Problem Set 7: Inference

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 15th (same day as problem set is released to help you study for the exam!).
- 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.**
- It is good practice to not allow your code to run off the page. To avoid this, 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.

- Useful mathematical notation in markdown:

$$\mu$$

$$\sigma$$

Part I: Confidence Intervals

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.

1. 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 PROMPT
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)

" # END PROMPT


# BEGIN SOLUTION
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_that("p1a", {
  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_that("p1b", {
  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
```

2. Interpret the confidence interval from question 1 in the context of this question.

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.

3. 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

4. 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 PROMPT
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)
" # END PROMPT


# BEGIN SOLUTION
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_that("p4a", {
  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_that("p4b", {
  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
```

5. 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 questions 1 and 4.

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.

6. Write the null and alternative hypotheses for the question above.

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

7. Calculate the z test statistic. Round your answer to 3 decimal places.

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


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

```
test_that("p7a", {
  expect_true(z_stat > 0 & z_stat <= 5)
  print("Checking: range of z_stat")
})
```

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

```
test_that("p7b", {
  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
```

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

8. Calculate the p-value as a decimal. Round your answer to 3 decimal places.

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


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

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

```
# END SOLUTION
```

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

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

```
test_that("p8b", {
  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 $= 0.017 = 1.7\%$

9. Interpret the p-value you calculated in the context of this question.

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

Part II: Proportions

Suppose we want to estimate the proportion of Americans who would be willing to get a vaccine for a new strain of COVID-19. 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 of those would be vaccinated is 0.50.

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

10. What is the expected value and the standard error of the sample proportion?

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

Standard Error = 0.05

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

(a) 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.

(b) 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.

(c) 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.

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

(e) 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.

(f) 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.

(d)

12. Fill in the blanks below.

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

μ

Please watch this short video about shifting the population distribution: https://www.youtube.com/watch?v=8BJNzH6_JpU

Read the 2001 reprint of the 1985 article "Sick Individuals and Sick Populations" by Geoffrey Rose.

Some things to think about from this article.

13. What is the issue Rose highlights with exposures that are very common in a population?

Very common exposures cannot explain much regarding the distribution of a disease within a population (because almost everyone is exposed).

14. What are the differences in how the high risk vs population strategies affect the distribution?

High risk intervention strategies truncate the upper tail of the risk distribution by focusing on worst individual cases. The population strategy, on the other hand, causes a downward shift in the entire distribution of risk.

15. The ban on smoking in public places (restaurants, bars, etc.) was argued legally based on the rights of staff in these locations to be free of second hand smoke. The impact has been a shift in the curve of tobacco exposure, smoking and smoking related health outcomes. Name another type or example of an intervention that has been promoted recently (for any outcome) based on this idea of shifting the curve in the population.

Many examples. Recently, hand-washing, mask wearing, getting vaccinated, etc. in relation to COVID-19.