Chapter 10: Comparing Two Populations or Groups

Section 10.2
Comparing Two Means

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Chapter 10
Comparing Two Populations or Groups

- 10.1 Comparing Two Proportions
- **10.2 Comparing Two Means**

+

Section 10.2
Comparing Two Means

Learning Objectives

After this section, you should be able to…

- ✓ DESCRIBE the characteristics of the sampling distribution of the difference between two sample means
- ✓ CALCULATE probabilities using the sampling distribution of the difference between two sample means
- ✓ DETERMINE whether the conditions for performing inference are met
- ✓ USE two-sample *t* procedures to compare two means based on summary statistics or raw data
- ✓ INTERPRET computer output for two-sample *t* procedures
- ✓ PERFORM a significance test to compare two means
- ✓ INTERPRET the results of inference procedures

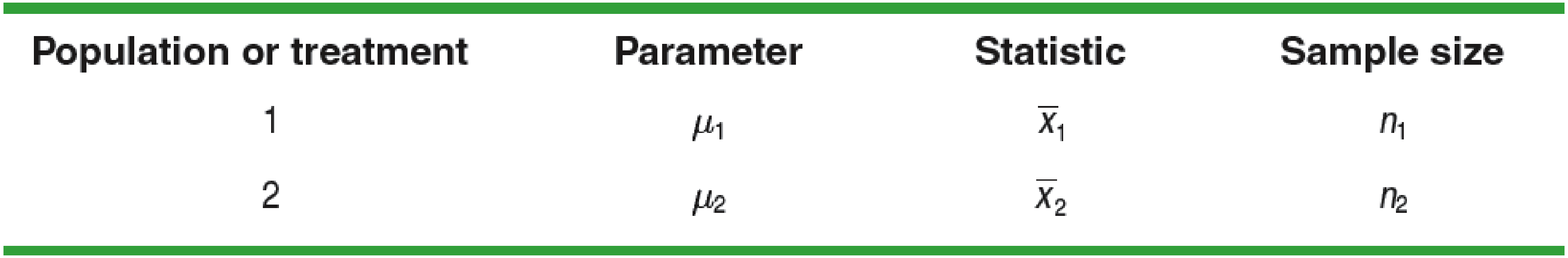
Introduction

In the previous section, we developed methods for comparing two proportions. What if we want to compare the mean of some quantitative variable for the individuals in Population 1 and Population 2?

Our parameters of interest are the population means μ_1 and μ_2. Once again, the best approach is to take separate random samples from each population and to compare the sample means.

Suppose we want to compare the average effectiveness of two treatments in a completely randomized experiment. In this case, the parameters μ_1 and μ_2 are the true mean responses for Treatment 1 and Treatment 2, respectively. 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

The Sampling Distribution of a Difference Between Two Means

In Chapter 7, we saw that the sampling distribution of a sample mean has the following properties:

Shape Approximately Normal if the population distribution is Normal or *n* ≥ 30 (by the central limit theorem).

Center $\mu_{\bar{x}} = \mu$

Spread $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ if the sample is no more than 10% of the populatior

To explore the sampling distribution of the difference between two means, let's start with two Normally distributed populations having known means and standard deviations.

Based on information from the U.S. National Health and Nutrition Examination Survey (NHANES), the heights (in inches) of ten-year-old girls follow a Normal distribution *N*(56.4, 2.7). The heights (in inches) of ten-year-old boys follow a Normal distribution *N*(55.7, 3.8).

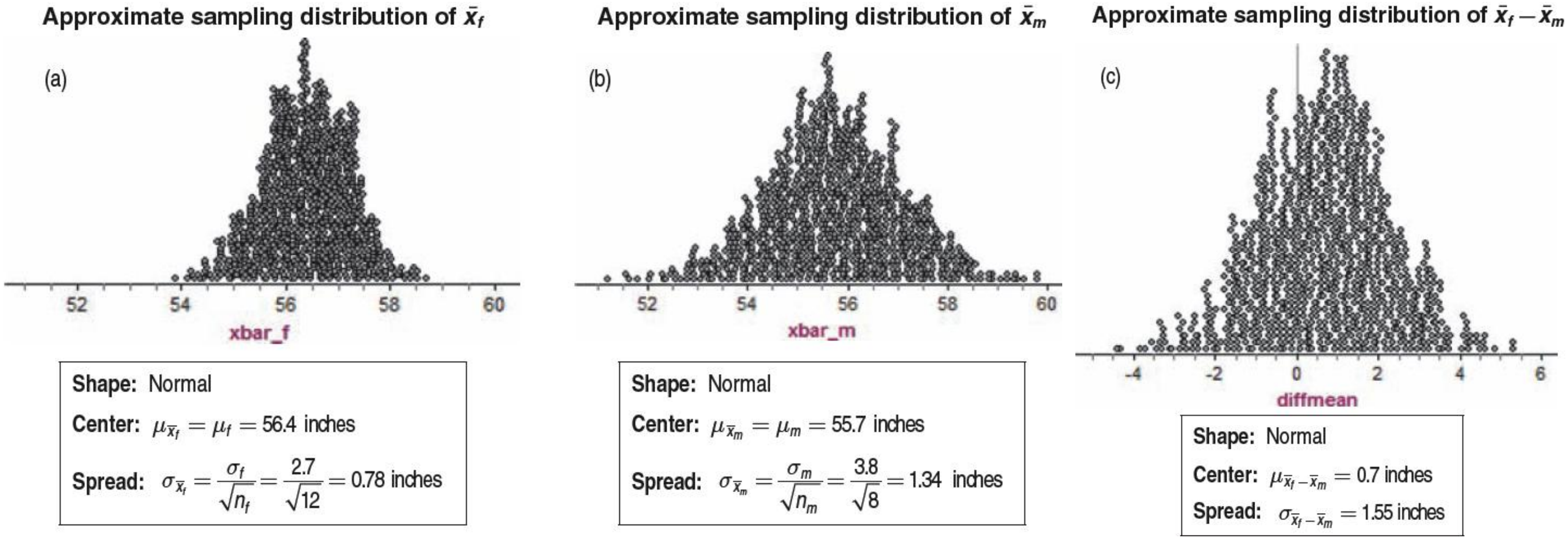
Suppose we take independent SRSs of 12 girls and 8 boys of this age and measure their heights.

What can we say about the difference $\bar{x}_f - \bar{x}_m$ in the average heights of the sample of girls and the sample of boys?

The Sampling Distribution of a Difference Between Two Means

Comparing Two Means

Using Fathom software, we generated an SRS of 12 girls and a separate SRS of 8 boys and calculated the sample mean heights. The difference in sample means was then calculated and plotted. We repeated this process 1000 times. The results are below:



Shape: Normal

Center: $\mu_{\bar{x}_f} = \mu_f = 56.4$ inches

Spread: $\sigma_{\bar{x}_f} = \frac{\sigma_f}{\sqrt{n_f}} = \frac{2.7}{\sqrt{12}} = 0.78$ inches

Shape: Normal

Center: $\mu_{\bar{x}_m} = \mu_m = 55.7$ inches

Spread: $\sigma_{\bar{x}_m} = \frac{\sigma_m}{\sqrt{n_m}} = \frac{3.8}{\sqrt{8}} = 1.34$ inches

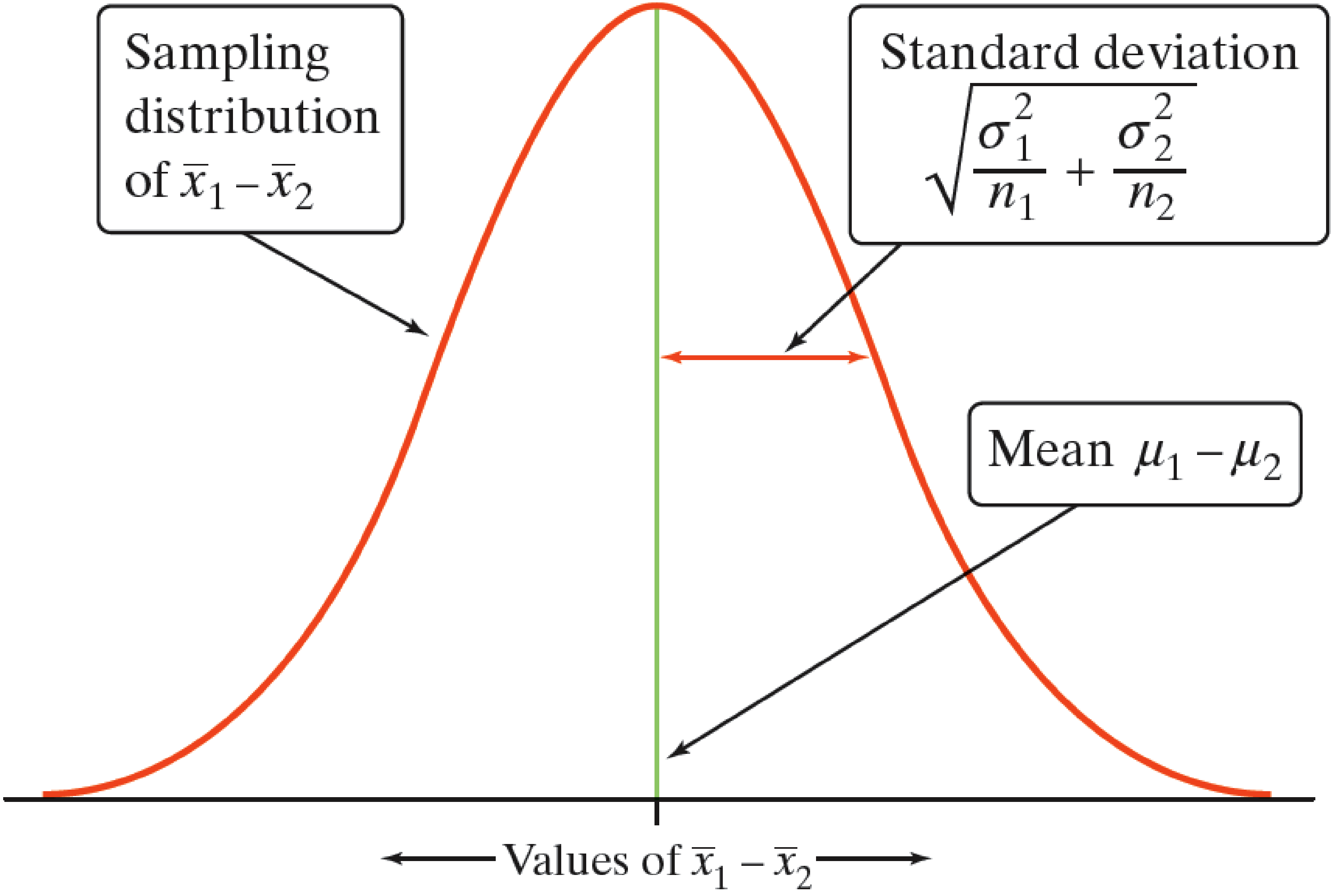
Shape: Normal

Center: $\mu_{\bar{x}_f - \bar{x}_m} = 0.7$ inches

Spread: $\sigma_{\bar{x}_f - \bar{x}_m} = 1.55$ inches

What do you notice about the shape, center, and spread of the sampling distribution of $\bar{x}_f - \bar{x}_m$?

- **The Sampling Distribution of a Difference Between Two Means**


Sampling distribution of $\bar{x}_1 - \bar{x}_2$
Standard deviation $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Mean $\mu_1 - \mu_2$
Values of $\bar{x}_1 - \bar{x}_2$

Comparing Two Means

Example: Who’s Taller at Ten, Boys or Girls?

- Based on information from the U.S. National Health and Nutrition Examination Survey (NHANES), the heights (in inches) of ten-year-old girls follow a Normal distribution *N*(56.4, 2.7)*.* The heights (in inches) of ten-year-old boys follow a Normal distribution *N*(55.7*,* 3.8)*.* A researcher takes independent SRSs of 12 girls and 8 boys of this age and measures their heights. After analyzing the data, the researcher reports that the sample mean height of the boys is larger than the sample mean height of the girls.

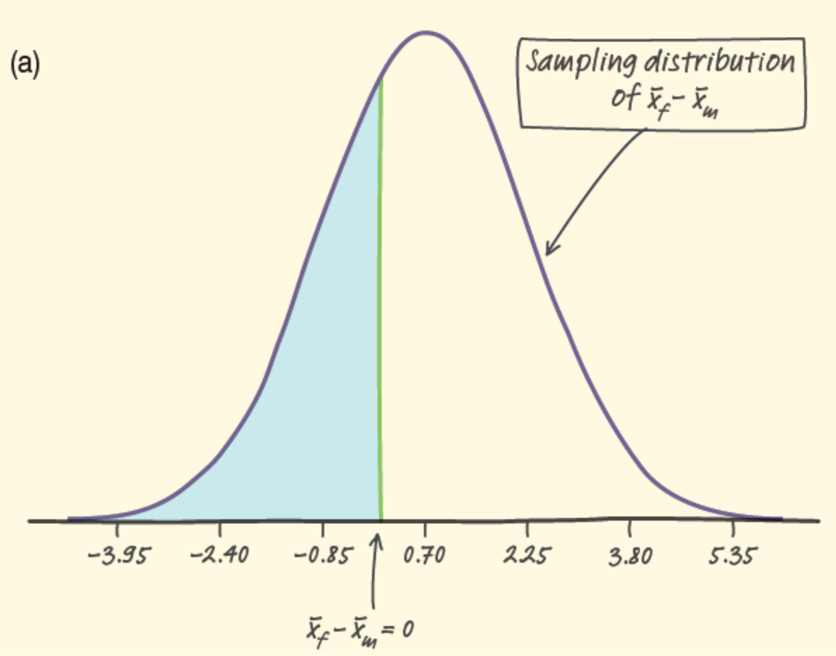
a) Describe the shape, center, and spread of the sampling distribution of $\bar{x}_f - \bar{x}_m$.

Because both population distributions are Normal, the sampling distribution of $\bar{x}_f - \bar{x}_m$ is Normal.

Its mean is $\mu_f - \mu_m = 56.4 - 55.7 = 0.7$ inches

Its standard deviation is

$$\sqrt{\frac{2.7^2}{12} + \frac{3.8^2}{8}} = 1.55 \text{ inches}$$



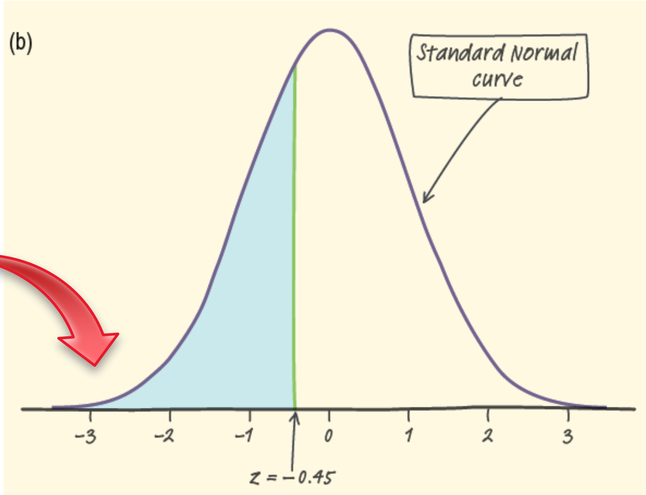
Example: Who's Taller at Ten, Boys or Girls?

b) Find the probability of getting a difference in sample means $\bar{x}_1 - \bar{x}_2$ that is less than 0.

Standardize: When $\bar{x}_1 - \bar{x}_2 = 0$,

$$z = \frac{0 - 0.70}{1.55} = -0.45$$

Use Table A: The area to the left of $z = -0.45$ under the standard Normal curve is 0.3264.



(c) Does the result in part (b) give us reason to doubt the researchers' stated results?

If the mean height of the boys is greater than the mean height of the girls, $\bar{x}_m > \bar{x}_f$, That is $\bar{x}_f - \bar{x}_m < 0$. Part (b) shows that there's about a 33% chance of getting a difference in sample means that's negative just due to sampling variability. This gives us little reason to doubt the researcher's claim.

Alternate Example: Potato chips

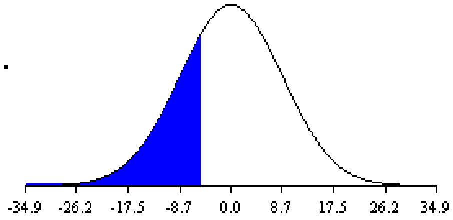
- A potato chip manufacturer buys potatoes from two different suppliers, Riderwood Farms and Camberley, Inc. The weights of potatoes from Riderwood Farms are approximately Normally distributed with a mean of 175 grams and a standard deviation of 25 grams. The weights of potatoes from Camberley, Inc. are approximately Normally distributed with a mean of 180 grams and a standard deviation of 30 grams. When shipments arrive at the factory, inspectors randomly select a sample of 20 potatoes from each shipment and weigh them. They are surprised when the average weight of the potatoes in the sample from Riderwood Farms $\bar{x}_r$ was higher than the average weight of the potatoes in the sample from Camberley, Inc. $\bar{x}_C$.

a) Describe the shape, center, and spread of the sampling distribution of $\bar{x}_c - \bar{x}_r$.

Because both population distributions are Normal, the sampling distribution of $\bar{x}_c - \bar{x}_r$ is Normal. Its mean is $\mu_c - \mu_r = 180 - 175 = 5$ grams. Its standard deviation is $\sqrt{\frac{25^2}{20} + \frac{30^2}{20}} = 8.73$ grams.

(b) Find the probability that the mean weight of the Riderwood sample is larger than the mean weight of the Camberley sample. Should the inspectors have been surprised?

If the mean of the Riderwood sample is larger, then $\bar{x}_c - \bar{x}_r$ must be negative. The graph shows the sampling distribution with the desired probability shaded.



$P(\bar{x}_c - \bar{x}_r < 0) = P\left(z < \frac{0-5}{8.73}\right) = P(z < -0.57) = 0.28.$ The inspectors shouldn't be surprised, because the Riderwood sample will have a higher mean more than one-fourth of the time.

Comparing Two Means

The Two-Sample *t* Statistic

When data come from two random samples or two groups in a randomize 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 resu

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:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}}$$

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

Confidence Intervals for $\mu_1 - \mu_2$

Comparing Two Means

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

When the Random, Normal, and Independent conditions are met, an approximate level C confidence interval for $(\bar{x}_1 - \bar{x}_2)$ 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 $n_2 - 1$.

Random The data are produced by a random sample of size n_1 from Population 1 and a random sample of size n_2 from Population 2 or by two groups of size n_1 and n_2 in a randomized experiment.

Normal Both population distributions are Normal OR both sample group sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$).

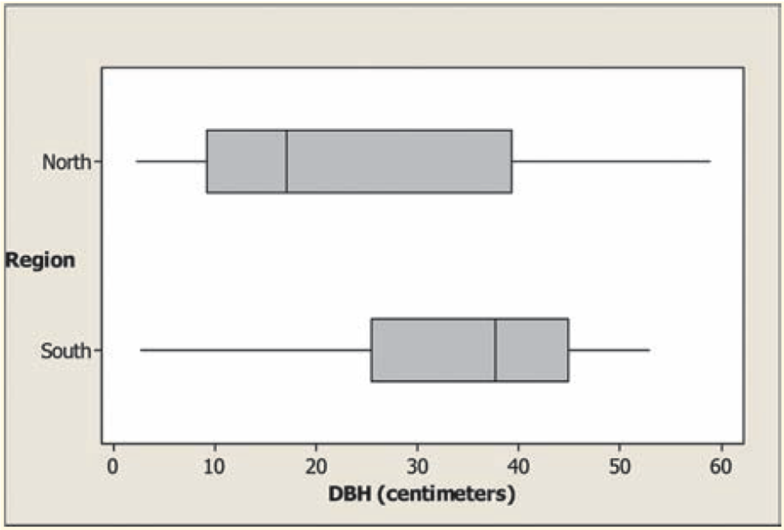
Independent Both the samples or groups themselves and the individual observations in each sample or group are independent. When sampling without replacement, check that the two populations are at least 10 times as large as the corresponding samples (the 10% condition).

Big Trees, Small Trees, Short Trees, Tall Trees

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
```



State: Our parameters of interest are μ_1 = the true mean DBH 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.

Big Trees, Small Trees, Short Trees, Tall Trees

Comparing Two Means

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

✓ **Random** The data come from a random samples of 30 trees each from the northern and southern halves of the forest.

✓ **Normal** The boxplots give us reason to believe that the population distributions of DBH measurements may not be Normal. However, since 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. Because sampling without replacement was used, there have to be at least 10(30) = 300 trees in each half of the forest. This is pretty safe to assume.

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.

Conclude: We are 90% confident that the interval from 3.83 to 17.83 centimeters captures the difference in the actual mean DBH of the southern trees and the actual mean DBH of the northern trees. This interval suggests that the mean diameter of the southern trees is between 3.83 and 17.83 cm larger than the mean diameter of the northern trees.

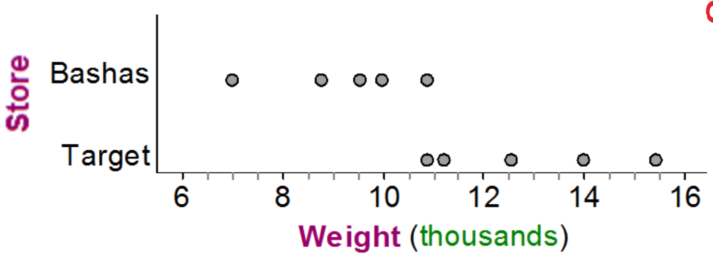
Alternate Example – Plastic grocery bags

Do plastic bags from Target or plastic bags from Bashas hold more weight? A group of AP Statistic students decided to investigate by filling a random sample of 5 bags from each store with common grocery items until the bags ripped. Then they weighed the contents of items in each bag to determine its capacity. Here are their results, in grams:

Target:	12,572	13,999	11,215	15,447	10,896
Bashas:	9552	10,896	6983	8767	9972

Problem:

(a) Construct and interpret a 99% confidence interval for the difference in mean capacity of plastic grocery bags from Target and Bashas.



State: We want to estimate μ_T - μ_B at the 99% confidence level where μ_T = the mean capacity of plastic bags from Target (in grams) and μ_B = the mean capacity of plastic bags from Bashas (in grams).

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

✓ **Random** The students selected a random sample of bags from each store.

✓ **Normal** Since the sample sizes are small, we must graph the data to see if it is reasonable to assume that the population distributions are approximately Normal. Since there is no obvious skewness or outliers, it is safe to use t procedures.

✓ **Independent** The samples were selected independently and it is reasonable to assume that there are more than 10(5) = 50 plastic grocery bags at each store.

STEP 4

Alternate Example – Plastic grocery bags

Do: For these data, $\bar{X}_T = 12{,}825.8$, $S_T = 1912.5$, $\bar{X}_B = 9234$, $S_B = 1474.2$. Using the conservative df of $5 - 1 = 4$, the critical value for 99% confidence is $t^* = 4.604$. Thus, the confidence interval is

$$(12{,}826\text{-}9234) \pm 4.604\sqrt{\frac{1474.2^2}{5} + \frac{1912.5^2}{5}} = 3592 \pm 4972 = (-1380, 8564).$$

With technology and df = 7.5, CI = (-101, 7285). Notice how much narrower this interval is.

Conclude: We are 99% confident that the interval from –1380 to 8564 grams captures the true difference in the mean capacity for plastic grocery bags from Target and from Bashas.

(b) Does your interval provide convincing evidence that there is a difference in the mean capacity among the two stores?

Since the interval includes 0, it is plausible that there is no difference in the two means. Thus, we do not have convincing evidence that there is a difference in mean capacity. However, if we increased the sample size we would likely find a convincing difference since it seems pretty clear that Target bags have a bigger capacity.

Comparing Two Means

Significance Tests for $\mu_1 - \mu_2$

An observed difference between two sample means can reflect an actual difference in the parameters, or it may just be due to chance variation in random sampling or random assignment. Significance tests help us decide which explanation makes more sense. The null hypothesis has the general form

$$H_0: \mu_1 - \mu_2 = \text{hypothesized value}$$

We’re often interested in situations in which the hypothesized difference is 0. Then the null hypothesis says that there is no difference between the two parameters:

$$\boldsymbol{H_0: \mu_1 - \mu_2 = 0 \text{ or, alternatively, } H_0: \mu_1 = \mu_2}$$

The alternative hypothesis says what kind of difference we expect.

$$\boldsymbol{H_a: \mu_1 - \mu_2 > 0,\ H_a: \mu_1 - \mu_2 < 0, \text{ or } H_a: \mu_1 - \mu_2 \neq 0}$$

If the Random, Normal, and Independent conditions are met, we can proceed with calculations.

Significance Tests for $\mu_1 - \mu_2$

To do a test, standardize $\bar{x}_1 - \bar{x}_2$ to get a two-sample t statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}}$$

To find the *P*-value, use the *t* distribution with degrees of freedom given by technology or by the conservative approach (df = smaller of n_1 - 1 *and* n_2 - 1).

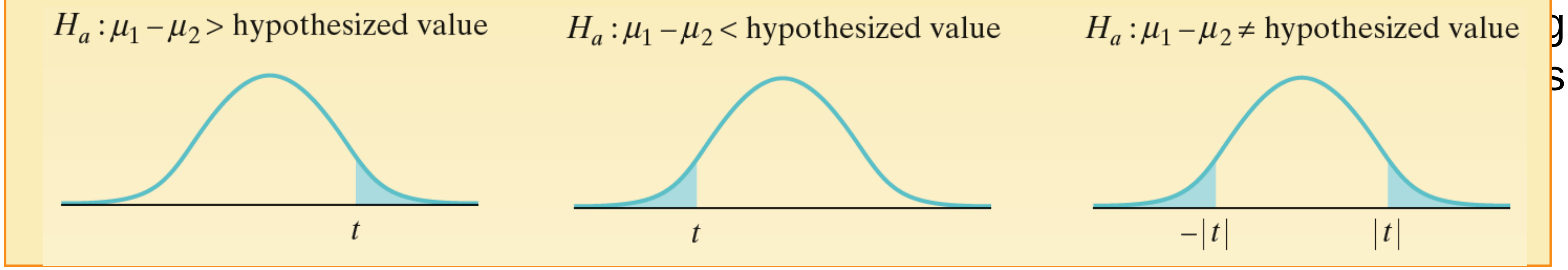
Two-Sample *t* Test for The Difference Between Two Means

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

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}}$$

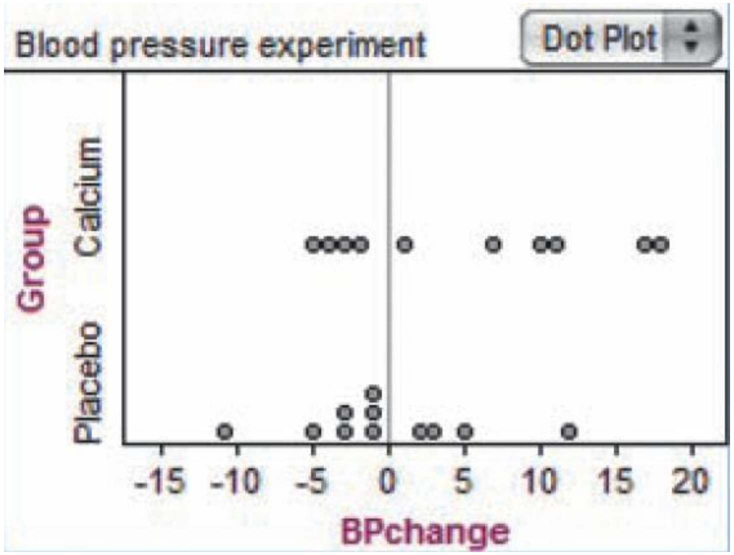
Find the P - value by calculating the probabilty of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.



Example: Calcium and Blood Pressure

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. The relationship was strongest for black men. Such observational studies do not establish causation. Researchers therefore designed a randomized comparative experiment. The subjects were 21 healthy black 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



State: We want to perform a test of

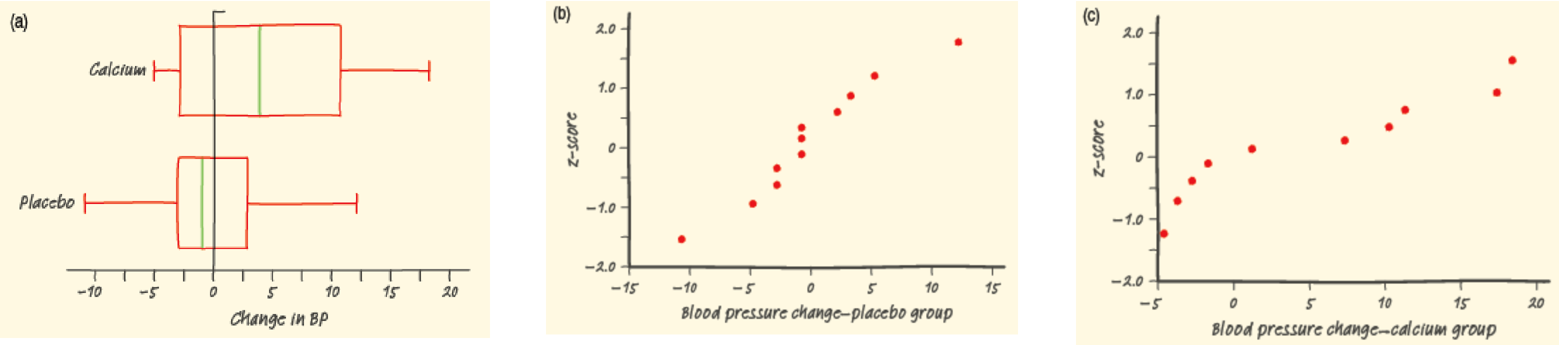
$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 > 0$$

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

Example: Calcium and Blood Pressure

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** With such small sample sizes, we need to examine the data to see if it's reasonable to believe that the actual distributions of differences in blood pressure when taking calcium or placebo are Normal. Hand sketches of calculator boxplots and Normal probability plots for these data are below:



The boxplots show no clear evidence of skewness and no outliers. The Normal probability plot of the placebo group's responses looks very linear, while the Normal probability plot of the calcium group's responses shows some slight curvature. 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. Individual observations in each group should also be independent: knowing one subject's change in blood pressure gives no information about another subject's response.

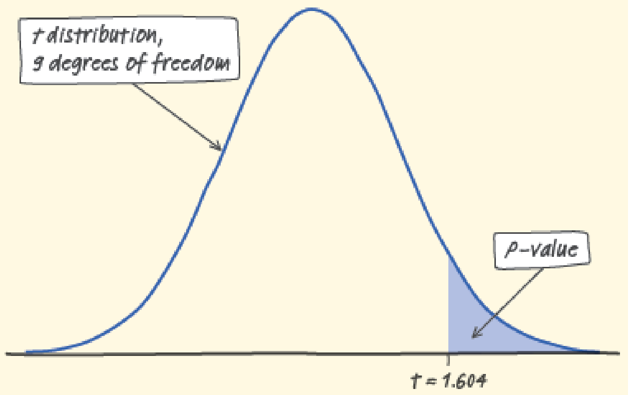
Example: Calcium and Blood Pressure

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 $\alpha = 0.05$, we fail to reject H_0. The experiment provides some evidence that calcium reduces blood pressure, but the evidence is not convincing enough to conclude that calcium reduces blood pressure more than a placebo.

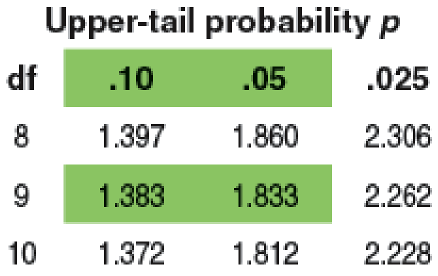
Comparing Two Means

Example: Calcium and Blood Pressure

We can estimate the difference in the true mean decrease in blood pressure for the calcium and placebo treatments using a two-sample *t* interval for $\mu_1 - \mu_2$. To get results that are consistent with the one-tailed test at $\alpha = 0.05$ from the example, we'll use a 90% confidence level. The conditions for constructing a confidence interval are the same as the ones that we checked in the example before performing the two-sample *t* test.

With df = *9,* the critical value for a 90% confidence interval is $t^* = 1.833$.

The interval is:



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

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = [5.000 - (-0.273)] \pm 1.833\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}$$

$$= 5.273 \pm 6.027$$

$$= (-0.754, 11.300)$$

We are 90% confident that the interval from -0.754 to 11.300 captures the difference in true mean blood pressure reduction on calcium over a placebo. Because the 90% confidence interval includes 0 as a plausible value for the difference, we cannot reject $H_0: \mu_1 - \mu_2 = 0$ against the two-sided alternative at the $\alpha = 0.10$ significance level or against the one-sided alternative at the $\alpha = 0.05$ significance level.

Alternate Example: The stronger picker-upper?

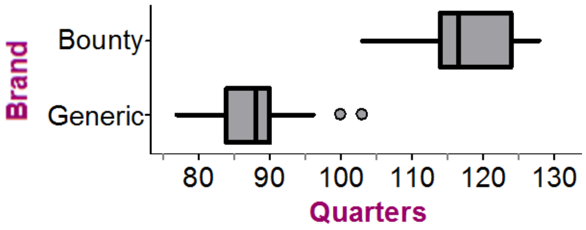
In commercials for Bounty paper towels, the manufacturer claims that they are the "quicker picker-upper." But are they also the stronger picker upper? Two AP Statistics students, Wesley and Maverick, decided to find out. They selected a random sample of 30 Bounty paper towels and a random sample of 30 generic paper towels and measured their strength when wet. To do this, they uniformly soaked each paper towel with 4 ounces of water, held two opposite edges of the paper towel, and counted how many quarters each paper towel could hold until ripping, alternating brands. Here are their results:

- **Bounty:** 106, 111, 106, 120, 103, 112, 115, 125, 116, 120, 126, 125, 116, 117, 114
  118, 126, 120, 115, 116, 121, 113, 111, 128, 124, 125, 127, 123, 115, 114
- **Generic:** 77, 103, 89, 79, 88, 86, 100, 90, 81, 84, 84, 96, 87, 79, 90
  86, 88, 81, 91, 94, 90, 89, 85, 83, 89, 84, 90, 100, 94, 87

Problem:

(a) Display these distributions using parallel boxplots and briefly compare these distributions. Based only on the boxplots, discuss whether or not you think the mean for Bounty is significantly higher than the mean for generic.

The five-number summary for the Bounty paper towels is (103, 114, 116.5, 124, 128) and the five-number summary for the generic paper towels is (77, 84, 88, 90, 103).



Both distributions are roughly symmetric, but the generic brand has two high outliers. The center of the Bounty distribution is much higher than the center of the generic distribution. Although the range of each distribution is roughly the same, the interquartile range of the Bounty distribution is larger.

Since the centers are so far apart and there is almost no overlap in the two distributions, the Bounty mean is almost certain to be significantly higher than the generic mean. If the means were really the same, it would be virtually impossible to get so little overlap.

Alternate Example: The stronger picker-upper?

(b) Use a significance test to determine if there is convincing evidence that wet Bounty paper towels can hold more weight, on average, than wet generic paper towels.

State: We want to perform a test of $H_0: \mu_B - \mu_G = 0$

$H_a: \mu_B - \mu_G > 0$

at the 5% level of significance where μ_B = the mean number of quarters a wet Bounty paper towel can hold and μ_G = the mean number of quarters a wet generic paper towel can hold.

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

- **Random** The students used a random sample of paper towels from each brand.
- **Normal** Even though there were two outliers in the generic distribution, both distributions were reasonably symmetric and the sample sizes are both at least 30, so it is safe to use *t* procedures.
- **Independent** The samples were selected independently and it is reasonable to assume there are more than 10(30) = 300 paper towels of each brand.

Alternate Example: The stronger picker-upper?

Do: For these data, $\overline{x}_B$ = 117.6, S_B = 6.64, $\overline{x}_G$ = 88.1, and S_G = 6.30.

Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(117.6 - 88.1) - 0}{\sqrt{\frac{6.64^2}{30} + \frac{6.30^2}{30}}} = 17.64$$

P-value Using either the conservative df = 30 – 1 = 29, or from technology, df = 57.8, the P-value is approximately 0.

Conclude: Since the *P*-value is less than 0.05, we reject H_0. There is very convincing evidence that wet Bounty paper towels can hold more weight, on average, than wet generic paper towels.

(c) Interpret the *P*-value from (b) in the context of this question.

Since the *P*-value is approximately 0, it is almost impossible to get a difference in means of at least 29.5 quarters by random chance, assuming that the two brands of paper towels can hold the same amount of weight when wet.

Using Two-Sample *t* Procedures Wisely

The two-sample *t* procedures are more robust against non-Normality than the one-sample *t* methods. When the sizes of the two samples are equal and the two populations being compared have distributions with similar shapes, probability values from the *t* table are quite accurate for a broad range of distributions when the sample sizes are as small as $n_1 = n_2 = 5$.

Using the Two-Sample *t* Procedures: The Normal Condition

•*Sample size less than 15:* Use two-sample *t* procedures if the data in both samples/groups appear close to Normal (roughly symmetric, single peak, no outliers). If the data are clearly skewed or if outliers are present, do not use *t*.

• *Sample size at least 15:* Two-sample *t* procedures can be used except in the presence of outliers or strong skewness.

• *Large samples:* The two-sample *t* procedures can be used even for clearly skewed distributions when both samples/groups are large, roughly $n \geq 30$*.*

Using Two-Sample *t* Procedures Wisely

Here are several cautions and considerations to make when using two-sample *t* procedures.

✓ **In planning a two-sample study, choose equal sample sizes if you can.**

✓ **Do not use “pooled” two-sample *t* procedures!**

✓ **We are safe using two-sample *t* procedures for comparing two means in a randomized experiment.**

✓ **Do not use two-sample *t* procedures on paired data!**

✓ **Beware of making inferences in the absence of randomization. The results may not be generalized to the larger population of interest.**

Alternate Example: Testing with distractions

- Suppose you are designing an experiment to determine if students perform better on tests when there are no distractions, such as a teacher talking on the phone. You have access to two classrooms and 30 volunteers that are willing to participate in your experiment.

(a) Design an experiment so that a two-sample *t* test would be the appropriate inference method.

On 15 index cards write “A” and on 15 index cards write “B”. Shuffle the cards and hand them out at random to the 30 volunteers. All 30 subjects will take the same reading comprehension test. Subjects that receive A cards will go to a classroom with no distractions and subjects that receive B cards will go to a classroom that will have the proctor talking on the phone during the test. At the end of the experiment, compare the mean score for subjects in room A with the mean score for subjects in room B.

(b) Design an experiment so that a paired *t* test would the appropriate inference method.

Using the same procedure in part (a), divide the subjects into two rooms and give them the same reading comprehension test. One room will be distraction free and the other room will have a proctor talking on the phone. Then, after a short break, give all 30 subjects a similar reading comprehension test, but have the distraction in the opposite room. At the end of the experiment, calculate the difference in the two reading comprehension scores for each subject and compare the mean difference to 0.

(c) Which experimental design is better? Explain.

The experimental design in part (b) is better since it eliminates an important source of variability—the reading comprehension skills of the individual subjects.

Section 10.2
Comparing Two Means

Summary

In this section, we learned that…

- ✓ Choose an SRS of size n_1 from Population 1 and an independent SRS of size n_2 from Population 2. The sampling distribution of the difference of sample means has:

  Shape Normal if both population distributions are Normal; approximately Normal otherwise if both samples are large enough $n (\geq 30)$.

  Center The mean $\mu_1 - \mu_2$.

  Spread As long as each sample is no more than 10% of its populati

  (10% condition), its standard deviation is $\sqrt{\frac{\sigma_1^{\;2}}{n_n} + \frac{\sigma_2^{\;2}}{n_2}}$.

- ✓ Confidence intervals and tests for the difference between the means of two populations or the mean responses to two treatments $\mu_1 - \mu_2$ are based on the difference between the sample means.

- ✓ If we somehow know the population standard deviations σ_1 and σ_2, we can use a *z* statistic and the standard Normal distribution to perform probability calculations.

Section 10.2
Comparing Two Means

Summary

- Since we almost never know the population standard deviations in practice, we use the **two-sample *t* statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}}$$

- where *t* has approximately a *t* distribution with degrees of freedom found by technology or by the conservative approach of using the smaller of $n_1 - 1$ and $n_2 - 1$.

- The conditions for two-sample *t* procedures are:

Random The data are produced by a random sample of size n_1 from Population 1 and a random sample of size n_2 from Population 2 or by two groups of size n_1 and n_2 in a randomized experiment.

Normal Both population distributions (or the true distributions of respons to the two treatments) are Normal OR both sample/group sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$).

Independent Both the samples or groups themselves and the individua observations in each sample or group are independent. When samplin without replacement,check that the two populations are at least 10 time as large as the corresponding samples (the 10% condition).

+

Section 10.2
Comparing Two Means

Summary

- ✓ The level C **two-sample *t* interval for *μ*$_1$ – *μ*$_2$** 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 conservative approach.

- ✓ To test *H*$_0$: *μ*$_1$ - *μ*$_2$ = hypothesized value, use a **two-sample *t* test for *μ*$_1$ - *μ*$_2$**. The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\;2}}{n_1} + \frac{s_2^{\;2}}{n_2}}}$$

P-values are calculated using the *t* distribution with degrees of freedom from either technology or the conservative approach.

Section 10.2
Comparing Two Means

Summary

- ✓ The two-sample *t* procedures are quite robust against departures from Normality, especially when both sample/group sizes are large.

- ✓ Inference about the difference μ_1 - μ_2 in the effectiveness of two treatments in a completely randomized experiment is based on the **randomization distribution** of the difference between sample means. When the Random, Normal, and Independent conditions are met, our usual inference procedures based on the sampling distribution of the difference between sample means will be approximately correct.

- ✓ Don’t use two-sample *t* procedures to compare means for paired data*.*

+

Looking Ahead…

In the next Chapter…

We'll learn how to perform inference for distributions of categorical data.

We'll learn about

- ✓ **Chi-square Goodness-of-Fit tests**
- ✓ **Inference for Relationships**