Week 1

S.E=std of error of mean is less than the pop std ,because we expect the means of sampling dist to be less variable than the pop

LO 1. Define sample statistic as a point estimate for a population parameter, for example, the sample mean is used to estimate the population mean, and note that point estimate and sample statistic are synonymous.

LO 2. Recognize that point estimates (such as the sample mean) will vary from one sample to another, and define this variability as sampling variability (sometimes also called sampling variation).

LO 3. Calculate the sampling variability of the mean, the standard error, as $SE = \frac{\sigma}{\sqrt{n}}$ where σ is the population standard deviation.

- Note that when the population standard deviation σ is not known (almost always), the standard error SE can be estimated using the sample standard deviation *s*, so that $SE = \frac{s}{\sqrt{n}}$.

LO 4. Distinguish standard deviation (σ or s) and standard error (SE): standard deviation measures the variability in the data, while standard error measures the variability in point estimates from different samples of the same size and from the same population, i.e. measures the sampling variability.

S.E

LO 5. Recognize that when the sample size n increases we would expect the sampling variability to decrease.

- Conceptually: Imagine taking many samples from the population. When the size of each sample is large, the sample means will be much more consistent across samples than when the sample sizes are small.
- Mathematically: Remember $SE = \frac{\sigma}{\sqrt{n}}$. Then, when *n* increases SE will decrease since n is in the denominator.

Conditions that go along with SE

https://gallery.shinyapps.io/CLT_mean/

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

12:56 / 20:59

THE SECOND CONDITION HERE IS AT MOST 10% AND NOT ATLEAST 10%

LO 4. Recall that independence of observations in a sample is provided by random sampling (in the case of observational studies) or random assignment (in the case of experiments).

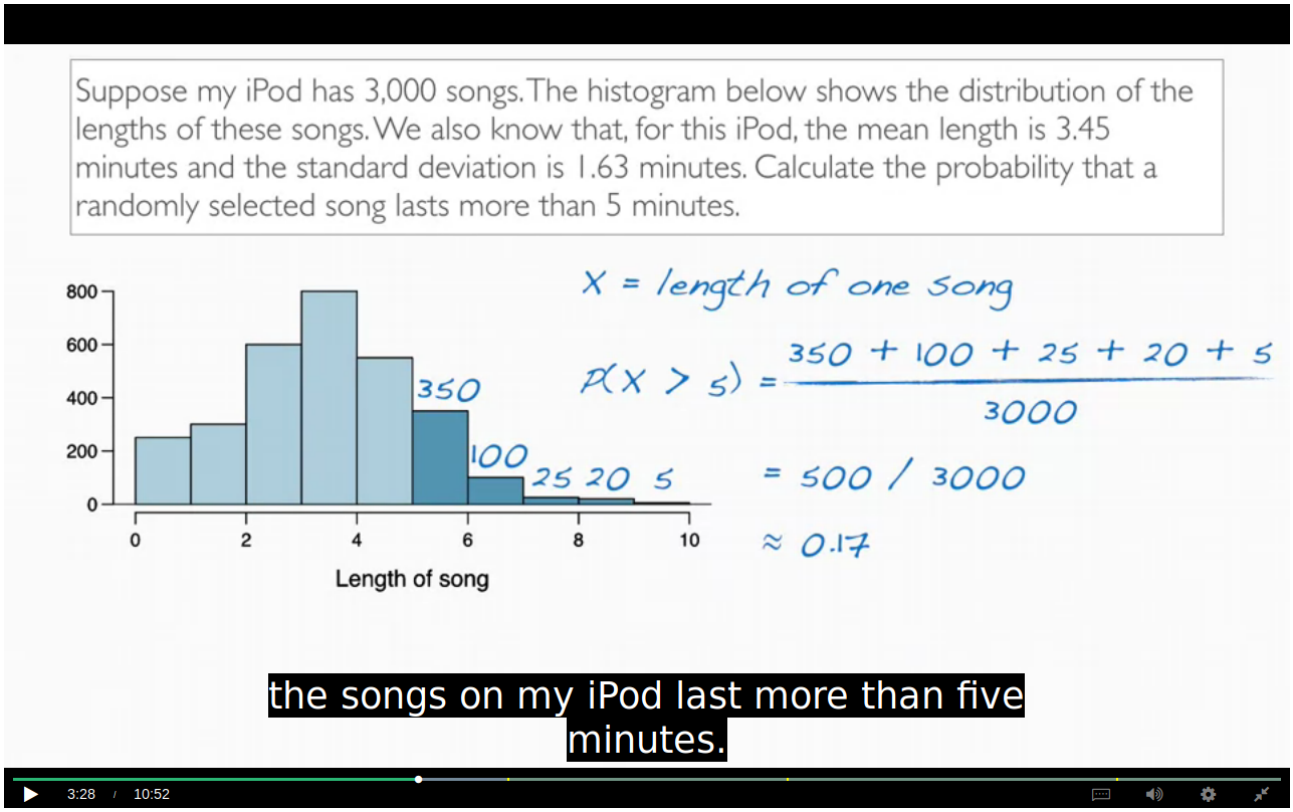
In addition, the sample should not be too large compared to the population, or more precisely, should be smaller than 10% of the population, since samples that are too large will likely contain observations that are not independent.

if the pop dist is normal then the sampling dist is also going to be normal regardless of the sample size however if it is not then we would have to take the sample size>= 30. the more skewed a dist is the larger the s.s will be needed in order to follow the CLT.

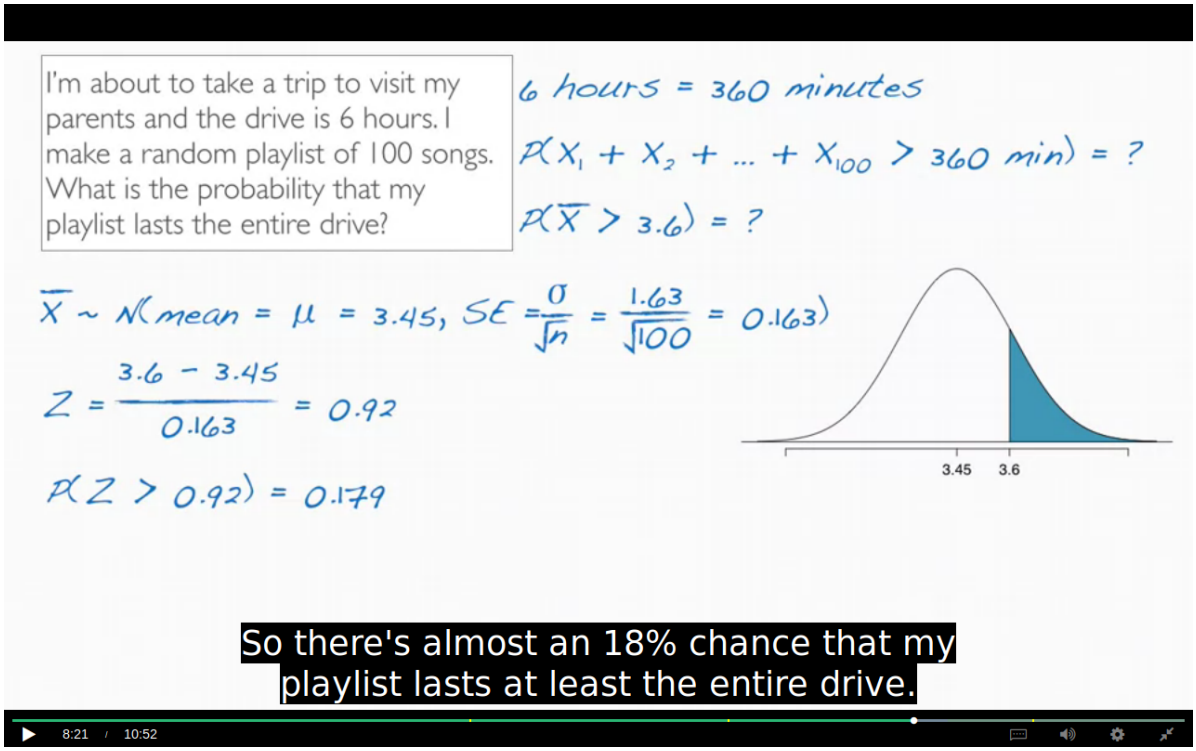
Often times We do not know what dist our population is following,so for those scenarios we plot our sample data and assume that both follow the same dist.

We need to concentrate on sample dist before doing any calculations ,for instance the dist of the incomes can't be the normal as there's no certain limit to what a person can earn.

CASE STUDY FOR SAMPLE DIST: AVG INTAKE OF SNACKS OF DISTRACTED EATERS. Suppose The sample mean is 52 and sd=45.1gm there's a natural boundary at 0 since one cannot eat less than 0 grams of biscuit. So the 68, 95, 99.7 rule is just not going to apply here. Because if we were go more than one standard deviation below the mean, we're going to hit that natural boundary of 0 grams. Therefore that data are likely right skipped



The methods for calculating the prop using Z table can;t be applied here as the dist is right skewed.



Think why this is true ?

Which of the following is equivalent to the desired probability?

$P(X_1 + X_2 + ... + X_100 > 360min) = ?$

- P(each song on the playlist lasts at least 3.6 minutes)
- P(average song length is at least 3.6 minutes)

Correct

Q. When do I use sigma /sqrt(n)?

A. You *always* divide by sqrt(n). However, occasionally the square root of n sometimes equals 1 (making it just σ in the denominator. for example, if you are choosing one person and trying to figure out the probability their weight is under x pounds, then n=1. In other words, if you are calculating a z-score, you can *always* use √(n).

We measure the variability of individual observations with standard deviations. We measure the variability of sample means with standard errors.So whatever the observation is that you plug in in the numerator in your Z-score, its variability belongs in the denominator.
In other words, our observation is an X bar, and not an X.

CONFIDENCE INTERVAL:

LO 1. Define a confidence interval as the plausible range of values for a population parameter.

LO 2. Define the confidence level as the percentage of random samples which yield confidence intervals that capture the true population parameter.

LO 3. Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal.

- In the case of the mean the CLT tells us that if

(1a) the sample size is sufficiently large (n ≥ 30 or larger if the data are considerably skewed), or

(1b) the population is known to have a normal distribution, and

(2) the observations in the sample are independent,

then the distribution of the sample mean will be nearly normal, centered at the true population mean and with a standard error of $\frac{\sigma}{\sqrt{n}}$:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

When the population distribution is unknown, condition (1a) can be checked using a histogram or some other visualization of the distribution of the observed data in the sample.

The larger the sample size (n), the less important the shape of the distribution becomes, i.e. when n is very large the sampling distribution will be nearly normal regardless of the shape of the population distribution.

LO 4. Recall that independence of observations in a sample is provided by random sampling (in the case of observational studies) or random assignment (in the case of experiments).

In addition, the sample should not be too large compared to the population, or more precisely, should be smaller than 10% of the population, since samples that are too large will likely contain observations that are not independent.

LO 5. Recognize that the nearly normal distribution of the point estimate (as suggested by the CLT) implies that a confidence interval can be calculated as

$$point\ estimate \pm z^\star \times SE,$$

,

where $z^\star$ corresponds to the cutoff points in the standard normal distribution to capture the middle XX% of the data, where XX% is the desired confidence level.

- For means this is: $\bar{x} \pm z^\star \frac{\sigma}{\sqrt{n}}$
- Note that $z^\star$ is always positive.

LO 6. Define margin of error as the distance required to travel in either direction away from the point estimate when constructing a confidence interval, i.e. $z^\star \frac{\sigma}{\sqrt{n}}$.

- Notice that this corresponds to half the width of the confidence interval.

LO 7. Interpret a confidence interval as "We are XX% confident that the true population parameter is in this interval", where XX% is the desired confidence level.

- Note that your interpretation must always be in context of the data – mention what the population is and what the parameter is (mean or proportion).

Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$

THE CONDITIONS ARE SAME AS OF CLT BUT THE SECOND IS MORE STRICT HERE
THE SECOND CONDITION HERE IS AT MOST 10% AND NOT ATLEAST 10%

What is the critical value for the 98% confidence level?

You may need to refer to the

normal probability table

- Z = 2.05
- Z = 1.96
- Z = 2.33

  Correct
  Z = 2.33 is the positive cutoff for the middle 98% of the standard normal distribution.

- Z = −2.33

Given that the critical value for a 95% confidence interval is 1.96, for a 98% confidence interval is 2.33, and for a 99% confidence interval is 2.58, what happens to the width of the confidence interval as the confidence level increases (all else held constant)?

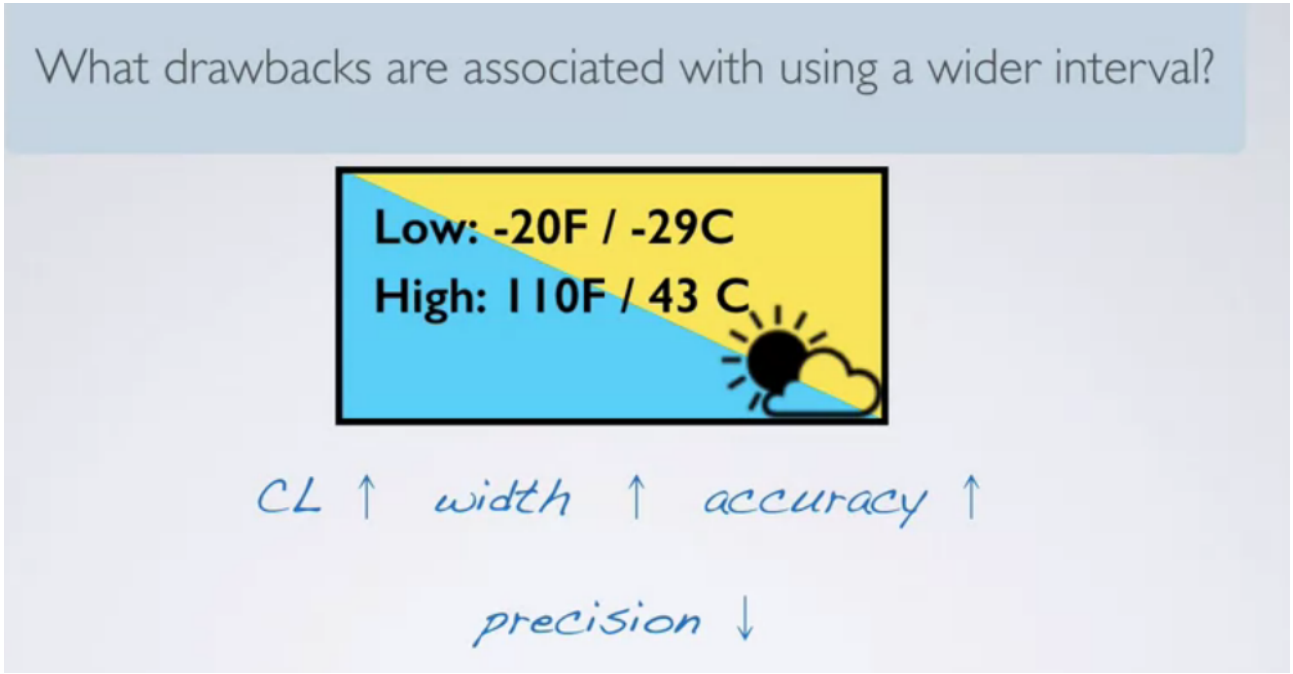
As the confidence level increases the interval gets wider.

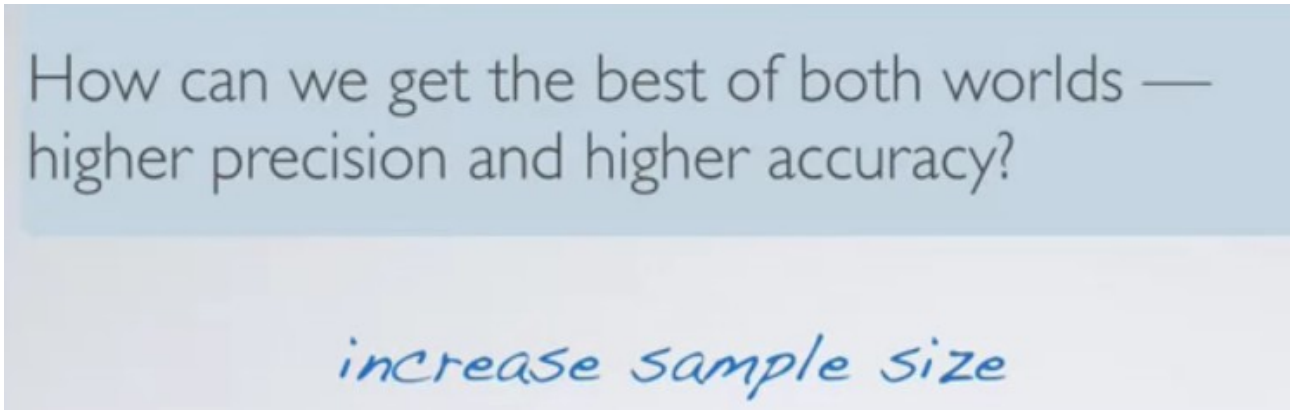
Correct

The higher the confidence level, the larger the critical value, hence the larger the margin of error, and hence the width of the confidence interval.

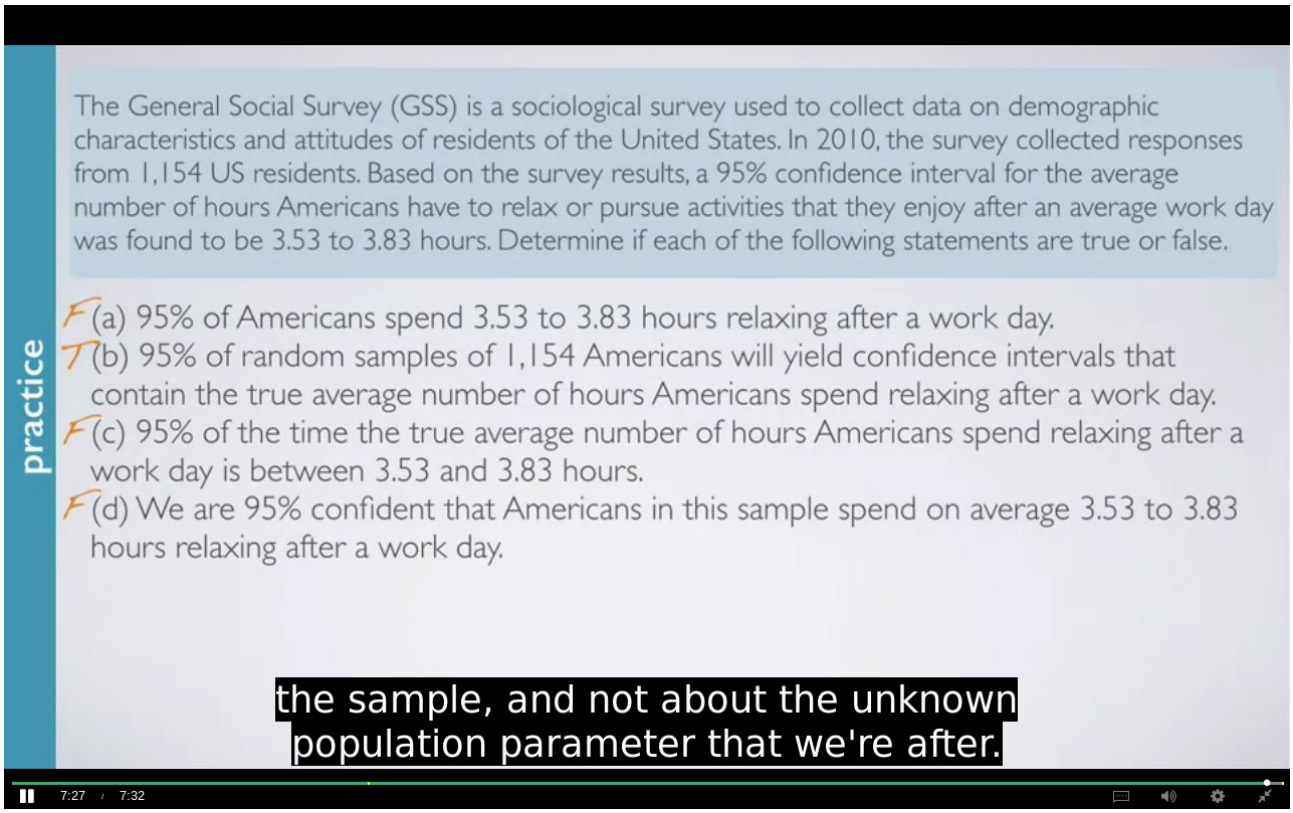
As the confidence level increases the interval gets narrower.

As the accuracy increases the precision decreases





as we increase our sample size the SE decreases and hence our precision will not decrease drastically.



MOST IMPORTANT PART

A says that 95% of Americans spend between
3.53 to 3.83 hours relaxing after a work day.
This is not true because remember that the confidence interval is not about
individuals in the population but instead about the true population parameter.
B says that 95% of ***random samples of 1,154 Americans(and not one individual)***
will yield confidence intervals that contain the true average
number of hours Americans spend relaxing after a work day.
This is indeed the definition of the confidence level.
The percentage of random samples that will yield confidence intervals
that contain the true population parameter.
So this is true.
C says 95% of the time the true average
number of hours Americans spend relaxing after
a work day is between 3.53 and 3.83 hours.

This is not true because the population parameter is not this moving target
that is sometimes within an interval and sometimes outside of it.

And lastly, D says we are 95% confident that Americans ***in this sample***
spend on average 3.53 to 3.83 hours relaxing after a work day.

This is not true because remember that the confidence interval
is not about the sample mean, but is instead about the population mean.
We know exactly what the sample mean is.
It has to be between these values because we construct the confidence interval
around the sample mean.
Therefore, we could actually say that we are 100% confident that Americans
in this sample spend on average between 3.53 and
3.83 hours relaxing after an average work day.

EXAMPLE:

The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. Based on the survey results, a 95% confidence interval for the average number of hours Americans have to relax or pursue activities that you enjoy after an average work day is 3.53 to 3.83 hours. Which of the following is false?

Increasing the confidence level would result in a more accurate but less precise confidence interval.

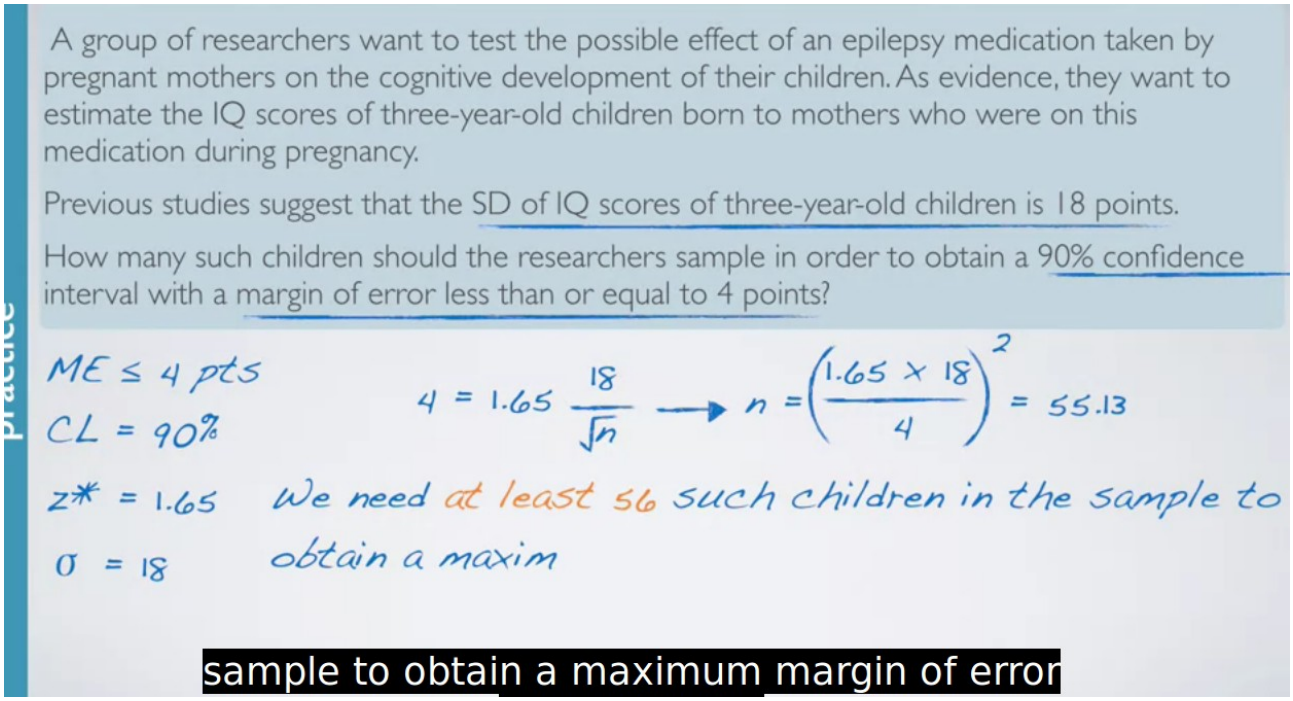
The standard error is approximately 0.075 hours.

The sample mean is 3.68 hours.

The margin of error is 0.3 hours.

Correct

The margin of error is (3.83 – 3.53) / 2 = 0.3 / 2 = 0.15 hours.



Since 55..13 was minimum we will round it to **56**

We found that we needed at least 56 children in the sample to achieve a maximum margin of error of 4 points. How would the required sample size change if we want to further decrease the margin of error to 2 points?

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{n}} \frac{1}{2}$$

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{4n}}$$

the reation between **n** and **me** is inverse and exponential as well as we have to take the square of the number by which we want our **me** to be decreased.

The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. Interpret this interval in context of the data.

We are 95% confident that Americans on average have 3.40 to 4.24 bad mental health days per month.

In this context, what does a 95% confidence level mean?

95% of random samples of 1,151 Americans will yield CIs that capture the true population mean of number of bad mental health days per month.

Which of the following is a condition that needs to be met to calculate a confidence interval for a population mean using methods that rely on the Central Limit Theorem?

A) The population distribution must be nearly normal.

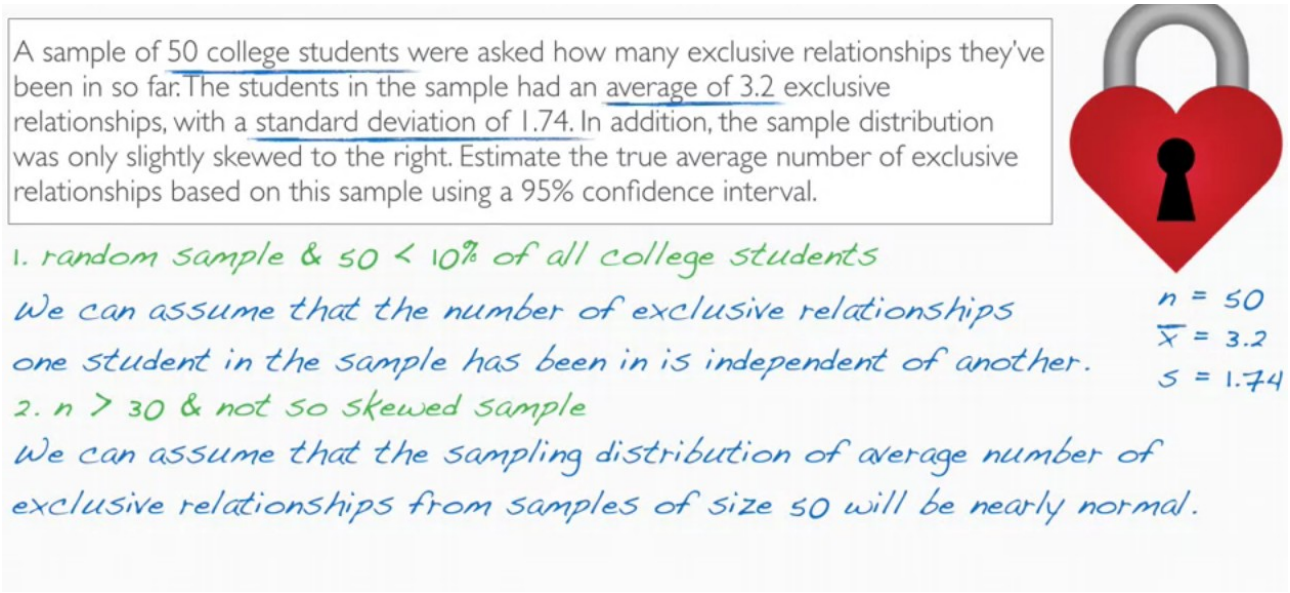
B) At least 10% of the population must be sampled.

C) The sampled observations must be independent with respect to the variable in question.

Correct

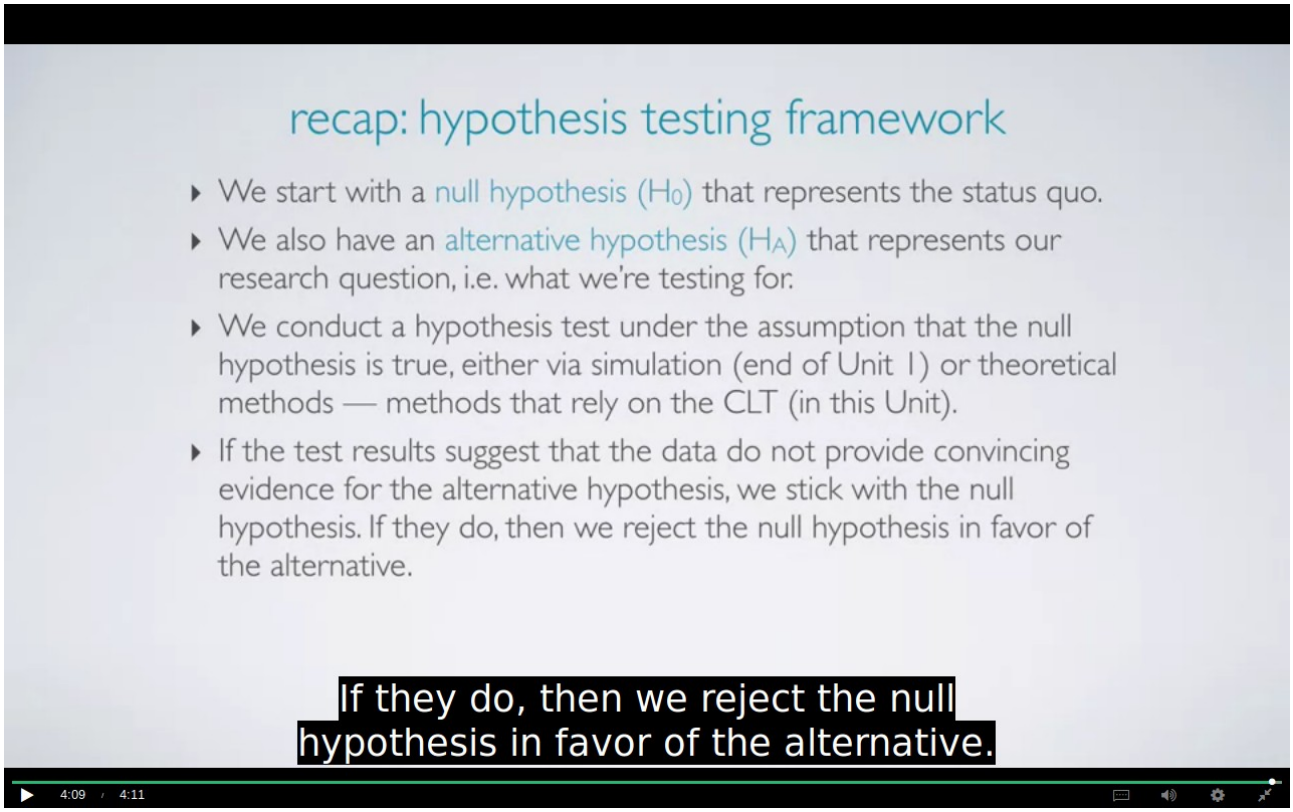
The population distribution does not necessarily need to be nearly normal, as the CLT will hold if the sample size is large even if the population distribution is skewed. At most, not at least, 10% of the population must be sampled. The success-failure condition is useful for categorical variables, not numerical. So the only choice that is correct is “The sampled observations must be independent with respect to the variable in question.”

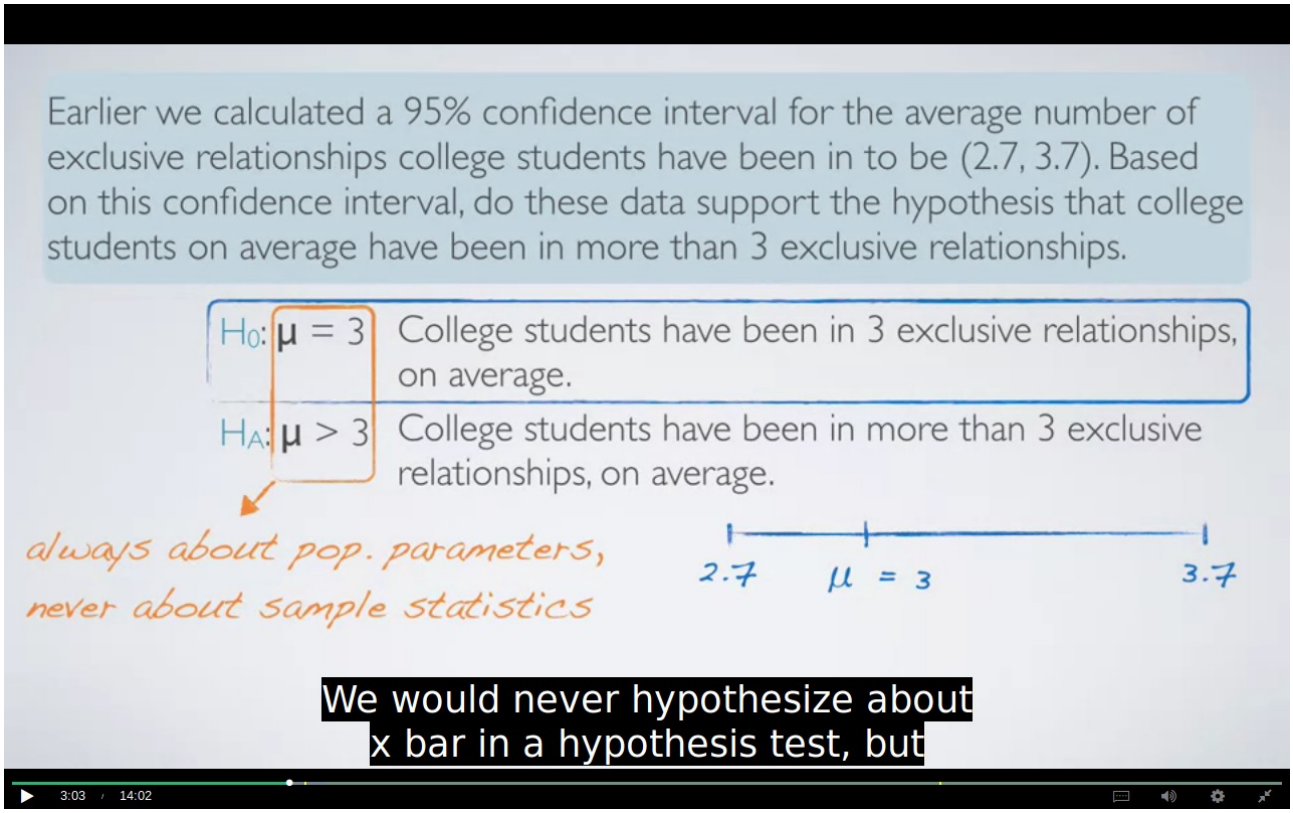
D) There should be at least 10 successes and 10 failures



WEEK 2

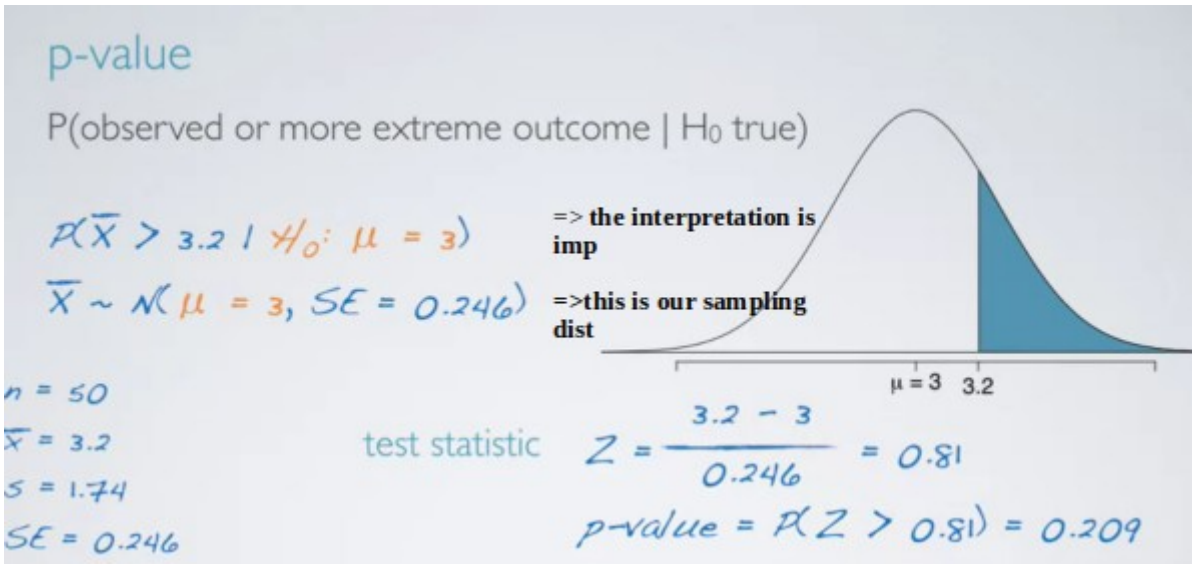
HYPOTHESIS TESTING:





WE NEVER DO HYPOTHESIS TESTING FOR SAMPLE STATISTICS

IMPORTANT:

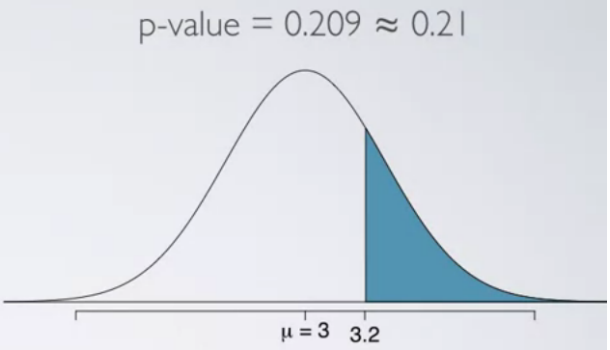


decision based on the p-value

- We used the test statistic to calculate the p-value, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0.
- If the p-value is high (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject H_0.

interpreting the p-value

- If in fact college students have been in 3 exclusive relationships on average, there is a 21% chance that a random sample of 50 college students would yield a sample mean of 3.2 or higher.
- This is a pretty high probability, so we think that a sample mean of 3.2 or more exclusive relationships is likely to happen simply by chance.



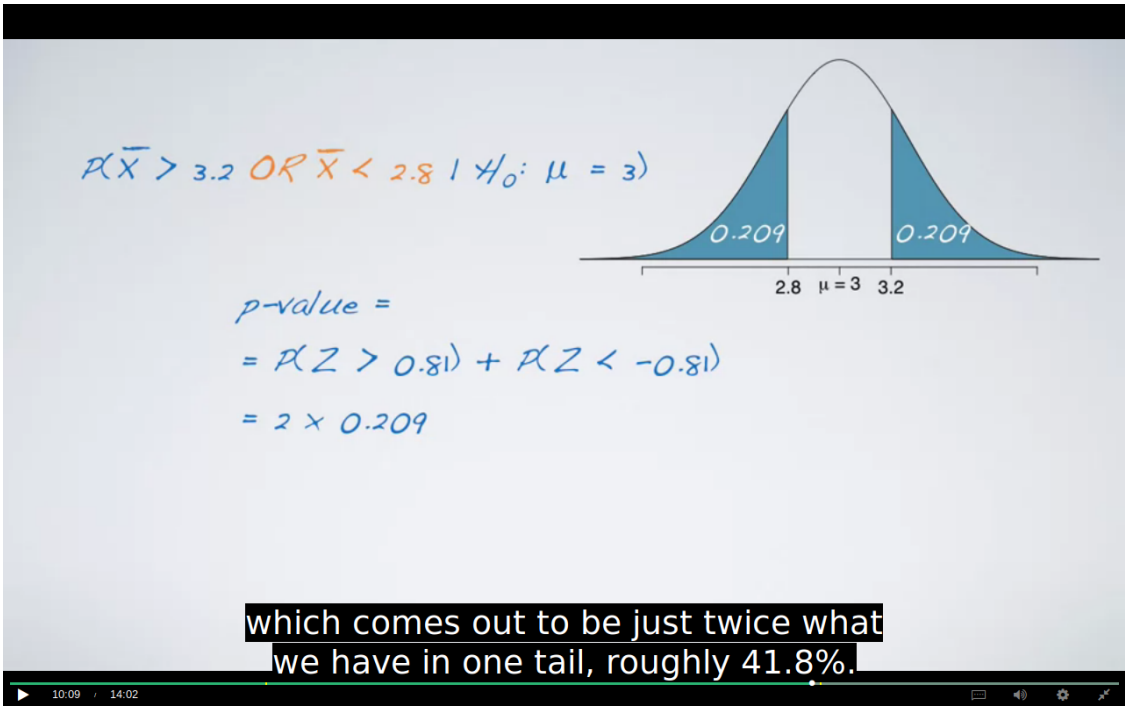
more exclusive relationships is
likely to happen simply by chance.

making a decision

- Since p-value is high (higher than 5%) we fail to reject H_0.
- These data do not data provide convincing evidence that college students have been in more than 3 relationships on average.
- The difference between the null value of 3 relationships and the observed sample mean of 3.2 relationships is due to chance or sampling variability.

two-sided tests

- Often instead of looking for a divergence from the null in a specific direction, we might be interested in divergence in any direction.
- We call such hypothesis tests two-sided (or two-tailed).
- The definition of a p-value is the same regardless of doing a one or two-sided test, however the calculation is slightly different since we need to consider "at least as extreme as the observed outcome" in both directions.



Hypothesis testing for a single mean:

1. Set the hypotheses: $H_0 : \mu = null\ value$
   $H_A : \mu < or > or \neq null\ value$
2. Calculate the point estimate: $\bar{x}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   2. ***Sample size/skew:*** n ≥ 30, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\bar{x} - \mu}{SE}$, $SE = \frac{s}{\sqrt{n}}$
5. Make a decision, and interpret it in context of the research question:
   - If p-value < α, reject H_0; the data provide convincing evidence for H_A.
   - If p-value > α, fail to reject H_0 the data *do not* provide convincing evidence for H_A.

You would fail to reject it and determine that the data do not provide convincing

13:55 / 14:02

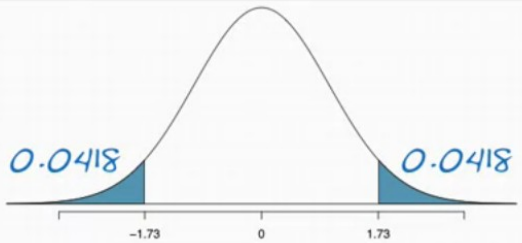
IMPORTANT QUESTION:

A statistics student interested in sleep habits of domestic cats took a random sample of 144 cats and monitored their sleep. The cats slept an average of 16 hours / day. According to online resources domestic dogs sleep, on average, 14 hours day. We want to find out if these data provide convincing evidence of different sleeping habits for domestic cats and dogs with respect to how much they sleep. The test statistic is 1.73.

$\bar{x} = 16$

$H_0: \mu = 14$

$H_A: \mu \neq 14$



p-value = 0.0418 × 2

= 0.0836

What is the interpretation of this p-value in context of these data?

= P(observed or more extreme outcome | H_0 true)

=P(obtaining a random sample of 144 cats that sleep 16 hours or more or 12 hours or less, on average, if in fact cats truly slept 14 hours per day on average) = 0.0836

$n = 144$

$\bar{x} = 16$

$H_0: \mu = 14$

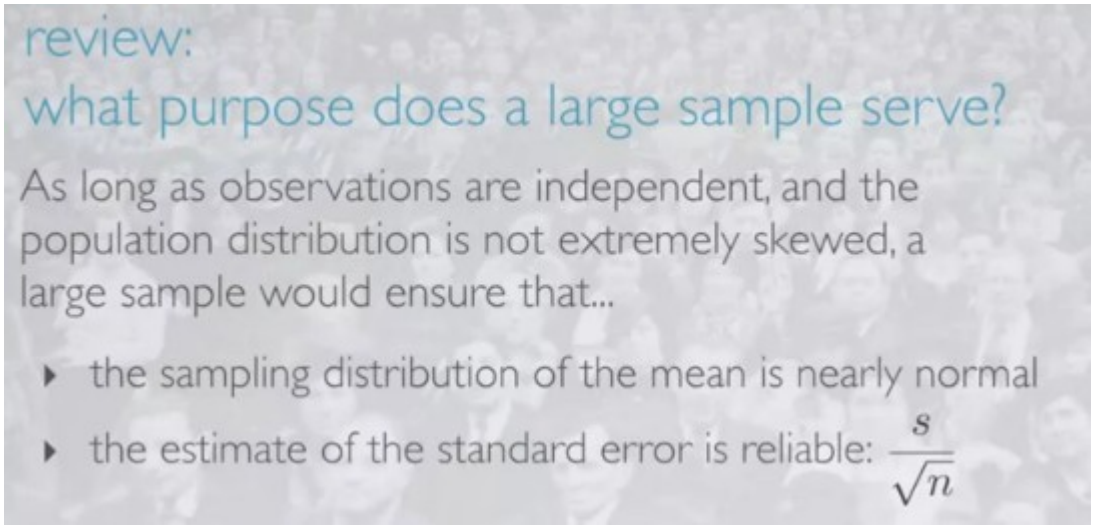
$H_A: \mu \neq 14$

null hypothesis being true is basically saying, if in fact

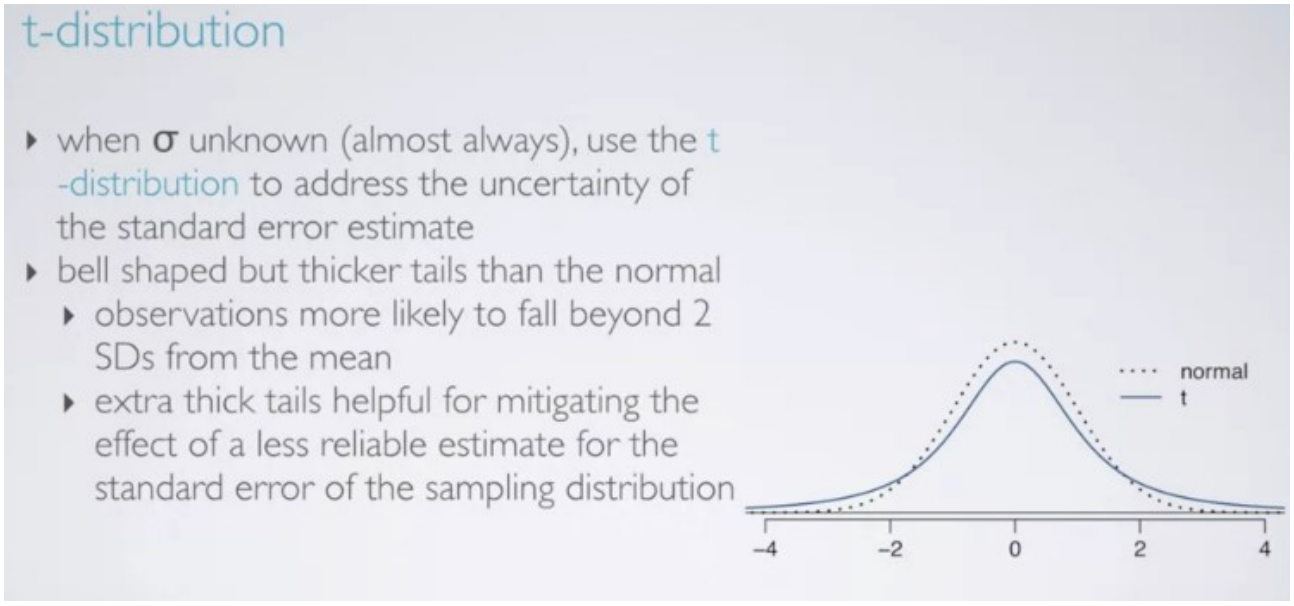
WEEK 3

In a nut shell, the t-distribution is useful for describing the distribution of the sample mean when the population standard deviation, sigma, is unknown,

if a std of pop is known and the sample size is less than 30 than we will still use Z distribution

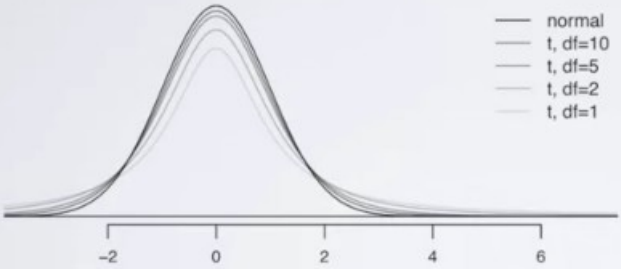


the uncertainty of the SE estimate is addressed by the t distribution which happens because of small sample size.



t distribution

- always centered at 0 (like the standard normal)
- has one parameter: degrees of freedom (df) - determines thickness of tails
  - remember, the normal distribution has two parameters: mean and SD



What happens to the shape of the t-distribution as degrees of freedom increases?

approaches the normal dist.

t statistic

- for inference on a mean where
  - σ unknown, which is almost always
- calculated the same way

$$T = \frac{obs - null}{SE}$$

- p-value (same definition)
  - one or two tail area, based on H_A
  - using R, applet, or table

R

```
> pnorm(2, lower.tail = FALSE) * 2
[1] 0.0455
> pt(2, df = 50, lower.tail = FALSE) * 2
[1] 0.0509
```

lower the df the harder it gets for the Ho to be rejected

Find the following probabilities.

a. $P(|Z| > 2)$ 0.0455

b. $P(|t_{df=50}| > 2)$ 0.0509

c. $P(|t_{df=10}| > 2)$ 0.0734

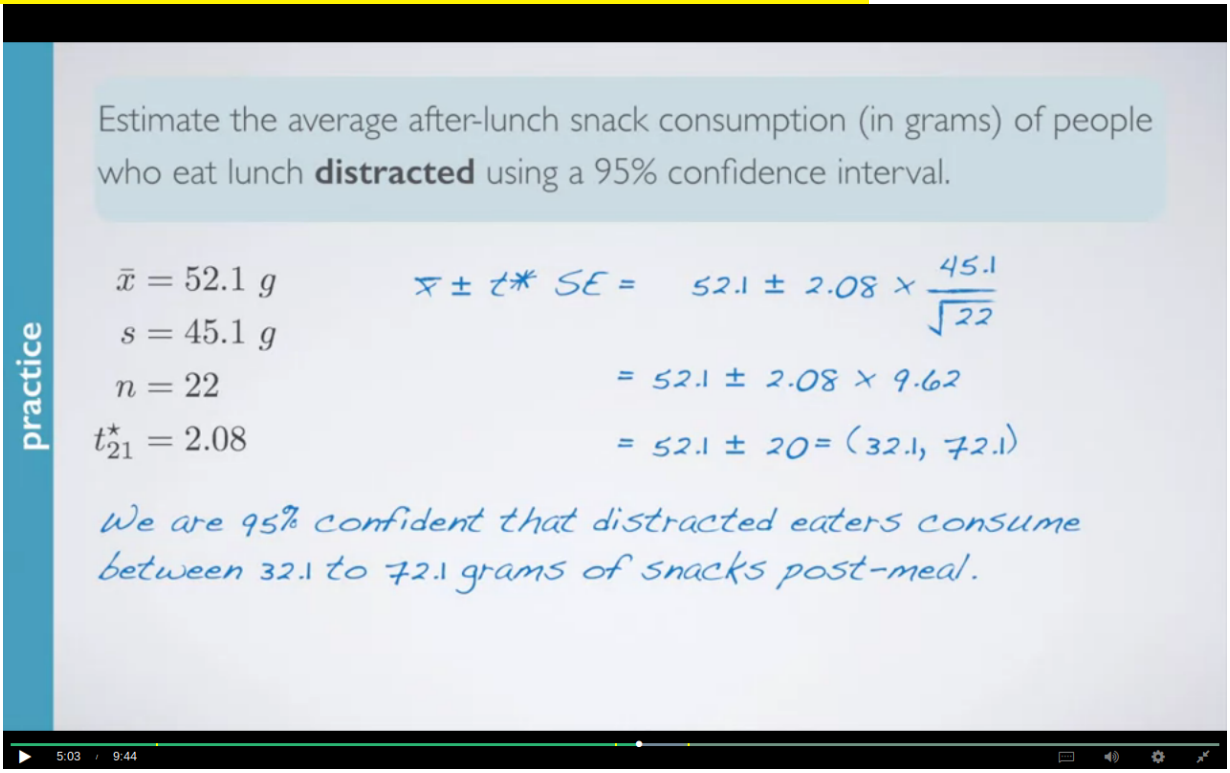
IMPORTANT

Which of the following is the true interpretation based on this confidence interval?

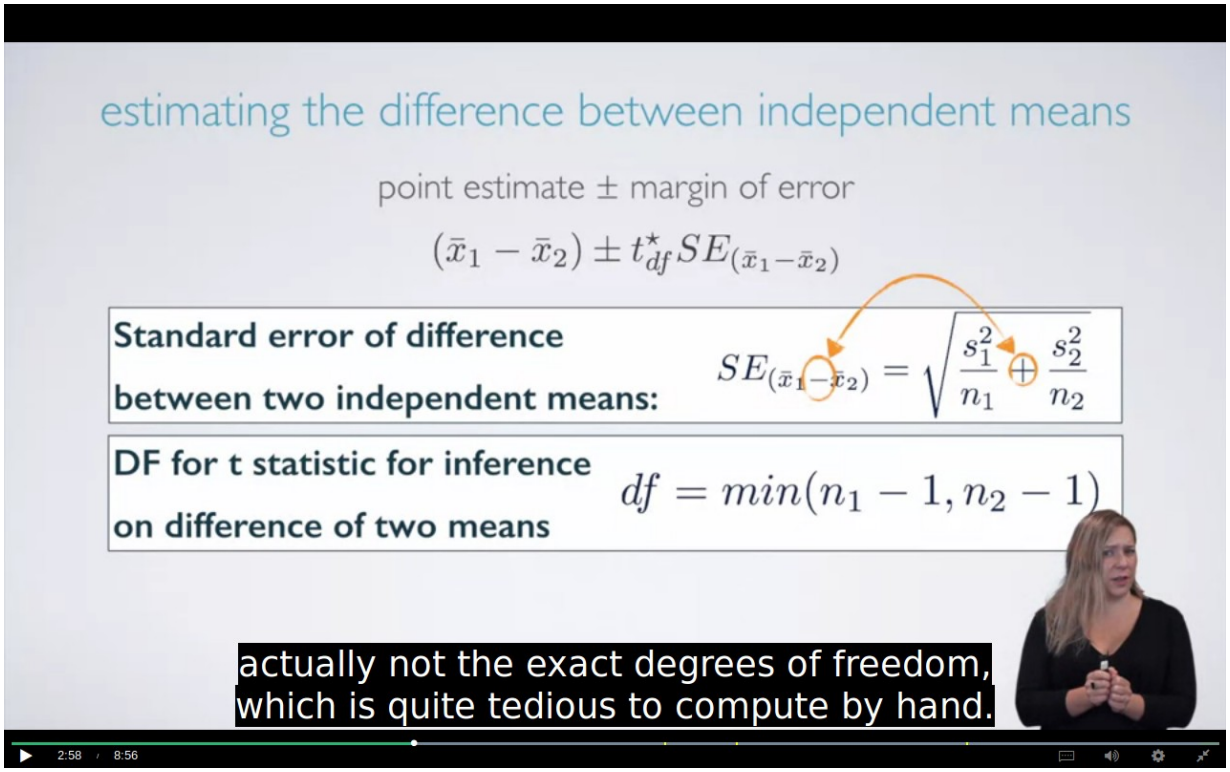
- 95% of distracted eaters consume between 32.1 g to 72.1 g of snacks after lunch.
- 95% of the time the true average snack consumption of distracted eaters is between 32.1 g and 72.1 g.
- 95% of random samples of 22 distracted eaters will yield average post-lunch snack consumption of 32.1 g and 72.1 g.
- We are 95% confident that the 22 distracted eaters in this sample consumed between 32.1 g to 72.1 g of snacks after lunch.
- None of the above.

Correct
Go back to the video for correct interpretation.

WE HAVEN'T TALKED ABOUT HOW CONFIDENT WE ARE IN THE ABOVE QUESTION EXCEPT THE LAST WHICH IS TALKING ABOUT THE SAMPLE



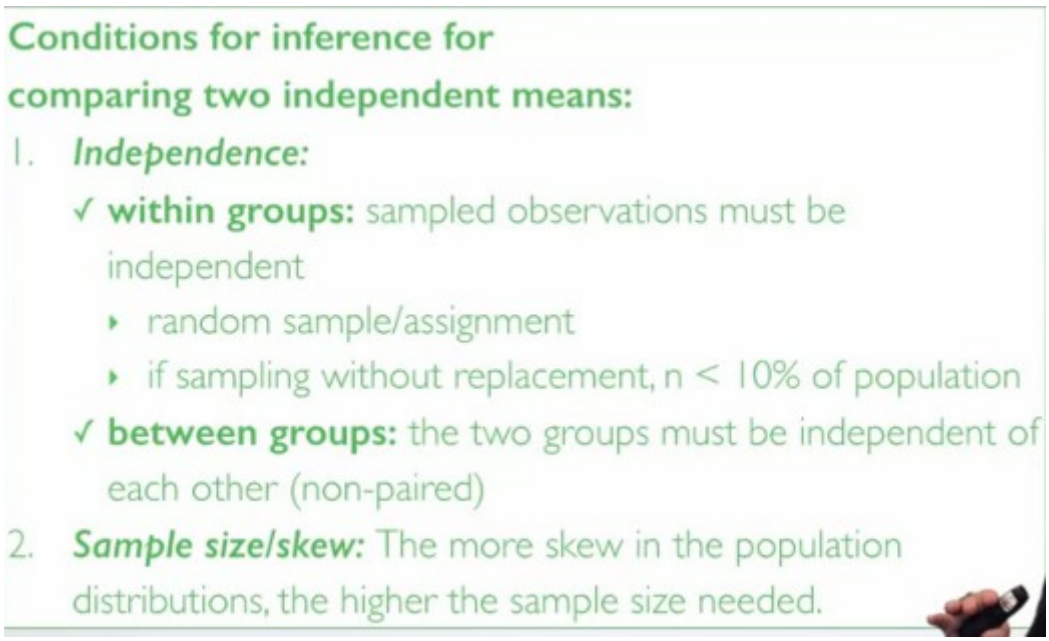
When you bring two unknowns together, the result should always be more variable, regardless of whether you're adding them or subtracting them.



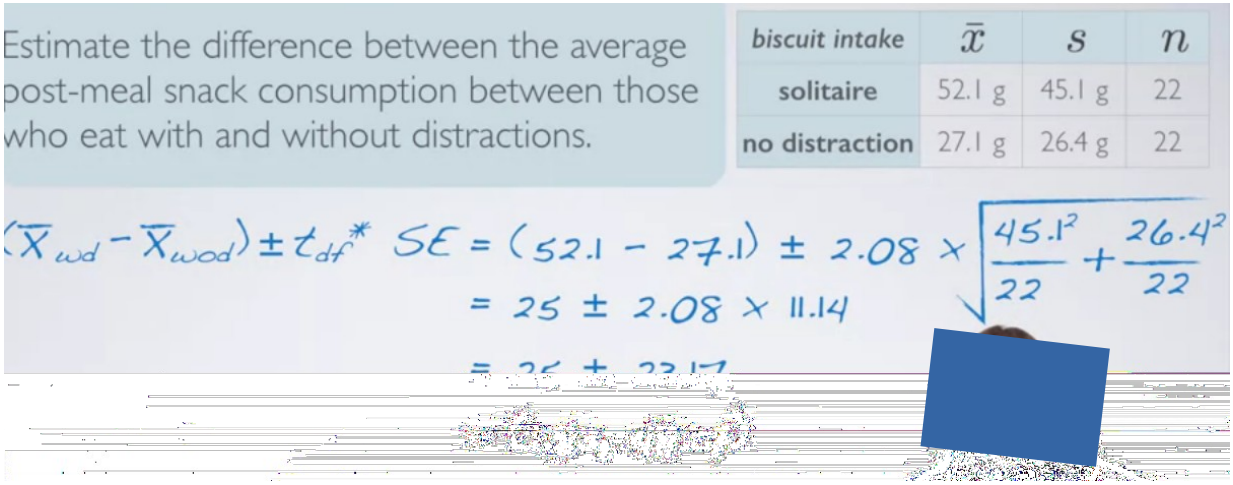
A conservative test always keep the probability of rejecting the null hypothesis well below the significance level. Let's say you're running a hypothesis test where you set the alpha level at 5%. That means that the test will (falsely) give you a significant result 1 out of 20 times. This is called the Type I error rate. A conservative test would always control the Type I error rate at a level much smaller than 5%, which means your chance of getting it wrong will be well below 5% (perhaps 2%).

minimum in calculating the df is basically setting our test ,conservative

if we are taking the df smaller the t distribution will have the wider hands and the probability of any observation lying within 95% of the data increases which lowers the chances of null hypothesis being rejected.



BOTH N1 and N2 should be the 10% of their respective population.



IMPORTANT:

Which of the following is the best interpretation for this interval?

We are 95% confident that...

A) the difference between the average snack consumption of those who eat with and without distractions is between 1.83 g and 48.17 g.

B) those who eat with distractions consume 1.83 g and 48.17 g more snacks than those who eat without distractions, on average.

Correct

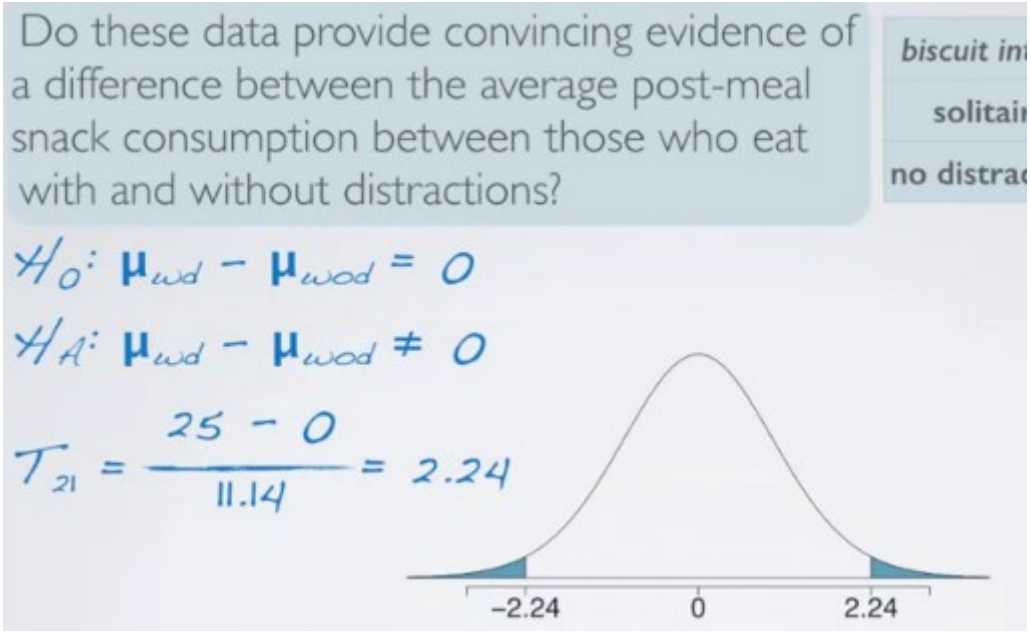
The interpretation should indicate the direction of the relationship (i.e. which group is higher) and the difference between with and without distractions is positive.

C) those who eat with distractions consume 1.83 g and 48.17 g less snacks than those who eat without distractions, on average.

D) those who eat with distractions consume 1.83 g less to 48.17 g more snacks than those who eat without distractions, on average.

E) those who eat with distractions consume 1.83 g more to 48.17 g less snacks than those who eat without distractions, on average.95% of distracted eaters consume between 32.1 g to 72.1 g of snacks after lunch.

HYPOTHESIS TESTING FOR DIFF OF INDEPENDENT MEANS:



P VALUE: BETWEEN 0.02 AND 0.05

just because we observe a difference in the sample means,doesn't necessarily mean that there is something going on that is statistically significant in the actual populations. So, we use statistical inference tools to evaluate if this apparent relationship between distracted eating and snacking more provide evidence of a real difference at the population level. Note that we have a randomized control trial here, so if we do indeed find a significant result, we could then talk about a causal relationship between these two variables.

The confidence interval for the average difference was 1.83 to 48.17 and the hypothesis test evaluating a different between the two means yielded a p-value of roughly 4%.
Which means that we would reject the null hypothesis and conclude that these data do indeed provide convincing evidence that there is a difference between the average snack intake of distracted and non-distracted eaters.

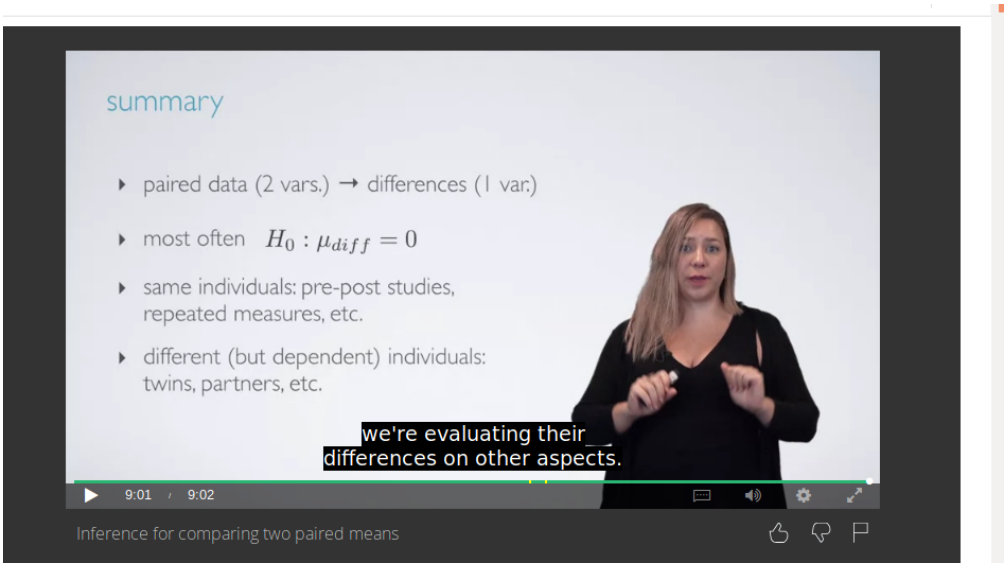
COMPARING PAIRED MEANS:

the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations. Common applications of the paired sample t-test include case-control studies or repeated-measures designs. Suppose you are interested in evaluating the effectiveness of a company training program. One approach you might consider would be to measure the performance of a sample of employees before and after completing the program, and analyze the differences using a paired sample t-test.

In situations where we do inference for pair data, most often the null hypothesis sets the average difference between the two paired means equal to zero.
Indicating no difference between them.

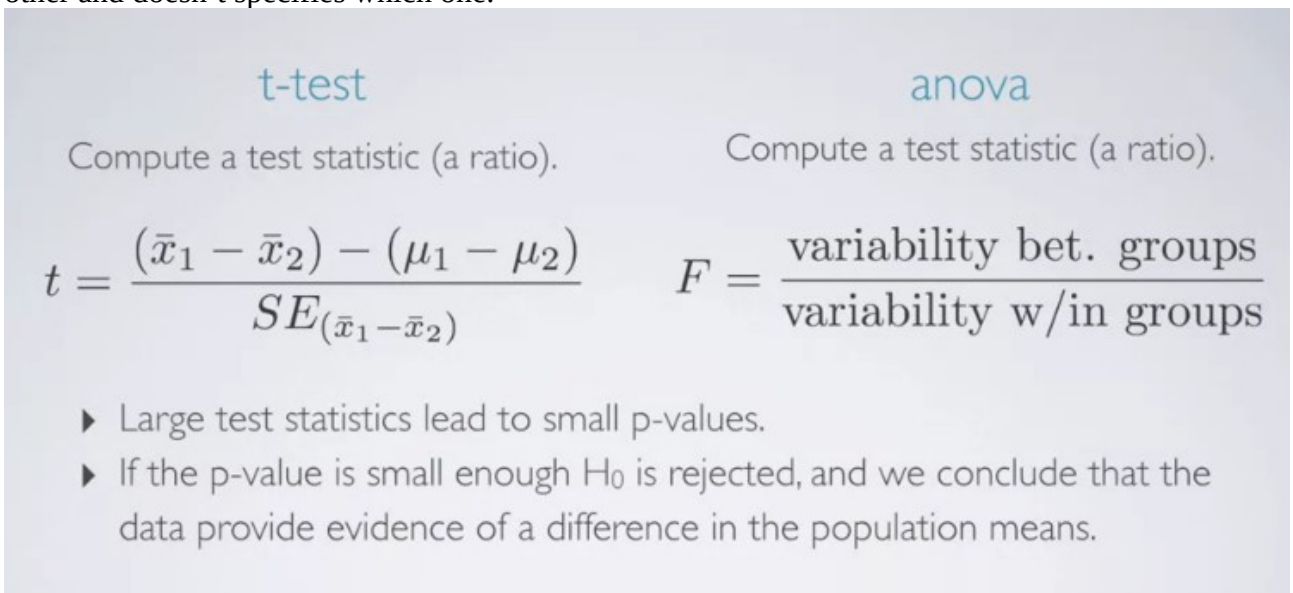
Paired data can happen when we have a set of data from the same set of people.
Like in this case or in cases of pre-post studies.
A weight-loss study, for example, is a good example here.

The post-weight of an individual after a diet regimen
will necessarily be dependent on their pre-weight.
Other studies might also take repeated measures on the same set of people.
For example, you might measure reaction time of the same set of people after they
have spent the recommended amount of 7.5 hours the previous night or
if they've only spent two hours.
We might also use paired approaches when we have different sets of subjects to
begin with.
But for some reason these subjects we believe to be not independent.
Twin studies is an obvious example for these or studies on partner A and
partner B who are in a relationship.
We would design these studies as paired if we believe these individuals
in the two groups are similar on certain aspects and
we're evaluating their differences on other aspects.

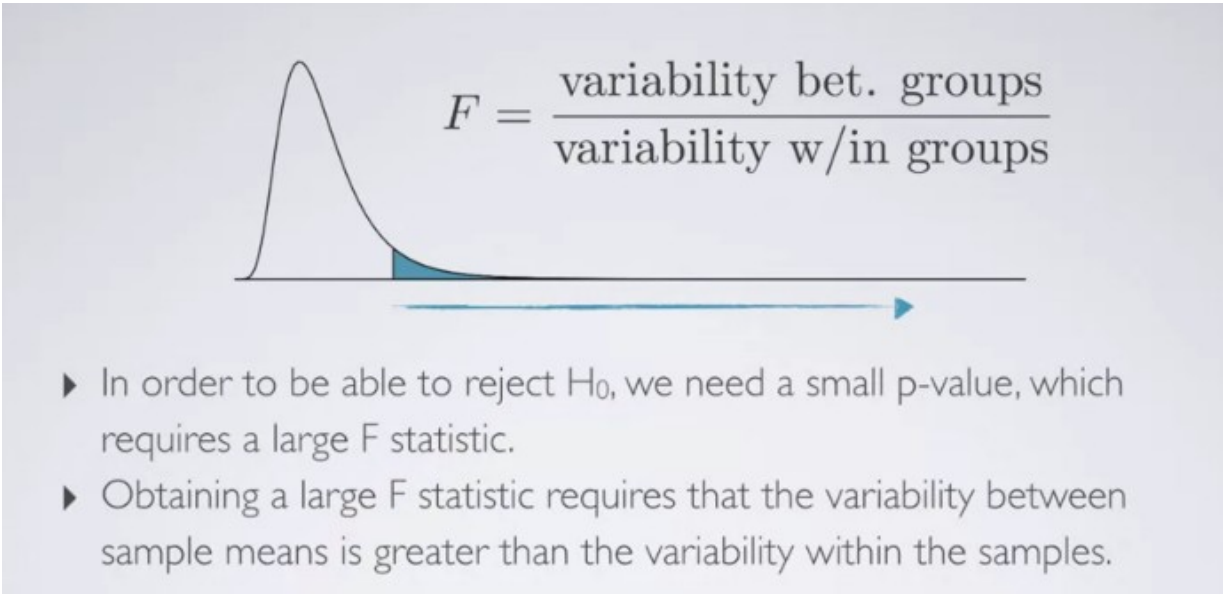


ANOVA:

we use anova to compare more than 2 means the null hypothesis says that there's nothing going on all the means are same while the alternate says that at least one pair of mean are diff from each other and doesn't specifies which one.



F DISTRIBUTION IS RIGHT SKEWED AND ALWAYS POSITIVE SINCE IT IS A RATIO OF VARIABILITY WHICH CAN NEVER BE NEGATIVE



Which of the following is false about the independence condition for this analysis?

Remember, data source is the General Social Survey (GSS), and there are 41 participants in lower, 407 in working, 331 in middle, and 16 in upper class.

Random sampling condition can be assumed to be met since the GSS likely selects participants randomly.

Number of participants in each class are less than 10% of their respective populations.

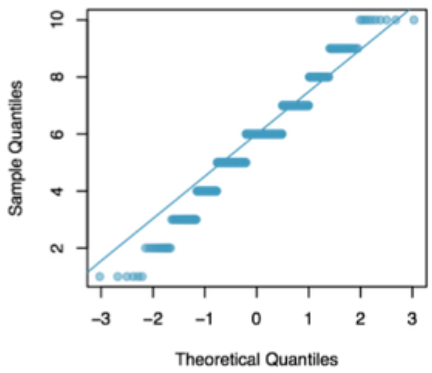
We can assume that one participant in a given class is independent of another participant in that class with respect to their vocabulary score.

The groups are not independent since they come from the same sample of 795 participants who took the GSS.

Correct

The data for the groups come from the same survey, but this doesn’t mean the groups aren’t independent. Lack of independence would be caused by having a correspondence between people in the groups.

Why are there jumps in the normal probability plots? Choose the best answer.



- vocabulary scores are rounded
- vocabulary scores can only be whole numbers since it's count data (number of questions answered correctly)

Correct
Reason for the jumps is that the theoretical (normal) distribution is defined on a continuous scale but the vocabulary scores are discrete.

- vocabulary scores are right skewed
- vocabulary scores can't be negative