STAT201A
Homeworks, Exams, Labs, & Lectures

Reece D. Huff
rdhuff@berkeley.edu

Fall 2024

Introduction to Probability at an Advanced Level
Yun S. Song
Department of Statistics
University of California, Berkeley

Contents

STAT201A: Introduction to Probability at an Advanced Level September 13th, 2024

Homework # 1: Basics, Independence, Conditioning, & Exchangeability

Reece D. Huff

Problems (Solutions)

1. (**Basic probability**) Assume that $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.7$ and $\mathbb{P}(C) = 0.8$.
   (a) Show that $0.3 \leq P(A \cap B) \leq 0.6$.
   (b) Show that $0.1 \leq \mathbb{P}(A \cap B \cap C) \leq 0.6$.

(a) From the inclusion-exclusion principle, we have that

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \implies 0 \leq \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] \leq 1.$$

Starting with the right inequality, we have

$$\mathbb{P}[A \cap B] \geq \mathbb{P}[A] + \mathbb{P}[B] - 1 = 1.3 - 1 = 0.3$$

thus establishing our lower bound. From the left inequality, we have

$$\mathbb{P}[A \cap B] \leq \mathbb{P}[A] + \mathbb{P}[B] \implies \mathbb{P}[A \cap B] \leq \min\{\mathbb{P}[A], \mathbb{P}[B]\},$$

which directly shows the upper bound. Therefore, we have

$$\boxed{0.3 \leq P(A \cap B) \leq 0.6.}$$

(b) To prove this inequality, we will use the Bonferroni Inequality (sometimes referred to as Boole's inequality). It states that for events $A_1, \ldots, A_n$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have

$$\mathbb{P}\left[\bigcap_{i=1}^{n} A_i\right] \geq \sum_{i=1}^{n} \mathbb{P}[A_i] - (n-1).$$

We will use it to solve for the lower bound, i.e.,

$$\mathbb{P}[A \cap B \cap C] \geq \mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[C] - (3-1) \implies \mathbb{P}[A \cap B \cap C] \geq 0.6 + 0.7 + 0.8 - 2 \implies \mathbb{P}[A \cap B \cap C] \geq 0.1.$$

Next, we note that the intersection is upper bounded when $A \subseteq B \subseteq C$ leading to

$$\mathbb{P}[A \cap B \cap C] \leq \min\{\mathbb{P}[A], \mathbb{P}[B], \mathbb{P}[C]\} \implies \mathbb{P}[A \cap B \cap C] \leq \mathbb{P}[A] = 0.6.$$

Thus, we have shown the desired result

$$\boxed{0.1 \leq \mathbb{P}(A \cap B \cap C) \leq 0.6.}$$

2. (**Independence**) Suppose we roll an unbiased six-sided die $n \geq 3$ times. Let E_{ij} denote the event that the ith and the jth rolls produce the same number. Show that the events $\{E_{ij} \mid 1 \leq i < j \leq n\}$ are pairwise independent but not independent as a family.

In this problem, we are asked to show that events E_{ij} are pairwise independent but not independent as a family. For simplicity, let us consider $n = 3$. In this setting, the events are pairwise independent if

$$\mathbb{P}(E_{12} \cap E_{13}) = \mathbb{P}(E_{12}) \times \mathbb{P}(E_{13}), \quad \mathbb{P}(E_{13} \cap E_{23}) = \mathbb{P}(E_{13}) \times \mathbb{P}(E_{23}), \quad \text{and} \quad \mathbb{P}(E_{12} \cap E_{23}) = \mathbb{P}(E_{12}) \times \mathbb{P}(E_{23}).$$

Starting with the first, we have

$$\mathbb{P}(E_{12} \cap E_{13}) = \mathbb{P}(E_{12} \big| E_{13}) \times \mathbb{P}(E_{13}) = \frac{1}{6} \times \frac{1}{6} \implies \mathbb{P}(E_{12} \cap E_{13}) = \mathbb{P}(E_{12}) \times \mathbb{P}(E_{13}) \quad \checkmark$$

It is easy to verify that the same holds for the other two showing that events $\{E_{ij} \mid 1 \leq i < j \leq n\}$ are pairwise independent.

Moving onto showing that the events are not independent as a family, we again begin with $n = 3$. Independence would imply that

$$\mathbb{P}(E_{12} \cap E_{13} \cap E_{23}) = \mathbb{P}(E_{12}) \times \mathbb{P}(E_{13}) \times \mathbb{P}(E_{23}).$$

However, starting with the left hand side, we have

$$\mathbb{P}(E_{12} \cap E_{13} \cap E_{23}) = \mathbb{P}(E_{12} \cap E_{13} \big| E_{23}) \times \mathbb{P}(E_{23}) = \frac{1}{6} \times \frac{1}{6} \neq \frac{1}{6^3} \implies \mathbb{P}(E_{12} \cap E_{13} \cap E_{23}) \neq \mathbb{P}(E_{12}) \times \mathbb{P}(E_{13}) \times \mathbb{P}(E_{23}).$$

Thus, we have shown that the events are not independent as a family.

3. (**Expectation, joint distribution, uniform distribution**) Let X be a random variable with values $\{1, 2\}$ and Y a random variable with values $\{0, 1, 2\}$. Initially we have the following partial information about their joint probability mass function.

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 1$	1/8		
$X = 2$		0	

Subsequently we learn that $E[XY] = \frac{13}{9}$ and that Y has uniform distribution. Use this information to fill in the missing values of the joint probability mass function table.

To begin, we will leverage the uniformity of Y. We have that

$$\mathbb{P}(Y = y) = \sum_{x=1}^{2} \mathbb{P}(Y = y \cap X = x) = \frac{1}{3} \quad \text{for all } y \in \{0, 1, 2\}.$$

From the above expression, we have immediately infer

$$\mathbb{P}(Y = 0) = \mathbb{P}(Y = 0 \cap X = 1) + \mathbb{P}(Y = 0 \cap X = 2) \implies \frac{1}{3} = \frac{1}{8} + \mathbb{P}(Y = 0 \cap X = 2) \implies \mathbb{P}(Y = 0 \cap X = 2) = \frac{5}{24},$$
$$\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1 \cap X = 1) + \mathbb{P}(Y = 1 \cap X = 2) \implies \frac{1}{3} = \mathbb{P}(Y = 1 \cap X = 1) + 0 \implies \mathbb{P}(Y = 1 \cap X = 1) = \frac{1}{3}.$$

From the uniformity of Y, we also have the relationship,

$$\mathbb{P}(Y = 2) = \mathbb{P}(Y = 2 \cap X = 1) + \mathbb{P}(Y = 2 \cap X = 2) = \frac{1}{3}. \tag{1}$$

Next, we will leverage the expectation $\mathbb{E}[XY] = 13/9$. We have that

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x=1}^{2} \sum_{y=0}^{2} \mathbb{P}(X = x \cap Y = y) \cdot x \cdot y = \frac{13}{9} \\
&= \sum_{x=1}^{2} \sum_{y=1}^{2} \mathbb{P}(X = x \cap Y = y) \\
&= \mathbb{P}(X = 1 \cap Y = 1) \cdot 1 \cdot 1 + \mathbb{P}(X = 1 \cap Y = 2) \cdot 1 \cdot 2 + \mathbb{P}(X = 2 \cap Y = 1) \cdot 2 \cdot 1 + \mathbb{P}(X = 2 \cap Y = 2) \cdot 2 \cdot 2 \\
&= \mathbb{P}(X = 1 \cap Y = 1) + 2\,\mathbb{P}(X = 1 \cap Y = 2) + 2\,\mathbb{P}(X = 2 \cap Y = 1) + 4\,\mathbb{P}(X = 2 \cap Y = 2) \\
&= \frac{1}{3} + 2\,\mathbb{P}(X = 1 \cap Y = 2) + 2(0) + 4\,\mathbb{P}(X = 2 \cap Y = 2) \\
\frac{13}{9} - \frac{1}{3} &= 2\,\mathbb{P}(X = 1 \cap Y = 2) + 4\,\mathbb{P}(X = 2 \cap Y = 2) = \frac{10}{9}
\end{aligned} \tag{2}$$

Combining Equation (1) and Equation (2), we have that

$$2\,\mathbb{P}(X = 1 \cap Y = 2) + 4\,\mathbb{P}(X = 2 \cap Y = 2) = \frac{10}{9} \quad \text{and} \quad \mathbb{P}(X = 1 \cap Y = 2) + \mathbb{P}(X = 2 \cap Y = 2) = \frac{1}{3}$$

from which we have $\mathbb{P}(X = 1 \cap Y = 2) = 1/9$ and $\mathbb{P}(X = 2 \cap Y = 2) = 2/9$. Thus our table becomes

	$Y=0$	$Y=1$	$Y=2$
$X=1$	$1/8$	$1/3$	$1/9$
$X=2$	$5/24$	0	$2/9$

4. (**Conditioning, cumulative distribution function**) You flip a fair coin. If you get tails, you choose a uniformly random number on the interval $[0,2]$. If you get heads, you choose the number 1. Let X be the random variable describing the outcome of that experiment.

   (a) Using the law of total probabilities, calculate $\mathbb{P}(X \leq 1/2)$ and $\mathbb{P}(X \leq 3/2)$.

   (b) Find the cumulative distribution function F_X of X.

   (c) Is X a discrete random variable? Is X a continuous random variable?

   (a) From the law of total probability, we have that

$$\mathbb{P}(X \leq x) = \mathbb{P}(X \leq x \,|\, H)\,\mathbb{P}(H) + \mathbb{P}(X \leq x \,|\, T)\,\mathbb{P}(T).$$

   We can apply this to both $\mathbb{P}(X \leq 1/2)$ and $\mathbb{P}(X \leq 3/2)$ to get

$$\mathbb{P}(X \leq 1/2) = \mathbb{P}(X \leq 1/2 \,|\, H)\,\mathbb{P}(H) + \mathbb{P}(X \leq 1/2 \,|\, T)\,\mathbb{P}(T) = 0 \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} \quad \Longrightarrow \quad \boxed{\mathbb{P}(X \leq 1/2) = \frac{1}{8}}$$

$$\mathbb{P}(X \leq 3/2) = \mathbb{P}(X \leq 3/2 \,|\, H)\,\mathbb{P}(H) + \mathbb{P}(X \leq 3/2 \,|\, T)\,\mathbb{P}(T) = 1 \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} \quad \Longrightarrow \quad \boxed{\mathbb{P}(X \leq 3/2) = \frac{7}{8}}$$

   (b) From our solution to part (a.), we know $\mathbb{P}(X \leq 1/2) = 1/8$ and $\mathbb{P}(X \leq 3/2) = 7/8$. Next, we calculate $\mathbb{P}(X \leq 1)$ as

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X \leq 1 \,|\, H)\,\mathbb{P}(H) + \mathbb{P}(X \leq 1 \,|\, T)\,\mathbb{P}(T) = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \quad \Longrightarrow \quad \boxed{\mathbb{P}(X \leq 1) = \frac{3}{4}}$$

   Stitching our results together it is clear that we have

$$F_X(x) = \begin{cases} 0 & \text{when } x < 0 \\ \frac{x}{4} & \text{when } 0 \leq x < 1 \\ \frac{x}{4} + \frac{1}{2} & \text{when } 1 \leq x < 2 \\ 1 & \text{when } x \geq 2 \end{cases}$$

   A plot of $F_X(x)$ is below:

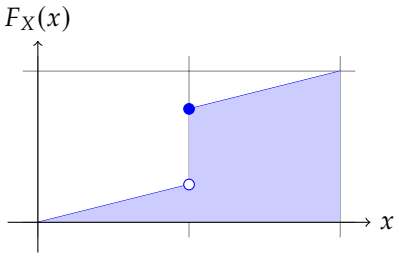


Figure 1: Plot of $F_X(x)$ for $0 \leq x \leq 2$.

   (c) X is clearly a discrete random variable as its cumulative distribution function is not continuous.

5. (**Bounding even moments**) Let X be a random variable. Show that $\mathbb{E}[X^{2k}] \geq (\mathbb{E}[X])^{2k}$ for all positive integers k.

   This result follows directly from Jensen's inequality. Let X be a $\mathbb{R}$-values random variable and g be a convex function $g : X \mapsto \mathbb{R}$, then we have

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

   We note that X^{2k} is convex for all positive integers k. Then from Jensen's inequality, we have

$$\mathbb{E}[X^{2k}] \geq (\mathbb{E}[X])^{2k} \quad \text{for all positive integers } k.$$

6. (**Continuous distributions, probability density function, independence**) Pick a uniformly chosen random point (X, Y) inside the sector delimited by the x-axis, the y-axis and the parabola given by the equation $y = 1 - x^2$.

   (a) Verify that the area of that sector is $2/3$.

   (b) What is the probability that the distance of this point to the y-axis is less than $1/2$?

   (c) What is the probability that the distance of this point to the origin is more than $1/2$?

   (d) Find the p.d.f. of X.

   (e) Find the p.d.f. of Y.

   (f) Are X and Y independent?

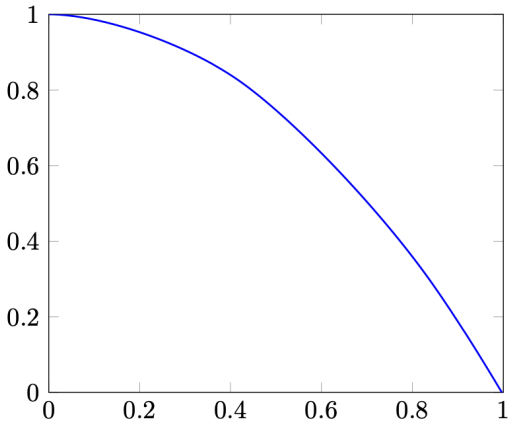


Figure 2: Graph of $y = 1 - x^2$

(a) The area of the sector is given by the integral,

$$\int_0^1 1 - x^2 dx = \left[x - \frac{1}{3}x^3\right]\Big|_{x=0}^{x=1} = 1 - \frac{1}{3} = \frac{2}{3}. \quad \checkmark$$

(b) The probability that the distance of this point to the y-axis is less than $1/2$ is equal to $\mathbb{P}(X < 1/2)$, which we write as

$$\begin{aligned}\mathbb{P}(X < 1/2) &= (\text{Area of sector for } 0 \le x < 1/2) \div (\text{Area of sector}) \\ &= \int_0^{1/2} 1 - x^2 dx \div \frac{2}{3} = \left[x - \frac{1}{3}x^3\right]\Big|_{x=0}^{x=1/2} \div \frac{2}{3} = \left(\frac{1}{2} - \frac{1}{3}\cdot\frac{1}{8}\right) \div \frac{2}{3}\end{aligned}$$

$$\boxed{\mathbb{P}(X < 1/2) = \frac{11}{16}.}$$

(c) Similar to the previous part, we can write the probability that the distance of this point to the origin is more than $1/2$

$$\begin{aligned}\mathbb{P}(\sqrt{X^2 + Y^2} > 1/2) &= 1 - \mathbb{P}(\sqrt{X^2 + Y^2} \le 1/2) = 1 - \frac{1}{4} \cdot \left(\text{Area of circle with radius } r = \frac{1}{2}\right) \div (\text{Area of sector}) \\ &= 1 - \frac{1}{4} \cdot \pi \left(\frac{1}{2}\right)^2 \div \frac{2}{3} = 1 - \frac{\pi}{16} \cdot \frac{3}{2}\end{aligned}$$

$$\boxed{\mathbb{P}(\sqrt{X^2 + Y^2} > 1/2) = 1 - \frac{3\pi}{32}.}$$

(d) The pdf is the derivative of the cdf. Based on the result from part (b), we have that

$$\mathbb{P}(X \le x) = \frac{3}{2}\int_0^x 1 - x^2 dx = F_X(x)$$

From inspection, it is clear that the pdf of X is

$$\boxed{f_X(x) = \frac{3}{2}(1-x^2)}$$

We verify that the area under this curve is equal to 1, i.e.,

$$\frac{3}{2}\int_0^1 1 - x^2 dx = \frac{3}{2}\cdot\frac{2}{3} = 1 \quad \checkmark$$

(e) Next, we rearrange the expression $y = 1 - x^2$ to be in terms of x, i.e, $x = \sqrt{1-y}$. It follows that

$$F_Y(y) = \mathbb{P}(Y \le y) = \frac{3}{2}\int_0^y \sqrt{1-y}dy$$

implying the pdf for Y is

$$\boxed{f_Y(y) = \frac{3}{2}\sqrt{1-y}}$$

Again, we verify that the area under this curve is equal to 1, i.e.,

$$\frac{3}{2}\int_0^1 \sqrt{1-y}dy = \frac{3}{2}\cdot\left[-\frac{2}{3}(1-y)^{3/2}\right]\Big|_{y=0}^{y=1} = 1 \quad \checkmark$$

(f) No, X and Y are not independent. Let's say we picked X u.a.r from $x \in [0,1]$. And then we pick Y. We are not free to pick Y u.a.r. $y \in [0,1]$ because Y must satisfy the constraint $y \le 1-x^2$. Because of this constraint, there is a dependence between X and Y.

7. (**Events, indicators and basic probability inequalities**) Recall that for an event A, we denote the corresponding indicator random variable by $\mathbb{1}\{A\}$ (i.e., $\mathbb{1}\{A\}$ takes value 1 when A occurs and the value 0 when A does not occur). Also recall that the probability $\mathbb{P}(A)$ of A equals the expectation of the random variable $\mathbb{E}[\mathbb{1}\{A\}]$.

(a) Given events $A_1, \ldots, A_n$, show that $\mathbb{1}\left\{\cup_{i=1}^n A_i\right\} = \max_{1\le i\le n} \mathbb{1}\{A_i\}$.

(b) Using the fact observed above (and the following ordering property of expectation: $X \le Y$ implies that $\mathbb{E}[X] \le \mathbb{E}[Y]$), show that

$$\mathbb{P}(\cup_{i=1}^n A_i) \le \sum_{i=1}^n \mathbb{P}(A_i).$$

Note: This is known as the union bound and used quite frequently.

(c) For every event A, show that $\mathbb{1}\{A^c\} = 1 - \mathbb{1}\{A\}$ where A^c denotes the event that A does not occur.

(d) For events $A_1, \ldots, A_n$, show that $\mathbb{1}\left\{\cap_{i=1}^n A_i\right\} = \prod_{i=1}^n \mathbb{1}\{A_i\}$.

(e) Using the above two facts, prove the inclusion-exclusion formula: For events $A_1, \ldots, A_n$,

$$\mathbb{P}(\cup_{i=1}^n A_i) = \Sigma_1 - \Sigma_2 + \Sigma_3 - \Sigma_4 + \cdots + (-1)^{n-1}\Sigma_n$$

where

$$\Sigma_k := \sum_{1\le i_1<i_2<\cdots<i_k\le n} \mathbb{P}(A_{i_1}A_{i_2}\cdots A_{i_k}).$$

(a) Given events $A_1, \ldots, A_n$, we have

$$\mathbb{1}\left\{\cup_{i=1}^n A_i\right\} = \begin{cases} 1 & \text{when in } \cup_{i=1}^n A_i, \\ 0 & \text{when not in } \cup_{i=1}^n A_i. \end{cases}$$

Clearly, the $\max\left\{\cup_{i=1}^n A_i\right\} = 1$. Thus, so long as the event is in one of the A_i for all $i \in [1,n]$, we have $\mathbb{1}\left\{\cup_{i=1}^n A_i\right\} = 1$. Thus, we have

$$\mathbb{1}\left\{\cup_{i=1}^n A_i\right\} = \max_{1\le i\le n} \mathbb{1}\{A_i\}.$$

(b) For this proof, we begin by noting that the max function is convex, i.e., for events $A_1, \ldots, A_n$ and events $B_1, \ldots, B_n$

$$A_i \leq \max_{1\leq i\leq n} A_i \quad \text{and} \quad (1-\lambda)B_i \leq (1-\lambda)\max_{1\leq i\leq n} B_i \quad \Longrightarrow \quad \max_{1\leq i\leq n}\{\lambda A_i + (1-\lambda)B_i\} \leq \lambda \max_{1\leq i\leq n} A_i + (1-\lambda)\max_{1\leq i\leq n} B_i.$$

From this fact, we have that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{E}\left[\mathbb{1}\left\{\cup_{i=1}^{n} A_i\right\}\right] = \mathbb{E}\left[\max_{1\leq i\leq n} \mathbb{1}\{A_i\}\right] \leq \max_{1\leq i\leq n} \mathbb{E}\left[\mathbb{1}\{A_i\}\right] = \max_{1\leq i\leq n} \mathbb{P}(A_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

$$\Longrightarrow \quad \boxed{\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)}$$

where the first inequality comes from Jensen's inequality and the second inequality is because the sum includes the max and all of the other probabilities.

(c) To prove that for every event A, we have $\mathbb{1}\{A^c\} = 1 - \mathbb{1}\{A\}$ where A^c denotes the event that A does not occur, we will use the definition of the indicator function, i.e.,

$$\mathbb{1}\{A\} = \begin{cases} 1 & \text{when } \in A \\ 0 & \text{when } \notin A \end{cases} \quad = \begin{cases} 1 & \text{when } \notin A^c \\ 0 & \text{when } \in A^c \end{cases}$$

From the expression on the right, we see that taking $1 - \mathbb{1}\{A\}$ will result in $\mathbb{1}\{A^c\}$, i.e.,

$$1 - \mathbb{1}\{A\} = 1 - \begin{cases} 1 & \text{when } \notin A^c \\ 0 & \text{when } \in A^c \end{cases} \quad = \begin{cases} 1 & \text{when } \in A^c \\ 0 & \text{when } \notin A^c \end{cases} \quad = \mathbb{1}\{A^c\}. \quad \checkmark$$

(d) For events $A_1, \ldots, A_n$, we have that

$$\mathbb{1}\left\{\cap_{i=1}^{n} A_i\right\} = \begin{cases} 1 & \text{when } \in \cap_{i=1}^{n} A_i \\ 0 & \text{when } \notin \cap_{i=1}^{n} A_i \end{cases} \quad \Longrightarrow \mathbb{1}\left\{\cap_{i=1}^{n} A_i\right\} = \begin{cases} 1 & \text{when } \in A_1 \text{ and } \in A_2 \ldots \text{ and } \in A_n \\ 0 & \text{otherwise} \end{cases}$$

Clearly from our definition above, $\mathbb{1}\left\{\cap_{i=1}^{n} A_i\right\} = 1$ only when the event is in $A_1, \ldots, A_n$, which can be written as

$$\boxed{\mathbb{1}\left\{\cap_{i=1}^{n} A_i\right\} = \prod_{i=1}^{n} \mathbb{1}\{A_i\}.}$$

(e) From the two facts above and de Morgan's law, we have that for events $A_1, \ldots, A_n$,

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \mathbb{E}\left[\mathbb{1}\left\{\cup_{i=1}^{n} A_i\right\}\right] = \mathbb{E}\left[1 - \mathbb{1}\left\{\overline{\cup_{i=1}^{n} A_i}\right\}\right] = \mathbb{E}\left[1 - \mathbb{1}\left\{\cap_{i=1}^{n} A_i^c\right\}\right] = \mathbb{E}\left[1 - \prod_{i=1}^{n} \mathbb{1}\left\{A_i^c\right\}\right] = \mathbb{E}\left[1 - \prod_{i=1}^{n}(1 - \mathbb{1}\{A_i\})\right].$$

We next to evaluate $\prod_{i=1}^{n}(1 - \mathbb{1}\{A_i\})$. It follows that

$$\begin{aligned}
\prod_{i=1}^{n}(1 - \mathbb{1}\{A_i\}) &= (1 - \mathbb{1}\{A_1\} - \mathbb{1}\{A_2\} + \mathbb{1}\{A_1 \cap A_2\}) \prod_{i=3}^{n}(1 - \mathbb{1}\{A_i\}) \\
&= (1 - \mathbb{1}\{A_1\} - \mathbb{1}\{A_2\} - \mathbb{1}\{A_3\} + \mathbb{1}\{A_1 \cap A_2\} + \mathbb{1}\{A_2 \cap A_3\} + \mathbb{1}\{A_1 \cap A_3\} \\
&\qquad - \mathbb{1}\{A_1 \cap A_2 \cap A_3\}) \prod_{i=4}^{n}(1 - \mathbb{1}\{A_i\}) \\
&= \quad \vdots \\
\prod_{i=1}^{n}(1 - \mathbb{1}\{A_i\}) &= 1 - \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\leq i_1<i_2<\cdots<i_k\leq n} \mathbb{1}\left\{A_{i_1} \cap \ldots \cap A_{i_k}\right\}.
\end{aligned}$$

Plugging our result back into our original expression results in

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \mathbb{E}\left[1 - \prod_{i=1}^{n}(1 - \mathbb{1}\{A_i\})\right] = \mathbb{E}\left[\sum_{k=1}^{n}(-1)^{k-1} \sum_{1\leq i_1<i_2<\cdots<i_k\leq n} \mathbb{1}\left\{A_{i_1} \cap \ldots \cap A_{i_k}\right\}\right]$$

and by linearity of expectation, we have

$$\mathbb{P}(\cup_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \Sigma_1 - \Sigma_2 + \Sigma_3 - \Sigma_4 + \cdots + (-1)^{n-1}\Sigma_n$$

where

$$\Sigma_k := \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}(A_{i_1} A_{i_2} \cdots A_{i_k}).$$

8. (**Hypergeometric and exchangeability**) We have an urn with R red balls and $N - R$ white balls, where $0 < R < N$. We draw n balls in sequence from the urn without replacement. Let R_i denote the proposition that the i^{th} draw results in a red ball.

   (a) Calculate $\mathbb{P}(R_i)$ for each $i = 1, \ldots, n$.

   (b) Show that $\mathbb{P}(R_j \mid R_k) = \mathbb{P}(R_k \mid R_j)$ for every $1 \le j, k \le n$.

   (c) Calculate $\mathbb{P}(R_k \mid \bigcup_{i=k+1}^{n} R_i)$ for a fixed $1 \le k < n$.

   (d) Let X be the random variable representing the minimum number of draws required to get at least one red ball. Calculate $E[X]$, the expected value of X. (Hint: Use exchangeability to simplify the calculation.)

   (e) Suppose that instead of only two colors, the urn has balls of k different colors: N_1 of color 1, N_2 of color 2, $\ldots$, N_k of color k. Let $N = N_1 + \cdots + N_k$. Argue that the probability of drawing r_1 balls of color 1, r_2 balls of color 2, $\ldots$, r_k balls of color k in $n = r_1 + \cdots + r_k$ draws without replacement is given by

$$\frac{\binom{N_1}{r_1} \cdots \binom{N_k}{r_k}}{\binom{N}{n}}.$$

(a) We can think of this as a sequence of random variables, i.e.,

$$\omega = \mathbb{1}\{1\}, \mathbb{1}\{2\}, \ldots, \mathbb{1}\{n\} \qquad \text{where } \mathbb{1}\{i\} = \begin{cases} 1 & \text{when } i \text{ is a } R \text{ ball} \\ 0 & \text{otherwise} \end{cases}$$

We are interested in whether $\mathbb{1}\{i\}$ is 1. Then by exchangeability, we have that probability of the i^{th} draw being a red ball does not depend on i. Thus we have

$$\boxed{\mathbb{P}(R_i) = \mathbb{P}(R_1) = \frac{R}{N} \quad \text{for all } i = 1, \ldots, n.}$$

(b) This result follows directly from Bayes' rule and part (a),

$$\mathbb{P}(R_j \mid R_k) = \mathbb{P}(R_k \mid R_j) \overset{\text{(Bayes')}}{\Longrightarrow} \frac{\mathbb{P}(R_j \cap R_k)}{\mathbb{P}(R_k)} = \frac{\mathbb{P}(R_k \cap R_j)}{\mathbb{P}(R_j)}$$

where $\mathbb{P}(R_j \cap R_k) = \mathbb{P}(R_k \cap R_j)$ holds by commutative of the intersection, and $\mathbb{P}(R_j) = \mathbb{P}(R_k)$ from part (a). Thus we have

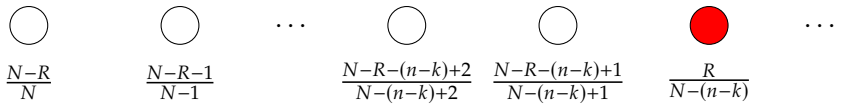
$$\boxed{\mathbb{P}(R_j \mid R_k) = \mathbb{P}(R_k \mid R_j) \quad \text{for every } 1 \le j, k \le n.}$$

(c) We are asked to compute $\mathbb{P}(R_k \mid \bigcup_{i=k+1}^{n} R_i)$ for a fixed $1 \le k < n$. This is the probability that the k^{th} draw results in a red ball, given that at least one red ball is drawn in the remaining draws $(k+1)^{th}, \ldots, n^{th}$. To begin, we write

$$\mathbb{P}\left(R_k \mid \bigcup_{i=k+1}^{n} R_i\right) = \frac{\mathbb{P}(R_k \cap \bigcup_{i=k+1}^{n} R_i)}{\mathbb{P}(\bigcup_{i=k+1}^{n} R_i)}.$$

Starting with the denominator, we note that by de Morgan's law, we have

$$\mathbb{P}\left(\bigcup_{i=k+1}^{n} R\right) = 1 - \mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c\right).$$



By exchangeability, we have that $\mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c\right)$ is equal to the probability of the first $n-k$ balls drawn being white. From the diagram, we have

$$\mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c\right) = \frac{(N-R)!}{(N-R-(n-k))!}\frac{(N-(n-k))!}{N!} = \frac{\binom{N-R}{n-k}}{\binom{N}{n-k}}.$$

Now we focus on the numerator. We have

$$\mathbb{P}\left(R_k \cap \bigcup_{i=k+1}^{n} R_i\right) = \mathbb{P}\left(\bigcup_{i=k+1}^{n} R_i \mid R_k\right)\mathbb{P}(R_k) = \left(1 - \mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c \mid R_k\right)\right)\mathbb{P}(R_k) = \left(1 - \frac{\mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c \cap R_k\right)}{\mathbb{P}(R_k)}\right)\mathbb{P}(R_k)$$

$$\mathbb{P}\left(R_k \cap \bigcup_{i=k+1}^{n} R_i\right) = \left(\mathbb{P}(R_k) - \mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c \cap R_k\right)\right).$$

Referring to our diagram, it clear that

$$\mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c \cap R_k\right) = \frac{(N-R)!}{(N-R-(n-k))!}\frac{(N-(n-k))!}{N!}\frac{R}{N-(n-k)} = \frac{R}{N}\frac{(N-R)!}{(N-R-(n-k))!}\frac{(N-1-(n-k))!}{(N-1)!}$$

$$\mathbb{P}\left(\bigcap_{i=k+1}^{n} R_i^c \cap R_k\right) = \frac{R}{N}\frac{\binom{N-R}{n-k}}{\binom{N-1}{n-k}}$$

Put our results together, we have

$$\boxed{\mathbb{P}\left(R_k \mid \bigcup_{i=k+1}^{n} R_i\right) = \frac{\frac{R}{N}\left(1 - \frac{\binom{N-R}{n-k}}{\binom{N-1}{n-k}}\right)}{1 - \frac{\binom{N-R}{n-k}}{\binom{N}{n-k}}}}$$

(d) Let X be the random variable representing the minimum number of draws required to get at least one red ball. We are asked to calculate the expected value $E[X]$. Let us define an indicator variable that the j-th white ball will be drawn before any of the red balls, i.e.,

$$\mathbb{1}\{j\}(\omega) = \begin{cases} 1 & \text{when the } j\text{-th white ball will be drawn before any of the red balls} \\ 0 & \text{otherwise} \end{cases}$$

Then we can write X as

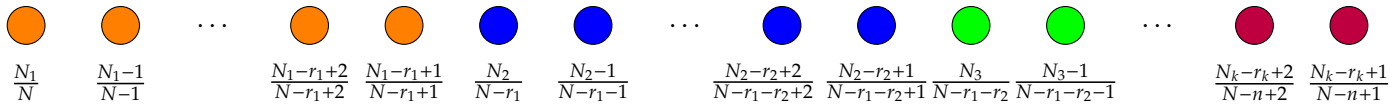
$$X = 1 + \sum_{j=1}^{W} \mathbb{1}\{j\}$$

By exchangeability, we have that the probability that the j-th white ball is drawn before any of the red balls as

$$\mathbb{P}(\mathbb{1}\{j\} = 1) = \mathbb{E}[\mathbb{1}\{j\}] = \frac{1}{R+1}.$$

It then follows that

$$\mathbb{E}[X] = 1 + \sum_{j=1}^{W} \mathbb{E}[\mathbb{1}\{j\}] = 1 + \frac{W}{R+1} \implies \boxed{\mathbb{E}[X] = \frac{N+1}{R+1}.}$$



(e) Consider the diagram below

It is clear we have

$$\mathbb{P}(X=\omega)=\frac{\left(\frac{N_1}{(N_1-r_1)!}\right)\left(\frac{N_2}{(N_2-r_2)!}\right)\cdots\left(\frac{N_k}{(N_k-r_k)!}\right)}{\frac{N!}{(N-n)!}},$$

and this can be arranged in $\frac{n!}{r_1!r_2!\ldots r_k!}$ ways leading to

$$\boxed{\mathbb{P}(r_1 \text{ balls of color 1}, r_2 \text{ balls of color 2}, \ldots, r_k \text{ balls of color } k)=\frac{\binom{N_1}{r_1}\cdots\binom{N_k}{r_k}}{\binom{N}{n}}}$$

Homework # 2: Concentration Inequalities & Moment Generating Functions

Reece D. Huff

Problems (Solutions)

1. (**Binomial tail bounds**) Let S_n have the Binomial(n,p) distribution of the number of successes in n independent Bernoulli(p) trials. Use a suitable computational environment to evaluate the right tail probabilities

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right)$$

for $n = 100$ and $p_i = \frac{i}{10}$ for $i = 1, 2, \ldots, 9$, and $\epsilon = \frac{1}{10}$, together with various approximations and upper bounds as indicated. In each case:

- Give an exact mathematical formula for the function of i you are computing;
- Indicate suitable code for evaluating the formula in your preferred environment and **attach the code at the end of the homework**;
  Code is written in `Python` and is attached at the end of the document.
- Give the numerical values correct to two significant decimal places.

(a) The exact probabilities.
We begin by plugging in our values of n, p_i, and ϵ, i.e.,

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right) = \mathbb{P}\left(\frac{S_{100}}{100} \geq \frac{i}{10} + \frac{1}{10}\right) = \mathbb{P}\left(S_{100} \geq 10(i+1)\right).$$

From the definition of the Binomial distribution, we have

$$\boxed{\mathbb{P}\left(S_{100} \geq 10(i+1)\right) = \sum_{k=10(i+1)}^{100} \binom{100}{k} p_i^k (1-p_i)^{100-k} = 1 - \sum_{k=0}^{10(i+1)-1} \binom{100}{k} p_i^k (1-p_i)^{100-k}.}$$

p_i	$p_1 = 0.1$	$p_2 = 0.2$	$p_3 = 0.3$	$p_4 = 0.4$	$p_5 = 0.5$	$p_6 = 0.6$	$p_7 = 0.7$	$p_8 = 0.8$	$p_9 = 0.9$
$\mathbb{P}(S_n - np_i \geq n\epsilon) =$	0.002	0.011	0.021	0.027	0.028	0.025	0.016	0.0057	2.7e-05

(b) Markov's upper bounds for these probabilities.
Recall Markov's inequality,

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c} \quad \text{for any arbitrary r.v. } X : \Omega \to \mathbb{R} \text{ and constant } c > 0.$$

Additionally, recall the expectation of the Binomial distribution is $\mathbb{E}[S_n] = np$, which is simply $100(i/10) = 10i$ for all $i \in \{1, \ldots, 9\}$. Thus, we have

$$\boxed{\mathbb{P}\left(S_n - np_i \geq n\epsilon\right) \leq \frac{\mathbb{E}[S_{100}]}{10(i+1)} = \frac{i}{i+1} \quad \text{for all } i \in \{1, \ldots, 9\}}$$

p_i	$p_1 = 0.1$	$p_2 = 0.2$	$p_3 = 0.3$	$p_4 = 0.4$	$p_5 = 0.5$	$p_6 = 0.6$	$p_7 = 0.7$	$p_8 = 0.8$	$p_9 = 0.9$
$\mathbb{P}\left(S_n - np_i \geq n\epsilon\right) \leq$	0.50	0.67	0.75	0.8	0.83	0.86	0.88	0.89	0.90

(c) Chebyshev's upper bounds for these probabilities (which can be halved for $i = 5$ only: explain why).

Recall Chebyshev's inequality,

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2} \quad \text{for an r.v. } X : \Omega \to \mathbb{R} \text{ with } \mathbb{E}[X] = \mu < \infty \text{ and constant } c > 0.$$

Additionally, recall the variance of the Binomial distribution is $\text{Var}(S_n) = np(1-p)$, which is simply $100(i/10)(1 - i/10) = 100(i/10 - i^2/100) = 10i - i^2$. Thus, we have

$$\boxed{\mathbb{P}(|S_{100} - \mathbb{E}[S_{100}]| \geq n\epsilon) \leq \frac{10i - i^2}{(n\epsilon)^2} = \frac{10i - i^2}{100} \quad \text{for all } i \in \{1, \ldots, 9\} \ / 5}$$

Note that for $i = 5$, we can halve the probability by symmetry, i.e.,

$$\mathbb{P}(S_{100} - \mathbb{E}[S_{100}] \geq n\epsilon) = \mathbb{P}(S_{100} - \mathbb{E}[S_{100}] \leq -n\epsilon) = \frac{1}{2}\,\mathbb{P}(S_{100} - \mathbb{E}[S_{100}] \geq n\epsilon)$$

$$\implies \boxed{\mathbb{P}(|S_{100} - \mathbb{E}[S_{100}]| \geq n\epsilon) \leq \frac{10i - i^2}{200} \quad \text{for } i = 5}$$

p_i	$p_1 = 0.1$	$p_2 = 0.2$	$p_3 = 0.3$	$p_4 = 0.4$	$p_5 = 0.5$	$p_6 = 0.6$	$p_7 = 0.7$	$p_8 = 0.8$	$p_9 = 0.9$
$\mathbb{P}(\lvert S_n - np_i\rvert \geq n\epsilon) \leq$	0.090	0.16	0.21	0.24	0.25	0.24	0.21	0.16	0.090

(d) Hoeffding's upper bounds.

Recall Hoeffding's inequality: Let $X_1, \ldots, X_n$ be independent r.v.'s with $\mathbb{E}[X_i] = \mu_i < \infty$ and $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$ for constants $a_i, b_i \in \mathbb{R}$. Let $S_n = X_1 + \cdots + X_n$. Then,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \text{for any } \epsilon > 0.$$

Note that in our case, $X_i \in \{0, 1\}$ for all $i \in \{1, \ldots, n\}$. Then we have $a_i = 0$ and $b_i = 1$ for all $i \in \{1, \ldots, n\}$. Then our Hoeffding bounds are

$$\boxed{\mathbb{P}(S_n - \mathbb{E}[S_n] \geq n\epsilon) \leq \exp\left(\frac{-2n^2\epsilon^2}{n}\right) = \exp\left(-2n\epsilon^2\right) = \exp(-2)}$$

p_i	$p_1 = 0.1$	$p_2 = 0.2$	$p_3 = 0.3$	$p_4 = 0.4$	$p_5 = 0.5$	$p_6 = 0.6$	$p_7 = 0.7$	$p_8 = 0.8$	$p_9 = 0.9$
$\mathbb{P}(S_n - np_i \geq n\epsilon) \leq$	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

(e) Chernoff's upper bounds.

Recall Chernoff's inequality,

$$\mathbb{P}(X \geq c) \leq \min_{t>0}\left\{\frac{M_X(t)}{e^{tc}}\right\} \quad \text{for any } t > 0 \text{ and } c \in \mathbb{R}.$$

Recall that we derived in called the Chernoff bound of a Binomial random variable as

$$\mathbb{P}(X \geq an) \leq \left(\frac{1-p}{1-a}\right)^{(1-a)n}\left(\frac{p}{a}\right)^{an} \quad \text{for any } t > 0 \text{ and any constant } a \in \mathbb{R}.$$

We set $a = (p_i + \epsilon)$ to arrive at

$$\mathbb{P}(S_n \geq (p_i + \epsilon)n) = \mathbb{P}(S_n - np_i \geq n\epsilon) \leq \left(\frac{1-p_i}{1-p_i-\epsilon}\right)^{(1-p_i-\epsilon)n}\left(\frac{p_i}{p_i+\epsilon}\right)^{(p_i+\epsilon)/n}$$

$$\implies \boxed{\mathbb{P}(S_{100} \geq 10(i+1)) \leq \left(\frac{10-i}{9-i}\right)^{90-10i}\left(\frac{i}{i+1}\right)^{10(i+1)}}$$

p_i	$p_1 = 0.1$	$p_2 = 0.2$	$p_3 = 0.3$	$p_4 = 0.4$	$p_5 = 0.5$	$p_6 = 0.6$	$p_7 = 0.7$	$p_8 = 0.8$	$p_9 = 0.9$
$\mathbb{P}(S_n - np_i \geq n\epsilon) \leq$	0.012	0.06	0.10	0.13	0.13	0.12	0.076	0.026	2.7e-05

2. (**LLN**) Suppose that $X_1, X_2, \ldots$ form an i.i.d. sequence of random variables with $\mathbb{E}[X_i] = \mu < \infty$ and $\text{Var}[X_i] = \sigma^2 < \infty$. Evaluate

$$\lim_{n\to\infty} \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} (X_i - X_j)^2.$$

We apply the law of large numbers to the random varible $Z := \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} (X_i - X_j)^2$ to arrive at

$$\lim_{n\to\infty} \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} (X_i - X_j)^2 = \mathbb{E}\left[\frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} (X_i - X_j)^2\right] = \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} \mathbb{E}\left[X_i^2 - 2X_iX_j + X_j^2\right]$$

$$= \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} \sigma^2 + \mu^2 - 2\mu^2 + \sigma^2 + \mu^2 = \frac{1}{\binom{n}{2}} \frac{n(n+1)}{2} 2\sigma^2$$

$$\boxed{\lim_{n\to\infty} \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} (X_i - X_j)^2 = 2\sigma^2.}$$

3. (**Chebyshev & CLT**) Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with mean zero and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Determine the limits below, with precise justifications.

   (a) $\lim_{n\to\infty} \mathbb{P}(S_n \geq 0.01n)$.

   (b) $\lim_{n\to\infty} \mathbb{P}(S_n \geq 0)$.

   (c) $\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n)$.

To begin, we note that we can rewrite the expression $\lim_{n\to\infty} \mathbb{P}(S_n \geq x)$ as

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq x) = 1 - \lim_{n\to\infty} \mathbb{P}(S_n \leq x).$$

This follows from our discrete r.v. S_n becoming continuous as $n \to \infty$, i.e.,

$$\mathbb{P}(S_n \geq x) + \mathbb{P}(S_n \leq x) - \mathbb{P}(S_n = x) = 1 \implies \lim_{n\to\infty} \{\mathbb{P}(S_n \geq x) + \mathbb{P}(S_n \leq x) - \mathbb{P}(S_n = x)\} = 1$$

$$\implies \lim_{n\to\infty} \mathbb{P}(S_n \geq x) + \lim_{n\to\infty} \mathbb{P}(S_n \leq x) - \cancelto{0}{\lim_{n\to\infty} \mathbb{P}(S_n = x)} = 1$$

$$\implies \lim_{n\to\infty} \mathbb{P}(S_n \geq x) = 1 - \lim_{n\to\infty} \mathbb{P}(S_n \leq x)$$

Next, we recall the Central Limit Theorem.

Theorem 1 (Central Limit Theorem)**.** Let $X_i, \ldots, X_n$ be a sequence of i.i.d. r.v.'s with finite mean μ and finite variance σ^2. Let $S_n := X_i + \cdots + X_n$. Then,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right) \leq x\right] = \Phi(x), \quad \text{for all } x \in \mathbb{R} \text{ where } \quad \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \quad \text{(the c.d.f. of } \mathcal{N}(0,1)).$$

Then for $X_1, X_2, X_3, \ldots$ i.i.d. random variables with mean zero and finite variance σ^2, and $S_n = X_1 + \cdots + X_n$, we have

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{S_n}{\sigma\sqrt{n}} \leq x\right] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

It then follows for part (a), (b), and (c):

(a) $\lim_{n\to\infty} \mathbb{P}(S_n \geq 0.01n)$.

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq 0.01n) = 1 - \lim_{n\to\infty} \mathbb{P}(S_n \leq 0.01n) = 1 - \lim_{n\to\infty} \mathbb{P}\left[\frac{S_n}{\sigma\sqrt{n}} \leq \frac{0.01\sqrt{n}}{\sigma}\right] = 1 - \lim_{n\to\infty} \int_{-\infty}^{\frac{0.01\sqrt{n}}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq 0.01n) = 1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt = 1 - 1$$

$$\boxed{\lim_{n\to\infty} \mathbb{P}(S_n \geq 0.01n) = 0.}$$

(b) $\lim_{n\to\infty} \mathbb{P}(S_n \geq 0)$.

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq 0) = 1 - \lim_{n\to\infty} \mathbb{P}(S_n \leq 0) = 1 - \lim_{n\to\infty} \mathbb{P}\left[\frac{S_n}{\sigma\sqrt{n}} \leq 0\right] = 1 - \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq 0) = 1 - \frac{1}{2}$$

$$\boxed{\lim_{n\to\infty} \mathbb{P}(S_n \geq 0) = \frac{1}{2}.}$$

(c) $\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n)$.

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n) = 1 - \lim_{n\to\infty} \mathbb{P}(S_n \leq -0.01n) = 1 - \lim_{n\to\infty} \mathbb{P}\left[\frac{S_n}{\sigma\sqrt{n}} \leq \frac{-0.01\sqrt{n}}{\sigma}\right]$$

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n) = 1 - \lim_{n\to\infty} \int_{-\infty}^{-\frac{0.01\sqrt{n}}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

$$\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n) = 1 - \int_{-\infty}^{-\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt = 1 - 0$$

$$\boxed{\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n) = 1.}$$

4. (**Convolution & MGF**) The Laplace distribution has density $f_Z(z) = \frac{\lambda}{2}\exp(-\lambda|z|)$ and MGF $M_Z(t) = \frac{\lambda^2}{\lambda^2 - t^2}$, where $\lambda > 0$. Let $X, Y \overset{iid}{\sim} \text{Exp}(\lambda)$. Prove that $Z = X - Y$ follows a Laplace distribution by using:

(a) Moment generating functions.

Recall the MGF of the exponential distribution,

$$M_X(t) = \mathbb{E}\left[\exp(tX)\right] = \frac{\lambda}{\lambda - t} \quad \text{for a r.v. } X \sim \text{Exp}(X) \text{ and } \lambda > 0, t \in \mathbb{R}.$$

Now we simply apply the MGF to the random variable $Z = X - Y$, i.e.,

$$\begin{aligned} M_Z(t) &= \mathbb{E}\left[\exp(tZ)\right] = \mathbb{E}\left[\exp(t(X-Y))\right] = \mathbb{E}\left[\exp(tX)\exp(-tY)\right] \\ &= \mathbb{E}\left[\exp(tX)\right]\mathbb{E}\left[\exp(-tY)\right] && \text{(by independence of } X \text{ and } Y\text{)} \\ &= \left(\frac{\lambda}{\lambda - t}\right)\left(\frac{\lambda}{\lambda + t}\right) && \text{(MGF of } X, Y \overset{iid}{\sim} \text{Exp}(\lambda)\text{)} \end{aligned}$$

$$\boxed{M_Z(t) = \frac{\lambda^2}{\lambda^2 - t^2}}$$

(b) The convolution formula.

Recall the density of the exponential distribution,

$$f_X(x) = \begin{cases} \lambda\exp(-\lambda x) & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad \text{for a r.v. } X \sim \text{Exp}(X) \text{ and } \lambda > 0.$$

Additionally, recall the convolution formula,

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx \quad \Leftrightarrow \quad f_{X-Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(x-z)dx \quad \text{for continuous r.v.'s } X, Y.$$

Using these formulae, we have that for a r.v. $Z = X - Y$,

$$f_Z(z) = f_{X-Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(x-z)dx = \int_{-\infty}^{\infty} \lambda \exp(-\lambda x) \cdot \lambda \exp(-\lambda(x-z))dx$$

Note from our integral is non-zero only when $x - z \geq 0$. Therefore, we consider two cases: when $z \geq 0$ and when $z < 0$. For each case we will changes the bounds of integration. As such, we solve the integral above for general bounds of integration, a and b:

$$f_Z(z) = \int_a^b \lambda \exp(-\lambda x) \cdot \lambda \exp(-\lambda(x-z))dx = \lambda^2 \int_a^b \exp(-\lambda x)\exp(-\lambda x)\exp(\lambda z)dx = \lambda^2 \exp(\lambda z) \int_a^b \exp(-2\lambda x)dx$$

$$f_Z(z) = \lambda^2 \exp(\lambda z)\left[\frac{-1}{2\lambda}\exp(-2\lambda x)\right]_a^b = \frac{-\lambda}{2}\exp(\lambda z)\left[\exp(-2\lambda x)\right]_a^b$$

(i) $f_Z(z)$ when $z \geq 0$:
When $z \geq 0$, we have that density is nonzero for $x \geq z$, so we evaluate our integral from $x = z$ to $x = \infty$, i.e.,

$$f_Z(z) = \frac{-\lambda}{2}\exp(\lambda z)\left[\exp(-2\lambda x)\right]_z^{\infty} = \frac{-\lambda}{2}\exp(\lambda z)\left[0 - \exp(-2\lambda z)\right] = \frac{\lambda}{2}\exp(\lambda z)\exp(-2\lambda z) = \frac{\lambda}{2}\exp(-\lambda z)$$

(ii) $f_Z(z)$ when $z < 0$:
When $z < 0$, we have that density is nonzero for $x \geq 0$ as $x - z \geq 0$ for $x \geq 0$, so we evaluate our integral from $x = 0$ to $x = \infty$, i.e.,

$$f_Z(z) = \frac{-\lambda}{2}\exp(\lambda z)\left[\exp(-2\lambda x)\right]_0^{\infty} = \frac{-\lambda}{2}\exp(\lambda z)\left[0 - \exp(-2\lambda 0)\right] = \frac{\lambda}{2}\exp(\lambda z)$$

Taken together, we have that the density of $Z = X - Y$ is

$$f_Z(z) = \begin{cases} \frac{\lambda}{2}\exp(-\lambda z) & \text{if } z \geq 0, \\ \frac{\lambda}{2}\exp(\lambda z) & \text{otherwise} \end{cases} \quad \Leftrightarrow \quad \boxed{f_Z(z) = \frac{\lambda}{2}\exp(-\lambda|z|)}$$

5. (**Moments & MGF**) Let X be a random variable with p.d.f. given by

$$f_X(x) = \begin{cases} \frac{2}{9}, & 0 \leq x \leq 1, \\ \frac{4-|4-2x|}{9}, & 1 < x \leq 4, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Verify that this is actually a p.d.f.

We verify that $f_X(x)$ is actually a p.d.f. by checking $\int_{-\infty}^{\infty} f_X(x)dx = 1$, i.e.,

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 \frac{2}{9}dx + \int_1^4 \frac{4-|4-2x|}{9}dx = \frac{2}{9} + \frac{4}{9}\int_1^4 1dx - \frac{2}{9}\int_1^4 |2-x|dx$$

$$\int_{-\infty}^{\infty} f_X(x)dx = \frac{2}{9} + \frac{12}{9} - \frac{2}{9}\left[\frac{-|2-x|(2-x)}{2}\right]_1^4 = \frac{14}{9} - \frac{2}{9}\left[\frac{-|2-4|(2-4)}{2} + \frac{|2-1|(2-1)}{2}\right] = \frac{14}{9} - \frac{2}{9}\left[\frac{5}{2}\right] = \frac{14}{9} - \frac{5}{9}$$

$$\boxed{\int_{-\infty}^{\infty} f_X(x)dx = 1 \quad \checkmark}$$

(b) Find the moment generating function of X.

The MGF for X is

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx = \int_0^1 e^{tx}\frac{2}{9}dx + \int_1^4 e^{tx}\frac{4-|4-2x|}{9}dx$$

$$M_X(t) = \frac{2}{9}\int_0^1 e^{tx}dx + \frac{4}{9}\int_1^4 e^{tx}dx - \frac{2}{9}\int_1^4 e^{tx}|x-2|dx$$
$$M_X(t) = \frac{2}{9t}\left[e^{tx}\right]_0^1 + \frac{4}{9t}\left[e^{tx}\right]_1^4 - \frac{2}{9}\int_1^4 e^{tx}|x-2|dx$$
$$M_X(t) = \frac{2}{9t}\left[e^t - 1\right] + \frac{4}{9t}\left[e^{4t} - e^t\right] + \frac{2}{9}\int_1^2 e^{tx}(x-2)dx - \frac{2}{9}\int_2^4 e^{tx}(x-2)dx$$

By integration by parts with $u = x-2$ and $dv = e^{tx}dx$, we have

$$\int_a^b u\,dv = [uv]_a^b - \int_a^b v\,du \quad \Longrightarrow \quad \int_a^b e^{tx}(x-2)dx = \frac{1}{t}\left[e^{tx}(x-2)\right]_a^b - \frac{1}{t}\int_a^b e^{tx}dx = \frac{1}{t}\left[e^{tx}(x-2)\right]_a^b - \frac{1}{t^2}\left[e^{tx}\right]_a^b$$

Then we have

$$M_X(t) = \frac{2}{9t}\left[e^t - 1\right] + \frac{4}{9t}\left[e^{4t} - e^t\right] + \frac{2}{9}\frac{1}{t}\left[e^{tx}(x-2)\right]_1^2 - \frac{2}{9}\frac{1}{t^2}\left[e^{tx}\right]_1^2 - \frac{2}{9}\frac{1}{t}\left[e^{tx}(x-2)\right]_2^4 + \frac{2}{9}\frac{1}{t^2}\left[e^{tx}\right]_2^4$$
$$M_X(t) = \frac{2e^t - 2}{9t} + \frac{4e^{4t} - 4e^t}{9t} + \frac{2}{9t}\left[0 - e^t(1-2)\right] - \frac{2}{9t^2}\left[e^{2t} - e^t\right] - \frac{2}{9t}\left[e^{4t}(4-2) - 0\right] + \frac{2}{9t^2}\left[e^{4t} - e^{2t}\right]$$
$$M_X(t) = \frac{2e^t - 2}{9t} + \frac{\cancel{4e^{4t}} - 4e^t}{9t} + \frac{2e^t}{9t} + \frac{2e^t - 2e^{2t}}{9t^2} - \frac{\cancel{4e^{4t}}}{9t} + \frac{2e^{4t} - 2e^{2t}}{9t^2}$$
$$M_X(t) = \frac{\cancel{2e^t} - 2 - \cancel{4e^t} + \cancel{2e^t}}{9t} + \frac{2e^t - 4e^{2t} + 2e^{4t}}{9t^2}$$
$$M_X(t) = \frac{2e^t - 4e^{2t} + 2e^{4t} - 2t}{9t^2}$$
$$\boxed{M_X(t) = \frac{2(e^t - 2e^{2t} + e^{4t} - t)}{9t^2}.}$$

(c) Find $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

The expectation of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 x\frac{2}{9}dx + \int_1^4 x\frac{4-|4-2x|}{9}dx = \frac{2}{9}\int_0^1 x\,dx + \frac{4}{9}\int_1^4 x\,dx - \frac{2}{9}\int_1^4 x|x-2|dx$$
$$\mathbb{E}[X] = \frac{2}{9}\frac{1}{2}\left[x^2\right]_0^1 + \frac{4}{9}\frac{1}{2}\left[x^2\right]_1^4 + \frac{2}{9}\int_1^2 x(x-2)dx - \frac{2}{9}\int_2^4 x(x-2)dx$$
$$\mathbb{E}[X] = \frac{1}{9} + \frac{2}{9}[16-1] + \frac{2}{9}\int_1^2 x^2 - 2x\,dx - \frac{2}{9}\int_2^4 x^2 - 2x\,dx$$
$$\mathbb{E}[X] = \frac{1}{9} + \frac{30}{9} + \frac{2}{9}\left[\frac{x^3}{3} - x^2\right]_1^2 - \frac{2}{9}\left[\frac{x^3}{3} - x^2\right]_2^4$$
$$\mathbb{E}[X] = \frac{31}{9} + \frac{2}{9}\left[\frac{8}{3} - 4 - \frac{1}{3} + 1\right] - \frac{2}{9}\left[\frac{64}{3} - 16 - \frac{8}{3} + 4\right]$$
$$\mathbb{E}[X] = \frac{31}{9} + \frac{2}{9}\left[\frac{8}{3} - \frac{12}{3} - \frac{1}{3} + \frac{3}{3}\right] - \frac{2}{9}\left[\frac{64}{3} - \frac{48}{3} - \frac{8}{3} + \frac{12}{3}\right]$$
$$\mathbb{E}[X] = \frac{31}{9} + \frac{2}{9}\left[-\frac{2}{3}\right] - \frac{2}{9}\left[\frac{20}{3}\right] = \frac{93}{27} - \frac{4}{27} - \frac{40}{27}$$
$$\boxed{\mathbb{E}[X] = \frac{49}{27}}$$

In order to calculate the variance, we first calculate $\mathbb{E}[X^2]$ as $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Then we have

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^1 x^2\frac{2}{9}dx + \int_1^4 x^2\frac{4-|4-2x|}{9}dx = \frac{2}{9}\int_0^1 x^2dx + \frac{4}{9}\int_1^4 x^2dx - \frac{2}{9}\int_1^4 x^2|x-2|dx$$
$$\mathbb{E}[X^2] = \frac{2}{9}\frac{1}{3}\left[x^3\right]_0^1 + \frac{4}{9}\frac{1}{3}\left[x^3\right]_1^4 + \frac{2}{9}\int_1^2 x^2(x-2)dx - \frac{2}{9}\int_2^4 x^2(x-2)dx$$

$$\mathbb{E}[X^2] = \frac{2}{27} + \frac{4}{27}[64-1] + \frac{2}{9}\int_1^2 x^3 - 2x^2 dx - \frac{2}{9}\int_2^4 x^3 - 2x^2 dx$$

$$\mathbb{E}[X^2] = \frac{2}{27} + \frac{252}{27} + \frac{2}{9}\left[\frac{x^4}{4} - \frac{2x^3}{3}\right]_1^2 - \frac{2}{9}\left[\frac{x^4}{4} - \frac{2x^3}{3}\right]_2^4$$

$$\mathbb{E}[X^2] = \frac{254}{27} + \frac{2}{9}\left[\frac{16}{4} - \frac{16}{3} - \frac{1}{4} + \frac{2}{3}\right] - \frac{2}{9}\left[\frac{256}{4} - \frac{128}{3} - \frac{16}{4} + \frac{16}{3}\right]$$

$$\mathbb{E}[X^2] = \frac{254}{27} + \frac{2}{9}\left[\frac{48}{12} - \frac{64}{12} - \frac{3}{12} + \frac{8}{12}\right] - \frac{2}{9}\left[\frac{768}{12} - \frac{512}{12} - \frac{48}{12} + \frac{64}{12}\right]$$

$$\mathbb{E}[X^2] = \frac{254}{27} + \frac{2}{9}\left[-\frac{11}{12}\right] - \frac{2}{9}\left[\frac{272}{12}\right] = \frac{1016}{108} - \frac{22}{108} - \frac{544}{108} = \frac{450}{108} = \frac{25}{6}$$

Now that we have solved for $\mathbb{E}[X^2]$, we can solve for the variance of X, i.e.,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{25}{6} - \left(\frac{49}{27}\right)^2 = \frac{25}{6} - \frac{2401}{729} \quad \Longrightarrow \quad \boxed{\mathrm{Var}(X) = \frac{1273}{1458} \approx 0.873}$$

(d) Find a formula for the moments of X.

We define the k-th moment of X as

$$\mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k f_X(x)dx = \int_0^1 x^k \frac{2}{9}dx + \int_1^4 x^k \frac{4-|4-2x|}{9}dx = \frac{2}{9}\int_0^1 x^k dx + \frac{4}{9}\int_1^4 x^k dx - \frac{2}{9}\int_1^4 x^k|x-2|dx$$

$$\mathbb{E}[X^k] = \frac{2}{9}\left[\frac{x^{k+1}}{k+1}\right]_0^1 + \frac{4}{9}\left[\frac{x^{k+1}}{k+1}\right]_1^4 + \frac{2}{9}\int_1^2 x^k(x-2)dx - \frac{2}{9}\int_2^4 x^k(x-2)dx$$

$$\mathbb{E}[X^k] = \frac{2}{9(k+1)} + \frac{4}{9(k+1)}\left[4^{k+1}-1\right] + \frac{2}{9}\int_1^2 x^k(x-2)dx - \frac{2}{9}\int_2^4 x^k(x-2)dx$$

Again, we apply integration by parts with $u = x-2$ and $dv = x^k dx$, i.e.,

$$\int_a^b x^k(x-2)dx = \left[(x-2)\frac{x^{k+1}}{k+1}\right]_a^b - \int_a^b \frac{x^{k+1}}{k+1}dx = \left[(x-2)\frac{x^{k+1}}{k+1}\right]_a^b - \left[\frac{x^{k+2}}{(k+1)(k+2)}\right]_a^b.$$

Then we have

$$\mathbb{E}[X^k] = \frac{2+4(4^{k+1}-1)}{9(k+1)} + \frac{2}{9}\left(\left[(x-2)\frac{x^{k+1}}{k+1}\right]_1^2 - \left[\frac{x^{k+2}}{(k+1)(k+2)}\right]_1^2\right) - \frac{2}{9}\left(\left[(x-2)\frac{x^{k+1}}{k+1}\right]_2^4 - \left[\frac{x^{k+2}}{(k+1)(k+2)}\right]_2^4\right)$$

$$\mathbb{E}[X^k] = \frac{2+4^{k+2}-4}{9(k+1)} + \frac{2}{9}\left(\left[0-(1-2)\frac{1}{k+1}\right] - \left[\frac{2^{k+2}}{(k+1)(k+2)} - \frac{1}{(k+1)(k+2)}\right]\right)$$
$$- \frac{2}{9}\left(\left[(4-2)\frac{4^{k+1}}{k+1} - 0\right] - \left[\frac{4^{k+2}}{(k+1)(k+2)} - \frac{2^{k+2}}{(k+1)(k+2)}\right]\right)$$

$$\mathbb{E}[X^k] = \frac{4^{k+2}-2}{9(k+1)} + \frac{2}{9}\left(\frac{1}{k+1} + \frac{1-2^{k+2}}{(k+1)(k+2)}\right) - \frac{2}{9}\left(\frac{2\cdot 4^{k+1}}{k+1} + \frac{2^{k+2}-4^{k+2}}{(k+1)(k+2)}\right)$$

$$\mathbb{E}[X^k] = \frac{4^{k+2}-2}{9(k+1)} + \frac{2}{9(k+1)} + \frac{2(1-2^{k+2})}{9(k+1)(k+2)} - \frac{2\cdot 2\cdot 4^{k+1}}{9(k+1)} - \frac{2(2^{k+2}-4^{k+2})}{9(k+1)(k+2)}$$

$$\mathbb{E}[X^k] = \frac{4^{k+2}-2}{9(k+1)} + \frac{2}{9(k+1)} - \frac{4^{k+2}}{9(k+1)} + \frac{2-2^{k+3}}{9(k+1)(k+2)} - \frac{2^{k+3}-2\cdot 4^{k+2}}{9(k+1)(k+2)}$$

$$\mathbb{E}[X^k] = \frac{\cancel{4^{k+2}} - \cancel{2} + \cancel{2} - \cancel{4^{k+2}}}{9(k+1)} + \frac{2-2^{k+3}-2^{k+3}+2\cdot 4^{k+2}}{9(k+1)(k+2)}$$

$$\boxed{\mathbb{E}[X^k] = \frac{2(1-2^{k+3}+4^{k+2})}{9(k+1)(k+2)}}$$

We can verify that it holds for the first and second moments:

$$\mathbb{E}[X] = \frac{2(1-2^4+4^3)}{9(2)(3)} = \frac{2(1-2^4+4^3)}{54} = \frac{98}{54} = \frac{49}{27} \quad \checkmark$$

$$\mathbb{E}[X^2] = \frac{2(1-2^5+4^4)}{9(4)(5)} = \frac{450}{108} = \frac{25}{6} \quad \checkmark$$

6. **(Distribution of sums using MGFs)** Let $S_n := X_1 + \cdots + X_n$ for independent $X_1, \ldots, X_n$. Use MGFs to find the distribution of S_n:

   (a) For X_i with Normal (μ_i, σ_i^2) distribution.

   Recall the MGF of the Normal distribution with mean μ_i and variance σ_i^2,

   $$M_{X_i}(t) = \exp\left(t\mu_i + \frac{1}{2}\sigma_i^2 t^2\right) \quad \text{for a r.v. } X_i \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

   Then the MGF of $S_n = X_1 + \cdots + X_n$ for for independent $X_1, \ldots, X_n$ is

   $$M_{S_n}(t) = \prod_{i=1}^n \exp\left(t\mu_i + \frac{1}{2}\sigma_i^2 t^2\right) = \exp\left(\sum_{i=1}^n t\mu_i + \frac{1}{2}\sigma_i^2 t^2\right) = \exp\left(t\sum_{i=1}^n \mu_i + \frac{t^2}{2}\sum_{i=1}^n \sigma_i^2\right)$$

   Therefore, S_n follows the Normal distribution with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$, i.e.,

   $$\boxed{S_n \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right) \quad \text{with density} \quad f_{S_n}(x) = \frac{1}{\sqrt{2\pi\sum_{i=1}^n \sigma_i^2}} \exp\left(-\frac{(x-\sum_{i=1}^n \mu_i)^2}{2\sum_{i=1}^n \sigma_i^2}\right)}$$

   (b) For X_i with Gamma (r_i, λ) distribution.

   Recall the MGF of a Gamma distribution with shape parameter r_i and rate parameter λ is given by:

   $$M_{X_i}(t) = \left(1 - \frac{t}{\lambda}\right)^{-r_i}, \quad \text{for } t < \lambda.$$

   Since $S_n = X_1 + X_2 + \cdots + X_n$ is the sum of independent Gamma random variables with the same rate parameter λ, we have:

   $$M_{S_n}(t) = \prod_{i=1}^n \left(1 - \frac{t}{\lambda}\right)^{-r_i} = \left(1 - \frac{t}{\lambda}\right)^{-\sum_{i=1}^n r_i}.$$

   This is the MGF of a Gamma distribution with shape parameter $\sum_{i=1}^n r_i$ and rate parameter λ. Hence, S_n follows the Gamma distribution:

   $$\boxed{S_n \sim \text{Gamma}\left(\sum_{i=1}^n r_i, \lambda\right) \quad \text{with density} \quad f_{S_n}(x) = \frac{\lambda^{\sum_{i=1}^n r_i} x^{\sum_{i=1}^n r_i - 1} e^{-\lambda x}}{\Gamma\left(\sum_{i=1}^n r_i\right)}, \quad x > 0..}$$

   (c) For $X_i = Z_i^2$ with $Z_i \sim \text{Normal}(0,1)$.

   The MGF and density of a standard normal random variable $Z_i \sim \text{Normal}(0,1)$ is:

   $$M_{Z_i}(t) = \exp\left(\frac{t^2}{2}\right) \quad \text{and} \quad f_{Z_i}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

   We note that the MGF of Z_i^2 is not as simple as $M_{Z_i}(t)^2$. The MGF of $X_i = Z_i^2$ is as follows,

   $$M_{X_i}(t) = \mathbb{E}\left[e^{tZ_i^2}\right] = \int_{-\infty}^{\infty} \exp\left(tz^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(tz^2 - \frac{z^2}{2}\right) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{z^2}{2}(1-2t)\right) dz$$

   In order for the integral to be convergent, we require that $1 - 2t > 0$ or equivalently $t < \frac{1}{2}$. Then we have

   $$M_{X_i}(t) = \frac{1}{\sqrt{1-2t}}, \quad \text{for } t < \frac{1}{2}.$$

Now we can derive $M_{S_n}(t)$. Since $X_1, X_2, \ldots, X_n$ are independent, we have:

$$M_{S_n}(t) = \left(\frac{1}{\sqrt{1-2t}}\right)^n = (1-2t)^{-\frac{n}{2}}, \quad \text{for } t < \frac{1}{2}.$$

This is the MGF of a chi-squared distribution with n degrees of freedom. Therefore, S_n follows a chi-squared distribution with n degrees of freedom:

$$S_n \sim \chi^2(n) \quad \text{with density} \quad f_{S_n}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-x/2}, \quad x > 0.$$

STAT201A: Introduction to Probability at an Advanced Level October 11th, 2024

Homework # 3: PDF's, CDF's, PGF's, and Transformations

Reece D. Huff

Problems (Solutions)

1. (**Approximating Binomial Distributions**)

   The goal of this question is to empirically verify three approximations to the exact Binomial probability $\mathbb{P}(X = k)$, where $X \sim \text{Binomial}(n, p)$:

   - $\mathbb{P}(Y = k)$, where $Y \sim \text{Poisson}(np)$, the Poisson approximation with rate parameter np;
   - The normal approximation

   $$\phi(k; np, np(1-p)) := \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{(k-np)^2}{2np(1-p)}\right\};$$

   - The entropic approximation

   $$\text{Ent}(k; n, p) := \frac{1}{\sqrt{2\pi nf(1-f)}} \exp\left(-n\text{KL}(f \parallel p)\right),$$

   where $f = \frac{k}{n}$ and $\text{KL}(f \parallel p) = f \log\left(\frac{f}{p}\right) + (1-f)\log\left(\frac{1-f}{1-p}\right)$.

   For this problem, I elected to plot the absolute and relative errors for each distribution making it easier to understand the Binomial approximation accuracy of the Poisson, the Normal, and the Entropic distribution. For completeness, I attach the tables corresponding to part (a), (b), (c), (d) as well as the code for generating the plots and tables to the end of this document.

   In the analysis that follows, recall these properties about each approximation:

   - **Poisson Approximation**: The absolute error is bounded by $2np^2$. This is because when $p = \lambda/n$, we have $np^2 = \lambda^2/n$ which will be small when n is large.
   - **Normal Approximation**: Only accurate when $f = k/n$ is close to p.
   - **Entropic Approximation**: Accurate as long as the Stirling Approximation is accurate for $n-k$ and k (and the Stirling approximation is quite accurate even for small integers).

   (a) Take $n = 30$ and $p = 0.05$. Create a table (31 rows and 3 columns) containing the absolute errors for each approximation:

   $$|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|, \quad |\mathbb{P}(X = k) - \phi(k; np, np(1-p))|, \quad \text{and} \quad |\mathbb{P}(X = k) - \text{Ent}(k; n, p)|$$

   for $k = 0, 1, \ldots, 30$. (Note: The entropic approximation does not exist for $k = 0$ and $k = 30$, so only list it for $k = 1, \ldots, 29$). Based on the table, comment on the accuracy of each of the three approximations for the Binomial distribution.

   (b) Create a similar table for the relative errors:

   $$\frac{|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|}{\mathbb{P}(X = k)}, \quad \frac{|\mathbb{P}(X = k) - \phi(k; np, np(1-p))|}{\mathbb{P}(X = k)}, \quad \text{and} \quad \frac{|\mathbb{P}(X = k) - \text{Ent}(k; n, p)|}{\mathbb{P}(X = k)}$$

   for $k = 0, 1, \ldots, 30$. Based on this table, comment on the accuracy of each of the three approximations for the Binomial.

   The absolute and relative errors for $k = 0, 1, \ldots, 30$ are listed in Table 1 and Table 2. The errors have also been plotted in Figure 3.

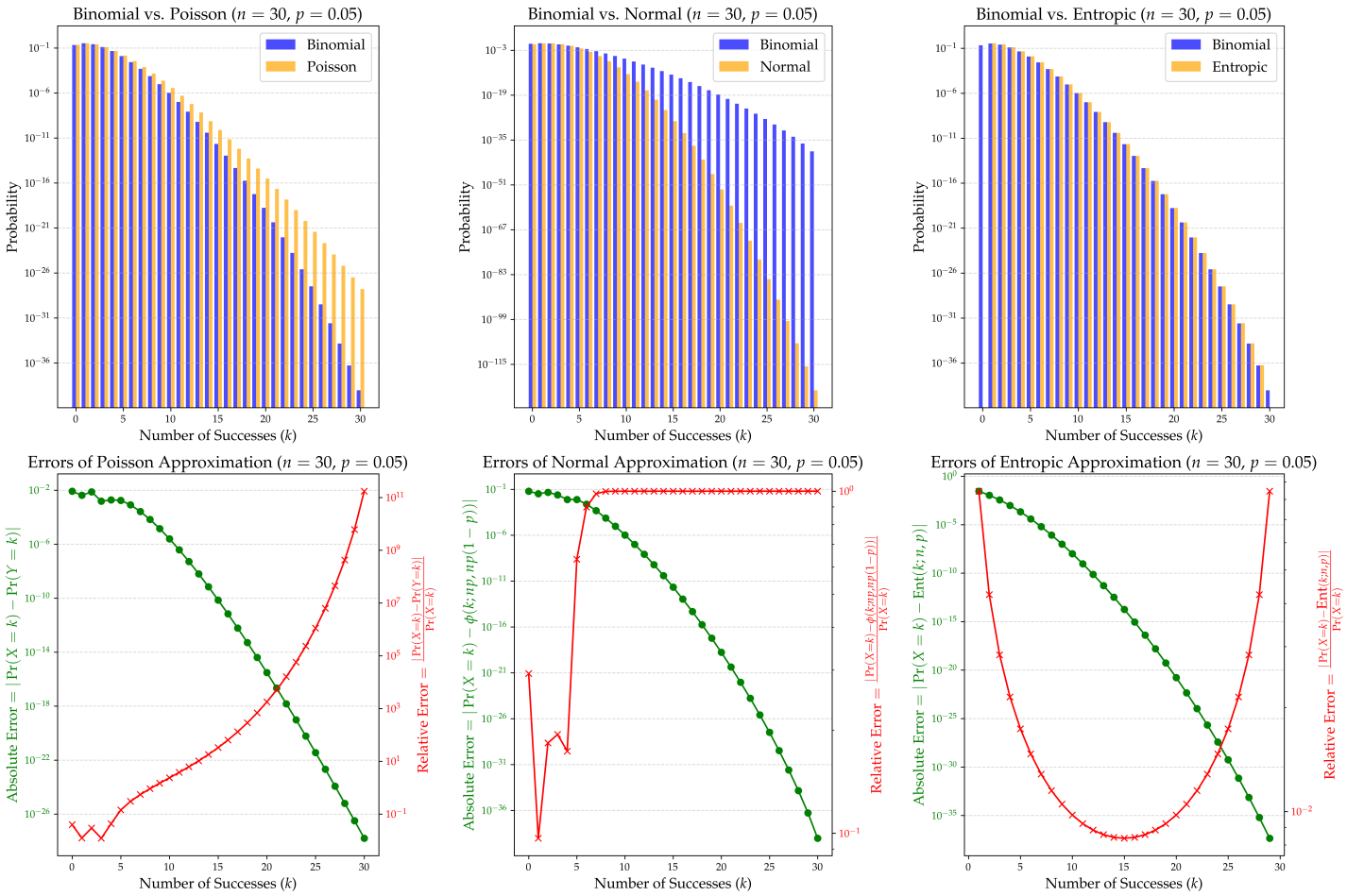


FIGURE 3: Probabilities (left) and absolute errors and relative errors (right) for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.05$.

Commentary for part (a) & (b):

- **Poisson Approximation**: We would expect the Poisson approximation only to work when k is near the mean (np). We see the absolute errors increase as k moves away from the mean.
- **Normal Approximation**: The normal approximation is only accurate when $f = k/n$ is close to p (i.e., $k = np = 1.5$). As k increases, the error blows up. This follows from the approximation going to zero much faster with such a small p value.
- **Entropic Approximation**: The Entropic approximation is clearly the best approximation in both absolute and relative error compares to the other approximations. Interestingly, the absolute error is smallest when $k = 1$ and the relative error is minimized when $k = n/2$.

(c) Repeat exercises (a) and (b) for $n = 30$ and $p = 0.25$.

The absolute and relative errors for $k = 0, 1, \ldots, 30$ are listed in Table 3 and Table 4. The errors have also been plotted in Figure 4.

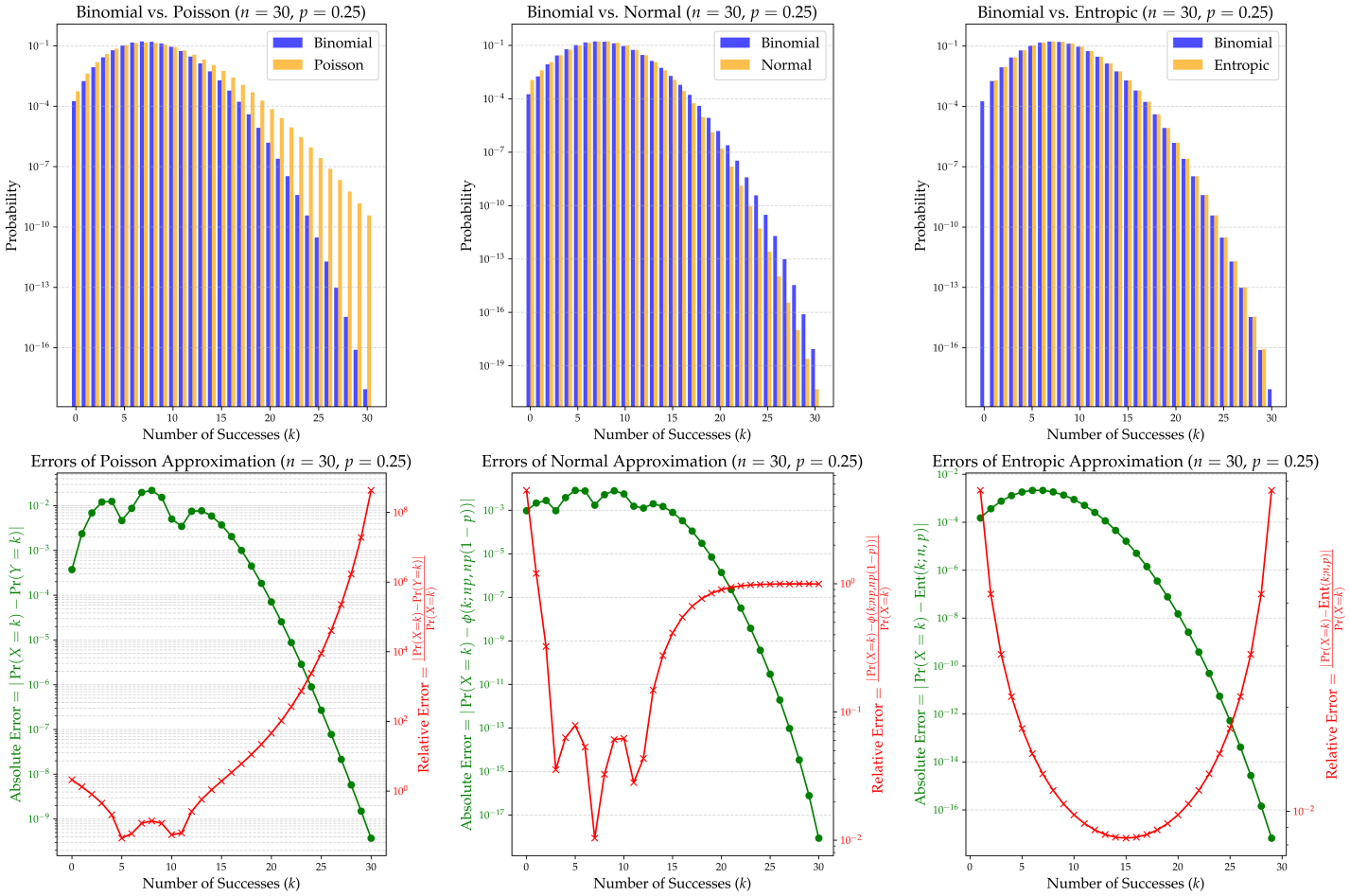


Figure 4: Probabilities (left) and absolute errors and relative errors (right) for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.25$.

Commentary for part (c):

- **Poisson Approximation**: Again, the Poisson approximation works well when np^2 is small, and now p is 5 times larger. The approximation is still able to do well when k is near the mean. However, as k increases, the absolute errors blows up even though the absolute errors become quite small.
- **Normal Approximation**: Much like before, the Normal approximation performs well when k is close to the mean ($np = 7.5$). The approximation hangs on for larger values of k and interestingly, it does very poorly for small values of k. This is becomes the exponential has had a chance to decay such that the absolute error is greater than the probability of the Binomial.
- **Entropic Approximation**: The Entropic approximation once again proves to be the most accurate across the entire range of k. It consistently provides smaller absolute and relative errors compared to the other two approximations. We now notice that the absolute errors are smallest when k is close to the mean.

(d) Repeat exercises (a) and (b) for $n = 30$ and $p = 0.5$.

The absolute and relative errors for $k = 0, 1, \ldots, 30$ are listed in Table 5 and Table 6. The errors have also been plotted in Figure 5.

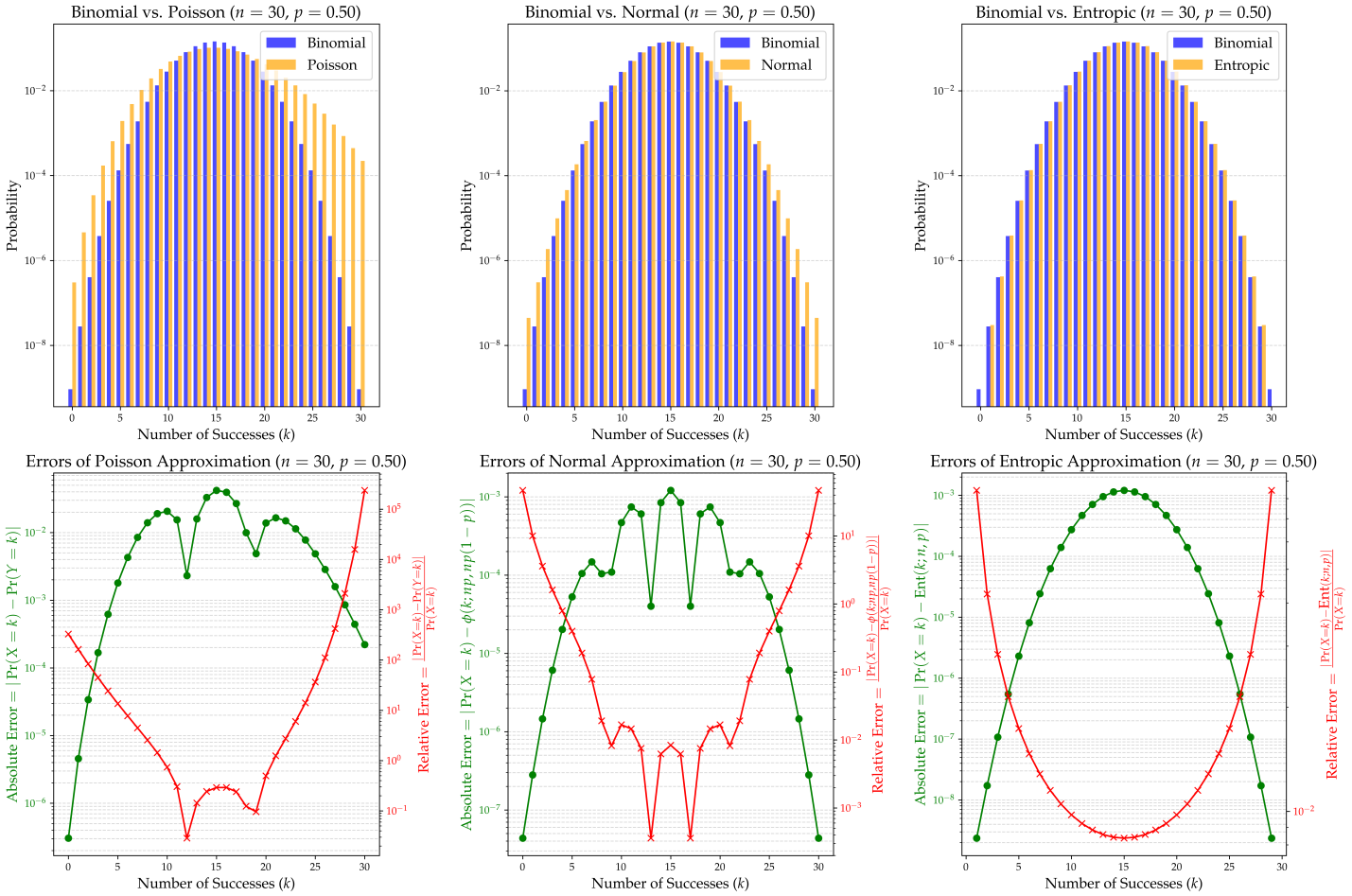


Figure 5: Probabilities (left) and absolute errors and relative errors (right) for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.25$.

Commentary for part (d):

- **Poisson Approximation**: As expected, the Poisson approximation becomes even worse now that $p = 0.50$. The approximation significantly overestimates probabilities when k is away from the mean. It still does reasonably well when k is close to the mean but not as well as before.
- **Normal Approximation**: The Normal approximation performs better when $p = 0.50$ compared to the earlier cases simply because there are more values of k that are near the mean. Still when k is small or large, the Normal approximation falls apart resulting large absolute errors (much larger than 1 in some cases).
- **Entropic Approximation**: The Entropic approximation (to no surprise) provides the best accuracy. It shows small absolute and relative errors across the entire range of k, especially when k is near the mean. It is striking that the absolute error is minimized at $k = n/2$ regardless of the p.

2. **(KL-Divergence, Multinomial)**

Let X and Y be discrete random variables with distributions p and q, respectively. So $p(k) = \mathbb{P}(X = k)$ and $q(k) = \mathbb{P}(Y = k)$. Recall that the Kullback-Leibler divergence is defined by

$$\mathrm{KL}(p \parallel q) := \mathbb{E}_p\left[\ln\left(\frac{p(X)}{q(X)}\right)\right] = \sum_k p(k)\ln\left(\frac{p(k)}{q(k)}\right).$$

(a) Show that when $q(k)$ is a Poisson distribution with parameter $\lambda > 0$, the KL-divergence is minimized by setting λ to be the mean of $p(k)$.

Let $q(k)$ follow a Poisson distribution, i.e., $q(k) = \mathbb{P}(Y = k) = e^{-\lambda}\lambda^k/k!$ for $k \in \{0, 1, ..., n\}$, $\lambda > 0$. Then the KL divergence can be written as

$$\mathrm{KL}(p \parallel q) = \sum_{k\in\{0,\dots,n\}} p(k)\log\left(\frac{p(k)}{q(k)}\right) = \sum_{k\in\{0,\dots,n\}} p(k)\log\left(\frac{p(k)k!}{e^{-\lambda}\lambda^k}\right) = \frac{\partial}{\partial\lambda}\sum_{k\in\{0,\dots,n\}} p(k)\log\left(p(k)k!e^{\lambda}\lambda^{-k}\right)$$

Setting the derivate of the KL divergence with respect to λ equal to 0 results in

$$\frac{\partial}{\partial\lambda}\mathrm{KL}(p \parallel q) = \frac{\partial}{\partial\lambda}\sum_{k\in\{0,\dots,n\}} p(k)\log\left(\frac{p(k)k!}{e^{-\lambda}\lambda^k}\right) = \sum_{k\in\{0,\dots,n\}} p(k)\left(\frac{1}{p(k)k!e^{\lambda}\lambda^{-k}}\right)\frac{\partial}{\partial\lambda}\left(p(k)k!e^{\lambda}\lambda^{-k}\right)$$

$$\frac{\partial}{\partial\lambda}\mathrm{KL}(p \parallel q) = \sum_{k\in\{0,\dots,n\}} p(k)\left(\frac{1}{e^{\lambda}\lambda^{-k}}\right)\left(e^{\lambda}\lambda^{-k} - ke^{\lambda}\lambda^{-k-1}\right) = \sum_{k\in\{0,\dots,n\}} p(k)\left(1 - \frac{ke^{\lambda}\lambda^{-k-1}}{e^{\lambda}\lambda^{-k}}\right)$$

$$\frac{\partial}{\partial\lambda}\mathrm{KL}(p \parallel q) = \sum_{k\in\{0,\dots,n\}} p(k)\left(1 - \frac{k\lambda^{-k}\lambda^{-1}}{\lambda^{-k}}\right) = \sum_{k\in\{0,\dots,n\}} p(k)\left(1 - \frac{k}{\lambda}\right) = 0$$

$$\implies \lambda \cancelto{1}{\sum_{k\in\{0,\dots,n\}} p(k)} = \sum_{k\in\{0,\dots,n\}} p(k)k$$

$$\implies \boxed{\lambda^\star = \mathbb{E}[X]}$$

(b) Remember that the entropy $H(p)$ is defined to be $H(p|q) := -\mathbb{E}_p[\ln(p(X))]$. Assume that we need to place n balls into d bins. The number of ways to place the balls, resulting in k_i total balls in bin i, for $i = 1, \dots, d$, is given by the combinatorial expression

$$\binom{n}{k_1, k_2, \dots, k_d}.$$

Now, consider the empirical distribution of the balls. Its probability mass function is $p(i) = k_i/n$. Let N_p denote the number of configurations with empirical distribution p. Show that

$$\ln(N_p) = nH(p) + O(\ln(n)),$$

where $H(p)$ is the entropy of p.

In other words, there are many more high-entropy configurations than low-entropy configurations. This suggests the intuition that, if we consider a physical system at a "macro level" (such as the distribution of gas particles in a container) then we should expect it to drift toward high-entropy configurations.

Hint: Recall Stirling's approximation

$$\ln(n!) = n\ln(n) - n + O(\ln(n)).$$

We begin by taking the log of combinatorial expression $\binom{n}{k_1,k_2,\dots,k_d}$. We have that

$$\log(N_p) = \log\binom{n}{k_1, k_2, \dots, k_d} = \log\left(\frac{n!}{k_1!k_2!\dots k_d!}\right) = \log(n!) - \log(k_1!k_2!\dots k_d!) = \log(n!) - \sum_{i=1}^{d}\log(k_i!)$$

Next, we apply Stirling's approximation to $\log(n!)$ and $\log(k_i!)$ to arrive at

$$\log(N_p) = n\log(n) - n + O(\log(n)) - \sum_{i=1}^{d}\left(k_i\log(k_i) - k_i + O(\log(k_i))\right)$$

$$\log(N_p) = n\log(n) - n - \sum_{i=1}^{d}\left(np(i)\log(np(i)) - np(i)\right) + O(\log(n)) - \sum_{i=1}^{d} O(\log(k_i))$$

$$\log(N_p) = n\log(n) - n - \sum_{i=1}^{d} np(i)\log(np(i)) + n\sum_{i=1}^{d} p(i) + O(\log(n)) - \sum_{i=1}^{d} O(\log(k_i))$$

$$\log(N_p) = n\log(n) - n - n\log(n)\underbrace{\sum_{i=1}^{d} p(i)}_{1} - n\underbrace{\sum_{i=1}^{d} p(i)\log(p(i))}_{-H(p)} + n\underbrace{\sum_{i=1}^{d} p(i)}_{1} + O(\log(n)) - \sum_{i=1}^{d} O(\log(k_i))$$

$$\log(N_p) = n\log(n) - n - n\log(n) + nH(p) + n + O(\log(n)) - \sum_{i=1}^{d} O(\log(k_i))$$

$$\boxed{\log(N_p) = nH(p) + O(\log(n))}$$

3. **(Poisson)**

Let $K = X_1 + X_2 + \cdots + X_N$, where $N \sim \text{Poisson}(\lambda)$ and $X_1, X_2, \ldots$ are independent Bernoulli(p) random variables. Assuming that N and $\{X_i\}_{i\in\mathbb{N}}$ are mutually independent, find the distribution of K.

To solve this problem, we utilize probability generating functions (p.g.f.'s). Recall that the p.g.f. of the Poisson distribution

$$G_N(t) = \mathbb{E}[t^N] = \sum_{n=0}^{\infty} t^n\,\mathbb{P}[N=n] = \sum_{n=0}^{\infty} t^n\left(\frac{e^{-\lambda}\lambda^n}{n!}\right) = e^{-\lambda}\sum_{n=0}^{\infty}\left(\frac{(t\lambda)^n}{n!}\right) \overset{1}{=} e^{-\lambda}e^{t\lambda} = e^{\lambda(t-1)} \quad \text{for all } |t|\le 1, \lambda > 0.$$

Additionally, recall the p.g.f. for Bernoulli random variables

$$G_{X_i}(t) = \mathbb{E}[t^{X_i}] = \sum_{n=0}^{\infty} t^n\,\mathbb{P}[X_i = n] = 1 - p + tp \quad \text{for all } |t|\le 1, p\in[0,1].$$

Then by the Compunding theorem from Lecture 11, we have that

$$G_K(t) = G_N(G_X(t)) = \sum_{n=0}^{\infty}(G_X(t))^n\,\mathbb{P}[N=n] = \sum_{n=0}^{\infty}(1-p+pt)^n\frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{[(1-p+pt)\lambda]^n}{n!} = e^{-\lambda}e^{(1-p+pt)\lambda}$$

$$G_K(t) = e^{\lambda-\lambda p+\lambda pt-\lambda} = e^{\lambda p(t-1)}$$

We notice the p.g.f. of K has the same form as the Poisson distribution with parameter λp. Thus we conclude

$$\boxed{K \sim \text{Poisson}(\lambda p)}$$

4. **(Joint densities)**

Let the joint density function of (X, Y) be

$$f(x,y) = \begin{cases} 3xy(x+y), & \text{if } (x,y)\in[0,1]^2, \\ 0, & \text{else.} \end{cases}$$

Calculate the covariance $\text{Cov}(X, Y)$.

Recall the covariance can be written as $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. We calculate each of these terms starting with $\mathbb{E}[X]$,

$$\mathbb{E}[X] = \int_0^1\int_0^1 xf(x,y)dxdy = \int_0^1\int_0^1 3x^2y(x+y)dxdy = 3\int_0^1\int_0^1 x^3ydxdy + 3\int_0^1\int_0^1 x^2y^2dxdy$$

[1] Recall the power series expansion of the exponential function, $\exp(x) = \sum_{n=0}^{\infty}\frac{x^n}{n!}$.

$$\mathbb{E}[X] = 3\int_0^1 \frac{y}{4}dy + 3\int_0^1 \frac{y^2}{3}dy = \frac{3}{8} + \frac{3}{9} = \frac{17}{24}.$$

By symmetry, we have that $\mathbb{E}[X] = \mathbb{E}[Y] = \frac{17}{24}$. We then calculate $\mathbb{E}[XY]$ as

$$\mathbb{E}[XY] = \int_0^1\int_0^1 xyf(x,y)dxdy = \int_0^1\int_0^1 3x^2y^2(x+y)dxdy = 3\int_0^1\int_0^1 x^3y^2dxdy + 3\int_0^1\int_0^1 x^2y^3dxdy$$
$$\mathbb{E}[XY] = \frac{3}{12} + \frac{3}{12} = \frac{1}{2}.$$

Taken together, we have the $\mathrm{Cov}(X,Y)$ is

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \frac{1}{2} - \left(\frac{17}{24}\right)^2 \quad\Longrightarrow\quad \boxed{\mathrm{Cov}(X,Y) = -\frac{1}{576}.}$$

5. **(Transformation of random variables)**

(a) Suppose X has the Cauchy distribution with density:

$$f_X(x) := \frac{1}{\pi(1+x^2)}.$$

Show that $1/X$ has the same distribution as X.
Let $Y = T(X) = 1/X$ and $X = T^{-1}(Y) = 1/Y$. Then we have

$$f_Y(y) = f_X\left(T^{-1}(y)\right)\left|\frac{\mathrm{d}T^{-1}(y)}{\mathrm{d}y}\right| = f_X\left(1/y\right)\left|-\frac{1}{y^2}\right| = \frac{1}{\pi\left(1+\left(\frac{1}{y}\right)^2\right)}\left(\frac{1}{y^2}\right) = \frac{y^2}{\pi(y^2+1)}\left(\frac{1}{y^2}\right) = \frac{1}{\pi(1+y^2)}$$

Since $f_X(x) = f_{1/X}(x)$, we have that $\boxed{X \overset{d}{=} 1/X}$ where $\overset{d}{=}$ is that the random variables are equal in distribution.

(b) Suppose $Y \sim \mathrm{Exp}(1)$. Find a function $g : (0,\infty) \to (-\infty,\infty)$ such that $g(Y)$ has the Cauchy distribution with density given by (a).

Note that any r.v. A with c.d.f. F_A has the following property: $F_A(A) \overset{d}{=} U$ where $U \sim \mathrm{Uniform}(0,1)$. Let $X = g(Y)$ for a function $g : (0,\infty) \to (-\infty,\infty)$. Recall that $X \sim \mathrm{Cauchy}(0,1)$ has a p.d.f. of $f_X(x) = \frac{1}{\pi(1+x^2)}$ and c.d.f. of $F_X(x) = \frac{1}{\pi}\tan^{-1}(x) + \frac{1}{2}$ for all $x \in (-\infty,\infty)$. Additionally, recall $Y \sim \mathrm{Exp}(1)$ has a p.d.f. of $f_Y(y) = e^{-y}$ and c.d.f. of $F_Y(y) = 1 - e^{-y}$ for all $y \geq 0$. It then follows that

$$F_Y(Y) \overset{d}{=} U \overset{d}{=} F_X(X) \overset{d}{=} F_X(g(Y)) \quad\Longrightarrow\quad F_Y(Y) \overset{d}{=} F_X(g(Y)) \quad\Longrightarrow\quad 1 - e^{-Y} \overset{d}{=} \frac{1}{\pi}\tan^{-1}(g(Y)) + \frac{1}{2}$$

Solving for $g(Y)$ results in

$$g(Y) = \tan\left(\frac{\pi}{2} - \pi e^{-Y}\right) = \frac{\sin\left(\frac{\pi}{2} - \pi e^{-Y}\right)}{\cos\left(\frac{\pi}{2} - \pi e^{-Y}\right)} = \frac{\cos\left(\pi e^{-Y}\right)}{\sin\left(\pi e^{-Y}\right)} = \cot\left(\pi e^{-Y}\right) \quad\Longrightarrow\quad \boxed{g(Y) = \cot\left(\pi e^{-Y}\right)}$$

(c) Suppose $Z \sim \mathrm{Exp}(\lambda)$, where $\lambda > 0$. Show that the distribution of $W := \lceil Z \rceil$ (here $\lceil z \rceil$ is the smallest integer that is larger than or equal to z) is Geometric. Explicitly express the parameter of the Geometric distribution in terms of λ.
To begin, we recall that $Z \sim \mathrm{Exp}(\lambda)$ has a p.d.f. of $f_Z(z) = \lambda e^{-\lambda z}$ and c.d.f. of $F_Z(z) = 1 - e^{-\lambda z}$ for all $z \geq 0$. Additionally, recall that $W \sim$ Geometric has $\mathbb{P}[W = k] = (1-p)^{k-1}p$ for all $k = 1, 2, 3, \ldots$. It then follows that

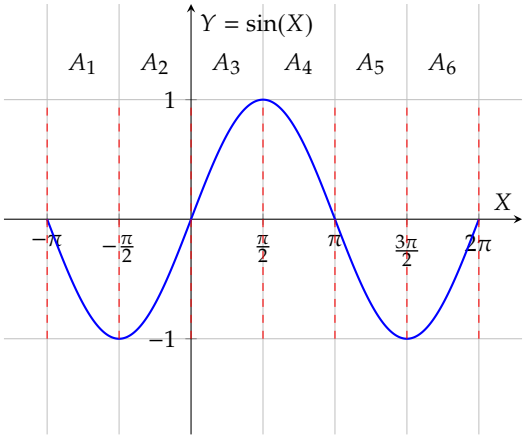
$$\mathbb{P}[W = k] = \mathbb{P}[k-1 < Z \leq k] = F_Z(k) - F_Z(k-1) = 1 - e^{-\lambda k} - \left(1 - e^{-\lambda(k-1)}\right) = e^{-\lambda(k-1)} - e^{-\lambda k} = e^{-\lambda(k-1)} - \frac{e^{-\lambda k}e^{\lambda}}{e^{\lambda}}$$
$$\mathbb{P}[W = k] = e^{-\lambda(k-1)} - \frac{e^{-\lambda k+\lambda}}{e^{\lambda}} = e^{-\lambda(k-1)} - \frac{e^{-\lambda(k-1)}}{e^{\lambda}} = e^{-\lambda(k-1)}\left(1 - \frac{1}{e^{\lambda}}\right) = \underbrace{(e^{-\lambda})^{(k-1)}}_{(1-p)^{k-1}}\underbrace{\left(1 - e^{-\lambda}\right)}_{p}$$

By inspection, we see that $\boxed{W = \lceil Z \rceil \sim \text{Geometric with a parameter } p = 1 - e^{-\lambda}.}$

6. **(Transformation of random variables)**

Suppose $X \sim \text{Uniform}[-\pi, 2\pi]$. Find the p.d.f. of $Y = \sin(X)$.

This problem involves what we call a *many-to-one* transformation T. Therefore, we begin by splitting the domain of X into regions of monotonicity of $Y = \sin(X)$.



We then define the inverse transforms for each partition as

$$T_1^{-1}(y) = \begin{cases} \sin^{-1}(y) - \pi/2 & \text{for } x \in A_1 = [-\pi, -\pi/2] \quad y \in [-1, 0] \\ \sin^{-1}(y) & \text{for } x \in A_2 = [-\pi/2, 0] \quad y \in [-1, 0] \\ \sin^{-1}(y) + \pi/2 & \text{for } x \in A_3 = (0, \pi/2] \quad y \in (0, 1] \\ \sin^{-1}(y) + \pi & \text{for } x \in A_4 = [\pi/2, \pi) \quad y \in (0, 1] \\ \sin^{-1}(y) + 3\pi/2 & \text{for } x \in A_5 = [\pi, 3\pi/2] \quad y \in [-1, 0] \\ \sin^{-1}(y) + 2\pi & \text{for } x \in A_6 = [3\pi/2, 2\pi] \quad y \in [-1, 0] \end{cases}$$

Then for $y \in [-1, 0]$, we have

$$f_Y(y) = \sum_{i \in \{1,2,5,6\}} f_X\left(T_i^{-1}(y)\right) \left| \frac{dT_i^{-1}(y)}{dy} \right| = \sum_{i \in \{1,2,5,6\}} \frac{1}{3\pi} \frac{1}{\sqrt{1-y^2}} = \frac{4}{3\pi} \frac{1}{\sqrt{1-y^2}}$$

Similarly, for $y \in (0, 1]$, we have

$$f_Y(y) = \sum_{i \in \{3,4\}} f_X\left(T_i^{-1}(y)\right) \left| \frac{dT_i^{-1}(y)}{dy} \right| = \sum_{i \in \{3,4\}} \frac{1}{3\pi} \frac{1}{\sqrt{1-y^2}} = \frac{2}{3\pi} \frac{1}{\sqrt{1-y^2}}$$

Taken together we have that

$$\boxed{f_Y(y) = \begin{cases} \frac{4}{3\pi\sqrt{1-y^2}} & \text{for } y \in [-1, 0], \\ \frac{2}{3\pi\sqrt{1-y^2}} & \text{for } y \in (0, 1], \\ 0 & \text{otherwise.} \end{cases}}$$

(a) Take $n = 30$ and $p = 0.05$. Create a table (31 rows and 3 columns) containing the absolute errors for each approximation:

$$|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|, \quad |\mathbb{P}(X = k) - \phi(k; np, np(1-p))|, \quad \text{and} \quad |\mathbb{P}(X = k) - \text{Ent}(k; n, p)|$$

for $k = 0, 1, \ldots, 30$. (Note: The entropic approximation does not exist for $k = 0$ and $k = 30$, so only list it for $k = 1, \ldots, 29$). Based on the table, comment on the accuracy of each of the three approximations for the Binomial distribution.

Table 1: Absolute errors for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.05$.

k	$\lvert\mathbb{P}(X = k) - \mathbb{P}(Y = k)\rvert$	$\lvert\mathbb{P}(X = k) - \phi(k; np, np(1-p))\rvert$	$\lvert\mathbb{P}(X = k) - \text{Ent}(k; n, p)\rvert$
0	0.0085	0.063	
1	0.0042	0.033	0.029
2	0.0076	0.047	0.011
3	0.0015	0.025	0.0036
4	0.0019	0.0078	0.00097
5	0.0018	0.0078	0.00021
6	0.00082	0.0024	4.0×10^{-5}
7	0.00027	0.00048	6.3×10^{-6}
8	6.8×10^{-5}	7.4×10^{-5}	8.5×10^{-7}
9	1.4×10^{-5}	9.5×10^{-6}	1.0×10^{-7}
10	2.5×10^{-6}	1.1×10^{-6}	1.0×10^{-8}
11	3.8×10^{-7}	1.0×10^{-7}	9.3×10^{-10}
12	5.2×10^{-8}	8.4×10^{-9}	7.4×10^{-11}
13	6.4×10^{-9}	6.1×10^{-10}	5.2×10^{-12}
14	7.1×10^{-10}	3.9×10^{-11}	3.3×10^{-13}
15	7.3×10^{-11}	2.2×10^{-12}	1.8×10^{-14}
16	6.9×10^{-12}	1.1×10^{-13}	9.1×10^{-16}
17	6.1×10^{-13}	4.7×10^{-15}	4.0×10^{-17}
18	5.1×10^{-14}	1.8×10^{-16}	1.6×10^{-18}
19	4.1×10^{-15}	5.9×10^{-18}	5.5×10^{-20}
20	3.0×10^{-16}	1.7×10^{-19}	1.7×10^{-21}
21	2.2×10^{-17}	4.3×10^{-21}	4.5×10^{-23}
22	1.5×10^{-18}	9.3×10^{-23}	1.1×10^{-24}
23	9.7×10^{-20}	1.7×10^{-24}	2.2×10^{-26}
24	6.1×10^{-21}	2.6×10^{-26}	3.8×10^{-28}
25	3.6×10^{-22}	3.3×10^{-28}	5.7×10^{-30}
26	2.1×10^{-23}	3.3×10^{-30}	7.1×10^{-32}
27	1.2×10^{-24}	2.6×10^{-32}	7.4×10^{-34}
28	6.2×10^{-26}	1.5×10^{-34}	6.2×10^{-36}
29	3.2×10^{-27}	5.3×10^{-37}	4.5×10^{-38}
30	1.6×10^{-28}	9.3×10^{-40}	

(b) Create a similar table for the relative errors:

$$\frac{|\mathbb{P}(X=k)-\mathbb{P}(Y=k)|}{\mathbb{P}(X=k)}, \quad \frac{|\mathbb{P}(X=k)-\phi(k;np,np(1-p))|}{\mathbb{P}(X=k)}, \quad \text{and} \quad \frac{|\mathbb{P}(X=k)-\text{Ent}(k;n,p)|}{\mathbb{P}(X=k)}$$

for $k = 0, 1, \ldots, 30$. Based on this table, comment on the accuracy of each of the three approximations for the Binomial.

Table 2: Relative errors for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.05$.

k	$\frac{\lvert\mathbb{P}(X=k)-\mathbb{P}(Y=k)\rvert}{\mathbb{P}(X=k)}$	$\frac{\lvert\mathbb{P}(X=k)-\phi(k;np,np(1-p))\rvert}{\mathbb{P}(X=k)}$	$\frac{\lvert\mathbb{P}(X=k)-\text{Ent}(k;n,p)\rvert}{\mathbb{P}(X=k)}$
0	0.040	0.29	
1	0.012	0.097	0.085
2	0.029	0.18	0.042
3	0.012	0.19	0.028
4	0.043	0.17	0.021
5	0.14	0.63	0.017
6	0.30	0.90	0.015
7	0.55	0.98	0.013
8	0.92	1.0	0.011
9	1.5	1.0	0.011
10	2.4	1.0	0.0098
11	3.8	1.0	0.0092
12	6.2	1.0	0.0088
13	10.	1.0	0.0086
14	18.	1.0	0.0084
15	33.	1.0	0.0084
16	64.	1.0	0.0084
17	$1.3 \times 10^{+2}$	1.0	0.0086
18	$2.9 \times 10^{+2}$	1.0	0.0088
19	$6.9 \times 10^{+2}$	1.0	0.0092
20	$1.8 \times 10^{+3}$	1.0	0.0098
21	$5.1 \times 10^{+3}$	1.0	0.011
22	$1.6 \times 10^{+4}$	1.0	0.011
23	$5.7 \times 10^{+4}$	1.0	0.013
24	$2.3 \times 10^{+5}$	1.0	0.015
25	$1.1 \times 10^{+6}$	1.0	0.017
26	$6.3 \times 10^{+6}$	1.0	0.021
27	$4.5 \times 10^{+7}$	1.0	0.028
28	$4.3 \times 10^{+8}$	1.0	0.042
29	$6.1 \times 10^{+9}$	1.0	0.085
30	$1.7 \times 10^{+11}$	1.0	

(c) Repeat exercises (a) and (b) for $n = 30$ and $p = 0.25$.

Table 3: Absolute errors for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.25$.

k	$\lvert \mathbb{P}(X = k) - \mathbb{P}(Y = k) \rvert$	$\lvert \mathbb{P}(X = k) - \phi(k; np, np(1-p)) \rvert$	$\lvert \mathbb{P}(X = k) - \text{Ent}(k; n, p) \rvert$
0	0.00037	0.00095	
1	0.0024	0.0021	0.00015
2	0.0069	0.0028	0.00037
3	0.012	0.00095	0.00076
4	0.012	0.0038	0.0013
5	0.0046	0.0082	0.0018
6	0.0087	0.0077	0.0021
7	0.020	0.0017	0.0021
8	0.022	0.0052	0.0018
9	0.015	0.0079	0.0014
10	0.0050	0.0056	0.00089
11	0.0035	0.0015	0.00051
12	0.0075	0.0013	0.00026
13	0.0077	0.0020	0.00011
14	0.0059	0.0015	4.6×10^{-5}
15	0.0037	0.00080	1.6×10^{-5}
16	0.0020	0.00033	5.1×10^{-6}
17	0.0010	0.00011	1.4×10^{-6}
18	0.00045	3.1×10^{-5}	3.5×10^{-7}
19	0.00018	7.1×10^{-6}	7.7×10^{-8}
20	7.1×10^{-5}	1.4×10^{-6}	1.5×10^{-8}
21	2.6×10^{-5}	2.3×10^{-7}	2.6×10^{-9}
22	8.7×10^{-6}	3.2×10^{-8}	3.8×10^{-10}
23	2.9×10^{-6}	3.8×10^{-9}	5.0×10^{-11}
24	8.9×10^{-7}	3.7×10^{-10}	5.5×10^{-12}
25	2.7×10^{-7}	3.0×10^{-11}	5.2×10^{-13}
26	7.7×10^{-8}	1.9×10^{-12}	4.1×10^{-14}
27	2.2×10^{-8}	9.5×10^{-14}	2.7×10^{-15}
28	5.8×10^{-9}	3.4×10^{-15}	1.4×10^{-16}
29	1.5×10^{-9}	7.8×10^{-17}	6.6×10^{-18}
30	3.7×10^{-10}	8.6×10^{-19}	

Table 4: Relative errors for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.25$.

k	$\frac{\lvert \mathbb{P}(X=k)-\mathbb{P}(Y=k)\rvert}{\mathbb{P}(X=k)}$	$\frac{\lvert \mathbb{P}(X=k)-\phi(k;np,np(1-p))\rvert}{\mathbb{P}(X=k)}$	$\frac{\lvert \mathbb{P}(X=k)-\mathrm{Ent}(k;n,p)\rvert}{\mathbb{P}(X=k)}$
0	2.1	5.3	
1	1.3	1.2	0.085
2	0.80	0.32	0.042
3	0.45	0.035	0.028
4	0.21	0.063	0.021
5	0.044	0.078	0.017
6	0.060	0.053	0.015
7	0.12	0.010	0.013
8	0.14	0.033	0.011
9	0.12	0.061	0.011
10	0.055	0.062	0.0098
11	0.063	0.028	0.0092
12	0.26	0.043	0.0088
13	0.57	0.15	0.0086
14	1.1	0.28	0.0084
15	1.9	0.41	0.0084
16	3.4	0.55	0.0084
17	6.1	0.67	0.0086
18	11.	0.77	0.0088
19	22.	0.84	0.0092
20	46.	0.90	0.0098
21	$1.0 \times 10^{+2}$	0.94	0.011
22	$2.6 \times 10^{+2}$	0.96	0.011
23	$7.4 \times 10^{+2}$	0.98	0.013
24	$2.4 \times 10^{+3}$	0.99	0.015
25	$8.9 \times 10^{+3}$	0.99	0.017
26	$4.0 \times 10^{+4}$	0.99	0.021
27	$2.3 \times 10^{+5}$	1.0	0.028
28	$1.7 \times 10^{+6}$	1.0	0.042
29	$1.9 \times 10^{+7}$	1.0	0.085
30	$4.3 \times 10^{+8}$	0.99	

(d) Repeat exercises (a) and (b) for $n = 30$ and $p = 0.50$.

Table 5: Absolute errors for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.50$.

k	$\lvert \mathbb{P}(X = k) - \mathbb{P}(Y = k) \rvert$	$\lvert \mathbb{P}(X = k) - \phi(k; np, np(1-p)) \rvert$	$\lvert \mathbb{P}(X = k) - \text{Ent}(k; n, p) \rvert$
0	3.0×10^{-7}	4.4×10^{-8}	
1	4.6×10^{-6}	2.8×10^{-7}	2.4×10^{-9}
2	3.4×10^{-5}	1.5×10^{-6}	1.7×10^{-8}
3	0.00017	6.1×10^{-6}	1.1×10^{-7}
4	0.00062	2.0×10^{-5}	5.5×10^{-7}
5	0.0018	5.3×10^{-5}	2.3×10^{-6}
6	0.0043	0.00010	8.1×10^{-6}
7	0.0085	0.00015	2.4×10^{-5}
8	0.014	0.00010	6.3×10^{-5}
9	0.019	0.00011	0.00014
10	0.021	0.00047	0.00027
11	0.015	0.00074	0.00047
12	0.0023	0.00061	0.00071
13	0.016	4.0×10^{-5}	0.00096
14	0.033	0.00084	0.0011
15	0.042	0.0012	0.0012
16	0.039	0.00084	0.0011
17	0.027	4.0×10^{-5}	0.00096
18	0.0099	0.00061	0.00071
19	0.0049	0.00074	0.00047
20	0.014	0.00047	0.00027
21	0.017	0.00011	0.00014
22	0.015	0.00010	6.3×10^{-5}
23	0.011	0.00015	2.4×10^{-5}
24	0.0077	0.00010	8.1×10^{-6}
25	0.0048	5.3×10^{-5}	2.3×10^{-6}
26	0.0028	2.0×10^{-5}	5.5×10^{-7}
27	0.0016	6.1×10^{-6}	1.1×10^{-7}
28	0.00085	1.5×10^{-6}	1.7×10^{-8}
29	0.00044	2.8×10^{-7}	2.4×10^{-9}
30	0.00022	4.4×10^{-8}	

Table 6: Relative errors for the between the Binomial distribution and the Poisson, Normal, and Entropic distribution for $n = 30$ and $p = 0.50$.

k	$\frac{\lvert \mathbb{P}(X=k)-\mathbb{P}(Y=k)\rvert}{\mathbb{P}(X=k)}$	$\frac{\lvert \mathbb{P}(X=k)-\phi(k;np,np(1-p))\rvert}{\mathbb{P}(X=k)}$	$\frac{\lvert \mathbb{P}(X=k)-\mathrm{Ent}(k;n,p)\rvert}{\mathbb{P}(X=k)}$
0	$3.3 \times 10^{+2}$	47.	
1	$1.6 \times 10^{+2}$	10.	0.085
2	84.	3.6	0.042
3	45.	1.6	0.028
4	24.	0.79	0.021
5	14.	0.40	0.017
6	7.8	0.19	0.015
7	4.5	0.078	0.013
8	2.6	0.019	0.011
9	1.4	0.0082	0.011
10	0.74	0.017	0.0098
11	0.30	0.015	0.0092
12	0.029	0.0075	0.0088
13	0.14	0.00036	0.0086
14	0.24	0.0062	0.0084
15	0.29	0.0084	0.0084
16	0.29	0.0062	0.0084
17	0.24	0.00036	0.0086
18	0.12	0.0075	0.0088
19	0.096	0.015	0.0092
20	0.49	0.017	0.0098
21	1.2	0.0082	0.011
22	2.7	0.019	0.011
23	6.0	0.078	0.013
24	14.	0.19	0.015
25	37.	0.40	0.017
26	$1.1 \times 10^{+2}$	0.79	0.021
27	$4.2 \times 10^{+2}$	1.6	0.028
28	$2.1 \times 10^{+3}$	3.6	0.042
29	$1.6 \times 10^{+4}$	10.	0.085
30	$2.4 \times 10^{+5}$	47.	

Code for Question 1

```
import os
from glob import glob
import numpy as np
from math import factorial, comb, exp, pi, sqrt, log
import re
from PIL import Image
import matplotlib.pyplot as plt
import matplotlib as mpl
mpl.rcParams['text.usetex'] = True
mpl.rcParams['text.latex.preamble'] = r'\usepackage{mathpazo}'
mpl.rcParams["font.family"] = "Palatino"

# Create a figure dir if it does not already exists
FIG_DIR = os.path.join(os.path.abspath( os.path.dirname( __file__ ) ), "figures")
if not os.path.exists(FIG_DIR):
    os.makedirs(FIG_DIR, exist_ok=True)

#%% Printing function
def replace_sci_notation(input_str):
    return re.sub(r'(\d\.\d)e([+-])0*([0-9]+)', r'\1 \\times 10^{\2\3}', input_str)

def np_to_content(arr):
    output = []
    for k, r in enumerate(arr):
        l = "\t\t${:d}$ & ${:#.2g}$ & ${:#.2g}$ & ${:#.2g}$ \\\\".format(int(r[0]), float(r[1]),
            float(r[2]), float(r[3]))
        l = replace_sci_notation(l)
        if k == 0 or k == (arr.shape[0]-1):
            l = l[:l.rfind("&")] + "& \\\\"
        output.append(l)
    return output

def print_table(data_in, title, caption, table):
    # \begin{table}[!ht]
    #     \centering
    #     \begin{tabular}{C{1cm} C{5cm} C{5cm} C{5cm}}
    #         \toprule
    #         $k$
    #         & $|\Pr(X = k) - \Pr(Y = k)|$
    #         & $|\Pr(X = k) - \phi(k; np, np(1 - p))|$
    #         & $|\Pr(X = k) - \textnormal{Ent}(k; n, p)|$
    #         \\
    #         \midrule
    #         0 & 0.01 & 0.01 & 0.01 \\
    #         \bottomrule
    #     \end{tabular}
    # \end{table}
    # Update the caption to include a \label
    error_type = caption[caption.find("{")+1:caption.find("errors")-1]
    inds  = [i for i, c in enumerate(list(caption)) if c == "$"]
    n_str = caption[inds[0]+1:inds[1]].replace("=","_")
    p_str = caption[inds[2]+1:inds[3]].replace("=","_")
    caption += "\n\t\label{%s_%s_%s}" % (error_type, n_str, p_str)

    frontmatter = [
    "\\begin{table}[!ht]",
        "\t\centering",
        "\t\\begin{tabular}{C{1cm} C{5cm} C{5cm} C{5cm}}",
            "\t\t\\toprule"
    ]
    frontmatter.insert(2, caption)

    middlematter = [
```

```
        "\t\t\\\\",
        "\t\t\midrule",
    ]
    content = np_to_content(data_in)
    endmatter = [
            "\t\t\\bottomrule",
        "\t\end{tabular}",
    "\end{table}"
    ]
    # endmatter.insert(-1, caption)

    print("\n")
    print("="*50)
    print(table)
    print("="*50)
    for line in frontmatter + title + middlematter + content + endmatter:
        print(line)
    print("="*50)
    print("\n")

def plot_approximation(approx_probs, abs_errors, rel_errors, ylabels):

    # Plotting the Two-Panel Plot for Approximation
    # Compute PMFs
    k_values      = np.arange(0, len(abs_errors))
    binom_probs   = [binomial(n,p,k) for k in k_values]

    # Create the Two-Panel Plot

    # Option 1
    # _, axes = plt.subplots(1, 2, figsize=(12, 6))

    # Option 2
    _, axes = plt.subplots(2, 1, figsize=(6, 12))

    # --- Left Panel: Binomial vs Poisson ---
    axes[0].bar(k_values - 0.2, binom_probs, width=0.4, label='Binomial', alpha=0.7, color='blue',
        log=True)
    axes[0].bar(k_values + 0.2, approx_probs, width=0.4, label=ylabels[0], alpha=0.7,
        color='orange', log=True)
    axes[0].set_xlabel(r'Number of Successes ($k$)', fontsize=14)
    axes[0].set_ylabel(r'Probability', fontsize=14)
    axes[0].set_title(r'Binomial vs. %s ($n=%d$, $p=%0.2f$)' % (ylabels[0],n,p), fontsize=16)
    axes[0].legend(fontsize=14, loc='upper right')
    axes[0].grid(True, linestyle='--', alpha=0.5, which='both', axis='y')

    # --- Right Panel: Absolute and Relative Errors --- #
    ax1 = axes[1]
    ax2 = ax1.twinx()

    # Plot Absolute Errors
    ax1.plot(k_values, abs_errors, color='green', marker='o')
    ax1.set_xlabel(r'Number of Successes ($k$)', fontsize=14)
    ax1.set_ylabel(r'Absolute Error $ = %s$' % ylabels[1], color='green', fontsize=14)
    ax1.tick_params(axis='y', labelcolor='green')
    ax1.set_yscale('log')
    ax1.grid(True, linestyle='--', alpha=0.5, which='both', axis='y')

    # Plot Relative Errors
    ax2.plot(k_values, rel_errors, color='red', marker='x')
    ax2.set_ylabel(r'Relative Error $ = %s$' % ylabels[2], color='red', fontsize=14)
    ax2.tick_params(axis='y', labelcolor='red')
    ax2.set_yscale('log')

    # Title
```

```
    axes[1].set_title(r'Errors of %s Approximation ($n=%d$, $p=%0.2f$)' % (ylabels[0],n,p),
        fontsize=16)

    plt.tight_layout()
    fig_name = "%s_%d_%0.2f.png" % (ylabels[0],n,p)
    fig_path = os.path.join(FIG_DIR, fig_name)
    plt.savefig(fig_path, dpi=600)

def plot_data(n,p,abs_data,rel_data):

    # Poisson labels and approximate probabilities
    ylabels = [
        'Poisson',
        r'|\Pr(X = k) - \Pr(Y = k)|',
        r'\frac{|\Pr(X = k) - \Pr(Y = k)|}{\Pr(X = k)}'
    ]
    approx_probs  = [poisson(n*p, int(k)) for k in abs_data[:,0]]
    plot_approximation(approx_probs, abs_data[:,1], rel_data[:,1], ylabels)

    # Normal labels and approximate probabilities
    ylabels = [
        'Normal',
        r'|\Pr(X = k) - \phi(k; np, np(1 - p))|',
        r'\frac{|\Pr(X = k) - \phi(k; np, np(1 - p))|}{\Pr(X = k)}'
    ]
    approx_probs  = [normal(n,p,int(k)) for k in abs_data[:,0]]
    plot_approximation(approx_probs, abs_data[:,2], rel_data[:,2], ylabels)

    # Entropic labels and approximate probabilities
    ylabels = [
        'Entropic',
        r'|\Pr(X = k) - \textnormal{Ent}(k; n, p)|',
        r'\frac{|\Pr(X = k) - \textnormal{Ent}(k; n, p)|}{\Pr(X = k)}'
    ]
    approx_probs  = [entropic(n,p,int(k)) for k in abs_data[:,0]]
    plot_approximation(approx_probs, abs_data[:,3], rel_data[:,3], ylabels)

    # Join the figures as one png
    join_figs(n,p)

def join_figs(n,p):

    dists  = ['Poisson', 'Normal', 'Entropic']
    search = os.path.join(FIG_DIR, "*%d_%.2f.png" % (n,p))
    paths  = glob(search)
    images = []
    for dist in dists:
        for path in paths:
            if dist in path:
                print(path)
                images.append(Image.open(path))
    im_size = images[0].size

    # Option 1
    # new_im  = Image.new('RGB', (im_size[0],3*im_size[1]), (255,255,255))
    # new_im.paste(images[0], (0,0))
    # new_im.paste(images[1], (0,1*im_size[1]))
    # new_im.paste(images[2], (0,2*im_size[1]))
    # new_im_path = os.path.join(FIG_DIR, "n%d_p%.2f.png" % (n,p))
    # new_im.save(new_im_path, "PNG")

    # Option 2
    new_im  = Image.new('RGB', (3*im_size[0],im_size[1]), (255,255,255))
    new_im.paste(images[0], (0,0))
    new_im.paste(images[1], (1*im_size[0], 0))
```

```
    new_im.paste(images[2], (2*im_size[0], 0))
    new_im_path = os.path.join(FIG_DIR, "n%d_p%.2f.png" % (n,p))
    new_im.save(new_im_path, "PNG")


#%% Distributions
def binomial(n,p,k):
    return comb(n,k) * (p**k) * ((1-p)**(n-k))

def poisson(lambda_, k):
    return ( exp(-lambda_) * (lambda_**k) ) / (factorial(k))

def normal(n,p,k):
    return ( 1/sqrt(2*pi*n*p*(1-p)) ) * ( exp( - ((k-n*p)**2) / (2*n*p*(1-p)) ) )

def entropic(n,p,k):
    if k == 0 or k == n:
        return np.nan
    else:
        f  = k/n
        KL = f*log(f/p) + (1-f)*log((1-f)/(1-p))
        return ( 1/sqrt(2*pi*n*f*(1-f)) ) * ( exp(-n*KL) )

#%% Error functions
def absolute_error(n,p,k_min,k_max):
    out = np.zeros((k_max-k_min+1, 4))
    for i, k in enumerate(range(k_min, k_max+1)):
        out[i,0]   = k
        binom_k    = binomial(n,p,k)
        poisson_k  = poisson(n*p,k)
        normal_k   = normal(n,p,k)
        entropic_k = entropic(n,p,k)
        out[i,1]   = abs(binom_k-poisson_k)
        out[i,2]   = abs(binom_k-normal_k)
        out[i,3]   = abs(binom_k-entropic_k)
    title_out = [
        "\t\t$k$",
        "\t\t& $|\Pr(X = k) - \Pr(Y = k)|$",
        "\t\t& $|\Pr(X = k) - \phi(k; np, np(1 - p))|$",
        "\t\t& $|\Pr(X = k) - \\textnormal{Ent}(k; n, p)|$"
    ]
    caption_out = "\t\caption{Absolute errors for the between the Binomial distribution and the
        Poisson, Normal, and Entropic distribution for $n=%d$ and $p=%.2f$.}" % (n,p)
    table_out  = "Absolute errors for $n=%d$ and $p=%.2f$" % (n,p)
    return out, title_out, caption_out, table_out

def relative_error(n,p,k_min,k_max):
    out = np.zeros((k_max-k_min+1, 4))
    for i, k in enumerate(range(k_min, k_max+1)):
        out[i,0]   = k
        binom_k    = binomial(n,p,k)
        poisson_k  = poisson(n*p,k)
        normal_k   = normal(n,p,k)
        entropic_k = entropic(n,p,k)
        out[i,1]   = abs(binom_k-poisson_k)/binom_k
        out[i,2]   = abs(binom_k-normal_k)/binom_k
        out[i,3]   = abs(binom_k-entropic_k)/binom_k
    title_out = [
        "\t\t$k$",
        "\t\t& $\\frac{|\Pr(X = k) - \Pr(Y = k)|}{\Pr(X = k)}$",
        "\t\t& $\\frac{|\Pr(X = k) - \phi(k; np, np(1 - p))|}{\Pr(X = k)}$",
        "\t\t& $\\frac{|\Pr(X = k) - \\textnormal{Ent}(k; n, p)|}{\Pr(X = k)}$"
    ]
    caption_out = "\t\caption{Relative errors for the between the Binomial distribution and the
        Poisson, Normal, and Entropic distribution for $n=%d$ and $p=%.2f$.}" % (n,p)
```

```
253     table_out  = "Relative errors for $n=%d$ and $p=%.2f$" % (n,p)
254     return out, title_out, caption_out, table_out
255 
256 #%% Part (a) and (b)
257 n     = 30
258 p     = 0.05
259 abs_error, title, caption, table = absolute_error(n,p,0,30)
260 print_table(abs_error, title, caption, table)
261 rel_error, title, caption, table = relative_error(n,p,0,30)
262 print_table(rel_error, title, caption, table)
263 plot_data(n,p,abs_error,rel_error)
264 
265 #%% Part (c)
266 n     = 30
267 p     = 0.25
268 abs_error, title, caption, table = absolute_error(n,p,0,30)
269 print_table(abs_error, title, caption, table)
270 rel_error, title, caption, table = relative_error(n,p,0,30)
271 print_table(rel_error, title, caption, table)
272 plot_data(n,p,abs_error,rel_error)
273 
274 #%% Part (d)
275 n     = 30
276 p     = 0.5
277 abs_error, title, caption, table = absolute_error(n,p,0,30)
278 print_table(abs_error, title, caption, table)
279 rel_error, title, caption, table = relative_error(n,p,0,30)
280 print_table(rel_error, title, caption, table)
281 plot_data(n,p,abs_error,rel_error)
```

STAT201A: Introduction to Probability at an Advanced Level November 5th, 2024

Homework # 4: Ordered Statistics and Conditional Expectations & Variances

Reece D. Huff

Problems (Solutions)

1. (**Order statistics**) Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exp(λ) distribution, where $\lambda > 0$, and let $X_{(i)}$ be the order statistics for $i = 1, \ldots, n$.

   (a) Find the distribution of $X_{(1)}$.

   Recall the density of j-th order statistic is

   $$f_{X_{(j)}}(x) = n\binom{n-1}{j-1} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{n-j}$$

   Then the density of the 1st order statistic is

   $$f_{X_{(1)}}(x) = n f_X(x) \left[1 - F_X(x)\right]^{n-1}$$

   and since $f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$, we have

   $$f_{X_{(1)}}(x) = n\lambda e^{-\lambda x} \left[1 - (1 - e^{-\lambda x})\right]^{n-1} = n\lambda e^{-\lambda x} \left[e^{-\lambda x}\right]^{n-1} = n\lambda e^{-n\lambda x}$$

   which shows that distribution of $X_{(1)}$ is exponential with parameter $n\lambda$, i.e.,

   $$\boxed{X_{(1)} \sim \text{Exp}(n\lambda)}$$

   (b) Using the memoryless property, find the distribution of $X_{(i+1)} - X_{(i)}$ for $i = 1, \ldots, n-1$.

   Recall that the memoryless property of the exponential distribution states that $\mathbb{P}[X > s + t \mid X > t] = \mathbb{P}[X > s]$. Now we define the gaps between them order statistics as

   $$L_i = \begin{cases} X_{(i+1)} & \text{for } i = 0, \\ X_{(i+1)} - X_{(i)} & \text{for } i = 1, \ldots, n-1. \end{cases}$$

   We additionally define sets of indices. Let $\mathcal{S}$ be the set of indices above index (i). Then there exists some subset of indices of the original iid random variables X_j that correspond to the order statistics indices in $\mathcal{S}$. We define this set as $\mathcal{A}$. We all define the $\mathcal{S}'$ and $\mathcal{A}'$ in the same way as $\mathcal{S}$ and $\mathcal{A}$, but now we include index (i). We have that

   $\mathcal{S} = \{(i+1), (i+2), \ldots, (n)\}$ that correspond to some set of indices of X_j $\mathcal{A} = \{j_1, j_2, \ldots, j_{n-i}\}$

   $\mathcal{S}' = \{(i), (i+1), \ldots, (n)\}$ that correspond to some set of indices of X_j $\mathcal{A}' = \{j_0, j_1, \ldots, j_{n-i}\}$

   Note that the indices within $\mathcal{A}$ and $\mathcal{A}'$ are not unique. Their uniqueness is not critical to what follows. Importantly, the cardinality of $\mathcal{A}$ is $|\mathcal{A}| = n - i$. It then follows that for all $i = 1, \ldots, n-1$,

   $$\mathbb{P}[L_i > x] = \mathbb{P}[X_{(i+1)} - X_{(i)} > x] = \mathbb{P}[X_{(i+1)} - X_{(i)} > x | X_{(i)} > t] = \mathbb{P}[X_{(i+1)} > x + t | X_{(i)} > t]$$

   $$\mathbb{P}[L_i > x] = \mathbb{P}\left[\bigcap_{j\in\mathcal{A}} X_j > x + t \,\middle|\, \bigcap_{j\in\mathcal{A}'} X_j > t\right] = \prod_{j\in\mathcal{A}} \mathbb{P}\left[X_j > x + t \,\middle|\, \bigcap_{j\in\mathcal{A}'} X_j > t\right] \qquad \text{(by independence)}$$

   $$\mathbb{P}[L_i > x] = \prod_{j\in\mathcal{A}} \mathbb{P}\left[X_j > x\right] \qquad \text{(by memoryless property)}$$

   $$\mathbb{P}[L_i > x] = \prod_{j\in\mathcal{A}} (1 - \mathbb{P}\left[X_j \le x\right]) = \prod_{j\in\mathcal{A}} 1 - F_{X_j}(x) = \prod_{j\in\mathcal{A}} 1 - (1 - e^{-\lambda x}) = \prod_{j\in\mathcal{A}} e^{-\lambda x} = e^{-\lambda(n-i)x}.$$

   To conclude, we have that

   $$\mathbb{P}[L_i > x] = 1 - \mathbb{P}[L_i \le x] = 1 - e^{-\lambda(n-i)x} \quad \Longrightarrow \quad \boxed{L_i \sim \text{Exp}((n-i)\lambda) \text{ for } i = 0, 1, \ldots, n-1}$$

(c) Use the previous item to show that each $X_{(i)}$ has the same distribution as a sum of i independent random variables.

To begin, we note that

$$\begin{aligned}
X_{(1)} &= L_0 && = X_{(1)} \\
X_{(2)} &= L_0 + L_1 && = X_{(1)} + \big(X_{(2)} - X_{(1)}\big) \\
X_{(3)} &= L_0 + L_1 + L_2 && = X_{(1)} + \big(X_{(2)} - X_{(1)}\big) + \big(X_{(3)} - X_{(2)}\big) \\
&\vdots \\
X_{(i)} &= \sum_{k=1}^{i} L_{k-1} && = X_{(1)} + \big(X_{(2)} - X_{(1)}\big) + \cdots + \big(X_{(i-1)} - X_{(i-2)}\big) + \big(X_{(i)} - X_{(i-1)}\big)
\end{aligned}$$

where L_k are independent events for all $k = 0, 1, \ldots, n-1$. Note their independence follows from our result in part (b). Since each $X_{(i+1)} - X_{(i)}$ depends only on the residual lifetimes (i.e., the X_j's in $\mathcal{A}$) and not on any earlier times, the spacings L_k are independent. Then we conclude that $X_{(i)}$ has the same distribution as a sum of i independent random variables,

$$\boxed{X_{(i)} = \sum_{k=1}^{i} L_{k-1} \quad \text{where} \quad L_{k-1} \sim \text{Exp}((n-k+1)\lambda) \text{ for } k = 1, \ldots, i.}$$

(d) Calculate the expectation and the variance of $X_{(i)}$ for $i = 1, \ldots, n$.

To begin we recall that the expectation and variance of $X \sim \text{Exp}(\lambda)$ is $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\text{Var}[X] = \frac{1}{\lambda^2}$. Then, $\mathbb{E}[L_{k-1}] = \frac{1}{\lambda(n-k+1)}$ and $\text{Var}[L_{k-1}] = \frac{1}{\lambda^2(n-k+1)^2}$. The expectation of $X_{(i)}$ is

$$\mathbb{E}[X_{(i)}] = \mathbb{E}\left[\sum_{k=1}^{i} L_{k-1}\right] = \sum_{k=1}^{i} \mathbb{E}[L_{k-1}] = \sum_{k=1}^{i} \frac{1}{\lambda(n-k+1)} \qquad \text{(by independence)}$$

$$\boxed{\mathbb{E}[X_{(i)}] = \frac{1}{\lambda} \sum_{k=1}^{i} \frac{1}{(n-k+1)} \quad \text{for } i = 1, \ldots, n,}$$

and the variance is

$$\text{Var}[X_{(i)}] = \text{Var}\left[\sum_{k=1}^{i} L_{k-1}\right] = \sum_{k=1}^{i} \text{Var}[L_{k-1}] = \sum_{k=1}^{i} \frac{1}{\lambda^2(n-k+1)^2} \qquad \text{(by independence)}$$

$$\boxed{\text{Var}[X_{(i)}] = \frac{1}{\lambda^2} \sum_{k=1}^{i} \frac{1}{(n-k+1)^2} \quad \text{for } i = 1, \ldots, n.}$$

2. (**Joint and conditional densities**) Let X, Y be two random variables with the following properties. Y has density function $f_Y(y) = 3y^2$ for $0 < y < 1$ and zero elsewhere. For $0 < y < 1$, given that $Y = y$, X has conditional density function $f_{X|Y}(x|y) = \frac{2x}{y^2}$ for $0 < x < y$ and zero elsewhere.

(a) Find the joint density function $f_{X,Y}(x, y)$ of X, Y. Be precise about the values (x, y) for which your formula is valid. Check that the joint density function you find integrates to 1.

To find $f_{X,Y}(x, y)$, we use $f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y)$. Then we have that

$$f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) = \frac{2x}{y^2} \cdot 3y^2 = 6x \quad \text{for } y \in (0, 1) \text{ and } x \in (0, y).$$

To check that this is a valid density function, we verify that it integrates to 1,

$$\int_0^1 \int_0^y 6x \, dx \, dy = \int_0^1 3y^2 \, dy = 1 \quad \checkmark$$

(b) Find the conditional density function of Y, given $X = x$. Be precise about the values of x and y for which the answer is valid. Identify the conditional distribution of Y by name.

To find the conditional density function of Y given $X = x$, we use:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

First, we need to calculate $f_X(x)$. We have that

$$f_X(x) = \int_0^1 f_{X,Y}(x,y)\,dy = \int_x^1 6x\,dy = 6x(1-x)$$

Then we have that

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{6x}{6x(1-x)} = \frac{1}{1-x} \quad \text{for } 0 < x < y < 1.$$

The conditional distribution of Y given $X = x$ is Uniform on the interval $(x, 1)$, i.e.,

$$\boxed{Y \mid X \sim \text{Uniform}(x,1) \quad \text{for } 0 < x < y < 1.}$$

3. (**Model selection**) Given data $x_1, \ldots, x_n$, consider the problem of selecting between the two models:

Model One : $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$

Model Two : $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$ for an unknown μ.

To use probability to solve this problem, let us introduce an additional random variable Θ that has the Bernoulli distribution with parameter 0.5. Assume that the conditional distribution of $X_1, \ldots, X_n$ given $\Theta = \theta$ is given by the following

$$X_1, \ldots, X_n | \Theta = 0 \overset{\text{i.i.d.}}{\sim} N(0,1)$$

and

$$X_1, \ldots, X_n | \mu, \Theta = 1 \overset{\text{i.i.d.}}{\sim} N(\mu, 1) \text{ and } \mu | \Theta = 1 \sim N(0, \tau^2).$$

Here τ is a parameter which you can treat as a fixed constant in this exercise.

(a) Using the formula

$$f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) = \int f_{X_1,\ldots,X_n|\mu,\Theta=1}(x_1,\ldots,x_n) f_{\mu|\Theta=1}(\mu) d\mu \tag{3}$$

prove that

$$f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{1+n\tau^2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2}\right) \exp\left(\frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}\right),$$

where $\overline{x}$ is the mean of $x_1, \ldots, x_n$.

Using the formula:

$$f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) = \int f_{X_1,\ldots,X_n|\mu,\Theta=1}(x_1,\ldots,x_n) f_{\mu|\Theta=1}(\mu)\,d\mu$$

Recall that if $Z \sim N(\mu, \sigma^2)$, then

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

By independence of $X_1, \ldots, X_n$, we have that

$$f_{X_1,\ldots,X_n|\mu,\Theta=1}(x_1,\ldots,x_n) = \prod_{i=1}^n f_{X_i|\mu,\Theta=1}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right)$$

Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Expanding the sum inside the exponent, we get:

$$-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{1}{2}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\mu + \mu^2\right) = -\frac{1}{2}\sum_{i=1}^{n}x_i^2 + \mu\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2} = -\frac{1}{2}\sum_{i=1}^{n}x_i^2 + \mu n\overline{x} - \frac{n\mu^2}{2}$$

Since $\mu \mid \Theta = 1 \sim N(0, \tau^2)$, we have:

$$f_{\mu|\Theta=1}(\mu) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\mu^2}{2\tau^2}\right)$$

Plugging our results into Equation (3), we have that

$$\begin{aligned}
f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) &= \int f_{X_1,\ldots,X_n|\mu,\Theta=1}(x_1,\ldots,x_n) f_{\mu|\Theta=1}(\mu)d\mu \\
f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) &= \int \left[\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2 + \mu n\overline{x} - \frac{n\mu^2}{2}\right)\right]\left[\frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\mu^2}{2\tau^2}\right)\right]d\mu \\
f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right)\underbrace{\int \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(\mu n\overline{x} - \frac{n\mu^2}{2} - \frac{\mu^2}{2\tau^2}\right)d\mu}_{(*)}
\end{aligned}$$

Now our goal is to rewrite the terms in the exponent in the form of $-\frac{(\mu-\alpha)^2}{2\beta^2}$ + constant for some α and β^2 that would effectively represent the mean and variance of a Gaussian. We have that

$$-\frac{n\mu^2}{2} + \mu n\overline{x} - \frac{\mu^2}{2\tau^2} = -\frac{1}{2}\left(n + \frac{1}{\tau^2}\right)\mu^2 + \mu n\overline{x} = -\frac{\mu^2}{2\tau^2}\left(1+n\tau^2\right) + \mu n\overline{x}. = -\frac{1}{2\tau^2}\left(1+n\tau^2\right)\left(\mu^2 - \frac{2\mu\cdot n\tau^2\overline{x}}{1+n\tau^2}\right).$$

Completing the square inside the parentheses gives

$$-\frac{n\mu^2}{2} + \mu n\overline{x} - \frac{\mu^2}{2\tau^2} = -\frac{1}{2\tau^2}\left(1+n\tau^2\right)\left(\mu - \frac{n\tau^2\overline{x}}{1+n\tau^2}\right)^2 + \frac{1}{2\tau^2}\frac{\left(n\tau^2\overline{x}\right)^2}{1+n\tau^2} = -\frac{1}{2\tau^2}\left(1+n\tau^2\right)\left(\mu - \frac{n\tau^2\overline{x}}{1+n\tau^2}\right)^2 + \frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}.$$

Thus, we identify α and β^2 as: $\alpha = \frac{n\tau^2\overline{x}}{1+n\tau^2}$ and $\beta^2 = \frac{\tau^2}{1+n\tau^2}$. With this in mind, we can rewrite our integral $(*)$ as

$$\begin{aligned}
\int \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(\mu n\overline{x} - \frac{n\mu^2}{2} - \frac{\mu^2}{2\tau^2}\right)d\mu &= \int \frac{1}{\sqrt{2\pi\tau^2}}\frac{\sqrt{2\pi\beta^2}}{\sqrt{2\pi\beta^2}}\exp\left(-\frac{(\mu-\alpha)^2}{2\beta^2} + \frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)d\mu \\
\int \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(\mu n\overline{x} - \frac{n\mu^2}{2} - \frac{\mu^2}{2\tau^2}\right)d\mu &= \left(\frac{\beta}{\tau}\right)\exp\left(\frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)\int \frac{1}{\sqrt{2\pi\beta^2}}\exp\left(-\frac{(\mu-\alpha)^2}{2\beta^2}\right)d\mu \quad \to 1 \\
\int \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(\mu n\overline{x} - \frac{n\mu^2}{2} - \frac{\mu^2}{2\tau^2}\right)d\mu &= \left(\frac{1}{\sqrt{1+n\tau^2}}\right)\exp\left(\frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)
\end{aligned}$$

Plugging this back in, we arrive at the desired result

$$\boxed{f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{1+n\tau^2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right)\exp\left(\frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)}$$

(b) Calculate the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$.

By Bayes rule, we have that

$$\begin{aligned}
f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta) &= \frac{f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n)\,\mathbb{P}[\Theta=\theta]}{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)} = \frac{f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n)\,\mathbb{P}[\Theta=\theta]}{f_{X_1,\ldots,X_n|\Theta=0}(x_0)\,\mathbb{P}[\Theta=0] + f_{X_1,\ldots,X_n|\Theta=1}(x_0)\,\mathbb{P}[\Theta=1]} \\
f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta) &= \frac{f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n)}{f_{X_1,\ldots,X_n|\Theta=0}(x_1,\ldots,x_n) + f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n)} \qquad (4)
\end{aligned}$$

where the last simplification come from the symmetry of Θ, i.e.,

$$\mathbb{P}[\Theta = \theta] = \begin{cases} \frac{1}{2} & \text{when } \Theta = 0 \\ \frac{1}{2} & \text{when } \Theta = 1 \end{cases}$$

In part (a) we showed that

$$f_{X_1,\dots,X_n|\Theta=1}(x_1,\dots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)\exp\left(\frac{n^2\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)\left(\frac{1}{\sqrt{1+n\tau^2}}\right)$$

So now we need to solve for $f_{X_1,\dots,X_n|\Theta=0}(x_1,\dots,x_n)$. We have that

$$f_{X_1,\dots,X_n|\Theta=0}(x_1,\dots,x_n) = \prod_{i=1}^n f_{X_i|\Theta=0}(x_i) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n x_i^2\right\}$$

Then the denominator in Equation (4) becomes

$$f_{X_1,\dots,X_n|\Theta=0}(x_1,\dots,x_n) + f_{X_1,\dots,X_n|\Theta=1}(x_1,\dots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n x_i^2\right\} \times \left(1 + \frac{1}{\sqrt{1+n\tau^2}}\exp\left(\frac{n\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)\right)$$

Thus,

$$f_{\Theta|X_1=x_1,\dots,X_n=x_n}(\theta) = \begin{cases} \frac{\sqrt{1+n\tau^2}}{\sqrt{1+n\tau^2}+\exp\left(\frac{n\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)} & \text{when } \Theta = 0 \\ \frac{\exp\left(\frac{n\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)}{\sqrt{1+n\tau^2}+\exp\left(\frac{n\tau^2\overline{x}^2}{2(1+n\tau^2)}\right)} & \text{when } \Theta = 1 \end{cases}$$

(c) Intuitively, we would prefer Model Two over Model One when $\overline{x}$ is far from zero. Is this intuition reflected in your conditional distribution from the previous part?
Yes this intuition reflected in your conditional distribution from the previous part, because $f_{\Theta=1|X_1=x_1,\dots,X_n=x_n}(\theta) \to 1$ as $\overline{x}^2$ goes away from zero. Meaning, the probability of using Model Two will increase $\overline{x}$ goes away from zero.

4. (**Gamma-Poisson**) Consider random variables $\Theta, X_1, \dots, X_n$ such that

$$\Theta \sim \text{Gamma}(\alpha, \lambda) \quad \text{and} \quad X_1,\dots,X_n|\Theta = \theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$$

Recall the following about the Gamma distribution

$$f_x(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} \quad \text{where} \quad \Gamma(r) = \begin{cases} (r-1)! & \text{when } r \in \mathbb{N}^+ \\ \int_0^\infty t^{r-1}e^{-t}\,dt & \text{holds for all } r > 0 \end{cases}$$

Three other useful properties of the Gamma function are

$$\int_0^\infty e^{-\alpha t}t^{\beta-1}\,dt = \frac{1}{\alpha^\beta}\Gamma(\beta), \qquad \Gamma(z+1) = z\Gamma(z), \quad \text{and} \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Additionally, recall that for $Y \sim \text{Poisson}(\lambda)$, we have

$$\mathbb{P}[Y = k] = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k \in \{0,1,2,\dots\}$$

(a) Find the conditional distribution of Θ given $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$.
To find $f_{\Theta|X_1,\dots,X_n}(\theta)$, we will use

$$f_{\Theta|X_1,\dots,X_n}(\theta) = \frac{\mathbb{P}[X_1 = x_1,\dots,X_n = x_n \mid \Theta = \theta]f_\Theta(\theta)}{f_{X_1,\dots,X_n}(x_1,\dots,x_n)}$$

We start with

$$\mathbb{P}[X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta] = e^{-n\theta} \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} = e^{-n\theta}\theta^{n\bar{x}} \prod_{i=1}^{n} \frac{1}{x_i!}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We know that the marginal distribution of Θ is $f_\Theta(\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}$. Then we have

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta] f_\Theta(\theta)\, d\theta = \prod_{i=1}^{n} \frac{1}{x_i!} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-(n+\lambda)\theta}\theta^{n\bar{x}+\alpha-1}\, d\theta$$

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{x_i!} \cdot \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(n\bar{x}+\alpha)}{(n+\lambda)^{n\bar{x}+\alpha}}$$

Plugging our results into $f_{\Theta|X_1,\ldots,X_n}(\theta)$, we have

$$f_{\Theta|X_1,\ldots,X_n}(\theta) = \frac{e^{-n\theta}\theta^{n\bar{x}}\theta^{\alpha-1}e^{-\lambda\theta}}{\frac{\Gamma(n\bar{x}+\alpha)}{(n+\lambda)^{n\bar{x}+\alpha}}} = \frac{(n+\lambda)^{n\bar{x}+\alpha}}{\Gamma(n\bar{x}+\alpha)}\theta^{n\bar{x}+\alpha-1}e^{-\theta(n+\lambda)} \quad \Longrightarrow \quad \boxed{f_{\Theta|X_1,\ldots,X_n}(\theta) \sim \text{Gamma}(\alpha + n\bar{x}, \lambda + n)}$$

(b) Find $\mathbb{E}[\Theta|X_1 = x_1, \ldots, X_n = x_n]$.

To find $E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$, we will use

$$E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \int_0^\infty \theta f_{\Theta|X_1=x_1,\ldots,X_n=x_n}(\theta)\, d\theta$$

Since $\Theta \mid X_1 = x_1, \ldots, X_n = x_n \sim \text{Gamma}(\alpha + n\bar{x}, \lambda + n)$, we have:

$$E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{(n+\lambda)^{n\bar{x}+\alpha}}{\Gamma(n\bar{x}+\alpha)} \int_0^\infty \theta \cdot \theta^{n\bar{x}+\alpha-1}e^{-\theta(n+\lambda)}\, d\theta = \frac{(n+\lambda)^{n\bar{x}+\alpha}}{\Gamma(n\bar{x}+\alpha)} \int_0^\infty \theta^{(n\bar{x}+\alpha+1)-1}e^{-\theta(n+\lambda)}\, d\theta$$

$$E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{(n+\lambda)^{n\bar{x}+\alpha}}{\Gamma(n\bar{x}+\alpha)} \cdot \frac{\Gamma(n\bar{x}+\alpha+1)}{(n+\lambda)^{n\bar{x}+\alpha+1}} = \frac{1}{n+\lambda} \cdot \frac{(n\bar{x}+\alpha)\Gamma(n\bar{x}+\alpha)}{\Gamma(n\bar{x}+\alpha)}$$

$$\boxed{E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{n\bar{x}+\alpha}{n+\lambda}}$$

(c) Write $\mathbb{E}[\Theta|X_1 = x_1, \ldots, X_n = x_n]$ as a weighted linear combination of $\left(\frac{x_1+\cdots+x_n}{n}\right)$ and the mean of the marginal distribution (i.e., prior mean) of Θ and argue that the weight of the prior mean goes to zero as $n \to \infty$.

To begin, we note that our derivation of $E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$ shows that the expectation of a random variable $X \sim \text{Gamma}(\alpha, \lambda)$ is $\mathbb{E}[X] = \frac{\alpha}{\lambda}$. Therefore, $E[\Theta] = \frac{\alpha}{\lambda}$.

Now we rewrite $E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$ as a weighted linear combination of $\bar{x}$ and the mean of the marginal distribution $E[\Theta]$. We have that

$$E[\Theta \mid X_1 = x_1, \ldots, X_n = x_n] = \frac{n\bar{x}+\alpha}{n+\lambda} = \left(\frac{n}{n+\lambda}\right)\bar{x} + \left(\frac{\lambda}{n+\lambda}\right)\frac{\alpha}{\lambda}$$

It is really clear that as $n \to \infty$, the weight on the prior mean also goes to zero, i.e., $\frac{\lambda}{\lambda+n} \to 0$.

5. **(Law of total expectation)**

Let the joint probability mass function (p.m.f.) of (X, Y) be

$$p_{X,Y}(k, n) = \begin{cases} \frac{1}{n+1}\left(1 - \frac{1}{n+1}\right)^{k-1}\frac{1}{2^n}, & \text{for } 1 \le n < \infty \text{ and } 1 \le k < \infty, \\ 0, & \text{else.} \end{cases}$$

Throughout this problem, we will leverage the convergence of a geometric series,

$$\sum_{n=0}^{\infty} \alpha r^n = \frac{\alpha}{1-r} \quad \text{and} \quad \sum_{n=0}^{\infty} \alpha n r^{n-1} = \frac{\alpha}{(1-r)^2} \quad \text{for } |r| < 1.$$

(a) Find the p.m.f. $p_Y(n)$ of Y and the conditional p.m.f $p_{X|Y}(k|n)$.

We obtain the marginal distribution $P_Y(n)$ by summing over k,

$$P_Y(n) = \sum_{k=1}^{\infty} p_{X,Y}(k,n) = \sum_{k=1}^{\infty} \left(\frac{1}{n+1}\right)\left(\frac{1}{2^n}\right)\left(1-\frac{1}{n+1}\right)^{k-1} = \left(\frac{1}{n+1}\right)\left(\frac{1}{2^n}\right)\sum_{k=1}^{\infty}\left(\frac{n}{n+1}\right)^{k-1}$$

$$P_Y(n) = \left(\frac{1}{n+1}\right)\left(\frac{1}{2^n}\right)\sum_{k=0}^{\infty}\left(1-\frac{1}{n+1}\right)^{k} = \left(\frac{1}{n+1}\right)\left(\frac{1}{2^n}\right)\cdot\frac{1}{1-\left(1-\frac{1}{n+1}\right)} = \frac{1}{2^n}$$

$$\boxed{P_Y(n) = \frac{1}{2^n}}$$

For the conditional distribution $P_{X|Y}(k \mid n)$, we have

$$P_{X|Y}(k \mid n) = \frac{P_{X,Y}(k,n)}{P_Y(n)} = \frac{\frac{1}{n+1}\frac{1}{2^n}\left(1-\frac{1}{n+1}\right)^{k-1}}{\frac{1}{2^n}} = \frac{1}{n+1}\left(1-\frac{1}{n+1}\right)^{k-1}$$

$$\boxed{P_{X|Y}(k \mid n) = \frac{1}{n+1}\left(1-\frac{1}{n+1}\right)^{k-1}}$$

(b) Calculate $\mathbb{E}[Y]$.

We simply sum over all n to obtain $\mathbb{E}[Y]$,

$$\mathbb{E}[Y] = \sum_{n=1}^{\infty}\frac{n}{2^n} = \sum_{n=1}^{\infty} n\left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty} n\left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty}\left(\frac{1}{2}\right) n\left(\frac{1}{2}\right)^{n-1} = \frac{\frac{1}{2}}{(1-\frac{1}{2})^2} = 2$$

Note that we can change the sum from 1 to ∞ to 0 to ∞ because the first term is zero. Our result is

$$\boxed{\mathbb{E}[Y] = 2.}$$

(c) Find the conditional expectation $\mathbb{E}[X|Y]$.

We simply sum over all k to obtain $\mathbb{E}[X \mid Y]$,

$$E[X \mid Y] = \sum_{k=1}^{\infty} k\,P(X=k \mid Y) = \sum_{k=1}^{\infty} k\cdot\frac{1}{Y+1}\cdot\left(1-\frac{1}{Y+1}\right)^{k-1} = \frac{1}{Y+1}\sum_{k=0}^{\infty} k\left(1-\frac{1}{Y+1}\right)^{k-1} = \frac{\frac{1}{Y+1}}{\left(1-(1-\frac{1}{Y+1})\right)^2}$$

$$E[X \mid Y] = Y+1$$

Again, we can change the sum from 1 to 0 because the first term is zero. We have

$$\boxed{E[X \mid Y] = Y+1}$$

(d) Use parts (a) and (c) to calculate $\mathbb{E}[X]$.

Use the law of total expectation, we have that

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[Y+1] = \mathbb{E}[Y] + 1 = 2 + 1 = 3 \quad \Longrightarrow \quad \boxed{\mathbb{E}[X] = 3}$$

6. (**Expected number of coin tosses**)

Consider a sequence of coin tosses.

(a) On average, how many tosses of a fair coin does it take to see two heads in a row?

Let N be the total number of tosses needed to get two heads in a row. Additionally, let

- $\mathbb{E}[N]$: the expected number of tosses to get two heads in a row.
- $\mathbb{E}[N \mid H_1]$: the expected number of tosses given that the first toss is a head.

Using the law of total expectation on the first toss:

$$\mathbb{E}[N] = 1 + \mathbb{P}(H_1)\,\mathbb{E}[N \mid H_1] + \mathbb{P}(\text{not } H_1)\,\mathbb{E}[N \mid \text{not } H_1]$$
$$\mathbb{E}[N] = 1 + \frac{1}{2}\,\mathbb{E}[N \mid H_1] + \frac{1}{2}\,\mathbb{E}[N]$$

as this is a fair coin and the number of tosses doesnt change after we don't get one, i.e., $\mathbb{E}[N] = \mathbb{E}[N \mid \text{not } H]$. Now, using the law of total expectation again for $\mathbb{E}[N \mid H_1]$:

$$\mathbb{E}[N \mid H_1] = 1 + \mathbb{P}(H_2 \mid H_1)\,\mathbb{E}[N \mid H_1 \cap H_2] + \mathbb{P}(\text{not } H_2 \mid H_1)\,\mathbb{E}[N \mid H_1 \cap \text{not } H_2]$$
$$\mathbb{E}[N \mid H_1] = 1 + \mathbb{P}(H_2)\,\mathbb{E}[N \mid H_1 \cap H_2] + \mathbb{P}(\text{not } H_2)\,\mathbb{E}[N \mid H_1 \cap \text{not } H_2]$$
$$\mathbb{E}[N \mid H_1] = 1 + \frac{1}{2}\,\mathbb{E}[N]$$

Now we note (1.) that $\mathbb{E}[N \mid H_1 \cap \text{not } H_2] = \mathbb{E}[N]$ as soon as we hit a tail, we reset, (2.) $\mathbb{E}[N \mid H_1 \cap H_2] = 0$ as we reached two heads and we are done, and (3.) $\mathbb{P}(H_2 \mid H_1) = \mathbb{P}(H_2)$ as the probability of the coin landing on heads is independent of the flip.
Then we have two equations and two unknowns and we can solve for $\mathbb{E}[N]$.

$$\mathbb{E}[N] = 1 + \frac{1}{2}\,\mathbb{E}[N \mid H_1] + \frac{1}{2}\,\mathbb{E}[N] \quad \text{and} \quad \mathbb{E}[N \mid H_1] = 1 + \frac{1}{2}\,\mathbb{E}[N]$$
$$\mathbb{E}[N] = 1 + \frac{1}{2}\,\mathbb{E}[N] + \frac{1}{2}\left(1 + \frac{1}{2}\,\mathbb{E}[N]\right)$$
$$\frac{1}{4}\,\mathbb{E}[N] = \frac{3}{2}$$
$$\boxed{\mathbb{E}[N] = 6}$$

(b) How many tosses on average to see the sequence HTH for the first time?

We learned from part (a.) that the flips are independent so we do not need to worry about conditional probabilities. We also saw that $E[N] = \mathbb{E}[N \mid \text{not } H_1] = \mathbb{E}[N \mid H_1 \cap H_2]$, so as soon as we deviate from the desired path, we reset.
Then for clarity, we define

- E_0 be the expected number of tosses to reach HTH (the reset state and the expectation we want to solve for).
- E_H be the expected number of tosses to reach HTH after seeing H.
- E_{HT} be the expected number of tosses to see HTH after seeing HT.
- E_{HTH} be the expected number of tosses to see HTH after seeing HTH.

Then, we have that

$$E_0 = 1 + \frac{1}{2}E_H + \frac{1}{2}E_0$$
$$E_H = 1 + \frac{1}{2}E_{HT} + \frac{1}{2}E_H$$
$$E_{HT} = 1 + \frac{1}{2}E_{HTH} + \frac{1}{2}E_0$$
$$E_{HTH} = 0$$

Now we solve for E_0,

$$\frac{1}{2}E_0 = 1 + \frac{1}{2}E_H$$
$$-\frac{1}{2}E_{HT} = 1 - \frac{1}{2}E_H$$
$$\frac{1}{2}E_{HT} - \frac{1}{4}E_0 = \frac{1}{2}$$
$$\frac{1}{4}E_0 = \frac{5}{2}$$
$$\boxed{E_0 = 10}$$

(c) How does our answer change if we have an unfair coin?

Let p be the probability of landing on heads.

- **Seeing 2H**:

$$\begin{aligned} E_0 &= 1 + pE_H + (1-p)E_0 \\ E_H &= 1 + (1-p)E_0 \end{aligned} \implies \begin{aligned} E_0 &= 1 + pE_H + (1-p)E_0 \\ 0 &= p - pE_H + p(1-p)E_0 \end{aligned}$$

$$\begin{aligned} E_0 &= 1 + p + (1-p)E_0 + p(1-p)E_0 \\ pE_0 &= 1 + p + p(1-p)E_0 \\ [p - p(1-p)]E_0 &= 1 + p \\ p^2 E_0 &= 1 + p \\ E_0 &= \frac{1+p}{p^2} \end{aligned}$$

As p increases the expected number of flips decreases which makes sense.

- **Seeing HTH**:

$$\begin{aligned} E_0 &= 1 + pE_H + (1-p)E_0 \\ E_H &= 1 + pE_H + (1-p)E_{HT} \\ E_{HT} &= 1 + (1-p)E_0 \\ (1-p)E_H &= 1 + (1-p)\big[1 + (1-p)E_0\big] \\ E_H &= \frac{1}{1-p} + 1 + (1-p)E_0 \\ pE_0 &= 1 + p\left[\frac{1}{1-p} + 1 + (1-p)E_0\right] \\ E_0 &= \frac{1}{p} + \frac{1}{1-p} + 1 + (1-p)E_0 \\ pE_0 &= \frac{1}{p} + \frac{1}{1-p} + 1 \\ E_0 &= \frac{1}{p^2} + \frac{1}{p(1-p)} + \frac{1}{p} = \frac{1-p}{p^2(1-p)} + \frac{p}{p^2(1-p)} + \frac{p(1-p)}{p^2(1-p)} = \frac{1+p-p^2}{p^2(1-p)} \\ E_0 &= \frac{1}{p^2} + \frac{1}{p(1-p)} + \frac{1}{p} \end{aligned}$$

We calculate the p that results in the fewest number of flips on average by taking the derivative.

$$\begin{aligned} \frac{dE_0}{dp} = \frac{d}{dp}\left(\frac{1+p-p^2}{p^2(1-p)}\right) = \frac{(1-2p)[p^2(1-p)] - (2p-3p^2)[1+p-p^2]}{p^4(1-p)^2} &= 0 \\ (1-2p)[p^2(1-p)] - (2p-3p^2)[1+p-p^2] &= 0 \\ (1-2p)[p(1-p)] - (2-3p)[1+p-p^2] &= 0 \\ (1-2p)[p(1-p)] &= (2-3p)[1+p-p^2] \\ [p-p^2] - 2p[p-p^2] &= 2[1+p-p^2] - 3p[1+p-p^2] \\ p - p^2 - 2p^2 + 2p^3 &= 2 + 2p - 2p^2 - 3p - 3p^2 + 3p^3 \\ 2p + 2p^2 - p^3 &= 2 \\ p[2 + 2p - p^2] &= 2 \end{aligned}$$

This leads to a $p^\star = 0.688$. So the expectation will decrease as it approaches 0.688 and then increase as it increases.

STAT201A: Introduction to Probability at an Advanced Level November 20th, 2024

Homework # 5: Multivariate Normal and Gaussian Process

Reece D. Huff

Problems (Solutions)

1. **(Multivariate normal)** Suppose $Y \sim \mathcal{N}_n(\mu, \Sigma)$ in this problem.

   (a) If a is any fixed vector in $\mathbb{R}^n$, show that

   $$\frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}} \sim \mathcal{N}(0,1).$$

   We begin by taking the expectation of the random variable $Z = \frac{a^\top(Y-\mu)}{\sqrt{a^\top \Sigma a}}$,

   $$\mathbb{E}[Z] = \mathbb{E}\left[\frac{a^\top(Y-\mu)}{\sqrt{a^\top \Sigma a}}\right] = \frac{1}{\sqrt{a^\top \Sigma a}}\Big(\mathbb{E}[a^\top(Y-\mu)]\Big) = \frac{1}{\sqrt{a^\top \Sigma a}} a^\top \Big(\mathbb{E}[Y-\mu]\Big) = \frac{1}{\sqrt{a^\top \Sigma a}} a^\top \cancel{\Big(\mathbb{E}[Y]-\mu\Big)} = 0 \quad \checkmark$$

   where we can pull out the constants $a^\top$ and $\frac{1}{\sqrt{a^\top \Sigma a}}$ by the linearity of expectation and note that $\mathbb{E}[Y] = \mu$. Next, we compute the variance of Z,

   $$\mathrm{Var}[Z] = \mathrm{Var}\left[\frac{a^\top(Y-\mu)}{\sqrt{a^\top \Sigma a}}\right] = \frac{1}{a^\top \Sigma a}\mathrm{Var}[a^\top(Y-\mu)] = \frac{1}{a^\top \Sigma a}\mathrm{Var}[a^\top Y - a^\top \mu] = \frac{1}{a^\top \Sigma a}\Big(\mathrm{Var}[a^\top Y] + \cancel{\mathrm{Var}[a^\top \mu]}\Big)$$
   $$\mathrm{Var}[Z] = \frac{1}{a^\top \Sigma a} a^\top \mathrm{Var}[Y] a = \frac{1}{a^\top \Sigma a} a^\top \Sigma a = 1 \quad \checkmark$$

   where we used that for any random vector $\vec{X}$, constant vector $\vec{\alpha}$, and scalar β, we have

   $$\mathrm{Var}[\vec{\alpha}^\top \vec{X}] = \vec{\alpha}^\top \mathrm{Var}[\vec{X}]\vec{\alpha} \quad \text{and} \quad \mathrm{Var}[\beta \vec{X}] = \beta^2 \mathrm{Var}[\vec{X}].$$

   Therefore, we have shown that Z is distributed according to $\mathcal{N}(0,1)$.

   (b) If A is now a random vector that is independent of Y, then show again that

   $$\frac{A^\top(Y-\mu)}{\sqrt{A^\top \Sigma A}}$$

   is distributed according to $\mathcal{N}(0,1)$ and that it is independent of A.

   Similar to part (a), we can compute the expectation of the random variable $Z = \frac{A^\top(Y-\mu)}{\sqrt{A^\top \Sigma A}}$,

   $$\mathbb{E}[Z] = \mathbb{E}\left[\frac{A^\top(Y-\mu)}{\sqrt{A^\top \Sigma A}}\right] = \frac{1}{\sqrt{A^\top \Sigma A}}\Big(\mathbb{E}[A^\top(Y-\mu)]\Big) = \frac{1}{\sqrt{A^\top \Sigma A}} A^\top \Big(\mathbb{E}[Y-\mu]\Big) = \frac{1}{\sqrt{A^\top \Sigma A}} A^\top \cancel{\Big(\mathbb{E}[Y]-\mu\Big)} = 0 \quad \checkmark$$

   and the variance of Z, we have that

   $$\mathrm{Var}[Z] = \mathrm{Var}\left[\frac{A^\top(Y-\mu)}{\sqrt{A^\top \Sigma A}}\right] = \frac{1}{A^\top \Sigma A}\mathrm{Var}[A^\top(Y-\mu)] = \frac{1}{A^\top \Sigma A}\mathrm{Var}[A^\top Y - A^\top \mu] = \frac{1}{A^\top \Sigma A}\Big(\mathrm{Var}[A^\top Y] + \cancel{\mathrm{Var}[A^\top \mu]}\Big)$$
   $$\mathrm{Var}[Z] = \frac{1}{A^\top \Sigma A} A^\top \mathrm{Var}[Y] A = \frac{1}{A^\top \Sigma A} A^\top \Sigma A = 1 \quad \checkmark$$

   and we can conclude that $Z \sim \mathcal{N}(0,1)$.

   To show that $Z \perp\!\!\!\perp A$, we will use our result from part (a) that $\frac{a^\top(Y-\mu)}{\sqrt{a^\top \Sigma a}} \sim \mathcal{N}(0,1)$. We have that

   $$(Z \mid A = a) = \frac{a^\top(Y-\mu)}{\sqrt{a^\top \Sigma a}} \sim \mathcal{N}(0,1).$$

   In words, $Z|A = a$ follows the same distribution as Z. Therefore, $Z \perp\!\!\!\perp A$.

(c) Using the above result, show that if $Y \sim \mathcal{N}_3(0, I_3)$, then

$$\frac{Y_1 e^{Y_3} + Y_2 \log|Y_3|}{\sqrt{e^{2Y_3} + (\log|Y_3|)^2}} \sim \mathcal{N}(0,1).$$

To begin, we note that in the case of a Multivariate Normal distribution, $\Sigma_{ij} = 0 \Leftrightarrow Y_i \perp\!\!\!\perp Y_j$ for all $i \neq j$. Therefore, we have that $Y_1 \perp\!\!\!\perp Y_2 \perp\!\!\!\perp Y_3$.
Next, we notice that the numerator $Y_1 e^{Y_3} + Y_2 \log|Y_3|$ is a linear combination of the random vector $Y' = [Y_1, Y_2]^\top$. Therefore, we construct the random vector $A' = [e^{Y_3}, \log|Y_3|]^\top$ and note that the random vector A' is independent of Y' (since $Y_1 \perp\!\!\!\perp Y_2 \perp\!\!\!\perp Y_3$). We can now use the result from part (a) to show that $Z = \frac{A^\top(Y-\mu)}{\sqrt{A^\top \Sigma A}} \sim \mathcal{N}(0,1)$. We have that

$$Z = \frac{A'^\top(Y' - \mu')}{\sqrt{A'^\top \Sigma' A'}} = \frac{\begin{bmatrix} e^{Y_3} & \log|Y_3| \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} - \begin{bmatrix} e^{Y_3} & \log|Y_3| \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}}{\sqrt{\begin{bmatrix} e^{Y_3} & \log|Y_3| \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} e^{Y_3} \\ \log|Y_3| \end{bmatrix}}} = \frac{Y_1 e^{Y_3} + Y_2 \log|Y_3|}{\sqrt{e^{2Y_3} + (\log|Y_3|)^2}}$$

where $\mu' = \mathbb{E}[Y'] = [0,0]^\top$ and $\Sigma' = \mathrm{Var}[Y'] = I_2$. Therefore, we have that $Z \sim \mathcal{N}(0,1)$.

2. **(Marginally normal but not bivariate normal)** Give an example of a 2×1 random vector $Y = (Y_1, Y_2)^\top$ with a positive definite covariance matrix such that each Y_1 and Y_2 is standard normal but Y is not bivariate normal.

To begin, let $Y_1 \sim \mathcal{N}(0,1)$ and $Y_2 = WY_1$ where W is a Rademacher random variable that is independent of Y_1 and has the following distribution $\mathbb{P}(W = k) = \begin{cases} 1/2, & \text{if } k = 1, \\ 1/2, & \text{if } k = -1. \end{cases}$ with $\mathbb{E}[W] = 0$ and $\mathrm{Var}[W] = 1$.

By our construction, clearly Y_1 is not independent of Y_2 as Y_2 is a function of Y_1.

Next, we show that the marginal distributions of Y_1 and Y_2 are standard normal. Y_1 is standard normal by construction. Y_2 is also standard normal as

$$\mathbb{P}[Y_2 \leq y] = \mathbb{P}[WY_1 \leq y] = \mathbb{P}[WY_1 \leq y | W = 1]\,\mathbb{P}[W = 1] + \mathbb{P}[WY_1 \leq y | W = -1]\,\mathbb{P}[W = -1]$$

$$\mathbb{P}[Y_2 \leq y] = \frac{1}{2}\left(\mathbb{P}[Y_1 \leq y] + \mathbb{P}[-Y_1 \leq y]\right) = \mathbb{P}[Y_1 \leq y]$$

where the last equality follows from the symmetry of the standard normal distribution (i.e., if $Y_1 \sim \mathcal{N}(0,1)$ then $-Y_1 \sim \mathcal{N}(0,1)$).

Next we calculate the covariance matrix of Y. We have that

$$\Sigma = \mathrm{Var}[Y] = \mathrm{Cov}[Y, Y] = \mathbb{E}[YY^\top] - \mathbb{E}[Y]\,\mathbb{E}[Y]^\top = \begin{pmatrix} \mathrm{Var}[Y_1] & \mathrm{Cov}[Y_1, Y_2] \\ \mathrm{Cov}[Y_2, Y_1] & \mathrm{Var}[Y_2] \end{pmatrix}$$

The covariance between Y_1 and Y_2 is

$$\mathrm{Cov}[Y_1, Y_2] = \mathrm{Cov}[Y_2, Y_1] = \mathbb{E}[Y_1 Y_2] - \mathbb{E}[Y_1]\,\mathbb{E}[Y_2] = \mathbb{E}[Y_1 W Y_1] - \cancel{\mathbb{E}[Y_1]}\,\mathbb{E}[WY_1] = \mathbb{E}[Y_1^2 W] - 0 = \mathbb{E}[Y_1^2]\cancel{\mathbb{E}[W]} = 0$$

Thus, the covariance matrix of Y is

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Finally, we show that Y is not bivariate normal. To do this, recall that two random variables X and Y are said to be **bivariate normal** (or two **jointly normal** random variables), if $aX + bY$ has a normal distribution for all $a, b \in \mathbb{R}$.

In our case, we have that Y_1 and Y_2 are standard normal, but $Z = Y_1 + Y_2$ is not standard normal. To see this, note that

$$Z = Y_1 + Y_2 = \begin{cases} 2Y_1, & \text{with probability } 1/2, \\ 0, & \text{with probability } 1/2. \end{cases}$$

In other words, $\mathbb{P}[Z = 0] = 1/2$ which is not the characteristic of a standard normal random variable. Therefore, Y is not bivariate normal.

3. **(Conditional distribution)** Consider three random variables Y_1, Y_2, and Y_3 that are independent and standard normal. Let

$$X_1 = Y_2 + Y_3,$$
$$X_2 = Y_1 + Y_3,$$
$$X_3 = Y_1 + Y_2.$$

Find the conditional distribution of X_1 given $X_2 = X_3 = 0$.

To begin, recall that we can partition the multivariable normal distribution $X \sim \mathcal{N}_n(\mu, \Sigma)$ into $X_a \in \mathbb{R}^k$ and $X_b \in \mathbb{R}^{n-k}$, such that $X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$, where μ_a and μ_b are the means of X_a and X_b, respectively, and Σ_{aa}, Σ_{ab}, Σ_{ba}, and Σ_{bb} are the corresponding covariance matrices. We that the marginal distributions of X_a and X_b are given by

$$X_a \sim \mathcal{N}_k(\mu_a, \Sigma_{aa}), \quad \text{and} \quad X_b \sim \mathcal{N}_{n-k}(\mu_b, \Sigma_{bb}).$$

and the conditional distribution of X_a given X_b is given by

$$X_a \mid X_b = x_b \sim \mathcal{N}_k(\mu_{a|b}, \Sigma_{a|b}), \quad \text{where} \quad \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \quad \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}. \tag{5}$$

Now, we can apply this result to the given problem. Let us calculate the expectation and covariance matrix of $X = [X_1, X_2, X_3]^\top$. We have that

$$\mathbb{E}[X_i] = \mathbb{E}[Y_j] + \mathbb{E}[Y_k] = 0, \quad \operatorname{Cov}[X_i, X_i] = \operatorname{Var}[X_i] = \operatorname{Var}[Y_j] + \operatorname{Var}[Y_k] = 2 \quad \text{and}$$
$$\operatorname{Cov}[X_i, X_j] = \operatorname{Cov}[Y_j + Y_k, Y_i + Y_k] = \cancel{\operatorname{Cov}[Y_j, Y_i]} + \cancel{\operatorname{Cov}[Y_j, Y_k]} + \cancel{\operatorname{Cov}[Y_k, Y_i]} + \operatorname{Cov}[Y_k, Y_k] = \operatorname{Var}[Y_k] = 1$$

for all $i, j, k \in \{1, 2, 3\}$ such that $i \neq j \neq k$. Thus we have

$$X = \begin{pmatrix} X_a = \begin{bmatrix} X_1 \end{bmatrix} \\ X_b = \begin{bmatrix} X_2 \\ X_3 \end{bmatrix} \end{pmatrix}, \quad \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \mu_a = \begin{bmatrix} 0 \end{bmatrix} \\ \mu_b = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} = \begin{pmatrix} \Sigma_{aa} = [2] & \Sigma_{ab} = \begin{bmatrix} 1 & 1 \end{bmatrix} \\ \Sigma_{ba} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} & \Sigma_{bb} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 2 \end{bmatrix} \end{pmatrix},$$

Noting that $\Sigma_{bb}^{-1} = \dfrac{1}{\det(\Sigma_{bb})} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = \dfrac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$, we use Equation (5) to show

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) = [0] + \left(\frac{1}{3}\right) \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \left(\begin{bmatrix} X_2 \\ X_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) = \left(\frac{1}{3}\right) \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} X_2 \\ X_3 \end{bmatrix} = \frac{1}{3}\Big(X_2 + X_3\Big)$$

and

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} = [2] - \left(\frac{1}{3}\right) \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 2 - \left(\frac{1}{3}\right) \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 2 - \left(\frac{1}{3}\right)\Big(1 + 1\Big) = 2 - \frac{2}{3} = \frac{4}{3}$$

Therefore,

$$\boxed{(X_1 \mid X_2 = 0, X_3 = 0) \sim \mathcal{N}_1\left(0, \frac{4}{3}\right)}$$

4. **(More on jointly Gaussian distributions)** Let X and Y be independent standard normal variables.

(a) For a constant k, find $\mathbb{P}[X > kY]$.

Let $Z_a = X - kY$. Then we have

$$\mathbb{E}[Z_a] = \mathbb{E}[X - kY] = \mathbb{E}[X] - k\,\mathbb{E}[Y] = 0 \quad \text{and} \quad \operatorname{Var}[Z_a] = \operatorname{Var}[X - kY] = \operatorname{Var}[X] + k^2 \operatorname{Var}[Y] = 1 + k^2$$

Therefore, $Z_a \sim \mathcal{N}(0, 1 + k^2)$ and

$$\mathbb{P}[X > kY] = \mathbb{P}[X - kY > 0] = \mathbb{P}[Z_a > 0] \quad \Longrightarrow \quad \boxed{\mathbb{P}[X > kY] = \frac{1}{2}}$$

(b) If $U = \sqrt{3}X + Y$ and $V = X - \sqrt{3}Y$, find $\mathbb{P}[U > kV]$. First, we note that

$$\mathbb{E}[U] = \mathbb{E}[\sqrt{3}X + Y] = \sqrt{3}\,\mathbb{E}[X] + \mathbb{E}[Y] = 0 \quad \text{and} \quad \mathbb{E}[V] = \mathbb{E}[X - \sqrt{3}Y] = \mathbb{E}[X] - \sqrt{3}\,\mathbb{E}[Y] = 0$$

and

$$\mathrm{Var}[U] = \mathrm{Var}[\sqrt{3}X + Y] = 3\,\mathrm{Var}[X] + \mathrm{Var}[Y] = 4, \quad \mathrm{Var}[V] = \mathrm{Var}[X - \sqrt{3}Y] = \mathrm{Var}[X] + 3\,\mathrm{Var}[Y] = 4 \quad \text{and}$$
$$\mathrm{Cov}[U, V] = \mathrm{Cov}[\sqrt{3}X + Y, X - \sqrt{3}Y] = \sqrt{3}\,\mathrm{Var}[X] - \sqrt{3}\,\mathrm{Var}[Y] = 0$$

Let $Z_b = U - kV$. Then we have

$$\mathbb{E}[Z_b] = \mathbb{E}[U - kV] = \mathbb{E}[U] - k\,\mathbb{E}[V] = 0 \quad \text{and} \quad \mathrm{Var}[Z_b] = \mathrm{Var}[U - kV] = \mathrm{Var}[U] + k^2\,\mathrm{Var}[V] = 4(1 + k^2)$$

Therefore, $Z_b \sim \mathcal{N}\big(0, 4(1 + k^2)\big)$ and

$$\mathbb{P}[U > kV] = \mathbb{P}[U - kV > 0] = \mathbb{P}[Z_b > 0] \quad \Longrightarrow \quad \boxed{\mathbb{P}[U > kV] = \frac{1}{2}}$$

(c) Find $\mathbb{P}[U^2 + V^2 < 1]$.

For this problem, we will use polar coordinates. Let $R^2 = X^2 + Y^2$ such that

$$U^2 + V^2 = (\sqrt{3}X + Y)^2 + (X - \sqrt{3}Y)^2 = 3X^2 + 2\sqrt{3}XY + Y^2 + X^2 - 2\sqrt{3}XY + 3Y^2 = 4(X^2 + Y^2) = 4R^2.$$

We also let $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$ for a $\Theta \sim \mathrm{Uniform}[0, 2\pi]$. It then follows that

$$\mathbb{P}[U^2 + V^2 < 1] = \mathbb{P}[4(X^2 + Y^2) < 1] = \mathbb{P}[4R^2 < 1] = \mathbb{P}\left[R^2 < \frac{1}{4}\right] = \mathbb{P}\left[R < \frac{1}{2}\right]$$

$$\mathbb{P}[U^2 + V^2 < 1] = \int_0^{2\pi} \int_0^{\frac{1}{2}} \frac{1}{2\pi} \exp\left\{\frac{-r^2}{2}\right\} r\,dr\,d\theta$$

Let $u = \frac{r^2}{2}$ and $du = rdr$. We change the bound of $r = \frac{1}{2}$ to $u = \frac{1}{8}$. Then we have

$$\mathbb{P}[U^2 + V^2 < 1] = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\frac{1}{8}} e^{-u}\,du\,d\theta = \int_0^{\frac{1}{8}} e^{-u}\,du = -\Big[e^{-u}\Big]_0^{\frac{1}{8}} = 1 - e^{-\frac{1}{8}}$$

$$\boxed{\mathbb{P}[U^2 + V^2 < 1] = 1 - e^{-\frac{1}{8}}}$$

(d) Find the conditional distribution of X given $V = v$.

Let us consider the following: Let $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and X and V are jointly normal random variables,

$$\begin{bmatrix} X \\ V \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -\sqrt{3} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Then the two dimensional version of Equation (5) is that $\big(X \mid V = v\big) \sim \mathcal{N}_1\big(\mathbb{E}[X \mid V], \mathrm{Var}[X \mid V]\big)$ where

$$\mathbb{E}[X \mid V] = \mathbb{E}[X] + \frac{\mathrm{Cov}[X, V]}{\mathrm{Var}[V]}(V - \mathbb{E}[V]), \quad \mathrm{Var}[X \mid V] = \mathrm{Var}[X] - \frac{(\mathrm{Cov}[X, V])^2}{\mathrm{Var}[V]}. \tag{6}$$

We have that $\mathbb{E}[X] = \mathbb{E}[V] = 0$, $\mathrm{Var}[X] = 1$, and $\mathrm{Var}[V] = 4$ (see part (b) for the calculation of $\mathrm{Var}[V]$) We calculate $\mathrm{Cov}[X, V]$

$$\mathrm{Cov}[X, V] = \mathrm{Cov}[X, X - \sqrt{3}Y] = \mathrm{Var}[X] - \sqrt{3}\cancel{\mathrm{Cov}[X, Y]} = 1$$

We calculate $\mathbb{E}[X \mid V]$ and $\mathrm{Var}[X \mid V]$,

$$\mathbb{E}[X \mid V = v] = \cancel{\mathbb{E}[X]} + \frac{\cancel{\mathrm{Cov}[X, V]}^{\,1}}{\cancel{\mathrm{Var}[V]}^{\,4}}(v - \cancel{\mathbb{E}[V]}) = \frac{v}{4} \quad \text{and} \quad \mathrm{Var}[X \mid V] = \cancel{\mathrm{Var}[X]}^{\,1} - \frac{1\,\cancel{(\mathrm{Cov}[X, V])^2}^{\,1}}{\cancel{\mathrm{Var}[V]}^{\,4}} = \frac{3}{4}.$$

Thus, the conditional distribution is given by

$$\boxed{\Big(X \mid V = v\Big) \sim \mathcal{N}_1\left(\frac{v}{4}, \frac{3}{4}\right)}$$

5. **(Wigner's surmise)** Let $X = \begin{pmatrix} X_1 & X_3 \\ X_3 & X_2 \end{pmatrix}$ with X_1 and X_2 independent $\mathcal{N}(0,1)$ and X_3 another independent $\mathcal{N}(0,1/2)$. Let λ_1 and λ_2 be two eigenvalues of X and $s = |\lambda_1 - \lambda_2|$.

(a) Prove that $s = \sqrt{(X_1 - X_2)^2 + 4X_3^2}$.

To begin we determine the eigenvalues of X. We have that

$$\det(X - \lambda I_2) = \det\begin{pmatrix} X_1 - \lambda & X_3 \\ X_3 & X_2 - \lambda \end{pmatrix} = (X_1 - \lambda)(X_2 - \lambda) - X_3^2 = \lambda^2 - \lambda(X_1 + X_2) + (X_1X_2 - X_3^2) = 0.$$

Setting the characteristic equation to zero and solving for λ, we have

$$\lambda = \frac{(X_1 + X_2) \pm \sqrt{(X_1 + X_2)^2 - 4(X_1X_2 - X_3^2)}}{2} = \frac{(X_1 + X_2) \pm \sqrt{(X_1 - X_2)^2 + X_3^2}}{2}.$$

Therefore, we have the s is

$$s = |\lambda_1 - \lambda_2| = \left| \frac{(X_1 + X_2) + \sqrt{(X_1 - X_2)^2 + X_3^2}}{2} - \frac{(X_1 + X_2) - \sqrt{(X_1 - X_2)^2 + X_3^2}}{2} \right| = \sqrt{(X_1 - X_2)^2 + X_3^2} \quad \checkmark$$

(b) Find the density of s.

To determine the density of s, we let $Z_1 = \frac{X_1 - X_2}{\sqrt{2}}$ and $Z_2 = \sqrt{2}X_3$ such that $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$[2] . Then we have

$$s = \sqrt{(X_1 - X_2)^2 + X_3^2} = \sqrt{2Z_1^2 + 2Z_2^2} = \sqrt{2(Z_1^2 + Z_2^2)}.$$

We showed in class that the sum of k sqaured independent, standard normal random variables is distributed according to the chi-squared distribution with k degrees of freedom χ_k^2. Let $W = Z_1^2 + Z_2^2$. Then $W \sim \chi_2^2$ with a pdf of

$$f_W(w) = \frac{1}{2}e^{-w/2} \quad \text{where} \quad W = Z_1^2 + Z_2^2 \sim \chi_2^2.$$

Now we apply change of variable to determine the distribution of s. Let $S = \sqrt{2W}$ (i.e., $W = \frac{S^2}{2}$). Then we have

$$f_S(s) = f_W\left(\frac{s^2}{2}\right)\left|\frac{dW}{dS}\right| = \frac{1}{2}e^{-\frac{s^2}{4}}|s| = \frac{s}{2}e^{-\frac{s^2}{4}} \quad \text{for all } s \geq 0$$

where $s \sim \text{Rayleigh}(\sqrt{2})$.

(c) Plot the density function of s. What do you observe regarding the eigenvalues of the random matrix X?

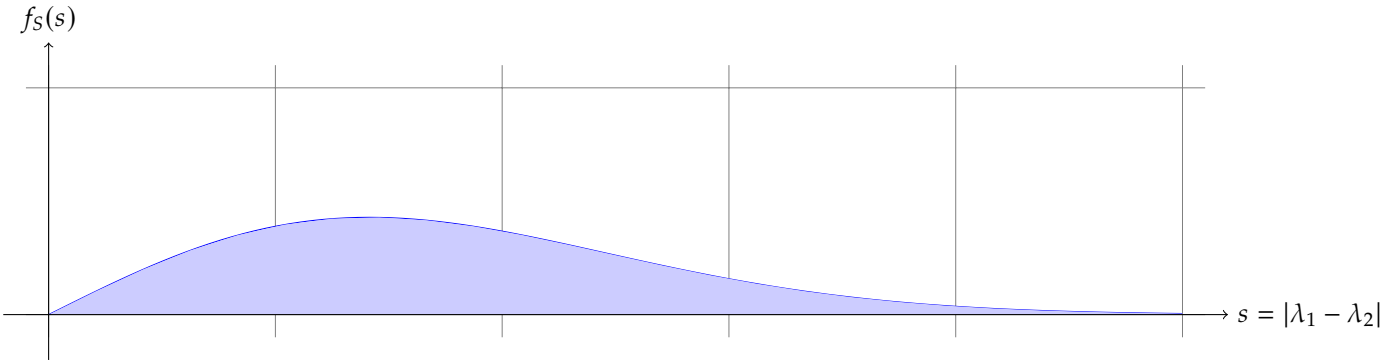


Figure 6: Plot of $f_S(s)$ for $0 \leq s \leq 5$.

[2]We can easily show that $\mathbb{E}(Z_1) = \mathbb{E}(Z_2) = 0$, $\text{Var}(Z_1) = \frac{1}{2}\text{Var}(X_1) + \frac{1}{2}\text{Var}(X_2) = 1$ and $\text{Var}(Z_2) = \text{Var}(\sqrt{2}X_3) = 2\,\text{Var}(X_3) = 1$. The independence of Z_1 and Z_2 follows from the independence of X_1, X_2, and X_3.

Comments:

- The density function $f_S(s) = \frac{s}{2}e^{-s^2/4}$ is zero at $s = 0$, indicating that the eigenvalues are almost surely distinct.
- The eigenvalue spacings is non-uniform and follow a Rayleigh distribution that peaks at approximately $s = 1.4$.
- The eigenvalues of the random matrix X are typically spread out rather than clustered together (e.g., $\mathbb{P}[s \leq 1] \leq \mathbb{P}[s \geq 1]$).

6. **(1D Gaussian process)** In this problem, you will implement a 1D Gaussian process that predicts outputs based on noisy training data. You will be given (noisy) 1D training data pairs $D_{\text{train}} = \{(x_1, y_1), (x_2, y_2), \dots\}$. Your task is to predict the output for a set of test queries $D_{\text{test}} = \{x_1^*, x_2^*, \dots\}$, conditioned on the training data. Implement two separate kernel functions, namely the

   - **Squared Exponential Kernel:** This is the kernel we discussed in class.

   $$k(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{(x_i - x_j)^T M (x_i - x_j)}{2}\right)$$

   where σ_f is a scale factor for the kernel and M is a metric measuring distance between two input vectors. In the 1D case, $M = \frac{1}{l^2}$ where l is the length scale of the kernel.
   - **Matérn Kernel:** This kernel is used commonly in many machine learning applications.

   $$k(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right)$$

   where ν and l are (positive) parameters of the kernel and $r = |x_i - x_j|$. K_ν is a modified Bessel function and Γ is the gamma function. Good parameter settings for ν are 0.25 - 3. You can use `scipy.special.kv()` in `Python` or `besselK()` in `R` for implementing K_ν.

   (a) Implement the squared exponential and Matérn kernel functions to compute similarity between any pair of inputs. The output for each function should be a kernel matrix K.

   (b) Using your kernel functions, implement a Gaussian process regression function to predict the posterior mean and variance of test data $\vec{y}^*$.

   (c) The simulation function and plotting function are provided in the file `ps5_GP_1D.ipynb`. Vary the kernel parameters (e.g., σ_f, l, and ν) and observe how they affect the predictive mean and variance. What impact do these parameters have on the smoothness and uncertainty of your GP predictions?

   Note: It's recommended to use `Python` (Jupyter notebook) and submit a PDF file including code, plots, and comments. If you prefer using another coding language, please ensure the data simulation is consistent with the provided code.

   - **Squared Exponential Kernel:**
     - **Varying σ_f:**
       - **Predictive Mean:** σ_f doesn't really affect the predicted mean.
       - **Predictive Variance:** Increases with larger σ_f, resulting in wider uncertainty bands (notice how the bounds of the axes changes as σ_f increases).
     - **Varying l:**
       - **Predictive Mean:** Smaller l values allow the function to change rapidly, fitting closely to training data and resulting jagged bands; larger l yields smoother functions.
       - **Predictive Variance:** Smaller l leads to lower variance near training points but higher variance away from them.
   - **Matérn Kernel:**
     - **Varying ν:**
       - **Predictive Mean:** Smaller ν values ($0.25 \leq \nu \leq 3$) allow for rougher, less smooth functions; larger ν results in smoother functions.
       - **Predictive Variance:** Smaller ν increases uncertainty between data points as the function is more jagged.
     - **Varying l:**
       - Again, smaller l permits allows for a more jagged function. So the same holds here for the predicted mean and variance in the Matérn Kernel as the Squared Exponential Kernel.

   See the attached Jupyter notebook for the solution.

STAT201A: Introduction to Probability at an Advanced Level December 6th, 2024

Homework # 6: Discrete Markov Chains

Reece D. Huff

Notation

Let $\{X_t, t \in \mathbb{N}_0\}$ represent a Galton-Watson process where X_t represents the number of particles at time t and $\mathbb{N}_0$ respresents the set of natural numbers including 0, i.e., $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Let $B_i^{(t)}$ represent the number of offspring of branch i at time t. Then we have

$$X_t = B_1^{(t-1)} + \cdots + B_{X_{t-1}}^{(t-1)}.$$

The probability of branch i has k offspring at time t, $\mathbb{P}[B_i^{(t)} = k]$, is given by an offspring number distribution $\mathcal{F} = \{p_k, k \in \mathbb{N}_0\}$ with mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Here, p_k is shorthand for $\mathbb{P}[B_i^{(t)} = k]$. We note that each particle gives birth to $k \in \mathbb{N}_0$ children with probability independently of other particles in the past and the present, i.e.,

$$B_1^{(t-1)}, \ldots, B_{X_{t-1}}^{(t-1)} \overset{\text{i.i.d.}}{\sim} \mathcal{F} \quad \text{and} \quad \perp\!\!\!\perp \quad \text{from} \quad X_t.$$

In class, we showed that $\mathbb{E}[X_t] = \mu^t$ and $\text{Var}[X_t] = \sigma^2\left(\mu^{t-1} + \mu^t + \cdots + \mu^{2t-2}\right)$.
We defined the extinction time as $\tau = \{t \in \mathbb{N}_0 \mid X_t = 0\}$ and the extinction probability as $\mathbb{P}[\tau < \infty]$. We also introduced the notation

$$\varphi(s) = \mathbb{E}[s^B] = \sum_{k=0}^{\infty} s^k \, \mathbb{P}[B = k] = \sum_{k=0}^{\infty} s^k p_k \quad \text{and} \quad \varphi_t(s) = \mathbb{E}[s^{X_t}] = \sum_{k=0}^{\infty} s^k \, \mathbb{P}[X_t = k]$$

We let $e_t = \mathbb{P}[X_t = 0]$, the probability of extinction by time t, from which we showed $e_t = \varphi(e_{t-1})$. We derived that the probability of extinction is the smallest non-negative solution of the fixed point equation, $s = \varphi(s)$.

Markov Chain Definitions in Simple Terms

- **Irreducible**: A Markov Chain is irreducible if it is possible to visit every other state regardless of where you start. In most cases, we determine if an MC is irreducible by looking for terminal states: if a chain has a terminal state, it's definitely not irreducible (unless the state space consists of that one state only). However, the absence of a single terminal state does not necessarily guarantee irreducibility, because it's possible to have multiple disjoint sets of states that don't communicate with each other (no terminal states, but the chain still isn't fully "connected").
- A state $i \in \mathcal{S}$ is...
  1. **recurrent** or **persistent** if when starting from i, the probability of eventually returning to i in the future is 1.
     (a) **positive recurrent** Imagine you're in a particular state (a place) that you can return to again and again. If, on average, it doesn't take you too long to get back there—meaning the expected time until you return is finite—then this state is positive recurrent. It's like a bus that comes around regularly and quickly enough that you never have to wait forever.
     (b) **null recurrent** Now imagine you're still guaranteed to return to that same state eventually (just like above), but this time, you might have to wait a very, very long time on average—so long that the average waiting time is actually infinite. In other words, it's certain you'll get back there if you wait long enough, but there is no meaningful "average" return time because it stretches out to infinity. Such a state is called null recurrent. It's like a bus that will definitely come, but you can't put a reasonable number on how long you'll have to wait.
  2. **transient** if when starting from i, the probability of eventually returning to i in the future is not 1.
- **Periodicity**:
  1. **Periodic**: If I know for a fact, that after n steps, I will be at some state j.
  2. **Aperiodic**: If I do not know for a fact, that after n steps, I will be at some state j.
- **Ergodicity**: A Markov chain is **ergodic** if it is irreducible, aperiodic, and (usually) positive recurrent (for finite state spaces, irreducible + aperiodic automatically gives positive recurrence).

Problems (Solutions)

1. **(Branching process)** A branching process starts with one individual, i.e. $X(0) = 1$, who reproduces according to the following principle:

# of children	0	1	2
probability	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Individuals reproduce independently of each other and independently of the number of their sisters and brothers. Determine

(a) the probability that the population becomes extinct;

Recall in class, we derived that the probability of extinction is the smallest non-negative solution of the fixed point equation, $s = \varphi(s)$. Then, we have that

$$s = \varphi(s) = \sum_{k=0}^{\infty} s^k p_k = s^0 p_0 + s^1 p_1 + s^2 p_2 \quad \Longrightarrow \quad s = \frac{1}{6} + \frac{1}{3}s + \frac{1}{2}s^2 \quad \Longrightarrow \quad \frac{1}{2}s^2 - \frac{2}{3}s + \frac{1}{6} = 0$$
$$\Longrightarrow \quad s^2 - \frac{4}{3}s + \frac{1}{3} = 0 \quad \Longrightarrow \quad (s-1)(s-1/3) = 0 \quad \Longrightarrow \quad s_1 = 1, s_2 = \frac{1}{3}.$$

Clearly, $s_2 = \frac{1}{3}$ is the smallest non-negative solution of the fixed point equation, $s = \varphi(s)$, so we have

$$\boxed{\mathbb{P}[\text{extinction}] = \mathbb{P}[\tau < \infty] = \frac{1}{3}.}$$

(b) the probability that the population has become extinct in the second generation, i.e. $\mathbb{P}[X(2) = 0]$;

By the law of total probability[3], we have that

$$\mathbb{P}[X_2 = 0] = \sum_{k=0}^{\infty} \mathbb{P}[X_2 = 0 \mid X_1 = k]\,\mathbb{P}[X_1 = k]$$

$$\mathbb{P}[X_2 = 0] = \mathbb{P}[X_2 = 0 \mid X_1 = 0]\,\mathbb{P}[X_1 = 0] + \mathbb{P}[X_2 = 0 \mid X_1 = 1]\,\mathbb{P}[X_1 = 1] + \mathbb{P}[X_2 = 0 \mid X_1 = 2]\,\mathbb{P}[X_1 = 2]$$

($\mathbb{P}[X_2 = 0 \mid X_1 = 0] = 1$)

$$\mathbb{P}[X_2 = 0] = \left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right)\left(\frac{1}{2}\right)$$

$$\boxed{\mathbb{P}[X_2 = 0] = \frac{17}{72}}$$

(c) the expected number of children given that there are no grandchildren.

In this problem, we are asked to calculate $\mathbb{E}[X_1 | X_2 = 0]$. We have that

$$\mathbb{E}[X_1 | X_2 = 0] = \sum_{k=0}^{\infty} k\,\mathbb{P}[X_1 = k | X_2 = 0] = \sum_{k=0}^{\infty} k \left(\frac{\mathbb{P}[X_2 = 0 | X_1 = k]\,\mathbb{P}[X_1 = k]}{\mathbb{P}[X_2 = 0]}\right)$$
$$\mathbb{E}[X_1 | X_2 = 0] = 1\left(\frac{\mathbb{P}[X_2 = 0 | X_1 = 1]\,\mathbb{P}[X_1 = 1]}{\mathbb{P}[X_2 = 0]}\right) + 2\left(\frac{\mathbb{P}[X_2 = 0 | X_1 = 2]\,\mathbb{P}[X_1 = 2]}{\mathbb{P}[X_2 = 0]}\right)$$

[3]Note that this approach is no different from the approach in class, where we showed $e_t = \varphi(e_{t-1})$ with $e_t = \mathbb{P}[X_t = 0]$ representing the probability of extinction by time t, i.e.,

$$\mathbb{P}[X_2 = 0] = e_2 = \varphi(e_1) = \sum_{k=0}^{\infty} (e_1)^k p_k = (e_1)^0 p_0 + (e_1)^1 p_1 + (e_1)^2 p_2 = (e_1)^0 p_0 + (e_1)^1 p_1 + (e_1)^2 p_2$$

$$\mathbb{P}[X_2 = 0] = \left(\frac{1}{6}\right) + \left(\frac{1}{3}\right) e_1 + \left(\frac{1}{2}\right)(e_1)^2 = \left(\frac{1}{6}\right) + \left(\frac{1}{3}\right)\left(\sum_{\ell=0}^{\infty} (e_1)^\ell p_\ell\right) + \left(\frac{1}{2}\right)\left(\sum_{\ell=0}^{\infty} (e_1)^\ell p_\ell\right)^2$$

$$\mathbb{P}[X_2 = 0] = \left(\frac{1}{6}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{6}\right)^2 = \frac{17}{72}$$

$$\mathbb{E}[X_1|X_2=0] = 1\left(\frac{(1/6)(1/3)}{(17/72)}\right) + 2\left(\frac{(1/6)^2(1/2)}{(17/72)}\right) = \frac{72}{17}\left[1\left(\frac{1}{6}\right)\left(\frac{1}{3}\right) + 2\left(\frac{1}{6}\right)^2\left(\frac{1}{2}\right)\right]$$

$$\boxed{\mathbb{E}[X_1|X_2=0] = \frac{6}{17}}$$

2. **(Random walk)** Random walk on $\{0,1,2,3\}$. Consider the Markov chain (X_n) with transition matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

started with $X_0 = 0$. Define T_j as $\min\{n \geq 1 : X_n = j\}$. Find explicitly the following distributions and expectations:

(a) The distribution of X_2.

From class, we defined the row vector $\vec{u}_n := \left[\mathbb{P}[X_n = i]\right]$ for all $i \in \mathcal{S} := \{0,1,2,3\}$ such that $\vec{u}_n = \vec{u}_0 P^n$. It follows that

$$\vec{u}_2 = \vec{u}_0 P^2 = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

Thus, the distribution of X_2 is given by

$$\boxed{\mathbb{P}[X_2 = i] = \begin{cases} \frac{1}{2} & \text{when } i = 0 \\ \frac{1}{4} & \text{when } i = 1 \\ \frac{1}{4} & \text{when } i = 2 \\ 0 & \text{when } i = 3 \end{cases}}$$

(b) The limit distribution of X_n as $n \to \infty$.

Since this Markov Chain has a finite state space $\mathcal{S}$ and is irreducible and aperiodic, the limit distribution of X_n as $n \to \infty$ is given by

$$\lim_{n\to\infty} P^n = \mathbf{1}\pi \quad \text{where } \pi \text{ is the unique stationary distribution from } \pi = \pi P.$$

It follows that

$$\pi = \pi P \Leftrightarrow \pi^\top = P^\top \pi^\top \quad \Longrightarrow \quad \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\pi_0 + \frac{1}{2}\pi_1 \\ \frac{1}{2}\pi_0 + \frac{1}{2}\pi_2 \\ \frac{1}{2}\pi_1 + \frac{1}{2}\pi_3 \\ \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3 \end{bmatrix} \quad \Longrightarrow \quad \pi_0 = \pi_1 = \pi_2 = \pi_3.$$

Since $\pi_0 = \pi_1 = \pi_2 = \pi_3$ and we must have an eigenvalue of 1, we ensure that $\sum_{i\in\mathcal{S}} \pi_i = 1$. Then the limit distribution of X_n as $n \to \infty$.

$$\boxed{\lim_{n\to\infty} [P^n]_{ij} = \pi_j, \forall i,j \in \mathcal{S} \quad \text{where } \pi = (\pi_i)_{i\in\mathcal{S}} \text{ is the unique stationary distribution} \quad \pi_j = \begin{cases} \frac{1}{4} & \text{when } j = 0 \\ \frac{1}{4} & \text{when } j = 1 \\ \frac{1}{4} & \text{when } j = 2 \\ \frac{1}{4} & \text{when } j = 3 \end{cases}}$$

(c) $\mathbb{E}[T_0]$

Since our Markov chain is irreducible and finite, then π is the unique stationary distribution and $\pi_i = 1/r_i, \ \forall i \in \mathcal{S}$ where $r_i = \mathbb{E}[T_i]$ is the mean recurrence time. It follows that

$$\pi_0 = \frac{1}{r_0} \quad \Longrightarrow \quad r_0 = \frac{1}{1/4} = 4. \quad \Longrightarrow \quad \boxed{\mathbb{E}[T_0] = 4.}$$

We can equivalently show that for $E_i = \mathbb{E}[T_0 \mid X_0 = i]$ for all $i \in \mathcal{S}$, we have

$$\begin{aligned} E_0 &= 1 + 1/2 E_1 \\ E_1 &= 1 + 1/2 E_2 \\ E_2 &= 1 + 1/2 E_1 + 1/2 E_3 \\ E_3 &= 1 + 1/2 E_2 + 1/2 E_3 \end{aligned} \quad \Longrightarrow \quad E_2 = 1 + \frac{1}{2}E_1 + \left(1 + \frac{1}{2}E_2\right) \quad \Longrightarrow \quad E_1 = 1 + \left(2 + \frac{1}{2}E_1\right) \quad \Longrightarrow \quad E_0 = 1 + (3) = 4$$

(d) $\mathbb{E}[T_3]$

Similarly, let $F_i = \mathbb{E}[T_3 \mid X_0 = i]$ for all $i \in \mathcal{S}$. Then we have

$$\begin{aligned} \mathbb{E}[T_3 \mid X_0 = 0] &= \mathbb{E}[T_3 \mid X_0 = 0, X_1 = 0]\,\mathbb{P}[X_1 = 0 \mid X_0 = 0] + \mathbb{E}[T_3 \mid X_0 = 0, X_1 = 1]\,\mathbb{P}[X_1 = 1 \mid X_0 = 0] \\ \mathbb{E}[T_3 \mid X_0 = 1] &= \mathbb{E}[T_3 \mid X_0 = 1, X_1 = 0]\,\mathbb{P}[X_1 = 0 \mid X_0 = 1] + \mathbb{E}[T_3 \mid X_0 = 1, X_1 = 2]\,\mathbb{P}[X_1 = 2 \mid X_0 = 1] \\ \mathbb{E}[T_3 \mid X_0 = 2] &= \mathbb{E}[T_3 \mid X_0 = 2, X_1 = 1]\,\mathbb{P}[X_1 = 1 \mid X_0 = 2] + \mathbb{E}[T_3 \mid X_0 = 2, X_1 = 3]\,\mathbb{P}[X_1 = 3 \mid X_0 = 2] \end{aligned}$$

$$\begin{aligned} &\Leftrightarrow \begin{aligned} F_0 &= 1 + 1/2 F_0 + 1/2 F_1 \\ F_1 &= 1 + 1/2 F_0 + 1/2 F_2 \\ F_2 &= 1 + 1/2 F_1 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} F_0 &= 2 + F_1 \\ F_1 &= 1 + 1/2 F_0 + 1/2 F_2 \\ F_1 &= 2F_2 - 2 \end{aligned} \quad \Longrightarrow \quad 1 + 1/2 F_0 + 1/2 F_2 = 2F_2 - 2 \quad \Longrightarrow \quad F_0 = 3F_2 - 6 \\ &\Longrightarrow \quad F_0 = 2 + (2F_2 - 2) = 2F_2 \quad \Longrightarrow \quad F_2 = 6 \quad \Longrightarrow \quad F_0 = 12 \end{aligned}$$

Thus, we have

$$\boxed{\mathbb{E}[T_3] = 12.}$$

(e) $\mathbb{P}[T_3 < T_0]$

The probability that $T_3 < T_0$ is given by

$$\mathbb{P}[T_3 < T_0] = \mathbb{P}[T_3 < T_0 \mid X_0 = 0]\,\mathbb{P}[X_0 = 0]$$

We can find the probability of $T_3 < T_0$ by conditioning on the first step of the Markov chain. We have

$$\begin{aligned} \mathbb{P}[T_3 < T_0 \mid X_0 = 0] &= \mathbb{P}[T_3 < T_0 \mid X_0 = 0, X_1 = 0]\,\mathbb{P}[X_1 = 0 \mid X_0 = 0] + \mathbb{P}[T_3 < T_0 \mid X_0 = 0, X_1 = 1]\,\mathbb{P}[X_1 = 1 \mid X_0 = 0] \\ \mathbb{P}[T_3 < T_0 \mid X_0 = 1] &= \mathbb{P}[T_3 < T_0 \mid X_0 = 1, X_1 = 0]\,\mathbb{P}[X_1 = 0 \mid X_0 = 1] + \mathbb{P}[T_3 < T_0 \mid X_0 = 1, X_1 = 2]\,\mathbb{P}[X_1 = 2 \mid X_0 = 1] \\ \mathbb{P}[T_3 < T_0 \mid X_0 = 2] &= \mathbb{P}[T_3 < T_0 \mid X_0 = 2, X_1 = 1]\,\mathbb{P}[X_1 = 1 \mid X_0 = 2] + \mathbb{P}[T_3 < T_0 \mid X_0 = 2, X_1 = 3]\,\mathbb{P}[X_1 = 3 \mid X_0 = 2] \end{aligned}$$

Let $G_i = \mathbb{P}[T_3 < T_0 \mid X_0 = i]$ for all $i \in \mathcal{S}$. Then we have

$$\begin{aligned} G_0 &= 1/2 G_1 \\ G_1 &= 1/2 G_2 \\ G_2 &= 1/2 G_1 + 1/2 \end{aligned} \quad \Longrightarrow \quad \begin{aligned} 2G_0 &= G_1 \\ G_1 &= 1/2\,(1/2 G_1 + 1/2) \end{aligned} \quad \Longrightarrow \quad 2G_0 = 1/2\,(G_0 + 1/2) \quad \Longrightarrow \quad 3/2 G_0 = 1/4 \quad \Longrightarrow \quad \boxed{\mathbb{P}[T_3 < T_0] = 1/6}$$

3. **(The average number of jobs)**

Jennifer is employed for one day at a time. When she is out of work, she visits the job agency in the morning to see if there is work for that day. There is a job for her with probability 1/2. If there is no work, she comes back the next day. When she has a job, she will be called back to the same job for the next day with probability 2/3. When she is not called back, she goes to the job agency again the next morning to look for a new job that she had not had previously. Approximate the average number of jobs Jennifer works in a year.

Let the state in which Jennifer is not working and the state in which she is working be the 0th and 1st states of a 2-state Markov chain, respectively. Let X_n be the state of Jennifer on day n with $X_0 = 1$. The transition matrix is given by

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \end{bmatrix}$$

Recall in class, we showed that the fraction of time spent in state j during the steps $0, 1, 2, \ldots, n$ is given by

$$H_j^{(n)} := \frac{1}{n+1} \sum_{k=0}^{n} \mathbf{1}\{X_k = j\}$$

The expected value given the initial state $X_0 = i$ is

$$\mathbb{E}[H_j^{(n)} \mid X_0 = i] = \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}[X_k = j \mid X_0 = i] = \frac{1}{n+1} \sum_{k=0}^{n} [P^k]_{ij} \quad \overset{n\to\infty}{\longrightarrow} \quad \pi_j$$

We use this result to find the average number of jobs Jennifer works in a year. We approximate that 365 days is enough for the chain to reach its stationary distribution. Then, we have

$$\mathbb{E}[\text{Jobs worked in a year}] = 365 \times \mathbb{E}[H_j^{(365)} \mid X_0 = 0] \approx 365 \times \pi_1$$

where $\pi = (\pi_0, \pi_1)$ is the unique stationary distribution of the chain. We find π by solving the equation $\pi = \pi P$ to get

$$\pi = P\pi \iff \pi^\top = P^\top \pi^\top \implies \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} = \begin{bmatrix} 1/2\pi_0 + 1/3\pi_1 \\ 1/2\pi_0 + 2/3\pi_1 \end{bmatrix} \implies \begin{bmatrix} \pi_0 \\ \pi_1 \end{bmatrix} = \begin{bmatrix} 2/5 \\ 3/5 \end{bmatrix}$$

Therefore, the average number of jobs Jennifer works in a year is

$$\boxed{\mathbb{E}[\text{Jobs worked in a year}] \approx 219.}$$

4. **(Rain or no rain)** Suppose that at day 0 it is not raining. Then each new day, if it rained yesterday, it will rain with probability 0.7; if it did not rain yesterday, it will rain with probability 0.2.

Let the state in which it is not raining and the state in which it is raining be the 0th and 1st states of a 2-state Markov chain, respectively. Let X_n be the state of the weather on day n with $X_0 = 0$. The transition matrix is given by

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}$$

(a) Find the stationary distribution.

We showed in class that the unique stationary distribution $\pi = (\pi_0, \pi_1)$ of the chain is given by

$$\pi = \frac{1}{a+b}[b, a]$$

where a and b are the off-diagonal entries of the transition matrix P. Therefore, we have

$$\boxed{\pi = \frac{1}{0.2+0.3}[0.3, 0.2] = \left[\frac{3}{5}, \frac{2}{5}\right]}$$

(b) How many days should we expect to wait to have rain for the first time?

For this part, we are interested in finding the mean first passage time from state 0 to state 1. We defined the fundamental matrix of irreducible MC as

$$Z = (I - P + \mathbf{1}\pi)^{-1}$$

In this case, we have

$$Z = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} + \begin{bmatrix} 0.6 & 0.4 \\ 0.6 & 0.4 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}^{-1} = \frac{1}{0.5} \begin{bmatrix} 0.7 & -0.2 \\ -0.3 & 0.8 \end{bmatrix} = \begin{bmatrix} 1.4 & -0.4 \\ -0.6 & 1.6 \end{bmatrix}$$

The mean first passage time is given by

$$\mathbb{E}[T_j \mid X_0 = i] = \frac{Z_{jj} - Z_{ij}}{\pi_j}$$

Therefore, we have

$$\boxed{\mathbb{E}[T_1 \mid X_0 = 0] = \frac{Z_{11} - Z_{01}}{\pi_1} = \frac{1.6 - (-0.4)}{2/5} = 5 \text{ days}}$$

5. **(The game of roulette)** A gambler plays the game of roulette, betting X dollars on red or black. The gambler wins X dollars with probability $p = 18/38$ or loses the bet with probability $q = 20/38$. Suppose that the gambler starts the game with \$500 in his pocket and an upper limit on winnings is 1000.

Before we proceed, we derive the Gambler's ruin formula which that the probability of ruin for a gambler starting with i units of money is given by

$$\mathbb{P}[\text{ruin} \mid X_0 = i] = \frac{\left(\frac{q}{p}\right)^i - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N}$$

where N is the upper limit on winnings. We will use this formula to compute the probability of ruin for the gambler.

Proof. Suppose we normalize our Markov chain such that the gambler starts with i units of money where $0 \leq i \leq N$. Let X_n be the gambler's fortune at time n with $X_0 = i$. The gambler's fortune evolves as

$$X_{n+1} = \begin{cases} X_n + 1 & \text{with probability } p \\ X_n - 1 & \text{with probability } q \end{cases}$$

The gambler's ruin is the event that the gambler loses all his money, i.e., $X_n = 0$. We are interested in finding the probability of ruin, i.e., $\mathbb{P}[X_n = 0 \mid X_0 = i]$. We begin by writing the recursive relation

$$\mathbb{P}[X_n = 0 \mid X_0 = i] = p\,\mathbb{P}[X_{n-1} = 1 \mid X_0 = i] + q\,\mathbb{P}[X_{n-1} = -1 \mid X_0 = i]$$

Let $r_i = \mathbb{P}[X_n = 0 \mid X_0 = i]$. We can write the above equation as

$$r_i = pr_{i+1} + qr_{i-1}.$$

We then have

$$x^i = px^{i+1} + qx^{i-1} \quad \Longrightarrow \quad px^2 - x + q = 0 \quad \Longrightarrow \quad (x-1)(x-q/p) = 0$$

The roots of the above equation are $x = 1$ and $x = q/p$. Therefore, the general solution to the above equation is given By

$$r_i = A + B\left(\frac{q}{p}\right)^i$$

We can solve for A and B by using the boundary conditions $r_0 = 1$ and $r_N = 0$. We have

$$r_0 = A + B = 1 \quad \text{and} \quad r_N = A + B\left(\frac{q}{p}\right)^N = 0$$

Solving the above equations, we get

$$A = \frac{-\left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N} \quad \text{and} \quad B = \frac{1}{1 - \left(\frac{q}{p}\right)^N}$$

Therefore, we have

$$r_i = \frac{\left(\frac{q}{p}\right)^i - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N}$$

This completes the proof. □

(a) Compute the probability of the gambler's ruin for $X = \$10$.

Now that we have the formula for the probability of ruin, we can plug in the values of i and N to get the probability of ruin for the gambler. We normalize our values such that the gambler starts with i units of money where $0 \leq i \leq N$. We have $i = 50$ and $N = 100$. Also, note that $q/p = 20/38/18/38 = 20/18 = 10/9$. Therefore, we have

$$\mathbb{P}[\text{ruin} \mid X_0 = \$500] = \mathbb{P}[\text{ruin} \mid X_0 = 50] = \frac{\left(\frac{q}{p}\right)^{50} - \left(\frac{q}{p}\right)^{100}}{1 - \left(\frac{q}{p}\right)^{100}} \approx 0.9949$$

(b) Compute the probability of the gambler's ruin for $X = \$100$.

We do the same for $X = \$100$. We have $i = 5$ and $N = 10$. Therefore, we have

$$\mathbb{P}[\text{ruin} \mid X_0 = \$500] = \mathbb{P}[\text{ruin} \mid X_0 = 5] = \frac{\left(\frac{q}{p}\right)^5 - \left(\frac{q}{p}\right)^{10}}{1 - \left(\frac{q}{p}\right)^{10}} \approx 0.6287$$

(c) Compare the above results with the probability of ruin in the case the gambler bets everything on a single turn of the wheel.

If we bet it all on a single turn of the wheel, the probability of ruin is simply the probability of losing the bet, i.e., $q = {}^{20}/_{38} \approx 0.5263$. This is much higher than the probability of ruin in the previous cases.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 1

Due: 10:00pm, Friday, September 13, 2024 (via Gradescope)

1. (**Basic probability**) Assume that $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.7$ and $\mathbb{P}(C) = 0.8$.

   (a) Show that $0.3 \leq \mathbb{P}(A \cap B) \leq 0.6$.

   (b) Show that $0.1 \leq \mathbb{P}(A \cap B \cap C) \leq 0.6$.

2. **(Independence)** Suppose we roll an unbiased six-sided die $n \geq 3$ times. Let E_{ij} denote the event that the ith and the jth rolls produce the same number. Show that the events $\{E_{ij} \mid 1 \leq i < j \leq n\}$ are pairwise independent but not independent as a family.

3. (**Expectation, joint distribution, uniform distribution**) Let X be a random variable with values $\{1, 2\}$ and Y a random variable with values $\{0, 1, 2\}$. Initially we have the following partial information about their joint probability mass function.

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 1$	$1/8$		
$X = 2$		0	

   Subsequently we learn that $\mathbb{E}[XY] = \frac{13}{9}$ and that Y has uniform distribution. Use this information to fill in the missing values of the joint probability mass function table.

4. (**Conditioning, cumulative distribution function**) You flip a fair coin. If you get tails, you choose a uniformly random number on the interval $[0, 2]$. If you get heads, you choose the number 1. Let X be the random variable describing the outcome of that experiment.

   (a) Using the law of total probabilities, calculate $\mathbb{P}(X \leq 1/2)$ and $\mathbb{P}(X \leq 3/2)$.

   (b) Find the cumulative distribution function F_X of X.

   (c) Is X a discrete random variable? Is X a continuous random variable?

5. (**Bounding even moments**) Let X be a random variable. Show that $\mathbb{E}[X^{2k}] \geq (\mathbb{E}[X])^{2k}$ for all positive integers k.

6. (**Continuous distributions, probability density function, independence**) Pick a uniformly chosen random point (X, Y) inside the sector delimited by the x-axis, the y-axis and the parabola given by the equation $y = 1 - x^2$; see Figure 1.

   (a) Verify that the area of that sector is $2/3$.

   (b) What is the probability that the distance of this point to the y-axis is **less** than $1/2$?

   (c) What is the probability that the distance of this point to the origin is **more** than $1/2$?

   (d) Find the p.d.f. of X.

   (e) Find the p.d.f. of Y.

   (f) Are X and Y independent?

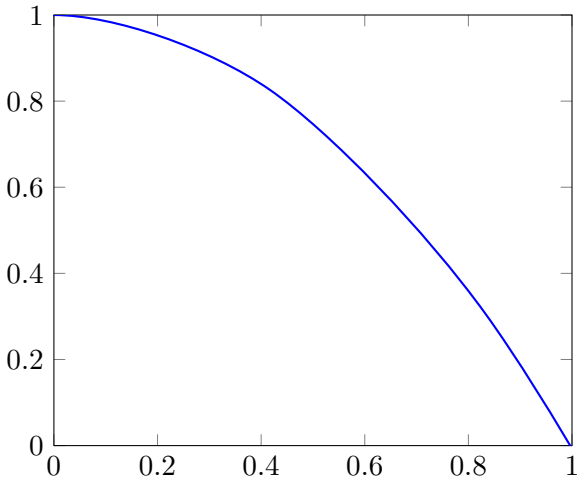


Figure 1: Graph of $y = 1 - x^2$

7. (**Events, indicators and basic probability inequalities**) Recall that for an event A, we denote the corresponding indicator random variable by $I(A)$ (i.e., $I(A)$ takes value 1 when A occurs and the value 0 when A does not occur). Also recall that the probability $\mathbb{P}(A)$ of A equals the expectation of the random variable $\mathbb{E}(I(A))$.

   (a) Given events $A_1, \ldots, A_n$, show that $I\left(\cup_{i=1}^n A_i\right) = \max_{1 \leq i \leq n} I\left(A_i\right)$.

   (b) Using the fact observed above (and the following ordering property of expectation: $X \leq Y$ implies that $\mathbb{E}(X) \leq \mathbb{E}(Y)$), show that

   $$\mathbb{P}\left(\cup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}\left(A_i\right).$$

   Note: This is known as the union bound and used quite frequently.

   (c) For every event A, show that $I\left(A^c\right) = 1 - I(A)$ where A^c denotes the event that A does not occur.

   (d) For events $A_1, \ldots, A_n$, show that $I\left(\cap_{i=1}^n A_i\right) = \prod_{i=1}^n I\left(A_i\right)$.

   (e) Using the above two facts, prove the inclusion-exclusion formula: For events $A_1, \ldots, A_n$,

   $$\mathbb{P}\left(\cup_{i=1}^n A_i\right) = \Sigma_1 - \Sigma_2 + \Sigma_3 - \Sigma_4 + \cdots + (-1)^{n-1} \Sigma_n$$

   where

   $$\Sigma_k := \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mathbb{P}\left(A_{i_1} A_{i_2} \cdots A_{i_k}\right).$$

8. (**Hypergeometric and exchangeability**) We have an urn with R red balls and $N - R$ white balls, where $0 < R < N$. We draw n balls in sequence from the urn without replacement. Let R_i denote the proposition that the i^{th} draw results in a red ball.

   (a) Calculate $\mathbb{P}\left(R_i\right)$ for each $i = 1, \ldots, n$.

   (b) Show that $\mathbb{P}\left(R_j \mid R_k\right) = \mathbb{P}\left(R_k \mid R_j\right)$ for every $1 \leq j, k \leq n$.

   (c) Calculate $\mathbb{P}\left(R_k \mid \bigcup_{i=k+1}^n R_i\right)$ for a fixed $1 \leq k < n$.

(d) Let X be the random variable representing the minimum number of draws required to get at least one red ball. Calculate $\mathbb{E}[X]$, the expected value of X. (Hint: Use exchangeability to simplify the calculation.)

(e) Suppose that instead of only two colors, the urn has balls of k different colors: N_1 of color $1, N_2$ of color $2, \ldots, N_k$ of color k. Let $N = N_1 + \cdots + N_k$. Argue that the probability of drawing r_1 balls of color $1, r_2$ balls of color $2, \ldots, r_k$ balls of color k in $n = r_1 + \cdots + r_k$ draws without replacement is given by

$$\frac{\binom{N_1}{r_1} \cdots \binom{N_k}{r_k}}{\binom{N}{n}}.$$

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 2

Due: 10:00pm, Friday, September 27, 2024 (via Gradescope)

1. **(Binomial tail bounds)** Let S_n have the Binomial(n, p_i) distribution of the number of successes in n independent Bernoulli(p) trials. Use a suitable computational environment to evaluate the right tail probabilities

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right)$$

for $n = 100$ and $p_i = i/10$ for $i = 1, 2, \ldots, 9$, and $\epsilon = 1/10$, together with various approximations and upper bounds as indicated. In each case,

- give an exact mathematical formula for the function of i you are computing;
- indicate suitable code for evaluating the formula in your preferred environment and **attach the code at the end of the homework**;
- give the numerical values correct to two significant decimal place.

   (a) The exact probabilities.
   (b) Markov's upper bounds for these probabilities.
   (c) Chebychev's upper bounds for these probabilities (which can be halved for $i = 5$ only: explain why).
   (d) Hoeffding's upper bounds.
   (e) Chernoff's upper bounds.

2. **(LLN)** Suppose that $X_1, X_2, \ldots$ form an i.i.d. sequence of random variables with $\mathbb{E}[X_i] = \mu < \infty$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Evaluate

$$\lim_{n\to\infty} \frac{1}{\binom{n}{2}} \sum_{i,j:1\leq i<j\leq n} (X_i - X_j)^2.$$

3. **(Chebyshev & CLT)** Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with mean zero and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Determine the limits below, with precise justifications.

   (a) $\lim_{n\to\infty} \mathbb{P}(S_n \geq 0.01n)$.
   (b) $\lim_{n\to\infty} \mathbb{P}(S_n \geq 0)$.
   (c) $\lim_{n\to\infty} \mathbb{P}(S_n \geq -0.01n)$.

4. **(Convolution & MGF)** The Laplace distribution has density $f_Z(z) = \frac{\lambda}{2}\exp(-\lambda|z|)$ and MGF $M_Z(t) = \frac{\lambda^2}{\lambda^2 - t^2}$, where $\lambda > 0$. Let $X, Y \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$. Prove that $Z = X - Y$ follows a Laplace distribution by using:

   (a) Moment generating functions.
   (b) The convolution formula.

5. **(Moments & MGF)** Let X be a random variable with p.d.f. given by

$$f_X(x) = \begin{cases} 2/9, & \text{if } 0 \le x \le 1, \\ (4 - |4 - 2x|)/9, & \text{if } 1 < x \le 4, \\ 0, & \text{otherwise.} \end{cases}$$

   (a) Verify that this is actually a p.d.f.

   (b) Find the moment generating function of X.

   (c) Find $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

   (d) Find a formula for the moments of X.

6. **(Distribution of sums using MGFs)** Let $S_n := X_1 + \cdots + X_n$ for independent $X_1, \ldots, X_n$. Use MGFs to find the distribution of S_n

   (a) For X_i with Normal $\left(\mu_i, \sigma_i^2\right)$ distribution.

   (b) For X_i with Gamma (r_i, λ) distribution.

   (c) For $X_i = Z_i^2$ with $Z_i \sim \mathrm{Normal}(0, 1)$.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 3

Due: 10:00pm, Friday, October 11, 2024 (via Gradescope)

1. **(Approximating Binomial Distributions)** The goal of this question is to empirically verify three approximations to the exact Binomial probability $\mathbb{P}(X = k)$, where $X \sim \text{Binomial}(n, p)$:

   - $\mathbb{P}(Y = k)$, where $Y \sim \text{Poisson}(np)$, the Poison approximation with rate parameter np;
   - The normal approximation

   $$\phi(k; np, np(1-p)) := \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{(k-np)^2}{2np(1-p)}\right\}$$

   - The entropic approximation

   $$\text{Ent}(k; n, p) := \frac{1}{\sqrt{2\pi nf(1-f)}} \exp(-n\text{KL}(f\|p))$$

   where $f = \frac{k}{n}$ and $\text{KL}(f\|p) = f \log \frac{f}{p} + (1-f) \log \frac{1-f}{1-p}$.

   (a) Take $n = 30$ and $p = 0.05$. Create a table (31 rows and 3 columns) containing the absolute errors for each approximation,

   $$|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|, \quad |\mathbb{P}(X = k) - \phi(k; np, np(1-p))|$$

   and

   $$|\mathbb{P}(X = k) - \text{Ent}(k; n, p)|$$

   for $k = 0, 1, \ldots, 30$. (Note: The entropic approximation does not exist for $k = 0$ and $k = 30$, so only list it for $k = 1, \ldots, 29$).
   Based on the table, comment on the accuracy of each of the three approximations for the Binomial distribution.

   (b) Create a similar table for the relative errors

   $$\frac{|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|}{\mathbb{P}(X = k)}, \quad \frac{|\mathbb{P}(X = k) - \phi(k; np, np(1-p))|}{\mathbb{P}(X = k)}$$

   and

   $$\frac{|\mathbb{P}(X = k) - \text{Ent}(k; n, p)|}{\mathbb{P}(X = k)}$$

   for $k = 0, 1, \ldots, 30$. Based on this table, comment on the accuracy of each of the three approximations for the Binomial.

   (c) Repeat exercises (a) and (b) above for $n = 30$ and $p = 0.25$.

   (d) Repeat exercises (a) and (b) above for $n = 30$ and $p = 0.5$.

2. **(KL-Divergence, Multinomial)** Let X and Y be discrete random variables with distributions p and q respectively (So $p(k) = \mathbb{P}(X = k)$ and $q(k) = \mathbb{P}(Y = k)$). Remember that the Kullback–Leibler divergence is defined by

$$\mathrm{KL}(p\|q) := \mathbb{E}_p\Big[\ln\big(\frac{p(X)}{q(X)}\big)\Big] = \sum_k p(k)\ln\big(\frac{p(k)}{q(k)}\big).$$

   (a) Show that when $q(k)$ is a Poisson distribution with parameter $\lambda > 0$, then the KL-divergence is minimized by setting λ to be the mean of $p(k)$.

   (b) Rememebr that the entropy $H(p)$ is defined to be $H(p|q) := -\mathbb{E}_p[\ln(p(X))]$. Assume that we need to place n balls into d bins. The number of ways to place the balls resulting in k_i total balls in bin i, for $i = 1, \ldots, d$, is given by the combinatorial expression $\binom{n}{k_1, k_2, \ldots, k_d}$. Now consider the empirical distribution of the balls. Its probability mass function is $p(i) = k_i/n$. Let N_p denote the number of configurations with empirical distribution p, show that

$$\ln(N_p) = nH(p) + O(\ln(n)),$$

   where $h(p)$ is the entropy of p.

   In other words, there are many more high-entropy configurations than low-entropy configurations. This suggests the intuition that, if we consider a physical system at a "macro level" (such as the distribution of gas particles in a container) then we should expect it to drift toward high-entropy configurations.

   Hint: Recall Stirling's approximation

$$\ln(n!) = n\ln(n) - n + O(\ln(n)).$$

3. **(Poisson)** Let $K = X_1 + X_2 + \cdots + X_N$, where $N \sim \mathrm{Poisson}\,(\lambda)$ and $X_1, X_2, \cdots$ are independent $\mathrm{Bernoulli}\,(p)$ random variables. Assuming that N and $\{X_i\}_{i\in\mathbb{N}}$ are mutually independent, find the distribution of K.

4. **(Joint densities)** Let the joint density function of (X, Y) be

$$f(x,y) = \begin{cases} 3xy(x+y), & \text{if } (x,y) \in [0,1]^2, \\ 0, & \text{else.} \end{cases}$$

   Calculate the covariance $\mathrm{Cov}(X, Y)$.

5. **(Transformation of random variables)**

   (a) Suppose X has the Cauchy distribution with density:

$$f_X(x) := \frac{1}{\pi(1+x^2)}.$$

   Show that $1/X$ has the same distribution as X.

   (b) Suppose $Y \sim \mathrm{Exp}(1)$. Find a function $g : (0,\infty) \to (-\infty,\infty)$ such that $g(Y)$ has the Cauchy distribution with density given by (a).

   (c) Suppose $Z \sim \mathrm{Exp}(\lambda)$, where $\lambda > 0$. Show that the distribution of $W := \lceil Z \rceil$ (here $\lceil z \rceil$ is the smallest integer that is larger than or equal to z) is Geometric. Explicitly express the parameter of the Geometric distribution in terms of λ.

6. **(Transformation of random variables)** Suppose $X \sim \mathrm{Uniform}[-\pi, 2\pi]$. Find the p.d.f. of $Y = \sin(X)$.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 4

Due: 10:00pm, Tuesday, November 5, 2024 (via Gradescope)

1. **(Order statistics)** Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exp(λ) distribution, where $\lambda > 0$, and let $X_{(i)}$ be the order statistics for $i = 1, \ldots, n$.
   (a) Find the distribution of $X_{(1)}$.
   (b) Using the memoryless property, find the distribution of $X_{(i+1)} - X_{(i)}$ for $i = 1, \ldots, n-1$.
   (c) Use the previous item to show that each $X_{(i)}$ has the same distribution as a sum of i independent random variables.
   (d) Calculate the expectation and the variance of $X_{(i)}$ for $i = 1, \ldots, n$.

2. **(Joint and conditional densities)** Let X, Y be two random variables with the following properties. Y has density function $f_Y(y) = 3y^2$ for $0 < y < 1$ and zero elsewhere. For $0 < y < 1$, given that $Y = y$, X has conditional density function $f_{X|Y}(x|y) = \frac{2x}{y^2}$ for $0 < x < y$ and zero elsewhere.
   (a) Find the joint density function $f_{X,Y}(x, y)$ of X, Y. Be precise about the values (x, y) for which your formula is valid. Check that the joint density function you find integrates to 1.
   (b) Find the conditional density function of Y, given $X = x$. Be precise about the values of x and y for which the answer is valid. Identify the conditional distribution of Y by name.

3. **(Model selection)** Given data $x_1, \ldots, x_n$, consider the problem of selecting between the two models:
$$\text{Model One}: X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(0, 1)$$
and
$$\text{Model Two}: X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, 1) \text{ for an unknown } \mu.$$
To use probability to solve this problem, let us introduce an additional random variable Θ that has the Bernoulli distribution with parameter 0.5. Assume that the conditional distribution of $X_1, \ldots, X_n$ given $\Theta = \theta$ is given by the following
$$X_1, \ldots, X_n \mid \Theta = 0 \overset{\text{i.i.d}}{\sim} N(0, 1)$$
and
$$X_1, \ldots, X_n \mid \mu, \Theta = 1 \overset{\text{i.i.d}}{\sim} N(\mu, 1) \text{ and } \mu \mid \Theta = 1 \sim N\left(0, \tau^2\right).$$
Here τ is a parameter which you can treat as a fixed constant in this exercise.
   (a) Using the formula
$$f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n) = \int f_{X_1,\ldots,X_n|\mu,\Theta=1}(x_1, \ldots, x_n)\, f_{\mu|\Theta=1}(\mu) d\mu$$
   prove that
$$f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{1+n\tau^2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2}\right) \exp\left(\frac{n^2\tau^2\bar{x}^2}{2\left(1+n\tau^2\right)}\right),$$
   where $\bar{x}$ is the mean of $x_1, \ldots, x_n$.

(b) Calculate the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$.

(c) Intuitively, we would prefer Model Two over Model One when $\bar{x}$ is far from zero. Is this intuition reflected in your conditional distribution from the previous part?

4. **(Gamma-Poisson)** Consider random variables $\Theta, X_1, \ldots, X_n$ such that

$$\Theta \sim \text{Gamma}(\alpha, \lambda) \quad \text{and} \quad X_1, \ldots, X_n \mid \Theta = \theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$$

(a) Find the conditional distribution of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

(b) Find $\mathbb{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$.

(c) Write $\mathbb{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$ as a weighted linear combination of $(x_1 + \cdots + x_n)/n$ and the mean of the marginal distribution (i.e., prior mean) of Θ and argue that the weight of the prior mean goes to zero as $n \to \infty$.

5. **(Law of total expectation)** Let the joint probability mass function (p.m.f.) of (X, Y) be

$$p_{X,Y}(k,n) = \begin{cases} \dfrac{1}{n+1}\left(1 - \dfrac{1}{n+1}\right)^{k-1} \dfrac{1}{2^n}, & \text{for } 1 \le n < \infty \text{ and } 1 \le k < \infty, \\ 0, & \text{else.} \end{cases}$$

(a) Find the p.m.f. $p_Y(n)$ of Y and the conditional p.m.f $p_{X|Y}(k|n)$.

(b) Calculate $\mathbb{E}[Y]$.

(c) Find the conditional expectation $\mathbb{E}[X|Y]$.

(d) Use parts (a) and (c) to calculate $\mathbb{E}[X]$.

6. **(Expected number of coin tosses)** Consider a sequence of coin tosses.

(a) On average, how many tosses of a fair coin does it take to see two heads in a row?

(b) How many tosses on average to see the sequence HTH for the first time?

(c) How does our answer changes if we have an unfair coin?

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 5

Due: 10:00pm, Wednesday, November 20, 2024 (via Gradescope)

1. **(Multivariate normal)** Suppose $Y \sim \mathcal{N}_n(\mu, \Sigma)$ in this problem.

   (a) If a is any fixed vector in $\mathbb{R}^n$, show that

   $$\frac{a^T(Y-\mu)}{\sqrt{a^T\Sigma a}} \sim \mathcal{N}(0,1).$$

   (b) If A is now a random vector that is independent of Y, then show again that

   $$\frac{A^T(Y-\mu)}{\sqrt{A^T\Sigma A}}$$

   is distributed according to $\mathcal{N}(0,1)$ and that it is independent of A.

   (c) Using the above result, show that if $Y \sim \mathcal{N}_3(0, I_3)$, then

   $$\frac{Y_1 e^{Y_3} + Y_2 \log|Y_3|}{\sqrt{e^{2Y_3} + (\log|Y_3|)^2}} \sim \mathcal{N}(0,1).$$

2. **(Marginally normal but not bivariate normal)** Give an example of a 2×1 random vector $Y = (Y_1, Y_2)^T$ with positive definite covariance matrix such that each Y_1 and Y_2 is standard normal but Y is not bivariate normal.

3. **(Conditional distribution)** Consider three random variables Y_1, Y_2 and Y_3 that are independent and standard normal. Let

   $$X_1 = Y_2 + Y_3,$$
   $$X_2 = Y_1 + Y_3,$$
   $$X_3 = Y_1 + Y_2.$$

   Find the conditional distribution of X_1 given $X_2 = X_3 = 0$.

4. **(More on jointly Gaussian distributions)** Let X and Y be independent standard normal variables.

   (a) For a constant k, find $\mathbb{P}(X > kY)$.
   (b) If $U = \sqrt{3}X + Y$, and $V = X - \sqrt{3}Y$, find $\mathbb{P}(U > kV)$.
   (c) Find $\mathbb{P}(U^2 + V^2 < 1)$.
   (d) Find the conditional distribution of X given $V = v$.

5. **(Wigner's surmise)** Let $X = \begin{pmatrix} X_1 & X_3 \\ X_3 & X_2 \end{pmatrix}$ with X_1 and X_2 independent $\mathcal{N}(0,1)$ and X_3 another independent $\mathcal{N}(0, 1/2)$. Let λ_1 and λ_2 be two eigenvalues of X and $s = |\lambda_1 - \lambda_2|$.

(a) Prove that $s = \sqrt{(X_1 - X_2)^2 + 4X_3^2}$.

(b) Find the density of s.

(c) Plot the density function of s. What do you observe respect to the eigenvalues of the random matrix X?

6. **(1D Gaussian process)** In this problem, you will implement a 1D Gaussian process that predicts outputs based on noisy training data. You will be given (noisy) 1D training data pairs $D_{\text{train}} = \{(x_1, y_1), (x_2, y_2) \ldots\}$. Your task is to predict the output for a set of test queries $D_{\text{test}} = \{x_1^*, x_2^*, \ldots\}$, conditioned on the training data. Implement two separate kernel functions, namely the

   - **Squared Exponential Kernel**: This is the kernel we discussed in class.

$$k(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{(x_i - x_j)^T M (x_i - x_j)}{2}\right)$$

     where σ_f is a scale factor for the kernel and M is a metric measuring distance between two input vectors. In the 1D case, $M = \frac{1}{l^2}$ where l is the length scale of the kernel.

   - **Matérn Kernel**: This kernel is used commonly in many machine learning applications.

$$k(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu} r}{l}\right)$$

     where ν and l are (positive) parameters of the kernel and $r = |x_i - x_j|$. K_ν is a modified bessel function and Γ is the gamma function. Good parameters settings for ν are 0.25 - 3. You can use scipy.special.kv() in Python or besselK() in R for implementing K_ν.

   (a) Implement the squared exponential and Matérn kernel functions to compute similarity between any pair of inputs. The output for each function should be a kernel matrix K.

   (b) Using your kernel functions, implement a Gaussion process regression function to predict the posterior mean and variance of test data $\vec{y}^*$.

   (c) The simulation function and plotting function are provided in the file `ps5_GP_1D.ipynb`. Vary the kernel parameters (e.g., σ_f, l, and ν) and observe how they affect the predictive mean and variance. What impact do these parameters have on the smoothness and uncertainty of your GP predictions?

   Note: It's recommended to use Python (Jupyter notebook) and submit a pdf file including code, plots and comments. If you prefer using another coding language, please make sure the data simulation is the same with the provided code.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 6

Due: 10:00pm, Friday, December 6, 2024 (via Gradescope)

1. (**Branching process)** A branching process starts with one individual, i.e. $X(0) = 1$, who reproduces according to the following principle:

# of children	0	1	2
probability	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Individuals reproduce independently of each other and independently of the number of their sisters and brothers. Determine

   (a) the probability that the population becomes extinct;

   (b) the probability that the population has become extinct in the second generation, i.e. $\mathbb{P}(X(2) = 0)$;

   (c) the expected number of children given that there are no grandchildren.

2. (**Random walk)** Random walk on $\{0, 1, 2, 3\}$. Consider the Markov chain (X_n) with transition matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

started with $X_0 = 0$. Define T_j as $\min\{n \geq 1 : X_n = j\}$. Find explicitly the following distributions and expectations.

   (a) The distribution of X_2.

   (b) The limit distribution of X_n as $n \to \infty$.

   (c) $\mathbb{E}[T_0]$

   (d) $\mathbb{E}[T_3]$

   (e) $\mathbb{P}[T_3 < T_0]$

3. (**The average number of jobs)** Jennifer is employed for one day at a time. When she is out of work, she visits the job agency in the morning to see if there is work for that day. There is a job for her with probability 1/2. If there is no work, she comes back the next day. When she has a job, she will be called back to the same job for the next day with probability 2/3. When she is not called back, she goes to the job agency again the next morning to look for a new job that she had not had previously. Approximate the average number of jobs Jennifer works in a year.

4. (**Rain or no rain)** Suppose that at day 0 it is not raining. Then each new day, if it rained yesterday, it will rain with probability 0.7; if it did not rain yesterday, it will rain with probability 0.2.

   (a) Find the stationary distribution.

(b) How many days should we expect to wait to have rain for the first time?

5. (**The game of roulette)** A gambler plays the game of roulette, betting X dollars on red or black. The gambler wins X dollars with probability $p = 18/38$ or loses the bet with probability $q = 20/38$. Suppose that the gambler starts the game with \$500 in his pocket and upper limit on winnings is \$1000.

   (a) Compute the probability of the gambler's ruin for $X = \$10$.

   (b) Compute the probability of the gambler's ruin for $X = \$100$.

   (c) Compare the above results with the probability of ruin in the case the gambler bets everything on a single turn of the wheel.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 1 Solutions

1. (**Basic probability**) Assume that $\mathbb{P}(A) = 0.6$, $\mathbb{P}(B) = 0.7$ and $\mathbb{P}(C) = 0.8$.

   (a) Show that $0.3 \leq \mathbb{P}(A \cap B) \leq 0.6$.
   For the second inequality, since $A \cap B \subseteq A$ then $\mathbb{P}(A \cap B) \leq \mathbb{P}(A) = 0.6$. For the first inequality note that $\mathbb{P}(A \cup B) \leq 1$. Using the principle of inclusion-exclusion on B and C we have that

   $$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$$
   $$\geq 0.6 + 0.7 - 1 = 0.3$$

   We conclude that $0.3 \leq \mathbb{P}(A \cap B) \leq 0.6$.

   (b) Show that $0.1 \leq \mathbb{P}(A \cap B \cap C) \leq 0.6$.
   For the second inequality, since $A \cap B \cap C \subseteq A$ then $\mathbb{P}(A \cap B \cap C) \leq \mathbb{P}(A) = 0.6$. Note that $\mathbb{P}((A \cap B) \cup C) \leq 1$. Using the principle of inclusion-exclusion again on C and $A \cap B$ we have that

   $$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \cap B) + \mathbb{P}(C) - \mathbb{P}((A \cap B) \cup C)$$
   $$\geq 0.3 + 0.8 - 1 = 0.1$$

2. **(Independence)** Suppose we roll an unbiased six-sided die $n \geq 3$ times. Let E_{ij} denote the event that the ith and the jth rolls produce the same number. Show that the events $\{E_{ij} \mid 1 \leq i < j \leq n\}$ are pairwise independent but not independent as a family.
   Remark that $\mathbb{P}(E_{ij}) = 1/6$. We also have that $\mathbb{P}(E_{ij} \cap E_{k\ell}) = 1/36$ and $\mathbb{P}(E_{ij} \cap E_{ik}) = 1/36$. Since $\mathbb{P}(E_{ij} \cap E_{k\ell}) = \mathbb{P}(E_{ij})\mathbb{P}(E_{k\ell})$ in all cases, we conclude that the events are pairwise independent. On the other hand, remark that $\mathbb{P}(E_{12})\mathbb{P}(E_{13})\mathbb{P}(E_{23}) = 1/6^3$ while $\mathbb{P}(E_{12} \cap E_{13} \cap E_{23}) = 1/6^2$. Hence the events are not independent.

3. (**Expectation, joint distribution, uniform distribution**) Let X be a random variable with values $\{1, 2\}$ and Y a random variable with values $\{0, 1, 2\}$. Initially we have the following partial information about their joint probability mass function.

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 1$	$1/8$		
$X = 2$		0	

   Subsequently we learn that $\mathbb{E}[XY] = \frac{13}{9}$ and that Y has uniform distribution. Use this information to fill in the missing values of the joint probability mass function table.

   The missing values on the table are $a = \mathbb{P}(X = 1, Y = 1)$, $b = \mathbb{P}(X = 1, Y = 2)$, $c = \mathbb{P}(X = 2, Y = 0)$ and $d = \mathbb{P}(X = 2, Y = 2)$. We know this must be a joint PMF so

   $$1/8 + a + b + c + d = 1.$$

We also know that

$$\mathbb{E}[XY] = a + 2b + 4d = 13/9,$$

and since Y is uniform we have that

$$1/8 + c = a = b + d.$$

Using the last equation on the first two equations we obtain $3b + 3d = 1$ and $3b + 5d = 13/9$. By solving the system of equations we obtain $b = 1/9$, $d = 2/9$ and finally using the last equation again we conclude $a = 1/3$ and $c = 5/24$.

	Y=0	Y=1	Y=2
X=1	1/8	**1/3**	**1/9**
X=2	**5/24**	0	**2/9**

4. (**Conditioning, cumulative distribution function**) You flip a fair coin. If you get tails, you choose a uniformly random number on the interval $[0, 2]$. If you get heads, you choose the number 1. Let X be the random variable describing the outcome of that experiment.

   (a) Using the law of total probabilities, calculate $\mathbb{P}(X \leq 1/2)$ and $\mathbb{P}(X \leq 3/2)$.

   (b) Find the cumulative distribution function F_X of X.

   (c) Is X a discrete random variable? Is X a continuous random variable?

   Let T be the event in which we got tails and H be the event in which we got heads.

   (a) We have that

$$\begin{aligned}\mathbb{P}(X \leq 1/2) &= \mathbb{P}(X \leq 1/2|T)\mathbb{P}(T) + \mathbb{P}(X \leq 1/2|H)\mathbb{P}(H) \\ &= \frac{1}{4} \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{1}{8},\end{aligned}$$

   and that

$$\begin{aligned}\mathbb{P}(X \leq 3/2) &= \mathbb{P}(X \leq 3/2|T)\mathbb{P}(T) + \mathbb{P}(X \leq 3/2|H)\mathbb{P}(H) \\ &= \frac{3}{4} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{7}{8}.\end{aligned}$$

   (b) We want to find $F_X(s) = \mathbb{P}(X \leq s)$. We will proceed exactly as in part 1. If $s < 0$ then directly $\mathbb{P}(X \leq s) = 0$. If $s > 2$ then directly $\mathbb{P}(X \leq s) = 1$. If $0 < s < 1$ then

$$\begin{aligned}\mathbb{P}(X \leq s) &= \mathbb{P}(X \leq s|T)\mathbb{P}(T) + \mathbb{P}(X \leq s|H)\mathbb{P}(H) \\ &= \frac{s}{2} \times \frac{1}{2} + 0 \times \frac{1}{2} = \frac{s}{4},\end{aligned}$$

   If $1 \leq s \leq 2$ then

$$\begin{aligned}\mathbb{P}(X \leq s) &= \mathbb{P}(X \leq s|T)\mathbb{P}(T) + \mathbb{P}(X \leq s|H)\mathbb{P}(H) \\ &= \frac{s}{2} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{2+s}{4}.\end{aligned}$$

(c) Following the definitions given in lecture, this is neither a continuous or a discrete random variable. It is not continuous since we have $\mathbb{P}(X = 1) = 1/2 \neq 0$. It is not discrete since the sum of probabilities of possible values X can take with positive probability is $1/2$ instead of 1. There are other ways to argue. For example showing that F_X is not continuous, that X don't have a p.d.f., that the cardinality of possible values X can take is infinite uncountable, etc.

5. (**Bounding even moments**) Let X be a random variable. Show that $\mathbb{E}[X^{2k}] \geq (\mathbb{E}[X])^{2k}$ for all positive integers k. This is a direct application of Jensen's inequality with the function $\varphi(x) = x^{2k}$. To verify that φ is convex we can calculate the second derivative and verify it is nonnegative.

6. (**Continuous distributions, probability density function, independence**) Pick a uniformly chosen random point (X, Y) inside the sector delimited by the x-axis, the y-axis and the parabola given by the equation $y = 1 - x^2$; see Figure 1.

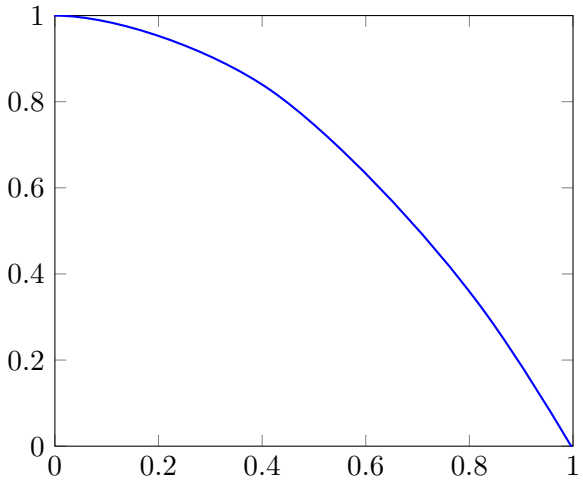


Figure 1: Graph of $y = 1 - x^2$

(a) Verify that the area of that sector is $2/3$.
(b) What is the probability that the distance of this point to the y-axis is **less** than $1/2$?
(c) What is the probability that the distance of this point to the origin is **more** than $1/2$?
(d) Find the p.d.f. of X.
(e) Find the p.d.f. of Y.
(f) Are X and Y independent?

(a) Calculate $\int_0^1 1 - x^2\, dx = [x - x^3/3]_0^1 = 2/3$.
(b) Let A be the described event, given that we are choosing a point uniformly, the value of $\mathbb{P}(A)$ is given by the ratios of area described in event A and the total area of the delimited sector. Given that, let's note that for the distance between the y-axis and the point to be less than $1/2$ then the point must be in the sector delimited by the y-axis, the x-axis, the equation $y = 1 - x^2$ and the line $x = 1/2$. The area of this sector is given by $\int_0^{1/2} 1 - x^2\, dx = \frac{11}{24}$. Finally $\mathbb{P}(A) = \frac{11/24}{2/3} = \frac{11}{16}$.

(c) We proceed as in part 2., let B be the described event. the are we are looking for correspond to the area of the original sector minus a quarter of disk of radius 1/2. More precisely $\mathbb{P}(B) = \dfrac{2/3 - \pi/16}{2/3} = 1 - \dfrac{3\pi}{32} \approx 0.705...$

(d) The pdf of X is the only function $p_X(t)$ such that $P(a \leq X \leq b) = \int_a^b p_X(t)\,dt$. We get $p_X(t) = \frac{3}{2}(1-t^2)$ for $0 \leq t \leq 1$ and 0 in other case.

(e) Similarly, the pdf of Y is $p_Y(t) = \frac{3}{2}(\sqrt{1-t})$ for $0 \leq t \leq 1$ and 0 in other case.

(f) Just taking $I = [4/5, 1]$, remark that $\mathbb{P}(X \in I) \neq 0$ and $\mathbb{P}(Y \in I) \neq 0$, however $\mathbb{P}(X \in I, Y \in I) = 0$. Since $\mathbb{P}(X \in I, Y \in I) = 0 \neq \mathbb{P}(X \in I)\mathbb{P}(Y \in I)$, we conclude that X and Y are not independent.

7. (**Events, indicators and basic probability inequalities**) Recall that for an event A, we denote the corresponding indicator random variable by $I(A)$ (i.e., $I(A)$ takes value 1 when A occurs and the value 0 when A does not occur). Also recall that the probability $\mathbb{P}(A)$ of A equals the expectation of the random variable $\mathbb{E}(I(A))$.

(a) Given events $A_1, \ldots, A_n$, show that $I\left(\cup_{i=1}^n A_i\right) = \max_{1 \leq i \leq n} I\left(A_i\right)$.
Recall the definition of the indicator function $I(A)$ for an event A :

$$I(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur} \end{cases}$$

The event $\cup_{i=1}^n A_i$ occurs if at least one of the A_i occurs, meaning that:

$$I\left(\cup_{i=1}^n A_i\right) = \begin{cases} 1 & \text{if at least one } A_i \text{ occurs} \\ 0 & \text{if none of the } A_i \text{ occurs.} \end{cases}$$

The maximum of the individual indicators:

$$\max_{1 \leq i \leq n} I\left(A_i\right) = \begin{cases} 1 & \text{if } I\left(A_i\right) = 1 \text{ for at least one } i \\ 0 & \text{if } I\left(A_i\right) = 0 \text{ for all } i \end{cases}$$

By definition, $I\left(A_i\right) = 1$ if and only if event A_i occurs. Therefore, the maximum $\max_{1 \leq i \leq n} I\left(A_i\right)$ takes the value 1 if at least one of the events A_i occurs, and 0 if none of the events occur. This shows

$$I\left(\cup_{i=1}^n A_i\right) = \max_{1 \leq i \leq n} I\left(A_i\right).$$

(b) Using the fact observed above (and the following ordering property of expectation: $X \leq Y$ implies that $\mathbb{E}(X) \leq \mathbb{E}(Y)$), show that

$$\mathbb{P}\left(\cup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}\left(A_i\right).$$

Note: This is known as the union bound and used quite frequently.
We know that for any collection of non-negative random variables $X_1, X_2, \ldots, X_n$, the maximum of these random variables is always less than or equal to the sum:

$$\max_{1 \leq i \leq n} X_i \leq \sum_{i=1}^n X_i$$

Applying this to our indicator random variables and from result from a) , we have:

$$I\left(\cup_{i=1}^{n} A_i\right) = \max_{1 \le i \le n} I\left(A_i\right) \le \sum_{i=1}^{n} I\left(A_i\right)$$

We then have

$$\begin{aligned}
\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) &= \mathbb{E}\left[I\left(\cup_{i=1}^{n} A_i\right)\right] \\
&\le \mathbb{E}\left[\sum_{i=1}^{n} I\left(A_i\right)\right] \quad \text{(i)} \\
&= \sum_{i=1}^{n} \mathbb{E}\left[I\left(A_i\right)\right] \quad \text{(ii)} \\
&= \sum_{i=1}^{n} \mathbb{P}\left(A_i\right)
\end{aligned}$$

where (i) follows the ordering property of expectation and (ii) holds because of linearity of expectation.

(c) For every event A, show that $I\left(A^c\right) = 1 - I(A)$ where A^c denotes the event that A does not occur.
By definition:

$$I(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

$$I\left(A^c\right) = \begin{cases} 1 & \text{if } A^c \text{ occurs (i.e., } A \text{ does not occur)} \\ 0 & \text{if } A^c \text{ does not occur (i.e., } A \text{ occurs)} \end{cases}$$

Thus, we observe that $I\left(A^c\right) = 1$ when $I(A) = 0$, and $I\left(A^c\right) = 0$ when $I(A) = 1$. Hence,

$$I\left(A^c\right) = 1 - I(A).$$

(d) For events $A_1, \ldots, A_n$, show that $I\left(\cap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} I\left(A_i\right)$.
The indicator $I\left(\cap_{i=1}^{n} A_i\right)$ is 1 if and only if all events A_i occur simultaneously, otherwise it is 0. This can be expressed as

$$I\left(\cap_{i=1}^{n} A_i\right) = \begin{cases} 1 & \text{if } I\left(A_i\right) = 1 \text{ for all } i \\ 0 & \text{if } I\left(A_i\right) = 0 \text{ for some } i \end{cases}$$

This is exactly the product of the indicators:

$$I\left(\cap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} I\left(A_i\right)$$

(e) Prove the inclusion-exclusion formula: For events $A_1, \ldots, A_n$,

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \Sigma_1 - \Sigma_2 + \Sigma_3 - \Sigma_4 + \cdots + (-1)^{n-1} \Sigma_n$$

where

$$\Sigma_k := \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}\left(A_{i_1} A_{i_2} \cdots A_{i_k}\right).$$

(Approach 1: induction) **Base cases.** For $n = 1$, the formula is simply $\mathbb{P}(A_1) = \mathbb{P}(A_1)$ and for $n = 2$, the formula is

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2),$$

which holds as the standard inclusion-exclusion formula for the union of two events. Hence, the base case holds for both $n = 1$ and $n = 2$.
Inductive Step. Assume that the formula holds for $n = k$. That is, assume

$$\mathbb{P}\left(\cup_{i=1}^{k} A_i\right) = \sum_{1 \le i \le k} \mathbb{P}(A_i) - \sum_{1 \le i_1 < i_2 \le k} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) + \ldots + (-1)^{k-1} \mathbb{P}\left(\cap_{i=1}^{k} A_i\right).$$

We need to show that the formula also holds for $n = k + 1$, i.e.,

$$\mathbb{P}\left(\cup_{i=1}^{k+1} A_i\right) = \sum_{1 \le i \le k+1} \mathbb{P}(A_i) - \sum_{1 \le i_1 < i_2 \le k+1} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) + \ldots + (-1)^{k} \mathbb{P}\left(\cap_{i=1}^{k+1} A_i\right).$$

We can express the union of the $k + 1$ events as

$$\begin{aligned}
\mathbb{P}\left(\cup_{i=1}^{k+1} A_i\right) &= \mathbb{P}\left(\left(\cup_{i=1}^{k} A_i\right) \cup A_{k+1}\right) \\
&= \mathbb{P}\left(\cup_{i=1}^{k} A_i\right) + \mathbb{P}(A_{k+1}) - \mathbb{P}\left(\left(\cup_{i=1}^{k} A_i\right) \cap A_{k+1}\right) \\
&= \mathbb{P}\left(\cup_{i=1}^{k} A_i\right) + \mathbb{P}(A_{k+1}) - \mathbb{P}\left(\cup_{i=1}^{k} \left(A_i \cap A_{k+1}\right)\right).
\end{aligned}$$

As we assume the formula holds for $n = k$, we can expand $\mathbb{P}\left(\cup_{i=1}^{k} \left(A_i \cap A_{k+1}\right)\right)$ as

$$\mathbb{P}\left(\cup_{i=1}^{k} \left(A_i \cap A_{k+1}\right)\right) = \sum_{1 \le i \le k} \mathbb{P}\left(A_i \cap A_{k+1}\right) - \sum_{1 \le i_1 < i_2 \le k} \mathbb{P}\left(A_{i_1} \cap A_{i_2} \cap A_{k+1}\right) + \cdots + (-1)^{k-1} \mathbb{P}\left(\cap_{i=1}^{k+1} A_i\right).$$

Substituting this into the expression for $\mathbb{P}\left(\cup_{i=1}^{k+1} A_i\right)$, we obtain

$$\mathbb{P}\left(\cup_{i=1}^{k+1} A_i\right) = \sum_{1 \le i \le k+1} \mathbb{P}(A_i) - \sum_{1 \le i_1 < i_2 \le k+1} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right) + \ldots + (-1)^{k} \mathbb{P}\left(\cap_{i=1}^{k+1} A_i\right).$$

This shows that the formula holds for $n = k + 1$, completing the induction.
(Approach 2: Direct proof) From parts (c) and (d) it is clear that:

$$\begin{aligned} I\left(\cup_{i=1}^{n} A_i\right) &= 1 - I\left(\cap_{i=1}^{n} A_i^c\right) \\ &= 1 - \prod_{i=1}^{n} I\left(1 - A_i\right) \\ &\text{(expand this product)} \\ &= 1 - \left(1 - \sum_{i=1}^{n} I\left(A_i\right) + \sum_{i<j} I\left(A_i, A_j\right) - \sum_{i<j<k} I\left(A_i, A_j, A_k\right) + \ldots\right) \\ &= \sum_{i=1}^{n} I\left(A_i\right) - \sum_{i<j} I\left(A_i, A_j\right) + \sum_{i<j<k} I\left(A_i, A_j, A_k\right) - \ldots \end{aligned}$$

Taking the expectation of both sides:

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}\left(A_i\right) - \sum_{i<j} \mathbb{P}\left(A_i, A_j\right) + \sum_{i<j<k} \mathbb{P}\left(A_i, A_j, A_k\right) - \ldots$$

Note that the RHS can be written more simply such as:

$$\Sigma_k = \sum_{i \le i_1 < i_2 < \cdots < i_k \le n} \mathbb{P}\left(A_{i_1}, A_{i_2}, \ldots, A_{i_k}\right)$$

Thus

$$\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) = \Sigma_i - \Sigma_2 + \Sigma_3 - \Sigma_4 + \ldots + (-1)^{n-1}\Sigma_n$$

8. (**Hypergeometric and exchangeability**) We have an urn with R red balls and $N - R$ white balls, where $0 < R < N$. We draw n balls in sequence from the urn without replacement. Let R_i denote the proposition that the i^{th} draw results in a red ball.

   (a) Calculate $\mathbb{P}\left(R_i\right)$ for each $i = 1, \ldots, n$.
   Since there are R red balls out of N total balls, we have

   $$\mathbb{P}\left(R_1\right) = \frac{R}{N}.$$

   By exchangability, it follows that when we consider one i at a time, we have

   $$\mathbb{P}\left(R_i\right) = \frac{R}{N}, \quad \forall i \in \{1, \ldots, n\}.$$

   Exchanging the order in which we consider i does not change the underlying distribution.

   (b) Show that $\mathbb{P}\left(R_j \mid R_k\right) = \mathbb{P}\left(R_k \mid R_j\right)$ for every $1 \le j, k \le n$.
   Consider the definition of conditional probabilities:

   $$\begin{aligned} \mathbb{P}\left(R_i \mid R_j\right) &= \frac{\mathbb{P}\left(R_i \cap R_j\right)}{\mathbb{P}\left(R_j\right)} \\ &= \frac{\mathbb{P}\left(R_i \cap R_j\right)}{\mathbb{P}\left(R_i\right)} \quad \text{(by part (a))} \\ &= \mathbb{P}\left(R_j \mid R_i\right) \end{aligned}$$

(c) Calculate $\mathbb{P}\left(R_k \mid \bigcup_{i=k+1}^{n} R_i\right)$ for a fixed $1 \leq k < n$.
For fixed $1 \leq k < n$,

$$\begin{aligned}
\mathbb{P}\left(R_k \mid \bigcup_{i=k+1}^{n} R_i\right) &= \frac{\mathbb{P}\left(R_k \cap \left(\bigcup_{i=k+1}^{n} R_i\right)\right)}{\mathbb{P}\left(\bigcup_{i=k+1}^{n} R_i\right)} \\
&= \frac{\mathbb{P}\left(R_1 \cap \left(\bigcup_{i=k+1}^{n} R_i\right)\right)}{\mathbb{P}\left(\bigcup_{i=1}^{n-k} R_i\right)} \quad (\text{ by exchangability }) \\
&= \frac{\mathbb{P}\left(R_1\right)\mathbb{P}\left(\bigcup_{i=2}^{n-k+1} R_i \mid R_1\right)}{\mathbb{P}\left(\bigcup_{i=1}^{n-k} R_i\right)} \\
&= \frac{\left(\frac{R}{N}\right)\left(1 - \mathbb{P}\left(\begin{smallmatrix}\text{draw } n-k \text{ white balls} \\ \text{from a urn with } N-1 \text{ balls}\end{smallmatrix}\right)\right)}{1 - \mathbb{P}\left(\begin{smallmatrix}\text{all of the first } n-k \\ \text{draws are are white}\end{smallmatrix}\right)} \\
&= \frac{\left(\frac{R}{N}\right)\left[1 - \frac{\binom{N-R}{n-k}}{\binom{N-1}{n-k}}\right]}{1 - \frac{\binom{N-R}{n-k}}{\binom{N}{n-k}}}
\end{aligned}$$

(d) Let X be the random variable representing the minimum number of draws required to get at least one red ball. Calculate $\mathbb{E}[X]$, the expected value of X. (Hint: Use exchangeability to simplify the calculation.)
Label the white balls as $1, 2, \ldots, N-R$. Define the indicator variable I_j for each white ball j, where $I_j = 1$ if white ball j is drawn before any red ball, and $I_j = 0$ otherwise.
The probability that a specific white ball j is drawn before any red ball is given by

$$\mathbb{P}(I_j = 1) = \frac{1}{R+1}$$

This is because, when considering the order in which one specific white ball and all red balls are drawn, all possible orders are equally likely.
Let Y represent the number of white balls drawn before the first red ball. Then Y is simply the sum of all indicator variables:

$$Y = \sum_{j=1}^{N-R} I_j.$$

Thus, the expected value of Y is

$$\mathbb{E}[Y] = \sum_{j=1}^{N-R} \mathbb{E}[I_j] = \frac{N-R}{R+1}$$

Since we are interested in the expected number of total draws X to get the first red ball, we have $X = Y + 1$ (as the next draw after all white balls have been drawn must be a red ball). Therefore,

$$\mathbb{E}[X] = \mathbb{E}[Y] + 1 = \frac{N+1}{R+1}.$$

(e) Suppose that instead of only two colors, the urn has balls of k different colors: N_1 of color $1, N_2$ of color $2, \ldots, N_k$ of color k. Let $N = N_1 + \cdots + N_k$. Argue that the probability of drawing r_1 balls of color $1, r_2$ balls of color $2, \ldots, r_k$ balls of color k in $n = r_1 + \cdots + r_k$ draws without replacement is given by

$$\frac{\binom{N_1}{r_1} \cdots \binom{N_k}{r_k}}{\binom{N}{n}}.$$

Use the concept of combinatorial counting.

Total Number of Possible Outcomes: The total number of ways to draw n balls from an urn containing N balls (where $N = N_1 + N_2 + \cdots + N_k$) without considering the color is given by the binomial coefficient $\binom{N}{n}$.

Number of Favorable Outcomes: The number of ways to choose r_1 balls from the N_1 balls of color 1 is $\binom{N_1}{r_1}$. Similarly, the number of ways to choose r_2 balls from the N_2 balls of color 2 is $\binom{N_2}{r_2}$, and so on. The total number of ways to achieve this specific configuration is given by the product of these individual combinations

$$\binom{N_1}{r_1} \cdot \binom{N_2}{r_2} \cdot \ldots \cdot \binom{N_k}{r_k}$$

Probability:

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} = \frac{\binom{N_1}{r_1} \cdot \binom{N_2}{r_2} \cdot \ldots \cdot \binom{N_k}{r_k}}{\binom{N}{n}}$$

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 2

Due: 10:00pm, Friday, September 27, 2024 (via Gradescope)

1. **(Binomial tail bounds)** Let S_n have the Binomial(n, p_i) distribution of the number of successes in n independent Bernoulli(p_i) trials. Use a suitable computational environment to evaluate the right tail probabilities

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right)$$

for $n = 100$ and $p_i = i/10$ for $i = 1, 2, \ldots, 9$, and $\epsilon = 1/10$, together with various approximations and upper bounds as indicated. In each case,

- give an exact mathematical formula for the function of i you are computing;
- indicate suitable code for evaluating the formula in your preferred environment and **attach the code at the end of the homework**;
- give the numerical values correct to two significant decimal place.

(a) The exact probabilities.
For a binomial random variable $S_n \sim \text{Binomial}\,(n, p_i)$, the exact probability is

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) = \sum_{k=[n(p_i+\epsilon)]}^{n} \binom{n}{k} p_i^k \left(1 - p_i\right)^{n-k}$$

(b) Markov's upper bounds for these probabilities.
Markov's inequality provides a simple upper bound on the tail probability, which is

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \frac{\mathbb{E}\left[S_n\right]}{n\left(p_i + \epsilon\right)} = \frac{np_i}{n\left(p_i + \epsilon\right)} = \frac{p_i}{p_i + \epsilon}$$

(c) Chebychev's upper bounds for these probabilities (which can be halved for $i = 5$ only: explain why).
Chebyshev's inequality uses the variance of S_n, which is $\text{Var}\left(S_n\right) = np_i\left(1 - p_i\right)$. It provides an upper bound on the tail probability as

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \frac{\text{Var}\left(S_n\right)}{n^2\epsilon^2} = \frac{p_i\left(1 - p_i\right)}{n\epsilon^2}$$

When $p_i = 0.5$, the probability distribution for S_n is symmetric around its expectation $0.5n$

$$\mathbb{P}\left(\left|\frac{S_n}{n} - 0.5\right| \geq \epsilon\right) = \mathbb{P}\left(S_n \geq n(0.5 + \epsilon)\right) + \mathbb{P}\left(S_n \leq n(0.5 - \epsilon)\right)$$

$$= 2\mathbb{P}\left(S_n \geq n(0.5 + \epsilon)\right).$$

When $p_i \neq 0.5$, the distribution is not symmetric and that's why the bound can be halved for $i = 5$ only.

(d) Hoeffding's upper bounds.
Hoeffding's inequality provides an upper bound for the sum of independent bounded random variables, such as a binomial variable S_n. The upper bound is

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right) \leq \exp\left(-2n\epsilon^2\right)$$

(e) Chernoff's upper bounds.
Here we use KL divergence form for Chernoff's bound

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \exp\left(-nD\left(p_i + \epsilon \| p_i\right)\right)$$

where the KL divergence $\mathrm{KL}(q\|p)$ between two Bernoulli distributions with parameters q and p is given by

$$\mathrm{KL}(q\|p) = q\log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

For $q = p_i + \epsilon$ and $p = p_i$, the bound becomes:

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \exp\left(-n\left[\left(p_i+\epsilon\right)\log\left(\frac{p_i+\epsilon}{p_i}\right) + \left(1-p_i-\epsilon\right)\log\left(\frac{1-p_i-\epsilon}{1-p_i}\right)\right]\right)$$

Remark: The proof of equivalence of the standard form and the KL divergence form of Chernoff's bound in Binomial distribution.
The standard form of Chernoff's bound gives an exponential decay for the upper tail of a sum of independent random variables. For $S_n \geq n\left(p_i + \epsilon\right)$, the bound is

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \min_{t>0}\left(\mathbb{E}\left[e^{tS_n}\right] e^{-\operatorname{tn}(p_i+\epsilon)}\right)$$

For binomial random variables, the moment generating function $\mathbb{E}\left[e^{tS_n}\right]$ is

$$\mathbb{E}\left[e^{tS_n}\right] = \left(p_i e^t + (1-p_i)\right)^n$$

Thus, the bound can be expressed as

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \left(p_i e^t + (1-p_i)\right)^n e^{-\operatorname{tn}(p_i+\epsilon)}$$

Optimizing for t using calculus, we get that the right-hand side is minimized if

$$e^t = \frac{\left(1-p_i\right)\left(p_i+\epsilon\right)}{p_i\left(1-p_i-\epsilon\right)}$$

Substituting this back into the bound, we obtain the KL divergence form of the Chernoff bound, which is equivalent to

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \exp\left(-nD\left(p_i + \epsilon \| p_i\right)\right)$$

2. **(LLN)** Suppose that $X_1, X_2, \ldots$ form an i.i.d. sequence of random variables with $\mathbb{E}[X_i] = \mu < \infty$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Evaluate

$$\lim_{n\to\infty} \frac{1}{\binom{n}{2}} \sum_{i,j:1\le i<j\le n} (X_i - X_j)^2.$$

First notice that $\sum_{1\le i<j\le n}(X_i - X_j)^2 = n\sum_{k=1}^n X_k^2 - \left(\sum_{k=1}^n X_k\right)^2$. Now using the law of larges numbers we get

$$\binom{n}{2}^{-1} n\sum_{k=1}^n X_k^2 = \frac{2n}{n-1}\frac{\sum_{k=1}^n X_k^2}{n} \to 2(\sigma^2+\mu^2).$$

and that

$$\binom{n}{2}^{-1} \left(\sum_{k=1}^n X_k\right)^2 = \frac{2n}{n-1}\left(\frac{\sum_{k=1}^n X_k}{n}\right)^2 \to 2\mu^2.$$

We then have that $\binom{n}{2}^{-1}\sum_{1\le i<j\le n}(X_i - X_j)^2 \to 2(\sigma^2+\mu^2-\mu^2) = 2\sigma^2$.

3. **(Chebyshev & CLT)** Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with mean zero and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Determine the limits below, with precise justifications.

   (a) $\lim_{n\to\infty} \mathbb{P}(S_n \ge 0.01n)$.
   (b) $\lim_{n\to\infty} \mathbb{P}(S_n \ge 0)$.
   (c) $\lim_{n\to\infty} \mathbb{P}(S_n \ge -0.01n)$.

   (a) Using Chebyshev's inequality, we get that

$$\mathbb{P}\left(S_n \ge 0.01n\right) \le \frac{n\sigma^2}{n^2/10^4} \to 0.$$

   (b) Using CLT, we get that

$$\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \ge 0\right) \to \frac{1}{2}.$$

   (b) Using Chebyshev's inequality, we get that

$$\mathbb{P}\left(S_n \le -0.01n\right) \le \frac{n\sigma^2}{n^2/10^4} \to 0.$$

   Therefore,

$$\mathbb{P}\left(S_n \ge -0.01n\right) \to 1.$$

4. **(Convolution & MGF)** The Laplace distribution has density $f_Z(z) = \frac{\lambda}{2}\exp(-\lambda|z|)$ and MGF $M_Z(t) = \frac{\lambda^2}{\lambda^2 - t^2}$, where $\lambda > 0$. Let $X, Y \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$. Prove that $Z = X - Y$ follows a Laplace distribution by using:

   (a) Moment generating functions.

(b) The convolution formula.

(a) Our approach is to find the MGF of $Z = X - Y$ and match it to the moment generating function of the Laplace distribution, which is given to be $\frac{\lambda^2}{\lambda^2 - t^2}$. We have:

$$\begin{aligned} M_Z(t) &= M_X(t) M_{-Y}(t) \\ &= M_X(t) \mathbb{E}[e^{-tY}] \\ &= M_X(t) M_Y(-t) \\ &= \frac{\lambda}{\lambda - t} \frac{\lambda}{\lambda - (-t)} = \frac{\lambda^2}{\lambda^2 - t^2}, \end{aligned}$$

and therefore Z is Laplace.

(b) Using the convolution formula:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_{-Y}(z - x)\, dx \\ &= \int_{-\infty}^{\infty} \lambda e^{-\lambda x} I_{(x>0)} \lambda e^{-\lambda(-(z-x))} I_{(z-x<0)} dx \\ &= \int_{-\infty}^{\infty} \lambda e^{-\lambda x} I_{(x>0)} \lambda e^{-\lambda(x-z)} I_{(x-z>0)} dx \\ &= \lambda^2 e^{\lambda z} \int_{-\infty}^{\infty} e^{-2\lambda x} I_{(x>0)} I_{(x-z>0)} dx \end{aligned}$$

We can simplify the indicators:

$$I_{(x>0)} \cdot I_{(x-z>0)} = I_{(x>0)\&(x>z)} = I_{x>\max(0,z)}.$$

Therefore:

$$\begin{aligned} f_Z(z) &= \lambda^2 e^{\lambda z} \int_{\max(z,0)}^{\infty} e^{-2\lambda x} dx \\ &= \lambda^2 e^{\lambda z} \left[-\frac{1}{2\lambda} e^{-2\lambda x} \right]_{x=\max(z,0)}^{x=\infty} \\ &= \lambda^2 e^{\lambda z} \cdot \frac{1}{2\lambda} e^{-2\lambda \max(z,0)} \\ &= \frac{\lambda}{2} e^{\lambda z} \cdot e^{-2\lambda \max(z,0)} \\ &= \frac{\lambda}{2} e^{-\lambda(2\max(z,0) - z)} \\ &= \frac{\lambda}{2} e^{-\lambda|z|} \end{aligned}$$

which is the Laplace(λ) PDF.

5. **(Moments & MGF)** Let X be a random variable with p.d.f. given by

$$f_X(x) = \begin{cases} 2/9, & \text{if } 0 \le x \le 1, \\ (4 - |4 - 2x|)/9, & \text{if } 1 < x \le 4, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Verify that this is actually a p.d.f.

(b) Find the moment generating function of X.

(c) Find $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

(d) Find a formula for the moments of X.

First note that the p.d.f. can also be written as

$$f_X(x) = \begin{cases} 2/9 \text{ if } 0 \le x \le 1 \\ 2x/9 \text{ if } 1 < x \le 2 \\ (8 - 2x)/9 \text{ if } 2 < x \le 4 \\ 0 \text{ else.} \end{cases}$$

(a) Let's verify that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)\,dx &= \int_0^1 2/9\,dx + \int_1^2 2x/9\,dx + \int_2^4 (8-2x)/9\,dx \\ &= \frac{2}{9}\Big([x]_0^1 + [x^2/2]_1^2 + [4x - x^2/2]_2^4 \Big) \\ &= \frac{2}{9}\Big((1-0) + (2 - 1/2) + (8-6) \Big) = 1. \end{aligned}$$

(b) Let's do the full calculation,

$$\begin{aligned} \mathbb{E}[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx \\ &= \frac{2}{9}\Big(\int_0^1 e^{tx}\,dx + \int_1^2 x e^{tx}\,dx + \int_2^4 (4-x)e^{tx}\,dx \Big) \\ &= \frac{2}{9}\Big([\frac{e^{tx}}{t}]_0^1 + [\frac{(tx-1)e^{tx}}{t^2}]_1^2 + [\frac{(1+4t-tx)e^{tx}}{t^2}]_2^4 \Big) \\ &= \frac{2}{9}\Big(\frac{e^t - 1}{t} + \frac{(2t-1)e^{2t} - (t-1)e^t}{t^2} + \frac{(1+4t-4t)e^{4t} - (1+4t-2t)e^{2t}}{t^2} \Big) \\ &= \frac{2}{9} \cdot \frac{e^{4t} + e^t - 2e^{2t} - t}{t^2} \end{aligned}$$

(c) One way to solve this problem is to directly calculate $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ using the formulas $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$ and $\mathbb{E}[X] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx$ and then $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Another way to solve this part is using part 4. We obtain $\mathbb{E}[X] = \frac{49}{27}$, $\mathbb{E}[X^2] = \frac{25}{6}$ and then $\mathrm{Var}(X) = \frac{1273}{1458}$.

(d) Let's use the Taylor expansion for the exponential. We have that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Replacing the series on the formula obtained in part 2. we obtain the following.

$$\begin{aligned}
\mathbb{E}[e^{tx}] &= \frac{2}{9} \cdot \frac{e^{4t} + e^t - 2e^{2t} - t}{t^2} \\
&= \frac{2}{9} \cdot \frac{\sum_{n=0}^{\infty} \frac{4^n t^n}{n!} + \sum_{n=0}^{\infty} \frac{t^n}{n!} - 2\sum_{n=0}^{\infty} \frac{2^n t^n}{n!} - t}{t^2} \\
&= \frac{2}{9} \cdot \frac{\sum_{n=0}^{\infty} \frac{1}{n!}(1 + 4^n - 2^{n+1})t^n - t}{t^2} \\
&= \frac{2}{9} \sum_{n=2}^{\infty} \frac{1 + 4^n - 2^{n+1}}{n!} t^{n-2} = \frac{2}{9} \sum_{n=0}^{\infty} \frac{1 + 4^{n+2} - 2^{n+3}}{(n+2)!} t^n
\end{aligned}$$

We conclude from this that the general formula for the moments is

$$\mathbb{E}(X^n) = \frac{2(1 + 4^{n+2} - 2^{n+3})}{9(n+1)(n+2)}.$$

6. **(Distribution of sums using MGFs)** Let $S_n := X_1 + \cdots + X_n$ for independent $X_1, \ldots, X_n$. Use MGFs to find the distribution of S_n

(a) For X_i with Normal (μ_i, σ_i^2) distribution.
The MGF of a normal random variable $X_i \sim$ Normal (μ_i, σ_i^2) is given by

$$M_{X_i}(t) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right)$$

Thus, for the sum $S_n = X_1 + X_2 + \cdots + X_n$, the MGF is

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right) = \exp\left(t \sum_{i=1}^{n} \mu_i + \frac{t^2}{2} \sum_{i=1}^{n} \sigma_i^2\right)$$

This is the MGF of a normal distribution with mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$. Therefore,

$$S_n \sim \text{Normal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(b) For X_i with Gamma (r_i, λ) distribution.
The MGF of a gamma random variable $X_i \sim$ Gamma (r_i, λ) (where r_i is the shape parameter and λ is the rate parameter) is given by

$$M_{X_i}(t) = \left(1 - \frac{t}{\lambda}\right)^{-r_i}, \quad \text{for } t < \lambda$$

Thus, for the sum $S_n = X_1 + X_2 + \cdots + X_n$, the MGF is

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \left(1 - \frac{t}{\lambda}\right)^{-r_i} = \left(1 - \frac{t}{\lambda}\right)^{-\sum_{i=1}^{n} r_i}$$

This is the MGF of a gamma distribution with shape parameter $\sum_{i=1}^{n} r_i$ and rate parameter λ. Therefore,

$$S_n \sim \text{Gamma}\left(\sum_{i=1}^{n} r_i, \lambda\right)$$

(c) For $X_i = Z_i^2$ with $Z_i \sim \text{Normal}(0, 1)$.
The MGF of X_i is the expectation of $e^{tZ_i^2}$. Applying the density function for Normal distribution, we have

$$\begin{aligned} M_{X_i(t)} &= \mathbb{E}\left[e^{tZ_i^2}\right] = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}(1-2t)} dz \\ &= \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{2\pi}{1-2t}} \\ &= \frac{1}{\sqrt{1-2t}}, \end{aligned}$$

which is valid for $t < \frac{1}{2}$. This is exactly the MGF for χ_1^2.

$$M_{S_n}(t) = \left((1-2t)^{-\frac{1}{2}}\right)^n = (1-2t)^{-\frac{n}{2}}, \quad \text{for } t < \frac{1}{2}$$

Therefore, S_n follows a chi-squared distribution with n degrees of freedom $S_n \sim \chi_n^2$.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 3

Due: 10:00pm, Friday, October 11, 2024 (via Gradescope)

1. **(Approximating Binomial Distributions)** The goal of this question is to empirically verify three approximations to the exact Binomial probability $\mathbb{P}(X = k)$, where $X \sim \text{Binomial}(n, p)$:

   - $\mathbb{P}(Y = k)$, where $Y \sim \text{Poisson}(np)$, the Poison approximation with rate parameter np;
   - The normal approximation

   $$\phi(k; np, np(1-p)) := \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left\{-\frac{(k-np)^2}{2np(1-p)}\right\}$$

   - The entropic approximation

   $$\text{Ent}(k; n, p) := \frac{1}{\sqrt{2\pi nf(1-f)}} \exp(-n\text{KL}(f\|p))$$

   where $f = \frac{k}{n}$ and $\text{KL}(f\|p) = f \log \frac{f}{p} + (1-f) \log \frac{1-f}{1-p}$.

   (a) Take $n = 30$ and $p = 0.05$. Create a table (31 rows and 3 columns) containing the absolute errors for each approximation,

   $$|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|, \quad |\mathbb{P}(X = k) - \phi(k; np, np(1-p))|$$

   and

   $$|\mathbb{P}(X = k) - \text{Ent}(k; n, p)|$$

   for $k = 0, 1, \ldots, 30$. (Note: The entropic approximation does not exist for $k = 0$ and $k = 30$, so only list it for $k = 1, \ldots, 29$).
   Based on the table, comment on the accuracy of each of the three approximations for the Binomial distribution.

   (b) Create a similar table for the relative errors

   $$\frac{|\mathbb{P}(X = k) - \mathbb{P}(Y = k)|}{\mathbb{P}(X = k)}, \quad \frac{|\mathbb{P}(X = k) - \phi(k; np, np(1-p))|}{\mathbb{P}(X = k)}$$

   and

   $$\frac{|\mathbb{P}(X = k) - \text{Ent}(k; n, p)|}{\mathbb{P}(X = k)}$$

   for $k = 0, 1, \ldots, 30$. Based on this table, comment on the accuracy of each of the three approximations for the Binomial.

   (c) Repeat exercises (a) and (b) above for $n = 30$ and $p = 0.25$.

   (d) Repeat exercises (a) and (b) above for $n = 30$ and $p = 0.5$.

See attached PDF.

2. **(KL-Divergence, Multinomial)** Let X and Y be discrete random variables with distributions p and q respectively (So $p(k) = \mathbb{P}(X = k)$ and $q(k) = \mathbb{P}(Y = k)$). Remember that the Kullback–Leibler divergence is defined by

$$\mathrm{KL}(p\|q) := \mathbb{E}_p\Big[\ln\big(\frac{p(X)}{q(X)}\big)\Big] = \sum_k p(k)\ln\big(\frac{p(k)}{q(k)}\big).$$

(a) Show that when $q(k)$ is a Poisson distribution with parameter $\lambda > 0$, then the KL-divergence is minimized by setting λ to be the mean of $p(k)$.

(b) Rememebr that the entropy $H(p)$ is defined to be $H(p|q) := -\mathbb{E}_p[\ln(p(X))]$. Assume that we need to place n balls into d bins. The number of ways to place the balls resulting in k_i total balls in bin i, for $i = 1, \ldots, d$, is given by the combinatorial expression $\binom{n}{k_1,k_2,\ldots,k_d}$. Now consider the empirical distribution of the balls. Its probability mass function is $p(i) = k_i/n$. Let N_p denote the number of configurations with empirical distribution p, show that

$$\ln(N_p) = nH(p) + O(\ln(n)),$$

where $h(p)$ is the entropy of p.

In other words, there are many more high-entropy configurations than low-entropy configurations. This suggests the intuition that, if we consider a physical system at a "macro level" (such as the distribution of gas particles in a container) then we should expect it to drift toward high-entropy configurations.

Hint: Recall Stirling's approximation

$$\ln(n!) = n\ln(n) - n + O(\ln(n)).$$

(a) if $q(k)$ is Poisson with parameter λ, then $q(k) = e^{-\lambda}\lambda^k/k!$. Let's calculate now the KL-divergence.

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \sum_k p(k)\ln\big(\frac{p(k)}{q(k)}\big) \\
&= \sum_k p(k)\ln\big(p(k)\big) - \sum_k p(k)\ln\big(e^{-\lambda}\lambda^k/k!\big) \\
&= \sum_k p(k)\ln\big(p(k)\big) + \lambda\sum_k p(k) - \sum_k p(k)\Big(k\ln\big(\lambda\big) - \ln\big(k!\big)\Big) \\
&= \sum_k p(k)\ln\big(p(k)\big) + \lambda - \ln\big(\lambda\big)\sum_k kp(k) + \sum_k p(k)\ln\big(k!\big)
\end{aligned}$$

Taking the derivative we can verify that there is a minimum precisly when $\lambda = \sum_k kp(k)$.

(b) As stated at the beginning of the problem, $N_p = \frac{n!}{k_1!k_2!\cdots k_d!}$, we can now use Stirling approxi-

mation and do a calculation.

$$\begin{aligned}
\log(N_p) &= \log(n!) - \sum_{i=1}^{d} \log(k_i!) \\
&= n\log(n) - n + O(\log n) - \Big(\sum_{i=1}^{d} k_i \log(k_i) - k_i + O(\log k_i)\Big) \\
&= n\log(n) - n + O(\log n) - \Big(\sum_{i=1}^{d} k_i \log(k_i) - k_i + O(\log n)\Big) \\
&= n\log(n) - n - \Big(\sum_{i=1}^{d} k_i \log(k_i) - k_i\Big) + O(\log n) \\
&= \Big(\sum_{i=1}^{d} k_i\Big)\log(n) - \Big(\sum_{i=1}^{d} k_i \log(k_i)\Big) + \Big(\sum_{i=1}^{d} k_i\Big) - n + O(\log n) \\
&= -\Big(\sum_{i=1}^{d} k_i\big(\log(k_i) - \log(n)\big)\Big) + O(\log n) \\
&= -n\Big(\sum_{i=1}^{d} \frac{k_i}{n} \log\big(\frac{k_i}{n}\big)\Big) + O(\log n) \\
&= nh(p) + O(\log n)
\end{aligned}$$

The relevant identities we used are that $k_i \leq n$ and hence $\log(k_i) = O(\log n)$, that $\sum_{i=1}^{d} k_i = n$ and since d is fixed and finite, $O(\log n)\sum_{i=1}^{d} O(\log n)$ is still $O(\log n)$.

3. **(Poisson)** Let $K = X_1 + X_2 + \cdots + X_N$, where $N \sim \text{Poisson}\,(\lambda)$ and $X_1, X_2, \cdots$ are independent Bernoulli (p) random variables. Assuming that N and $\{X_i\}_{i\in\mathbb{N}}$ are mutually independent, find the distribution of K.

$$M_N(t) = \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}e^{ti} = \sum_{i=0}^{\infty} e^{-\lambda}\frac{\left(\lambda e^t\right)^i}{i!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda\left(e^t-1\right)},$$

Here we used that $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$. We have

$$\begin{aligned} M_K(t) &= \mathbb{E}\left[e^{tK}\right] = \sum_{n=0}^{\infty} \mathbb{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] \mathbb{P}(N=n) \\ &= \sum_{n=0}^{\infty} \left(q+pe^t\right)^n e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \sum_{n=0}^{\infty} e^{-\lambda} \frac{\left(\lambda\left(q+pe^t\right)\right)^n}{n!} \\ &= e^{-\lambda} e^{\lambda\left(q+pe^t\right)} \\ &= e^{\lambda\left(q+pe^t-1\right)} \\ &= e^{\lambda\left(pe^t-p\right)} \\ &= e^{\lambda p\left(e^t-1\right)}, \end{aligned}$$

we used the fact $q-1=-p$. This is the same as the MGF of $Poisson(\lambda p)$, thus $K \sim Poison(\lambda p)$.

4. **(Joint densities)** Let the joint density function of (X,Y) be

$$f(x,y) = \begin{cases} 3xy(x+y), & \text{if } (x,y) \in [0,1]^2, \\ 0, & \text{else.} \end{cases}$$

Calculate the covariance $\mathrm{Cov}(X,Y)$. We want to calculate $\mathbb{E}[X]$, $\mathbb{E}[Y]$ and $\mathbb{E}[XY]$. In all cases we need to apply the formula

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y)\, dx\, dy.$$

Let's do each one of the calculations and then conclude.

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\, dx\, dy \\ &= \int_0^1 \int_0^1 3x^2 y(x+y)\, dx\, dy \\ &= \int_0^1 \int_0^1 3x^3 y + 3x^2 y^2\, dx\, dy \\ &= \int_0^1 \frac{3}{4} y + y^2\, dy \\ &= \frac{17}{24}. \end{aligned}$$

By symmetry we also have $\mathbb{E}[Y] = \frac{17}{24}$.

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dx\,dy \\
&= \int_0^1\int_0^1 3x^2y^2(x+y)\,dx\,dy \\
&= \int_0^1\int_0^1 3x^3y^2 + 3x^2y^3\,dx\,dy \\
&= \int_0^1 \frac{3}{4}y^2 + y^3\,dy \\
&= \frac{1}{2}.
\end{aligned}$$

We finally conclude using the formula $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 1/2 - (17/24)^2 = -1/576$.

5. **(Transformation of random variables)**

   (a) Suppose X has the Cauchy distribution with density:

$$f_X(x) := \frac{1}{\pi(1+x^2)}.$$

   Show that $1/X$ has the same distribution as X.
   We aim to find the probability density function of $Y = \frac{1}{X}$. Using the change of variables formula, we have

$$\begin{aligned}
f_Y(y) &= f_X\left(\frac{1}{y}\right)\left|\frac{dx}{dy}\right| \\
&= \frac{1}{\pi\left(1+\left(\frac{1}{y}\right)^2\right)}\cdot\left|-\frac{1}{y^2}\right| \\
&= \frac{1}{\pi(1+y^2)}.
\end{aligned}$$

   Therefore, $Y = \frac{1}{X}$ has the same distribution as X.

   (b) Suppose $Y \sim \mathrm{Exp}(1)$. Find a function $g : (0,\infty) \to (-\infty,\infty)$ such that $g(Y)$ has the Cauchy distribution with density given by (a).
   The CDF of X in (a) is

$$F_X(x) = \frac{1}{\pi}\left(\arctan(x) + \frac{\pi}{2}\right).$$

   The CDF of Y is

$$F_Y(y) = P(Y \le y) = 1 - e^{-y}, \quad y > 0.$$

   Suppose g is a strictly increasing function, we can relate the CDFs of Y and X as

$$F_Y(y) = P(Y \le y) = P(g(Y) \le g(y)) = F_X(g(y))$$

Substituting the CDFs, we have

$$1 - e^{-y} = \frac{1}{\pi}\left(\arctan(g(y)) + \frac{\pi}{2}\right)$$

Simplifying the above, we get

$$g(y) = \tan\left(\pi\left(1 - e^{-y}\right) - \frac{\pi}{2}\right)$$

which is strictly increasing when $y > 0$. This is the required transformation that ensures $g(Y)$ follows the Cauchy distribution.

(c) Suppose $Z \sim \text{Exp}(\lambda)$, where $\lambda > 0$. Show that the distribution of $W := \lceil Z \rceil$ (here $\lceil z \rceil$ is the smallest integer that is larger than or equal to z) is Geometric. Explicitly express the parameter of the Geometric distribution in terms of λ.

Suppose X has an exponential distribution with rate parameter λ, i.e., the density of X is given by

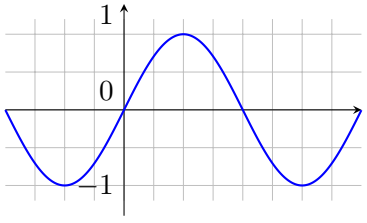
$$f_x(x) = \lambda e^{-\lambda x}; \quad x > 0.$$

Now consider $Y = \lceil X \rceil$. Since the ceiling function returns the smallest integer that is larger than or equal to X, Y is a discrete random variable. The PMF of Y can be expressed as

$$\begin{aligned}
\mathbb{P}[Y = y] &= \mathbb{P}[y - 1 < x \leq y] \\
&= \int_{y-1}^{y} \lambda e^{-\lambda x} dx \\
&= -e^{-\lambda x}\Big|_{y-1}^{y} \\
&= e^{-\lambda(y-1)}\left[1 - e^{-\lambda}\right] \\
&= \left(1 - \left(1 - e^{-\lambda}\right)\right)^{y-1}\left(1 - e^{-\lambda}\right)
\end{aligned}$$

The geometric distribution has the PMF $\mathbb{P}[X = x] = (1-p)^{x-1}p$. Recognizing $p = 1 - e^{-\lambda}$, we now have the PMF of $Y = \lceil X \rceil$ in the form of a geometric distribution with parameter $p = 1 - e^{-\lambda}$. Therefore, for $X \sim \text{Exp}(\lambda)$, we have $Y = \lceil X \rceil \sim \text{Geo}(1 - e^{-\lambda})$.

6. **(Transformation of random variables)** Suppose $X \sim \text{Uniform}[-\pi, 2\pi]$. Find the p.d.f. of $Y = \sin(X)$.



Plot of $y = \sin(x)$.

Here we have a transformation of the form $Y = g(X)$ for $g(x) = \sin(x)$. While X takes values on $[-\pi, 2\pi]$, Y takes values on $[-1, 1]$. A picture of the function g on $[-\pi, 2\pi]$ show us that we have

to make a distinction in two cases, either $y \in [-1,0]$ or $y \in (0,1]$. We want to use the formula $f_Y(y) = \sum_{\substack{g(x)=y,\\ g'(x)\neq 0}} f_X(x)\frac{1}{|g'(x)|}$. Note that $f_X(x) = 1/(3\pi)$ for $x \in [-\pi, 2\pi]$. $g'(x) = \cos(x)$. Now we separate in two cases.

(a) If $y \in (-1,0)$ then y has 4 preimages, we obtain

$$\begin{aligned} f_Y(y) &= \sum_{\substack{g(x)=y,\\ g'(x)\neq 0}} f_X(x)\frac{1}{|g'(x)|} \\ &= \sum_{\substack{g(x)=y,\\ g'(x)\neq 0}} \frac{1}{3\pi} \times \frac{1}{\cos(\arcsin(y)} = \frac{4}{3\pi\sqrt{1-y^2}}. \end{aligned}$$

(b) If $y \in (1,0)$ then y has 2 preimages, we obtain

$$\begin{aligned} f_Y(y) &= \sum_{\substack{g(x)=y,\\ g'(x)\neq 0}} f_X(x)\frac{1}{|g'(x)|} \\ &= \sum_{\substack{g(x)=y,\\ g'(x)\neq 0}} \frac{1}{3\pi} \times \frac{1}{\cos(\arcsin(y)} = \frac{2}{3\pi\sqrt{1-y^2}}. \end{aligned}$$

We shouldn't be concerned by the cases $y = -1, 0, 1$ since Y is a continuous random variable and we can modify the p.d.f at a finite number of points without any repercussion. We conclude that

$$f_Y(y) = \begin{cases} \dfrac{4}{3\pi\sqrt{1-y^2}} \text{ if } y \in (-1,0), \\ \dfrac{2}{3\pi\sqrt{1-y^2}} \text{ if } y \in (0,1), \\ 0 \text{ else.} \end{cases}$$

STAT201A_PS3_code

2024-10-06

Q1a

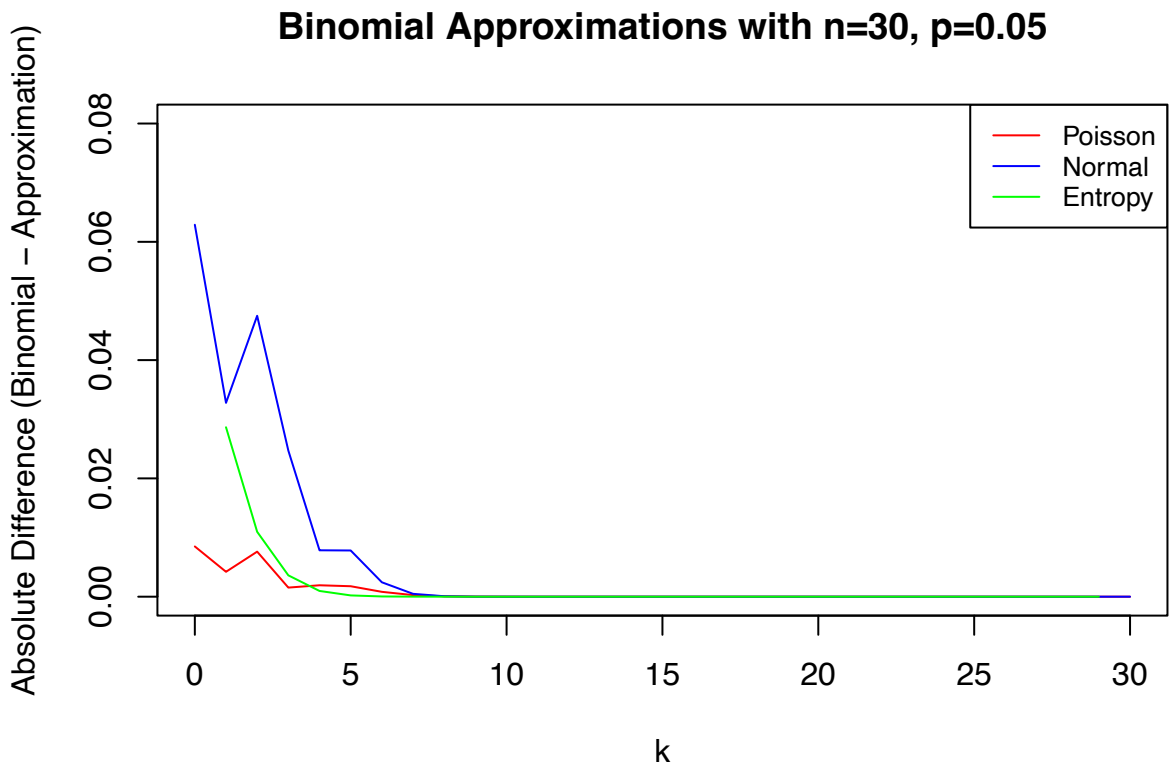
```
## Given values
n <- 30
p <- 0.05
k <- 0:30
## Binomial Calculation
## P(Bin(n,p) = k)
binom <- dbinom(k, n, p)
## Poisson Approximation
## P(Pois(np) = k)
pois <- dpois(k, n*p)
## Normal Approximation
## phi(k; np, np(1-p))
norm <- dnorm(k, n*p, sqrt(n*p*(1-p)))
#n*abs(k/n - p)^3 ## normal approx is good if this value is small

## Entropy Approximation
## Ent(k; n, p)
## Note that f=k/n and the entropy approximation DNE for k=0 and k=30
f <- k/n
entr <- 1/(sqrt(2*pi*n*f*(1-f)))*exp(-n*(f*log(f/p) + (1-f)*log((1-f)/(1-p))))
## Error terms
## Binomial - Poisson
pois_diff <- abs(binom - pois)
## Binomial - Normal
norm_diff <- abs(binom - norm)
## Binomial - Entropy
entr_diff <- abs(binom - entr)
cbind(pois_diff, norm_diff, entr_diff)
```

```
##          pois_diff    norm_diff    entr_diff
##  [1,] 8.491396e-03 6.288532e-02          NaN
##  [2,] 4.208071e-03 3.277279e-02 2.865137e-02
##  [3,] 7.615308e-03 4.749378e-02 1.096979e-02
##  [4,] 1.538910e-03 2.470382e-02 3.605901e-03
##  [5,] 1.930467e-03 7.845186e-03 9.679277e-04
##  [6,] 1.766931e-03 7.810490e-03 2.143120e-04
##  [7,] 8.209923e-04 2.434696e-03 3.976033e-05
##  [8,] 2.675847e-04 4.806307e-04 6.268669e-06
##  [9,] 6.786047e-05 7.384760e-05 8.496632e-07
## [10,] 1.412178e-05 9.515639e-06 9.992828e-08
## [11,] 2.493920e-06 1.051824e-06 1.027277e-08
## [12,] 3.828577e-07 1.006534e-07 9.284106e-10
## [13,] 5.205110e-08 8.387780e-09 7.409012e-11
```

```
## [14,] 6.362462e-09 6.112552e-10 5.237982e-12
## [15,] 7.081188e-10 3.906518e-11 3.287950e-13
## [16,] 7.252526e-11 2.193133e-12 1.834906e-14
## [17,] 6.896636e-12 1.082138e-13 9.107894e-16
## [18,] 6.133846e-13 4.690382e-15 4.019292e-17
## [19,] 5.132796e-14 1.782894e-16 1.574848e-18
## [20,] 4.060356e-15 5.926516e-18 5.466524e-20
## [21,] 3.047997e-16 1.715570e-19 1.675528e-21
## [22,] 2.177936e-17 4.299675e-21 4.514869e-23
## [23,] 1.485157e-18 9.257674e-23 1.063399e-24
## [24,] 9.686240e-20 1.694769e-24 2.173290e-26
## [25,] 6.053980e-21 2.601619e-26 3.818433e-28
## [26,] 3.632400e-22 3.286255e-28 5.701307e-30
## [27,] 2.095617e-23 3.326169e-30 7.132860e-32
## [28,] 1.164232e-24 2.593504e-32 7.360840e-34
## [29,] 6.236956e-26 1.462502e-34 6.203044e-36
## [30,] 3.226012e-27 5.308539e-37 4.487914e-38
## [31,] 1.613006e-28 9.313226e-40          NaN
```

```r
plot(0:30, pois_diff, type="l", col="red", ylim=c(0,0.08),
ylab="Absolute Difference (Binomial - Approximation)", xlab="k",
main="Binomial Approximations with n=30, p=0.05") ### poisson best here since p is small
lines(0:30, norm_diff, col="blue")
lines(0:30, entr_diff, col="green")
legend("topright", legend=c("Poisson", "Normal", "Entropy"),
col=c("red", "blue", "green"), lty=c(1,1,1), cex=0.8)
```



Based on this table, it seems that the Poisson approximation is doing a good job of approximating the Binomial distribution for small k. As k increases, we see that the error between the actual Binomial values and the approximations gets closer and closer to zero for all three approximations. It is important to note

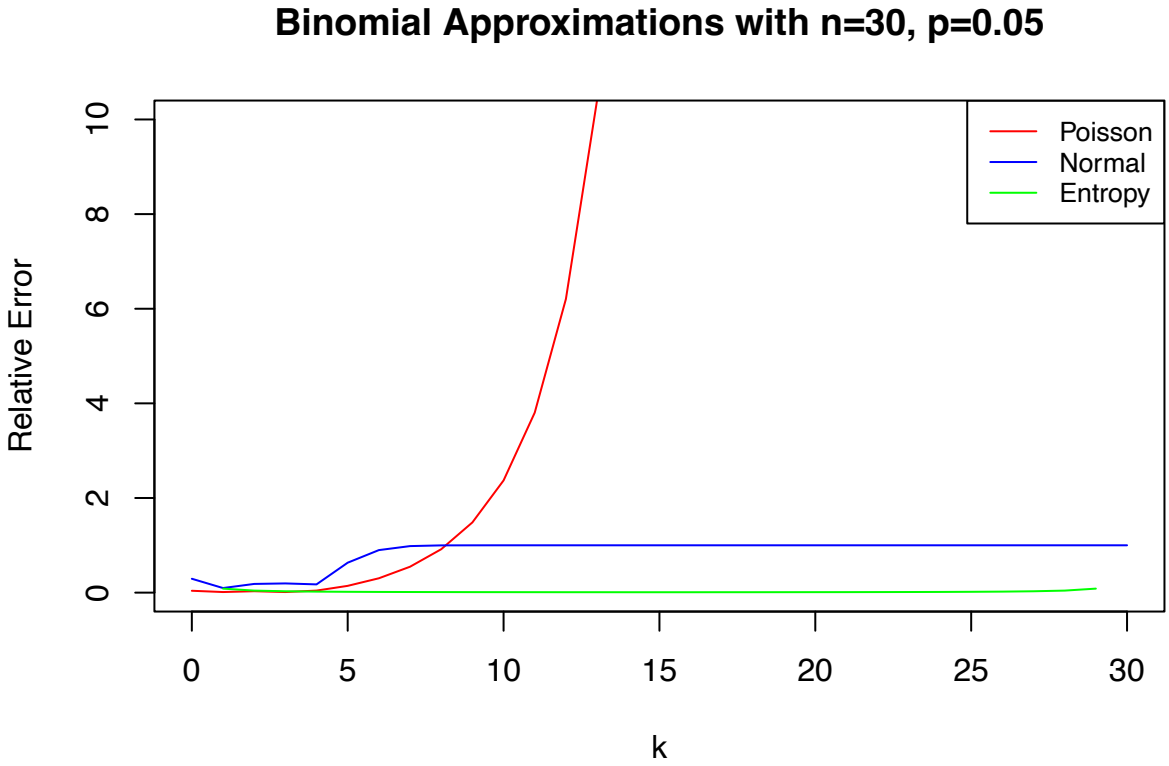
here though that the true Binomial values are inherently small, so the approximation differences are small. The Entropy approximation is not valid when $n - k$ or k is very small, which is why we have NaN values for $k = 0$ and $k = n$.

Q1b

```
## Relative Errors
pois_rel_error <- pois_diff / binom
norm_rel_error <- norm_diff / binom
entr_rel_error <- entr_diff / binom
cbind(pois_rel_error, norm_rel_error, entr_rel_error)
```

```
##       pois_rel_error norm_rel_error entr_rel_error
##  [1,]   3.956134e-02     0.29298210            NaN
##  [2,]   1.241673e-02     0.09670249    0.084541418
##  [3,]   2.944403e-02     0.18363122    0.042413904
##  [4,]   1.211267e-02     0.19444231    0.028381831
##  [5,]   4.276996e-02     0.17381198    0.021444669
##  [6,]   1.430363e-01     0.63227346    0.017348947
##  [7,]   3.030614e-01     0.89874474    0.014677142
##  [8,]   5.473854e-01     0.98320354    0.012823520
##  [9,]   9.174123e-01     0.99835288    0.011486679
## [10,]   1.483921e+00     0.99990583    0.010500489
## [11,]   2.371035e+00     0.99999689    0.009766593
## [12,]   3.803725e+00     0.99999994    0.009223841
## [13,]   6.205587e+00     1.00000000    0.008833103
## [14,]   1.040885e+01     1.00000000    0.008569223
## [15,]   1.812660e+01     1.00000000    0.008416573
## [16,]   3.306925e+01     1.00000000    0.008366598
## [17,]   6.373157e+01     1.00000000    0.008416573
## [18,]   1.307750e+02     1.00000000    0.008569223
## [19,]   2.878913e+02     1.00000000    0.008833103
## [20,]   6.851169e+02     1.00000000    0.009223841
## [21,]   1.776666e+03     1.00000000    0.009766593
## [22,]   5.065349e+03     1.00000000    0.010500489
## [23,]   1.604244e+04     1.00000000    0.011486679
## [24,]   5.715375e+04     1.00000000    0.012823520
## [25,]   2.327005e+05     1.00000000    0.014677142
## [26,]   1.105331e+06     1.00000000    0.017348947
## [27,]   6.300392e+06     1.00000000    0.021444669
## [28,]   4.489030e+07     1.00000000    0.028381831
## [29,]   4.264579e+08     1.00000000    0.042413904
## [30,]   6.077024e+09     1.00000000    0.084541418
## [31,]   1.731952e+11     1.00000000            NaN
```

```
plot(0:30, pois_rel_error, type="l", col="red", ylim=c(0,10),
ylab="Relative Error", xlab="k",
main="Binomial Approximations with n=30, p=0.05") ### poisson best here since p is small
lines(0:30, norm_rel_error, col="blue")
lines(0:30, entr_rel_error, col="green")
legend("topright", legend=c("Poisson", "Normal", "Entropy"),
col=c("red", "blue", "green"), lty=c(1,1,1), cex=0.8)
```



The relative error is pretty small for all three approximations until we hit about $k = 5$. Overall, the relative Entropy error is best and is consistently small. The Poisson approximation holds when we look at absolute difference, but it does not hold in the context of relative error, which is why see a spike in relative error for Poisson. For the Normal relative error, it stays at about 1 for $k = 8$ and beyond.

Q1c

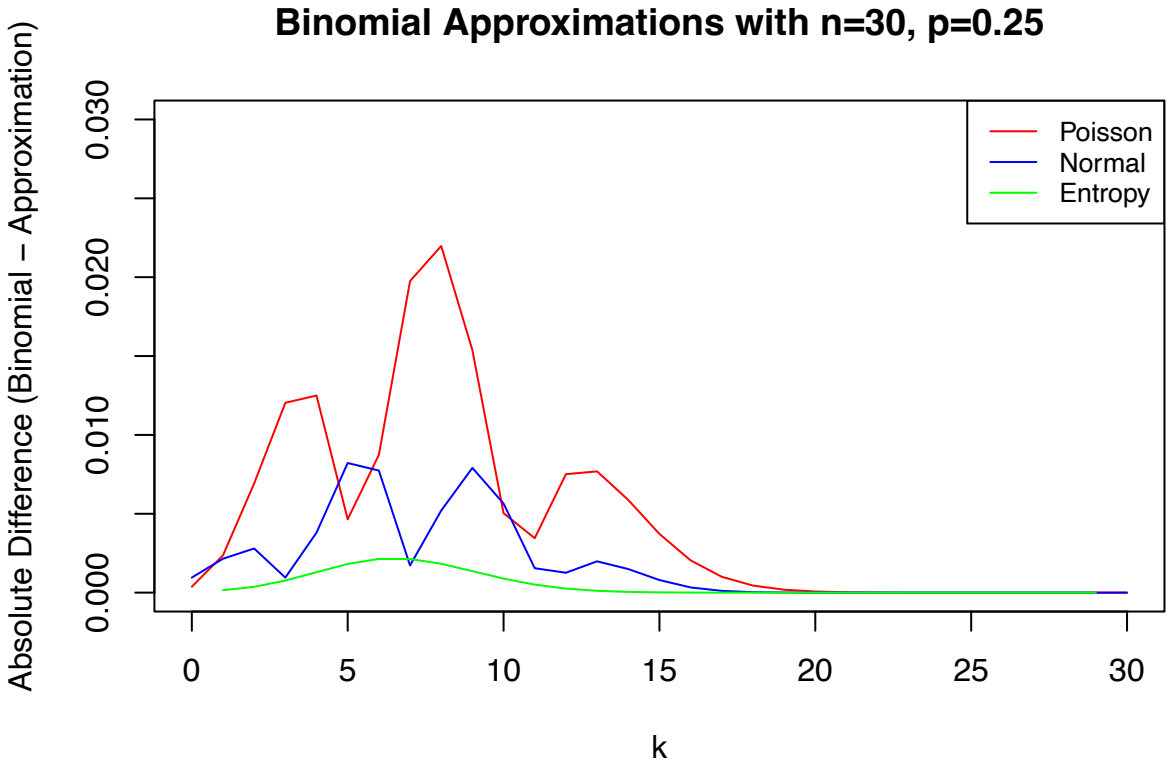
```
## Repeat this process for n = 30 and p = 0.25
n <- 30
p <- 0.25
k <- 0:30
## Binomial Calculation
## P(Bin(n,p) = k)
binom2 <- dbinom(k, n, p)
## Poisson Approximation
## P(Pois(np) = k)
pois2 <- dpois(k, n*p)
## Normal Approximation
## phi(k; np, np(1-p))
norm2 <- dnorm(k, n*p, sqrt(n*p*(1-p)))
#n*abs(k/n - p)^3 ## normal approx is good if this value is small

## Entropy Approximation
## Ent(k; n, p)
## Note that f=k/n and the entropy approximation DNE for k=0 and k=30
f2 <- k/n
entr2 <- 1/(sqrt(2*pi*n*f*(1-f)))*exp(-n*(f*log(f/p) + (1-f)*log((1-f)/(1-p))))
## Error terms
## Binomial - Poisson
```

```r
pois_diff2 <- abs(binom2 - pois2)
## Binomial - Normal
norm_diff2 <- abs(binom2 - norm2)
## Binomial - Entropy
entr_diff2 <- abs(binom2 - entr2)
cbind(pois_diff2, norm_diff2, entr_diff2)
```

```
##         pois_diff2   norm_diff2   entr_diff2
##  [1,] 3.745023e-04 9.548001e-04          NaN
##  [2,] 2.362312e-03 2.148155e-03 1.509758e-04
##  [3,] 6.924030e-03 2.799400e-03 3.660942e-04
##  [4,] 1.203529e-02 9.512783e-04 7.621502e-04
##  [5,] 1.249612e-02 3.802816e-03 1.295693e-03
##  [6,] 4.646120e-03 8.218080e-03 1.816929e-03
##  [7,] 8.737971e-03 7.738469e-03 2.134882e-03
##  [8,] 1.975184e-02 1.723586e-03 2.131727e-03
##  [9,] 2.198059e-02 5.202900e-03 1.829934e-03
## [10,] 1.536699e-02 7.910260e-03 1.363042e-03
## [11,] 5.034870e-03 5.645154e-03 8.874438e-04
## [12,] 3.450864e-03 1.547615e-03 5.079555e-04
## [13,] 7.510803e-03 1.259906e-03 2.567309e-04
## [14,] 7.686768e-03 1.983581e-03 1.149514e-04
## [15,] 5.874565e-03 1.495681e-03 4.569911e-05
## [16,] 3.721567e-03 7.971628e-04 1.615209e-05
## [17,] 2.046132e-03 3.299491e-04 5.077679e-06
## [18,] 1.003255e-03 1.104228e-04 1.419153e-06
## [19,] 4.471579e-04 3.054174e-05 3.521686e-07
## [20,] 1.838540e-04 7.073812e-06 7.742043e-08
## [21,] 7.055401e-05 1.382503e-06 1.502894e-08
## [22,] 2.550318e-05 2.287576e-07 2.564804e-09
## [23,] 8.744228e-06 3.202122e-08 3.825936e-10
## [24,] 2.858378e-06 3.772368e-09 4.952123e-11
## [25,] 8.940745e-07 3.702489e-10 5.510502e-12
## [26,] 2.683049e-07 2.978268e-11 5.210901e-13
## [27,] 7.740240e-08 1.915055e-12 4.128901e-14
## [28,] 2.150111e-08 9.472807e-14 2.698550e-15
## [29,] 5.759247e-09 3.385663e-15 1.440258e-16
## [30,] 1.489461e-09 7.782203e-17 6.599519e-18
## [31,] 3.723653e-10 8.625467e-19          NaN
```

```r
plot(0:30, pois_diff2, type="l", col="red", ylim=c(0,0.03),
ylab="Absolute Difference (Binomial - Approximation)", xlab="k",
main="Binomial Approximations with n=30, p=0.25")
lines(0:30, norm_diff2, col="blue")
lines(0:30, entr_diff2, col="green") ## entropy is best here
legend("topright", legend=c("Poisson", "Normal", "Entropy"),
col=c("red", "blue", "green"), lty=c(1,1,1), cex=0.8)
```

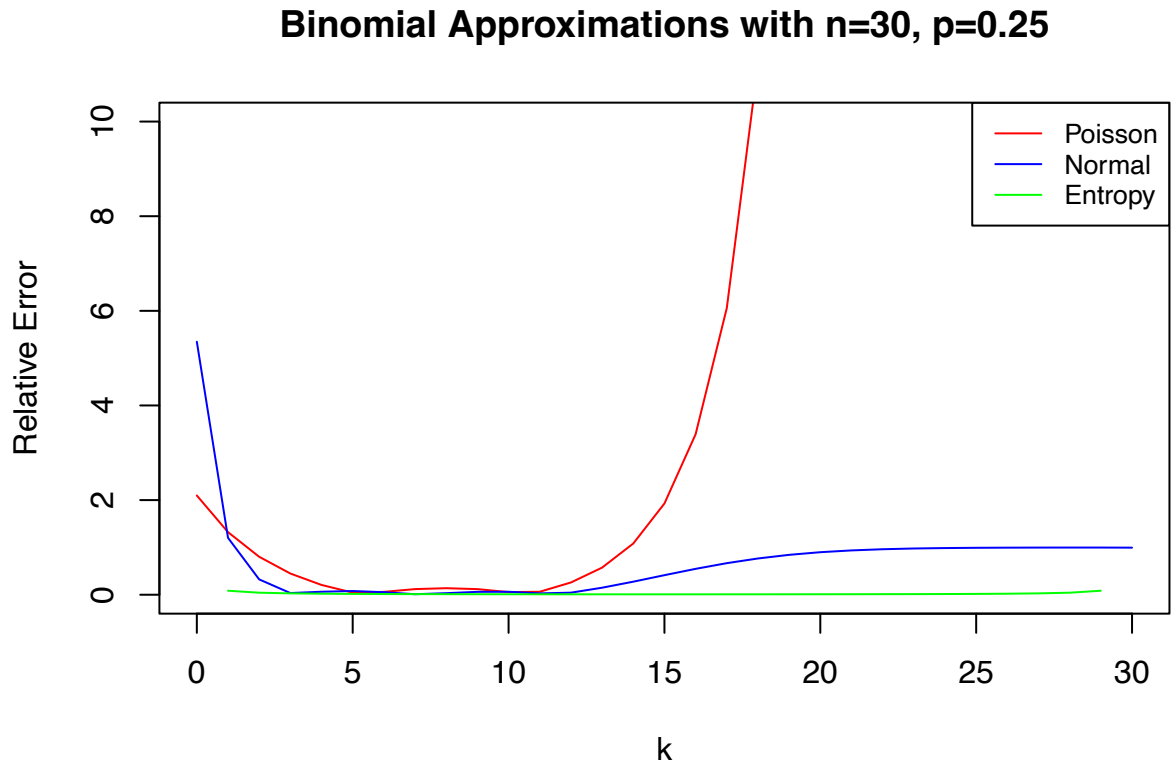



```
## Relative Errors
pois_rel_error2 <- pois_diff2 / binom2
norm_rel_error2 <- norm_diff2 / binom2
entr_rel_error2 <- entr_diff2 / binom2
cbind(pois_rel_error2, norm_rel_error2, entr_rel_error2)
```

```
##       pois_rel_error2 norm_rel_error2 entr_rel_error2
##  [1,]    2.097088e+00      5.34656148             NaN
##  [2,]    1.322816e+00      1.20289525     0.084541418
##  [3,]    8.021846e-01      0.32432490     0.042413904
##  [4,]    4.481840e-01      0.03542480     0.028381831
##  [5,]    2.068200e-01      0.06293940     0.021444669
##  [6,]    4.436348e-02      0.07847035     0.017348947
##  [7,]    6.007286e-02      0.05320136     0.014677142
##  [8,]    1.188183e-01      0.01036833     0.012823520
##  [9,]    1.379744e-01      0.03265913     0.011486679
## [10,]    1.183829e-01      0.06093839     0.010500489
## [11,]    5.541030e-02      0.06212666     0.009766593
## [12,]    6.266342e-02      0.02810277     0.009223841
## [13,]    2.584172e-01      0.04334841     0.008833103
## [14,]    5.730215e-01      0.14786901     0.008569223
## [15,]    1.081940e+00      0.27546514     0.008416573
## [16,]    1.927728e+00      0.41292112     0.008366598
## [17,]    3.391593e+00      0.54691136     0.008416573
## [18,]    6.057917e+00      0.66676229     0.008569223
## [19,]    1.121563e+01      0.76604869     0.008833103
## [20,]    2.190430e+01      0.84277120     0.009223841
## [21,]    4.584970e+01      0.89842291     0.009766593
## [22,]    1.044118e+02      0.93654972     0.010500489
## [23,]    2.625296e+02      0.96137894     0.011486679
```

```
## [24,]    7.401769e+02      0.97685460     0.012823520
## [25,]    2.381354e+03      0.98615246     0.014677142
## [26,]    8.932828e+03      0.99157154     0.017348947
## [27,]    4.020123e+04      0.99464042     0.021444669
## [28,]    2.261365e+05      0.99629656     0.028381831
## [29,]    1.696030e+06      0.99703786     0.042413904
## [30,]    1.908035e+07      0.99691874     0.084541418
## [31,]    4.293080e+08      0.99444867             NaN
```

```r
plot(0:30, pois_rel_error2, type="l", col="red", ylim=c(0,10),
ylab="Relative Error", xlab="k",
main="Binomial Approximations with n=30, p=0.25") ### poisson best here since p is small
lines(0:30, norm_rel_error2, col="blue")
lines(0:30, entr_rel_error2, col="green")
legend("topright", legend=c("Poisson", "Normal", "Entropy"),
col=c("red", "blue", "green"), lty=c(1,1,1), cex=0.8)
```



As shown in the above plot and tables, the entropy approximation is behaving best here with $p = 0.25$. The absolute difference for entropy is consistently small and the overall relative errors are smaller for entropy when compared to the other two approximations. At $k = 7$, we actually do see that the Normal approximation has a smaller absolute difference than the Entropy approximation, but overall the Entropy approximations behave the best. In terms of relative error, for small k Normal has the worst relative error. For k greater than 14, the Poisson relative error is the worst. The relative error for Entropy is consistently small.

Q1d

```r
## Repeat this process for n = 30 and p = 0.5
n <- 30
p <- 0.5
k <- 0:30
```

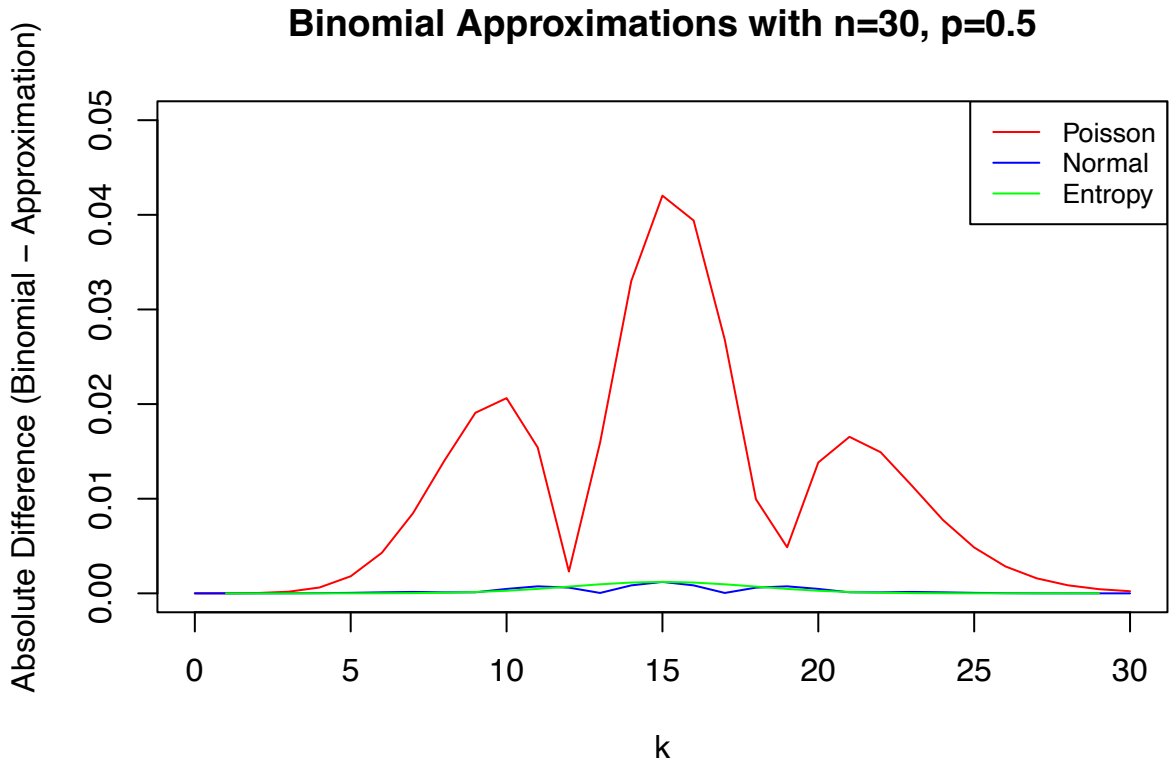
```r
## Binomial Calculation
## P(Bin(n,p) = k)
binom3 <- dbinom(k, n, p)
## Poisson Approximation
## P(Pois(np) = k)
pois3 <- dpois(k, n*p)
## Normal Approximation
norm3 <- dnorm(k, n*p, sqrt(n*p*(1-p)))
#n*abs(k/n - p)^3 ## normal approx is good if this value is small

## Entropy Approximation
## Ent(k; n, p)
## Note that f=k/n and the entropy approximation DNE for k=0 and k=30
f3 <- k/n
entr3 <- 1/(sqrt(2*pi*n*f*(1-f)))*exp(-n*(f*log(f/p) + (1-f)*log((1-f)/(1-p))))
## Error terms
## Binomial - Poisson
pois_diff3 <- abs(binom3 - pois3)
## Binomial - Normal
norm_diff3 <- abs(binom3 - norm3)
## Binomial - Entropy
entr_diff3 <- abs(binom3 - entr3)
cbind(pois_diff3, norm_diff3, entr_diff3)
```

```
##         pois_diff3   norm_diff3   entr_diff3
##  [1,] 3.049710e-07 4.363042e-08          NaN
##  [2,] 4.560595e-06 2.800940e-07 2.362060e-09
##  [3,] 3.400889e-05 1.458370e-06 1.718295e-08
##  [4,] 1.682889e-04 6.085087e-06 1.073165e-07
##  [5,] 6.197398e-04 2.019336e-05 5.473300e-07
##  [6,] 1.803069e-03 5.266949e-05 2.302536e-06
##  [7,] 4.286474e-03 1.049484e-04 8.116402e-06
##  [8,] 8.474307e-03 1.475841e-04 2.431322e-05
##  [9,] 1.399334e-02 1.040396e-04 6.261344e-05
## [10,] 1.908260e-02 1.094041e-04 1.399145e-04
## [11,] 2.062915e-02 4.675016e-04 2.732849e-04
## [12,] 1.541175e-02 7.416804e-04 4.692688e-04
## [13,] 2.306141e-03 6.059874e-04 7.115337e-04
## [14,] 1.592824e-02 4.012203e-05 9.557687e-04
## [15,] 3.299955e-02 8.428050e-04 1.139902e-03
## [16,] 4.202858e-02 1.208676e-03 1.208676e-03
## [17,] 3.940180e-02 8.428050e-04 1.139902e-03
## [18,] 2.679950e-02 4.012203e-05 9.557687e-04
## [19,] 9.940133e-03 6.059874e-04 7.115337e-04
## [20,] 4.871436e-03 7.416804e-04 4.692688e-04
## [21,] 1.382870e-02 4.675016e-04 2.732849e-04
## [22,] 1.653993e-02 1.094041e-04 1.399145e-04
## [23,] 1.491120e-02 1.040396e-04 6.261344e-05
## [24,] 1.138368e-02 1.475841e-04 2.431322e-05
## [25,] 7.746798e-03 1.049484e-04 8.116402e-06
## [26,] 4.847157e-03 5.266949e-05 2.302536e-06
## [27,] 2.847483e-03 2.019336e-05 5.473300e-07
## [28,] 1.592333e-03 6.085087e-06 1.073165e-07
## [29,] 8.546561e-04 1.458370e-06 1.718295e-08
```

```
## [30,] 4.422451e-04 2.800940e-07 2.362060e-09
## [31,] 2.211356e-04 4.363042e-08          NaN
```

```r
plot(0:30, pois_diff3, type="l", col="red", ylim=c(0,0.05),
ylab="Absolute Difference (Binomial - Approximation)", xlab="k",
main="Binomial Approximations with n=30, p=0.5") ### poisson best here since p is small
lines(0:30, norm_diff3, col="blue")
lines(0:30, entr_diff3, col="green")
legend("topright", legend=c("Poisson", "Normal", "Entropy"),
col=c("red", "blue", "green"), lty=c(1,1,1), cex=0.8)
```

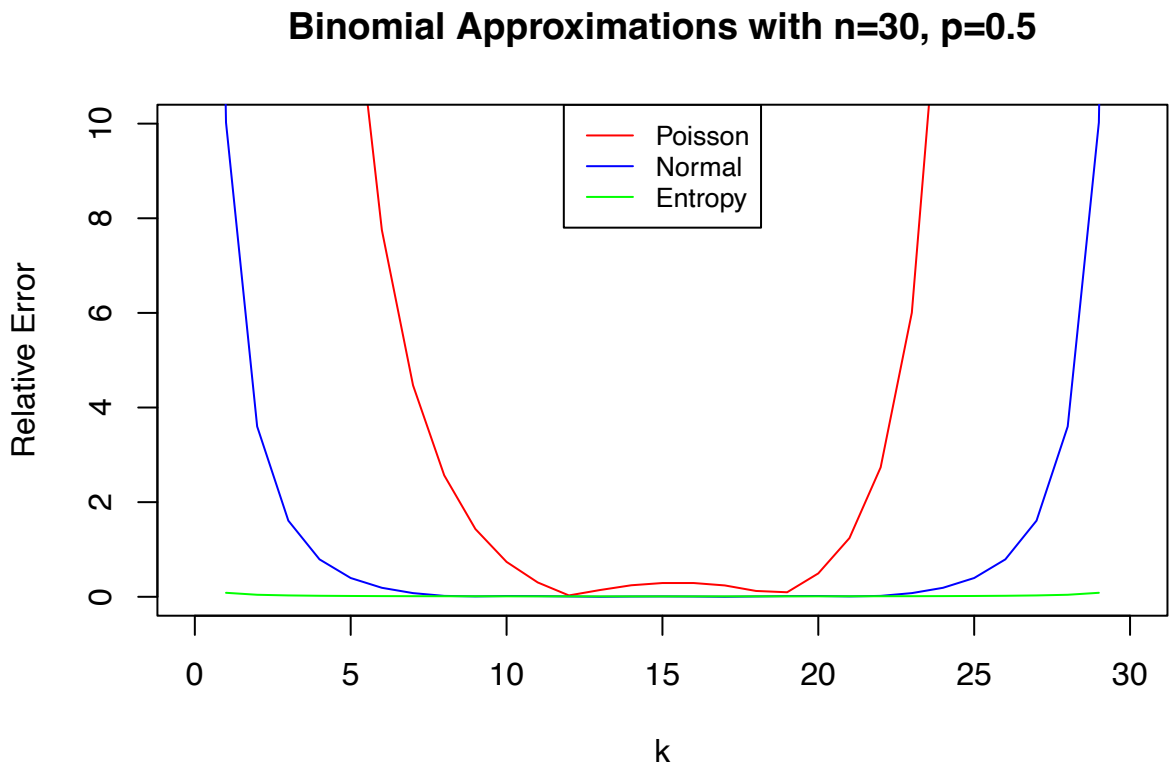



```r
pois_rel_error3 <- pois_diff3 / binom3
norm_rel_error3 <- norm_diff3 / binom3
entr_rel_error3 <- entr_diff3 / binom3
cbind(pois_rel_error3, norm_rel_error3, entr_rel_error3)
```

```
##       pois_rel_error3 norm_rel_error3 entr_rel_error3
##  [1,]    3.274601e+02    4.684781e+01             NaN
##  [2,]    1.632301e+02    1.002495e+01     0.084541418
##  [3,]    8.394658e+01    3.599799e+00     0.042413904
##  [4,]    4.450710e+01    1.609313e+00     0.028381831
##  [5,]    2.428172e+01    7.911860e-01     0.021444669
##  [6,]    1.358561e+01    3.968495e-01     0.017348947
##  [7,]    7.751365e+00    1.897814e-01     0.014677142
##  [8,]    4.469603e+00    7.784027e-02     0.012823520
##  [9,]    2.567132e+00    1.908647e-02     0.011486679
## [10,]    1.432136e+00    8.210703e-03     0.010500489
## [11,]    7.372398e-01    1.670746e-02     0.009766593
## [12,]    3.029299e-01    1.457830e-02     0.009223841
## [13,]    2.862884e-02    7.522833e-03     0.008833103
## [14,]    1.428093e-01    3.597258e-04     0.008569223
```

```
## [15,]    2.436553e-01    6.222929e-03     0.008416573
## [16,]    2.909268e-01    8.366598e-03     0.008366598
## [17,]    2.909268e-01    6.222929e-03     0.008416573
## [18,]    2.402787e-01    3.597258e-04     0.008569223
## [19,]    1.233985e-01    7.522833e-03     0.008833103
## [20,]    9.575183e-02    1.457830e-02     0.009223841
## [21,]    4.942070e-01    1.670746e-02     0.009766593
## [22,]    1.241311e+00    8.210703e-03     0.010500489
## [23,]    2.735518e+00    1.908647e-02     0.011486679
## [24,]    6.004096e+00    7.784027e-02     0.012823520
## [25,]    1.400878e+01    1.897814e-01     0.014677142
## [26,]    3.652194e+01    3.968495e-01     0.017348947
## [27,]    1.115658e+02    7.911860e-01     0.021444669
## [28,]    4.211218e+02    1.609313e+00     0.028381831
## [29,]    2.109609e+03    3.599799e+00     0.042413904
## [30,]    1.582857e+04    1.002495e+01     0.084541418
## [31,]    2.374425e+05    4.684781e+01              NaN
```

```r
plot(0:30, pois_rel_error3, type="l", col="red", ylim=c(0,10),
ylab="Relative Error", xlab="k",
main="Binomial Approximations with n=30, p=0.5") ### poisson best here since p is small
lines(0:30, norm_rel_error3, col="blue")
lines(0:30, entr_rel_error3, col="green")
legend("top", legend=c("Poisson", "Normal", "Entropy"),
col=c("red", "blue", "green"), lty=c(1,1,1), cex=0.8)
```



The absolute error for Poisson is the worst among our approximations. The Poisson absolute errors have a multi-modal shape, indicating that the Poisson approximations are only suitable for select values of k. The Normal and Entropy approximations are small overall. Within the range of $k = 12, 13, 14, 16, 17, 18$ we see that the Normal approximation is actually lower than Entropy in terms of absolute error. In this scenario,

Normal is the best approximation. For relative errors, Entropy again is the best. The Entropy relative errors are consistently small. From $k = 7$ to $k = 23$, the Normal relative errors are pretty small. Outside of this range, the approximation isn't appropriate. For $k = 12$ and $k = 19$, the Poisson relative error is small, but outside of this range, the approximation worsens.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 4

Due: 10:00pm, Tuesday, November 5, 2024 (via Gradescope)

1. **(Order statistics)** Let $X_1, \ldots, X_n$ be i.i.d. random variables with Exp(λ) distribution, where $\lambda > 0$, and let $X_{(i)}$ be the order statistics for $i = 1, \ldots, n$.

   (a) Find the distribution of $X_{(1)}$.

   (b) Using the memoryless property, find the distribution of $X_{(i+1)} - X_{(i)}$ for $i = 1, \ldots, n-1$.

   (c) Use the previous item to show that each $X_{(i)}$ has the same distribution as a sum of i independent random variables.

   (d) Calculate the expectation and the variance of $X_{(i)}$ for $i = 1, \ldots, n$.

   (a) Since $X_{(1)} = \min_{i \in \{1,\ldots,n\}} X_i$, $\mathbb{P}(X_{(1)} > t) = \mathbb{P}(X_i > t)^n = e^{-n\lambda t}$. Hence $X_{(1)}$ is distributed as an exponential random variable with parameter $n\lambda$.

   (b) For each i, consider the $n-i$ random variables $Y_k = X_k - X_{(i)}|_{X_k > X_{(i)}}$. The key observation is that these random variables have exponential distributions, an application of the memoryless property gives

   $$\mathbb{P}(Y_k > t) = \mathbb{P}(X_k - X_{(i)} > t | X_k > X_{(i)}) = \mathbb{P}(X_k > t).$$

   Moreover, $X_{(i+1)} - X_{(i)}$ corresponds to the minimum of the random variables Y_k and hence has exponential distribution with parameter $(n-i)\lambda$.

   (c) It follows from our prveious argument that we can write $X_i = \sum_{k=1}^{i} X_{(k)} - X_{(k-1)}$ where $X_{(0)} = 0$. Hence we can describe $X_{(i)} = \sum_{k=1}^{i} Z_k$ where Z_k is a collection of independent random variables, Z_k with exponential distribution with parameter $(n-k+1)\lambda$.

   (d) Finally, using our previous formula we have that

   $$\mathbb{E}[X_{(i)}] = \sum_{k=1}^{i} \mathbb{E}[Z_k] = \sum_{k=1}^{i} \frac{1}{(n-k+1)\lambda}.$$

   Similarly for the variance,

   $$\mathrm{Var}[X_{(i)}] = \sum_{k=1}^{i} \mathrm{Var}[Z_k] = \sum_{k=1}^{i} \frac{1}{(n-k+1)^2\lambda^2}.$$

2. **(Joint and conditional densities)** Let X, Y be two random variables with the following properties. Y has density function $f_Y(y) = 3y^2$ for $0 < y < 1$ and zero elsewhere. For $0 < y < 1$, given that $Y = y$, X has conditional density function $f_{X|Y}(x|y) = \frac{2x}{y^2}$ for $0 < x < y$ and zero elsewhere.

   (a) Find the joint density function $f_{X,Y}(x, y)$ of X, Y. Be precise about the values (x, y) for which your formula is valid. Check that the joint density function you find integrates to 1.

(b) Find the conditional density function of Y, given $X = x$. Be precise about the values of x and y for which the answer is valid. Identify the conditional distribution of Y by name.

(a) The joint density is given by

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = \frac{2x}{y^2}1_{0<x<y} \cdot 3y^2 1_{0<y<1} = 6x \cdot 1_{0<x<y<1}.$$

We have that

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f_{X,Y}(x,y)\,dx\,dy = \int_0^1 \int_0^y 6x\,dx\,dy = 1.$$

(b) We first calculate the maringal density of X, $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy = \int_x^1 6x\,dx = 6x(1-x)\cdot 1_{0<x<1}$. We can now calculate the conditional density of Y given X.

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{1-x}\cdot 1_{x<y<1}.$$

We conclude that the conditional distribution of Y given X is $\mathcal{U}(X,1)$, uniform on the interval $(X,1)$.

3. **(Model selection)** Given data $x_1,\ldots,x_n$, consider the problem of selecting between the two models:

$$\text{Model One}: X_1,\ldots,X_n \overset{\text{i.i.d}}{\sim} N(0,1)$$

and

$$\text{Model Two}: X_1,\ldots,X_n \overset{\text{i.i.d}}{\sim} N(\mu,1) \text{ for an unknown } \mu.$$

To use probability to solve this problem, let us introduce an additional random variable Θ that has the Bernoulli distribution with parameter 0.5. Assume that the conditional distribution of $X_1,\ldots,X_n$ given $\Theta = \theta$ is given by the following

$$X_1,\ldots,X_n \mid \Theta = 0 \overset{\text{i.i.d}}{\sim} N(0,1)$$

and

$$X_1,\ldots,X_n \mid \mu,\Theta = 1 \overset{\text{i.i.d}}{\sim} N(\mu,1) \text{ and } \mu \mid \Theta = 1 \sim N\left(0,\tau^2\right).$$

Here τ is a parameter which you can treat as a fixed constant in this exercise.

(a) Using the formula

$$f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) = \int f_{X_1,\ldots,X_n|\mu,\Theta=1}(x_1,\ldots,x_n)\, f_{\mu|\Theta=1}(\mu)d\mu$$

prove that

$$f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{1+n\tau^2}} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2}\right)\exp\left(\frac{n^2\tau^2\bar{x}^2}{2\left(1+n\tau^2\right)}\right),$$

where $\bar{x}$ is the mean of $x_1, \ldots, x_n$.

$$
\begin{aligned}
f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) &= \int f_{X_1,\ldots,X_n|\mu,\Theta=1}\,(x_1,\ldots,x_n)\, f_{\mu|\Theta=1}(\mu)d\mu \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\mu^2}{2\tau^2}\right) d\mu \\
&= \frac{1}{\tau}\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu n\bar{x} + n\mu^2\right)\right) \exp\left(-\frac{\mu^2}{2\tau^2}\right) d\mu \\
&= \frac{1}{\tau}\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\left(n+\frac{1}{\tau^2}\right)\mu^2 - 2n\mu\bar{x}\right)\right) d\mu \\
&= \frac{1}{\tau}\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \sqrt{\frac{2\pi}{n+\frac{1}{\tau^2}}} \exp\left(\frac{n^2\tau^2\bar{x}^2}{2(1+n\tau^2)}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^{n} \frac{1}{\sqrt{1+n\tau^2}} \exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2}\right) \exp\left(\frac{n^2\tau^2\bar{x}^2}{2(1+n\tau^2)}\right).
\end{aligned}
$$

(b) Calculate the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$.
From Model One, we know

$$
f_{X_1,\ldots,X_n|\Theta=0}\,(x_1,\ldots,x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2}\right).
$$

By Bayes' theorem,

$$
\begin{aligned}
&\mathbb{P}\left(\Theta = 1 \mid X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \frac{\mathbb{P}(\Theta=1) f_{X_1,\ldots,X_n|\Theta=1}\,(x_1,\ldots,x_n)}{\mathbb{P}(\Theta=0) f_{X_1,\ldots,X_n|\Theta=0}\,(x_1,\ldots,x_n) + \mathbb{P}(\Theta=1) f_{X_1,\ldots,X_n|\Theta=1}\,(x_1,\ldots,x_n)} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=1}\,(x_1,\ldots,x_n)}{f_{X_1,\ldots,X_n|\Theta=0}\,(x_1,\ldots,x_n) + f_{X_1,\ldots,X_n|\Theta=1}\,(x_1,\ldots,x_n)} \\
&= \frac{\frac{1}{\sqrt{1+n\tau^2}} \exp\left(\frac{n^2\tau^2\bar{x}^2}{2(1+n\tau^2)}\right)}{1 + \frac{1}{\sqrt{1+n\tau^2}} \exp\left(\frac{n^2\tau^2\bar{x}^2}{2(1+n\tau^2)}\right)}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{P}\left(\Theta = 0 \mid X_1 = x_1, \ldots, X_n = x_n\right) &= 1 - \mathbb{P}\left(\Theta = 1 \mid X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \frac{1}{1 + \frac{1}{\sqrt{1+n\tau^2}} \exp\left(\frac{n^2\tau^2\bar{x}^2}{2(1+n\tau^2)}\right)}
\end{aligned}
$$

(c) Intuitively, we would prefer Model Two over Model One when $\bar{x}$ is far from zero. Is this intuition reflected in your conditional distribution from the previous part?

Yes. When $\bar{x}$ is close to zero, the exponential term $\exp\left(\frac{n^2\tau^2\bar{x}^2}{2(1+n\tau^2)}\right)$ is approximately 1, and the square root term $\frac{1}{\sqrt{1+n\tau^2}}$ is small for large n, so $(\Theta = 1)$ remains small, meaning we favor Model One. As $\bar{x}$ moves away from zero, the exponential term grows rapidly, dominating the expression, so $\mathbb{P}(\Theta = 1)$ increases and close to 1, favoring Model Two.

4. **(Gamma-Poisson)** Consider random variables $\Theta, X_1, \ldots, X_n$ such that

$$\Theta \sim \text{Gamma}(\alpha, \lambda) \quad \text{and} \quad X_1, \ldots, X_n \mid \Theta = \theta \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$$

(a) Find the conditional distribution of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.
The conditional distribution of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is

$$\begin{aligned}
f_{\Theta|X_1=x_1,X_2=x_2,\ldots,X_n=x_n}(\theta) &\propto f_{X_1,X_2,\ldots,X_n|\Theta=\theta}(x_1, x_2, \ldots, x_n) f_\Theta(\theta) \\
&\propto \left(\prod_{i=1}^n f_{X_i|\Theta=\theta}(x_i)\right) f_\Theta(\theta) \\
&= \left(\prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} 1\{x_i \in \mathbb{N}_0\}\right) \left(\frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta} 1\{\theta > 0\}\right) \\
&= \frac{\theta^{\sum x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} 1\{x_i \in \mathbb{N}_0\} \frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta} 1\{\theta > 0\} \\
&\propto \theta^{\alpha-1+\sum x_i} e^{-(n+\lambda)\theta} 1\{x_i \in \mathbb{N}_0\} 1\{\theta > 0\}.
\end{aligned}$$

The above is in the form of a Gamma $(\alpha + \sum x_i, n + \lambda)$ distribution. Therefore, the exact distribution of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is

$$f_{\Theta|X_1=x_1,X_2=x_2,\ldots,X_n=x_n}(\theta) = \frac{(n+\lambda)^{\alpha+\sum x_i}}{\Gamma\left(\alpha + \sum x_i\right)} \theta^{\alpha+\sum x_i-1} e^{-(n+\lambda)\theta} 1\{\theta > 0\},$$

where $x_i \in \mathbb{N}_0$.

(b) Find $\mathbb{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$.
We know that for $Y \sim \text{Gamma}(\alpha, \beta)$, $\mathbb{E}(Y) = \frac{\alpha}{\beta}$. Since $\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \sim \text{Gamma}(\alpha + \sum x_i, n + \lambda)$, we have

$$\mathbb{E}[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \frac{\alpha + \sum x_i}{n + \lambda}.$$

(c) Write $\mathbb{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$ as a weighted linear combination of $(x_1 + \cdots + x_n)/n$ and the mean of the marginal distribution (i.e., prior mean) of Θ and argue that the weight of the prior mean goes to zero as $n \to \infty$.
We can express $\mathbb{E}[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]$ as a weighted linear combination of the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and the prior mean $\frac{\alpha}{\lambda}$. Specifically, we have

$$\mathbb{E}[\Theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \frac{\alpha + \sum_{i=1}^n x_i}{n + \lambda} = \frac{\lambda}{n+\lambda} \cdot \frac{\alpha}{\lambda} + \frac{n}{n+\lambda} \cdot \bar{x}.$$

As $n \to \infty$, the weight on the prior mean, $\frac{\lambda}{n+\lambda}$, tends to 0, meaning the prior mean becomes less influential. Conversely, the weight on the sample mean, $\frac{n}{n+\lambda}$, tends to 1, meaning the sample

mean dominates as n increases. Thus, as $n \to \infty$, the conditional expectation of Θ approaches the sample mean $\bar{x}$, which aligns with the intuition that with more data, the influence of the prior diminishes, and the posterior is dominated by the data.

5. **(Law of total expectation)** Let the joint probability mass function (p.m.f.) of (X, Y) be

$$p_{X,Y}(k,n) = \begin{cases} \frac{1}{n+1}\left(1-\frac{1}{n+1}\right)^{k-1}\frac{1}{2^n}, & \text{for } 1 \le n < \infty \text{ and } 1 \le k < \infty, \\ 0, & \text{else.} \end{cases}$$

(a) Find the p.m.f. $p_Y(n)$ of Y and the conditional p.m.f $p_{X|Y}(k|n)$.

(b) Calculate $\mathbb{E}[Y]$.

(c) Find the conditional expectation $\mathbb{E}[X|Y]$.

(d) Use parts (a) and (c) to calculate $\mathbb{E}[X]$.

(a) We start with a calculation.

$$\begin{aligned} p_Y(n) &= \sum_k p_{X,Y}(k,n) \\ &= \sum_{k=1}^{\infty} \frac{1}{n+1}\big(1-\frac{1}{n+1}\big)^{k-1}\frac{1}{2^n} \\ &= \frac{1}{(n+1)2^n}\sum_{k=1}^{\infty}\big(1-\frac{1}{n+1}\big)^{k-1} \\ &= \frac{1}{(n+1)2^n}(n+1) = \frac{1}{2^n}. \end{aligned}$$

Now we have that $p_{X|Y}(k|n) = \dfrac{p_{X,Y}(k,n)}{p_Y(n)} = \dfrac{1}{n+1}\big(1-\dfrac{1}{n+1}\big)^{k-1}$. We are able to recognize this distributions.

$$Y \sim \text{Geom}(1/2) \quad \text{and} \quad X|_{Y=n} \sim \text{Geom}(1/(n+1)).$$

(b) We automatically conclude from (a) that $\mathbb{E}[Y] = 2$.

(c) Since $X|_{Y=n} \sim \text{Geom}(1/(n+1))$ we automatically conclude that $\mathbb{E}[X|Y=n] = n+1$. It follows that $\mathbb{E}[X|Y] = Y+1$.

(d) We finally have that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[Y+1] = \mathbb{E}[Y]+1 = 3$.

6. **(Expected number of coin tosses)** Consider a sequence of coin tosses.

(a) On average, how many tosses of a fair coin does it take to see two heads in a row?

(b) How many tosses on average to see the sequence HTH for the first time?

(c) How does our answer changes if we have an unfair coin?

(a) Let X be the random variable describing the number of tosses needed to see tho heads in a row. Let A be the even that the first toss is tails, B the event that the first two tosses are heads and C the event that the first toss is head and the second is tails. Observe that the following equation is satisfied

$$\mathbb{E}[X] = \mathbb{E}[X|A]\mathbb{P}(A) + \mathbb{E}[X|B]\mathbb{P}(B) + \mathbb{E}[X|C]\mathbb{P}(C).$$

Now the key observation is that $\mathbb{E}[X|A] = \mathbb{E}[X] + 1$, $\mathbb{E}[X|B] = 2$ and $\mathbb{E}[X|C] = \mathbb{E}[X] + 2$. We can now solve the equation

$$\mathbb{E}[X] = \frac{\mathbb{E}[X] + 1}{2} + \frac{2}{4} + \frac{\mathbb{E}[X] + 2}{4}.$$

We get $\mathbb{E}[X] = 6$.

(b) We repeat the previous argument. Let x represents the expected number of tosses to get HTH, y the expected number of tosses to get HTH given that our last toss is H and z the expected number of tosses to get HTH given that our last toss is HT. We then obtain the following system of equations. $a = \frac{a+1}{2} + \frac{b+1}{2}$. $b = \frac{b+1}{2} + \frac{c+1}{2}$ and finally $c = \frac{1}{2} + \frac{a+1}{2}$. Solving the system of equations gives $a = 10$.

(c) This is completly analogous to part (a) and (b) only that the probability to get H is now p. Solving the equations we get that the expected number of tosses to get HH is $\frac{1+p}{p^2}$ while the expected number of tosses to get HTH is $\frac{1+p-p^2}{p^2(1-p)}$.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 5

Due: 10:00pm, Wednesday, November 20, 2024 (via Gradescope)

1. **(Multivariate normal)** Suppose $Y \sim \mathcal{N}_n(\mu, \Sigma)$ in this problem.

   (a) If a is any fixed vector in $\mathbb{R}^n$, show that

   $$\frac{a^\top (Y - \mu)}{\sqrt{a^\top \Sigma a}} \sim \mathcal{N}(0, 1).$$

   (b) If A is now a random vector that is independent of Y, then show again that

   $$\frac{A^\top (Y - \mu)}{\sqrt{A^\top \Sigma A}}$$

   is distributed according to $\mathcal{N}(0, 1)$ and that it is independent of A.

   (c) Using the above result, show that if $Y \sim \mathcal{N}_3(0, I_3)$, then

   $$\frac{Y_1 e^{Y_3} + Y_2 \log |Y_3|}{\sqrt{e^{2Y_3} + (\log |Y_3|)^2}} \sim \mathcal{N}(0, 1).$$

(a) Suppose $Y \sim \mathrm{N_n}(\mu, \Sigma)$ and $a \in \mathbb{R}^n$. Then, since Y is Multivariate Normal,

$$a^\top Y \sim \mathrm{N}\left(a^\top \mu, a^\top \Sigma a\right)$$

$$\Rightarrow a^\top Y - a^\top \mu \sim \mathrm{N}\left(0, a^\top \Sigma a\right)$$

$$\Rightarrow \frac{a^\top (Y - \mu)}{\sqrt{a^\top \Sigma a}} \sim \mathrm{N}(0, 1).$$

(b) We note that from part (a), when a is fixed, $\frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}} \sim \mathrm{N}(0, 1)$. Hence, its moment generation function is given by

$$M_{\frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}}}(t) = \mathbb{E}\left[e^{t \frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}}}\right] = e^{\frac{1}{2}t^2}.$$

Now, when a is a random vector independent from Y, we have

$$\begin{aligned} M_{\frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}}}(t) &= \mathbb{E}\left[e^{t \frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}}}\right] \\ &= \mathbb{E}\left(\mathbb{E}\left[e^{t \frac{a^\top (Y-\mu)}{\sqrt{a^\top \Sigma a}}} \,\middle|\, a\right]\right) \\ &= \mathbb{E}\left(e^{\frac{1}{2}t^2}\right) \\ &= e^{\frac{1}{2}t^2}. \end{aligned}$$

(c) Let $Y = (Y_1, Y_2, Y_3)^T \sim \mathrm{N}_3\left((0,0,0)^T, I\right)$, then $W = (Y_1, Y_2)^T \sim \mathrm{N}_2\left((0,0)^T, I\right)$. Letting $a = \left(e^{Y_3}, \log|Y_3|\right)$ we see that aW. Then,

$$\frac{a^\top (W-\mu)}{\sqrt{a^\top \Sigma a}} \sim \mathrm{N}(0,1)$$

$$\implies \frac{a^\top W}{\sqrt{a^\top a}} \sim \mathrm{N}(0,1)$$

$$\implies \frac{Y_1 e^{Y_3} + Y_2 \log|Y_3|}{\sqrt{e^{2Y_3} + (\log|Y_3|)^2}} \sim \mathrm{N}(0,1).$$

2. **(Marginally normal but not bivariate normal)** Give an example of a 2×1 random vector $Y = (Y_1, Y_2)^T$ with positive definite covariance matrix such that each Y_1 and Y_2 is standard normal but Y is not bivariate normal.
This was covered in lecture 19. See the first slide. Take $Y_1 \sim_d \mathcal{N}(0,1)$ and X a random variable taking value -1 and 1 with probability $1/2$ in each case. Then $Y_= XY_1$ is also a normal. Moreover the covariance matrix of (Y_1, Y_2) is the identity matrix, so it is positive definite. However This random vector is not bivariate normal since $Y_1 + Y_2$ is not normal. For instance, $\mathbb{P}(Y_1 + Y_2 = 0) = 1/2$.

3. **(Conditional distribution)** Consider three random variables Y_1, Y_2 and Y_3 that are independent and standard normal. Let

$$X_1 = Y_2 + Y_3,$$
$$X_2 = Y_1 + Y_3,$$
$$X_3 = Y_1 + Y_2.$$

Find the conditional distribution of X_1 given $X_2 = X_3 = 0$.
This problem is an application of the formulas for conditional normal random variables given in lecture 19. Since $Y \sim_d \mathcal{N}\left((0,0,0)^T, I_3\right)$, taking the matrix $A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$, we have that $(X_1, X_2, X_3)^T = A(Y_1, Y_2, Y_3)^T$. Hence

$$(X_1, X_2, X_3) \sim_d \mathcal{N}\left((0,0,0)^T, AI_3A^T\right).$$

Here $AI_3A^T = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$. Using the formulas given in lecture we obtain for $Z = X_1|_{X_2=X_3=0}$, $\mathbb{E}[Z] = 0$ and $\mathrm{Var}(Z) = 2 - (1,1)\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^{-1}(1,1)^T = \frac{4}{3}$.

4. **(More on jointly Gaussian distributions)** Let X and Y be independent standard normal variables.
   (a) For a constant k, find $\mathbb{P}(X > kY)$.
   (b) If $U = \sqrt{3}X + Y$, and $V = X - \sqrt{3}Y$, find $\mathbb{P}(U > kV)$.
   (c) Find $\mathbb{P}(U^2 + V^2 < 1)$.

(d) Find the conditional distribution of X given $V = v$.

(a) We can use the radial symmetry of the joint distribution of two standard independent random variables. Since the line $x = ky$ goes through 0 it divides the plane in two sections where the total density is the same. We automatically have $P(X > kY) = 1/2$. Alternatively, $X - kY \sim N(0, 1 + k^2)$ and $P(X - kY > 0) = 1/2$.

(b) Notice that $U \sim N(0, 4)$ and $V \sim N(0, 4)$. Furthermore, $Cov(U, V) = Cov(\sqrt{3}X + Y, X - \sqrt{3}Y)$. Using bilinearity properties of covariance this is $\sqrt{3}Var(X) - 3Cov(X, Y) + Cov(Y, X) - \sqrt{3}Var(Y) = -2Cov(X, Y) = 0$. It follows that the joint (U, V) is uncorrelated bivariate normal and that $P(U > kV) = \frac{1}{2}$ by radial symmetry of uncorrelated bivariate normal. Alternatively, you can check that $U - kV \sim N(0, (\sqrt{3} - k)^2 + (1 - k\sqrt{3})^2)$ and thus $P(U - kV > 0) = \frac{1}{2}$.

(c) $U, V \sim iid\, N(0, 4)$ so $U/2, V/2 \sim iid\, N(0, 1)$. It follows that

$$\left(\frac{U}{2}\right)^2 + \left(\frac{V}{2}\right)^2 \sim Exp\left(\frac{1}{2}\right).$$

Then

$$P(U^2 + V^2 < 1) = P(\left(\frac{U}{2}\right)^2 + \left(\frac{V}{2}\right)^2 < \frac{1}{4}) = 1 - e^{-\frac{1}{2}(\frac{1}{2})^2} = 1 - e^{-\frac{1}{8}}.$$

(d) $Cov(X, V) = Cov(X, X - \sqrt{3}Y) = Var(X)$. It follows that

$$Corr(X, V) = \frac{Cov(X, V)}{SD(X)SD(V)} = \frac{Var(X)}{\sqrt{Var(X)Var(V)}} = \frac{1}{\sqrt{4}} = \frac{1}{2}.$$

Hence $(X, V) \sim BVN(0, 0, 1, 4, \rho = \frac{1}{2})$ which implies that $(X, \frac{V}{2}) \sim BVN(0, 0, 1, 1, \rho = \frac{1}{2})$. Hence $X|(\frac{V}{2} = \frac{v}{2}) \sim N(\rho\frac{v}{2}, 1 - \rho^2) = N(\frac{1}{4}v, \frac{3}{4})$.

5. **(Wigner's surmise)** Let $X = \begin{pmatrix} X_1 & X_3 \\ X_3 & X_2 \end{pmatrix}$ with X_1 and X_2 independent $\mathcal{N}(0, 1)$ and X_3 another independent $\mathcal{N}(0, 1/2)$. Let λ_1 and λ_2 be two eigenvalues of X and $s = |\lambda_1 - \lambda_2|$.

(a) Prove that $s = \sqrt{(X_1 - X_2)^2 + 4X_3^2}$.

(b) Find the density of s.

(c) Plot the density function of s. What do you observe respect to the eigenvalues of the random matrix X?

Start by noticing that since all the entries of the random matrix are continuous, the probability of the eigenvalues to be equal is 0.

(a) Since this is a 2×2 matrix an explicit calculation of the characteristic polynomial gives

$$p(t) = t^2 - (X_1 + X_2)t + X_1X_2 - X_3^2.$$

The roots are

$$\lambda_1 = \frac{X_1 + X_2 + \sqrt{(X_1 + X_2)^2 - 4X_1X_2 + 4X_3^2}}{2} \text{ and } \lambda_1 = \frac{X_1 + X_2 - \sqrt{(X_1 + X_2)^2 - 4X_1X_2 + 4X_3^2}}{2}.$$

This gives $s = \sqrt{(X_1 - X_2)^2 + 4X_3^2}$.

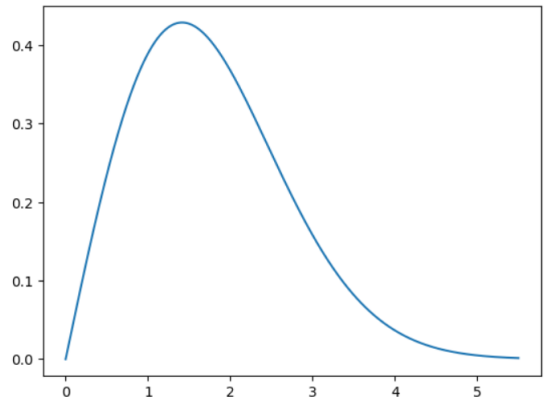


Figure 1: Wigner's surmise

(b) Notice that $X_1 - X_2 \sim_d \mathcal{N}(0,2)$ and $X_3 \sim_d \mathcal{N}(0,2)$. Hence $s \sim_d \sqrt{2}\sqrt{U^2+V^2}$ where U and V are independent standar normals. We get that $s \sim_d \sqrt{2}\chi(2)$, so s has the same distribution as a reescaled by $\sqrt{2}$ chi-2 distribution. The pdf of a chi-2 distribution is given by $p(x) = xe^{-x^2/2}I_{x\geq 0}$. We conclude that the density function of s is given by $p(s) = \frac{s}{2}e^{s^2/4}I_{s\geq 0}$.

(c) Using python we obtain a nice graph. The observation is that while the probability of the eigenvalues being far away decreases exponentially, the probability of them being arbitrarily close goes to 0.

6. **(1D Gaussian process)** In this problem, you will implement a 1D Gaussian process that predicts outputs based on noisy training data. You will be given (noisy) 1D training data pairs $D_{\text{train}} = \{(x_1, y_1), (x_2, y_2) \ldots\}$. Your task is to predict the output for a set of test queries $D_{\text{test}} = \{x_1^*, x_2^*, \ldots\}$, conditioned on the training data. Implement two separate kernel functions, namely the

- **Squared Exponential Kernel**: This is the kernel we discussed in class.

$$k(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{(x_i - x_j)^T M (x_i - x_j)}{2}\right)$$

where σ_f is a scale factor for the kernel and M is a metric measuring distance between two input vectors. In the 1D case, $M = \frac{1}{l^2}$ where l is the length scale of the kernel.

- **Matérn Kernel**: This kernel is used commonly in many machine learning applications.

$$k(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right)$$

where ν and l are (positive) parameters of the kernel and $r = |x_i - x_j|$. K_ν is a modified bessel function and Γ is the gamma function. Good parameters settings for ν are 0.25 - 3. You can use scipy.special.kv() in Python or besselK() in R for implementing K_ν.

(a) Implement the squared exponential and Matérn kernel functions to compute similarity between any pair of inputs. The output for each function should be a kernel matrix K.

(b) Using your kernel functions, implement a Gaussion process regression function to predict the posterior mean and variance of test data $\vec{y}^*$.

(c) The simulation function and plotting function are provided in the file `ps5_GP_1D.ipynb`. Vary the kernel parameters (e.g., σ_f, l, and ν) and observe how they affect the predictive mean and variance. What impact do these parameters have on the smoothness and uncertainty of your GP predictions?

Note: It's recommended to use Python (Jupyter notebook) and submit a pdf file including code, plots and comments. If you prefer using another coding language, please make sure the data simulation is the same with the provided code.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 6

Due: 10:00pm, Friday, December 6, 2024 (via Gradescope)

1. (**Branching process**) A branching process starts with one individual, i.e. $X(0) = 1$, who reproduces according to the following principle:

# of children	0	1	2
probability	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Individuals reproduce independently of each other and independently of the number of their sisters and brothers. Determine

(a) the probability that the population becomes extinct;
The probability generating function of the number of offspring is

$$\phi(s) = \sum_{k=0}^{2} s^k p_k = \frac{1}{6} + \frac{1}{3}s + \frac{1}{2}s^2.$$

The probability of extinction is the smallest solution s to the equation

$$s = \phi(s).$$

Solving this equation, the probability of extinction is $\frac{1}{3}$.

(b) the probability that the population has become extinct in the second generation, i.e. $\mathbb{P}(X(2) = 0)$;

$$\begin{aligned}\mathbb{P}(X(2)=0) &= \mathbb{P}(X(2)=0 \mid X(1)=0)\cdot\mathbb{P}(X(1)=0)\\ &\quad + \mathbb{P}(X(2)=0 \mid X(1)=1)\cdot\mathbb{P}(X(1)=1)\\ &\quad + \mathbb{P}(X(2)=0 \mid X(1)=2)\cdot\mathbb{P}(X(1)=2)\\ &= 1\cdot\frac{1}{6} + \frac{1}{6}\cdot\frac{1}{3} + \left(\frac{1}{6}\right)^2\cdot\frac{1}{2}\\ &= \frac{17}{72}.\end{aligned}$$

(c) the expected number of children given that there are no grandchildren.

$$\begin{aligned}\mathbb{E}[X(1) \mid X(2)=0] &= \mathbb{P}(X(1)=1 \mid X(2)=0)\cdot 1 + \mathbb{P}(X(1)=2 \mid X(2)=0)\cdot 2\\ &= \frac{\mathbb{P}(X(2)=0 \mid X(1)=1)\cdot\mathbb{P}(X(1)=1)}{\mathbb{P}(X(2)=0)}\cdot 1\\ &\quad + \frac{\mathbb{P}(X(2)=0 \mid X(1)=2)\cdot\mathbb{P}(X(1)=2)}{\mathbb{P}(X(2)=0)}\cdot 2\\ &= \frac{\left(\frac{1}{6}\cdot\frac{1}{3}\right)}{\frac{17}{72}}\cdot 1 + \frac{\left(\frac{1}{36}\cdot\frac{1}{2}\right)}{\frac{17}{72}}\cdot 2\\ &= \frac{6}{17}.\end{aligned}$$

2. (**Random walk)** Random walk on $\{0,1,2,3\}$. Consider the Markov chain (X_n) with transition matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

started with $X_0 = 0$. Define T_j as $\min\{n \geq 1 : X_n = j\}$. Find explicitly the following distributions and expectations.

(a) The distribution of X_2.

$$P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$$

$P(X_2 = 0) = \frac{1}{2}, P(X_2 = 1) = P(X_2 = 2) = \frac{1}{4}, P(X_2 = 3) = 0.$

(b) The limit distribution of X_n as $n \to \infty$.

By solving $\pi P = \pi$, we have $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Since the MC has finite state space S and it's irreducible and aperiodic, in limiting distribution theorem, we know for $i, j \in 0, 1, 2, 3$, $\lim_{n\to\infty} [P^n]_{ij} = \frac{1}{4}$. Therefore the limit distribution of X_n is also π.

(c) $\mathbb{E}[T_0]$

Define $h_k = \mathbb{E}[\min\{n \geq 0 : X_n = 0\} \mid X_0 = k]$. By one-step analysis, we derive the following system of equations

$$\begin{cases} h_0 = 0, \\ h_1 = \frac{1}{2}h_0 + \frac{1}{2}h_2 + 1, \\ h_2 = \frac{1}{2}h_1 + \frac{1}{2}h_3 + 1, \\ h_3 = \frac{1}{2}h_2 + \frac{1}{2}h_3 + 1. \end{cases}$$

Solving this system, we find

$$\begin{cases} h_0 = 0, \\ h_1 = 6, \\ h_2 = 10, \\ h_3 = 12. \end{cases}$$

Finally, the expected value of T_0 is given by:

$$\mathbb{E}[T_0] = 1 + \sum_{k=0}^{3} h_k P(X_1 = k) = \frac{1}{2} \times 0 + \frac{1}{2} \times 6 + 1 = 4.$$

(d) $\mathbb{E}[T_3]$

Define $g_k = \mathbb{E}\left[\min\{n \geq 0 : X_n = 3\} \mid X_0 = k\right]$. Similar to (c), by one-step analysis, we derive the following system of equations

$$\begin{cases} g_0 = \frac{1}{2}g_0 + \frac{1}{2}g_1 + 1, \\ g_1 = \frac{1}{2}g_0 + \frac{1}{2}g_2 + 1, \\ g_2 = \frac{1}{2}g_1 + \frac{1}{2}g_3 + 1, \\ g_3 = 0. \end{cases}$$

Solving this system, we find

$$\begin{cases} g_0 = 12, \\ g_1 = 10, \\ g_2 = 6, \\ g_3 = 0. \end{cases}$$

Finally, since it definitely takes more than one step from 0 to 3,

$$\mathbb{E}[T_3] = \mathbb{E}[\min\{n \geq 1 : X_n = 3\}|X_0 = 0] = \mathbb{E}[\min\{n \geq 0 : X_n = 3\}|X_0 = 0] = 12.$$

(e) $\mathbb{P}[T_3 < T_0]$

Define $f_k = \mathbb{P}\left(T_3 < T_1 \mid X_1 = k\right)$. We can derive the following system of equations by conditional probabilities

$$\begin{cases} f_0 = 0, \\ f_1 = \frac{1}{2}f_0 + \frac{1}{2}f_2, \\ f_2 = \frac{1}{2}f_1 + \frac{1}{2}f_3, \\ f_3 = 1. \end{cases}$$

Solving this system, we find

$$\begin{cases} f_0 = 0, \\ f_1 = \frac{1}{3}, \\ f_2 = \frac{2}{3}, \\ f_3 = 1. \end{cases}$$

Finally,

$$\mathbb{P}(T_3 < T_1) = \mathbb{P}\left(T_3 < T_1 \mid X_1 = 1\right) P(X_1 = 1) + \mathbb{P}\left(T_3 < T_1 \mid X_1 = 0\right) P(X_1 = 0) = 1/6.$$

3. (**The average number of jobs)** Jennifer is employed for one day at a time. When she is out of work, she visits the job agency in the morning to see if there is work for that day. There is a job for her with probability 1/2. If there is no work, she comes back the next day. When she has a job, she will be called back to the same job for the next day with probability 2/3. When she is not called back, she goes to the job agency again the next morning to look for a new job that she had not had previously. Approximate the average number of jobs Jennifer works in a year.

There are multiple solutions to this problem. A short approximation consist in calculating the average number of days on the same job and the average number of days without a job. Let X be the random variable representing the number of days on a fixed job, X has geometric distribution

with parameter 1/3. Let Y be a random variable representing the number of days without job, Y also has geometric distribution but starts at 0, with parameter 1/2. We get $\mathbb{E}[X] = 3$ and $\mathbb{E}[Y] = 1$. Hence $\mathbb{E}[X + Y] = 4$. We obtain $365/4 \approx 91.25$.

4. (**Rain or no rain**) Suppose that at day 0 it is not raining. Then each new day, if it rained yesterday, it will rain with probability 0.7; if it did not rain yesterday, it will rain with probability 0.2.

   (a) Find the stationary distribution.
   This is an irreducible finite state Markov chain. Hence an stationary distribution exists and is unique. It is enough to find π_r and π_n, the stationary distribution for rain and not rain respectively, by solving the system of equations $-0.3\pi_r + 0.2\pi_n = 0$, $0.3\pi_r - 0.2\pi_n = 0$, $\pi_r + \pi_n = 1$. We get $(\pi_r, \pi_n) = (2/5, 3/5)$.

   (b) How many days should we expect to wait to have rain for the first time?
   Suppose we start from a non-rainy day. Then by considering non-rainy days as failures and rainy days as successes. Then for X the random variable giving the first day with rain, X is geometric with parameter $0.2 = 1/5$. Hence $\mathbb{E}[x] = 5$.

5. (**The game of roulette**) A gambler plays the game of roulette, betting X dollars on red or black. The gambler wins X dollars with probability $p = 18/38$ or loses the bet with probability $q = 20/38$. Suppose that the gambler starts the game with \$500 in his pocket and upper limit on winnings is \$1000.

   (a) Compute the probability of the gambler's ruin for $X = \$10$.
   This problem can be solved by considering a Markov chain on $\{0, 1, \ldots, N\}$, where N is a positive integer; 0 and N are absorbing boundaries; and for $j = 1, \ldots, N-1$,

$$\mathbb{P}[X_{t+1} = j - 1 \mid X_t = j] = q = 1 - p,$$
$$\mathbb{P}[X_{t+1} = j + 1 \mid X_t = j] = p.$$

   Let R denote the event that you hit the boundary 0 before hitting the boundary N. Define $u_j := \mathbb{P}[R \mid X_0 = j]$. Then, $u_0 = 1$ and $u_N = 0$, while for $j = 1, \ldots, N-1$,

$$\begin{aligned} u_j &= \mathbb{P}[R \mid X_0 = j] \\ &= \mathbb{P}[R \mid X_0 = j, X_1 = j-1]\mathbb{P}[X_1 = j-1 \mid X_0 = j] \\ &\qquad + \mathbb{P}[R \mid X_0 = j, X_1 = j+1]\mathbb{P}[X_1 = j+1 \mid X_0 = j] \\ &= (1-p)\, u_{j-1} + p\, u_{j+1}. \end{aligned}$$

   Now, rewrite the left hand side as $p\mu_j + (1-p)\mu_j$ and rearrange terms to get

$$p[\mu_{j+1} - \mu_j] = (1-p)[\mu_j - \mu_{j-1}].$$

Define $r := \frac{1-p}{p}$ and $\Delta_j := \mu_j - \mu_{j-1}$. Then, we obtain

$$\begin{aligned}
\Delta_2 &= r\Delta_1 \\
\Delta_3 &= r\Delta_2 = r^2\Delta_1 \\
\Delta_4 &= r\Delta_3 = r^2\Delta_2 = r^3\Delta_1 \\
&\vdots \\
\Delta_N &= r^{N-1}\Delta_1.
\end{aligned}$$

Further, for all $j = 1, \ldots, N$, we have $\Delta_1 + \Delta_2 + \cdots + \Delta_j = \mu_j - \mu_0 = \mu_j - 1$, where we have used the boundary condition in the last equality. Hence,

$$\mu_j = 1 + \Delta_1 + \Delta_2 + \cdots + \Delta_j = 1 + \Delta_1[1 + r + \cdots + r^{j-1}].$$

Since $\mu_N = 0$, we obtain $\Delta_1 = -1/[1 + r + \cdots + r^{N-1}]$, so

$$\mu_j = 1 - \frac{1 + r + \cdots + r^{j-1}}{1 + r + \cdots + r^{N-1}} = \begin{cases} 1 - \dfrac{j}{N}, & \text{if } r = 1, \\ \dfrac{r^j - r^N}{1 - r^N}, & \text{if } r \neq 1. \end{cases} \tag{1}$$

Using $N = 100$ and $X_0 = 50$ in (1) gives $\mathbb{P}[R \mid X_0 = 50] \approx 0.995$.

(b) Compute the probability of the gambler's ruin for $X = \$100$.

Similarly, this is equivalent to having $N = 10$ and $X_0 = 5$ in the ruin problem, and we immediately obtain $\mathbb{P}[R \mid X_0 = 5] \approx 0.629$.

(c) Compare the above results with the probability of ruin in the case the gambler bets everything on a single turn of the wheel.

If the gambler bets everything on a single turn of the wheel, the probability of ruin is $q = 1 - p = 20/38 \approx 0.526$. This probability is lower than either one of the cases above.

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Final Practice Problems

1. (**True/False**) Determine whether each of the following claims is true or false. If true, provide an argument or proof; and if false, give a counterexample.

   (a) Suppose X and Y are distributed Uniform in $[0,1]$. Then, (X,Y) is uniform on $[0,1]^2$.

   (b) For every random variable X, one can find a function g such that $g(Z)$ has the same distribution as X (here $Z \sim U(0,1)$).

   (c) For every random variable X, one can find a function h such that $h(X) \sim N(0,1)$.

   (d) For any random variables X and Y, $\mathrm{Var}[X] = \mathrm{Var}[\mathrm{Var}[X \mid Y]]$.

2. (**Short questions, no explanation is needed.**) A random variable X has moment generating function $M_X(t) < \infty$ for $|t| < \epsilon$ for some $\epsilon > 0$.

   (a) Suppose $M_X(t) = e^{3t+t^2}$. What is the distribution of X? Explain.

   (b) Find $M_X(t)$ for $X = W_1 + \cdots + W_k$ where the W_k are IID with $P(W_k > w) = e^{-w}$ for $w > 0$.

   (c) Suppose $X_1, \ldots, X_n$ are independent normal (μ, σ^2) random variables. Let $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. What is the distribution of $\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$?

   (d) Suppose $U_1, \ldots, U_n$ are independent uniform $[0, 1]$ variables. Determine a constant c, where $(\prod_{i=1}^n U_i)^{1/n} \xrightarrow{d} c$.

   (e) Suppose X_i for $1 \leq i \leq 6$ is the number of dice showing face i when a Poisson(μ) number of fair six-sided dice are rolled. Describe the joint distribution of $(X_1, \ldots, X_6)$.

3. Suppose a coin (about which you know nothing) has been tossed n times (n is a large number) and it shows heads 75% of the time. What probability should one assign to the coin showing heads in the next toss? Answer this question using the following model. Let $X_1, \ldots, X_{n+1}$ be the results of the $n+1$ coin tosses ($X_i = 1$ if the ith toss is heads and $X_i = 0$ if tails). Assume that

$$X_1, \ldots, X_n, X_{n+1} \mid \theta \sim \text{Ber}(\theta) \quad \text{and} \quad \theta \sim \text{Unif}(0,1)$$

   (a) Calculate the conditional density of θ given $X_1 = x_1, \ldots, X_n = x_n$.

   (b) Calculate the conditional distribution of X_{n+1} given $X_1 = x_1, \ldots, X_n = x_n$.

   (c) What happens to $\mathbb{P}\{X_{n+1} = 1 \mid X_1 = x_1, \ldots, X_n = x_n\}$ when n is large and $(x_1 + \cdots + x_n)/n = 0.75$?

4. Let $F_n(x) := \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \leq x\right)$ be the empirical cdf for $X_1, \ldots, X_n$ assumed to be independent with common cdf $F(x) := P\left(X_i \leq x\right)$. Suppose $x < y$.

   (a) Find a simple formula for $\mathrm{Cov}\left(F_n(x), F_n(y)\right)$.

   (b) Find the conditional distribution of $nF_n(x)$ given $nF_n(y) = k$ for each $0 \leq k \leq n$.

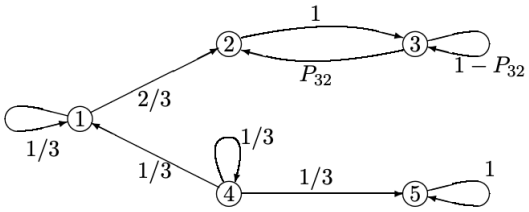
   (c) Suppose F is continuous. What is the distribution of $F\left(\max_{1\leq i\leq n} X_i\right) - F\left(\min_{1\leq i\leq n} X_i\right)$? Explain and evaluate the mean of this distribution.

5. Let $X_1, X_2, \ldots$ be i.i.d. random variables with positive integer values (on the set $\{1, 2, \ldots\}$). Assume that $\mathbb{P}(X_1 = 1) > 0$. Let $Z_n = \min\{X_1, X_2, \ldots, X_n\}$. Show that

$$\lim_{n \to \infty} \mathbb{P}(Z_n = 1) = 1.$$

How does this change if the random variables are not identically distributed?

6. Consider the following finite-state Markov chain.



(a) Identify the transient states and identify each class of recurrent states.

(b) Is the stationary distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_5)$ unique in above Markov Chain? If it's unique, explain the reason and calculate the stationary distribution. If not, give at least two stationary distributions.

(c) Find the n-step transition probabilities $p_{ij}^{(n)} = \mathbb{P}\{X_n = j \mid X_0 = i\}$ as a function of n. Give a brief explanation of each.

i. $p_{44}^{(n)}$

ii. $p_{43}^{(n)} + p_{42}^{(n)}$

iii. $\lim_{n\to\infty} p_{43}^{(n)}$

7. You may use the approximation $\frac{1}{2^{2n}}\binom{2n}{n} \sim \frac{1}{\sqrt{2\pi n}}$ and that $\sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} = \infty$ for $\alpha \leq 1$.

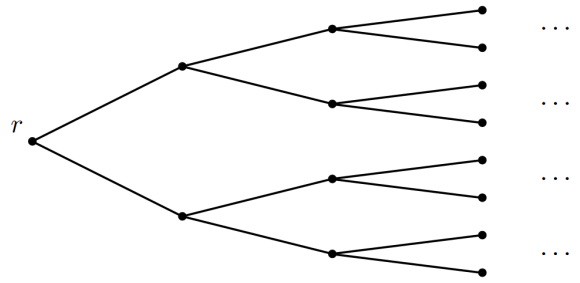
   (a) Show that the one-dimensional symmetric random walk is recurrent.

   (b) Using item (a), show the two-dimensional symmetric random walk is recurrent.

8. You may use the approximation $\frac{2^n}{9^n}\binom{2n}{n} \sim \frac{8^n}{9^n\sqrt{2\pi n}}$ and that $\sum_{n=1}^{\infty} \frac{x^n}{\sqrt{n}} < \infty$ for $0 \leq x < 1$.

   (a) Suppose that you have a one-dimensional biased random walk with bias $p = \frac{2}{3}$ to the right. Show that this random walk is transient.

   (b) Consider a symmetric random walk on the infinite binary tree with root r (depicted below) starting at r. Is it recurrent or transient?



Final Exam Solutions

7:00-10:00pm, December 20, 2024

Your First Name: **Your Last Name:**

SIGN Your Name: **Your SID Number:**

Instructions:

(a) *As soon as the exam starts, please write your student ID in the space provided at the top of every page! (We will remove the staple when scanning your exam.)*

(b) *There are* **8 double-sided** *sheets (15 numbered pages) on the exam. Notify a proctor immediately if a sheet is missing.*

(c) *We will not grade anything outside of the space provided for a question (i.e., either a designated box if it is provided, or otherwise the white space immediately below the question).* **Be sure to write your full answer in the box or space provided!** *Scratch paper is provided on request; however, please bear in mind that nothing you write on scratch paper will be graded!*

(d) ***You may use, without proof, theorems and lemmas that were proved in lecture and/or in homework.***

(e) *You may consult two double-sided "cheat sheets" of notes. Apart from that, you may not look at any other materials. Calculators, phones, computers, and other electronic devices are NOT permitted.*

(f) *You have 180 minutes: there are 7 questions on this exam worth a total of 120 points.*

Problem	Points
1	14
2	30
3	16
4	11
5	18
6	16
7	15

(g) *On questions 1-2, you need only give the answer in the format requested (e.g., True/False, an expression, a statement.) An expression may simply be a number or an expression with a relevant variable in it. For short answer questions, correct, clearly identified answers will receive full credit with no justification. Incorrect answers may receive partial credit.*

(h) *On questions 3-7, you should give arguments, proofs or clear descriptions* ***if requested****. If there is a box, you must use it for your answer; answers written outside the box may not be graded!*

[exam starts on next page]

1. True/False [*No justification; answer by shading the correct bubble. 2 points per correct answer; total of 14 points. No penalty for incorrect answers.*]

Indicate which of the following statements is **TRUE** or **FALSE** by shading the appropriate bubble.

TRUE	FALSE		
○	●	The c.d.f. F_X of a random variable X is a random variable.	*2pts*
○	●	If $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$, then $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}$, where F_{X_n} and F_X denote the cumulative distribution functions of X_n and X, respectively.	*2pts*
○	●	If X and Y are normal random variables, then (X, Y) is bivariate normal.	*2pts*
●	○	If a square matrix M is positive definite, then there exists a square matrix A such that $M = A^2$.	*2pts*
●	○	If $X \sim \text{Exp}(\lambda)$ for some $\lambda > 0$, then $e^{-\lambda X}$ is uniformly distributed over $(0, 1)$.	*2pts*
○	●	Let $\hat{F}_n$ denote the empirical c.d.f. from $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and define $D_n := \sup_x \lvert\hat{F}_n(x) - F(x)\rvert$. Then, $\sqrt{n}D_n \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.	*2pts*
○	●	It k_1, k_2, k_3 are valid kernels, then a function g defined as $g(x, x') = k_1(x, x')k_2(x, x') - k_3(x, x')$ is also a valid kernel.	*2pts*

[exam continued on next page]

2. Short Answers [*Answer is a single number or expression; write it in the box provided; no justification necessary. Total of 30 points. No penalty for incorrect answers.*]

(a) Let $X \sim \text{Uniform}(0,2)$ and $Y \sim \text{Uniform}(0,3)$ be independent random variables. Find the p.d.f. $f_Z(z)$ of $Z = X + Y$. *4pts*

$$f_Z(z) = \begin{cases} z/6, & 0 < z \le 2, \\ 1/3, & 2 < z \le 3, \\ (5-z)/6, & 3 < z < 5, \\ 0, & \text{otherwise.} \end{cases}$$

[This follows from convolution: $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = I\{z \in (0,5)\} \int_{\max(0,z-3)}^{\min(2,z)} \frac{1}{ab} dx$.]

(b) Let $X, Y \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$ for some $\lambda > 0$. Find $\mathbb{P}[X \le 1 \mid X + Y = 3]$. *3pts*

$\frac{1}{3}$. [This was covered in **Lab 8**, Problem 2, which showed that the conditional distribution of X given $X + Y = a$ is uniform over $[0, a]$.]

(c) Consider an urn containing N balls, B of which are blue and R are red, with $N = B + R$. Suppose $n < N$ balls are sampled uniformly at random from the urn *without* replacement. What is the *expected* number of color changes in the sequence of observed balls? (Note: If the observed sequence is $RBBBRB$ for $n = 6$, the number of color changes is 3.) *3pts*

$\frac{2BR(n-1)}{N(N-1)}$. [Let X_i denote the color of the ith ball. Then the number of color changes is given by $\sum_{i=1}^{n-1} I\{X_i \neq X_{i+1}\}$. By exchangeability, for all $i = 1, \ldots, n-1$, $\mathbb{P}[X_i \neq X_{i+1}] = \mathbb{P}[X_1 \neq X_2] = \frac{2BR}{N(N-1)}$, so $\mathbb{E}[\sum_{i=1}^{n-1} I\{X_i \neq X_{i+1}\}] = \frac{2BR(n-1)}{N(N-1)}$.]

(d) Alice and Bob are playing a game. Alice initially has 10 marbles and Bob has 7 marbles. In each round, a fair coin is tossed. If it shows heads, then Bob gives 1 marble to Alice; if it shows tails, then Alice gives 1 marble to Bob. They keep playing until either one has no marble left, when the game ends. What is the probability that Alice ends up winning all the marbles? *3pts*

$\frac{10}{17}$. [This problem is essentially the same as Problem 3 from **Lab 9**.]

(e) Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with p.d.f. $f(x) = \begin{cases} 2x, & \text{if } x \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$ *3pts*

Find the p.d.f. of the second order statistic $X_{(2)}$.

$2n(n-1)x^3(1-x^2)^{n-2} I\{x \in [0,1]\}$. [The c.d.f. corresponding to this problem is $F(x) = x^2$, so the answer follows from the general formula $f_{X_{(2)}}(x) = n(n-1) f(x) F(x) [1 - F(x)]^{n-2}$, which was covered in **Lecture 15**. It should be straightforward to derive this formula from scratch using the approach discussed in the lecture.]

(f) Let $f_{X,Y}(x,y)$ denote the joint density of random variables X and Y. For $U = X^2$ and $V = X + Y$, find the joint density $f_{U,V}(u,v)$ in terms of $f_{X,Y}$. *4pts*

$\frac{1}{2\sqrt{u}} \left[f_{X,Y}(\sqrt{u}, v - \sqrt{u}) + f_{X,Y}(-\sqrt{u}, v + \sqrt{u}) \right] I\{u \in [0, \infty)\}$

[This follows from the fact that the transformation $T(X,Y) = (U,V)$ is two-to-one and that $|\det(J)| = \frac{1}{2\sqrt{u}}$ for each preimage.]

[Q2 continued on next page]

(g) Suppose $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2(\vec{\mu}, \Sigma)$, where $\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. Find the mean and covariance of $\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. *3pts*

$\mathbb{E}[\vec{Y}] = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 + \mu_2 \end{pmatrix}$

$\mathrm{Cov}[\vec{Y}] = \begin{pmatrix} 0 & 0 \\ 0 & 4 \end{pmatrix}$

[These results follow from the fact that if $\vec{Y} = A\vec{X}$, then $\mathbb{E}[\vec{Y}] = A\,\mathbb{E}[\vec{X}]$ and $\mathrm{Cov}(\vec{Y}) = A\,\mathrm{Cov}(\vec{X})A^T$.]

(h) Let $\vec{X}_1, \ldots, \vec{X}_n$ be a sequence of i.i.d. random vectors in $\mathbb{R}^2$ with $\mathbb{E}[\vec{X}_i] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\mathrm{Cov}(X_i) = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$ for all $i \in \{1, \ldots, n\}$. Find $\lim_{n\to\infty} \mathbb{P}\left[\frac{\vec{X}_1 + \cdots + \vec{X}_n}{\sqrt{n}} \le \begin{pmatrix} a \\ b \end{pmatrix}\right]$, where a, b are real constants. Your answer may be left as an integral. *3pts*

$\frac{1}{\pi\sqrt{3}} \int_{-\infty}^{b} \int_{-\infty}^{a} e^{-\frac{2}{3}(x_1^2 - x_1x_2 + x_2^2)} dx_1 dx_2$.

[This follows from the multivariate CLT discussed in Lecture 20 and the p.d.f. of a bivariate normal distribution with mean zero and covariance $\Sigma = \mathrm{Cov}(X_i)$, for which $|\det(\Sigma)| = \frac{4}{3}$ and the precision matrix is $\Sigma^{-1} = \frac{4}{3}\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$.]

(i) Consider a discrete-time branching process $\{X_n, n \in \mathbb{N}_0\}$ with $X_0 = 1$ and the geometric offspring number distribution $P_k = q(1-q)^k$ for $k \in \mathbb{N}_0$, where $q \in (0, 1)$. Find the extinction probability. *4pts*

$\begin{cases} 1, & \text{if } q \in [\frac{1}{2}, 1), \\ \frac{q}{1-q}, & \text{if } q \in (0, \frac{1}{2}). \end{cases}$

[The probability generation function for this case is given by $\varphi(s) = \sum_{k=1} q(1-q)^k s^k = \frac{q}{1-s(1-q)}$. Letting $s = \varphi(s)$, we obtain $(s-1)[s(1-q) - q] = 0$, so the two roots are $s = 1$ and $s = q/(1-q)$. The smaller non-negative root corresponds to the extinction probability.]

[exam continued on next page]

3. Tail bounds [*Total of 16 points.*]
Let $X_1, \ldots, X_n$ be independent random variables taking values in $\{0, 2\}$, with $\mathbb{P}[X_i = 0] = \frac{3}{4}$ and $\mathbb{P}[X_i = 2] = \frac{1}{4}$, for all $i = 1, \ldots, n$. Let $S_n = X_1 + \cdots + X_n$.

(a) Find an upper bound on $\mathbb{P}[S_n \geq n]$ using Markov's inequality. No justification required. *3pts*

$\frac{1}{2}$.

(See Lecture 5) We have that $\mathbb{E}[X_1] = 0 \cdot \frac{3}{4} + 2 \cdot \frac{1}{4} = \frac{1}{2}$. Hence $\mathbb{E}[S_n] = n/2$. Markov's inequality gives

$$\mathbb{P}[S_n \geq n] \leq \frac{\mathbb{E}[S_n]}{n} = \frac{n/2}{n} = \frac{1}{2}.$$

(b) Find an upper bound on $\mathbb{P}[S_n \geq n]$ using Chebyshev's inequality. No justification required. *4pts*

$\frac{3}{n}$.

(See Lecture 5) We have that $\mathbb{E}[X_1^2] == 0^2 \cdot \frac{3}{4} + 2^2 \cdot \frac{1}{4} = 1$, so $\text{Var}[X_1] = \mathbb{E}[X_1]^2 - \mathbb{E}[X_1]^2 = \frac{3}{4}$ and $\text{Var}[S_n] = \frac{3n}{4}$. Chebyshev's inequality gives

$$\begin{aligned}\mathbb{P}[S_n \geq n] &= \mathbb{P}[S_n - \mathbb{E}[S_n] \geq n - \mathbb{E}[S_n]] \\ &\leq \mathbb{P}[|S_n - \mathbb{E}[S_n]| \geq n/2] \\ &\leq \frac{\text{Var}[S_n]}{(n/2)^2} = \frac{3n/4}{n^2/4} = \frac{3}{n}.\end{aligned}$$

(c) Find the moment generating function of S_n. No justification required. *4pts*

$M_{S_n}(t) = \left(\frac{3}{4} + \frac{1}{4}e^{2t}\right)^n$.

(See Lecture 7) Since the random variables X_i are IID we have that $M_{S_n}(t) = M_{X_1}(t)^n$. It is then enough to compute $M_{X_1}(t)$.

$$\begin{aligned}M_{X_1}(t) &= \mathbb{E}[e^{tX_1}] \\ &= e^{t\cdot 0}\frac{3}{4} + e^{t\cdot 2} \cdot \frac{1}{4} \\ &= \frac{3}{4} + \frac{1}{4}e^{2t}.\end{aligned}$$

[Q3 continued on next page]

(d) Find the *best* upper bound on $\mathbb{P}[S_n \geq n]$ using the Chernoff bound. Write your final answer in the box below, **and justify your answer in the space provided.** *5pts*

$\left(\frac{\sqrt{3}}{2}\right)^n$.

(See Lecture 8) Chernoff's bound gives

$$\mathbb{P}[S_n \geq n] \leq \inf_{t \geq 0} M_{S_n}(t)e^{-tn}.$$

So we need to minimize $M_{S_n}(t)e^{-tn}$, to this end we compute the logarithmic derivative to simplify calculations.

$$\begin{aligned}\frac{d}{dt}\ln\left(M_{S_n}(t)e^{-tn}\right) &= \frac{d}{dt}\left[n\ln\left(\frac{3}{4}+\frac{1}{4}e^{2t}\right)-tn\right] \\ &= \frac{2ne^{2t}}{3+e^{2t}} - n.\end{aligned}$$

We want to solve $\frac{2ne^{2t}}{3+e^{2t}} - n = 0$,

$$\begin{aligned}\frac{2ne^{2t}}{3+e^{2t}} = n &\iff 2e^{2t} = 3+e^{2t} \\ &\iff e^{2t} = 3 \\ &\iff t = \frac{\ln(3)}{2}.\end{aligned}$$

We can check that $\frac{d}{dt}\ln\left(M_{S_n}(t)e^{-tn}\right)$ is actually increasing, then $\frac{\ln(3)}{2}$ gives a minima of $M_{S_n}(t)e^{-tn}$. Hence,

$$\begin{aligned}\mathbb{P}[S_n \geq n] &\leq \inf_{t\geq 0} M_{S_n}(t)e^{-tn} \\ &= M_{S_n}(\frac{\ln(3)}{2})e^{-n\cdot\frac{\ln(3)}{2}} \\ &= \left(\frac{3}{4}+\frac{1}{4}e^{2\cdot\frac{\ln(3)}{2}}\right)^n e^{-n\cdot\frac{\ln(3)}{2}} \\ &= \left(\frac{3}{4}+\frac{3}{4}\right)^n 3^{-n/2} \\ &= \frac{3^n}{2^n}\cdot 3^{-n/2} = \left(\frac{\sqrt{3}}{2}\right)^n.\end{aligned}$$

[exam continued on next page]

4. A run of consecutive heads [*Total of 11 points.*]
A biased coin shows heads with probability $p \in (0,1)$. Let X_n denote the number of tosses until a run of n consecutive heads is obtained.

(a) Find $\mathbb{E}[X_1]$. No justification required. *3pts*

$\frac{1}{p}$. [This follows from the fact that $X_1 \sim$ Geometric(p).]

(b) For $n \geq 2$, find an equation involving $\mathbb{E}[X_n]$ and $\mathbb{E}[X_{n-1}]$. No justification required. *4pts*

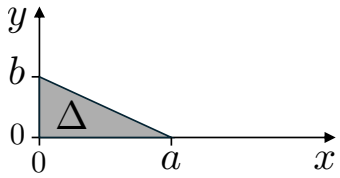
$\mathbb{E}[X_n] = \frac{1}{p} + \frac{1}{p}\mathbb{E}[X_{n-1}]$. [This result follows from $\mathbb{E}[X_n] = \mathbb{E}[\mathbb{E}[X_n|X_{n-1}]] = p(\mathbb{E}[X_{n-1}]+1) + (1-p)(\mathbb{E}[X_{n-1}]+1+\mathbb{E}[X_n])$, simplifying which yields the answer.]

(c) Find a closed-form expression for $\mathbb{E}[X_n]$ for $n \geq 1$. No justification required. *4pts*

$\frac{\frac{1}{p^n}-1}{1-p}$. [Parts (a) and (b) imply $\mathbb{E}[X_n] = \sum_{k=1}^{n} \frac{1}{p^k} = \frac{1-\frac{1}{p^{n+1}}}{1-\frac{1}{p}} - 1 = \frac{\frac{1}{p}-\frac{1}{p^{n+1}}}{1-\frac{1}{p}} = \frac{\frac{1}{p^n}-1}{1-p}$.]

[exam continued on next page]

5. Uniform distribution over a triangle [*Total of 18 points.*]
Consider a solid triangle $\Delta \in \mathbb{R}^2$ with corners at $(0,0)$, $(a,0)$, and $(0,b)$, where $a, b > 0$, as shown in the figure below. Suppose (X,Y) is uniformly distributed over Δ.



(a) Find the joint density $f_{X,Y}(x,y)$. No justification required. *3pts*

$\frac{2}{ab} I\{(x,y) \in \Delta\}$. [Since (X,Y) is uniformly distributed over Δ, $f_{X,Y}$ should be a constant function over Δ. The normalization constant is determined by $\int f_{X,Y}(x,y)dxdy = 1$, yielding $f_{X,Y}(x,y) = \frac{1}{\text{Area}(\Delta)} I\{(x,y) \in \Delta\} = \frac{2}{ab} I\{(x,y) \in \Delta\}$.]

(b) Find the marginal density $f_Y(y)$. No justification required. *3pts*

$\frac{2}{b}\big(1 - \frac{y}{b}\big) I\{0 \le y \le b\}$. [The line going through $(a,0)$ and $(0,b)$ is defined by the equation $y = b - \frac{b}{a}x$, so $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = I\{0 \le y \le b\} \int_0^{a - \frac{a}{b}y} \frac{2}{ab} dx = \frac{2}{ab}\big(a - \frac{a}{b}y\big) I\{0 \le y \le b\}$, which simplifies to the desired result.]

(c) Find $f_{X|Y=y}(x)$, the conditional density of X given $Y = y$. No justification required. *3pts*

Given $y \in [0,b]$, $f_{X|Y=y}(x) = \frac{1}{a\big(1 - \frac{y}{b}\big)} I\{0 \le x \le a\big(1 - \frac{y}{b}\big)\}$ [This result follows from combining the results from parts (a) and (b): $f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$.]

(d) Show $\mathbb{E}[X] = \frac{a}{2}\Big(1 - \frac{\mathbb{E}[Y]}{b}\Big)$. **Justify your answer in the space provided.** *5pts*

$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)dx = \int_0^{a(1-\frac{y}{b})} x \frac{1}{a(1-\frac{y}{b})} dx = \frac{1}{2}a(1 - \frac{y}{b})$. So, by the Law of Total Expectation, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[\frac{1}{2}a\big(1 - \frac{Y}{b}\big)] = \frac{1}{2}a\big(1 - \frac{\mathbb{E}[Y]}{b}\big)$, where the last equality follows from the linearity of expectation.

(e) Find $\mathbb{E}[X]$ and $\mathbb{E}[Y]$. No justification required. *4pts*

$\mathbb{E}[X] = \frac{a}{3}$

$\mathbb{E}[Y] = \frac{b}{3}$

In part (d), we showed $\mathbb{E}[X] = \frac{a}{2}\big(1 - \frac{\mathbb{E}[Y]}{b}\big)$. By symmetry, we have $\mathbb{E}[Y] = \frac{b}{2}\big(1 - \frac{\mathbb{E}[X]}{a}\big)$. Solving this coupled system of linear equations for $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ yields the above results.

[exam continued on next page]

6. Brownian motion and bridge [*Total of 16 points.*]
A stochastic process $\{X_t, t \geq 0\}$ is said to have *independent* increments if, for every choice of times $0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n < \infty$, $\{X_{t_i} - X_{s_i}, i = 1, \ldots, n\}$ are jointly independent. Furthermore, increments are called *stationary* if, for all $0 < s, t < \infty$, $X_{t+s} - X_s$ has the same distribution as $X_t - X_0$. Let $\{B_t, t \geq 0\}$ be a Brownian motion.

(a) Show that $\{B_t, t \geq 0\}$ has independent increments. **Justify your answer in the space provided.** *4pts*

(See Lab 10) Because the distribution of the increments $\{B_{t_i} - B_{s_i}, i = 1, \ldots, n\}$ is a multivariate Gaussian, it is enough to prove that their covariance is 0 to verify independence. Given $a < b \leq c < d$ we want to compute $\text{Cov}(B_d - B_c, B_b - B_a)$ by using the bilinearity of the covariance and that the kernel of a Brownian motion $\text{Cov}(B_s, B_t) = \min(s, t)$ **(See Lecture 21)**.

$$\begin{aligned}\text{Cov}(B_d - B_c, B_b - B_a) &= \text{Cov}(B_d, B_b) - \text{Cov}(B_d, B_a) - \text{Cov}(B_c, B_b) + \text{Cov}(B_c, B_a) \\ &= \min(d, b) - \min(d, a) - \min(c, b) + \min(c, a) \\ &= b - a - b + a = 0.\end{aligned}$$

This proves that the increments are independent.

(b) Show that $\{B_t, t \geq 0\}$ has stationary increments. **Justify your answer in the space provided.** *3pts*

(See Lab 10) Similarly to (a), we know that the increments are Gaussian with mean 0, it is then enough to verify they have the same variance.

$$\text{Var}(B_{t+s} - B_s) = \text{Cov}(B_{t+s} - B_s, B_{t+s} - B_s).$$

Using part (a) with $a = c = s$ and $b = d = t + s$ we obtain

$$\text{Var}(B_{t+s} - B_s) = t.$$

Since the result doesn't depend on s we conclude that the Brownian motion has stationary increments.

(c) Show that $X_t = B_t - tB_1$ is a Brownian bridge for $t \in [0, 1]$. **Justify your answer in the space provided.** *4pts*

(See Labs 10 and 11) Since each X_t is a linear combination of the B_t we get that $\{X_t, 0 \leq t \leq 1\}$ is a Gaussian process. Hence, it is enough to compute it's kernel (*i.e.* Covariance function) to determine if it is a Brownian bridge.

$$\begin{aligned}\text{Cov}(X_t, X_s) &= \text{Cov}(B_t - tB_1, B_s - sB_1) \\ &= \text{Cov}(B_t, B_s) - s\text{Cov}(B_t, B_1) - t\text{Cov}(B_1, B_s) + st\text{Cov}(B_1, B_1) \\ &= \min(t, s) - st - st + st = \min(t, s) - st.\end{aligned}$$

This is precisely the kernel of a Brownian bridge **(See Lecture 21)**.

(d) Find the conditional density $f_{B_3|(B_1,B_2)=(x,y)}(z)$ of B_3 given $(B_1, B_2) = (x, y)$. No justification required. *5pts*

$$f_{B_3|(B_1,B_2)=(x,y)}(z) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(z-y)^2}{2}\Big).$$

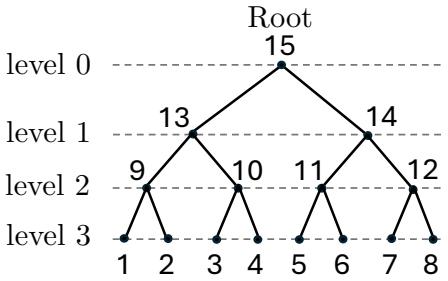
(See Lecture 19 and Homework 5) Intuitively, By parts (a) and (b) the distribution of a Brownian motion at time 3 will only depend on the information we have at time 2 and not at time 1, we should have that B_3 has normal distribution centered at 0 with variance given the the distance between 2 and 3, hence $\mathcal{N}(y, 1)$.

Formally, by part (a), $B_3 - B_2$ is independent of $B_2 = B_2 - B_0$ and of $B_1 = B_1 - B_0$. since, conditioned to $(B_1, B_2) = (x, y)$, $B_3 = B_3 - B_2 + y$ we have that B_3 given $(B_1, B_2) = (x, y)$ is distributed as $B_3 - B_2 + y$ without any conditional. By part (b) we know that the distribution of $B_3 - B_2$ is the same as $B_1 - B_0 = B_1 \sim_d \mathcal{N}(0, 1)$. We conclude that B_3 given $(B_1, B_2) = (x, y)$ is distributed as $\mathcal{N}(y, 1)$.

An alternative solution consists in using the conditional distribution formulas for multivariate Gaussians given in lecture 19.

[exam continued on next page]

7. Random walk on an undirected binary tree [*Total of 15 points.*]
Consider a perfect rooted binary tree of depth d, which has $n = 2^d$ leaves. Let V denote the vertex set. Shown below is an example for $d = 3$.



In this problem, we will analyze the following random walk $\{X_n, n \in \mathbb{N}_0\}$ on the tree: From a vertex $v \in V$ of degree $\deg(v)$, one moves to a specific neighbor with probability $\frac{1}{\deg(v)}$. (The degree of a vertex refers to the number of edges attached to the vertex. The root has degree 2, while all other internal nodes have degree 3. All leaves have degree 1.) Let P denote the transition probability matrix for this Markov chain.

(a) Show that $\pi_v = \frac{\deg(v)}{2(2^{d+1}-2)}$ for $v \in V$ is a stationary distribution of the Markov chain. *4pts*

We need to prove both that $\pi P = \pi$ and that π is a probability measure. We start by verifying $\pi P = \pi$. Each vertex v has $\deg(v)$ vertex neighborhood $\{v_1, \ldots, v_{\deg(v)}\}$. Additionally the transition probabilities $P(v_i \to v) = \frac{1}{\deg(v_i)}$. We get

$$\begin{aligned}\sum_{i=1}^{\deg(v)} \pi_{v_i} P(v_i \to v) &= \sum_{i=1}^{\deg(v)} \frac{\deg(v_i)}{2(2^{d+1}-2)} \cdot \frac{1}{\deg(v_i)} \\ &= \sum_{i=1}^{\deg(v)} \frac{1}{2(2^{d+1}-2)} \\ &= \frac{\deg(v)}{2(2^{d+1}-2)} = \pi_v.\end{aligned}$$

Let's verify it is a probability measure, i.e. we want to verify that $\sum_{v\in V} \pi_v = 1$.

$$\begin{aligned}\sum_{v\in V} \pi_v &= \sum_{v\in V} \frac{\deg(v)}{2(2^{d+1}-2)} \\ &= \frac{1}{2(2^{d+1}-2)} \sum_{v\in V} \deg(v) \\ &= \frac{1}{2(2^{d+1}-2)} \cdot (2\#\{\text{Edges}\}) \\ &= \frac{1}{2(2^{d+1}-2)} \cdot 2 \cdot \left(2 + 4 + \cdots + 2^d\right) \\ &= \frac{1}{2(2^{d+1}-2)} \cdot 4 \cdot \left(1 + 2 + \cdots + 2^{d-1}\right) \\ &= \frac{1}{2(2^{d+1}-2)} \cdot \frac{4(2^d-1)}{2-1} = 1.\end{aligned}$$

(b) Can there be other stationary distributions? Shade the correct bubble. ○ Yes ● No *2pts*
This finite state Markov chain is irreducible so a stationary distribution must be unique.

(c) For $u, v \in V$, does $\lim_{n\to\infty}[P^n]_{uv}$ exist? Shade the correct bubble. ○ Yes ● No *2pts*
If YES, what does it converge to? If NO, leave the box blank.

This Markov chain has period 2, h't have convergence.

[Q7 continued on next page]

(d) Let A denote the set of 2^d leaves and define $T_A = \min\{n \in \mathbb{N}_0 : X_n \in A\}$. We wish to compute $\mathbb{E}[T_A \mid X_0 = \text{Root}]$, the expected time of hitting A given that the walk starts from the root. Let $h_{i,j}$ denote the expected hitting time from level i to level j of the tree. Then, note that $h_{0,d} = \mathbb{E}[T_A \mid X_0 = \text{Root}]$. Find a system of equations relating $h_{i,j}$, together with a suitable boundary condition that will allow us to find $h_{0,d}$. No justification required. *7pts*

$$h_{d,d} = 0 \tag{1}$$
$$h_{0,d} = 1 + h_{1,d} \tag{2}$$
$$h_{i,d} = 1 + \frac{1}{3}h_{i-1,d} + \frac{2}{3}h_{i+1,d}, \quad \text{for } i \in \{1, \dots, d-1\} \tag{3}$$

Let $X_{i,j}$ the random variable denoting the hitting time starting from level i to level d of the tree, then $\mathbb{E}[X_{i,d}] = h_{i,d}$. Let A the event in which we move a level up as a first move and $B = A^c$ the event in which we move a level down as a first move. For $i < d$ Conditioning on A and B we have that

$$\mathbb{E}[X_{i,d}] = \mathbb{E}[X_{i,d}|A]\mathbb{P}(A) + \mathbb{E}[X_{i,d}|B]\mathbb{P}(B).$$

Notice that $\mathbb{E}[X_{i,d}|A] = \mathbb{E}[X_{i-1,d}] + 1$ and $\mathbb{E}[X_{i,d}|B] = \mathbb{E}[X_{i+1,d}] + 1$ Additionally, $X_{d,d} = 0$, if $i = 0$, then $P(A) = 0$ and $P(B) = 1$ and if $0 < i < d$, then $P(A) = 1/3$ and $P(B) = 2/3$ and $X_{d,d} = 0$. We obtain the equations (1), (2) and (3) respectively.

Intuitively, this system of equations allow us to calculate $h_{0,d}$ since it provides a recurrent formula. We have that $h_{i,d} = 3h_{i+1,d} - 2h_{i+2,d} - 3$ for $0 \le i \le d-2$, since $h_{d,d}$=0. Let $g_{i,d} = h_{i,d} + 3i - 3d$, we recover the relations $g_{d,d} = 0$, $g_{0,d} = g_{1,d} - 2$ and

$$\begin{aligned} g_{i,d} &= h_{i,d} + 3i - 3d \\ &= 3h_{i+1,d} - 2h_{i+2,d} - 3 + 3i - 3d \\ &= 3(h_{i+1,d} + 3(i+1) - 3d) - 2(h_{i+2,d} + 3(i+2) - 3d) \\ &= 3g_{i+1,d} - 2g_{i+2,d}. \end{aligned}$$

This relation gives $g_{d-i,d} = (2^i - 1)g_{d-1,d}$ for $0 \le i \le d$. In particular we have $g_{0,d} = (2^d - 1)g_{d-1,d}$ and $g_{1,d} = (2^{d-1} - 1)g_{d-1,d}$. Using equation (2) gives $(2^d - 1)g_{d-1,d} = (2^{d-1} - 1)g_{d-1,d} - 2$ and $g_{d-1,d} = -1/2^{d-2}$. Hence $g_{0,d} = -(2^d - 1)/2^{d-2} = 1/2^{d-2} - 4$ and finally $h_{0,d} = 3d - 4 + 1/2^{d-2}$.

[End of Exam!]

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Practice Problems for the Midterm Exam

Note: You are not expected to solve all these problems in just 80 minutes.

1. Determine whether each of the following claims is true or false. Provide reasons in each case.

   (a) It is often said that $\text{Bin}(n, p)$ is well-approximated by the $N(np, np(1-p))$ distribution. When $n = 3710$ and $p = 0.2$, this would mean that $\text{Bin}(3710, 0.2)$ is well-approximated by $N(742, 593.6)$. Therefore

   $$\frac{\mathbb{P}(\text{Bin}(3710, 0.2) \geq 941)}{\mathbb{P}\{N(742, 593.6) \geq 941\}}$$

   should be close to 1 (you might note here that $941/3710 \approx 0.254$).

   (b) Suppose X has the Negative Binomial distribution with parameters k and p (for example, X can be thought of as the distribution of the number of independent tosses of a coin with probability of heads p required to get the k^{th} head). Let $F_X(\cdot)$ denote the cdf of X. Then $F_X(X)$ has the uniform distribution on $(0, 1)$.

   (c) Suppose X has the geometric distribution with parameter p(X can be thought of as the number of independent tosses of a coin with probability of heads p to get the first head). Then

   $$\mathbb{P}\{X > 3.5 + 1.5 \mid X > 1.5\} = \mathbb{P}\{X > 3.5\}$$

   (d) We can generate a random variable having any specified distribution by first generating a uniformly distributed random variable on $(0, 1)$ and then by applying an appropriate transformation to the uniform random variable.

2. Short questions.

(a) There are 8 parents, 24 students and 3 teachers in a room. If a person is selected at random, what is the probability that it is a teacher or a student?

(b) Find the probability to see 3 or less tails in 4 flips of a coin.

(c) Suppose that A and B are independent, $\mathbb{P}(A) = 1/3$ and $\mathbb{P}(B) = 1/7$. Calculate $\mathbb{P}(A \cap B^c)$.

(d) Suppose a box has 4 red marbles and 3 black ones. We select 2 marbles. What is the probability that second marble is red given that the first one is red?

(e) Suppose the random variable X has possible values $\{1, 2, 3\}$ and probability mass function of the form $\mathbb{P}(X = k) = ck$. Find c. Find $\mathbb{E}[X]$. Find $\mathrm{Var}(X)$.

(f) Let X be a random variable with exponential distribution with parameter 2. Find $\mathbb{P}(X > 14 \mid X > 4)$.

(g) Russel has a biased coin for the which the probability of getting tails is an unknown p. He decide to flip the coin n and writes the total number of times X he gets tails. How large should n be in order to know with at least 0.95 certainty that the true p is within 0.1 of the estimate X/n? What if he wants 0.99 certainty?

(h) Let X and Y be independent random variables with exponential distribution with parameter λ, find $\mathbb{P}(X > Y)$.

(i) Let X be a random variable with m.g.f. $M_X(t) = e^{5t} - e^{3t}$. Find a formula for the moments of X.

(j) Let X be a non-negative random variable with $\mathbb{E}[X] = 2$ and $\mathbb{E}[X^2] = 5$. Use Markov's inequality to find an upper bound for $\mathbb{P}(X > 10)$. Use Chebyshev's inequality to find an upper for $\mathbb{P}(X > 10)$.

3. Consider the urn setting that we discussed in lecture. We have an urn with R red balls and $N - R$ white balls. We draw balls in sequence from the urn without replacement.

   (a) Calculate $\mathbb{P}(F)$ where F denotes the proposition that the first red ball is drawn before the third white ball.

   (b) Calculate $\mathbb{P}(E)$ where E denotes the proposition that, when we draw n balls, our sample contains at least one red ball and at least two white balls.

4. Take random variables $X_1, X_2, X_3, \ldots$ such that each of them has mean μ and variance 1.

   (a) Suppose that X_i are *negatively correlated*, i.e. $\mathrm{Cov}(X_i, X_j) < 0$ for all i, j. Set $S_n = X_1 + \cdots + X_n$. Show that (IMPORTANT: X_i are not independent!)

$$\mathrm{Var}\left(\frac{S_n}{n}\right) \leq \frac{1}{n}. \tag{1}$$

   (b) Assume instead that X_i are *positively correlated*, i.e. $\mathrm{Cov}(X_i, X_j) > 0$ for all i and j. Is (1) still true? Either give a proof or provide a counterexample.

5. (a) In Bernoulli (p) trials let V_n be the number of trials required to produce either n successes or n failures, whichever comes first. Find the distribution of V_n.

   (b) Suppose n balls are thrown independently at random into b boxes. Let X be the number of boxes left empty. Find expressions for $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

6. Suppose X and Y are independent random variables with X having the Exponential distribution with rate parameter λ and Y having the Standard Cauchy distribution. Let

$$U := \frac{Y\sqrt{X}}{\sqrt{1+Y^2}} \quad \text{and} \quad V := \frac{\sqrt{X}}{\sqrt{1+Y^2}}$$

   (a) Find the joint density of U and V.

   (b) Find the marginal densities of U and V.

   (c) Are U and V independent? Why or why not?

Midterm Exam Solutions

3:40-5:00pm, October 17, 2024

Your First Name: **Your Last Name:**

SIGN Your Name: **Your SID Number:**

Instructions:

(a) *As soon as the exam starts, please write your student ID in the space provided at the top of every page! (We will remove the staple when scanning your exam.)*

(b) *There are* **5 double-sided** *sheets (10 numbered pages) on the exam. Notify a proctor immediately if a sheet is missing.*

(c) *We will not grade anything outside of the space provided for a question (i.e., either a designated box if it is provided, or otherwise the white space immediately below the question).* **Be sure to write your full answer in the box or space provided!** *Scratch paper is provided on request; however, please bear in mind that nothing you write on scratch paper will be graded!*

(d) ***You may use, without proof, theorems and lemmas that were proved in lecture and/or in homework.***

(e) *You may consult a single two-sided "cheat sheet" of notes. Apart from that, you may not look at any other materials. Calculators, phones, computers, and other electronic devices are NOT permitted.*

(f) *You have 80 minutes: there are 4 questions on this exam worth a total of 75 points.*

Problem	Points
1	12
2	21
3	20
4	22

(g) *On questions 1-2, you need only give the answer in the format requested (e.g., True/False, an expression, a statement.) An expression may simply be a number or an expression with a relevant variable in it. For short answer questions, correct, clearly identified answers will receive full credit with no justification. Incorrect answers may receive partial credit.*

(h) *On questions 3-4, you should give arguments, proofs or clear descriptions* ***if requested****. If there is a box, you must use it for your answer; answers written outside the box may not be graded!*

[exam starts on next page]

1. True/False [*No justification; answer by shading the correct bubble. 2 points per answer; total of 12 points. No penalty for incorrect answers.*]

Indicate which of the following statements is **TRUE** or **FALSE** by shading the appropriate bubble.

In this problem, let $(\Omega, \mathscr{F}, \mathbb{P})$ denote a probability space.

TRUE	**FALSE**		
●	○	If $A_i \in \mathscr{F}$ for all $i \in \mathbb{N}$, then $\cap_{i=1}^{\infty} A_i \in \mathscr{F}$.	*2pts*
○	●	Let X be a non-negative random variable and $c > 0$ some constant. Chebyshev's inequality always gives a stronger bound on $\mathbb{P}[X > c]$ than that given by Markov's inequality.	*2pts*
○	●	If $X_1, X_2, X_3, \ldots$ are i.i.d. random variables, then the sample average $\frac{S_n}{n} := \frac{X_1 + \cdots + X_n}{n}$ satisfies $\lim_{n\to\infty} \mathbb{P}\big[\big\lvert \frac{S_n}{n} - \mathbb{E}[X_1]\big\rvert < 0.01\big] = 1$.	*2pts*
○	●	Let $X_1, X_2, X_3, \ldots$ and X be random variables on the sample probability space, and suppose $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$. Then, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega$.	*2pts*
○	●	Let $X_1, X_2, X_3, \ldots$ and X be random variables on the sample probability space. It is possible to have $\lim_{n\to\infty} \mathbb{E}\big[\,(X_n - X)^4\,\big] = 0$ while X_n does not converge in distribution to X as $n \to \infty$.	*2pts*
●	○	For all random variables X with $\mathbb{E}[X] = \mu < \infty$, their moment generating functions $M_X(t)$ satisfy $e^{t\mu} \le M_X(t)$ for all $t \in \mathbb{R}$.	*2pts*

[exam continued on next page]

2. Short Answers [*Answer is a single number or expression; write it in the box provided; no justification necessary. Total of 21 points. No penalty for incorrect answers.*]

(a) Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, suppose $A, B \in \mathscr{F}$. If $\mathbb{P}[A] = 1/2$, $\mathbb{P}[B] = 1/2$ and $\mathbb{P}[A \cap B] = 1/5$, what is $\mathbb{P}[A^c \cap B^c]$? *3pts*

$\frac{1}{5}$. [Note that $\mathbb{P}[A^c \cap B^c] = \mathbb{P}[(A \cup B)^c]$], while $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] = \frac{1}{2} + \frac{1}{2} - \frac{1}{5} = \frac{4}{5}$. Hence, $\mathbb{P}[(A \cup B)^c] = 1 - \mathbb{P}[(A \cup B)] = \frac{1}{5}$.]

(b) Suppose X_1, X_2, X_3, X_4 are i.i.d. Normal$(0, 1)$ random variables. Find $\mathbb{P}[X_3 < X_1 < X_2 \mid X_4 > 0]$. *3pts*

$\frac{1}{6}$. [This problem is similar to Q1(a) of Lab 2. Independence implies $\mathbb{P}[X_3 < X_1 < X_2 \mid X_4 > 0] = \mathbb{P}[X_3 < X_1 < X_2]$, while exchangeability implies $\mathbb{P}[X_3 < X_1 < X_2] = \frac{1}{3!}$.]

(c) Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim}$ Bernoulli$(\frac{1}{3})$. To what does $\frac{1}{n}\sum_{i=1}^n I\{X_i \le \frac{1}{2}\}$ converge almost surely as $n \to \infty$, where $I\{X_i \le x\}$ denote indicator random variables. *3pts*

$\frac{2}{3}$. [SLLN implies $\frac{1}{n}\sum_{i=1}^n I\{X_i \le x\} \xrightarrow{\text{a.s.}} F(x)$ as $n \to \infty$, where F denotes the c.d.f. of Bernoulli$(\frac{1}{3})$, and $F(1/2) = 2/3$.]

(d) Let $X \sim$ Poisson(λ), where $\lambda > 0$. Recall that $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}[X] = \lambda$. Find $\lim_{\lambda\to\infty} \mathbb{P}[X - \lambda \le \sqrt{\lambda}]$. Write your answer as an integral; you do not need to evaluate it. *4pts*

$\int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. [This follows from $\frac{X-\lambda}{\sqrt{\lambda}} \xrightarrow{d}$ Normal$(0, 1)$ as $n \to \infty$, which in turn follows from the CLT, together with the fact that a sum of independent Poisson random variables is also Poisson with rate given by the sum of individual rates.]

(e) Let $X \sim$ Normal$(0, 1)$ and $Y = \frac{1}{2}|X|$. Find the pdf of Y. *4pts*

$$f_Y(y) = \begin{cases} \frac{4}{\sqrt{2\pi}} e^{-2y^2}, & y \ge 0, \\ 0, & y < 0. \end{cases}$$

$\left[f_Y(y) = f_X(2y)\left|\frac{d(2y)}{dy}\right| + f_X(-2y)\left|\frac{d(-2y)}{dy}\right| = 4 f_X(2y). \right]$

(f) Consider the linear transformation $\begin{pmatrix} U \\ V \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Find the joint density $f_{U,V}(u, v)$ in terms of $f_{X,Y}$. (**Hint**: Note that $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ is an orthogonal matrix.) *4pts*

$f_{U,V}(u, v) = f_{X,Y}\left(\frac{1}{\sqrt{2}}(u - v + 1), \frac{1}{\sqrt{2}}(u + v - 1)\right)$. $\left[M_T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$ and its inverse is $M_S = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. Hence, $\begin{pmatrix} X \\ Y \end{pmatrix} = M_S \left[\begin{pmatrix} U \\ V \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right] = \frac{1}{\sqrt{2}} \begin{pmatrix} U - V + 1 \\ U + V - 1 \end{pmatrix}$. Also, note that $|\det M_T| = 1$. $\right]$

[exam continued on next page]

3. Lazy random walk [*Total of 20 points.*]
A particle takes a lazy random walk in an infinite 1-dimensional lattice $\mathbb{Z} := \{\ldots, -3, -2, -1, 0, +1, +2, +3, \ldots\}$, starting at the origin at time 0. Then, at every $\frac{1}{n}$ second, it moves one step to the right with probability p, moves one step to the left with probability q, or stays at the current position with probability $1 - p - q$.

(a) What is the conditional probability that the particle moves to the right given that it moves? No justification required. *3pts*

$\frac{p}{p+q}$.

(b) What is the expected waiting time (in seconds) until the particle makes its first move? No justification required. *3pts*

$\frac{1}{n(p+q)}$. [The number of trials until the particle moves is distributed as Geometric$(p+q)$, which has mean $\frac{1}{p+q}$. One trial is made every $\frac{1}{n}$ second, so the expected waiting time in seconds is $\frac{1}{n(p+q)}$.]

IMPORTANT: For the remainder of this problem, parts (c)-(f), assume $p = \frac{\lambda}{n}$ and $q = \frac{\mu}{n}$, where λ and μ are positive constants, and consider the limit as $n \to \infty$. In this limit, let R (respectively, L) denote the total number of right (respectively, left) steps taken in the time interval $[0, t]$ measured in seconds.

(c) What is the distribution of $R + L$? Write your final answer in the box below, **and justify your answer in the space provided**. *6pts*

Poisson$\big((\lambda + \mu)t\big)$. [First, note that this problem is a slight variant of the convergence result covered in Lecture 11 (see pages 2-3 of Lecture 11 Notes). At every $\frac{1}{n}$ second, the probability that the particle moves is $p + q = \frac{\lambda+\mu}{n}$. Let Y_n denote the number of times the particle moves in 1 second, which has n trials. Then, $Y_n \sim$ Binomial$(n, \frac{\lambda+\mu}{n})$, and it was proved in lecture that $Y_n \xrightarrow{d} Y \sim$ Poisson$(\lambda + \mu)$ as $n \to \infty$. Hence, in the time interval $[0, t]$, the total number of moves is distributed as Poisson$((\lambda + \mu)t)$.]

(d) Are R and L independent? Shade the correct bubble. ● Yes ○ No *2pts*
[Recall Poissonization of the multinomial covered in Lecture 11 (pages 3-4).]

(e) What is the expected position (which is given by $R - L$) of the particle at time t? No justification required. *3pts*

$(\lambda - \mu)t$. [Again, by Poissonization of the multinomial (in fact, binomial in this case), $R \sim$ Poisson(λt) and $L \sim$ Poisson(μt), and they are independent (independence is not needed to solve this problem, however). In summary, $\mathbb{E}[R - L] = \mathbb{E}[R] - \mathbb{E}[L] = \lambda t - \mu t = (\lambda - \mu)t$.]

(f) In the $n \to \infty$ limit described above, let X denote the waiting time (in seconds) until the particle makes a move to the right. What is the distribution of X? No justification required. *3pts*

$X \sim$ Exp(λ). [This follows from Application 1 (page 5) of Lecture 7.]

[exam continued on next page]

4. High-Dimensional Random Vectors [*Total of 22 points.*]

Let $\vec{X} = (X_1, \ldots, X_n)$ and $\vec{Y} = (Y_1, \ldots, Y_n)$ denote random vectors in $\{-1, +1\}^n$, where $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are i.i.d random variables with $\mathbb{P}[X_i = -1] = \mathbb{P}[X_i = +1] = \mathbb{P}[Y_i = -1] = \mathbb{P}[Y_i = +1] = \frac{1}{2}$ for all $i = 1, \ldots, n$. The Euclidean norm $\|\vec{X}\|$ is defined as $\sqrt{\vec{X} \cdot \vec{X}} = \sqrt{X_1^2 + \cdots + X_n^2}$.

(a) Find $\mathbb{E}[\|\vec{X} - \vec{Y}\|^2]$. No justification required. *3pts*

$2n$. $\left[\mathbb{E}[(X_i - Y_i])^2] = \frac{1}{2}(2)^2 = 2$, so $\mathbb{E}[\|\vec{X} - \vec{Y}\|^2] = \sum_{i=1}^n \mathbb{E}[(X_i - Y_i)^2] = 2n.\right]$

(b) Find $\mathrm{Var}[\|\vec{X} - \vec{Y}\|^2]$. No justification required. *3pts*

$4n$. $\Big[\, \mathbb{E}[(X_i - Y_i)^4] = \frac{1}{2}(2)^4 = 8$. Hence, $\mathrm{Var}[(X_i - Y_i)^2] = \mathbb{E}[(X_i - Y_i)^4] - (\mathbb{E}[(X_i - Y_i)^2])^2 = 8 - 2^2 = 4$. By independence, $\mathrm{Var}[\sum_{i=1}^n (X_i - Y_i)^2] = \sum_{i=1}^n \mathrm{Var}[(X_i - Y_i)^2] = 4n.\, \Big]$

(c) Let $\mu = \mathbb{E}[\|\vec{X} - \vec{Y}\|^2]$. Find Hoeffding's bound on $\mathbb{P}\big[\, \big|\|\vec{X} - \vec{Y}\|^2 - \mu\big| \geq \varepsilon n\big]$, where $\varepsilon > 0$ is a constant. No justification required. *4pts*

$2e^{-\varepsilon^2 n/8}$. $[(X_i - Y_i)^2 \in \{0, 4\}$, so using $(b_i - a_i)^2 = (4 - 0)^2 = 16$ in Hoeffding's inequality gives $\mathbb{P}\big[\, \big|\|\vec{X} - \vec{Y}\|^2 - \mu\big| \geq \varepsilon n\big] \leq 2e^{-2\varepsilon^2 n^2/(16n)} = 2e^{-\varepsilon^2 n/8}$.$]$

(d) For any fixed non-zero vector $\vec{v} \in \mathbb{R}^n$ and any constant $c > 0$, prove that $\mathbb{P}[|\vec{v} \cdot \vec{X}| \geq c\|\vec{v}\|] \leq 2e^{-c^2/2}$. (**Hint:** Use Hoeffding's inequality. Alternatively, you can also prove this result using the Chernoff's inequality together with the identity $\frac{1}{(2k)!} \leq \frac{1}{2^k k!}$ for $k = 0, 1, 2, \ldots$) *6pts*

Approach 1. Proof using Hoeffding's inequality: Define $Y_i := v_i X_i$ and $S_n := \vec{v} \cdot \vec{X} = \sum_{i=1}^n Y_i$, and note that Y_i are bounded random variables; more precisely, $\mathbb{P}[-v_i \leq Y_i \leq v_i] = 1$. Since $\mathbb{E}[Y_i] = \frac{1}{2}(v_i - v_i) = 0$, we have $\mathbb{E}[S_n] = 0$. Hence, Hoeffding's inequality implies that for any $\varepsilon > 0$,

$$\mathbb{P}[|S_n| \geq \varepsilon] \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^n (2v_i)^2}. \tag{1}$$

Noting $\sum_{i=1}^n (2v_i)^2 = 4\|\vec{v}\|^2$ and setting $\varepsilon = c\|\vec{v}\|$ in (2) yields the desired result.

Approach 2. Proof using Chernoff's inequality: Chernoff's inequality implies that for all $t > 0$,

$$\mathbb{P}[\vec{v} \cdot \vec{X} \geq c\|\vec{v}\|] \leq \frac{\mathbb{E}[e^{t\vec{v} \cdot \vec{X}}]}{e^{tc\|\vec{v}\|}}.$$

Since $X_1, \ldots, X_n$ are independent,

$$\mathbb{E}[e^{t\vec{v} \cdot \vec{X}}] = \prod_{i=1}^n \mathbb{E}[e^{tv_i X_i}] = \prod_{i=1}^n \frac{1}{2}(e^{tv_i} + e^{-tv_i}) = \prod_{i=1}^n \sum_{k=0}^\infty \frac{(tv_i)^{2k}}{(2k)!},$$

where the last equality follows from the fact that odd terms cancel when you Taylor expand $e^{tv_i} + e^{-tv_i}$. Now, the algebraic identity mentioned in the hint gives $\sum_{k=0}^\infty \frac{(tv_i)^{2k}}{(2k)!} \leq \sum_{k=0}^\infty \frac{(tv_i)^{2k}}{2^k k!} = e^{t^2 v_i^2/2}$, which implies $\mathbb{E}[e^{t\vec{v} \cdot \vec{X}}] \leq e^{t^2\|\vec{v}\|^2/2}$. In summary, we have

$$\mathbb{P}[\vec{v} \cdot \vec{X} \geq c\|\vec{v}\|] \leq e^{\frac{1}{2}t^2\|\vec{v}\|^2 - tc\|\vec{v}\|}.$$

Now, define $g(t) = e^{\frac{1}{2}t^2\|\vec{v}\|^2 - tc\|\vec{v}\|}$ and note that $g'(t) = g(t)(t\|\vec{v}\|^2 - c\|\vec{v}\|) = 0$ at $t = t^* := c/\|\vec{v}\| > 0$ (since $c > 0$). Furthermore, $g''(t) = \|\vec{v}\|^2 g(t) + g'(t) > 0$ at $t = t^*$, implying that $g(t)$ is minimized at $t = t^*$. Finally, since $g(t^*) = -c^2/2$, we have $\mathbb{P}[\vec{v} \cdot \vec{X} \geq c\|\vec{v}\|] \leq e^{-c^2/2}$. By symmetry, $\mathbb{P}[\vec{v} \cdot \vec{X} = a] = \mathbb{P}[\vec{v} \cdot \vec{X} = -a]$ for all $a \in \mathbb{R}$, so $\mathbb{P}[\vec{v} \cdot \vec{X} \leq -c\|\vec{v}\|] = \mathbb{P}[\vec{v} \cdot \vec{X} \geq c\|\vec{v}\|]$. Hence, $\mathbb{P}[|\vec{v} \cdot \vec{X}| \geq c\|\vec{v}\|] = \mathbb{P}[\vec{v} \cdot \vec{X} \geq c\|\vec{v}\|] + \mathbb{P}[\vec{v} \cdot \vec{X} \leq -c\|\vec{v}\|] = 2\,\mathbb{P}[\vec{v} \cdot \vec{X} \geq c\|\vec{v}\|] \leq 2e^{-c^2/2}$.

(e) Let Θ denote the angle between $\vec{X}$ and $\vec{Y}$. Prove that $\mathbb{P}[|\cos\Theta| \geq \varepsilon] \leq 2e^{-\varepsilon^2 n/2}$, where $\varepsilon > 0$ is an arbitrary constant. [**Hint**: Use the fact that $\vec{X} \cdot \vec{Y} = \|\vec{X}\|\,\|\vec{Y}\|\cos(\Theta)$. **Remark**: This result shows that the probability of two independent random vectors in $\{-1, +1\}^n$ being orthogonal *6pts*

quickly approaches 1 as $n \to \infty$.]

Approach 1. Proof using Law of total probability: Note that $\|\vec{X}\| = \|\vec{Y}\| = \sqrt{n}$, which implies $\vec{X} \cdot \vec{Y} = n \cos\Theta$. Hence,

$$\begin{aligned}
\mathbb{P}[|\cos\Theta| \geq \varepsilon] = \mathbb{P}\Big[\frac{1}{n}|\vec{X} \cdot \vec{Y}| \geq \varepsilon\Big] &= \mathbb{P}[|\vec{X} \cdot \vec{Y}| \geq \varepsilon n] \\
&= \sum_{\vec{v} \in \{-1,+1\}^n} \mathbb{P}[|\vec{X} \cdot \vec{Y}| \geq \varepsilon n \mid \vec{Y} = \vec{v}]\, \mathbb{P}[\vec{Y} = \vec{v}] \\
&= \sum_{\vec{v} \in \{-1,+1\}^n} \mathbb{P}[|\vec{v} \cdot \vec{X}| \geq \varepsilon\sqrt{n}\|\vec{v}\|]\, \mathbb{P}[\vec{Y} = \vec{v}] \\
&\leq 2e^{-\varepsilon^2 n/2} \sum_{\vec{v} \in \{-1,+1\}^n} \mathbb{P}[\vec{Y} = \vec{v}] \\
&= 2e^{-\varepsilon^2 n/2},
\end{aligned}$$

where the third line follows from the fact that $\|\vec{v}\| = \sqrt{n}$ for any $\vec{v} \in \{-1,+1\}^n$, while the fourth line follows from the result from part (d).

Approach 2. Proof using Hoeffding's inequality: Note that $\|\vec{X}\| = \|\vec{Y}\| = \sqrt{n}$, which implies $\vec{X} \cdot \vec{Y} = n \cos\Theta$. Using $Z_i := X_i Y_i$ and $S_n := \vec{X} \cdot \vec{Y} = \sum_{i=1}^n X_i Y_i$, and note that Z_i are bounded random variables; more precisely, $\mathbb{P}[-1 \leq Z_i \leq 1] = 1$. Since $\mathbb{E}[Z_i] = 0$, we have $\mathbb{E}[S_n] = 0$. Hence, Hoeffding's inequality implies that for any $\varepsilon > 0$,

$$\mathbb{P}[|S_n| \geq \varepsilon n] \leq 2e^{-2\varepsilon^2 n^2 / \sum_{i=1}^n 2^2}. \tag{2}$$

Noting $\sum_{i=1}^n 2^2 = 4n$ yields the desired result.

[End of Exam!]

Stat 201A: Lab 1

Conceptual review

- What is a probability space? Give an example.
- What is a random variable X? How to characterize it?
- How are Bernoulli distributions related to binomial, geometric, negative binomial ditributions?

Problem 1

(a) Three events A, B and C satisfy the following: A and B are independent, C is a subset of B, C is disjoint from A, $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/4$ and $\mathbb{P}(C) = 1/10$. Compute $\mathbb{P}(A \cup B \cup C)$.

(b) Suppose that a rapid COVID test is 99% accurate when someone doesn't have COVID but only 90% accurate if someone has COVID. Suppose that 1% of the population has covid. If someone tests positive, what is the probability they have covid?

Problem 2

Consider the experiment of drawing a point uniformly at random from the unit interval $[0, 1]$. Let Y be the first digit after the decimal point of the chosen number.

(a) Explain why Y is discrete and find its probability mass function.

(b) Find the expectation of Y. Find the variance of Y.

Problem 3

We have a system that has two independent components. Both components must function in order for the system to function. The first component has 8 independent elements that each work with probability 0.95. If at least 6 of the elements are working then the first component will function. The second component has 4 independent elements that each work with probability 0.90. If at least 3 of the elements are working then the second component will function.

a. What is the probability that the system functions?

b. Suppose the system is not functioning. Given that information, what is the probability that the second component is not functioning?

Problem 4 (Illustration of different distributions)

See lab1_code.Rmd

Problem 5 (Inverse Transform Sampling)

(a) Prove that for any random variable $X \in \mathbb{R}$, the random variable $F_X^{-1}(U)$ has the same distribution as X, where F_X^{-1} is the inverse of the cumulative distribution function F_X of X, and U is uniform on $[0, 1]$. For simplicity, prove this for continuous random variable X.

(b) Consider an exponential distribution with rate parameter $\lambda = 0.5$, where $X \sim \text{Exp}(0.5)$. Using the inverse CDF method, simulate the exponential random variable $X = F_X^{-1}(U)$, where U is uniformly distributed on $[0, 1]$.

Hint: The CDF for exponential distribution $\text{Exp}(\lambda)$ is

$$F_X(x) = 1 - e^{-\lambda x} \quad x \geq 0$$

(c) (Bonus) Simulate any distribution you like using inverse transform sampling.

Stat 201A, Fall 2024: Lab 2

Conceptual review

- When is a sequence of random varibales exchangeable?
- How are the Markov inequality, Chebyshev inequality and the Weak LLN related?

Problem 1

(a) Let X_1, X_2, X_3 be independent $\text{Exp}(\lambda)$ distributed random variables. Find the probability that $\mathbb{P}(X_1 < X_2 < X_3)$.

(b) We deal five cards, one by one, from a standard deck of 52. (Dealing cards from a deck means sampling without replacement.)

(i) Find the probability that the second card is an ace and the fourth card is a king.

(ii) Find the probability that the first and the fifth cards are both spades.

(iii) Find the conditional probability that the second card is a king given that the last two cards are both aces.

Problem 2

Chebyshev's inequality does not always give a better estimate than Markov's inequality. Let X be a random variable with $\mathbb{E}[X] = 2$ and $\mathrm{Var}(X) = 9$. Find the values of t where Markov's inequality gives a better bound for $\mathbb{P}(X > t)$ than Chebyshev's inequality.

Problem 3

A cereal company is performing a promotion, and they have put a toy in each box of cereal they make. There are n different toys altogether and each toy is equally likely to show up in any given box, independently of the other boxes. Let T_n be the number of boxes we need to buy in order to collect the complete set of n toys.

(a) The random variable W_k is the number of boxes we need to open to see a new toy after we have collected k distinct toys. What is the distribution of W_k? Prove that $T_n = 1 + W_1 + W_2 + \cdots + W_{n-1}$.

(b) Calculate the limits $\lim_{n\to\infty} \frac{\mathbb{E}[T_n]}{n\ln(n)}$ and $\lim_{n\to\infty} \frac{\mathrm{Var}[T_n]}{n^2}$.

(c) Use Chebyshev's inequality to estimate $\mathbb{P}(|T_n - \mathbb{E}[T_n]| > \varepsilon n)$.

(d) Show that for any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\frac{T_n}{n\ln(n)} - 1\Big| > \varepsilon\Big) = 0.$$

This is a weak law of large numbers for the coupon collector's problem.

(e) Using the union bound for the event $E_i^{cn\log(n)}$ that the i-th coupon was not picked in the first $cn\log(n)$ trials, prove that $\mathbb{P}(T_n \geq cn\log(n)) \leq n^{1-c}$.

Problem 4

Cantelli's inequality provides a sharper one-sided bound compared to Chebyshev's inequality. Let X be a random variable with mean μ and variance σ^2, and let $b > 0$.

- Chebyshev's inequality (one-sided):

$$P(X \geq \mu + b) \leq \frac{\sigma^2}{b^2}$$

- Cantelli's inequality:

$$P(X \geq \mu + b) \leq \frac{\sigma^2}{\sigma^2 + b^2}$$

(a) Prove Cantelli's inequality using Markov's inequality.

Hint: Let $Y = X - \mu$. For any $u > 0$,

$$P(Y \geq b) = P(Y + u \geq b + u) \leq P\left((Y + u)^2 \geq (b + u)^2\right).$$

Then use Markov's inequality and find u that minimizes the resulting bound.

(b) Cantelli's inequality implies

$$P(|X - \mu| \geq b) \leq \frac{2\sigma^2}{\sigma^2 + b^2}$$

Comment of the value of this inequality compared to Chebychev's.

Problem 5

The standard Cauchy distribution has the probability density function:

$$f(x) = \frac{1}{\pi\,(1+x^2)}$$

(a) Does the Weak Law of Large Numbers hold for the Cauchy distribution? Explain why or why not.

(b) Simulate N samples from the standard Cauchy distribution for $N = 10^2, 10^3, 10^4, 10^5$. Calculate the sample averages as N increases. How do the results relate to your explanation in (a)?

Stat 201A, Fall 2024: Lab 3

Conceptual review

- How are the LLN and CLT different? How are they related?
- How are the exponential and gamma distributions related?

Problem 1

Noodle decide to improve her ability to calculate integrals. Each day she flips a coin until she gets tails. If she gets tails in 3 or less flips, she will calculate 10 integrals. If she needs strictly more than 3 flips to get tails she will calculate 60 integrals. After a full year passes, estimate the probability that Noodle has solved more than 6000 integrals.

Problem 2

Here is a limit theorem that one can prove without complicated tools. Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables with distribution $Exp(1)$, and let $M_n = \max(X_1, \ldots, X_n)$. Show that for any $x \in \mathbb{R}$ we have

$$\lim_{n\to\infty} \mathbb{P}(M_n - \ln(n) \leq x) = \exp(-e^{-x}).$$

This is called the *Gumbel distribution*.

Problem 3

Prove that the Exponential Distribution is the only distribution on $(0, \infty)$ that satisfies:

$$\mathbb{P}(X > a + b | X > b) = \mathbb{P}(X > a), \, \forall \, a, b > 0.$$

Problem 4

In the lectures, we used the MGF to prove the following results regarding convergence in distribution.

- Let $G_n \sim \text{Geometric}\left(\frac{\lambda}{n}\right)$, where $\lambda > 0$, and $n = 1, 2, 3, \ldots$. Define $X_n = \frac{G_n}{n}$. As $n \to \infty$, X_n converges in distribution to an Exponential distribution with rate λ:

$$X_n \xrightarrow{d} \text{Exp}(\lambda)$$

- Let $F_{r,n} \sim \text{NB}\left(r, \frac{\lambda}{n}\right)$, where $\lambda > 0$ and $n = 1, 2, 3, \ldots$. Define $X_n = \frac{F_{r,n}}{n}$. As $n \to \infty$, X_n converges in distribution to a Gamma distribution with shape r and rate λ:

$$X_n \xrightarrow{d} \text{Gamma}(r, \lambda)$$

1. Apply MGF and use similar strategies discussed in the lecture to prove the following.
   Let $W_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right)$ represent a binomial random variable with probability $\frac{\lambda}{n}$, then W_n converges in distribution to a Poisson distribution with parameter λ:

$$W_n \xrightarrow{d} \text{Poi}(\lambda)$$

   .

2. Conduct simulations to show the convergence in distribution for three results above.

Problem 5

The MGF of a random variable X is given by

$$M_X(t) = \frac{c}{1-t} - \frac{2}{1-t} \quad \text{for } |t| < 1$$

1. Find the value of c.
2. Find $\mathbb{E}[X]$.
3. Find $\mathbb{E}[X^2]$.

Stat 201A, Fall 2024: Lab 4

Conceptual review

- If X and Y are independent continuous random variables with probability density functions f and g. What is the probability density function of $X + Y$?
- What is the statement of CLT for Binomail distribution?
- What does the Kullback–Leibler divergence describes?

Problem 1

1. Roll a fair die 720 times. Estimate the probability that we have exactly 113 sixes.
2. You flip a fair coin 10,000 times. Approximate the probability that the difference between the number of heads and number of tails is at most 100.

Problem 2

Suppose we have a biased coin and we do not know the true probability p that it lands on heads. How can we estimate p? Can we estimate the error of our approximation?

Problem 3

Mitchell and Alex are competing together in a 2-mile relay race. The time Mitchell takes to finish (in hours) is $X \sim \text{Unif}(0, 2)$ and the time Alex takes to finish his mile (in hours) is continuous $Y \sim \text{Unif}(0, 1)$. Alex starts immediately after Mitchell finishes his mile, and their performances are independent. What is the distribution of $Z = X + Y$, the total time they take to finish the race?

Problem 4

Suppose $p_1, \ldots, p_k$ be a set of nonnegative numbers that sum to one. Suppose $f_1, \ldots, f_k$ are another set of nonnegative numbers that sum to one. The Kullback-Leibler divergence between these two sets of numbers is given by

$$KL(f\|p) := \sum_{i=1}^{k} f_i \log \frac{f_i}{p_i}$$

1. Show that $KL(f\|p)$ is always nonnegative.
2. Show that $KL(f\|p) = 0$ if and only if $f_i = p_i$ for each $i = 1, \ldots, k$.
3. Suppose that a coin toss can give three different results: H (heads), T (tails) and edge (when the coin just stands on its edge). Suppose that a person A assigns probabilities

   $$p_1^A = 0.499, \quad p_2^A = 0.499, \quad p_3^A = 0.002$$

   to the three outcomes and another person B assigns probabilities

   $$p_1^B = p_2^B = p_3^B = \frac{1}{3}$$

   to the three outcomes. Suppose that an experiment is performed by tossing the coin a bunch of times and this led to the observed proportions

   $$f_1 = \frac{14}{29}, \quad f_2 = \frac{14}{29}, \quad f_3 = \frac{1}{29}$$

   of the three outcomes. Calculate the Kullback-Leibler divergences $KL\left(f\|p^A\right)$ and $KL\left(f\|p^B\right)$. Which of $KL\left(f\|p^A\right)$ and $KL\left(f\|p^B\right)$ is smaller and does that seem reasonable?

Problem 5

Compare real binomial probabilities with entropy and normal approximations for $n = 100$ and $p = 0.5$ and $p = 0.05$, using plots to visualize. You may play around with different n, p and k to compare.

Stat 201A, Fall 2024: Lab 5

Conceptual review

- When to use Poisson approximation instead of normal approximation for a Binomial distribution?
- How are the binomial and multinomials distributions related?

Problem 1 (Poisson approximation)

1. A large company has a large fleet of cars. On average, there are 3 accidents each week. What is the probability that at most 2 accidents happens next week?
2. Every evening Murdoc goes to the local casino. There is a 1% chances that he wins \$10000 and 99% he loses \$100. Define a random variable X_k representing the winning/losing outcome of Murdoc after each day k. After a full year passes, estimate the probability that Murdoc wins at least least \$1000.

Problem 2 (Exchangeability and multinomial distribution)

Suppose an urn contains 2 green, 3 red and 4 yellow balls. Six balls are chosen **with** replacement. Find the probability that green appeared 1 times, red 2 times, and yellow 3 times. Six balls are chosen **without** replacement. Find the probability that the 3rd ball chosen is green, given that the 5th ball chosen is yellow?

Problem 3 (Poisson distribution)

1. Let $X \sim \text{Geom}(1/3)$ and $Y \sim \text{Poisson}(2)$ be independent random variables. Calculate $\mathbb{P}(X = Y + 2)$.
2. Suppose that $X \sim \text{Poisson}(\lambda)$. Find the probability $\mathbb{P}(X \text{ is even})$.
3. Let $X \sim \text{Poisson}(\mu)$. Compute $\mathbb{E}\left(\frac{1}{1+X}\right)$.

Problem 4

Let $N \sim \text{Poisson}(\lambda)$, and let $X_1, X_2, \ldots$ be a sequence of i.i.d. geometric random variables with parameter p, where $X_i \sim \text{Geometric}(p)$. Define $S_N = X_1 + X_2 + \cdots + X_N$. N is independent of the X_i 's.

1. Find the probability generating function (PGF) of S_N
   Hint: Use the compounding theorem discussed in Lecture 11. The PGF of a Poisson random variable $N \sim \text{Poisson}(\lambda)$ is given by

   $$G_N(t) = e^{\lambda(t-1)}$$

   and the PGF of a geometric random variable $X \sim \text{Geometric}(p)$ is

   $$G_X(t) = \frac{p}{1-(1-p)t}, \quad |t| < \frac{1}{1-p}$$

2. Suppose $p = 0.5$, $\lambda = 1$. Calculate the probability $\mathbb{P}(S_N = 1)$.

3. (Bonus) Verify the PGF of a Poisson random variable and a geometric random variable through explicit calculation.

Problem 5

Recall: When $N \sim \text{Poisson}(\lambda)$ and $(X_1, \ldots, X_m) \mid N \sim \text{Multinomial}(N, p_1, \ldots, p_m)$, the joint distribution of $X_1, \ldots, X_m$ follows independent Poisson distributions, i.e., $X_j \sim \text{Poisson}(p_j\lambda)$. We can show this result through simulation.

1. Simulate 10,000 samples of the Poissonized Multinomial and Independent Poisson distributions with $\lambda = 10$ and $\mathbf{p} = (0.3, 0.5, 0.2)$.

2. Visualize and compare the joint distribution of X_1 and X_2 from two data simulation procedure using either a 2D density plot or an overlayed scatter plot.

Stat 201A, Fall 2024: Lab 7

Conceptual review

- Given X and Y continuous random variables with joint density $f_{X,Y}$. How to compute $\mathbb{E}[g(X,Y)]$?
- Given a univariate real function f, how to find the tangent equation to a point in the curve $(x, f(x))$? What is the equivalent if f is bivariate?
- Explain the relations between geometric, exponential, gamma, Poisson and beta distributions.

Problem 1

Suppose the joint density of X and Y is given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi} \exp\left(\frac{-(x^2+y^2)}{2}\right)$$

Find the joint density of $U = X + Y$ and $V = X - Y$.

Problem 2

Let X and Y be independent $Geom(p)$ random variables. Let $V = min(X, Y)$ and

$$W = \begin{cases} 0, & if \quad X < Y \\ 1, & if \quad X = Y \\ 2, & if \quad X > Y \end{cases}$$

Find the join probability mass function of V and W and show that V and W are independent.

Problem 3

Let the random variables X, Y have joint density function

$$f(x,y) = \begin{cases} 3(2-x)y \text{ if } 0 < y < 1 \text{ and } y < x < 2-y, \\ 0 \text{ otherwise.} \end{cases}$$

1. Find the marginal density functions f_X and f_Y .
2. Calculate the probability that $X + Y \leq 1$.
3. Find the joint density of $(W, Z) = (XY, (1-Y)X)$.

Problem 4 (Joint density change Under a Non-invertible transformation)

In class, we looked at the Jacobian formula for calculating the joint density of a transformed set of continuous random variables in terms of the joint density of the original random variables. This formula assumed that the transformation is invertible. However, the general method based on change of variant principles works fine. This is illustrated in the following example.

Suppose X and Y have joint density $f_{X,Y}$. What is the joint density of $U = \min(X, Y)$ and $V = \max(X, Y)$?

Stat 201A, Fall 2024: Lab 8

Conceptual review

- Let $X_1, X_2, \ldots, X_n$ iid continuous random variables. What is the density of $X_{(k)}$?
- How do we get $f_{Y|X}(y|x)$ from $f_X(x)$ and $f_{X,Y}(x,y)$?

Problem 1

Let $U_{(1)}, \ldots, U_{(n)}$ be the values of n iid. $\mathrm{U}(0,1)$ variables arranged in increasing order. For $0 < x < y < 1$, find a simple formula for:

a. $P(U_{(1)} > x, U_{(n)} < y)$

b. $P(U_{(1)} > x, U_{(n)} > y)$

c. $P(U_{(1)} < x, U_{(n)} < y)$

d. $P(U_{(1)} < x, U_{(n)} > y)$

Problem 2

Let X and Y independent $\text{Exp}(\lambda)$ random variables. Describe the distribution of $X|_{X+Y}$.

Problem 3

Let $X \sim \text{Exp}(\lambda)$, and let $Y \sim \text{Poisson}\,(X)$ (that is, given $X = x$, Y follows the Poisson (x) distribution).

a. Find $P(X \in dx, Y = y)$.

b. Use (a) to find the unconditional distribution of Y.

c. Given $Y = y$, what is the conditional density of X?.

Problem 4 (Beta, Gamma distribution)

1. Let $B \sim \text{Beta}(a, b)$. Find the distribution of $1 - B$.
2. Let $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$, with X and Y independent. Is the ratio X/Y independent of the sum $X + Y$?
3. The F-test is a very widely-used statistical test based on the $F(m, n)$ distribution, which is the distribution of $\frac{X/m}{Y/n}$ with $X \sim \text{Gamma}\left(\frac{m}{2}, \frac{1}{2}\right), Y \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$. Find the distribution of $mV/(n + mV)$ for $V \sim F(m, n)$.
4. Let $U_1, \ldots, U_n$ be i.i.d. $\text{Unif}(0, 1)$. Find the mean and variance of the j th order statistic $U_{(j)}$.

Problem 5

Fred is waiting for a bus, but the waiting time X depends on some environmental factor Y, which affects the bus schedule. The environmental factor Y represents the bus delay rate and is not fixed but follows a Gamma distribution. Specifically,

- Given the environmental factor $Y = y$, the waiting time X follows an exponential distribution with rate y, meaning $X \mid Y = y \sim \text{Exp}(y)$.
- The environmental factor Y itself is random and follows a Gamma distribution, $Y \sim \text{Gamma}(\alpha, \beta)$, where α is the shape parameter and β is the rate parameter.

Find the overall distribution of the waiting time X.
(Hint: $\int_0^\infty y^\alpha e^{-cy} dy = \frac{\Gamma(\alpha+1)}{c^{\alpha+1}}$).

Stat 201A, Fall 2024: Lab 9

Conceptual review

- Can you explain what $\mathbb{E}[X|Y]$ represents? Why do we have $\mathbb{E}\big[\mathbb{E}[X|Y]\big] = \mathbb{E}[X]$?
- What is the purpose of the loss function in risk minimization?
- What is Wald's identity and the law of total variance?

Problem 1

Let $X_1, X_2, \ldots$ be i.i.d. exponential random variables with parameter λ. Let N be a Geom(p) random variable (with $0 < p < 1$) independent of the X_i random variables. Define the random sum $S_N = X_1 + X_2 + \cdots + X_N$.

1. Find the mean $\mathbb{E}[S_N]$
2. Find the probability distribution of S_N.

Problem 2

1. Suppose that X is a discrete random variable. Find an estimator d that minimizes the risk for the loss function $1_{\{X \neq d\}}$.
2. Suppose that X is a continuous random variable. Prove that the mean absolute error minimizer is given by the median.
3. What loss function should we use so that the estimator d that minimizes the risk is given by the γ quantile?

Problem 3

Jack and Jill are playing a game. Each will start with $ 5 and $ 10 respectively and play a game by making a series of $ 1 bets until one of them loses all their money. For each bet they flip 1 fair coin. If it's tail, Jack wins. If it's head, Jill wins.

1. Find the probability that Jack wins the game.
2. Find the expected length of the game.

Problem 4 (Law of total variance)

We have a sample of 100 normally distributed payments, with mean=1000 dollars and standard deviation= 100 dollars. 10% of these payments were made in error and should be refunded their full payment amount. The other 90% will have a refund amount of 0 dollars. What is the variance of the refunded amount?

Problem 5

Let Y be the number of heads in n spins of a coin, whose probability of heads is θ. Suppose our prior distribution for θ is Uniform$\sim [0, 1]$,

1. Derive the posterior distribution of $\theta \mid Y = y$.
2. Show that the posterior mean of the posterior $\theta \mid Y = y$ always lies between the mean of the prior for θ and the observed relative frequency of heads y/n.
3. Show that the posterior variance of $\theta \mid X$ is always less than the prior variance.

Stat 201A, Fall 2024: Lab 11

Conceptual review

- Review of Gaussian Process.
- Review of Branching Process.

Problem 1

1. Given n independently sampled values from an unknown distribution. We believe this unknown distribution to have CDF F, how do we use the Kolmogorov–Smirnov test?
2. How many points do we need to sample to get a 99% confidence level of the empirical distribution to be at distant at most 0.01 to the real distribution?
3. After sampling a big number of values from an unknown distribution, we guess it to be Exponential(0.1). The real distribution turns out to be Exponential(0.05). Suppose that we sampled 100 values, what is our confidence level for our guessed distribution? What if we had sampled 10 000 values?

Problem 2

1. Suppose that $\{B_t\}_t$ for $t \in [0,1]$ is a Brownian motion. What is the distribution of the process $X_t = B_t - tB_1$? what about $Y_t = e^{-t}B_{e^{2t}}$?
2. Suppose that $\{B_t\}_t$ for $t \in [0,1]$ is a Brownian bridge. Let Z a standard normal independent of $\{B_t\}_t$. Show that $X_t = B_t + tZ$ is a Brownian motion.

Problem 3

The growth dynamics of pollen cells can be modeled by binary splitting as follows: After one unit of time, a cell either splits into two or dies. The new cells develop according to the same law independently of each other. The probabilities of dying and splitting are 0.46 and 0.54 , respectively.

1. Determine the maximal initial size of the population in order for the probability of extinction to be at least 0.3.
2. What is the probability that the population is extinct after two generations if the initial population is the maximal number obtained in (a)?

Problem 4

The following model can be used to describe the number of women (mothers and daughters) in a given area. The number of mothers is a random variable $X \sim \text{Poisson}(\lambda)$. Independently of the others, every mother gives birth to a $\text{Poisson}(\mu)$-distributed number of daughters. Let Y be the total number of daughters and hence $Z = X + Y$ be the total number of women in the area.

1. Find the generating function of Z.
2. Compute $\text{E}(Z)$ and $\text{Var}(Z)$.

Stat 201A, Fall 2024: Lab 12

Conceptual review

- When is a Markov chain reducible, when is it irreducible? Give examples.
- Explain the difference between recurrent and transient Markov chains.
- How to find a hitting probability, a hitting time, a stationary distribution?

Problem 1

Consider a two state Markov chain with all transition probabilities equal to 1/2.

1. Is this Markov chain irreducible?
2. Is it recurrent? transient?
3. Find the stationary distribution.
4. Now suppose that for one state both transitions probabilities are 1/2 but for the other point the probability to stay is 1. Find all hitting probabilities and the hitting times.

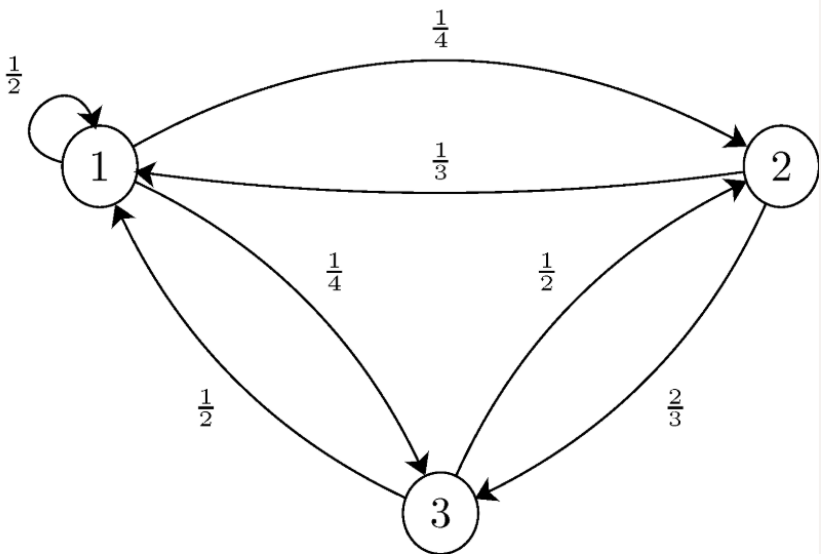
Problem 2

Three cards labeled 1, 2, 3 are laid in a row in that order, forming the three-digit number 123 when read from left to right. A swap consists of picking two distinct cards, and then swapping them. After two swaps, the cards form a new three-digit number n when read from left to right.

1. Find the probability p that the digit in any given place will be the same as it was at the start
2. Compute the expected value of n.
3. How to generalize this to x cards and y swaps?

Problem 3

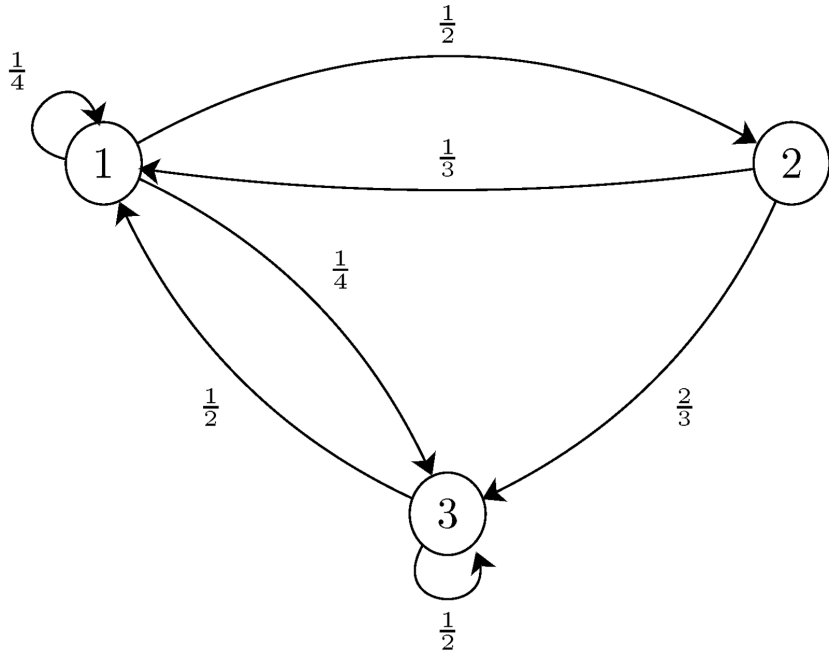
Consider the Markov chain shown below



1. Is this chain irreducible?
2. Find the stationary distribution for this chain.

Problem 4

Consider the Markov chain shown below.



Assume $X_0 = 1$, and let R be the first time that the chain returns to state 1, i.e.,

$$R = \min\{n \geq 1 : X_n = 1\}$$

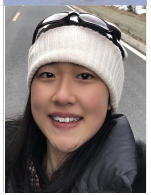
Find $E[R \mid X_0 = 1]$

STAT201A: Introduction to Probability at an Advanced Level

Fall 2024, UC Berkeley

Lecture 1
August 29, 2024

Teaching Staff

Course admin e-mail:		stat201a-fa24@lists.berkeley.edu Your e-mail must be sent from a Berkeley e-mail address; otherwise, it will get rejected automatically.
Instructor :		Prof. Yun S. Song (yss@berkeley.edu) Office Hours: TuTh 5-6pm, 304B Stanley
GSI (20hr/week):		Gabriel Ramirez Raposo (raposo@berkeley.edu) Office Hours: M 4-5pm, Th 11:30am-12:30pm, 444 Evans
GSI (10hr/week):		Fanding Zhou (zhoufd@berkeley.edu) Office Hours: W 5-6pm, F 10-11am, 444 Evans

Please use this e-mail or Ed Discussion for most course-related correspondence.

Use this e-mail **only if** you need to communicate with me about a private matter.

Time and Place

Lecture:

- **Time:** 3:30-5:00pm
- **Place:** Stanley 106

Discussion Sections:

- M: 12-2pm (Evans 330) by Gabriel
- M: 2-4pm (Evans 340) by Gabriel
- M: 4-6pm (Evans 330) by Fanding
- You may attend any section, **provided that there is space**. Students registered for the section will have priority.

Registration

- Available seats: 90
- Current enrollment size: 83
- More than 25 Concurrent Enrollment students have applied, but unfortunately we can accommodate only 7 of you. Priority has been given to graduate students.

Requirements

- We will use bCouses to share course material and post announcements.
  - https://bcourses.berkeley.edu/courses/1537153

Send us e-mail if you need to be added to this site.

- **Prerequisites**:
  - Undergraduate probability (at the level of Berkeley's Statistics 134)
  - Multivariable calculus (at the level of Berkeley's Mathematics 53)
  - Linear algebra (at the level of Berkeley's Mathematics 54)
- **Textbook**:
  - There is no required textbook for the class. You may use the following books as general references:
    - *An Intermediate Course in Probability*, 2nd edition by Allan Gut. (Available in bCourses)
    - *Stochastic Processes: Theory for Applications* by Robert G. Gallager.

Grading and Exams

- **Grades** will be determined as follows:
  - Homework: 50% (there will be six homework assignments)
  - Midterm: 20%
  - Final: 30%
  - **NOTE:** We will **drop** the lowest Homework score.
  - No additional allowances will be made for late or missed homework: **please do not contact us about missed homework or late submissions.**
- If you are on the waitlist (Concurrent Enrollment), you should submit Homework.
- **Exams**:
  - Midterm is on Thursday, October 17, in class. **No makeup exam** will be offered.
  - The Final is on Friday, December 20, 7–10 pm. We are **unable to accommodate exam conflicts**.

Course Policies

- It is **required** that you read the Course Policies detailed in **bCourses**.
- Please use **Ed Discussion** for all **technical** questions.
  - Please read the **Ed Discussion Etiquette** section.
  - Posting can be made anonymously to students, but will not be anonymous to instructors.
  - **Think first before posting a question!** A few students tend to abuse Ed Discussion by asking an excessive number of questions.
- For personal administrative questions, please either use a **private post on Ed Discussion** (visible to course staff only) or send email to the course administrative account: stat201a-fa24@lists.berkeley.edu
- **Gradescope**:
  - All homework will be submitted through Gradescope.
- Please DO NOT post any material (lecture notes, discussion section material, exams, homework, solutions, etc) on the internet.

Collaboration

- You are welcome to work on homework problems in study groups of two to four people.
- However, you must **always write up the solutions on your own**.
- Similarly, you may use books or online resources to help solve homework problems, but you must **always credit all such sources in your writeup and you must never copy material verbatim**.
- We believe that most students can distinguish between helping other students and cheating.

Advice

- **Take longhand notes:**
  - You might think that it's old school, but taking longhand notes can facilitate your learning, as supported by this study:
    - The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking
      https://doi.org/10.1177/0956797614524581
  - Please read this NPR article and listen to the accompanying 3-minute interview, if you prefer a quick summary:
    - https://www.npr.org/2016/04/17/474525392/attention-students-put-your-laptops-away
  - We are confident that the pen is also mightier than screenshots! Taking longhand notes will help you summarize and process the lectures better.

Academic Dishonesty

- Three types of students I have encountered:
  1. Students who value integrity more than their grades, and don't cheat.
  2. Students who acknowledge that cheating is wrong, but might give in to temptation.
  3. Students who think that cheating is okay and that it would be to their disadvantage if they did not cheat.

- We have a **zero-tolerance policy** for cheating. Consequences of cheating include: **negative points for the corresponding assignment, a failing grade in the class, and/or a referral to the Office of Student Conduct.**
- Your attention is drawn to **Berkeley Honor Code**:
  - "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."
- In particular, you should be aware that copying or sharing solutions, in whole or in part, from other students in the class (or any other source without acknowledgment) constitutes cheating.

Introduction

Probability

- **Why should you learn probability?**
- **Probability is ubiquitous.**
  - Mathematics
  - Statistics (estimation and inference from data, prediction)
  - Physics (statistical physics, quantum physics)
  - Chemistry
  - Climate Science
  - Economics and Finance
  - Biology (cellular dynamics, signaling, development, evolution)
  - Medicine (clinical trials, drug discovery)
  - Computer Science and Engineering

Probability

- **Q**: Why should computer scientist care about probability?
- **A**: Over the past decade, we have seen a tremendous increase in the use of probability theory in computing. Examples include:
  - Machine Learning and Artificial Intelligence
  - Massive data analysis and data mining
  - Randomized numerical linear algebra
  - Graph theory
  - Cryptography
  - Program verification
  - Packet routing in networks
  - Design of ethernet cards
- These successful applications rely on algorithms that involve clever probabilistic and statistical ideas.

Randomized Algorithms

Design (Randomized)

- Make random choices during execution.
  - e.g., Monte Carlo
- **Pros**:
  - Can be **significantly more efficient** than the best deterministic solution
  - Often **simpler and easier to implement**
- **Cons**:
  - The answer may be incorrect with some probability (acceptable if it's small)
  - Efficiency is guaranteed only with some probability.

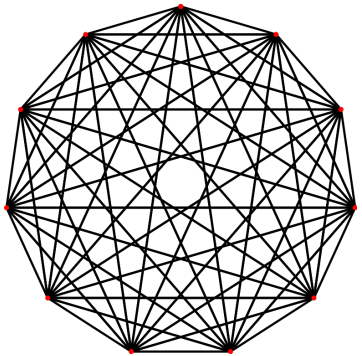
Analysis (Probabilistic)

- View the input as being randomly selected.
- "Hard" instances may appear with relatively small probability.
- So, often "hard" problems are easy to solve in practice.
- **Computational complexity** concerns the worst-case scenario. Some NP-hard problems might admit algorithms that are extremely efficient on **almost** all inputs.

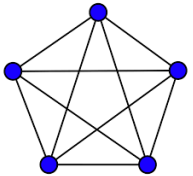
The Probabilistic Method

- **Key Idea:** Prove **existence** by showing that

$\mathbb{P}$[a randomly chosen object has the required property] > 0

K_{11} 

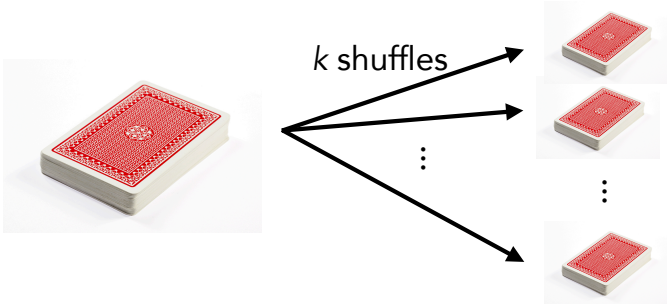
Is it possible to color the edges of K_{11} using two colors such that there exists no monochromatic K_5?

K_5 

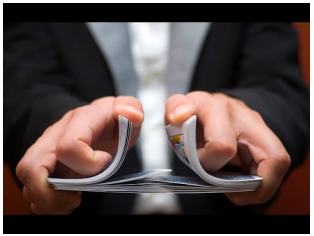
Theorem: If $\binom{n}{k} < 2^{\binom{k}{2}-1}$, then there exists a 2-coloring of K_n (a complete graph with n vertices) edges such that it contains no monochromatic K_k subgraph.

Card Shuffling

- Suppose you have a deck of n distinct cards. How many times do you need to shuffle the deck for the cards to be "well mixed"?



- More precisely, you want the order of cards, $(X_1, \ldots, X_n)$, to be close to being uniformly distributed over the space of $n!$ permutations of the cards.
- **Random-to-top:** Pick any card and move it to the top of the deck. $O(n \log n)$
- **Random transposition:** Pick two cards uniformly at random and swap them. $O(n \log n)$
- **Riffle shuffle:**

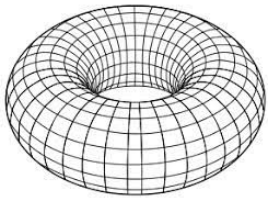


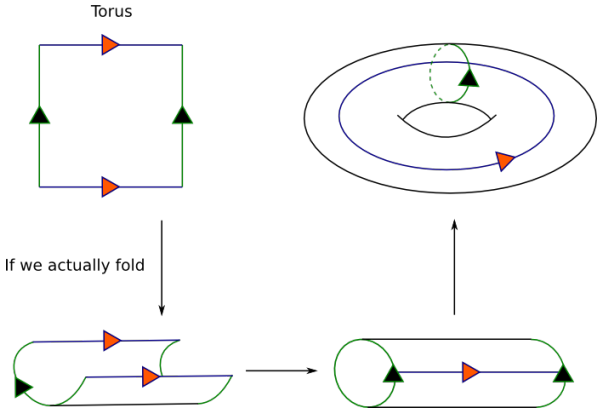
$O(\log n)$

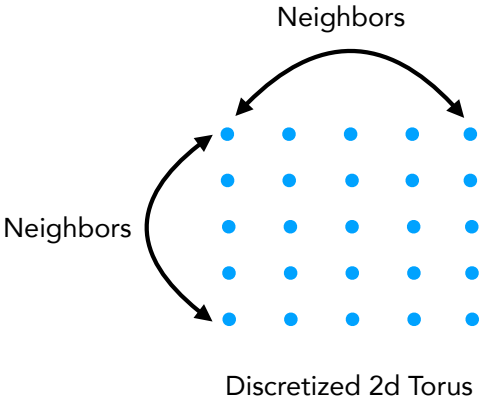
For $n = 52$, about 7 riffle shuffles are "sufficient."

199

Ising Model on a 2-dimensional Torus

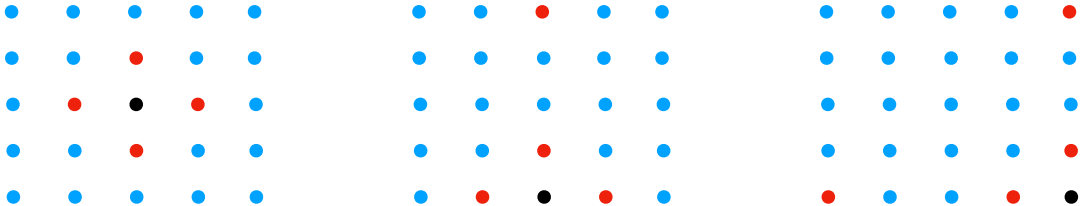






http://pi.math.cornell.edu/~mec/Winter2009/Victor/part1.htm

red = a neighbor of black



200

Ising Model on a 2-dimensional Torus

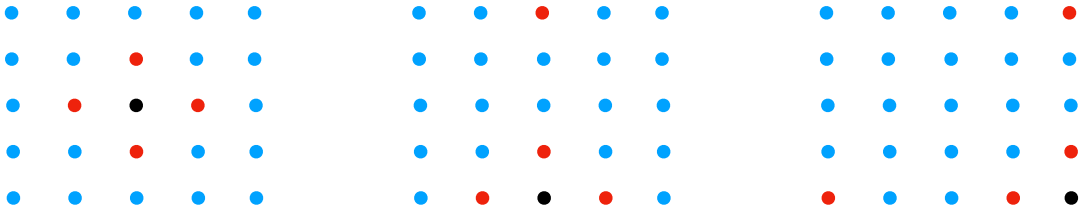
+	+	−	+	−
−	−	+	−	+
+	−	−	+	−
+	+	−	+	−
−	−	+	−	−

$$S_i \in \{-1, +1\}, \forall i$$

$$\mathbb{P}(\mathbf{S}) = \frac{1}{Z(T)} \exp\left(\frac{1}{T} \sum_{i,j \text{ neighbors}} S_i S_j\right)$$

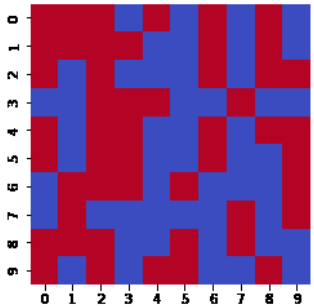
- Computing the normalization constant $Z(T)$, called the partition function, is hard.
- How can we sample from $\mathbb{P}(\mathbf{S})$?

red = a neighbor of black



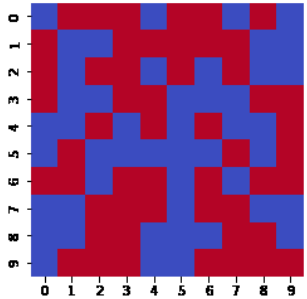
Ising Model on a 2-dimensional Torus

Markov chain Monte Carlo



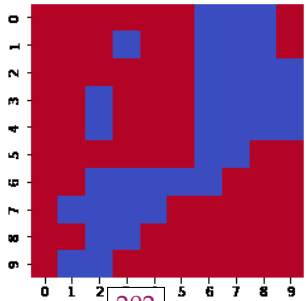
Samples from the target distribution

T = 10

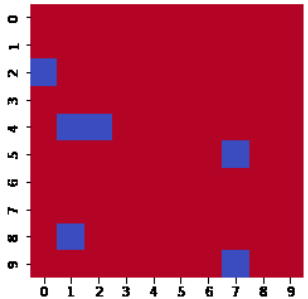


$$\mathbb{P}(\mathbf{S}) = \frac{1}{Z(T)} \exp\left(\frac{1}{T} \sum_{i,j \text{ neighbors}} S_i S_j\right)$$

T = 2.5



T = 2



202

Topics covered in this course

- Basic probability theory (review of undergraduate probability + some new material)
- Tail bounds: Markov, Chebyshev, Chernoff, Hoeffding
- Convergence of random variables
- Law of Large Numbers (weak and strong)
- Generating functions
- Proof of the central limit theorem
- Transformation of several random variables
- Multivariate Normal
- Gaussian processes
- Branching processes
- Poisson processes
- Markov chains

Homework #1

- Homework #1 is available on bCourses.
- Due in ~2 weeks: Friday, September 13, 2024, 10pm via Gradescope.

Consider two coin tosses.
Possible outcomes $\{HH, HT, TH, TT\} =: \Omega$

Some events:
$E_1 :=$ first toss is H $= \{HH, HT\} \subset \Omega$
$E_1^c :=$ " " " T $= \{TT, TH\} \subset \Omega$
$E_2 :=$ second toss is H $= \{HH, TH\} \subset \Omega$
$E_2^c :=$ " " " T $= \{TT, HT\} \subset \Omega$
$E_1 \cap E_1^c = \phi$, $E_1 \cup E_1^c = \Omega$
$E_1 \cup E_2 = \{HH, HT, TH\}$
$E_1 \cap E_2 = \{HH\}$
$(E_1 \cup E_2)^c = \{TT\} = E_1^c \cap E_2^c$ (De Morgan's law)

Probability of events
$\mathbb{P}[E_1] = p$, $\mathbb{P}[E_2] = q$ ($p=q \Leftrightarrow$ Identically distributed)
$\mathbb{P}[E_1^c] = 1-p$, $\mathbb{P}[E_2^c] = 1-q$
$\mathbb{P}[E_1 \cap E_2] = r$ ($r=pq \Leftrightarrow$ independent)

In general, $0 \le r \le \min\{p, q\}$
since $E_1 \cap E_2 \subset E_1, E_2$

Probability space $(\Omega, \mathcal{F}, \mathbb{P})$
Ω = sample space, the set of all outcomes.
$\mathcal{F}$ = a σ-algebra (aka σ-field) on Ω
A set of subsets of Ω satisfying certain properties
$\mathbb{P}$ = probability measure

Def: (σ-algebra of Measurable sets)
Given a set S, $\mathcal{F} \subseteq \mathcal{P}(S)$ (Power set) is called a σ-alg (aka σ-field) on S if
a) $\phi, S \in \mathcal{F}$
b) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
c) $A_i \in \mathcal{F}$, $i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Closed under countable union

$A \in \mathcal{F}$ is called $\mathcal{F}$-measurable

e.g. 1) $\mathcal{F} = \{\phi, S\}$ the smallest σ-alg
2) $\mathcal{F} = \mathcal{P}(S)$ the largest σ-alg (Power set, the set of all subsets of S)

Def Measurable space $(S, \mathcal{F})$ — S: a set, $\mathcal{F}$: a σ-alg on S

Def (Measure)
Let $(S, \mathcal{F})$ be a measurable space. A non-negative set function

$$\mu : \mathcal{F} \to [0, \infty]$$

is called a measure on $(S, \mathcal{F})$ if

a) $\mu(\phi) = 0$

b) $\forall (A_i \in \mathcal{F}, i \in \mathbb{N})$ s.t. $A_i \cap A_j = \phi$, for $i \neq j$,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

countably additive or σ-additive

Remarks:

1) $(S, \mathcal{F}, \mu)$ is called a measure space.
2) If $\mu(S) = 1$, μ is called a probability measure, often denoted by $\mathbb{P}$.
3) Specifying $(S, \mathcal{F})$ constrains the possible measure that can be defined on it.

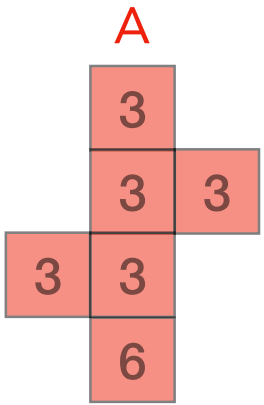
e.g.) (Not Lebesgue measurable)
Consider a measure λ on $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ satisfying
i) $\lambda([a,b]) = b - a$, for $b > a$
ii) $\lambda(x + A) = \lambda(A)$, for $x \in \mathbb{R}$, $A \in \mathcal{P}(\mathbb{R})$
$\exists$ a subset $V \in \mathcal{P}(\mathbb{R})$ for which $\lambda(V)$ cannot be defined consistantly.
(e.g. Vitaly set)

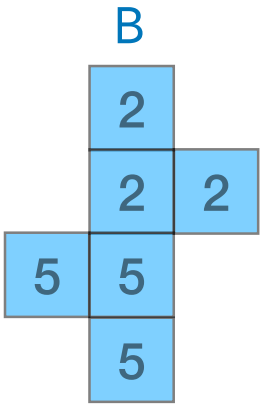
e.g.) Consider a unit ball $B \subset \mathbb{R}^3$ and drop a pt x u.a.r. on B. For any subset $A \subset B$ can we define

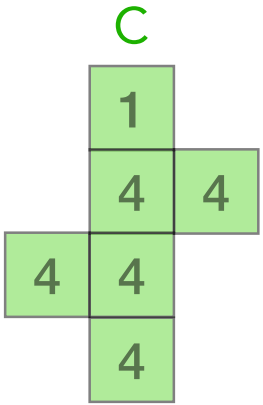
$$\mathbb{P}[x \in A] = \frac{\text{Volume}(A)}{\frac{4}{3}\pi} ?$$

No! (Banach-Tarski Paradox)
some $A \in \mathcal{P}(B)$ are not Lebesgue measurable

Problem of the Day







1. Bob chooses a die first.
2. Alice then chooses a die from the remaining two dice.
3. Each person rolls their die and the person with a higher number wins the round.
4. 11 rounds will be played with the same chosen dice.
5. Who has the advantage?

207

Answer:

Event "A beats B" = $\{(6,5), (6,2), (3,2)\}$ (A, B)

$\mathbb{P}[(6,5)] = \frac{1}{6} \cdot \frac{3}{6} = \frac{1}{12} = \mathbb{P}[(6,2)]$

$\mathbb{P}[(3,2)] = \frac{5}{6} \cdot \frac{1}{2} = \frac{5}{12}$

$\boxed{\mathbb{P}[\text{A beats B}] = \frac{7}{12} \approx 0.58}$

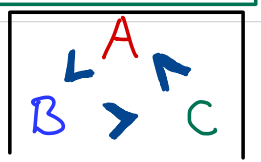
"B beats C" = $\{(5,4), (5,1), (2,1)\}$ (B, C)

$\boxed{\mathbb{P}[\text{B beats C}] = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{6}\right) = \frac{7}{12}}$

"C beats A" = $\{(4,3)\}$ (C, A)

$\boxed{\mathbb{P}[\text{C beats A}] = \frac{5}{6} \cdot \frac{5}{6} = \frac{25}{36} \approx 0.69}$

Non-transitive Dice



Probabilistic Rock-Paper-Scissors

$$\mathbb{P}\left(\begin{matrix}\text{Alice wins}\\ \text{the Game}\end{matrix}\right) = \sum_{k=6}^{11} \binom{11}{k} p^k (1-p)^{11-k},$$

where p = prob of winning a single round.

- If Bob chooses A, Alices can choose C
  $\Rightarrow p = \frac{25}{36} \Rightarrow \mathbb{P}\left(\begin{matrix}\text{Alice wins}\\ \text{the Game}\end{matrix}\right) = \boxed{0.915}$
- If Bob chooses B, Alices can choose A
  $\Rightarrow p = \frac{7}{12} \Rightarrow \mathbb{P}\left(\begin{matrix}\text{Alice wins}\\ \text{the Game}\end{matrix}\right) = \boxed{0.715}$
- If Bob chooses C, Alices can choose B
  $\Rightarrow p = \frac{7}{12} \Rightarrow \mathbb{P}\left(\begin{matrix}\text{Alice wins}\\ \text{the Game}\end{matrix}\right) = \boxed{0.715}$

Q Suppose n rounds are played instead $\lim_{n\to\infty} \mathbb{P}(\text{Alice wins the Game})$?

Lecture 2

Def (Random Variable)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a RV is a function $X: \Omega \to \mathbb{R}$ s.t.

$$\text{"}X \le x\text{"} := \{\omega \in \Omega \mid X(\omega) \le x\} \in \mathcal{F}$$

for all $x \in \mathbb{R}$. ($\mathcal{F}$-measurable)

Def (Distribution)

Given a RV X, its (cumulative) distribution function F_X is defined as

$$F_X(x) = \mathbb{P}[X \le x], \quad x \in (-\infty, \infty)$$

1) **Discrete RV**: (Image of X) Range(X) is either finite or countably infinite.

e.g.) For $A \in \mathcal{F}$, $I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$ Indicator RV

e.g.) Recall the coin toss example. $X(\omega) = \#$ Heads in $\omega \in \Omega$.

$(X=0) := \{TT\}$

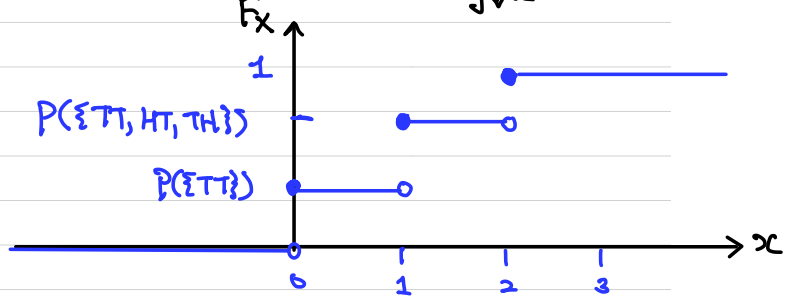
$(X=1) := \{HT, TH\}$ ← Disjoint since X is a function.

$(X \le 1) := \{TT, HT, TH\}$

$(X=2) := \{HH\}$

a) $F_X(x) = \sum_{y \le x} \mathbb{P}(X=y)$ (by σ-additivity)

b) F_X is right-continuous: $\lim_{y \downarrow x} F_X(y) = F_X(x)$



Other examples.

$X \sim$ Binomial(n, p), Range$(X) = \{0, 1, \dots, n\}$

$X \sim$ Poisson(λ), Range$(X) = \{0, 1, 2, \dots\}$

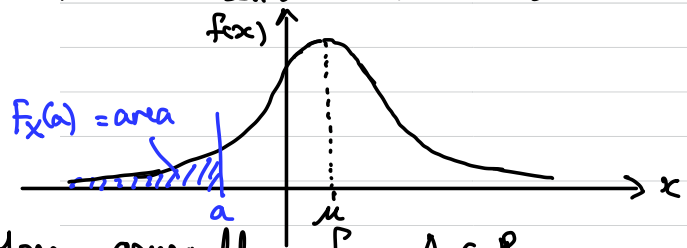
2) **Continuous RV**: $F_X(x)$ is cts $\forall\, x \in \mathbb{R}$

We will consider cts RVs with densities: $F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy$ ← density function

$$\frac{dF_X(x)}{dx} = f_X(x)$$

e.g.) $X \sim \text{Normal}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad x \in \mathbb{R}$$



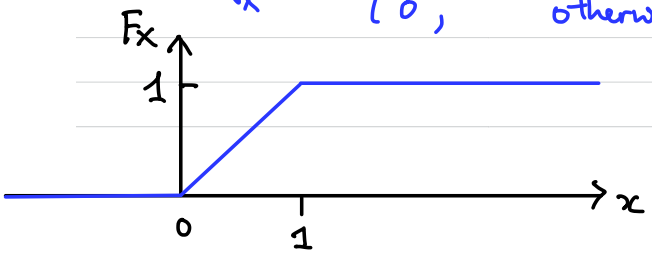
More generally, for $A \subseteq \mathbb{R}$,

$$\mathbb{P}[X \in A] = \int_A f_X(x)\,dx .$$

"$X \in A$" $= \{\omega \in \Omega \mid X(\omega) \in A\}$

e.g.) $X \sim \text{Uniform}[0,1]$,

$$f_X(x) = \begin{cases} 1, & x \in [0,1], \\ 0, & \text{otherwise}. \end{cases}$$



Def: (Expectation)

$g: \mathbb{R} \to \mathbb{R}$. (e.g. $g(x) = x^k$)

Discrete RV:

$$\mathbb{E}[g(X)] = \sum_{x \in \text{Range}(X)} g(x)\,\mathbb{P}[X = x],$$

Continuous RV: (f_X: PDF of X)

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx,$$

Provided that $\mathbb{E}[|g(X)|] < \infty$

absolutely convergent

Why is this condition needed?

Consider the discrete case:

What does it mean by

$\sum_{n=1}^{\infty} a_n$ converges to c?

$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + a_3 + a_4 + a_5 + \cdots$ "

S_1

S_2

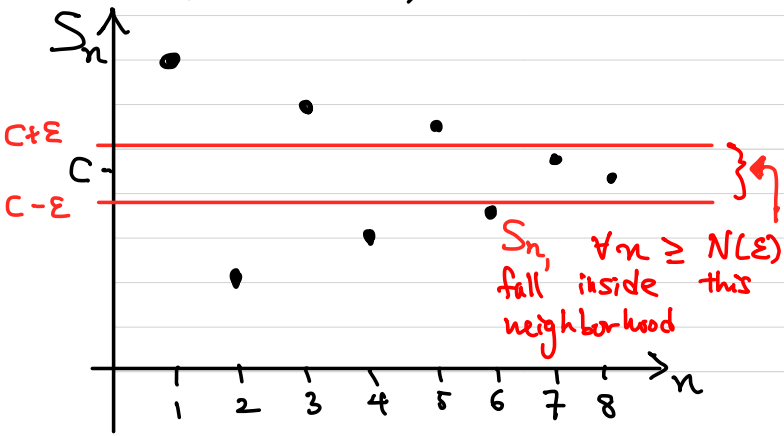
S_3

n^{th} Partial Sum

$S_n = \sum_{m=1}^{n} a_m$

$S_1, S_2, S_3, \ldots, S_n, \ldots$ $\lim_{n\to\infty} S_n = c$

For any $\varepsilon > 0$, $\exists$ an int $N(\varepsilon) > 0$ s.t. $|S_n - c| \le \varepsilon$, $\forall n \ge N(\varepsilon)$



Theorem (Riemann rearrangement theorem)
If $\sum_{n=1}^{\infty} a_n$ converges but $\sum_{n=1}^{\infty} |a_n|$ diverges, then for any given $r \in [-\infty, \infty]$, $\exists$ a permutation π of $\mathbb{N}$ s.t.
$\sum_{n=1}^{\infty} a_{\pi(n)} = r$.

$\Rightarrow$ Need $\mathbb{E}[|g(X)|] < \infty$ for $\mathbb{E}[g(X)]$ to be well defined.

Def (Mean) $\mathbb{E}[X]$

Def (Variance) $\mathbb{E}[(X - \mathbb{E}[X])^2]$

Claim (Linearity of Expectation)
Let $X_1, \ldots, X_n$ be RVs defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}[X_i]$, $\forall i = 1, \ldots, n$ are well defined. Then for all constants $c_1, \ldots, c_n$,
$\mathbb{E}[\sum_{i=1}^{n} c_i X_i] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i]$.

$$\mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$$

linearity of $\mathbb{E}$ $\Rightarrow$ $= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2$

$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

Def (Covariance)

A measure of association b/w X & Y.

Let X, Y be RVs on the same probability space. Then

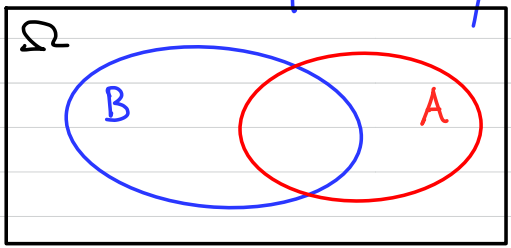
$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

linearity of $\mathbb{E}$ $\Rightarrow$ $= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Exercise: Show

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

Conditional probability



$(\Omega, \mathcal{F}, \mathbb{P})$

Suppose $\mathbb{P}(B) > 0$

Given that B happens, what is the probability that A also happens?

Want to consider a new probability space $(B, \mathcal{F}_B, \mathbb{P}_B)$.

How should we define $\mathbb{P}_B$ so that it is consistent with $\mathbb{P}$?

For all $E_1, E_2 \in \mathcal{F}$ s.t. $E_1 \cap B \neq \phi$ and $E_2 \cap B \neq \phi$, we want

$$\frac{\mathbb{P}[E_1 \cap B]}{\mathbb{P}[E_2 \cap B]} = \frac{\mathbb{P}_B(E_1 \cap B)}{\mathbb{P}_B(E_2 \cap B)}$$

$\Rightarrow \mathbb{P}_B = c\,\mathbb{P}$, c constant.

$\mathbb{P}_B(B) = 1 \Rightarrow C = \frac{1}{\mathbb{P}(B)}$

$\Rightarrow$ For all $A, B \in \mathcal{F}$ s.t. $\mathbb{P}[B] > 0$,

$$\mathbb{P}[A|B] := \mathbb{P}_B[A \cap B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

If A, B are independent, then

$$\mathbb{P}[A|B] = \mathbb{P}[A],$$
$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B].$$

Def (Independence of Events)

Events $E_1, \dots, E_n$ are independent iff for every $k = 2, \dots, n$ and for every k-subset $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$,

$$\mathbb{P}\left[\bigcap_{j=1}^{k} E_{i_j}\right] = \prod_{j=1}^{k} \mathbb{P}[E_{i_j}].$$

e.g.) $n = 3$. Need to check

$$\mathbb{P}[E_1 \cap E_2 \cap E_3] = \mathbb{P}[E_1]\,\mathbb{P}[E_2]\,\mathbb{P}[E_3]$$
$$\mathbb{P}[E_1 \cap E_2] = \mathbb{P}[E_1]\,\mathbb{P}[E_2]$$
$$\mathbb{P}[E_1 \cap E_3] = \mathbb{P}[E_1]\,\mathbb{P}[E_3]$$
$$\mathbb{P}[E_2 \cap E_3] = \mathbb{P}[E_2]\,\mathbb{P}[E_3]$$

Def (Independence of RVs)

RVs X, Y on the same prob space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be independent ($\perp\!\!\!\perp$) iff

$$\mathbb{P}[(X \le x) \cap (Y \le y)] = \mathbb{P}[X \le x]\,\mathbb{P}[Y \le y],$$
$\forall x, y \in \mathbb{R}$.

(Equivalent conditions: $\forall x, y \in \mathbb{R}$

Discrete case: $\mathbb{P}[(X = x) \cap (Y = y)] = \mathbb{P}[X = x]\,\mathbb{P}[Y = y]$

Continuous case: $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$

↑ more in later lectures.)

RVs $X_1, \dots, X_n$ on the same prob space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be mutually $\perp\!\!\!\perp$ iff

$$\mathbb{P}\left[\bigcap_{i=1}^{n} (X_i \le x_i)\right] = \prod_{i=1}^{n} \mathbb{P}[X_i \le x_i],$$

$\forall x_1, \dots, x_n \in \mathbb{R}$.

Remark: (Mutual $\perp\!\!\!\perp$) $\Rightarrow$ (pairwise $\perp\!\!\!\perp$)

$\not\Leftarrow$

e.g.) See HW1, Q2

Remarks

1) $X \perp\!\!\!\perp Y \Rightarrow \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

$\not\Leftarrow$

e.g.

(x,y)	$\mathbb{P}[(X=x)\cap(Y=y)]$
$-1, 0$	$1/3$
$0, 1$	$1/3$
$1, 0$	$1/3$

$X \not\perp\!\!\!\perp Y$ since

$\mathbb{P}[(X=1)\cap(Y=1)] = 0$

$\mathbb{P}[X=1]\,\mathbb{P}[Y=1] = \frac{1}{9}$

$\mathbb{E}[XY] = 0$
$\mathbb{E}[X] = 0$
$\mathbb{E}[Y] = \frac{1}{3}$

$\Big\}\ \mathbb{E}[XY] = 0 = \mathbb{E}[X]\,\mathbb{E}[Y]$

2) $X \perp\!\!\!\perp Y \Rightarrow \mathrm{Cov}(X,Y) = 0$

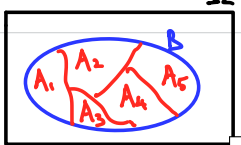
$[\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)]$

$\not\Leftarrow$

Def (Partition)

$A_1, \ldots, A_n \in \mathcal{F}$ partitions $B \in \mathcal{F}$ if

1) $B = A_1 \cup A_2 \cup \cdots \cup A_n$
2) $A_i \cap A_j = \phi$ if $i \neq j$.

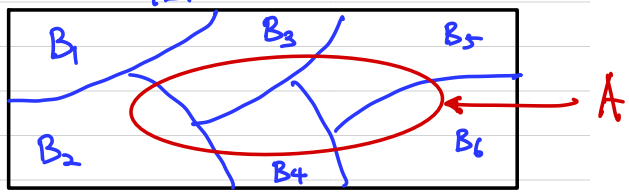


Thm (Law of Total Probability)

Suppose $B_1, \ldots, B_n \in \mathcal{F}$ is a partition of Ω s.t. $\mathbb{P}[B_i] > 0\ \forall i$. Then, for any $A \in \mathcal{F}$,

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]$$

Pf



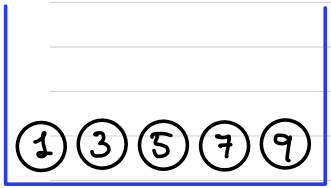
$B_1, \ldots, B_n$ a partition of Ω

$\Rightarrow (B_1 \cap A), \ldots, (B_n \cap A)$ a partion of A

$$\Rightarrow \mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \cap B_i] \quad \text{(by additivity of measure)}$$
$$= \sum_{i=1}^{n} \mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i] \quad \text{(by definition of conditional prob)}$$

Remark: A similar result applies to a countably infinite partition of Ω.

Problem of the Day



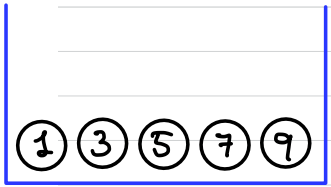
a) Draw a ball uniformly at random (u.a.r.)
b) record the number on the ball
c) return the ball to the urn
d) repeat n times.

Let S_n = Sum over the observed numbers.

Q $\mathbb{P}[S_n$ is divisible by $5]$?

Lecture 3

Problem from Lecture 2



← Urn

a) Draw a ball uniformly at random (u.a.r.)
b) record the number on the ball
c) return the ball to the urn
d) repeat n times.

Let S_n = sum over the observed numbers

Q $\mathbb{P}[S_n \text{ is divisible by } 5]$?

X_i = # from the ith draw
$S_n = X_1 + \cdots + X_n$
Let $\mathcal{R} = \{1, 3, 5, 7, 9\}$

Law of Total Probability ⇒

$$\mathbb{P}[S_n \text{ is divisible by } 5] = \sum_{a \in \mathcal{R}} \mathbb{P}[S_n \text{ is divisible by } 5 \mid X_n = a]\,\mathbb{P}[X_n = a]$$

$S_{n-1} + a$ is divisible 5 ; $\mathbb{P}[X_n = a] = \frac{1}{5}$

a	$S_{n-1} \bmod 5$
1	4
3	2
5	0
7	3
9	1

$$= \frac{1}{5} \sum_{k=0}^{4} \mathbb{P}[S_{n-1} = k \bmod 5]$$

Call this event E_k.

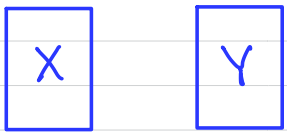
$E_0, E_1, \ldots, E_4$ partition Ω ⇒

$$= \frac{1}{5}\underbrace{\mathbb{P}[\Omega]}_{1} = \boxed{\frac{1}{5}}$$

Problem of the Day

Alice and Bob play the following guessing game.

1) Bob writes down two different numbers on two separate cards:



Say $X, Y > 0$

2) Alice picks one of the cards uniformly at random and looks at the number.

3) Alice wins if she correctly guesses which of the two cards has a larger number.

Q Can Alice do better than random guess?

Bayes' formula

Let $B_1, \dots, B_n \in \mathcal{F}$ be a partition of Ω. Then, for any $A \in \mathcal{F}$ with $\mathbb{P}[A] > 0$,

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[B_i \cap A]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}[A \mid B_j]\,\mathbb{P}[B_j]}$$

e.g.) Suppose you get tested for a disease and the test result comes back "+".

Q Should you worry?

Suppose disease prevalence = p

"D" = event of having the disease

$\mathbb{P}[D] = p$ (prior)

$\mathbb{P}[D^c] = 1-p$

test result

$\mathbb{P}(+ \mid D^c) = \text{FPR}$

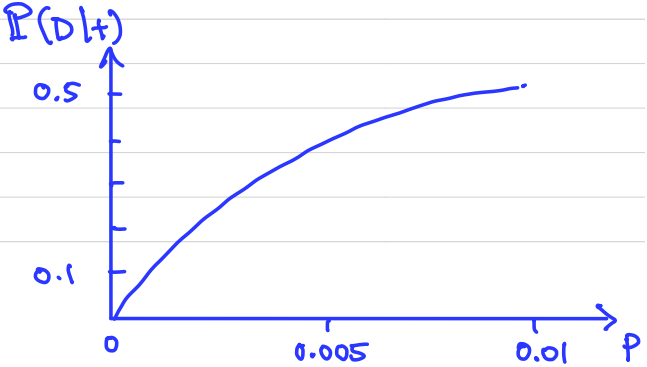
$\mathbb{P}(- \mid D) = \text{FNR}$

$$\mathbb{P}[D \mid +] = \frac{\mathbb{P}[+ \mid D]\,\mathbb{P}[D]}{\mathbb{P}[+ \mid D]\,\mathbb{P}[D] + \mathbb{P}[+ \mid D^c]\,\mathbb{P}[D^c]}$$ (posterior)

$$= \frac{(1-\text{FNR})\,p}{(1-\text{FNR})\,p + \text{FPR}\,(1-p)}$$

$$= \frac{(1-\text{FNR})\,p}{\text{FPR} + p\,(1-\text{FNR}-\text{FPR})}$$

If FPR = FNR = 0.01, then



$X \sim$ Bernoulli(p), $0 < p < 1$

Success Fail

$\mathbb{P}[X=1] = p$, $\mathbb{P}[X=0] = 1-p$.

$\mathbb{E}[X] = 1 \cdot p + 0(1-p) = p$

$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$= p - p^2 = p(1-p)$

Bernoulli process

independent and identically distributed

$X_1, X_2, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(p)

0 0 1 0 0 0 1 1 0 0 0 ... 1 0 1 0 1

S_n = total number of 1s

$= X_1 + \cdots + X_n$

$S_n \sim$ Binomial(n, p)

$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$

$\mathbb{E}[S_n] = \mathbb{E}[X_1 + \cdots + X_n]$

by linearity of $\mathbb{E}$ $= \sum_{i=1}^{n} \mathbb{E}[X_i] = np$

$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$ (by $\perp\!\!\!\perp$ of $X_1, \ldots, X_n$)

$= np(1-p)$

0 0 1 0 0 0 1 1 0 0 0 ... 1 0 1 0 1

$W_1 = 3$ $W_2 = 4$ $W_3 = 1$

W_i = waiting time between the $(i-1)^{th}$ & i^{th} successes

$W_1, W_2, W_3, \ldots$ are $\perp\!\!\!\perp$

$W_i \sim$ Geometric(p) $\forall\, i$

$\mathbb{P}[W = k] = (1-p)^{k-1} p$, $k = 1, 2, 3, \ldots$

$\mathbb{E}[W] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = \frac{1}{p}$

$\mathrm{Var}[W] = \mathbb{E}[W^2] - \mathbb{E}[W]^2$

$= \left(\sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p\right) - \left(\frac{1}{p}\right)^2$

$= \frac{1-p}{p^2}$

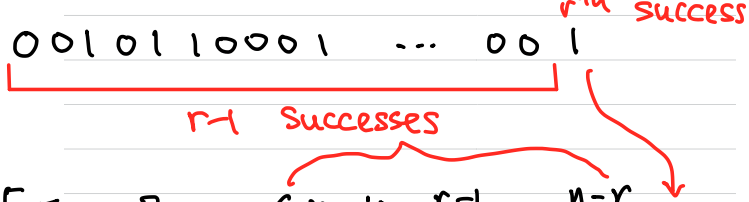
These moments can be computed more easily using generating functions (Later lectures)

$r \in \mathbb{N} = \{1, 2, 3, \dots\}$ fixed positive int.

T_r = Total waiting time to the r^{th} success

$T_r = W_1 + W_2 + \cdots + W_r$

where $W_1, \dots, W_r \overset{iid}{\sim} \text{Geometric}(p)$



$$\mathbb{P}[T_r = n] = \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r} \cdot p$$

$$= \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

F_r = # failures before r^{th} success

$F_r + r = T_r$

$$\mathbb{P}[F_r = k] = \binom{r+k-1}{r-1} p^r (1-p)^k$$

$$= \binom{r+k-1}{k} p^r (1-p)^k$$

$F_r \sim \text{Negative Binomial}(r, p)$

We can compute $\mathbb{E}[F_r]$ & $\text{Var}[F_r]$ without using the prob. mass function directly.

$$\mathbb{E}[F_r] = \mathbb{E}[T_r] - r$$

$$= \frac{r}{p} - r = \frac{r(1-p)}{p}$$

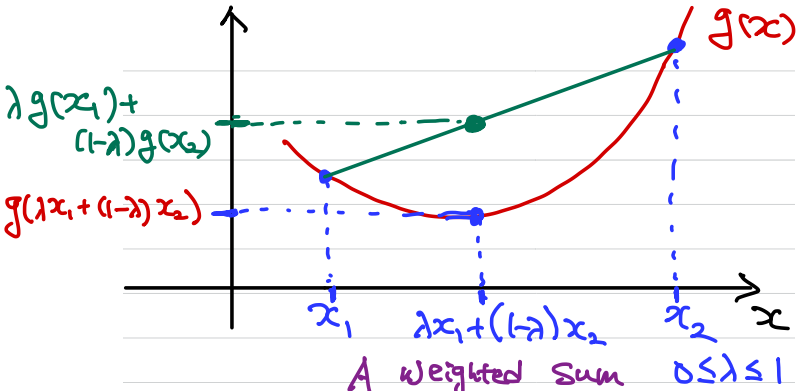
$$\text{Var}[F_r] = \text{Var}[T_r - r]$$

$$= \text{Var}[T_r]$$

$$= r \, \text{Var}[W_1]$$

$$= \frac{r(1-p)}{p^2}$$

NB distribution is widely used in single-cell genomics



Def (Convex function)
A function $g:(a,b)\to\mathbb{R}$ is said to be convex if

$$g(\lambda x_1+(1-\lambda)x_2)\le \lambda g(x_1)+(1-\lambda)g(x_2)$$

$\forall\, x_1, x_2 \in (a,b)$ and $\forall\, 0\le\lambda\le 1$.

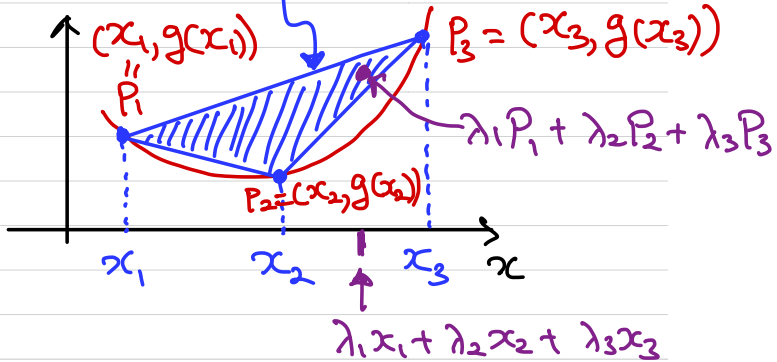
g is called strictly convex if the equality holds only for $\lambda=0$, $\lambda=1$ or $x_1=x_2$.

(The graph of g over (a,b) contains no straight lines.)

$P_1, P_2, P_3 \in \mathbb{R}^2$

Convex hull of $\{P_1, P_2, P_3\}$

$$=\left\{(x,y)\in\mathbb{R}^2 \;\middle|\; (x,y)=\lambda_1P_1+\lambda_2P_2+\lambda_3P_3,\ \lambda_1+\lambda_2+\lambda_3=1,\ \lambda_i\ge 0\right\}$$



$$g(\lambda_1x_1+\lambda_2x_2+\lambda_3x_3)\le \lambda_1g(x_1)+\lambda_2g(x_2)+\lambda_3g(x_3)$$

$\forall\, x_1,x_2,x_3\in\mathbb{R}$ and $\forall \sum_{i=1}^{3}\lambda_i=1$, $\lambda_i\in[0,1]$

g strictly convex $\Leftrightarrow$ equality holds only for 1) $\lambda_i=1$ and $\lambda_j=0$ for $j\neq i$ or 2) $x_1=x_2=x_3$.

Thm (Jensen's Inequality)
A convex function $g:(a,b)\to\mathbb{R}$ satisfies $g(\sum_{i=1}^n \lambda_i x_i) \le \sum_{i=1}^n \lambda_i g(x_i)$,

$\forall \lambda_1,\dots,\lambda_n$ satisfying $\sum_{i=1}^n \lambda_i = 1$, $\lambda_i \ge 0$ $\forall i$.

Corollary Let X be a $\mathbb{R}$-valued discrete RV and g a convex function. Then,

$$g(\mathbb{E}[X]) \le \mathbb{E}[g(X)]$$

If g is strictly cvx and X is not constant, then

$$g(\mathbb{E}[X]) < \mathbb{E}[g(X)].$$

Pf Let $\lambda_i = \mathbb{P}(X=x_i)$ in Jensen's inequality. □

(Holds for cts RVs also.)

Pf of Jensen's inequality

Induction on n.

- Base case: $n=2$. True by cvx def.
- Induction Hyp: Assume true for all $n=2,\dots,k$
- Show for $n=k+1$:

$$g\Big(\sum_{i=1}^{k+1} \lambda_i x_i\Big) = g\Big(\sum_{i=1}^{k} \lambda_i x_i + \lambda_{k+1} x_{k+1}\Big)$$

$$= g\left((1-\lambda_{k+1}) \underbrace{\frac{\sum_{i=1}^k \lambda_i x_i}{(1-\lambda_{k+1})}} + \lambda_{k+1} x_{k+1}\right)$$

Def of Convexity of g $\Rightarrow$

$$\le (1-\lambda_{k+1})\, g\left(\frac{\sum_{i=1}^k \lambda_i x_i}{1-\lambda_{k+1}}\right) + \lambda_{k+1} g(x_{k+1})$$

↓ By induction hyp.

$$\le (1-\lambda_{k+1}) \left[\sum_{i=1}^k \frac{\lambda_i}{1-\lambda_{k+1}} g(x_i)\right] + \lambda_{k+1} g(x_{k+1})$$

$$= \sum_{i=1}^{k+1} \lambda_i\, g(x_i)$$ □

Suppose $\mathbb{P}$ & $\mathbb{Q}$ are two probability measures on $(\Omega, \mathcal{F})$, and let X be a discrete RV s.t.

$$\mathbb{P}[X = x] = p(x), \quad x \in \text{Range}(X)$$

$$\mathbb{Q}[X = x] = q(x), \quad x \in \text{Range}(X)$$

Def (Shannon entropy)

$$H(\mathbb{P}) = -\sum_x p(x) \log p(x)\, dx = -\mathbb{E}_p[\log p(X)]$$

Def (Cross entropy)

Cross entropy of $\mathbb{Q}$ relative to $\mathbb{P}$:

$$H(\mathbb{P}, \mathbb{Q}) = -\sum_x p(x) \log q(x)\, dx$$

$$= -\mathbb{E}_p[\log q(X)]$$

Def (KL Divergence)

The Kullback-Leibler divergence of $\mathbb{P}$ from $\mathbb{Q}$ is defined as

$$KL(\mathbb{P} \| \mathbb{Q}) = H(\mathbb{P}, \mathbb{Q}) - H(\mathbb{P})$$

$$= -\sum_x p(x) \log \left[\frac{q(x)}{p(x)}\right]$$

$$= -\mathbb{E}_p\left[\log \frac{q(X)}{p(X)}\right]$$

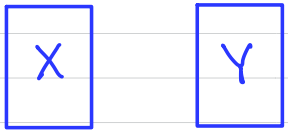
For continuous RV, replace $p(x)$ with pdf and $\sum_x$ with $\int dx$

Has lots of applications in machine learning (e.g., loss function, EM algorithm, VAE)

Problem from Lecture 3

Alice and Bob play the following guessing game.

1) Bob writes down two different numbers on two separate cards:



Say $X, Y > 0$

2) Alice picks one of the cards uniformly at random and looks at the number.

3) Alice wins if she correctly guesses which of the two cards has a larger number.

Q Can Alice do better than random guess?

Ans Yes! Here is a strategy.

Suppose Bob's numbers are $X < Y$.

Let A = number Alice picked.

Generate a random number $R \in \mathbb{R}_+$

If $R > A$, call the other card as having a larger number.

If $R < A$, call A to be the larger number.

$\mathbb{P}[R<X] = a$ $\mathbb{P}[X<R<Y] = b$ $\mathbb{P}[R>Y] = c$

Alice can't compute these probabilities since she doesn't know how Bob generated X and Y. Nevertheless, she can show:

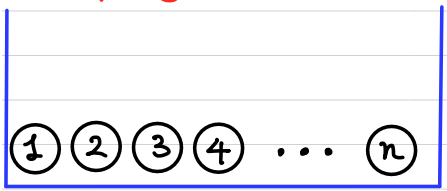
$$\mathbb{P}[\text{Correct}] = \underbrace{\mathbb{P}[\text{Correct} \mid A=X]}_{b+c}\underbrace{\mathbb{P}[A=X]}_{\frac{1}{2}} + \underbrace{\mathbb{P}[\text{Correct} \mid A=Y]}_{a+b}\underbrace{\mathbb{P}[A=Y]}_{\frac{1}{2}}$$

$$= \frac{1}{2}(a+b+b+c) = \frac{1}{2} + \frac{b}{2} > \frac{1}{2}$$

since $b > 0$.

Problem of the Day

Sampling without replacement



Suppose you sample $k \le n$ numbers u.a.r. from $\{1, 2, \dots, n\}$ without replacement. What is the expected sum of the top $k-1$ numbers?

Hint: Use the Tail Sum formula.

Thm (Tail Sum Formula)
Let X be a RV with range $\{0, 1, 2, \dots\}$. Then,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}[X \ge k]$$

Pf

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\, \mathbb{P}[X=k]$$

$$= \mathbb{P}[X=1] + \mathbb{P}[X=2] + \mathbb{P}[X=2] + \mathbb{P}[X=3] + \mathbb{P}[X=3] + \mathbb{P}[X=3] + \mathbb{P}[X=4] + \mathbb{P}[X=4] + \mathbb{P}[X=4] + \mathbb{P}[X=4]$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$= \mathbb{P}[X \ge 1] \quad \mathbb{P}[X \ge 2] \quad \mathbb{P}[X \ge 3] \quad \mathbb{P}[X \ge 4]$$

these all non-negative numbers, so can be summed in any order

Evidence Lower Bound (ELBO)

X = observed data
Z = latent variable
q = a distribution over Z

log-likelihood ↓ ← parameters

$$\log p(X|\theta) = \mathcal{L}(q, X, \theta) + KL(q \| p),$$

where

$$\mathcal{L}(q, X, \theta) = \mathbb{E}_q\left\{\log\left[\frac{p(X,Z|\theta)}{q(Z)}\right]\right\} \quad \text{(ELBO)}$$

$$KL(q\|p) = -\mathbb{E}_q\left\{\log\left[\frac{p(Z|X,\theta)}{q(Z)}\right]\right\}$$

Key Facts

1) $\mathcal{L}(q, X, \theta)$ is a lower bound on the log-likelihood since $KL(q\|p) \geq 0$ for all p and q.

2) ELBO is much easier to compute & optimize than $\log p(X|\theta)$.

Claim

1) $KL(p\|q) \geq 0$
2) $KL(p\|q) = 0$ iff $p(x) = q(x)$ $\forall x$

pf $-\log$ is a cvx function, so Jensen's inequality for RV $Z = \frac{q(X)}{p(X)}$ implies

$$-\log \mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] \leq -\mathbb{E}_p\left[\log \frac{q(X)}{p(X)}\right] \quad (*)$$

$$\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] = \sum_x p(x)\frac{q(x)}{p(x)} = \sum_x q(x) = 1$$

$\Rightarrow$ LHS $= 0$

In fact, $-\log$ is a strictly cvx function, so equality in $(*)$ holds iff $p(x) = q(x)$ $\forall x$

Remark: In general, $KL(p\|q) \neq KL(q\|p)$

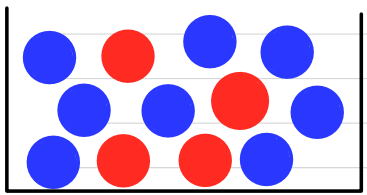
Hypergeometric Distribution

- Often used in enrichment analysis

e.g. N = total # genes assayed
M = # genes involved in a particular biological pathway
n = # significant genes
k = # significant genes in the pathway

Q: Is the pathway over-represented? Can we provide a quantitative measure of enrichment?

Mathematical Model



B blue balls
R red balls
$N = B + R$

Sample size $n < N$

1) Sample n balls u.a.r. with replacement.
Sample space $\Omega = \{\bullet, \bullet\}^n$, $|\Omega| = 2^n$.
Let $X(\omega)$ = # blue balls in $\omega \in \Omega$
Then, $X \sim \text{Binomial}(n, p)$ where $p = \frac{B}{N}$.

$$\mathbb{P}_1[X=k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, 1, 2, \cdots, n\}$$

2) Sample n balls u.a.r. without replacement.
Let $X(\omega)$ = # blue balls in $\omega \in \Omega$

$\mathbb{P}_2[X=k] = 0$ if $k > B$ or $n-k > R$

For $\max\{0, n-R\} \le k \le B$?

$\omega =$ (k blue, then $n-k$ red)

$$\frac{B}{N}\left[\frac{B-1}{N-1}\right] \cdots \left[\frac{B-k+1}{N-k+1}\right]\left[\frac{R}{N-k}\right]\left[\frac{R-1}{N-k-1}\right] \cdots \left[\frac{R-(n-k)+1}{N-n+1}\right]$$

$$\mathbb{P}_2[\{\omega\}] = \frac{\frac{B!}{(B-k)!}\,\frac{R!}{(R-(n-k))!}}{\frac{N!}{(N-n)!}} =: q$$

Key observation Another $\omega' \in \Omega$ with k blue balls and $n-k$ red balls has $\mathbb{P}_2[\{\omega'\}] = \mathbb{P}_2[\{\omega\}]$.

Permutation invariant.

There are $\binom{n}{k}$ distinct sequences with k blue balls and $n-k$ red balls

$$\Rightarrow \mathbb{P}_2[X=k] = \binom{n}{k} q$$

$$= \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}}.$$

This distribution is called Hypergeometric (N, B, n)

(Note: $\mathbb{P}_2[X=k] \to \binom{n}{k} p^k (1-p)^{n-k}$ as $B \to \infty$, $N \to \infty$ s.t. $\frac{B}{N} \to p$.)

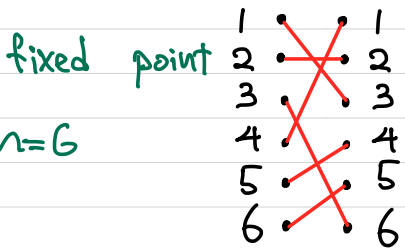
Exchangeability

Permutation: one-to-one map

$$\pi : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$$

There are $n!$ distinct permutations of $\{1, \dots, n\}$

e.g.



Def A sequence $(X_1, \dots, X_n)$ of n RVs on the same probability space is said to be exchangeable if $(X_{\pi(1)}, \dots, X_{\pi(n)})$ has the same joint distribution as $(X_1, \dots, X_n)$ for all permutations π of $\{1, \dots, n\}$.

e.g.) $n=2$, discrete RVs.

$\mathbb{P}[X_1 = a, X_2 = b] = \mathbb{P}[X_2 = a, X_1 = b]$

$\forall\ a, b \in \mathbb{R}$.

Claim If $(X_1, \ldots, X_n)$ is exchangeable, then all subsequences $(X_{i_1}, \ldots, X_{i_m})$ of a given length $m \in \{1, 2, \ldots, n\}$ have the same joint distribution.

e.g. $n=3$, discrete RVs

(X_1, X_2, X_3) exchangeable $\Rightarrow$

$$\mathbb{P}[X_1=a, X_2=b, X_3=c] = \mathbb{P}[X_1=a, X_3=b, X_2=c]$$
$$= \mathbb{P}[X_2=a, X_1=b, X_3=c] = \mathbb{P}[X_2=a, X_3=b, X_1=c]$$
$$= \mathbb{P}[X_3=a, X_2=b, X_1=c] = \mathbb{P}[X_3=a, X_1=b, X_2=c]$$

Sum over c $\Rightarrow$

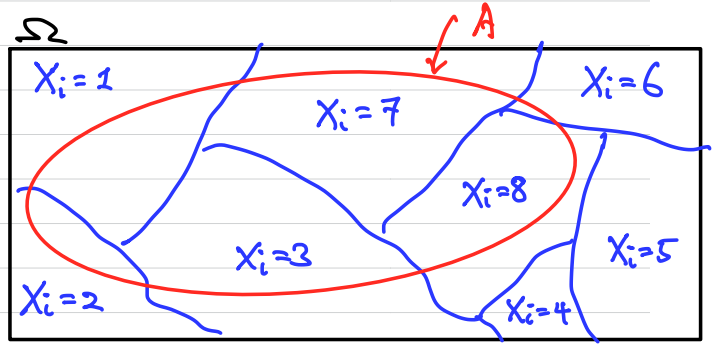
$$\mathbb{P}[X_1=a, X_2=b] = \mathbb{P}[X_1=a, X_3=b]$$
$$= \mathbb{P}[X_2=a, X_1=b] = \mathbb{P}[X_2=a, X_3=b]$$
$$= \mathbb{P}[X_3=a, X_2=b] = \mathbb{P}[X_3=a, X_1=b]$$
$$\forall\ a, b \in \mathbb{R}.$$

Sum over b $\Rightarrow$

$$\mathbb{P}[X_1=a] = \mathbb{P}[X_2=a] = \mathbb{P}[X_3=a]$$
$$\forall\ a \in \mathbb{R}$$

Why are these statements true?

Since $\{(X_i=c), c \in \text{Range}(X_i)\}$ partitions Ω, for any $A \in \mathcal{F}$, $\{A \cap (X_i=c), c \in \text{Range}(X_i)\}$ partitions A



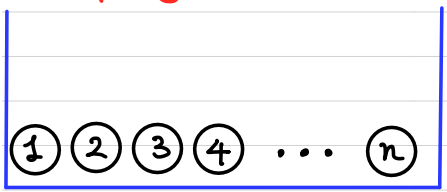
$$\Rightarrow \mathbb{P}[A] = \mathbb{P}\left[\bigcup_c A \cap (X_i=c)\right]$$

by σ-additivity of $\mathbb{P}$ $= \sum_c \mathbb{P}[A \cap (X_i=c)]$

This procedure is referred to as marginalizing out X_i

Problem from Lecture 4

Sampling without replacement



Suppose you sample $k \le n$ numbers u.a.r. from $\{1, 2, \dots, n\}$ without replacement. What is the expected sum of the top $k-1$ numbers?

Hint: Use the Tail Sum formula.

Thm (Tail Sum Formula)
Let X be a RV with range $\{0, 1, 2, \dots\}$. Then,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}[X \ge k]$$

Pf

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\, \mathbb{P}[X=k]$$

$$= \mathbb{P}[X=1] + \mathbb{P}[X=2] + \mathbb{P}[X=2] + \mathbb{P}[X=3] + \mathbb{P}[X=3] + \mathbb{P}[X=3] + \mathbb{P}[X=4] + \mathbb{P}[X=4] + \mathbb{P}[X=4] + \mathbb{P}[X=4]$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$= \mathbb{P}[X \ge 1] \quad \mathbb{P}[X \ge 2] \quad \mathbb{P}[X \ge 3] \quad \mathbb{P}[X \ge 4]$$

these all non-negative numbers, so can be summed in any order

Solution

Let $X_1, \dots, X_K$ denote the sample.

$M = \min\{X_1, \dots, X_K\}$

Want $\mathbb{E}[X_1 + \cdots + X_K - M] = \sum_{i=1}^{K} \mathbb{E}(X_i) - \mathbb{E}[M]$

For $a_1, \dots, a_K$ distinct elements of $\{1, \dots, n\}$,

$$\mathbb{P}[X_1 = a_1, X_2 = a_2, \dots, X_K = a_K] = \frac{1}{n}\left(\frac{1}{n-1}\right)\cdots\left(\frac{1}{n-K+1}\right)$$

$\Rightarrow (X_1, \dots, X_K)$ is exchangeable

$\Rightarrow X_1, \dots, X_K$ are identically distributed

$\forall i$, $\mathbb{E}[X_i] = \mathbb{E}[X_1] = \sum_{a=1}^{n} a\,\mathbb{P}[X_1 = a] = \frac{n+1}{2}$ (since $\mathbb{P}[X_1 = a] = \frac{1}{n}$)

$\mathbb{P}[M = a] \neq 0$ for $a = 1, \dots, n-K+1$.

$$\mathbb{P}[M \geq a] = \mathbb{P}[X_1 \geq a, \dots, X_K \geq a] = \frac{\binom{a-1}{0}\binom{n-a+1}{K}}{\binom{n}{K}}$$

$$\mathbb{E}[M] = \sum_{a=1}^{n-K+1} \mathbb{P}[M \geq a] \quad \text{(Tail sum)}$$

$$= \frac{1}{\binom{n}{K}} \sum_{a=1}^{n-K+1} \binom{n-a+1}{K}$$

$$\binom{n}{k} + \binom{n-1}{k} + \cdots + \binom{k}{k} = \binom{n+1}{k+1}$$

Hockey-stick identity

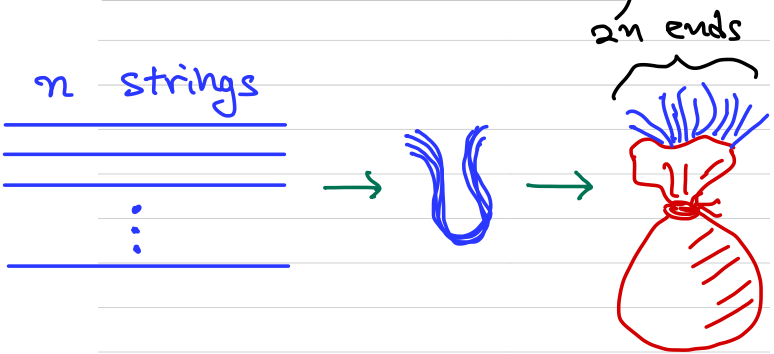
$$\Rightarrow \mathbb{E}[M] = \frac{\binom{n+1}{K+1}}{\binom{n}{K}} = \frac{n+1}{K+1}$$

So,

$$\mathbb{E}[X_1 + \cdots + X_K - M] = \frac{K(n+1)}{2} - \frac{n+1}{K+1}$$

$$= \boxed{(n+1)\left[\frac{K}{2} - \frac{1}{K+1}\right]}$$

Problem of the day



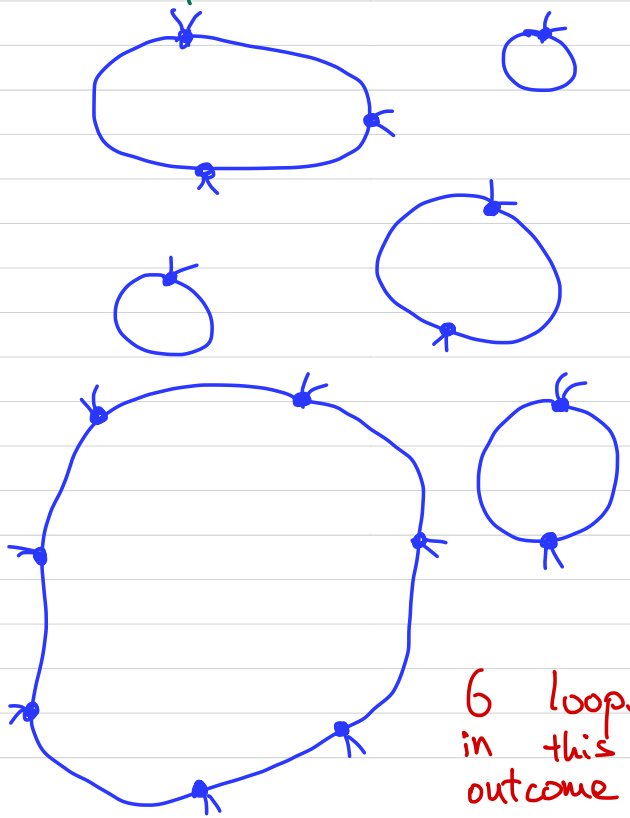
Suppose you tie two free ends u.a.r. and continue until no more free ends remain.

Q What is the expected number of loops formed?

Min = 1

max = n

Example: n=16



6 loops in this outcome

Back to $X \sim \text{Hypergeometric}(N, B, n)$

$\mathbb{E}[X] = \sum_k k\, \mathbb{P}[X=k]$ is not straightforward to evaluate.

Define $I_i(\omega) = \begin{cases} 1, & \text{if the } i\text{th draw is blue for } \omega \in \Omega \\ 0, & \text{otherwise} \end{cases}$

Indicator RV

$X = I_1 + I_2 + \cdots + I_n$

$(I_1, \ldots, I_n)$ is exchangeable! (see page 4-3)

$\Rightarrow \mathbb{P}[I_i = 1] = \mathbb{P}[I_1 = 1] \quad \forall i.$

$\mathbb{P}[I_1 = 1] = \frac{B}{N} = \mathbb{E}[I_1]$

$\Rightarrow \mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[I_i] = n\,\mathbb{E}[I_1] = \frac{nB}{N}$

Note that this is the same as the expectation of Binomial$(n, \frac{B}{N})$

How about variance?

$\text{Var}(X) = \text{Cov}(I_1 + \cdots + I_n, I_1 + \cdots + I_n)$

by bilinearity of Cov

$= \sum_{i=1}^{n} \text{Var}(I_i) + \sum_{i,j:\, i \neq j} \text{Cov}(I_i, I_j)$

by exchangeability $= n\,\text{Var}(I_1) + n(n-1)\,\text{Cov}(I_1, I_2)$

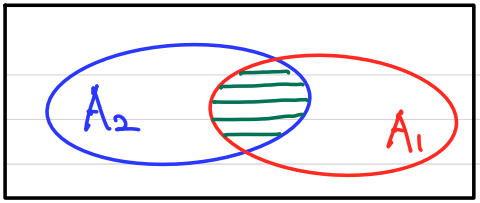
$\text{Var}(I_1) = \frac{B}{N}\left(1 - \frac{B}{N}\right)$

$\text{Cov}(I_1, I_2) = \mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\,\mathbb{E}[I_2]$

$\mathbb{P}[I_1 = 1, I_2 = 1] = \frac{B}{N}\left(\frac{B-1}{N-1}\right) = \mathbb{E}[I_1 I_2]$

$\Rightarrow \text{Cov}(I_1, I_2) = -\frac{B(N-B)}{N^2(N-1)} < 0$

$\Rightarrow \text{Var}(X) = n \underbrace{\frac{B}{N}\left(\frac{N-B}{N}\right)}_{\text{Variance of Binomial}(n, \frac{B}{N})} \underbrace{\left(\frac{N-n}{N-1}\right)}_{\leq 1}$



$$\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2] - \mathbb{P}[A_1 \cap A_2]$$

Thm (Union Bound, Boole's inequality)
Let $A_1, \cdots, A_n \in \mathcal{F}$ be a collection of events on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then,

$$\mathbb{P}[A_1 \cup A_2 \cup \cdots \cup A_n] \le \sum_{i=1}^{n} \mathbb{P}[A_i]$$

Pf Use induction on n.

Remark: denoted Σ_1 in Q7 of PS#1.

$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] \ge \Sigma_1 - \Sigma_2$$
$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] \le \Sigma_1 - \Sigma_2 + \Sigma_3$$
$$\mathbb{P}\left[\bigcup_{i=1}^{n} A_i\right] \ge \Sigma_1 - \Sigma_2 + \Sigma_3 - \Sigma_4$$
and so on

Thm (Cauchy–Schwarz Inequality)
Let X, Y be two RVs on the same probability space. Then,

$$(\mathbb{E}[XY])^2 \le \mathbb{E}[X^2]\,\mathbb{E}[Y^2]$$

$$\Rightarrow [\mathrm{Cov}(X,Y)]^2 \le \mathrm{Var}(X)\,\mathrm{Var}(Y)$$

which is used to prove many results e.g. Cramér-Rao Inequality in theoretical Statistics

Pf For constants $a, b \in \mathbb{R}$,

$$0 \le \mathbb{E}[(aX - bY)^2] = \mathbb{E}[a^2X^2 + b^2Y^2 - 2abXY]$$
$$\Rightarrow 2ab\,\mathbb{E}[XY] \le a^2\mathbb{E}[X^2] + b^2\mathbb{E}[Y^2]$$

Let $a = \sqrt{\mathbb{E}[Y^2]}$, $b = \sqrt{\mathbb{E}[X^2]}$

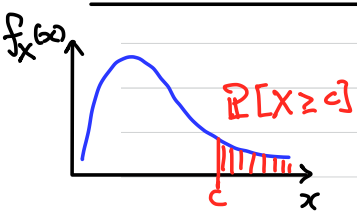
$$\Rightarrow 2\sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}\;\mathbb{E}[XY] \le 2\,\mathbb{E}[X^2]\,\mathbb{E}[Y^2]$$
$$\Rightarrow \mathbb{E}[XY] \le \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$$

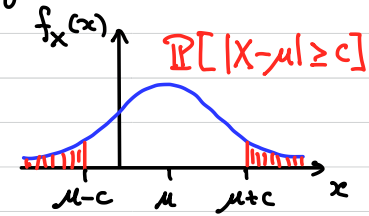
Similarly, $0 \le \mathbb{E}[(aX + bY)^2]$ implies

$$-\sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]} \le \mathbb{E}[XY]$$
$$\Rightarrow |\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$$

Squaring yield the desired result □

Concentration Inequalities (Tail bounds)





Useful for
- proving convergence results
- bounding failure probabilities
- probabilistic bounds on runtimes

Thm (Markov's Inequality)
Let X be a non-negative RV with $\mathbb{E}[|X|] < \infty$. Then, for any constant $c > 0$,

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}.$$

Lemma
X, a non-neg RV $\Rightarrow \forall \omega \in \Omega, \forall c > 0,$

$$X(\omega) \geq c\, I_{\{X \geq c\}}(\omega) \quad (*)$$

Recall Indicator RV
For $A \in \mathcal{F}$, $I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$

$(I_A = 1) = \{\omega \in \Omega \mid I_A(\omega) = 1\} = A$

So, $\mathbb{E}[I_A] = \mathbb{P}[I_A = 1] = \mathbb{P}[A]$

Pf (Lemma)

$(X \geq c) = \{\omega \in \Omega \mid X(\omega) \geq c\}$

- $X(\omega) < c \Rightarrow I_{\{X \geq c\}}(\omega) = 0$
  X non-neg $\Rightarrow X(\omega) \geq 0 \ \forall \omega \in \Omega$
  $\Rightarrow (*)$ holds
- $X(\omega) \geq c \Rightarrow I_{\{X \geq c\}}(\omega) = 1 \Rightarrow (*)$ holds □

Pf (Markov's inequality)

Lemma $\Rightarrow \forall \omega \in \Omega, c > 0, \ X(\omega) \geq c\, I_{\{X \geq c\}}(\omega)$
Taking expectation yields

$$\mathbb{E}[X] \geq \mathbb{E}[c\, I_{\{X \geq c\}}]$$
$$= c\, \mathbb{E}[I_{\{X \geq c\}}]$$
$$= c\, \mathbb{P}[X \geq c]$$

□

Thm (Generalized Markov's Ineq.)
Let $X: \Omega \to \mathbb{R}$ be an arbitrary RV.
Then, for all constants $c > 0$ and $k > 0$,

$$\mathbb{P}[|X| \geq c] \leq \frac{\mathbb{E}[|X|^k]}{c^k}$$

Pf Similar argument as above applied to $|X(\omega)|^k \geq c^k I_{\{|X| \geq c\}}(\omega)$.

Thm (Chebyshev's inequality)
For all RVs X with $\mathbb{E}[X] = \mu < \infty$ and for all constants $c > 0$,

$$\mathbb{P}[|X-\mu| \geq c] \leq \frac{\mathrm{Var}(X)}{c^2}$$

Pf since $|X-\mu| \geq c \iff |X-\mu|^2 \geq c^2$,

$$\mathbb{P}[|X-\mu| \geq c] = \mathbb{P}[|X-\mu|^2 \geq c^2]$$

Makov's Ineq $\Rightarrow \leq \frac{\mathbb{E}[|X-\mu|^2]}{c^2} = \frac{\mathrm{Var}(X)}{c^2}$ □

Thm (Weak Law of Large Numbers)
Let $X_1, X_2, \ldots$ be a sequence of iid RVs with a finite mean μ and finite variance σ^2.
Let $S_n = X_1 + \cdots + X_n$. Then, $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right] = 0.$$

($\frac{S_n}{n}$: Sample average)

Pf $\mathrm{Var}\left(\frac{S_n}{n}\right) = \mathbb{E}\left[\left(\frac{S_n}{n}\right)^2\right] - \mathbb{E}\left[\frac{S_n}{n}\right]^2 = \frac{\mathrm{Var}(S_n)}{n^2}$

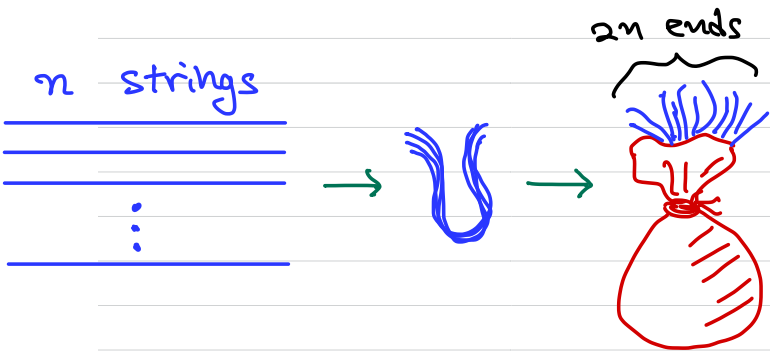
$\mathrm{Var}(X_i) = \sigma^2$ and $X_1, \ldots, X_n \perp\!\!\!\perp \Rightarrow = \frac{\sigma^2}{n}$

Chebyshev's inequality $\Rightarrow$

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty$$ □

Remarks:
1) Finite variance is not required for WLLN.
2) We will later discuss a stronger version of LLN (SLLN).

Problem from Lecture 5



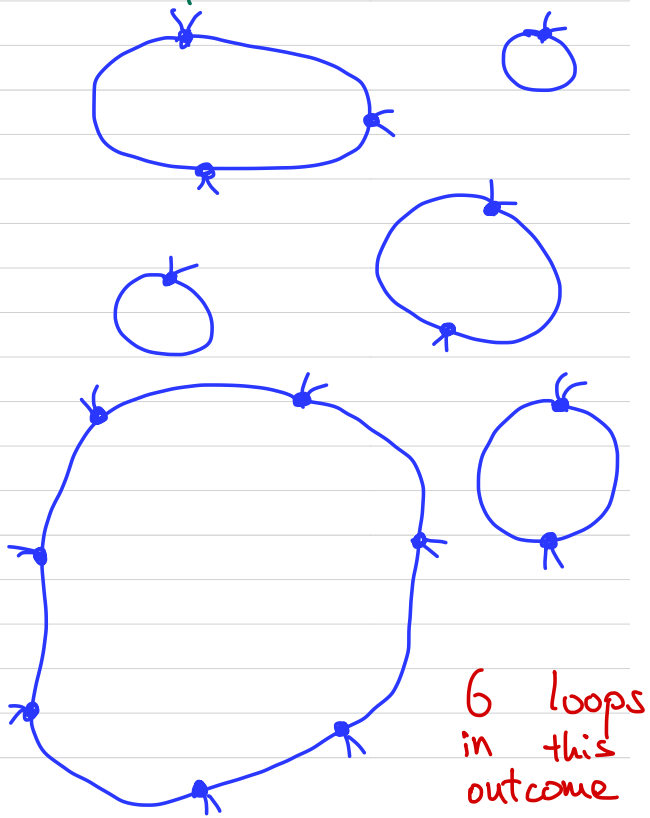
Suppose you tie two free ends u.a.r. and continue until no more free ends remain.

Q What is the expected number of loops formed?

Min = 1

max = n

Example: n=16

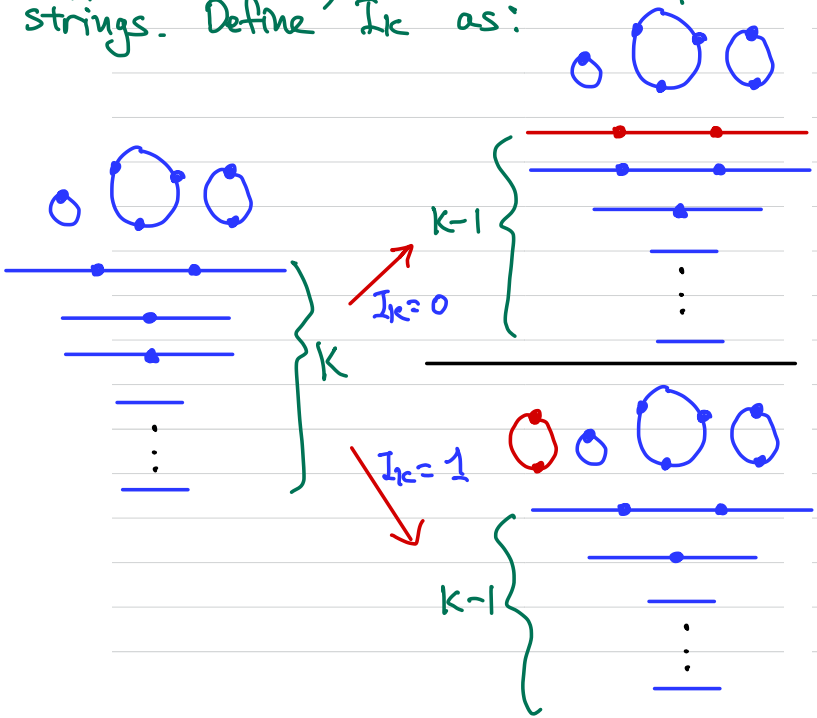


6 loops in this outcome

Solution:

Let L = # loops formed.

Suppose currently there are k open strings. Define I_k as:



$$L = I_n + I_{n-1} + \cdots + I_1$$

$$\mathbb{E}[L] = \sum_{k=1}^{n} \mathbb{E}[I_k]$$

$$= \sum_{k=1}^{n} \mathbb{P}[I_k = 1]$$

$2k$ free ends $\Rightarrow$

$$\mathbb{P}[I_k = 1] = \frac{k}{\binom{2k}{2}} = \frac{1}{2k-1}$$

$$\boxed{\mathbb{E}[L] = \sum_{k=1}^{n} \frac{1}{2k-1}}$$

which is $\sim \frac{1}{2} \ln(n)$ for large n.

Lecture 6

$X_1, X_2, X_3, \ldots$ a sequence iid of RVs with a finite mean $\mu = \mathbb{E}[X_i]$.

LLN states that the sample average $\frac{S_n}{n}$ converges to the expectation $\mathbb{E}[X_i] = \mu$ as the sample size n gets large. (See Demo)

What can we say about the fluctuation (distribution) of $\frac{S_n}{n}$?

Suppose $\mathrm{Var}(X_i) = \sigma^2 < \infty$.

$\mathbb{E}\left(\frac{S_n}{n}\right) = \mu$, $\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$

$$\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right)$$ Standardized

mean 0

variance 1

Thm (Central Limit Theorem)

Let $X_1, X_2, X_3, \ldots$ be a sequence of iid RVs with finite mean μ and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right) \le x\right] = \Phi(x), \quad \forall x \in \mathbb{R}.$$

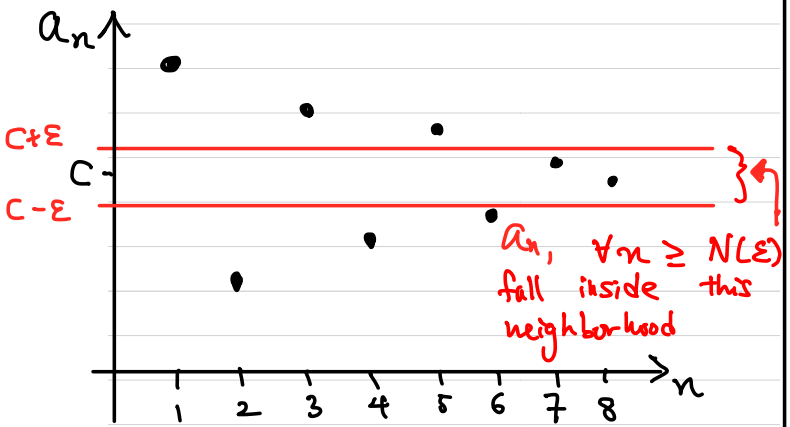
where $\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$ denotes the c.d.f. of $\mathcal{N}(0,1)$

Remark: For large n, $\frac{S_n}{n}$ is well approximated by $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

It is clear what it means for a sequence of real numbers $a_1, a_2, a_3, \ldots$ to converge to a number c:

$$\lim_{n\to\infty} a_n = c$$

For any $\varepsilon > 0$, $\exists$ an integer $N(\varepsilon) > 0$ s.t. $|a_n - c| < \varepsilon$, $\forall n \geq N(\varepsilon)$.



However, a RV $X: \Omega \to \mathbb{R}$ is a function, so we need to define what it means for a sequence of functions to converge.

$X_1, X_2, X_3, \ldots$ a sequence of RVs
X another RV

$\omega \in \Omega$	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$\cdots$	$X_n(\omega)$	$\cdots$	$X(\omega)$
ω_1	a_{11}	a_{12}	a_{13}	$\cdots$	a_{1n}	$\cdots$ $\lvert a_{1n} - c_1 \rvert$	c_1
ω_2	a_{21}	a_{22}	a_{23}	$\cdots$	a_{2n}	$\cdots$ $\lvert a_{2n} - c_2 \rvert$	c_2
ω_3	a_{31}	a_{32}	a_{33}	$\cdots$	a_{3n}	$\cdots$ $\lvert a_{3n} - c_3 \rvert$	c_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_m	a_{m1}	a_{m2}	a_{m3}	$\cdots$	a_{mn}	$\cdots$ $\lvert a_{mn} - c_m \rvert$	c_m
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$F_{X_n}(x) = \mathbb{P}[X_n \leq x]$$

$$\mathbb{P}[|X_n - X| < \varepsilon]$$

$X_1, X_2, X_3, \ldots$ a sequence of RVs } Defined on the same prob. space unless said otherwise
X another RV }

Pointwise Convergence

$\lim_{n\to\infty} X_n(\omega) = X(\omega), \quad \forall\ \omega \in \Omega.$

This notion of convergence turns out to be too strong.

Almost sure convergence. ($X_n \xrightarrow{a.s.} X$ as $n\to\infty$)

(a.k.a. strong convergence, convergence with probability 1 (w.p. 1))

$$\mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = \mathbb{P}\left[\{\omega\in\Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right] = 1$$

$$\left\{\omega\in\Omega \;\middle|\; \begin{array}{l}\text{For every } \varepsilon>0, \exists \text{ an int } N(\omega,\varepsilon) \text{ s.t.} \\ |X_n(\omega)-X|<\varepsilon, \text{ for all } n \geq N(\omega,\varepsilon)\end{array}\right\}$$

$$\mathbb{P}\left[\{\omega\in\Omega \mid \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}\right] = 0$$

"Measure zero"

Example:

Thm (Strong LLN) Let $X_1, X_2, \ldots$ be a sequence of iid RVs with finite mean μ. Then, $\frac{S_n}{n} \xrightarrow{a.s.} \mu$ as $n\to\infty$.

Convergence in probability ($X_n \xrightarrow{p} X$ as $n\to\infty$)

for every $\varepsilon>0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0.$$

$$\{\omega\in\Omega \mid |X_n(\omega) - X(\omega)| > \varepsilon\}$$

Example: WLLN. $\frac{S_n}{n} \xrightarrow{p} \mu$ as $n\to\infty$

Convergence in r^{th}-mean ($X_n \xrightarrow{r} X$ as $n\to\infty$)

For $r>0$,

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0.$$

Convergence in distribution. ($X_n \xrightarrow{d} X$ as $n\to\infty$)

a.k.a. **Convergence in Law**

Here, X does not need to be defined on the same prob space as $X_1, X_2, \ldots$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \forall\ x \in C(F_X),$$

where $C(F_X) = \{x\in\mathbb{R} \mid F_X(x) \text{ is continuous at } x\}$

Example: CLT. $\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right) \xrightarrow{d} X \sim N(0,1)$ as $n\to\infty$.

Equal in Distribution vs Equal in Probability

Take a fair coin and toss it twice. Assume the two tosses are independent.

$\Omega = \{HH, HT, TH, TT\}$

Let X_1, X_2 be RVs defined as

$X_i(\omega) = \begin{cases} 1, & \text{if the } i^{th} \text{ toss shows } H \\ 0, & \text{" " " " " } T \end{cases}$

Then, $X_1 \overset{d}{=} X_2$ (Equal in Law or distribution)

But, $\mathbb{P}[X_1 = X_2] = \mathbb{P}[\{HH, TT\}]$

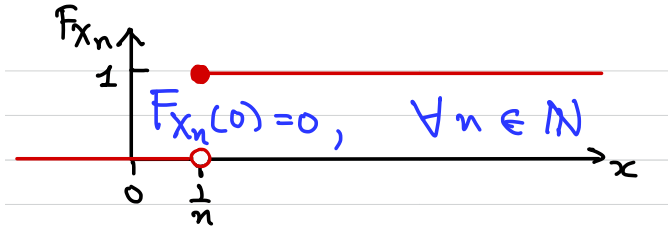
$= \frac{1}{2}$

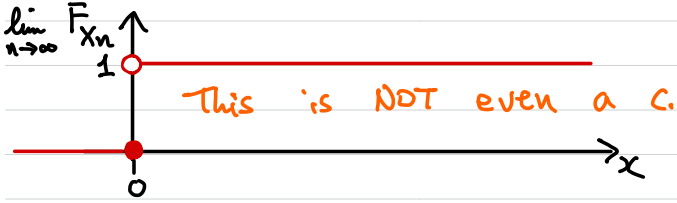
So X_1 and X_2 are not equal in probability.

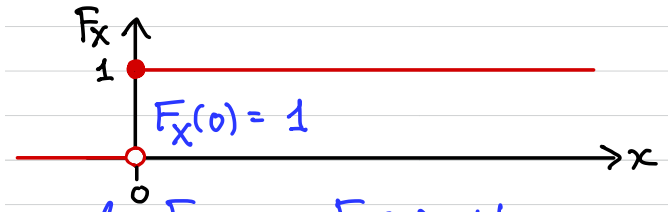
Example: Convergence in distribution

Let $X_1, X_2, X_3, \ldots$ be RVs with $\mathbb{P}[X_n = \frac{1}{n}] = 1$

and X a RV with $\mathbb{P}[X = 0] = 1$.







$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \forall \; x \neq 0$

$\lim_{n\to\infty} F_{X_n}(x) \neq F_X(x)$ at $x = 0$

But this does not affect $\xrightarrow{d}$ since $F_X(x)$ is not continuous at $x = 0$.

$X_n \xrightarrow{d} X$ as $n \to \infty$

Thm (Relations between different Convergence concepts)

$$(X_n \xrightarrow{a.s.} X) \underset{(3)}{\Rightarrow} (X_n \xrightarrow{P} X) \underset{(4)}{\Rightarrow} (X_n \xrightarrow{d} X)$$

$$\Uparrow (2)$$

$$(X_n \xrightarrow{s} X) \underset{(1)}{\Rightarrow} (X_n \xrightarrow{r} X), \quad 0 < r < s$$

Pf Next lecture.

(Converses DO NOT hold in general.)

Example:

$(X_n \xrightarrow{P} X)$ does not imply $(X_n \xrightarrow{a.s.} X)$

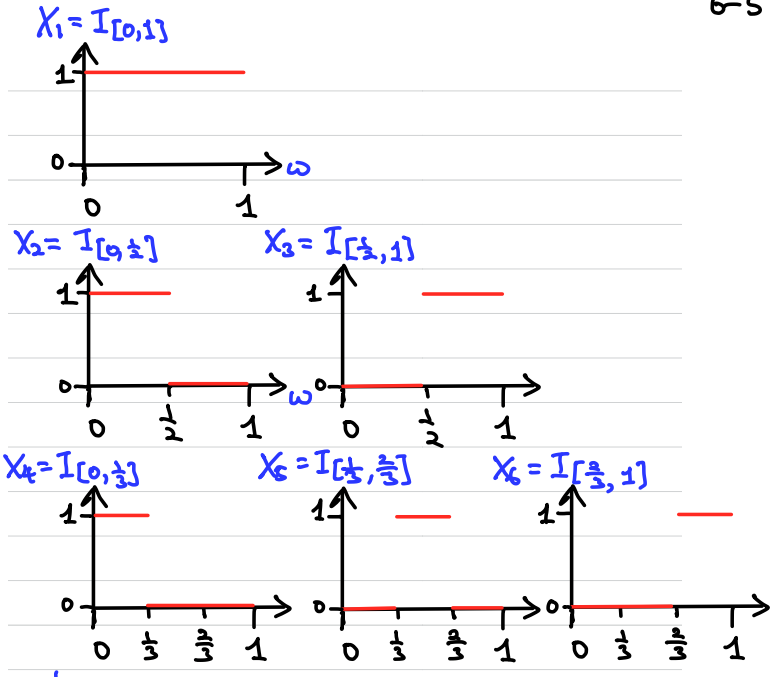
Suppose $\Omega = [0,1]$

$$P([a,b]) = b-a, \quad \forall\, 0 \le a < b \le 1$$

Let $X(\omega) = 0$ and define $X_1, X_2, X_3, \ldots$ as shown on the right.

For any $\varepsilon > 0$, $\lim_{n\to\infty} P[|X_n - X| > \varepsilon] = 0$.

$X_n \xrightarrow{P} X$ as $n \to \infty$



and so on.

But for any $\omega \in \Omega$, there exist infinitely many values of n for which $X_n(\omega) = 1$.

$\Rightarrow \lim_{n\to\infty} X_n(\omega) \neq X(\omega), \quad \forall\, \omega \in \Omega$

Almost sure convergence does not hold.

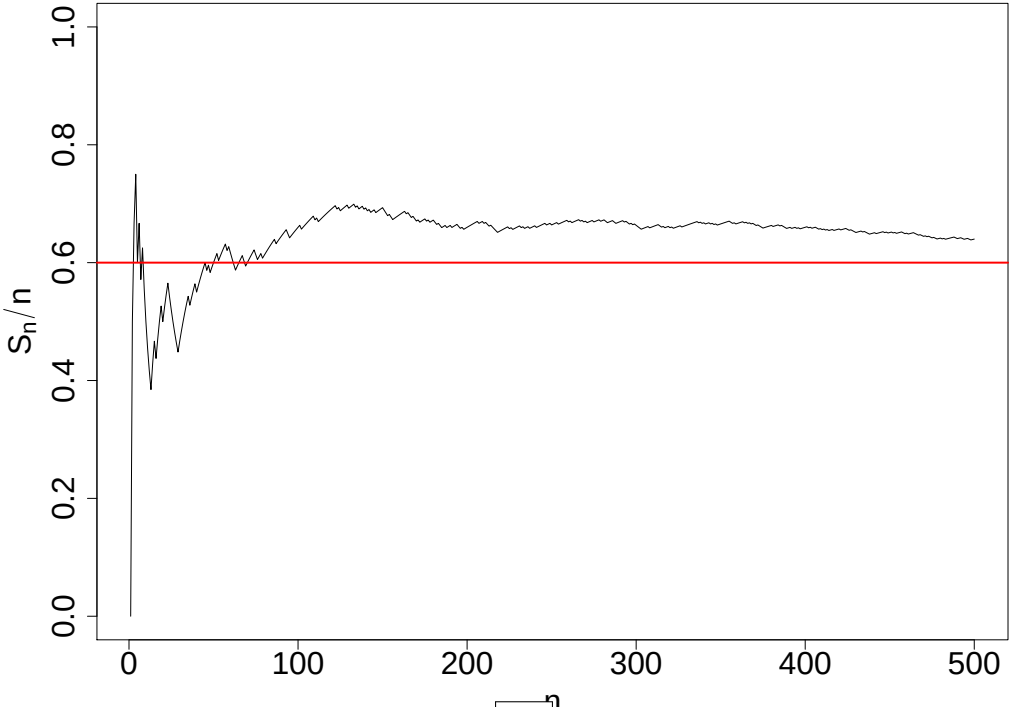
Stat 201A Fall 2024

Lecture 6
LLN Demo

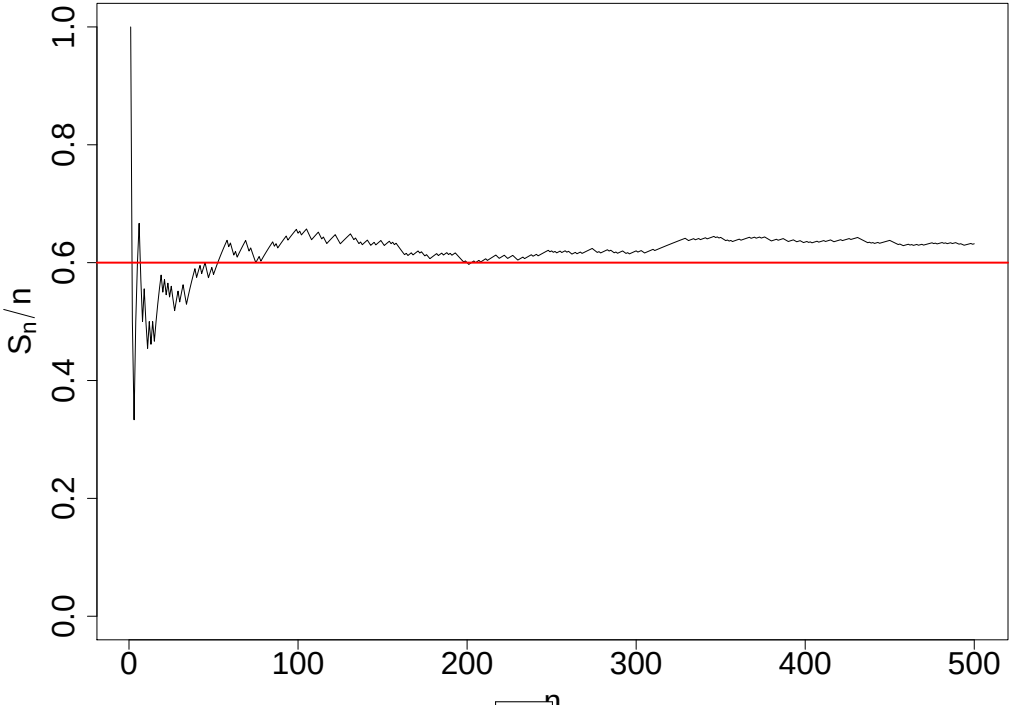
September 17, 2024

- Let $I_1, \ldots, I_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$.
- $\mathbb{E}[I_i] = p$ for all $i = 1, \ldots, n$.
- Let $S_n = I_1 + I_2 + \cdots + I_n$.
- $\mathbb{E}[S_n] = np$, so $\mathbb{E}[\frac{S_n}{n}] = p$.

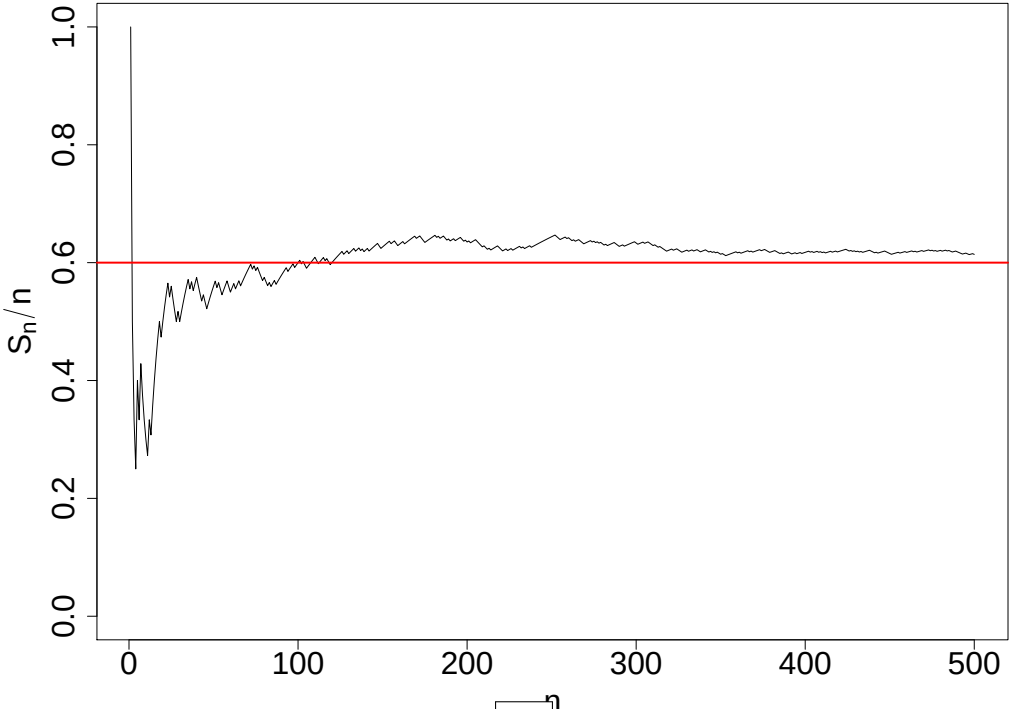
Sample path of S_n/n as a function of n for $p = 0.6$



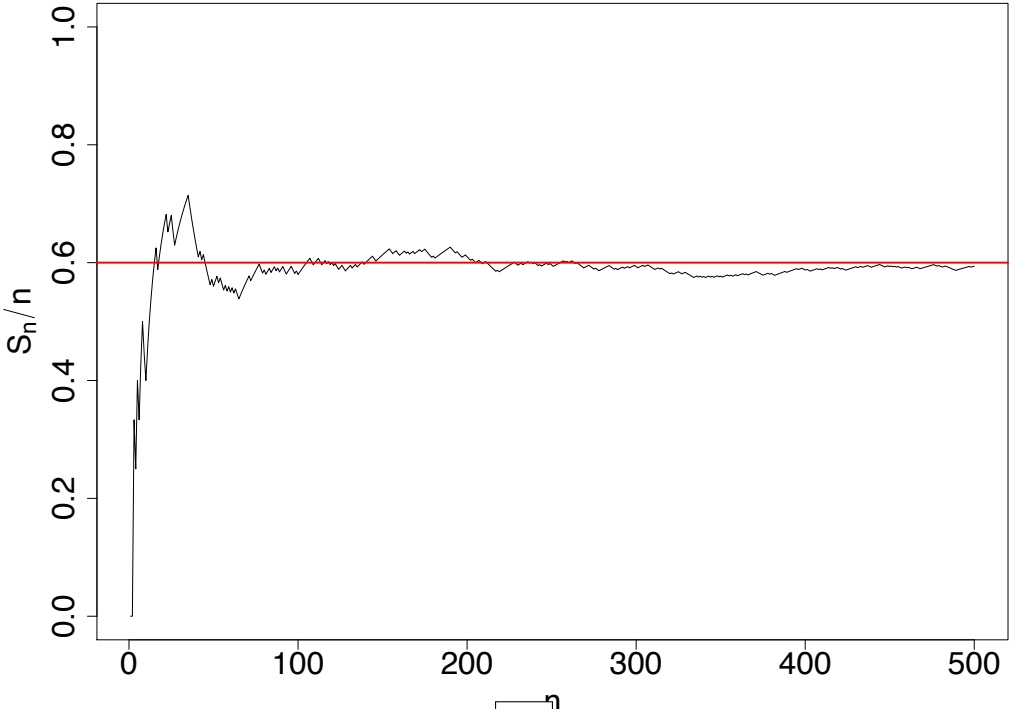
Sample path of S_n/n as a function of n for $p = 0.6$



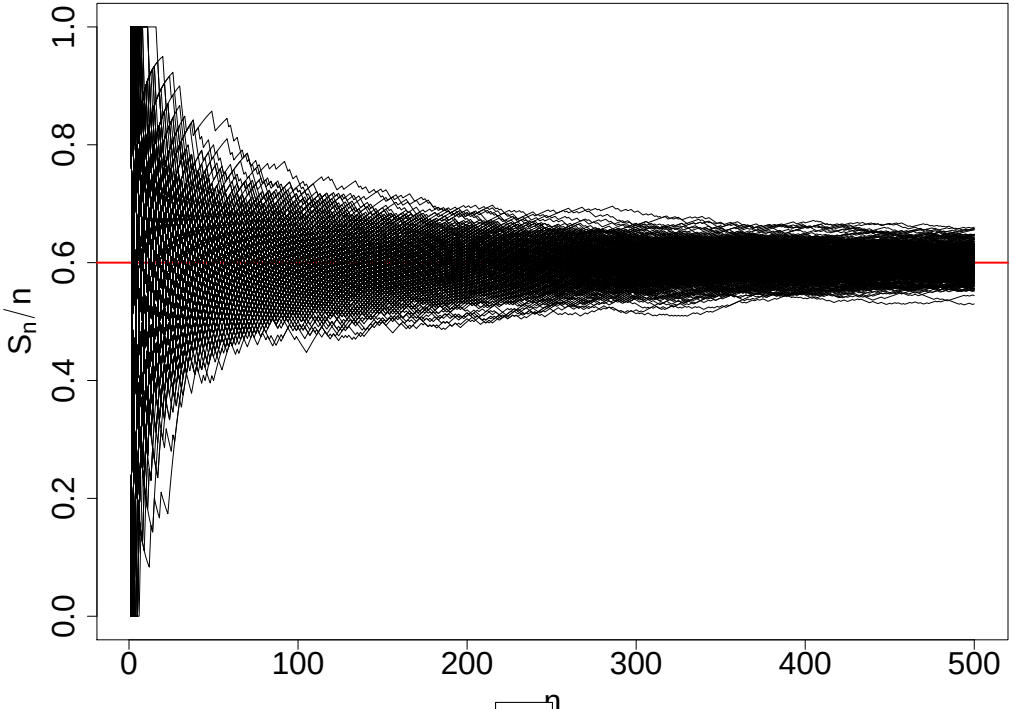
Sample path of S_n/n as a function of n for $p = 0.6$



Sample path of S_n/n as a function of n for $p = 0.6$

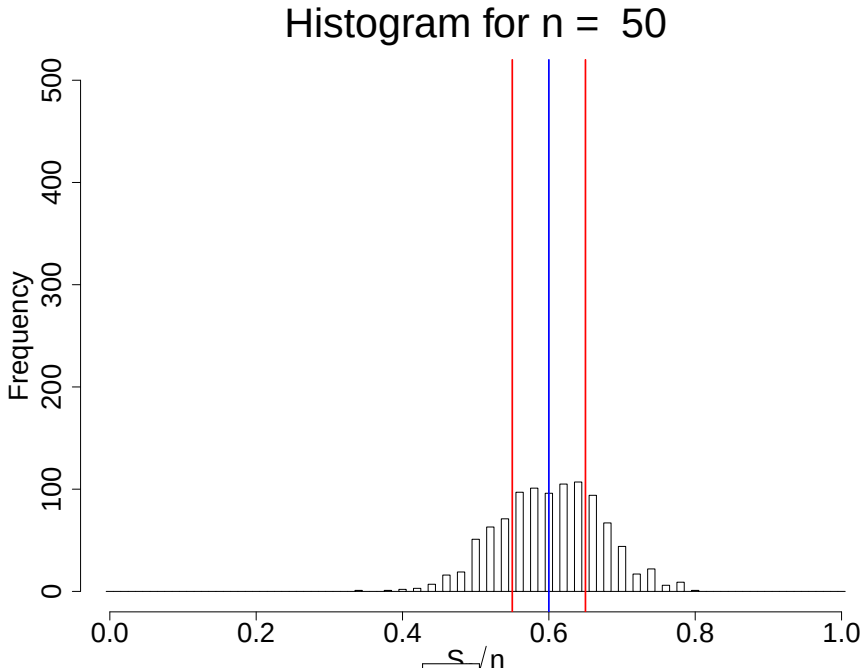


500 independent sample paths of S_n/n for $p = 0.6$



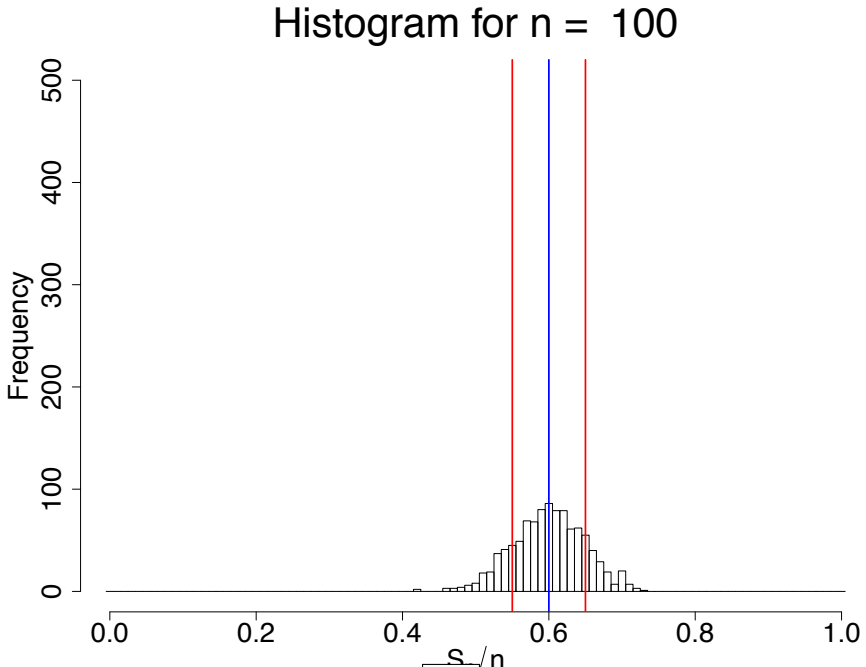
Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



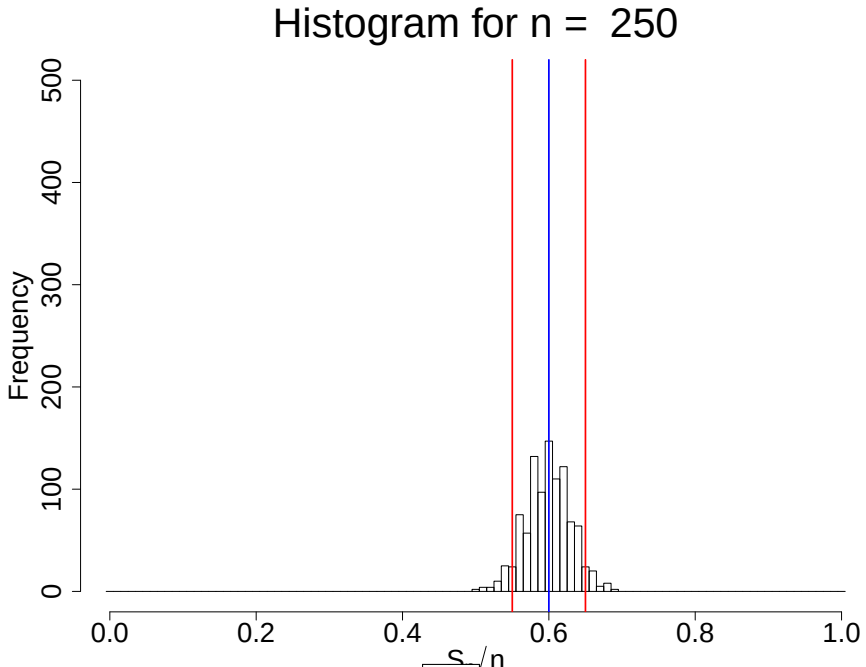
Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



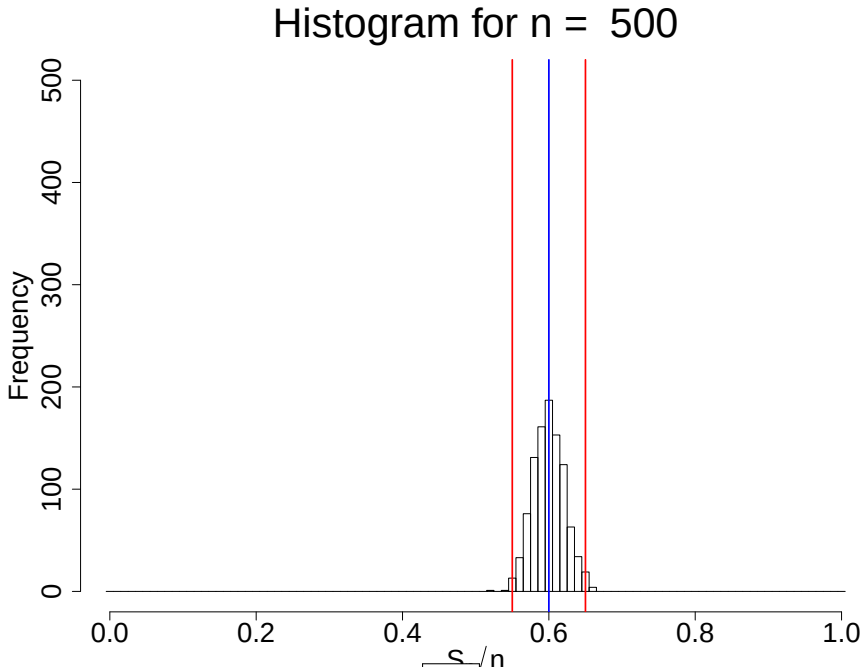
Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



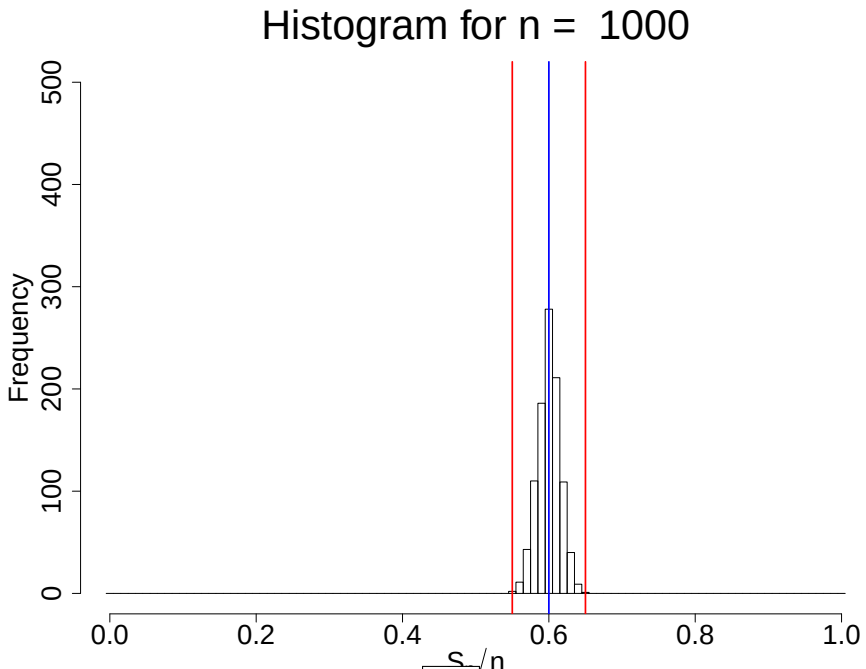
Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



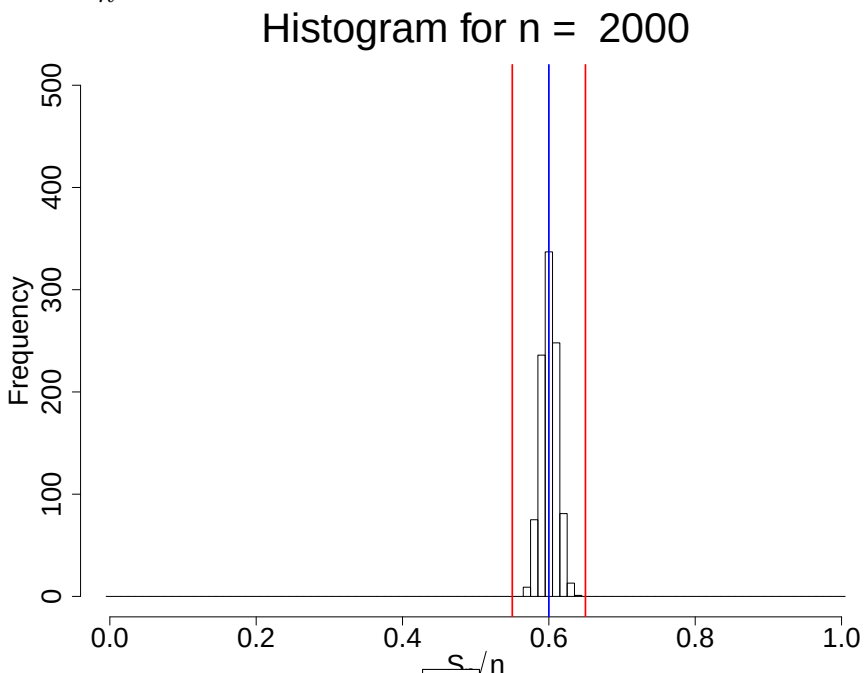
Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



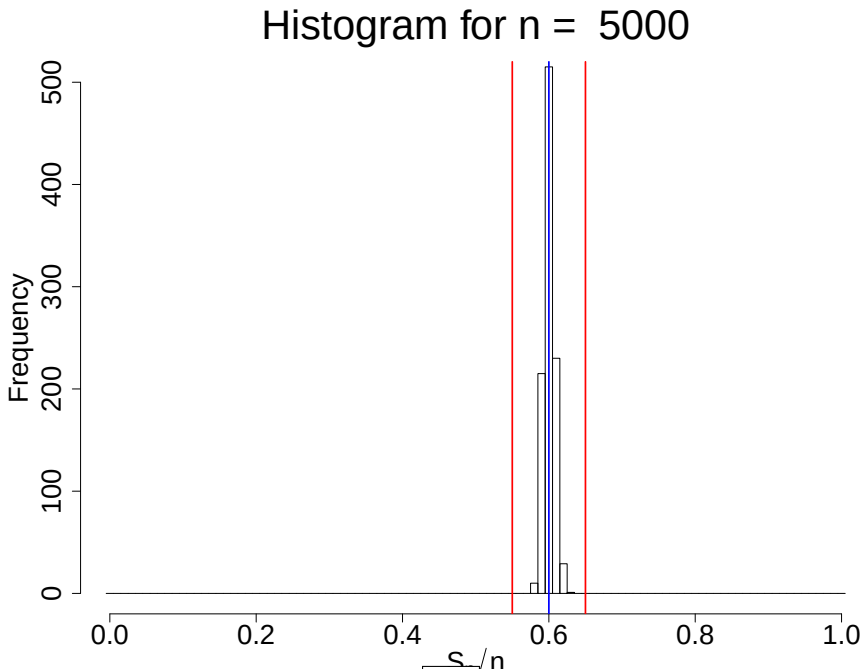
Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



Law of Large Numbers in action for $p = 0.6$

Red lines correspond to $\frac{S_n}{n} = p \pm 0.05$.



Thm

$(X_n \xrightarrow{a.s.} X) \underset{(3)}{\Rightarrow} (X_n \xrightarrow{p} X) \underset{(4)}{\Rightarrow} (X_n \xrightarrow{d} X)$

$\Uparrow$ (2)

$(X_n \xrightarrow{s} X) \underset{(1)}{\Rightarrow} (X_n \xrightarrow{r} X), \quad 0 < r < s$

(No other implications in general.)

Lemma 1 For $0 < r < s$, any RV Y satisfies

$$(\mathbb{E}[|Y|^r])^{\frac{1}{r}} \le (\mathbb{E}[|Y|^s])^{\frac{1}{s}}$$

Pf of Lemma 1:

Let $g(x) = |x|^{\frac{s}{r}}$. This function is cvx since $|x|^a$ is cvx for $a \ge 1$.

Jensen's Inequality $\Rightarrow$

$$g(\mathbb{E}[|Y|^r]) \le \mathbb{E}[g(|Y|^r)]$$

$$\Rightarrow (\mathbb{E}[|Y|^r])^{s/r} \le \mathbb{E}[|Y|^s]$$

$$\Rightarrow (\mathbb{E}[|Y|^r])^{\frac{1}{r}} \le (\mathbb{E}[|Y|^s])^{\frac{1}{s}}$$

□

Pf of (1):

Lemma 1 ⇓

$$0 \le \mathbb{E}[|X_n - X|^r] \le \mathbb{E}[|X_n - X|^s]^{\frac{r}{s}}$$

↑ since $|X_n - X|^r$ is non-negative.

So, if $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^s] = 0$ for RHS,

then $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0$. □

Pf of (2):

For any $\varepsilon > 0$ and $r > 0$,

Generalized Markov's Inequality (Note 5-6)

$$0 \le \mathbb{P}[|X_n - X| \ge \varepsilon] \le \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r}.$$

So, $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0$ for RHS

$$\Rightarrow \lim_{n\to\infty} \mathbb{P}[|X_n - X| \ge \varepsilon] = 0$$

□

Lemma 2

$X_n \xrightarrow{a.s.} X$ as $n \to \infty$

$\Updownarrow$ if and only if

for every $\varepsilon > 0$,

$\lim_{n\to\infty} \mathbb{P}\left(|X_m - X| < \varepsilon \text{ for all } m \geq n\right) = 1.$

Pf of Lemma 2 (Optional reading)

$$\left(\lim_{n\to\infty} X_n = X\right) = \left\{\omega \in \Omega \,\middle|\, \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}$$

$$= \left\{\omega \in \Omega \,\middle|\, \begin{array}{l}\text{for every } \varepsilon > 0, \exists \text{ an } N(\omega,\varepsilon), \text{ s.t.} \\ |X_m(\omega) - X(\omega)| < \varepsilon, \ \forall\ m \geq N(\omega,\varepsilon)\end{array}\right\}$$

$$= \bigcap_{\varepsilon>0} \bigcup_{n=1}^{\infty} \underbrace{\left\{\omega \in \Omega \,\middle|\, |X_m(\omega) - X(\omega)| < \varepsilon, \ \forall\ m \geq n\right\}}_{\text{denote this set by } A_{n,\varepsilon}}$$

$\left(X_n \xrightarrow{a.s.} X \text{ as } n \to \infty\right)$

$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1 \iff \mathbb{P}\left(\bigcap_{\varepsilon>0} \bigcup_{n=1}^{\infty} A_{n,\varepsilon}\right) = 1$

$\iff \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_{n,\varepsilon}\right) = 1, \ \forall \varepsilon > 0$

Lastly, $\forall \varepsilon > 0$

$A_{1,\varepsilon} \subset A_{2,\varepsilon} \subset \cdots \subset \bigcup_{n=1}^{\infty} A_{n,\varepsilon}$

$\Rightarrow \lim_{n\to\infty} \mathbb{P}(A_{n,\varepsilon}) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_{n,\varepsilon}\right)$

This completes the proof. □

Proof of (3):

Suppose $X_n \xrightarrow{a.s.} X$ as $n \to \infty$

$1 \geq \mathbb{P}[|X_n - X| < \varepsilon]$

$\geq \mathbb{P}[|X_m - X| < \varepsilon, \forall\ m \geq n] \to 1$ as $n \to \infty$ (By Lemma 2)

since $\{|X_m - X| < \varepsilon, \forall\ m \geq n\} = \bigcap_{m=n}^{\infty} \{|X_m - X| < \varepsilon\}$

and $\mathbb{P}(A) \geq \mathbb{P}(B)$ if $B \subseteq A$.

$\Rightarrow \lim_{n\to\infty} \mathbb{P}[|X_n - X| < \varepsilon] = 1.$ □

Proof of (4) (Optional reading)

Assume $X_n \xrightarrow{P} X$ as $n \to \infty$

i) $X_n(\omega)$ falls here

If $X_n(\omega) \le x$, then either
$X(\omega) \le x + \varepsilon$ or $|X(\omega) - X_n(\omega)| > \varepsilon$

$\{\omega \in \Omega \mid X_n(\omega) \le x\} \subseteq$
$\{\omega \in \Omega \mid X(\omega) \le x + \varepsilon\} \cup \{\omega \in \Omega \mid |X(\omega) - X_n(\omega)| > \varepsilon\}$

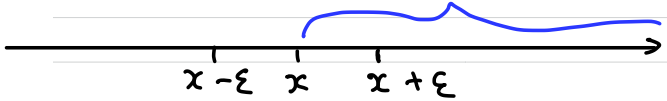
Union Bound $\Rightarrow$

$\mathbb{P}[X_n \le x] \le \mathbb{P}[X \le x + \varepsilon] + \mathbb{P}[|X - X_n| > \varepsilon]$

$F_{X_n}(x) \le F_X(x + \varepsilon) + \mathbb{P}[|X - X_n| > \varepsilon]$

ii) Similarly,

$X_n(\omega)$ falls here



If $X_n(\omega) > x$, then either
$X(\omega) > x - \varepsilon$ or $|X(\omega) - X_n(\omega)| > \varepsilon$

$\Rightarrow \mathbb{P}[X_n > x] \le \mathbb{P}[X > x - \varepsilon] + \mathbb{P}[|X - X_n| > \varepsilon]$

$1 - F_{X_n}(x) \le 1 - F_X(x - \varepsilon) + \mathbb{P}[|X - X_n| > \varepsilon]$

$X_n \xrightarrow{P} X$ as $n \to \infty$ $\Rightarrow$

$\lim_{n \to \infty} \mathbb{P}[|X - X_n| > \varepsilon] = 0, \quad \forall\, \varepsilon > 0$

So, (i) & (ii) together imply

$F_X(x - \varepsilon) \le \liminf_{n \to \infty} F_{X_n}(x)$

$\le \limsup_{n \to \infty} F_{X_n}(x)$

$\le F_X(x + \varepsilon), \quad \forall\, \varepsilon > 0$

If $x \in C(F_X)$, then

$\lim_{\varepsilon \to 0} F_X(x - \varepsilon) = F_X(x) = \lim_{\varepsilon \to 0} F_X(x + \varepsilon)$

$\Rightarrow \lim_{n \to \infty} F_{X_n}(x) = F_X(x)$ if $x \in C(F_X)$ □

Def (Moment Generating Function) The MGF of a RV X is a function $M_X: \mathbb{R} \to [0, \infty)$ given by

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}.$$

Note: If $X_1, \dots, X_n$ are $\perp\!\!\!\perp$ RVs, and $S_n = X_1 + \cdots + X_n$, then

$$M_{S_n}(t) = \prod_{k=1}^{n} M_{X_k}(t).$$

Thm If $M_X(t) < \infty$, $\forall\, t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then

1) $\left.\frac{d^k}{dt^k} M_X(t)\right|_{t=0} = \mathbb{E}[X^k]$, for all $k = 1, 2, 3, \dots$

2) For t within the radius of convergence,

$$M_X(t) = \sum_{k=0}^{\infty} \mathbb{E}[X^k] \frac{t^k}{k!}.$$

Thm (Uniqueness) Suppose X, Y are two RVs with well defined MGFs. If $M_X(t) = M_Y(t)$, $\forall\, t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then

$$X \overset{d}{=} Y.$$

Thm (Convergence in distribution) Suppose $X_1, X_2, \dots$ is a sequence of RVs with MGF $M_{X_n}(t)$ $\forall\, n$ well defined for $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$.

If $M_{X_n}(t) \longrightarrow M(t)$ as $n \to \infty$ for every $t \in (-\varepsilon, \varepsilon)$, then $M(t) = M_X(t)$, where M_X is the MGF of a RV s.t.

$$X_n \xrightarrow{d} X \text{ as } n \to \infty.$$

Application 1

Let $G \sim \text{Geometric}(p)$

$\mathbb{P}[G = k] = (1-p)^{k-1} p, \quad k = 1, 2, 3, \ldots$

$\mathbb{E}[G] = \frac{1}{p}$

$M_G(t) = \mathbb{E}[e^{tG}] = \sum_{k=1}^{\infty} e^{tk} (1-p)^{k-1} p$

$= p e^t \sum_{k=1}^{\infty} [e^t (1-p)]^{k-1}$

For $|x| < 1$,

$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \Rightarrow$

$= \frac{p e^t}{1 - (1-p) e^t}$

Let $G_n \sim \text{Geometric}\left(\frac{\lambda}{n}\right)$, where $\lambda > 0$

$n = 1, 2, 3, \ldots$

Let $X_n = \frac{G_n}{n}$. Then $\mathbb{E}[X_n] = \frac{1}{\lambda}$

$M_{X_n}(t) = \mathbb{E}[e^{t X_n}] = \mathbb{E}[e^{\frac{t}{n} G_n}]$

$= \frac{\frac{\lambda}{n} e^{t/n}}{1 - (1 - \frac{\lambda}{n}) e^{t/n}} \to \frac{\lambda}{\lambda - t}$ as $n \to \infty$

Let $X \sim \text{Exp}(\lambda)$, $\lambda > 0$.

pdf: $f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

$\mathbb{E}[X] = 1/\lambda$.

$M_X(t) = \mathbb{E}[e^{tX}] = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx$

$= \frac{\lambda}{\lambda - t}$.

So, $X_n \xrightarrow{d} X$ as $n \to \infty$.

Application 2

Recall Bernoulli process from Lecture 3

$\mathbb{P}(\text{Success}) = p$, $\mathbb{P}(\text{Failure}) = 1-p$

T_r = total # trials until the r^{th} success

F_r = " " failures " the r^{th} success

$F_r + r = T_r$

$F_r \sim NB(r,p)$

$M_{F_r}(t)$?

$T_r = W_1 + W_2 + \cdots + W_r$, where

$W_1, \ldots, W_r \overset{iid}{\sim}$ Geometric(p)

$$\Rightarrow M_{T_r}(t) = \left[\frac{p\,e^t}{1-(1-p)e^t}\right]^r$$

$$M_{F_r}(t) = \mathbb{E}[e^{tF_r}] = \mathbb{E}[e^{t(T_r - r)}]$$

$$= e^{-tr}\,\mathbb{E}[e^{tT_r}]$$

$$= \left[\frac{p}{1-(1-p)e^t}\right]^r$$

Let $X_n = \frac{F_{r,n}}{n}$, where $F_{r,n} \sim NB(r, \frac{\lambda}{n})$

$$M_{X_n}(t) = \mathbb{E}\left[e^{t\frac{F_{r,n}}{n}}\right] = \left[\frac{\frac{\lambda}{n}}{1-(1-\frac{\lambda}{n})e^{t/n}}\right]^r$$

$$M_{X_n}(t) \longrightarrow \left(\frac{\lambda}{\lambda - t}\right)^r \quad \text{as } n \to \infty$$

Let $X = Y_1 + Y_2 + \cdots + Y_r$, where

$Y_1, \ldots, Y_r \overset{iid}{\sim} Exp(\lambda)$.

Then, $X \sim Gamma(r, \lambda)$ (shape r, rate λ)

pdf $f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$

$$M_X(t) = \prod_{i=1}^{n} M_{Y_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^r$$

$\Rightarrow X_n \xrightarrow{d} X$ as $n \to \infty$

Sums of Random Variables

Let X, Y be discrete RVs on the same probability space.

Q: Find $\mathbb{P}[X+Y=c]$.

$$\mathbb{P}[X+Y=c] = \sum_{(a,b):\, a+b=c} \mathbb{P}[X=a, Y=b]$$

$$= \sum_{a} \mathbb{P}[X=a, Y=c-a]$$

If $X \perp\!\!\!\perp Y$, then

$$\mathbb{P}[X+Y=c] = \sum_{a} \mathbb{P}[X=a]\,\mathbb{P}[Y=c-a]$$

This is called **convolution.**

e.g. $X_i \sim \text{Geometric}(p_i)$, $i=1,2$, $X_1 \perp\!\!\!\perp X_2$, $p_1 \neq p_2$

For $c = 2, 3, 4, \ldots$

$$\mathbb{P}[X_1+X_2=c] = \sum_{a=1}^{c-1} \mathbb{P}[X_1=a]\,\mathbb{P}[X_2=c-a]$$

$$= \sum_{a=1}^{c-1} (1-p_1)^{a-1} p_1 (1-p_2)^{c-a-1} p_2$$

$$= \left[\frac{p_2}{p_2-p_1}\right] p_1 (1-p_1)^{c-1} + \left[\frac{p_1}{p_1-p_2}\right] p_2 (1-p_2)^{c-1}$$

Convolution for continuous RVs.

X, Y continuous RVs:

p.d.f. $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx$

e.g. For $i=1,2$, $X_i \sim \text{Exp}(\lambda_i)$, $\lambda_i > 0$, $\lambda_1 \neq \lambda_2$, $X_1 \perp\!\!\!\perp X_2$

$$f_{X_i}(x) = \begin{cases} \lambda_i e^{-\lambda_i x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

For $z > 0$,

$$f_{X_1+X_2}(z) = \int_{-\infty}^{\infty} f_{X_1}(x)\, f_{X_2}(z-x)\, dx$$

$$= \int_0^z f_{X_1}(x)\, f_{X_2}(z-x)\, dx$$

$$= \int_0^z \lambda_1 e^{-\lambda_1 x}\, \lambda_2 e^{-\lambda_2 (z-x)}\, dx$$

$$= \lambda_1 \lambda_2 e^{-\lambda_2 z} \left. \frac{e^{-x(\lambda_1-\lambda_2)}}{-(\lambda_1-\lambda_2)} \right|_0^z$$

$$= \left[\frac{\lambda_2}{\lambda_2-\lambda_1}\right] \lambda_1 e^{-\lambda_1 z} + \left[\frac{\lambda_1}{\lambda_1-\lambda_2}\right] \lambda_2 e^{-\lambda_2 z}$$

Empirical Distribution

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, where $F(x) = \mathbb{P}[X \le x]$ is an unknown c.d.f.

Goal: Estimate $F: \mathbb{R} \to [0,1]$ (Deterministic function, not a RV)

A natural estimator is the empirical distribution

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x)$$

This is a function of RVs $X_1, \ldots, X_n$

$\hat{F}_n: \mathbb{R} \times \Omega \to [0,1]$

where, for $\omega \in \Omega$,

$$I(X_i \le x)(\omega) = \begin{cases} 1, & \text{if } X_i(\omega) \le x, \\ 0, & \text{otherwise.} \end{cases}$$

Remarks:

1) $X_i(\omega) \le x \iff \omega \in X_i^{-1}(-\infty, x]$

2) $\mathbb{E}[I(X_i \le x)] = \mathbb{P}[I(X_i \le x) = 1]$
$= \mathbb{P}[\{\omega \in \Omega \mid X_i(\omega) \le x\}]$
$= \mathbb{P}[X_i \le x]$
$= F(x) \le 1.$

So, SLLN $\Rightarrow$ for each $x \in \mathbb{R}$,

$\hat{F}_n(x) \xrightarrow{a.s.} F(x)$ as $n \to \infty$

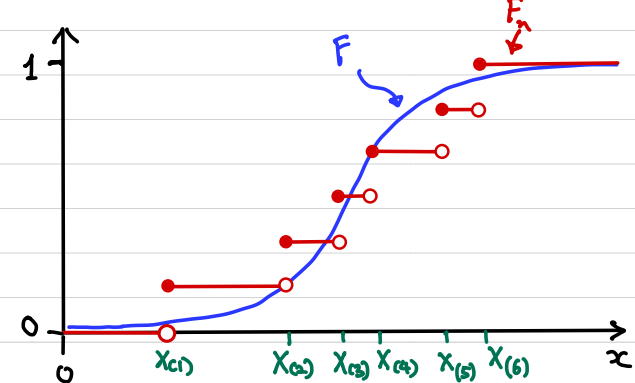
RV — Number

For every given $x \in \mathbb{R}$,

$$\mathbb{P}\left[\lim_{n\to\infty} \hat{F}_n(x) = F(x)\right] = 1$$

Sorted $X_1, \ldots, X_n$: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$

order statistics.



As the sample size n increases, one obtains a more accurate estimate of $F(x)$ for every $x \in \mathbb{R}$.

But as a function of x, how well does $\hat{F}_n$ approximate F?

e.g. (Pointwise but not uniform convergence)

Let $f_n : [0,1] \to \mathbb{R}$, $f_n(x) = x^n$

$$\lim_{n\to\infty} f_n(x) = f(x) = \begin{cases} 1, & x = 1, \\ 0, & x \in [0,1). \end{cases}$$

$\Rightarrow$ f_n is cts on $[0,1]$ $\forall n$, but the pointwise limit f is not cts.

$|f_n(x) - f(x)| = x^n$

For a given $0 < \varepsilon < 1$, there does not exist a positive integer n s.t. $x^n < \varepsilon$ for all $x \in [0,1)$

x^n converges to 0 at an arbitrarily slow rate for x close to 1.

Def (Supremum) Let $S \subset \mathbb{R}$.

Supremum of S = least upper bound of S in $\mathbb{R}$.

$\sup(S)$

If $\sup(S) \in S$, then $\sup(S) = \max(S)$.

e.g. $S = \{x \in \mathbb{R} \mid |x| < 1\}$.

$\max(S)$ does not exist.

$\sup(S) = 1$.

For every $x \in \mathbb{R}$,

$$\mathbb{P}\left[\lim_{n\to\infty} \hat{F}_n(x) = F(x)\right] = 1$$

$$\left\{\omega \in \Omega \;\middle|\; \begin{array}{l} \text{For every } \varepsilon > 0, \exists \text{ an integer } N(x,\omega,\varepsilon) \\ \text{s.t. } n \geq N(x,\omega,\varepsilon) \Rightarrow |\hat{F}_n(x,\omega) - F(x)| < \varepsilon \end{array}\right\}$$

Point-wise convergence

In fact, one can obtain even a stronger result:

Thm (Glivenko-Cantelli Theorem)

Suppose $X_1, \dots, X_n \overset{iid}{\sim} F$. Then

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

In other words,

$$\mathbb{P}\left[\lim_{n\to\infty} \sup_x |\hat{F}_n(x) - F(x)| = 0\right] = 1.$$

$$\left\{\omega \in \Omega \;\middle|\; \begin{array}{l} \text{For every } \varepsilon > 0, \exists \text{ an integer } N(\omega,\varepsilon) \text{ s.t.} \\ n \geq N(\omega,\varepsilon) \Rightarrow |\hat{F}_n(x,\omega) - F(x)| < \varepsilon, \forall x \in \mathbb{R} \end{array}\right\}$$

Uniform convergence

Pf Later in the course.

Chernoff Inequalities: A one-parameter family of bounds derived as follows:

1) For any $t>0$ and $c \in \mathbb{R}$,

$$\mathbb{P}[X \geq c] = \mathbb{P}[tX \geq tc] \quad \text{since exp is monotonically increasing}$$
$$= \mathbb{P}[e^{tX} \geq e^{tc}]$$
$$\leq \frac{\mathbb{E}[e^{tX}]}{e^{tc}} \quad \text{By Markov's ineq.}$$

Right tail

$$\Rightarrow \mathbb{P}[X \geq c] \leq \min_{t>0} \frac{M_X(t)}{e^{tc}}$$

2) For any $t<0$ and $c \in \mathbb{R}$,

$$\mathbb{P}[X \leq c] = \mathbb{P}[tX \geq tc] \leq \frac{M_X(t)}{e^{tc}}$$

Left tail

$$\Rightarrow \mathbb{P}[X \leq c] \leq \min_{t<0} \frac{M_X(t)}{e^{tc}}$$

These lead to exponentially decreasing tail bounds.

---

e.g. $X \sim \text{Binomial}(n,p) \Rightarrow \mathbb{E}[X] = np$

$I \sim \text{Bernoulli}(p)$

$$M_X(t) = [M_I(t)]^n = [e^t p + (1-p)]^n$$

For $t>0$, $\mathbb{P}[X \geq an] \leq \dfrac{[e^t p + (1-p)]^n}{e^{ant}}$

$$\frac{d}{dt} RHS(t) = \frac{n[e^t p + (1-p)]^{n-1}[e^t p]}{e^{tan}} - \frac{an[e^t p + (1-p)]^n}{e^{ant}}$$

$$\frac{d\,RHS(t)}{dt} = 0 \Rightarrow n e^t p = an(e^t p + (1-p))$$
$$\Rightarrow e^t = \frac{a(1-p)}{p(1-a)}$$

Can check that this minimizes RHS(t)

$$\mathbb{P}[X \geq an] \leq \left[\frac{1-p}{1-a}\right]^{(1-a)n} \left(\frac{p}{a}\right)^{an}$$

For $p=\frac{1}{2}$ & $a=\frac{3}{4}$, $\mathbb{P}[X \geq \frac{3}{4}n] \leq \left(\frac{16}{27}\right)^{\frac{n}{4}}$

This bound is much stronger than

Markov's ineq: $\mathbb{P}[X \geq \frac{3}{4}n] \leq \frac{\frac{1}{2}n}{(\frac{3}{4}n)} = \frac{2}{3}$

Chebyshev's ineq: $\mathbb{P}[X \geq \frac{3}{4}n] = \mathbb{P}[X - \frac{1}{2}n \geq \frac{1}{4}n]$

$$\leq \mathbb{P}[|X - \tfrac{n}{2}| \geq \tfrac{1}{4}n] \leq \frac{\frac{n}{4}}{(\frac{n}{4})^2} = \frac{4}{n}$$

Hoeffding's Inequality: Bounded RVs

Let $X_1, \dots, X_n$ be ind RVs with $\mathbb{E}[X_i] = \mu_i < \infty$ and

$$\mathbb{P}[a_i \le X_i \le b_i] = 1$$

for some constants $a_i, b_i \in \mathbb{R}$.

Let $S_n = X_1 + \cdots + X_n$. Then

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \ge \varepsilon\right] \le 2\exp\left[\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right].$$

Pf See next lecture.

Hoeffding's Inequality: Bounded RVs

Let $X_1, \dots, X_n$ be ⫫ RVs with $\mathbb{E}[X_i] = \mu_i < \infty$ and $\mathbb{P}[a_i \le X_i \le b_i] = 1$ for some constants $a_i, b_i \in \mathbb{R}$.
Let $S_n = X_1 + \cdots + X_n$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \ge \varepsilon\right] \le 2 \exp\left[\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right].$$

Equivalently,

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right| \ge \varepsilon\right] \le 2 \exp\left[\frac{-2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right]$$

We will prove this general result.

First, recall Normal RV $X \sim N(\mu, \sigma^2)$

p.d.f. $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

$\forall t \in \mathbb{R}$,

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{tx} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

$$tx - \frac{(x-\mu)^2}{2\sigma^2} = -\frac{(x - \mu - t\sigma^2)^2}{2\sigma^2} + t\mu + \frac{t^2\sigma^2}{2}$$

$$\Rightarrow M_X(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}$$

Def (Sub-Gaussian) A RV X with $\mathbb{E}[X] = \mu < \infty$ is said to be sub-Gaussian with variance proxy σ^2 (sub-Gaussian parameter $\sigma > 0$) if

$$\mathbb{E}\left[e^{t(X-\mu)}\right] \le e^{\frac{t^2\sigma^2}{2}}, \quad \forall t \in \mathbb{R}.$$

Lemma 1 Let X be a sub-Gaussian RV with $\mathbb{E}[X] = \mu < \infty$ and variance proxy σ^2. Then, for all $\varepsilon > 0$,

$$\mathbb{P}[|X - \mathbb{E}[X]| \ge \varepsilon] \le 2 e^{\frac{-\varepsilon^2}{2\sigma^2}}.$$

Lemma 2 (Hoeffding's Lemma)
Let X be a RV s.t. $\mathbb{E}[X] = \mu < \infty$ and $\mathbb{P}[a \le X \le b] = 1$ for some constants $a, b \in \mathbb{R}$, $a < b$. Then,

$$\mathbb{E}\left[e^{t(X-\mu)}\right] \le e^{\frac{t^2}{2}\left(\frac{b-a}{2}\right)^2}, \quad \forall t \in \mathbb{R}$$

i.e. sub-Gaussian with variance proxy $\sigma^2 = \left(\frac{b-a}{2}\right)^2$

Proof of Hoeffding's Inequality

1) $\mathbb{E}\left[e^{t(S_n - \sum_{i=1}^n \mu_i)}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{t(X_i - \mu_i)}\right]$

$X_1, \ldots, X_n$ independent $\Rightarrow = \prod_{i=1}^n \mathbb{E}\left[e^{t(X_i - \mu_i)}\right]$

$0 \le \mathbb{E}\left[e^{t(X_i - \mu_i)}\right]$ and Lemma 2 $\Rightarrow \le \prod_{i=1}^n e^{\frac{t^2}{2}\left(\frac{b_i - a_i}{2}\right)^2}$

$= e^{\frac{t^2}{2}\sum_{i=1}^n \left(\frac{b_i - a_i}{2}\right)^2}$

$\Rightarrow$ S_n is sub-Gaussian with variance proxy $\sigma^2 = \sum_{i=1}^n \frac{(b_i - a_i)^2}{4}$

2) Let $X = S_n$ in Lemma 1. $\square$

Hoeffding's bound only depends on the range of X_i, not on its distribution over $[a_i, b_i]$.

Using more information about X_i can lead to a sharper bound. e.g. Chernoff.

Proof of Lemma 1:

Chernoff inequality $\Rightarrow$ $\forall\, t>0$,

$$\mathbb{P}[X-\mu>\varepsilon] \le \frac{\mathbb{E}[e^{t(X-\mu)}]}{e^{t\varepsilon}} \le \frac{e^{t^2\sigma^2/2}}{e^{t\varepsilon}}$$

Assumption of the lemma; call this $g(t)$.

$g'(t) = (t\sigma^2-\varepsilon)g(t) = 0 \Rightarrow t=\frac{\varepsilon}{\sigma^2}$ is a critical pt

$g''(t) = \sigma^2 g(t) + (t\sigma^2-\varepsilon)^2 g(t) > 0 \Rightarrow g(t)$ is minimized at $t=\frac{\varepsilon}{\sigma^2}$

$g(\frac{\varepsilon}{\sigma^2}) = e^{\frac{-\varepsilon^2}{2\sigma^2}}$.

Similarly, $\forall\, t<0$,

$$\mathbb{P}[X-\mu<-\varepsilon] \le \frac{\mathbb{E}[e^{t(X-\mu)}]}{e^{-t\varepsilon}} \le \frac{e^{t^2\sigma^2/2}}{e^{-t\varepsilon}}$$ is minimized at $t=-\frac{\varepsilon}{\sigma^2}$.

$$\mathbb{P}[|X-\mu|\ge\varepsilon] = \mathbb{P}[(X-\mu\ge\varepsilon)\cup(X-\mu\le-\varepsilon)]$$

Disjoint events $\Rightarrow$ $= \mathbb{P}[X-\mu\ge\varepsilon] + \mathbb{P}[X-\mu\le-\varepsilon]$

+ additivity property $\le 2e^{-\varepsilon^2/(2\sigma^2)}$ □

Proof of Lemma 2:

MGF $M_Y(t)$

Let $Y = X-\mu$ and $\varphi(t) = \log \mathbb{E}[e^{tY}]$

Note: $\varphi(0) = \log 1 = 0$

$$\frac{d}{dt}\varphi(t) = \frac{1}{M_Y(t)}\mathbb{E}[Ye^{tY}],\quad \frac{d^2}{dt^2}\varphi(t) = \frac{\mathbb{E}[Y^2e^{tY}]}{M_Y(t)} - \left[\frac{d\varphi}{dt}\right]^2$$

Note: $\frac{d}{dt}\varphi(0) = 0$ since $\mathbb{E}[Y]=\mathbb{E}[X-\mu]=0$.

Define a prob measure ($\mathbb{P}$ = original measure.)

$$\mathbb{P}_t[y\le Y\le y+dy] = \frac{e^{ty}}{M_Y(t)}\mathbb{P}[y\le Y\le y+dy]$$

Note: $\mathbb{P}_0 = \mathbb{P}$ ($\mathbb{E}_t$ = expectation w.r.t. $\mathbb{P}_t$)

Then, $\frac{d}{dt}\varphi(t) = \mathbb{E}_t[Y]$

$\frac{d^2}{dt^2}\varphi(t) = \mathbb{E}_t[Y^2] - (\mathbb{E}_t[Y])^2$

$= \mathrm{Var}_t(Y)$

Var does not change under constant shift. $= \mathrm{Var}_t(Y - \frac{a+b}{2} + \mu)$

since $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$ $\le \mathbb{E}_t[(X-\frac{a+b}{2})^2] \le (\frac{b-a}{2})^2$

$1 = \mathbb{P}[a\le X\le b] = \mathbb{P}[-\frac{(b-a)}{2} \le X - \frac{a+b}{2} \le \frac{(b-a)}{2}]$

$= \mathbb{P}[|X-\frac{a+b}{2}| \le (\frac{b-a}{2})]$

$\varphi(0) = 0 = \frac{d\varphi}{dt}(0)$ and $\frac{d^2\varphi(t)}{dt^2} \le (\frac{b-a}{2})^2$

$\Rightarrow \varphi(t) \le \frac{t^2}{2}(\frac{b-a}{2})^2 \Rightarrow \mathbb{E}[e^{tY}] \le e^{\frac{t^2}{2}(\frac{b-a}{2})^2}$

since exp is monotonically increasing □

Approximating Binomial Distribution

Reference: Lec Note 8 from Fall 2022

$S_n \sim$ Binomial (n,p), $0<p<1$.

$S_n \overset{d}{=} I_1 + \cdots + I_n$ where $I_1, \ldots, I_n \sim$ Bernoulli(p)

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n$$

Factorials are cumbersome to work with.

Key tool: Stirling Approximation.

$$e^{\frac{1}{12n+1}} < \frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} < e^{\frac{1}{12n}}.$$

(Laplace Method for approximating integrals)

Thm (Entropy Approximation)

Let $S_n \sim$ Binomial (n,p) and define $f = \frac{k}{n}$.

Then, for $k = 1, 2, \ldots, n-1$,

$$\left[1 - \frac{1}{12nf(1-f)}\right] \frac{1}{\sqrt{2\pi n f(1-f)}} e^{-n\,KL(f\|p)} < \mathbb{P}(S_n = k) < \frac{1}{\sqrt{2\pi n f(1-f)}} e^{-n\,KL(f\|p)}$$

where $KL(f\|p) = -f \log\left(\frac{p}{f}\right) - (1-f)\log\left(\frac{1-p}{1-f}\right)$

Normal Approximation

viewed as a function of f

1) Taylor expansion of $KL(f\|p)$ about $f = p$ gives

$$KL(f\|p) = \frac{(f-p)^2}{2p(1-p)} + \frac{2g-1}{6g^2(1-g)^2}(f-p)^3$$

for some g between f and p.

2) To obtain a normal approximation, need the following quantity to be small.

$$\text{Remainder} := \left|\frac{n(2g-1)}{6g^2(1-g)^2}(f-p)^3\right| \le \frac{n|f-p|^3}{6[\min(f,p)\min(1-f,1-p)]^2}$$

- If p is away from both 0 and 1 and $n|f-p|^3$ is small, then this upper bound will be small.
- For $n|f-p|^3$ to be small, we need $|f-p|$ to decay faster than $\frac{1}{n^{1/3}}$ as $n \to \infty$.

Good News! $I_1, \ldots, I_n \overset{iid}{\sim}$ Bernoulli(p)

SLLN $\Rightarrow \frac{S_n}{n} \overset{d}{=} \frac{1}{n}(I_1 + \cdots + I_n) \overset{a.s.}{\to} p$ as $n \to \infty$

Normal approximation:

$$\mathbb{P}[S_n=k] \sim \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-\frac{(k-np)^2}{2np(1-p)}}$$

Accurate if $\dfrac{n|f-p|^3}{6[\min(f,p)\,\min(1-f,1-p)]^2} \ll 1$

In general, Normal approx. is less accurate than the entropy approximation.

$\Rightarrow n|f-p|^3 = O\left(\frac{1}{\sqrt{n}}\right) \to 0$ as $n\to\infty$

$\Rightarrow$ the normal approx. for $\mathbb{P}[S_n=k]$ is accurate for large n.

Thm (CLT for Binomial Distribution)

Let $S_n \sim$ Binomial (n,p) for $0<p<1$.

Then, $\forall\, a,b \in \mathbb{R}$ where $a<b$,

$$\lim_{n\to\infty} \mathbb{P}\left[a \le \sqrt{n}\,\frac{\left(\frac{S_n}{n}-p\right)}{\sqrt{p(1-p)}} \le b\right] = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{\frac{-t^2}{2}}\, dt.$$

Let $t_k = \dfrac{k-np}{\sqrt{np(1-p)}}$. Then,

$$t_k - t_{k-1} = \frac{1}{\sqrt{np(1-p)}} \to 0 \text{ as } n\to\infty$$

Pf

$$\mathbb{P}\left[a \le \sqrt{n}\,\frac{\left(\frac{S_n}{n}-p\right)}{\sqrt{p(1-p)}} \le b\right] = \sum_{k=np+a\sqrt{np(1-p)}}^{np+b\sqrt{np(1-p)}} \mathbb{P}[S_n=k] \sim \sum_{t_k\in[a,b]} (t_k - t_{k-1})\,\frac{1}{\sqrt{2\pi}}\, e^{-\frac{t_k^2}{2}}$$

In this range of k, $\left|\frac{k}{n}-p\right| = O\left(\frac{1}{\sqrt{n}}\right)$ ($\frac{k}{n} \to f$)

which is a Riemann sum converging to $\int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-t}\, dt$ as $n\to\infty$ □

Application of the CLT for Binomial(n,p)

p = proportion of the population supporting Kamala Harris.

n = # people polled u.a.r. from the population.

S_n = # people in the sample who support Harris.

$\hat{p}_n = \frac{S_n}{n}$ = estimator of p.

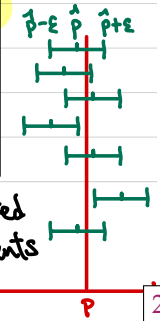
Q How large should n be s.t.

$$\mathbb{P}\left(p \in [\hat{p}_n - \varepsilon,\ \hat{p}_n + \varepsilon]\right) \geq 1-\alpha$$

(error; confidence level or nominal coverage probability)

$100(1-\alpha)\%$ confidence interval

for given $\varepsilon > 0$ and $0 < \alpha < 1$?



These intervals need to cover p at least $100(1-\alpha)\%$ of the time.

$$I_j = \begin{cases} 1, & \text{if the } j^{th} \text{ person polled supports Harris} \\ 0, & \text{otherwise.} \end{cases}$$

$I_1, \dots, I_n \overset{iid}{\sim} \text{Bernoulli}(p)$

$S_n = I_1 + \cdots + I_n \sim \text{Binomial}(n,p)$

$$\text{CLT} \Rightarrow \sqrt{n}\,\frac{(\hat{p}_n - p)}{\sqrt{p(1-p)}} \xrightarrow[\text{as } n\to\infty]{d} Z \sim N(0,1)$$

$$\mathbb{P}\left[\hat{p}_n - \varepsilon \leq p \leq \hat{p}_n + \varepsilon\right] = \mathbb{P}\left[-\varepsilon \leq \hat{p}_n - p \leq \varepsilon\right]$$

$$= \mathbb{P}\left[\underbrace{\frac{-\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}}_{-z} \leq \sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \leq \underbrace{\frac{+\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}}_{z}\right]$$

$$\text{CLT} \Rightarrow \approx \Phi(z) - \Phi(-z) = 1 - 2\Phi(-z)$$

$$1 - 2\Phi(-z) \geq 1-\alpha \Rightarrow \Phi(-z) \leq \alpha/2$$

$$\Rightarrow n \geq \left[\frac{-\Phi^{-1}\left(\frac{\alpha}{2}\right)}{\varepsilon}\right]^2 p(1-p)$$

since $p(1-p) \leq \frac{1}{4}$ $\forall\, p \in [0,1]$,

$$n \geq \frac{1}{4}\left[\frac{-\Phi^{-1}\left(\frac{\alpha}{2}\right)}{\varepsilon}\right]^2 \Rightarrow \text{the above inequality is satisfied}$$

$\varepsilon = 0.05,\ \alpha = 0.05 \Rightarrow n \geq 385$

$\varepsilon = 0.01,\ \alpha = 0.05 \Rightarrow n \geq 9{,}604$

Recall from Lecture 7: aka Continuity Thm

Recall that the MGF of $G \sim N(\mu, \sigma^2)$ is given by $M_G(t) = \mathbb{E}[e^{tX}] = e^{\mu t + \frac{\sigma^2 t^2}{2}}$.

$\mathbb{E}[e^{t(X-\mu)}] = 1 + \left(\frac{\sigma^2 t^2}{2}\right) + \frac{1}{2!}\left(\frac{\sigma^2 t^2}{2}\right)^2 + \cdots + \frac{1}{k!}\left(\frac{\sigma^2 t^2}{2}\right)^k + \cdots$

$\Rightarrow \mathbb{E}[(X-\mu)^p] = \begin{cases} 0, & \text{if } p \text{ is odd} \\ \frac{\sigma^p\, p!}{2^{p/2}\left(\frac{p}{2}\right)!} = \sigma^p (p-1)!!, & \text{if } p \text{ is even} \end{cases}$

Central moment

$(p-1)!! = (p-1)(p-3)\cdots 5 \cdot 3 \cdot 1$

Thm (Sum of ⫫ Normal RVs)

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ with $X_1 \perp\!\!\!\perp X_2$. Then,

$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Pf

$M_{X_1+X_2}(t) = M_{X_1}(t)\, M_{X_2}(t)$ by ⫫ of X_1, X_2

$= e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}} e^{\mu_2 t + \frac{\sigma_2^2 t^2}{2}}$

$= e^{(\mu_1+\mu_2)t + (\sigma_1^2+\sigma_2^2)\frac{t^2}{2}}$

↑ MGF of $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

Lecture 7, uniqueness of MGF

$\Rightarrow X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ □

Thm (Convergence in distribution)

Suppose $X_1, X_2, \ldots$ is a sequence of RVs with MGF $M_{X_n}(t)$ $\forall n$ well defined for $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$.

If $M_{X_n}(t) \longrightarrow M(t)$ as $n \to \infty$ for every $t \in (-\varepsilon, \varepsilon)$, then $M(t) = M_X(t)$, where M_X is the MGF of a RV s.t. $X_n \xrightarrow{d} X$ as $n \to \infty$.

Thm (Central Limit Theorem)

Let $X_1, X_2, X_3, \ldots$ be a sequence of iid RVs with finite mean μ and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then,

$$\lim_{n\to\infty} \mathbb{P}\left[\underbrace{\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right)}_{\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}}} \le x\right] = \Phi(x), \quad \forall x \in \mathbb{R}.$$

where $\Phi(x)$ denotes the c.d.f. of $N(0,1)$.

Pf of CLT

Let $Y_i = \frac{X_i - \mu}{\sigma}$. Then, $\mathbb{E}[Y_i] = 0$, $\mathrm{Var}[Y_i] = 1$.

Let $Z_n = \sqrt{n}\left(\frac{S_n}{n} - \mu\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$

Want to show $Z_n \xrightarrow{d} Z \sim N(0,1)$ as $n \to \infty$

$$M_{Z_n}(t) = \mathbb{E}\left[e^{\frac{t}{\sqrt{n}} \sum_{i=1}^{n} Y_i}\right] = \left[M_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

by ⫫ of $Y_1, \ldots, Y_n$

$$\log M_{Z_n}(t) = n \log M_{Y_1}\left(\frac{t}{\sqrt{n}}\right)$$

Define $L\left(\frac{t}{\sqrt{n}}\right)$

By Continuity Thm, MGF of $N(0,1)$

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{t^2}{2}},$$

or equivalently $\lim_{n\to\infty} n L\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2}$

implies $Z_n \xrightarrow{d} Z$ as $n \to \infty$.

$M_{Y_1}(0) = 1 \Rightarrow L(0) = 0.$

$L'(t) = \frac{M'_{Y_1}(t)}{M_{Y_1}(t)} \Rightarrow L'(0) = \mathbb{E}[Y_1] = 0.$

$L''(t) = \frac{M''_{Y_1}(t)}{M_{Y_1}(t)} - \left[\frac{M'_{Y_1}(t)}{M_{Y_1}(t)}\right]^2 \Rightarrow L''(0) = \mathbb{E}[Y_1^2] = 1$

Apply L'Hôpital's rule:

$$\lim_{n\to\infty} \frac{L\left(\frac{t}{\sqrt{n}}\right)}{\frac{1}{n}} = \lim_{n\to\infty} \frac{-\frac{1}{2} n^{-\frac{3}{2}} t\, L'\left(\frac{t}{\sqrt{n}}\right)}{-n^{-2}}$$

$$\frac{t\, L'\left(\frac{t}{\sqrt{n}}\right)}{2 n^{-\frac{1}{2}}}$$

$$= \lim_{n\to\infty} \frac{-\frac{1}{2} n^{-\frac{3}{2}} t^2 L''\left(\frac{t}{\sqrt{n}}\right)}{-n^{-\frac{3}{2}}}$$

$$\frac{t^2}{2} L''\left(\frac{t}{\sqrt{n}}\right)$$

$$= \frac{t^2}{2}$$

□

Rate of Convergence

Thm (Berry-Eséen Theorem)
There exists a constant C s.t. if $X_1, X_2, \dots, X_n$ are iid RVs with finite mean μ, finite variance σ^2, and finite $\rho = \mathbb{E}[|X_i - \mu|^3]$, then $\forall\, n \in \mathbb{N}$,

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \le \frac{C\rho}{\sigma^3 \sqrt{n}},$$

where $S_n = X_1 + \cdots + X_n$,

$$Z_n = \sqrt{n}\left(\frac{S_n}{n} - \mu\right)\Big/\sigma, \quad F_n(x) = \mathbb{P}[Z_n \le x].$$

Remarks:
1) This implies uniform convergence.
2) C does not depend on the distribution of X_i.
$0.4097 \le C \le 0.4748$

Extensions of CLT.

1) $\perp\!\!\!\perp$ but not identically distributed RVs satisfying additional conditions Lyapunov or Lindeberg condition (Stronger)

- Special case: (Bounded RVs)
  $X_1, X_2, \dots$ $\perp\!\!\!\perp$ RVs with $\mathbb{E}[X_i] = \mu_i < \infty$ and $\mathrm{Var}[X_i] = \sigma_i^2 < \infty$
  $S_n = X_1 + \cdots + X_n$, n^{th} partial sum.
  a) $\exists$ a constant $M > 0$ s.t. $\mathbb{P}[|X_i| < M] = 1 \ \forall\, i \in \mathbb{N}$.
  b) $\lim_{n\to\infty} \mathrm{Var}(S_n) = \infty$ ($\mathrm{Var}(S_n) = \sum_{i=1}^{n} \sigma_i^2$ since $X_1, \dots, X_n$ $\perp\!\!\!\perp$.)

Then,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} < x\right] = \Phi(x), \ \forall\, x \in \mathbb{R}$$

2) Identically distributed but not $\perp\!\!\!\perp$
Stationary m-dependent sequence of RVs

We have used the following theorem on multiple occasions to obtain several convergence results.

Thm (Convergence in distribution)
Suppose $X_1, X_2, \ldots$ is a sequence of RVs with MGF $M_{X_n}(t)$ $\forall n$ well defined for $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$.

If $M_{X_n}(t) \longrightarrow M(t)$ as $n \to \infty$ for every $t \in (-\varepsilon, \varepsilon)$, then $M(t) = M_X(t)$, where M_X is the MGF of a RV s.t. $X_n \xrightarrow{d} X$ as $n \to \infty$.

However, the MGF $M_X(t)$ may not exist for some RVs.
e.g. For $X \sim$ Cauchy distribution, $M_X(t) = \infty$, $\forall t \neq 0$.

Fortunately, the Fourier Transform of a RV X, defined as

$$\varphi_X(t) = \mathbb{E}[e^{itX}], \text{ where } i = \sqrt{-1},$$

is finite $\forall t \in \mathbb{R}$. This function is called the Characteristic function.

Furthermore, we have the following useful theorem!

Thm (Levy's Continuity Theorem)

$$X_n \xrightarrow{d} X \text{ as } n \to \infty$$

$$\Updownarrow$$

$$\varphi_{X_n}(t) \to \varphi_X(t) \text{ as } n \to \infty, \ \forall t \in \mathbb{R}.$$

pointwise convergence

We will study another useful convergence result today.

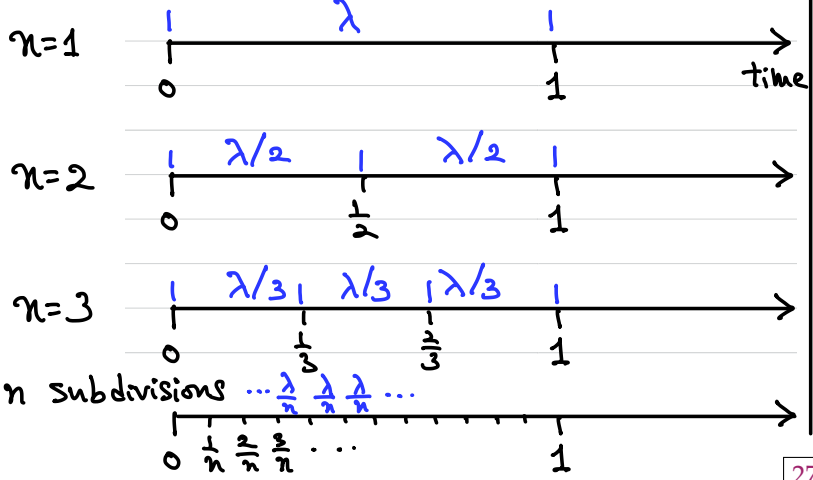
Recall that in Lecture 7 we showed:

Let $G_n \sim \text{Geometric}\left(\frac{\lambda}{n}\right)$, where $\lambda > 0$ and define $X_n = \frac{G_n}{n}$.

Then, $X_n \xrightarrow{d} X \sim \text{Exp}(\lambda)$ as $n \to \infty$

Discrete Continuous

Subdivide both the parameter and the time interval between trials finer and finer.



successes in the original unit interval:

$Y_n \sim \text{Binomial}(n, p_n)$, where $p_n = \frac{\lambda}{n}$

$\mathbb{E}[Y_n] = n p_n = \lambda$.

$Y_n \overset{d}{=} I_1 + \cdots + I_n$ where $I_1, \ldots, I_n \overset{iid}{\sim} \text{Bernoulli}(p_n)$

$\Rightarrow$ MGF: $M_{Y_n}(t) = \left[p_n e^t + (1-p_n)\right]^n$

$\lim_{n\to\infty} M_{Y_n}(t) = \left[1 + \frac{\lambda}{n}(e^t - 1)\right]^n \to e^{\lambda(e^t-1)}$

What does this correspond to?

Aside:

$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$ if $\lim_{n\to\infty} a_n = a$

See page 126 of Lecture Notes from Fall 2022

$Y \sim \text{Poisson}(\lambda)$, $\lambda > 0$

$M_Y(t) = \mathbb{E}[e^{tY}] = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!}$

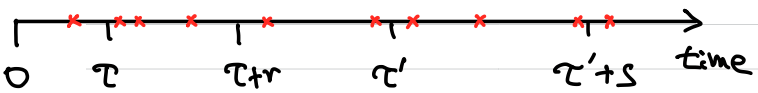
$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!}(\lambda e^t)^k = e^{-\lambda} e^{\lambda e^t}$

$\Rightarrow Y_n \xrightarrow{d} Y \sim \text{Poisson}(\lambda)$ as $n \to \infty$.

λ = intensity per unit time.

(Successes)
arrivals in this interval $\sim$ Poisson(λr)

$Y \sim$ Poisson(λr) $\quad Z \sim$ Poisson(λs)



$Y \perp\!\!\!\perp Z$ iff the intervals do not overlap.

Poissonization of the Multinomial

Consider repeated trials each with m types of outcome.
(For Bernoulli trials, $m=2$)
e.g. $m=6$: roll a die repeatedly

N = # trials
X_i = # times type i is observed.
Denote this by $\vec{X}$

$(X_1, \dots, X_m) \mid N=n \sim$ Multinomial$(n, p_1, \dots, p_m)$
NOT $\perp\!\!\!\perp$ given $N=n$ $\qquad \sum_{j=1}^m p_j = 1, \quad p_j \in [0,1]$

For $\vec{a} = (a_1, \dots, a_m)$ s.t. $\sum_{j=1}^m a_j = n$ and $a_j \in \{0, 1, \dots, n\}$,

$$P[\vec{X} = \vec{a} \mid N=n] = \binom{n}{a_1, \dots, a_m} p_1^{a_1} p_2^{a_2} \cdots p_m^{a_m}$$

$$\binom{n}{a_1, \dots, a_m} = \frac{n!}{a_1! \, a_2! \cdots a_m!}$$

Now, suppose the # trials is a RV $N \sim$ Poisson(λ), $\lambda > 0$.

What is the distribution of $(X_1, \dots, X_m)$?

$\vec{a} = (a_1, \ldots, a_m)$, $a_j \in \{0, 1, 2, \ldots\}$

$$\mathbb{P}[\vec{X} = \vec{a}] = \sum_{n=0}^{\infty} \mathbb{P}[\vec{X} = \vec{a} \mid N = n]\, \mathbb{P}[N = n]$$

$$= \sum_{n=0}^{\infty} \frac{\cancel{n!}}{a_1! a_2! \cdots a_m!} p_1^{a_1} p_2^{a_2} \cdots p_m^{a_m} \mathbb{1}\{\sum_{j=1}^{m} a_j = n\} \times \frac{\lambda^n}{\cancel{n!}} e^{-\lambda}$$

$$= \frac{p_1^{a_1} \cdots p_m^{a_m}}{a_1! a_2! \cdots a_m!} \lambda^{a_1 + a_2 + \cdots + a_m} e^{-\lambda(p_1 + \cdots + p_m)} \quad (e^{-\lambda(p_1 + \cdots + p_m)} = 1)$$

$$= \left[\frac{(p_1\lambda)^{a_1}}{a_1!} e^{-p_1\lambda}\right]\left[\frac{(p_2\lambda)^{a_2}}{a_2!} e^{-p_2\lambda}\right] \cdots \left[\frac{(p_m\lambda)^{a_m}}{a_m!} e^{-p_m\lambda}\right]$$

$\Rightarrow$ If $N \sim \text{Poisson}(\lambda)$,

1) $X_1, \ldots, X_m$ are $\perp\!\!\!\perp$

and 2) $X_j \sim \text{Poisson}(p_j \lambda)$

Def (Probability Generation Function) Let X be a non-negative integer valued RV. Then, for $|t| \le 1$,

$$G_X(t) := \mathbb{E}[t^X] = \sum_{n=0}^{\infty} t^n \mathbb{P}[X = n].$$

e.g.) $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$.

$$G_X(t) = \sum_{n=0}^{\infty} t^n \frac{\lambda^n e^{-\lambda}}{n!} = e^{\lambda(t-1)}$$

Thm (Uniqueness) Suppose X, Y are non-negative RVs. If $G_X = G_Y$, then $X \overset{d}{=} Y$.

Thm Let $X_1, \ldots, X_n$ be $\perp\!\!\!\perp$, non-neg integer valued RVs. Then, the PGF for $S_n = X_1 + \cdots + X_n$ satisfies

$$G_{S_n}(t) = \prod_{k=1}^{n} G_{X_i}(t)$$

Pf $G_{S_n}(t) = \mathbb{E}(t^{X_1 + \cdots + X_n}) = \mathbb{E}\left[\prod_{k=1}^{n} e^{tX_k}\right]$

by $\perp\!\!\!\perp$ $= \prod_{k=1}^{n} \mathbb{E}[e^{tX_k}] = \prod_{k=1}^{n} G_{X_k}(t)$.

e.g. $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, $X \perp\!\!\!\perp Y$

$G_{X+Y}(t) = e^{\lambda(t-1)} e^{\mu(t-1)} = e^{(\mu+\lambda)(t-1)}$

Uniqueness Theorem $\Rightarrow X+Y \sim \text{Poisson}(\mu+\lambda)$

A useful theorem:

Thm (Compounding) Let $X_1, X_2, \cdots$ be a sequence of iid non-negative integer-valued RVs with common PGF G_X, while N is a non-negative integer-valued RV $\perp\!\!\!\perp$ of $X_1, X_2, \cdots$ with PGF G_N.

Let $S_N = X_1 + \cdots + X_N$. Then,

$$G_{S_N}(t) = G_N(G_X(t))$$

$$= \sum_{n=0}^{\infty} (G_X(t))^n \, \mathbb{P}[N=n]$$

Pf

$$G_{S_N}(t) = \sum_{k=0}^{\infty} t^k \, \mathbb{P}[S_N = k]$$

$$= \sum_{k=0}^{\infty} t^k \sum_{n=0}^{\infty} \mathbb{P}[S_N = k \mid N=n] \mathbb{P}[N=n]$$

$X_1, X_2, \cdots \perp\!\!\!\perp N \Rightarrow = \mathbb{P}[S_n = k]$

interchange the order of summations (ok to do this here.)

$$= \sum_{n=0}^{\infty} \left[\sum_{k=0}^{\infty} t^k \, \mathbb{P}[S_n = k] \right] \mathbb{P}[N=n]$$

$G_{S_n}(t) = [G_X(t)]^n$

□

Lecture 12

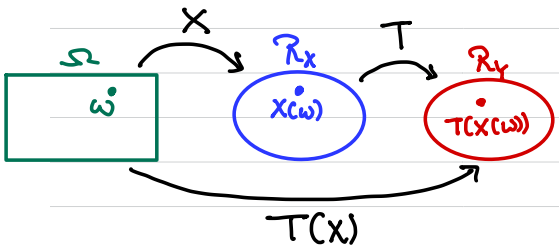
Transformations of RVs

Probability space $(\Omega, \mathcal{F}, \mathbb{P})$

RV $X: \Omega \to \mathcal{R}_X \subseteq \mathbb{R}^n$

$T: \mathcal{R}_X \to \mathcal{R}_Y \subseteq \mathbb{R}^m$

RV $T(X): \Omega \to \mathcal{R}_Y$



Technical detail: T needs to be a measurable function.

σ-algebra on $\mathcal{R}_X$

Measurable spaces $(\mathcal{R}_X, \Sigma_X)$, $(\mathcal{R}_Y, \Sigma_Y)$

For all $A \in \Sigma_Y$, require $T^{-1}(A) \in \Sigma_X$

$T^{-1}(A) := \{x \in \mathcal{R}_X \mid T(x) \in A\}$ is the preimage of A

e.g. $X = (X_1, X_2)$, $X_1, X_2 \overset{iid}{\sim} \mathcal{N}(0,1)$

$T(X) = X_1^2 + X_2^2$

Note: This transformation is NOT invertible. But preimage $T^{-1}(A)$ is well defined.

Q What is the distribution of $Y = T(X)$?

Change of Variable Principle:

For a given T, the distribution of $T(X)$ is determined by the distribution of X.

For $A \subseteq \mathcal{R}_Y$, (technically $A \in \Sigma_Y$)

$$\begin{aligned}\mathbb{P}[T(X) \in A] &= \mathbb{P}\{\omega \in \Omega \mid T(X(\omega)) \in A\} \\ &= \mathbb{P}\{\omega \in \Omega \mid X(\omega) \in T^{-1}(A)\} \\ &= \mathbb{P}[X \in T^{-1}(A)] \qquad (*)\end{aligned}$$

Discrete Case:

$$\mathbb{P}[T(X) \in A] = \sum_{x \in \mathcal{R}_X : T(x) \in A} \mathbb{P}[X = x]$$

Continuous Case:

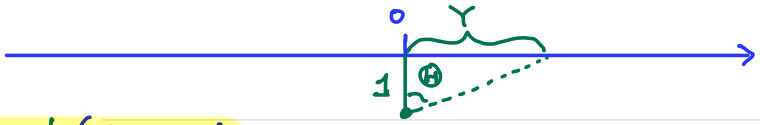
Define $Y = T(X)$

$$(*) \Rightarrow \int_A f_Y(y)\,dy = \int_{T^{-1}(A)} f_X(x)\,dx$$

p.d.f. is given by $f_Y(y) = \lim_{\delta \downarrow 0} \frac{\mathbb{P}[y < Y < y+\delta]}{\delta}$

For $\delta \ll 1$, $\mathbb{P}[y < Y < y+\delta] \approx f_Y(y)\,\delta$

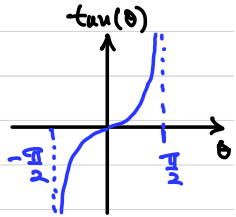
Note: $\mathbb{P}[Y = y] = 0$ for continuous RV Y.



e.g. (Cauchy)

$\Theta \sim \text{Uniform}(-\frac{\pi}{2}, \frac{\pi}{2})$

$Y = \tan(\Theta) = \frac{\sin(\Theta)}{\cos(\Theta)}$



For $y \in \mathbb{R}$,

$$\mathbb{P}[y < \tan(\Theta) < y + dy] \approx f_Y(y)\,dy$$

(since $\tan^{-1}$ is monotonically increasing)

$$= \mathbb{P}[\tan^{-1}(y) < \Theta < \tan^{-1}(y+dy)]$$

(Taylor approximation)

$$\approx \mathbb{P}[\tan^{-1}(y) < \Theta < \tan^{-1}(y) + dy \underbrace{\frac{d}{dy}(\tan^{-1}(y))}_{\frac{1}{1+y^2}}]$$

$$\approx \underbrace{f_\Theta(\tan^{-1}(y))}_{1/\pi} \frac{dy}{1+y^2}$$

$$\Rightarrow f_Y(y) = \frac{1}{\pi(1+y^2)}, \quad y \in \mathbb{R}$$

Recall $\mathbb{E}[Y^n] = \infty$ for all $n \in \mathbb{N} = \{1, 2, 3, \ldots\}$

In general, if T is invertible and T^{-1} is differentiable, the p.d.f. of $Y = T(X)$ is given by

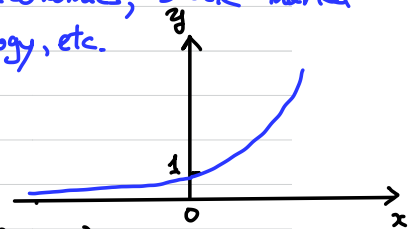
$$f_Y(y) = f_X(T^{-1}(y)) \left| \frac{dT^{-1}(y)}{dy} \right|$$

e.g. (Log-normal)

Commonly used in economics, stock market analysis, engineering, biology, etc.

$X \sim N(\mu, \sigma^2)$

$Y = T(X) = e^X$



$T^{-1}(y) = \log(y)$

$$f_Y(y) = f_X(\log(y)) \frac{d(\log(y))}{dy}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y} e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}}, \quad y > 0$$

All moments of Y exist:

$$\mathbb{E}[Y^n] = \int_0^\infty y^n f_Y(y)\,dy = e^{n\mu + \frac{n^2\sigma^2}{2}}$$

However,

$$\mathbb{E}[e^{tY}] = \int_0^\infty e^{ty} f_Y(y)\,dy = \infty, \quad \forall t > 0.$$

$\Rightarrow$ There does not exist a neighborhood $(-\varepsilon, \varepsilon)$, $\varepsilon > 0$ where MGF exists.

Log-normal

Mode = $e^{\mu-\sigma^2}$

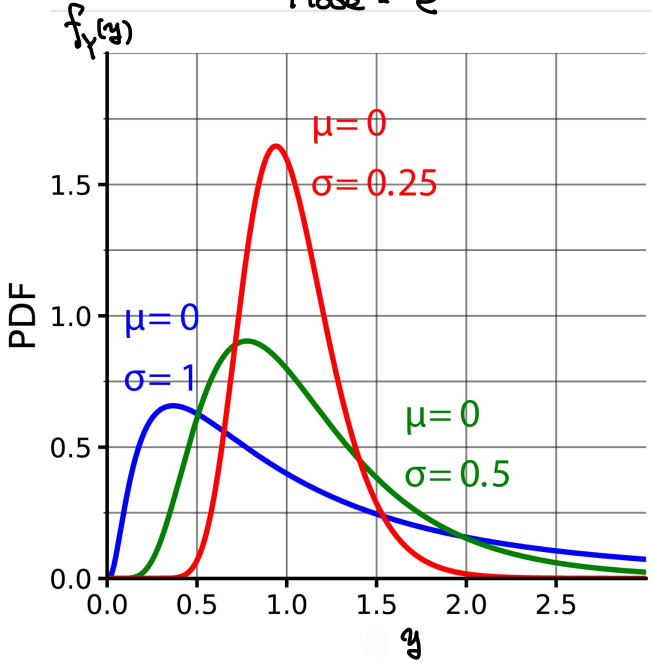


Figure from Wikipedia

e.g. (Chi-square or Gamma)

(Non-invertible)

$X \sim N(0,1)$, $Y = T(X) = X^2$.

For $y > 0$,

$\mathbb{P}[y < Y < y+dy] \approx f_Y(y)\,dy$

$= \mathbb{P}[\sqrt{y} < X < \sqrt{y+dy}] + \mathbb{P}[-\sqrt{y+dy} < X < -\sqrt{y}]$

$\approx \mathbb{P}[\sqrt{y} < X < \sqrt{y} + dy\,\frac{1}{2\sqrt{y}}] + \mathbb{P}[-\sqrt{y} - dy\,\frac{1}{2\sqrt{y}} < X < -\sqrt{y}]$

$\approx f_X(\sqrt{y})\,\frac{dy}{2\sqrt{y}}$ $\qquad$ $\approx f_X(-\sqrt{y})\left(\frac{dy}{2\sqrt{y}}\right)$

$f_X(-\sqrt{y}) = f_X(\sqrt{y})$

$\Rightarrow f_Y(y) = \frac{f_X(\sqrt{y})}{\sqrt{y}} = \frac{1}{\sqrt{2\pi}}\, y^{-\frac{1}{2}}\, e^{-\frac{y}{2}}$, for $y>0$

This is χ^2_1 or Gamma$(\frac{1}{2}, 2)$ distribution

$M_Y(t) = \mathbb{E}[e^{tY}] = \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{y(t-\frac{1}{2})}\,dy$

$u = (1-2t)\,y$

$du = (1-2t)\,dy$

$= \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}(1-2t)}\,dy$

$= \frac{1}{\sqrt{1-2t}} \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{u}} e^{-\frac{u}{2}}\,du$ (the integral = 1)

$= \frac{1}{\sqrt{1-2t}}$

for $t < \frac{1}{2}$

(Many-to-one T) Suppose $T^{-1}(y)$ of each $y \in R_Y$ consists of a finite or countably infinite set of points $\{T_i^{-1}(y)\}$ (e.g. $\pm\sqrt{y}$ above), where T_i are differentiable. Then,

$$f_Y(y) = \sum_i f_X(T_i^{-1}(y)) \left| \frac{dT_i^{-1}(y)}{dy} \right|$$

---

e.g. (Length of a standard Gaussian vector in n-dim)

$$V = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{bmatrix} \qquad X_1, \dots, X_n \overset{iid}{\sim} N(0,1)$$

$$S_n = X_1^2 + X_2^2 + \cdots + X_n^2$$

$$M_{S_n}(t) = [M_{X_1}(t)]^n = \frac{1}{(1-2t)^{n/2}}$$

$Y \sim \chi^2_n$ or Gamma $(\frac{n}{2}, 2)$ has p.d.f.

$$f_Y(x) = \frac{1}{\Gamma(\frac{n}{2})} \frac{1}{2^{n/2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x \geq 0$$

$$M_Y(t) = \frac{1}{(1-2t)^{n/2}}, \quad t < \tfrac{1}{2}$$

$$\Rightarrow X_1^2 + \cdots + X_n^2 \sim \chi^2_n$$

Let $R_n = \sqrt{X_1^2 + X_2^2 + \cdots + X_n^2}$. Then

$$f_{R_n}(r) = f_{X_1^2+\cdots+X_n^2}(r^2) \left| \frac{dr^2}{dr} \right|$$

$$= \begin{cases} \frac{1}{\Gamma(\frac{n}{2})} \frac{1}{2^{\frac{n}{2}-1}} r^{n-1} e^{-\frac{r^2}{2}}, & r \geq 0, \\ 0, & r < 0. \end{cases}$$

$$\mathbb{E}[R_n] = \sqrt{2}\,\Gamma(\tfrac{n}{2} + \tfrac{1}{2}) / \Gamma(\tfrac{n}{2})$$

(Chi Distribution)

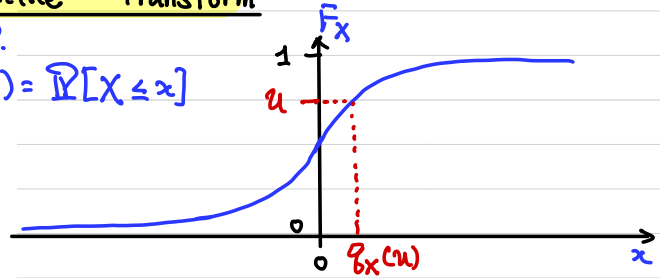
$n=2$: **Rayleigh Distribution**

$n=3$: **Maxwell-Boltzmann Distribution** — Used to describe particle speeds in ideal gas

Quantile Transform

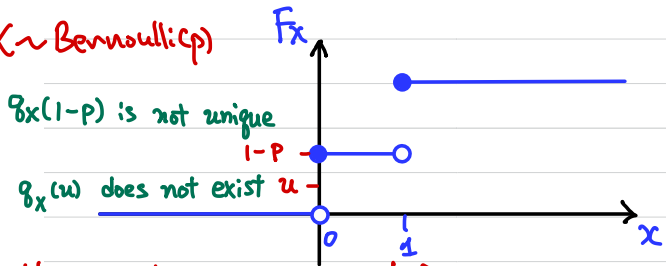
c.d.f.

$F_X(x) = \mathbb{P}[X \le x]$



For $u \in (0,1)$, the u-quantile $q_X(u) \in \mathbb{R}$ of F_X satisfies $F_X(q_X(u)) = u$, provided that it exists.

$X \sim \text{Bernoulli}(p)$



To address these issues, define q_X as:

Def (Quantile Function) For $u \in (0,1)$,

$$q_X(u) = \inf\{x \in \mathbb{R} \mid F_X(x) \ge u\}$$

"infimum" greatest lower bound

For $X \sim \text{Bernoulli}(p)$,

$$q_X(u) = \begin{cases} 0, & \text{if } 0 < u \le 1-p, \\ 1, & \text{if } 1-p < u < 1, \end{cases}$$

Let $U \sim \text{Uniform}(0,1)$

1) X either discrete or continuous

Then, $q_X(U) \overset{d}{=} X$

2) F_X continuous

Then, $F_X(X) \overset{d}{=} U$

$F_X(x) = \mathbb{P}[X \le x] = \mathbb{P}\{\omega \in \Omega \mid X(\omega) \le x\}$

$X : \Omega \to \mathcal{R}_X$, $F_X : \mathcal{R}_X \to [0,1]$

$F_X(X) : \Omega \to [0,1]$

$F_X(X(\omega)) = \mathbb{P}\{\omega' \in \Omega \mid X(\omega') \le X(\omega)\} = \int_{-\infty}^{X(\omega)} f_X(x)\,dx$

F_X continuous $\Rightarrow$ Inverse exists and

$q_X(u) = F_X^{-1}(u)$ for all $u \in (0,1)$

For $u \in (0,1)$

$\mathbb{P}[F_X(X) \le u] = \mathbb{P}[F_X^{-1}(F_X(X)) \le F_X^{-1}(u)]$ (since F_X^{-1} is strictly increasing)

$= \mathbb{P}[X \le F_X^{-1}(u)]$

Def. of CDF $= F_X(F_X^{-1}(u)) = u$

STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Lecture 13

It is natural to consider multiple random variables and try to understand their interaction. For instance, random vectors, random matrices, point processes, among many other interesting objects. Here we focus on the case of **bivariate jointly continuous random variables**.

Definition: We say that (X, Y) are *jointly continuous* if there exist a *joint probability density function* f on $\mathbb{R}^2$ such that for every measurable $B \subset \mathbb{R}^2$.

$$\mathbb{P}\big((X, Y) \in B\big) = \int \int_B f(x, y)\, dx\, dy.$$

A function f in $\mathbb{R}^2$ defines a joint probability density function (also called joint density function) if

$$f(x, y) \geq 0 \text{ and } \int \int_{\mathbb{R}^2} f(x, y)\, dx\, dy = 1.$$

Example 1: (Uniform)
Let D be a subset of $\mathbb{R}^2$ with finite non-zero area. A random point (X, Y) is uniformly distributed on D if it's joint density is

$$f_{X,Y}(x, y) = \frac{I_D(x, y)}{\text{Area}(D)} = \begin{cases} \frac{1}{\text{Area}(D)}, & \text{if } (x, y) \in D, \\ 0, & \text{else.} \end{cases}$$

Example 2: (Bivariate normal with correlation 0)
Take (X, Y) with joint probability density function $f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$.

Remark: In the one dimensional case, for continuous random variable, the probability density function represent the infinitesimal probability of a random variable to take a specific value. More concretely,

$$\mathbb{P}(X \in [x, x + \varepsilon)) \approx f_X(x)\varepsilon$$

Similarly, in the two dimensional case, take a small neighborhood Δ containing a point (x, y), we then have

$$\mathbb{P}\big((X, Y) \in \Delta\big) \approx f_{X,Y}(x, y)\text{Area}(\Delta).$$

Question: Given a joint density for (X, Y), how can one recover the *marginal density* for X or Y?

Proposition: Let (X, Y) be jointly continuous with joint density $f_{X,Y}$ and let f_X and f_Y be the densities for X and Y respectively. Then

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy,$$
$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dx.$$

Proof. We will provide a proof for f_X, the other case being analogous.

$$\begin{aligned}\mathbb{P}(X \leq t) &= \mathbb{P}(X \leq t, -\infty<Y<\infty) \\ &= \int_{-\infty}^{t}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.\end{aligned}$$

Differentiating both sides of the identity gives the statment of the proposition. □

Remark: We can have $X \overset{d}{=} W$ and $Y \overset{d}{=} Z$ but $(X,Y) \overset{d}{\neq} (W,Z)$. This is precisely why it is interesting to study joint distributions.

Example 3: Let X, W, V be independent Exp(1) random variables and Z be another independent Gamma$(2,1)$. Let $Y = X + V$. It follows from previous lectures that $Y \overset{d}{\sim}$ Gamma$(2,1)$. We then have $X \overset{d}{=} W$ and $Y \overset{d}{=} Z$. However, we always have $X \leq Y$, while the event $W>Z$ is possible with positive probability, so $(X,Y) \overset{d}{\neq} (W,Z)$.

Question: Given a transformation T and the joint density of (X,Y), what is the joint density of $(W,Z) = T(X,Y)$?

Polar coordinates.
Let (R,Θ) be polar coordinates for the point (X,Y), then

$$f_{R,\Theta}(r,\theta) = r f_{X,Y}(r\cos(\theta), r\sin(\theta)).$$

Proof. On one hand, $\mathbb{P}(r \leq R \leq r+\delta, \theta \leq \Theta \leq \theta+\varepsilon) \approx f_{R,\Theta}(r,\theta)\,\delta\varepsilon$. On the other hand, by describing this event using (X,Y) (Draw a picture) we have that

$$\begin{aligned}\mathbb{P}(r \leq R \leq r+\delta, \theta \leq \Theta \leq \theta+\varepsilon) &= \mathbb{P}\big((X,Y) \in \Delta\big) \\ &\approx f_{X,Y}(x,y)\text{Area}(\Delta) \\ &= f_{X,Y}(r\cos(\theta), r\sin(\theta))\, r\delta\varepsilon.\end{aligned}$$

Putting together both approximations we conclude. □

Example 4: As an application of this result for (X,Y) bivariate normal with correlation 0 (See example 2) we get that

$$f_{R,\Theta}(r,\theta) = \frac{r}{2\pi} e^{-\frac{r^2}{2}}.$$

This joint density doesn't depend on θ: the normal distribution is *radially symmetric*.

Definition: We say that a function $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a *linear transformation* if

$$T\begin{pmatrix}x\\y\end{pmatrix} = M_T\begin{pmatrix}x\\y\end{pmatrix} + P_T.$$

Here M_T is a 2×2 matrix and P_T is a 2×1 vector.

Example 4: Consider $T\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}x-y\\x+y+1\end{pmatrix}$. Here $M_T = \begin{pmatrix}1 & -1\\1 & 1\end{pmatrix}$ while $P_T = \begin{pmatrix}0\\1\end{pmatrix}$.

Some properties: (Without proof)

1. T is invertible if and only if M_T is invertible.
2. if P is a parallelogram in $\mathbb{R}^2$, then $T(P)$ is also a parallelogram.
3. For every parallelogram P, $\text{Area}(T(P)) = \text{Area}(P)\,|\det(M_T)|$.

Remark: Property 1 is giving us a criteria to determine when a linear transformation is invertible (for example, we can calculate the determinant of the 2×2 matrix M_T). Property 2 is saying that linear transformation are well behaved from a geometric point of view. Property 3 allow us to understand how areas are deformed through linear transformations.

We want understand how joint distributions change after applying linear transformations. Given $f_{X,Y}$, what is $f_{T(X,Y)}$?

Invertible linear transformations:
Given T a linear transformation with inverse S, let $(W, Z) = T(X, Y)$. Let P be a small parallelogram containing a point (w, z). On one hand,

$$\mathbb{P}\big((W, Z) \in P\big) \approx f_{W,Z}(w, z)\,\text{Area}(P).$$

Similarly,

$$\begin{aligned}\mathbb{P}\big((W, Z) \in P\big) &= \mathbb{P}\big(T(X, Y) \in P\big)\\ &= \mathbb{P}\big((X, Y) \in S(P)\big)\\ &\approx f_{X,Y}(x, y)\,\text{Area}(S(P))\\ &= f_{X,Y}(S(w, z))\,\text{Area}(P)\,|\det(M_S)|\end{aligned}$$

We conclude that for invertible transformations T,

$$f_{W,Z}(w, z) = f_{X,Y}(S(w, z))\,|\det(M_S)|.$$

Rotations:
Given (X, Y) in the plane, we may be interested in the distribution we obtain after rotation by angle of θ, of the new coordinates (X_θ, Y_θ) in counterclockwise direction. This is given by the linear transformation T_θ.

$$T_\theta \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Using our previous result in this situation provides the following

$$f_{X_\theta,Y_\theta}(w, z) = f_{X,Y}(w\cos(\theta) + z\sin(\theta),\, z\cos(\theta) - w\sin(\theta)).$$

Sums and differences:
Say we would like to obtain the joint density for $(X + Y, X - Y)$. This can be obtained using the linear transformation

$$T \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Again, a direct application of our result for invertible linear transformations provides

$$f_{X+Y,X-Y}(w, z) = \frac{1}{2} f_{X,Y}\Big(\frac{w + z}{2}, \frac{w - z}{2}\Big).$$

Orthogonal transformations:

Definition: An *orthogonal transformation* T is linear transformation that preserves the inner product. So it satisfies

$$\langle \vec{v}, \vec{u} \rangle = \langle T\vec{v}, T\vec{u} \rangle.$$

This definition is quite abstract, so we will give a few facts to get to know them better.

1. It preserves Euclidean norm, $\|\vec{v}\| = \|T\vec{v}\|$. Orthogonal transformations preserve angles, lengths and areas.
2. For T an orthogonal transformation, $P_T = \vec{0}$ and M_T is an orthogonal matrix, so $M_T^T = M_T^{-1}$. In particular $\det(M_T) = \pm 1$.
3. In two dimensions, orthogonal transformations are rotations, reflections or composition of rotation and reflections.

Example 5: The following are reflections, $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ or $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Example 6: If (X, Y) has radial symmetry, so the density function $f_{X,Y}(x, y)$ is of the form $g(x^2 + y^2)$, then the joint density is unchanged under orthogonal transformation. For instance, let (X, Y) be two independent standard normal random variables. Then $\left(\frac{X+Y}{\sqrt{2}}, \frac{X-Y}{\sqrt{2}}\right)$ are also two independent normal random variables.

Lecture 14

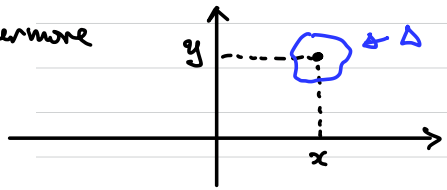
Recall from last lecture:

RVs X and Y have joint density $f_{X,Y}$

$\Rightarrow$ for every measurable $A \subset \mathbb{R}^2$,

$$\mathbb{P}[(X,Y) \in A] = \iint_A f_{X,Y}(x,y)\,dx\,dy.$$

Furthermore



$$f_{X,Y}(x,y) = \lim_{\Delta \downarrow (x,y)} \frac{\mathbb{P}[(X,Y) \in \Delta]}{\text{Area}(\Delta)}.$$

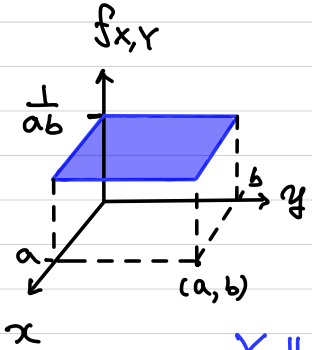
The precise shape of Δ does not matter.

Independence of X, Y

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y), \quad \forall (x,y) \in \mathbb{R}^2$$

f_X: a function of x only; f_Y: a function of y only

e.g. $f_{X,Y}(x,y) = I\{0 \le x \le a\} \cdot I\{0 \le y \le b\}$



$f_X(x) = I\{0 \le x \le a\}$

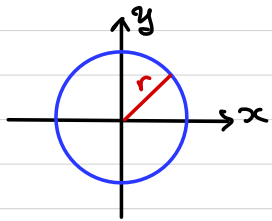
$f_Y(y) = I\{0 \le y \le b\}$

Knowing the value of X does not tell us anything about Y, and vice versa.

$X \perp\!\!\!\perp Y$.

---

e.g. How about $f_{X,Y}(x,y) = \frac{1}{\pi r^2} I\{x^2+y^2 \le r^2\}$?



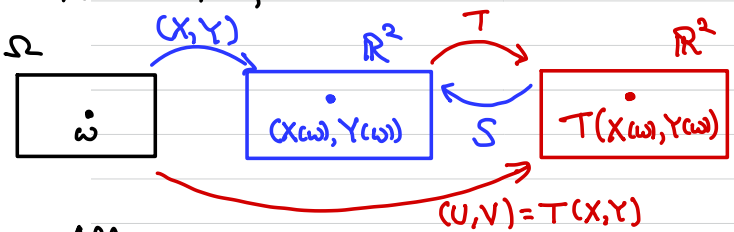
$$f_X(x) = \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} f_{X,Y}(x,y)\,dy = \frac{2}{\pi r^2}\sqrt{r^2-x^2}\, I\{-r \le x \le r\}$$

Similarly for $f_Y(y)$.

$X \not\perp\!\!\!\perp Y$.

$T: \mathbb{R}^2 \to \mathbb{R}^2$ invertible affine transformation.

$T(X,Y) = (U,V)$, Inverse $S(U,V) = (X,Y)$



Affine = Linear + Translation

$$T(X,Y) = M_T \begin{bmatrix} X \\ Y \end{bmatrix} + P_T = \begin{bmatrix} U \\ V \end{bmatrix}$$

M_T: 2×2 invertible matrix; P_T: 2×1 vector

$$S\,T(X,Y) = S(U,V) = (X,Y)$$

$$S(U,V) = M_T^{-1} \begin{bmatrix} U \\ V \end{bmatrix} - M_T^{-1} P_T$$

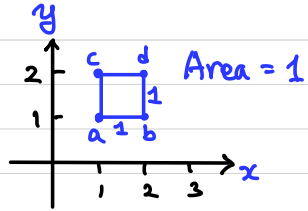
$$=: M_S \begin{bmatrix} U \\ V \end{bmatrix} + P_S$$

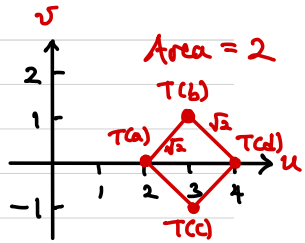
$$\mathbb{P}[(U,V) \in B] = \mathbb{P}[(X,Y) \in S(B)]$$

$$\iint_B f_{U,V}(u,v)\,du\,dv = \iint_{S(B)} f_{X,Y}(x,y)\,dx\,dy$$

$$\boxed{f_{U,V}(u,v) = f_{X,Y}(S(u,v))\,|\det M_S|}$$

e.g. $U = X+Y$, $V = X-Y$ $\Rightarrow$ $X = \frac{1}{2}(U+V)$, $Y = \frac{1}{2}(U-V)$





Δ = a small region containing (u,v)

$$\mathbb{P}[(U,V) \in \Delta] = \mathbb{P}[(X,Y) \in S(\Delta)]$$

$\approx$ $f_{U,V}(u,v)\,\text{Area}(\Delta)$ $\approx$ $f_{X,Y}(\frac{1}{2}(u+v), \frac{1}{2}(u-v))\,\underbrace{\text{Area}(S(\Delta))}_{\frac{1}{2}\text{Area}(\Delta)}$

Algebraically, $T\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} U \\ V \end{pmatrix}$

$M_T = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $M_S = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $M_S M_T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$|\det M_S| = \left|\frac{1}{4}(-2)\right| = \frac{1}{2}$

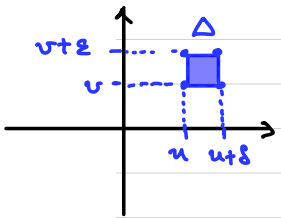
$$f_{U,V}(u,v) = \frac{1}{2} f_{X,Y}(\tfrac{1}{2}(u+v), \tfrac{1}{2}(u-v))$$

General Invertible Transformations.

$T: \mathbb{R}^2 \to \mathbb{R}^2$ Assume differentiable. but not necessarily affine.

$T(X,Y) = (U,V)$, $S(U,V) = (X,Y)$

$ST(X,Y) = (X,Y)$



$f_{U,V}(u,v)\,\text{Area}(\Delta) \approx$

$\mathbb{P}[(U,V) \in \Delta] = \mathbb{P}[(X,Y) \in S(\Delta)]$

$\approx f_{X,Y}(S(u,v))\,\text{Area}(S(\Delta))$

What is $S(\Delta)$?

If S is affine, then $S(\Delta)$ is a parallelogram

For general S, if $\delta, \varepsilon \ll 1$, then $S(\Delta)$ can be approximated by a parallelogram, since S can be approximated by an affine transformation on Δ.

Let $S\begin{pmatrix} u \\ v \end{pmatrix} = \begin{bmatrix} S_1(u,v) \\ S_2(u,v) \end{bmatrix} \in \mathbb{R}^2$, where

$S_i: \mathbb{R}^2 \to \mathbb{R}$ are differentiable functions. $i = 1, 2$

Then, for any point (a,b) near (u,v),

The Taylor expansion in 2-d gives

$$S_i(a,b) \approx S_i(u,v) + (a-u)\frac{\partial S_i}{\partial u}(u,v) + (b-v)\frac{\partial S_i}{\partial v}(u,v)$$

In matrix notation,

$$S\begin{bmatrix} a \\ b \end{bmatrix} \approx \begin{bmatrix} S_1(u,v) \\ S_2(u,v) \end{bmatrix} + \underbrace{\begin{bmatrix} \frac{\partial S_1}{\partial u}(u,v) & \frac{\partial S_1}{\partial v}(u,v) \\ \frac{\partial S_2}{\partial u}(u,v) & \frac{\partial S_2}{\partial v}(u,v) \end{bmatrix}}_{J_S(u,v)} \begin{bmatrix} a-u \\ b-v \end{bmatrix}$$

$J_S(u,v)$ Jacobian matrix of S at (u,v)

↑ an affine transformation.

Hence,

$$\mathbb{P}[(X,Y) \in S(\Delta)] \approx f_{X,Y}(S(u,v)) \underbrace{|\det J_S(u,v)|\,\text{Area}(\Delta)}_{\approx \text{Area}(S(\Delta))}$$

$$\Rightarrow \boxed{f_{U,V}(u,v) = f_{X,Y}(S(u,v))\,|\det J_S(u,v)|}$$

e.g. $X \sim \text{Gamma}(\alpha_1, \beta)$, $Y \sim \text{Gamma}(\alpha_2, \beta)$ and $X \perp\!\!\!\perp Y$. (α: shape, β: rate)

Let $U = X+Y$, $V = \frac{X}{X+Y}$ (since $X, Y > 0$, $V \in (0,1)$)

$\Rightarrow X = UV$, $Y = U - X = U(1-V)$

$S_1(u,v) = uv$

$S_2(u,v) = u(1-v)$

$J_S(u,v) = \begin{bmatrix} v & u \\ 1-v & -u \end{bmatrix}$

$|\det J_S(u,v)| = |-uv - u(1-v)| = |-u|$

$f_{U,V}(u,v) = f_{X,Y}(uv, u(1-v))\, u\, I\{u \in (0,\infty)\}\, I\{v \in (0,1)\}$

$X \perp\!\!\!\perp Y \Rightarrow f_X(uv)\, f_Y(u(1-v))$

$= \frac{\beta^{\alpha_1}(uv)^{\alpha_1-1} e^{-\beta uv}}{\Gamma(\alpha_1)} \frac{\beta^{\alpha_2}[u(1-v)]^{\alpha_2-1} e^{-\beta(u(1-v))}}{\Gamma(\alpha_2)} \times u\, I\{u \in (0,\infty)\}\, I\{v \in (0,1)\}$

$= \left[\frac{\beta^{\alpha_1+\alpha_2} u^{\alpha_1+\alpha_2-1} e^{-\beta u}}{\Gamma(\alpha_1+\alpha_2)}\right] \left[\frac{\Gamma(\alpha_1+\alpha_2) v^{\alpha_1-1}(1-v)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\right]$

$\times I\{u \in (0,\infty)\}$ $\times I\{v \in (0,1)\}$

p.d.f. for $\text{Gamma}(\alpha_1+\alpha_2, \beta)$ p.d.f. for $\text{Beta}(\alpha_1, \alpha_2)$

$\Rightarrow f_{U,V}(u,v) = f_U(u) f_V(v), \quad \forall (u,v) \in \mathbb{R}^2$

$\Rightarrow F_{U,V}(u,v) = F_U(u) F_V(v), \quad \forall (u,v) \in \mathbb{R}^2$

$\Rightarrow U \perp\!\!\!\perp V$

In summary,

$X+Y \sim \text{Gamma}(\alpha_1+\alpha_2, \beta)$

$\frac{X}{X+Y} \sim \text{Beta}(\alpha_1, \alpha_2)$

and $X+Y \perp\!\!\!\perp \frac{X}{X+Y}$

Lecture 15

Problem of the day

- Alice and Bob agree to meet for lunch but both forget the exact agreed time.

- Each person arrives at the cafeteria u.a.r. between 12 pm and 1 pm, and is willing to wait for 10 minutes.

- Q: What is the probability that they meet?

Recall the Beta distribution from Lecture 14. For $\alpha, \beta, \lambda > 0$,

suppose $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$

(α: shape, λ: rate)

and $X \perp\!\!\!\perp Y$. (Note Gamma$(1,\lambda)$ is the same as Exp(λ).)

Then, $X+Y \perp\!\!\!\perp \dfrac{X}{X+Y} =: Z$, $Z \sim \text{Beta}(\alpha, \beta)$

p.d.f. (Does not depend on λ)

$$f_Z(z) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1} \times \mathbb{1}\{z \in (0,1)\}$$

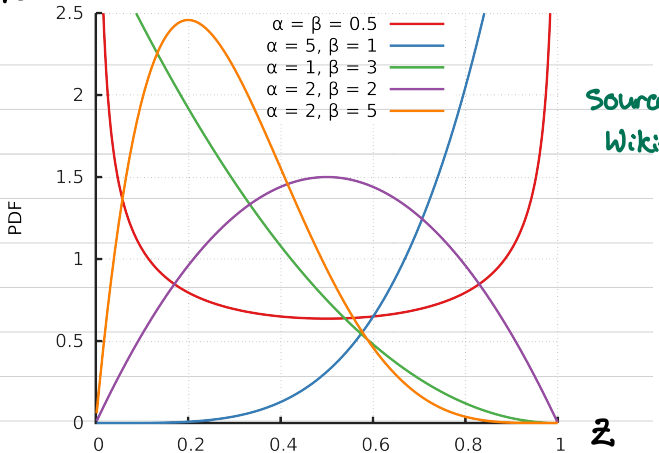
$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt.$$ Gamma function

For $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$

$$\mathbb{E}[Z] = \frac{\alpha}{\alpha+\beta}.$$

A very useful family of continuous distributions on $[0,1]$.

Application: Bayesian Inference. (Later)



Source: Wikipedia

$f_Z(z)$ can take on quite different shapes for different α, β values.

Generalization: $X_i \sim \text{Gamma}(\alpha_i, \lambda)$, $i = 1, \dots, n$

$X_1, \dots, X_n \perp\!\!\!\perp$.

Then, $(X_1 + \cdots + X_n) \perp\!\!\!\perp \dfrac{(X_1, \dots, X_n)}{(X_1 + \cdots + X_n)}$ and

$$\frac{(X_1, \dots, X_n)}{X_1 + \cdots + X_n} \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_n)$$ with p.d.f.

$$f(x_1, \dots, x_n) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)} x_1^{\alpha_1 - 1} \cdots x_n^{\alpha_n - 1} \mathbb{1}\{\sum_{i=1}^n x_i = 1\} \times \prod_{i=1}^n \mathbb{1}\{x_i > 0\}$$

Order Statistics:

$(x_1, x_2, \ldots, x_n)$ = a list of real numbers.

The order statistics of $(x_1, \ldots, x_n)$ is a permutation of the list s.t.

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)},$$

where $x_{(j)} = x_{\pi(j)}$ for some permutation π of $\{1, \ldots, n\}$.

If $x_1, \ldots, x_n$ are all distinct, then

$x_{(j)}$ = the j^{th} smallest element of $\{x_1, \ldots, x_n\}$.

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} F$, Some continuous distribution

Let $f(x) = \frac{dF(x)}{dx}$ denote the p.d.f.

Claim The p.d.f. of $X_{(1)} = \min\{X_1, \ldots, X_n\}$ is given by $f_{X_{(1)}}(x) = n\, f(x)\,[1 - F(x)]^{n-1}$.

Pf $\mathbb{P}[X_{(1)} > x] = \mathbb{P}[\bigcap_{i=1}^{n} (X_i \ge x)]$

$\Downarrow \Rightarrow$ $= \prod_{i=1}^{n} \mathbb{P}[X_i \ge x] = [1 - F(x)]^n$

$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = \frac{d}{dx}(1 - \mathbb{P}[X_{(1)} > x]) = n[1-F(x)]^{n-1} \frac{dF(x)}{dx}$

Claim The p.d.f. of $X_{(n)} = \max\{X_1, \ldots, X_n\}$ is given by $f_{X_{(n)}}(x) = n\, f(x)\,[F(x)]^{n-1}$.

Pf $F_{X_{(n)}}(x) = \mathbb{P}[X_{(n)} \le x] = \mathbb{P}[\bigcap_{i=1}^{n} (X_i \le x)]$

$= [F(x)]^n$

$f_{X_{(n)}} = \frac{dF_{X_{(n)}}(x)}{dx} = n[F(x)]^{n-1} \frac{dF(x)}{dx}$ □.

Claim The p.d.f. of the j^{th} order statistic $X_{(j)}$ is given by

$$f_{X_{(j)}}(x) = n \binom{n-1}{j-1} f(x) [F(x)]^{j-1} [1-F(x)]^{n-j}$$

for $j = 1, \ldots, n$.

Pf

The event $(X_{(j)} \le x)$ corresponds to at least j of $\{X_1, \ldots, X_n\}$ falling in $(-\infty, x]$

$F_{X_{(j)}}(x) =$

$$\mathbb{P}[X_{(j)} \le x] = \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1-F(x)]^{n-k}$$

$$f_{X_{(j)}}(x) = \frac{dF_{X_{(j)}}(x)}{dx}$$

$$= \sum_{k=j}^{n} \binom{n}{k} f(x) \left\{ k [F(x)]^{k-1} [1-F(x)]^{n-k} - (n-k)[F(x)]^{k} [1-F(x)]^{n-k-1} \right\}$$

$$= n \binom{n-1}{j-1} f(x) [F(x)]^{j-1} [1-F(x)]^{n-j}$$ by telescoping.

A more intuitive derivation.

$$f_{X_{(j)}}(x) = \lim_{\delta \downarrow 0} \frac{\mathbb{P}[x < X_{(j)} < x+\delta]}{\delta} \quad (*)$$

The leading contribution to $\mathbb{P}[x < X_{(j)} < x+\delta]$ is given by the following event:

$$\binom{n-1}{j-1} [F(x)]^{j-1} \; n f(x) \delta \; [1-F(x+\delta)]^{n-j}$$

n choices for the point falling in $(x, x+\delta)$

Events with two or more points in $(x, x+\delta)$ are of higher order in δ and hence do not contribute to $(*)$.

Uniform order statistics.

Suppose $U_1, \dots, U_n \overset{iid}{\sim} \text{Uniform}(0,1)$.

Then, $f(x) = 1\{x \in (0,1)\}$ and $F(x) = x$.

$$\Rightarrow f_{U_{(j)}}(u) = n \binom{n-1}{j-1} u^{j-1} (1-u)^{n-j}$$

$$= \frac{n!}{(j-1)!\,(n-j)!} u^{j-1} (1-u)^{(n-j+1)-1}$$

$$= \frac{\Gamma(n+1)}{\Gamma(j)\,\Gamma(n-j+1)} u^{j-1} (1-u)^{(n-j+1)-1}$$

$$\Rightarrow U_{(j)} \sim \text{Beta}(j, n-j+1)$$

General order statistics.

Recall from Lecture 12:

Def (Quantile Function) Given a RV X with c.d.f. F_X, $q_X(u) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq u\}$ for $u \in (0,1)$.

Furthermore, for $U \sim \text{Uniform}(0,1)$, $q_X(U) \overset{d}{=} X$.

$\Rightarrow$ For general $X_1, \dots, X_n \overset{iid}{\sim} F_X$,

$X_{(j)} \overset{d}{=} q_X(U_{(j)})$, where $U_{(j)} \sim \text{Beta}(j, n-j+1)$ for $j = 1, \dots, n$.

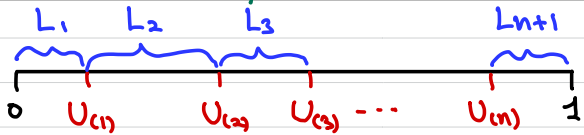
Joint Distribution

Thm Suppose $(X_1, \dots, X_n)$ is exchangeable with joint density $f(x_1, \dots, x_n)$ (which is symmetric in $x_1, \dots, x_n$). Then, the joint density of their order statistics $(X_{(1)}, \dots, X_{(n)})$ is

$$f_{(X_{(1)}, \dots, X_{(n)})}(x_1, \dots, x_n) = n! \, f(x_1, \dots, x_n) \, \mathbb{1}\{x_1 < x_2 < \cdots < x_n\}.$$

Pf See Gut Chapter 4.3.

Gaps:



$L_1 = U_{(1)}$ Transformation of variables.

$L_j = U_{(j)} - U_{(j-1)}$, $j = 2, \dots, n$ $\quad |\det J| = 1$

$L_{n+1} = 1 - U_{(n)}$

$$f_{(L_1, \dots, L_n)}(l_1, \dots, l_n) = n! \, \mathbb{1}\{l_1 + \cdots + l_n < 1\} \prod_{i=1}^{n} \mathbb{1}\{l_i > 0\}$$

RHS = the joint density of $\dfrac{(W_1, \dots, W_n)}{W_1 + \cdots + W_{n+1}}$

where $W_1, \dots, W_{n+1} \overset{iid}{\sim} \text{Exp}(\lambda) \equiv \text{Gamma}(1, \lambda)$

$$(L_1, L_2, \dots, L_{n+1}) \overset{d}{=} \frac{(W_1, \dots, W_{n+1})}{(W_1 + \cdots + W_{n+1})}.$$

$(L_1, L_2, \dots, L_{n+1}) \sim \text{Dirichlet}(1, \dots, 1)$

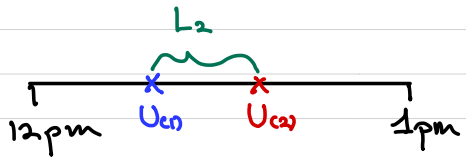
$$f(l_1, \dots, l_{n+1}) = n! \, \mathbb{1}\{\textstyle\sum_{i=1}^{n+1} l_i = 1\} \prod_{i=1}^{n} \mathbb{1}\{l_i > 0\}$$

The gaps are **not independent**, but $(L_1, \dots, L_{n+1})$ is **exchangeable**.

$\Rightarrow L_1 \overset{d}{=} L_2 \overset{d}{=} \cdots \overset{d}{=} L_{n+1}$

$L_1 = X_{(1)} \sim \text{Beta}(1, n)$ $\qquad \forall\, i = 1, \dots, n+1, \quad \mathbb{E}[L_i] = \dfrac{1}{1+n}$

Back to the Alice & Bob problem.



$n = 2$

10 min $= \frac{1}{6}$ hr

$L_2 \overset{d}{=} L_1 = X_{(1)}$

$$\mathbb{P}[L_2 \le \tfrac{1}{6}] = \int_0^{\frac{1}{6}} \frac{\Gamma(1+2)}{\Gamma(1)\Gamma(2)} (1-u) \, du$$

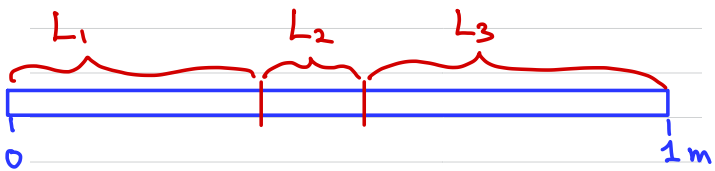
$$= 2 \frac{(1-u)^2}{-2} \Big|_0^{\frac{1}{6}}$$

$$= -\left(\frac{5}{6}\right)^2 + 1 = \boxed{\frac{11}{36}}$$

Lecture 16

Problem of the day

Sample two points independently and u.a.r. from a meter stick, thereby obtaining 3 segments.



Q: What is the probability that the three segments can form a triangle?

Lecture 16

Linear Model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

Y: response (e.g., disease risk) $\quad \varepsilon \sim N(0, \sigma^2)$

$\vec{X} = (1, X_1, \ldots, X_{p-1})$ = covariates or features (e.g., age, BMI, genetic data)

$\vec{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ = unknown coefficients

Training Data: $\{(\vec{X}^{(i)}, Y^{(i)}), i = 1, \ldots, n\}$

$$\underset{n\times 1}{\vec{Y}} = \begin{bmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{bmatrix} \qquad \underset{n\times p}{\mathbb{X}} = \begin{bmatrix} \vec{X}^{(1)} \\ \vdots \\ \vec{X}^{(n)} \end{bmatrix}$$

Design matrix

Assume $n \gg p$.

If $(\mathbb{X}^T\mathbb{X})$ is non-singular, then the MLE or least-squares estimate of $\vec{\beta}$ is

$$\hat{\vec{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\vec{Y}$$

Common Hypothesis Testing Task:

Null H_0: $\beta_j = 0$

Alternative H_1: $\beta_j \neq 0$

Key Theorem: Under H_0,

$$S := \frac{\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \Big/ \sqrt{\|\vec{e}\|^2/[(n-p)\sigma^2]} \sim t_{n-p}$$

t-distribution with $n-p$ degrees of freedom

where $\vec{e} = \vec{Y} - \mathbb{X}\hat{\vec{\beta}}$ is the residual.

Facts:

1) Under H_0, $\dfrac{\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \sim N(0,1)$.

2) $\dfrac{\|\vec{e}\|^2}{\sigma^2} \sim \chi^2_{n-p}$.

3) $\|\vec{e}\|^2 \perp\!\!\!\perp \hat{\vec{\beta}}$.

Today we will show that these facts imply $S \sim t_{n-p}$.

Conditional densities of continuous RVs

X, Y cts RVs with joint density $f_{X,Y}$. For a given measurable set $A \in \mathcal{F}$, we want to find

$$\mathbb{P}[X \in A \mid Y = y_0].$$

Subtlety: $\mathbb{P}[Y = y] = 0, \ \forall\, y \in \mathbb{R}$.

For $\delta \ll 1$,

$$\mathbb{P}[X \in A \mid y_0 < Y < y_0 + \delta] = \frac{\mathbb{P}[X \in A, \ (y_0 < Y < y_0 + \delta)]}{\mathbb{P}[y_0 < Y < y_0 + \delta]}$$

$$= \frac{\int_A \int_{y_0}^{y_0+\delta} f_{X,Y}(x,y)\, dy\, dx}{\int_{y_0}^{y_0+\delta} f_Y(y)\, dy} \approx \frac{\int_A f_{X,Y}(x, y_0)\, \delta\, dx}{f_Y(y_0)\, \delta}$$

$$= \int_A \boxed{\frac{f_{X,Y}(x,y_0)}{f_Y(y_0)}}\, dx$$

Conditional density of X given $Y = y_0$, denoted $f_{X|Y=y_0}(x) = \dfrac{f_{X,Y}(x,y_0)}{f_Y(y_0)}$

$f_{X|Y=y_0}(x)$ is well defined as long as $f_Y(y_0) > 0$.

Independence:

$X \perp\!\!\!\perp Y \iff f_{X|Y=y}(x) = f_X(x)$, $\forall\, x \in \mathbb{R}$ and $\forall\, y \in \mathbb{R}$ s.t. $f_Y(y) > 0$.

$\iff f_{X,Y}(x,y) = f_X(x) f_Y(y)$ $\forall\, x, y \in \mathbb{R}$.

Law of Total Probability:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

$$= \int_{-\infty}^{\infty} f_{X|Y=y}(x)\, f_Y(y)\, dy$$

Application 1:

$X, Y \overset{iid}{\sim} N(0,1)$

Let $R = \frac{X}{Y}$.

$$f_R(r) = \int_{-\infty}^{\infty} f_{R|Y=y}(r)\, f_Y(y)\, dy$$

$$(R \mid Y=y) \overset{d}{=} \left(\frac{X}{y} \,\middle|\, Y=y\right) \overset{d}{=} \frac{X}{y}$$

by $\perp\!\!\!\perp$ of X, Y

$T(x) = \frac{x}{y}$ is an invertible and differentiable transformation.

Recalling Lecture 12-2, we obtain

$$f_{\frac{X}{y}}(r) = f_X(ry)\left|\frac{d(ry)}{dr}\right| = |y|\, f_X(ry)$$

$$f_R(r) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(ry)^2}{2}} |y| \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$$

$$= \frac{2}{2\pi}\int_0^{\infty} y\, e^{-\frac{y^2}{2}(1+r^2)}\, dy = \frac{1}{\pi(1+r^2)}$$

Cauchy!

Application 2:

$X, Y_1, \cdots, Y_k \overset{iid}{\sim} N(0,1)$

$$R = \frac{X}{\sqrt{\frac{Y_1^2 + \cdots + Y_k^2}{k}}}$$

Recall: $Y_1^2 + \cdots + Y_k^2 \sim \text{Gamma}\left(\frac{k}{2}, \frac{1}{2}\right)$ a.k.a. χ^2_k

Define $G = Y_1^2 + \cdots + Y_k^2$.

$$f_R(r) = \int_0^{\infty} f_{R|G=g}(r)\, f_G(g)\, dg$$

$$(R \mid G=g) \overset{d}{=} \left(\frac{X}{\sqrt{\frac{g}{k}}} \,\middle|\, G=g\right) \overset{d}{=} \frac{X}{\sqrt{g/k}}$$

by $\perp\!\!\!\perp$ of X, G

$$f_{\frac{X}{\sqrt{g/k}}}(r) = f_X(r\sqrt{g/k})\left|\frac{d(r\sqrt{g/k})}{dr}\right| = \sqrt{g/k}\, f_X(r\sqrt{g/k})$$

$$f_R(r) = \int_0^{\infty} \sqrt{\frac{g}{k}}\, f_X(r\sqrt{g/k})\, f_G(g)\, dg$$

$$= \int_0^{\infty} \sqrt{\frac{g}{k}} \frac{1}{\sqrt{2\pi}} e^{-\frac{g}{2k}r^2} \frac{\left(\frac{1}{2}\right)^{k/2}}{\Gamma\left(\frac{k}{2}\right)} g^{\frac{k}{2}-1} e^{-g/2}\, dg$$

$$= \frac{1}{\sqrt{2\pi k}} \frac{1}{2^{k/2}} \frac{1}{\Gamma\left(\frac{k}{2}\right)} \int_0^{\infty} g^{\frac{1}{2}(k+1)-1} e^{-\frac{g}{2}\left(\frac{r^2}{k}+1\right)}\, dg$$

Recall $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

$$f_R(r) = \frac{1}{\sqrt{2\pi K}} \frac{1}{2^{K/2}} \frac{1}{\Gamma(\frac{K}{2})} \Gamma\left(\frac{K+1}{2}\right) 2^{\frac{K+1}{2}} \frac{1}{\left(1+\frac{r^2}{K}\right)^{\frac{K+1}{2}}}$$

$$= \frac{\Gamma(\frac{K+1}{2})}{\Gamma(\frac{K}{2})} \frac{1}{\sqrt{\pi K}} \frac{1}{\left(1+\frac{r^2}{K}\right)^{\frac{(K+1)}{2}}}$$

density for t-distribution with K degrees of freedom.

If $K=1$, t_K is equivalent to Cauchy

What happens as $K \to \infty$?

$Y_1^2, Y_2^2, \ldots \overset{iid}{\sim} \text{Gamma}(\frac{1}{2}, \frac{1}{2})$

$\mathbb{E}[Y_i^2] = 1$

SLLN $\Rightarrow \frac{Y_1^2 + \cdots + Y_K^2}{K} \xrightarrow{a.s.} 1$ as $K \to \infty$.

$t_K \to N(0,1)$ as $K \to \infty$.

Bayes Rule for continuous RVs

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

$$= \frac{f_{X|Y=y}(x)\, f_Y(y)}{f_X(x)}$$

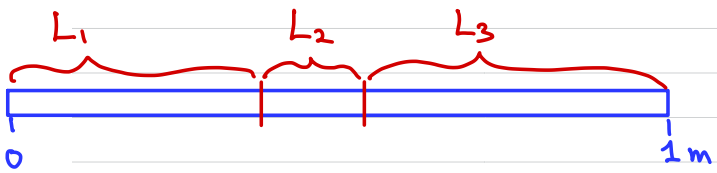
by Law of total probability $= \frac{f_{X|Y=y}(x)\, f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y=y}(x)\, f_Y(y)\, dy}$

Applications next week.

Lecture 17

Problem from Lecture 16

Sample two points independently and u.a.r. from a meter stick, thereby obtaining 3 segments.



Q: What is the probability that the three segments can form a triangle?

Can form a triangle $\Rightarrow$

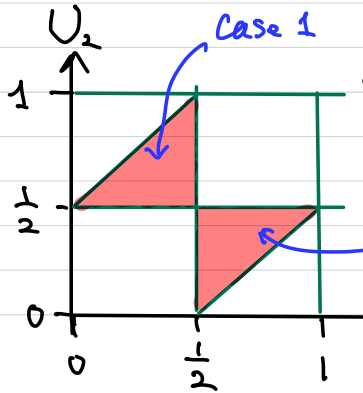
$L_1 < L_2 + L_3 = 1 - L_1 \Rightarrow L_1 < \frac{1}{2}$

$L_2 < L_1 + L_3 = 1 - L_2 \Rightarrow L_2 < \frac{1}{2}$

$L_3 < L_1 + L_2 = 1 - L_3 \Rightarrow L_3 < \frac{1}{2}$

Case 1 Need $U_1 < \frac{1}{2}$, $U_2 - U_1 < \frac{1}{2}$, $1 - U_2 < \frac{1}{2}$

Case 2 Need $U_2 < \frac{1}{2}$, $U_1 - U_2 < \frac{1}{2}$, $1 - U_1 < \frac{1}{2}$



$\mathbb{P}[(U_1, U_2) \in \text{Shaded region}] = \boxed{\frac{1}{4}}$

Problem of the day

Suppose $U_1, U_2, \ldots \overset{iid}{\sim} \text{Uniform}(0,1)$ and define

$$N = \min\{n \mid U_1 + \cdots + U_n > 1\}.$$

What is $\mathbb{E}[N]$?

Bayesian Inference

X = observed data
Θ = unknown parameter(s)
} continuous RVs

Law of Total Probability:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|\Theta=\theta}(x)\, f_\Theta(\theta)\, d\theta$$

Bayes Rule for continuous RVs

Likelihood → $f_{X|\Theta=\theta}(x)$, prior → $f_\Theta(\theta)$

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x)\, f_\Theta(\theta)}{\int_{-\infty}^{\infty} f_{X|\Theta=\theta'}(x)\, f_\Theta(\theta')\, d\theta'}$$

← posterior

If X or Θ is discrete, use probability mass function instead of p.d.f.

$$(X|\Theta=\theta) \sim \text{Binomial}(n,\theta)$$

$$\mathbb{P}[X=x|\Theta=\theta] = \binom{n}{x}\theta^x(1-\theta)^{n-x} \times 1\{x\in\{0,\dots,n\}\}$$

$$f_{\Theta|X=x}(\theta) \propto \mathbb{P}[X=x|\Theta=\theta]\, f_\Theta(\theta)$$

Suppose we want the prior and the posterior to have the same functional form. How should we choose $f_\Theta(\theta)$?

If $f_\Theta(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, then
$f_{\Theta|X=x}(\theta) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$

That is $\Theta \sim \text{Beta}(\alpha,\beta)$ ← Conjugate prior for the parameter of Binomial

$\Rightarrow (\Theta|X=x) \sim \text{Beta}(\alpha+x, \beta+n-x)$

---

d dimensions: $\vec{X} = (X_1,\dots,X_d)$
$\vec{\theta} = (\theta_1,\dots,\theta_d)$

$$(\vec{X}|\vec{\Theta}=\vec{\theta}) \sim \text{Multinomial}(n,\theta_1,\dots,\theta_d)$$

$$\mathbb{P}[\vec{X}=(x_1,\dots,x_d)|\vec{\Theta}=\vec{\theta}] = \binom{n}{x_1,\dots,x_d}\theta_1^{x_1}\cdots\theta_d^{x_d} \times 1\{x_1+\cdots+x_d=n\}\prod_{i=1}^{d} 1\{x_i\in\{0,\dots,n\}\}$$

Q: Conjugate prior for θ?

A: $f_{\vec{\Theta}}(\vec{\theta}) \propto \prod_{i=1}^{d}\theta_i^{\alpha_i-1}$ Dirichlet$(\alpha_1,\dots,\alpha_d)$

Gaussian $X_1, \dots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$

Case 1: μ unknown, σ^2 known

$\vec{X} = (X_1, \dots, X_n)$, M = mean

$$f_{\vec{X}|M=\mu}(x_1, \dots, x_n) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

Conjugate prior: $M \sim \text{Normal}(\mu_0, \sigma_0^2)$

$$f_M(\mu) \propto \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

$$f_{M|\vec{X}=\vec{x}}(\mu) \propto f_{\vec{X}|M=\mu}(\vec{x})\, f_M(\mu)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right\}$$

where $\mu_n = \left(\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\right)\mu_0 + \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\mu_{ML}$ ← convex combination

$\mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ← Maximum Likelihood Estimate

precision: $\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$

prior precision + data precision

Remarks:

1) $\mu_n \to \mu_{ML}$ as $n \to \infty$
2) Precisions are additive.
3) Precision ↑ as n ↑. less informative prior
4) For a finite n, if $\sigma_0^2 \to \infty$, then $\mu_n \to \mu_{ML}$ and $\sigma_n^2 \to \frac{\sigma^2}{n}$

Case 2: μ Known, σ^2 unknown.

More convenient to work with precision $\Lambda = \frac{1}{\sigma^2}$

$$f_{\vec{X}|\Lambda=\lambda}(x_1, \dots, x_n) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

Conjugate prior: $\Lambda \sim \text{Gamma}(\alpha_0, \beta_0)$

$$f_\Lambda(\lambda) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\lambda^{\alpha_0-1}e^{-\beta_0\lambda}$$

$$f_{\Lambda|\vec{X}=\vec{x}}(\lambda) \propto f_{\vec{X}|\Lambda=\lambda}(\vec{x})\, f_\Lambda(\lambda)$$

$$\propto \lambda^{\alpha_0+\frac{n}{2}-1}\exp\left\{-\lambda\left[\beta_0 + \underbrace{\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\frac{n}{2}\sigma_{ML}^2}\right]\right\}$$

$$\sigma_{ML}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

$(\Lambda \mid \vec{X}=\vec{x}) \sim \text{Gamma}(\alpha_n, \beta_n)$

$\alpha_n = \alpha_0 + \frac{n}{2}$

$\beta_n = \beta_0 + \frac{n}{2}\sigma^2_{ML}$

Case 3: both μ and σ^2 unknown.

$f_{\vec{X} \mid M=\mu, \Lambda=\lambda}(\vec{x})$

$\propto \left[\lambda^{\frac{1}{2}} e^{-\frac{\lambda\mu^2}{2}}\right]^n \exp\left\{\lambda\mu \sum_{i=1}^n x_i - \frac{\lambda}{2}\sum_{i=1}^n x_i^2\right\}$

$f_{M,\Lambda}(\mu,\lambda) = f_{M \mid \Lambda=\lambda}(\mu)\, f_\Lambda(\lambda)$

$\uparrow$ Normal $(\mu_0, \frac{1}{c\lambda})$ $\qquad$ $\uparrow$ Gamma(α, β)

where

$\mu_0 = \frac{a}{c}$

$\alpha = 1 + \frac{c}{2}$

$\beta = b - \frac{a^2}{2c}$

} a, b, c are constants s.t. $\alpha, \beta, c > 0$

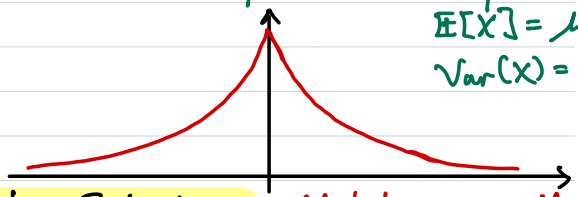
Double exponential

$X \sim \text{Laplace}(\mu, \beta)$, $\beta > 0$

$f_X(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right)$

$\mathbb{E}[X] = \mu$

$\text{Var}(X) = 2\beta^2$



Model Selection Model 0 vs. Model 1

$\Theta \sim \text{Bernoulli}(\frac{1}{2})$ ← prior

Prior odds $= \frac{\mathbb{P}[\Theta=0]}{\mathbb{P}[\Theta=1]}$

$X_1, \ldots, X_n \mid \Theta=0 \overset{iid}{\sim} \text{Normal}(0,1)$

pdf $f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$X_1, \ldots, X_n \mid \Theta=1 \overset{iid}{\sim} \text{Laplace}(0, \sqrt{\frac{\pi}{2}})$

pdf $f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-|x|\sqrt{\frac{2}{\pi}}}$

Bayes Factor $= \frac{f_{\vec{X} \mid \Theta=0}(\vec{x})}{f_{\vec{X} \mid \Theta=1}(\vec{x})}$ ← Model Evidence a.k.a. Marginal Likelihood

Posterior odds = BF × Prior odds

Data sampled from $N(0,1)$:

$\{-0.7, -0.5, 0.18, 0.35, 0.66\}$

$$BF = \frac{f_{\vec{X}|\Theta=0}(\vec{x})}{f_{\vec{X}|\Theta=1}(\vec{x})} = 3.46$$

Additional data pt: $X_6 = 4$

$$BF = \frac{f_{\vec{X}|\Theta=0}(\vec{x})}{f_{\vec{X}|\Theta=1}(\vec{x})} = 0.028$$

Now suppose

$X_1, \ldots, X_n | \Theta=0 \overset{iid}{\sim}$ Normal $(0, \alpha^2)$, $\alpha > 0$ unknown

$X_1, \ldots, X_n | \Theta=1 \overset{iid}{\sim}$ Laplace $(0, \beta)$, $\beta > 0$ unknown

Put priors on α and β.

RVs A B

Example:

$\log A \mid \Theta=0 \sim$ Uniform$(-c, c)$

$\log B \mid \Theta=1 \sim$ Uniform$(-c, c)$

$$f_{\vec{X}|\Theta=0}(\vec{x}) = \int_0^\infty f_{\vec{X}|\Theta=0, A=\alpha}(\vec{x})\, f_{A|\Theta=0}(\alpha)\, d\alpha$$

$Z = \log A$, $A = T(Z) = e^Z$

$$f_{A|\Theta=0}(\alpha) = f_{\log A|\Theta=0}(\log \alpha) \left| \frac{d}{d\alpha} \log \alpha \right|$$

$$= \frac{1}{\alpha} f_{\log A|\Theta=0}(\log \alpha)$$

$$= \frac{1\{-c < \log \alpha < c\}}{2c\alpha}$$

$$f_{\vec{X}|\Theta=1}(\vec{x}) = \int_0^\infty f_{\vec{X}|\Theta=1, B=\beta}(\vec{x})\, f_{B|\Theta=1}(\beta)\, d\beta$$

$$f_{B|\Theta=1}(\beta) = \frac{1\{-c < \log \beta < c\}}{2c\beta}$$

$$BF = \frac{f_{\vec{X}|\Theta=0}(\vec{x})}{f_{\vec{X}|\Theta=1}(\vec{x})}$$

← c cancels out in the Bayes Factor.

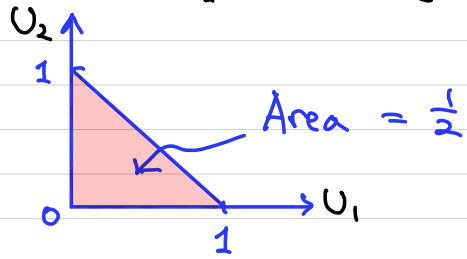
Problem from Lecture 17

Suppose $U_1, U_2, \ldots \overset{iid}{\sim} \text{Uniform}(0,1)$ and define

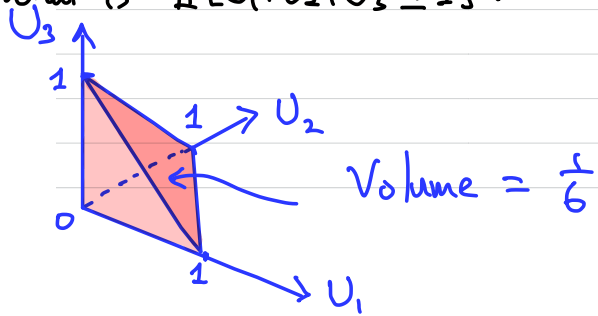
$$N = \min\{n \mid U_1 + \cdots + U_n > 1\}.$$

What is $\mathbb{E}[N]$?

What is $\mathbb{P}[U_1 + U_2 \le 1]$?



What is $\mathbb{P}[U_1 + U_2 + U_3 \le 1]$?



Consider the following polytope in $\mathbb{R}^n$

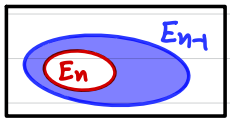
$$\Delta_n = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \begin{array}{l} 0 \le u_i \le 1 \quad \forall\, i \text{ and} \\ u_1 + \cdots + u_n \le 1 \end{array} \right\}$$

$$\text{Vol}(\Delta_n) = \frac{1}{n!}$$

$$\mathbb{P}[U_1 + \cdots + U_n \le 1] = \frac{1}{n!}$$

Define E_k = event $U_1 + \cdots + U_k \le 1$

$$\mathbb{P}[N = n] = \mathbb{P}[E_{n-1} \cap E_n^c]$$



$$= \mathbb{P}[E_{n-1}] - \mathbb{P}[E_{n-1} \cap E_n]$$

$E_n \subset E_{n-1} \Rightarrow E_n$

$$= \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}$$

$$\mathbb{E}[N] = \sum_{n=2}^{\infty} n\,\mathbb{P}[N=n] = \sum_{n=2}^{\infty} \frac{1}{(n-2)!}$$

$$= \boxed{e^1}$$

Lecture 18

Conditional Expectation

- Defining conditional expectation in full generality is beyond the scope of Stat201A.
- We will assume that joint density $f_{X,Y}(x,y)$ of a pair of RVs (X,Y) exists.

Def If Y is continuous,

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y\, f_{Y|X=x}(y)\, dy.$$

More generally, for suitable $g: \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}[g(Y)|X=x] = \int_{-\infty}^{\infty} g(y)\, f_{Y|X=x}(y)\, dy,$$

and for suitable $h: \mathbb{R}^2 \to \mathbb{R}$

$$\mathbb{E}[h(X,Y)|X=x] = \mathbb{E}[h(x,Y)|X=x] = \int_{-\infty}^{\infty} h(x,y)\, f_{Y|X=x}(y)\, dy.$$

If Y is discrete,

$$\mathbb{E}[Y|X=x] = \sum_y y\, \mathbb{P}[Y=y|X=x].$$

and so on.

e.g. Die 0 fair, Die 1 loaded

Y = # shown, X = Die chosen

$\mathbb{P}[Y=6|X=1] = p$

$\mathbb{P}[Y=j|X=1] = \frac{1-p}{5}$, $j = 1,\dots,5$

$$\mathbb{E}[Y|X=0] = \sum_{j=1}^{6} j\left(\frac{1}{6}\right) = \frac{7}{2}$$

$$\mathbb{E}[Y|X=1] = 6p + \sum_{j=1}^{5} j\left(\frac{1-p}{5}\right) = 3(1+p).$$

Remarks:

1) For a fixed x, $\mathbb{E}[Y|X=x]$ satisfies the usual properties of expectation. (e.g., linearity)

2) $\mathbb{E}[Y|X=x] = \psi(x)$ is a function of x.

3) $\mathbb{E}[Y|X] = \psi(X)$ is a RV.

$\psi(X): \Omega \to \mathbb{R}$.

$[\psi(X)](\omega) = \psi(X(\omega))$

Thm (Law of Total Expectation / Law of Iterated Expectation / Tower Property)

For any RV Y s.t. $\mathbb{E}[|Y|] < \infty$,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]].$$

(outer $\mathbb{E}$: expectation over X; inner $\mathbb{E}$: expectation over Y)

Pf

$$\mathbb{E}[\mathbb{E}[Y|X]] = \int \mathbb{E}[Y|X=x] f_X(x)\,dx$$

$$= \int \left[\int y\, f_{Y|X=x}(y)\,dy\right] f_X(x)\,dx$$

Interchange the order of integration

$$= \int y \left[\int f_{Y|X=x}(y) f_X(x)\,dx\right] dy$$

Law of total probability $\Rightarrow$

$$= \int y\, f_Y(y)\,dy = \mathbb{E}[Y].$$

□

In the discrete case,

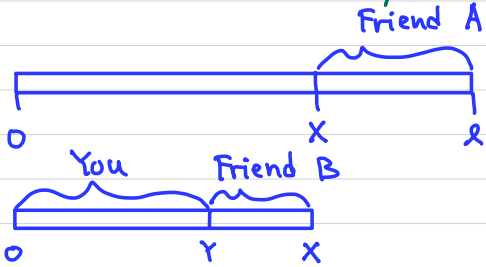
$$\mathbb{E}[\mathbb{E}[Y|X]] = \sum_x \mathbb{E}[Y|X=x]\,\mathbb{P}[X=x].$$

e.g.) In the two dice example earlier, suppose $X \sim$ Bernoulli($\frac{1}{2}$)

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \frac{7}{2}(1-\tfrac{1}{2}) + 3(1+p)\tfrac{1}{2}$$

---

e.g.) Suppose you get a long candy bar of length ℓ for Halloween and you share it with two of your friends as follows.



How much candy do expect to get for yourself?

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}\left[\frac{X}{2}\right] = \frac{1}{2}\left(\frac{\ell}{2}\right) = \frac{\ell}{4}$$

Thm (Wald's Identity)

Suppose $X_1, X_2, \ldots$ is a sequence of iid RVs with $\mathbb{E}[X_i] = \mu < \infty$, and N is a positive integer valued RV s.t. $N \perp\!\!\!\perp X_1, X_2, \ldots$ and $\mathbb{E}[N] < \infty$.

Let $S_N = X_1 + \cdots + X_N$. Then

$$\mathbb{E}[S_N] = \mu\, \mathbb{E}[N].$$

Pf

$$\mathbb{E}[S_N] = \mathbb{E}[\mathbb{E}[S_N \mid N]]$$

μN since $N \perp\!\!\!\perp X_1, X_2, \ldots$

$$= \mathbb{E}[\mu N] = \mu\, \mathbb{E}[N].$$

Statistical Risk Minimization

Y, a RV of interest.

$g(X)$, prediction of Y

$L(Y, g(X))$ Loss function

$R(g) = \mathbb{E}[L(Y, g(X))]$ Risk.

↑ expectation over the joint distribution of X and Y.

Goal: Find $g^* = \underset{g}{\operatorname{argmin}}\, R(g)$

MSE

Example: (Mean Squared Error Minimizer)

$L(Y, g(X)) = (Y - g(X))^2$

$$R(g) = \mathbb{E}[(Y - g(X))^2] = \mathbb{E}[\mathbb{E}[(Y - g(X))^2 \mid X]]$$

↑ expectation over X, Y ↑ expectation over X

Consider $h(c) = \mathbb{E}[(Z - c)^2]$ where Z is a RV and c is a const.

Then, $h(c) = \mathbb{E}[Z^2] - 2c\,\mathbb{E}[Z] + c^2$

$h'(c) = -2\,\mathbb{E}[Z] + 2c$, which vanishes at $c = \mathbb{E}[Z]$

$h''(c) = 2$

$\Rightarrow c^* = \underset{c}{\operatorname{argmin}}\, h(c) = \mathbb{E}[Z]$

$\Rightarrow g^*(X) = \mathbb{E}[Y \mid X]$

Regression is about modeling this conditional expectation.

Def (Median) For a RV X, a median of the distribution of X is any value m s.t.

$$\mathbb{P}[X \le m] \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}[X \ge m] \ge \frac{1}{2}.$$

Not =

Remarks:

1) Every distribution has at least one median
2) Median may not be unique.

e.g. (Unique median)

x	1	2	3	4
$\mathbb{P}[X=x]$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

$\mathbb{P}[X \le 3] = \frac{6}{10}$, $\mathbb{P}[X \ge 3] = \frac{7}{10}$

So, 3 is a median.

For any $\varepsilon > 0$, $\mathbb{P}[X \ge 3+\varepsilon] < \frac{1}{2}$, $\mathbb{P}[X \le 3-\varepsilon] < \frac{1}{2}$

$\Rightarrow$ 3 is the unique median.

e.g. (Infinitely many medians)

x	1	2	3	4
$\mathbb{P}[X=x]$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$

Any $m \in [2,3]$ is a median.

MAE

Thm (Mean Absolute Error Minimizer) Let Z be a RV with a finite median m. Then, m minizes $h(c) = \mathbb{E}[|Z-c|]$.

Back to Risk minimization.

$$L(Y, g(X)) = |Y - g(X)|$$
$$R(g) = \mathbb{E}[|Y-g(X)|]$$

Any function g s.t. $g(x)$ is a conditional median of Y given $X=x$ minimizes $R(g)$

Conditional Variance.

$$\mathrm{Var}(Y|X=x) = \mathbb{E}\left[(Y-\mathbb{E}[Y|X=x])^2 \,\middle|\, X=x\right]$$

$$= \mathbb{E}[Y^2|X=x] - \left(\mathbb{E}[Y|X=x]\right)^2$$

Claim (Law of Total Variance)

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$$

- $\mathbb{E}$: expectation over X
- $\mathrm{Var}(Y|X)$: conditional variance of Y given X
- $\mathbb{E}[Y|X]$: conditional expectation of Y given X.
- Var: variance over X

Pf

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$$

Tower Property $\Rightarrow$

$$= \mathbb{E}[\mathbb{E}[Y^2|X]] - (\mathbb{E}[\mathbb{E}[Y|X]])^2$$

$$= \mathbb{E}[\mathbb{E}[Y^2|X]] - \mathbb{E}[(\mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X])^2] - (\mathbb{E}[\mathbb{E}[Y|X]])^2$$

$$\mathbb{E}[(\mathbb{E}[Y|X])^2] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 = \mathrm{Var}(\mathbb{E}[Y|X])$$

linearity of $\mathbb{E}$ $\Rightarrow$

$$= \mathbb{E}\left[\mathbb{E}[Y^2|X] - (\mathbb{E}[Y|X])^2\right] + \mathrm{Var}(\mathbb{E}[Y|X])$$

$$= \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$$

Application

Suppose $X_1, X_2, \ldots$ is a sequence of iid RVs with $\mathbb{E}[X_i] = \mu < \infty$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, and N is a positive integer valued RV s.t. $N \perp\!\!\!\perp X_1, X_2, \ldots$ and $\mathbb{E}[N] < \infty$.

Let $S_N = X_1 + \cdots + X_N$. Then

$$\mathrm{Var}(S_N) = \mathbb{E}[\mathrm{Var}(S_N|N)] + \mathrm{Var}(\mathbb{E}[S_N|N])$$

$X_1, X_2, \ldots$ iid and $N \perp\!\!\!\perp X_1, X_2, \ldots$ $\Rightarrow$ $\mathrm{Var}(S_N|N) = N\sigma^2$, $\mathbb{E}[S_N|N] = \mu N$

$$= \sigma^2 \mathbb{E}[N] + \mu^2 \mathrm{Var}(N)$$

Suppose $Z \sim N(0,1)$. Then, for fixed constants $a, b \in \mathbb{R}$,

$X = aZ + \mu \sim N(\mu, a^2)$.

Def (Jointly-Normal RVs or Multivariate Normal)

$\{X_1, \dots, X_n\}$ is called a set of jointly-normal RVs if

n-variate normal

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = A \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$$

where $Z_1, \dots, Z_m \overset{iid}{\sim} N(0,1)$, and $A \in \mathbb{R}^{n \times m}$ and $(\mu_1, \dots, \mu_n)^T \in \mathbb{R}^n$ are constant matrices.

Remark

1) Each X_i is a normal RV.
2) Being jointly-normal is more strict than just being normal marginally. Each $(X_i - \mu_i)$ has to be a linear combination of the same set of iid $N(0,1)$ RVs.

Q Does $Z_1, Z_2 \sim N(0,1) \Rightarrow Z_1 + Z_2$ normal?

No, not in general. $\perp\!\!\!\perp$ is important.

e.g.) Suppose $Z_1 \sim N(0,1)$ and $Z_1 \perp\!\!\!\perp X$ with $\mathbb{P}[X=+1] = \mathbb{P}[X=-1] = \frac{1}{2}$.

Define $Z_2 = XZ_1$.

For any $a \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}[Z_2 \le a] &= \tfrac{1}{2}\left(\mathbb{P}[Z_2 \le a \mid X=1] + \mathbb{P}[Z_2 \le a \mid X=-1]\right) \\ &= \tfrac{1}{2}\left(\mathbb{P}[Z_1 \le a] + \mathbb{P}[Z_1 \ge -a]\right) \\ &= \mathbb{P}[Z_1 \le a]. \end{aligned}$$

$\Rightarrow Z_2 \sim N(0,1)$.

However, $Z_1 + Z_2$ is not normal.

Also, $\mathbb{E}[Z_1 Z_2] = \mathbb{E}[X Z_1^2] = \mathbb{E}[X]\,\mathbb{E}[Z_1^2] = 0$ ($\mathbb{E}[X] = 0$)

$\Rightarrow \mathrm{Cov}(Z_1, Z_2) = 0$.

However, $Z_1 \not\perp\!\!\!\perp Z_2$.

Covariance Matrices

Random Vector $\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$

$$\mathrm{Cov}(\vec{Y}) = \mathbb{E}[\vec{Y}\vec{Y}^T] - \mathbb{E}[\vec{Y}]\,\mathbb{E}[\vec{Y}]^T$$
$$= \text{an } n\times n \text{ matrix whose } (i,j)^{\text{th}} \text{ entry is } \mathrm{Cov}(Y_i, Y_j).$$

Note:

1) $[\mathrm{Cov}(\vec{Y})]^T = \mathrm{Cov}(\vec{Y})$

i.e., $\mathrm{Cov}(\vec{Y})$ is symmetric since $\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(Y_j, Y_i)$

2) $\mathrm{Cov}(A\vec{Y}+\vec{b})$
$= \mathbb{E}[(A\vec{Y}+\vec{b})(A\vec{Y}+\vec{b})^T] - \mathbb{E}[A\vec{Y}+\vec{b}]\,\mathbb{E}[(A\vec{Y}+\vec{b})^T]$
$= A\,\mathbb{E}[\vec{Y}\vec{Y}^T]A^T - A\,\mathbb{E}[\vec{Y}]\,\mathbb{E}[\vec{Y}^T]A^T$
$= A\,\mathrm{Cov}(\vec{Y})\,A^T$

Let $Z_1, \dots, Z_n \overset{iid}{\sim} N(0,1)$

$\vec{Z} = (Z_1, \dots, Z_n)^T$

Let $\underset{n\times 1}{\vec{X}} = A\underset{n\times 1}{\vec{Z}} + \vec{\mu}$, and A is invertible

$A \in \mathbb{R}^{n\times n}$, $\vec{\mu} \in \mathbb{R}^{n\times 1}$ fixed.

$$\mathbb{E}[\vec{X}] = A\,\mathbb{E}[\vec{Z}] + \vec{\mu} = \vec{\mu} \quad (\mathbb{E}[\vec{Z}] = \vec{0})$$
$$\mathrm{Cov}(\vec{X}) = A\,\underbrace{\mathrm{Cov}(\vec{Z})}_{I}\,A^T = AA^T =: \Sigma$$

Then,

$$f_{\vec{X}}(\vec{x}) = f_{\vec{Z}}(A^{-1}(\vec{x}-\vec{\mu}))\,|\det(A^{-1})|$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}[A^{-1}(\vec{x}-\vec{\mu})]^T[A^{-1}(\vec{x}-\vec{\mu})]} \cdot |\det(A^{-1})|$$
$$\left((A^{-1})^T = (A^T)^{-1}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T(AA^T)^{-1}(\vec{x}-\vec{\mu})} \cdot |\det(A^{-1})|$$

$\det(\Sigma) = \det(AA^T) = \det(A)\det(A^T) = (\det(A))^2$

$\det(A^{-1}) = \frac{1}{\det(A)} = \frac{1}{\sqrt{\det(\Sigma)}}$

$$f_{\vec{X}}(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

$\vec{X} \sim N_n(\vec{\mu}, \Sigma)$, $\Sigma = AA^T$.

Bivariate Normal ($n=2$)

$$\mathrm{Var}(X_i) = \sigma_i^2$$

$$\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1) = \rho \sigma_1 \sigma_2$$

$-1 < \rho < +1, \quad \sigma_1 > 0, \quad \sigma_2 > 0$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

singular if $\rho = \pm 1$

$$\det(\Sigma) = \sigma_1^2 \sigma_2^2 (1-\rho^2)$$

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]}$$

See demonstrations.wolfram.com/TheBivariateNormalDistribution

$X_1 \perp\!\!\!\perp X_2 \iff \rho = 0$

(Not true for general RVs)

Theorem (Maxwell)

Let X and Y be $\perp\!\!\!\perp$ RVs with finite variance and define $\begin{bmatrix} X_\theta \\ Y_\theta \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$.

Then,

$X_\theta \perp\!\!\!\perp Y_\theta \iff X, Y$ are both Normal RVs with the same variance.

Claim Let $\vec{Z} \sim N_m(\vec{0}, I)$ and $\vec{X} = A\vec{Z} + \vec{\mu}$, $A \in \mathbb{R}^{n\times m}$ and $\vec{\mu} \in \mathbb{R}^n$ s.t. $\Sigma := \mathrm{Cov}(\vec{X}) = AA^T$ is invertible.

Then,

$$f_{\vec{X}}(\vec{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})}$$

$$\vec{X} \sim N_n(\vec{\mu}, \Sigma)$$

Note: For $M \in \mathbb{R}^{n\times m}$, $\mathrm{Rank}(M) \le \min(n,m)$

$\Rightarrow$ $(MM^T)_{n\times n}$ invertible $\Rightarrow \mathrm{Rank}(MM^T) = n$

$\Rightarrow$ need $m \ge n$.

Suppose $\vec{X} \sim N_n(\vec{\mu}, \Sigma)$

$$\vec{X} = \begin{bmatrix} \vec{X}_a \\ \vec{X}_b \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix}, \quad \vec{\mu} = \begin{bmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{bmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{matrix} k \\ n-k \end{matrix}$$

(k, n−k)

Precision matrix $\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} = \Sigma^{-1}$

A useful result:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad M^{-1} = \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} \begin{bmatrix} S^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix}$$

where $S = (A - BD^{-1}C)$ is the Schur complement of D in M

Marginal Distribution

$$\vec{X}_a \sim N_k(\vec{\mu}_a, \Sigma_{aa})$$
$$\vec{X}_b \sim N_{n-k}(\vec{\mu}_b, \Sigma_{bb})$$

Conditional Distribution

$$(\vec{X}_a \mid \vec{X}_b = \vec{x}_b) \sim N_k(\vec{\mu}_{a|b}, \Sigma_{a|b}),$$

where $\vec{\mu}_{a|b} = \vec{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\vec{x}_b - \vec{\mu}_b)$

$= \vec{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\vec{x}_b - \vec{\mu}_b)$

linear dependence on $\vec{x}_b$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} = \Lambda_{aa}^{-1}$$

Independence!

For $i, j = 1, 2, \ldots, n$:

$X_i \perp\!\!\!\perp X_j \iff \Sigma_{ij} = 0$

Affine transformation

$$\vec{Y} = B\vec{X} + \vec{v} = B(A\vec{z} + \vec{\mu}) + \vec{v}$$

$B \in \mathbb{R}^{n \times n}$, invertible $= BA\vec{z} + (B\vec{\mu} + \vec{v})$

Then, $\vec{Y} \sim N_n(B\vec{\mu} + \vec{v}, B\Sigma B^T)$

Example: $n = 2$, $k = 1$.

$$\mu_{1|2} = \mu_1 + \rho\sigma_1\sigma_2\left(\frac{1}{\sigma_2^2}\right)(x_2 - \mu_2)$$
$$= \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2)$$

The sign of this shift depends on the sign of ρ and whether $x_2 > \mu_2$.

$$\Sigma_{1|2} = \sigma_1^2 - \rho\sigma_1\sigma_2\left(\frac{1}{\sigma_2^2}\right)(\rho\sigma_1\sigma_2)$$
$$= \sigma_1^2 - \rho^2\sigma_1^2$$
$$= \sigma_1^2(1 - \rho^2) \leq \mathrm{Var}(X_1) = \sigma_1^2$$

Lecture 20

Recall covariance matrix for an n-dimensional random vector $\vec{X}$:

$$Cov(\vec{X}) = \mathbb{E}[(\vec{X}-\vec{\mu})(\vec{X}-\vec{\mu})^T], \quad \vec{\mu} = \mathbb{E}[\vec{X}]$$

Then, for all fixed $\vec{v} \in \mathbb{R}^n$,

$$\vec{v}^T Cov(\vec{X})\vec{v} = \mathbb{E}[\underbrace{\vec{v}^T(\vec{X}-\vec{\mu})}_{\text{scalar } Y}\underbrace{(\vec{X}-\vec{\mu})^T\vec{v}}_{Y}] = \mathbb{E}[Y^2] \geq 0$$

$\Rightarrow$ Covariance matrices are positive semi-definite

Def (Positive (semi-) definite matrix)

Let $M \in \mathbb{R}^{n\times n}$ be a real-valued, symmetric $n\times n$ matrix. ($M = M^T$).

1) M is called positive definite if

$$\vec{x}^T M \vec{x} > 0, \quad \forall\ \vec{x} \in \mathbb{R}^n \setminus \{\vec{0}\}$$

- All eigenvalues are real and positive.
- M is non-singular. ($\det M > 0$)

2) M is called positive semi-definite if

$$\vec{x}^T M \vec{x} \geq 0, \quad \forall\ x \in \mathbb{R}^n.$$

All eigenvalues are real and non-negative. (Some eigenvalues can be zero)

Theorem

1) $M = AA^T$ where A is some real square matrix. $\Leftrightarrow$ M is positive semi-definite.

2) $M = AA^T$ where A is some real non-singular square matrix. $\Leftrightarrow$ M is positive definite.

Pf $\Rightarrow$ This direction is easy to show. Exercise.

$\Leftarrow$ This direction can be proved using the fact the eigenvectors of a real $n\times n$ symmetric matrix can be chosen to be an orthonormal basis $\{\vec{q}_1, \dots, \vec{q}_n\}$ of $\mathbb{R}^n$:

$M = Q\Lambda Q^T$, where i^{th} column of $Q = \vec{q}_i$, $Q^TQ = I$, $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$ eigenvalues

$\lambda_i \geq 0\ \forall i \Rightarrow$ can define $\Lambda^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix}$

$$M = Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q^T = \underbrace{Q\Lambda^{\frac{1}{2}}Q^T}_{A = M^{\frac{1}{2}}}\underbrace{Q\Lambda^{\frac{1}{2}}Q^T}_{A^T = M^{\frac{1}{2}}}$$

square-root matrix of M

Let Σ be an $n \times n$ positive semi-definite matrix and let A be its square-root matrix. Then,

$$\Sigma = \text{Cov}(\vec{X}),$$

where $\vec{X} = A\vec{Z}$ and $\vec{Z} = \mathcal{N}_n(\vec{0}, I_n)$

$$\text{Cov}(\vec{X}) = A\,\text{Cov}(\vec{Z})\,A^T = A\,I\,A^T = AA^T = \Sigma$$

More generally,

Σ is a covariance matrix $\iff$ Σ is positive semi-definite

Σ is a non-singular covariance matrix $\iff$ Σ is positive definite

$\Downarrow$

Σ is invertible

$\Sigma = Q \Lambda Q^T$

$\Sigma^{-1} = Q \Lambda^{-1} Q^T = Q \Lambda^{-\frac{1}{2}} Q^T \; Q \Lambda^{-\frac{1}{2}} Q^T$

$\Sigma^{-\frac{1}{2}}$ $\quad$ $\Sigma^{-\frac{1}{2}}$

$\lambda_i > 0 \;\; \forall \; i = 1, \dots, n$

$$\Lambda^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_n} \end{bmatrix} \qquad \Lambda^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\lambda_n}} \end{bmatrix}$$

So, if $\vec{X} \sim \mathcal{N}_n(\vec{\mu}, \Sigma)$, where Σ is positive definite, then

$$\vec{Z} = \Sigma^{-\frac{1}{2}}(\vec{X} - \vec{\mu}) \sim \mathcal{N}_n(\vec{0}, I_n)$$

Standardization of n-variate normal.

$$\text{Cov}(\vec{Z}) = \Sigma^{-\frac{1}{2}}\,\text{Cov}(\vec{X})\,(\Sigma^{-\frac{1}{2}})^T = \Sigma^{-\frac{1}{2}}\,\Sigma\,\Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}}\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}} = I \cdot I = I$$

Moment Generating Function.

- Recall in 1-dim, $X \sim \mathcal{N}(\mu, \sigma^2)$

  $t \in \mathbb{R}$, $M_X(t) = \mathbb{E}[e^{tX}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

- n-dim: $\vec{X} \sim \mathcal{N}_n(\vec{\mu}, \Sigma)$

  $\vec{t} \in \mathbb{R}^n$, $M_{\vec{X}}(\vec{t}) = \mathbb{E}[e^{\vec{t}^T \vec{X}}]$

  $= e^{\vec{\mu}^T \vec{t} + \frac{1}{2}\vec{t}^T \Sigma \vec{t}}$

- Why is this true?

  $\vec{t}^T \vec{X}$ is a univariate normal RV with (cf 19-4)

  $\mathbb{E}[\vec{t}^T \vec{X}] = \vec{t}^T \mathbb{E}[\vec{X}] = \vec{t}^T \vec{\mu}$.

  $\text{Var}(\vec{t}^T \vec{X}) = \text{Cov}(\vec{t}^T \vec{X}, \vec{t}^T \vec{X})$

  $= \vec{t}^T \text{Cov}(\vec{X})\,\vec{t} = \vec{t}^T \Sigma \vec{t}$

  ($\text{Cov}(\vec{X})$: covariance matrix.)

Multivariate Normal and χ^2 distributions.

Def (Idempotent) An $n \times n$ matrix M is said to be idempotent if $M^2 = M$.

Fact $M \in \mathbb{R}^{n \times n}$ is a symmetric and idempotent matrix of rank r if and only if

$$M = \vec{q}_1 \vec{q}_1^T + \cdots + \vec{q}_r \vec{q}_r^T$$

where $\{\vec{q}_1, \ldots, \vec{q}_r\}$ are r orthonormal vectors in $\mathbb{R}^n$.

Theorem Suppose $\vec{X} \sim N_n(\vec{\mu}, I_n)$ and M is an $n \times n$ symmetric matrix.

M idempotent with rank r $\Rightarrow$ $\underbrace{(\vec{X}-\vec{\mu})^T M (\vec{X}-\vec{\mu})}_{\text{Quadratic form}} \sim \chi^2_r$

(In fact, the converse is also true.)

Pf $M = \sum_{i=1}^{r} \vec{q}_i \vec{q}_i^T$, where $\{\vec{q}_1, \ldots, \vec{q}_r\}$ are orthonormal.

$$(\vec{X}-\vec{\mu})^T M (\vec{X}-\vec{\mu}) = \sum_{i=1}^{r} [\vec{q}_i^T (\vec{X}-\vec{\mu})]^2$$

$$\mathbb{E}[\vec{q}_i^T (\vec{X}-\vec{\mu})] = \vec{q}_i^T \mathbb{E}[\vec{X}-\vec{\mu}] = \vec{0}$$

$$\mathrm{Var}[\vec{q}_i^T (\vec{X}-\vec{\mu})] = \vec{q}_i^T \mathrm{Cov}(\vec{X}) \vec{q}_i = \vec{q}_i^T \vec{q}_i = 1$$

$$\Rightarrow \vec{q}_i^T (\vec{X}-\vec{\mu}) \sim N(0,1)$$

The theorem now follows from the fact that $Z_1^2 + \cdots + Z_r^2 \sim \chi^2_r$ if $Z_1, \ldots, Z_r \overset{iid}{\sim} N(0,1)$. $\square$

The above result has many applications in statistics.

e.g.] $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$S_n = X_1 + \cdots + X_n$. Then, the above result can be used to show

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \frac{S_n}{n}\right)^2 \sim \chi^2_{n-1}.$$

($\frac{S_n}{n}$: Sample mean)

Multivariate CLT

Thm Let $\vec{X}_1, \dots, \vec{X}_n$ be a sequence of iid $\mathbb{R}^K$-valued random vectors where

$$\vec{X}_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{iK} \end{bmatrix}, \quad \mathbb{E}[\vec{X}_i] = \vec{\mu} \text{ and, } \mathbb{E}[X_{ij}^2] < \infty \ \forall j = 1, \dots, K. \quad \mathrm{Cov}(\vec{X}_i) = V$$

Let $\vec{S}_n = \vec{X}_1 + \cdots + \vec{X}_n$

Then, $\sqrt{n}\left(\frac{\vec{S}_n}{n} - \vec{\mu}\right) \xrightarrow{d} \mathcal{N}_K(\vec{0}, V)$ as $n \to \infty$

If V is non-singular (positive definite), then

$$\sqrt{n} V^{-\frac{1}{2}} \left(\frac{\vec{S}_n}{n} - \vec{\mu}\right) \xrightarrow{d} \mathcal{N}_K(\vec{0}, I_K) \text{ as } n \to \infty$$

1-dim : $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli(p).

$$\sqrt{n}\left(\frac{\frac{S_n}{n} - p}{\sqrt{p(1-p)}}\right) \xrightarrow{d} N(0,1) \text{ as } n \to \infty$$

---

general k-dimensional : $\mathbb{P}[X = j] = p_j$, $j = 1, \dots, K$ $\quad \sum_{j=1}^{K} p_j = 1$.

$$C_{n,j} = \sum_{i=1}^{n} \mathbb{1}\{X_i = j\}$$

$\vec{C}_n = (C_{n,1}, \dots, C_{n,K})^T \sim$ Multinomial$(n, p_1, \dots, p_K)$

$\mathbb{E}[C_{n,j}] = np_j$, $\mathrm{Var}(C_{n,j}) = np_j(1-p_j)$

$\mathrm{Cov}(C_{n,i}, C_{n,j}) = -np_ip_j$, for $i \neq j$.

Thm (Pearson)

$$\sum_{i=1}^{K} \frac{(C_{n,j} - np_j)^2}{np_j} \xrightarrow{d} \chi^2_{K-1} \text{ as } n \to \infty.$$

Proof sketch :

multivariate CLT $\Rightarrow$ as $n \to \infty$,

$$\left(\frac{C_{n,1} - np_1}{\sqrt{np_1}}, \dots, \frac{C_{n,K} - np_K}{\sqrt{np_K}}\right) \xrightarrow{d} \mathcal{N}_K(\vec{0}, M)$$

where M is a $K \times K$ matrix with

$$M_{ij} = \mathrm{Cov}\left(\frac{C_{n,i}}{\sqrt{np_i}}, \frac{C_{n,j}}{\sqrt{np_j}}\right) = \begin{cases} 1 - p_i, & \text{if } i = j, \\ -\sqrt{p_ip_j}, & \text{if } i \neq j. \end{cases}$$

Fact! M is idempotent and symmetric with $\mathrm{rank}(M) = K - 1$.

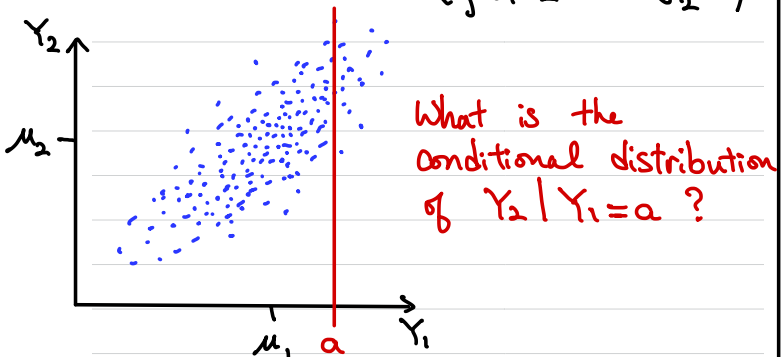
Apply the theorem on page 20-3.

Lecture 21

Recall Bivariate normal:

$$(Y_1, Y_2) \sim N_2(\vec{\mu}, \Sigma)$$

$$\vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$



What is the conditional distribution of $Y_2 \mid Y_1 = a$?

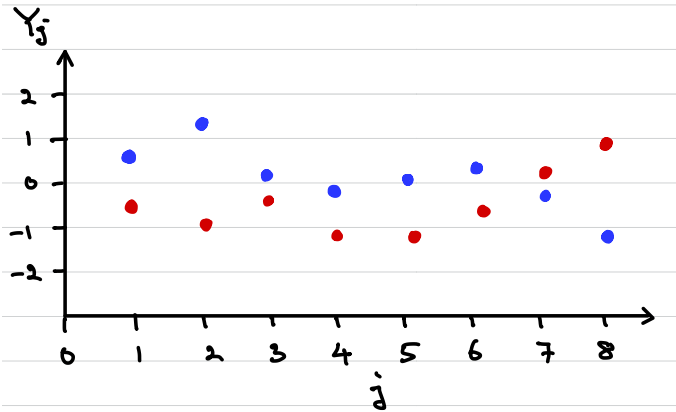
$$Y_2 \mid Y_1 = a \sim N(\mu_{2|1}, \Sigma_{2|1})$$

See Lecture 19-4, which implies

$$\mu_{2|1} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(a - \mu_1)$$

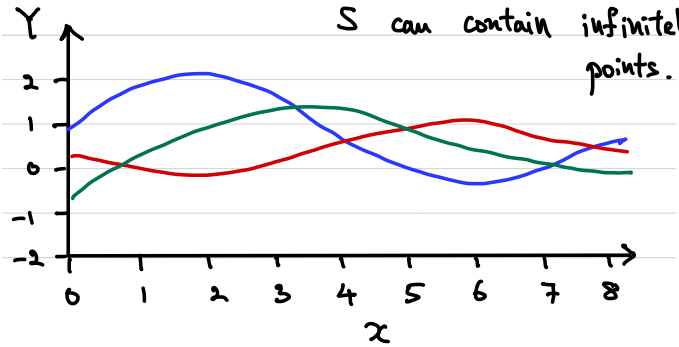
$$\Sigma_{2|1} = \sigma_2^2(1-\rho^2) < \sigma_2^2 \text{ if } \rho \neq 0.$$

How can we visualize $Y_1, \ldots, Y_d$ for $d > 2$?



Gaussian Processes generalize this concept to **functions**: $\{Y(x) \in \mathbb{R} \text{ for } x \in S \subset \mathbb{R}\}$

S can contain infinitely many points.



More generally, $Y = Y(x) \in \mathbb{R}$ for $x \in S \subset \mathbb{R}^n$. ($S$: index set)

A stochastic process is a collection of RVs. $\{Y(x),\ x \in S\}$

Def (Gaussian Process) A Gaussian Process is a stochastic process s.t. any finite collection $\{Y(x_1), \ldots, Y(x_n)\}$ is multivariate normal.

$$Y(\cdot) \sim GP(m(\cdot), K(\cdot,\cdot))$$

$m(\cdot)$: mean function; $K(\cdot,\cdot)$: covariance function (Kernel).

$$\mathbb{E}[Y(x)] = m(x)$$

$$\mathrm{Cov}(Y(x), Y(x')) = K(x, x')$$

$$(Y(x_1), \ldots, Y(x_n)) \sim N_n(\vec{m}, K)$$

$$\vec{m} = \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \quad K = \begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) \\ \vdots & \ddots & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_n) \end{bmatrix}$$

- **Mean function:**

  Can be anything

  Popular choices: $m(x) = 0 \quad \forall x$

  $m(x) = \beta^T x$

- **Covariance function:**

  Must be positive semi-definite.

  Stationary: $\mathrm{Cov}(Y(x), Y(x')) = K(x - x')$

  Isotropic: $\mathrm{Cov}(Y(x), Y(x')) = K(\|x - x'\|)$ ← Kernel

  Note that the RHS depends on x, x' but not on $Y(x)$ or $Y(x')$.

Example:

- **Radial Basis Function (RBF)**

$$K_{RBF}(x, x') = \sigma^2 \exp\left\{-\frac{1}{2\ell^2}|x - x'|^2\right\}$$

length scale $\ell > 0$

  - Positive definite.
  - Infinitely differentiable.
  - Possibly the most widely used kernel.

Q How can one sample a function from $GP(m(\cdot), k(\cdot,\cdot))$?

1) Discretize S as $\{x_1, \dots, x_D\}$
2) Sample $Y(x_1) \sim N(m(x_1), k(x_1, x_1))$.
3) For $n = 1, 2, \dots, D$
$(Y(x_1), \dots, Y(x_{n+1})) \sim N_{n+1}(m, K)$

$\vec{c}_{n+1}$ ← Call this K_{n-1}

$$m = \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \\ m(x_{n+1}) \end{bmatrix}, \quad K = \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) & k(x_1,x_{n+1}) \\ \vdots & & \vdots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) & k(x_n,x_{n+1}) \\ \vec{c}_{n+1}^T & & & k(x_{n+1},x_{n+1}) \end{bmatrix}$$

Sample $Y(x_{n+1})$ from the following conditional distribution:

$$Y(x_{n+1}) \mid Y(x_1), \dots, Y(x_n) \sim N(\mu_{n+1}, \sigma_{n+1}^2)$$

where

$$\mu_{n+1} = m(x_{n+1}) - \vec{c}_{n+1}^T K_n^{-1} \begin{bmatrix} Y(x_1) - m(x_1) \\ \vdots \\ Y(x_n) - m(x_n) \end{bmatrix}$$

$$\sigma_{n+1}^2 = k(x_{n+1}, x_{n+1}) - \vec{c}_{n+1}^T K_n^{-1} \vec{c}_{n+1}$$

Warning! K_n can be ill conditioned, leading to numerical issues.

A Solution:

Replace K with $K + \tau^2 I$ where $\tau^2 I$ is a constant diagonal matrix.

This is similar to the regularization in Ridge regression.

Interpretation: Noisy observation.

Each $Y(x_i)$ is observed with some additive noise independent of GP.

$$Y(x_i) = g(x_i) + \varepsilon_i$$

$g(x_i)$: from GP; $\varepsilon_1, \varepsilon_2, \dots \overset{iid}{\sim} N(0, \tau^2)$

Called the nugget of GP

See slides for example samples from these GPs.

Matérn Kernel

$\nu = p + \frac{1}{2}$ where $p \in \mathbb{N}$.

$$K^{\text{Matérn}}_{p+\frac{1}{2}}(x,x') = \sigma^2 \frac{p!}{(2p)!} \sum_{j=0}^{p} \frac{(p+j)!}{j!(p-j)!} \left[\frac{2\sqrt{2p+1}}{\ell} |x-x'|\right]^{p-j}$$

A polynomial in $|x-x'|$ of degree p

$$\times \exp\left(-\frac{\sqrt{2p+1}}{\ell} |x-x'|\right)$$

GP with $K^{M}_{\nu}(\cdot,\cdot)$ kernel is $\lceil \nu \rceil - 1$ times differentiable in the "mean square" sense.

$$\lim_{p\to\infty} K^{\text{Matérn}}_{p+\frac{1}{2}}(x,x') = K_{RBF}(x,x') = \sigma^2 \exp\left(-\frac{|x-x'|^2}{2\ell^2}\right)$$

Brownian Motion (Non-stationary)

$$K_{BM}(x,x') = \min(x,x')$$
$$x, x' \in (0, \infty)$$

Brownian Bridge (Non-stationary)

$$K_{BB}(x,x') = \min(x,x') - xx'$$
$$x, x' \in [0,1]$$

This corresponds to a stochastic process Conditioned on $Y(0) = 0$ and $Y(1) = 0$.

Ornstein - Uhlenbeck

$$K_{OU}(x,x') = \frac{\sigma^2 \ell}{2} e^{-\frac{1}{\ell}|x-x'|}$$
$$x, x' \in (0, \infty)$$

This model has applications in physics, biology, finance, etc.

Linear (Non-Stationary)

Bayesian Linear Regression

Suppose $Y(x_i) = \beta_0 + \beta_1 (x_i - c)$

where $\beta_0 \sim N(0, \sigma_0^2)$, $\beta_1 \sim N(0, \sigma_1^2)$ and $\beta_0 \perp\!\!\!\perp \beta_1$

Then, for $x \neq x'$,

$$\begin{aligned} \text{Cov}(Y(x), Y(x')) &= \text{Cov}(\beta_0 + \beta_1(x-c), \beta_0 + \beta_1(x'-c)) \\ &= \sigma_0^2 + \sigma_1^2 (x-c)(x'-c) \end{aligned}$$

So, the corresponding kernel is

$$k_{\text{Linear}}(x,x') = \sigma_0^2 + \sigma_1^2 (x-c)(x'-c)$$

Prediction

Training data $D = \{(x_i, y_i),\ i = 1, \ldots, n\}$.

How can we predict $Y(x)$ for a new sample point x?

Ans: Use the algorithm on page 21-3. See slides.

Stat 201A Fall 2024

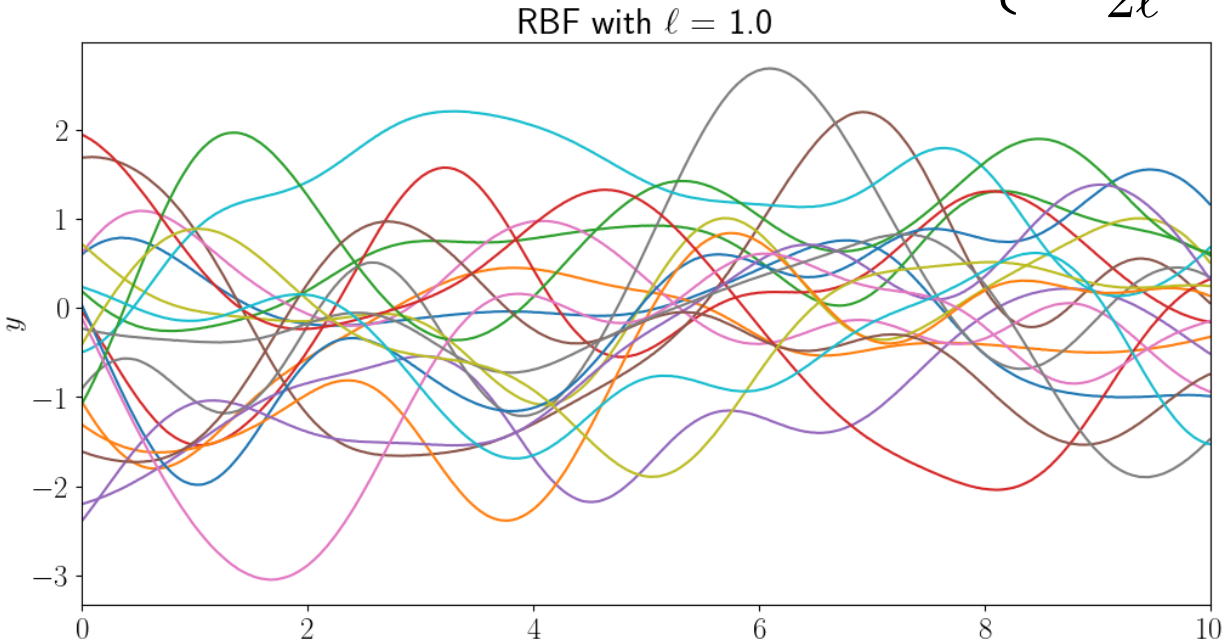
Prof. Yun S. Song

Lecture 21

Gaussian Processes Demonstration

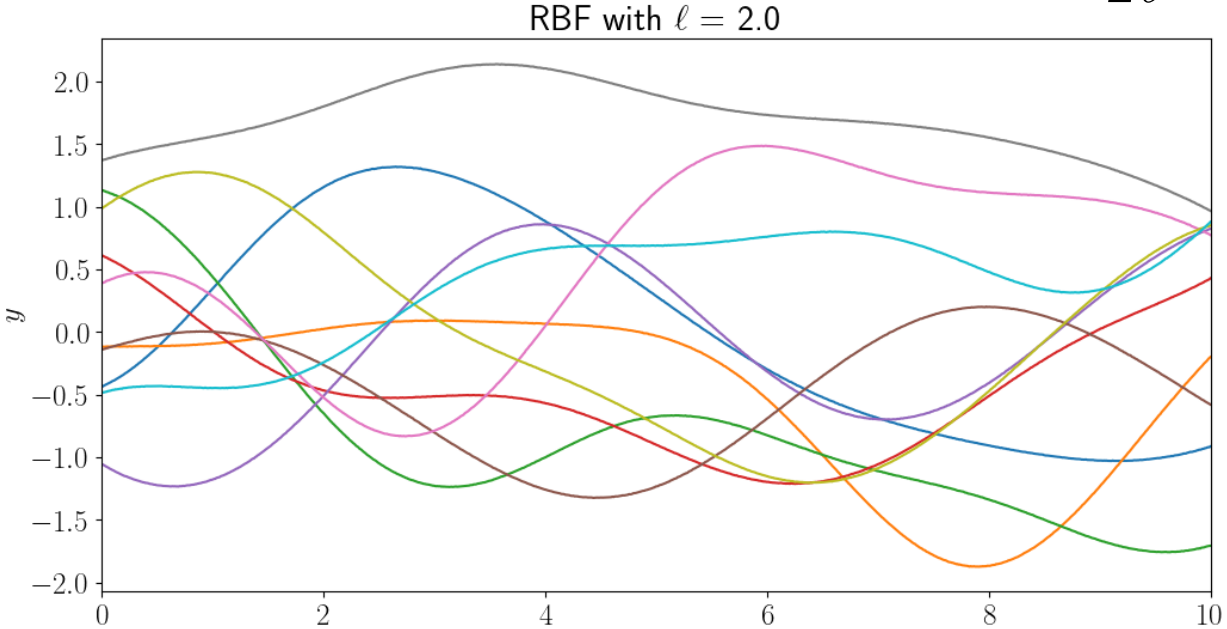
Radial Basis Function Kernel

$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



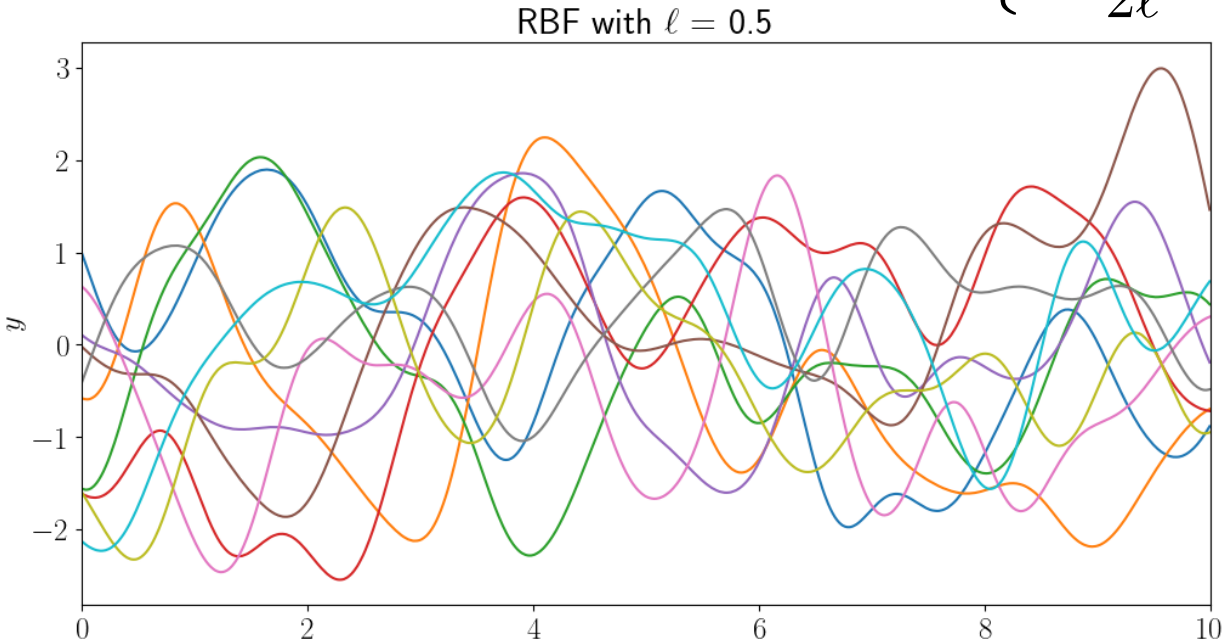
Radial Basis Function Kernel

$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



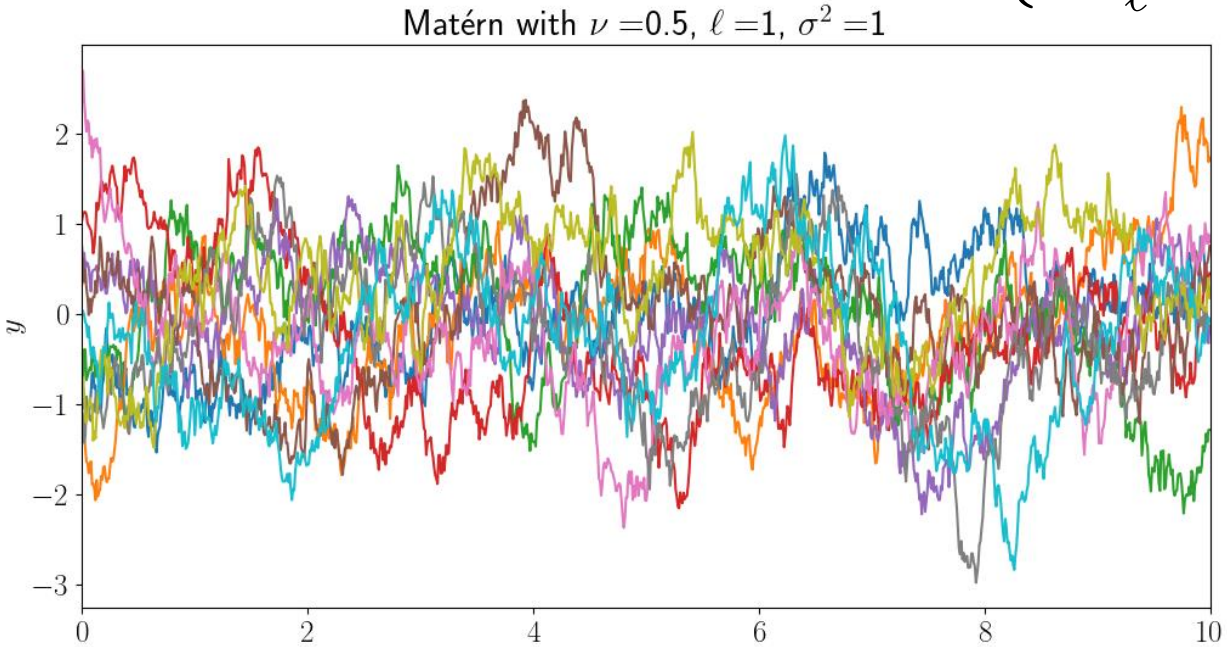
Radial Basis Function Kernel

$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



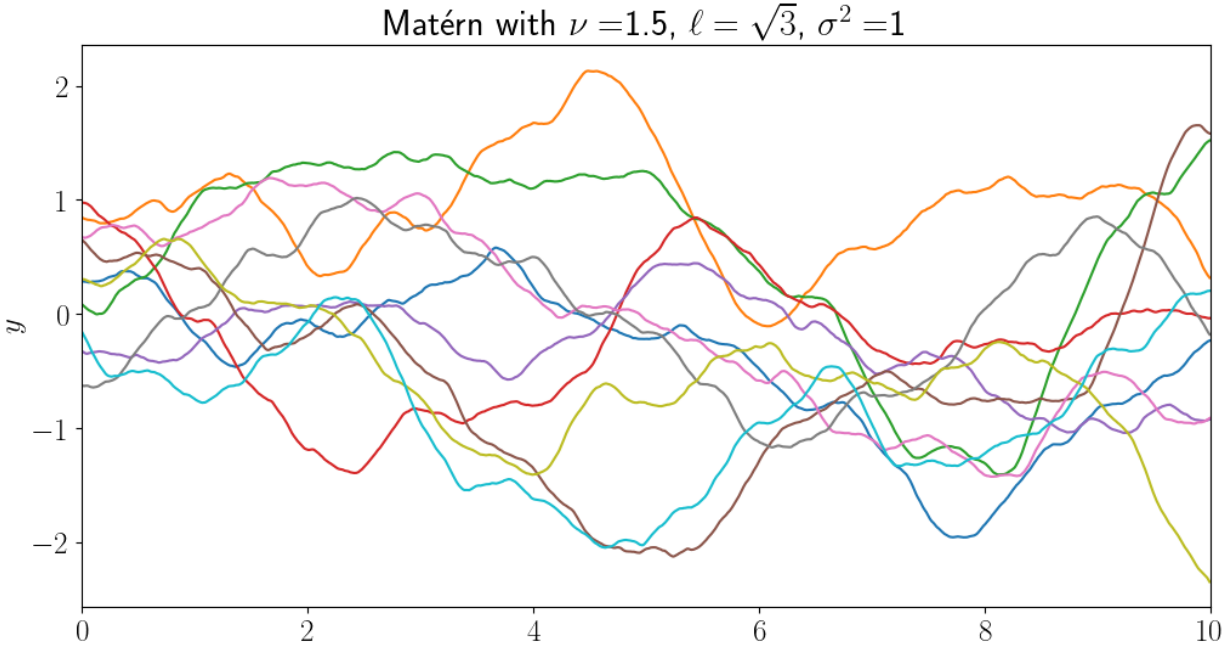
Matérn Kernel

$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{\ell}|x - x'| \right\}$$



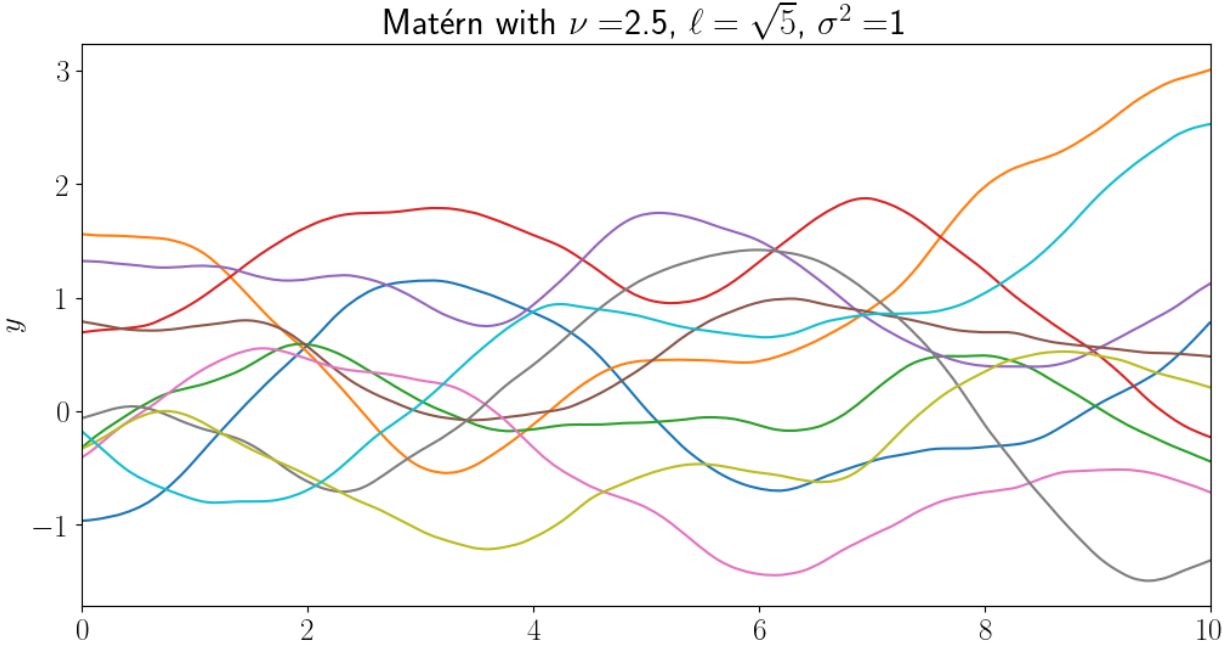
Matérn Kernel

$$k(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3}}{\ell}|x - x'|\right) \exp\left\{ - \frac{\sqrt{3}}{\ell}|x - x'|\right\}$$



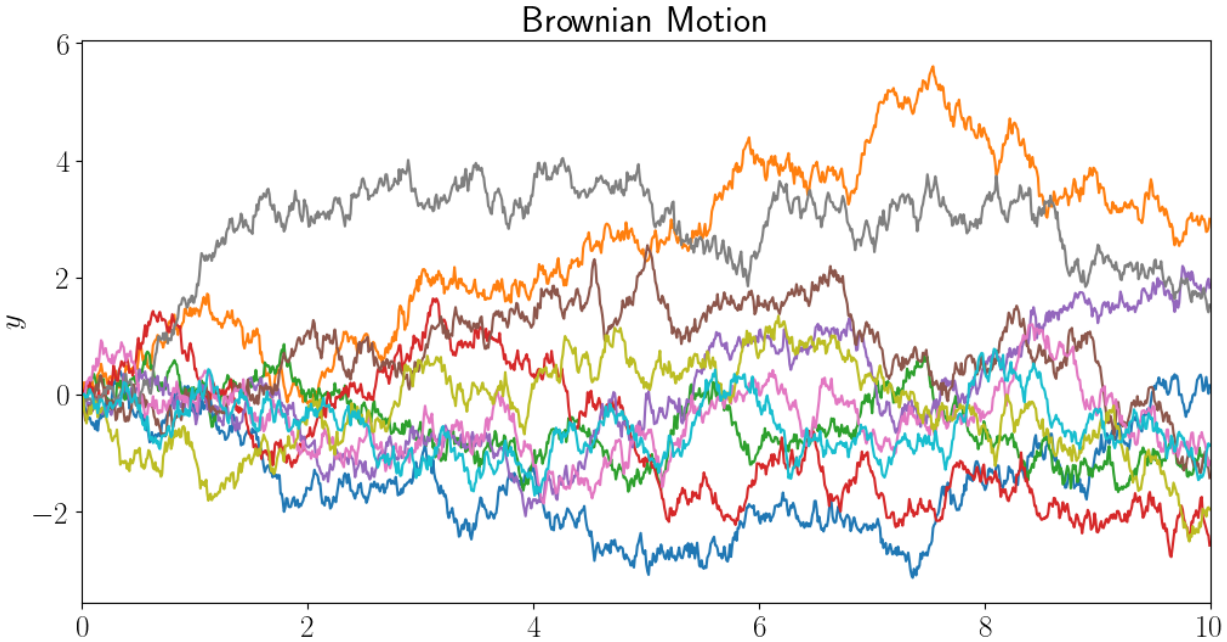
Matérn Kernel

$$k(x, x') = \sigma^2 \left(1 + \frac{\sqrt{5}}{\ell}|x - x'| + \frac{5}{3\ell^2}|x - x'|^2\right) \exp\left\{-\frac{\sqrt{5}}{\ell}|x - x'|\right\}$$



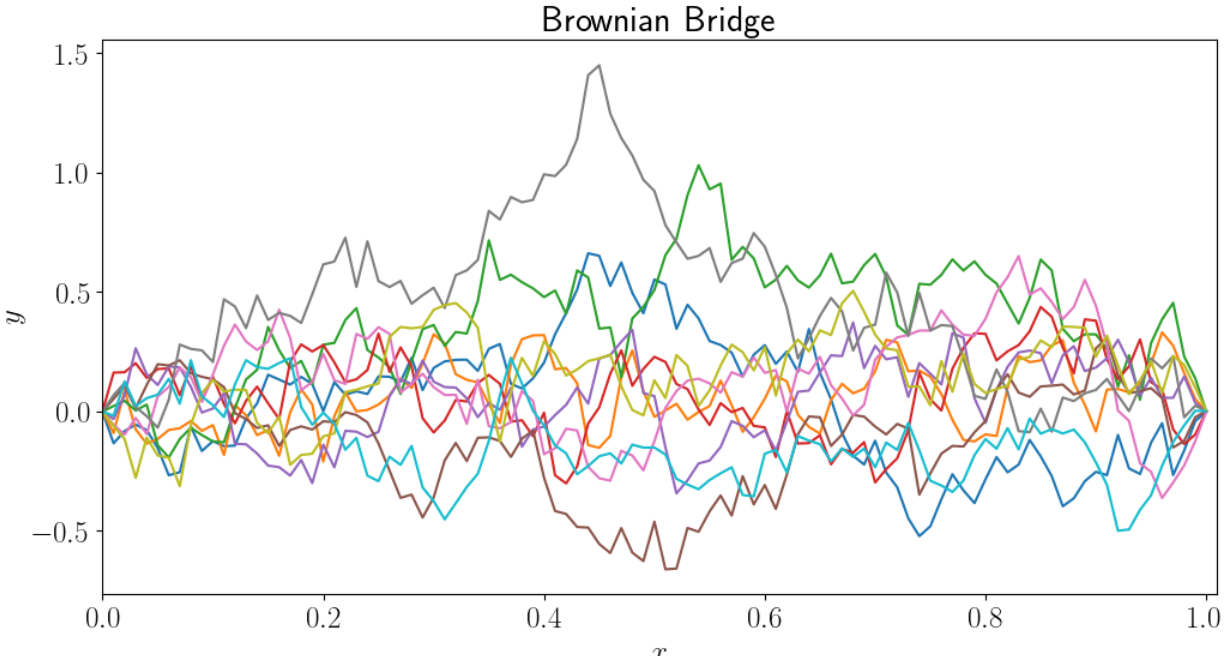
Brownian Motion Kernel

$$k(x, x') = \min\{x, x'\}$$



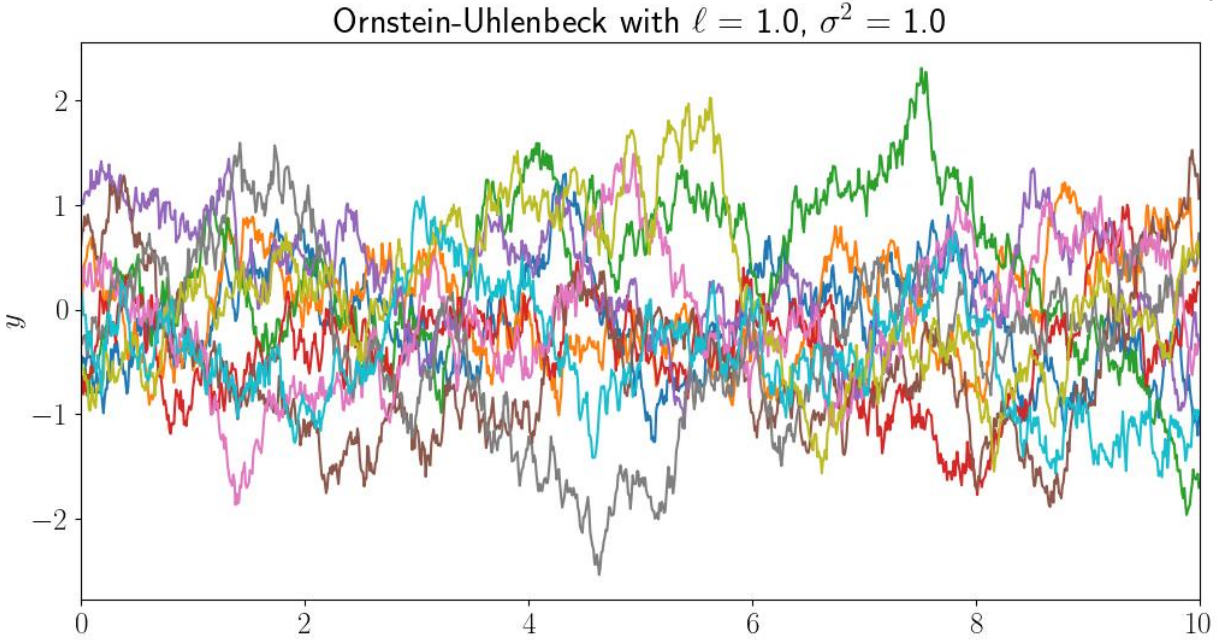
Brownian Bridge Kernel

$$k(x, x') = \min\{x, x'\} - xx'$$



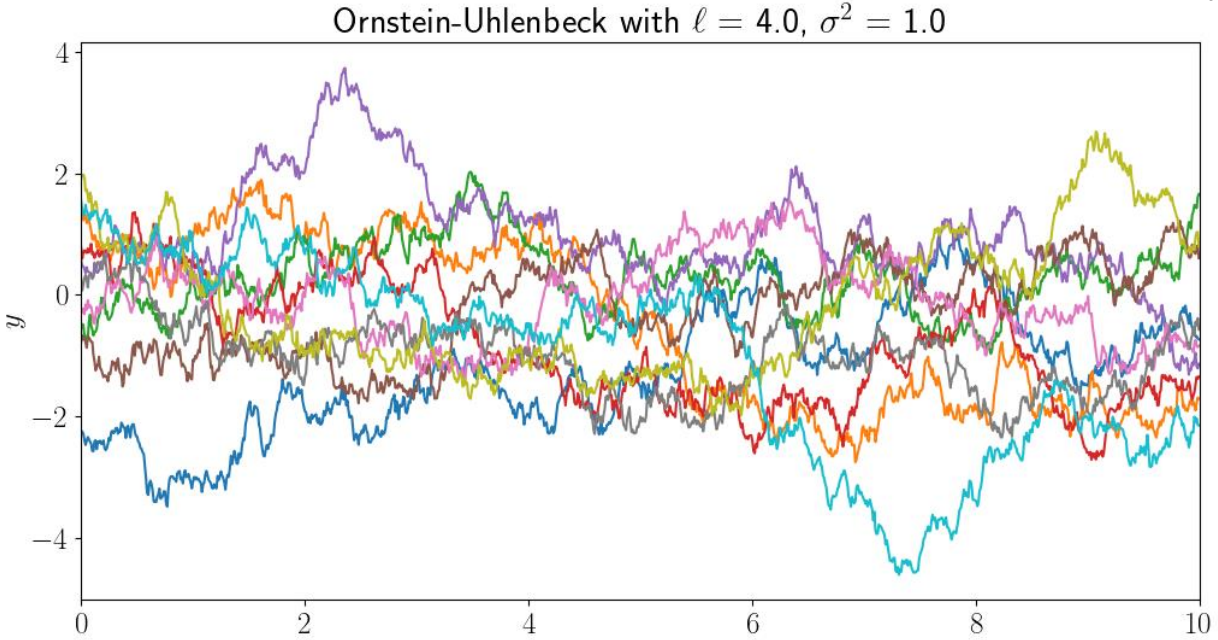
Ornstein-Uhlenbeck Kernel

$$k(x, x') = \frac{\sigma^2 \ell}{2} \exp\left\{ -\frac{1}{\ell}|x - x'| \right\}$$



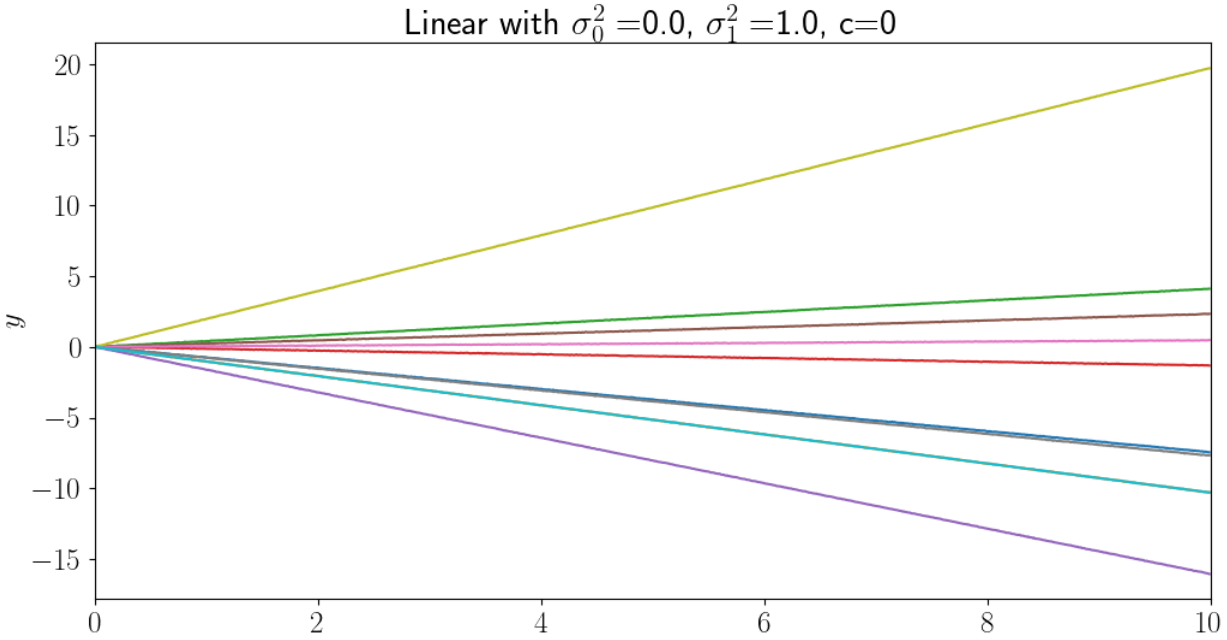
Ornstein-Uhlenbeck Kernel

$$k(x, x') = \frac{\sigma^2 \ell}{2} \exp\left\{ -\frac{1}{\ell}|x - x'| \right\}$$



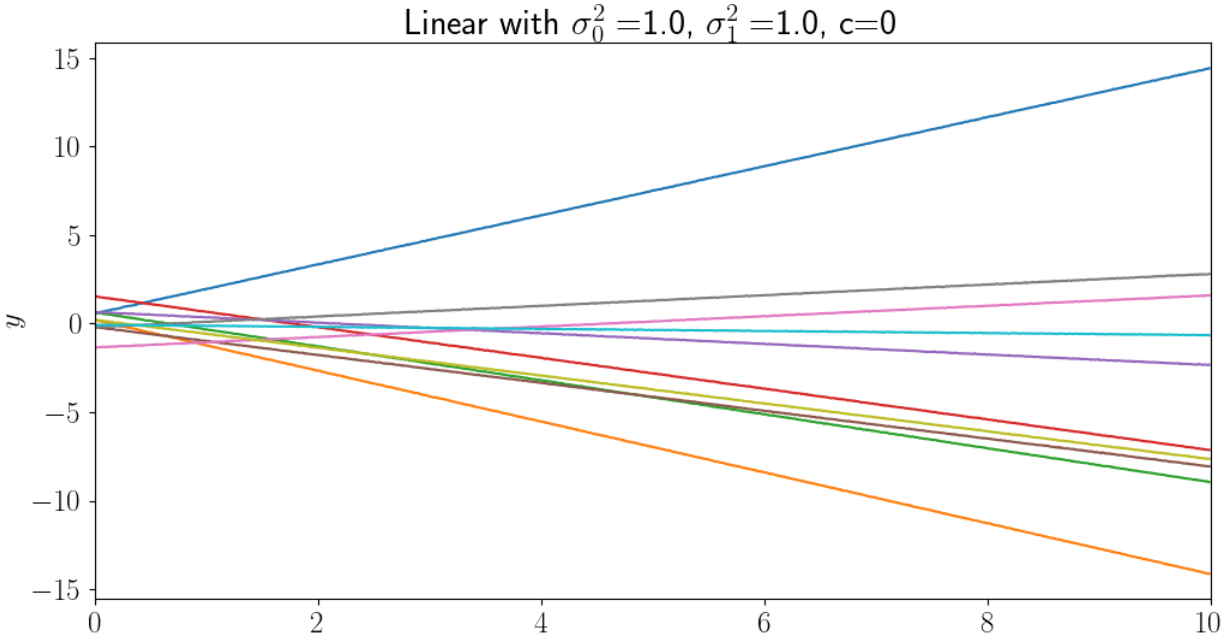
Linear Kernel

$$k(x, x') = \sigma_0^2 + \sigma_1^2 (x - c)(x' - c)$$



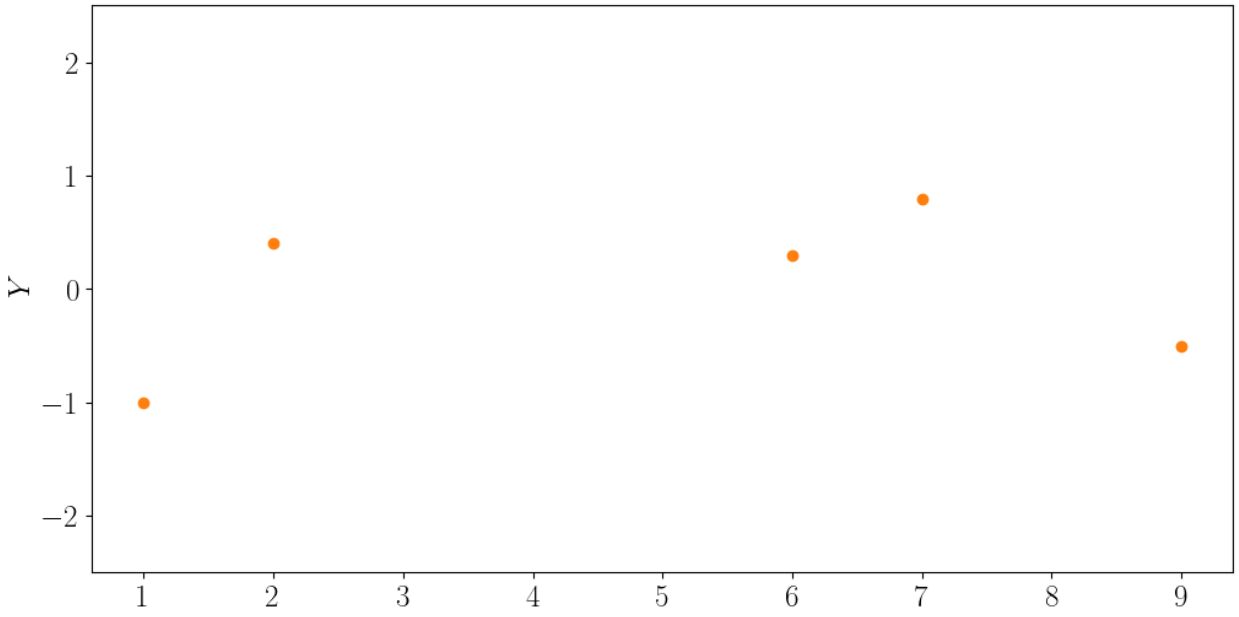
Linear Kernel

$$k(x, x') = \sigma_0^2 + \sigma_1^2 (x - c)(x' - c)$$



Prediction of $Y(x)$ for a new point x?

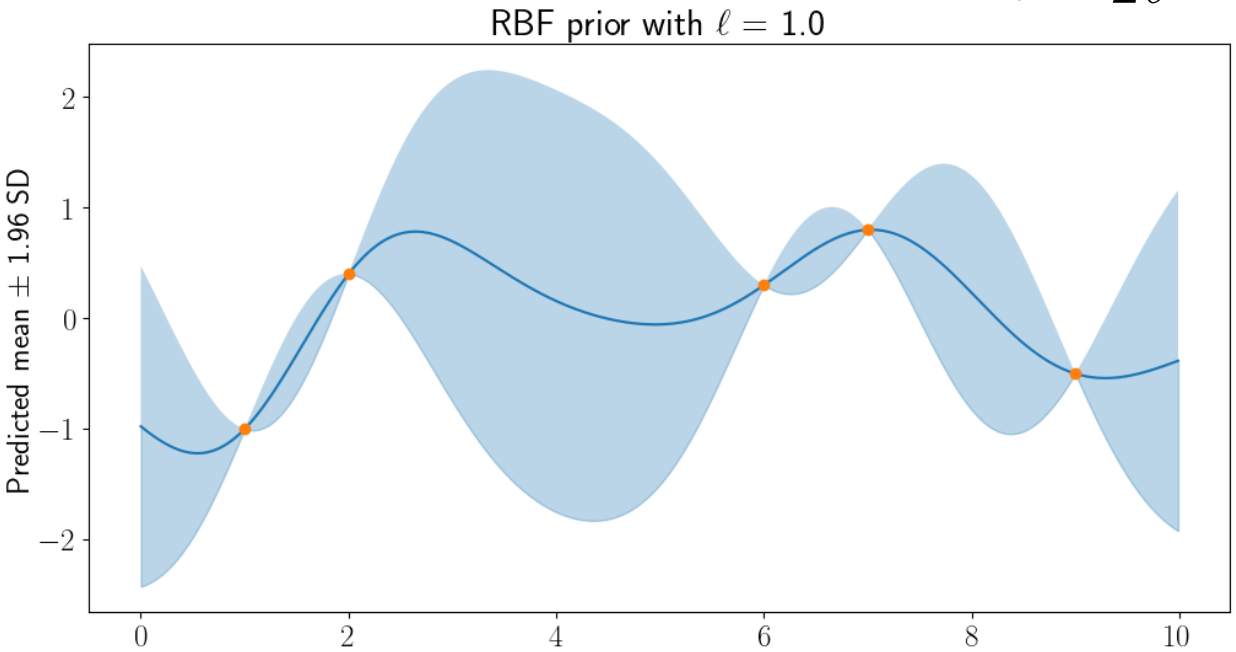
Observed data =
[[1,-1],
[2,0.4],
[6,0.3],
[7,0.8],
[9,-0.5]]



Posterior Distribution

Observed data =
[[1,-1],
[2,0.4],
[6,0.3],
[7,0.8],
[9,-0.5]]

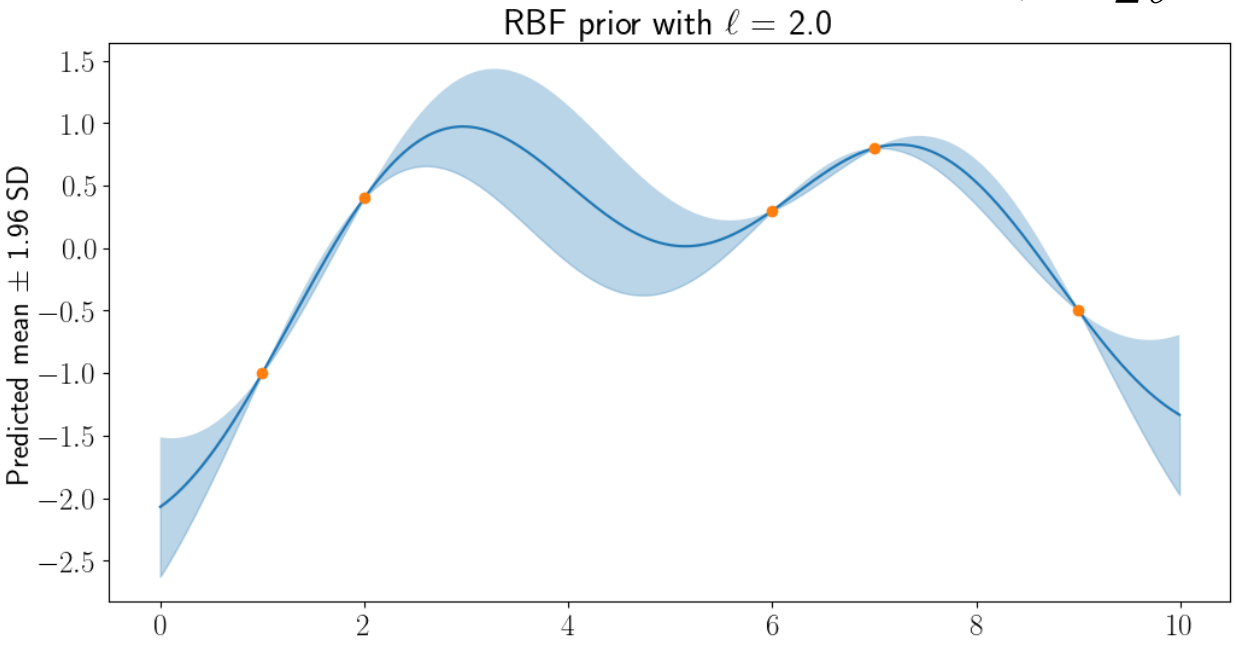
$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



Posterior Distribution

Observed data =
[[1,-1],
[2,0.4],
[6,0.3],
[7,0.8],
[9,-0.5]]

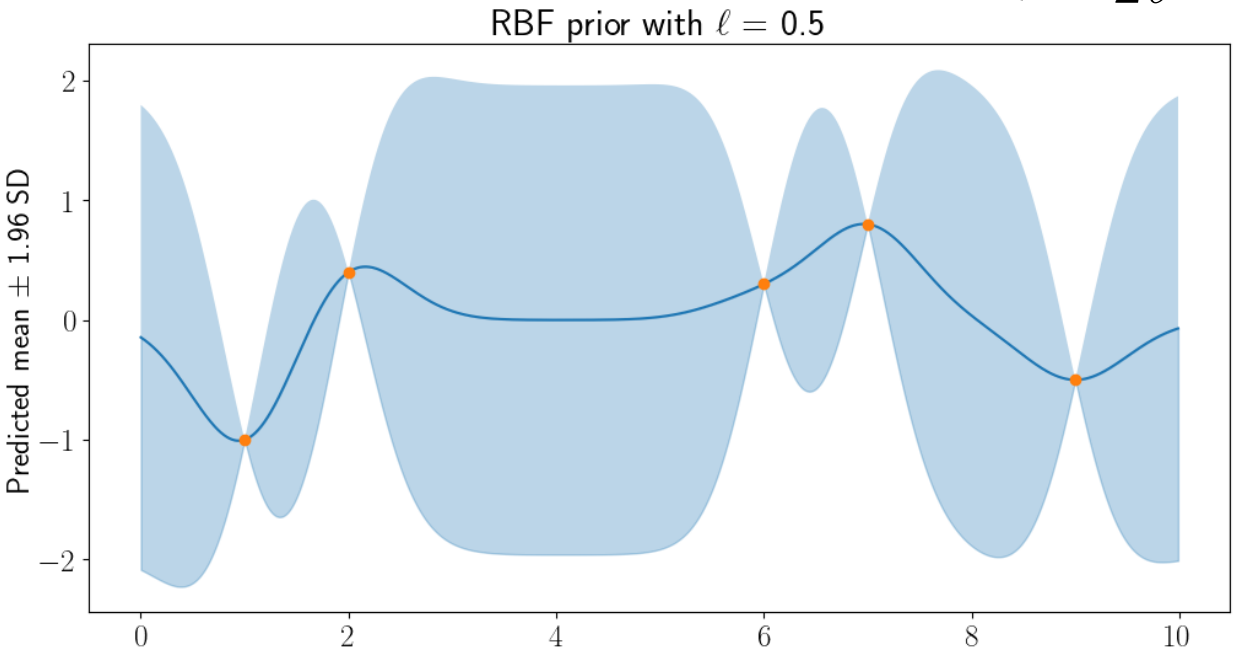
$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



Posterior Distribution

Observed data =
[[1,-1],
[2,0.4],
[6,0.3],
[7,0.8],
[9,-0.5]]

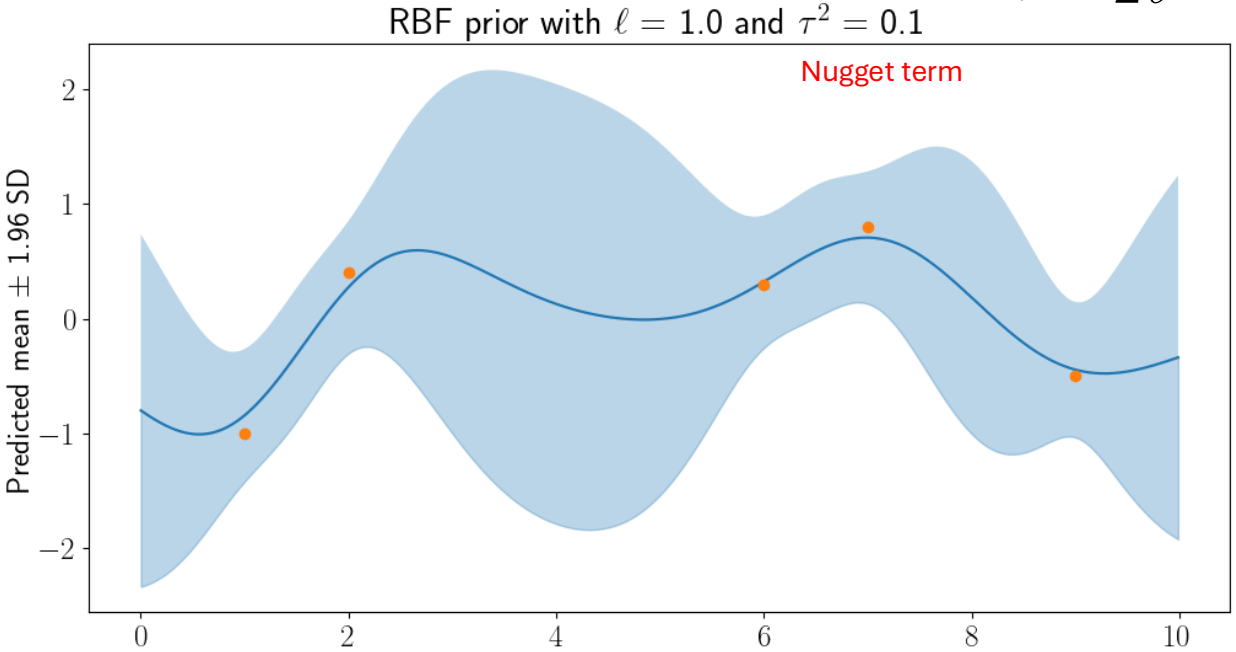
$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



Posterior Distribution

Observed data =
[[1,-1],
[2,0.4],
[6,0.3],
[7,0.8],
[9,-0.5]]

$$k(x, x') = \sigma^2 \exp\left\{ -\frac{1}{2\ell^2}|x - x'|^2 \right\}$$



Recall the empirical CDF from Lecture 8:

$X_1, X_2, \dots, X_n \overset{iid}{\sim} F$

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}$$

Strong LLN $\Rightarrow$ For every $x \in \mathbb{R}$, $\hat{F}_n(x) \xrightarrow{a.s.} F(x)$ as $n \to \infty$.

Define $D_n := \sup_x |\hat{F}_n(x) - F(x)|$.

Thm (Glivenko–Cantelli) $D_n \xrightarrow{a.s.} 0$ as $n \to \infty$.

Today we will prove a weaker version of this theorem: $D_n \xrightarrow{P} 0$ as $n \to \infty$.

Lemma 1 The distribution of D_n is the same for all continuous F.

Pf

Recall the quantile function from Lecture 12

$$q(u) = \inf\{x \in \mathbb{R} \mid F(x) \ge u\}$$

$U \sim \text{Uniform}(0,1)$, $X \sim F$, continuous

Then, $q(U) \overset{d}{=} X$ and $F(X) \overset{d}{=} U$

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$$

$$= \sup_{u \in (0,1)} |\hat{F}_n(q(u)) - F(q(u))|$$

$$= \sup_{u \in (0,1)} |\hat{F}_n(q(u)) - u|$$

$$\hat{F}_n(q(u)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le q(u)\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{F(X_i) \le u\}$$

But $F(X_1), \dots, F(X_n) \overset{iid}{\sim} \text{Unif}(0,1)$, which proves the claim. □

Hence, if we show that $D_n \xrightarrow{P} 0$ as $n \to \infty$ for Uniform(0,1), then Lemma 1 implies that it holds true for all continuous F.

In what follows, let $U_1, \dots, U_n \overset{iid}{\sim} \text{Uniform}$.

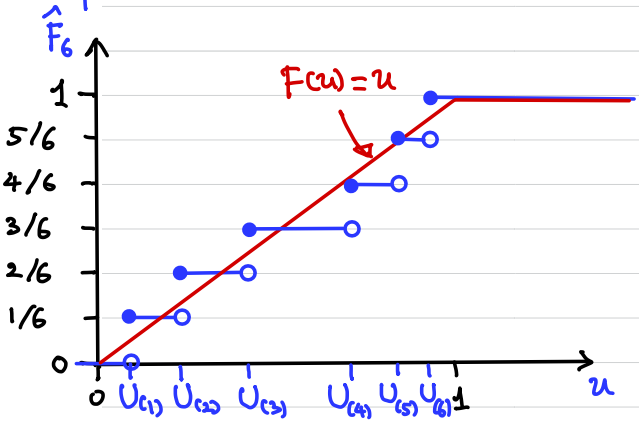
Recall order statistics from Lecture 15.

$$U_{(1)} < U_{(2)} < \cdots < U_{(n)}$$

$U_{(j)} = j^{th}$ smallest element of $\{U_1, \dots, U_n\}$

$$\hat{F}_n(u) = \frac{1}{n}\sum_{a=1}^{n} \mathbb{1}\{U_a \le u\}$$

example for $n=6$



$\sup_{u\in(0,1)} |\hat{F}_n(u) - u|$ occurs at either

$U_{(k)}$ or $U_{(k)}^-$ for some $k \in [n] := \{1,\dots,n\}$

$\hat{F}_n(U_{(k)}) = \frac{k}{n}$, $\hat{F}_n(U_{(k)}^-) = \frac{k-1}{n}$.

$$D_n = \max\left\{ \max_{k\in[n]} \left|\frac{k}{n} - U_{(k)}\right|, \max_{k\in[n]} \left|\frac{k-1}{n} - U_{(k)}\right| \right\}$$

Recall from Lecture 15 that

$U_{(k)} \sim \text{Beta}(k, n-k+1)$

$$\mathbb{E}[U_{(k)}^m] = \frac{n!}{(n+m)!}\frac{(k+m-1)!}{(k-1)!} = \frac{(k+m-1)(k+m-2)\cdots k}{(n+m)(n+m-1)\cdots(n+1)}$$

For any $\varepsilon > 0$,

$$\mathbb{P}\left[\max_{k\in[n]} |U_{(k)} - \mathbb{E}[U_{(k)}]| \ge \varepsilon\right]$$

$$= \mathbb{P}\left[\bigcup_{k\in[n]} \{|U_{(k)} - \mathbb{E}[U_{(k)}]| \ge \varepsilon\}\right] \quad \text{Union}$$

$$\le \sum_{k\in[n]} \mathbb{P}[|U_{(k)} - \mathbb{E}[U_{(k)}]| \ge \varepsilon] \quad \text{By Union Bound}$$

$$= \mathbb{P}[|U_{(k)} - \mathbb{E}[U_{(k)}]|^4 \ge \varepsilon^4]$$

$$\le \frac{\mathbb{E}[|U_{(k)} - \mathbb{E}[U_{(k)}]|^4]}{\varepsilon^4} \quad \text{By Markov's Ineq.}$$

$$\le \frac{c}{n^2\varepsilon^4}$$ ← constant that does not depend on k or n.

$$\le n\frac{c}{n^2\varepsilon^4} = \frac{c}{n\varepsilon^4} \to 0 \text{ as } n\to\infty.$$

So, $\max_{k\in[n]} |U_{(k)} - \mathbb{E}[U_{(k)}]| \xrightarrow{p} 0$ as $n\to\infty$.

Furthermore, $\mathbb{E}[U_{(k)}] = \frac{k}{n+1}$ and for every $k \in [n]$,

$$\left|\mathbb{E}[U_{(k)}] - \frac{k}{n}\right| = \left|\frac{k}{n(n+1)}\right| \to 0 \text{ as } n \to \infty.$$

$$\left|\mathbb{E}[U_{(k)}] - \frac{k-1}{n}\right| = \left|\frac{n-k+1}{n(n+1)}\right| \to 0 \text{ as } n \to \infty.$$

Hence, $D_n \xrightarrow{p} 0$ as $n \to \infty$,
since $0 \le |x-z| \le |x-y| + |y-z|$, $\forall x, y, z \in \mathbb{R}$.

This proves the Glivenko-Cantelli Theorem □

Multivariate CLT (Lecture 20) implies

Again, assume Uniform(0,1) distn. $F(u) = u$, $\forall u \in (0,1)$

For any $0 < u_1 < u_2 < \cdots < u_k < 1$,

$$\sqrt{n}\left(\begin{bmatrix} \hat{F}_n(u_1) \\ \vdots \\ \hat{F}_n(u_k) \end{bmatrix} - \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix}\right) \xrightarrow{d} N_k(\vec{0}, \Sigma), \quad \text{as } n \to \infty$$

$$\| \quad \text{← a } k\text{-dim random vector}$$

$$\frac{1}{n}\sum_{a=1}^{n} \begin{bmatrix} \mathbb{1}\{U_a \le u_1\} \\ \vdots \\ \mathbb{1}\{U_a \le u_k\} \end{bmatrix}$$

where $\Sigma = (\Sigma_{ij})_{i,j=1,\cdots,k}$ with

$U \sim \text{Unif}(0,1)$

$$\begin{aligned}\Sigma_{ij} &= \text{Cov}(\mathbb{1}\{U \le u_i\}, \mathbb{1}\{U \le u_j\}) \\ &= \mathbb{E}[\mathbb{1}\{U \le u_i\}\mathbb{1}\{U \le u_j\}] \\ &\quad - \mathbb{E}[\mathbb{1}\{U \le u_i\}]\mathbb{E}[\mathbb{1}\{U \le u_j\}] \\ &= \min\{u_i, u_j\} - u_i u_j\end{aligned}$$

since

$$\begin{aligned}&\mathbb{E}[\mathbb{1}\{U \le u_i\}\mathbb{1}\{U \le u_j\}] \\ &= \mathbb{P}[U \le u_i, U \le u_j] = \mathbb{P}[U \le \min\{u_i, u_j\}] \\ &= \min\{u_i, u_j\}\end{aligned}$$

$$\mathbb{E}[\mathbb{1}\{U \le u_i\}] = \mathbb{P}[U \le u_i] = u_i$$

$$\mathbb{E}[\mathbb{1}\{U \le u_j\}] = u_j$$

This holds true for all k.

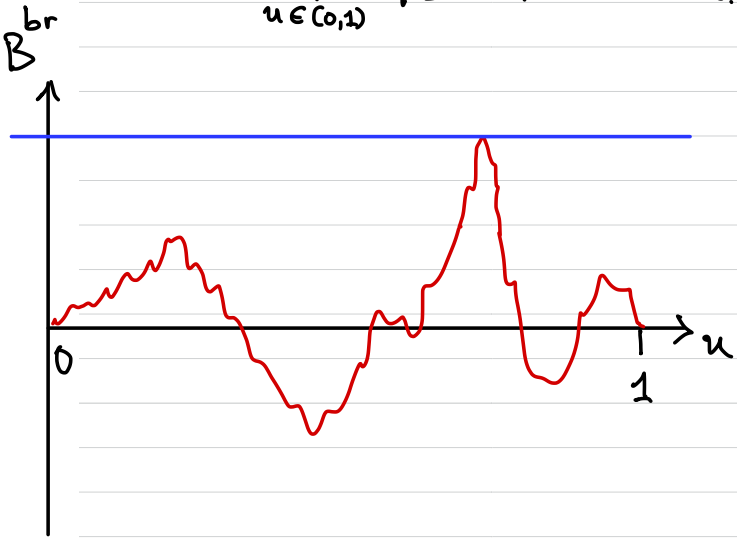
So, the RHS corresponds to a Gaussian Process with the Brownian Bridge kernel (Lecture 21)!

$$\{B^{br}(u),\ u \in (0,1)\}$$

Hence,

$$\sqrt{n}\, D_n = \sup_{u \in (0,1)} \sqrt{n}\, |\hat{F}_n(u) - u|$$

$$\xrightarrow{d} \sup_{u \in (0,1)} |B^{br}(u)| \quad \text{as } n \to \infty.$$



A lot is known about Brownian Bridges

For example,

Thm (Kolmogorov - Smirnov Distribution)

$$\mathbb{P}\left[\sup_{u \in (0,1)} |B^{br}(u)| > x\right] = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2x^2}$$

The first term $2e^{-2x^2}$ alone is very accurate.

So, if n is large,

$$\mathbb{P}\left[D_n > \frac{x}{\sqrt{n}}\right] \approx 2e^{-2x^2}$$

This can be used to find an asymptotic level-α confidence interval for estimating $F(u)$ simultaneously for all u.

$$2e^{-2x^2} = \alpha \Rightarrow x = \sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$$

$$\frac{x}{\sqrt{n}} = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}$$

Def A random process is a family $\{X_t, t \in T\}$ of RVs indexed by some set T.

$X_t : \Omega \to S$ (S: state space.)

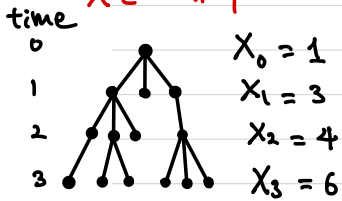
Interpretation: X_t "evolves" as time passes in a random but prescribed way.

Galton-Watson Branching Process.

Historical context: Model for
- Family name propagation (Galton, 1889)
- Free neutrons in nuclear fission reactions (1930's) (critical mass needed to sustain chain reaction.)

$T = \mathbb{N}_0 = \{0,1,2,\dots\}$, $S = \mathbb{N}_0$

X_t = # particles at time t



Each particle gives birth to $k \in \mathbb{N}_0$ children with probability p_k, independently of other particles in the past and present.

Let $F = \{p_k, k \in \mathbb{N}_0\}$ denote the offspring number distribution with mean $\mu < \infty$ and variance $\sigma^2 < \infty$.

$B_1^{(t)}, \dots, B_{X_t}^{(t)} \overset{\text{iid}}{\sim} F$ and $\perp\!\!\!\perp$ from X_t

$X_{t+1} = B_1^{(t)} + \cdots + B_{X_t}^{(t)}$ $\quad \mathbb{P}[B_i^{(t)} = k] = p_k$

Wald's Identity (Lecture note 18-3)

$$\Rightarrow \mathbb{E}[X_t] = \mu\, \mathbb{E}[X_{t-1}] = \mu(\mu\, \mathbb{E}[X_{t-2}]) = \cdots = \mu^t\, \mathbb{E}[X_0]$$

($\mathbb{E}[X_t]$ determined by the initial condition. Typically, $X_0 = 1$.)

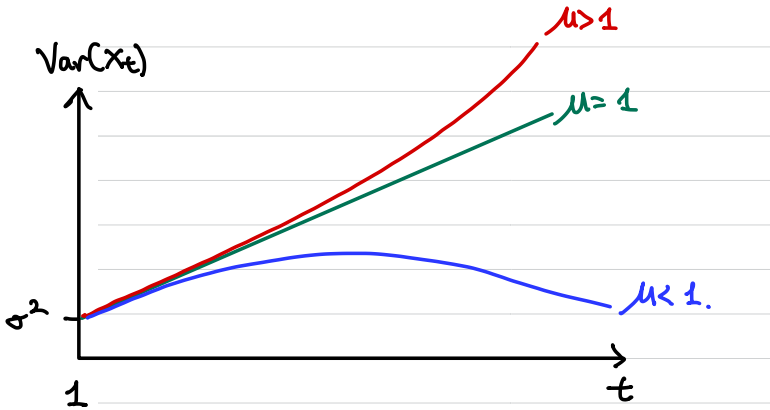
So, $\mathbb{E}[X_t]$ increases geometrically if $\mu > 1$ (supercritical)

" decreases " " $\mu < 1$ (subcritical)

$\mathbb{E}[X_t]$ remains constant if $\mu = 1$ (critical)

Law of Total Variance (Lecture note 18-5)

$$\Rightarrow \mathrm{Var}[X_t] = \sigma^2\, \mathbb{E}[X_{t-1}] + \mu^2\, \mathrm{Var}[X_{t-1}]$$
$$= \sigma^2 \mu^{t-1} + \mu^2(\sigma^2 \mu^{t-2} + \mu^2\, \mathrm{Var}[X_{t-2}])$$
$$= \sigma^2[\mu^{t-1} + \mu^t + \cdots + \mu^{2t-2}]$$
$$= \begin{cases} \sigma^2 t, & \text{if } \mu = 1, \\ \sigma^2 \mu^{t-1}\left(\frac{1-\mu^t}{1-\mu}\right), & \text{if } \mu \neq 1. \end{cases}$$



For all GW processes, the state 0 is the absorbing state.

$$(X_t = 0) \Rightarrow (X_{t+1} = 0)$$

Assumptions:

1) There is no single $k \in \mathbb{N}_0$ s.t. $p_k = 1$.
2) $p_k > 0$ for some $k \geq 2$.
3) $\mu > 0$ and $\sigma^2 > 0$, both finite.

Def (Extinction Time) $\tau = \min\{t \in \mathbb{N}_0 \mid X_t = 0\}$.
$\tau = \infty$ if there exists no such t.

Def (Extinction probability) $\mathbb{P}[\tau < \infty]$.

Claim $\mathbb{P}[\tau > t] \leq \mu^t$.

Pf $\mathbb{P}[\tau > t] = \mathbb{P}[X_t \geq 1] \leq \mathbb{E}[X_t] = \mu^t$ (by Markov's inequality) □

Hence, if $\mu < 1$, then extinction occurs with probability 1.

Key tool for computing the extinction probability is the probability generating function (Lecture 11)

Consider a Galton-Watson process $\{X_t, t \in \mathbb{N}_0\}$ with offspring number distribution $F = \{p_k, k \in \mathbb{N}_0\}$

For $B \sim F$, define

$$\varphi(s) = \mathbb{E}[s^B] = \sum_{k=0}^{\infty} s^k p_k.$$

Assumptions above $\Rightarrow \varphi(s)$ is non-linear

1) $\varphi(0) = p_0$, $\varphi(1) = 1$
2) $\varphi(s)$ is strictly increasing for $s \in (0,1)$
3) $\varphi(s)$ is strictly convex

PGF for X_t

Claim Let $\varphi_t(s) = \mathbb{E}[s^{X_t}] = \sum_{k=0}^{\infty} s^k \mathbb{P}[X_t = k]$

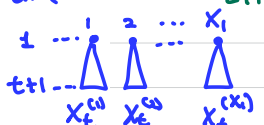
Then, $\varphi_{t+1}(s) = \varphi(\varphi_t(s)) = \varphi_t(\varphi(s))$, $\forall t \in \mathbb{N}_0$

If $X_0 = 1$, then this implies that $\varphi_t(s)$ is the t-fold composition of φ; i.e.,

$$\varphi_t(s) = \underbrace{\varphi(\varphi(\cdots(\varphi(s))\cdots)}_{t \text{ times}}$$

Pf First, note that

$$X_{t+1} \overset{d}{=} \underbrace{X_t^{(1)} + X_t^{(2)} + \cdots + X_t^{(X_1)}}_{\text{iid}}$$



$$\Rightarrow \varphi_{t+1}(s) = \mathbb{E}[(\varphi_t(s))^{X_1}] = \varphi(\varphi_t(s))$$

Furthermore,

$$\varphi_{t+1}(s) = \mathbb{E}[s^{X_{t+1}}] = \sum_{k=0}^{\infty} \underbrace{\mathbb{E}[s^{X_{t+1}} \mid X_t = k]}_{[\varphi(s)]^k} \mathbb{P}[X_t = k]$$

PGF for X_1 with variable $\varphi_t(s)$

$$= \varphi_t(\varphi(s))$$

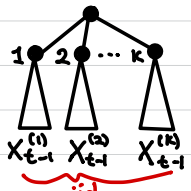
PGF for X_t with variable $\varphi(s)$

□

Claim Let $e_t = \mathbb{P}[X_t = 0]$, the prob. of extinction by time t. Then,

$$e_t = \varphi(e_{t-1})$$

Pf

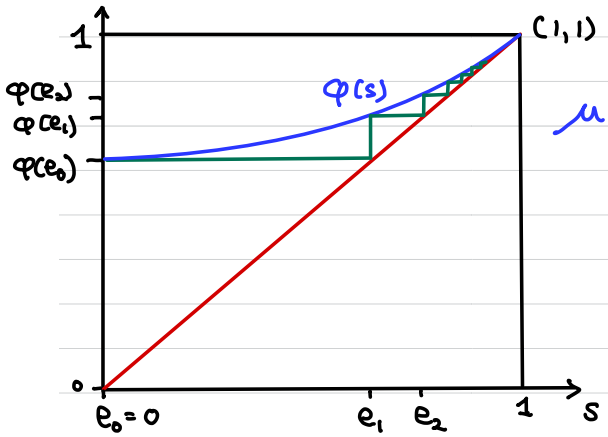


$$e_t = \sum_{k=0}^{\infty} \mathbb{P}[X_{t-1}^{(1)} = 0, \cdots, X_{t-1}^{(X_1)} = 0 \mid X_1 = k] \cdot \mathbb{P}[X_1 = k]$$

$$= \sum_{k=0}^{\infty} (e_{t-1})^k p_k = \varphi(e_{t-1})$$

□

Note: $e_0 = 0$ and $e_t = \varphi_t(0)$. Extinction prob $= \lim_{t\to\infty} \varphi_t(0)$



$\mu \leq 1$

Claim The probability of extinction $\xi = \mathbb{P}[T < \infty]$ is the smallest non-negative solution of the fixed point equation

$$s = \varphi(s). \qquad (*)$$

Pf

$$\xi = p_0 + \sum_{k=1}^{\infty} \xi^k p_k = \varphi(\xi)$$

So, ξ satisfies $(*)$

Note that $\mathbb{P}[X_t = 0] = \varphi_t(0) \leq 1$

$(X_t = 0) \subseteq (X_{t+1} = 0)$

$\Rightarrow \varphi_t(0) \leq \varphi_{t+1}(0)$ non-decreasing sequence bounded from above by 1.

$\Rightarrow \lim_{t \to \infty} \varphi_t(0)$ exists and it is equal to ξ.

Suppose $s \geq 0$ satisfies $(*)$. Then, $\varphi_t(s) = s, \forall t \in \mathbb{N}$.

Monotonicity of $\varphi_t \Rightarrow \varphi_t(0) \leq \varphi_t(s) = s.$

Take the limit of each side as $t \to \infty$

$\Rightarrow \xi \leq s.$

So, ξ is the smallest non-neg solution to $(*)$. □

Thm Unless $p_k = 1$, the fixed-point equation has either one or two solutions.

1) Supercritical ($\mu > 1$) case has a unique solution ξ less than 1.
2) Critical ($\mu = 1$) and subcritical ($\mu < 1$) cases have only one solution $\xi = 1$.

Thm Suppose $\{X_t, t \in \mathbb{N}_0\}$ is a Galton-Watson process with offspring number distribution $F = \{p_k, k \in \mathbb{N}_0\}$.

1) If $\mu < 1$, then $\mathbb{P}[\tau > t] \sim C_F \mu^t$ as $t \to \infty$, where $C_F \in (0, \infty)$ is a constant that depends on F.

(Geometrically decaying tail)

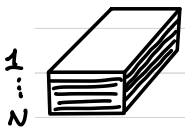
2) If $\mu = 1$, then $\mathbb{P}[\tau > t] \sim \frac{2}{\sigma^2 t}$ as $t \to \infty$

(Fat tail)

$\Rightarrow \mathbb{E}[\tau]$ is infinite.

Lecture 24

Problem of the day



A deck of N cards each with a number written on one side, facing down

Find a strategy with success probability $\approx \frac{1}{e} \approx 37\%$.

Assume:
- N is large and all numbers are distinct.
- The deck is well suffled.

Goal: Get the largest number

Rules:
1) Reveal one card at a time starting from top.
2) STOP at the current card or reveal the next card.
3) If you pass on a card (i.e., do not stop), then you can't return to it.

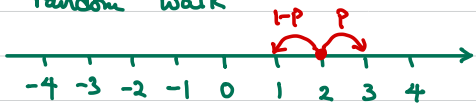
Lecture 24

Def (Discrete-time Markov Chain)

$\{X_n, n \in \mathbb{N}_0\}$ is a Markov Chain if $\forall n \in \mathbb{N}_0$ and $\forall s_0, \dots, s_{n+1} \in S$,

$$\mathbb{P}[X_{n+1} = s_{n+1} \mid X_0 = s_0, \dots, X_n = s_n] = \mathbb{P}[X_{n+1} = s_{n+1} \mid X_n = s_n].$$

e.g. 1d random walk



- Transition probability: $\mathbb{P}[X_{n+1} = j \mid X_n = i]$
- Homogeneous if $\mathbb{P}[X_{n+1} = j \mid X_n = i] = p_{ij}$, $\forall n$ (does not depend on n)
  - assume this in what follows.
- Transition matrix $P = (p_{ij})_{i,j \in S}$
- For every $i \in S$, $\sum_{j \in S} \mathbb{P}[X_{n+1} = j \mid X_n = i] = 1$

  $\Rightarrow$ each row of P sums to 1.
- $\mathbb{P}[X_n = j] = \sum_{i \in S} \mathbb{P}[X_n = j \mid X_{n-1} = i] \mathbb{P}[X_{n-1} = i]$, where $\mathbb{P}[X_n = j \mid X_{n-1} = i] = p_{ij}$

Row vector $\vec{u}_n := (\mathbb{P}[X_n = i])_{i \in S}$

$\vec{u}_n = \vec{u}_{n-1} P \Rightarrow \vec{u}_n = \vec{u}_0 P^n$

$\mathbb{P}[X_n = j \mid X_0 = i] = \mathbb{P}[X_{n+m} = j \mid X_m = i] = [P^n]_{ij} = p_{ij}^{(n)}$

n-step transition probability.

Claim (Chapman-Kolmogorov Equation)

$\forall n, m \in \mathbb{N}$, $p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)}$

Pf

$p_{ij}^{(m+n)} = \mathbb{P}[X_{m+n} = j \mid X_0 = i]$

Law of Total Prob $= \sum_{k \in S} \mathbb{P}[X_{m+n} = j \mid X_0 = i, X_m = k] \mathbb{P}[X_m = k \mid X_0 = i]$

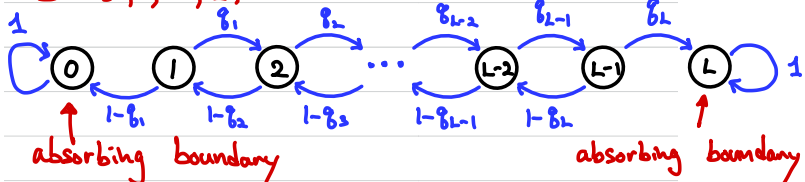
$\mathbb{P}[X_{m+n} = j \mid X_m = k]$ by Markov Property

$= \sum_{k \in S} p_{kj}^{(n)} p_{ik}^{(m)}$.

Alternatively, it follows from $[P^{n+m}]_{ij} = [P^n P^m]_{ij}$ □

A graphical representation of a homogeneous MC

$S = \{0, 1, \dots, L\}$



Some questions of interest: Given $X_0 = i \in \{1, 2, \dots, L-1\}$,

1) What is the probability of hitting 0 before hitting L?
2) How long does it take?
3) How many times is state j visited?

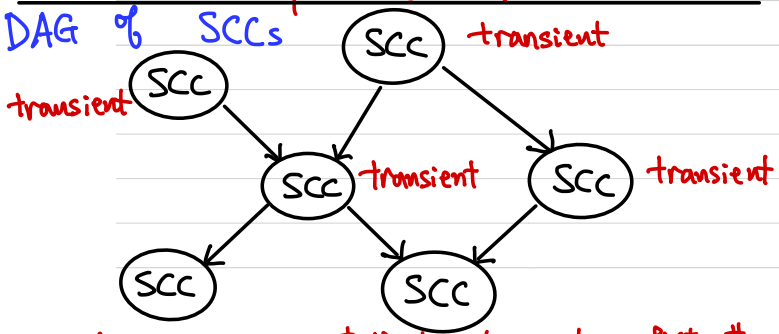
Classification of States

Def $(i \to j)$ State j is **accessible** from state i if $p_{ij}^{(n)} > 0$ for some $n \in \mathbb{N}$. (There is a path from i to j in the graphical representation)

$(i \leftrightarrow j)$ States i, j **intercommunicate** if $i \to j$ and $j \to i$.

- $\leftrightarrow$ defines an equivalence relation.
- The state space S can be partitioned into equivalence classes of $\leftrightarrow$.
- A subset $C \subset S$ is called **irreducible** if $i \leftrightarrow j$ for all $i, j \in C$. C is a strongly connected component (SCC).

DAG of SCCs



Terminal SCC is recurrent if it contains only a finite # states. Otherwise, it can be either transient or recurrent

A Markov chain is said to be **irreducible** if $i \leftrightarrow j \ \forall \ i, j \in S$.

Def (First passage probability)
$$f_{ij}^{(n)} = \mathbb{P}[X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j \mid X_0 = i]$$

Def (Return probability) $f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)}$

Def (Mean recurrence time) $r_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$

Def A state $i \in S$ is called

1) **recurrent or persistent** if $f_{ii} = 1$.
   a) **Null recurrent** if $r_i = \infty$
   b) **Positive recurrent** if $r_i < \infty$
2) **transient** if $f_{ii} < 1$.

These are class properties.

Thm

1) $\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty \iff j$ is recurrent $\Rightarrow \forall i \in S$ s.t. $i \to j$, $\sum_{n=1}^{\infty} p_{ij}^{(n)} = \infty$.

2) $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty \iff j$ is transient $\Rightarrow \forall i \in S$, $\sum_{n=1}^{\infty} p_{ij}^{(n)} < \infty$ and $\lim_{n \to \infty} p_{ij}^{(n)} = 0$

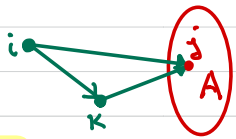
Suppose $S = \{1,2,\cdots,m\} \cup \{m+1,\cdots,n\}$ (B: Transient, A: Absorbing)

Transition probability matrix

$$P = \begin{matrix} B \\ A \end{matrix}\begin{bmatrix} Q & R \\ 0 & S \end{bmatrix} \quad \begin{matrix} B & A \end{matrix}$$

Def (Hitting Probability) for $i \in B$ and $j \in A$,
$h_{ij} = \mathbb{P}[\text{Enter } A \text{ through } j \in A \mid X_0 = i]$

Claim $h_{ij} = p_{ij} + \sum_{k \in B} p_{ik} h_{kj}$



Pf First-step analysis:

$$h_{ij} = \sum_{k \in S} \underbrace{\mathbb{P}[\text{Hit } j \in A \mid X_0 = i, X_1 = k]}_{\mathbb{P}[\text{Hit } j \in A \mid X_1 = k]} \underbrace{\mathbb{P}[X_1 = k \mid X_0 = i]}_{p_{ik}}$$

$$\mathbb{P}[\text{Hit } j \in A \mid X_1 = k] = \begin{cases} \delta_{jk}, & \text{if } k \in A, \\ h_{kj}, & \text{if } k \in B. \end{cases}$$

Kronecker delta $\delta_{jk} = \begin{cases} 1, & \text{if } j = k \\ 0, & \text{otherwise} \end{cases}$

□

In matrix form, $H = (h_{ij})$ satisfies $H = R + QH \Rightarrow H = (I-Q)^{-1}R$ if $(I-Q)^{-1}$ exists.

Lemma Suppose M is a square matrix with $\lim_{n\to\infty} M^n = 0$. Then, $(I-M)^{-1}$ exists and $(I-M)^{-1} = \sum_{n=0}^{\infty} M^n$.

Pf

$$(I-M)(I+M+M^2+\cdots+M^{n-1}) = I - M^n \quad (*)$$

So, $\det(I-M)\det(I+M+\cdots+M^{n-1}) = \det(I-M^n)$ (**)

det is a continuous function $\Rightarrow$

$$\lim_{n\to\infty} \det(I-M^n) = \det\left[\lim_{n\to\infty}(I-M^n)\right] = \det I = 1$$

$\Rightarrow \det(I-M^n) > 0$ for some n

This result and (**) $\Rightarrow \det(I-M) \neq 0$

$\Rightarrow (I-M)^{-1}$ exists.

Taking $n \to \infty$ in (*) gives $(I-M)^{-1} = \sum_{n=0}^{\infty} M^n$.

$$P^n = \begin{bmatrix} Q^n & * \\ 0 & S^n \end{bmatrix} \Rightarrow p_{ik}^{(n)} = q_{ik}^{(n)} \text{ for } i,k \in B$$

k transient $\Rightarrow \lim_{n\to\infty} p_{ik}^{(n)} = 0 \Rightarrow \lim_{n\to\infty} Q^n = 0$

So Lemma $\Rightarrow (I-Q)^{-1}$ exists.

Def $(I-Q)^{-1} = \sum_{n=0}^{\infty} Q^n$ is called the **fundamental matrix** of the absorbing Markov Chain.

Claim (Hitting Time) Let t_i = expected number of steps it takes to hit A given $X_0 = i \in B$.

$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix} = (I-Q)^{-1} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Pf First, note that $t_i = \sum_{j \in B} t_{ij}$, where for $i, j \in B$,

$$t_{ij} := \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{1}\{X_n = j\} \mid X_0 = i\right]$$

$$= \sum_{k \in S} \underbrace{\mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{1}\{X_n = j\} \mid X_0 = i, X_1 = k\right]}_{= \begin{cases} \delta_{ij}, & \text{if } k \in A \\ \delta_{ij} + t_{kj}, & \text{if } k \in B \end{cases}} \underbrace{\mathbb{P}[X_1 = k \mid X_0 = i]}_{P_{ik}}$$

$$= \sum_{k \in S} \delta_{ij} P_{ik} + \sum_{k \in B} t_{kj} P_{ik} = \delta_{ij} + \sum_{k \in B} Q_{ik} t_{kj}$$

In matrix form $T = (t_{ij})$, we have $T = I + QT$

$$\Rightarrow T = (I-Q)^{-1}$$

So, t_i = i^{th} row of $(I-Q)^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

Def (Stationary distribution) The row vector $\pi = (\pi_i)_{i \in S}$ is called a stationary distribution of the Markov chain with transition probability matrix P if

1) $\pi_i \geq 0 \ \forall i \in S$, $\sum_{i \in S} \pi_i = 1$
2) $\pi P = \pi$.

(left eigenvector of P with eigenvalue 1)

If $X_0 \sim \pi$, then $X_n \sim \pi \ \forall n \in \mathbb{N}$.

Q Does every MC have a stationary dist'n? When it exists, is it unique?

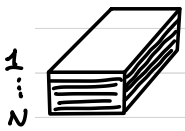
Thm (Perron-Frobenius Theorem)

1) Every MC with finite S has a stationary distribution π.
2) If in addition the chain is irreducible, then π is unique, and $\pi_i = \frac{1}{r_i} \ \forall i \in S$, where

$r_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$ is the mean recurrence time for state $i \in S$.

Lecture 25

Problem from Lecture 24



A deck of N cards each with a number written on one side, facing down

Assume:
- N is large and all numbers are distinct.
- The deck is well suffled.

Goal: Get the largest number

Rules:
1) Reveal one card at a time starting from top.
2) STOP at the current card or reveal the next card.
3) If you pass on a card (i.e., do not stop), then you can't return to it.

Find a strategy with success probability $\approx \frac{1}{e} \approx 37\%$.

Strategy:
1) Reveal a certain proportion, say p, of the cards and record the largest number (denoted M) you have seen.
2) Then, STOP if you see a number larger than M.

Q What is the optimal p?

Success ≡ found the largest number

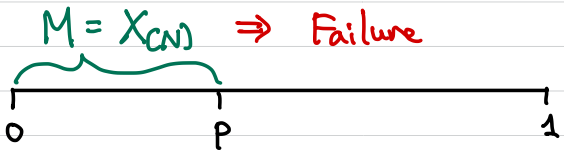
Let $X_1, \ldots, X_N$ denote the numbers.

Order statistics!

$$X_{(1)} < X_{(2)} < \cdots < X_{(N)}$$

Case 1:

$M = X_{(N)} \Rightarrow$ Failure

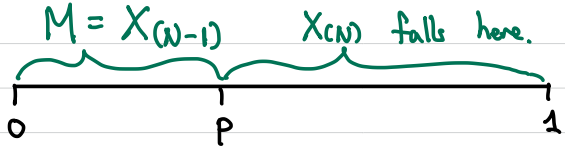


$\mathbb{P}[M = X_{(N)}] = p$

$\mathbb{P}[\text{Success} \mid M = X_{(N)}] = 0$

Case 2:

$M = X_{(N-1)}$ $X_{(N)}$ falls here.

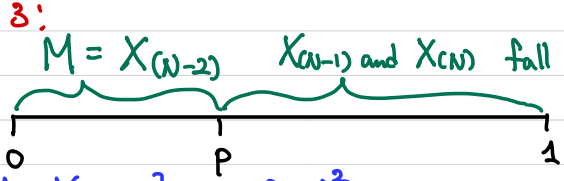


$\mathbb{P}[M = X_{(N-1)}] \approx p(1-p)$

$\mathbb{P}[\text{Success} \mid M = X_{(N-1)}] = 1$

Case 3:

$M = X_{(N-2)}$ $X_{(N-1)}$ and $X_{(N)}$ fall here.



$\mathbb{P}[M = X_{(N-2)}] \approx p(1-p)^2$

$\mathbb{P}[\text{Success} \mid M = X_{(N-2)}]$

$= \mathbb{P}[X_{(N)} \text{ appears before } X_{(N-1)}] = \frac{1}{2}$

General Case:

$\mathbb{P}[M = X_{(N-k)}] \approx p(1-p)^k$

$\mathbb{P}[\text{Success} \mid M = X_{(N-k)}] = \frac{1}{k}$

$$\mathbb{P}[\text{Success}] = \sum_{k=0}^{(1-p)N} \mathbb{P}[\text{Success} \mid M = X_{(N-k)}]\,\mathbb{P}[M = X_{(N-k)}]$$

$$\approx \sum_{k=1}^{(1-p)N} p(1-p)^k \left(\frac{1}{k}\right) \approx -p \ln p$$

which is maximized when $\boxed{p = \frac{1}{e}}$

Lecture 25

The Long Run Behavior of MC

Consider a 2-state Markov Chain $\{X_n, n \in \mathbb{N}_0\}$ with $S = \{1, 2\}$ and transition probability matrix

$$P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}, \text{ where } a, b \in (0, 1].$$

One can show

$$P^n = \frac{1}{a+b}\begin{bmatrix} b & a \\ b & a \end{bmatrix} + \frac{(1-a-b)^n}{a+b}\begin{bmatrix} a & -a \\ -b & b \end{bmatrix}.$$

Suppose $a = 1 = b$. Then,

$$P^n = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \frac{(-1)^n}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \begin{cases} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, & \text{if } n \text{ is odd,} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \text{if } n \text{ is even.} \end{cases}$$

For all other cases, $|1-a-b| < 1$, so $\lim_{n\to\infty} (1-a-b)^n = 0$

$\Rightarrow$ For $(a,b) \neq (1,1)$, $\lim_{n\to\infty} P^n = \frac{1}{a+b}\begin{bmatrix} b & a \\ b & a \end{bmatrix}$.

$\Rightarrow$ In the limit as $n \to \infty$, the distribution of X_n becomes independent of the initial condition X_0.

In fact, $\pi = \frac{1}{a+b}[b \;\; a]$ is the unique stationary distribution of the Markov chain. ($\pi P = \pi$)

In general, every finite-state Markov Chain has a stationary distribution π and

irreducibility $\Rightarrow$ π is unique.

Q Under what condition is $\lim_{n\to\infty} [P^n]_{ij} = \lim_{n\to\infty} \mathbb{P}[X_n = j \mid X_0 = i] = \pi_j, \; \forall\, i, j \in S$?

Def (Period) The period $d(i)$ of a state $i \in S$ is defined as

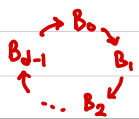
$$d(i) = \gcd\{n \in \mathbb{N} \mid [P^n]_{ii} > 0\}.$$

e.g. In the 2-state MC example with $a = b = 1$, $d(1) = d(2) = 2$.

Def A state $i \in S$ is $\begin{cases} \text{periodic}, & \text{if } d(i) > 1, \\ \text{aperiodic}, & \text{if } d(i) = 1. \end{cases}$

Claim If two states i, j intercommunicate ($i \leftrightarrow j$), then $d(i) = d(j)$.

Thm Let C be a class in a finite-state Markov Chain with period d. Then, C can be partitioned into d subsets $C = B_0 \sqcup B_1 \sqcup \dots \sqcup B_{d-1}$ such that all transitions from $B_k \pmod d$ go to $B_{k+1 \pmod d}$



Thm If a MC with finite state space S is irreducible and aperiodic, then

1) (Regular or ergodic)
$\exists$ a positive integer k, s.t. $[P^k]_{ij} > 0$, $\forall i,j \in S$

2) (Limiting distribution)
$\lim_{n\to\infty} [P^n]_{ij} = \pi_j$, $\forall i,j \in S$ where
$\pi = (\pi_i)_{i\in S}$ is the unique stationary distribution.

$$\left(\lim_{n\to\infty} P^n = \underset{\text{outer product}}{1\pi} = \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix} \right)$$

Another interpretation of the stationary distribution for an irreducible finite MC.

Lemma Consider an irreducible finite MC with transition matrix P and the unique stationary distribution π. Then

$$\lim_{n\to\infty} \frac{I + P + P^2 + \cdots + P^n}{n+1} = \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix}.$$

$$H_j^{(n)} := \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{1}\{X_k = j\}$$

= Fraction of time spent in state j during steps $0, 1, 2, \dots, n$

$$\mathbb{E}\left[H_j^{(n)} \,\middle|\, X_0 = i \right] = \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}[X_k = j \mid X_0 = i]$$

$$= \frac{1}{n+1} \sum_{k=0}^{n} [P^k]_{ij}$$

Lemma $\longrightarrow \pi_j$ as $n \to \infty$.

Thm (WLLN for irreducible finite MC)
For every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left| H_j^{(n)} - \pi_j \right| > \varepsilon \right] = 0,$$

independent of the starting state $X_0 = i$.

A stronger result:

Thm (Ergodic Theorem) Let $\{X_n, n \in \mathbb{N}_0\}$ be an irreducible, positive recurrent MC on state space S and stationary distribution π. Suppose $g: S \to \mathbb{R}$ be a function s.t. $\sum_{i \in S} g(i)\pi_i < \infty$. Then, for any initial distribution for X_0,

$$\frac{1}{n} \sum_{k=1}^{n} g(X_k) \xrightarrow{a.s.} \sum_{i \in S} g(i)\pi_i \text{ as } n \to \infty.$$

Def (First passage time) For $i \in S$,

$$T_i = \min\{n \in \mathbb{N}_0 \mid X_n = i\}$$

Def (Fundamental matrix of irreducible MC)

$$Z = (I - P + 1\pi)^{-1}$$

Remarks:

1) Z is well defined
2) If the MC is aperiodic, then

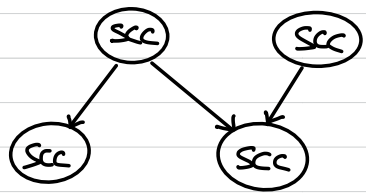
$$Z = I + \sum_{n=1}^{\infty} (P^n - 1\pi)$$

Thm (Mean first passage time)

For an irreducible finite Markov chain with the fundamental matrix $Z = (z_{ij})$ and the unique stationary distribution π,

$$\mathbb{E}[T_j \mid X_0 = i] = \frac{z_{jj} - z_{ij}}{\pi_j}.$$

Suppose a finite MC is not irreducible. (has more than one SCC)



P restricted to each terminal SCC is a valid transition matrix.

$\Rightarrow$ There exists a stationary distribution for each terminal SCC.

Forward process: $X_0 \xrightarrow{(P_{ij})} X_1 \to \cdots \to X_{N-1} \to X_N$

Reversed Process: $Y_N \xleftarrow{(q_{ji})} Y_{N-1} \leftarrow \cdots \leftarrow Y_1 \leftarrow Y_0$

Thm Let $\{X_n, 0 \le n \le N\}$ be an irreducible, positive recurrent MC with transition matrix $P = (P_{ij})$ and unique stationary distribution π. Further, suppose $X_n \sim \pi \quad \forall \ 0 \le n \le N$. Then, the reversed process $\{Y_n = X_{N-n}, 0 \le n \le N\}$ is a MC with transition matrix $Q = (q_{ij})$, where $q_{ij} = \frac{\pi_j P_{ji}}{\pi_i}$.

Pf

$$\begin{aligned}
&\mathbb{P}[Y_{n+1} = j \mid Y_0 = i_0, \dots, Y_n = i_n] \\
&= \frac{\mathbb{P}[Y_0 = i_0, \dots, Y_n = i_n, Y_{n+1} = j]}{\mathbb{P}[Y_0 = i_0, \dots, Y_n = i_n]} \\
&= \frac{\mathbb{P}[X_N = i_0, \dots, X_{N-n} = i_n, X_{N-n-1} = j]}{\mathbb{P}[X_N = i_0, \dots, X_{N-n} = i_n]} \\
&= \frac{\pi_j P_{j,i_n} P_{i_n,i_{n-1}} \cdots P_{i_1,i_0}}{\pi_{i_n} P_{i_n,i_{n-1}} \cdots P_{i_1,i_0}} \\
&= \mathbb{P}[X_{N-n-1} = j \mid X_{N-n} = i_n] = \mathbb{P}[Y_{n+1} = j \mid Y_n = i_n]
\end{aligned}$$

□

Def (Reversibility) The chain $\{X_n\}$ is said to be reversible if $q_{ij} = P_{ij} \ \forall \ i,j$. That is,

(Detailed Balance) $\pi_j P_{ji} = \pi_i P_{ij}, \quad \forall \ i,j$

Thm Let $\{X_n\}$ be an irreducible MC with transition matrix $P = (P_{ij})$ and suppose $\exists$ a distribution $\nu = (\nu_i)$ s.t.

$$\nu_i P_{ij} = \nu_j P_{ji} \quad \forall \ i,j.$$

Then, ν is a stationary distribution of the chain and $\{X_n\}$ is reversible in equilibrium.

Pf

$$\sum_i \nu_i P_{ij} = \sum_i \nu_j P_{ji} = \nu_j \underbrace{\sum_i P_{ji}}_{1} = \nu_j \Rightarrow \nu \text{ is a stationary distn.}$$

Reversibility of $\{X_n\}$ follows from definition. □

STAT 201A Fall 2024

Lecture 25
Supplementary Material on Markov Chains

December 3, 2024

Communication Class Property

Theorem 1 (Class property). Suppose two states $i, j \in S$ of a Markov chain inter-communicate ($i \longleftrightarrow j$), i.e., they belong to the same strongly connected component (SCC) in the graphical representation of the Markov chain. Then,

1. i and j have the same period.
2. i is transient if and only if j is transient.
3. i is null recurrent if and only if j is null recurrent.
4. i is positive recurrent if and only if j is positive recurrent.

Theorem 2. All finite Markov chains have the following properties:

1. At least one state is recurrent.
2. All recurrent states are positive recurrent.
3. If the Markov chain is irreducible (i.e., the corresponding graph consists of a single SCC), then all states are positive recurrent.

Existence and Uniqueness of Stationary Distribution

Theorem 3 (Finite state space S).

1. Every finite Markov chain has a stationary distribution $\vec{\pi}$.
2. Furthermore, if the Markov chain is irreducible, then $\vec{\pi}$ is the unique stationary distribution and $\pi_i = \frac{1}{r_i}, \forall i \in S$, where r_i is the mean recurrence time.

Theorem 4 (Countably infinite state space S).

1. A Markov chain with a countably infinite state space has a stationary distribution $\vec{\pi}$ **if and only if** at least one state is positive recurrent.
2. Furthermore, if the Markov chain is irreducible, then $\vec{\pi}$ (which exists **if and only if** all states are positive recurrent) is unique and $\pi_i = \frac{1}{r_i}, \forall i \in S$, where r_i is the mean recurrence time.

Limit Theorems

For the n-step transition matrix P^n to converge as $n \to \infty$, aperiodicity is crucial.

Recall that $r_j = \infty$ if j is either transient or null recurrent, while $r_j < \infty$ if j is positive recurrent.

Theorem 5 (Limiting distributions).

1. For any aperiodic state $j \in S$ of a Markov chain,

$$\lim_{n\to\infty} [P^n]_{jj} = \frac{1}{r_j} \text{ and } \lim_{n\to\infty} [P^n]_{ij} = \frac{f_{ij}}{r_j}, \forall i \neq j, \text{ where } f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)}.$$

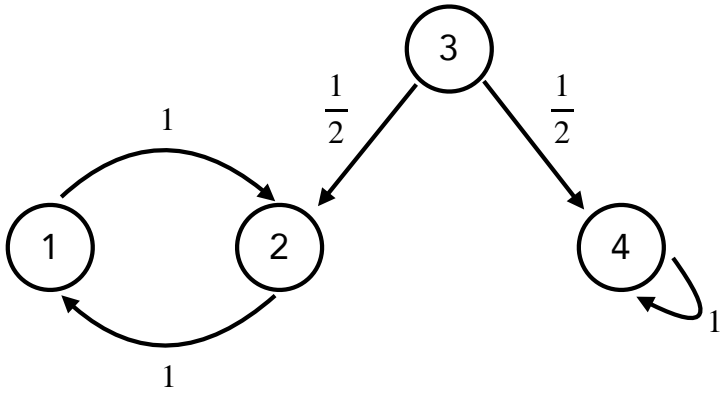
2. If a Markov chain is irreducible and aperiodic, then

$$\lim_{n\to\infty} [P^n]_{ij} = \frac{1}{r_j}, \forall i, j \in S. \text{ (i.e., the limit does not depend on the starting state } i.$$

3. If a Markov chain is irreducible and ergodic (aperiodic and positive recurrent), then

$$\lim_{n\to\infty} [P^n]_{ij} = \frac{1}{r_j} = \pi_j, \forall i, j \in S.$$

Markov Chain 1 (periodic)

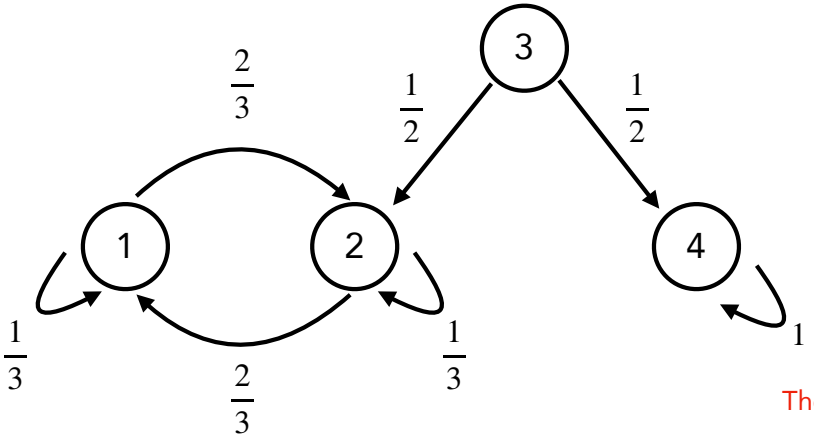


$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$P^{2k+1} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad P^{2k} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- States 1 and 2 are periodic, so P^n does **not** converge as $n \to \infty$.
- State 3 is transient.
- States 1, 2, and 4 are positive recurrent.
- {1,2} is a terminal strongly connected component (SCC), so the transition matrix restricted to {1,2} is a valid transition matrix for a Markov chain on {1,2}.
- By Theorem 3, there exists a unique stationary distribution corresponding to this SCC. More precisely, $\vec{\pi_1} = [1/2,\ 1/2,\ 0,\ 0]$ is the unique stationary distribution for this SCC.
- {4} also is a terminal SCC and the unique stationary distribution corresponding to this SCC is $\vec{\pi}_2 = [0,\ 0,\ 0,\ 1]$.

Markov Chain 2 (aperiodic but not irreducible)



$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Theorem 5, part 1 applies

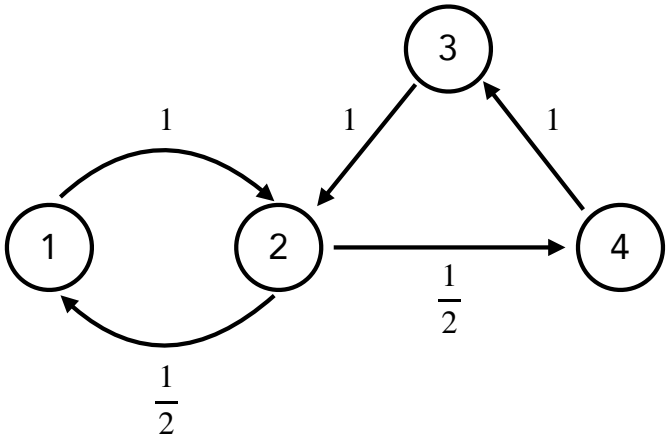
$$\lim_{n\to\infty} P^n = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- All states are aperiodic, so P^n converges as $n \to \infty$.
- State 3 is transient.
- States 1, 2, and 4 are positive recurrent.
- {1,2} is a terminal SCC, so the transition matrix restricted to {1,2} is a valid transition matrix for a Markov chain on {1,2}.
- By Theorem 3, there exists a unique stationary distribution corresponding to this SCC. More precisely, $\vec{\pi}_1 = [1/2,\ 1/2,\ 0,\ 0]$ is the stationary distribution for this SCC.
- {4} also is a terminal SCC and the unique stationary distribution corresponding to this SCC is $\vec{\pi}_2 = [0,\ 0,\ 0,\ 1]$.

Markov Chain 2 (aperiodic but not irreducible)

- **Question:** Let $\vec{\nu} = (\mathbb{P}[X_0 = i])_{i \in S}$ denote the distribution of the initial state. Then, what does $\vec{\nu}P^n$ converge to as $n \to \infty$?
- **Answer:** Since the chain is not irreducible, the answer depends on the choice of ν. For example:
  - If $\vec{\nu} = (a, 1-a, 0, 0)$ where $0 \leq a \leq 1$, then $\lim_{n\to\infty} \vec{\nu}P^n = (1/2, 1/2, 0, 0)$, which is the stationary distribution $\vec{\pi}_1$ corresponding to the terminal SCC {1,2}.
  - If $\vec{\nu} = (0, 0, 0, 1)$, then $\lim_{n\to\infty} \vec{\nu}P^n = (0, 0, 0, 1)$, which is the stationary distribution $\vec{\pi}_2$ corresponding to the terminal SCC {4}.
  - More generally, if $\vec{\nu} = (a, b, c, d)$ where $a, b, c, d \in [0,1]$ such that $a + b + c + d = 1$, then $\lim_{n\to\infty} \vec{\nu}P^n = \left[(a+b) + \frac{c}{2}\right]\vec{\pi}_1 + \left[d + \frac{c}{2}\right]\vec{\pi}_2$.

Markov Chain 3 (irreducible and ergodic)



$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Theorem 5, part 3 applies

$$\lim_{n\to\infty} P^n = \begin{bmatrix} \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} \end{bmatrix}$$

All rows are equal to the stationary distribution $\vec{\pi}$

- All states are aperiodic, so P^n converges as $n \to \infty$.
- This Markov chain is irreducible ({1,2,3,4} is a SCC), so all of its states are positive recurrent (by Theorem 2).
- By Theorem 3, there exists a unique stationary distribution corresponding to this Markov chain. More precisely, $\vec{\pi} = [1/5,\ 2/5,\ 1/5,\ 1/5]$ is the unique stationary distribution.

Markov Chain 3 (irreducible and ergodic)

- In this case, every row of the limiting transition matrix $\lim_{n\to\infty} P^n$ is equal to the unique stationary distribution $\vec{\pi}$.
- Hence, given an arbitrary initial distribution $\vec{\nu} = (\nu_1, \nu_2, \nu_3, \nu_4)$, we obtain $\lim_{n\to\infty} \sum_{i\in S} \nu_i [P^n]_{i,j} = \sum_{i\in S} \nu_i \pi_j = \pi_j \sum_{i\in S} \nu_i = \pi_j.$
- In other words, irrespective of how you initialize the chain at time zero (i.e., how you set X_0), the distribution of X_n will converge to the unique stationary distribution as $n \to \infty$.

Markov Chain 4 (irreducible and aperiodic, but not ergodic)

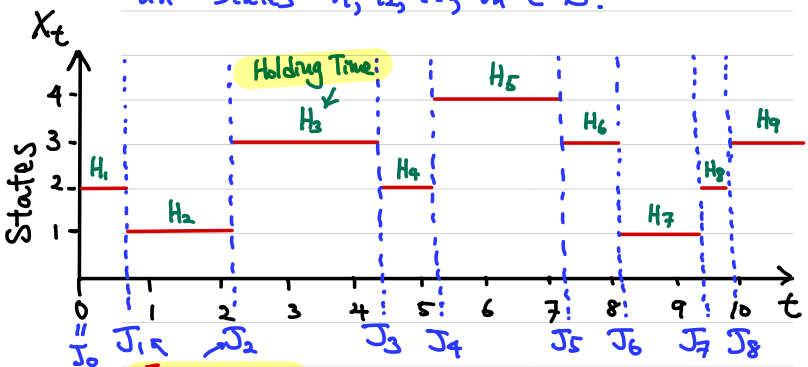
- A finite Markov chain cannot have a null recurrent state and all states of a finite irreducible Markov chain must be positive recurrent (see Theorem 2).
- Hence, for a Markov chain to be irreducible and aperiodic but not ergodic, it must have an infinite state space.

Continuous-Time Markov Chain (CTMC)

$\{X_t, t \geq 0\}$ on finite or countably infinite state space S satisfying

$$\mathbb{P}[X_{t_n} = i_n \mid X_{t_1} = i_1, X_{t_2} = i_2, \dots, X_{t_{n-1}} = i_{n-1}] = \mathbb{P}[X_{t_n} = i_n \mid X_{t_{n-1}} = i_{n-1}]$$

for all times $0 \leq t_1 < t_2 < \cdots < t_n$ and all states $i_1, i_2, \dots, i_n \in S$.



Homogeneous if

$$\mathbb{P}[X_{t+u} = j \mid X_u = i] = \mathbb{P}[X_t = j \mid X_0 = i] =: p_{ij}(t)$$

for all $t, u \geq 0$ and all $i, j \in S$.

$p_{ij}(t)$: transition probability

$P(t) = (p_{ij}(t))_{i,j \in S}$ Transition probability matrix.

Since $p_{ij}(0) = \delta_{ij}$, $P(0) = I$ identity matrix

Claim (Chapman-Kolmogorov Equation)

For homogeneous MC,

$$P(t)\,P(u) = P(t+u), \quad \text{for all } t, u \geq 0.$$

Pf $\forall\, i, j \in S$

$$p_{ij}(t+u) = \mathbb{P}[X_{t+u} = j \mid X_0 = i]$$

$$= \sum_{k \in S} \underbrace{\mathbb{P}[X_{t+u} = j \mid X_u = k, X_0 = i]}_{\text{Markov property} \Rightarrow \mathbb{P}[X_{t+u} = j \mid X_u = k]} \underbrace{\mathbb{P}[X_u = k \mid X_0 = i]}_{p_{ik}(u)}$$

Homogeneity $\Rightarrow = p_{kj}(t)$ □

Def $\{P(t)\}$ is called **standard** if $\lim_{h \downarrow 0} P(h) = I$

Claim $\{P(t)\}$ standard $\Rightarrow$ $p_{ij}(t)$ is a continuous function of t for all $i, j \in S$.

Pf

$$|p_{ij}(t+h) - p_{ij}(t)| = \left| \sum_{k \in S} p_{ik}(h)\, p_{kj}(t) - p_{ij}(t) \right|$$

By CK Eq

$$= \left| (p_{ii}(h) - 1)\, p_{ij}(t) + \sum_{k \in S:\, k \neq i} p_{ik}(h)\, p_{kj}(t) \right| \quad (p_{kj}(t) \leq 1)$$

$$\leq (1 - p_{ii}(h))\, p_{ij}(t) + \sum_{k \in S:\, k \neq i} p_{ik}(h) \quad (\text{sum} = 1 - p_{ii}(h))$$

$\to 0$ as $h \downarrow 0$ since $\lim_{h \downarrow 0} p_{ii}(h) = 1$ □

($\lim_{h\to 0} P(h) = I$)

Thm Let $\{P(t)\}$ be standard transition matrices. Then, for all $i,j \in S$, the following rates exist:

1) $q_i \overset{\text{def}}{=} \lim_{h\to 0} \frac{1 - P_{ii}(h)}{h} \in [0, \infty]$

2) $q_{ij} \overset{\text{def}}{=} \lim_{h\to 0} \frac{P_{ij}(h)}{h} \in [0, \infty]$

Def 1) With $q_{ii} = -q_i$, $Q = (q_{ij})_{i,j\in S}$ is called the generator of the Markov chain.

2) $\{X_t\}$ is called stable if $q_i < \infty \ \forall i \in S$

3) $\{X_t\}$ is called Conservative if $q_i = \sum_{j\neq i} q_{ij}$ for all $i \in S$.

(Every row of Q sums to zero).

In what follows, we consider standard, stable, conservative Markov chains, for which we can write

$$P_{ii}(h) = 1 - q_i h + o(h)$$
$$P_{ij}(h) = q_{ij} h + o(h)$$

Kolmogorov Forward Equation By CK Eq.

$$\frac{dP(t)}{dt} = \lim_{h\to 0} \frac{P(t+h) - P(t)}{h} = \lim_{h\to 0} \frac{P(t)P(h) - P(t)}{h}$$
$$= P(t) \lim_{h\to 0} \frac{[P(h) - I]}{h} = P(t)Q$$

Kolmogorov Backward Equation.

$$\frac{dP(t)}{dt} = \lim_{h\to 0} \frac{P(t+h) - P(t)}{h} = \lim_{h\to 0} \frac{P(h)P(t) - P(t)}{h}$$
$$= \left[\lim_{h\to 0} \frac{[P(h) - I]}{h}\right] P(t) = Q P(t)$$

Initial Condition: $P(0) = I$.

Unique solution:

$$P(t) = e^{tQ} = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}.$$

Claim $\sum_{j\in S} [P(t)]_{ij} = 1, \quad \forall i \in S.$

Pf $\frac{dP(t)}{dt} \cdot \vec{1} = P(t)Q\vec{1} = \vec{0}$ (By KFE) since row sums of Q are all zero.

($\vec{1}$: column vector of 1s)

$\Rightarrow P(t)\cdot\vec{1} = \vec{C}$, a constant vector

$P(0) = I \Rightarrow \vec{C} = \vec{1}$. $\Rightarrow$ Every row sum of $P(t)$ equals 1. $\square$

Claim (Holding Time) Suppose $X(t) = i \in S$. Then,
$H = \inf\{u > 0 : X(t+u) \neq i\} \sim \text{Exp}(q_i)$.

Pf For $a, b > 0$,

$\mathbb{P}[H > a+b \mid H > a] = \mathbb{P}[H > a+b \mid X(t+u) = i, \forall u \in [0,a]]$

By the Markov property & homogeneity $= \mathbb{P}[H > b]$

Exp dist'n is the only continuous distribution with the memoryless property.

$H \sim \text{Exp}(\lambda)$, for some $\lambda > 0$, so

$F_H(u) = \mathbb{P}[H \leq u] = 1 - e^{-\lambda u} \Rightarrow F_H'(0) = \lambda$

Need to find λ:

$1 - F_H(u) = \mathbb{P}[H > u] = [P(u)]_{ii}$

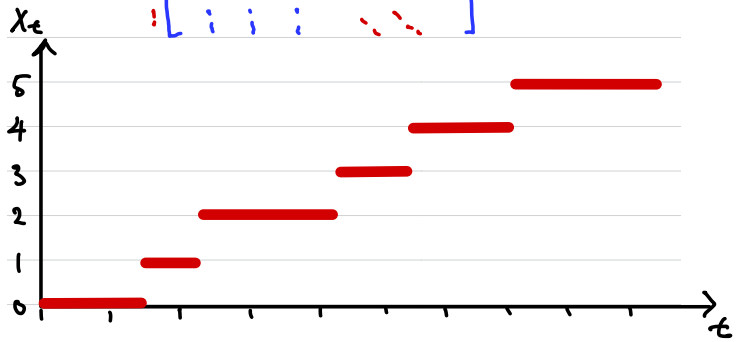
$\Rightarrow F_H'(u) = -[P'(u)]_{ii} \overset{\text{KFE}}{=} -[P(u)Q]_{ii}$

$P(0) = I \Rightarrow \lambda = F_H'(0) = -[IQ]_{ii} = -q_{ii} = q_i$ □

Example 1: Poisson Process

$\{X_t\}$ on $S = \mathbb{N}_0 = \{0, 1, 2, \dots\}$

$$Q = \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{matrix} \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ \vdots & \vdots & \vdots & & \ddots & \ddots & \end{bmatrix}$$



$\mathbb{P}(X_t = j \mid X_0 = 0) = [P(t)]_{0j}$

$P'(t) = P(t)Q \Rightarrow [P'(t)]_{00} = [P(t)]_{00} q_{00} = -\lambda [P(t)]_{00}$ (KFE)

$[P(0)]_{00} = [I]_{00} = 1$

$\Rightarrow [P(t)]_{00} = e^{-\lambda t}$

For $j \geq 1$, $[P'(t)]_{0j} = [P(t)]_{0j} \underbrace{q_{jj}}_{-\lambda} + [P(t)]_{0,j-1} \underbrace{q_{j-1,j}}_{\lambda}$

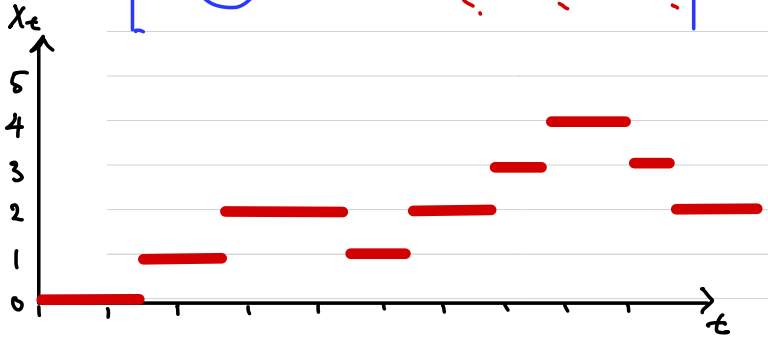
$[P(0)]_{0j} = 0$

$\Rightarrow [P(t)]_{0j} = \frac{(\lambda t)^j}{j!} e^{-\lambda t}$ (Can show by induction.)

Example 2: Birth Death Processes

$\{X_t\}$ on $S = \mathbb{N}_0$

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & & \\ & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \\ & & \mu_3 & -(\lambda_3+\mu_3) & \lambda_3 \\ & & & \ddots & \ddots \end{bmatrix} \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{matrix}$$



A system of coupled ODEs from KBE

$$P'_{0j}(t) = -\lambda_0 P_{0j}(t) + \lambda_0 P_{1j}(t)$$

$$P'_{ij}(t) = \mu_i P_{i-1,j}(t) - (\mu_i+\lambda_i) P_{ij}(t) + \lambda_i P_{i+1,j}(t), \quad i \geq 1$$

Boundary condition $P_{ij}(0) = \delta_{ij}$

- Closed form solutions are known for various special cases.
- Otherwise the system can be solved numerically.

Jump process, jump chain, embedded chain.

Discrete-time MC $\{Y_n, n \in \mathbb{N}_0\}$ given by

$$Y_n = X_{J_n},$$

where J_n = the n^{th} jump time (see fig on page 1)

Thm The transition matrix $\tilde{P} = (\tilde{p}_{ij})$ of the embedded jump chain is given by

$$\tilde{p}_{ii} = \begin{cases} 0, & \text{if } q_{ii} \neq 0, \\ 1, & \text{if } q_{ii} = 0. \end{cases}$$

$$\tilde{p}_{ij} = \frac{q_{ij}}{q_i} = \frac{q_{ij}}{-q_{ii}}, \quad \forall\, i,j \in S \text{ such that } i \neq j.$$

Furthermore, $H_n \perp\!\!\!\perp Y_n$.

H_n: the n^{th} holding time