Data 100, Spring 2023

Homework 1B

Due Date: Thursday, January 26, 11:59pm

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Thursday, January 26, 11:59pm**. 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.

There are two parts to this assignment listed on Gradescope:

- **Homework 01 Coding**: Submit your Jupyter notebook zip file for Homework 1A, which can be generated and downloaded from DataHub by using the grader.export() cell provided.
- **Homework 01 Written**: Submit a single PDF to Gradescope that contains both (1) your answers to all manually graded questions from the Homework 1A Jupyter Notebook, and (2) your answers to all questions in this Homework 1B document.

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 1A questions in one of three ways:
    - *Automatically create PDF (recommended)*: We have provided a cell to generate your written response in Homework 1A notebook for you. Run the cell and click to download the generated PDF. This function will extract your response to the manually graded questions and put them on separate pages. 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, on DataHub, you can try downloading the pdf by clicking on `File->SaveandExportNotebookAs...->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 below Homework 1B written questions in one of many ways:

   - Type your answer. We recommend LaTeX, the math typesetting language. Overleaf is a great tool to type in LaTeX.
   - Download this PDF, print it out, and write directly on these pages. If you have a tablet, you may save this PDF and write directly on it.
   - Write your answers on a blank sheet of physical or digital paper.
   - Note: If you write your answers on physical paper, use a scanning application (e.g., CamScanner, Apple Notes) to generate a PDF.

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 (https://smallpdf.com/merge-pdf) or Apple Preview (https://support.apple.com/en-us/HT202945).

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.

Collaborators

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 submission.

0.0.1 Question 1a

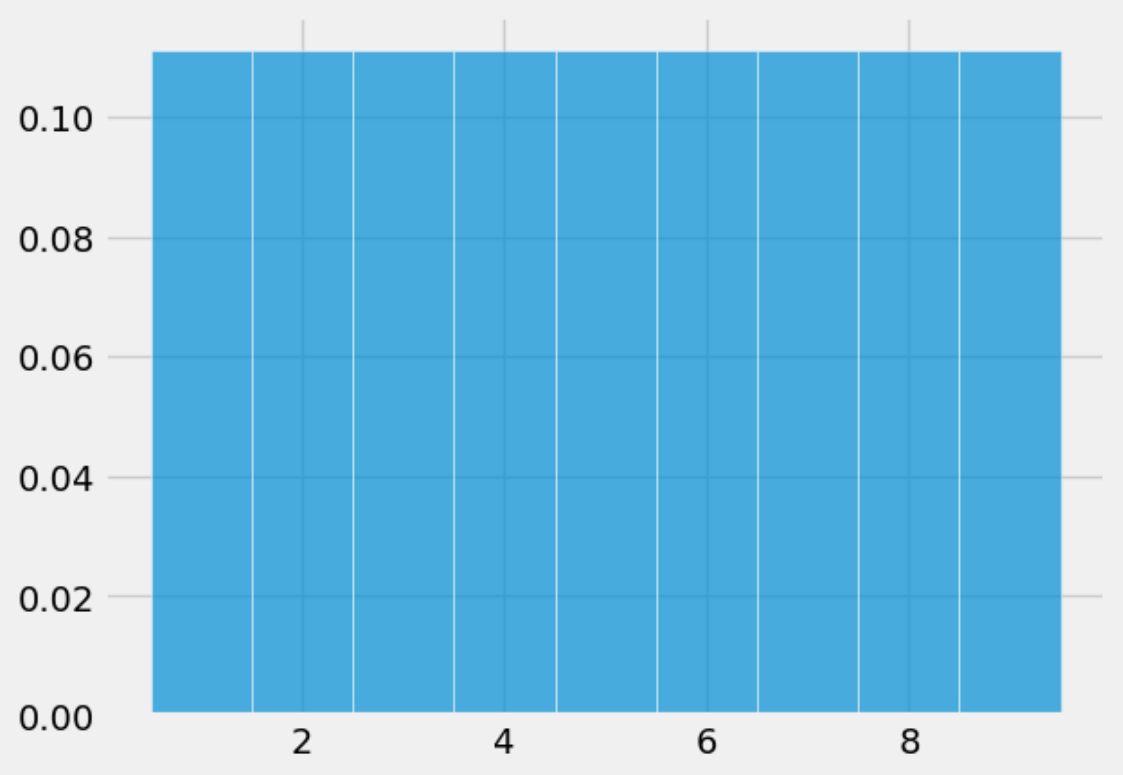
Define a function `integer_distribution` that takes an array of integers and draws the histogram of the distribution using unit bins centered at the integers and white edges for the bars. The histogram should be drawn to the density scale, and opacity should be 70%. The left-most bar should be centered 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. No title or labels are required for this question.

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.

Documentation: `plt.hist()` link

```
In [57]: def integer_distribution(arr_ints):
             unit_bins = np.arange(min(arr_ints)-0.5, max(arr_ints)+1.5, 1)
             plt.hist(arr_ints, bins=unit_bins, ec='white', density=True, alpha=0.7)
         faces = range(1, 10)
         integer_distribution(faces)
```


0.10
0.08
0.06
0.04
0.02
0.00
2
4
6
8

0.0.2 Question 1c

Before we write any code, let's review the idea of hypothesis testing with the permutation test. We first simulate the experiment many times (say, 10,000 times) through random permutation (i.e. without replacement). Assuming that the null hypothesis holds, this process will produce an empirical distribution of a predetermined test statistic. Then, we use this empirical distribution to compute an empirical p-value, which is then compared against a particular cutoff threshold in order to accept or reject our null hypothesis.

In the below cell, answer the following questions: * What does an empirical p-value from a permutation test mean in this particular context of birthweights and maternal smoking habits? * Suppose the resulting empirical p-value $p \leq 0.01$, where 0.01 is our p-value cutoff threshold. Do we accept or reject the null hypothesis? Why?

1. The empirical p-value means out of the many permutation tests, how many times that we see birthweights difference between random samples is larger (or smaller in the negative case) than the observed mean difference between moms with or without maternal smoking habits.
2. With empirical p-value lower than our p-value cutoff threshold, we reject the null hypothesis and accept the alternative hypothesis. It means the chance of random sampling matching the observed difference is lower than 1%. Hence the result is statistically significant.

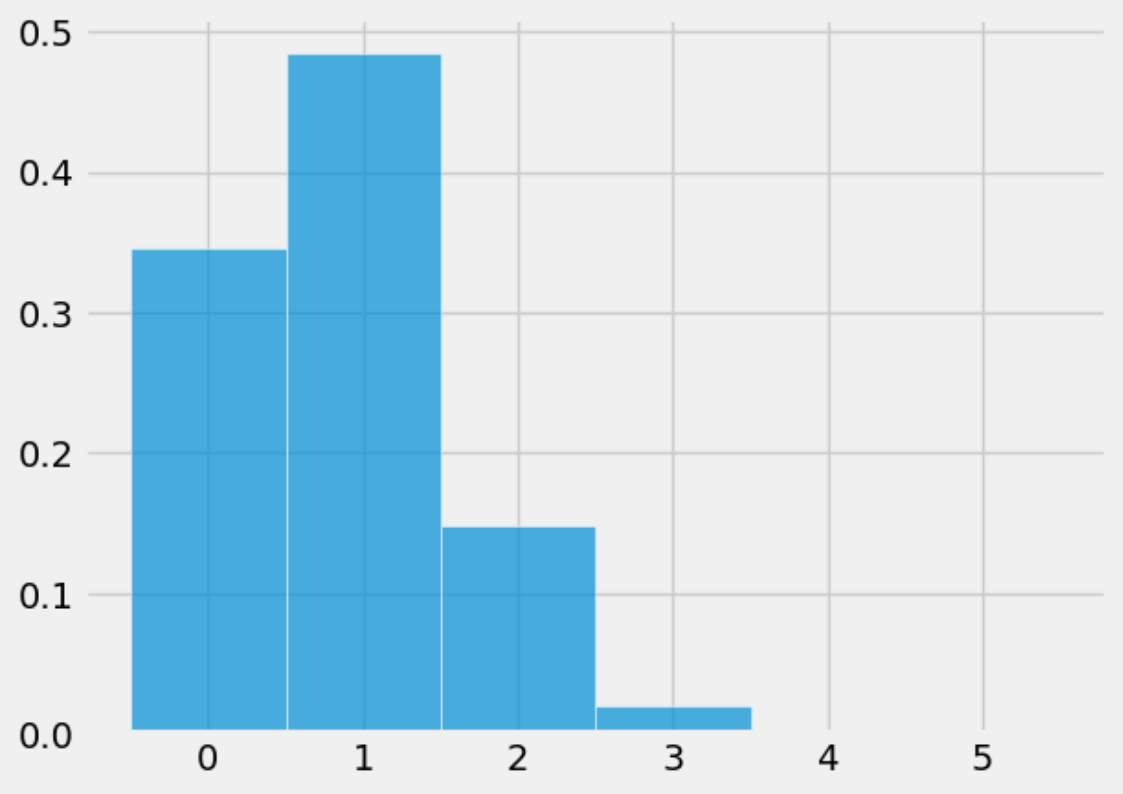
0.0.3 Question 1e

The array `differences` is an empirical distribution of the test statistic simulated under the null hypothesis. This is a prediction about the test statistic, based on the null hypothesis.

Use the `integer_distribution` function you defined in an earlier part to plot a histogram of this empirical distribution. Because you are using this function, your histogram should have unit bins, with bars centered at integers. No title or labels are required for this question.

Hint: This part should be very straightforward.

```
In [60]: integer_distribution(differences)
```



0.0.4 Question 1g

Based on your computed empirical p-value, do we accept or reject the null hypothesis? Be sure to include a reasonable p-value cutoff threshold, if any.

I reject the null hypothesis with a p-value of 0.01. The result is statistically significant.

Homework 1A Manually Graded Questions

0. This is not a question. This is a reminder to include your Homework 1A manually graded questions (automatically generated into a PDF) in your single written PDF submission to Gradescope.

Calculus and Algebra

1. (3 points) In this question we will review calculus by proving some fundamental properties of the logistic function, which will be discussed when we talk more about logistic regression in the latter half of the class. The logistic function (also called the sigmoid function) is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$.

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

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

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

   (b) (2 points) Show that the derivative can be written as:

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

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

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

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

$$= \frac{1}{2+e^{x}+e^{-x}}$$

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

Minimization: A Least Squares Predictor

2. (5 points) **A Least Squares Predictor.** Let the list of numbers $(x_1, x_2, \ldots, x_n)$ 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

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

(a) (1 point) 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 (Data 8, Chapter 14.2). Now prove that fact. That is, show that $\sum_{i=1}^{n}(x_i - \mu) = 0$.

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

$$= \mu \times n - \mu \times n = 0.$$

(b) (1 point) 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 (Data 8, Chapter 14.2). 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.

Step 1 average : $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

Step 2 deviation from average, for ith element, $x_i - \mu$

Step 3 squared deviation: for ith element, $(x_i - \mu)^2$,

Step 4 Variance : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

(c) (3 points) **Mean Squared Error.** 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 some number c as your *predictor*.

The *error* in your prediction is $x_i - c$. Thus the **mean squared error** (MSE) of your predictor c over the entire list of n data points can be written as:

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

You may already see some similarities to your definition of variance from above! You then start to wonder—if you picked your favorite number $c = \mu$ as the predictor, would it be "better" than other choices $c \neq \mu$?

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

Using calculus, determine the value of c that minimizes $MSE(c)$. You must justify that this is indeed a minimum, and not a maximum.

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

$$= \frac{1}{n}\left[(x_1^2 - 2cx_1 + c^2) + (x_2^2 - 2cx_2 + c^2) + \cdots + (x_n^2 - 2cx_n + c^2)\right]$$

$$= \frac{1}{n}\left[(x_1^2 + x_2^2 + \cdots + x_n^2) - 2c(x_1 + x_2 + \cdots + x_n) + n \times c^2\right]$$

$$= \overline{x_i^2} - 2c\,\overline{x_i} + c^2.$$

To minimize $MSE(c)$, $\frac{dMSE(c)}{dc} = 0$, therefore,

$$\frac{dMSE(c)}{dc} = 2c - 2\overline{x_i} = 0, \quad c = \overline{x_i}$$

Therefore, the least squares predictor of the number array is indeed the mean value $\mu / \overline{x_i}$ of the array.

In Data 8 you found (numerically) the least squares *linear* predictor of a variable y based on a related variable x (Data 8, Chapter 15.3). In this course, we will prove your findings using a generalization of your calculation in the previous question. Stay tuned!

Probability and Statistics

3. (2 points) 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 (Chapter 9.5), with particular attention to what's being multiplied in the multiplication rule.

   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. The article is here: https://www.pewresearch.org/fact-tank/2019/03/22/public-confidence-in-scientists-has-remained-stable-for-decades/

   Pick one of the options (1) or (2) to answer the question. If you pick (1), tell us what p is below; if you pick (2), tell us why.

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

   1. is equal to p%.
   2. cannot be found with the information in the article.

1. In 2020, the probability of great public trust in science leaders is:

$P_{scientist}(2020) = 44\%$;

The probability of trust in religions leaders is:

$P_{religion}(2020) = 17\%$;

The chance of public trusting both is

$P\% = P_{scientist}(2020) \times P_{religion}(2020) = 7.48\%$

Content Warning:
This question includes discussion about cancer. If you feel uncomfortable with this topic, **please contact your GSI or the instructors**.

4. (3 points) 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? (Note: You must show all of your work, and also simplify your final answer to 3 decimal places.)

There are 4 scenarios:

$\{\text{True positive}\} \equiv \{TP\}$; $\{\text{False positive}\} \equiv \{FP\}$

$\{\text{True negative}\} \equiv \{TN\}$; $\{\text{False negative}\} \equiv \{FN\}$

What's known:

$\{TP\} + \{FN\} = 1\%$; $\frac{\{TP\}}{\{TP\}+\{FN\}} = 80\%$; $\frac{\{FP\}}{\{FP\}+\{TN\}} = 9.6\%$

$\{TP\} + \{FP\} + \{TN\} + \{FN\} = 100\%$

Need to calculate: $\frac{\{TP\}}{\{TP\}+\{FP\}}$.

$\{TP\} = 1\% \times 80\% = 0.8\%$; $\{FN\} = 1\% - 0.8\% = 0.2\%$

$\{FP\} + \{TN\} = 99\%$, $\{FP\} = 99\% \times 9.6\% = 9.504\%$

$$\frac{\{TP\}}{\{TP\}+\{FP\}} = \frac{0.8\%}{0.8\% + 9.504\%} = 7.764\%$$

Linear Algebra

5. (6 points) A common representation of data uses matrices and vectors, so it is helpful to familiarize ourselves with linear algebra notation, as well as some simple operations.

   Define a vector $\vec{v}$ to be a column vector. Then, the following properties hold:

   - $c\vec{v}$ with c some constant, is equal to a new vector where every element in $c\vec{v}$ is equal to the corresponding element in $\vec{v}$ multiplied by c. For example, $2\begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}2\\4\end{bmatrix}$.
   - $\vec{v}_1 + \vec{v}_2$ is equal to a new vector with elements equal to the elementwise addition of $\vec{v}_1$ and $\vec{v}_2$. For example, $\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}-3\\4\end{bmatrix} = \begin{bmatrix}-2\\6\end{bmatrix}$.

   The above properties form our definition for a **linear combination** of vectors. $\vec{v}_3$ is a linear combination of $\vec{v}_1$ and $\vec{v}_2$ if $\vec{v}_3 = a\vec{v}_1 + b\vec{v}_2$, where a and b are some constants.

   Oftentimes, we stack column vectors to form a matrix. Define the **rank** of a matrix A to be equal to the maximal number of linearly independent columns in A. A set of columns is **linearly independent** if no column can be written as a linear combination of any other column(s) within the set.

   For example, let A be a matrix with 4 columns. If three of these columns are linearly independent, but the fourth can be written as a linear combination of the other three, then $\text{rank}(A) = 3$.

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

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

   A is full rank.

   $\vec{v}_1$ & $\vec{v}_2$ linearly independent.

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

rank (B) = 1, not full rank.

$\vec{v}_1 \times 0 = \vec{v}_2$

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

rank (C) = 2, not full rank

$\vec{v}_1 \times 10 + \vec{v}_2 \times 2 = \vec{v}_3$

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

rank (D) = 2, not full rank

$\vec{v}_1 + \vec{v}_2 \times (-1) = \vec{v}_3$