Chapter ~~19~~ Objectives

- Describe the conditions necessary for inference
- Check the conditions necessary for inference
- Perform two-sample *t* procedures
- Describe the robustness of the *t* procedures

Two-Sample Problems

Comparing two populations or two treatments is one of the most common situations encountered in statistical practice. We call such situations two–sample problems.
Suppose we want to compare the mean of some quantitative variable for the individuals in two populations—Population 1 and Population 2.

Unlike the matched pairs designs studied earlier, there is no matching of the individuals in the two samples, and the two samples can be of different sizes. Inference procedures for two–sample data differ from those for matched pairs.

Our parameters of interest are the population means μ_1 and μ_2. The best approach is to take separate random samples from each population and to compare the sample means.
We use the mean response in the two groups to make the comparison.
Here's a table that summarizes these two situations:

Population or treatment	Parameter	Statistic	Sample size
1	μ_1	$\bar{x}_1$	n_1
2	μ_2	$\bar{x}_2$	n_2

Two Sample Problems

To study the change in attitude towards statistics over the course of the semester, a professor selected a simple random sample of students. She administered a questionnaire to the students at the beginning and again at the end of the semester.

What type of problem is this?

a) Matched pairs
b) Two independent samples

Two Sample Problems

The National Park Service compared the average amount of money spent by visitors in two different national parks. They sampled visitors on the same day in each of the two parks.

What type of problem is this?

a) Matched pairs
b) Two independent samples

Conditions for Inference Comparing Two Means

Conditions for Inference Comparing Two Means

- We have two SRSs from two distinct populations. The samples are independent. That is, one sample has no influence on the other. Matching violates independence, for example. We measure the same response variable for both samples.
- Both populations are Normally distributed. The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.

The Two-Sample *t* Statistic

When data come from two random samples or two groups in a randomized experiment, the statistic $\bar{x}_1 - \bar{x}_2$ is our best guess for the value of $\mu_1 - \mu_2$.

When the Independent condition is met, the standard deviation of the statistic $\bar{x}_1 - \bar{x}_2$ is:

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

Since we don't know the values of the parameters σ_1 and σ_2, we replace them in the standard deviation formula with the sample standard deviations. The result is the **standard error** of the statistic $\bar{x}_1 - \bar{x}_2$:

$$\sqrt{\frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}}$$

If the Normal condition is met, we standardize the observed difference to obtain a *t* statistic that tells us how far the observed difference is from its mean in standard deviation units:

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

The two-sample *t* statistic has approximately a *t* distribution. We can use technology to determine degrees of freedom OR we can use the smaller of $n_1 - 1$ *and* $n_2 - 1$ for the degrees of freedom.

Confidence Interval for μ_1 - μ_2

Two-Sample *t* Interval for a Difference Between Means

When the Random, Normal, and Independent conditions are met, a level C confidence interval for $(\mu_1 - \mu_2)$ is

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

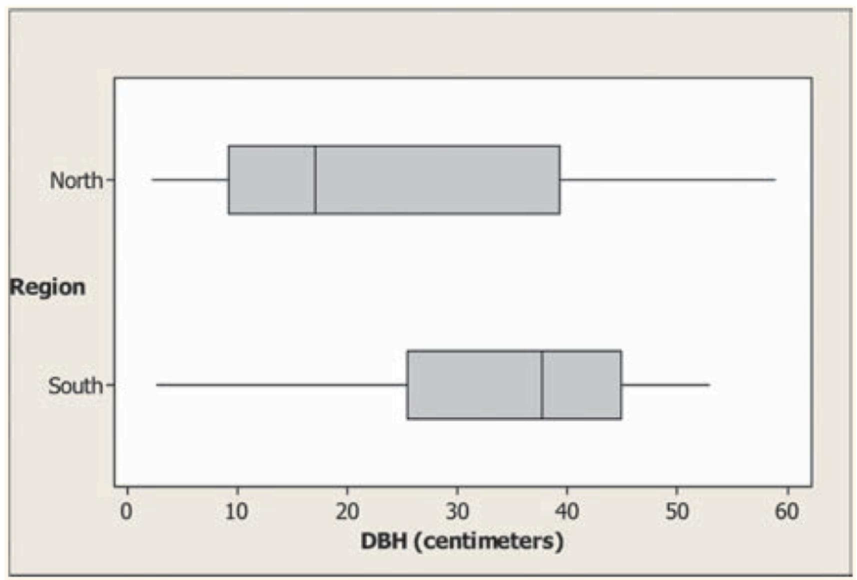
where t^* is the critical value for confidence level C for the t distribution with degrees of freedom from either technology or the smaller of $n_1 - 1$ and $n_2 - 1$.

Example

The Wade Tract Preserve in Georgia is an old-growth forest of longleaf pines that has survived in a relatively undisturbed state for hundreds of years. One question of interest to foresters who study the area is “How do the sizes of longleaf pine trees in the northern and southern halves of the forest compare?” To find out, researchers took random samples of 30 trees from each half and measured the diameter at breast height (DBH) in centimeters. Comparative boxplots of the data and summary statistics from Minitab are shown below. Construct and interpret a 90% confidence interval for the difference in the mean DBH for longleaf pines in the northern and southern halves of the Wade Tract Preserve.

Descriptive Statistics: North, South

Variable	N	Mean	StDev
North	30	23.70	17.50
South	30	34.53	14.26



Example

State: Our parameters of interest are μ_1 = the true mean DBH (diameter at breast height) of all trees in the southern half of the forest and μ_2 = the true mean DBH of all trees in the northern half of the forest. We want to estimate the difference $\mu_1 - \mu_2$ at a 90% confidence level.

Plan: We should use a two-sample *t* interval for $\mu_1 - \mu_2$ if the conditions are satisfied.

✓ **Random** The data come from random samples of 30 trees, one from the northern half and one from the southern half of the forest.

✓ **Normal** The boxplots give us reason to believe that the population distributions of DBH measurements may not be Normal.

However, because both sample sizes are at least 30, we are safe using *t* procedures.

✓ **Independent** Researchers took independent samples from the northern and southern halves of the forest.

Example

Do: Since the conditions are satisfied, we can construct a two-sample *t* interval for the difference $\mu_1 - \mu_2$. We'll use the conservative *df* = 30 – 1 = 29.

Upper-tail probability *p*

df	.10	.05	.025
28	1.313	1.70	2.048
29	1.311	1.699	2.045
30	1.310	1.697	2.042
	80%	90%	95%

Confidence level *C*

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (34.5 - 23.70) \pm 1.699\sqrt{\frac{14.26^2}{30} + \frac{17.50^2}{30}}$$

$$= 10.83 \pm 7.00 = (3.83,\ 17.83)$$

Conclude: We are 90% confident the **true difference in mean** DBH of the southern trees and mean DBH of the northern trees is between 3.83 and 17.83 centimeters.

Two-Sample *t* Test

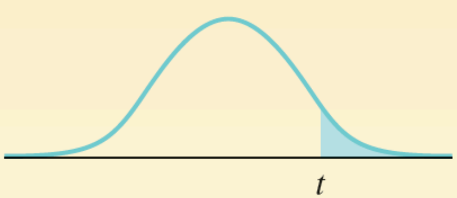
Two-Sample *t* Test for the Difference Between Two Means

Suppose the Random, Normal, and Independent conditions are met. To test the hypothesis $H_0: \mu_1 - \mu_2 =$ hypothesized value, compute the t statistic

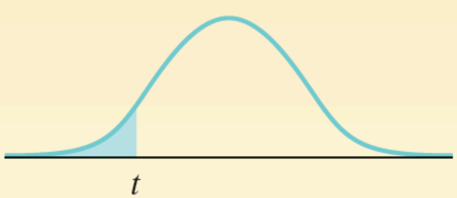
$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Find the P-value by calculating the probabilty of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.

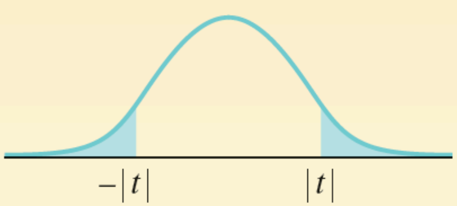
$H_a: \mu_1 - \mu_2 >$ hypothesized value



$H_a: \mu_1 - \mu_2 <$ hypothesized value



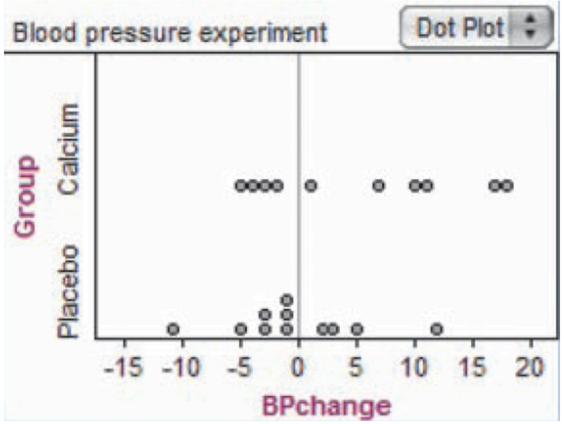
$H_a: \mu_1 - \mu_2 \neq$ hypothesized value



Example

Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. Researchers designed a randomized comparative experiment. The subjects were 21 healthy men who volunteered to take part in the experiment. They were randomly assigned to two groups: 10 of the men received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The experiment was double-blind. The response variable is the decrease in systolic (top number) blood pressure for a subject after 12 weeks, in millimeters of mercury. An increase appears as a negative response. Here are the data:

Group 1 (calcium):	7	−4	18	17	−3	−5	1	10	11	−2	
Group 2 (placebo):	−1	12	−1	−3	3	−5	5	2	−11	−1	−3



Example

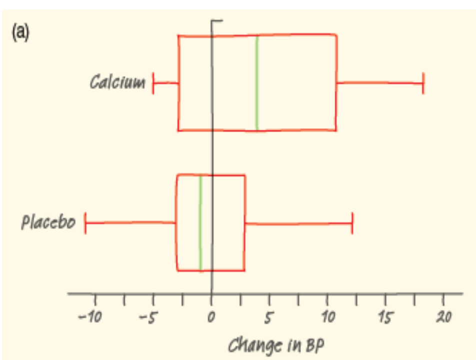
State: We want to perform a test of:

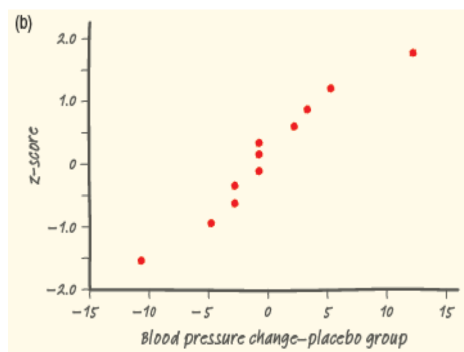
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

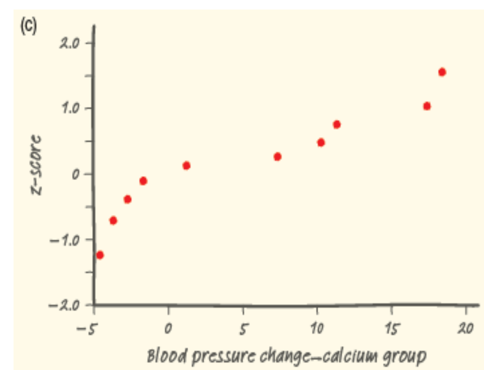
where μ_1 = the true mean decrease in systolic blood pressure for healthy men like the ones in this study who take a calcium supplement, and μ_2 = the true mean decrease in systolic blood pressure for healthy men like the ones in this study who take a placebo. We will use $\alpha = 0.05$.

Plan: If conditions are met, we will carry out a two-sample t test for $\mu_1 - \mu_2$.

- **Random** The 21 subjects were randomly assigned to the two treatments.
- **Normal** Boxplots and Normal probability plots for these data are below:







The boxplots show no clear evidence of skewness and no outliers. With no outliers or clear skewness, the t procedures should be pretty accurate.

- **Independent** Due to the random assignment, these two groups of men can be viewed as independent.

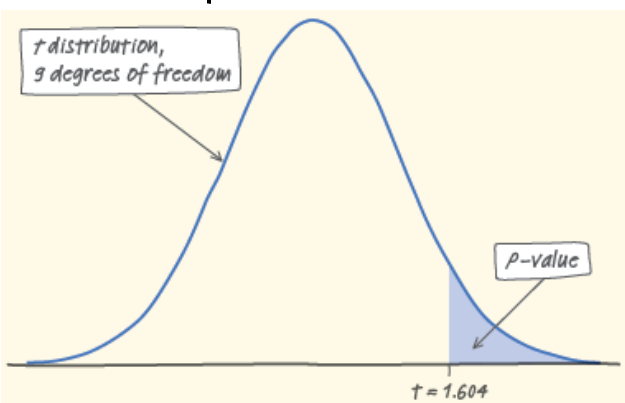
Example

Do: Since the conditions are satisfied, we can perform a two-sample *t* test for the difference $\mu_1 - \mu_2$.

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{[5.000 - (-0.273)] - 0}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = 1.604$$

	Upper tail probability *p*		
df	**.10**	**.05**	**.025**
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228



***P*-value** Using the conservative *df* = 10 – 1 = 9, we can use Table B to show that the *P*-value is between 0.05 and 0.10.

Conclude: Because the *P*-value is greater than α = 0.05, we fail to reject H_0. There is no evidence that calcium reduces blood pressure, but the evidence is not convincing enough to conclude that calcium reduces blood pressure more than a placebo

Example

Community service and attachment to friends

STATE: Do college students who have volunteered for community service work differ from those who have not? A study obtained data from 57 students who had done service work and 17 who had not. One of the response variables was a measure of "attachment to friends," with larger values indicating greater attachment. Here are the results:

Group	Condition	n	$\bar{x}$	s
1	Service	57	105.32	14.68
2	No service	17	96.82	14.26

PLAN: The investigator had no specific direction for the difference in mind before looking at the data, so the alternative is two-sided. We will test the hypotheses

$$H_0\colon \mu_1 = \mu_2$$
$$H_a\colon \mu_1 \neq \mu_2$$

SOLVE: The investigator says that the individual scores, examined separately in the two samples, appear roughly Normal. There is a serious problem with the more important condition that the two samples can be regarded as SRSs from two student populations. We will discuss that after we illustrate the calculations.

The two–sample *t* statistic is

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{105.32 - 96.82}{\sqrt{\frac{14.68^2}{57} + \frac{14.26^2}{17}}} = \frac{8.5}{3.9677} = 2.142$$

The degrees of freedom is 16, the smaller of

$$n_1 - 1 = 57 - 1 = 56$$
$$n_2 - 1 = 17 - 1 = 16$$

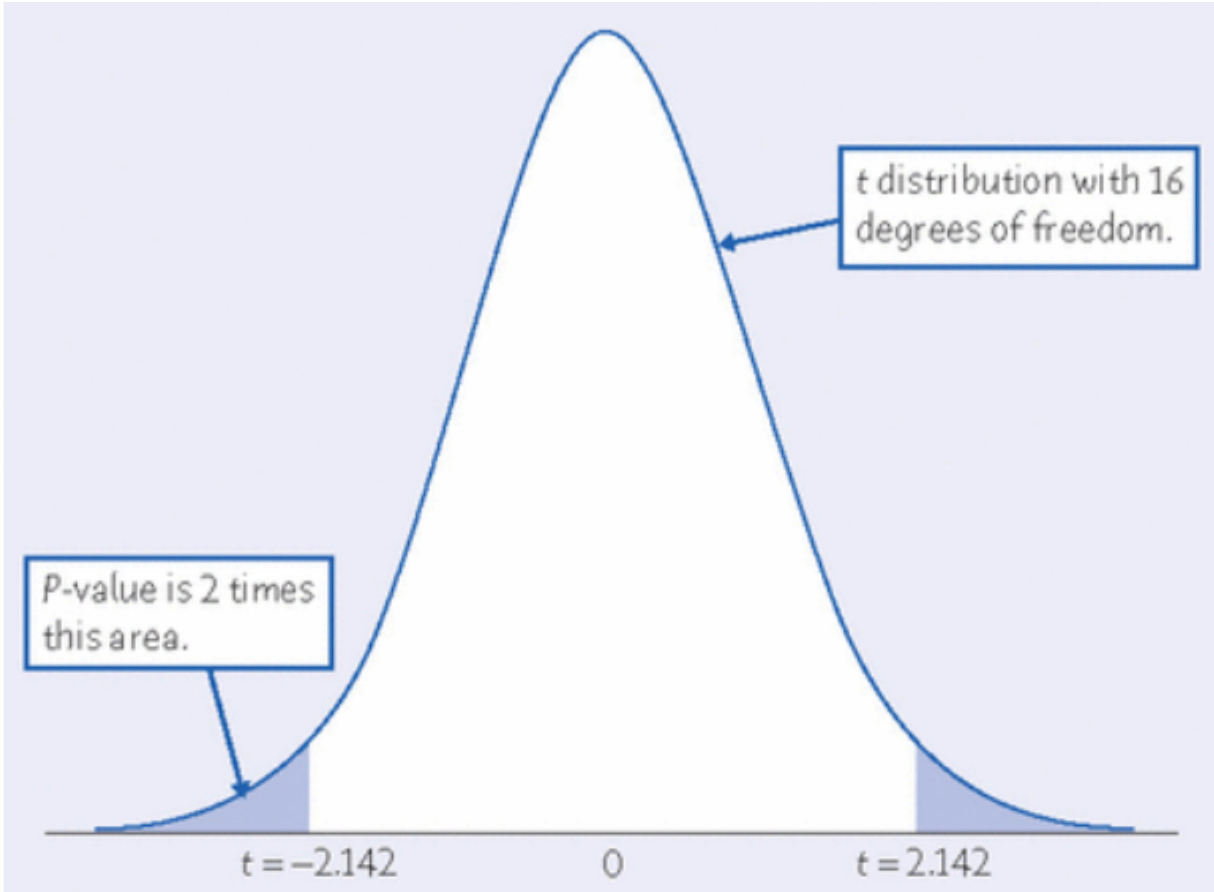
How we can find the P-value?

t = 2.142 and df = 16, the alternative hypothesis is two-sided $H_a: \mu_1 \neq \mu_2$

df = 16

t^*	2.120	2.235
Two-sided *P*	.05	.04

Table C shows that the *P*–value is between 0.05 and 0.04.



CONCLUDE: The data give moderately strong evidence ($P < 0.05$) that students who have engaged in community service are, on the average, more attached to their friends.

Robustness Again

The two-sample *t* procedures are more robust than the one-sample *t* methods, particularly when the distributions are not symmetric.

Using the *t* Procedures

- Except in the case of small samples, the condition that the data are SRSs from the populations of interest is more important than the condition that the population distributions are Normal.
- *Sum of the sample sizes less than 15:* Use *t* procedures if the data appear close to Normal. If the data are clearly skewed or if outliers are present, do not use *t*.
- *Sum of the sample size at least 15:* The *t* procedures can be used except in the presence of outliers or strong skewness.
- *Large samples:* The *t* procedures can be used even for clearly skewed distributions when the sum of the sample sizes is large.

Is the t procedure safe for these data?

STATE: Do obese people average more time lying down than lean people?

PLAN: We test

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

where μ_1 is the mean time spent lying down for the lean group, and μ_2 is the mean time for the obese group.

SOLVE: We assume that the data come from SRSs of the two populations.

Back-to-back stemplots do not indicate non Normal data.

Lean		Obese
9	3	
	4	1
	4	
5	4	44
6	4	6
8	4	
10	5	001
33	5	23
5	5	
6	5	6

Thus, we can use the t test for two samples.

We can calculate the sample mean and sample standard deviation for each sample:

$$\overline{x_1} = 501{,}6461$$
$$\overline{x_2} = 491{,}7426$$
$$s_1 = 52.0449$$
$$s_2 = 46.5932$$

Also the size of each sample is 10, thus

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{501.6461 - 491.7426}{22.0898} = 0.448$$

Using df as the smaller of

$$n_1 - 1 = 10 - 1 = 9$$
$$n_2 - 1 = 10 - 1 = 9$$

df = 9

From table C P-value > 0.50 (using software, P = 0.6596). P-value is very large, there is no evidence to reject the null hypothesis.

CONCLUDE: There is no evidence to support a conclusion that lean people spend less time laying down (on average) than obese people.

Two-Sample t-Procedures

A dairy scientist compared milk production of cows fed 2 diets. He randomly divided a set of 10 cows into two groups. One group was fed diet 1 for a month and the other was fed diet 2 for a month. Here are the milk production results (lb/week):

group	n	$\bar{x}$	s
diet 1	10	385.7	25.7
diet 2	10	398.2	43.1

What is the standard error of $\bar{x}_1 - \bar{x}_2$?

a) $25.7/10 + 43.1/10$

b) $\sqrt{25.7/10 - 43.1/10}$

c) $\sqrt{25.7^2/10 + 43.1^2/10}$

d) $\sqrt{25.7^2/10 - 43.1^2/10}$

Two-Sample *t*-Procedures

A dairy scientist compared milk production of cows fed 2 diets. He randomly divided a set of 10 cows into two groups. One group was fed diet 1 for a month and the other was fed diet 2 for a month. Here are the milk production results (lb/week):

group	n	$\bar{x}$	s
diet 1	10	385.7	25.7
diet 2	10	398.2	43.1

Using Option 2, the 2-sample *t*-statistic for this situation has an approximate *t*-distribution with how many degrees of freedom?

a) 9
b) 10
c) 18
d) 20

Two-Sample *t*-Procedures

A dairy scientist compared milk production of cows fed 2 diets. He randomly divided a set of 10 cows into two groups. One group was fed diet 1 for a month and the other was fed diet 2 for a month. Here are the milk production results (lb/week):

group	n	$\bar{x}$	s
diet 1	10	385.7	25.7
diet 2	10	398.2	43.1

What is the 2-sample *t*-statistic for testing H_0: $\mu_1 = \mu_2$?

a) $385.7 + 398.2) \big/ \sqrt{25.7^2 / 10 + 43.1^2 / 10}$

b) $385.7 + 398.2) \big/ \sqrt{25.7^2 / 10 - 43.1^2 / 10}$

c) $385.7 - 398.2) \big/ \sqrt{25.7^2 / 10 + 43.1^2 / 10}$

d) $385.7 - 398.2) \big/ \sqrt{25.7^2 / 10 - 43.1^2 / 10}$

Two-Sample *t*-Procedures

A dairy scientist compared milk production of cows fed 2 diets. He randomly divided a set of 10 cows into two groups. One group was fed diet 1 for a month and the other was fed diet 2 for a month. Here are the milk production results (lb/week):

group	n	$\bar{x}$	s
diet 1	10	385.7	25.7
diet 2	10	398.2	43.1

The 2-sample *t*-statistic for testing H_0: $\mu_1 = \mu_2$ is -0.79. If the test is one-sided, what is the *P*-value?

a) 0.40 to 0.50
b) 0.20 to 0.25
c) 0.15 to 0.20
d) 0.10 to 0.15

TABLE C *t* distribution critical values

d.f.				
8	0.703	0.883	1.100	1.383
9	0.700	0.897	1.093	1.372
10	0.697	0.876	1.088	1.363
1-sided P	.25	.20	.15	.10