PSP – Raport 4

Nela Tomaszewicz

Czerwiec 2020

1 Zadanie 1

W zadaniu generujemy parami różne próby z rozkładu normalnego, a następnie konstruujemy 95% przedziały ufności dla różnicy dwóch średnich $\mu_2 - \mu_1$. Przedział na poziomie ufności $1 - \alpha$ dla różnicy dwóch średnich wyrażamy następującym wzorem:

$$(\bar{y_2} - \bar{y_1}) \pm t_{(df, 1-\frac{\alpha}{2})} SE_{\bar{y_2}-\bar{y_1}}, \tag{1}$$

gdzie $\bar{y_1}$ oznacza średnią próbkową z pierwszej próby, $\bar{y_2}$ średnią próbkową z drugiej próby, α oznacza poziom istotności, $t_{(df,1-\frac{\alpha}{2})}$ jest kwantylem rzędu $1 - \frac{\alpha}{2}$ z rozkładu t–Studenta z liczbą stopni swobody df określoną następującym wzorem:

$$df = \frac{(SE_1^2 + SE_2^2)^2}{\frac{SE_1^4}{n_1 - 1} + \frac{SE_2^4}{n_2 - 1}}, \tag{2}$$

gdzie n_1 i n_2 to liczności pierwszej i drugiej próby, natomiast SE_1 i SE_2 to standardowe błędy średniej $\left(SE_1 = \frac{s_1}{n_1}, SE_2 = \frac{s_2}{n_2}\right)$. Ostatnią nieomówioną częścią wzroru (1) jest standardowy błąd dla różnicy dwóch średnich, czyli $SE_{\bar{y_2}-\bar{y_1}}$. Może on być wyrażony na dwa sposoby: jako błąd nieuśredniony (niełączony, ang. *unpooled*) lub uśredniony (łączony, ang. *pooled*). SE nieuśrednione wyrażamy wzorem:

$$(N)SE = \sqrt{SE_1^2 + SE_2^2}. \tag{3}$$

Do wyznaczenia uśrednionego SE na początek musimy wyznaczyć sumę kwadratów odchyleń dla obydwu prób, czyli $SS_1 = \sum (y_{1,i} - \bar{y_1})^2$ oraz $SS_2 = \sum (y_{2,i} - \bar{y_2})^2$. Następnie, wyznaczamy uśrednioną wariancję, czyli $s_c^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$. Teraz możemy skonstruować uśrednione SE :

$$(U)SE = \sqrt{sc^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{4}$$

Pierwszą część zadania polegającą na wyznaczeniu przedziałów ufności wykonam przy pomocy następującej funkcji.

```
set.seed(1)
```

```r
conf_interval <- function(n1, mi1, sd1, n2, mi2, sd2) {
  norm_a_1 <- rnorm(n1, mi1, sd1)
  norm_a_2 <- rnorm(n2, mi2, sd2)
  alfa <- 0.05

  SE_1 <- sd(norm_a_1)/sqrt(n1)
  SE_2 <- sd(norm_a_2)/sqrt(n2)

  NSE <- sqrt(SE_1^2 + SE_2^2)

  mean_1 <- mean(norm_a_1)
  mean_2 <- mean(norm_a_2)
  SS_1 <- sum((norm_a_1 - mean_1)^2)
  SS_2 <- sum((norm_a_2 - mean_2)^2)

  sc <- sqrt((SS_1 + SS_2)/(length(norm_a_1) + length(norm_a_2) - 2))

  USE <- sc*sqrt(1/length(norm_a_1) + 1/length(norm_a_2))

  df <- (SE_1^2 + SE_2^2)^2/((SE_1^4/(n1 - 1)) + (SE_2^4/(n2 - 1)))

  two_means <- mean_2 - mean_1
  t <- qt(df = df, p = 1 - alfa/2)

  # dla NSE
  left_N <- two_means - t*NSE
  right_N <- two_means + t*NSE

  # dla USE
  left_U <- two_means - t*USE
  right_U <- two_means + t*USE

  cat("Przedzial ufnosci z NSE: [", left_N, ",", right_N, "]", "\n")
  cat("Przedzial ufnosci z USE: [", left_U, ",", right_U, "]")
}
```

Funkcja przyjmuje parametry obydwu rozkładów i zwraca odpowiednie przedziały ufności.

1.1 Podpunkt (a)

Generujemy 5–elementową próbę z rozkładu $N(0, 1)$ oraz 10–elementową próbę z rozkładu $N(20, 10)$ i konstruujemy dwa przedziały ufności (jeden z $(N)SE$, a drugi z $(U)SE$) dla różnicy średnich.

```
## Przedzial ufnosci z NSE: [ 12.87627 , 28.59777 ]
## Przedzial ufnosci z USE: [ 9.475468 , 31.99858 ]
```

Przedział ufności skonstruowany przy użyciu $(U)SE$ jest szerszy.

1.2 Podpunkt (b)

Generujemy 5–elementową próbę z rozkładu $N(0,1)$ oraz 10–elementową próbę z rozkładu $N(20,1)$ i wyznaczamy przedziały ufności.

```
## Przedzial ufnosci z NSE: [ 18.6143 , 20.19922 ]
## Przedzial ufnosci z USE: [ 18.4176 , 20.39592 ]
```

Przedziały ufności są do siebie bardzo podobne.

1.3 Podpunkt (c)

Generujemy 10–elementową próbę z rozkładu $N(0,1)$ oraz 10–elementową próbę z rozkładu $N(20,10)$ i wyznaczamy przedziały ufności.

```
## Przedzial ufnosci z NSE: [ 16.94289 , 25.49839 ]
## Przedzial ufnosci z USE: [ 16.94289 , 25.49839 ]
```

Przedziały są takie same.

1.4 Powtórzenie 1000 razy

Aby lepiej zrozumieć badane przedziały, powtarzamy 1000 razy eksperyment polegający na losowaniu odpowiednich prób i wyznaczaniu przedziałów ufności i porównujemy prawdopodobieństwo pokrycia $\mu_2 - \mu_1$ oraz średnie długości dla obu przedziałów. W tym celu przerobiłam funkcję używaną poprzednio na taką, która zwraca wektor zawierający kolejno: lewy i prawy koniec dla przedziału konstruowanego przy użyciu $(N)SE$ oraz lewy i prawy koniec dla przedziału konstruowanego przy użyciu $(U)SE$. Następnie zaimplementowałam następującą funkcję powtarzającą eksperyment 1000 razy i zwracającą odpowiednie wartości dla badanych parametrów.

```
conf_interval_rep <- function(n1, mi1, sd1, n2, mi2, sd2) {
  prob_N <- 0
  prob_U <- 0

  v_N <- c()
  v_U  <- c()

  for(i in 1:1000) {
    interval <- conf_interval(n1, mi1, sd1, n2, mi2, sd2)
    v_N[i] <- interval[2] - interval[1]
    v_U[i] <- interval[4] - interval[3]

    real_mean <- mi2 - mi1

    if(interval[1] <= real_mean & interval[2] >= real_mean){
      prob_N = prob_N + 1
    }
```

```
    if(interval[3] <= real_mean & interval[4] >= real_mean) {
      prob_U = prob_U + 1
    }
  }

  cat("Dlugosc przedzialu z NSE: ", mean(v_N), ", p-stwo pokrycia: ",
      prob_N/1000, "\n")
  cat("Dlugosc przedzialu z USE: ", mean(v_U), ", p-stwo pokrycia: ",
      prob_U/1000)
}
```

Dla prób z podpunktu 1.1 mamy następujące rezultaty.

```
## Dlugosc przedzialu z NSE:  14.08733 , p-stwo pokrycia:  0.956
## Dlugosc przedzialu z USE:  20.13295 , p-stwo pokrycia:  0.995
```

Dla prób z podpunktu 1.2:

```
## Dlugosc przedzialu z NSE:  2.511464 , p-stwo pokrycia:  0.951
## Dlugosc przedzialu z USE:  2.506996 , p-stwo pokrycia:  0.961
```

Nastomiast dla prób z podpunktu 1.3:

```
## Dlugosc przedzialu z NSE:  14.08783 , p-stwo pokrycia:  0.949
## Dlugosc przedzialu z USE:  14.08783 , p-stwo pokrycia:  0.949
```

Dla pierwszego przypadku możemy zaobserwować największą różnicę w prawdopodobieństwie pokrycia przedziału oraz w jego długości. Wynika to z tego, że próby różnią się od siebie najbardziej wśród wszystkich przykładów. Druga badana próba w podpunkcie 1.1 ma 10 razy większe odchylenie standardowe oraz jej wielkość jest dwa razy większa od pierwszej próby. Skutkuje to szerszym przedziałem, a co za tym idzie – większym prawdopodobieństwem pokrycia średniej przez przedział. W podpunkcie 1.2, różnice są już bardzo niewielkie, odchylenia standardowe dla obydwu prób są takie same, natomiast druga próba znów ma dwa razy więcej elementów, co skutkuje odrobinę większym prawdopodobieństwem pokrycia przedziału, chociaż różnica w długości jest mała. W ostatnim przykładzie, mamy próby o różnym odchyleniu standardowym, ale o tej samej liczności. W taki przypadku obydwie metody wyznaczania przedziałów ufności dają ten sam wynik.

1.5 Podsumowanie i wnioski

Sprawdziliśmy, że metoda wyznaczania przedziałów ufności przy wykorzystaniu nieuśrednionego i uśrednionego SE daje ten sam rezultat dla równolicznych prób, nawet jeśli mają one różne ochylenia standardowe.

2 Zadanie 2

Badamy zbiór danych z pliku `chol.txt` zawierający dane dotyczące poziomu cholesterolu pacjentów kilka dni po zawale. Celem zadania jest przetestowanie hipotezy mówiącej o

tym, że poziom cholesterolu u osób z grupy kontrolnej (pacjenci nie będący po zawale) różni się od poziomu cholesterolu u pacjentów po zawale. Wykorzystamy test t-Studenta dla dwóch niezależnych prób. W R skorzystamy z funkcji `t.test()`, ale matematycznie test Studenta dla dwóch niezależnych prób opiera się na statystyce testowej:

$$t_s = \frac{\bar{y_1} - \bar{y_2}}{SE}, \tag{5}$$

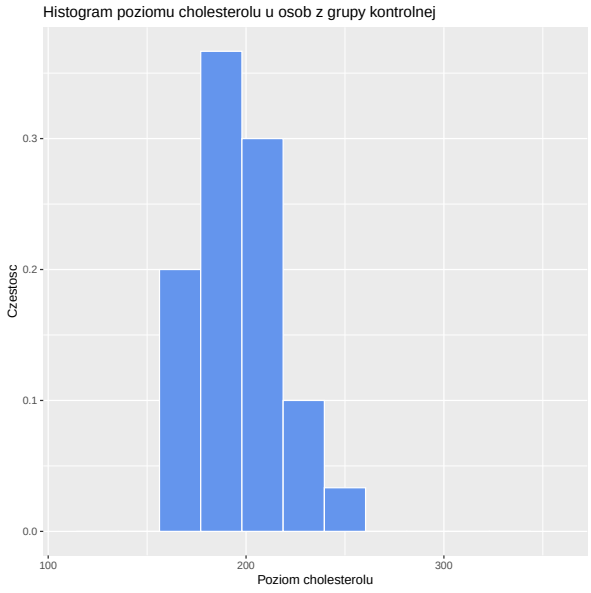
gdzie SE może być uśrednione lub nieuśrednione o wzorach takich jak w zadaniu 1. Każdy test wykonujemy na pewnym poziomie istotności α (prawdopodobieństwo popełnienia błędu I-go rodzaju) i to właśnie przy pomocy poziomu istotności konstruujemy obszar krytyczny (obszar odrzuceń), który determinuje decyzje o odrzuceniu lub przyjęciu hipotezy zerowej H_0. Obszar krytyczny tworzony jest przy pomocy kwantyli rzędu α lub $\frac{\alpha}{2}$ (w zależności od tego, czy badamy alternatywę obustronną lub jednostronną) z rozkładu t-Studenta. Ważną badaną wielkością dotyczącą testów statystycznych jest też p–wartość, czyli graniczny poziom istotności, najmniejszy, przy którym zaobserwowana wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej (1). Jeśli p–wartość jest mniejsza bądź równa poziomowi istotności α, odrzucamy H_0 na rzecz H_1. W przeciwnym razie przyjmujemy hipotezę zerową. Zarówno statystyka testowa jak i p–wartość są obliczane przez funkcję `t.test()`.

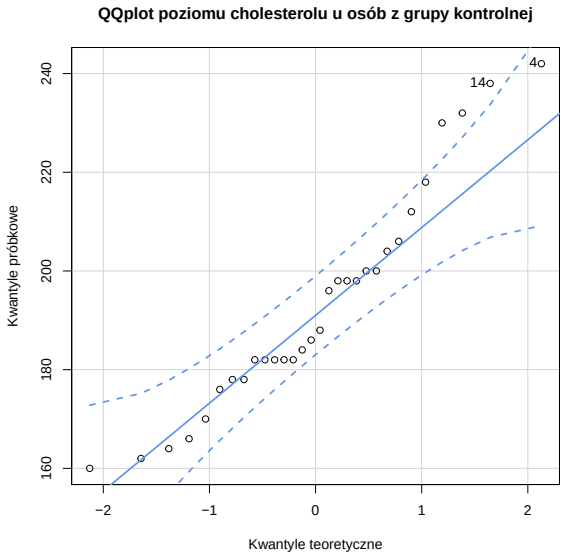
2.1 Porównanie histogramów i wykresów kwantylowych

Test Studenta ma swoje założenia dotyczące badanych danych. Pierwszym z nich jest założenie dotyczące rozkładu, mówiące o tym, że rozkład wyników zmiennej zależnej w każdej z analizowanych grup ma być zbliżony do normalnego (2). To założenie zbadamy w tym podpunkcie, na początku metodą graficzną wykorzystującą histogramy i wykresy kwantylowe, a następnie testem Shapiro–Wilka, będącym dobrym testem normalności dla małych prób. W zbiorze jest 28 obserwacji dla pacjentów po zawale i 30 dla osób z grupy kontrolnej.

2.1.1 Grupa kontrolna

Wyznaczamy histogram oraz wykres kwantylowy poziomu cholesterolu dla osób z grupy kontrolnej.





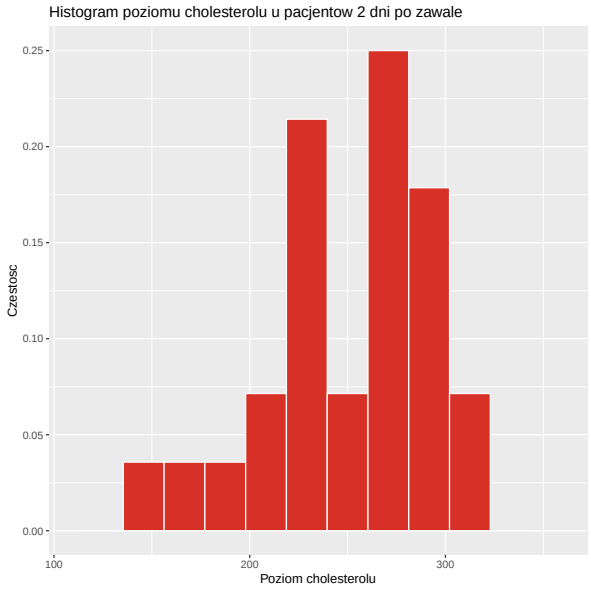
Zarówno histogram jak i wykres kwantylowy pokazują, że rozkład nie jest normalny. Na wykresie kwantylowym możemy zauważyć obserwacje znajdujące się poza przedziałem ufności. Sprawdźmy zatem testem Shapiro–Wilka, czy z puntu widzenia statystyki jesteśmy w stanie przyjąć, że rozkład poziomu cholesterolu u osób z grupy kontrolnej jest w przybliżeniu normalny. W tym celu używamy funkcji `shapiro.test()`. Przyjmujemy poziom istotności $\alpha = 0.05$.

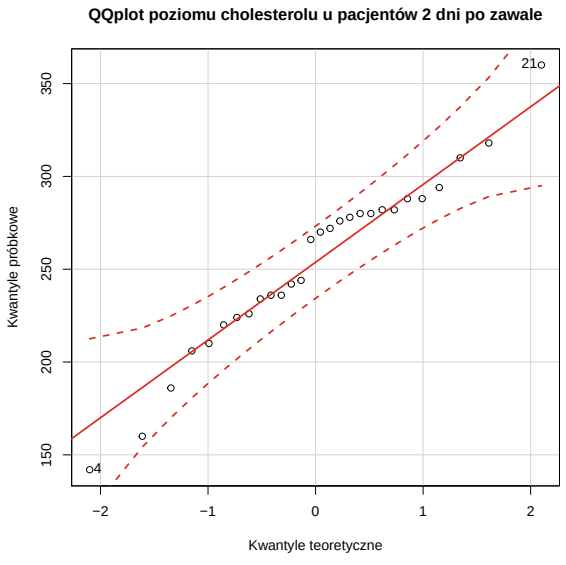
```
##
##  Shapiro-Wilk normality test
##
## data:  chol2$chol
## W = 0.93943, p-value = 0.08778
```

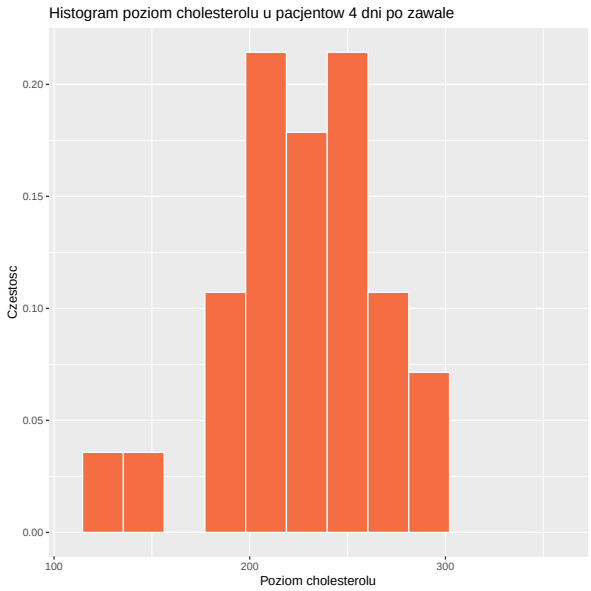
Interpretacja p–wartości dla testu Shapiro–Wilka mówi o tym, że jeśli jest ona większa od poziomu istotności α, to możemy przyjąć, że dane w przybliżeniu pochodzą z rozkładu normalnego. W naszym przypadku p–wartość $\approx 0.09 > 0.05$, stąd przyjmujemy, że rozkład poziomu cholesterolu dla osób z grupy kontrolnej jest w przybliżeniu normalny.

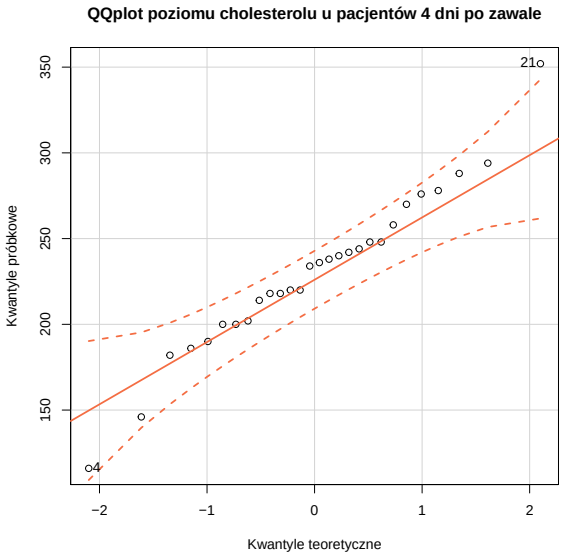
2.1.2 Pacjenci po zawale

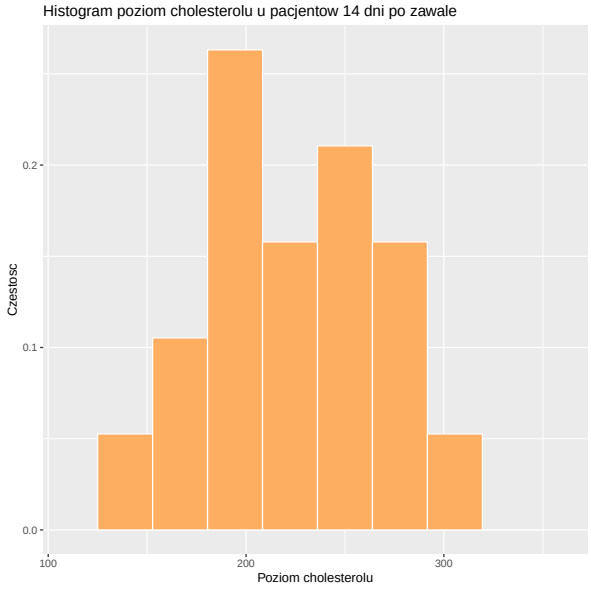
Badamy dane dotyczące poziomu cholesterolu u pacjentów 2, 4 i 14 dni po zawale.

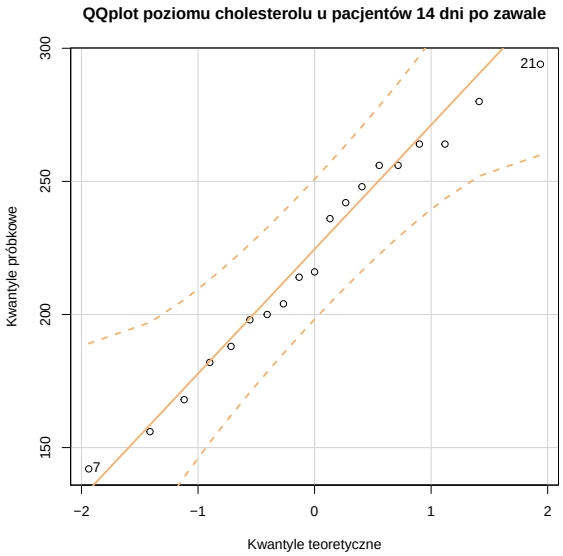












W przypadku pacjentów po zawale możemy zauważyć, że jedynie dla pacjentów 4 dni po zawale jedna obserwacja znajduje się poza przedziałem ufności na wykresie kwantylowym. Jeśli chodzi o same pomiary poziomu cholesterolu, możemy zauważyć, że najwyższy bin „przesuwa" się od poziomu pomiędzy 200 a 300 do poziomu około 200 wraz z mijającym czasem. Zgodnie z (3) prawidłowy poziom cholesterolu we krwi wg. WHO powinien wynosić 180 mg/dl. Możemy zauważyć, że bin zawierający taką wartość faktycznie jest najwyższy na histogramie dla grupy kontrolnej. 14 dni po zawale zbliżamy się do tej wartości. Najgorsze wskazania poziomu cholesterolu są u pacjentów 2 dni po zawale– dochodzą do wartości nawet przekraczających 300 mg/dl.

Ponieważ w późniejszej części zadania będziemy wykonywać test Studenta dla osób zdrowych i pacjentów dwa dni po zawale, to sprawdzimy także normalność tej drugiej grupy używając testu Shapiro–Wilka na poziomie istotności $\alpha = 0.05$.

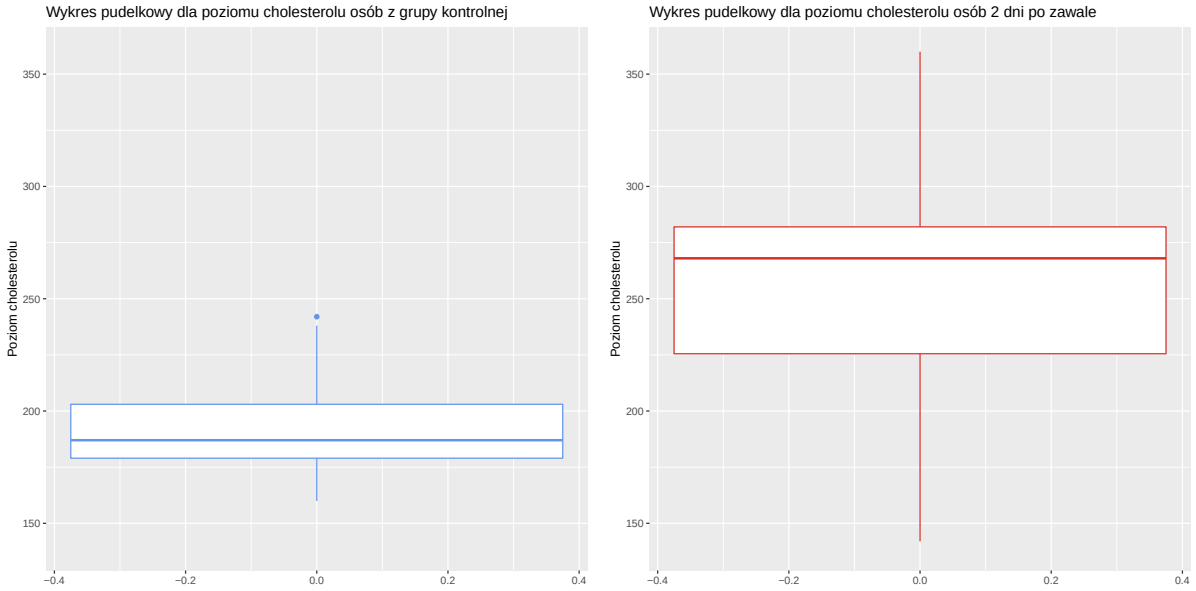
```
##
##  Shapiro-Wilk normality test
##
## data:  chol1$V8
```

```
## W = 0.96894, p-value = 0.5525
```

P–wartość jest większa od α, zatem przyjmujemy, że wartości poziomu cholesterolu dla pacjentów 2 dni po zawale mają w przybliżeniu rozkład normalny.

2.2 Wykresy pudełkowe

Drugim założeniem testu Studenta jest homogeniczność wariancji, czyli założenie, że wariancje w porównywanych grupach są do siebie podobne. W tym celu skonstruujemy wykresy pudełkowe (ang. *boxplot*) dla grupy kontrolenej i pacjentów dwa dni po zawale. Wykresy pudełkowe służą do ocenienia wartości typowych i nietypowych dla badanych danych. Z wykresu możemy odczytać 5 wartości : minimalną wartość typową, pierwszy kwartyl, medianę, trzeci kwartyl oraz maksymalną watość typową. Na wykresie pudełkowym również możemy zaznaczyć wartości odstające. Otrzymując od diagramu te informacje możemy wyciągnąć wnioski dotyczące, między innymi, rozrzutu danych, oceniając wysokość „pudełka" na wykresie, która wyznacza rozstęp miedzykwartylowy (IQR).



Odpowiednie dopasowanie osi Y pokazuje, że rozrzut danych w próbach jest różny. IQR dla danych dotyczących poziomu cholesterolu u pacjentów 2 dni po zawale jest ponad dwa razy większy niż dla grupy kontrolnej, co również pokazują wyliczenia mówiące, że $IQR_{kontrolna} = 24$, natomiast $IQR_{2-dni} = 56.5$. Stąd, test który tak naprawdę jest przeprowadzony to nie test t-Studenta, a **test t-Welcha**, będący uogólnieniem testu t-Studenta na populacje o różnych wariancjach (4).

2.3 Testowanie hipotez

Zanim przejdziemy do testowania należałoby omówić jeszcze kilka założeń dla (już teraz) testu t-Welcha. Poza założeniem dotyczącego homogeniczności wariancji, pozostałe założenia pozostają takie jak w przypadku testu t-Studenta. Zgodnie z (5) założeniem jest to, że dane są ciągłe (zgadza się, pomiary poziomu cholesterolu są ciągłe) oraz, że próby są niezależne i losowe. Miejmy nadzieję, że ten kto przeprowadził badanie dotyczące cholesterolu zadbał o to, aby w grupie kontrolnej faktycznie znalazły się osoby zdrowe, a w grupie osób 2 dni po zawale inne osoby, które rzeczywiście przeszły tę chorobę. To samo dotyczy losowości.

Mając sprawdzone założenia, przechodzimy do przetestowania hipotezy H_0 mówiącej o tym, że średni poziom cholesterolu u osób dwa dni po zawale jest taki sam jak u osób zdrowych, przeciwko hipotezie alternatywnej H_1 stanowiącej, że średni poziom cholesterolu jest różny. Oznaczmy jako μ_1 średni poziom cholesterolu u osób z grupy kontrolnej, a jako μ_2 średni poziom cholesterolu u osób 2 dni po zawale. Na poziomie istotności $\alpha = 0.05$ mamy:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned} \tag{6}$$

Przeprowadzamy test t-Welcha.

```
##
##  Welch Two Sample t-test
##
## data:  chol2$chol and chol1$chol2days
## 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:
##  -80.82835 -40.76212
## sample estimates:
## mean of x mean of y
##  193.1333  253.9286
```

Bardzo mała p–wartość (mniejsza od α) daje nam podstawę do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Możemy stwierdzić, że na poziomie istotności $\alpha = 0.05$ faktycznie średni poziom cholesterolu dla grupy kontrolnej jest inny niż dla osób 2 dni po zawale.

Zadajmy w takim razie inne pytanie: czy średni poziom cholesterolu u osób 2 dni po zawale jest **wyższy** niż u osób zdrowych. Mamy wtedy następujące hipotezy na poziomie istotności α:

$$\begin{aligned} H_0 &: \mu_1 \geqslant \mu_2 \\ H_1 &: \mu_1 < \mu_2. \end{aligned} \tag{7}$$

Przeprowadzamy test.

```
##
##  Welch Two Sample t-test
##
## data:  chol2$chol and chol1$chol2days
## t = -6.1452, df = 37.675, p-value = 1.86e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -44.1124
## sample estimates:
## mean of x mean of y
##  193.1333  253.9286
```

Znów, p–wartość jest znacząco mniejsza od poziomu istotności α, zatem możemy powiedzieć, że na poziomie istotności 0.05 badanie potwierdza, że średni poziom cholesterolu jest wyższy u osób 2 dni po zawale niż u osób zdrowych.

2.4 Podsumowanie i wnioski

Zadanie przeprowadziło nas przez proces testowania hipotez dla dwóch niezależnych prób testem t-Welcha zaczynając od sprawdzenia założeń testu metodami graficznymi i testem Shapiro–Wilka, a kończąc na wykonaniu samego testu i podjęciu decyzji o przyjęciu lub odrzuceniu hipotezy zerowej.

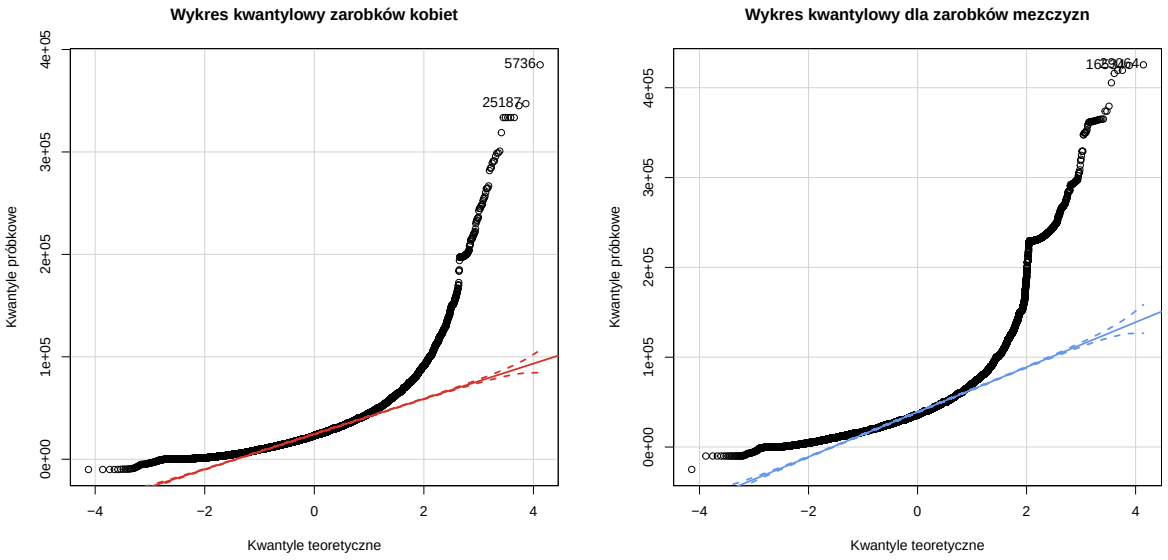
3 Zadanie 3

W zadaniu wracamy do zbioru danych `income.txt` zawierającego wyniki ankiet przeprowadzonych przez Bureau of Labor Statistics w USA na próbie 55 899 osób. Dla każdej osoby mamy podane:

- wiek (w latach),
- wykształcenie (1 – podstawowe, 2 – niepełne średnie, 3 – średnie, 4 – niepełne wyższe, 5 – wyższe (licencjat), 6 – wyższe (magisterium)),
- płeć (1–mężczyzna, 2–kobieta),
- roczne zarobki (w dolarach),
- sektor zatrudnienia (5 – sektor prywatny, 6 – sektor publiczny, 7 – samozatrudnienie).

3.1 Wykresy kwantylowe zarobków dla kobiet i mężczyzn

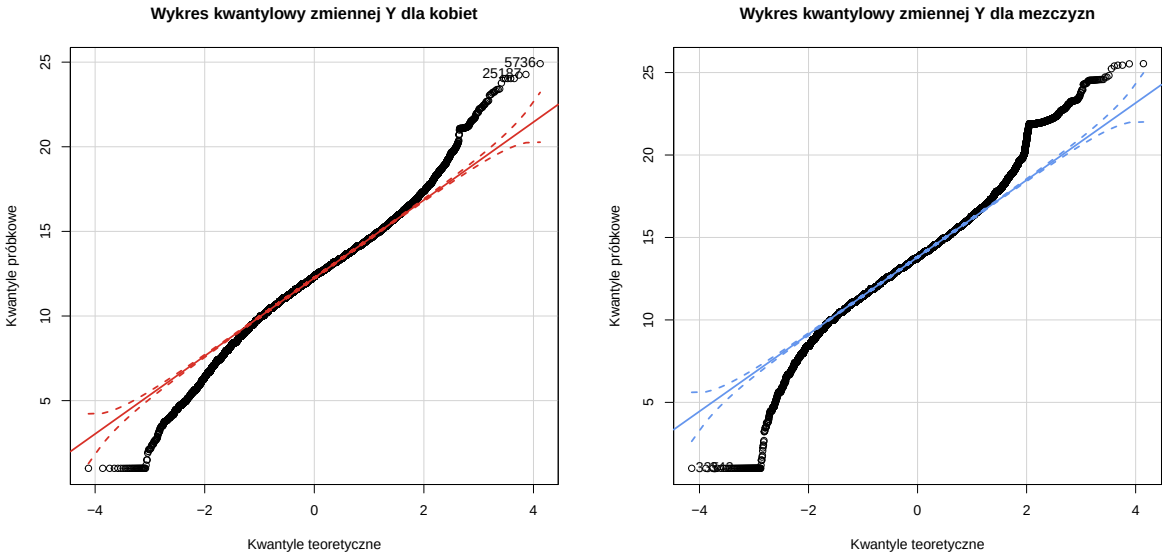
Ponieważ celem jest, podobnie jak w zadaniu 2, przeprowadzenie testu t-Studenta, musimy sprawdzić założenia dotyczące badanych danych. Zaczynamy od sprawdzenia normalności dla danych dotyczących zarobków dla kobiet i mężczyzn. Wykonamy to przy pomocy wykresu kwantylowego.



Możemy zauważyć, że bardzo duża część danych nie zawiera się w przedziale ufności wyznaczonym przez funkcję `qqPlot()`. Spróbujemy przekształcić dane tak, aby były trochę bardziej podobne do rozkładu normalnego.

3.2 Przekształcenie $Y = D^{0.25}$

Tworzymy nową zmienną $Y = D^{0.25}$, gdzie D oznacza dochód i konstruujemy dla niej wykresy kwantylowe, sprawdzając, czy rozkład zarobków dla kobiet i mężczyzn przypomina rozkład normalny.



Możemy zauważyć, że faktycznie, większa część wartości zmiennej Y zawiera się w przedziale ufności niż w przypadku zmiennej nieprzekształconej. Ponieważ próba jest bardzo duża, nie sprawdzimy dla niej normalności testem Shapiro–Wilka. Moglibyśmy wykonać to przy pomocy, na przykład, testu Lilliefors. Jednakże dzięki liczności próby skorzystamy z Centralnego Twierdzenia Granicznego i przeprowadzimy test t-Studenta uznając, że dane w przybliżeniu pochodzą z rozkładu normalnego.

3.3 Test Studenta

Badane dane są ciągłe i losowe, w przybliżeniu normalne, zatem sprawdzimy jeszcze homogeniczność wariancji. Aplikując funkcję `t.test()` możemy zobaczyć, czy został wykonany test Welcha czy Studenta, jednakże możemy również sprawdzić równość wariancji używając F-testu (mamy 2 próby) przyjmującego H_0 mówiącą o tym, że wariancje w obydwu próbach są równe. Wykonamy F-test na poziomie istotności $\alpha = 0.05$:

```
##
##  F test to compare two variances
##
## data:  women$Y and men$Y
## F = 0.8852, num df = 26596, denom df = 29131, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8646425 0.9062656
## sample estimates:
## ratio of variances
##          0.8852038
```

P–wartość jest znacząco mniejsza od poziomu istotności, jednakże zgodnie z (6) F-test jest bardzo wrażliwy na zaburzenia normalności danych. Przyjmiemy, że wariancje

na poziomie istotności 0.05 nie są równe.

Przeprowadzimy test Studenta (Welcha) na poziomiomie istotności = 0.05, sprawdzający, czy średnia wielkość zmiennej Y jest istotnie większa u mężczyzn niż u kobiet. Niech μ_1 oznacza średnią wielkość zmiennej Y u mężczyzn, a μ_2 średnią wielkość zmiennej Y u kobiet. Mamy:

$$\begin{aligned} H_0 &: \mu_1 \geqslant \mu_2 \\ H_1 &: \mu_1 < \mu_2. \end{aligned} \tag{8}$$

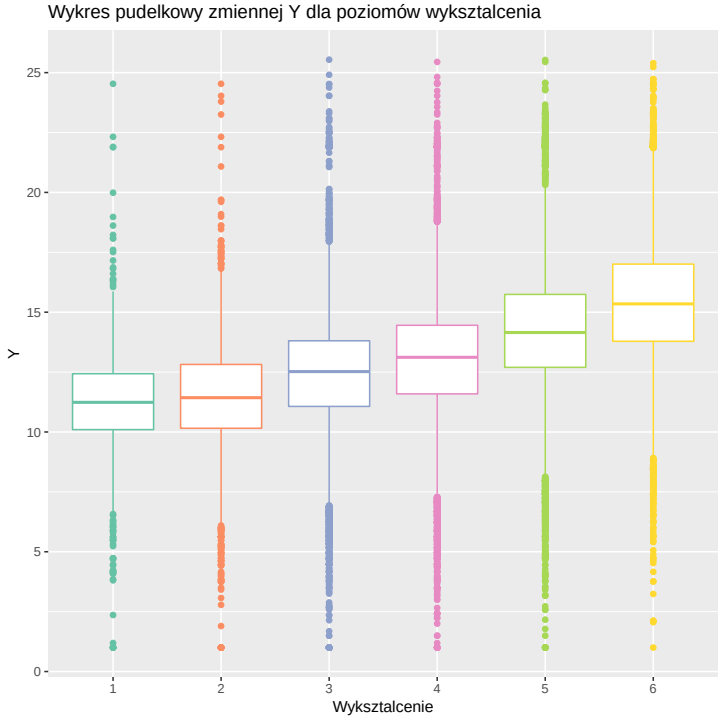
Wykonujemy test:

```
##
##  Welch Two Sample t-test
##
## data:  women$Y and men$Y
## t = -72.576, df = 55677, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.625103
## sample estimates:
## mean of x mean of y
##  12.22203  13.88481
```

P–wartość jest znacząco mniejsza od poziomu ufności, zatem odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Możemy powiedzieć, że na poziomie ufności 0.05 wartość zmiennej Y dla mężczyzn jest znacząco większa niż dla kobiet.

3.4 Wykres pudełkowy zmiennej Y dla różnych poziomów wykształcenia

Konstruujemy wykres pudełkowy zmiennej Y dla sześciu poziomów wykształcenia zawartych w danych.



Mediana wyznaczona przez poziomy odcinek w każdym z pudełek jest wyższa wraz ze wzrostem wykształcenia. IQR jest dość podobne, aczkolwiek najmniejsza wysokość wykresu pudełkowego jest dla wykształcenia podstawowego. Może być to związane z najmniejszą licznością grupy osób posiadających wykształcenie podstawowe.

3.5 Test Studenta

Przeprowadzamy test do sprawdzenia, czy średnia wielkość zmiennej Y u osób z magisterium jest istotnie większa niż u osób z licencjatem. IQR jest bardzo podobne dla obydwu grup. Mamy 10 991 zmiennych dla osób z licencjatem oraz 5601 dla osób z magisterium, zatem na mocy Centralnego Twierdzenia Granicznego przyjmiemy, że założenie o normalności danych jest spełnione. Niech μ_1 oznacza średnią wartość Y dla osób z licencjatem, a μ_2 średnią wartość Y dla osób z magisterium. Na poziomie ufności $\alpha = 0.05$ tetsujemy:

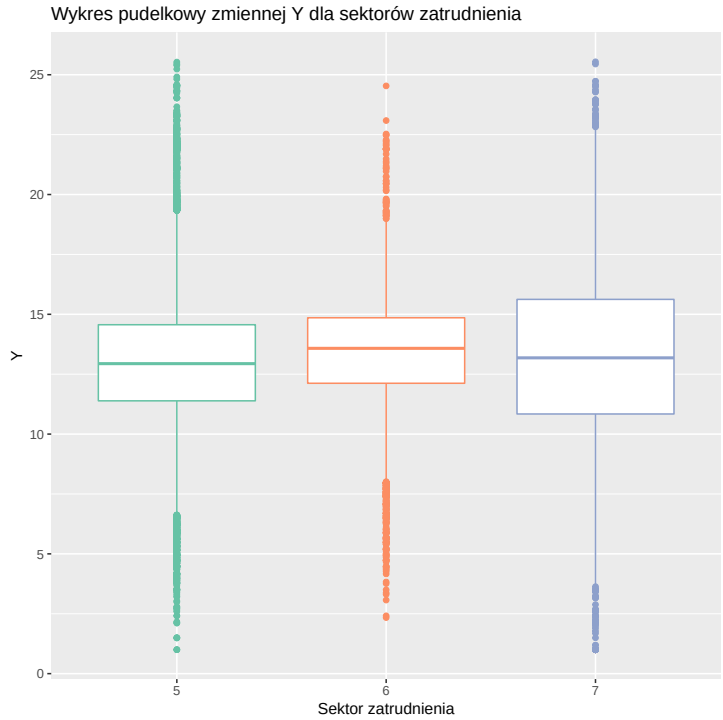
$$\begin{aligned} H_0 &: \mu_1 \geqslant \mu_2 \\ H_1 &: \mu_1 < \mu_2. \end{aligned} \tag{9}$$

```
##
##  Welch Two Sample t-test
##
## data:  lic$Y and mgr$Y
## t = -27.079, df = 10611, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.234471
## sample estimates:
## mean of x mean of y
##  14.16716  15.48147
```

P–wartość jest znacząco mniejsza od poziomu istotności, zatem mamy podstawy do odrzucenia H_0 na rzecz H_1. Na poziomie istotności 0.05 możemy stwierdzić, że osoby z magisterium statystycznie zarabiają więcej od osób z licencjatem.

3.6 Wykres pudełkowy zmiennej Y dla różnych sektorów zatrudnienia

Konstruujemy wykres pudełkowy dla porównania wartości zmiennej Y w trzech różnych sektorach zatrudnienia.



Na wykresie widzimy diagramy pudełkowe dla sektora publicznego (5), prywatnego (6) i samozatrudnienia (7). Najwyższa mediana jak i najmniejsze IQR zachodzi dla sektora prywatnego. Sektor publiczny ma trochę większe IQR i odrobinę niższą medianę, natomiast sektor samozatrudnienia ma zdecydowanie największe IQR i medianę podobną do sektora publicznego.

3.7 Test Studenta

Zastosujemy test Studenta do udzielenia odpowiedzi na pytanie czy istnieje różnica między średnią wartością zmiennej Y dla osób zatrudnionych w sektorze publicznym i prywatnym. Niech μ_1 oznacza średnią wartość Y w sektorze publicznym, a μ_2 średnią wartość Y dla sektora prywatnego. Na poziomie istotności $\alpha = 0.5$ mamy:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned} \tag{10}$$

Wykonujemy test.

```
##
##  Welch Two Sample t-test
##
## data:  public$Y and private$Y
## t = -12.754, df = 15705, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3091918
## sample estimates:
## mean of x mean of y
##  13.01557  13.37054
```

Wykonany test jest konkretniej testem Welcha. P–wartość jest znacząco mniejsza od poziomu istotności, zatem mamy podstawy, aby odrzucić hipotezę zerową na rzecz alternatywnej i stwierdzić (na poziomie istotności $\alpha = 0.05$), że istnieje różnica między średnią wartością zmiennej Y dla osób zatrudnionych w sektorze publicznym i prywatnym.

3.8 Podsumowanie i wnioski

W zadaniu przeprowadziliśmy transformację zmiennej ilościowej, która pozwoliła nam na wykonanie różnych testów Studenta (i Welcha) dla danych dotyczących zarobków. Mogliśmy statystycznie „potwierdzić" przypuszczenia z poprzednich raportów, dotyczące przykładowo, Gender Pay Gap i różnicy w zarobkach zależnej od wykształcenia.

Literatura

[1] P-wartość.

[2] Założenia testu t-Studenta.

[3] Markery ryzyka zawałowego.

[4] Test t-Welcha.

[5] Dokument z NCSS Statistical Software z założeniami dotyczącymi testu t-Studenta.

[6] F-test w R.