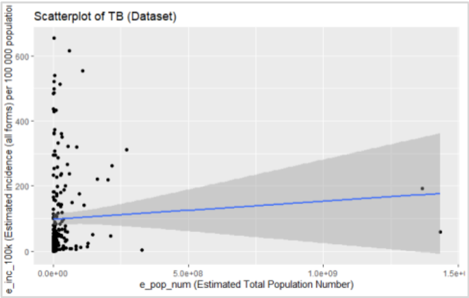
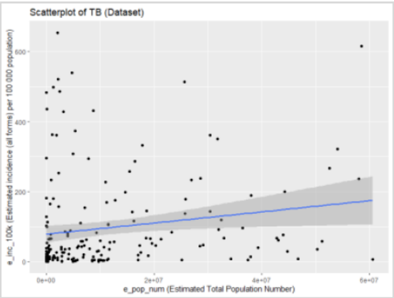
Data

From pair assignment – 2, we analysed the data set from the World Health Organization (WHO) which contains WHO-generated estimates of Tuberculosis (TB) mortality, incidence (including disaggregation by age, sex, risk factors).

The default data set comes with 4,272 observations and 50 variables. Each variable has a column, and each observation on those variables has one row. There are 215 observations and 4 variables. With the year filtered as “2019”, we have taken ‘Estimated Total Population Number’ and ‘(Estimated incidence (all forms) per 100 000 population’ as our correlation variables. For an integer value, the dataset is reloaded to explicitly parse those values as an integer. Moreover, we checked for “NA” values in the data set and it returned zero. As seen from the plots, there are 2 outliers and removed. Outlier plots are generated as follows:

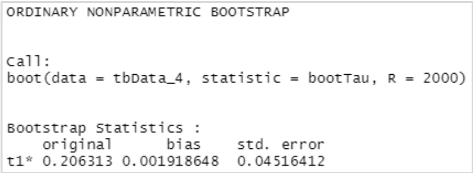




Planning

Bootstrapping will be implemented based on the 2 variables, which were considered partially correlated in previous assignment, **‘e_pop_num’** and **‘e_inc_100k’**. Bootstrapping will test using random sampling with replacement and assigns measures of accuracy like bias, confidence interval to sample estimates. So, here assuming that the sample population is similar to the general population and randomly resample from the sample with replacement to produce estimates of test statistic by taking the mean as estimate and quantiles for a 95% confidence interval. At first, Kendall’s Tau test will be performed, which is a non-parametric measure of relationships between columns of ranked data. The reason to perform this test is to check whether our data of number of TB incidence is good representation of general population of total number of populations. It has been observed that the variables are not normally distributed or the relationship between the variables is not linear, it will be more appropriate to use the Spearman rank correlation method. However, we will analyse the result of the Spearman’s rho and Pearson’s r test as well since it measures the linear relationship between two continuous random variables.

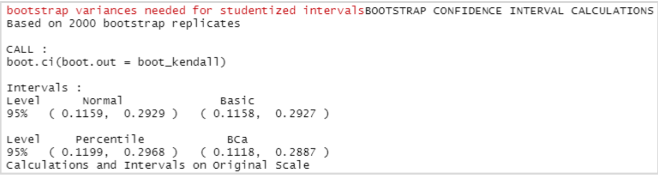
Analysis



```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = tbData_4, statistic = bootTau, R = 2000)


Bootstrap Statistics :
    original      bias    std. error
t1* 0.206313 0.001918648  0.04516412
```



```
bootstrap variances needed for studentized intervalsBOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_kendall)

Intervals :
Level      Normal              Basic
95%   ( 0.1159,  0.2929 )   ( 0.1158,  0.2927 )

Level     Percentile            BCa
95%   ( 0.1199,  0.2968 )   ( 0.1118,  0.2887 )
Calculations and Intervals on Original Scale
```

On running the Kendall Tau on our data set, we get original stat and confidence interval aforementioned. To summarize, we have found the boot_kendall$t0 = 0.206313 as the correlation coefficient for 2000 estimates of correlation.

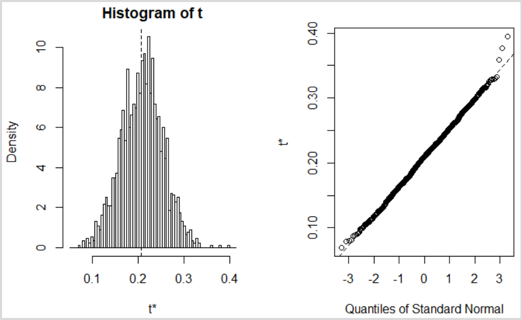
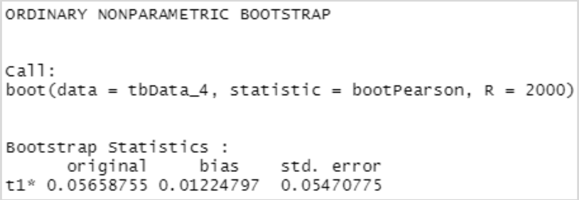


Figure 1: Histogram and QQ Plot for Kendall Tau Bootstrap

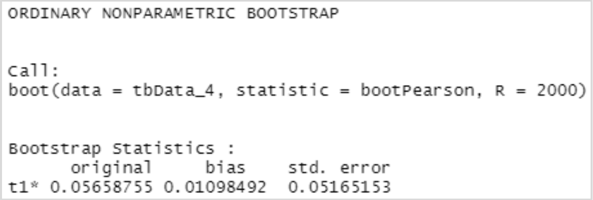
In the previous Individual Assignment 1, Pearson's R Correlation and Spearman's rho test were performed and to verify the same, here have been performed bootstrap using Pearson's and Spearman's Coefficient. The original stats and confidence interval are as follows:



```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = tbData_4, statistic = bootPearson, R = 2000)


Bootstrap Statistics :
      original      bias    std. error
t1* 0.05658755 0.01224797  0.05470775
```



```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = tbData_4, statistic = bootPearson, R = 2000)


Bootstrap Statistics :
      original      bias    std. error
t1* 0.05658755 0.01098492  0.05165153
```

We found the boot_spearman$t0 = 0.05658755 and boot_spearman$t0 = 0.05658755 as the correlation coefficient for 2000 estimates of correlation.

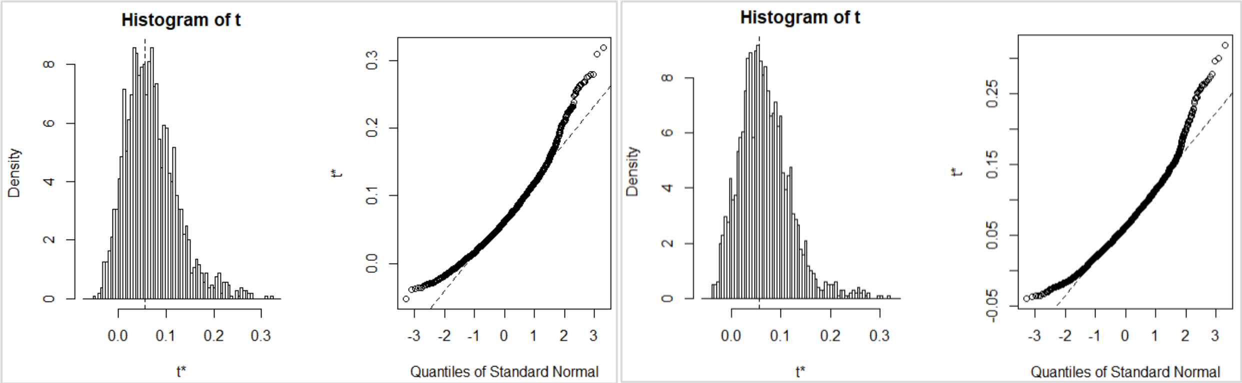


Figure 2: Histogram and QQ Plot for Spearman's rho and Pearson's R Bootstrap

To summarize, the results show that the Spearman's rho and Pearson's corelation coefficient are signifying a medium effect of the 2 chosen variables on each other. There is a medium strength relation between the total number of incidences of TB within the total number of populations who were working in the country in the public and private sector and the number of Estimated incidence (all forms) per 100 000 population.

Bootstrap is also an appropriate way to control and check the stability of the results. Although for most issues, it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality. Therefore, bootstrapping has been used to evaluate the mentioned correlations and found the aforementioned results.

Bibliographic Reference:

Kabakoff, R. I. (2017) *Bootstrapping* Retrieved from https://www.statmethods.net/advstats/bootstrapping.html

(Anonymous, 2021) *Bootstrapping in R – Single guide for all concepts* Retrieved from https://data-flair.training/blogs/bootstrapping-in-r/

(Anonymous, 2021) *HOW CAN I GENERATE BOOTSTRAP STATISTICS IN R? | R FAQ* Retrieved from https://stats.idre.ucla.edu/r/faq/how-can-i-generate-bootstrap-statistics-in-r/