Kamil Marciniak

Raport 5

20 maja 2021

Spis treści

1. Zadanie 1

1.1. przedziały ufności

```
t.test(X,Y)
```

```
##
##  Welch Two Sample t-test
##
## data:  X and Y
## t = -6.4156, df = 9.313, p-value = 0.0001057
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -26.71078 -12.83723
## sample estimates:
##  mean of x  mean of y
##  0.4146278 20.1886360
```

Zacznijmy od skonstruowania odpowiednich przedziałów ufności dla podpunktu a. 95% przedział ufności dla różnicy średnich skonstruowany korzystając z nieuśrednionego SE wynosi [-26.7107843, -12.8372321] i jak widać zgadza się z wynikiem obliczonym przy użyciu funkcji t.test

```
t.test(X,Y,var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  X and Y
## t = -3.3215, df = 13, p-value = 0.005515
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.874019  -5.904046
## sample estimates:
##  mean of x  mean of y
##  0.5066316 17.3956640
```

95% przedział ufności dla różnicy średnich skonstruowany korzystając z uśrednionego SE wynosi [-27.8740192, -5.9040455] i jak widać zgadza się z wynikiem obliczonym przy użyciu funkcji t.test z odpowiednim parametrem.

Po zaimplementowaniu testów ręcznie od teraz będę już korzystał z wbudowanej funkcji.

Skonstruuję teraz przedziały ufności dla podpunktu b

95% przedział ufności dla różnicy średnich skonstruowany korzystając z nieuśrednionego SE wynosi [-21.2444109, -18.6207523] 95% przedział ufności dla różnicy średnich skonstruowany korzystając z uśrednionego SE wynosi [-21.1494193, -18.715744]

Skonstruuję teraz przedziały ufności dla podpunktu c

95% przedział ufności dla różnicy średnich skonstruowany korzystając z nieuśrednionego SE wynosi [-24.4779138, -13.6475807]

95% przedział ufności dla różnicy średnich skonstruowany korzystając z uśrednionego SE wynosi [-24.1017082, -14.0237864]

1.2. Prawdopodobieństwa pokrycia i szerokość przedziałów

Zajmijmy się teraz prawdopodobieństwem pokrycia dla podpunktu a

Dla metody NSE prawdopodobieństwo pokrycia wynosi 0.957 a dla metody USE wynosi 0.989. Jako że dla prób z podpunktu a odchylenia standardowe są różne to zgodnie z oczekiwaniem dla metody NSE otrzymujemy prawdopodobieństwo pokrycia bliskie 95%. Dla metody USE nie mamy spełnionych założeń i widzimy, że prawdopodobieństwo pokrycia jest większe niż dla metody NSE. Średnia szerokość przedziału ufności dla metody NSE wynosi 14.0945077 a dla metody USE wynosi 19.353929. Widać, że dla metody która jest poprawna w tym przypadku(NSE) średnia szerokość przedziału jest mniejsza.

Wykonajmy teraz to samo dla b

Dla metody NSE prawdopodobieństwo pokrycia wynosi 0.945 a dla metody USE wynosi 0.942. Jako że dla prób z podpunktu b odchylenia standardowe są równe to na podstawie teorii oczekujemy, że poprawna będzie tutaj metoda USE. Jak widać z naszych wyników dla metody NSE również otrzymujemy prawdopodobieństwo pokrycia bliskie 95% . Czyli w tym wypadku prawdopodobieństwo pokrycia obu metod jest podobne. Średnia szerokość przedziału ufności dla metody NSE wynosi 2.4658723 a dla metody USE wynosi 2.3074667. Widać, że dla metody USE w tym przypadku średnia szerokość przedziału jest mniejsza. Co nie dziwi ponieważ jest to poprawna metoda dla tych danych a mając dodatkową informację o równych odchyleniach standardowych test Studenta skutkuje w dokładniejszych przedziałach.

Wykonajmy teraz to samo dla c

Dla metody NSE prawdopodobieństwo pokrycia wynosi 0.95 a dla metody USE wynosi 0.932. Jako że dla prób z podpunktu c odchylenia standardowe są różne to na podstawie teorii oczekujemy, że poprawna będzie tutaj metoda NSE. Jak widać z naszych wyników dla metody NSE otrzymujemy prawdopodobieństwo pokrycia bliskie 95% a dla metody USE trochę mniejsze. Średnia szerokość przedziału ufności dla metody NSE wynosi 13.8622393 a dla metody USE wynosi 12.9168624. Widać, że dla metody USE w tym przypadku średnia szerokość przedziału jest mniejsza. Ale zasadniczo przedziały te mają podobną długość. Odnotujmy, że podpunkt c różni się od a tylko licznością jednej z prób a skutkuje to bardzo dużą różnicą szerokości przedziału jak i prawdopodobieństwa pokrycia dla metody USE.

2. Zadanie 2

2.1. Wstęp i opis danych

Warto rozpocząć od pewnych wiadomości wstępnych. Bazując na[1] przyjmijmy, że dla dorosłego człowieka poprawny poziom cholesterolu to mniej niż 200 miligramów na decylitr. A korzystając również z [2] możemy stwierdzić, że wysoki poziom cholesterolu zwiększa ryzyko chorób serca i zawału. Dowiedziałem się również, że dla pacjentów z chorobami serca którzy przeszli zawał lekarze rekomendują leczenie mające na celu obniżenie cholesterolu.[3] Okazuje się również, że przy stosowaniu takiego leczenia poziom cholesterolu może spaść w przeciągu kilku tygodni.[4]

Przechodząc do naszych danych, mamy 28 obserwacji w grupie badawczej i 30 obserwacji w grupie kontrolnej. Z treści zadania wiemy również, że osobom po zawale zmierzono poziom cholesterolu po 2,4, 14 dniach od zawału serca. Przyjmijmy, że w grupie kontrolnej są osoby które nie przebyły zawału serca i im zmierzono poziom cholesterolu tylko raz. Narysujmy teraz histogramy poziomu cholesterolu.

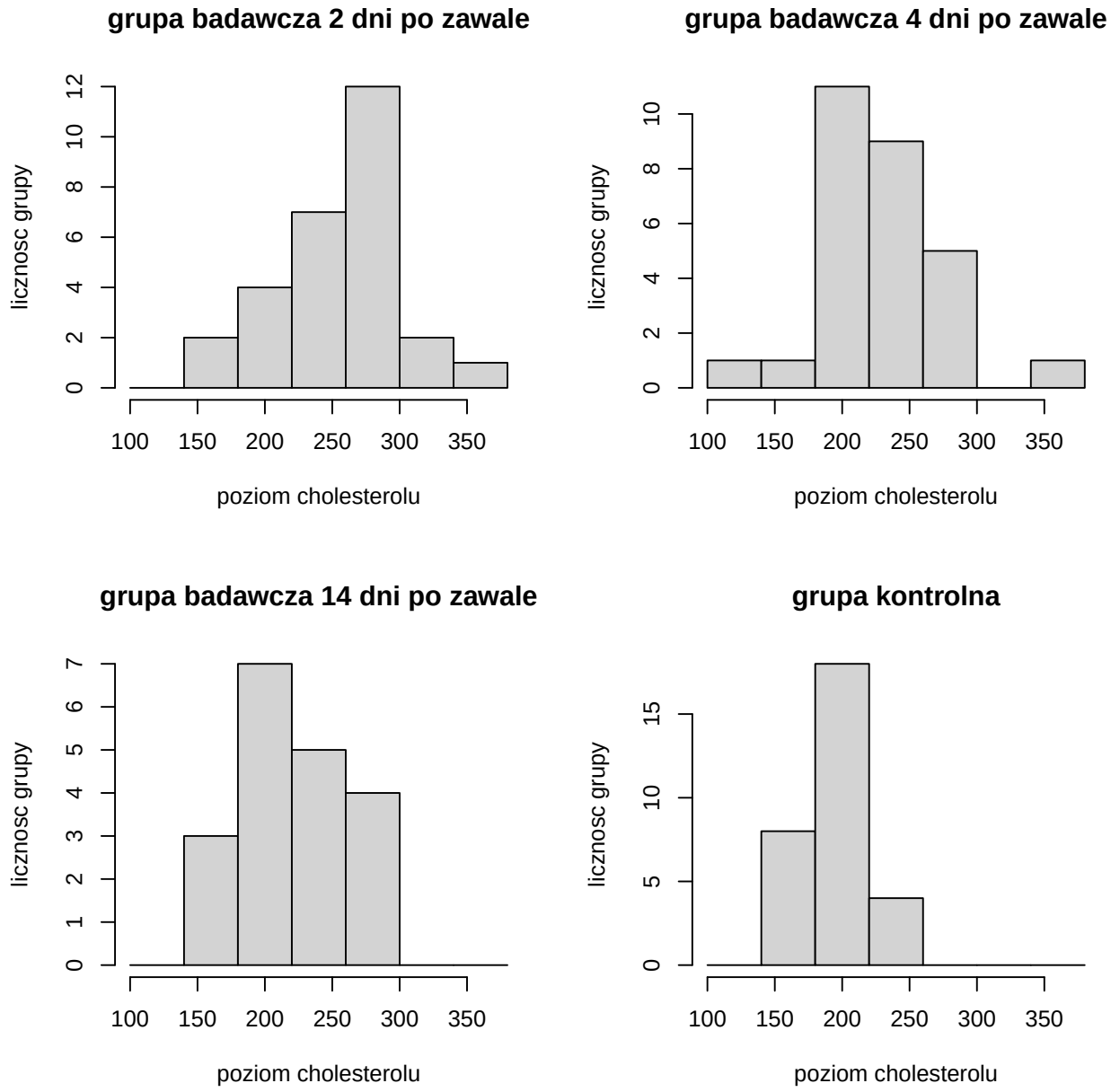
[1] https://www.medicalnewstoday.com/articles/315900#recommended-levels

[2] https://uhs.umich.edu/cholesterol

[3] https://www.medicinenet.com/cholesterol_levels_after_heart_attack_and_bypass/ask.htm

[4] https://www.medicalnewstoday.com/articles/how-long-does-it-take-to-lower-cholesterol#how-long-to-reduce-cholesterol

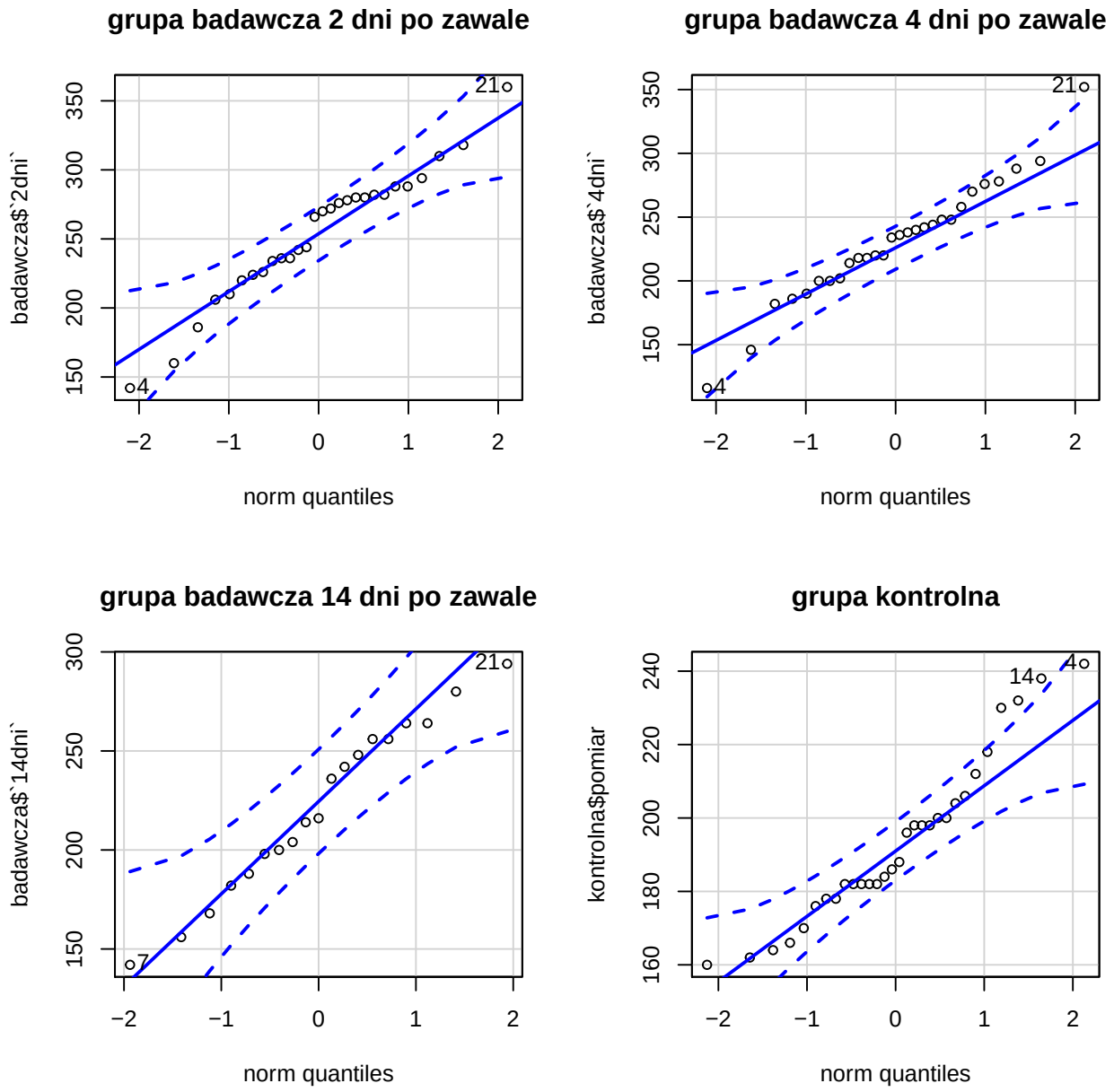
2.2. Histogramy



Postanowiłem nie ujednolicać osi y na histogramach ponieważ moim zdaniem przy tak małej ilości obserwacji czytelność wykresów się pogorszy. Ujednoliciłem za to osie x. Dzięki temu łatwo zauważyć, że poziom cholesterolu w grupie kontrolnej jest dość jednolity i nie ma wartości ekstremalnych, a już dla osób z grupy po przebytym zawale mamy sporo dużych wartości. Przypomnijmy, że poziom cholesterolu powyżej 200 już jest uznawany za niezdrowy a mamy tu bardzo wiele takich obserwacji. Warto również odnotować, że wraz z upływem czasu po zawale zmniejsza się liczba osób mających skrajnie wysoki poziom cholesterolu.

2.3. Wykresy kwantylowe

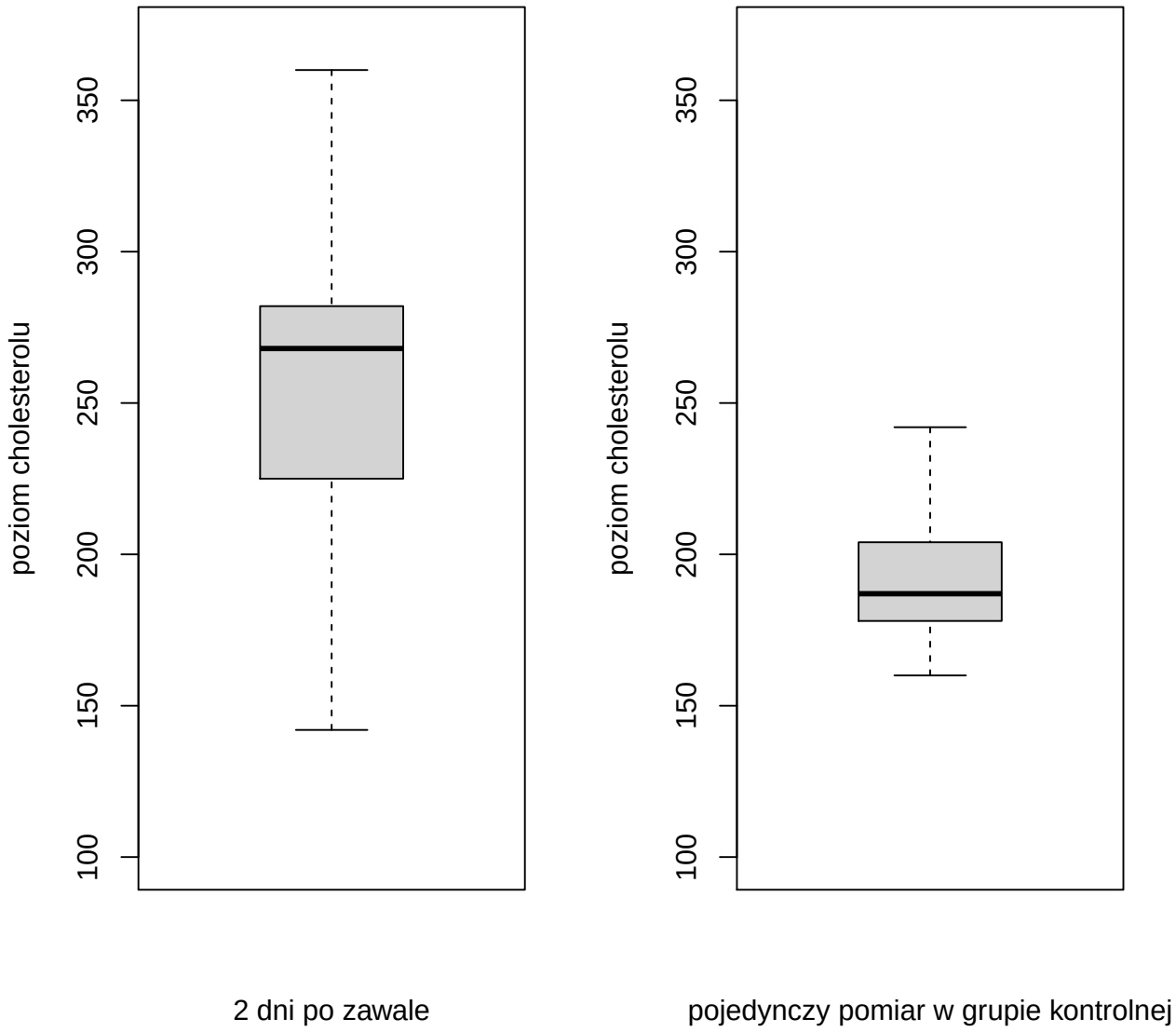
```
## [1]  4 21
## [1] 21  4
## [1]  7 21
```




```
## [1]  4 14
```

Z wykresów kwantylowych widzimy, że kwantyle naszych rozkładów mieszczą się między przerywanymi liniami więc nasze rozkłady możemy przybliżać rozkładem normalnym. Dzięki temu w podpunkcie c będziemy mogli użyć testu Studenta. Co prawda w grupie kontrolnej te kwantyle trochę wychodzą poza oczekiwany pas ale mimo to testowana hipoteza powinna dobrze przybliżać rzeczywiste dane.

2.4. Wykresy pudełkowe



Na wykresach pudełkowych również potwierdza się, że w grupie kontrolnej poziom cholesterolu jest niższy od grupy badawczej. Przy czym różnice są kolosalne. Wartość mediany dla poziomu cholesterolu 2 dni po zawale jest większa od wartości maksymalnej w grupie kontrolnej. Zauważmy również, że w grupie badawczej 2 dni po zawale zdecydowana większość osób ma podwyższony poziom cholesterolu poza normę. A w grupie kontrolnej około 75% osób ma prawidłowy poziom.

2.5. Test Studenta

Przejdźmy teraz do testu Studenta średniego poziomu cholesterolu w grupie badawczej 2 dni po zawale i w grupie kontrolnej.

```
t.test(badawcza$`2dni`,kontrolna$pomiar)
```

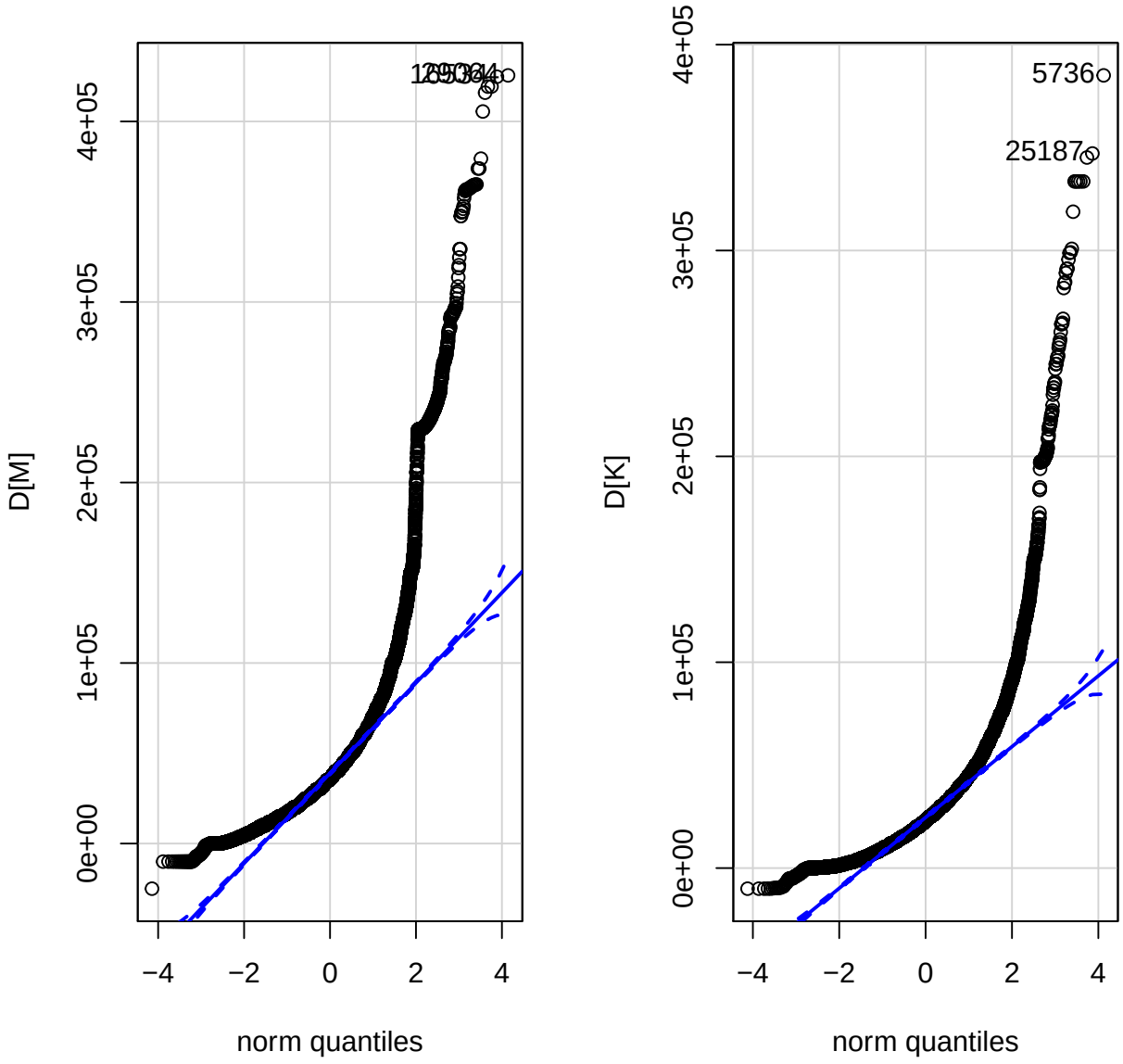
```
##
##  Welch Two Sample t-test
##
## data:  badawcza$`2dni` and kontrolna$pomiar
## t = 6.1452, df = 37.675, p-value = 3.721e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  40.76212 80.82835
## sample estimates:
## mean of x mean of y
##  253.9286  193.1333
```

W naszym wypadku hipoteza zerowa to twierdzenie, że średnie są równe, a hipoteza alternatywna że są różne. Jak widzimy na poziomie istotności 0.05 odrzucamy hipotezę zerową ponieważ p-wartość jest mniejsza niż 0.05. Na podstawie przedziału ufności możemy również ocenić, że większą średnią ma grupa badawcza 2 dni po zawale co zgadza się z moimi wcześniejszymi analizami.

3. Zadanie 3

3.1. Wykresy kwantylowe

```
## [1] 29064 16534
```



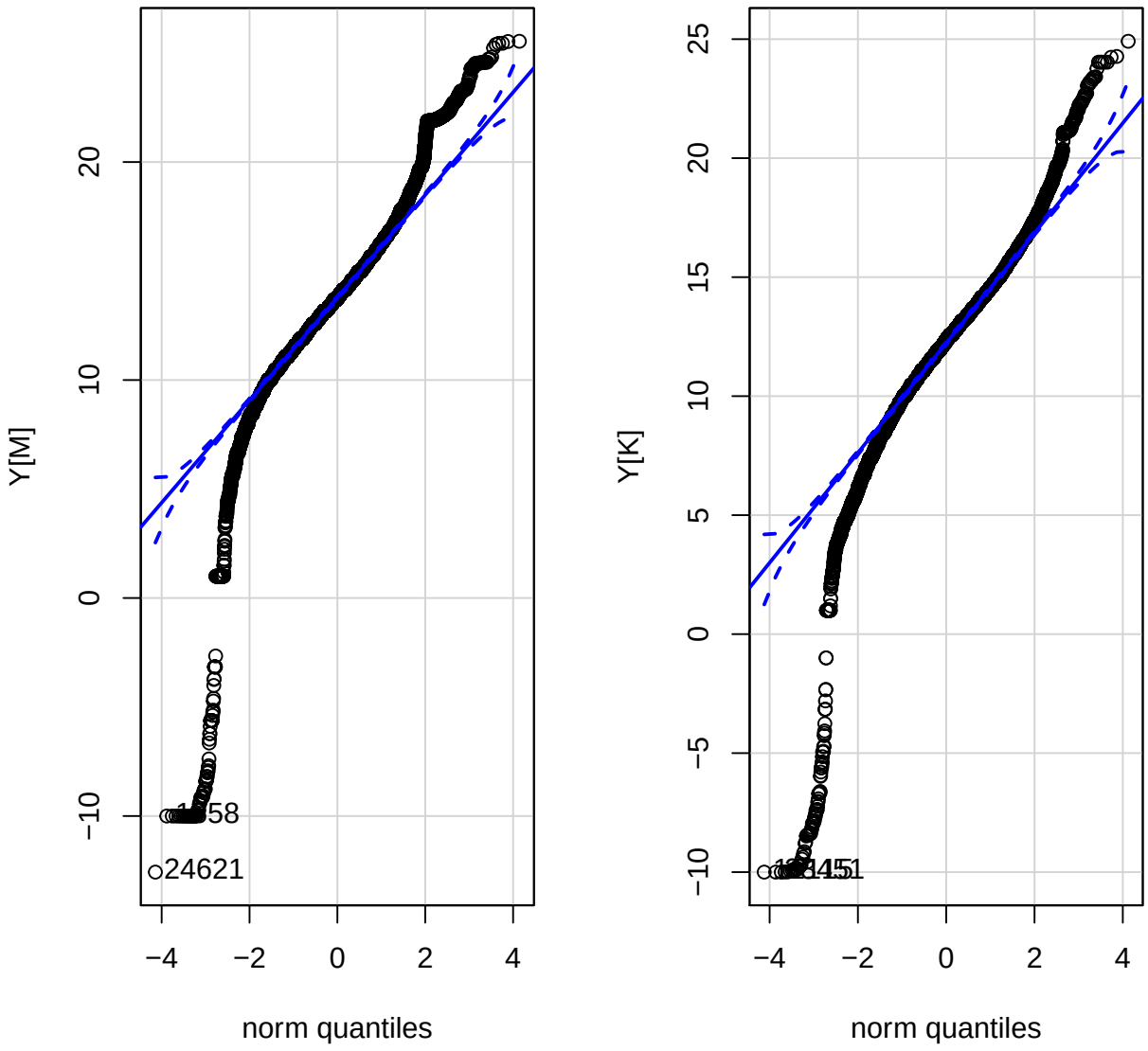
```
## [1]  5736 25187
```

Tak jak można się było spodziewać kwantyle dochodów nie są zgodne z rozkładem normalnym dla obu płci. Między innymi dlatego, że jak analizowałem na pierwszej liście rozkład dochodów jest skośny w prawo.

Narysujmy jeszcze wykresy kwantylowe dla zmiennej Y

3.2. Wykresy kwantylowe po modyfikacji

```
## [1] 24621  1558
```



```
## [1] 11115 23451
```

Pomimo tej modyfikacji nadal rozkład dochodów przeskalowany przez odpowiedni pierwiastek nie jest rozkładem normalnym ani dla kobiet ani dla mężczyzn.Ale mimo to, dzięki temu że nasza próba jest bardzo duża to w oparciu o CTG skorzystamy z testu Studenta.

3.3. Test Studenta

```
t.test(Y[M],Y[K])
```

```
##
##  Welch Two Sample t-test
##
## data:  Y[M] and Y[K]
## t = 67.308, df = 55834, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.617287 1.714303
## sample estimates:
## mean of x mean of y
##  13.82452  12.15872
```

Ponieważ p-wartość jest mniejsza od 0.05 to na poziomie istotności 0.05 odrzucamy hipotezę zerową. Czyli możemy wnioskować że jest istotna statystycznie różnica między zarobkami kobiet i mężczyzn. Jak wiemy z przeprowadzonego testu jak i analiz w pierwszym raporcie to mężczyźni mają większe zarobki.