CS 215 : Data Analysis and Interpretation

(Instructor : Suyash P. Awate)

Mid-Semester Examination (Closed Book)

Roll Number: ______________________________

Name: ______________________________

For all questions, if you feel that some information is missing, make justifiable assumptions, state them clearly, and answer the question.

Relevant Formulae

• Poisson: $P(k|\lambda) := \lambda^k \exp(-\lambda)/(k!)$

• Exponential: $P(x;\lambda) = \lambda \exp(-\lambda x); \forall x > 0$

• Univariate Gaussian:
$$P(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5\frac{(x-\mu)^2}{\sigma^2}\right)$$

---

1. [30 points]

For a random variable X, define $M_X(t) := E[e^{tX}]$ where $t \in \mathbb{R}$. For the range of values of t when $M_X(t)$ exists (e.g., when the function value is finite), $M_X(t)$ is a function of t.

Let the d-th partial derivative of $M_X(t)$ with respect to t be denoted by $M_X^d(t) := \frac{\partial^d}{\partial t^d} M_X(t)$.

(a) (2 points)

For a Bernoulli random variable X, find the closed-form mathematical expression (without any term involving integrals or summations) for $M_X(t)$.

Evaluate $M_X^1(t)$ at $t = 0$, and evaluate $M_X^2(t)$ at $t = 0$.

---

https://proofwiki.org/wiki/Moment_Generating_Function_of_Bernoulli_Distribution

---

(b) (3 points)

For a Binomial random variable X, find the closed-form mathematical expression (without any term involving integrals or summations) for $M_X(t)$.

Evaluate $M_X^1(t)$ at $t = 0$, and evaluate $M_X^2(t)$ at $t = 0$.

---

https://proofwiki.org/wiki/Moment_Generating_Function_of_Binomial_Distribution

---

(c) (3 points)

For a Poisson random variable X, find the closed-form mathematical expression (without any term involving integrals or summations) for $M_X(t)$.

Evaluate $M_X^1(t)$ at $t = 0$, and evaluate $M_X^2(t)$ at $t = 0$.

---

https://proofwiki.org/wiki/Moment_Generating_Function_of_Poisson_Distribution

---

(d) (4 points)

Let X and Y be independent random variables.

Let random variable $Z := X + Y$. Find a mathematical expression for $M_Z(t)$ in terms of $M_X(t)$ and $M_Y(t)$.

---

https://www.le.ac.uk/users/dsgp1/COURSES/MATHSTAT/5binomgf.pdf

---

(e) (6 points)

For a real-valued random variable X and a real-valued scalar a, derive an inequality between the functions $f(a) := P(X \geq a)$ and $g(a; t) := M_X(t)/e^{ta}$ that is valid for all positive values of t. For example, for all a, is $f(a) <$ or $\leq$ or $>$ or $\geq g(a; t), \forall t > 0$?

Similarly, derive an inequality that is valid for all negative values of t.

---

https://en.wikipedia.org/wiki/Chernoff_bound

https://www.probabilitycourse.com/chapter6/6_2_3_chernoff_bounds.php

---

(f) (12 points)

Property: It is well known that if X and Y are two random variables for whom $M_X(t)$ and $M_Y(t)$ are well-defined and identical, then X and Y have the same cumulative distribution function.

Use the aforementioned property (and other properties covered in class, if need be) to find a closed-form analytical expression (without any terms involving integration or differentiation) for the convolution of two Gaussian probability density functions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. State each property used clearly and provide a mathematical justification/proof for that property.

`https://proofwiki.org/wiki/Moment_Generating_Function_of_Gaussian_Distribution`

sum of independent random variables leads/relates to convolution of their PDFs (using and proving this: 4 points)

sum of independent random variables multiplies their MGFs (using this: 2 points; proved above)

find MGF of each Gaussian and take their product (4 points)

see that the product of MGFs is the same as the MGF of a Gaussian PDF with mean $\mu_1 + \mu_2$ and variance $(\sigma_1)^2 + (\sigma_2)^2$. (1 point)

use MGF uniqueness theorem to conclude (1 point)

---

2. [30 points]

Two independent random walkers (Mr. A and Mr. B) start their random walk on the real line (as covered in class) at time $t = 0$ when they are both at location $x = 0$. During their random walk, each of their steps has a length of 1 foot (i.e., 12 inches), and they both independently take one step each second either to the left or to the right (with equal probability).

For all the following questions, when performing simulation, the only random-number generator function you can use is one that simulates a draw from a uniform distribution over the real-line interval $(0, 1)$.

(a) (8 points)

• After taking N steps, what will be the probability mass function (PMF) of the location of each walker ? Give an exact analytical form, without any approximations. What will be the mean and variance of the PMF ?

(4 points)

modeling locations of walkers is equivalent to modeling number of steps taken towards right by each walker

location = 2 * numOfStepsTakenRight - N

numOfStepsTakenRight, for both walkers, are modeled by independent random variables X and Y with Binomial PMFs

Each Binomial PMF, on numOfStepsTakenRight, has mean $N/2$ and variance $N/4$.

So, location will have mean 0 and variance N

• Write pseudocode to simulate both the random walkers, and get the empirical distribution of their locations after N steps.

(4 points)

rng (0); N=1e1; S=1e6;

locations1 = sum (2*(rand(N,S)>0.5)-1);

locations2 = sum (2*(rand(N,S)>0.5)-1);

(b) (12 points)

• After taking N steps, what will be the PMF of the displacement (signed difference in locations) between the random walkers ? Give an exact analytical form, without any approximations. What will be the mean and variance of the PMF ?

(10 points)

locationX = 2 * numOfStepsTakenRightX - N
locationY = 2 * numOfStepsTakenRightY - N
locationX - locationY = 2 * (numOfStepsTakenRightX - numOfStepsTakenRightY)

they meeting together is modeled by the event where random variable $Z := X - Y$ takes a value of 0.

Z has mean 0 and variance $N/2$.

random variable $Y' := N - Y$ (taking non-negative values) has a binomial PMF Bin$(N, 1/2)$.

random variable $Z' := Z + N = X + Y'$ (taking non-negative values) has a binomial PMF Bin$(2N, 1/2)$ with mean N and variance $N/2$.

Z has a PMF that is a shifted version of the PMF of Z' (i.e., $P(Z = a) = P(Z' = a + N)$), with mean 0 and variance $N/2$

so, locationX-locationY will have mean 0 and variance $2N$

• Using the pseudocode simulating the random walkers, write pseudocode to get an empirical estimate of the aforementioned PMF after N steps.

(2 points)

```
rng (0); N=1e1; S=1e6;
locations1 = sum (2*(rand(N,S)>0.5)-1);
locations2 = sum (2*(rand(N,S)>0.5)-1);
displacement = locations1-locations2;
x = [-20:2:20]; px = hist(displacement,x); px=px/S;
plot(x,px,'ro'), grid on; pause, close
```

(c) (10 points)

• After taking N steps, what is the probability of the event that both walkers will be at the same location ? Give an exact analytical form, without any approximations. When $N \to \infty$, is this probability small, medium, or large ?

(4 points)

as $N \to \infty$, $P(Z = 0) \to 0$ because PMF has mean 0 with variance tending to infinity.

• Write pseudocode to evaluate the theoretical probability using the exact analytical form; your code must work even (i.e., not suffer from limitations of finite-precision arithmetic) for large $N > 100$.

(3 points)

```
N=1e3;
probZeroTheory = 2*N/N*0.5*0.5;
for i=1:N-1; probZeroTheory = probZeroTheory *(2*N-i)/(N-i)*0.5*0.5; end; probZeroTheory
probMinusTwoTheory = probZeroTheory * N/(N+1)
```

probPlusTwoTheory = probZeroTheory * N/(N+1)

- Using the pseudocodes written earlier, write pseudocode to empirically evaluate (via simulation) the aforementioned probability after N steps.

(3 points)

```
rng (0); N=1e1; S=1e6;
locations1 = sum (2*(rand(N,S)>0.5)-1); locations2 = sum (2*(rand(N,S)>0.5)-1);
displacement = locations1-locations2;
x = [-20:2:20]; px = hist(displacement,x); px=px/S;
plot(x,px,'ro'), grid on; pause, close

px (x== 0)
px (x== -2)
px (x== +2)
nchoosek(2*N,N) * (0.5)^(2*N)
probZeroSimulation = sum((locations1-locations2) == 0) / S
probMinusOneSimulation = sum((locations1-locations2) == -1) / S
probPlusOneSimulation = sum((locations1-locations2) == +1) / S
probMinusTwoSimulation = sum((locations1-locations2) == -2) / S
probPlusTwoSimulation = sum((locations1-locations2) == +2) / S

probZeroTheory = 2*N/N*0.5^2;
for i=1:N-1; probZeroTheory = probZeroTheory*(2*N-i)/(N-i)*0.5^2; end;
probZeroTheory
probMinusTwoTheory = probZeroTheory * N/(N+1)
probPlusTwoTheory = probZeroTheory * N/(N+1)
```

---

3. [10 points]

For the following questions, the only random-number generator function you can use is one that simulates a draw from a uniform distribution over the real-line interval $(0, 1)$.

(a) (4 points)

Consider a triangular region of the 2D Euclidean plane denoted by the vertices located at $(0, 0)$, $(1, 0)$, and $(1, 4)$. State an algorithm (e.g., via pseudocode) to uniformly sample over this triangular region. Justify why the algorithm should work.

Step 1: Generate, independently, (i) $a \sim U(0, 1)$ and (ii) $b = 4b'$ where $b' \sim U(0, 1)$.

Then (a, b) is uniformly distributed over the rectangular region, and PDF $U(a, b) = U(a)U(b) = 1 * 0.25 = 0.25$.

Step 2: select those (a, b) that fall within the aforementioned triangular region that is a subset of the rectangular region.

PDF $P((a, b)|(a, b)\text{lies within triangle})$

$= P((a, b)\text{and}(a, b)$lies within triangle$)/P((a, b)\text{lies within triangle})$

$= 0.25/0.5 = 0.5$

(b) (6 points)

Given a uniform PDF over the aforementioned triangle, and assuming it is the joint PDF of the random-variable pair (X, Y), mathematically derive the marginal PDFs $P(X)$ and $P(Y)$. Also, provide an algorithm to take as input draws from $P(X, Y)$ and generate draws from $P(X)$ and $P(Y)$.

5+1 points

Marginal $P(X = x) = \int_{y=0}^{y=4x} P(X = x, Y = y)dy = \int_{y=0}^{y=4x} 0.5dy = 2x$

check that PDF integrates to 1: $\int_{x=0}^{x=1} 2xdx = [x^2]_0^1 = 1$

Marginal $P(Y = y) = \int_{x=y/4}^{x=1} P(X = x, Y = y)dx = \int_{x=y/4}^{x=1} 0.5dy = 0.5 - y/8$

check that PDF integrates to 1: $\int_{y=0}^{y=4}(0.5 - y/8)dy = 2 - 1 = 1$

draw $(x, y) \sim P(x, y)$ and then take x: this gives a draw from marginal $P(X)$

draw $(x, y) \sim P(x, y)$ and then take y: this gives a draw from marginal $P(Y)$

---

4. [10 points]

For the exponential probability density function, derive a closed-form mathematical expression (without any term involving integrals or summations) for $E[X^n]$ where n is a positive integer.

`https://en.wikipedia.org/wiki/Exponential_distribution#Mean,_variance,_moments,_and_median`

`https://towardsdatascience.com/moment-generating-function-explained-27821a739035`

(5 points) seeing that $E[X^n]$ can be expressed as $M_X^n(t = 0)$

(3 points) evaluating the MGF of the exponential $M_X(t) = \lambda/(\lambda - t)$

(1 points) deriving the term for the n-th derivative of the MGF, and substituting $t = 0$

(1 points) answer: $E[X^n] = n!/\lambda^n$
