0.1 Question 1: Calculus

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

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

0.1.1 Question 1a

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

Note, again: 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 *Type your answer here, replacing this text.*

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

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

0.1.2 Question 1b

Show that the derivative of the sigmoid function can be written as:

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

This PDF has some handy LaTeX.

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

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.

0.1.3 Question 2a

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) This information cannot be found in the source.

The question posed askes us to find the percent of adults who had a great deal of confidence in both scientists and religious leaders, so $P(S \cap R)$. However, while we do have the information on what percent of people have a great deal of confidence in scientists and religious leaders individually, we do not know and are not offered the intersection of those two events nor are we offered the probability that one event happens given the another event happened. From the multiplication rule we know that $P(S \cap R) = P(S) \cdot P(R|S)$. We have P(S) but we cannot assume the $P(R|S)$ from the table.

0.1.4 Question 2d

(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.

(iv) is true because when the true positive rate is at it's lowest (0) the P(N | Tn) is still greater than 0.95. Likewise, when the true positive rate is at its highest (1) then P(N | Tn) is 1. Therefore, P(N | Tn) must be between what P(N | Tn) is when p = 0 and 1. Testing the case where the p = 0 shows that the minimun P(N | Tn) could be is approximately 0.99894. Then, (iii) is true, (i) is not true, and (ii) is also not true.

0.1.5 Question 2e

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.

Despite the Lamborghini being flashier and notoriously fast cars, I would have to say the Suzuki because the probability of Suzuki's in the county is likely way higher than that of the Lambo. We need to take into account the base rate of the suzuki because it also affects P(Suzuki | Speeding).

0.1.6 Question 3a

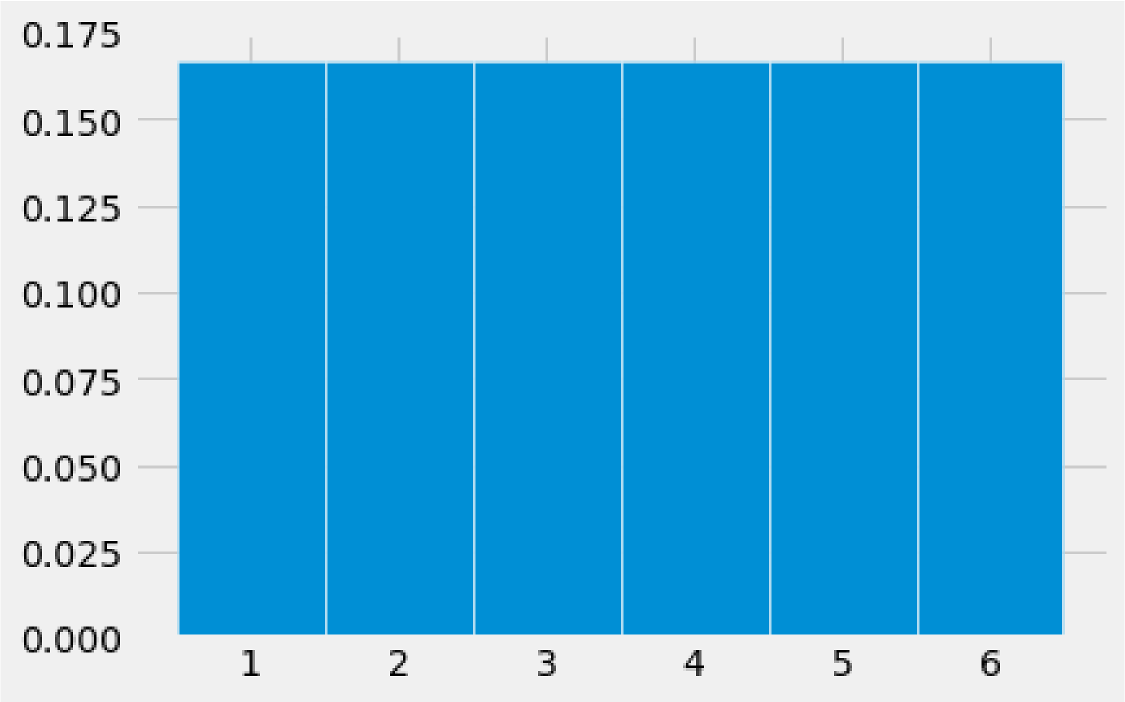
Define a function `integer_distribution` that takes an array of integers and draws the histogram of the distribution using unit bins centered at the integers and white edges for the bars. The histogram should be drawn to the density scale. The left-most bar should be centered at the smallest integer 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 [15]: def integer_distribution(x):
             #takes array of integers and draws the histogram of the distribution
             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)
```

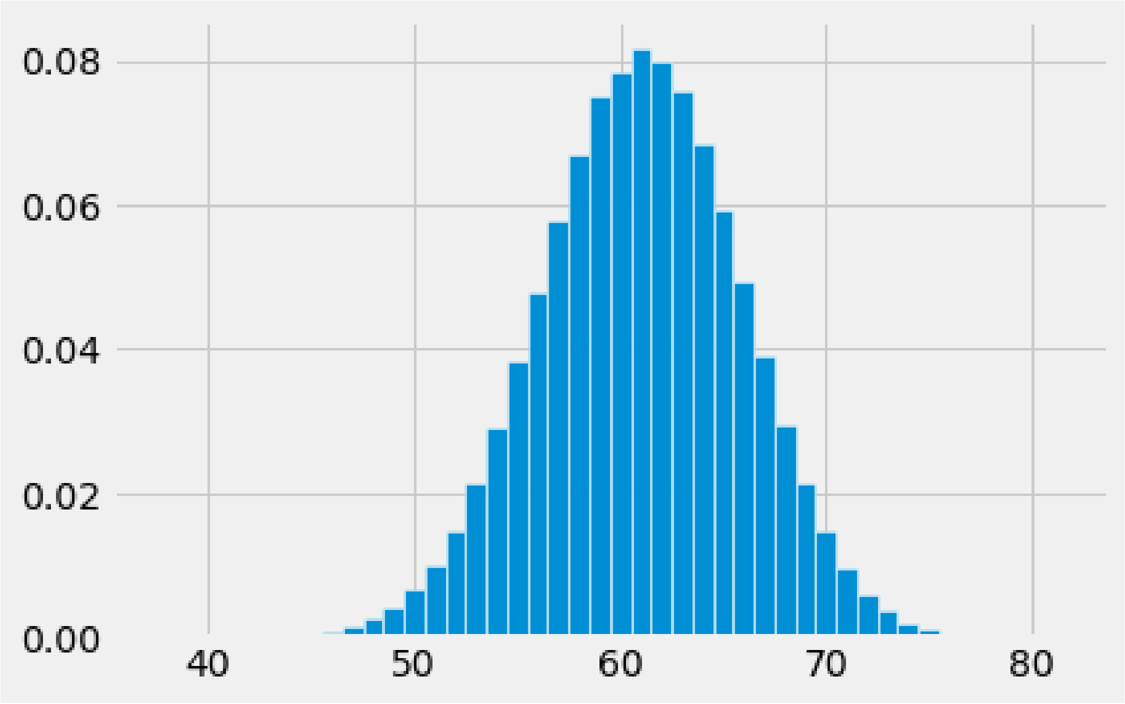


0.1.7 Question 3c

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 voters for Roosevelt in 100 draws made at random with replacement from the voting population.

After you have drawn the histogram, you might want to take a moment to recall the conclusion reached by the *Literary Digest*, a magazine that—while having successfully predicted the outcome of many previous presidential elections—failed to correctly predict the winner of the 1936 presidential election. In their survey of 10 million individuals, they predicted the popular vote as just 43% for Roosevelt and 57% for Landon. Based on our simulation, there was most definitely sampling bias in the *Digest*'s sampling process.

```
In [18]: simulated_counts = np.random.multinomial(100, [0.61, 0.37, 0.02], 500000)[:, 0]
         integer_distribution(simulated_counts)
```



0.1.8 Question 3d

As you know, the count of Roosevelt voters in a sample of 100 people drawn at random from the eligible population is expected to be 61. 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 Roosevelt voters in a random sample of 100 people drawn from the eligible population is closest to

(i) 1.9

(ii) 4.9

(iii) 10.9

(iv) 15.9

(ii) The standard deviation just based on looking at that histogram seems to be about 4.9. The histogram shows a bell shaped curve and the place that looks like the point of inflection is about halfway through 60 and 70.

0.1.9 Question 3e

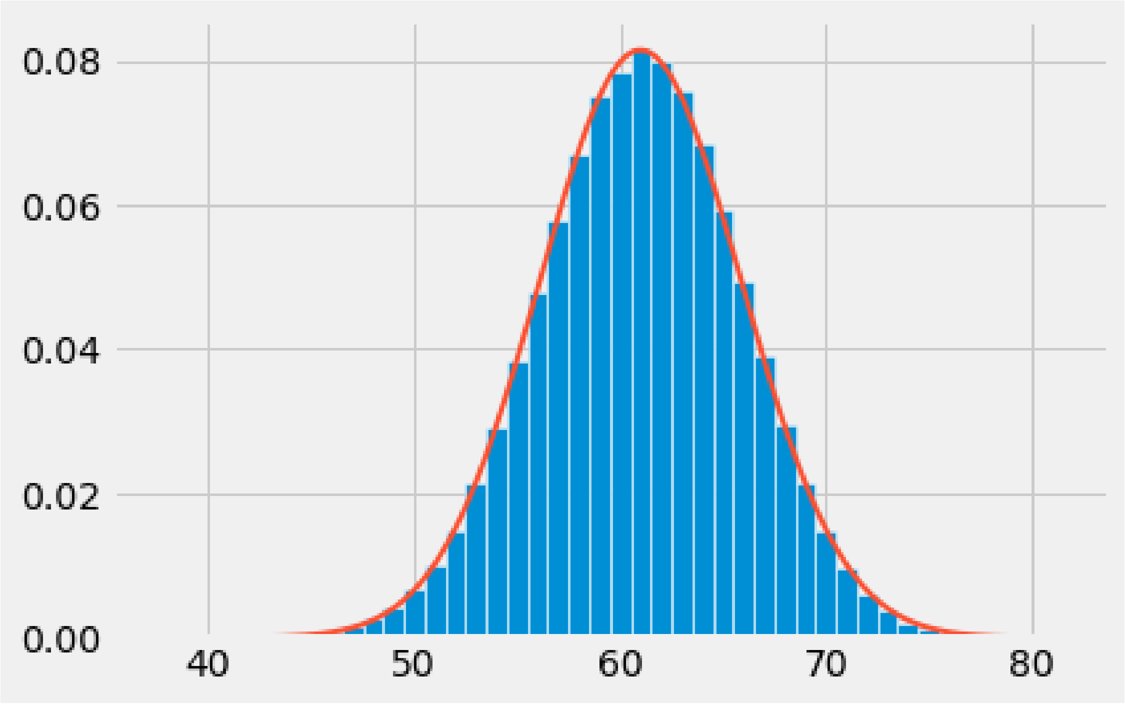
The *normal curve with mean μ and SD σ* is defined by

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

Redraw your histogram from Part **c** and overlay the normal curve with $\mu = 61$ 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 [19]: mu = 61
         sigma = 4.9
         x = np.linspace(40, 80, 200)
         f_x = (1/((2 * np.pi) ** 0.5 * sigma)) * (np.e ** ((-1/2) * ((x - mu)/sigma) ** 2))
         integer_distribution(simulated_counts)
         plt.plot(x, f_x, lw=2);
```



For each part below, you will be presented with a set of vectors, and a matrix consisting of those vectors stacked in columns. 1. State the rank of the matrix, and whether or not the matrix is full rank. 1. If the matrix is *not* full rank, state a linear relationship among the vectors—for example: $\vec{v}_1 = 2\vec{v}_2$.

0.1.10 Question 4a

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, A = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix}$$

1. The rank of the matrix is 2 and it is a full rank because both vectors are linearly independent.

0.1.11 Question 4b

$$\vec{v}_1 = \begin{bmatrix} 3 \\ -4 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, B = \begin{bmatrix} | & | \\ \vec{v}_1 & \vec{v}_2 \\ | & | \end{bmatrix}$$

1. The rank of the matrix is 1 and it is not a full rank.
2. v2 can be written as v1(0).

0.1.12 Question 4c

$$\vec{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 5 \\ 0 \end{bmatrix}, \vec{v}_3 = \begin{bmatrix} 10 \\ 10 \end{bmatrix}, C = \begin{bmatrix} | & | & | \\ \vec{v}_1 & \vec{v}_2 & \vec{v}_3 \\ | & | & | \end{bmatrix}$$

1. The rank of the matrix is 2, and it is not a full rank.
2. v3 can be written as a linear combination of $v3 = 5(v1) + v2$.

0.1.13 Question 4d

$$\vec{v}_1 = \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} -2 \\ -2 \\ 5 \end{bmatrix}, \vec{v}_3 = \begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix}, D = \begin{bmatrix} | & | & | \\ \vec{v}_1 & \vec{v}_2 & \vec{v}_3 \\ | & | & | \end{bmatrix}$$

1. The rank of the matrix is 2, and it is not a full rank.
2. The vectors are linearly dependent, $v3 = v1 + -(v2)$

0.2 Question 5: A Least Squares Predictor

Let the list of numbers $(x_1, x_2, \dots, x_n)$ be data. 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** μ of the list to be

$$\mu \ = \ \frac{1}{n}\sum_{i=1}^{n} x_i.$$

0.2.1 Question 5a

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

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

0.2.2 Question 5b

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 $(x_1 \ldots x_n)$ 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.

First write notation for average in terms of data: $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

Then calculate all the variations: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i - \mu$

0.2.3 Question 5c

One common approach to defining a "best" predictor is as predictor that *minimizes* the MSE on the data $(x_1, \dots, x_n)$.

In this course, we commonly use calculus to find the predictor c as follows: 1. Define MSE to be a function of c, i.e., $MSE(c)$ as above. Assume that the data points $x_1, x_2, ..., x_n$ are fixed, and that c is the only variable. 2. Determine the value of c that minimizes $MSE(c)$. 3. Justify that this is indeed a minimum, not a maximum.

Step 1 is done for you in the problem statement; follow steps 2 and 3 to show that μ is the value of c that minimizes $MSE(c)$. You must do both steps.

1. probably do some optimixation
2. Assume $MSE(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$.

- $\frac{1}{n}\frac{d}{dc}(\sum_{i=1}^{n}(x_i - c)^2) = 0$
- $\frac{1}{n}(-2\sum_{i=1}^{n}(x_i - c)) = 0$
- $\frac{-2}{n}(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}c) = 0$
- $\sum_{i=1}^{n}x_i - nc = 0$
- $\sum_{i=1}^{n}x_i = nc$
- $\sum_{i=1}^{n}x_i = nc$
- $c = \frac{\sum_{i=1}^{n}x_i}{n}$
- $c = \frac{\sum_{i=1}^{n}x_i}{n}$
- $c = $

2. Solve for $MSE''(c) > 0$

- $\frac{-2}{n}\frac{d}{dc}(\sum_{i=1}^{n}(x_i) - nc) > 0$
- $\frac{-2}{n}(0 - n)) > 0$
- $\frac{-2}{n}(-n) > 0$
- $2 > 0$