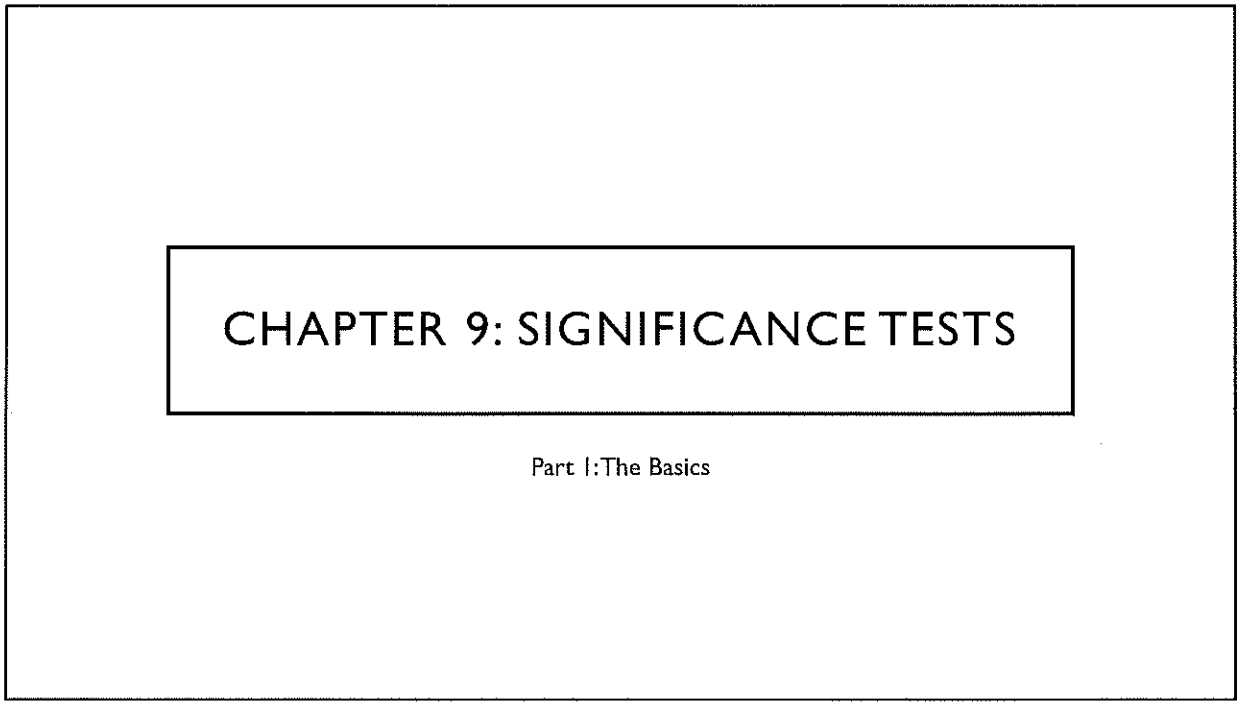
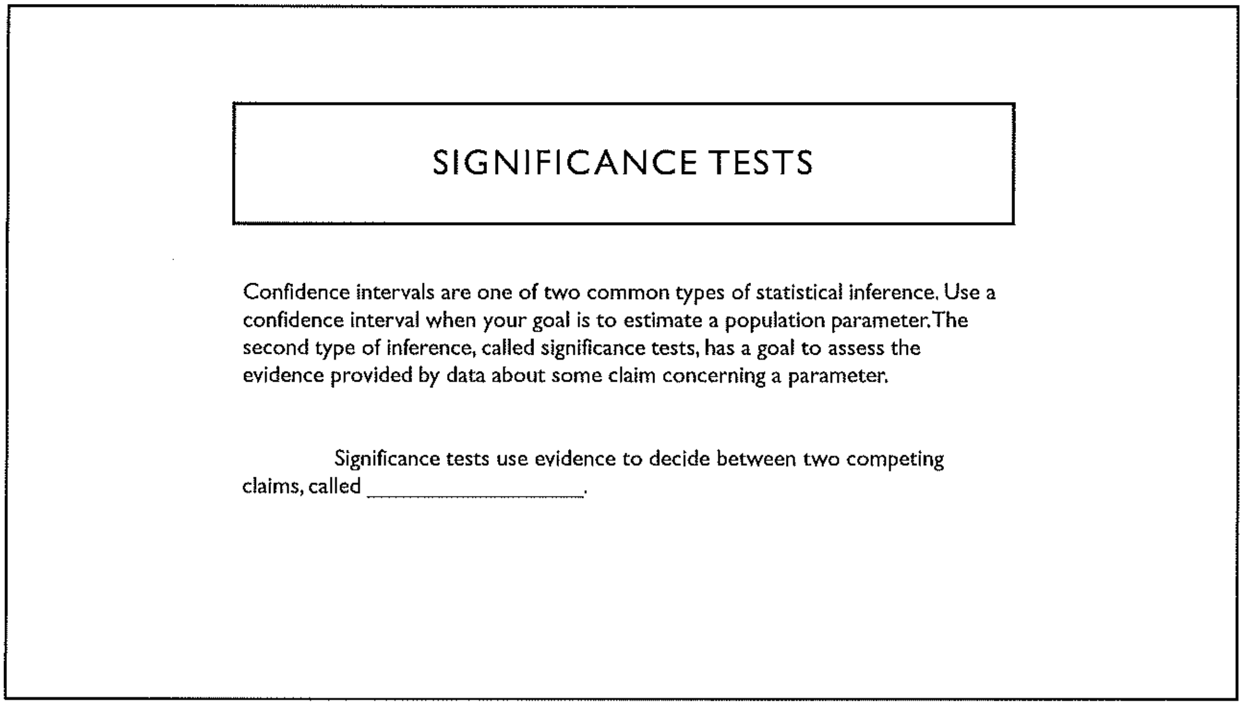


CHAPTER 9: SIGNIFICANCE TESTS

Part 1:The Basics

Open with activity pg 538

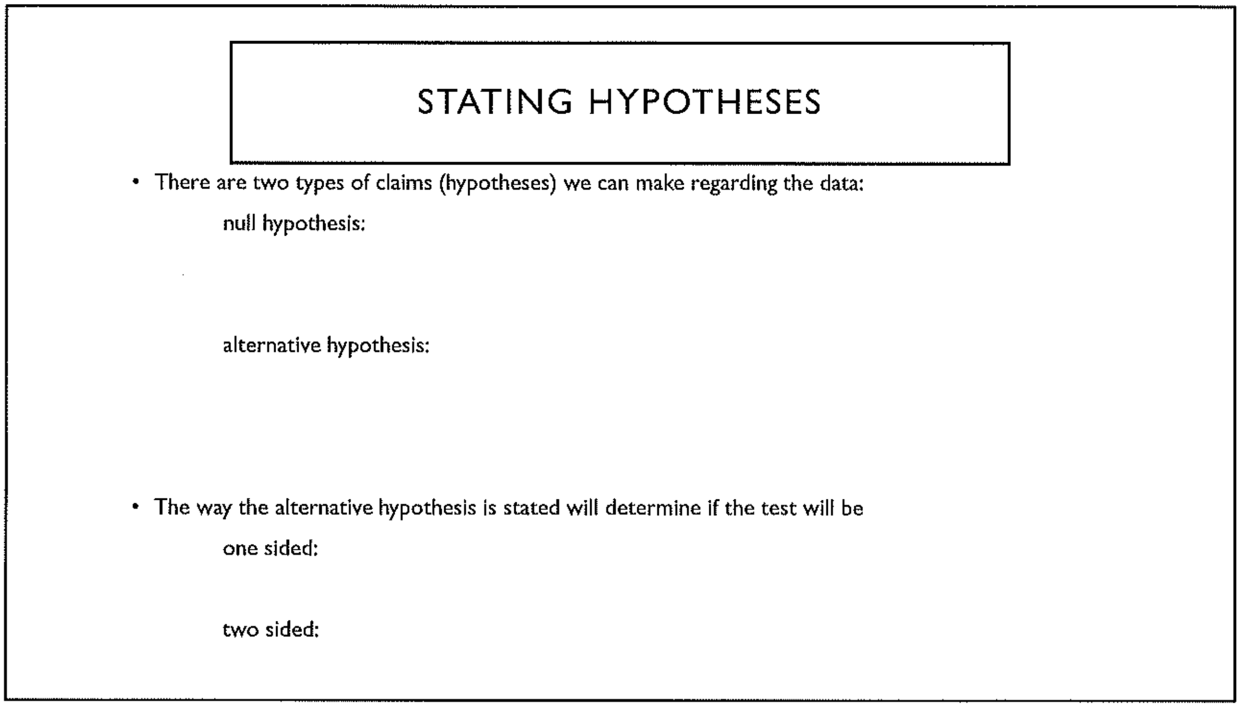


SIGNIFICANCE TESTS

Confidence intervals are one of two common types of statistical inference. Use a confidence interval when your goal is to estimate a population parameter.The second type of inference, called significance tests, has a goal to assess the evidence provided by data about some claim concerning a parameter.

Significance tests use evidence to decide between two competing claims, called ____________________.

i.e.



STATING HYPOTHESES

- There are two types of claims (hypotheses) we can make regarding the data:

  null hypothesis:

  alternative hypothesis:

- The way the alternative hypothesis is stated will determine if the test will be

  one sided:

  two sided:

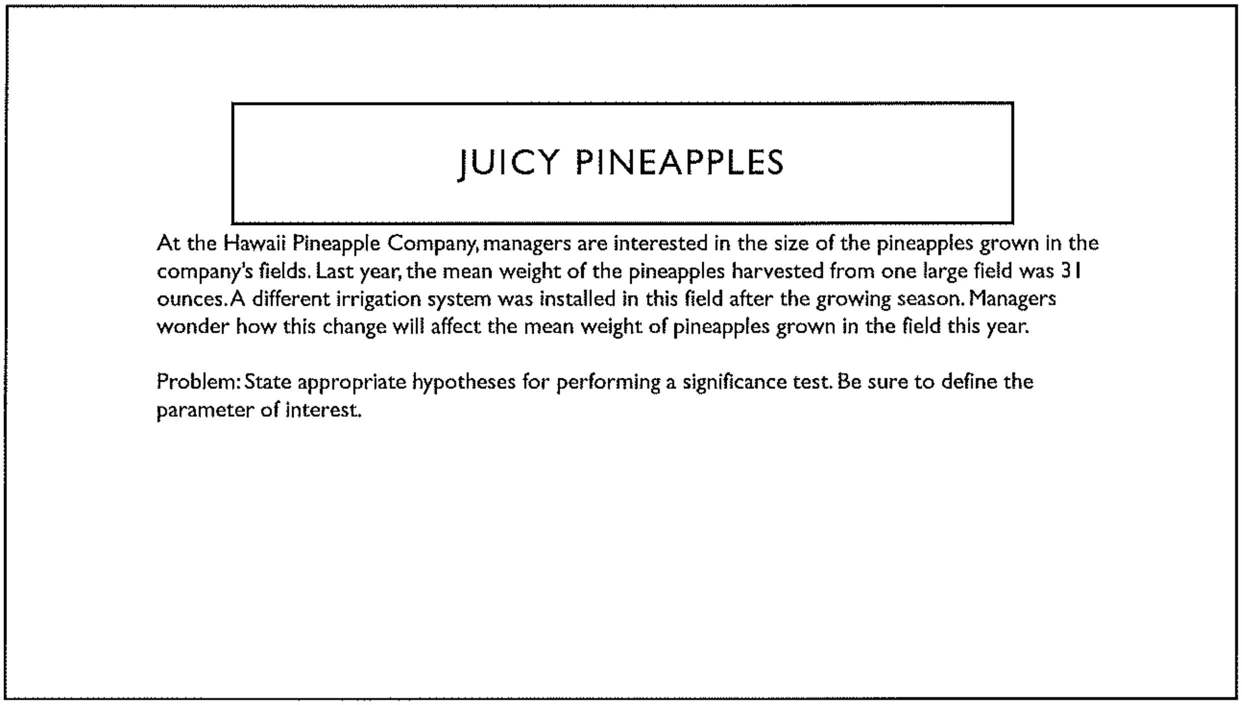
Null: H_0 → "no different from what's claimed" ; what we weight evidence against

Alternative: H_a → "the claim we suspect to be true instead" ; what we try to find evidence for

For the free throw shooter:
H_0: $p=.80$
H_A: $p<.80$

One sided: parameter is either < or > the null value
Two sided: =/= the null value

**HYPOTHESES ALWAYS HAVE TO DO WITH POPULATION PARAMETERS, SO WRITE THEM AS SUCH



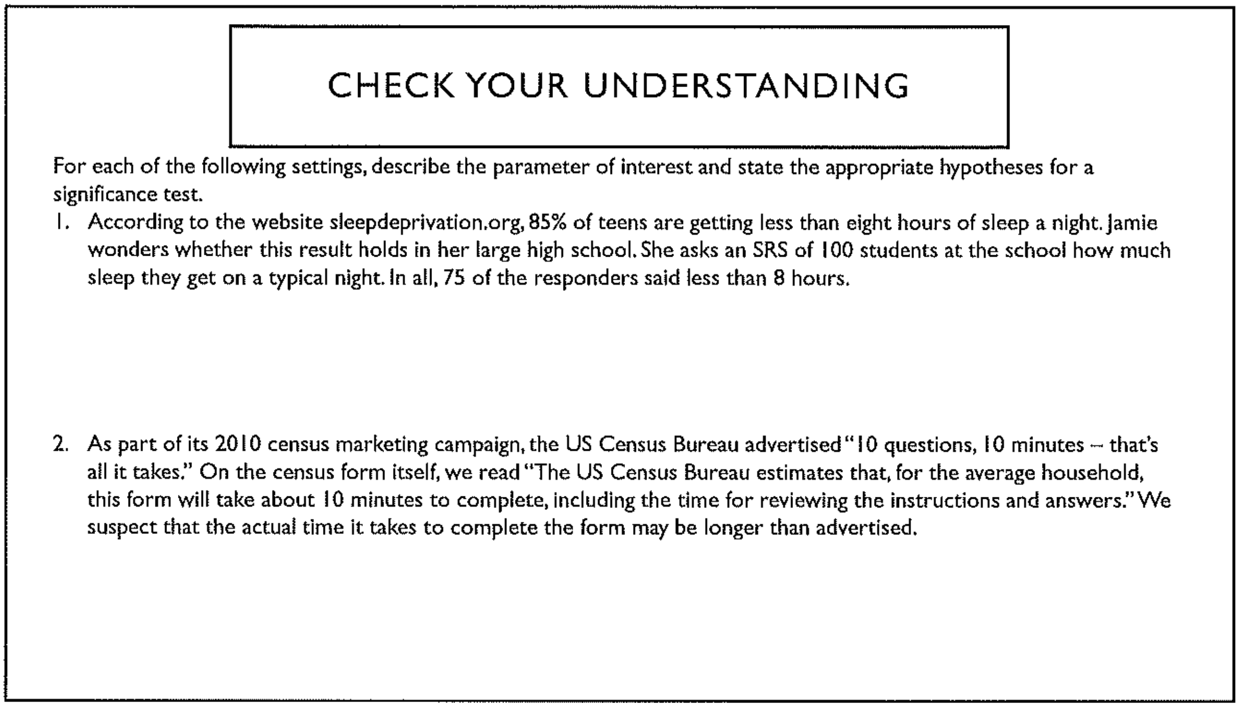
JUICY PINEAPPLES

At the Hawaii Pineapple Company, managers are interested in the size of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A different irrigation system was installed in this field after the growing season. Managers wonder how this change will affect the mean weight of pineapples grown in the field this year.

Problem: State appropriate hypotheses for performing a significance test. Be sure to define the parameter of interest.

Solution: **Note that they are not looking for growth or decrease in the pineapples, only change, so this will be two-sided

The parameter of interest is the mean weight mu of all pineapples grown in the field this year. Therefore,
H_0: mu = 31
H_a: mu =/= 31



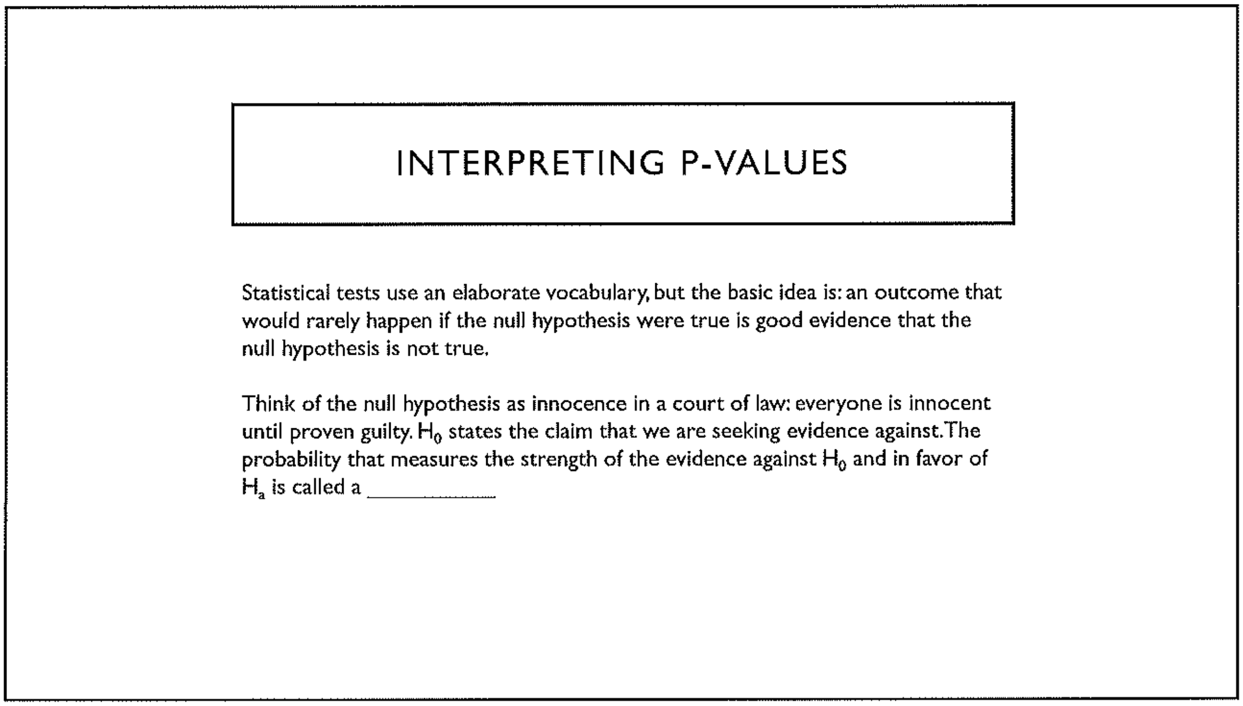
CHECK YOUR UNDERSTANDING

For each of the following settings, describe the parameter of interest and state the appropriate hypotheses for a significance test.

1. According to the website sleepdeprivation.org, 85% of teens are getting less than eight hours of sleep a night. Jamie wonders whether this result holds in her large high school. She asks an SRS of 100 students at the school how much sleep they get on a typical night. In all, 75 of the responders said less than 8 hours.
2. As part of its 2010 census marketing campaign, the US Census Bureau advertised "10 questions, 10 minutes – that's all it takes." On the census form itself, we read "The US Census Bureau estimates that, for the average household, this form will take about 10 minutes to complete, including the time for reviewing the instructions and answers." We suspect that the actual time it takes to complete the form may be longer than advertised.

1. P = proportion of students who get less than 8 hrs of sleep; H0: p=0.85, Ha: p =/= 0.85
2. Mu = true mean time for the average household to complete the survey; H0: mu=10, Ha: mu>10

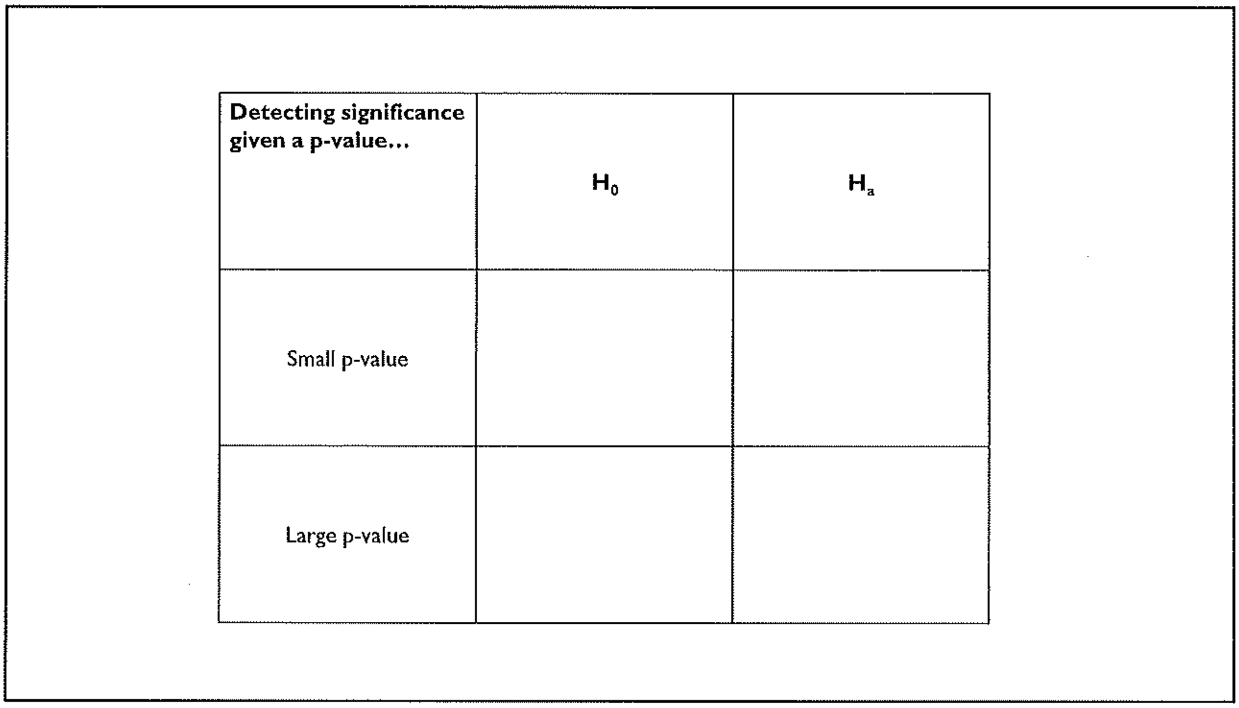
**Activity pg 542



INTERPRETING P-VALUES

Statistical tests use an elaborate vocabulary, but the basic idea is: an outcome that would rarely happen if the null hypothesis were true is good evidence that the null hypothesis is not true.

Think of the null hypothesis as innocence in a court of law: everyone is innocent until proven guilty. H_0 states the claim that we are seeking evidence against.The probability that measures the strength of the evidence against H_0 and in favor of H_a is called a ____________

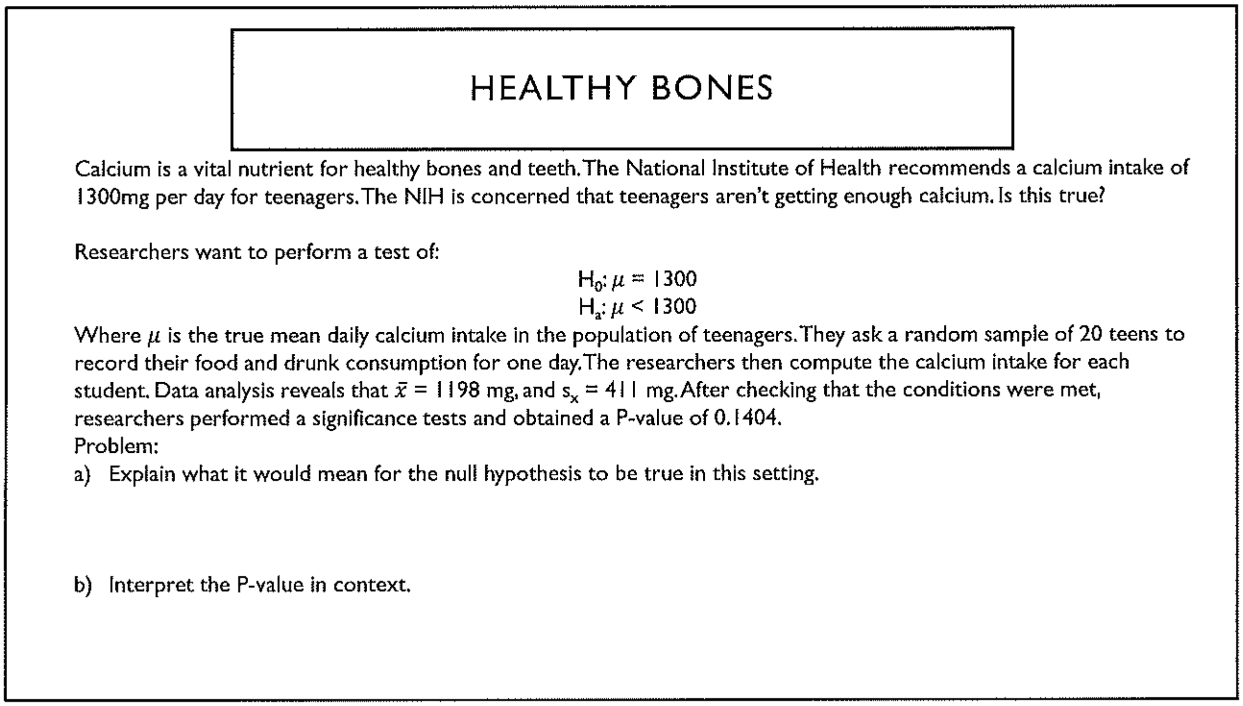


Detecting significance given a p-value…	**H_0**	**H_a**
Small p-value		
Large p-value		

Small p-values: evidence AGAINST H0 because they say that the observed result is unlikely to occur when H0 is true.

Large p-values: fail to give evidence against H0 because they say that the observed result is likely to occur by chance alone when H0 is true.

**remember that P means proportion – we are looking at the probability that these events happen just by chance alone in multiple repetitions of the same sample. If we were to take many many samples of the same size and measuring the same thing, we would have a sampling distribution with some amount of variability, just by chance alone. If the variability is not actually by chance then, there's evidence that the null hypothesis claim in not true.



HEALTHY BONES

Calcium is a vital nutrient for healthy bones and teeth.The National Institute of Health recommends a calcium intake of 1300mg per day for teenagers.The NIH is concerned that teenagers aren't getting enough calcium. Is this true?

Researchers want to perform a test of:

$$H_0: \mu = 1300$$
$$H_a: \mu < 1300$$

Where μ is the true mean daily calcium intake in the population of teenagers.They ask a random sample of 20 teens to record their food and drunk consumption for one day.The researchers then compute the calcium intake for each student. Data analysis reveals that $\bar{x}$ = 1198 mg, and s_x = 411 mg.After checking that the conditions were met, researchers performed a significance tests and obtained a P-value of 0.1404.

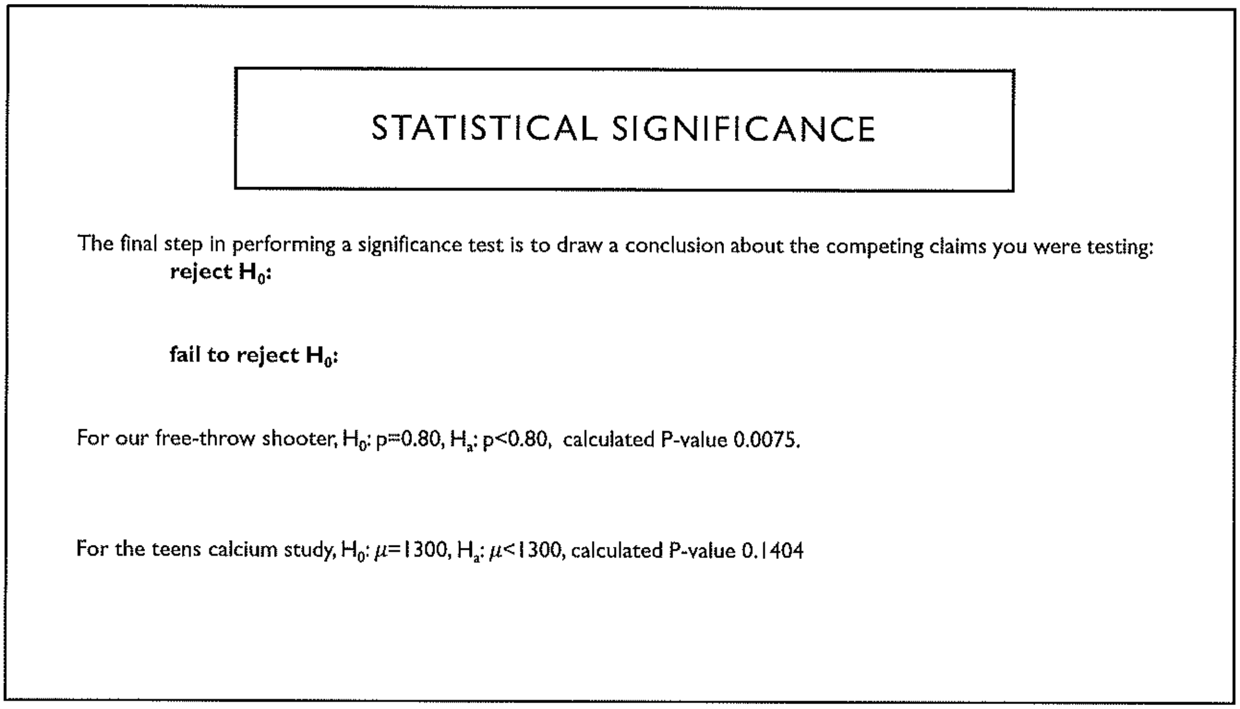
Problem:

a) Explain what it would mean for the null hypothesis to be true in this setting.

b) Interpret the P-value in context.

a) In this setting, H0: mu = 1300 says that the mean daily calcium intake in the population of teenagers is 1300 mg. If H0 is true, then teenagers are getting enough calcium, on average.

b) Assuming that the mean daily calcium intake in the teen population is 1300 mg, there's a 0.1404 probability of getting a sample mean of 1198mg or less jut by chance in a random sample of 20 teens.



STATISTICAL SIGNIFICANCE

The final step in performing a significance test is to draw a conclusion about the competing claims you were testing:

reject H_0:

fail to reject H_0:

For our free-throw shooter, H_0: $p=0.80$, H_a: $p<0.80$, calculated P-value 0.0075.

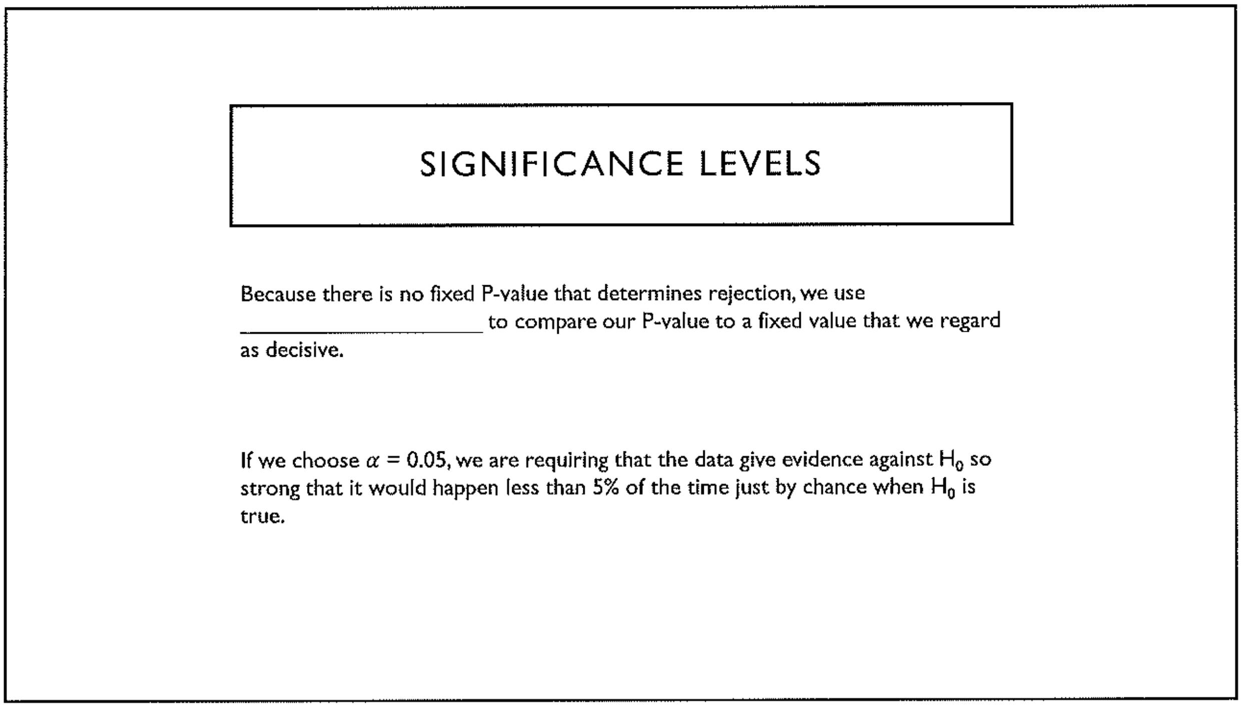
For the teens calcium study, H_0: $\mu=1300$, H_a: $\mu<1300$, calculated P-value 0.1404

Reject: our sample result is too unlikely to have happened by chance assuming H0 is true; convincing evidence for Ha

Fail: our sample result is likely enough to happen by chance assuming H0 is true

Free throw shooter: reject the null hypothesis; we have convincing evidence that the player makes fewer than 80% of his free throws.
Calcium: fail to reject the null hypothesis; we do not have convincing evidence that teens are not getting 1300 mg of calcium daily.

**Whether or not the null hypothesis is rejected is up to the researchers, based on the chance they're willing to take. We normally want something like P-values less than or equal to 0.10. Good practice is .05



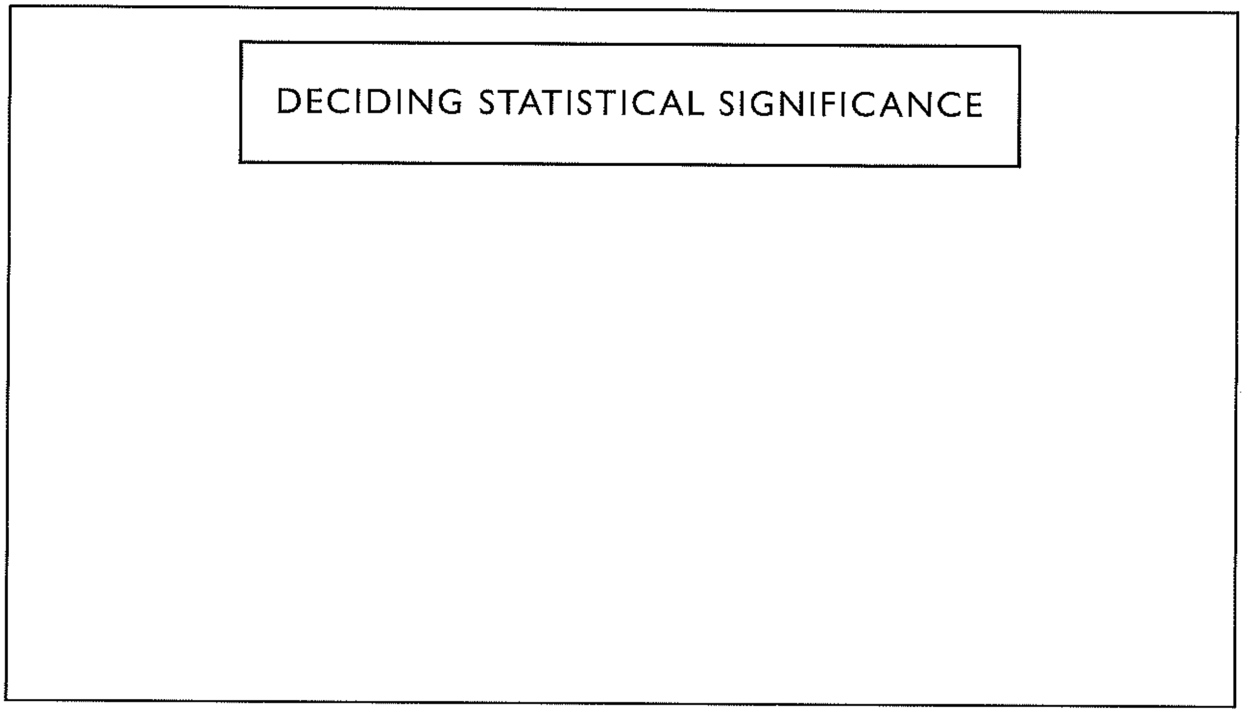
SIGNIFICANCE LEVELS

Because there is no fixed P-value that determines rejection, we use ____________________ to compare our P-value to a fixed value that we regard as decisive.

If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen less than 5% of the time just by chance when H_0 is true.

*Recall from chapter 4 that "statistical significance" means that it would rarely occur by chance alone. When our P-value is less than the chosen alpha, we can say that our result is statistically significant

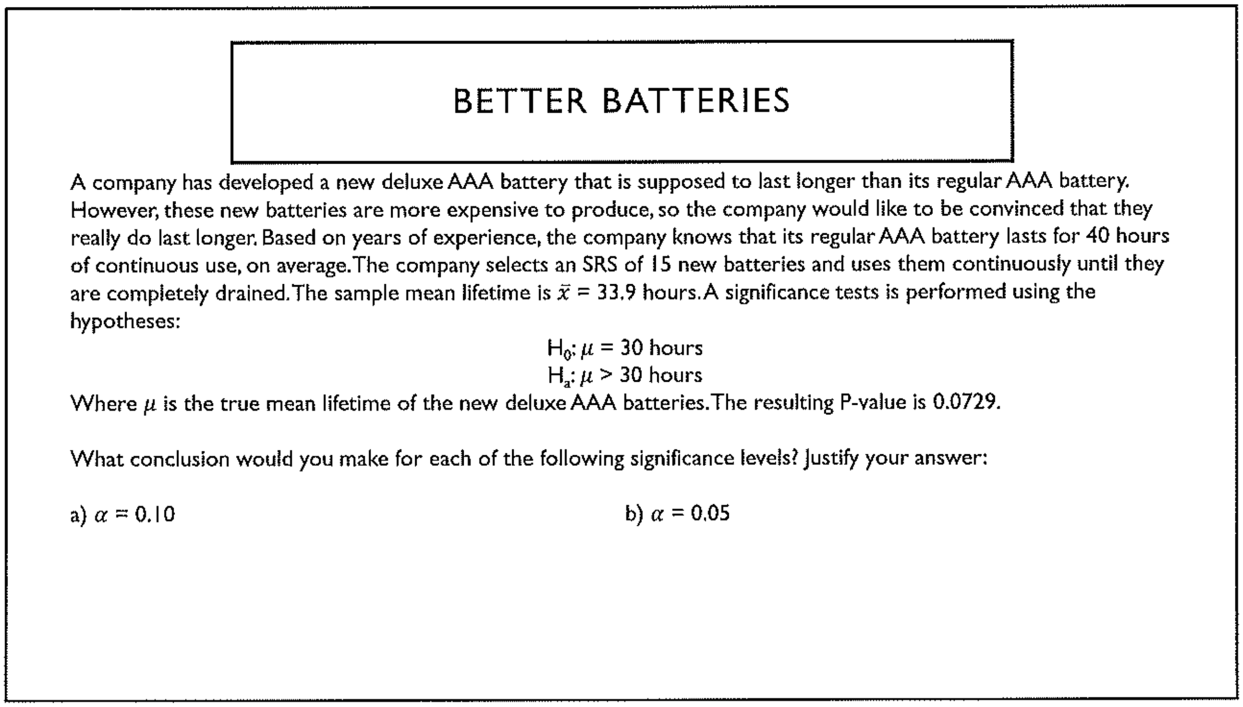
α MUST BE STATED BEFORE DATA ARE PRODUCED



DECIDING STATISTICAL SIGNIFICANCE

Flow chart:

1. Determine parameter of interest and hypotheses
2. Obtain data and conduct statistical tests; calculate p-value
3. Compare p-value to alpha
4. P value is small → smaller than alpha? Yes = reject the null (convincing evidence for Ha), No = fail to reject the null OR pick a larger alpha
5. P value is large → fail to reject the null (not convincing evidence for Ha)



BETTER BATTERIES

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 battery lasts for 40 hours of continuous use, on average. The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained. The sample mean lifetime is $\bar{x}$ = 33.9 hours. A significance tests is performed using the hypotheses:

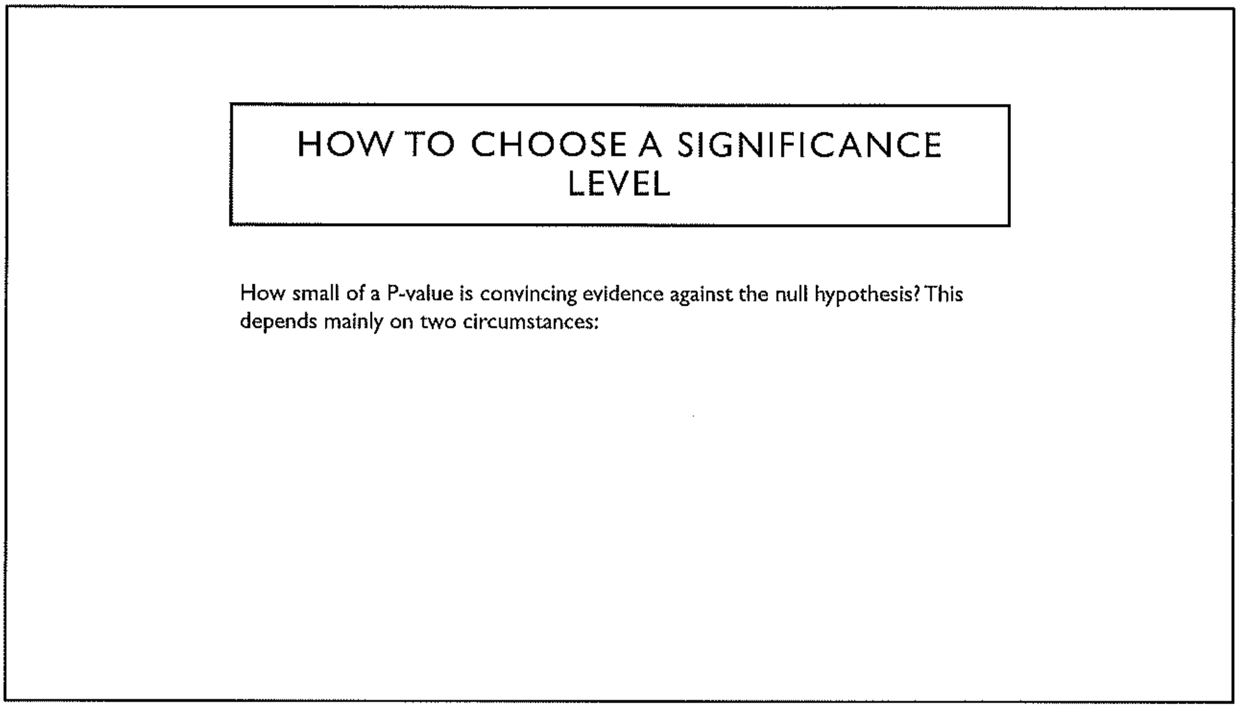
$$H_0: \mu = 30 \text{ hours}$$
$$H_a: \mu > 30 \text{ hours}$$

Where μ is the true mean lifetime of the new deluxe AAA batteries. The resulting P-value is 0.0729.

What conclusion would you make for each of the following significance levels? Justify your answer:

a) $\alpha = 0.10$ b) $\alpha = 0.05$

a) Reject the null, P < α
b) Fail to reject the null, P > α

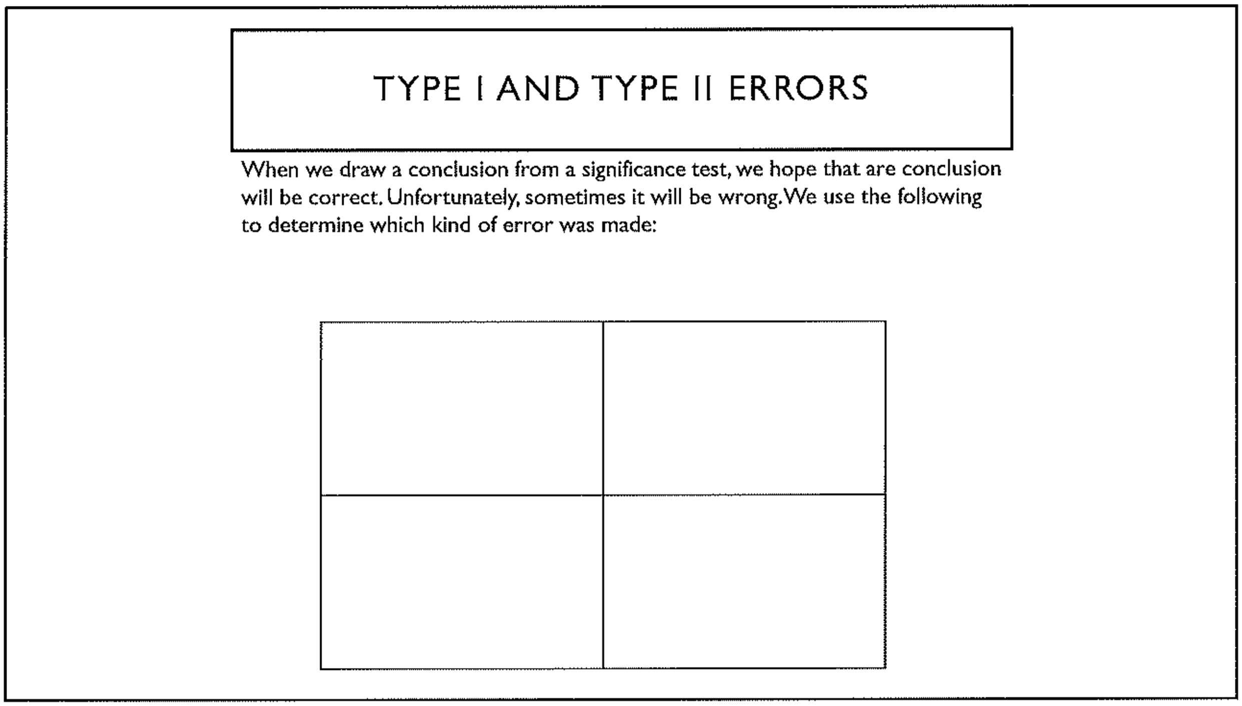


HOW TO CHOOSE A SIGNIFICANCE LEVEL

How small of a P-value is convincing evidence against the null hypothesis? This depends mainly on two circumstances:

1. How plausible is H0? If H0 represents an assumption that the people you must concince have believed for years, strong evidence (a very small P-value) will be needed to persuade them.
2. What are the consequences of rejecting H0? If rejecting H0 in favor of Ha means making an expensive change of some kind, you need strong evidence that the change will be beneficial

Yes it's all subjective. You will get better at recognizing w/ practice.



TYPE I AND TYPE II ERRORS

When we draw a conclusion from a significance test, we hope that are conclusion will be correct. Unfortunately, sometimes it will be wrong. We use the following to determine which kind of error was made:

Reject H0 → H0 is actually true = type 1 error
Fail to reject H0 → Ha is actually true = type 2 error
*see table pg 548

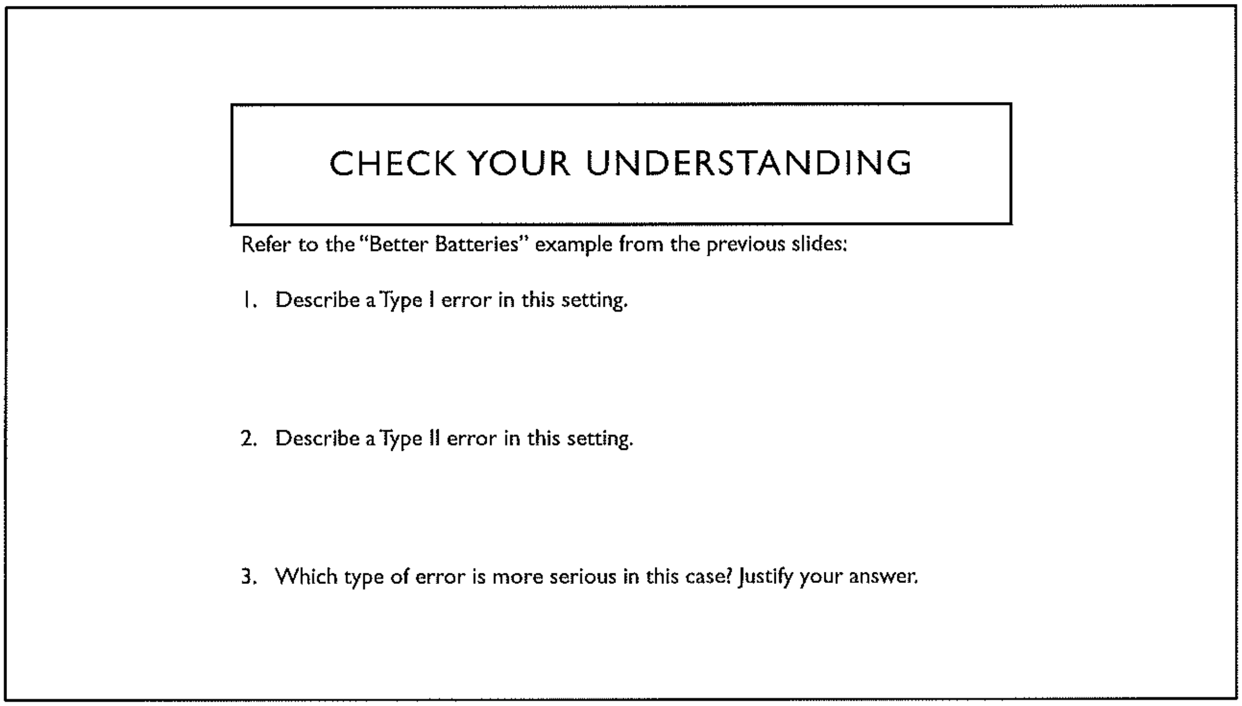
PERFECT POTATOES

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
H_a: p > 0.08

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

Problem: Describe a Type I and a Type II error in this setting, and explain the consequence of each



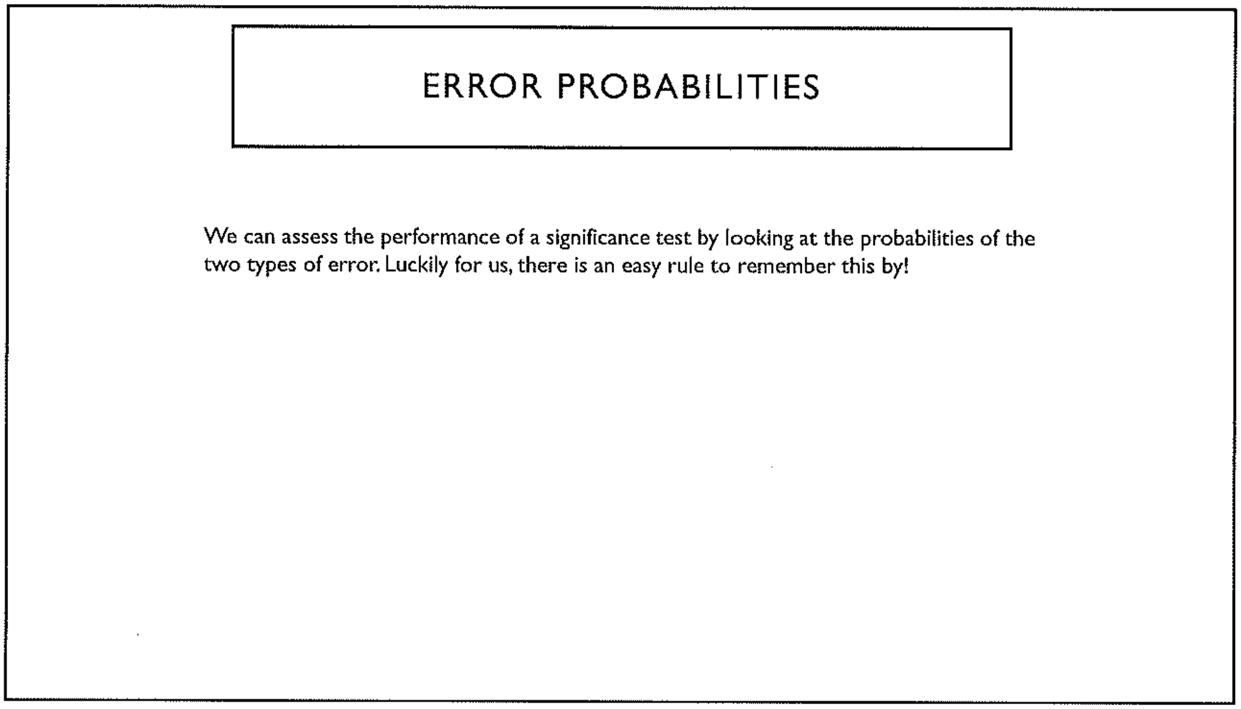
CHECK YOUR UNDERSTANDING

Refer to the "Better Batteries" example from the previous slides:

1. Describe a Type I error in this setting.

2. Describe a Type II error in this setting.

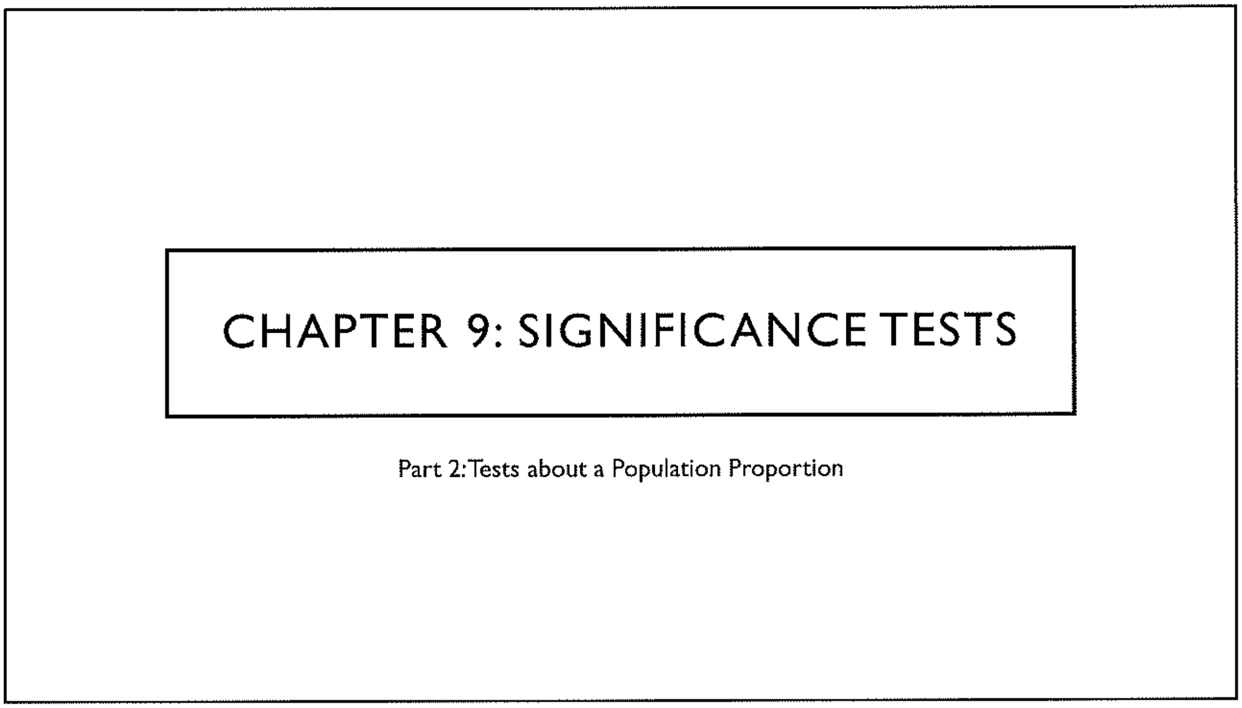
3. Which type of error is more serious in this case? Justify your answer.



ERROR PROBABILITIES

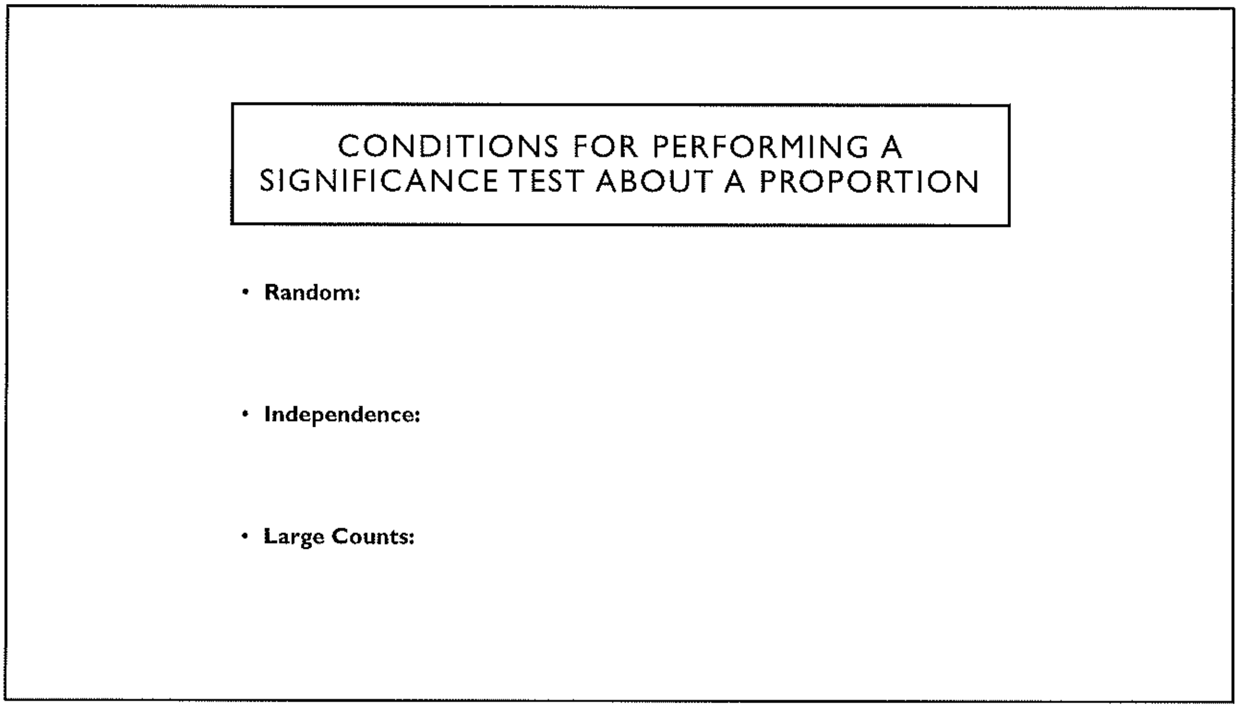
We can assess the performance of a significance test by looking at the probabilities of the two types of error. Luckily for us, there is an easy rule to remember this by!

The significance level alpha of any fixed-level test is the probability of a Type I error. That is, alpha is the probability that the test will reject the null hypothesis H0 when H0 is actually true. Consider the consequences of a Type I error before choosing a significance level.



CHAPTER 9: SIGNIFICANCE TESTS

Part 2: Tests about a Population Proportion



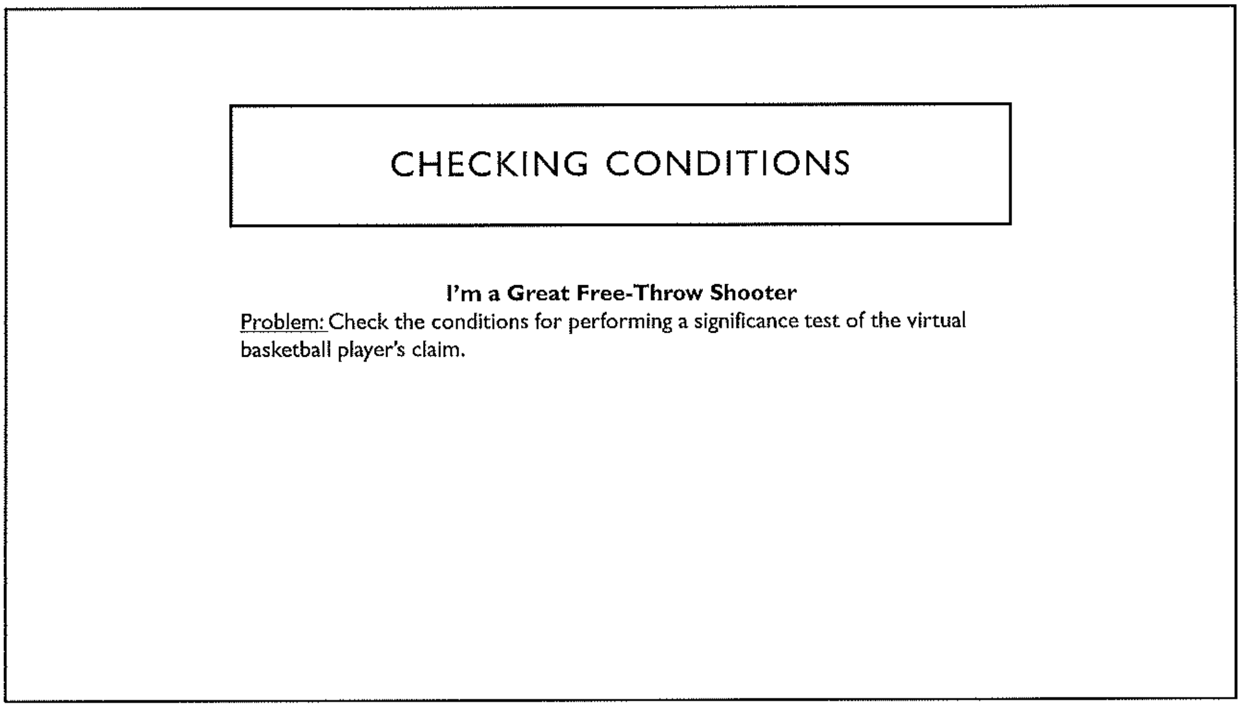
CONDITIONS FOR PERFORMING A SIGNIFICANCE TEST ABOUT A PROPORTION

- **Random:**
- **Independence:**
- **Large Counts:**

Must be an SRS

10% condition

Large Counts: assuming H_0 is true, test np and n(1-p)

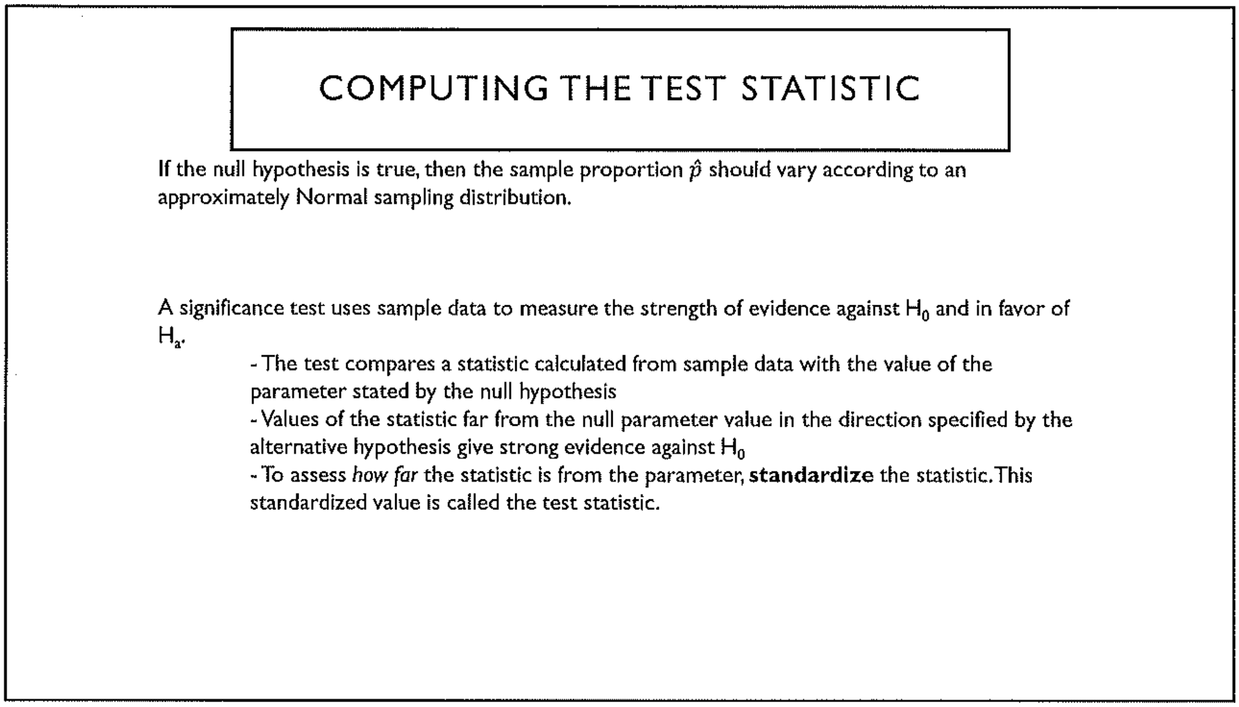


CHECKING CONDITIONS

I'm a Great Free-Throw Shooter

Problem: Check the conditions for performing a significance test of the virtual basketball player's claim.

The required conditions are:
Random – We can view the set of 50 computer generated shots as a simple random sample from the population of all possible shots that the cirtual shooter takes.
10% We're not sampling without repalcement from a finite population, so we don't need to check the 10% condition
Large Counts– Assuming H_0 is true p = 0.80. Then np = 0.80 = 50(0.80) = 40, and n(1-p) = 50(.2) = 10 are both at least 10, so this condition is met.



COMPUTING THE TEST STATISTIC

If the null hypothesis is true, then the sample proportion $\hat{p}$ should vary according to an approximately Normal sampling distribution.

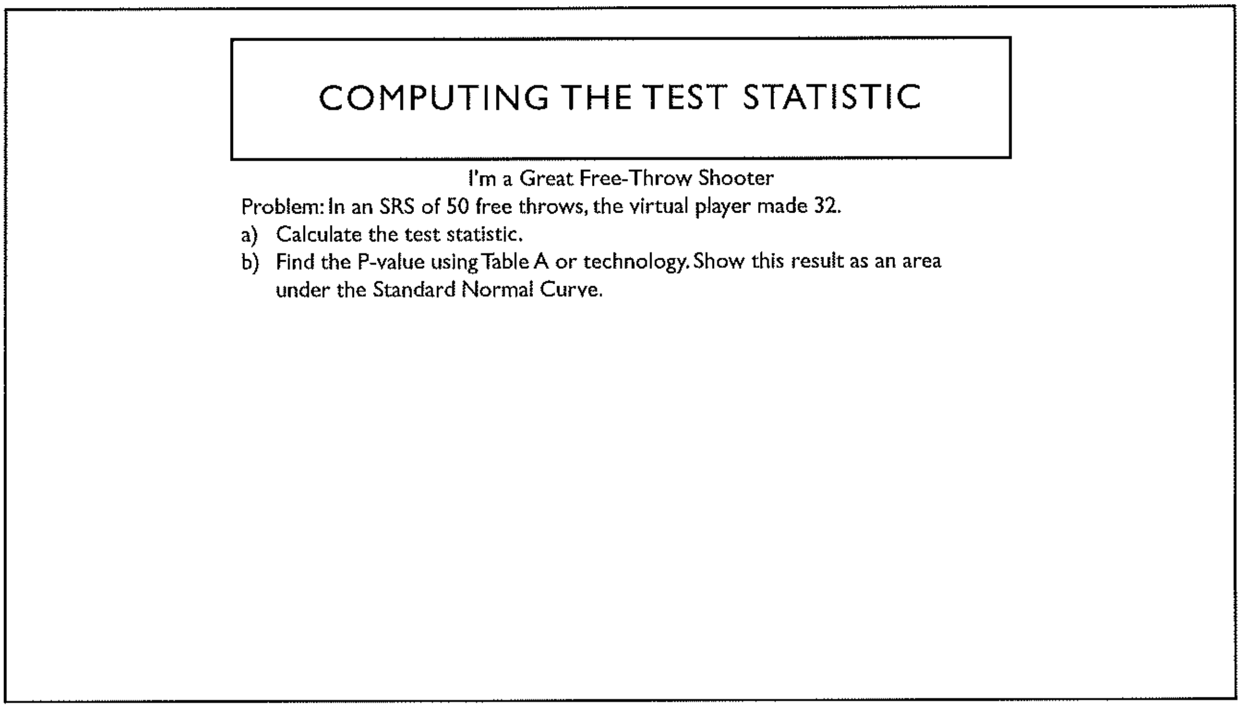
A significance test uses sample data to measure the strength of evidence against H_0 and in favor of H_a.

- The test compares a statistic calculated from sample data with the value of the parameter stated by the null hypothesis
- Values of the statistic far from the null parameter value in the direction specified by the alternative hypothesis give strong evidence against H_0
- To assess *how far* the statistic is from the parameter, **standardize** the statistic. This standardized value is called the test statistic.

Sampling distribution of a hypothesis is N(mu_p, $sigma_p$) where mu_p = p and $sigma_p$ = sqrt(p(1-p)/n)

*Recall standardize = z score!

Test statistic = stat – parameter / standard deviation of stat



COMPUTING THE TEST STATISTIC

I'm a Great Free-Throw Shooter

Problem: In an SRS of 50 free throws, the virtual player made 32.

a) Calculate the test statistic.
b) Find the P-value using Table A or technology. Show this result as an area under the Standard Normal Curve.

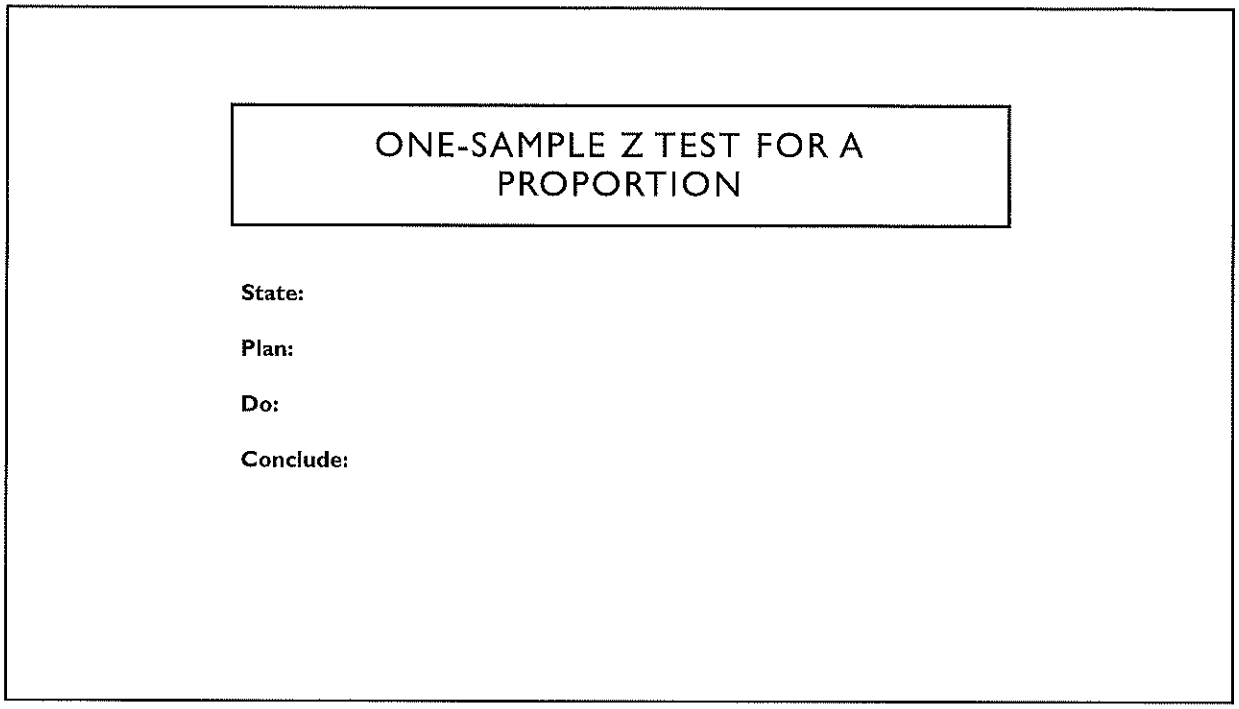
Solution:

A) Phat = 32/50 = .64. Standardizing, we get:

Test stat = stat – parameter / sd of stat → z = 0.64 – 0.80/ sqrt(p(1-p)/n) = 0.64 – 0.80/.0566 = -2.83.

b) The shaded area under the curve shows the P-value. From table A, P(z≤-2.83) = .0023. **DRAW CURVE**

If H_0 is true, and the player makes 80% of his free throws in the long run, there's only about a .0023 probability that he would make 32 or fewer of 50 shots by chance alone. This gives us evidence to reject the null hypothesis and believe that the player is exaggerating.



ONE-SAMPLE Z TEST FOR A PROPORTION

State:

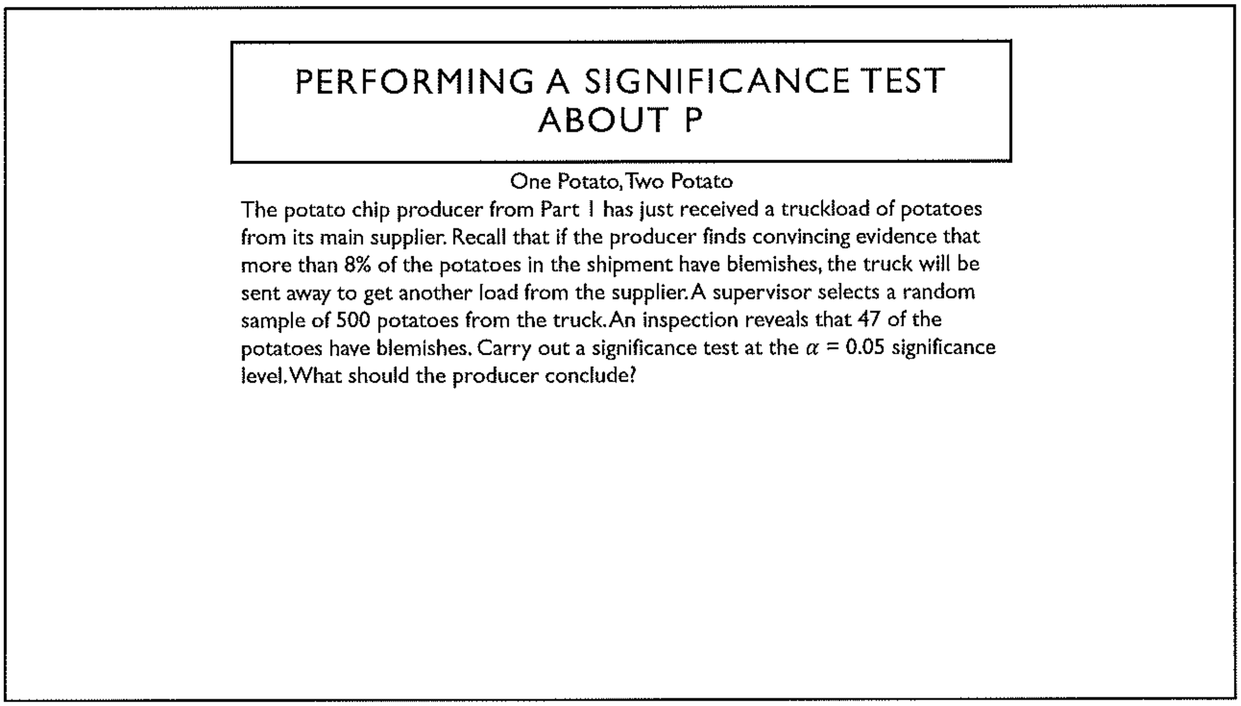
Plan:

Do:

Conclude:

State: what hypotheses do you want to test, and at what significance level? Define any parameters you use.
Plan: Choose the appropriate inference method. Check conditions.
    proportion → z score
    mean → t score
Do: If conditions are met, perform calculations:
    compute the test statistic
    find the p-value
Conclude: Make a decision about the hypotheses in the context of the problem

**When conditions are met, the sampling distribution of phat is approximately normal with mean mu$_p$ = p and sigma$_p$ = sqrt(p(1-p)/n)
** For a *proportion*: z = phat – p$_0$ / sqrt(p$_0$(1-p$_0$)/n)
** Check for the direction of H$_a$ – if greater than, then 1-stat; less than = as is, two tailed = double areas



PERFORMING A SIGNIFICANCE TEST ABOUT P

One Potato, Two Potato

The potato chip producer from Part 1 has just received a truckload of potatoes from its main supplier. Recall that if the producer finds convincing evidence that more than 8% of the potatoes in the shipment have blemishes, the truck will be sent away to get another load from the supplier. A supervisor selects a random sample of 500 potatoes from the truck. An inspection reveals that 47 of the potatoes have blemishes. Carry out a significance test at the α = 0.05 significance level. What should the producer conclude?

State: We want to perform a test of
H_0: p = 0.08
H_a: p > 0.08
Where p is the actual proportion of potatoes in this shipment with blemishes. We'll use an α = .05 significance level.

Plan: If conditions are met, we should do a one-sample z test for the population proportion p.

- random: the supervisor took a random sample of 500 potatoes from the shipment
- 10% condition: it seems reasonable to assume that there are at least 10(500) = 5000 potatoes in the shipment.
- Large counts: Assuming H_0: p = 0.08 is true, the expected counts of blemished and unblemished potatoes are np_0 = 500(.08) = 40, and $n(1-p_0)$ = 500(.92) = 460 respectively. Because both values are at least 10, we should be safe using normal calculations.

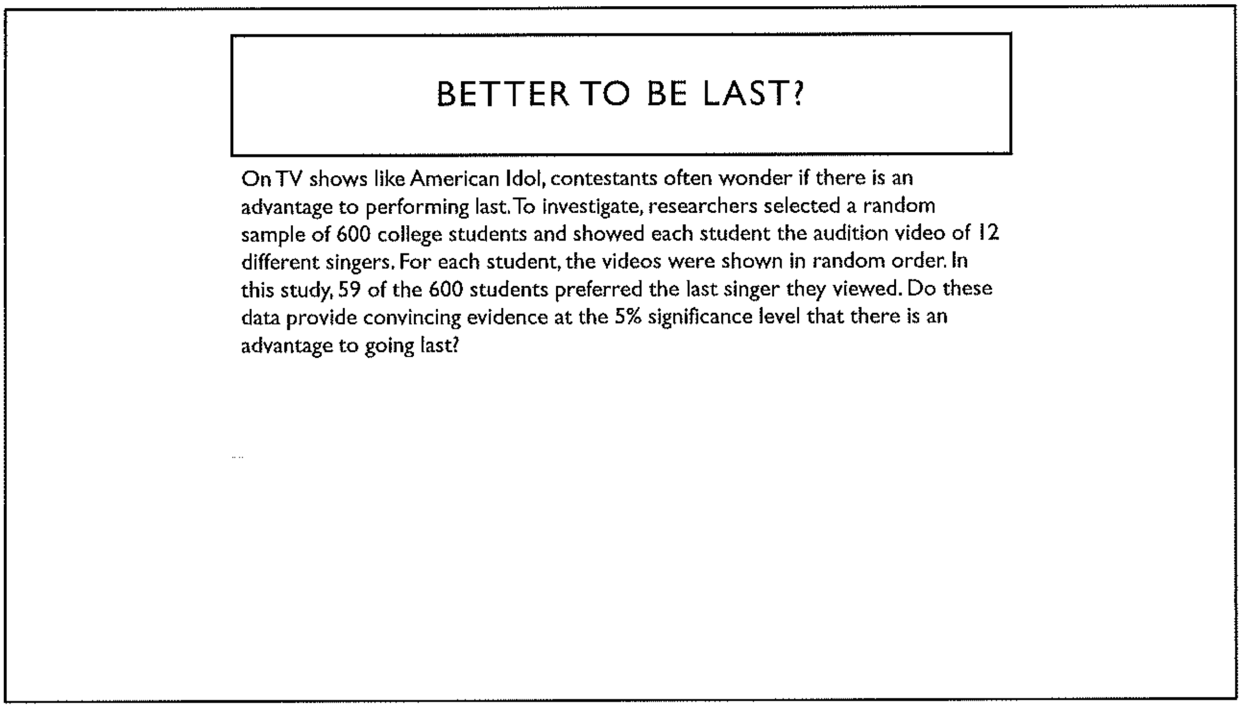
Do: The sample proportion of blemished potatoes is phat = 47/500 = .094.
Test stat: z = .094 - .08/ sqrt(.08(.92)/500) = 1.15

P value: Table A gives the P value as $P(z \geq 1.15) = 1 - .8749 = .1251$.

Using technology: normalcdf(lower: 1.15, upper: 10000, mu: 0, sigma: 1) also gives p value .1251.

Conlcude: Because our p-value, .01251, is greater than $\alpha = .05$, we fail to reject H_0. There is not concinving evidence that the shipment contains more than 8% blemished potatoes. As a result, the producer will use this truckload of potatoes to make potato chips.

** YOU MAY NOT SAY THAT THE TRUE PROPORTION IS .08!!! WE DON'T KNOW WHAT IT TRULY IS, WE JUST DON'T HAVE EVIDENCE THAT IT'S ANYTHING DIFFERENT.



BETTER TO BE LAST?

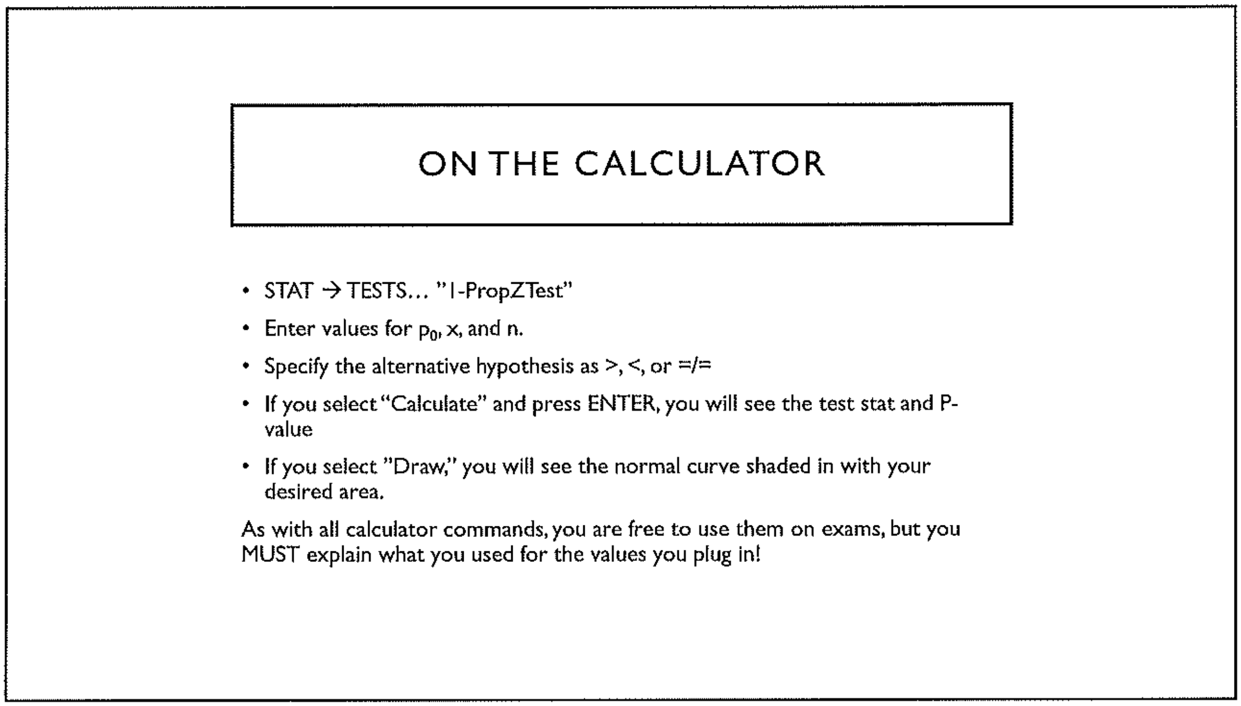
On TV shows like American Idol, contestants often wonder if there is an advantage to performing last. To investigate, researchers selected a random sample of 600 college students and showed each student the audition video of 12 different singers. For each student, the videos were shown in random order. In this study, 59 of the 600 students preferred the last singer they viewed. Do these data provide convincing evidence at the 5% significance level that there is an advantage to going last?

State: We want to test the following hypotheses at the α =.05 level:
H_0: p = 1/12 (.083)
H_a: p>1/12
Where p = the true proportion of students who prefer the last singer they see.

Plan: If conditions are met, we will perform a one-sample z test for p:
Random: a random sample of students was selected, and the order in which the videos were viewed was randomized for each subject
10%: It is reasonable to assume that there are more than 10(600) = 6000 students.
Large Counts: np_0 = 600(1/12) = 50 $\geq$ 10, and $n(1-p_0)$ = 600(11/12) = 550 $\geq$ 10. We are safe conducting normal tests.

Do: The sample proprotion of fans who preferred the last contestant is phat = 59/600=.098.
Test stat: z = .098 - .083/ sqrt(.083*.917)/600) = 1.33
P-Value: P(z > 1.33) = 1-9082 = .0918.

Conclude: Because the P-value .0918 is greater than α =.05, we fail to reject the null hypothesis. There is not convincing evidence that the true proportion of students who prefer the final singer is greater than p=1/12. That is, there is not convincing evidence that there is an advantage to going last.

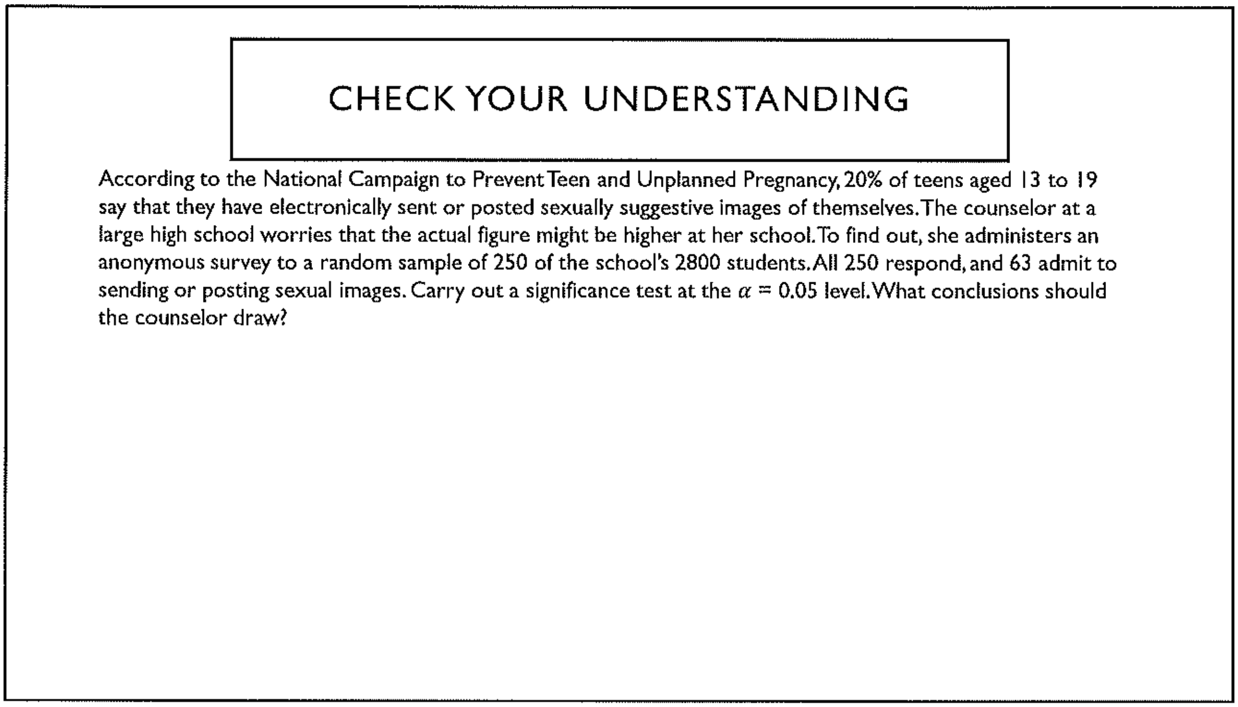


ON THE CALCULATOR

- STAT → TESTS… "1-PropZTest"
- Enter values for p_0, x, and n.
- Specify the alternative hypothesis as >, <, or =/=
- If you select "Calculate" and press ENTER, you will see the test stat and P-value
- If you select "Draw," you will see the normal curve shaded in with your desired area.

As with all calculator commands, you are free to use them on exams, but you MUST explain what you used for the values you plug in!

Do with the potato example



CHECK YOUR UNDERSTANDING

According to the National Campaign to Prevent Teen and Unplanned Pregnancy, 20% of teens aged 13 to 19 say that they have electronically sent or posted sexually suggestive images of themselves. The counselor at a large high school worries that the actual figure might be higher at her school. To find out, she administers an anonymous survey to a random sample of 250 of the school's 2800 students. All 250 respond, and 63 admit to sending or posting sexual images. Carry out a significance test at the $\alpha = 0.05$ level. What conclusions should the counselor draw?

State: H_0: p = .20

H_a: p>.20

p = true proportion of students who send sexually explicit images, at the .05 level

Plan: 1-prop z test if conditions are met:

Random: yes

10%: 250(10)=2500

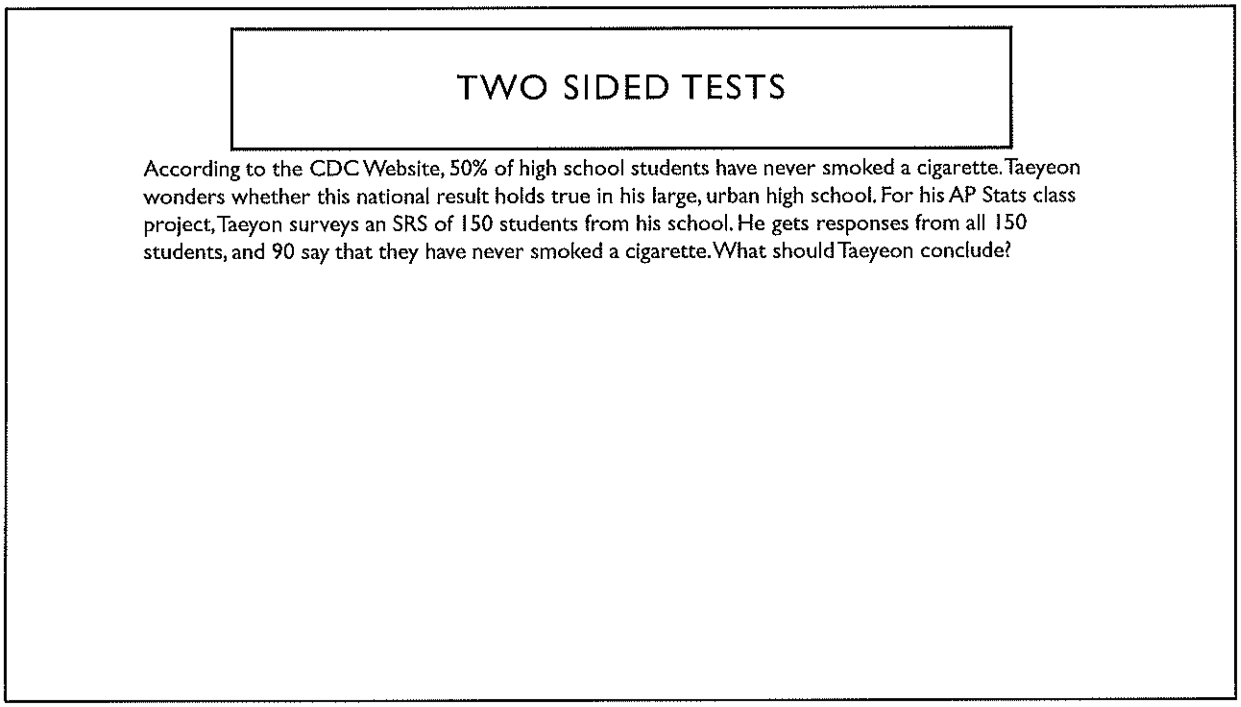
Large Counts: np = 50, n(1-p) = 200

Do: phat = .252

Z = .252-.20/sqrt(.20(.8)/250) = 2.06

P(Z$\geq$2.06) = 1 - .9803 = .0197

Conclude: Because P value .0197 < α = .05, we reject H_0. We have convincing evidence that more than 20% of teens in the school would say that have sent or posted sexually suggestive images of themselves.



TWO SIDED TESTS

According to the CDC Website, 50% of high school students have never smoked a cigarette. Taeyeon wonders whether this national result holds true in his large, urban high school. For his AP Stats class project, Taeyeon surveys an SRS of 150 students from his school. He gets responses from all 150 students, and 90 say that they have never smoked a cigarette. What should Taeyeon conclude?

State: We want to perform a significance test using the hypotheses
H_0: p = .50
H_a: p=/= .50
Where p = the proportion of all students in Taeyeon's school who would say that they have never smoked. Because no significance level was stated, we will use α =.05

Plan: If conditions are met, we'll do a one-sample z test for the population proportion p.

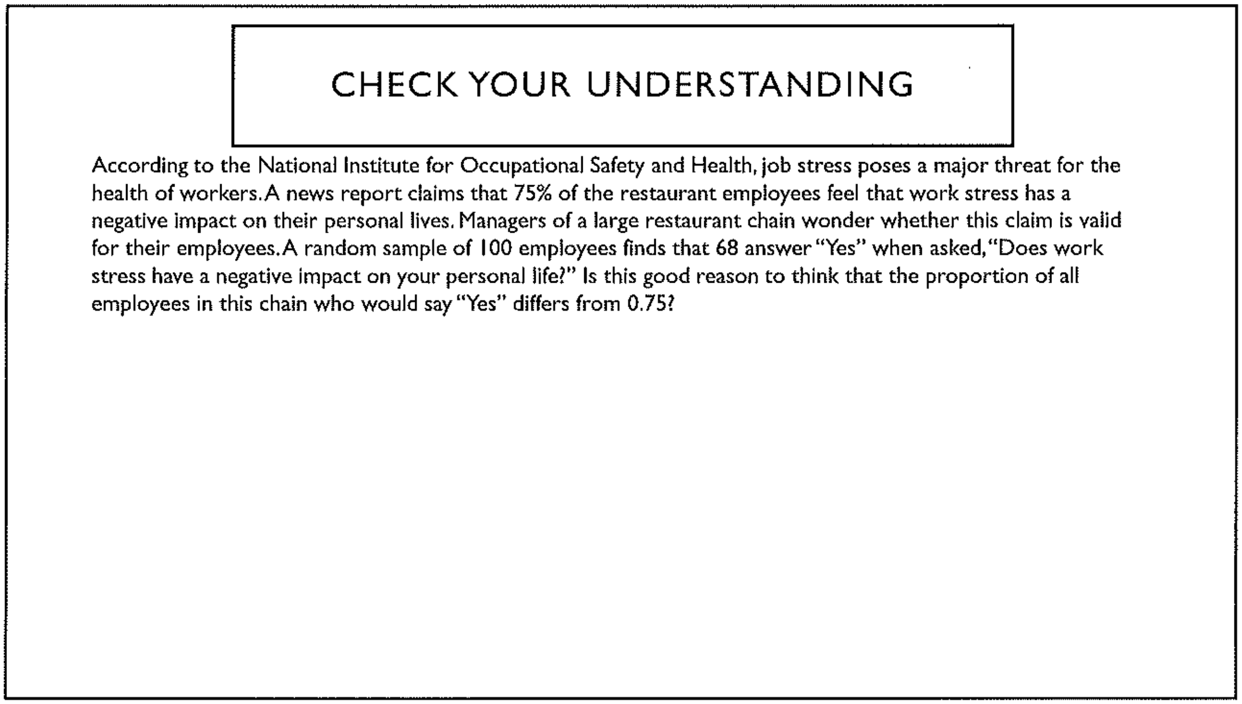
Random: Taeyeon surveyed an SRS of 150 students from his school

10%: It seems reasonable to assume that there are at least 10(150) = 1500 students in a large high school

Large Counts: Assuming H_0: p=.50 is true, np_0 = 150(.50) = 75 and $n(1-p_0)$ = 150(.50) = 75. Both values are greater than 10.

Do: phat = 90/150 = .60
Test stat: z = .60-.50/sqrt(.5•.5/150) = 2.45.
P(z >2.45) = .0071. This is just the right tailed area, so 2(.0071) = .0142 gives the two-tailed probability.

Conclude: Because P = .0143 < α =.05, we reject H_0. We have convincing evidence that the proportion of all students at Taeyeon's school who would say that they have never smoked differs from the national result of .50.



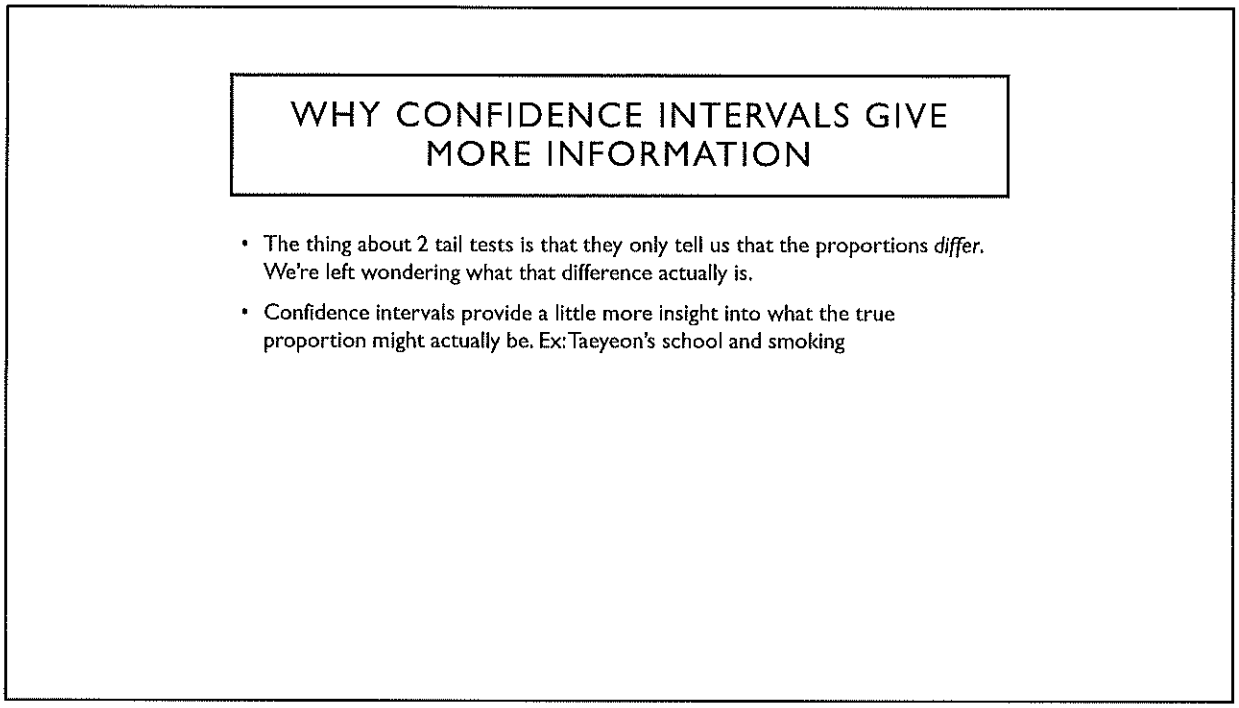
CHECK YOUR UNDERSTANDING

According to the National Institute for Occupational Safety and Health, job stress poses a major threat for the health of workers. A news report claims that 75% of the restaurant employees feel that work stress has a negative impact on their personal lives. Managers of a large restaurant chain wonder whether this claim is valid for their employees. A random sample of 100 employees finds that 68 answer "Yes" when asked, "Does work stress have a negative impact on your personal life?" Is this good reason to think that the proportion of all employees in this chain who would say "Yes" differs from 0.75?

H_0: p = .75, H_a: p =/= .75, use alpha = .05

Conditions all met, use 1prop z test

Phat = .68
Z = .68-.75/sqrt(.75•.25/100) = -1.62
2P(Z $\leq$-1.62) = .1052.
Because P>α we fail to reject H_0



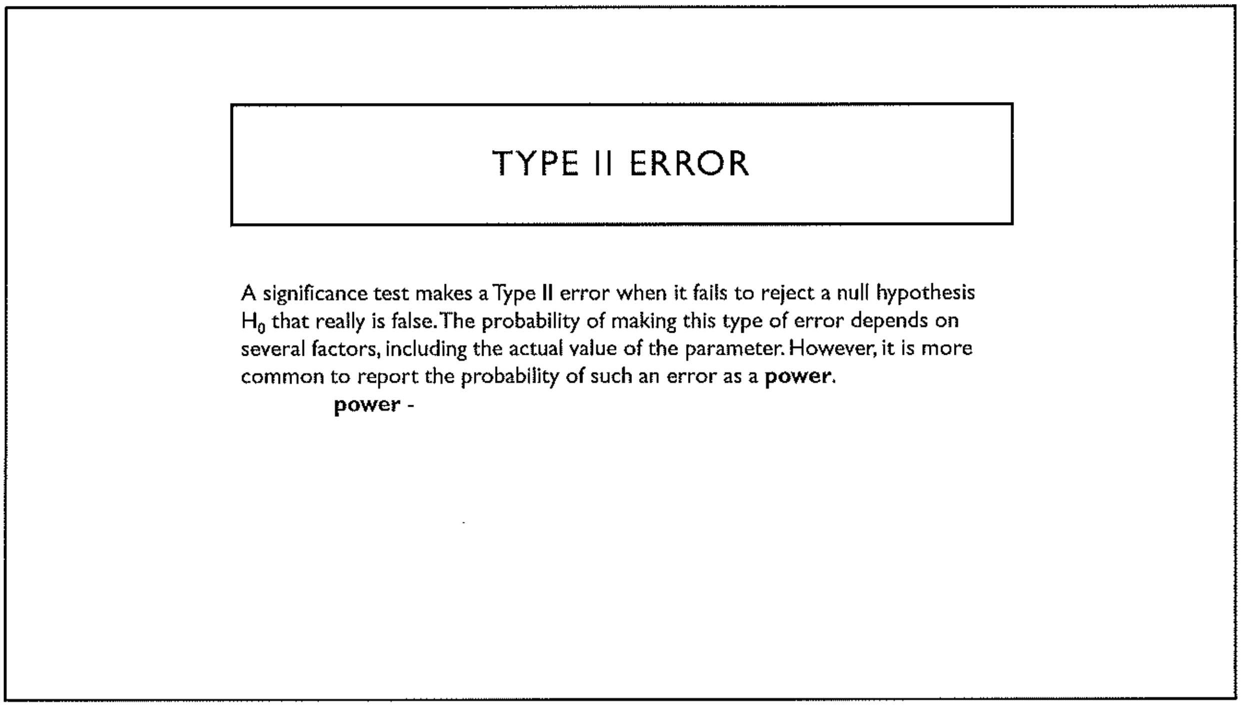
WHY CONFIDENCE INTERVALS GIVE MORE INFORMATION

- The thing about 2 tail tests is that they only tell us that the proportions *differ.* We're left wondering what that difference actually is.
- Confidence intervals provide a little more insight into what the true proportion might actually be. Ex:Taeyeon's school and smoking

Confidence interval for smoking example: use 95% confidence (like using α =.05)

Phat ± z* sqrt(phat(1-phat)/n) = .60 ± 1.96•sqrt((.6 •.4)/150) = .60 ± .078 = (0.522, 0.678).
We are 95% confident that the interval from 0.552 to 0.678 captures the true proportion of students at the school who would say that they have never smoked cigarettes. It's actually higher!

*****COOL THING: For a two sided test, you can use a confidence interval instead! For a one-sided test, you still have to do a significance test. Still write out hypotheses and check conditions, but the do/conclude part will look a little different



TYPE II ERROR

A significance test makes a Type II error when it fails to reject a null hypothesis H_0 that really is false. The probability of making this type of error depends on several factors, including the actual value of the parameter. However, it is more common to report the probability of such an error as a **power.**

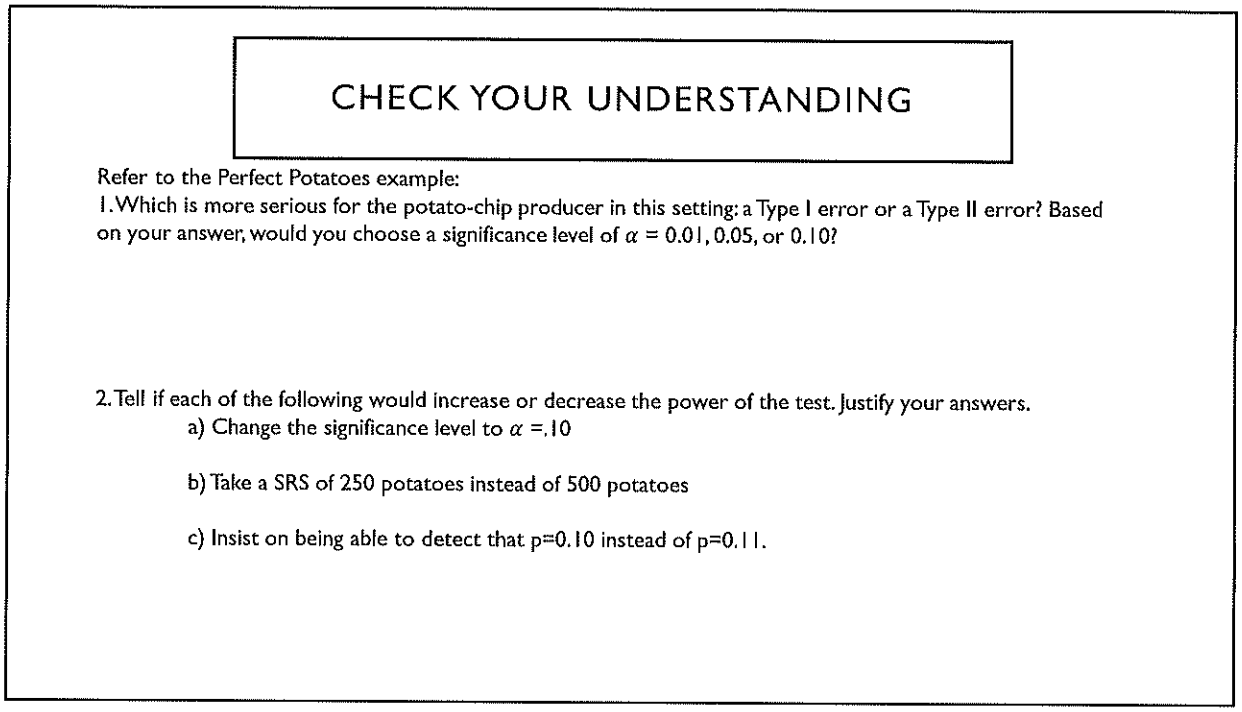
power -

Power – the power of a test against a specific alternative is the probability that the test will reject H_0 at a chosen α when the specified alternative value of the parameter is true.

*power = probability of avoiding a type II error

Best way to avoid type II error:

- increase the sample size
- increase the alpha
- increase the difference between the null and alternative parameter values (easier to detect large differences)



CHECK YOUR UNDERSTANDING

Refer to the Perfect Potatoes example:

1. Which is more serious for the potato-chip producer in this setting: a Type I error or a Type II error? Based on your answer, would you choose a significance level of α = 0.01, 0.05, or 0.10?

2. Tell if each of the following would increase or decrease the power of the test. Justify your answers.

   a) Change the significance level to α =.10

   b) Take a SRS of 250 potatoes instead of 500 potatoes

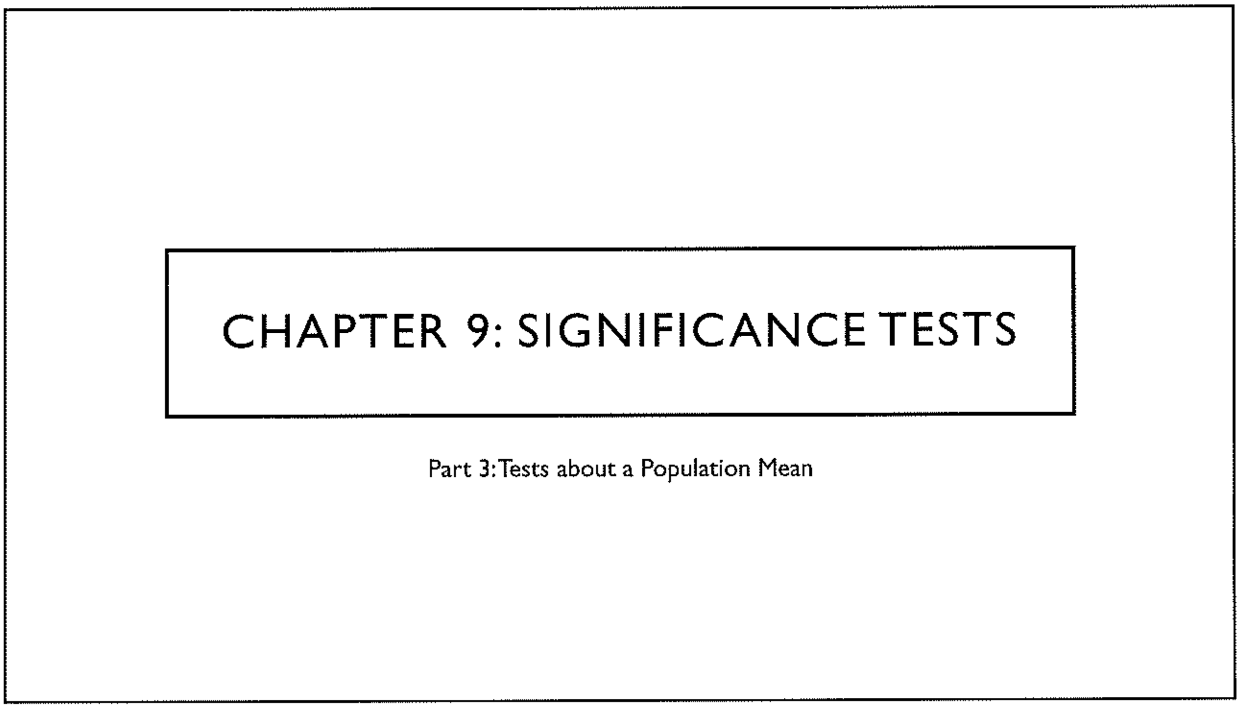
   c) Insist on being able to detect that p=0.10 instead of p=0.11.

1. Type II. To minimize Type II error, choose a large significance level such as α =.10

2. a) increase

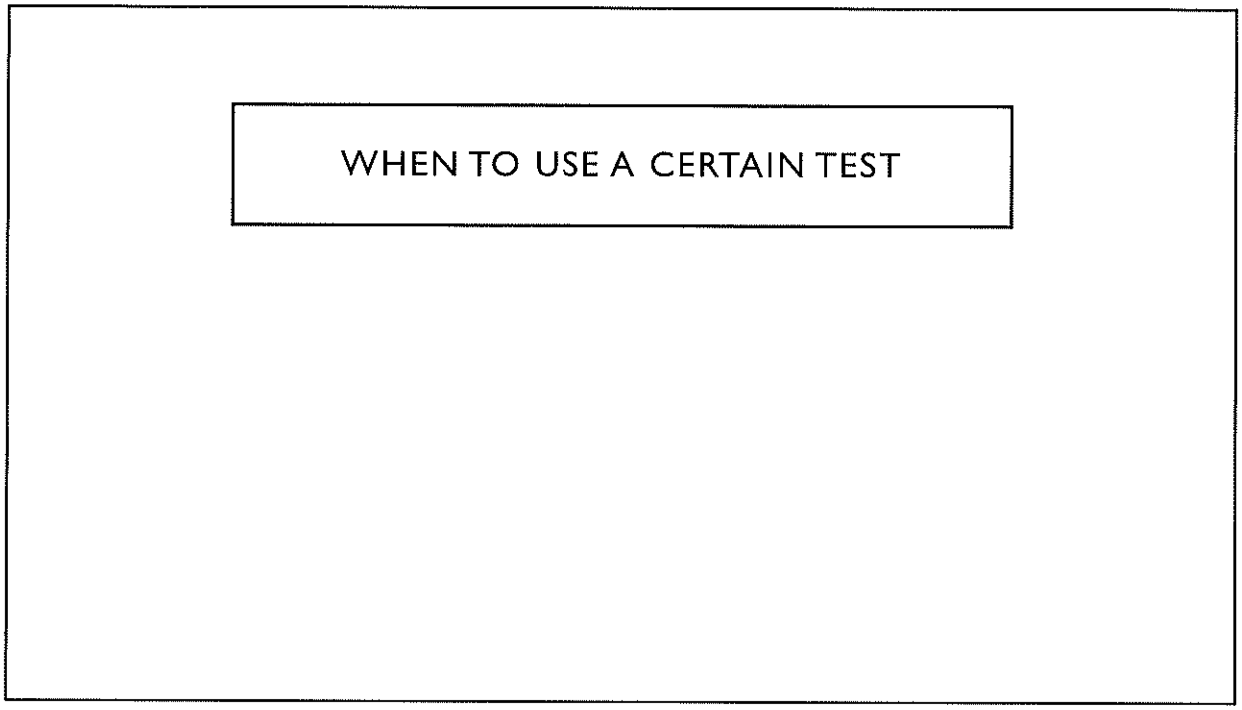
   b) decrease

   c) decrease



CHAPTER 9: SIGNIFICANCE TESTS

Part 3: Tests about a Population Mean



WHEN TO USE A CERTAIN TEST

Parameter = proportion

- Confidence interval
    - “true proportion”
    - CI = phat ± z*sqrt((p)(1-p)/n)
- Z test
    - Hypothesis testing for a sample proportion
        - Z = phat – p_0/sqrt(p(1-p)/n)
    - Proportion/percentile of given value
        - Z = x – mu/sigma

        **difference is knowing sigma and mu

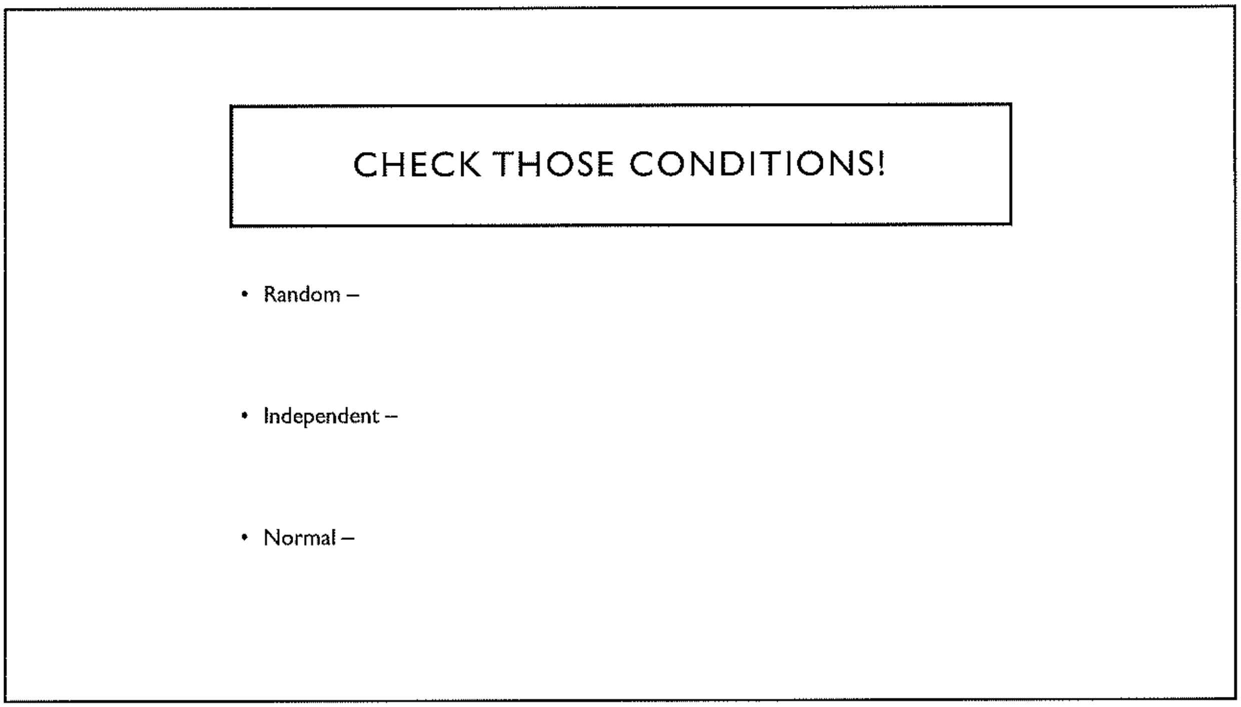
Parameter = mean

Confidence intervals w/ t

CI = xbar ± t*s_x/sqrt(n)

T test → hypothesis testing for a sample mean

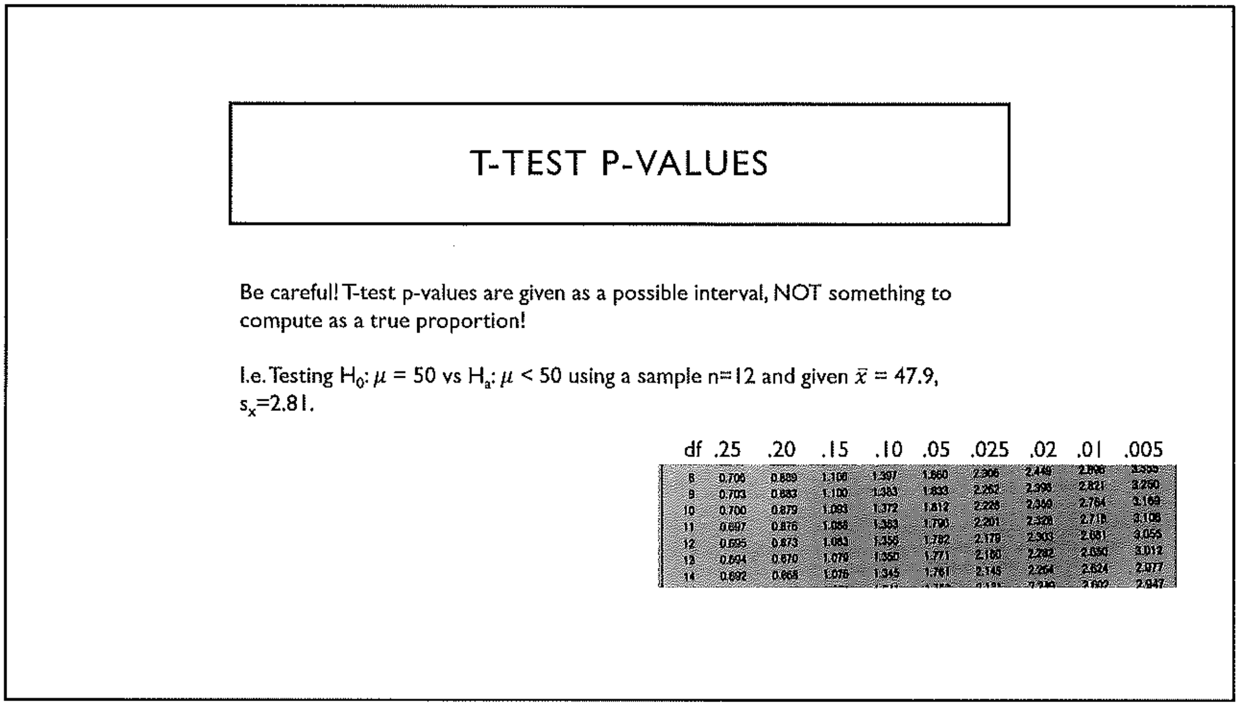
t = xbar – mu_0/ s_x/sqrt(n) for df = n-1



CHECK THOSE CONDITIONS!

- Random –
- Independent –
- Normal –

*random – data come from SRS or well designed experiment
*10% n≤1/10(N)
*Normal/Large Sample – sample size is large (n≥30) OR mentioned shaped normal; if neither, gotta draw a histogram and check it (don't use t procedure if the graph shows strong skewness or outliers)



T-TEST P-VALUES

Be careful! T-test p-values are given as a possible interval, NOT something to compute as a true proportion!

I.e. Testing H_0: $\mu = 50$ vs H_a: $\mu < 50$ using a sample n=12 and given $\bar{x} = 47.9$, s_x=2.81.

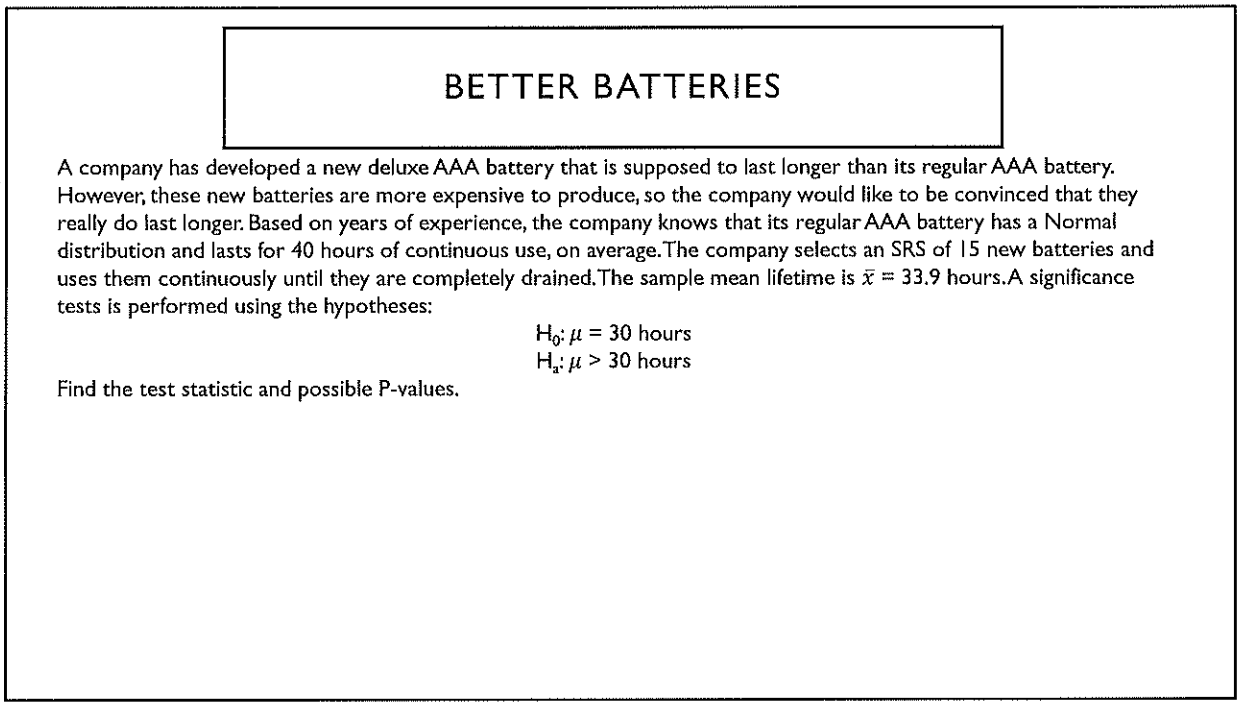
df	.25	.20	.15	.10	.05	.025	.02	.01	.005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.602	2.947

t = 47.9 – 50/2.81/sqrt(12) = -2.59

On table with df = 12 – 1 = 11

Look for what 2.59 is between → it's between p = .01 and p=.02

So we say $.01 \leq P \leq .02$



BETTER BATTERIES

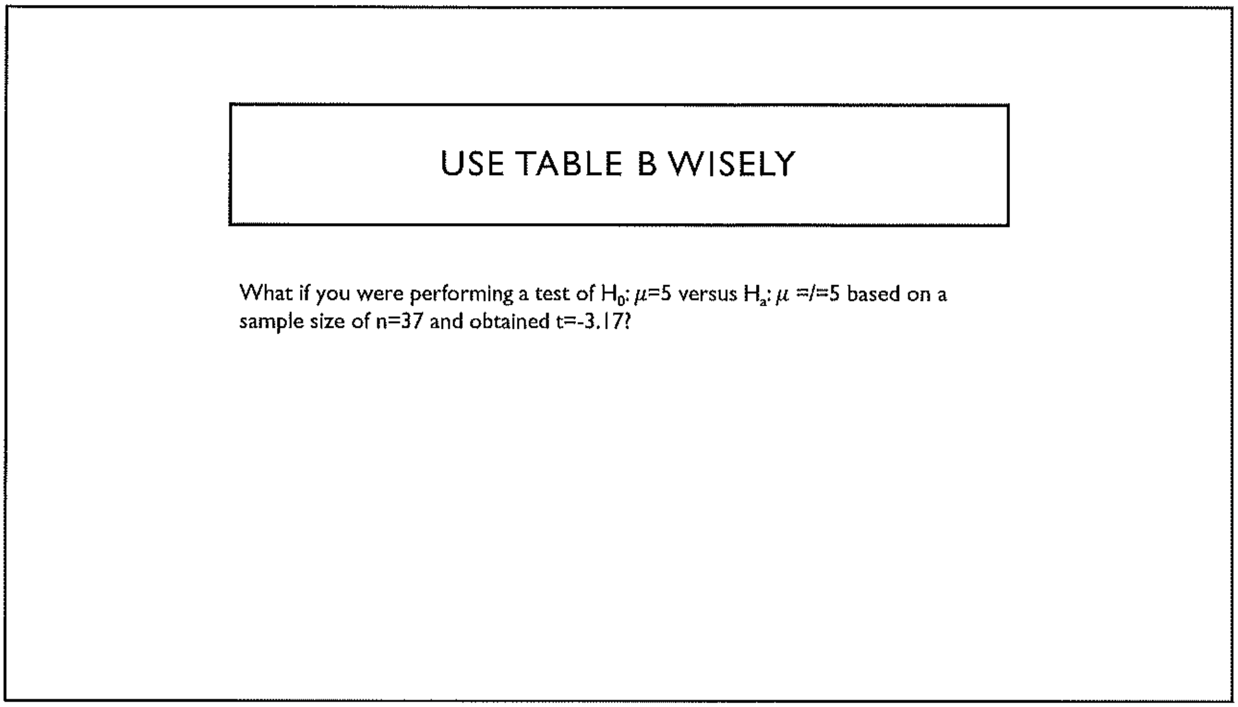
A company has developed a new deluxe AAA battery that is supposed to last longer than its regular AAA battery. However, these new batteries are more expensive to produce, so the company would like to be convinced that they really do last longer. Based on years of experience, the company knows that its regular AAA battery has a Normal distribution and lasts for 40 hours of continuous use, on average.The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained.The sample mean lifetime is $\bar{x}$ = 33.9 hours.A significance tests is performed using the hypotheses:

H_0: μ = 30 hours
H_a: μ > 30 hours

Find the test statistic and possible P-values.

Check conditions first –
Random is good
10% good
Large counts – no good, but mentions normality

T = 33.9 – 30/9.8/sqrt(15) = 1.54
Df = 15-1 = 14
Since we're testing > 30, P(t≥1.54) is between 0.05 and 0.10.

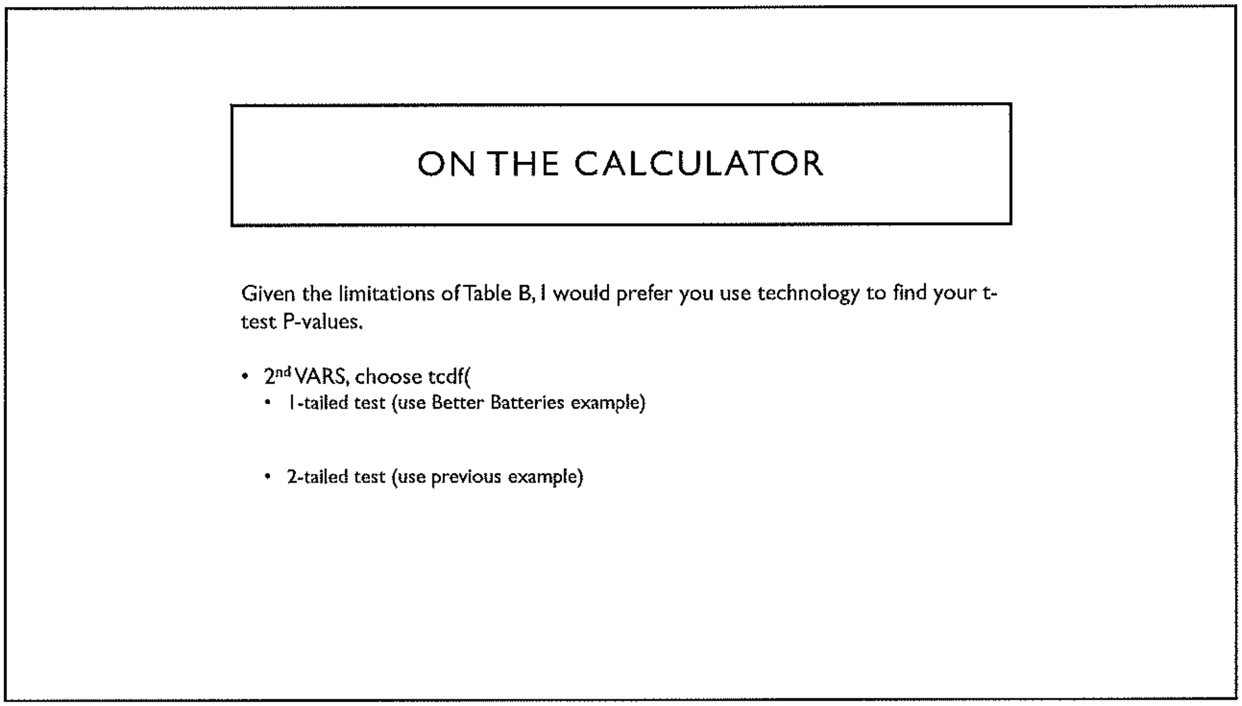


USE TABLE B WISELY

What if you were performing a test of H_0: μ=5 versus H_a: μ =/=5 based on a sample size of n=37 and obtained t=-3.17?

This means you want t>3.17 and t<-3.17. Since t-values fall on a normal distribution, we can use symmetry and just use what we find at 3.17 for twice that. (this works for z-dist too)

We cant get to df=37-1=36, so take the next one below (30). Do not round up!!! It means you're assuming your n is bigger than it really is, and will make your p-value smaller than it really is, leading you to be more likely to make a type 1 error



ON THE CALCULATOR

Given the limitations of Table B, I would prefer you use technology to find your t-test P-values.

- 2nd VARS, choose tcdf(
  - 1-tailed test (use Better Batteries example)
  - 2-tailed test (use previous example)

1 tailed:

lower: 1.54
upper:10000 (just something big)
df: 14
Paste

2 tailed:

lower: -100000 **you can switch these two**
upper: -3.17
df=36
Paste

CHECK YOUR UNDERSTANDING

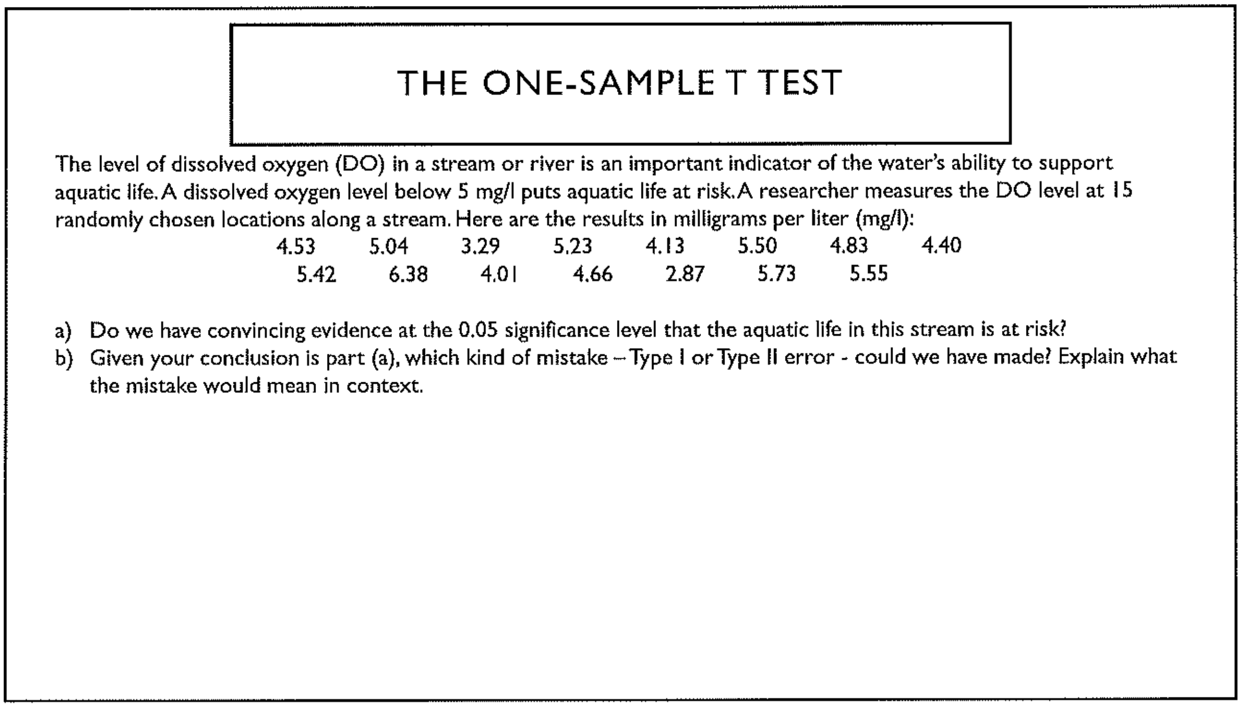
The makers of Aspro brand aspirin want to be sure their tablets contain the right amount of active ingredient (acetylsalicylic acid). So they inspect a random sample of 36 tables from a batch in production. When the production process is working properly, Aspro tablets have an average of $\mu = 370$ mg of active ingredient. The amount of active ingredient in the 36 selected tablets has mean 319 mg and standard deviation 3 mg.

1. State the appropriate hypotheses for a significance test in this setting.

2. Check that the conditions are met for carrying out the test.

3. Calculate the test statistic. Show your work.

4. Use Table B to find the P-value. Then use technology to get a more accurate result. What conclusion would you draw?



THE ONE-SAMPLE T TEST

The level of dissolved oxygen (DO) in a stream or river is an important indicator of the water's ability to support aquatic life. A dissolved oxygen level below 5 mg/l puts aquatic life at risk. A researcher measures the DO level at 15 randomly chosen locations along a stream. Here are the results in milligrams per liter (mg/l):

4.53 5.04 3.29 5.23 4.13 5.50 4.83 4.40
5.42 6.38 4.01 4.66 2.87 5.73 5.55

a) Do we have convincing evidence at the 0.05 significance level that the aquatic life in this stream is at risk?
b) Given your conclusion is part (a), which kind of mistake – Type I or Type II error - could we have made? Explain what the mistake would mean in context.

State: We want to test the following:
H0: μ = 5
Ha: μ < 5, where μ is the true mean dissolved oxygen level in this stream at the alpha = .05 level.

Plan: If the conditions are met, we will perform a one-sample t test for μ:

Random – Measures the DO level at 15 randomly selected locations

Independent: There are an infinite number of possible locations to test along this stream, so it is unneccessary to check this condition.

Normal – We don't know the shape of this distribution, and with a small sample (n=15), we need to graph the data to see if it's safe to use t procedure. ACTUALLY DRAW SOMETHING (histogram might be easy here). It check out, let's do this.

Do: We enter data into the calculator and compute 1-VarStats to obtain xbar = 4.771 and s_x=0.9396. Therefore, the test statistic is:

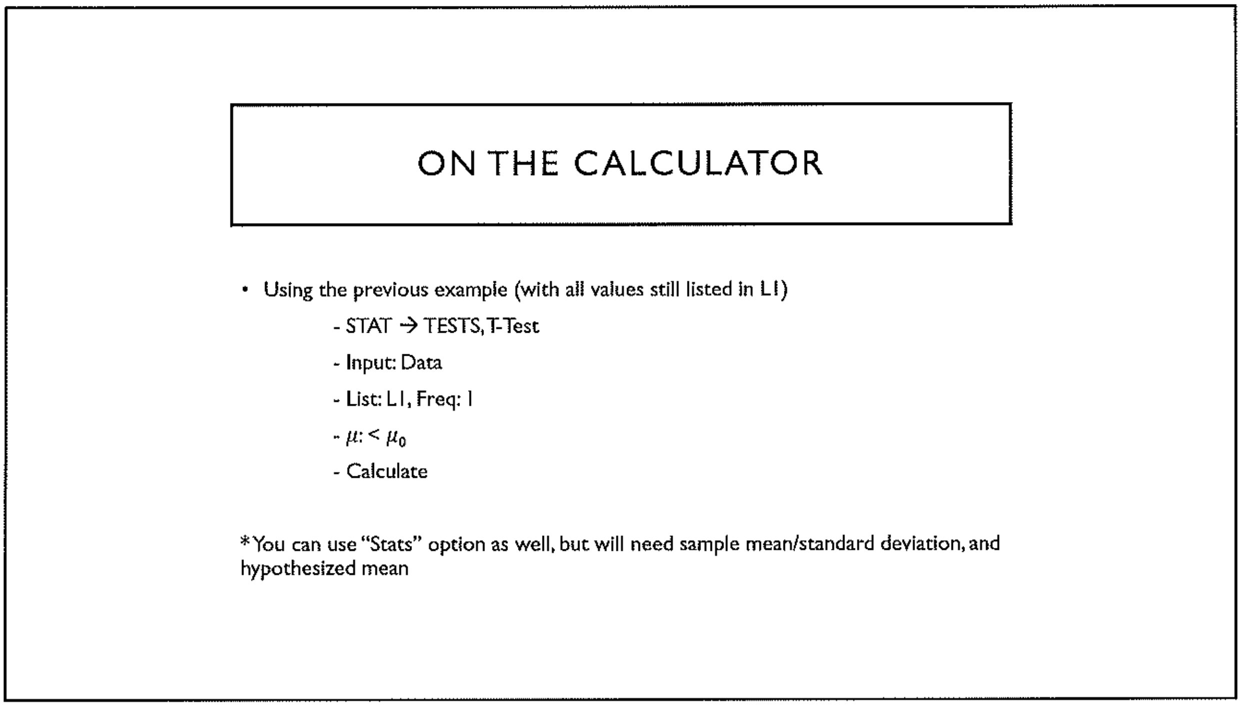
t = 4.771 – 5/0.9396/sqrt(5) = -.94

The P-value for this statistic with degrees of freedom 15-1=14 relates to P(t$\geq$0.94), which is between 0.15 and 0.20. We can find the exact P-value using the calculator:

tcd(lower:-100, upper:-0.94, df:14) = 0.1816.

Conclude: Because the P-value = 0.1816 is greater than alpha = 0.05, we fail to reject Ho. We don't have convincing evidence that the mean DO level in the stream is less than 5 mg/l.

b) Because we decided not to reject H0, we could have made a Type II error (failing to reject Ho when Ho is false). If we did, then the mean DO level μ in the stream is actually less than 5mg, but we didn't find that with our significance test.



ON THE CALCULATOR

- Using the previous example (with all values still listed in L1)
  - STAT → TESTS, T-Test
  - Input: Data
  - List: L1, Freq: 1
  - μ: < μ_0
  - Calculate

*You can use "Stats" option as well, but will need sample mean/standard deviation, and hypothesized mean

CHECK YOUR UNDERSTANDING

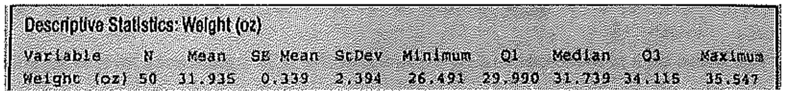
A college professor suspects that students at his school are getting less than 8 hours of sleep a night, on average. To test his belief, the professor asks a random sample of 28 students, "How much sleep did you get last night?" Here are the data (in hours):

9 6 8 6 8 8 6 6.5 6 7 9 4 3 4 5 6 11 6 3 6 6 10 7 8 4.5 9 7 7

Do these data provide convincing evidence at the 0.05 significance level in support of the professor's suspicion?

TWO TAILED T TESTS

At the Hawaii Pineapple Company, managers are interested in the size of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A different irrigation system was installed in this field after the growing season. Managers wonder how this change will affect the mean weight of pineapples grown in the field this year. To find out, they select and weight a random sample of 50 pineapples from this year's crop. The summary data indicates the following:



```
Descriptive Statistics: Weight (oz)
Variable      N    Mean  SE Mean  StDev  Minimum     Q1  Median      Q3  Maximum
Weight (oz)  50  31.935    0.339  2.394   26.491 29.990  31.739  34.115   35.547
```

a) Do these data give convincing evidence that the mean weight of pineapples produced in the field has changed this year?
b) Can we conclude that the different irrigation caused a change in the mean weight of pineapples produced? Explain.

State: We want to test H0: mu = 31, Ha: mu=/= 31 were mu = mean weight in ounces of all pineapples frown in the field this year. Because no sig level is given, we'll use .05

Plan: If conditions are met, we will use a one-sample t test for mu.

Random: yes

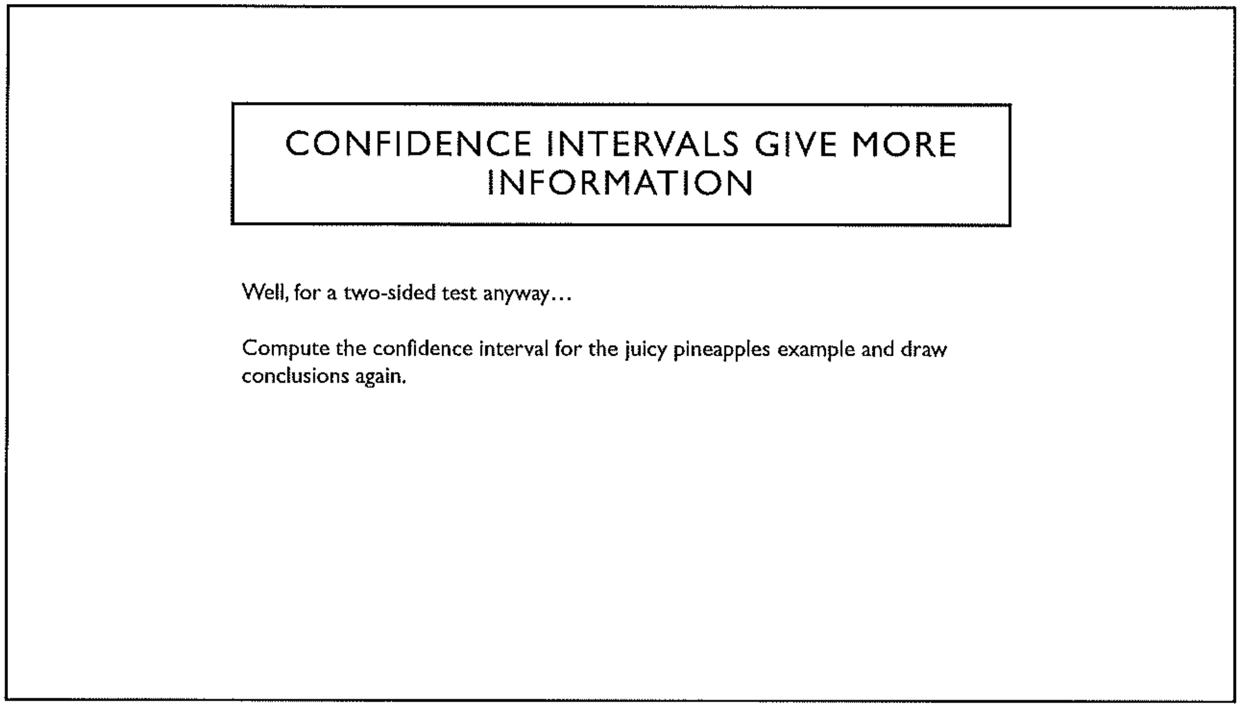
Independent: We can assume there are more than 10(50) = 500 pineapples grown in a "large field"

Normal: Because n=50 $\geq$ 30, we are safe using normal procedures.

Do: The test statistic:
Using the calculator, we use the T-Test function, where xbar = 31.935, sx = 2.394, and df = 49. This renders t=2.762 and P-value = 0.0081.

Conclude: Because the P-value .0081 < alpha = .05, we reject Ho. We have convincing evidence that the mean weight of the pineapples this year is not 31 ounces.

b) No. This was not a comparative experiment, so we cannot infer causation. It is possible that other factors besides the irrigation system changed from last year's growing season (weather, pesticides, etc.)

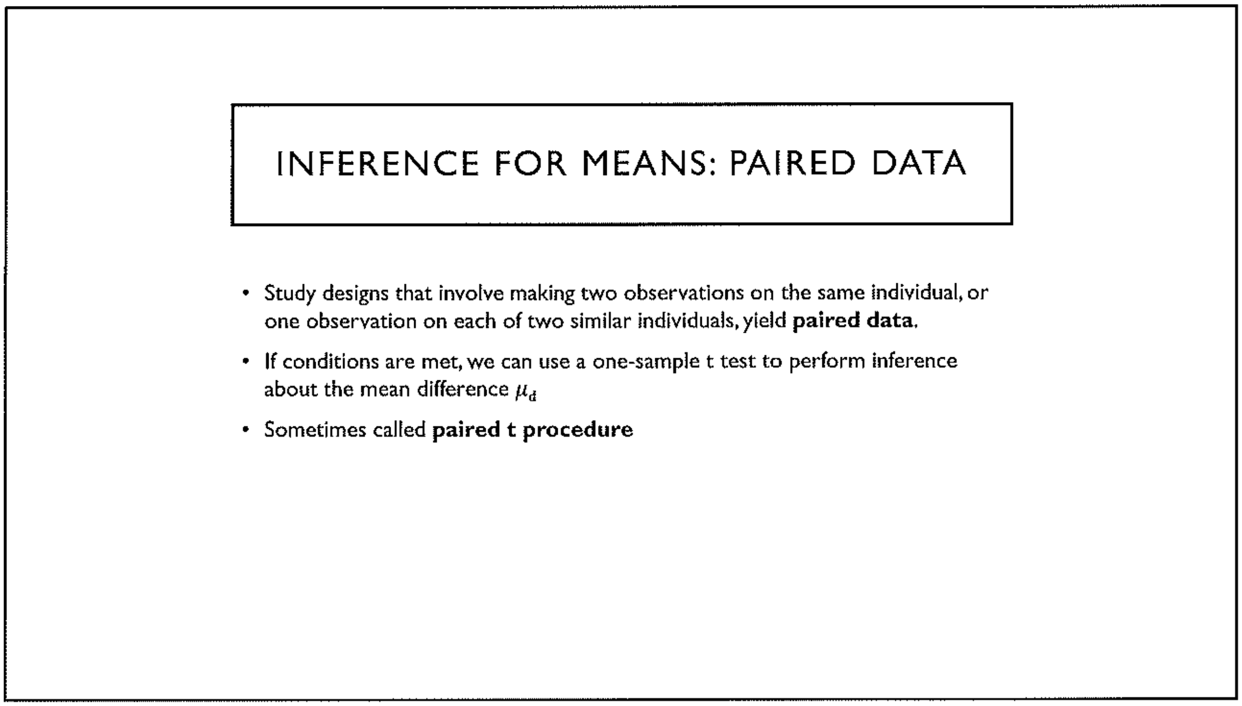


CONFIDENCE INTERVALS GIVE MORE INFORMATION

Well, for a two-sided test anyway…

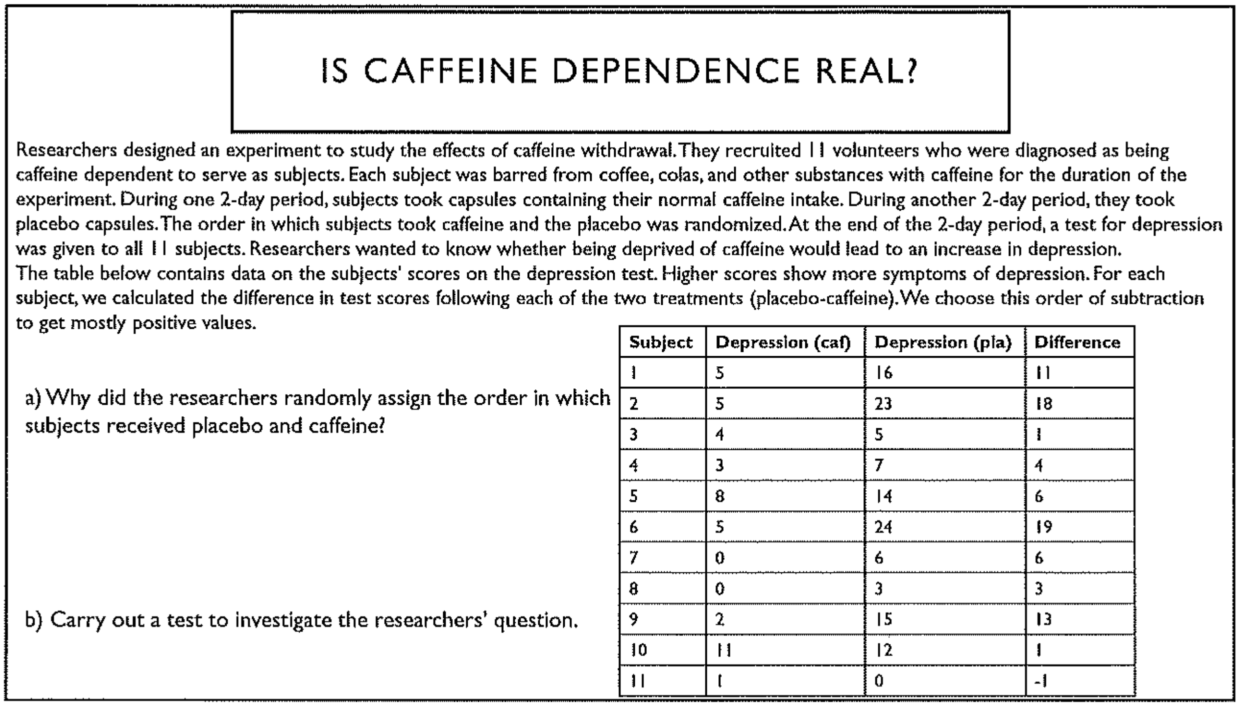
Compute the confidence interval for the juicy pineapples example and draw conclusions again.

Even writing the same hypotheses, we can reject the Ho because the confidence interval does not include the hypothesized value.



INFERENCE FOR MEANS: PAIRED DATA

- Study designs that involve making two observations on the same individual, or one observation on each of two similar individuals, yield **paired data.**
- If conditions are met, we can use a one-sample t test to perform inference about the mean difference μ_d
- Sometimes called **paired t procedure**



IS CAFFEINE DEPENDENCE REAL?

Researchers designed an experiment to study the effects of caffeine withdrawal. They recruited 11 volunteers who were diagnosed as being caffeine dependent to serve as subjects. Each subject was barred from coffee, colas, and other substances with caffeine for the duration of the experiment. During one 2-day period, subjects took capsules containing their normal caffeine intake. During another 2-day period, they took placebo capsules. The order in which subjects took caffeine and the placebo was randomized. At the end of the 2-day period, a test for depression was given to all 11 subjects. Researchers wanted to know whether being deprived of caffeine would lead to an increase in depression.
The table below contains data on the subjects' scores on the depression test. Higher scores show more symptoms of depression. For each subject, we calculated the difference in test scores following each of the two treatments (placebo-caffeine). We choose this order of subtraction to get mostly positive values.

Subject	Depression (caf)	Depression (pla)	Difference
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	-1

a) Why did the researchers randomly assign the order in which subjects received placebo and caffeine?

b) Carry out a test to investigate the researchers' question.

a) To be able to conclude statistical significance is due to the treatments themselves, not some other variable (or chance).

b) State: If caffeine deprivation has no effect on depression, then we would expect the actual mean difference in depression scores to be 0. Therefore, we want to test hypotheses:

H_0: $\mu_d = 0$

H_a: $\mu_d > 0$

Where μ is the true mean difference (placebo – caffeine) in depression score for subjects like these. We will use the significance level alpha = 0.05.

Plan: If conditions are met, we will conduct a paired t test for μ_d:

Random: Randomly assigned treatments to the subjects →

Independence: no need to test because we are not sampling

Normal: We don't know the shape of the distribution, nor is the sample size large enough (n=11 < 30). We graph the DIFFERENCE data (since that's what we're measuring) to assess shape (box and whisker looks easiest here). It's approx normal.

Do: Entering the difference list into List 1 on the calculator, we use the t test command with degrees of freedom 11-1 =10. The test statistic t=3.53 with P-value = 0.0027.

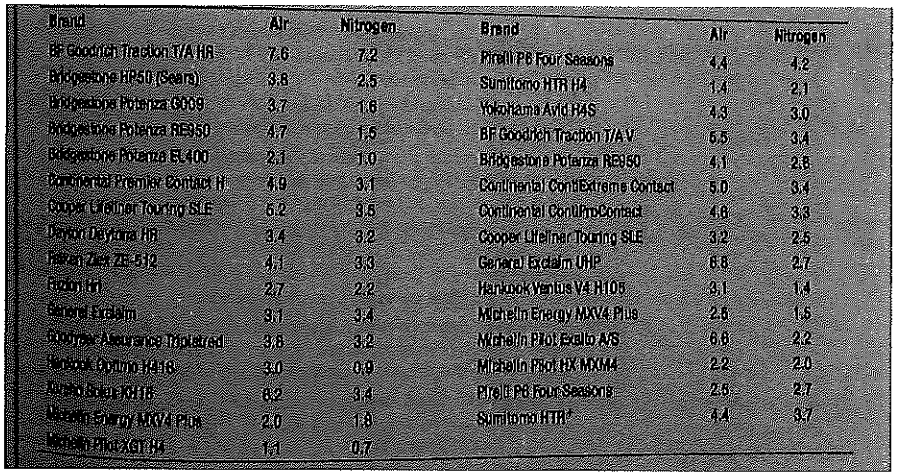
Conclude: Because P=0.0027 < alpha = 0.05, we reject Ho. We have convincing evidence that the true mean difference (placebo – caffeine) in depression score is positive for subjects like these.

*********RANDOM SELECTION (SRS) IS FOR MAKING INFERENCE ABOUT A POPULATION. RANDOM ASSIGNMENT IS FOR MAKING CAUSE AND EFFECT CONCLUSIONS**********

**Subjects not chosen randomly, so we can't generalize to the population. We can, however, say these results are for "people like these guys."

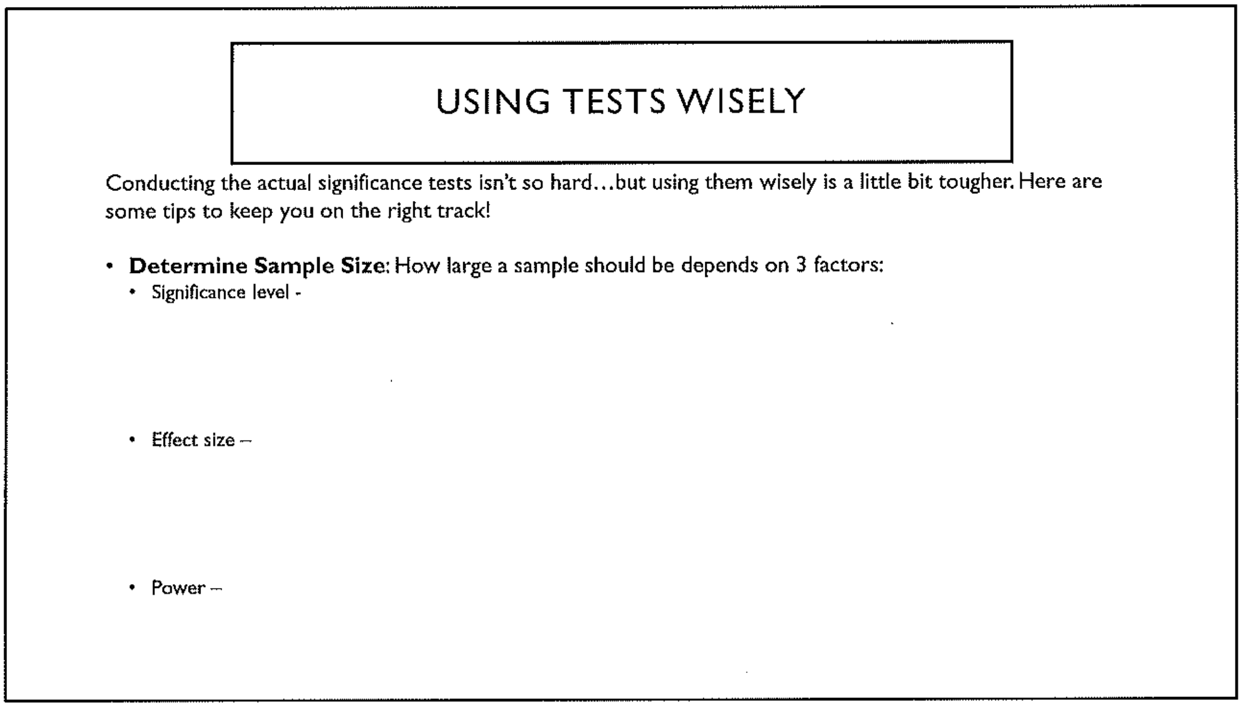
CHECK YOUR UNDERSTANDING

Given the values below, do the data give convincing evidence at the 0.05 significance level that filling tires with nitrogen instead of air decreases pressure loss?



Brand	Air	Nitrogen	Brand	Air	Nitrogen
BF Goodrich Traction T/A HR	7.6	7.2	Pirelli P6 Four Seasons	4.4	4.2
Bridgestone HP50 (Sears)	3.8	2.5	Sumitomo HTR H4	1.4	2.1
Bridgestone Potenza G009	3.7	1.6	Yokohama Avid H4S	4.3	3.0
Bridgestone Potenza RE950	4.7	1.5	BF Goodrich Traction T/A V	5.5	3.4
Bridgestone Potenza EL400	2.1	1.0	Bridgestone Potenza RE950	4.1	2.8
Continental Premier Contact H	4.9	3.1	Continental ContiExtreme Contact	5.0	3.4
Cooper Lifeliner Touring SLE	5.2	3.5	Continental ContiProContact	4.8	3.3
Dayton Daytona HR	3.4	3.2	Cooper Lifeliner Touring SLE	3.2	2.5
Falken Ziex ZE-512	4.1	3.3	General Exclaim UHP	6.8	2.7
Fuzion Hri	2.7	2.2	Hankook Ventus V4 H105	3.1	1.4
General Exclaim	3.1	3.4	Michelin Energy MXV4 Plus	2.5	1.5
Goodyear Assurance Tripletred	3.8	3.2	Michelin Pilot Exalto A/S	6.8	2.2
Hankook Optimo H418	3.0	0.9	Michelin Pilot HX MXM4	2.2	2.0
Kumho Solus KH16	6.2	3.4	Pirelli P6 Four Seasons	2.5	2.7
Michelin Energy MXV4 Plus	2.0	1.8	Sumitomo HTR+	4.4	3.7
Michelin Pilot XGT H4	1.1	0.7			

Ho: mu(d) = 0 vs Ha: mu(d) > 0 where mu(d) = true mean difference (air – nitrogen).
Conditions met
D: xbar = 1.252, sx = 1.202, t=5.80, P =0.000…
We reject Ho, we have convincing evidence that the true mean difference in pressure lost > 0

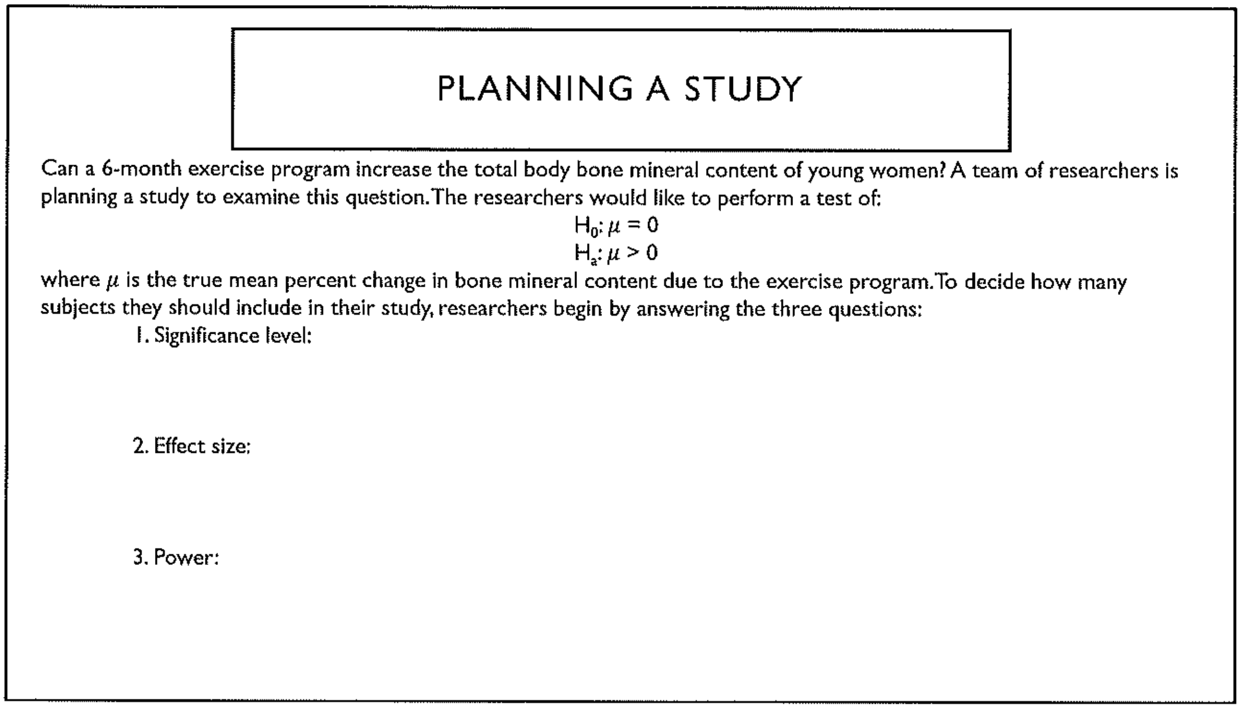


USING TESTS WISELY

Conducting the actual significance tests isn't so hard…but using them wisely is a little bit tougher. Here are some tips to keep you on the right track!

- **Determine Sample Size:** How large a sample should be depends on 3 factors:
  - Significance level -
  - Effect size –
  - Power –

1. If consequences are serious, opt for a stricter alpha (less likely to make Type I error). If you want to avoid Type II error, use a bigger alpha. Recall, though, that a larger n reduces chance for Type II error. (So increase n, decrease alpha, works for both)
2. Determine how large the difference between null and actual parameter value is important for us to detect (remember, larger difference means easier to detect)
3. Probability of avoiding type 2 error



PLANNING A STUDY

Can a 6-month exercise program increase the total body bone mineral content of young women? A team of researchers is planning a study to examine this queštion. The researchers would like to perform a test of:

$H_0: \mu = 0$

$H_a: \mu > 0$

where μ is the true mean percent change in bone mineral content due to the exercise program. To decide how many subjects they should include in their study, researchers begin by answering the three questions:

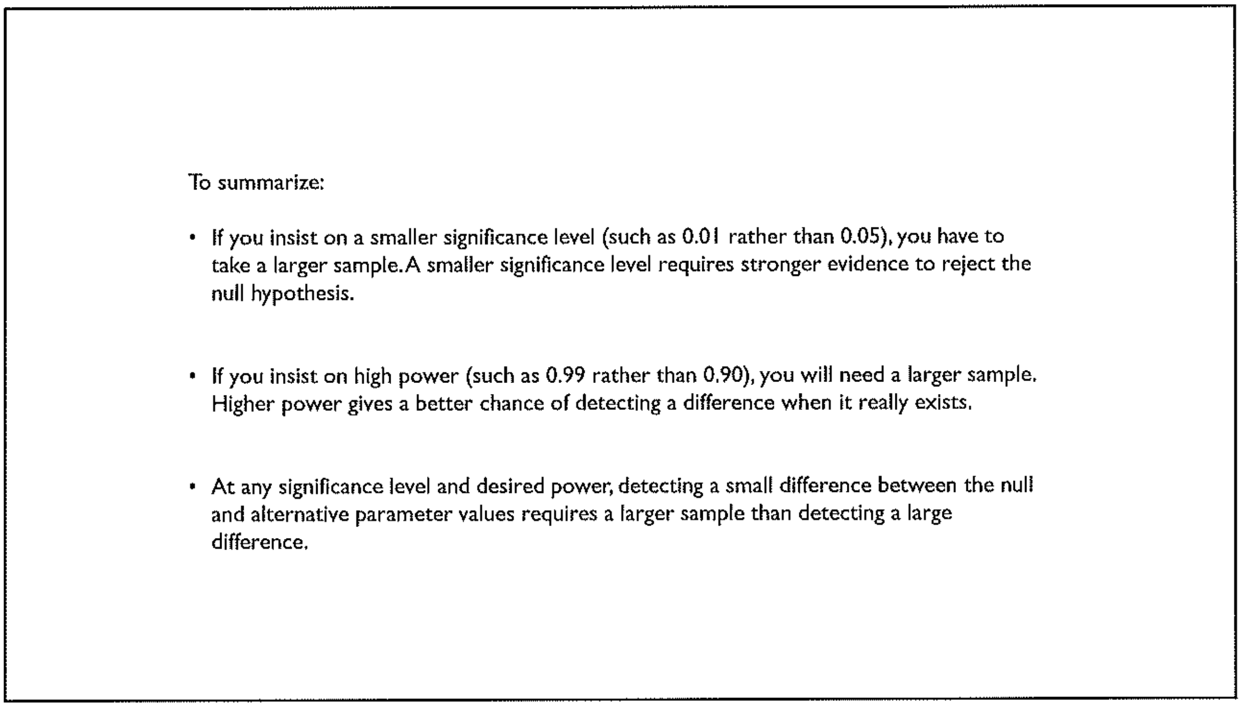
1. Significance level:

2. Effect size:

3. Power:

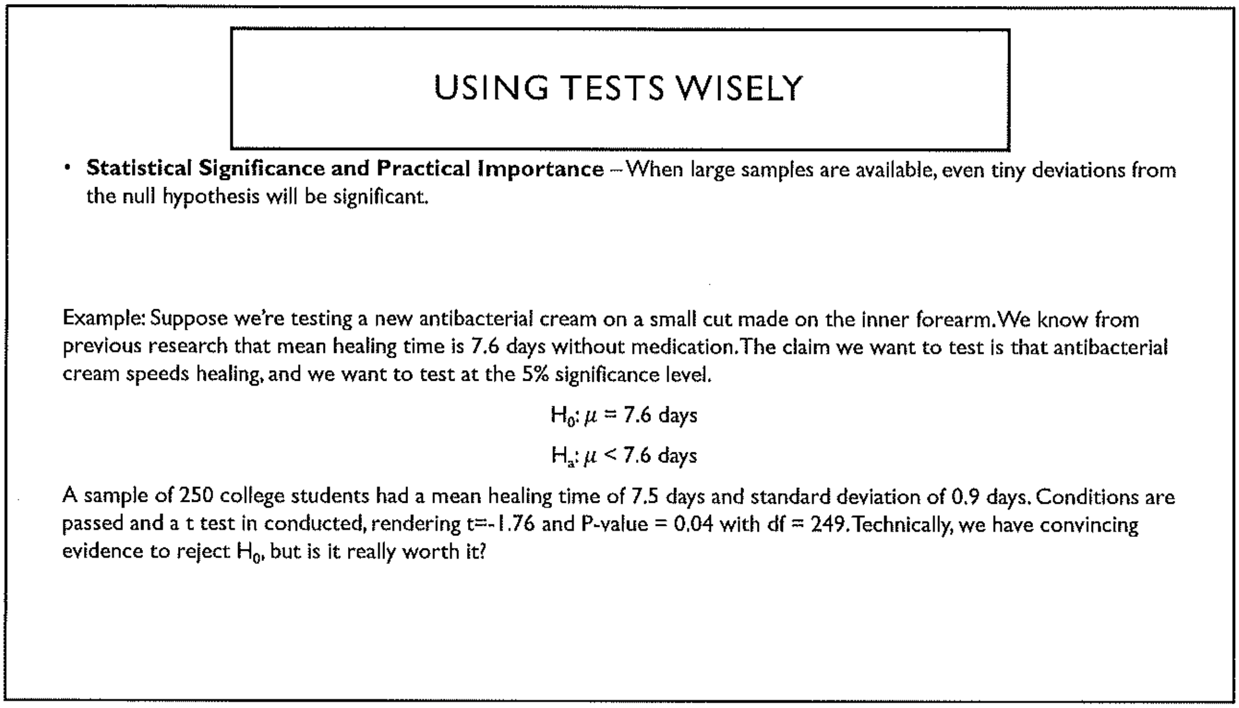
1. Alpha = 0.05 is best because it gives enough protection against declaring that the exercise program increases content when it really doesn't (type I error)
2. A mean increase of 1% would be considered important to detect
3. Want probability of at least 0.9 that a test at the chosen alpha will reject the null when the truth is $\mu = 1$

**ACTIVITY: pg 590 investigate power



To summarize:

- If you insist on a smaller significance level (such as 0.01 rather than 0.05), you have to take a larger sample. A smaller significance level requires stronger evidence to reject the null hypothesis.
- If you insist on high power (such as 0.99 rather than 0.90), you will need a larger sample. Higher power gives a better chance of detecting a difference when it really exists.
- At any significance level and desired power, detecting a small difference between the null and alternative parameter values requires a larger sample than detecting a large difference.



USING TESTS WISELY

- **Statistical Significance and Practical Importance** – When large samples are available, even tiny deviations from the null hypothesis will be significant.

Example: Suppose we're testing a new antibacterial cream on a small cut made on the inner forearm. We know from previous research that mean healing time is 7.6 days without medication. The claim we want to test is that antibacterial cream speeds healing, and we want to test at the 5% significance level.

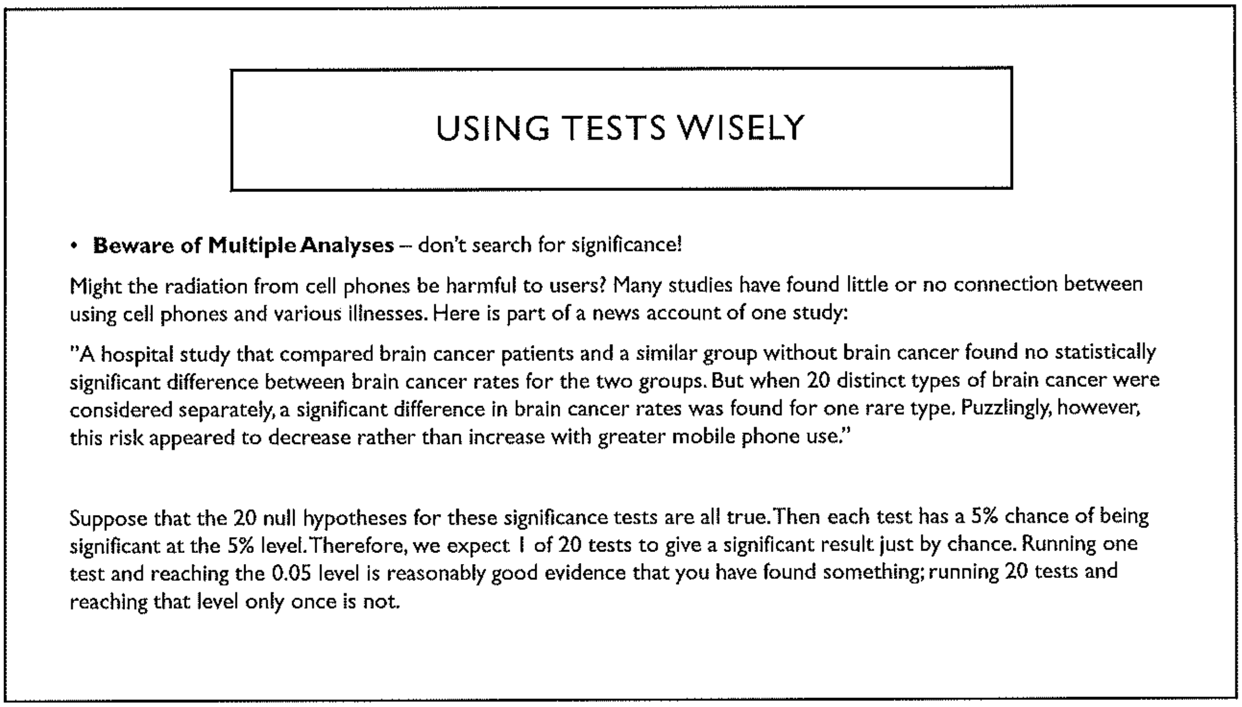
$$H_0: \mu = 7.6 \text{ days}$$

$$H_a: \mu < 7.6 \text{ days}$$

A sample of 250 college students had a mean healing time of 7.5 days and standard deviation of 0.9 days. Conditions are passed and a t test in conducted, rendering t=-1.76 and P-value = 0.04 with df = 249. Technically, we have convincing evidence to reject H_0, but is it really worth it?

Large sample? Small effect size OR smaller alpha

**REMEMBER: statistical significance is not the same as practical importance!



USING TESTS WISELY

- **Beware of Multiple Analyses** – don't search for significance!

Might the radiation from cell phones be harmful to users? Many studies have found little or no connection between using cell phones and various illnesses. Here is part of a news account of one study:

"A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant difference between brain cancer rates for the two groups. But when 20 distinct types of brain cancer were considered separately, a significant difference in brain cancer rates was found for one rare type. Puzzlingly, however, this risk appeared to decrease rather than increase with greater mobile phone use."

Suppose that the 20 null hypotheses for these significance tests are all true. Then each test has a 5% chance of being significant at the 5% level. Therefore, we expect 1 of 20 tests to give a significant result just by chance. Running one test and reaching the 0.05 level is reasonably good evidence that you have found something; running 20 tests and reaching that level only once is not.