Homework #4

APPM 4570/5570, Statistical Methods, Fall 2017

Due in class on Friday October 6, 2017. *Instructions for "theoretical" questions: Answer all of the following questions. The "theoretical" problems should be neatly numbered, written out, and solved. Do not turn in messy work. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work. You should always justify your answer.*

Theoretical Questions

1. Individuals A and B play a sequence of chess games until one player wins 10 games. A wins an individual game with probability p, and B wins a game with probability $1-p$ (i.e., there are no draws). Let X denote the number of games played.

   (a) What are the possible values of X?

   The minimum possible value is $X = 10$, where one individual wins 10 consecutive games. The maximum possible value is $X = 19$, where one individual wins after the series is tied 9-9. Thus, the range is: $10 \leq X \leq 19$

   (b) Obtain an expression for $P(X = x)$.

   Letting X = # of games played $\epsilon\{10, 11, ...19\}$ and S_a = #of games won by A, S_b = #of games won by B, the probability can be interpreted as,

   $$P(X = x) = P(\text{A wins } 10^{th}\text{game on game x} \cup \text{B wins } 10^{th}\text{game on game x})$$
   $$= P(\text{A wins } 10^{th}\text{game on game x}) + P(\text{B wins } 10^{th}\text{game on game x})$$
   $$= P(S_a = 9 \text{ in x-1 games} \cap \text{A wins game x}) + P(S_b = 9 \text{ in x-1 games} \cap \text{B wins game x})$$
   $$= \text{Bin(x-1,p)} \cdot \text{(p)} + \text{Bin(x-1,1-p)} \cdot \text{(1-p)}$$
   $$= \binom{x-1}{9} p^9 (1-p)^{x-1-9} \cdot p + \binom{x-1}{9} (1-p)^9 p^{x-1-9} \cdot (1-p)$$
   $$= \binom{x-1}{9} p^{10} (1-p)^{x-1-9} + \binom{x-1}{9} (1-p)^{10} p^{x-1-9}$$

   (c) Let $p = 0.5$. Find $P(X = 12)$.

   $$P(X = 12) = \binom{12-1}{9} 0.5^{10} (1-0.5)^{12-1-9} + \binom{12-1}{9} (1-0.5)^{10} 0.5^{12-1-9}$$
   $$= \binom{11}{9} 0.5^{10} (0.5)^2 + \binom{11}{9} (0.5)^{10} 0.5^2$$
   $$= 2\binom{11}{9} (0.5)^{12}$$
   $$\approx 0.0269 = 2.69\%$$

2. (a) A traffic office wishes to monitor the number of vehicles crossing a certain bridge in the city in any given day. What family of distributions will they most likely be observing?

A Poisson distribution is appropriate for this case.

(b) If it is estimate that an average of 300 cars cross the bridge per day, what is the probability that less than 150 cars will cross the bridge on any given day?

A Poisson distribution follows this equation,

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

where λ is equal to the average. To determine the probability that less than 150 cars will cross the bridge on any given day, the pdf is summed over an interval from 1 to 149,

$$P(X < 150) = \sum_{i=1}^{149} \frac{300^i e^{-300}}{i!}$$
$$\approx 3.291\text{e-}22$$

The probability is effectively zero.

3. A couple wishes to have exactly two female children in their family. They will have children until this condition is fulfilled. Assume that male and female births are equally likely.

(a) What is the probability that the family has x male children?

Using the negative binomial distribution, and letting X be the number of males before the condition of 2 females is met, the probability can be interpreted as,

$$P(X = x) = \text{NBin(x,r,p)} = P(X = x) = \binom{x+1}{r-1}(1-p)^r p^x$$
$$= \text{NBin(x, 2, 0.5)} = P(X = x) = \binom{x+1}{1}(0.5)^2(0.5)^x$$

(b) What is the probability that the family has four children?

Here, the number of male children, x, is 2, and the condition where we stop counting (at 2 females), r, is also 2.

$$P(X = x) = \binom{x+1}{1}(1-p)^r p^x$$
$$P(X = 2) = \binom{3}{1}(0.5)^2(0.5)^2 = 3(\frac{1}{2})^4$$
$$= 0.1875$$

(c) What is the probability that the family has at most four children?

$$\begin{aligned} P(X \leq 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= \binom{1}{1}(0.5)^2 0.5^0 + \binom{2}{1}(0.5)^2 0.5^1 + \binom{3}{1}(0.5)^2(0.5)^2 \\ &= (\frac{1}{2})^2 + 2(\frac{1}{2})^3 + 3(\frac{1}{2})^4 \\ &= \frac{11}{16} = 0.6875 \end{aligned}$$

4. A certain lab machine has 22 rings. With each use, these rings can fail, and an oil leak occurs. The probability of any ring failing during machine use is 8%. The rings are independent, and the failure of one ring does not impact the probability of failing for other rings. The machine gets serviced after each use, so any damaged rings are repaired after each use.

   (a) If 4 or more rings fail, the entire machine will shut down. What is the probability of the machine shutting down on any one use?

   Let X be a random variable such that X: number of ring failures and that X $\in \{0, 1, 2, ...22\}$. For this case X can be described with a binomial distribution,

   $$X \sim \text{Binom}(22, 0.08) : P(X = x) = \binom{n}{x} p_r^x (1 - p_r)^{n-x}$$

   where p_r is the probability of a ring failure ($p_r = 0.08$). We want to find the probability of 4 or more rings failing,

   $$\begin{aligned} P(X=4) + P(X=5) + ... + P(X=22) &= P(X \geq 4) \\ &= \sum_{i=4}^{22} P(X=i) \\ &= \sum_{i=4}^{22} \frac{22!}{x!(22-x)!}(0.08)^x (0.92)^{22-x} \\ &= 0.0941 = 9.41\% \end{aligned}$$

   (b) What is the probability that the machine runs successfully at least 5 times before shutting down?

   Let Y be a random variable such that Y: number of successful runs. For this case, Y can be described with a geometric distribution,

   $$Y \sim \text{Geom}(0.0941) : P(Y = y) = (1 - p_m)^{y+1}$$

   where p_m is the probability of the machine shutting down from part a. We want to find the probability of 5 or more successful runs. A cdf can be used to calculate this probability,

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y \leq 4) \\ &= 1 - [1 - (1 - p_m)^{4+1}] \\ &= (1 - 0.0941)^5 \\ &= 0.6101 = 61.01\% \end{aligned}$$

5. Let $X =$ the leading digit of a randomly selected number from a large accounting ledger. So, for example, if we randomly draw the number \$20,695, then $X = 2$. People who make up numbers to commit accounting fraud tend to give X a (discrete) uniform distribution, i.e., $P(X = x) = 1/9$, for $x \in \{1, ..., 9\}$. However, there is empirical evidence that suggests that "naturally occurring" numbers (e.g., numbers in a non-fraudulent accounting ledger) have leading digits that do not follow a uniform distribution. Instead, they follow a distribution defined by:

$$f(x) = \log_{10}\left(\frac{x+1}{x}\right), \quad x = 1, 2, ..., 9.$$

   (a) Show that $f(x) = P(X = x)$ is, in fact, a probability distribution function.

   Probability distribution functions need to satisfy two conditions:

   1. $f(x) \geq 0$, for all x
   2. For discrete distributions : $\sum_x P(X = x) = 1$

   By inspection, the first condition is satisfied since the terms within the log function will always be greater than 1; thus, the log function will always be positive. Since $x + 1 > x$,

$$\log_{10}\left(\frac{x+1}{x}\right) > \log_{10}(1) = 0, \forall x \in \{1, ..., 9\}$$

   For the second condition, a summation of the function can be calculated using the quotient property of logarithms,

$$\begin{aligned} \sum_{i=1}^{9} \log_{10}\left(\frac{i+1}{i}\right) &= \log_{10}(2) + \log_{10}\left(\frac{3}{2}\right) + \log_{10}\left(\frac{4}{3}\right) + ... + \log_{10}\left(\frac{10}{9}\right) \\ &= \log_{10}(2) + [\log_{10}(3) - \log_{10}(2)] + ... + [\log_{10}(10) - \log_{10}(9)] \\ &= \log_{10}(10) = 1 \end{aligned}$$

   Thus, both conditions are satisfied.

   (b) Compute the individual probabilities for $x \in \{1, ..., 9\}$, and compare them to the corresponding discrete uniform distribution (i.e., $P(X = x) = 1/9$). What do you notice?

Table 1: Probability Table of pdf and Uniform Distribution

Function	**X=1**	**X=2**	**X=3**	**X=4**	**X=5**	**X=6**	**X=7**	**X=8**	**X=9**
"Benford's Law"	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
Uniform	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

Looking at the table above, it is seen that the pdf follows a decaying function and is different from the uniform function.

(c) Obtain the cdf of X.

A summation function of the pdf can be used to represent the cdf since the pdf is discrete,

$$\begin{aligned} F(X=x) &= \sum_{i=1}^{x} \log_{10}\left(\frac{i+1}{i}\right) \\ &= \sum_{i=1}^{x} [\log_{10}\left(i+1\right) - \log_{10}\left(i\right)] \\ &= \log_{10}(2) + [\log_{10}(3) - \log_{10}(2)] + ... + \log_{10}(x+1) - \log_{10}(x) \\ &= \log_{10}(x+1) \end{aligned}$$

(d) Using the cdf, what is the probability that the leading digit is at most 4? At least 5?

Using the cdf from the previous part, these probabilities can be calculated,

$$\begin{aligned} P(X \leq 4) = F(X=4) &= \log_{10}(x+1) \\ &= \log_{10}(5) \approx 0.699 \end{aligned}$$

Using the previous probability, the next probability can be calculated,

$$P(X \geq 5) = 1 - F(X=4) = 1 - \log_{10} 5 \approx 0.301$$

Computational Questions

Instructions for "computational" questions: Your work should be neatly done and include all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do not put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. If you turn in something that is messy or out of order, it will be returned to you with a zero. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

1. Let X follow a geometric distribution, where the probability of success is 0.2.

   (a) What is the probability that 7 failures occur before the first success?

For the following problems, X:success, with P(X) = 0.2 The probability that 7 failures occur before the first success can be determined with a geometric pdf. Using R, this gives a probability of 0.0419.

```
> P_X7 = dgeom(7,0.2)
```

(b) What is the probability that 7 or more failures occur before the first success?

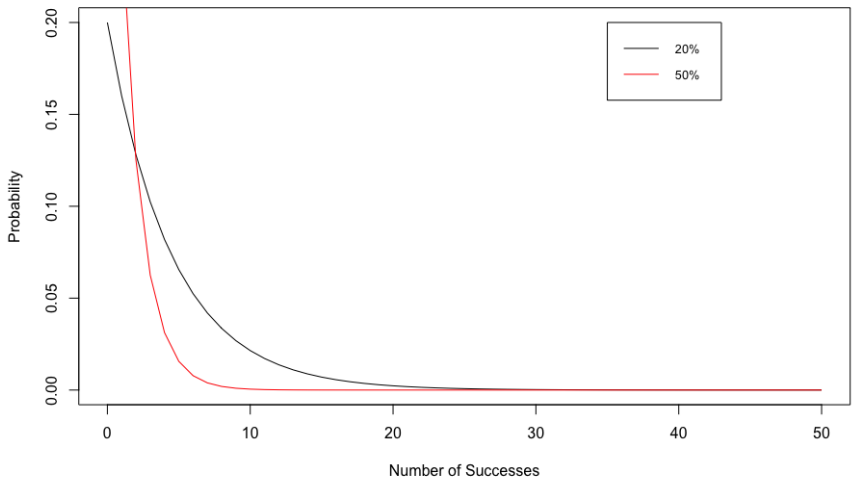
The probability that 7 or more failures occur before the first success can be determined with a geometric cdf.

$$P(X \geq 7) = 1 - P(x \leq 6) = 0.2097$$

```
> P_X7more = 1 - pgeom(6,0.2)
```

(c) Plot this pdf for $X = 1, ..., 50$, using the X values on the x-axis, and $P(X = x)$ on the y-axis. To this same plot, add a red line showing the probabilities for $X = 1, ..., 50$ when the probability of success is 0.5. What do you notice?

Plotting the pdf for two different probabilities, a figure is created below.

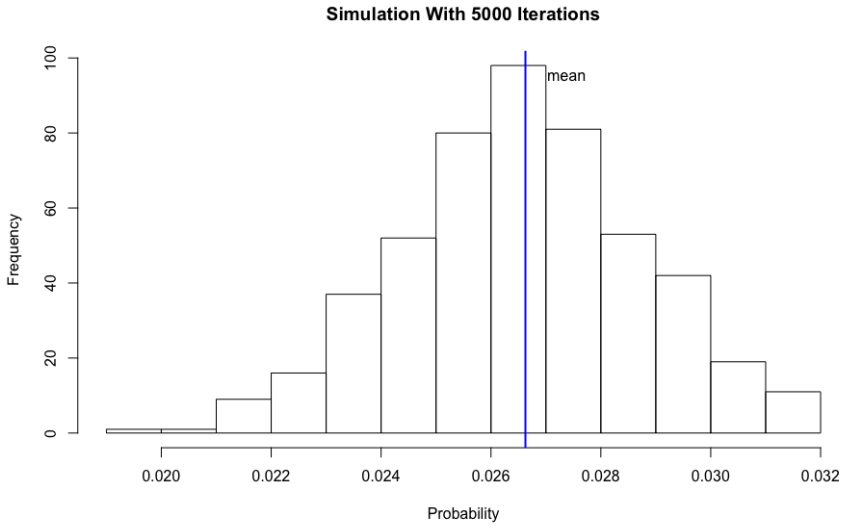



```
> P_X7more = 1 - pgeom(6,0.2)
> #PDFs
> P = dgeom(0:50,0.2)
> P2 = dgeom(0:50,0.5)
> #Plot
> plot(0:50,P,type="l")
> points(0:50,P2, type="l", col="red")
```

From the figure above, it is seen that both probabilities follow a geometric distribution. At first the 50% probability has a higher probability due to the smaller amount of trials; however, an intersection occurs at some point where the 20% is higher. A higher probability of success causes a lower probability of obtaining a success after a large amount of failures. Thus, the 20% probability becomes higher than the 50% due to this fact for higher amounts of trials.

2. Run a simulation that estimates the probability calculation in theoretical question 1 (c).

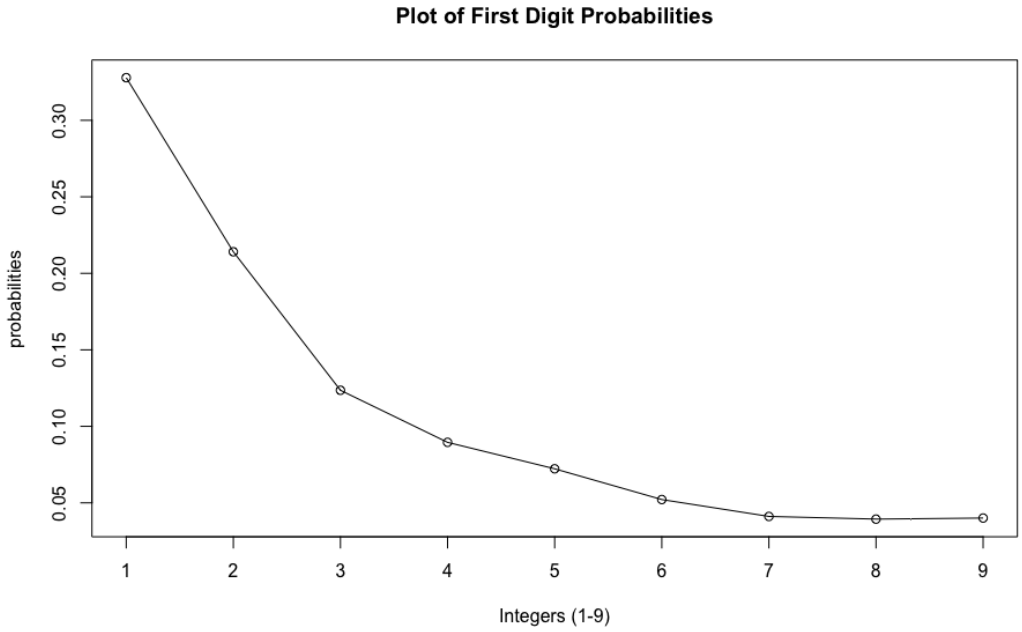
   Shown below, a histogram of the simulated probabilities is created to show the results,




```
> ### Run Simulation
> n = 5000
> #f returns on sequence of length 12
> f = function(){
+       x = sample(c(0,1), size = 12, replace = TRUE)
+       return(x)
+ }
> #create n sequences of length 12
> prob = matrix(NA, ncol = 500, nrow = 1)
> for (i in 1:500){
+       x = t(replicate(n, expr = f()))
+ #count sequences of winning the final 10th game
+ y = x[(x[,12] == 1 & rowSums(x) == 10) ...
+ ...|(x[,12] == 0 & rowSums(x) == 2) ,]
+ prob[i] = dim(y)[1]/n
+ }
```

   From R, the mean of the results from the simulation is : $\mu \approx 0.02663$, close to the calculated value from theoretical question 1 (c).

3. Using `tax.txt`, a dataset containing the taxable incomes for individuals in 1978, and the information given in theoretical problem 5, decide whether this dataset is fraudulent.




```
#Read Data File
> dataset = read.table('tax.txt')
#Preallocate zeros
> holder = rep(0,9)
> count = 0
#Determine probabilities of first digits
> for (i in 1 : length(dataset[,2])){
+ if (dataset[i,2] != 0){
+ first = as.numeric(strsplit(as.character(dataset[i,2]),"")[[1]])
+ holder[first[1]] = holder[first[1]] + 1
+ count = count + 1}
+ }
> probabilities = holder/count
#Check
> sum(probabilities)
#Plot
> plot(c(1:9),probabilities , xlab = "Integers (1-9)" , main = "Plot of First D
> axis(side = 1, at = c(1:9),labels = T)
> lines(c(1:9),probabilities)
```

The plot above shows the probabilities of the first digits within the tax data file. It is seen that the distribution follows that same distribution of the pdf in theoretical problem 5; thus, the taxable incomes from this data file is not fraudulent.