Lab 05 Answer Sketch and Rubric

431 Staff

2020-10-28

Set Up

```
knitr::opts_chunk$set(comment=NA)

library(broom)
library(janitor)
library(knitr)
library(magrittr)
library(patchwork)
library(tidyverse)

options(dplyr.summarise.inform = FALSE)
theme_set(theme_bw())
```

Data Load

```
data_url <-
  "https://raw.githubusercontent.com/THOMASELOVE/431-2020/master/labs/lab05/data/lab05.csv"

lab05 <- read_csv(data_url)
```

Load the `Love-boost.R` script

```
Love_boost_URL <-
  "https://raw.githubusercontent.com/THOMASELOVE/431-data/master/Love-boost.R"

source(Love_boost_URL)
```

Description of the `lab05.csv` file

The available data are described below, and may be found at this link.

Variable	Description
`county_number`	Identification Code for County (C-001 through C-487)
`county_type`	NCHS county classification (6 levels listed above)
`drive_in_minutes`	Driving time (minutes) to certified OTP from county center

You will have to do at least some rearrangement work (perhaps including creating new variables, pivoting, transformation, rearrangement, etc.) in order to answer questions that relate to this study. The data set will require some management on your part. It's up to you to figure out what needs to be done. Be sure you check for missingness in the data, and note that each of the three footnotes in this Lab are worth your attention.

1 Question 1. (10 points)

> What is the outcome under study? What are the treatment/exposure groups?

1.1 Outcome (5 points)

The outcome is drive time in minutes to the nearest certified opioid treatment program (OTP) from the center of the county.

1.2 Exposure Groups (5 points)

The exposure groups being compared are

1. urban counties (those with `county_type` of small metro, medium metro, large fringe metro or large central metro)
2. rural counties (those with `county_type` of micropolitan or noncore)

2 Question 2. (10 points)

> Are the data a random sample from the population(s) of interest? Is there at least a reasonable argument for generalizing from the sample to the population(s) or is there insufficient information provided on this point? How do you know?

The five states included in this sample were Indiana, Kentucky, Ohio, Virginia, and West Virginia. Data was unavailable for two counties, so those are excluded from the `lab05` data.

If we choose as our target population of interest the US as a whole, then clearly this isn't a random sample of states from that population, nor is it particularly representative of that population. These states were selected specifically because they have the highest county rates of opioid-related overdose mortality, based on a 2018 analysis. If our target population is taken to be US counties which are at risk of high rates of opioid-related overdose mortality, then it's still clearly not a random sample, but it may be possible to describe it as fairly representative.

2.1 Grading Question 2

We're looking for a reasonable description of what the target population is (there are several possible reasonable answers) and a clear statement that this is not a *random* sample. You should also provide an argument as to why you think it is or is not representative of your target population.

3 Question 3. (10 points)

> What is the significance level (or, the confidence level) we require here? Are we doing one-tailed or two-tailed confidence interval generation? How do you know?

There are two good responses, and one other that we expect to see quite frequently that we don't like.

- We clearly require a 95% confidence level, to match the specification provided in footnote 3. That implies a 5% significance level (or $\alpha = 0.05$).
- The PI is focused on whether rural counties actually have detectably **longer** drive times than urban counties.

As a result, there is definitely a reasonable argument for completing a one-tailed or one-sided 95% confidence interval, by providing only a lower bound, and that will be our focus down below. That's certainly the most straightforward approach in this Lab.

If you want to instead build a two-tailed confidence interval (which would match with Dr. Love's general recommendations) you would want to explain why you were providing this interval to the PI. One reasonable argument (in fact, the one Dr. Love would make in this case) would be to provide a 90% CI with two tails, since that can also provide a 95% CI with one tail by just looking at the lower bound. We'll provide that interval below, as well, in our responses, and if you take this approach, we will look on that favorably, too.

We expect that some folks will instead build a two-tailed 95% confidence interval and that is hard to justify. You'll have to convince us as to why it's a good idea that you're not matching the request of the PI, and that will likely lead to you losing credit here, but not below, if your results in later questions are consistent with your Question 3 decision.

3.1 Grading Question 3

This is an essay question, where we expect you to identify the one-tailed 95% CI, or perhaps a two-tailed 90% approach that also yields the one-tailed 95% bound in order to receive full credit.

4 Question 4. (15 points)

> What do you need to do (if anything) to manage the data, create new variables, or rearrange the data for analyses? Perform these data management tasks using R code, and describe what you've done, and why you've done it.

The most important thing that you'll need to do is create a new factor variable (which I'll call `county2`) which re-categorizes each county as **rural** or **urban** on the basis of the current `county_type`. We'll do this as follows:

```
lab05 <- lab05 %>%
  mutate(county2 =
           fct_collapse(county_type,
                        rural = c("micropolitan", "noncore"),
                        urban = c("large central metro",
                                  "large fringe metro",
                                  "medium metro", "small metro")))

lab05 %>%
  tabyl(county_type, county2) %>%
  adorn_totals(where = c("row")) %>%
  adorn_title()
```

```
                    county2
         county_type   urban rural
 large central metro      25     0
  large fringe metro      42     0
        medium metro      50     0
```

```
micropolitan       0   110
     noncore       0   160
 small metro     100     0
       Total     217   270
```

From the above tabyl, we see that there are now 217 urban and 270 rural counties, as suggested in the instructions.

While footnote 2 in the instructions indicated that counties with incomplete data were removed, let's double check for good measure.

```
mosaic::favstats(drive_in_minutes ~ county2, data = lab05)
```

```
Registered S3 method overwritten by 'mosaic':
  method                           from
  fortify.SpatialPolygonsDataFrame ggplot2
```

```
  county2 min Q1 median Q3 max     mean       sd   n missing
1   urban   1 15     24 32  86 26.74654 19.13755 217       0
2   rural   1 37     44 55 100 45.82593 16.98817 270       0
```

Both `urban` and `rural` counties have missing n = 0, so we are good!

4.1 Grading Question 4

The key issue was creating the new factor variable. There were other ways to do this besides fct_collapse, and you were welcome to use any of those strategies. It would also have been OK if you created a variable called rural that was 1 for the rural counties, and 0 for the urban ones, or a variable called urban that was a 1/0 indicator in the other direction.

We didn't require you to show the verification that you now have exactly 217 and 270 counties, but you really should have. We also didn't require you to check that the data had no missing values.

5 Question 5. (10 points)

> Were the data collected using matched / paired samples or independent samples? How do you know?

The data are gathered using independent samples. There is no pairing or matching of urban to rural counties, and one easy way to describe that is that we do not have equal sample sizes in the two exposure groups, so we cannot possibly have paired data.

5.1 Grading Question 5

You're basically either correct or incorrect here. If you chose paired samples, that was definitely a mistake, if for no other reason than Dr. Love told you in class that these were independent samples.

6 Question 6. (15 points)

> Answer either part 1 or 2 of this question, whichever is appropriate in light of your response to Question 4.

1. If we have paired samples, what does the distribution of sample paired differences tell us about which inferential procedure to use? Display an appropriate visualization that motivates your conclusions, and then describe those conclusions in English.

We do not have paired samples, so we will skip this.

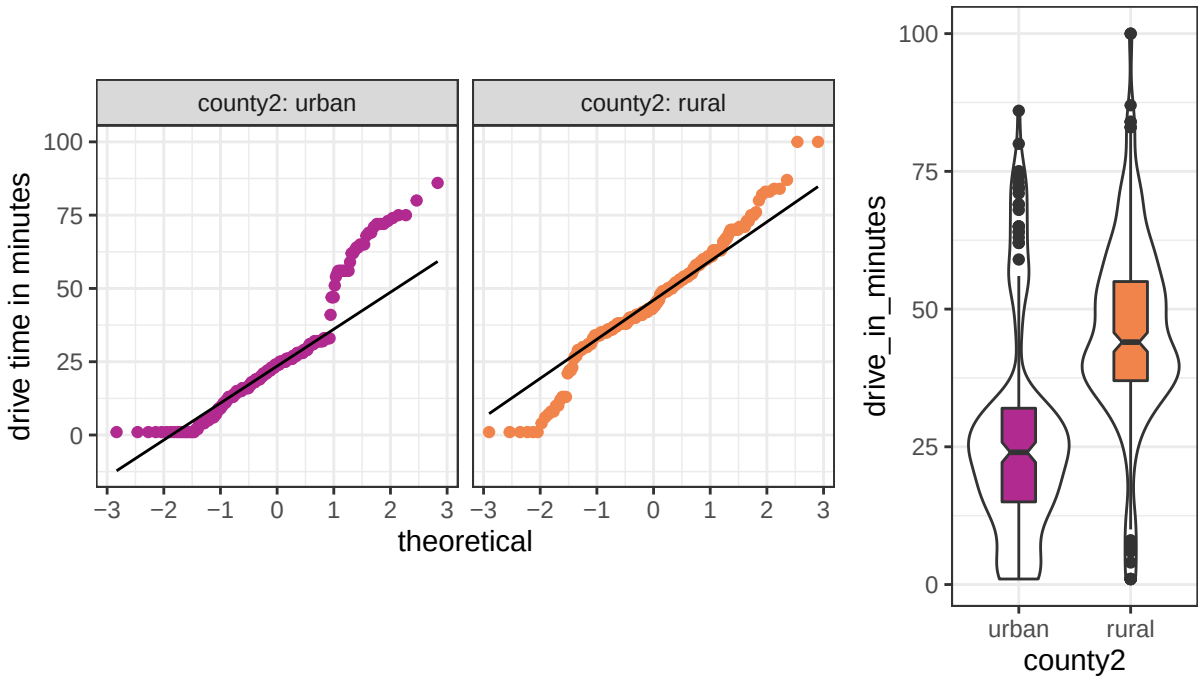
2. If we have independent samples, what does the distribution of each individual sample tell us about which inferential procedure to use? Display an appropriate visualization that motivates your conclusions, and then describe those conclusions in English.

Here's a visualization using Normal Q-Q plots, and boxplots and violin plots.

```
p1 <-
  ggplot(lab05, aes(sample = drive_in_minutes)) +
  geom_qq(aes(col = county2)) + geom_qq_line(col = "black") +
  facet_wrap(~ county2, labeller = "label_both") +
  scale_color_viridis_d(option = "C", begin = 0.4, end = 0.7) +
  theme(aspect.ratio = 1) +
  labs(y = "drive time in minutes") +
  guides(col = FALSE)

p2 <-
  ggplot(lab05, aes(x = county2, y = drive_in_minutes)) +
  geom_violin() +
  geom_boxplot(aes(fill = county2), notch = TRUE, width = 0.3) +
  scale_fill_viridis_d(option = "C", begin = 0.4, end = 0.7) +
  guides(fill = FALSE)

p1 + p2 + plot_layout(widths = c(3,1)) +
  plot_annotation(title = "Comparing Drive Time to Nearest Opioid Treatment Program",
                  subtitle = "Rural and Urban Counties in 5 States (IN, KY, OH, VA, WV)")
```



I don't think you can make a good case for treating these as well described by a Normal model.

The rural data might be symmetric, but appears to be heavy-tailed. The urban data show a floor effect near 0m and a very heavy right tail, suggesting some problems with assuming symmetry, let alone a Normal distribution.

- A bootstrap is likely to be a stronger choice for comparing means here than one based on a t-distribution.
- A rank-based approach is problematic because it's hard to assume symmetry.

6.1 Grading Question 6

- Building an appropriate, well-labeled plot (which might just be the Normal Q-Q plots, or might just be a boxplot or pair of histograms, although the combination is best) is worth 8 points.
- The conclusions about what the data show and what approach to use are worth the remaining 7 points.

7 Question 7. (15 points)

> Produce an appropriate confidence interval for a relevant population *mean* that addresses the key question from the study, following the requirements of the principal investigator. Be sure to show and describe the R code that led to your selected confidence interval, and describe how your responses to prior Questions led you to select this approach.

A one-tailed 95% confidence interval using the bootstrap may be found by calculating the 90% two-sided confidence interval, and using its upper bound. Using `4312020` as my seed, I get:

```
set.seed(4312020)

lab05 %$%
  bootdif(drive_in_minutes, county2, conf.level = 0.90)
```

```
Mean Difference            0.05            0.95
       19.07938        16.35268        21.73906
```

Note that we usually use `message = FALSE` the first time we call `bootdif` to avoid an unhelpful message.

- So the 95% one-tailed confidence interval here would have lower bound 16.35 minutes.
- Alternatively, the 90% two-tailed CI is (16.35, 21.74) minutes.

7.1 Grading Question 7

- It was important to get the units (minutes) right in presenting your interval here.
- It was important to fit a confidence interval that matches your responses to questions 3 and 6, assuming you didn't call these paired samples.
- It is critical in Question 8 to recognize the direction here (whether this difference is rural - urban or the reverse) but you didn't need to discuss that in your Question 7 response explicitly.

7.1.1 What if we used a t-based procedure instead?

I'll just emphasize that I don't think this is particularly justifiable, since the data in neither sample is well-approximated by a Normal model.

```
model2 <- lm(drive_in_minutes ~ county2, data = lab05)

tidy(model2, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low)
```

```
# A tibble: 2 x 3
  term         estimate conf.low
  <chr>           <dbl>    <dbl>
1 (Intercept)      26.7     24.7
2 county2rural     19.1     16.4
```

The one-sided 95% confidence interval based on the t distribution uses the same lower bound as the two-sided 90% confidence interval, and that bound is 16.38 minutes.

To fit this directly, we could use the pooled t test:

```
lab05 %$% t.test(drive_in_minutes ~ county2,
                 conf.level = 0.95,
                 var.equal = TRUE,
                 alt = "less")
```

```

    Two Sample t-test

data:  drive_in_minutes by county2
t = -11.641, df = 485, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -16.37834
sample estimates:
```

```
mean in group urban mean in group rural
           26.74654            45.82593
```

Note that the sample sizes (217 vs. 270) and sample variances (19.14^2 = 366.34, and 16.99^2 = 288.66 are somewhat different from each other.

```
mosaic::favstats(drive_in_minutes ~ county2, data = lab05)
```

```
  county2 min Q1 median Q3 max     mean       sd   n missing
1   urban   1 15     24 32  86 26.74654 19.13755 217       0
2   rural   1 37     44 55 100 45.82593 16.98817 270       0
```

So we might instead prefer to use the Welch t method to obtain our confidence interval:

```
lab05 %$% t.test(drive_in_minutes ~ county2,
                 conf.level = 0.95,
                 alt = "less")
```

```
    Welch Two Sample t-test

data:  drive_in_minutes by county2
t = -11.491, df = 435.85, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -16.34259
sample estimates:
mean in group urban mean in group rural
           26.74654            45.82593
```

8 Question 8. (15 points)

> Interpret your confidence interval from Question 7 in the context of the request by the project's principal investigator using complete English sentences.

Lets interpret our Bootstrap CI calculated using the `bootdif` function.

The one-sided 95% confidence interval would have a lower bound of 16.35 minutes for the population mean driving time difference comparing rural to urban counties. Specifically we are estimating $\mu_{Rural} - \mu_{Urban}$ here, and we know this from the listed Mean Difference (which is 45.83 - 26.75 and not the reverse.)

We are 95% confident that this method of creating confidence intervals will produce a result containing an accurate lower bound for the true population mean rural - urban difference in mean drive times to OTP clinics.

8.1 Grading Question 8

- It was important to indicate the direction of the difference. (Rural - Urban instead of the reverse, if you used our approach.)
- It was important to match your response to questions 3 and 6 in your interval, and to match your statement about that interval back to the original description of the data.
- It was important that your statement about what a confidence interval means be sufficiently correct. For example, something like "we are 95% sure that the true mean is in the interval" isn't correct.

Miscellaneous Issues

Submission Problems

- Failure to submit a functioning R Markdown file and HTML result derived from that R Markdown file by the deadline specified in the course calendar will cost 10 points.
- Failure to submit a functioning R Markdown file and HTML result derived from that R Markdown file by noon on the day after the deadline will cost 25 points.
- Work submitted more than one week after the deadline will not be accepted.

Original Footnotes

1. Joudrey PJ, Edelman EJ, Wang EA. Drive Times to Opioid Treatment Programs in Urban and Rural Counties in 5 US States. *JAMA*. 2019;322(13):1310-1312. https://doi.org/10.1001/jama.2019.12562
2. The five states were Indiana, Kentucky, Ohio, Virginia, and West Virginia. Data was unavailable for two counties, so those are excluded from the `lab05` data set.
3. The Joudrey et al. paper presented a mean drive time across all counties of 37.3 minutes, with a 95% confidence interval of (35.5, 39.1) minutes. You should be able to verify that the simulated data in `lab05` matches those results. The simulation mirrors this particular result, and it also mirrors the means by classification shown in the Table accompanying the original letter, but it does not mirror any other elements of that data set.