

Sprawozdanie 2

Yana Negulescu

Spis treści:

1. Wstęp
2. Opis testów:
 - Test t-Studenta przy założeniu $\sigma_1 = \sigma_2$
 - Testu t-Welcha
 - Testu sumy rang Wilcoxona
3. Próba wyboru testu jednostajnie najmocniejszego spośród wybranych:
 - Zadanie 1
 - Zadanie 2
 - Zadanie 3
4. Próba wyboru testu jednostajnie najmocniejszego:
 - Test t-Studenta
 - Testu t-Welcha
 - Testu sumy rang Wilcoxona
5. Wnioski

1. Wstęp

Celem niniejszego raportu jest analiza i weryfikacja hipotez dotyczących równości średnich dwóch populacji, przy zastosowaniu różnych testów statystycznych na poziomie istotności $\alpha = 0.05$. Hipotezy badane w tym kontekście to hipoteza zerowa $H_0 : \mu_1 = \mu_2$, przeciwstawiona hipotezie alternatywnej $H_1 : \mu_1 \neq \mu_2$. W raporcie przyjęto założenie o równości lub różności wariancji w populacjach, co doprowadziło do zastosowania testu t-Studenta, testu t-Welcha oraz testu sumy rang Wilcoxona.

Idea testowania opiera się na ocenie, czy obserwowane różnice między grupami są statystycznie znaczące, czy mogą być wynikiem przypadkowej zmienności. Kluczowym aspektem analizy jest funkcja mocy testu, która informuje o prawdopodobieństwie odrzucenia hipotezy zerowej, gdy jest ona fałszywa. Funkcja mocy pozwala ocenić skuteczność testu w identyfikowaniu istotnych różnic.

Test jednostajnie najmocniejszy to taki, który dla danego poziomu istotności α maksymalizuje moc testu dla wszystkich możliwych wartości parametru określającego hipotezę alternatywną, zapewniając najwyższą skuteczność w odrzucaniu fałszywej hipotezy zerowej.

Metoda Monte Carlo, wykorzystana do symulacji w niniejszym raporcie, opiera się na generowaniu dużych próbek danych zgodnie z określonymi rozkładami prawdopodobieństwa, aby numerycznie oszacować właściwości statystyczne testów, w tym funkcję mocy. Jest to narzędzie umożliwiające zrozumienie zachowania testów statystycznych w różnych scenariuszach, co jest kluczowe dla właściwej interpretacji wyników analizy statystycznej.

2. Opis poszczególnych testów

2.1 Test t-Studenta przy założeniu $\sigma_1 = \sigma_2$

Opis testu:

Test t-Studenta dla niezależnych prób stosuje się do porównania średnich z dwóch grup. Założeniem jest równość wariancji w obu grupach.

- **Statystyka testowa:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

gdzie $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ to skorygowana wariancja, s_1^2, s_2^2 to wariancje próbek, a n_1, n_2 to rozmiary próbek.

- **Rozkład statystyki testowej przy H_0 :**

Rozkład t-Studenta z $n_1 + n_2 - 2$ stopniami swobody.

- **Obszar odrzucenia:**

Dla testu dwustronnego, odrzucamy H_0 jeśli $|t| > t_{\alpha/2, n_1+n_2-2}$, gdzie $t_{\alpha/2, n_1+n_2-2}$ jest wartością krytyczną rozkładu t-Studenta.

Przykładowe wywołanie testu:

```

# Generowanie danych symulacyjnych
set.seed(123) # Dla reprodukowalności
n1 <- 100
n2 <- 200
mu1 <- 1
mu2 <- 2
sigma <- 2

data1 <- rnorm(n1, mean = mu1, sd = sigma)
data2 <- rnorm(n2, mean = mu2, sd = sigma)

# Wykonanie testu t-Studenta
test_result <- t.test(data1, data2, var.equal = TRUE)

# Wypisanie wyników
test_result

```

Two Sample t-test

```

data: data1 and data2
t = -3.589, df = 298, p-value = 0.000388
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.2883817 -0.3758313
sample estimates:
mean of x mean of y
 1.180812  2.012918

```

Na podstawie przedstawionych wyników testu t-Studenta dla dwóch prób:

Statystyka testowa: $t = -3.589$

Stopnie swobody: $df = 289$

p-wartość: $p = 0.000388$

Przyjmując poziom istotności $\alpha = 0.05$, p-wartość jest znacznie mniejsza niż α , co oznacza, że odrzucamy hipotezę zerową. Wynik wskazuje na istotną statystycznie różnicę między średnimi obu grup. Przedział ufności $[-1.2883817 - 0.3758313]$ dla różnicy średnich nie obejmuje zera, co dodatkowo potwierdza znaczącą różnicę.

Oszacowanie funkcji mocy:

```
library(ggplot2)

mu_diffs <- seq(-2, 2, by = 0.1)
alpha <- 0.05
n_simulations <- 1000
powers <- numeric(length(mu_diffs))

# Symulacja
for (i in seq_along(mu_diffs)) {
  mu2 <- mu1 + mu_diffs[i]
  rejections <- 0

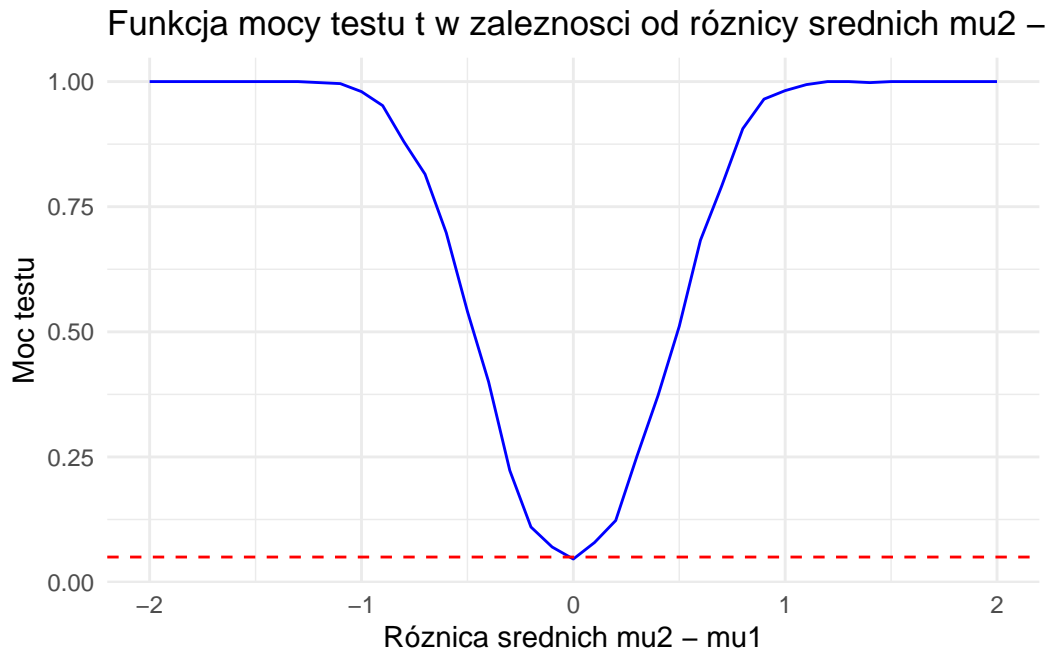
  for (j in 1:n_simulations) {
    sample1 <- rnorm(n1, mu1, sigma)
    sample2 <- rnorm(n2, mu2, sigma)
    t_test <- t.test(sample1, sample2, var.equal = TRUE)

    if (t_test$p.value < alpha) {
      rejections <- rejections + 1
    }
  }

  powers[i] <- rejections / n_simulations
}

# Tworzenie danych dla wykresu
data_for_plot <- data.frame(mu_diff = mu_diffs, power = powers)

# Tworzenie wykresu
ggplot(data_for_plot, aes(x = mu_diff, y = power)) +
  geom_line(color = "blue") +
  geom_hline(yintercept = alpha, linetype="dashed", color = "red") +
  labs(title = "Funkcja mocy testu t w zależności od różnicy średnich mu2 - mu1",
       x = "Różnica średnich mu2 - mu1",
       y = "Moc testu") +
  theme_minimal()
```



2.2 Test t-Welcha

Opis testu:

Test t-Welcha jest modyfikacją testu t-Studenta, który nie zakłada równości wariancji próbek.

- **Statystyka testowa:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Rozkład statystyki testowej przy H_0 :** Rozkład t-Studenta, lecz stopnie swobody są obliczane za pomocą wzoru Welcha-Satterthwaite'a:

$$v \approx \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1})^2}{n_2 - 1}$$

- **Obszar odrzucenia:** Analogicznie do testu t-Studenta, ale z uwzględnieniem obliczonych stopni swobody.

Przykładowe wywołanie testu:

```
# Parametry prób
mu1 <- 1
mu2 <- 2
sigma <- 2
n1 <- 100
n2 <- 200

# Generowanie danych
set.seed(123) # Dla reprodukowalności wyników
data1 <- rnorm(n1, mean = mu1, sd = sigma)
data2 <- rnorm(n2, mean = mu2, sd = sigma)

# Wykonanie testu t-Welcha
test_result <- t.test(data1, data2, var.equal = FALSE)

# Wyświetlenie wyników
test_result
```

Welch Two Sample t-test

```
data: data1 and data2
t = -3.6535, df = 207.8, p-value = 0.0003275
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.2811098 -0.3831032
sample estimates:
mean of x mean of y
 1.180812  2.012918
```

Na podstawie przedstawionych wyników testu t-Studenta dla dwóch prób:

Statystyka testowa: $t = -3.6535$

Stopnie swobody: $df = 207.8$

p-wartość: $p = 0.0003275$

Przyjmując poziom istotności $\alpha = 0.05$, p-wartość jest znacznie mniejsza niż α , co oznacza, że odrzucamy hipotezę zerową. Wynik wskazuje na istotną statystycznie różnicę między średnimi obu grup. Przedział ufności $[-1.180812 - 3831032]$ dla różnicy średnich nie obejmuje zera, co dodatkowo potwierdza znaczącą różnicę.

Oszacowanie funkcji mocy:

```
mu_diffs <- seq(-2, 2, by = 0.1)
alpha <- 0.05
n_simulations <- 1000
powers <- numeric(length(mu_diffs))

for (i in seq_along(mu_diffs)) {
  mu2 <- mu1 + mu_diffs[i]
  rejections <- 0

  for (j in 1:n_simulations) {
    sample1 <- rnorm(n1, mean = mu1, sd = sigma)
    sample2 <- rnorm(n2, mean = mu2, sd = sigma)

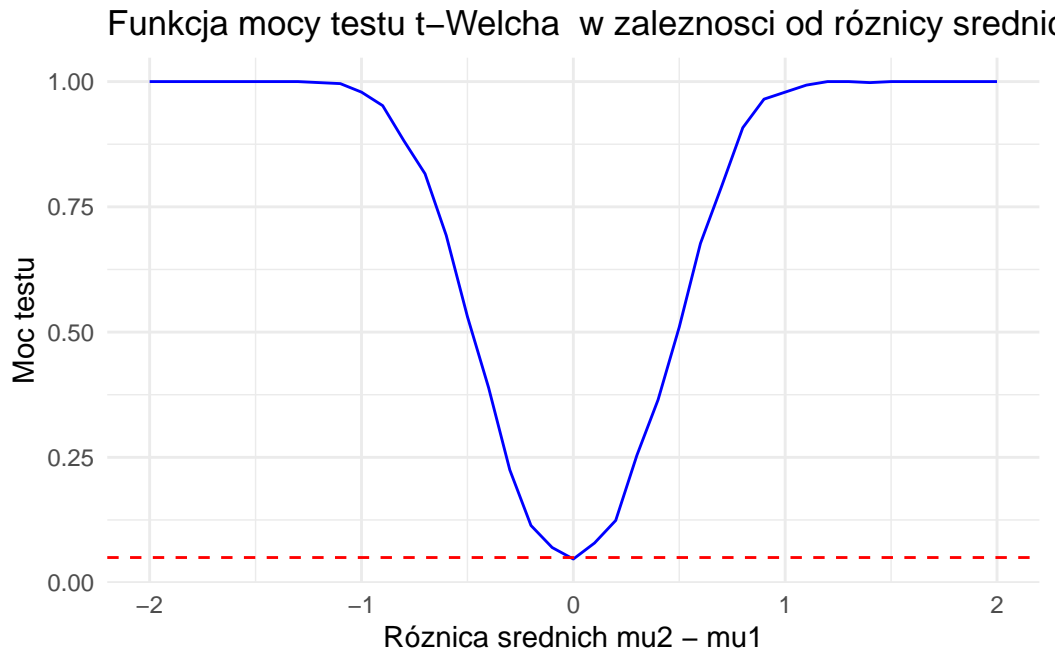
    # Test t-Welcha
    t_test <- t.test(sample1, sample2, var.equal = FALSE)

    if (t_test$p.value < alpha) {
      rejections <- rejections + 1
    }
  }

  powers[i] <- rejections / n_simulations
}

# Tworzenie danych dla wykresu
data_for_plot <- data.frame(mu_diff = mu_diffs, power = powers)

# Tworzenie wykresu
ggplot(data_for_plot, aes(x = mu_diff, y = power)) +
  geom_line(color = "blue") +
  geom_hline(yintercept = alpha, linetype="dashed", color = "red") +
  labs(title = "Funkcja mocy testu t-Welcha w zależności od różnicy średnich mu2 - mu1",
       x = "Różnica średnich mu2 - mu1",
       y = "Moc testu") +
  theme_minimal()
```



2.3 Test sumy rang Wilcoxona

Opis testu:

Test sumy rang Wilcoxona jest nieparametrycznym testem używanym do porównania dwóch niezależnych prób.

- **Statystyka testowa:** Rangi są przypisywane wszystkim obserwacjom od najmniejszej do największej, a następnie sumowane dla każdej z grup. Statystyka testowa U jest mniejsza z dwóch sum rang.

$$U = R_1 - \frac{n_1(n_1 + 1)}{2}$$

gdzie R_1 to suma rang dla pierwszej próbki, a n_1 to liczba obserwacji w tej próbce.

- **Rozkład statystyki testowej przy H_0 :** Dla dużych prób rozkład U może być przybliżony przez rozkład normalny z odpowiednią średnią i wariancją.

Średnia i wariancja statystyki U , które są używane do jej normalizacji, gdy próbki są duże, są obliczane jako:

$$\mu_U = \frac{n_1 n_2}{2} \quad \text{oraz} \quad \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

- **Obszar odrzucenia:** Dla testu dwustronnego, odrzucamy H_0 jeśli wartość U jest znacznie mniejsza lub większa od wartości oczekiwanej pod hipotezą zerową.

Przy dużych rozmiarach prób (U) może być przybliżone przez rozkład normalny, co pozwala na wykorzystanie standardowych metod statystycznych do oceny znaczenia.

Przykładowe wywołanie testu:

```
mu1 <- 1
mu2 <- 2
sigma <- 2
n1 <- 100
n2 <- 200

# Generowanie danych
set.seed(123) # Dla reprodukowalności wyników
data1 <- rnorm(n1, mean = mu1, sd = sigma)
data2 <- rnorm(n2, mean = mu2, sd = sigma)

# Wykonanie testu sumy rang Wilcoxona
test_result <- wilcox.test(data1, data2)

# Wyświetlenie wyników
test_result
```

Wilcoxon rank sum test with continuity correction

data: data1 and data2

W = 7679, p-value = 0.001052

alternative hypothesis: true location shift is not equal to 0

- Statystyka testowa W: 7679
- Wartość p: 0.001052
- Hipoteza alternatywna: prawdziwa różnica lokalizacji nie jest równa 0

Hipoteza zerowa testu Wilcoxona mówi, że mediany dwóch porównywanych grup są równe. Hipoteza alternatywna sugeruje, że istnieje różnica w lokalizacji rozkładów tych dwóch grup. p-wartość jest wyznacznikiem siły przeciwko hipotezie zerowej. Przy standardowym poziomie istotności $\alpha = 0.05$, p-wartość jest znacznie mniejsza od α , co oznacza, że odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej. Oznacza to, że istnieją statystycznie istotne dowody sugerujące różnicę w medianach pomiędzy dwiema grupami.

Oszacowanie funkcji mocy:

```
mu_diffs <- seq(-2, 2, by = 0.1)
alpha <- 0.05
n_simulations <- 1000
powers <- numeric(length(mu_diffs))

for (i in seq_along(mu_diffs)) {
  mu2 <- mu1 + mu_diffs[i]
  rejections <- 0

  for (j in 1:n_simulations) {
    sample1 <- rnorm(n1, mean = mu1, sd = sigma)
    sample2 <- rnorm(n2, mean = mu2, sd = sigma)

    # Test Wilcoxona
    wilcox_test <- wilcox.test(sample1, sample2)

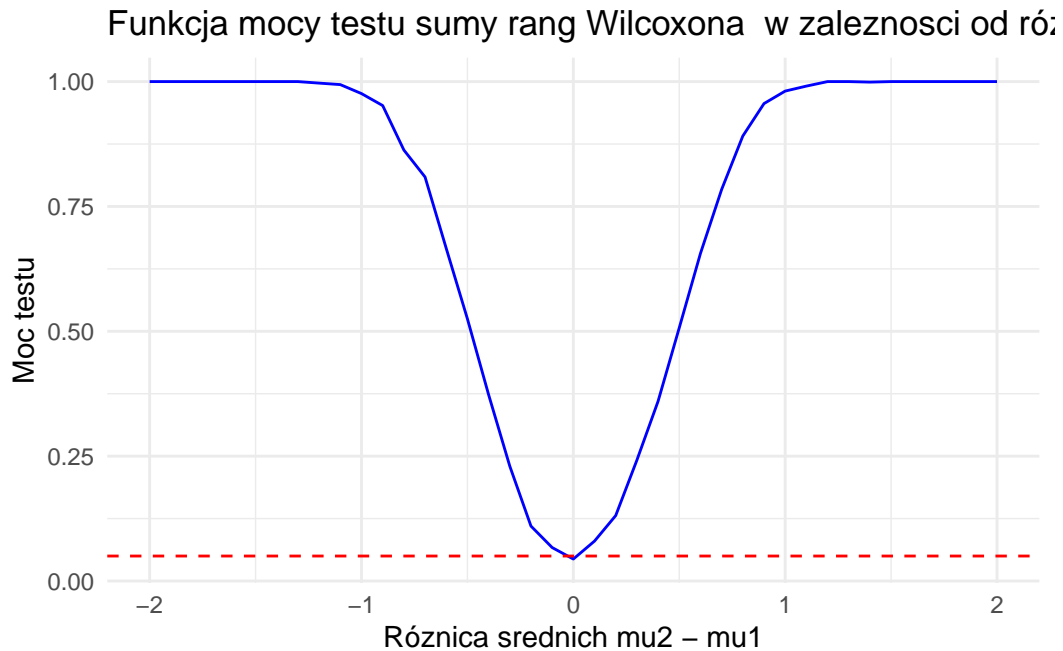
    if (wilcox_test$p.value < alpha) {
      rejections <- rejections + 1
    }
  }

  powers[i] <- rejections / n_simulations
}

# Tworzenie wykresu

data_for_plot <- data.frame(mu_diff = mu_diffs, power = powers)

# Tworzenie wykresu
ggplot(data_for_plot, aes(x = mu_diff, y = power)) +
  geom_line(color = "blue") +
  geom_hline(yintercept = alpha, linetype="dashed", color = "red") +
  labs(title = "Funkcja mocy testu sumy rang Wilcoxona w zależności od różnicy średnich m",
       x = "Różnica średnich mu2 - mu1",
       y = "Moc testu") +
  theme_minimal()
```



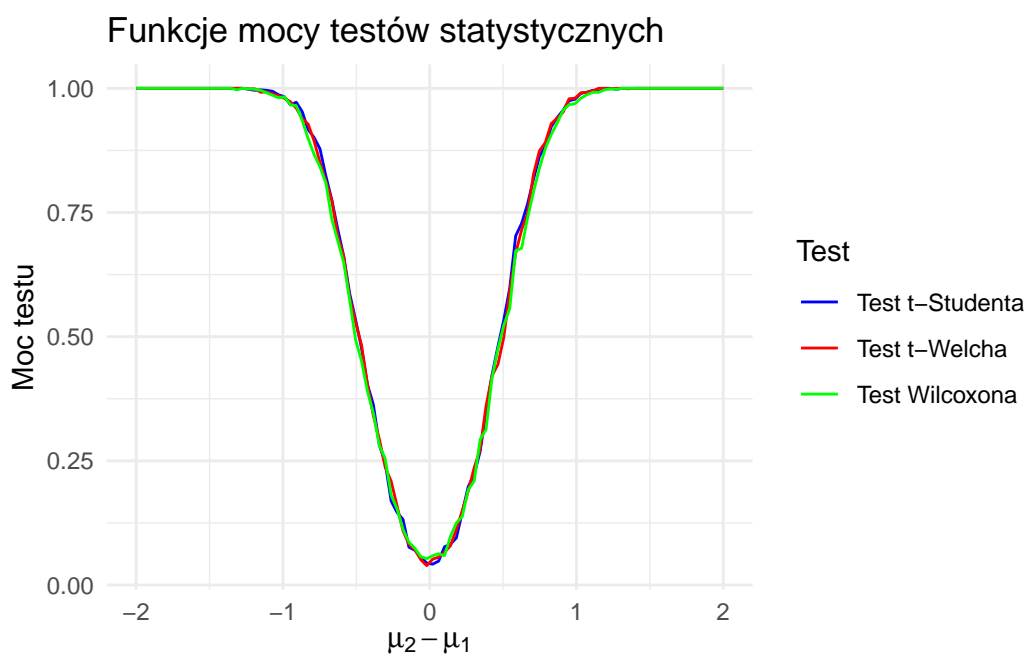
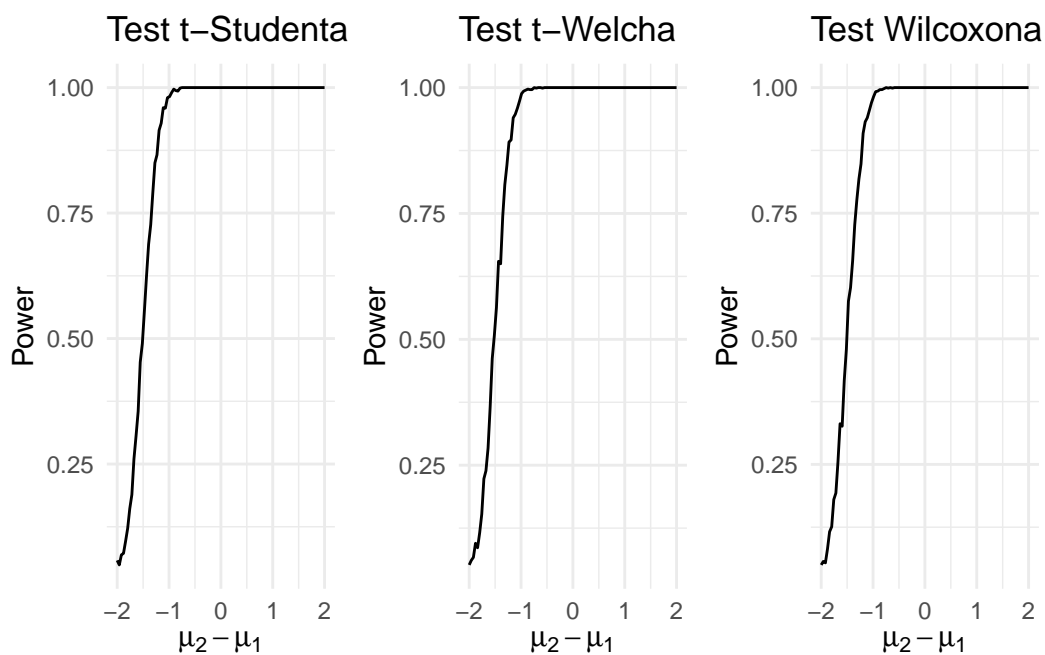
3. Próba wyboru testu jednostajnie najmocniejszego

Zadanie 1

Rozważmy próbę (X_1, \dots, X_{100}) z rozkładu normalnego $\mathcal{N} \sim (\mu_1, 2^2)$ oraz drugą próbę (Y_1, \dots, Y_{200}) z rozkładu normalnego $\mathcal{N} \sim (\mu_2, 2^2)$. Korzystając z symulacji Monte Carlo wykonamy wykres funkcji mocy w zależności od $\mu_2 - \mu_1$ na przedziale $(-2, 2)$ dla wszystkich trzech testów.

Warning: 'gridExtra' R 4.3.2

Warning: `qplot()` was deprecated in ggplot2 3.4.0.



Podobieństwo funkcji mocy: Funkcje mocy dla testu t-Studenta, testu t-Welcha i testu Wilcoxona są bardzo zbliżone na całym przedstawionym zakresie różnicy średnich. Oznacza to, że w kontekście tych symulacji wszystkie trzy testy mają podobną zdolność do wykrywania różnicy między średnimi populacji.

Brak jednoznacznie najmocniejszego testu: Na podstawie wykresu nie można wskazać testu, który byłby jednostajnie najmocniejszy (czyli mający największą moc testu we wszystkich punktach zakresu) dla każdej wartości różnicy średnich. Wszystkie trzy testy wykazują podobne właściwości w zakresie mocy, szczególnie w obszarach, gdzie różnica średnich jest bliska 0.

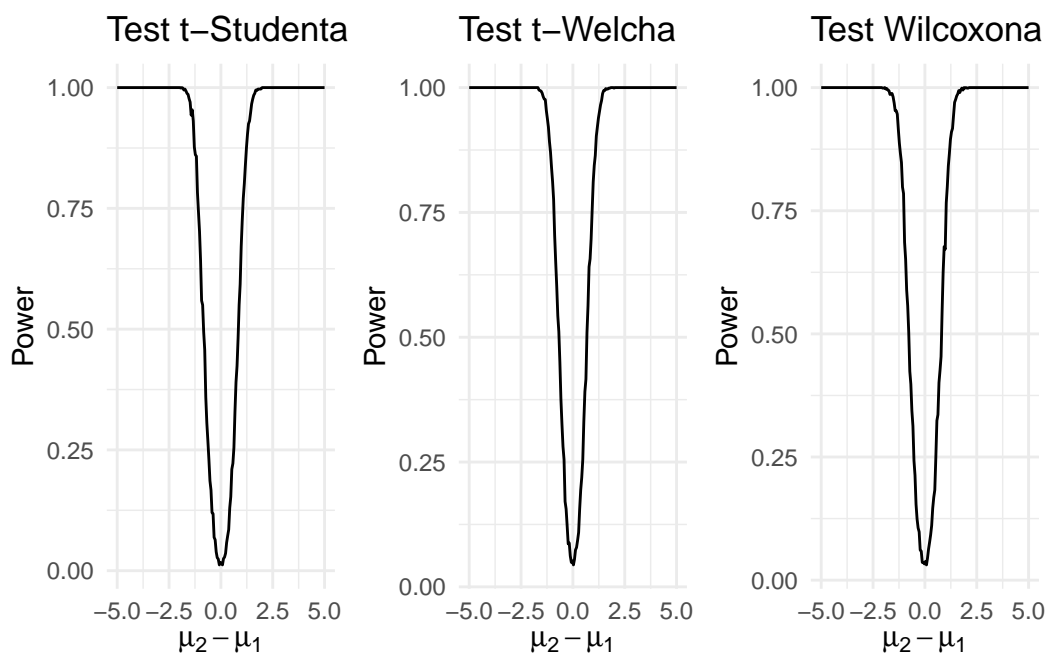
Moc testów przy braku różnicy średnich: Zauważamy, że gdy różnica średnich jest równa 0 ($\mu_2 - \mu_1 = 0$), moc wszystkich testów jest najniższa. Jest to oczekiwane, ponieważ przy braku rzeczywistej różnicy między grupami, testy powinny rzadziej doprowadzać do odrzucenia hipotezy zerowej.

Moc testów przy istotnej różnicy średnich: Gdy różnica średnich zwiększa się (zarówno w kierunku dodatnim, jak i ujemnym), moc testów wzrasta, zbliżając się do 1. Oznacza to, że testy są coraz bardziej skuteczne w wykrywaniu istotnych różnic między średnimi.

W kontekście danych symulacyjnych i przyjętych założeń, żaden z testów nie wykazuje się jako jednostajnie najmocniejszy we wszystkich rozważanych scenariuszach. Wszystkie trzy testy mają podobną moc w wykrywaniu istotnych różnic średnich, z tendencją do zwiększania mocy w miarę wzrostu bezwzględnej wartości różnicy średnich.

Zadanie 2

Rozważmy próbę (X_1, \dots, X_{100}) z rozkładu normalnego $\mathcal{N} \sim (\mu_1, 2^2)$ oraz drugą próbę (Y_1, \dots, Y_{200}) z rozkładu normalnego $\mathcal{N} \sim (\mu_2, 4^2)$. Wykonajmy wykres funkcji mocy na przedziale $(-5, 5)$.



Podobieństwo mocy testów: Wszystkie trzy testy mają bardzo podobne funkcje mocy w całym zakresie różnicy średnich. To wskazuje, że ich zdolność do odrzucenia hipotezy zerowej przy istniejącej różnicy średnich jest porównywalna.

Brak jednostajnie najmocniejszego testu: Żaden z testów nie wyróżnia się jako jednostajnie

najmocniejszy we wszystkich badanych przypadkach. Wartości mocy są bardzo bliskie dla wszystkich trzech testów, co oznacza, że wybór konkretnego testu powinien zależeć od innych czynników, takich jak założenia o danych, wielkość próbki i preferencje analityka.

Moc wokół hipotezy zerowej: Dla wartości $\mu_2 - \mu_1$ bliskich zeru, wszystkie testy mają niską moc, co jest zgodne z oczekiwaniami, ponieważ różnica średnich jest niewielka, a więc trudniejsza do wykrycia.

Moc dla znaczącej różnicy średnich: Gdy wartość bezwzględna różnicy średnich $\mu_2 - \mu_1$ wzrasta, moc wszystkich testów również wzrasta, zbliżając się do wartości 1. Oznacza to, że wszystkie testy są bardzo skuteczne w wykrywaniu większych różnic między średnimi.

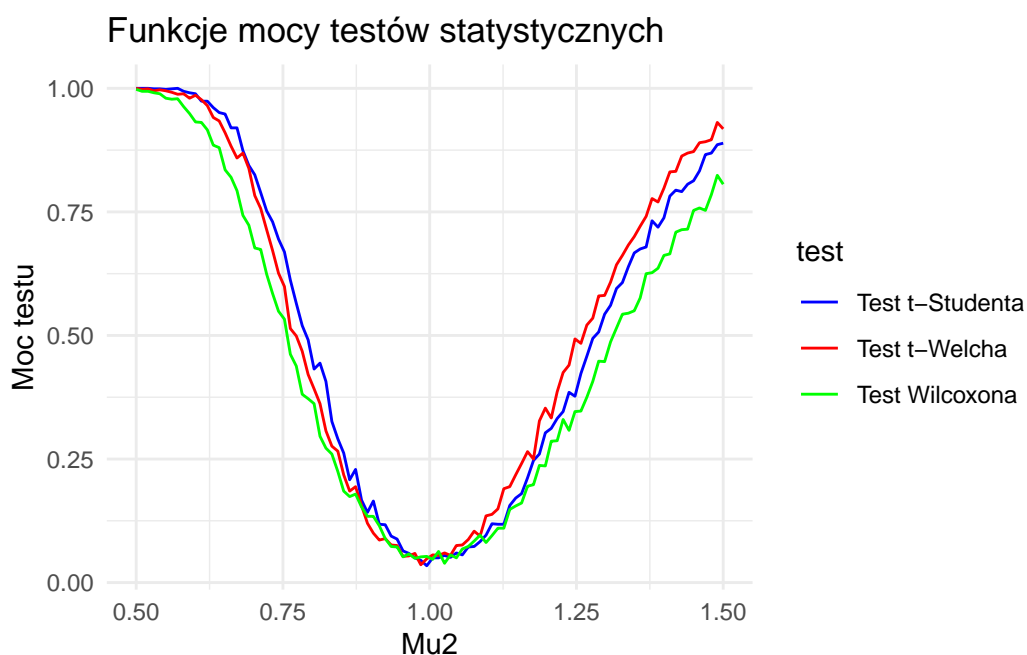
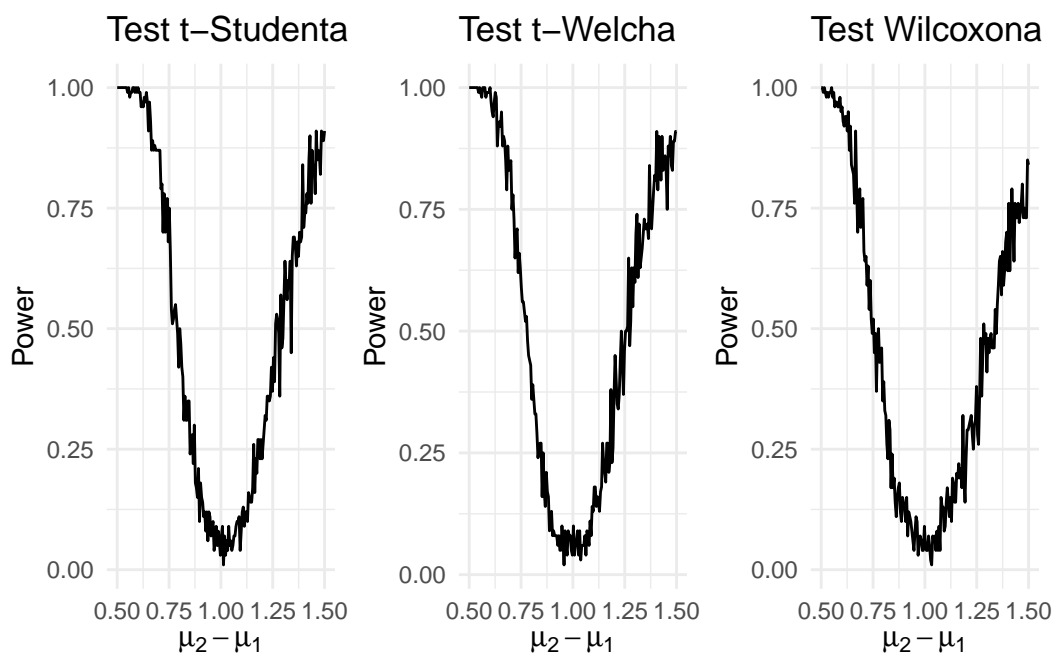
Wrażliwość testu Wilcoxona: Test Wilcoxona jest testem nieparametrycznym i jest mniej wrażliwy na założenia o normalności rozkładów, więc jego podobna moc do testów parametrycznych wskazuje, że może być dobrym wyborem w praktyce, zwłaszcza gdy nie można spełnić założeń testów parametrycznych.

Wpływ różnych wariancji: Test t-Welcha jest zaprojektowany do pracy z próbkami mającymi różne wariancje, a mimo to jego moc jest bardzo zbliżona do testu t-Studenta, który zakłada równość wariancji. To sugeruje, że w tej konkretnej symulacji różnica w wariancjach między grupami nie miała dużego wpływu na moc testu.

Podsumowując, na podstawie tego wykresu nie można jednoznacznie stwierdzić, że istnieje test jednostajnie najmocniejszy we wszystkich badanych scenariuszach. Wybór testu powinien zatem bazować na innych kryteriach, takich jak założenia dotyczące danych, równość wariancji oraz rozmiar próbki.

Zadanie 3

Rozważmy próbę (X_1, \dots, X_{100}) z rozkładu wykładniczego $\mathcal{Exp}(1/\mu_1)$ oraz drugą próbę (Y_1, \dots, Y_{200}) z rozkładu wykładniczego $\mathcal{Exp}(1/\mu_2)$. Wykonajmy wykres funkcji mocy na przedziale $(0.5, 1.5)$.



Moc testów w zależności od średniej: Wszystkie trzy testy mają najniższą moc, gdy wartości μ_2 są bliskie wartości μ_1 . To jest zgodne z oczekiwaniami, gdyż mniejsza różnica między średnimi dwóch rozkładów wykładniczych sprawia, że trudniej jest statystycznie wykryć różnicę.

Zwiększenie mocy testów: Moc wszystkich trzech testów zwiększa się w miarę oddalania

wartości μ_2 od wartości μ_1 . To pokazuje, że testy stają się bardziej skuteczne w odrzucaniu fałszywej hipotezy zerowej, gdy rzeczywista różnica między średnimi jest większa.

Porównanie testów: Test Wilcoxona, który jest testem nieparametrycznym, wydaje się mieć nieco niższą moc niż testy parametryczne (test t-Studenta i t-Welcha) w okolicy, gdzie μ_2 jest znacznie większe lub mniejsze od μ_1 . Jest to szczególnie widoczne, gdy μ_2 jest większe od μ_1 .

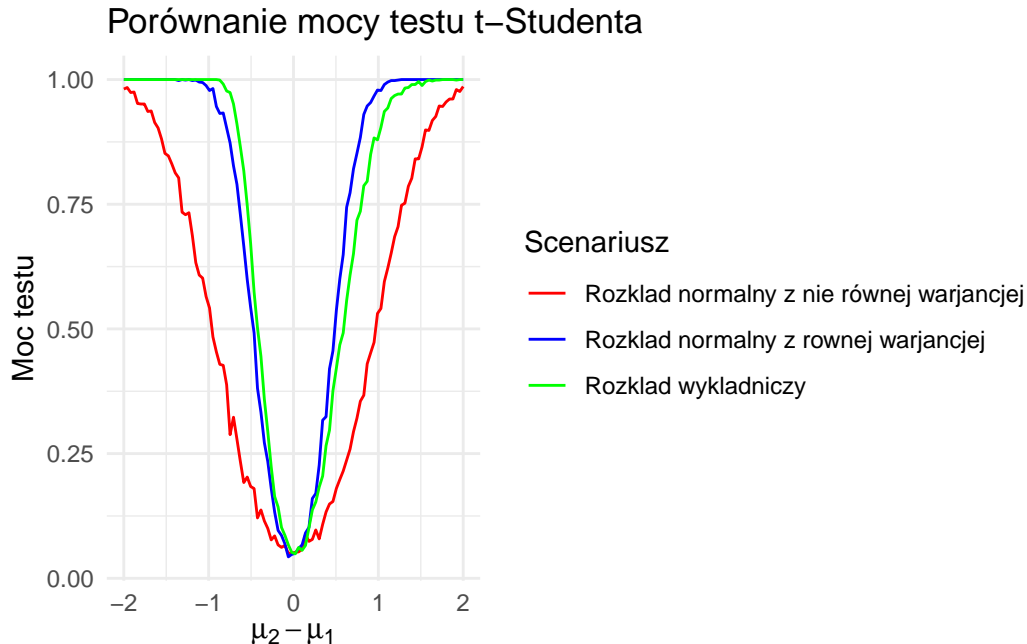
Brak jednostajnie najmocniejszego testu: Nie ma testu, który byłby jednostajnie najmocniejszy we wszystkich punktach zakresu μ_2 . Test t-Studenta i test t-Welcha mają bardzo zbliżoną moc, podczas gdy test Wilcoxona ma nieco niższą moc, szczególnie dla większych wartości μ_2 .

Podsumowując, wykres wskazuje, że testy parametryczne mają lepszą moc w wykrywaniu różnic w rozkładach wykładniczych, gdy różnice są większe, ale nie ma jednego testu, który byłby najlepszy we wszystkich sytuacjach. Wybór testu powinien zatem zależeć od założeń dotyczących danych oraz od specyfiki badanego problemu.

4. Próba wyboru testu jednostajnie najmocniejszego

4.1 Test t-Studenta

Wykonajmy wykres funkcji mocy dla testu t-Studenta na przedziale $(-2, 2)$, dla trzech prób które już symulowaliśmy w sekcji 3.

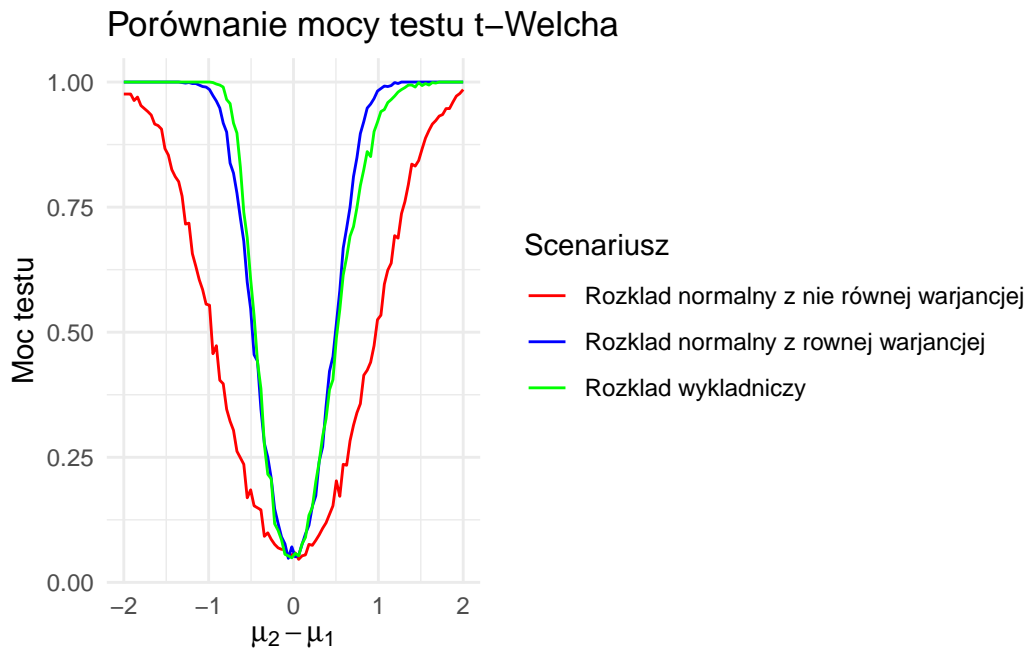


Test t-Studenta wydaje się być stosunkowo stabilny w obu scenariuszach z rozkładem normalnym, niezależnie od równości wariancji.

W przypadku rozkładu wykładniczego, moc testu jest generalnie niższa, co sugeruje, że test t-Studenta może nie być najlepszym wyborem dla danych, które nie są normalnie rozłożone.

4.2 Test t-Welcha

Wykonajmy wykres funkcji mocy dla testu t-Welcha na przedziale $(-2, 2)$, dla trzech prób które już symulowaliśmy w sekcji 3.



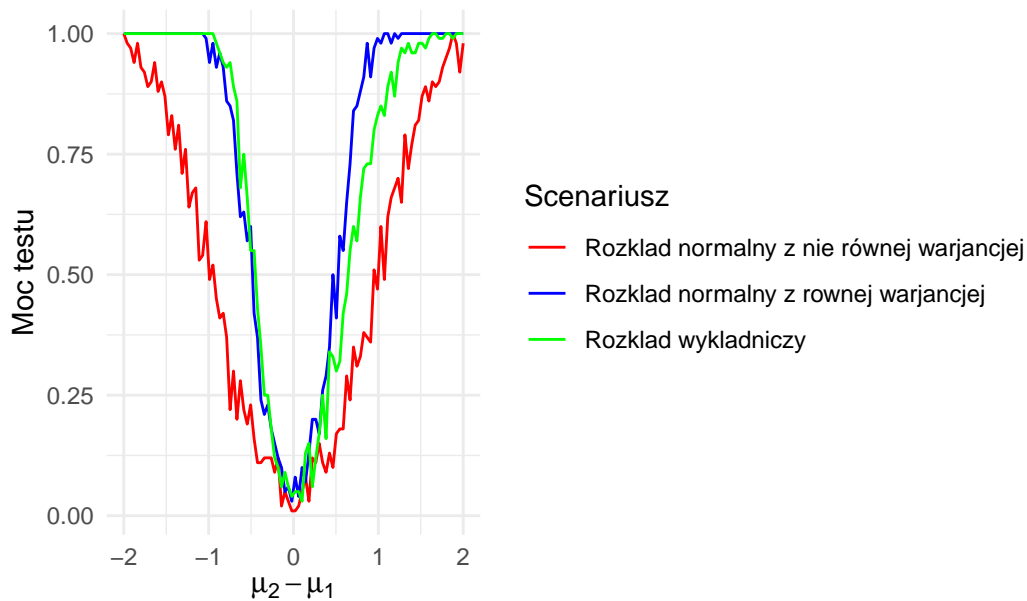
Test t-Welcha, który jest bardziej odpowiedni dla prób o różnych wariancjach, pokazuje większą moc w scenariuszu z nierównymi wariancjami, co jest oczekiwanym wynikiem.

Podobnie jak w przypadku testu t-Studenta, dla rozkładu wykładniczego moc testu t-Welcha jest niższa.

4.3 Test sumy rang Wilcoxona

Wykonajmy wykres funkcji mocy dla testu sumy rang Wilcoxona na przedziale $(-2, 2)$, dla trzech prób które już symulowaliśmy w sekcji 3.

Porównanie mocy testu sumy rang Wilcoxona



Test Wilcoxona, jako test nieparametryczny, pokazuje lepszą zdolność do radzenia sobie z danymi z rozkładu wykładniczego w porównaniu do testów parametrycznych.

W przypadku rozkładów normalnych, test Wilcoxona ma porównywalną moc do testu t-Studenta i testu t-Welcha, co sugeruje, że jest dość uniwersalnym testem, ale może nie być najmocniejszym wyborem w przypadku danych dobrze pasujących do założeń testów parametrycznych.

5. Wnioski

Na podstawie wykresów nie ma jednoznacznego testu, który byłby jednostajnie najmocniejszy w każdym scenariuszu. Wybór testu powinien zależeć od rozkładu danych:

- Dla danych normalnie rozłożonych z równymi wariancjami, test t-Studenta i test t-Welcha są podobnie skuteczne.
- Dla danych normalnie rozłożonych z różnymi wariancjami, test t-Welcha wydaje się być bardziej odpowiedni.
- Dla danych z rozkładu wykładniczego, test Wilcoxona może być lepszym wyborem ze względu na mniejsze założenia dotyczące rozkładu danych.

Wybór odpowiedniego testu powinien uwzględniać charakterystykę danych, wielkość próby oraz założenia dotyczące rozkładów próbek.