Data 100/200 Homework 1 Written

Jackie Hu

TOTAL POINTS

19.5 / 27

QUESTION 1

Question 1 7 pts

1.1 1a 1 / 1

✓ **+ 1 pts** **Valid algebraic proof**

+ 0 pts Incorrect, incomplete, blank, or not enough work shown

1.2 1b 1 / 1

✓ **+ 1 pts** **Correct expression for variance (OK if denominator is $$n-1$$)**

+ 0 pts Incorrect/blank

1.3 1c 2 / 2

✓ **+ 2 pts** **Correct expression for MSE**

+ 1 pts Partial credit (e.g. forgot to square $$x_i - c$$)

+ 0 pts Incorrect/blank

1.4 1d 1 / 3

+ 3 pts Correct

+ 2 pts Correctly used the hint to decompose $$x_i - c$$ into $$(x_i - \mu) + (\mu - c)$$ and split the sum, but no further progress (does not recognize that the middle sum simplifies to $$0$$)

✓ **+ 1 pts** **Correctly used the hint to decompose $$x_i - c$$ into $$(x_i - \mu) + (\mu - c)$$**

+ 0 pts Incorrect/blank

QUESTION 2

Question 2 3 pts

2.1 2a 1 / 1

✓ **+ 1 pts** **Correct**

+ 0 pts Incorrect/blank

+ 0.5 pts Minor Error (wrong deduction, skep steps, no conclusions, incomplete answers etc.)

2.2 2b 2 / 2

✓ **+ 2 pts** **Correct**

+ 1 pts Made initial progress or had some algebraic errors

+ 0 pts Incorrect/blank

QUESTION 3

6 pts

3.1 3b 2 / 2

✓ **+ 2 pts** **Correct plot**

+ 1 pts Plot peaks at $$0.6$$ instead of $$0.4$$

+ 0 pts Incorrect/blank

3.2 3c 1 / 1

✓ **+ 1 pts** **Correct interpretation as the observed proportion**

+ 0 pts Incorrect/blank

3.3 3d 1 / 1

✓ **+ 1 pts** **Correct argument based on the statement that $$\log$$ is a monotonically increasing function**

+ 0 pts Incorrect/blank

3.4 3e 0 / 2

+ 2 pts Correct

+ 1 pts Takes correct derivative, but makes an algebraic error when solving for $$\hat{p}$$

✓ **+ 0 pts** **Incorrect/blank**

QUESTION 4

Question 4 4 pts

4.1 4a 1 / 1

✓ + 1 pts **Correct (we only have measures of confidence in each category separately)**

+ 0 pts Incorrect/blank

4.2 4d 0.5 / 2

+ 2 pts Correct

+ 1 pts Argues that (i) and (ii) are incorrect because they cannot be true for all values of $$p$$, but does not argue that (iv) is correct

+ 0 pts Incorrect/blank

+ 0.5 Point adjustment

Please explain your answer through algebraic or probabilistic reasoning next time for full credit!

4.3 4e 0 / 1

+ 1 pts Correct

✓ **+ 0 pts Incorrect/blank**

QUESTION 5

Question 5 7 pts

5.1 5a 2 / 2

✓ **+ 2 pts Correct**

+ 1 pts Not centered at integers, does not have white edges, or other minor error

+ 0 pts Incorrect/blank

5.2 5c 1 / 1

✓ **+ 1 pts Correct**

+ 0 pts Incorrect/blank

5.3 5d 2 / 2

✓ **+ 2 pts Correct answer using the fact that the distribution is roughly normal (may be implied, e.g. 68% being 1 std, or the point of inflection)**

+ 1 pts Correct answer without mentioning the distribution being roughly normal

+ 0 pts Incorrect/blank

5.4 5e 1 / 2

+ 2 pts Correct (using the sigma found in 5.d.)

✓ **+ 1 pts Small error in implementing the Gaussian PDF**

+ 0 pts Incorrect/blank

0.0.1 Question 1

Let $x_1, x_2, \ldots, x_n$ be a list of numbers. You can think of each index i as the label of a household, and the entry x_i as the annual income of Household i. Define the *mean* or *average* of the list to be $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Question 1a) The ith *deviation from average* is the difference $x_i - \mu$. In Data 8 you saw in numerical examples that the sum of all these deviations is 0. Now prove that fact. That is, show that $\sum_{i=1}^{n}(x_i - \mu) = 0$.

Note: In this class, you must always put your answer in the cell that immediately follows the question. DO NOT create any cells between this one and the one that says *Write your answer here, replacing this text.*

$\sum_{i=1}^{n}(x_i - \mu) = (x_1 + x_2 + x_3 + ... + x_n) - (\mu + \mu + \mu + ... + \mu) = (x_1 + x_2 + x_3 + ... + x_n) - n(\mu)$
$= (x_1 + x_2 + x_3 + ... + x_n) - n(\frac{1}{n}\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$

Question 1b) Recall that the *variance* of a list is defined as the *mean squared deviation from average*, and that the *standard deviation* (SD) of the list is the square root of the variance. The SD is in the same units as the data and measures the rough size of the deviations from average.

Denote the variance of the list by σ^2. Write a math expression for σ^2 in terms of the data $(x_1 \dots x_n)$ and μ. We recommend building your expression by reading the definition of variance from right to left. That is, start by writing the notation for "average", then "deviation from average", and so on.

$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$

Question 1c) Suppose you have to predict the value of x_i for some i, but you don't get to see i and you certainly don't get to see x_i. You decide that whatever x_i is, you're just going to use your favorite number μ as your predictor.

The *error* in your prediction is $x_i - \mu$, which is your old friend the deviation from average. Thus the *mean squared error* (MSE) of your predictor μ over the entire list of n data points is the mean squared deviation from average, which is your old friend the variance. So we will write $\sigma^2 = MSE(\mu)$.

Now suppose I decide that whatever x_i is, I'm just going to use *my* favorite number as my predictor, and my favorite number is c. Write a math expression for $MSE(c)$. Again, go from right to left: first c, then the error, and so on.

$MSE(c) = \frac{\sum_{i=1}^{n}(x_i - c)^2}{n}$

Question 1d) Whose predictor is better? It seems reasonable to guess that your predictor μ is better than my favorite but possibly weird c. Show that $MSE(c) > MSE(\mu)$ for all $c \neq \mu$, by the method indicated below.

- Write the error $x_i - c$ as $x_i - c = (x_i - \mu) + (\mu - c)$.
- Substitute this expression for $x_i - c$ in your formula for $MSE(c)$.
- Expand the square and use properties of sums; don't forget what you showed in Part **a**.

This shows that μ is the *least squares* constant predictor. In Data 8 you found (numerically) the least squares linear predictor of a variable y based on a related variable x. We will return to that later in this course, using a generalization of the calculation in this exercise.

$$MSE(c) = \frac{\sum_{i=1}^{n}((x_i-\mu)+(\mu-c))^2}{n} = \frac{\sum_{i=1}^{n}(x_i-c)^2}{n} = \frac{\sum_{i=1}^{n}(x_i^2-2x_ic+c^2)}{n}$$

$$MSE(\mu) = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} = \frac{\sum_{i=1}^{n}(0)}{n} = 0$$

$$MSE(c) > MSE(\mu)$$

0.0.2 Question 2

In this question we will review some fundamental properties of the sigmoid function, which will be discussed when we talk more about logistic regression in the latter half of the class. The sigmoid function is defined to be

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

Question 2a) Show that $\sigma(-x) = 1 - \sigma(x)$

$$\begin{aligned}\sigma(x) &= \frac{e^x}{1+e^x} \\ &= \frac{e^x + 1 - 1}{1+e^x} \\ &= \frac{e^x+1}{e^x+1} - \frac{1}{1+e^x} \\ &= 1 - \sigma(-x)\end{aligned}$$

therefore

$$1 - \sigma(x) = \sigma(-x)$$

Question 2b) Show that the derivative of the sigmoid function can be written as:

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

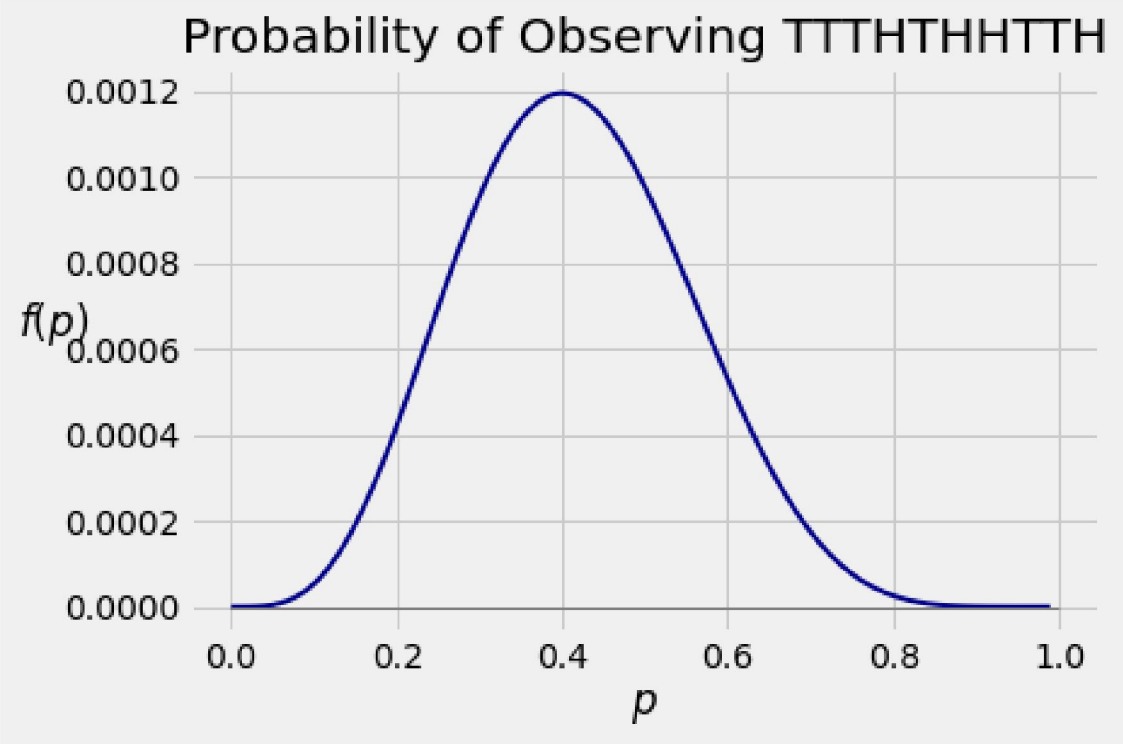
$$\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) \\
&= \frac{d}{dx}\left(1+\mathrm{e}^{-x}\right)^{-1} \\
&= -(1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \frac{e^{-x}}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}} \\
&= \frac{(1+e^{-x})-1}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}} \\
&= \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) \cdot \frac{1}{1+e^{-x}} \\
&= \left(1 - \frac{1}{1+e^{-x}}\right) \cdot \frac{1}{1+e^{-x}} \\
&= (1-\sigma(x)) \cdot \sigma(x) \\
&= \sigma(x) \cdot (1-\sigma(x))
\end{aligned}$$

Question 3b) I have a coin that lands heads with an unknown probability p. I toss it 10 times and get the sequence TTTHTHHTTH.

If you toss this coin 10 times, the chance that you get the sequence above is a function of p. That function, which we will call f, determines how likely the sequence TTTHTHHTTH is depending on p

Plot the graph of f as a function of p for $p \in [0, 1]$.

```
In [3]: p = np.arange(0,1,0.01)
        f = ((1-p) ** 6 )* (p ** 4)
        plt.plot(p, f, lw=2, color='darkblue') # lw is line width
        plt.plot([0, 1], [0, 0], lw=1, color='grey')    # horizontal axis
        plt.xlabel('$p$')
        plt.ylabel('$f(p)$', rotation=0)
        plt.title('Probability of Observing TTTHTHHTTH');
```



Question 3c) Among all values of p, we want to find the one that makes the observed data the most likely, which we will call $\hat{p}$. Please provide the value of $\hat{p}$ and also a simple interpretation of that value in terms of the data TTTHTHHTTH.

$\hat{p} = 0.4$

The probability of getting the head that will have the maximum probability to get TTTHTHHTTH is 0.4. We want 6 times on tails, 4 times on head, therefore a pobability of slightly lower than 0.5 for lading on head will give us the highest probability.

Question 3d) Explain why the value, $\hat{p}$, at which the function f attains its maximum is the same as the value at which the function $\log(f)$ attains its maximum. To clarify, $\log(f)$ is the composition of log and f: $\log(f)$ at p is $\log(f(p))$. Even though it doesn't make a difference for this problem, log is now and forevermore the log to the base e, not to the base 10.

It might help to compare $\log(x_1)$ and $\log(x_2)$ for $x_1 < x_2$.

The observation in this exercise is hugely important in data science because many probabilities are products and the log function turns products into sums. It's much simpler to work with a sum than with a product.

$f(\hat{p}) > f(p)$ for all p

Since log is increasing

$\log(f(\hat{p})) > \log(f(p))$ for all p

So $\hat{p}$ Maximizes $\log(f)$ as well

Question 3e) Instead of using the graph, this time use Part c and calculus to find $\hat{p}$. Using Part d makes the calculus much easier. You don't have to check that the value you've found produces a max and not a min – we'll spare you that step.

$f(p) = p^4(1-p)^6$

$\log(f(p)) = 4p + 6(1-p) = 4p + 6 - 6p = 6 - 2p = 0.3$

0.0.3 Question 4

Much of data analysis involves interpreting proportions – lots and lots of related proportions. So let's recall the basics. It might help to start by reviewing the main rules from Data 8, with particular attention to what's being multiplied in the multiplication rule.

Question 4a) The Pew Research Foundation publishes the results of numerous surveys, one of which is about the trust that Americans have in groups such as the military, scientists, and elected officials to act in the public interest. A table in the article summarizes the results.

Pick one of the options (i) and (ii) to answer the question below; if you pick (i), fill in the blank with the percent. Then, explain your choice.

The percent of surveyed U.S. adults who had a great deal of confidence in both scientists and religious leaders

(i) is equal to _____.

(ii) cannot be found with the information in the article.

(ii) we need the percent of the ones who had a great deal of condidence in the religious leaders given that they trust scientists, or vice versa.

0.0.4

Question 4d) (This part is a continuation of the previous two.) Pick all of the options (i)-(iv) that are true for all values of p. Explain by algebraic or probailistic reasoning; you are welcome to use your function `no_disease_given_negative` to try a few cases numerically. Your explanation should include the reasons why you *didn't* choose some options.

$P(N \mid T_N)$ is

(i) equal to 0.95.

(ii) equal to 0.999×0.95.

(iii) greater than 0.999×0.95.

(iv) greater than 0.95.

(iii)(iv)

Question 4e) Suzuki is one of most commonly owned makes of cars in our county (Alameda). A car heading from Berkeley to San Francisco is pulled over on the freeway for speeding. Suppose I tell you that the car is either a Suzuki or a Lamborghini, and you have to guess which of the two is more likely.

What would you guess, and why? Make some reasonable assumptions and explain them (data scientists often have to do this), justify your answer, and say how it's connected to the previous parts.

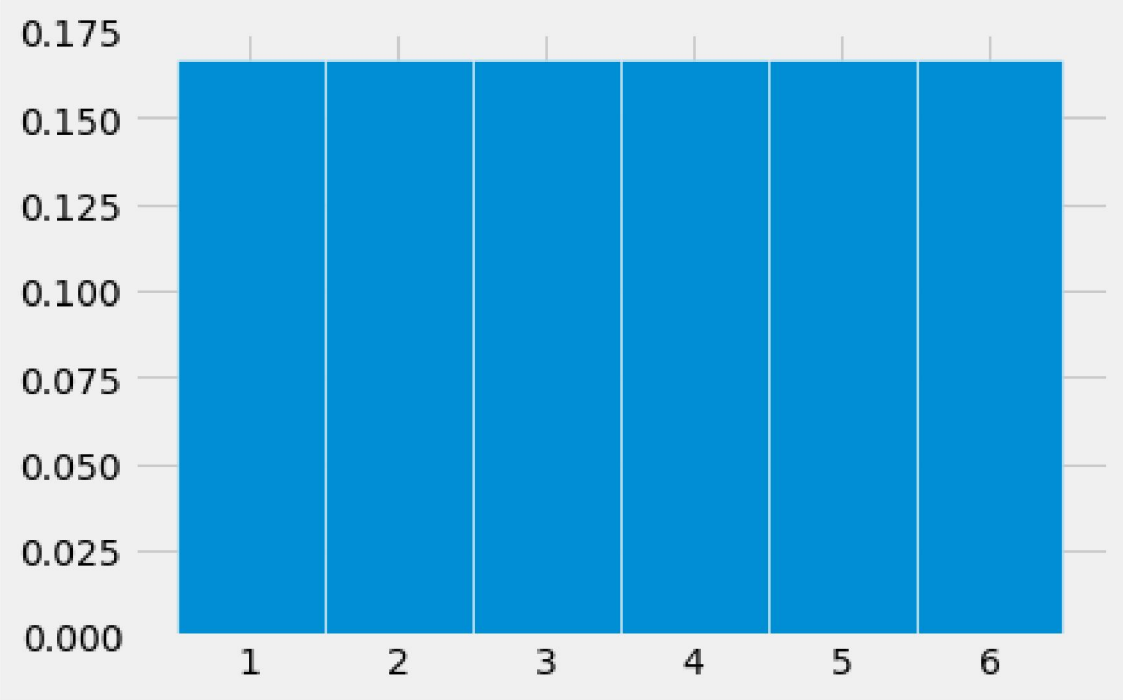
Equally possibility, since there's only 2 options given for this specific car that's being pulled over. The previous message about Suzuki being the most common makes in the county is not useful anymore.

Question 5a) Define a function `integer_distribution` that takes an array of integers and draws the histogram of the distribution using unit bins centered at the integers and white edges for the bars. The histogram should be drawn to the density scale. The left-most bar should be centered at the smallest integer in the array, and the right-most bar at the largest.

Your function does not have to check that the input is an array consisting only of integers. The display does not need to include the printed proportions and bins.

If you have trouble defining the function, go back and carefully read all the lines of code that resulted in the probability histogram of the number of spots on one roll of a die. Pay special attention to the bins.

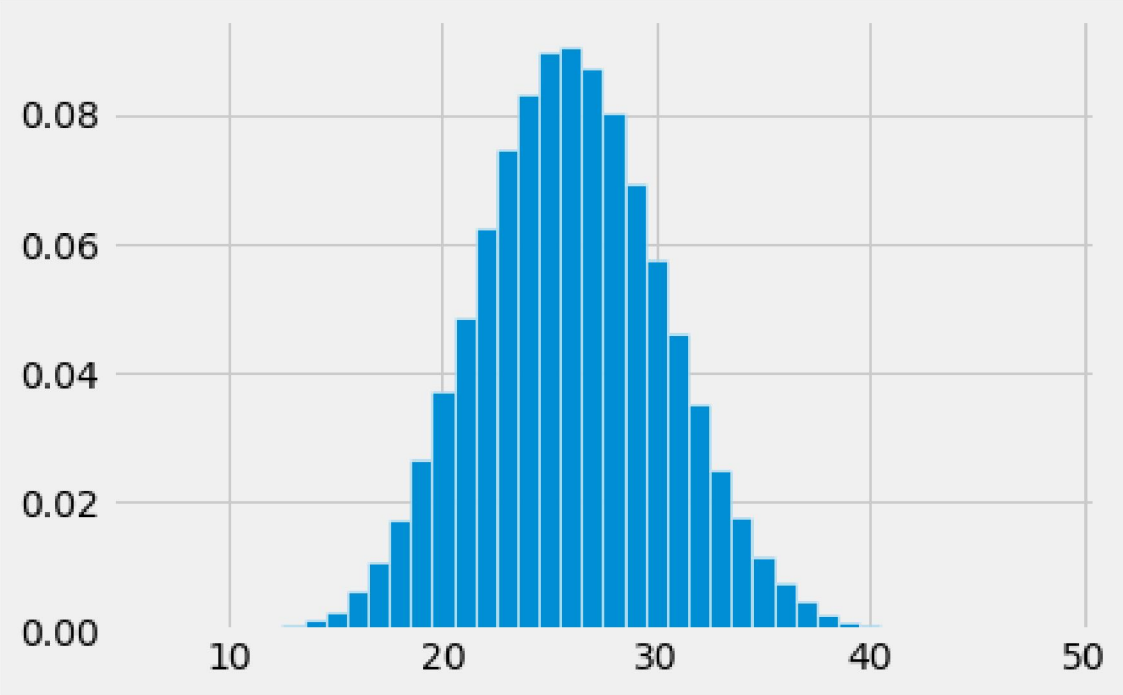
```
In [34]: def integer_distribution(x):
             ...

             unit_bins = np.arange(min(x) - 0.5, max(x) + 0.6)

             plt.hist(x, bins = unit_bins, ec='white', density=True)

         integer_distribution(faces)
```



Question 5c) Replace the "..." in the code cell below with a Python expression so that the output of the cell is an empirical histogram of 500,000 simulated counts of black people in 100 draws made at random with replacement from the population eligible for Swain's jury panel.

After you have drawn the histogram, you might want to take a moment to recall the conclusion reached in Data 8 based on the data that Swain's panel had 8 black people in it.

```
In [57]: sample2 = np.random.multinomial(100, [0.26,0.74],500000)[:,[0]]
         simulated_counts = sample2.reshape(1, 500000) [0]
         integer_distribution(simulated_counts)
```



Question 5d) As you know, the count of black people in a sample of 100 people drawn at random from the eligible population is expected to be 26. Just by looking at the histogram in Part **c**, and **no other calculation**, pick the correct option and explain your choice. You might want to refer to the Data 8 textbook again.

The SD of the distribution of the number of black people in a random sample of 100 people drawn from the eligible population is closest to

(i) 1.4

(ii) 4.4

(iii) 7.4

(iv) 10.4

(ii) the mean is around 25 - 26, the first "dip" is around 31-32, the closest guess for the SD is 4.4.

Question 5e) The *normal curve with mean μ and SD σ* is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

Redraw your histogram from Part **c** and overlay the normal curve with $\mu = 26$ and σ equal to the choice you made in Part **d**. You just have to call `plt.plot` after `integer_distribution`. Use `np.e` for e. For the curve, use 2 as the line width, and any color that is easy to see over the blue histogram. It's fine to just let Python use its default color.

Now you can see why centering the histogram bars over the integers was a good idea. The normal curve peaks at 26, which is the center of the corresponding bar.

```
In [82]: import math
         mu = 26
         sigma = 4.4
         x = np.linspace(0, 50, 200)
         f_x = 1/ math.sqrt(2 * math.pi)/ sigma * np.e** (-0.5 *((((x - mu) ** 2)/(sigma))**2))

         integer_distribution(simulated_counts)

         plt.plot(x, f_x, linewidth=2)
```

```
Out[82]: [<matplotlib.lines.Line2D at 0x7f1ca350c880>]
```