Assignment 6

Your name and student ID

Today's date

```
BEGIN ASSIGNMENT
requirements: requirements.R
generate: true
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(testthat)
```

```
##
## Attaching package: 'testthat'
```

```
## The following object is masked from 'package:dplyr':
##
##     matches
```

Instructions

- Solutions will be released on Tuesday, March 9
- 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.*

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.

```
BEGIN QUESTION
name: p1
manual: false
points: 1
```

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

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

```
## Test ##
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 ##
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. [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.

```
BEGIN QUESTION
name: p2
manual: false
points: 1
```

```
probability <- NULL # YOUR CODE HERE
probability
```

```
## NULL
```

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

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

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

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

```
## Test ##
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. [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.

```
BEGIN QUESTION
name: p3
manual: true
```

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

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

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

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

4. [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!:>

```
BEGIN QUESTION
name: p4
manual: true
```

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

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

```
BEGIN QUESTION
name: p5
manual: false
points: 1
```

```
probability_109_or_more <- NULL # YOUR CODE HERE
probability_109_or_more
```

```
## NULL
```

```
# BEGIN SOLUTION NO PROMPT
probability_109_or_more <- 0
probability_109_or_more
```

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

```
# for GSIs
#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 ##
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 ##
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. [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?)

```
BEGIN QUESTION
name: p6
manual: true
```

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

```
BEGIN QUESTION
name: p7
manual: true
```

[1 point for correct conclusion. 1 point for explanation] 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 my conclusions that the true mean is increasing.

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?

```
BEGIN QUESTION
name: p8
manual: false
points: 1
```

```
sampling_dist_mean <- 0.53 # SOLUTION
sampling_dist_mean
```

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

```
## Test ##
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 ##
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. [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.

```
BEGIN QUESTION
name: p9
manual: false
points: 1
```

```
sampling_dist_sd <- 0.025 # SOLUTION
sampling_dist_sd
```

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

```
## Test ##
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 ##
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. [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.

```
BEGIN QUESTION
name: p10
manual: true
```

- 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.

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?

```
BEGIN QUESTION
name: p11
manual: true
```

Sampling error occurs here because the survey of voters is based on a sample of the total voting population.

Yes, the margin of error accounts for sampling variation.

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.

```
BEGIN QUESTION
name: p12
manual: true
```

Any of:

- "Frame error": mismatch between people who were polled vs. true target population.
- Nonresponse error: likelihood of responding is systematically related to how one would have answered the survey.
- "Analysis error": pollsters are performing the analysis wrong.

Note: students may not use these exact terms, but we're looking for them to describe one of these three errors.

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"

```
BEGIN QUESTION
name: p13
manual: false
points: 1
```

```
nytimes_answer <- "REPLACE WITH a, b, c, or d. Keep the quotes"
nytimes_answer
```

```
## [1] "REPLACE WITH a, b, c, or d. Keep the quotes"
```

```
# BEGIN SOLUTION NO PROMPT
nytimes_answer <- "a"
nytimes_answer
```

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

```
# END SOLUTION
```

```
## Test ##
test_that("p13", {
  expect_true(nytimes_answer == "a")
  print("Checking: nytimes_answer choice")
})
```

```
## [1] "Checking: nytimes_answer choice"
## Test passed
```

Note: (b) is incorrect because increasing sample size does nothing to remove systematic biases like confounding or mismeasurement.

Note: The following section corresponds to lectures 19 and 20 (March 8 and 10, respectively).

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.

```
BEGIN QUESTION
name: p14
manual: false
points: 1
```

```
lower_tail <- "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
upper_tail <- "REPLACE WITH YOUR ANSWER FOR THE UPPER BOUND"
ci_99 <- c(lower_tail, upper_tail)

# BEGIN SOLUTION NO PROMPT
ci_99 <- c(85.592, 96.608)
ci_99
```

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

```
# for the GSIs
known.sigma <- 8
critical.value <- 2.576
lower_sol <- 91.1 - critical.value*(8/sqrt(14)) #lower bound
upper_sol <- 91.1 + critical.value*(8/sqrt(14)) #upper bound

# END SOLUTION
```

```
## Test ##
test_that("p14a", {
  expect_true(all.equal(ci_99[1], 85.592, tol = 0.001))
  print("Checking: first value of ci_99")
})
```

```
## [1] "Checking: first value of ci_99"
## Test passed
```

```
## Test ##
test_that("p14b", {
  expect_true(all.equal(ci_99[2], 96.608, tol = 0.001))
  print("Checking: second value of ci_99")
})
```

```
## [1] "Checking: second value of ci_99"
## Test passed
```

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

```
BEGIN QUESTION
name: p15
manual: true
```

Our 99% CI for this population of deer mice lengths is 85.59mm to 96.61mm. This means that if we were to take 100 samples using this same method, 99 of them would contain the true value μ in the underlying population and 1 would not.

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?

```
BEGIN QUESTION
name: p16
manual: true
```

- Reduce the level of confidence from 99% to 95% or to 90% even.
- Increase the sample size

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.

```
BEGIN QUESTION
name: p17
manual: false
points: 1
```

```
lower_tail95 <- "REPLACE WITH YOUR ANSWER FOR THE LOWER BOUND"
upper_tail95 <- "REPLACE WITH YOUR ANSWER FOR THE UPPER BOUND"
ci_95 <- c(lower_tail95, upper_tail95)

# BEGIN SOLUTION NO PROMPT
ci_95 <- c(86.909, 95.291)
ci_95
```

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

```
# for the GSIs
known.sigma <- 8
critical.value <- 1.96
lower_sol <- 91.1 - critical.value*(8/sqrt(14)) #lower bound
upper_sol <- 91.1 + critical.value*(8/sqrt(14)) #upper bound

# END SOLUTION
```

```
## Test ##
test_that("p17a", {
  expect_true(all.equal(ci_95[1], 86.909, tol = 0.001))
  print("Checking: first value of ci_95")
})
```

```
## [1] "Checking: first value of ci_95"
## Test passed
```

```
## Test ##
test_that("p17b", {
  expect_true(all.equal(ci_95[2], 95.291, tol = 0.001))
  print("Checking: second value of ci_95")
})
```

```
## [1] "Checking: second value of ci_95"
## 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.

```
BEGIN QUESTION
name: p18
manual: true
```

The 95% confidence interval is from 86.91mm to 95.29mm. Thus, there is evidence against $H_0 : \mu = 86$, because 86mm is not contained within this 95% confidence level. We know that the p-value is greater then 0.01 but less than 0.05 because 86mm is outside of the 95% confidence interval but inside the 99% confidence interval.

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.

```
BEGIN QUESTION
name: p19
manual: true
```

Null: $H_0 : \mu = 86$. Alternative: $H_a : \mu \neq 86$

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

```
BEGIN QUESTION
name: p20
manual: false
points: 1
```

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

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

```
## Test ##
test_that("p20a", {
  expect_true(z_stat > 0 & z_stat <= 5)
  print("Checking: range of z_stat")
})
```

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

```
## Test ##
test_that("p20b", {
  expect_true(all.equal(z_stat, 2.385, tol = 0.001))
  print("Checking: value of z_stat")
})
```

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

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{91.1-86}{8/\sqrt{14}} = 2.385307$

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

```
BEGIN QUESTION
name: p21
manual: false
points: 1
```

```
p_val <- round(2*pnorm(2.385, mean = 0, sd = 1, lower.tail = F),3) # SOLUTION
p_val
```

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

```
## Test ##
test_that("p21a", {
  expect_true(p_val >= 0 & p_val <= 1)
  print("Checking: range of p_val")
})
```

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

```
## Test ##
test_that("p21b", {
  expect_true(all.equal(p_val, 0.017, tol = 0.001))
  print("Checking: value of p_val")
})
```

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

p-value $= 0.017 = 1.7\%$

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

```
BEGIN QUESTION
name: p22
manual: true
```

Thus, there is a 1.7% chance of seeing a result this or more extreme under the null hypothesis of no difference in the means.