Probability & Statistics for EECS: Homework #01

Due on 2023-10-15 at 23:59

Name:
Student ID:

Problem 1

(Story Proof) Define $\left\{ \begin{array}{c} n \\ k \end{array} \right\}$ as the number of ways to partition $\{1, 2, \ldots, n\}$ into k non-empty subsets, or the number of ways to have n students split up into k groups such that each group has at least one student. For example, $\left\{ \begin{array}{c} 4 \\ 2 \end{array} \right\} = 7$ because we have the following possibilities:

- $\{1\}, \{2, 3, 4\}$
- $\{2\}, \{1, 3, 4\}$
- $\{3\}, \{1, 2, 4\}$
- $\{4\}, \{1, 2, 3\}$
- $\{1, 2\}, \{3, 4\}$
- $\{1, 3\}, \{2, 4\}$
- $\{1, 4\}, \{2, 3\}$

Prove the following identities:
(a)

$$\left\{ \begin{array}{c} n+1 \\ k \end{array} \right\} = \left\{ \begin{array}{c} n \\ k-1 \end{array} \right\} + k \left\{ \begin{array}{c} n \\ k \end{array} \right\}.$$

Hint: I'm either in a group by myself or I'm not.
(b)

$$\sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) \left\{ \begin{array}{c} j \\ k \end{array} \right\} = \left\{ \begin{array}{c} n+1 \\ k+1 \end{array} \right\}.$$

Hint: First decide how many people are not going to be in my group.

Solution:
(a) When n turns to $n+1$, there may be two possible case

1, the new item becomes a new subset which is independent to the subset of $\left\{ \begin{array}{c} n \\ k-1 \end{array} \right\}$, then in this case we have $\left\{ \begin{array}{c} n \\ k-1 \end{array} \right\}$ because the new item is fixed in one subset.

2, the new item belongs to one subset of $\left\{ \begin{array}{c} n \\ k \end{array} \right\}$. since item can join any any one of the existing k sunsets, in this case we have $k \left\{ \begin{array}{c} n \\ k \end{array} \right\}$ possibilities. In conclusion, we can prove that

$$\left\{ \begin{array}{c} n+1 \\ k \end{array} \right\} = \left\{ \begin{array}{c} n \\ k-1 \end{array} \right\} + k \left\{ \begin{array}{c} n \\ k \end{array} \right\} \tag{1}$$

(b) According to the identities in sub-problem(a), if we extend k to $k+1$, then we have:

$$\left\{ \begin{array}{c} n+1 \\ k+1 \end{array} \right\} = \left\{ \begin{array}{c} n \\ k \end{array} \right\} + (k+1) \left\{ \begin{array}{c} n \\ k+1 \end{array} \right\} \tag{2}$$

then the identity to be proved in sub-problem(b) is equivalent to

$$\sum_{j=k}^{n-1} \left(\begin{array}{c} n \\ j \end{array} \right) \left\{ \begin{array}{c} j \\ k \end{array} \right\} = (k+1) \left\{ \begin{array}{c} n \\ k+1 \end{array} \right\} \tag{3}$$

Then we try to observe the generation process of $\left\{ \begin{array}{c} n \\ k+1 \end{array} \right\}$. Firstly consider then situation that we have k subsets, obviously we need at least $j = k$ items to make sure that each subset contains at least one item. j should be at most $n-1$ because we require at least one item to fill the last($k+1$th) subset. For each j, obviously we have $\left(\begin{array}{c} n \\ k-1 \end{array} \right)$ choices. Noting that in this process there may be repetitive situation because

any one of the $k+1$ subsets can be seen as the last subset which is not considered in the potability discussion of $\left\{ \begin{array}{c} j \\ k \end{array} \right\}$, which means that $\sum_{j=k}^{n-1} \left(\begin{array}{c} n \\ j \end{array} \right) \left\{ \begin{array}{c} j \\ k \end{array} \right\}$ repeats $k+1$ times of $\left\{ \begin{array}{c} n \\ k+1 \end{array} \right\}$, therefore we can prove the identity in equ3, which is equivalent to the original identity in problem(b).

Problem 2

A *norepeatword* is a sequence of at least one (and possibly all) of the usual 26 letters a, b, c, .., z, with repetitions not allowed. For example, "course" is a norepeatword, but "statistics" is not. Order matters, e.g., "course" is not the same as "source". A norepeatword is chosen randomly, with all norepeatwords equally likely. Show that the probability that it uses all 26 letters is very close to $1/e$.

Solution:
Let n represents the length of the norepeatword, $N(n)$ repesents the number of norepeatword whose length is n. Obviously $N(n)$ is equal to the num of ordered sample with no replacement, which is equal to $\frac{K!}{(K-n)!}$, where $K = 26$. Thus the probability that the length of the selected norepeatword letter is 26 can be written as:

$$\frac{K!}{\sum_{j=1}^{K} \frac{K!}{(K-j)!}} = \frac{1}{\sum_{j=1}^{K} \frac{1}{(K-j)!}} = \frac{1}{\sum_{i=0}^{K-1} \frac{1}{i!}} \tag{4}$$

noting that the limitation of the numerator in above number is $1/e$ because the limitation of $\sum_{i=1}^{K} \frac{1}{i!}$ is $e-1$. Thus we can show that the probability that it uses all 26 letters is very close to $1/e$.

Problem 3

Given $n \geq 2$ numbers $(a_1, a_2, \ldots, a_n)$ with no repetitions, a bootstrap sample is a sequence $(x_1, x_2, \ldots, x_n)$ formed from the a_j 's by sampling with replacement with equal probabilities. Bootstrap samples arise in a widely used statistical method known as the bootstrap. For example, if $n = 2$ and $(a_1, a_2) = (3, 1)$, then the possible bootstrap samples are $(3, 3), (3, 1), (1, 3)$, and $(1, 1)$.
(a) How many possible bootstrap samples are there for $(a_1, \ldots, a_n)$?
(b) How many possible bootstrap samples are there for $(a_1, \ldots, a_n)$, if order does not matter (in the sense that it only matters how many times each a_j was chosen, not the order in which they were chosen)?
(c) One random bootstrap sample is chosen (by sampling from $a_1, \ldots, a_n$ with replacement, as described above). Show that not all unordered bootstrap samples (in the sense of (b)) are equally likely. Find an unordered bootstrap sample $\mathbf{b}_1$ that is as likely as possible, and an unordered bootstrap sample $\mathbf{b}_2$ that is as unlikely as possible. Let p_1 be the probability of getting $\mathbf{b}_1$ and p_2 be the probability of getting $\mathbf{b}_2$ (so p_i is the probability of getting the specific unordered bootstrap sample $\mathbf{b}_i$). What is p_1/p_2 ? What is the ratio of the probability of getting an unordered bootstrap sample whose probability is p_1 to the probability of getting an unordered sample whose probability is p_2 ?

Solution:
(a) It is obvious that every time we can choose any one of the n numbers, with n times, so there are n^n possible samples in total.
(b) According to Bose-Einstein Counting, the result is given by $\binom{n+n-1}{n} = \binom{2n-1}{n}$
(c) Since the order does not matter, so a bootstrap with all the elements different has the highest probability to be chosen, since there are many ordered permutations can be merged to it. On the contrary, bootstraps with all the elements the same have the lowest probability to be chosen. According to the statement, p_1 and p_2 can be respectively given by $p_1 = n!/n^n$ and $p_2 = 1/n^n$. Thus, we have $p_1/p_2 = n!$. Because there are n bootstraps which contains the identical number in it, so the probability of getting a bootstrap whose probability is p_1 to the probability of getting a bootstrap whose probability is p_2 is given by $P = n!/n = (n-1)!$.

Problem 4

(Geometric Probability) You get a stick and break it randomly into three pieces. What is the probability that you can make a triangle using such three pieces?

Solution:
It is denoted that the length of the three pieces are x, y, and $1-x-y$, respectively. It is obvious that we have $0 \leq x \leq 1$, $0 \leq y \leq 1$ and $0 \leq (1-x-y) \leq 1$. In order to make a triangle successfully, it is necessary that

- $x+y > 1-x-y \Longrightarrow x+y > 1/2$;
- $x+1-x-y > y \Longrightarrow y < 1/2$;
- $y+1-x-y > x \Longrightarrow x < 1/2$.

With the help of Figure 1, the probability is 1/4.

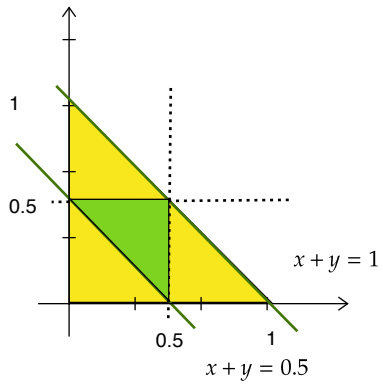


Figure 1: Problem 4.

Problem 5

In the birthday problem, we assumed that all 365 days of the year are equally likely (and excluded February 29). In reality, some days are slightly more likely as birthdays than others. For example, scientists have long struggled to understand why more babies are born 9 months after a holiday. Let $\mathbf{p} = (p_1, p_2, \ldots, p_{365})$ be the vector of birthday probabilities, with p_j the probability of being born on the j th day of the year (February 29 is still excluded, with no offense intended to Leap Dayers). The k th elementary symmetric polynomial in the variables $x_1, \ldots, x_n$ is defined by

$$e_k(x_1, \ldots, x_n) = \sum_{1 \leq j_1 < j_2 < \cdots < j_k \leq n} x_{j_1} \ldots x_{j_k}.$$

This just says to add up all of the $\begin{pmatrix} n \\ k \end{pmatrix}$ terms we can get by choosing and multiplying k of the variables. For example, $e_1(x_1, x_2, x_3) = x_1 + x_2 + x_3$, $e_2(x_1, x_2, x_3) = x_1x_2 + x_1x_3 + x_2x_3$, and $e_3(x_1, x_2, x_3) = x_1x_2x_3$
Now let $k \geq 2$ be the number of people.
(a) Find a simple expression for the probability that there is at least one birthday match, in terms of $\mathbf{p}$ and

an elementary symmetric polynomial.
(b) Explain intuitively why it makes sense that P (at least one birthday match) is minimized when $p_j = \frac{1}{365}$ for all j, by considering simple and extreme cases.
(c) The famous arithmetic mean-geometric mean inequality says that for $x, y \geq 0$

$$\frac{x+y}{2} \geq \sqrt{xy}.$$

This inequality follows from adding $4xy$ to both sides of $x^2 - 2xy + y^2 = (x-y)^2 \geq 0$ Define $\mathbf{r} = (r_1, \ldots, r_{365})$ by $r_1 = r_2 = (p_1 + p_2)/2, r_j = p_j$ for $3 \leq j \leq 365$. Using the arithmetic mean-geometric mean bound and the fact, which you should verify, that

$$e_k(x_1, \ldots, x_n) = x_1 x_2 e_{k-2}(x_3, \ldots, x_n) + (x_1 + x_2) e_{k-1}(x_3, \ldots, x_n) + e_k(x_3, \ldots, x_n)$$

show that $P($ at least one birthday match $\mid \mathbf{p}) \geq P($ at least one birthday match $\mid \mathbf{r})$ with strict inequality if $\mathbf{p} \neq \mathbf{r}$, where the given $\mathbf{r}$ notation means that the birthday probabilities are given by $\mathbf{r}$. Using this, show that the value of $\mathbf{p}$ that minimizes the probability of at least one birthday match is given by $p_j = \frac{1}{365}$ for all j.

Solution:
(a) It is easier to consider the problem inversely, i.e., what is the probability that no person shares the same birthday. Since $e_k(\mathbf{p})$ multiplies and sums k **different** days, and there are $k!$ ways to map each condition to everyone, so the probability for no matching can be given by $k! e_k(\mathbf{p})$, thereby the probability for at least one birthday matching can be given by $1 - k! e_k(\mathbf{p})$.
(b) By considering the simple case where $k = 2$, we have

$$\begin{aligned}
&P(\text{at least one birthday matches}) \\
=&1 - 2e_2(\mathbf{p}) \\
=&(\sum_{i=1}^{365} p_i)^2 - 2 \sum_{1 \leq i < j \leq n} p_i p_j \\
=&\sum_{i=1}^{365} p_i^2 \\
\geq&365 \cdot \left(\frac{\sum_{i=1}^{365} p_i}{365} \right)^2 \\
=&\frac{1}{365},
\end{aligned} \tag{5}$$

where the equation holds if and only if $\forall i, p_i = 1/365$.
(c) For each terms in the expansion of $e_k(x_1, \cdots, x_n)$, it either contains x_1 or not. For those terms not containing x_1, their sum can be written as $e_k(x_2, \cdots, x_n)$. For those terms containing x_1, by extracting the common factor (i.e., x_1), they can be written as $x_1 \cdot e_{k-1}(x_2, \cdots, x_n)$ (this is correct, because there are $\binom{n}{k}$ terms in $e_k(x_1, \cdots, x_n)$, $\binom{n-1}{k}$ terms in $e_k(x_2, \cdots, x_n)$, $\binom{n-1}{k-1}$ terms in $e_{k-1}(x_2, \cdots, x_n)$, and $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$). By applying the same steps to x_2, this conclusion holds.

$$\begin{aligned}
&P(\text{at least one birthday match}|\mathbf{p}) \\
=&1 - k! e_k(\mathbf{p}) \\
=&1 - k!(p_1 p_2 e_{k-2}(p_3, \cdots, p_n) + (p_1 + p_2) e_{k-1}(p_3, \cdots, p_n) + e_k(p_3, \cdots, p_n)) \\
\geq&1 - k! \left(\frac{(p_1 + p_2)^2}{4} e_{k-2}(p_3, \cdots, p_n) + (p_1 + p_2) e_{k-1}(p_3, \cdots, p_n) + e_k(p_3, \cdots, p_n) \right) \\
=&1 - k!(r_1 r_2 e_{k-2}(r_3, \cdots, r_n) + (r_1 + r_2) e_{k-1}(r_3, \cdots, r_n) + e_k(r_3, \cdots, r_n)) \\
=&1 - k! e_k(\mathbf{r}) \\
=&P(\text{at least one birthday match}|\mathbf{r})
\end{aligned} \tag{6}$$

Problem 5 continued on next page...

The proposition can be then proved by contradiction. Assume that $\mathbf{p} = (365^{-1}, \cdots, 365^{-1})$, and $\mathbf{p}' = (p'_1, \cdots, p'_n) \neq \mathbf{p}$ satisfies that $\mathbf{p}'$ minimizes the probability. Since $\mathbf{p}' \neq \mathbf{p}$, there are at least two elements, saying p'_i and p'_j, satisfy that $p'_i \neq p'_j$. Then, a corresponding $\mathbf{r}' = (r'_1, \cdots, r'_n)$ can be further given by $r'_i = r'_j = (p'_i + p'_j)/2$, $r'_k = p'_k (k \neq i, j)$. It is obvious that we have $P(\text{at least one birthday match}|\mathbf{p}') \geq P(\text{at least one birthday match}|\mathbf{r}')$, which contradicts with the assumption. Therefore, the probability is minimized only when $\mathbf{p} = (365^{-1}, \cdots, 365^{-1})$.

Problem 6

(Coupon Collection) If each box of a brand of crispy instant noodle contains a coupon, and there are 108 different types of coupons. Given $n \geq 200$, what is the probability that buying n boxes can collect all 108 types of coupons? You also need to plot a figure to show how such probability changes with the increasing value of n. When such probability is no less than 95%, what is the minimum number of n ?

Solution:
To sample an arbitrary type of coupons from all 108 types for n times, there are 108^n possibilities in total. In order to collect all the 108 types with n boxes, it is equivalent to find a division of n, where it is divided into 108 non-empty subsets and each type of coupon is placed into the corresponding subsets, whose result is given by $\left\{ {n \atop 108} \right\}$. Since the order of these types does not matter, there are 108! permutations in total, so the final probability can be written as $P = 108! \left\{ {n \atop 108} \right\} / 108^n$.
With the help of the formula of Stirling number, i.e., $\left\{ {n \atop m} \right\} = \frac{1}{m!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} (m-k)^n = \sum_{k=0}^{m} (-1)^k \frac{(m-k)^n}{k!(m-k)!}$, the probability can be expanded by $P = \sum_{k=0}^{108} (-1)^k \frac{108!}{k!(108-k)!} \cdot \left(\frac{108-k}{108}\right)^n$.
With numerical experiments, the minimal number of boxes for $P \geq 0.95$ is 823, as shown in Figure 2.

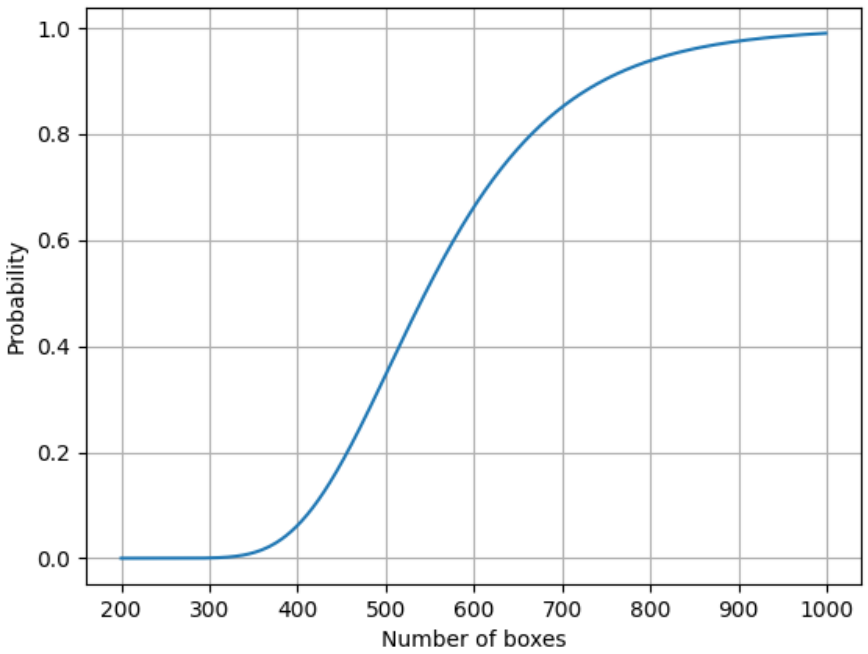


Figure 2: Problem 6.

Probability & Statistics for EECS: Homework #02

Due on Oct 22, 2023 at 23:59

Name:
Student ID:

Oct 17, 2023

Problem 1

Alice is trying to communicate with Bob, by sending a message (encoded in binary) across a channel.
(a) Suppose for this part that she sends only one bit (a 0 or 1), with equal probabilities. If she sends a 0, there is a 5% chance of an error occurring, resulting in Bob receiving a 1; if she sends a 1, there is a 10% chance of an error occurring, resulting in Bob receiving a 0. Given that Bob receives a 1, what is the probability that Alice actually sent a 1?
(b) To reduce the chance of miscommunication, Alice and Bob decide to use a repetition code. Again Alice wants to convey a 0 or a 1, but this time she repeats it two more times, so that she sends 000 to convey 0 and 111 to convey 1. Bob will decode the message by going with what the majority of the bits were. Assume that the error probabilities are as in (a), with error events for different bits independent of each other. Given that Bob receives 110, what is the probability that Alice intended to convey a 1?

Solution:

(a) Suppose event A represents: Alice sends 1, event A^c represents: Alice sends 0, event B represents: Bob receives 1 event B^c represents: Bob receives 0.

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{p(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{0.5 \cdot 0.9}{0.5 \cdot 0.05 + 0.5 \cdot 0.9} \\ &\approx 0.9474. \end{aligned}$$

(b) Suppose event A_1 represents Alice sends 111, event A_0 represents Alice sends 000, and event B represents Bob receives 110.

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} \\ &= \frac{0.9^2 \cdot 0.1 \cdot 0.5}{0.9^2 \cdot 0.1 \cdot 0.5 + 0.05^2 \cdot 0.95 \cdot 0.5} \\ &\approx 0.9715. \end{aligned}$$

Problem 2

Fred decides to take a series of n tests, to diagnose whether he has a certain disease (any individual test is not perfectly reliable, so he hopes to reduce his uncertainty by taking multiple tests). Let D be the event that he has the disease, $p = P(D)$ be the prior probability that he has the disease, and $q = 1 - p$. Let T_j be the event that he tests positive on the jth test.

(a) Assume for this part that the test results are conditionally independent given Fred's disease status. Let $a = P(T_j \mid D)$ and $b = P(T_j \mid D^c)$, where a and b don't depend on the jth test. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the n tests.

(b) Suppose that Fred tests positive on all n tests. However, some people have a certain gene that makes them always test positive. Let G be the event that Fred has the gene. Assume that $P(G) = 1/2$ and that D and G are independent. If Fred does not have the gene, then the test results are conditionally independent given his disease status. Let $a_0 = P(T_j \mid D, G^c)$ and $b_0 = P(T_j \mid D^c, G^c)$, where a_0 and b_0 don't depend on j. Find the posterior probability that Fred has the disease, given that he tests positive on all n of the tests.

Solution

(a) We need to calculate $P\left(D \mid \cap_{j=1}^n T_j\right)$. Use Bayes' formula, LOTP and conditional independence of T_j (if D is given) to obtain following

$$\begin{aligned} P\left(D \mid \cap_{j=1}^n T_j\right) &= \frac{P(D)P\left(\cap_{j=1}^n T_j \mid D\right)}{P\left(\cap_{j=1}^n T_j\right)} = \frac{P(D)\prod_{j=1}^n P(T_j \mid D)}{P\left(\cap_{j=1}^n T_j \mid D\right)P(D) + P\left(\cap_{j=1}^n T_j \mid D^c\right)P(D^c)} \\ &= \frac{P(D)\prod_{j=1}^n P(T_j \mid D)}{P(D)\prod_{j=1}^n P(T_j \mid D) + P(D^c)\prod_{j=1}^n P(T_j \mid D^c)} \\ &= \frac{p \cdot a^n}{p \cdot a^n + (1-p)b^n}. \end{aligned}$$

(b) Again, we need to calculate $P\left(D \mid \cap_{j=1}^n T_j\right)$. Same as in (a), obtain that is

$$P\left(D \mid \cap_{j=1}^n T_j\right) = \frac{P(D)P\left(\cap_{j=1}^n T_j \mid D\right)}{P(D)P\left(\cap_{j=1}^n T_j \mid D\right) + P(D^c)P\left(\cap_{j=1}^n T_j \mid D^c\right)}.$$

Use LOTP and independence of G and D to calculate

$$\begin{aligned} P\left(\cap_{j=1}^n T_j \mid D\right) &= P\left(\cap_{j=1}^n T_j \mid D, G\right)P(G \mid D) + P\left(\cap_{j=1}^n T_j \mid D, G^c\right)P(G^c \mid D) \\ &= P\left(\cap_{j=1}^n T_j \mid D, G\right)P(G) + P\left(\cap_{j=1}^n T_j \mid D, G^c\right)P(G^c) \\ &= \frac{1}{2}P\left(\cap_{j=1}^n T_j \mid D, G\right) + \frac{1}{2}P\left(\cap_{j=1}^n T_j \mid D, G^c\right) \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \prod_{j=1}^n P(T_j \mid D, G^c) \\ &= \frac{1}{2} + \frac{1}{2}{a_0}^n. \end{aligned}$$

Similarly we get that

$$P\left(\cap_{j=1}^n T_j \mid D^c\right) = \frac{1}{2} + \frac{1}{2}{b_0}^n.$$

Plug all these information in to obtain that

$$P\left(D \mid \cap_{j=1}^n T_j\right) = \frac{p\left(\frac{1}{2} + \frac{1}{2}{a_0}^n\right)}{p\left(\frac{1}{2} + \frac{1}{2}{a_0}^n\right) + (1-p)\left(\frac{1}{2} + \frac{1}{2}{b_0}^n\right)}.$$

Problem 3

We want to design a spam filter for email. A major strategy is to find phrases that are much more likely to appear in a spam email than in a no spam email. In that exercise,we only consider one such phrase:free money. More realistically, suppose that we have created a list of 100 words or phrases that are much more likely to be used in spam than in non-spam. Let W_j be the event that an email contains the jth word or phrase on the list. Let

$$p = P(spam), p_j = P(W_j|spam), r_j = P(W_j|not \quad spam)$$

where spam is shorthand for the event that the email is spam.
Assume that $W_1, ..., W_{100}$ are conditionally independent given that the email is spam, and also conditionally independent given that it is not spam. A method for classifying emails (or other objects) based on this kind of assumption is called a naive Bayes classifier. (Here naive refers to the fact that the conditional independence is a strong assumption, not to Bayes being naive. The assumption may or may not be realistic, but naive Bayes classifiers sometimes work well in practice even if the assumption is not realistic.)
Under this assumption we know, for example, that

$$P(W_1, W_2, W_3^c, W_4^c, ..., W_{100}^c|spam) = p_1p_2(1-p_3)(1-p_4)...(1-p_{100}).$$

Without the naive Bayes assumption, there would be vastly more statistical and computational difficulties since we would need to consider $2100 \approx 1.31030$ events of the form$A1 \cap A2... \cap A100$ with each A_j equal to either W_j orW_j^c . A new email has just arrived, and it include the 23rd, 64th, and 65th words or phrases on the list(but not the other 97). So we want to compute

$$P(spam|W_1^c, ..., W_{22}^c, W_{23}, W_{24}^c, ..., W_{63}^c, W_{64}, W_{65}, W_{66}^c, ..., W_{100}^c).$$

Note that we need to condition on all the evidence, not just the fact that $W_{23} \cap W_{64} \cap W_{65}$ occurred. Find the condition probability that the new email is spam (in terms of p and the p_j and r_j).

Solution:
Let W represents $W_1^c, ..., W_{22}^c, W_{23}, W_{24}^c, ..., W_{63}^c, W_{64}, W_{65}, W_{66}^c, ..., W_{100}^c$. Then using Baye's formula, we have:

$$p(spam|W) = \frac{p(spam) \cdot p(W|spam)}{p(spam) \cdot p(W|spam) + p(spam^c) \cdot p(W|spam^c)},$$

Because:
$p(W|spam) = p(W_1^c|spam) \cdot ... \cdot p(W_{22}^c|spam) \cdot p(W_{23}|spam) \cdot p(W_{24}^c|spam) \cdot ... \cdot p(W_{64}|spam) \cdot p(W_{65}|spam) \cdot p(W_{66}^c|spam) \cdot ... \cdot p(W_{100}^c|spam) = (1-p_1)...(1-p_{22})p_{23}(1-p_{24})...p_{64}p_{65}(1-p_{66})...(1-p_{100}),$

$p(W|spam^c) = p(W_1^c|spam^c) \cdot ... \cdot p(W_{22}^c|spam^c) \cdot p(W_{23}|spam^c) \cdot p(W_{24}^c|spam^c) \cdot ... \cdot p(W_{64}|spam^c) \cdot p(W_{65}|spam^c) \cdot p(W_{66}^c|spam^c) \cdot ... \cdot p(W_{100}^c|spam^c) = (1-r_1)...(1-r_{22})r_{23}(1-r_{24})...r_{64}r_{65}(1-r_{66})...(1-r_{100}),$
the equation above can be written as

$$\begin{aligned} &P(spam|W) \\ =&\frac{p(1-p_1)...(1-p_{22})p_{23}(1-p_{24})...p_{64}p_{65}(1-p_{66})(1-p_{100})}{p(1-p_1)...p_{23}(1-p_{24})...p_{65}(1-p_{66})(1-p_{100}) + (1-p)(1-r_1)...(1-r_{22})r_{23}(1-r_{24})...r_{64}r_{65}(1-r_{66})(1-r_{100})} \\ =&\frac{P}{P+Q}, \end{aligned}$$

where,
$P = p(1-p_1)...(1-p_{22})p_{23}(1-p_{24})...p_{64}p_{65}(1-p_{66})...(1-p_{100}),$
$Q = (1-p)(1-r_1)...(1-r_{22})r_{23}(1-r_{24})...r_{64}r_{65}(1-r_{66})...(1-r_{100}).$

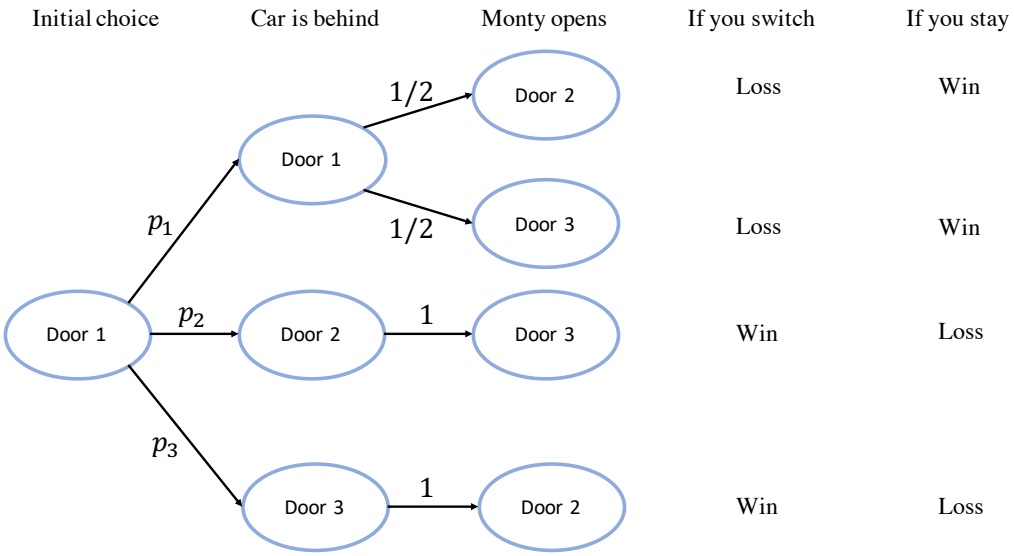
Problem 4

In Monty Hall problem, now suppose the car is not placed randomly with equal probability behind the three doors. Instead, the car is behind door one with probability p_1, behind door two with probability p_2, and behind door three with probability p_3. Here $p_1 + p_2 + p_3 = 1$ and $p_1 \geq p_2 \geq p_3 > 0$. You are to choose one of the three doors, after which Monty will open a door he knows to conceal a goat. Monty always chooses randomly with equal probability among his options in those cases where your initial choice is correct. What strategy should you follow?

Solution:
We define

1. $P_i^{\text{switch}}(\text{win})$ as the probability of winning if "choosing door i first and then switching".
2. $P_i^{\text{stay}}(\text{win})$ as the probability of winning if "choosing door i first and then sticking to initial choice".

When we choose door 1 first, outcomes are shown in the following diagram:



Therefore, we have

$$P_1^{\text{switch}}(\text{win}) = p_2 + p_3,$$
$$P_1^{\text{stay}}(\text{win}) = p_1.$$

Similarly, when we choose door 2 first,

$$P_2^{\text{switch}}(\text{win}) = p_1 + p_3,$$
$$P_2^{\text{stay}}(\text{win}) = p_2.$$

When we choose door 3 first,

$$P_3^{\text{switch}}(\text{win}) = p_1 + p_2,$$
$$P_3^{\text{stay}}(\text{win}) = p_3.$$

Remind that $p_1 \geq p_2 \geq p_3 > 0$. It's not difficult to find that $P_3^{\text{switch}}(\text{win})$ has the maximum winning probability. Hence, in this case, the optimal strategy should be: **choose door 3 first, then switch to the unopened door** after Monty opens some door. Intuitively, we are actually choosing the most unlikely door at the beginning, and then switch to the surviving door, which is the most likely one to be our target.

Problem 5

Consider the Monty Hall problem, except that Monty enjoys opening door 2 more than he enjoys opening door 3, and if he has a choice between opening these two doors, he opens door 2 with probability p, where $1/2 \leq p \leq 1$. To recap: there are three doors, behind one of which there is a car (which you want), and behind the other two of which there are goats (which you don't want). Initially, all possibilities are equally likely for where the car is. You choose a door, which for concreteness we assume is door 1.
(a) Find the unconditional probability that the strategy of always switching succeeds (unconditional in the sense that we do not condition on which of doors 2 or 3 Monty opens).
(b) Find the probability that the strategy of always switching succeeds, given that Monty opens door 2.
(c) Find the probability that the strategy of always switching succeeds, given that Monty opens door 3.

Solution

(a) Let C_j be the event that the car is hidden behind door j and let W be the event that we win using the switching strategy. Using the law of total probability, we can find the unconditional probability of winning in the same way as in class:

$$\begin{aligned} P(W) &= P(W \mid C_1)\,P(C_1) + P(W \mid C_2)\,P(C_2) + P(W \mid C_3)\,P(C_3) \\ &= 0 \cdot 1/3 + 1 \cdot 1/3 + 1 \cdot 1/3 = 2/3. \end{aligned}$$

(b) A tree method works well here (delete the paths which are no longer relevant after the conditioning, and reweight the remaining values by dividing by their sum), or we can use Bayes' rule and the law of total probability (as below).
Let D_i be the event that Monty opens Door i. Note that we are looking for $P(W \mid D_2)$, which is the same as $P(C_3 \mid D_2)$ as we first choose Door 1 and then switch to Door 3. By Bayes' rule and the law of total probability,

$$\begin{aligned} P(C_3 \mid D_2) &= \frac{P(D_2 \mid C_3)\,P(C_3)}{P(D_2)} \\ &= \frac{P(D_2 \mid C_3)\,P(C_3)}{P(D_2 \mid C_1)\,P(C_1) + P(D_2 \mid C_2)\,P(C_2) + P(D_2 \mid C_3)\,P(C_3)} \\ &= \frac{1 \cdot 1/3}{p \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} \\ &= \frac{1}{1+p}. \end{aligned}$$

(c) The structure of the problem is the same as part (b) (except for the condition that $p \geq 1/2$, which was no needed above). Imagine repainting doors 2 and 3, reversing which is called which. By part (b) with $1-p$ in place of p, $P(C_2 \mid D_3) = \frac{1}{1+(1-p)} = \frac{1}{2-p}$.

Problem 6

A/B testing is a form of randomized experiment that is used by many companies to learn about how customers will react to different treatments. For example, a company may want to see how users will respond to a new feature on their website (compared with how users respond to the current version of the website) or compare two different advertisements. As the name suggests, two different treatments, Treatment A and Treatment B, are being studied. Users arrive one by one, and upon arrival are randomly assigned to one of the two treatments. The trial for each user is classified as "success" (e.g., the user made a purchase) or "failure". The probability that the n-th user receives Treatment A is allowed to depend on the outcomes for the previous users. This set-up is known as a two-armed bandit. Many algorithms for how to randomize the treatment assignments have been studied. Here is an especially simple (but fickle) algorithm, called a "stay-with-a-winner" procedure:
(i) Randomly assign the first user to Treatment A or Treatment B, with equal probabilities.
(ii) If the trial for the n-th user is a success, stay with the same treatment for the $(n+1)$-st user; otherwise, switch to the other treatment for the $(n+1)$-st user.
Let a be the probability of success for Treatment A, and b be the probability of success for Treatment B. Assume that $a \neq b$, but that a and b are unknown (which is why the test is needed). Let p_n be the probability of success on the n-th trial and a_n be the probability that Treatment A is assigned on the n-th trial (using the above algorithm).
(a) Show that

$$p_n = (a-b)a_n + b,\ a_{n+1} = (a+b-1)a_n + 1 - b.$$

(b) Use the results from (a) to show that p_{n+1} satisfies the following recursive equation:

$$p_{n+1} = (a+b-1)p_n + a + b - 2ab.$$

(c) Use the result from (b) to find the long-run probability of success for this algorithm, $\lim_{n\to+\infty} p_n$, assuming that this limit exists.
Solution:
(a)

$$\begin{aligned} p_n =& P\{n\text{-th trial succeed}\} \\ =& P\{n\text{-th trial succeed}|\text{Treatment A is assigned on the } n\text{-th trial}\}P\{\text{Treatment A is assigned on the } n\text{-th trial}\} \\ &+ P\{n\text{-th trial succeed}|\text{Treatment B is assigned on the } n\text{-th trial}\}P\{\text{Treatment B is assigned on the } n\text{-th trial}\} \\ =& a \cdot a_n + b \cdot (1-a_n) \\ =& (a-b)a_n + b. \end{aligned} \tag{1}$$

$$\begin{aligned} a_{n+1} =& P\{\text{Treatment A is assigned on the } (n+1)\text{-th trial}\} \\ =& P\{\text{Treatment A is assigned on the } n\text{-th trial}\}P\{n\text{-th trial succeed}\} \\ &+ P\{\text{Treatment B is assigned on the } n\text{-th trial}\}P\{n\text{-th trial failed}\} \\ =& a_n a + (1-a_n)(1-b) \\ =& (a+b-1)a_n + 1 - b. \end{aligned} \tag{2}$$

(b)

$$\begin{aligned}
p_{n+1} =& P\{(n+1)\text{-th trial succeed}\} \\
=& P\{(n+1)\text{-th trial succeed}|\text{Treatment A is assigned on the } (n+1)\text{-th trial}\} \\
& \cdot P\{\text{Treatment A is assigned on the } (n+1)\text{-th trial}\} \\
& + P\{(n+1)\text{-th trial succeed}|\text{Treatment B is assigned on the } (n+1)\text{th trial}\} \\
& \cdot P\{\text{Treatment B is assigned on the } (n+1)\text{-th trial}\} \\
=& a a_{n+1} + b(1 - a_{n+1}) \\
=& (a-b)[(a+b-1)a_n + 1 - b] + b \\
=& (a+b-1)p_n + a + b - 2ab.
\end{aligned} \tag{3}$$

(c) It is denoted that $\lim_{n\to+\infty} p_n = p$. Since the limitation exists, we have

$$p = (a+b-1)p + a + b - 2ab, \tag{4}$$

that is,

$$p = \frac{a+b-2ab}{2-a-b}. \tag{5}$$

Problem 7

(a) An event E_{n+1} is mutually independent of the set of events $E_1, \ldots, En$ if for any subset $I \subseteq [1, n]$

$$P\left(E_{n+1}|\bigcap_{j\in I} E_j\right) = P(E_{n+1}).$$

(b) A dependence graph for the set of events $E_1, \ldots, En$ is a graph $G = (V, E)$ such that $V = \{1, ..., n\}$, and for $i = 1, \ldots, n$, event E_i is mutually independent of the events $\{E_j|(i, j) \notin E\}$.
(c) Assume there exist real numbers $x_1, \ldots, x_n \in [0, 1]$ such that, for any $i(1 \le i \le n)$,

$$P(E_i) \le x_i \prod_{j:(i,j)\in E} (1 - x_j).$$

Then show the following inequality hold:

$$P\left(\bigcap_{i=1}^{n} E_i^c\right) \ge \prod_{i=1}^{n}(1 - x_i).$$

(d) Find the possible applications of the above inequality in the field of EECS.
Solution:
(c) Firstly we assume that when $n < k + 1$, there is $P\left(\bigcap_{i=1}^{n} E_i^c\right) \ge \prod_{i=1}^{n}(1 - x_i)$. Then when $n = k + 1$, we have:

$$\begin{aligned} P\left(\bigcap_{i=1}^{k+1} E_i^c\right) &= P\left(E_{k+1}^c|\bigcap_{i=1}^{k} E_i^c\right) P(\bigcap_{i=1}^{k} E_i^c) \\ &= P\left(E_{k+1}^c|\bigcap_{j:(i,j)\in E} E_i^c\right) P(\bigcap_{i=1}^{k} E_i^c) \\ &\ge \prod_{i=1}^{k}(1 - x_i) P\left(E_{k+1}^c|\bigcap_{j:(i,j)\in E} E_i^c\right) \end{aligned} \tag{6}$$

Thus, the proof of $P\left(\bigcap_{i=1}^{k+1} E_i^c\right) \ge \prod_{i=1}^{k+1}(1 - x_i)$ is equivalent to prove

$$\prod_{i=1}^{k}(1 - x_i) P\left(E_{k+1}^c|\bigcap_{j:(i,j)\in E} E_i^c\right) \ge \prod_{i=1}^{k+1}(1 - x_i) \tag{7}$$

$$P\left(E_{k+1}^c|\bigcap_{j:(i,j)\in E} E_i^c\right) \ge 1 - x_{k+1} \tag{8}$$

$$1 - P\left(E_{k+1}|\bigcap_{j:(i,j)\in E} E_i^c\right) \ge 1 - x_{k+1} \tag{9}$$

$$P\left(E_{k+1}|\bigcap_{j:(i,j)\in E} E_i^c\right) \le x_{k+1} \tag{10}$$

$$\frac{P\left(\bigcap_{j:(i,j)\in E} E_i^c|E_{k+1}\right) P(E_{k+1})}{P\left(\bigcap_{j:(i,j)\in E} E_i^c\right)} \le x_{k+1} \tag{11}$$

Since the inequality $P(E_{k+1}) \leq x_{k+1} \prod_{j:(k+1,j)\in E}(1-x_j)$, the proof can be further transformed to:

$$\frac{P\left(\bigcap_{j:(i,j)\in E} E_i^c | E_{k+1}\right) x_{k+1} \prod_{j:(k+1,j)\in E}(1-x_j)}{P\left(\bigcap_{j:(i,j)\in E} E_i^c\right)} \leq x_{k+1} \tag{12}$$

$$\frac{P\left(\bigcap_{j:(i,j)\in E} E_i^c | E_{k+1}\right)}{P\left(\bigcap_{j:(i,j)\in E} E_i^c\right)} \leq \frac{1}{\prod_{j:(k+1,j)\in E}(1-x_j)} \tag{13}$$

Finally, the problem we have to prove can be transformed to

$$\frac{P\left(\bigcap_{j:(i,j)\in E} E_i^c\right)}{P\left(\bigcap_{j:(i,j)\in E} E_i^c | E_{k+1}\right)} \geq \prod_{j:(k+1,j)\in E}(1-x_j) \tag{14}$$

Obviously $\{j : (i,j) \in E\}$ in $P\left(\bigcap_{j:(i,j)\in E} E_i^c\right)$ is a subset of k, so there is $P\left(\bigcap_{j:(i,j)\in E} E_i^c\right) \geq \prod_{j:(k+1,j)\in E}(1-x_j)$. Since $0 \leq P\left(\bigcap_{j:(i,j)\in E} E_i^c | E_{k+1}\right) \leq 1$, inequality(14) can be proved. Therefore, we can know that when $n = k+1$, the assumption still holds.
Next, we have to prove when $n = 1$ and $n = 2$, the assumption holds.
When $n = 1$, there is $P(E_1) \leq 1 - x_1$. We can easily obtain $P(E_1^c) \geq 1 - x_1$.
When $n = 2$, we should discuss if there exists edge between E_1 and E_2.
(1) no edge: there are $P(E_1) \leq x_1$ and $P(E_2) \leq x_2$, we can easily obtain that $P(E_1^c \cap E_2^c) \geq (1-x_1)(1-x_2)$ since E_1 and E_2 are independent.
(2) exist edge: there are $P(E_1) \leq x_1(1-x_2)$ and $P(E_2) \leq x_2(1-x_1)$. Then there is

$$\begin{aligned} P\left(\bigcap_{i=1}^{2} E_i^c\right) &= P\left((\bigcup_{i=1}^{2} E_i)^c\right) = 1 - P\left(\bigcup_{i=1}^{2} E_i\right) \\ &\geq 1 - (P(E_1) + P(E_2)) \geq 1 - (x_2(1-x_1) + x_2(1-x_1)) \\ &= (1-x_1)(1-x_2) + x_1 x_2 \geq (1-x_1)(1-x_2) \end{aligned} \tag{15}$$

Therefore when $n = 1, 2$ and $n < k+1$, the assumptions always holds. According to strong mathematical induction, we can prove the inequality $P\left(\bigcap_{i=1}^{n} E_i^c\right) \geq \prod_{i=1}^{n}(1-x_i)$.
(d) For instance, in a machine learning system that has several components, each pair of them may exist underlying dependency. So when we try to evaluate the ML system, let $P(E_i)$ represent the test error of the ith component, which is bounded by real numbers related to the ith component and its neighboring components, and the upper bound is $P(E_i) \leq x_i \prod_{j:(i,j)\in E}(1-x_j)$.. We can obtain the lower bound of the probability that the ML system runs correctly by the inequality $P(\cap_{i=1}^{n} E_i^c) \geq \Pi_{i=1}^{n}(1-x_i)$.

Probability & Statistics for EECS: Homework #03

Due on Oct 10, 2023 at 23:59

Name:
Student ID:

Problem 1

狼来了：从前有个放羊娃，每天都把羊群带到山上去吃草, 山里有狼出没．第一天，放羊娃觉得无聊，想要作弄山下耕作的村民。他朝着山下大喊“狼来了!狼来了”，村民们信以为真，冲上山来准备帮助他，发现被欺骗了，大家很生气。第二天，放羊娃故技重施，村民们虽然有点迟疑，但还是冲上山来准备打狼，结果又一次发现被欺骗了，大家非常生气．第三天，狼真的来了，此时放羊娃慌了，哭着向山下大喊“狼来了!狼来了!”，请求村民的帮助．但这一次村民们认为他又在撒谎，无人相信他。最后他所有的羊都被狼吃掉了。

Solution

It is denoted that event A: there come wolves; event B: the boy yells. After the boy plays with farmers, the farmers know that $P(B|A^C)$ is high. With the Bayesian formula, we have

$$P(A^C|B) = \frac{P(B|A^C)P(A^C)}{P(B)}. \tag{1}$$

With $P(A)$ and $P(B)$ fixed, this probability increases with $P(B|A^C)$, so the farmers have a higher probability of not trusting the boy.

Problem 2

A fair die is rolled repeatedly, and a running total is kept (which is, at each time, the total of all the rolls up until that time). Let p_n be the probability that the running total is ever exactly n (assume the die will always be rolled enough times so that the running total will eventually exceed n, but it may or may not ever equal n).
(a) Write down a recursive equation for p_n (relating p_n to earlier terms p_k in a simple way). Your equation should be true for all positive integers n, so give a definition of p_0 and p_k for $k < 0$ so that the recursive equation is true for small values of n.
(b) Find p_7.
(c) Give an intuitive explanation for the fact that $p_n \ \ 1/3.5 = 2/7$ as $n \to \infty$.

Solution

(a) For an arbitrary integer n, it can be rolled by rolling a $(n-1)$ and a 1, or a $(n-2)$ and a 2, ..., or a $(n-6)$ and a 6, where temporarily ignore the limitation of positive integers. Thus we have

$$p_n = \frac{1}{6}(p_{n-1} + p_{n-2} + p_{n-3} + p_{n-4} + p_{n-5} + p_{n-6}) \tag{2}$$

On one hand, it is easy to find that $p_1 = 1/6$. On the other hand, $p_1 = 1/6 \cdot (p_0 + p_{-1} + p_{-2} + p_{-3} + p_{-4} + p_{-5})$. Consider practical scenarios, where a 0 can always be rolled without rolling it, and a negative number can never be rolled. Thereby, we have

$$\begin{cases} p_0 = 1, \\ p_i = 0, \quad i = -1, \ldots, -5. \end{cases} \tag{3}$$

(b) Since

$$p_7 = \frac{1}{6}(p_6 + p_5 + p_4 + p_3 + p_2 + p_1) \tag{4}$$

We have

$$\begin{aligned}
p_1 &= \frac{1}{6} * p_0 = \frac{1}{6} \\
p_2 &= \frac{1}{6} * (p_0 + p_1) = \frac{1}{6} * (1 + \frac{1}{6}) = \frac{7}{36} \\
p_3 &= \frac{1}{6} * (p_0 + p_1 + p_2) = \frac{1}{6} * (1 + \frac{1}{6} + \frac{7}{36}) = \frac{49}{216} \\
p_4 &= \frac{1}{6} * (p_0 + p_1 + p_2 + p_3) = \frac{1}{6} * (1 + \frac{1}{6} + \frac{7}{36} + \frac{49}{216}) = \frac{343}{1296} \\
p_5 &= \frac{1}{6} * (p_0 + p_1 + p_2 + p_3 + p_4) = \frac{1}{6} * (1 + \frac{1}{6} + \frac{7}{36} + \frac{49}{216} + \frac{343}{1296}) = \frac{2401}{7776} \\
p_6 &= \frac{1}{6} * (p_0 + p_1 + p_2 + p_3 + p_4 + p_5) = \frac{1}{6} * (1 + \frac{1}{6} + \frac{7}{36} + \frac{49}{216} + \frac{343}{1296} + \frac{2401}{7776}) = \frac{16807}{46656}
\end{aligned}$$

Then, we have

$$p_7 = \frac{70993}{279936}$$

(c) As $n \to +\infty$, the gap of rolling different numbers is negligible, where the expectation for each "increment" is given by $1/6 \cdot (1 + 2 + 3 + 4 + 5 + 6) = 7/2$, so the probability for rolling a certain number is given by $p_n \approx 1/(7/2) = 2/7$.

Problem 3

A sequence of $n \geq 1$ independent trials is performed, where each trial ends in "success" or "failure" (but not both). Let p_i be the probability of success in the i^{th} trial, $q_i = 1 - p_i$, and $b_i = q_i - 1/2$, for $i = 1, 2, ..., n$. Let A_n be the event that the number of successful trials is even.
(a) Show that for $n = 2$, $P(A_2) = 1/2 + 2b_1b_2$.
(b) Show by induction that $P(A_n) = 1/2 + 2^{n-1}b_1b_2...b_n$ (This result is very useful in cryptography. Also, note that it implies that if n coins are flipped, then the probability of an even number of Heads is $1/2$ if and only if at least one of the coins is fair.) Hint: Group some trials into a super-trial.
(c) Check directly that the result of (b) is true in the following simple cases: $p_i = 1/2$ for some i; $p_i = 0$ for all i; $p_i = 1$ for all i.
solution:

(a) Suppose, S is the trail ends "success", F is the trail ends "fail". We have $P(A_2) = P(S, S) + P(F, F)$.

$$P(A_2) = \ q_1 \cdot q_2 + p_1 \cdot p_2.$$

According to the equation, it is easy to get $q_i = b_i + 1/2$, $p_i = 1/2 - b_i$.Thus,

$$\begin{aligned} P(A_2) = \ & (b_1 + 1/2) \cdot (b_2 + 1/2) + (1/2 - b_1) \cdot (1/2 - b_2) \\ = \ & 1/2 + 2b_1b_2. \end{aligned}$$

(b)

$$P(A_1) = q_1 = 1/2 + b_1$$

$$P(A_2) = 1/2 + 2b_1b_2$$

Assume $P(A_{n-1}) = 1/2 + 2^{n-2}b_1b_2...b_{n-1}$, $n \geq 2$.

$$\begin{aligned} P(A_n) &= P(A_{n-1})q_n + [1 - P(A_{n-1})]p_n \\ &= (1/2 + 2^{n-2}b_1 \cdots b_{n-1})(1/2 + b_n) \\ &\quad + (1/2 - 2^{n-2}b_1 \cdots b_{n-1})(1/2 - b_n) \\ &= 1/2 + 2^{n-1}b_1 \cdots b_n \end{aligned}$$

(c)

- $p_i = 1/2 = q_i$, $b_i = 0$, for all i, which means the probability of success or failure is the same. So, $P(A_n) = 1/2 = 1/2 + 0$, True.

- $p_i = 0$, $b_i = 1/2$ for all i, which means the probability of success is 0, $A_n = 0$. So, $P(A_n) = 1 = 1/2 + 2^{n-1}(1/2)^n$, True.

- $p_i = 1$, $b_i = -1/2$, for all i, which means the probability of success is 1, $A_n = n$. So, when n is even, $P(A_n) = 1 = 1/2 + 2^{n-1}(-1/2)^n$. When n is odd, $P(A_n) = 0 = 1/2 + 2^{n-1}(-1/2)^n$, True.

Problem 4

A message is sent over a noisy channel.The message is a sequence $x_1, x_2, ..., x_n$ of n bits ($x_i \in \{0, 1\}$). Since the channel is noisy, there is a chance that any bit might be corrupted, resulting in an error ($a_0$becomes a_1 or vice versa). Assume that the error events are independent. Let p be the probability that an individual bit has an error($0 < p < 1/2$).Let $y_1, y_2, ..., y_n$ be the received message (so $y_i = x_i$ if there is no error in that bit,but $y_i = 1 - x_i$ if there is an error there).
To help detect errors, the n th bit is reserved for a parity check:x_n is defined to be 0 if $x_1 + x_2 + ... + x_{n-1}$ is even, and 1 if $x_1 + x_2 + ... + x_{n-1}$ is odd. When the message is received, the recipient checks whether y_n has the same parity as $y_1 + y_2 + ... + y_{n_1}$. If the parity is wrong, the recipient knows that at least one error occurred; otherwise, the recipient assumes that there were no errors.
(a) For $n = 5, p = 0.1$, what is the probability that the received message has errors which go undetected?
(b) For general n and p, write down an expression (as a sum) for the probability that the received message has errors which go undetected.
(c) Give a simplified expression, not involving a sum of a large number of terms, for the probability that the received message has errors which go undetected.

solution:

(a) If the error message is undetected, the total number of error bits should be even: if the number of error bits in $x_1, x_2, \ldots, x_{n-1}$ is even(> 0), then the last bit x_n should be right; if the number of error bits in $x_1, x_2, \ldots, x_{n-1}$ is odd, then the last bit x_n should be wrong.

$$\begin{aligned} P(\text{undetected error message}) &= \binom{5}{4} p^4(1-p) + \binom{5}{2} p^2(1-p)^3 \\ &= 0.07335 \end{aligned}$$

(b) According to (a), we have

$$P(\text{undetected error message}) = \sum_{k \text{ is even}}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

(c)

$$\begin{aligned} P(\text{undetected error message}) &= \frac{(p + (1-p))^n + (-1)^n (p - (1-p))^n}{2} - (1-p)^n \\ &= \frac{1 + (1-2p)^n}{2} - (1-p)^n \end{aligned}$$

Problem 5

For x and y binary digits (0 or 1), let $x \oplus y$ be 0 if $x = y$ and 1 if $x \neq y$ (this operation is called exclusive or (often abbreviated to XOR), or addition mod 2).
(a) Let $X \sim Bern(p)$ and $Y \sim Bern(1/2)$, independently. What is the distribution of $X \oplus Y$
(b) With notation as in sub-problem(a), is $X \oplus Y$ independent of X? Is $X \oplus Y$ independent of Y? Be sure to consider both the case $p = 1/2$ and the case $p \neq 1/2$.
(c) Let $X_1, ..., X_n$ be i.i.d. (i.e., independent and identically distributed) Bern(1/2) R.V.s. For each nonempty subset J of $\{1, 2, ..., n\}$, let

$$Y_J = \oplus_{Y \in J} X_J.$$

Show that Y_J $Bern(1/2)$ and that these $2^n - 1$ R.V.s are pairwise independent, but not independent.
Solution:

1. Let $Z = X \oplus Y$. When $Z = 1$ is the same as $X = 1, Y = 0$ or $X = 0, Y = 1$, also $X \sim Bern(p)$ and $Y \sim Bern(1/2)$, thus we can get:

$$\begin{aligned} p(Z = 1) = &\ p(X = 1, Y = 0) + p(X = 0, Y = 1) \\ = &\ p(X = 1)p(Y = 0) + +p(X = 0)p(Y = 1) \\ = &\ p * 1/2 + (1 - p) * 1/2 \\ = &\ 1/2. \end{aligned}$$

   For the same reason, we can get $p(Z = 0) = 1/2$. Therefor, $Z \sim Bern(1/2)$.

2. To show whether Z is independent of X, it is the same to verify whether:

$$p(Z = z, X = x) = p(Z = z)p(X = x)$$

   Let's consider the left side respectively:

$$p(Z = 0, X = 1) = p(Y = 1, X = 1) = \frac{1}{2}p,$$

$$p(Z = 0, X = 0) = p(Y = 0, X = 0) = \frac{1}{2}(1 - p),$$

$$p(Z = 1, X = 1) = p(Y = 0, X = 1) = \frac{1}{2}p,$$

$$p(Z = 1, X = 0) = p(Y = 1, X = 0) = \frac{1}{2}(1 - p).$$

   Then consider the right side:

$$p(Z = 0)p(X = 1) = \frac{1}{2}p,$$

$$p(Z = 0)p(X = 0) = \frac{1}{2}(1 - p),$$

$$p(Z = 1)p(X = 1) = \frac{1}{2}p,$$

$$p(Z = 1)p(X = 0) = \frac{1}{2}(1 - p),$$

   No matter the value of p, the equation above is always true. Thus, Z is independent of X for all the case.

   To show whether Z is independent of Y, it is very similar to the above. For the left side:

$$p(Z = 0, Y = 1) = \frac{1}{2}p,$$

$$p(Z=0,Y=0)=\frac{1}{2}(1-p),$$
$$p(Z=1,Y=1)=\frac{1}{2}(1-p),$$
$$p(Z=1,Y=0)=\frac{1}{2}p.$$

For the right side:

$$p(Z=z)p(Y=Y)=\frac{1}{2}*\frac{1}{2}=\frac{1}{4}.$$

To make sure the equation is always true, we should guarantee $\frac{1}{2}p=\frac{1}{2}(1-p)=\frac{1}{4}$ for all time. This is true only when $p=\frac{1}{2}$.Thus, only when $p=\frac{1}{2}$, Z is independent of Y.

3. Let l denotes the length of subset J. Then use Mathematical induction to prove the equation. As we know:
When $l=1$, $p(Y_J=1)=p(X_J=1)=\frac{1}{2}$, $Y_J\sim Bern(1/2)$.
Suppose $l=k$, $Y_J\sim Bern(1/2)$.
Then when $l=k+1$, let $\hat{J}=J\cup\{\hat{j}\}$, where the length of J is k, Thus:

$$\begin{aligned}p(Y_{\hat{J}}=1)=&\ p(Y_J=1,X_{\hat{j}}=0)+p(Y_J=0,X_{\hat{j}}=1)\\=&\ \frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\\=&\ 1/2.\end{aligned}$$

Therefore, $Y_J\sim Bern(1/2)$.
To show that they are pairwise independent,but not independent, is equal to verify:

$$p(Y_mY_n)=p(Y_m)p(Y_n),\forall m,n,$$
$$p(Y_1Y_2...Y_{2^n-1})\neq p(Y_1)p(Y_2)...p(Y_{2^n-1}).$$

For the first equation, there are two occasions.
$Y_m\cap Y_n=\emptyset$:
as $X_1,X_2...,X_n$ are IID distribution, Y_m and Y_n are obviouly independent.
$Y_m\cap Y_n\neq\emptyset$:
let $p=Y_m\cap Y_n$, $s=Y_m-Y_m\cap Y_n$, $q=Y_n-Y_m\cap Y_n$,

$$\begin{aligned}P(Y_m=1,Y_n=1)=&\ =P(p=1)p(s=0)p(q=0)+P(p=0)p(s=1)p(q=1)\\=&\ \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\\=&\ 1/4\\=&\ P(Y_m=1)P(Y_n=1).\end{aligned}$$

Thus, they are pairwise independent.

For the second equation, we can list a counterexample:

$$P(Y_1,Y_2,Y_3)=\frac{1}{4},$$
$$P(Y_1)P(Y_2)P(Y_3)=\frac{1}{8},$$

where $Y_1=x_1\oplus x_2$, $Y_2=x_2\oplus x_3$, $Y_3=x_3\oplus x_1$. Therefore, they are not independent.

Problem 6

By LOTP for problems with recursive structure, we generate many difference equations. To solve the difference equation in the form of

$$f_{i+1} = b \cdot f_i + a \cdot f_{i-1}, i \geq 1. \tag{5}$$

where a and b are constants, we turn to the so-called characteristic equation:

$$x^2 = bx + a. \tag{6}$$

If such equation has two distinct roots r_1 and r_2, then the general form of f_i is

$$f_i = c \cdot r_1^i + d \cdot r_2^i, \tag{7}$$

If there is only one distinct root r, then the general form of f_i is

$$f_i = c \cdot r^i + d \cdot i \cdot r^i. \tag{8}$$

Show the mathematical principle behind the method of characteristic equation.

Solution

If we multiply x^{n-2} on both sides of the characteristic equation, there is $x^n = ax^{n-1} + bx^{n-1}$, which satisfies the recursive structure.

Firstly we consider the case that two distinct roots r_1, r_2 are different. What we have to prove is that for any constant c, d, $f_i = cr_1^i + dr_2^i$ is the solution of the recursive equation $f_i = af_{i-1} + bf_{i-2}$.

Noting that

$$\begin{aligned} & af_{n-1} + bf_{n-2} \\ &= a(cr_1^{n-1} + dr_2^{n-1}) + b(cr_1^{n-2} + dr_2^{n-2}) \\ &= c(ar_1^{n-1} + br_1^{n-2}) + d(ar_2^{n-1} + br_2^{n-2}) \\ &= cr_1^n + dr_2^n \\ &= f_n \end{aligned} \tag{9}$$

Thus, when $n \geq 2$, any linear combination of r_1^n, r_2^n is sufficient to represent the solution of the recursive equation.

When $n = 0, 1$, given initial value of f_0 and f_1, there is

$$\begin{aligned} c + d &= f_0 \\ cr_1 + dr_2 &= f_1 \end{aligned} \tag{10}$$

Hence if $r_1 \neq r_2$, there exists a unique solution to the linear equations. Thus we complete the proof that for any c, d, $f_n = cr_1^n + dr_2^n$ is the general solution of the recursive equation.

Next, we discuss the situation when $r_1 = r_2 = r$. obviously $f_n = xr^n$ is one solution of the recursive equation $f_n = af_{n-1} + bf_{n-2}$. Since r is the only one distinct root of $x^2 = ax + b$, it is also the only one distinct root of $x^n = ax^{n-1} + bx^{n-2}$, take the derivative from both sides, we have

$$nx^n = a(n-1)x^{n-1} + b(n-2)x^{n-2} \tag{11}$$

therefore, $f_n = nr^n$ is also one special solution of the recursive equation. Then similar to the process of proof under $r_1 \neq r_2$, we can prove that the linear combination of r^n and nr^n is the solution of the recursive equation when $n \geq 2$. When $n = 0, 1$, given initial value of f_0 and f_1, there is

$$\begin{aligned} c &= f_0 \\ cr + dr &= f_1 \end{aligned} \tag{12}$$

obviously, there is a unique solution of c and d.

To sum up, both cases have been proved.

Probability & Statistics for EECS: Homework #04

Due on Oct 10, 2023 at 23:59

Name:
Student ID:

Problem 1

Let X have PMF

$$P(X = k) = cp^k/k \ for \ k = 1, 2, ...,$$

where p is a parameter with $0 < p < 1$ and c is a normalizing constant. We have $c = -1/log(1-p)$, as seen from the Taylor series

$$-log(1-p) = p + \frac{p^2}{2} + \frac{p^3}{3} + .$$

This distribution is called the Logarithmic distribution (because of the log in the above Taylor series), and has often been used in ecology. Find the mean and variance of X.

Solution:

1.

$$\begin{aligned} E[X] = & \sum_{k=1}^{\infty} k \cdot p(X = k) \\ = & \ c \cdot \sum_{k=1}^{\infty} k \cdot \frac{p^k}{k} \\ = & \ c \cdot \sum_{k=1}^{\infty} p^k \\ = & \ c \cdot \frac{p}{1-p} \\ = & \ -\frac{p}{(1-p)log(1-p)}. \end{aligned}$$

2.

$$\begin{aligned} E[X^2] = & \sum_{k=1}^{\infty} k^2 \cdot c \cdot \frac{p^k}{k} \\ = & \sum_{k=1}^{\infty} kcp^k. \end{aligned}$$

As $\sum_{k=1}^{\infty} p^k = \frac{p}{1-p}$, we can get $\sum_{k=1}^{\infty} kp^{k-1} = \frac{1}{(1-p)^2}$.

$$\begin{aligned} E[X^2] = & \ cp \sum_{k=1}^{\infty} k \cdot p^{k-1} \\ = & \ \frac{cp}{(1-p)^2}. \end{aligned}$$

$$\begin{aligned} Var[X] = & \ E[X^2] - (E[X])^2 \\ = & \ \frac{cp(1-cp)}{(1-p)^2}. \end{aligned}$$

Problem 2

Nick and Penny are independently performing independent Bernoulli trials. For concreteness, assume that Nick is flipping a nickel with probability p_1 of Heads and Penny is flipping a penny with probability p_2 of Heads. Let $X_1, X_2, \ldots$ be Nick's results and $Y_1, Y_2, \ldots$ be Penny's results, with $X_i \sim Bern(p_1)$ and $Y_j \sim Bern(p_2)$.
(a) Find the distribution and expected value of the first time at which they are simultaneously successful, i.e., the smallest n such that $X_n = Y_n = 1$.
Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.
(b) Find the expected time until at least one has a success (including the success).
Hint:Define a new sequence of Bernoulli trials and use the story of the Geometric.
(c)For $p_1 = p_2$, find the probability that their first successes are simultaneous, and use this to find the probability that Nick's first success precedes Penny's.
Solution:

1. Let $Z_i = 1$, if $X_i = Y_i = 1$, otherwise, $Z_i = 0$.

$$p(Z = k) = (1 - P_1P_2)^{K-1}P_1P_2 \sim Geom(p_1p_2).$$

   Thus, $E(Z) = \frac{1}{p_1p_2}$.

2. Let $Z_i = 0$, if $X_i = Y_i = 0$, otherwise, $Z_i = 1$.

$$p(Z = k) \sim Geom\left(1 - (1 - p_1)(1 - p_2)\right).$$

   Thus, $E(Z) = \frac{1}{p_1+p_2-p_1p_2}$.

3. Let X denotes Nick first success and Y denotes Penny first success, and $p_1 = p_2 = p$,$q = 1 - p$.

$$\begin{aligned} p(X = Y) &= \sum_{k=1}^{\infty} p^2(q^2)^{k-1} \\ &= p^2 \sum_{k=0}^{\infty} (q^2)^k \\ &= \frac{p^2}{1 - q^2} \\ &= \frac{p}{2 - p}. \end{aligned}$$

   Based on above,

$$\begin{aligned} p(X > Y) &= \frac{1 - \frac{p}{2-p}}{2} \\ &= \frac{1 - p}{2 - p}. \end{aligned}$$

Problem 3

A building has n floors, labeled $1, 2, \ldots, n$. At the first floor, k people enter the elevator, which is going up and is empty before they enter. Independently, each decides which of floors $2, 3, \ldots, n$ to go to and presses that button (unless someone has already pressed it).
(a) Assume for this part only that the probabilities for floors $2, 3, \ldots, n$ are equal. Find the expected number of stops the elevator makes on floors $2, 3, \ldots, n$.
(b) Generalize (a) to the case that floors $2, 3, \ldots, n$ have probabilities $p_2, \ldots, p_n$ (respectively); you can leave your answer as a finite sum.
solution:

(a) Let X be the number of stops. $X = X_1 + X_2 + \ldots + X_n$, where $X_i = 1$ if someone stops at ith floor, otherwise $X_i = 0$.

$$\begin{aligned} E(X_i) &= 1 \cdot \mathrm{P}(\text{at least one person stop at } i\text{th floor}) \\ &= 1 - (\frac{n-2}{n-1})^k. \end{aligned}$$

Thus,

$$\begin{aligned} E(X) &= E(X_1 + X_2 + \cdots + X_n) \\ &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= (n-1)\left[1 - (\frac{n-2}{n-1})^n\right] \end{aligned}$$

(b) According to (a)

$$\begin{aligned} E(X_i) &= 1 \cdot \mathrm{P}(\text{at least one person stop at } i\text{th floor}) \\ &= 1 - (1-p_i)^k. \end{aligned}$$

Thus, we have

$$E(X) = \sum_{i=2}^{n} 1 - (1-p_i)^k = n - 1 - (\sum_{i=2}^{n}(1-p_i)^k)$$

Problem 4

(a) Use LOTUS to show that for $X \sim \text{Pois}(\lambda)$ and any function g, $E(Xg(X)) = \lambda E(g(X+1))$. This is called the *Stein-Chen identity* for the Poisson.
(b) Find the third moment $E(X^3)$ for $X \sim \text{Pois}(\lambda)$ by using the identity from (a) and a bit of algebra to reduce the calculation with the fact that X has mean λ and variance λ.
solution:

(a) From $X \sim \text{Poisson}(\lambda)$ we have $P(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}$, $k \in \mathbb{N}$. Denote $f(x) = Xg(X)$, we have

$$\begin{aligned} E[Xg(X)] =& \sum_{x=0}^{+\infty} f(x)P(X = x) \\ =& \sum_{x=0}^{+\infty} xg(x)\frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \\ =& \lambda \sum_{x=0}^{+\infty} g(x)\frac{\lambda^{(x-1)} \mathrm{e}^{-\lambda}}{(x-1)!} \end{aligned}$$

Denote $Y = X - 1$, we have

$$\begin{aligned} E[Xg(X)] =& \lambda \sum_{x=0}^{+\infty} g(x)\frac{\lambda^{(x-1)} \mathrm{e}^{-\lambda}}{(x-1)!} \\ =& \lambda \sum_{y=0}^{+\infty} g(y+1)\frac{\lambda^{(y)} \mathrm{e}^{-\lambda}}{(y)!} \\ =& \lambda E(g(Y)) \\ =& \lambda E(g(X+1)) \end{aligned}$$

(b) Let $g(X) = X^3$

$$\begin{aligned} E(X^3) =& \lambda E[(X+1)^2] \\ =& \lambda[E(X^2) + E(2X) + 1] \\ =& \lambda(\lambda + \lambda^2 + 2\lambda + 1) \\ =& \lambda^3 + 3\lambda^2 + \lambda. \end{aligned}$$

Problem 5

People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, i.e., before person X arrives there are no two people with the same birthday, but when person X arrives there is a match. Assume for this problem that there are 365 days in a year, all equally likely. By the result of the birthday problem form Chapter 1, for 23 people there is a 50.7% chance of a birthday match (and for 22 people there is a less than 50% chance). But this has to do with the median of X; we also want to know the mean of X, and in this problem we will find it, and see how it compares with 23.

(a) A median of an r.v. Y is a value m for which $P(Y \leq m) \geq 1/2$ and $P(Y \geq m) \geq 1/2$. Every distribution has a median, but for some distributions it is not unique. Show that 23 is the unique median of X.

(b) Show that $X = I_1 + I_2 + \cdots + I_{366}$, where I_j is the indicator r.v. for the event $X \geq j$. Then find $E(X)$ in terms of p_j's defined by $p_1 = p_2 = 1$ and for $3 \leq j \leq 366$,

$$p_j = (1 - \frac{1}{365})(1 - \frac{2}{365}) \cdots (1 - \frac{j-2}{365})$$

(c) Compute $E(X)$ numerically.

(d) Find the variance of X, both in terms of the p_j's and numerically.

Hint:What is I_i^2, and what is $I_i I_j$ for $i < j$? Use this to simplify the expansion

$$X^2 = I_1^2 + \cdots + I_{366}^2 + 2\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_i I_j.$$

Note: In addition to being an entertaining game for parties, the birthday problem has many applications in computer science, such as in a method called the birthday attack in cryptography. It can be shown that if there are n days in a year and n is large, then $E(X) \approx \sqrt{\pi n/2}$. In Volume 1 of his masterpiece The Art of Computer Programming, Don Knuth shows that an even better approximation is

$$E(X) \approx \sqrt{\frac{\pi n}{2}} + \frac{2}{3} + \sqrt{\frac{\pi}{288n}}.$$

Solution:

(a) For an arbitrary pair of people, the probability of having the same birthday is 1/365. It is denoted that the number of birthday match is Z. Since in the corresponding number of samples is relatively large and the probability is small, we have

$$P(\text{At least one birthday match}) = P(Z \geq 1) = 1 - P(Z = 0) \approx 1 - \mathrm{e}^{\lambda}, \tag{1}$$

where $\lambda = \binom{m}{2}p$, m is the number of people, and p is the probability. Therefore, we have

$$P(X \leq 23) \approx 1 - \mathrm{e}^{\lambda} \approx 0.5002 \geq 0.5. \tag{2}$$

On the other hand, we have

$$\begin{aligned} P(X \geq 23) =& P(\text{No match before 23}) \approx \mathrm{e}^{\lambda} \\ =& \mathrm{e}^{\binom{22}{2}\cdot\frac{1}{365}} \approx 0.531 \geq 0. \end{aligned} \tag{3}$$

Thus, 23 is the unique median of X.

(b) For X, it can always be expressed with the sum of binary indicators since it is not decreasing. Then we have

$$\begin{aligned} E(X) =& E(I_1 + I_2 + \cdots + I_{366}) \\ =& E(I_1) + E(I_2) + \cdots + E(I_{366}) \\ =& \sum_{i=1}^{366} p_j. \end{aligned} \tag{4}$$

(c)

$$\begin{aligned} E(X) =& \sum_{i=1}^{366} p_j \\ =& 1 + 1 + \left(1 - \frac{1}{365}\right) + \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) + \cdots + \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{364}{365}\right) \\ \approx& 24.62 \end{aligned} \tag{5}$$

(d)

$$\begin{aligned} E(X^2) =& E\left(I_1^2 + I_2^2 + I_3^2 + ... + I_{366}^2 + 2\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_i I_j\right) \\ =& E(I_1^2) + E(I_2^2) + E(I_3^2) + ... + E(I_{366}^2) + 2E\left(\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_j\right) \\ =& \sum_{j=1}^{366} p_j + 2\sum_{j=1}^{366}(j-1)E(I_j) \\ =& \sum_{j=1}^{366}(2j-1)p_j. \end{aligned} \tag{6}$$

Thus, we have

$$D(X) = E(X^2) - [E(X)]^2 \approx 148.64. \tag{7}$$

Problem 6

Suppose a fair coin is tossed repeatedly, and we obtain a sequence of H and T (H denotes Head and T denotes Tail). Let N denote the number of tosses to observe the first occurrence of the pattern "HHH". Find $E(N)$ and $Var(N)$.

Solution:

Suppose $P(H) = p, q = 1 - p, P_k = P(N = k), k = 0, 1, 2, 3, \cdots$, and $P_0 = P_1 = P_2 = 0, P_3 = p^3, P_4 = qp^3$. Suppose $S_i =$ result of the i th toss, $S_i = H$ or T. According to Law of Total Probability, when $k \geq 4$

$$\begin{aligned}
P_k &= P(N = k) \\
&= P(N = k \mid S_1 = H) P(S_1 = H) + P(N = k \mid S_1 = T) P(S_1 = T) \\
&= P(N = k, S_1 = H) + P(N = k, S_1 = T) \\
P(N = k, S_1 = H) &= P(N = k, S_1 = H \mid S_2 = H) P(S_2 = H) + P(N = k, S_1 = H \mid S_2 = T) P(S_2 = T) \\
&= P(N = k, S_1 = H, S_2 = H) + P(N = k, S_1 = H, S_2 = T) \\
P(N = k, S_1 = H, S_2 = H) &= P(S_1 = H) P(S_2 = H) P(S_3 = T) P(N = k - 3) \\
&= qp^2 P_{k-3} \\
P(N = k, S_1 = H, S_2 = T) &= P(S_1 = H) P(S_2 = T) P(N = k - 2) \\
&= qpP_{k-2} \\
P(N = k, S_1 = T) &= P(S_1 = T) P_{k-1} \\
&= qP_{k-1}
\end{aligned}$$

So $P_k = qp^2 P_{k-3} + qpP_{k-2} + qP_{k-1}$. The PGF of N is :

$$\begin{aligned}
g(t) = E\left(t^N\right) &= \sum_{k=0}^{\infty} P_k t^k \\
&= \sum_{k=3}^{\infty} P_k t^k \\
&= P_3 t^3 + \sum_{k=4}^{\infty} P_k t^k \\
&= p^3 t^3 + \sum_{k=4}^{\infty} P_k t^k
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\sum_{k=4}^{\infty} P_k t^k &= \sum_{k=4}^{\infty} \left(qp^2 P_{k-3} + qpP_{k-2} + qP_{k-1}\right) t^k \\
&= \sum_{k=4}^{\infty} qp^2 P_{k-3} t^k + \sum_{k=4}^{\infty} qpP_{k-2} t^k + \sum_{k=4}^{\infty} qP_{k-1} t^k \\
&= qp^2 t^3 \sum_{k=4}^{\infty} P_{k-3} t^{k-3} + qpt^2 \sum_{k=4}^{\infty} P_{k-2} t^{k-2} + qt \sum_{k=4}^{\infty} P_{k-1} t^{k-1} \\
&= qp^2 t^3 \sum_{k=1}^{\infty} P_k t^k + qpt^2 \sum_{k=2}^{\infty} P_k t^k + qt \sum_{k=3}^{\infty} P_k t^k \\
&= \left(qp^2 t^3 + qpt^2 + qt\right) g(t)
\end{aligned}$$

Problem 6 continued on next page...

$$
\begin{aligned}
g(t) &= p^3t^3 + \left(qp^2t^3 + qpt^2 + qt\right) g(t) \\
g(t) &= \frac{p^3t^3}{1 - (qp^2t^3 + qpt^2 + qt)} \\
&= \frac{t^3}{8 - t^3 - 2t^2 - 4t} \\
E(N) &= g'(t)|_{t=1} = g'(1) = 14 \\
\mathrm{Var}(N) &= g''(1) + g'(1) - (g'(1))^2 = 142
\end{aligned}
$$

Problem 7

Show the following theorems:

1. Given a complete graph $K_n (n \geq 3)$, if $\binom{n}{m} 2^{-\binom{m}{2}+1} < 1$, then it is possible to color the edges of K_n with two colors so that it has no monochromatic K_m subgraph $(1 < m < n)$.

2. Let $M \in F(x_1, x_2, \ldots, x_n)$ be a non-zero polynomial of total degree $d \geq 0$ over a field F. Let S be a finite subset of F and let $r_1, r_2, \ldots, r_n$ be selected at random independently and uniformly from S. Then
$$P[M(r_1, r_2, \ldots, r_n) = 0] \leq \frac{d}{|S|} \tag{8}$$

Solution:

1. Here we color all the edges with the two colors randomly. Define event K_m: **subgraph k_m is not monochromatic**, the probability of it is $P(K_m) = 1 - 2 * (\frac{1}{2})^{\binom{m}{2}}$, since the probability of $P(K_m^c)$ is $2 * (\frac{1}{2})^{\binom{m}{2}}$.

   The probability of the event that **there is no monochromatic subgraph in the graph** equals to $P(K_1 \cap K_2 \cdots K_{\binom{n}{m}})$, since we must ensure that every subgraph is not monochromatic. Then according to the following theorem:
$$P(A_1 \cap A_2 \cap \cdots \cap A_n) \geq P(A_1) + P(A_2) + \cdots + P(A_n) - (n-1), \tag{9}$$
   We have
$$P(K_1 \cap K_2 \cdots \cap K_{\binom{n}{m}}) \geq \binom{n}{m}(1 - 2 * (\frac{1}{2})^{\binom{m}{2}}) - (\binom{n}{m} - 1) > 0 \tag{10}$$
   Thus we can obtain that $\binom{n}{m} 2 * (\frac{1}{2})^{\binom{m}{2}} < 1$, so we complete the proof.

2. When $n = 1$, $M(x_1)$ at most have d roots, so $P(M(r_1) = 0) \leq \frac{d}{|S|}$ holds. Assume that when $n = k$, the inequality holds. Then when $n = k + 1$, there is
$$M(x_1, x_2, \ldots, x_{k+1}) = \sum_{i=1}^{d} x_{k+1}^i M_i(x_1, x_2, \ldots, x_k) \tag{11}$$
   Since the degree of $x_{k+1} M_i(x_1, x_2, \ldots, x_k)$ is at most d, the degree of M_i is less or equal to $d - i$. Therefore we can further know that $P(M_i(r_1, \ldots, r_n) = 0) \leq \frac{d-i}{|S|}$.
   When $M_i(r_1, \ldots, r_k) \neq 0$, the degree of $M(r_1, \ldots, r_k, x_{k+1})$ is i. Therefore there is
$$P(M(r_1, \ldots, r_k, r_{k+1}) = 0 | M_i(r_1, \ldots, r_k) \neq 0) \leq \frac{i}{|S|} \tag{12}$$
   Then according to LOTP, there is
$$\begin{aligned}
&P(M(r_1, \ldots, r_k, r_{k+1}) = 0) \\
&= P(M(r_1, \ldots, r_k, r_{k+1}) = 0, M_i(r_1, \ldots, r_k) \neq 0) \\
&\quad + P(M(r_1, \ldots, r_k, r_{k+1}) = 0, M_i(r_1, \ldots, r_k) = 0) \\
&= P(M_i(r_1, \ldots, r_k) \neq 0) P(M(r_1, \ldots, r_k, r_{k+1}) = 0 | M_i(r_1, \ldots, r_k) \neq 0) \\
&\quad + P(M_i(r_1, \ldots, r_k) = 0) P(M(r_1, \ldots, r_k, r_{k+1}) = 0 | M_i(r_1, \ldots, r_k) = 0) \\
&\leq P(M_i(r_1, \ldots, r_k) = 0) + P(M(r_1, \ldots, r_k, r_{k+1}) = 0 | M_i(r_1, \ldots, r_k) \neq 0) \\
&\leq \frac{d-i}{|S|} + \frac{i}{|S|} \\
&= \frac{d}{|S|}
\end{aligned} \tag{13}$$

Thus we complete the proof.

Probability & Statistics for EECS: Homework #5 Solutions

Professor Ziyu Shao

Problem 1

(a) The Cauchy distribution has PDF

$$f(x) = \frac{1}{\pi\,(1+x^2)}$$

for all x. Find the CDF of a random variable with the Cauchy PDF.

Hint: Recall that the derivative of the inverse tangent function $\tan^{-1}(x)$ is $\frac{1}{1+x^2}$.

(b) The Pareto distribution with parameter $a > 0$ has PDF

$$f(x) = \frac{a}{x^{a+1}}$$

for $x \geq 1$ (and 0 otherwise). This distribution is often used in statistical modeling. Find the CDF of a Pareto r.v. with parameter a; check that it is a valid CDF.

Solution

(a) Given that the PDF of the Cauchy distribution is

$$f(x) = \frac{1}{\pi(1+x^2)},$$

and the hint that the derivative of the inverse tangent function $\tan^{-1}(x)$ is $\frac{1}{1+x^2}$, we can calculate the CDF of the Cauchy distribution by definition, *i.e.*, integrating the PDF over range $(-\infty, x]$.

Therefore, the CDF of the Cauchy distribution $F(x)$ is as follows:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \frac{1}{\pi}\tan^{-1}(t)\Big|_{-\infty}^{x} = \frac{1}{\pi}\tan^{-1}(x) - \frac{1}{\pi}(-\frac{\pi}{2}) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x),$$

where $x \in (-\infty, \infty)$.

(b) Given that the PDF of the Pareto distribution with parameter $\alpha > 0$ is

$$f(x) = \begin{cases} \frac{a}{x^{a+1}}, & x \geq 1 \\ 0, & \text{Otherwise} \end{cases},$$

we can calculate the CDF of the Pareto distribution by definition, *i.e.*, integrating the PDF over range $(-\infty, x]$.

Therefore, the CDF of the Pareto distribution $F(x)$ is as follows:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{1}^{x} \frac{a}{t^{a+1}}dt = -\frac{1}{t^a}\Big|_{1}^{x} = 1 - \frac{1}{x^a},$$

where $x \in [1, \infty)$. When $x \in (-\infty, 1)$, by definition, $F(x) = 0$.

We then check if $F(x)$ is a valid CDF as follows:

- Increasing: Due to the fact that $\frac{1}{x^a}, a > 0$ is decreasing over $[1, \infty)$, CDF $F(x) = 1 - \frac{1}{x^a}$ is increasing over the corresponding support $[1, \infty)$.
- Right-continuous: Due to the fact that $1 - \frac{1}{x^a}, a > 0$ is continuous over $[1, \infty)$, CDF $F(x)$ is right-continuous over the corresponding support $[1, \infty)$.
- Convergence to 0 and 1 in the limits: Due to the fact that $F(x) = 0, x < 1$ and $\lim_{x\to\infty} \frac{1}{x^a} = 0$ when $a > 0$, CDF $F(x)$ have its limits as follows:

$$\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1 - 0 = 1.$$

In summary, the CDF $F(x)$ is valid.

Problem 2

The Exponential is the analog of the Geometric in continuous time. This problem explores the connection between Exponential and Geometric in more detail, asking what happens to a Geometric in a limit where the Bernoulli trials are performed faster and faster but with smaller and smaller success probabilities.
Suppose that Bernoulli trials are being performed in continuous time; rather than only thinking about first trial, second trial, etc., imagine that the trials take place at points on a timeline. Assume that the trials are at regularly spaced times 0, Δt, $2\Delta t$,... , where Δt is a small positive number. Let the probability of success of each trial be $\lambda\Delta t$, where λ is a positive constant. Let G be the number of failures before the first success (in discrete time), and T be the time of the first success (in continuous time).

(a) Find a simple equation relating G to T. Hint: Draw a timeline and try out a simple example.

(b) Find the CDF of T. Hint: First find $P(T > t)$.

(c) Show that as $\Delta t \to 0$, the CDF of T converges to the Expo(λ) CDF, evaluating all the CDFs at a fixed $t \geq 0$.

Solution

(a) $T = G\Delta t$.

(b) For $t \geq 0$, $P(T > t) = P\left(G > \frac{t}{\Delta t}\right) = P\left(\text{no success in the first } \left\lfloor \frac{t}{\Delta t} \right\rfloor \text{ trials}\right) = (1-\lambda\Delta t)^{\left\lfloor \frac{t}{\Delta t} \right\rfloor + 1}$. Thus The CDF of T is

$$P(T \leq t) = 1 - P(T > t) = 1 - (1-\lambda\Delta t)^{\left\lfloor \frac{t}{\Delta t} \right\rfloor + 1}.$$

(c) As $\Delta t \to 0$,

$$\begin{aligned}\lim_{\Delta t \to 0} P(T \leq t) &= \lim_{\Delta t \to 0} \left[1 - (1-\lambda\Delta t)^{\left\lfloor \frac{t}{\Delta t} \right\rfloor + 1}\right] = 1 - \lim_{\Delta t \to 0} (1-\lambda\Delta t)^{\frac{t}{\Delta t}} \\ &= 1 - \lim_{\Delta t \to 0} \left[(1-\lambda\Delta t)^{\frac{1}{\lambda\Delta t}}\right]^{\lambda t} = 1 - e^{-\lambda t}.\end{aligned}$$

Thus for $t \geq 0$, the CDF of T converges to the Expo(λ) CDF as $\Delta t \to 0$.

Problem 3

Let X be a Pois (λ) random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The bias of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, *i.e.*, how far off the estimate is on average; the estimator is unbiased if its bias is 0.

(a) For estimating λ, the r.v. X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

(b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it turns out to be the only unbiased estimator for θ.)

(c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

$$|h(X) - \theta| \leq |g(X) - \theta|$$

with the inequality sometimes strict.

Solution

(a) The estimator is biased, with bias given by

$$\begin{aligned} E\left[e^{-3X}\right] - \theta =& \sum_{k=0}^{\infty} e^{-3k} \frac{\lambda^k}{k!} e^{-\lambda} - e^{-3\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(e^{-3}\lambda\right)^k}{k!} - e^{-3\lambda} \\ =& e^{-\lambda} e^{e^{-3}\lambda} - e^{-3\lambda} = e^{-3\lambda}\left(e^{\left(2+e^{-3}\right)\lambda} - 1\right) \neq 0. \end{aligned}$$

(b) The estimator $g(X) = (-2)^X$ is unbiased since

$$E\left[(-2)^X\right] - \theta = e^{-\lambda} \sum_{k=0}^{\infty} (-2)^k \frac{\lambda^k}{k!} - e^{-3\lambda} = e^{-\lambda} e^{-2\lambda} - e^{-3\lambda} = 0.$$

(c) The estimator $g(X)$ is silly in the sense that it is sometimes negative, whereas $e^{-3\lambda}$ is positive. One simple way to get a better estimator is to modify $g(X)$ to make it non-negative, by letting $h(X) = 0$ if $g(X) < 0$ and $h(X) = g(X)$ otherwise. Better yet, note that $e^{-3\lambda}$ is between 0 and 1 since $\lambda > 0$, so letting $h(X) = 0$ if $g(X) < 0$ and $h(X) = 1$ if $g(X) > 0$ is clearly more sensible than using $g(X)$.

Problem 4

Elk dwell in a certain forest. There are N elk, of which a simple random sample of size n is captured and tagged (so all $\binom{N}{n}$ sets of n elk are equally likely). The captured elk are returned to the population, and then a new sample is drawn. This is an important method that is widely used in ecology, known as capture-recapture. If the new sample is also a simple random sample, with some fixed size, then the number of tagged elk in the new sample is Hypergeometric.

For this problem, assume that instead of having a fixed sample size, elk are sampled one by one without replacement until m tagged elk have been recaptured, where m is specified in advance (of course, assume that $1 \leq m \leq n \leq N$). An advantage of this sampling method is that it can be used to avoid ending up with a very small number of tagged elk (maybe even zero), which would be problematic in many applications of capture-recapture. A disadvantage is not knowing how large the sample will be.

(a) Find the PMFs of the number of untagged elk in the new sample (call this X) and of the total number of elk in the new sample (call this Y).

(b) Find the expected sample size $\mathrm{E}Y$ using symmetry, linearity, and indicator r.v.s.

Hint: We can assume that even after getting m tagged elk, they continue to be captured until all N of them have been obtained; briefly explain why this can be assumed. Express $X = X_1 + \cdots + X_m$, where X_1 is the number of untagged elk before the first tagged elk, X_2 is the number between the first and second tagged elk, etc. Then find $\mathrm{E}X_j$ by creating the relevant indicator r.v. for each untagged elk in the population.

(c) Suppose that m, n, N are such that $\mathrm{E}Y$ is an integer. If the sampling is done with a fixed sample size equal to $\mathrm{E}Y$ rather than sampling until exactly m tagged elk are obtained, find the expected number of tagged elk in the sample. Is it less than m, equal to m, or greater than m (for $n < N$)?

Solution

(a) When $X = x$, it implies that there are x untagged elks and $m-1$ tagged elks in the first $x+m-1$ samples, and the $x+m$th sample is a tagged elk. Let A denotes event "there are x untagged elks and $m-1$ tagged elks in the first $x+m-1$ samples", B denotes event "the $x+m$th sample is a tagged elk", then for $x = 0, 1, ..., N-n$,

$$\begin{aligned} P(X=x) =& P(A,B) = P(A)P(B|A) \\ =& \frac{\binom{N-n}{x}\binom{n}{m-1}}{\binom{N}{x+m-1}} \cdot \frac{n-(m-1)}{N-(x+m-1)}. \end{aligned}$$

It follows that for $y = 0, 1, \ldots, N-n+m$,

$$P(Y=y) = P(X=y-m) = \frac{\binom{N-n}{y-m}\binom{n}{m-1}}{\binom{N}{y-1}} \cdot \frac{n-(m-1)}{N-(y-1)}.$$

(b) Suppose that we chose elks one at the time and we order them in the row until the all elks have been captured, not only until the $m$mth pre-tagged elk has been captured. We can consider our problem also in this way since the permutation after mth pre-tagged captured elk does not affect our probabilities. Now, consider following: put tagged elks (n of them) in the row and consider spots at the both ends of the row and between these elks. So, there exist $n+1$ empty spots. Now we have to arrange non-tagged elks on these spots. Because of the fact that all permutation are equally likely, there on average will be $\frac{N-n}{n+1}$ elks on each spot. So, using this idea, symmetry and linearity, we can write that

$$\begin{aligned} E(X) &= E(X_1) + \cdots + E(X_m) = mE(X_1) = m \cdot \frac{N-n}{n+1}, \\ E(Y) &= E(X) + m = m \cdot \frac{N+1}{n+1} \end{aligned}$$

(c) Let Z_i be the indicator of the event "the ith sample is a tagged elk". Then the number of tagged elk in the sample is $Z = Z_1 + \cdots + Z_{\mathrm{E}Y}$. By symmetry,

$$P(Z_i = 1) = \frac{n}{N},$$

thus

$$\mathrm{E}(Z) = \mathrm{E}(Z_1 + \cdots + Z_{\mathrm{E}Y}) = \mathrm{E}(Z_1) + \cdots + \mathrm{E}(Z_{\mathrm{E}Y}) = \frac{n}{N}\mathrm{E}Y.$$

For $n < N$, we have

$$\frac{n}{N}\mathrm{E}Y = \frac{n}{N} \cdot \frac{m\,(N+1)}{n+1} = \frac{nN+n}{nN+N} \cdot m < m.$$

Another solution

Let Z be the r.v. of the sample size. Then form the story we know that $Z \sim Geom(\frac{m(N+1)}{n+1}, n, N)$.

$$E(Z) = \frac{m(N+1)}{n+1}\frac{n}{N} \leq m.$$

Problem 5

The legendary Caltech physicist Richard Feynman and two editors of *The Feynman Lectures on Physics* (Michael Gottlieb and Ralph Leighton) posed the following problem about how to decide what to order at a restaurant. You plan to eat m meals at a certain restaurant, where you have never eaten before. Each time, you will order one dish (without replacement).

The restaurant has n dishes on the menu, with $n \geq m$. Assume that if you had tried all the dishes, you would have a definite ranking of them from 1 (your least favorite) to n (your favorite). If you knew which your favorite was, you would be happy to order it always (you never get tired of it).

Before you've eaten at the restaurant, this ranking is completely unknown to you. After you've tried some dishes, you can rank those dishes amongst themselves, but don't know how they compare with the dishes you haven't yet tried. There is thus an *exploration-exploitation trade-off*: should you try new dishes, or should you order your favorite among the dishes you have tried before?

A natural strategy is to have two phases in your series of visits to the restaurant: an exploration phase, where you try different dishes each time, and an exploitation phase, where you always order the best dish you obtained in the exploration phase. Let k be the length of the exploration phase (so $m - k$ is the length of the exploitation phase). Your goal is to maximize the expected sum of the ranks of the dishes you eat there (the rank of a dish is the "true" rank from 1 to n that you would give that dish if you could try all the dishes). Show that the optimal choice is

$$k = \sqrt{2(m+1)} - 1$$

or this rounded up or down to an integer if needed. Do this in the following steps:

(a) Let X be the rank of the best dish that you find in the exploration phase. Find the expected sum of the ranks of the dishes you eat, in terms of $E[X]$.

(b) Find the PMF of X, as a simple expression in terms of binomial coefficients.

(c) Show that

$$E[X] = \frac{k(n+1)}{k+1}.$$

(d) Use calculus to find the optimal value of k.

Solution

(a) Suppose that we pick k different meals out of n available in our exploration phase. Every picked meal has average rating $\frac{n+1}{2}$ (arithmetic mean), so, using the linearity of expectation we can obtain that the average sum of rankings in first k days is $k \cdot \frac{n+1}{2}$. Now it's starting our exploitation phase: in remaining $m - k$ days we always pick our best meal that we have chosen in first phase. Thus, the average sum of rankings of these meals is simply $(m-k) \cdot E(X)$. Finally, the total average sum of rankings is

$$k \cdot \frac{n+1}{2} + (m-k) \cdot E(X)$$

(b) When $X = u$ (the highest ranked dish we find in the exploration phase is u), then $P(X = u)$ is what we want, for $k \leq u \leq n$; *i.e.*, the probability that the highest ranking is u (note that the probability is zero when $1 \leq u < k$). Between 1 and n, we choose k integers. Note that

- The total number of selecting sequence is $k!\binom{n}{k}$.
- When u is fixed, then the number of all possible combinations is given by $k!\binom{u-1}{k-1}$.

As a result,

$$P(X=u)=\frac{k!\binom{u-1}{k-1}}{k!\binom{n}{k}}=\frac{\binom{u-1}{k-1}}{\binom{n}{k}},\ k\le u\le n.$$

(c) By the definition of expectation, we have

$$E[X]=\sum_{u=1}^{n}u\cdot P(X=u)=\sum_{u=k}^{n}u\cdot P(X=u)=\sum_{u=k}^{n}u\cdot\frac{\binom{u-1}{k-1}}{\binom{n}{k}}$$
$$\overset{(a)}{=}\sum_{u=k}^{n}k\frac{\binom{u}{k}}{\binom{n}{k}}=\frac{k}{\binom{n}{k}}\sum_{u=k}^{n}\binom{u}{k}\overset{(b)}{=}\frac{k}{\binom{n}{k}}\binom{n+1}{k+1}=\frac{k(n+1)}{k+1},$$

where we have the (a) by following Example 1.5.2, and the (b) by following the hockey stick identity.

(d) Plugging our result in (a), we obtain the final expression for expected sum of values. According to that, define the function f with

$$f(k)=k\cdot\frac{n+1}{2}+(m-k)\cdot\frac{k(n+1)}{k+1}.$$

Derive this function and set it equal to zero:

$$\frac{d}{dk}f(k)=\frac{n+1}{2}-\frac{k(n+1)}{2}+(m-k)\frac{(n+1)(k+1)-k(n+1)}{(k+1)^2}=0.$$

Use basic algebra to obtain that this equation is equal to

$$k^2+2k-(2m+1)=0.$$

Therefore, maximizing it gives rise to

$$k=\sqrt{2(m+1)}-1.$$

Problem 6

(a) What is the probability that four points selected uniformly at random on a circle lie on the same semicircle?

(b) What is the probability that $n \geq 2$ points selected uniformly at random on a circle lie on the same semicircle?

(c) Suppose $n \geq 2$ points selected uniformly at random on the surface of d-dimension ($d \geq 3$) unit sphere, what is the probability that all points lie on the same hemisphere?

Solution

Hint: Check the paper *"A Problem in Geometric Probability" by J.G. Wendel* for the original derivation and also the this link to the wiki page for Wendel's theorem.
In the following we only prove the most general case in (c).
Let $x_1, x_2, \ldots, x_n$ be d-dimensional random vectors such that the x_j is uniformly and independently distributed over the surface of the unit sphere. We denote $p_{d,n}$ as the probability that all x_j lie in on the same hemisphere, i. e. that for some vector y the inner products $y^\top x_j$ are all positive.
In the following, we focus on proving the recurrence relation

$$p_{d,n} = \frac{1}{2}\Big(p_{d,n-1} + p_{d-1,n-1}\Big),$$

together with the evident boundary conditions

$$p_{1,n} = 2^{-n+1}, p_{d,n} = 1 \text{ if } n \leq d,$$

this will give rise to the final answer that

$$p_{d,n} = 2^{-n+1} \sum_{k=0}^{d-1} \binom{n-1}{k}$$

It is sufficient to evaluate the conditional probability when the x_j are non-zero and lie on fixed lines through the origin. Suppose that y is perpendicular to none of these lines. Then the sequence $s_y = \{\text{sgn}(y^\top x_j)\}$ is a random point in the set $S = \{s\}$ of all ordered n-tuples consisting of plus and minus signs. A specific s is said to occur if there is a y such that $s_y = s$. Let A_s be the event that s occurs, and let $\mathbb{I}_s$ be the indicator of A_s. By definition,

$$p_{d,n} = \mathbb{P}(A_{s^*}), s^* = (+, +, \ldots, +).$$

Since any s can be changed into any other be appropriately reflecting x_j through the origin it follows that all A_s are equally likely. Therefore, we have

$$2^n p_{d,n} = \sum_{s^*} \mathbb{P}(A_{s^*}) = \mathbb{E}\left[\sum_{s^*} I_{s^*}\right] = \mathbb{E}[Q_{d,n}],$$

where $Q_{d,n}$ is the number of different s^* that occurs.
Let X_j be the $(d-1)$-dimensional hyperplane perpendicular to vector x_j. Then $Q_{d,n}$ is just the number of complementary components that are perpendicular to all the X_j in the d-dimensional ambient space at the same time. In order to count such complementary components, consider the effect of deleting one hyperplane, say X_n. There remain $n-1$ hyperplanes, with $Q_{d,n-1}$ complementary components. These components are of two kinds:

(i) those which intersect X_n and

(ii) those not intersecting X_n.

In an obvious notation we have $Q_{d,n-1} = Q^{(i)} + Q^{(ii)}$. When X_n is restored (i.e., adding back) it cuts each component of type (i) into two and does not disturb the others. Therefore $Q_{d,n} = 2Q^{(i)} + Q^{(ii)}$. It follows that

$$Q_{d,n} = Q_{d,n-1} + Q^{(i)}.$$

Note a claim now that

$$Q^{(i)} = Q_{d-1,n-1}.$$

In fact, the $\{X_j \cap X_n\}$ are "new" $(d-2)$-dimensional hyperplanes in the $(d-1)$-dimensional space, i.e., the space X_n (note that a hyperplane is also a space), and their normals are linearly independent. Therefore the set $X_n - \bigcup_{j=1}^{n-1}(X_j \cap X_n)$ has $Q_{d-1,n-1}$ components that are complementary to all $(d-2)$-dimensional hyperplanes $\{X_j \cap X_n\}$ (which are resided in X_n), and it is easy to see that these are just the intersections of the original type (i) components with X_n, establishing the claim. Substituting the above results and recalling that $Q_{d,n} = 2^N p_{d,n}$, we obtain the recursive equation mentioned above. This completes the proof.

Probability & Statistics for EECS: Homework #06

Due on Oct 29, 2023 at 23:59

Name:
Student ID:

Oct 22, 2023

Problem 1

The *Beta distribution* with parameters $a = 3$, $b = 2$ has PDF

$$f(x) = 12x^2(1-x), \text{ for} 0 < x < 1.$$

Let X have this distribution.

(a) Find the CDF of X.

(b) Find $P(0 < X < 1/2)$.

(c) Find the mean and variance of X (without quoting results about the Beta distribution).

Solution:

(a) The CDF of X is

$$\begin{aligned} F(X) &= \int_0^x f(t)dt = \int_0^x 12t^2(1-t)dt \\ &= \int_0^x 12t^2dt - \int_0^x 12t^3dt \\ &= 4t^3|_0^x - 3t^4|_0^x \\ &= x^3(4-3x), \quad \text{for } 0 < x < 1 \end{aligned}$$

(b) According to CDF $F(x)$, $P(0 < x < 1/2) = F(1/2) = \frac{5}{16}$.

(c) According to PDF, the mean of X is

$$\begin{aligned} E(X) &= \int_0^1 xf(x)dx = \int_0^1 12x^2(1-x)dx \\ &= \int_0^1 12x^3dx - \int_0^1 12x^4dx \\ &= \frac{3}{5} \end{aligned}$$

We have

$$\begin{aligned} E(X^2) &= \int_0^1 x^2f(x)dx = \int_0^1 12x^4(1-x)dx \\ &= \int_0^1 12x^4dx - \int_0^1 12x^5dx \\ &= \frac{2}{5} \end{aligned}$$

Thus, we have

$$Var(X) = E(X^2) - EX^2 = \frac{1}{25}$$

Problem 2

Let $U_1, \ldots, U_n$ be i.i.d. Unif$(0,1)$, and $X = \max(U_1, \ldots, U_n)$.

(a) What is the PDF of X?

(b) What is $E[X]$?

Solution:

(a) The CDF of X is

$$\begin{aligned} P(X \leq x) &= P(U_1 \leq x, \ldots, U_n \leq x) \\ &= P(U_1 \leq x) \cdots P(U_n \leq x) \\ &= x^n. \end{aligned}$$

Thus, the PDF of X is $f(x) = nx^{n-1}$ $(0 < x < 1)$.

(b) Then we have

$$\begin{aligned} E[X] &= \int_0^1 x f(x) dx = \int_x \cdot n x^{n-1} dx \\ &= n \int_0^1 x^n dx \\ &= \frac{n}{n+1} \end{aligned}$$

Problem 3

the *Laplace distribution* has PDF

$$f(x) = \frac{1}{2}e^{-|x|}$$

for all real x. The Laplace distribution is also called a *symmetrized Exponential* distribution. Explain this in the following two ways.

(a) Plot the PDFs and explain how they relate.

(b) Let $X \sim Expo(1)$ and S be a random sign (1 or -1, with equal probabilities), with S and X independent. Find the PDF of SX (by first finding the CDF), and compare the PDF of SX and the Laplace PDF.

solution:

(a) The figure is shown below.
When $x \geq 0$, $2X \sim Expo(1)$.
When $x \leq 0$, $-2X \sim Expo(1)$.

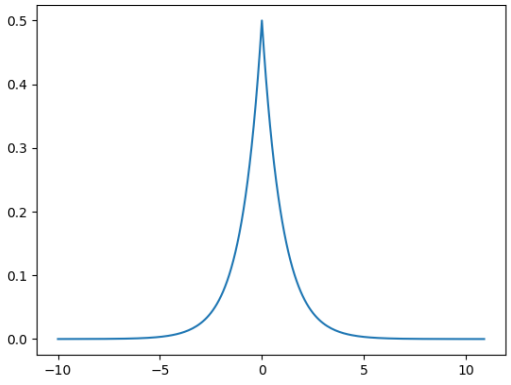


Figure 1: PDF of Laplace distribution

(b)

$$\begin{aligned} F(X) = &\ P(SX \leq x) \\ = &\ P(S=-1)P(SX \leq x|S=-1) + P(S=1)P(SX \leq x|S=1) \\ = &\ \frac{1}{2}\left(P(-X \leq x) + P(X \leq x)\right). \end{aligned}$$

When $x \geq 0$, $P(X \leq x) = 1 - e^{-x}$, $P(-X \leq x) = 1$,

$$F(x) = 1 - \frac{e^{-x}}{2},$$

$$f(x) = \frac{e^{-x}}{2}.$$

When $x < 0$, $P(X \leq x) = 0$, $P(-X \leq x) = P(X \neq -x) = 1 - P(X \leq x) = e^{x}$,

$$F(x) = \frac{e^{x}}{2},$$

$$f(x) = \frac{e^x}{2}.$$

Therefore, $f(x) = \frac{1}{2}e^{-|x|}$.

Problem 4

The *Gumbel distribution* is the distribution of $-logX$ with $X \sim \text{Expo}(1)$.

(a) Find the CDF of the Gumbel distribution.

(b) Let $X_1, X_2, \ldots$ be i.i.d. $Expo(1)$ and let $M_n = max(X_1, \ldots, X_n)$. Show that $M_n - \log n$ converges in distribution to the Gumbel distribution, i.e., as $n \to \infty$ the CDF of $Mn - \log n$ converges to the Gumbel CDF.

solution:

(a) Let G be Gumbel and $X \sim \text{Expo}(1)$. The CDF of G is

$$\begin{aligned} P(G \le t) &= P(-\log X \le t) \\ &= P(X \ge e^{-t}) \\ &= e^{-e^{-t}}. \end{aligned}$$

(b) CDF of $M_n - \log n$ can be written as

$$P(M_n - \log n \le t) = P(X_1 \le t + \log n, X_2 \le t + \log n, ..., X_n \le t + \log n) = P(X_1 \le t + \log n)^n.$$

According to (a) and $(1 + \frac{x}{n})^n \to e^x$, when $n \to \infty$. Thus,

$$\begin{aligned} CDF &= (1 - e^{-(t+\log n)})^n \\ &= (1 - \frac{e^{-t}}{n})^n \\ &= e^{-e^{-t}}. \end{aligned}$$

Problem 5

Let $Z \sim \mathcal{N}(0,1)$, and c be a nonnegative constant. Find $E(\max(Z-c,0))$, in terms of the standard Normal CDF Φ and PDF φ.

Let φ be the PDF of $\mathcal{N}(0,1)$, then we have

$$\begin{aligned} E(\max(Z-c,0)) &= \int_{-\infty}^{\infty} \max(z-c,0)\varphi(z)\,\mathrm{d}z \\ &= \int_{c}^{\infty} (z-c)\varphi(z)\,\mathrm{d}z \\ &= \int_{c}^{\infty} z\varphi(z)\,\mathrm{d}z - c\int_{c}^{\infty} \varphi(z)\,\mathrm{d}z \\ &= \frac{-1}{\sqrt{2\pi}} e^{-z^2/2}\Big|_{c}^{\infty} - c(1-\Phi(c)) \\ &= \frac{1}{\sqrt{2\pi}} e^{-c^2/2} - c(1-\Phi(c)) \\ &= \varphi(c) + c\Phi(c) - c \end{aligned} \tag{1}$$

Problem 6

Suppose $X \sim N(m, \sigma^2)$, where m is an integer and σ is a real number. Let $Y = \lfloor X \rfloor$ be the integer part of X.

1. Find the PMF of Y
2. Find $\mathbb{E}[Y]$
3. Find $\text{Var}[Y]$

Solution:

1. Recall the PDF and CDF of a Gaussian random variable $X \sim \mathcal{N}(m, \sigma^2)$ as follows:

$$f(x) = \phi\left(\frac{x-m}{\sigma}\right)\frac{1}{\sigma}, F(x) = \Phi\left(\frac{x-m}{\sigma}\right).$$

Therefore, we have

$$\begin{aligned} P(Y = y) &= P(\lfloor X \rfloor = y) \\ &= P(y \le X \le y+1) \\ &= F(y+1) - F(y) \\ &= \Phi\left(\frac{y+1-m}{\sigma}\right) - \Phi\left(\frac{y-m}{\sigma}\right). \end{aligned} \tag{2}$$

2. Recall that we can equivalently write

$$X = m + \sigma Z, Z \sim \mathcal{N}(0,1).$$

Therefore, we have

$$\mathbb{E}[Y] = \mathbb{E}[\lfloor m + \sigma Z \rfloor] = m + \mathbb{E}[\lfloor \sigma Z \rfloor],$$

where the last equality holds since m is an integer.

Substituting $W = \lfloor \sigma Z \rfloor$ into subproblem 1, we have

$$P(W = w) = \Phi\left(\frac{w+1}{\sigma}\right) - \Phi\left(\frac{w}{\sigma}\right).$$

By the definition of expectation, we have

$$\begin{aligned} \mathbb{E}[W] &= \sum_{w=-\infty}^{\infty} w\left(\Phi\left(\frac{w+1}{\sigma}\right) - \Phi\left(\frac{w}{\sigma}\right)\right) \\ &= \sum_{w=0}^{\infty} w\left(\Phi\left(\frac{w+1}{\sigma}\right) - \Phi\left(\frac{w}{\sigma}\right)\right) - (w+1)\left(\Phi\left(\frac{-w}{\sigma}\right) - \Phi\left(\frac{-w-1}{\sigma}\right)\right) \\ &= -\sum_{w=0}^{\infty} \Phi\left(\frac{-w}{\sigma}\right) - \Phi\left(\frac{-w-1}{\sigma}\right) \\ &= -\frac{1}{2} \end{aligned}$$

Therefore, we have

$$\mathbb{E}[Y] = m + \mathbb{E}[W] = m - \frac{1}{2}.$$

3. **(Approximation via Discrete Fourier Transform)** Recall that we define $Y = \lfloor X \rfloor$ and we further define Y^* as the fractional part of X. Therefore, we have

$$X = Y + Y^*.$$

Intuitively, since the knowledge of X is fully known, we may analyze the properties of Y^* first and then show the variance of Y. According to paper "The Density of the Fractional Part of a Normal Distribution", the density of Y^* can be written as

$$f_\sigma(y - m), 0 \le y \le 1,$$

where

$$f_\sigma(y) = \sum_{k=-\infty}^{\infty} \phi_\sigma(y + k), \phi_\sigma(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{y^2}{2\sigma^2}).$$

Since the density of Y^* is only on the $(0, 1)$ interval since it is the fractional part, we calculate the discrete Fourier transform to get the individual Fourier coefficients:

$$\begin{aligned} c_n &= \int_0^1 f_\sigma(y) \exp(2\pi i n y) dy = \int_0^1 \sum_{k=-\infty}^{\infty} \phi_\sigma(y + k) \exp(2\pi i n y) dy \\ &= \sum_{k=-\infty}^{\infty} \int_0^1 \phi_\sigma(y + k) \exp(2\pi i n y) dy = \sum_{k=-\infty}^{\infty} \int_k^{k+1} \phi_\sigma(y) \exp(2\pi i n y) dy \\ &= \int_{-\infty}^{\infty} \phi_\sigma(y) \exp(2\pi i n y) dy = \int_{-\infty}^{\infty} \phi_\sigma(y) \cos(2\pi n y) dy \\ &= \exp\left(-(2\pi n\sigma)^2/2\right). \end{aligned}$$

Since ϕ is an even function, we are able to replace the complex exponential with just a cosine in the penultimate step above, giving us a discrete cosine transform, instead of the Fourier transform. Adding up all the Fourier terms then gives the result

$$\begin{aligned} f_\sigma(y) &= \sum_{k=-\infty}^{\infty} c_k \cos(2\pi k y) = \sum_{k=-\infty}^{\infty} \exp\left(-(2\pi k\sigma)^2/2\right) \cos(2\pi k y) \\ &= 1 + 2\sum_{k=1}^{\infty} \exp\left(-2(\pi k\sigma)^2\right) \cos(2\pi k y). \end{aligned}$$

Therefore, the density of Y^* is defined as

$$\begin{aligned} &1 + 2\sum_{k=1}^{\infty} \exp\left(-2(\pi k\sigma)^2\right) \cos(2\pi k(y - m)) \\ &= 1 + 2\sum_{k=1}^{\infty} \exp\left(-2(\pi k\sigma)^2\right) \cos(2\pi k y), \end{aligned} \tag{3}$$

where the equality holds since m is an integer.

From the equation (3), we know that when $\sigma \ge 1$, the the density of Y^* is approximately to that of a uniform distribution over interval $(0, 1)$. Therefore, the variance of a rounded Gaussian random variable Y is approximately

$$\sigma^2 + \frac{1}{12}.$$

Check the following links for more information

- Paper: "The Density of the Fractional Part of a Normal Distribution".
- Does rounding introduce variance into estimates?
- Finding the mean and variance of a distribution.
- Statistics of a Gaussian random variable with the floor function transformation.
- Fractional part of normally distributed variable.

Probability & Statistics for EECS: Homework #7 Solutions

Professor Ziyu Shao

Problem 1

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x)=\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x)=\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x)=\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x)=\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

- X discrete, Y continuous:
  According to the continuous Bayes' rule, we have

$$P(Y\in(y-\varepsilon,y+\varepsilon)|X=x)=\frac{P(X=x|Y\in(y-\varepsilon,y+\varepsilon)P(Y\in(y-\varepsilon,y+\varepsilon))}{P(X=x)}.$$

  By letting $\varepsilon\to 0$, we have

$$\lim_{\varepsilon\to 0}\ P(Y\in(y-\varepsilon,y+\varepsilon)|X=x)=\lim_{\varepsilon\to 0}\ f_Y(y|X=x)\cdot 2\varepsilon,$$

  and

$$\lim_{\varepsilon\to 0}\ \frac{P(X=x|Y\in(y-\varepsilon,y+\varepsilon)P(Y\in(y-\varepsilon,y+\varepsilon))}{P(X=x)}=\lim_{\varepsilon\to 0}\ \frac{P(X=x|Y=y)f_Y(y)\cdot 2\varepsilon}{P(X=x)}.$$

  Therefore, we can finish the proof by canceling the term 2ε in the following equation:

$$\begin{aligned}\lim_{\varepsilon\to 0}\ f_Y(y|X=x)\cdot 2\varepsilon&=\lim_{\varepsilon\to 0}\ \frac{P(X=x|Y=y)f_Y(y)\cdot 2\varepsilon}{P(X=x)}\\ \Rightarrow f_Y(y|X=x)&=\frac{P(X=x|Y=y)f_Y(y)}{P(X=x)}.\end{aligned}$$

- X continuous, Y discrete:

$$\begin{aligned}P(Y=y|X=x)&=\lim_{\varepsilon\to 0}\ P(Y=y|X\in(x-\varepsilon,x+\varepsilon))\\&=\lim_{\varepsilon\to 0}\ \frac{P(X\in(x-\varepsilon,x+\varepsilon)|Y=y)P(Y=y)}{P(X\in(x-\varepsilon,x+\varepsilon))}\\&=\lim_{\varepsilon\to 0}\ \frac{2\varepsilon\cdot f_X(x|Y=y)P(Y=y)}{2\varepsilon\cdot f_X(x)}\\&=\frac{f_X(x|Y=y)P(Y=y)}{f_X(x)}\end{aligned}$$

	Y **discrete**	Y **continuous**
X **discrete**	$P(X=x)=\sum_y P(X=x\|Y=y)P(Y=y)$	$P(X=x)=\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$f_X(x)=\sum_y f_X(x\|Y=y)P(Y=y)$	$f_X(x)=\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

- X discrete, Y continuous:

$$P(X=x|Y\in(y-\varepsilon,y+\varepsilon))=\frac{P(Y\in(y-\varepsilon,y+\varepsilon)|X=x)P(X=x)}{P(Y\in(y-\varepsilon,y+\varepsilon))}.$$

By letting $\varepsilon\to 0$, we have

$$\lim_{\varepsilon\to 0}P(X=x|Y\in(y-\varepsilon,y+\varepsilon))=P(X=x|Y=y),$$

and

$$\begin{aligned}&\lim_{\varepsilon\to 0}\frac{P(Y\in(y-\varepsilon,y+\varepsilon)|X=x)P(X=x)}{P(Y\in(y-\varepsilon,y+\varepsilon))}\\ =&\lim_{\varepsilon\to 0}\frac{f_Y(y|X=x)\cdot 2\varepsilon\cdot P(X=x)}{f_Y(y)\cdot 2\varepsilon}\\ =&\frac{f_Y(y|X=x)P(X=x)}{f_Y(y)}.\end{aligned}$$

By combining the two equations, we can get

$$\begin{aligned}P(X=x|Y=y)&=\frac{f_Y(y|X=x)P(X=x)}{f_Y(y)}\\ \Rightarrow P(X=x|Y=y)f_Y(y)&=f_Y(y|X=x)P(X=x).\end{aligned}$$

By integrating on both sides of the equation with respect y, we can get

$$\begin{aligned}\int_{-\infty}^{\infty}P(X=x|Y=y)f_Y(y)dy&=\int_{-\infty}^{\infty}f_Y(y|X=x)P(X=x)dy\\ &=P(X=x)\int_{-\infty}^{\infty}f_Y(y|X=x)dy\\ &=P(X=x).\end{aligned}$$

- X continuous, Y discrete:

$$P(X\in(x-\varepsilon,x+\varepsilon))=\sum_y P(X\in(x-\varepsilon,x+\varepsilon)|Y=y)P(Y=y).$$

By letting $\varepsilon\to 0$, we have

$$\lim_{\varepsilon\to 0}P(X\in(x-\varepsilon,x+\varepsilon))=\lim_{\varepsilon\to 0}f_X(x)\cdot 2\varepsilon,$$

and

$$\lim_{\varepsilon\to 0}\sum_y P(X\in(x-\varepsilon,x+\varepsilon)|Y=y)P(Y=y)=\lim_{\varepsilon\to 0}\sum_y f_X(x|Y=y)\cdot 2\varepsilon\cdot P(Y=y).$$

By combining the two equations, we can get

$$\begin{aligned}\lim_{\varepsilon\to 0}f_X(x)\cdot 2\varepsilon&=\lim_{\varepsilon\to 0}\sum_y f_X(x|Y=y)\cdot 2\varepsilon\cdot P(Y=y)\\ f_X(x)&=\sum_y f_X(x|Y=y)P(Y=y).\end{aligned}$$

Problem 2

A chicken lays a Pois(λ) number N of eggs. Each egg hatches a chick with probability p, independently. Let X be the number which hatch, and Y be the number which do NOT hatch.

(a) Find the joint PMF of N, X, Y. Are they independent?

(b) Find the joint PMF of N, X. Are they independent?

(c) Find the joint PMF of X, Y. Are they independent?

(d) Find the correlation between N (the number of eggs) and X (the number of eggs which hatch). Simplify; your final answer should work out to a simple function of p (the λ should cancel out).

Solution

Using the chicken-egg story, we can obtain that X is distributed Pois($p\lambda$) and Y similarly Pois($q\lambda$) with $q = 1 - p$ and that these random variables are independent!

(a) For non-negative integer i, j, n, if $i + j \neq n, P(X = i, Y = j, N = n) = 0$.
If $i + j = n$, then

$$P(X = i, Y = j, N = n) = P(X = i, Y = j \mid N = n)P(N = n) = \binom{n}{i} p^i q^{n-i} \cdot \frac{\lambda^n}{n!} e^{-\lambda}.$$

X, Y, and N are not independent because we have that for $i, j, n > 0$ such that $i + j \neq n$

$$P(X = i, Y = j, N = n) = 0.$$

But obviously we have that

$$P(X = i)P(Y = j)P(N = n) > 0.$$

(b) For $n \geq i \geq 0$,

$$P(X = i, N = n) = P(X = i \mid N = n)P(N = n) = \binom{n}{i} p^i q^{n-i} \cdot \frac{\lambda^n}{n!} e^{-\lambda}, n \geq i \geq 0.$$

Otherwise, $P(X = i, N = n) = 0$.

X and N are not independent since from the story, $X \sim \text{Pois}(p\lambda)$, then we have

$$P(X = i)P(N = n) = \frac{(\lambda p)^i}{i!} e^{-\lambda p} \cdot \frac{\lambda^n}{n!} e^{-\lambda}$$

which is obviously not equal to the joint PMF. (We could also see this by observing that for $i > n$ we have that $P(X = i, N = n) = 0$.)

(c) As we know from the chicken-egg story, we have that X and Y are independent, so the joint distribution is

$$P(X = i, Y = j) = P(X = i)P(Y = j) = \frac{(\lambda p)^i}{i!} e^{-\lambda p} \frac{(\lambda q)^j}{j!} e^{-\lambda q}, i, j \geq 0.$$

(d) By the property of covariance,

$$\text{Cov}(N, X) = \text{Cov}(X + Y, X) = \text{Cov}(X, X) + \text{Cov}(X, Y) = \text{Var}(X) = \lambda p$$

Since $N \sim \text{Pois}(\lambda), \text{Var}(N) = \lambda$, we have

$$\text{Corr}(N, X) = \frac{\text{Cov}(N, X)}{\sqrt{\text{Var}(N)\text{Var}(X)}} = \frac{\lambda p}{\sqrt{\lambda \cdot \lambda p}} = \sqrt{p}.$$

Problem 3

Let X and Y be i.i.d. Expo(λ), and $T = X + Y$.

(a) Find the conditional CDF of T given $X = x$. Be sure to specify where it is zero.

(b) Find the conditional PDF $f_{T|X}(t \mid x)$, and verify that it is a valid PDF.

(c) Find the conditional PDF $f_{X|T}(x \mid t)$, and verify that it is a valid PDF.

Hint: This can be done using Bayes' rule without having to know the marginal PDF of T, by recognizing what the conditional PDF is up to a normalizing constant-then the normalizing constant must be whatever is needed to make the conditional PDF valid.

(d) In Example 8.2.4, we will show that the marginal PDF of T is $f_T(t) = \lambda^2 t e^{-\lambda t}$, for $t > 0$. Give a short alternative proof of this fact, based on the previous parts and Bayes' rule.

Solution

(a)

$$F_{T|X}(t|x) = P(T \le t|X = x) = P(X + Y \le t|X = x) = P(Y \le t - x) = \left(1 - e^{-\lambda(t-x)}\right) \cdot \chi_{t \ge x}.$$

P.S., view $\chi_{\{\cdot\}}$ as the indicator function $\mathbb{I}\{\cdot\}$.

(b) Take derivative from $F_{T|X}$ respective to t.

$$f_{T|X}(t|x) = \frac{\partial}{\partial t} F_{T|X}(t|x) = \frac{\partial}{\partial t}\left[\left(1 - e^{-\lambda(t-x)}\right) \cdot \chi_{t \ge x}\right] = \lambda e^{-\lambda(t-x)} \cdot \chi_{t \ge x}.$$

- Non-negativity:

$$f_{T|X}(t|x) = \lambda e^{-\lambda(t-x)} \cdot \chi_{t \ge x} = \begin{cases} 0 \ge 0, & t < x \\ \lambda e^{-\lambda(t-x)} \ge 0, & t \ge x \end{cases}$$

- Integrates to 1:

$$\int_{-\infty}^{\infty} f_{T|X}(t|x)dt = \int_{x}^{\infty} \lambda e^{-\lambda(t-x)}dt = -\, e^{-\lambda(t-x)}\Big|_{t=x}^{t=\infty} = 1$$

Therefore, $f_{T|X}(t|x)$ is valid PDF.

(c)

$$\begin{aligned} f_{X|T}(x|t) =& \frac{f_{T|X}(t|x) f_X(x)}{f_T(t)} \\ =& \frac{1}{f_T(t)} \lambda e^{-\lambda(t-x)} \cdot \lambda e^{-\lambda x} \cdot \chi_{t \ge x} \\ =& \frac{1}{f_T(t)} \lambda^2 e^{-\lambda t} \cdot \chi_{t \ge x} \end{aligned}$$

- Non-negativity:

$$f_{X|T}(x|t) = \frac{1}{f_T(t)} \lambda^2 e^{-\lambda t} \cdot \chi_{t \ge x} = \begin{cases} 0 \ge 0, & t < x \\ \dfrac{1}{f_T(t)} \lambda^2 e^{-\lambda t} \ge 0, & t \ge x \end{cases}$$

- Integrates to 1: Note that $f_{X|T}(x|t) = \frac{1}{f_T(t)} \lambda^2 e^{-\lambda t} \cdot \chi_{t \ge x}$ is constant with respect to x. In particular, $f_{X|T}(x|t)$ is a non-zero constant respect to x over support $(0, t)$ and zero otherwise. By definition of Uniform distribution, we have $X|T = t \sim \text{Unif}(0, t)$, hence a valid PDF $f_{X|T}(x|t)$ over support $(0, t)$.

(d) Recall in part (c) that $X|T = t \sim \text{Unif}(0, t)$, hence $f_{X|T}(x|t) = \frac{1}{t} \cdot \chi_{t \geq x}$. Therefore, we have

$$f_T(t) = \frac{f_{T|X}(t|x) f_X(x)}{f_{X|T}(x|t)} = \frac{\lambda e^{-\lambda(t-x)} \cdot \chi_{t \geq x} \cdot \lambda e^{-\lambda x}}{\frac{1}{t} \cdot \chi_{t \geq x}} = \lambda^2 t e^{-\lambda t}, t > 0.$$

Another "solution" to (c) and (d)

Using Bayes' rule we have that

$$f_{X|T}(x \mid t) = \frac{f(x,t)}{f_T(t)} = \frac{f_{T|X}(t|x) f_X(x)}{f_T(t)} = \alpha f_{T|X}(t \mid x) f_X(x) = \alpha \lambda e^{-\lambda(t-x)} \lambda e^{-\lambda x} \cdot \chi_{t \geq x} = \alpha \lambda^2 e^{-\lambda t} \cdot \chi_{t \geq x}$$

for some $\alpha > 0$. Observe that $f_{X|T}(x \mid t)$ is a constant function respective to x. In order to be a valid PDF, $f_{X|T}(x \mid t)$ has to satisfy following

$$1 = \int_{\mathbb{R}} f_{X|T}(x \mid t) dx = \int_0^t \alpha \lambda^2 e^{-\lambda t} dx = t \alpha \lambda^2 e^{-\lambda t}$$

So, for every $t > 0$ there has to be

$$\alpha = \frac{1}{t \lambda^2 e^{-\lambda t}}$$

and in this case it is a valid PDF.
Observe that in part (c) we have that in fact $f_T(t) = \frac{1}{\alpha}$. So, we can easily obtain that

$$f_T(t) = \lambda^2 t e^{-\lambda t}.$$

Problem 4

Let U_1, U_2, U_3 be i.i.d. Unif$(0,1)$, and let $L = \min(U_1, U_2, U_3)$, $M = \max(U_1, U_2, U_3)$.

(a) Find the marginal CDF and marginal PDF of M, and the joint CDF and joint PDF of L, M.
Hint: For the latter, start by considering $P(L \geq l, M \leq m)$.

(b) Find the conditional PDF of M given L.

Solution

(a) The event $M \leq m$ is the same as the event that all 3 of the U_j are at most m, so the CDF of M is $F_M(m) = m^3$ and the PDF is $f_M(m) = 3m^2$, for $0 \leq m \leq 1$. The event $L \geq l$, $M \leq m$ is the same as the event that all 3 of the U_j are between l and m (inclusive), so

$$P(L \geq l, M \leq m) = (m-l)^3$$

for $m \geq l$ with $m, l \in [0,1]$. By the axioms of probability, we have

$$P(M \leq m) = P(L \leq l, M \leq m) + P(L > l, M \leq m)$$

So the joint CDF is

$$P(L \leq l, M \leq m) = m^3 - (m-l)^3,$$

for $m \geq l$ with $m, l \in [0,1]$. The joint PDF is obtained by differentiating this with respect to l and then with respect to m (or vice versa):

$$f(l,m) = 6(m-l),$$

for $m \geq l$ with $m, l \in [0,1]$. As a check, note that getting the marginal PDF of M by finding $\int_0^m f(l,m)dl$ does recover the PDF of M (the limits of integration are from 0 to m since the min can't be more than the max).

(b) The marginal PDF of L is $f_L(l) = 3(1-l)^2$ for $0 \leq l \leq 1$ since $P(L > l) = P(U_1 > l, U_2 > l, U_3 > l) = (1-l)^3$ (alternatively, use the PDF of M together with the symmetry that $1 - U_j$ has the same distribution as U_j, or integrate out m in the joint PDF of L, M). So the conditional PDF of M given L is

$$f_{m|L}(m|l) = \frac{f(l,m)}{f_L(l)} = \frac{2(m-1)}{(1-l)^2},$$

for all $m, l \in [0,1]$ with $m \geq l$.

Problem 5

Let X, Y, Z be r.v.s such that $X \sim \mathcal{N}(0,1)$ and conditional on $X = x, Y$ and Z are i.i.d. $\mathcal{N}(x,1)$

(a) Find the joint PDF of X, Y, Z.

(b) By definition, Y and Z are conditionally independent given X. Discuss intuitively whether or not Y and Z are also unconditionally independent.

(c) Find the joint PDF of Y and Z. You can leave your answer as an integral, though the integral can be done with some algebra (such as completing the square) and facts about the Normal distribution.

Solution

(a) We given independence to obtain that

$$\begin{aligned} f_{X,Y,Z}(x,y,z) = f_{Y,Z|X}(y,z \mid x) f_X(x) &= f_{Y|X}(y \mid x) f_{Z|X}(z \mid x) f_X(x) \\ &= \varphi(y-x)\varphi(z-x)\varphi(x) \end{aligned}$$

where φ is the PDF of the standard normal distribution.

(b) Y and Z are not unconditionally independent since we don't even know their distribution in that case (their distributions depend on X). So it is pretty silly to discus about the independence.

(c) Using the part (a) and LOTP, we have that

$$f_{Y,Z}(y,z) = \int_{\mathbb{R}} f_{Y,Z|X}(y,z \mid x) f_X(x) dx = \int_{\mathbb{R}} \varphi(y-x)\varphi(z-x)\varphi(x) dx$$

Problem 6

This problem explores a visual interpretation of covariance. Data are collected for $n \geq 2$ individuals, where for each individual two variables are measured (e.g., height and weight). Assume independence across individuals (e.g., person l's variables gives no information about the other people), but not within individuals (e.g., a person's height and weight may be correlated).

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the n data points. The data are considered here as fixed, known numbers-they are the observed values after performing an experiment. Imagine plotting all the points (x_i, y_i) in the plane, and drawing the rectangle determined by each pair of points. For example, the points $(1, 3)$ and $(4, 6)$ determine the rectangle with vertices $(1, 3), (1, 6), (4, 6), (4, 3)$.

The signed area contributed by (x_i, y_i) and (x_j, y_j) is the area of the rectangle they determine if the slope of the line between them is positive, and is the negative of the area of the rectangle they determine if the slope of the line between them is negative. (Define the signed area to be 0 if $x_i = x_j$ or $y_i = y_j$, since then the rectangle is degenerate.) So the signed area is positive if a higher x value goes with a higher y value for the pair of points, and negative otherwise. Assume that the x_i are all distinct and the y_i are all distinct.

(a) The sample covariance of the data is defined to be

$$r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

are the sample means. (There are differing conventions about whether to divide by $n-1$ or n in the definition of sample covariance, but that need not concern us for this problem.)

Let (X, Y) be one of the (x_i, y_i) pairs, chosen uniformly at random. Determine precisely how $\text{Cov}(X, Y)$ is related to the sample covariance.

(b) Let (X, Y) be as in (a), and $(\tilde{X}, \tilde{Y})$ be an independent draw from the same distribution. That is, (X, Y) and $(\tilde{X}, \tilde{Y})$ are randomly chosen from the n points, independently (so it is possible for the same point to be chosen twice).

Express the total signed area of the rectangles as a constant times $E((X - \tilde{X})(Y - \tilde{Y}))$. Then show that the sample covariance of the data is a constant times the total signed area of the rectangles.

Hint: Consider $E((X - \tilde{X})(Y - \tilde{Y}))$ in two ways: as the average signed area of the random rectangle formed by (X, Y) and $(\tilde{X}, \tilde{Y})$, and using properties of expectation to relate it to $\text{Cov}(X, Y)$. For the former, consider the n^2 possibilities for which point (X, Y) is and which point $(\tilde{X}, \tilde{Y})$; note that n such choices result in degenerate rectangles.

(c) Based on the interpretation from (b), give intuitive explanations of why for any r.v.s W_1, W_2, W_3 and constants a_1, a_2, covariance has the following properties:

(i) $\text{Cov}(W_1, W_2) = \text{Cov}(W_2, W_1)$;

(ii) $\text{Cov}(a_1 W_1, a_2 W_2) = a_1 a_2 \, \text{Cov}(W_1, W_2)$;

(iii) $\text{Cov}(W_1 + a_1, W_2 + a_2) = \text{Cov}(W_1, W_2)$;

(iv) $\text{Cov}(W_1, W_2 + W_3) = \text{Cov}(W_1, W_2) + \text{Cov}(W_1, W_3)$.

Solution

(a) Since (X, Y) is chosen uniformly at random, we have

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i P(X = x_i) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}; \qquad \mathrm{E}(Y) = \sum_{i=1}^{n} y_i P(Y = y_i) = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}.$$

By definition, we know

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] \\ &= \sum_{i=1}^{n}(x_i - \mathrm{E}(X))(y_i - \mathrm{E}(Y))P(X = x_i, Y = y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = r.\end{aligned}$$

Thus we prove that $\mathrm{Cov}(X, Y)$ equals the sample covariance r.

(b)
- Denote the total signed area of the rectangles as S, then

$$S = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j).$$

Since (X, Y) and $(\tilde{X}, \tilde{Y})$ are independent, we have

$$\begin{aligned}\mathrm{E}((X - \tilde{X})(Y - \tilde{Y})) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j)P(X = x_i, Y = y_i)P(\tilde{X} = x_j, \tilde{Y} = y_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j) = \frac{S}{n^2}.\end{aligned}$$

Thusly we have $S = n^2\mathrm{E}((X - \tilde{X})(Y - \tilde{Y}))$.

- By the properties of expectation and considering that (X, Y) and $(\tilde{X}, \tilde{Y})$ are identically and independently sampled, we have

$$\begin{aligned}\mathrm{E}((X - \tilde{X})(Y - \tilde{Y})) &= \mathrm{E}(XY) - \mathrm{E}(\tilde{X}Y) - \mathrm{E}(X\tilde{Y}) + \mathrm{E}(\tilde{X}\tilde{Y}) \\ &= \mathrm{E}(XY) - \mathrm{E}(\tilde{X})\mathrm{E}(Y) - \mathrm{E}(X)\mathrm{E}(\tilde{Y}) + \mathrm{E}(\tilde{X}\tilde{Y}) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(\tilde{X})\mathrm{E}(\tilde{Y}) + \mathrm{E}(\tilde{X}\tilde{Y}) \\ &= 2[\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)] \\ &= 2\,\mathrm{Cov}(X, Y) = 2r.\end{aligned}$$

Thusly we have $r = \frac{S}{2n^2}$.

(c)
- The claim (i) is true because it doesn't matter what is the base and what is the height of the rectangle, we can switch them.
- The claim (ii) is true because rescaling the one coordinate by the factor c yields that the total area of the rectangle rescales for c.
- The claim (iii) is true since the area of the rectangle is invariant on linear translation.
- The claim (iv) is true because the distributive property of the area: it doesn't matter if we calculate two areas with the same base and then sum them or first we add heights and then calculate the total area.

Problem 7

We use the notation $X \perp\!\!\!\perp Y \mid Z$ to represent the statement: random variables X and Y are conditionally independent given random variable Z. Now given any four continuous random variables X, Y, Z, W, show the following properties of conditional independence:

1. Symmetry:
$$X \perp\!\!\!\perp Y \mid Z \iff Y \perp\!\!\!\perp X \mid Z.$$

2. Decomposition:
$$X \perp\!\!\!\perp (Y, W) \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z.$$

3. Weak Union:
$$X \perp\!\!\!\perp (Y, W) \mid Z \Rightarrow X \perp\!\!\!\perp (Y, W) \mid (Z, W).$$

4. Contraction:
$$X \perp\!\!\!\perp Y \mid Z \ \& \ X \perp\!\!\!\perp W \mid (Y, Z) \iff X \perp\!\!\!\perp (Y, W) \mid Z.$$

5. Intersection: For any positive joint PDF of X, Y, Z, W,
$$X \perp\!\!\!\perp Y \mid (Z, W) \ \& \ X \perp\!\!\!\perp Z \mid (Y, W) \iff X \perp\!\!\!\perp (Y, Z) \mid W.$$

In fact, these properties are found by Judea Pearl, who won 2011 Turing Award for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning. As Judea Pearl commented: "Exploiting conditional independence to generate fast probabilistic computations is one of the main contributions CS has made to probability theory."

Solution

1. According to the problem, proving one side is enough. By definition of conditional densities, given $X \perp\!\!\!\perp Y \mid Z$, we have
$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z &\Leftrightarrow f_{XY|Z}(x, y, z) = f_{X|Z}(x, z) f_{Y|Z}(y, z). \\ &\Leftrightarrow f_{XYZ}(x, y, z) f_Z(z) = f_{XZ}(x, z) f_{YZ}(y, z). \\ &\Leftrightarrow \exists f, a, b : f(x, y, z) = a(x, z) b(y, z). \end{aligned}$$
This immediately proves the symmetry:
$$\begin{aligned} & X \perp\!\!\!\perp Y \mid Z \\ \Leftrightarrow \ & f(x, y, z) = a(x, z) b(y, z). \\ \Leftrightarrow \ & \exists a^*, b^* : f(x, y, z) = a^*(x, z) b^*(y, z). \\ \Leftrightarrow \ & Y \perp\!\!\!\perp X \mid Z \end{aligned}$$

2. Given $X \perp\!\!\!\perp (Y, W) \mid Z$, we have $f(x, y, z, w) = a(x, z) b(y, z, w)$. By definition, we have
$$\begin{aligned} f_{XYZ}(x, y, z) = & \int_w f_{XYZW}(x, y, z, w) = \int_w f(x, y, z, w) \\ = \ & a(x, z) \int_w b(y, z, w) = a^*(x, z) b^*(y, z), \end{aligned}$$
which shows $X \perp\!\!\!\perp Y \mid Z$.

3. Given $X \perp\!\!\!\perp (Y, W) \mid Z$, we have $f(x, y, z, w) = a(x, z)b(y, z, w)$. Therefore, we have

$$\begin{aligned} f(x, y, z, w) &= a(x, z)b(y, z, w) \\ &= a^*(x, z, w)b^*(y, z, w), \end{aligned}$$

where the last equality holds by defining $a^*(x, z, w) \propto a(x, z), \forall w$. Therefore, this shows $X \perp\!\!\!\perp (Y, W) \mid (Z, W)$.

4. • $\Rightarrow$ Given $X \perp\!\!\!\perp Y \mid Z$, we have $f_{XY|Z}(x, y, z) = f_{X|Z}(x, z)f_{Y|Z}(y, z)$; $X \perp\!\!\!\perp W \mid (Y, Z)$ means $f_{XW|YZ}(x, y, z, w) = f_{X|YZ}(x, y, z)f_{W|YZ}(y, z, w)$. Therefore, by definition, we have

$$\begin{aligned} f_{XYW|Z}(x, y, z, w) &= f_{XW|YZ}(x, y, z, w)f_{Y|Z}(y, z) \\ &= f_{X|YZ}(x, y, z)f_{W|YZ}(y, z, w)f_{Y|Z}(y, z) \\ &\overset{X \perp\!\!\!\perp Y|Z}{=} f_{X|Z}(x, z)f_{YW|Z}(y, z, w), \end{aligned}$$

which shows $X \perp\!\!\!\perp (Y, W) \mid Z$.

• $\Leftarrow$ Given $X \perp\!\!\!\perp (Y, W) \mid Z$, we have

$$X \perp\!\!\!\perp (Y, W) \mid Z \overset{\text{Decomposition}}{\longrightarrow} X \perp\!\!\!\perp Y \mid Z$$
$$X \perp\!\!\!\perp (Y, W) \mid Z \overset{\text{WeakUnion}}{\longrightarrow} X \perp\!\!\!\perp (Y, W) \mid (Y, Z) \overset{\text{Decomposition}}{\longrightarrow} X \perp\!\!\!\perp W \mid (Y, Z)$$

5. • $\Rightarrow$ Given $X \perp\!\!\!\perp Y \mid (Z, W)$, we have $f(x, y, z, w) = a(x, z, w)b(y, z, w)$. Similarly, $X \perp\!\!\!\perp Z \mid (Y, W)$ means $f(x, y, z, w) = g(x, y, w)h(y, z, w)$. If $f(x, y, z, w) > 0$ for all (x, y, z, w), it follows that

$$g(x, y, w) = \frac{a(x, z, w)b(y, z, w)}{h(y, z, w)}.$$

Since the left-hand side does not depend on z, So for fixed $z = z_0$, we have

$$g(x, y, w) = \tilde{a}(x, w)\tilde{b}(y, w).$$

Insert this into the second expression for f to get

$$f(x, y, z, w) = \tilde{a}(x, w)\tilde{b}(y, w)h(y, z, w) = a^*(x, w)b^*(y, z, w),$$

which shows $X \perp\!\!\!\perp (Y, Z) \mid W$.

• $\Leftarrow$ Given $X \perp\!\!\!\perp (Y, Z) \mid W$, we have

$$X \perp\!\!\!\perp (Y, Z) \mid W \overset{\text{WeakUnion}}{\longrightarrow} X \perp\!\!\!\perp (Y, Z) \mid (Z, W) \overset{\text{Contraction}}{\longrightarrow} X \perp\!\!\!\perp Y \mid (Z, W)$$
$$X \perp\!\!\!\perp (Y, Z) \mid W \overset{\text{WeakUnion}}{\longrightarrow} X \perp\!\!\!\perp (Y, Z) \mid (Y, W) \overset{\text{Contraction}}{\longrightarrow} X \perp\!\!\!\perp Z \mid (Y, W)$$

Probability & Statistics for EECS: Homework #08

Due on Dec 2, 2023 at 23:59

Name:
Student ID:

Problem 1

Let X and Y be two continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} cx^2y, & \text{if } 0 \le y \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the value of constant c.

(b) Find the conditional probability $P(Y \le X/4 \mid Y \le X/2)$.

Solution:

(a) According to the statement, we have

$$\begin{aligned} 1 =& \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ =& \int_0^1\int_0^x cx^2y\,\mathrm{d}y\,\mathrm{d}x \\ =& \int_0^1 \frac{c}{2}x^4\,\mathrm{d}x \\ =& \frac{c}{10} \end{aligned} \tag{1}$$

So that, $c = 10$.

(b)

$$\begin{aligned} P\left(Y \le \frac{X}{4} \mid Y \le \frac{X}{2}\right) =& \frac{P(Y \le \frac{X}{4}, Y \le \frac{X}{2})}{P(Y \le \frac{X}{2})} \\ =& \frac{P(Y \le \frac{X}{4})}{P(Y \le \frac{X}{2})} \\ =& \frac{\int_0^1\int_0^{\frac{x}{4}} 10x^2y\,\mathrm{d}y\,\mathrm{d}x}{\int_0^1\int_0^{\frac{x}{2}} 10x^2y\,\mathrm{d}y\,\mathrm{d}x} \\ =& \frac{\int_0^1 \frac{x^4}{32}\,\mathrm{d}y\,\mathrm{d}x}{\int_0^1 \frac{x^4}{8}\,\mathrm{d}y\,\mathrm{d}x} \\ =& \frac{1}{4}. \end{aligned} \tag{2}$$

Problem 2

Let X and Y be two integer random variables with joint PMF

$$P_{X,Y}(x,y) = \begin{cases} \frac{1}{6\cdot 2^{\min(x,y)}}, & \text{if } x,y \geq 0, |x-y| \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the marginal distributions of X and Y.

(b) Are X and Y independent?

(c) Find $P(X = Y)$.

Solution:

(a) The marginal distributions of X is

$$P_X(X) = \sum_{y=0}^{\infty} P_{X,Y}(X,Y).$$

When $X = 0$, we have

$$P(X=0) = P(X=0,Y=0) + P(X=0,Y=1) = \frac{1}{3}.$$

When $X \neq 0$, we have

$$P(X=x) = P(X=x,Y=x-1) + P(X=x,Y=x) + P(X=x,Y=x+1) = \frac{1}{6\cdot 2^{x-2}}.$$

Thus, the marginal distribution of X is

$$P_X(X) = \begin{cases} \frac{1}{3}, & x = 0 \\ \frac{1}{6\cdot 2^{x-2}}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

According to the symmetric, the marginal distribution of Y is

$$P_Y(Y) = \begin{cases} \frac{1}{3}, & y = 0 \\ \frac{1}{6\cdot 2^{y-2}}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(b) Since that

$$P_{X,Y}(0,0) = \frac{1}{6}, \tag{3}$$

and

$$P(X=0)P(Y=0) = \frac{1}{9}, \tag{4}$$

X and Y are not independent.

(c) According to symmetric, we have $P(X=Y) = P(X=Y-1) = P(X=Y+1)$ and $P(X=Y)+P(X=Y-1)+P(X=Y+1) = 1$. Thus, we have

$$P(X=Y) = \frac{1}{3}.$$

Problem 3

Let X and Y be i.i.d. $\mathcal{N}(0,1)$, and let S be a random sign (1 or -1, with equal probabilities) independent of (X,Y).

(a) Determine whether or not $(X, Y, X+Y)$ is Multivariate Normal.

(b) Determine whether or not $(X, Y, SX+SY)$ is Multivariate Normal.

(c) Determine whether or not (SX, SY) is Multivariate Normal.

Solution:

(a) Yes, $(X, Y, X+Y)$ is Multivariate Normal, because for any $a, b, c \in R$,

$$aX + bY + c(X+Y) = (a+c)X + (b+c)Y,$$

and any linear combination of independent normally distributed variables are Normal.

(b) Denote $Z = X + Y + SX + SY = (1+S)X + (1+S)Y$.
$Z = 0$ is in fact $S = -1$, hence, we have that

$$P(Z=0) = P(S=-1) = \frac{1}{2}.$$

Hence, Z is not normally distributed.

(c) Observe that random vector (X,Y) is identically distributed as (-X,-Y). So,

$$\begin{aligned} P(SX+SY \le k) &= P(SX+SY \le k, S=1) + P(SX+SY \le k, S=-1) \\ &= P(SX+SY \le k | S=1)P(S=1) + P(SX+SY \le k | S=-1)P(S=-1) \\ &= \frac{1}{2}P(X+Y \le k) + \frac{1}{2}P(X+Y \ge -k) \\ &= \frac{1}{2}P(X+Y \le k) + \frac{1}{2}P(X+Y \le k) \\ &= P(X+Y \le k). \end{aligned}$$

So, (SX, SY) is equally distributed as (X, Y), and (X, Y) is Bivariate normal. Hence, (SX, SY) is Multivariate Normal.

Problem 4

Let Z_1, Z_2 be two i.i.d. random variables satisfying standard normal distributions, i.e., $Z_1, Z_2 \sim \mathcal{N}(0.1)$. Define

$$X = \mathbf{\Sigma}_X Z_1 + \mu_X;$$
$$Y = \mathbf{\Sigma}_Y(\rho Z_1 + \sqrt{1-\rho^2} Z_2) + \mu_Y,$$

where $\mathbf{\Sigma}_X > 0$, $\mathbf{\Sigma}_Y > 0$, $-1 < \rho < 1$.

(a) Show that X and Y are bivariate normal.

(b) Find the correlation coefficient between X and Y, i.e., $\text{Corr}(X, Y)$.

(c) Find the joint PDF of X and Y.

Solution:

(a) For $a, b \in \mathbb{R}$, we have

$$aX + bY = (a\mathbf{\Sigma}_X + b\mathbf{\Sigma}_Y \rho) Z_1 + b\sqrt{1-\rho^2}\mathbf{\Sigma}_Y Z_2 + a\mu_X + b\mu_Y.$$

Since the linear combination of two Normal distribution follows Normal distribution, X and Y are bivariate normal.

(b) Since $Z_1, Z_2 \sim \mathcal{N}(0,1)$. We have $\rho Z_1 + \sqrt{1-\rho^2} Z_2 \sim \mathcal{N}(0,1)$. So $X \sim \mathcal{N}(\mu_X, \mathbf{\Sigma}_X)$, $Y \sim \mathcal{N}(\mu_Y, \mathbf{\Sigma}_Y)$. Thus, we have

$$\begin{aligned} Cov(X,Y) &= Cov(\mathbf{\Sigma}_X Z_1 + \mu_X, \mathbf{\Sigma}_Y(\rho Z_1 + \sqrt{1-\rho^2} Z_2) + \mu_Y) \\ &= \mathbf{\Sigma}_X \mathbf{\Sigma}_Y Cov(Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= \mathbf{\Sigma}_X \mathbf{\Sigma}_Y \left(\rho Var(Z_1) + \sqrt{1-\rho^2} Cov(Z_1, Z_2)\right) \\ &= \mathbf{\Sigma}_X \mathbf{\Sigma}_Y \rho. \end{aligned}$$

Then correlation coefficient between X and y is

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\mathbf{\Sigma}_X \mathbf{\Sigma}_Y \rho}{\mathbf{\Sigma}_X \mathbf{\Sigma}_Y}.$$

(c) Since Z_1 and Z_2 are i.i.d., we have

$$f_{Z_1,Z_2}(z_1, z_2) = f_{Z_1}(z_1) f_{Z_2}(z_2) = \frac{1}{2\pi} e^{-\frac{z_1^2 + z_2^2}{2}}.$$

Since $X = \mathbf{\Sigma}_X Z_1 + \mu_X$, $Y = \mathbf{\Sigma}_Y(\rho Z_1 + \sqrt{1-\rho^2} Z_2) + \mu_Y$, we have

$$Z_1 = \frac{X - \mu_X}{\mathbf{\Sigma}_X}$$

and

$$Z_2 = \frac{Y - \mu_Y}{\sqrt{1-\rho^2}\mathbf{\Sigma}_Y} - \rho \frac{X - \mu_X}{\sqrt{1-\rho^2}\mathbf{\Sigma}_X}.$$

Problem 4 continued on next page...

Thus,

$$
\begin{aligned}
f_{X,Y}(x,y) &= \left|\frac{\partial(Z_1,Z_2)}{\partial(X,Y)}\right| f_{Z_1,Z_2}(z_1,z_2) \\
&= \frac{1}{\begin{vmatrix} \frac{\partial z_1}{\partial x} & \frac{\partial z_1}{\partial y} \\ \frac{\partial z_2}{\partial x} & r\frac{\partial z_2}{\partial x} \end{vmatrix}} f_{Z_1,Z_2}(z_1,z_2) \\
&= \frac{1}{\begin{vmatrix} \frac{1}{\mathbf{\Sigma}_X} & 0 \\ \frac{-\rho}{\sqrt{1-\rho^2}\mathbf{\Sigma}_X} & \frac{1}{\sqrt{1-\rho^2}\mathbf{\Sigma}_Y} \end{vmatrix}} f_{Z_1,Z_2}(z_1,z_2) \\
&= \frac{1}{\mathbf{\Sigma}_X\mathbf{\Sigma}_Y\sqrt{1-\rho^2}} f_{Z_1,Z_2}(z_1,z_2) \\
&= \frac{1}{\mathbf{\Sigma}_X\mathbf{\Sigma}_Y\sqrt{1-\rho^2}} f_{Z_1,Z_2}(\frac{x-\mu_X}{\mathbf{\Sigma}_X}, \frac{y-\mu_Y}{\sqrt{1-\rho^2}\mathbf{\Sigma}_Y} - \rho\frac{x-\mu_X}{\sqrt{1-\rho^2}\mathbf{\Sigma}_X}) \\
&= \frac{1}{2\pi\mathbf{\Sigma}_X\mathbf{\Sigma}_Y\sqrt{1-\rho^2}} e^{-\frac{(\frac{x-\mu_X}{\mathbf{\Sigma}_X})^2 + (\frac{y-\mu_Y}{\sqrt{1-\rho^2}\mathbf{\Sigma}_Y} - \rho\frac{x-\mu_X}{\sqrt{1-\rho^2}\mathbf{\Sigma}_X})^2}{2}} \\
&= \frac{1}{2\pi\mathbf{\Sigma}_X\mathbf{\Sigma}_Y\sqrt{1-\rho^2}} e^{-\frac{(\frac{x-\mu_X}{\mathbf{\Sigma}_X})^2 - \frac{2\rho(x-\mathbf{\Sigma}_X)(Y-\mathbf{\Sigma}_Y)}{\mathbf{\Sigma}_X\mathbf{\Sigma}_Y} + (\frac{y-\mu_Y}{\mathbf{\Sigma}_Y})^2}{2(1-\rho^2)}}.
\end{aligned}
$$

Problem 5

Given a random vector $\mathbf{X} = (X_1, \ldots, X_n)^\top$, which satisfies a multivariate normal (Gaussian) distribution, i.e., $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. When $\boldsymbol{\Sigma}$ is positive definite, the probability density function is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{5}$$

Now we divide random vector $\mathbf{X}(\mathbf{x})$ into two parts:

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}_A \\ \mathbf{X}_B\end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{bmatrix} \tag{6}$$

and split $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly:

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B\end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB}\end{bmatrix} \tag{7}$$

Show the following results:

1. Marginal distribution of $\mathbf{X}_A$ and $\mathbf{X}_B$ are still normal, i.e $\mathbf{X}_A \sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA})$.
2. $\boldsymbol{\Sigma}_{AB} = \mathbf{0}$ if and only if $\mathbf{X}_A$ and $\mathbf{X}_B$ are independent.
3. Given $\mathbf{X}_B$, the conditional distribution of $\mathbf{X}_A$ is still normal, i.e.

$$\mathbf{X}_A|\mathbf{X}_B \sim \mathcal{N}(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B}), \tag{8}$$

   where

$$\begin{aligned}\boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{X}_B - \boldsymbol{\mu}_B) \\ \boldsymbol{\Sigma}_{A|B} &= \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}.\end{aligned} \tag{9}$$

4. If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{X}' \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ are independent, then

$$\mathbf{X} + \mathbf{X}' \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\mu}', \boldsymbol{\Sigma} + \boldsymbol{\Sigma}'). \tag{10}$$

Solution

1. Suppose that the dimension of $\mathbf{X}_A$ is m, we then define matrix $\mathbf{A} = [\mathbf{I}_m, 0] \in \mathbb{R}^{m\times n}$ with identity matrix $\mathbf{I}_m$. Accordingly, we have $\mathbf{X}_A = \mathbf{A}\mathbf{X}$, i.e., $\mathbf{X}_A$ is a linear transformation of MVN $\mathbf{X}$, thus it is also a MVN itself. Besides, we have $\mathbf{X} = \mathbf{A}\mathbf{X} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top) = \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA})$. Similar reasoning also applies to $\mathbf{X}_B$.
2. • $\Rightarrow$ When $\boldsymbol{\Sigma}_{AB} = \boldsymbol{\Sigma}_{BA} = 0$, we have

$$\det(\boldsymbol{\Sigma}) = \det\left(\begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB}\end{bmatrix}\right) = \det(\boldsymbol{\Sigma}_{\boldsymbol{AA}})\det(\boldsymbol{\Sigma}_{BB}), \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix}\boldsymbol{\Sigma}_{AA}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{BB}^{-1}\end{bmatrix}.$$

   Therefore, we further have

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &= \frac{1}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma}_{AA})}} \frac{1}{\sqrt{(2\pi)^{n-m}\det(\boldsymbol{\Sigma}_{BB})}} \exp\left(-\frac{1}{2}([\mathbf{x}_A-\boldsymbol{\mu}_A, \mathbf{x}_B-\boldsymbol{\mu}_B])^\top\boldsymbol{\Sigma}^{-1}([\mathbf{x}_A-\boldsymbol{\mu}_A, \mathbf{x}_B-\boldsymbol{\mu}_B])\right) \\ &= \frac{1}{\sqrt{(2\pi)^m \det\boldsymbol{\Sigma}_{AA}}} \exp\left(-\frac{1}{2}(\mathbf{x}_A-\boldsymbol{\mu}_A)^\top\boldsymbol{\Sigma}_{AA}^{-1}(\mathbf{x}_A-\boldsymbol{\mu}_A)\right) \\ &\quad \cdot \frac{1}{\sqrt{(2\pi)^{n-m}\det(\boldsymbol{\Sigma}_{BB})}} \exp\left(-\frac{1}{2}(\mathbf{x}_B-\boldsymbol{\mu}_B)^\top\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{x}_B-\boldsymbol{\mu}_B)\right) \\ &= f_{\mathbf{X}_A}(\mathbf{x}_A) f_{\mathbf{X}_B}(\mathbf{x}_B). \end{aligned} \tag{11}$$

Problem 5 continued on next page...

- $\Leftarrow$ Since $\mathbf{X}_A$ and $\mathbf{X}_B$ are independent, we have

$$\begin{aligned}\boldsymbol{\Sigma}_{i,j} &= [\boldsymbol{\Sigma}_{AB}]_{ij} \\ &= [\mathrm{Cov}(\mathbf{X}_A, \mathbf{X}_B)]_{ij} \\ &= E[([\mathbf{X}_A]_i - [\boldsymbol{\mu}_A]_i)([\mathbf{X}_B]_j - [\boldsymbol{\mu}_B]_j)] \\ &= E[([\mathbf{X}_A]_i - [\boldsymbol{\mu}_A]_i)]E[([\mathbf{X}_B]_j - [\boldsymbol{\mu}_B]_j)] \\ &= \mathbf{0}\end{aligned} \tag{12}$$

3. By Schur complement, We first have

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{M} & -\mathbf{M}\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1} \\ -\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}\mathbf{M} & \mathbf{N} \end{bmatrix} \tag{13}$$

where $\mathbf{M} = (\boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA})^{-1}$ and $\mathbf{N} = \boldsymbol{\Sigma}_{BB}^{-1} + \boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}M\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}$. We then have

$$\begin{aligned}&f_{\mathbf{X}_A|\mathbf{X}_B}(\mathbf{X}_A = \mathbf{x}_A|\mathbf{X}_B = \mathbf{x}_B) \\ &= \frac{f_{\mathbf{X}_A,\mathbf{X}_B}(\mathbf{X}_A = \mathbf{x}_A, \mathbf{X}_B = \mathbf{x}_B)}{f_{\mathbf{X}_B}(\mathbf{X}_B = \mathbf{x}_B)} = \frac{f_{\mathbf{X}}(\mathbf{X} = \mathbf{x})}{f_{\mathbf{X}_B}(\mathbf{X}_B = \mathbf{x}_B)} \\ &= \frac{\frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)}{\frac{1}{\sqrt{(2\pi)^{n-m} \det(\boldsymbol{\Sigma}_{BB})}} \exp\left(-\frac{1}{2}(\mathbf{x}_B - \boldsymbol{\mu}_B)^\top \boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B)\right)} \\ &= \frac{1}{\sqrt{(2\pi)^m \frac{\det(\boldsymbol{\Sigma})}{\det(\boldsymbol{\Sigma}_{BB})}}} \exp\left(-\frac{1}{2}\left[(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) - (\mathbf{x}_B - \boldsymbol{\mu}_B)^\top \boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B)\right]\right)\end{aligned} \tag{14}$$

Note that

$$\begin{aligned}\det\left(\begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB}\end{bmatrix}\right)[\det(\boldsymbol{\Sigma}_{BB})]^{-1} &= \det\left(\begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB}\end{bmatrix}\right)\det(\boldsymbol{\Sigma}_{BB}^{-1}) \\ &= \det\left(\begin{bmatrix}\boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB}\end{bmatrix}\right)\det\left(\begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{BB}^{-1}\end{bmatrix}\right) \\ &= \det(\boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}) \\ &= \det(\mathbf{M})\end{aligned} \tag{15}$$

We denote $\hat{\mathbf{x}}_A = \mathbf{x}_A - \boldsymbol{\mu}_A$ and $\hat{\mathbf{x}}_B = \mathbf{x}_B - \boldsymbol{\mu}_B$, then we have:

$$\begin{aligned}f_{\mathbf{X}_A|\mathbf{X}_B}(\mathbf{x}_A|\mathbf{x}_B) &= \frac{1}{\sqrt{(2\pi)^m \det(\mathbf{M})}} \exp\Bigg(-\frac{1}{2}\Big[\hat{\mathbf{x}}_A^\top \mathbf{M}\hat{\mathbf{x}}_A + \hat{\mathbf{x}}_B^\top \boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}\mathbf{M}\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\hat{\mathbf{x}}_B \\ &\quad - \hat{\mathbf{x}}_B^\top \boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}\mathbf{M}\hat{\mathbf{x}}_A - \hat{\mathbf{x}}_A^\top \mathbf{M}\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\hat{\mathbf{x}}_B\Big]\Bigg) \\ &= \frac{1}{\sqrt{(2\pi)^m \det(\mathbf{M})}} \exp\left(-\frac{1}{2}\left[(\hat{\mathbf{x}}_A - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\hat{\mathbf{x}}_B)^\top \mathbf{M}(\hat{\mathbf{x}}_A - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\hat{\mathbf{x}}_B)\right]\right)\end{aligned} \tag{16}$$

Thus we can know that $X_A|X_B$ is still a MVN, and

$$\begin{aligned}\boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{X}_B - \boldsymbol{\mu}_B) \\ \boldsymbol{\Sigma}_{A|B} &= \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}\end{aligned} \tag{17}$$

4. Due to the following Theorems, we know linear combination of two independent MVN is still a MVN.

Theorem

If $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ are MVN vectors with X independent of Y, then the concatenated random vector $W = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is Multivariate Normal.

Definition

A random vector $X = (X_1, \ldots, X_k)$ is said to have a *Multivariate Normal* (MVN) distribution if every linear combination of the X_j has a Normal distribution. That is, we require

$$t_1 X_1 + \ldots + t_k X_k$$

to have a Normal distribution for any choice of constants $t_1, \ldots, t_k$. If $t_1 X_1 + \cdots + t_k X_k$ is a constant (such as when all $t_i = 0$), we consider it to have a Normal distribution, albeit a degenerate Normal with variance 0. An important special case is $k = 2$; this distribution is called the *Bivariate Normal*(BVN).

Denote $\mathbf{Y} = \mathbf{X} + \mathbf{X}'$, we have the following by transformation

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}) + E(\mathbf{X}') = \boldsymbol{\mu} + \boldsymbol{\mu}' \\ \mathrm{Cov}(\mathbf{Y}) &= \mathrm{Cov}(\mathbf{X} + \mathbf{X}') = \mathrm{Cov}(\mathbf{X}) + \mathrm{Cov}(\mathbf{X}') = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}' \end{aligned} \tag{18}$$

Probability and Statistics for EECS: Homework #9 Solutions

Professor Ziyu Shao

Problem 1

Let X and Y be i.i.d. Expo(λ), and transform them to $T = X + Y$, $W = X/Y$. Find the marginal PDFs of T and W, and the joint PDF of T and W.

Solution

By the statement, we have $X = \frac{WT}{W+1}, Y = \frac{T}{W+1}$. Then

$$\begin{aligned}
&\frac{\partial(x,y)}{\partial(w,t)} = \begin{vmatrix} \frac{t}{(w+1)^2} & \frac{w}{w+1} \\ \frac{-t}{(w+1)^2} & \frac{1}{w+1} \end{vmatrix} = \frac{t}{(w+1)^2} > 0 \\
\Longrightarrow\ & f_{W,T}(w,t) = f_{X,Y}(x,y)\left|\frac{\partial(x,y)}{\partial(w,t)}\right| \\
\Longrightarrow\ & f_{W,T}(w,t) = f_X(x)f_Y(y)\frac{t}{(w+1)^2} = \lambda e^{-\lambda x}\cdot\lambda e^{-\lambda y}\frac{t}{(w+1)^2}, (x \geq 0, y \geq 0) \\
=& \lambda^2 e^{-\lambda(x+y)}\frac{t}{(w+1)^2}, = \lambda^2 e^{-\lambda t}\frac{t}{(w+1)^2}, (x \geq 0, y \geq 0).
\end{aligned}$$

The marginal PDF of W is

$$f_W(w) = \int_0^\infty f_{W,T}(w,t)dt = \frac{\lambda}{(w+1)^2}\int_0^\infty t\lambda e^{-\lambda t}dt = \frac{\lambda}{(w+1)^2}\frac{1}{\lambda} = \frac{1}{(w+1)^2}, (w \geq 0).$$

Otherwise, the marginal PDF of W is 0.

The marginal PDF of T is

$$f_T(t) = \int_0^\infty f_{W,T}(w,t)dw = \lambda^2 te^{-\lambda t}\int_0^\infty \frac{1}{(w+1)^2}dw = \lambda^2 te^{-\lambda t}\int_1^\infty \frac{1}{u^2}du = \lambda^2 te^{-\lambda t}, (t \geq 0).$$

Otherwise, the marginal PDF of T is 0.

Problem 2

Let X, Y, Z be i.i.d. Unif$(0, 1)$, and $W = X + Y + Z$. Find the PDF of W.

Solution

Denote $T = X + Y$. It's obvious that $f_T(t) = 0$ for $t < 0$ or $t > 2$. For $0 \le t \le 2$, using the convolution formula, we have

$$f_T(t) = \int_{\mathbb{R}} f_X(x) f_Y(t - x) dx.$$

Observe that the function under the integral is non-trivial for

$$0 \le x \le 1,$$

$$0 \le t - x \le 1 \implies t - 1 \le x \le t.$$

Divide the problem into two cases. For $0 \le t \le 1$, we have

$$f_T(t) = \int_0^t 1 dx = t.$$

For $1 < t \le 2$, we have

$$f_T(t) = \int_{t-1}^1 1 dx = 2 - t.$$

Therefore, the PDF of T is

$$f_T(t) = \begin{cases} t, & 0 \le t \le 1; \\ 2 - t, & 1 < t \le 2; \\ 0, & \text{otherwise.} \end{cases}$$

Observe that $W = T + Z$, we have $f_W(w) = 0$ for $w < 0$ or $w > 3$. For $0 \le w \le 3$, using the convolution formula, we have that

$$f_W(w) = \int_{\mathbb{R}} f_T(t) f_Z(w - t) dt.$$

Observe that the function under the integral is non-trivial for

$$0 \le t \le 2$$
$$0 \le w - t \le 1 \implies w - 1 \le t \le w$$

Divide our problem into three cases. For $0 \le w \le 1$, we have that

$$f_W(w) = \int_0^w t dt = \frac{w^2}{2}.$$

For $1 < w \le 2$, we have that

$$f_W(w) = \int_{w-1}^1 t dt + \int_1^w (2 - t) dt = -w^2 + 3w - \frac{3}{2}.$$

For $2 < w \le 3$, we have that

$$f_W(w) = \int_{w-1}^2 (2 - t) dt = \frac{(w - 3)^2}{2}.$$

Finally, we obtain the PDF of W as

$$f_W(w) = \begin{cases} \frac{w^2}{2}, & 0 \le w \le 1; \\ -w^2 + 3w - \frac{3}{2}, & 1 < w \le 2; \\ \frac{(w-3)^2}{2}, & 2 < w \le 3; \\ 0, & \text{otherwise.} \end{cases}$$

Problem 3

Let (X, Y) be Bivariate Normal with $X \sim \mathcal{N}(0, \sigma_1^2)$ and $Y \sim \mathcal{N}(0, \sigma_2^2)$ marginally and with $\text{Corr}(X, Y) = \rho$. Find a constant c such that $Y - cX$ is independent of X.

Solution

Let's find c such that $\text{Corr}(Y - cX, X) = 0 \implies \text{Cov}(Y - cX, X) = 0$.
We have following

$$0 = \text{Cov}(Y - cX, X) = \text{Cov}(Y, X) - c\,\text{Cov}(X, X) = \rho\sigma_1\sigma_2 - c\sigma_1^2$$

Solving this equation we get that the only possibility for c is $c = \rho\frac{\sigma_2}{\sigma_1}$. Further, observe that for every $a, b \in \mathbb{R}$ we have that

$$a\left(Y - \rho\frac{\sigma_2}{\sigma_1}X\right) + bX = aY + \left(-\rho\frac{\sigma_2}{\sigma_1} + b\right)X$$

Using the fact that (X, Y) is Bivariate Normal, we have shown that also $\left(Y - \rho\frac{\sigma_2}{\sigma_1}X, X\right)$ is Bivariate normal. Then, there exist some μ_1, μ_2, o_1, o_2 such that

$$\begin{aligned} f\left(Y - \rho\frac{\sigma_2}{\sigma_1}X, X\right)(x, y) =& \frac{1}{2\pi o_1 o_2}\exp\left[-\frac{1}{2}\left(\frac{(x-\mu_1)^2}{o_1^2} + \frac{(y-\mu_2)^2}{o_2^2}\right)\right] \\ =& \frac{1}{\sqrt{2\pi}o_1}\exp\left[-\frac{1}{2}\frac{(x-\mu_1)^2}{o_1^2}\right] \cdot \frac{1}{\sqrt{2\pi}o_2}\exp\left[-\frac{1}{2}\frac{(x-\mu_2)^2}{o_2^2}\right] \end{aligned}$$

Observe that this joint PDF of $Y - \rho\frac{\sigma_2}{\sigma_1}X$ and X can be factored as product of two PDFs. Hence, we have proved that $Y - \rho\frac{\sigma_2}{\sigma_1}X$ and X are independent.

Problem 4

(a) Let $U_1, \ldots, U_n$ be i.i.d. Unif$(0,1)$. Let $U_{(j)}$ be the jth order statistic, $U_{(k)}$ be the kth order statistic, where $1 \leq j < k \leq n$. Find the joint PDF of $U_{(j)}$ and $U_{(k)}$.

(b) Let $X \sim \text{Bin}(n,p)$ and $B \sim \text{Beta}(j, n-j+1)$, where n is a positive integer and j is a positive integer with $j \leq n$. Show using a story about order statistics that

$$P(X \geq j) = P(B \leq p)$$

This shows that the CDF of the continuous r.v. B is closely related to the CDF of the discrete r.v. X, and is another connection between the Beta and Binomial.

(c) Show that

$$\int_0^x \frac{n!}{(j-1)!(n-j)!} t^{j-1}(1-t)^{n-j} dt = \sum_{k=j}^{n} \binom{n}{k} x^k (1-x)^{n-k},$$

without using calculus, for all $x \in [0,1]$ and j, n positive integers with $j \leq n$.

Solution

Figure 1: The joint PDF of $U_{(j)}$ and $U_{(k)}$.

(a) To have $U_{(j)}$ be in a tiny interval around a and $U_{(k)}$ be in a tiny interval around b, where $a < b$, we need to have one of the $U_{(j)}$'s be almost exactly at a, another be almost exactly at b, $j-1$ of them should be to the left of a, $n-k$ should be to the right of b, and the remaining $k-j-1$ should be between a and b, as shown in the picture. This gives that the PDF is

$$\begin{aligned} f_{(j),(k)}(a,b) &= \frac{n!}{(j-1)!(k-j-1)!(n-k)!} F(a)^{j-1} f(a) (F(b)-F(a))^{k-j-1} f(b) (1-F(b))^{n-k} \\ &= \frac{n!}{(j-1)!(k-j-1)!(n-k)!} a^{j-1}(b-a)^{k-j-1}(1-b)^{n-k}, (0 \leq a < b \leq 1) \end{aligned}$$

Otherwise, $f_{(j),(k)}(a,b) = 0$. The coefficient in front counts the number of ways to put the U_i's into the 5 categories "left of a", "at a", "between a and b", "at b", and "right of b" with the desired number in each category (which is the same idea used to find the coefficient in front of the Multinomial PMF). Equivalently, we could write the coefficient as $n(n-1)\binom{n-2}{j-1}\binom{n-j-1}{k-j-1}$.

(b) Let $U_1, \ldots, U_n$ be i.i.d. Unif$(0,1)$. Think of these as Bernoulli trials, where U_j is defined to be "successful" if $U_j \leq p$ (so the probability of success is p for each trial). Let X be the number of successes. Then $X \geq j$ is the same event as $U_{(j)} \leq p$, so $P(X \geq j) = P(U_{(j)} \leq p) = P(B \leq p)$ where the last equality is due to the Example 8.6.5 (Order statistics of Uniforms), *i.e.*, $U_{(j)} \sim \text{Beta}(j, n-j+1)$.

(c) Let $B \sim \text{Beta}(j, n-j+1), Y \sim \text{Bin}(n,x)$, by (b), we have

$$P(B \leq x) = P(Y \geq j) \implies \int_0^x \frac{n!}{(j-1)!(n-j)!} t^{j-1}(1-t)^{n-j} dt = \sum_{k=j}^{n} \binom{n}{k} x^k (1-x)^{n-k}.$$

Problem 5

(a) Let $p \sim \text{Beta}(a, b)$, where a and b are positive real numbers. Find $E(p^2(1-p)^2)$, fully simplified (Γ should not appear in your final answer).

Two teams, A and B, have an upcoming match. They will play five games and the winner will be declared to be the team that wins the majority of games. Given p, the outcomes of games are independent, with probability p of team A winning and $(1-p)$ of team B winning. But you don't know p, so you decide to model it as an r.v., with $p \sim \text{Unif}(0,1)$ a priori (before you have observed any data).
To learn more about p, you look through the historical records of previous games between these two teams, and find that the previous outcomes were, in chronological order, $AAABBAABAB$. (Assume that the true value of p has not been changing over time and will be the same for the match, though your beliefs about p may change over time.)

(b) Does your posterior distribution for p, given the historical record of games between A and B, depend on the specific order of outcomes or only on the fact that A won exactly 6 of the 10 games on record? Explain.

(c) Find the posterior distribution for p, given the historical data.

The posterior distribution for p from (c) becomes your new prior distribution, and the match is about to begin!

(d) Conditional on p, is the indicator of A winning the first game of the match positively correlated with, uncorrelated with, or negatively correlated of the indicator of A winning the second game of the match? What about if we only condition on the historical data?

(e) Given the historical data, what is the expected value for the probability that the match is not yet decided when going into the fifth game (viewing this probability as an r.v. rather than a number, to reflect our uncertainty about it)?

Solution

(a) Since the PDF of r.v. p is $f_p(x) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}$ for $p \in (0,1)$, we have

$$\begin{aligned} \mathrm{E}\left(p^2(1-p)^2\right) &= \int_0^1 x^2(1-x)^2 \cdot \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}\, dx \\ &= \frac{1}{\beta(a,b)} \int_0^1 x^{a+2-1}(1-x)^{b+2-1}\, dx. \end{aligned}$$

Because of the equation $\int_0^1 \frac{1}{\beta(a+2,b+2)} x^{a+2-1}(1-x)^{b+2-1}\, dx = 1$, we further have

$$\mathrm{E}\left(p^2(1-p)^2\right) = \frac{1}{\beta(a,b)} \cdot \beta(a+2, b+2).$$

Moreover, by the fact that $\beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ we can write the result as

$$\mathrm{E}\left(p^2(1-p)^2\right) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b+2)}{\Gamma(a+b+4)}.$$

Substitute $\Gamma(n) = (n-1)!$, we get

$$\mathrm{E}\left(p^2(1-p)^2\right) = \frac{(a+b-1)!}{(a-1)!(b-1)!} \cdot \frac{(a+1)!(b+1)!}{(a+b+3)!} = \frac{ab(a+1)(b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)}.$$

(b) The posterior distribution for p does not depend on the specific order of outcomes. The reason is that the probability of all possible orders are the same when p is given, *i.e.*, orders don't influence our belief about p.

(c) The prior distribution of p is $\text{Unif}(0,1)$, which is equivalent to $\text{Beta}(1,1)$. By Beta-Binomial conjugacy, the posterior distribution of p is $\text{Beta}(1+6, 1+4)$, *i.e.*, $\text{Beta}(7,5)$.

(d) Let I_1 be the indicator of A winning the first game of the match, and I_2 be the indicator of A winning the second game of the match. Conditional on p, I_1 and I_2 are uncorrelated since they are i.i.d. Bernoulli r.v.s when p is given.

If we only condition on the historical data, *i.e.*, $p \sim \text{Beta}(7,5)$, we have

$$\text{E}(I_1) = \Pr(I_1 = 1) = \int_0^1 \Pr(I_1 = 1 | p = x) f_p(x)\, dx = \int_0^1 x f_p(x)\, dx = \text{E}(p),$$

where the third equation is because $I_1 | p \sim \text{Bern}(p)$. Similarly we also have $\text{E}(I_2) = \text{E}(p)$. On the other hand,

$$\begin{aligned}\text{E}(I_1 I_2) &= \Pr(I_1 I_2 = 1) = \Pr(I_1 = 1, I_2 = 1) \\ &= \int_0^1 \Pr(I_1 = 1, I_2 = 1 | p = x) f_p(x)\, dx = \int_0^1 x^2 f_p(x)\, dx = \text{E}\left(p^2\right)\end{aligned}$$

Since I_1 and I_2 are independent Bernoulli r.v.s with distribution $\text{Bern}(p)$ given p. Therefore

$$\text{Cov}(I_1, I_2) = \text{E}(I_1 I_2) - \text{E}(I_1)\,\text{E}(I_2) = \text{E}\left(p^2\right) - \text{E}^2(p) = \text{Var}(p) > 0,$$

which implies that I_1 and I_2 are positively correlated.

(e) Let X be the number of games that A win in the first 4 games of the match, then we have $X | p \sim \text{Bin}(4, p)$. The probability that the match is not yet decided when going into the fifth game given p is

$$\Pr(X = 2 | p) = \binom{4}{2} p^2 (1-p)^2.$$

Given the historical data, *i.e.*, $p \sim \text{Beta}(7,5)$, the expected value of probability that the match is not yet decided when going into the fifth game is

$$\text{E}\left(\binom{4}{2} p^2 (1-p)^2\right) = \binom{4}{2} \text{E}\left(p^2 (1-p)^2\right).$$

From (a) we know that when $p \sim \text{Beta}(7,5)$, $\text{E}\left(p^2(1-p)^2\right) = \frac{2}{39}$. Thus the expected value of probability is $\binom{4}{2} \times \frac{2}{39} = \frac{4}{13}$.

Problem 6

If $X \sim \text{Pois}(\lambda), Z \sim \text{Gamma}(k+1, 1)$, where k is a nonnegative integer. Use two different methods to show the Poisson-Gamma Duality holds:

$$P(X \leq k) = P(Z > \lambda).$$

Solution

Let the r.v. $B \sim \text{Beta}(k+1, n-(k+1)+1), Y \sim \text{Bin}(n, p)$ and we have the following due to the above identity in (c),

$$P(B \leq p) = P(Y \geq k+1) \implies P(B \geq p) = P(Y \leq k)$$

Therefore, we have

$$P(Y \leq k) = \int_p^1 \frac{n!}{k!(n-k-1)!} t^k (1-t)^{n-k-1} dt$$

Let $t = x/n$,

$$\begin{aligned}
P(Y \leq k) =& \frac{n!}{k!(n-k-1)!} \times \int_{np}^{n} \left(\frac{x}{n}\right)^k \left(1-\frac{x}{n}\right)^{n-k-1} \frac{1}{n} dx \\
=& \frac{(n-1)!}{k!(n-1-k)!} \times \int_{np}^{n} \left(\frac{x}{n}\right)^k \left(1-\frac{x}{n}\right)^{n-k-1} dx \\
=& \int_{np}^{n} \binom{n-1}{k} \left(\frac{x}{n}\right)^k \left(1-\frac{x}{n}\right)^{n-1-k} dx \\
=& \int_{np}^{n} P(A=k) dx,
\end{aligned}$$

where $A \sim \text{Bin}(n-1, x/n)$ and $P(A=k)$ is its PMF (with value at point k).
Now we consider the "Poisson approximation to Binomial". That is, as $n \to \infty, p \to 0$ and fix $\lambda = np$, we have the following approximations:

- Approximate $Y \sim \text{Bin}(n, p)$ with $X \sim \text{Pois}(\lambda)$;
- Approximate $A \sim \text{Bin}(n-1, \frac{x}{n})$ with $C \sim \text{Pois}(x)$ (since $\frac{x}{n} \to 0$ and $n-1 \approx n$ when $n \to \infty$)

Therefore, we have

$$\begin{aligned}
&\text{LHS} = P(Y \leq k) = P(X \leq k); \\
&\text{RHS} = \int_\lambda^\infty P(A=k) dx = \int_\lambda^\infty P(C=k) dx = \int_\lambda^\infty \frac{e^{-x} x^k}{k!} dx.
\end{aligned}$$

Since $Z \sim \text{Gamma}(k+1, 1)$, its PDF satisfies

$$f_Z(x) = \frac{1}{\Gamma(k+1)} \times (1 \cdot x)^{k+1} e^{-1 \cdot x} \times \frac{1}{x} = \frac{e^{-x} x^k}{k!}.$$

Therefore, we have

$$P(X \leq k) = \int_\lambda^\infty \frac{e^{-x} x^k}{k!} dx = \int_\lambda^\infty f_Z(x) dx = P(Z > \lambda).$$

In a Poisson process where the arrivals preserve the unit arriving rate:

1. The number of arrivals within a certain interval can be modeled as a Poisson random variable. That is, $X \sim \text{Pois}(\lambda)$ can be interpreted as the number of arrivals in the interval $[0, \lambda]$. Accordingly, $P(X \leq k)$ focuses on the event that "number of arrivals in interval $[0, \lambda]$ is less or equal to k".

2. The waiting time between arrivals can be model as exponential random variables with a unit rate. Note that we know Z is a summation of $(k+1)$ i.i.d. exponential r.v.s with unit rate, *i.e.* $Z = X_1 + \cdots + X_{k+1} \sim \text{Gamma}(k+1, 1)$, where $X_i \sim \text{Expo}(1), i = 1, \ldots, k+1$. By viewing X_i as the waiting time between the $(i-1)$th and the ith arrival, Z is the total waiting time for the $(k+1)$th arrivals. Accordingly, $P(Z > \lambda)$ focuses on the event that "the waiting time for the $(k+1)$th arrival is longer than λ".

Clearly, two interpreted events above are equivalent since once we limit the number of arrivals in interval $[0, \lambda]$ to be less or equal to k, the waiting time for the $(k+1)$th arrival is naturally larger than λ, vice versa. This proves the identity immediately.

Probability and Statistics for EECS: Homework #10 Solutions

Professor Ziyu Shao

Problem 1

Use the methods of inverse transform sampling to obtain samples from each of the following continuous distributions:

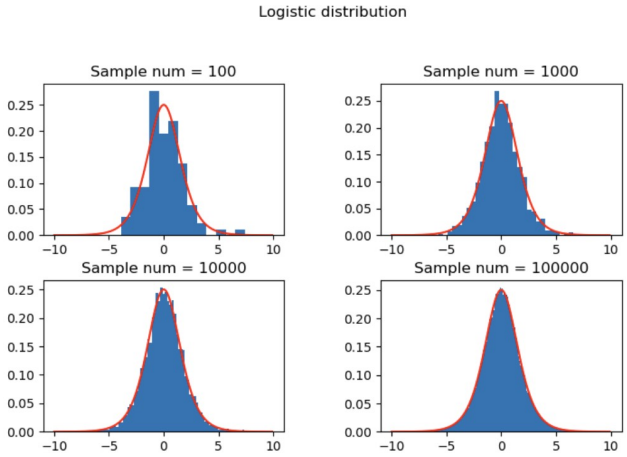
(a) Logistic distribution with CDF $F(x) = e^x/(1+e^x), x \in \mathbb{R}$.

(b) Rayleigh distribution with CDF $F(x) = 1 - e^{-x^2/2}, x > 0$.

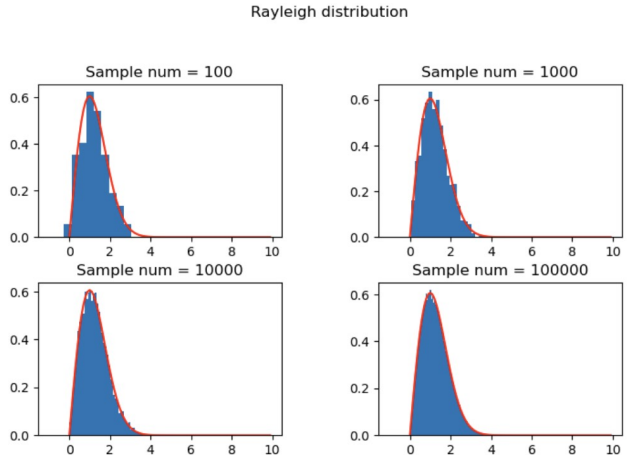
(c) Exponential distribution with CDF $F(x) = 1 - e^{-x}, x > 0$.

Solution

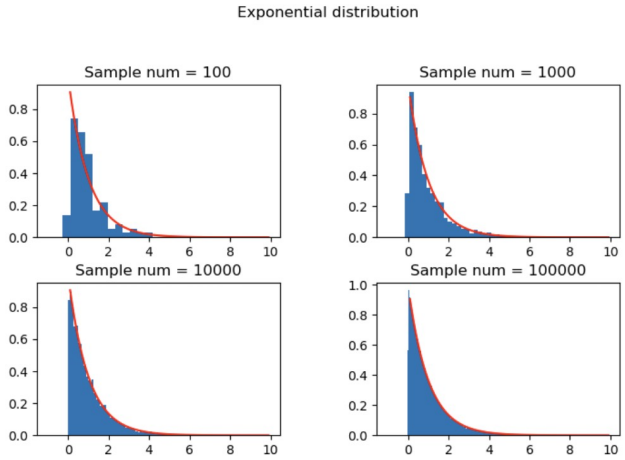
(a) The CDF of logistic distribution can be represented by $F(x) = \frac{1}{1+e^{-x}}, \forall x \in \mathbb{R}$, which is continuous and strictly increasing on the support of the distribution. So we can obtain its inverse function $F^{-1}(x) = -\log(1/x - 1)$ and the PDF is $f(x) = \frac{e^x}{(1+e^x)^2}$.



(b) The CDF $F(x) = 1 - e^{-x^2/2}, \forall x > 0$, which is continuous and strictly increasing on the support of the distribution. So we can obtain its inverse function $F^{-1}(x) = \sqrt{-2\log(1-y)}$ and the PDF is $f(x) = xe^{-x^2/2}$.



(c) The CDF $F(x) = 1 - e^{-x}, \forall x > 0$, which is continuous and strictly decreasing on the support of the distribution. So we can obtain its inverse function $F^{-1}(x) = \sqrt{-2\log(1-y)}$ and the PDF is $f(x) = -\log(1-x)$.



Problem 2

Use the methods of inverse transform sampling to obtain samples from each of the following discrete distributions:

(a) Bernoulli distribution Bern(0.5).

(b) Binomial distribution Bin(20, 0.5).

(c) Geometric distribution Geom(0.5).

(d) Negative Binomial distribution NBin(10, 0.5).

(e) Poisson distribution Pois(1).

Solution

Theorem

Let F be a CDF which is a continuous function and strictly increasing on the support of the distribution. This ensures that the inverse function F^{-1} exists, as a function from $(0,1)$ to $\mathbb{R}$. We then have the following results.

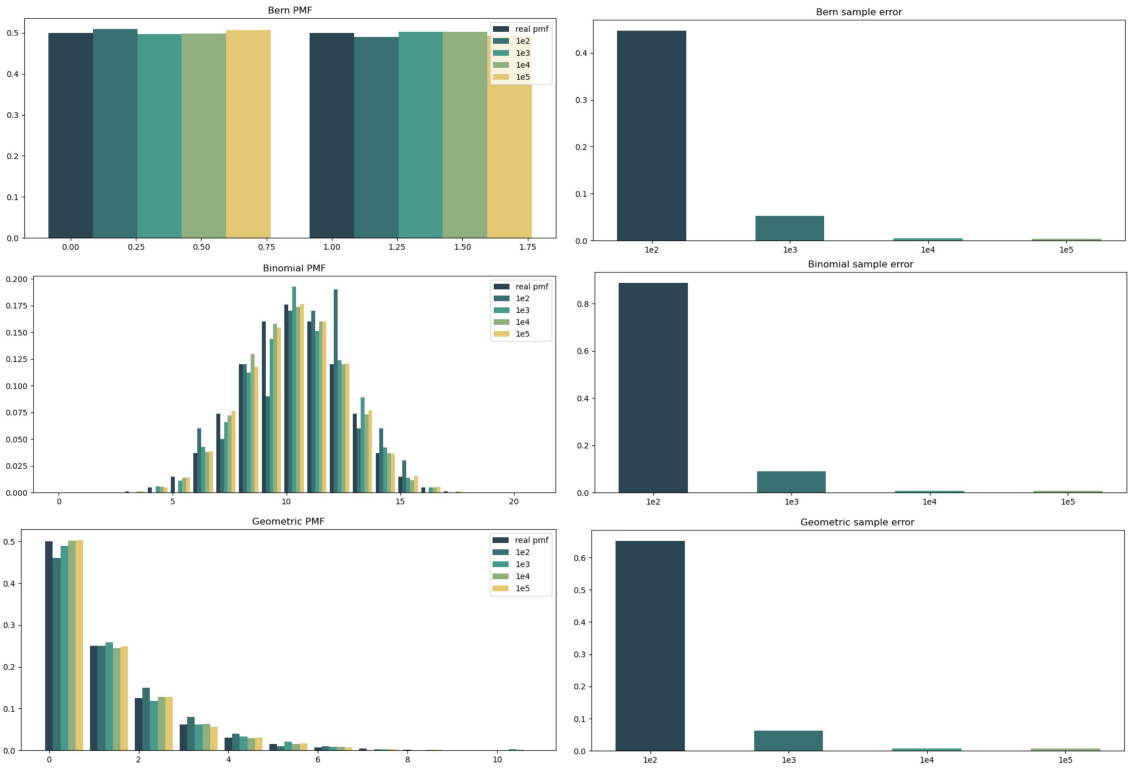
1. *Let $U \sim \text{Unif}(0,1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.*
2. *Let X be an r.v. with CDF F. Then $F(X) \sim \text{Unif}(0,1)$.*

Algorithm Inverse-Transform Method: PMF Case

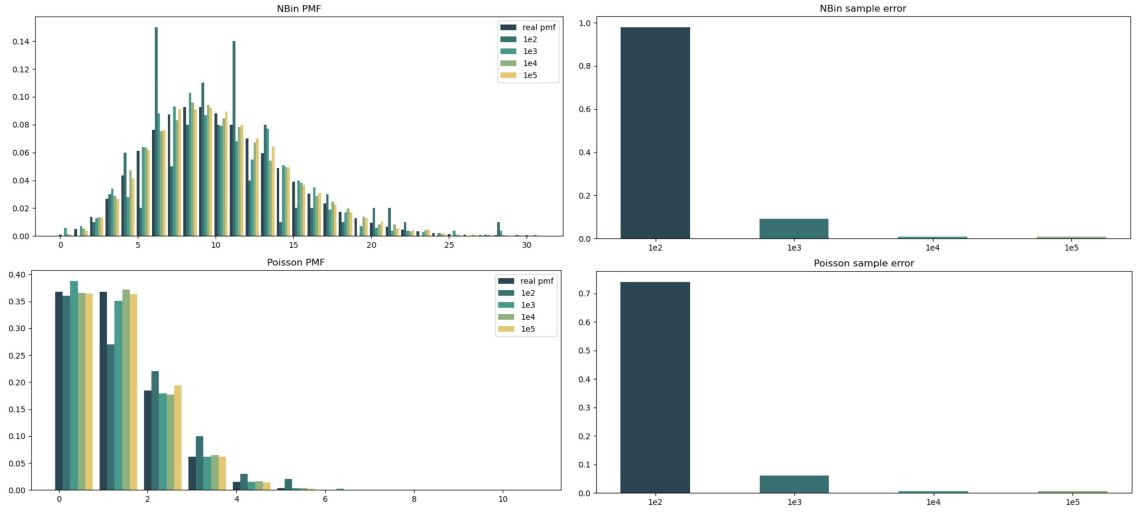
input: Discrete cumulative distribution function F with monotonic sequence $\{x_j\}$
output: Discrete random variable X distributed according to F.

1: Generate $U \sim \text{Unif}(0,1)$.
2: Find the smallest positive integer, k, such that $U \leq F(x_k)$. Let $X \leftarrow x_k$.
3: **return** X

The following figures cover the answers from (a)-(c).



The following figures cover the answers from (d) and (e).



Problem 3

(a) Use the Box-Muller method to obtain samples from the standard Normal distribution $\mathcal{N}(0,1)$.

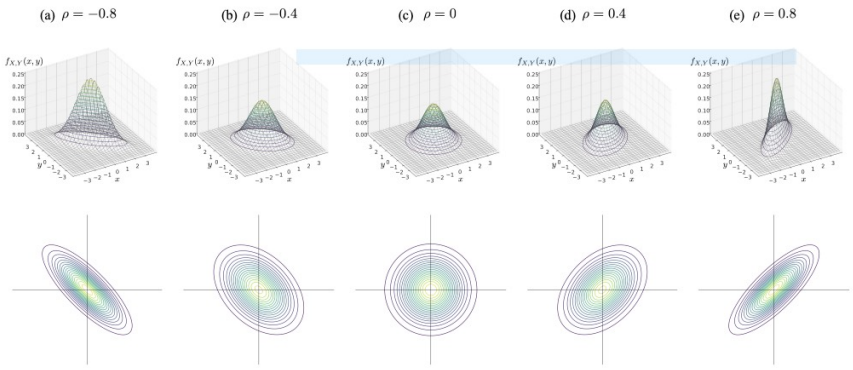
(b) Use the Acceptance-Rejection method to obtain samples from standard Normal distribution $\mathcal{N}(0,1)$.

(c) Compare the pros and cons of the above two methods in terms of factors like variance, sampling efficiency, running speed, etc.

(d) Use the following transformation to generate samples from bivariate Normal distribution with correlation coefficient ρ:

$$X = Z$$
$$Y = \rho Z + \sqrt{1-\rho^2}W,$$

where $-1 < \rho < 1, Z$ and W are i.i.d. random variables following $\mathcal{N}(0,1)$. The joint pdf function of bivariate Normal distribution with correlation coefficient ρ and the corresponding contour (or isocontour) are shown in the following figure for your reference:



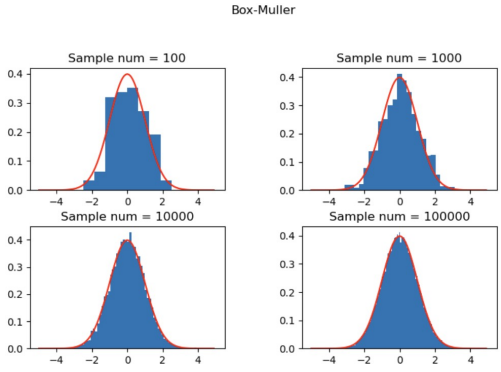
Solution

(a) The following is the pseudocode and simulation results of Box-Muller method.

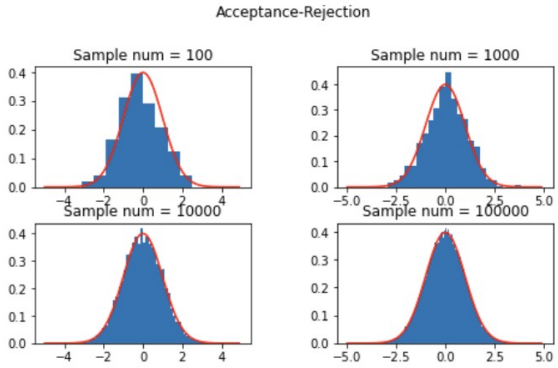
Let $U \sim \text{Unif}(0, 2\pi)$, and let $T \sim \text{Expo}(1)$ be independent of U. Define $X = \sqrt{2T}\cos U$ and $Y = \sqrt{2T}\sin U$. Then X and Y are independent, and their marginal distributions are standard normal distribution.

Algorithm Normal Random Variable Generation: Box-Muller Approach

output: Independent standard normal random variables X and Y.
1: Generate two independent random variables, U_1 and U_2, from $Unif(0,1)$.
2: $X \leftarrow (-2\ln U_1)^{1/2}\cos(2\pi U_2)$
3: $Y \leftarrow (-2\ln U_1)^{1/2}\sin(2\pi U_2)$
4: **return** X, Y



(b)
- Generate Exponential distribution with parameter $\lambda = 1$ from Unif(0,1).
  The CDF of Expo(1) is $F(x) = 1 - e^{-x}, x \geq 0$. The inverse of the CDF is $F^{-1}(u) = -\ln(1-u)$. If $u \sim \text{Unif}(0,1)$, then $y = -\ln(1-U) \sim \text{Expo}(1)$.
- Generate $U \sim \text{Unif}(0,1)$.



- Let $Z \sim N(0,1)$, we will generate $X \sim |Z|$ firstly.
  The pdf of X is $p(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, x \geq 0$. The pdf of Expo(1) is $q(x) = e^{-x}, x \geq 0$.
  Let $c = \sup_x \frac{p(x)}{q(x)} = \sup_x \sqrt{\frac{2}{\pi}} e^{x-\frac{1}{2}x^2} = \sqrt{\frac{2e}{\pi}}$.
  If $u < \frac{p(y)}{cq(y)} = e^{-\frac{1}{2}(y-1)^2}$, set $x = y$; otherwise go back to step 1.
- $Z = \begin{cases} x, & \text{w.p. } 0.5 \\ -x, & \text{w.p. } 0.5 \end{cases}$. Then $Z \sim N(0,1)$.

(c) In terms of sampling Normal distribution, their variance are similar, while the sample efficiency of BoxMuller is higher with also higher running speed.

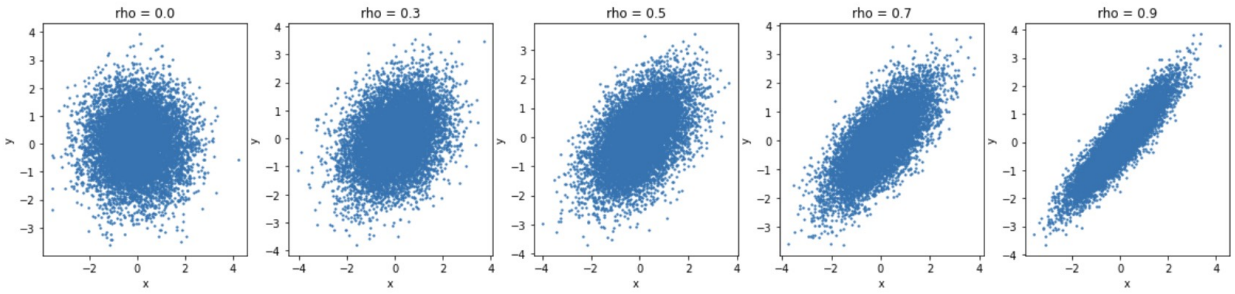
Box-Muller:

- Pros: It is easy to implement, and the method only uses Unif$(0,1)$ as the basis data sample, which is simple to sample.
- Cons: Only the standard normal distribution can be sampled by this method.

Acceptance-Rejection:

- Pros: It can sample many kinds of probability distribution including many distributions that is difficult to sample directly.
- Cons: The domain of function $g(x)$ must cover the domain of function $f(x)$. If c is closed to 1, the basis distribution g is still difficult to sample; while if c is closed to 0, the probability of acceptance success will be small, which will cause low efficiency.

(d) The following is the simulation result of sampling BVN:



Problem 4

(a) Use the Acceptance-Rejection Method to generate a random variable with distribution Beta$(2, 4)$.

(b) Use Monte Carlo methods to evaluate the integration

$$\int_0^1 \frac{4}{1+x^2} dx.$$

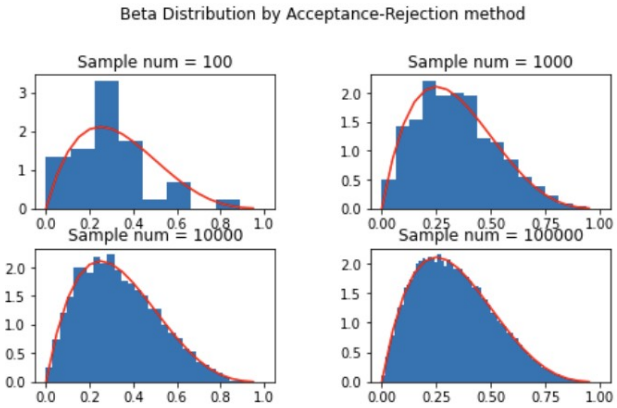
(c) Use Monte Carlo methods to evaluate the integration

$$\int_0^4 \sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x}}}} dx.$$

(d) Use Monte Carlo methods to estimate the value of π.

(e) Use importance sampling method to evaluate the probability of rare event $c = P(Y > 8)$, where $Y \sim \mathcal{N}(0, 1)$.

Solution



(a) We use the Acceptance-Rejection method to generate the Beta distribution.

- Generate two independent random variables, U_1 and U_2, from Unif(0,1). Set $V_1 = U_1^{1/a}$ and $V_2 = U_2^{1/b}$.
- If $V_1 + V_2 \leq 1$, set $X = V_1/(V_1 + V_2)$; otherwise go back to step 1.

Proof. Let $W = V_1 + V_2$. Then the CDF of X is

$$F(x) = P(X \leq x) = P\left(\frac{V_1}{W} \leq x | W \leq 1\right) = \frac{P(V_1 \leq xW, W \leq 1)}{P(W \leq 1)}.$$

By changing of variables, we have

$$f_{V_1}(v_1) = f_{U_1}(v_1^a)|av_1^{a-1}| = av_1^{a-1}, \quad 0 < v_1 < 1,$$
$$f_{V_2}(v_2) = f_{U_2}(v_2^b)|bv_2^{b-1}| = bv_2^{b-1}, \quad 0 < v_2 < 1.$$

The Jacobian matrix of transmission for $V_1, V_2 = W - V_1$ is

$$J = \begin{pmatrix} \frac{\partial v_1}{\partial v_1} & \frac{\partial v_1}{\partial w} \\ \frac{\partial v_2}{\partial v_1} & \frac{\partial v_2}{\partial w} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

Then the joint pdf of V_1 and W is

$$f_{V_1,W}(v_1,w) = f_{V_1,V_2}(v_1,v_2)|J| = abv_1^{a-1}(w-v_1)^{b-1}, \quad 0 < v_1 < 1, v_1 < w < v_1 + 1.$$

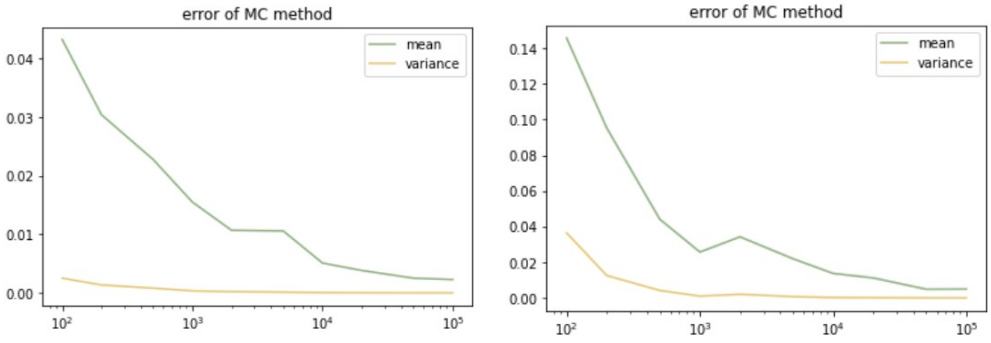
Therefore, we have

$$\begin{aligned} F(x) &= \frac{P(V_1 \le xW, W \le 1)}{P(W \le 1)} \\ &= \frac{\int_0^1 \int_0^{wx} abv_1^{a-1}(w-v_1)^{b-1} dv_1 dw}{P(W \le 1)} \\ &= \frac{\int_0^1 \int_0^{wx} v_1^{a-1}(w-v_1)^{b-1} dv_1 dw}{c}, \end{aligned}$$

where c is a constant. Taking the derivate and applying the fundamental theorem of calculus, we have

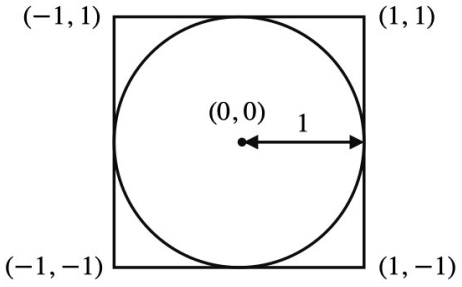
$$f(x) = \frac{1}{c}\int_0^1 w(wx)^{a-1}(w-wx)^{b-1}dw = \frac{1}{c(a+b)}x^{a-1}(1-x)^{b-1}.$$

The following figures cover the results from (b)-(c)



- Ground truth 3.1415926535897936, Estimation with 2×10^6 samples: 3.14160782958
- Ground truth: 7.6766100019, Estimation with 2×10^6 samples: 7.6765401290

(d) Ground truth: 3.1415926535, Estimation with 2×10^6 samples: 3.141808



- Indicator: bridge between expectation and probability
- Given event A:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{Otherwise} \end{cases}.$$

- For random variable X:

$$\begin{aligned} P(X \in A) &= 1 \cdot P(X \in A) + 0 \cdot P(X \notin A) \\ &= E(I_A(X)) \\ &\approx \frac{1}{n}\sum_{i=1}^{n} I_A(X_i). \end{aligned}$$

(e)
- Without importance sampling

$$c = P(Y > 8) = \mathbb{E}[I(Y > 8)] = \int_{-\infty}^{\infty} I(Y > 8) f(y) dy, f \sim \mathcal{N}(0,1),$$

$$c \approx \frac{1}{n}\sum_{j=1}^{n} I(Y_j > 8), Y_j \sim \mathcal{N}(0,1).$$

- With importance sampling

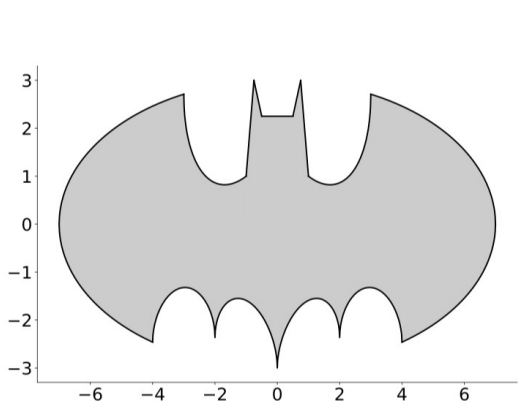
$$c \approx \frac{1}{n}\sum_{j=1}^{n}\frac{h(Y_j)f(Y_j)}{g(Y_j)} = \frac{1}{n}\sum_{j=1}^{n} I(Y_j > 8)\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}Y_j^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(Y_j-8)^2}} = \frac{1}{n}\sum_{j=1}^{n} I(Y_j > 8)e^{-8Y_j+32}, Y_j \sim g = \mathcal{N}(8,1)$$

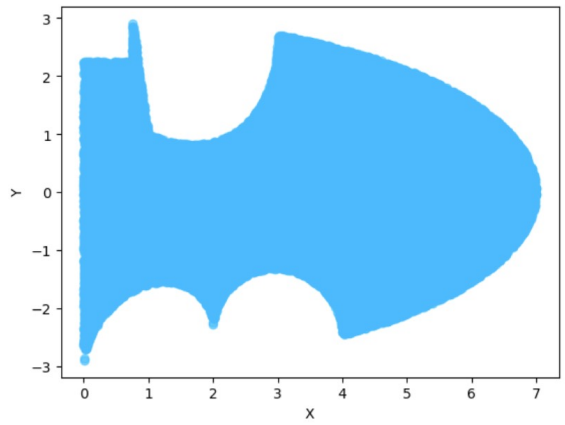
The following is the simulation result with 10^7 samples:

- With importance sampling: 6.228×10^{-16}
- Without importance sampling: 0.0

Problem 5

Use Monte Carlo methods to compute the area of Batman Curve and compare it with the exact value. Please refer to the following webpage for more details: https://mathworld.wolfram.com/BatmanCurve.html





Solution

1. Upper bound:

$$0 \geq x < \frac{3}{4} \Rightarrow y \leq 2\frac{1}{4}$$

$$\frac{3}{4} \geq x < 1 \Rightarrow y \leq -8x + 9$$

$$1 \geq x < 3 \Rightarrow y \leq 6\frac{\sqrt{10}}{7} + \frac{3}{2} - \frac{1}{2}x - \frac{3\sqrt{10}}{7}\sqrt{4 - (x-1)^2}$$

$$3 \geq x \leq 7 \Rightarrow y \leq 3\sqrt{1 - \frac{x^2}{49}}$$

2. Lower bound:

$$0 \geq x < 4 \Rightarrow y \geq \frac{1}{2}x + \sqrt{1 - (|x-2| - 1)^2} - \frac{3\sqrt{33} - 7}{112}x^2 - 3$$

$$4 \geq x \leq 7 \Rightarrow y \geq -3\sqrt{1 - \frac{x^2}{49}}$$

The probability of landing in the betman sign = The area of betman sign / The sample space

- Find the probability using upper and lower bounds via method similar for sampling the value of π
- The half sample space has the area $6 * 7 = 42$
- The final result of the total area of the betman sign is about 48.09084.

Probability & Statistics for EECS: Homework #11

Due on Dec 2, 2023 at 23:59

Problem 1

Let X be a discrete r.v. whose distinct possible values are $x_0, x_1, \ldots$, and let $p_k = P(X = x_k)$. The entropy of X is $H(X) = \sum_{k=0}^{\infty} p_k \log_2(1/p_k)$.

(a) Find $H(X)$ for $X \sim \text{Geom}(p)$.

(b) Let X and Y be i.i.d. discrete r.v.s. Show that $P(X = Y) \geq 2^{-H(X)}$. Hint: Jensen's Inequality.

Solution

(a) The PMF of X is $P(X = k) = p(1-p)^k$ since there is $X \sim \text{Geom}(p)$. Thus we have

$$\begin{aligned} H(X) &= -\sum_{k=0}^{\infty} p(1-p)^k \log_2 \left(p(1-p)^k\right) \\ &= -p\sum_{k=0}^{\infty} k(1-p)^k \log_2(1-p) - p\log_2 p \sum_{k=0}^{\infty}(1-p)^k \\ &= -\log_2 p - \frac{1-p}{p}\log_2(1-p) \end{aligned}$$

(b) Since X and Y are i.i.d random variables, via LOTP, we have

$$\begin{aligned} P(X = Y) &= \sum_{k=0}^{\infty} P(X = Y | Y = k) \cdot P(Y = k) \\ &= \sum_{k=0}^{\infty} P(X = k) \cdot P(Y = k) = \sum_{k=0}^{\infty} p_k^2 \end{aligned}$$

Denote Z as a new discrete random variable such that $P(Z = p_k) = p_k$, then we have:

$$E(Z) = \sum_{k=0}^{\infty} p_k \times p_k = P(X = Y)$$

Since $\log(\cdot)$ is a convex function, according to Jensen's inequality, we have $E(\log(Z)) \leq \log(E(Z))$, thus there is

$$\begin{aligned} &\sum p_k \log_2 p_k \leq \log_2 \sum p_k^2 \\ \Leftrightarrow &- H(X) \leq \log_2 P(X = Y) \\ \Leftrightarrow &P(X = Y) \geq 2^{-H(X)}. \end{aligned}$$

Problem 2

Let $X \sim \text{Pois}(\lambda)$. The conditional distribution of X, given that $X \geq 1$, is called a truncated Poisson distribution.

(a) Find $E(X|X \geq 1)$.

(b) Find $\text{Var}(X|X \geq 1)$.

Solution

(a) Using LOTE, we have that

$$E(X) = E(X \mid X \geq 1)P(X \geq 1) + E(X \mid X = 0)P(X = 0).$$

Now, use that $E(X) = \lambda$ and that simply $E(X \mid X = 0) = 0$. Also, we have that

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}.$$

Finally we have that

$$E(X \mid X \geq 1) = \frac{E(X)}{P(X \geq 1)} = \frac{\lambda}{1 - e^{-\lambda}}.$$

(b) For variance, we have that

$$\text{Var}(X \mid X \geq 1) = E\left(X^2 \mid X \geq 1\right) - \Big(E(X \mid X \geq 1)\Big)^2.$$

Similarly as before, we have that

$$E\left(X^2 \mid X \geq 1\right) = \frac{E\left(X^2\right)}{P(X \geq 1)} = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}},$$

and

$$\Big(E(X \mid X \geq 1)\Big)^2 = \left(\frac{\lambda}{1 - e^{-\lambda}}\right)^2,$$

so the answer is

$$\text{Var}(X \mid X \geq 1) = \frac{\lambda + \lambda^2}{1 - e^{-\lambda}} - \left(\frac{\lambda}{1 - e^{-\lambda}}\right)^2.$$

Problem 3

Let $X_1 \sim \text{Expo}(\lambda_1), X_2 \sim \text{Expo}(\lambda_2)$ and $X_3 \sim \text{Expo}(\lambda_3)$ be independent.

(a) Find $E(X_1 \mid X_1 > 2023)$

(b) Find $E(X_1 + X_2 + X_3 \mid X_1 > 2023, X_2 > 2024, X_3 > 2025)$ in terms of $\lambda_1, \lambda_2, \lambda_3$.

Solution

(a) According to the memoryless property of exponential distribution, we have $E(X - 2023|X > 2023) = E(X)$. Thus we can obtain the conditional expectation as follows:

$$E(X|X > 2023) = 2023 + E(X - 2023|X > 2023) = 2023 + E(X) = 2023 + \frac{1}{\lambda_1}.$$

(b) Since X_1, X_2, X_3 are independent, we have

$$\begin{aligned}
& E(X_1 + X_2 + X_3|X1 > 2023, X_2 > 2024, X_3 > 2025) \\
= \; & E(X_1|X_1 > 2023, X_2 > 2024, X_3 > 2025) \\
& + E(X_2|X1 > 2023, X_2 > 2024, X_3 > 2025) \\
& + E(X_3|X_1 > 2023, X_2 > 2024, X_3 > 2025) \\
= \; & E(X_1|X_1 > 2023) + E(X_2|X_2 > 2024) + E(X_3|X_3 > 2025) \\
= \; & E(X_1 - 2023|X_1 > 2023) + E(X_2 - 2024|X_2 > 2024) + E(X_3 - 2025|X_3 > 2025) + 6072 \\
= \; & E(X_1) + E(X_2) + E(X_3) + 6072 \\
= \; & \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} + 6072
\end{aligned}$$

Problem 4

Let X and Y be two continuous random variables with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 6xy & \text{if } 0 \le x \le 1, 0 \le y \le \sqrt{x} \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the marginal distributions of X and Y. Are X and Y independent?

(b) Find $E[X \mid Y = y]$ and $\text{Var}[X \mid Y = y]$ for $0 \le y \le 1$.

(c) Find $E[X \mid Y]$ and $\text{Var}[X \mid Y]$.

Solution

(a) The supports of X and Y are both $[0,1]$. In this way, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_0^{\sqrt{x}} 6xydy \\ &= 3xy^2\Big|_{y=0}^{y=\sqrt{x}} \\ &= 3x^2, \end{aligned}$$

and

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \int_{y^2}^{1} 6xydx \\ &= 3yx^2\Big|_{x=y^2}^{x=1} \\ &= 3y - 3y^5. \end{aligned}$$

Therefore,

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$f_Y(y) = \begin{cases} 3y - 3y^5 & \text{if } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $f_{X,Y}(x,y) \neq f_X(x)f_Y(y)$, X and Y are not independent.

(b) Since

$$E[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx,$$

to calculate $E[X|Y=y]$, we need to first calculate $f_{X|Y}(x|y)$.

If $y^2 \le x \le 1$,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2x}{1-y^4}.$$

In this way,

$$f_{X|Y}(x|y) = \begin{cases} \frac{2x}{1-y^4} & \text{if } y^2 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Problem 4 continued on next page...

Therefore,

$$\begin{aligned} E[X|Y=y] &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \\ &= \int_{y^2}^{1} x\frac{2x}{1-y^4}dx \\ &= \frac{2}{3(1-y^4)}x^3\Big|_{x=y^2}^{x=1} \\ &= \frac{2(1-y^6)}{3(1-y^4)} \\ &= \frac{2}{3}\cdot\frac{1+y^2+y^4}{1+y^2} \end{aligned}$$

Since

$$\mathrm{Var}[X|Y=y] = E[X^2|Y=y] - (E[X|Y=y])^2,$$

to calculate $\mathrm{Var}[X|Y=y]$, we need to first calculate $E[X^2|Y=y]$.

Since

$$\begin{aligned} E[X^2|Y=y] &= \int_{-\infty}^{\infty} x^2 f_{X|Y}(x|y)dx \\ &= \int_{y^2}^{1} x^2\frac{2x}{1-y^4}dx \\ &= \frac{1}{2(1-y^4)}x^4\Big|_{x=y^2}^{x=1} \\ &= \frac{1-y^8}{2(1-y^4)} \\ &= \frac{1+y^4}{2}, \end{aligned}$$

we have,

$$\begin{aligned} \mathrm{Var}[X|Y=y] &= E[X^2|Y=y] - (E[X|Y=y])^2 \\ &= \frac{1+y^4}{2} - \left(\frac{2(1-y^6)}{3(1-y^4)}\right)^2 \\ &= \frac{1+y^4}{2} - \frac{4}{9}\cdot\frac{(1+y^2+y^4)^2}{(1+y^2)^2} \end{aligned}$$

(c) According to the result in question(b), we have

$$\begin{aligned} E[X|Y] &= \frac{2}{3}\cdot\frac{1+Y^2+Y^4}{1+Y^2}, \\ \mathrm{Var}[X|Y] &= \frac{1+Y^4}{2} - \frac{4}{9}\cdot\frac{(1+Y^2+Y^4)^2}{(1+Y^2)^2}. \end{aligned}$$

Problem 5

Instead of predicting a single value for the parameter, we give an interval that is likely to contain the parameter: A $1-\delta$ confidence interval for a parameter p is an interval $[\hat{p}-\epsilon, \hat{p}+\epsilon]$ such that $P(p \in [\hat{p}-\epsilon, \hat{p}+\epsilon]) \geq 1-\delta$. Now we toss a coin with probability p landing heads and probability $1-p$ landing tails. The parameter p is unknown and we need to estimate its value from experiment results. We toss such coin N times. Let $X_i = 1$ if the ith result is head, otherwise 0. We estimate p by using

$$\hat{p} = \frac{X_1 + \ldots + X_N}{N}.$$

Find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(a) Method 1: Adopt Chebyshev inequality to find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(b) Method 2: Adopt Hoeffding bound to find the $1-\delta$ confidence interval for p, then discuss the impacts of δ and N.

(c) Discuss the pros and cons of the above two methods.

Solution

Since $X_i \sim \text{Bern}(p), X_i \in \{0,1\}$, we have $\mathbb{E}[X_i] = p$ and $\mathbb{V}[X_i] = p(1-p)$. Therefore, we have

$$\mathbb{E}[\hat{p}] = p, \mathbb{V}[\hat{p}] = \frac{p(1-p)}{N}.$$

Besides, we know that

$$P(p \in [\hat{p}-\varepsilon, \hat{p}+\varepsilon]) \geq 1-\delta \Leftrightarrow P(|\hat{p}-p| \geq \varepsilon) \leq \delta.$$

(a) Applying Chebyshev's inequality on random variable $\hat{p}$, we have

$$P(|\hat{p}-p| \geq \epsilon) \leq \frac{p(1-p)}{N\epsilon^2} \Rightarrow \delta = \frac{p(1-p)}{N\epsilon^2}, \epsilon = \sqrt{\frac{p(1-p)}{N\delta}}$$

Therefore, we know that δ negatively correlates with ϵ, i.e., given a fixed number of samples N, there is natural trade-off between accuracy and confidence. Besides, 1) Fix the confidence interval parametrized by δ, reducing the estimation error ϵ requires increasing the number of samples N. 2) Fix the estimation error ϵ, narrowing the confidence interval requires increasing the number of samples N. That is, the impacts of N is on both the "estimation accuracy" and "estimation confidence".

(b) Applying Hoeffding's inequality on random variable $\hat{p}$, we have

$$P(|\hat{p}-p| \geq \epsilon) \leq 2e^{-2N\epsilon^2} \Rightarrow \delta = 2e^{-2N\epsilon^2}, \epsilon = \sqrt{\frac{\ln(2/\delta)}{2N}}$$

The effects of δ and N are similarly discussed as in (a).

(c) Chebyshev's inequality (see Cantelli's inequality for the one-side improvement):

- Pros: 1) sharp bound and cannot be improved in general (given no extra assumption). 2) can be improved with extra distributional information on polynomial moments.
- Cons: 1) requires the existence of moments until the second order. 2) quadratic convergence rate.

Hoeffding's inequality (see Theorem 2.8 and 2.9 of paper "old and new concentration inequalities" for the one-side improvement):

- Pros: 1) exponential convergence rate. 2) does not require assumption on moments.
- Cons: 1) works only for sub-Gaussian (e.g., bounded random variables). 2) in general not sharp when the variance is small (e.g., see popoviciu's inequality on variances and Bernstein's inequality).

Problem 6

A coin with probability p of Heads is flipped repeatedly. For (a) and (b), suppose that p is a known constant, with $0 < p < 1$.

(a) What is the expected number of flips until the pattern HT is observed?

(b) What is the expected number of flips until the pattern HH is observed?

(c) What is the expected number of flips until the pattern HTH is observed?

(d) Now suppose that p is unknown, and that we use a Beta(a, b) prior to reflect our uncertainty about p (where a and b are known constants and are greater than 2). In terms of a and b, find the corresponding answers to $(a), (b)$ and (c) in this setting.

Solution

(a) This can be thought of as "Wait for the first Head, then wait for the first Tail afterwards," so the expected value is $\frac{1}{p} + \frac{1}{q}$ by the story of first success distributions, with $q = 1 - p$.

(b) Let X be the waiting time for HH and condition on the first toss, writing H for the occurrence of Head and T for the occurrence of Tail:

$$E[X] = E[X|\mathrm{H}]p + E[X|\mathrm{T}]q = E[X|\mathrm{H}]p + (1 + E[X])q.$$

To find $E[X|H]$, condition on the second toss:

$$E[X|\mathrm{H}] = E[X|\mathrm{HH}]p + E[X|\mathrm{HT}]q = 2p + (2 + E[X])q.$$

Solving for $E[X]$, we have

$$E[X] = \frac{1}{p} + \frac{1}{p^2}.$$

(c) Let Y be the waiting time for HTH. Since now we do not assume as a known constant probability p, we view it as an unknown constant instead and apply the first step analysis as in (b) with an extra condition. In fact, only given a fixed probability p, the sequence is independent across tosses, i.e., conditional independence.

For convenience, we denote $\widehat{E}[\cdot] = E[\cdot|p]$. Therefore, we have

$$\begin{aligned}
\widehat{E}[Y] &= \widehat{E}[Y|\mathrm{H}]p + \widehat{E}[Y|\mathrm{T}]q \\
&= p\Big(\widehat{E}[Y|\mathrm{HH}]p + \widehat{E}[Y|\mathrm{HT}]q\Big) + \Big(1 + \widehat{E}[Y]\Big)q \\
&= p\Big\{\Big(2 + \widehat{E}[Y]\Big)p + \Big(3p + \Big(3 + \widehat{E}[Y]\Big)q\Big)q\Big\} + \Big(1 + \widehat{E}[Y]\Big)q
\end{aligned}$$

Therefore, the final answer is

$$E[Y|p] = \frac{2}{p} + \frac{1}{p^2} + \frac{1}{1-p}.$$

Note that it makes no sense by taking outer expectation with respect to an unknown "constant", we thus only obtain a conditional expectation for this subproblem.

(d) Let Z_1, Z_2 and Z_3 be the number of flips until HT, HH and HTH, respectively. Therefore, we have

$$\begin{aligned}
E[Z_1] &= E\Big[E[Z_1|p]\Big] = E\left[\frac{1}{p}\right] + E\left[\frac{1}{1-p}\right] \\
E[Z_2] &= E\Big[E[Z_2|p]\Big] = E\left[\frac{1}{p}\right] + E\left[\frac{1}{p^2}\right] \\
E[Z_3] &= E\Big[E[Z_3|p]\Big] = E\left[\frac{2}{p}\right] + E\left[\frac{1}{p^2}\right] + E\left[\frac{1}{1-p}\right]
\end{aligned}$$

By LOTUS,

$$\begin{aligned}
E\left[\frac{1}{p}\right] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 p^{a-2}(1-p)^{b-1}dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a-1)\Gamma(b)}{\Gamma(a+b-1)} = \frac{a+b-1}{a-1},\\
E\left[\frac{1}{1-p}\right] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 p^{a-1}(1-p)^{b-2}dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a)\Gamma(b-1)}{\Gamma(a+b-1)} = \frac{a+b-1}{b-1},\\
E\left[\frac{1}{p^2}\right] &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 p^{a-3}(1-p)^{b-1}dp = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a-2)\Gamma(b)}{\Gamma(a+b-2)} = \frac{(a+b-1)(a+b-2)}{(a-1)(a-2)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
E[Z_1] &= \frac{a+b-1}{a-1} + \frac{a+b-1}{b-1},\\
E[Z_2] &= \frac{a+b-1}{a-1} + \frac{(a+b-1)(a+b-2)}{(a-1)(a-2)}\\
E[Z_3] &= \frac{2(a+b-1)}{a-1} + \frac{a+b-1}{b-1} + \frac{(a+b-1)(a+b-2)}{(a-1)(a-2)}
\end{aligned}$$

Probability & Statistics for EECS: Homework #12

Due on Dec 2, 2023 at 23:59

Name:
Student ID:

Problem 1

Given a coin with the probability p of landing heads. p is unknown and we need to estimate its value through data. In our data collection model, we have n independent tosses, result of each toss is either Head or Tail. Let X denote the number of heads in the total n tosses. Now we conduct experiments to collect data and find $X = k$. Then we need to find $\hat{p}$, the estimation of p.

(a) Assume p is an unknown constant. Find $\hat{p}$ through the MLE (Maximum Likelihood Estimation) rule.

(b) Assume p is a random variable with a prior distribution $p \sim \text{Beta}(a, b)$, where a and b are known constants. Find $\hat{p}$ through the MAP (Maximum a Posterior Probability) rule.

(c) Assume p is a random variable with a prior distribution $p \sim \text{Beta}(a, b)$, where a and b are known constants. Find $\hat{p}$ through the MMSE (Minimal Mean Squared Error) rule.

Solution

(a) Let X_i be the outcome of ith toss. Then $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \text{Bern}(p)$, where p is an unknown constant. The PMF of X_i can be formulated as

$$P_{X_i}(x_i; p) = p^{x_i}(1-p)^{1-x_i}$$

since

$$p^{x_i}(1-p)^{1-x_i} = \begin{cases} p, & \text{if } x_i = 1, \\ 1-p, & \text{if } x_i = 0. \end{cases}$$

The likelihood function is

$$P_X(x; p) = \prod_{i=1}^{n} P_{X_i}(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^k(1-p)^{n-k}$$

So the corresponding log-likelihood function is

$$g(p) = \log P_X(x; p) = \log p^{S_n}(1-p)^{n-S_n} = S_n \log p + (n - S_n)\log(1-p)$$

Now we try to find $\widehat{p}_{\text{MLE}}$ such that $g(\widehat{p}_{\text{MLE}})$ is the maximum of $g(p)$. We have

$$\begin{aligned} g'(p) &= \frac{k}{p} - \frac{n-k}{1-p}, \\ g''(p) &= -\frac{k}{p^2} - \frac{n-k}{(1-p)^2} \leq 0 \end{aligned}$$

Let $g'(p) = 0$, we can get $p = \frac{k}{n}$. Since $g''(p) \leq 0$, then we know that

$$\widehat{p}_{\text{MLE}} = \frac{k}{n}$$

is the MLE of p.

(b) We know the posterior distribution

$$f_{p|X=k} \propto p^{a+k-1}(1-p)^{b+n-k-1},\ p \in (0,1)$$

by Beta-Binomial conjugacy. Then the MAP estimator

$$\widehat{p}_{\text{MAP}} = \arg\max_p f_{\theta|X=k} = \arg\max_p \log(f_{p|X=k})$$

Problem 1 continued on next page...

since logarithmic function is monotonically increasing. Let

$$g(p) = \log(f_{p|X=k}) = (a+k-1)\log p + (b+n-k-1)\log(1-p),$$

where we don't consider the proportional constant. Our goal is to find p^* such that $g(p^*)$ is maximum of $g(p)$. We have

$$\begin{aligned} g'(p) &= \frac{a+k-1}{p} - \frac{b+n-k-1}{1-p}, \\ g''(p) &= -\frac{a+k+1}{p^2} - \frac{b+n-k-1}{(1-p)^2} < 0. \end{aligned}$$

Let $g'(p^*) = 0$. We have $p^* = \frac{a+k-1}{a+b+n-2}$, and $g(p^*)$ is maximum of $g(p)$ since $g''(p) < 0$.

Then we can get the MAP estimate

$$\widehat{p}_{\text{MAP}} = \arg\max_p f_{p|X=k} = \arg\max_p \log(f_{\theta|X=k}) = p^* = \frac{a+k-1}{a+b+n-2}.$$

(c) Since the prior distribution is $p \sim \text{Beta}(a,b)$ and the conditional distribution of X given p is $X|p \sim \text{Bin}(n,p)$, we can get the posterior distribution

$$\Theta|X = k \sim \text{Beta}(a+k, b+n-k)$$

by Beta-Binomial conjugacy. It follows that

$$E(p|X=k) = \frac{a+k}{a+b+n},$$

so the MMSE estimation of Θ is

$$\widehat{p}_{\text{MMSE}} = E(p|X=k) = \frac{a+k}{a+b+n}.$$

Problem 2

Let X be the height of a randomly chosen adult man, and Y be his father's height, where X and Y have been standardized to have mean 0 and standard deviation 1. Suppose that (X, Y) is Bivariate Normal, with $X, Y \sim \mathcal{N}(0,1)$ and $\text{Corr}(X,Y) = \rho$.

(a) Let $y = ax + b$ be the equation of the best line for predicting Y from X (in the sense of minimizing the mean squared error), e.g., if we were to observe $X = 1.3$ then we would predict that Y is $1.3a + b$. Now suppose that we want to use Y to predict X, rather than using X to predict Y. Give and explain an intuitive guess for what the slope is of the best line for predicting X from Y.

(b) Find a constant c (in terms of ρ) and an r.v. V such that $Y = cX + V$, with V independent of X. Hint: Start by finding c such that $\text{Cov}(X, Y - cX) = 0$.

(c) Find a constant d (in terms of ρ) and an r.v. W such that $X = dY + W$, with W independent of Y.

(d) Find $E(Y \mid X)$ and $E(X \mid Y)$.

(e) Reconcile (a) and (d), giving a clear and correct intuitive explanation.

Solution

(a) Since the parameter ρ tells us what is the rate of change of second variable respective to the first one, we can assume that ρ is the slope of the line, i.e. $a = \rho$. Now, in order to predict X from Y, we just have to consider the line that is inverse to the original line. From the basic algebra, we know that inverse has slope one over the original slope. Thus, the required slope is $\frac{1}{\rho}$.

(b) Since we have to find $V = Y - cX$ such that is independent from X, using the given hint, we have that

$$0 = \text{Cov}(X, Y - cX) = \text{Cov}(X,Y) - c\,\text{Var}(X) = \rho - c.$$

Hence, let's define $c = \rho$ and it is the only candidate for the constant c.

Let's check that X and V are independent. Observe that $Y - \rho X$ is also Normal (as the linear combination of two Bivariate Normals). So, the fact that two Normals that construct Bivariate Normal are independent is equivalent to the fact that they are uncorrelated. Since we have the last information, we have found the required.

(c) With the same calculation and discussion as in part (b), we have that the answer is also $d = \rho$.

(d) Using the definition of conditional density function, we have that

$$\begin{aligned} f_{Y|X}(y \mid x) = &\frac{f(x,y)}{f(x)} = \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{1}{2(1-\rho^2)}\left(x^2 + y^2 - 2xy\rho\right)\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)} \\ = &\frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}} \exp\left(-\frac{x^2 + y^2 - 2xy\rho}{2\left(1-\rho^2\right)} + \frac{x^2}{2}\right) \\ = &\frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}} \exp\left(-\frac{(y-\rho x)^2}{2\left(1-\rho^2\right)}\right). \end{aligned}$$

Now, we see that

$$Y \mid X = x \sim \mathcal{N}\left(\rho x, 1 - \rho^2\right).$$

Hence, $E(Y \mid X) = \rho X$. Because of the symmetry, we also have that $E(X \mid Y) = \rho Y$.

Problem 2 continued on next page...

(e) Since we know that means of X and Y are zero, we have that

$$X = \alpha Y$$

for some α. Applying the conditional expectation $E(\cdot \mid X)$ to the both sides, we have that

$$X = \alpha E(Y \mid X) = \alpha \rho X.$$

Because of the fact that $X \neq 0$ almost certainly, we can conclude that $\alpha = \frac{1}{\rho}$. Hence, we have proved the claimed.

Problem 3

Two chess players, Vishy and Magnus, play a series of games. Given p, the game results are i.i.d. with probability p of Vishy winning, and probability $q = 1-p$ of Magnus winning (assume that each game ends in a win for one of the two players). But p is unknown, so we will treat it as an r.v. To reflect our uncertainty about p, we use the prior $p \sim \text{Beta}(a, b)$, where a and b are known positive integers and $a \geq 2$.

(a) Find the expected number of games needed in order for Vishy to win a game (including the win). Simplify fully; your final answer should not use factorials or Γ.

(b) Explain in terms of independence vs. conditional independence the direction of the inequality between the answer to (a) and $1 + E(G)$ for $G \sim \text{Geom}\left(\frac{a}{a+b}\right)$.

(c) Find the conditional distribution of p given that Vishy wins exactly 7 out of the first 10 games.

Solution

(a) Denote N as the number of games needed for Vishy to win the game for one time, then there is $N|p \sim \text{FS}(p)$. Via Adam's law, we have:

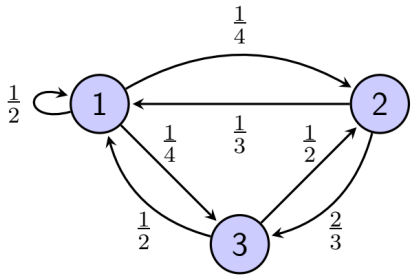
$$\begin{aligned}
\text{E}(N) &= \text{E}(\text{E}(N|p)) \\
&= \text{E}\left(\frac{1}{p}\right) \\
&= \int_0^1 \frac{1}{\beta(a,b)} \frac{1}{p} p^{a-1} (1-p)^{b-1} \, dp \\
&= \frac{\beta(a-1,b)}{\beta(a,b)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a-1)\Gamma(b)}{\Gamma(a+b-1)} \\
&= \frac{a+b-1}{a-1}
\end{aligned}$$

(b) $1 + E(G) = \frac{a+b}{a} < \frac{a+b-1}{a-1} = E(N)$ means that the games are conditionally independent given p while not dependent between each other. Since $1 + E(G)$ can been seen as the expectation of the number of trials for Vishy to win the first game given $p = \frac{a}{a+b}$. While p is unknown, and each time of Vishy's loss will decrease the probability of wining the game. Thus the expectation estimated by conditional probability is larger than the expectation given by prior distribution. Therefore $1 + E(G) < E(N)$.

(c) Via Beta-Binomial conjugacy, the conditional distribution of p give that Vish wins exactly 7 out out of the first 10 games is $\text{Beta}(a + 7, b + 3)$.

Problem 4

Given a Markov chain with state-transition diagram shown as follows:



(a) Is this chain irreducible?

(b) Is this chain aperiodic?

(c) Find the stationary distribution of this chain.

(d) Is this chain reversible?

Solution

The transition matrix of the Markov chain is

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

(a) Yes, because the elements in the matrix are $Q_{1,2}, Q_{2,1}, Q_{1,3}, Q_{3,1}, Q_{2,3}, Q_{3,2}$ are all non-zero.

(b) Yes, the diagram we know that 1 is a possible return time for state 1, thus $d(1) = 1$ since both $2, 3$ are possible for state 2 and 3, $d(2) = d(3) = 1$ because the chain is irreducible and $d(1) = d(2) = d(3) = 1$. Therefore the chain is aperiodic.

(c) Denote π as the stationary distribution for the chain, then there is $\pi Q = \pi$. Then we solve the problem $\pi(Q - I) = 0$, as follows:

$$\begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & -1 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix} = 0$$

where $\sum_{i=1}^{3} \pi_i$. The solution is $\pi = (\frac{16}{35}, \frac{9}{35}, \frac{2}{7})$.

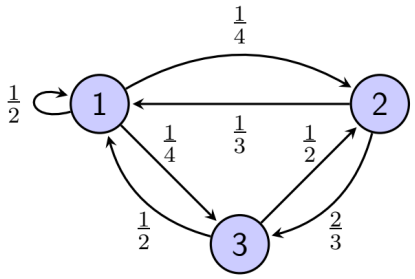
(d) No. If the chain is reversible, there exists a distribution π which satisfy:

$$\begin{cases} \pi_1 \cdot \frac{1}{4} = \pi_2 \cdot \frac{1}{3} \\ \pi_1 \cdot \frac{1}{4} = \pi_3 \cdot \frac{1}{2} \\ \pi_2 \cdot \frac{2}{3} = \pi_3 \cdot \frac{1}{2} \end{cases}$$

The solution of the above problem is $\pi_1 = \pi_2 = \pi_3 = 0$, which cannot satisfy the constraint $\sum \pi = 1$.

Problem 5

Given a Markov chain with state-transition diagram shown as follows:



(a) Find $P(X_3 = 3 \mid X_2 = 2)$ and $P(X_4 = 1 \mid X_3 = 2)$.

(b) If $P(X_0 = 2) = \frac{2}{5}$, find $P(X_0 = 2, X_1 = 3, X_2 = 1)$.

(c) Find $P(X_2 = 1 \mid X_0 = 2), P(X_2 = 2 \mid X_0 = 2)$, and $P(X_2 = 3 \mid X_0 = 2)$.

(d) Find $E(X_2 \mid X_0 = 2)$.

Solution

The transition matrix of the Markov chain is

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

(a) From the state-transition diagram, we have

$$\begin{aligned} P(X_3 = 3|X_2 = 2) &= \frac{2}{3} \\ P(X_4 = 1|X_3 = 2) &= \frac{1}{3} \end{aligned} \tag{1}$$

(b)

$$\begin{aligned} P(X_0 = 2, X_1 = 3, X_2 = 1) &= P(X_1 = 3, X_2 = 1|X_0 = 2)P(X_0 = 2) \\ &= P(X_2 = 1|X_1 = 3, X_0 = 2)P(X_1 = 3|X_0 = 2)P(X_0 = 2) \\ &= P(X_2 = 1|X_1 = 3)P(X_1 = 3|X_0 = 2)P(X_0 = 2) \\ &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{5} = \frac{2}{15} \end{aligned} \tag{2}$$

Problem 5 continued on next page...

(c)

$$
\begin{aligned}
&P(X_2 = 1|X_0 = 2) = \sum_{i=1}^{3} P(X_2 = 1|X_1 = i, X_0 = 2)P(X_1 = i|X_0 = 2) \\
&= \sum_{i=1}^{3} P(X_2 = 1|X_1 = i)P(X_1 = i|X_0 = 2) \\
&= \frac{1}{6} + 0 + \frac{2}{6} = \frac{1}{2} \\
&P(X_2 = 2|X_0 = 2) = \sum_{i=1}^{3} P\big(X_2 = 2\big|X_1 = i, X_0 = 2\big)P\big(X_1 = i\big|X_0 = 2\big) \\
&= \sum_{i=1}^{3} P(X_2 = 2|X_1 = i)P(X_1 = i|X_0 = 2) \\
&= \frac{1}{12} + 0 + \frac{2}{6} = \frac{5}{12} \\
&P(X_2 = 3|X_0 = 2) = \sum_{i=1}^{3} P(X_2 = 3|X_1 = i, X_0 = 2)P(X_1 = i|X_0 = 2) \\
&= \sum_{i=1}^{3} P(X_2 = 3|X_1 = i)P(X_1 = i|X_0 = 2) \\
&= \frac{1}{12} + 0 + 0 = \frac{1}{12}.
\end{aligned}
\tag{3}
$$

(d) The expectation is

$$
\mathrm{E}(X_2|X_0 = 2) = \sum_{i=1}^{3} iP(X_2 = i|X_0 = 2) = \frac{1}{2} + \frac{10}{12} + \frac{3}{12} = \frac{19}{12}.
$$

Problem 6

A fair coin is flipped repeatedly. We use H to denote "Head appeared" and T to denote the "Tail appeared".

(a) What is the expected number of flips until the pattern HTHT is observed?

(b) What is the expected number of flips until the pattern THTT is observed?

(c) What is the probability that pattern HTHT is observed earlier than THTT?

Solution

(a) Denote $t(\cdot)$ as the time for transferring from the current state to the ending state. The state space is $\{H, T, HT, HTH, HTHT\}$, and the transition relationship between is can be demonstrated as follows:

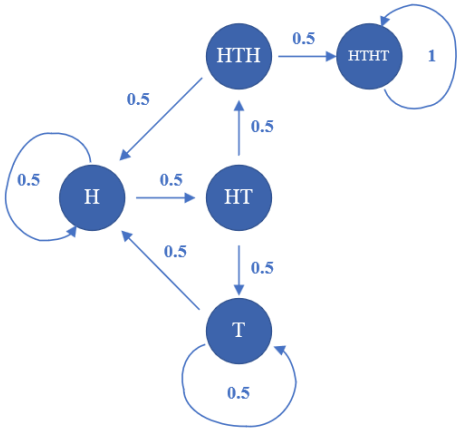


Figure 1: 6(1)

Then the expectation of step numbers for transferring from one state to the ending state can be listed as follows:

$$\begin{cases} E(t(H)) = \frac{1}{2}E(t(H)) + \frac{1}{2}E(t(HT)) + 1 \\ E(t(T)) = \frac{1}{2}E(t(T)) + \frac{1}{2}E(t(H)) + 1 \\ E(t(HT)) = \frac{1}{2}E(t(T)) + \frac{1}{2}E(t(HTH)) + 1 \\ E(t(HTH)) = \frac{1}{2}E(t(HTHT)) + \frac{1}{2}E(t(H)) + 1 \\ E(t(HTHT)) = 0 \end{cases}$$

Then we can obtain that $E(t(H)) = 18$ and $E(t(T)) = 20$. Therefore, the expected numbers of flips from starting is $\frac{1}{2}E(H) + \frac{1}{2}E(T) + 1 = 20$.

(b) Similarly, The state space is $\{H, T, HT, HTH, HTHT\}$, and the transition relationship between is can be demonstrated as follows: Then the expectation of step numbers for transferring from one state to the ending state can be listed as follows:

$$\begin{cases} E(t(T)) = \frac{1}{2}E(t(T)) + \frac{1}{2}E(t(TH)) + 1 \\ E(t(TH)) = \frac{1}{2}E(t(H)) + \frac{1}{2}E(t(THT)) + 1 \\ E(t(H)) = \frac{1}{2}E(t(H)) + \frac{1}{2}E(t(T)) + 1 \\ E(t(THT)) = \frac{1}{2}E(t(TH)) + \frac{1}{2}E(t(THTT)) + 1 \\ E(t(THTT)) = 0 \end{cases} \tag{4}$$

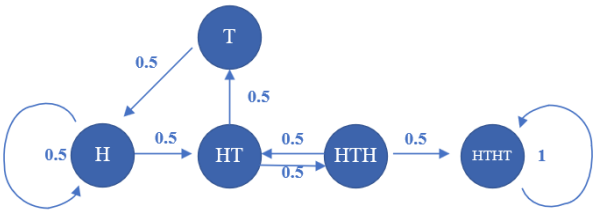


Figure 2: 6(2)

Then we can obtain that $E(t(H)) = 18$ and $E(t(T)) = 16$. Therefore, the expected numbers of flips from starting is $\frac{1}{2}E(H) + \frac{1}{2}E(T) + 1 = 18$.

(c) The state space in this problem can be conclude by $\{H, HT, HTH, HTHT, T, TH, THT, THTT\}$, and the relationship between each state can be demonstrated as follows:

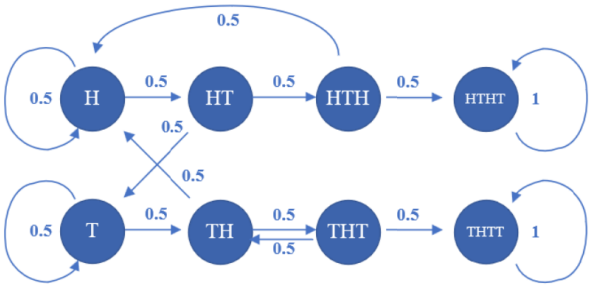


Figure 3: 6(3)

The relationship between the probability of each state finally ends up in $HTHT$ can be listed as follows:

$$\begin{cases} P(HTHT) = 1 \\ P(THTT) = 0 \\ P(HTH) = \frac{1}{2}P(HTHT) + \frac{1}{2}P(H) \\ P(THT) = \frac{1}{2}P(THTT) + \frac{1}{2}P(TH) \\ P(HT) = \frac{1}{2}P(HTH) + \frac{1}{2}P(T) \\ P(TH) = \frac{1}{2}P(THT) + \frac{1}{2}P(H) \\ P(H) = \frac{1}{2}P(H) + \frac{1}{2}P(HT) \\ P(T) = \frac{1}{2}P(T) + \frac{1}{2}P(TH). \end{cases} \tag{5}$$

Finally we can get that $P(H) = \frac{5}{7}$ and $P(T) = \frac{4}{7}$. Thus the probability of pattern $HTHT$ observed earlier than $THTT$ is $\frac{1}{2}P(H) + \frac{1}{2}P(T) = \frac{9}{14}$ by assuming the same initial state of entering states H and T.