HOMEWORK 2: BRAIN CONDITIONING STATS 📊

DUE: *Oct 02, 2025 @ 11:59 PM*

24-HR LATE DUE DATE WITH A 15% PENALTY: *Oct 3, 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.

Overview:

- This assignment is worth 120 points.
- There are 3 mandatory parts to complete in this section (DON'T REMOVE ANY PART OF THE QUESTIONS).
- Make a local copy, read all of the instructions carefully, and complete all tasks.

Submission:

- Submit your completed notebook (.ipynb) file and PDF to the "HW2 - Brain Conditioning Stats" section in Gradescope.

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.

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

No Hardcoding either 😊📉

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 13,000 users. There are 5,500 bloggers, 7,000 shoppers, and 6,500 reviewers. The users could be in multiple categories. 2,200 of the bloggers are shoppers, 1,800 of the bloggers are reviewers, and 3,000 shoppers are also reviewers.

Answer the following questions in the designated boxes:

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.)

> X = 1000

2. (3 POINTS) Calculate the probability that a randomly selected shopper is also a blogger. Round to the nearest hundredth. (Hint: Use Bayes Theorem)

> 0.31

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

> 0.31

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 questions in the designated boxes:

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

> 7

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

> 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 questions in the designated boxes:

1. (3 POINTS) What is the probability that X takes a value between 6 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/

> 0.81

2. (3 POINTS) What is the probability that X takes a value greater than 15?

> 0.01

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` dictionary object 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. The order of the keys in the Counter does not matter, unsorted is fine.

In [34]:
```python
from collections import Counter

def count_streaks(xs):
    streak_active = 0
    ys = []
    for num in xs:
        if num == 1:
            streak_active += 1
        elif streak_active != 0:
            ys.append(streak_active)
            streak_active = 0

    if streak_active != 0:
        ys.append(streak_active)

    return Counter(ys)
```

In [35]:
```python
# 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)
display(count_streaks(np.random.randint(0,2,1000000)))
```

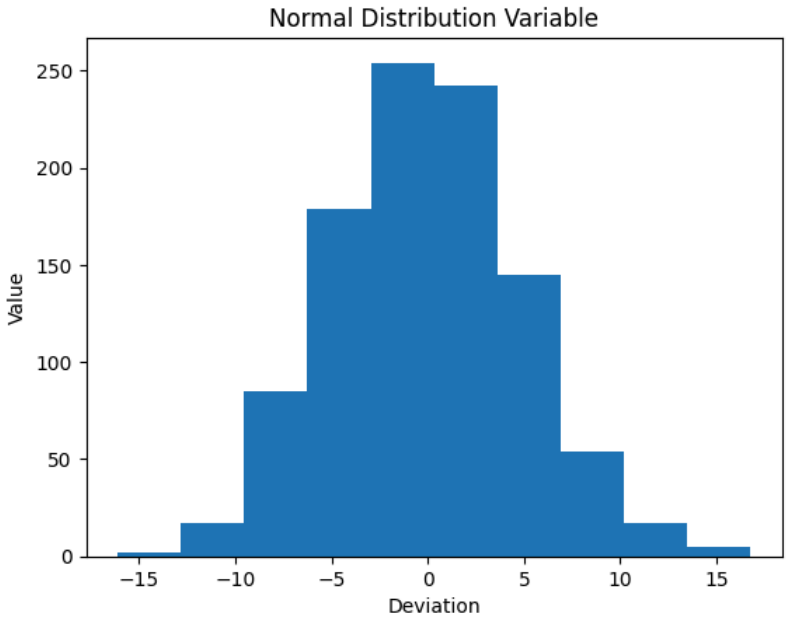
```
Counter({1: 2, 3: 2, 2: 1})
Counter({1: 125036,
         2: 62589,
         3: 31100,
         4: 15859,
         5: 7699,
         6: 3893,
         7: 1921,
         8: 946,
         9: 470,
         10: 245,
         11: 126,
         12: 45,
         13: 29,
         14: 11,
         15: 9,
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

In [55]:
```python
import numpy as np
import matplotlib.pyplot as plt

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

# Generate random samples
samples = np.random.normal(0, 5, 1000)

# Plot the histogram
plt.hist(samples)

# Labels and title
plt.title("Normal Distribution Variable")
plt.xlabel("Deviation")
plt.ylabel("Value")


# Show plot
plt.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:

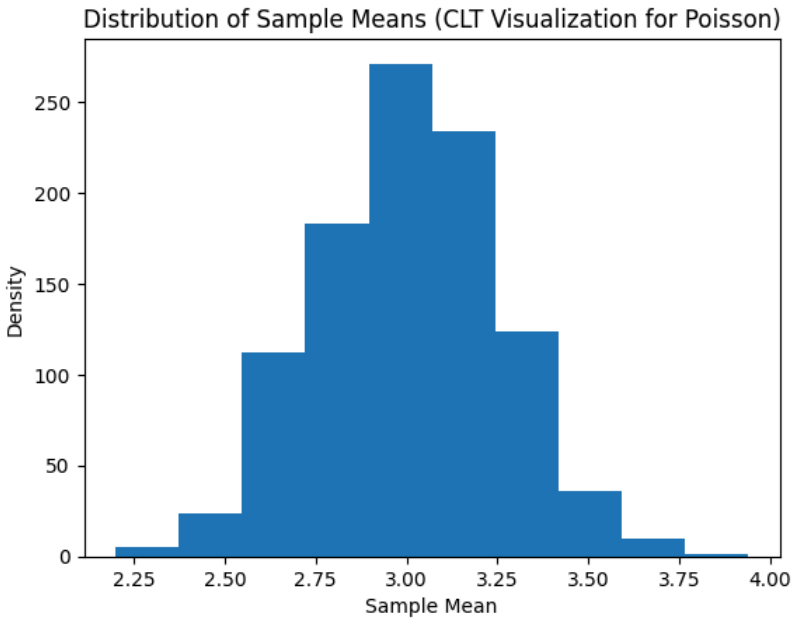
1. Generate a population with a Poisson distribution (check: https://numpy.org/doc/stable/reference/random/generated/numpy.random.poisson.html).
2. Draw multiple samples and calculate the mean of each sample.
3. Return these means as an iterable.

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

In [37]:
```python
import numpy as np

def poisson_clt_simulator(sample_size, num_samples, lambda_):
    sample_means = []
    for _ in range(num_samples):
        sample = np.random.poisson(lam=lambda_, size=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.

In [56]:
```python
import matplotlib.pyplot as plt

# Parameters
sample_size = 50
num_samples = 1000
lambda_ = 3

# Simulate and get sample means
sample_means = poisson_clt_simulator(50, 1000, 3)

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

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

# Show plot
plt.show()
```



Q3) (18 POINTS) More on Distributions

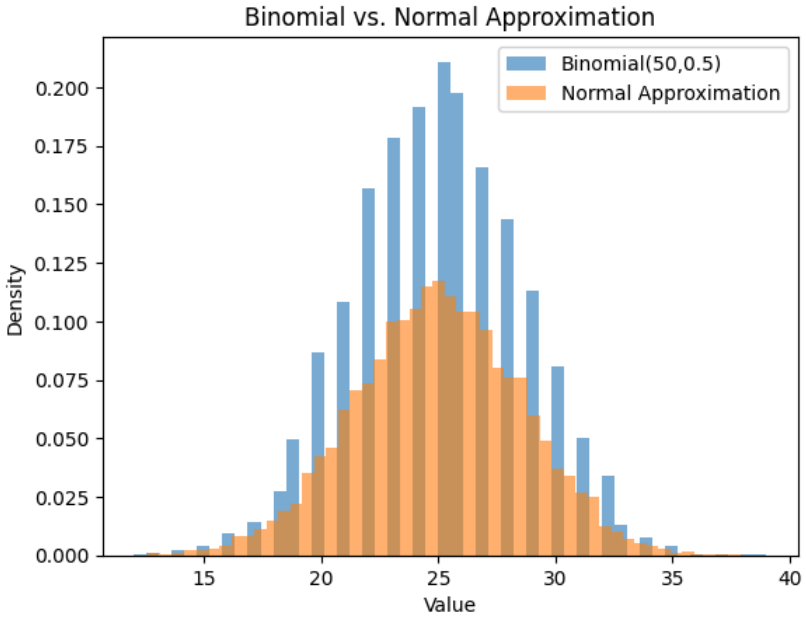
You can't get around with distributions while data sciencing. Let's explore how distributions are related to each other.

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.

In [ ]:
```python
import numpy as np
import matplotlib.pyplot as plt

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

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

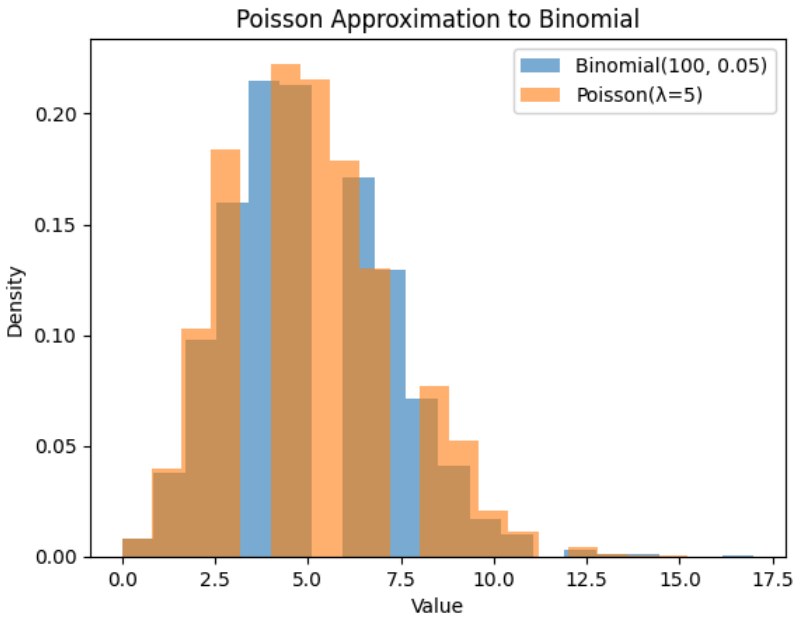


2. (6 POINTS) Now with Poisson

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.

In [58]:
```
size = 10000
n, p = 100, 0.05  # np = 5, small p
binomial_samples = np.random.default_rng().binomial(n, p, size=size)
poisson_samples = np.random.default_rng().poisson(lam=n*p, size=size)

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

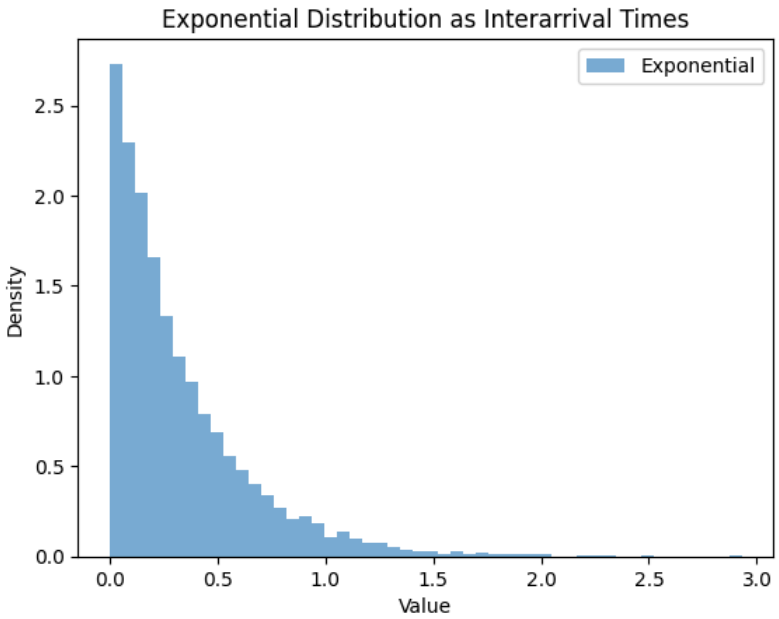
*Check out https://numpy.org/doc/stable/reference/random/generated/numpy.random.exponential.html *

> **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.

In [59]:
```
size = 10000
lambda_exp = 3  # rate for Poisson

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

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



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. Answer the questions in the designated boxes:

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

1.1 (2 POINTS) Chi-Squared Test

> This test determines if observed frequencies differ significantly from expected frequencies in a table.

1.2 (2 POINTS) Z-Test

> Z test is used to find the difference between a sample and population mean. Used when the sample data is normally distributed or is large enough for the CLT to apply, and the standard dev. of the population is known.

1.3 (2 POINTS) T-Test

> T test is used when the std. dev of the population is unknown and the sample size is small

1.4 (2 POINTS) Man-Whitney U Test

> This test is used when the data is ordinal, non-normal, or as an alternative to an independent T-test.

1.5 (2 POINTS) ANOVA Test

> ANOVA test determines if there are significant differences between the means of three or more groups

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 calculated value based on tests, the p-value is the probability of getting results as deviant as what is expected, and is compared against the significance level to determine if the null hypothesis is rejected.

Q2) (2 POINTS) Create a DataFrame and Display

In [42]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

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.

In [43]:
```
df = pd.read_csv("colleges.csv")
display(df)
```

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	collegeTy
0	A leading global research university, MIT attr...	1	Massachusetts Institute of Technology	MA	12195	Urban	173700.0	-71.093539	42.359006	http://web.mit.edu	...	1861.0	MA	Private n for-pro
1	Stanford University sits just outside of Palo ...	2	Stanford University	CA	20961	Suburban	173500.0	-122.168924	37.431370	http://www.stanford.edu	...	1891.0	CA	Private n for-pro
2	One of the top public universities in the coun...	2	University of California, Berkeley	CA	45878	Urban	154500.0	-122.258393	37.869236	http://www.berkeley.edu	...	1868.0	CA	Pub
3	Princeton is a leading private research univer...	4	Princeton University	NJ	8532	Urban	167600.0	-74.659119	40.349855	http://www.princeton.edu	...	1746.0	NJ	Private n for-pro
4	Located in upper Manhattan, Columbia Universit...	5	Columbia University	NY	33882	Urban	148800.0	-73.961288	40.806515	http://www.columbia.edu	...	1754.0	NY	Private n for-pro
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
493	St. Joseph's College is a private institution ...	494	St. Joseph's College (NY)	NY	5901	Urban	100900.0	-73.968304	40.690548	http://www.sjcny.edu	...	1916.0	NY	Private n for-pro
494	A liberal arts college founded by the Moravian...	495	Moravian University	PA	2961	Urban	109800.0	-75.381596	40.630303	http://www.moravian.edu	...	1742.0	PA	Private n for-pro
495	Lawrence Technological University in Southfiel...	496	Lawrence Technological University	MI	3163	Urban	119900.0	-83.278458	42.450606	http://https://www.ltu.edu	...	NaN	MI	Private n for-pro
496	Saint Martin's University in Lacey, WA, one of...	497	Saint Martin's University	WA	1980	Urban	102100.0	NaN	NaN	NaN	...	NaN	WA	Private n for-pro
497	The University of Memphis is a large public re...	498	University of Memphis	TN	25128	Urban	90700.0	-89.939618	35.118453	http://www.mephis.edu	...	1912.0	TN	Pub

498 rows × 25 columns

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

In [44]:
```
df.dropna(inplace=True)
display(df)
```

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	collegeType
0	A leading global research university, MIT attr...	1	Massachusetts Institute of Technology	MA	12195	Urban	173700.0	-71.093539	42.359006	http://web.mit.edu	...	1861.0	MA	Private not for-profi
1	Stanford University sits just outside of Palo ...	2	Stanford University	CA	20961	Suburban	173500.0	-122.168924	37.431370	http://www.stanford.edu	...	1891.0	CA	Private not for-profi
2	One of the top public universities in the coun...	2	University of California, Berkeley	CA	45878	Urban	154500.0	-122.258393	37.869236	http://www.berkeley.edu	...	1868.0	CA	Publi
3	Princeton is a leading private research univer...	4	Princeton University	NJ	8532	Urban	167600.0	-74.659119	40.349855	http://www.princeton.edu	...	1746.0	NJ	Private not for-profi
4	Located in upper Manhattan, Columbia Universit...	5	Columbia University	NY	33882	Urban	148800.0	-73.961288	40.806515	http://www.columbia.edu	...	1754.0	NY	Private not for-profi
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
490	Loyola University New Orleans provides student...	491	Loyola University New Orleans	LA	4972	Urban	102300.0	-90.077714	29.953690	http://www.loyno.edu	...	1904.0	LA	Private not for-profi
491	Xavier University is a Jesuit Catholic school ...	492	Xavier University	OH	8079	Urban	104900.0	-84.476379	39.149037	http://www.xavier.edu	...	1831.0	OH	Private not for-profi
493	St. Joseph's College is a private institution ...	494	St. Joseph's College (NY)	NY	5901	Urban	100900.0	-73.968304	40.690548	http://www.sjcny.edu	...	1916.0	NY	Private not for-profi
494	A liberal arts college founded by the Moravian...	495	Moravian University	PA	2961	Urban	109800.0	-75.381596	40.630303	http://www.moravian.edu	...	1742.0	PA	Private not for-profi
497	The University of Memphis	498	University of Memphis	TN	25128	Urban	90700.0	-89.939618	35.118453	http://www.mephis.edu	...	1912.0	TN	Publi

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	collegeTyp
	is a large public re...													

422 rows × 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 $\alpha = 0.05$.

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.

In [45]:
```
df_cont = pd.crosstab(df["region"], df["collegeType"])
display(df_cont)
```

collegeType	Private not-for-profit	Public
region		
Midwest	54	37
Northeast	109	39
South	37	57
West	37	52

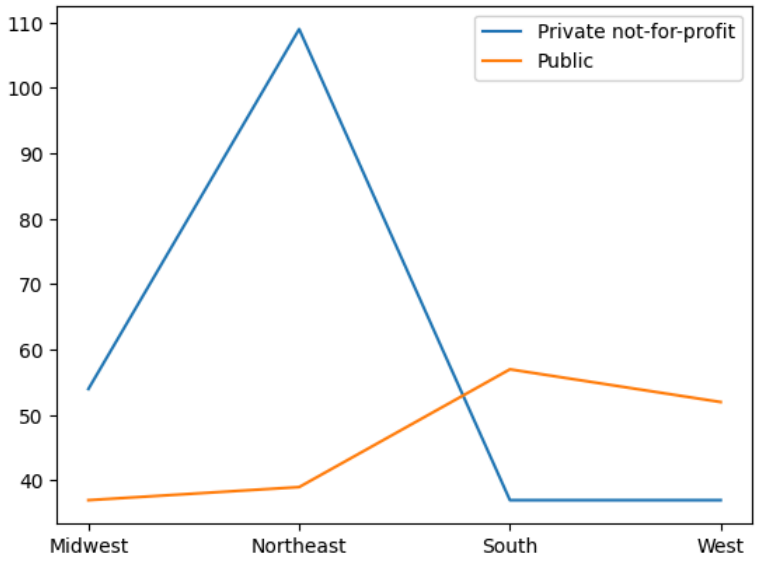
TASK 3.2 (2 POINTS): Why would we consider using a chi-squared test specifically (as opposed to some other hypothesis test)?

> We are analyzing frequencies in a table, and the chi-squared test analyzes expected vs. actual frequencies in a table.

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

In [46]:
```
plt.plot(df_cont["Private not-for-profit"], label="Private not-for-profit")
plt.plot(df_cont["Public"], label="Public")
plt.legend()
```

Out[46]:
```
<matplotlib.legend.Legend at 0x21e73dadd00>
```



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

> There is a noticeable difference in types of colleges per region: the northeast has the highest amount of private universities, while the south and west have the least.

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.

In [47]:
```
import scipy
display(scipy.stats.contingency.chi2_contingency(df_cont).pvalue)
```

```
np.float64(4.1236859547961256e-08)
```

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

> The p-value is extremely small, far smaller than the signifiance level, meaning that the null hypothesis is rejected.

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:

> Null hypothesis (HO) = The campus setting has no effect on total student population. HA = The campus setting has an effect on total student population.

Q6) (7 POINTS) Hypothesis Testing

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

In [48]:
```
for setting, setting_df in df.groupby(["campusSetting"]):
    display(setting_df)
```

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	collegeTy
6	Located in rural Williamstown, MA, Williams Co...	7	Williams College	MA	2307	Rural	152600.0	-73.208078	42.712389	http://www.williams.edu	...	1793.0	MA	Private n for-pr
13	The smallest Ivy League school, Dartmouth Coll...	14	Dartmouth College	NH	7171	Rural	161300.0	-72.289499	43.700465	http://www.dartmouth.edu	...	1769.0	NH	Private n for-pr
43	Colgate University is a leading liberal arts s...	44	Colgate University	NY	3112	Rural	154400.0	-75.536415	42.821191	http://www.colgate.edu	...	1819.0	NY	Private n for-pr
47	Located in the town of Brunswick, ME, Bowdoin ...	48	Bowdoin College	ME	1973	Rural	145600.0	-69.963975	43.906764	http://www.bowdoin.edu	...	1794.0	ME	Private n for-pr
54	Middlebury College is a small private liberal ...	55	Middlebury College	VT	4616	Rural	138100.0	-73.167117	44.014999	http://www.middlebury.edu	...	1800.0	VT	Private n for-pr
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
445	One of six senior military colleges in the U.S...	446	University of North Georgia	GA	23141	Rural	97200.0	-83.986084	34.531943	http://www.ung.edu	...	2013.0	GA	Pu
448	John Brown University is a private institution...	449	John Brown University	AR	2749	Rural	87800.0	-94.558494	36.187260	http://www.jbu.edu	...	1919.0	AR	Private n for-pr
454	The University of South Dakota, a public unive...	455	University of South Dakota	SD	12276	Rural	93100.0	-96.925776	42.782510	http://www.usd.edu	...	1862.0	SD	Pu
468	Saint Mary's University of Minnesota is a smal...	469	Saint Mary's University of Minnesota	MN	6947	Rural	107100.0	-91.673367	44.045336	http://www.smumn.edu	...	1912.0	MN	Private n for-pr
484	Sam Houston State is a large, public research ...	485	Sam Houston State University	TX	24116	Rural	98600.0	-95.547926	30.714614	http://shsu.edu	...	1879.0	TX	Pu

61 rows × 25 columns

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	colle
1	Stanford University sits just outside of Palo ...	2	Stanford University	CA	20961	Suburban	173500.0	-122.168924	37.431370	http://www.stanford.edu	...	1891.0	CA	Priva fc
22	The second-oldest member of the University of ...	23	University of California, Davis	CA	41236	Suburban	134800.0	-121.747976	38.540631	http://www.ucdavis.edu	...	1908.0	CA	
23	A top liberal arts school, Amherst is located ...	24	Amherst College	MA	1940	Suburban	148700.0	-72.533204	42.370772	http://https://www.amherst.edu	...	1821.0	MA	Priva fc
26	A private research university, Washington Univ...	27	Washington University in St. Louis	MO	17893	Suburban	136000.0	-90.301291	38.647812	http://www.wustl.edu	...	1853.0	MO	Priva fc
28	This public research university of Charlottesv...	29	University of Virginia	VA	29237	Suburban	137300.0	-78.581033	38.078711	http://www.virginia.edu	...	1819.0	VA	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
472	Ohio Wesleyan University is a private liberal ...	473	Ohio Wesleyan University	OH	1695	Suburban	115500.0	-83.068078	40.295043	http://www.owu.edu	...	1842.0	OH	Priva fc
473	About 27 miles northwest of Philadelphia in Co...	474	Ursinus College	PA	1492	Suburban	123300.0	-75.458534	40.191492	http://www.ursinus.edu	...	1869.0	PA	Priva fc
475	The only Edmundite college in the world, Saint...	476	Saint Michael's College	VT	2689	Suburban	111600.0	-73.165081	44.492760	http://www.smcvt.edu	...	1904.0	VT	Priva fc
483	Southern Illinois University, Edwardsville ope...	484	Southern Illinois University Edwardsville	IL	15204	Suburban	97100.0	-89.944006	38.841447	http://www.siue.edu	...	1957.0	IL	
489	The College of Idaho was founded in 1891 with ...	490	College of Idaho	ID	1149	Suburban	113500.0	-116.675961	43.654855	http://www.collegeofidaho.edu	...	1884.0	ID	Priva fc

106 rows × 25 columns

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	collegeType
0	A leading global research university, MIT attr...	1	Massachusetts Institute of Technology	MA	12195	Urban	173700.0	-71.093539	42.359006	http://web.mit.edu	...	1861.0	MA	Private not for-profi
2	One of the top public universities in the coun...	2	University of California, Berkeley	CA	45878	Urban	154500.0	-122.258393	37.869236	http://www.berkeley.edu	...	1868.0	CA	Publi
3	Princeton is a leading private research univer...	4	Princeton University	NJ	8532	Urban	167600.0	-74.659119	40.349855	http://www.princeton.edu	...	1746.0	NJ	Private not for-profi
4	Located in upper Manhattan, Columbia Universit...	5	Columbia University	NY	33882	Urban	148800.0	-73.961288	40.806515	http://www.columbia.edu	...	1754.0	NY	Private not for-profi
5	The University of California, Los Angeles is t...	6	University of California, Los Angeles	CA	46947	Urban	137200.0	-118.437855	34.073903	http://ucla.edu	...	1919.0	CA	Publi
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
490	Loyola University New Orleans provides student...	491	Loyola University New Orleans	LA	4972	Urban	102300.0	-90.077714	29.953690	http://www.loyno.edu	...	1904.0	LA	Private not for-profi
491	Xavier University is a Jesuit Catholic school ...	492	Xavier University	OH	8079	Urban	104900.0	-84.476379	39.149037	http://www.xavier.edu	...	1831.0	OH	Private not for-profi
493	St. Joseph's College is a private institution ...	494	St. Joseph's College (NY)	NY	5901	Urban	100900.0	-73.968304	40.690548	http://www.sjcny.edu	...	1916.0	NY	Private not for-profi
494	A liberal arts college founded by the Moravian...	495	Moravian University	PA	2961	Urban	109800.0	-75.381596	40.630303	http://www.moravian.edu	...	1742.0	PA	Private not for-profi
497	The University	498	University of Memphis	TN	25128	Urban	90700.0	-89.939618	35.118453	http://www.mephis.edu	...	1912.0	TN	Publi

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...	yearFounded	stateCode	collegeTyp
	of Memphis is a large public re...													

255 rows × 25 columns

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

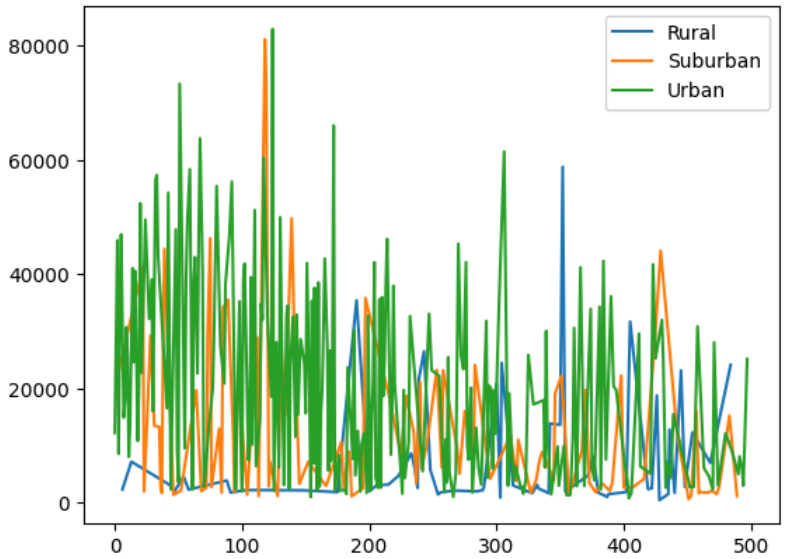
In [49]:
```
setting1, setting2, setting3 = [group_df for _, group_df in df.groupby(["campusSetting"])]
display(scipy.stats.f_oneway(setting1["totalStudentPop"], setting2["totalStudentPop"], setting3["totalStudentPop"]).pvalue)
```

```
np.float64(5.371995673405588e-09)
```

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

In [50]:
```
plt.plot(setting1["totalStudentPop"], label="Rural")
plt.plot(setting2["totalStudentPop"], label="Suburban")
plt.plot(setting3["totalStudentPop"], label="Urban")
plt.legend()
```

Out[50]:
```
<matplotlib.legend.Legend at 0x21e738b8320>
```



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

> The p-value is far less than 0.05, meaning that the null hypothesis should be rejected.

Q7) (2 POINTS) Post Hoc Tests

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

> We can determine which groups have the largest difference between the others.

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

In [51]:
```
# 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 $\alpha = 0.05$).

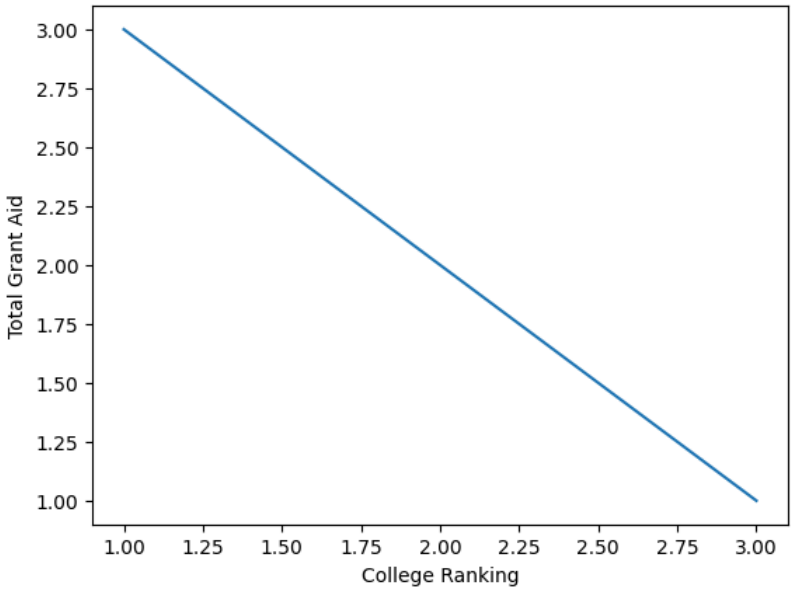
TASK 8.1 (3 POINTS): Write down the null and alternative hypotheses below.

> The null hypothesis is that the total grant aid has an impact on college ranking. The alternate hypothesis is that total grant aid has an impact on college ranking.

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

In [52]:
```
y = [3,2,1]
x = [1,2,3]

plt.plot(x,y)
plt.xlabel("College Ranking")
plt.ylabel("Total Grant Aid")
```

Out[52]:
```
Text(0, 0.5, 'Total Grant Aid')
```



TASK 8.3 (3 POINTS): Apply an appropriate hypothesis test and find the p-value of it. Only for this question, you are allowed to apply a hypothesis test that we haven't covered in the Hypothesis Testing class (hints: how about we consider finding some "relation" between them)?

In [53]:
```
display(scipy.stats.pearsonr(df["rank"], df["totalGrantAid"]).pvalue)
```

```
np.float64(1.6476144412424221e-19)
```

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

> The p value is far smaller than 0.05, meaning the null hypothesis should be rejected.

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?

> This cannot be concluded as all we know is that the null hypothesis should be rejected. We can input an alternative hypothesis to the pearson correlation test and see what the p-value comes to be.

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.

> Left-tail

In [54]:
```
private, public = [group_df for _, group_df in df.groupby(["collegeType"])]
display(scipy.stats.ttest_ind(private["studentFacultyRatio"], public["studentFacultyRatio"], alternative="less").pvalue)
```

```
np.float64(2.6905666569529506e-75)
```

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

> The p value is clearly less than 0.05, meaning the null hypothesis should be rejected.

THE END!