|  |
| --- |
| Hospital Readmissions and Initial Length of Stay |
| Western Governors University |
| Heather Rauch |

Table of Contents

[A: Real-World Organizational Situation or Issue 2](#_Toc114686067)

[A1: Question for Analysis 2](#_Toc114686068)

[A2: Benefit from Analysis 2](#_Toc114686069)

[A3: Data Identification 2](#_Toc114686070)

[B: Describe the Data 2](#_Toc114686071)

[B1: Code 2](#_Toc114686072)

[B2: Output 4](#_Toc114686073)

[B3: Justification 6](#_Toc114686074)

[C: Univariate Statistics 6](#_Toc114686075)

[Categorical Variable Distributions 6](#_Toc114686076)

[Continuous Variable Distributions 6](#_Toc114686077)

[C1: Visual of Findings 7](#_Toc114686078)

[D: Bivariate Statistics 7](#_Toc114686079)

[Two Categorical Variables 7](#_Toc114686080)

[Two Continuous Variables 8](#_Toc114686081)

[D1: Visual of Findings 8](#_Toc114686082)

[Two Categorical Variable Distribution 8](#_Toc114686083)

[Two Continuous Variable Distribution 9](#_Toc114686084)

[E: Implications 9](#_Toc114686085)

[E1: Results of Analysis 9](#_Toc114686086)

[E2: Limitations of Analysis 10](#_Toc114686087)

[E3: Recommended Course of Action 10](#_Toc114686088)

[F: Video 10](#_Toc114686089)

[G: Sources for Third-Party Code 10](#_Toc114686090)

[H: Sources 11](#_Toc114686091)

# A: Real-World Organizational Situation or Issue

## A1: Question for Analysis

Research Question: Do patients who are readmitted to the hospital have a longer initial stay?

Hypotheses:

H0: The mean initial stay for patients who were readmitted to the hospital is the same as the initial stay for patients who were not readmitted to the hospital

HA: The mean initial stay for patients who were readmitted to the hospital is larger than the initial stay of patients who were not readmitted to the hospital.

## A2: Benefit from Analysis

The national medical chain for whom this analysis is being performed is interested in determining to what extent readmission is an issue in their health system so that they may reduce their rates to better align with national standards. If hospital administration were made to better understand how the length of patients’ initial stays in the hospital varies by whether they are readmitted, better decisions could be made about how and where to prioritize resources to reduce the chances of readmission.

## A3: Data Identification

The data used to answer the research question contains 2 columns and 10000 observations. The independent variable, ReAdmis, is categorical and contains two groups, “Yes” and “No”, representing whether they were readmitted to the hospital after their initial stay. The dependent variable, Initial\_days, is continuous and represents the number of days the patient was admitted during their initial stay. To answer the research question, the Initial\_days variable will be split into two groups based on whether ReAdmis is “Yes” or “No” and the mean of Initial\_days will be compared. The difference will be checked for statistical significance by using a two-sample right-tailed t-test with a significance level of 0.05.

# B: Describe the Data

## B1: Code

The following code was used to prepare for, perform, and display the findings of the hypothesis test. This and the remaining code used to perform the univariate and bivariate analyses can be found in the attached file, “task1.py”.

Note: Just after the imports there is a line that reads “pd.set\_option(‘chained\_assignment’, None)”. This line prevents a warning issued by the pandas library indicating that the code may not be performing the intended action ((https://stackoverflow.com/users/644898/jeff), 2014). Before adding that line, all code was tested to ensure the intended actions were performed.

**import** pandas **as** pd  
**import** matplotlib.pyplot **as** plt  
**import** seaborn **as** sns  
**from** scipy.stats **import** ttest\_ind   
  
pd.set\_option(**'chained\_assignment'**, **None**)  
  
MEDICAL\_CLEAN = **r'C:\Users\user\OneDrive\Documents\Education\Western Govenors University\MS - Data Analytics\D207 - '** \  
 **r'Exploratory Data Analysis\medical\_clean.csv'***# Import the data*medical\_clean = pd.read\_csv(MEDICAL\_CLEAN)  
data = medical\_clean[[**'ReAdmis'**, **'Initial\_days'**]]  
  
*# View Initial\_days distribution*fig, ax = plt.subplots(figsize=(6, 4))  
fig.suptitle(**'Initial\_days Distribution'**)  
ax.hist(data.Initial\_days, bins=int(max(data.Initial\_days) / 10))  
ax.set\_xlabel(**'Initial\_days'**)  
ax.set\_ylabel(**'Frequency'**)  
plt.savefig(**'./output/initial\_days\_distribution.png'**)  
  
*# Assess variance by ReAdmis groups*a = data.query(**'ReAdmis == "Yes"'**).Initial\_days  
b = data.query(**'ReAdmis == "No"'**).Initial\_days  
print(**f'The variance of Group A is: {**a.var()**}\n'  
 f'The variance of Group B is: {**b.var()**}\n'**)  
  
*# View transformed distribution*data[**'Initial\_days\_transformed'**] = abs(data.Initial\_days - data.Initial\_days.mean())  
fig, ax = plt.subplots(figsize=(6, 4))  
fig.suptitle(**'Transformed Initial\_days Distribution'**)  
ax.hist(data.Initial\_days\_transformed)  
ax.set\_xlabel(**'abs(Initial\_days - mean(Initial\_days))'**)  
ax.set\_ylabel(**'Frequency'**)  
plt.savefig(**'./output/initial\_days\_distribution\_transformed.png'**)  
  
*# Assess variance by ReAdmis groups after transformation*a\_transformed = data.query(**'ReAdmis == "Yes"'**).Initial\_days\_transformed  
b\_transformed = data.query(**'ReAdmis == "No"'**).Initial\_days\_transformed  
print(**f'The variance of Group A is: {**a\_transformed.var()**}\n'  
 f'The variance of Group B is: {**b\_transformed.var()**}\n'**)  
  
*# View group distributions*fig, ax = plt.subplots(figsize=(6, 4))  
fig.suptitle(**'Transformed Initial\_days Distribution by ReAdmis'**)  
ax.hist(a\_transformed, color=**'r'**, alpha=0.5, label=**'Yes'**)  
ax.hist(b\_transformed, color=**'b'**, alpha=0.5, label=**'No'**)  
ax.set\_xlabel(**'abs(Initial\_days - mean(Initial\_days))'**)  
ax.set\_ylabel(**'Frequency'**)  
ax.legend(title=**'ReAdmis'**)  
plt.savefig(**'./output/initial\_days\_distribution\_by\_readmis\_transformed.png'**)  
  
*# Two-sample t-test right-tailed*result = ttest\_ind(a\_transformed, b\_transformed, alternative=**'greater'**)  
print(**f'two-sample right-tailed t-test\n'  
 f'\tt-statistic: {**result.statistic**}\n'  
 f'\tp-value: {**result.pvalue**}\n'**)  
  
*# Boxplots*fig, ax = plt.subplots(figsize=(6, 4))  
fig.suptitle(**'Transformed Initial\_days by ReAdmis'**)  
ax.boxplot([a\_transformed, b\_transformed], labels=[**'Yes'**, **'No'**], sym=**''**)  
ax.set\_xlabel(**'ReAdmis'**)  
ax.set\_ylabel(**'abs(Initial\_days - mean(Initial\_days))'**)  
plt.savefig(**'./output/initial\_days\_by\_readmis\_transformed.png'**)

## B2: Output

Chart, histogram

Description automatically generated

The variance of Group A is: 25.789959313868575

The variance of Group B is: 286.8674853268104

Chart, histogram

Description automatically generated

The variance of Group A is: 25.789959313868575

The variance of Group B is: 57.03060432372243

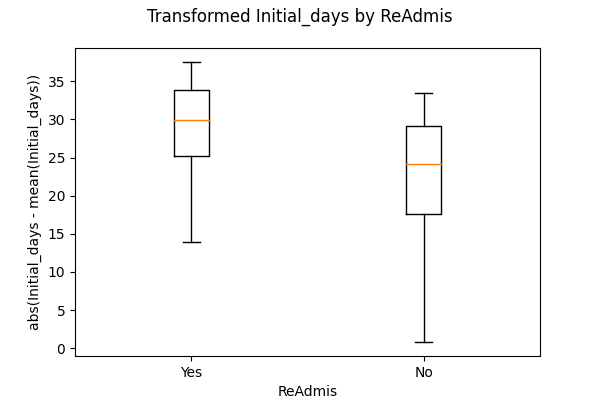
Chart, histogram

Description automatically generated

two-sample right-tailed t-test

t-statistic: 47.097790625495236

p-value: 0.0



## B3: Justification

A two-sample t-test was selected for this analysis because the goal was to check the difference between two independent group means from the same sample for statistical significance. The test was performed as a right-tailed test because the alternate hypothesis states that the mean Initial\_days for patients who have been readmitted will be greater than the mean of those who were not.

# C: Univariate Statistics

## Categorical Variable Distributions

|  |  |  |
| --- | --- | --- |
|  | **Initial\_admin** | **Services** |
| **Count** | 10000 | 10000 |
| **Unique Count** | 3 | 4 |
| **Top** | Emergency Admission | Blood Work |
| **Frequency** | 5060 | 5265 |
| **Unique Values** | Emergency Admission, Elective Admission, Observation Admission | Blood Work, Intravenous, CT Scan, MRI |

## Continuous Variable Distributions

|  |  |  |
| --- | --- | --- |
|  | **Initial\_days** | **TotalCharge** |
| **Count** | 10000.000000 | 10000.000000 |
| **Mean** | 34.455299 | 5312.172769 |
| **Standard Deviation** | 26.309341 | 2180.393838 |
| **Minimum** | 1.001981 | 1938.312067 |
| **25th Percentile** | 7.896215 | 3179.374015 |
| **Median** | 35.836244 | 5213.952000 |
| **75th Percentile** | 61.161020 | 7459.699750 |
| **Maximum** | 71.981490 | 9180.728000 |
| **Skewness** | 0.070286 | 0.069661 |
| **Kurtosis** | -1.754525 | -1.668267 |
| **Shape** | Bimodal | Bimodal |

## C1: Visual of Findings

Diagram

Description automatically generated with medium confidence

# D: Bivariate Statistics

## Two Categorical Variables

Initial\_admin vs Services

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Initial\_admin** | | | |
| **Services** | **Emergency Admission** | **Elective Admission** | **Observation Admission** | **Total** |
| **Blood Work** | 2652 (26.52%) | 1316 (13.16%) | 1297 (12.97%) | **5265 (52.65%)** |
| **CT Scan** | 632 (6.32%) | 322 (3.22%) | 271 (2.71%) | **1225 (12.25%)** |
| **Intravenous** | 1575 (15.75%) | 771 (7.71%) | 784 (7.84%) | **3130 (31.3%)** |
| **MRI** | 201 (2.01%) | 95 (0.95%) | 84 (7.84%) | **380 (3.8%)** |
| **Total** | **5060 (50.6%)** | **2504 (25.04%)** | **2436 (24.36)** | **10000 (100%)** |

## Two Continuous Variables

Initial\_days vs TotalCharge

Correlation Coefficient: 0.9876402655398173

R-squared: 0.9754332941155611

Line of best fit:

## D1: Visual of Findings

### Two Categorical Variable Distribution

Chart, bar chart

Description automatically generated

### Two Continuous Variable Distribution

Chart, scatter chart

Description automatically generated

# E: Implications

## E1: Results of Analysis

This analysis set out to answer the research question, “Do patients who are readmitted to the hospital have a longer initial stay?” From this question, the following were formed for the null (H0) and alternate (HA) hypotheses:

H0: The mean initial stay for patients who were readmitted to the hospital is the same as the initial stay for patients who were not readmitted to the hospital

HA: The mean initial stay for patients who were readmitted to the hospital is larger than the initial stay of patients who were not readmitted to the hospital.

To test these hypotheses, a two-sample, right-tailed, t-test was performed with an alpha of 0.05. The two samples were created based on the outcome variable, ReAdmis. Observations where ReAdmis was “Yes” were placed in group A, and observations where ReAdmis was “No” were placed in group B. Before performing the t-test, a data transformation was required to account for the bimodal distribution of Initial\_days as well as the large gap in variance between the two groups (A: 25.79, B: 286.87). The transformation performed was ((https://stats.stackexchange.com/users/2958/henry), 2016):

This resulted in a distribution that was much closer to normal with variances that were more similar than previously (A: 25.79, B: 57.03).

Once the test was performed, it was found that Initial\_days was greater in patients who were readmitted than those who were not, t(9998) = 47.098, p < .05. Because there was less than a five percent chance of this observation occurring if H0 were true, H0 can be rejected in favor of HA. This means that patients who were readmitted to the hospital tend to have longer initial stays on average.

## E2: Limitations of Analysis

The two-sample t-test has two assumptions that must be met: that the dependent variable is normally distributed and that each group’s variance is the same (Sewell, 2021) (Elrod, 2004). While transformations were performed to help correct violations of both assumptions, the problems were not irradicated. The post-transformation distribution appears to be slightly skewed in the negative direction. Additionally, the variance difference between the two groups is still approximately 25, meaning group B’s variance is twice that of group A’s.

## E3: Recommended Course of Action

Based on the results of the hypothesis test, it can be said that patients who stay in the hospital longer are more likely to be readmitted. Knowing this and the hospital administration’s ultimate goal of reducing readmission rates, the next step should be to collect additional information about the subset of patients who had an extended initial stay in the hospital. Information like the admitting diagnosis, whether they had surgery and/or complications from surgery, or even preexisting conditions could be used to tune a classification algorithm to predict whether a patient is likely to be readmitted. If an algorithm such as this were implemented in an inpatient hospital setting, it could give staff the awareness needed to reduce individual patients’ likelihood of readmission.

# F: Video

The Panopto recording titled ‘eda\_task1’ can be found at this link which is also attached to the submission: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e082e30a-61d4-4284-9427-af180174cbc6>

# G: Sources for Third-Party Code

(https://stackoverflow.com/users/644898/jeff), J. (2014, January 30). *python - Pandas: Chained assignments.* Retrieved from stackoverflow.com: https://stackoverflow.com/a/21463854

(https://stats.stackexchange.com/users/2958/henry), H. (2016, April 25). *What transformation should I use for a bimodal distribution?* (C. Validated, Editor) Retrieved from stats.stackexchange.com: https://stats.stackexchange.com/q/209261

# H: Sources

Elrod, D. (2004). *Assumptions for the t-test.* Retrieved from cisic.cornell.edu: http://www.csic.cornell.edu/Elrod/t-test/t-test-assumptions.html#:~:text=Assumptions%20for%20the%20t-test%201%20Bivariate%20independent%20variable,from%20each%20other%20%28across%20i%29.%20...%20More%20items

Sewell, W. (2021, May 26). *William Sewell's meeting-20210526 2159-1.* Retrieved from wgu.webex.com: https://wgu.webex.com/recordingservice/sites/wgu/recording/8ebcbb20a09b1039bf7f0050568116ea/playback