Activity #16: Hypothesis Testing (single sample)

In the previous activity, we used parametric, bootstrap, and Bayesian methods to construct confidence intervals for population means (μ,with known and unknown σ) and population proportions.

Another method of statistical inference is *null hypothesis significance testing* (NHST). We've already done some NHST:

- Activity #1: Buzz and Doris (Can dolphins communicate?) and the bank supervisor sexism study
- Assignment #1: Do kissing couples lean to the right? Did Phil Mickelson putt worse than average?
- Activity #3: Are yawns contagious? Should I shoot free throws underhanded?
- Assignment #3: Is dolphin therapy effective in reducing depression? Nurse Gilbert.
- Activity #7: Do homing pigeons really know how to return home? Do dogs resemble their owners?

In those examples, we used randomization methods (activity #1), the hypergeometric distribution (activity #3), or the binomial distribution (activity #7) to estimate p-values from data involving proportions.

In this activity, we'll begin to formalize hypothesis testing and apply it to population means. Resist the temptation to try to turn NHST into a step-by-step process. Instead, focus on the logic and concepts behind NHST.

Also, I want you to be aware that not everyone thinks NHST is worthwhile. We'll discuss many of the limitations of NHST in class, but check out the following links if you want to know more about the problems with NHST (and some possible alternatives): http://warnercnr.colostate.edu/~gwhite/fw663/testing.pdf
http://www.indiana.edu/~kruschke/AnOpenLetter.htm
http://www.andrews.edu/~rbailey/Chapter%20two/7217331.pdf

Recall from Activity #1 the logic behind hypothesis testing:

- We state two competing hypotheses (the null and alternative hypotheses) and collect some data
- Assuming the null hypothesis is true, we estimate the likelihood of observing our data (or more extreme data)
- If the likelihood (p-value) is smaller than a predetermined level), we have evidence against our null hypothesis

Another way of thinking about the logic is: We state a prior belief, collect some data, and update our belief.

In doing NHST, we'll discuss some experimental design issues (sampling methods, random assignment) and introduce, review, or formalize some concepts, such as:

- α (or Type I) error
- β (or Type II) error
- Power
- Sampling distributions
- Sample statistics vs. population parameters
- p-values
- Effect sizes

We'll also learn how to conduct NHST using randomization-based methods, theory-based methods (t-tests), and confidence intervals (theory-based and bootstrap-based).

1. Before we begin, explain how bootstrap methods work. If we have a sample of data, how can we use bootstrap methods to construct a 90% confidence interval for a population mean?

Scenario: The vision of St. Ambrose includes being "recognized as a leading Midwestern university…" Suppose we believe a leading university would attract above-average students. To determine if we're reaching our vision, we might compare ACT scores of St. Ambrose students to the average ACT score in Iowa.

In 2013, the average ACT score for the 22,526 students in Iowa who took the ACT was $\mu = 22.3$ with $\sigma = 6$. The ACT is designed to yield scores that follow a normal distribution.

The average ACT score for the 510 freshmen who entered St. Ambrose in 2014 was 22.8.

Sources: http://www.act.org/newsroom/data/2012/states.html ; 2014 MAP-Works dataset

2. For this scenario, identify the following:

Dependent variable: ________________________ Independent variable: __________________________

Parameter of interest: ________________________ Observed statistic (estimator): ___________________

Target population: ___________________________ Sample: ______________________________________

Random assignment? ________________________ Random sampling? _____________________________

3. Write out the null and alternative hypotheses for this scenario. Remember, the null hypothesis (H_0) is the "dull hypothesis" that typically corresponds to a default position.

The alternative hypothesis typically asserts a particular believe about the treatment effect. If we don't know what to expect, we might write out a <u>two-tailed</u> (≠) alternative hypothesis. If we have a prior belief, we may choose to write a <u>one-tailed</u> (> or <) alternative hypothesis.

H_0 : ______________________________________ H_1 : ______________________________________

4. Before we do anything else, take a look at your null hypothesis. What's the probability it's true?

5. As we first did in activity #3, let's evaluate the potential outcomes (and errors we could make) from this study. Fill-in the following table:

	Decision: Retain H_0	**Decision: Reject H_0**
Reality: H_0 is true		
Reality: H_0 is false		

6. Express in plain language, what alpha, beta, and power represent in this scenario. Identify a possible consequence of each error.

alpha: ____

beta: ____

power: ____

7. As we'll see, the researcher has direct control over the probability of making an α-error. We'll also discover the probability of making an α-error is inversely related to the probability of making a β-error.

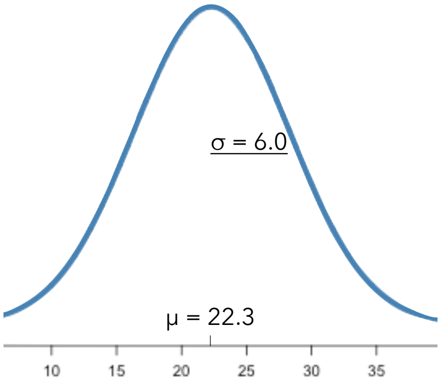
In this scenario, which error do you believe is more costly? Should we set a relatively high or low α-level?

The more costly error is: **alpha** **beta**. We should set α at a relatively **low** **high** level.

Suppose we set α = 0.05. Express what this represents in this scenario.

α = 0.05 represents: ____

8. Under our null hypothesis, we're assuming the ACT scores for our 510 freshmen represent a random sample from a normal population with μ = 22.3 and σ = 6.



The key question is: **If our null hypothesis were true (and the distribution to the left is the population distribution), how likely were we to have observed an average ACT score of 22.8 or higher?**

To estimate this likelihood, we can use randomization-based or theory-based methods. We'll first investigate one randomization-based method that's based on the bootstrap method we used to construct confidence intervals in the previous activity.

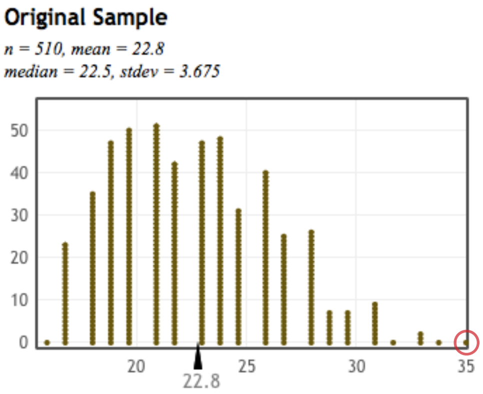
1. Copy the data from: http://www.bradthiessen.com/html5/stats/m300/act.txt
2. Open the applet at: http://lock5stat.com/statkey/randomization_1_quant/randomization_1_quant.html
3. Click the **EDIT DATA** button at the top, paste the data in the box, and click **OK**.
4. You should now see your sample data in the box at the top-right. You can verify the sample mean is 22.8.
5. At the top, we need to change our null hypothesis to be: **μ = 22.3**.
6. Generate one sample, look at that sample at the bottom-right of the screen, and see if you can figure out what the applet is doing. We'll discuss this method in just a minute.
7. Generate at least 10,000 samples and estimate the likelihood of observing a sample mean ≥ 22.8.

Record the estimated p-value here: p = ______________

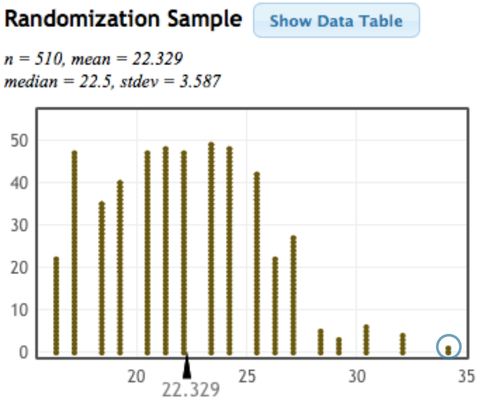
Randomization-based test for a single mean (an explanation of the method)

A. Our null hypothesis states that our data come from a population with a mean of 22.3. Our sample data has a mean of 22.8, so the computer (invisibly) shifts each observation in our sample to the left by 0.5.

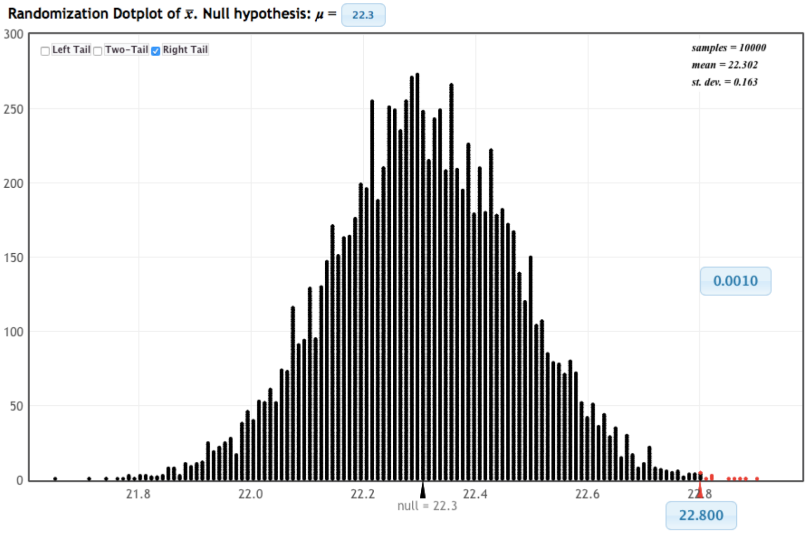
B. Now that our sample data reflect our null hypothesis, the computer selects a bootstrap sample. In this scenario, a bootstrap sample is a sample of n = 510 observations (sampled <u>with</u> replacement).



C. If you look at our sample data (top-right), you'll notice we had one freshman with an ACT score of 35 (circled in red). The distribution for one bootstrap sample is displayed below. Notice (circled in blue) the highest ACT was shifted to the left by 0.5 and was selected twice (because of sampling with replacement). The mean of this bootstrap sample was 22.329.



D. Means from 9,999 more bootstrap samples were then plotted to form a randomization-based sampling distribution (displayed below).

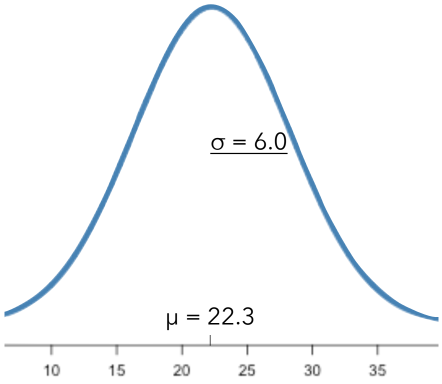


E. To estimate the p-value (likelihood of observing results more extreme than what we observed), the computer simply has to count the proportion of sample means greater than or equal to our observed sample mean of 22.8. In the example displayed to the left, the p-value was estimated to be 0.001.

9. What could we conclude from our p-value?

Now let's go back and try a theory-based method…

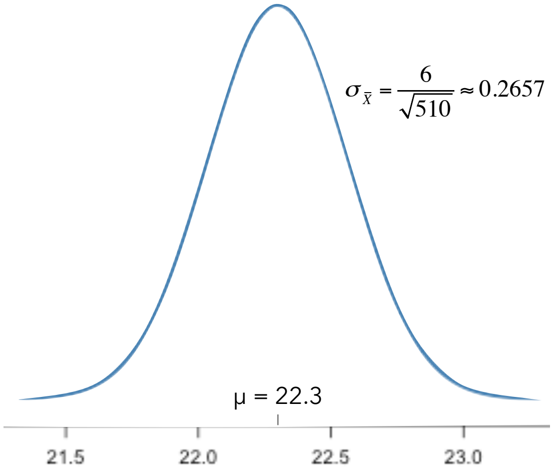
10. As we stated in question #8, under our null hypothesis, we're assuming the ACT scores for our 510 freshmen represent a random sample from a normal population with μ = 22.3 and σ = 6.



We want to generalize beyond our single sample of 510 freshmen in 2014, so we can consider (perhaps incorrectly) that these 510 freshmen are a random sample of all possible SAU freshmen over time. Thus, the average ACT score from these 510 freshmen is just one many different averages we could have observed if we would have chosen a different random sample.

The key question is: **If our null hypothesis were true (and the distribution to the left is the population distribution), how likely were we to have observed an average ACT score of 22.8 or higher?**

Assuming the null hypothesis is true, sketch the sampling distribution of the sample mean ACT scores for 510 SAU freshmen. How do we know the center and std. error of this distribution? Explain what the distribution represents.



Theory-based test for a single mean (an explanation of the method)

A. Our null hypothesis states that our data come from a population with a mean of 22.3 (and a standard deviation of 6). If the null hypothesis is true, our data represent a random sample from this population.

B. The Central Limit Theorem tells us what to expect if we were to repeatedly sample 510 observations from the population and calculate the mean of each sample. We can sketch this expected sampling distribution.

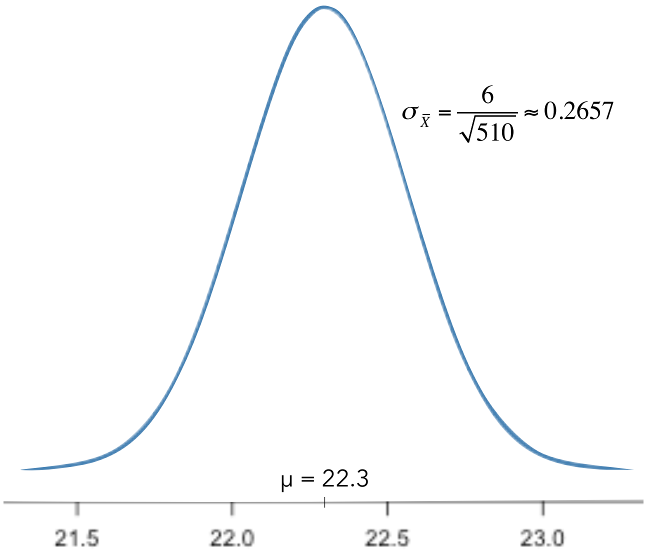
C. To estimate a p-value, we simply need to calculate the probability of observing a sample mean more extreme than what we actually observed. In other words, we need to calculate $P\left(\bar{X} \geq \bar{X}_{\text{observed}}\right) = P\left(\bar{X} \geq 22.8\right)$

D. If that p-value is small, it means our sample mean probably did not come from this sampling distribution. We would then reject this sampling distribution (which was generated under the belief that the null hypothesis was true).

   A large p-value, on the other hand, would mean it was likely our sample data came from this sampling distribution (so we do not have evidence to reject our null hypothesis).

Calculate this p-value and record it here: p-value = ________________. What do we compare this p-value to?

Here, again, is the distribution of average ACT scores we would get if we repeatedly took samples of size n=510.



11. Earlier, we decided to set α = 0.05. In other words, we decided we would reject the sampling distribution pictured above if the mean ACT score we calculate from our sample of 510 freshmen falls in the top 5% of the distribution.

    Let's locate the top 5% of this distribution. The critical value is the ACT score that cuts-off the top 5% of our sampling distribution. Find this value, label it on the above distribution, and shade-in the critical (rejection) region.

    We know the z-score that cuts-off 5% of the distribution would be z = _______________.

    We can convert that z-score to an ACT score: ACT = ____.

    Or we can calculate the critical value directly via a normal distribution applet:
    http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#normal

    Critical value: ____

12. Now, draw an arrow on the distribution representing your (actual) observed sample mean of 22.8. From this observed sample average, what do you conclude about the null hypothesis?

    Decision regarding the null hypothesis: ____

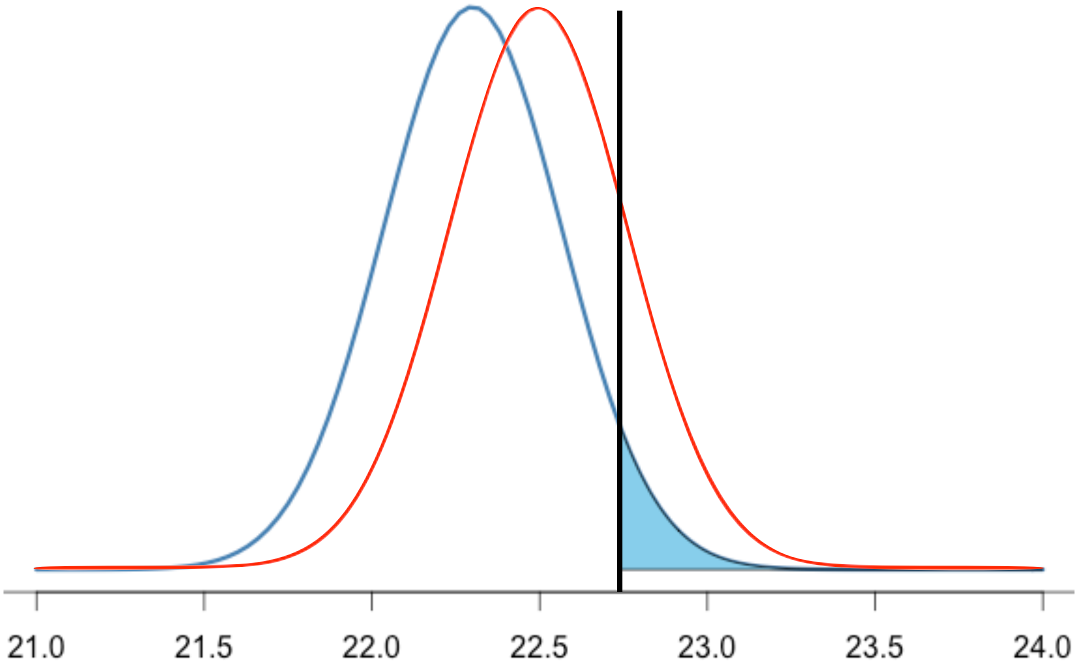
13. Using the above distribution, visualize what the p-value represents. How do we interpret this p-value?

    p-value: ____

14. Once again, the sampling distribution (assuming a true null hypothesis) is displayed below (in blue). The critical region has been shaded-in. Remember that the size of this critical region was entirely up to us (when we chose our alpha-level).

    Suppose the null hypothesis was, in fact, false. Suppose that the population mean ACT score for St. Ambrose freshmen is actually 22.5. If this is true, our sampling distribution should have been the red one pictured below.

    Shade and label α, β, and power in the two distributions displayed below. Then, calculate the probability of each.



Alpha = ____

Beta = ____

Power = ____

You've just conducted a hypothesis test of a single population mean using the z-distribution. The process was:

1. State two competing hypotheses about a population parameter
2. Consider potential decision errors and choose an appropriate level of significance (α)
3. Determine the statistic you will calculate from your sample to estimate the parameter of interest
4. Sketch the sampling distribution of that statistic
    a. Use theory- or randomization-based methods
    b. Assume the null hypothesis is true
5. Estimate a p-value (the probability of observing data as or more extreme given a true null hypothesis)
6. Make a decision to reject or retain the null hypothesis

15. In the previous example, we used a normal distribution because we assumed we knew the population standard deviation. What do we do if we do not know the population variance? I'm glad you asked…

Scenario: In the 1980s, some companies experimented with flextime, allowing employees to choose their work schedules within broad limits set by management. It was believed that flextime would reduce absenteeism.

Suppose a company is willing to try flextime for one year to see if it reduces absenteeism. Based on historical data, this company knows its employees have averaged 6.3 days absent. The company chose 100 employees at random to try flextime for one year. During that year, those 100 employees averaged 5.5 days absent with a standard deviation of 2.9 days.

16. State the competing hypotheses.

H_0 : ____ H_1 : ____

17. Express the consequences of both Type I and Type II errors. Set an appropriate alpha-level.

alpha: ____

beta: ____

18. According to the scenario, the distribution of "days absent" for the entire company has μ=6.3. We are not told the shape or the standard deviation of that distribution. We know even less about the distribution of days absent for flextime employees (we do not know the mean, shape, or standard deviation!).

For our sample of 100 employees, the average was 5.5 with a standard deviation of 2.9. What distribution does this sample average come from? In other words, if we could repeatedly sample 100 employees at random, have them try flextime for a year, and calculate their average days absent, what would the distribution of those averages look like? What would be the mean and standard error of that distribution? What assumptions are you making? Sketch and label the sampling distribution below

.

19. Suppose we set α=0.05. Find the critical value(s) corresponding to this α-level and shade-in the appropriate critical region

Our sampling distribution approximates a _____ - distribution with ____________________ degrees of freedom.

The value that cuts-off 5% of this distribution would be t = __________________.
http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#t

We can convert that to a number of days absent = ____

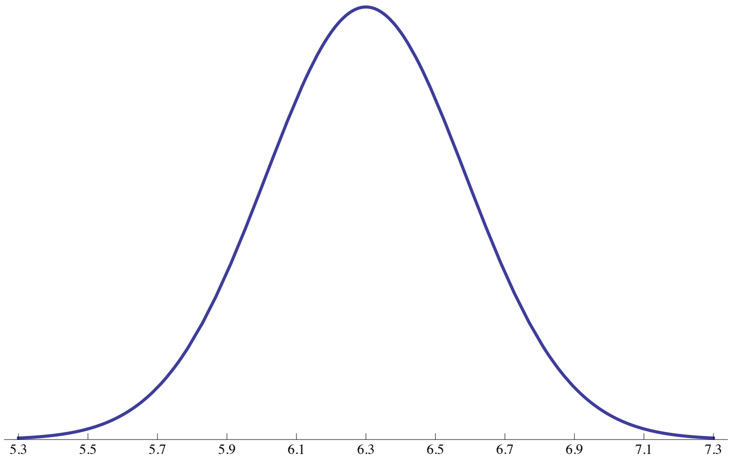
20. Locate the observed sample mean of 5.5 on your sampling distribution. Convert this to a t-score.

Observed sample mean = 5.5. Observed t-score = ____.

21. Use the online t-distribution applet to estimate the p-value from this scenario. What can you conclude? Does this p-value represent the probability that the null hypothesis is true?

p-value = ___________________________.

22. The sampling distribution of our sample averages (assuming a true null hypothesis) has been drawn below. Once again, draw an arrow to locate our observed mean of 5.5. Find the critical values for α=0.001, α=0.01, α=0.05, and α=0.10. What happens as we increase α? Does our conclusion about flextime change?



23. Circle the correct word:

In a hypothesis test, we **reject** **retain** the null hypothesis if our **p-value < α**.

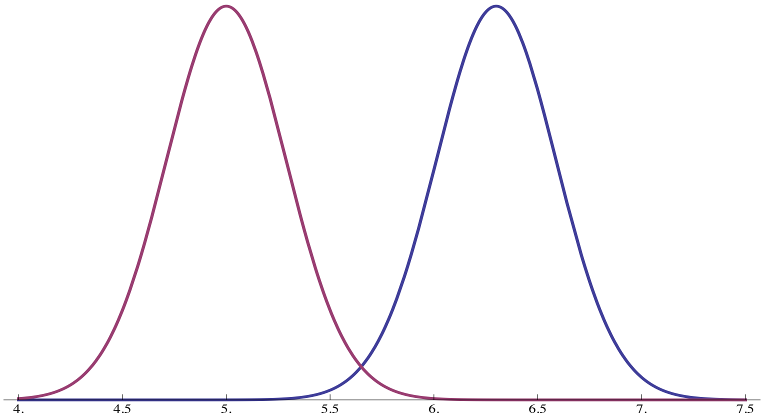
24. True or False:

a. A p-value tells us the probability that the null hypothesis is true.

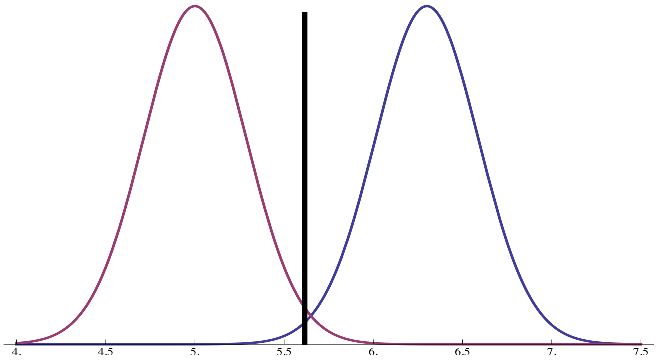
b. A p-value tells us the probability that the null hypothesis is false.

c. A smaller p-value means the observed sample mean was farther away from the (null) hypothesized mean.

25. Suppose, in reality, flextime actually reduces absenteeism by 1.3 days (in other words, if we could have sampled our entire population, we would have found flextime employees at this company only miss 5 days each year). Given this reality (where the null hypothesis is no longer true), calculate the power of this study. Use the curves below for help.



26. Let's examine this concept of power in more depth. Below, I've drawn sampling distributions for both the null and alternative hypotheses. Using α=0.05, the critical value is also displayed. Shade-in α, β, and power..



a. If we choose a larger value for α, power **increases** **decreases** **does not change**

b. If we increase our sample size, power **increases** **decreases** **does not change**