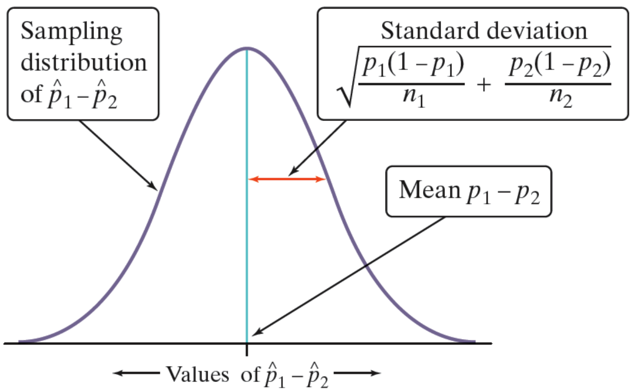
Often we want to compare the proportions of individuals with a certain characteristic in Population 1 and Population 2 (parameters of interest p_1 and p_2).

The Sampling Distribution of $\hat{p}_1 - \hat{p}_2$:

Choose an SRS of size n_1 from Population 1 with proportion of successes p_1 and an independent SRS of size n_2 from Population 2 with proportion of successes p_2.



- SHAPE: When n_1p_1 , $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.

- CENTER: The mean of the sampling distributions is $p_1 - p_2$. That is, the difference in sample proportions is an unbiased estimator of the difference in population proportions.

- SPREAD: The standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is:
  Formula:

  As long as each sample is no more than 10% of its population (the 10% condition)

Example: Who Does More Homework?
Suppose that there are two large high schools, each with more than 2000 students, in a town. At HS 1, 70% of students did their homework last night. Only 50% of the students at HS 2 did their homework last night. The counselor at HS 1 takes an SRS of 100 students and records the proportion $\hat{p}_1$ that did homework. HS 2's counselor takes an SRS of 200 students and records the proportion $\hat{p}_2$ that did homework. HS 1's counselor and HS 2's counselor meet to discuss the results of their homework surveys. After the meeting, they both report to their principals that $\hat{p}_1 - \hat{p}_2 = 0.10$
a) Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.
b) Find the probability of getting a difference in sample proportions $\hat{p}_1 - \hat{p}_2$ of 0.10 or less from the two surveys. Show your work.
c) Does the result in (b) give us reason to doubt the counselors' reported value? Explain.
d) Suppose that two counselors at School 1, Michelle and Julia, independently take a random sample of 100 students from their school and record the proportion of students who did their homework last night. When they finished, they find that the difference in proportions, $\widehat{p_M - p_J}$, is 0.08. They were surprised to get a difference this big, considering that they were sampling from the same population. Find the probability of getting two proportions that are at least 0.08 apart.

CYU: Pg.608

TWO-SAMPLE *z* INTERVAL FOR A DIFFERENCE BETWEEN TWO PROPORTIONS:

Formula:

This formula is used when the three conditions are met:
1) _______________: The data are produced by a random sample or by two groups in a randomized experiment.

2) _______________l: The counts of successes and failures in each sample are at least 10.

3) _______________: 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 (10% condition)

Constructing and interpreting a confidence interval:

Example: Teens and Adults on Social Networking Sites.
As part of the Pew Internet and American Life Project, researchers conducted two surveys in late 2009. The first survey asked a random sample of 800 U.S. teens about their use of social media and the Internet. A second survey posed similar questions to a random sample of 2253 U.S. adults. In these two studies, 73% of teens and 47% of adults said that they use social-networking sites. Use these results to construct and interpret a 95% confidence interval for the difference between the proportion of all U.S. teens and adults who use social-networking sites.

**It is valid to take one random sample from the population and then separate the individuals into two independent groups.

CYU: Pg.611

Two-Sample z Test for the Difference between Two Proportions
State your hypotheses: H_o: $p_1 - p_2 = 0$ or $p_1 = p_2$ (null hypothesis is usually a statement of "no difference") H_a: $p_1 - p_2 < 0$ or $p_1 - p_2 > 0$ or $p_1 - p_2 \neq 0$ (alternative hypothesis is stated according to the researchers 'suspicions') Check conditions: 1. RANDOM: The data are produced by taking random samples from two populations or by two groups in a randomized experiment. 2. NORMAL: The counts of successes and failures in each sample or group are at least 10. 3. 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 (10% condition) Find the pooled proportion: $\hat{p}_c$ = Compute the z statistic: Find the P-value by calculating the probability of getting a z statistic this large or larger in the direction specified by the alternative hypothesis:

Example: Hungry Children (Significance Test for $p_1 - p_2$)

Researchers designed a survey to compare the proportions of children who come to school without eating breakfast in two low-income elementary schools. An SRS of 80 students from School 1 found that 19 had not eaten breakfast. At School 2, an SRS of 150 students included 26 who had not had breakfast. More than 1500 students attend each school. Do these data give convincing evidence of a difference in the population proportions? Carry out a significance test at the $\alpha = 0.05$ level to support your answer.

Example: Cholesterol and Heart Attacks (Significance test in an experiment)
High levels of cholesterol in the blood are associated with higher risk of heart attacks. Will using a drug to lower blood cholesterol reduce heart attacks? The Helsinki Heart Study recruited middle-aged men with high cholesterol but no history of other serious medical problems to investigate this question. The volunteer subjects were assigned at random to one of two treatments: 2051men took the drug gemfibrozil to reduce their cholesterol levels, and a control group of 2030 men took a placebo. During the next five years, 56 men in the gemfibrozil group and 84 men in the placebo group had heart attacks. Is the apparent benefit of gemfibrozil statistically significant? Perform an appropriate test to find out.

CYU: Pg.619

AP Statistics – Chapter 10 Notes §10.2 Comparing Two Means

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

Choose an SRS of size n_1 from Population 1 with mean μ_1 and standard deviation σ_1 and an independent SRS of size n_2 from Population 2 with mean μ_2 and standard deviation σ_2.

- SHAPE: When the population distributions are Normal, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is Normal. In other cases, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be approximately Normal if the sample sizes are large enough ($n_1 \geq 30\ \ and\ \ n_2 \geq 30$)
- CENTER: The mean of the sampling distributions is $\bar{x}_1 - \bar{x}_2$. That is, the difference in sample means is an unbiased estimator of the difference in population means.
- SPREAD: The standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is:
  Formula:

  As long as each sample is no more than 10% of its population (the 10% condition)

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

Based on information from the U.S. National Health and Nutrition Examination Survey, the heights of ten-year-old girls follow a Normal distribution with mean μ_f = 56.4 inches and standard deviation σ_f = 2.7 inches. The heights of ten-year-old boys follow a Normal distribution with μ_m = 55.7 inches and standard deviation σ_m = 3.8 inches. A researcher takes a random sample of 12 girls and a separate sample of 8 boys. After analyzing the data, the researcher reports that the mean height of the boys is larger than the mean height of the girls.

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

b) Find the probability of getting a difference in sample means that is less than zero. Show your work.

c) Does the result in part (a) give us reason to doubt the researcher's stated results? Explain.

CYU: Pg.632

TWO-SAMPLE *t* INTERVAL FOR A DIFFERENCE BETWEEN TWO MEANS:

Formula:

Conditions for a two-sample *t* Interval for the difference between two means

RANDOM: The data are produced by a random sample or by two groups in a randomized experiment.

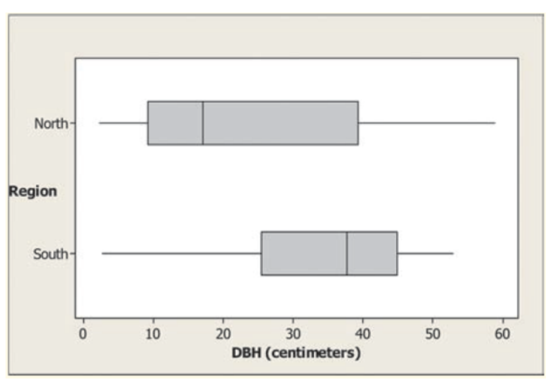
NORMAL: Both population distributions are Normal OR both sample/group sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$)

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 (10% condition)

Example: Big Trees, Small Trees, Short Trees, Tall Trees (Confidence Interval for $\mu_1 - \mu_2$)
The Wade Tract Preserve in Georgia is an old-growth forest of long-leaf 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 of breast height in cm. Comparative boxplots of the data and summary statistics from Minitab are shown below.

Descriptive Statistics: North, South

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



a) Based on the graph and numerical summaries, write a few sentences comparing the sizes of longleaf pine trees in the two halves of the forest.

b) Construct and interpret a 90% confidence interval for the difference in the mean DBH of longleaf pines in the northern and southern halves of the Wade Tract Preserve.

CYU: Pg.638

Two-Sample t Test for the Difference between Two Means
State the hypotheses: H_o: $\mu_1 - \mu_2$ = hypothesized value (usually zero) H_a: $\mu_1 - \mu_2 <$ value or $\mu_1 - \mu_2 >$ value or $\mu_1 - \mu_2 \neq$ value Check conditions: 1. RANDOM: The data are produced by a random sample or by two groups in a randomized experiment. 2. NORMAL: Both population distributions are Normal OR both sample/group sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$) 3. 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 (10% condition) Compute the two-sample t statistic: Then, Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis. 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 (Comparing two means)
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 mm of mercury. An increase appears a negative response. The data is shown below:

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									

Do the data provide sufficient evidence to conclude that a calcium supplement reduces blood pressure more than the placebo? Carry out an appropriate test to support your answer.

CYU: Pg.644

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 sued except in the presence of outliers or strong skewness.

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

**In planning a two-sample study, choose equal sample sizes if you can. This makes the procedure more robust against non-Normality and the P-values are more accurate.

Do I perform a two-sample t test or a paired t test?

In an experiment, if groups were formed using a completely randomized design, then a two-sample t test is the right choice.

If subjects were paired and then split at random into the two treatment groups, or if each subject received both treatments, then a paired t test is appropriate.

**The proper method of analysis depends on the design of the study.
DO NOT USE TWO-SAMPLE *t* PROCEDURES ON PAIRED DATA

Example: Comparing Tires and Comparing Workers
In each of the following settings, decide whether you should use paired t procedures or two-sample t procedures to perform inference. Explain your choice.

A) To test the wear characteristics of two tire brands, A and B, one Brand A tire is mounted on one side of each car in the rear, while a Brand B tire is mounted on the other side. Which side gets which brand is determined by flipping a coin. The same procedure is used on the front.

B) Can listening to music while working increase productivity? Twenty factory workers agree to take part in a study to investigate this question. Researchers randomly assign 10 workers to do a repetitive task while listening to music and the other 10 workers to do the task in silence.