Lecture 13

Paul Holaway, Abhi Thanvi

July 13th, 2022

Lecture 12 Review

Before we move on to the new material, we will do a quick review of Lecture 12 content. Last time we learned how to make confidence intervals for proportions (p) and means (μ). We will begin today doing a quick review of that.

Example 1; CI Review

Suppose you are taking a sample of people where you ask two things…

1. Do they have siblings? (0 = No, 1 = Yes)
2. How many siblings do they have?

The data is below.

```
#DO NOT DELETE THIS CODE!!
set.seed(143572)
sibling = sample(0:1, size = 15, replace = TRUE)
number = rep(0,15)
for(i in 1:15){
  if(sibling[i] != 0){
    number[i] = sample(1:5, size = 1, replace = TRUE, prob = c(0.55,0.25,0.10,0.08,0.02))
  } else {
    number[i] = 0
  }
}
data = data.frame(sibling,number)
data
```

```
##    sibling number
## 1        1      4
## 2        1      1
## 3        1      2
## 4        0      0
## 5        1      2
## 6        1      1
## 7        0      0
## 8        1      1
## 9        1      1
## 10       1      2
## 11       1      1
## 12       0      0
## 13       0      0
## 14       0      0
## 15       1      2
```

Create two 95% confidence intervals. One for the proportion of people with siblings, another for the average number of siblings they have. Look back at Lecture 12 notes if you do not remember the formulas.

```
#Proportion CI
n = length(data$sibling)
phat = sum(data$sibling)/n
c = (1-0.95)/2
Z = qnorm(c,lower.tail = FALSE)
phat + c(-1,1)*Z*sqrt(phat*(1-phat)/n)
```

```
## [1] 0.4281074 0.9052259
```

```
#Mean CI
xbar = mean(data$number)
s = sd(data$number)
df = n - 1
t = qt(c,df,lower.tail = FALSE)
xbar + c(-1,1)*t*s/sqrt(n)
```

```
## [1] 0.5100728 1.7565939
```

Okay, now we can move onto the next portion of lecture content.

Hypothesis Testing

Today we are going to cover the other most important topics in basic statistics, hypothesis testing. You have probably heard the word **hypothesis** used in many different contexts. In statistics, a **hypothesis** is a declaration/claim, in the form of a mathematical statement, about the value of a specific (or several) population parameter(s). There are four parts of a hypothesis test…

1. **Null Hypothesis** (H_0): This is the claim (about a population parameter) assumed to be true, what is believed to be true, or the hypothesis to be tested. You can think of this as the status quo. This may seem strange, but we usually try to *reject* the null hypothesis. This is because it is actually easier in statistics to disprove a hypothesis than to prove it is true.
2. **Alternative Hypothesis:** (H_a): This statement identifies other possible values of the population parameter, or simply a possibility not included in H_0. H_a indicates the possible values of the parameter if H_0 is false. Experiments are often designed to determine whether there is evidence in favor of H_a. The alternative hypothesis represents change in the status quo.
3. **Test Statistic** (TS): This statistic is a rule, related to H_0, involving the information in a sample. The *value* of the TS will be used to determine which hypothesis is more likely to be true, H_0 or H_a.
4. **p-value**: This will be a number that tells you the probability of rejecting H_0, but in reality, it was correct all along. The larger the p-value, the less likely it is you should reject H_0. We will set a maximum value permitting the p-value to be for us to reject H_0. This is known as the significance level (α).

Now we can do multiple types of hypothesis testing, however, we are limited on time. Therefore, we will just be focusing on single sample hypothesis testing for population means. For this process, let's walk through each of the four steps before doing an example.

Null Hypothesis

The null hypothesis is again, what we assume to be the status quo. Another way you can think about it as what you believe to be true. The null hypothesis is always written like this… $\mu = \mu_0$, where μ_0 is the mean value that you believe to be true.

Alternative Hypothesis

The alternative hypothesis is again, the other possible values of the parameter if H_0 is false. It is what you are testing against H_0. The alternative hypothesis can be written in three different ways…

1. $\mu < \mu_0$
2. $\mu > \mu_0$
3. $\mu \neq \mu_0$

Which one you use depends on what you are trying to investigate. Thankfully, which one you use where is intuitive. If you are trying to prove that the parameter is on average smaller than what you think, you use #1. If you are trying to prove that the parameter is on average larger than what you think, you use #2. If you are trying to prove that the parameter is on average not what you think, you use #3.

Test Statistic

Think of the test statistic as a tool to help you do your hypothesis test. There are two different formulas for the test statistic just like how there are two different formulas for the confidence intervals.

$$Z_{TS} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$t_{TS} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- $\bar{x}$ is the sample mean
- μ_0 is the value you are testing in your hypothesis
- n is the population/sample size
- σ is the population standard deviation
- s is the sample standard deviation

You are probably wondering when you use which formula. Thankfully, it is just like when making a confidence interval.

	Known Pop. SD	Unknown Pop. SD
Large Sample	Z	Z or t
Small Sample	Z	t

p-value

The p-value can be calculated in multiple different ways, but we will use `RStudio` because it is the easiest. How you calculate the p-value depends on your *alternative hypothesis* and whether you are using Z or t.

1. $\mu < \mu_0$

- Z: `pnorm(Zts,lower.tail = TRUE)`
- t: `pt(tts,df,lower.tail = TRUE)`

2. $\mu > \mu_0$

- Z: `pnorm(Zts,lower.tail = FALSE)`
- t: `pt(tts,df,lower.tail = FALSE)`

3. $\mu \neq \mu_0$

- Z: `2*pnorm(abs(Zts),lower.tail = FALSE)`
- t: `2*pt(abs(tts),df,lower.tail = FALSE)`

Remember, the output from this code is the probability that we reject H_0, but in reality it is correct. If the probability outputted by `RStudio` is less than the significance level, then we would reject H_0. Otherwise, we would fail to reject H_0.

- $p - value \leq \alpha \implies$ Reject H_0
- $p - value > \alpha \implies$ Fail to Reject H_0

Okay, now that we have done a massive info dump, let's go through some examples to finish out lecture.

Example 2; Settling An Arguement

Alex and Bob are baseball fans who love to go to go watch a local college summer league team because it is much cheaper than hyper-inflated MLB prices. Bob claims that the average number of hits each player on the team has so far is about 25 hits per player. Alex disagrees, he does not think that the average number of hits a player has on the team so far is about 25. Perform the hypothesis test for them to see who is correct. Use 0.05 as the significance level. Pretend that the `baseball` data set is the team's stats so far.

$$H_0 : \mu = 25$$

$$H_a : \mu \neq 25$$

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

Answer: We would use Z as we have the population of all players on the team.

```
#Setup
baseball <- read.csv("~/Classes/DPISu22/Data Sets/baseball.csv", stringsAsFactors=TRUE)
#TS Calculation
xbar = mean(baseball$H)
n = length(baseball$H)
sigma = sd(baseball$H)*sqrt((n-1)/n)
Zts = (xbar - 25)/(sigma / sqrt(n))
Zts
```

```
## [1] -3.822099
```

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

```
## [1] 0.0001323208
```

Question: Do we reject or fail to reject H_0?

Answer: We would reject H_0 because the p-value < 0.05.

Example 3; Train Delays

NJ Transit has recently received multiple complaints from riders that trains are showing up more than 5 minutes late. (They have no patience.) NJ Transit would like to know if this is really the case because they pride themselves on punctuality. They took a sample of their trains and figured out how many minutes late the train was. Perform the hypothesis test for them to see if trains are indeed running more than 5 minutes late. Use 0.05 as the significance level. Pretend that the `AMTK_NJT Performance 5-20` data set is the sample of trains taken. Recall that there are `NA`s in this data, so we will have to remove them.

$$H_0 : \mu = 5$$

$$H_a : \mu > 5$$

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

Answer: We can use either as we have a large population with an unknown population standard deviation.

```
#Setup
AMTK_NJT <- read.csv("~/Classes/DPISu22/Data Sets/AMTK_NJT Performance 5-20.csv",
                     stringsAsFactors=TRUE)
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()
```

```
NJT = AMTK_NJT %>% filter(type == "NJ Transit") %>% drop_na(delay_minutes)
#Needed for both
xbar = mean(NJT$delay_minutes)
n = length(NJT$delay_minutes)
#Needed for Zts
sigma = sd(NJT$delay_minutes)*sqrt((n-1)/n)
#Needed for tts
s = sd(NJT$delay_minutes)
df = n - 1
#Zts
Zts = (xbar - 5)/(sigma / sqrt(n))
Zts
```

```
## [1] -49.49241
```

```
#tts
tts = (xbar - 5)/(s / sqrt(n))
tts
```

```
## [1] -49.49213
```

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

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

```
pt(tts,df,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. Much large actually, the p-value is 1.

Example 4; F

The University of Urbana-Champaign is worried that their professors are not failing enough students (More Fs = More Time In College = More Money For UIUC). They figure that there should be on average 1 F per class. Perform the hypothesis test for them to see if there is less than 1 F per class. Use 0.05 as the significance level. Use the gpa data set and use Spring 2020 courses for the sample.

```
gpa <- read.csv("~/Classes/DPISu22/Data Sets/gpa.csv", stringsAsFactors=TRUE)
#DO NOT DELETE THIS!! This renames all columns for convenience.
gpa = gpa %>% rename("A+" = "A.") %>% rename("A-" = "A..1") %>% rename("B+" = "B.")
gpa = gpa %>% rename("B-" = "B..1") %>% rename("C+" = "C.") %>% rename("C-" = "C..1")
gpa = gpa %>% rename("D+" = "D.") %>% rename("D-" = "D..1") %>% rename("F" = `F`)
```

$$H_0 : \mu = 1$$
$$H_a : \mu < 1$$

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

Answer: We can use either as we have a large population with an unknown population standard deviation.

```
#Sample
gpa_S20 = gpa %>% filter(YearTerm == "2020-sp")
#Needed for both
xbar = mean(gpa_S20$F)
n = length(gpa_S20$F)
#Needed for Zts
sigma = sd(gpa_S20$F)*sqrt((n-1)/n)
#Needed for tts
s = sd(gpa_S20$F)
df = n - 1
#Zts
Zts = (xbar - 1)/(sigma / sqrt(n))
Zts
```

```
## [1] -0.329337
```

```
#tts
tts = (xbar - 1)/(s / sqrt(n))
tts
```

```
## [1] -0.3292585
```

```
#p-values
pnorm(Zts,lower.tail = TRUE)
```

```
## [1] 0.3709505
```

```
pt(tts,df,lower.tail = TRUE)
```

```
## [1] 0.3709966
```

Question: Do we reject or fail to reject H_0?

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

Okay, that is all we will cover for hypothesis testing. There are a wide variety of them out there, this is just a snowflake on the tip of the iceberg. However, even what you have learned today can still be useful in your project. If you are interested in learning more about hypothesis testing, please ask your instructors. Next time we will be looking into creating simple linear models in `RStudio`. It will also be our last lecture and lab, so hang in there, we are almost at the end.

End of Lecture 13 Notes