

Chi-Square Hypothesis Testing on Patient Readmission and Hospital Services

Joanne Senoren

Master of Science, Data Analytics

Table of Contents

A1. Research Question.....	3
A2. Reasoning for Test Selection	3
A3. Data for Analysis	3
B1. Programming Code for Selected Analysis	4
B2. Output of Analysis	4
B3. Chi-Square Selection.....	4
C. Univariate Statistics.....	5
C1. Univariate Visualizations	6
D. Bivariate Statistics.....	8
D1. Bivariate Visualizations	9
E1. Overview of Hypothesis Test Results	11
E2. Data Analysis Limitations	12
E3. Actionable Recommendations.....	12
F. Sources for Third-Party Code	13
G. References.....	13

A1. Research Question

The following exploratory data analysis is written as supporting documentation for the inquiry posed in the provided dictionary, asking to predict patient readmissions based on factors surrounding the patient. The subsequent hypothesis test and data analyses are geared towards a more specific and related question: *Is there an association between patient readmission and services provided to the patient?*

For this hypothesis test, H_0 is defined as there is no association between the two variables, 'ReAdmis' and 'Services.' Alternatively, H_1 concludes that there is an association between patient readmission and services provided (*LibGuides: SPSS Tutorials: Chi-Square Test of Independence*, n.d.).

A2. Reasoning for Test Selection

By checking if there is a statistically significant association between readmission type and services provided to a patient, providers can determine whether to look further into how these variables affect readmissions. For example, the analytics team can investigate what specific illnesses were examined using services and analyze further to look for a pattern that can impact a patient's readmission status. Doing so will help them focus on refining initial treatments for those specific illnesses to lessen the number of readmissions.

Consequently, health providers doing this initial hypothesis test can lead them to explore the possibility that patients are getting side effects from their hospitalized service that may cause them to be readmitted, such as radiation poisoning. This deep dive could lead to a more serious inquiry: whether these services performed during hospitalization hurt patients more than help them.

A3. Data for Analysis

Two variables from the medical dataset are necessary to complete this exploratory analysis: 'ReAdmis' and 'Services.' 'ReAdmis' is a categorical variable with values that include yes and no on whether the patient was readmitted a month after being released from the original hospitalization.

'Services' is the primary service facilitated for the patient while hospitalized. 'Services' consist of the following values: blood work, intravenous, CT scan, MRI. These variable definitions were listed in the provided dictionary.

B1. Programming Code for Selected Analysis

Please see the attached Jupyter notebook file titled 'D207_EDA_Code_Resubmission' to view the executed code.

B2. Output of Analysis

Figure A

Chi-Square Hypothesis Test Results & Contingency Table

```
from scipy import stats

# Establish alpha value
alpha = 0.05

# Create crosstab to pass in chi2 test
table = pd.crosstab(df['ReAdmis'], df['Services'])

print(table)

# Perform chi2 test for independent variables

stat, p, dof, expected = stats.chi2_contingency(table)
if p < alpha:
    print('Variables indicate a correlation \nReject null hypothesis')
else:
    print('Variables do not indicate a correlation \nAccept null hypothesis')

print('Alpha: {} \nP-Value: {}'.format(alpha, p))

Services  Blood Work  CT Scan  Intravenous  MRI
ReAdmis
No          3335      737       2027    232
Yes         1930      488       1103    148
Variables indicate a correlation
Reject null hypothesis
Alpha: 0.05
P-Value: 0.03075281113212747
```

B3. Chi-Square Selection

The chi-square test of independence checks for one of the following hypotheses:

H_0 – the categorical variables are independent of each other

H_1 – the categorical variables are dependent on each other

Additionally, the test requires that variables are categorical, have at least two or more groups in the variable, are not known to be related, and consist of a sizable sample (*LibGuides: SPSS Tutorials: Chi-Square Test of Independence*, n.d.).

The stakeholder's goal is to lower readmissions, so I selected 'Readmis' as one of the variables for this hypothesis test. 'Readmis' is a categorical variable with yes and no values, separating the sample into two groups based on their readmission status. 'Services' is also a categorical variable with four groups based on the treatments provided. Readmission is not currently known to be related to services provided during the patient's initial hospitalization. Furthermore, all 10,000 rows from the dataset will be utilized for the analysis. Therefore, the chi-square test is the most appropriate for these two variables.

C. Univariate Statistics

The first two visualizations in Figure B and Figure C demonstrate two continuous variables' summary statistics and distribution.

Based on the generated histogram in Figure B for Vitamin D levels, the variable's distribution is normal, with a mean value of about 18 out of 26 (maximum value). The normal distribution is shown in Figure B with a bell-shaped curve, starting low at the left side, then exhibiting a peak around the mean, and going down again towards the right (Middleton, n.d.). We see that these values are between ~10 and ~26. It would be helpful to discuss with someone who has domain knowledge of Vitamin D levels how these values are defined into specific categories such as low, medium, or high levels.

The patient income distribution in Figure C demonstrates a right-skewed distribution because most of the sample is concentrated on the left side of the graph and gradually decreases towards the right side, creating a tail (Middleton, n.d.). This tells us that more patients fall within the lower side of income

around the mean, which is \$40,500. We would most likely see hospitalized patients who make less than \$50,000.

The next two visualizations, Figure D and Figure E, consist of pie graphs summarizing statistics and the distribution of two categorical variables – 'Initial_admin' and 'Complication_risk.' The distribution for a categorical variable provides the count or the percentage of the sample in each category (Nosedal, 2017).

The pie chart in Figure D shows the distribution of the groups within the categorical variable 'Initial_admin' as frequency percentages. The group mode or highest frequency is 'Emergency Admission' with 50.6% of the sample, followed by 'Elective Admission' and 'Observation Admission' at about 25% each. This distribution tells us that nearly half of the patients hospitalized entered with an 'Emergency Admission' status. We should note that the dataset only provides these three initial admission reasons, and it would be sensible to inquire about other potential reasons if they are available.

The distribution from the pie graph in Figure E shows that most patients have a 'Medium' complication risk level, followed by 'High' and then 'Low.' The top frequency in this variable, or mode, is 'Medium,' with 4,517 out of 10,000 patients from the sample. Obtaining a more precise definition of what risks are assessed per level would add more depth to the analysis.

C1. Univariate Visualizations

Figure B

'VitD-levels' Distribution Visualization and Summary Statistics

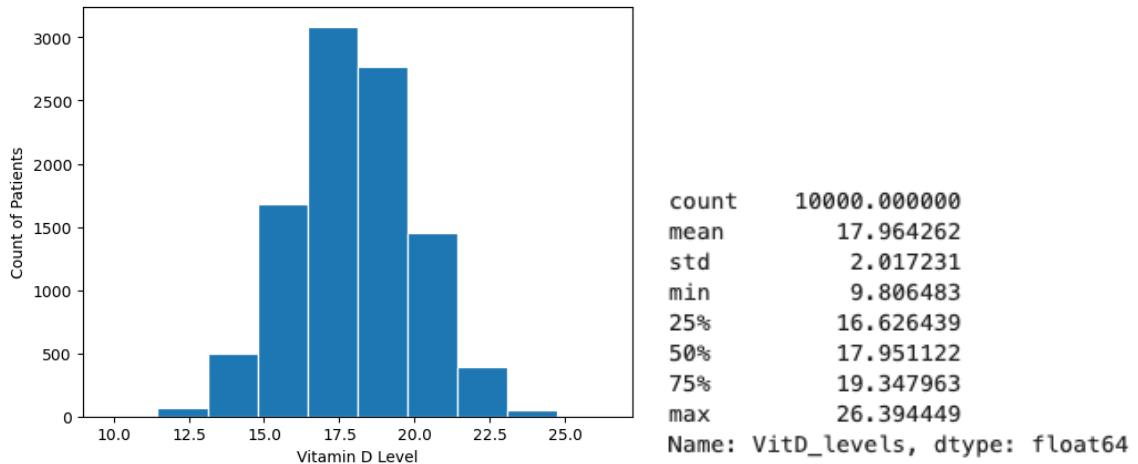


Figure C

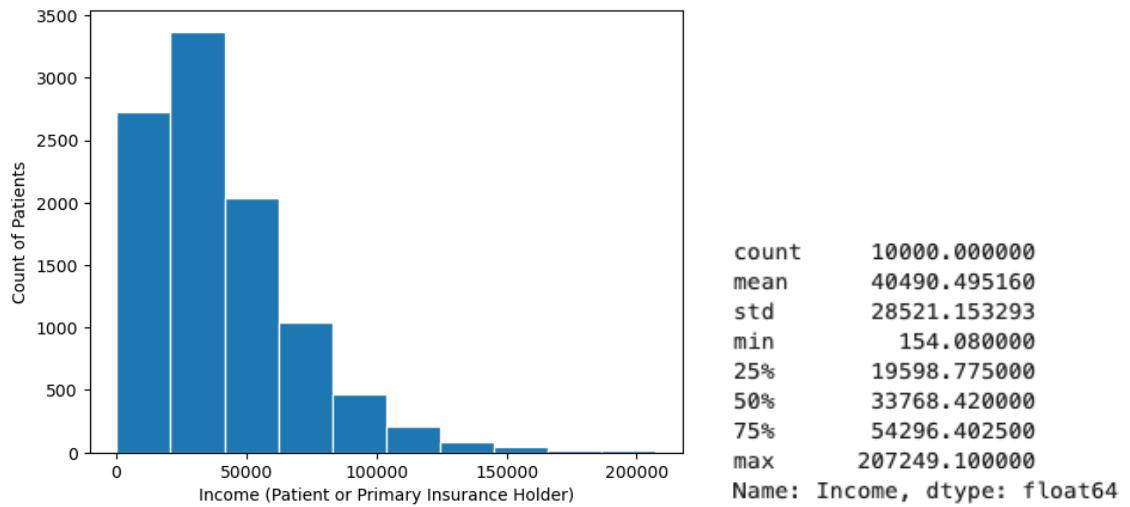
Income Distribution Visualization and Summary Statistics

Figure D

'Initial_Admin' Distribution Visualization and Summary Statistics

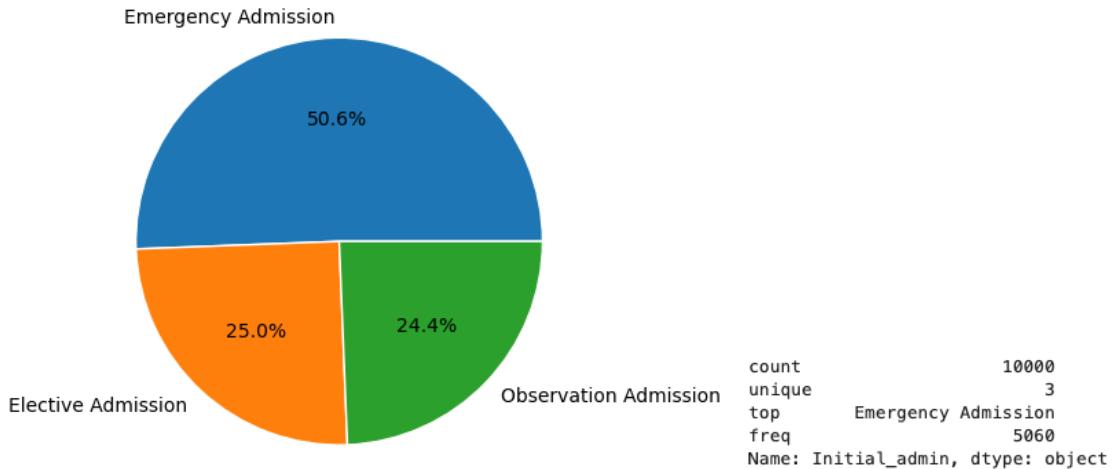
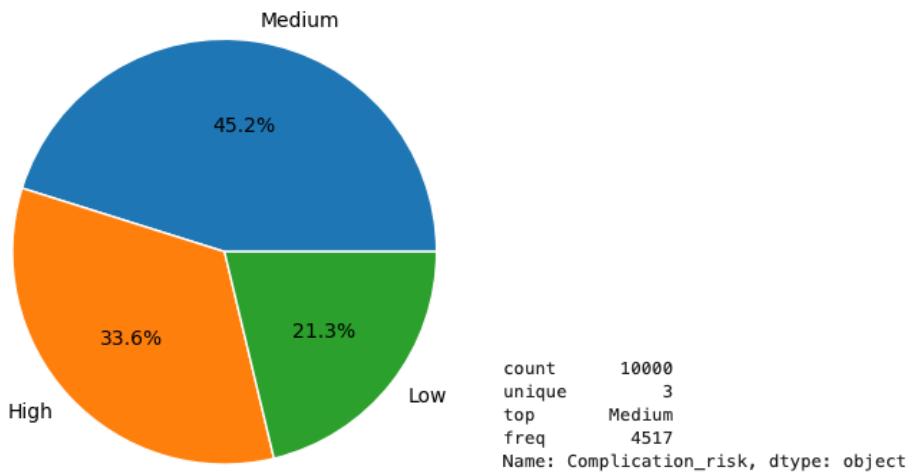


Figure E

'Complication_risk' Distribution Visualization and Summary Statistics



D. Bivariate Statistics

Figure F below is a scatterplot demonstrating a relationship between two continuous variables, 'Initial_days' and 'TotalCharge.' We see a linear distribution and positive correlation between total daily charges and initial days, as shown by the concentration of points starting at the bottom left of the graph and going to the top right (Middleton, n.d.). The calculated correlation coefficient in Figure G informs us that the relationship is very strong and statistically significant (Luna & datacamp, 2022). If we were to

take the dictionary definition that 'TotalCharge' is the average daily charge, we would expect a uniform distribution since 'TotalCharge' theoretically could stay the same per patient regardless of the length of days. But that is not the case here. The graph shows a pattern of increased average daily charges against longer days in the hospital, so we need to discuss with the team and ask why longer stays increase the daily average charge. We should also note a charge gap between 30 and 40 days of hospitalization and ask someone who might know why no charges are reflected for these durations. We must further examine what other variable(s) might be affecting the relationship of these two variables.

The chart in Figure H examines the distribution of readmissions by gender. It is necessary to order stacked bar charts from largest to smallest (Atlassian, n.d.). These charts also commonly represent the counts and frequencies of comparison between two categorical variables (Hazra & Gogtay, 2016). The gender distribution is concentrated between male and female patients and is nearly uniform, whereas the nonbinary count is minuscule. The cross-tabulation or contingency table (Figure H) is also used to simultaneously display counts and distribution for two or more categorical variables (Hazra & Gogtay, 2016). The percentage summary of the charted distribution shows of those readmitted, 49.4% are female, 48.3% are male, and 2.3% are nonbinary. The chi-square test for independence and the generated p-value of 0.45 in Figure I inform us that the relationship between the variables is not statistically significant. Based on the sensitivity of chi-square tests regarding sample sizes, we should consider checking whether this current sample is large enough.

D1. Bivariate Visualizations

Figure F

'Initial_days' vs. 'TotalCharge' Scatterplot and Summary Statistics

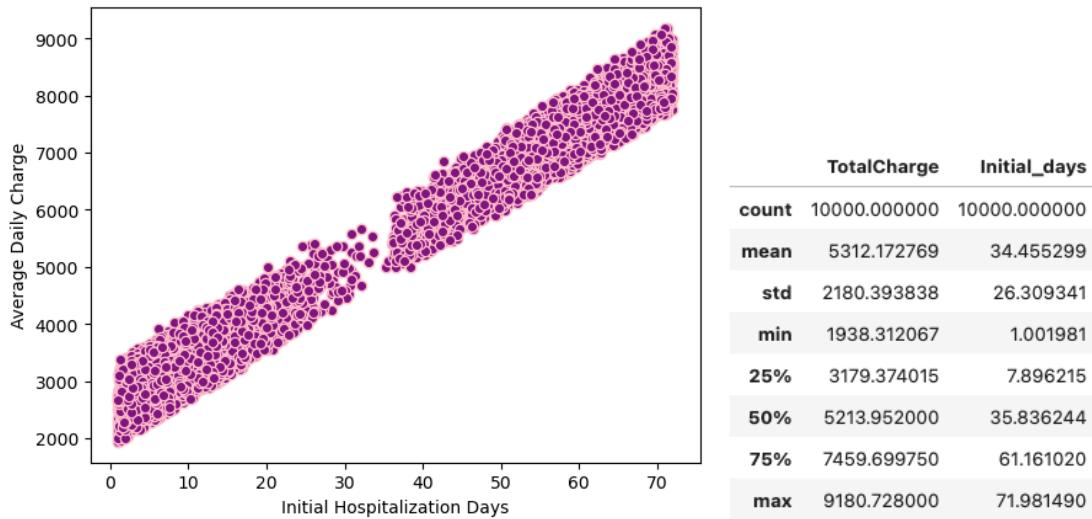


Figure G

'Initial_days' vs. 'TotalCharge' Correlation Coefficient Calculation

```

: # Generate correlation coefficient and statistical significance

# Extract Results
statistic, pvalue = stats.pearsonr(df['Initial_days'], df['TotalCharge'])

print('Correlation Coefficient: {}'.format(statistic))
print('Alpha: 0.05')
print('P-Value: {}'.format(pvalue))

if statistic >= .90:
    print('Correlation coefficient indicates a very strong relationship between target variables')
elif statistic >= .70:
    print('Correlation coefficient indicates a strong relationship between target variables')
elif statistic >= .50:
    print('Correlation coefficient indicates a moderate relationship between target variables')
elif statistic >= .30:
    print('Correlation coefficient indicates a weak relationship between target variables')
else:
    print('Correlation coefficient does not indicate a relationship between target variables')

```

Correlation Coefficient: 0.9876402655398173
Alpha: 0.05
P-Value: 0.0
Correlation coefficient indicates a very strong relationship between target variables

Figure H

'Gender' and 'ReAdmis' Distribution Visualization and Cross-Tabulation

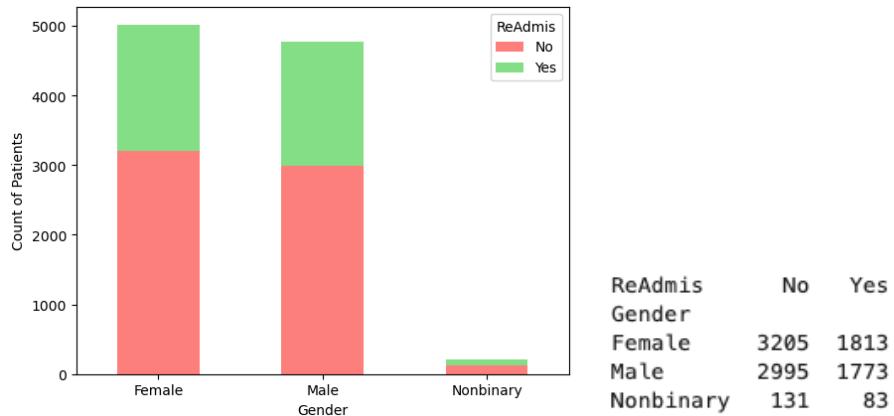


Figure I

Chi-Square Test for 'Gender' and 'ReAdmis'

```
# Chi-Square Test
gender_table = pd.crosstab(df['Gender'], df['ReAdmis'])
stat, p, dof, expected = stats.chi2_contingency(gender_table)

print('Alpha: {} \nP-Value: {}'.format(alpha, p))

if p < alpha:
    print('Variables indicate a correlation')
else:
    print('Variables do not indicate a correlation')

Alpha: 0.05
P-Value: 0.4525370014241822
Variables do not indicate a correlation
```

E1. Overview of Hypothesis Test Results

The p-value generated from the chi-square test is 0.03, less than the pre-determined alpha of 0.05.

It indicates statistically significant dependent variables. Therefore, we must reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1), which states there is an association between the variables in question (*LibGuides: SPSS Tutorials: Chi-Square Test of Independence*, n.d.). With that said, we can assume that from the sample of 10,000 patients, we will see an association between 'Readmis' and 'Services' when extrapolating the test to the population.

E2. Data Analysis Limitations

There are several limitations to the chi-square independence test, the variables, and the dataset regarding how to move forward with the hypothesis results. First, the chi-square test only examines the association among the proportions of the groups and values between the two variables. It does not provide a strong sense of causation, just that there is a perceived relationship. Some of the frequencies in the generated cross-tabulation table are small, and the chi-square test is sensitive to sample size (Hazra & Gogtay, 2016). Thus, a larger sample could attain a more accurate p-value. Additionally, the pre-determined alpha is 0.05, which means there is a five percent chance that this test resulted in a false positive.

The dictionary cannot specify certain aspects of variables that might affect analysis. For example, have these patients been readmitted multiple times? If patients have been readmitted numerous times in the sample, would there be an association to that specific group versus those that have only been readmitted once? The 'Services' variable vaguely assumes that all patients had participated in a hospitalized service. How can we compare this against those who didn't have services performed or those who had unlisted services completed? Looking into these questions can help us analyze and generate more substantive and well-informed results.

E3. Actionable Recommendations

The chi-square test indicated an association between the two variables 'Services' and 'ReAdmis.' However, we need to get additional information about the dataset, confirm the sample size is sound, and consider other variables that influence 'Services' and 'ReAdmis.' Recommended actions include prioritizing further investigation into the variables' definitions, adding to the hypothesis test sample size, and testing for confounding variables that may affect both readmissions and services. After further exploratory analysis, the team can consider a more astute direction for finding and refining treatments (services) to potentially lower hospital readmissions.

F. Sources for Third-Party Code

1. *Chi-square test | Python.* (n.d.). DataCamp.
<https://campus.datacamp.com/courses/performing-experiments-in-python/the-basics-of-statistical-hypothesis-testing?ex=10>
2. Coder, R. (2022, September 28). *Stacked bar chart in matplotlib.* PYTHON CHARTS | the Definitive Python Data Visualization Site. <https://python-charts.com/part-whole/stacked-bar-chart-matplotlib/>
3. GeeksforGeeks. (2023, April 26). *Python Pearson's Chi-Square Test.* GeeksforGeeks.
<https://www.geeksforgeeks.org/python-pearsongs-chi-square-test/>
4. *Pie charts — Matplotlib 3.9.0 documentation.* (n.d.).
https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html
5. *Stacked Bar Charts with Labels in Matplotlib.* (2021, January 22).
<https://www.pythontutorials.com/matplotlib/stacked-bar-charts-labels/>

G. References

1. Atlassian. (n.d.). *Stacked Bar Charts: A detailed breakdown | Atlassian.*
<https://www.atlassian.com/data/charts/stacked-bar-chart-complete-guide>
2. Hazra, A., & Gogtay, N. (2016). Biostatistics series module 4: Comparing groups - categorical variables. *Indian Journal of Dermatology/Indian Journal of Dermatology*, 61(4), 385. <https://doi.org/10.4103/0019-5154.185700>
3. *LibGuides: SPSS Tutorials: Chi-Square Test of Independence.* (n.d.).
<https://libguides.library.kent.edu/SPSS/ChiSquare>

4. Luna, J. & datacamp. (2022, February). *Python details on correlation tutorial*.
www.datacamp.com. Retrieved June 30, 2024, from
<https://www.datacamp.com/tutorial/tutorial-datails-on-correlation>
5. Middleton, K. (n.d.). *Data Types, Distributions and Univariate Imputation*.
<https://app.kyronlearning.com/>. Retrieved June 17, 2024, from
https://app.kyronlearning.com/video_player/631
6. Nosedal, A. (2017). *Displaying and describing categorical data* (pp. 1–79).
<https://mcs.utm.utoronto.ca/~nosedal/sta215/sta215-chap2.pdf>