Homework 4 Answers

EPI 525

Book exercises

3.38 Stenographer's typos

A very skilled court stenographer makes one typographical error (typo) per hour on average.

a

What are the mean and the standard deviation of the number of typos this stenographer makes in an hour?

Let X be the number of typos this stenographer makes in an hour. Then X can be modeled by a Poisson random variable with $\lambda = 1$ error per hour:

$$X \sim Pois(\lambda = 1)$$

For a Poisson r.v., the mean and variance are both equal to λ:

$$\mu = E(X) = \lambda = 1 \tag{1}$$

$$\sigma = \sqrt{Var(X)} = \sqrt{\lambda} = \sqrt{1} = 1 \tag{2}$$

The mean number of typos the stenographer makes in an hour is 1 with standard deviation 1.

b

Calculate the probability that this stenographer makes at most 3 typos in a given hour.

$$X \sim Pois(\lambda = 1)$$

$$P(X \leq 3) = \sum_{k=0}^{3} \frac{e^{-1}(1^k)}{k!}$$

This can be calculated "directly" in R using the formula:

```
# vector of x values whose probabilities need to be added
x <- 0:3
# vector of respective Poisson prob's of x values
(Poisson_prob_0_3 <- exp(-1)*(1^x)/factorial(x))
```

```
[1] 0.36787944 0.36787944 0.18393972 0.06131324
```

```
# add up the probabilities
sum(Poisson_prob_0_3)
```

```
[1] 0.9810118
```

Using the R command `ppois` we have:

```
(Pleq3 <- ppois(3,1))
```

```
[1] 0.9810118
```

The probability that the stenographer made no more than 3 typos in an hour is 0.981.

Note: the assignment did not specify which way to calculate the probability (ppois vs. formula), and thus either one is correct. Make sure you know how to calculate it both ways though.

c

Calculate the probability that this stenographer makes at least 5 typos over 3 hours.

Let X be the number of typos this stenographer makes in 3 hours. Then X can be modeled by a Poisson random variable with $\lambda = 3$ errors per 3 hours since $\lambda_{t=3} = \lambda_{t=1} \cdot 3 = 3$:

$$X \sim Pois(\lambda = 3)$$

The probability that this stenographer makes at least 5 typos over 3 hours is:

$$\begin{aligned} P(X \geq 5) &= 1 - P(X < 5) \qquad (3)\\ &= 1 - P(X \leq 4) \qquad (4)\\ &= 1 - \sum_{k=0}^{4} \frac{e^{-3}(3)^k}{k!} \qquad (5) \end{aligned}$$

There are various ways this can be calculated.

- Calculated "directly" in R using the formula:

```
# vector of x values whose probabilities need to be added
x <- 0:4
# vector of respective Poisson prob's of x values
(Poisson_prob_0_4 <- exp(-3)*(3^x)/factorial(x))
```

```
[1] 0.04978707 0.14936121 0.22404181 0.22404181 0.16803136
```

```
# add up the probabilities
1 - sum(Poisson_prob_0_4)
```

```
[1] 0.1847368
```

Using the R command `ppois` we have two options depending on whether we use `lower.tail = TRUE` or `lower.tail = FALSE`:

```
# P(X >= 5) for Pois(lambda = 3) random variable
# P(X >= 5) = 1 - P(X <= 4)
```

```r
1 - ppois(q = 4, lambda = 3, lower.tail = TRUE)
```

```
[1] 0.1847368
```

```r
# P(X >= 5) for Pois(lambda = 3) random variable
# P(X >= 5) = P(X > 4)
(Pgeq5 <- ppois(q = 4, lambda = 3, lower.tail = FALSE))
```

```
[1] 0.1847368
```

The probability that the stenographer made 5 or more typos over 3 hours is 0.185.

3.40 Osteosarcoma in NYC

Osteosarcoma is a relatively rare type of bone cancer. It occurs most often in young adults, age 10 - 19; it is diagnosed in approximately 8 per 1,000,000 individuals per year in that age group. In New York City (including all five boroughs), the number of young adults in this age range is approximately 1,400,000.

- Let X be the number of young adults diagnosed with osteosarcoma in NYC in a given year
  - The probability of diagnosis is $p = 8/1,000,000$.
  - The number of young adults in this age range living in New York City is $n = 1,400,000$.
- Then $X \sim Bin(n = 1,400,000, p = \frac{8}{1{,}000{,}000}$ assuming that all of the young adults in NYC get osteosarcoma independently of each other.
- Based on the rule of thumb shown in class ($\frac{1}{10} \leq np(1-p) \leq 10$)
  - X does not satisfy the criteria to be approximated by a Poisson random variable since $np(1-p)$ is bigger than 10:

```r
n <- 1400000
p <- 8/1000000

(npq <- n*p*(1-p))
```

```
[1] 11.19991
```

$$np(1-p) = 1,400,000 \cdot \frac{8}{1,000,000} \cdot \frac{999,992}{1,000,000} = 11.2$$

- However,
  - since
    - the textbook doesn't mention this rule of thumb,
    - 11.2 is only slightly bigger than 10, and
    - this exercise is in the Poisson distribution section of the book,
  - we will proceed to use the Poisson distribution to model osteosarcoma cases.
- But!!
  - if this problem were on the exam,
  - you should check the rule of thumb and then decide that the Binomial distribution is a more appropriate model for these questions.

a

What is the expected number of cases of osteosarcoma in NYC in a given year?

Note: Regardless of whether X is being modeled by a binomial or Poisson distribution, the expected number of cases is np.

Let X be a Poisson random variable with $\lambda = np = 1,400,000 \cdot \frac{8}{1,000,000}$ cases per year:

$$X \sim Pois(\lambda = 1,400,000 \cdot \frac{8}{1,000,000} = 11.2)$$

```
(np <- n*p)
```

```
[1] 11.2
```

- The expected value of a Poisson random variable is λ, which is 11.2.

The expected number of cases of osteosarcoma in NYC amongst young adults in a given year is 11.2.

b

What is the probability that 15 or more cases will be diagnosed in a given year?

We still are using $X \sim Pois(\lambda = 11.2)$.

The probability that 15 or more cases will be diagnosed in a given year is:

$$
\begin{aligned}
P(X \geq 15) &= 1 - P(X < 15) \tag{6} \\
&= 1 - P(X \leq 14) \tag{7} \\
&= 1 - \sum_{k=0}^{14} \frac{e^{-11.2}(11.2)^k}{k!} \tag{8}
\end{aligned}
$$

There are various ways this can be calculated.

- Calculated "directly" in R using the formula:

```
# vector of x values whose probabilities need to be added
x <- 0:14
# vector of respective Poisson prob's of x values
(Poisson_prob_0_14 <- exp(-11.2)*(11.2^x)/factorial(x))
```

```
 [1] 0.0000136742 0.0001531510 0.0008576456 0.0032018768 0.0089652551
 [6] 0.0200821714 0.0374867200 0.0599787520 0.0839702528 0.1044963146
[11] 0.1170358723 0.1191637973 0.1112195441 0.0958199149 0.0766559320
```

```
# add up the probabilities
1 - sum(Poisson_prob_0_14)
```

```
[1] 0.1608991
```

Using the R command `ppois` we have two options depending on whether we use `lower.tail = TRUE` or `lower.tail = FALSE`:

```
# P(X >= 15) for Pois(lambda = 11.2) random variable
# P(X >= 15) = 1 - P(X <= 14)
1 - ppois(q = 14, lambda = 11.2, lower.tail = TRUE)
```

```
[1] 0.1608991
```

```r
# P(X >= 15) for Pois(lambda = 11.2) random variable
# P(X >= 15) = P(X > 14)
(Pgeq15 <- ppois(q = 14, lambda = 11.2, lower.tail = FALSE))
```

```
[1] 0.1608991
```

The probability that 15 or more cases of osteosarcoma will be diagnosed in a given year in NYC amongst young adults is 0.161.

c

The largest concentration of young adults in NYC is in the borough of Brooklyn, where the population in that age range is approximately 450,000. What is the probability of 10 or more cases in Brooklyn in a given year?

First, calculate λ_B given that n=450,000 for Brooklyn.

```r
nB <- 450000
(nBp <- nB*p)
```

```
[1] 3.6
```

$$X_B \sim Pois(\lambda_B = 450,000 \cdot \frac{8}{1,000,000} = 3.6)$$

Then, the probability of 10 or more cases in Brooklyn in a given year is

$$\begin{aligned}
P(X_B \geq 10) &= 1 - P(X_B < 10) \tag{9}\\
&= 1 - P(X_B \leq 9) \tag{10}\\
&= 1 - \sum_{k=0}^{9} \frac{e^{-3.6}(3.6)^k}{k!} \tag{11}
\end{aligned}$$

There are various ways this can be calculated.

- Calculated “directly” in R using the formula:

```
# vector of x values whose probabilities need to be added
x <- 0:9
# vector of respective Poisson prob's of x values
(Poisson_prob_0_9 <- exp(-3.6)*(3.6^x)/factorial(x))
```

```
 [1] 0.027323722 0.098365401 0.177057721 0.212469266 0.191222339 0.137680084
 [7] 0.082608051 0.042484140 0.019117863 0.007647145
```

```
# add up the probabilities
1 - sum(Poisson_prob_0_9)
```

```
[1] 0.004024267
```

Using the R command `ppois` we have two options depending on whether we use `lower.tail = TRUE` or `lower.tail = FALSE`:

```
# P(X >= 10) for Pois(lambda = 3.6) random variable
# P(X >= 10) = 1 - P(X <= 9)
1 - ppois(q = 9, lambda = 3.6, lower.tail = TRUE)
```

```
[1] 0.004024267
```

```
# P(X >= 10) for Pois(lambda = 3.6) random variable
# P(X >= 10) = P(X > 9)
(Pgeq10 <- ppois(q = 9, lambda = 3.6, lower.tail = FALSE))
```

```
[1] 0.004024267
```

The probability of observing 10 or more cases of osteosarcoma will be diagnosed in a given year in Brooklyn in a given year is 0.004.

d

Suppose that in a given year, 10 cases of osteosarcoma were observed in NYC, with all 10 cases occurring among young adults living in Brooklyn. An official from the NYC Public Health Department claims that the probability of this event (that is, the probability of 10 or more cases being observed, and all of them occurring in Brooklyn) is what was calculated in part c). Is the official correct? Explain your answer. You may assume that your answer to part c) is correct. This question can be answered without doing any calculations.

- The official is not correct since the probability calculated in (c) is restricted to only cases within Brooklyn and in particular depended on the size of Brooklyn. This probability ignored all other boroughs.
- If boroughs are independent of each other, then the probability of 10 or more cases being observed, and all of them occurring in Brooklyn is

$$P(Y = 0 \text{ in boroughs other than Brooklyn}) \cdot P(X \geq 10 \text{ in Brooklyn}),$$

where Y is a Poisson distribution with its rate λ dependent on the sample size of all the boroughs excluding Brooklyn.

The official would be correct if $P(Y = 0 \text{ in boroughs other than Brooklyn}) = 1$, but we do not expect this to the case.

e

Suppose that over five years, there was one year in which 10 or more cases of osteosarcoma were observed in Brooklyn. Is the probability of this event equal to the probability calculated in part c)? Explain your answer.

- No, these are not the same probabilities.
- Assuming cases in different years are independent of each other, then the probability of 10 or more cases in Brooklyn in a single year, over 5 different years is modeled by a
  - binomial distribution with
  - $n = 5$ years and
  - p being the probability of 10 or more cases in Brooklyn in a single year.

Note that p is the probability that was calculate in part (c):

$$p = P(X_B \geq 10) = 0.004.$$

Let $Y \sim Bin(n = 5, p = 0.004)$.

Then, the probability that over five years, there was one year in which 10 or more cases of osteosarcoma were observed in Brooklyn is

```r
(Py1 <- choose(5,1) * Pgeq10^1 * (1-Pgeq10)^4)
```

```
[1] 0.01979939
```

$$P(Y = 1) = \binom{5}{1}(0.004)^1(0.996)^{5-1} = 0.02$$

Or using R:

```r
# P(Y = 1) for Bin(n=5, p=Pgeq10) random variable
Pgeq10
```

```
[1] 0.004024267
```

```r
dbinom(x = 1, size = 5, prob = Pgeq10)
```

```
[1] 0.01979939
```

R exercises

R exercise 1

Below you will be using a dataset from Gapminder to complete a few R exercises.

Load all the packages you need below here

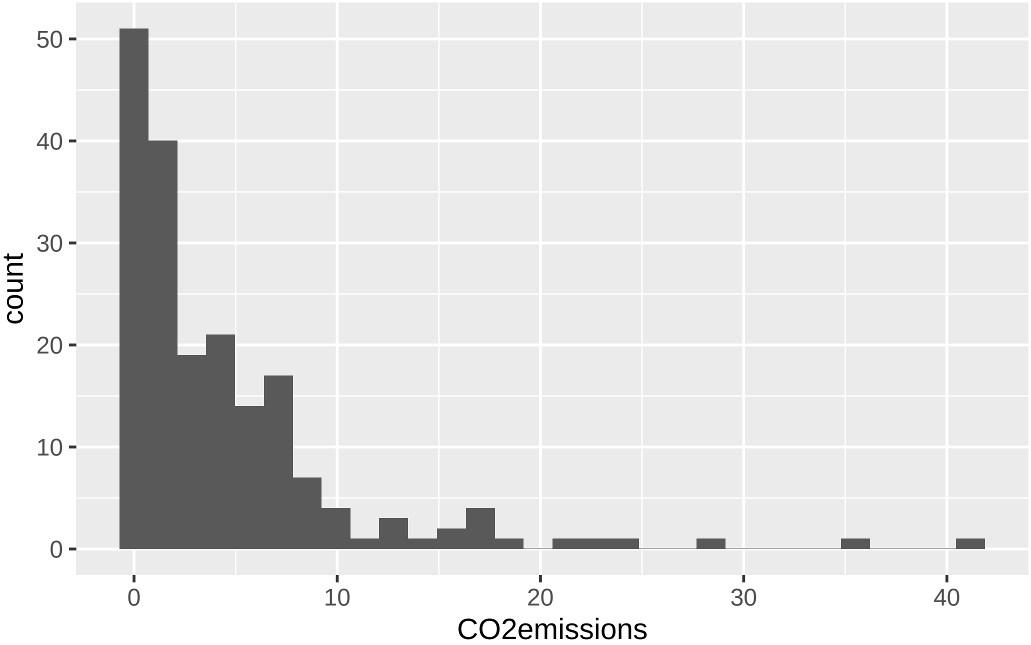
You don't need to do it all at once, you can add more libraries as you realize you need them.

Import dataset

Import the dataset called "Gapminder_2011_LifeExp_CO2.csv" You can find it in the student files under Data then Homework. You will need to download the file onto your computer, and use the correct file path to import the data.

Make a histogram

Using `ggplot2`, make a histogram of the variable `CO2emissions`.


count
CO2emissions