HW 1: Math Review and Plotting

Due Date: Mon 4/1, 11:59 PM

Collaboration Policy: You may talk with others about the homework, but we ask that you **write your solutions individually**. If you do discuss the assignments with others, please **include their names** in the following line.

Collaborators: *list collaborators here (if applicable)*

This Assignment

This homework is to help you diagnose your preparedness for the course. The rest of this course will assume familiarity with the programming and math concepts covered in this homework. Please consider reviewing prerequisite material if you struggle with this homework.

Score Breakdown

Question	Points
1	1
2a	1
2b	1
2c	1
2d	1
3a	4
3b	2
4a	2
4b	2

Question	Points
5	2
6a	2
6b	1
6c	1
7	5
Total	30

Here are some useful Jupyter notebook keyboard shortcuts. To learn more keyboard shortcuts, go to **Help -> Keyboard Shortcuts** in the menu above.

Here are a few we like:

1. `ctrl` + `return` : *Evaluate the current cell*
2. `shift` + `return` : *Evaluate the current cell and move to the next*
3. `esc` : *command mode* (may need to press before using any of the commands below)
4. `a` : *create a cell above*
5. `b` : *create a cell below*
6. `dd` : *delete a cell*
7. `m` : *convert a cell to markdown*
8. `y` : *convert a cell to code*

Initialize your environment

This cell should run without error if you have **set up your personal computer correctly**.

In [5]:
```
import numpy as np
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')

from IPython.display import display, Latex, Markdown
```

Python

Question 1 (1 pt)

Recall the formula for population variance below:

$$\textbf{Mean}(\mathbf{x}) = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\textbf{Var}(\mathbf{x}) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

Complete the functions below to compute the variance of `population`, an array of numbers. For this question, do not use built-in NumPy functions (i.e. `np.mean` and `np.var`); instead we will use NumPy to verify your code.

In [6]:
```
def mean(population):
    """
    Compute the mean of population (mu).

    Args:
        population: a numpy array of numbers of shape [N,]
    Returns:
        the mean of population (mu).
    """
    # Calculate the mean of a population
    # BEGIN YOUR CODE
    # ----------------------
    return np.sum(population) / len(population)
    # ----------------------
    # END YOUR CODE
```

```
def variance(population):
    """
    Compute the variance of population (sigma squared).

    Args:
        population: a numpy array of numbers of shape [N,]
    Returns:
        the variance of population
    """
    # Calculate the variance of a population
    # BEGIN YOUR CODE
    # ----------------------
    return np.sum((population - mean(population)) ** 2) / len(population)
    # ----------------------
    # END YOUR CODE
```

In [7]:
```
population_0 = np.random.randn(100)
assert np.isclose(mean(population_0), np.mean(population_0), atol=1e-6)
assert np.isclose(variance(population_0), np.var(population_0), atol=1e-6)
population_1 = 3 * np.random.randn(100) + 5
assert np.isclose(mean(population_1), np.mean(population_1), atol=1e-6)
assert np.isclose(variance(population_1), np.var(population_1), atol=1e-6)
```

NumPy

You should be able to understand the code in the following cells. If not, please review the following:

- UC Berkeley DS100 NumPy Review
- Stanford Condensed NumPy Review
- The Official NumPy Tutorial

Jupyter pro-tip: Pull up the docs for any function in Jupyter by running a cell with the function name and a `?` at the end:

In [8]:
```
np.arange?
```

You can close the window at the bottom by pressing `esc` several times.

Another Jupyter pro-tip: Pull up the docs for any function in Jupyter by typing the function name, then `<Shift>-<Tab>` on your keyboard. This is super convenient when you forget the order of the arguments to a function. You can press `<Tab>` multiple times to expand the docs and reveal additional information.

Try it on the function below:

In [9]:
```
np.linspace
```

Out[9]:
```
<function numpy.linspace(start, stop, num=50, endpoint=True, retstep=False, dtype=None, axis=0)>
```

Now, let's go through some linear algebra coding questions with NumPy. In this question, we'll ask you to use your linear algebra knowledge to fill in NumPy matrices. To conduct matrix multiplication in NumPy, you should write code like the following:

In [10]:
```
# A matrix in NumPy is a 2-dimensional NumPy array
matA = np.array([
    [1, 2, 3],
    [4, 5, 6],
])

matB = np.array([
    [10, 11],
    [12, 13],
    [14, 15],
])

# The notation B @ v means: compute the matrix multiplication Bv
matA @ matB
```

Out[10]:
```
array([[ 76,  82],
       [184, 199]])
```

You can also use the same syntax to do matrix-vector multiplication or vector dot products. Handy!

In [11]:
```
matA = np.array([
    [1, 2, 3],
    [4, 5, 6],
])
```

```
# A vector in NumPy is simply a 1-dimensional NumPy array
some_vec = np.array([ 10, 12, 14, ])
another_vec = np.array([ 10, 20, 30 ])

print(matA @ some_vec)
print(some_vec @ another_vec)
```

```
[ 76 184]
760
```

Question 2 (4 pt)

Question 2a

Joey, Deb, and Sam are shopping for fruit at K-Bowl. K-Bowl, true to its name, only sells fruit bowls. A fruit bowl contains some fruit and the price of a fruit bowl is the total price of all of its individual fruit.

Berkeley Bowl has apples for $2.00, bananas for \$1.00, and cantaloupes for $4.00 (expensive!). The price of each of these can be written in a vector:

$$v = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$$

K-Bowl sells the following fruit bowls:

1. 2 of each fruit
2. 5 apples and 8 bananas
3. 2 bananas and 3 cantaloupes
4. 10 cantaloupes

Create a 2-dimensional numpy array encoding the matrix B such that the matrix-vector multiplication

$$Bv$$

evaluates to a length 4 column vector containing the price of each fruit bowl. The first entry of the result should be the cost of fruit bowl #1, the second entry the cost of fruit bowl #2, etc.

In [12]:
```
v = np.array([2,1,4])

# BEGIN YOUR CODE
# ----------------------
B = np.array([[2, 2, 2],
              [5, 8, 0],
              [0, 2, 3],
              [0, 0, 10]])
# ----------------------
# END YOUR CODE

# The notation B @ v means: compute the matrix multiplication Bv
B @ v
```

Out[12]:
```
array([14, 18, 14, 40])
```

In [13]:
```
assert B.shape == (4, 3)
assert np.allclose(B @ v, np.array([14, 18, 14, 40]))
```

Question 2b

Joey, Deb, and Sam make the following purchases:

- Joey buys 2 fruit bowl #1s and 1 fruit bowl #2.
- Deb buys 1 of each fruit bowl.
- Sam buys 10 fruit bowl #4s (he really like cantaloupes).

Create a matrix A such that the matrix expression

$$ABv$$

evaluates to a length 3 column vector containing how much each of them spent. The first entry of the result should be the total amount spent by Joey, the second entry the amount sent by Deb, etc.

Note that the tests for this question do not tell you whether your answer is correct. That's up to you to determine.

In [14]:
```
A = np.array([
    [2, 1, 0, 0],
    # Finish this!
```

```
    # BEGIN YOUR CODE
    # ----------------------
    [1, 1, 1, 1],
    [0, 0, 0, 10]
    # ----------------------
    # END YOUR CODE
])

A @ B @ v
```

Out[14]:
```
array([ 46,  86, 400])
```

In [15]:
```
assert A.shape == (3, 4)
assert np.allclose(A @ B @ v , np.array([ 46,  86, 400]))
```

Question 2c

Who spent the most money? Assign `most` to a string containing the name of this person.

In [191…
```
# BEGIN YOUR CODE
# ----------------------
names = ["Joey", "Deb", "Sam"]
max_idx = np.argmax(A @ B @ v)
most = names[max_idx]
# ----------------------
# END YOUR CODE
```

In [192…
```
assert most in ["Joey", "Deb", "Sam"]
```

Question 2d

Let's suppose K-Bowl changes their fruit prices, but you don't know what they changed their prices to. Joey, Deb, and Sam buy the same quantity of fruit baskets and the number of fruit in each basket is the same, but now they each spent these amounts:

$$x = \begin{bmatrix} 80 \\ 80 \\ 100 \end{bmatrix}$$

Use `np.linalg.inv` and the above final costs to compute the new prices for the individual fruits as a vector called `new_v`.

In [18]:
```
# BEGIN YOUR CODE
# ----------------------
x = np.array([80, 80, 100])
new_v = np.linalg.inv(A @ B) @ x
# ----------------------
# END YOUR CODE
new_v
```

Out[18]:
```
array([5.5       , 2.20833333, 1.        ])
```

In [19]:
```
assert new_v.shape == (3,)
assert np.allclose(new_v, np.array([ 5.5,  2.20833333, 1.]))
```

Multivariable Calculus, Linear Algebra, and Probability

The following questions ask you to recall your knowledge of multivariable calculus, linear algebra, and probability. We will use some of the most fundamental concepts from each discipline in this class, so the following problems should at least seem familiar to you.

If you have trouble with these topics, we suggest reviewing:

- Khan Academy's Multivariable Calculus
- Khan Academy's Linear Algebra
- Khan Academy's Statistics and Probability

LaTeX

For the following problems, you should use LaTeX to format your answer. If you aren't familiar with LaTeX, not to worry. It's not hard to use in a Jupyter notebook. Just place your math in between dollar signs:

\$ f(x) = 2x \$ becomes $f(x) = 2x$.

If you have a longer equation, use double dollar signs to place it on a line by itself:

$\$ \sum_{i=0}^n i^2 $\$ becomes:

$$\sum_{i=0}^n i^2$$

.

Here is some handy notation:

Output	Latex
x^{a+b}	`x^{a + b}`
x_{a+b}	`x_{a + b}`
$\frac{a}{b}$	`\frac{a}{b}`
$\sqrt{a+b}$	`\sqrt{a + b}`
$\{\alpha, \beta, \gamma, \pi, \mu, \sigma^2\}$	`\{ \alpha, \beta, \gamma, \pi, \mu, \sigma^2 \}`
$\sum_{x=1}^{100}$	`\sum_{x=1}^{100}`
$\frac{\partial}{\partial x}$	`\frac{\partial}{\partial x}`
$\begin{bmatrix} 2x + 4y \\ 4x + 6y^2 \end{bmatrix}$	`\begin{bmatrix} 2x + 4y \\ 4x + 6y^2 \\ \end{bmatrix}`

For more about basic LaTeX formatting, you can read this article.

Question 3a (4 pt)

Suppose we have the following scalar-valued function:

$$f(x,y) = x^2 + 4xy + 2y^3 + e^{-3y} + \ln(2y)$$

Compute the partial derivative $\frac{\partial}{\partial x} f(x,y)$:

Answer: $2x + 4y$

Now compute the partial derivative $\frac{\partial}{\partial y} f(x,y)$:

Answer: $4x + 6y^2 - 3e^{-3y} + \frac{1}{y}$

Finally, using your answers to the above two parts, compute $\nabla f(x,y)$. Also what is the gradient at the point (x, y) = (2, -1):

Note that ∇ represents the gradient.

Answer:

$\nabla f(x,y) = (2x + 4y, 4x + 6y^2 - 3e^{-3y} + \frac{1}{y})$,

gradient at the point (x, y) = (2, -1) : $(0, 13 - 3e^3)$

Question 3b (2 pt)

Find the value(s) of x which minimizes the expression below. Justify why it is the minimum.

$\sum_{i=1}^{10} (i - x)^2$

Answer: Because $\sum_{i=1}^{10} (i - x)^2 = \sum_{i=1}^{10} (i^2 - 2ix + x^2) = 385 - 110x + 10x^2 = 10(x - \frac{11}{2})^2 + \frac{165}{2}$, $x = \frac{11}{2}$ is the value which minimizes the expression.

Question 4a (2 pt)

Let $\sigma(x) = \frac{1}{1+e^{-x}}$.

Show that $\sigma(-x) = 1 - \sigma(x)$.

Answer: $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$. Thus, $\sigma(-x) = \frac{1}{1+e^x} = \frac{1+e^x}{1+e^x} - \frac{e^x}{1+e^x} = 1 - \frac{e^x}{1+e^x} = 1 - \sigma(x)$

Question 4b (2 pt)

Show that the derivative can be written as:

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$$

Answer: $\frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})(1+e^{-x})} = \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\sigma(-x) = \sigma(x)(1-\sigma(x))$

Question 5 (2 pt)

Consider the following scenario:

Only 1% of 40-year-old women who participate in a routine mammography test have breast cancer. 80% of women who have breast cancer will test positive, but 9.6% of women who don't have breast cancer will also get positive tests.

Suppose we know that a woman of this age tested positive in a routine screening. What is the probability that she actually has breast cancer?

Hint: Use Bayes' rule.

Answer: Let A = have breast cancer, B = test positive.

$P(A) = 0.01, P(B|A) = 0.8, P(B|A^C) = 0.096$

We should solve $P(A|B)$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} = \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.09504} = 0.0776...$$

Thus, the answer is about 0.0776.

Question 6

Consider (once again) a sample of size n drawn at random with replacement from a population in which a proportion p of the individuals are called successes.

Let S be the random variable that denotes the number of successes in our sample. Then, the probability that the number of successes in our sample is **at most** s (where $0 \leq s \leq n$) is

$$P(S \leq s) = P(S = 0) + P(S = 1) + \ldots + P(S = s) = \sum_{k=0}^{s} \binom{n}{k} p^k (1-p)^{n-k}$$

We obtain this by summing the probability that the number of successes is exactly k, for each value of $k = 0, 1, 2, \ldots, s$.

Question 6a (2pt)

Please fill in the function `prob_at_most` which takes n, p, and s and returns $P(S \leq s)$ as defined above. If the inputs are invalid: for instance, if p > 1 or s > n then return 0."

Hint: One way to compute the binomial coefficients is to use SciPy module, which is a collection of Python-based software for math, probability, statistics, science, and enginnering. Feel free to use scipy.special.comb**

In [20]:
```
from scipy import special

def prob_at_most(n, p, s):
    """
    returns the probability of S <= s
    Input n: sample size; p : proportion; s: number of successes at most
    """
    if p > 1 or s > n:
        return 0
```

```
    # BEGIN YOUR CODE
    # ----------------------
    all_probs = [special.comb(n, i)*(p**i)*((1-p)**(n-i)) for i in range(s+1)]

    # ----------------------
    # END YOUR CODE

    return sum(all_probs[:s+1])
```

In [21]:
```
assert prob_at_most(3, 0.4, 1) >= 0
assert prob_at_most(5, 0.6, 3) <= 1
assert prob_at_most(2, 3, 4) == 0
```

Question 6b (1pt)

In an election, supporters of Candidate C are in a minority. Only 45% of the voters in the population favor the candidate.

Suppose a survey organization takes a sample of 200 voters at random with replacement from this population. Use `prob_at_most` to write an expression that evaluates to the chance that a majority (more than half) of the sampled voters favor Candidate C.

In [22]:
```
# BEGIN YOUR CODE
# ----------------------
p_majority = 1 - prob_at_most(200, 0.45, 100)
# ----------------------
# END YOUR CODE
p_majority
```

Out[22]:
```
0.06807524986263847
```

In [23]:
```
assert p_majority >= 0 and p_majority <= 1
```

Question 6c (1pt)

Suppose each of five survey organizations takes a sample of voters at random with replacement from the population of voters in Part **b**, independently of the samples drawn by the other organizations.

- Three of the organizations use a sample size of 200
- One organization uses a sample size of 300

- One organization uses a sample size of 400

Write an expression that evaluates to the chance that in at least one of the five samples the majority of voters favor Candidate C. You can use any quantity or function defined earlier in this exercise.

In [24]:
```
# BEGIN YOUR CODE
# ----------------------
prob_6c = 1 - (prob_at_most(200, 0.45, 100) ** 3 * prob_at_most(300, 0.45, 150) * prob_at_most(400, 0.45, 200))
# ----------------------
# END YOUR CODE
prob_6c
```

Out[24]:
```
0.23550361568442357
```

In [25]:
```
assert prob_6c >= 0 and prob_6c <= 1
```

The US Presidential Election

The US president is chosen by the Electoral College, not by the popular vote. Each state is alotted a certain number of electoral college votes, as a function of their population. Whomever wins in the state gets all of the electoral college votes for that state.

There are 538 electoral college votes (hence the name of the Nate Silver's site, FiveThirtyEight).

Pollsters correctly predicted the election outcome in 46 of the 50 states. For these 46 states Trump received 231 and Clinton received 232 electoral college votes.

The remaining 4 states accounted for a total of 75 votes, and whichever candidate received the majority of the electoral college votes in these states would win the election.

These states were Florida, Michigan, Pennsylvania, and Wisconsin.

State	Electoral College Votes
Florida	29
Michigan	16
Pennsylvania	20

State	Electoral College Votes
Wisconsin	10

For Donald Trump to win the election, he had to win either:

- Florida + one (or more) other states
- Michigan, Pennsylvania, and Wisconsin

The electoral margins were very narrow in these four states, as seen below:

State	Trump	Clinton	Total Voters
Florida	49.02	47.82	9,419,886
Michigan	47.50	47.27	4,799,284
Pennsylvania	48.18	47.46	6,165,478
Wisconsin	47.22	46.45	2,976,150

Those narrow electoral margins can make it hard to predict the outcome given the sample sizes that the polls used.

Question 7a (1pt)

For your convenience, the results of the vote in the four pivotal states is repeated below:

State	Trump	Clinton	Total Voters
Florida	49.02	47.82	9,419,886
Michigan	47.50	47.27	4,799,284
Pennsylvania	48.18	47.46	6,165,478
Wisconsin	47.22	46.45	2,976,150

Using the table above, write a function `draw_state_sample(N, state)` that returns a sample with replacement of N voters from the given state. Your result should be returned as a list, where the first element is the number of Trump votes, the second element is the number of Clinton

votes, and the third is the number of Other votes. For example, `draw_state_sample(1500, "florida")` could return `[727, 692, 81]`. You may assume that the state name is given in all lower case.

You might find `np.random.multinomial` useful.

In [26]:
```
def draw_state_sample(N, state):
    # BEGIN YOUR CODE
    # ----------------------
    l = ['florida', 'michigan', 'pennsylvania', 'wisconsin']
    idx = l.index(state)
    arr = np.array([
        [0.4902, 0.4782, 1 - 0.4902 - 0.4782],
        [0.4750, 0.4727, 1 - 0.4750 - 0.4727],
        [0.4818, 0.4746, 1 - 0.4818 - 0.4746],
        [0.4722, 0.4645, 1 - 0.4722 - 0.4645]
    ])
    return np.random.multinomial(N, arr[idx])
    # ----------------------
    # END YOUR CODE
```

In [31]:
```
assert len(draw_state_sample(1500, "florida")) == 3
assert sum(draw_state_sample(1500, "michigan")) == 1500
q7a_penn = draw_state_sample(1500, "pennsylvania")
trump_win_penn = (q7a_penn[0] - q7a_penn[1]) / 1500
abs(trump_win_penn - 0.007) <= 0.12
```

Out[31]:
```
True
```

Question 7b (1pt)

Now, create a function `trump_advantage` that takes in a sample of votes (like the one returned by `draw_state_sample`) and returns the difference in the proportion of votes between Trump and Clinton. For example `trump_advantage([100, 60, 40])` would return `0.2`, since Trump had 50% of the votes in this sample and Clinton had 30%.

In [33]:
```
def trump_advantage(voter_sample):
    # BEGIN YOUR CODE
    # ----------------------
    diff = voter_sample[0] - voter_sample[1]
    return diff / sum(voter_sample)
```

```
    # ----------------------
    # END YOUR CODE
```

In [36]:
```
assert -1 < trump_advantage(draw_state_sample(1500, "wisconsin")) < 1
assert np.isclose(trump_advantage([100, 60, 40]), 0.2)
assert np.isclose(trump_advantage([10, 30, 10]), -0.4)
```

Question 7c (1pt)

Simulate Trump's advantage across 100,000 samples of 1500 voters for the state of Pennsylvania and store the results of each simulation in a list called `simulations`.

That is, `simulations[i]` should be Trump's proportion advantage for the `i+1` th simple random sample.

In [48]:
```
# BEGIN YOUR CODE
# ----------------------
simulations = list(map(lambda x: trump_advantage(x), [draw_state_sample(1500, 'pennsylvania') for _ in range(100000)]))
# ----------------------
# END YOUR CODE
```

In [50]:
```
assert len(simulations) == 100000
assert sum([-1 < x < 1 for x in simulations]) == len(simulations)
assert abs(np.mean(simulations) - 0.007) <= 0.016
```

Question 7d (1pt)

Now write a function `trump_wins(N)` that creates a sample of N voters for each of the four crucial states (Florida, Michigan, Pennsylvania, and Wisconsin) and returns 1 if Trump is predicted to win based on these samples and 0 if Trump is predicted to lose.

Recall that for Trump to win the election, he must either:

- Win the state of Florida and 1 or more other states
- Win Michigan, Pennsylvania, and Wisconsin

In [162…
```
def trump_wins(N):
    # BEGIN YOUR CODE
    # ----------------------
    voters = 0
```

```
    l = ['florida', 'michigan', 'pennsylvania', 'wisconsin']
    ECVotes = [29, 16, 20, 10]
    for idx, state in enumerate(l):
        tmp = draw_state_sample(N, state)
        maximum = max(tmp)
        if maximum == tmp[0]:
            voters += ECVotes[idx]
    if voters > sum(ECVotes)/2 :
        return 1
    else:
        return 0
    # ----------------------
    # END YOUR CODE
```

In [173…
```
assert trump_wins(1000) in [0, 1]
```

Question 7e

If we repeat 100,000 simulations of the election, i.e. we call `trump_wins(1500)` 100,000 times, what proportion of these simulations predict a Trump victory? Give your answer as `proportion_trump`.

This number represents the percent chance that a given sample will correctly predict Trump's victory *even if the sample was collected with absolutely no bias*.

In [184…
```
# BEGIN YOUR CODE
# ----------------------
l = [trump_wins(1500) for _ in range(100000)]
proportion_trump = sum(l) / len(l)
# ----------------------
# END YOUR CODE
proportion_trump
```

Out[184]:
```
0.70737
```

In [186…
```
assert 0 < proportion_trump < 1
assert abs(proportion_trump - 0.695) <= 0.02
```

Congratulations! You have completed HW1.

Make sure you have run all cells in your notebook in order before running the cell below, so that all images/graphs appear in the output.,

Please save before submitting!

Please generate pdf as follows and submit it to Gradescope.

File > Print Preview > Print > Save as pdf