Assignment 2

Shahzeb Naveed (20789222) | Zaryab Javaid (20852202)) | Muhammad Mohsin Tahir (20812155))

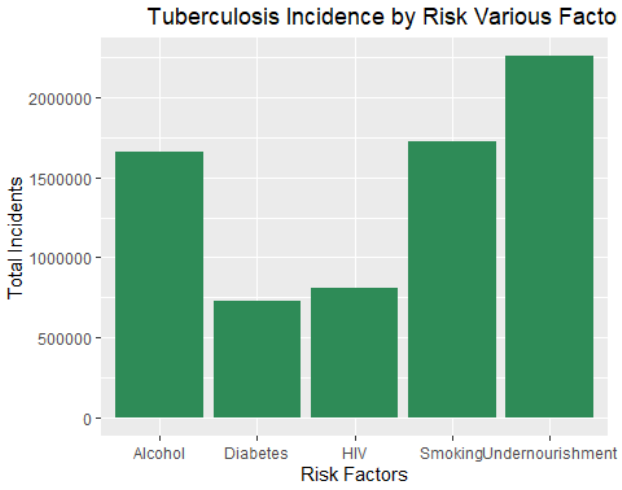
Introduction

Tuberculosis (TB) remains an important global health issue. Awareness about the disease, its diagnosis, and treatment among public will help in controlling this killer disease. This study is related to finding a contributing factor which is highly co-related with TB.

Data:

After exploration of datasets provided by WHO, we wanted to know more about the contribution of various factors that can potentially cause this fatal disease. The following variables of interest were plotted from the 'Dissegregated Estimates' dataset that inititally had **7310 obervations** and **13 variables**.

Variable	Description	Type	Min	Median	Max
best	TB Incidents (Best Estimate)	Integer	0	7832	2690000
risk_factor	Risk factor for TB	char	-	-	-



As clearly seen in the graph to the right, a person can develop TB due to many possible risk factors with undernourishment being the most common and diabetes being the least common causes. Based on data availablity, we decided to expore the correlation between HIV infections and TB cases. For this purpose, we found that the dataset 'TB_notification' contained total number of independent HIV cases reported from 1980 till 2018 in 218 different countries. For total number of TB cases reported in different years, we used the "TB_burden_countries" dataset and left-joined both the datasets. We then, **left-joined** the two datasets using (country, year) as the primary key.

Summary of variables of interest:

The final dataset had **5 variables** and **862 observations**.

Variable	Description	Type	Min	Median	Max
e_inc_num	TB Incidents	Integer	1	5400	3200000
e_pop_num	Country Population	Integer	1.012e+04	6.950e+06	1.399e+09
hiv_reg	HIV Incidents	Integer	1	1984	4277683
e_inc_tbhiv_num	HIV-positive TB Cases	Integer	0	210	332000
e_mort_num	Deaths by TB	Integer	0	300	735000
year	Year	char	-	-	-
country	Country	char	-	-	-

Data Cleaning :

Out of the three selected datasets, only "TB_notification" contained around **3142 missing values** in the column of hiv_reg which were dropped from analysis. As far as **outlier detection** is concerned, the very high values(TB cases) might appear to be "abnormally high" but we are including those in our analysis as they might correspond to countries like China/India having high incidents in proportion to their large sizes. Apart from this the data we used was consistent, correct and there were no structural errors.

Planning

Before building any hypothesis, it's important to check whether the hypothesis makes any practical sense. After doing extensive research, we came to know that TB is an opportunistic infection (OI)[1]. OIs are infections that occur more often or are more severe in people with weakened immune systems than in people with healthy immune systems. HIV weakens the immune system, increasing the risk of TB in people with HIV. So, it definitely makes sense to investigate this correlation in detail.

Hypothesis :

Null Hypothesis: The correlation coefficent between number of TB and number HIV cases reported is 0.
Alternative Hypothesis: The correlation coefficent between number of TB and number HIV cases reported is not 0.

Data Manipulation :

The three variables which we selected are HIV cases, TB cases and Total Population for each country (as **Control Variable**) were present in two different datasets. Before merging, we filtered data from **2000 to 2018** from the Notificants dataset as we only had data for TB for that duration in the TB_Burden_Countries dataset.

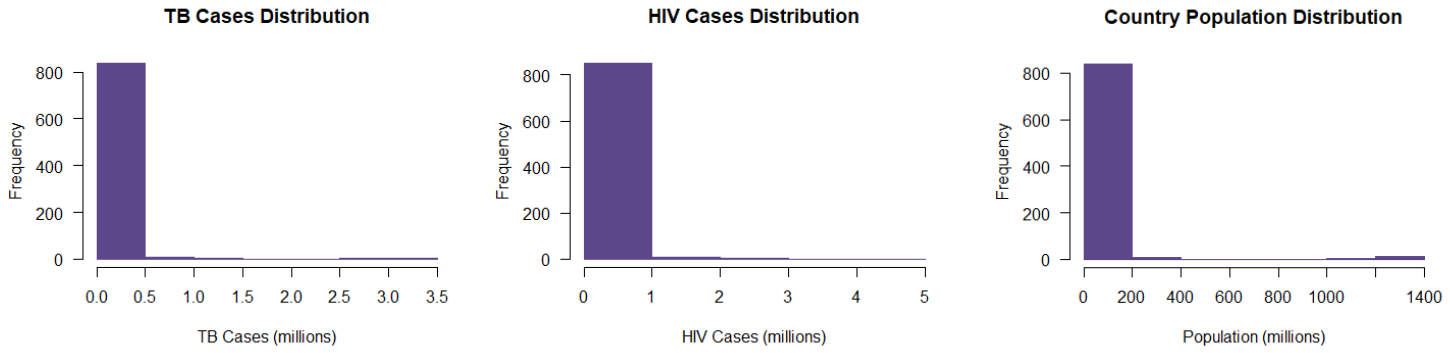
For simplifying our analysis, we categorized countries as Large or Small with countries with populations greater than 10 times the mean population size being labelled as Large.

Assumptions

Testing Assumptions for Pearson Correlation:

1. Normality
2. Homeoscedacity
3. Linearity

From the histograms below, it is evident that all three variables are **highly right-skewed**. However, we confirm this by Shapiro-Wilk Normality test giving beneath the hisograms:



Shapiro-Wilk Normality Test

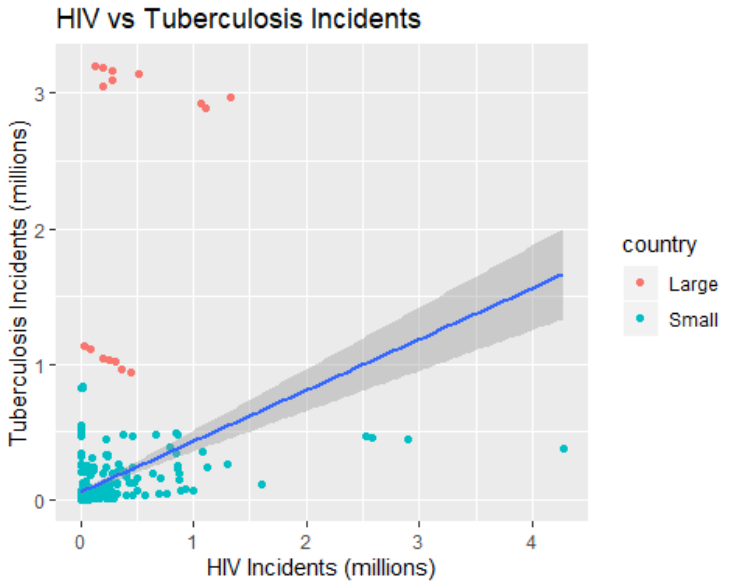
Variable	Test Statistic W	p-Value	Result
TB Incidents	0.22547	2.2e-16	Non-normal
HIV Incidents	0.23568	2.2e-16	Non-normal
Population	0.26553	2.2e-16	Non-normal

As far as homoscedasticity is concerned, we can observe from the scatter plots plotted below that **fans out thus violaiting the assumption for homoscedacity**. It is evident from scatter plots that data is **not linear** as well.

Therefore, we cannot use Pearson corrleation method, we will have to use non-parameteric method **Kendall Rank Correlation**.

Analysis

Correlation Coefficeint Calcualtions and Hypothesis Testing



Now, let's plot the two variables: the HIV incidents and Total Tuberculoses incidents. As seen visually, there exists a clear positive correlation i.e as number of infections increase, the TB incidents also increase. (For more insights, see "Conclusion and Key Insights" section)

Kendall's Correlation Coefficient:

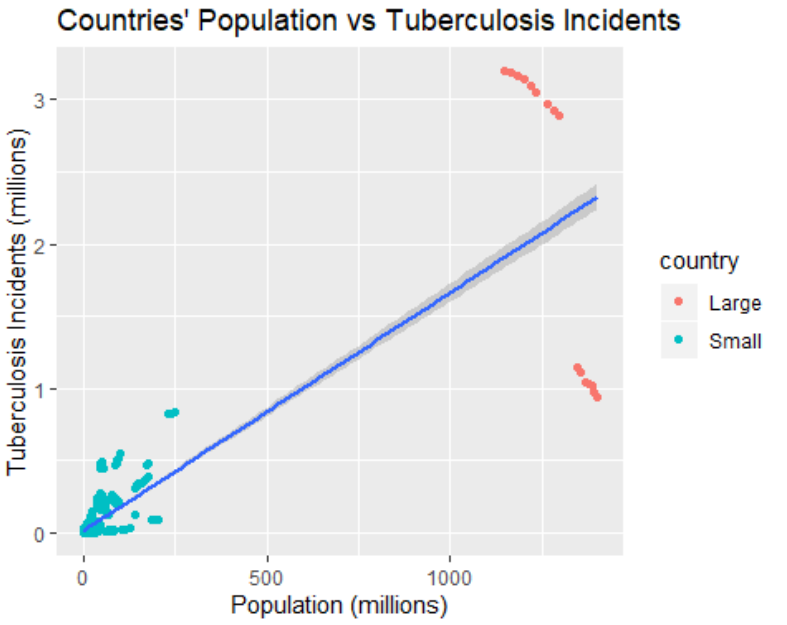
```
##            e_inc_num   hiv_reg
## e_inc_num 1.0000000 0.5201556
## hiv_reg   0.5201556 1.0000000
```

The correlation coefficient comes out to be **0.5201**. To test the statistical significance (and our null hypothesis), we apply cor.test() function:

```
## z = 22.803, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.5201556
```

With a p-value of 2.2e-16, we have a strong evidence to reject the null hypothesis at 0.05 significance level and conclude that the correlation between HIV cases and TB cases exists.

Partial-Correlation Coefficient Calculations:



We will find partial correlation in order to measure the true strength of the relationship by **controlling for the effect of Population sizes**. The plot below shows that correlation exists between population and number of TB cases reported and thus, the Population must be "controlled" using partial-correlation. We can also confirm it by performing Kendell Rank Correation Coefficient between population and TB cases.

Kendall's Correlation Coefficient:

```
##            e_inc_num e_pop_num
## e_inc_num 1.0000000 0.6765424
## e_pop_num 0.6765424 1.0000000
```

The correlation coefficient of **0.677** shows that there is a high positive correlation between population and number of TB cases reported. So it means we will have to nullify the effect of population on TB cases when testing our hypothesis and will have to treat it as a **control variable**.

Kendall's Partial Correlation Coefficient:

We got a **0.301** value for Kendall's partial correlation coefficient.It shows a moderate positive correlation between number of TB cases and HIV cases while controlling the population. We further tested with pcor.test() and got the following results.

Partial Correlation Test

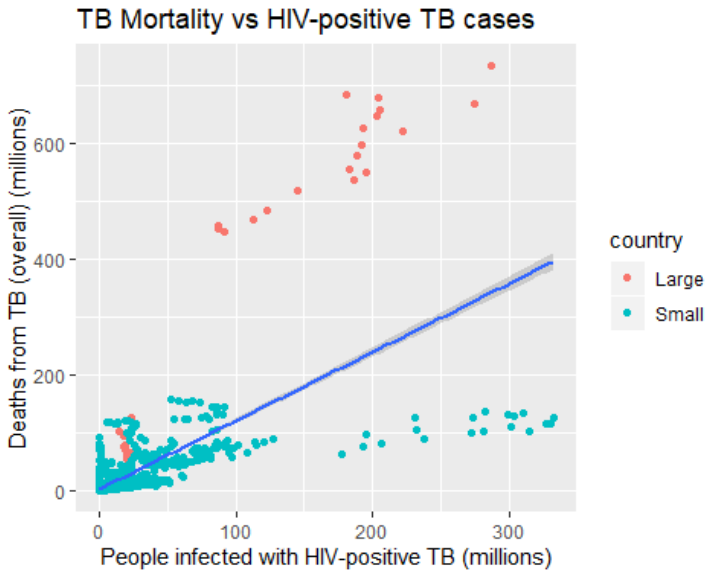
```
##    estimate      p.value statistic   n gp  Method
## 1 0.3011637 6.000311e-40  13.22859 862  1 kendall
```

With a p-value of 6.000311e-40, we have a strong evidence to reject the null hypothesis at 0.05 significance level and conclude that the a direct corelation between HIV cases and TB cases exists and it can be considered as moderate positive association.

Follow-up:

Digging further into it, we analyzed contribution to "deaths by TB" by HIV as a factor. To determine the true correlation between TB Mortality and HIV-incuded TB mortality, we "excluded" the effect of other possible factors of TB (such as alcohol, diabetes etc) through **partial correlation**. In partial correlation test, we get correlation coeffcient of **0.402** as shown below. (For more insights, see "Conclusion and Key Insights" section)

```
##    estimate       p.value statistic    n gp  Method
## 1 0.4019858 1.392255e-272  35.27556 3427  1 kendall
```



Conclusion and Key Insights:

1. Undernourishment is the most common and diabetes is the least common cause of Tuberculosis.
2. We conclude that a moderate positive correlation exists between the number of TB and HIV cases reported, however we cannot conclude anything about the causality here.
3. In large countries, the increase in TB Mortality with an increase in HIV-positive is more rapid.
4. In large countries, even with large number of TB incidents, the contribution of HIV to TB is lower than that in small countries. This implies that in these countries, other factors such as alcohol consumption or diabetes might be the more contributing causes of TB.
5. From our analysis, we also conclude that if any future studies about Tuberculosis and HIV are to be carried out, then the effect of population should be taken into account.

Recommendation:

TB/HIV co-infection is one of the serious health problems as mortality rates are quite high. Thus, collaborative TB/HIV activities that reduce the co-morbidities and mortalities should be addressed. There is also an urgent need for increased public funding toward TB health care services that have long-term effectiveness in high HIV-prevalence settings.

References:

[1] https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/26/90/hiv-and-tuberculosis--tb-
[2] https://www.cdc.gov/tb/topic/basics/tbhivcoinfection.htm
[3] https://www.unaids.org/en/topic/tuberculosis