Properties of Linear Regression Residuals

1. (10 points) In the lecture, we spent a great deal of time talking about Simple Linear Regression (SLR), which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \theta_0 + \theta_1 x$.

   In Lecture 10, we saw that the $\theta_0 = \hat{\theta}_0$ and $\theta_1 = \hat{\theta}_1$ that minimize the average L_2 loss (or Mean Squared Error - MSE) for the simple linear regression model are:

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$
$$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$$

   Or, rearranging terms, our predictions $\hat{y}$ are:

$$\hat{y} = \bar{y} + r\sigma_y \frac{x - \bar{x}}{\sigma_x}$$

   (a) (3 points) As we saw in the lecture, a residual e_i, for data point $i \in \{1, \dots, n\}$, is defined to be the difference between a true response y_i and predicted response $\hat{y}_i$. Specifically, $e_i = y_i - \hat{y}_i$. Note that there are n data points, and each data point is denoted by (x_i, y_i).

   Prove, using the equation for $\hat{y}$ above, that $\sum_{i=1}^{n} e_i = 0$.

   (b) (2 points) Prove that $\bar{y} = \bar{\hat{y}}$. You may use your result from part (a).

   (c) (2 points) Show that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

1a) $e = y - \hat{y}$

→ substitute for $\hat{y}$ with given equation

$$e = y - \left(\bar{y} + r\sigma_y\left(\frac{x-\bar{x}}{\sigma_x}\right)\right)$$

→ take summation of both sides

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\left(y_i - \left(\bar{y} + r\sigma_y\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\right)\right)$$

$$= \sum_{i=1}^{n} y_i - \bar{y} - \sum_{i=1}^{n} r\sigma_y\left(\frac{x_i-\bar{x}}{\sigma_x}\right)$$

$$\sum_{i=1}^{n} y_i - \bar{y} = 0$$

Summation of the deviations of y from its mean equals 0

$$-\sum_{i=1}^{n} r\sigma_y\left(\frac{x_i-\bar{x}}{\sigma_x}\right)$$

r, σ_y, σ_x are all constants, but the summation of x from its mean is 0

$$-\sum_{i=1}^{n} r\sigma_y\left(\frac{0}{\sigma_x}\right) = 0$$

$$\sum_{i=1}^{n} e_i = 0 - 0 = 0$$

1b. $\hat{y} = \bar{y} + r\sigma_y\left(\frac{x-\bar{x}}{\sigma_x}\right)$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{y} + r\sigma_y\left(\frac{x-\bar{x}}{\sigma_x}\right)\right)$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}\bar{y} + \sum_{i=1}^{n} r\sigma_y\left(\frac{x-\bar{x}}{\sigma_x}\right)\right]$$

$$= \frac{1}{n}\left[n\bar{y} + \sum_{i=1}^{n} r\sigma_y\left(\frac{x-\bar{x}}{\sigma_x}\right)\right]$$

$$= \bar{y} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} r\sigma_y\left(\frac{x-\bar{x}}{\sigma_x}\right)}$$

↳ proved in 1a this equals 0

$$= \bar{y} + 0$$

$$\bar{\hat{y}} = \bar{y}$$

1c. $\hat{y} = \theta_0 + \theta_1 x$

① mean of $\hat{y}$ ② sub

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} (\theta_0 + \theta_1 x)$$

↓ sub $\bar{x}$

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} (\theta_0 + \theta_1 \bar{x})$$

constants

$$\bar{y} = \theta_0 + \theta_1 \bar{x} \left(\frac{1}{n} \cdot n\right)$$

$\bar{y} = \theta_0 + \theta_1 \bar{x}$ → therefore $(\bar{x}, y)$ is on the line

(d) (3 points) Show that the residuals are uncorrelated with the predictor variable, that is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i-\bar{e}}{\sigma_e}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)=0,$$

where $\bar{e}=\frac{1}{n}\sum_{i=1}^{n}e_i$, $\sigma_e^2=\frac{1}{n}\sum_{i=1}^{n}(e_i-\bar{e})^2$, and $\sigma_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$. You may assume that σ_e, σ_x, and at least one residual are not exactly zero. Use the properties of estimating equations derived in the lecture.

$\bar{e}_i = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$

(proved in 1a)

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i-\bar{e}}{\sigma_e}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)$$

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i-\frac{1}{n}\sum_{i=1}^{n}e_i}{\sigma_e}\right)\left(\frac{x_i-\frac{1}{n}\sum_{i=1}^{n}x_i}{\sigma_x}\right)$$

constant

$$= \frac{1}{n}\sum_{i=1}^{n}(e_i(x_i-\bar{x}))$$

$$= \frac{1}{n}\sum_{i=1}^{n} e_i x_i - e_i\bar{x}$$

constant

$$= \frac{1}{n}\left[\sum_{i=1}^{n} e_i x_i - \sum_{i=1}^{n} e_i \bar{x}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} e_i x_i - \bar{x}\,\frac{1}{n}\sum_{i=1}^{n} e_i$$

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

orthogonal

$$r = 0$$

Properties of a Linear Model With No Constant Term

2. (4 points) Suppose that we don't include an intercept term in our model. That is, our model is now

$$\hat{y} = \theta x,$$

where θ is the single parameter for our model that we need to optimize. (In this equation, x is a scalar, corresponding to a single observation.)

As usual, we are looking to find the value $\hat{\theta}$ that minimizes the average L_2 loss (MSE) across our observed data $\{(x_i, y_i)\}, for\ i \in \{1, \dots, n\}$:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)^2$$

The estimating equations derived in the lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model, and we'll also explore whether or not our properties from the previous problem still hold.

Use calculus to find the minimizing $\hat{\theta}$.

That is, you may prove that:

$$\hat{\theta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Hint: You may start by following the format of SLR in lecture 10 and replace the SLR model with the model defined above.

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)^2$$

$$\frac{\partial}{\partial\theta}R(\theta) = \frac{-2}{n}\sum_{i=1}^{n}(y_i - \theta x_i)x_i = 0$$

$$\sum_{i=1}^{n} y_i x_i - \theta\sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

MSE "Minimizer"

3. (10 points) Recall from calculus that given some function $g(x)$, the x you get from solving $\frac{dg(x)}{dx} = 0$ is called a *critical point* of g – this means it could be a minimizer or a maximizer for g. In this question, we will explore some basic properties and build some intuition on why, for certain loss functions such as squared L_2 loss, the critical point of the empirical risk function (defined as an average loss on the observed data) will always be the minimizer.

   Given some linear model $f(x) = \theta x$ for some real scalar θ, we can write the empirical risk of the model f given the observed data $\{x_i, y_i\},\ for\ i \in \{1, \ldots, n\}$ as the average L_2 loss (MSE):

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

   (a) (3 points) Let's investigate one of the n functions in the summation in the MSE. Define $g_i(\theta) = \frac{1}{n}(y_i - \theta x_i)^2$ for $i \in \{1, \ldots, n\}$. In this case, note that the MSE can be written as $\sum_{i=1}^{n} g_i(\theta)$.

   Recall from calculus that we can use the 2nd derivative of a function to describe its curvature about a certain point (if it is facing concave up, down, or possibly a point of inflection). You can take the following as a fact: A function is convex if and only if the function's 2nd derivative is non-negative on its domain. Based on this property, verify that $g_i(\theta)$ is a **convex function**.

   (b) (2 points) Briefly explain intuitively in words why given a convex function $g(\theta)$, the critical point we get by solving $\frac{dg(\theta)}{d\theta} = 0$ minimizes g. You can assume that $\frac{dg(\theta)}{d\theta}$ is a function of θ (and not a constant).

   (c) (3 points) Now that we have shown that each term in the summation of the MSE is a convex function, one might wonder if the entire summation is convex, given

3a) $\frac{\partial}{\partial\theta} g_i(\theta) = \frac{-2}{n}(y_i - \theta x_i)x_i$

$\frac{\partial^2}{\partial\theta} g_i(\theta) = \frac{2}{n} x_i > 0$

if second derivative > 0, then it's a convex function

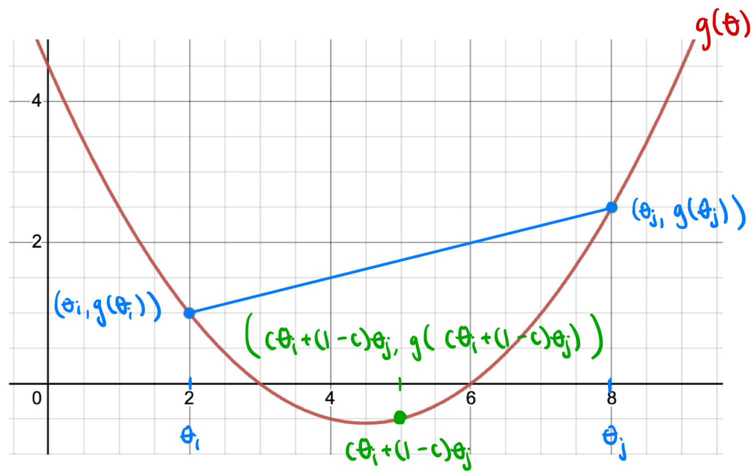
3b) If your given a convex function, then when you take the derivative, the resulting value (when the derivative is set equal to zero) will be the minimum. This is because the first derivative of any function gives the slope. Where the slope is zero is where there is the bottom (or top if its a concave function) of the curve is (an inflection point).

that it is a sum of convex functions.

Let's look at the formal definition of a **convex function**. Algebraically speaking, a function $g(\theta)$ is convex if for any two points $(\theta_i, g(\theta_i))$ and $(\theta_j, g(\theta_j))$ on the function,

$$g(c \times \theta_i + (1-c) \times \theta_j) \leq c \times g(\theta_i) + (1-c) \times g(\theta_j)$$

for any real constant $0 \leq c \leq 1$.
The function g evaluated on any point between θ_i and θ_j will always lie at or below the secant line connecting $g(\theta_i)$ and $(g(\theta_j)$



See a graph in this Wikipedia article `https://en.wikipedia.org/wiki/Convex_function`.

Intuitively, the above definition says that, given the plot of a convex function $g(\theta)$, if you connect 2 randomly chosen points on the function, the line segment will always lie on or above $g(\theta)$ (try this with the graph of $g(\theta) = \theta^2$).

i. (2 points) Using the definition above, show that if $g(\theta)$ and $h(\theta)$ are both convex functions, their sum $g(\theta) + h(\theta)$ will also be a convex function.

$g(ax + (1-a)y) + h(ax + (1-a)y) \leq a\, g(x) + (1-a) g(y)$

$h(ax + (1-a)y) \leq a\, h(x) + (1-a) h(y)$

adding the two equations gives $g(\theta) + h(\theta)$ is convex

ii. (1 point) Based on what you have shown in the previous part, explain intuitively why a (finite) sum of n convex functions is still a convex function when $n > 2$.

Because the second derivatives of the convex functions are non-negative. Additionally, so are the sums.

hw05

October 8, 2023

```
[1]: # Initialize Otter
import otter
grader = otter.Notebook("hw05.ipynb")
```

1 Homework 5A: Sampling

1.1 Due Date: Thursday, October 5th, 11:59 PM

You must submit this assignment to Gradescope by the on-time deadline, Thursday, October 5th, 11:59 PM. Please read the syllabus for the grace period policy. No late submissions beyond the grace period will be accepted. While course staff is happy to help you if you encounter difficulties with submission, we may not be able to respond to last-minute requests for assistance (TAs need to sleep, after all!). **We strongly encourage you to plan to submit your work to Gradescope several hours before the stated deadline.** This way, you will have ample time to reach out to staff for submission support.

This is part of a two-part assignment. After completing this part ("Homework 5A"), please read all instructions carefully to combine manually graded questions from Homework 5A with Homework 5B, and submit your work to both the coding and written portals of Gradescope.

1.2 Collaboration Policy

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you **write your solutions individually**. If you do discuss the assignments with others please **include their names** at the top of your notebook.

Collaborators: *list collaborators here*

1.3 This Assignment

The purpose of this assignment is for you to understand the process of sampling: including convenience samples, random samples; as well as sampling biases. We will also practice drawing samples from a categorical distribution using `NumPy`.

For the purposes of clarity, we will refer to all variations of this class (Data 100/200/200A/200S) as Data 100 in this homework.

1.4 Score Breakdown

Question	Manual?	Points
1ai	No	1
1aii	No	1
1aiii	No	1
1bi	No	1
1bii	No	1
1biii	No	1
1biv	No	1
2a	No	2
2b	Yes	2
2c	Yes	3
2d	Yes	2
Total	3	16

1.4.1 Initialize your environment

[2]:
```
import numpy as np
np.random.seed(200)
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
```

1.5 Data 100 Cutest Pets Contest

Welcome to the Data 100 Cutest Pets Contest, Fall 2023 edition! Course staff nominate their pets to participate in this contest. Students will vote on the cutest one among the nominations in the final exam.

The nominees are:

(From left to right) **Appa** (Matthew's cats), **Pishi** (Professor Norouzi's cat), and **Mimi** (Shiny's dog).

Earlier in the semester, before the actual contest, however, course staff would like to predict the results by surveying students in the class. This process is similar to polling that occurs before a political election.

In this section, you are going to explore different sampling methods.

1.6 Question 1

1.6.1 Question 1a

Since her dog, Mimi, is nominated, Shiny would like to understand the class opinion before the contest. This coming week, she decided to survey all students enrolled in Data 100 this Fall semester (Fall 2023) by sending out an Ed announcement via email that asked students to choose the cutest from the three pets. You may assume no other students/users receive the survey. Shiny closes the survey 12 hours after sending it out.

You can assume that all, and only, enrolled students are on Ed.

Part 1. In Shiny's survey, which of the following is the population of interest? Assign your answer choice to `q1ai` (e.g., `'A'`, `'B'`, etc.).

A. All UC Berkeley students B. All students enrolled in Data 100 across all semesters (Fall 2023 and previous) C. All students enrolled in Data 100 for this semester (Fall 2023) D. All students who fill out Shiny's survey

[3]:
```
q1ai = 'C'
```

[4]:
```
grader.check("q1ai")
```

[4]:
```
q1ai results: All test cases passed!
```

Part 2. In Shiny's survey, which of the following is the sampling frame? Assign your answer choice to `q1aii` (e.g., `'A'`, `'B'`, etc.).

A. UC Berkeley students B. All students enrolled in Data 100 across all semesters (Fall 2023 and previous) C. All students enrolled in Data 100 for this semester (Fall 2023) D. All students who fill out Shiny's survey

[5]:
```
q1aii = 'C'
```

[6]:
```
grader.check("q1aii")
```

[6]:
```
q1aii results: All test cases passed!
```

Part 3. Which of the following is the sample? Assign your answer choice to `q1aiii` (e.g., `'A'`, `'B'`, etc.).

A. UC Berkeley students B. All students enrolled in Data 100 across all semesters (Fall 2023 and previous) C. All students enrolled in Data 100 for this semester (Fall 2023) D. All students who fill out Shiny's survey

[7]:
```
q1aiii = 'D'
```

[8]:
```
grader.check("q1aiii")
```

[8]:
```
q1aiii results: All test cases passed!
```

1.6.2 Question 1b

In practice, we cannot get a 100% survey response rate, often because our population is too large, or because there is a time limit. In this case, very few students answered Shiny's survey before she closed it.

To get more data to predict the answer to the original question ("Which pet will win the Data 100 Cutest Pet Contest?"), Shiny decides on a different strategy: **She conducts the pre-contest survey in person in her discussion section that same week**. She then asks every student who attends the discussion that week for their opinion on the cutest of the three pets, by presenting the following slide:

Part I: In this sampling scheme, which of the following is the population of interest? Assign your answer choice to `q1bi` (e.g., `'A'`, `'B'`, etc.).

A. UC Berkeley students B. All students enrolled in Data 100 across all semesters (Fall 2023 and previous) C. All students enrolled in Data 100 for this semester (Fall 2023) D. All students enrolled in Shiny's discussion section E. All students who fill out Shiny's pre-contest survey

[9]:
```
q1bi = 'C'
```

[10]:
```
grader.check("q1bi")
```

[10]:
```
q1bi results: All test cases passed!
```

Part II: In this sampling scheme, which of the following is the sampling frame? Assign your answer choice to `q2bii` (e.g., `'A'`, `'B'`, etc.).

A. UC Berkeley students B. All students enrolled in Data 100 across all semesters (Fall 2023 and previous) C. All students enrolled in Data 100 for this semester (Fall 2023) D. All students enrolled in Shiny's discussion section E. All students who fill out Shiny's pre-contest survey

[11]:
```
q1bii = 'D'
```

[12]:
```
grader.check("q1bii")
```

[12]:
```
q1bii results: All test cases passed!
```

Part III: Which of the following is the sample? Assign your answer choice to `q2biii` (e.g., `'A'`, `'B'`, etc.).

A. UC Berkeley students B. All students enrolled in Data 100 across all semesters (Fall 2023 and previous) C. All students enrolled in Data 100 for this semester (Fall 2023) D. All students enrolled in Shiny's discussion section E. All students who fill out Shiny's pre-contest survey

[13]:
```
q1biii = 'E'
```

[14]:
```
grader.check("q1biii")
```

[14]:
```
q1biii results: All test cases passed!
```

Part IV: Which of the following best characterizes the sample? Assign your answer choice to `q1biv` (e.g., `'A'`, `'B'`, etc.)..

A. Convenience Sample B. Simple Random Sample C. Probability Sample

[15]:
```
q1biv = 'A'
```

[16]:
```
grader.check("q1biv")
```

[16]:
```
q1biv results: All test cases passed!
```

1.7 Question 2

Shiny was able to sample 50 students in her section; the results are as follows:

Pet	Vote Share
Appa	2%
Pishi	20%
Mimi	78%

Based on this result, she predicts that her dog Mimi will win the contest.

Fast-forward to the end of the semester, when the contest has actually taken place. Assume that after tallying every student's votes, **the true popularity of each pet** is:

Dog	Vote Share
Appa	10%
Pishi	82%
Mimi	8%

The true winner was actually Pishi! Shiny was devastated: what went wrong?

1.7.1 Question 2a

Pishi ultimately wins the Cutest Pet Contest, but from Shiny's results, it was predicted that Mimi would win. Perhaps there was some underlying issue with Shiny's sampling method.

Probability samples can help us quantify sampling bias and chance error. Put briefly, if we assume that a sample distribution was selected at random from a known population, then we can quantify how likely that sample is to have arisen due to random chance (**chance error**). If the difference in sample and population distributions is too great, then we suspect that the given sample has **bias** in how it was selected from the population.

Write one line of code that runs 1000 independent simulations, where each simulation finds the proportion of voters who voted for **Pishi** in a sample of size 50, selected uniformly at random from the **true population**. You may assume that the true population is large enough such that the sample is a random sample with replacement. The output `samples` should be an array with 1000 elements, each of which is the proportion of **Pishi** votes in that simulated sample.

Hint:

- Use `np.random.multinomial` (documentation).
- Use `array[:, i]` to select the i-th column of a 2D `NumPy` array
- Feel free to print out `samples` or `samples.shape` to check the output!

[ ]:

```

```

[17]:

```
samples = np.random.multinomial(50, np.array([0.1, 0.82, 0.08]), size = 1000)[:
↪, 1]/50
samples[:5]
```

[17]:

```
array([0.78, 0.82, 0.88, 0.78, 0.74])
```

[18]:

```
grader.check("q2a")
```

[18]:

```
q2a results: All test cases passed!
```

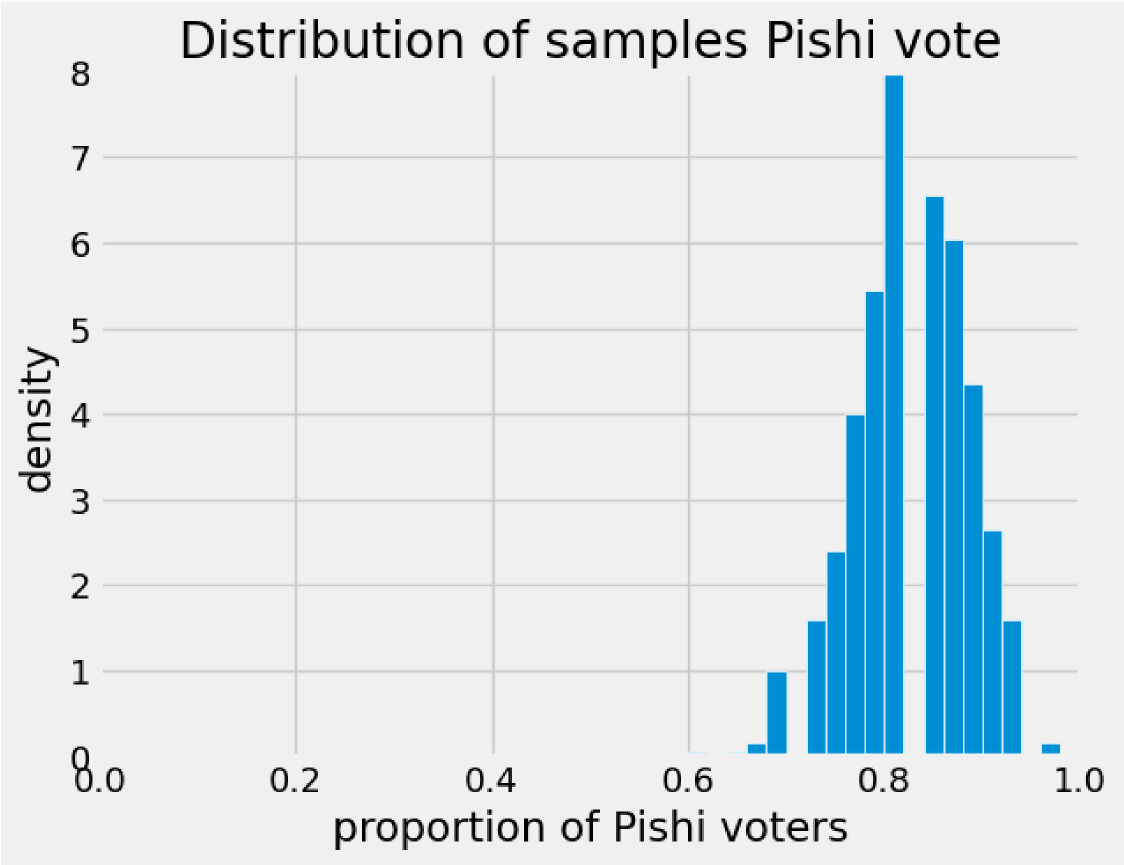
1.7.2 Question 2b

Create a plot using any `seaborn` and/or `matplotlib.pyplot` functions of your choice to visualize `samples`, which is the simulated distribution of Pishi votes using a sample of size 50. Include descriptive titles and labels. An example is included below. The total area under the plot must be normalized to 1. Your plot may not match exactly ours due to randomness of the data generating process in `np.random.multinomial`.

Hint: use `plt.xlim(left, right)` (documentation) to specify the left and right limits of the x-axis.

[31]:

```
plt.hist(samples, density = True, edgecolor = 'white', bins = np.arange(0, 1.1,␣
↪0.02))
plt.xlim(0.0, 1.0)
plt.ylim(0,8)
plt.xlabel('proportion of Pishi voters')
plt.ylabel('density')
plt.title('Distribution of samples Pishi vote')
plt.show()
```



1.7.3 Question 2c

According to Shiny's 50-person sample, 20% of her discussion section reported that they would vote for Pishi in the end-of-semester contest.

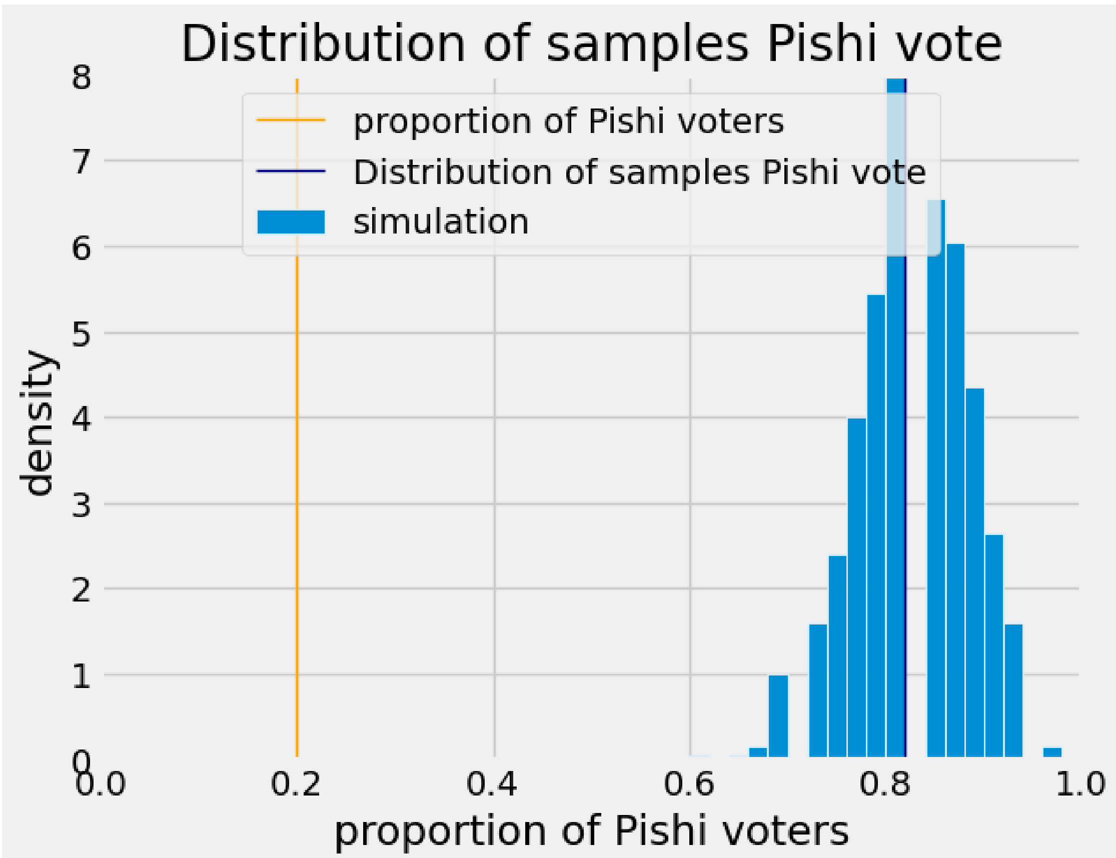
In the cell below, create a plot using any `seaborn` and/or `matplotlib.pyplot` functions of your choice to visualize Shiny's sample statistic superimposed on the simulated sample distribution you plotted in the previous part. In other words, include - a vertical line that passes through 20%, - a vertical line that passes through the mean of the simulated sample distribution, and - the simulated sample distribution itself.

You should choose contrasting colors and include a descriptive title, labels, and a legend if needed. An example is included below.

[32]:
```
mean_sample_dis = np.mean(samples)
mean_sample_dis
```

[32]: 0.8184600000000001

[45]:
```
plt.hist(samples, density = True, edgecolor = 'white', bins = np.arange(0, 1.1,␣
  ↪0.02))
plt.xlim(0.0, 1.0)
plt.ylim(0,8)
plt.xlabel('proportion of Pishi voters')
plt.ylabel('density')
plt.title('Distribution of samples Pishi vote')
plt.axvline(0.2, color="orange", lw=1)
plt.axvline(np.mean(samples), color="navy", lw=1)
plt.legend(['proportion of Pishi voters', 'Distribution of samples Pishi vote',␣
  ↪'simulation'], loc = 'upper center');
```



1.7.4 Question 2d

Based on your analysis above, could Shiny's result have arisen due to chance alone? If not, what could be a potential source of bias?

No, it does not seem that Shiny's result could have arisen due to chance. It is much too far from the distribution of the simulated results. Her erronous result could be due to bias in her sampling meathod: most notably because she only asked students in her section and they might have felt pressured into choosing her pet.

1.8 Congratulations! You have finished Homework 5A!

1.8.1 Submission Instructions

Below, you will see two cells. * Running the first cell will automatically generate a PDF of your answers to all questions that need to be manually graded. * Running the second cell will automatically generate a zip with your autograded answers.

You are responsible for combining this resulting Homework 5A PDF with your answers to Homework 5B, then submitting **both** the coding and written portions of Homework 5 to their respective Gradescope portals: * **Homework 05 Coding**: Submit your Jupyter notebook zip file for Homework 5A, which can be generated and downloaded from DataHub by using the `grader.export()` cell provided below. * **Homework 05 Written**: Submit a single PDF to Gradescope that contains both (1) your answers to all manually graded questions from this Homework 5A Jupyter Notebook, and (2) your answers to all questions in Homework 5B.

To receive credit on this assignment, **you must submit both your coding and written portions to their respective Gradescope portals.** Your written submission (a single PDF) can be generated as follows:

1. Access your answers to manually graded Homework 5A questions in one of three ways:
    - *Automatically create PDF (recommended)*: Run the first cell below and download the generated PDF. This function will extract your response to the manually graded questions and put them on a separate page. This process may fail if your answer is not properly formatted; if this is the case, check out common errors and solution described on Ed or follow either of the two ways described below.
    - *Manually download PDF*: If there are issues with automatically generating the PDF in the first cell, you can try downloading the notebook as a PDF by clicking on `File -> Save and Export Notebook As... -> PDF`. If you choose to go this route, you must take special care to ensure all appropriate pages are chosen for each question on Gradescope.
    - *Take screenshots*: If that doesn't work either, you can take screenshots of your answers (and your code if present) to manually graded questions and include them as images in a PDF. The manually graded questions are listed at the top of the Homework 1A notebook.
2. Answer the Homework 5B written questions.
3. Combine these two sets of answers together into the same PDF, and submit to the appropriate Gradescope written portal. You can use PDF merging tools, e.g., Adobe Reader, Smallpdf or Apple Preview.
4. **Important**: When submitting on Gradescope, you **must tag pages to each question correctly** (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our readers. Failure to do this may result in a score of 0 for untagged questions.

You are responsible for ensuring your submission follows our requirements. We will

not be granting regrade requests nor extensions to submissions that don't follow instructions. If you encounter any difficulties with submission, please don't hesitate to reach out to staff prior to the deadline.

1.9 Submission

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. The cell below will generate a zip file for you to submit. **Please save before exporting!**

After you have run the cell below and generated the zip file, you can open the PDF here.

[46]:
```
# Save your notebook first, then run this cell to export your submission.
grader.export(run_tests=True)
```

```
Running your submission against local test cases…



Your submission received the following results when run against available test
cases:

    q1ai results: All test cases passed!

    q1aii results: All test cases passed!

    q1aiii results: All test cases passed!

    q1bi results: All test cases passed!

    q1bii results: All test cases passed!

    q1biii results: All test cases passed!

    q1biv results: All test cases passed!

    q2a results: All test cases passed!
<IPython.core.display.HTML object>
```

[ ]:
```
```

(d) (2 points) Remember from part (a) that the MSE can be written as:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)^2 = \sum_{i=1}^{n}\frac{1}{n}(y_i - \theta x_i)^2 = \sum_{i=1}^{n} g_i(\theta)$$

Explain why solving for the critical point of the MSE by taking the gradient with respect to the parameter θ and setting that expression to 0, is guaranteed that the solution we find will minimize the MSE.

B/C MSE is a convex function, the critical point is a minimum, meaning the solution minimizes MSE. That means (x,y) are the best values for parameters for the model. → gives the smallest difference in error (between predicted + true values)

Closing note: In this question, we have discussed only the simple linear model with no constant term—a single-variable function. However, the above properties extend more generally to all multivariable linear regression models; this proof is beyond the scope of this course and is left to a future you.

Congratulations! You have finished Homework 5B!