Estadística avanzada para ciencia de datos

Contrastes de Hipótesis - Ejercicio 2 Tests Estadísticos

Jakub Maciążek

Exercise description

> **Scenario** It is desired to study the possible relationship between water hardness and mortality in Great Britain. To do this, we want to explicitly check if the hardness of the water influences mortality in general throughout the country. Is there a correlation between both variables?

Instructions

Prepare an RMarkdown file explaining the steps to follow to study the proposed case. Analysis of results. Explanations. Deliver in the CV in a task the .Rmd file together with the output.

- Download the water.csv file from the CV datasets directory and import it into R.
- The dataset has information in the North and South of Great Britain on mortality and water hardness.
- Graphically analyzes (scatterplot, histogram, etc.) the mortality, differentiating the dataset points according to the zone (North or South).
- Analyze the normality of the data.
- Check through a correlation analysis if the hardness of the water influences mortality in general throughout the country.
- Check if the same thing happens if we separate the data according to the zone. Is there a correlation between both variables if we analyze it by area? Use the statistical tests seen in class. Explain the results.

Study overview

Given above scenario and instructions, following steps will be taken in order to study the case:

- Required data will be loaded, it format presented and explained.
- Several graphs will be presented for initial graphical analysis
- Data normality will be tested
- T-tests and correlation will be calculated for an in depth analysis
- Based on above, with regard to the total sample and with distinctions by regions, conclusions will be presented.

Data and presentation

Data set analyzed in the study contains records corresponding to several cities in UK, each with following four fields:

- location - label North or South, representing location of the city
- town - name of the city

- mortality - number of deceased people in given city
- hardness - hardness of water recorded at given city

```
water_df <- read.csv("S:/0_Universidad_de_Malaga/MI_Ingenieria_y_ciencia_de_datos/Estatistica_avanzada_
```

```
head(water_df)
##   X location       town mortality hardness
## 1 1    South       Bath      1247      105
## 2 2    North Birkenhead      1668       17
## 3 3    South Birmingham      1466        5
## 4 4    North  Blackburn      1800       14
## 5 5    North  Blackpool      1609       18
## 6 6    North     Bolton      1558       10
```

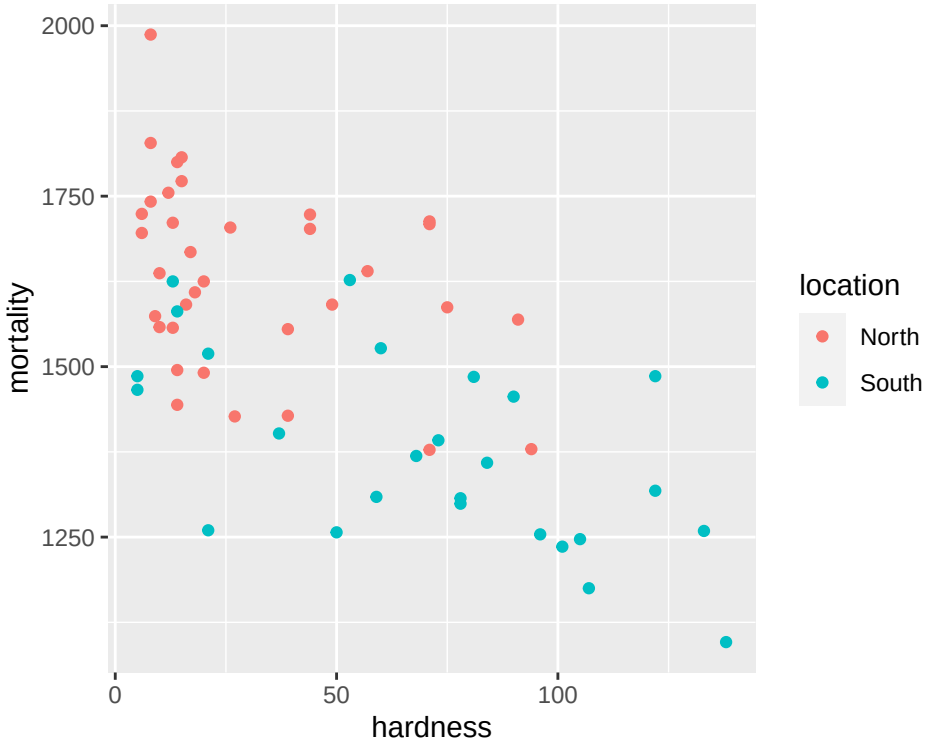
Graphical analysis

In the next step, several plots will be presented for initial graphical analysis.

Scatter plot

First presented graph is a scatter plot, representing relation of mortality rate to water hardness. Each point is also colored, representing region of the city, where measurement was taken.

```
ggplot(data = water_df, aes(x=hardness, y=mortality)) + geom_point(aes(colour=location))
```



On the graph inversely proportional trend can be observed, as in general points follow direction from upper left to lower right angle, meaning that cities with more hard water measured smaller mortality rates.

Moreover, cities in the North predominate at the beginning of this line, presenting higher mortality rates, while in contrary, cities form the South overweight end of that line, with smaller mortality rates.

Based on those observations, possibility of correlation between water hardness and mortality, or city location and mortality may be assumed. (Same applies for correlation between water hardness and city location.)

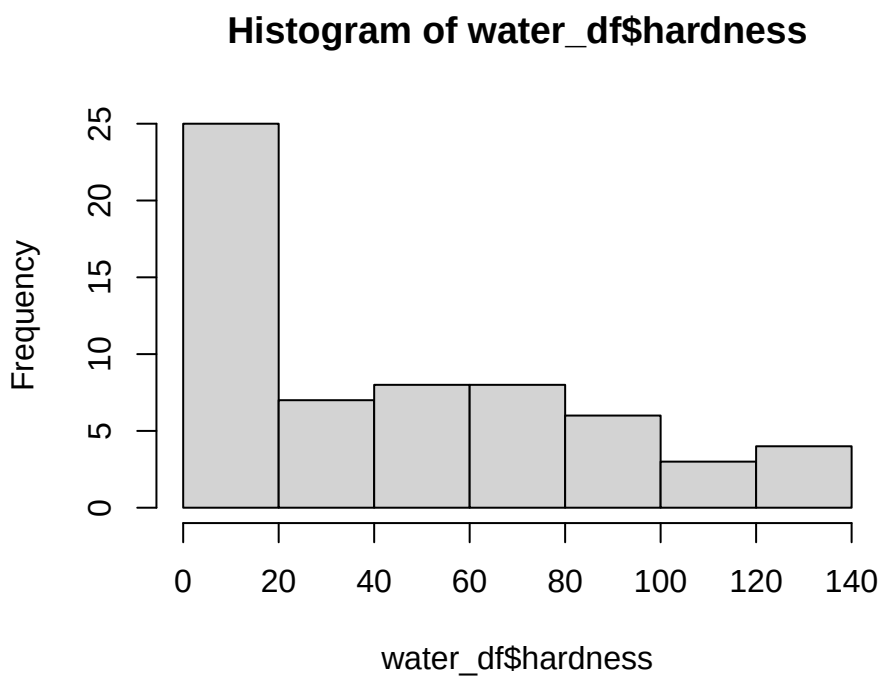
Histograms

This section presents histograms both for water hardness and mortality rates, based on which dominating values and normality can be assumed.

Water hardness

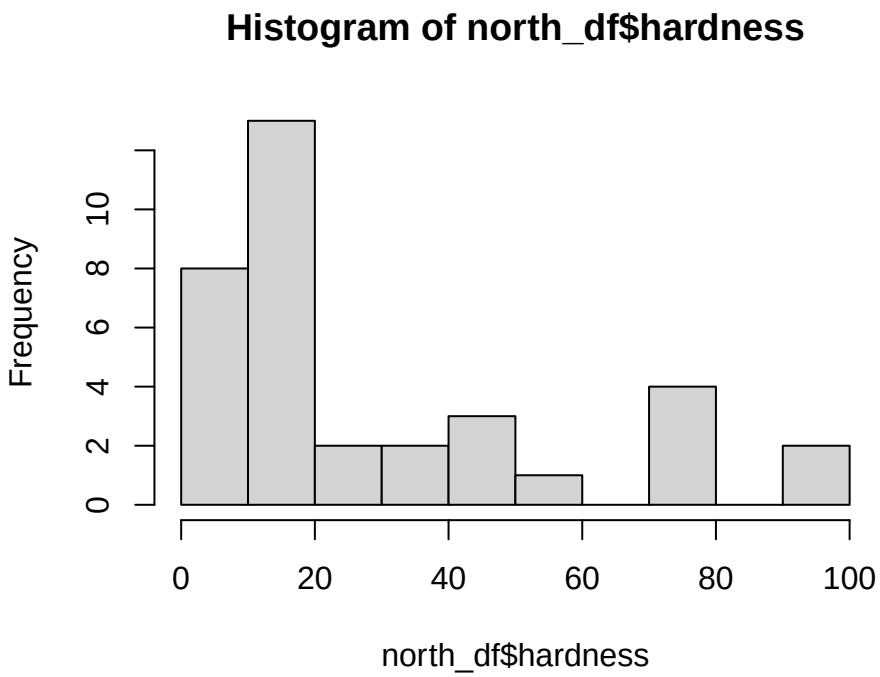
Based on the histograms of water hardness, following can be sated; In general hard water predominates in measurements. However, after separately examining data by regions, it can be seen that while almost all measurements from the Noth present hard water, water hardness is almost equal throught the South. Moreover neither general sample, nor considered by regions, follows normal distiribution.

```
hist(water_df$hardness)
```

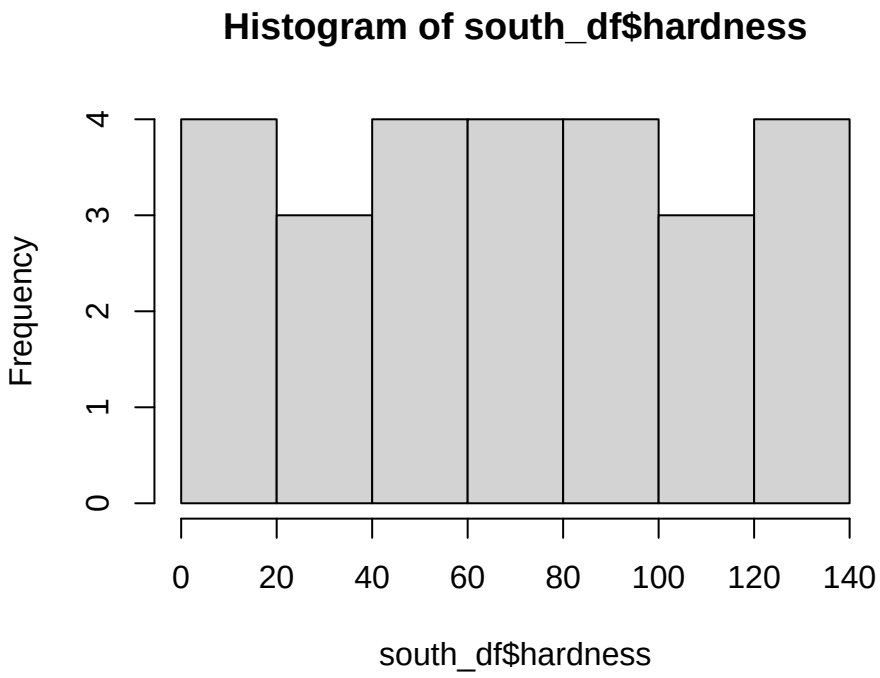



```
north_df <- water_df[water_df$location %in% c("North"),]
south_df <- water_df[water_df$location %in% c("South"),]

hist_hard_north <- hist(north_df$hardness)
```



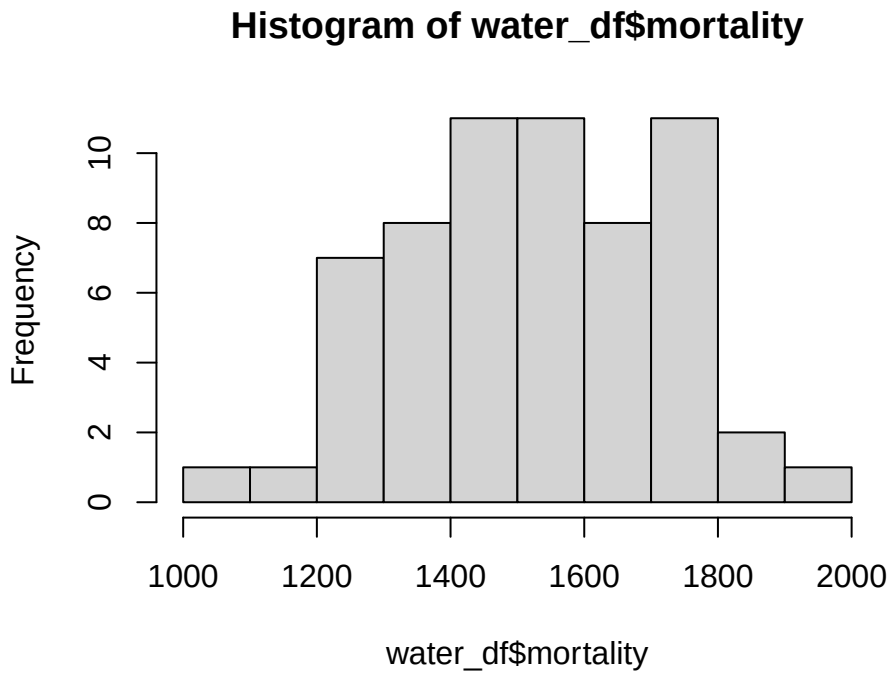

```
hist_hard_south <- hist(south_df$hardness)
```



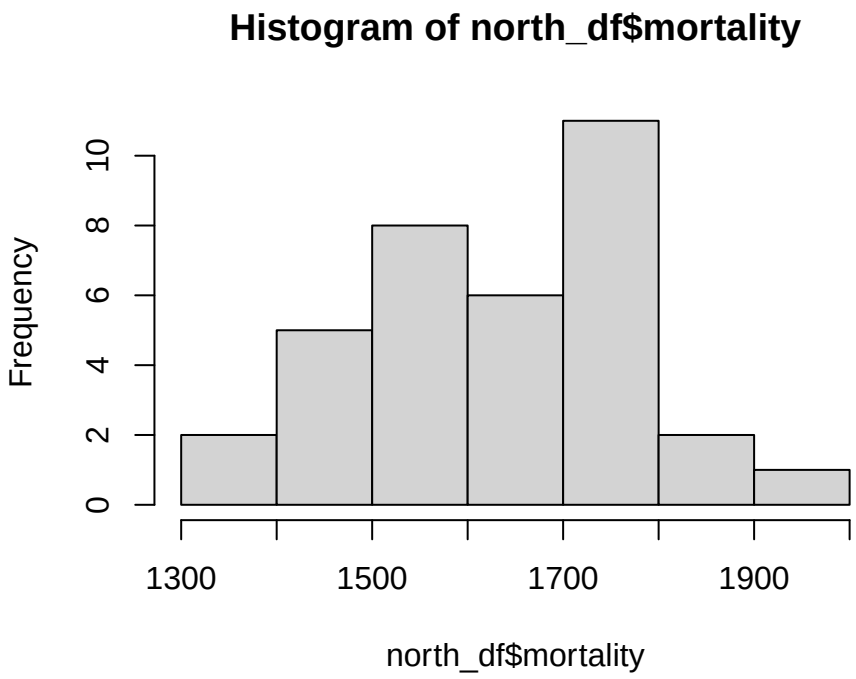
Mortality rates

Based on the shapes of mortality rate histograms, possibility of normal distribution of data can be assumed both for general sample, and those distincted by regions. From given graphs it is hard to assume whether any of the samples is skewed, however on comparison we can see that mortality rates are smaller in the South in comparison to North, which supports previous assumptions about possible correlation.

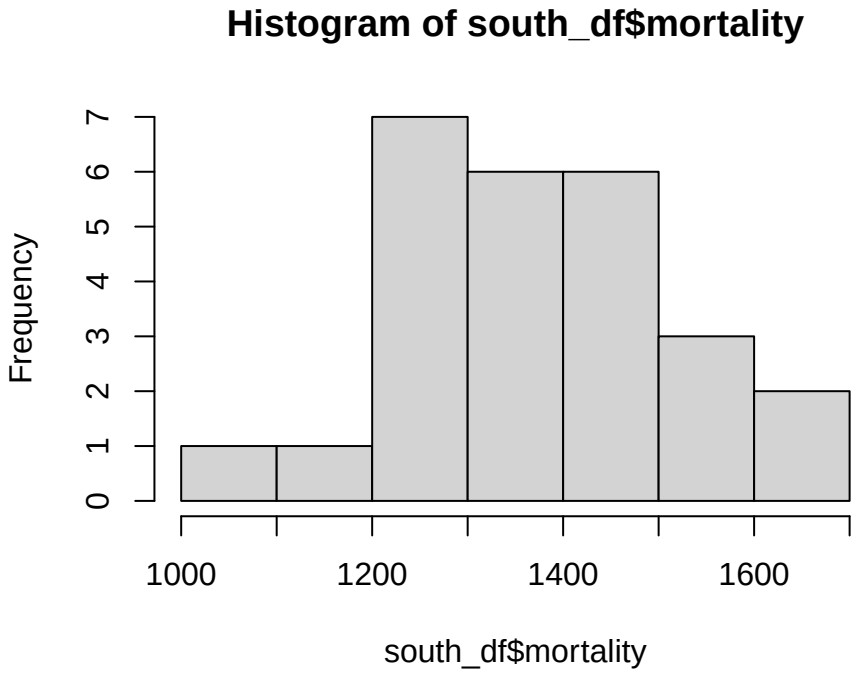
```
hist(water_df$mortality)
```



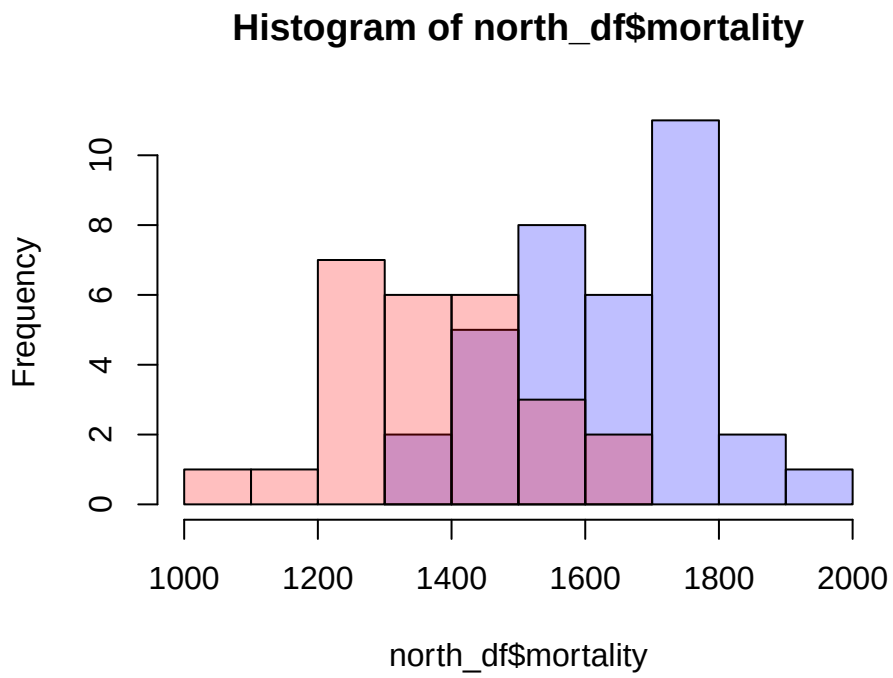

```
hist_mort_north <- hist(north_df$mortality)
```




```
hist_mort_south <- hist(south_df$mortality)
```



```
plot( hist_mort_north, col=rgb(0,0,1,1/4), xlim=c(1000,2000))
plot( hist_mort_south, col=rgb(1,0,0,1/4), xlim=c(1000,2000), add=T)
```

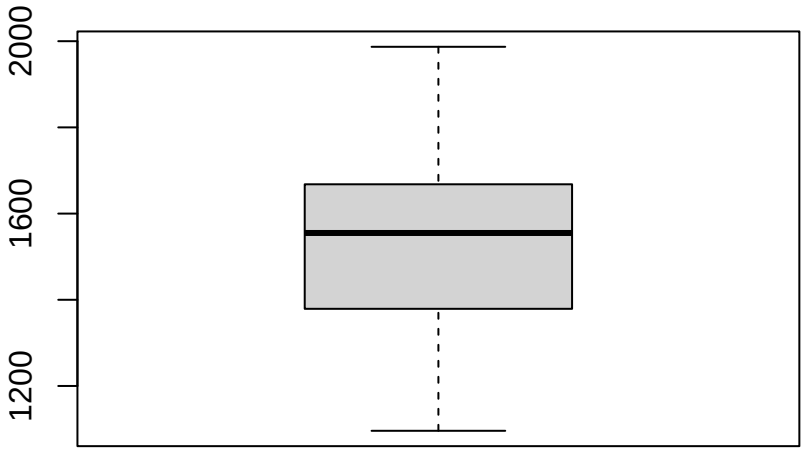


Box plots

Based on general morality box plot following can be stated:

- Minimum mortality rate was little over 1100, while maximum was almost 2000, giving the range of 900.
- 50% of values corresponds to IQR range from about 1400 to 1650, giving the span of 250.
- Median mortality rate was around 1550.

```
boxplot(water_df$mortality)
```

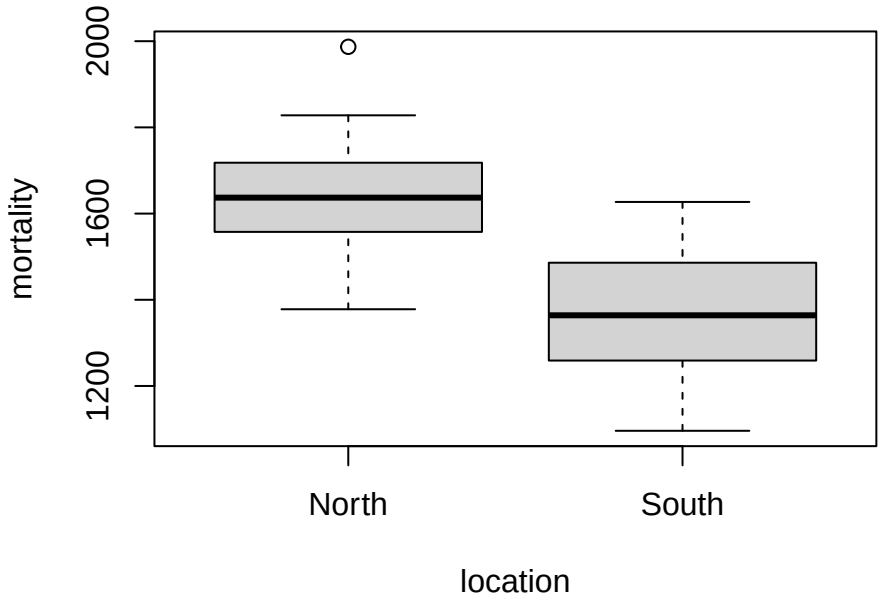


Much more information can be derived from box plots separated by regions:

- outlier for the North region means that more accurate total range from previous point should be said between 110 and 1850.
- There is a big difference between medians of regions, ~1650 for the north and ~1400 for the South. Moreover, median of North is similar to South maximum value, and South median is similar to North median value.
- IQR ranges between samples do not overlap. They share only about 50% of values.

Based on those information, possibly significant difference and correlation between mortality rates in different regions can be observed.

```
boxplot(mortality~location, data=water_df)
```



Data normlality analysis

Based on the histograms presented in previous section, lack of normal distribution was observed for water hardness and possible normal distribution for mortality rates. For morality rates also box plots suggest normal distribution, with both IQR and full ranges symmetrically centered around medians.

This means that T-Test can not be calculated neither for water hardness, to compare it between regions, neither for comparison of mortality rates. However, T-test can be run for mortality rates with location used as a label, in order to test for equal values of means between the samples.

To ensure that normal distribution, Shapiro-Wilk test was also conducted. In each test, the p-value was greater than 0.05 implying that the distribution of the data are not significantly different from normal distribution, meaning normality can be assumed for each of the samples.

```
shapiro.test(water_df$mortality)
##
##  Shapiro-Wilk normality test
##
## data:  water_df$mortality
## W = 0.98554, p-value = 0.6884

shapiro.test(north_df$mortality)
##
##  Shapiro-Wilk normality test
##
## data:  north_df$mortality
## W = 0.97554, p-value = 0.6117
```

```
shapiro.test(south_df$mortality)
##
##  Shapiro-Wilk normality test
##
## data:  south_df$mortality
## W = 0.96579, p-value = 0.518
```

Small sample size could have affected the test, however as same conclusions were derived from visual inspection, normal distribution for mortality samples (both general, and divided by regions) was confirmed.

T-tests and correlation calculations

Following section presents calculations which aim to support and prove assumptions established in the previous sections.

T-Tests

Paired T-Test was used to confirm, whether and how means of mortality rates differ significantly between North and the South. As variances are unknown, Welch T-test was calculated.

```
t.test(north_df$mortality, south_df$mortality)
##
##  Welch Two Sample t-test
##
## data:  north_df$mortality and south_df$mortality
## t = 7.1427, df = 53.29, p-value = 2.584e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  184.6919 328.8928
## sample estimates:
## mean of x mean of y
##  1633.600  1376.808
```

Recived p-value is smaller than 0.05, which means H_0 **(no difference between means)** can be rejected.

```
t.test(north_df$mortality, south_df$mortality, alternative = "greater")
##
##  Welch Two Sample t-test
##
## data:  north_df$mortality and south_df$mortality
## t = 7.1427, df = 53.29, p-value = 1.292e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  196.6111      Inf
## sample estimates:
## mean of x mean of y
##  1633.600  1376.808
```

Moreover, the next test confirms, that mean mortality rate in the North is significantly greater than in the South.

Correlation

To evaluate possible relationship between the variables, correlation test was calculated.

Because water hardness does not satisfy normal distribution requirement, it is impossible to use Pearson correlation. Instead, non-parametric Spearman rank-based correlation test was calculated, as it allows data to have distribution other than normal.

```
cor.test(water_df$mortality,
         water_df$hardness,
         method = "spearman")
## Warning in cor.test.default(water_df$mortality, water_df$hardness, method =
## "spearman"): Cannot compute exact p-value with ties
##
##  Spearman's rank correlation rho
##
## data:  water_df$mortality and water_df$hardness
## S = 61710, p-value = 4.795e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.6316646
```

P-value from correlation test for the whole country was calculated to be **4.795e-08**, which is lower than usually required 0.05, therefore results of the test can be considered significant.

Rho was calculated to have value of **-0.6316646**, meaning there is a strong negative correlation, proving earlier assumptions of decreasing mortality rates for rising water hardness.

```
cor.test(north_df$mortality,
         north_df$hardness,
         method = "spearman")
## Warning in cor.test.default(north_df$mortality, north_df$hardness, method =
## "spearman"): Cannot compute exact p-value with ties
##
##  Spearman's rank correlation rho
##
## data:  north_df$mortality and north_df$hardness
## S = 10026, p-value = 0.01603
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4042079

cor.test(south_df$mortality,
         south_df$hardness,
         method = "spearman")
## Warning in cor.test.default(south_df$mortality, south_df$hardness, method =
## "spearman"): Cannot compute exact p-value with ties
##
##  Spearman's rank correlation rho
##
## data:  south_df$mortality and south_df$hardness
## S = 4667.5, p-value = 0.001322
## alternative hypothesis: true rho is not equal to 0
```

```
## sample estimates:
##        rho
## -0.5957229
```

When considering correlation in scope of North and South regions, it is also significant with acquired **p-values** of **0.01603** and **0.001322** accordingly, and **rho** values **-0.4042079** and **-0.5957229**. This means that for both regions there is a significant negative correlation, however it is ~1.5 times stronger for the South region.

Conclusions

Following paper provided step by step study of the possible relationship between water hardness and mortality rates in Great Britan.

Beginning with graphical analysis of the data, both general and scoped to regions, possibility of negative correlation in all 3 cases was established. Those assumptions were later proved through correlation tests, that confirmed the correlation between water hardness and mortality rates.

Therefore it can be concluded, that the hardness of the water is negatively correlated to mortality, both in general throughout the country, and when taking into account separately North and South regions.

However, correlation does not equal causation, therefore possible influence requires further studies.