Problem Set 6

Your name and student ID

Today's date

Run this chunk of code to load the autograder package!

Instructions

- Solutions will be released by Wednesday, March 8th
- This semester, homework assignments are for practice only and will not be turned in for marks.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.*

Part I

Oklahoma is not historically known for experiencing earthquakes. Up until 2008, Oklahoma experienced a constant rate of about 1.5 perceptible earthquakes per year on average.

1. Assuming that earthquakes are random and independent, with a constant rate of 1.5 per year, the count of perceptible earthquakes per year in Oklahoma should have a Poisson distribution with mean 1.5. What is the standard deviation of the number of earthquakes per year? Round your answer to 3 decimal places.

```
. = " # BEGIN PROMPT
sd_earthquake <- NULL # YOUR CODE HERE
sd_earthquake
" # END PROMPT


# BEGIN SOLUTION
sd_earthquake <- round(sqrt(1.5), 3)
# END SOLUTION
```

```
test_that("p1a", {
  expect_true(sd_earthquake > 0 & sd_earthquake < 2)
  print("Checking: range of sd_earthquake")
})
```

```
## [1] "Checking: range of sd_earthquake"
## Test passed
```

```
test_that("p1b", {
  expect_true(all.equal(sd_earthquake, 1.225, tol = 0.001))
  print("Checking: value of sd_earthquake")
})
```

```
## [1] "Checking: value of sd_earthquake"
## Test passed
```

2. Using the same assumptions from question 1, use one or two R functions to compute the probability of seeing less than two earthquakes per year. Round your answer to three decimal places.

```
. = " # BEGIN PROMPT
probability <- NULL # YOUR CODE HERE
probability
" # END PROMPT


# BEGIN SOLUTION NO PROMPT
probability <- round(ppois(q = 1, lambda = 1.5), 3)

# See both solution options below:
# Cumulative probability of seeing 1 or less:
# option_1 <- ppois(q = 1, lambda = 1.5)

# Sum of the probability of seeing exactly 1 and the probability of seeing 0:
# option_12 <- dpois(x = 1, lambda = 1.5) + dpois(x = 0, lambda = 1.5)

# END SOLUTION
```

```
test_that("p2a", {
  expect_true(probability > 0 & probability < 1)
  print("Checking: range of probability")
})
```

```
## [1] "Checking: range of probability"
## Test passed
```

```
test_that("p2b", {
  expect_true(all.equal(probability, 0.558, tol = 0.001))
  print("Checking: value of probability")
})
```

```
## [1] "Checking: value of probability"
## Test passed
```

3. Repeat same calculation as above, this time using only a hand calculator. Show your work and round your answer to two decimal places.

$P(X = k) = \frac{e^{-\mu}\mu^k}{k!}$

$P(X = 0) = \frac{e^{-\mu}\mu^0}{0!} = e^{-1.5} = 0.2231302$

$P(X = 1) = \frac{e^{-1.5}1.5^1}{1!} = 0.3346952$

Thus: $P(X < 2) = P(X = 0) + P(X = 1) = 0.2231302 + 0.3346952 = 0.5578254 = 55.78\%$

4. In 2013, Oklahoma experienced 109 perceptible earthquakes (an average of two per week). Assuming the same model as above, write an equation to show how the chance of experiencing 109 earthquakes or more can be written as a function of the probability at or below some k.

(Note: You can write these equations using pen and paper. You can also write the equations using plain text (i.e., P(X >= k)). If you would like to use math equations that render when you knit the pdf, (i.e., $P(X >= k)$) you need to be **very careful** with your symbols. For example, to get the symbol for "greater than or equal to" you cannot copy and paste it into R from the slides or another document. This will cause errors! Instead, you need to write $P(X \geq k)$.

$P(X \geq 109) = 1 - P(X \leq 108)$

5. Use R to calculate the probability of observing 109 perceptible earthquakes or more. Round your answer to the nearest whole number.

```
. = " # BEGIN PROMPT
probability_109_or_more <- NULL # YOUR CODE HERE
probability_109_or_more
" # END PROMPT


# BEGIN SOLUTION
probability_109_or_more <- 0

# See both solution options below:

# solution A (at or above k=109 is equal 1 - at or below k = 108):
# option_1 <- 1 - ppois(q = 108, lambda = 1.5, lower.tail = T)

# solution B (use the upper tail probability at or above 109):
# option_2 <- ppois(q = 109, lambda = 1.5, lower.tail = F)

# END SOLUTION
```

```
test_that("p5a", {
  expect_true(probability_109_or_more >= 0 & probability_109_or_more <= 1)
  print("Checking: range of probability_109_or_more")
})
```

```
## [1] "Checking: range of probability_109_or_more"
## Test passed
```

```
test_that("p5b", {
  expect_true(all.equal(probability_109_or_more, 0, tol = 0.001))
  print("Checking: value of probability_109_or_more")
})
```

```
## [1] "Checking: value of probability_109_or_more"
## Test passed
```

6. Based on your answer to question 5, write a sentence describing the chance of seeing such an event assuming the specified Poisson distribution (i.e., is it rare or common?)

The chance of seeing the event is rare because the probability of the above happening is almost 0.

7. Based on your answer to question 5, would you conclude that the mean number of perceptible earthquakes has increased? Why or why not? Would knowing that the number of perceptible earthquakes was 585 in 2014 support your conclusion?

Yes, the mean number of perceptible earthquakes has increased. The probability of observing such a high number of earthquakes is essentially 0 when the true mean is 1.5 earthquakes per year. Yes, observing 585 earthquakes in 2014 supports the conclusion that the true mean is increasing.

Part II

To track epidemics, the Center for Disease Control and Prevention requires physicians to report all cases of important transmissible diseases. In 2014, a total of 350,062 cases of gonorrhea were officially reported, 53% of which were individuals in their 20s. Assume this 53% stays the same every year. Researchers plan to take a simple random sample of 400 diagnosed cases of gonorrhea to study the risk factors associated with the disease. Call $\hat{p}$ the proportion of cases in the sample corresponding to individuals in their 20s.

8. What is the mean of the sampling distribution of $\hat{p}$ in random samples of size 400?

```
. = " # BEGIN PROMPT
sampling_dist_mean <- NULL # YOUR CODE HERE
sampling_dist_mean
" # END PROMPT


# BEGIN SOLUTION
sampling_dist_mean <- 0.53
# sample mean is an unbiased estimator of $p$
# END SOLUTION
```

```
test_that("p8a", {
  expect_true(sampling_dist_mean >= 0 & sampling_dist_mean <= 1)
  print("Checking: range of sampling_dist_mean")
})
```

```
## [1] "Checking: range of sampling_dist_mean"
## Test passed
```

```
test_that("p8b", {
  expect_true(all.equal(sampling_dist_mean, 0.53, tol = 0.01))
  print("Checking: value of sampling_dist_mean")
})
```

```
## [1] "Checking: value of sampling_dist_mean"
## Test passed
```

mean = 0.53, since the sample mean is an unbiased estimator of p

9. What is the standard deviation of the sampling distribution of $\hat{p}$ in random samples of size 400? Round your answer to 3 decimal places.

```
. = " # BEGIN PROMPT
sampling_dist_sd <- NULL # YOUR CODE HERE
sampling_dist_sd
" # END PROMPT


# BEGIN SOLUTION
sampling_dist_sd <- 0.025
# Standard deviation = sqrt(p(1-p)/n) = sqrt(0.53(1-0.53)/n) = 0.02495496
# The standard deviation is approximately 0.025 when the sample is size 400.
# END SOLUTION
```

```
test_that("p9a", {
  expect_true(sampling_dist_sd >= 0 & sampling_dist_sd <= 1)
  print("Checking: range of sampling_dist_sd")
})
```

```
## [1] "Checking: range of sampling_dist_sd"
## Test passed
```

```
test_that("p9b", {
  expect_true(all.equal(sampling_dist_sd, 0.025, tol = 0.001))
  print("Checking: value of sampling_dist_sd")
})
```

```
## [1] "Checking: value of sampling_dist_sd"
## Test passed
```

standard deviation = sqrt(p(1-p)/n) = sqrt(0.53(1-0.53)/n) = 0.02495496

The standard deviation is approximately 0.025 when the sample is size 400.

10. Describe the conditions required for the sampling distribution of $\hat{p}$ to be Normally distributed. Use the numbers provided in the question to check if the conditions are met.

- The population is expected to be at least 20 times larger than the sample. Using the 2014 data, the population of >350k cases is much much larger that a sample of size 400
- $400 \times 0.53 = 212$, and $400 \times (1 - 0.53) = 188$ are both greater than 10, implying that n is large enough and p is not too rare or too common.
- Yes the conditions are met for the distribution of $\hat{p}$ to be Normally distributed.

END