Part A

1. Research question: Is there a statistically significant difference between the mean duration of initial hospital stays and patients who are or are not readmitted?
2. This data-driven approach is valuable for hospitals because it provides effective and valuable insight into potential impact of the duration of hospital stays preventing patient readmissions. By understanding the potential relationship between initial hospital stay duration and readmission probabilities, hospitals can develop and enhance their reduction strategies, which can contribute to effective resource allocation, patient management, and increased overall healthcare quality.
3. Clean medical dataset
    a. ReAdmis variable: If the patient was readmitted within a month of release (yes, no)
    b. Initial_days variable: Number of days the patient stayed in the hospital during the initial visit

Part B

1. Two-sample t-test: will test the likelihood that there is a difference between two means of populations
    a. Two sample t-test code:
        - Code to import the CSV file data:

```
import pandas as pd
import numpy as np
medical_data_clean = pd.read_csv
('/Users/jasminemoniquecooper/Downloads/medical_clean.csv')
pd.set_option('display.max_columns', None)
medical_data_clean.head(10)
```

        - Code to conduct the t-test:

```
from scipy import stats
Group_ReAdmitted = medical_data_clean[medical_data_clean.ReAdmis == "Yes"]
Group_Not_ReAdmitted = medical_data_clean[medical_data_clean.ReAdmis ==
"No"]

Group_ReAdmitted_Initial_days = Group_ReAdmitted.Initial_days
Group_Not_ReAdmitted_Initial_days = Group_Not_ReAdmitted.Initial_days

t_result = stats.ttest_ind(Group_ReAdmitted_Initial_days,
Group_Not_ReAdmitted_Initial_days)

alpha = 0.05

if (t_result[1] < alpha):
   print("Reject the null hypothesis. The difference between samples represents a
real difference between the two groups.")
else:
   print("Fail to reject the null hypothesis. There is no significant difference
between the two groups.")
```

Jasmine Cooper
D207
Data Cleaning
September 11, 2023
Western Governor's University

```
print("P-value:", t_result[1])
```

2. Two-sample t-test output:

```
Reject the null hypothesis. The difference between samples represents a real difference between the two gr
oups.
P-value: 0.0
```

3. I selected the two-sample t-test analysis technique because it assesses whether a statistically significant difference exists between the means of two independent groups. It compares the means of two groups when you have a continuous outcome variable, such as the duration of initial hospital stays. By utilizing the two-sample t-test, I can effectively evaluate whether I should reject the null hypothesis, that there is no statistical difference between the mean duration of hospital stays between patients who are readmitted and those who are not.

In text citation:
(“Student’s t-test”, n.d.)

Part C

1. Two continuous variables using univariate statistics:
    a. Total_charge: Upon visually inspecting the histogram, the distribution appears to be either bimodal or multimodal, however, since I can not rely on visual inspection alone, I have decided to use the summary statistics to help identify the distribution. Since, the mean ($5,312) and the median ($5,213) are relatively close in value, this suggests symmetry and could indicate that the Total Charge distribution is a normal distribution.
    b. Initial_days: Similarly, for the Initial days distribution, it is difficult to visually identify the distribution. Since the mean (34) and the median (36) are relatively close the Initial distribution also suggests symmetry and could indicate that the Initial days distribution is a normal distribution.
2. Two categorical variables using univariate statistics
    a. Readmis: By analyzing the mode of the readmission variable, it appears to be a skewed discrete distribution because the amount of patients that were not readmitted to the hospital is almost double the amount that were.
    b. Initial_admin: By examining the Initial admin variable, it also appears to be a skewed discrete distribution since the amount of patients that were admitted to the hospital for emergencies is approximately twice as high as patients admitted for elective or observation reasons.
3. See summary statistics, visualizations, and distributions on excel file attached (Univariate tab)

In-text citations:
(“Measures of center”, n.d.)
(“Measures of spread”, n.d.)
(“Continuous distributions”, n.d.)
(“Discrete distributions”, n.d.)
(Gudikandula,2018)

Jasmine Cooper
D207
Data Cleaning
September 11, 2023
Western Governor’s University

Part D

1. Two continuous variables using bivariate statistics
    a. Initial days vs. Total Charges: By analyzing the joint histogram, it is evident that the distribution is skewed. It conveys that as patients spend more time during their initial hospital stay that their total charge is higher. Examining the scatter plot, I can specifically identify that this is a positive right skewed distribution.
2. Two categorical variables using bivariate statistics
    a. Readmis vs. Initial admin: By examining the contingency table and the bar chart, it appears the distribution is skewed and discrete. The data conveys that patients who are admitted for emergencies are twice as likely to be readmitted than patients who are admitted for elective or observation reasons.
3. See summary statistics, visualizations, and distributions on excel file attached (Bivariate tab)

In-text citations:
("Measures of center", n.d.)
("Measures of spread", n.d.)
("Continuous distributions", n.d.)
("Discrete distributions", n.d.)
(Gudikandula,2018)

Part E

1. The result of the hypothesis suggests that there is a statistical difference between the mean duration of hospital stays between patients who are readmitted and those who are not. This could imply that the number of days a patient stays in the hospital initially affects whether they will be readmitted to the hospital. Although, the duration of the initial hospital stay does not determine the causation of a patients' potential readmission, this implication can be a starting point for root cause further investigation.
2. One of the limitations of my data analysis include outliers that involve reasons that are out of the patients control to be readmitted. Also, we are limited in granularity required to complete further investigations and draw conclusions that could help the hospital improve their reduction strategies.
3. Based on the results, a recommended course of action would involve conducting further investigations into additional variables. For example, it would be valuable to assess how readmission correlates with the initial reason for admission and risk of complications. This aims to determine whether the hospital should redefine what are less and/or more critical conditions. By adding more granularity to the data, the hospital can better assess the correlation between the duration of initial hospital stay and readmission. This may lead to the establishment of redefinition of reason for admissions and complication levels, and tighter intervals for patient discharge. Once the hospital implements this strategy, the team can monitor its effectiveness and decide whether further refinements or additional reduction strategies are necessary.

Part F

Panopto video recording

Jasmine Cooper
D207
Data Cleaning
September 11, 2023
Western Governor's University

Part G

Citations for code:

DataCamp. (n.d.). Student's t-test [Video file]. Retrieved from https://campus.datacamp.com/courses/experimental-design-in-python/the-basics-of-statistical-hypothesis-testing?ex=5

DataCamp. (n.d.). Measures of spread [Video file]. Retrieved from https://campus.datacamp.com/courses/introduction-to-statistics-in-python/summary-statistics-1?ex=7

DataCamp. (n.d.). Measures of center [Video file]. Retrieved from https://campus.datacamp.com/courses/introduction-to-statistics-in-python/summary-statistics-1?ex=4

Part H

Citations for content:

DataCamp. (n.d.). Discrete distributions [Video file]. Retrieved from https://campus.datacamp.com/courses/introduction-to-statistics-in-python/random-numbers-and-probability-2?ex=5

DataCamp. (n.d.). Continuous distributions [Video file]. Retrieved from https://campus.datacamp.com/courses/introduction-to-statistics-in-python/random-numbers-and-probability-2?ex=9

DataCamp. (n.d.). Measures of center [Video file]. Retrieved from https://campus.datacamp.com/courses/introduction-to-statistics-in-python/summary-statistics-1?ex=4

Gudikandula, P. (2018, November 9). Exploratory Data Analysis (beginner), Univariate, Bivariate and Multivariate – Habberman dataset. purnasaigudikandula.medium.com. Retrieved from https://purnasaigudikandula.medium.com/exploratory-data-analysis-beginner-univariate-bivariate-and-multivariate-habberman-dataset-2365264b751

Jasmine Cooper
D207
Data Cleaning
September 11, 2023
Western Governor's University