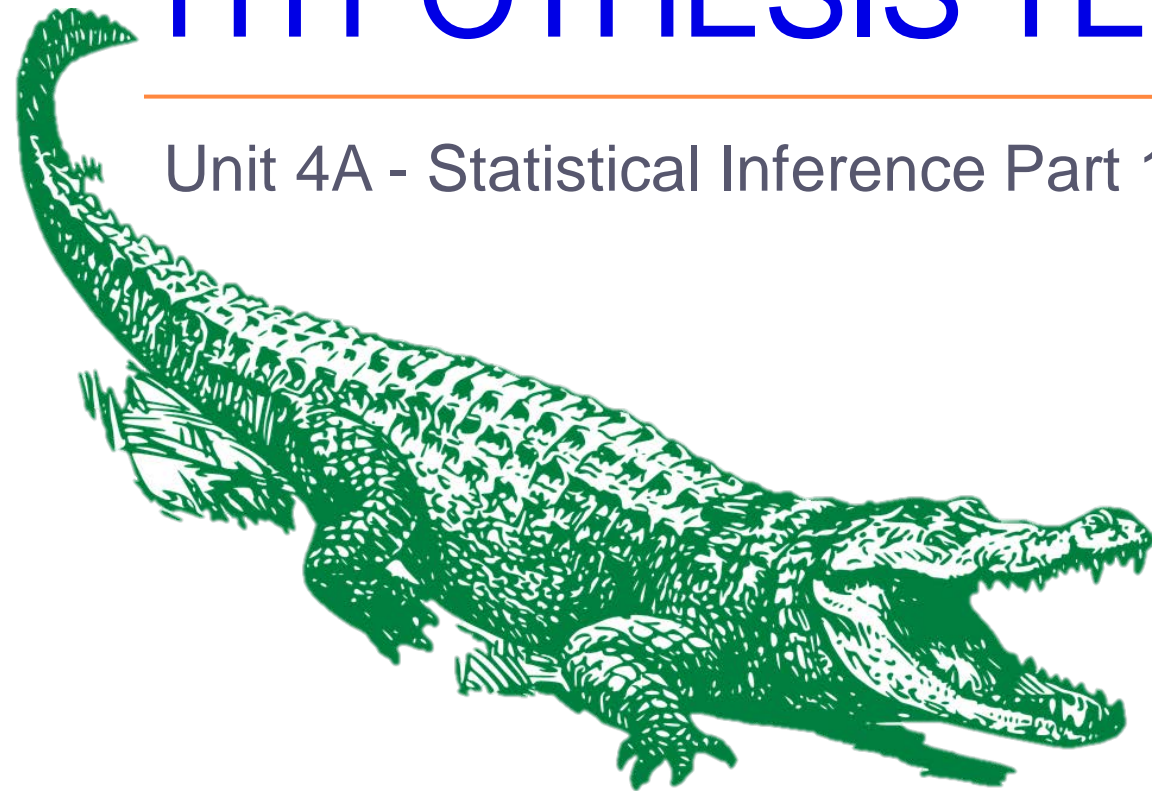
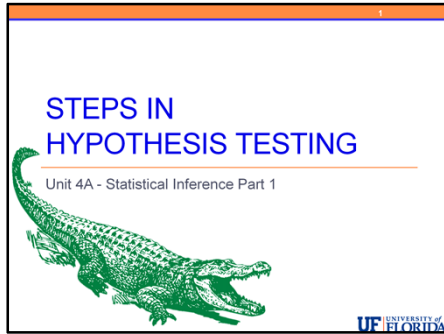




STEPS IN HYPOTHESIS TESTING

Unit 4A - Statistical Inference Part 1





Now we will look closer at each step using the three motivating examples discussed in the materials in the introduction to hypothesis testing.

Let's review each situation.



Example 1: Smoking a GU

- A recent study estimated that 20% of all college students in the United States smoke. The head of Health Services at GU suspects that the proportion of smokers may be lower at GU. In hopes of confirming her claim, the head of Health Services chooses a random sample of 400 GU students, and finds that 70 of them are smokers.

Example 1: Smoking at GU

- A recent study estimated that 20% of all college students in the United States smoke. The head of Health Services at GU suspects that the proportion of smokers may be lower at GU. In hopes of confirming her claim, the head of Health Services chooses a random sample of 400 GU students, and finds that 70 of them are smokers.

UF UNIVERSITY OF FLORIDA

Example 1 – Smoking at GU

A recent study estimated that 20% of all college students in the United States smoke.

The head of Health Services at GU suspects that the proportion of smokers may be lower at GU.

In hopes of confirming her claim, the head of Health Services chooses a random sample of 400 GU students, and finds that 70 of them are smokers.

Example 2: Allergy Medication

- A certain prescription allergy medicine is supposed to contain an average of 245 parts per million (ppm) of a certain chemical. If the concentration is higher than 245 ppm, the drug will likely cause unpleasant side effects, and if the concentration is below 245 ppm, the drug may be ineffective. The manufacturer wants to check whether the mean concentration in a large shipment is the required 245 ppm or not. To this end, a random sample of 64 portions from the large shipment is tested, and it is found that the sample mean concentration is 250 ppm with a sample standard deviation of 12 ppm.

Example 2: Allergy Medication

- A certain prescription allergy medicine is supposed to contain an average of 245 parts per million (ppm) of a certain chemical. If the concentration is higher than 245 ppm, the drug will likely cause unpleasant side effects, and if the concentration is below 245 ppm, the drug may be ineffective. The manufacturer wants to check whether the mean concentration in a large shipment is the required 245 ppm or not. To this end, a random sample of 64 portions from the large shipment is tested, and it is found that the sample mean concentration is 250 ppm with a sample standard deviation of 12 ppm.

UF UNIVERSITY OF FLORIDA

Example 2: Allergy Medication

A certain prescription allergy medicine is supposed to contain an average of 245 parts per million (ppm) of a certain chemical.

If the concentration is higher than 245 ppm, the drug will likely cause unpleasant side effects, and if the concentration is below 245 ppm, the drug may be ineffective.

The manufacturer wants to check whether the mean concentration in a large shipment is the required 245 ppm or not.

To this end, a random sample of 64 portions from the large shipment is tested, and it is found that the sample mean concentration is 250 ppm with a sample standard deviation of 12 ppm.

Example 3: SAT vs. Gender

- Is there a relationship between gender and combined scores (Math + Verbal) on the SAT exam?
- Following a report on the College Board website, which showed that in 2003, males scored generally higher than females on the SAT exam, an educational researcher wanted to check whether this was also the case in her school district. The researcher chose random samples of 150 males and 150 females from her school district, collected data on their SAT performance and found:

| Females | | |
|---------|------|--------------------|
| n | mean | standard deviation |
| 150 | 1010 | 206 |

| Males | | |
|-------|------|--------------------|
| n | mean | standard deviation |
| 150 | 1025 | 212 |

Example 3: SAT vs. Gender

- Is there a relationship between gender and combined scores (Math + Verbal) on the SAT exam?
- Following a report on the College Board website, which showed that in 2003, males scored generally higher than females on the SAT exam, an educational researcher wanted to check whether this was also the case in her school district. The researcher chose random samples of 150 males and 150 females from her school district, collected data on their SAT performance and found:

| Females | | | Males | | |
|---------|------|--------------------|-------|------|--------------------|
| n | mean | standard deviation | n | mean | standard deviation |
| 150 | 1010 | 206 | 150 | 1025 | 212 |

UNIVERSITY OF
FLORIDA

Example 3: SAT vs. Gender

Is there a relationship between gender and combined scores (Math + Verbal) on the SAT exam?

Following a report on the College Board website, which showed that in 2003, males scored generally higher than females on the SAT exam, an educational researcher wanted to check whether this was also the case in her school district.

The researcher chose random samples of 150 males and 150 females from her school district, collected data on their SAT performance and found:

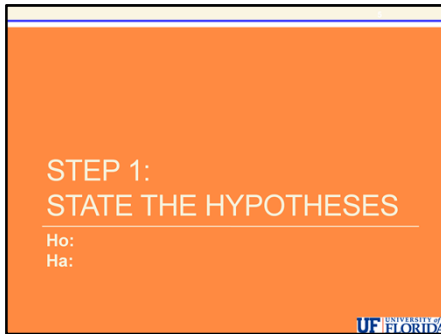
- The mean for females is 1010 and the standard deviation is 206
- The mean for males is 1025 and the standard deviation is 212



STEP 1: STATE THE HYPOTHESES

Ho:

Ha:



In step 1, we need to set up our statistical hypotheses.

We will now use the statistical notation of H-zero or H-naught to represent the null hypothesis and H-a to represent the alternative hypothesis.

The null hypothesis suggests nothing special is going on; in other words, there is no change from the status quo, no difference from the traditional state of affairs, or no relationship – depending on the situation at hand.

In contrast, the alternative hypothesis disagrees with this, stating that something is going on, or there is a change from the status quo, or there is a difference from the traditional state of affairs.

The alternative hypothesis, H_a , usually represents what we want to check or what we suspect is really going on.



STEP 1 – Examples

Example 1:

- **H₀:** The proportion of smokers at GU is 0.20.
- **H_a:** The proportion of smokers at GU is less than 0.20.

Example 2:

- **H₀:** The mean concentration in the shipment is the required 245 ppm.
- **H_a:** The mean concentration in the shipment is not the required 245 ppm.

Example 3:

- **H₀:** Performance on the SAT is not related to gender (males and females score the same).
- **H_a:** Performance on the SAT is related to gender – males score higher.

8

STEP 1 – Examples

Example 1:

- **Ho:** The proportion of smokers at GU is 0.20.
- **Ha:** The proportion of smokers at GU is less than 0.20.

Example 2:

- **Ho:** The mean concentration in the shipment is the required 245 ppm.
- **Ha:** The mean concentration in the shipment is not the required 245 ppm.

Example 3:

- **Ho:** Performance on the SAT is not related to gender (males and females score the same).
- **Ha:** Performance on the SAT is related to gender – males score higher.

UNIVERSITY OF FLORIDA

In each of our examples we have the following hypotheses using the Ho and Ha notation.

In example 1:

- **Ho:** The proportion of smokers at GU is 0.20.
- **Ha:** The proportion of smokers at GU is less than 0.20.

In example 2:

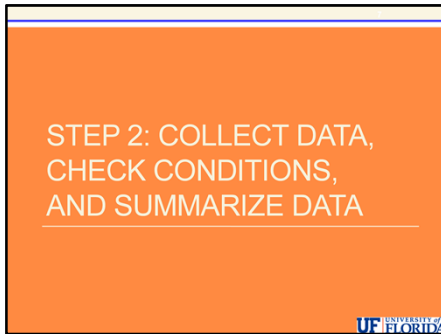
- **Ho:** The mean concentration in the shipment is the required 245 ppm.
- **Ha:** The mean concentration in the shipment is not the required 245 ppm.

In example 3:

- **Ho:** Performance on the SAT is not related to gender (males and females score the same).
- **Ha:** Performance on the SAT is related to gender – males score higher.



STEP 2: COLLECT DATA, CHECK CONDITIONS, AND SUMMARIZE DATA



In Step 2: We Collect data, check conditions, and Summarize Data

We look at sampled data in order to draw conclusions about the entire population.

In the case of hypothesis testing, based on the data, you draw conclusions about whether or not there is enough evidence to reject H_0 .

There is, however, one detail that we would like to add here. In this step we collect data and **summarize** it.

Note that in order to summarize the data we used simple sample statistics such as the sample proportion (\hat{p}), sample mean (\bar{x}) and the sample standard deviation (s).

In practice, you go a step further and use these sample statistics to summarize the data with what's called a **test statistic**.

This step will also involve checking any conditions or assumptions required to use the test.

We are not going to go into any details right now, as we will discuss test statistics when we go through the specific tests.



STEP 2 – Examples

Example 1:

- $n = 400$
- $p\text{-hat} = 70/400 = 0.175$
- Test statistic = ? (later...)

Example 2:

- $n = 64$
- $\bar{x} = 250$
- $s = 12$
- Test statistic = ? (later...)

Example 3:

- Females: $\bar{x} = 1010$ and $s = 206$
- Males: $\bar{x} = 1025$ and $s = 212$
- Test statistic = ? (later...)

STEP 2 – Examples

| | |
|-------------------|---|
| Example 1: | <ul style="list-style-type: none"> • n = 400 • $p\text{-hat}$ = $70/400 = 0.175$ • Test statistic = ? (later...) |
| Example 2: | <ul style="list-style-type: none"> • n = 64 • \bar{x} = 250 • s = 12 • Test statistic = ? (later...) |
| Example 3: | <ul style="list-style-type: none"> • Females: $\bar{x} = 1010$ and $s = 206$ • Males: $\bar{x} = 1025$ and $s = 212$ • Test statistic = ? (later...) |

UNIVERSITY OF FLORIDA

Here is a review of the summary statistics we have for these examples. We will add the test statistic when we get to each type of test.

When we are interested in the population proportion such as the population proportion of smokers at GU:

- **We gather the sample size and the sample proportion.**

When we are interested in the population mean such as the population mean concentration in a shipment:

- **We gather the sample size, the sample mean, and the sample standard deviation.**
- From this point, we will rarely assume that we know the population standard deviation again.

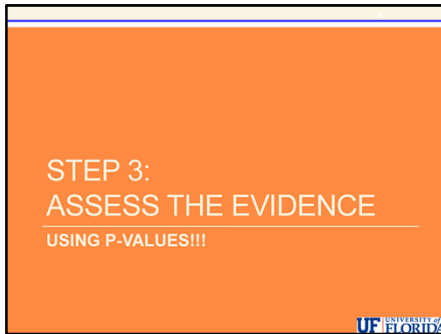
When we are interested in comparing two population means, such as the population mean SAT scores for males and females:

- **We gather the sample size, sample mean, and sample standard deviation for each gender.**



STEP 3: ASSESS THE EVIDENCE

USING P-VALUES!!!



In Step 3 We Assess the Evidence

This is the step where we calculate how likely is it to get data like that observed (or more extreme) ASSUMING H_0 is true.

In a sense, this is the heart of the process, since we draw our conclusions based on this probability.

- If this **probability is very small** (see example 2), then that means that it would be very surprising to get data like that observed (or more extreme) if H_0 were true.
 - The fact that we did observe such data is therefore **evidence against H_0** , and we should reject it.
- On the other hand, if this **probability is not very small** (see example 3) this means that observing data like that observed (or more extreme) is not very surprising if H_0 were true.
 - The fact that we observed such data **does not provide evidence against H_0** .

These statements may take a while to sink in!

Think about the examples carefully and try to make this probability argument make sense to you... and ask if you have questions.

This crucial probability has a special name. It is called the **p-value** of the test.



STEP 3 – Examples

Example 1:

- P-VALUE = 0.106

Example 2:

- P-VALUE = 0.0007

Example 3:

- P-VALUE = 0.29

10

STEP 3 – Examples

Example 1: • P-VALUE = 0.106

Example 2: • P-VALUE = 0.0007

Example 3: • P-VALUE = 0.29

UF UNIVERSITY OF FLORIDA

We aren't looking yet at how to obtain these p-values so don't worry about that now, however, in our three examples, the p-values are:

- **Example 1:** p-value = 0.106
- **Example 2:** p-value = 0.0007
- **Example 3:** p-value = 0.29

The smaller the p-value, the more surprising it is to get data like ours (or more extreme) when H_0 is true, and therefore, the stronger the evidence the data provide against H_0 .

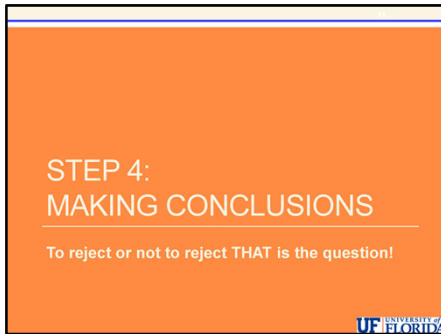
Looking at the three p-values of our three examples, we see that:

- the data that we observed in example 2 provide the strongest evidence against the null hypothesis
- followed by example 1
- while the data in example 3 provides the least evidence against H_0 .



STEP 4: MAKING CONCLUSIONS

To reject or not to reject THAT is the question!



Step 4: Making Conclusions

Since our conclusion is based on how small the p-value is, which measures how surprising our data are when H_0 is true, we need a guideline or cutoff that will help determine how small the p-value must be, or how “rare” (unlikely) our data must be when H_0 is true, for us to conclude that we have enough evidence to reject H_0 .

This cutoff exists, and because it is so important, it has a special name. It is called the **significance level of the test** and is usually denoted by the Greek letter α (alpha).

The most commonly used significance level is α (alpha) = 0.05 (or 5%) which we will use in this course. This means that:

- if the p-value $< \alpha$ (alpha) (usually 0.05), then the data we obtained is considered to be “rare (or surprising) enough” assuming H_0 is true, and we say that:
 - The data provide significant evidence against H_0
 - So we reject H_0 and accept H_a .
 - In this case we can say the results are “statistically significant.”
- if the p-value $> \alpha$ (alpha) (usually 0.05), then our data are not considered to be “surprising enough” assuming H_0 is true, and we say that:
 - Our data do not provide enough evidence to reject H_0
 - (or, equivalently, that the data do not provide enough evidence to accept H_a).
 - In this case we could say the results are “not statistically significant.”

The most important part of this step will be translating the results into the context of the current problem.



STEP 3 – Examples

Example 1:

- Using our cutoff of 0.05, we fail to reject H_0 .
- **Conclusion:** There **IS NOT** enough evidence that the proportion of smokers at GU is less than 0.20

Example 2:

- Using our cutoff of 0.05, we reject H_0 .
- **Conclusion:** There **IS** enough evidence that the mean concentration in the shipment is not the required 245 ppm.

Example 3:

- Using our cutoff of 0.05, we fail to reject H_0 .
- **Conclusion:** There **IS NOT** enough evidence that males score higher on average than females on the SAT.

12

STEP 3 – Examples

| | |
|------------|--|
| Example 1: | <ul style="list-style-type: none"> Using our cutoff of 0.05, we fail to reject H_0. Conclusion: There IS NOT enough evidence that the proportion of smokers at GU is less than 0.20 |
| Example 2: | <ul style="list-style-type: none"> Using our cutoff of 0.05, we reject H_0. Conclusion: There IS enough evidence that the mean concentration in the shipment is not the required 245 ppm. |
| Example 3: | <ul style="list-style-type: none"> Using our cutoff of 0.05, we fail to reject H_0. Conclusion: There IS NOT enough evidence that males score higher on average than females on the SAT. |

UNIVERSITY OF FLORIDA

Now that we have a cutoff to use, here are the appropriate conclusions for each of our examples based upon the p-values we were given.

- **Example 1:** **p-value = 0.106** Using our cutoff of 0.05, we fail to reject H_0 .

Conclusion: There **IS NOT** enough evidence that the proportion of smokers at GU is less than 0.20

- **Example 2:** **p-value = 0.0007** Using our cutoff of 0.05, we reject H_0 .

Conclusion: There **IS** enough evidence that the mean concentration in the shipment is not the required 245 ppm.

- **Example 3:** **p-value = 0.29** Using our cutoff of 0.05, we fail to reject H_0 .

Conclusion: There **IS NOT** enough evidence that males score higher on average than females on the SAT.

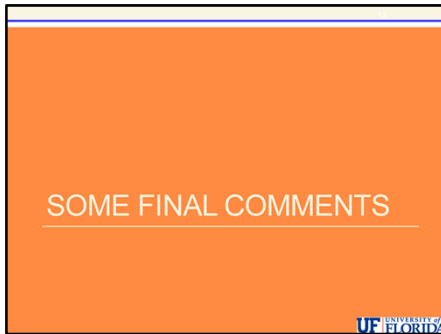
Notice that all of the above conclusions are written in terms of the alternative hypothesis and are given in the context of the situation.

In no situation have we claimed the null hypothesis is true.

Be very careful of this and be sure to provide a conclusion in the words of the scenario.



SOME FINAL COMMENTS



We have given a broad overview of the process of hypothesis testing in statistics.

We will be working with specific tests for the remainder of the semester and will be going back through this process in each of these settings.

You shouldn't feel that you can conduct any test yourself at this point but hopefully you have a good picture of the overall process.

Try to assess what you most need to work on and focus on that as needed during the coming weeks, returning to this material as needed and asking questions.

Before we finish, we have a few important points to mention.


Comments

- Cutoff of 0.05 for p-values is not “MAGIC” and is open to interpretation especially around 0.05
 - **P-values: 0.049 vs. 0.051**
- Conclusions:
 - **BEST WRITTEN IN TERMS OF ALTERNATIVE HYPOTHESIS:**
Is there or is there not evidence that ***H_a is true***?
 - **IN CONTEXT:** Write ***H_a is true*** in the words of the situation
 - **NEVER say** that you accept the null hypothesis is true

14

Comments

- Cutoff of 0.05 for p-values is not "MAGIC" and is open to interpretation especially around 0.05
 - P-values: 0.049 vs. 0.051
- Conclusions:
 - BEST WRITTEN IN TERMS OF ALTERNATIVE HYPOTHESIS:
Is there or is there not evidence that ***H_a is true?***
 - IN CONTEXT: Write ***H_a is true*** in the words of the situation
 - NEVER say that you accept the null hypothesis is true



First, the cutoff of 5% is not a magic number and is open to some interpretation, especially around this value.

- For example, what is the difference between a p-value of 0.049 and 0.051 in terms of evidence?
 - They are both about the same.
 - I am not more or less excited about either of these values
 - The only difference is that one of them let's me legitimately reject the null using the 5% cutoff
 - But both offer very similar evidence.
 - We much prefer to see very small p-values where the conclusion is clear.

Next, when writing conclusions to any hypothesis test, especially in this course,

- It is best to write the conclusion in terms of the alternative hypothesis.
 - **Is there or is there not enough evidence that *H_a is True?***
- You must write your conclusion in context – it must be clear in the words of the problem as well as using accurate statistical terminology as appropriate.
- We can never say that we accept the null hypothesis. We can only say that we have failed to disprove it. Be very careful on this point. There is an example illustrating this point in the course materials if you need further convincing.



Common Terminology in Journals

- “The results are statistically significant” ($p\text{-value} < \alpha$)
 - “The results are not statistically significant” ($p\text{-value} > \alpha$)
-
- $0.01 \leq p\text{-value} < 0.05$ \Rightarrow (statistically) *significant*.
 - $0.001 \leq p\text{-value} < 0.01$ \Rightarrow *highly significant*.
 - $p\text{-value} < 0.001$ \Rightarrow *very highly significant*.
 - $p\text{-value} > 0.05$ \Rightarrow *not statistically significant* (NS).
 - $0.05 \leq p\text{-value} < 0.10$ \Rightarrow *marginally significant*.

18

Common Terminology in Journals

- "The results are statistically significant" ($p\text{-value} < \alpha$)
- "The results are not statistically significant" ($p\text{-value} > \alpha$)

| | |
|--------------------------------------|---|
| • $0.01 \leq p\text{-value} < 0.05$ | ⇔ (statistically) <i>significant</i> . |
| • $0.001 \leq p\text{-value} < 0.01$ | ⇔ <i>highly significant</i> . |
| • $p\text{-value} < 0.001$ | ⇔ <i>very highly significant</i> . |
| • $p\text{-value} > 0.05$ | ⇔ <i>not statistically significant (NS)</i> . |
| • $0.05 \leq p\text{-value} < 0.10$ | ⇔ <i>marginally significant</i> . |

UNIVERSITY OF FLORIDA

Often in journals you will read statements such as

- "The results are statistically significant" – when the $p\text{-value} < \alpha$ (alpha) or
- "The results are not statistically significant" – when the $p\text{-value} > \alpha$ (alpha).

Sometimes statistical significance is qualified based upon the evidence for example:

- If the $p\text{-value}$ is between 0.001 and 0.01 we could say the result is highly significant
- If the $p\text{-value}$ is < 0.001 we could say it is VERY highly significant and
- If it is between 0.05 and 0.10 we sometimes call this marginally significant or trending towards significance.


P-VALUE

- = probability of obtaining results like those of our data (or more extreme) GIVEN the null hypothesis is true
- **P(Obtaining results like ours or more extreme | H_0 is True).**
- We are asking “Assuming the null hypothesis is true, how rare is it to observe something as or more extreme than what I have found in my data?”
- **Can also be thought of as the probability, assuming the null hypothesis is true, that the result we have seen is solely due to random error (or chance).**

16

P-VALUE

- = probability of obtaining results like those of our data (or more extreme) GIVEN the null hypothesis is true
- $P(\text{Obtaining results like ours or more extreme} \mid H_0 \text{ is True})$.
- We are asking "Assuming the null hypothesis is true, how rare is it to observe something as or more extreme than what I have found in my data?"
- Can also be thought of as the probability, assuming the null hypothesis is true, that the result we have seen is solely due to random error (or chance).



It is important that you find a definition of the p-value which makes sense to you.

It is also very important to remember that we begin by assuming the null hypothesis is true in order to find this probability.

The p-value is the probability of obtaining results like those of our data (or more extreme) GIVEN the null hypothesis is true

In probability notation we can write: $P(\text{Obtaining results like ours or more extreme} \mid H_0 \text{ is True})$.

We are asking "Assuming the null hypothesis is true, how rare is it to observe something as or more extreme than what I have found in my data?"

We can also think of the p-value as the probability, assuming the null hypothesis is true, that the result we have seen is solely due to random error (or chance).

No matter how we look at it, a small probability gives us evidence that the null hypothesis is not true and thus our alternative is true – which is usually what we are hoping to show.



Remember:

- We infer that the alternative hypothesis is true ONLY by rejecting the null hypothesis
- A statistically significant result is one that has a very low probability of occurring if the null hypothesis is true

17

Remember:

- We infer that the alternative hypothesis is true ONLY by rejecting the null hypothesis
- A statistically significant result is one that has a very low probability of occurring if the null hypothesis is true

UF UNIVERSITY OF FLORIDA

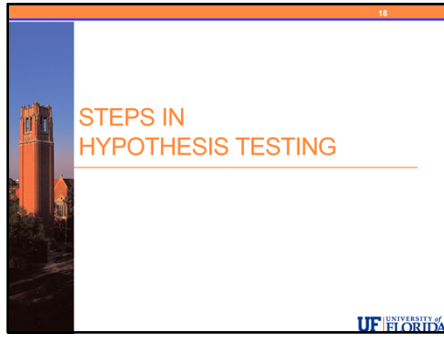
Remember:

We infer that the alternative hypothesis is true ONLY by rejecting the null hypothesis

A statistically significant result is one that has a very low probability of occurring if the null hypothesis is true.



STEPS IN HYPOTHESIS TESTING



Next we will discuss the difficult concept of ERRORS and POWER after which we will begin working some examples ourselves.