Individual Assignment 1

Ali Rafieepouralavialavijeh, 20900871

Quick Summary of the Dataset:

WHO periodically publishes datasets about various health issues around the world. In pair assignment 2 and in this assignment, the pivotal research question is ***whether the ratio of expenditure to budget of each country has a correlation with mortality rate of the disease in that country***. To figure out the answer and for making grounds for further investigation certain datasets are chosen including *TB_Budget* (because it contains data about budgets of each country throughout years 2018-2019), *TB_Expenditure* (because it contains data about expenditure of each country throughout years 2017-2019), and *TB_burden_countries* (because it contains estimates of infection and death of each country in years 2017-2019).

Based on the TB Dictionary (the dataset including the definitions of all columns and variables) and the question we are concerned with, definition of selected variables for this study is as follows: - *budget_tot*: Total budget required (US Dollars, numeric) - *exp_tot*: Total actual expenditure (US Dollars, numeric) - *e_mort_100k*: Estimated mortality of TB cases (all forms) per 100,000 population (numeric) - *e_inc_100k*: Estimated incidence (all forms) per 100,000 population .Further looking into the datasets, it is now noticeable that studies are done in different years and therefore the resulting dataset lacks data for the year 2017 for certain countries. Here, I decided to remove the year 2017 to make a consistent dataset. The type of each column perfectly matches the tests that we are going to hold and therefore need no change. The next step is to check whether there are certain rows that are abnormal or in other words outlier, and as a result highly affect the result of our study. As such, I decided to remove the outliers to reduce their effect on our result [please see details at Appendix 1].

Planning:

To start with our analysis, we need to create two new columns. The first is the ratio of expenditure to budget (*exp_budget*), this shows how much more or less in comparison to the budget allocated to each country is spent on this disease. We need this column because making a comparison between different countries which come with different populations and budgets would otherwise be impossible due to the difference in the scale. The second variable is the ratio between number of deaths in 100,000, and number of incidents in 100,000 cases (*mort_inc_100k*). This variable is important to our analysis because the number of instances in each country loses its significance if we do not consider the size and population of that country. Figure 1 illustrates the distribution of these two variables in different years.

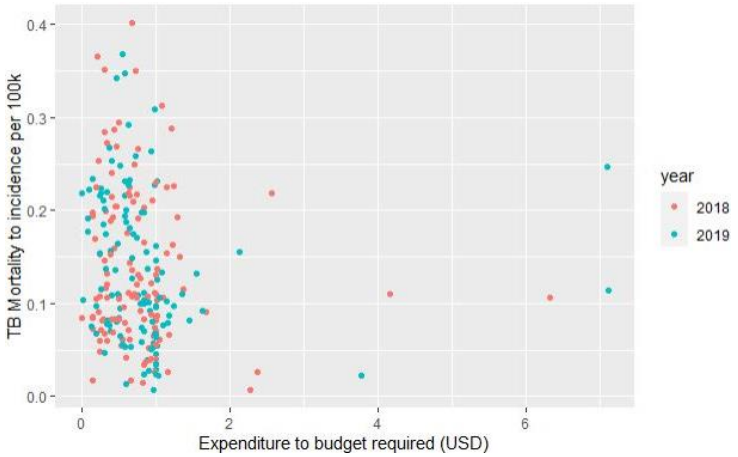


Figure 1

In order to select the correct correlation, test we need to consider and check a few assumptions. The two columns we are working on are both of "double" type and therefore of "interval" type. Now, let's test the normality of our columns.

The Shapiro-Wilk normality test for (*exp_budget*) and (*mort_inc_100k*) shows that at 95 percent significance level we do have enough evidence to reject the null hypothesis (which is that the column has a normal distribution), and therefore conclude that they are not normal (p~=0) (Please see Appendix 2 for more details). To further check and ensure this result, QQ plots are calculated and the results are shown at Figure 2 and 3. Based on these plots, it seems reasonable to conclude

that none of these columns are following normal distribution. As such, either we have to change our data or our test. Transformations such as log, sqrt, or 1/x can now be applied to these columns, however the results of either of the tests, Shapiro-Wilk and QQ plots do not demonstrate any improvement in favor of normality (details of the resulting transformations and their plots can be found in Appendix 3).

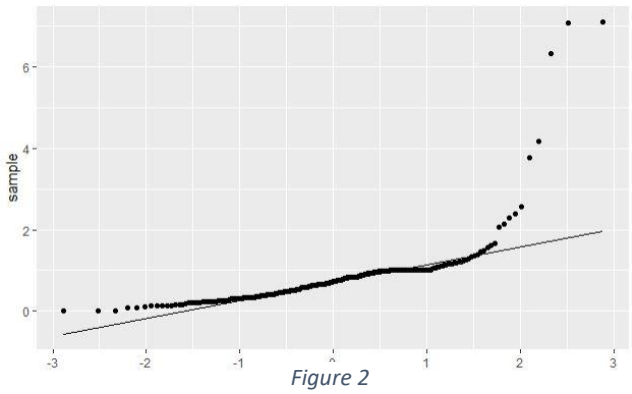


Figure 2

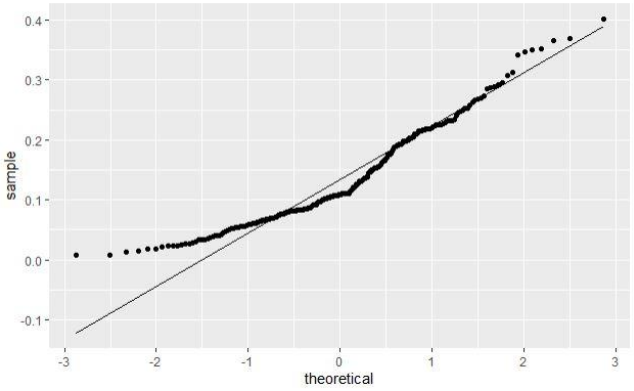


Figure 3

All in all, for testing the correlation of these two columns we must employ non-parametric tests. As such Spearman rank-order correlation test is selected for the purpose of this study.

Analysis:

Based on previous step, we should use non-parametric tests for calculating the correlation. Figures 4 and 5 below illustrate the resulting output for two different non-parametric tests. As can be noticeable, both tests unanimously agree about existence of a negative correlation between two columns. In other words, we have enough evidence to reject the null hypothesis of these tests about correlation between our variables.

```
        Spearman's rank correlation rho

data:  TB_df$exp_budget and TB_df$mort_inc_100k
S = 3130471, p-value = 0.0002366
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.2314392
```

```
        Kendall's rank correlation tau

data:  TB_df$exp_budget and TB_df$mort_inc_100k
z = -3.6255, p-value = 0.0002884
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
-0.1549698
```

Figure 4 shows this correlation within a plot.

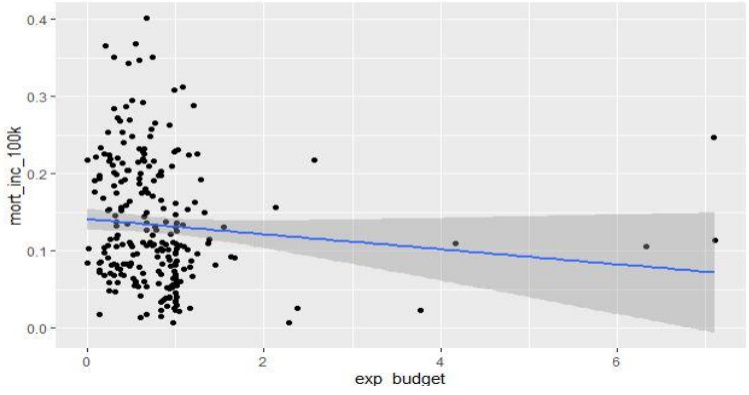


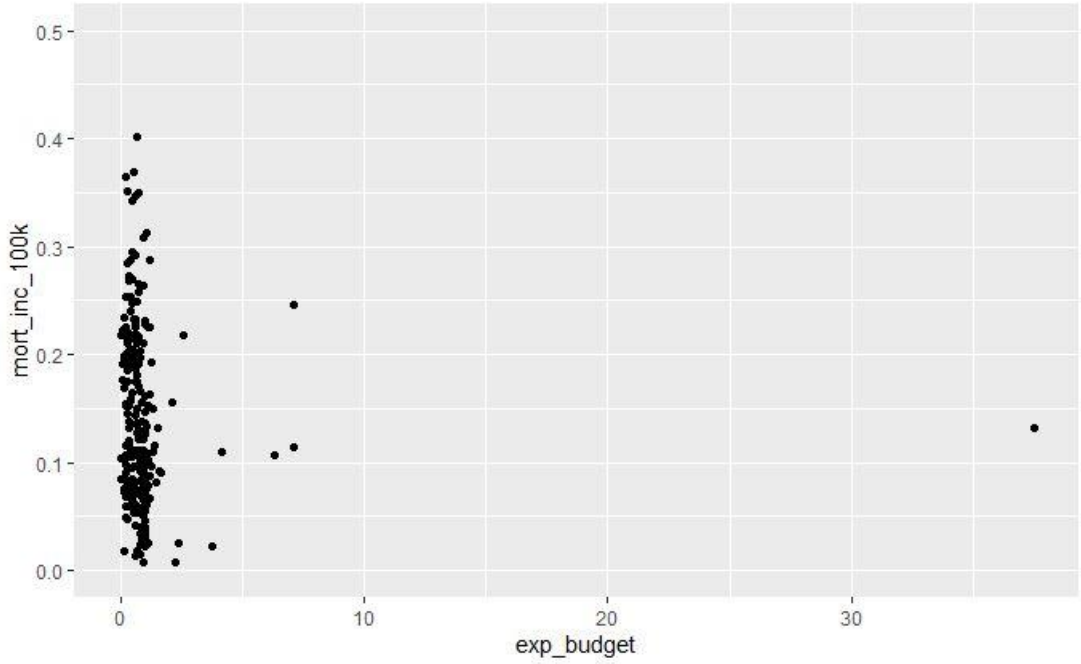
Figure 4

Conclusion:

Based on the results of the previous section, it is intuitive to conclude there is a negative correlation between ratio of expenditure to budget and ratio of deaths to instances of disease. The higher (*exp_budget*) is the lower (*mort_inc_100k*) would be. This correlation, however, is not very strong as is known from both the plot and the results.

Appendix 1:

Figure below shows that there is a country whose data might be mistaken because it is too out of scale. So I tried removing it from my dataset.



"Vanuatu" country has value of 37.47 expenditure to budget rate which seems out of scale.

Appendix 2:

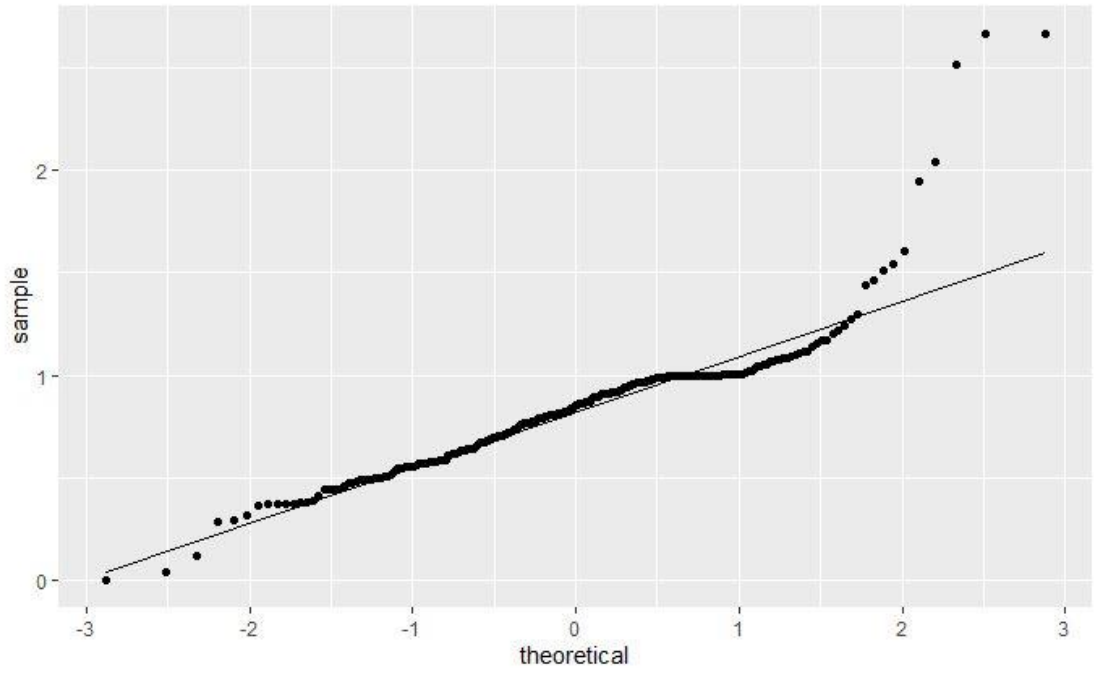
The result below is for untransformed data and Shapiro-Wilk normality test.

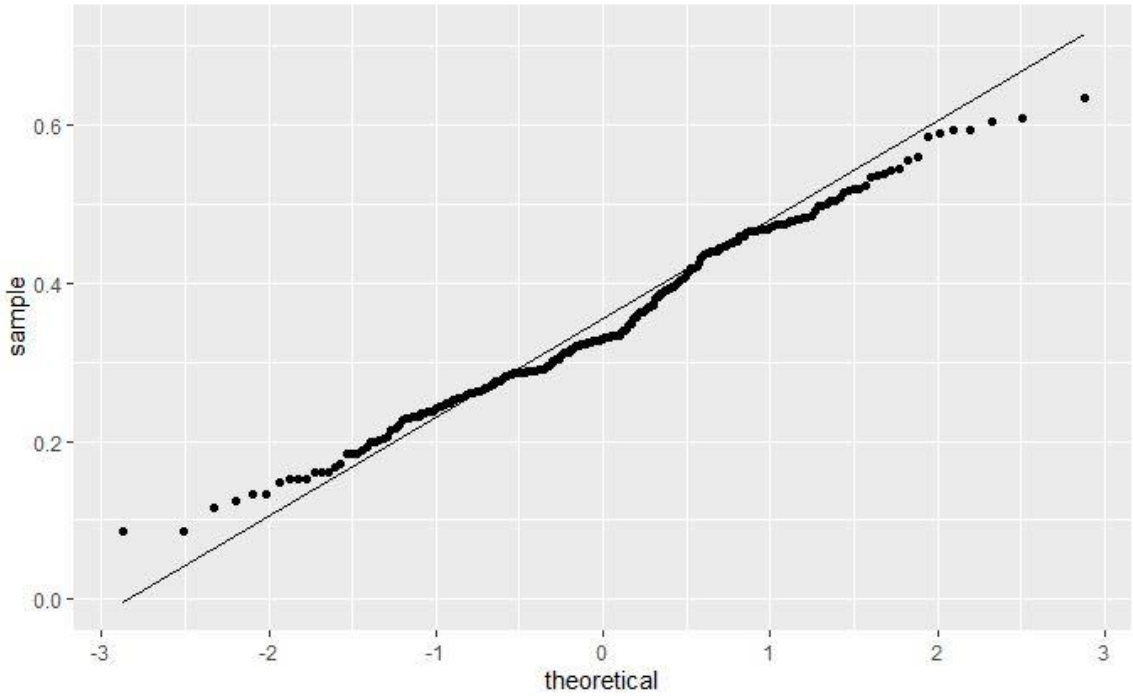
```
        Shapiro-Wilk normality test

data:  TB_df$exp_budget
W = 0.53508, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  TB_df$mort_inc_100k
W = 0.93585, p-value = 6.325e-09
```

If we transform the data with “sqrt” transformation, the results below are taken:

```
        Shapiro-Wilk normality test

data:  sqrt(TB_df$exp_budget)
W = 0.84317, p-value = 2.892e-15


        Shapiro-Wilk normality test

data:  sqrt(TB_df$mort_inc_100k)
W = 0.98661, p-value = 0.02035
```

The QQ plots are also as follows:





Even though transforming the data, still there is not enough evidence to consider these two columns normal. So we tried changing our test.