje, że rozkłady wyników badanych testów chi-kwadrat Pearsona oraz gładkiego testu Neymana są bardzo bliskiemu odpowiedniego rozkładu asymptotycznego. Kształty otrzymanych histogramów są zbliżone do krzywych odpowiednich rozkładów. Jedynie wartości otrzymane dla testu P_8 świadczą o wynikach, które nie odpowiadają dokładnie rozkładowi asymptotycznemu.

Niemniej jednak powyższe wyniki świadczą o tym, że wykonując badane testy, możemy podejmować decyzje na podstawie porównania wyników testów z kwantylami asymptotycznymi.

Zadanie 2

W zadaniu omówimy **metodę eliminacji** umożliwiającą generację rozkładu F zadanego przez gęstość f. Metoda opiera się na oszacowaniu funkcji gęstości f przez gęstość g pochodzącą z rozkładu, z którego umiemy generować proste próby losowe, tzn. gęstość g musi spełniać warunek

$$f(s) \leq M \cdot g(s) \quad \text{dla wszystkich } s \quad \text{oraz pewnej stałej } M.$$

Algorytm postępowania:

1. Generujemy niezależne zmienne $X \sim g$ i $U \sim U[0,1]$.
2. Jeśli $U \leq f(X)/(M \cdot g(X))$, przyjmujemy $Y = X$.
3. W przeciwnym wypadku powracamy do punktu 1.

Otrzymana zmienna Y ma gęstość f.

Uzasadnienie:

$$P(Y \leq y) = P\left(X \leq y | U \leq \frac{f(X)}{M \cdot g(X)}\right) = \frac{P\left(X \leq y,\ U \leq \frac{f(X)}{M \cdot g(X)}\right)}{P(U \leq \frac{f(X)}{M \cdot g(X)})} =$$

$$\frac{\int_{-\infty}^{y}\left(\int_0^{f(x)/Mg(x)} du\right) g(x)dx}{\int_{-\infty}^{\infty}\left(\int_0^{f(x)/Mg(x)} du\right) g(x)dx} = \frac{\frac{1}{M}\int_{-\infty}^{y} f(x)dx}{\frac{1}{M}\int_{-\infty}^{\infty} f(x)dx} = \int_{-\infty}^{y} f(x)dx = F(y).$$

Dzięki powyższej metodzie możemy losować próby losowe z dowolnych rozkładów, które potrafimy oszacować z góry przez rozkład, z którego umiemy generować proste próby losowe. Liczba potrzebnych kroków algorytmu do uzyskania próby danego rozmiaru zależy od dokładności oszacowania $U \leq f(X)/(M \cdot g(X))$.

Implementacja

```
# gdy G ~ U(0,1)
elimination_method <- function(n, f, M){
  Y <- c()
  while(length(Y)<n){
    X <- runif(1)
    U <- runif(1)
    if(U <= (f(X)/M)) Y <- c(Y,X)
  }
  return(Y)
}
```

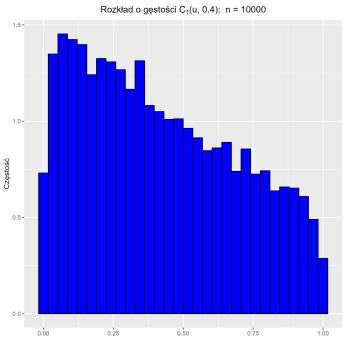
Zadanie 3

Zadanie polega na wygenerowaniu n obserwacji z rozkładu o gęstości

$$C_1(u, 0.4) = 1 + 0.4\cos(\pi u), \quad u \in (0, 1).$$

Na podstawie wylosowanych prób losowych wyznaczymy następnie empirycznie (1000 powtórzeń eksperymentu) moce testów z zadania 1.

```
C_1 <- function(u){1 + 0.4*cos(pi*u)}
ex_3 <- function(n){elimination_method(n, C_1, 1.4)}
vec_ex_3 <- Vectorize(ex_3)
```



Analiza powyższego histogramu pokazuje, że badany rozkład NIE ma rozkładu, który można w przybliżeniu uznać za rozkład jednostajny $U(0, 1)$. Rozkład ten jest jednak bliski jednostajnemu - w próbach małych rozmiarów odstępstwo od rozkładu jednostajnego może być niewykrywalne. Stąd spodziewamy się, że testy

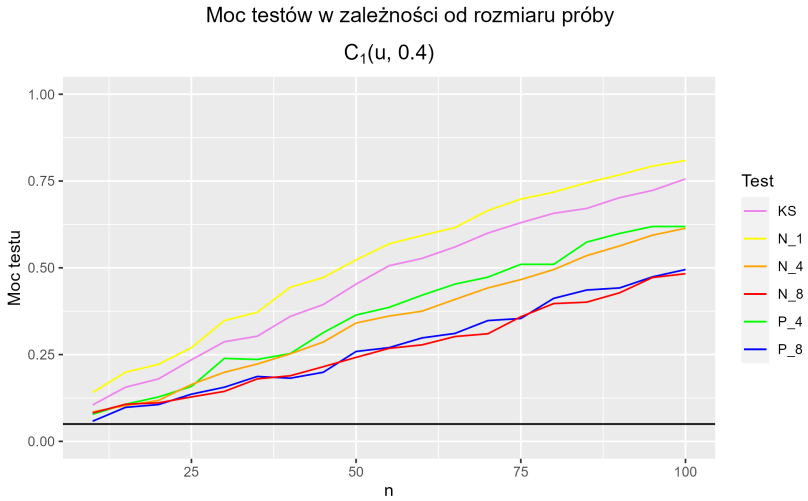
badające, czy dany rozkład jest rozkładem $U(0,1)$ będę miały coraz większą moc wraz ze wzrostem rozmiaru próby.

Wyznaczmy moce poszczególnych testów badających, czy otrzymany rozkład o gęstości $C_1(u, 0.4) = 1 + 0.4\cos(\pi u), u \in (0,1)$ ma rozkład $U(0,1)$.

```
vector_of_statistics <- function(X){
  P_4 <- P_k(X, 4)
  P_8 <- P_k(X, 8)
  N_1 <- N_k(X, 1)
  N_4 <- N_k(X, 4)
  N_8 <- N_k(X, 8)
  KS <- KS(X)
  return(c(P_4, P_8, N_1, N_4, N_8, KS))
}

power_of_tests <- function(n){
  X <- ex_3(n)
  return(vector_of_statistics(X) > vec_q)
}

M <- 1000
power_of_tests_n <- function(n){
  set.seed(1)
  results_tmp <- replicate(M, power_of_tests(n))
  return(apply(results_tmp, 1, mean))
}
vec_power_of_tests_n <- Vectorize(power_of_tests_n)
results_power <- vec_power_of_tests_n(vec_n)
```



Analiza powyższego wykresu pokazuje, że:

- wartości mocy wszystkich badanych testów rosną wraz ze wzrostem rozmiaru próby (spodziewany wynik). Zatem empirycznie pokazuje to, że wszystkie badane testy rzeczywiście badają jednostajność rozkładu danej próby.
- Moce poszczególnych testów rosną w sposób w przybliżeniu monotoniczny od wartości bliskiej założonemu poziomowi istotności $\alpha = 0.05$. Maskymalana moc dla prób rozmiaru 100 wynosi ok. 0,8.
- Największą moc ma w tym przypadku gładki test Neymana N_1, nieco gorszą moc wykazuje test

Kołmogorowa-Smirnowa.

- Najmniejszą moc mają testy N_8 i P_8. Jest to interesujący fakt, bowiem zwiększenie dokładności testów (większa liczba wielomianów w teście Neymana czy gęstsza partycja odcinka (0,1) w teście Pearsona) intuicyjnie powinno spowodować zwiększenie mocy testów.

Zadanie 4

Powtórzmy eksperyment numeryczny z poprzedniego zadania dla rozkładów o gęstościach

$$C_j(u,\rho) = 1 + \rho\cos(j\pi u), \quad u \in (0,1)$$

i) $j = 2, \rho = 0.5$,

ii) $j = 3, \rho = 0.5$,

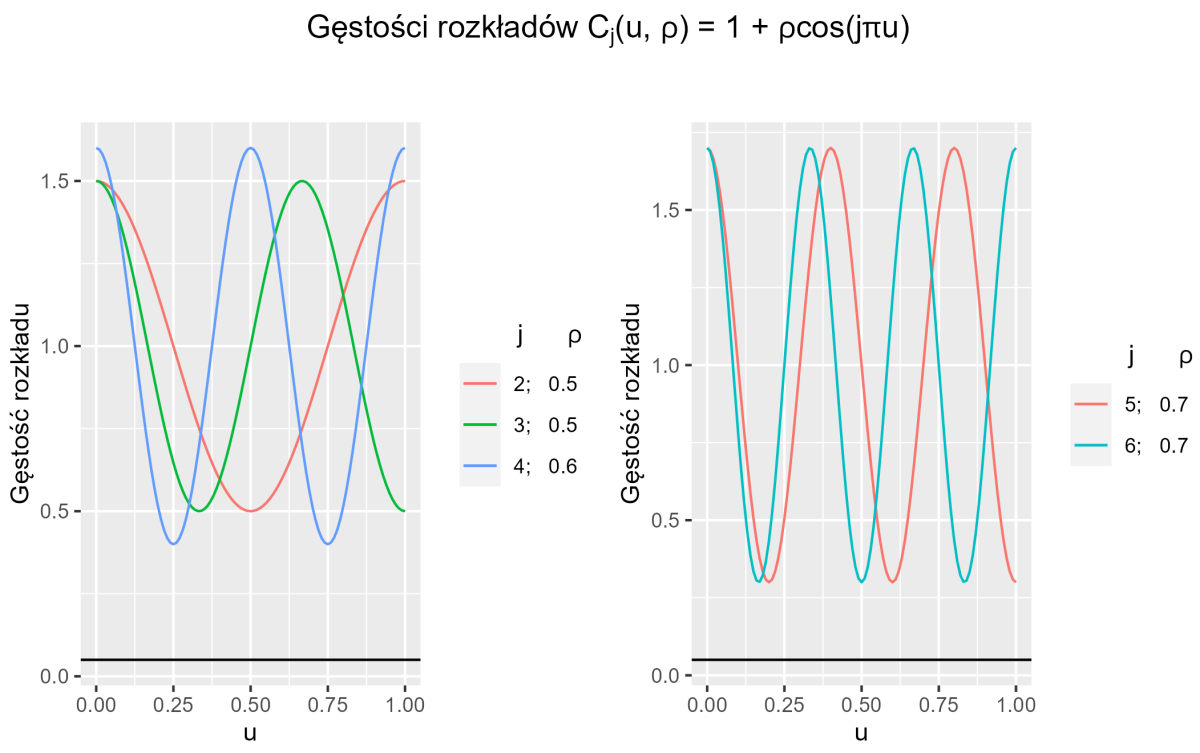
iii) $j = 4, \rho = 0.6$,

iv) $j = 5, \rho = 0.7$,

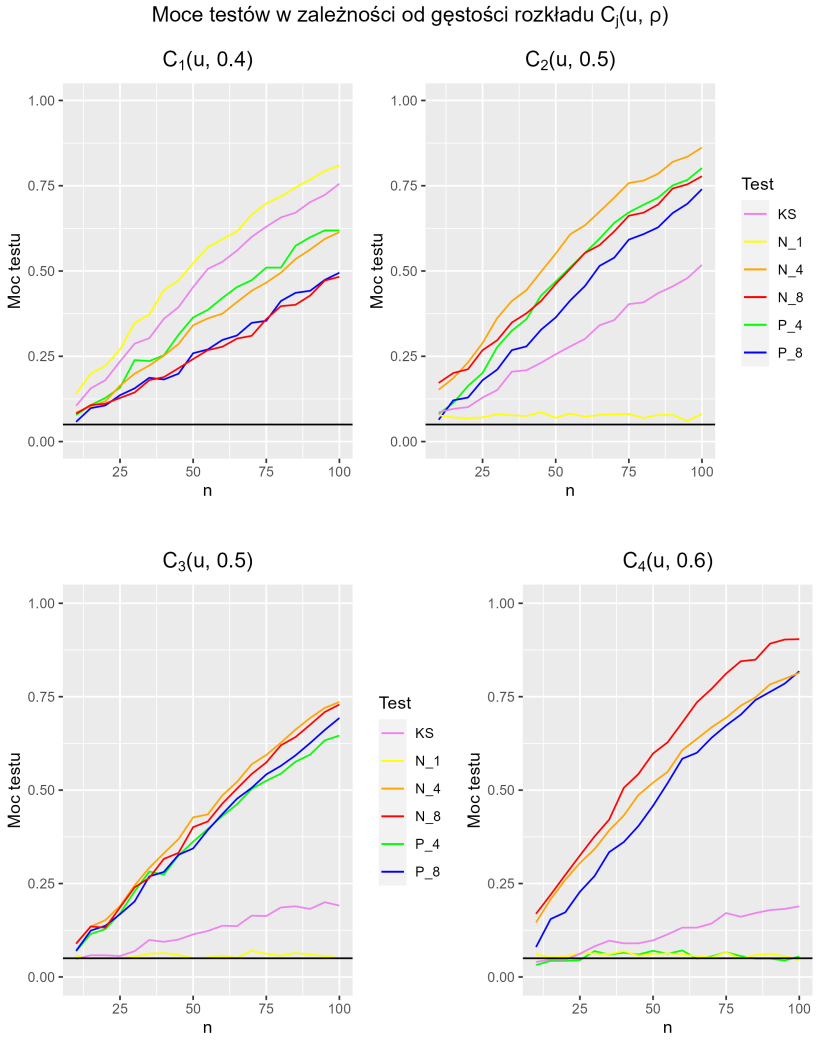
v) $j = 6, \rho = 0.7$.

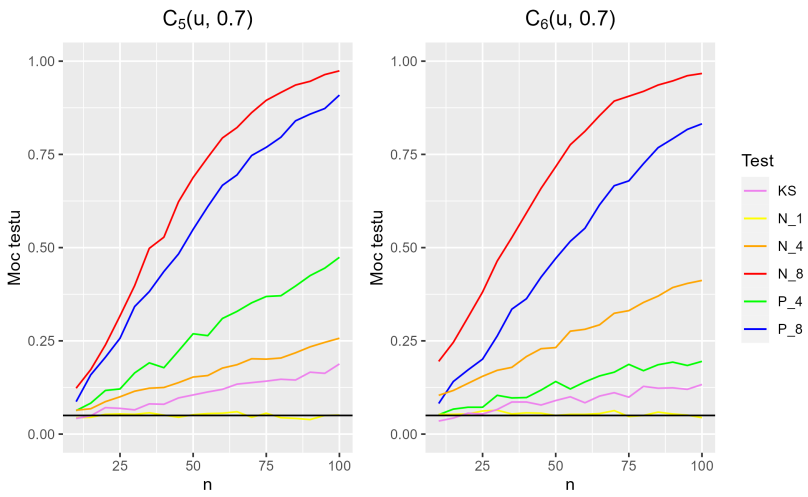
```
C_i <- function(u){1 + 0.5*cos(2*pi*u)}
C_ii <- function(u){1 + 0.5*cos(3*pi*u)}
C_iii <- function(u){1 + 0.6*cos(4*pi*u)}
C_iv <- function(u){1 + 0.7*cos(5*pi*u)}
C_v <- function(u){1 + 0.7*cos(6*pi*u)}

ex_4i <- function(n){elimination_method(n, C_i, 1.5)}
ex_4ii <- function(n){elimination_method(n, C_ii, 1.5)}
ex_4iii <- function(n){elimination_method(n, C_iii, 1.6)}
ex_4iv <- function(n){elimination_method(n, C_iv, 1.7)}
ex_4v <- function(n){elimination_method(n, C_v, 1.7)}
```



Analiza funkcji gęstości $C_j(u, \rho)$ pokazuje, że żadna z tych funkcji nie może zostać uznana za funkcję stałą (gęstość rozkładu jednostajnego). Zauważmy jednak, że dla duzych wartości obu parametrów funkcje gęstości kilkukrotnie oscylują wokół wartości ekstremalnych. Z tego powodu zbyt mało gęsty podział odcinka (0, 1) może spowodować stracenie informacji o istocie tych rozkładów.





Analiza powyższych wykresów pokazuje, że:

- Test N_1, który okazał się testem o największej mocy dla $C_1(u, 0.4)$ w pozostałych przypadkach osiąga poziom zbliżony do poziomu istotności, a zatem nie odrzuca hipotezy mówiącej, że próba pochodzi z rozkładu $U(0, 1)$ (a tak nie jest). Zatem test N_1 nie daje wiarygodnych wyników i nie powinniśmy go stosować. Nie jest to zaskakujące, gdyż wzięcie tylko jednego wielomianu z szeregu nieskończonego nie może gwarantować dokładności wykonania testu.
- Z kolei moc testu N_4 jest stosukowo wysoka w pierwszych czterech przypadkach. Dla gęstości $C_5(u, 0.7)$ i $C_6(u, 0.7)$ moc testu jest wyraźnie mniejsza. Ostatnie dwa przypadki obejmują duże wartości zarówno parametru ρ odpowiadającego za rozrzut wartości funkcji gęstosci $(C_j(u, \rho) \in [1 - \rho, 1 + \rho])$, jak i parametru j odpowiadającego za zagęszczenie funkcji *cosinus*, więc także zagęszczenie $C_j(u, \rho)$. W takich przypadkach test N_4 może nakładać oba efekty na raz, nie rozróżniając wówczas rozkładu od rozkładu jednostajnego.
- Test N_8 przeciwnie do poprzednich testów Neymana wykazuje wzrost wartości mocy wraz ze wzrostem obu parametrów ρ oraz j. Test ten zbudowany na bazie 8 wielomianów Legendre'a okazuje się bardziej czuły na zmiany rozkłady - dokładniej wykrywa odstępstwa od rozkładu jednostajnego. Test ten osiąga najWiększą moc spośród badanych testów.
- Podobna sytuacja ma miejsce w przypadku testów chi-kwadrat Pearsona. Moc testu P_4 spada wraz ze wzrostem parametrów ρ oraz j, natomiast moc testu P_8 wzrasta. Analogicznie do testów Neymana test P_4 oparty jest na zbyt mało dokładnej metodzie (podział przedziału (0,1) na zaledwie 4 części), by jego wyniki w wiarygodny sposób potwierdzały lub nie jednostajność rozkładu. Dużo bardziej wiaryogodne wyniki daje dokładniejszy test P_8.
- Test Kołmogorowa-Smirnowa wykazuje spadek mocy wraz ze wzrostem obu parametrów. Test ten okazuje się również nie być optymalnym/wiarygodnym testem.

Podsumowując, jedynym testem, który nie zauważa różnic pomiedzy rozkładami (poza jednym przypadkiem) okazał się test N_1, który z tego powodu nie powinien być stosowany.

Wszystkie pozostałe testy stwierdzają odstępstwa od rozkładu jednostajnego (z różną mocą), zatem możemy z nich korzytać.

Największą moc w przypadku zmiany obu parametrów rozkładu o gęstości $C_j(u, \rho)$ mają testy P_8 oraz N_8. Z tego powodu to właśnie te testy uznajemy za najdokładniejsze i dające najbardziej wiarygodne wyniki. Nie jest to zaskakujące, gdyż konstrukcje tych testów są najbardziej skomplikowane, przez co testy te są w stanie wykryć drobne odstępstwa od jednostajności rozkładu.