CSE 312 : Spring 2023 Final Exam Solutions

Name:	**NetID:** @uw.edu

Instructions

- You have 110 minutes to complete this exam.
- You are permitted one piece of 8.5x11 inch paper with handwritten notes (notes are allowed on both sides of the paper).
- You may not use a calculator or any other electronic devices during the exam.
- We will be scanning your exams before grading them. Please write legibly, and avoid writing up to the edge of the paper.
- If you run out of room, you may also use the last page for extra space, but tell us where to find your answer if it's not right below the problem.
- Since you don't have a calculator, you are generally free to **not** simplify expressions (though you may if you think it will be helpful).
- In general, show us the work you used to get to an answer, and explanations will help us reward partial credit, but we do not expect explanations at the level we usually require on homeworks.

Advice

- Writing a few words about where an expression came from is often very helpful for awarding partial credit.
- Remember to take deep breaths.

Question	**Max points**
Multiple Choice	30
Small Problems	11
Zoo	25
Cabbages	20
Concentration (Berries)	15
MLE (Baseball)	13
Morale	1
Total	**115**

1. [Multiple Choice] $0.25^{10} \approx 0.000001$ [30 points]

(a) A maximum likelihood estimator is

○ Always consistent, but only sometimes unbiased.
○ Always unbiased, but only sometimes consistent.
○ Always both unbiased and consistent.
○ None of the above.

Solution:

Always consistent, but only sometimes unbiased.

(b) Let f_X be the pdf of a continuous random variable. Which of the following must be true?

○ $\int_0^\infty f_X(z)dz = 1$
○ $f_X(z) \leq 1$ for all z
○ f_X is continuous
○ None of the above. **Solution:**

None of the above. Some continuous random variables can take on negative values, which aren't covered by $\int_0^\infty f_X(z)dz$. Densities are not probabilities, so they can be greater than 1. And pdfs can have jump discontinuities. For example, a continuous Unif$(0.5, 0)$ random variable displays all three of these things.

(c) Which of the following is true of continuous random variables?

○ $f_X(z)$ gives the probability that X takes on the value z.
○ $\int_a^b f_X(z)\ dz$ gives $\mathbb{P}(a < X \leq b)$, for values $a < b$
○ Conditioning is impossible for continuous random variables, because all outcomes have probability 0.

Solution:

$\int_a^b f_X(z)\ dz$ gives $\mathbb{P}(a < X \leq b)$. $\mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a)$
$= \int_{-\infty}^b f_X(z)\ dz - \int_{-\infty}^a f_X(z)\ dz = \int_a^b f_X(z)\ dz$

(d) How do you calculate the marginal value $p_X(x)$ from the joint pmf $p_{X,Y}(x,y)$?

○ $\sum_x p_{X,Y}(x,y)$
○ $\sum_y p_{X,Y}(x,y)$
○ $\int_{-\infty}^\infty p_{X,Y}(x,y)dy$
○ You cannot calculate the marginal from the joint pmf. **Solution:**

$\sum_y p_{X,Y}(x,y)$. By LTP, $p_X(x) = \mathbb{P}(X = x) = \sum_{y \in \Omega_Y} \mathbb{P}(X = x \cap Y = y) = \sum_y p_{X,Y}(x,y)$

(e) We saw a "Las Vegas" algorithm; a version of quicksort that uses randomness. What is true about this algorithm?

○ One should never use this algorithm as success cannot be guaranteed.
○ One should run the algorithm (with independent randomness), repeatedly to increase success probability.
○ One should run the algorithm (just once, as is). **Solution:**

One should run the algorithm (just once, as is).

(f) Cov$(X, Y) > 0$ indicates

○ X and Y are independent.

○ X being positive tends to indicate that Y is likely to be negative.

○ X being more than its expectation tends to indicate that Y is likely to be more than its expectation.

○ X being more than its expectation tends to indicate that Y is likely to be less than its expectation. **Solution:**

X being more than its expectation tends to indicate that Y is likely to be more than its expectation.

Each of the questions below has exactly one **correct** answer.
Fully fill in the circle of the best option below.

(g) Recall that $\Phi(z) = \mathbb{P}(Z \leq z)$, where $Z \sim \mathcal{N}(0, 1)$. If $k > 0$, select which of the following is equal to $\Phi(-k)$?

○ $\Phi(k)$

○ $1 - \Phi(k)$

○ $\Phi(k + 1)$

Solution:

$1 - \Phi(k)$

(h) Suppose we have random samples $X_1, X_2, \dots, X_n$ from a $\mathcal{N}(\theta, \sigma^2)$ random variable. Which of the following estimators is **biased**?

I. $\hat{\theta_1} = X_1$

II. $\hat{\theta_2} = \frac{1}{n-1} \sum_{i=1}^{n} X_i$

○ $\hat{\theta_1}$ only

○ $\hat{\theta_2}$ only

○ $\hat{\theta_1}$ and $\hat{\theta_2}$

○ Neither

Solution:

$\hat{\theta_2}$ only.

$\mathbb{E}[\hat{\theta_1}] = \mathbb{E}[X_1] = \theta$, so unbiased.

$\mathbb{E}[\hat{\theta_2}] = \mathbb{E}[\frac{1}{n-1} \sum_{i=1}^{n} X_i] = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n-1} \cdot n \cdot \theta \neq \theta$, so biased.

(i) Let $X_1, X_2, \dots, X_n$ be independent Bernoulli random variables, with $X = \sum_{i=1}^{n} X_i$ being the sum of these random variables. And suppose we know $\mathbb{E}[X]$ but do not know the parameters for each individual Bernoulli random variable. If we want to bound $P(X \geq t)$ for some fixed t, which of the following tail bounds can we **not** use here?

○ Markov's inequality

○ Chebyshev's inequality

○ Chernoff bound

○ More than one of the above are not usable here

○ None of the above; all of the above tail bounds are usable here **Solution:**

We accepted two answers to this question. If t is negative, none of the bounds work; thus we accepted "More than one..."

We also accepted "Chebyshev's inequality." Since we know $\mathbb{E}[X]$ and that X is non-negative, we can use Markov's inequality. Since we know $\mathbb{E}[X]$ and that X is a sum of independent Bernoulli rvs, we can use the Chernoff bound. But since we don't know the parameters of these Bernoulli rvs, we don't know Var(X), so we can't use Chebyshev's inequality.

(j) When applying the Central Limit Theorem to approximate the sum of n (independent) Bernoulli random variables with a Gaussian (Normal) distribution, we use a continuity correction because:

○ The CLT only applies to a sum of continuous i.i.d random variables.

○ The continuity correction guarantees we only ever over-estimate the probability of the event we're interested in.

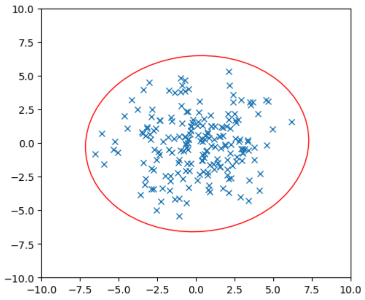
○ The continuity correction tends to make estimates more accurate, especially for small n.

○ The continuity correction tends to make computations simpler. **Solution:**

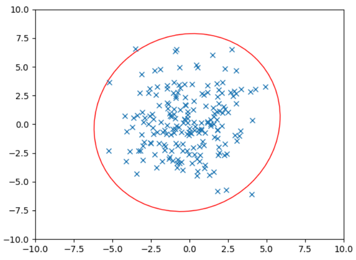
The continuity correction tends to make estimates more accurate, especially for small n.

2. Small Problems [11 points]

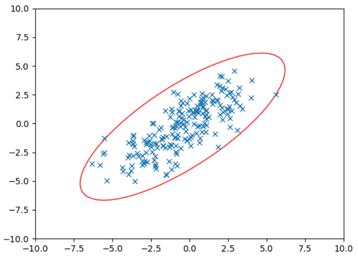
(a) One of the images below is 200 independent draws from a 2D Gaussian X, Y with mean $\begin{bmatrix}\mathbb{E}[X]\\ \mathbb{E}[Y]\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$ and covariance matrix $\begin{bmatrix}4 & -3\\ -3 & 7\end{bmatrix}$. Each image shows the range $[-10, 10]$ on both axes.
Fill in the circle next to the correct image. [3 points]



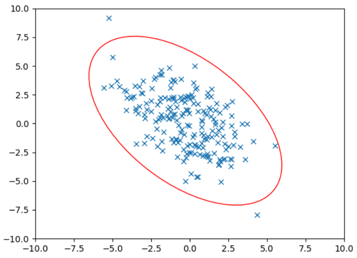
○



○



○



○

Solution:

Since $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ are negative, X and Y are negatively correlated. This means that as X gets greater than $\mathbb{E}[X]$, Y gets more likely to be less than $\mathbb{E}[Y]$. The bottom-right graph is the only one to display this relationship.

(b) Let $X \sim \mathcal{N}(1, 4)$. Give $\mathbb{P}(X \le 2)$ as a decimal (you must give a decimal here, you may not use Φ notation in your final answer). [4 points] **Solution:**

$\mathbb{P}(X \le 2) = \mathbb{P}(Z \le \frac{2-1}{2}) = \Phi(0.5) = 0.69146$

(c) A donut shop sells boxes of 12 donuts; they have flavors of chocolate, strawberry, blueberry, and coconut. It's nearing closing time, so there are only two strawberry donuts left (but an unlimited number of the others). How many distinct boxes are possible?
Your answer should be relatively simple; summation symbols or "..." are not simple, but a few plus signs are fine to leave unsimplified. [4 points] **Solution:**

The number of ways to choose the box if the box have no strawberry donuts is $\binom{14}{2}$; The number of ways to choose the box if the box have 1 strawberry donut is $\binom{13}{2}$; The number of ways to choose the box if the box have 2 strawberry donuts is $\binom{12}{2}$; so the total number of ways is $\binom{14}{2} + \binom{13}{2} + \binom{12}{2}$

3. [RV Zoo] Wacky Races [25 points]

Robbie and three TAs decide to have a friendly race around Seattle, to decide who will have to do all of the final exam grading. Contestant race in their own wacky vehicles that can finish the race in different amounts of time, independent of each other. Here are the contestants and their wacky vehicles:

- Robbie with his Rubbernoulli-Burner finishes in $R = 10X + 30$ minutes, where $X \sim \text{Ber}(p)$. (Treat p as an unknown constant).
- Charles with his Binomiobile finishes in $B \sim \text{Bin}(128, 0.25)$ minutes.
- Allie with her Gaussoline-Guzzler finishes in $G \sim \mathcal{N}(\mu = 33, \sigma^2 = 9)$ minutes.
- Edward with his Exponentiator 9000 finishes in $E \sim \text{Exp}(\frac{3}{100})$ minutes.

(a) Give the PMF of R. [4 points] **Solution:**

$$\mathbb{P}(R = r) = p_R(r) = \begin{cases} 1 - p & \text{if } r = 30 \\ p & \text{if } r = 40 \\ 0 & \text{otherwise} \end{cases}$$

(b) What is the probability that Robbie with his Rubbernoulli-Burner finishes the race before Edward with his Exponentiator 9000, that is, $\mathbb{P}(R < E)$? Use the law of total probability, partitioning on the two possible values of R. Do **not** evaluate or simplify; leave your expression in terms of expressions like "$\mathbb{P}(R = 30)$" (4 points) **Solution:**

$\mathbb{P}(R < E) = \mathbb{P}(R = 30)\mathbb{P}(R < E \mid R = 30) + \mathbb{P}(R \neq 30)\mathbb{P}(R < E \mid R \neq 30)$, using Law of Total Probability.

(c) What is the probability that Edward with his Exponentiator 9000 takes over 30 minutes to finish the race, that is, $\mathbb{P}(E > 30)$? Your answer must be an expression that could be evaluated with just a scientific calculator (e.g., no integrals). [3 points] **Solution:**

$$\mathbb{P}(E > 30) = 1 - \mathbb{P}(E \leq 30) = 1 - (1 - e^{-0.03 \cdot 30}) = e^{-0.03 \cdot 30}$$

(d) Suppose Robbie with his Rubbernoulli-Burner finished the race before Edward with his Exponentiator 9000. Use Bayes' Theorem to calculate the probability that Robbie finished in 30 minutes, that is, $\mathbb{P}(R = 30|R < E)$. Your final answer can be given in terms of p. You do not need to fully simplify, but only expressions involving p and constants that can be evaluated with a calculator may remain (still no integrals). [10 points] **Solution:**

$$
\begin{aligned}
&\mathbb{P}(R = 30 \mid R < E) \\
&= \frac{\mathbb{P}(R = 30)\mathbb{P}(R < E \mid R = 30)}{\mathbb{P}(R < E)} && \text{Bayes' rule} \\
&= \frac{\mathbb{P}(R = 30)\mathbb{P}(R < E \mid R = 30)}{\mathbb{P}(R = 30)\mathbb{P}(R < E \mid R = 30) + \mathbb{P}(R = 40)\mathbb{P}(R < E \mid R = 40)} && \text{LTP or (b)} \\
&= \frac{\mathbb{P}(R = 30)\mathbb{P}(30 < E \mid R = 30)}{\mathbb{P}(R = 30)\mathbb{P}(30 < E \mid R = 30) + \mathbb{P}(R = 40)\mathbb{P}(40 < E \mid R = 40)} && \text{fixed } R\text{-values} \\
&= \frac{\mathbb{P}(R = 30)\mathbb{P}(30 < E)}{\mathbb{P}(R = 30)\mathbb{P}(30 < E) + \mathbb{P}(R = 40)\mathbb{P}(40 < E)} && \text{independence} \\
&= \frac{(1-p)e^{-0.03 \cdot 30}}{(1-p)e^{-0.03 \cdot 30} + (p)e^{-0.03 \cdot 40}}
\end{aligned}
$$

(e) Give the expectation of the time for each of the vehicles to finish. Put your final answer directly on the line, in as simplified a form as you can. [4 points total]

- Robbie with his Rubbernolli-Burner finishes $R = 10X + 30$ minutes, where $X \sim$ Ber(p)
- Charles with his Binomiobile finishes in $B \sim$ Bin$(128, 0.25)$ minutes.
- Allie with her Gaussoline-Guzzler finishes in $G \sim \mathcal{N}(\mu = 33, \sigma^2 = 9)$ minutes.
- Edward with his Exponentiator 9000 finishes in $E \sim$ Exp$(\frac{3}{100})$ minutes.

Solution:

- $\mathbb{E}[R] = 10\mathbb{E}[X] + 30 = 10p + 30$ minutes
- $\mathbb{E}[B] = 128 \cdot 0.25 = 32$ minutes
- $\mathbb{E}[G] = \mu = 33$ minutes
- $\mathbb{E}[E] = \frac{1}{(\frac{3}{100})} = \frac{100}{3}$ minutes

4. My Cabbages! [20 points]

A cabbage merchant tries to sell cabbages from his cart in Omashu. On typical days, he earns 6 gold pieces from selling his cabbages. However, if the Avatar is in Omashu on a day, the merchant will lose his entire cart of cabbages, causing a loss of 15 gold pieces (and no earnings). The Avatar spends $\frac{1}{3}$ of his time in Omashu, independently choosing each day whether to be in the city.

Let X be the net profit the merchant earns in 20 days.

(a) What is the expected net profit the merchant earns in 20 days, i.e., $\mathbb{E}[X]$? [5 points] **Solution:**

Let X_i be the net profit the merchant earns on the i^{th} day.

$\mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{20} X_i] = \sum_{i=1}^{20} \mathbb{E}[X_i] = \sum_{i=1}^{20}(6 \cdot \mathbb{P}(X_i = 6) - 15 \cdot \mathbb{P}(X_i = -15) = \sum_{i=1}^{20} \left(6 \cdot \frac{2}{3} - 15 \cdot \frac{1}{3}\right) = \sum_{i=1}^{20} -1 = -20.$

(b) What is $\text{Var}\,(X)$? [5 points] **Solution:**

As the X_is are independent, we can calculate $\text{Var}\,(X)$ as follows.

Note that $\mathbb{E}[X_i^2] = 6^2 \cdot \frac{2}{3} + (-15)^2 \cdot \frac{1}{3} = 99$

$\text{Var}\,(X) = \text{Var}\left(\sum_{i=1}^{20} X_i\right) = \sum_{i=1}^{20} \text{Var}\,(X_i) = \sum_{i=1}^{20} \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \sum_{i=1}^{20} 99 - (-1)^2 = \sum_{i=1}^{20} 98 = 20 \cdot 98$

The cabbage merchant decides to stock two types of cabbages: red and green. On a particular day, he stocks 10 red cabbages and 4 green cabbages. Every customer randomly chooses one of the **not-yet-purchased** cabbages to buy.

Let R be the event that the **first** cabbage bought is red. Let G be the event that the **second** cabbage bought is green.

(c) What is $\mathbb{P}(G|R)$? [3 points] **Solution:**

$\mathbb{P}(G|R) = \frac{4}{13}$

(d) What is $\mathbb{P}(G)$? [4 points] **Solution:**

Using LTP, we find that $\mathbb{P}(G) = \mathbb{P}(G|R)\mathbb{P}(R) + \mathbb{P}(G|R')\mathbb{P}(R') = \frac{4}{13} \cdot \frac{10}{14} + \frac{3}{13} \cdot \frac{4}{14}$.

(e) Are R and G independent? Justify your answer in reference to your results from (c) and (d) or a new computation. [3 points]

Solution:

R and G are not independent, since $\mathbb{P}(G|R) \neq \mathbb{P}(G)$.

5. [Concentration] Wild Berries [15 points]

You've stumbled upon the perfect opportunity to mesh two of your favorite hobbies - hiking and baking. There are multiple scenic hikes in Washington that are known for their abundance of berries. In order to bake a blueberry pie, you need at least 260 juicy blueberries.

(a) The first hike you go on, is the Shannon Ridge hike which is near the Mt. Baker area. On average, people have reported collecting 220 blueberries on this seven mile hike. Use Markov's inequality to bound the probability that you collect at least 260 berries on this hike. **Solution:**

> Let X be the amount of berries collected on this hike. We know $E[X] = 220$, and we can apply Markov's inequality since X is non-negative. $P(X \geq 260) \leq \frac{220}{260}$.

(b) After doing more research, you learn that the number of blueberries collected on this hike is distributed with variance 80. Use the Chebyshev's inequality to bound the same probability from part (a). In your answer you must precisely state the event which you are using Chebyshev to bound. **Solution:**

> We know that $\mathbb{E}[X] = 220$. And we know that $P(X \geq 260) = P(X - 220 \geq 260 - 220) = P(X - \mathbb{E}[X] \geq 40) \leq P(|X - \mathbb{E}[X]| \geq 40)$. We can bound this using Chebyshev's inequality:
> $P(X \geq 260) \leq P(|X - \mathbb{E}[X]| \geq 40) \leq \frac{\text{Var}(X)}{40^2} = \frac{80}{40^2} = 0.05$

(c) Over the summer months, you plan to go on four more hikes (so five hikes in total), where the number of berries collected in each hike is distributed the same way as the Shannon Ridge hike (but each hike is independent). That is, you will get five independent trails with average 220 berries and variance 80.

Bound the probability that none of your hikes have enough berries for you to make a pie. (Since berries taste best when fresh, all blueberries used in the pie must be collected on the same hike). Write an inequality, and name the tool you used to get it (Hint: use your answer from (b)).

Solution:

> Let X_i be the event that you find enough berries on the i'th hike. We want to bound $\mathbb{P}(X_1 \cup ... \cup X_5)$ which is, by the union bound, at most $\mathbb{P}(X_1) + \cdots + \mathbb{P}(X_5)$. Since, from part (b), each $\mathbb{P}(X_i)$ is at most 0.05, we know that $\mathbb{P}(X_1) + \cdots + \mathbb{P}(X_5)$ is at most $0.05 \cdot 5 = 0.25$. But since we want to bound the probability that none of our hikes have enough berries to make a pie, we are bounding 1 - $\mathbb{P}(X_1 \cup ... \cup X_5)$, which must be at least 1 - 0.25.

6. [MLE] Swing For The Fences [13 points]

You are a big baseball fan, who loves watching home runs (and therefore hates waiting to watch home runs). Luckily, you are a fan of a very consistent team. Each player has identical hitting skills, which are independent of all their teammates.

If it's a **windy** day, it's easy to hit home runs. Each player hits a home run with probability $3/5$.
If it's a **calm** day, it's hard to hit home runs. Each player hits a home run with probability only $1/5$.

You may assume the only two types of weather are calm and windy.

You watch batters hit, and record how many players hit until the first time one hits a home run. Today, the fourth batter was the first to hit a home run. In this problem, you will use this information to determine the maximum likelihood estimator for the state of the weather (i.e., whether it was windy or calm). In this problem, you are estimating a Boolean parameter ($\{\texttt{windy}, \texttt{calm}\}$) which is different than what we've done in examples from class.

(a) What is the likelihood of it taking (exactly) 4 hitters to see the first home run on a windy day? You do not have to simplify. [4 points] **Solution:**

> Let X be the number of hitters until you see the first home run (i.e., X is geometric once you know which type of day it is).
>
> $L(X = 4; \text{windy}) = (2/5)^3(3/5)$

(b) What is the likelihood of it taking (exactly) 4 hitters to see the first home run on a calm day? You do not have to simplify. [4 points] **Solution:**

> $L(X = 4; \text{calm}) = (1/5)^3(4/5)$

(c) After observing it take (exactly) 4 hitters to see the first home run, describe the process you would use to determine the MLE if you had access to a calculator (and/or computer). You should describe in enough detail that someone who doesn't know what an MLE is (but does have a calculator and computer) could perform the computation for you. [3 points] **Solution:**

> Evaluate the expressions in (a) and (b). If the answer in (a) is larger, report `windy`, otherwise report `calm`.

(d) Parts a and b sound a lot like a Bayes' Rule problem! You don't have enough information to use Bayes' Rule here. What information are you missing that you would need to find $\mathbb{P}(\text{windy}|\text{exactly } 4 \text{ hitters})$? You may give your answer in notation or in English, but if you use notation be sure we'll understand any variables or events you refer to. [2 points] **Solution:**

> You would need to know the probability of it being a windy day. (i.e., the prior)

7. Grading Morale [1 point]

This time, it's **your** turn to give us a probability problem. Come up with your own probability question for us to answer; the wackier, the better. (**As long as you don't leave this page blank, you'll get full credit for this part.**)