Homework #5

APPM 4570/5570, Statistical Methods, Fall 2017

Due in class on Friday October 13, 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. Suppose that the number of plants of a particular type found in a rectangular sampling region (called a quadrat by ecologists) in a certain geographic area is a random variable X with pmf

$$P(X = x) = \begin{cases} \frac{c}{x^a} & x = 1, 2, 3... \\ 0 & \text{otherwise.} \end{cases}$$

(a) Let $a = 3$. Is $E(X)$ finite? Justify your answer.

$$\begin{aligned} E(X) &= \sum x \cdot P(X = x) \\ &= \sum_{x=1}^{\infty} x \cdot \frac{c}{x^3} \\ &= \sum_{x=1}^{\infty} \frac{c}{x^2} = c \sum_{x=1}^{\infty} \frac{1}{x^2} \end{aligned}$$

The final expression above contains an infinite series; in particular, it is a p-series: $\sum \frac{1}{n^p}$. If $p > 1$, then it is said that the function converges; therefore, the final expression suggests that $E(X)$ converges and is finite.

(b) Let $a = 2$. Is $E(X)$ finite? Justify your answer.

$$\begin{aligned} E(X) &= \sum x \cdot P(X = x) \\ &= \sum_{x=1}^{\infty} x \cdot \frac{c}{x^2} \\ &= \sum_{x=1}^{\infty} \frac{c}{x} = c \sum_{x=1}^{\infty} \frac{1}{x} \end{aligned}$$

Likewise, the final expression above contains an infinite series. Since $p = 1$, the series does not converge; therefore, the final expression suggests that $E(X)$ does not converge and is not finite.

2. Let X = the outcome when a fair die is rolled once. Suppose that, before the die is rolled, you are offered a choice: Option #1: a guarantee of $\frac{1}{3.5}$ dollars (whatever the outcome of the roll); Option #2: $h(X) = 1/X$ dollars. Which option would you prefer? Justify your answer.

   In choosing between options, we will use the following rule:

   Principle of Maximizing Expected Value/Utility: If an act, A, has higher expected value than another act, B, then prefer A to B.

   We start with the expected value of Option #1: $\frac{1}{3.5} \approx 0.29$. The expected value of Option #2 is

$$E(h(X)) = \sum_{x=1}^{6} (1/x) \cdot (1/6)$$
$$E(h(X)) = \frac{1}{6} * (1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6}) = \frac{49}{120}$$
$$\frac{49}{120} \approx 0.41.$$

   We can see that, on average, Option #1 yields $\approx$ 29 cents, while Option #2 yields $\approx$ 41 cents (these values represent what we would expect to have after choosing these options over and over again). So according to our principle, we prefer Option 2 to to Option #1.

3. Recall the distribution from HW #4: X = the leading digit of a randomly selected number from a large accounting ledger. The pmf was defined by:

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

   Find $E(X)$.

   From HW #4, a table of the pmf is shown below.

Table 1: Probability Table of pmf

	X=1	**X=2**	**X=3**	**X=4**	**X=5**	**X=6**	**X=7**	**X=8**	**X=9**
P(X=x)	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

   Using the probabilities above, the expected value of X or mean can be calculated,

$$
\begin{aligned}
E(X) &= \sum x \cdot P(X = x) \\
&= \sum_{x=1}^{9} x \cdot \log_{10}\left(\frac{x+1}{x}\right) \\
&= 1(.0301) + 2(0.176) + 3(0.125) + \ldots + 8(0.051) + 9(0.046) \\
&\approx 3.44
\end{aligned}
$$

4. (a) Suppose that $E(X) = 5$ and $E[X(X-1)] = 27.5$. What is:
   i. $E(X^2)$?
   ii. $\text{Var}(X)$?

   i)

$$
\begin{aligned}
&E[X(X-1)] = 27.5 \\
&E[X^2 - X] = 27.5 \\
&E[X^2] - E[X] = 27.5 \\
&E[X^2] = 27.5 + 5 \\
&E[X^2] = 32.5
\end{aligned}
$$

   ii)

$$
\begin{aligned}
&\text{Var}[X] = E[X^2] - E[X]^2 \\
&\text{Var}[X] = 32.5 - 5^2 \\
&\text{Var}[X] = 7.5
\end{aligned}
$$

   (b) Suppose that $a \leq X \leq b$. Prove that $a \leq E(X) \leq b$ (consider both the discrete and the continuous cases).

   Suppose that $a \leq X \leq b$. Then, $E(a) \leq E(X) \leq E(b)$, where, in the continuous case,

$$
\begin{aligned}
E(X) &= \int_x x f(x) dx, \\
E(a) &= \int_x a f(x) dx = a \int_x f(x) dx = a, \text{ and} \\
E(b) &= \int_x b f(x) dx = b \int_x f(x) dx = b.
\end{aligned}
$$

   Thus, $a \leq E(X) \leq b$. The discrete case is similar.

5. Using the definition $E(X) = \sum_{x=0}^{\infty} xP(X = x)$, show that if $X \sim P(\lambda)$, then $E(X) = \lambda$ (HINT: What is the series representation of e^λ?).

Note that, since $X \sim P(\lambda)$, $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$. Thus,

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x P(X = x) \\ &= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \end{aligned}$$

Since the first element $(x = 0)$ within the sum expression equates to zero, the element can be taken out. Simplifying further,

$$\begin{aligned} \underbrace{\sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}}_{\text{first term is 0}} = \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} &= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x(x-1)!} \\ &= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\ &= e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \\ &= e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \end{aligned}$$

Recognizing that the summing expression is now that Taylor series for an exponential function,

$$\begin{aligned} &= e^{-\lambda} \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \lambda e^{\lambda} = \lambda. \end{aligned}$$

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. (**APPM 5570 Only**) Perform a simulation to further justify your answer to theoretical question 2. The simulation should answer the question: which option produces the better outcome if you were offered the choice over and over again?

   We can sample the experiment of winning (1/X) dollars where X is the roll of a fair die, and we confirm our results that:

   $$E(1/X) \approx 0.409 > 0.286 \approx E(1/3.5)$$

```
n = 5000
f = function(n){
x = sample(1:6,n, replace = TRUE)
e = mean(1/x)
return(e)
}

f(n);

# 0.4096

#To see the distribution of E(1/x):
h = replicate(n, expr = f());  hist(h)
```

2. One really cool (and useful!) application of random variables is approximating integrals/the area under a curve. Consider $f(x) = e^x$ on the interval $0 \leq x \leq 1$. Let's use uniform random variables to approximate the area under $f(x)$ on this interval. Note that this general idea is used often to solve really important but hard integrals!

   (a) By hand, and using the `integrate()` function in R, calculate the true area under $f(x)$.

   $$\begin{aligned} \int_0^1 e^x dx &= e^x \Big|_0^1 \\ &= e^1 - e^0 \\ &= e - 1 \approx 1.718 \end{aligned}$$

   Using the `integrate()` function in R,

```
> integrate(exp, 0,1)
[1] 1.718282 with absolute error < 1.9e-14
```

(b) Generate $n = 5,000$ uniform random (x, y) coordinates in the rectangle $x \in [0, 1]$, $y \in [0, e]$.

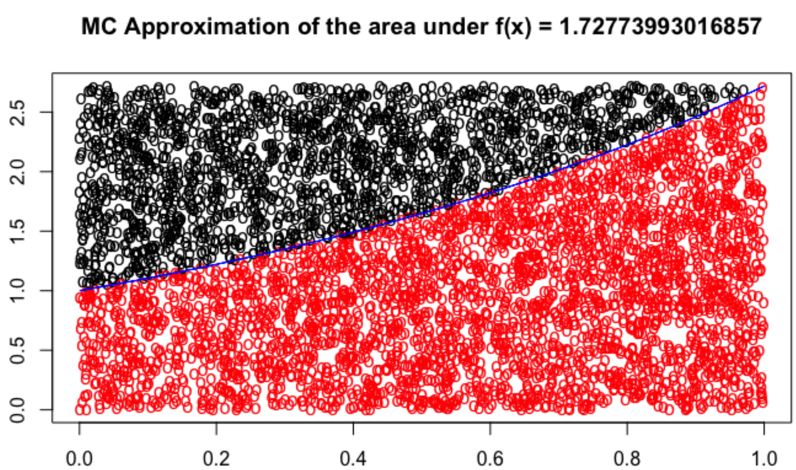
```
> n = 5000; x = runif(n,0,1); y = runif(n, 0, exp(1))
```

(c) Calculate the proportion of points from (b) that fall below $f(x)$ and use this proportion to approximate the area under $f(x)$.

```
> under_curve = y <= exp(x)
> aHat = (sum(under_curve)/n)*exp(1);
[1] 1.72774
```

(d) Plot all of the randomly generated points, giving different colors for point above and below $f(x)$.

```
> plot(x,y,pch='o',col=ifelse(under_curve,"red","black"), xlab='',
> grid = seq(0,1,length.out = n)
> lines(grid,exp(grid), col = "blue")
```



(e) Find the error between our approximation and the true area calculated in (a). How can we make this error smaller?

```
> error = abs(a\$value - aHat); print(error)
[1] 0.009458102
```

To reduce the error, a larger number of samples can be taken,

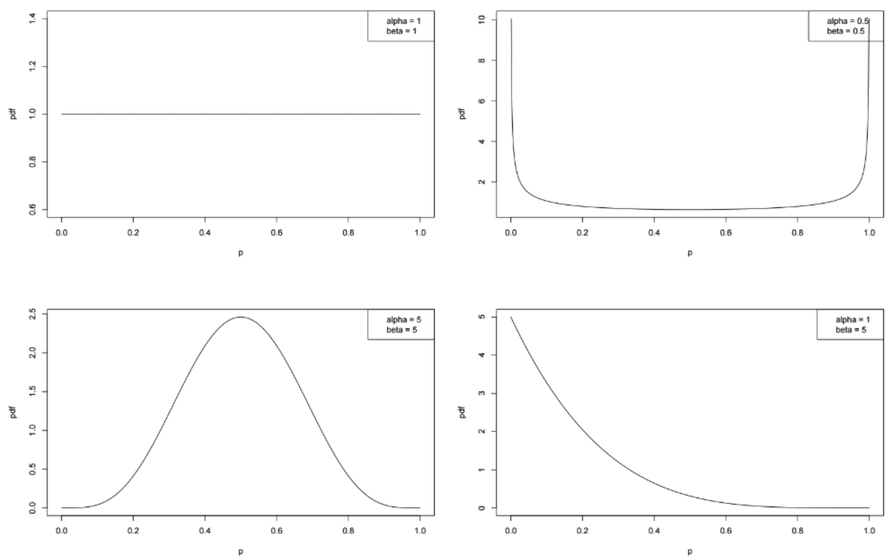
```
> n = 10000; x = runif(n,0,1); y = runif(n, 0, exp(1))
> under_curve = y <= exp(x)
> aHat = (sum(under_curve)/n)*exp(1); aHat
> error = abs(a\$value - aHat); print(error)
[1] 0.001958682
```

3. Assume the 'subjective' interpretation of probability for this problem. That is, probability models your degree of belief in an event or claim.

   (a) In the same window, plot four different beta distributions (pdf):

      i. beta(1,1)
      ii. beta(0.5,0.5)
      iii. beta(5,5)
      iv. beta(1,5)

   What is another name for (i)?



   Another name for (i) is a uniform distribution.

   (b) Now, suppose your friend has a coin whose probabilities are unknown to you. Since the beta distribution has support $[0, 1]$, it is well-suited to model your beliefs about the probability of heads. Match the following descriptions with the beta distributions from the previous part.

      i. I am pretty confident that this is a two-headed or two-tailed coin.
      ii. I have no idea what the probability of heads is.
      iii. I believe that the probability of heads is low.
      iv. I believe that the coin is close to fair.

   (i) : beta(0.5,0.5)
   (ii) : beta(1,1)
   (iii) : beta(1,5)
   (iv) : beta(5,5)