Assignment 1

Bio-Statistics (DSE 401)

Descriptive and Inferential Statistics

Name: Ajay Choudhury **Roll no.:** 18018 **Date:** 13th Feb 2022

Task: Use the data provided as an attachment to find and describe the most striking pattern that you can find. Write a report that describes your pattern, how you found it and what it means. Provide the R code used by you and the plots and statistics generated by you within the report.

Selected Countries: The selected countries for descriptive and inferential statistics here are:

1. United States
2. Germany
3. United Kingdom
4. China
5. India
6. Japan

Setting directory for loading datasets: First of all, for loading all datasets, the working directory is set to the directory where all the datasets are located. The code for the same in R is:

```
# set directory and load data
setwd("E:\\8_Eighth Sem\\Biostatistics\\Load_data\\data")
```

Life Expectancy

Code:

```
# load data set
lifeExpectancyInYears <- read.csv("life_expectancy_years.csv", header = T, check.names = F) # remove X from X1800...

# Plot
na.omit(as.numeric(unlist(lifeExpectancyInYears[lifeExpectancyInYears$country=="United States",])))->USLife
na.omit(as.numeric(unlist(lifeExpectancyInYears[lifeExpectancyInYears$country=="Germany",])))->GermanLife
na.omit(as.numeric(unlist(lifeExpectancyInYears[lifeExpectancyInYears$country=="United Kingdom",])))->UKLife
na.omit(as.numeric(unlist(lifeExpectancyInYears[lifeExpectancyInYears$country=="China",])))->ChinaLife
na.omit(as.numeric(unlist(lifeExpectancyInYears[lifeExpectancyInYears$country=="India",])))->IndiaLife
na.omit(as.numeric(unlist(lifeExpectancyInYears[lifeExpectancyInYears$country=="Japan",])))->JapanLife

min(USLife)
max(USLife)
sd(USLife)
mean(USLife)
median(USLife)
```

```
min(GermanLife)
max(GermanLife)
sd(GermanLife)
mean(GermanLife)
median(GermanLife)

min(UKLife)
max(UKLife)
sd(UKLife)
mean(UKLife)
median(UKLife)

min(ChinaLife)
max(ChinaLife)
sd(ChinaLife)
mean(ChinaLife)
median(ChinaLife)

min(IndiaLife)
max(IndiaLife)
sd(IndiaLife)
mean(IndiaLife)
median(IndiaLife)

min(JapanLife)
max(JapanLife)
sd(JapanLife)
mean(JapanLife)
median(JapanLife)
```

Output:

Country	Minimum	Maximum	Mean	Median	Std. Deviation
United States	31	88.5	62.81096	68.2	18.42187
Germany	29.1	90.7	62.15681	67	20.15504
United Kingdom	37.3	90.6	64.09136	68.5	18.59152
China	22.4	88.6	53.31993	39.5	23.26529
India	8.16	81.8	46.01681	35.2	22.53328
Japan	30.7	93.7	61.25581	59.7	23.03061

The above table shows the minimum, maximum, mean, median and standard deviation of the life expectancy in years in various countries over the period of 301 years i.e. from 1800 to 2100. The dataset here is an extrapolation of existing data till 2100 based on various factors.

Code:

```
par(mfrow=c(3,2))
hist(USLife)
abline(v=mean(USLife))
```

```
text(x=mean(USLife), y=75, label=mean(USLife), col=2)

hist(GermanLife)
abline(v=mean(GermanLife))
text(x=mean(GermanLife), y=75, label=mean(GermanLife), col=2)

hist(UKLife)
abline(v=mean(UKLife))
text(x=mean(UKLife), y=50, label=mean(UKLife), col=2)

hist(ChinaLife)
abline(v=mean(ChinaLife))
text(x=mean(ChinaLife), y=75, label=mean(ChinaLife), col=2)

hist(IndiaLife)
abline(v=mean(IndiaLife))
text(x=mean(IndiaLife), y=75, label=mean(IndiaLife), col=2)

hist(JapanLife)
abline(v=mean(JapanLife))
text(x=mean(JapanLife), y=75, label=mean(JapanLife), col=2)

# box plot
boxplot(USLife, IndiaLife, GermanLife, UKLife, ChinaLife, JapanLife, names = c("US",
"India", "Germany", "UK", "China", "Japan"))
```

Output:

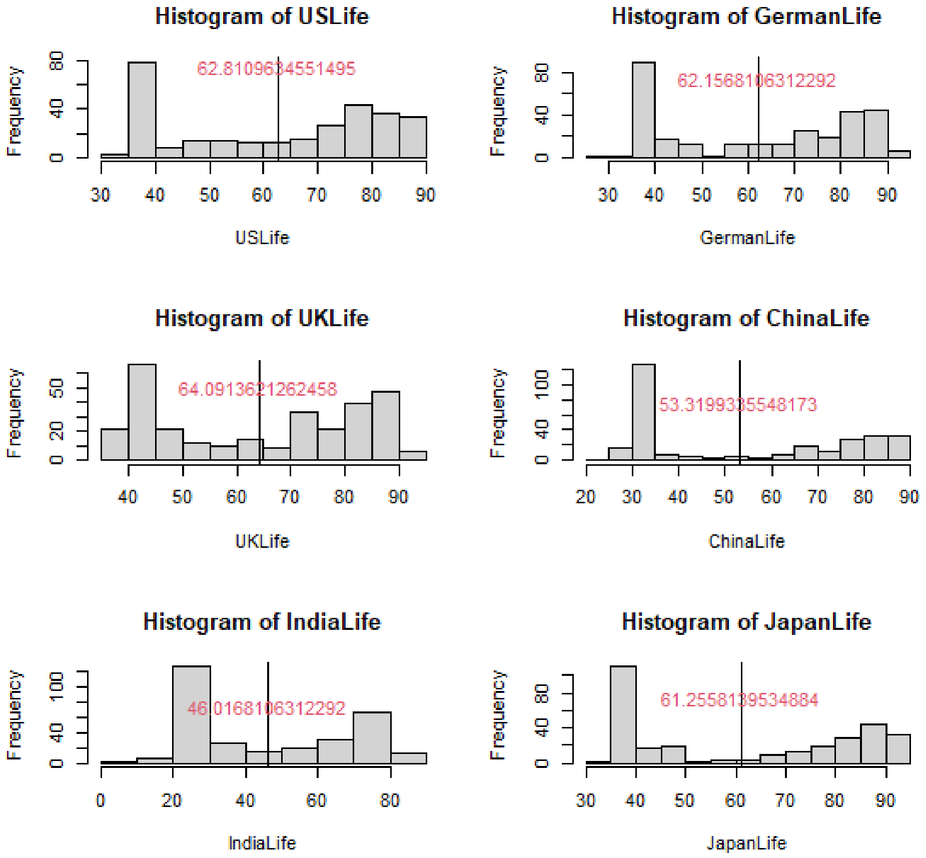

Fig.1: Histograms on life expectancies vs years (with mean)

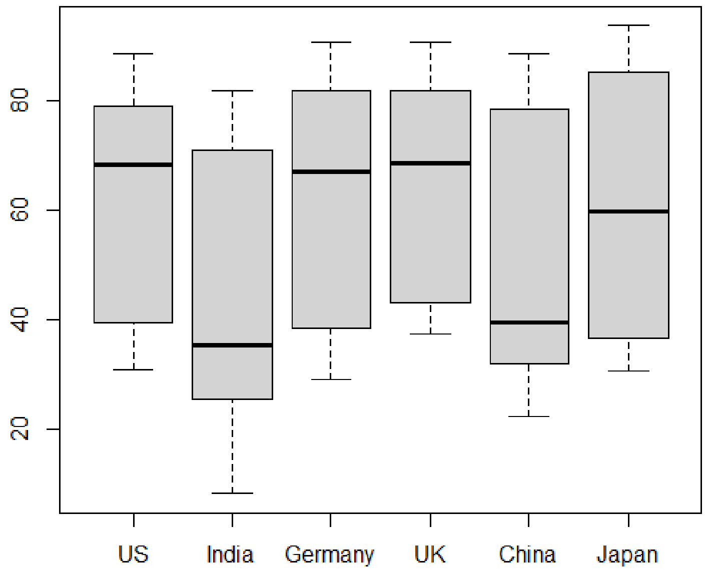

Fig.2: Boxplots of life expectancies in years (with median)

Code:

```
par(mfrow=c(1,1))
plot(c(1800:2100),USLife,col="red",main="Life Expectancy in
Years",pch=15,ylim=c(0,100),xlab="Years",ylab="Age", lwd=2.0, type = "l")
lines(c(1800:2100),GermanLife,col="skyblue", lwd=2.0)
lines(c(1800:2100),UKLife,col="green", lwd=2.0)
lines(c(1800:2100),ChinaLife,col="orange", lwd=2.0)
lines(c(1800:2100),IndiaLife,col="black", lwd=2.0)
lines(c(1800:2100),JapanLife,col="yellow", lwd=2.0)

legend(x = "topleft",                    # Position
       legend = c("United States", "Germany", "United Kingdom", "China", "India",
"Japan"),  # Legend texts
       fill = c("red", "skyblue","green", "orange", "black", "yellow"))              #
Colors

abline(v=1918)
abline(v=2022)
abline(h=20)
```

Output:

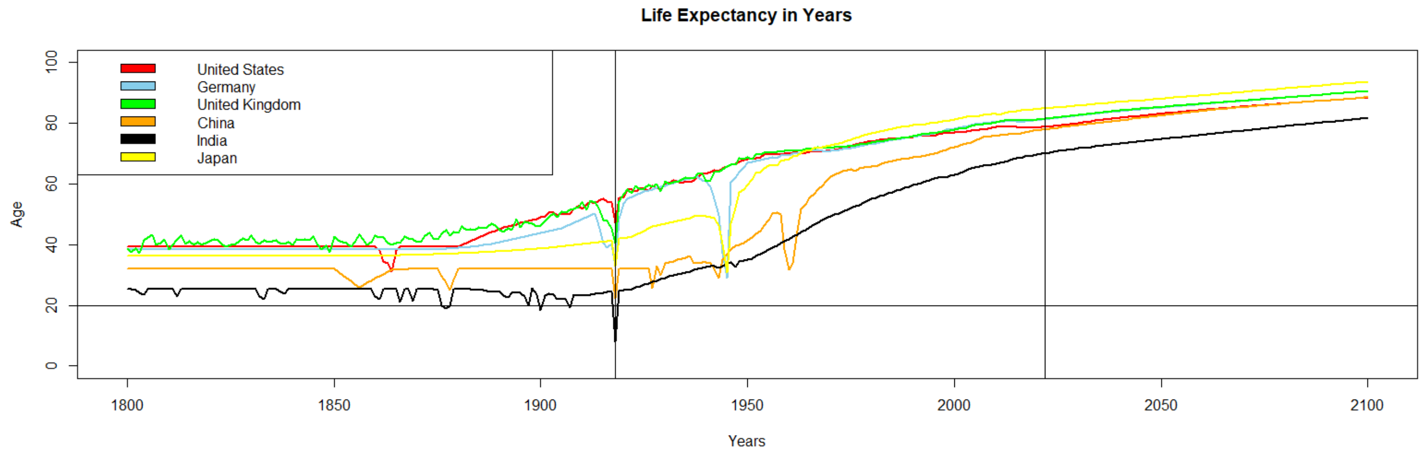

Fig.3: Trend in life expectancy from 1800 to 2100.

Trend:

a. There is a sharp drop in every country's life expectancy in the year 1918 due to the Spanish Flu Pandemic, which was responsible for death of millions of people in India and other countries as well.

```
wilcox.test(IndiaLife[119], IndiaLife[122], paired = T, alternative = "greater")
```

Null hypothesis: India's life expectancy in 1918 is lesser than that of India in 1921.
Here, we get a **p-value of 1** in Wilcoxon test (which is greater than 0.05), that means, the **null hypothesis cannot be rejected** and India's life expectancy in 1918 is less than that in 1921 due to the spread of Spanish flu.

b. Life expectancy of Japan and Germany drops because of the second world war.
c. Every country's life expectancy increases but there was an abrupt drop in the life expectancy of China between the year 1959 and 1962 due to Great Chinese Femine.

```
wilcox.test(ChinaLife[160:163], ChinaLife[167:170], paired = T, alternative =
"greater")
```

Null hypothesis: China's life expectancy between 1959 and 1962 is less than it's life expectancy between 1966 and 1969.
Here, after the Wilcoxon test, we get the **p-value of 1**(which is greater than 0.05), that means, **the null hypothesis cannot be rejected** and hence, China's life expectancy between 1959 and 1962 is lesser it's life expectancy between 1966 and 1969 due to the great chinese femine.

d. Over the last several decades, life expectancy in the countries has improved substantially. As medical care improved and more individuals gained access to healthcare, life expectancy has generally increased.

Inferential Statistics:

- **Shapiro-Wilk Normality Test**

Null Hypothesis: Here the null hypothesis is that the distribution of life expectancies is normal.

Code:

```
# Shapiro test
shapiro.test(USLife) # null hypothesis- the distribution is normal (rejected for all
the data below)
shapiro.test(GermanLife)
shapiro.test(UKLife)
shapiro.test(ChinaLife)
shapiro.test(IndiaLife)
shapiro.test(JapanLife)
```

Output:

Country	p-value	W-value
United States	1.439e-15	0.86482
Germany	< 2.2e-16	0.84367
United Kingdom	4.19e-15	0.87234

China	< 2.2e-16	0.7916
India	< 2.2e-16	0.82507
Japan	< 2.2e-16	0.80376

Since, for all countries' data, the p-value obtained is much lesser than 0.05, **the distributions are not normal and hence t-tests cannot be performed** on these distributions as a common assumption made during a t-test is the normality of distribution.

Here, we can perform Wilcoxon signed-rank test on the distribution as it is a non-parametric test.

- **Wilcoxon Signed-Rank Test**

A few Wilcoxon tests are performed on some countries' life expectancy distribution.

Code:

```
wilcox.test(USLife, GermanLife, paired = T, alternative = "greater")
wilcox.test(IndiaLife,JapanLife, paired = T, alternative = "less")
```

Output:

1. **Null Hypothesis:** The life expectancy in the United States is lesser than that in Germany. Here, after the Wilcoxon test, the p-value obtained is **0.3136** (not less than 0.05), hence the **null hypothesis cannot be rejected** and life expectancy in the US is lesser than that in Germany.
2. **Null Hypothesis:** The life expectancy in India is greater than that in Japan. Here, after the Wilcoxon test, the p-value obtained is less than **2.2e-16**, hence the **null hypothesis can be rejected** and the alternative hypothesis is true i.e. life expectancy in India is lesser than that in Japan.

Total Population

Code:

```
# Population
population <- read.csv("population_total.csv", header = T, check.names = F) # remove X
from X1800...

# Plot
na.omit(as.numeric(unlist(population[population$country=="United States",])))->USLife
na.omit(as.numeric(unlist(population[population$country=="Germany",])))->GermanLife
na.omit(as.numeric(unlist(population[population$country=="United Kingdom",])))->UKLife
na.omit(as.numeric(unlist(population[population$country=="China",])))->ChinaLife
na.omit(as.numeric(unlist(population[population$country=="India",])))->IndiaLife
na.omit(as.numeric(unlist(population[population$country=="Japan",])))->JapanLife

min(USLife)
max(USLife)
sd(USLife)
mean(USLife)
median(USLife)

min(GermanLife)
max(GermanLife)
```

```
sd(GermanLife)
mean(GermanLife)
median(GermanLife)

min(UKLife)
max(UKLife)
sd(UKLife)
mean(UKLife)
median(UKLife)

min(ChinaLife)
max(ChinaLife)
sd(ChinaLife)
mean(ChinaLife)
median(ChinaLife)

min(IndiaLife)
max(IndiaLife)
sd(IndiaLife)
mean(IndiaLife)
median(IndiaLife)

min(JapanLife)
max(JapanLife)
sd(JapanLife)
mean(JapanLife)
median(JapanLife)
```

Output:

Country	Minimum	Maximum	Mean	Median	Std. Deviation
United States	6e+06	4.34e+08	190774817	1.59e+08	147790947
Germany	1.8e+07	83900000	61485714	70700000	20918981
United Kingdom	10800000	78100000	49991362	50600000	19594291
China	3.3e+08	1.46e+09	799392027	5.54e+08	428489229
India	2.01e+08	1.65e+09	746053156	3.93e+08	566445494
Japan	2.8e+07	1.29e+08	74394352	76900000	35783482

The above table shows the minimum, maximum, mean, median and standard deviation of the total population in various countries over the period of 301 years i.e. from 1800 to 2100. The dataset here is an extrapolation of existing data till 2100 based on various factors.

Code:

```
par(mfrow=c(3,2))
hist(USLife)
abline(v=mean(USLife))
text(x=mean(USLife), y=75, label=mean(USLife), col=2)
```

```r
hist(GermanLife)
abline(v=mean(GermanLife))
text(x=mean(GermanLife), y=40, label=mean(GermanLife), col=2)

hist(UKLife)
abline(v=mean(UKLife))
text(x=mean(UKLife), y=30, label=mean(UKLife), col=2)

hist(ChinaLife)
abline(v=mean(ChinaLife))
text(x=mean(ChinaLife), y=70, label=mean(ChinaLife), col=2)

hist(IndiaLife)
abline(v=mean(IndiaLife))
text(x=mean(IndiaLife), y=75, label=mean(IndiaLife), col=2)

hist(JapanLife)
abline(v=mean(JapanLife))
text(x=mean(JapanLife), y=75, label=mean(JapanLife), col=2)

# box plot
boxplot(USLife, IndiaLife, GermanLife, UKLife, ChinaLife, JapanLife, names = c("US",
"India", "Germany", "UK", "China", "Japan"))
```

Output:

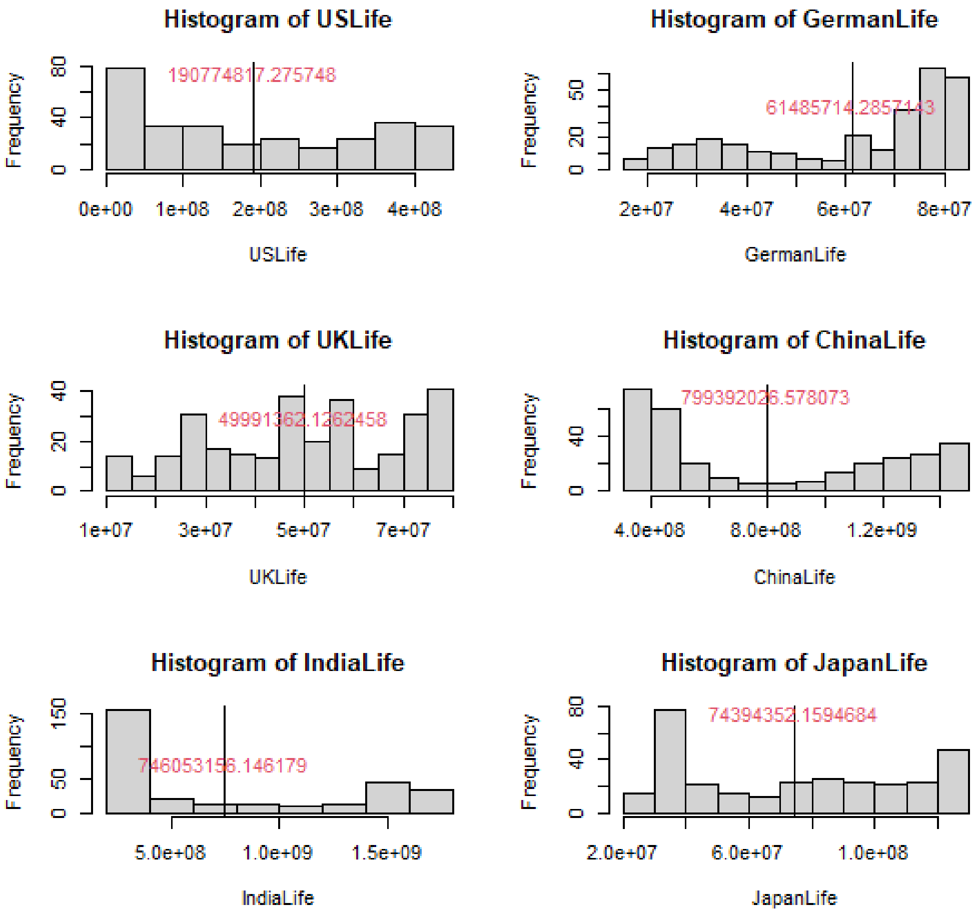

Fig.4: Histograms on total population vs years (with mean)

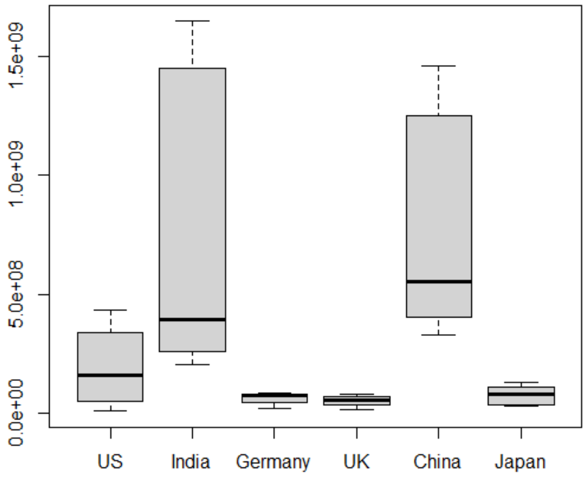

Fig.5: Boxplots of total population in years (with median)

Code:

```
# trend plot
par(mfrow=c(1,1))
plot(c(1800:2100),USLife,col="red",main="Total Population",pch=15,ylim=c(0,1.7E+09),xlab="Years",ylab="Population", lwd=2.0, type = "l")
lines(c(1800:2100),GermanLife,col="skyblue", lwd=2.0)
lines(c(1800:2100),UKLife,col="green", lwd=2.0)
lines(c(1800:2100),ChinaLife,col="orange", lwd=2.0)
lines(c(1800:2100),IndiaLife,col="black", lwd=2.0)
lines(c(1800:2100),JapanLife,col="yellow", lwd=2.0)

legend(x = "topleft",                    # Position
       legend = c("United States", "Germany", "United Kingdom", "China", "India", "Japan"),  # Legend texts
       fill = c("red", "skyblue","green", "orange", "black", "yellow"))             # Colors

abline(v=1950)
abline(v=2022)
```

Output:

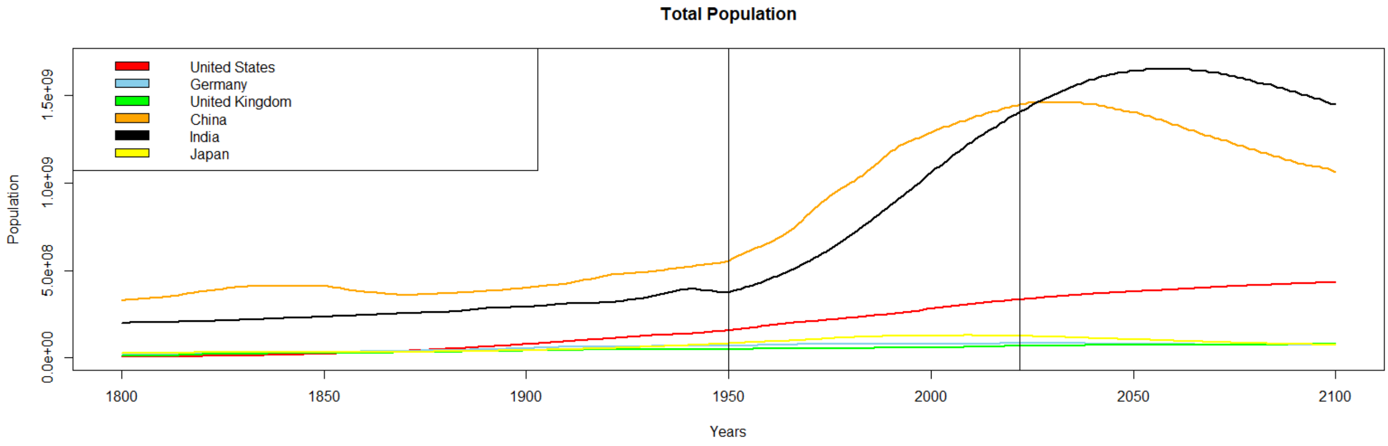

Fig.6: Trend in total populations from 1800 to 2100.

Trend:

a. Decline in the death rates and a persistently high birth rate leads to increse in the total population after the year 1950 in every country.

```
wilcox.test(USLife[131], USLife[161], paired = T, alternative = "greater")
```

Null hypothesis: Population of US in 1930 is lesser than that of US in 1960.
Here, by Wilcoxon test, we get the **p-value of 1**, thus **null hypothesis cannot be rejected**. So, Population of US started increasing rapidly after 1950's i.e. after second industrial revolution.

b. The decline in the death rate has come about due to an increase in life expectancy.
c. From all the 6 countries I have taken only United States doesn't have any saturation point of population while other countries population set to decline after a certain period of time.

Inferential Statistics:

- **Shapiro-Wilk Normality Test**

Null Hypothesis: Here the null hypothesis is that the distribution of life expectancies is normal.

Code:

```
# Shapiro test
shapiro.test(USLife) # null hypothesis- the distribution is normal (rejected for all the data below)
shapiro.test(GermanLife)
shapiro.test(UKLife)
shapiro.test(ChinaLife)
shapiro.test(IndiaLife)
shapiro.test(JapanLife)
```

Output:

Country	**p-value**	**W-value**
United States	8.169e-14	0.89169
Germany	< 2.2e-16	0.84521
United Kingdom	2.904e-09	0.94412
China	< 2.2e-16	0.81458
India	< 2.2e-16	0.78301
Japan	3.029e-14	0.88548

Since, for all countries' data, the p-value obtained is much lesser than 0.05, **the distributions are not normal and hence t-tests cannot be performed** on these distributions as a common assumption made during a t-test is the normality of distribution.

Here, we can perform Wilcoxon signed-rank test on the distribution as it is a non-parametric test.

- **Wilcoxon Signed-Rank Test**

A few Wilcoxon tests are performed on some countries' life expectancy distribution.

Code:

```
wilcox.test(USLife, GermanLife, paired = T, alternative = "greater")
wilcox.test(IndiaLife,JapanLife, paired = T, alternative = "less")
```

Output:

1. **Null Hypothesis:** The total population in the United States is lesser than that in Germany. Here, after the Wilcoxon test, the p-value obtained is **less than 2.2e-16** (less than 0.05), hence the **null hypothesis can be rejected** and alternative hypothesis is true, i.e. total population in the US is greater than that in Germany.
2. **Null Hypothesis:** The total population in India is greater than that in Japan. Here, after the Wilcoxon test, the p-value obtained is 1, hence the **null hypothesis cannot be rejected** and the total population in India is greater than that in Japan.

GDP Per Capita

Code:

```
# GDP Per Capita
gdp <- read.csv("income_per_person_gdppercapita_ppp_inflation_adjusted.csv", header =
T, check.names = F) # remove X from X1800...

# Plot
na.omit(as.numeric(unlist(gdp[gdp$country=="United States",])))->USLife
na.omit(as.numeric(unlist(gdp[gdp$country=="Germany",])))->GermanLife
na.omit(as.numeric(unlist(gdp[gdp$country=="United Kingdom",])))->UKLife
na.omit(as.numeric(unlist(gdp[gdp$country=="China",])))->ChinaLife
na.omit(as.numeric(unlist(gdp[gdp$country=="India",])))->IndiaLife
na.omit(as.numeric(unlist(gdp[gdp$country=="Japan",])))->JapanLife

min(USLife)
max(USLife)
sd(USLife)
mean(USLife)
median(USLife)

min(GermanLife)
max(GermanLife)
sd(GermanLife)
mean(GermanLife)
median(GermanLife)

min(UKLife)
max(UKLife)
sd(UKLife)
mean(UKLife)
median(UKLife)

min(ChinaLife)
max(ChinaLife)
sd(ChinaLife)
mean(ChinaLife)
median(ChinaLife)
```

```
min(IndiaLife)
max(IndiaLife)
sd(IndiaLife)
mean(IndiaLife)
median(IndiaLife)

min(JapanLife)
max(JapanLife)
sd(JapanLife)
mean(JapanLife)
median(JapanLife)
```

Output:

Country	Minimum	Maximum	Mean	Median	Std. Deviation
United States	1970	80000	19022.03	8130	20969.09
Germany	1990	66200	15935.93	6560	17336.51
United Kingdom	3040	56500	14975.48	7970	14188.17
China	560	34800	3906.071	785	7895.737
India	705	15400	2216.353	890	3299.305
Japan	1010	55200	11805.39	2620	15782.87

The above table shows the minimum, maximum, mean, median and standard deviation of the GDP per capita income in various countries over the period of 241 years i.e. from 1800 to 2040. The dataset here is an extrapolation of existing data till 2040 based on various factors.

Code:

```
par(mfrow=c(3,2))
hist(USLife)
abline(v=mean(USLife))
text(x=mean(USLife), y=75, label=mean(USLife), col=2)

hist(GermanLife)
abline(v=mean(GermanLife))
text(x=mean(GermanLife), y=75, label=mean(GermanLife), col=2)

hist(UKLife)
abline(v=mean(UKLife))
text(x=mean(UKLife), y=50, label=mean(UKLife), col=2)

hist(ChinaLife)
abline(v=mean(ChinaLife))
text(x=mean(ChinaLife), y=75, label=mean(ChinaLife), col=2)

hist(IndiaLife)
abline(v=mean(IndiaLife))
text(x=mean(IndiaLife), y=75, label=mean(IndiaLife), col=2)
```

```
hist(JapanLife)
abline(v=mean(JapanLife))
text(x=mean(JapanLife), y=75, label=mean(JapanLife), col=2)

# box plot
boxplot(USLife, IndiaLife, GermanLife, UKLife, ChinaLife, JapanLife, names = c("US",
"India", "Germany", "UK", "China", "Japan"), outline = F)
```

Output:

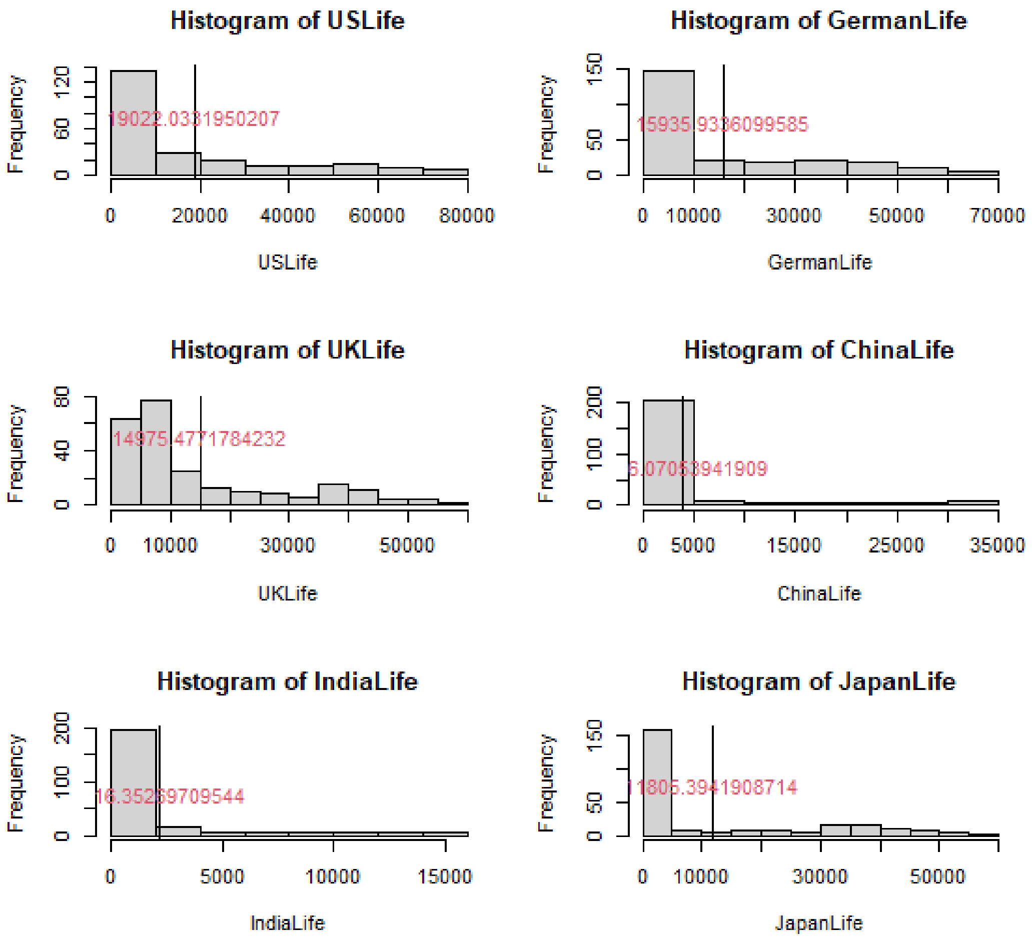

Fig.7: Histograms on GDP Per Capita vs years (with mean)

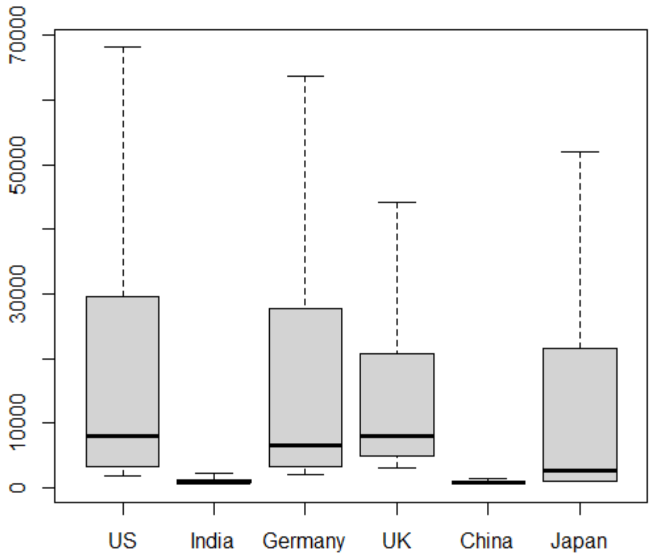

Fig.8: Boxplots of GDP Per Capita in years (with median)

Code:

```
par(mfrow=c(1,1))
plot(c(1800:2040),USLife,col="red",main="GDP Per
Capita",pch=15,ylim=c(0,90000),xlab="Years",ylab="GDP Per Capita", lwd=2.0, type =
"l")
lines(c(1800:2040),GermanLife,col="skyblue", lwd=2.0)
lines(c(1800:2040),UKLife,col="green", lwd=2.0)
lines(c(1800:2040),ChinaLife,col="orange", lwd=2.0)
lines(c(1800:2040),IndiaLife,col="black", lwd=2.0)
lines(c(1800:2040),JapanLife,col="yellow", lwd=2.0)

legend(x = "topleft",                    # Position
       legend = c("United States", "Germany", "United Kingdom", "China", "India",
"Japan"),  # Legend texts
       fill = c("red", "skyblue","green", "orange", "black", "yellow"))             #
Colors

abline(v=1950)
abline(v=2022)
```

Output:

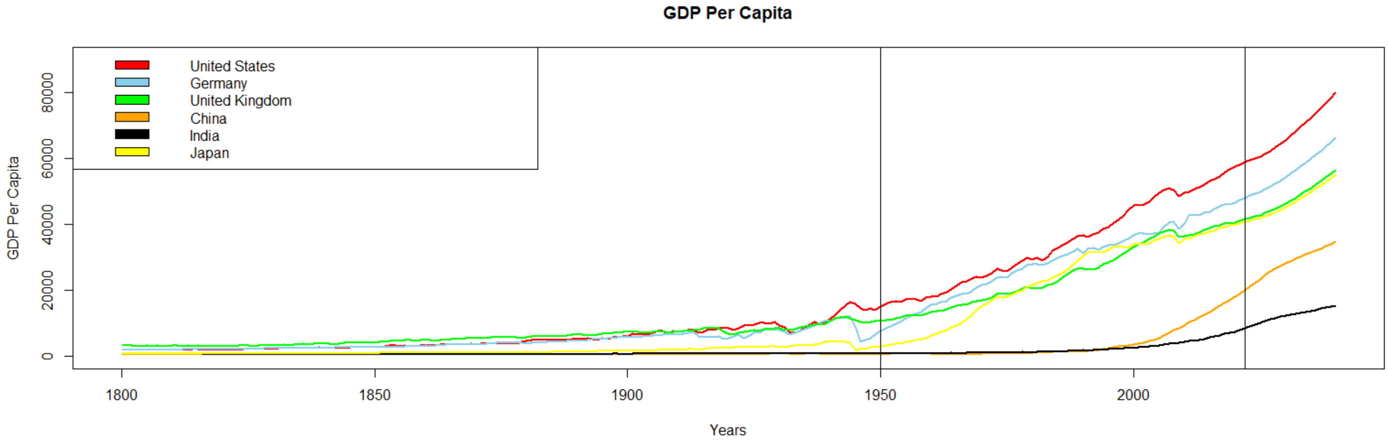

Fig.9: Trend in GDP Per Capita from 1800 to 2040.

Trend:

a. From the start of the 1800s, every countries population tend to increase but in the United States, the Great Depression occurred(the worst economic downturn in the history of the industrialized world), lasting from 1929 to 1939. By 1933, when the Great Depression reached its lowest point, some 15 million Americans were unemployed and nearly half the country's banks had failed.

```
wilcox.test(USLife[131:136],USLife[146:151], paired = T, alternative = "greater")
```

Null hypothesis: GDP per capita between 1930 and 1935 in US is lesser than GDP per capita in US between 1945 and 1950.
Here, by the Wilcoxon test, we get the **p-value of 1**, which means **the null hypothesis cannot be rejected**. Hence, it is true that due to the great depression, the GDP per capita fell for the US between 1930 and 1935 approximately.

b. Also during the time of the 2nd world war, the GDP per capita of Germany and Japan falls abruptly.

c. From the start of the 1800s until the 1990s India's and China's GDP per capita almost followed a similar trend but from the early 1990s, China's GDP per capita increased drastically.

Inferential Statistics:

- **Shapiro-Wilk Normality Test**

Null Hypothesis: Here the null hypothesis is that the distribution of life expectancies is normal.

Code:

```
# Shapiro test
shapiro.test(USLife) # null hypothesis- the distribution is normal (rejected for all
the data below)
shapiro.test(GermanLife)
shapiro.test(UKLife)
shapiro.test(ChinaLife)
shapiro.test(IndiaLife)
shapiro.test(JapanLife)
```

Output:

Country	**p-value**	**W-value**
United States	< 2.2e-16	0.7832
Germany	< 2.2e-16	0.77256
United Kingdom	< 2.2e-16	0.77845
China	< 2.2e-16	0.46196
India	< 2.2e-16	0.49894
Japan	< 2.2e-16	0.70306

Since, for all countries' data, the p-value obtained is much lesser than 0.05, **the distributions are not normal and hence t-tests cannot be performed** on these distributions as a common assumption made during a t-test is the normality of distribution.

Here, we can perform Wilcoxon signed-rank test on the distribution as it is a non-parametric test.

- **Wilcoxon Signed-Rank Test**

A few Wilcoxon tests are performed on some countries' life expectancy distribution.

Code:

```
wilcox.test(USLife, GermanLife, paired = T, alternative = "greater")
wilcox.test(IndiaLife,JapanLife, paired = T, alternative = "less")
```

Output:

3. **Null Hypothesis:** The GDP per capita income in the United States is lesser than that in Germany. Here, after the Wilcoxon test, the p-value obtained is **less than 2.2e-16** (less than 0.05), hence the **null hypothesis can be rejected** and the alternative hypothesis is true i.e. GDP per capita income in the US is greater than that in Germany.
4. **Null Hypothesis:** The GDP per capita income in India is greater than that in Japan. Here, after

the Wilcoxon test, the p-value obtained is **less than 2.2e-16**, hence the **null hypothesis can be rejected** and the alternative hypothesis is true, i.e. GDP per capita income in India is lesser than that in Japan.

Children Per Woman

Code:

```
# Children per Woman
CPWoman <- read.csv("children_per_woman_total_fertility.csv", header = T, check.names = F) # remove X from X1800...

# Plot
na.omit(as.numeric(unlist(CPWoman[CPWoman$country=="United States",])))->USLife
na.omit(as.numeric(unlist(CPWoman[CPWoman$country=="Germany",])))->GermanLife
na.omit(as.numeric(unlist(CPWoman[CPWoman$country=="United Kingdom",])))->UKLife
na.omit(as.numeric(unlist(CPWoman[CPWoman$country=="China",])))->ChinaLife
na.omit(as.numeric(unlist(CPWoman[CPWoman$country=="India",])))->IndiaLife
na.omit(as.numeric(unlist(CPWoman[CPWoman$country=="Japan",])))->JapanLife

min(USLife)
max(USLife)
sd(USLife)
mean(USLife)
median(USLife)

min(GermanLife)
max(GermanLife)
sd(GermanLife)
mean(GermanLife)
median(GermanLife)

min(UKLife)
max(UKLife)
sd(UKLife)
mean(UKLife)
median(UKLife)

min(ChinaLife)
max(ChinaLife)
sd(ChinaLife)
mean(ChinaLife)
median(ChinaLife)

min(IndiaLife)
max(IndiaLife)
sd(IndiaLife)
mean(IndiaLife)
median(IndiaLife)

min(JapanLife)
max(JapanLife)
```

```
sd(JapanLife)
mean(JapanLife)
median(JapanLife)
```

Output:

Country	Minimum	Maximum	Mean	Median	Std. Deviation
United States	1.74	7.03	3.434319	2.57	1.744066
Germany	1.31	5.46	2.998272	2.1	1.620489
United Kingdom	1.67	6.02	2.996777	2.1	1.404009
China	1.49	7.41	3.951063	5.4	1.8218
India	1.77	5.95	4.448173	5.73	1.794985
Japan	1.3	5.35	3.104419	3.34	1.41446

The above table shows the minimum, maximum, mean, median and standard deviation of the Children per woman in various countries over the period of 301 years i.e. from 1800 to 2100. The dataset here is an extrapolation of existing data till 2100 based on various factors.

Code:

```
par(mfrow=c(3,2))
hist(USLife)
abline(v=mean(USLife))
text(x=mean(USLife), y=75, label=mean(USLife), col=2)

hist(GermanLife)
abline(v=mean(GermanLife))
text(x=mean(GermanLife), y=75, label=mean(GermanLife), col=2)

hist(UKLife)
abline(v=mean(UKLife))
text(x=mean(UKLife), y=50, label=mean(UKLife), col=2)

hist(ChinaLife)
abline(v=mean(ChinaLife))
text(x=mean(ChinaLife), y=75, label=mean(ChinaLife), col=2)

hist(IndiaLife)
abline(v=mean(IndiaLife))
text(x=mean(IndiaLife), y=75, label=mean(IndiaLife), col=2)

hist(JapanLife)
abline(v=mean(JapanLife))
text(x=mean(JapanLife), y=75, label=mean(JapanLife), col=2)

# box plot
boxplot(USLife, IndiaLife, GermanLife, UKLife, ChinaLife, JapanLife, names = c("US",
"India", "Germany", "UK", "China", "Japan"))
```

Output:

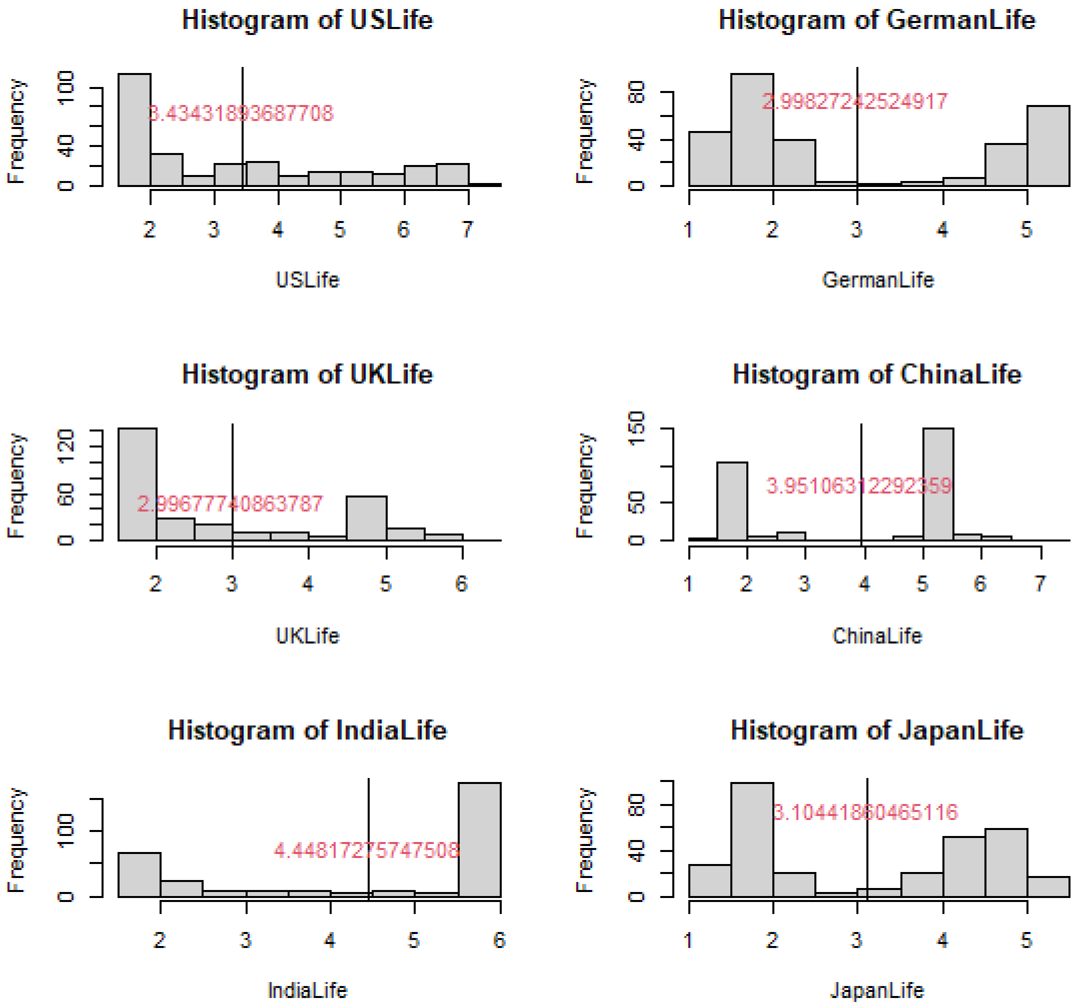

Fig.10: Histograms on Children per woman vs years (with mean)

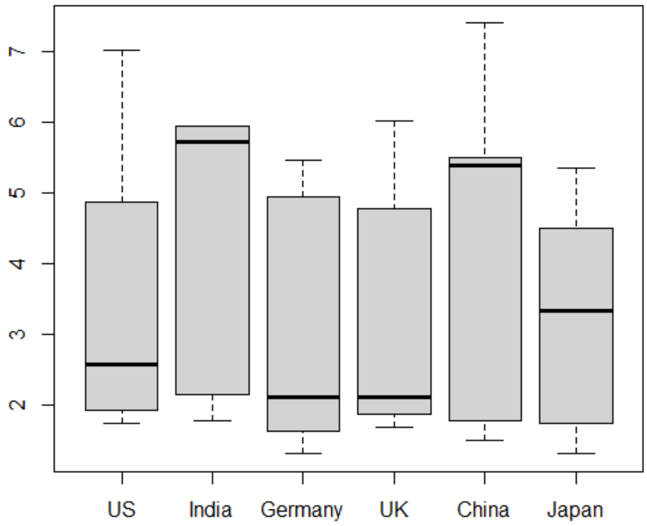

Fig.11: Boxplots of Children per woman in years (with median)

Code:

```
par(mfrow=c(1,1))
```

```
plot(c(1800:2100),USLife,col="red",main="Children Per
Woman",pch=15,ylim=c(0,9),xlab="Years",ylab="Population", lwd=2.0, type = "l")
lines(c(1800:2100),GermanLife,col="skyblue", lwd=2.0)
lines(c(1800:2100),UKLife,col="green", lwd=2.0)
lines(c(1800:2100),ChinaLife,col="orange", lwd=2.0)
lines(c(1800:2100),IndiaLife,col="black", lwd=2.0)
lines(c(1800:2100),JapanLife,col="yellow", lwd=2.0)

legend(x = "topright",                    # Position
       legend = c("United States", "Germany", "United Kingdom", "China", "India",
"Japan"),  # Legend texts
       fill = c("red", "skyblue","green", "orange", "black", "yellow"))              #
Colors

abline(v=1960)
abline(v=2022)
```

Output:

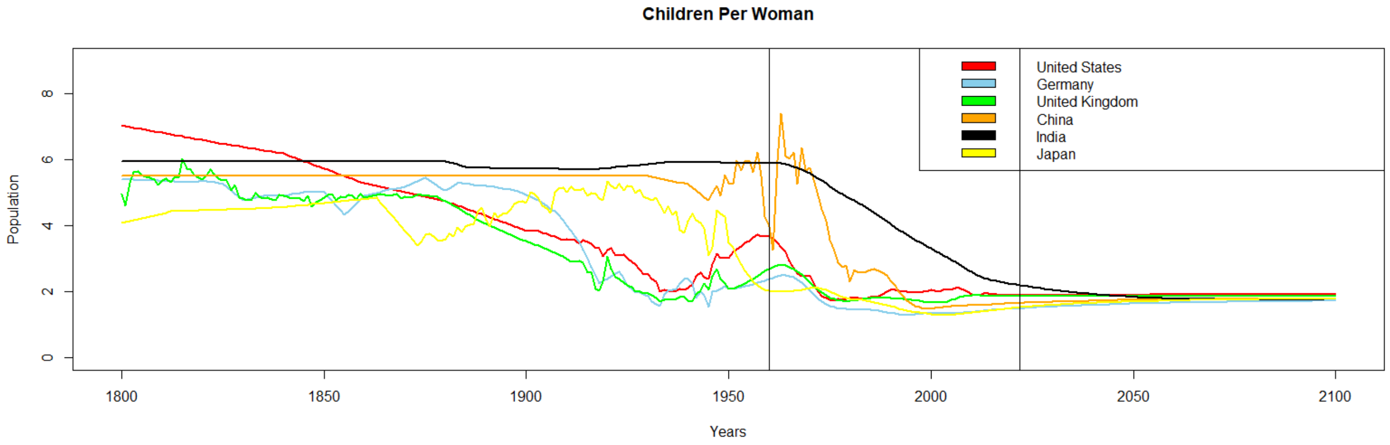

Fig.12: Trend in Children per woman from 1800 to 2100.

Trend:

a. According to different government's different child policies in China. Until 1960s, the government mostly encouraged families to have as many children as possible in China. After that there came 1 child policy in china so women per child drops.

```
wilcox.test(ChinaLife[160:162], ChinaLife[167:169], paired = T, alternative =
"greater")
```

Null hypothesis: Children per woman between 1959 and 1961 in China decreased as compared to that in between in 1966 and 1968.
Here, by the Wilcoxon test, we get the **p-value as 1**, which means **the null hypothesis cannot be rejected**. Due to government's one child policy, the number of children per woman decreased substantially between 1959 and 1961 in China.

b. In india Until 1960s almost 6 women per child was there but after that due to better contraception initiatives and government health and family welfare schemes it drops abruptly.

Inferential Statistics:

- **Shapiro-Wilk Normality Test**

Null Hypothesis: Here the null hypothesis is that the distribution of life expectancies is normal.

Code:

```
# Shapiro test
shapiro.test(USLife) # null hypothesis- the distribution is normal (rejected for all the data below)
shapiro.test(GermanLife)
shapiro.test(UKLife)
shapiro.test(ChinaLife)
shapiro.test(IndiaLife)
shapiro.test(JapanLife)
```

Output:

Country	**p-value**	**W-value**
United States	< 2.2e-16	0.81867
Germany	< 2.2e-16	0.76956
United Kingdom	< 2.2e-16	0.7769
China	< 2.2e-16	0.72573
India	< 2.2e-16	0.7169
Japan	< 2.2e-16	0.81992

Since, for all countries' data, the p-value obtained is much lesser than 0.05, **the distributions are not normal and hence t-tests cannot be performed** on these distributions as a common assumption made during a t-test is the normality of distribution.

Here, we can perform Wilcoxon signed-rank test on the distribution as it is a non-parametric test.

- **Wilcoxon Signed-Rank Test**

A few Wilcoxon tests are performed on some countries' life expectancy distribution.

Code:

```
wilcox.test(USLife, GermanLife, paired = T, alternative = "greater")
wilcox.test(IndiaLife,JapanLife, paired = T, alternative = "less")
```

Output:

5. **Null Hypothesis:** Children per woman in the United States is lesser than that in Germany. Here, after the Wilcoxon test, the p-value obtained is **less than 2.2e-16** (less than 0.05), hence the **null hypothesis can be rejected** and the alternative hypothesis is true i.e. Children per woman in the US is greater than that in Germany.
6. **Null Hypothesis:** Children per woman in India is greater than that in Japan. Here, after the Wilcoxon test, the p-value obtained is **1**, hence the **null hypothesis cannot be rejected** and the alternative hypothesis is false, i.e. Children per woman in India is greater than that in Japan.

Child Mortality

Code:

```r
# Child Mortality
ChildMortality <- read.csv("child_mortality_0_5_year_olds_dying_per_1000_born.csv", header = T, check.names = F) # remove X from X1800...

# Plot
na.omit(as.numeric(unlist(ChildMortality[ChildMortality$country=="United States",])))->USLife
na.omit(as.numeric(unlist(ChildMortality[ChildMortality$country=="Germany",])))->GermanLife
na.omit(as.numeric(unlist(ChildMortality[ChildMortality$country=="United Kingdom",])))->UKLife
na.omit(as.numeric(unlist(ChildMortality[ChildMortality$country=="China",])))->ChinaLife
na.omit(as.numeric(unlist(ChildMortality[ChildMortality$country=="India",])))->IndiaLife
na.omit(as.numeric(unlist(ChildMortality[ChildMortality$country=="Japan",])))->JapanLife

min(USLife)
max(USLife)
sd(USLife)
mean(USLife)
median(USLife)

min(GermanLife)
max(GermanLife)
sd(GermanLife)
mean(GermanLife)
median(GermanLife)

min(UKLife)
max(UKLife)
sd(UKLife)
mean(UKLife)
median(UKLife)

min(ChinaLife)
max(ChinaLife)
sd(ChinaLife)
mean(ChinaLife)
median(ChinaLife)

min(IndiaLife)
max(IndiaLife)
sd(IndiaLife)
mean(IndiaLife)
median(IndiaLife)

min(JapanLife)
max(JapanLife)
```

```
sd(JapanLife)
mean(JapanLife)
median(JapanLife)
```

Output:

Country	Minimum	Maximum	Mean	Median	Std. Deviation
United States	1.7	329	132.5814	37.6	141.6974
Germany	0.86	539	170.609	60.2	186.2398
United Kingdom	0.86	329	114.5514	36.6	122.1242
China	2.17	500	227.5	317	191.2179
India	6.13	537	264.1654	266	207.9043
Japan	0.49	363	163.0502	91.3	163.7896

The above table shows the minimum, maximum, mean, median and standard deviation of the Children mortality in various countries over the period of 301 years i.e. from 1800 to 2100. The dataset here is an extrapolation of existing data till 2100 based on various factors.

Code:

```
par(mfrow=c(3,2))
hist(USLife)
abline(v=mean(USLife))
text(x=mean(USLife), y=75, label=mean(USLife), col=2)

hist(GermanLife)
abline(v=mean(GermanLife))
text(x=mean(GermanLife), y=75, label=mean(GermanLife), col=2)

hist(UKLife)
abline(v=mean(UKLife))
text(x=mean(UKLife), y=50, label=mean(UKLife), col=2)

hist(ChinaLife)
abline(v=mean(ChinaLife))
text(x=mean(ChinaLife), y=75, label=mean(ChinaLife), col=2)

hist(IndiaLife)
abline(v=mean(IndiaLife))
text(x=mean(IndiaLife), y=75, label=mean(IndiaLife), col=2)

hist(JapanLife)
abline(v=mean(JapanLife))
text(x=mean(JapanLife), y=75, label=mean(JapanLife), col=2)

# box plot
boxplot(USLife, IndiaLife, GermanLife, UKLife, ChinaLife, JapanLife, names = c("US",
"India", "Germany", "UK", "China", "Japan"))
```

Output:

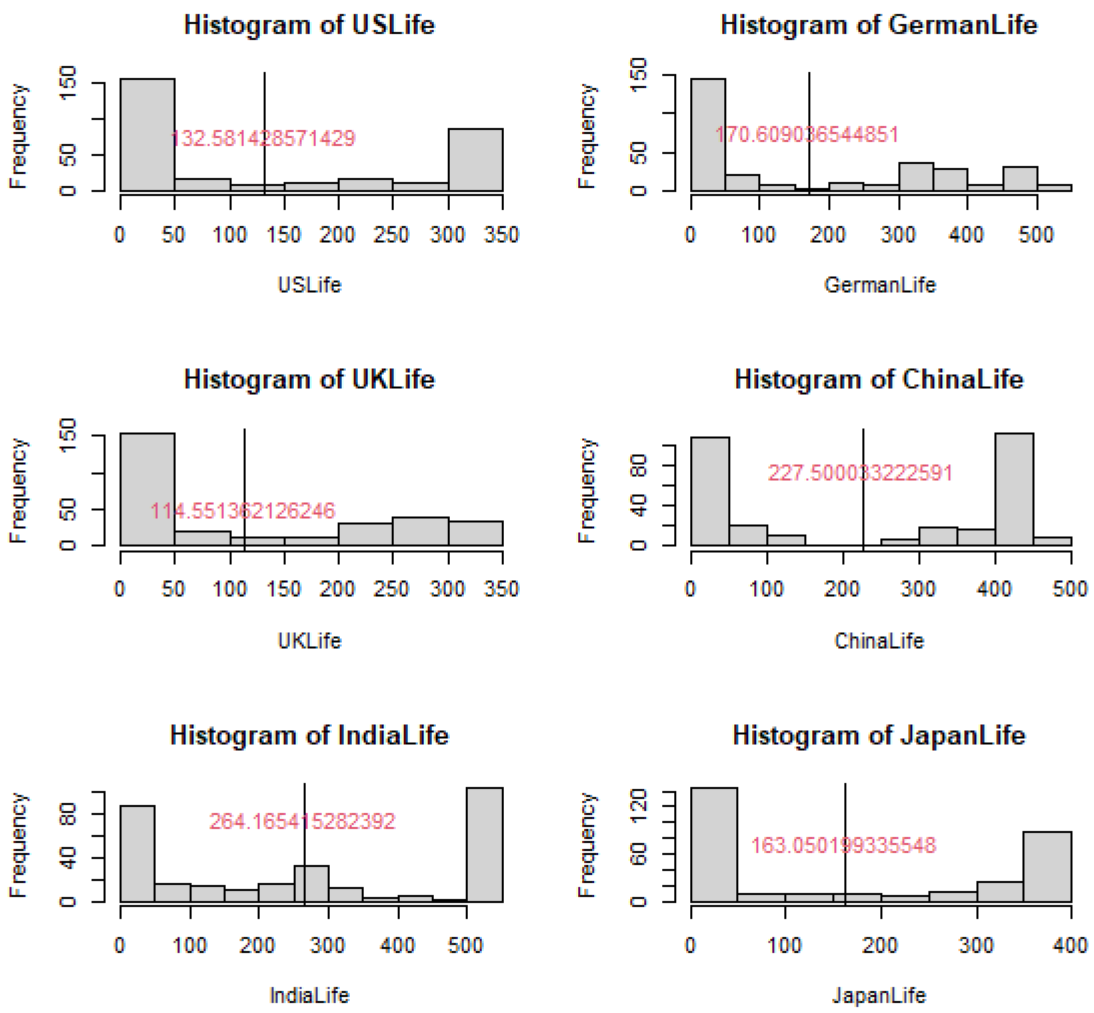

Fig.13: Histograms on Child mortality vs years (with mean)

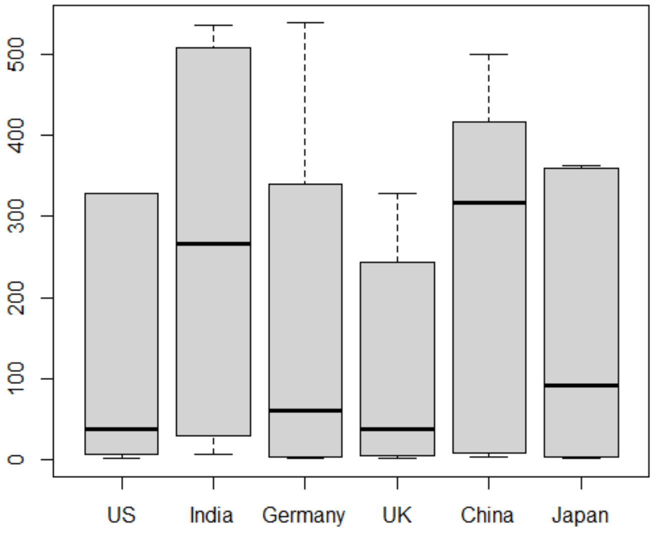

Fig.14: Boxplots of Child mortality in years (with median)

Code:

```
par(mfrow=c(1,1))
```

```
plot(c(1800:2100),USLife,col="red",main="Child
Mortality",pch=15,ylim=c(0,600),xlab="Years",ylab="Child Mortality", lwd=2.0, type =
"l")
lines(c(1800:2100),GermanLife,col="skyblue", lwd=2.0)
lines(c(1800:2100),UKLife,col="green", lwd=2.0)
lines(c(1800:2100),ChinaLife,col="orange", lwd=2.0)
lines(c(1800:2100),IndiaLife,col="black", lwd=2.0)
lines(c(1800:2100),JapanLife,col="yellow", lwd=2.0)

legend(x = "topright",                    # Position
       legend = c("United States", "Germany", "United Kingdom", "China", "India",
"Japan"),  # Legend texts
       fill = c("red", "skyblue","green", "orange", "black", "yellow"))              #
Colors

abline(v=1918)
abline(v=2022)
```

Output:

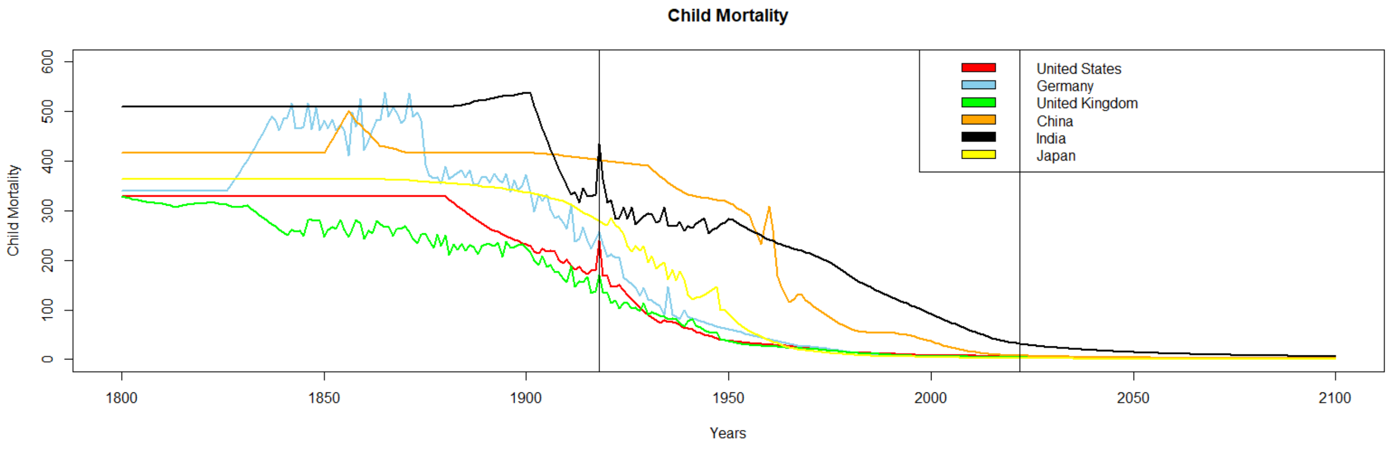

Fig.15: Trend in Child mortality from 1800 to 2100.

Trend:

a. Child mortality rate increased in India, which were as a result of the Spanish Flu pandemic in the 1918, and in the 1950s as India adjusted to its newfound independence.

```
wilcox.test(IndiaLife[119], IndiaLife[131], paired = T, alternative = "greater")
```

Null hypothesis: Child mortality in India in 1918 is lesser than child mortality in India in 1930. Here, by the Wilcoxon test, we get the **p-value as 0.5**(which is greater than 0.05), which means the **null hypothesis cannot be rejected** and spanish flu had an adverse effect on child mortality in India.

b. United States also has increase in the child mortality during the 1918 due to the spanish flu.
c. In China the sharpest decrease came between 1950 and 1955, as the Chinese Civil War ended, and the country began to recover from the Second World War.
d. The decline then stopped between 1955 and 1965, due to famines caused by Chairman Mao Zedong's attempted Great Leap Forward.The Taiping Rebellion 1850 to 1864 of China was a civil war in southern China caused the lives of millions of people also the child mortality.

Inferential Statistics:

- **Shapiro-Wilk Normality Test**

Null Hypothesis: Here the null hypothesis is that the distribution of life expectancies is normal.

Code:

```
# Shapiro test
shapiro.test(USLife) # null hypothesis- the distribution is normal (rejected for all the data below)
shapiro.test(GermanLife)
shapiro.test(UKLife)
shapiro.test(ChinaLife)
shapiro.test(IndiaLife)
shapiro.test(JapanLife)
```

Output:

Country	**p-value**	**W-value**
United States	< 2.2e-16	0.7443
Germany	< 2.2e-16	0.79689
United Kingdom	< 2.2e-16	0.79424
China	< 2.2e-16	0.75854
India	< 2.2e-16	0.82893
Japan	< 2.2e-16	0.73649

Since, for all countries' data, the p-value obtained is much lesser than 0.05, **the distributions are not normal and hence t-tests cannot be performed** on these distributions as a common assumption made during a t-test is the normality of distribution.

Here, we can perform Wilcoxon signed-rank test on the distribution as it is a non-parametric test.

- **Wilcoxon Signed-Rank Test**

A few Wilcoxon tests are performed on some countries' life expectancy distribution.

Code:

```
wilcox.test(USLife, GermanLife, paired = T, alternative = "greater")
wilcox.test(IndiaLife,JapanLife, paired = T, alternative = "less")
```

Output:

7. **Null Hypothesis:** Child mortality in the United States is lesser than that in Germany. Here, after the Wilcoxon test, the p-value obtained is **1** (more than 0.05), hence the **null hypothesis cannot be rejected** and the alternative hypothesis is falsei.e. Child mortality in the US is lesser than that in Germany.
8. **Null Hypothesis:** Child mortality in India is greater than that in Japan. Here, after the Wilcoxon test, the p-value obtained is **1**, hence the **null hypothesis cannot be rejected** and the alternative hypothesis is false, i.e. Child mortality in India is greater than that in Japan.

The compiled code used for the descriptive and inferential statistics is available here.