Homework 4

Autumn 2022

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday Nov 2 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. (Therapy) In the past, a person afflicted with a certain neurological disease had a 30% chance of complete recovery. A radically different therapy has been tested on 10 patients, 7 of whom recovered. Let the random variable X denote the number (in a sample of 10) who recover using the new therapy.

a. What is the distribution of X assuming the new therapy is no better than the old one? State the name of the distribution and also the values of its parameters. Be sure to state any assumptions you are making when deciding on the distribution.

Let the random variable X denote the number (in a sample of 10) who recover using the new therapy. X is distributed binomial with a size of 10, and a probability of success of 0.3, $X \sim Binom(n = 10, \pi = 0.3)$.

Assumptions made when deciding the distribution:

(i) There are a fixed number of trials, n = 10.

(ii) We have only two mutually exclusive outcomes, recovery (success) and non-recovery (failure).

(iii) The probability of a success, $\pi = 0.3$, is the same in each trial.

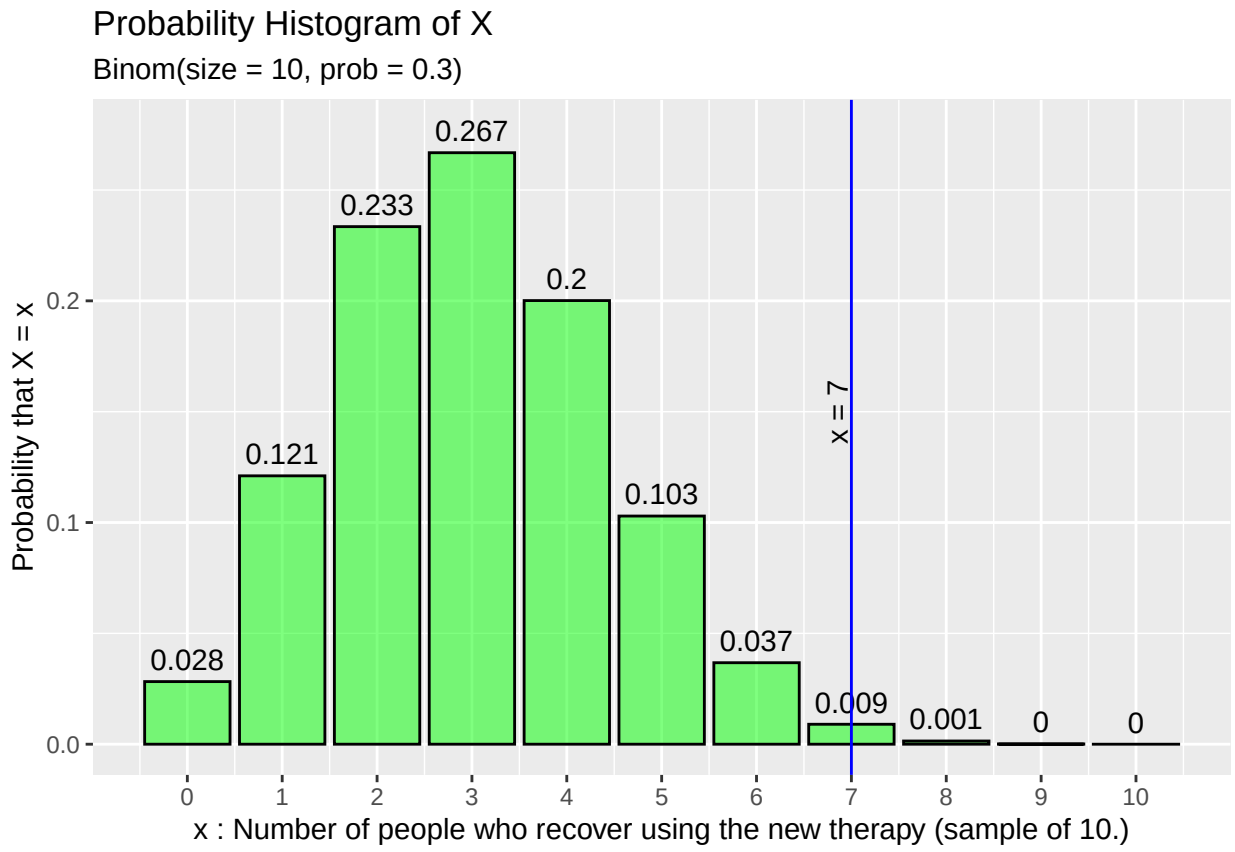
(iv) Each trial is independent of the other trials, meaning that one person recovering/not recovering does not

affect any of the other individuals from recovering or not

b. Make a probability histogram of the distribution in part a. Add a vertical line at $x = 7$ to the histogram using the `geom_vline` layer. Be sure to echo your code chunk. (*Hint* Type `? geom_vline` for help.)

```
library(tidyverse)
disease <- tibble(
  x = 0:10,
  f = dbinom(x, size = 10, prob = 0.3)
)

ggplot(data = disease, mapping = aes(x = x, y = f)) +
  geom_col(fill = "green", color = "black", alpha = 0.5) +
  geom_text(mapping = aes(label = round(f, 3), y = f + 0.01)) +
  labs(x = "x : Number of people who recover using the new therapy (sample of 10.)",
       y = "Probability that X = x",
       title = "Probability Histogram of X",
       subtitle = "Binom(size = 10, prob = 0.3)"
  ) +
  scale_x_continuous(breaks = c(0, 1, 2, 3, 4, 5, 6 , 7 ,8, 9, 10)) +
  geom_vline(xintercept = 7, color = "blue", size = 0.5)  +
  annotate("text", x = 6.85, y = 0.15, label = "x = 7", angle = 90)
```



c. How *extreme* is a value of 7 under the presumed distribution in part a? The one-sided p-value which we will learn more about in STAT 342 is $P(X \geq 7)$. Calculate the one-sided p-value and report your answer in a complete sentence using inline code.

```
prob_geq_7 <- pbinom(6, size = 10, prob = 0.3, lower.tail = FALSE)
```

The extremity of the value 7 under the presumed distribution in part a is represented by $P(X \geq 7)$. This one-sided p-value is: $P(X \geq 7) = 0.0105921$.

2. (Pooling blood) Suppose that fifty people are to be given a blood test to see who has a certain disease. The obvious laboratory procedure is to examine each person's blood individually, meaning that fifty tests would eventually be run. An alternative strategy is to divide each person's blood sample into two parts—say, A and B. All of the A's would then be mixed together and treated as one sample. If that "pooled" sample proved to be negative for the disease, all fifty individuals must necessarily be free of the infection, and no further testing would need to be done. If the pooled sample gave a positive reading, of course, all fifty B samples would have to be analyzed separately.

Let the random variable X denote the number of tests which will need to be performed. Also let π denote

the probability that a randomly selected person is infected with the disease.

a. Write the PMF of X in a tabular format. You may assume independence of outcomes among people. (*Hint: X only has 2 values: 1, 51*).

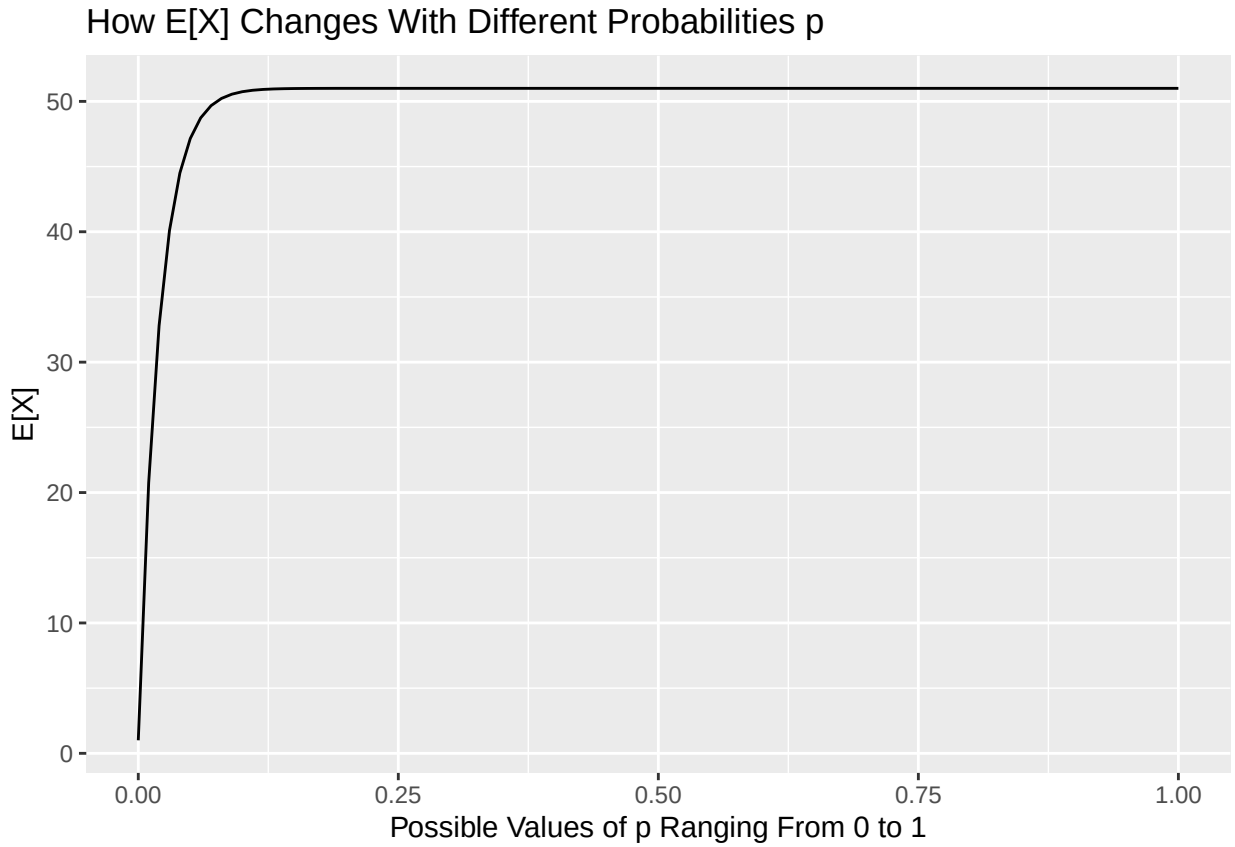
x	1	51
$f(x)$	$(1-\pi)^{50}$	$1-(1-\pi)^{50}$

b. Give an expression for $E[X]$. Show your steps beginning with the definition of an expected value.

$$\begin{aligned} E[X] &= \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x) \\ &= 1(1-\pi)^{50} + 51(1-(1-\pi)^{50}) \\ &= (1-\pi)^{50} + 51 - 51(1-\pi)^{50} \\ &= 51 - 50(1-\pi)^{50} \end{aligned}$$

c. Make a line plot of $E[X]$ versus p. Be sure to echo the code chunk. Does the graph make sense intuitively? Explain briefly. (You can use the `geom_function` layer as we did to graph the variance of a Bernoulli random variable)

```
library(tidyverse)
ggplot() +
  geom_function(fun = function(x){51 - (50 * (1-x)^50)}, xlim = c(0, 1)) +
  labs(x = "Possible Values of p Ranging From 0 to 1",
       y =  "E[X]",
       title = "How E[X] Changes With Different Probabilities p",
  )
```



Since π denotes the probability that a randomly selected person is infected with the disease, this means that $P(X = 1)$ decreases as π increases. Furthermore, when $P(X = 1)$ decreases, the value of $P(X = 51)$ increases since $f(51) = 1 - (1 - \pi)^{50}$. Since the values of $f(51)$ is increasing while $f(1)$ is decreasing, we will then expect $E[X]$ to start converging to 51 as π gets closer and closer to 1. Therefore, the above graph makes sense intuitively.

3. (Proof) Suppose $X \sim Binom(n, \pi)$.

a. Prove the following identity for any $x = 2, 3, \ldots, n$

$$x \cdot (x-1) \cdot \binom{n}{x} = n \cdot (n-1) \binom{n-2}{x-2}.$$

To show that $x \cdot (x-1) \cdot \binom{n}{x} = n \cdot (n-1)\binom{n-2}{x-2}$. We must start with the left hand side.

$$\begin{aligned} x \cdot (x-1) \cdot \binom{n}{x} &= x \cdot (x-1) \cdot \frac{n!}{x!(n-x)!} \quad \text{(Definition of Binomial Coefficient)} \\ &= \frac{x \cdot (x-1)n!}{x \cdot (x-1) \cdot (x-2)!(n-x)!} \\ &= \frac{n!}{(x-2)!(n-x)!} \\ &= \frac{n \cdot (n-1)(n-2)!}{(x-2)!(n-2-(x-2))!} \\ &= n \cdot (n-1) \cdot \binom{n-2}{x-2} \end{aligned}$$

b. Find $E\left[X(X-1)\right]$. (*Hint: follow the derivation for* $E\left[X\right]$ *in Theorem 7.1 for the binomial random variable*)

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{n} x \cdot (x-1) \cdot f(x) = \sum_{x=0}^{n} x(x-1)\binom{n}{x}\pi^x(1-\pi)^{n-x} \\ &= \sum_{x=2}^{n} x(x-1)\binom{n}{x}\pi^x(1-\pi)^{n-x} \\ &= n(n-1)\pi^2 \sum_{x=2}^{n} \binom{n-2}{x-2}\pi^{x-2}(1-\pi)^{n-x} \quad \text{(By Problem 3a.)} \\ &= n(n-1)\pi^2 \sum_{y=0}^{n-2} \binom{n-2}{y}\pi^{y}(1-\pi)^{n-2-y} \quad \text{(Let } y = x-2\text{)} \\ &= \pi^2(n^2-n) \quad \text{(By Fact 2)} \end{aligned}$$

c. Use your result from part b. to show that $Var(X) = n\pi(1-\pi)$. (*Hint:*

$$\sum x(x-1)f(x) = \sum x^2 f(x) - \sum x f(x).$$

By Definition 7.2 we get that $Var(X) = E[X^2] - E[X]^2$. To find what $Var(X)$ equals we must find $E[X^2]$ and $E[X]^2$.

a.) Find $E[X^2]$:

$$\begin{aligned} E[X^2] &= E[X(X-1)] + E[X] \quad \text{(Problem 3a. and Lemma 7.2)} \\ &= \sum x(x-1)f(x) + \sum x f(x) \quad \text{(Definition 7.1)} \\ &= \pi^2(n^2-n) + n\pi \quad \text{(Problem 3a. and Theorem 7.1)} \end{aligned}$$

b.) Find $E[X]^2$:

$E[X]^2 = (E[X])^2 = (n\pi)^2 = n^2\pi^2$ By Theorem 7.1.

Therefore, $Var(X) = \pi^2(n^2 - n) + n\pi - n^2\pi^2 = n\pi(1 - \pi)$.

4. (Chebyshev) Suppose $X \sim Binom(n, \frac{1}{2})$.

    a. What is the mean μ and standard deviation σ of X?

$E[X] = \mu = n\pi = n \cdot \frac{1}{2} = \frac{n}{2}$

$Var[X] = \sigma^2 = n\pi(1 - \pi) = \frac{n}{2}(1 - \frac{1}{2}) = \frac{n}{4}$

$SD[X] = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{n}{4}} = \frac{\sqrt{n}}{2}$

    b. Use Chebyshev's inequality to find the smallest n in order for

$$P\left(0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2}\right)$$

    to be at least 90%.

(Hint: Show that the event

$$0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2}$$

can be rewritten as

$$|X - \mu| < 0.1 \times \sqrt{n} \times \sigma.$$

Then apply Chebyshev's inequality.)

First we will show that the event the $0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2}$ can be rewritten as $|X - \mu| < 0.1 \times \sqrt{n} \times \sigma$:

$$0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2}$$

$$0.9\mu < X < 1.1\mu$$

$$\mu(1 - 0.1) < X < \mu(1 + 0.1)$$

$$\mu - 0.1\mu < X < \mu + 0.1\mu$$

$$\mu - 0.1 \times \sqrt{n} \times \sigma < X < \mu + 0.1 \times \sqrt{n} \times \sigma$$

$$|X - \mu| < 0.1 \times \sqrt{n} \times \sigma$$

Thus, $P(0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2})$ can be rewritten as $P(|X - \mu| < 0.1 \times \sqrt{n} \times \sigma)$.

Chebychev's inequality states that $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$. This is equivalent to saying $P(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}$.

In our case, we want to know when $1 - \frac{1}{k^2} = 0.9$. Observe that $k = 0.1\sqrt{n}$ and $\frac{1}{k^2} = \frac{1}{(0.1\sqrt{n})^2} = \frac{100}{n}$. Thus to find n:

$$\begin{aligned} 1 - \frac{1}{k^2} &= 0.9 \\ 1 - \frac{100}{n} &= 0.9 \\ 0.1 &= \frac{100}{n} \\ n &= \frac{100}{0.1} \\ n &= 1000 \end{aligned}$$

Therefore, the smallest n in order for $P\left(0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2}\right)$ to be at least 90% is 1000.