# 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 $\mu$ 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 $i^{\text{th}}$ observation. We take the sum of the differences between $x_i$ and the the $\mu$, 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.

## 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, $\mu_1$, $s_1$ and $n_1$ are the sample mean, sample standard deviation and number of observations from the first data set. Next, $\mu_2$, $s_2$ 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 $\nu$ 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.

## 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 $i^{\text{th}}$ observations associated with variable 1 and 2, $\mu_1$ and $\mu_2$ 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.