KEY

Activity 9.1a Test of Hypothesis Basics and Errors

There are 2 types of statistical inference

- Confidence Interval-CI (Chapter 8):
  - estimates all the plausible values for a population parameter. It gives us more information. Our course covers "p" and "μ."
- Significance Tests-TOH (Chapters 9-12):
  - are formal procedure for comparing observed data with a claim (hypothesis) whose truth we want to assess.
  - We express the results of a significance test in terms of a probability (p-value) that measures how well the data and the claim agree.

9.1 CONCEPTS YOU MUST KNOW

DEFINE HYPOTHESES:

- State the parameter of interest
- Null Hypothesis
- Alternative Hypothesis

STATISTICAL INFERENCE:

- Claim
- 2 Outcomes of a Statistical Test
- P-Value
- Significance Level
- Statistically Significant

Error:

- Type I
- Type II
- Power (covered next class)

Example: The Basketball Player

Setting up Significance Tests

EXAMPLE:

Ben claims that he makes 80% of his free-throws.

1. **What is the population parameter we want to test?**

   p = TRUE PROPORTION OF FREE-THROWS BEN MADE.

2. **What is our first claim that we are seeking to gather evidence against? This is the null hypothesis. Express in symbols and words:**

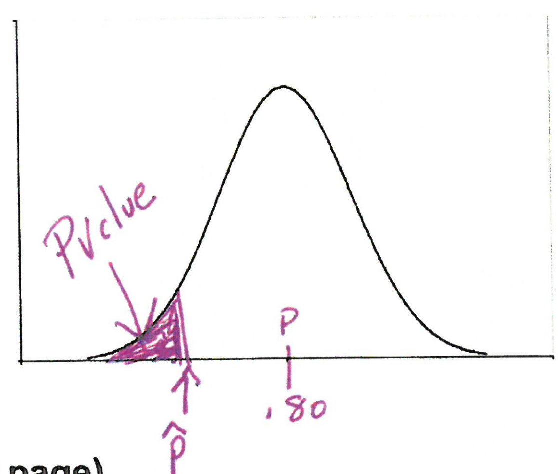
   $H_0: p = .80$ (BEN'S TRUE FOUL SHOOTING IS 80%)

3. **What is our second claim that we suspect to be true instead of the null hypothesis. There are 3 possible scenarios to consider for the alternative hypothesis.**

State the alternate hypothesis and sketch the graph:

- **Alternate Scenario #1:** We think Ben is exaggerating and can't possibly shoot that well.

  $H_A: p < .80$ (LEFT TAIL TEST)



(next page)

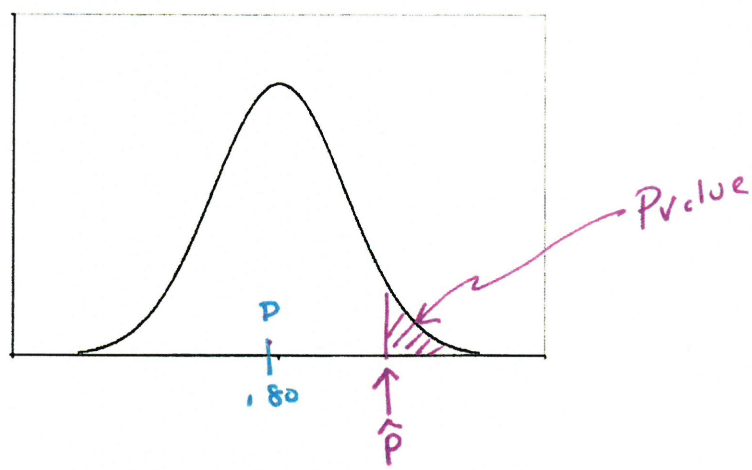
- **Example**: The Basketball Player
- ***Setting up Significance Tests (cont.)***

EXAMPLE: Ben claims that he makes 80% of his free-throws.

State the alternate hypothesis and sketch the graph:

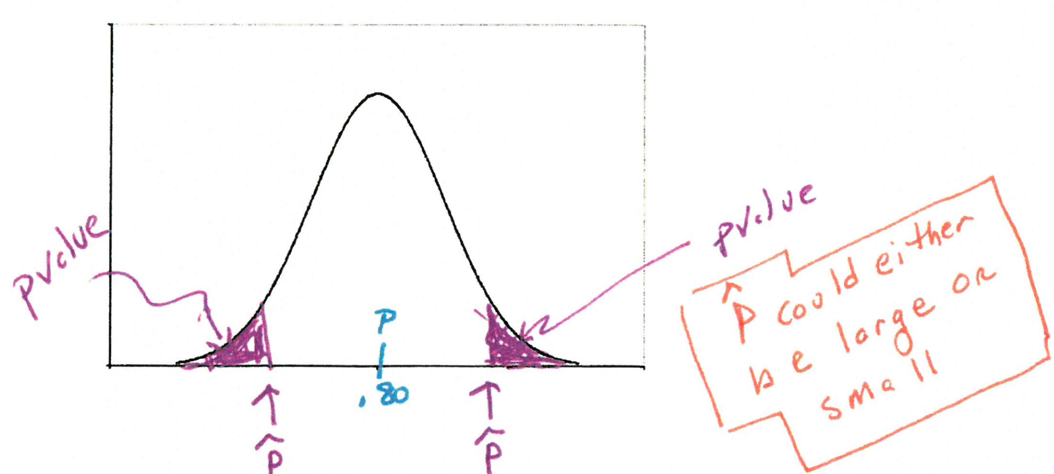
- **Alternate Scenario #2:** We think Ben is being modest, and is the best free-throw shooter in the state.

H_A: $p > .80$ (RIGHT TAIL TEST)



- **Alternate Scenario #3:** We simply think Ben is lying.

H_A: $p \neq .80$ (2 TAIL TEST)



Example: The Basketball Player

The Reasoning of Significance Tests

EXAMPLE: Now we gather evidence. We do not think Ben makes 80% of his free throws. So, we have him attempt 50 free-throws. He makes 32 of them. His sample proportion of made shots is 32/50 = 0.64. What can we conclude about the claim based on this sample data?

Option 1: What hypothesis do we want to test if we think Ben is exaggerating?

$H_0: p = .8$

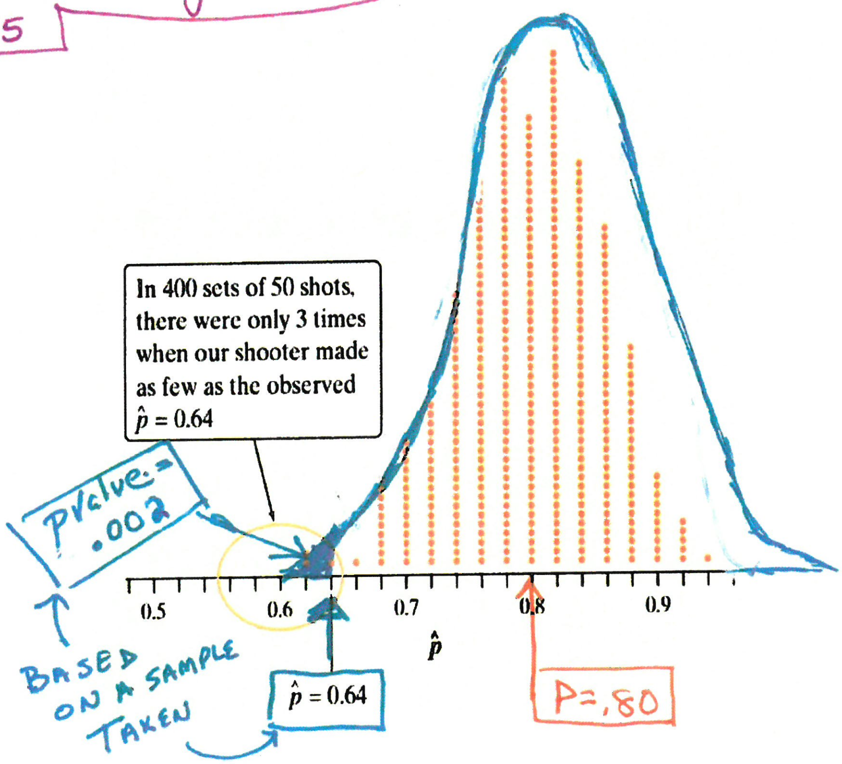
$H_A: p < .8$

Do we have enough evidence to reject our null hypothesis?

We typiclly set the significcnce level to $\alpha = .05$

Basic Idea

An outcome that would rarely happen if a claim were true is good evidence that the claim is not true.



CONCLUSION

Since the pvalue (.002) is VERY SMALL (EXTREME VALUE) AND LESS THAN OUR PREDETERMINED SIGNIFICANCE LEVEL ($\alpha = .05$) WE HAVE CONVINCING EVIDENCE TO REJECT H_0 AND HAVE SUFFICIENT EVIDENCE BEN SHOOTS LESS THAN 80%.

■ Example: The Basketball Player

■ *Would a Confidence Interval Provide the Evidence?*

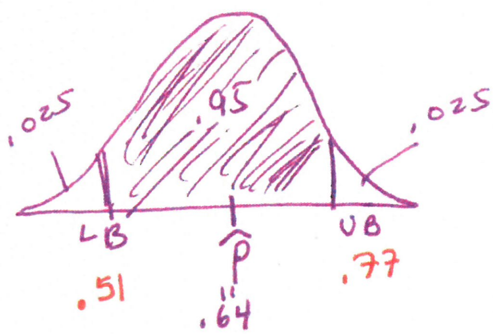
<u>Option 2</u>: **Let's try a different hypothesis. We don't believe Ben is an 80% free-throw shooter.**

1. What hypothesis do we want to test?

$H_0 : p = .80$

$: p \neq .80$

2. What evidence do we have (assume conditions of random, independent and normal are met)? $\boxed{\textbf{Create a 95\% CI.}}$



$\hat{p} = .64 \quad n = 50 \quad z^* = 1.96$

$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$.64 \pm 1.96 \cdot \sqrt{\frac{(.64)(.36)}{50}}$

$.64 \pm 1.96\,(.068)$ $SE(\hat{p})$

$.64 \pm .13$ ME

$LB = .64 - .13 = \boxed{.51}$

$UB = .64 + .13 = \boxed{.77}$

3. Do we have enough evidence to reject our null hypothesis?

<u>OUR CI</u>: WE ARE 95% CONFIDENT THAT THE TRUE FOUL SHOOTING PROPORTION IS BETWEEN 51% AND 77%.

$\boxed{\text{DO WE HAVE CONVINCING EVIDENCE.}}$

$\boxed{\text{YES}}$ SINCE OUR CI DOES NOT INCLUDE .80, WE HAVE CONVINCING EVIDENCE THAT BEN IS NOT AN 80% FREE THROW SHOOTER. (WE WOULD HAVE REJECTED H_0)

SUMMARY:

Stating Hypotheses In any significance test

1) The **null hypothesis** has the form

$$H_0: \underline{\text{PARAMETER}} = \underline{\text{VALUE}}$$

(μ, p)

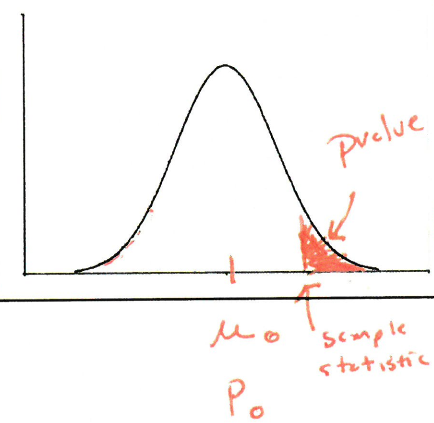
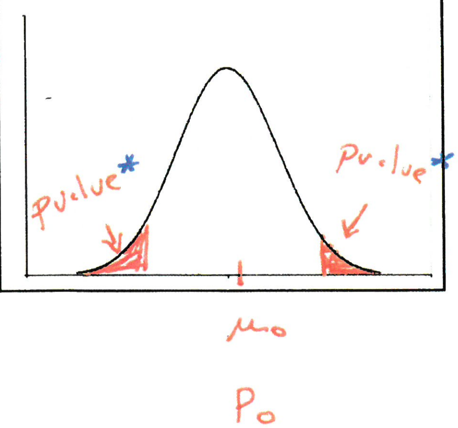
EX

$H_0: \mu = 5$

$H_0: p = .2$

2) The **alternative hypothesis** has one of 3 forms

- To determine the correct form of H_a, read the problem carefully!!
- Determine the symbol (=,>,<,≠)
- Label the P-value.

Left-sided Test - H_a : parameter $\underline{<}$ value	Right-sided Test - H_a : parameter $\underline{>}$ value	Two-sided Test - H_a : parameter $\underline{\neq}$ value
		

* pvalue combines both areas

Stating Hypotheses – Try this practice problem

Studying Job Satisfaction - see *Activity 9.1A (answer key)*

EXAMPLE: Does the job satisfaction of assembly-line workers differ when their work is machine-paced rather than self-paced? One study chose 18 subjects at random from a company with over 200 workers who assembled electronic devices. Half of the workers were assigned at random to each of two groups. Both groups did similar assembly work, but one group was allowed to pace themselves while the other group used an assembly line that moved at a fixed pace. After two weeks, all the workers took a test of job satisfaction. Then they switched work setups and took the test again after two more weeks. The response variable is the difference in satisfaction scores, self-paced minus machine-paced.

a) Describe the parameter of interest in this setting.

μ = DIFFERENCE IN SATISFACTION SCORES [SELF PACED − MACHINE PACED]

b) State appropriate hypotheses for performing a significance test. (in symbols and words)

$H_0: \mu = 0$ (No difference in job satisfaction scores)

$H_A: \mu \neq 0$ (there is a difference in job satisfaction scores)

NOTE: Workers could either be MORE OR LESS SATISFIED

Significance Tests: The Basics

- A **significance test** assesses the evidence provided by data against a **null hypothesis** H_0 in favor of an **alternative hypothesis** H_a.
- **We use the** ***P*****-value of a test** <u>**and**</u> **our predetermined** α **(alpha - the significance level), to make decisions regarding our hypothesis.**
- The ***P*****-value** of a test is the probability, computed supposing H_0 to be true, that the statistic will take a value at least as extreme as that actually observed in the direction specified by the alternate hypothesis H_a.
  - <u>**Small *P*-values**</u> indicate strong evidence against H_0. To calculate a *P*-value, we must know the sampling distribution of the test statistic when H_0 is true. There is no universal rule for how small a *P*-value in a significance test provides convincing evidence against the null hypothesis.
  - If the *P*-value is smaller than a specified value α (called the **significance level**), the data are **statistically significant** at level α. In that case, we can reject H_0. If the *P*-value is greater than or equal to α, we fail to reject H_0.

- **General Rule:**
  - <u>**Small *P*-values**</u> we <u>REJECT</u> the null hypothesis.
  - <u>**Large *P*-values**</u> we <u>FAIL TO REJECT</u> the null hypothesis.
  - We <u>NEVER accept</u> the null hypothesis.

BAD

Statistically Significance at level α – Try this practice problem

Better Batteries – see Activity 9.1A (answer key)

A company has developed a new deluxe AAA battery that is supposed to last longer than its regular AAA battery. However, these new batteries are more expensive to produce, so the company would like to be convinced that they really do last longer. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained. A significance test is performed using the hypotheses

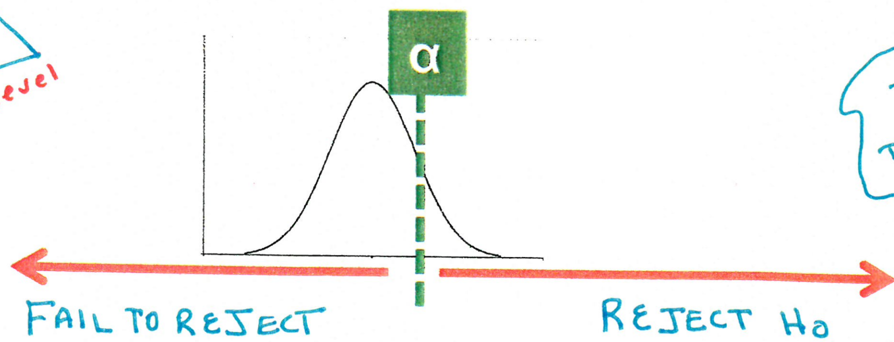
$$H_0: \mu = 30 \text{ hours}$$
$$H_a: \mu > 30 \text{ hours}$$

where μ is the true mean lifetime of the new deluxe AAA batteries.
The resulting $\boxed{P\text{-value is } 0.0276}$. ← GIVEN

For each, provide a supporting graph labeling the mean, P-value, and alpha(α).

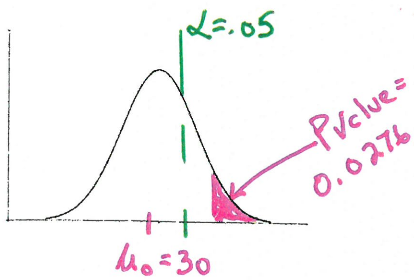
Researcher SETS ALPHA

α = significance level



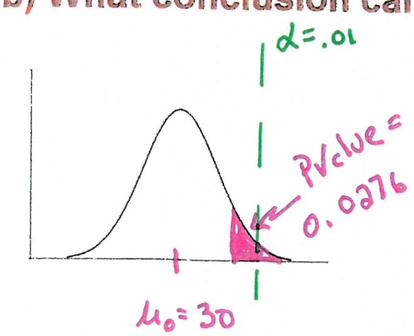
DECISIONS BASED ON P-VALUE

a) What conclusion can you make for the significance level $\alpha = 0.05$?



Since the pvalue (0.0276) is LESS THAN α=.05, we reject Ho,
WE HAVE SUFFICIENT EVIDENCE TO CONCLUDE the AAA BATTERIES LAST LONGER THAN 30 HOURS

b) What conclusion can you make for the significance level $\alpha = 0.01$?



Since the pvalue (0.0276) is GREATER THAN α=.05, we fail to reject Ho. WE DO NOT HAVE ENOUGH EVIDENCE TO CONCLUDE THE AAA BATTERIES LAST LONGER THAN 30 HOURS.

Introduction to Type I and Type II Errors:

EXAMPLE: Our Court System "O.J. Analogy"

Understand Hypothesis Testing

- In our jury system, you are innocent until proven guilty. This is how we are going to set up our statistical test of hypothesis statements:

  H_0: p = O.J. not guilty (innocent) ← Null hypothesis. H_o -> "H not"

  H_a: $p \neq$ O.J. guilty ← Alternate hypothesis

- The lawyers give evidence to prove their case (we will do the same by taking a sample).
- The jury comes back with the verdict based on whether this was a criminal or civil trial.
  - Criminal Trial evidence must be convincing "Beyond a reasonable doubt."
  - Civil Trial evidence must be convincing "By a preponderance of the evidence"
  - Which has a lower threshold? This threshold is comparable to our significance level(α). We predetermine "α" based on how much of an error we are willing to make. Typically, α=.05 or α=.01.

- The Jury decides:
  1. "GUILTY," if they have enough evidence. *We will do the same… If we have enough evidence, we "**REJECT Ho.**"*
  2. Or the jury says "NOT GUILTY." *We will do the same… If we do not have enough evidence, we "**Fail to reject H_o.**"*
  3. The Jury **never** says "INNOCENT," because OJ will never tell us the truth. We **NEVER accept** the null hypothesis because we have a chance of making a mistake.

- **Discussion Questions**:
  a) OJ was found "not guilty" in the criminal trial. He was found "guilty" in the civil trial. Why?

  b) What are 2 possible errors that could happen in our jury system?

  TYPE 1: JURY FINDS OJ Guilty, BUT HE IS INNOCENT

  TYPE 2: JURY FIND OJ NOT Guilty, WHEN HE IS GUILTY

Type I and Type II Errors

- When we draw a conclusion from a significance test, we hope our conclusion will be correct. But sometimes it will be wrong.
- **There are two types of mistakes we can make and it is very important how recognize and interpret these errors!!!!**
  - If WE REJECT H_0 when H_0 is TRUE, we have committed a Type I error. (+,+)
  - If WE FAIL TO REJECT H_0 when H_0 is FALSE (OR H_a is true), we have committed a Type II error. (−,−)
  - Fill in table:

Truth about the population

		H_0 true	H_0 false (H_a true)
Conclusion based on sample	Reject H_0	**Type I error** (α)	CORRECT ☺ conclusion (Power = 1 − β)
	Fail to reject H_0	CORRECT ☺ conclusion	**Type II error** (β)

STATED ANOTHER WAY:

$\boxed{\text{TYPE I ERROR}}$ "The null hypothesis is TRUE, but ... WRONG DECISION"

$\boxed{\text{TYPE II ERROR}}$ "The alternative hypothesis is TRUE, but ... WRONG DECISION"

Example "Perfect Potatoes"

A potato chip producer and its main supplier agree that each shipment of potatoes must meet certain quality standards. If the producer determines that more than 8% of the potatoes in the shipment have "blemishes," the truck will be sent away to get another load of potatoes from the supplier. Otherwise, the entire truckload will be used to make potato chips. To make the decision, a supervisor will inspect a random sample of potatoes from the shipment. The producer will then perform a significance test using the hypotheses

$H_0: p = 0.08$ POTATOES MET THE STANDARD

$H_a: p > 0.08$ TOO MANY BAD POTATOES

where p is the actual proportion of potatoes with blemishes in a given truckload

Describe Type I & Type II error in this setting; explain consequences of each:

- A **Type I error** would occur if… THE COMPANY FINDS CONVINCING EVIDENCE THERE WERE TOO MANY BAD POTATOES, WHEN IN FACT THE PROPORTION OF BAD POTATOES IS 8%

(Reject H_0, TRUE)

- *Consequence*: TRUCKS OF POTATOES ARE SENT AWAY WITH GOOD POTATOES AND THE COMPANY LOSES MONEY.

- A **Type II error** would occur if… THE COMPANY DOES NOT FIND CONVINCING EVIDENCE THE POTATOES WERE BAD, WHEN IN FACT THE PROPORTION OF BAD POTATOES IS GREATER THAN 8%

[FAIL TO REJECT, H_0 FALSE]

- *Consequence*: THE COMPANY WILL MAKE CHIPS WITH BAD POTATOES AND UPSET CUSTOMERS.

WHICH CONSEQUENCE IS WORSE?

* You can argue either way.

IF TYPE I IS MORE SERIOUS → lower α

IF TYPE II IS MORE SERIOUS → increase α

Type I and II Errors– Try this practice problem

Faster fast food?" - see Activity 9.1B (answer key)

Example "*Faster fast food?*" The manager of a fast-food restaurant want to reduce the proportion of drive-through customers who have to wait more than 2 minutes to receive their food once their order is placed. Based on store records, the proportion of customers who had to wait at least 2 minutes was p = 0.63. To reduce this proportion, the manager assigns an additional employee to assist with drive-through orders. During the next month the manager will collect a random sample of drive-through times and test the following hypotheses:

$H_0: p = 0.\cancel{0}63$ Customer Service did NOT improve

$H_a: p < 0.63$ Customer Service did improve

where *p = the true proportion of drive-through customers who have to wait more than 2 minutes after their order is placed to receive their food.*

Describe Type I & Type II error in this setting; explain consequences of each:

- A **Type I error** would occur if… THE MANAGER FINDS CONVINCING EVIDENCE CUSTOMER SERVICE HAS IMPROVED, WHEN IN FACT 63% OF CUSTOMERS WAIT LONGER THAN 2 MINUTES

[Reject H_0 when H_0 TRUE
$P(\text{TYPE I}) = \alpha$]

- Consequence:

↳ MANAGER IS SPEND MORE $'s FOR ADDITIONAL EMPLOYEE AND CUSTOMER SERVICE HAS NOT IMPROVED

- A **Type II error** would occur if… THE MANAGER DOES NOT FIND CONVINCING EVIDENCE CUSTOMER SERVICE HAS IMPROVED WITH EXTRA EMPLOYEE, WHEN IN FACT 63% OF CUSTOMERS WAIT LESS THAN 2 MINUTES

[FAIL TO REJECT H_0, WHEN H_A TRUE]

- Consequence:

↳ MANAGER FIRES ADDITIONAL EMPLOYEE AND UPSET CUSTOMERS WITH POOR SERVICE

AP Stats Calculating Power, Type I and Type II Errors

WHAT YOU NEED TO KNOW !

Quote from AP Statistics Teacher Forum

"Do not try to teach any calculations about Type II error or power. Not only is that not required, it can be confusing and it distracts students from understanding the concepts. They need to know what the two types of error are and what power is. They need to be able to explain them in the context of the questions. And they need to understand the interactions among the errors, power, sample size and effect size. But no calculations!"