**D207 Essay**

Darien Nguyen

Western Governors University

D207: Exploratory Data Analysis

December 3, 2022

**A.  Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:**

**1.  Provide one question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).**

Is there a correlation between readmission rate and complication risk for medical patients? Our hypothesis would be there is a correlation between readmission rate and complication risk. We believe that there is a higher chance of readmission if the patient has higher complication risk.

**2.  Explain how stakeholders in the organization could benefit from an analysis of the data.**

Stakeholders can spend more time on higher risk patients to lower their admission rate and save more time with lower risk patients. This will save cost and resources by reallocation of doctors and supplies based off patient risk.

This information can be used to identify potential risk factors for complications and readmissions among the hospital clients, and to develop strategies for reducing the risk of these outcomes. It can also help the hospital to better allocate resources and provide more targeted care to patients who are at higher risk of complications and readmissions.

**3.  Identify *all* of the data in your data set that are relevant to answering your question in part A1.**

 The data used will be patient readmission rate and complication risk.

**B.  Describe the data analysis by doing the following:**

**1.  Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:**

•   chi-square

•   t-test

•   ANOVA

newdf = pd.crosstab(df['Complication\_risk'], df['ReAdmis'])

print (newdf)

observed = newdf.values

print('observed:\n', observed)

stat, p, dof, expected = chi2\_contingency(observed)

print('dof=%d \n' % dof)

print (expected)

prob = 0.95

critical = chi2.ppf(prob, dof)

print ('probability = %.3f, critical = %.3f, stat = %.3f' % (prob, critical, stat))

if abs(stat) >= critical:

print('Dependent (reject HO)')

else:

print ('Independent (fail to reject HO)')

alpha = 1.0 - prob

print ('significance =%.3f, p =%.3f' % (alpha, p))

if p <= alpha:

print('Dependent (reject HO)')

else:

print ('Independent (fail to reject HO)')

Table

Description automatically generated**2.  Provide the output and the results of *any* calculations from the analysis you performed.**

Graphical user interface, application

Description automatically generated

Text

Description automatically generated with medium confidenceGraphical user interface, text, application

Description automatically generatedText

Description automatically generated

**3.  Justify why you chose this analysis technique.**

The chi-square test is a widely used and well-respected statistical tool for analyzing the relationship between categorical variables. One of the benefits of using the chi-square test to analyze the relationship between complication risk and readmission rate in patients is its simplicity, which makes it accessible to researchers with a basic knowledge of statistics. It is also a flexible tool that can be used to compare frequencies between different groups of patients, such as those with and without complications or those who were readmitted and those who were not. In addition, the chi-square test is relatively robust to violations of its assumptions, provided that the sample size is large. Overall, since the hypothesis is analyzing categorical data and a large data set, it fits the requirements of the chi-square test and was the best choice for analysis.

**C.  Identify the distribution of two continuous variables and two categorical variables using univariate statistics from your cleaned and prepared data.**

1. Chart, bar chart

   Description automatically generated*Chart, histogram

   Description automatically generated*Chart, box and whisker chart

   Description automatically generatedChart, bar chart

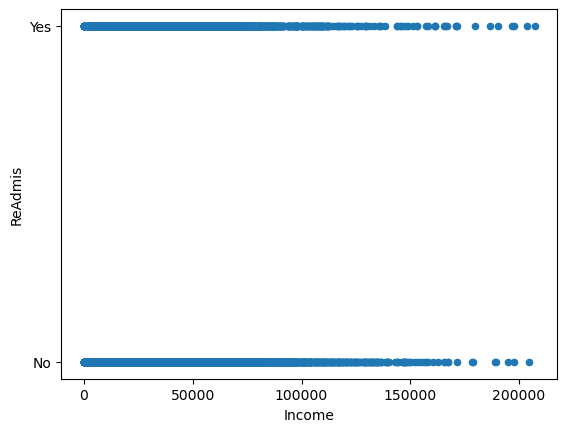
   Description automatically generated**Chart, box and whisker chart

   Description automatically generated**Chart, histogram

   Description automatically generated**Represent your findings in Part C, visually as part of your submission.**

**D.  Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data.**

1. **A picture containing chart

   Description automatically generated****Graphical user interface, application

   Description automatically generated**A picture containing text

   Description automatically generated**Represent your findings in Part D, visually as part of your submission.**

**E.  Summarize the implications of your data analysis by doing the following:**

**1.  Discuss the results of the hypothesis test.**

The chi-square test was conducted, and the resulting p-value was 0.924. Our significance level was set at 0.050. Based on the p-value being greater than the significance level, we conclude that there is no statistically significant correlation between complication risk and readmission rates. In other words, the two categorical data points are independent of each other and do not influence each other.

For the univariate tests, income showcases the average income of patients being around 50,000 with outliers being seen past 100,000. Age is a uniform graph giving us the interpretation that all ages are seen at the hospital in an even amount. With readmission rate, we observe approximately 3,700 patients are seen to readmit, while 6,000 patients are found not to readmit. We then see that bloodwork is the most common test being done on patients with intravenous being second.

For bivariate tests, it shows the relationship between the continuous variables of income and age to the categorical readmissions and services. Regarding age, we see a uniform range of services and readmissions to all age groups with no discrimination between the variables. However, income suggests that those who make over 150,000 are less likely be readmitted to the hospital.

**2.  Discuss the limitations of your data analysis.**

There are several limitations to using a chi-square test to analyze the relationship between readmission rate and complication risk in patients:

1. The chi-square test is only suitable for categorical (nominal or ordinal) data and is not appropriate for continuous data (Zibran, 2007, p.6).
2. The sample size must be sufficiently large. The larger the sample size, the more reliable the results of the chi-square test will be (Zibran, 2007, p.6).
3. The expected frequencies should be at least 5 for each category. If the expected frequencies are too small, the chi-square test may not be reliable (Zibran, 2007, p.6).

Overall, it is important to carefully consider the assumptions and limitations of the chi-square test when using it to analyze the relationship between readmission rate and complication risk in patients. If the assumptions of the test are not met, the results may not accurately reflect the relationship between the variables being analyzed.

**3.  Recommend a course of action based on your results.**

Since the null hypothesis was not rejected, we can conclude that there is no significant association between patient complication risk and readmission rate. This means that levels of complication do not have a direct association with patient readmission rates. As a result, the hospital should treat patient complications equally and fairly based on their needs.

**F.  Panopto Video**

[**https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bc75a571-a3b9-440a-a9b3-af6c014c5214&query=d207**](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=bc75a571-a3b9-440a-a9b3-af6c014c5214&query=d207)

**G.  Coding Sources**

No sources used.

**H.  Intext Sources**

Zibran, M. F. (2007). Chi-squared test of independence. *Department of Computer Science, University of Calgary, Alberta, Canada*, *1*(1), 1-7.