For Live Session Assignment (FLS)

Unit 2

Question 1: Quick Quiz Questions

Quick Quiz Question (QQQ 1)

True or False: If a sample size is large, then the shape of a histogram of the sample data will be approximately normal, regardless of the shape of the original population distribution.

Select one:

- a. True
- b. False

QQQ 2

Suppose the following statement is made in a statistical summary:

A comparison of breathing capacities of individuals in households with low nitrogen dioxide levels and individuals in households with high nitrogen dioxide levels indicated that there is no difference in the means (two-sided p-value = 0.24).

What is wrong with this statement?

Select one:

- a. The researchers should have reported a one-sided p-value.
- b. Since we find in favor of the null hypothesis that the difference in means is equal to 0, the statement is correct.
- c. Finding in favor of the null hypothesis does not necessarily mean that that the difference in mean nitrogen dioxide levels between the two groups is 0, it means that the difference could plausibly be zero.
- d. The researchers did not account for confounding variables in the study.

QQQ 3

The data in Display 2.14 (page 53 in the text) are survival times (in days) of guinea pigs that were randomly assigned to a control group or to a treatment group that received a dose of tubercle bacilli. Which of the following models would be most appropriate for these data?

Select one:

- a. The additive treatment effect model (see section 1.3.1 of the text for a refresher) using a normal approximation with equal variances as a test statistic.
- b. Use an additive treatment effect model, but calculate a confidence interval instead of performing a hypothesis test.
- c. The model would need to take into account a shift in mean as well as changes in variability between the treatment groups.
- d. A permutation test must be used for these data because we do not know the shapes of the underlying populations of data.

QQQ 4

What is the formal definition of the ”pvalue”?

End Question 1:
Quick Quiz Questions

Question 2

Question from Concept Check 2.13 in Asynch

- Assume you wanted to test whether a new marketing strategy is increasing the mean total sales per day. You sample days at random under the old marketing system and 10 days at random under the new marketing system and record the mean sales for each group. You find that the 95 percent confidence interval for the difference of mean sales per day (μ_new-μ_old) is ($1.23, $1.60). You note that the new marketing strategy would cost $5.00 per day to implement. Is this result statistically significant? Why? Is this study practically significant? Why?
- Please respond below and add this question and your response to your FLS slide deck.

End Question 2

Question 3
($\leq$ 1 hour)

CLT Activity (From Section 2.2 of Asynch)

Consider again the Sleep 1 data from the asynch material. Using the same app:
http://www.rossmanchance.com/applets/OneSample.html?population=gettysburg

1. Now, take 500 random samples, each of size 5. What is the mean of the 500 means?
2. What is the range of the 500 sample means(smallest value and largest value)?
3. What is the standard deviation of the 500 sample means? You can find this value in the upper right corner of the graph on the lower right.
4. Describe the distribution of 500 sample means from samples of size 5 (think shape, center, and spread).
5. Run the following simulations using the applet. Click "reset" before each simulation!! For each simulation, obtain the mean and standard deviation of the 500 sample means.
    a. Number of samples = 500, sample size = 10
    b. Number of samples = 500, sample size = 20
    c. Number of samples = 500, samples size = 50
6. For each of a, b, and c, above, describe the shape of the distribution of the 500 sample means relative to the original population distribution.
7. For each of a, b, and c, calculate sigma divided by the square root of n and compare this to the standard deviation of the distribution of the sample means from your simulation.

CLT Activity

8. How do the distributions of simulated samples of the same sample size compare to one another (think center, shape, and spread). These are the samples that you see in the middle histogram.
9. Describe the distributions of the sample means for samples of size 5, 10, 20, and 50. How do these distributions compare to one another, and to the population distribution?
10. What would you expect to happen to the distribution of the sample means if we changed the sample size again, to a larger number?
11. What is the pattern that you see?
12. Fore Shadowing for the next section: Do you think the same pattern would result if we took samples and made distributions of sampling statistics for other types of populations (not sleep times)? Try it with “Pennies” and “Change”. (Note, the "Variable" needs to be set in the "Show Sampling Options" section.) What do you see?
13. Do you think the same pattern would result if we took samples and made distributions of other sampling statistics (not the mean)? Try it by punching the radio buttons for the median and standard deviation over the plot at the bottom right.

End Question 3

Question / Activity 4

Review the following … this is a huge concept in tying everything together and will help with HW, Midterm, Final and most importantly, you application of these methods.

Confidence Interval

The following are ages of 7 randomly selected patrons at the Beach Comber in South Mission Beach at 7pm! We assume that the data come from a normal distribution and would like to ***construct and interpret*** a 95% confidence interval for the actual mean age of patrons at the Comber. Assume we don't know the population standard deviation and have estimated it to be 7.08 years. In addition, the t multiplier (t critical value) for this problem is 2.447. The other statistics needed to construct the interval will need to be derived from the data. Show and fully explain your work.

25, 19, 37, 29, 40, 28, 31

95% confidence interval for mean age

Sample Ages: 25, 19, 37, 29, 40, 28, 31

Let's say we do NOT know σ (population standard deviation). We must estimate it using s (sample standard deviation).

$n = 7$

$\bar{x} = 29.86$

$s = 7.08$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$t_{\alpha/2,\, n-1} = 2.447$

$$\bar{x} - E < \mu < \bar{x} + E, \text{ where}$$

$$E = t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}} = (2.447)\frac{(7.08)}{\sqrt{7}} = 6.55$$

$$29.86 - 6.55 < \mu < 29.86 + 6.55$$

$$23.31 < \mu < 36.41$$

IMPORTANT: These are the ***plausible*** values of the mean given the data!

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in the interval (23.31 yrs., 36.41 yrs.).

95% confidence interval for mean age

Sample Ages: 25, 19, 37, 29, 40, 28, 31

Let's say we know σ (population standard deviation).

$n = 7$

$\bar{x} = 29.86$

$\sigma = 7.08$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

$$\bar{x} - E < \mu < \bar{x} + E, \text{ where}$$

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)\frac{(7.08)}{\sqrt{7}} = 5.24$$

$$29.86 - 5.24 < \mu < 29.86 + 5.24$$

$$24.62 < \mu < 35.10$$

IMPORTANT: These are the ***plausible*** values of the mean given the data!

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in the interval (24.62 years, 35.10 years).

Comparison of z to t

$n = 7$

$\bar{x} = 29.86$

$\sigma = 7.08$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

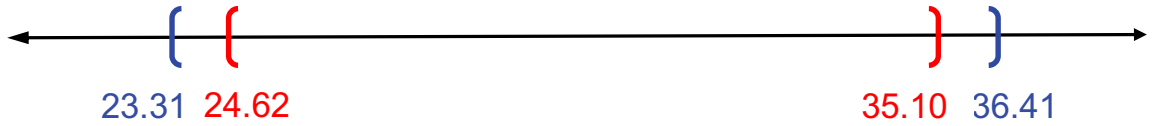
$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)(7.08) = 5.24 \qquad \sqrt{7}$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$29.86 - 5.24 < \mu < 29.86 + 5.24$

$24.62 < \mu < 35.10$

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in the interval (24.62 years, 35.10 years).



$n = 7$

$\bar{x} = 29.86$

$s = 7.08$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$t_{\alpha/2,\, n-1} = 2.447$

$$E = t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}} = (2.447)(7.08) = 6.55 \qquad \sqrt{7}$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$29.86 - 6.55 < \mu < 29.86 + 6.55$

$23.31 < \mu < 36.41$

We are 95% confident that the mean age of Beach Comber patrons at 7pm is contained in the interval (23.31 years, 36.41 years).

1 Sample Hypothesis Testing: The 6 Steps

1. Identify H_0 and H_a.
2. Find the Critical Value(s) and Draw and Shade.
3. Calculate the Test – Statistic. (The evidence!)
4. Calculate the p-value.
5. Make a decision… Reject H_0 or FTR H_0.
6. Write a clear conclusion in the context of the problem…. Use mostly non-statistical terms but always report the p-value! Add a confidence interval if appropriate. End this conclusion with a statement about the scope.

Example: 1 Sample t-test

The following are ages of 7 randomly chosen patrons seen leaving the Beach Comber in South Mission Beach at 7pm! We assume that the data come from a normal distribution and would like to test the claim that the mean age of the distribution of Comber patrons is different than 21. Conduct a 6 step one sample 2-sided t-test with alpha = .05 to test this claim. Provide enough explanation to fully describe each step. Recall that Step 2 is finding the critical value which is the same as the t multiplier in the corresponding 95% confidence interval.

25, 19, 37, 29, 40, 28, 31

Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is different than 21.

Step 1: Identify the null (H_0) and alternative (H_a) hypothesis.

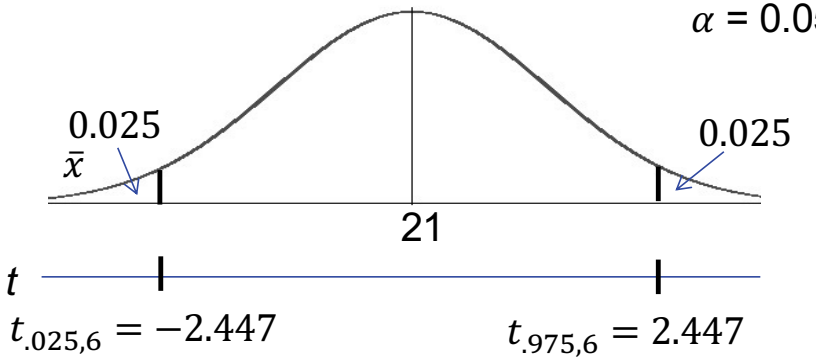
$$H_0: \mu = 21$$
$$H_a: \mu \neq 21$$

Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is different from 21.
To do this, we take a sample of size n = 7.

Step 1: Identify the null (H_0) and alternative (H_a) hypothesis. $H_0: \mu = 21$ $H_a: \mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



α = 0.05 = significance level.

df = 7 – 1 = 6

```
data critval;
 p = quantile("T",.975,6)
 ;
proc print data = critval;
 run;
```

Obs	p
1	2.44691

```
> qt(0.05/2, 7-1)
[1] -2.446912
```

Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 7 and find that $\bar{x}$ = 29.86 years and s = 7.08 years.

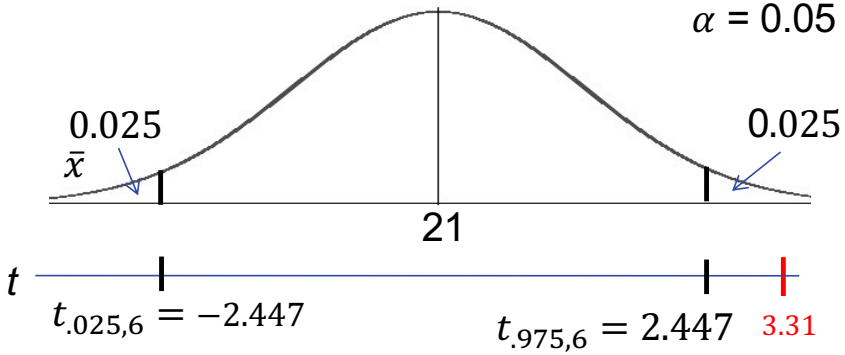
Step 1: Identify the null (H_0) and alternative (H_a) hypothesis. $H_0: \mu = 21$

$H_a: \mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.

α = 0.05 = significance level.

df = 7 – 1 = 6



Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{29.86 - 21}{\frac{7.08}{\sqrt{7}}} = 3.31$$

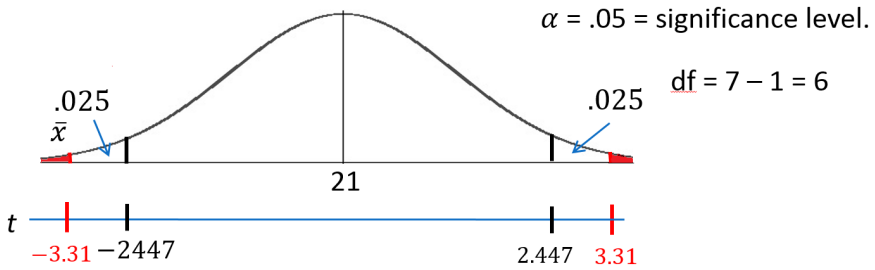
Let's Formalize This Test Into 6 Steps!

We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 8 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (H_0) and alternative (H_a) hypothesis. $H_0: \mu = 21$

$H_a: \mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.) $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{29.86 - 21}{\frac{7.09}{\sqrt{7}}} = 3.31$

Step 4: Find the p-value: *The probability of observing by random chance something as extreme or more extreme than what was observed under the assumption that the null hypothesis is true.* (Usually found with software.) The red shaded region above is 0.0162 (sum of both red areas)

Let's Formalize This Test Into 6 Steps!

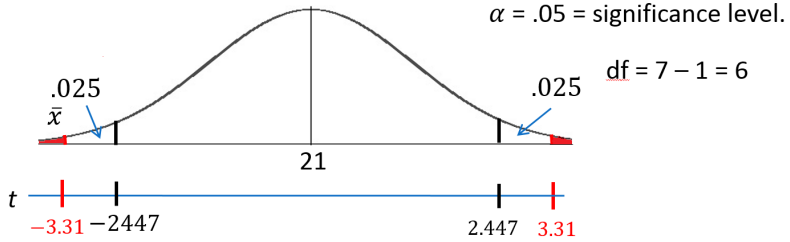
We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 8 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (H_0) and alternative (H_a) hypothesis.

H_0: $\mu = 21$

H_a: $\mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{29.86 - 21}{\frac{7.09}{\sqrt{7}}} = 3.31$$

Step 4: Find the p-value: P-value 0.0162< .05

Step 5: Key! The sample mean we found is very unusual under the assumption that the true mean age is 21. So we Reject the assumption that the true mean age is 21. That is, we REJECT H_0.

Let's Formalize This Test Into 6 Steps!

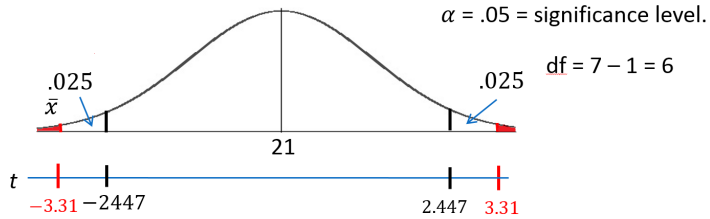
We would like to test the claim that the population mean is not equal to 21. To do this, we take a sample of size n = 8 and find that $\bar{x}$ = 29.86 years and s = 7.09 years.

Step 1: Identify the null (H_0) and alternative (H_a) hypothesis.

$H_0: \mu = 21$

$H_a: \mu \neq 21$

Step 2: Draw and Shade and Find the Critical Value.



Step 3: Find the test statistic. (The t value for the data.)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{29.86 - 21}{\frac{7.09}{\sqrt{7}}}$$

$$= \mathbf{3.31}$$

Step 4: Find the p-value: P-value 0.0162 < 0.05

Step 5: REJECT H_0

Step 6: There is sufficient evidence to conclude that the true mean age of patrons at the Comber at 7pm is not equal to 21 (p-value =0.0162 from a t-test). We could also say that there is sufficient evidence to conclude that the true mean is greater than 21. (Consider the red area in the right most tail.) This was not a random sample of all times, only at 7pm; thus, the result cannot be applied to the bar at all times. The results are nevertheless intriguing.

Finding the P-value – more detail

Step 4: Find the p-value: p-value < 0.05

You could use Stat Trek / or the t-table.

OR

Software like SAS:

```
data comber;
 input age @@;
 datalines;
 25 19 37 29 40 28 31
 ;

proc print data = comber;
 run;

proc ttest data = comber h0 = 21 sides = 2 alpha = .05;
 var age;
 run;
```

* The @@ symbol tells SAS to read the data “sideways” in a row.

Confidence interval

The TTEST Procedure

Variable: age

N	Mean	Std Dev	Std Err	Minimum	Maximum
7	29.8571	7.0812	2.6764	19.0000	40.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
29.8571	23.3082	36.4061	7.0812	4.5631	15.5932

DF	t Value	Pr > |t|
6	3.31	0.0162

Finding the P-value – more detail

Step 4: Find the p-value: p-value < 0.05

Using R:

```
> age = c(25, 19, 37, 29, 40, 28, 31)
> t.test(x=age, mu = 21, conf.int = "TRUE", alternative = "two.sided")

        One Sample t-test

data:  age
t = 3.3093, df = 6, p-value = 0.01622
alternative hypothesis: true mean is not equal to 21
95 percent confidence interval:
 23.30816 36.40613
sample estimates:
mean of x
 29.85714
```

Confidence interval

End Question/Activity 4

Question 5
($\leq$ 2 hours)

From The Homework! (Q 2)

1. In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below:

Fired

34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56

Not fired

27 33 36 37 38 38 39 42 42 43 43 44 44 44 45 45 45 45 46 46 47 47 48 48 49 49 51 51 52 54

a. Perform a permutation test to test the claim that there is age discrimination. Provide the Ho and Ha, the p-value, and full statistical conclusion, including the scope (inference on population and causal inference). Note: this was similar to an example in Live Session 1. You may start from scratch or use the sample code and PowerPoints from Live Session 1.

b. Now run a two sample t-test appropriate for this scientific problem. (Use SAS.) *(Note: we may not have talked much about a two-sided versus a one-sided test. If you would like to read the discussion on pg. 44 (Statistical Sleuth), you can run a one-sided test if it seems appropriate. Otherwise, just run a two-sided test as in class. There are also examples in the Statistics Bridge Course.)* Be sure to include all six steps, a statistical conclusion, and scope of inference.

c. Compare this p-value to the randomized p-value found in the previous sub-question.

d. The jury wants to see a range of plausible values for the difference in means between the fired and not fired groups. Provide them with a confidence interval for the difference of means and an interpretation.

e. Given the sample standard deviations from SAS, calculate by hand

   i. Pooled standard deviation (s_p)

   ii. The standard error of ($X_{FIRED} - X_{Not\ Fired}$)

f. Inspect and run this R Code and compare the results (t statistic, p-value, and confidence interval) to those you found in SAS. To run the code, simply copy and paste the code below into R.

```
Fired = c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)

Not_fired = c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)

t.test(x = Fired, y = Not_fired, conf.int = .95, var.equal = TRUE, alternative = "two.sided")
```

End Question 5

Question 6: Takeaways! (~ 1 Hour)

Please provide at least 4 takeaways from this section and any questions that you may have. These questions will help design the live session for this unit.

This question will be the last question of every For Live Session Assignment. The idea is that this deck will serve as a document that you can reference in the future to remember what was covered in this section. For instance, this may come in handy for the Capstone and will hopefully become useful in your career. Most immediately, it may become handy as a quick reference for your Midterm and Final! Some students find it very useful and spend a few slides summarizing the asynch material while others learn different ways and only had the minimum 4 takeaways. Either is fine and will earn you full credit for this question.

Question 7: Questions!

Please provide any question or topics of discussion that came up in this Unit! These will help help us optimize our live session for maximum learning and takeaways!

This question will be the last question of every For Live Session Assignment. The idea is that this deck will serve as a document that you can reference in the future to remember what was covered in this section. For instance, this may come in handy for the Capstone and will hopefully become useful in your career. Most immediately, it may become handy as a quick reference for your Midterm and Final! Some students find it very useful and spend a few slides summarizing the asynch material while others learn different ways and only had the minimum 4 takeaways. Either is fine and will earn you full credit for this question.

End For Live Session Assignment Unit 2!

DataScience@SMU