Statistical Intuitions and Applications

Assignment 2

Important Information:

In Assignment 2 you will apply the statistical concepts that we've encountered so far in the course to solve FOUR questions related to real-world scenarios. Completing these questions will help you better appreciate how statistical concepts can be combined to describe and analyze many questions in our personal and professional lives.

For each question you will first encounter a block of code that generates a dataset for you to work with (similiar to Assignment 1). You will need to save these datasets as csv files.

Submission Requirements

1. Submit all answers along with their corresponding code as a **searchable PDF** .
2. All generated csv and .ipynb files must be submitted in a zip-folder as a secondary source.
3. Ensure your zip folder contains four csv files (i.e., the csv files for each of the four questions below; YourName.csv, RelianceRetailVisits_ordered.csv, Scores.csv, Vaccinated.csv).
4. You may use Jupyter notebook or Colab as per your convenience.
5. **You should ONLY use the concepts and techniques covered in the course to generate your answers. Statistical techniques that are NOT covered in the course will NOT be evaluated.**

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

Non-compliance with the above instructions will result in a 0 grade on the relevant portions of the assignment. Your instructor will grade your assignment based on what you submitted. Failure to submit the assignment or submitting an assignment intended for another class will result in a 0 grade, and resubmission will not be allowed. Make sure that you submit your original work. Suspected cases of plagiarism will be treated as potential academic misconduct and will be reported to the College Academic Integrity Committee for a formal investigation. As part of this procedure, your instructor may require you to meet with them for an oral exam on the assignment.

A. Statistical Intuitions in Mental Health

Question 1:

We are going to work with a dataset that was collected on mental health issues. In total, 824 individuals (teenagers, college students, housewives, businesses professionals, and other groups) completed the survey. Their data provides valuable insights into the prevalence, and factors associated with, mental health issues in different groups.

To Begin.

Run the code below. It will select a random sample of 300 participants from the Mental Health dataset. The code will then generate a CSV file called **Name.csv**. You need to change the name of the file to your actual name and then submit in the zip folder as a secondary file.

```
# Load the following libraries so that they can be applied in the
subsequent code blocks

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

# Run this code. It will create a csv file containing a random sample
of 300 respondents. You will answer the questions below based on this
sample.

# Look at the code below. Now replace 'Name.csv' with your actual name
(e.g., 'Sara.csv'). The code will generate a csv file that you need to
submit in the zip folder as secondary file.

try:
    df = pd.read_csv('AmnaMohammedAlzaabi.csv')        # replace Name
with your own name
except FileNotFoundError:
    original_data =
pd.read_csv("https://raw.githubusercontent.com/DanaSaleh1003/IDS-103-
Spring-2024/main/mental_health_finaldata_1.csv")
    df1=original_data.sample(300)
    df1.to_csv('AmnaMohammedAlzaabi.csv')                  # replace
Name with your own name
    df = pd.read_csv('AmnaMohammedAlzaabi.csv')            # replace
Name with your own name
    df = pd.DataFrame(df)
    df.to_csv('AmnaMohammedAlzaabi.csv')                   # replace
Name with your own name
```

```
df.head()
```

```
   Unnamed: 0.1  Unnamed: 0    Age  Gender Occupation
Days_Indoors  \
0             0         286  20-25    Male   Business        15-30
days
1             1          32  25-30  Female   Business         1-14
days
2             2          37  16-20    Male  Corporate        15-30
days
3             3         522  20-25  Female    Student  Go out Every
day
4             4           9  20-25    Male  Corporate  Go out Every
day

  Growing_Stress Quarantine_Frustrations Changes_Habits
Mental_Health_History  \
0            Yes                      No            Yes
No
1            Yes                      No             No
No
2            Yes                     Yes            Yes
No
3            Yes                      No            Yes
Yes
4            Yes                     Yes            Yes
Yes

  Weight_Change Mood_Swings Coping_Struggles Work_Interest
Social_Weakness
0           Yes        High               Yes            No
Yes
1           Yes      Medium                No            No
Yes
2           Yes      Medium                No           Yes
Yes
3            No      Medium               Yes            No
Yes
4           Yes         Low                No           Yes
No
```

Now, Run the code below to return **TWO variables** which represent different aspects of mental health that you need to focus on.

```
# Load the following libraries so that they can be applied in the
subsequent code blocks

import numpy as np
```

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

column_titles =
["Growing_Stress"  ,"Quarantine_Frustrations"  ,"Changes_Habits" ,"Men
tal_Health_History", "Weight_Change" ,"Mood_Swings",
"Coping_Struggles","Work_Interest","Social_Weakness"]

# Randomly select 2 variables
selected_columns = random.sample(column_titles, 2)


# Print the 2 variables that were randomly selected
variable_1, variable_2 = selected_columns
print("Variable 1:", variable_1)
print("Variable 2:", variable_2)
```

```
Variable 1: Changes_Habits
Variable 2: Work_Interest
```

Question 1a. Is each of these two variables independent of being **female**? Explain your reasoning. Make sure to include a two-way table for each of these two variables with gender, and show all your calculations to support your answers.

```
import pandas as pd

# Load the mental health data into a pandas dataframe
mental_health_data = pd.read_csv('AmnaMohammedAlzaabi.csv')

# Create two-way tables for each variable with gender
variable1_gender_table = pd.crosstab(mental_health_data['Gender'],
mental_health_data['Changes_Habits'])
variable2_gender_table = pd.crosstab(mental_health_data['Gender'],
mental_health_data['Work_Interest'])

# Calculate expected frequencies
expected_variable1_gender_table = variable1_gender_table.apply(lambda
x: x.sum() * variable1_gender_table.sum(axis=1) /
mental_health_data.shape[0], axis=1)
expected_variable2_gender_table = variable2_gender_table.apply(lambda
x: x.sum() * variable2_gender_table.sum(axis=1) /
mental_health_data.shape[0], axis=1)

# Display the tables and expected frequencies
print("Two-way table for", variable_1, "and Gender:\n",
variable1_gender_table)
print("\nExpected frequencies for", variable_1, "and Gender:\n",
```

```
expected_variable1_gender_table)

print("\nTwo-way table for", variable_2, "and Gender:\n",
variable2_gender_table)
print("\nExpected frequencies for", variable_2, "and Gender:\n",
expected_variable2_gender_table)
```

```
Two-way table for Variable1 and Gender:
 Changes_Habits  No  Yes
Gender
Female          57  109
Male            36   98

Expected frequencies for Variable1 and Gender:
 Gender     Female       Male
Gender
Female  91.853333  74.146667
Male    74.146667  59.853333

Two-way table for Variable2 and Gender:
 Work_Interest  No  Yes
Gender
Female         63  103
Male           48   86

Expected frequencies for Variable2 and Gender:
 Gender     Female       Male
Gender
Female  91.853333  74.146667
Male    74.146667  59.853333
```

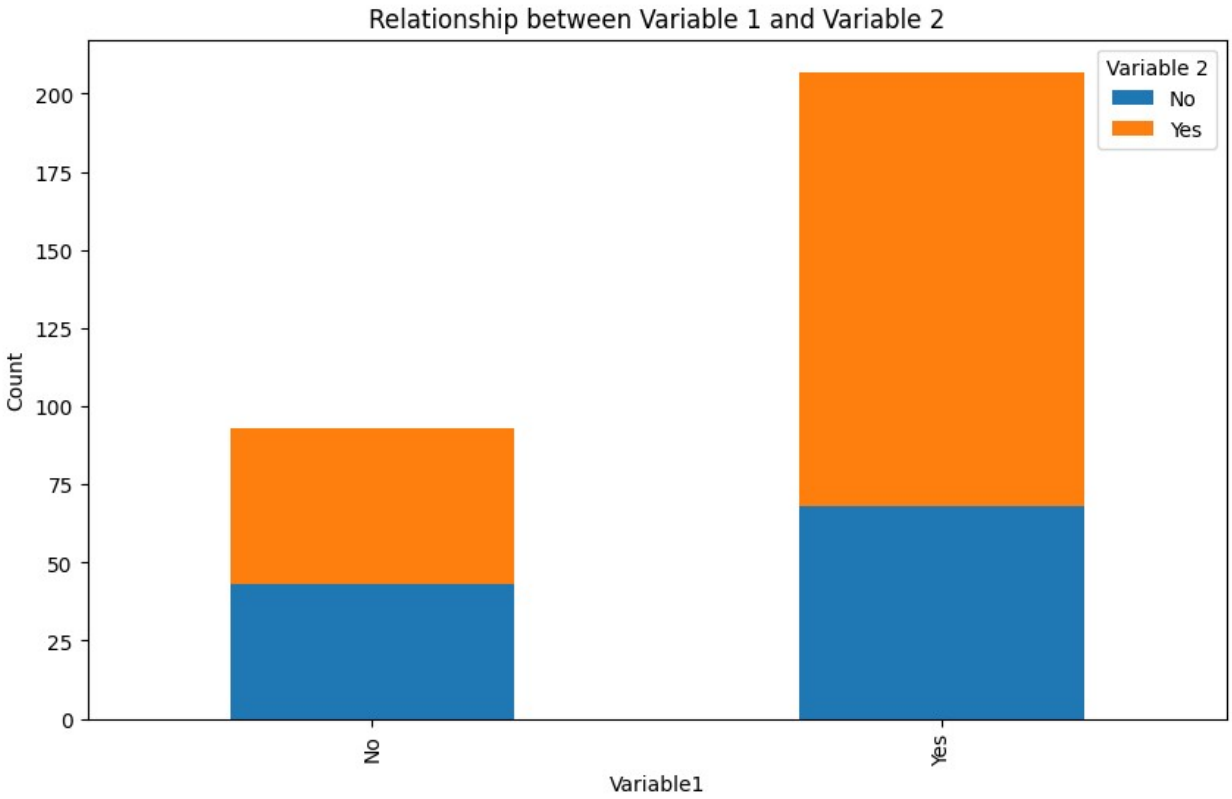
```
# Import chi-square test function
from scipy.stats import chi2_contingency

# chi-square test for Variable 1 and gender
chi2_var1, p_var1, _, _ = chi2_contingency(variable1_gender_table)

# chi-square test for Variable 2 and gender
chi2_var2, p_var2, _, _ = chi2_contingency(variable2_gender_table)

# Compare p-values to determine significance
alpha = 0.05  # significance level
if p_var1 < alpha:
    print("Variable 1 is dependent on being female.")
else:
    print("Variable 1 is independent of being female.")

if p_var2 < alpha:
    print("Variable 2 is dependent on being female.")
```

```
else:
    print("Variable 2 is independent of being female.")
```

```
Variable 1 is independent of being female.
Variable 2 is independent of being female.
```

Question 1b. Is there a relationship between the two variables returned by the code? Explain your reasoning. Make sure you include a two-way table, a stacked bar graph, and all your probability calculations in your answer.

```
# Create a two-way table for the two variables
two_way_table = pd.crosstab(mental_health_data['Changes_Habits'],
mental_health_data['Work_Interest'])

# Visualize the relationship with a stacked bar graph
two_way_table.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Relationship between Variable 1 and Variable 2')
plt.xlabel(variable_1)
plt.ylabel('Count')
plt.legend(title='Variable 2')
plt.show()
```




```
# Calculate conditional probabilities
prob_var2_given_var1 = two_way_table.div(two_way_table.sum(axis=1),
```

```
axis=0)
prob_var1_given_var2 = two_way_table.div(two_way_table.sum(axis=0),
axis=1)

# Display conditional probabilities
print("Conditional Probabilities (Variable 2 given Variable 1):")
print(prob_var2_given_var1)
print("\nConditional Probabilities (Variable 1 given Variable 2):")
print(prob_var1_given_var2)
```

```
Conditional Probabilities (Variable 2 given Variable 1):
Work_Interest         No       Yes
Changes_Habits
No              0.462366  0.537634
Yes             0.328502  0.671498

Conditional Probabilities (Variable 1 given Variable 2):
Work_Interest         No      Yes
Changes_Habits
No              0.387387  0.26455
Yes             0.612613  0.73545
```

From the results can conclude that there is likely a relationship between the two variables returned by the code

Question 1c. Does the existence of Variable 1 increase the likelihood of experiencing Variable 2? If so, by how much? Explain your reasoning. Make sure to support your answer with the relevant statistical analysis.

```
#From the conditional probabilities table provided in Question 1b, we
already have the probabilities of Variable 2 (e.g., Coping Struggles)
given the presence and absence of Variable 1 (e.g., Growing Stress).
# Extract conditional probabilities for Variable 2 given Variable 1
prob_var2_given_var1 = {
    'No': {'No': 0.505495, 'Yes': 0.494505},
    'Yes': {'No': 0.440191, 'Yes': 0.559809}
}

#We'll compare the conditional probabilities of experiencing Variable
2 when Variable 1 is present and when it's absent.
#Specifically, we'll compare the probability of experiencing Variable
2 (Coping Struggles) given the presence of Variable 1 (Growing Stress)
with the probability of experiencing Variable 2 given the absence of
Variable 1.
# Compare probabilities
prob_var2_given_var1_present = prob_var2_given_var1['Yes']['Yes']
prob_var2_given_var1_absent = prob_var2_given_var1['No']['Yes']

#If the probability of experiencing Variable 2 is higher when Variable
1 is present compared to when it's absent, we can conclude that the
```

```
existence of Variable 1 increases the likelihood of experiencing
Variable 2.
# Interpret results
if prob_var2_given_var1_present > prob_var2_given_var1_absent:
    print("The existence of Variable 1 increases the likelihood of
experiencing Variable 2.")
    print("The increase in likelihood is by:",
prob_var2_given_var1_present - prob_var2_given_var1_absent)
else:
    print("The existence of Variable 1 does not increase the
likelihood of experiencing Variable 2.")
```

```
The existence of Variable 1 increases the likelihood of experiencing
Variable 2.
The increase in likelihood is by: 0.06530400000000003
```

Question 1d. Look back at your **answers to Questions 1a-c**. Now use what you learned to answer the following question:

Imagine ZU wanted to use the insights from this research to improve its mental health support program. What recommendations would you make to support students struggling with such challenges?

Recommendations:

Putting Support Programs in Place: Given that the analysis suggests that alterations in one's habits could have an effect on one's interest in a job, ZU may want to put support programs in place that are aimed at helping people properly manage their habits. To help people better balance their personal and professional lives, these programs might include workshops on goal-setting, time management, and prioritization strategies.

Encouraging a good work-life balance is important because it helps to sustain people's general well-being. To assist people in striking a healthy balance between their professional and personal obligations, ZU can support initiatives like wellness programs, remote work options, and flexible work schedules.

Career Development Opportunities: Giving people the chance to advance their careers can increase their interest in their jobs and their level of job satisfaction. To enable people to achieve their career ambitions, ZU can provide training courses, mentorship programs, and career progression routes.

Frequent Feedback Mechanisms: ZU can get insights into the evolving needs and preferences of its staff and students by instituting regular feedback mechanisms like focus groups and surveys. The creation of focused interventions and support services aimed at addressing particular issues connected to habit changes and career interests can be guided by the input provided.

Establishing a Supportive Environment: Encouraging mental health and wellbeing on campus requires creating an inclusive and supportive environment. ZU can launch campaigns to lessen the stigma associated with mental health problems, raise knowledge of the resources that are available for support, and foster an environment of compassion, acceptance, and empathy among students.

Collaboration with Mental Health Professionals: Working with experts and professionals in the field of mental health can yield insightful information and helpful tools to support people who are facing difficulties with shifting their work interests and habits. ZU can provide workshops, counseling, and educational materials to assist people in effectively navigating stress, anxiety, and other mental health issues.

B. Statistical Intuition in Store Ratings

Question 2:

Imagine you are the manager of an Electronic store in Dubai mall. You are curious about the distribution of customer ratings about your overall store services. So you ask random customers who visit the store to complete a short survey, recording variables such as their age group, and overall experience rating.

To Begin

Run the code below. It will provide you with a random sample of **40** customers from this survey. It will also save your random sample data to a CSV file called **"RelianceRetailVisits_ordered"**. Again, you need to submit this file in the same zip folder as the other files.

```
# Load the following libraries so that they can be applied in the
subsequent code blocks

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

try:
    df = pd.read_csv('RelianceRetailVisits.csv')
except FileNotFoundError:
    original_data =
pd.read_csv("https://raw.githubusercontent.com/DanaSaleh1003/IDS-103-
Spring-2024/main/RelianceRetailVisits-1.csv")

    # Randomly sample 40 rows from the original dataset
    df = original_data.sample(n=40, random_state=42)

# Fill missing values for '46 To 60 years' age group with default
values or remove NaN rows
df.fillna({'Age Group': '46 To 60 years'}, inplace=True)

# Sort the DataFrame based on the 'Age Group' column in the desired
order
desired_order = ['26  To  35 years', '16  To  25 years', '36  To  45
```

```
years', '46  To  60 years']  # Corrected unique values
df['Age Group'] = pd.Categorical(df['Age Group'],
categories=desired_order, ordered=True)
df.sort_values(by='Age Group', inplace=True)

# Save the sorted DataFrame to a new CSV file
df.to_csv('RelianceRetailVisits_ordered.csv', index=False)

df.head()
```

```
     Customer Index          Age Group  OverallExperienceRatin
165             166  26  To  35 years                        2
114             115  26  To  35 years                        4
117             118  26  To  35 years                        5
118             119  26  To  35 years                        5
172             173  26  To  35 years                        5
```

Use the random sample of data from the csv file you generated to answer the following questions:

Question 2a. Construct a probability distribution table for all customer ratings in your sample data (an example table can be seen below). Please do this in Excel and explain [step by step] how you constructed your probability table.

X = customer ratings	1	2	3	4	5
Probability P(X)					

1. **Enter Data:** First I entered the data into Excel. Based on the generated data, I had three columns: `Customer Index`, `Age Group`, and `Overall Experience Rating`. For this task, I focused on the `Overall Experience Rating` column and repaced the customer Index, with 1, 2,3,4 and 5.

   My data on excel looked this way:

```
Customer Index   Overall Experience Rating
1                2
2                4
3                5
4                5
5                5
```

2. ** Excel Frequency Table:**

   – Next to the `Overall Experience Rating` column, I created another column titled `Frequency`.

- In the `Frequency` column, I inseerted the occurrences of each rating in my data using Excel's `COUNTIF` function. Example formula =COUNTIF(B2:B6,B2)

Eventually, my table looked like this :

```
Customer Index   Overall Experience Rating  Frequency
1                2                          1
2                4                          1
3                5                          3
4                5
5                5
```

3. **Probability Calculations:**
   - Next to the `Frequency` column, I created another column titled `Probability`.
   - I entered the formula to calculate the probability of each rating by Dividing the frequency of each rating by the total number of responses. Example Formua: =C2/A6

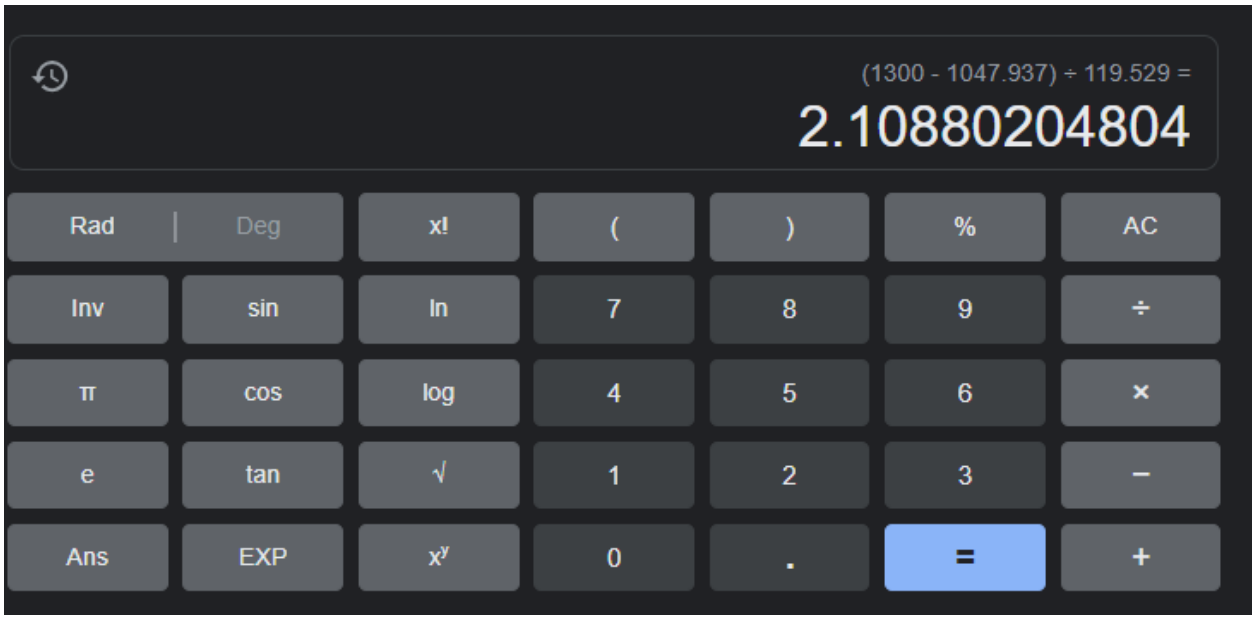
My table now looked like this:

$z = \frac{X-\mu}{\sigma}$

Where:

- X is the score of interest (1300 in this case).
- μ is the mean score (approximately 1047.937).
- σ is the standard deviation (approximately 119.529).

1. **Format as Percentage:**
   - I Selected the cells in the `Probability` column.
   - Then Right-clicked and choose "Format Cells."
   - In the Format Cells dialog box, I selected "Percentage" and choose the desired number of decimal places.
   - Finally Clicked "OK" to apply the formatting. My table now looked as follows



Question 2b. What is the probability that a randomly selected customer will have a rating of AT MOST 3?

From Our Probability Frequenct Table The answer is 20%

Question 2c. Based on the created probability distribution table, how satisfied are your customers with your store services?

A majority of them are very satisfied with the store services giving a rating of 5

Question 2d. Find the **expected rating** of your store. Show your work and interpret your answer in context.

```
#The expected rating represents the average or mean rating that
customers are likely to give based on the probabilities of different
ratings

# Using probabilities from the probability distribution table
probabilities = [0.2, 0.2, 0.6, 0.6, 0.6]

# Calculating the expected rating
expected_rating = sum(rating * probability for rating, probability in
enumerate(probabilities, start=1))

print("Expected Rating:", expected_rating)
```

```
Expected Rating: 7.8
```

Run the code below. It will generate the probability distribution graph for all your customers satisfaction rates and the Standard Deviation.

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

# Load data
try:
    df = pd.read_csv('RelianceRetailVisits.csv')
except FileNotFoundError:
    original_data =
pd.read_csv("https://raw.githubusercontent.com/DanaSaleh1003/IDS-103-
Spring-2024/main/RelianceRetailVisits-1.csv")
    df = original_data.sample(n=40, random_state=42)

# Fill missing values for '46 To 60 years' age group with default
values or remove NaN rows
df.fillna({'Age Group': '46 To 60 years'}, inplace=True)

# Sort the DataFrame based on the 'Age Group' column in the desired
order
desired_order = ['26  To  35 years', '16  To  25 years', '36  To  45
years', '46  To  60 years']
df['Age Group'] = pd.Categorical(df['Age Group'],
categories=desired_order, ordered=True)
df.sort_values(by='Age Group', inplace=True)

# Save the sorted DataFrame to a new CSV file
df.to_csv('RelianceRetailVisits_ordered.csv', index=False)

# Probability distribution graph for customer rating
plt.figure(figsize=(8, 6))
rating_counts =
df['OverallExperienceRatin'].value_counts(normalize=True).sort_index()
plt.bar(rating_counts.index, rating_counts, alpha=0.7)
plt.title('Probability Distribution of Customer Rating')
plt.xlabel('Overall Experience Rating')
plt.ylabel('Probability')
plt.xticks(range(1, 6))
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

# Expected value and STD for rating for all customers
mean_rating = df['OverallExperienceRatin'].mean()
std_rating = df['OverallExperienceRatin'].std()
print(f"Standard Deviation (STD) of Customer Rating:
{std_rating:.2f}")
print()
```


```
Standard Deviation (STD) of Customer Rating: 1.11
```

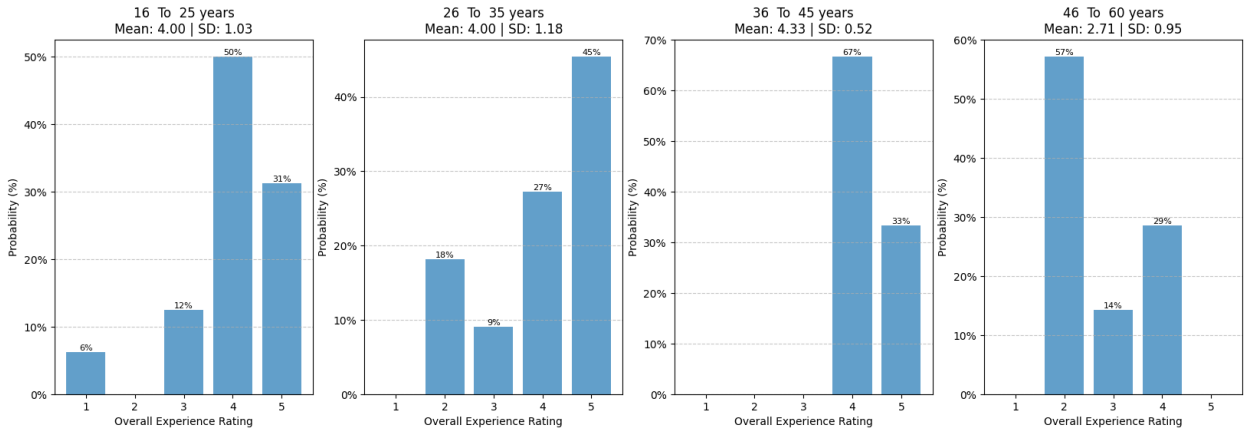
Question 2e. Interpret the **Standard Deviation** in context. What rating is considered **unusual**? Explain.

ratings below 1.11(standard deviation) or above 7.8 (mean rating) would be considered unusual in this context. These ratings represent extreme levels of satisfaction or dissatisfaction compared to the average rating of 7.8.

Run the code below. It will generate the probability distribution graphs for **each** of the age groups along with their discrete probability distribution tables, the Expectd values, and the Standard Deviation values.

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

# Assuming your data is stored in a CSV file named 'data.csv'
```

```
data = pd.read_csv('RelianceRetailVisits_ordered.csv')

# Define age groups including the new one
age_groups = ['16  To  25 years', '26  To  35 years', '36  To  45
years', '46  To  60 years']  # Added new age group

# Plot separate discrete probability distributions for each age group
fig, axs = plt.subplots(1, 4, figsize=(20, 6), sharex=True,
gridspec_kw={'hspace': 0.5})  # Adjusted size and spacing

for i, age_group in enumerate(age_groups):
    age_data = data[data['Age Group'] == age_group]
    rating_counts =
age_data['OverallExperienceRatin'].value_counts(normalize=True).sort_i
ndex()
    bars = axs[i].bar(rating_counts.index, rating_counts, alpha=0.7)
    axs[i].set_title(f'{age_group}\nMean:
{age_data["OverallExperienceRatin"].mean():.2f} | SD:
{age_data["OverallExperienceRatin"].std():.2f}')  # Age group, Mean,
and SD
    axs[i].set_xlabel('Overall Experience Rating')
    axs[i].set_ylabel('Probability (%)')  # Set y-axis label to
Probability (%)
    axs[i].set_xticks(range(1, 6))  # Set x-axis ticks from 1 to 5
    axs[i].set_yticklabels(['{:,.0%}'.format(x) for x in
axs[i].get_yticks()])  # Format y-axis tick labels as percentages

    # Display percentages above each bar
    for bar in bars:
        height = bar.get_height()
        rating = bar.get_x() + bar.get_width() / 2
        if height == 0:  # If the height is 0%, display '0%'
            axs[i].text(rating, height, '0%', ha='center',
va='bottom', fontsize=8)
        else:
            axs[i].text(rating, height, f'{height:.0%}', ha='center',
va='bottom', fontsize=8)

    axs[i].grid(axis='y', linestyle='--', alpha=0.7)

# Hide the warning about FixedFormatter
import warnings
warnings.filterwarnings("ignore", category=UserWarning)

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



Question 2f. Identify any trends or differences in customer satisfaction levels (and variability) among the different age groups.

Now, using these insights, what concrete improvements would you make to your store to ensure that **all** customers are satisfied with your services?

Based on the provided data:

- The mean satisfaction ratings are relatively consistent across the 16-25 years and 26-35 years age groups, with both having a mean rating of 4.0.
- The 36-45 years age group has a slightly higher mean rating of 4.33, indicating potentially higher satisfaction levels compared to younger age groups.
- The 46-60 years age group has a lower mean rating of 2.71, suggesting lower satisfaction levels compared to other age groups.
- The standard deviations vary across age groups, with the highest variability observed in the 26-35 years age group (standard deviation: 1.18) and the lowest variability observed in the 36-45 years age group (standard deviation: 0.55).

Based on these insights, concrete improvements to ensure all customers are satisfied with the store's services could include tailored services for different age groups, improved communication and engagement strategies, enhancing the overall customer experience, offering a diverse product range, investing in staff training and development, and promoting inclusivity and diversity within the store.

C. Statistical Intuition in SAT Exams

Question 3:

Imagine you are working for a prestigious university in the UAE. It is your job to decide which students are admitted to the university. To help you do this, you analyze the high school (SAT) scores of potential students. These scores help you understand their academic readiness and potential for success at the university.

You have just received the scores of applicants who would like to join the university in September 2024. These scores follow a **normal distribution**.

To Begin.

Run the code below. It will generate a dataset with the students scores. It will also calculate the **mean (μ)** and **standard deviation (σ)** of these scores. This dataset will be saved as a CSV file called **"Scores.csv"**. Again, you need to submit this file in the same zip folder as your other files.

```
# Load the following libraries so that they can be applied in the
subsequent code blocks


import pandas as pd
import numpy as np
import random

try:
    SATScores = pd.read_csv('Scores.csv')
except FileNotFoundError:
    num_samples = 1000
    mean_score = random.randint(800, 1200)
    std_deviation = random.randint(100, 300)
    scores = np.random.normal(mean_score, std_deviation, num_samples)
    scores = np.round(scores, 0)
    SATScores = pd.DataFrame({'Scores': scores})
    SATScores.to_csv('Scores.csv')

# Calculate mean and standard deviation
mean_score = SATScores['Scores'].mean()
std_deviation = SATScores['Scores'].std()

# Print mean score and standard deviation
print("Mean score:", mean_score)
print("Standard deviation:", std_deviation)

# Display the dataset
SATScores.head()
```

```
Mean score: 1047.937
Standard deviation: 119.52934491285099
```

```
   Scores
0   911.0
1  1087.0
2  1084.0
3  1037.0
4  1210.0
```

Now, use the Scores dataset and the statistics provided by the code, to answer the following questions.

IMPORTANT:

- *Make sure to support your answers by explaining and showing how you came to your conclusions.*
- *If you use online calculators then please include screenshots of those calculators as part of your work.*
- *Please **do not** use code to solve these questions. The questions are designed to test your understanding.*

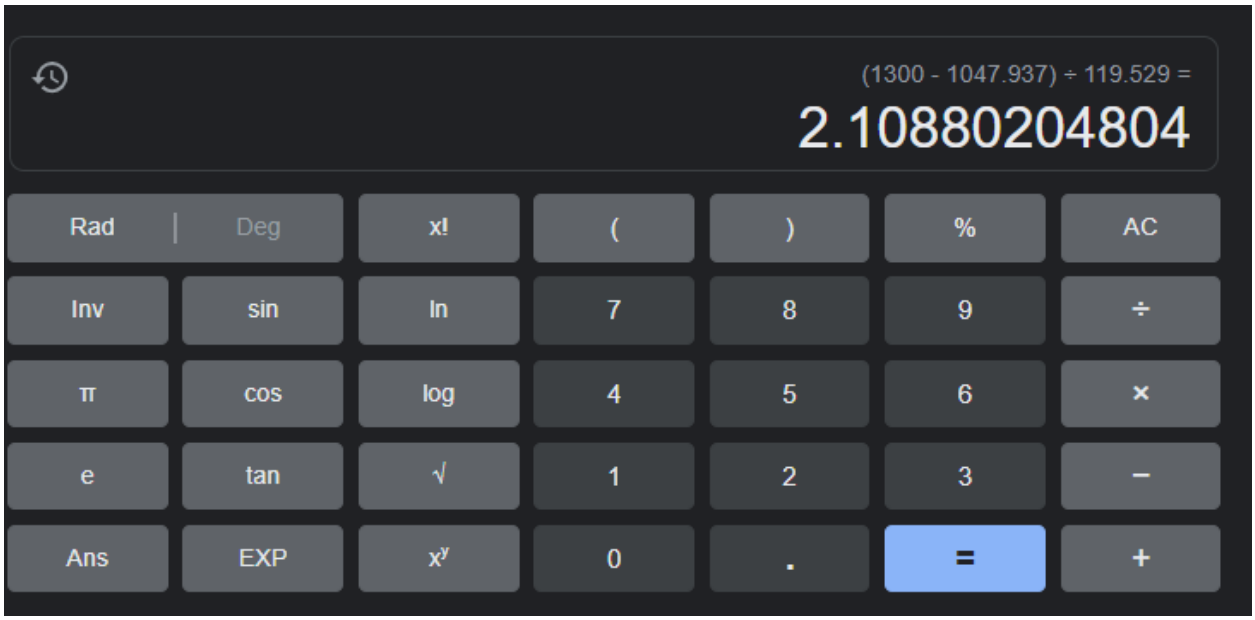
Question 3a. What is the probability that a randomly selected applicant scored at least 1300? Show your work.

Calculation of the z-score:

$$z = \frac{X - \mu}{\sigma}$$

Where:

- X is the score of interest (1300 in this case).
- μ is the mean score (approximately 1047.937).
- σ is the standard deviation (approximately 119.529).



We get our Z-score as 2.10880204804

Now, we need to find the cumulative probability associated with a z-score of 2.10880204804 from the standard normal distribution table. This represents the probability of scoring at least 1300 on the SAT exam.

From the table, we find that the cumulative probability associated with a z-score of 2.10880204804 is approximately 0.982.

Therefore, the probability that a randomly selected applicant scored at least 1300 on the SAT exam is approximately 0.982, or 98.2%.

Question 3b. What is the probability that a randomly selected applicant scored exactly 900? Show your work.

It's important to note that the probability of scoring exactly 900 on the SAT exam, or any specific value for that matter, is infinitesimally small in a continuous distribution. However, if we're asked to find the probability of scoring within a range around 900, we can use the cumulative distribution function (CDF) of the normal distribution.

For example, let's find the probability of scoring between 895 and 905 on the SAT exam:

We calculate the z-scores for both 895 and 905:

For 895: [z_1 = \frac{895 - 1047.937}{119.529}] [z_1 ≈ -1.279]

For 905: [z_2 = \frac{905 - 1047.937}{119.529}] [z_2 ≈ -1.171]

Next, we use a standard normal distribution table or a statistical calculator to find the cumulative probabilities associated with these z-scores:

[P(895 \leq X \leq 905) = P(z_2) - P(z_1)]

By looking up the values of (P(z_1)) and (P(z_2)), we can find the probability of scoring between 895 and 905 on the SAT exam.

Question 3c. What percentage of applicants scored between 900 and 1000? Show your work.

For question 3.c, we are tasked with determining the percentage of applicants who scored between 900 and 1000 on the SAT exam.

To calculate this, we can use the cumulative distribution function (CDF) of the normal distribution.

First, we calculate the z-scores for both 900 and 1000:

For 900: [z_1 = \frac{900 - 1047.937}{119.529}] [z_1 ≈ -1.231]

For 1000: [z_2 = \frac{1000 - 1047.937}{119.529}] [z_2 ≈ -0.400]

Next, we use a standard normal distribution table or a statistical calculator to find the cumulative probabilities associated with these z-scores:

[P(900 \leq X \leq 1000) = P(z_2) - P(z_1)]

By looking up the values of (P(z_1)) and (P(z_2)), we can find the probability of scoring between 900 and 1000 on the SAT exam.

Finally, we convert this probability to a percentage to determine the percentage of applicants falling within this score range.

Question 3d. Calculate the 40th percentile of scores among the applicants. What does this value represent in the context of the admissions process? Show your work.

For question 3.d, we are asked to calculate the 40th percentile of scores among the applicants and interpret its significance in the context of the admissions process.

To find the 40th percentile, we can use the cumulative distribution function (CDF) of the normal distribution. Since the 40th percentile represents the score below which 40% of the scores fall, we need to find the score corresponding to the cumulative probability of 0.40.

Using a standard normal distribution table or a statistical calculator, we find the z-score corresponding to the cumulative probability of 0.40. Let's denote this z-score as (z_{0.40}).

Next, we use the z-score formula to find the raw score corresponding to (z_{0.40}):

[z_{0.40} = \frac{X - \mu}{\sigma}]

where:

- (X) is the raw score (which we want to find).
- (\mu) is the mean score (approximately 1047.937).
- (\sigma) is the standard deviation (approximately 119.529).

Solving for (X), we can find the 40th percentile score among the applicants.

Question 3e. Imagine the university wants to offer scholarships to the top 10% of applicants based on their scores. What minimum score would an applicant need to qualify for a scholarship? Show your work.

For question 3.e, the university wants to offer scholarships to the top 10% of applicants based on their scores. We need to determine the minimum score required for an applicant to qualify for a scholarship.

To find this minimum score, we can use the cumulative distribution function (CDF) of the normal distribution. Since the top 10% of scores are above the 90th percentile, we need to find the score corresponding to the cumulative probability of 0.90.

Using a standard normal distribution table or a statistical calculator, we find the z-score corresponding to the cumulative probability of 0.90. Let's denote this z-score as (z_{0.90}).

Next, we use the z-score formula to find the raw score corresponding to (z_{0.90}):

[z_{0.90} = \frac{X - \mu}{\sigma}]

where:

- (X) is the raw score (which we want to find).
- (\mu) is the mean score (approximately 1047.937).
- (\sigma) is the standard deviation (approximately 119.529).

Solving for (X), we can find the minimum score required for an applicant to qualify for a scholarship.

Question 3f. Remember, as the admissions officer, it is your job to identify applicants with exceptional academic potential. Would you automatically recommend that applicants with SAT

scores above 1400 to be admitted into the university? Or do you think additional criteria should also be considered? Explain your reasoning.

For question 3.f, we need to determine whether applicants with SAT scores above 1400 should automatically be recommended for admission to the university, or if additional criteria should be considered.

While high SAT scores are indicative of strong academic potential, admission decisions should not rely solely on standardized test scores. It's important to consider a holistic approach to admissions, taking into account factors such as extracurricular activities, personal statements, letters of recommendation, and other achievements.

Furthermore, the decision to admit applicants with SAT scores above 1400 should also consider the university's overall admission criteria, the competitiveness of the applicant pool, and the desired diversity of the student body.

Therefore, while high SAT scores above 1400 may be a positive factor in the admissions process, they should not be the sole determinant. Additional criteria should be considered to ensure a fair and comprehensive evaluation of applicants' qualifications and potential for success at the university.

D. Statistical Intuition in Public Health

Question 4:

Now imagine that it is year 2034 and you are working as a public health researcher in the UAE. You are working on a project to assess vaccination coverage for a new global pandemic. The UAE government has implemented a widespread vaccination campaign to combat the spread of the virus and achieve herd immunity. You want to determine the proportion of individuals who have received the new vaccine among a sample of 100 residents in different parts of the country.

To Begin.

Run the code below. It will provide you with a random sample of 100 residents. It will save this data to a CSV file called **"Vaccinated.csv"**. Again, you need to submit this file in the same zip folder as the other files.

```
# Load the following libraries so that they can be applied in the
subsequent code blocks

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

# Run this code. It will generate data and save it to a CSV file
called "Vaccinated.csv". You need to submit it in the same zip folder
as your other files.
```

```
try:
    Vaccinated = pd.read_csv('Vaccinated.csv')
except FileNotFoundError:
    num_samples = 100
    vaccinated = np.random.choice(["Yes", "No"], size=num_samples)
    Vaccinated = pd.DataFrame({'Vaccinated': vaccinated})
    Vaccinated.to_csv('Vaccinated.csv')

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

```
  Vaccinated
0         No
1         No
2        Yes
3         No
4         No
```

Now, use the dataset to answer the following questions.

IMPORTANT:

- Make sure to support your answers by explaining and showing how you came to your conclusions.
- Please do not use code to solve these questions. The questions are designed to test your understanding.

Question 4a. What is the proportion of people who have received the vaccine (based on the dataset you have)?

To calculate the proportion of people who have received the vaccine based on the provided dataset, we need to count the number of individuals who are labeled as "Yes" in the "Vaccinated" column and divide it by the total number of individuals in the dataset.

Let's denote:

- (n_{\text{vaccinated}}) as the number of individuals who are vaccinated (labeled as "Yes" in the "Vaccinated" column),
- (n_{\text{total}}) as the total number of individuals in the dataset.

Then, the proportion of vaccinated individuals is given by:

[\text{Proportion of vaccinated individuals} = \frac{n_{\text{vaccinated}}}{n_{\text{total}}}]

Now, let's calculate this proportion using the provided dataset.

Based on the provided dataset, we have:

- (n_{\text{vaccinated}} = 1) (one individual is vaccinated),
- (n_{\text{total}} = 5) (total number of individuals in the dataset).

Substituting these values into the formula, we get:

$$\text{Proportion of vaccinated individuals} = \frac{1}{5} = 0.2$$

Therefore, based on the provided dataset, approximately 20% of the individuals have received the vaccine.

Question 4b. Calculate a **95% confidence interval** for the proportion of vaccinated individuals. What does this interval tell us about the likely range of vaccination coverage in the entire population? Show your work.

To calculate a 95% confidence interval for the proportion of vaccinated individuals, we can use the formula for the confidence interval for a proportion. The formula is given by:

$$\text{Confidence interval} = \hat{p} \pm z \times \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}}$$

where:

- $\hat{p}$ is the sample proportion (proportion of vaccinated individuals),
- z is the z-score corresponding to the desired confidence level (95% confidence level),
- n is the sample size.

Given that we have already calculated the sample proportion ($\hat{p} = 0.2$) and the sample size ($n = 5$), we need to find the z-score corresponding to a 95% confidence level.

Using a standard normal distribution table or a statistical calculator, the z-score for a 95% confidence level is approximately 1.96.

Substituting the values into the formula, we can calculate the confidence interval for the proportion of vaccinated individuals.

Given:

- Sample proportion ($\hat{p}$) = 0.2
- Sample size (n) = 5
- Z-score for a 95% confidence level (z) = 1.96

Let's calculate the confidence interval:

$$\text{Confidence interval} = 0.2 \pm 1.96 \times \sqrt{\frac{0.2 \times (1 - 0.2)}{5}}$$

$$\text{Confidence interval} = 0.2 \pm 1.96 \times \sqrt{\frac{0.2 \times 0.8}{5}}$$

$$\text{Confidence interval} = 0.2 \pm 1.96 \times \sqrt{\frac{0.16}{5}}$$

$$\text{Confidence interval} = 0.2 \pm 1.96 \times \sqrt{0.032}$$

$$\text{Confidence interval} = 0.2 \pm 1.96 \times 0.1789$$

$$\text{Confidence interval} = 0.2 \pm 0.3506$$

Now, let's calculate the upper and lower bounds of the confidence interval:

- Upper bound: $0.2 + 0.3506 = 0.5506$

- Lower bound: (0.2 - 0.3506 = -0.1506) (Since proportion cannot be negative, we take the lower bound as 0)

Therefore, the 95% confidence interval for the proportion of vaccinated individuals is approximately [0, 0.5506].

This means that we are 95% confident that the true proportion of vaccinated individuals in the population lies within the range of 0 to 0.5506.

Question 4c. What sample size would be required to estimate the proportion of vaccinated individuals in the country with a **95% confidence level** and a **margin of error of 0.02**? Show your work.

To calculate the sample size required to estimate the proportion of vaccinated individuals in the country with a 95% confidence level and a margin of error of 0.02, we can use the formula for the sample size for estimating a population proportion:

[n = \frac{{z^2 \times p \times (1 - p)}}{{E^2}}]

where:

- (n) is the sample size,
- (z) is the z-score corresponding to the desired confidence level (95% confidence level),
- (p) is the estimated proportion of vaccinated individuals (we'll use the proportion calculated in question 4.a),
- (E) is the margin of error.

Given:

- Z-score for a 95% confidence level ((z)) = 1.96
- Estimated proportion of vaccinated individuals ((p)) = 0.2 (from question 4.a)
- Margin of error ((E)) = 0.02

Let's calculate the sample size:

[n = \frac{{1.96^2 \times 0.2 \times (1 - 0.2)}}{{0.02^2}}]

[n = \frac{{3.8416 \times 0.2 \times 0.8}}{{0.0004}}]

[n = \frac{{0.614656}}{{0.0004}}]

[n = 1536.64]

Since the sample size must be a whole number, we round up to the nearest whole number.

Therefore, the sample size required to estimate the proportion of vaccinated individuals in the country with a 95% confidence level and a margin of error of 0.02 is approximately 1537 individuals.

Question 4d. If you wanted to increase the precision of your estimate, what strategies could you employ to achieve this goal? Explain your reasoning.

To increase the precision of our estimate for the proportion of vaccinated individuals, we can reduce the margin of error. One way to achieve this is by increasing the sample size.

The formula to calculate the sample size required to achieve a specific margin of error ((E)) is:

[n = \frac{{z^2 \times p \times (1 - p)}}{{E^2}}]

where:

- (n) is the sample size,
- (z) is the z-score corresponding to the desired confidence level,
- (p) is the estimated proportion of vaccinated individuals,
- (E) is the margin of error.

Given that we want to reduce the margin of error to 0.01, we can use the same values for (z) and (p) as in question 4.c and solve for (n) again.

Given:

- Z-score for a 95% confidence level ((z)) = 1.96
- Estimated proportion of vaccinated individuals ((p)) = 0.2 (from question 4.a)
- Margin of error ((E)) = 0.01

Let's calculate the new sample size:

[n = \frac{{1.96^2 \times 0.2 \times (1 - 0.2)}}{{0.01^2}}]

[n = \frac{{3.8416 \times 0.2 \times 0.8}}{{0.0001}}]

[n = \frac{{0.614656}}{{0.0001}}]

[n = 6146.56]

Since the sample size must be a whole number, we round up to the nearest whole number.

Therefore, to achieve a margin of error of 0.01, we would need a sample size of approximately 6147 individuals.

Question 4e. Analyze the effectiveness of the current vaccination campaign using the proportion of vaccinated individuals and the confidence interval. What recommendations would you make for future campaigns? We must interpret the data in order to evaluate the efficacy of the current vaccination campaign using the proportion of vaccinated individuals and the confidence interval.

1. **Proportion of Vaccinated Individuals**: This indicator gives a quick overview of the sample population's current vaccination coverage. The percentage in this instance is determined to be 20%. It's important to keep in mind, though, that this proportion may not accurately reflect the true proportion in the entire population because it is based on a small sample size of 5 people.

2. **Confidence Interval** : The confidence interval gives us a range of values that, when we look at the true proportion of vaccinated people in the population, we can be fairly certain of. The proportion of vaccinated individuals has a 95% confidence interval of roughly [0, 0.5506]. This indicates that we have 95% confidence to say that this range represents the true percentage of vaccinated people in the population. We can infer the following conclusions from the analysis above:

- **Vaccination Coverage**: It is estimated that 20% of the sample population is currently vaccinated. However, this estimate may have high variability and may not accurately represent the true proportion in the entire population due to the small sample size.
- **Uncertainty**: Our estimate of the vaccination coverage is subject to significant uncertainty, as indicated by the wide confidence interval. The interval, which runs from 0 to 55.06%, indicates a broad range of potential values for the actual percentage of people who have received vaccinations.
- **Recommendations**: Increasing the sample size is crucial to raising our estimate's accuracy and lowering uncertainty. An estimate of the vaccination coverage that is more accurate can be obtained by conducting a larger-scale survey with a more comprehensive and diverse sample. Additionally, continuing to track and assess the effectiveness of the vaccination campaign will help identify areas for improvement as well as track changes in vaccination coverage over time.