TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Birte Hellwig
Dr. Paul Wiemann
M. Sc. Hendrik Dohme

Author: Amirreza Khamehchin Khiabani

Group number: 10

Group members: Amirreza Khamehchin Khiabani, Rama Kassoumeh, Javiera Riffo Torres

November 11, 2021

Contents

1 Introduction

Accessing the International Data Base (IDB) of the U.S. Census Bureau provides us information regarding various geographical data(from 1950 to 2060) of all regions and subregions of the World by the U.S. Department of State. For database contains, information from state institutions including censuses, surveys, or administrative records, and estimates and projections from the U.S. Census Bureau are concerned. This Project dataset is a csv file, which is applied during the whole project. This excel file includes life expectancy at birth and total fertility rates for 228 countries for 2001 and 2021.The purpose of this project is to employ statistical methods to solve four tasks regarded to obtain more information from these raw data. In Section 2, those problems are represented to find a better resolution of the data and the information, which provides better solutions to tackle these issues. In Section 3, the methods applied for the statistical analysis are introduced and more details are represented. Additionally, the method of frequency distributions of the variables and Bivariate correlations are defined in a statistical way. The main differences between homogeneous and heterogeneous are proposed in this part. Furthermore, the granted statistical methods are applied to the presented dataset and the results are deciphered with statistical plots, particularly histogram. The final part contains a summary of the conclusions, the information and relationships are obtained from our raw data, and new data for which we can use further projects.

The statistical analysis of the data contains four parts. The first three, main goals are for the year 2021. For this year, analyzing the frequency distributions of the variables and their bivariate correlations are examined. For the third task, it is investigated whether the variables are homogeneous within regions and heterogeneous between them. In the end, the last part of the statistical analysis deals with the change in the values of the variables during the 20 years, examining the 2001 values with those of 2021. The main statistical methods employed in investigating frequency distributions are frequency distribution graphs for discrete variables and histograms for continuous numeric variables with including statistical measures for better understanding data distribution. By employing scatter plots and computing correlation coefficients for two variables at a time, bivariate correlations are examined. For the third part, box plots and the interquartile range of the variables for each of the subregions are applied to traverse homogeneity within them. Heterogeneity between the subregions is studied by the median values of the variables for each region. Similarly, to compare the values of the variables in 2001 to

those in 2021, another time histograms are applied for difference between distributions of data between 2001 and 2021 for each of the variables.

2 Problem statement

The International Data Base has been presenting demographic data for the world and on all states and regions population recognized by the U.S Department of State, which has a population of 5000 or more. This subset of dataset from the International Data Base is constituted of 9 variables, containing 228 countries and additional column GENC(FGDC Structure and Federal Agency and Bureau Representation), representing the names of countries in two capital letters. Moreover, countries are split among 21 subregions and 5 regions and with information of the total fertility rate, female life expectancy, male life expectancy, and the life expectancy for both sexes, granted for two years, 2001 and 2021. The whole dataset has 456 records for the two years mentioned. Country, GENC, subregion, and region are nominal variables, whereas year is a discrete numeric variable with two values 2001 and 2021. The values of the total fertility rate are provided with one to four decimal places. By contrast, values of life expectancy for both sexes, life expectancy males, and life expectancy females are shown with two decimal places. These late values are measured by year unit, starting from birth time to death. The data quality is consistent, but there are some missing values particularly for 2001 including 6 countries or territories. For the first three tasks, only the values of 2021 are concerned and no data value is missing for statistical analysis. For the fourth part of the analysis, missing values for 6 countries or territories of 228 of 2001 is more than 2 percent, which is considered ignorable.

The total fertility rate is defined as the average number of children a woman has, assuming that current birth rates remain constant throughout her childbearing years which are usually considered to be between ages 15 to 49. On the other hand, the life expectancy at birth is the average number of years that a newborn is expected to live if current mortality rates continue to apply.

The total fertility rate is described by asking a woman that the number of children she has, with additional information like remaining birth rates during her childbearing years, usually regarding ages 15 to 49 (The United States Census Bureauu). In contrast, the life expectancy at birth is the average number of years that a newborn is assumed to live if current mortality rates continue to apply. If birth dates are not

available, estimation for statistical databases is applied(International Programs, Population Division, U.S. Census Bureau, p. 15-16). In more developed countries, basic data are obtained from vital registration systems and official estimates of life expectancy at birth derived by national statistical offices are employed for the life expectancy at birth. Nationally-representative household questionnaires, censuses, and essential registration data for most statistically less developed countries, the U.S. Census Bureau develops base mortality estimates adopting a composite of data sources (International Programs, Population Division, U.S. Census Bureau, p. 11-12). The purposes of this report are to define the frequency distributions of the variables and besides involve both sexes and their bivariate correlations for the year 2021. Furthermore, homogeneity within subregions and heterogeneity between those values in the individual subregions and then compare the values between different subregions. In the final step, variations changing from 2001 and 2021 are determined.

3 Statistical methods

In Statistical methods section, various types of statistical methods are represent and they are applied in different cases. All calculations and visualizations of plots is applied via the software R (Version 4.1.0, R Development Core Team).

3.1 Basic statistical tools

Mean is considered as a measure of location or central value for a continuous variable. The mean is mostly performed where the data is symmetrical and without outliers. To make it more precisely, consider a sample of n observations $x_1, x_2, x_3, ..., x_n$ the mean is given by this formula: This is a formula:

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

(B. S. EVeritt and A. Skrondal, p. 274)

Rth moment of X as a random variable can be defined as follow if the expectation value does exist:

$$\mu_r^{'} := E[X^r]$$

(Alexander M. Mood and Franklin A. Graybill, p. 73)

By definition of mean and moment, variance is the second moment about the mean. An unbiased estimator of the population value is provided by s^2 can be formulated by the formula below, where $x_1, x_2, ..., x_n$ are the n sample observations and $\bar{x}$ is the sample mean.(B. S. EVeritt and A. Skrondal, p. 445)

$$\sigma^2 := \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Following of this method, quartile is defined as a division of a probability distribution or frequency distribution into equal, ordered subgroups, for instance, quartiles or percentiles (B. S. EVeritt and A. Skrondal, p. 348) .

Mode is the most frequently occurring value in a set of observations. Occasionally used as a measure of location. See also mean and median.(B. S. EVeritt and A. Skrondal, p. 281)

One of the most basic definition is median.This splits the data into two parts of the same size. When the number of observations is even, the measure is computed as the average of the two central values. When there is an odd number of observations the median is the middle value. It grants a measure of the location of a sample, which is fit for asymmetric distributions and is relatively insensitive to the presence of outliers. Median is defined as second quartile of n observations(B. S. EVeritt and A. Skrondal, p. 275).

$$m(x) = x_{n+1/2}$$

n odd,

$$m(x) = 1/2(x_{n/2} + x_{n+1/2})$$

Otherwise. The first and third quartile are defined q_3 and q_1 (or $q_0.75$ and $q_0.25$) , respectively. The range between these two quartile is defined the interquartile range.

$$IQR = q_3 - q_1$$

(B. S. EVeritt and A. Skrondal, p. 220)

the minimum value, 1st quartile, median, 3rd quartile, and the maximum value are all called Five-number Summary (B. S. EVeritt and A. Skrondal, p. 169).

Skewness is the lack of symmetry in a probability distribution. Usually quantified by the index, s, given by

$$s = \frac{\mu_3}{\mu_2^{3/2}}$$

where µ2 and µ3 are the second and third moments about the mean. The index takes the value zero for a symmetrical distribution. A distribution is said to have positive skewness when it has a long thin tail to the right, and to have negative skewness.(B. S. EVeritt and A. Skrondal, p. 397)

3.2 Graphical tools and their measures

Box plot is a graphical method, displaying the five-number summary of a set of observations. The interquartile range and the 'whiskers' extend to 1.5 times the interquartile range on both sides of the box is covered. Additionally, points outside the whiskers are called outliers.(B. S. EVeritt and A. Skrondal, p. 61) .

Another graphical method, histogram, illustrates a set of observations. Class frequencies are represented by the areas of rectangles centred on the class interval. If the following are all equal in their value, the heights of the rectangles are also proportional to the observed frequencies (B. S. EVeritt and A. Skrondal, p. 206).

Correlation Coefficient is a parameter quantifying the linear relationship between two variables by Consideration of its value between 1 and 1 inclusively. Its sign indicates the direction of the relationship in addition to the numerical measurement. The value of zero indicates the lack of any linear relationship (B. S. EVeritt and A. Skrondal, p. 107).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

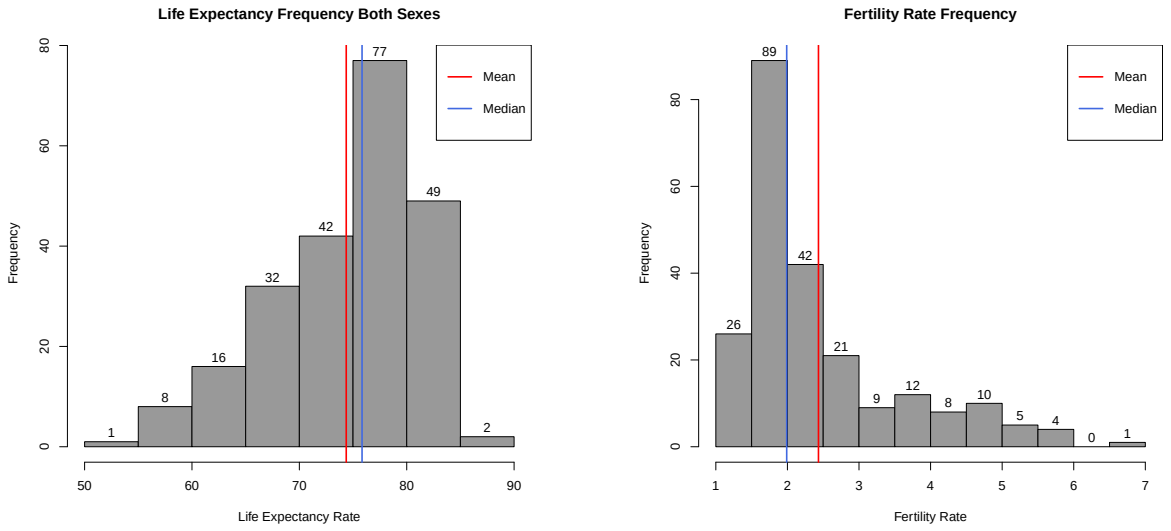
$\bar{x}$ and $\bar{y}$ are means of x and y as random variables, respectively.

4 Statistical analysis

In this section, the statistical methods described in the previous section are implemented to the dataset and the four tasks are solved and interpreted. Subsections 4.1 to 4.3 only consider values of 2021 whereas for subsection 4.4 both values in 2001 and 2021 data are indicated.

4.1 Frequency Distributions of the Variables in 2021

Frequency distribution presents us with a shortened grouping of data divided into mutually exclusive classes and the number of events in a class. Figure 1 shows two histograms with frequency of life expectancy rate and fertility rate.

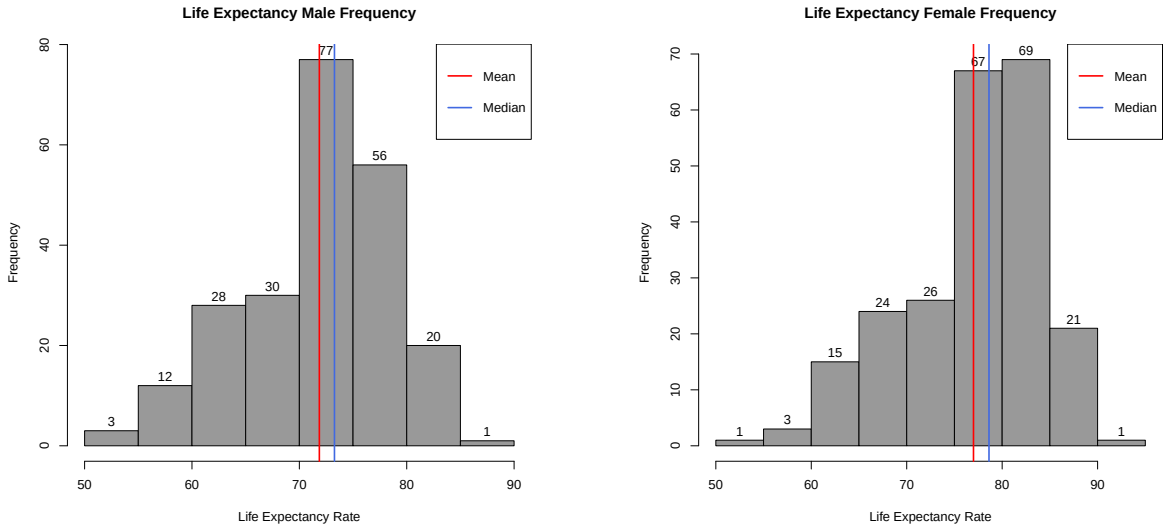


(a) (b)

Figure 1: Frequency distribution for life expectancy rate and fertility rate

It could be described that for the first figure(life expectancy frequency) x-axis is divided into 8 bins, which represented different age groups of life expectancy of both sexes. It could be easily interpreted that groups of ages between 70 to 85 have the most frequency among all countries or territories. By contrast, age groups below 65 and above 85 to 90, which is the oldest group among all of this deviation, have smaller frequencies in compare others. Total 77 countries or territories belong to the highest frequency while only one country or territory belongs to below 55, which is both frequency and least age group, respectively. Mean and median are both less than the mode. The mean is less than the median(with values of 74.36 and 75.84 for mean and median, respectively), which makes the histogram skewed to the left (The exact values of mean and median, other quartiles, and variance for each variables and distributions are listed in Table 1 and 2 in Appendix). Another figure(fertility rate frequency) x-axis is partitioned into 12 bins. The fertility rate frequency is far higher in the group between 1.5 to 2 by value 89. Others have similar values, only subgroup 2 to 2.5 has significant value among others. Mean and median values are both more than the mode. The Median is less than the

mean, which makes the histogram skewed to the right. Figure 6 shows two histograms with frequency of life expectancy males and females.



(a) (b)

Figure 2: Life expectancy male frequency versus life expectancy female frequency

For both figures life expectancy male and life expectancy female, the x-axis is separated into arbitrarily 9 bins for different age groups of life expectancy and the y-axis is represented as the number of frequency. The interpretation describes that groups of ages between 70 to 80 have the most frequency among all males, whereas for the next plot age groups with values from 75 to 85 have the most height on the plot. Mean and median and mode are so near together for life expectancy for males. The plot is not skewed, and distribution is Uniform. The same description is for the life expectancy female, mean and median have the similar values according to their lines with mode. Skewness is not like previous histograms. However, it could be slightly considerable. The mean is more than the median, which makes the histogram skewed to the left (2.43 and 1.99 for mean and median, respectively). We can interpreted from two plots that females have more life expectancy than males and frequencies for this sub age groups as mentioned above are the same values. Information regarding the histograms and their measures for other variables are available in Appendix.

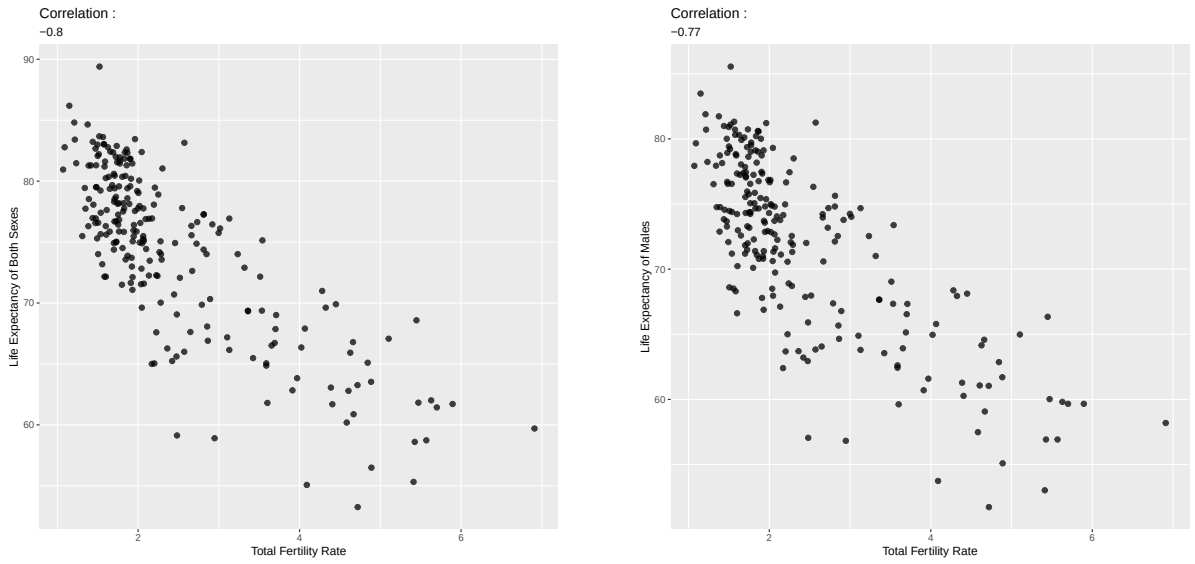
4.2 Bivariate analysis between the variables

Figure 3 part (a) presents the correlation between the life expectancy of both sexes and the total fertility rate. The correlation coefficient for the two variables is 0.80, which indicates they have a strong negative correlation. In general, the higher the life expectancy for both sexes, the lower the fertility rate. Other scatter plots regarding the life expectancy of males and females concerning fertility are prepared in Appendix. Typically, when total fertility reaches 3 children, it can be seen that life expectancy is below 70. They are those countries that life expectancy is not as higher as other parts of the words and the rate of bearing is higher than average. One interesting point is Monaco, which has the highest life expectancy in both sexes. Another points represented for Afghanistan and Niger, which shown the lowest life expectancy and highest fertility rate, respectively, are concerned. The correlation is not so much strong on the right part of the graph, which are most countries from Africa continent. By contrast, countries from Europe and East Asia are more correlated on the upper left side. Correlation Coefficient is calculated by its formula, which is described in previous section. Table 1 of values of correlations for numerical variables (a), (b) , (c), and (d) correspond to fertility rate, life expectancy both sexes, life expectancy males, and finally life expectancy females is represented.

Table 1: Values of the correlations

-	(a)	(b)	(c)	(d)
(a)	1.00	-0.80	0.77	0.82
(b)	0.80	1.00	0.99	0.99
(c)	0.77	0.99	1.00	0.97
(d)	0.82	0.99	0.97	1.00

As can be seen, the correlations for life expectancy for total fertility rate and life expectancy for males stay approximately the same with the value of the both sexes, Showing strong enough correlations between two values that are already concerned. Another Correlation that would be possible to be interpreted is the correlation between life expectancy between males and females, which is strong enough(near to 1 and positive). Hence, instead of examining the male and female life expectancy individually, the life expectancy of both sexes would be enough. Other figures of this sections are represented in appendix.



(a) (b)

Figure 3: Scatter plots and correlations, a)fertility rate and the life expectancy of the both sexes and b)fertility rate and life expectancy of males

4.3 Homogeneous within subregions and Heterogeneous between different subregions

In this section, homogeneous within subregions and heterogeneous between different subregions are examined. First, the fertility rate in each subregion is described. It can be seen that approximately all subregions of Europe, America, and interestingly Australia/New Zealand subregion box plots are not significantly widened in comparison with other subregions. As a result, they are more homogeneous within themselves, which means a smaller interquartile for each boxplot and quartiles are so close together. By contrast, all subregions belonging to Africa except the southern part and West Asia have the least homogeneity. Middle Africa has the largest interquartile range and consequently the least homogeneity within itself among all boxplots and therefore subregions. Figure 4 shows its related figure.

Describing heterogeneity, subregions belonging to Europe and America have the least heterogeneous values and their interquartile are approximately computed the same value. Even though there are some outliers for Southern Europe. On the other hand, for Oceania, except Australia/Newzealand, homogeneity is considerable. However, heterogeneity is significant between Africa subregions in the same continent and additionally with other subregions all around the world. By analysing the level of median of each subre-

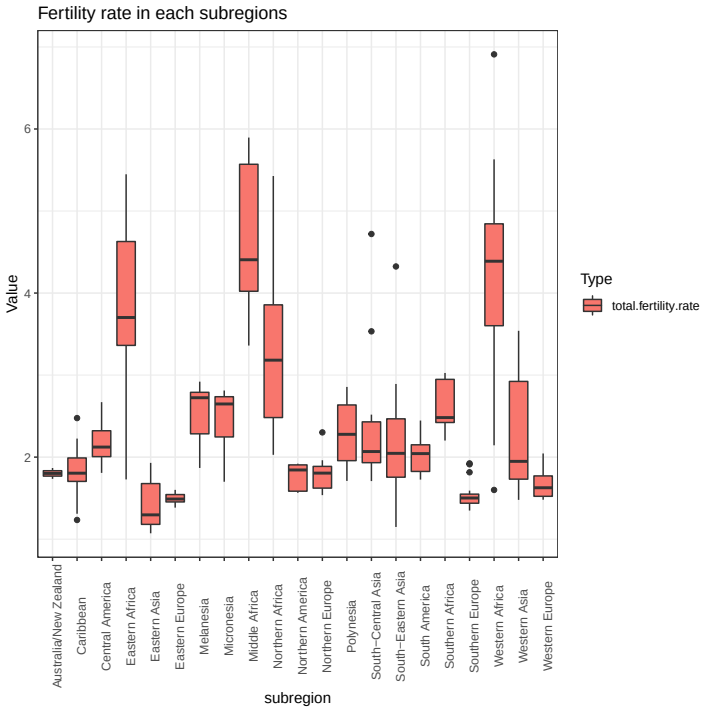


Figure 4: Figure of box plots related to individual subregions for analysing homogeneity and heterogeneity for fertility rate

gion and comparing their results with others on the figure, African subregions have the most heterogeneous to each other and other subregions.

In the next figure, life expectancy for both sexes, Australia/New Zealand, Northern America, Northern Europe, and West Europe have the most homogeneous(with variance value). In comparison, Eastern Asia, Northern Africa, SouthEastern Asia have the least values of homogeneity with the interpretation that mentioned above for the previous figure by using techniques of interquartile. Heterogeneity on the other side for this figure could be described that Australia/New Zealand, Northern America, Northern Europe, and Western Europe are homogeneous to each other (using the comparison the median level of each subregions). Other subregions are the most heterogeneous between themselves. African regions are more heterogeneous in comparison to other subregions from other continents Since the value of median for life expectancy for African subregion is much lower than median values for other subregions of continents.

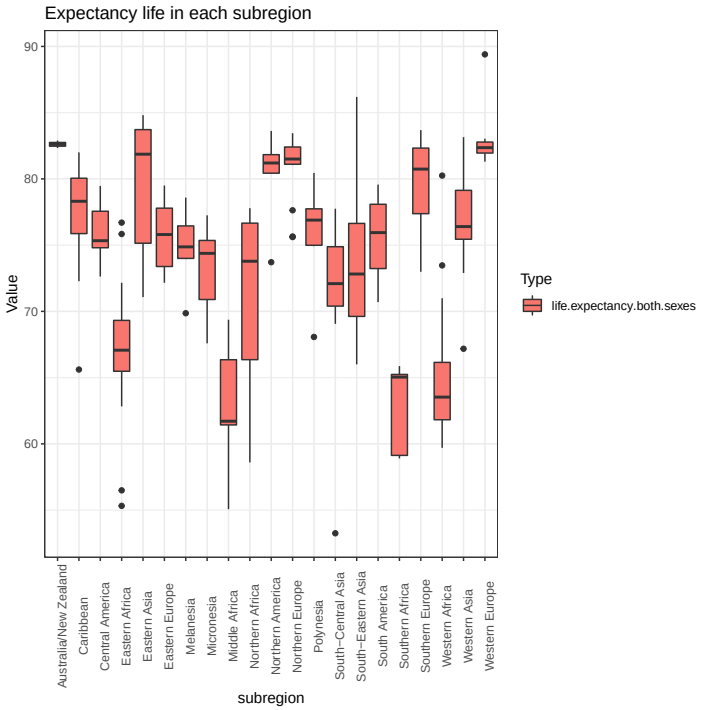
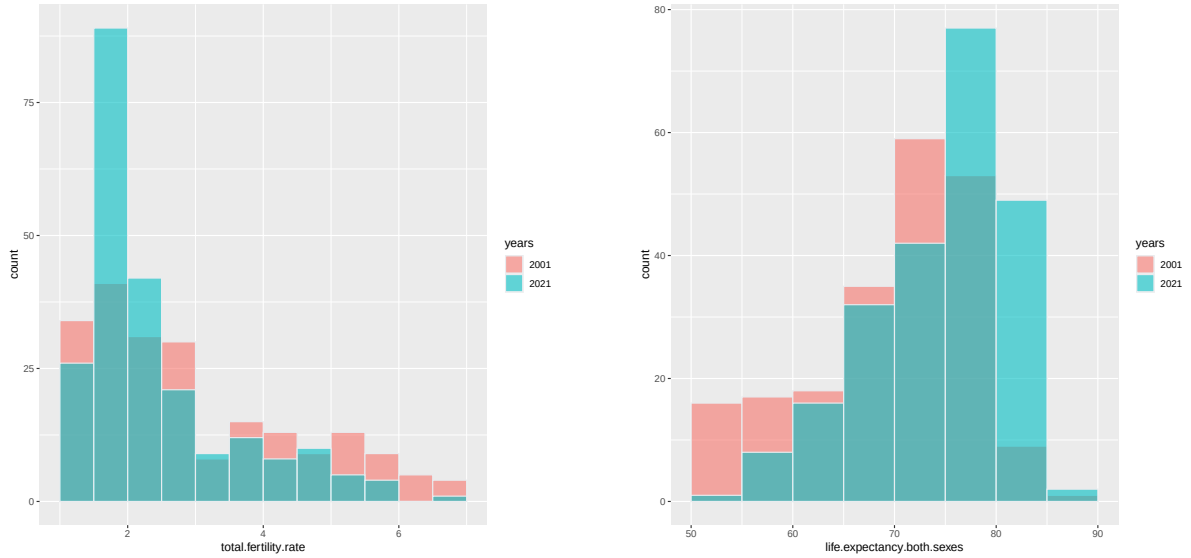


Figure 5: Figure of box plots related to individual subregions for analysing homogeneity and heterogeneity for life expectancy for both sexes

4.4 Comparing the values of the Variables in 2001 with 2021

For this task, the differences between the values of the fertility rate of 2001 and 2021 and the life expectancy of both sexes separately are considered. In the problem statement section, information regarding missing values is described. However, in 2021 those values are filled with the new data. The first figure illustrates that the frequencies of groups over 4 for fertility rate(having more than 4 children) and most other groups of the figure have been decreasing over these 20 years. Consequently, the most frequencies in this range can see the decline in the number of countries. In contrast, there is a significant increase in the amount of 1 to 2 children, more than 200 percent rise in the frequency, which interpreted that most countries and their people tend to have around two children. Statistically point of view, the skewness reminded the same during this gap.



(a) (b)

Figure 6: compare two frequency distributions of fertility rate and life expectancy for both sexes of 2001 and 2021

In comparison to the figure related to life expectancy for both sexes, although first age groups below 75 have seen a reduction in their frequencies, a notable rise are shown in the plot for age groups from 75 to 85. It is desirable to specify that there is an escalating in height for the age group 85 to 90. As a result, life expectancy has been increasing over 20 years for all countries around the world. For this graph skewness shifted right.

5 Summary

The goal of this report is to describe and use statistical methods for the analysis of a dataset represented by the International database. The data includes 9 variables from 228 countries and two different years, 2001 and 2021. The total fertility rate, life expectancy for both sexes, male life expectancy, and the life expectancy for females are numerical values corresponding to different years. In the first task, analysing the frequency distributions of the variables for the year 2021 is performed.

The fertility rate frequency is skewed right. However, life expectancy for both sexes is a left-skewed distribution. It is shown that for life expectancy for males and females, approximately the degree of asymmetry observed in a probability distribution of them is around zero. Following the first task, in task two bivariate correlations between the

variables are examined for the year 2021. It is illustrated by numbers and plots that the correlation between the life expectancy of both sexes and fertility rate is strong enough to interpret and divide countries for its correlation. In particular, the correlation of life expectancy of males and fertility rate is shown that its value is approaching both sexes correlation and it can be considered for females as well.

In the third part, the homogeneity within the subregions and heterogeneity between them for particularly 2021 is examined. In general, for the female fertility rate, European, American, and Austria/Newzealand subregions are more homogeneous within themselves. However, all subregions belonging to Africa except the southern part and West Asia are widened within and less homogeneous. Heterogeneity is notable between Africa subregions and additionally with other subregions all around the world. Moreover, for the life expectancy of both sexes for 2021, Australia/New Zealand, Northern America, Northern Europe, and West Europe have the most homogeneous. African regions are more heterogeneous in comparison to other subregions from other continents.

In the last part, the data for the years 2001 and 2021 is differentiated. The frequencies of groups over 4 children for fertility rate and other groups declined after 20 years. However, there is a significant increase in the amount of 1 to 2 children, explaining that most countries and their people prefer to have almost two children. Comparing life expectancy for both sexes between these two years, first age groups under 75 experienced a decrease in the number of countries, age group over that are displayed have an increase in their values. In conclusion, life expectancy is much higher this year for all countries. For the final overview, by applying statistical methods and their measures, and proper graph, it is decided to have better results and views about the dataset and the four tasks. For instance, applying histograms performs these possibilities to have a more approving view over distributions and scatterplots and boxplots for revealing the correlations and homogeneity, respectively. Since two different times are used for task 4, it might be possible to mention that after these years, these results could be helpful for further decisions of every government for tackle their problems regarding these topics.

Bibliography

[1] Alexander M. Mood and Franklin A. Graybill. *Introduction to the Theory of Statistics*. McGraw-Hill,Inc, 1325 Avenue of the Americas New York City, US, 1974.

[2] B. S. EVeritt and A. Skrondal. *THE Cambridge Dictionary OF Statistics*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK, 2010.

[3] FGDC Structure and Federal Agency and Bureau Representation. Fgdc endorses geopolitical entities, names, and codes (genc) standard edition. `https://www.fgdc.gov/standards/news/GENC/`. (visited on 06/25/2020).

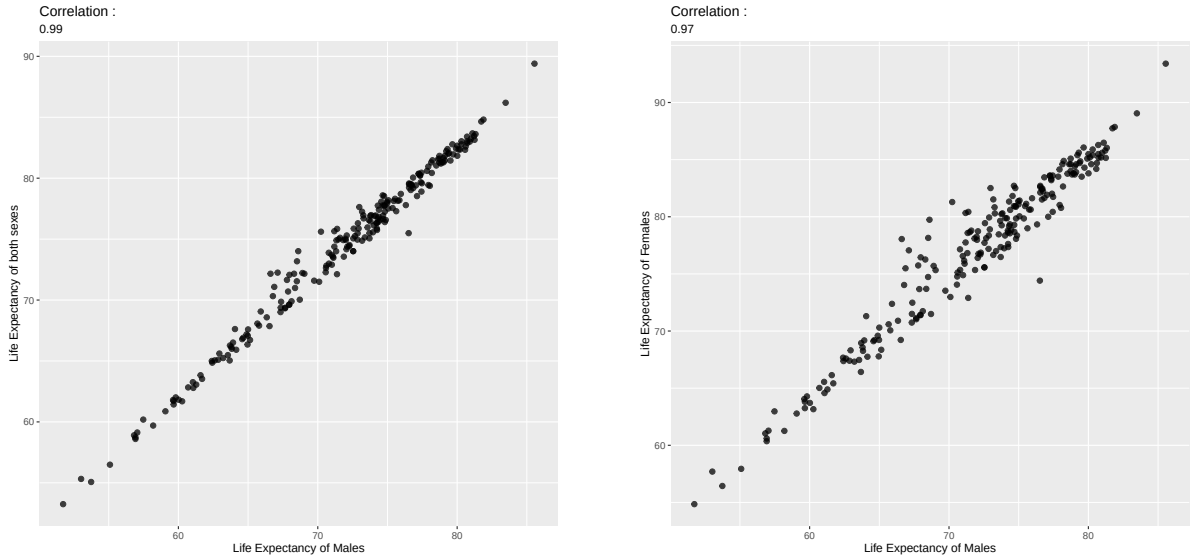
[4] International Programs, Population Division, U.S. Census Bureau. International data base:population estimates and projections methodology. `https://www2.census.gov/programs-surveys/international-programs/technical-documentation/methodology/idb-methodology.pdf/`, Last Updated: December 2020. (visited on 06/25/2020).

[5] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[6] The United States Census Bureauu. The united states census bureau, about. `https://www.census.gov/topics/health/fertility/about.html/`. (visited on 06/25/2020).

Appendix

A Additional figures



(a) (b)

Figure 7: Scatter plots of expectancy life of both sexes versus males and expectancy life of males versus females

B Additional tables

Table 2: Values of the descriptive distributions of variables

Variables	Mean	Standard deviation
Total Fertility Rate	2.43	1.12
Life Expectancy Both Sexes	74.36	6.92
Life Expectancy Males	71.86	6.75
Life Expectancy Females	76.99	7.21

Table 3: Values of the descriptive distributions of variables

Quantile of Variables	q_0	q_1	q_2	q_3	q_4
Total Fertility Rate	1.07	1.70	1.99	2.81	6.90
Life Expectancy Both Sexes	53.25	69.96	75.84	79.45	89.40
Life Expectancy Males	51.73	67.65	73.27	76.94	85.55
Life Expectancy Females	54.85	72.43	78.62	82.43	93.40