DS BMED 200, Fall 2024
Problem Set 2: Probability
Due November 7th, 2024 at 11:59pm PST

Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- Submit through the BruinLearn course website under the Gradescope tab on the left. You can add yourself to the course Gradescope site by going to gradescope.com, clicking "Add a course" and entering the following entry code: GPBBKD.
- Please provide short and concise answers. Long, cumbersome, or unclear answers will not be checked.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you plan to submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

1 Events [10 pts]

(a) [2 pts] If we flip a coin 3 times, what is our sample space Ω?

{HHH, HHT, HTT, HTH, TTT, THH, THT, TTH}

(b) [2 pts] If we flip a coin 3 times, what is the complement of the set $A = \{HHH, HHT, HTH\}$?

{HTT, TTT, THH, THT, TTH}

(c) [2 pts] We flip a fair coin 3 times. Let $A = \{HHH, HHT, HTH, THH, TTH\}$ and $B = \{THT\}$ be two events, calculate $P(A \cap B)$.

The probability is zero since A and B are mutually exclusive.

(d) [2 pts] We roll a fair 12-sided dice once. Let $A = \{1, 2, 5\}, B = \{1, 3, 5, 6\}$ be two events, calculate $P(A \cup B)$.

A U B = {1,2,3,5,6} = 5/12

(e) [2 pts] We roll a fair 12-sided dice just once. Are the events $A = \{1, 2, 5, 7, 11, 12\}, B = \{1, 3, 5, 6\}$ independent?

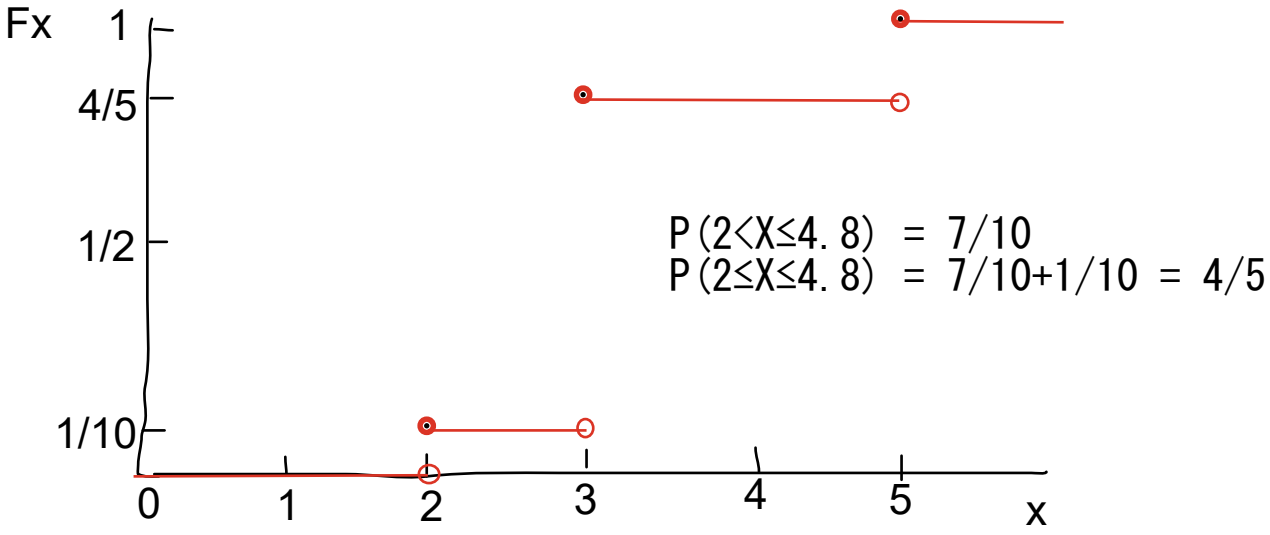
P(A) = 1/2, P(B) = 1/3
P(A)*P(B) = 1/6, P(A n B) = P{1,5} = 1/6
P(A)*P(B) = P(A n B)
A and B are independent.

2 Density/Mass functions [8 pts]

(a) [4 pts] Let X be a random variable with a probability function defined by:

$$P(X = 2) = 1/10, \quad P(X = 3) = 7/10, \quad P(X = 5) = 2/10$$

Plot the CDF F_X. Use F_X to find $P(2 < X \leq 4.8)$ and $P(2 \leq X \leq 4.8)$.



(b) [4 pts] Is the following a valid PDF?

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 2e^{-2x} & \text{otherwise} \end{cases} \tag{1}$$

1)f(x) = 2e^(-2x) is greater than zero for all x>0, thus f(x)>=0 for all x.
2) the integral of f(x)dx for x<= 0 is 0;
the integral of f(x)dx for x from 0 to infinity is 1.

Thus f(x) is a valid PDF.

3 Joint and Conditional Probability [8 pts]

Consider a scenario where we repeatedly flip a coin with probability p for heads (i.e., $P(X_i = H) = p$, where X_i is the random variable representing the i-th flip). Assume different flips are independent. Please express each of your answers to the following questions as a function of p.

(a) [2 pts] Suppose we observe $x_1 = H, x_2 = H, x_3 = H, x_4 = H$ for the first four flips. Calculate the conditional probability of observing an H on the fifth flip, that is, $P(X_5 = H|x_1 = H, x_2 = H, x_3 = H, x_4 = H)$.

Since we assume difference flips are independent, P(x5 = H) = p.

(b) [2 pts] Suppose we flip the coin five times. Calculate the probability of observing the exact sequence $THTTH$.

P = (1-p)(p)(1-p)(1-p)(p)=(1-p)^3*p^2

(c) [2 pts] Suppose we flip the coin five times. Calculate the probability of observing two heads and three tails across five flips.

P = 5!/(2!(5-2)!) * p^2 * (1-p)^(5-2) = 10*p^2*(1-p)^3

(d) [2 pts] Calculate the probability of observing at least three heads in a sequence of five coin flips.

P = P(X=3) + P(X=4) + P(X=5)
P(X=3) = 10*p^3*(1-p)^2
P(X=4) = 5*p^4*(1-p) = 5p^4 - 5p^5
P(X=5) = p^5
P = 10*p^3*(1-p)^2 + 5*p^4 -4p^5

4 Bayes Rule [4 pts]

(a) [4 pts] Suppose there is a rare disease that can only be found in 0.1% 0f the population. A company has developed a test with probability 0.96 for a *true positive* result (true positive rate) and probability 0.03 for a *false positive* result (false positive rate). If a person gets a positive result, what is the probability of them actually having the disease?

Recall: The probability of a true positive result is the probability of receiving a positive result when the patient is indeed a case for the disease; the probability of a false positive result is the probability of receiving a positive result when the patient is in fact NOT a case for the disease.

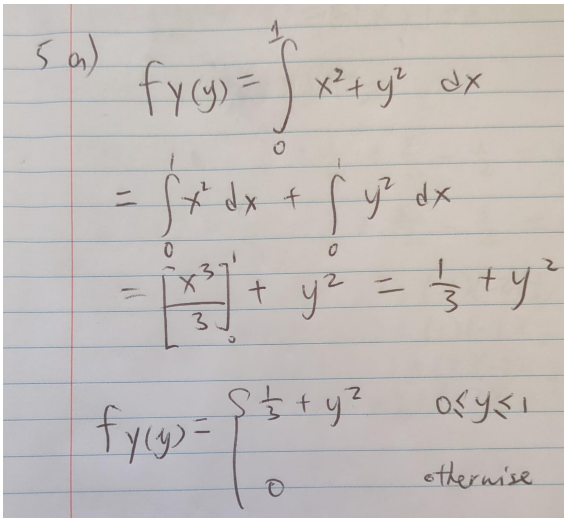
P(disease|+) = P(+|disease)*P(disease)/P(+)
P(+) = P(+|disease)*P(disease) + P(+|healthy)*P(healthy)
=0.96*0.001 + 0.03*0.999 = 0.03093
P(disease|+) = 0.96*0.001/0.03093 = 0.031

5 Marginal Probability [4 pts]

(a) [4 pts] Let X, Y be two random variables with the following joint density function:

$$f(x,y) = \begin{cases} x^2 + y^2 & \text{if } 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Derive $f_Y(y)$, the marginal probability density function of Y.

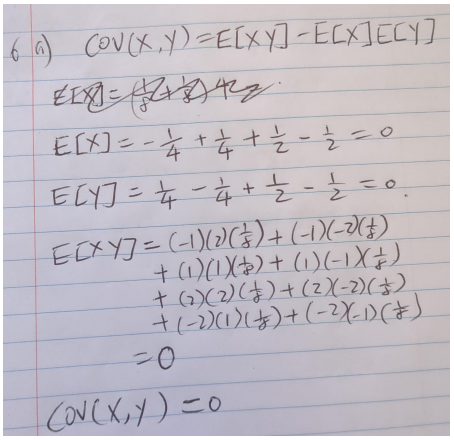


6 Covariance [4 pts]

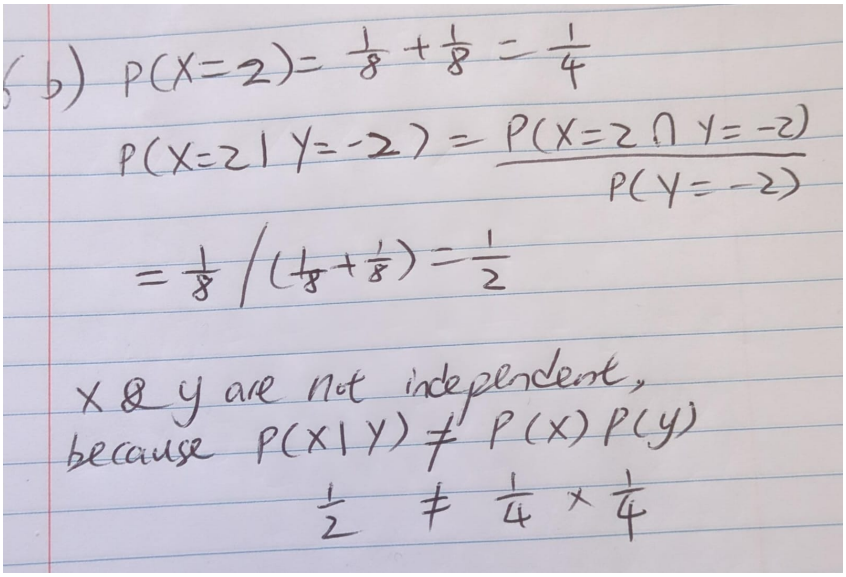
Let X, Y be two random variables with the following joint probability mass function:

$f(X,Y)$	$Y=1$	$Y=-1$	$Y=2$	$Y=-2$
$X=-1$	0	0	$\frac{1}{8}$	$\frac{1}{8}$
$X=1$	$\frac{1}{8}$	$\frac{1}{8}$	0	0
$X=2$	0	0	$\frac{1}{8}$	$\frac{1}{8}$
$X=-2$	$\frac{1}{8}$	$\frac{1}{8}$	0	0

(a) [3 pts] Calculate $\text{COV}(X,Y)$.



(b) [1 pts] What is $P(X=2)$? How about $P(X=2|Y=-2)$? What can you conclude about the relationship between X and Y? Are X and Y independent?

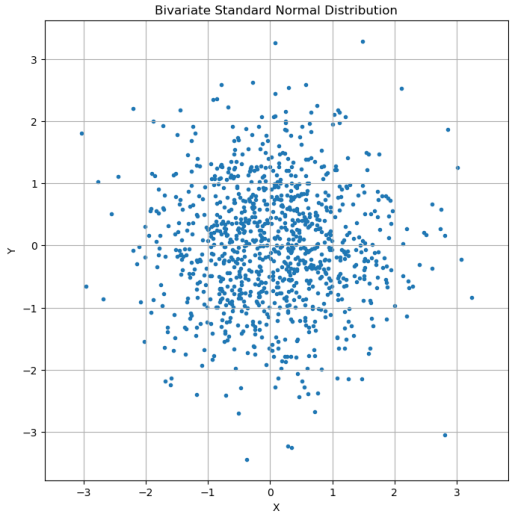


7 Implementation: Multivariate Gaussian [8 pts]

(a) [3 pts] Sample 1,000 values (i.e., 1,000 vectors of length 2) from the bi-variate standard normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1, 0 \\ 0, 1 \end{bmatrix} \right)$$

Visualize the 1,000 values by a scatter plot (i.e., the values drawn from X on one axis and those drawn from Y on the second axis). (Hint: you can use the function `numpy.random.multivariate_normal` in python or `MASS::mvrnorm` in R.)



(b) [1 pts] Calculate $P(X \geq 0)$ and $P(X \geq 0 \,|\, Y \geq 0)$ based on the values you drew in (a) by calculating the frequencies of these cases in your 1,000 sampled values.

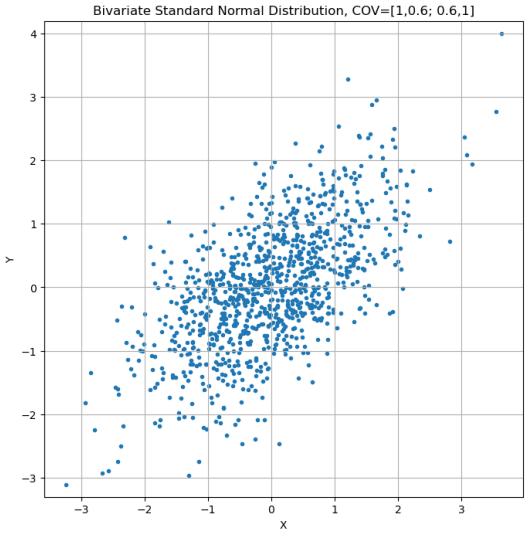
```
#Count samples where X >= 0
count_X = np.sum(X_values >= 0)
P_X = count_X / sample_size

#Count samples where X >= 0 and Y >= 0
count_XY = np.sum((X_values >= 0) & (Y_values >= 0))

#Count samples where Y >= 0
count_Y = np.sum(Y_values >= 0)

#Calculate conditional probability P(X >= 0 | Y >= 0)
P_XY = count_XY / count_Y if count_Y > 0 else 0

print("P(X ≥ 0):", P_X)
print("P(X ≥ 0 | Y ≥ 0):", P_XY)
```

```
[52] ✓ 0.0s
P(X ≥ 0): 0.517
P(X ≥ 0 | Y ≥ 0): 0.5060975609756098
```

(c) [3 pts] Sample 1,000 values (i.e., 1,000 vectors of length 2) from the following bi-variate normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1, 0.6 \\ 0.6, 1 \end{bmatrix} \right)$$

Visualize the 1,000 values by a scatter plot.



(d) [1 pts] Calculate $P(X \geq 0)$ and $P(X \geq 0 \,|\, Y \geq 0)$ based on the values you drew in (a) by calculating the frequencies of these cases in your 1,000 sampled values. Based on what you observe, are X and Y independent?

```
#Count samples where X >= 0
count_X = np.sum(X_values >= 0)
P_X = count_X / sample_size

#Count samples where X >= 0 and Y >= 0
count_XY = np.sum((X_values >= 0) & (Y_values >= 0))

#Count samples where Y >= 0
count_Y = np.sum(Y_values >= 0)
P_Y = count_Y/sample_size

#Calculate conditional probability P(X >= 0 | Y >= 0)
P_XY = count_XY / count_Y if count_Y > 0 else 0

print("P(X ≥ 0):", P_X)
print("P(Y ≥ 0):", P_Y)
print("P(X ≥ 0 | Y ≥ 0):", P_XY)
```

[55] ✓ 0.0s

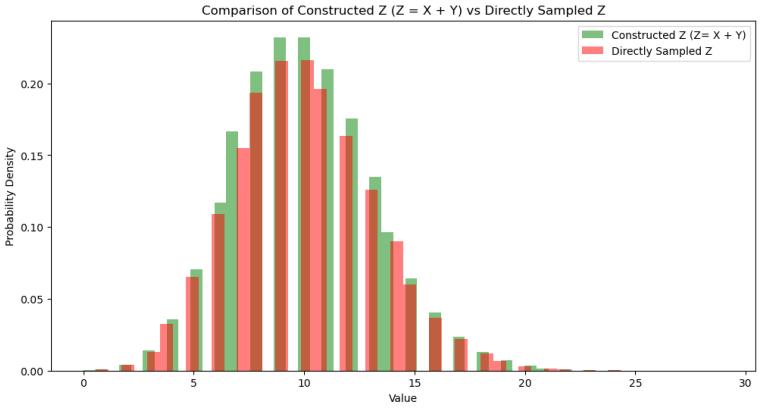
```
P(X ≥ 0): 0.481
P(Y ≥ 0): 0.498
P(X ≥ 0 | Y ≥ 0): 0.6807228915662651
```

P(X ≥ 0)*P(Y ≥ 0)= 0.240
P(X ≥ 0 | Y ≥ 0)= 0.681
X and Y are not indepedent.

8 Implementation: Poisson and binomial [8 pts]

(a) [4 pts] If $X \sim \textbf{Poisson}(\lambda_1)$ and $Y \sim \textbf{Poisson}(\lambda_2)$ are two independent random variables, then their sum $Z = X + Y$ satisfies: $Z \sim \textbf{Poisson}(\lambda_1 + \lambda_2)$.

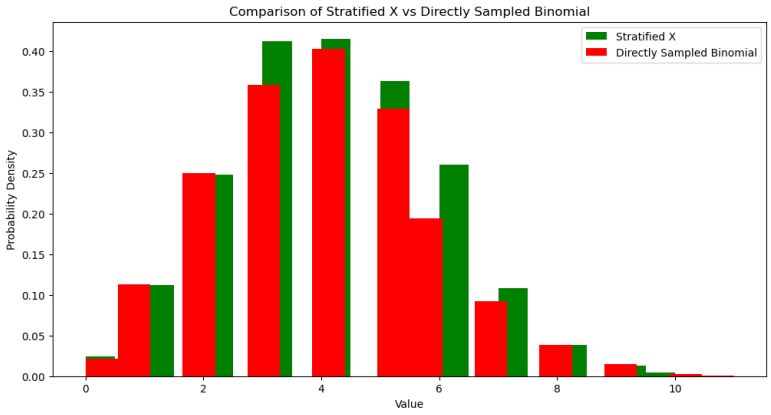
We shall verify this relation empirically. Let $X \sim \textbf{Poisson}(\lambda_1 = 2)$, $Y \sim \textbf{Poisson}(\lambda_2 = 8)$. First, draw 10^6 pairs of values from X, Y and use their sums to construct each pair's corresponding value of Z. In addition, draw a second sample of 10^6 values directly from $\tilde{Z} \sim \textbf{Poisson}(\lambda = \lambda_1 + \lambda_2 = 10)$. Compare the distribution of the values from our constructed Z with the distribution of the values from $\tilde{Z}$ by plotting the histogram of each of the two samples (both histograms in the same figure). Comment on the similarity or dissimilarity of the histograms.



The two histograms are very similar in distribution, suggesting that if X and Y are two independent Poisson distributions λ1 and λ2, then their sum Z = X+Y would also satisfy the sum of the two independent Poisson distributions (λ1 + λ2).

(b) [4 pts] If $X \sim$ **Poisson**(λ_1) and $Y \sim$ **Poisson**(λ_2) are two independent random variables, then the conditional distribution of X given $X + Y = t$ follows **Binomial**(t, p) where $p = \lambda_1/(\lambda_1 + \lambda_2)$.

We shall verify this relation empirically. Our conditional distribution criterion (i.e. $X+Y = t$) can be met by stratifying the data used in part (a) to pairs that sum up to precisely t. We will arbitrarily pick $t = 20$ in this exercise. In addition, draw the same amount of values directly from the desired Binomial distribution $\tilde{X} \sim$ **Binomial**$(t = 20,\ p = \lambda_1/(\lambda_1 + \lambda_2) = 1/5)$. Compare the two sets of values by plotting the histogram of each set. Comment on the similarity or dissimilarity of the histograms.



The two histograms are similar in distribution, suggesting that if X and Y are two independent Poisson distributions λ1 and λ2, the conditional distribution of X given X+Y= t indeed follows Binomial(t,p) where p = λ1/(λ1 + λ2).