Performance Assessment D207 – Exploratory Data Analysis

Doug Haunsperger

College of IT, Western Governors University

October 2, 2023

# Part I. Research Question

## A1. Question Description

I am choosing to examine the question, “Is there a significant difference in the Vitamin D blood serum levels of patients who are readmitted to the hospital versus those who are not?”.

## A2. Analysis Benefit

The answer to this question could lead to improvements in patient treatment. For example, if it is shown that patients with higher levels of Vitamin D present tend to be readmitted less frequently, the hospital could attempt Vitamin D supplementation as a regular practice. Alternatively, if it is shown that there is no significant difference in Vitamin D level between the readmitted and non-readmitted populations, the hospital could save money on ineffective supplementation.

## A3. Data Identification

To answer this question, I will utilize the ‘ReAdmis’ and ‘VitD\_levels’ columns of the data set.

# Part II. Data Analysis

## B1. Code

Following is code to perform a two-sample t-test comparing the Vitamin D levels of the patient populations that were and were not readmitted:

Figure 1  
*Input code screenshot*

# A screenshot of a computer program Description automatically generated

This code is also included in the attached Jupyter notebook.

## B2. Output

Figure 2  
*Output screenshot*



The p-value returned is 0.683, which is much greater than the standard alpha = 0.05 threshold for significance (Hayden, Lesson 1, n.d.).

## B3. Justification

My research question is looking for a difference in a continuous variable (Vitamin D levels) between two populations based on a differentiating factor (whether or not the patient was readmitted). I chose to use a two sample Student’s t-test. “The [independent] T-test is a parametric test used to test for a statistically significant difference in the means between 2 groups. … [T]here are certain conditions that need to be met in order for the test results to be considered reliable.” (Python for Data Science, LLC [PDS], 2020)

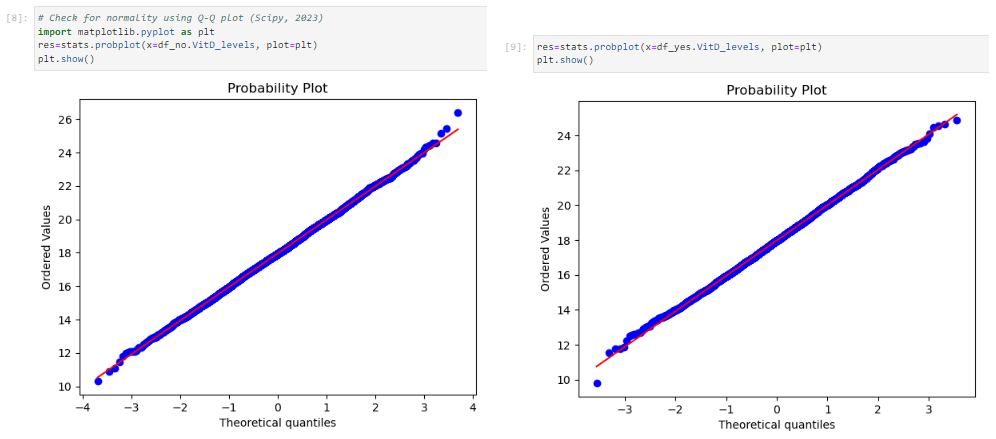
As given by PDS, a two-sample t-test is valid when three conditions are met: the population distributions are normal, samples have equal variances, and the two samples are independent. To check for normality, one can visually examine the distributions, and also use a Q-Q plot (Hayden, Lesson 4, n.d.)

Figure 3  
*Histogram plot of VitD\_levels, both non-readmitted and readmitted populations*

A graph of a number of levels

Description automatically generated with medium confidence

Figure 4  
*Q-Q plots of VitD\_levels, both non-readmitted and readmitted populations*



The histogram plot shows the normal distribution of both populations, confirmed by the linear nature of the Q-Q plots.

The second condition to check for is equal variance. It is simple to use the built-in std function to calculate standard deviation, which is the square root of variance (Matsui, n.d.):

Figure 5  
*Standard deviations of VitD\_levels, both non-readmitted and readmitted populations*

A screenshot of a computer code

Description automatically generated

The standard deviations are equal to within 1%, satisfying the second condition for validity of the t-test.

The final condition, independence of the samples, is guaranteed by the data cleaning process, that there is no duplicate patient information. The set of people who were readmitted is disjoint from the set who were not readmitted. Since all three conditions are met, therefore the t-test is a valid measure of whether the population means are statistically different. It is possible to say with high certainty that there is no meaningful difference in readmission based on Vitamin D blood serum levels.

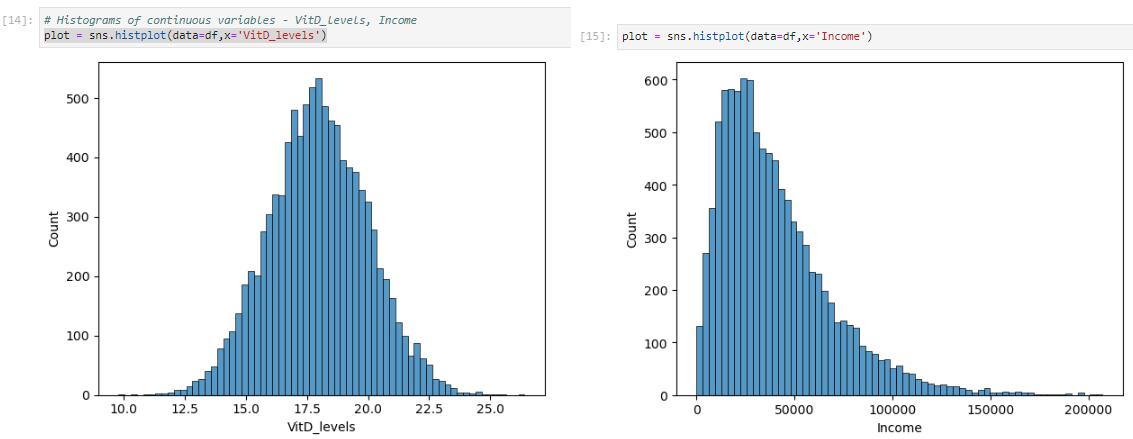
# Part III. Distribution Identification

## C./C1. Univariate Statistics & Visuals

In this section, I look at the distributions of two continuous and two categorical variables, plotting histograms for each to show the distributions visually.

First, I have chosen the continuous variables VitD\_levels and Income.

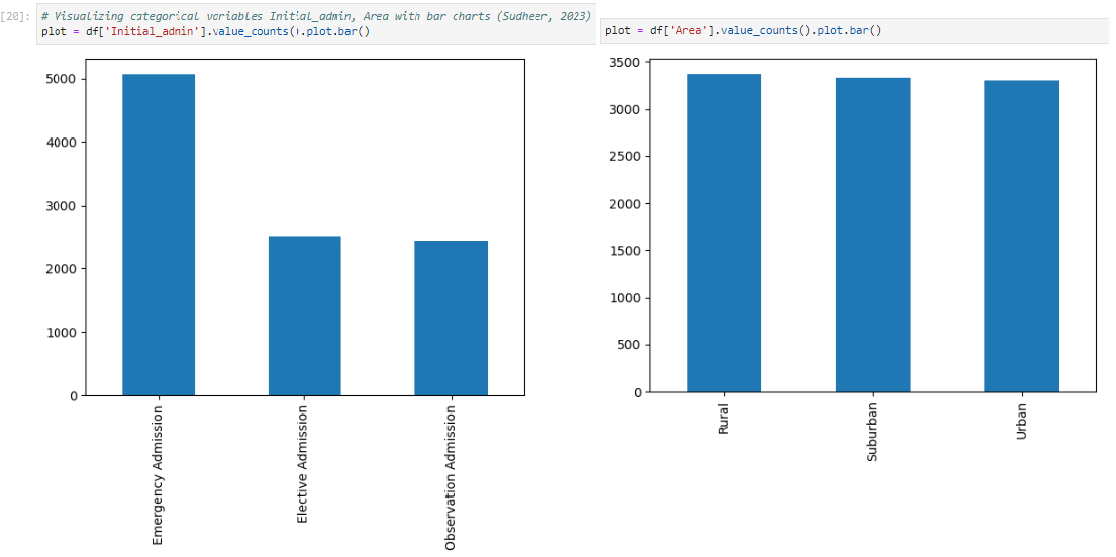
Figure 6  
*Histogram plots of Vitamin D levels and patient income*



VitD\_levels follows a normal distribution, while Income shows a rightward skew.

Next, I have chosen to look at the categorical variables Initial\_admin and Area.

Figure 7  
*Bar charts of Initial admission type (‘Initial\_admin’) and patient location type (‘Area’)*

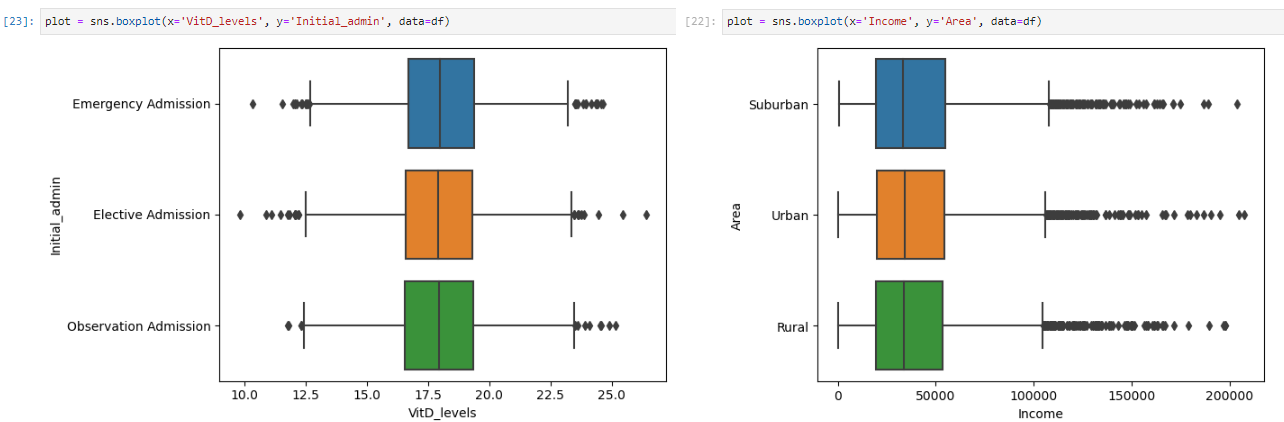


Initial\_admin shows roughly twice as many emergency admissions versus elective or observation, while Area shows a nearly equal distribution between patients.

## D./D1. Bivariate Statistics & Visuals

For bivariate analysis, I will plot the same variables as in part C, but now compared against one another – VitD\_levels vs. Initial\_admin, and Income vs. Area. The code I used is based on reference code from Kumar, 2022.

Figure 8  
*Box plots of Vitamin D levels vs. initial admission type, and patient income vs. location type*



# Part IV. Analysis Summary

## E1. Analysis Results

As explained above in part B, the data met the validity conditions for using a Student’s t-test. The p-value of 0.683 calculated by the test does not meet the threshold for statistical significance. Therefore, I conclude that there is no meaningful difference in hospital readmission based on patient Vitamin D blood serum levels.

## E2. Analysis Limitations

Even had I found a significant difference, it is not possible to conclude causality from a t-test. That is, a positive result would not necessarily mean that abnormally low or high blood serum levels caused hospital readmission, only that a correlation was present.

It is also possible that our negative result is a Type II error, i.e. a false negative (Hayden, Lesson 3, n.d.). Using the method given by Hayden, I calculate that the effect size, or Cohen’s d, that can be differentiated at a 0.05 alpha level with 95% certainty given the data set size is 0.084. Glen (n.d.) characterizes this Cohen’s d value as “trivial”, meaning that we can say with high confidence that there is no Type II error that would affect our conclusions.

Figure 9  
*Code snippet to calculate expected detectable effect size*

A screenshot of a computer code

Description automatically generated

## E3. Recommended Course of Action

Since no significant relationship was seen between Vitamin D blood serum levels and whether a patient was readmitted, the hospital can explore other avenues to determine how best to avoid readmissions. It is possible that Vitamin D supplementation is unnecessary, at least as a general rule, and money can be saved by discontinuing this practice except where otherwise medically necessary.

# Part V. Supporting Documentation

## F. Demonstration Video

## A video describing my methods and code can be found at: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b333a68a-7b8e-4251-87b2-b08f01264ff6>

## G. Third-party Code Sources

Hayden, L. (n.d.). Lesson 1: The Basics of Statistical Hypothesis Testing. *Performing Experiments in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/experimental-design-in-python/the-basics-of-statistical-hypothesis-testing?ex=5>

Hayden, L. (n.d.) Lesson 3: Sample Size, Power Analysis, and Effect Size. *Performing Experiments in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/experimental-design-in-python/sample-size-power-analysis-and-effect-size?ex=4>

Kumar, A. (2022, March 22). *A Quick Guide to Bivariate Analysis in Python*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/>

Scipy (2023). *Scipy.stats.probplot.* Scipy Documentation, v. 1.11.3. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html>

Sudheer, S (2023, July 21). *12 Univariate Data Visualizations With Illustrations in Python.* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/07/univariate-analysis-visualization-with-illustrations-in-python/>

## H. References

Glen, S. (n.d.) *Cohen’s D: Definition, Examples, Formulas*. Statistics How To. <https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/cohens-d/>

Hayden, L. (n.d.) Lesson 1: The Basics of Statistical Hypothesis Testing. *Performing Experiments in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/experimental-design-in-python/the-basics-of-statistical-hypothesis-testing?ex=5>

Hayden, L. (n.d.) Lesson 3: Sample Size, Power Analysis, and Effect Size. *Performing Experiments in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/experimental-design-in-python/sample-size-power-analysis-and-effect-size?ex=4>

Hayden, L. (n.d.) Lesson 4: Testing Normality: Parametric and Non-parametric Tests. *Performing Experiments in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/experimental-design-in-python/testing-normality-parametric-and-non-parametric-tests?ex=1>

Matsui, M (n.d.) Lesson 1: Summary Statistics. *Introduction to Statistics in Python* [MOOC]. DataCamp. <https://campus.datacamp.com/courses/introduction-to-statistics-in-python/summary-statistics-1?ex=1>

Python for Data Science, LLC (2020). *Independent T-Test.* <https://www.pythonfordatascience.org/independent-samples-t-test-python/>