Homework 5

Autumn 2022

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday Nov 9 by midnight PST.
- Please answer the following questions in the order in which they are posed.
- Don't forget to 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.

---

1. (Linear transformations) Suppose that the PMF of X is given by:

x	a	b
$f(x)$	π	$(1-\pi)$

where a and b are some numbers. Answer the following. Be sure to support your answers.

a. Find $E[X]$.

$$\begin{aligned} E[X] &= \sum_x x \cdot f(x) \\ &= a \cdot \pi + b \cdot (1-\pi) \\ &= \pi(a-b) + b \end{aligned}$$

b. Find $Var[X]$.

$$\begin{aligned} Var[X] = E[X^2] - E[X]^2 &= \sum_x x^2 \cdot f(x) - (\sum_x x \cdot f(x))^2 \\ &= \pi a^2 + (1-\pi)b^2 - (\pi(a-b)+b)^2 \\ &= \pi a^2 - \pi b^2 - \pi^2(a-b)^2 - 2\pi b(a-b) \\ &= \pi(a^2-b^2) - \pi^2(a-b)^2 - 2\pi b(a-b) \\ &= \pi(a+b)(a-b) - \pi^2(a-b)^2 - 2\pi b(a-b) \\ &= \pi(a-b)((a+b) - \pi(a-b) - 2b) \\ &= \pi(a-b)((a-b) - \pi(a-b)) \\ &= \pi(a-b)^2(1-\pi) \end{aligned}$$

c. Find $E\left[\frac{X-b}{a-b}\right]$.

$$\begin{aligned}
E[\frac{X-b}{a-b}] &= E[\frac{1}{a-b}X - \frac{b}{a-b}] \\
&= \frac{1}{a-b}E[X] - \frac{b}{a-b} \quad \text{(Lemma 7.1)} \\
&= \frac{1}{a-b}(\pi(a-b)+b) - \frac{b}{a-b} \\
&= \frac{\pi(a-b)+b}{a-b} - \frac{b}{a-b} \\
&= \pi + \frac{b}{a-b} - \frac{b}{a-b} \\
&= \pi
\end{aligned}$$

d. Find $Var\left[\frac{X-b}{a-b}\right]$.

$$\begin{aligned}
Var[\frac{X-b}{a-b}] &= Var[\frac{1}{a-b}X - \frac{b}{a-b}] \\
&= (\frac{1}{a-b})^2 Var[X] \quad \text{(Lemma 7.4)} \\
&= \frac{1}{(a-b)^2}Var[X] \\
&= \frac{1}{(a-b)^2}(\pi(a-b)^2(1-\pi)) \\
&= \pi(1-\pi) \\
&= \pi - \pi^2
\end{aligned}$$

2. (Rh negative) 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 he examines the blood types of successive newborn infants until he finds an Rh negative infant.

a. How many Rh positive infants should they expect to type before they find their first Rh negative? Be sure to set up the problem (random variable, distribution, assumptions you need to make) and clearly **state** any results you are using before just using them.

Random Variable X and the Distribution of X:
Let the random variable X denote the number of failures (Rh positive infant) until the medical researcher obtains 1 success (Rh negative infant). X is distributed negative binomial with the number of successes being 1, and a probability of success being 0.15. $X \sim Binom(s = 1, \pi = 0.15) = Geom(\pi = 0.15)$.

Assumptions made when deciding the distribution:
(i) There are a fixed number of successes, s = 1, and trials will be repeated until we have the 1 success.
(ii) We have only two mutually exclusive outcomes, Rh negative infant (success) and Rh positive infant (failure).
(iii) The probability of a success, $\pi = 0.15$, is the same in each trial.
(iv) Each trial is independent of the other trials, meaning that one infant being Rh positive or Rh negative does not affect any of the other infants being Rh positive or Rh negative.

Calculating E[X]:
Since $X \sim Geom(\pi = 0.15)$ we can use Theorem 8.2 to find $E[X]$. Thus by Theorem 8.2 we see that:
$E[X] = \frac{1-\pi}{\pi} = \frac{1-0.15}{0.15} = \frac{0.85}{0.15} = 5.66667$

b. Calculate the probability that they will type more Rh positive infants than expected. This probability can be calculated exactly as shown in class. Give the closed form expression for this probability and then calculate it in R. Be sure to echo your code chunk.

```
prob_geq_ex = pnbinom(q = 5.6667, size = 1, prob = 0.15, lower.tail = FALSE)
prob_geq_ex
```

```
## [1] 0.3771495
```

$$
\begin{aligned}
P(X > 5.66667) &= P(X \geq 6) \\
&= P(X = 6) + P(X = 7) + P(X = 8) + \dots \\
&= (0.85)^6(0.15) + (0.85)^7(0.15) + (0.85)^8(0.15) + \dots \\
&= (0.85)^6[0.15 + (0.85)(0.15) + (0.85)^2(0.15) + \dots] \\
&= (0.85)^6 \sum_{x=0}^{\infty} (0.15)(0.85)^x = (0.85)^6 \cdot 1 \\
&= 0.37715
\end{aligned}
$$

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

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

a. Let k be some non-negative integer. What is $P(X \geq k)$? (*Hint:* we did this in class, I want you to do it again.)

$$
\begin{aligned}
P(X \geq k) &= P(X = k) + P(X = k + 1) + P(X = k + 2) + \dots \\
&= (1 - \pi)^k(\pi) + (1 - \pi)^{k+1}(\pi) + (1 - \pi)^{k+2}(\pi) + \dots \\
&= (1 - \pi)^k[\pi + (1 - \pi)(\pi) + (1 - \pi)^2(\pi) + \dots] \\
&= (1 - \pi)^k \sum_{x=0}^{\infty} (\pi)(1 - \pi)^x = (1 - \pi)^k \cdot 1 \\
&= (1 - \pi)^k
\end{aligned}
$$

b. Show that for all non-negative integers $x(\geq k)$

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

To show that $P(X \geq x|X \geq k) = P(X \geq x - k)$, observe that we can write $P(X \geq x|X \geq k)$ as:

$$
\begin{aligned}
P(X \geq x|X \geq k) &= \frac{P(X \geq x \cap X \geq k)}{P(X \geq k)} \\
&= \frac{P(X \geq x) + P(X \geq k) - P(X \geq x \cup X \geq k))}{P(X \geq k)} \quad \text{(Theorem 2.2)} \\
&= \frac{P(X \geq x) + P(X \geq k) - P(X \geq k)}{P(X \geq k)} \quad \text{(Since } x \geq k\text{)} \\
&= \frac{P(X \geq x)}{P(X \geq k)} \\
&= \frac{(1 - \pi)^x}{(1 - \pi)^k} \\
&= (1 - \pi)^{x-k} \\
&= P(X \geq x - k)
\end{aligned}
$$

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

Memoryless Property Explained:
Memoryless is an appropriate name for the above property as $P(X \geq x|X \geq k) = P(X \geq x-k)$ shows that the geometric distribution is independent of its past events. When I say "independent of its past events," I mean that the geometric distribution has "no memory" of its previous trials. A brief example of this phenomenon would be flipping a coin until we see our first heads, in this case X can represent the number of tails (failures) we get until our first heads is obtained. Thus, $X \sim Geom(\pi = \frac{1}{2})$. If we have already flipped 6 tails (6 failures) and we decide to make two additional flips, then the probability that we will need to wait more than 8 flips until we see our first heads given that we have already waited 6 flips, is simply the probability we will have to wait for more than another 2 flips. Thus, the geometric distribution "has no memory" of the flips we have already made, it only cares about how long your remaining interval is.

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.

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

total_leaves <- nrow(apple_trees)
mean_mites <- mean(apple_trees$mites)
sd_mites <- sd(apple_trees$mites)
```

a. Print the number of observations (leaves) in the data set along with the mean $\bar{x}$ and standard deviation s of the `mites` variable. Be sure to show your code.

Calculating mean and standard deviation of the mites variable:
The number of observations (leaves) in the data set is 150. The mean of the `mites` variable is $\bar{x} = 1.1466667$ and the standard deviation of the `mites` variable is $s = 1.5078616$.

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

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

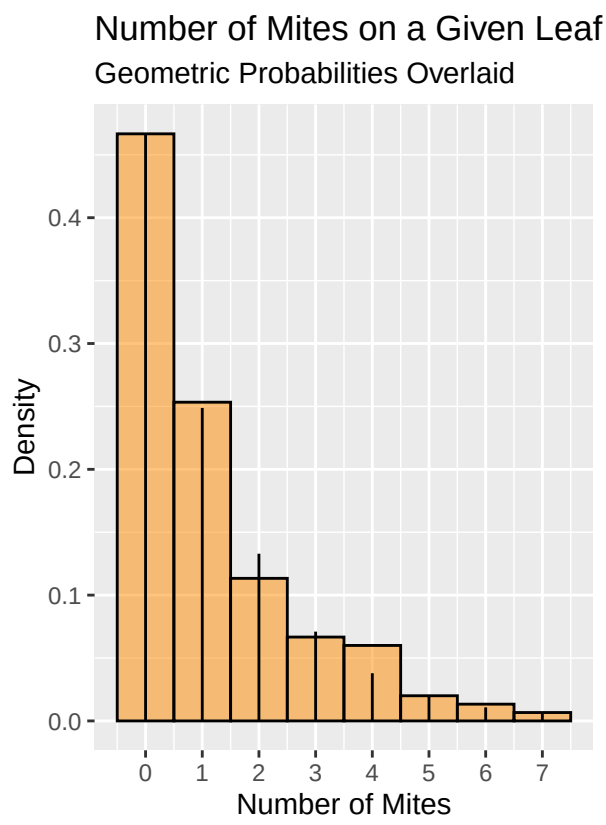
where $\bar{x}$ is the number you calculated from part a. for the mean.

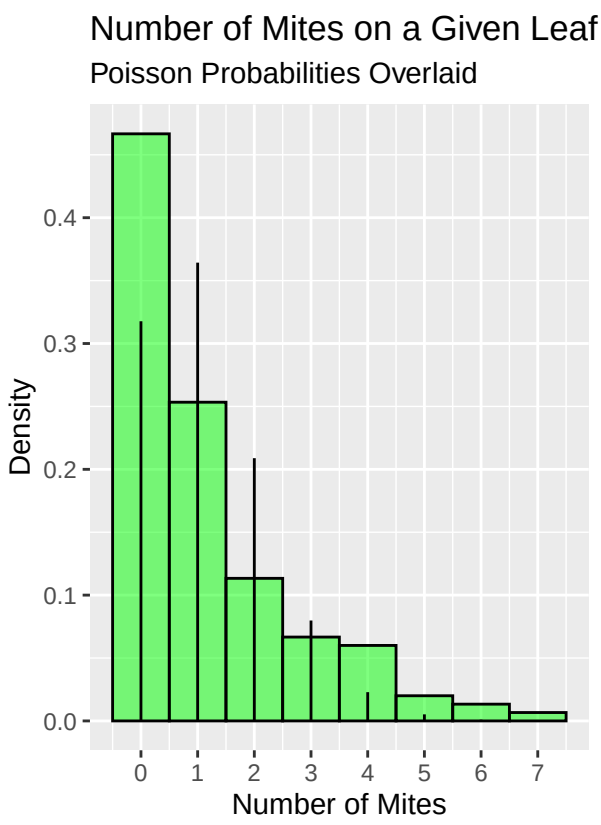
b. Make side-by-side plots showing how well each model fits. Each plot must have the histogram of the data with the probabilities according to the model overlaid on it. Which model appears to fit the data better? (Be sure to echo your code. Don't forget labels, titles etc.)

Hint: For help with the code for side-by-side plots, look at Section 8.1 on the negative binomial. For help on overlaying theoretical probabilities on a histogram, look at section 8.2 on the Poisson distribution.

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

pois_fit <- tibble(
  num_mites = 0:7,
  f = dpois(num_mites, lambda = mean(apple_trees$mites))
)

geom_fit <- tibble(
```

```
  num_mites = 0:7,
  f = dnbinom(x = num_mites, size = 1, prob = 1 / (1 + mean(apple_trees$mites)))
)

pois_plot <- ggplot() +
               geom_histogram(data = apple_trees,
                              mapping = aes(x = mites, y = ..density..),
                              binwidth = 1, fill = "green", color = "black",
                              alpha = 0.5) +
               geom_segment(data = pois_fit,
                            mapping = aes(x = num_mites, xend = num_mites,
                                          y = 0, yend = f)) +
               labs(x = "Number of Mites",
                    y = "Density",
                    title = "Number of Mites on a Given Leaf",
                    subtitle = "Poisson Probabilities Overlaid") +
               scale_x_continuous(breaks = c(0, 1, 2, 3, 4, 5, 6, 7))

geom_plot <- ggplot() +
               geom_histogram(data = apple_trees,
                              mapping = aes(x = mites, y = ..density..),
                              binwidth = 1, fill = "darkorange", color = "black",
                              alpha = 0.5) +
               geom_segment(data = geom_fit,
                            mapping = aes(x = num_mites, xend = num_mites,
                                          y = 0, yend = f)) +
               labs(x = "Number of Mites",
                    y = "Density",
                    title = "Number of Mites on a Given Leaf",
                    subtitle = "Geometric Probabilities Overlaid") +
               scale_x_continuous(breaks = c(0, 1, 2, 3, 4, 5, 6, 7))

geom_plot + pois_plot
```





Which Distribution Fits the Data Better?:
Based on the above graphs with the density histogram of the mites variable with the geometric and poisson probabilities overlaid, it appears that the geometric distribution fits the data better.

5. Free throw Freddy is an 80% shooter. At the end of practice, he shoots until he makes 10 baskets.

a. Let X be the number of misses until he makes 10 baskets. What is the distribution of X? Be sure to give any assumptions you are making.

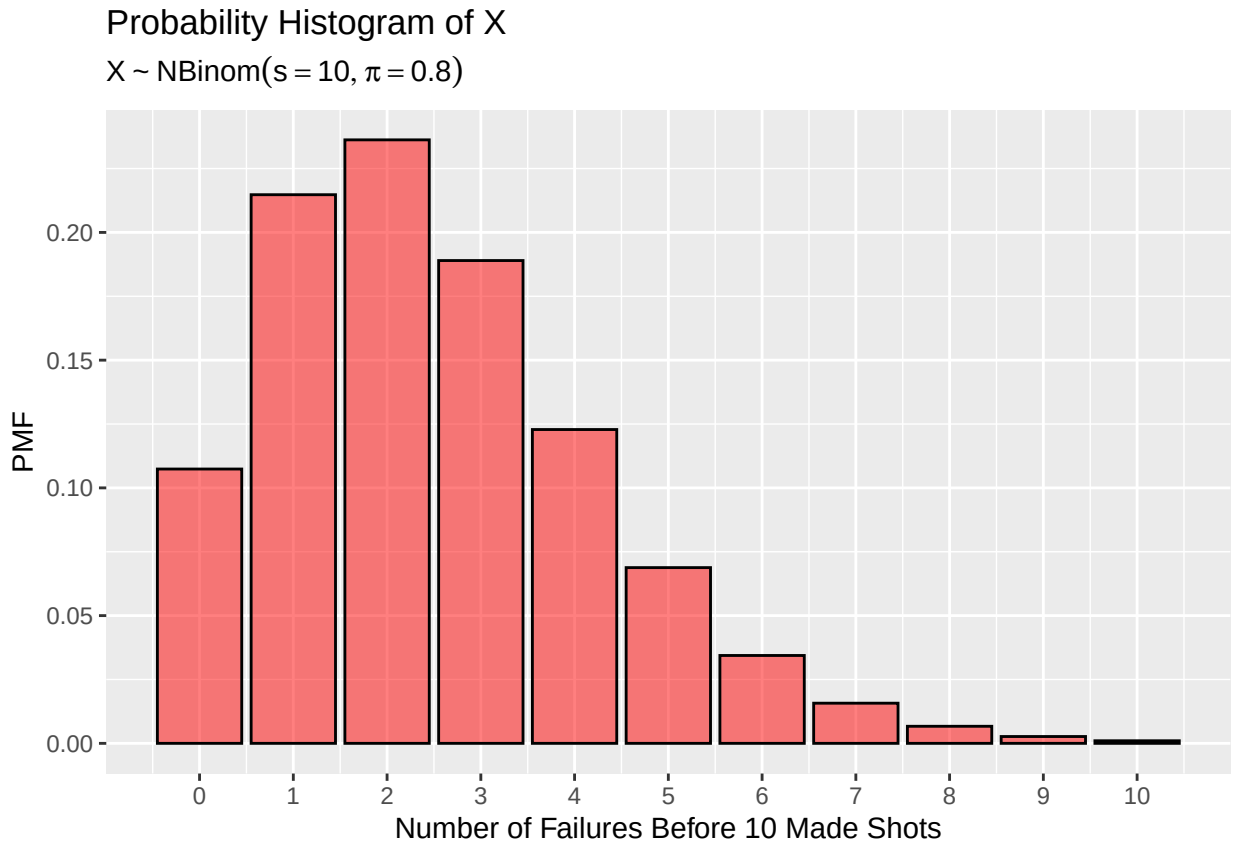
Random Variable X and the Distribution of X:
Let the random variable X denote the number of failures (misses) until Freddy obtains 10 successes (makes). X is distributed negative binomial with the number of successes being 10,and a probability of success being 0.8. $X \sim NBinom(s = 10, \pi = 0.8)$.

Assumptions made when deciding the distribution:
(i) There are a fixed number of successes, s = 10, and trials will be repeated until we have 10 success.
(ii) We have only two mutually exclusive outcomes, Freddy makes the basket (success) and Freddy misses the basket (failure).
(iii) The probability of a success, $\pi = 0.8$, is the same in each trial.
(iv) Each trial is independent of the other trials, meaning that Freddy missing or making a shot won't affect the probability of missing or making any future shots.

b. Draw the probability histogram of X over a reasonable range of values. Be sure to add a title, axis labels. Show your code.

```
nbinom_fit <- tibble(
  num_shots = 0:10,
  f = dnbinom(x = num_shots, size = 10, prob = 0.8)
)
```

```
nbinom_plot <- ggplot() +
               geom_col(data = nbinom_fit,
                             mapping = aes(x = num_shots, y = f),
                             fill = "red", color = "black",
                             alpha = 0.5) +
               labs(x = "Number of Failures Before 10 Made Shots",
                    y = "PMF",
                    title = "Probability Histogram of X",
                    subtitle = expression(X %~% NBinom(s == 10, pi == 0.8))) +
               scale_x_continuous(breaks = c(0:10))

nbinom_plot
```



c. Calculate the probability that Freddy needs 15 or more attempts to make the 10 baskets. Calculate the probability in a code chunk. Show the code. Then write the probability - rounded to 4 digits - in a complete sentence using inline code. (Type ?round for help.)

```
fifteen_prob <- round(1 - pnbinom(q = 4, size = 10, prob = 0.8), 4)
fifteen_prob
```

```
## [1] 0.1298
```

Probability of 15 or more attempts:
The probability that Freddy needs 15 or more attempts to make the 10 baskests is 0.1298.