

Projekt - moce testów normalności rozkładu

Karolina Grzanka

Wprowadzenie

Tematem projektu jest zbadanie mocy testów normalności rozkładu w przypadku, gdy dane pochodzą z rozkładu t-Studenta.

Celem projektu jest przygotowanie grafik ilustrujących zmianę mocy testów w zależności od:

- długości próby
- ilości stopni swobody w rozkładzie t
- poziomu istotności

```
#import niezbędnych bibliotek
library(dplyr)
library(tidyrr)
library(ggplot2)
library(tseries)
```

Testy

W projekcie będą analizowane 3 testy normalności rozkładu:

1. Test Shapiro-Wilka: ***shapiro.test(x)*** Jest to jeden z najpopularniejszych testów wykorzystywanych do badania normalności danych.
Hipoteza zerowa i alternatywna w teście Shapiro-Wilka ma następującą postać:
H0: Próba pochodzi z populacji o rozkładzie normalnym
H1: Próba nie pochodzi z populacji o rozkładzie normalnym.
2. Test Kołmogorowa-Smirnowa: ***ks.test(x, y)*** Jest to jeden z nieparametrycznych testów zgodności dla rozkładów ciągłych. Pozwala on zbadać zgodność rozkładu cechy 'x' z wybranym rozkładem teoretycznym 'y' (w tym projekcie będzie to rozkład normalny - *pnorm*).
Hipoteza zerowa i alternatywna w teście Kołmogorowa-Smirnowa ma następującą postać:
H0: Porównywane próby mają taki sam rozkład
H1: Porównywane próby nie mają takiego samego rozkładu
3. Test Jarque-Bera: ***jarque.bera.test(x)*** Jest to test, który testuje wartość normalności przy użyciu statystyki Jarque-Bera. Test ten wykorzystuje wartości trzeciego i czwartego momentu centralnego.
Hipoteza zerowa i alternatywna w teście Jarque-Bera ma następującą postać:
H0: Próba pochodzi z rozkładu normalnego
H1: Próba nie pochodzi z rozkładu normalnego

Dane

Do projektu użyte zostały 3 wektory:

- **lengths** -> wektor długości próby (będzie to zakres od 10 do 40 zwiększający się o 3)
- **dfs** -> wektor ilości stopni swobody (będzie to zakres od 5 do 25 zwiększający się o 2)
- **alphas** -> wektor poziomu istotności testu (będą to typowe wartości dla analizy 0.02 0.05, 0.1, 0.2)

- oraz wartość **N** oznaczającą ilość symulacji

```
lengths <- seq(10, 40, by = 3)
dfs <- seq(5, 25, by = 2)
alphas <- c(.02, .05, .1, .2)
N <- 1000
```

Cele oraz Hipotezy

W projekcie poprzez analizę będą ustalane poprawności następujących hipotez:

- **W zależności od zmiany długości próby:**
H0: wraz ze zwiększaniem długości próby rośnie moc testu
H1: wraz ze zwiększaniem długości próby maleje moc testu
- **W zależności od zmiany ilości stopni swobody:**
H0: wraz ze zwiększaniem ilości stopni swobody maleje moc testu
H1: wraz ze zwiększaniem ilości stopni swobody rośnie moc testu
- **W zależności od zmiany poziomu istotności:**
H0: wraz ze zwiększaniem poziomu istotności rośnie moc testu
H1: wraz ze zwiększaniem poziomu istotności maleje moc testu

(We wszystkich analizach poprawnym wynikiem powinno być brak podstaw do odrzucenia hipotez zerowych.)

Ramka danych

Za pomocą funkcji **expand.grid** tworzona jest ramka danych wszystkich kombinacji o nazwie 'params'.

```
params <- expand.grid(length = lengths, df = dfs, alpha = alphas)
```

Test Shapiro-Wilka

Dla wszystkich kombinacji (z pomocą odpowiedniej funkcji **shapiro.test**) wyliczona zostaje moc testu Shapiro-Wilka. Następnie z pomocą funkcji **sapply** tworzony jest wektor **powers_sw**

```
powers_sw <- sapply(1:nrow(params),
  function(i){
    length <- params[i,1]
    df <- params[i, 2]
    alpha <- params[i, 3]
    p_vector <- sapply(rep(df, N),
      function(x){
        my_sample <- rt(length, df)
        shapiro.test(my_sample)$p.value
      })
    mean(p_vector < alpha)
  })
```

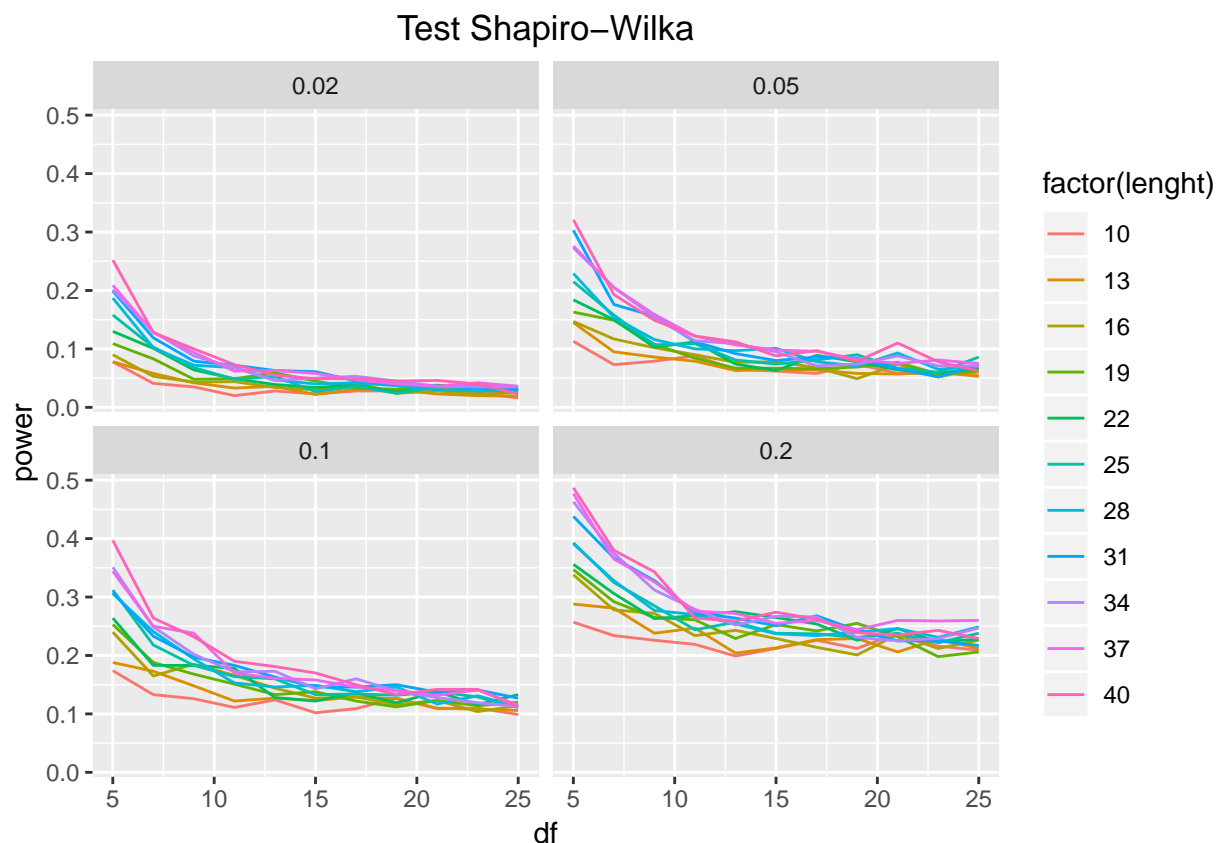
Tworzenie ramki danych dla testu Shapiro-Wilka

```
shapiro <- bind_cols(params, power = powers_sw)
```

Wizualizacja badania Testu Shapiro-Wilka

Wykresy przedstawiają wyniki analizy testu Shapiro-Wilka. Na ilustracji możemy zobaczyć 4 wykresy, każdy przedstawia wyniki dla innej wartości istotności testu. Na osi X znajdują się stopnie swobody, natomiast na osi Y moc testów. W zależności od wielkości próby na wykresie widnieje inny kolor.

```
ggplot(shapiro, aes(x = df, y = power, color = factor(lenght))) +
  labs(title = "Test Shapiro-Wilka") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line() +
  facet_wrap(~ alpha, nrow = 2)
```



Test Kołmogorowa-Smirnowa

Dla wszystkich kombinacji (z pomocą odpowiedniej funkcji **ks.test**) wyliczona zostaje moc testu Kołmogorowa-Smirnowa. Następnie z pomocą funkcji **sapply** tworzony jest wektor **powers_ks**

```
powers_ks <- sapply(1:nrow(params),
  function(i){
    length <- params[i,1]
    df <- params[i, 2]
```

```

alpha <- params[i, 3]
p_vector <- sapply(rep(df, N),
  function(x){
    my_sample <- rt(length, df)
    ks.test(my_sample, pnorm)$p.value
  })
mean(p_vector < alpha)
})

```

Tworzenie ramki danych dla testu Kołmogorowa-Smirnowa

```
kolmogorow <- bind_cols(params, power = powers_ks)
```

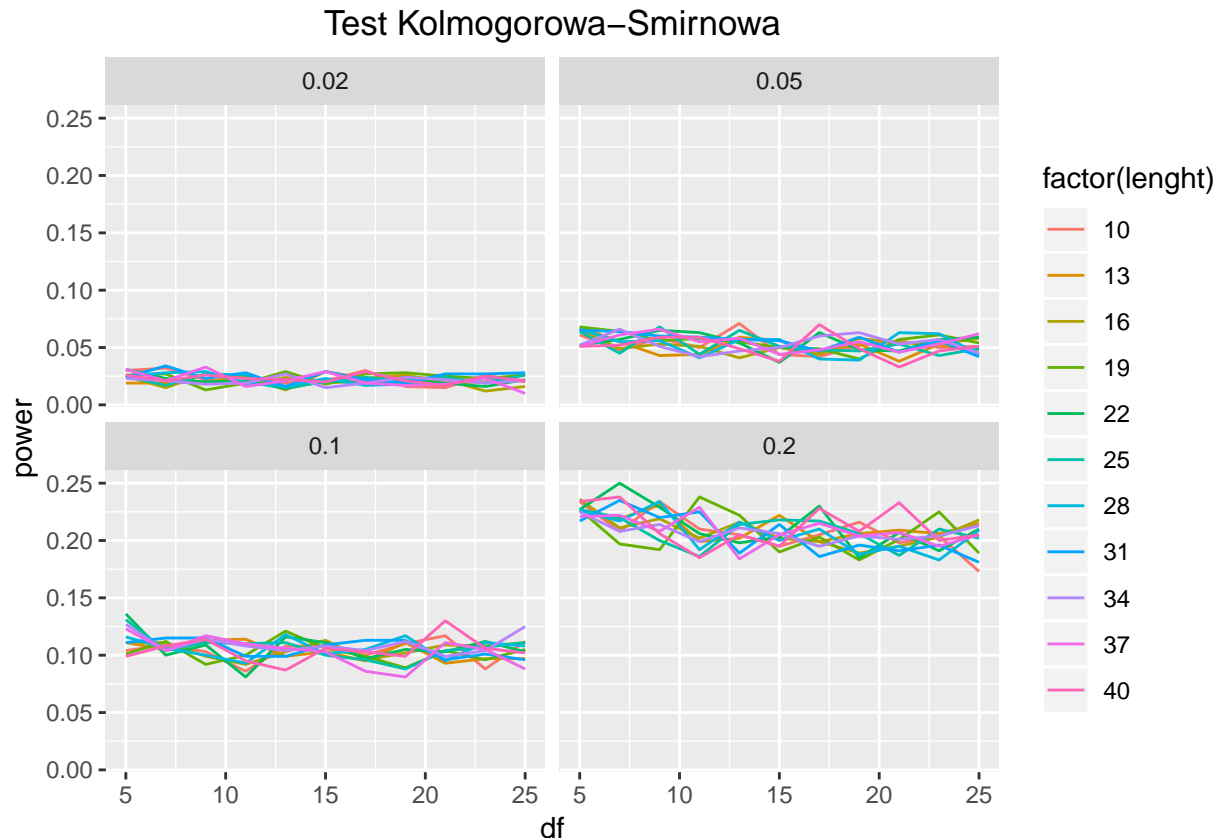
Wizualizacja badania Testu Kołmogorowa-Smirnowa

Wykresy przedstawiają wyniki analizy testu Kołmogorowa-Smirnowa. Na ilustracji możemy zobaczyć 4 wykresy, każdy przedstawia wyniki dla innej wartości poziomu istotności testu. Na osi X znajdują się stopnie swobody, natomiast na osi Y moc testów. W zależności od wielkości próby na wykresie widnieje inny kolor.

```

ggplot(kolmogorow, aes(x = df, y = power, color = factor(lenght))) +
  labs(title = "Test Kolmogorowa-Smirnowa") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line() +
  facet_wrap(~ alpha, nrow = 2)

```



Test Jarque-Bera

Dla wszystkich kombinacji (z pomocą odpowiedniej funkcji `jarque.bera.test`) wyliczona zostaje moc testu Kołmogorowa-Smirnowa. Następnie z pomocą funkcji `sapply` tworzony jest wektor `powers_jb`

```
powers_jb <- sapply(1:nrow(params),
  function(i){
    length <- params[i,1]
    df <- params[i, 2]
    alpha <- params[i, 3]
    p_vector <- sapply(rep(df, N),
      function(x){
        my_sample <- rt(length, df)
        jarque.bera.test(my_sample)$p.value
      })
    mean(p_vector < alpha)
  })
```

Tworzenie ramki danych dla testu Jarque-Bera

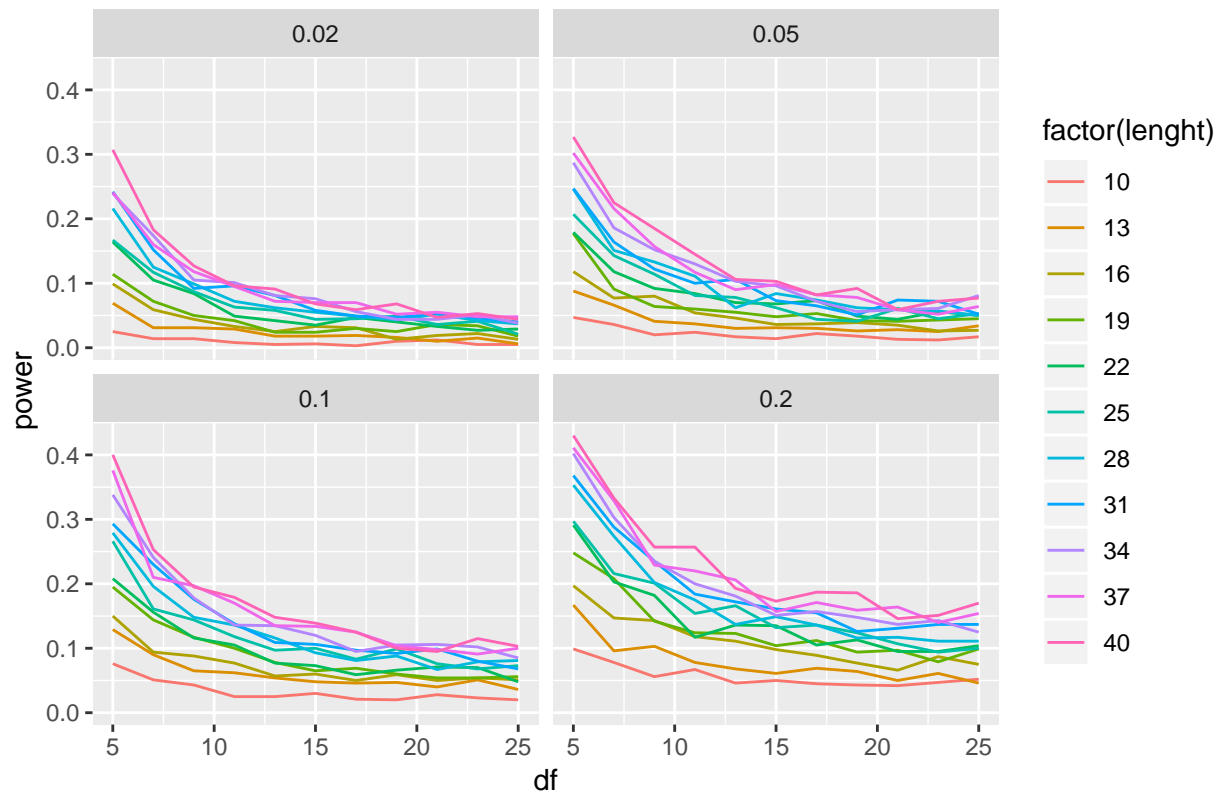
```
jarque <- bind_cols(params, power = powers_jb)
```

Wizualizacja badania Testu Jarque-Bera

Wykresy przedstawiają wyniki analizy testu Jarque-Bera. Na ilustracji możemy zobaczyć 4 wykresy, każdy przedstawia wyniki dla innej wartości istotności testu. Na osi X znajdują się stopnie swobody, natomiast na osi Y moc testów. W zależności od wielkości próby na wykresie widnieje inny kolor.

```
ggplot(jarque, aes(x = df, y = power, color = factor(length))) +
  labs(title = "Test Jarque-Bera") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_line() +
  facet_wrap(~ alpha, nrow = 2)
```

Test Jarque–Bera



Wizualizacja porównawcza

Wykres porównawczy mocy testów dla $\alpha = 5\%$ i $\text{length} = 16$. Na osi X znajdują się stopnie swobody, natomiast na osi Y moc testów.

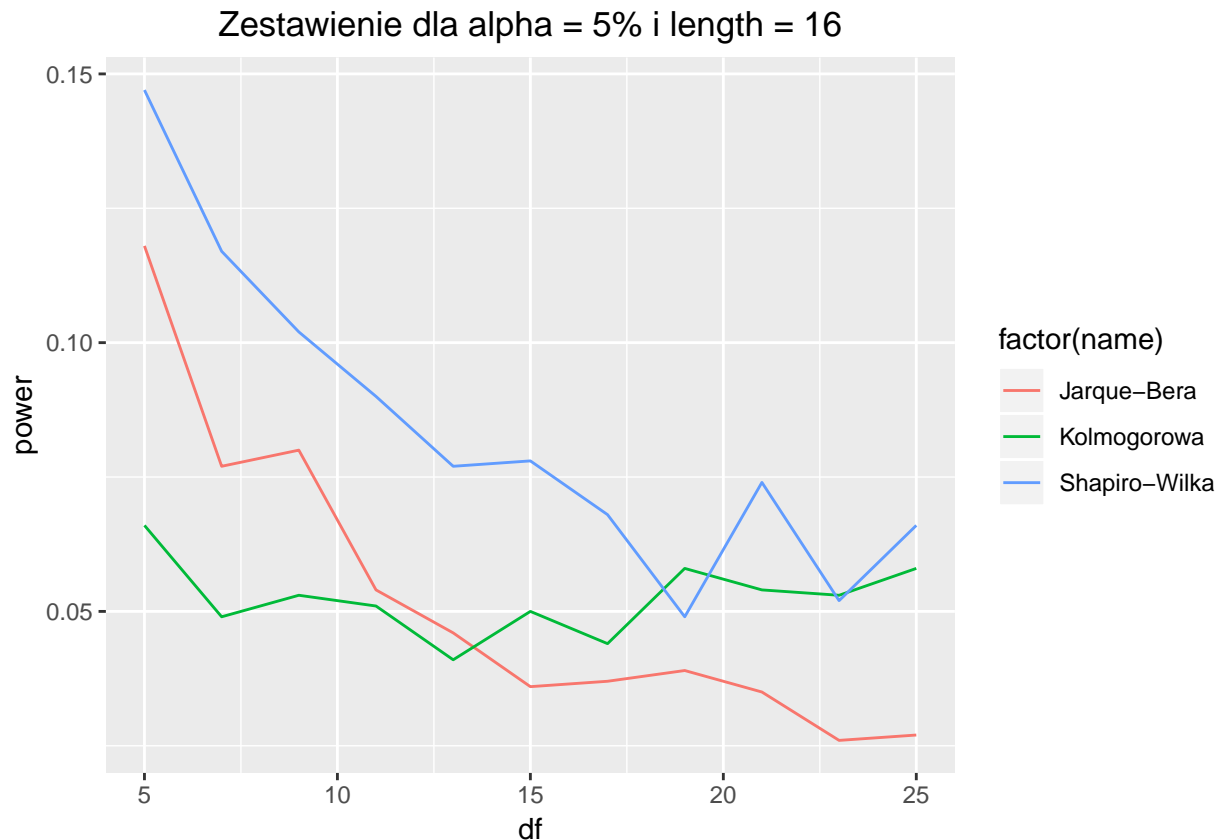
```
shapiro_wide <- spread(shapiro, lenght, power)
kolmogorow_wide <- spread(kolmogorow, lenght, power)
jarque_wide <- spread(jarque, lenght, power)

shapiro_wide <- shapiro_wide[which(shapiro_wide[,2] == .05), c(1,2,5)]
shapiro_wide$name <- "Shapiro-Wilka"
colnames(shapiro_wide)[3] = "power"
kolmogorow_wide <- kolmogorow_wide[which(kolmogorow_wide[,2] == .05), c(1,2,5)]
kolmogorow_wide$name <- "Kolmogorowa"
colnames(kolmogorow_wide)[3] = "power"
jarque_wide <- jarque_wide[which(jarque_wide[,2] == .05), c(1,2,5)]
jarque_wide$name <- "Jarque-Bera"
colnames(jarque_wide)[3] = "power"

all_wide <- data.frame(shapiro_wide)
all_wide <- rbind(all_wide, kolmogorow_wide, jarque_wide)

ggplot(all_wide, aes(x = df, y = power, color = factor(name))) +
  labs(title = "Zestawienie dla alpha = 5% i length = 16") +
  theme(plot.title = element_text(hjust = 0.5)) +
```

```
geom_line()
```



Wykres porównawczy mocy testów dla $\alpha = 5\%$ i $df = 13$. Na osi X znajdują się stopnie swobody, natomiast na osi Y moc testów.

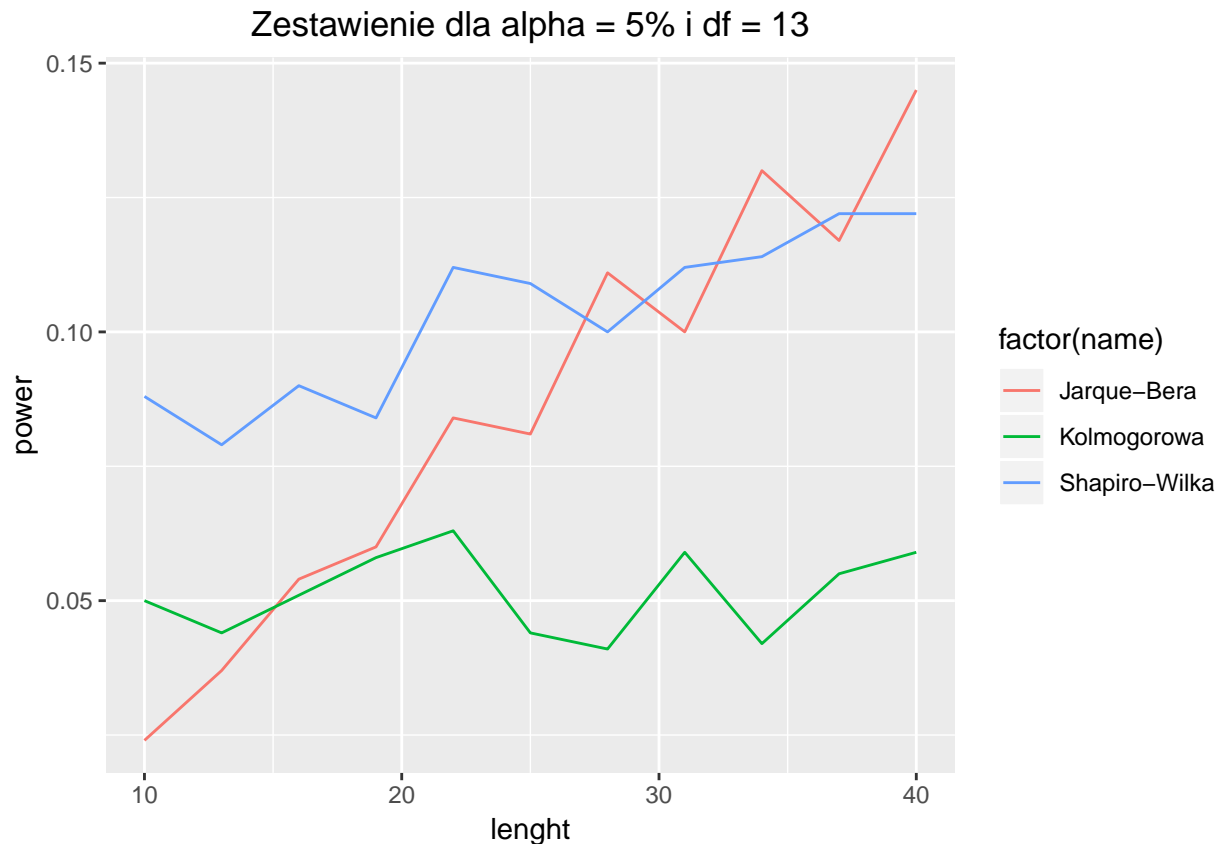
```
shapiro_wide <- spread(shapiro, df, power)
kolmogorow_wide <- spread(kolmogorow, df, power)
jarque_wide <- spread(jarque, df, power)

shapiro_wide <- shapiro_wide[which(shapiro_wide[,2] == .05), c(1,2,6)]
shapiro_wide$name <- "Shapiro-Wilka"
colnames(shapiro_wide)[3] = "power"
kolmogorow_wide <- kolmogorow_wide[which(kolmogorow_wide[,2] == .05), c(1,2,6)]
kolmogorow_wide$name <- "Kolmogorowa"
colnames(kolmogorow_wide)[3] = "power"
jarque_wide <- jarque_wide[which(jarque_wide[,2] == .05), c(1,2,6)]
jarque_wide$name <- "Jarque-Bera"
colnames(jarque_wide)[3] = "power"

all_wide <- data.frame(shapiro_wide)
all_wide <- rbind(all_wide, kolmogorow_wide, jarque_wide)

ggplot(all_wide, aes(x = length, y = power, color = factor(name))) +
  labs(title = "Zestawienie dla alpha = 5% i df = 13") +
  theme(plot.title = element_text(hjust = 0.5)) +
```

```
geom_line()
```



Wykres porównawczy mocy testów dla length = 16 i df = 13. Na osi X znajdują się stopnie swobody, natomiast na osi Y moc testów.

```
shapiro_wide <- spread(shapiro, df, power)
kolmogorow_wide <- spread(kolmogorow, df, power)
jarque_wide <- spread(jarque, df, power)

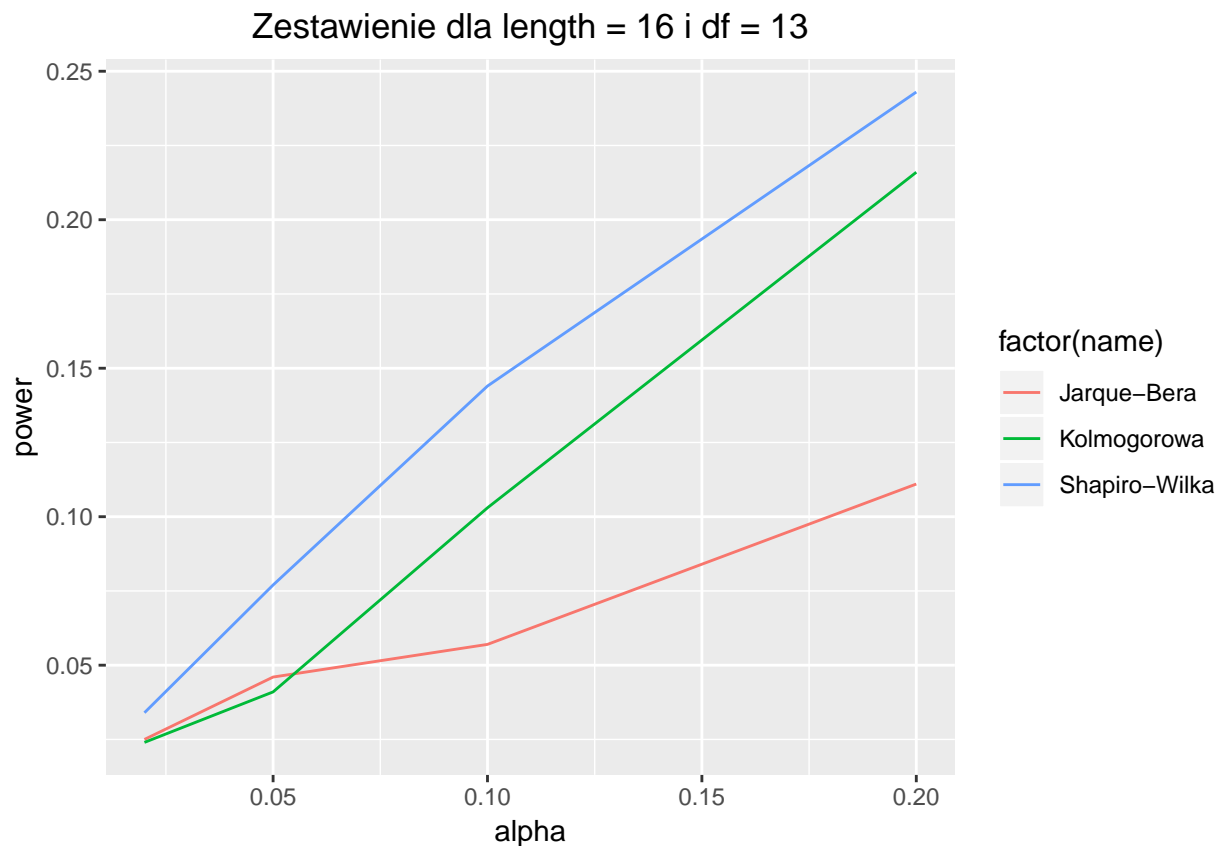
shapiro_wide <- shapiro_wide[which(shapiro_wide[,1] == 16), c(1,2,7)]
shapiro_wide$name <- "Shapiro-Wilka"
colnames(shapiro_wide)[3] = "power"
kolmogorow_wide <- kolmogorow_wide[which(kolmogorow_wide[,1] == 16), c(1,2,7)]
kolmogorow_wide$name <- "Kolmogorowa"
colnames(kolmogorow_wide)[3] = "power"
jarque_wide <- jarque_wide[which(jarque_wide[,1] == 16), c(1,2,7)]
jarque_wide$name <- "Jarque-Bera"
colnames(jarque_wide)[3] = "power"

all_wide <- data.frame(shapiro_wide)
all_wide <- rbind(all_wide, kolmogorow_wide, jarque_wide)

ggplot(all_wide, aes(x = alpha, y = power, color = factor(name))) +
  labs(title = "Zestawienie dla length = 16 i df = 13") +
  theme(plot.title = element_text(hjust = 0.5)) +
```



```
geom_line()
```



Wnioski

Po przeanalizowaniu wszystkich badań oraz wyników, można stwierdzić, iż we wszystkich postawionych na początku hipotezach - nie ma podstaw do odrzucenia hipotezy zerowych. Oznacza to, że:

1. Wielkość próby (length) ma wpływ na moc testu oraz wzrost ze wzrostem tej wartości rośnie moc testu.
2. Ilość stopni swobody (df) ma wpływ na moc testu oraz wzrost ze wzrostem tej wartości maleje moc testu.
3. Poziom istotności ma wpływ na moc testu oraz wraz ze zwiększaniem poziomu istotności rośnie moc testu.

Porównanie Testów

Z powyższych wyników oraz wizualizacji wynika, iż:

- W teście Shapiro-Wilka widać zmianę mocy testu w zależności na długość próby. Widać również jak moc testu zdecydowanie maleje w zależności od ilości stopni swobody. Porównując wszystkie 4 wykresy możemy zaobserwować iż poziom istotności ma znaczny wpływ na moc testu.
- W teście Kołmogorowa-Smirnowa widać iż wszystkie linie pokazujące równe wielkości próby są bardzo ponakładane na siebie, oznacza to, że ten czynnik znacznie mniej wpływa na moc testu. Również analizując wykresy pod względem zmiany stopni swobody widać, że wykres nie zmienia się dynamicznie. Porównując wszystkie 4 wykresy dokładnie widać jakie znaczenie ma poziom istotności.

- W teście Jarque-Bera widać wyraźnie wpływ długości próby na moc testu oraz wpływ ilości stopni swobody. Natomiast mniejszy wpływ na zmianę poziomu istotności (nadal jest, jednak mniejszy niż w powyższych testach).
- Na ostatnich wykresach widać, iż test Shapiro-Wilka ma największą moc spośród badanych testów co pokrywa się z wynikami analizy porównawczej przy użyciu metod Monte-Carlo.

Na analizę mogły mieć takie czynniki jak:

- Mała długość wektora długość (length)
- Mała długość wektora stopni swobody (df)
- Uproszczone porównanie ze sobą testów