Please pull out your class notes from yesterday (The Preschool Problem)

AP® Exam Tip

The formula for the two-sample z statistic for a test about $p_1 - p_2$ often leads to calculation errors by students.

As a result, your teacher may recommend using the calculator's 2-PropZTest feature to perform calculations on the AP® Statistics exam.
(I Do !)

Be sure to name the procedure (two-sample z test for $p_1 - p_2$) in the "Plan" step and **report the standardized test statistic (z = -2.32) and P-value (0.98) in the "Do" step.**

Example of A Significance Test for p_1-p_2

Preschool - To study the long-term effects of preschool programs for poor children, researchers designed an experiment. They recruited 123 children who had never attended preschool from low-income families in Michigan. Researchers randomly assigned 62 of the children to attend preschool (paid for by the study budget) and the other 61 to serve as a control group who would not go to preschool. One response variable of interest was the need for social services as adults. Over a 10-year period, 38 children in the preschool group and 49 in the control group have needed social services.

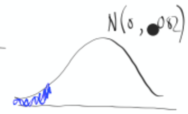
1. Do these data provide convincing evidence that preschool reduces the later need for social services for children like the ones in this study? Justify your answer.

62 attended preschool (38 used Social Services as adults)
61 no pre-school (49 used Social Services as adults)

DO

$$\text{Test Stat} = \frac{\text{Stat} - \text{Null}}{\text{SD}}$$

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{.71(.29)}{62} + \frac{.71(.29)}{61}} \approx .082$$



$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sigma_{\hat{p}_1 - \hat{p}_2}}$$

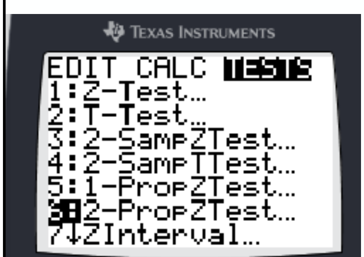
$$z = \frac{-.19 - 0}{.082}$$

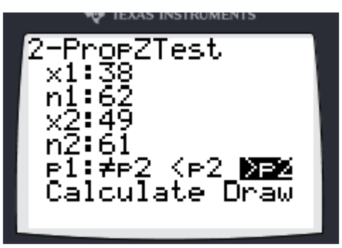
$$z = -2.32$$

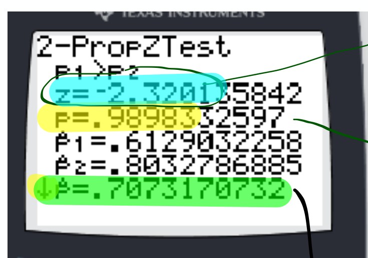
$P(z < -2.32) = .0102$

normalcdf

don't forget to multiply by 2 in some problems







Statistic

P-Value

Combined proportion

$\hat{p}$

This disadvantage of doing this on every problem on every Significance Test for a difference of proportions (including HW) is that you might not develop/practice some of the details for multiple choice questions.

The same formulas will apply in Ch. 12.

but it is recommended you use the technology on the AP EXAM

Warm Up

EXPERIMENTAL DESIGN 8

Which of the following is most useful in establishing cause-and-effect relationships?

(A) A complete census

(B) A least squares regression line showing high correlation

(C) A simple random sample (SRS)

(D) A well-designed, well-conducted survey incorporating chance to ensure a representative sample

(E) A controlled experiment

Regression lines show association, not causation, Surveys suggest relationships, which controlled experiments can help show to be cause and effect.

Answer: **E**

EXPERIMENTAL DESIGN 12

Sampling error occurs

(A) when interviewers make mistakes resulting in bias.

(B) when interviewers use judgment instead of random choice in picking the sample.

(C) when samples are too small.

(D) because a sample statistic is used to estimate a population parameter.

(E) in all of the above cases.

Different samples give different statistics, all of which are estimates for the same population parameter, and so error, called sampling error, is naturally present.

Answer: **D**

Ch.8 Confidence Intervals
σ
μ

Ch.9 Hypothesis Test
σ
μ

Ch.10 Confidence Intervals Hypotheses Tests $\sigma - \sigma$ } 10.1

Confidence Intervals Hypotheses Tests $\mu - \mu$ } 10.2

which is hardest ?

✓DESCRIBE the shape, center, and variability of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

✓DETERMINE whether the conditions are met for doing inference about a difference between two means.

✓CONSTRUCT and INTERPRET a confidence interval for a difference between two means.

TWO GROUPS
(heights)

both groups
Population
Normally distributed

$\mu_G = 56.4$ in

$\sigma_G = 2.7$ in

10 year old
Girls

$\mu_B = 55.7$ in

$\sigma_B = 3.8$ in

10 year old
boys

Suppose we take
independent SRS's
of 12 girls and 8 boys

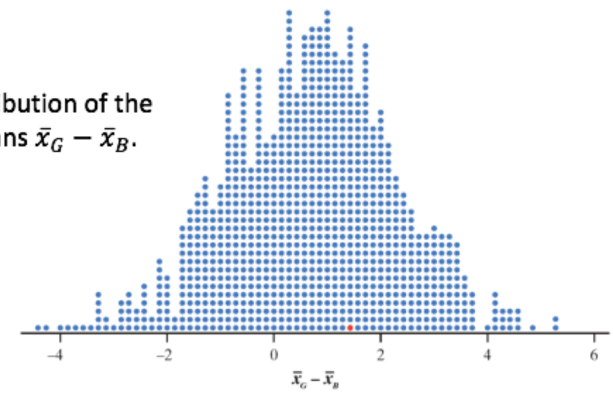
and measure their heights

	Sampling distribution of $\bar{x}_G$	Sampling distribution of $\bar{x}_B$
Shape	Approximately Normal, because the population distribution is approximately Normal	Approximately Normal, because the p distribution is approximately Normal
Center	$\mu_{\bar{x}_G} = \mu_G = 56.4$ inches	$\mu_{\bar{x}_B} = \mu_B = 55.7$ inches
Variability	$\sigma_{\bar{x}_G} = \frac{\sigma_G}{\sqrt{n_G}} = \frac{2.7}{\sqrt{12}} = 0.78$ inch because 12 < 10% of all 10-year-old girls in the United States.	$\sigma_{\bar{x}_B} = \frac{\sigma_B}{\sqrt{n_B}} = \frac{3.8}{\sqrt{8}} = 1.34$ inches because 8 < 10% of all 10-year-ol in the United States.

but, what can we say about the difference in sample means $\bar{X}_G - X_B$

What can we say about the difference $\bar{x}_G - \bar{x}_B$ in the sample means?

Simulated sampling distribution of the difference in sample means $\bar{x}_G - \bar{x}_B$.



Ch. 7 Sampling Distributions

Sample Means

Important ideas:	
Sampling Distrib. of $\bar{x}$ $\mu_{\bar{x}} = \mu$ $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ provide 10% condition is met $n < 0.1N$	If a popul. is approx Normal... the sampling distrib of $\bar{x}$ will also be approx normal

6.2

Combining Probability Distributions

Important ideas:

Adding & Subtracting Random Variables

Normal Probab. Distribution

$\mu_{X+Y} = \mu_X + \mu_Y$ | $\mu_{X-Y} = \mu_X - \mu_Y$

$\sigma_{x+y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$ | $\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

NOTICE

6.2

Combining Probability Dis

Important ideas:

Adding & Subtracting Random Variables

$\mu_{X+Y} = \mu_X + \mu_Y$ | $\mu_{X-Y} = \mu_X - \mu_Y$

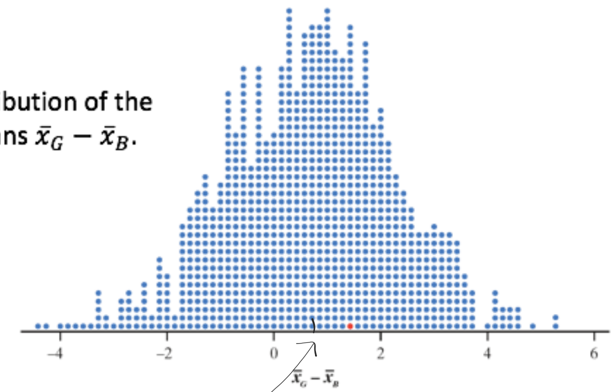
$\sigma_{x+y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$ | $\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

NOTICE

Add the Vairances then $\sqrt{\ }$

What can we say about the difference $\bar{x}_G - \bar{x}_B$ in the sample means?

Simulated sampling distribution of the difference in sample means $\bar{x}_G - \bar{x}_B$.



$\mu_{\bar{x}_G - \bar{x}_B} = \mu_G - \mu_B$

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is **Normal** if both population distributions are Normal. It is **approximately Normal** if both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$) or if one population is Normally distributed and the other sample size is large.

Shape

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is **Normal** if both population distributions are Normal. It is **approximately Normal** if both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$) or if one population is Normally distributed and the other sample size is large.
- **The mean of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is**

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$$

Center

The Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is **Normal** if both population distributions are Normal. It is **approximately Normal** if both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$) or if one population is Normally distributed and the other sample size is large.
- The **mean** of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$$

- The **standard deviation** of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

as long as the *10% condition* is met for both samples: $n_1 < 0.10N_1$ and $n_2 < 0.10N_2$.

Variability

So.......

When creating C.I.'s

for $\mu_1 - \mu_2$

Conditions for Constructing a Confidence Interval About a Difference in Means

Random: The data come from two independent random samples or from two groups in a randomized experiment.

10%: When sampling without replacement, $n_1 < 0.10N_1$ and $n_2 < 0.10N_2$.

Large Counts: Normal/Large Sample: For each sample, the corresponding population distribution (or the true distribution of response to the treatment) is Normal or the sample size is large ($n \geq 30$). For each sample, if the population (treatment) distribution has unknown shape and $n < 30$, a graph of the sample data shows no strong skewness or outliers.

Two Sample t Interval for $\mu_1 - \mu_2$

Assuming the conditions have been met

Important ideas:

Statistic $\pm$ (critical value)(Std. Dev of Statistic)

Two Sample t Interval for $\mu_1 - \mu_2$

Assuming the conditions have been met

Important ideas:

Statistic $\pm$ (critical value)(Std. Dev of Statistic)

$(\bar{x}_1 - \bar{x}_2)$

statistic ± (critical value) · (standard deviation of statistic)

$(\bar{x}_1 - \bar{x}_2) \pm$ (critical value) · (standard deviation of statistic)

don't know population std. deviations most of the time

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\boxed{s_1^2}}{n_1} + \frac{\boxed{s_2^2}}{n_2}}$$

We use s_1 and s_2 to estimate σ_1 and σ_2. We call this the **standard error** of $\bar{x}_1 - \bar{x}_2$.

Two Sample t Interval for $\mu_1 - \mu_2$

Assuming the conditions have been met

Important ideas:

Statistic $\pm$ (critical value)(Std. Dev of Statistic)

$(\bar{x}_1 - \bar{x}_2)$

$$(\bar{x}_1 - \bar{x}_2) \pm (\text{critical value})(SE_{\bar{x}_1 - \bar{x}_2})$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the critical value with $C\%$ of the area between $-t^*$ and t^* for the t distribution with degrees of freedom using either

- Option 1 (technology) or
- Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$)

smaller df would yield larger which is more conservative

Important ideas:

Statistic $\pm$ (critical value)(Std. Dev of Statistic)

$(\bar{x}_1 - \bar{x}_2)$

$$(\bar{x}_1 - \bar{x}_2) \pm (\text{critical value})(SE_{\bar{x}_2 - \bar{x}_1})$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$$

where t^* is the critical value with $C\%$ of the area between $-t^*$ and t^* for the t distribution with degrees of freedom using either

- Option 1 (technology) or
- Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$)

smaller df would yield larger which is more conservative

(III) Inferential Statistics

Standardized test statistic: $\dfrac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$

Confidence interval: statistic $\pm$ (critical value) • (standard deviation of statistic)

Single-Sample

Statistic	Standard Deviation of Statistic
Sample Mean	$\dfrac{\sigma}{\sqrt{n}}$
Sample Proportion	$\sqrt{\dfrac{p(1-p)}{n}}$

Two-Sample

Statistic	Standard Deviation of Statistic
Difference of sample means	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ Special case when $\sigma_1 = \sigma_2$ $\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$
Difference of sample proportions	$\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$ Special case when $p_1 = p_2$ $\sqrt{p(1-p)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$

← 10.2 means

← 10.1 proportions

Which cookie has the most chips?

VS **Treasure Chips**

Is there a difference in the number of chocolate chips in **Chips Ahoy** cookies versus the number of chocolate chips in **Treasure Chips** cookies? Each pair of students will count the number of chocolate chips in 1 Chips Ahoy cookie and 1 Treasure Chips cookie. Due to the factories processes, we can assume the population distributions of # of chips are approximately normal and that the samples are random.

1. Record the number of chocolate chips in each cookie. Write them on the board.

\# in Chips Ahoy = _______ # in Treasure Chips = _______

# of chips in your cookie

18 17 21 18
23 18 16
24 21 22
14 19 20
15 19 21

$\bar{x}_1 = 19.1$

$s_1 = 2.83$

Treasure Chips

31 16 22 21
25 32 18
27 21 21
21 31 24
24 23 32
20 20

zeke

$\bar{x}_2 = 23.3$

~~$s_2 = 2$~~

$s_2 = 4.6$

$n = 18$

2. Find the class totals find the mean number of chocolate chips for each type of cookie, the standard deviation and the difference.

Chips Ahoy: $\bar{x}_1 = 19.1$ Treasure Chips: $\bar{x}_2 = 23.3$ Difference: $\bar{x}_1 - \bar{x}_2 = 19.1 - 23.3$

$= -4.2$

$s_1 = 2.83$ $s_2 = 4.6$

$n = 16$

3. If we repeated this process many times and created a dotplot, we would have the sampling distribution of $\bar{x}_1 - \bar{x}_2$. Describe the shape, center and variability of the sampling distribution.

Shape: Approx Normal because Population is Approx. Normal

Center:

$$\mu_{\bar{x}_c - \bar{x}_T} = \mu_c - \mu_T = -4.2$$

Variability:

$$\sigma_{\bar{x}_c - \bar{x}_T} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{2.83^2}{16} + \frac{4.6^2}{18}} = 1.29$$

4. With at least one other person in the class, discuss and confirm tthat the conditions for constructing a confidence interval been met? (don't have to write details)

5. Construct a 95% confidence interval for the true difference in the mean number of chocolate chips in Chips Ahoy and Treasure Chips. (you don't need to write the general or specific formulas in this particular unless you want) but do show the "work".

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(19.1 - 23.3) \pm 1.753 \sqrt{\frac{2.83^2}{16} + \frac{4.6^2}{18}}$$

$$-4.2 \pm 2.26$$

$$= (-6.46, -1.94)$$

degrees of freedom

It's complicated

Option 1 (technology): Just let technology compute the df. This will most likely not be a whole number.

Option 2 (conservative): smaller of $n_1 - 1$ and $n_2 - 2$

degrees of freedom

It's complicated

Best Option

Option 1 (technology): Just let technology compute the df. This will most likely not be a whole number.

Option 2 (conservative): smaller of $n_1 - 1$ and $n_2 - 2$

where t^* is the critical value with $C\%$ of the area between $-t^*$ and t^* for the t distribution with degrees of freedom using either

- Option 1 (let technology calculate the degrees of freedom) or
- Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$)

smaller df would yield larger which is more conservative

$n_1 =$ $n_2 =$

TABLE B

C 95%

$df = n-1$

Tail Prob. 0.05

$df =$

$t^* =$

4. With at least one other person in the class, discuss and confirm tthat the conditions for constructing a confidence interval been met? (don't have to write details)

5. Construct a 95% confidence interval for the true difference in the mean number of chocolate chips in Chips Ahoy and Treasure Chips. (you don't need to write the general or specific formulas in this particular unless you want) but do show the "work".

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\pm \sqrt{\quad}$$

What if we swithed the order of subtraction to $(P_2 - P_1)$?

example: $(\overset{\hat{p}_1}{34.53} - \overset{\hat{p}_2}{23.70})$ becomes $(\overset{\hat{p}_2}{23.70} - \overset{\hat{p}_1}{34.53})$

10.83 becomes -10.83

confidence interval (3.83, 17.83) gets flipped (−17.83, −3.83)

In answers in the back of your textbook:

It's OK if the endpoints of the endpoints of the interval in your answer to have opposite signs.

Either subtraction order is correct as long as you clearly identify the order you are using.

4. With at least one other person in the class, discuss and confirm tthat the conditions for constructing a confidence interval been met? (don't have to write details)

5. Construct a 95% confidence interval for the true difference in the mean number of chocolate chips in Chips Ahoy and Treasure Chips. (you don't need to write the general or specific formulas in this particular unless you want) but do show the "work".

Option 1 (Tech)

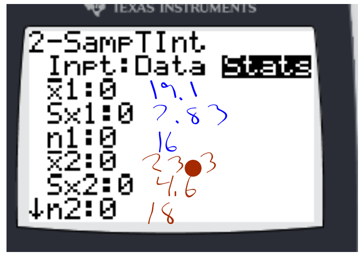
If using Technology for the "DO" step, you must:

- have already named the procedure in the in the "Plan" step
- Give the interval (3.9632, 17.724) for example
- and report the degrees of freedom.

It's a cinch

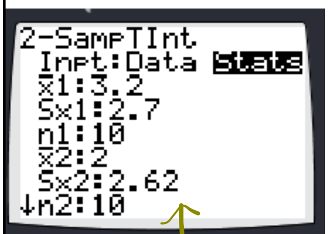
$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

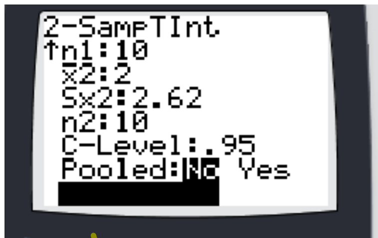
Just kidding

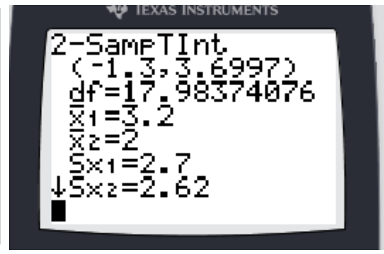


for 2 sample t procedures

No to pooling







example

4. With at least one other person in the class, discuss and confirm tthat the conditions for constructing a confidence interval been met? (don't have to write details)

5. Construct a 95% confidence interval for the true difference in the mean number of chocolate chips in Chips Ahoy and Treasure Chips. (you don't need to write the general or specific formulas in this particular unless you want) but do show the "work".

$\bar{x}_1 =$

$s_1 =$

$n_1 =$

$\bar{x}_2 =$

$s_2 =$

$n_2 =$

2-SampTInt gives

(-6.849, -1.551)

with df = 28.6

Should I Tech or Table?

↑ ↑ ↑
recommended
on FR

The two methods will not produce the same answer

If you write out the formula with numbers substituted in •

- a) leave t^* • in the formula instead of using conservative value •
- b) Then use 2 SampT Int and report the calculator's interval and df.

6. Do we have evidence that there is a difference in the average number of chocolate chips in a Chips Ahoy and a Treasure cookie?

(− , −) → Treasure C has more

(+ , +) → Chips Ahoy has more

(− , +) → No difference

What's expected on the AP exam for CI for diff. of means

State
- Defind both paramters
- State Confidence level

Plan Plan
- Idetify the procedure
- State and check conditions

Do
Calculate the Cofidence Interval

1. State sample $\bar{x}$'s and s_x's

2. Caluculate the interval with option 1 or 2 (but not both)

Option 1 – include *for example:* $2 - SampTInt$ gives $(3.9632, 17.724)$

Option 2 - include *for example:* $df = 15 \quad t^* = 1.26$ and

$$CI = (43.24 - 38.26) \pm 1.26\sqrt{\frac{14.26^2}{30} + \frac{17.50^2}{34}} = \ldots\ldots$$

It's not a must to show the general & specific formula but the longer you continue the better you will know it for M/C (and ch. 12)

Conclude

- Report Confidence level
- Report *computed interval*
- Give interpretation in context

Now Put it all together

Pulse Rates

Pulse Rates:

Mr. Cedarlund's class performed an experiment to investigate whether drinking a caffeinated beverage would increase pulse rates. Twenty students in the class volunteered to take part in the experiment. All of the students measured their initial pulse rates (in beats per minute). Then Mr. Wilcox randomly assigned the students into two groups of 10. Each student in the first group drank 12 ounces of cola with caffeine. Each student in the second group drank 12 ounces of caffeine-free cola. All students then measured their pulse rates again. The table displays the change in pulse rate for the students in both groups.

	Change in pulse rate (Final pulse rate − Initial pulse rate)										Mean change
Caffeine	8	3	5	1	4	0	6	1	4	0	3.2
No caffeine	3	−2	4	−1	5	5	1	2	−1	4	2.0

Construct and interpret a 95% confidence interval for the difference in true mean change in pulse rate for subjects like these who drink caffeine versus who drink no caffeine.

Pulse Rates:

Mr. Cedarlund's class performed an experiment to investigate whether drinking a caffeinated beverage would increase pulse rates. Twenty students in the class volunteered to take part in the experiment. All of the students measured their initial pulse rates (in beats per minute). Then Mr. Cedarlund randomly assigned the students into two groups of 10. Each student in the first group drank 12 ounces of cola with caffeine. Each student in the second group drank 12 ounces of caffeine-free cola. All students then measured their pulse rates again. The table displays the change in pulse rate for the students in both groups.

	Change in pulse rate (Final pulse rate − Initial pulse rate)										Mean change	
Caffeine	8	3	5	1	4	0	6	1	4	0	3.2	$s_1 = 2.70$
No caffeine	3	−2	4	−1	5	5	1	2	−1	4	2.0	$s_2 = 2.62$

Construct and interpret a 95% confidence interval for the difference in true mean change in pulse rate for subjects like these who drink caffeine versus who drink no caffeine.

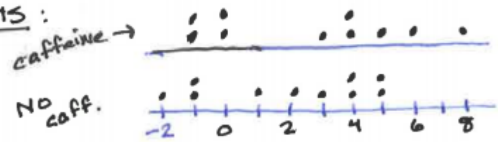
$\boxed{\text{STATE}}$ 95% CI for $\mu_1 - \mu_2$ where μ_1 = the true mean change in pulse rate for students like these after drinking 12 oz. of cola w/caffeine

and μ_2 = the true mean change in pulse rate for students like these after drinking 12 oz of caffeine free cola.

PLAN Two-Sample t interval for $\mu_1 - \mu_2$

✓ Random
Randomly assigned to groups

✓ Large Counts :



No strong skew / No outliers

DO $\bar{x}_1 = 3.2$ $\bar{x}_2 = 2$
$s_1 = 2.70$ $s_2 = 2.62$
$n = 10$ $n = 10$

2 - Samp T Int gives $(-1.302, 3.702)$
using $df = 17.986$

Option 2

$df = 9$ $t^* = 2.262$

$$(3.2 - 2) \pm 2.262 \sqrt{\frac{2.70^2}{10} + \frac{2.62^2}{10}}$$

$$1.2 \pm 2.691$$

$$(-1.491, 3.891)$$

CONCLUDE

We are 95% confident that the interval from −1.302 to 3.702 beats per minute captures $\mu_1 - \mu_2$, the difference in the true mean change in pulse rate for for all students like these after drinking 12 oz of cola with caffeine versus w/o caffeine.

10.2.....37, 39, 41, 45, 49

study pp. 645 - 654