Michael Segaline

D207: Exploratory Data Analysis

Dr. Sewell

Aug 9, 2021

Abstract

Bottom line up front: The data suggests, overweight patients have no correlation with hospital readmission rates.

Stakeholders from a major hospital chain pondered what variables impact the readmission rate of patients. The readmission rate is problematic due to the costly implications imposed by third party regulatory agencies. The findings of this analysis can help save time via asset reallocation and eliminating variable inspection not related to hospital readmissions. Given the common knowledge, that being overweight leads to adverse health conditions, the question posited by the author is: *What is the statistical significance between overweight patients and readmission rates?* To answer the question the author needed to examine two variables, that of overweight patients and patient readmission; A chi-square test for independence was implemented, finding no statistical significance between the two variables. This analysis also conducted univariate analysis on the continuous variables of the of children and income; The univariate analysis categorical variables of services received, and overweight patients will be examined. Furthermore, bivariate analysis between patient income levels and number of children per patient will be explored. Lastly, the bivariate analysis between overweight patients and patient readmission will be examined. The graphic distributions of all these variables will be displayed for further study. A recommended course of action is to increase the amount of data to be explored and asses other variables. Lastly, for the hospital to possibly disregard overweight patients are a risk factor for readmission.

To discover the statistical significance (if any) between overweight patients and hospital readmission, the relevant variables from the dataset will be “Overweight” and “ReAdmis” (indicating readmission of patients to the hospital). The null hypothesis being there is no relationship between readmission of patients due to them being overweight. The alternative hypothesis being: Yes, there is a relationship. To test the null hypothesis, a Chi square test for independence was utilized because both variables are independent, categorical, assumed to be random, and each cell has a count more than five. The hypothesis test and exploratory data analysis was conducted via Python code. The following error-free code was used to execute this hypothesis test:

#-- Importing all possibly nessesary liberaries --#

import scipy.stats as stats

import statsmodels.api as sm

import numpy as np

import statistics

from scipy.stats import f\_oneway

from bioinfokit.analys import stat

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

import pandas as pd

from scipy.stats import chi2\_contingency

from scipy.stats import chi

from scipy.stats import chi2

import statsmodels.api as sm

from statsmodels.formula.api import ols

#-- reading the csv and converting it to a pandas dataframe--#

df =pd.read\_csv('/Users/lindasegalini/Desktop/WGU/New Program/Data Exporation/medical\_clean.csv')

#--Creating a new dataframe for chi-sqaure to analyis of readmission and overweight patients--#

df\_chi = pd.crosstab(index =df['ReAdmis'],

columns = df['Overweight'])

#-- Calling chi-square function with the new dataframe--#

#--Conducting chi-square test for independence--#

stat, p, dof, expected = chi2\_contingency(df\_chi)

#--Displaying the degrees of freedom--#

print("degrees of freedom =", dof)

print()

print("Expected = ",expected)

print()

# --Displaying the pvalue --#

print('P-value = ', p)

#--interpreting the test-statistic--#

prob =0.95

critical = chi2.ppf(prob, dof)

print( 'Alpha level =',prob ,'Critical Value =',critical, 'Test statistic =',stat)

if abs(stat) >= critical:

print('Dependent(reject H0)')

else:

print('Independent (fail to reject H0)')

#-- Interperating the p-value --#

alpha = 1.0 - prob

print('significance = ', alpha, p)

if p <= alpha:

print('Dependent (reject H0)')

else:

print('Independent (fail to reject H0)')

*Displaying the result*:

degrees of freedom = 1

Expected = [[1839.7886 4491.2114]

[1066.2114 2602.7886]]

P-value = 0.4032948387365497

Alpha level = 0.95 Critical Value = 3.841458820694124 Test statistic = 0.6984802059617877

Independent (fail to reject H0)

significance = 0.050000000000000044 0.4032948387365497

Independent (fail to reject H0)

With only one degree of freedom, the resulting p-value of, 0.403 is much bigger than the alpha level of .005. The alpha threshold levels are typically five percent or one percent; In this instance, the five percent probability was chosen with an alpha level of ninety-five percent with a resulting critical value of 0.05. The alpha level is how rare a probability must be for statistical significance. Plus, the Test Statistic is drastically smaller than the Critical Value. Therefore, there is a forty percent chance that the observed frequency distributions are independent and not correlated with one another. Ergo, there is no statically significant difference between the distributions and the null hypothesis is accepted. An item of consideration according to *Practical Statistics for Data Scientists*, “A p-value does not provide a good measure of evidence regarding a model or hypothesis. Plus, scientific conclusions and business policy decisions should not be based only on p-value. A p-value is a useful metric in situations where you want to know whether a model appears interesting or useful and is merely a point of information bearing on a decision” (Bruce and Gedeck 2020).

The following pages will display the distributions from exploratory analysis in Python. The first graph will demonstrate univariate analysis resulting in the displayed distributions of number of children per patient and patient income; Both distributions are arranged side by side. The second visual layout is frequency distribution of services received by patients. The third visual exhibit is of overweight vs non -overweight patients. The fourth visual displays the bivariate analysis as a Kernel Density Estimate between number of children per patient and patient income. The final image is a bivariate analysis between overweight patents and patient readmission.

Now showing the distribution of the number of children per patient and income level:

Chart, histogram

Description automatically generated

The distributions of both are right skewed.

Now displaying the categorical distribution of services received by patients.

Chart, bar chart

Description automatically generated

Blood Work being the most common and Intravenous a trailing second.

Now displaying the distribution of overweight patients:

Chart, bar chart

Description automatically generated

Overweight patients are more than double that of normal weight patients. One of the reasons behind the alternative hypothesis that overweight patients may be statistically significant to readmission rates.

Now showing the bivariate distribution between the number of children per patient and income level:

Text

Description automatically generated with medium confidence

This Kernel Density Estimate shows the majority of patients have less than two children and earn less than $50,000 a year.

Now showing the bivariate distribution between overweight patients and hospital readmission:

Chart, bar chart

Description automatically generated

The bivariate analysis gives some visual credence to the alternative hypothesis. Of the non-overweight patients, approximately one thousand where readmitted to the hospital. In contrast, the amount to overweight patients readmitted where more than double.

In conclusion, the hospital’s administration, that generated the analyzed dataset, could consider discounting overweight patients as a higher risk of hospital readmission. Furthermore, researchers may review other variables for possible correlation with readmission. Additionally, the analysis with the provided dataset is a generalization from a sample. More data from previous years and other hospitals may yield a different result and between the studied factors. Other variables could be explored for correlations with readmission rates and include variables not seen in the dataset. Lastly, the conclusion to the Chi-Square hypothesis test is there is no statistical connection between overweight patients and hospital readmission rates.

References

Bruce, P., Bruce, A., & Gedeck, P. (2020, May 19). *Practical Statistics for Data Scientists : 50+ Essential Concepts Using R and Python*. Ebook Central. https://www.ebookcentral.proquest.com/lib/westerngovernors-ebooks/reader.action?docID=6173908&ppg=66.