Statistical Intuitions and Applications

Assignment - 2

This assignment is targeted towards the applications of statistical intution across different disciplines which are of interdisciplinary in nature. To ease out the sampling step for the students, most of the datasets will generate once you run the codes and auto-save as csv files. You need to make sure following during submission:

1. All answers along with their codes should be submitted as **searchable PDF **.
2. Pictures or snapshots of your work will not be accepted.
3. All generated csv and .ipynb files must be submitted in a zip-folder as a secondary source.
4. Ensure the zip-folder has four csv files (Your Name, college_data, Height, Sustainability).
5. You may use Jupyter notebook or Colab as per your convenience.

Note: Reach out to your instructor for any question regarding csv files, codes or zip-folder.

```
# Following libraries will be loaded so that these can be applied in codes.

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
import scipy.stats as stats
```

A - Statistical Intuition in Medical Science

Question 1:

The data is sourced from the 2020 annual CDC survey of 400k US adults regarding their health status. The dataset is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to comprehensively collect health-related information from residents across the United States.

The code given below select a random sample of 300 respondents from the original data set which you need to use to analyze the data for those 300 respondents.

Use the Heart Disease dataset to answer the following questions:

a. Is having heart disease independent of whether the respondent is diabetic? Include the two-way table and all your calculations in your answer.

b. Is there a relationship between having heart disease and the general health condition of the respondent? Include the two-way table, stacked bar graph, and all your calculations in your answer.

c. Is there a relationship between having heart disease and the gender of the respondent? Include the two-way table, stacked bar graph, and all your calculations in your answer.

d. Does doing physical activity lower the risk of having heart disease? If so, by how much?

```
# This code gives you a random sample of 300 respondents to analyze and solve above questions.

# Replace Name with your name. The code will generate a csv file that you need to submit in the zip folder as secondary file.

try:
    college = pd.read_csv('Ayesha.csv')        # replace Name with your own name
except FileNotFoundError:
    original_data = pd.read_csv("https://raw.githubusercontent.com/zu-math/SIA-Fall-2023-Dataset/main/heart_2020_cleaned.csv")
    df1=original_data.sample(300)
    df1.to_csv('Ayesha.csv')                   # replace Name with your own name
    df = pd.read_csv('Ayesha.csv')             # replace Name with your own name
    df = pd.DataFrame(df)
    df.to_csv('Ayesha.csv')                    # replace Name with your own name

df.head()
```

	Unnamed: 0	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race
0	309438	No	25.80	Yes	No	No	0.0	0.0	No	Male	70-74	White
1	53386	No	17.63	No	No	No	4.0	7.0	No	Female	18-24	White
2	294677	No	21.29	Yes	No	No	0.0	15.0	No	Female	18-24	White
3	159902	No	28.89	No	No	No	3.0	0.0	Yes	Female	70-74	White
4	63651	No	26.61	Yes	No	No	0.0	0.0	No	Male	65-69	Asian

Answer: Add more "markdown" code cells below as per your need.

```
# Create a two-way table for Heart Disease and Diabetes
diabetes_heart_crosstab = pd.crosstab(df['Diabetic'], df['HeartDisease'])


# Perform the chi-squared test
from scipy.stats import chi2_contingency
chi2, p, _, _ = chi2_contingency(diabetes_heart_crosstab)


# Calculate the expected frequencies
expected_freq = pd.DataFrame(chi2_contingency(diabetes_heart_crosstab)[3])


# Display the results
print("Two-way table:")
print(diabetes_heart_crosstab)
print("\nChi-squared test:")
print("Chi-squared value:", chi2)
print("p-value:", p)
print("Expected frequencies:")
print(expected_freq)
```

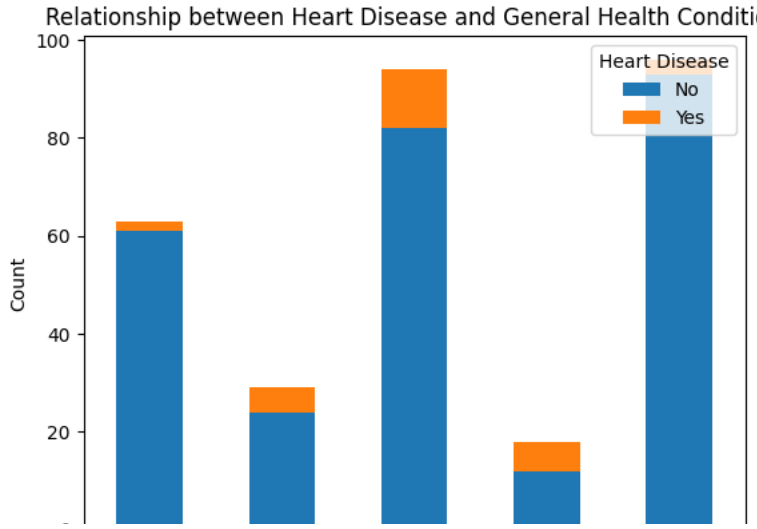
```
Two-way table:
HeartDisease              No  Yes
Diabetic
No                       224   15
No, borderline diabetes    4    1
Yes                       41   12
Yes (during pregnancy)     3    0

Chi-squared test:
Chi-squared value: 14.713279327678023
p-value: 0.00207881182084043
Expected frequencies:
              0          1
0  216.693333  22.306667
1    4.533333   0.466667
2   48.053333   4.946667
3    2.720000   0.280000
```

Based on data analyzed there is no strong relationship between diabetes and heart disease.

```
# Create a two-way table for Heart Disease and General Health
health_heart_crosstab = pd.crosstab(df['GenHealth'], df['HeartDisease'])


# Plot a stacked bar graph
health_heart_crosstab.plot(kind='bar', stacked=True)
plt.title("Relationship between Heart Disease and General Health Condition")
plt.xlabel("General Health Condition")
plt.ylabel("Count")
plt.xticks(rotation=0)  # To keep x-axis labels readable
plt.legend(title="Heart Disease", loc='upper right')
plt.show()
```




```
# Perform the chi-squared test
from scipy.stats import chi2_contingency
chi2, p, _, _ = chi2_contingency(health_heart_crosstab)


# Calculate the expected frequencies
expected_freq = pd.DataFrame(chi2_contingency(health_heart_crosstab)[3])


# Display the results
print("Two-way table:")
print(health_heart_crosstab)
print("\nChi-squared test:")
print("Chi-squared value:", chi2)
print("p-value:", p)
print("Expected frequencies:")
print(expected_freq)
```

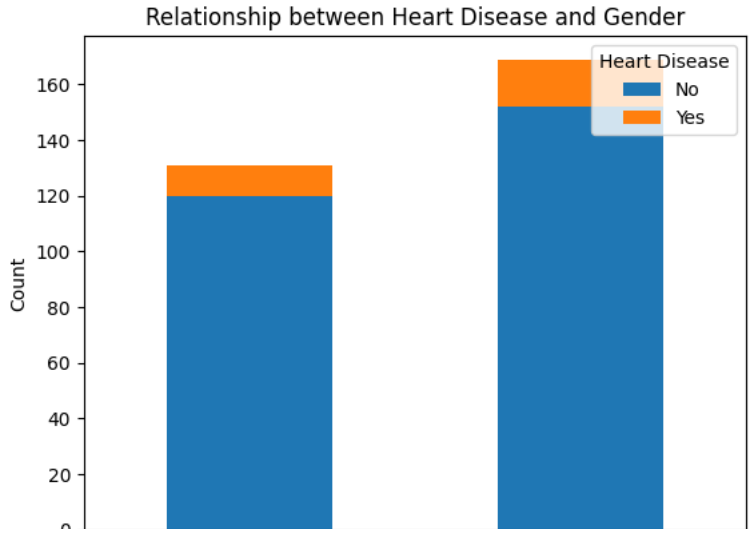
```
Two-way table:
HeartDisease  No  Yes
GenHealth
Excellent     61    2
Fair          24    5
Good          82   12
Poor          12    6
Very good     93    3

Chi-squared test:
Chi-squared value: 22.900516417590516
p-value: 0.00013255447745448792
Expected frequencies:
           0         1
0  57.120000  5.880000
1  26.293333  2.706667
2  85.226667  8.773333
3  16.320000  1.680000
4  87.040000  8.960000
```

From data we can conclude there is a strong relationship between general health and heart disease.

```
# Create a two-way table for Heart Disease and Gender
gender_heart_crosstab = pd.crosstab(df['Sex'], df['HeartDisease'])


# Plot a stacked bar graph
gender_heart_crosstab.plot(kind='bar', stacked=True)
plt.title("Relationship between Heart Disease and Gender")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.xticks(rotation=0)  # To keep x-axis labels readable
plt.legend(title="Heart Disease", loc='upper right')
plt.show()
```




```
# Perform the chi-squared test
from scipy.stats import chi2_contingency
chi2, p, _, _ = chi2_contingency(gender_heart_crosstab)


# Calculate the expected frequencies
expected_freq = pd.DataFrame(chi2_contingency(gender_heart_crosstab)[3])


# Display the results
print("Two-way table:")
print(gender_heart_crosstab)
print("\nChi-squared test:")
print("Chi-squared value:", chi2)
print("p-value:", p)
print("Expected frequencies:")
print(expected_freq)
```

```
Two-way table:
HeartDisease   No  Yes
Sex
Female        120   11
Male          152   17

Chi-squared test:
Chi-squared value: 0.08455694938890684
p-value: 0.7712143993738413
Expected frequencies:
              0          1
0  118.773333  12.226667
1  153.226667  15.773333
```

Based on data there is no significant relationship between gender and heart disease.

```
# Calculate the percentage of heart disease in active and inactive individuals
active_heart_disease = (df[df['PhysicalActivity'] == 'Yes']['HeartDisease'] == 'Yes').mean()
inactive_heart_disease = (df[df['PhysicalActivity'] == 'No']['HeartDisease'] == 'Yes').mean()


# Calculate risk reduction
risk_reduction = (inactive_heart_disease - active_heart_disease) / inactive_heart_disease * 100


print("Percentage of heart disease in active individuals:", active_heart_disease * 100, "%")
print("Percentage of heart disease in inactive individuals:", inactive_heart_disease * 100, "%")
print("Risk reduction:", risk_reduction, "%")
```

```
Percentage of heart disease in active individuals: 6.578947368421052 %
Percentage of heart disease in inactive individuals: 18.055555555555554 %
Risk reduction: 63.56275303643725 %
```

B - Statistical Intuition in Data Science

Question 2:

a. Compare the probability distributions of any three classes of your choice and identify the key differences. Your answers should have proper justification.

b. Explain how will you determine which class (or classes) has the lowest number of students who fail.

c. In a class (or classes) that has the highest number of students fail, which grade is the highest?

d. Explain how will you calculate the number of students with B grade in Class_2 and Class_4.

e. After comparing students with C grade in Class_1 and Class_4, Amna concluded that Class_4 has 5% more students with C grade than in Class_1. Provide an evidence if she is right or wrong.

```
# Following code is going to give you dataset of the five-classes in a college. The dataset will not change once you run the code

try:
    college = pd.read_csv('college_data.csv')
except FileNotFoundError:
    num_rows = 20
    num_columns = 5
    discrete_values = ['A', 'B', 'C', 'D', 'F']
    data = np.random.choice(discrete_values, size=(num_rows, num_columns))
    college = pd.DataFrame(data, columns=[f'Class_{i+1}' for i in range(num_columns)])
    college.to_csv('college_data.csv', index=False)
# Have a look at college dataset.
college.head()


# This code provides you probability distributions for each of the classes.

column_names = college.columns
num_colors = len(column_names)
colors = plt.cm.viridis(np.linspace(0, 1, num_colors))
fig, axes = plt.subplots(1, len(column_names), figsize=(15, 4))
for i, column_name in enumerate(column_names):
    data = college[column_name]
    probability_distribution = data.value_counts(normalize=True).sort_index()
    axes[i].bar(probability_distribution.index, probability_distribution, color=colors[i])
    axes[i].set_title(column_name)
    axes[i].set_xlabel("Grades Distribution")
    axes[i].set_ylabel("Probability")

plt.tight_layout()
plt.show()
```

Answer: Add more "markdown" code cells below as per your need.

C - Statistical Intuition in Health Sciences

Question 3:

Dubai Health Authority (DHA) contacted you to analyze a particular situation in the two hospitals: Dubai Hospital and Rashid Hospital which is to assess their performance in terms of resources allocation (Staffing, medical equipment, space etc). DHA provided you the most updated information available to them which can be obtained through the code below. It gives you the probability distributions of number of patients in these hospitals who have been diagnosed with a certain medical condition along with expected values and standard deviations.

a. How do you interpret the differences in both the mean and standard deviation when assessing the patient load and healthcare demands at the two hospitals?

b. How does this difference in expected patient counts affect resource allocation, such as staffing and medical supplies, for the two hospitals?

```
# This code gives you the number of patients diagonse with a disease in Dubai and Rashid hospitals. Once you made your mind
# to solve this question, only then run the code once. Perform your analysis on the values obtained.
```

```
s1=random.randint(8,12)
s2=random.randint(8,12)
num_patients_A = 100
patient_counts_A = np.random.poisson(s1, num_patients_A)
num_patients_B = 100
patient_counts_B = np.random.poisson(s2, num_patients_B)
Dubai_hospital = pd.DataFrame({'Patients_A': patient_counts_A})
Rashid_hospital = pd.DataFrame({'Patients_B': patient_counts_B})
expected_value_A = Dubai_hospital['Patients_A'].mean()
standard_deviation_A = Dubai_hospital['Patients_A'].std()
expected_value_B = Rashid_hospital ['Patients_B'].mean()
standard_deviation_B = Rashid_hospital ['Patients_B'].std()
probability_distribution_A = Dubai_hospital['Patients_A'].value_counts(normalize=True).sort_index()
probability_distribution_B = Rashid_hospital ['Patients_B'].value_counts(normalize=True).sort_index()
plt.figure(figsize=(10, 4))
plt.subplot(1, 2, 1)
plt.bar(probability_distribution_A.index, probability_distribution_A.values, width=0.5, align='center')
plt.xlabel("Number of Patients (Dubai_hospital)")
plt.ylabel("Probability")
plt.title("Probability Distribution - Dubai_hospital")
plt.subplot(1, 2, 2)
plt.bar(probability_distribution_B.index, probability_distribution_B.values, width=0.5, align='center')
plt.xlabel("Number of Patients (Rashid_hospital)")
plt.ylabel("Probability")
plt.title("Probability Distribution - Rashid_hospital ")
plt.tight_layout()
plt.show()

print("Dubai_hospital  - Expected Value (Mean):", expected_value_A)
print("Dubai_hospital  - Standard Deviation:", standard_deviation_A)


print("Rashid_hospital  - Expected Value (Mean):", expected_value_B)
print("Rashid_hospital - Standard Deviation:", standard_deviation_B)
```

Answer: Add more "markdown" code cells below as per your need.

D - Statistical Intuition in Social Innovation

Here we will consider a simple dataset from social sciences that records the height of individuals.

Question 4:

Consider a dataset from social sciences which is related to the heights of individuals in a population that often forms a normal distribution.

a. Explain how would you calculate the probability that a randomly selected individual has a height greater than 185 cm?

b. Explain how would you calculate the probability that a randomly selected individual has a height between 160 cm and 180 cm?

c. Calculate the 90th percentile of the height distribution. What does this value represent in the context of heights?

d. In a random sample of 49 individuals from this population, what is the probability that the sample mean height is within 1 cm of the population mean?

e. Suppose we want to find the height range that contains the middle 95% of the individuals in the population. Explain how do you obtain the lower and upper bounds of this range?

f. If you needed to cast a basketball team with the tallest players from this dataset, how many players would you need to ensure they win every game without jumping? Explain.

```
# In social sciences, one often comes across a dataset on the height of individuals.
# This dataset will be saved as csv file "Height" and you need to submit it in the same zip folder.

try:
    Social_science = pd.read_csv('Height.csv')
except FileNotFoundError:
```

```
    num_samples = 1000
    mean_height=random.randint(160,175)
    std_deviation = random.randint(4,12)
    heights = np.random.normal(mean_height, std_deviation, num_samples)
    heights = np.round(heights, 2)
    Social_science = pd.DataFrame({'Height': heights})
    Social_science.to_csv('Height.csv')
# Have a look at college dataset.
Social_science.head()
```

Answer: Add more "markdown" code cells below as per your need.

E - Statistical Intuition in Sustainability: COP-28

Imagine you have been invited to participate in the upcoming COP-28 which gives you an opportunity to contribute to the community using your statistical skills. You conducted a survey within your area and observed the recycling behavior in your community and recorded whether individuals in the community recycle ("1" for yes and "0" for no). You managed to get 100 records.

Question 5:

The answers to the following questions will provide you valuable insights that can be presented in **COP-28**.

a. What is the proportion of individuals in the community who recycle, based on the dataset?

b. Calculate a 95% confidence interval for the proportion of recyclers in the community. What does this interval tell us about the likely range of recycling behavior in the entire community?

c. If the 95% confidence interval for the proportion of recyclers is (0.55, 0.75), how would you interpret this interval? What level of confidence does it represent, and what can we conclude about the community's recycling behavior?

d. If a separate survey conducted earlier found that the recycling rate in the community was 70%, how does this rate compare to the point estimate obtained from the dataset, and what additional insights can you provide?

```
# This code gives you a community based dataset in Sustainability.
# This dataset will be saved as csv file "Sustainability" and you need to submit it in the same zip folder.

try:
    Sustainability = pd.read_csv('Sustainability.csv')
except FileNotFoundError:
    num_samples = 100
    recycling_behavior = np.random.choice([0, 1], size=num_samples)
    Sustainability = pd.DataFrame({'Recycles': recycling_behavior})
    Sustainability.to_csv('Sustainability.csv')

# Have a look at Sustainability dataset.
Sustainability.head()
```

Answer: Add more "markdown" code cells below as per your need.