Homework 4

Autumn 2023

KEY

2023-11-07

Instructions

- This homework is due in Gradescope on Wednesday Nov 1 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. (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.

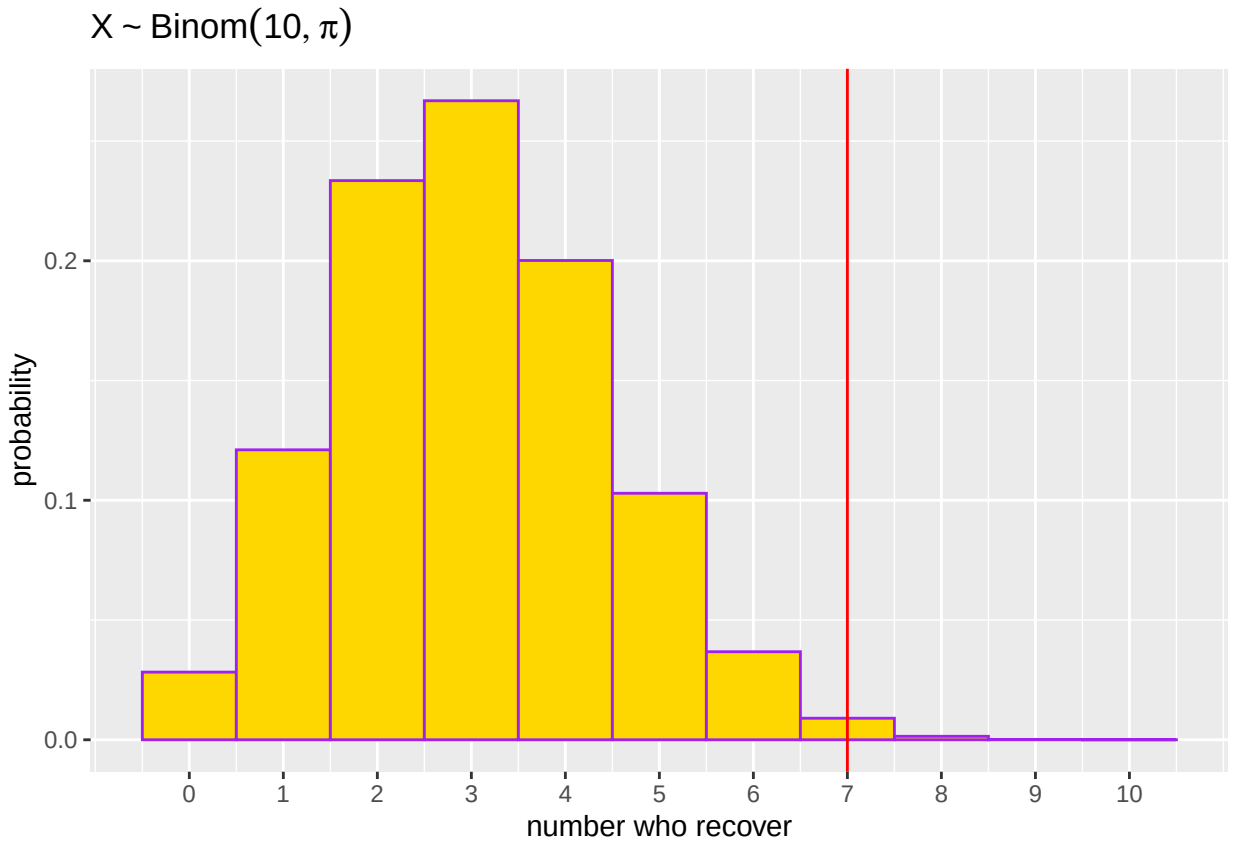
a. What is the distribution of X assuming the new therapy is no better than the old one? State the name of the distribution and also the values of its parameters.

The distribution of X is binomial with $n = 10$ and $\pi = 0.3$. The assumptions of the binomial experiment stated in context are:

- Each trial is the act of testing the therapy on a patient and there are $n = 10$ trials,
- On each trial, there are two possible outcomes: patient recovers (success) and patient does not recover (failure).
- The probability of recovery is the same for each patient.
- Whether or not a patient recovers is independent of another patient.

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. (*Hint* Type `? geom_vline` in the Console for help.)

```r
binom_df <- tibble(
  x = 0:10,
  prob = dbinom(x, size = 10, prob = 0.3)
)
ggplot(data = binom_df,
       mapping = aes(x = x, y = prob))+
  geom_col(width =1,
           fill = "gold",
           color = "purple") +
  geom_vline(xintercept = 7, color = "red")+
  labs(x = "number who recover",
       y = "probability",
       title = expression(X %~% Binom(10, pi)))+
  scale_x_continuous(breaks=0:10)
```



c. How *extreme* is a value of 7 under the presumed distribution in part a? The one-sided P-value is $P(X \geq 7)$. The smaller this probability, the more unusual the value is to have been observed. Calculate this number and report your answer (rounded to 4 decimals) in a complete sentence using inline code.

```r
p_value <- pbinom(q = 6, size = 10, prob = 0.3, lower.tail = F)
```

The p-value corresponding to the observed data of $x = 7$ recoveries is 0.0106. It tells us that the probability of observing something at least as extreme as our observed data is fairly small under the

hypothesized model from part a.

d. Do you believe the presumption that the new therapy is no better than the old one? Why or why not?

An observed $x = 7$ is quite unusual under this assumption since it has a very low probability of being observed. In fact, all but 3 values - 8, 9, 10 - are more unusual. So we can conclude that either we observed something unusual, or our assumption about the success rate with the new therapy being only 30% is not correct. The data certainly points to it being much higher than that.

2. (Pooled Testing) 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*).

Table 1: PMF of X: number of tests

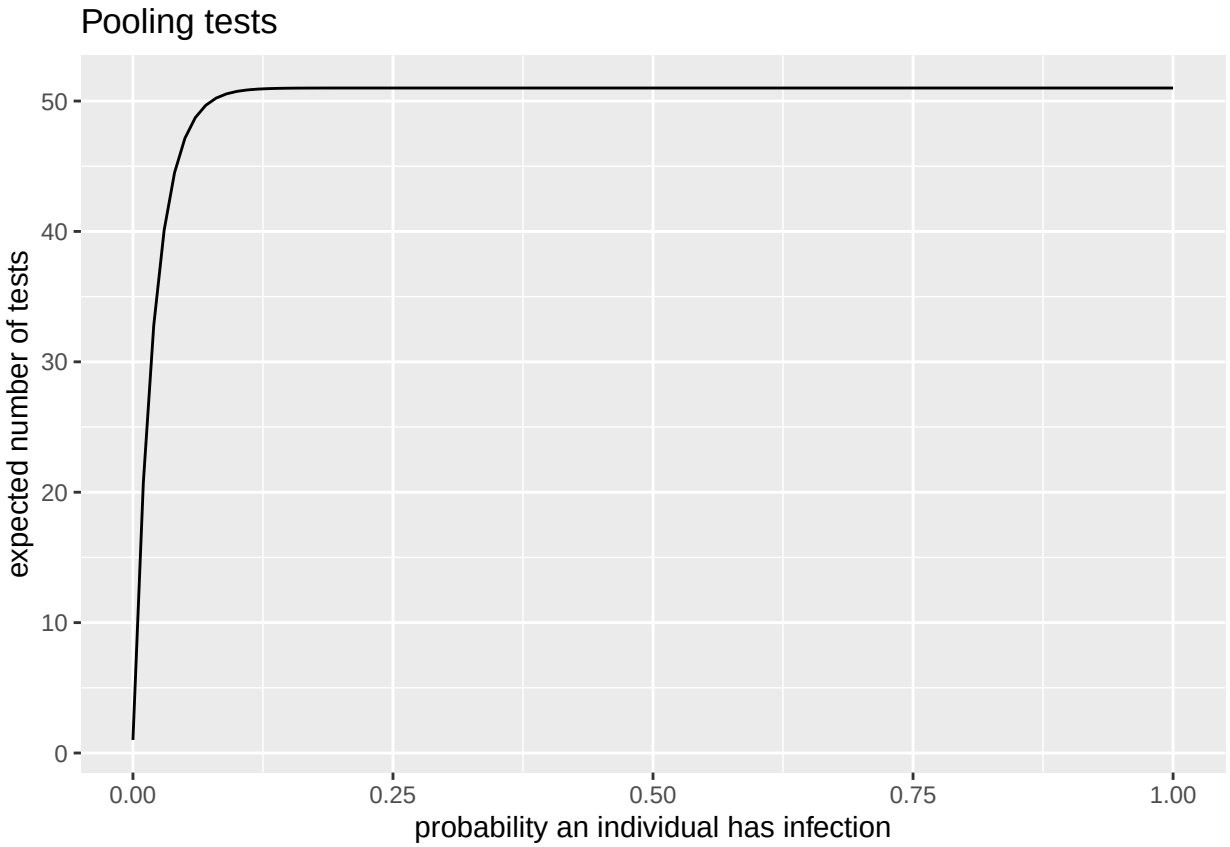
x	1	51
$f(x)$	$(1-\pi)^{50}$	$1-(1-\pi)^{50}$

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

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

c. Make a plot of $E[X]$ versus π. 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 in Example 7.7)

```
ggplot() +
  geom_function( fun = function(x){(1-x)^(50) +51*(1-(1-x)^(50))},
                 xlim=c(0,1) ) +
  labs(x = "probability an individual has infection",
       y = "expected number of tests",
       title = "Pooling tests")
```



Yes, the graph makes sense since it shows that as π, the probability of an individual being infected increases, so will the number of tests we will need to perform on average. Of course, this number tops out at 51 since we will never need to do more than 51 tests.

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

   a. What is the mean μ and standard deviation σ of X? Just cite results and use them.

   Using the formulas from class (Theorem 7.1 and the comment following Example 7.7 concerning the mean and variance of a binomial random variable), we have:

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

   and

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

   Therefore the standard deviation is

$$\sigma = SD[X] = \sqrt{Var[X]} = \frac{\sqrt{n}}{2}.$$

   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 $|X - \frac{n}{2}| < 0.1 \times \sqrt{n} \times \sigma$. Then apply Chebyshev's inequality.)

Since $\mu = \frac{n}{2}$ and $\sigma = \frac{\sqrt{n}}{2}$ we have

$$\begin{aligned}
P(0.9 \times \frac{n}{2} < X < 1.1 \times \frac{n}{2}) &= P(0.9 \times \frac{n}{2} - \frac{n}{2} < X - \frac{n}{2} < 1.1 \times \frac{n}{2} - \frac{n}{2}), \\
&= P(-0.1\frac{n}{2} < \left(X - \frac{n}{2}\right) < 0.1\frac{n}{2}), \\
&= P(|X - \frac{n}{2}| < 0.1 \cdot \frac{n}{2}), \\
&= 1 - P\left(|X - \frac{n}{2}| \geq \underbrace{0.1\sqrt{n}}_{k} \underbrace{\frac{\sqrt{n}}{2}}_{\sigma}\right), \\
&\geq 1 - \frac{1}{0.1^2 n}. \text{ Chebyshev's inequality}
\end{aligned}$$

We want to find n so that the right hand side of the last equation is at least 0.9. Therefore we want

$$1 - \frac{1}{0.1^2 n} \geq 0.9 \quad \Rightarrow \frac{1}{0.1^2 n} \leq 0.1 \Rightarrow n \geq \frac{1}{0.1^3} = 1,000.$$

The smallest n is 1,000.

4.(Golfers) Two golfers are playing sudden death to decide a tournament. The first one wins a hole with probability p, the second one wins with probability q, and holes are tied with probability r. Holes are independent, and the game stops the first time someone wins a hole. What is the probability that the first player wins? Show your work step by step and cite any rules/results you use.

Hint: Let X be the number of games it takes for player A to win. Then X can be 1 (win), 2 (tie, win), 3(tie, tie, win), and so on. The probability that player A wins is $P(X = 1 \cup X = 2 \cup \dots)$

Let X be the number of holes it takes for player A to win. Then X can be 1, 2 (tie, win), 3(tie, tie, win), ...

By axiom 3 (countable additivity) we know that

P(Player A Wins) = $P(X = 1 \cup X = 2 \cup \dots) = \sum_{i=1}^{\infty} P(X = i)$ since the events $X = 1$, $X = 2$, etc. are disjoint, meaning only one of them can occur.

The probability $P(X = x)$ that player A wins in $X = x$ games is $r^{x-1}p$, where $x = 1, 2, \dots$ since the first $(x - 1)$ games must be ties and also the games are independent.

Thus P(Player A Wins) = $\sum_{x=1}^{\infty} r^{x-1}p$

So we have that:

$$
\begin{aligned}
P(A) = & \\
& = \sum_{x=1}^{\infty} r^{x-1} p \\
& = \sum_{x=1}^{\infty} r^{x-1} p \\
& = (p + pr + pr^2 + \dots) \\
& = p \frac{1}{1-r}
\end{aligned}
$$

where the last equation follows from an application of the geometric series with $a = p$ and r is the ratio: $\sum_{k=0}^{\infty} ar^k = a + ar + ar^2 \cdots = \frac{a}{1-r}$ for $r < 1$.