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 use Z here, because our population is the STAT107 class, and we have data for every person in this class.

```
#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()
```

```
hello <- read.csv("~/Desktop/DPI 2022/hello.csv", stringsAsFactors=TRUE)
#TS Calculation
xbar = mean(hello$Texts)
n = nrow(hello)
sigma = sd(hello$Texts)*sqrt((n-1)/n)
Zts = (xbar - 30)/(sigma / sqrt(n))
Zts
```

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

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

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

Question: Do we reject or fail to reject H_0?

Answer: We fail to reject H_0, because our p value is 1, which is bigger than our significance level ($\alpha = 0.05$). This means that we cannot claim that the mean number of texts is *not* 30 or less - in other words, it could be (and probably is) less than or equal to 30.

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 use t for the test statistic, because we're pretending that this data set is a sample of apple_sales reports (meaning we don't know our population standard deviation) and our filtered data set is tiny (it has a size of 18, which is less than 30).

```
#Setup
apple_sales <- read.csv("~/Desktop/DPI 2022/apple_sales.csv", stringsAsFactors=TRUE)

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

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

#Needed for tts
s = sd(filtered_sales$iPhone)
```

```r
df = n - 1

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

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

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

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

Question: Do we reject or fail to reject H_0?

Answer: We just barely fail to reject H_0 here, because our p value greater than 0.05. This means that we fail to reject H_0 by a little bit, so with a 95% confidence level, we cannot reject the possibility of average iPhone sales being less than or equal to 25.

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: I don't think the Z test is possible here, since the data set is too small (less than 30) and the population standard deviation is unknown. In fact, because of the small sample size, the use of a Z test would not account for the discrepancies caused by a small sample size (however, because of the augmented nature of the t distribution, a t test would account for these discrepancies and would therefore be a better test here).

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: If you know the population variance (σ^2), then you know the population standard deviation (σ), since it's just the square root - in this case, since $\sigma^2 = 25$, $\sigma = 5$. Since you KNOW the population standard deviation now (you know that it's 5 now!), you should resort immediately to a Z test.

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: At a significance level of $\alpha = 0.05$, you can definitely reject your null hypothesis here, because $0.00231 < 0.05$.

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: Our new p value will be even larger. This is because by bringing the alternative hypothesis closer to the mean, you're bringing your p-value closer to 0.5, because for alternative hypotheses close to the actual mean, it should be pretty much equally likely for something in the population to be greater or less it. Since you started with a small number, bringing it closer to 0.5 would increase its value, and therefore, p will be larger.

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: If a confidence interval with confidence level $(1 - \alpha)$ includes the null value (μ_0), then the p value must be greater than (or equal to) the significance level (α), because if μ_0 is within the confidence interval, that means that it must be pretty accurate, and therefore shouldn't be rejected, thus making its p value greater than the significance level (α). For example, if your μ_0 value is within a 95% confidence interval, then its p value must be greater than or equal to 0.05.

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.