Problem Set 6

Your name and student ID

Today's date

```
library(dplyr)
```

```
##
## 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 Friday, October 6th.
- This semester, problem sets 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!
- To avoid code running off the page, have a look at your knitted PDF and ensure all the code fits in the file. If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

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 3 decimal places.

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

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p1.R")
```

```
##
## All tests passed!
```

2. [1 point] Using the same assumptions from 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
```

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p2.R")
```

```
##
## All tests passed!
```

3. Repeat the same calculation as above, this time using only a scientific calculator. Show your work and round your final percentage 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 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 $P(X \geq k)$.

<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!>

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

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

```
# 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 = 108, lambda = 1.5, lower.tail = F)
probability_109_or_more <- 0
probability_109_or_more
```

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p5.R")
```

```
##
## All tests passed!
```

6. 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 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?

[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 whom 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
# sample mean is an unbiased estimator of $p$
sampling_dist_mean
```

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

```
. = ottr::check("tests/p8.R")
```

```
##
## All tests 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 <- 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.
sampling_dist_sd
```

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p9.R")
```

```
##
## All tests passed!
```

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 likely 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.

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. Explain sampling variation in the context of this article. Does the 3 percentage point margin of error account for sampling variation?

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

12. The authors provide several reasons as to why the true margin of error is larger than three percent. Describe one of the primary reasons provided in 1-2 sentences.

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 the 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"
```

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p13.R")
```

```
##
## All tests passed!
```

Deer mice are small rodents native in North America. Their adult body lengths (excluding their tails) 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 scientific calculator). Round your final values to three decimal places.

```
ci_99 <- c(85.592, 96.608)

known.sigma <- 8
critical.value <- 2.576
lower_tail <- 91.1 - critical.value*(8/sqrt(14))
upper_tail <- 91.1 + critical.value*(8/sqrt(14))
lower_tail
```

```
## [1] 85.59228
```

```
upper_tail
```

```
## [1] 96.60772
```

```
ci_99
```

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p14.R")
```

```
##
## All tests passed!
```

15. Interpret the confidence interval from Problem 14.

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 of the samples would not.

16. 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?

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

17. [1 point] You decide to calculate a 95% confidence interval rather than use the 99% confidence interval you just calculated. Perform this calculation below and round your answer to 3 decimal places.

```
ci_95 <- c(86.909, 95.291)

known.sigma <- 8
critical.value <- 1.96
lower_tail95 <- 91.1 - critical.value*(8/sqrt(14))
upper_tail95 <- 91.1 + critical.value*(8/sqrt(14))

lower_tail95
```

```
## [1] 86.90934
```

```
upper_tail95
```

```
## [1] 95.29066
```

```
ci_95
```

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p17.R")
```

```
##
## All tests passed!
```

18. Based on this 95% CI, is there evidence against the null 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 is the range of values for this p-value in a two-sided hypothesis test of H_0?

Hint: Use information from questions 14 and 17.

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 that the true mean length is not equal to 86mm. In the next four problems, we will conduct a z-test step by step.

19. Write the null and alternative hypotheses for the above problem using statistical notation.

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.

```
z_stat <- round((91.1-86)/(8/sqrt(14)), 3)
# $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{91.1-86}{8/\sqrt{14}} = 2.385307$
z_stat
```

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p20.R")
```

```
##
## All tests passed!
```

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

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

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

```
# YOUR CODE HERE
```

```
. = ottr::check("tests/p21.R")
```

```
##
## All tests passed!
```

22. Interpret your above p-value in the context of this problem.

Assuming the null hypothesis is true (that the mean length of deer mice is 86 mm), there is a 1.7% chance of seeing deer mice with a mean length this extreme or more extreme (91.1 mm or greater). Thus, there is evidence to reject the null that the mean deer mice length is 86 mm.

Central Limit Theorem (CLT)

After a vaccine is created for SARS-CoV-2, the next important step would be understanding how many Americans will actually get the vaccine.

23. Suppose we want to estimate the proportion of Americans who would get the vaccine if it were available. We interview a random sample of 100 Americans about whether they would choose to be vaccinated if it were an option. Unknown to us, the true population proportion who would be vaccinated is 0.50. What is the expected value and the standard error of the sample proportion?

Note: This sample proportion is only an estimate but reflects the proportion of Americans willing to accept the hypothetical vaccine in a recent study.

$\mathbb{E}[Vaccination] = 0.50$

Standard Error = 0.05

24. Which of the following is an appropriate statement of the central limit theorem? Select just one.

(1) The central limit theorem states that if you take a large random sample from a population and the data in the population are normally distributed, the data in your sample will be normally distributed.

(2) The central limit theorem states that if you take a large random sample from a population, the data in your sample will be normally distributed.

(3) The central limit theorem states that if you take many large random samples from a population and the data in the population are normally distributed, the sample means will be normally distributed.

(4) The central limit theorem states that if you take many large random samples from a population, the sample means will be normally distributed.

(5) The central limit theorem states that if you take many large random samples from a population and the data in the population are normally distributed, the data from the pooled samples will be normally distributed.

(6) The central limit theorem states that if you take many large random samples from a population, the data from the pooled samples will be normally distributed.

(7) is an appropriate statement of the CLT.

25. Fill in the blanks below.

As n increases the estimate $\bar{x}$ gets closer to μ .

END