Chapter 6

Section 6.3-D: Distribution of a Difference in Proportions

Recall that in chapter 3.1 we already learned about the distribution of the sample proportion.

A Sampling Ditribution is the distribution of sample statistics computed for different samples of the same size from the same population. A sampling distribution shows us how the sample statistic varies from sample to sample.

Properties of Sampling distribution:

- Center: If samples are randomly selected, the sampling distribution will be centered around the population parameter. (for difference of two population proportion: p1- p2)
- Shape: For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.

In Chapter 5 we saw the generic formula for the sampling distribution:

Statistic ~ N(parameter, SE)

We can change the generic formula to describe the sampling distribution of differences in sample proportions:

$$\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, SE)$$

Notice that we need the standard error. The formula for the standard error when describing the sampling distribution of a difference in proportions is …

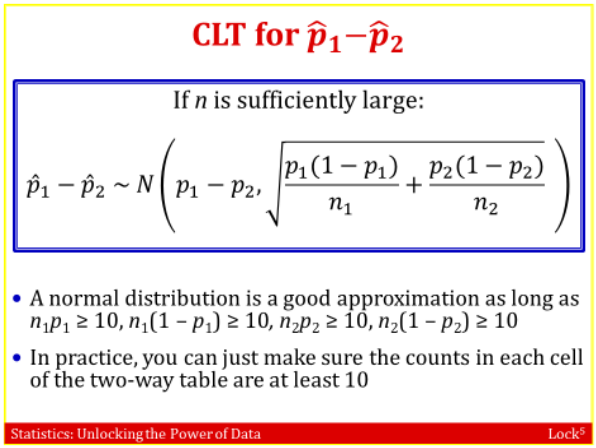
$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

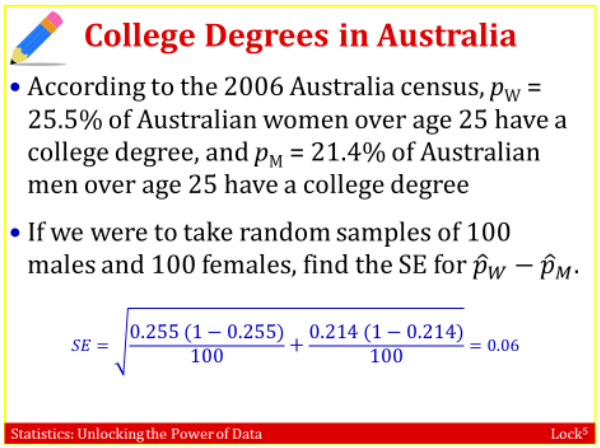
Distribution for a Difference in Two Sample Proportions

When choosing random samples of size $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ from populations with proportions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, respectively, the distribution of the differences in the sample proportions, $\hat{p}_1 - \hat{p}_2$, is centered at the difference in population proportions, $\boldsymbol{p}_1 - \boldsymbol{p}_2$, has standard error for $\hat{p}_1 - \hat{p}_2$ given by

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

and is reasonably normally distributed if $\boldsymbol{n}_1\boldsymbol{p}_1 \geq 10$ and $\boldsymbol{n}_1(1 - \boldsymbol{p}_1) \geq 10$ and $\boldsymbol{n}_2\boldsymbol{p}_2 \geq 10$ and $\boldsymbol{n}_2(1 - \boldsymbol{p}_2) \geq 10$.





Example 1: Proportion of Foreign-Born Residents, Alabama and Arizona

From the 2010 US Census, we learn that 13.9% of the residents of Arizona were born outside the US while 3.4% of the residents of Alabama were born outside the US. If we take random samples of 500 residents from each state and calculate the difference in the proportion of foreign-born residents (Arizona – Alabama), describe the shape, mean, and standard error of the distribution of differences in proportions.

Recall that in Chapter 2.1 we learned about distinguishing between 1 and 2 proportions.

- 2 proportions: the situation involves comparing proportions for two groups.
- 1 proportion: the situation involves comparing two proportions from the same group.

The methods of this section only apply to difference in proportions between samples taken from two distinct groups.

Example 2: Coke/Pepsi Taste Test

Suppose 500 people participate in a blind Coke/Pepsi taste test, and 285 of them prefer Coke while the other 215 of them prefer Pepsi.

(a) If we conduct inference (creating a confidence interval or conducting a hypothesis test) using this data, should we use the formulas for a single proportion or a difference in proportions?

(b) If we want to test whether the preferences are equally split between Coke and Pepsi, what is the null hypothesis?

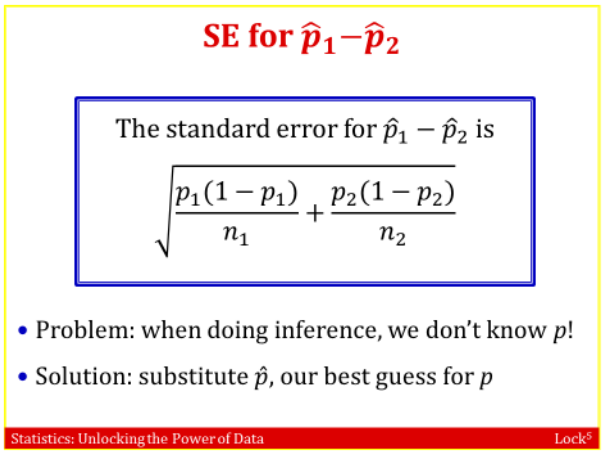
Section 6.3-CI: Confidence Interval for a Difference in Proportions

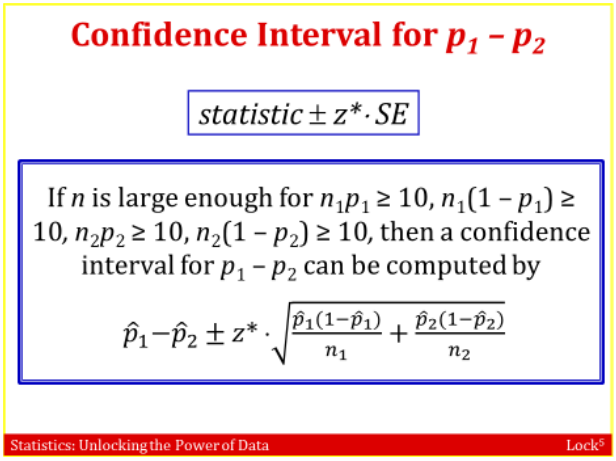
In Chapter 5 we saw the generic formula for a confidence interval:

Statistic ± Critical Value · SE

We can change the generic formula to describe a confidence interval for a difference in proportions:

$$\hat{p}_1 - \hat{p}_2 \pm z^* \cdot SE$$





Example 1: Mobile Connections to Libraries

In a random sample of 2,252 Americans age 16 and older, 11% of the 1,059 men and 16% of the 1,193 women said they have accessed library services via a mobile device. Find a 95% confidence interval for the difference in proportion accessing libraries via mobile devices, between men and women.

The conditions are met for using the normal distribution. The confidence interval is given by:

$$Statistic \pm z^* \cdot SE$$

$$(\hat{p}_M - \hat{p}_F) \pm z^* \cdot \sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \frac{\hat{p}_F(1-\hat{p}_F)}{n_F}}$$

$$(0.11 - 0.16) \pm 1.96 \cdot \sqrt{\frac{0.11(1-0.11)}{1059} + \frac{0.16(1-0.16)}{1193}}$$

$$-0.05 \pm 0.028$$

–0.078 to –0.022

We are 95% sure that the proportion of men who access libraries via mobile devices is between 0.078 and 0.022 less than the proportion of women who access libraries via mobile devices. Note that if we had subtracted the other way, the interval would be positive, but the interpretation would be the same.

Quick Self-Quiz: Smoking and Pregnancy Rate?

Does smoking negatively affect a person's ability to become pregnant? A study collected data on 678 women who were trying to get pregnant. The two-way table shows the proportion who successfully became pregnant during the first cycle trying and smoking status. Find a 90% confidence interval for the difference in proportion of women who get pregnant, between smokers and non-smokers. Interpret the interval in context.

	Smoker	Non-smoker	Total
Pregnant	38	206	244
Not pregnant	97	337	434
Total	135	543	678

Section 6.3-HT: Hypothesis Test for a Difference in Proportions

In Chapter 5 we saw the generic formula for a test statistic:

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Value under } H_0}{\text{SE}}$$

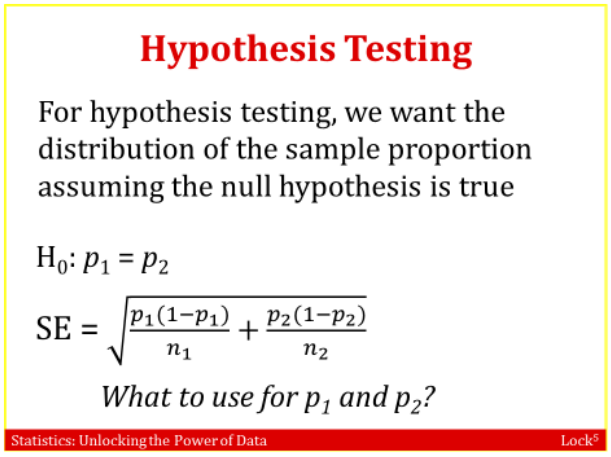
We can change the generic formula to describe a test statistic for a difference in proportions:

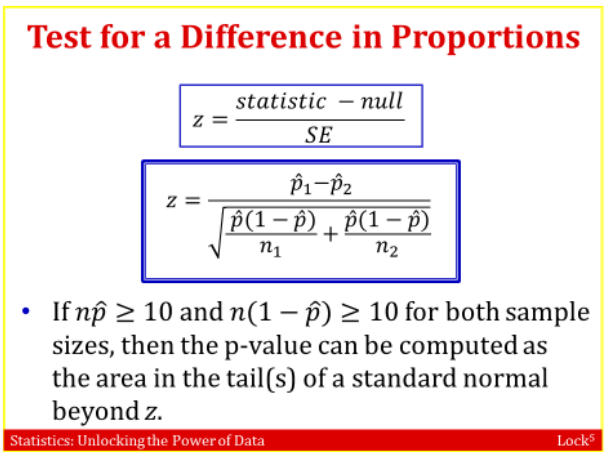
$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{SE}}$$

Notice that we need the standard error. The formula for the standard error when describing a test statistic for a difference in proportions is …

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Notice the formula involves $\hat{p}$ this is known as the pooled portion. A pooled proportion is calculated by combing both groups into one big group, and use the overall proportion.





Example 1: Accuracy of Lie Detectors

Participants in a study to evaluate the accuracy of lie detectors were divided into two groups, with one group reading true material and the other group reading false material, while connected to a lie detector. Both groups received electric shocks to add stress. The two way table indicates whether the participants were lying or telling the truth and also whether the lie detector indicated they were lying or not.

	Detector says lying	Detector says not	Total
Person lying	31	17	48
Person not lying	27	21	48
Total	58	38	96

(a) Are the conditions met for using the normal distribution?
Yes (all cell counts at least 10)

(b) Find the three sample proportions for the proportion of times the lie detector says the person is lying (the proportion for the lying people, the proportion for the truthful people, and the pooled proportion).

We see that the proportion for the lying people is $\hat{p}_L = \frac{31}{48} = 0.6458$*, the proportion for the not lying people is* $\hat{p}_N = \frac{27}{48} = 0.5625$*, and the pooled proportion for all 96 people is* $\hat{p} = \frac{58}{96} = 0.6042$.

(c) Test to see if there is a difference in the proportion of times the lie detector says the person is lying, depending on whether the person is lying or telling the truth. Show all details of the test.

We are testing $H_0: p_L = p_N$ *vs* $H_a: p_L \neq p_N$.

(d) *The test statistic is* $z = \frac{Statistic - Null}{SE} = \frac{(\hat{p}_L - \hat{p}_N) - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_L} + \frac{\hat{p}(1-\hat{p})}{n_N}}} = \frac{0.6458 - 0.5625}{\sqrt{\frac{0.6042(1-0.6042)}{48} + \frac{0.6042(1-0.6042)}{48}}} = 0.834.$

(e) *This is a two-tail test, and the area to the right of 0.834 in a normal distribution is 0.202, so the p-value is 2(0.202) = 0.404.*

(f) *We fail to reject* H_0 *and conclude that there is not enough evidence that a lie detector can tell whether a person is lying or telling the truth.*

Quick Self-Quiz: Tagging Penguins

A study was conducted to see if tagging penguins with metal tags harms them. In the study, 100 penguins were randomly assigned to receive a metal tag or (as a control group) an electronic tag. One of the variables studied is survival rate ten years after the penguins were tagged. The scientists observed that 10 of the 50 metal tagged penguins survived while 18 of the 50 electronic penguins survived.

(a) Create a two-way table of the data. Include row and column totals.

	Lived	*Died*	*Total*
Metal tag	*10*	*40*	*50*
Electronic tag	*18*	*32*	*50*
Total	*28*	*72*	*100*

(b) Are the conditions met for using the normal distribution?

(c) Test to see if the survival rate is lower for metal-tagged penguins than for electronic-tagged penguins. Do metal tags appear to reduce survival rate in penguins?