ECS 132 Spring 2024: Assignment 1
Solution

1 Problem

Consider the Manhattan grid network shown in Figure 1. Suppose that starting at the point labelled A, you can go one step up (denoted u) or one step to the right (denoted r) at each move. This procedure is is continued until the point labelled B is reached. For instance, the path along the upper left corner would be denoted (u u u r r r r). How many different paths from A to B are possible?

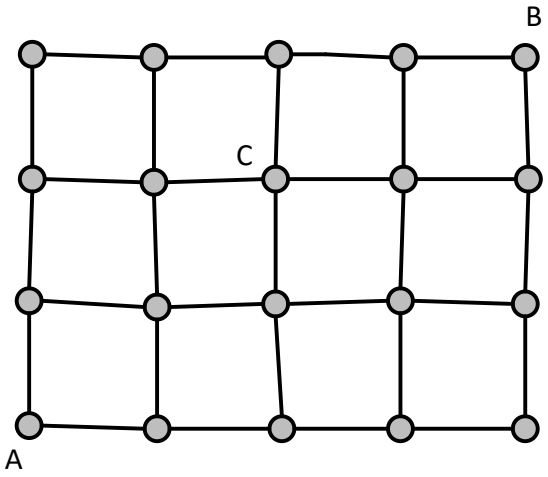


Figure 1: The Manhattan grid network.

Answer

Each path from A to B is a permutation of 7 steps consisting of 4 r's (rights) and 3 u's (ups). So there are 7 positions and we choose 4 of them to be r's and the remaining 3 to be u's. Once the r's are selected, the u's are automatically selected. The total number of ways is $\binom{7}{4} = \binom{7}{3}$. Note that order is not relevant because the r's are indistinguishable from one another as are the u's. For example, if the r's are labelled r_1, r_2, r_3, r_4, and the u's are labeled u_1, u_2, u_3 then $(r_1u_1r_2u_2r_3u_3r_4) = (r_2u_1r_3u_3r_4u_2r_1)$. Also, note that $\binom{n}{k}$ is the same as $\binom{n}{n-k}$.

2 Problem

Consider a pool of six I/O buffers. Assume that any buffer is just as likely to be available (or occupied) as any other. Compute the probabilities associated with the following events

1. A = "A least two but no more than five buffers are occupied"
2. B = "At least one buffer is occupied"

Answer

There are 6 I/O buffers and each o f them can be in one of two states available (0) and occupied (1). The state space S consist of 2^6 possible outcomes of the I/O buffer pool. For example $(0, 0, 0, 0, 0, 0)$ corresponds to the outcome that all 6 I/O buffers are available. Similarly, $(0, 1, 0, 0, 0, 1)$ correspond to the outcome that 2 I/O buffers (specifically 2 and 6 in this case) are occupied.

1. For event A we are looking for the outcomes in which 2, 3, 4, and 5 I/O buffers are occupied. The number of outcomes that have i I/O buffers occupied is $\binom{6}{i}$. Thus,

$$P(A) = \frac{\binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5}}{2^6} \tag{1}$$

2. For the event B, we can follow the same logic as before and

$$P(B) = \frac{\binom{6}{1} + \binom{6}{2} + \binom{6}{3} + \binom{6}{4} + \binom{6}{5} + +\binom{6}{6}}{2^6} \tag{2}$$

   The other way think about is to consider the complement of event B denoted as $B^{\complement}$ which is the event that 0 I/O buffers are available:

$$\begin{aligned} P(B) &= 1 - P(B^{\complement}) && (3)\\ &= 1 - \frac{\binom{6}{0}}{2^6} && (4)\\ &= 1 - \frac{1}{2^6} && (5) \end{aligned}$$

3 Problem

Suppose your MP3 player contains 100 songs. 10 of those songs are by your favorite group (say Cake, which is a band from Sacramento). Suppose the shuffle feature is used to play the songs in random order. Note that after a shuffle all the 100 songs are played before the next shuffle.

1. What is the probability that the first Cake song heard is the 5th song played?
2. Similar to the **simulation** of the Birthday paradox that we discussed in class, write an R code to simulate the problem and determine the probability that the first Cake song heard is the 5th song played.

Answer

This is a sampling problem. Always ask yourself these two questions: (i) does order matter or not matter? (ii) is sampling with our without replacement. Here the answers are (i) yes order matters, (ii) sampling is without replacement. So we will use a permutation ${}^nP_k$.

1. What is the probability that the first Cake song heard is the 5th song played? The number of ways in which the first 5 songs can be played is $100 \times 99 \times 98 \times 97 \times 96 = {}^{100}P_5$. The number of ways in which the first 4 songs are non-Cake songs is $90 \times 89 \times 88 \times 87 = {}^{90}P_4$. And the number of ways in which the 5th song is a Cake song is 10. Thus, P(5th song is Cake) is given by

$$\begin{aligned} P(\text{5th song is Cake}) &= \frac{90 \times 89 \times 88 \times 87 \times 10}{100 \times 99 \times 98 \times 97 \times 96} \\ &= \frac{10 \times {}^{90}P_4}{{}^{100}P_5} \\ &= .0679 \end{aligned}$$

2.
```
# m is the number of trials
# n is total number of songs.
# We can assume that songs 1 through 10 are Cake songs,
# without loss of generality (WLOG).
# vector x records the outcomes.
# x(i)=1 if during trial i the 5th song is the 1st Cake song

 m = 100000
 n = 100
 x = numeric(m)
  for (i in 1:m)
 {
 b = sample(1:100, n, repl=F) # a random shuffle of 100 songs

 if ((b[1] > 10) & (b[2] > 10) & (b[3] > 10) & (b[4] >10) & (b[5] <= 10)) {
     x[i] = 1
     }
 else {
     x[i] = 0
     }
```

```
}
pmean = mean(x == 1)
print(pmean) # We take advantage of the mean function as a shorthand
             # to calculate number of times x(i)=1 divided by m.
```

4 Problem

Suppose that you want to backup the 100 songs on your MP3 player onto 4 different external drives each of different size. The first drive can store 20 songs, the second can store 30 songs, the third can store 40 songs, and the forth can store 10 songs. How many different ways can you arrange the songs on the four external drives.

Answer

This is an example of a multinomial coefficient. A set of n distinct objects is to be divided into r distinct groups of sizes $n_1, n_2, \ldots n_r$ such that $\sum_{i=1}^{r} n_i = n$. The total number of ways is given by $T = \frac{n!}{n_1! n_2! \ldots n_r!}$. Applying this here:

$$T = \frac{100!}{20!\ 30!\ 40!\ 10!}$$

5 Problem

Consider two specific memory locations that are within the address space of a computer program when it executes. With probability 0.5, the program will access the first location; with probability 0.4 it will access the second location and with probability 0.3 it will access both locations. What is the probability that the program will access neither location?

Answer

Let B_i denote the event that the computer program access location $i, i = 1, 2$. Then the probability that the program accesses at least one of the locations is given by

$$\begin{aligned} P(B_1 \cup B_2) &= P(B_1) + P(B_2) - P(B_1 B_2) && (6)\\ &= 0.5 + 0.4 - 0.3 && (7)\\ &= 0.6 && (8) \end{aligned}$$

The event that the computer program will access neither location is

$$\begin{aligned} P(B_1^{\complement} \cap B_2^{\complement}) &= P((B_1 \cup B_2)^{\complement}) && (9)\\ &= 1 - P(B_1 \cup B_2) && (10)\\ &= 1 - 0.6 && (11)\\ &= 0.4 && (12) \end{aligned}$$

6 Problem

John is 31 years old, single, outspoken, and has multiple talents. He majored in Computer Science. As a student, he was deeply concerned with issues of discrimination and social justice, and environment. Which of the following scenarios is more probable? **A**) John is a computer programmer. **B**) John is a computer programmer and is an activist in the environmental movement. Give a reason for your answer.

Answer

The event that John is both a computer programmer and environmental activist is a subset of the event that John is a computer programmer. Thus, it cannot be more likely than the event that John is a computer programmer.

If event X is a subset of event Y then

$$P(X) \leq P(Y)$$

Here, $P(X) < P(Y)$ since there is some non-zero probability that John is a computer programmer but not active in the environmental movement.

A similar problem was made famous by Amos Tversky and Daniel Kahneman, whose 1983 study showed that 85% of respondents erroneously thought it was more likely that John is a computer programmer and is an activist in the environmental movement. These are related to mental/implicit biases that can/should be avoided and probability provides a framework for analyzing and avoiding them `brilliant.org`. Note Kahneman won the 2002 Nobel Prize in Economics, in part for his work establishing Prospect Theory and he is well known for his book "Thinking fast and slow" published in 2011. Wikipedia has quite a lot of information on this.

ECS 132 Spring 2024: Assignment 2
Solution

1 Problem

Consider a universe where it is equally likely that a child is born with brown or green eyes. A couple has two children.

1. What is the probability that both have green eyes?
2. What is the probability that both have green eyes if the older of the two has green eyes?

Solution

1. The sample space $S = \{(B, G), (B, B), (G, B), (G, G)\}$ where the first element of the tuple is the elder child. There is only one state with both G. So $P(\text{both green}) = 1/4$.
2. We have the same initial sample space, but given the information that the older of the two child is a girl, the constrained sample space becomes $S_c = \{(G, B), (G, G)\}$. One of out the two states has both G, so the probability $P(\text{both green}) = 1/2$.

2 Problem

A person has n keys, of which only one will open her door.

1. If she tries the keys at random, discarding those that don't work, what is the probability that she will open the door on her kth try?
2. If she does not discard previously tried keys what is the probability that she will open the door on her kth try?

Solution

1. On the first try, the probability of getting the right key is $\frac{1}{n}$ and of not getting the right key is $\frac{n-1}{n}$. On the second try, there are $n-1$ remaining keys so the probability of getting the right key is $\frac{1}{n-1}$ and of not getting the right key is $\frac{n-2}{n-1}$. On the kth try there are $n-k+1$ remaining keys. Let X denote the random variable which is the number of attempts to get the right key. Hence,

$$P\{X = k\} = \left(\frac{n-1}{n}\right)\left(\frac{n-2}{n-1}\right)\cdots\left(\frac{n-k+1}{n-k+2}\right)\left(\frac{1}{n-k+1}\right)$$

2. What if she does not discard previously tried keys?

   On each try the probability that she will get the right key is $1/n$. Let X be the random variable which denotes the number of attempts needed to get the right key. The probability that she will get the right key on the kth try implies that she did not get the right key on the first $k-1$ attempts. Hence,

$$P\{X = k\} = \frac{1}{n}\left(\frac{n-1}{n}\right)^{k-1} \quad \text{where } k = 1, 2, \ldots$$

3 Problem

Consider that three types of requests arrive at an IT help desk: urgent (colored red), important (colored green), and routine (colored black). What is the probability that the IT specialist receives 10 red requests among the first 20 requests received on a particular day?

Solution

Treat this as a binomial process. Filling each slot is a Bernoulli process where the probability of "success" (getting red) is 1/3. There are 20 independent slots to fill, and we want to know what is the probability that 10 of them are red. Considering how many independent Bernoulli slots are successful is a Binomial process with $p = 1/3$ and $n = 20$ evaluated at $k = 10$.

$$P(10 \text{ reds}) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{20}{10}\left(\frac{1}{3}\right)^{10}\left(\frac{2}{3}\right)^{10} = 0.054259$$

4 Problem

Bridge is a card game in which the deck of 52 cards is randomly and equally divided among the 4 players. So each player gets 13 cards. What is the probability that one player receives all the 13 hearts?

As an aside, did you know that one of the greatest bridge players in the world is a (retired) faculty of the Computer Science Department here at UCDavis? Yes! Prof. Charles U. (Chip) Martel who retired a few years ago but still does research (https://web.cs.ucdavis.edu/~martel/main/ has been the world champion many times[1].

Solution

Let H_i denote the event that hand i has all 13 hearts, $i = 1, 2, 3, 4$. There is only one way for that hand to arise. The number of possible distinct 5 card hands is $\binom{52}{13}$. So $P(H_i)$ is given by

$$P(H_i) = \frac{1}{\binom{52}{13}} \quad i = 1, 2, 3, 4$$

Note there are four players and any one of them could get the hand. But if player 1 gets the had with all hearts, then no other play can get such a hand, so the events $H_i, i = 1, 2, 3, 4$ are mutually exclusive. We

[1] https://alum.mit.edu/slice/meet-chip-martel-75-one-worlds-greatest-bridge-players

need to calculate $P(\cup_{i=1}^{4} H_i)$, the probability that any of the four players received the hand.

$$\begin{aligned} P(\cup_{i=1}^{4} H_i) &= \sum_{i=1}^{4} P(H_i) \\ &= 4 * \frac{1}{\binom{52}{13}} \\ &= \sim 6.3 \times 10^{-12} \end{aligned}$$

5 Problem

In class we looked at an example of wireless sensor networks where each sensor can be active or asleep during each time slot of a timeframe. Here we will consider 2 sensors and 3 time slots. Each sensor is active with probability 1/2 in a given slot, and the sensors are independent of one another. Let A_i denote the event that both sensors are active in slot i.

1. What is $p(A_i)$, the probability that both sensors are active during the i-th time slot?
2. What is $p(A_i \cap A_j)$, for $i \neq j$? This is the probability that both sensors are active during two time slots.
3. What is $p(A_1 \cap A_2 \cap A_3)$, the probability that both sensors are active during all three time slots?
4. What is $p(A_1 \cup A_2 \cup A_3)$, the probability that they are both active in at least one time slot? Give the mathematical expression and then the numerical answer.

Solution

1. For a given slot i there are two bits present: one for the first sensor, and one for the second sensor. The state space $S = \{(0,0),(0,1),(1,0),(1,1)\}$. There is only one state with both sensors on so $P(A_i) = 1/4$.
2. Since events A_i and A_j are independent for $i \neq j$, then $p(A_i \cap A_j) = p(A_i)\, p(A_j) = (1/4)^2 = 1/16$. (Note we say $i \neq j$ for mathematical correctness since if $i = j$, then $A_i \cap A_i = A_i$.)
3. Since events A_1, A_2, A_3 are independent, $p(A_1 \cap A_2 \cap A_3) = p(A_1)\, p(A_2)\, p(A_3) = (1/4)^3 = 1/64$.
4. The events are **independent so they are not disjoint** and we use the inclusion-exclusion principle:

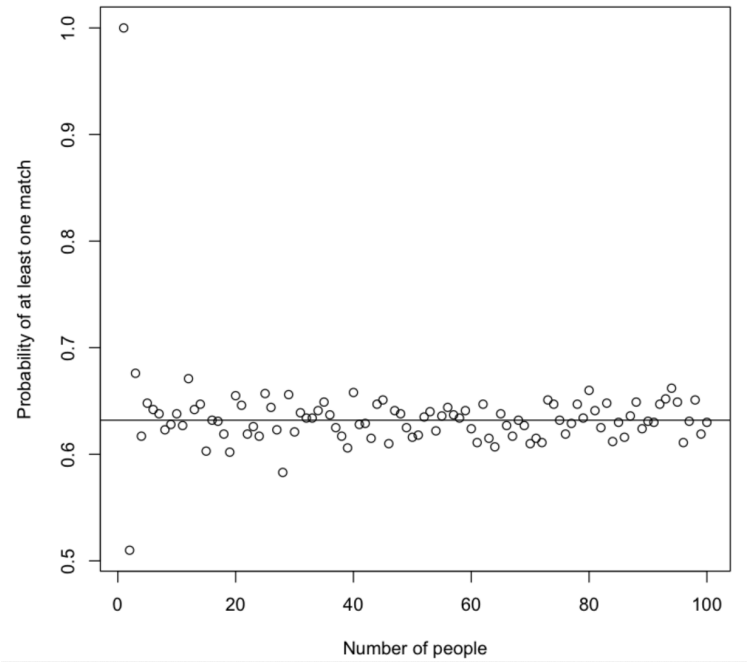
$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) && (1) \\ &\quad -P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) && (2) \\ &\quad +P(A_1 \cap A_2 \cap A_3) && (3) \\ &= 3\left(\frac{1}{4}\right) - 3\left(\frac{1}{16}\right) + \left(\frac{1}{64}\right) = 0.578. && (4) \end{aligned}$$

6 Problem

Write a simulation of the (hat) matching problem discussed in class. The simulation should calculate the probability of at least one match for different values of n (the number of people). For each value of n do 1000 experiments to calculate the probability. Plot the probability as a function of n. In the same plot, add a line showing the asymptotic value, $1 - \frac{1}{e}$. Notice how quickly the asymptotic value is reached.

Solution

```
# m is number of experiments
# n is number of people in room
# x is the vector of outcomes

 m = 1000
 n = 100
 x = numeric(m)
 p = numeric(n)

 for (k in 1:n)
     {
         for (i in 1:m)
         {
             matches = 0                  # initialize the match count
             b = sample(1:k, k, repl=F)  # n random permutations of 1 to k
             for (j in 1:k)
             {
                 if (b[j] == j) {
                  matches = matches +  1
                 }
             }
             if (matches > 0)
             {
                 x[i] = 1
             }

         }
         p[k] = mean(x==1)      # probability of at least 2 match
         x = numeric(m)
     }

 plot(p, ylab = "Probability of at least one match", xlab = "Number of people")
 abline(h = 0.632)
```


Probability of at least one match
1.0
0.9
0.8
0.7
0.6
0.5
0
20
40
60
80
100
Number of people

ECS 132 Spring 2024: Assignment 3
Solution

This problem set covers:

- Conditional probability, problems 1-2.
  Problem 1 should be straightforward and 2 is more involved.
- Discrete random variables and the Bernoulli, Binomial and Geometric distributions, problems 3-5.
  Problem 3 should be very simple, and 4 and 5 more involved.

1 Problem

Suppose the university has designed a e-mail spam filter that attempts to identify spam by looking for commonly occurring phrases in spam. E-mail analysis has shown that 80% of email is spam. Suppose that 10% of the spam email contain the phrase "Large inheritance", whereas this phrase is only used in 1% of non-spam emails. Suppose a new email is received containing the phrase "Large inheritance", what is the probability that it is spam?

Answer

We define the following 2 events

- S: event that an e-mail is spam
- F: event that the email has the phrase "Large inheritance"

Using Bayes' Rule we find

$$\begin{aligned} P(S|F) &= \frac{P(F|S)P(S)}{P(F)} \\ &= \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|S^c)P(S^c)} \\ &= \frac{0.1 \times 0.8}{0.1 \times 0.8 + 0.01 \times 0.2} \\ &\approx 0.9756 \end{aligned}$$

2 Problem

A binary communication channel carries data as one of two types of signals denoted by 0 and 1. Owing to noise, a transmitted 0 is sometimes received as a 1 and a transmitted 1 is sometimes received as a 0. For a given channel, assume a probability of 0.94 that a transmitted 0 is correctly received as a 0 and a probability 0.91 that a transmitted 1 is received as a 1. Further assume a probability of 0.45 of transmitting a 0. If a signal is sent, determine

1. Probability that a 1 is received.
2. Probability that a 0 is received.
3. Probability that a 1 was transmitted given that a 1 was received.
4. Probability that a 0 was transmitted given that a 0 was received.
5. Probability of an error (this means either a transmitted 0 was received as a 1, or a transmitted 1 was received as a 0).

Answer

Define the following events:

- T_0 : A 0 is transmitted
- R_0 : A 0 is received
- $T_1 : \overline{T_0}$
- $R_1 : \overline{R_0}$

Then the statements given in the problem formulation can be translated to the following:

- $P(R_0|T_0) = 0.94$
- $P(R_1|T_1) = 0.91$
- $P(T_0) = 0.45$

From those statements we can directly deduce:

- $P(R_1|T_0) = 0.06$
- $P(R_0|T_1) = 0.09$
- $P(T_1) = 0.55$

This provides all the information we need.

1) The law of total probability tells us that:

$$P(R_1) = P(R_1|T_0)P(T_0) + P(R_1|T_1)P(T_1) = (0.06)(0.45) + (0.91)(0.55) = 0.5275$$

2) The law of total probability tells us that:

$$P(R_0) = P(R_0|T_0)P(T_0) + P(R_0|T_1)P(T_1) = (0.94)(0.45) + (0.09)(0.55) = 0.4725$$

3)

$$P(T_1|R_1) = \frac{P(T_1)P(R_1|T_1)}{P(R_1)} = \frac{(0.55)(0.91)}{0.5275} = 0.949$$

4)

$$P(T_0|R_0) = \frac{P(T_0)P(R_0|T_0)}{P(R_0)} = \frac{(0.45)(0.94)}{0.4725} = 0.895$$

5) Probability of an error:

$$\begin{aligned} P(\text{ Error}) &= P(R_0 \cap T_1) + P(R_1 \cap T_0) \\ &= P(R_0|T_1)P(T_1) + P(R_1|T_0)P(T_0) \\ &= (0.09)(0.55) + (0.06)(0.55) = 0.0765 \end{aligned}$$

3 Problem

You have two sensors who can be active or asleep in 3 different time slots. Each sensor is independent and active with probability p=1/2 in each slot, and the slots are independent of one another. Consider the following:

- Event A_i = both sensors are active in slot i.
- Events A_i and A_j are independent if $i \neq j$.
- Event B = the two sensors are active in all three time slots.
- The random variable $X = 1$ if event B occurs, and $X = 0$ otherwise.

Answer these questions:

1. What is $P(X = 1)$?
2. To what family of distributions does the distribution of X belong? Provide this answer as X $\sim$ distribution_name(parameters). See the lecture notes for examples.
3. What is the value of $E(X)$?
4. What is the value of $Var(X)$?

Answer:

1. $X = 1$ corresponds to the event that both sensors are active in all three time slots. Recall HW2, Problem 4 where we derived this probability. We start by considering the probability that both sensors are active in a slot i which is $P(A_i) = 1/4$ since $S = \{(0,0),(0,1),(1,0),(1,1)\}$ and only one of those 4 states corresponds to both sensors on. The probability the sensors are both active in three **independent** slots is $P(A_1)\ P(A_2)\ P(A_3) = (1/4)^3 = (1/64)$. So

$$P(X = 1) = 1/64$$

2. This is a single trial with one of two outcomes: either they are both on in three slots or they are not. The former (which we call "success") happens with $p = 1/64$. So

$$X \sim \text{Bern}(p = 1/64)$$

3. For a Bernoulli distribution $E(X) = p = \frac{1}{64}$.
4. For a Bernoulli distribution $Var(x) = p(1-p) = \frac{1}{64} \cdot \frac{63}{64} = \frac{63}{4096} = 0.0154$.

4 Problem

The probability that a patient recovers from a rare blood disease is 0.4 and a total of 10 people are known to have contracted this disease. Let X denote the random variable which corresponds to the number of patients who survive the disease. Assume that the patient's recoveries are all independent of one another.

1. What is the equation for the probability that X = k of the ten people survive?
2. To what family of distributions does the distribution of X belong? Provide this answer as X $\sim$ distribution_name(parameters). See the lecture notes for examples.

For the rest of the problem, use R to do the things listed below. You can use the built-in functions for this family of distributions. Include your code in the pdf writeup you submit to Gradescope and submit your code to Canvas as a .R or .ipynb file.

3. Plot the probability mass function (pmf) of X.
4. Plot the cumulative distribution function (cdf) of X.
5. What is the probability that at least 8 survive, i.e., $P\{X \geq 8\}$?
6. What is the probability that 3 to 8 survive, i.e., $P\{3 \leq X \leq 8\}$?

Answers

1. $P(X = k) = \binom{10}{k} p^k (1-p)^{10-k} = \binom{10}{k} (0.4)^k (0.6)^{10-k}$.
2. The binomial family, X $\sim$ Binom(n=10, p=0.4).
3. Plot the probability mass function (pmf) of X. We can use R to get all the values and plot them.

```
n <- 10
k <- seq(0,n)
p <- 0.4
pr <- dbinom(k, n, p)
pr
[1] 0.0060466176 0.0403107840 0.1209323520 0.2149908480 0.2508226560
[6] 0.2006581248 0.1114767360 0.0424673280 0.0106168320 0.0015728640
[11] 0.0001048576
barplot(pr, names.arg=k, ylab = "p(i)", xlab = "i")
```

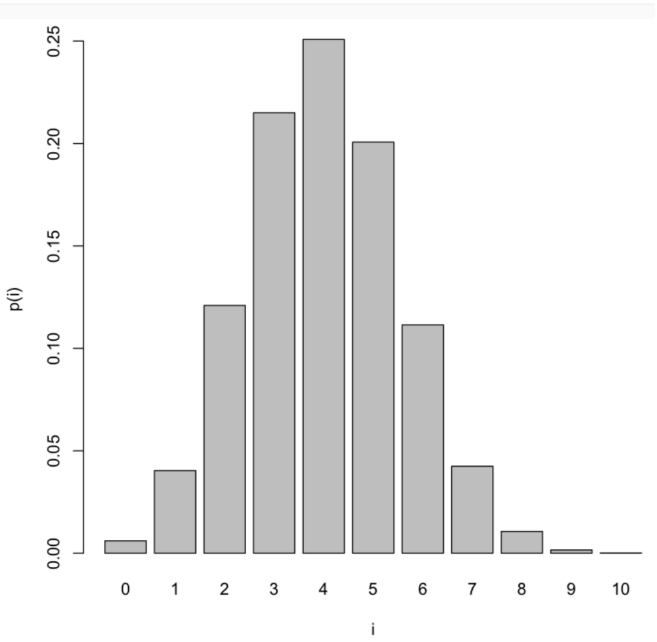


Figure 1: The probability mass function of Binomial random variable with parameter $n = 10$ and $p = 0.4$.

4. Plot the cumulative distribution function (cdf) of X.

```
pr <- pbinom(k, n, p)
[1] 0.006046618 0.046357402 0.167289754 0.382280602 0.633103258 0.833761382
[7] 0.945238118 0.987705446 0.998322278 0.999895142 1.000000000
barplot(pr, names.arg=k, ylab = "P(X <= i)", xlab = "i")
```

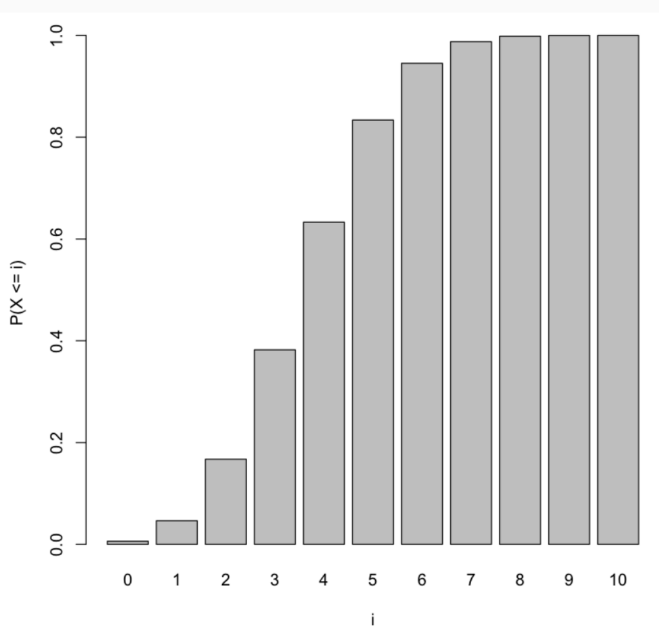


Figure 2: The cumulative distribution function of Binomial random variable with parameter $n = 10$ and $p = 0.4$.

5. What is the probability that at least 8 survive, i.e., $P\{X \geq 8\}$?

   $P\{X \geq 8\} = 1 - P\{X \leq 7\}$. Using R, we can do the following

```
pr = pbinom(7, n, 0)
ans <- 1 - pr
ans
0.01229455
```

6. What is the probability that 3 to 8 survive, i.e., $P\{3 \leq X \leq 8\}$? The required probability is

$$\begin{aligned} P\{3 \leq X \leq 8\} &= P\{X = 3\} + P\{3 < X \leq 8\} \\ &= P\{2 < X \leq 8\} \\ &= \text{pbinom}(8, n, p) - \text{pbinom}(2, n, p) \end{aligned}$$

   Again, we can use R

```
pbinom(8, n, p) - pbinom(2, n, p)
```

5 Problem

Consider the following program statement consisting of a **while** loop

$$\text{w}\mathit{hile}\ \neg\text{B}\ \mathit{do}\ S$$

Assume that the Boolean expression B takes the value true with probability p and the value false with probability q. Assume that the successive test on B are independent.

1. Find the probability that the loop will be executed k times.
2. Find the expected number of times the loop will be executed.
3. Considering the same above assumptions, suppose the loop is now changed to

   $$\text{r}\mathit{epeat}\ \text{S}\ \mathit{until}\ B$$

   What is the expected number of times that the repeat loop will be executed?

Answer

1. Let random variable X denotes the times the loop will be executed. X is distributed according to geometric* and can take values $\{0, 1, 2, \ldots, \}$. X takes the value 0 if B takes the value True the first time. Similarly, X will take the value 1 if B takes value False the first time and then the value True on the second time. And so on. The pmf is given by

   $$\begin{aligned} P(X = k) &= (1-p)^k p \qquad k \in \{0, 1, 2, \ldots, \} \\ &= q^k p \end{aligned}$$

   Consequently, $X \sim \text{Geom}^*(p)$ and represents number of failure until success for independent Bernoulli trials.

2. You can go to the Jupyter notebook and see that it is stated as a fact that for $X \sim \text{Geom}^*(p)$, that

   $$E[X] = \frac{1-p}{p} = \frac{q}{p}.$$

   But here we want to provide the full derivation. First recall how to sum a geometric series where a is an arbitrary constant:

   $$S = \sum_{k=0}^{\infty} aq^k = a\sum_{k=0}^{\infty} q^k$$

   To achieve a closed-form solution we first multiply this by q:

   $$qS = \sum_{k=0}^{\infty} aq^{k+1} = a\sum_{k=0}^{\infty} q^{k+1}$$

   Subtracting the second from the first:

   $$(1-q)S = a\left[\sum_{k=0}^{\infty} q^k - \sum_{k=0}^{\infty} q^{k+1}\right] = ax^0 = a$$

So we find:

$$S = \frac{a}{(1-q)} \text{ for } |q| < 1$$

Now consider the expectation value $X \sim \text{Geom}^*(p)$. By definition:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} P(X=k)k \\ E[X] &= p\sum_{k=0}^{\infty} q^k k \end{aligned}$$

Note we can write the this in terms of a derivative of the geometric series:

$$\begin{aligned} E[X] &= p\sum_{k=0}^{\infty} q^k k = p\left(q\frac{d}{dq}S\right) \\ &= pq\frac{d}{dq}\left[\frac{1}{(1-q)}\right] = pq\frac{1}{(1-q)^2}. \end{aligned}$$

Recall that $p = (1-q)$ so the equation simplifies:

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

3. Let random variable Y denote the times the repeat loop will be executed. Since the check is done at the end of the loop, the loop will be executed at least once. It is easy to see that the random variable Y takes value $\{1, 2, 3, \ldots\}$. $Y \sim \text{Geom}(p)$. In the Jupyter notebooks it is stated that for $Y \sim \text{Geom}(p)$ that

$$E[Y] = \frac{1}{p}.$$

We provide the full derivation here.

$$\begin{aligned} P(Y=k) &= (1-p)^{k-1}p \qquad k \in \{1, 2, \ldots, \} \\ &= q^{k-1}p \end{aligned}$$

We can think of Y as the number of trials until success. Thus, when $X = k$, the first $k-1$ trials must

be failure and the kth trial must be success.

$$\begin{aligned}
E(Y) &= \sum_{k=1}^{\infty} kP(Y=k) \\
&= \sum_{k=1}^{\infty} kq^{k-1}p \\
&= p\sum_{k=1}^{\infty} kq^{k-1} \\
&= p\sum_{j=0}^{\infty}(j+1)q^{j} \quad \text{this by setting } j = k-1 \text{ in the above eq.} \\
&= p\left(\sum_{j=0}^{\infty} jq^{j} + \sum_{j=0}^{\infty} q^{j}\right). \\
&= p\left(\frac{q}{p^2} + \frac{1}{1-q}\right). \\
&= \frac{1}{p}
\end{aligned}$$

ECS 132 Spring 2024: Assignment 4
Solution

This problem set covers:

- Problem 1 deals with the coupon collector problem
- Problems 2-4 concern Poisson processes
- Problem 5 is about quantiles, boxplots and data analysis

1 Problem

Recall the coupon collector problem discussed in class which has many applications in computer science. Consider a bag that contains N different types of coupons (say coupons numbered $1 \ldots N$). There are infinite number of each type of coupon. Each time a coupon is drawn from the bag, it is independent of the previous selection and equally likely to be any of the N types. Since there is an infinite number of each type, one can view this as sampling with replacement. Let T correspond to the random variable that denotes the number of total coupons that needed to be collected in order to obtain a complete set of at least one of each type of coupon. Write a R simulation code to compute E(T) considering the following:

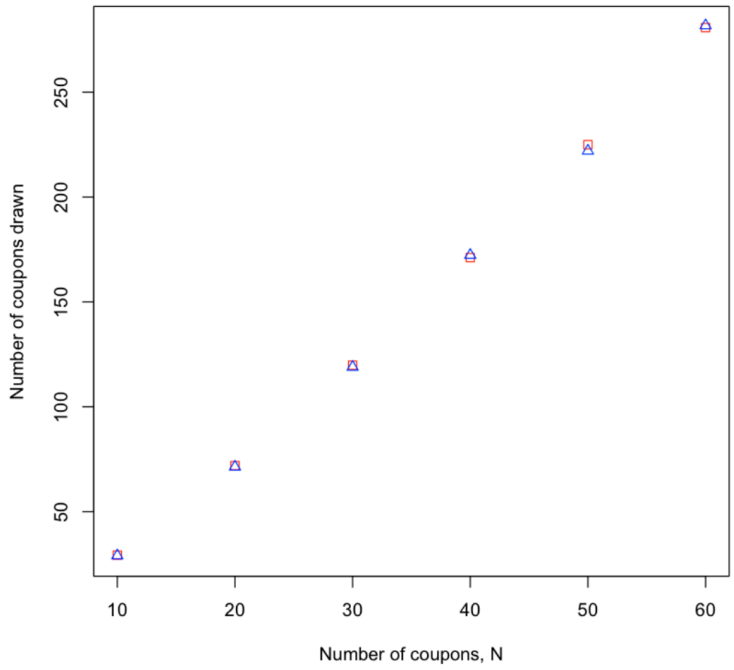
- N denotes the total number of coupons. Run your numerical simulation to develop an estimate of E(T) and plot E(T) for $N = 10, 20, 30, 40, 50, 60$. (Use 1000 trials or more, i.e, Nsim ≥ 1000.)
- We showed in class that for large N, $E(T)$ can be approximated by $N \log(N) + 0.577N + 0.5$. In the same plot show the theoretical value and summarize your observation regarding the accuracy of the approximation.

Answer

Here is the code.

```
N=c(10,20,30,40,50,60)     # different numbers of coupons to consider
T=rep(0,length(N))         # T[j]=mean number of draws needed for N[j] coupons
Theory=rep(0,length(N))    # Theory[j]=theoretical number for N[j] coupons
NSim=1000                  # The number of simulations to perform for each N value

for (j in 1:length(N)){
    num=rep(0,NSim)    # initialize fresh for each N[j] value
    for (i in 1:NSim) {
        trials=rep(0,0)
        while (length(unique(as.vector(trials))) < N[j]) {
            trials<-cbind(sample(1:N[j],1),trials)
            num[i]=num[i]+1
```

```
            }
        }
        T[j]=mean(num)
        Theory[j] = N[j]*log(N[j]) + 0.5771*N[j] + 0.5
}

plot(N,Theory,col="red",pch=0)
points(N,T,col="blue",pch=2)
```



Observations: There are two reasons why the simulated values differ from the theoretical value and they trade-off against each other: (1) the theoretical value is valid for large N, so the agreement should improve with N, but (2) the larger the sample size, greater the NSim value needed to converge to a stable value. (We will learn about error bars a little later.)

2 Problem

Consider that the entire human genome can be written in the form of a long book broken into pages. Suppose that the number of mutations on a single page of this book has a Poisson distribution with parameter $\lambda = 1$ (i.e., on average we expect 1 mutation per page).

1. For a given page, calculate the probability that there are at least 2 mutations on the page.
2. Now consider three pages, calculate the probability that there are no more than 2 mutations over all three of those pages.
3. For a given page, calculate the probability that there are at least 2 mutations on the page given that you have already spotted one mutation.

Answer

1. Let X denote the number of mutations on the page. Then we need to find $P(X \geq 2)$.
Recall the Poisson distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-1} - e^{-1} = 1 - 2e^{-1} \\ &= 0.264 \end{aligned}$$

2. Solve for $P(X \leq 2)$ over three pages. We were given that the expected number per page $\lambda = 1$, so the expected number over three pages $\lambda_3 = 3$ which is what we will use in the Poisson distribution. Recall the Poisson distribution:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

So,

$$\begin{aligned} P(X \leq 2) &= P(0) + P(1) + P(2) \\ &= e^{-3} + 3e^{-3} + \frac{9e^{-3}}{2} \end{aligned}$$

3 . Solve for $P(X \geq 2|X \geq 1)$.

$$\begin{aligned} P(X \geq 2|X \geq 1) &= \frac{P(X \geq 2)P(X \geq 1|X \geq 2)}{P(X \geq 1)} \\ &= \frac{P(X \geq 2)}{P(X \geq 1)} \\ &= \frac{1 - 2e^{-1}}{1 - e^{-1}} = 0.418 \end{aligned}$$

3 Problem

For the Poisson process explicitly show that $\sum_{k=0}^{\infty} p_k = 1$ where $p_k = P(X = k)$. Write out the first few terms explicitly and recall the formula for the Taylor series expansion for e^x.

Answer

$$\begin{aligned}\sum_{k=0}^{\infty} P(X=k) &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= e^{-\lambda}\left[1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}+\cdots\right] \\ &= e^{-\lambda}\cdot e^{\lambda} = 1.\end{aligned}$$

Recall the Taylor series expansion:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

4 Problem

This question deals with the Poisson process and is formulated in terms of a subway station. It could as well be formulated in terms of a networking problem. A subway station where different train lines intersect is like switch/router in the communication network with the different train lines corresponding to what are called Labelled Switched Paths (LSPs). As for the trains, you can think of each one as a packet or a burst of packets. One can also draw this analogy in optical networks which forms the core of the Internet backbone.

Consider that two one-way subway lines, the A train line and the B train line, intersect at a transfer station. The A trains and B trains arrive at the station according to independently operating Poisson processes with rates $\lambda_A = 3\ trains/hr$ and $\lambda_B = 6\ trains/hr$. We assume that passenger boarding and un-boarding occurs essentially instantaneously. [Note that the superposition of two Poisson processes with rates λ_1 and λ_2 is also a Poisson process with rate $\lambda_1 + \lambda_2$.]

1. What is $P(X = 9)$, the probability that the station handles exactly 9 trains during any given hour?
2. An observer arrives at the station at 8:00am. At the top of each subsequent hour (i.e., 9:00am, 10am, 11am, noon, etc) they record the number of trains the station has handled during that last hour. What is the expected number of hours they will need to wait until they first count exactly 9 trains arriving in an hour? (Note each hour is assumed to be independent of the previous hour and success is defined as exactly 9 trains arrive in an hour.)

Answers

- The combined process $N(t)$ of the A trains and B trains is a Poisson process with rate $\lambda_A + \lambda_B = 9$. The probability that there are exactly 9 trains in 1 hour is given by

$$\begin{aligned}P\{X=9\} &= e^{-\lambda}\frac{\lambda^9}{9!} \\ &= e^{-9}\frac{9^9}{9!} \\ &= 0.132\end{aligned}$$

- Let success correspond to the event if there are exactly 9 trains in any given hour starting on the hour. The probability of success denoted by p is 0.132 as computed above. The probability of failure, which corresponds to the event that the number of trains in a hour is any number other than 9, is $1 - p = 0.868$. Now each hour can be considered as a Bernoulli random variable. Furthermore, if we consider a sequence of hours we have a sequence of Bernoulli random variables which are independent. The independence comes from the fact that the underlying arrival process is Poisson. In a Poisson process the number of events in non-overlapping intervals are independent.

  The number of hours, (which is the number of trials) denoted by N until first success (observe exactly 9 trains) is geometrically distributed, i.e.,

  $$P\{N = k\} \quad = \quad p^{k-1}p \quad k = 1, 2, \ldots, \infty$$

  and we can show that the $E[N] = 1/p$. Thus, the expected number of hours is 7.6 hours. Since the observer only measures at the top of the hour, they will have to wait 8 hours on average before seeing 9 trains in one hour.

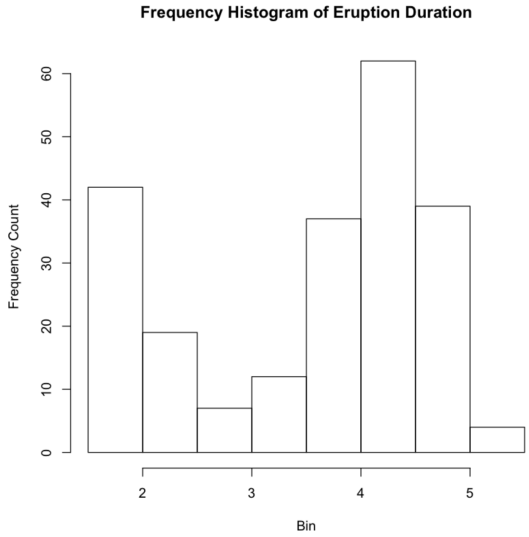
5 Problem

In lecture we discussed preliminary data analysis including the median, quantiles, quartiles, and boxplots using the inter-eruption time of the Old Faithful data. Use this data to answer the following questions concerning the eruption duration (the last column of the data set).

1. Plot the frequency histogram of the eruption duration with breaks of 1.
2. Draw the boxplot of the eruption duration, with the whiskers having the range=0.2 times the interquartile range.
3. What are the values for the 95, 97, and 99 quantiles of the eruption duration?
4. Suppose we classify the eruption duration using the following simple rule: if the duration is less than or equal to 3 mins then we classify it as a short eruption otherwise (i.e., if the duration is greater than 3 mins) it is a long eruption. Use the basic plot(x,y) function to draw a scatter plot that compares the current eruption duration (the x-axis) with the duration of the next eruption (the y-axis). In detail, suppose there are n data points $e[1] \ldots e[n]$. (You can find the number of data points using len(data[,3])). Then plot $(x = e[i], y = e[i+1])$ for $i = 1 \ldots n-1$. Note that both the x and y axes are in units time. On the scatter plot now draw a horizontal line at y=3 mins and a vertical line at x=3 mins. These two lines divide the area into 4 parts corresponding to long eruption followed by a long eruption, a long by a short, a short by a long, and short by a short.
5. Use R to analyze the data to estimate the probabilities that:
   - a long eruption is followed by a long eruption
   - a long eruption is followed by a short eruption
   - a short eruption is followed by a long eruption
   - a short eruption is followed by a short eruption

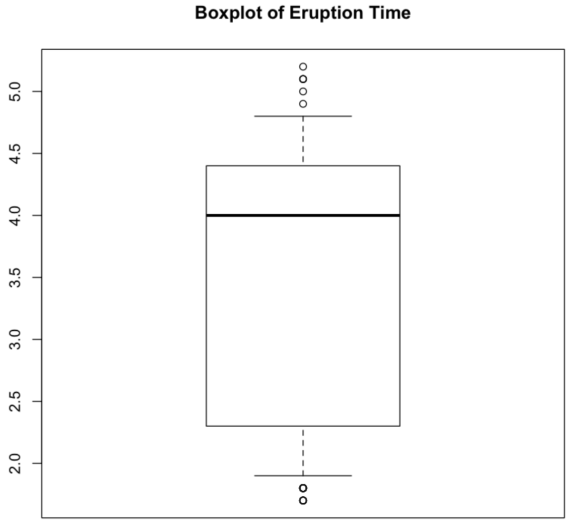
Answer

```
1) hist(data[,3], xlab ="Duration", ylab ="Frequency Count", main = "Eruption Durat
```



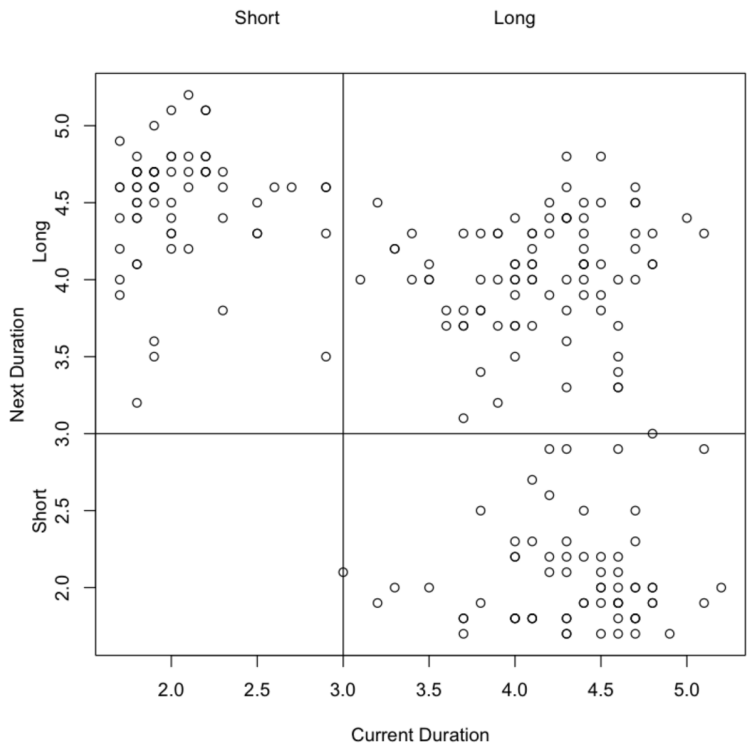

```
# Note the default hist function produces breaks of size 1 for this data set.
```

```
2) boxplot(data[,3], range=0.2, main = "Boxplot of Eruption Time")
```




```
3) q = c(.95, .97, .99)
quantile(data[,3], q)

Returns:
0.96      4.8
```

```
0.97      4.8
0.99      5.1

4 & 5)

data = read.table(file="./Data/Old_Faithful.txt", header=TRUE)

n = length(data[,1])
currentd = numeric(n-1)
nextd = numeric(n-1)
pss = 0
psl = 0
pll = 0
pls = 0

for (i in seq(1,n-1,1)) {
  currentd[i] = data[i,3]
  nextd[i] = data[i+1,3]
  if ((currentd[i] >= 3.0) & (nextd[i] >= 3.0))
      {pll = pll + 1}
    else if ((currentd[i] >= 3.0)  && (nextd[i] < 3.0))
        {pls = pls + 1}
    else if ((currentd[i] < 3.0)  && (nextd[i] < 3.0))
        {pss = pss + 1}
    else
        {psl = psl + 1}
}

plot(currentd,nextd, xlab = "Current Duration", ylab = "Next Duration")
mtext(c("Short","Long"),side=3,line=2,at=c(2.5,4))
mtext(c("Short","Long"),side=2,line=2,at=c(2.5,4))
abline(v=3)
abline(h=3)
```



5) see code above and then:

```
sprintf("Probability of Long Long is %f", pll/(n-1))
sprintf("Probability of Long Short is %f",pls/(n-1))
sprintf("Probability of Short Long is %f", psl/(n-1))
sprintf("Probability of Short Short is %f", pss/(n-1))
```

```
Returns:
Probability of Long Long is 0.398190
Probability of Long Short is 0.303167
Probability of Short Long is 0.298643
Probability of Short Short is 0.000000
```

ECS 132 Spring 2024: Assignment 5
Solutions

This problem set covers:

- Problem 1: Discrete random variables and conditional probability
- Problem 2: A general pdf and a discrete counting problem
- Problem 3: Minimum of two exponential processes
- Problem 4: Normal distribution and Z-scores
- Problem 5: Sensitivity and specificity

1 Problem

Recall the Medium Access Control (MAC) Protocol discussed in lecture (at the end of Sec V.6.) We saw that each time slot could be in one of three states: I (idle), C (collision) or S (Success). Consider here that there are 6 nodes who want to send packets and that all of the nodes act independently of one another (the example in lecture had 5 nodes). Each node has a probability $p = 0.4$ of being active in a time slot. Let Y be the random variable that denotes the number of nodes that are active in a time slot. For an idle slot $Y = 0$, for a successful slot $Y = 1$, and if $Y \geq 2$ the slot has a collision.

1. For a single time slot, calculate these two things: (a) $P(I)$ and (b) $P(S)$.
2. Now consider the event E, that there is an eavesdropper who sometimes monitors the time slot to read the the content being sent by the nodes, and that $P(E) = 0.5$. By looking through data logs from your security software you learn about the behavior of the eavesdropper. Specifically, that with probability 0.2 the the slot will be an idle slot given that the eavesdropper is listening, and that with probability 0.3 the slot will be a successful slot given that the eavesdropper is listening. What is the probability that the eavesdropper is listening given that the time slot is a collision? (Note, you calculated $P(I)$ and $P(S)$ above and are given information about conditional probabilities here.)

Answer

Since the sensors are independent the number of sensors that are active in any one time slot $Y \sim \text{binom}(6, p)$.

1. a) $P(I) = P(Y = 0) = \binom{6}{0} p^0(1-p)^6 = (1-p)^6 = 0.046656$
2. b) $P(S) = P(Y = 1) = \binom{6}{1} p(1-p)^5 = 6p(1-p)^5 = 0.186624$

3. We need to solve for $P(E|C)$.

   Givens:

   - $P(I)$ calculated above
   - $P(S)$ calculated above
   - $P(I|E) = 0.2$
   - $P(S|E) = 0.3$

   From these givens we can deduce:

   - $P(C) = 1 - P(I) - P(S) = 0.76672$
   - $P(C|E) = 1 - P(I|E) - P(S|E) = 0.5$

   Now to solve the problem:

$$\begin{aligned} P(E|C) &= \frac{P(E)P(C|E)}{P(C)} \\ &= \frac{(0.5)(0.5)}{0.76672} \\ &= 0.326 \end{aligned}$$

2 Problem

The probability density function of X, the lifetime of a certain type of electronic device (measured in hours) is given by

$$f(x) = \begin{cases} \frac{10}{x^2} & x > 10 \\ \\ 0 & x \leq 10 \end{cases}$$

1. Calculate the mathematical expression for the cumulative distribution function of X.
2. What is the probability that the device will fail within the first 15 hours?
3. Now consider that you have a large collection of these devices lined up in a row, numbered device1, device2, etc. They are all independent of one another. What is the probability that the 4th device will be the first one that does not fail within the first 15 hours?

Answers

1. Calculate the cumulative distribution function of X.

$$\begin{aligned} P\{X \leq a\} &= \int_{-\infty}^{a} f(x)dx = \int_{10}^{a} f(x)dx \\ &= \int_{10}^{a} \frac{10}{x^2}\, dx \\ &= -10x^{-1}\Big|_{10}^{a} \\ &= 1 - \frac{10}{a} \quad \text{for } a \geq 10 \text{ and } 0 \text{ otherwise} \end{aligned}$$

2. Solve for $F_X(15)$.

$$\begin{aligned} P\{X \leq 15\} &= 1 - \int_{10}^{15} \frac{10}{x^2} dx \\ &= -10x^{-1}\Big|_{10}^{15} \\ &= -(2/3) + 1 \\ &= 1/3 \end{aligned}$$

3. What is the probability that the 4th device will be the first one that does not fail within the first 15 hours?

   Each device is independent and follows a Bernoulli process where success is that the device does not fail within the first 15 hours, $p = P(X > 15)$. The random variable Y corresponding to the first success in a series of $X \sim Bern(p)$ processes is the geometric distribution, $Y \sim Geom(p)$.

   Using the part above we solve for $p = 1 - P(X \leq 15) = 2/3$. The probability that $Y = 4$ is

$$\begin{aligned} P(Y = 4) &= (1-p)^3 p \\ &= (1/3)^3(2/3) = 0.0247 \end{aligned}$$

3 Problem

You arrive at a bus stop to find there are no busses currently there. Two different lines, line A and line B service the bus stop and both will take you to your destination so you get onto the first bus that arrives. Both busses have arrival times corresponding to exponential distributions and the busses are independent of one another. Bus A on average comes by once every ten minutes. Bus B on average comes by once every 5 mins. What is the probability that you will have to wait more than 8 mins for a bus to arrive?

Answer

The arrival time of Bus A is random variable $X_A \sim \text{Expo}(\lambda_A)$ and the arrival time of Bus B is random variable $X_B \sim \text{Expo}(\lambda_B)$. Here the rates $\lambda_A = 1/10$ and $\lambda_B = 1/5$. Just like the example in lecture of the two cores handling arriving jobs, we want to define the random variable $W = \min(X_A, X_B)$ and calculate the probability that $W > 8$.

$$\begin{aligned} P(W > 8) &= P(X_A > 8 \cap X_B > 8) \\ &= P(X_A > 8) \cdot P(X_B > 8) \\ &= \exp(-\lambda_A \cdot 8) \cdot \exp(-\lambda_B \cdot 8) \\ &= \exp(-\frac{1}{10}8) \cdot \exp(-\frac{1}{5}8) \\ &= (0.449)(0.202) = 0.0907 \end{aligned}$$

Recall that the CDF for the exponential distribution $F_X(y) = 1 - \exp(-\lambda y)$. So $P(X > y) = 1 - F_X(y) = \exp(-\lambda y)$

4 Problem

Suppose the size of a network flow is a normal random variable with parameters $\mu = 71$ GBytes and $\sigma = 2.5$ GBytes. Use a Z-table to determine the following. (This requires mapping your normal distribution onto the standard normal distribution.)

1. What percentage of the flows are greater than 72 GBytes?
2. Let m denote the size of the flow. What is the value of m for which 88.30% of the flows are smaller than m?

Answer

1. Let X be a random variable that denotes the network flow size. $X \sim \mathrm{Norm}(71, 2.5)$. We want to find $P(X > 72)$ which is $1 - P(X \leq 72)$.

$$\begin{aligned} P(X \leq 72) &= P(\frac{X - 71}{2.5} \leq \frac{72 - 71}{2.5}) \\ &= \Phi(0.4) \\ &= 0.6554 \end{aligned}$$

   Thus the required probability is $1 - 0.6554 = 0.3446$.

2. We want to know for what value of m is it that case that $F_X(m) = 0.8830$. From the Z-table we read off that $\Phi(Z) = 0.8830$ for $Z = 1.19$. Now we translate that into the X value. Recall

$$Z = \frac{X - \mu}{\sigma} \quad \text{so} \quad X \ = \ Z\sigma + \mu$$

   And here we set $X = m$ so

$$m = \ Z\sigma + \mu \ = \ (1.19)(2.5) + 71 = 73.975$$

5 Problem

In a newspaper trivia column, L. M. Boyd (Boyd, L. M.: The Grab Bag (syndicated newspaper column), The San Francisco Chronicle (July 17, 1999)) ponders why lie detector results are not admissible in court. His answer is that "lie detector tests pass 10 percent of the liars and fail 20 percent of the truth-tellers." If you use these percentages and take $L = 1$ to mean being a liar and $F = 1$ to mean failing the test, what are the numerical values of the (1) sensitivity and (2) specificity for such a lie detector test?

Answer

We define the following two random variables

$$L = \begin{cases} 1 & \text{person is a liar} \\ 0 & \text{person is a truth teller} \end{cases}$$

$$F = \begin{cases} 1 & \text{fails lie detector test} \\ 0 & \text{passes lie detector test} \end{cases}$$

1. 10% of the liars pass test. This implies $P(F = 0)|(L = 1) = 0.1$, which further implies that $P(F = 1)|(L = 1) = 0.9$. Thus the sensitivity $\eta = 0.9$.

2. Similarly, 20% of truth tellers fail the test. This implies that $P(F = 1)|(L = 0) = 0.2$. This implies that the specificity $\theta = P(F = 0)|(L = 0) = 0.8$.

   (Remember $P(A^{\complement}|B) + P(A|B) = 1$.)