Projekt ze Statystycznych Metod Opracowania Danych II, analiza danych w języku R

Kamil Kalinowski

26 stycznia 2022

```
library(readr,nortest)
Sys.which("pdflatex")
```

```
##                                                                               pdflatex
## "C:\\Users\\kalin\\AppData\\Local\\Programs\\MiKTeX\\miktex\\bin\\x64\\pdflatex.exe"
```

Raport został przygotowany przy użyciu notatnika R Markdown z programu R Studio.

Zadanie 1. Gęstości asteroid

W tej sekcji zostanie poddana analizie próbka danych empirycznych. Są to gęstości różnych asteroid wraz z niepewnościami.

Podpunkt A

Poniżej przedstawiono podstawowe statystyki opisujące dane. Pod tabelką wyświetlono kolejno odchylenie standardowe pomiaru gęstości, odchylenie standardowe średniej gęstości, odchylenie standardowe niepewności gęstości, błąd mediany (średnie odchylenie bezwzględne) gęstości i błąd mediany niepewności.

```
Asteroids <- read.table("asteroid_dens.dat.txt", header = T) #wczytaj dane
summary(Asteroids) #wyświetl podstawowe statystyki
```

```
##    Asteroid              Dens            Err
##  Length:26          Min.   :0.800   Min.   :0.0300
##  Class :character   1st Qu.:1.343   1st Qu.:0.1350
##  Mode  :character   Median :2.060   Median :0.3000
##                     Mean   :2.182   Mean   :0.6073
##                     3rd Qu.:2.700   3rd Qu.:0.7500
##                     Max.   :4.900   Max.   :3.9000
```

```
S2Dens <- sd(Asteroids$Dens) #odchylenie standardowe pomiaru
S2AvrgDens <-S2Dens/sqrt(26) #odchylenie standardowe średniej
S2Err <- sd(Asteroids$Err)
print(S2AvrgDens)
```

```
## [1] 0.2053115
```

```
print(S2Dens)
```

```
## [1] 1.046888
```

```
print(S2Err)
```

```
## [1] 0.8179612
```

```
DensMedianError <- mad(Asteroids$Dens) #błąd mediany (średnie odchylenie bezwzględne)
ErrMedianError <- mad(Asteroids$Err)
print(DensMedianError)
```

```
## [1] 0.971103
```

```
print(ErrMedianError)
```

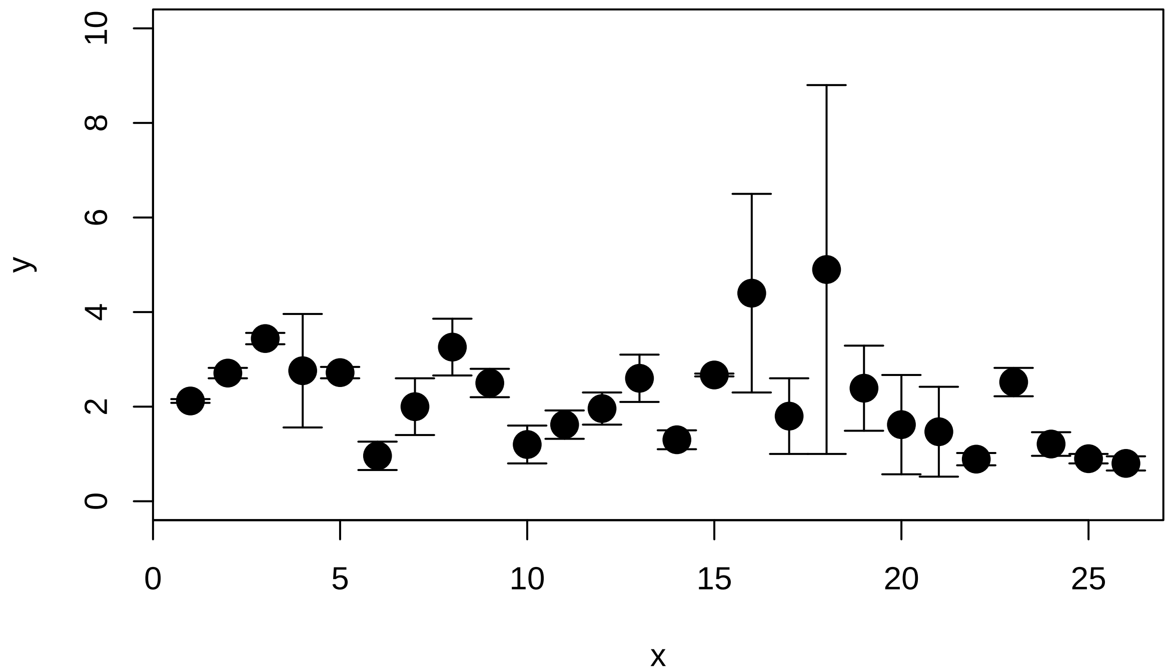
```
## [1] 0.289107
```

```
y <- Asteroids$Dens # Zmienna z gęstościami
y.sd <- Asteroids$Err # Zmienna z błędami
x <- 1:26 #Lista z kolejnymi liczbami naturalnymi od 1 do 26
```

Podpunkt B

Poniżej przedstawiono na wykresie wszystkie dane. W osi pionowej są to gęstości asteroid i ich niepewności, a w osi poziomej liczby porządkowe odpowiadające asteroidom. Wykorzystano taką formę wykresu, ponieważ próbka jest wystarczająco mała, aby wykres tego typu był czytelny. Taka forma nie pozwala na przeoczenie wyróżniających pojedyncze asteroidy cech, które można przeczyć na wykresach przedstawiających statystyki próbki.

```
plot(x, y, ylim=c(0, 10), xlab="x", ylab="y", pch=16, cex=2) #utwórz wykres
# wyświetl niepewności
arrows(x0=x, y0=y-y.sd, x1=x, y1=y+y.sd, code=3, angle=90, length=0.1)
```



Użyto funkcji plot, ponieważ wyświetla ona wykresy, natomiast powodem użycia funkcji arrows była chęć naniesienia niepewności na wykres.

Podpunkt C

Następnie sprawdzono za pomocą testów Shapiro-Wilka i Kołmogorova-Smirnova, czy dane można opisać rozkładem naturalnym.

Test Shapiro-Wilka

Jest on jednym z testów służących do sprawdzania, czy próbka pochodzi z populacji o rozkładzie normalnym. Hipoteza zerowa w teście Shapiro-Wilka to pochodzenie próby z populacji o rozkładzie normalnym.

Statystyką testową jest

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

gdzie x_i jest i-tą najmniejszą liczbą w próbce, a $\overline{x}$ jest średnią z próbki. Współczynniki a_i są dane przez wzór

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{C},$$

gdzie C jest normą wektora, daną wzorem

$$C = ||m^T V^{-1}|| = \sqrt{m^T V^{-1} V^{-1} m}$$

i $m = (m_1, \dots, m_n)^T$ zawiera obserwacje w kolejności niemalejącej.

Test Kołmogorova-Smirnova

Test ten jest kolejnym popularnym testem normalności. Hipotezą zerową jest normalność populacji, z jakiej pochodzi próbka.

Statystyką testową jest

$$D_n = \sup_x |F_0(x) - S_n(x)|,$$

gdzie $s_n(x)$ jest dystrybuantą empiryczną opisaną wzorem

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leqslant x},$$

gdzie X_i jest wartością i-tej obserwacji. $I_{X_i \leqslant x}$ jest funkcją charakterystyczną zbioru przyjmująca wartość 1, gdy $X_i \leqslant x$. W przeciwnym wypadku przyjmuje ona wartość 0.

```
library(nortest) #użyj biblioteki

AD<-Asteroids$Dens
AR<-Asteroids$Err

lillie.test(AD) #Wykonaj test KS
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  AD
## D = 0.13644, p-value = 0.2435
```

```
shapiro.test(AD) #Wykonaj test shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  AD
## W = 0.93021, p-value = 0.07841
```

```
lillie.test(AR)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  AR
## D = 0.24016, p-value = 0.0004775
```

```
shapiro.test(AR)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  AR
## W = 0.64214, p-value = 9.4e-07
```

Wyniki obu testów są zgodne: Przyjmując poziom istotności $\alpha = 0.05$, nie można odrzucić hipotez zerowych dla gęstości asteroid, ale należy je odrzucić dla niepewności tych gęstości.

Podpunkt C

Zbadano również, czy próbka zawierała dane powstałe w wyniku popełnienia błędu grubego. Implementacja w R testów przeprowadzonych w tym celu pochodzi z pakietu outliers.

Test Dixona

Jest jednym z testów służących do sprawdzenia, czy próbka zawiera dane powstałe w wyniku popełnienia błędu grubego.

Statystyką testową jest $Q = \frac{\text{gap}}{\text{range}}$, gdzie gap jest modułem z różnicy pomiędzą wartością podejrzanego pomiaru, a wartością pomiaru o najbliższej wartości, natomiast range jest różnicą pomiędzy największą wartości z próbki, a najmniejszą wartościa z próbki.

Test Grubbsa

Jest to kolejny test służący do sprawdzenia, czy próbka zawiera dane powstałe w wyniku popełnienia błędu grubego.

Statystyką testową jest

$$G = \frac{max|X_i - \overline{X}|}{\sigma}, \tag{1}$$

gdzie $\overline{X}$ to średnia, a σ jest odchyleniem standardowym. Statystyką Grubbsa jest największe odchylenie od średniej w zbiorze o rozkładzie normalnym.

```
library(outliers)
dixon.test(AD)
```

```
##
##  Dixon test for outliers
##
## data:  AD
## Q = 0.365, p-value = 0.1709
## alternative hypothesis: highest value 4.9 is an outlier
```

```
grubbs.test(AD)
```

```
##
##  Grubbs test for one outlier
##
## data:  AD
## G = 2.5967, U = 0.7195, p-value = 0.07011
## alternative hypothesis: highest value 4.9 is an outlier
```

Wyniki obu testów są zgodne: Przyjmując poziom istotności $\alpha = 0.05$, nie można odrzucić hipotez zerowych. Istotność statystyczna wyniku testu Grubbsa jest jednak niska (p-value jest bliske α).

Zadanie 2. Jasności gromad kulistych

W tej części raportu analizie poddano rozkłady jasności gromad kulistych w filtrze K z Drogi Mlecznej i z galaktyki M31.

Podpunkt A

Poniżej przedstawiono podstawowe statystyki opisujące dane.

```
Galaxies <- read.table("GlobClus_M31.dat.txt", header = T)
Galaxies2 <- read.table("GlobClus_MWG.dat.txt", header = T)
summary(Galaxies)
```

```
##     M31_GC                K        
##  Length:360         Min.   :10.75  
##  Class :character   1st Qu.:13.85  
##  Mode  :character   Median :14.54  
##                     Mean   :14.46  
##                     3rd Qu.:15.33  
##                     Max.   :18.05  
```

```
summary(Galaxies2)
```

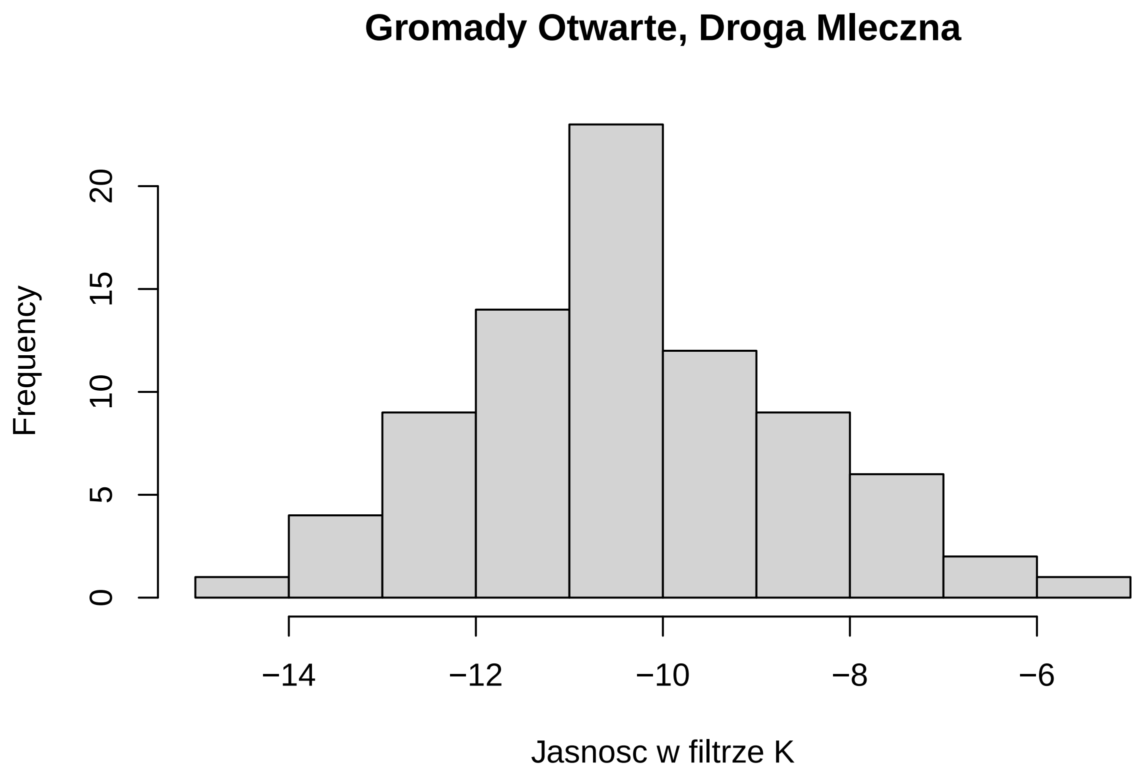
```
##     MWG_GC                K          
##  Length:81          Min.   :-14.205  
##  Class :character   1st Qu.:-11.478  
##  Mode  :character   Median :-10.557  
##                     Mean   :-10.324  
##                     3rd Qu.: -9.199  
##                     Max.   : -5.140  
```

```
y2 <- Galaxies$K
y3 <- Galaxies2$K
```

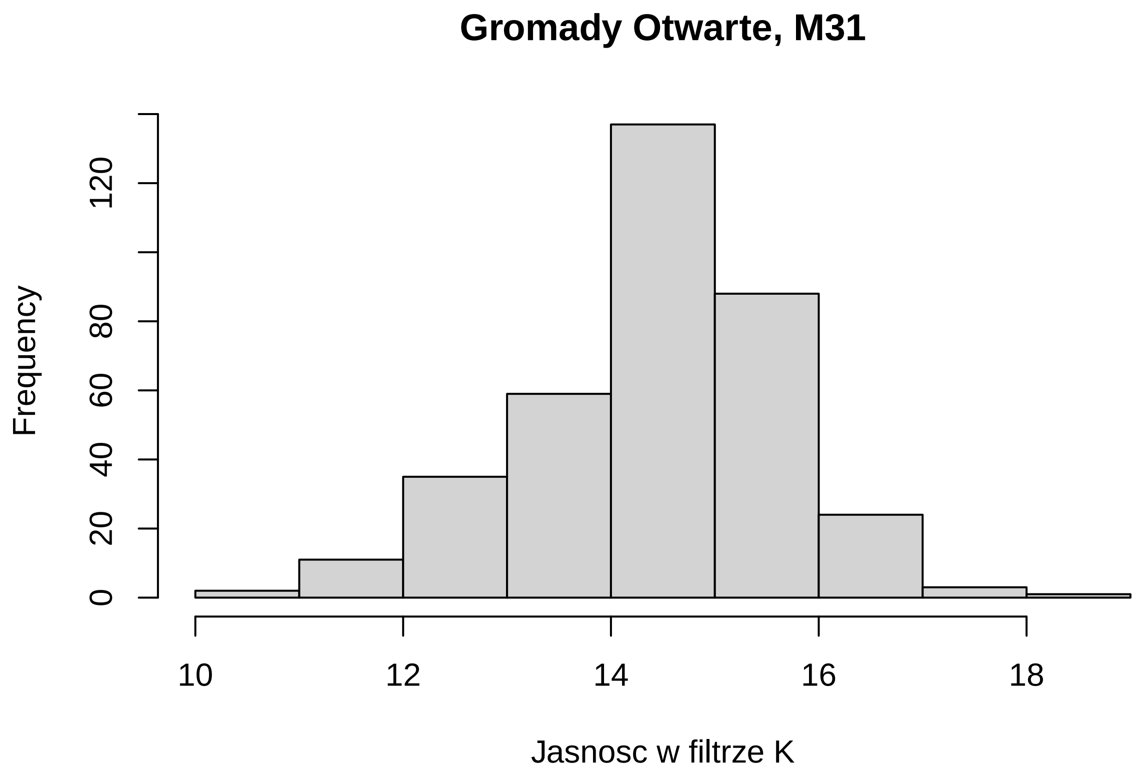
Podpunkt B

Poniżej przedstawiono histogramy dla obu zestawów danych. Wybrano ten typ wykresów, ponieważ pozwala on na zorientowanie się, jakie mniej więcej są cechy obu populacji, takie jak symetryczność rozkładu, wartość oczekiwana, skośność czy kurtoza. Wyświetlono również obie populacje na wykresie pudełkowym, ponieważ uwydania on różnicę w położeniu rozkładów (wartości oczekiwanej), a także ich kształt i rozproszenie. Wniosek z analizy tego wykresu jest taki, że w celu porównania tych próbek warto przesunąć je względem siebie. Zrobiono to w kolejnym podpunkcie.

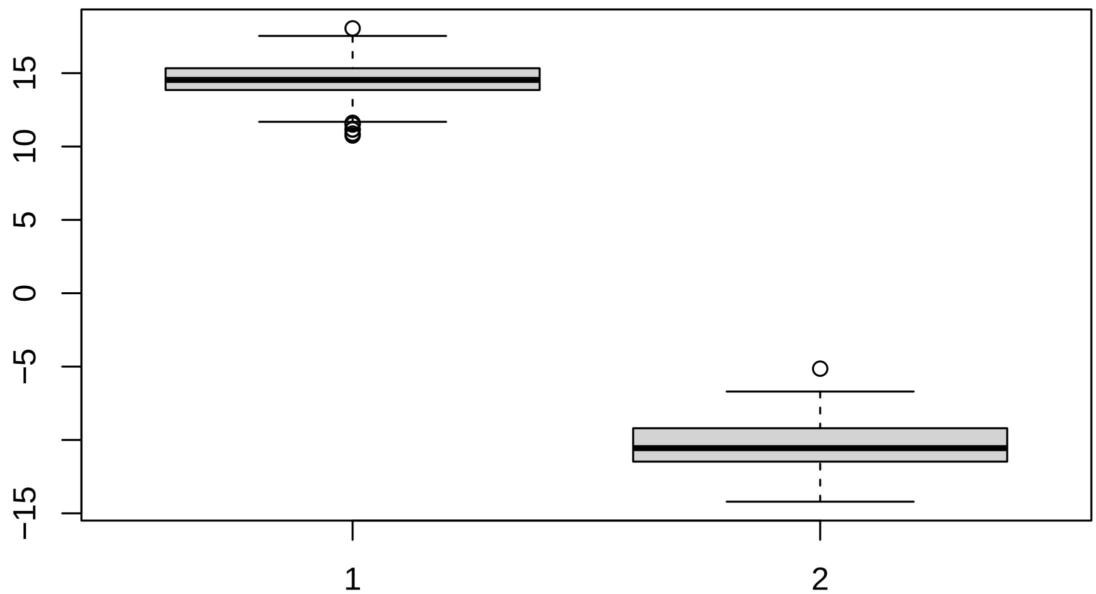
```
#wyświetl histogram
hist(y3, main="Gromady Otwarte, Droga Mleczna",xlab="Jasność w filtrze K")
```




```
hist(y2, main="Gromady Otwarte, M31",xlab="Jasność w filtrze K")
```




```
boxplot(y2,y3) #wyświetl wykres pudełkowy
```



Podpunkt C

Obliczono różnicę median jasności gromad pomiędzy galaktykami.

```
clusters_diff <- median(y3)-median(y2) #oblicz różnicę median
print(clusters_diff)
```

```
## [1] -25.0965
```

Wynik porównano z wynikiem testu rang Wilcoxona.

```
y2shifted=y2+clusters_diff #przesuń jasności gromad z M31
wilcox.test(y3,y2shifted)
```

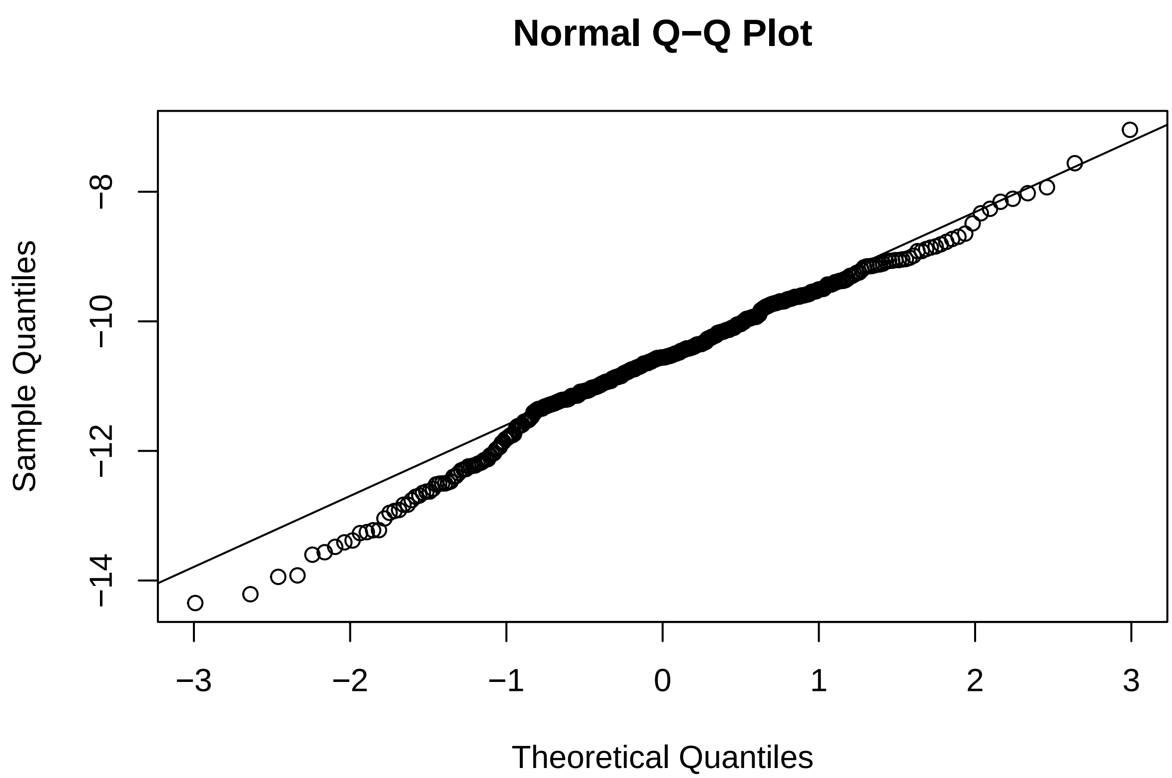
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  y3 and y2shifted
## W = 15669, p-value = 0.2936
## alternative hypothesis: true location shift is not equal to 0
```

Dla poziomu istotności $\alpha = 0.05$ nie można odrzucić hipotezy zerowej, że mediany rozkładów jasności z populacji obu galakyk po dodaniu ich różnicy do jasności z M31 są równe.

Podpunkt D

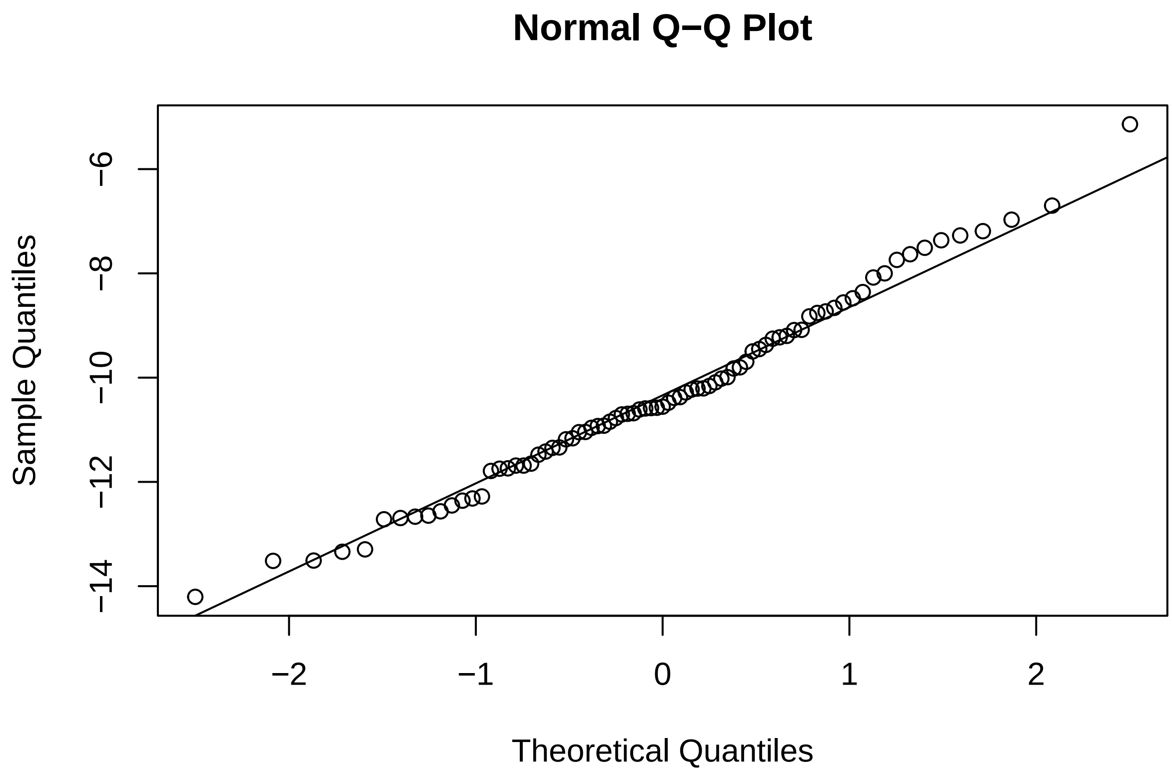
Wykres kwantyl-kwantyl (QQ) powstaje w 2-wymiarowej przestrzeni zmiennych losowych. Każdy punkt na wykresie powstaje dla wartości zmiennych losowych odpowiadających tym samym kwantylom ich rozkładu. Służy on do graficznego porównania dwóch rozkładów. Jeśli wszystkie punkty na wykresie są zawarte w prostej, zmienne losowe są opisane tym samym rozkładem.

```
qqnorm(y2shifted) #Wyświetl wykres QQ
qqline(y2shifted)
```



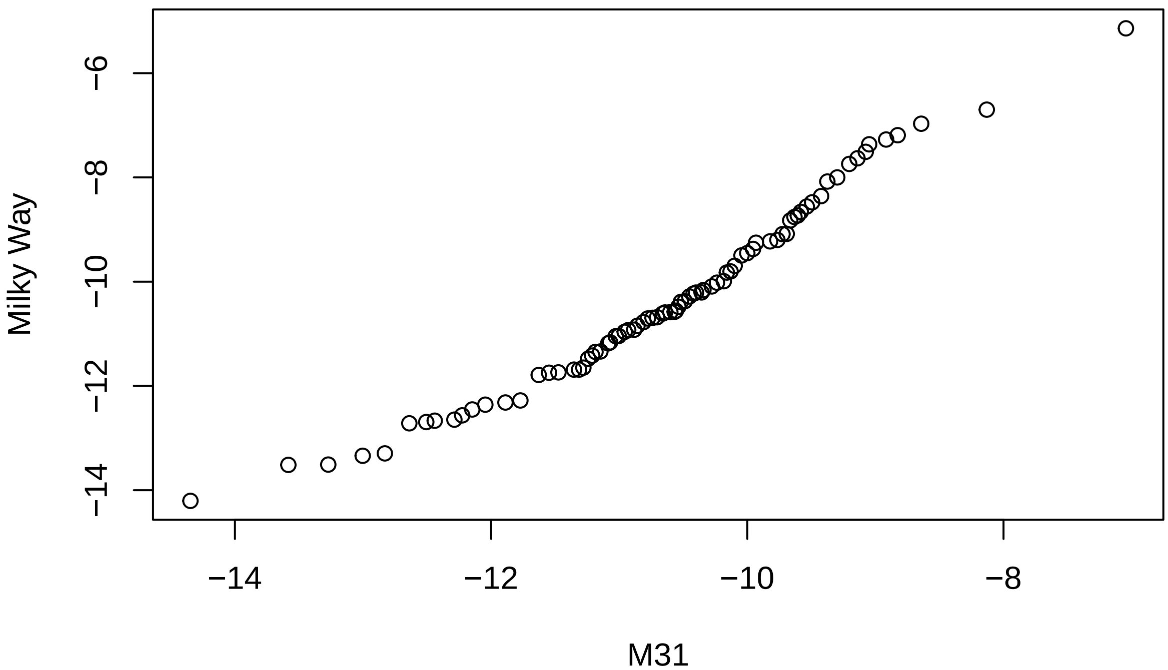
W powyższym przypadku na osi rzędnych przedstawiono empiryczną dystrybucję jasności gromad w M31 przesuniętą o różnicę median jasności gromad pomiędzy galaktykami, a na osi odciętych teoretyczną dystrybucję rozkładu normalnego.

```
qqnorm(y3)
qqline(y3)
```



W powyższym przypadku na osi rzędnych przedstawiono empiryczną dystrybucję jasności gromad w Drodze Mlecznej, a na osi odciętych teoretyczną dystrybucję rozkładu normalnego.

```
qqplot(y2shifted,y3, xlab = "M31", ylab = "Milky Way")
```



W powyższym przypadku na osiach są empiryczne dystrybucje jasności gromad w Drodze Mlecznej i w M31.

Test Kołmogorova-Smirnova dla dwóch prób

Jest to jeden z testów służących do sprawdzenia, rozkłady dwóch zmiennych losowych różnią się od siebie. Jest on wrażliwy na różnice w położeniu i kształcie dystrybuanty rozkładów. Można porównywać rozkład empiryczny z rozkładem teoretycznym (1. wariant), lub 2 rozkłady empiryczne (2. wariant).

Statystyką testową jest

$$D_{n,n'} = \sup |F_n(x) - F_{n'}(x)|,$$

a hipoteza zerowa (o nie różniących się rozkładach) jest odrzucana na poziomie α, gdy

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha,$$

gdzie K jest rozkładem Kołmogorowa.

```
ks.test(y2shifted, y3)
```

```
## Warning in ks.test(y2shifted, y3): p-value will be approximate in the presence
## of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
```

```
## data:  y2shifted and y3
## D = 0.18333, p-value = 0.02348
## alternative hypothesis: two-sided
```

Przyjmując poziom istotności $\alpha = 0.05$, należy odrzucić hipotezę zerową, że te rozkłady różnią się od siebie.

Zinterpretować ten wynik można w ten sposób, że gromady kuliste w obu galaktych powstały w podobnych warunkach, poprzez te same procesy fizyczne.

Zadanie 3. Styl życia studentów

W tej sekcji analizie poddano wyniki ankiety przeprowadzonej wśród studentów. Zapytano ich m. in. o płeć, wiek, o częstotliwość palenia i wykonywania ćwiczeń fizycznych.

Podpunkt A

Obliczono wartość estymatora punktowego dla średniej wzrostu respondentów.

```
library(MASS)
Wzrost <- survey$Height
SrWzrost<-mean(Wzrost, na.rm = TRUE)
print(SrWzrost)
```

```
## [1] 172.3809
```

Podpunkt B

Wykorzystując wzór wyprowadzony na ćwiczeniach, obliczono przedział ufności 95% i odpowiadający mu błąd. Poniższe wartości to kolejno lewostronna granica przedziału ufności, prawostronna granica przedziału ufności i błąd.

```
alpha <- 0.05 #przedział ufności
n <- 212 #liczba pomiarów
t <- qt(1 - alpha / 2, n - 1) #kwantyl rozkładu studenta
S2 <- sd(Wzrost, na.rm = TRUE) #odchylenie standardowe

lboundary <- SrWzrost - t * S2 / sqrt(n) #lewa granica
rboundary <- SrWzrost + t * S2 / sqrt(n) #prawa granica
print(lboundary)
```

```
## [1] 171.0476
```

```
print(rboundary)
```

```
## [1] 173.7141
```

```
error=(rboundary-lboundary)/2 #błąd
print(error)
```

```
## [1] 1.333231
```

Podpunkt C

wykorzystując test niezależności χ^2 sprawdzono, czy fakt, że studenci palą, jest niezależny od ich poziomu aktywności fizycznej.

```
Exercises=survey$Exer
Smoking=survey$Smoke
table(Exercises,Smoking) #wyświetl tabelę zbiorczą
```

```
##          Smoking
## Exercises Heavy Never Occas Regul
##      Freq     7    87    12     9
##      None     1    18     3     1
##      Some     3    84     4     7
```

```
chisq.test(Exercises,Smoking)
```

```
## Warning in chisq.test(Exercises, Smoking): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Exercises and Smoking
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

Na poziomie istotności $\alpha = 0.05$ odrzucono hipotezę zerową, że częstość palenia nie jest skorelowana z częstością wykonywania ćwiczeń fizycznych.

Zadanie 4. Erupcje wulkanów

W tej części zostaną przeanalizowane dane dotyczące wulkanów. Zestaw danych zawiera czas oczekiwania i odpowiadającą mu liczbę erupcji.

Podpunkt A

Do danych dopasowano prostą za pomocą regresji klasycznej. Poniżej wypisano parametry dopasowania oraz wyświetlone dane i dopasowaną prostą na wykresie.

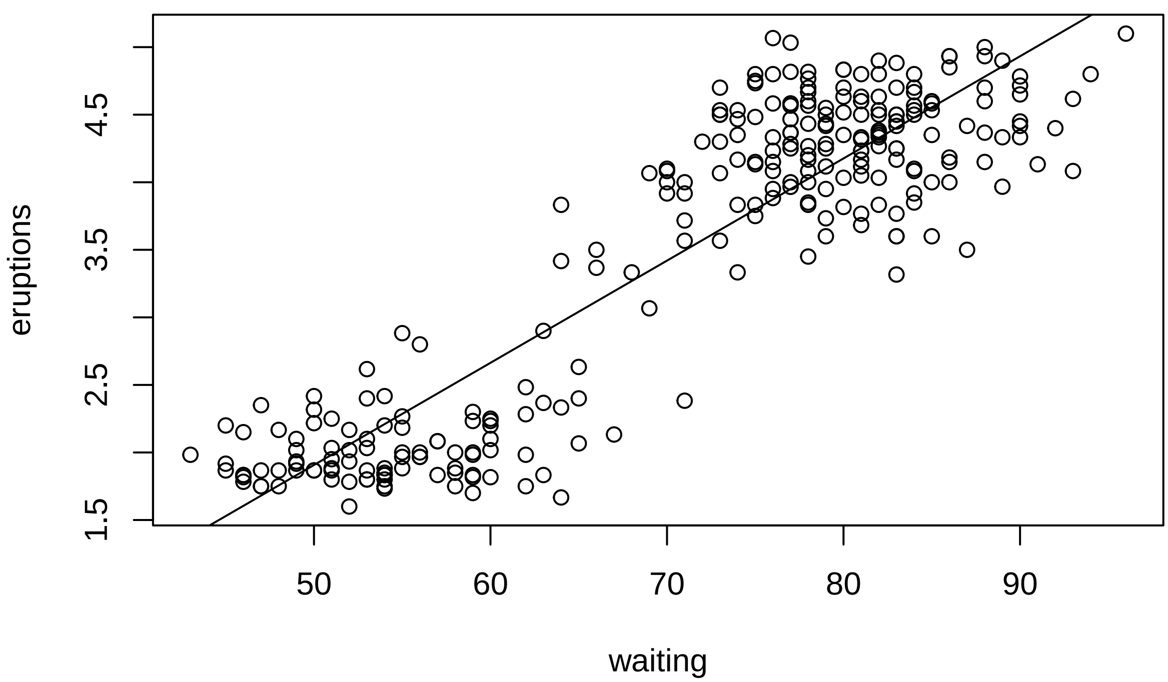
```
regression <- lm(eruptions~waiting,data=faithful) #wykonaj regresję klasyczną
summary(regression)
```

```
##
## Call:
## lm(formula = eruptions ~ waiting, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

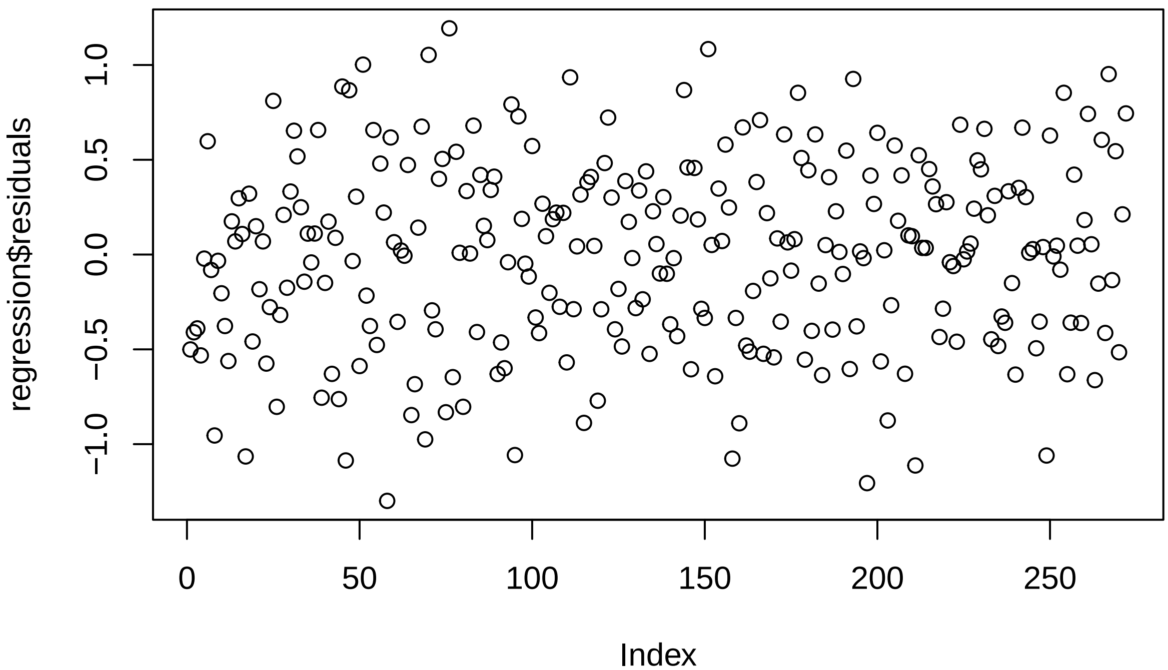
```r
waiting<-faithful$waiting
eruptions<-faithful$eruptions
```

```r
plot(waiting,eruptions)
abline(regression) #wyświetl na wykresie prostą regresji
```



Podpunkt B

```r
plot(regression$residuals) #wyświetl residua
```



Na powyższym wykresie wyświetlono residua. Sprawiają one wrażenie rozmieszczonych losowo, zatem model liniowy wydaje się być właściwym do opisu badanej zależności. Zostanie to zweryfikowane w sposób ścisły w kolejnym podpunkcie.

Podpunkt C

Wyliczono kryteria informacyjne Akaike oraz Bayessowskie. Te wartości są estymatorami błędu i zgodności próbek z modelem. Służą one do wyboru najlepszych modeli opisujących określone dane. Niższa wartość kryterium oznacza lepszy model.

```
AIC(regression) #Oblicz kryterium Akaike
```

```
## [1] 395.0159
```

```
BIC(regression) #Oblicz kryterium Bayessowskie
```

```
## [1] 405.8333
```

Porówanie tych dwóch wartości jest bezcelowe. W celu znalezienia najlepszego modelu należy porównywać wartości tego samego kryterium dla różnych modeli.

Podpunkt D

Znaleziono 95% poziom ufności średniego czasu trwania erupcji dla czasu oczekiwania 80 min. Wyświetlono go poniżej.

```
temp <- data.frame(waiting=80) #zmienna z parametrem czekania 80 minut
prediction <- predict(regression, data.frame(waiting=80), interval = 'confidence') #oszacuj liczbę
#erupcji na podstawie modelu z regresji
print(prediction)
```

```
##        fit      lwr      upr
## 1 4.17622 4.104848 4.247592
```