Hypothesis Testing

Jawadul Chowdhury

March 4, 2025

Contents

1 Introduction

In this paper, we explore hypothesis testing by manually creating functions in python for conducting a one sample t-test, two sample t-test and the pearson's correlation. We explore such methods of hypothesis testing by exploring a dataset that we will discuss in subsection 1.1.

1.1 Dataset Description

For this paper, we work with a dataset from 2009 by the World Health Organization. This dataset tracks a number of useful health-related metrics aggregated at the country level. Some features that we will be looking at from the dataset is as follows:

- Name of the country
- Life Expectancy in the Country
- Infant Mortality in the country
- Physician density
- Density of Hospital Beds
- Total Expenditure on Health as Percentage of GDP
- Out of Pocket Expenditure as Percentage of Private Expenditure on Health
- Per Capita Total Expenditure on Health
- Total Fertility Rate
- Gross National Income Per Capita
- Name of the Region

When we looked at the dataset, we wanted to get more information using `.info()` and `.describe()` on the Pandas data frame we created using the `.csv` file. Here is the information as follows:

- There are a total of 193 rows and 267 columns of data in the dataset
- The data types of the features are `float64`, `int64` & `object`
- The dataset in whole takes up a total memory of 402.7+ KB

2 Methods

In this section, we would like to explore the methods of hypothesis testing that we will be using throughout the paper, as well as which kind of graphs we will be using for each kind of hypothesis testing and why.

2.1 One Sample t-test

A onne sample t-test is meant to compare a numerical variable aginst a fixed number which is specified by us. The goal is to assess whether the numerical variable is different from the number we've specified.

To perform the one sample t-test, we need to calculate the test statistic as specified in equation 1.

$$t = \frac{\mu - M}{\frac{s}{\sqrt{n}}} \tag{1}$$

Test statistic for one sample

This is where the standard deviation and μ is the sample mean, n is the number of observations and M is a fixed number is specified by us.

Next, we need to calculate the standard deviation, which is specified in equation 2.

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

Sample standard deviation

This is where x_i is the value of the variable for the i^{th} observation. We take the sum of the differences between x_i and the the μ, and then multiply with 1 over n-1 and then take the square root to find the standard deviation. The other variables are similar to the ones explained in equation 1.

Next, we need to calculate the p-value, which is specified in equation 3.

$$p = 2(1 - P(|t|)) \tag{3}$$

P-value

We calculate the p-value using $P(|t|)$ where it is the cumulative distribution function (CDF) for the t-distribution.

Lastly, we would like to visualize this. Since we're comparing a numerical variable against a fixed number, it would be fitting to use a boxplot to help data spread, as this includes the mean, median, lower and upper quartile as well as any potential outliers. We apply this plotting to the life expectancy of Europe as will be seen in the results section in Figure 2.

2.2 Two Sample t-test

A two-sample t-test is meant to compare a numerical variable against a categorical variable, as the goal is to assess whether the numerical variable is different across the categories.

To perform the two sample t-test, we need to calculate the test statistic as specified in equation 4.

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4}$$

Test Statistic for two samples

Here, μ_1, s_1 and n_1 are the sample mean, sample standard deviation and number of observations from the first data set. Next, μ_2, s_2 and n_2 are the sample mean, sample standard deviation, and number of observations from the second second dataset.

The standard deviation is computed using equation 2 and the p-value is computed using equation 3, with a difference being the degrees of freedom being used.

To calculate the degrees of freedom, we need to use equation 5 as specified below, where ν is the degrees of freedom.

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \tag{5}$$

Degree of Freedom

Lastly, we would like to visualize this. Since we're comparing a numerical variable against a categorical variable, it would make the most sense to use a violin plot. In the results section, we will create a violin plot of the life expetancy in Europe vs in Asia, as shown in Figure 3 in the Results section.

2.3 Pearson's Correlation

The Pearson's Correlation is meant to compare a numerical variale against another numerical variable. We use this to assess whether the two variables "move" together in a significantly related way.

To calculate the pearson's correlation, we need to use equation 6 as specified below.

$$R = \frac{\sum_{i=1}^{n}(x_{i,1} - \mu_1)(x_{i,2} - \mu_2)}{\sqrt{\sum_{i=1}^{n}(x_{i,1} - \mu_1)^2}\sqrt{\sum_{i=1}^{n}(x_{i,2} - \mu_2)^2}} \tag{6}$$

Pearson's Coefficient

In equation 6 $x_{i,1}$ and $x_{i,2}$ are the i^{th} observations associated with variable 1 and 2, μ_1 and μ_2 are the means of each variable, and n is the number of observations.

When we do hypothesis testing, we use equation 7 to make the test statistic as specified below.

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \tag{7}$$

Test Statistic Pearson's Coefficient

Once we compute the test statistic, we can then compute the p-value with the degrees of freedom set to $n-2$, as specified in equation 3.

Lastly, we would like to visualize this. Since we're comparing a numerical variable against another numerical variable, it would make the most sense to plot a scatter plot. In the results section, we will create a scatter plot of the life expectancy vs the infant mortality across the entire dataset, as shown in 4 of the Results section.

3 Results

In this section, we explore and analyze the graphs that have been created while performing the different types of hypothesis testing.

3.1 Histogram of Gross National Income

As shown in Figure 1, we plotted a histogram of the gross national income across all countries in the world that are tracked by the World Health Organization. We can observe that the histogram is unimodal and has a right skew. It can be observed that the mode, median and mean is somewhere between 0 to 10000.

It is also worth noting that there is a peak between 30000 and 40000, however, it might the peak is not high enough to consider that the histogram is bimodal.

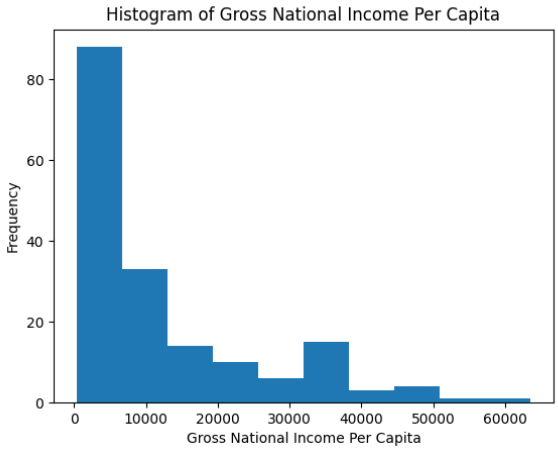

Figure 1: Histogram of Gross National Income

3.2 Boxplot of Life Expectancy in the European Region

As shown in Figure 2, we plotted a box plot of the life expectancy in the European region. The reason we chose a box plot is because we're comparing a numerical variable against a fixed number, which is 70 and 76. From this box plot, we can observe the following summary statistic:

- The Lowest Age: 66
- The Highest Age: 82
- The Lower Quartile: somwhere between 72 and 74
- The Upper Quartile: somewhere between 78 and 80
- The Median Age: almost around 78

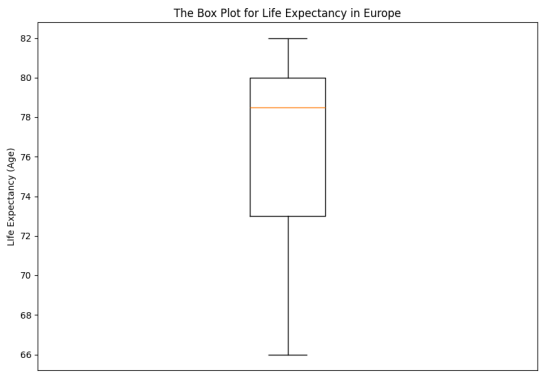

Figure 2: Box Plot of Life Expectancy in Europe

3.3 Violin Plot of Life Expectancy in Europe & Asia

As shown in Figure 3, we plotted a violin plot of the life expectancy in both Europe & Asia. The reason we chose a violin plot is because we're comparing a numerical variable against a categorical variable. From the violin plots, we can observe a number of things.

- It can be seen that around the median, the width of the violin plot is wider, which indicates that there are more data points concentrated in that range.
- A striking is difference is that the lowest life expectancy in Europe is somehwere near 60, however, in Asia, it can be as low as between 30 to 40.
- The violin plot for Asia has a negative skew, while the violin plot for Europe has a symmetric distribution
- It can be seen that Europe has a higher median than Asia, which suggests that the group has a higher central value

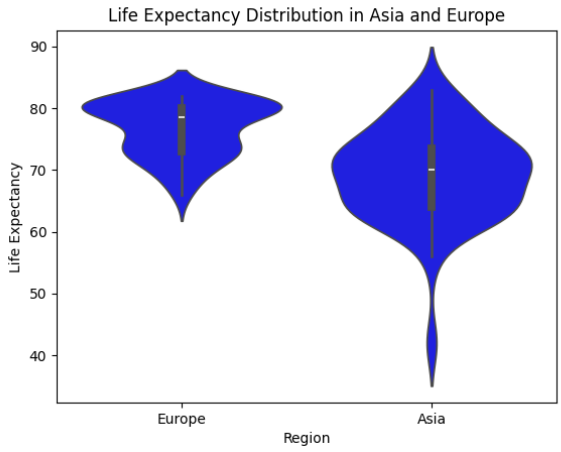

Figure 3: Violin of Life Expectancy in Europe & Asia

3.4 Scatterplot of Life Expectancy vs Infant Mortality

As shown in Figure 4, we plotted a scatter plot of the life expectance vs the infant mortality rate around the world. The reason we chose a scatter plot is because we're comparing a numerical variable against another numerical variable.

A look at the scatterplot will tell us a number of things:

- There is a negative correlation, where as the life expectancy increases, the infant mortality rate decreases
- If we were to draw a best fit line, we can see that there is a change in trend from the points being weakly clustered along the line to being strongly clustered as the life expectancy increases
- We can deduce that there is a linear relationship between the life expectancy & infant mortality rate

Overall, this suggests that countries that have a high life expectancy means they have low infant mortality rate, and countries that have a high infant mortality rate have a low life expectancy.

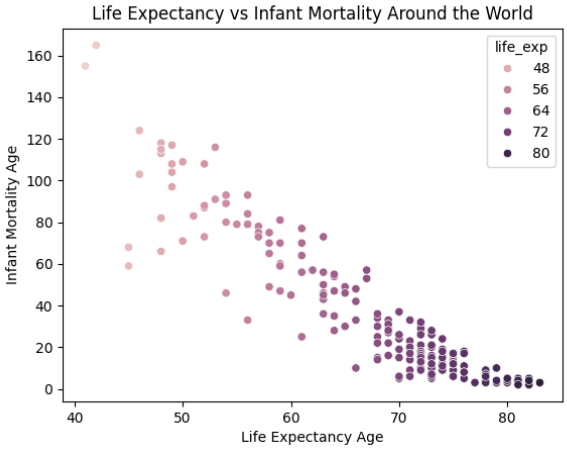

Figure 4: Scatterplot of Life Expectancy vs Infant Mortality

Moreover, we can draw a best fit line over the scatter plot to demonstrate the linear relationship as shown below in figure 5.

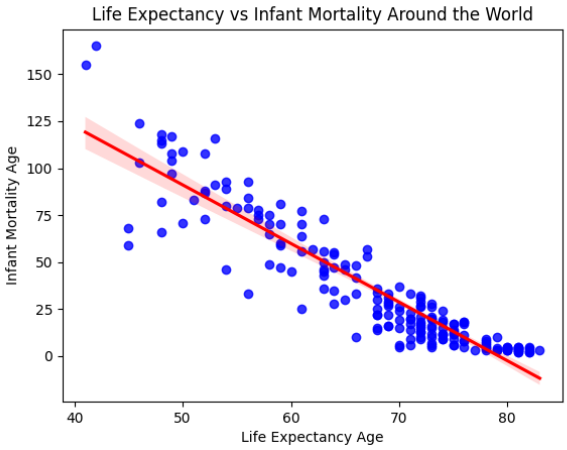

Figure 5: Scatterplot of Life Expectancy vs Infant Mortality with Best Fit Line

4 Discussion

In this last section, we would like to discuss some of the work that has been done in the paper, as well as what could've been done.

4.1 The Gross National Income

In figure 1, we created a histogram of the Gross National Income. As the shape of the distribution is right skewed and unimodal, it is worth noting that there was a presence of missing values. We can deal with such missing values by performing imputations. The imputations can be done using a predictive model such as a KNN, or they can be filled by calculating the mean, median and mode. A good choice would be to use a KNN model simply because it considers the relationships between variables.

4.2 The One Sample t-test

For our first hypothesis test, we conducted a one sample t-test. With regards to choosing the M value for our experiments, we chose 70 and 76 years because although we didn't know what the life expectancy was, we would like to see whether 70 or 76 would be a good estimate for determining what the mean age of the life expectancy is and how close we were to being right.

With regards to the value of M, it would make sense to set up a null and alternative hypothesis for the one sample t-test, which is outlined in the following bullet points:

- H_0: the mean of the life expectancy in Europe is the same as the value of M
- H_1: the mean of the life expectancy in Europe isn't the same as the value of M

With the null and alternative hypothesis being set, we move on to the p-value. As we look at equation 3, it can be seen that we find the probability $P(|t|)$ that the test statistic t falls under the observed range under the null hypothesis. Due to the presence of the absolute value symbol, we can tell that this is for a two tailed test, as we account for both tails of the probability distribution.

Next, we perform the calculations where M = 70 & M = 76. We get the following test statistic and p-values as such:

- The test statistic value for Life Expectancy of 70: 9.938016169764737
- The p value for Life Expectancy of 70: 1.760564922569916e-12
- The test statistic value for Life Expectancy of 76: 1.1814424817202844
- The p value for Life Expectancy of 76: 0.24423470312189977

Given the following test statistic and p-values, we set the signifigance level at 0.05. With this in mind, there are a few key takeaways:

- For age 70, the p value is 1.76e-12, which is smaller than 0.05, so we reject the null hypothesis
- For age 76, the p value is 0.24, which is larger than 0.05, so we fail to reject the null hypothesis

Lastly, we wanted to cross-check if our custom function produced the same results as the one from `scipy.stats`. As per the Jupyter Notebook, the function matched the results from `scipy.stats`.

4.3 The Two Sample t-test

For our second hypothesis test, we conducted a two sample t-test. For this test, we set up our null and alternative hypothesis as follows:

- H_0: the mean of the life expectancy is equal in both Europe & Asia
- H_1: the mean of the life expectancy is not equal in both Europe & Asia

We performed the calculations, and we get the following test statistic and p-value as such:

- The test statistic value of Life Expectancy in Europe vs Asia: -5.884174172667128
- The p value of Life Expectancy of Europe vs Asia: 1.0064936417641945e-07

For this experiment, we set our signifigance value to 0.05, which means that since the p-value is less than the signifigance level, this means we can reject the null hypothesis. As a result, this means we reject the null hypothesis. As a result, this means that the life expectancy in Europe is signifiantly different than in Asia.

Lastly, we wanted to cross-check if our custom function produced the same results as the one from `scipy.stats`. As per the Jupyter Notebook, the function matched the results from `scipy.stats`.

4.4 Pearson's Correlation

For our third hypothesis test, we conducted person's correlation. For this test, we set up our null and alternative hypothesis as follows:

- H_0: there is no correlation between the life expectancy and infant mortality rate
- H_1: there is a strong correlation between the life expectancy and infant mortality rate

We performed the calculations, and we get the following test statistics and p-value as such:

- The Pearson Coefficient: -0.9255371904433326
- The Test Statistic: -33.7804082017176
- The P Value: 0.0

For this experiment, we set our signifigance value to 0.05, which means that since the p-value is less than The signifigance level, we reject the null hypothesis. This relationship is a linear relationship because we are able to draw a best fit line as shown in Figure 5.

Lastly, we would like to compare the results derived from the function we made and the `lingress` method from `scipy.stats`. Adding on to the caluclations we performed, here are the results we obtained from using the `lingress` function:

- The Pearson Coefficient: -0.9255371904433327
- The P Value: 1.733832757295465e-82

When we compare the values, we can see that the pearson coefficients are equal. Although the `linegress` function doesn't return a test statistic, we can see that the p-values are different. However, we can conlude that both p-values from the function and the `lingress` is similar because 1.733832757295465e-82 is basically estimated as 0.

4.5 Further Reading