Lecture 12

Paul Holaway, Abhi Thanvi

July 12th, 2022

Lecture 11 Review

Before we move on to the new material, we will do a quick review of Lecture 11 content. Last time we learned about the Central Limit Theorem, random variables, and how to calculate their expected values and standard deviations. We info-dumped a few formulas on you, then taught you how to calculate the values in `RStudio` if the variable is Normal. Let's quickly review them before moving onto today's topic.

Example 1; Quick Review

Alex owns a race horse that is really good. It always finishes in the top 5. However, he wants to know where his horse will most likely finish (for various reasons). What is the expected place his horse will finish and the standard deviation of the finish.

Finish	Probability
1	0.15
2	0.25
3	0.35
4	0.20
5	0.05

$$\mathbb{E}(Finish) = 1 * 0.15 + 2 * 0.25 + 3 * 0.35 + 4 * 0.20 + 5 * 0.05 = 2.75 \approx 3rd$$
$$SD(Finish) = \sqrt{\mathbb{E}(Finish^2) - (\mathbb{E}(Finish)^2)}$$
$$\mathbb{E}(Finish^2) = 1^2 * 0.15 + 2^2 * 0.25 + 3^2 * 0.35 + 4^2 * 0.20 + 5^2 * 0.05 = 8.75$$
$$SD(Finish) = \sqrt{8.75 - (2.75)^2} \approx 1.0897$$

As for when you have a continuous random variable that is normal, remember that you can just use `mean(Data$Variable)` and `sd(Data$Variable)` to get the information you want. It is pretty simple to type into `RStudio`, so I will hold off on doing an example here. You will use both later today during our lecture time. Okay, now we can move onto the next portion of lecture content.

Confidence Intervals

Today we are going to cover one of the most useful tools in basic statistics, the confidence interval. A **Confidence Interval** is an interval of values constructed so that, with a specified degree of confidence, the value of the population parameter lies within it. Be careful, this does **NOT** say that our confidence interval contains the population parameter. It says that there is a certain percentage chance that the interval contains the population parameter. Remember that random sampling exists and what it can do to an estimate (#Lecture4). What the population parameter is depends on the distribution we are using.

- **Normal**
  - μ (Population Mean)
  - σ/σ^2 (Population Standard Deviation/Variance)
- **Binomial**
  - p (Population Probability)
  - n (Population Size)

We will only focus on μ for Normal and p for Binomial since those are the most commonly studied for those two distributions. We will first start with the confidence interval for p as it is much easier than the Normal confidence interval for μ.

Confidence Interval for a Population Percentage/Proportion

The confidence interval for a population percentage/proportion is ...

$$(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$$

- $\hat{p}$ is the estimated percentage/proportion
- n is the sample size; This must be a large number (Usually $n \geq 30$ works)
- $Z_{\frac{\alpha}{2}}$ is the quantile cutoff

You can probably already calculate the first two values, but what about the quantile cutoff? The manual calculation is quite complex, so we will use `RStudio` to do it for us. We will go over the different functions you can use in the examples. If this seems a bit overwhelming at first, that is fine. We will go through this during an example.

Example 2; Games Played

In NCAAF, most teams play only 12 games a season. However, some will play in their conference championship, bowl games, of the College Football Playoffs. However, just how many are there that do? Let's use the `cfb21` data set as the sample. Create a 90% and 95% confidence interval for the proportion of teams that play more than 12 games.

```
cfb21 <- read.csv("~/Classes/DPISu22/Data Sets/cfb21.csv", stringsAsFactors=TRUE)
```

Okay, now first we need to calculate the estimated proportion of teams that played more than 12 games. $\hat{p} = \frac{x}{n}$ where x is the number of teams that played more than 12 games. Let's calculate this in `RStudio`.

```
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()
```

```
cfb21_12 = cfb21 %>% filter(Games > 12)
n = length(cfb21$Games)
phat = length(cfb21_12$Games)/n
phat
```

```
## [1] 0.5769231
```

Okay, we now have our $\hat{p}$. The value we get makes sense due to the large number of bowl games these days. Now we can easily calculate n, but what about $Z_{\frac{\alpha}{2}}$? That can be calculated for the proportion/percentage interval using `qnorm(c,lower.tail = FALSE)`. Okay, but what is `c`? That value depends on the size of the confidence interval. It can be calculated using the following formula, $c = \frac{(1-a)}{2}$, where a is the confidence interval size. Let's calculate the two different c values.

```
c1 = (1-0.9)/2
c2 = (1-0.95)/2
c1
```

```
## [1] 0.05
```

```
c2
```

```
## [1] 0.025
```

Now let's find the $Z_{\frac{\alpha}{2}}$ values.

```
Z1 = qnorm(c1,lower.tail = FALSE)
Z2 = qnorm(c2,lower.tail = FALSE)
Z1
```

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

```
Z2
```

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

Okay, now let's make the two confidence intervals.

```
phat + c(-1,1)*Z1*sqrt(phat*(1-phat)/n)
```

```
## [1] 0.5056502 0.6481959
```

```
phat + c(-1,1)*Z2*sqrt(phat*(1-phat)/n)
```

```
## [1] 0.4919962 0.6618499
```

Note that `+ c(-1,1)` is just a fancy way for me to code in the interval without having to type the same expression twice (one with `-`, one with `+`) (#EffecientCode). Great, now we have the two different intervals. Notice how the 95% one is larger. This is because we want a higher certainty that the true population proportion of teams that played more than 12 games is captured in the interval.

Confidence Interval for a Population Mean (with Normal Distribution)

Okay, now we are onto the Normal case. In the normal distribution, it is a bit more complicated than the proportion.

In the Normal case, there are two different formulas for the confidence interval.

$$(\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$$

$$(\bar{x} - t_{\frac{\alpha}{2},df}\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},df}\frac{s}{\sqrt{n}})$$

- $\bar{x}$ is the sample mean
- n is the population/sample size
- σ is the population standard deviation
- s is the sample standard deviation
- $Z_{\frac{\alpha}{2}}$ or $t_{\frac{\alpha}{2},df}$ are the quantile cutoffs

Each one of these formulas is used in different circumstances. Notice how the only differences are population versus sample standard deviation and the quantile cutoff. To determine which formula to use, we need to look at two things.

1. Do we know the population standard deviation?
2. How large is our population/sample?

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

There are a few things to note here about coding into `RStudio`.

1. You can calculate $\bar{x}$ by using the `mean()` function like before.
2. If you are calculating the population standard deviation, you will have to do `sd(data$variable)*sqrt((n-1)/n)`. The `sd()` function calculates the sample standard deviation and the `*sqrt((n-1)/n)` is a correction. If you wish to know the math behind it, you can look it up or ask one of your instructors.
3. $Z_{\frac{\alpha}{2}}$ is still calculated using the same `qnorm(c,lower.tail = FALSE)` as before.
4. $t_{\frac{\alpha}{2},df}$ is calculated using `qt(c,df,lower.tail = FALSE)`

df stands for "degrees of freedom", and $df = n - 1$. You do not have to worry what df is, just know how to calculate it. Okay, let's do some examples to make things easier to comprehend.

Example 3; Hits

You are looking to create a 95% and 99% confidence interval for the average number of hits for a player on a baseball team. The data for this is in the `baseball` data set. Create the two confidence intervals.

1. Do we know the population standard deviation? Yes, we have the data for all players on the team.
2. How large is our population? Irrelevant as we always use Z cutoffs when the population standard deviation is known.

```
baseball <- read.csv("~/Classes/DPISu22/Data Sets/baseball.csv", stringsAsFactors=TRUE)
```

```
xbar = mean(baseball$H)
n = length(baseball$H)
sigma = sd(baseball$H)*sqrt((n-1)/n)
#Don't forget, we need the correction factor since we have the population SD.
c1 = (1-0.95)/2
c2 = (1-0.99)/2
Z1 = qnorm(c1,lower.tail = FALSE)
Z2 = qnorm(c2,lower.tail = FALSE)
#Confidence Intervals
xbar + c(-1,1)*Z1*sigma/sqrt(n)
```

```
## [1] 10.33750 20.27789
```

```
xbar + c(-1,1)*Z2*sigma/sqrt(n)
```

```
## [1]  8.775749 21.839635
```

Example 4; Z vs. t

Okay, now let's go back to the `cfb21` data set. Suppose you want to create a 95% confidence interval for the average number of touchdowns scored by the offence of NCAAF teams.

1. Do we know the population standard deviation? No, we only have 130 teams (specifically NCAAF FBS). This is not the entire population of NCAAF teams.
2. How large is our population? There are 130 teams in the sample, which is decently large.

This means we use Z or t. What's the difference though? Let's find out.

```
#Needed for both
xbar = mean(cfb21$Off.TDs)
n = length(cfb21$Off.TDs)
c = (1-0.95)/2
#Needed for Z CI
sigma = sd(cfb21$Off.TDs)*sqrt((n-1)/n) #Pop. SD
Z = qnorm(c,lower.tail = FALSE)
#Needed for t CI
s = sd(cfb21$Off.TDs) #Sample SD
df = n - 1
t = qt(c,df,lower.tail = FALSE)
#Z CI
xbar + c(-1,1)*Z*sigma/sqrt(n)
```

```
## [1] 40.76988 45.23012
```

```
#t CI
xbar + c(-1,1)*t*s/sqrt(n)
```

```
## [1] 40.74005 45.25995
```

Notice how they are practically the same. This is why you may use Z or t. Either one will get you approximately the same CI. From both, we can say with 95% certainty that the true average number of touchdowns scored by the offense of a NCAAF team will be between about 41 and 45 TDs.

Example 5; Z vs. t, But With A Small Sample

Okay, now let's repeat example 4, but say you only have this as your sample.

```r
set.seed(314439)
cfb21_sample = cfb21[sample(nrow(cfb21), size = 5, replace = FALSE),] #Lecture4
```

Now let's recalculate everything using this sample.

```r
#Needed for both
xbar = mean(cfb21_sample$Off.TDs)
n = length(cfb21_sample$Off.TDs)
c = (1-0.95)/2
#Needed for Z CI
sigma = sd(cfb21_sample$Off.TDs)*sqrt((n-1)/n) #Pop. SD
Z = qnorm(c,lower.tail = FALSE)
#Needed for t CI
s = sd(cfb21_sample$Off.TDs) #Sample SD
df = n - 1
t = qt(c,df,lower.tail = FALSE)
#Z CI
xbar + c(-1,1)*Z*sigma/sqrt(n)
```

```
## [1] 37.10156 52.89844
```

```r
#t CI
xbar + c(-1,1)*t*s/sqrt(n)
```

```
## [1] 32.49058 57.50942
```

Notice here how the two intervals are **NOT** the same. The one using t is wider by about 5 TDs. This is because when the sample size is small, there is a greater chance for variation in the sample. Therefore, you want to use the t interval here because it has a greater chance of containing the true average number of touchdowns scored by the offense of a NCAAF team. Remember, use t when you do not know the population standard deviation. You may use it if n is large, but you *must* use it when n is small.

End of Lecture 12 Notes