UNIVERSITY OF ENGINEERING AND TECHNOLOGY - VNU
FACULTY OF INFORMATION TECHNOLOGY


# MAT1101 - Probability & Statistics: Culminating Course Report

Submitted by:

**Hoang Bao An - 22024545**
**Nguyen Duc Huy - 22024528**
**Nguyen Anh Duc - 22024536**

Under the supervision of:
Dr. Hoang Thi Diep

Hanoi, 2023

# Abstract

[**?** ] Statistical techniques are employed in almost every phase of life. Surveys are designed to collect early returns on election day and forecast the outcome of an election. Consumers are sampled to provide information for predicting product preferences. Research physicians conduct experiments to determine the effect of various drugs and controlled environmental conditions on humans in order to infer the appropriate treatment for various illnesses. Engineers sample a product quality characteristic and various controllable process variables to identify key variables related to product quality. Newly manufactured electronic devices are sampled before shipping to decide whether to ship or hold individual lots. Economists observe various indices of economic health over a period of time and use the information to forecast the condition of the economy in the future.

Statistical methodologies assume a pivotal role in attaining the objectives within each of these practical scenarios. Within the context of this report, we have examined data pertaining to global GDP per capita and life expectancy, aiming to discern potential correlations between these two domains. All data have been sampled from gapminder.org. Also, we provide a Jupyter notebook file at https://github.com/hoangbaoan1901/stats-report/blob/main/analysis.ipynb and all the analysis can also be found at sheet 1, 2 and 3 of the additional Excel file.

# Table of Contents

# List of Figures

# Part 1

# Distributions, Estimations and Confidence Interval

In this section, we will visually represent the distribution of global GDP per capita and life expectancy. Subsequently, we will undertake the estimation of key statistical parameters, specifically the *mean* value and *standard deviation*. Furthermore, an analysis will be conducted to establish the 95% confidence interval for the mean value within each respective field.

## 1.1 Distributions & estimations

As stated in many textbooks, the sample mean $\overline{x}$ can be used to estimate the population mean $\mu$.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

And the unbiased standard deviation is approximated using this formula:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Excel offers the *AVG* and *STDEV* functions for the computation of the mean and standard deviation, thereby facilitating a straightforward approach to data analysis.

Additionally, we provided a Jupyter notebook file containing our analytical findings for visualization. Leveraging libraries such as pandas and NumPy, computations can be streamlined further. The employment of *mean* and *std* functions from NumPy serves to efficiently calculate these statistical values.

The following is the code in Python used for calculating these values.

```
gdp = df['GDP'].to_numpy()
lex = (pd.to_numeric(df['LEX'].str.replace(',', '.'), errors='coerce'))\
    .to_numpy()
size = gdp.size
gdp_mean = np.mean(gdp)
lex_mean = np.mean(lex)
gdp_sd = np.std(gdp)
lex_sd = np.std(lex)
print("Observations: ", size)
print("GDP(mean, standard deviation): ", gdp_mean, gdp_sd)
print("LEX(mean, standard deviation): ", lex_mean, lex_sd)
```

Result:

```
Observations:  195
GDP(mean, standard deviation):  20872.102564102563 24179.09720483299
LEX(mean, standard deviation):  72.39435897435897 7.319704279746025
```

Furthermore, we have visualized the data through the creation of a histogram, enhancing accessibility and comprehension of the dataset's distribution.

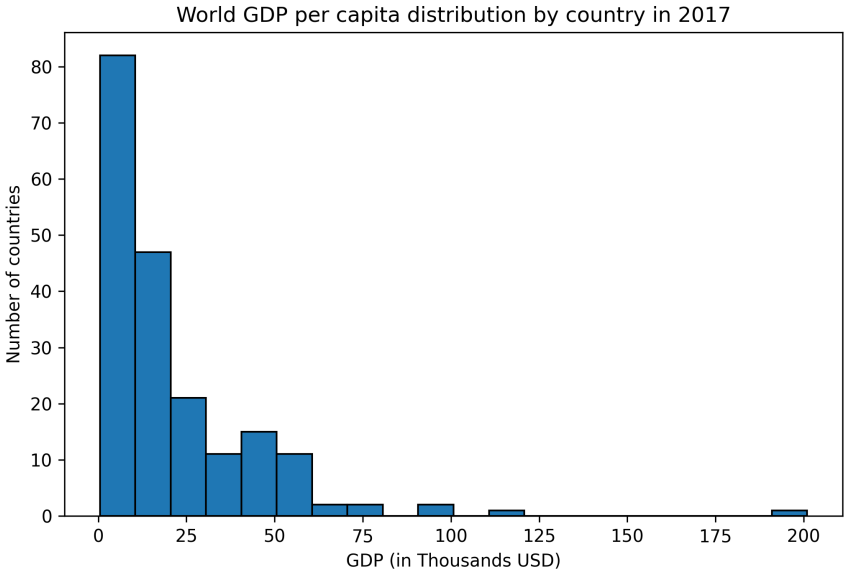


**Figure 1.1:** Distribution of GDP per capita in 2017

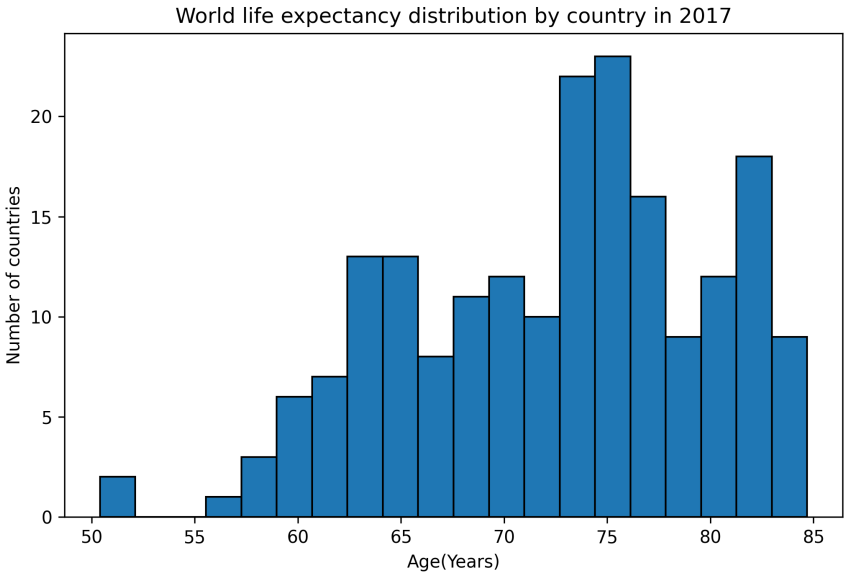


**Figure 1.2:** Distribution of Life expectancy in 2017

## 1.2 Popultion's mean value confidence interval

Given that our dataset comprises 195 observations, it is reasonable to presume a normal distribution with the preceeding estimations. The confidence interval can be calculated using this formula:

$$[\overline{x} - u_\beta \frac{s}{\sqrt{n}}, \overline{x} + u_\beta \frac{s}{\sqrt{n}}]$$

Nevertheless, when utilizing SciPy, it necessitates the specification of the "degree of freedom" parameter, which, in this instance, is determined as $195 - 1 = 194$. The subsequent code illustrates the implementation of the formula for calculating the confidence interval of the mean values pertaining to GDP and Life Expectancy

```
confidence_level = 0.95
dof = size - 1
alpha = (1 + confidence_level) / 2
critical_value = scipy.stats.t.ppf(alpha, dof)
print("c = ", critical_value)
gdp_moe = critical_value * gdp_sd / np.sqrt(size)
lex_moe = critical_value * lex_sd / np.sqrt(size)
gdp_confidence_interval = (gdp_mean - gdp_moe, gdp_mean + gdp_moe)
lex_confidence_interval = (lex_mean - lex_moe, lex_mean + lex_moe)
print("GDP 95% confidence interval: ", gdp_confidence_interval)
print("LEX 95% confidence interval: ", lex_confidence_interval)
```

Result:

```
c =  1.972267532579456
GDP 95% confidence interval:  (17457.119132515552, 24287.085995689573)
LEX 95% confidence interval:  (71.36054581633697, 73.42817213238096)
```

# Part 2

# Hypothesis Testing

In this part, we posited an inquiry: **Is it conceivable for people over the world to exceed a lifespan of 70 years?** To substantiate our hypothesis, we executed the ensuing examination, comprising these following procedural steps:

- States null hypothesis and alternative hypothesis
- Identify the level of test $\alpha$ (probability to make a *type I error*)
- Test statistic
- Rejection region
- Calculating the test statistic from observation sample
- Conclusion

## 2.1 Identifying null hypothesis and alternative hypothesis

Since we are verifying if the average world's life expectancy is more than 70 years, the null hypothesis shall be "the average life expectancy of people on the world IS 70":

$$H_0 : \mu = \mu_0 = 70$$

$$H_1 : \mu > \mu_0$$

## 2.2 Identify the level of test

In this test, we'll be using test level $\alpha = 5\%$. Hence, $c = z_\alpha = z_{0.05} = 1.64$

## 2.3 Test statistic

In this case, we'll be using the test:

$$T = \frac{(\overline{x} - \mu_0)\sqrt{n}}{s}$$

where n is the number of observations, $\overline{x}$ is the mean of Life Expectancy, and $s$ the standard deviation. These values has been calculated in the previous part:

```
Observations:  195
LEX(mean, standard deviation):  72.39435897435897 7.319704279746025
```

## 2.4 Rejection region

Since the alternative hypothesis is upper-tailed, the rejection region will be:

$$\Delta = \{T > c\}$$

## 2.5 Calculation

After some calculation:

```
c = 1.64
u0 = 70
T = (lex_mean - u0) * np.sqrt(size) / lex_sd
print("T = ", T)
print("T > c: ", T > c)
```

This is the result:

```
T =  4.5678626063375765
T > c:  True
```

## 2.6 Conclusion

As T resides within the rejection region, the null hypothesis is deemed invalid, leading to the rejection of the assertion that the average lifespan of individuals worldwide is less than or equal to 70 years.

# Part 3

# Correlation between GDP and expected lifespan

In this part, an exploration will be undertaken to discern whether there exists a correlation between a country's economic affluence and the life expectancy of its populace. Employing linear regression as a statistical tool to examine the potential relationship between these two variables is deemed appropriate for obtaining insights into this inquiry.

First we'll be listing the formula should be used for the calculation.

The correlation coefficient $\rho$ between 2 fields can be estimated using this formula:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

The slope and intercept can be calculated using this formula:

$$a = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b = \overline{y} - a\overline{x}$$

Utilizing NumPy, one has the capability to compute the correlation coefficient, as well as determine the slope and intercept of the regression line:

```
correl = np.corrcoef(gdp, lex)[0, 1]
print("r = ", correl)
slope, intercept = np.polyfit(gdp, lex, 1)
print("slope, intercept = ", slope, intercept)
```

Result:

```
r =  0.6402486572000062
slope, intercept =  0.00019382158053741545
68.34889506624556
```

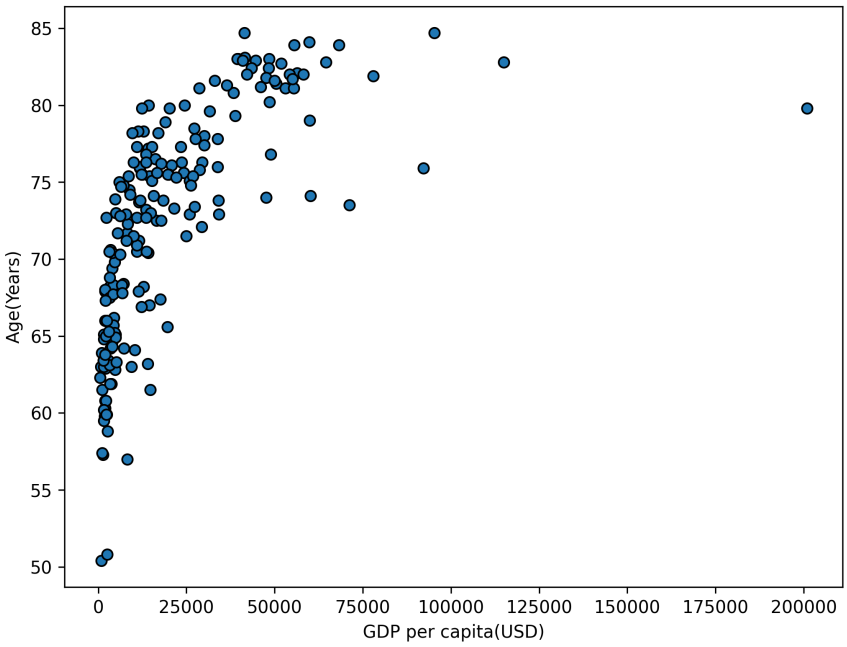


**Figure 3.1:** Correlation between GDP and Life Expectancy

Since $r = 0.6402486572000062$, we can conclude that there's a moderate linear correlation between GDP and life span expectancy. However, a glance at the scatter plot tells us that there suggests a decelerating growth logarithmic correlation between the two fields. So we decided to check if the correlation between the natural logarithmic value of GDP and the life expectancy is stronger than the current one.

```
correl2 = np.corrcoef(np.log(gdp), lex)[0, 1]
print(correl2)
```

And the result yields:

```
0.8380396098306463
```

It appears that a more robust linear correlation exists between the logarithmic values of GDP and life expectancy. Here's the visualization of the data.

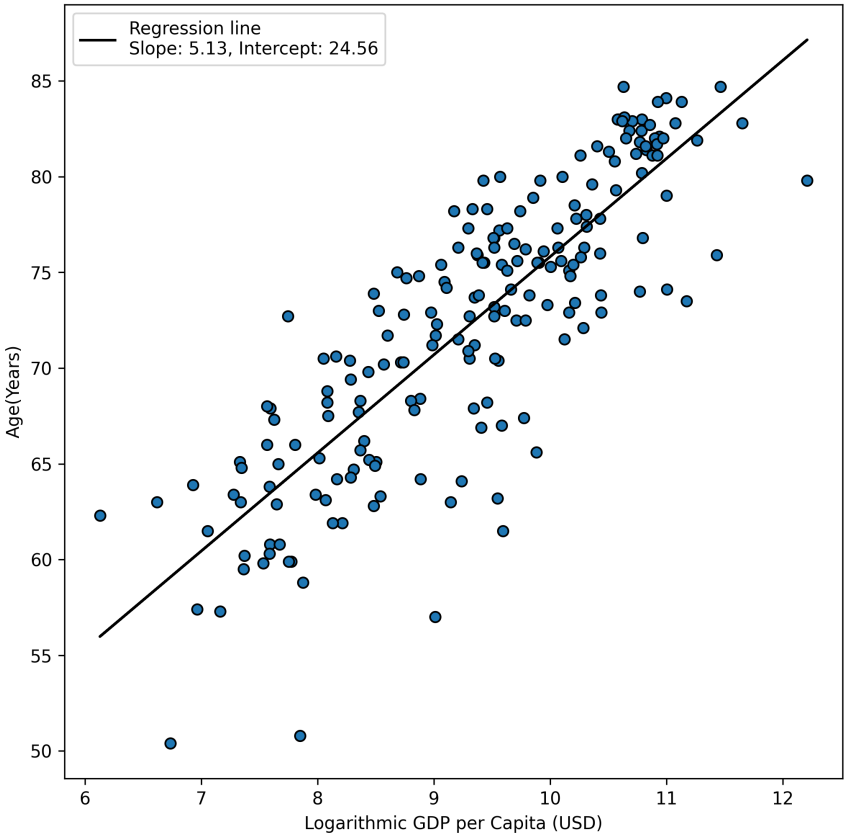


**Figure 3.2:** Correlation between ln(GDP) and Life Expectancy

With these calculation, we can reasonably predict that a country with GDP of 100000 USD is expected to have a life expectancy of $slope * ln(GDP) + intercept = 5.13 * ln(100000) + 24.56 = 83.62$ (years)

# Acknowledgements

We would like to express our sincere gratitude to our instructor Dr. Hoang Thi Diep, for imparting invaluable lessons and exhibiting an energetic attitude towards our learning. This project still harbors plenty opportunities for improvement, and we will remain committed to enhancing our problem-solving and mathematics skills in the future as a gesture of respect and appreciation for her guidance.