ECS 132 Fall 2020: Assignment 4 - 10 pts

October 30, 2020

Instructions

1. You may in no circumstances upload your homework to private tutoring websites such as CourseHero or Chegg. Remember all material related to this course is a property of the University of California and posting them is a violation of the copyright laws.
2. If you refer to a source (either a book or the internet), you must cite it.
3. You are highly urged to work on these problems on your own. See the homework grading policy. Not getting the answer correct has very low penalty. However, trying it and then figuring out where you went wrong will really help you understand the material better and you will be much better prepared for the exams. If you do discuss with others, you must list their names.
4. Write your answers in R Markdown and submit the knitted pdf on Gradescope; for due date and other details see the Homework Policy and Schedule.

1 Problem

We consider a noisy communication link in which the message is encoded in to binary digits (0,1) (bits) before being transmitted. We will denote the length of the encoded message by n. Since the channel is noisy, the bits can get flipped; a 0 to 1 or a 1 to 0. We will assume that each bit is flipped independently with probability p. In order to be able to detect that the received message is in error, a simple method is to add a parity bit at the transmitter. There can two types of parity - even parity and odd parity. If even (odd) parity is used , the parity bit is set such that total number of 1s in the encoded message is even (odd). For the sake of this problem, we will only consider even parity. Here are a few examples

```
1) message: 0001 + parity bit: 1 --> encoded message: 00011
2) message: 0101 + parity bit: 0 --> encoded message: 01010
3) message: 0111 + parity bit: 1 --> encoded message: 01111
```

At the receiver, the parity bit is computed from the message bits. If the received parity bit does not match the computed parity bit, then the message is flagged to be in error. For example, if the received message 1 is 10011, then it is in error.

1. Suppose $n = 7$ and $p = 0.1$, what is the probability that the received message has errors which go undetected?
2. For general n and p, write down an expression (as a sum) for the probability that the received message has errors which go undetected.

1.1 Answer

Notice that if there are even number of errors (and nonzero number of errors) then the error will go undetected. Now just use the Binomial distribution, Binom(n=7,p) to sum the probabilities for 2, 4, 6 errors.

$$P(error\ undetected) = \binom{7}{2}p^2(1-p)^5 + \binom{7}{4}p^4(1-p)^3 + \binom{7}{6}p^6(1-p)^1$$

Part 2 is just a generalization of the above. We find sum of even number (and non-zero number) of errors. This is given by

$$P(error\ undetected) = \sum_{k \geq 2\ and\ even} \binom{n}{k}p^k(1-p)^{n-k}$$

1.2 Points - 2pts

- 1 pt for setting up and finding the p(undetected error)
- 1 pt for properly setting up and solving the generalization

2 Problem

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

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

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

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

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

2.1 Answer

1. Let random variable X denotes the times the loop will be executed. X is geometrically distributed and can take values $\{0, 1, 2, \ldots, \}$. X takes the value 0 if the B takes the value True the first time. Similarly, X will take the value 1 is B first takes value False and then the value True. 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, $X \sim Geom(p)$ and represents number of failure until success for independent Bernoulli trials.

2. By definition:

$$\begin{aligned}
E[X] &= \sum_{k=0}^{\infty} P(X=k)k \\
E[X] &= p\sum_{k=0}^{\infty} q^k k \\
q \times E[X] &= p\sum_{k=0}^{\infty} q^{k+1}k \\
(1-q)E[X] &= p\sum_{k=1}^{\infty} q^k = pq\lim_{k\to\infty}\frac{1-q^{k+1}}{1-q} = pq\frac{1}{1-q} \\
E[X] &= \frac{q}{p}
\end{aligned}$$

3. Let random variable Y denotes 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 $\{1, 2, 3, \ldots\}$. $Y \sim \text{Geom}^*(p)$. Note that while X above and Y both have Geometric distributions, they are slightly different. The pmf of Y is given by

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

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.

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

2.2 Points - 4 pts

- 1 pt for part 1. Answer should determine that X is geometrically distributed by using appropriate equation, then properly identify PMF. Lose 1/2 pt for using other method to solve.

- 1 pt for part 2, using definition of Geometric distribution to compute the expected value.
- 2 pts for part 3. 1 pt for identifying the PMF of Y (or whatever you term this variable), and 1 pt for solving the expected value

3 Problem

The following problem is called the coupon collector problem and has many applications in computer science. Consider a bag that contains N different types of coupons (say coupons numbered $1 \dots N$. There are infinite number of each typ 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 are infinite numbers of each type, one can view this as sampling with replacement. Let T denote the random variable that denotes the number of coupons that needs to be collected until one obtains a complete set of atleast one of each type of coupon. Write a R simulation code to compute the E(T). Plot E(T) as for $N = 10, 20, 30, 40, 50, 60$.

We will show in class that for large N, $E(T)$ can be approximated by $N \times log(N) + 0.577 \times N + 0.5$. In the same plot show the theoretical value and summarize your observation regarding the accuracy of the approximation.

3.1 Answer

Here is the code for $N = 50$. Iterating over N and plotting is straightforward.

```
# N:      the number of unique coupons
# NSim:   the number of simulations that we will perform
# num:    record the number of trial needed to get a complete
#         set of one of each type of  coupon for each simulation

N=50
NSim=10000                   # Number of simulations
num=rep(0,NSim)                # This is  a vector initialized to 0
for (i in 1:NSim){
  trials <-rep(0,0)                  # for a simulation intialize trials to empty
  while (length(unique(as.vector(trials)))<N){    # until all coupons collected
    trials<-cbind(sample(1:N,1),trials)  # withdraw a coupon and add to trials
    num[i]=num[i]+1                      # increment trials
  }
}
results <- list(N=N,
          Nnumber_Simulations=NSim,
          Simulation_mean=mean(num),
          Theoretical_value=N*log(N) + 0.5771*N +0.5)
```

3.2 Points - 3 pts

- 2 pts for R-script that is syntactically correct and effectively models the system. (Partial credit if the script has errors)

- 1 pt for a plot that is properly setup and is consistent with the simulation you generated (you don't lose points twice for an error in the script) but it should contain both E(T) AND the approximation given in the problem. If it contains only one, lose 1/2 pt.

4 Problem

The entire human genome can be considered to 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 $\lambda = \frac{1}{2}$. For a given page, calculate that there are atleast 2 mutations on the page?

4.1 Answer

Let X denote the number of mutations on the page. Then we need to find $P(X \geq 2)$. This is given by

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-\frac{1}{2}} - \frac{1}{2} \times e^{-\frac{1}{2}} \\ &= 0.0902 \end{aligned}$$

4.2 Points - 1 pt

- 1/2 pt for properly setting up equation
- 1/2 pt for solving