# Lab 5: Hypothesis Testing (from scratch!)

In this lab we are going to work on hypothesis testing. In particular, we will focus on honing our expertise by implementing several hypothesis tests from scratch and then comparing the results we get to the outputs of built in python functions. To do this, we will work with World Health Organization data that tracks a number of useful health-related metrics aggregated at the country level.

## 1 Read in and prepare the data

1. First read the who2009.csv file into your python development environment using pandas. You'll need to import pandas and then apply the '.read_csv' method.

2. Assess the size of the data set using the .shape method. You should notice that there are a large number of columns in the data. We will only need a very small number of them. I have provided a codebook to go with this data (WHO2009SubsetCodebook.pdf). Drop columns from the data, keeping only those identified in the codebook.

3. Rename the columns in the reduced data set to names that are appropriate descriptors of the information contained in each variable according to the codebook. You may find the .rename method to be useful. Use the following variable names: country, life_exp, infant_mortality, phys_density, hospital_bed, health_exp_percent_GDP, OOP_percent_exp, health_exp_PC, fertility_rate, GNI_PC, regionname.

4. Answer the following questions in your writeup.

   (a) Plot the distribution of the GNI_PC variable. Given the shape of this distribution what methods would you apply to deal with those missing values?

## 2 Conduct a one-sample $t$-test from scratch

1. A one sample $t$-test is meant to compare a numerical variable against a fixed number (typically this is the null hypothesis) specified by you–the data scientist; the goal is to assess whether the numerical variable is "different" from the number you've specified.

2. Lets assess the life expectancy in Europe. Begin by creating a second data set that includes only rows in which regionname is "Europe."

3. The test statistic in a one sample $t$-test has the following form:

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

where:

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

is the sample standard deviation and where $\mu$ is the sample mean, $n$ is the number of observations in the sample, $x_i$ is the value of the variable for the $i$th observation, and $M$ is a number chosen by you which functions as the thing to compare the sample mean against. As in class, the $p$-value is given by:

$$p = 2(1 - P(|t|))$$

where $P(\cdot)$ is the cumulative distribution function (CDF) for the $t$ distribution.

4. Using the equations in 2.3 write a function that takes the data and the value $M$ as inputs and returns the test statistic and the $p$-value as outputs. To do this, you will need to work with the CDF of the $t$-distribution. I suggest you use from scipy.stats import t to load the $t$ distribution and then work with t.cdf with the degrees of freedom set to $n-1$. You may also find the .mean and .std methods from numpy to be helpful, but make sure you use ddof = 1 if you use .std.

5. Use the function you built in 2.4 to assess whether the life expectancy in Europe is significantly different from: 1) 70 years and 2) 76 years.

6. Finally, validate your code by comparing the results to the built in one-sample $t$-test. To do this, use the ttest_1samp method from scipy.stats.

7. Answer the following in your writeup:

   (a) How do you choose the fixed number $M$ in this test?

   (b) What are the null and alternative hypotheses for the one-sample $t$-test conducted here?

   (c) Explain the equation for the $p$-value. What is it doing?

   (d) Is life expectancy in Europe significantly different from 70 years or 76 years? How do you know?

   (e) Did your function produce the same results as the built in version?

# 3 Conduct a two-sample $t$-test from scratch

1. A two sample $t$-test is meant to compare a numerical variable against a categorical variable; the goal is to assess whether the numerical variable is "different" across the categories.

2. Lets compare the life expectancy in Europe against the life expectancy in Asia. You've already created a Europe-only data set. Create a third data set that includes only rows in which regionname is "Asia."

3. The test statistic in a two sample $t$-test has the following form:

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

where $\mu_1$, $s_1$, and $n_1$ are the sample mean, sample standard deviation, and number of observations from the first data set, $\mu_2$, $s_2$, and $n_2$ are the sample mean, sample standard deviation, and number of observations from the second data set, and the standard deviations are computed as in 2.3. The $p$-values are computed as in 2.3 with one important difference – you will need to compute the degrees of freedom for the $t$-distribution rather than just setting it to $n-1$. Use the following equation to compute the degrees of freedom and use this in the calculation of the $p$-value:

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

4. Using the equations in 3.3 write a function that takes two data sets inputs and returns the test statistic and the $p$-value as outputs.

5. Use the function you built in 3.4 to assess whether the life expectancy in Europe is significantly different from the list expectancy in Asia.

6. Finally, validate your code by comparing the results to the built in two-sample $t$-test. To do this, use the ttest_ind method from scipy.stats.

7. Answer the following in your writeup:

   (a) What are the null and alternative hypotheses for the two-sample $t$-test conducted here?

   (b) Is life expectancy in Europe significantly different from life expectancy in Asia? How do you know?

   (c) Did your function produce the same results as the built in version?

# 4 Pearson's Correlation from scratch

1. Pearson's correlation is meant to compare a numerical variable against another numerical variable; the goal is to assess whether the two variables "move" together in a significantly related way.

2. Pearson's correlation has the following form:

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

Here, $x_{i,1}$ and $x_{i,2}$ are the $i$th observations associated with variable 1 and variable 2, $\mu_1$ and $\mu_2$ are the means of each variable, and $n$ is the number of observations. Usually, when we do hypothesis testing we use this correlation to make the test statistic:

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

Once we have this test statistic, we can compute $p$-values as in 2.3 with the degrees of freedom set to $n-2$.

3. Using the equations in 4.2 write a function that takes two variables from a single data set as inputs and returns the test statistic and the $p$-value as outputs.

4. Use the function you built in 4.3 to assess whether life expectancy is correlated with infant mortality across the entire data set.

5. Finally, validate your code by comparing the results to the built in correlation test. To do this, use the linregress method from scipy.stats.

6. Answer the following in your writeup:

   (a) What are the null and alternative hypotheses for the correlation test conducted here?

   (b) Compare and contrast the results of your function vs linregress – do you see the quantities you developed in your function in linregress?

   (c) Would you say that there is a relationship between life expectancy and infant mortality? What is this relationship and how do you know?

# 5 Exploring Additional Statistical Tests

1. There are LOTS of hypothesis tests that have been developed for different data structures. In this final section we will look at two additional tests–the Kruskal-Wallis test and the $\chi^2$ test for independence. Read the associated entry for each in the Handbook for Biological Statistics.

2. Answer the following in your writeup:

   (a) List the two types of variables for which each test is appropriate. Indicate any assumptions that you would need to be aware of.

   (b) Write down the general forms of the null and alternative hypotheses in each test.

   (c) n your own words, write what it would mean if the test did and did not indicate statistical significance.

# 6 Submission Instructions

- Write up the answers to the questions in a short word document; aim for around 2 pages of text and include all graphics generated. Add footnotes identifying which sentence addresses which questions. Write in complete sentences organized into paragraphs – your goal is to explain what you've done and what you've learned to your audience (me!). Accordingly, you may seek to emulate some of the sections of the whale paper describing their data. Include the appropriate plots you've generated as mentioned above. Convert this to pdf and submit it. Submit your .ipynb file as well.
- The grading rubric for this assignment will be available in Canvas.
- NO OTHER SUBMISSION TYPES WILL BE ACCEPTED.
- **Late policy**: 5% of total points deducted per day for three days – after that no submissions allowed.