CMSC320_HW2_Spring2025

March 9, 2025

1 HOMEWORK 2: BRAIN CONDITIONING STATS

1.1 DUE: ***March 04, 2025 @ 11:59 PM***

1.2 24-HR LATE DUE DATE WITH A 15% PENALTY: ***March 5, 2025 @ 11:59 PM***

Objective:

The aim of this assignment is to deepen students' understanding of statistics and hypothesis testing using Python. By engaging with some theortical questions as well as practical exercises, students will apply statistical methods and perform hypothesis tests, using Python to code and execute these techniques. This approach will help solidify their grasp of statistical principles and their application in Python, bridging theoretical knowledge with practical skills.

1.2.1 Reminder: Please make sure your code runs before submitting your work. Code sections that do not run will receive 0 credits, no partials will be given. This is VERY important in real project development.

1.2.2 DO NOT REMOVE ANY PART OF ANY OF THE QUESTIONS OR YOU LOSE CREDIT

1.2.3 ***No Hardcoding either***

2 Part 1: Statistics Problem Solving

##Q1) (10 POINTS) Bayes Theorem

Suppose some hacker found a dataset on uselessdatasets.com containing information about three different types of users on an online platform: "bloggers", "shoppers", and "reviewers". The data has 10,000 users. There are 4,500 bloggers, 6,000 shoppers, and 5,500 reviewers. The users could be in multiple categories. 2,000 of the bloggers are shoppers, 1,800 of the bloggers are reviewers, and 3,000 shoppers are also reviewers.

Answer the following questions:

1. (3 POINTS) If X is a random variable that represents the users that were cross listed into all 3 categories, what is the value of X? (Hint: think of a Venn Diagram.)

   Using the inclusion-exclusion principle: 4500 + 6000 + 5500 - 2000 - 1,800 - 3,000 + X = 10,000 -> 9200 + X = 10000 -> X = 800

2. (3 POINTS) Calculate the probability that a randomly selected shopper is also a reviewer. (Hint: Use Bayes Theorem)

   P(R|S) = P(R S) / P(S)

   P(S) = |S|/10000 = 6000/10000 = 0.6

   P(R S) = 3000/10000 = 0.3

   P(R S) / P(S) = 0.3 / 0.6 = 0.5

   Thus, P(R|S) = 0.5

3. (4 POINTS) Calculate the probability that a random user is in exactly two categories but not all three.

   Users in B S but not in R = |B S| - |B S R| = 2,000 - 800 = 1200.

   Users in B R but not in S = |B R| - |B S R| = 1,800 - 800 = 1000.

   Users in R S but not in B = |R S| - |B S R| = 3,000 - 800 = 2200.

   1200 + 1000 + 2200 = 4400.

   The probability that a random user is in two categories but not all three would be 4400 / 10000 = 0.44 or 44%.

##Q2) (6 POINTS) Expected Values

Let T be the set of all sequences of two rolls of a dice. Let S be the set of all sequences of three rolls of a dice. Let X_n be the sum of the number of dots on n dice rolls.

Answer the following question:

1. (3 POINTS) What is $\mathbb{E}[X_2]$?

   $\mathbb{E}[X_1]$ = 1+2+3+4+5+6 / 6 = 21 / 6 = 3.5

   Since we know that these rolls are independent from each other, $\mathbb{E}[X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_1]$ = 3.5 + 3.5 = 7

2. (3 POINTS) What is $\mathbb{E}[X_3]$?

   Similar to the previous question, the rolls are independent of each other, therefore,

   $\mathbb{E}[X_3] = \mathbb{E}[X_1] + \mathbb{E}[X_1] + \mathbb{E}[X_1]$ = 3.5 + 3.5 + 3.5 = 10.5

##Q3) (6 POINTS) Probability distribution

Let X be a continuous random variable that follows a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 2$.

Answer the following question:

1. (3 POINTS) What is the probability that X takes a value between 8 and 12? Hints: You may have to utilize the standard normal table: https://math.arizona.edu/~jwatkins/normal-table.pdf

   How to read the "Standard Normal Cumulative Probability Table" table:

- Rows and Columns: The rows correspond to the first digit and first decimal place of z. The columns correspond to the second decimal place of z.
- Check out: https://byjus.com/maths/z-score-table/

  For X = 8, we get Z = 8 - 10/ 2 = -1

  For X = 12, we get Z = 12 - 10 / 2 = 1

  From the normal distribution table, P(8 <= x <= 12) = P(Z <= 1) - P(Z <= -1) = 0.8413 - 0.1587 = 0.6826

2. (3 POINTS) What is the probability that X takes a value greater than 14?

   For X = 14, we get Z = 14 - 10 / 2 = 2

   From the normal distribution table, P(Z > 2) = 1 - P(Z <= 2) = 1 - 0.9772 = 0.0228

3 Part 2: Python Warmups

##Q1) (10 POINTS) Bernoulli Trials

Consider a sequence of n Bernoulli trials with success probability p per trial. A string of consecutive successes is known as a *streak*.

> **Task to do:** Write a function that returns a `collections.Counter` that maps the length of a streak k to the number of times it is observed in an input sequence `xs`. For example, if `xs = [0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1]`, the output would be `Counter({1: 2, 2: 1, 3: 2})`. We have imported `Counter` from the Python `collections` library for you in the code block below.

[123]:

```python
from collections import Counter

def count_streaks(xs):
    """Count number of success runs of length k."""
    ys = []
    cnt = 0

    for i in xs:
      if i == 1:
        cnt += 1
      else:
        if cnt > 0:
          ys.append(cnt)
        cnt = 0

    if cnt > 0:
      ys.append(cnt)

    return Counter(ys)
```

[124]:

```
# Use this cell to test your answer. MAKE SURE YOUR RESULTS ARE SHOWN BELOW AFTER RUNNING THIS BOX
import numpy as np
print(count_streaks([0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1]))
np.random.seed(0)
count_streaks(np.random.randint(0,2,1000000))
```

```
Counter({1: 2, 3: 2, 2: 1})
```

[124]:

```
Counter({1: 125036,
         2: 62589,
         3: 31100,
         4: 15859,
         5: 7699,
         6: 3893,
         7: 1921,
         8: 946,
         9: 470,
         10: 245,
         11: 126,
         12: 45,
         13: 29,
         14: 11,
         15: 9,
         17: 6,
         16: 2,
         18: 1})
```

##Q2) (10 POINTS) Distribution and Visualization

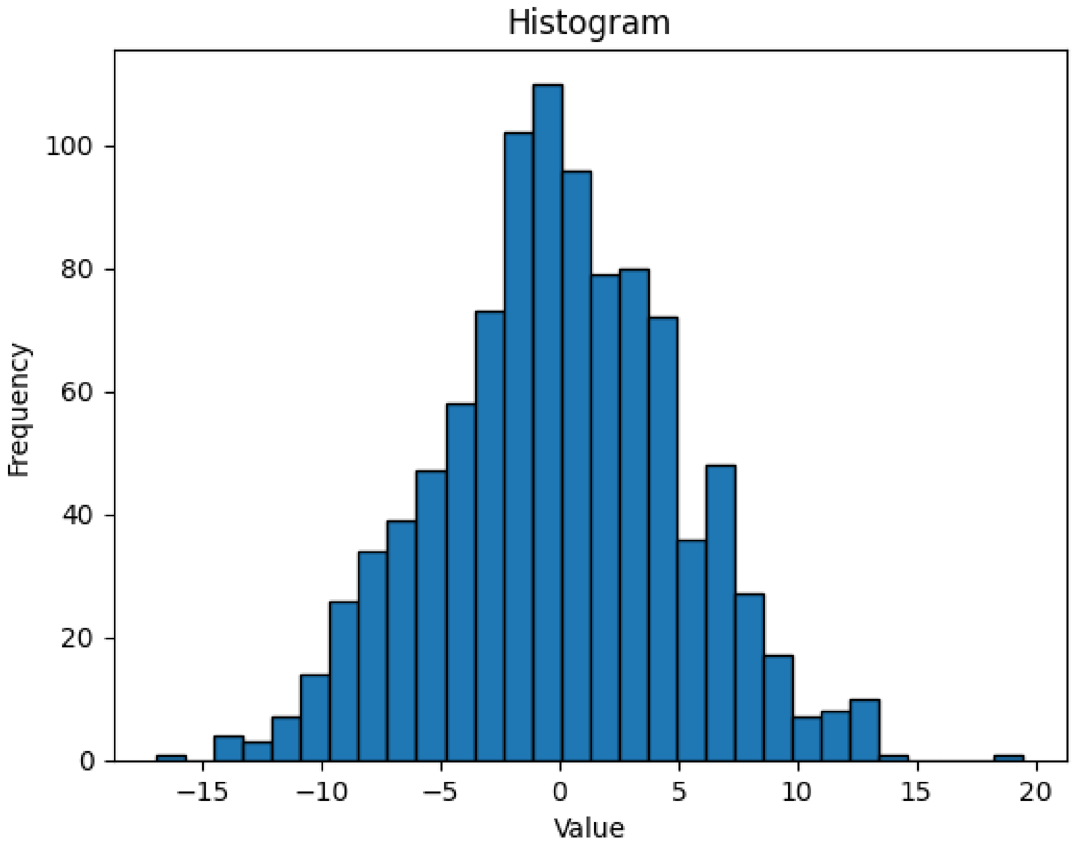
The goal of solving this problem is to become familiar with using built-in Python libraries to create various distributions. Plotting serves as an initial step toward data visualization.

1. (3 POINTS) Create a normally distributed random variable with mean $\mu = 0$, standard deviation $\sigma = 5$ and sample size $n = 1000$. Plot the histogram. Add labels and titles and other details as desired to make your plot understandable. You must use the packages `numpy` and `matplotlib`.

[125]:

```
import numpy as np
import matplotlib.pyplot as plot

# Parameters
mu = 0       # Mean
sigma = 5    # Standard deviation
size = 1000 # Number of samples

# Generate random samples
samples = np.random.normal(mu, sigma, size)
```

```
# Plot the histogram
plot.hist(samples, bins=30, edgecolor='black')

# Labels and title
plot.title('Histogram')
plot.xlabel('Value')
plot.ylabel('Frequency')

# Show plot
plot.show()
```



2. (7 POINTS) We are exploring the Central Limit Theorem (CLT) using a Poisson distribution. Suppose you have a population that follows a Poisson distribution with a rate parameter (or mean) $\lambda = 3$. You will draw multiple samples from this population and calculate the mean of each sample.

Write a Python function that simulates this process. The input of the function should be the sample size, the number of samples, and lambda. The function should: 1. Generate a population with a Poisson distribution (check: https://numpy.org/doc/stable/reference/random/generated/numpy.random.poisson.html). 2.

Draw multiple samples and calculate the mean of each sample. 3. Return these means as an iterable.

There will be no partial credit granted for this question. Any hardcoded results will receive a 0.

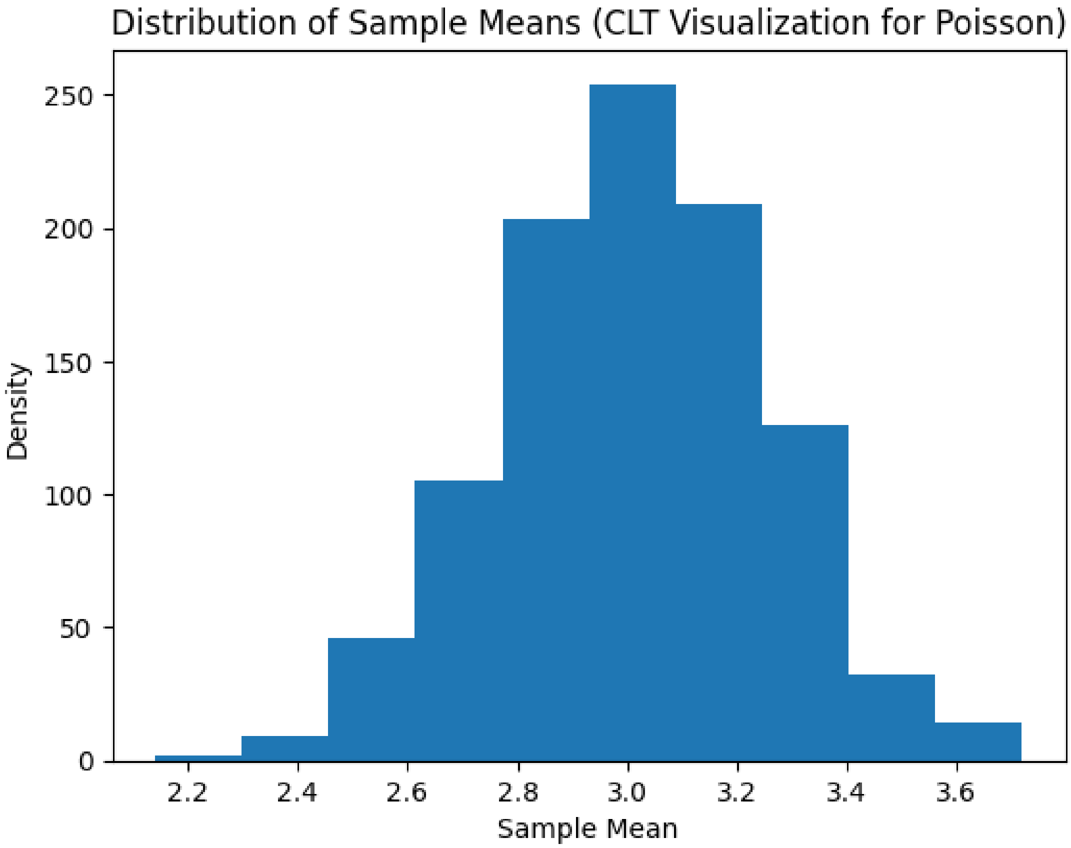
[126]:

```
import numpy as np

def poisson_clt_simulator(sample_size, num_samples, lambda_):
    sample_means = []
    for _ in range(num_samples):
        sample = np.random.poisson(lambda_, sample_size)
        sample_means.append(np.mean(sample)) # Think carefully what you are appending here, refer to variable name
    return sample_means
```

Now use the function to generate 1,000 sample means with sample size 50. Plot the distribution of these sample means to visualize the Central Limit Theorem. Add labels and titles and other details as desired to make your plot understandable.

[127]:

```
import matplotlib.pyplot as plot

# Parameters
sample_size = 50
num_samples = 1000
lambda_ = 3

# Simulate and get sample means
sample_means = poisson_clt_simulator(sample_size, num_samples, lambda_)

# Plot the distribution of sample means
plot.hist(sample_means)

# Add labels and title
plot.xlabel('Sample Mean')
plot.ylabel('Density')
plot.title('Distribution of Sample Means (CLT Visualization for Poisson)')

# Show plot
plot.show()
```



##Q3) (18 POINTS) More on Distributions

You can't get around with distributions while data sciencing. Let's explore how distributions are related to each other.

1. (6 POINTS) Since we have successfully demonstrated how CLT works, lets see what we can do with it.

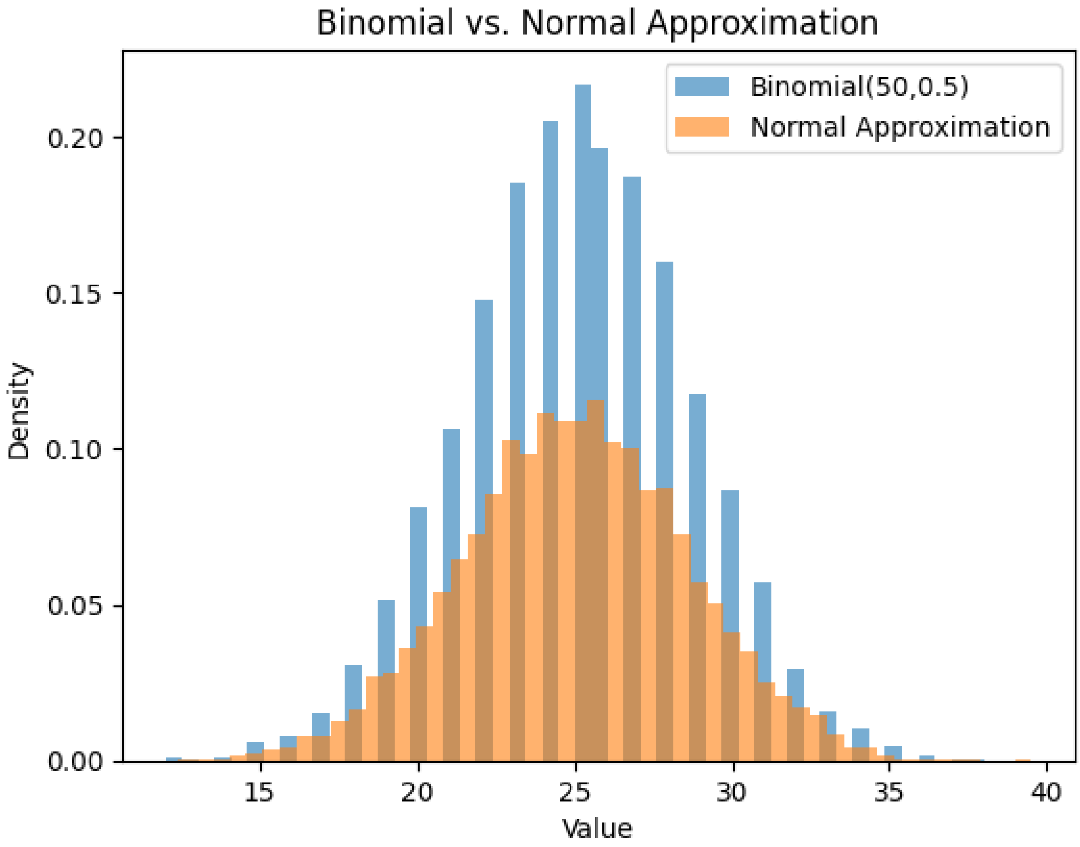
Check out https://numpy.org/doc/stable/reference/random/generated/numpy.random.binomial.html for how to create independent binomial distributions

TASK: Show that a Binomial(n, p) distribution approximates a Normal distribution when n is LARGE (due to CLT). Complete the following code according to comments.

[128]:

```
import numpy as np
import matplotlib.pyplot as plot

size = 10000
n, p = 50, 0.5  # Large n for normal approximation
binomial_samples = np.random.binomial(n, p, size)
normal_samples = np.random.normal(loc=n*p, scale=np.sqrt(n*p*(1-p)), size=size)␣
  ↪# Don't worry about this line unless you are interested
```

```
plot.hist(binomial_samples, bins=50, density=True, alpha=0.6,␣
 ↪label="Binomial(50,0.5)")
plot.hist(normal_samples, bins=50, density=True, alpha=0.6, label="Normal␣
 ↪Approximation")
plot.legend()
plot.title("Binomial vs. Normal Approximation")
plot.xlabel("Value")
plot.ylabel("Density")
plot.show()
```



2. (6 POINTS) Now with Poisson

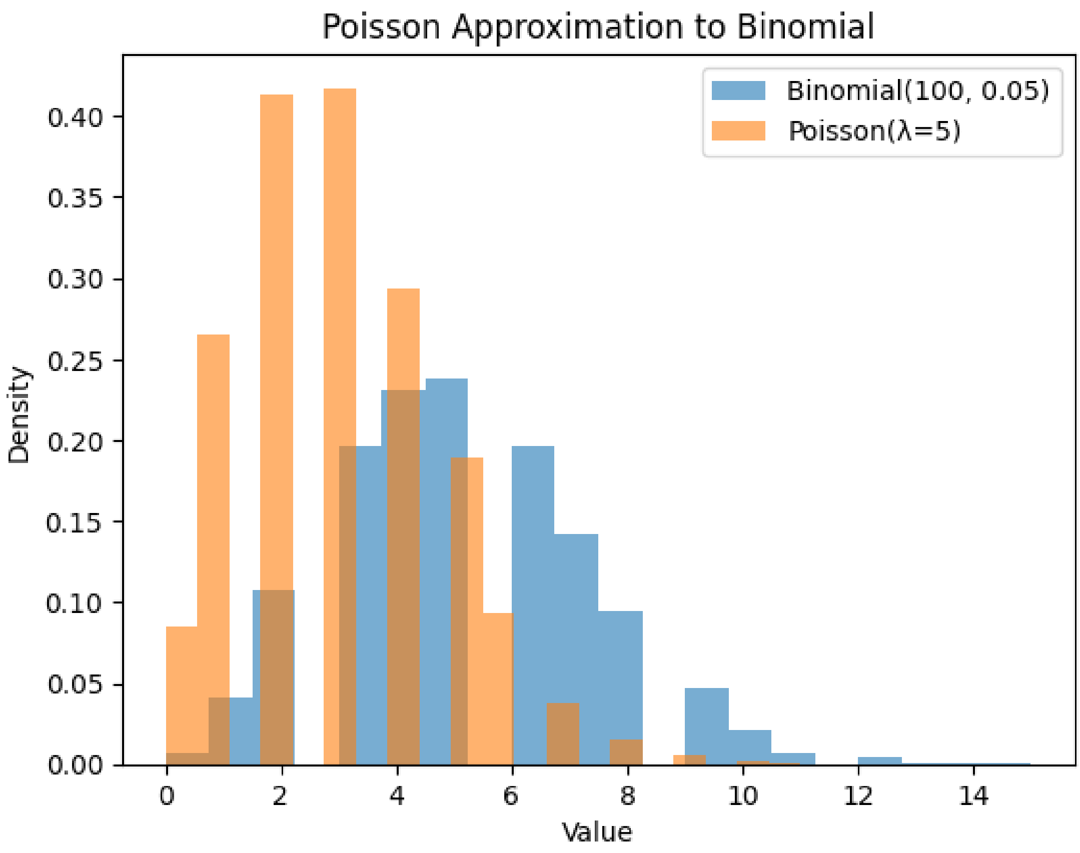
Check out https://numpy.org/doc/stable/reference/random/generated/numpy.random.poisson.html for how to create independent poisson distributions

TASK: Show that when n is large and p is small, a Binomial(n, p) distribution approximates a Poisson distribution with $\lambda = np$. Complete the following code according to comments.

[129]:

```
size = 10000
n, p = 100, 0.05  # np = 5, small p
```

```
binomial_samples = np.random.binomial(n, p, size)
poisson_samples = np.random.poisson(lambda_, size)

plot.hist(binomial_samples, bins=20, density=True, alpha=0.6, label="Binomial(100, 0.05)")
plot.hist(poisson_samples, bins=20, density=True, alpha=0.6, label="Poisson( =5)")
plot.legend()
plot.title("Poisson Approximation to Binomial")
plot.xlabel("Value")
plot.ylabel("Density")
plot.show()
```



3. (6 POINTS) Poisson and Exponential

We know that Poisson counts the number of arrivals, while Exponential models the time between them.

TASK: Plot a Poisson distribution and an Exponential distribution. You do not have to describe and justify your findings.

Check out https://numpy.org/doc/stable/reference/random/generated/numpy.random.exponential.html

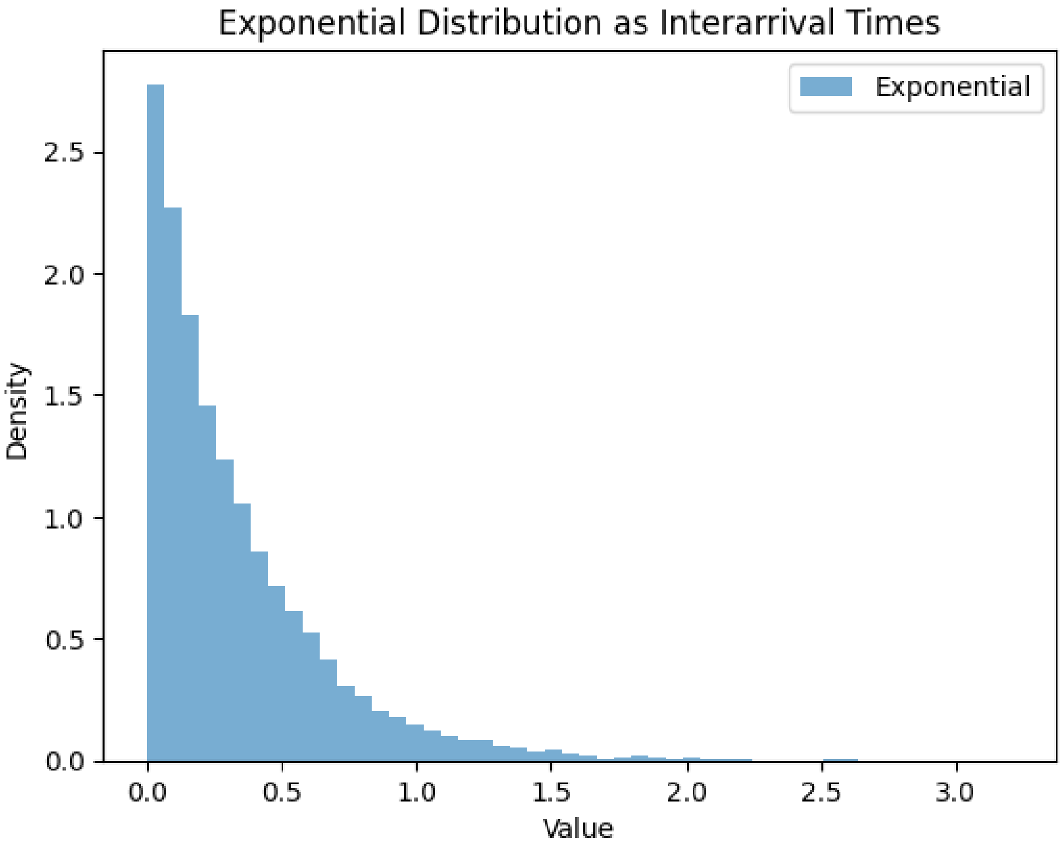
NOTES: If you dont know about Exponensial Distribution, check out:

- https://www.probabilitycourse.com/chapter4/4_2_2_exponential.php
- https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/business/probability/exponential-distribution.html

Complete the following code according to comments.

[130]:

```
size = 10000
lambda_exp = 3  # rate for Poisson

poisson_time_intervals = np.random.poisson(lambda_exp, size) # The variable␣
 ↪name might be tricky, but think carefully exactly what Poisson represents
exponential_samples = np.random.exponential(1/lambda_exp, size) # What is the␣
 ↪scale of exponential, and how is it related to lambda?

plot.hist(exponential_samples, bins=50, density=True, alpha=0.6,␣
 ↪label="Exponential")
plot.legend()
plot.title("Exponential Distribution as Interarrival Times")
plot.xlabel("Value")
plot.ylabel("Density")
plot.show()
```



4 Part 3: Hypothesis Testing

##Q1) (14 POINTS) Hypothesis Tests and P_value

TASK: For the next 5 problems, please describe when you would use each hypothesis test:

- Chi-Squared Test
- Z test
- T test
- Mann-Whitney U Test
- Anova

1.1 (2 POINTS) Chi-Squared Test

> You would use the Chi-Squared Test when you want to estimate the chances two sets of categorical data come from the same distribution.

1.2 (2 POINTS) Z-Test

> The Z-test compares a sample mean to a population mean. It is used when you have a large sample size (typically n > 30) and you know the population standard deviation ( ) or can make a reasonable assumption about it. The test assumes that the sample data

is normally distributed or that the sample size is large enough for the Central Limit Theorem to apply.

1.3 (2 POINTS) T-Test

T-test is a statistical test used to determine if there is a significant difference between the means of two groups. Contrary to the Z-Test, it is used when you have a smaller sample size (typically n < 30) or when you don't know the population standard deviation ( ) and must estimate it from the sample.

1.4 (2 POINTS) Man-Whitney U Test

The Man-Whitney U Test is a nonparametric test used to compare two independent groups and when the data does not meet the assumptions of normality required for a t-test.

1.5 (2 POINTS) ANOVA Test

The ANOVA Test is a powerful statistical test for comparing the means of multiple groups (three or more groups (more than two)) to determine if there are significant differences among them.

1.6 (4 POINTS) : Explain the statistical interpretation of a p-value. What is a p-value? What does it mean? Be sure to explain beyond just "rejecting or failing to reject the null hypothesis."

A p-value is the probability that we would see the observations defined in a "null hypothesis" or "alternative hypothesis" if the null hypothesis is true/correct. The p-value's significance doesn't define the importance of an effect, but how unlikely the data would be if the null hypothesis were true. For example, if the p-value was 0.05 or less, this would be interpreted as the possiblility of a failed null hypothesis, however, in reality it is simply strong evidence against the null hypothesis.

##Q2) (2 POINTS) Create a DataFrame and Display

[131]:

```
import pandas as pd
import matplotlib.pyplot as plt
```

4.0.1 We are creating a DataFrame `df`. Load `colleges.csv` and display the DataFrame below.

This college dataset contains a list of American colleges and their rankings, along with other details such as region, college type, student-to-faculty ratio, etc. In the sections below, you will develop hypotheses, test them, and draw conclusions.

[132]:

```
df = pd.read_csv('colleges.csv')
print(df.head())
```

```
                                         description  rank  \
0  A leading global research university, MIT attr…     1
1  Stanford University sits just outside of Palo …     2
2  One of the top public universities in the coun…     2
3  Princeton is a leading private research univer…     4
```

```
4  Located in upper Manhattan, Columbia Universit…    5

                       organizationName state  studentPopulation  \
0  Massachusetts Institute of Technology    MA              12195
1                    Stanford University    CA              20961
2     University of California, Berkeley    CA              45878
3                   Princeton University    NJ               8532
4                    Columbia University    NY              33882

  campusSetting  medianBaseSalary   longitude   latitude  \
0         Urban          173700.0  -71.093539  42.359006
1      Suburban          173500.0 -122.168924  37.431370
2         Urban          154500.0 -122.258393  37.869236
3         Urban          167600.0  -74.659119  40.349855
4         Urban          148800.0  -73.961288  40.806515

                 website  … yearFounded stateCode  \
0        http://web.mit.edu  …      1861.0        MA
1   http://www.stanford.edu  …      1891.0        CA
2   http://www.berkeley.edu  …      1868.0        CA
3  http://www.princeton.edu  …      1746.0        NJ
4   http://www.columbia.edu  …      1754.0        NY

             collegeType                             carnegieClassification  \
0  Private not-for-profit  Doctoral Universities: Very High Research Acti…
1  Private not-for-profit  Doctoral Universities: Very High Research Acti…
2                  Public  Doctoral Universities: Very High Research Acti…
3  Private not-for-profit  Doctoral Universities: Very High Research Acti…
4  Private not-for-profit  Doctoral Universities: Very High Research Acti…

  studentFacultyRatio  totalStudentPop undergradPop totalGrantAid  \
0                   3            12195         4582    35299332.0
1                   4            20961         8464    51328461.0
2                  19            45878        33208    64495611.0
3                   4             8532         5516    44871096.0
4                   6            33882         8689    44615007.0

  percentOfStudentsFinAid  percentOfStudentsGrant
0                    75.0                    60.0
1                    70.0                    55.0
2                    63.0                    53.0
3                    62.0                    61.0
4                    58.0                    54.0

[5 rows x 25 columns]
```

TASK 2.1 (2 POINTS): Some entries of the dataframe are NaN. remove those entries.

[133]:

```
df.dropna()
```

[133]:

```
                                           description  rank  \
0    A leading global research university, MIT attr…     1   
1    Stanford University sits just outside of Palo …     2   
2    One of the top public universities in the coun…     2   
3    Princeton is a leading private research univer…     4   
4    Located in upper Manhattan, Columbia Universit…     5   
..                                                 …   …   
490  Loyola University New Orleans provides student…   491   
491  Xavier University is a Jesuit Catholic school …   492   
493  St. Joseph’s College is a private institution …   494   
494  A liberal arts college founded by the Moravian…   495   
497  The University of Memphis is a large public re…   498   

                          organizationName state  studentPopulation  \
0    Massachusetts Institute of Technology    MA              12195   
1                      Stanford University    CA              20961   
2        University of California, Berkeley    CA              45878   
3                     Princeton University    NJ               8532   
4                      Columbia University    NY              33882   
..                                     …   …                 …   
490          Loyola University New Orleans    LA               4972   
491                      Xavier University    OH               8079   
493                  St. Joseph's College (NY)    NY               5901   
494                     Moravian University    PA               2961   
497                   University of Memphis    TN              25128   

    campusSetting  medianBaseSalary   longitude   latitude  \
0           Urban          173700.0  -71.093539  42.359006   
1        Suburban          173500.0 -122.168924  37.431370   
2           Urban          154500.0 -122.258393  37.869236   
3           Urban          167600.0  -74.659119  40.349855   
4           Urban          148800.0  -73.961288  40.806515   
..            …               …         …        …   
490         Urban          102300.0  -90.077714  29.953690   
491         Urban          104900.0  -84.476379  39.149037   
493         Urban          100900.0  -73.968304  40.690548   
494         Urban          109800.0  -75.381596  40.630303   
497         Urban           90700.0  -89.939618  35.118453   

                      website  … yearFounded stateCode  \
0          http://web.mit.edu  …      1861.0        MA   
1     http://www.stanford.edu  …      1891.0        CA   
2     http://www.berkeley.edu  …      1868.0        CA   
3    http://www.princeton.edu  …      1746.0        NJ   
4     http://www.columbia.edu  …      1754.0        NY   
```

```
..                        …  …         …       …
490      http://www.loyno.edu  …      1904.0        LA
491     http://www.xavier.edu  …      1831.0        OH
493      http://www.sjcny.edu  …      1916.0        NY
494   http://www.moravian.edu  …      1742.0        PA
497     http://www.mephis.edu  …      1912.0        TN

                collegeType  \
0    Private not-for-profit
1    Private not-for-profit
2                    Public
3    Private not-for-profit
4    Private not-for-profit
..                      …
490  Private not-for-profit
491  Private not-for-profit
493  Private not-for-profit
494  Private not-for-profit
497                  Public

                                carnegieClassification studentFacultyRatio  \
0    Doctoral Universities: Very High Research Acti…                  3
1    Doctoral Universities: Very High Research Acti…                  4
2    Doctoral Universities: Very High Research Acti…                 19
3    Doctoral Universities: Very High Research Acti…                  4
4    Doctoral Universities: Very High Research Acti…                  6
..                                                 …                 …
490                Doctoral/Professional Universities                  13
491  Master's Colleges & Universities: Larger Programs                  11
493  Master's Colleges & Universities: Medium Programs                  12
494      Baccalaureate Colleges: Arts & Sciences Focus                  11
497      Doctoral Universities: High Research Activity                  16

     totalStudentPop undergradPop totalGrantAid percentOfStudentsFinAid  \
0              12195         4582    35299332.0                    75.0
1              20961         8464    51328461.0                    70.0
2              45878        33208    64495611.0                    63.0
3               8532         5516    44871096.0                    62.0
4              33882         8689    44615007.0                    58.0
..                 …          …           …                     …
490             4972         3538    26114959.0                    99.0
491             8079         5473    28294277.0                   100.0
493             5901         4429    11919881.0                    99.0
494             2961         2268    12685943.0                   100.0
497            25128        20011    27575189.0                    98.0

     percentOfStudentsGrant
```

```
0                       60.0
1                       55.0
2                       53.0
3                       61.0
4                       54.0
..                       …
490                     99.0
491                    100.0
493                     99.0
494                    100.0
497                     97.0

[422 rows x 25 columns]
```

#Q3) (8 POINTS) Hypothesis Testing

Try to find relationships in this dataset through hypothesis testing. For each hypothesis test:

- First chose a null hypothesis, or a statement that there is no effect between different variables, that serves as a default assumption.
- Then chose an alternative hypothesis, or a statement that suggests that there is a correlation between different variables.

For the questions below, assume $\alpha = 0.05$.

4.1 First Hypothesis

- HO: The region of the college does not have an effect on the likelihood of the college type.
- HA: The region of the college does have an effect on the likelihood of the college type.

Our plan is to apply a chi-squared test. You may find it helpful to consult the `scipy.stats` library's documentation: https://docs.scipy.org/doc/scipy/reference/stats.html

Contingency table is a table used in statistics to display the frequency distribution of variables. It will help us perform a chi-squared test on our data. You can find more information on contingency table here - https://en.wikipedia.org/wiki/Contingency_table

TASK 3.1 (2 POINTS): Create a contingency table and display it.

[134]:

```python
contingency_table = pd.crosstab(df['region'], df['collegeType'])
print(contingency_table)
```

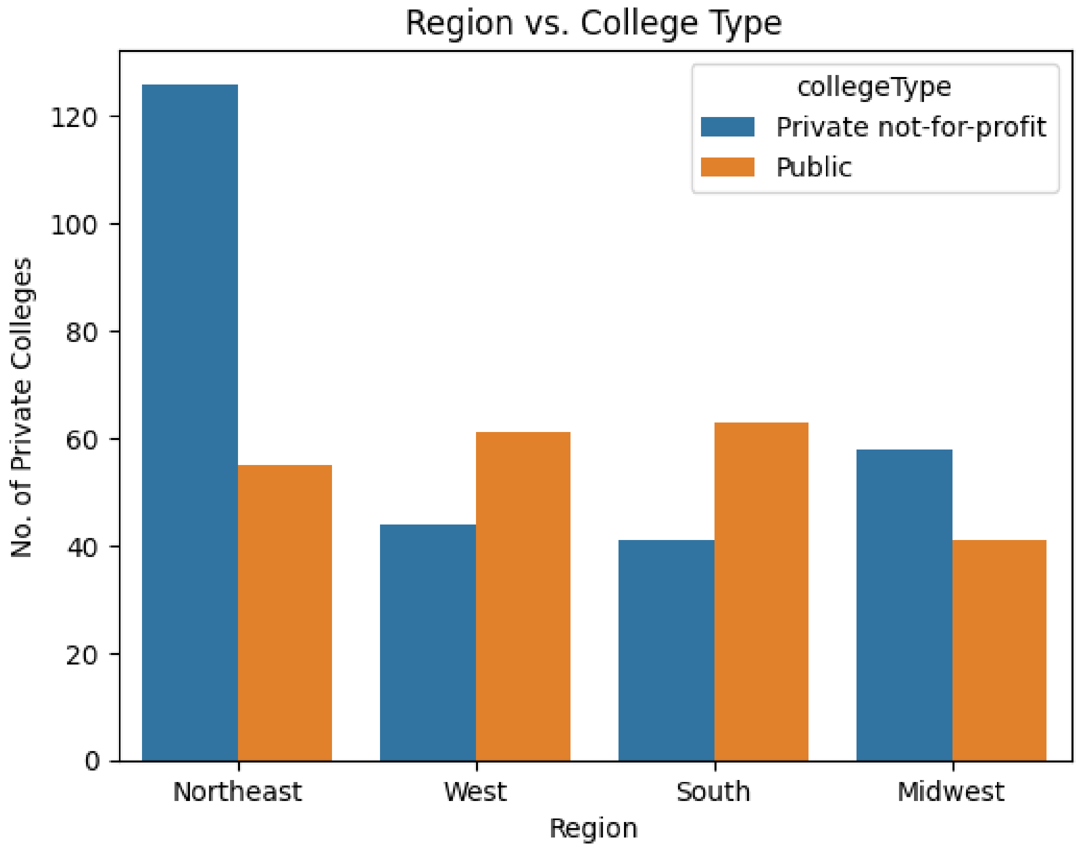
```
collegeType  Private not-for-profit  Public
region
Midwest                          58      41
Northeast                       126      55
South                            41      63
West                             44      61
```

TASK 3.2 (2 POINTS): Why would we consider using a chi-squared test specifically (as opposed to some other hypothesis test)?

We want to consider using a chi-squared test specifically as opposed to some other test because chi-squared tests estimate the chances two sets of categorical data come from the same distribution., which is exactly what we are doing - we want to see if the region affects the college type (which are both categorical data), which is the same as testing if they're from the same distribution.

TASK 3.3 (2 POINTS): Create a plot showing the relationship between the regions and the no. of private colleges in it.

[135]:
```python
import seaborn as sns
import matplotlib.pyplot as plot

# Any graph that can be easily identified as the prompt above is suffice
sns.countplot(data=df, x='region', hue='collegeType')
plot.title('Region vs. College Type')
plot.xlabel('Region')
plot.ylabel('No. of Private Colleges')
plot.show()
```



TASK 3.4 (2 POINTS): Explain what you can infer from your plot

The northeast has the most private not-for-profit colleges and has almost double its

> publi colleges, compared to the other regions where the numbers are more closer. The south has the least private not-for-profit colleges, the midwest has the least public colleges, and the south has the most public colleges out of the four regions.

#Q4) (5 POINTS) Conduct the chi-squared test

TASK: 4.1 (2 POINTS): Display the p-value of applying the chi-squared test using the `chi2_contingency()` function.

[136]:

```python
from scipy.stats import chi2_contingency

chi2, p, dof, expected = chi2_contingency(contingency_table)

print(f'P-value: {p}')

# Interpretation
alpha = 0.05

if p < alpha:
    print("Reject the null hypothesis. The region of the college does have an␣
  ↪effect on the likelihood of the college type.")
else:
    print("Fail to reject the null hypothesis. The region of the college does␣
  ↪not have an effect on the likelihood of the college type.")
```

```
P-value: 2.4020755468119133e-07
Reject the null hypothesis. The region of the college does have an effect on the
likelihood of the college type.
```

TASK: 4.2 (3 POINTS): Based on the p-value, determine whether to reject or fail to reject the null hypothesis. Explain your answer.

> Based on the p-value, we would reject the null hypothesis because our p-value of 2.4020755468119133e-07 is less than our significance level of 0.05, which indicates that there is a statistically significant correlation between the region of the college and the likelihood of the college type.

#Q5) (3 POINTS) A New Hypothesis

Now create a new hypothesis test for whether the campus setting has an effect on the total student population. (Assume $\alpha = 0.05$).

TASK 5.1 (3 POINTS): Write down your null and alternative hypotheses:

- HO: The campus setting does not have an effect on the total student population.
- HA: The campus setting does have an effect on the total student population.

#Q6) (7 POINTS) Hypothesis Testing

TASK 6.0: Split the data into 3 different dataframes based on campus setting.

[137]:

```
import pandas as pd
import scipy.stats as stats

urban = df[df['campusSetting'] == 'Urban']
suburban = df[df['campusSetting'] == 'Suburban']
rural = df[df['campusSetting'] == 'Rural']

print(urban.head())
print(suburban.head())
print(rural.head())
```

```
                                         description  rank  \
0  A leading global research university, MIT attr…     1   
2  One of the top public universities in the coun…     2   
3  Princeton is a leading private research univer…     4   
4  Located in upper Manhattan, Columbia Universit…     5   
5  The University of California, Los Angeles is t…     6   

                        organizationName state  studentPopulation  \
0  Massachusetts Institute of Technology    MA              12195   
2     University of California, Berkeley    CA              45878   
3                   Princeton University    NJ               8532   
4                    Columbia University    NY              33882   
5  University of California, Los Angeles    CA              46947   

  campusSetting  medianBaseSalary   longitude   latitude  \
0         Urban          173700.0  -71.093539  42.359006   
2         Urban          154500.0 -122.258393  37.869236   
3         Urban          167600.0  -74.659119  40.349855   
4         Urban          148800.0  -73.961288  40.806515   
5         Urban          137200.0 -118.437855  34.073903   

                   website  … yearFounded stateCode  \
0        http://web.mit.edu  …      1861.0        MA   
2   http://www.berkeley.edu  …      1868.0        CA   
3  http://www.princeton.edu  …      1746.0        NJ   
4   http://www.columbia.edu  …      1754.0        NY   
5           http://ucla.edu  …      1919.0        CA   

              collegeType                             carnegieClassification  \
0  Private not-for-profit  Doctoral Universities: Very High Research Acti…   
2                  Public  Doctoral Universities: Very High Research Acti…   
3  Private not-for-profit  Doctoral Universities: Very High Research Acti…   
4  Private not-for-profit  Doctoral Universities: Very High Research Acti…   
5                  Public  Doctoral Universities: Very High Research Acti…   

  studentFacultyRatio  totalStudentPop undergradPop totalGrantAid  \
0                   3            12195        4582    35299332.0
2                  19            45878       33208    64495611.0
3                   4             8532        5516    44871096.0
4                   6            33882        8689    44615007.0
5                  18            46947       33641    61100980.0

  percentOfStudentsFinAid  percentOfStudentsGrant
0                    75.0                    60.0
2                    63.0                    53.0
3                    62.0                    61.0
4                    58.0                    54.0
5                    73.0                    67.0

[5 rows x 25 columns]
                                          description  rank  \
1   Stanford University sits just outside of Palo …     2
22  The second-oldest member of the University of …    23
23  A top liberal arts school, Amherst is located …    24
26  A private research university, Washington Univ…    27
28  This public research university of Charlottesv…    29

                      organizationName state  studentPopulation campusSetting  \
1                  Stanford University    CA              20961      Suburban
22     University of California, Davis    CA              41236      Suburban
23                     Amherst College    MA               1940      Suburban
26  Washington University in St. Louis    MO              17893      Suburban
28              University of Virginia    VA              29237      Suburban

    medianBaseSalary   longitude   latitude                         website  \
1           173500.0 -122.168924  37.431370         http://www.stanford.edu
22          134800.0 -121.747976  38.540631          http://www.ucdavis.edu
23          148700.0  -72.533204  42.370772  http://https://www.amherst.edu
26          136000.0  -90.301291  38.647812            http://www.wustl.edu
28          137300.0  -78.581033  38.078711         http://www.virginia.edu

    … yearFounded stateCode             collegeType  \
1   …      1891.0        CA  Private not-for-profit
22  …      1908.0        CA                  Public
23  …      1821.0        MA  Private not-for-profit
26  …      1853.0        MO  Private not-for-profit
28  …      1819.0        VA                  Public

                               carnegieClassification studentFacultyRatio  \
1   Doctoral Universities: Very High Research Acti…                   4
22  Doctoral Universities: Very High Research Acti…                  20
23      Baccalaureate Colleges: Arts & Sciences Focus                   7
26  Doctoral Universities: Very High Research Acti…                   7
28  Doctoral Universities: Very High Research Acti…                  15
```

```
    totalStudentPop undergradPop totalGrantAid percentOfStudentsFinAid  \
1             20961         8464    51328461.0                    70.0  
22            41236        33181    72219528.0                    74.0  
23             1940         1940    15522081.0                    65.0  
26            17893         8909    39741443.0                    54.0  
28            29237        19253    34787367.0                    60.0  

    percentOfStudentsGrant  
1                     55.0  
22                    66.0  
23                    57.0  
26                    46.0  
28                    39.0  

[5 rows x 25 columns]
                                          description  rank  \
6   Located in rural Williamstown, MA, Williams Co…     7   
13  The smallest Ivy League school, Dartmouth Coll…    14   
43  Colgate University is a leading liberal arts s…    44   
47  Located in the town of Brunswick, ME, Bowdoin …    48   
54  Middlebury College is a small private liberal …    55   

      organizationName state  studentPopulation campusSetting  \
6     Williams College    MA               2307         Rural   
13   Dartmouth College    NH               7171         Rural   
43  Colgate University    NY               3112         Rural   
47     Bowdoin College    ME               1973         Rural   
54  Middlebury College    VT               4616         Rural   

    medianBaseSalary  longitude   latitude                   website  …  \
6           152600.0 -73.208078  42.712389    http://www.williams.edu  …   
13          161300.0 -72.289499  43.700465   http://www.dartmouth.edu  …   
43          154400.0 -75.536415  42.821191     http://www.colgate.edu  …   
47          145600.0 -69.963975  43.906764     http://www.bowdoin.edu  …   
54          138100.0 -73.167117  44.014999  http://www.middlebury.edu  …   

   yearFounded stateCode             collegeType  \
6       1793.0        MA  Private not-for-profit   
13      1769.0        NH  Private not-for-profit   
43      1819.0        NY  Private not-for-profit   
47      1794.0        ME  Private not-for-profit   
54      1800.0        VT  Private not-for-profit   

                               carnegieClassification studentFacultyRatio  \
6       Baccalaureate Colleges: Arts & Sciences Focus                   6   
13  Doctoral Universities: Very High Research Acti…                   7   
43      Baccalaureate Colleges: Arts & Sciences Focus                   9   
```

```
47      Baccalaureate Colleges: Arts & Sciences Focus                  8
54      Baccalaureate Colleges: Arts & Sciences Focus                  8

    totalStudentPop undergradPop totalGrantAid percentOfStudentsFinAid  \
6              2307         2251    15204855.0                     62.0
13             7171         4885    27997693.0                     58.0
43             3112         3098    16450893.0                     50.0
47             1973         1973    12574545.0                     58.0
54             4616         3833    12382994.0                     43.0

    percentOfStudentsGrant
6                     52.0
13                    45.0
43                    42.0
47                    52.0
54                    41.0

[5 rows x 25 columns]
```

TASK 6.1 (2 POINTS): Choose an appropriate hypothesis test and display the p-value of applying the that test.

[138]:

```
#Applying the ANOVA test
anova_result = stats.f_oneway(urban['studentPopulation'],␣
 ↪suburban['studentPopulation'], rural['studentPopulation'])

print(f'P-value: {anova_result.pvalue}')

# Interpretation
alpha = 0.05

if anova_result.pvalue < alpha:
    print("Reject the null hypothesis. The campus setting does have an effect␣
 ↪on the total student population.")
else:
    print("Fail to reject the null hypothesis. The campus setting does not have␣
 ↪an effect on the total student population.")
```

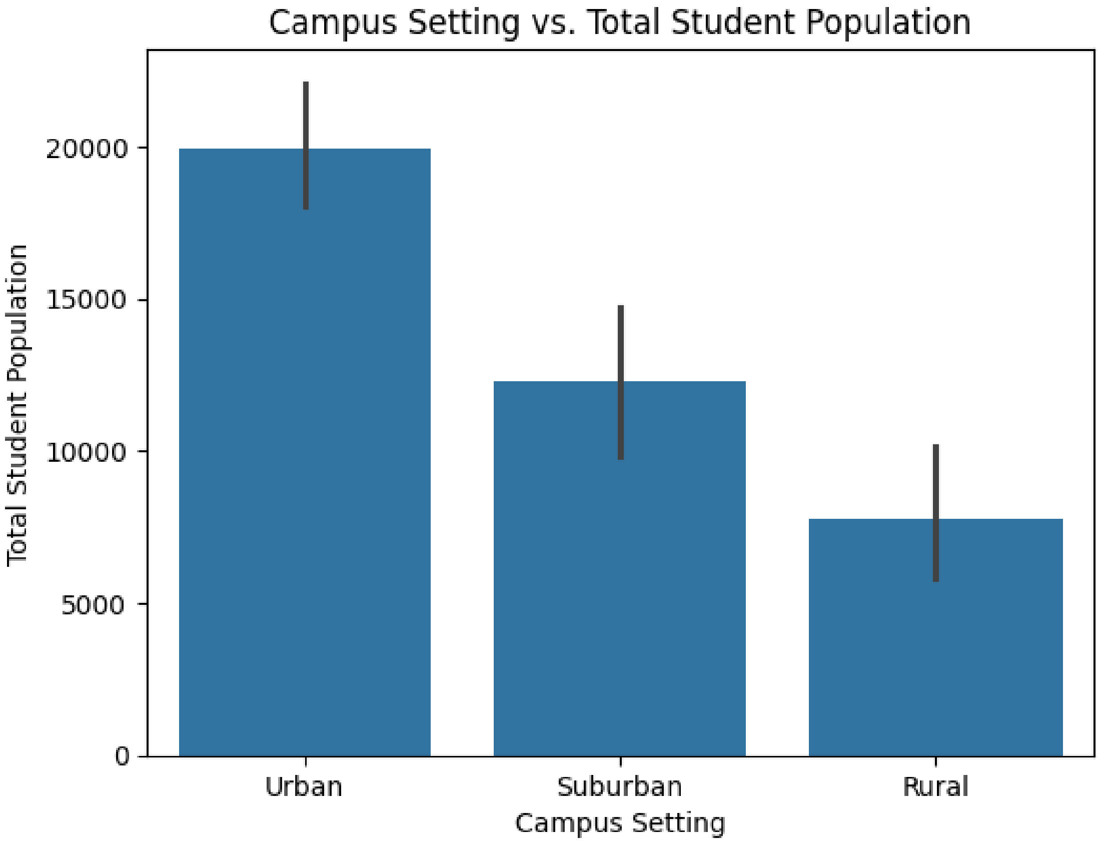
```
P-value: 1.8281832202730275e-10
Reject the null hypothesis. The campus setting does have an effect on the total
student population.
```

TASK 6.2 (2 POINTS): Create a graph(s) using `matplotlib` to show the relationship between campus setting and total student population.

[139]:

```
import matplotlib.pyplot as plot
import seaborn as sns

sns.barplot(data=df, x='campusSetting', y='studentPopulation')
```

```
plot.title('Campus Setting vs. Total Student Population')
plot.xlabel('Campus Setting')
plot.ylabel('Total Student Population')
plot.show()
```



TASK 6.3 (3 POINTS): Based on the p-value, determine whether to reject or fail to reject the null hypothesis. Explain your answer.

> Based on the p-value, we would reject the null hypothesis because our p-value of 1.8281832202730275e-10 is less than our significance level of 0.05, which indicates that there is a statistically significant correlation between the campus setting and total student population.

#*Q7)* (2 POINTS) Post Hoc Tests

TASK 7.1 (2 POINTS): Why might we need post-hoc tests in this scenario?

> We need post-hoc tests whenever we use the ANOVA test because when you use ANOVA, you find the main effect is significant, indicating that the main effect is different between groups. However, you might want to know which group is different than the other groups, which is what post-hoc tests do.

BONUS TASK 7.2 (2+1=3 POINTS): Apply a post-hoc test of your choice

```
[140]: # Your code here
```

Write your interpretation here

#*Q8)* (19 POINTS) Hypothesis Test

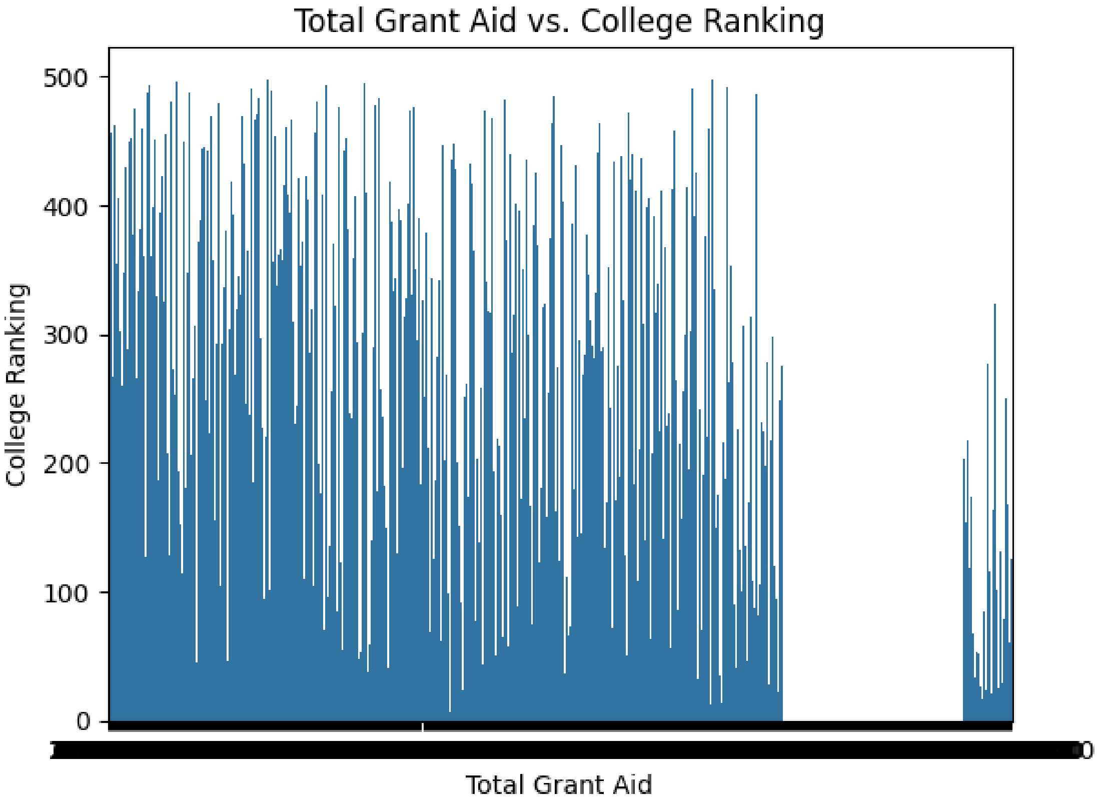
Now create a new hypothesis test for whether the total grant aid has an affect on college ranking. (Assume $\alpha = 0.05$).

TASK 8.1 (3 POINTS): Write down the null and alternative hypotheses below.

- HO: The total grant aid does not have an effect on college ranking.
- HA: The total grant aid does have an effect on college ranking.

TASK 8.2 (2 POINTS): Create a plot using `matplotlib` that visualizes your hypothesis.

```
[141]: import matplotlib.pyplot as plot
       import seaborn as sns

       sns.barplot(data=df, x='totalGrantAid', y='rank')
       plot.title('Total Grant Aid vs. College Ranking')
       plot.xlabel('Total Grant Aid')
       plot.ylabel('College Ranking')
       plot.show()
```



TASK 8.3 (3 POINTS): Apply an appropriate hypothesis test and find the p-value of the it.

[142]:

```python
# Applying the Pearson Correlation Test

import pandas as pd
import scipy.stats as stats

# Dropping rows with missing values
df_clean = df.dropna(subset=['totalGrantAid', 'rank'])

# Applying the Pearson Correlation Test
pearson_result = stats.pearsonr(df_clean['totalGrantAid'], df_clean['rank'])

print(f"P-value: {pearson_result.pvalue}")

# Interpretation
alpha = 0.05

if pearson_result.pvalue < alpha:
    print("Reject the null hypothesis. The total grant aid does have an effect␣
 ↪on college ranking.")
else:
    print("Fail to reject the null hypothesis. The total grant aid does not␣
 ↪have an effect on college ranking.")
```

```
P-value: 2.7231734688594448e-27
Reject the null hypothesis. The total grant aid does have an effect on college
ranking.
```

TASK 8.4 (3 POINTS): Based on the p-value, determine whether to reject or fail to reject the null hypothesis. Explain your answer.

> Based on the p-value, we would reject the null hypothesis because our p-value of 2.7231734688594448e-27 is less than our significance level of 0.05, which indicates that there is a statistically significant correlation between total grant aid and college ranking.

TASK 8.5 (3 POINTS): Based on your previous answer, can you conclude that increasing grant aid will change a college's ranking? What is experimental procedure required to reach this conclusion?

> Based on the previous answer, we can conclude that increasing grant aid is statistically highly correlated with college ranking. However, correlation does not imply causation, therefore we cannot conclude that increasing grant aid will change a college's ranking. To reach that conclusion, we will need to perform a controlled experiment.

TASK 8.6 (3 POINTS): What kind of t-test (right-tail or left-tail) would you use to verify the following hypothesis?

H0: There is no difference in student to faculty ratio between private and public colleges

HA: Private colleges have a smaller student to faculty ratio

Also perform the test and print your p value.

> I would use a left-tail test to verify the alternative hypothesis because it states that private colleges have a smaller student to faculty ratio, which implies that we are testing if the mean student to faculty ratio of private colleges is less than that of public colleges. This also means the rejection region is located to the extreme left of the distribution.

[143]:

```python
private_colleges = df[df['collegeType'] == 'Private not-for-profit']['studentFacultyRatio']
public_colleges = df[df['collegeType'] == 'Public']['studentFacultyRatio']

# Perform the left-tailed t-test
left_tail_t_test = stats.ttest_ind(private_colleges, public_colleges, alternative='less')

# Print the results
print(f"P-value: {left_tail_t_test.pvalue}")

# Interpretation
alpha = 0.05

if left_tail_t_test.pvalue < alpha:
    print("Reject the null hypothesis. Private colleges have a smaller student to faculty ratio.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference in student to faculty ratio between private and public colleges.")
```

```
P-value: 5.617896962995605e-73
Reject the null hypothesis. Private colleges have a smaller student to faculty
ratio.
```

TASK 8.7 (2 POINTS): Based on the p-value, determine whether to reject or fail to reject the null hypothesis. Explain your answer.

> Based on the p-value, we would reject the null hypothesis because our p-value of 5.617896962995605e-73 is less than our significance level of 0.05, which indicates that there is a statistically significant correlation between private colleges and a smaller student-to-faculty ratio.

5 THE END!