Homework 5

Autumn 2023

WRITE YOUR NAME HERE

2023-11-09

Instructions

- This homework is due in Gradescope on Wednesday Nov 8 by midnight PST. There is a 15 minute grace period and submissions made during this time will not be marked as late. Any work submitted past this period is considered late.
- Please answer the following questions in the order in which they are posed.
- Don't forget to (i) make a local copy this document for your work and to (ii) knit the document frequently to make sure there are no compilation errors.
- When you are done, download the PDF file as instructed in section and submit it in Gradescope.

Exercises

1. (Gambling) Independent tosses of a fair coin are made until a head appears. Let the random variable X denote the number of tails before the first head appears.

a. Write the PMF of X.

$$\begin{aligned} f(x) &= P(X = x), \\ &= \left(\frac{1}{2}\right)^{x} \times \frac{1}{2}, \\ &= \left(\frac{1}{2}\right)^{x+1}, x = 0, 1, 2, \ldots \end{aligned}$$

b. You will be given $Y = \left(\frac{1}{2}\right)^{X}$ dollars. How much money should you expect to get paid? Please show your steps. Cite any rules you use.

Hint Find $E[Y]$

$$
\begin{aligned}
E\left[Y\right] &= E\left[\left(\frac{1}{2}\right)^{X}\right], \\
&= \sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^{x} f(x) && \text{(law of unconscious probabilist)} \\
&= \sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^{2x+1}, \\
&= \frac{1}{2}\sum_{x=0}^{\infty}\left(\frac{1}{4}\right)^{x}, \\
&= \frac{1}{2}\times\frac{1}{(1-\frac{1}{4})}, && \text{(geometric series)} \\
&= \frac{2}{3}.
\end{aligned}
$$

I should expect to win $.67$ on a game.

2. (Blood type) In the US population, 85% have an agglutinating factor in their blood classifying them as Rh positive, while 15% lack the factor and are Rh negative. A medical researcher wants to analyze blood from a newborn Rh negative infant, so she examines the blood types of successive newborn infants until she finds an Rh negative infant.

a. Let the random variable X denote the number of Rh positive infants she needs to type before she finds the first negative case. Write the distribution of X. Be sure to state the values of any parameters. List the assumptions you need to make to answer this question.

   Each trial is the typing of an infant. On each trial we have two outcomes: Rh positive (failure) and Rh negative(success). The probability of a Rh negative, π, is the same for each infant (15%). Finally, we will *assume* that the outcome for an infant is independent of the outcome for another infant. **This means whether or not one infant is Rh negative has no relation to whether or not another infant is Rh negative.**

   Under these conditions,

$$X \sim Geom(\pi = 0.15).$$

b. How many Rh positive infants should she expect to type before she finds her first Rh negative? State the formula you are using and then report your answer in a complete sentence.

   From theorem 8.2, $E\left[X\right] = \frac{1-\pi}{\pi}$. This is calculated below.

```
pi <- 0.15
E_X <- (1-pi)/pi
```

   We should expect to type 5.6667 Rh positive infants before finding our first Rh negative.

c. What is the probability that she will type more Rh positive infants than expected? Give the expression for this probability here, do not calculate it yet.

   Hint see slide 8 from Chapter 8.1

Here we want to calculate $P(X > E_X)$. In other words:

$$\begin{aligned} P(X > E_X) &= P(X > 5.6667), \\ &= P(X \geq 6), \\ &= (1-\pi)^6 \end{aligned}$$

d. Calculate the probability from part c using a built in R function for working with the distribution from part a. Be sure to echo your code chunk and report your answer (rounded to 4 decimal places) using inline code.

```r
prob_more_than_expected <- pgeom(5, prob = 0.15, lower.tail=F)
```

The probability that they will type more Rh positive infants than expected is 0.3771.

3. (Memoryless) Let X be a geometric random variable with probability π. That is,

$$f(x) = \pi \cdot (1-\pi)^x, \quad x = 0, 1, 2 \ldots$$

a. Let k be some non-negative integer. Derive the expression $P(X \geq k)$. Be sure to show your steps.

Hint: We did this in class with $k = 19$ in Example 8.2 from Chapter 8.1. I want you to derive the steps again for generic k.

$$\begin{aligned} P(X \geq k) &= P(X = k) + P(X = k+1) + P(X = k+2) + \ldots \\ &= (1-\pi)^k \cdot \pi + (1-\pi)^{k+1} \cdot \pi + \ldots (1-\pi)^{k+2} \cdot \pi + \ldots \\ &= (1-\pi)^k \left(\pi + \pi(1-\pi) + \pi(1-\pi)^2 + \ldots\right) \\ &= (1-\pi)^k \frac{\pi}{1-(1-\pi)} \quad \text{geometric series} \\ &= (1-\pi)^k. \end{aligned}$$

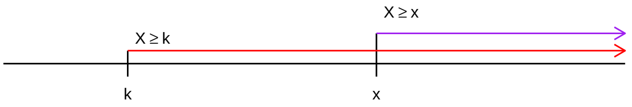
b. Show that for all non-negative integers x and k,

$$P(X \geq k + x | X \geq k) = P(X \geq x).$$

Before we begin with the proof, note that the set

$$X \geq x + k \subseteq X \geq k$$

because any x that satisfies the left hand condition also satisfies the right hand condition.



Therefore, the intersection of these two sets is simpler the smaller one:

$$P(X \geq x + k \cap X \geq x) = P(X \geq x + k).$$

Using the definition of conditional probability we have

$$\begin{aligned} P(X \geq x + k | X \geq k) &= \frac{P(X \geq x + k \cap X \geq k)}{X \geq k)}, \\ &= \frac{P(X \geq x + k)}{P(X \geq k)}, \\ &= \frac{(1-\pi)^{x+k}}{(1-\pi)^{k}}, \\ &= (1-\pi)^{x} = P(X \geq x). \end{aligned}$$

c. Because of the result in part b, we say that the geometric distribution is a *memoryless* distribution. Explain how this is an appropriate name for this property.

The word memoryless is appropriate for the result in part b since it says that if we have waited at least k trials for the first success, the probability that we have to wait at least another x trials is the same as initial probability of waiting at least x trials. That is, it is as the process has no memory that we have already waited for at least k trials.

4. (Apple trees) From each of 6 trees in an apple orchard, 25 leaves are selected. On each of the leaves, the number of adult mites are counted.

x	0	1	2	3	4	5	6	7	8+
count	70	38	17	10	9	3	2	1	0

The dataset can be created in R as shown below. Type `?rep` for help if you want to know what it does.

```
apple_trees<-tibble(
         mites=rep(0:7,
                    times=c(70,38,17,10,9,3,2,1) ))

##On your own, run this code chunk by clicking
##the green arrow and look ##at the dataset
##you have created by clicking on it in the
##Environment. Then try glimpse, slice_head,
##count in the Console like we did with the
##Fumbles data in Chapter 8.2.
##Nothing to turn in, just for fun.
```

a. Write code in the code chunk below to print the number of observations (leaves) in the data set along with the mean and standard deviation of the `mites` variable. (See sample code in Example 8.6 from Chapter 8.2; you will need to modify it for this context)

```
apple_trees %>% summarise(n = n(),
                          xbar = mean(mites),
                          s = sd(mites))
```

```
## # A tibble: 1 x 3
##         n  xbar     s
##   <int> <dbl> <dbl>
## 1   150  1.15  1.51
```

b. Let X denote the number of mites on a randomly selected leaf. In this question, you will consider two possible models for X:

- $X \sim Pois(lambda = \bar{x})$
- $X \sim Geom(prob = \frac{1}{1+\bar{x}})$

  where $\bar{x}$ is the number you calculated from part a. for the mean.

  Write code in the code chunk below to make side-by-side plots showing how well each model fits. Each plot must have the histogram of the data with the probabilities according to the model overlaid on it. Just by eyeballing, which model appears to fit the data better?

  Hint: For help with the code for side-by-side plots, look at slide 11 from the slidedeck Chapter 8.1. You will obviously need to modify this code for this example. For help with overlaying theoretical probabilities on a histogram, look at slide 26 from the slidedeck Chapter 8.2.

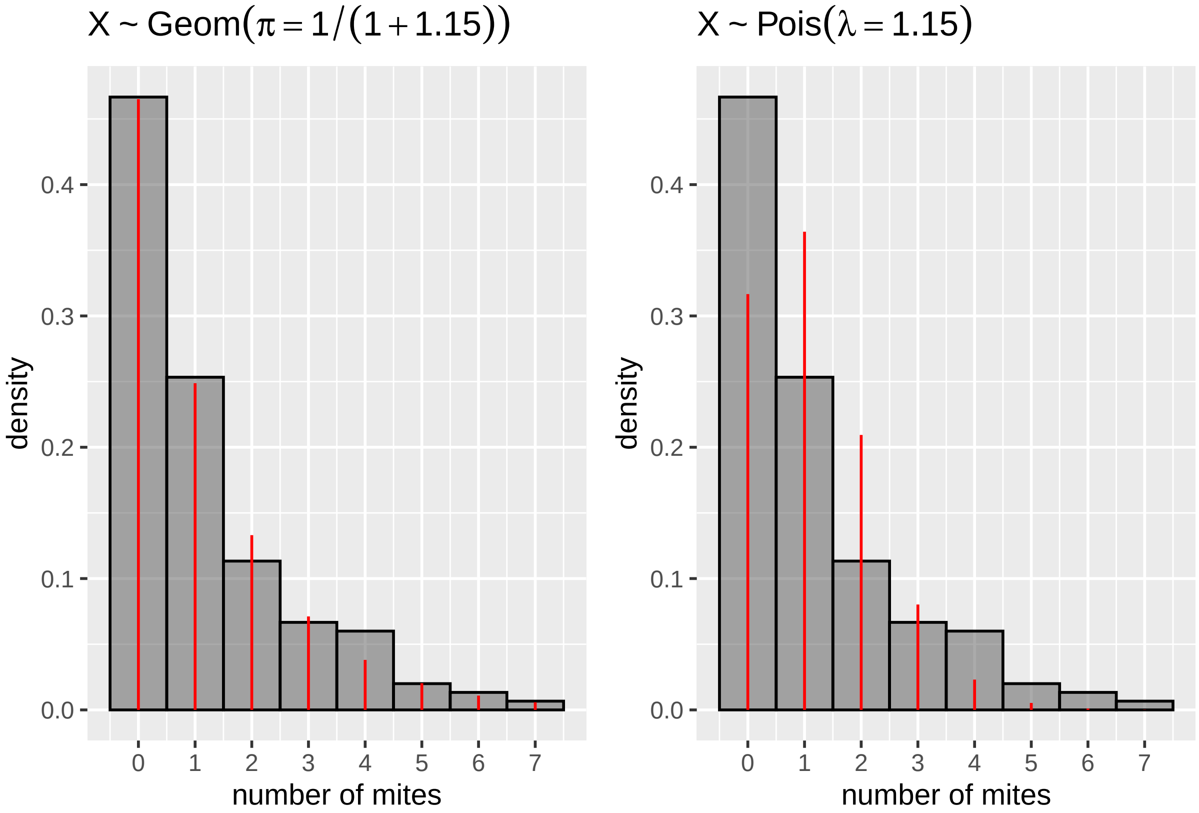
```
#add patchwork package in setup code chunk
```

```
Model_fits <- tibble(
  x = 0:7,
  f1 = dgeom(x, prob = 1/(1+1.15) ),
  f2 =dpois(0:7, lambda = 1.15)
)

p1 <- ggplot()+
  geom_histogram(data = apple_trees,
                 mapping = aes(x = mites,
                               y = after_stat(density)),
                 binwidth = 1,
                 alpha=0.5,
                 color = "black") +
  geom_segment( data = Model_fits,
                mapping = aes(x = x, xend = x,
                              y = 0, yend =f1),
                color = "red") +
  labs(x = "number of mites",
       title = expression(X %~% Geom(pi == 1/(1+1.15)))) +
  scale_x_continuous(breaks = 0:7)

p2 <- ggplot()+
  geom_histogram(data = apple_trees,
                 mapping = aes(x = mites,
                               y = after_stat(density)),
                 binwidth = 1,
                 alpha=0.5,
                 color = "black") +
  geom_segment( data = Model_fits,
                mapping = aes(x = x, xend = x,
                              y = 0, yend =f2),
                color = "red") +
  labs(x = "number of mites",
       title = expression(X %~% Pois(lambda == 1.15))) +
  scale_x_continuous(breaks = 0:7)

p1 + p2
```



The predicted probabilities for each model is shown in red as vertical lines on the probability histogram. The geometric provides a much better fit as evidenced by how well the predicted probabilities match with the actual data we observe.

“‘