Astronomical data project

Radosław Kluczewski

26.01.2022

Introduction to Asteroids data

In this short data project using R language Firstly it has been loaded data form AstroStatistics page. After that, the basic statistics are presented below:

```
##    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
```

The standard deviation σ has also been calculated and will be used in the following median standard error formula:

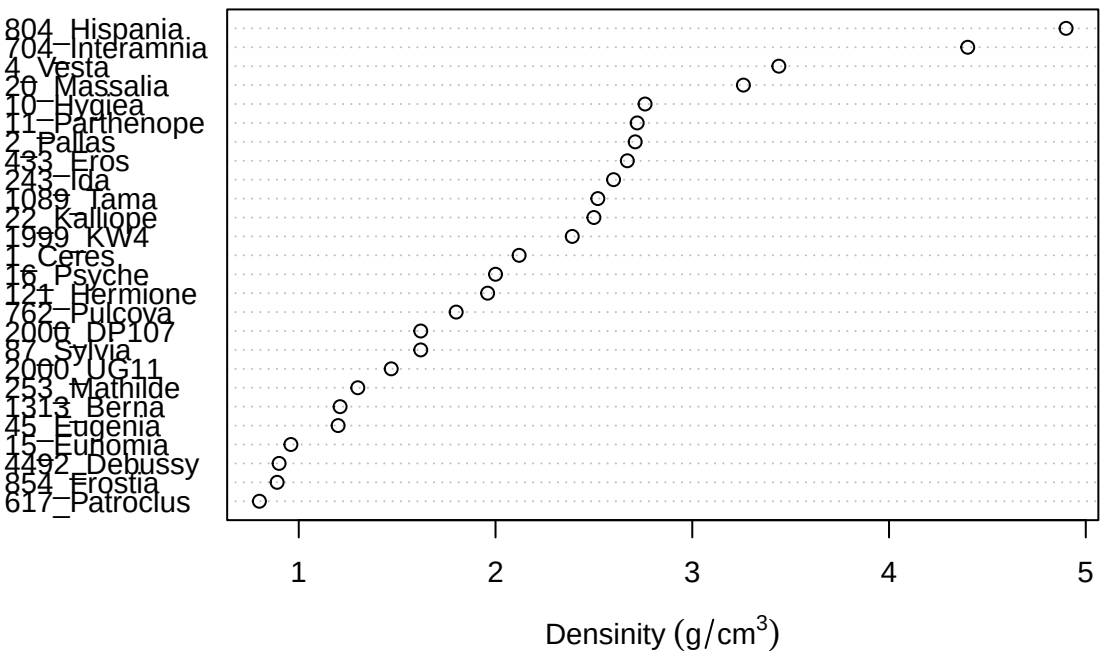
$$SE_{me} = \frac{1,253 \cdot \sigma}{\sqrt{N}}. \tag{1}$$

The results are presented below after substituting them to the above formula, where for the dens and err columns:

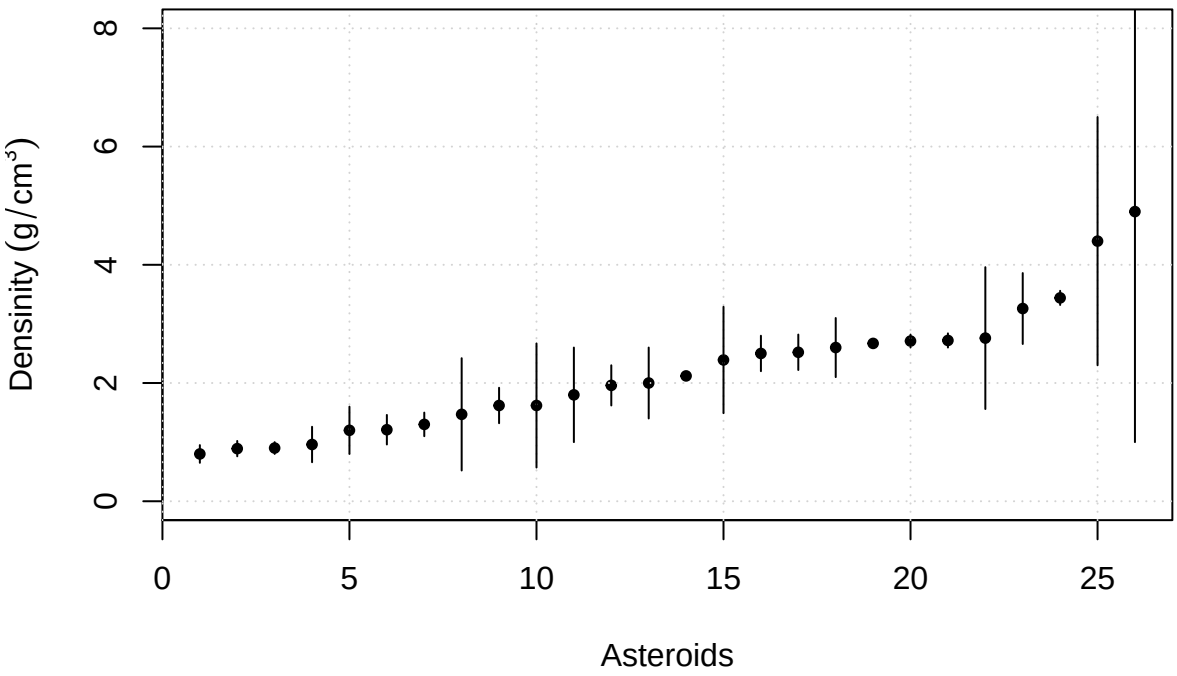
```
## [1] 0.2572554
```

```
## [1] 0.2010005
```

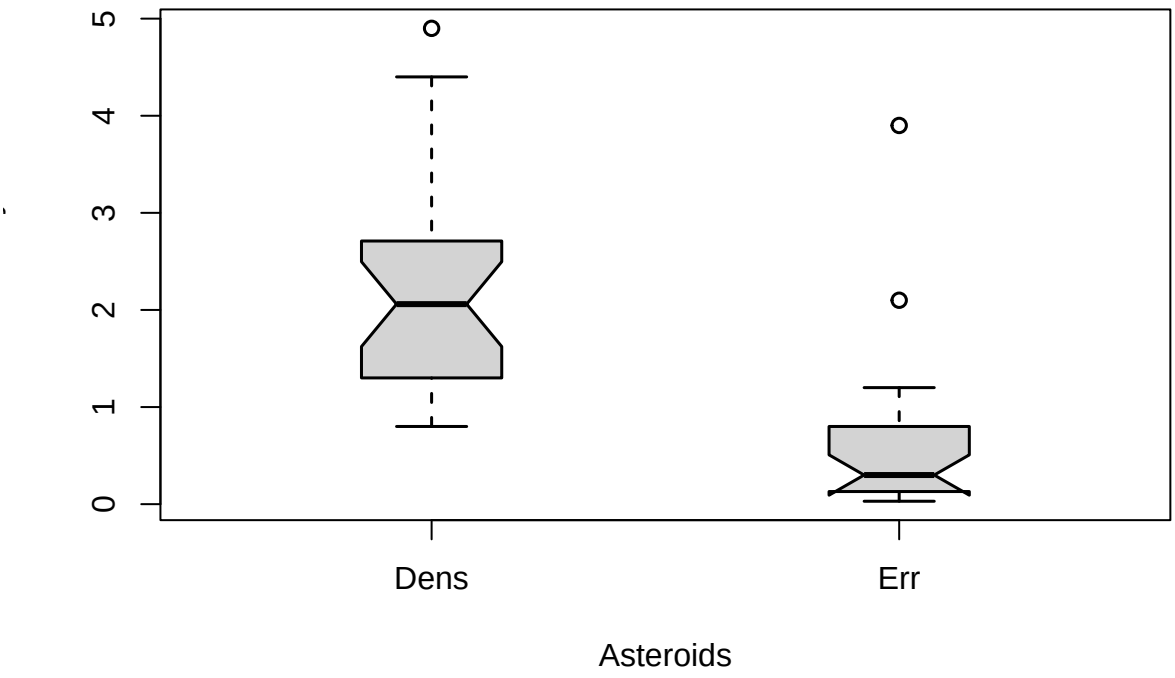
The figures below show the distribution of the data where errors are also included:



A drawing showing the density of individual asteroids. As can be seen, the highest density of data occurs around densities with values from 1 to 3. Very high densities were observed for the remaining data, which may raise suspicions. In order to check if the high-density data is true, the graph with errors was drawn, which is presented below:



The chart above shows very high uncertainties for large points densities. So, these points may not necessarily be so large density values. The last graph drawn is the comparison distribution of asteroid density to the distribution of errors in their measurements. This chart allows you to check whether the data is of similar density or not. In addition, it allows you to determine the size of the errors compared to the measurements densities and whether these errors are similar or outlier.



As we can read from the drawing, the data has a similar density, a errors are similar to each other.

To check whether the data can be described by a normal distribution two tests were decided on:

Kolmogorov-Smirnov test:

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

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

This test is one of the best-known collective tests on normality, where he tests the null hypothesis that the decomposition of our variable is close to normal. As a test statistic we accept:

$$D_n = \sup_x |F_0(x) - S_n(x)|, \tag{2}$$

where $s_n(x)$ is the empirical difference which is based on the sample order as follows:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leqslant x}, \tag{3}$$

where: X_i is the value of the X variable for the i-th observation. $I\_X_i \leqslant x$ is a characteristic function of the receiving set value one if $X_i \leqslant x$ and zero otherwise. Big the statistics values indicate a large discrepancy in the distribution function F_0 and S_n, so we construct a right-hand critical area. If the value of D_n falls into the critical area, we reject the null hypothesis on adopted significance level.

Assuming the significance level of $\alpha = 0.05$ and comparing go with the obtained p-value, we can say whether the hypothesis can be accept or reject. For density, the hypothesis is correct, because this value is much greater than the assumed value α, while for uncertainty we reject the hypothesis because value p-value is much smaller than α.

Shapiro-Wilk test:

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

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

This test verifies that the sample is from a population with a distribution normal. The test statistic is:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{4}$$

where: x_i is i - the smallest number in the sample, $\overline{x}$ is the sample mean, while \$ a_i \$ is expressed as pattern: $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C}$. C from the previous one of the formula is the norm of the vector, which is described by the formula: $C = ||m^T V^{-1}|| = \sqrt{m^T V^{-1} V^{-1} m}$, where $m = (m_1, \ldots, m_n)^T$ includes observations for non-decreasing values.

The null hypothesis is derived from samples from a normally distributed population. For the previously adopted significance level, we can say that We can assume that the hypothesis for the density of asteroids is correct, while the hypothesis for uncertainty must be rejected because it is much less p-value.

Based on the Dixon and Grubb tests, it was checked whether in data, there are outliers. Outliers it points where the measurement may be the result of a gross error. The first test is the Dixon test, the results of which were presented below:

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

The Dixon Test is a test to check whether a sample contains data resulting from committing a gross error. The test statistic is: $Q = \frac{\text{gap}}{\text{range}}$, where gap is the module from the difference between the suspect measurement and the value of the closest measurement. Range is the difference between the largest value of the sample and the smallest value of z samples.

Another test for detecting a gross error with an error is the test Grubbs, which consists in defining the H_0 : hypothesis of no deviation in the dataset and H_a : that is, is there a risk of a deviation in dataset. The test statistic is defined as:

$$G = \frac{max|X_i - \overline{X}|}{\sigma}, \tag{5}$$

where $\overline{X}$ - mean, σ - deviation standard. The Grubbs statistic is considered to be the greatest deviation from the mean in a normally distributed set. The test results are presented as follows:

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

Analyzing the test results, it can be concluded that we can not do any of the points considered a gross mistake. Especially the points that have the highest density.

Globular clusters

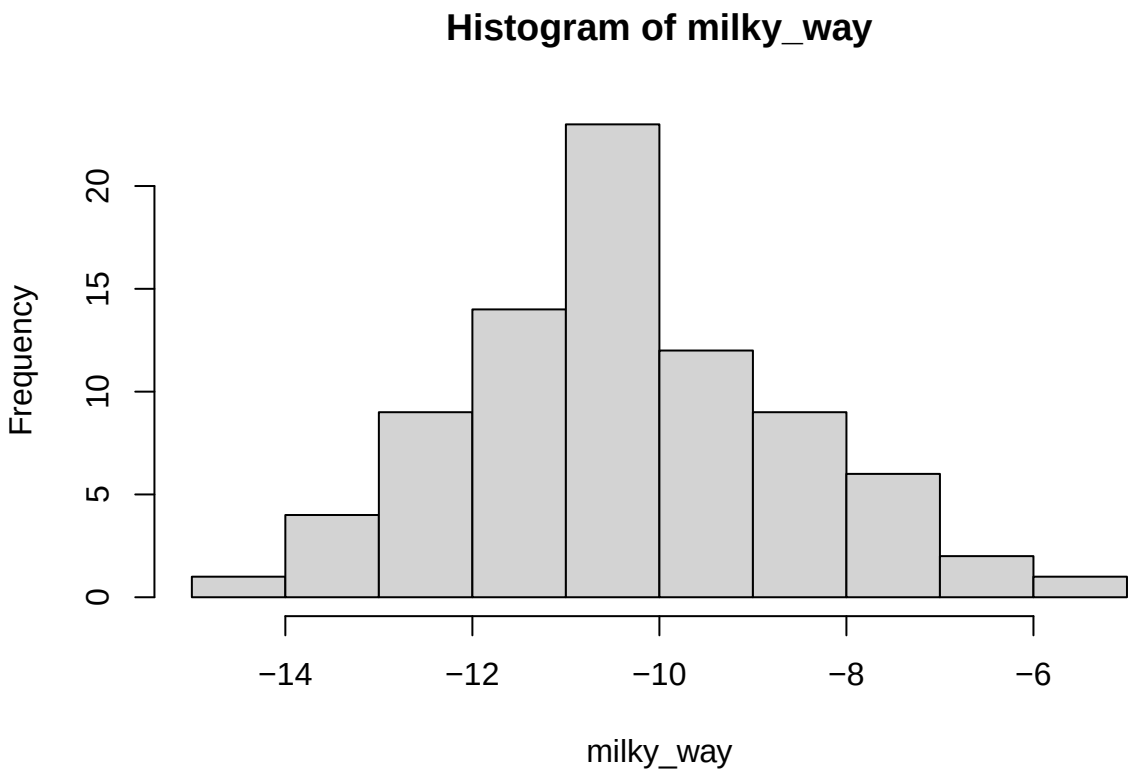
Below are the statistics for two datasets, where the first dataset is observations of globular clusters in the Milky Way:

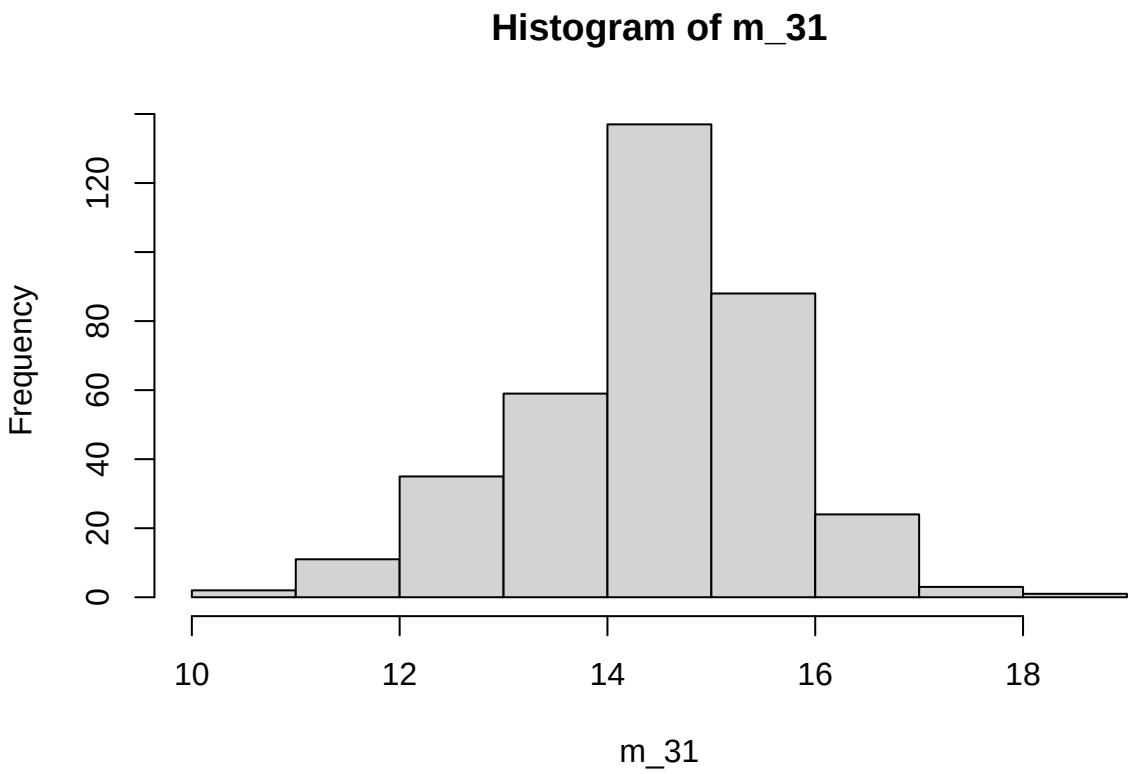
```
##     MWG_GC                K
##  Length:81          Min.   :-14.205
##  Class :character   1st Qu.:-11.478
##  Mode  :character   Median :-10.557
##                     Mean   :-10.324
##                     3rd Qu.: -9.199
##                     Max.   : -5.140
```

while the second is observations of the globular clusters in M31:

```
##     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
```

The data distributions for which the decision was made are presented below draw histograms for graphical representation the frequency of occurrence of brightness in a given range.

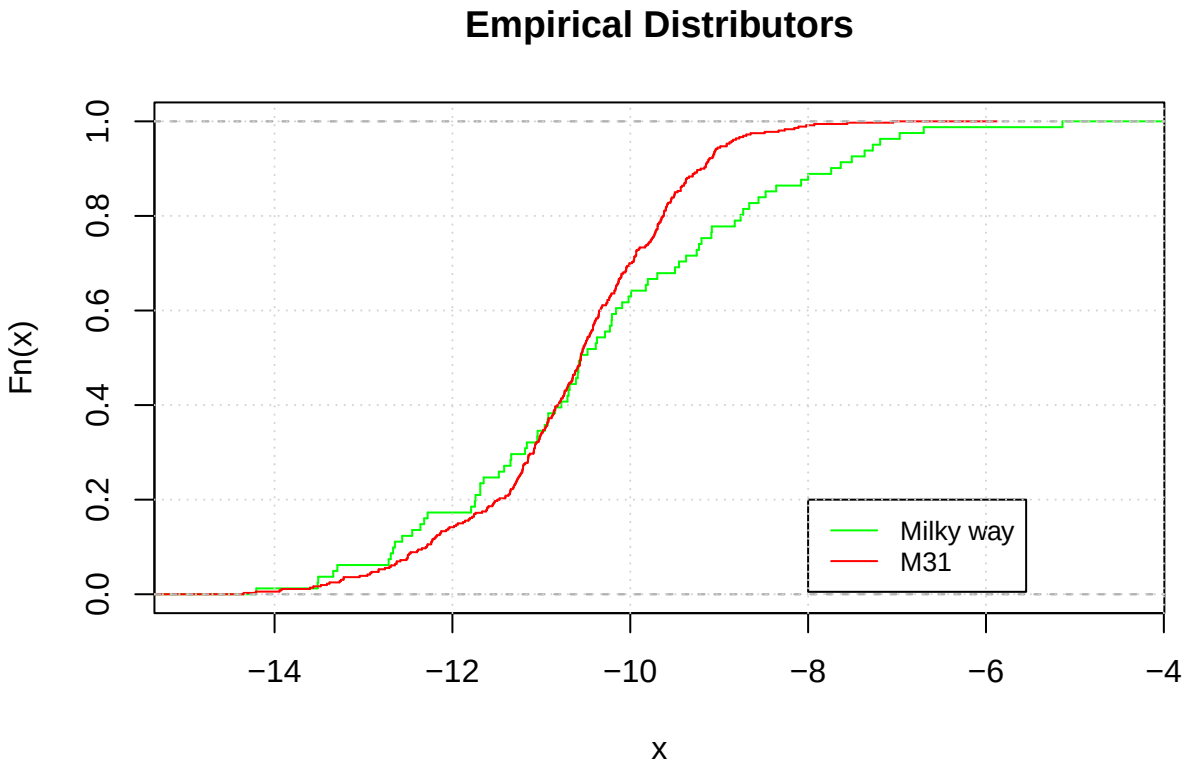

Histogram of milky_way
Frequency
0
5
10
15
20
−14
−12
−10
−8
−6
milky_way



It can be seen from the graphs that the brightness value is in both histograms completely different. If you want to present the distributions in one graph, you should take into account the correction by subtracting the difference of the medians of the Milky Way and M31. In for this purpose, the means were calculated and then subtracted from each other:

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

The result will be a similar brightness level to that observable in our galaxy. The results of the chart comparison remained presented below:



The result of the **module** value was compared with the result of the rank test Wilcoxon, which was calculated below:

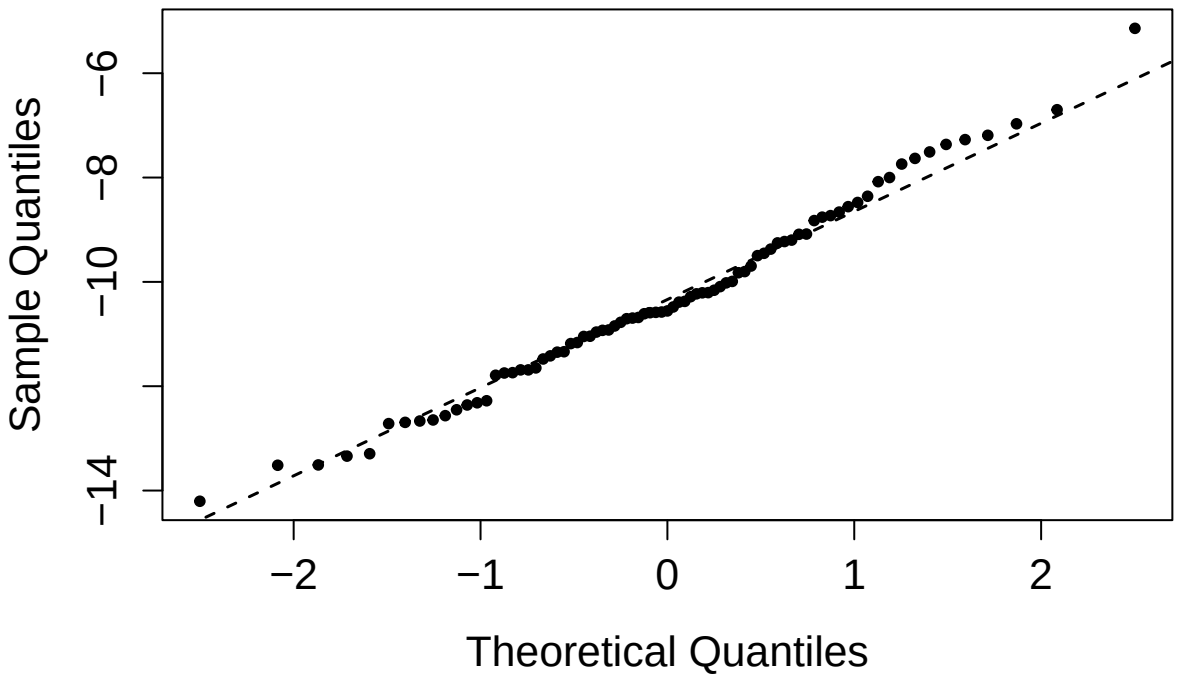
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  milky_way and m_31 + module
## W = 15669, p-value = 0.2936
## alternative hypothesis: true location shift is not equal to 0
```

As can be seen, the values determined by the median and the test are similar. Thus, it can be assumed that the median difference as a shift is satisfactorily sufficient..

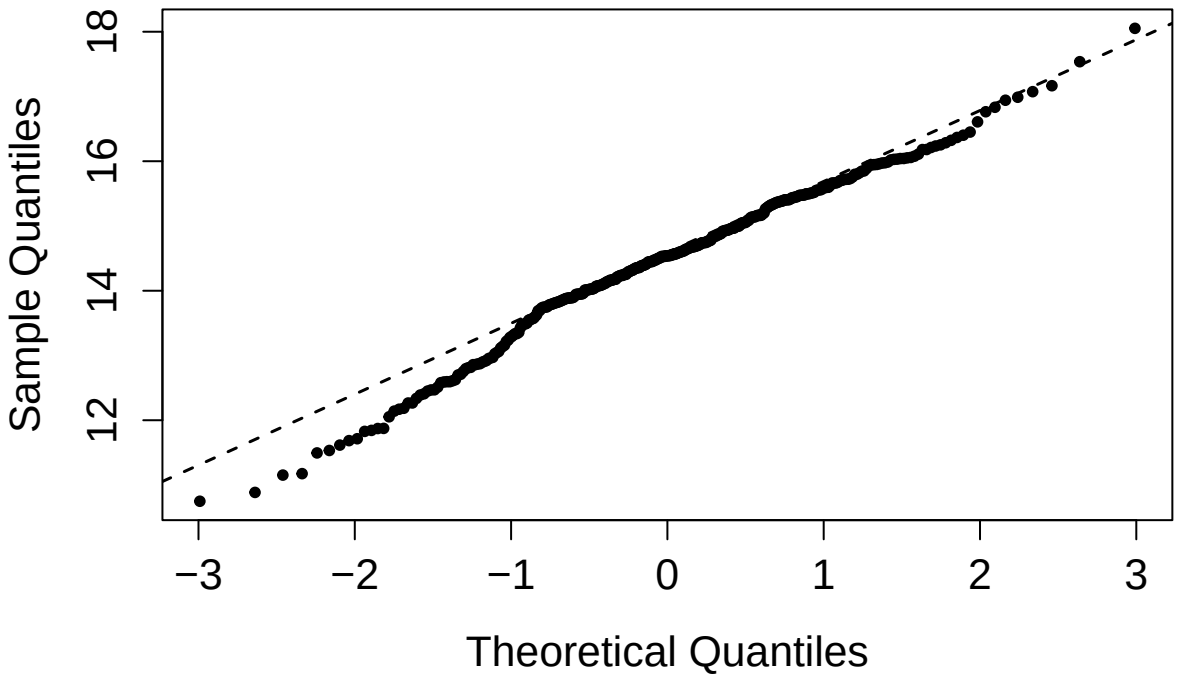
Q-Q - a quantile-quantile plot is used, among other things, to compare or the data distribution follows the model. Two-dimensional graph variables represents the points that arise on the plot as a result two corresponding quantile values of their random variables decomposition. it is worth noting that if all points correspond with the line straight line, the random variables in the plot are described by the same decomposition. The procedure for creating a chart is as follows as follows:

- ordering in ascending order residues, which are empirical quantiles,
- compute quantile orders,
- compute theoretical quantiles,
- sort residuo ascending for the second dataset,
- summary of quantiles in the plot.

The quantile graphs of the quantile are presented below:

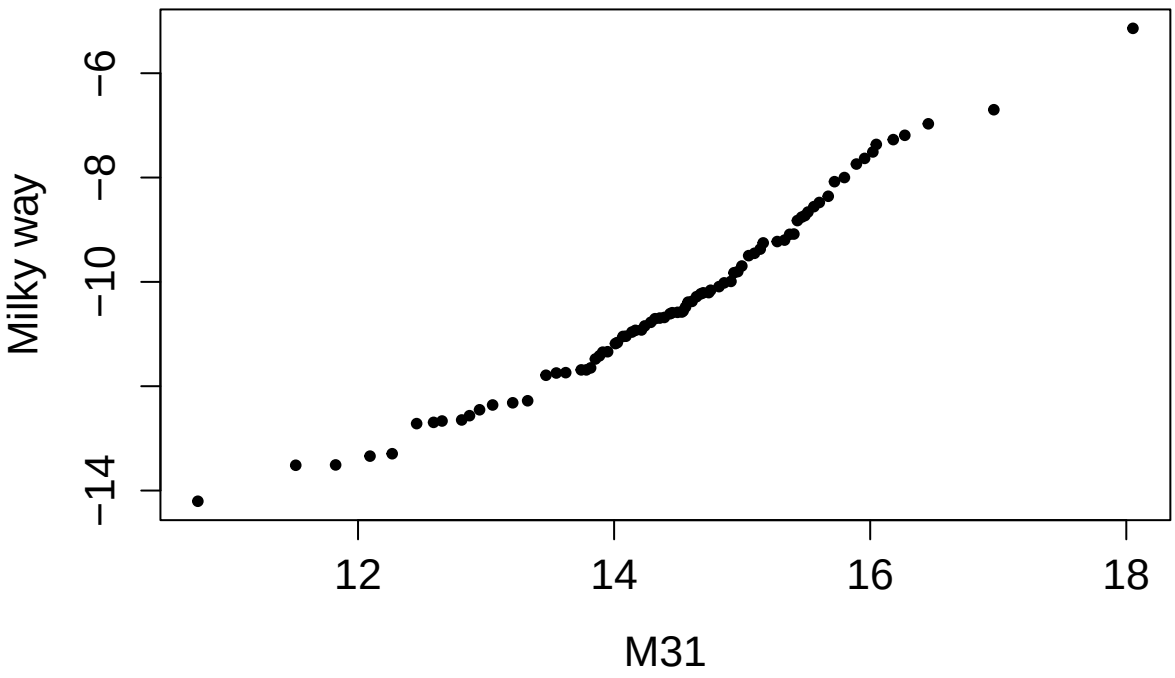


The OY axis of the above graph shows the empirical light distribution clusters in the Milky Way, and the OX axis theoretical distribution normal distribution. As can be seen, the empirical quantiles agree with the theoretical model.



Similarly as before, the OY axis of the above graph is shown empirical distribution of the brightness of the clusters in M31, while the OX axis theoretical distribution of the normal distribution. It can also be seen that the theoretical model does not agree with the lower values.

Comparing the distribution of two data sets, we get the following chart:



In the above case, the axes are empirical luminance distributions clusters in the Milky Way and in M31. In the figure above, we can observe a non-linear trend that leads to the conclusion that the distributions the brightness of the two galaxies is different.

In order to confirm the above conclusion, the following test should be carried out:

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  milky_way and m_31 + module
## D = 0.18333, p-value = 0.02348
## alternative hypothesis: two-sided
```

The test statistics have been described in the previous task. For the level significance $\alpha = 0.05$ the hipothesis should be rejected, but because the significance level can also be assumed at the level of $\alpha = 0.01$ no we are sure about the correctness of rejecting the hypothesis.

Estimators on survey data

A consistent estimator that is unbiased and the most efficient of the mean is the mean value. The value of the calculated mean is presented below:

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

For an unknown standard deviation and an unknown average value, we use student statistics. I am looking for the error value at 95% confidence.

Ultimately, the confidence interval is:

$$[171.1207; 173.641],$$

while the mean value together with the error from this interval is:

$$172.38 \pm 1.34[cm].$$

In order to test the null hypothesis, i.e. whether students smoke regardless of the amount chi-square test was used in the exercises. The statistics describing this test are:

$$\chi^2 = \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(n_{ik} - n \cdot p_{ik})^2}{n \cdot p_{ik}}, \tag{6}$$

where: n_{ik} - counts corresponding to a given pair categories, r = 3 - number of smoking frequency categories, s = 4 - number of exercise frequency categories, n - sum of n_{ik} and p_{ik} which is calculated from the formula:

$$p_{ik} = p_{i.}p_{.k} = \frac{n_{i.}}{n} \cdot \frac{n_{.k}}{k}. \tag{7}$$

In total, the statistics has 6 degrees of freedom, while the the critical area is on the right. The following result was obtained chi-square test:

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

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

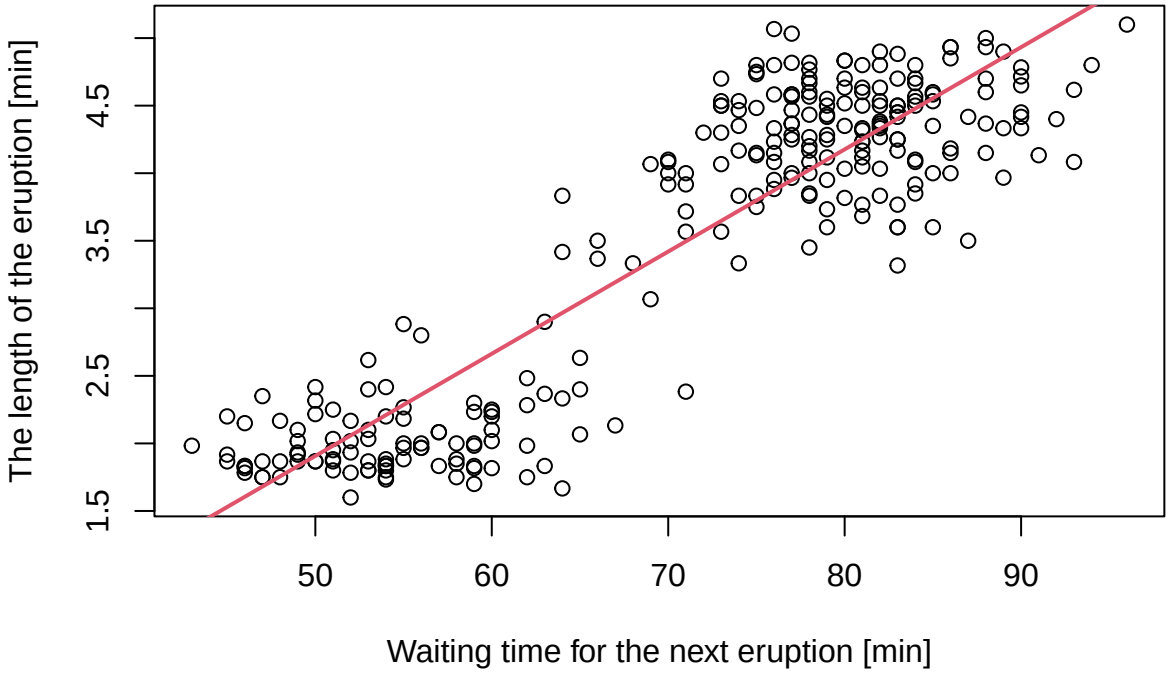
It can be concluded from the test result that for the significance level $\alpha = 0.05$ we have no grounds to reject the null hypothesis.

Linear Regression on faithful data

The following commands were executed to fit linear regression:

```
##
## Call:
## lm(formula = x ~ y)
##
## 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 ***
## y            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
```



Above are the statistics and the length dependence graph eruptions from waiting time for the next eruption along with the fitted straight using linear regression. Matching parameter values for straight:

$$y = a \cdot x + b, \tag{8}$$

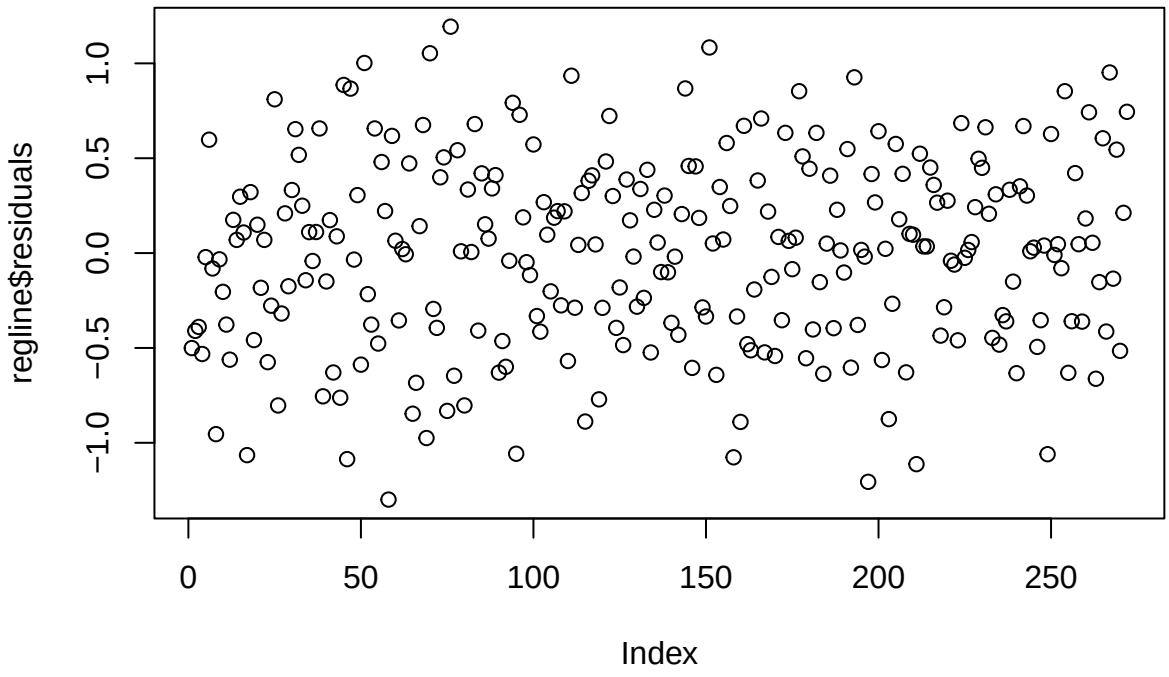
are, among others:

- $a = 0.0756 \pm 0.0023$,
- $b = -1.87 \pm 0.17$.

The above was used to estimate the duration of the next eruption formula with matching coefficients, where x = 80 min. After the substitution gave the following:

$$4.18 \pm 0.36.$$

The residual plot of the regression fit is drawn below:



As can be seen, the applied linear regression model is correct because the data is scattered and does not show any trend.

Akaike and **Bayesian** criteria are used for purpose comparison of the models used with the data. Values obtained from these Criteria are estimates of the error and compliance of the samples with the model. Their the value can indicate which model is a better choice where a lower value means that the model for each data is better. The values of these criteria for regression are presented below liner:

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

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

Using a 95% confidence level for the mean eruption duration for the time of 80 min, they were determined using the predict function predicted y values. Results of the action are presented below functions:

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

Ultimately, the confidence interval read from the above values is:

$$(4.1; 4.25).$$