# Elementary Statistics Project 2023-2024

Orfeo Terkuçi
*Faculty of Science*
*University of Antwerp*
Antwerp, Belgium
Orfeo.Terkuci@student.uantwerpen.be

***Abstract*— This is my report for the project of the subject “Elementary statistics” of the year 2023-2024 in the University of Antwerp.**

***Index terms*—Statistics, Heart infarcts**

## I. Introduction

The aim of the study was to describe factors associated with temporal trends in survival after hospital admission for acute myocardial infarction. We are providing a dataset with the following variables:

1. **`id`**: identification number
2. **`age`**: age of the patient at hospital admission
3. **`gender`**: gender of the patient: ‘0’ = man, ‘1’ = woman
4. **`hr`**: initial heart rate: beats per minute
5. **`bmi`**: BodyMassIndex: kg/m2
6. **`cvd`**: history of cardiovascular disease: ‘0’ = no, ‘1’ = yes
7. **`sho`**: cardiogenic shock: ‘0’ = no, ‘1’ = yes
8. **`mitype`**: type of myocardial infarction (pathological): ‘0’ = no presence of Q wavelengths, ‘1’ = presence of Q wavelengths
9. ***`los`*: *duration of hospital stay, in days***
10. **`dstat`**: hospital discharge status: ‘0’ = alive, ‘1’ = dead
11. **`lenfol`**: total duration of follow-up: number of days from hospital admission to the date of the last follow-up
12. **`fstat`**: status at last follow-up: ‘0’ = alive, ‘1’ = dead.

Each student has to look at the last three digits `ijk` of their student number and remove the following rows from the dataset: `k + 1`, `j + 1`, `i + 1`, `jk + 1`, `ij + 1`, `ik + 1`, `ijk + 1` en `i + j + k + 1`. My student number is `20213863`, thus I have **ijk = 863**.

The following rows will be removed from the dataset:

- `k + 1 = 4`
- `j + 1 = 7`
- `i + 1 = 9`
- `jk + 1 = 19`
- `ij + 1 = 49`
- `ik + 1 = 25`
- `ijk + 1 = 145`
- `i + j + k + 1 = 18`

The rows were removed with the command

```
data <- data[-c(k + 1, j + 1, i + 1, j*k + 1, i*j +
1, i*k + 1, i*j*k + 1),]
```

**Significant value $\alpha = 0.05$**

## II. The distribution of the variable `los`

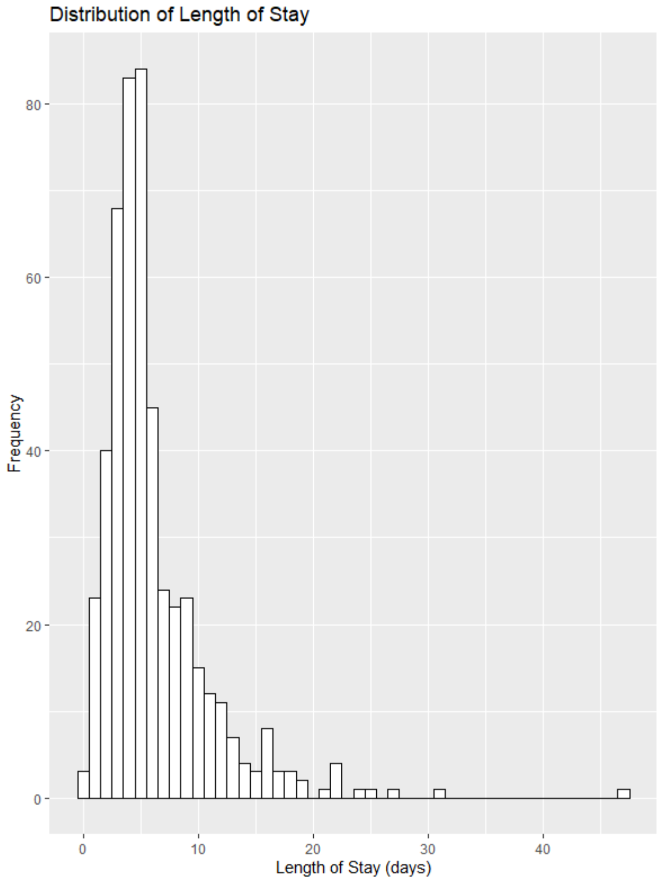
The first thing we needed to do was check if the variable `los` was normally distributed in our dataset.

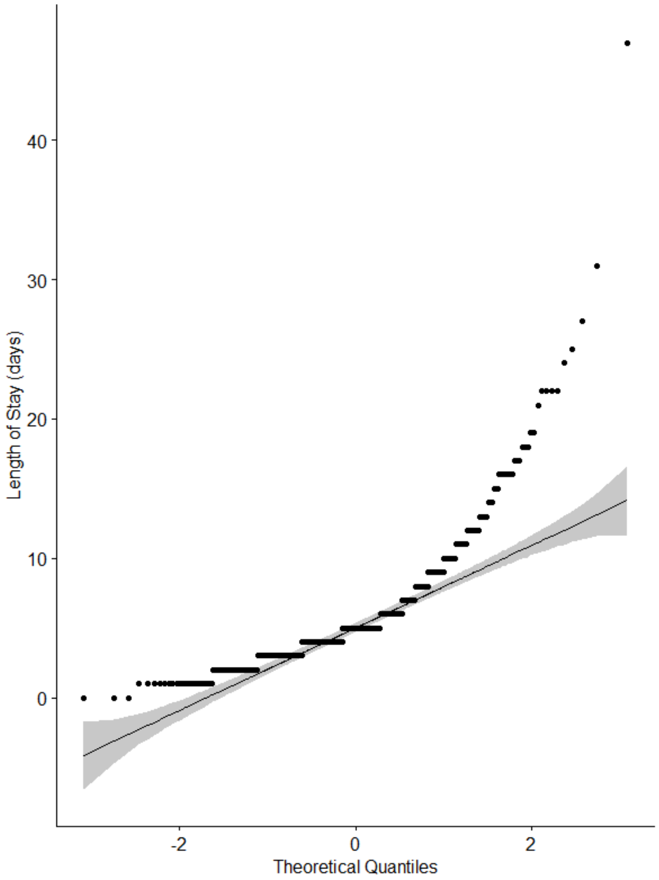
- **Null hypothesis**: The data is normally distributed
- **Alternative hypothesis**: The data is not normally distributed
- **Test statistic**: The W and the p-value obtained by `Q-Q plot` [1] followed by a `Shapiro-Wilk` [2] test and a `Kolmogorov-Smirnov` [3] test
- **Observed variable**: **`Los`**

The code that was used to plot the Q-Q graph and to do the `Shapiro-Wilk` test and the `Kolmogorov-Smirnov` test:

```
library(ggplot2)
library(ggpubr)
plot <- ggplot(data, aes(x=los)) +
        geom_histogram(binwidth=1,  color="black",
fill="white") +
    labs(x="Length of Stay (days)", y="Frequency",
title="Distribution of Length of Stay")
print(plot)
qqplot <- ggqqplot(data$los, ylab="Length of Stay
(days)", xlab="Theoretical Quantiles")
print(qqplot)
shapiro.test(data$los)

mean_los <- mean(data$los, na.rm = TRUE)
sd_los <- sd(data$los, na.rm = TRUE)
ks_result <- ks.test(data$los, "pnorm", mean_los,
sd_los)
print(ks_result)
```

### *A. Graphs*





*B. Is the data normally distributed?*

At first glance, as can be seen in the bar graph, the data is not normally distributed. It can be seen that most of the data is grouped at the left side of the graph, so it looks right-leaning. But just looking at a bar graph is not enough to conclude that it is indeed not normally distributed.

Furthermore, a `Shapiro-Wilk` test was done on the data with a significant value of 0.05.

The result of the `Shapiro-Wilk` test is as follows:

```
W = 0.76714, p-value < 2.2e-16
```

The **p-value** is **2.2e-16**, which is much smaller than 0.05. The **W** value tells how closely the data follows a normal distribution. The closer it is to 1, the more the data follows a normal distribution. And as you can see from that result, a **0.76714** is far from 1.

The result of the `Kolmogorov-Smirnov` test is as follows:

```
D = 0.21241, p-value < 2.2e-16
```

The **p-value** is 2.2e-16, much smaller than 0.05. The D-value, that is the test statistiec used in de KS test is 0.21241.

In both cases, de p-value is much smaller dan 0.05. Thus, the `los` variable is not normally distributed in our dataset.

1) *If not: in what way do the data deviate from normally distributed data?:*

When observing the Q-Q plot, it can be noticed that the plot is right skewed [4]. The points are above the 45° line. This is also visible in the histogram. The data in the left part of the graph, the smallest observations, are larger than would be expected with a normal distribution.

2) *Can you transform the data into normally distributed data?:*

I have tried seven different transformations:

1. Square root transformation
2. Cubic root transformation
3. Inverse transformation
4. Log transformation
5. Ln transformation
6. Arcsine transformation
7. Hyperbolic arcsine transformation

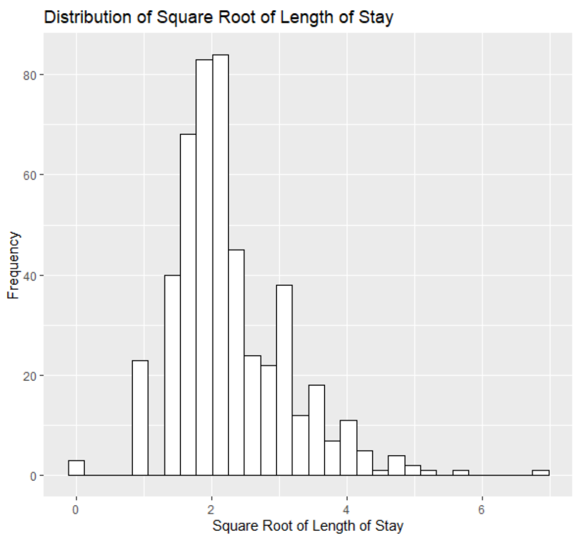
The used R-code is displayed below:

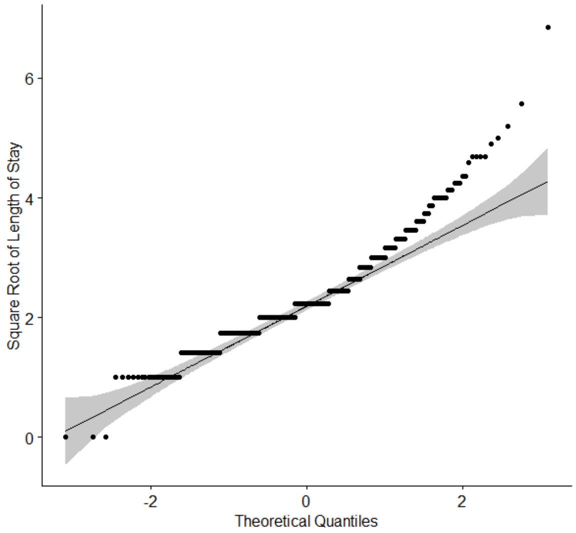
```
data$los_scaled <- 2 * (data$los - min(data$los,
na.rm = TRUE)) / (max(data$los, na.rm = TRUE) -
min(data$los, na.rm = TRUE)) - 1

data$los_square_root <- sqrt(data$los)
data$los_cube_root <- data$los^(1/3)
data$los_nat_log <- ifelse(data$los != 0,
log(data$los), 0)
data$los_log10 <- ifelse(data$los != 0,
log10(data$los), 0)
data$los_inverse <- ifelse(data$los != 0, 1 /
data$los, 0)
data$los_arcsine <- asin(data$los_scaled)
data$los_hyperbolic_arcsine <- log(data$los +
sqrt(data$los^2 + 1))
```

The data was rescaled to fit in the interval $[-1, 1]$, to be able to use it in the `bgsin(x)` function.

Here are the results of each transformation, shown in a bar graph and in a Q-Q graph. A `Shapiro-Wilk` test was also done after each transformation.

*Square root transformation:*

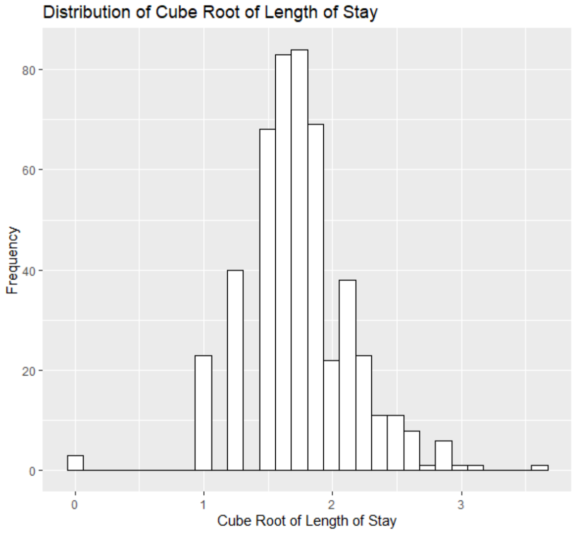


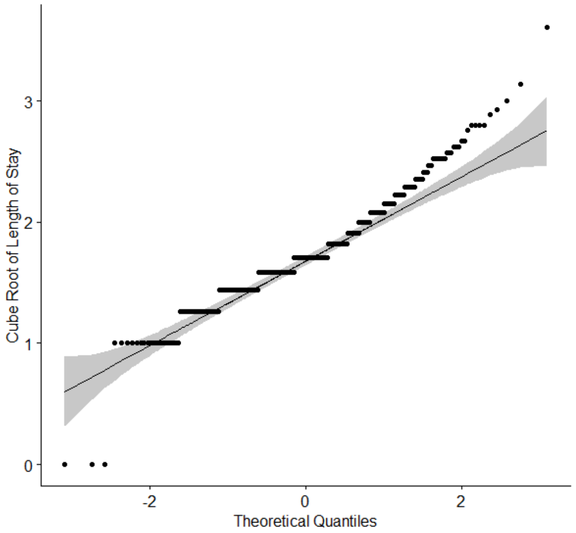


`Shapiro-Wilk` test results:

```
W = 0.93425, p-value = 6.413e-14
```

*Cubic root transformation:*

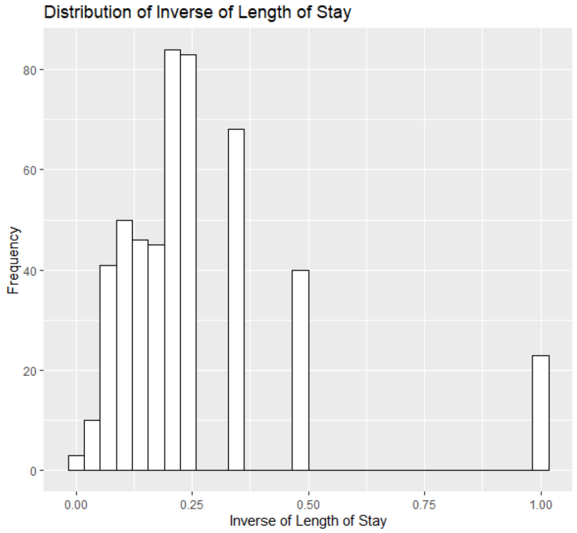


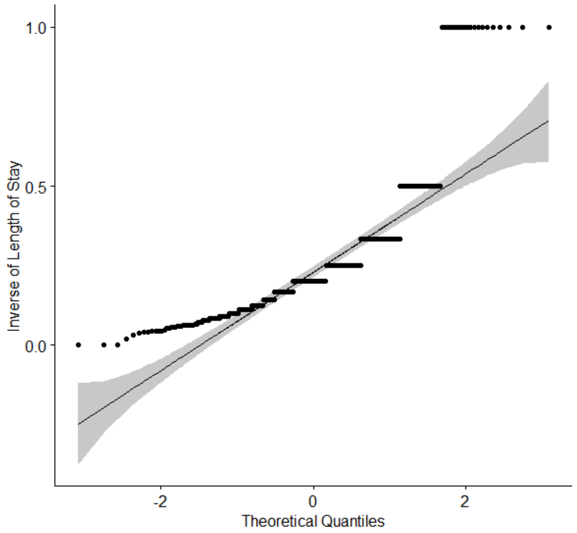


Shapiro-Wilk test results:

```
W = 0.95184, p-value = 1.371e-11
```

*Inverse transformation:*

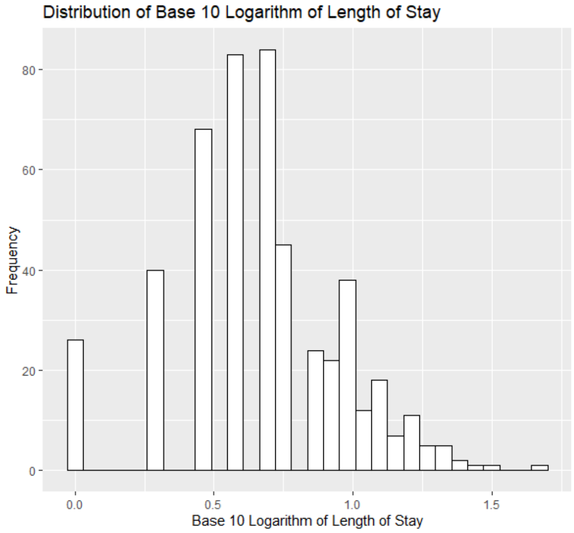


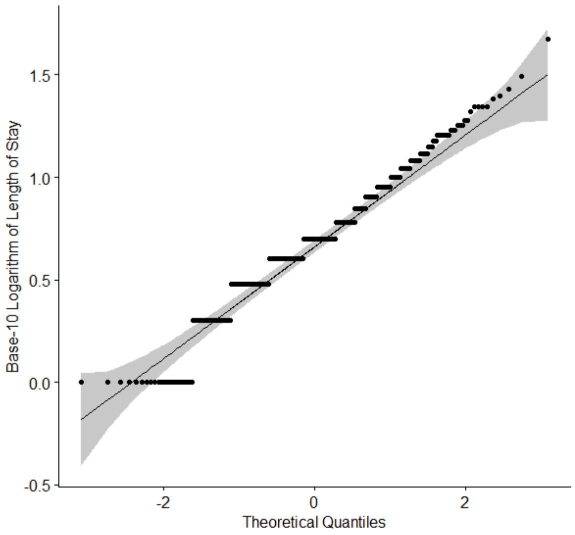


Shapiro-Wilk test results:

```
W = 0.73318, p-value < 2.2e-16
```

*Log transformation:*

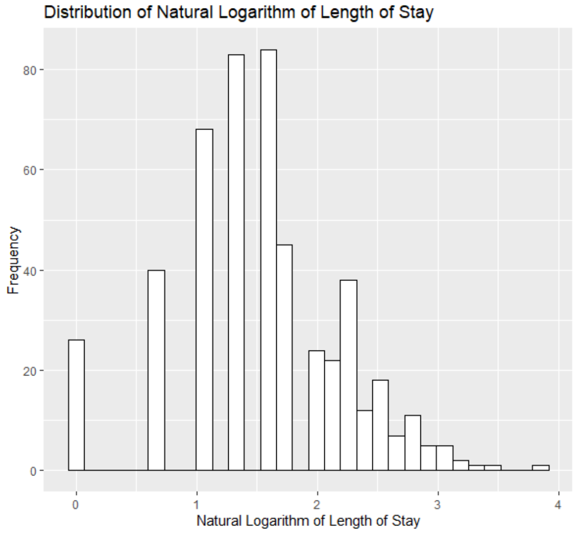


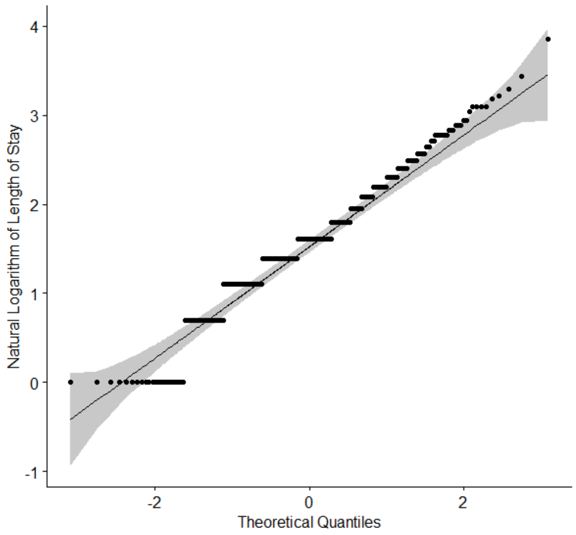


Shapiro-Wilk test results:

**W = 0.97247, p-value = 5.276e-08**

*Ln transformation:*

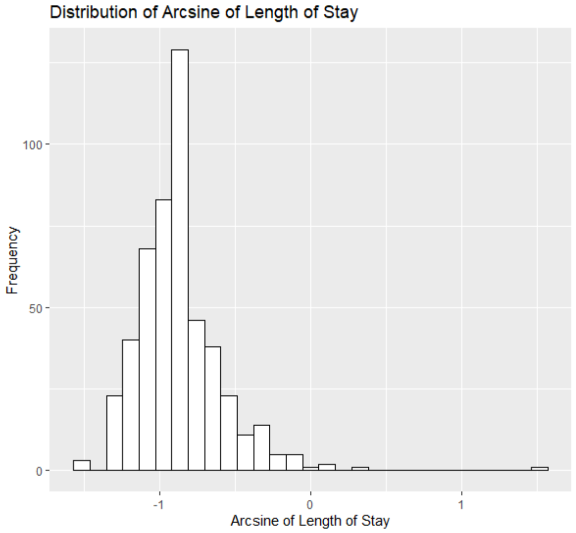


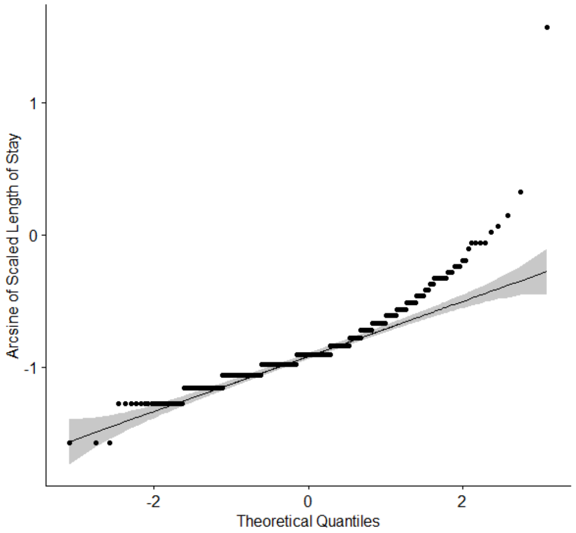


Shapiro-Wilk test results:

**W = 0.97247, p-value = 5.276e-08**

*Arcsin transformation:*

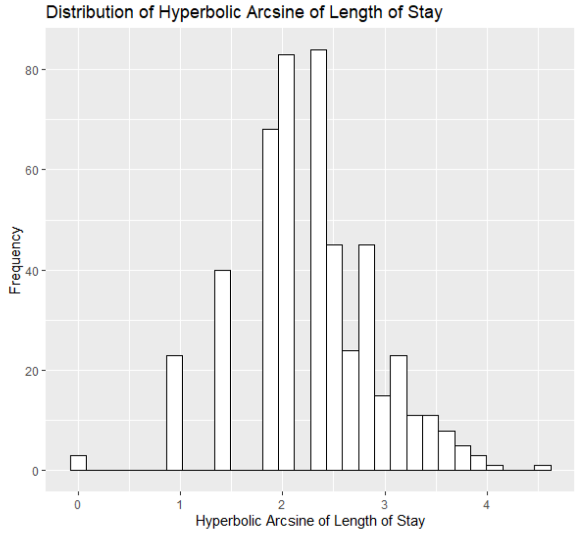


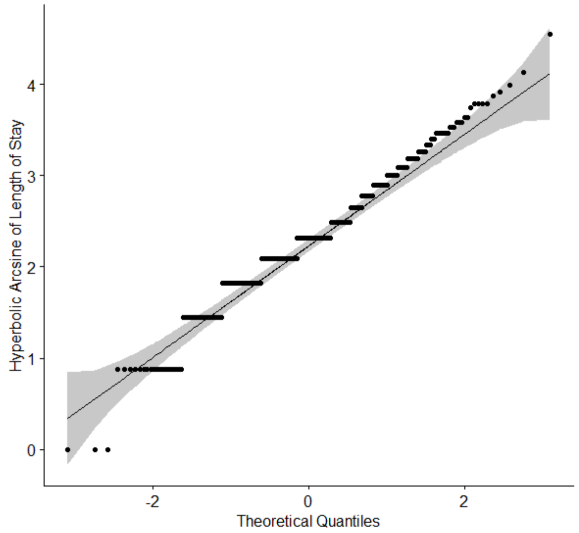


`Shapiro-Wilk` test results:

```
W = 0.86692, p-value < 2.2e-16
```

*Hyperbolic arcsin transformation:*





`Shapiro-Wilk` test results:

```
W = 0.97715, p-value = 5.603e-07
```

3) *Conclusion:*

Considering the results of the 'Shapiro-Wilk' test on each transformation, the transformed data is not normally distributed either. Thus, the null hypothesis that the `Los` variable (Length of Stay) is normally distributed is false.

## III. Is there a correlation between the type of myocardial infarct and the discharge status from the hospital?

To check whether there is a correlation between the type of myocardial infarct (the **`mitype`** variable) and the discharge status from the hospital (the **`dstat`** variable), three tests were conducted:

1. Pearson's chi-square test [5]
2. Kendall's rank correlation tau [6]

3. Spearman's rank correlation rho [6]

- **Null-hypothesis**: The two variables are independent of each other
- **Alternative hypothesis**: There is a correlation between the two variables
- **Observed variable**: **`mitype`** and **`dstat`**

### A. *Pearson's chi-square test*

R-code used:

```
chisq.test(data$mitype, data$dstat)
```

The result is as follows:

```
X-squared = 0.010877, df = 1, p-value = 0.9169
```

The p-value of 0.9169 is much greater than the significant value of 0.05, which means that we do not have enough evidence to reject the null hypothesis.

### B. *Kendall's rank correlation tau*

R-code used:

```
cor.test(data$mitype, data$dstat, method =
c("kendall"))
```

The result is as follows:

```
data:  data$mitype and data$dstat
z = -0.28568, p-value = 0.7751
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.0128795
```

The p-value of 0.7751 is much greater than the significant value of 0.05, which means that we do not have enough evidence to reject the null hypothesis.

### C. *Spearman's rank correlation rho*

R-code used:

```
cor.test(data$mitype, data$dstat, method =
c("spearman"))
```

The result is as follows:

```
data:  data$mitype and data$dstat
S = 20227653, p-value = 0.7754
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.0128795
```

The p-value of 0.7754 is much greater than the significant value of 0.05, which means that we do not have enough evidence to reject the null hypothesis.

### D. *Conclusion*

Based on the results of all three tests, we can conclude that the type of myocardial infarct and the patient's discharge status from the hospital.

## IV. Can you predict the BMI from the patient's age?

To check whether there is a correlation between the patient's age (the **`age`** variable) and their BMI (the **`bmi`** variable), the same three tests were conducted:

1. Pearson's chi-square test [5]
2. Kendall's rank correlation tau [6]
3. Spearman's rank correlation rho [6]

### A. *Pearson's chi-square test*

R-code used:

```
chisq.test(data$age, data$bmi)
```

The result is as follows:

```
data:  data$age and data$bmi
X-squared = 26611, df = 26390, p-value = 0.1677
```

The prerequisite for the Pearson's chi-square test is that the values are normally distributed. By running the Shapiro-Wilk test on the two variables we get the following results:

```
data:  data$age
W = 0.97329, p-value = 7.862e-08

data:  data$bmi
W = 0.98009, p-value = 2.838e-06
```

With a p-value much smaller than the significant value of 0.05, these variables are not normally distributed and thus we can disregard the null hypothesis.

### B. *Kendall's rank correlation tau*

R-code used:

```
cor.test(data$age, data$bmi, method = c("kendall"))
```

The result is as follows:

```
data:  data$age and data$bmi
z = -9.3438, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.2843868
```

A tau value of $-0.2843868$ signifies that there is a weak to moderate negative correlation between the two values. A p-value of 2.2e-16 is much smaller than the significant value, thus we have significant evidence to reject the null hypothesis.

### C. *Spearman's rank correlation rho*

R-code used:

```
cor.test(data$age, data$bmi, method = c("spearman"))
```

The result is as follows:

```
data:  data$age and data$bmi
S = 28280019, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.4160936
```

A rho value of $-0.4160936$ signifies that there is a moderate negative correlation between the two values. A p-value of $2.2e-16$ is much smaller than the significant value, thus we have significant evidence to reject the null hypothesis.

*D. Linear regression*

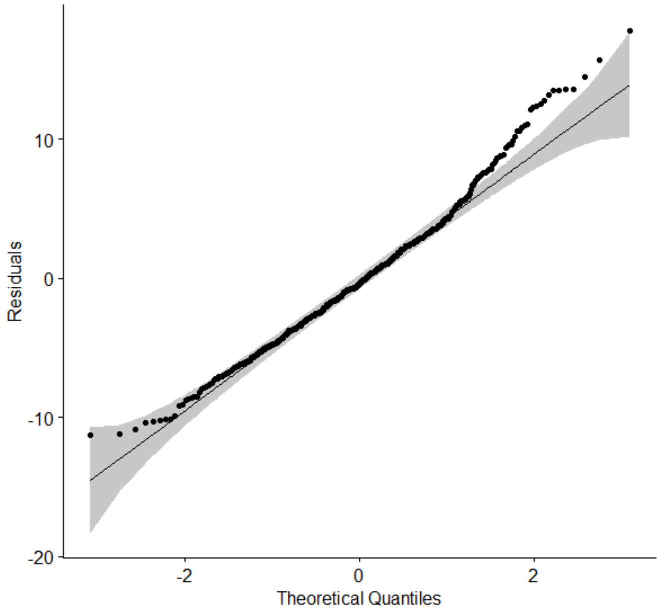
A linear regression was also performed to learn more about the relationship between Age and BMI.

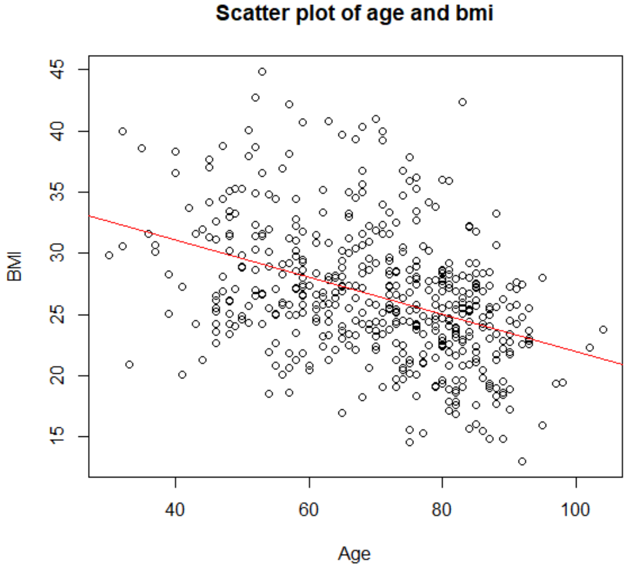
R-code used:

```
linear <- lm(data$bmi ~ data$age, data)
summary(linear)
```

The result of the linear regression is as follows:

```
Call:
lm(formula = data$bmi ~ data$age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2672  -3.4126  -0.3653   2.8127  17.8126

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2159     1.0922  34.075   <2e-16 ***
data$age     -0.1524     0.0153  -9.956   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 4.924 on 491 degrees of
freedom
Multiple R-squared:   0.168,  Adjusted R-squared:
0.1663
F-statistic: 99.11 on 1 and 491 DF,   p-value: <
2.2e-16
```





1) *Residuals analysis:*

The residuals are calculated as `residual = actual y value - predicted y value`. Having a negative residual means that the predicted value was too high, and similarly having a positive residual means that the predicted value was too low. The aim of the regression line is to minimize the sum of the residuals.

Let us look at some statistics of the residuals:

- **Min**: This is the minimum residual, the largest negative difference between the actual value and the predicted value. A minimum of **−11.2672** tells us that in at least one point the predicted BMI value was 11.2672 units higher than the actual BMI value from the dataset
- **Max**: Similarly, the maximum residual is the largest positive difference between the actual value and the predicted value. A maximum of **17.8126** tells us that in at least one point the predicted BMI value was 17.8126 units lower than the actual BMI value from the dataset.
- **Median**: The median is the middle value in a list of sorted residuals. A value of **−0.3653** tells us that half of the residuals have a value lower than −0.3653, which means that for half of the data the models predicted BMI was 0.3653 or more units higher than the actual BMI. The median is close to zero and this suggests that on average the models predictions are not systematically too high or too low.
- **1Q**: The first quartile (25%) of the residuals. A value of **−3.4126** means that 25% of the residuals are less than or equal to −3.4126, which means that for 25% of the values, the predicted BMI was 3.4126 or more units higher than the actual BMI.

- **3Q**: The third quartile (75%) of the residuals. A value of **2.8127** means that 75% of the residuals are less than or equal to 2.8127, which means that for 75% of the values, the predicted BMI was 2.8127 or more units lower than the actual BMI.

Knowing the scale of BMI, a minimum of $-11.2672$ and a maximum of 17.8126 is quite a wide deviation from the actual values in the dataset. This means that the predictions the model makes are far off from the actual values. A median of $-0.3653$, which is rather close to 0 tells us that the model does not have a significant systematic bias (higher or lower predictions).

The fact that Q1 is closer to zero than Q3 suggests that the residuals on the lower end (i.e., where the model overestimates) are generally closer to zero than those on the higher end (i.e., where the model underestimates), which means that when the model is off, the amount it overestimates tends to be smaller than the amount it underestimates.

Let us look at the coefficients:

- Intercept (37.2159): This is the estimated value of the BMI when the age is 0. In other words, it's the predicted BMI for an individual with an age of 0.
- `data$age (-0.1524)`: This is the estimated change in BMI for a one-unit increase in age. In other words, for each additional year of age, BMI is predicted to decrease by 0.1524 units, on average.

By looking at the significant codes and the estimated p-value, we can confidently conclude that age has a significant effect on the BMI.

The **Residual Standard Error (`RSE`)** is the variability of the residuals. An RSE of 4.924 means that on average the actual values of BMI deviate from the predicted values of the model by 4.924 units. It has a freedom of 491 degrees, based on the amount of total independent information available to the model (total amount of data).

The `Multiple R-squared` is the proportion of the variance in the BMI that is predictable from the age. A value of 0.168 means that 16.8% of the variance in BMI can be explained by age.

The `Adjusted R-squared` is a modified version of R-squared that has been adjusted for the number of predictors in the model. It takes into account the degrees of freedom and penalizes the addition of uninformative predictors to the model. In this case, since there's only one predictor (age), the adjusted R-squared is very close to the multiple R-squared.

And lastly, the F-statistic and its p-value tell us whether predictors in the regression model provide a significant improvement in the fit of the model. In our case a high **F-statistic of 99.11** and a very low **p-value of less than 2.2e-16** indicates that age significantly improves the model's ability to predict BMI, thus the model is statistically significant.

### *E. Conclusion*

Based on the results of the Kendall's and Spearman's rank correlation tests we can conclude that there is a moderate negative correlation [7] between age and BMI.

## References

[1] "Q-Q plot," [Online]. Available: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

[2] "Shapiro-Wilk Normality test," [Online]. Available: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

[3] "SPSS Kolmogorov Smirnov test for normality," [Online]. Available: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

[4] "Right-skewed data," [Online]. Available: https://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html#right-skewed-data

[5] "Pearson's chi-squared test," [Online]. Available: https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

[6] "Spearman's rank correlation coefficient," [Online]. Available: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

[7] "Negative Correlation," [Online]. Available: https://en.wikipedia.org/wiki/Negative_relationship