MA256 Lesson 7 - Estimation - How large is the effect? (3.1-3.4)

Chapter 3 Overview

Why should we care about the topics covered in Chapter 3?

So far, we can only say things like "We have strong evidence that the long-run probability Buzz pushes the correct button is larger than 0.5." Is this as helpful as it could be?

1) If we ask.... our answer is...

- Is _____ a plausible value for our parameter? Answer: answer is no (reject the null hypothesis) or yes (reject the null hypothesis)
- What are all plausible values for our parameter? Answer: answer is an interval (range of values)

2) For the buzz study, is there a better answer we can come up with?

If we know that the long-run probability is greater than 0.5, can we get a range of plausible values for the LR probability? Example: we are 95% confident that buzz will select the correct button 63 to 75% of the time.

3) Cutting the chase... when I walk out of class today, what do I need to know from Chapter 3?

good question. The big idea is how we calculate a confidence interval and how different things will influence the CI width.
In general the CI is calculated: $statistic \pm multiplier \times (SD\ of\ statistic)$

4) Making sense of the three methods described in the book:

-1: plausible values method: This method adjusts the parameter value under the null hypothesis utnil the p-value just exceeds the significance level (α). It is loosely related to the equation above.

-2: 2SD: the 95% CI is $statistic \pm 2 \times SD$. The SD can come from simulation or use the theory based equation for proportions.

-3: Theory based approach: For estimating a population proportion (categorical): $\hat{p} \pm multiplier \times \sqrt{\hat{p}(1-\hat{p})/n}$
For estimating a population mean: $\bar{x} \pm multiplier \times s/\sqrt{n}$

5) Looking at the equations above, what four things will affect the size of the confidence interval?

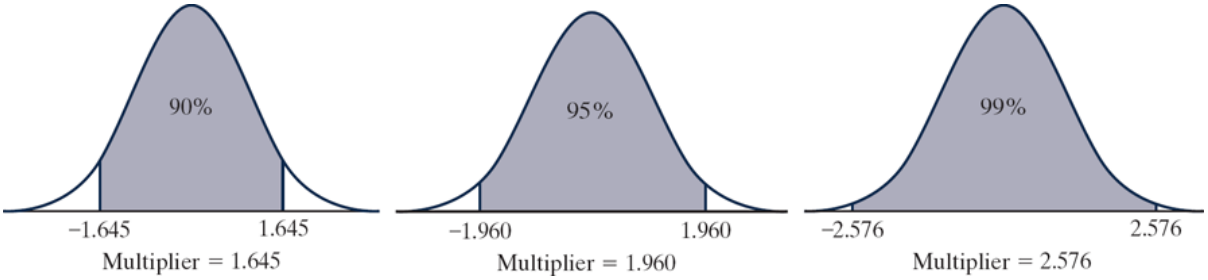
1. Confidence Level (More confidence requires wider CI)
2. Sample size (Larger samples have less SD & create narrower CI)
3. Standard deviation (More SD results in wider CI)
4. Distance $\hat{p}$ is from 0.5 (Farther $\hat{p}$ is from 0.5, less SE & narrower CI)

6) How are the *significance level*, *confidence level*, *Type I error*, and *Type II error* related?

α is what we typically use to refer to the significance level.
$1 - \alpha$ is the confidence level.
A Type I error is "rejecting the null hypothesis when it is actually true". The probability of this occurring is exactly the significance level (α).
A Type II error is "not rejecting the null hypothesis when it is actually false. The probability of this occuring is exactly the statistical power (β - beyond the scope of this course- take MA376.)

Null hypothesis is...	True	False
Rejected	Type I error. FP prob $= \alpha$	Correct Decision. TP prob $= 1 - \beta$
Not Rejected	Correct Decision. TN prob. $= 1 - \alpha$	Type II error. FN prob $= \beta$

7) Where does the "2" in the 2SD method come from? Figure 3.2.3. shows the change in the multiplier with a change in the confidence level. We can see that at the 95% confidence level, the multiplier is roughly 2. Use the `qnorm()` function to calculate the multiplier for a 90%, 95%, and 99% confidence levels. Compare your answers with using the $t - distribution$ for a sample size of $n = 20$. What is our significance level for each of the confidence levels?




```
# 90%
c(qnorm(0.95), qt(0.95, 19))
```

```
## [1] 1.644854 1.729133
```

```
# 95%
c(qnorm(0.975), qt(0.975, 19))
```

```
## [1] 1.959964 2.093024
```

```
# 99%
c(qnorm(0.995), qt(0.995, 19))
```

```
## [1] 2.575829 2.860935
```

the numbers for the t distribution are larger than the normal distribtuion at the same quantile. $\alpha = 0.1, 0.05, 0.01$

8) According to a 2018 report by the U.S. Department of Labor, civilian Americans spend 2.84 hours per day watching television. A faculty researcher, Dr. Sameer, at California Polytechnic State University (Cal Poly) conducts a study to see whether a different average applies to Cal Poly students. Suppose that for a random sample of 100 Cal Poly students, the mean and standard deviation of hours per day spent watching TV turns out to be 3.01 and 1.97 hours, respectively. There is not strong skew.

a) Is our statistic quantitative or categorical?

Quantitative

b) What is the null and alternative hypotheses in words and symbols?

H_0 : The average CalPoly student watches TV for 2.84 hours. $\mu = 2.84$

H_a : The average CalPoly student watches TV for something other than 2.84 hours. $\mu \neq 2.84$

c) In the context of this problem, what is a Type I error? Type II error?

A type I error would be saying that the average CP student doesn't watch TV for 2.84 hours, but they actually do. A type II error would be saying that the average CP student watches TV for 2.84 hours when they actually watch more or less TV.

d) What is the value of our statistic?

$\bar{x} = 3.01$, s=1.97

e) Do we meet our validity conditions?

Yes, we have at least 20 observations ($100 \geq 20$) and the data is not strongly skewed.

f) What is our 95% Confidence Interval for the true mean hours that Cal Poly students spend watching television per day?

```
xbar <- 3.01
s <- 1.97
n <- 100
c(xbar - qt(1 - 0.05/2, 99) * s / sqrt(n), xbar + qt(1 - 0.05/2, 99) * s / sqrt(n))
```

```
## [1] 2.619109 3.400891
```

Confidence Interval $= \bar{x} \pm qt(1 - \frac{\alpha}{2}, n-1) \times \sqrt{\frac{s^2}{n}}$ $= 3.01 \pm qt(1 - \frac{.05}{2}\ ,\ 99) \times \sqrt{\frac{1.97^2}{100}} = (2.6191, 3.4009)$

g) Given our confidence interval above, what do we know about the results of a strength of evidence test with a null hypothesis of $\mu = 2.84$ and an alternate hypothesis of $\mu \neq 2.84$?

We know that the p-value will be greater than 0.05, as 2.84 did "make the cut" and falls within our 95% confidence interval.

h) Report your standardized statistic (t or z) and p-value given the above data and a null hypothesis of $\mu = 2.84$ and an alternate hypothesis of $\mu \neq 2.84$.

```
mu <- 2.84
tstat <- (xbar - mu) / (s / sqrt(n))
pval <- 2*(1-pt(abs(tstat), 99))
c(tstat, pval)
```

```
## [1] 0.8629442 0.3902539
```

$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}} = \frac{3.01-2.84}{\sqrt{\frac{1.97^2}{100}}} = 0.8629$

p-value = 2*(1 - pt(abs(t), n-1)) = 2 * (1 - pt(abs(0.8629442), 99)) = 0.3903

i) Which of the following are INVALID interpretations of the 95% CI?

1. About 95% of all CP students spend between 2.619 and 3.401 hours/day watching TV.
2. There is a 95% chance that, on average, CP students spend between 2.619 and 3.401 hours/day watching TV.
3. We are 95% confident that, on average, these 100 CP students spend between 2.619 and 3.401 hours/day watching TV.
4. In the long run, 95% of the sample means will be between 2.619 and 3.401 hours.

they are all incorrect. A correct interpretation is always interpreted as: I am X% confident that the *insert parameter definition* is between (..., ...).
Option 1: The first option is saying 95% of all students' watch an average amount of TV that falls within the CI. This is not true, since CI's are about the population parameters which are about the overall proportion or average instead of individual observations. Think about 2 students: 1 who watches no TV and another who watches 6 hours. Neither falls into the interval, but their average does, which is what the confidence interval is estimating.
Option 2: Confidence intervals are not probabilities that the population average falls within a range, so 95% chance is not an accurate representation (See FAQ 3.2.1 for a more detailed explanation).
Options 3 & 4. Both are about samples instead of the population, so they do not accurately represent what a confidence interval is. The third option talks about these 100 CalPoly students and the fourth talks about all sample means.

9) Most people are right handed, and even the right eye is dominant for most people. Developmental biologists have suggested that late-stage human embryos tend to turn their heads to the right. In a study reported in Nature (2003), German bio-psychologist Onur Güntürkün conjectured that this tendency to turn to the right manifests itself in other ways as well, so he studied kissing couples to see which side they leaned their heads to while kissing. He and his researchers observed kissing couples in public places such as airports, train stations, beaches, and parks in Germany. They were careful not to include couples who were holding objects such as luggage that might have affected which direction they turned. For each kissing couple observed, the researchers noted whether the couple leaned their heads to the right or to the left. They observed 124 couples, ages 13 to 70 years.

a) Identify the observational units in this study.

Each kissing couple.

b) Identify the variable recorded in this study. Classify it as categorical or quantitative.

Whether each couple leaned right or left while kissing. This is a categorical variable.

c) Suppose we want to know the true long-run proportion of couples that kiss right in Germany. Would this be a statistic or a parameter? What symbol is used to represent this proportion of the population?

This would be a parameter, represented by π

d) Do we know the exact value of the long run proportion of couples kissing right based on the data? Explain.

No, we don't know the exact value of π. The population parameter is an unknown quantity, but we want to infer it using statistics for a best guess.

e) State the appropriate null and alternative hypotheses, both in words and in terms of the parameter π, for testing the conjecture that kissing couples tend to lean their heads to the right more often.

$H_0 : \pi = 50\%$. The null hypothesis is that the long run proportion of couples who lean their heads to the right is 50%.

$H_a : \pi > 50\%$. The alternate hypothesis is that the long run proportion of couples who lean their heads to the right is greater than 50%.

f) Calculate the sample proportion of the observed couples who leaned their heads to the right while kissing. Also indicate the symbol used to denote this value.

```
library(tidyverse)
library(janitor)
kiss <- read_csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Kissing.csv")
kiss %>% count(Direction)
```

```
## # A tibble: 2 x 2
##   Direction     n
##   <chr>     <int>
## 1 Left         44
## 2 Right        80
```

```
phat <- 80/124
phat
```

```
## [1] 0.6451613
```

$\hat{p} = \frac{80}{124} = 0.645$

g) Do we meet the validity conditions to conduct a theoretical test? Justify your answer.

Yes, we meet the validity conditions because 80 right and 44 left are both larger than 10.

h) If we wanted to do strength of evidence testing using theoretical methods, which test would we use?

As we are assessing quantitative (categorical) variables, we would use a one proportion z-test.

i) Use a theoretical test to assess the strength of evidence that the sample data provide for Güntürkün's conjecture that kissing couples tend to lean their heads to the right more often than they would by random chance. Report the approximate p-value and summarize your conclusion about this strength of evidence.

```
pi <- 0.5
z <- (phat - pi) / sqrt(pi*(1-pi)/ 124 )
z
```

```
## [1] 3.232895
```

```
1-pnorm(abs(z))
```

```
## [1] 0.000612712
```

Should be roughly 0 - this gives very strong evidence that the proportion is not equal to 0.5

j) Now use theoretical methods to test whether the data provide evidence that the probability that a couple leans their heads to the right while kissing (π) is different from 0.60. (Note that this question changes both the null hypothesis and the alternate hypothesis) Report the standardized statistic, p-value, and comment on the strength of evidence.

```
pi <- 0.6
z <- (phat - pi) / sqrt(pi*(1-pi)/ 124 )
z
```

```
## [1] 1.02653
```

```
2*(1-pnorm(abs(z)))
```

```
## [1] 0.3046419
```

z = 1.02653
p-value = 0.3046419
This is weak evidence against the null hypothesis, we would say that 0.6 is plausible for the true proportion of couples who kiss while leaning right.

k) Using theoretical methods, calculate the 95% confidence interval. Interpret your results.

```
c(phat - qnorm(0.975) * sqrt(phat*(1-phat)/124), phat + qnorm(0.975) * sqrt(phat*(1-phat)/124))
```

```
## [1] 0.5609468 0.7293758
```

Our confidence interval is $\hat{p} \pm$ M * SE $= 0.645 \pm qnorm(1 - \frac{0.05}{2}) * \sqrt{\frac{0.645*(1-0.645)}{124}} = (0.5609, 0.7294)$
We are 95% confident that the true parameter lies between 0.5609 and 0.7294.

l) Does your confidence interval include 0.50? Does it include 0.60? Explain how your answers relate to the strength of evidence tests conducted in i and j above.

Our interval does not include 0.5 but it does include 0.6. This makes sense since, with a signifance level of 0.05, we already concluded in 9 that 0.50 is not feasible, but 0.60 is.

m) Now suppose we were to use a significance level of 0.01 instead of 0.05. How would you expect the interval of plausible values to change: wider, narrower, or no change? Explain your reasoning.

We expect it to get wider, we are including more values in our confidence interval

n) Calculate the corresponding 99% confidence interval. Did it behave as expected?

```
c(phat - qnorm(0.995) * sqrt(phat*(1-phat)/124), phat + qnorm(0.995) * sqrt(phat*(1-phat)/124))
```

```
## [1] 0.5344847 0.7558379
```

Our confidence interval is $\hat{p} \pm$ M * SE $= 0.645 \pm qnorm(1 - \frac{0.01}{2}) * \sqrt{\frac{0.645*(1-0.645)}{124}} = (0.534, 0.756)$
This is a wider interval than our 95% confidence interval above, as expected.

o) Based on your 99% confidence interval, what can be said about the p-value for testing a null hypothesis of 0.78?

As it falls outside of our 99% confidence interval, we know that the p-value must be less than 0.01 for a two-sided test and a null hypothesis of $\pi = 0.78$.

p) Can your results be generalized? Explain your reasoning.

We cannot generalize our results as the couples selected were not truly random. We do not know what parts of Germany were surveyed to know what groups were or were not included. Additionally, outgoing coupls may have a different tendency than those who do not kiss in public places.