Statistical Modelling (STMO)

FINAL PROJECT

28/12/2020

—

Sata KAMARA

TABLE OF CONTENTS

INTRODUCTION

This project consists in studying a dataset with a statistical approach. The aim is to analyze the data with fundamental statistical tools, to interpret the results and to construct a model.

In view of the current health situation, I chose to work on a dataset related to effects of Covid-19 world pandemic on employment. For this purpose, I used the dataset "Labour productivity and unit labour costs" from the statistical office of the European Union (Eurostats). This dataset also contains data dating from before the pandemic.

Labor productivity is an economic measure that evaluate the workforce of a country. It is given by the following formula:

$$Labor\ productivity\ = \frac{Total\ output}{Total\ input}$$

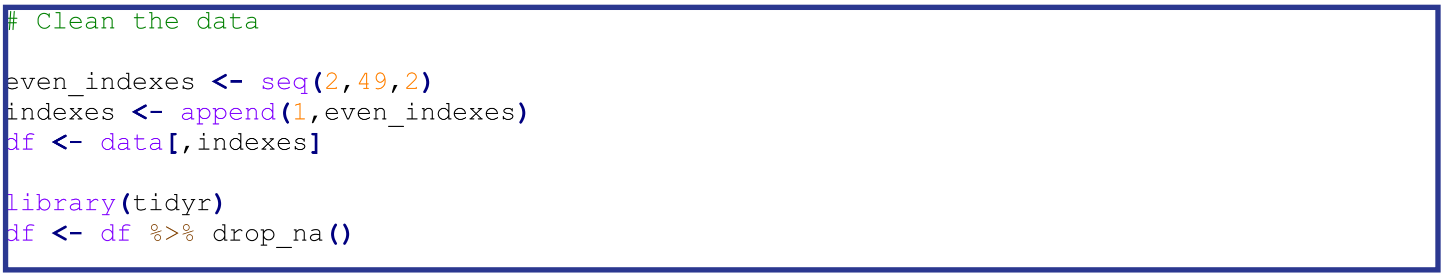
The aim of this project is to understand how the productivity evolves with time and what are the effects of the pandemic on this.

GENERAL ANALYSIS

In this section, I analyze the dataset of the Covid-19 period to understand how the productivity changes with time.

Description of the dataset

The initial dataset contains three sheets including two of presentation: I extract the sheet of interest for the project. The dataset is composed of 34 rows and 49 columns, representing the quarterly real labor productivity per person of 29 countries in Europe, the European Union (EU) and the Euro area. Before starting the analyze, I cleaned the data: this pre-processing step allows to keep only the relevant data for the study. The data cleansing is made by the following code and it can be illustrated by the following visualization objects:



```
# Clean the data

even_indexes <- seq(2,49,2)
indexes <- append(1,even_indexes)
df <- data[,indexes]

library(tidyr)
df <- df %>% drop_na()
```

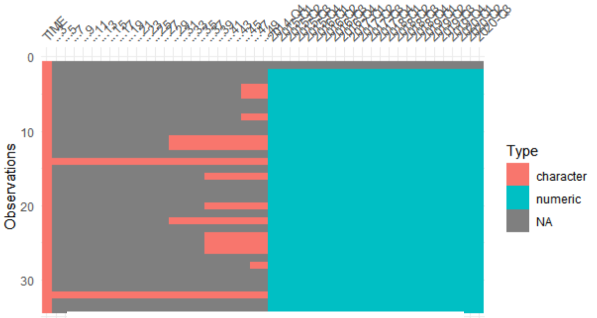


Figure 1.1: Data before cleansing

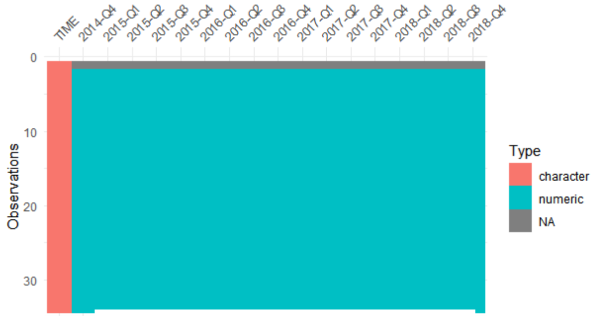
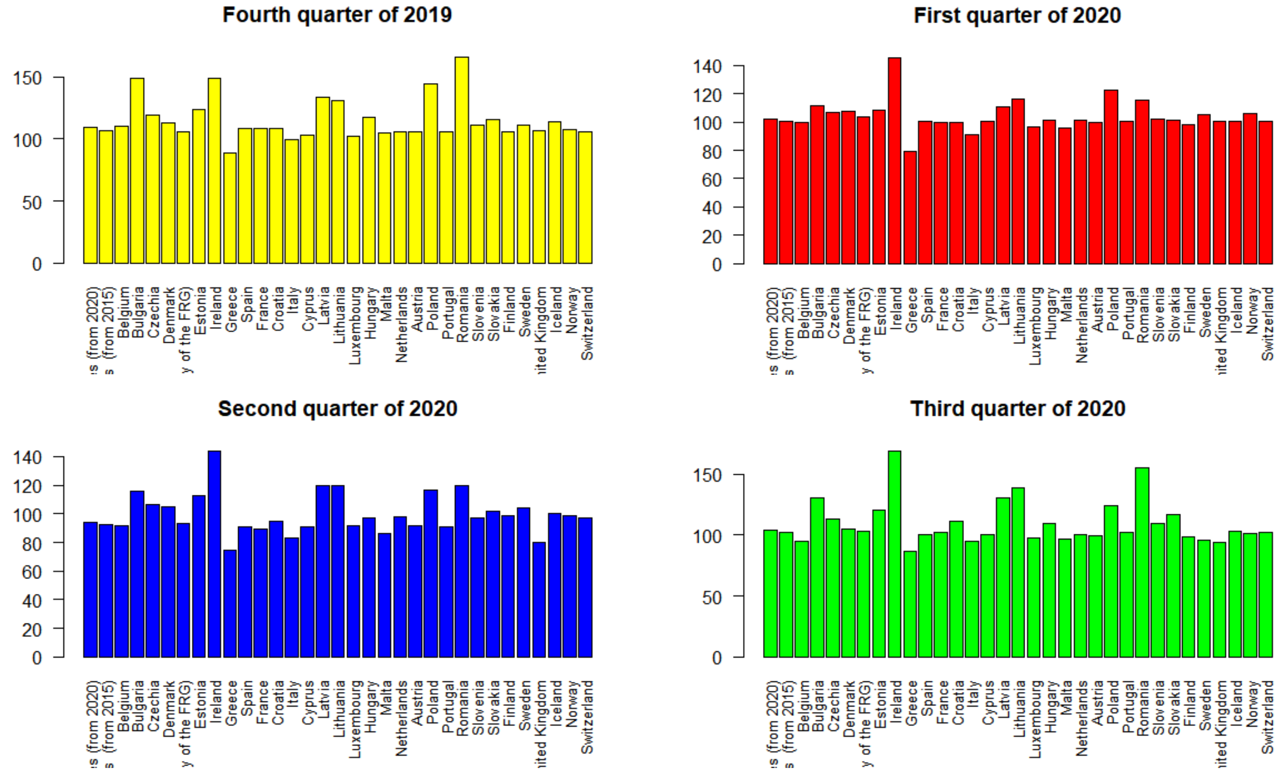


Figure 1.2: Data after cleansing

After that, I created a subset that only contains the data of the Covid-19 period, which means from the fourth quarter of 2019 to the third quarter of 2020[1].

Analysis of the Covid-19 subset

Firstly, I use the function `barplot()` to have an overview of the dataset:


Fourth quarter of 2019

First quarter of 2020

Second quarter of 2020

Third quarter of 2020

Figure 2: Barplots of each quarter of Covid-19 period

For each quarter, the labor productivity seems to be uniformly distributed according the countries. However, we can notice that this rate is globally lower for the second quarter of 2020: this can be explained by the fact that the pandemic has been declared global at the beginning of this period and this was followed by lockdowns for many countries in Europe.

[1] As the year 2020 is not finished yet, the data for the fourth quarter of 2020 are not available

Then I used statistical visualization tools to have more information about the distribution of each quarter: boxplots, and stem-and leaf diagrams.

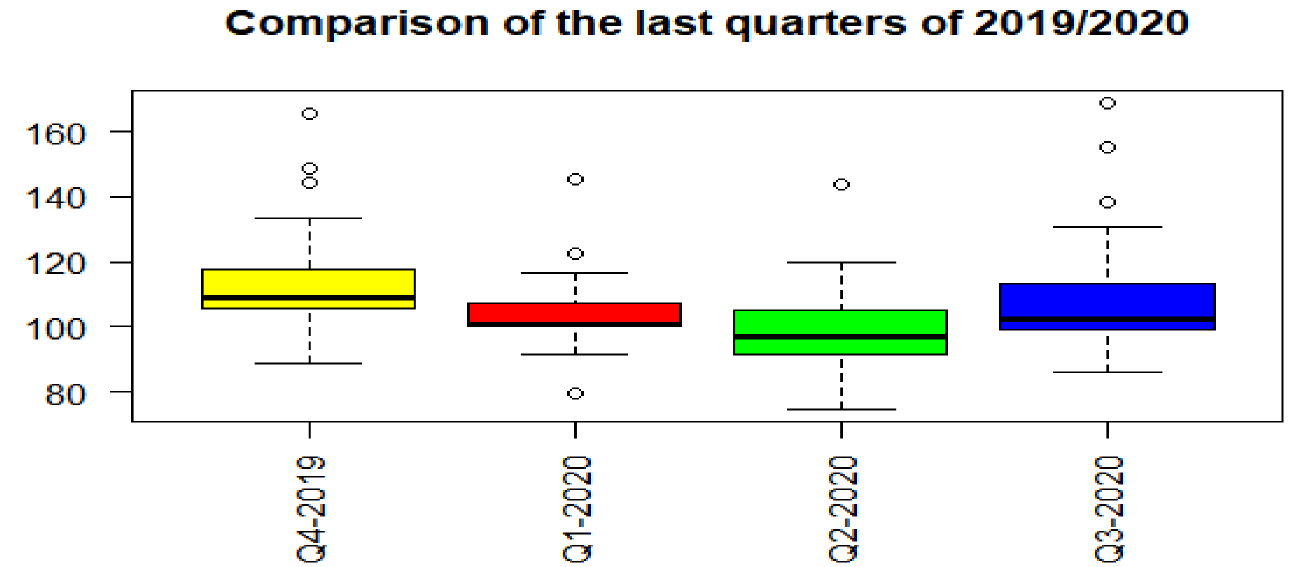
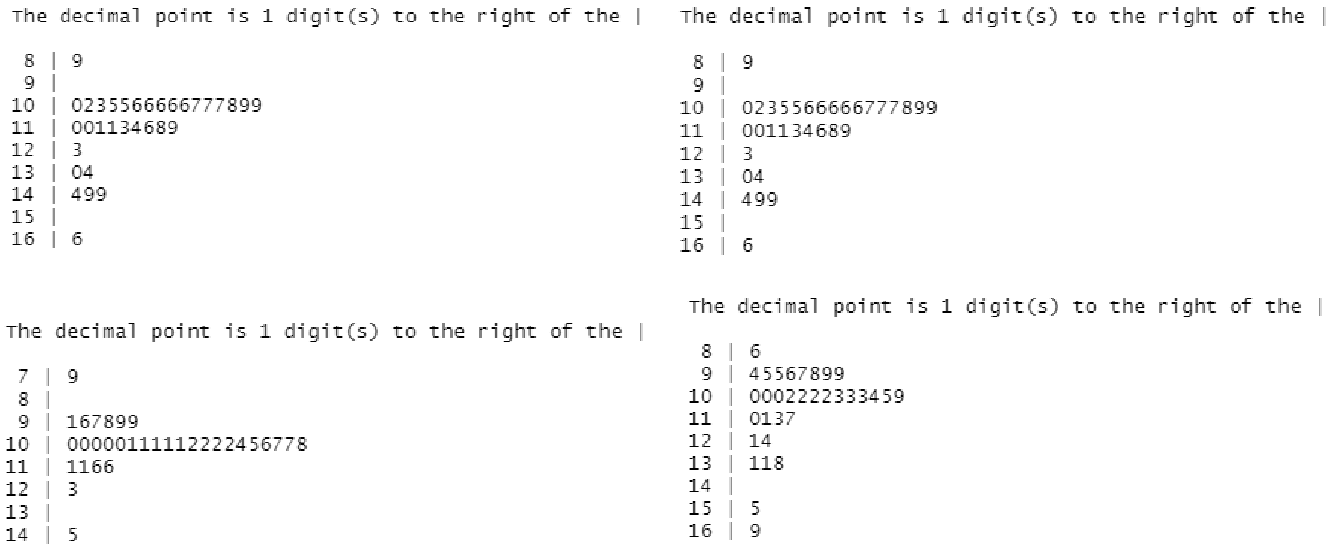


Figure 3: Boxplots of each quarter of Covid-19 period

According to the boxplot, the second quarter is indeed the trimester with the lowest labor productivity: the minimum, the median and the maximum are respectively the smallest among the four quarters.



```
  The decimal point is 1 digit(s) to the right of the |

   8 | 9
   9 |
  10 | 0235566666777899
  11 | 001134689
  12 | 3
  13 | 04
  14 | 499
  15 |
  16 | 6
```

```
  The decimal point is 1 digit(s) to the right of the |

   8 | 9
   9 |
  10 | 0235566666777899
  11 | 001134689
  12 | 3
  13 | 04
  14 | 499
  15 |
  16 | 6
```

```
  The decimal point is 1 digit(s) to the right of the |

   7 | 9
   8 |
   9 | 167899
  10 | 000001111222456778
  11 | 1166
  12 | 3
  13 |
  14 | 5
```

```
  The decimal point is 1 digit(s) to the right of the |

   8 | 6
   9 | 45567899
  10 | 0002222333459
  11 | 0137
  12 | 14
  13 | 118
  14 |
  15 | 5
  16 | 9
```

Figure 4: Stem-and-leaf diagrams (from the left-top to the right bottom corner: Q4-2019, Q1-2020, Q2-2020 and Q3-2020)

Finally, I summarize the whole period and I plot the histogram and the empirical cumulative distribution of the data:

Mean	Variance	Minimum	Maximum	Range	1st quartile	Median	3rd quartile	Interquartile range	Percentile 80	Mode
107.04	197.7069	82.30	151.57	69.27	99.72	102.58	108.95	9.23	114.36	102.575

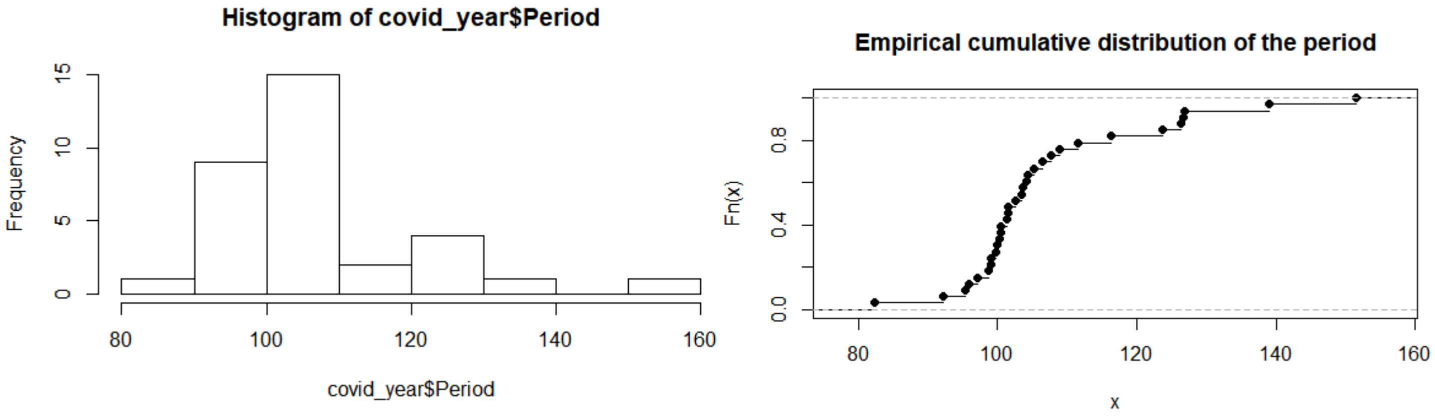


Figure 5: Histogram and cumulative distribution for the Covid-19 period

According to these figures, the distribution of labor productivity during the period of Covid-19 is a discrete distribution.

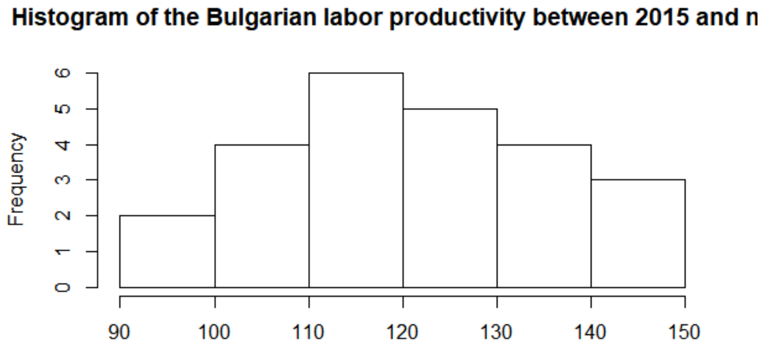
STUDY CASES FROM 2015 TO 2020

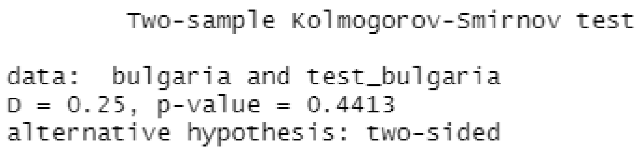
In this part of the project, I consider the case of five countries and the EU: for each one, I discuss about the methods that enable to determine the distributions of random variables. I also present how to estimate the moments of this variable.

Test of hypothesis for distribution

Each country is considered as a discrete random variable: after observing the histogram of these random variable, I perform a hypothesis test for determining the distribution. I chose to perform the Kolmogorov-Smirnov goodness-of-fit test with the function `ks.test()`, which allows to compute the distance between the empirical distribution functions of two samples, when it is two-sided.

Bulgaria

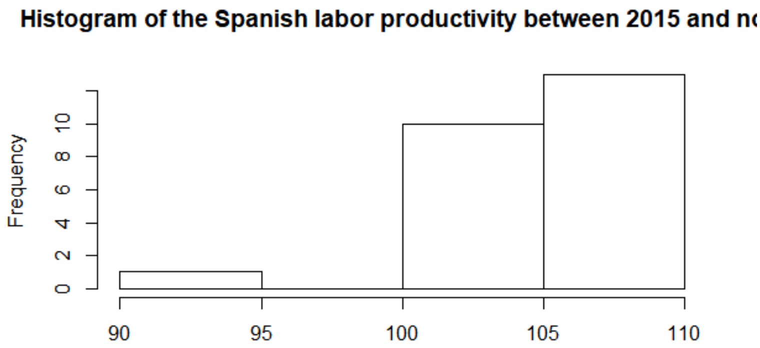


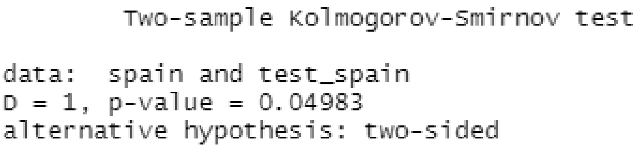


```
              Two-sample Kolmogorov-Smirnov test

data:  bulgaria and test_bulgaria
D = 0.25, p-value = 0.4413
alternative hypothesis: two-sided
```

Regarding the histogram, I performed a test on the variable `bulgaria` for a Poisson distribution of parameter the mean of `bulgaria`. The p-value of the test is above 0.1, which means that the null hypothesis cannot be rejected: the variable indeed follows a Poisson distribution.

Spain

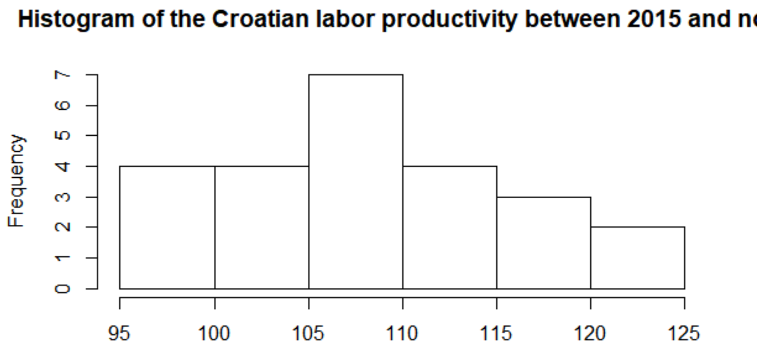


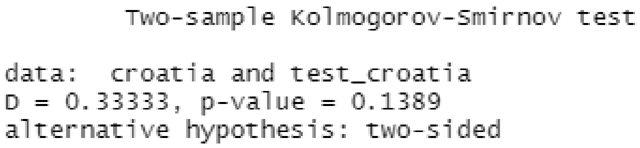


```
             Two-sample Kolmogorov-Smirnov test

data:  spain and test_spain
D = 1, p-value = 0.04983
alternative hypothesis: two-sided
```

Even though the variable is discrete, I performed a test on the variable `spain` a t-Student distribution of parameter $n = \frac{2\times var(spain)}{var(spain)-1}$. The p-value of the test is under 0.05 so we have to reject the null hypothesis.

Croatia

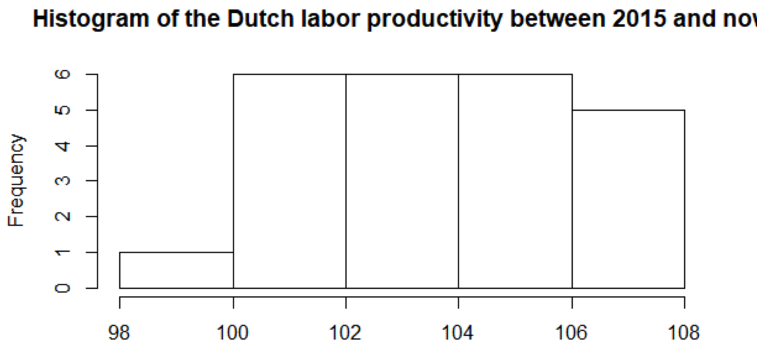


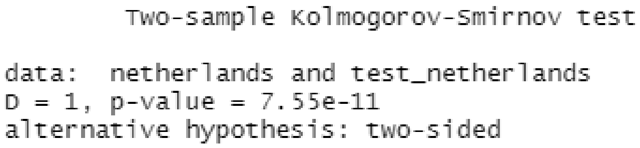


```
             Two-sample Kolmogorov-Smirnov test

data:  croatia and test_croatia
D = 0.33333, p-value = 0.1389
alternative hypothesis: two-sided
```

Regarding the histogram, I performed a test on the variable `croatia` for a Poisson distribution. As for the variable `bulgaria`, the p-value is above 0.1: the variable follows a Poisson distribution of parameter the mean of `croatia`.

Netherlands

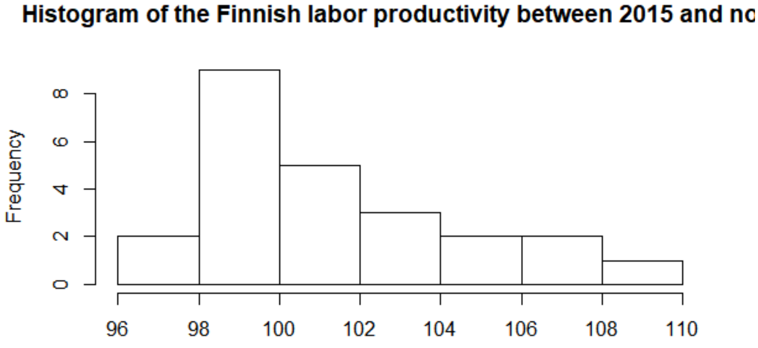


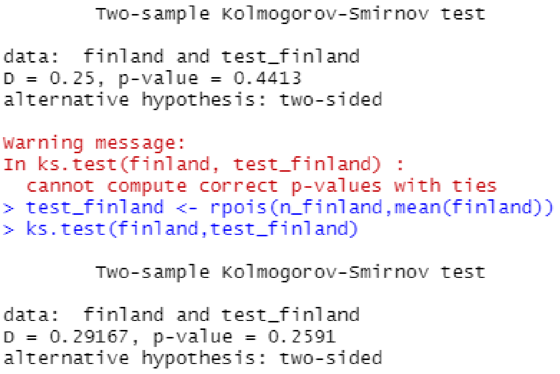


```
             Two-sample Kolmogorov-Smirnov test

data:  netherlands and test_netherlands
D = 1, p-value = 7.55e-11
alternative hypothesis: two-sided
```

Regarding the histogram, I performed a test on the variable `croatia` for a uniform distribution. The p-value is very low so we can conclude that the null hypothesis can be rejected.

Finland



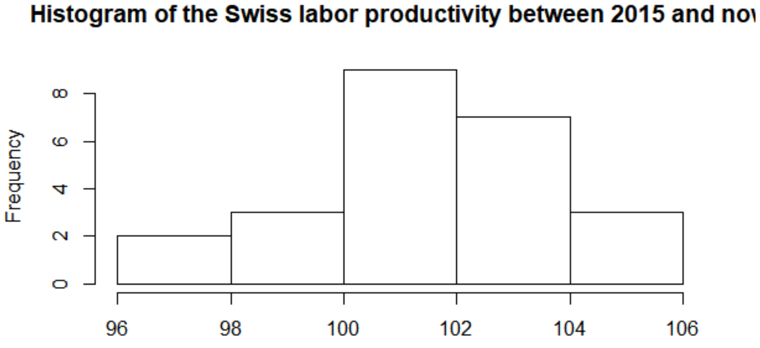


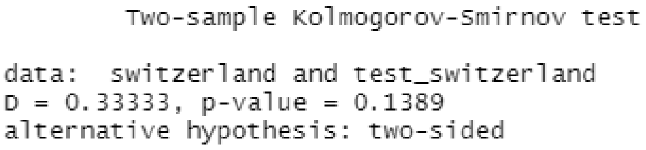
```
        Two-sample Kolmogorov-Smirnov test

data:  finland and test_finland
D = 0.25, p-value = 0.4413
alternative hypothesis: two-sided

Warning message:
In ks.test(finland, test_finland) :
  cannot compute correct p-values with ties
> test_finland <- rpois(n_finland,mean(finland))
> ks.test(finland,test_finland)

        Two-sample Kolmogorov-Smirnov test

data:  finland and test_finland
D = 0.29167, p-value = 0.2591
alternative hypothesis: two-sided
```

For Finland, I performed two tests: the first one was for a geometric distribution of parameter $p = \frac{1}{mean(finland)}$ and the second one for a Poisson distribution of parameter $\lambda = mean(finland)$.

Both tests have a higher p-value than 0.1, but the first test gave a higher p-value than the second one: we cannot reject the hypothesis of `finland` following a geometric distribution.

Switzerland

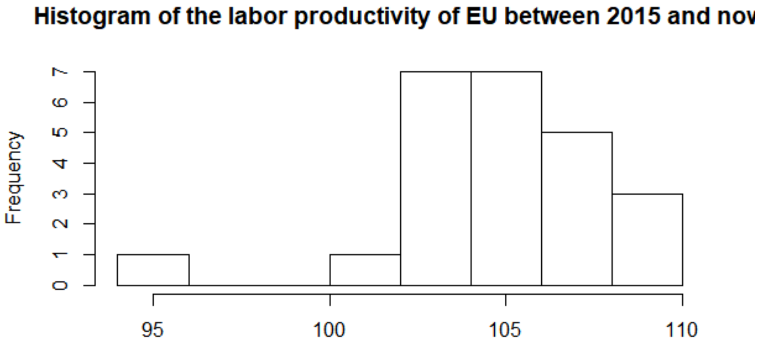


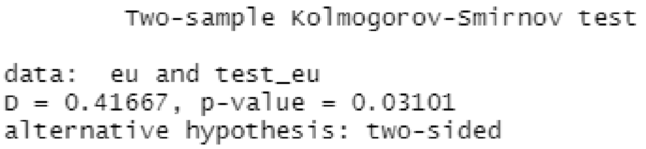


```
        Two-sample Kolmogorov-Smirnov test

data:  switzerland and test_switzerland
D = 0.33333, p-value = 0.1389
alternative hypothesis: two-sided
```

As the same way than Bulgaria and Croatia, we cannot reject the hypothesis of `switzerland` following a Poisson distribution.

European Union





```
        Two-sample Kolmogorov-Smirnov test

data:  eu and test_eu
D = 0.41667, p-value = 0.03101
alternative hypothesis: two-sided
```

I performed a test also for a Poisson distribution, but this time the p-value is under 0.05: we can reject the null hypothesis.

Depending on the country, the distribution of the variable representing the labor productivity is not the same. However, the Poisson distribution came back several times for this small sample of countries. With the central limit theorem, we can assume that for the long term (for a total of minimum 30 quarters or 7.5 years), after accumulating a lot of labor productivity data, the distribution of these variables tends to a normal distribution.

Focus on Switzerland: estimation and confidence intervals

I chose to focus on Switzerland to determine an estimation of the mean, the variance, and the standard deviation and to compute confidence intervals for each moment.

Estimation of the moments

For estimating the moments, I use the property of unbiasedness of the estimators:

$$E[\bar{X}] = \mu \text{ and } E[S^2] = \sigma^2$$

The aim is to determine the distribution followed by the variables $\bar{X}$ and S^2.

To obtain the distribution, I compute each mean and variance for several sets of random samples of same size obtained by resampling. I use the following R script [2]:

```r
n=10000
vector_mean<-rep(0,n)
vector_var<-rep(0,n)
i<-0
while(i<=n){
  resample_swiss<-sample(switzerland,replace=TRUE)
  vector_mean[i]<-mean(resample_swiss)
  vector_var[i]<-var(resample_swiss)
  i<-i+1
}
mean(vector_mean)
mean(vector_var)
sqrt(mean(vector_var))
```

The estimated moments are:

$$\begin{cases} E[switzerland] = 101.616 \\ Var(switzerland) = 4.768795 \\ \sigma_{switzerland} = 2.183757 \end{cases}$$

Confidence intervals

[2] The computational cost of this script is very high but for more precision it is necessary to have a high n

Now the objective is to determine confidence intervals of the three parameters that have been estimated previously. I chose to compute a 95% confidence interval which is large enough to have a least error. Even if we know that the variable may follows a Poisson distribution, I make no assumption on the distribution of the variable for computing the intervals. I just have a sample of the distribution so I can easily get the parameters of the sample. Nevertheless, there is no information about the population parameters.

For this reason, I used Bootstrapping method, which consists in resampling the data to determine the confidence intervals. I used the following script:

```
library(boot)

resample_boot <- sample(switzerland,replace=TRUE)
boot_mean <- function(switzerland, resample_boot){
  mean(switzerland[resample_boot])
}
boot_sd<-function(switzerland, resample_boot){
  sd(switzerland[resample_boot])
}
mean_results <- boot(switzerland,boot_mean,R=2000)
sd_results<-boot(switzerland,boot_sd,R=2000)
boot.ci(boot.out=mean_results,type="all")
boot.ci(boot.out=sd_results,type="all")
```

The confidence intervals I obtain are $[100.8, 102.5]$ for the mean and $[1.701,\ 2.958]$ for the standard deviation.

The moments that have been estimated are in these intervals, I can assume that the estimation has been correctly done.

Consequently, the variable `switzerland` follows the distribution $P(101.616)$.

MACHINE LEARNING MODELLING

The statistical tools enable to get information about the distribution of random variables and to estimate parameters, particularly the expected value. This section presents two methods of machine learning modelling to make prediction on the future values.

For both type of model, I must split the data into training and testing sets: the first one is for creating and training the model and the second is for testing and assessing it. I choose training sets containing 80% of the values of the total set.

Linear regression

The purpose of this linear regression is to predict the fourth quarter of a year, knowing the three previous quarters. The first one is called target and the latter are the predictors.

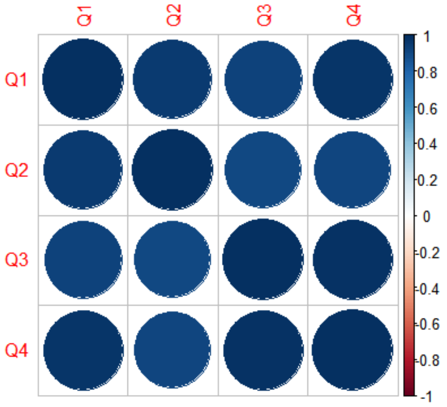
Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables)[3].

I performed two linear regression: one for Switzerland and one for the 29 countries.

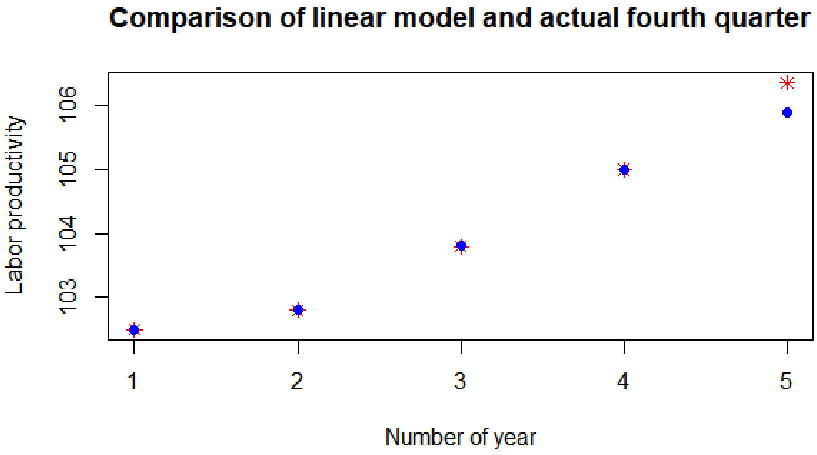
Case of Switzerland

I created a dataset containing each first, second, third, and fourth quarter of Switzerland, from 2015 to 2020: the dataset has 5 rows and 4 columns.

Before creating the model, I studied the correlation between each column. I obtained the following correlation matrix:



The variables are highly linearly correlated, so I can perform a linear regression. For this purpose, I used the function `lm()` to construct the model, fit it to the data and training it. The coefficients I obtained allow me to have the following plot:



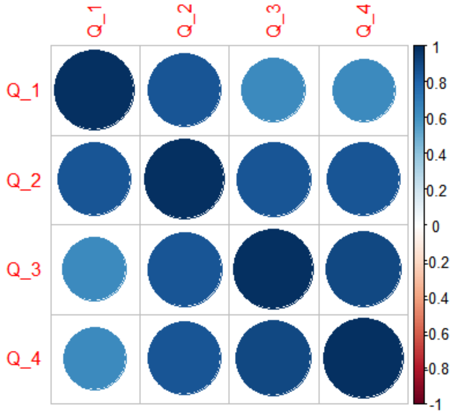
[3] Definition from Wikipedia

The model does not fit very well for the last datum, which is the unique value of the training set. In fact, the dataset does not contain enough information to construct a precise model. Using the estimated coefficients, I compute a prediction of the fourth quarter of 2020, and I obtained $Q4 - 2020 = 106.6638$. The prediction interval associated, determined with the function `predict()` is $[101, 110]$.

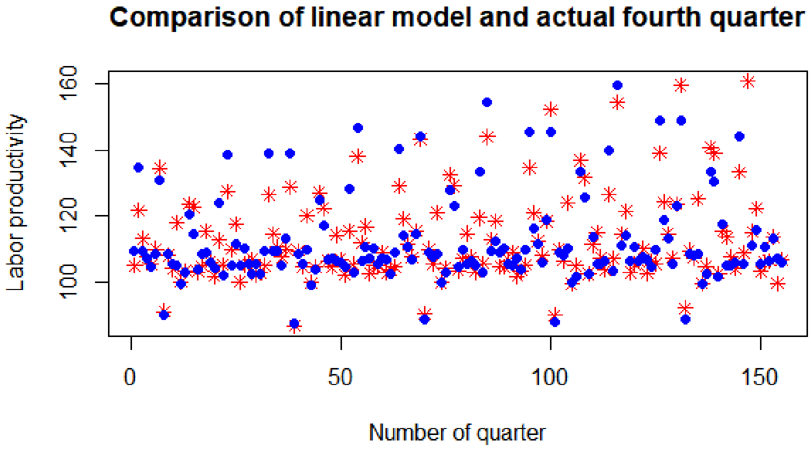
General case

I created a dataset containing each first, second, third, and fourth quarter of the 29 countries, from 2015 to 2020, which means that I have much more data to construct the model: the dataset has 155 rows and 4 columns.

I evaluated again the correlation between the columns:



The variables are also highly linearly correlated, so I can perform a linear regression. I repeated the same procedure than previously, and I obtained:



This plot does not provide relevant information because there are too many points.

The prediction for the fourth quarter of 2020 and the prediction interval are respectively 101.1693 and $[90.1, 113]$.

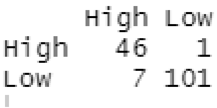
Classification

The mean of labor productivity is 108. I distinguish two classes of data: if the annual labor productivity is above the mean the country is considered as high productive and if not, it has a low productivity. There are several algorithms for classification, but I chose to focus on the simplest one: logistic regression. Logistic regression allows to construct a binary classifier. It uses the logistic (or logit, or sigmoid) function defined as:

$$g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

This function defines the threshold of the classifier: when $g(\theta^T x) < 0.5$, the country has a low labor productivity and when $g(\theta^T x) \geq 0.5$ the country has a high labor productivity.

I used the function `glm()` to construct the logistic regressor. After training and testing the model, I make a prediction for a value that I added in the dataset. Finally, I computed the confusion matrix, which is a metric that indicate the true and false positive and the true and false negative. I obtain the following matrix:



```
      High Low
High    46   1
Low      7 101
```

The actual values are the columns, and the predicted values are the rows: the numbers of true positive and true negative are both high, which means that the model is performing well.

Model assessment

For regression as for classification, I can determine whether a model fits with the data by performing a hypothesis test on the coefficients of the model.

These coefficients should minimize the Root Mean Squared Error (RMSE) of between the sample and the actual value, which is the sum of the variance and the squared of the bias.

For each model, I compute the Residual Sum of Squares (RSS), the Mean Squared Error (MSE) and the RMSE:

Model	RSS	MSE	RMSE
lm_swiss.fit	0	0	0
lm_europe.fit	3693	29.8	5.46
glm.fit	4.66e-25	3.76e-27	6.13e-14

The first regression minimizes the error, so the parameters of the model are optimal. For the classification, the error can be neglected because it is very small: the model also fits.

Conversely, the parameters of the second regression do not minimize the RMSE. To improve the performance, it is possible to add a regularizer in the construction of the model.

CONCLUSION

As a conclusion, the statistical study of the “Labour productivity and unit labour costs” dataset allows to determine that depending on the country, the variation of this economic metric is not the same. Using the machine learning techniques, I have been able to predict the next labor productivity. However, the models I constructed can be improved in various ways: for instance, if the values were monthly instead of quarterly, I could have more data and thus a smaller error for the test set. Another improvement can be the use of more ancient data.

REFERENCES

[1] Eurostat public database (2020)
Available at: https://ec.europa.eu/eurostat/web/covid-19/data
[2] Labor Productivity, Investopedia (2020)
Available at: https://www.investopedia.com/terms/l/labor-productivity.asp
[3] *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (Second edition) by Aurélien Géron, O’Reilly (2019)
[4] Statistical Modelling, Pedro J. Zufiria (2020)