Homework # 5

Elliott Pryor

10/5/2020

Problem 1

Statement

Let X be a random variable representing the number of earthquakes in the Mount Saint Helen's region per day.

a. Researchers claim that X can be modeled with a Poisson distribution. Explain why this is (or is not) a reasonable model for X.

b. Let the average number of earthquakes in a given day be 20. Using the Poisson model in (a), what is $E(X)$? Include an interpretation of the expectation in the context of this problem

c. Let Y be the number of earthquakes in the next 12 hrs

   i. What is the distribution of Y? *Be sure to specify parameters and their values*

   ii. What is $P(Y \geq 2)$? Write a mathematical expression for this value and evaluate the expression. *You may use R for the calculation, but provide the code you used to find the desired probability.*

Solution

a. This is a reasonable model. It does not severely break any of the three Poisson criteria. The independence of non-overlapping intervals is the most dubious. Since they are earthquakes, the likelihood of subsequent earthquakes would depend on previous ones. So independence may be questionable. However, the effect of this would be *very* difficult to measure so a reasonable start would be a Poisson distribution. There is no time correlation in earthquakes, so one is not more likely to occur at 6am than 6pm, so chopping up the day into n intervals the probability of one occurrence would be constant λ/n. Then we physically can't have 2 earthquakes happen at the same time, so the third Poisson criteria is met. Also, given that we have an expected number of earthquakes per day, the Poisson distribution is the best one for applying known information. Since we have occurences per unit effort. The Poisson distribution is a reasonable model for X with the given information.

b. For a Poisson distribution $E(X) = \lambda$, so in this case $E(X) = 20$. This means that we expect 20 earthquakes to occur on some day near Mount Saint Helen's.

c. i. Y is the number of earthquakes in 1/2 a day. So since Y = Number of successes in k units of effort implies $Y \sim Poisson(k\lambda)$. From the previous problem (b) we have that $\lambda = 20$ so $Y \sim Poisson(1/2 * 20 =) = Poisson(10)$

   ii. We know that $P(Y) \geq 2 = 1 - P(Y < 2) = 1 - (P(Y = 0) + P(Y = 1))$. We have the $P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$. So we have

$$P(Y) \geq 2 = 1 - (P(Y = 0) + P(Y = 1)) = 1 - (e^{-10} + 10e^{-10}) = 1 - 11e^{-10} = 0.9995006$$

   Where the final step using R as below

```r
1 - 11 * exp(-10)
```

```
## [1] 0.9995006
```

Problem 2

Statement

Suppose scientists expose 100 individual fish eggs from the same fish species to the same amount of a potentially harmful substance and then record how many fish have defects at age 6 weeks. The researchers are interested in estimating the probability of this species of fish developing defects after 6 weeks if they are exposed to this chemical.

a. Let Y be the number of fish at age 6 weeks that develop defects due to this experiment. Specify a reasonable distribution for the random variable Y.

b. Give at least one reason why your choice of distribution in (a) might be inadequate

c. Suppose the researchers inspect the 6-week-old fish one at a time until they observe the fourth fish with defects. Let Z be the number of fish inspected until the fourth fish with defects is identified. One of the researchers claims that it would be reasonable to model Z with a negative binomial distribution.

   i. Specify the appropriate value of r for the proposed model (i.e., write $Z \sim NegativeBinomial(r = ?, p)$

   ii. Explain to the researcher why this model is not appropriate for this situation.

Solution

a. I thing a Binomial distribution is reasonable in this case. Unlike Geometric or Negative Binomial it counts the total number of occurrences, which is what we are looking for: Y is the total number of fish that are defective. Also, whether or not a fish has defects is independent of the other fish. So it fits the binomial distribution criteria. For completeness, A Hypergeometric distribution would not make sense in this case as there isn't a fixed number of fish with defects. This is not a capture-recapture senario so Hypergeometric does not make sense. And Poisson also does not make sense as we don't know the mean number of occurrences.

b. It could be inadequate if the independence assumption is removed. Say depending on the genetic history of the fish some are more susceptible. Also if there is a finite amount of the hazardous substance, so ones exposed near the end of the trial are safer since most of the substance is gone.

c. i. We are looking for the fourth occurence, so $r = 4$. $Z \sim NegativeBinomial(r = 4, p)$

   ii. In this senario, the trials are not independent since the sample size would be large relative to the total population. Since there are a small number of total fish, drawing a fish out would significantly change the probability of selecting a defective fish next time. So the independence assumption is broken, and another distribution would have to be used to account for the finite population.

Problem 3

Statement

A hospital receives 15% of its flu vaccine shipments from Company A and the remainder of its shipments from Company B. Suppose each shipment contains exactly 20 vaccine vials.

- For each of Company A's shipments, suppose 20% of the vials are ineffective
- For each of Company B's shipments, 10% of the vials are ineffective.

- The hospital plans to randomly select a shipment (of exactly 20 vaccine vials) and test 3 randomly selected vials from the shipment.

a. Assuming the 20 vials in the shipment are independently made (i.e., they do not come from the same batch) what is the probability that a shipment from Company A would have exactly 2 ineffective vials out of the 3 vials that are randomly selected and tested? *Be sure to clearly define events and random variables*

b. Suppose another shipment is randomly selected and it is missing its label. That is, you do not know which company the shipment came from.

   i. What is the probability 2 of 3 randomly selected vials from a randomly selected shipment would be ineffective? That is, what is the overall probability that a random selection of 3 vials from a shipment will be ineffective. *Again, be sure to clearly define events and random variables.*

   ii. If 2 of 3 randomly selected vials from a randomly selected shipment are found to be ineffective, what is the probability the shipment came from Company A? *Again, be sure to clearly define events and random variables.*

Solution

a. Let X be the random variable denoting how many of the vials are ineffective. So $S_X = \{0, 1, 2, 3\}$. Since we can assume that they do not come from the same batch, we assume that all the vials are independent. So we use a Binomial distribution to model this $X \sim Binomial(3, 0.2)$ since we have 3 samples so $n = 3$ and the probability that the vial is ineffective is $20\% = 0.2$ (this is our event). Then

$$P(X = 2) = \binom{3}{2}(0.2)^2(1 - 0.2)^{3-2} = \frac{6}{(1)(2)}(0.2)^2(0.8) = 3 * 1/25 * 4/5 = 12/125 = 0.096$$

So there is a 9.6% chance that exactly two of the vials will be ineffective.

b. i. Let A be the event that the shipment is from company A and B be the event that the shipment is from company B. And X be the random variable denoting the number of defective vials in the sample coming from company A, and Y be the random variable denoting the number of defective vials in the sample coming from company B. If Z is the random variable denoting the number of defective samples from either company, we know that the expected number of ineffective vials from any sample is $P(Z = z) = P(X = x)P(A) + P(Y = y)P(B)$ where $x = y = z$. We are looking for $E(Z = 2) = P(X = 2)P(A) + P(Y = 2)P(B)$. We found $P(X = 2)$ in the last problem. We were also given $P(A)$ and $P(B)$ in the problem statement since we know the relative number of shipments from each company. $P(A) = 0.15$ $P(B) = 1 - P(A) = 0.85$. So we need to calculate $P(Y = 2)$ where $Y \sim Binomial(3, 0.1)$.

$$P(Y = 2) = \binom{3}{2}(0.1)^2(1 - 0.1)^{3-2} = \frac{6}{(1)(2)}(0.1)^2(0.9) = 3 * 1/100 * 9/10 = 18/1000 = 0.018$$

So now we have everything we need to calculate the probability. We have

$$E(Z = 2) = 0.096 * 0.15 + 0.018 * 0.85 = 0.0297$$

So the probability that we select two ineffective samples from a random shipment in the inventory is 2.97%.

ii. We are searching for $P(A|Z = 2)$ using the same events and variables as in part i. We can use Bayes rule to simplify this to the following

$$P(A|Z = 2)\frac{P(Z = 2|A)P(A)}{P(Z = 2|A)P(A) + P(Z = 2|B)P(B)}$$

From the previous part, we resolve some notation differences. By the definition of X and Y, we have $P(X = x) = P(Z = x|A)$. Thus we solved the denominator in part i. We also know from

part a and b.i that $P(Z = 2|A) = 0.096$ and $P(A) = 0.15$. So then plugging these values into the above equation results in

$$P(A|Z = 2) = \frac{0.096 * 0.15}{0.0297} = 0.4848$$

So there is a 48.5% chance that if there are two defective vials in a sample of three that they came from company A.

Problem 4

Statement

Consider the function: $p(y) = 1/5$ for $y \in \{3, 4, 5, 6, 7\}$

a. What is missing from p(y) to make it a legitimate probability function?

b. Create a clearly labeled plot for the pmf of Y

c. Create a clearly labeled plot for the CDF of Y

d. What is $Var(Y)$?

e. **Extra Credit** Find $E(\frac{1}{Y^2})$

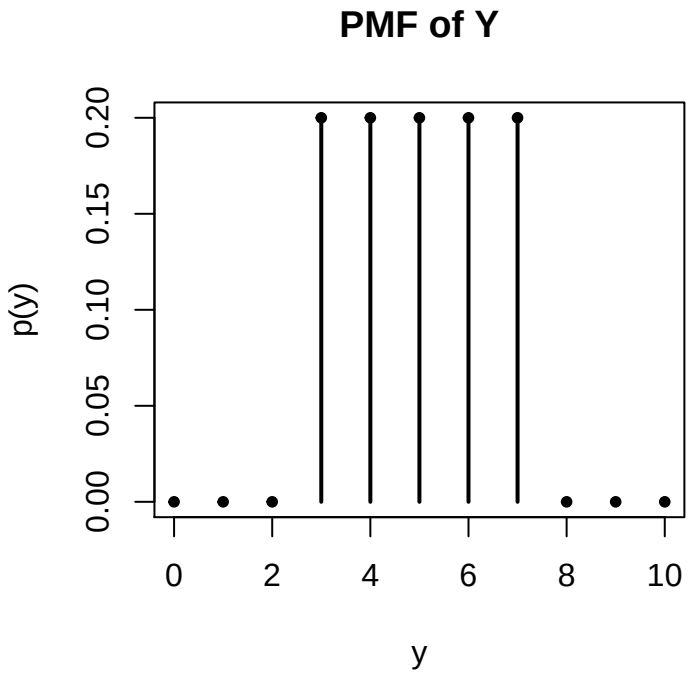
Solution

a. $p(y)$ is not defined for all $y \in \mathbb{R}$. So it needs an 'otherwise 0' So

$$p(y) = \begin{cases} 1/5 & y \in \{3, 4, 5, 6, 7\} \\ 0 & \text{otherwise} \end{cases}$$

b. We create a plot of the pmf.

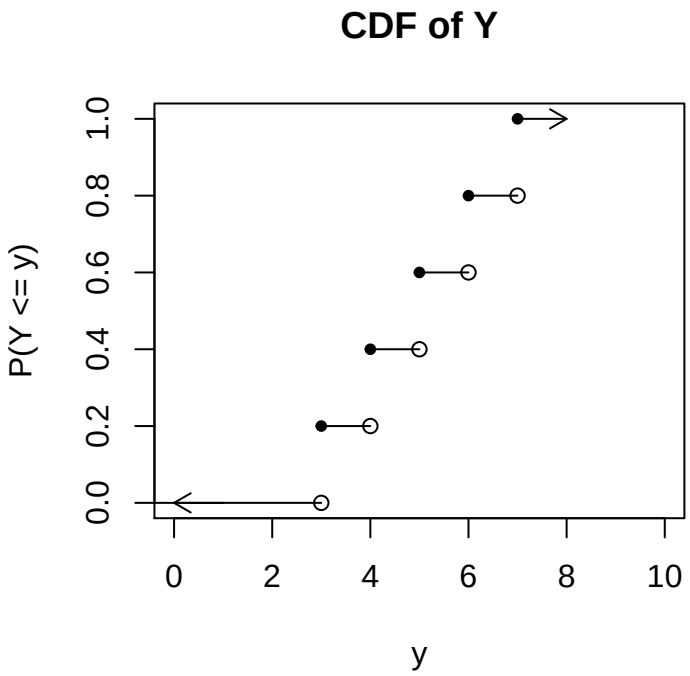
```
pmf <-function(y) {
  p <- 0
  if ((y == 3) | (y == 4) | (y == 5) | (y == 6) | (y == 7)){  # Ugly way to do it but I'm bad at R
      p <- 0.2
  }
  return(p)}

vals <- 0:10
p_vals <- lapply(vals, pmf)

# plot the pmfs (Copied from pmf-plotting-code.R)
plot(vals, p_vals, type = "h", xlab = "y",
     ylab = "p(y)", lwd = 2, main = "PMF of Y")
points(x = vals, y = p_vals, pch = 20)
```



c. We compute the CDF.

```
cdf <-function(y) {
  if(y < 3){
    p <- 0
  } else if ( y < 4){
    p <- 0.2
  } else if (y < 5){
    p <- 0.4
  } else if (y < 6){
    p <- 0.6
  } else if (y < 7){
    p <- 0.8
  } else{
    p <- 1
  }

  return(p)}

vals <- 0:10
p <- lapply(vals, cdf)

# plot the cdf (Copied form HW3 solutions)
plot(x = vals, y = p, type = "n", xlab = "y",
     ylab = "P(Y <= y)", lwd = 2, main = "CDF of Y", ylim = c(0,1))
segments(-5, 0, x1 = 3, y1 = 0)
points(3,0)
arrows(1, 0, x1 = 0, y1 = 0, length = 0.1)

for (i in 4:7){
  segments(i-1, cdf(i-1), x1 = i, y1 = cdf(i-1))
```

```
  points(i, cdf(i-1))
  points((i-1), cdf(i-1), pch = 20)
}
points(7, 1, pch = 20)
arrows(7, 1, x1 = 8, y1 = 1, length = 0.1)
```



d. We can then compute the variance of Y. We know that $Var(Y) = E(Y^2) - E(Y)^2$

First we compute $E(Y) = \sum_{y \in S_Y} y * P(Y = y) = 3 * 0.2 + 4 * 0.2 + 5 * 0.2 + 6 * 0.2 + 7 * 0.2 = 5$ This makes sense since every option is equally weighted so the expected value would be the simple average of $(3, 4, 5, 6, 7) = 5$.

Next we compute $E(Y^2) = \sum_{y \in S_Y} y^2 * P(Y = y) = 9 * 0.2 + 16 * 0.2 + 25 * 0.2 + 36 * 0.2 + 49 * 0.2 = 27$. This makes sense that the average is slightly over 5^2 since the later terms grow faster and have a higher contribution, so the avereage would shift slightly greater.

Now the variance is $Var(Y) = 27 - 25 = 2$

e. **Extra Credit** We compute $E(1/(Y^2))$. We do this the same way that we do it in part d. We also note that $1/y^2$ is defined for all $y \in S_Y$

$E(1/Y^2) = \sum_{y \in S_Y} 1/y^2 * P(Y = y) = 1/9 * 0.2 + 1/16 * 0.2 + 1/25 * 0.2 + 1/36 * 0.2 + 1/49 * 0.2 = 0.05235941$. I attempted to convert this into a fraction, but my calculator cannot and the common denominator is too big for me to reasonably work with. But the expected value is slightly greater than $1/20$. This makes sense for the same reasons in (b) but with the shift to move the average the other direction since we have an inverse quadratic.