<<Lecture8_Student.pdf>>

Stat 324 – Introduction to Statistics for Engineers

LECTURE 9: HYPOTHESIS TESTING DEFINITIONS AND A FIRST APPLICATION; 5.1, 5.4, 5.5, 5.6 OF OTT AND LONGNECKER.

Hypothesis Testing Big Idea

While confidence intervals are used to estimate a population parameter, hypothesis tests assess the evidence provided by data about some claim concerning a parameter.

E.g. A battery maker claims that its D battery lifetime has $\mu = 40$ and $\sigma = 5$ hours. Suppose a random sample of 100 batteries is selected.

a. If the company's claim is true, what is $P(\overline{X} \leq 36.7)$? Based on the makers claim, is seeing an average life time of 36.7 in a random sample of 100 unusually short? If $\overline{x} = 36.7$, is the claim plausible?

$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

$P(\overline{X} \leq 36.7) = P\left(Z \leq \frac{36.7-40}{5/\sqrt{100}}\right) = P(Z \leq -6.6) < .0003$

possible? yes but not likely

b. If the company's claim is true, what is $P(\overline{X} \leq 39.8)$? Based on the makers claim, is seeing an average life time of 39.8 in a random sample of 100 unusually short? If $\overline{x} = 39.8$, is the claim plausible?

possible and likely

$P(\overline{X} \leq 39.8) = P\left(Z \leq \frac{39.8}{5/\sqrt{100}}\right) = P(Z \leq -0.4) = 0.3446$

35% chance

Hypothesis Testing Vocabulary

A ***hypothesis test*** checks whether our observed sample data is consistent with a proposed value of a _parameter_.

A hypothesis test considers:

$\boldsymbol{H_0}$***: the null hypothesis,*** which asserts "any effect indicated by the sample is merely due to _chance_, and is _not an effect_ in the population "

*****$\boldsymbol{H_0}$ is _assumed true_ unless sufficient evidence to the contrary.

***** Often specifies a _single value_ for a parameter

****e.g.*** $\boldsymbol{\mu = 31}$

And the

$\boldsymbol{H_A}$*: the* _alternative_ ***hypothesis,*** which rejects $\boldsymbol{H_0}$***,*** saying the "effect observed in the sample is present in the _population_"

* Usually what we/scientists would like to show

* Often specifies a range of values for a parameter

* e.g. $\boldsymbol{\mu > 31}, \boldsymbol{\mu < 31}, \ \boldsymbol{\mu \neq 31}$

Hypothesis Testing Big Idea

After a hypothesis about a parameter (or relationship) is made, a **RV** that reflects the parameter or relationship, called a _test statistic_ is considered.

The specific formula for the test statistic will depend on the parameter/relationship being tested and the nature of the _sampling_.

The realization of the test statistic (relative to its distribution under the null) is evidence for deciding between $\boldsymbol{H_0}$ ***and*** $\boldsymbol{H_A}$.

"Likely" Values for Test Statistic, Assuming the null is true

Fail to reject H0

Test Statistic from sample

If the test statistic offers **insufficient** evidence against the null, we **fail to reject the** $\boldsymbol{H_0}$**.**

Notice we are not "accepting" the null.

"Unlikely" Values for Test Statistic, Assuming the null is true (Rejection Region)

Reject H0

Test Statistic from sample

If the observed test statistic is **unlikely** under the assumption of $\boldsymbol{H_0}$***,*** we say it falls in the **rejection region** and we **reject** the null.

Hypothesis Testing Ex:

Consider a fire alarm. The natural choices for $\boldsymbol{H_0}$ and $\boldsymbol{H_A}$ are:

$\boldsymbol{H_0}$: There is no fire $\boldsymbol{H_A}$: there is a fire

Possible test statistics might be concentration of smoke particles (S), temperature in room (T).

Higher values of S or T would be stronger evidence against the null.

Suppose research indicates that when there is no fire, temperatures stay below 110 F.
Then our rejection region would be T>110.

When the fire alarm collects data, if it measures room temperature:
t= 70, the test statistic is not in the rejection region, so we would fail to reject the null
*Notice, this doesn't necessarily mean there is no fire, just that we don't have enough evidence of a fire.

t=200, the test statistic is in the rejection region, so we would reject null and say evidence suggests that there is a fire (the alternative)

Hypothesis Testing Battery Ex Revisited:

Customer believes battery mean life time is too short…company claims battery lifetimes has $\mu = 40$ and $\sigma = 5$ hours
They want to find evidence for
alternative: $H_A: \mu < 40$ but start with the null assumption that H_o: $\mu = 40$

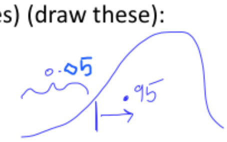
And determine "reasonable" values that estimator for population parameter ($\bar{X}$) could take based on its sampling distribution.

Assume: $X_1, X_2, .. X_{100}$ is a SRS from $N(\mu, \sigma)$ or n is large (CLT applies) and σ is known.

Then, under the null, $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(40, \frac{5}{\sqrt{100}}\right) = (40, 0.5)$

rejection region
Assuming the null is true, we can determine boundaries for our RR (critical values) (draw these):

95% of the sample means will fall above/5% will fall below: $\bar{X} = -1.645\,(0.5) + 40$
$= 39.178$



$\bar{X} = 39.178$

$Z_{.95} = Z_{.05} = -1.645$

$-1.645 = Z = \frac{\text{obs} - 40}{0.5}$

$\bar{x}$ values < 39.178
Z values < -1.645
are the most extreme 5%

Z Tests for means when population sd σ is known :

are the most extreme 5%

Z Tests for means when population sd $\boldsymbol{\sigma}$ is known :

When taking a random sample from a Normal population, or a large enough sample that we are confident the CLT ensures $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, and we know σ , we can use a **Z test** when interested in the population mean μ

Test Statistic : $\boxed{Z = \frac{\bar{X}-\mu_o}{\sigma/\sqrt{n}}}$

The **rejection region** tells us improbable values for realized z or $\bar{x}$ under the null and is one- or two-sided depending on the alternative hypothesis. Specifically,

z_α $H_a: \mu > \mu_o$ requires $RR: Z \geq z_\alpha \;\; \Rightarrow \;\; \bar{X} \geq z_\alpha * \frac{\sigma}{\sqrt{n}} + \mu_o$

$z_{1-\alpha}$ $H_a: \mu < \mu_o$ requires $RR: Z \leq z_{1-\alpha} \;\; \Rightarrow \bar{X} \leq z_{1-\alpha} \frac{\sigma}{\sqrt{n}} + \mu_0$

one sided

Batteries 1 sided Ex:
$\alpha = .05$
$z_{.95} = -1.645$ or
$\bar{X}$ =39.178

$\alpha/2$ $-z_{\frac{\alpha}{2}}$ $z_{\frac{\alpha}{2}}$ $\alpha/2$ $H_a: \mu \neq \mu_o$ requires $RR: Z \leq -z_{\frac{\alpha}{2}}$ *or* $Z \geq z_{\frac{\alpha}{2}} \Rightarrow \bar{X} \leq -z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \mu_0$

$\bar{X} \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \mu_0$

$\frac{\alpha}{2} \sqrt{n} + \mu_0$

Hypothesis Testing Procedure

Plan a Study:

1. Develop a null and alternative for a population parameter.

2. Choose a size for our rejection region, significance level α– what level of evidence do we want to require to reject H_o ?

3. Determine what effect size would be considered important to detect

4. Find an appropriate sample size so test has desired **power** to reject null when “truth” is important effect size.

Collect data according to study design

Analyze the Sample Data

Hypothesis Testing Errors and setting α :

Even when we do computations perfectly, because of sampling variability, we will sometimes draw the **incorrect** conclusion (not identify what is true in the population) based on the sample we see.

Error 1: There is no fire, but alarm thinks there is _enough_ **evidence** of fire and goes off

Error 2: There is a fire, but alarm thinks there is _not enough_ **evidence** and doesn't go off.

$\boldsymbol{H_0}$: There is no fire

$\boldsymbol{H_A}$: There is a fire

		Statistical Decision		
Real Truth		High Evidence so Reject H_0	Low Evidence so Fail to Reject H_0	
	H_0 True	Type_1_Error	No Error	no fire
	H_0 False	No Error	Type _2_Error	fire

Hypothesis Testing Errors and setting α :

Controlling Errors by defining "**enough evidence to reject** H_0 **"** with α. (significance level)

$$\mathrm{P}(Type\ I\ Error) = P(Reject\ H_o | H_o\ True) = \alpha$$
$$= P(Test\ statistic\ falls\ in\ rejection\ region | H_o\ True)$$

* Requires we understand distribution of test statistic under _null_ assumption

We want to limit Type 1 errors, so ideally α is _small_ (but there is a trade off).

		Statistical Decision	
Real Truth		High Evidence so Reject H_0	Low Evidence so Fail to Reject H_0
	H_0 True	Type I Error/α	No Error
	H_0 False	No Error	Type II Error

Court Example:
H_o: defendant innocent
H_a: defendant guilty find them guilty when actually innocent

Battery Example:
$H_0: \mu = 40$
$H_A: \mu < 40$

Hypothesis Testing Errors:

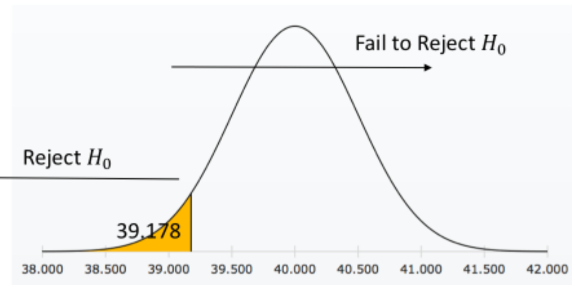
Controlling Errors by defining "**enough evidence to reject H_0** " with α

$$\text{P}(Type\ I\ Error) = P(Reject\ H_o | H_o\ True) = \alpha$$

Battery example:

$H_0: \mu = 40 \qquad H_A: \mu < 40$

If H_0 is true, a decision to reject H_0, based on the data is a Type _1_ error



$\alpha = .05$

Critical values:

Z test statistic value $< z_{.95} = -1.645$, or obs $\bar{x}$ < _39.178_ with a sample size of 100

will result in rejecting the null at $\alpha = .05$ (and a type 1 error if in fact H_0 is true)

Hypothesis Testing Errors:

Controlling Errors by defining "**enough evidence**" with α and β

$$\text{P}(Type\ II\ Error) = P(Fail\ to\ Reject\ H_o | H_o\ False) = \beta_a$$
$$= P(Test\ statistic\ does\ not\ fall\ in\ rejection\ region | H_o\ False)$$

* Requires we consider one value of parameter where H_o *False* to calculate probability

We want to limit Type II errors, so ideally β_a is small (but there is a trade off).

		Statistical Decision	
Real Truth		High Evidence so Reject H_0	Low Evidence so Fail to Reject H_0
	H_0 True	Type I Error/α	No Error
	H_0 False	No Error	Type II Error/ β_a

Court Example:

H_o: defendant innocent not enough evidence when they

H_a: defendant guilty

Battery Example:

$H_0: \mu = 40$

$H_A: \mu < 40$

Hypothesis Testing Errors:

$$P(Type\ II\ Error) = P(Fail\ to\ Reject\ H_o | H_o\ False\ and\ \mu_A\ true) = \beta_a$$

RR: $\bar{x} < 39.176$

$z < -1.645$

Battery example:

$H_0: \mu = 40 \qquad H_A: \mu < 40$

If $\mu_A = 39\ true$ (and $\underline{H_0}$ is false),

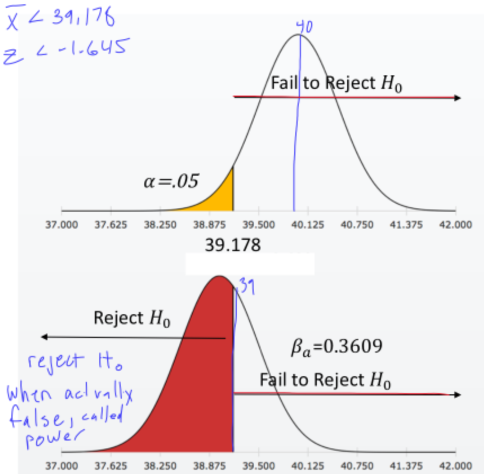
can only test for specific alternative

$\beta_{U_a=39}$ = P(Fail to Reject H_0 | $\mu_A = 39$)=

$P(\bar{x} > 39.176) = P\left(z > \frac{39.178 - 39}{5/\sqrt{100}}\right)$

$P(z > 0.356) = 1 - P(z \leq 0.356)$

$= 0.3609$



reject H_0 when actually false, called power

$\mu_A = 39$

36% chance to fail to reject when should

*notice we use the critical value[s] with μ_A distribution to calculate ~~power~~ type 2 error of z, $\bar{x}$ from RR

Power of a Test and Errors

The ***power*** of a test is $P(Reject\ H_o \mid H_o\ false\ and\ \mu_A\ true)$ = $\underline{1 - \beta_\alpha}$

(As with β_a, power can only be computed for a single value of the alternative)

Battery example:

$H_0: \mu = 40 \qquad \boxed{H_A: \mu < 40}$

If $\mu_A = 39\ true$ (and H_0 is false),

usually go for 0.80

80% power

Power= $\underline{1 - 0.3609 = 0.639}$

Power to reject null $H_0: \mu = 40$ with alternative $\mu_A = 39$ sample size n =100 and $\sigma = 5$ and $\alpha = 0.5$

$\mu_A = 39$

- Fire example: power is the probability that alarm goes off when fire
- Crime example: power is the probability that find defendent guilty when actually guilty

Power of a Test and Errors

For a fixed sample size, and specific $\mu_A = value$.

We want small α and small β /large power, unfortunately there is a trade off

- if we make our rejection region smaller by decreasing α and require stronger evidence to reject null
- then we increase β: probability of not having enough evidence to reject null and decrease power

When deciding on an appropriate rejection region/level of evidence before rejecting H_o we need to balance these concerns & decide which error is more important to control.

- In fire example, high power to sound alarm when there is a fire is more important $\alpha = 10-20\%$
- In crime example, low probability of sending innocent person to jail more important $\alpha = 1-5\%$

For fixed α to decrease β (increase power) we can increase our sample size.

Increasing Power (decreasing Type 1 Error Rate)

- http://digitalfirst.bfwpub.com/stats_applet/stats_applet_9_power.html
- To increase power,
  - Look for a larger effect size: $|\mu_0 - \mu_A|$
  - Increase the type I error rate, α (which means require less evidence to reject H_o
  - Increase the sample size, n
  - Decrease the population standard deviation σ

Power and sample size

To find a sample size n required to achieve power $1-\beta$ to reject H_o at level α when a particular H_a is true for a test of $H_0: \mu = \mu_o$ vs $H_a: \mu \neq \mu_o$, use $n \approx \left(\frac{\sigma(z_{\frac{\alpha}{2}}+z_\beta)}{\mu_o - \mu_a}\right)^2$

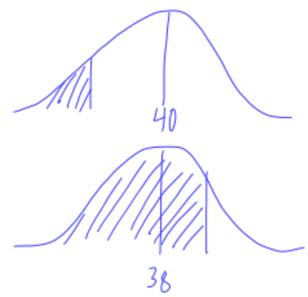
if the $\mu > \mu_0$ or $\mu < \mu_0$ $\quad z_\alpha$

For the battery example $H_0: \mu = 40 \quad H_A: \mu < 40$, consider $\sigma = 5$, and we seek the sample size, n required to have power 0.8 to reject H_o at level $\alpha = 0.05$ when the true mean is $\mu_a = 38$.
n=?

z = .8 in table

Since $\beta =$ 0.20, $z_{0.2} =$ 0.845, and $z_{0.05} =$ 1.645

so $n = \left(\frac{5(1.645+0.845)}{40-38}\right)^2 = 38.75$
$= 39$



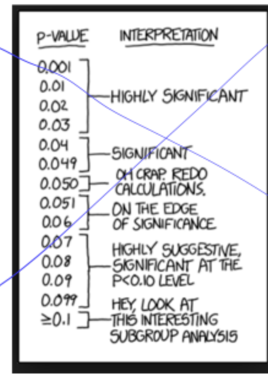
Hypothesis Testing Procedure

Plan a Study:
1. Develop a null and alternative for a population parameter.
2. Choose a significance level α– what level of evidence do we want to require to reject H_o
3. Determine what effect size would be considered important to detect
4. Find an appropriate sample size so test has desired **power** to reject null when "truth" is important effect size.

Collect data according to study design

Analyze the Sample Data
1. Calculate the statistic on sample data
2. Compare calculated statistic to critical value or
   - Calculate the p-value: probability of observing that statistic (under assumption null hypothesis is true)
3. If sample statistic is more extreme than critical value or p-value < α (significance level), reject H_o
4. Make conclusions in context of question

Hypothesis Testing Vocabulary



The ***p-value*** is defined to be the probability of a test statistic realizing to a value __as or more extreme__ than the one actually observed, under the assumption of the null hypothesis being true.

Smaller p-values indicate relatively __more__ evidence against the null hypothesis (evidence for the alternative).

The ***p-value*** required to cause a rejection of the null is called the **significance level (α)** of the test.

Typical significance levels of α are __0.05, 0.1, 0.01__. If p< α, we "Reject Null"

Setting a ***lower*** significance level α, requires __stronger evidence__ to reject the null which

*Lowers probability of __type 1 error__ *Increases probability of __type 2 error__

It is best practice to report the actual calculated p value instead of just saying "Reject" or "Fail to Reject" Null, as different readers may choose a different level of significance [evidence] required

Hypothesis Testing Battery Ex Revisited:

…customer …believes the mean life time is too short. They know the company claims battery lifetimes has $\mu = 40$ and $\sigma = 5$ hours.

They want to find evidence for:
alternative: $H_A: \mu < 40$ but start with the null assumption that $H_o: \mu = 40$

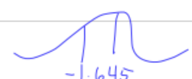
They choose a significance level $\alpha = 0.05$ (because ok with rejecting the null incorrectly 5% of time)

We collect an SRS of 100 lightbulbs and find a sample mean of 39.8
Calculate the p value: under $H_o: \mu = 40$

$p\ value$: $P(\bar{x} \leq 39.8) = P\left(Z \leq \frac{39.8 - 40}{5/\sqrt{100}}\right) = P(Z \leq -0.4) = 0.3446$

This p value $0.34 > .05$ so no evidence against null; evidence suggests insufficient.

* notice same conclusion we got by comparing our computed test statistic $z = -0.4$ not in our rejection region $RR: z < -1.645$



Hypothesis Testing Ex 2:

A powdered medicine is supposed to have a mean particle diameter of μ = 15 μm. Its manufacturing process is known to produce a mean particle diameter that occasionally drifts, while the standard deviation of diameters stays steady around 1.8 μm. A simple random sample of 87 particles had a mean diameter of 15.4 μm. Is this strong evidence that the powder does not meet its specification? (the manufacturing process needs to be recalibrated.)

Hypotheses: null : H_o: $\mu = 15$ alternative: H_A: $\mu \neq 15$

Type 1 Error in context: Reject the null when the null is true.
Calibrate machine when working fine

Type 2 Error in context: Fail to reject null when its false.
machine is fine when it needs to be recalibrated

Significance level $\alpha =$ Higher alpha because Type 2 is worse so ~ 0.05

Hypothesis Testing Ex 2:

medicine is supposed to have a mean diameter of μ = 15 μm. … standard deviation of diameters stays steady around 1.8 μm. A simple random sample of 87 particles had a mean diameter of 15.4 μm. Is this strong evidence that the powder does not meet its specification?

Hypotheses: null : $H_o: \mu = 15$ alternative: $H_A: \mu \neq 15$ Significance level $\alpha = 0.05$

Assumptions:
Suppose $X_1, X_2, .. X_{87}$ is a SRS from $N(\mu, \sigma)$ or n is large (so CLT applies) and σ is known.

Then, under the null, $\bar{X} \sim N\left(15, \frac{1.8}{\sqrt{87}}\right)$

We collect an SRS of 87 particles and find a sample mean of 15.4.
Calculate the p value: (draw this)

$p\ value$= $2 * P(\bar{x} \geq 15.4) = 2 * P\left(z \geq \frac{15.4-15}{\frac{1.8}{\sqrt{87}}}\right) = 2 * (0.0191)$
$= 0.0382$

This p value 0.0382 so moderate evidence against null; evidence suggests true mean may be different from $\mu = 15$

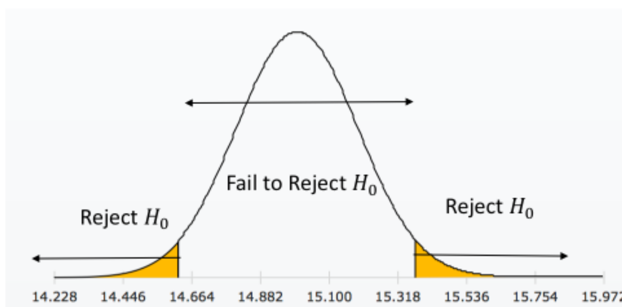
Hypothesis Testing Errors:

What if we continue to run tests at the α=*0.05 level with n=87?*

$$P(Type\ I\ Error) = P(Reject\ H_o | H_o\ True) = \alpha$$

Medicine example (if 2-sided alternative):
$H_0: \mu = 15$ $H_A: \mu \neq 15$

If H_0 is true, a decision to reject H_0, based on the data is a type I error



$\alpha/2$ $\alpha/2$=0.25

Critical Values
below $z_{.975} = -1.96$ or above $z_{.025} = 1.96$, *so*

Critical Values: $\bar{X} <$ $-1.96 * \frac{1.8}{\sqrt{87}} + 15 = 14.62$ or $\bar{X} >$ $1.96 * \frac{1.8}{\sqrt{87}} + 15 = 15.376$

with sample size n=87, $\sigma = 1.8$ will result in rejecting the null (a type 1 error if in fact H_0 is true)

Hypothesis Testing Ex 2:

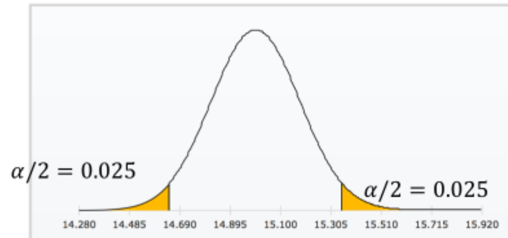
A powdered medicine is supposed to have a mean particle diameter of μ = 15 μm. .. standard deviation of diameters stays steady around 1.8 μm.

The company would like to have high power to detect mean thicknesses 0.2 um away from 15. With n=100, what power does this test have to detect when $u_a = 15.2$? Continue to use $\alpha = 0.05$.

Power=$1-\beta = P(Reject\ H_o | H_o\ false\ and\ u_a = 15.2)$

1. Rewrite H_o rejection region in terms of $\bar{X}$.

$z_{0.975} < -1.96 \quad or \quad z_{0.025} > 1.96.$ which correspond to



$\bar{X} < -1.96 * \frac{1.8}{\sqrt{100}} + 15 = 14.6472$

Or

$\bar{X} > 1.96 * \frac{1.8}{\sqrt{100}} + 15 = 15.3528$

Notice, with larger sample size, z critical values are the same, but critical $\bar{X}$ values are closer to 15

Hypothesis Testing Ex 2:

...medicine is supposed to have a mean particle diameter of μ = 15 μm. .. standard deviation of diameters stays steady around 1.8 μm. With n=100, what power does this test have to detect when $u_a = 15.2$? Continue to use $\alpha = 0.05$. Also calculate the probability of a Type II error.

Power=$1-\beta = P(Reject\ H_o | H_o\ false\ and\ u_a = 15.2)$

2. Calculate power of obtaining $\bar{X}$ larger than those specified above on alternative curve $u_a = 15.2$

Rejection Region

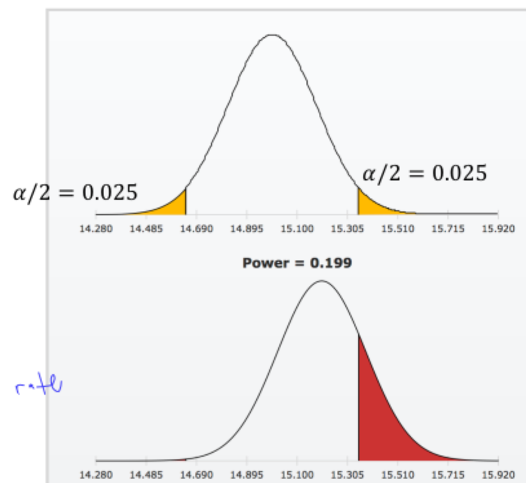
assume true

$Power(\mu_a = 15.2) =$

$P(\bar{x} < 14.65 \mid \mu = 15.2) + P(\bar{x} > 15.35 \mid \mu = 15.2)$

$0.001 + 0.198 = 0.199$

To high error rate



3. Type 2 Error:

$1 - 0.199 = 0.801$

Hypothesis Testing Reminders:

1. Statistical Significance vs Practical Importance
   There may be convincing statistical evidence of a difference or effect, however that difference may be very small and of little practical importance. When large samples are available, even tiny deviations from the null will be significant. Why?

2. Beware of Multiple Analyses