Homework 2: Probability (2)

UA CSC 380: Principles of Data Science, Fall 2023

Homework due at 11:59pm on Sep 15

Deliverables Your must make two submissions: (1) your homework as a SINGLE PDF file by the stated deadline to the gradescope (Homework 2). Include your code and output of the code as texts in the PDF. and (2) your codes in HW02.ipynb file to a separate submission (Homework 2 code). Each subproblem is worth 10 points. More instructions:

- You can hand-write your answers and scan them to make it a PDF. If you use your phone camera, I recommend using TurboScan (smartphone app) or similar ones to avoid uploading a slanted image or showing the background. Make sure you rotate it correctly.
- Watch the video and follow the instruction for the submission: https://youtu.be/KMPoby5g_nE
- **Show all work along with answers to get the full credit**.
- **Paste all your codes and outputs in the report to get full credit.**
- Place your final answer into an 'answer box' that can be easily identified.
- Map the questions with your solutions when submitting. Points will be deducted if not following this.
- There will be no late days. Late homeworks result in zero credit.

Failure to follow the submission instructions will result in a minor penalty in credit.

You can choose to work individually or in pairs.

- If you choose to work in pairs, you are free to discuss whatever you want with your partner; please make only one submission per group.
- Please do not discuss with people outside your group about the homework (refer to the academic integrity policy in Lecture 1).
- If you have clarification questions, please feel free to post on Piazza so that it can promote discussion.

Problem 1: Joint, Conditional, Marginal Probability

Suppose we throw a fair six-sided die twice in a row. Let A be a random variable representing the number on the first throw, and B be the number on the second throw. Let $M = A \times B$ be the product of both throws. What are the following probabilities?

a) $P(M = 5)$

(1,5),(5,1)

$\frac{2}{36} = \frac{1}{18}$

b) $P(M = 4)$

(2,2),(1,4),(4,1)

$\frac{3}{36} = \frac{1}{12}$

c) $P(A = 2, B = 3 \mid M = 6)$

(2,3),(3,2),(1,6),(6,1)

$\frac{1}{4}$

d) $P(M = 6 \mid A = 2, B = 3)$

(2,3)

1

e) $P(A = 1, B = 3 \mid M = 6)$

()

0

f) $P(A = 4 \mid M = 6)$

()

0

g) $P(A = 1 \mid M = 6)$

(2,3),(3,2),(1,6),(6,1)

$\mathbf{\frac{1}{4}}$

h) $P(A = 3 \mid M = 9)$

(3,3)

1

Problem 2: Discrete Approximation

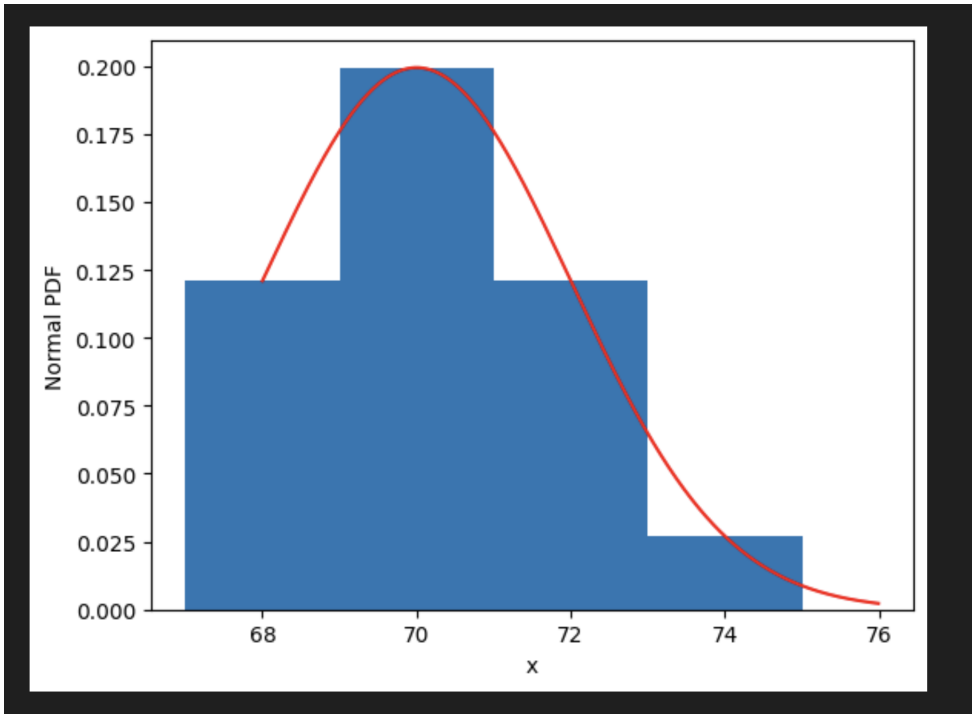
In continuous probability, we often need to solve messy integrals. For example, in this class we might need to use integrals to evaluate the probability of an event under a cumulative distribution function (CDF). Rather than solve this by hand, we can approximate it using discrete intervals. This problem will explore discrete approximation of integrals using a Gaussian model. Recall that the probability density function of a Gaussian is,

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

In the questions below, we will use Python to form a discrete approximation of this continuous distribution, and evaluate associated probabilities.

a) Form a discrete approximation of the Normal PDF with mean $\mu = 70$ and standard deviation $\sigma = 2$. To do this, create an array x of evenly spaced values in the range $[68, 76]$ at increments of 2 excluding 76 (this array will include 68 and 74). The function `numpy.arange` might be helpful. Create an array p containing values of the PDF at each location x. Plot the result as a bar chart (use `matplotlib.pyplot.bar`). In the same figure, overlay a PDF curve (use `matplotlib.pyplot.plot`) at more finely spaced intervals (e.g. 0.01). Paste your code here.

```
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> import scipy.stats as stats
>>> # generate array of x values
>>> x = np.arange(68, 76, 2)
>>> # generate array of normal PDF values at each x value
>>> p = [stats.norm.pdf(x_val, 70, 2) for x_val in x]
>>> # plot discrete estimate gaussian
>>> plt.bar(x, p, width=2)
<BarContainer object of 4 artists>
>>> # x and y labels
>>> plt.xlabel("x")
Text(0.5, 0, 'x')
>>> plt.ylabel("Normal PDF")
Text(0, 0.5, 'Normal PDF')
>>> # plot gaussian curve (as red)
>>> x1 = np.arange(68,76, 0.01)
>>> p1 = [stats.norm.pdf(x1_val, 70, 2) for x1_val in x1]
>>> plt.plot(x1, p1, color="red")
[<matplotlib.lines.Line2D object at 0x16a013fd0>]
```



b) The bar chart above is a discrete approximation of the continuous PDF. We will use it to approximate $P(68 < X \leq 76)$. Recall that $\mathcal{N}(x \mid \mu, \sigma^2)$ is the PDF of X, so

$$P(68 < X \leq 76) = \int_{68}^{76} \mathcal{N}(x \mid \mu, \sigma^2)\, dx.$$

We will approximate this integral using a Riemann sum (https://en.wikipedia.org/wiki/Riemann_sum). Let N be the number of grid points in your array x. The spacing between grid points is Δx and let the i^{th} point of array p be p_i. The Riemann sum approximation is,

$$P(68 < X \leq 76) \approx \sum_{i=1}^{N} p_i \, \Delta x$$

What is the value of the Riemann sum approximation to $P(68 < X \leq 76)$? Paste your code here.

```
>>> # insert your code here
>>> area_rect = [(p_val * 2) for p_val in p]
>>> print("%.4f\n" % sum(area_rect))
0.9369
```

c) Now, reduce the spacing $\Delta x = 0.01$ and recompute the discrete approximation of $P(68 < X \leq 76)$. How do the two approximations compare? What is the practical downside of smaller spacing?

```
>>> # insert your code here
>>> area_rect1 = [(p1_val * 0.01) for p1_val in p1]
>>> print("%.4f\n" % sum(area_rect1))
0.8406
```

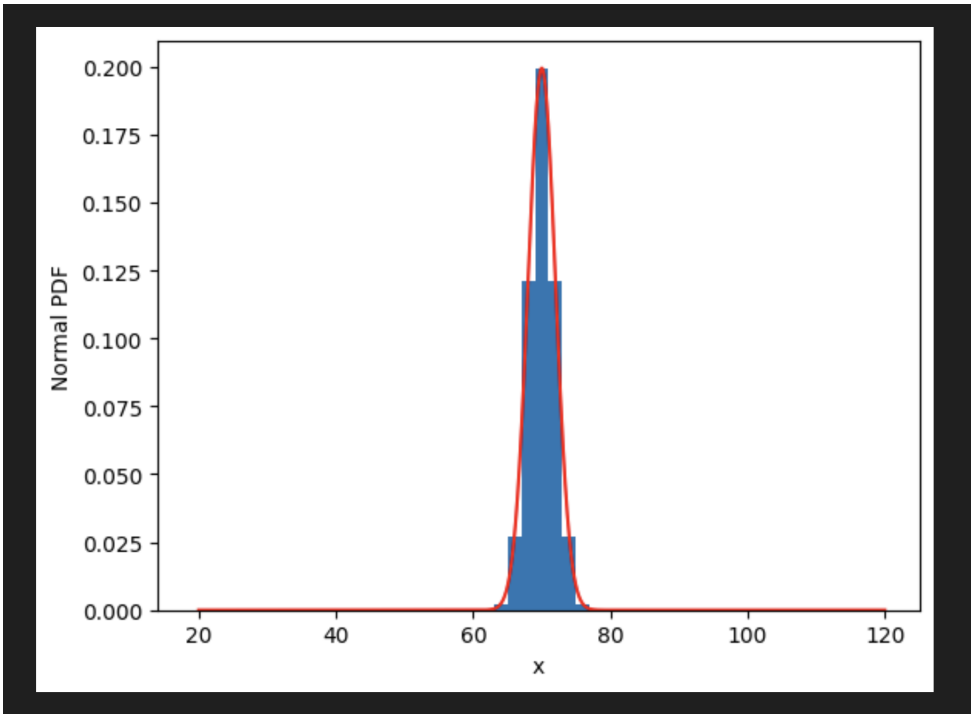
The $x_{delta} = 0.01$ riemann sum is a more accurate approximation of $P(68 < X \leq 76)$ than the $x_{delta} = 2$ riemann sum. This is because as x delta gets smaller, we have more partitions (rectangles) for our riemann sum so the area under the PDF gets more accurate. The practical downside of smaller spacing (meaning x delta is smaller I assume) is perhaps a larger computation since we are calculating and summing up the area of more rectangles than with a larger x delta value. However, a smaller x delta for the riemann sum leads to a more accurate approximation for the area under the PDF.

d) Repeat the steps above to show the distribution over the range $[20, 120]$ and compute $P(20 \leq X < 120)$. What is the value? This interval should contain almost all of the probability in this distribution, i.e. the event is almost certain.

```
>>> # insert your code here
>>> # generate array of x values
>>> x2 = np.arange(20, 120, 2)
>>> # generate array of normal PDF values at each x value
>>> p2 = [stats.norm.pdf(x2_val, 70, 2) for x2_val in x2]
>>> # plot discrete estimate gaussian
>>> plt.bar(x2, p2, width=2)
<BarContainer object of 50 artists>
>>> # x and y labels
>>> plt.xlabel("x")
Text(0.5, 0, 'x')
>>> plt.ylabel("Normal PDF")
Text(0, 0.5, 'Normal PDF')
>>> # plot gaussian curve (as red)
>>> x3 = np.arange(20, 120, 0.01)
>>> p3 = [stats.norm.pdf(x3_val, 70, 2) for x3_val in x3]
>>> plt.plot(x3, p3, color="red")
[<matplotlib.lines.Line2D object at 0x11207f290>]

>>> # riemann sum x_delta = 2
>>> area_rect2 = [(p_val * 2) for p_val in p2]
>>> print("riemann sum x_delta = 2:\n" + str(sum(area_rect2)))
riemann sum x_delta = 2:
1.000000005350576
```

```
>>> # riemann sum x_delta = 0.01
>>> area_rect3 = [(p_val * 0.01) for p_val in p3]
>>> print("riemann sum x_delta = 0.01:\n" + str(sum(area_rect3)))
riemann sum x_delta = 0.01:
0.9999999999998426
```



Problem 3: Conditional Independence

Let $A, B, C \in \{0, 1\}$ be three binary random variables with the following joint probability distribution:

a	b	c	$P(A = a, B = b, C = c)$
0	0	0	0.01
0	0	1	0.07
0	1	0	0.02
0	1	1	0.10
1	0	0	0.02
1	0	1	0.30
1	1	0	0.04
1	1	1	0.44

a) By direct calculation, compute the marginal $P(A, B)$ (recall that $P(A, B)$ is represented by 4 numbers: $P(A = 0, B = 0)$, $P(A = 0, B = 1)$, $P(A = 1, B = 0)$, $P(A = 1, B = 1)$).
$\boldsymbol{P(A = 0, B = 0) = 0.01 + 0.07 = 0.08}$
$\boldsymbol{P(A = 0, B = 1) = 0.02 + 0.10 = 0.12}$
$\boldsymbol{P(A = 1, B = 0) = 0.02 + 0.30 = 0.32}$
$\boldsymbol{P(A = 1, B = 1) = 0.44 + 0.04 = 0.48}$

b) By direct calculation compute the marginals $P(A)$ and $P(B)$.
$\boldsymbol{P(A = 0) = 0.01 + 0.07 + 0.02 + 0.10 = 0.20}$
$\boldsymbol{P(A = 1) = 0.02 + 0.30 + 0.04 + 0.44 = 0.80}$
$\boldsymbol{P(B = 0) = 0.01 + 0.07 + 0.02 + 0.30 = 0.40}$
$\boldsymbol{P(B = 1) = 0.10 + 0.02 + 0.04 + 0.44 = 0.60}$

c) Are the random variables A and B independent? Why or why not?

The random variables A and B are independent because the product of P(A) and P(B) equals P(A,B). For example, $\boldsymbol{P(A = 1) * P(B = 1) = 0.80 * 0.60 = 0.48 = P(A = 1, B = 1)}$ and so on.

d) Compute the conditional distribution $P(A, B \mid C)$. Note that this includes computing $P(A, B \mid C = 0)$ as well as $P(A, B \mid C = 1)$, each of which is represented by 4 numbers (in total 8 numbers).

$\boldsymbol{P(A = 0, B = 0 \mid C = 0) = \frac{0.01}{0.09} = \frac{1}{9}}$

$\boldsymbol{P(A = 0, B = 1 \mid C = 0) = \frac{0.02}{0.09} = \frac{2}{9}}$

$\boldsymbol{P(A = 1, B = 0 \mid C = 0) = \frac{0.02}{0.09} = \frac{2}{9}}$

$\boldsymbol{P(A = 1, B = 1 \mid C = 0) = \frac{0.04}{0.09} = \frac{4}{9}}$

$\boldsymbol{P(A = 0, B = 0 \mid C = 1) = \frac{0.07}{0.91} = \frac{1}{13}}$

$\boldsymbol{P(A = 0, B = 1 \mid C = 1) = \frac{0.10}{0.91} = \frac{10}{91}}$

$\boldsymbol{P(A = 1, B = 0 \mid C = 1) = \frac{0.30}{0.91} = \frac{30}{91}}$

$\boldsymbol{P(A = 1, B = 1 \mid C = 1) = \frac{0.44}{0.91} = \frac{44}{91}}$

e) Calculate $P(A = 0, B = 0 \mid C = 1)$ as well as $P(A = 0 \mid C = 1) \cdot P(B = 0 \mid C = 1)$. Are A and B conditionally independent given C? Why or why not?

$\boldsymbol{P(A = 0, B = 0 \mid C = 1) = \frac{0.07}{0.91} = \frac{1}{13} = 0.07692}$

$\boldsymbol{P(A = 0 \mid C = 1) = \frac{0.17}{0.91} = \frac{17}{91}}$

$\boldsymbol{P(B = 0 \mid C = 1) = \frac{0.37}{0.91} = \frac{37}{91}}$

$\boldsymbol{\frac{37}{91} * \frac{17}{91} = 0.07596}$

A and B are not conditionally independent given C because $\boldsymbol{P(A = 0, B = 0 \mid C = 1)}$ does not equal $\boldsymbol{P(A = 0 \mid C = 1) \cdot P(B = 0 \mid C = 1)}$.

Problem 4: Diagnostic Tests and Bayes' Rule

I have decided to get myself tested for COVID-19 antibodies. However, being comfortable with statistics, I am curious about what the test means for my actual status. Let's investigate these questions, showing all your work.

a) The antibody test I take has a *sensitivity* (a.k.a. true positive rate) of 97.5% and a *specificity* (a.k.a. true negative rate) of 99.1%. If you are not familiar with sensitivity vs specificity, please see Wikipedia. Assume that 4% of the population actually have COVID-19 antibodies. Write down the joint probability distribution $P(S, R)$ with events for antibody state $S \in \{\text{true}, \text{false}\}$ and test result $R \in \{\text{true}, \text{false}\}$.

		Test Result	Test Result	
		True	False	
Antibody	True	0.039	0.001	0.04
Antibody	False	0.00864	0.95136	0.96
		0.04764	0.95236	1

b) Assuming I receive a *positive* test result, use Bayes' rule to calculate the probability that I actually have COVID-19 antibodies.

$$\boldsymbol{P(S = true \mid R = true) = \frac{P(S=true, R=true)}{P(R=true)} = \frac{0.039}{0.04764} = 0.8186}$$

c) Assuming I receive a *negative* test result, what is the probability that I *do not* have COVID-19 antibodies?

$$\boldsymbol{P(S = false \mid R = false) = \frac{P(S=false, R=false)}{P(R=false)} = \frac{0.95136}{0.95236} = 0.9989}$$

d) Assume I take the test twice, and receive a positive result in the first test and a negative result in the second test. Assume that the two test results are conditionally independent given the existence of the antibody. What is the probability that I have COVID-19 antibodies according to Bayes' rule?

$$\boldsymbol{0.8186 * 0.18136 = 0.1485}$$

e) Now assume that only 1% of the population has COVID-19 antibodies. Repeat parts (b) and (c) with this revised prior belief.

		Test Result	Test Result	
		True	False	
Antibody	True	0.00975	0.00025	0.01
Antibody	False	0.00891	0.98109	0.99
		0.01866	0.98134	1

$$\boldsymbol{P(S = true \mid R = true) = \frac{P(S=true, R=true)}{P(R=true)} = \frac{0.00975}{0.01866} = 0.5225}$$
$$\boldsymbol{P(S = false \mid R = false) = \frac{P(S=false, R=false)}{P(R=false)} = \frac{0.98109}{0.98134} = 0.9997}$$