

Assignment2- Correlation -Appendix

Aishwarya (Student ID: 20909044)

Data source:

<https://extranet.who.int/tme/generateCSV.asp?ds=estimates>

Resources:

<https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf>

Discovering statistics Using r-Andy Field (Chapter 5 & 6)

Appendix-1

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1 4272
## [1] 50
```

Appendix-2

```
summary(df$e_pop_num)

##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.
## 1.126e+03 7.363e+05 5.760e+06 3.228e+07 2.050e+07 1.434e+09

#sd(df)
```

Appendix-3

```
summary(df$e_inc_num_lo)

##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.
##        0      160      2300    33631    12000  1870000
```

Appendix-4

In budget data set each column represents unique feature, each row represents unique observation and each cell has unique value so the data provided is already tidy.

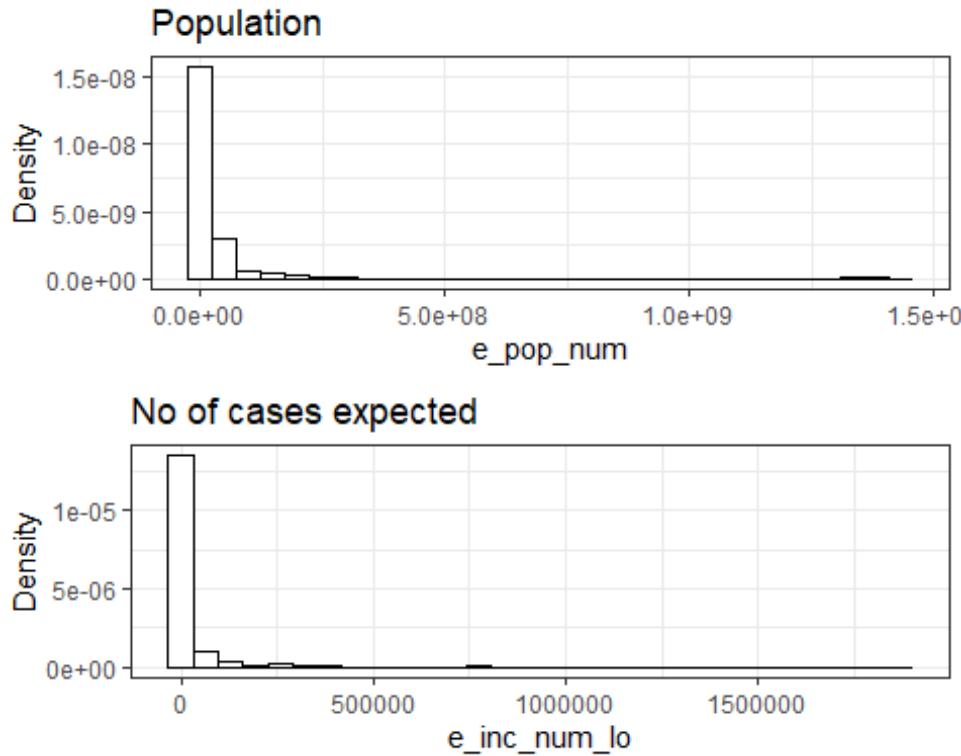
Appendix-5

```
missing_data<- df %>%
  select(e_pop_num,e_inc_num_lo) %>%
  summarise_all(~ sum(is.na(.)))
```

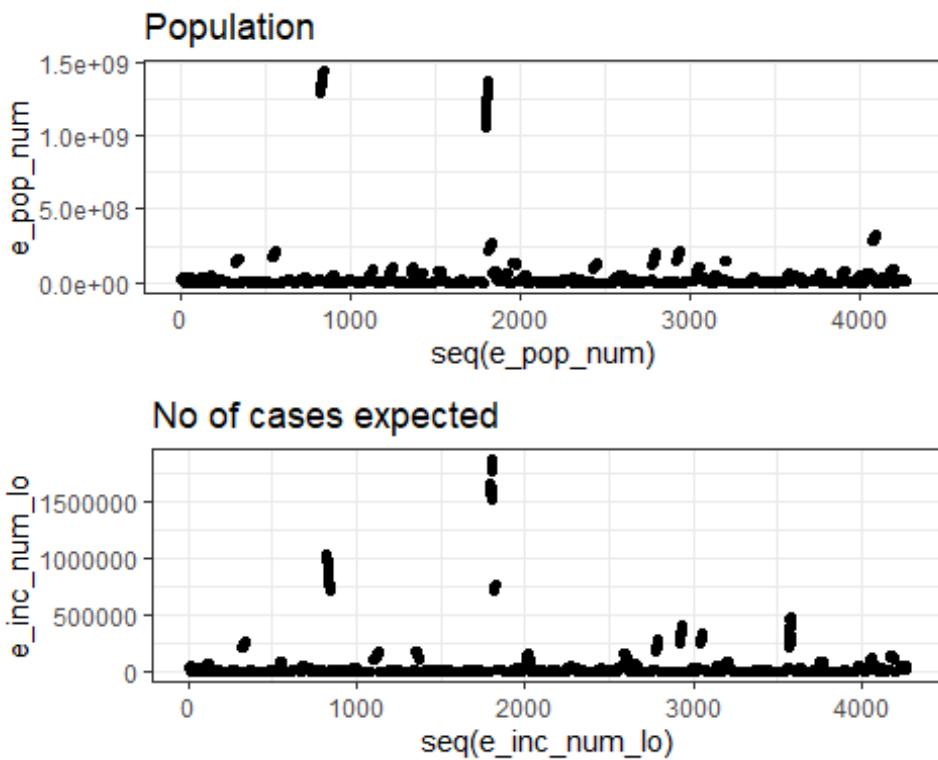
This data has several missing values and we may consider the budget to have normal distribution but at the same time this budget depends on several other factors such as

economy policies, geography and other medical conditions in vicinity, so due to these uncertainty we are not assuming anything to replace these values with any representative. ggplot will simply ignore missing values.

Appendix-6



Appendix-7



Appendix-8

```
e_pop_num_desc<-stat.desc(df$e_pop_num, basic=FALSE, norm=TRUE)
e_inc_num_lo_desc<-stat.desc(df$e_inc_num_lo, basic=FALSE, norm=TRUE)
e_pop_num_desc

##          median           mean        SE.mean CI.mean.0.95         var
std.dev
## 5.760274e+06 3.227658e+07 1.979156e+06 3.880173e+06 1.673367e+16
1.293587e+08
##      coef.var      skewness     skew.2SE      kurtosis     kurt.2SE
normtest.W
## 4.007818e+00 8.912870e+00 1.189543e+02 8.469490e+01 5.653158e+02
2.098049e-01
##      normtest.p
## 2.562658e-86

e_inc_num_lo_desc

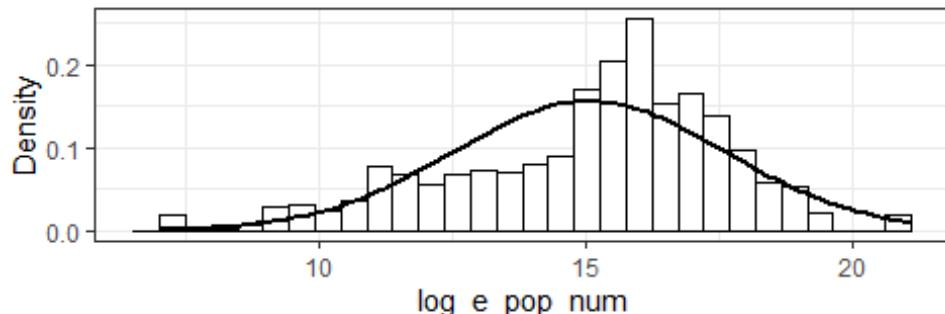
##          median           mean        SE.mean CI.mean.0.95         var
std.dev
## 2.300000e+03 3.363082e+04 2.226010e+03 4.364136e+03 2.116827e+10
1.454932e+05
##      coef.var      skewness     skew.2SE      kurtosis     kurt.2SE
normtest.W
## 4.326186e+00 8.450104e+00 1.127780e+02 8.293215e+01 5.535499e+02
```

```
2.191982e-01
## normtest.p
## 4.739115e-86
```

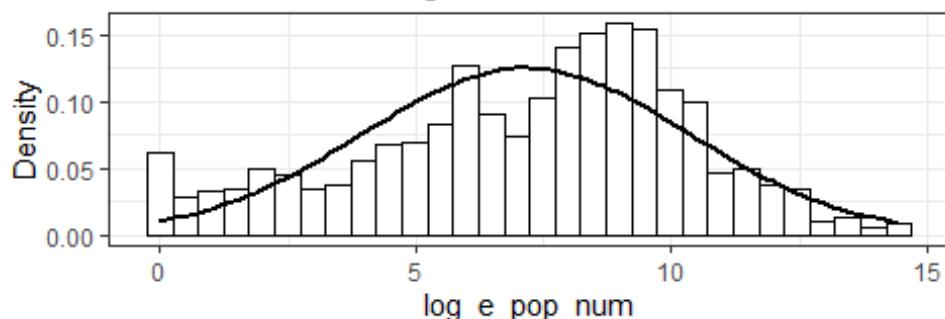
Appendix-9

Considering log transformation on budget and log transformation on (no. of cases+1) for handling skewness.

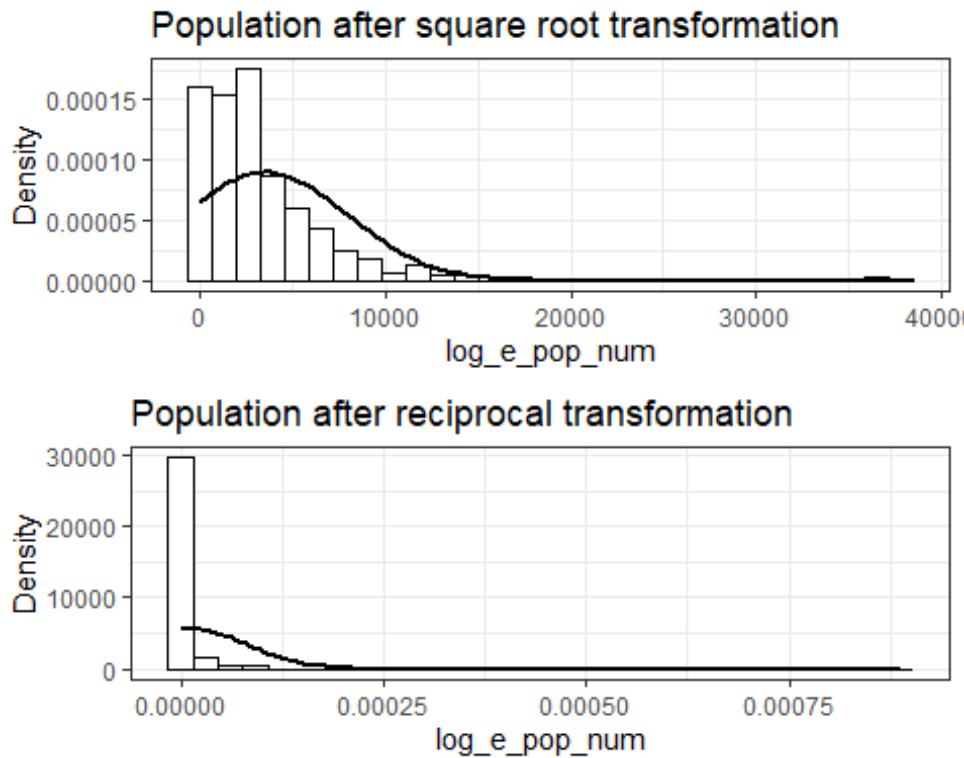
Population after log transformation



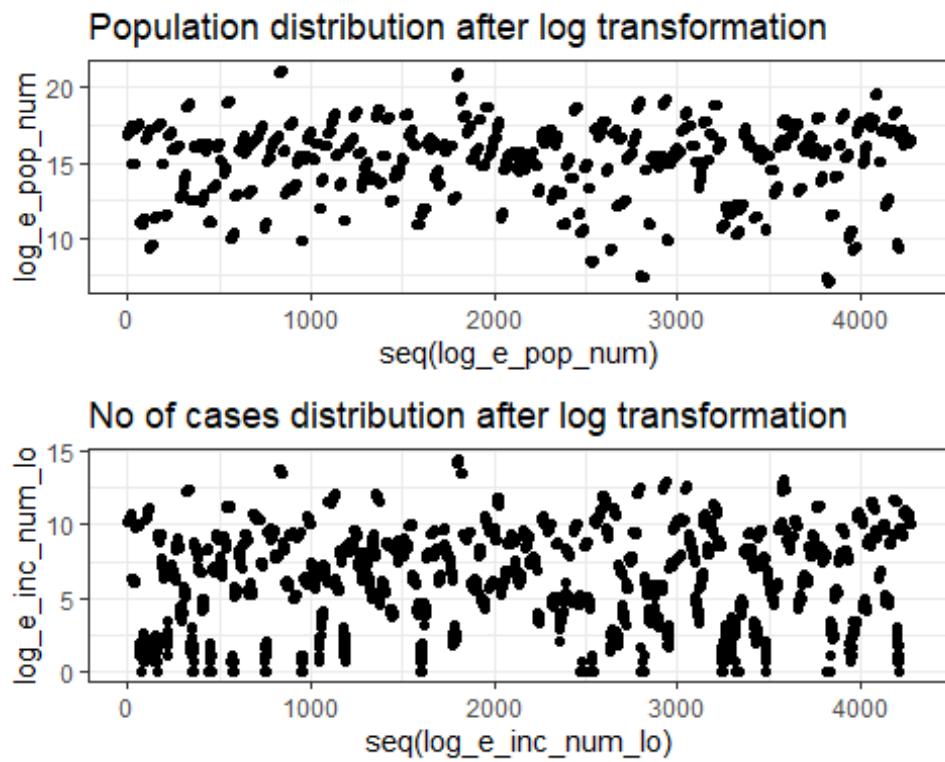
No of cases after log transformation



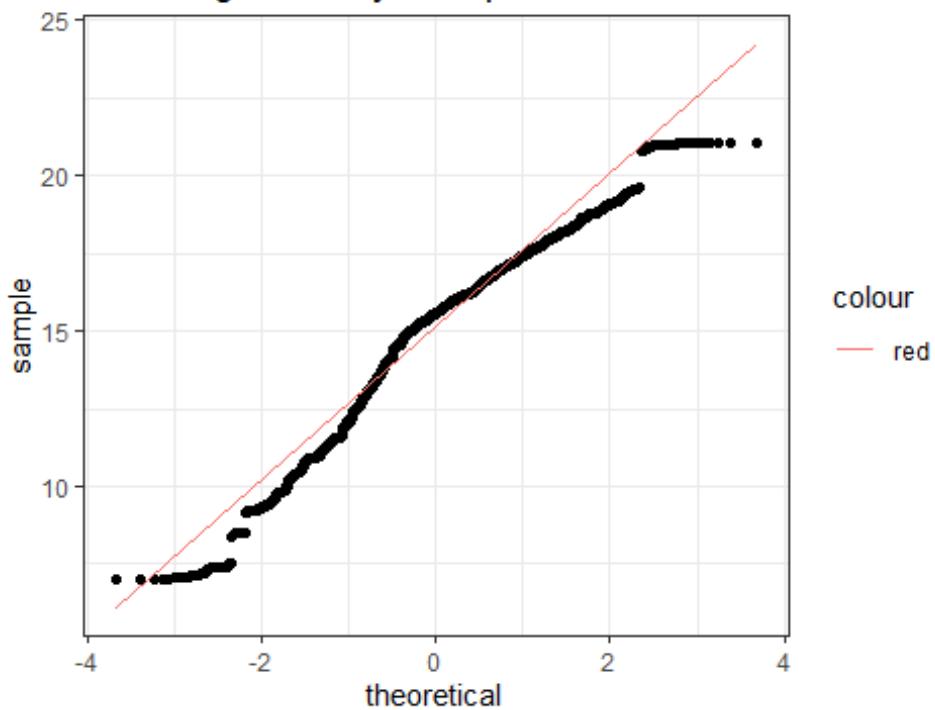
Appendix-10



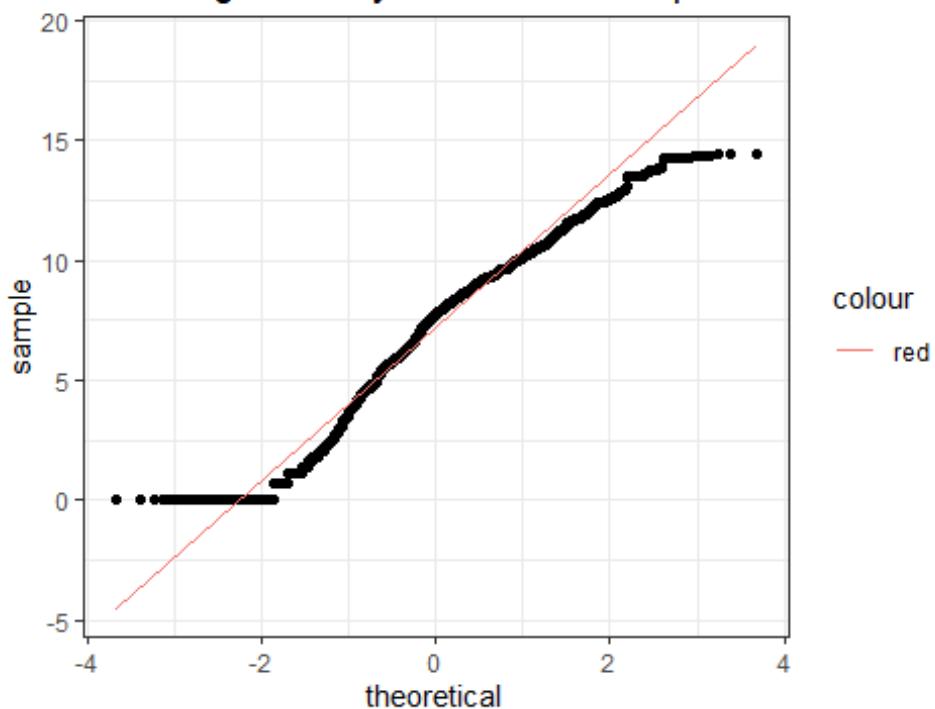
Appendix-11



Checking normality of Population



Checking normality of no of cases expected



```
stat.desc(df$log_e_pop_num, basic=FALSE, norm=TRUE)
```

```

##      median         mean        SE.mean    CI.mean.0.95          var
## 1.556650e+01 1.506601e+01 3.914722e-02 7.674890e-02 6.546861e+00
##      std.dev       coef.var      skewness     skew.2SE      kurtosis
## 2.558684e+00 1.698315e-01 -6.468572e-01 -8.633181e+00 1.130809e-01
##      kurt.2SE      normtest.W      normtest.p
## 7.547848e-01 9.617908e-01 3.648691e-32

stat.desc(df$log_e_inc_num_lo, basic=FALSE, norm=TRUE)

##      median         mean        SE.mean    CI.mean.0.95          var
## 7.741099e+00 7.146174e+00 4.895044e-02 9.596829e-02 1.023633e+01
##      std.dev       coef.var      skewness     skew.2SE      kurtosis
## 3.199427e+00 4.477119e-01 -4.277316e-01 -5.708654e+00 -4.201397e-01
##      kurt.2SE      normtest.W      normtest.p
## -2.804320e+00 9.716721e-01 2.784738e-28

```

Appendix-12

```

skew_e_pop_num<- stat.desc(df$log_e_pop_num, basic=FALSE,
norm=TRUE)[ "skew.2SE" ]
skew_e_inc_num_lo<- stat.desc(df$log_e_inc_num_lo, basic=FALSE,
norm=TRUE)[ "skew.2SE" ]

```

Appendix-13

```

##
## Shapiro-Wilk normality test
##
## data: df$log_e_pop_num
## W = 0.96179, p-value < 2.2e-16

##
## Shapiro-Wilk normality test
##
## data: df$log_e_inc_num_lo
## W = 0.97167, p-value < 2.2e-16

```

Appendix-14

```

library(car)
leveneTest(df$log_e_pop_num, df$year, center=mean)

## Levene's Test for Homogeneity of Variance (center = mean)
##          Df F value Pr(>F)
## group    19  0.0188      1
##          4252

leveneTest(df$log_e_inc_num_lo, df$year, center=mean)

## Levene's Test for Homogeneity of Variance (center = mean)
##          Df F value Pr(>F)
## group    19  0.0406      1
##          4252

```

Appendix-15

```
df %>%
  select(log_e_pop_num, log_e_inc_num_lo) %>%
  cor(use="complete.obs", method = "spearman")

##           log_e_pop_num log_e_inc_num_lo
## log_e_pop_num      1.0000000   0.8682392
## log_e_inc_num_lo     0.8682392   1.0000000
```

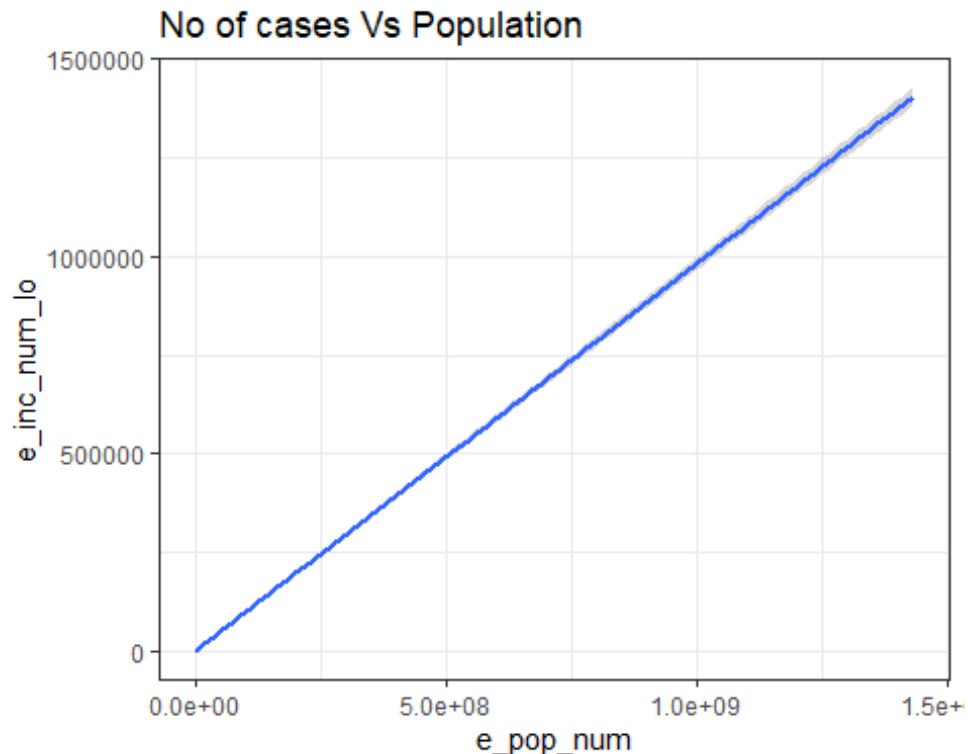
Appendix-16

```
cor.test(df$log_e_pop_num, df$log_e_inc_num_lo, method ="spearman")

##
##  Spearman's rank correlation rho
##
##  data:  df$log_e_pop_num and df$log_e_inc_num_lo
##  S = 1712098011, p-value < 2.2e-16
##  alternative hypothesis: true rho is not equal to 0
##  sample estimates:
##        rho
##  0.8682392
```

Appendix-17

```
e <- ggplot(budgetData, aes(e_pop_num, e_inc_num_lo))
e+geom_smooth(method = lm)+ggtitle("No of cases Vs Population")+theme_bw()
```



Appendix-18

<http://www.sthda.com/english/wiki/scatter-plots-r-base-graphs>