Project 2

Chad Chapnick

April 15, 2016

Part 1

Data Cleaning

The first step in the project was to prepare the data for analysis. Although the data were curated by the Population Reference Bureau, there are sources of potential error including data-entry, preprocessing, and conversion from a pdf to csv file. Due to the rather tedious and esoteric nature of this step, it has been absracted away by executing the `data_cleaner.R` script. The code is available with annotations regading major processing steps. The main components were:

- Providing meaningful column names.
- Converting data to the appropriate types
  - storing categorical variables as factors
  - storing numeric variables as numeric
  - diving columns with "<0.1" values into levels with an appropriate range of values
- Modifying the Ñ symbol from the data to the `NA`
- Labeling each country with the approriate continent and region, in columns stored as factors
- Removing the summary data for each continent and region from the original data frame, and saving them in separate data frames.

```
source('data_cleaner.R') # run the file cleaning script.
```

Featured below are the first six rows and first five coulmns of the data frame:

```
head(data[, c(1:5)])
```

```
##   COUNTRY          REGION CONTINENT LifeExp_Both LifeExp_Male
## 1 algeria NORTHERN AFRICA    AFRICA           74           72
## 2   egypt NORTHERN AFRICA    AFRICA           71           70
## 3   libya NORTHERN AFRICA    AFRICA           71           69
## 4 morocco NORTHERN AFRICA    AFRICA           74           73
## 5   sudan NORTHERN AFRICA    AFRICA           62           60
## 6 tunisia NORTHERN AFRICA    AFRICA           76           74
```

Part 2

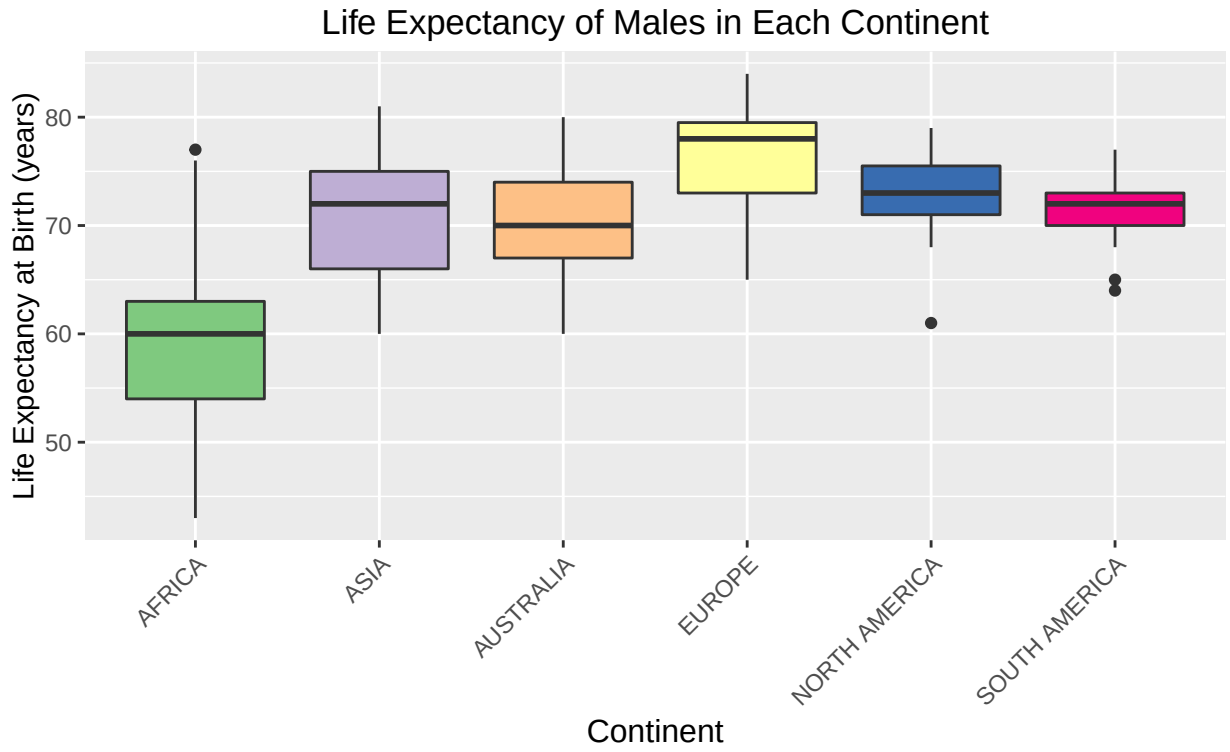
Examining Male Life Expectancy at Birth

To begin with, life expectancy at birth is a property of a population. Specifically, it is defined as the average number of years that a newborn is expected to live if the mortality rates of that population applied in the future (World Health Organization, 2006). The value for life expectancy at birth is traditionally calculated from a *life table* which depicts specific parameters such as the number of surivors, death rate, and the probability of dying of for each age group of a population (PB Silcocks, 2001). As a result, the calculated values for life expectancy at birth do not indicate how long any individual of a certain age is supposed to live, since they assume that the death rates at the time of measurement will not change. In effect, the figures for life expectancy at birth serve as a useful summary statistic to quantify the health of a population.

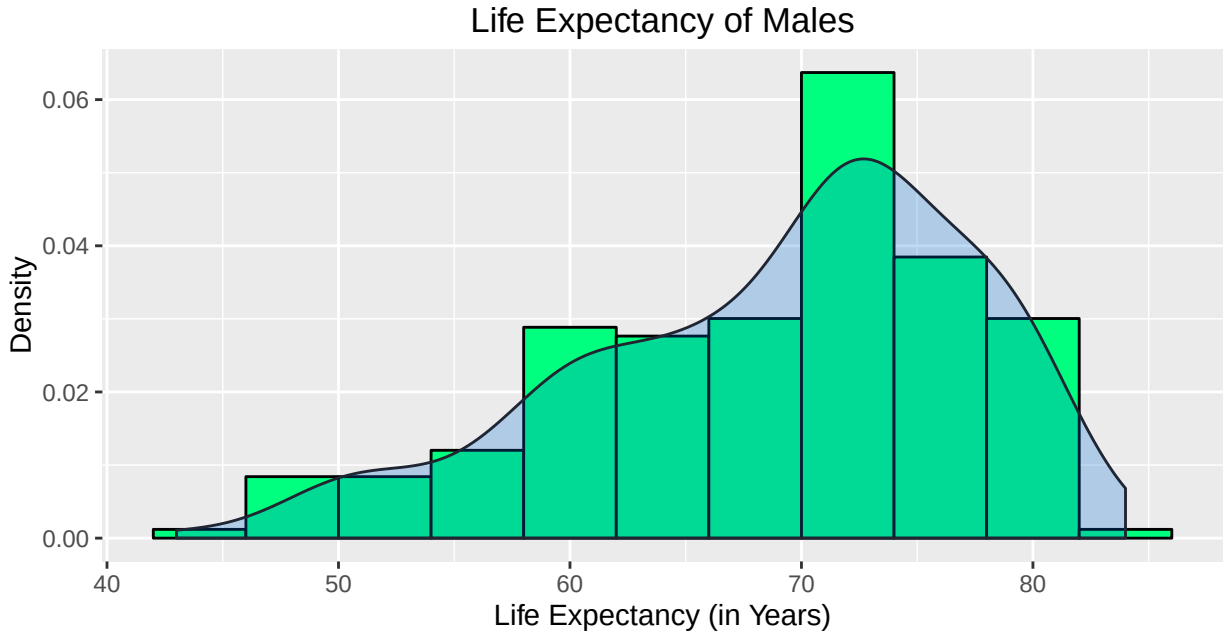
With this in mind, life expectancy at birth is a preferred indicator in epidemiological studies (Arias, 2014; World Health Organization, 2014). For this reason, a number of reasearchers have worked to accurately identify the shape of the sampling distribution of the estimated life expectancy at birth. In particular, Silcocks et al. used the Monte Carlo Simulation method to generate a sampling distribution of life expectancy at birth at the sub-national level (PB Silcocks, 2001). Barman applied Silcocks' techniques to study small tribal states of India (Prasanta Barman, 2010–2011). In both cases they found that the sampling distribution for life expextancy at birth is approximately normal.

Though it is valuable and informative, this finding is not unexpected given the central limit theorem. This fundamental theorem states that the distribution function of the sum of n mutually independent random variables is approximately normal. Examples of variables that might affect life exectancy at birth are access to safe water, immunization, nutritious food and adequate education. An important thing to note about the model for the sampling distribution of male life expectancy at birth is that it has been applied to relatively small populations. On a global scale, we are not guaranteed that this model will accurately predict the sampling distribution.

Below is a box plot of the data collected on expectancy of life at birth for males from the Population Reference Bureau, grouped by continent.



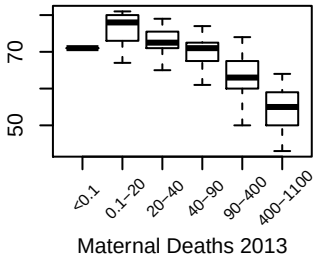
Using the `identify()` function, it was determined that Reunion, Haiti, Boivia, and Guyana were outliers for the life expectancy of males with respect to males in other countries on their respective continents (Africa, North America, and South America). Next, a historgram was created of male life expectancy at birth, with an estimate of the probability density function (pdf) overlayed in blue.

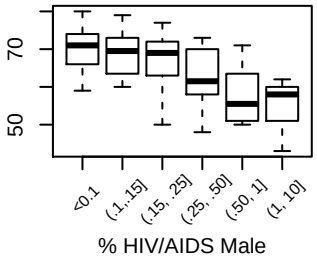


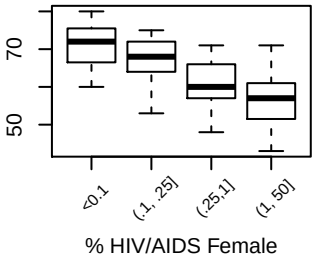
Looking at the histogram and the estimated pdf, the distribution appears to be skewed left. Given the background information presented above, it is reasonable to believe that the life expectancy of males is normally distributed. The central limit theorem suggests that parametric tests work well with large sample sizes, even if the population does not meet the assumptions of normality (Boston University School of Public Health, 2016). For this analysis, our sample size is *relatively* large, with observations from approximately 210 countries and only two missing values. Given the large sample size, it is reasonable to expect that parametric tests would perform reasonably weel, even though the assumptions of normality are not entirely met. This is something that will be kept in mind as the analysis continues.

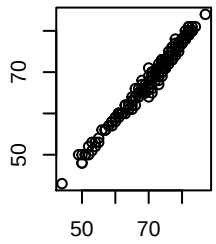
Relationships between Development Indicators

Below is a plot of male life expectancy at birth versus other columns in the data set. The ordinate axis of each plot is male life expectancy at birth (in years), however the labels were not included for aesthetic reasons.

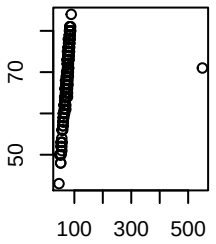




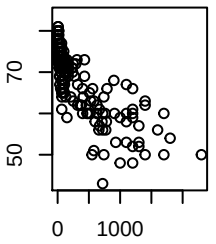




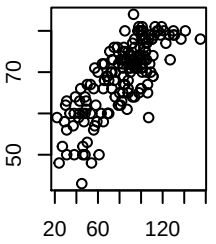
Life Expectancy of Both Males and Females



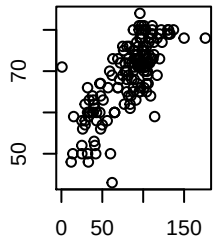
Life Expectancy of Females



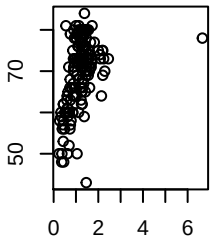
Maternal Deaths in 1990



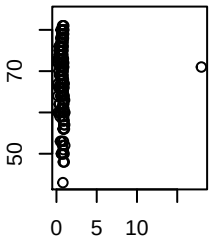
Secondary School Males



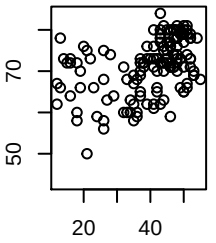
Secondary School Females



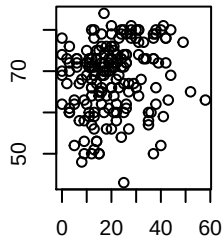
Tertiary School Gender Parity



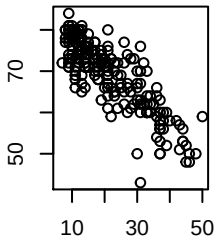
Gender Ratio Labor Force



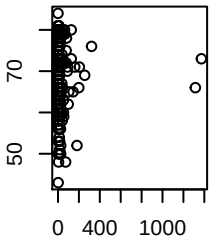
Female Share of Nonagricultural Wage Earner



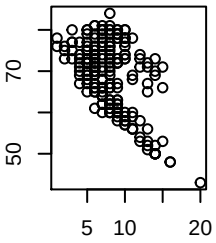
Femal Share of Parliament Members



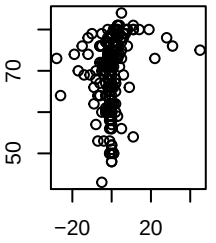
Births (per 100k Population)



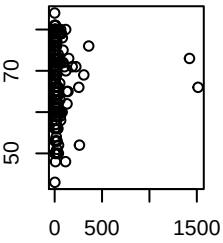
Population mid 2015 (millions)



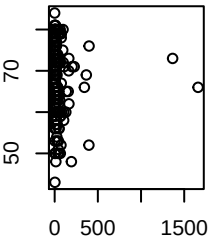
Deaths (per 100k Population)



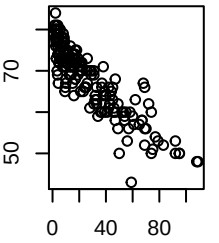
Net Migration Rate (per 100k)



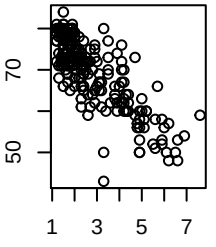
Population mid 2030 (millions)



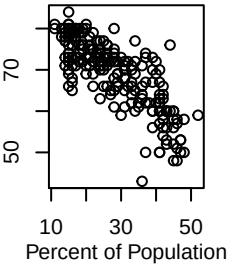
Population mid 2050 (millions)



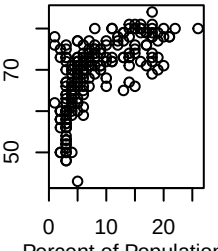
Infant Mortality Rate



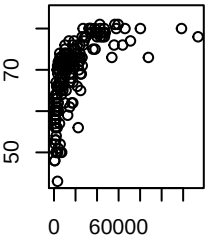
Total Fertility Rate



Percent of Population under 15

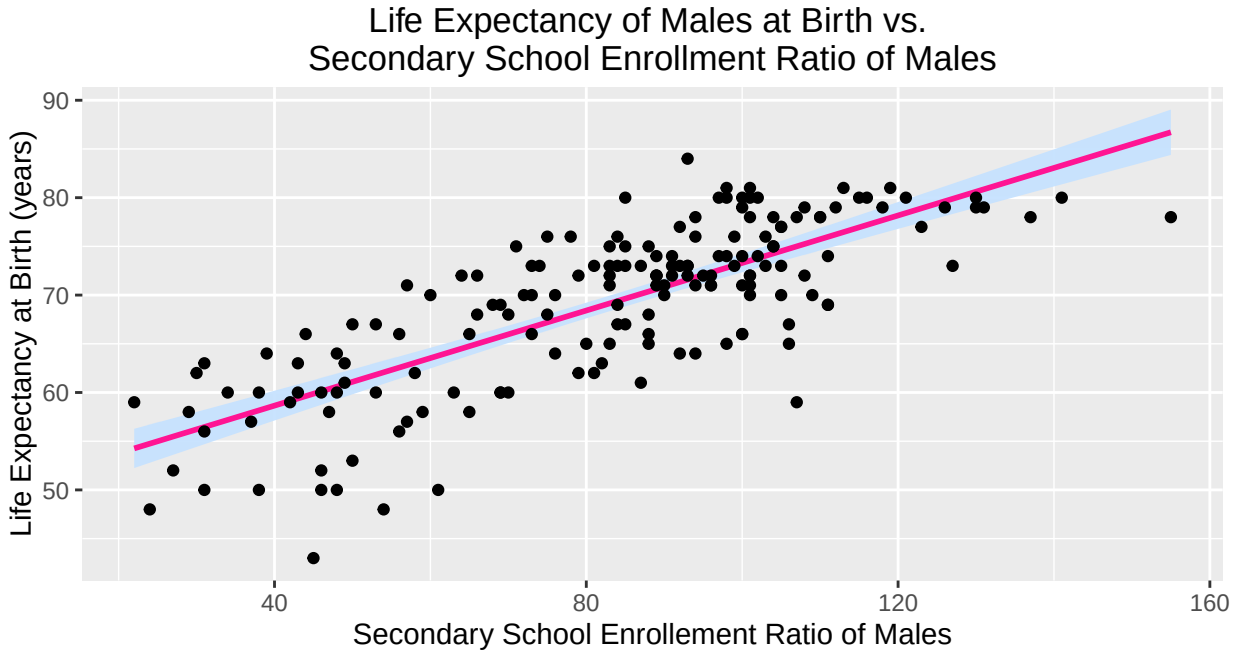


Percent of Population over 65

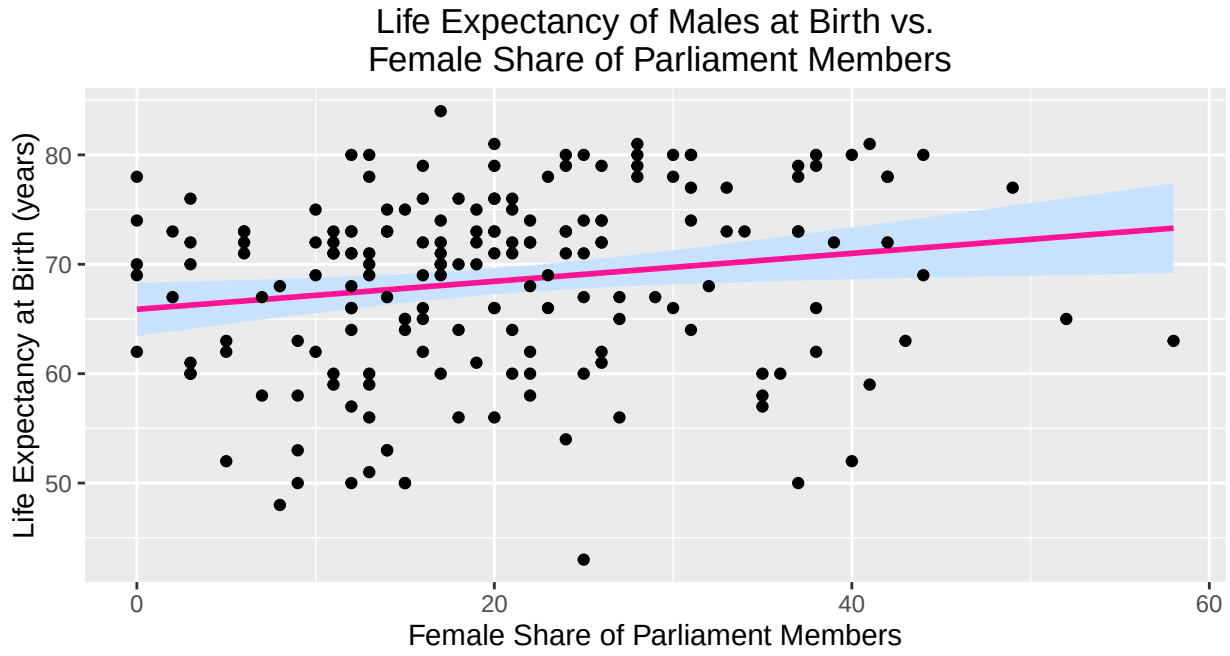


GNI per Capita in 2014

Looking at the figues above, there appears to be a relationship between the life expectancy of males at birth and many other development indicators in the data set. However, some variables seem to have little association with male life expectancy. In particular, secondary education of males seems to be very highly associated, while the female share of parilament memebers appears to be poorly associated. In order to quanatatively assess the degree to which these variables covary with the life expectancy of males at birth, we can perform a correlation test. In this case, a Spearman rank correlation was used due to the fact the test does not assume the data are normally distributed, and to reduce our assumptions about the underlying distributions of female share of parliament members and seconday education of males. Another benefit of using Spearman's rank correlation is that it is robust when outliers are present (Mukaka, 2012). The p-value of the test estimates the probability of obtaining a relationship at least as strong or stronger than that obtained for an uncorrelated system. Below is a plot of life expectancy of males at birth versus seconday education of males, and versus female share of parliament memebers, as well as the results of their respective Spearman rank correlation tests.




```
##
##  Spearman's rank correlation rho
##
## data:  data$Secondary_School_Male and data$LifeExp_Male
## S = 214950, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7509037
```




```
##
##  Spearman's rank correlation rho
##
## data:  data$Female_Share_of_Parliament_Members and data$LifeExp_Male
## S = 825410, p-value = 0.001562
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2303477
```

The results of the correlation tests give rho values of 0.751 and 0.230 for secondary school enrollement ratio of males and female share of parliament members, respectively. In both cases, the p-values are small (ie. less than 0.05). It is important to acknowledge that this only means that the relationship between the two variables is real and not due to random sampling. Moreover, the correlation coefficients describe the degree of association between the variables, and not a causal relationship. The results of the test indicate that the life expectancy of males at birth is strongly associated with secondary school enrollment for males, and weakly associated with the female share of parliament members.

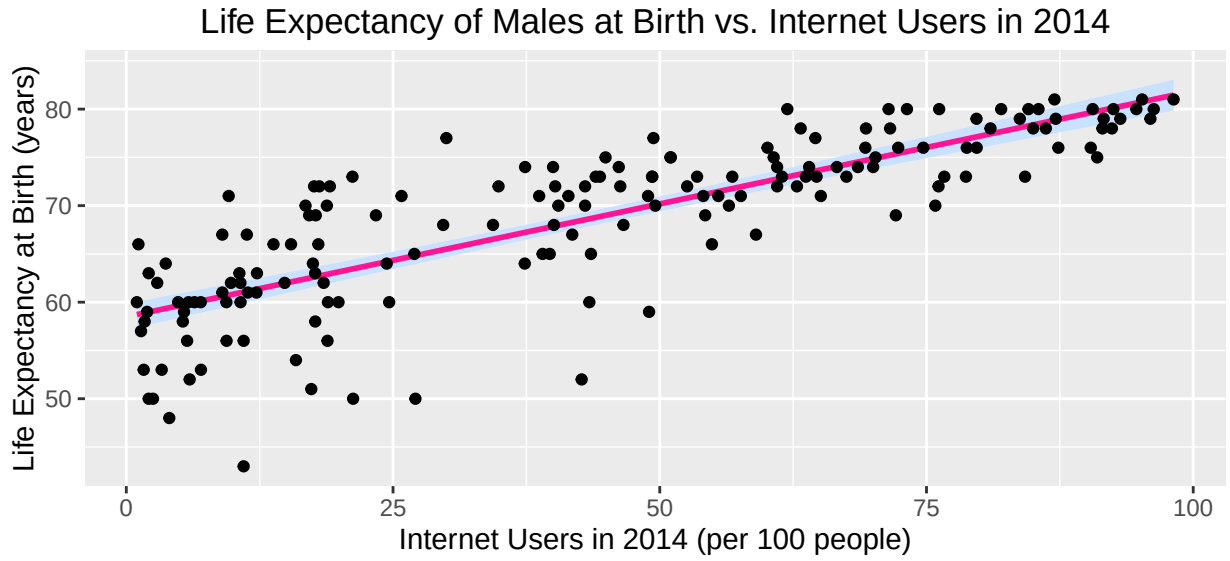
Part 3

Exploratory Analysis

```
source('get_WDIs.R')
```

Male life expetancy versus internet users

- http://data.worldbank.org/indicator/IT.NET.USER.P2




```
##
##  Spearman's rank correlation rho
##
## data:  data$INET_USRS_2014 and data$LifeExp_Male
## S = 111220, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8567144
```

References

Arias, E. (2014). United states life tables, 2009. *National Vital Statistics Reports*, *62*. Journal Article. Retrieved from http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_07.pdf

Boston University School of Public Health. (2016). Nonparametric tests. Retrieved from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/index.html

Mukaka, M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*, 69–71. Journal Article. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/

PB Silcocks, D. J. R. R. (2001). Life expectancy as a summary of mortality in a population: Statistical considerations and suitability for use by health authorities. *J Epidemiol Community Health*, *55*, 38–43. Journal Article. Retrieved from http://jech.bmj.com/content/55/1/38.full.pdf

Prasanta Barman, R. S., Labananda Choudhury. (2010–2011). Sub-state life expectancy estimation using the methodology for small population. Retrieved from https://asianpa.conference-services.net/reports/template/onetextabstract.xml?xsl=template/onetextabstract.xsl&conferenceID=4197&abstractID=864330

World Health Organization. (2006). Definitions and metadata. Online. Retrieved from http://www.who.int/whosis/whostat2006DefinitionsAndMetadata.pdf

World Health Organization. (2014). An overarching health indicator for the post-2015 development agenda. Retrieved from http://www.who.int/healthinfo/indicators/hsi_indicators_SDG_TechnicalMeeting_December2015_BackgroundPaper.pdf