Week Eleven: Summary of Inferential Statistics (CI and Testing)

Week Eleven Goals

- Normal Probability Plot and the Normality Test (with video)
- Numerical Data versus Categorical Data---A summary (with video)
- Distinguish Between CI and Test (numerical and categorical) and Solutions
- Multiple-Choice Practice (with video) and Solutions

Normal Probability Plot and The Normality Test

The following information is meant to give you background knowledge on the NPP. **You do not have to do anything differently with the NPP or Normality Test once you read this information.** You will still build an NPP and compare the P-value to 0.05 to determine if we can assume a normal model when working with numerical data and n < 30. The purpose of this page of information is to share the reasoning behind the P-value and our conclusion.

Recall: The **Normal Probability Plot (NPP)** is a graphical technique for assessing whether or not sample data come from an approximately normally distributed population (generally used when n < 30). The normal probability plot is a graph that plots observed data versus normal scores (the expected z-score of the data value, assuming that the distribution of the random variable is normal). If this graph is approximately linear, then we can assume the sample data come from a normally distributed population. **We have been using a P-value to determine if we can assume the normality assumption has been met.**

Normality Test

We had not yet learned about hypothesis testing when I first introduced you to the NPP. We needed to meet a normality assumption and it was efficient to use the P-value of the NPP. It is time to tell you about the hypothesis test that is being conducted. A normality test assesses whether or not there is sufficient evidence to show that the sample could have come from a population that is NOT normally distributed. We will ask Minitab/StatHelper to generate the normal probability plot with an Anderson-Darling (AD) P-value.

Normality Test Null and Alternative Hypotheses

H_o: The population distribution is normal (assumption of a normal population)
H_a: The population distribution is NOT normal (not normal population)

Normality Test Conclusion

If the AD P-value is less than .05, then we CAN reject the assumption of a normal population (reject Ho). We have sufficient evidence to say that the population distribution is not normal.
Therefore, ***the normality assumption has not been met.***

If the AD P-value is greater than .05, then we CANNOT reject the assumption of a normal population (cannot reject Ho and we must support Ha). We DO NOT have sufficient evidence to say that the population distribution is NOT normal. This double-negative statement allows us to say:
The normality assumption has been met.

Video Discussing the Ho and Ha of a Normality Test from Prof. Coffey

Here is a video about the normality test: https://youtu.be/tghFq5A33Wg

Here is the same video with interpreting:

Numerical Data versus Categorical Data ---A Summary

Numerical Data

When the data we want to study has been measured and those numerical values would make sense to study as a mean, then we know we have numerical data. Often, we have the numerical data—the raw data—in an Excel spreadsheet. Other times, the summarized statistics are provided to us: The sample size, n, the sample mean, $\bar{x}$, and the sample standard deviation, s.

Summary Statistics with Numerical Data

If we are only interested in summarizing our sample, then, in addition to the summary descriptive statistics mentioned above, we could choose to find the 5-number summary and build a graphical display such as a box plot or histogram in order to summarize what we know about the sample collected.

Inferential Statistics with Means

In this course, we will work with a **T-distribution** each time we study numerical data in order to make an inference about the population. We use the sample data collected to make an estimate about the population mean by constructing **confidence intervals about the true mean**. Building a confidence interval is one way to make an inference about the population mean. Another way is with a **hypothesis test**. We can make an assumption about the population mean and use sample data to test that assumption using hypothesis testing. In order to build a confidence interval or conduct a hypothesis test and trust the results, we determine if we have met the assumption of normality regarding the population data. Statistical technology is used to build the CI or run the test and we always choose a **1-sample T Test for Means**.

Categorical Data

When the data we want to study is categorical, then the responses we are interested in will get counted (this is called the number of events). If we are given the raw, categorical data, then the response of interest would need to get counted and technology can do that for you (e.g. You have 150 responses of Yes or No and you can ask technology to count the Yes responses, so you do not have to!)

Summary Statistics with Categorical Data

To study the counted responses, we always form a proportion. We look at the number of events (X) and divide by the number of trials (n). The sample proportion is denoted by $\hat{p} = \frac{x}{n}$. Constructing a pie chart or bar chart are convenient ways to display the results.

Inferential Statistics with Proportions

We will work with the **Standard Normal Z-distribution** each time we study categorical data in order to make an inference about the population. We use the sample data collected to make an estimate about the population proportion by constructing **confidence intervals about the true proportion**. Building a confidence interval is one way to make an inference about the population proportion. Another way is with a **hypothesis test**. We can make an assumption about the population proportion and use sample data to test that assumption using hypothesis testing. In order to build a confidence interval or conduct a hypothesis test and trust the results, we determine if we have met the assumption of normality regarding the population data. Statistical technology is used to build the CI or run the test and we always choose a **1-sample Z Proportion Test.**

Video Discussing the different procedures for numerical and categorical data from Prof. Coffey

Here is a video about numerical vs. categorical data: https://youtu.be/J54CafhBVtM

Here is the same video with interpreting:

Determine the Type of Variable??

Numerical
(work on T-curve)

STUDY MEANS
$\bar{x}$ is the sample mean
s is the sample standard deviation
n is the sample size

Estimate μ by building a
confidence interval about μ

$$\bar{x} \pm (t_{crit})\left({}^{s}/_{\sqrt{n}}\right)$$

Make an assumption about μ and
Conduct a **hypothesis test**

Define μ =
Ho: μ = ___
Ha: μ <, >, ≠ _____

The assumption of normality must be met to work with the t-distribution.

Categorical
(work on Z-curve)

STUDY PROPORTIONS
$\hat{p}$ is the sample proportion $\hat{p} = \frac{x}{n}$
where x is the number of trials
and n is the number of events.

Estimate p by building a
confidence interval about p

$$\hat{p} \pm (Z_{crit})\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Make an assumption about p and
Conduct a **hypothesis test**

Define p =
Ho: p = _____
Ha: p <, >, ≠ _____

The assumption of normality must be met
To work with the Z-distribution.

Distinguish Between CI and Test (numerical and categorical)

Practice Problems

In each situation, complete the following:

- Describe the variable being studied and determine if it is numerical or categorical.
- Decide if you need a CI or a test to answer the question

1. Do college students cheat? It is commonly thought that only a small proportion of college students have looked at notes during a "closed-notes" exam. A professor would like to convince college administrators that the proportion is actually greater than 20%.

2. How long to people avoid going to the dentist? A local dentist wants to estimate the mean time (months) since adults in Rochester last visited a dentist, with 90% confidence.

3. The Ithaca, NY police department believes that cars on Triphammer Road are going at or below the 30 mph speed limit, on average. One homeowner on Triphammer Road would like to convince the police that cars on this road have a mean speed above the speed limit.

4. How many Americans think the lottery will make them rich? A financial advisor would like to estimate, with 95% confidence, the proportion of all adult Americans who believe that playing the lottery is the best strategy for becoming wealthy.

Solutions to Practice Problems

1. The variable being studied is whether or not students cheat (look at notes during closed-notes exam) and the variable is categorical (yes/no answers that will be counted. We are given $p = .20$ as a proportion we can assume is true and we will test to see if $p > .20$. This is a hypothesis test.

2. The variable is the time since adults in Rochester last visited a dentist (measured in months) and it is numerical. (You can see the mention of 'mean months' right in the problem). We are asked to estimate with 90% confidence so we will build a CI about μ.

3. The variable being studied is the speed at which cars travel on Triphammer Rd. and the variable is numerical (measured speed in miles per hour, mph). We are provided a mean speed of 30 to assume is true and we will look for evidence to see if $\mu > 30$ mph. This is a hypothesis test.

   ****Note: I have used this question long enough to know that some students suggest this could be answered by collecting counts on how many cars speed and how many do not. In this case, it would be categorical data. Sure, a **<u>different</u>** study could be conducted this way, but this question did ask us to study speed 'on average' and there is mention of a 'mean' speed. This is clearly being studied numerically as I mentioned above.****

4. The variable being studied is whether or not Americans think the lottery will make them rich. This a categorical variable (count the responses). We are asked to estimate the proportion and this is done with a confidence interval.

Multiple-Choice Practice

There are video discussions of the questions along with the answers at the end of the document.

Video Discussing Questions 1 -11 from Prof. Coffey

Here is a video on questions 1 – 11: https://youtu.be/5F9TIzVjbcM

Here is the same video with interpreting:

1. You take a random sample from some population and form a 95% confidence interval for the population mean, μ. Which quantity is guaranteed to be in the interval that you form?

a) 0 b) μ c) $\bar{x}$ d) .95

2. Suppose you conduct a significance test for the population proportion and your P-value is 0.184. Given a 0.10 level of significance (i.e. α = .10), which of the following should be your conclusion?

a) accept Ho b) accept Ha c) Fail to reject Ha d) Fail to reject Ho e) Reject Ho

3. Decreasing the sample size, while holding the confidence level the same, will do what to the length of your confidence interval?

a) make it bigger b) make it smaller

c) it will stay the same d) cannot be determined from the given information

4. Decreasing the confidence level, while holding the sample size the same, will do what to the length of your confidence interval?

a) make it bigger b) make it smaller

c) it will stay the same d) cannot be determined from the given information

5. If you increase the sample size and confidence level at the same time, what will happen to the length of your confidence interval?

a) make it bigger b) make it smaller

c) it will stay the same d) cannot be determined from the given information

6. Which of the following is true about P-values?

a) A P-value must be between 0 and 1.

b) If a P-value is greater than .01 you will never reject Ho.

c) P-values can be negative or positive.

d) None of the above are true.

7. Suppose that we wanted to estimate the true mean number of eggs a queen bee lays with 95% confidence. The margin of error we are willing to accept is 0.5. Suppose we also know that s is about 10. What sample size should we use?

a) 1536 b) 1537 c) 2653 d) 2650

8. What should be the value of z used in a 93% confidence interval?

a) 2.70 b) 1.40 c) 1.81 d) 1.89

9. Why do we use inferential statistics?

a) to help explain the outcomes of random phenomena

b) to make informed predictions about parameters we don't know

c) to describe samples that are normal and large enough (n>30)

d) to generate samples of random data for a more reliable analysis

10. A 95% confidence interval for the mean number of televisions per American household is (1.15, 4.20). For each of the following statements about the above confidence interval, **choose true or false.**

a) The probability that μ is between 1.15 and 4.20 is .95. True or False?

b) We are 95% confident that the true mean number of televisions per American household is between 1.15 and 4.20. True or False?

c) 95% of all samples should have x-bars between 1.15 and 4.20. True or False?

d) 95% of all American households have between 1.15 and 4.20 televisions. True or False?

e) Of 100 intervals calculated the same way (with 95% confidence), we expect 95 of them to capture the population mean. True or False?

f) Of 100 intervals calculated the same way (with 95% confidence), we expect 100 of them to capture the sample mean. True or False?

11. When doing a significance test, a student gets a P-value of 0.003. This means that:

I. Assuming Ho were true, this sample's results were an unlikely event.

II. 99.97% of samples should give results which fall in this interval.

III. We reject Ho at any reasonable alpha level.

a) II only b) III only c) I and III d) I, II, and III

Video Discussing Questions 12 - 20 from Prof. Coffey

Here is a video on questions 12 – 20: https://youtu.be/sNbDmkHyzms

Here is the same video with interpreting:

12. Parameters and statistics...

a) Are both used to make inferences about $\bar{x}$

b) Describe the population and the sample, respectively.

c) Describe the sample and the population, respectively.

d) Describe the same group of individuals.

13. The workers at Richter, Inc. took a random sample of 800 manhole covers and found that 40 of them were defective. What is the 95% CI for p, the true proportion of defective manhole covers, based on this sample?

a) (37.26, 42.74) b) (.035, .065) c) (.047, .053) d) (.015, .085)

14. Researchers are designing a study to determine whether the age of the victim is a factor in reported scams. The researchers are testing to see if more than half of all reported scams victimize the elderly (> 80 years old). They randomly sample 350 reported scams over the past 10 years from the Better Business Bureau archives, and note that, for 187 of them, the victim is elderly. Match the following symbols with the correct number. Use technology to get a test statistic and P-value.

a) 0.50

b) 1.28

c) 350

d) 0.53

e) 187

f) .09977

1. x
2. n
3. $\hat{p}$
4. p_0
5. $P - value$
6. Z test statistic

15. When are p-values negative?

a) when the test statistic is negative.

b) when the sample statistic is smaller than the proposed value of the parameter

c) when the confidence interval includes only negative values

d) when we fail to reject the null hypothesis

e) never

16. Suppose the P-value for a test is .028. Which of the following is true?

a) At alpha = .05, we will not reject Ho

b) At alpha = .05, we will reject Ho

c) At alpha = .01, we will reject Ho

d) None of the above is true.

17. A random sample of married people were asked "Would you remarry your spouse if you were given the opportunity for a second time?"; Of the 150 people surveyed, 127 of them said that they would do so. Find a 95% confidence interval for the proportion of married people who would remarry their spouse.

a) 0.847 ± 0.002

b) 0.847 ± 0.029

c) 0.847 ± 0.048

d) 0.847 ± 0.058

18. A survey was conducted to get an estimate of the proportion of smokers among the graduate students. A past report says 38% of them are smokers. Prof. Coffey doubts the result and thinks that the actual proportion is much less than this. Choose the correct choice for the null and alternative hypothesis.

a) Ho: $p = .38$ versus Ha: $p \le .38$.

b) Ho: $p = .38$ versus Ha: $p > .38$.

c) Ho: $p = .38$ versus Ha: $p < .38$.

d) None of the above.

Questions 19 & 20: Suppose we are interested in finding a 95% confidence interval for the mean SAT Verbal score of students at a certain high school. We can assume a normal distribution. **Five students are sampled, and their SAT Verbal scores are 560, 500, 470, 660, and 640.**

19. What is the standard error of the sample mean? a) 16.71 b) 37.36 c) 83.55 d) 113.2

20. What is the 95% confidence interval for the population mean?

a) (462.3, 669.7) b) (469.9, 662.1) c) (486.3, 645.7) d) (492.8, 639.2)

ANSWERS to Multiple Choice Practice Above

1.	C
2.	D
3.	A
4.	B
5.	D
6.	A
7.	B
8.	C
9.	B
10a.	F
10b.	T
10c.	F
10d.	F
10e.	T
10f.	F
11.	C
12.	B
13.	B
14a.	4 (P_0)
14b.	6 (z)
14c.	2 (n)
14d.	3 ($\hat{p}$)
14e.	1 (x)
14f.	5 (p-value)
15.	E
16.	B
17.	D
18.	C
19.	B
20.	A