TB Population correlation Report

Data

In burden data set we have 4272 records about 217 countries, it consist of 50 variables as[appx-1]. The data is divided across different categories like hiv, tb which are contributing factors in both mortality and cases. Along with this case detection ratios are provided. There are some outliers associated with each column mainly because of denser distribution of population in few regions and the spread of diseases in few regions. Looking at distribution of population and no of cases:

```
##                      mean           sd  min        max      range
## e_pop_num    32276580.77 129358675.5 1126 1433783692 1433782566
## e_inc_num_lo    33630.82    145493.2    0    1870000    1870000
```

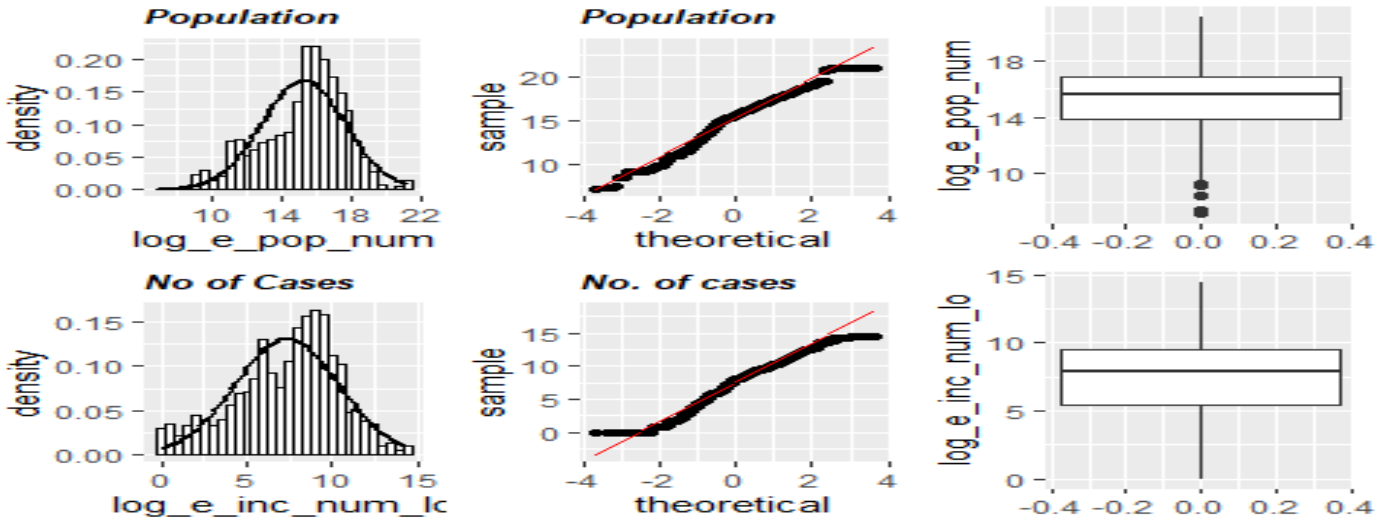
In these results, the mean population is a lot larger than the median is. The data appear to be skewed to the right[appx-2], which explains why the mean is greater than the median.(similar is applicable for no of cases). Here we are selecting population and no of cases for correlation testing because both the variables are monotonically increasing with respect to year,(i.e. a linear relationship) along with that since these variables are significantly skewed, both of them are transformed using log transformation in order to achieve normality required for parametric correlation test. Which can be crossverified using different normality tests.

Planning

Checking assumption: Normal distribution

A) Visual inspection: Both the boxplots are not showing ideal normal distribution; the median is slightly close to the upper quartile of the box and the whiskers are slightly unequal. Population has outliers (due to few countries having dense population). Along with box plot QQ plot is showing some linearity in middle of the graph however actual and expected values are not same for higher and lower values for both variables. So, only visual inspection is not sufficient, we shall evaluate skewness coefficients.

Fig 1- Visual inspection for normality



B) Quantifying normality with numbers:

1. Measuring skewness

```
##          log_e_pop_num log_e_inc_num_lo
## skew.2SE     -7.049447        -5.415768
```

Because skewness of Population(after transformation) -7.0494468 and skewness of no of Cases (after transformation) -5.4157681 is not between - 3.29 and 3.29 [ref-1], we can conclude that the skewness for both is different from 0 (at p < .001.).

2. Shapiro–Wilk test: Since we have observations<5000, we can perform Shapiro–Wilk test to test normality.

```
##      method                          data.name    statistic p.value
## [1,] "Shapiro-Wilk normality test" "population" 0.9686663 3.74686e-29
## [2,] "Shapiro-Wilk normality test" "cases"      0.9776878 5.320147e-25
```

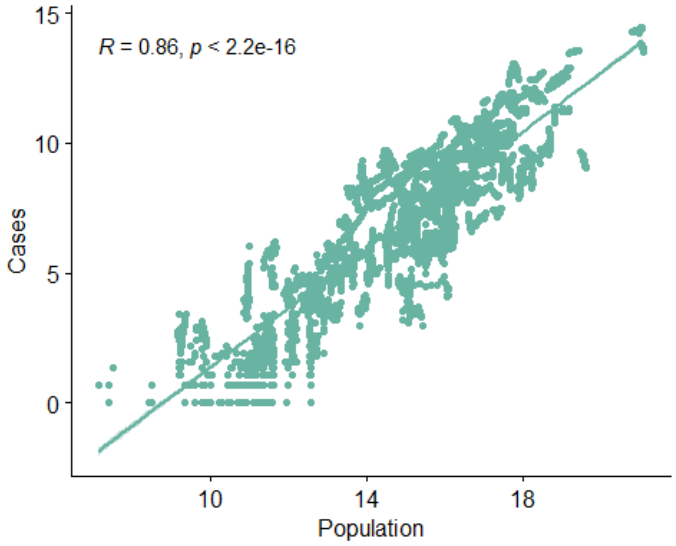
Shapiro–Wilk test for both features is p (very much) less than 0.05 (and, therefore, significant), and the numeracy scores (w), for Population is W=0.96 and Cases W=0.97, data is significantly non-normal.

Analysis:

correlation test: Since both the variable under consideration are not normally distributed, we cannot perform parametric test. However, we can calculate **Spearman's rho** since we have interval data (monotonically related), there is no assumption of normality (as it is a nonparametric statistic). For correlation test let's assume the null hypothesis as there is zero correlation between population and cases and hence alternative hypothesis is there is non-zero correlation in these 2 variables.

```
##  Spearman's rank correlation rho
##
## data:  population and cases
## S = 1701522213, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8563325
```

It reiterates that the correlation between Population and number of Cases was 0.86, but tells us that this is significantly different from zero, p < 2.2e-16 is very close to zero. A typical threshold is 0.05, anything smaller counts as statistically significant. Most important, In all likelihood, Population and number of Cases are, in reality, positively related. Hence we reject null hypothesis and accept alternative hypothesis. This positive correlation between Population and no. of Cases can be visualized from fig 3



Conclusion:

"A Spearman's correlation coefficient was computed to assess the relationship between the total population of the country and number of cases across country. There was a positive correlation between the two variables, rho = .86 (p-value < 2.2e-16). A scatterplot summarizes the results (Figure 3). Overall, there was a strong, positive correlation between total population and number of cases. Increases in total population were correlated with increases in number of cases."