Rabbani Nur Kumoro
21/472599/PA/20310
CSB - Introduction to Statistics

Assignment 1

I. Introduction

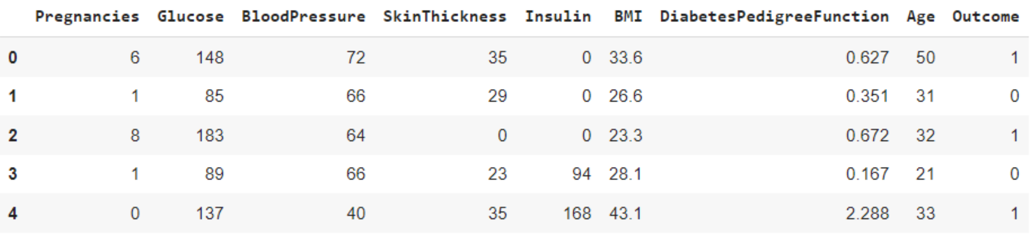
In this report, we will discuss Kaggle's Pima Indians Diabetes Dataset from the National Institute of Diabetes and Digestive and Kidney Diseases:

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

The Dataset will consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

II. Data Visualizations and Distributions

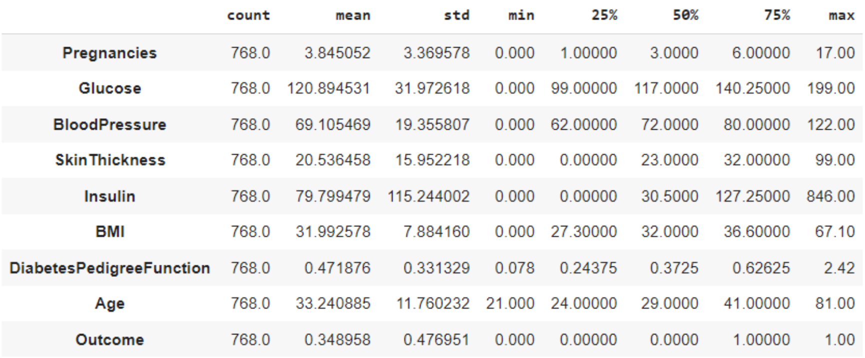
In this section, we will use tools such as Google Colab and Python Programming Language to Visualize and Distribute the Datasets.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. II.1 Data Visualization

In the first **Fig. II.1**, we describe the Data using Pandas DataFrame. By using the 'head' function to get the first n rows. This function returns the first n rows for the object based on position. So we can further understand the insights of the data and also verify if the object has the right type of data in it.

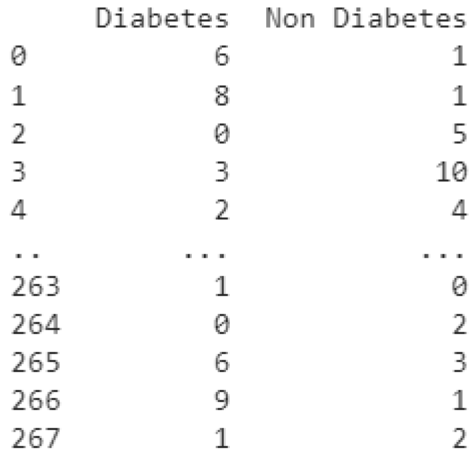


	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Fig. II.2 Descriptive Statistics of the Data

In the second **Fig. II.2**, we gave brief descriptive statistics of the data that summarize the central tendency, dispersion, and shape of a dataset's distribution. By using the ‘describe’ function to calculate some statistical data like percentile, mean, and standard deviation of the numerical values of the data frame. It analyzes both numeric and objects series, also the data frame column sets of mixed data types.

Now on, we will choose two particular data from the column and present the data distributions of the datasets. I decided to focus on two variables that are **Diabetic and Non-Diabetic Patients.**



```
     Diabetes  Non Diabetes
0           6             1
1           8             1
2           0             5
3           3            10
4           2             4
..        ...           ...
263         1             0
264         0             2
265         6             3
266         9             1
267         1             2
```

Fig. II.3 Visualization of Two Particular Data

As we separate the two particular data from the dataset, we merge the Diabetic and Non-Diabetic Patients' Data into one particular column output. We will try to compare and analyze the Pregnancy Months of those patients. The numbers that are displayed on the output are the expected Pregnancy Months.

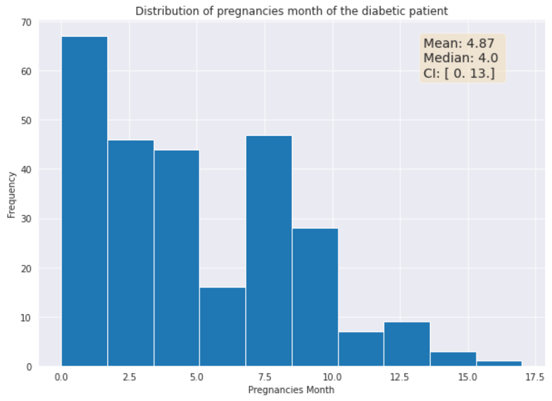


Fig. II.4 Histogram Distribution of Pregnancy Month for Diabetic Patients

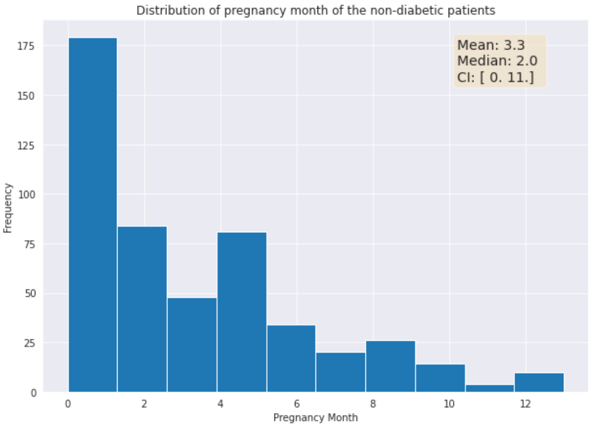


Fig. II.5 Histogram Distribution of Pregnancy month for Non-Diabetic Patients

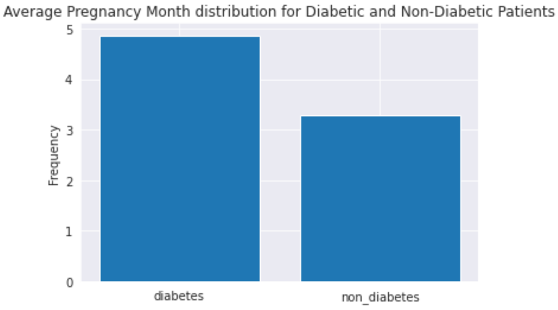


Fig. II.6 Histogram Distribution of the Average Pregnancy Month for both Patients

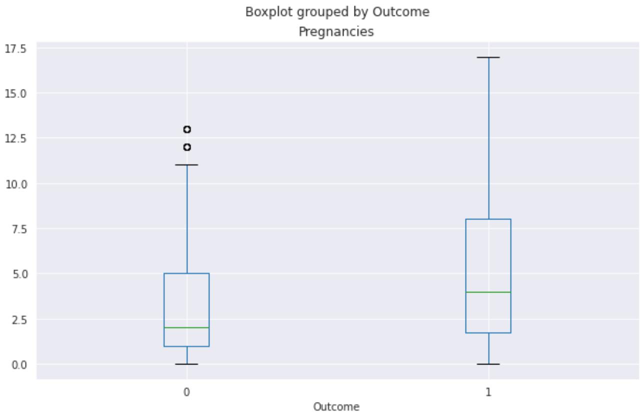


Fig. II.7 BoxPlot Distribution of the Average Pregnancy Month for both Patients

The BoxPlot represents a distribution of the average pregnancy month for both patients. The average month with the “1” type which represents Diabetic Patients has a longer average Month of Pregnancy than Non-Diabetic patients “0”. The outliers are represented with dots. The height of the boxes is proportional to how much the values are spread out. Thus, taller boxes indicate more variance.

III. Statements and Hypothesis Tests

In this section, we will discuss the problem statements and prove them through the hypothesis tests in the final statement:

1. Statement #1

What is the Average Month of Pregnancy for Patients with Diabetes?

Our Parameter of Interest is to calculate a population Mean Pregnancy Month of Patients with Diabetes and our task is to construct a 95% Confidence Interval for a Population Mean Pregnancy Months for all Patients with Diabetes.

The Formula to determine the Interval Estimation is listed as below:

$$x - z \times \frac{s}{\sqrt{n}} < \mu < x + z \times \frac{s}{\sqrt{n}} \ldots(1)$$

Where the x is the best point estimation for the Mean Pregnancy Months and $z \times \frac{s}{\sqrt{n}}$ is the Margin of Error for Mean Pregnancy Months.

The Best Point Estimate for the Mean Pregnancy Months of Patients with Diabetes:
= 4.865671641791045

Estimated Standard Error for the Mean Pregnancy Months of Patients with Diabetes:
= 0.2285325476073356

The Margin of Error for the Mean Pregnancy Months of Patients with Diabetes:
= 0.44838085840559244

Therefore, by using manual calculation the 95% Confidence Interval for the Mean Pregnancy Months of Patients with Diabetes:
(4.865671641791045 - 0.44838085840559244, 4.865671641791045 + 0.44838085840559244)
= (4.417290783385453, 5.314052500196637)

By using statsmodels Library the 95% Confidence Interval:
= (4.417756079185482, 5.313587204396608)

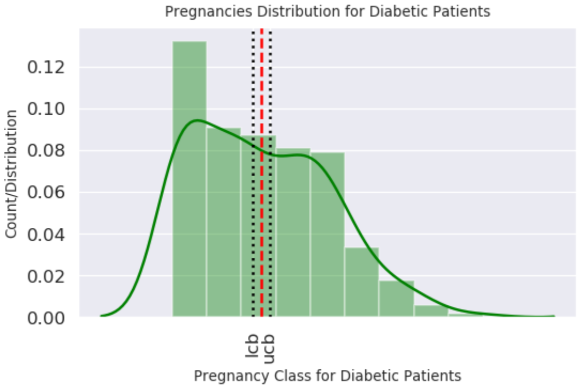


Fig. III.1 Pregnancies Distribution for Diabetic Patients

From this Confidence Interval, we can infer that with 95% confidence, the Population Mean Pregnancy for all Patients with Diabetes is estimated to be between 4.417 Months and 5.314 Months.

2. **Statement #2**

What is the Average Month of Pregnancy for Non-Diabetic Patients?

Our Parameter of Interest is to calculate a population Mean Pregnancy Month for Non-Diabetic Patients and our task is to construct a 95% Confidence Interval for a Population Mean Pregnancy Months for all Non-Diabetic Patients.

The Best Point Estimate for the Mean Pregnancy Months Non-Diabetic Patients:

= 3.298

Estimated Standard Error for Mean Pregnancy Months of Non-Diabetic Patients:

= 0.13493259654813752

The Margin of Error for Mean Pregnancy Months of Non-Diabetic Patients:

= 0.2647377544274458

From the first Formula Equation …(1) above in the First Statement and the value that we have obtained. By using manual calculation, the 95% Confidence Interval for Mean Pregnancy Months of Non-Diabetic Patients:

(3.298 - 0.2647377544274458, 3.298+ 0.2647377544274458)

= (3.0332622455725544, 3.5627377544274457)

By using statsmodels Library the 95% Confidence Interval:

= (3.033536970425177, 3.562463029574823)

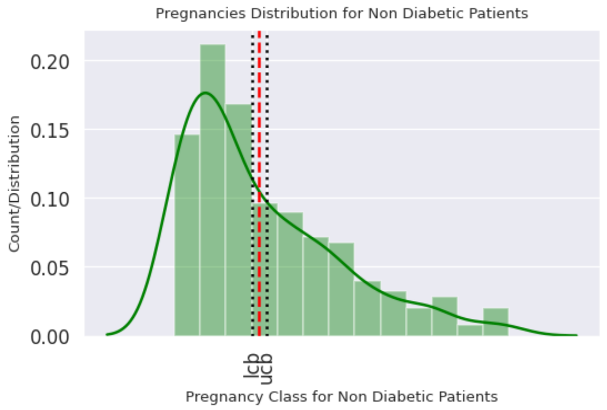


Fig. III.2 Pregnancies Distribution for Non-Diabetic Patients

From this Confidence Interval, we can infer that with 95% confidence, the Population Mean Pregnancy for all Non-Diabetic Patients is estimated to be between 3.033 Months and 3.562 Months. The Best Point Estimate is the center of the Confidence Interval.

3. Statement #3 and Hypothesis Testing

Considering Diabetic-Non Diabetic Patients, do Diabetic and Non-Diabetic Patients differ significantly in mean Pregnancy Months?

$$z = \frac{(\mu_3 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_3^2}{n}}}$$

Our Parameter of Interest is ($\mu 1 - \mu 3$): Pregnancy Months

μ_1: Mean Pregnancy Months of Diabetic Patients

μ_3: Mean Pregnancy Months of Non-Diabetic Patients

σ_1: Standard Deviation of Diabetic Patients

σ_3: Standard Deviation of Non - Diabetic Patients

Based on the Parameter, we could make a hypothesis testing as below :

H_0: There is no difference between Pregnancy Months of Diabetic and Non-Diabetic Patients

H_1: There exist difference on Pregnancy Months between Diabetic and Non-Diabetic Patients

From the program, we know that the best point estimate for (μ1 − μ3):

Pregnancy Months: 1.56767164179104

Moreover, the standard error is 0.4988975585853227

Therefore, we could obtain $z = \frac{(\mu_3 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_3^2}{n}}} = \frac{1.56767164179104}{0.4988975585853227} = 3.14$

Because we use confidence interval 95%, the critical value is:

If $-z_{0.025} \leq z \leq z_{0.025}$, then H_0 is accepted

Otherwise, H_0 is rejected

From Normal Distribution Table, we know that $z_{0.025}$ = 1.96, so if z located between -1.96 and 1.96, H_0 will be accepted. However, we obtain z = 3.14, which is outside that interval. Hence, we deduce that H_0 is rejected, implying that there is a difference between the Pregnancy Months of Diabetic and Non-Diabetic Patients.

We also want to know the Interval of Pregnancy Months difference between Diabetic and Non-Diabetic Patients. To calculate the mean difference confidence interval we will use the Pooled Approach which the variance of the two populations is assumed to be equal for both groups:

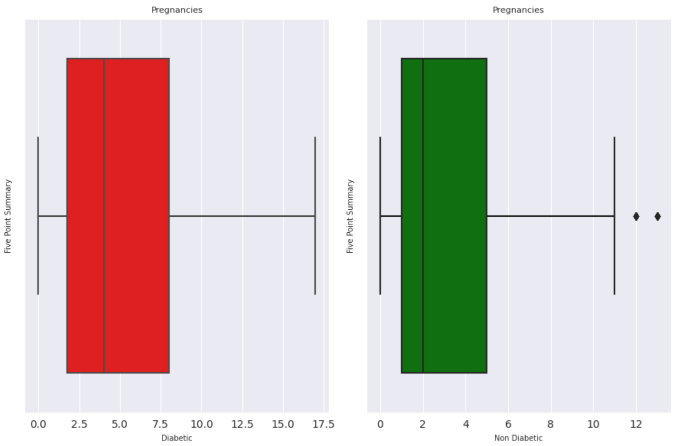


Fig. III.3 BoxPlot of Pregnancies Distribution for both Diabetic and Non-Diabetic Patients

Since the Variance of both Diabetic and Non-Diabetic Patients are nearly similar or the Interquartile Range is also almost the same. Thus we can proceed with Pooled Approach:

The Best Point Estimate for (μ1 − μ3):
Pregnancy Months: 1.567671641791045

Estimated Standard Error for (μ1 − μ3):
Pregnancy Months: 0.4988975585853227

The Margin of Error for (μ1 − μ3):
Pregnancy Months: 0.987817165998939

95% Confidence Interval for (μ1 − μ3):
Pregnancy Months: (0.579854475792106, 2.5554888077899838)

From this Confidence Interval, we can infer that with 95% confidence, the difference in Mean Pregnancy Months between Diabetic and Non-Diabetic Patients is estimated to be between 0.5798 Months and 2.5554 Months.

IV. References

[1] Bhandari, P., 2020. An introduction to inferential statistics. [online] Scribbr. Available at: <https://www.scribbr.com/statistics/inferential-statistics/> [Accessed 19 May 2022].

[2] Gaber, M., 2020. Hypothesis Testing Intuitively Explained using the Titanic Dataset in Python.. [online] Medium. Available at: <https://medium.datadriveninvestor.com/hypothesis-testing-intuitively-explained-using-the-titanic-dataset-in-python-5afa1e580ba6> [Accessed 18 May 2022].

[3] Yıldırım, S., 2021. Data Visualization with Pandas. [online] Medium. Available at: <https://towardsdatascience.com/data-visualization-with-pandas-1571bbc541c8> [Accessed 18 May 2022].