Assignment 6

Felicia Liu

10/09/2020

- Solutions released: Tuesday, October 13.

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.*

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. [1 point] 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 to the nearest 3 decimal places.

```
sd_earthquake <- round(sqrt(1.5), 3)
sd_earthquake
```

```
## [1] 1.225
```

```
check_problem1()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 1
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

2. [1 point] Making the same assumptions as in part (a), 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.

```
probability <- round(ppois(q = 1, lambda = 1.5), 3)
probability
```

```
## [1] 0.558
```

```
check_problem2()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 2
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

3. [2 points] Do the same calculation as above, this time using only a hand calculator. Show your work and round your final percentage to two decimal places.

$P(X < 2) = P(X = 0) + P(X = 1) = (e\hat{}(-1.5) * 1.5\hat{}0)/(0!) + (e\hat{}(-1.5) * 1.5\hat{}1)/(1!) = 0.56$

4. [1 point] 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 experience 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 and upload the image if you'd like. 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 $P(X \geq k)$. Again, use any of these three methods (hand-written, plain text in R, or "math equations between dollar signs", and you will get points so long as it is human-readable.)

<Note: If you are uploading an image (this is optional), use the following code, or delete if not using.BE SURE TO REMOVE THE OPTION "eval = F" if using this code OR IT WON'T RUN when you knit the file!:>

P(X >= 2) = 1 - P(X <= 1) = 1 - P(X = 0) - P(X = 1)

5. [1 point] Using R, calculate the probability of observing 109 perceptible earthquakes or more. Round your answer to the nearest whole number.

```
probability_109_or_more <- round(1 - ppois(q = 2, lambda = 2), 0)
probability_109_or_more
```

```
## [1] 0
```

```
check_problem5()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: Correct"
##
## Problem 5
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

6. [1 point] Based on your answer to Problem 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 such an event assuming the specified Poisson distribution is rare since the chance is 0.

7. [2 points] Based on your answer in question (e), 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?

I would conclude that the mean number of perceptible earthquakes has increased because there is with 109 perceptible earthquakes a year, the average number increased from 1.5 to 2. With 585 earthquakes in 2015, the mean is approximately 11.25 perceptible earthquakes per week, which would support my conclusion that the mean number of perceptible earthquakes has increased.

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 [1 point] What is the mean of the sampling distribution of $\hat{p}$ in random samples of size 400?

```
sampling_dist_mean <- 0.53
sampling_dist_mean
```

```
## [1] 0.53
```

```
check_problem8()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: Correct"
##
## Problem 8
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

9. [1 point] 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.

```
sampling_dist_sd <- round(sqrt((0.53*0.47)/400), 3)
sampling_dist_sd
```

```
## [1] 0.025
```

```
check_problem9()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 9
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

10. [3 points] 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 likely met.

The central limit theorem states that if you have a population with mean and standard deviation and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. Therefore, the average of our sample means will be the population mean (0.53) and the standard deviation of the sample means equals the standard error of the population mean (0.025).

Read this short article in the New York Times Upshot from 2016. (All Berkeley students should have access to a free NY Times subscription.)

11. [2 points] Explain sampling variation, in the context of this article. Does the 3 percentage point margin of error account for sampling variation?

The 3 percentage point margin of error does account for the sampling variation. This error occurs because surveys are based on only a subset of the full population of likely voters. Even if this sample of respondents is selected randomly from the full population, it is not a perfect representation of attitudes in the full population.

12. [1 point] The authors provides several reasons why the true margin of error is larger than three percent. Describe one of the primary reasons provided in 1-2 sentences.

One of the primary reasons is that there are other forms of error that are not accounted for, such as frame error. Frame error occurs when there is a mismatch between the people who are possibly included in the poll (the sampling frame) and the true target population.

13. [1 point] Based on the information in article, if we're doing a study in public health, choose the answer that is most correct:

    (a) The confidence interval accounts for random error only. If a study suffers from other sources of bias (i.e., confounding, or mismeasurement) the CI will not account for this limitation.
    (b) Increasing the sample size will reduce the chance of other sources of bias (i.e., confounding, or mismeasurement), which is why a larger sample is better.
    (c) both (a) and (b)
    d) neither (a) or (b)

Assign your letter choice as a string. Example: nytimes_answer <- "c"

```
nytimes_answer <- "a"
nytimes_answer
```

```
## [1] "a"
```

```
check_problem13()
```

```
## [1] "Checkpoint 1 Passed: Correct"
##
## Problem 13
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Note: This next section is from the lecture taught on October 14th. The notes should be available soon (if they aren't already)

Deer mice are small rodents native in North America. Their adult body lengths (excluding tail) are known to vary approximately Normally, with mean $\mu = 86$ mm and standard deviation $\sigma = 8$ mm. It is suspected that depending on their environment, deer mice may adapt and deviate from these usual lengths. A random sample of $n = 14$ deer mice in a rich forest habitat gives an average body length of $\bar{x} = 91.1$ mm. Assume that the standard deviation σ of all deer mice in this area is 8 mm.

14. [1 point] Calculate a 99% confidence interval based on this information (you can use R as a calculator to perform the calculation, or use a hand calculator). Round your final values to three decimal places.

```
lower_tail <- round(91.1 - 2.576*8/sqrt(14), 3)
upper_tail <- round(91.1 + 2.576*8/sqrt(14), 3)
ci_99 <- c(lower_tail, upper_tail)
ci_99
```

```
## [1] 85.592 96.608
```

```
check_problem14()
```

```
## [1] "Checkpoint 1 Passed: Both your answer are numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 14
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

15. [1 point] Interpret the confidence interval from Problem 14.

If our model assumptions are correct and there is only random error affecting the estimate, then 95% of the intervals we make will contain the true value μ.

16. [2 points] Suppose deer mice researchers thought your CI was too wide to be useful. Given that you cannot change the standard deviation, what two things could you do to provide a narrower confidence interval?

Decrease the confidence interval or increase the sample size n.

17. [1 point] You decide to create a 95% confidence interval, rather than a 99% confidence interval. Perform this calculation below and round your answer to 3 decimal places.

```
lower_tail95 <- round(91.1 - 1.96*8/sqrt(14), 3)
upper_tail95 <- round(91.1 + 1.96*8/sqrt(14), 3)
ci_95 <- c(lower_tail95, upper_tail95)
ci_95
```

```
## [1] 86.909 95.291
```

```
check_problem17()
```

```
## [1] "Checkpoint 1 Passed: Both your answer are numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 17
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

18. [2 points] Based on this 95% CI, is there evidence against the hypothesis H_0 that these mice have a significantly different mean length compared to the population described in the first part of the question? Without performing a calculation, what amounts do you know the p-value to be bounded between for a two-sided hypothesis test of H_0?

Hint: Use information from question 17 and from question 14.

There is not evidence against the hypotheses H_0 that these mice have a significantly different mean length compared to the population described in the first part of the question because the population mean is 86 mm, which is out of the confidence interval of 95% range. I know that the p-value to be bounded between 86.909 and 95.291 for a two-sided hypothesis test of H_0.

We want to perform a z-test with the two-sided alternative hypothesis the true mean length is not equal to 86mm. In the next four problems, we will conduct a z-test step by step.

19. [1 point] Write out the null and alternative hypotheses for the above problem using notation.

H_0: $\mu = 86$ mm

H_A: μ is not equal to 86 mm

20. [1 point] Calculate the z test statistic. Round your answer to 3 decimal places.

```
z_stat <- round((91.1 - 86)/(8/sqrt(14)), 3)
z_stat
```

```
## [1] 2.385
```

```
check_problem20()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 20
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

21. [1 point] Calculate the p-value as a decimal. Round your answer to 3 decimal places.

```
p_val <- round(pnorm(q = 91.1, mean = 86, sd = 8/sqrt(14), lower.tail = F)*2, 3)
p_val
```

```
## [1] 0.017
```

```
check_problem21()
```

```
## [1] "Checkpoint 1 Passed: You answer is numeric"
## [1] "Checkpoint 2 Passed: You have rounded the answer to 3 decimals"
## [1] "Checkpoint 3 Passed: Correct"
##
## Problem 21
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

22. [1 point] Interpret your above p-value.

There is a 1.7% chance of observing a sample mean of this value or more extreme (in either direction) under the null hypothesis. Thus, this sample mean could be chosen under the null hypothesis and there is no evidence against the null.

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##              Test Points_Possible       Type
## Problem 1  PASSED               1 autograded
## Problem 2  PASSED               1 autograded
## Problem 3  FAILED               0 autograded
## Problem 4  FAILED               0 autograded
## Problem 5  PASSED               1 autograded
## Problem 6  FAILED               0 autograded
## Problem 7  FAILED               0 autograded
## Problem 8  PASSED               1 autograded
## Problem 9  PASSED               1 autograded
## Problem 10 FAILED               0 autograded
## Problem 11 FAILED               0 autograded
## Problem 12 FAILED               0 autograded
## Problem 13 PASSED               1 autograded
## Problem 14 PASSED               1 autograded
## Problem 15 FAILED               0 autograded
## Problem 16 FAILED               0 autograded
## Problem 17 PASSED               1 autograded
## Problem 18 FAILED               0 autograded
## Problem 19 FAILED               0 autograded
## Problem 20 PASSED               1 autograded
## Problem 21 PASSED               1 autograded
## Problem 22 FAILED               0 autograded
```