**D207 Task 1: Exploratory Data Analysis**

**Joseph Duszynski**

**ID #011237236**

**A.1. Question for analysis**

Using the medical data set (WGU, 2024 [1]), is there any correlation between patients with high blood pressure and assessed complication risk? Specifically, the analysis seeks to accept or reject the null hypothesis that patients with high blood pressure have the same proportions of complication risk (in categories of ‘Low’, ‘Medium’, and ‘High’) as patients in the general population (i.e. testing for independence on the two variables). The alternative hypothesis is that at least one of these proportions is different (on the basis of blood pressure). This will be assessed through a χ2-test with α = 0.05.

**A.2. Benefit from analysis**

Finding a correlation can assist medical staff in improving diagnoses, assessments, and ideal level of caution needed for each patient on the basis of their blood pressure. If there is no such correlation, medical staff can benefit from knowing high blood pressure alone plays no statistically significant role when assigning patient complication risks. A follow-up analysis (that would be outside of the scope of this one) could examine the accuracy of doctor’s complication risk assessments by comparing to patient outcomes.

**A.3. Data identification**

Testing the hypothesis in A.1. requires the variables ‘HighBlood’ and ‘Complication\_risk’ from the data set. ‘HighBlood’ is a binary ‘Yes’/’No’ qualitative (string/object) data element for patients with high blood pressure. ‘Complication\_risk’ is a qualitative (string/object) data element for a patient’s complication risk as assessed by a physician, assigned into one of three categories ‘Low’, ‘Medium’, or ‘High’.

**B.1. Code**

See code below for the chi-square test. Alternatively, see the attached files “med\_hypothesis\_testing.py” and the plain text file “med\_hypothesis\_testing.py”.

*import* numpy *as* np  
*import* pandas *as* pd  
*import* matplotlib.pyplot *as* plt  
*import* seaborn *as* sns  
*from* scipy.stats *import* chi2\_contingency  
  
  
df\_med = pd.read\_csv('D207\_medical\_data\medical\_clean.csv')  
  
cont\_table = pd.crosstab(df\_med['HighBlood'], df\_med['Complication\_risk'])  
*print*(cont\_table)  
chi2, p, dof, expected = chi2\_contingency(cont\_table)  
*print*(f"\nchi2 is {chi2} with {dof} degrees of freedom, giving p-value: {p}\n")  
df\_expect = pd.DataFrame(data=expected, index=['No', 'Yes'], columns=['High', 'Low', 'Medium'])  
*print*(f"Expected values were:\n{df\_expect}\n")

**B.2. Output**

A screenshot of a computer

Description automatically generated

From the above output, the χ2-test has two degrees of freedom, a value of 7.812, and a p-value of 0.020. The p-value is below α = 0.05, so the null hypothesis is rejected and the alternative hypothesis is accepted. High blood pressure influences at least one of the complication risk proportions compared to the general population proportions. They’re not independent.

**B.3. Justification**

The analysis question is testing for independence on two categorical variables through a contingency table. Pearson’s χ2-test is best suited for testing the hypothesis of independence between categorical variables. Additionally, it doesn’t require the variables’ distributions to be normal.

**C. Univariate statistics**

The categorical variables ‘Services’ (the patient’s primary service among ‘Blood work’, ‘Intravenous’, ‘CT Scan’, or ‘MRI’) and ‘Initial\_admin’ (the patient’s initial type of admission from ‘Emergency Admission’, ‘Elective Admission’, or ‘Observation Admission’), as well as the numerical variables ‘Initial\_days’ (the length of the patient’s initial stay in days) and ‘Additional\_charges’ (charges to the patient for miscellaneous billable expenses), were examined.

A screenshot of a computer

Description automatically generated

The ’Initial\_days’ distribution is bimodal, with the left-most peak at ~4.1 days having a positive skew and the right-most peak at ~67.3 days having approximately half its height and a negative skew. Patients’ initial stays cluster into “short” (under 30 days) or “long” (over 30 days).

The ‘Additional\_charges’ distribution is approximately “semi-uniform” at two different levels (resembling a step function). More specifically, the density was approximately uniform for additional charges between $6600 and $14200 (density ~ 7 \* 10-5) with a second level from $16100 to $28100 (density ~ 2 \* 10-5).

**C.1. Visual of findings**

A graph of a number of blue rectangular bars

Description automatically generated with medium confidence

A graph of a number of patients

Description automatically generated

**A graph of a number of days during a visit

Description automatically generated**

**A graph showing a line

Description automatically generated with medium confidence**

**D. Bivariate statistics**

The categorical variables ‘Overweight’ (‘Yes’ or ‘No’ to the patient being overweight) and ‘ReAdmis’ (‘Yes’ or ‘No’ to the patient being readmitted), as well as the numerical variables ‘TotalCharge’ (total expenses billed to the patient averaged across the number of days spent in the hospital) and ‘Income’ (the patient’s income), were examined.

Using a stacked bar chart for ‘Overweight’ and ‘ReAdmis’, the readmission rate for patients who weren’t overweight was 37.3% (1085/2906). For overweight patients, the readmission rate was 36.4% (2584/7094).

A scatter plot of ‘TotalCharge’ vs. ‘Income’ shows two clusters, one centered around a ‘TotalCharge’ of ~$3500 and the other at ~$7000. High complication risks and emergency admissions have disproportionately high daily charge values in each cluster.

**D.1. Visual of findings**

A chart with green and orange squares

Description automatically generated

A blue and white graph

Description automatically generated with medium confidence

**A graph showing the amount of charge

Description automatically generated with medium confidence**

**A chart of different colored circles

Description automatically generated**

**A chart with many colored dots

Description automatically generated with medium confidence**

**E.1. Results of analysis**

As discussed in section B.2., the χ2-test had a value of 7.812 and a p-value of 0.020. The null hypothesis set α = 0.05, and since the p-value is under 0.05, the null hypothesis of independence between high blood pressure and complication risk is rejected. The analysis accepts the alternative hypothesis that they are not independent.

**E.2. Limitations of analysis**

Testing the hypothesis of independence can confidently conclude that they’re not independent, but provides little insight into the strength of the correlation or any possible causality. The deviations from expected values show that the expected proportion of patients with high blood pressure having a risk level of ‘Low’ was reduced and spread out among the ‘Medium’ and ‘High’ levels among the observed patients. However, it’s possible the complication risk level assessed by physicians is informed by measuring the patient’s blood pressure, creating a trivial correlation from the physicians’ assessments. How the complication risk level is determined isn’t discussed in the data dictionary or other documentation.

Unfortunately, the blood pressure is a binary variable rather than a numerical pressure reading. Emergency situations with extremely high blood pressure that are determined to be at high risk of complication could be a cluster of patients that explain the majority of the discrepancies between expected and observed values. Or, there could be other correlations on other variables that remain undiscovered. A better understanding of these potential relationships would require additional data not currently found in this data set.

**E.3. Recommended course of action**

Since the null hypothesis of independence is rejected, we can conclude that there is a correlation between a patient’s blood pressure and their risk of complications. However, given the limitations of this analysis discussed above and in the interest of caution, we recommend a follow-up analysis that includes numerical blood pressure readings and specifics on how patient complication risk levels are assigned. Actual complication rates would be very helpful in determining the accuracy of physicians’ assessments. At this stage, informing medical staff of the correlation could alter their assessments to crudely assign risk levels primarily on the basis of blood pressure, contaminating any future data collected with a change in policy that would be hard to quantify and control for.

**F. Video**

See attached link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=ca25cf9e-9394-4126-a7bd-b15400072770>

**G. Sources for third-party code**

**1.** WGU. 2024. D207 Exploratory Data Analysis “Medical Data Dictionary and Data Set”. Medical Data and Dictionary Files. Retrieved April 15, 2024, from <https://access.wgu.edu/ASP3/aap/content/g9rke9s0rlc9ejd92md0.html>.

**H. Sources**

No additional sources were used.