Lab Rejections

Abhi Thanvi, Paul Holaway

July 13th, 2022

Contents

Lab Rejections

Welcome

Just like learning a new spoken language, you will not learn the language without practice. Labs are an important part of this course. Collaboration on labs is **extremely encouraged**. If you find yourself stuck for more than a few minutes, ask a neighbor or course staff for help. When you are giving help to your neighbor, explain the **idea and approach** to the problem without sharing the answer itself so they can figure it out on their own. This will be better for them and for you. For them because it will stick more and they will have a better understanding of the concept. For you because if you can explain it to other students, that means you understand it better too.

The Idea of this Lab

After learning the topic of Confidence Interval, we are now moving on a topic of Hypothesis Testing. This topic is the second major half of Statistical Inference. Hypothesis Testing allows statisticians to understand if the statistic they get from their sample is significant or not. The significance depends on p-values. This topic could be pretty helpful for your project, if you do decide to go along this path for your project. I truly think this lab is going to be really helpful for you guys, and let's just dive in!

"Usually you want to be right, but in hypothesis testing, you want to be wrong (or rejecting stuff) because it means the p-value is significant" - Helpful Abhi

Problem 1: Stalker Alert

Question 1: Joe Goldberg wants to stalk people's text messages for obvious reasons. He wants to avoid stalking group of people that send only few text messages, according him, it's not worth taking the risk with them. Joe has hired us as Data Scientist to investigate if STAT 107 class will be a good group to stalk on. I told Joe 30 texts is the average we should look at. Joe claims that the average number of texts sent by each person for a "good" group for stalking would be more than 30. Perform the hypothesis test for Joe to see if STAT107 would be good group to be stalked by Joe. Use 0.05 as the significance level. Pretend that the `hello` data set is the group Joe wants to stalk.

$$H_0 : \mu = 30$$

$$H_a : \mu > 30$$

Question: Do we use Z or t for the test statistic here?

Answer: We can use either as we have a large sample size so $Z \approx t$.

```
#Setup
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
hello <- read.csv("~/Desktop/DPISu22/Data Sets/hello.csv", stringsAsFactors=TRUE)
#TS Calculation
xbar = mean(hello$Texts)
n = length(hello$Texts)
sigma = sd(hello$Texts)*sqrt((n-1)/n)
Zts = (xbar - 30)/(sigma / sqrt(n))
Zts
```

```
## [1] -7.720009
```

```r
#p-value Calculation
pnorm(Zts,lower.tail = FALSE)
```

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

Question: Do we reject or fail to reject H_0?

Answer: We would fail to reject H_0 because the p-value > 0.05.

Question 2: Turns out Joe Goldberg was Abhi. He just wanted to test how good the data scientist he hired were. Sneaky Sneaky. Anyways, we still get paid so let's just do what he says...again (insert eye roll). He mentioned he wanted us to see if the average iPhone sales was equal to 25, that's what we think is correct. Abhi doesn't know the units, but he has a hunch on the number 25. He thinks that the iPhone sales should be more than 25 since iPhone is super popular. Can we test this? Of course! Perform the hypothesis test for Abhi to see if the iPhone sales more than 25. Use 0.05 as the significance level. Pretend that the `apple_sales` data set is the sample of apple_sales reports.

$$H_0 : \mu = 25$$

$$H_a : \mu > 25$$

Question: Do we use Z or t for the test statistic here?

Answer: We can use t as we have a sample with unknown population standard deviation.

```r
#Setup
library(tidyverse)
apple_sales <- read.csv("~/Desktop/DPISu22/Data Sets/apple_sales.csv",
                        stringsAsFactors=TRUE)

#DO NOT DELETE THIS
filtered_sales <- apple_sales %>% drop_na(iPhone) %>% drop_na(iPad) %>% drop_na(iPod) %>% drop_na(Mac)

#TS Calculation
xbar = mean(filtered_sales$iPhone)
n = length(filtered_sales$iPhone)

#Needed for tts
s = sd(filtered_sales$iPhone)
df = n - 1

#tts
tts = (xbar - 25)/(s / sqrt(n))
tts
```

```
## [1] 1.725753
```

```
#p-values
pt(tts,df,lower.tail = FALSE)
```

```
## [1] 0.05126024
```

Question: Do we reject or fail to reject H_0?

Answer: We would fail to reject H_0 because the p-value > 0.05.

Problem 2: Am I Wrong?

Question 1: I am trying to do a z-test with a sample data set that has 20 rows. We do not know the population standard deviation of the population. Do you think z-test is possible?

Answer: (Student Response Here) Nope, we have a small sample with unknown standard deviation so using t would be better.

Question 2: I am trying to do a t-test with a data set that has 5000 rows. The population variance is 25. Do you think we should be doing a z or t-test?

Answer: (Student Response Here) You can use Z-interval as it is most likely going to be the most accurate.

Question 3: Wow wrong two times, but somehow I did one successful hypothesis test. I got a p-value is 0.00231. Could I be rejecting the null hypothesis, explain?

Answer: (Student Response Here) Yes you can reject the null hypothesis since the p-value is less than 0.05 (common yet acceptable significance level)

Question 4: The true mean of the population is 5.0 and our alternative hypothesis states the mean is less than 3. Our p-value was some small number. We then decide to make our alternative hypothesis to state that the mean is less than 4.5. Will our new p-value be larger or smaller? Feel free to play around with numbers or Google. Just make sure to explain your answers.

Answer: (Student Response Here) The p-value would be larger because we are getting closer to the true mean which means our test statistic would be smaller. In other words, we wouldn't be rejecting the null hypothesis, because our null hypothesis is close to the the true mean which means we wouldn't be rejecting the null hypothesis. So that's two ways of thinking about it.

Problem 3: Searching for answers

Question 1: What is the relation between confidence intervals and p-value? Discuss with your group and/or search for Google if necessary. This is a skill that would be helpful for your project and future data science career.

DO NOT COPY ANSWERS FROM GOOGLE!!!!

Answer: (Student Response Here) *Answers will vary and will be reviewed by instructors.* The confidence interval is the "acceptable range of values" and p-value tells us the likelihood of having as extreme values as you got. The width of the confidence interval and the size of the p value are related, the narrower the interval, the smaller the p value. There are more relations that could be possible, thus the instructors will review them as appropriate.

Project Questions

Feel free to work on your project if there is any time left after the labs. Paul and I are here to answer any questions during the second half of the lab times to answer mainly project related questions, but general questions are more than welcome too. Feel free to discuss among your group about any project ideas or help each other out. Remember collaboration is promoted, plagiarism is not! :)

Submission

Once you have finished your lab...

1. Go to the top left and click `File` and `Save`.
2. Click on the `Knit` button to convert this file to a PDF.
3. Submit **BOTH** the `.Rmd` file and `.pdf` file to Blackboard by 11:59 PM tonight.