Notes Ch 11-12

Section 11.1 Significance tests- basics

A ***significance test*** is a formal procedure for comparing observed data/ testing claims

Null Hypothesis: H_0 The statement being tested- the significance test itself is set up to asses the strength of evidence **AGAINST the Null Hypothesis.** Usually the null is the "no change" "equal to" "no difference" statement. The "status quo" hypothesis

Alternative Hypothesis: H_a: The **claim** about the population we are trying to find **evidence FOR**. Suggests something HAS changed.

One sided claim: suggests a change from the Null in one direction. This is usually the < or > symbol in the ALTERNATIVE Hyp (H_a).

Two Sided claim: suggests a change in either direction. Usually indicated with a $\neq$ in the H_a.

Hypotheses must be stated in a population parameter (pop mean, pop proportion, pop var or standard dev)(μ, ρ, σ^2, σ,)

When setting up a significance test, it's best to start with the Alternative Hyp (H_a). It's not always straight forward or clear but establishing this then doing the null may be easier.

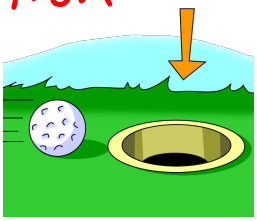
NOTE: in the real world, the hypotheses are set up BEFORE data is collected so be sure not to use the sample data first in setting up hypotheses.

Example: Mike is an avid golfer who would like to improve his play. A friend suggests finding new clubs and even let's Mike try his 7-iron. Based on years of experience, Mike has established that the mean distance that balls travel with his old 7-iron is =175 yards and =15 yards. **He is hoping that his new clubs will make his 7-iron shots more consistent (less variability)** so he goes to the driving range and hits 50 shots with the new 7-iron.

(a) Describe the parameter of interest in this setting.

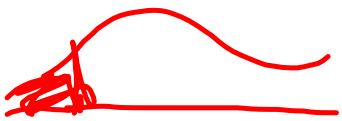
The parameter of interest is standard deviation as it relates to variability of a new 7-iron

(b) State an appropriate Null and Alternative hypothesis.



$H_0: \sigma = 15$

$H_a: \sigma < 15$ (claim)



Example: According to a website, 85% of teens are getting less than 8 hours of sleep a night. Jannie wonders whether this result holds to her large high school. She takes an SRS of 100 students at her school and asks them how much sleep they get on a typical night. In all, 75 of the responders said they get less than 8 hours of sleep.

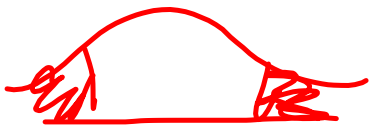
(a) Describe the parameter of interest

The parameter of interest is the proportion of students who get less than 8 hrs of sleep

(b) State appropriate hypotheses for a significance test.

$H_0: p = .85$

$H_a: p \neq .85$ claim



Conditions for Significance testing are the same 3 conditions that keep coming up

1. SRS
2. Independence
3. Normality

The P-Value

The smaller the P-value, the stronger the evidence is against the Null Hypothesis (H_0).

Where does the P-Value come from????? Z scores and Z charts!(yay!!!)

The P-Value is a quantitative measure of just how unlikley a given finding is, assuming the null hypothesis is true. We may compare this value to a significance level (α) in order to decide whether or not the finding is significantly different from what was expected

Test Statistic → $z = \frac{\bar{x} - \mu_o}{\sigma / \sqrt{n}}$

hypothesized value

Draw a DIAGRAM!!!! you have to see it on the normal curve.

note: For two sided P- value (on H_a) double the P for total area comparison.

Example: *7-iron continued*

When Mike was testing the new 7-iron, the hypotheses were:

H_0: = 15

H_a: < 15 (claim)

where is the true standard deviation of the distance over which Mike hits golf balls using the new 7-iron. Based on 50 shots with the new 7-iron, the standard deviation was s_x=10.9 yards. A significance test using the sample data prodiced a P-value of 0.002 (a) Interpret this P-value in this context. (b) Do the data provide convincing evidence against the null hypothesis. Explain.

At .2% chance of the St. Deviation being 10.9 just by chance alone is significant enough value to reject the Null & support the claim that S.D. < 15

Statistical Significance

Significance level ( ): The data gives evidence so strong that it would happen no more than (%) of the time. If the P-value is as small or smaller than alpha, we say the data are statistically significant at a level . And therefore it would be unlikely for the hypothesized value to be true.

Alpha usually takes on values of .05, .01, or .10. (which happens to correspond to our confidence levels). Alpha can take on other values, but these are most common. 0.01 would be the strongest.

So, if =0.05, that means that we are looking at data against H_0 so strong that it could not happen more than 5% of the time.

Remember, ***Significant*** *in statistics does not mean "important". It means "**not likely to happen just by chance.**"*

Interpreting Results.....To Reject or Fail to Reject the H_0 (null)

Important: Failing to reject does NOT mean Accept!!!!

Conclusions should have a CLEAR connection to your calculations and should be stated in the context of the problem.

- We reject the H_0 (null) if our sample result is too unlikely to have occurred by chance under the assumption that H_0 is true. We reject H_0 if our result is statistically significant as compared to our significance level . Always compare your P-value to the alpha and decide if it's reject or fail to reject the Null (H_0)
- If we *reject the* H_0, and our claim is H_a, then we support the claim.
- If we *fail to reject the* H_0, then we do not have enough evidence to support our claim (in the H_a) because we can't reject the possibility of the null (H_0) being true.
- Your alpha level must be **pre-established** before calcs are done. Otherwise it can be viewed as deceptive statistics.

Example: *A company has developedd a new delux 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 significence test is performed.*

P-Val = .0276

Write the hypostheses, perform a significance test at an α=0.05, *draw a conclusion and interpret the results.*

The parameter of interest is mean (hours) for AAA batteries

$H_0: \mu = 30$

$H_a: \mu > 30$

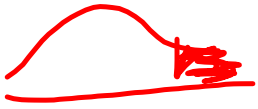
Since $P = .0276 < .05$ we have sufficient evidence to reject the null ($\mu = 30$) and support the claim that the mean battery life for new AAA batt will be longer than 30 hr.

If $\alpha = .01$

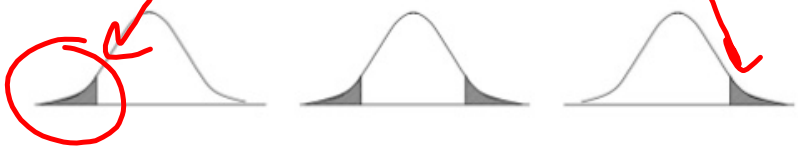
$.0276 \not< .01$

One Sided vs Two Sided tests:

Summary of One- and Two-Tail Tests…



One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_a : \mu < \mu_0$	$H_a : \mu \neq \mu_0$	$H_a : \mu > \mu_0$



11.37

The Alternative Hypothesis (H_a)determines 1 or 2 sided.....

Example: *For his second semester project in AP statistics, Benny decided to investigate whether students in his school prefer name-brand potato chips over generic. He randomly selects 50 students and had each student try both types of chips, in random order. Overall, 34 of the students preferred the name-brand chips. Benny performed a significance test. Determine the hypotheses, calculations, and make a proper conclusion based on results. Use =.01.* $S=.5$

How would it be different if you used $\alpha=.05$.

Bonus: How Confidence intervals can be used as a significance test:

Find the interval then determine if the hypothesized amount is in the interval. We can reject the Null (H_o) if it falls outside the interval at an alpha level of 1- %confidence

Ex: *A 95% confidence interval for a population mean is 31.5± 3.5.* (28, 35)

(a) Can you reject the Null hypothesis that =34 at a 5% significance level? Why/ Why not? At a 5% significance level, we cannot reject The H_o of $\mu=34$ as it is contained in our 95% conf. interval

(b) Can you reject the Null hypothesis that $\mu=36$ *at a 5% significance level? Why/ Why not?* At 5% Signf. Level, we can reject the H_o of $\mu=36$ as it is outside the 95% conf. Int.

EX: *The P-value for a one sided test of* H_o: $\mu=30$ *is .04. Would the 95% confidence interval for* μ *include 30? Explain.* pval = .04 Since .04 < .05 $\alpha=.05$ then we can reject the $H_o=30$ B/C not in the conf. interval

Section 11.2ish

Carrying out Significant tests

Steps:

1.Identify the population of interestand the parameter and write the hypotheses.

2. Choose an appropriate inference procedure and check conditions for using it.

3. Calculate- carry out your procedure

- calc test statistic
- Find P-Value
- Use Z test

4. Interpret your results in the context of the problem and write your conclusion with respect to the context.

Alternative: Z TESTS for ONE SAMPLE Population Mean

1. Write your H_o and H_a

2. Draw your bell cuve and shade the appropriate section as related to your H_a.

Determine if left tailed, right tailed or two tailed. Use as your boundary Z

*The shaded region is what is called- the **REJECTION REGION.***

3. Calculate your test statistic. Is the test statistic in the shaded region or the non shaded region.

- If test statistic in Shaded region: reject the Null (H_0)
- If test statistic in non shaded region: Fail to reject the Null(H_0)

Example: *A company that manufactures classroom chairs for high school students claims that the mean breaking strength of the chairs is 300 lbs with standard deviation of 5lbs. One of the chairs collapsed beneath a student that weighted less than 220 lbs last week. You suspect the manufacturer is exaggerating the breaking strength of the chairs so you would like to perform a test of their claim. A random sample of 40 chairs showed a mean breaking strength of 294 lbs. Use =0.05.*

Intro: parameter of interest, check conditions, etc

H_o/H_a with graph

Test statisitc

Conclusion

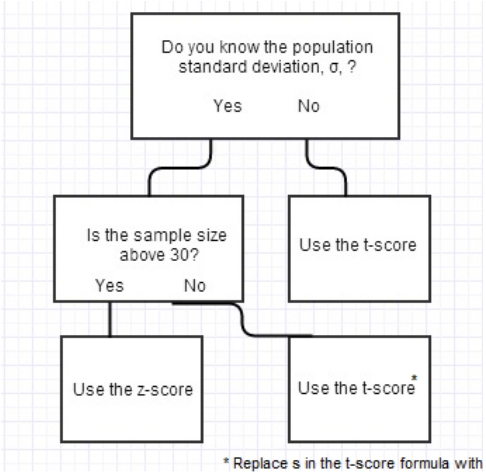
T-score vs. z-score: When to use a t score (these are both alternatives to P-value test)

The general rule of thumb for when to use a ***t score*** is when your sample:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Has a sample size below 30,
- Has an unknown population standard deviation.
- (also other rules for t table from last unit)

You must know the standard deviation of the population and your sample size should be above 30 in order for you to be able to use the z-score. Otherwise, use the t-score.



Example: *A tablet computer manufacturer claims that its batteries last an average of 10.5 hours when playing videos. The quality control dept randomly selects 20 tablets from each days production and tests the fully charged batteries by playing a video repeatedly until the battery dies. The quality control dept will discard batteries from the days production run if they find convincing evidence that the mean mattery life is less than 10.5 hrs. State the hypotheses for the quality control dept. test. Be sure to define the parameter. Check the conditions for t-test and perform a t-test.*

Example: *At the Hawaii Pinapple Company, the mean weight of pineapples harvested from one large field was 31 ounces last year. A different irrigation system was installed in the field after the growing season. Managers wonder if this change will affect the mean weight of future pineapples grown in the field. To find out, they select a random sample of 50 pineapples from this years crop.*

(a) State the appropriate hypotheses being sure to define the parameter fo interest

(b) Check conditions

(c) The mean of the sample was 31.9 oz with a standard deviation of 2.33 oz. Using an alpha of 0.05, perform the appropropriate test.

(d) Make a 95% confidence interval for this data. How could you use this interval to test this claim?

(e) Can we conclude the new irrigation system has was the cause in the change in the mean of the pineapple produced?

(c) Do the data give good evidence that the mean change in the population is greater than zero? Do a complete significance test.

PAIRED T/Z-test: exploring the mean difference and testing that difference.

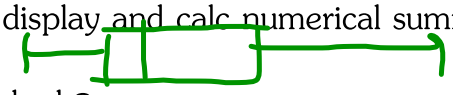
EX: We suspect that students will generally score higher the second time they take the SAT Math exam than on their first attempt.

Suppose we know that the changes in the score (Second-first) have population standard deviation $\sigma = 50$.

Here are the results for 46 randomly chosen high school students:

-30	24	47	70	-62	55	-41	-32	128	-11	-43	122	-10
56	32	-30	-28	-19	1	17	57	-14	-58	77	27	-33
51	17	-67	29	94	-11	2	12	-53	-49	49	8	-24
96	120	2	-33	-2	-39	99						

(a) Construct a graphical display and calc numerical summaries for the data.



How does the distribution look?

The distribution appear modestly Skewed Right Based on the Boxplot/Histo/NPP. The mean is 13.11 the median is 2 which also supports Skewness of data

(b) Based on your work from part (a) , do you belive that the population of differences in the Math SAT score is Normally distributed?

(c) Do the data give good evidence that the mean change in the population is greater than zero? Do a complete significance test.

Test for Population Proportion.

Different z formula- same rules apply

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

Check:

SRS, independence,

np >10, nq >10 (normality)

Example: *According to the CDC website, 68% of high school students have never smoked a cigarette. Paul wonders whether this national result holds true for is large urban school. For his AP stats project, Paul surveys a SRS of 150 students in his school. he gets responses to all 150 students and of those 90 say they have never smoked a cigarette. Is there confincing evidence that the CDC's claim does not hold true for Paul's school?*

The parameter of interest is proportion of students in Pauls school that have never smoked a cig.

SRS ✓ ind. ✓ $\hat{p} = \frac{90}{150} = .6$ $\alpha = .01$

$np > 10 \Rightarrow (150)(.6) > 10$ ✓ $(150)(.4) > 10$ ✓

$H_0: p = .68$

$H_a: p \neq .68$

$z = \frac{.6 - .68}{\sqrt{\frac{(.68)(.32)}{150}}} = -2.1$

$.0178 \times 2 = .0357$

Example: *Potato chips must meet a certain quaility standards. If a potato 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 of potatoes from the supplier. Otherwise the entire truck load will be used to make potato ships. The potato chip producer has just recieved a truckload of potatoes from the supplier. A supervisor selects a random sample of 500 potatoes and found 47 to have blemishes. Is there convincing evidence at a 0.10 significance level that more than 8% of the potatoes in the shipment have blemishes?*

use $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

The parameter of interest is proportion of potatoes that have blemish in a truck load.

SRS? ✓ Indep? ✓ $np > 10$ ✓ $nq > 10$

$H_0: p = .08$

$H_a: p > .08$

$z = 1.154$

$P\text{-val} = .123$

$.123 \not< .1$

FTR

Decis. FTR H_0

Insufficient evid. at $\alpha = .1$ to Reject H_0 & support $p > .08$

Keep truck load.

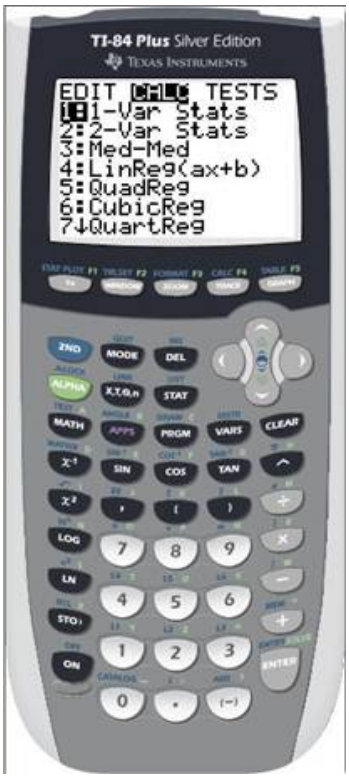
$z \geq 1.28$

Technology Time:

1-prop z test (STAT>>TESTS>>#5)

T- test (STAT>>TESTS>>#2)

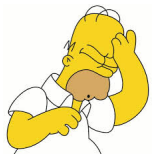
Z-test (STAT>>TESTS>>#1)



Accounting for Errors in Significance Testing:

Type I error: When we reject the Null (H_0) when in fact it is true.

Type II error: When we FTR the Null (H_0) when we should have rejected it.



Follow these guidelines:

If H_0 is true:

- Our conclusion is correct if we don't find convincing evidence that H_a is true
- We make a Type I error if we find convincing evidence that H_a is true

If H_0 is rejected (H_a is true):

- Our conclusion is correct if we find convincing evidence that H_a is not true
- We make a Type II error if we do not find convincing evidence that H_a is true

The *significance level* is the percent chance that a *type I error occurs*. Type II is far ore complicated in determing (will get to this).

$H_0: p = .08$

$H_a: p > .08$

What are the consequences to these errors?

EX: Let's go back to the potato chip problem. Tell what a Type I and Type II would look like.

Type I error: (Reject H_0 + should not have): we would end up to send the potato truck away erroneously thus missing out on a perfecty good batch of potatoes

Type II error (FTR H_0 + should have rejected it) we would have kept potato truck and used those potatoes but more than 8% are blemished and we can't use them

EX. *Your company markets a computerized device for detecting high blood pressure. The device measures an individual's blood pressure once per hour at a randomly selected time throughout a 12 hour period. Then it calculates the mean systolic (top number) pressure for the sample of measurements. Based on the sample results, the device determines whether there is significant evidence that the individuals actual mean systolic pressure is greater than 130. If so, it is recommended the person seek medical attention. Describe a type I and II error. (first state hypotheses)* $H_0: \mu = 130$ $H_a: \mu > 130$

Type I. Reject H_0 But should not have.: we would say $\mu > 130$ and thus have person seek medical attention

Type II: FTR H_0 But should have rejected it. not evid. for $\mu > 130$, so we don't send people who may need med attention to go get it.

The POWER of a Hypothesis test and Type II error- (super complicated- trying to make less complicated)

note: on the AP you will not be asked to calculate the Power of a test, but may be asked to interpret the Power value you are given

The POWER of a hypothesis test is nothing more than 1 minus the probability of a Type II error. Basically the **POWER OF A TEST** is the **probability that we make the right decision** when the null is not correct (i.e. we correctly reject the Ho).

If statistical power is high, the probability of making a Type II error, or concluding there is no effect (not rejecting Ho) when, in fact, there is one, goes down.

Bigger effects are easier to detect than smaller effects, *while large samples offer greater test sensitivity than small samples.* *(the last pages of this go into descriptions about what "effect" is)*

The power of any test of statistical significance will be affected by four main parameters:

- the effect size
- **the sample size (N). Increasing the sample size, decreases variability and improves x (bar) (sample mean) and thus will increase the power**
- **the alpha significance criterion (α). Increasing the alpha decreases the beta (β) which decreases the probability of a type II error (which is the error not supporting the Ha)**
- the chosen or implied beta (β)

All four parameters are mathematically related.

The power is relates to the detection of an true effect in what we are testing. For example, if we had a null where $\mu=15$ and some alternative $\mu= 25$ with a power of .3 (30%) and adjusted the alternative to $\mu=35$ would give a higher power bc the difference between 15 and 35 is greater, the effect we are looking to detect is more noticeable. (a computer program or applet would do the calc)

Why is this good to know?

If you knew prior to conducting a study that you had, at best, only a 30% chance of getting a statistically significant result, would you proceed with the study? Or,would you like to know in advance the minimum sample size required to have a decent chance of detecting the effect you are studying? These are the sorts of questions that power analysis can answer.

Let's take the first example where we want to know the prospective power of our study and, by association,
the implied probability of making a Type II error. In this type of analysis we would make statistical power the
outcome contingent on the other three parameters (sample size, alpha, effect) . This basically means that the probability of getting a statistically significant result will be high when the effect size is large, the N is large, and the chosen level of alpha is relatively high (or relaxed).

For example, if I had a sample of N = 100 and I expected to find an effect size equivalent to r = .30, a
quick calculation (on a computer program) would reveal that I have an 57% chance of obtaining a statistically significant result using a two-tailed test with alpha set at the conventional level of .05. If I had a sample twice as large, the probability that my results will turn out to be statistically significant would be 86%. (increasing n, increased my power)

Or let's say we want to know the minimum sample size required to give us a reasonable chance (.80) of
detecting an effect of certain size given a conventional level of alpha (.05). We can look up a power table or
plug the numbers into a power calculator to find out.

For example, if I desired an 80% probability of detecting an effect that I expect will be equivalent to r = .30
using a two-tailed test with conventional levels of alpha, a quick calculation reveals that I will need an N of at
least 84. If I decide a one-tailed test is sufficient, reducing my need for power, my minimum sample size falls
to 67.

You run a power analysis for many reasons, including:

- To find the number of trials needed to get an effect of a certain size. This is probably the most common use for power analysis–it tells you how many trials you need to do to **avoid incorrectly rejecting the null hypothesis.**
- To find the power, given an effect size and the number of trials available. This is often useful when you have a **limited budget**, for say, 100 trials, and you want to know if that number of trials is enough to detect an effect.
- To validate your research. Conducting power analysis is simply put–good science.

In Short: The power of a test, is the probability that you will make the right decision ***to support the alternative hypothesis (where your claim is what you are showing to be different from Ho-the no change hypothesis).*** Then, to find the probability for a type II error, we will call β.

P(type II error)=(β)

The Power is 1-β

$P = 1-\beta$

$\beta = 1-P$

.....phew 

Note: The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.

Let's use an illustration. Suppose we had the following hypotheses:

H_o: $\mu=30$

H_a: $\mu<30$

with power = 0.75. This means that there is a 75% chance of making the right conclusion that H_a is true when I reject the H_o.

Example: Let's refer back to the potato problem. Suppose we add the information to the problem that the power calculation for that was 0.732.

$\alpha=.1$

A) What is the probability for a **Type I** error?

P(Type I error) = .1

B) What does the power value mean?

with power value of .732, it means we have a 73.2% chance of correctly rejecting H_0, supporting the claim that $p>.08$

C) What is the probability of a Type II error? Address what this means in relation to the problem. P(Type II error) = 1 - .732 = .268

D) How can we increase the power?

→ increase sample size

→ increase α

EXAMPLE: Suppose the manager at a fast food restaurant wants to change some aspects of his study about the proportion of drive-thru customers who have to wait at least two minutes to recieve their food after placing their order.

Initially he test the following:

Ho: p=0.63

Ha: p<0.63

with p being the tru proprortion of drive-thru customers who have to wait at least 2 minutes to get their food after initial order. He used a significance level of 0.10.

$\alpha = .1$

$P(\text{Type I error}) = .1$

A) What would a type 1 error look like for this and what is the probability of making a Type I error?

Type I error would occur if we rejected null and should not have rejected it. This would falsely support claim that $p < .63$

B) The initial power value was 0.970. Explain what that means.

with power = .970, it means there is a 97% chance that we will correctly Reject H_0 ($p = .63$)

C) What would a type II error be in this case?

A type II error would occur if we FTR H_0 that $p = .63$ when we should have Rejected it!

D) What is the probability of a type II error.

$P(\text{Type II error}) = 1 - .97 = .03$

Example: You read that a statistical test at a significance level of $\alpha=.05$ has power of 0.78. What are the probabilities of type I and type II errors?

Example: You manufacture and sell a liquid product whose conductivity is supposed to be 5. You plan to make 6 measurements of the conductivity of each lot of product. If the product meets specifications, the mean of many measurements will be 5. You will therefor test:

Ho: $\mu=5$

Ha: $\mu\neq5$

If the true conductivity is 5.1 the liquid is not suitable for the intended use. You learn that the power of your test at a 5% significance level against the alternative $\mu=5.1$ is 0.23.

$\alpha=.05$

A) Explain in simple language what power = 0.23 means.

you have a 23% chance of correctly rejecting Ho & if $\mu=5$ and supporting $\mu\neq5$ (conductivity $\neq5$)

B) You could get higher power against the same alternative with the same alpha by changing the number of measurements. Should you make more measurements or fewer to increase power?

↑ sample size ⇒ decreases variability in $\bar{y}$

C) If you decide to use $\alpha=.10$ with no other changes in the test, will the power increase or decrease? Justify your answer. power will increase

5

D)If you adjust your alternative to 5.2, will that increase or decrease the power? Explain.

increase power B/C Diff is greater w/ 5.2 than 5.1

Complete these for HOMEWORK:

1. You are reviewing a research proposal that includes the sample size justification. A careful reading of that section includes that the power is 20% for detecting an effect that most people would consider important. Write a short explanation as to what that means and make a recommendation on whether the study should be done.

2. A one-sided test of the null hypothesis $\mu=50$ versus the alternative $\mu=70$ has power equal to 0.5. Will the power for the alternative $\mu=80$ increase or decrease the power. Explain.

What Affects Power?

There are four things that primarily affect the power of a test of significance. They are:

1.The significance level α of the test. If all other things are held constant, then as α increases, so does the power of the test. This is because a larger α means a larger rejection region for the test and thus a greater probability of rejecting the null hypothesis. That translates to a more powerful test. The price of this increased power is that as α goes up, so does the probability of a Type I error should the null hypothesis in fact be true.

2. The sample size n. As n increases, so does the power of the significance test. This is because a larger sample size narrows the distribution of the test statistic. The hypothesized distribution of the test statistic and the true distribution of the test statistic (should the null hypothesis in fact be false) become more distinct from one another as they become narrower, so it becomes easier to tell whether the observed statistic comes from one distribution or the other. The price paid for this increase in power is the higher cost in time and resources required for collecting more data. There is usually a sort of "point of diminishing returns" up to which it is worth the cost of the data to gain more power, but beyond which the extra power is not worth the price.

3. The inherent variability in the measured response variable. As the variability increases, the power of the test of significance decreases. One way to think of this is that a test of significance is like trying to detect the presence of a "signal," such as the effect of a treatment, and the inherent variability in the response variable is "noise" that will drown out the signal if it is too great. Researchers can't completely control the variability in the response variable, but they can sometimes reduce it through especially careful data collection and conscientiously uniform handling of experimental units or subjects. The design of a study may also reduce unexplained variability, and one primary reason for choosing such a design is that it allows for increased power without necessarily having exorbitantly costly sample sizes. For example, a matched-pairs design usually reduces unexplained variability by "subtracting out" some of the variability that individual subjects bring to a study. Researchers may do a preliminary study before conducting a full-blown study intended for publication. There are several reasons for this, but one of the more important ones is so researchers can assess the inherent variability within the populations they are studying. An estimate of that variability allows them to determine the sample size they will require for a future test having a desired power. A test lacking statistical power could easily result in a costly study that produces no significant findings.

4. The difference between the hypothesized value of a parameter and its true value. This is sometimes called the "magnitude of the effect" in the case when the parameter of interest is the difference between parameter values (say, means) for two treatment groups. The larger the effect, the more powerful the test is. This is because when the effect is large, the true distribution of the test statistic is far from its hypothesized distribution, so the two distributions are distinct, and it's easy to tell which one an observation came from. The intuitive idea is simply that it's easier to detect a large effect than a small one. This principle has two consequences that students should understand, and that are essentially two sides of the same coin. On the one hand, it's important to understand that a subtle but important effect (say, a modest increase in the life-saving ability of a hypertension treatment) may be demonstrable but could require a powerful test with a large sample size to produce statistical significance. On the other hand, a small, unimportant effect may be demonstrated with a high degree of statistical significance if the sample size is large enough. Because of this, too much power can almost be a bad thing, at least so long as many people continue to misunderstand the meaning of statistical significance. For your students to appreciate this aspect of power, they must understand that statistical significance is a measure of the strength of evidence of the presence of an effect. It is not a measure of the magnitude of the effect. For that, statisticians would construct a confidence interval.

More on EFFECT:

What is effect size?

When a difference is statistically significant, it does not necessarily mean that it is big, important, or helpful in decision-making. It simply means you can be confident that there is a difference. Let's say, for example, that you evaluate the effect of an EE activity on student knowledge using pre and posttests. The mean score on the pretest was 83 out of 100 while the mean score on the posttest was 84. Although you find that the difference in scores is statistically significant (because of a large sample size), the difference is very slight, suggesting that the program did not lead to a meaningful increase in student knowledge.

To know if an observed difference is not only statistically significant but also important or meaningful, you will need to calculate its effect size. Rather than reporting the difference in terms of, for example, the number of points earned on a test or the number of pounds of recycling collected, effect size is standardized. In other words, all effect sizes are calculated on a common scale -- which allows you to compare the effectiveness of different programs on the same outcome.

How do I calculate effect size?

There are different ways to calculate effect size depending on the evaluation design you use. Generally, effect size is calculated by taking the difference between the two groups (e.g., the mean of treatment group minus the mean of the control group) and dividing it by the standard deviation of one of the groups. For example, in an evaluation with a treatment group and control group, effect size is the difference in means between the two groups divided by the standard deviation of the control group.

$$\frac{\text{mean of treatment group} - \text{mean of control group}}{\text{standard deviation of control group}}$$

To interpret the resulting number, most social scientists use this general guide developed by Cohen:

< 0.1 = trivial effect

0.1 - 0.3 = small effect

0.3 - 0.5 = moderate effect

> 0.5 = large difference effect

How do I estimate effect size for calculating power?

Because effect size can only be calculated after you collect data from program participants, you will have to use an estimate for the power analysis. Common practice is to use a value of 0.5 as it indicates a moderate to large difference.