1	2	3	4	5	6	7	8	9	10	Total

CS 215: Data Interpretation and Analysis 2024, Mid-Semester exam

September 18, 2024.
1:30–3:30 pm

Roll: _____

Name: _____

Mode: Credit/Audit/Sit-through _____

Write all your answers in the space provided. Do not spend time/space giving irrelevant details or details not asked for. Use the marks as a guideline for the amount of time you should spend on a question.

Reference notes

- If $X \sim \text{Poi}(\lambda)$ then $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$. $\mathbb{E}[X] = \lambda$, MGF $\phi(t) = \mathbb{E}\left[e^{tX}\right] = e^{\lambda(e^t-1)}$
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $p(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$. $\mathbb{E}[X] = \mu$, MGF $\phi(t) = \mathbb{E}\left[e^{tX}\right] = e^{\mu t + \sigma^2 t^2/2}$
- If $X \sim U(a, b)$ then $p(X) = 1$ for $X \in [a, b]$ and 0 otherwise. $\mathbb{E}[X] = (a + b)/2$
- $Var(X) = E[X^2] - (E[X])^2$
- Jensen's inequality states that $\psi(\mathbb{E}[X]) \leq \mathbb{E}[\psi(X)]$ for convex functions ψ. For concave functions $\psi(\mathbb{E}[X]) \geq \mathbb{E}[\psi(X)]$.
- Markov's Inequality: Let X be any non-negative random variable and $a > 0$, then $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.
- One-sided Chebyshev inequality: For a random variable X with mean μ and variance σ^2: for every $\tau > 0$, we have $P(X - \mu \geq \tau) \leq \frac{\sigma^2}{\sigma^2+\tau^2}$, and for every $\tau < 0$, we have, $P(X - \mu \geq \tau) \geq 1 - \frac{\sigma^2}{\sigma^2+\tau^2}$.
- If $X_1, \ldots, X_n$ are independent random variables, then their joint density $f(X_1, \ldots, X_n) = \prod_j f(X_j)$.
- The mean value theorem states that if f is a continuous function on the closed interval $[a, b]$ and differentiable on the open interval (a, b) then there exists a point $c \in (a, b)$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$

1. Consider an infinitely long massless rod with holes at integer points. For each time $t = 1, 2, \cdots, T$, a sample x_i is drawn from Poi(λ) and a point mass of mass 1 unit is added to the x_i-th hole. Compute the approximate center of mass c of the system at time T, for large T. Recall that the center of mass of a distribution of mass in space is the unique point where the weighted relative position of the distributed mass sums to zero. Assume a hole can hold multiple point mass. ..2

   Since we have infinite trials, the relative mass at position k will be proportional to $P(X = k)$ which is Poisson distribution. By defn of COM: $\sum_i p_i(x_i - c) = 0$, This implies that $c = \mathbb{E}[X] = \lambda$

2. Let X be a discrete random variable with PMF P taking on finitely many values. Suppose m is a median of P. We showed in class that $\mathbb{E}[|X - m|] \leq \mathbb{E}[|X - c|]$ for any $c \in \mathbb{R}$. Now show that $|\mu - m| \leq \sigma$, where μ is the mean and σ is the standard deviation of X. [Use Jensen's inequality.] ..3

   Since $\psi(x) = |x|$ is a convex function, we apply Jensen's and get $|\mathbb{E}[X - m]| \leq \mathbb{E}[|X - m|]$ i.e., $|\mu - m| \leq \mathbb{E}[|X - m|]$.

   Next, $\mathbb{E}[|X - m|] \leq \mathbb{E}[|X - \mu|] = \mathbb{E}\left[\sqrt{(X-\mu)^2}\right] \leq \sqrt{\mathbb{E}[(X-\mu)^2]} = \sigma$ (Note: $\psi(x) = \sqrt{(x)}$ is a concave function.)

   Using Jensen on either of the function $|\mathbb{E}[X - m]|$ or $\mathbb{E}\left[\sqrt{(X-\mu)^2}\right]$ ... 1 mark each

   Using the above inequalities with $\mathbb{E}[|X - m|] \leq \mathbb{E}[|X - \mu|]$ to complete the proof ... 1 mark

3. Let S be a continuous random variable modeling the examination scores of students in a class. Scores are real-valued within the range $[0, 100]$. The mean of S is 60.

   - Provide a tight upper bound for $P(S \geq 90)$ [Hint: Use one of the inequalities listed on page 1.] ..1

     Using Markov inequality $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ we get the answer as $60/90 = 2/3$.

     Correct application of Markov inequality ... 1 mark

   - Show also that this bound is tight, i.e. provide a $P(S)$ where the bound is achieved while $\mathbb{E}[S] = 60$ ..3

     Here is one $P(S) = \begin{cases} 2/3 & \text{if } x = 90 \\ 1/3 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$

     Valid probability distribution and E[S] = 60 for P(S) ... 1.5 marks

     Tight bound achieved (equality established) with provided distribution ... 1.5 marks

   - Using an additional fact that the standard deviation of S is 10, derive a non-trivial upper bound for $P(S \leq 20)$. [Hint: Use one of the inequalities on page 1] ..2

     Apply One-sided Chebyshev inequality with $\tau = -40$ and we get $P(S \leq 20) = 100/(100 + 1600)$

     Correct application of One-sided Chebyshev inequality ... 1 mark

     Correct value of τ used ... 1 mark

4. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables distributed as Poi(λ). Show that the distribution of $Y = \sum_{i=1}^{n} X_i$ is Poisson. ..3

Use proof technique using MGF that we discussed for Gaussian distribution.

$$\mathbb{E}\left[(\right] e^{tY}) = \mathbb{E}\left[(\right] e^{t(X_1,X_2,\cdots,X_n)}) \quad (1)$$

$$= \prod_i \mathbb{E}\left[(\right] e^{t(X_i)}) \qquad \text{independence of } X_i\text{s} \quad (2)$$

$$= \prod_i e^{\lambda(1+e^t)} \quad (3)$$

The last term MGF of Poisson with mean $n\lambda$

(MGF Method)
Step 1 above ... 1 mark
Using independence assumption to simplify to step 2 above... 1 mark
Substituting MGF of X_i in the expression/ Write expression for MGF of single X_i ... 0.5 mark
Showing the final form is also MGF of Poisson with new λ' ... 0.5 marks

(Brute force calculation)
Proving X + Y is poisson, for two poisson RVs X and Y ... 2 marks
Induction for n variables ... 1 mark

OR

Correctly proving for n variables - simplifying using multinomial theorem ... 3 marks

(Partial marks deducted for small errors)

5. Let X, Y, Z be three independent continuous variables uniformly distributed in $[0,\ 1]$. What is the probability that $X + Y < Z$? ..2

   Answer. $\frac{1}{6}$

   We need to compute $\int_0^1 \int_0^z \int_0^{z-y} 1.dxdydz = 1/6$
   1m for right limits on integral
   1m for simplification
   Also considered other ways of finding volume of tetrahedron to be 1/6.

6. Let A and B be two discrete random variables for which the $\Pr(A|B)$ and $P(B)$ distributions are specified below. For example, from the table we know that $\Pr(A = 3|B = 2) = 0.2$

	B=1	B=2
A=1	0.4	0.3
A=2	0.3	0.5
A=3	0.3	0.2

B=1	B=2
0.4	0.6

   What is the probability that A is greater than B ..2

   Answer. $\Pr(B = 1)(\Pr(A = 2|B = 1) + \Pr(A = 3|B = 1)) + \Pr(B = 2)\Pr(A = 3|B = 2) = 0.4 * 0.6 + 0.6 * 0.2 = 0.36$

7. Let $\mathcal{N}(\mu, \sigma^2)$ denote a Gaussian distribution with mean μ and variance σ^2. For $X_1 \sim \mathcal{N}(2, 4)$ and $X_2 \sim \mathcal{N}(3, 9)$, where X_1 and X_2 are independent. For $Y_1 = 2X_1 + 3$ and $Z = Y_1 + X_2$, find the value of $M + \text{Var}(Z)$, where M denotes the mode of Z's pdf. ..2

Answer: 35

Solution: $\mathrm{E}[Z] = \mathrm{E}[2X_1+3+X_2] = 2\mathrm{E}[X_1]+3+\mathrm{E}[X_2] = 10$. $\mathrm{Var}(Z) = \mathrm{Var}(2X_1+3+X_2) = 4\mathrm{Var}(X_1)+\mathrm{Var}(X_2) = 25$. Y_1 and Z are both Gaussian random variables $\implies$ $M = \mathrm{E}[Z]$.

8. Consider a Gaussian random variable $X \sim \mathcal{N}(3, \pi^2)$. For $Y = X^3 + X$, find the value of $P(Y > 30)$. ..3

Answer: 0.5

Solution: $g(x) = x^3 + x$ is an increasing function and $g(3) = 30$. $P(Y > 30) = P(g(X) > 30) = P(X > 3) = 0.5$ (Gaussian pdf is symmetric about the mean).
+1.5 marks for explaining why $P(Y > 30) = P(X > 3)$
+1.5 marks for solving to get 0.5

9. Density estimation is the problem of reconstructing the probability density function using a set of given data points. Namely, we observe $D = \{X_1, \ldots, X_n\}$ and we want to recover the underlying probability density function generating our dataset. This is a central topic in statistical research. Here we will focus on the perhaps simplest approach: histogram. For simplicity, we assume that $X_i \in [0, 1]$ so $p(x)$ is non-zero only within $[0, 1]$. We also assume that $p(x)$ is smooth and $|p'(x)| \leq L$ for all x (i.e. the derivative is bounded). The histogram partitions the set $[0, 1]$ into M bins as follows:

$$B_1 = [0, \frac{1}{M}), \ldots B_j = [\frac{j-1}{M}, \frac{j}{M}) \ldots B_M = [\frac{M-1}{M}, 1]$$

Each bin B_j is associated with a parameter θ_j, and we use it to estimate a density for a $x \in [0, 1]$ as

$$\hat{p}(x) = \prod_{j=1}^{M} \theta_j^{I(x \in B_j)}$$

where $I(q)$ is the indicator function, which is 1, if the predicate q is true, else false. The intuition of this density estimator is that the histogram assigns equal density value to every point within the bin.

(a) If you use the dataset D to derive a maximum likelihood estimator for the θ_j parameters, show that the maximum likelihood estimator for the θ_js is:

$$\hat{\theta}_j = \frac{\text{number of observations within } B_j}{n} \times \frac{1}{\text{length of bin}} = \frac{M}{n} \sum_{i=1}^{n} I(X_i \in B_j)$$

Note you will have to ensure that $\int_{x=0}^{1} \hat{p}(x)dx = 1$ when estimating the θ parameters. ..4 Let $N_j = \sum_i I(X_i \in B_j)$ Log likelihood function is $LL(D) = \sum_j N_j \log \theta_j$ and the constraint $\int_{x=0}^{1} \hat{p}(x)dx = M \sum_j \theta_j = 1$. This implies that $\theta_1 = 1/M - \sum_{j=2}^{M} \theta_j$. Substituting we get $LL(D) = N_1 \log(1/M - \sum_{j=2}^{M} \theta_j) + \sum_{j=2}^{M} N_j \log \theta_j$.

Taking gradient wrt θ_j and equating to zero we get that $\frac{N_j}{\theta_j} = \frac{N_1}{1/M - \sum_{j=2}^{M} \theta_j}$.

This implies that $\theta_j \propto N_j$. Let $\theta_j = cN_j$. Substituting that back in the constraint gives the required result.

Writing the correct likelihood expression ... 1 mark

Simplifying the constraint on the θ_j ... 1 mark

Noting (by putting the constraint in the expression and differentiating, say) that $\theta_j \propto N_j$ ... 1 mark

Combining these to get the final answer ... 1 mark

(b) Calculate $\mathbb{E}\left[\hat{\theta}_j\right]$ in terms of the cumulative distribution function $F(x) = \int_0^x p(y)dy$ of X. ..2

The final expression should be of the form $\frac{1}{1/M}(F(j/M) - F((j-1)/M))$.

Fully correct answer ... 2 marks

1 mark deducted for incorrect/incomplete simplification

(c) The bias of $\hat{p}(x)$ is given as $\mathbb{E}\left[\hat{p}(x)\right] - p(x)$. Calculate an upper bound of the bias in terms of L and M. (**Hint**: Apply the mean value theorem twice.) ..4

Consider an $x \in B_j$. The $\mathbb{E}\left[\hat{\theta}_j\right] = (F(j/M) - F(j/M - 1/M))/(1/M)$

Applying mean value theorem on the rhs we can assume that there exists an $x_j \in B_j$ such that

$(F(j/M) - F(j/M - 1/M))/(1/M) = F'(x_j) = p(x_j)$

$p(x_j) - p(x) = p'(x*)(x_j - x)$ [Applying mean value theorem again]

RHS is $\leq L/M$

Using the previous part to simplify the expression ... 1 mark

Applying mean value theorem twice ... 2 marks

Using the bounds and getting the final answer ... 1 mark

An exercise for the reader: Calculate variance and MSE and find the optimal bin width. Feel free to ask Saksham Rathi (22B1003) if you are stuck.

10. A person is standing on an infinite 2D plane at the Origin $O(0,0)$. He decides to play a game consisting of n steps-

(a) At each step, he chooses a direction (either North or East) with an equal probability of $\frac{1}{2}$.

(b) Now, he moves in the chosen direction by a distance of 2^{-i} units in step i.

Let us assume for simplicity that North is the positive y direction and East is the positive x direction. Thus, a typical position for this person after 3 such steps can be $(0,0) \to (1/2,0) \to (1/2,1/4) \to (5/8,1/4)$ if he chooses the directions to be East, North, East. Now, if the person plays this game for a long time ($n \to \infty$),

(a) Find $E[(|x| + |y|)]$. (This is called the Manhattan distance of a point from the origin.) ..3

> We can observe that $x \geq 0$ and $y \geq 0$. Now $E[(|x| + |y|)] = E[x + y] = \sum_{i=1}^{\infty} 2^{-i} = 1$. (The observation that $x + y = 1$ is important for later parts)
> Alternative solution: $E[(|x| + |y|)] = E[x + y] = E[x] + E[y] = 2E[x]$ x, y are symmetric. Find $E[x] = \sum_{i=1}^{\infty} 2^{-i}/2 = \frac{1}{2}$. Hence $E(|x| + |y|) = 1$

(b) Find the Expected value of the square of the Euclidean distance. That is, evaluate, $E[x^2 + y^2]$. ..4

By linearity of expectation and symmetry of x and y,

$$E[x^2 + y^2] = E[x^2] + E[y^2] = 2E[x^2]$$

Now we know that,

$$E[x^2] = Var(x) + (E[x])^2$$

Maybe we can have the students prove this property of $Var(X) = E[X^2] - (E[X])^2$ in a subpart?
Now to find $Var(x)$, let us denote x as the sum of infinitely many bernoulli-like RVs (instead of attaining 1 they attain 2^{-i} with probability $\frac{1}{2}$) corresponding to each step $X_1, X_2, \ldots$. As all these RVs are independent,

$$Var(x) = \sum_{i=1}^{\infty} Var(X_i)$$

For any X_i,

$$Var(X_i) = E[X_i^2] - (E[X_i])^2 = \frac{2^{-2i}}{2} - (\frac{2^{-i}}{2})^2 = \frac{2^{-2i}}{4}$$

Thus,

$$Var(x) = \sum_{i=1}^{\infty} \frac{2^{-2i}}{4} = \frac{1}{12}$$

So,

$$E[x^2 + y^2] = 2E[x^2] = 2(\frac{1}{12} + \frac{1}{4}) = \frac{2}{3}$$

ALITER: Students may notice X is uniform and simply derive $E[x^2]$ by integration

(c) Find the Expected Value of the Euclidean Distance, evaluate $E[\sqrt{x^2 + y^2}]$. (Hard question, attempt only after you have done the easier ones. You can leave the integral without computing the closed form.) ..5

MISSTEP- $E[\sqrt{x^2+y^2}] \neq \sqrt{E[x^2+y^2]}$
Claim- X seems to follow a uniform distribution $U(0,1)$
Proof- For any two numbers a and $b \in [0,1)$, $P(x=a) = P(x=b)$. This property can be seen from that every distance $a \in [0,1)$ is attainable through exactly one combination of "moves" and each such combination is equally probable because it is equally likely whether or a move is executed or not.
This is equivalent to saying that every number $\in [0,1)$ has exactly one binary representation and as each bit is equally likely to be 0 or 1, every number is equally likely.
Induction can also be used to prove this fact, assume after a step i, $x = p$, then the possible range for x is $[p, p+2^{-i})$. Now in the next step, dependent on whether x chooses to move to the right or not, the range get halved with equal probability of landing on either $[p, p+2^{-i-1})$ or $[p+2^{-i-1}, p+2^{-i})$. Thus if both these ranges were to follow the uniform distribution, then the distribution of x is also uniform. We know that these ranges will be uniform as after all steps, there will only be an infinitesimal movement on either side with each step with equal probability.
Once it is suffficiently established that x is uniform and $y = 1-x$, we know that the PDF of $U(0,1) = 1$,

$$\begin{aligned} E[\sqrt{x^2+y^2}] &= E[\sqrt{x^2+(1-x)^2}] \\ &= \int_{x=0}^{x=1} \sqrt{x^2+(1-x)^2} f_X(x) dx \\ &= \int_{x=0}^{x=1} \sqrt{x^2+(1-x)^2} dx \end{aligned}$$

We should give the integral $\int \sqrt{u^2+1} du = \frac{1}{2}(ln(u+\sqrt{u^2+1}) + u\sqrt{u^2+1}$ or we could have them leave it as an integral

$$E[\sqrt{x^2+y^2}] = \frac{ln(\sqrt{2}+1)+\sqrt{2}}{2\sqrt{2}}$$

Total: 45