Math 140
Introductory Statistics

Professor Bernardo Ábrego
Lecture 26
Sections 8.3

8.3 A Confidence Interval for the Difference of two Proportions

- A very common and important situation involves taking two samples independently from two different populations with the goal of estimating the size of the difference between the proportion of successes in one population and the proportion of successes in the other.

Example

- A recent poll of 29,700 U.S. households found that 63% owned a pet. The percentage in 1994 was 56%. [*Source:* American Pet Products Manufacturers Association, www.appma.org.]
- The two populations are the households in the United States in 1994 and the households now. The question you will investigate is:
- "What was the change in the percentage of U.S. households that own a pet?"

Intuitive Formula

- A confidence interval for the difference of two proportions, $p_1 - p_2$, where p_1 is the proportion of successes in the first population and p_2 is the proportion of successes in the second population:

$$(\widehat{p}_1 - \widehat{p}_2) \pm z^* \cdot \text{standard error of}(\widehat{p}_1 - \widehat{p}_2)$$

$\widehat{p}_1$ and $\widehat{p}_2$ are the proportions of successes in the two samples.

- In our example, the 95% confidence interval for the difference between the proportion of U.S. households that own pets now and the proportion that owned pets in 1994:

$$(0.63 - 0.56) \pm 1.96 \cdot \text{standard error of}(\widehat{p}_1 - \widehat{p}_2)$$

Standard Error of the Difference

- The standard errors of the distributions of $\widehat{p}_1$ and $\widehat{p}_2$ can be estimated respectively as:

$$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}} \text{ and } \sqrt{\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

- And if we assume the proportion samples are *independent*, then the standard error of the difference can be approximated as

$$\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

The Formula

- The confidence interval for the difference, $p_1 - p_2$, of the proportion of successes in one population and the proportion of successes in the second population is,

$$\boxed{(\widehat{p}_1-\widehat{p}_2) \pm z^* \cdot \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}$$

- Where $\widehat{p}_1$ is the proportion of successes in a random sample of size n_1 taken from the first population, and $\widehat{p}_2$ is the proportion of successes in a random sample of size n_2 taken from the second population. (Sample sizes do not need to be equal.)

Conditions for use

- The conditions that must be met in order to use this formula are that
  - the two samples are taken randomly and independently from two populations.
  - each population is at least 10 times as large as its sample size.
  - $n_1\widehat{p}_1$, $n_1(1-\widehat{p}_1)$, $n_2\widehat{p}_2$, and $n_2(1-\widehat{p}_2)$ are all at least 5.

Example: A Difference in Pet Ownership?

- A recent pet ownership survey found that 63% of the 29,700 U.S. households sampled own a pet. A 1994 survey, taken by the same organization, found that 56% of the 6,786 U.S. households sampled owned a pet. Find and interpret a 95% confidence interval for the difference between the proportion of U.S. households that own a pet now and the proportion of U.S. households that owned a pet in 1994.

8.4/8.5 Significance Test for the Difference of two Proportions

- Often times we need to decide which is the greater of two proportions, or whether we can assume they are the same. For example.
  - Is snowboarding or skiing more likely to result in a serious injury?
  - Does a new treatment for AIDS result in fewer deaths than an old treatment?
  - Is Reggie Jackson's World Series record so much better than his play during the regular season that the difference can't reasonably be attributed to chance?

Example: Two AIDS treatments

- Consider a clinical trial experiment comparing two treatments for AIDS-related complex (ARC). The investigators want to find out if there is a significant difference on the survival rates of patients who had already developed AIDS. They have two treatments, patients who were given AZT and patients who were given AZT + ACV. Here's the data

Treated with

Survived?		AZT	AZT + ACV	Total
	No	28	13	41
	Yes	41	49	90
	Total	69	62	131

Notation:Proportions, Sample Proportions, and Sample Sizes.

$n_1 = 69$ sample size of patients treated with AZT

$n_2 = 62$ sample size of patients treated with AZT+ACV

$\hat{p}_1 = \frac{41}{69}$ proportion of patients in the sample treated with AZT that survived

$\hat{p}_2 = \frac{49}{62}$ proportion of patients in the sample treated with AZT+ACV that survived

p_1 True proportion of survival if all patients were treated with AZT (unknown)

p_2 True proportion of survival if all patients were treated with AZT+ACV (unknown))

Assumptions about the difference of two proportions.

- If we have obtain two **independent** sample proportions $\hat{p}_1$ and $\hat{p}_2$, then the distribution of the difference of the two proportions $\hat{p}_1 - \hat{p}_2$ is **approximately normal** as long as each proportion satisfies the following three conditions:
  - Each sample $\hat{p}_1$ and $\hat{p}_2$ is a simple random sample from a binomial population, and they are independent from each other. or, in case of experiments, subjects were randomly assigned to their treatments.
  - All the numbers $n_1\hat{p}_1, n_2\hat{p}_2, n_1(1-\hat{p}_1), n_2(1-\hat{p}_2)$ are at least 5.
  - Each of the two population sizes is at least 10 times the sample size.

Checking the conditions

- Check the conditions.
  - We assume that subjects were randomly assigned to treatments.
  - All of the following are greater than 5

$$n_1\hat{p}_1 = 69(41/69) = 41, n_2\hat{p}_2 = 62(49/62) = 49$$

$$n_1(1-\hat{p}_1) = 69(1-41/69) = 28$$

$$n_2(1-\hat{p}_2) = 62(1-49/62) = 13$$

  - The population size of AIDS patients that could potentially be treated with AZT is clearly greater than 10 times 69 = 690

    The population size of AIDS patients that could potentially be treated with AZT+ACV is clearly greater than 10 times 62 = 620

Writing the hypothesis

- The null hypothesis
  - H_0 : The new therapy (AZ+ACV) is as good as the old therapy (AZT). That is the survival rate if all patients were treated with AZT+ACV would be equal to the survival rate if all patients were treated with AZT. In symbols $p_1 = p_2$ or $p_1 - p_2 = 0$
- The alternate hypothesis
  - H_a : The new therapy (AZ+ACV) is better than the old therapy (AZT). That is the survival rate if all patients were treated with AZT+ACV would be greater than the survival rate if all patients were treated with AZT. In symbols $p_1 < p_2$ or $p_1 - p_2 < 0$