ENGR 301
Engineering Applications of Data Science

Laboratory 2
Topic: Biomedical Engineering; Subtopic: Diagnosis

Introduction: Inferential statistics and hypothesis testing is still used extensively in biomedical engineering and its subfields, as it helps explain observed phenomena instead of simply summarizing it as with descriptive statistics. In particular, in diagnostics, scientists can learn what signs accompany or indicate a given disease. In this dataset, anonymized medical data for a number of people with and without diabetes are provided; you will analyze if differences in these features between the two groups is statistically significant (i.e., if it could be used for diagnosis of type 2 diabetes). In order to reduce the need for standardization of background, all participants here identify as female, are at least 21 years old, and identify as being of Pima Indian heritage.

Discipline Specific Information: The data set contains the following information:

Column Name	**Description**
Pregnancies	Number of times a participant has been pregnant.
Glucose	The wind speed at the hub height of the turbine (units of meters per second).
BloodPressure	Blood pressure of a participant in units of mm Hg.
SkinThickness	The thickness of a participants skin measure at the triceps, in units of mm.
Insulin	The theoretical power the wind turbine *should* generate under the given conditions at each point in time (units of kilowatt hours).
BMI	Body mass index of a participant (weight/height^2) in units of kg/m^2.
DiabetesPedigree	A value indicating diabetes history in your relatives based on genetics.
Age	Age of a participant in years.
Outcome	1 indicates the participant has diabetes, 0 indicates they do not have diabetes.

If you need help with or want to know more about the different scipy.stats programs, check here:

normal_test: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html

ttest_1samp: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html

ttest_ind: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

Tasks

Note: For full credit, all plots must have axis labels and units on the x- and y-axes, and the discussion must be in complete and full sentences.

Part 1 – Exploratory Data Analysis and Data Cleaning

1. Import the appropriate packages (numpy, pandas, and matplotlib) into a Jupyter notebook, in addition to the SciPy Statistics package (scipy.stats). Also import the seaborn package. The data is provided in a CSV file called diabetes.csv. Use *pandas* to import this data set into a Jupyter notebook as a data frame.

2. How many measurements and features does this data set contain?

3. What is the mean age of the sample? What is the distribution of the outcomes (*i.e.*, how many people in the sample have diabetes and how many do not)? Hint: try the *pandas* command value_counts() on a column; what does that do?

4. Create a pairplot using seaborn (if you import seaborn as sns, you can use sns.pairplot(df) to do this). Is there something strange about the histograms for glucose, blood pressure, skin thickness, insulin, and BMI?

One critical skill a data scientist often has to deal with is incomplete information. In this data set, if certain information about a patient isn't known, it was replaced with a 0, which is of course unphysical for things such as glucose level or blood pressure. That will be the goal of this segment.

4a. Execute the first line of code in the provided Jupyter notebook. Create a new pairplot and compare the histograms with those from the previous part. What did this command do? Note that if you didn't call your data frame "df" this code will not work.

4b. Execute the second line of code in the provided Jupyter notebook. What does this command tell you about the data?

4c. Execute the third line of code in the provided Jupyter notebook. What does this command do? Do you think that performing this procedure is valid? Why or why not?

Now that we have "cleaned" our data, we can proceed with the exploratory analysis.

5. Which two features have the strongest correlation? Which feature has the strongest correlation with the outcome?

6. From this data frame, find the mean blood pressure of people with and without diabetes. How can you use *pandas* to sort data in this way? How different are these two means?

7. Use the groupby function in *pandas* to group the data frame by Outcome and find the mean of each column. What do you notice about each feature's mean when comparing them between the two outcomes?

Part 2 – Hypothesis Generation and Testing

1. First, examine the difference in blood pressure between the people in this study with and without diabetes.

Using a QQ plot (*i.e.*, a probability plot), does it appear that the blood pressure data is normally distributed? Why or why not?

ENGR 301
Engineering Applications of Data Science

Perform a normal test for the blood pressure data using the stats package (stats.normaltest(...)) with a significance level of 1%. What is the null and alternative hypothesis? What is the p-value? Given this, is the data normally distributed?

2. Create a 95% and 99% confidence interval for blood pressure using the percentiles command from numpy. What is different between the two? Does the mean fall in this range?

3. Assume that the mean blood pressure of the entire population is known to be 71 mm Hg. Perform two one sample t-tests (stats.ttest_1samp from the scipy.stats package) with a significance level of 1% (see below). For each test, state the null and alternative hypothesis, and compute the p-value. Based on this, reject or accept the null hypothesis and answer the question.

3a. Test 1: Is the blood pressure of people without diabetes in this sample significantly different than the blood pressure of the population?

3b. Test 2: Is the blood pressure of people with diabetes in this sample significantly different than the blood pressure of the population?

4. Perform a two sample t-test (stats.ttest_ind from the stats.scipy package) comparing the mean blood pressure of people with and without diabetes in this sample. Are they significantly different?

5. Choose one additional feature to compare between the people with and without diabetes and design a two sample t-test. What did you discover?