Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit (including R code!).***

Name: Imtiaz Rashid
Lab Day & Time: Tuesday, 3 to 4 PM

Introduction *(5pts)*

The main goal of my study is to investigate how healthcare expenditure per capita and human development by country impacts inequality in life expectancy. The population of interest is all the countries of the world, with available data representing differing economic backgrounds and diverse geographic locations. The stakeholders for my study are countries that display high inequality in life expectancy and individuals of lower socioeconomic status within those countries. Certain geographic regions of the world are stakeholders in the sense that they have a collective interest in reducing the gap of life expectancy within their populations for the goal of long-term development and to raise the standard of living. Finally, additional stakeholders includes businesses, corporations, international NGOS and stated-backed international organizations, who will all play a potential role in addressing this issue and will likely do it in such a manner that also benefits them i.e. profit, contracts, expanding services & products, labor, etc.

Research Questions *(4pts)*

RQ1 – Does a country's health expenditure per capita predict its lifespan inequality?
RQ2 – Is there a linear relationship between human development index & lifespan inequality?

Data Collection Summary *(9pts)*

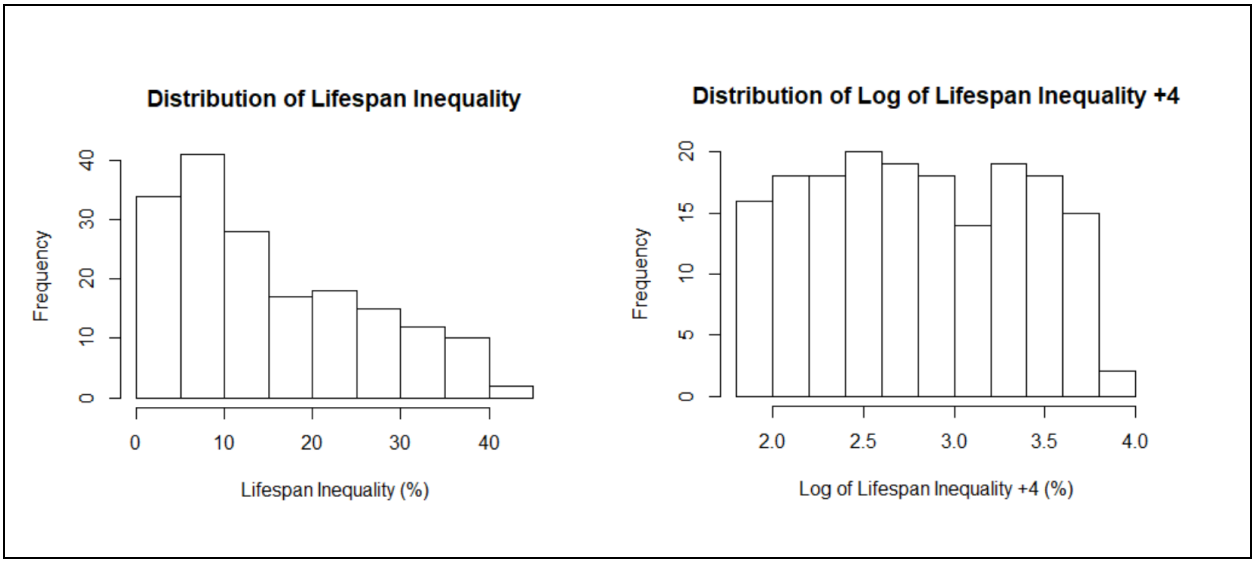
I collected my data from "Our World in Data", an online scientific publication that focuses on critical global issues such as poverty, disease, hunger, climate change, war, inequality and existential risks through the creation of numerous articles, charts and graphics. In my study, the sampling units are individual countries of the world. As my dataset originates from a pre-existing source, I'm unable to say with complete certainty that it'd be considered representative of my population of interest. However, as my study utilizes data from nearly all countries of the word, it's highly likely that the information collected would be representative of the world population. The final sample size is 177 and I did not have to remove any cases from my analysis.

Descriptive Analysis of Response Variable *(9pts) (include graph(s) of distribution here)*

My response variable is the inequality in life expectancy derived from the Atkinson index in percentages (%). The distribution is not normal as it is skewed positively or to the right. Due to the data appearing skewed, the median was used to describe the center whereas the interquartile range (IQR) was used to describe the spread. From the five-number summary, the median is 12.2% while the IQR is 17.9%. There are no outliers present in the data. To improve normality, several types of transformations were attempted, where eventually, taking the logarithm of the data and adding four made its distribution appear the most normal in comparison to other functions, which either skewed the distribution even more or produced no outputs (NaNs).

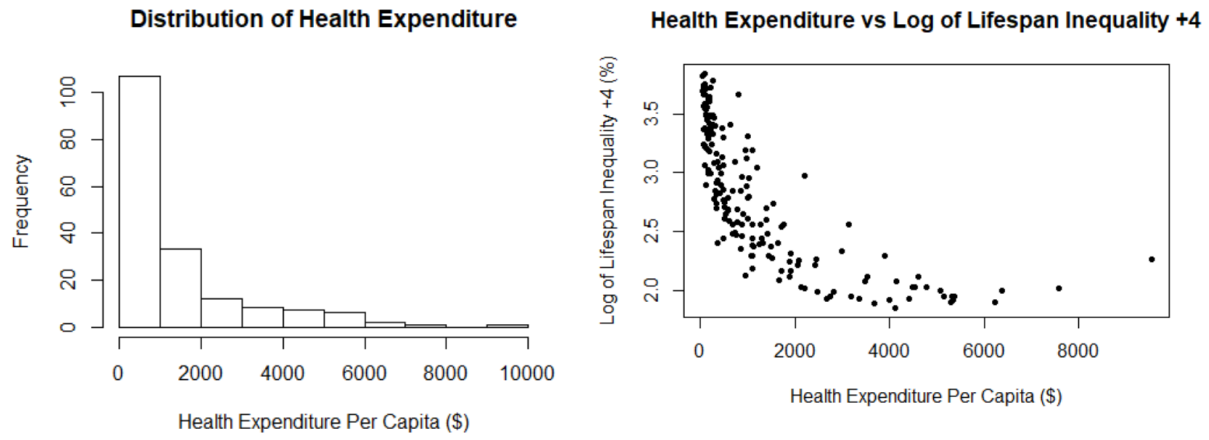
Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit (including R code!).***



<u>Investigation of Explanatory Variable 1</u> *(14pts) (include univariate and bivariate graphs here)*

My first explanatory variable is health expenditure per capita measured in purchasing power parity in current international dollars ($). The distribution is not normal as it is sharply skewed positively or to the right. Due to the data appearing skewed, the median was used to describe the center whereas the interquartile range (IQR) was used to describe the spread. From the five-number summary, the median is $677.466 per capita while the IQR is $1453.767 per capita. Based on the bivariate graph, the relationship between health expenditure per capita against lifespan inequality appears to be largely negative, but possibly quadratic due to a handful of countries spending a large amount on healthcare in comparison to most other countries. The relationship is largely what I expected as it makes sense that increase spending on healthcare would result in lower inequality in life expectancy. However, the possible quadratic relationship may be as a result of outliers due to a few countries spending an abnormally large amount on healthcare, yet still having a somewhat average lifespan inequality.

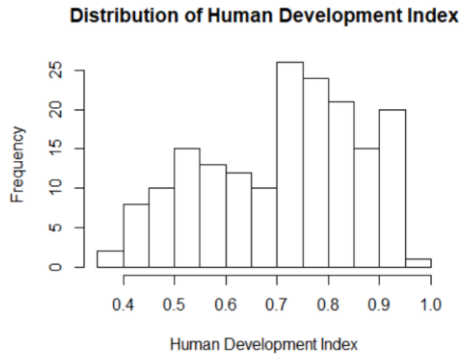


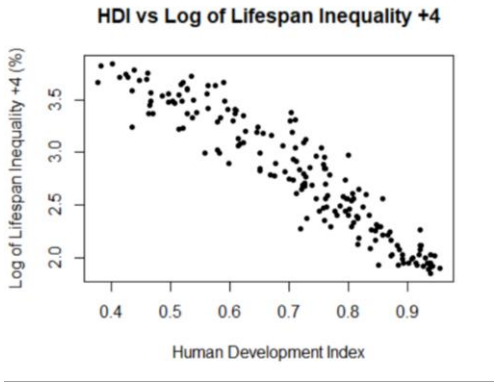
<u>Investigation of Explanatory Variable 2</u> *(14pts) (include univariate and bivariate graphs here)*

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit (including R code!).***

My second explanatory variable is Human Development Index, which has no units. The distribution is not normal as it is skewed negatively or to the left in a bimodal shape. Due to the data appearing skewed, the median was used to describe the center whereas the interquartile range (IQR) was used to describe the spread. From the five-number summary, the median is 0.726 while the IQR is 0.236. Based on the bivariate graph, the relationship between HDI against lifespan inequality appears to be negative. I expect this relationship because progress in human development tends to result in lower inequality in all systems of life.





R Code *(5pts) (organized by variable without output or extraneous syntax)*

Response:

```
hist(FinalDataset$Lifespan.inequality, xlab="Lifespan Inequality (%)", main="Distribution of Lifespan Inequality", right=FALSE)
Lifespan.inequalityT <- log(FinalDataset$Lifespan.inequality+4)
hist(Lifespan.inequalityT, xlab="Log of Lifespan Inequality +4 (%)", main="Distribution of Log of Lifespan Inequality +4", right=FALSE)
fivenum(FinalDataset$Lifespan.inequality)
```

Explanatory 1:

```
hist(FinalDataset$Health.expenditure, xlab="Health Expenditure Per Capita ($)", main="Distribution of Health Expenditure", right=FALSE)
fivenum(FinalDataset$Health.expenditure)
plot(FinalDataset$Health.expenditure, Lifespan.inequalityT, xlab="Health Expenditure Per Capita ($)", ylab="Log of Lifespan Inequality +4 (%)", main="Health Expenditure vs Log of Lifespan Inequality +4", pch=20)
```

Explanatory 2:

```
hist(FinalDataset$HDI, xlab="Human Development Index", main="Distribution of Human Development Index", right=FALSE)
fivenum(FinalDataset$HDI)
plot(FinalDataset$HDI, Lifespan.inequalityT, xlab="Human Development Index", ylab="Log of Lifespan Inequality +4 (%)", main="HDI vs Log of Lifespan Inequality +4", pch=20)
```

Biostatistics SDS 328M – Preliminary Analysis Report

Write in complete sentences using single-spaced, 12-pt font. Include figures in text. ***3-page limit (including R code!).***

Citation (if needed)

Our World in Data, United Nations Development Programme, World Bank, World Health Organization (2018-2019). *Inequality in life expectancy vs. health expenditure per capita, 2015*. Retrieved from https://ourworldindata.org/grapher/inequality-in-life-expectancy-vs-health-expenditure-per-capita