# Chapter 6

Overview:

Recall that one of the primary goals of this course it to learn how to use statistics to estimate parameters.

Previously we have looked at 3 different distributions that allow us to use information from the statistic to learn about the parameter:

- Sampling Distribution: Helped us visualize the behavior of statistics, where each statistic was calculated from a random sample taken from the population.
- Bootstrap Distribution: Helped us estimate the S.E. of the statistic by plotting bootstrap statistics.
- Randomization Distribution: Helped us calculate a p-value, by helping us visualize what statistics look like when the null hypothesis is true.
- Further we have seen that the statistic is rarely if ever exactly equal to the parameter and that the statistic varies from sample to sample. How the statistic changes from sample to sample is known as the standard error. When using the statistic to learn about the parameter we need to consider the SE of the statistic.

In chapter 5 we saw that we can approximate the distributions above using a Normal Distribution. In order to use the normal distribution, we need to know the appropriate mean/standard deviation. We also have 3 generic formulas that will help us:

| | |
|---|---|
| **Sampling Distribution** | Statistic ~ N (Parameter, SE) |
| **Confidence Interval** | Statistic ± Critical Value · SE |
| **Test Statistic** | $\text{Test Statistic} = \dfrac{\text{Statistic } - \text{Value under H}_0}{\text{SE}}$ |

The goal of chapter 6 is to learn different formulas for the Standard Error (SE). The formula we use for standard error will depend on 2 things:

1. What type of inference (sampling distribution, C.I., or Hypothesis Test).
2. Which statistic/parameter set we are working with.

Each of the sections in chapter 6 will look at a different statistic/parameter and type of inference.

## 6.1-D: Sampling Distribution of a Sample Proportion

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 population proportion: p)
- 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 for a sample proportion:

$$\hat{p} \sim \text{N (p, SE)}$$

When selecting random samples of size $\boldsymbol{n}$ from a population with proportion $\boldsymbol{p}$, the distribution of the sample proportions is centered at the population proportion $\boldsymbol{p}$, has standard error given by

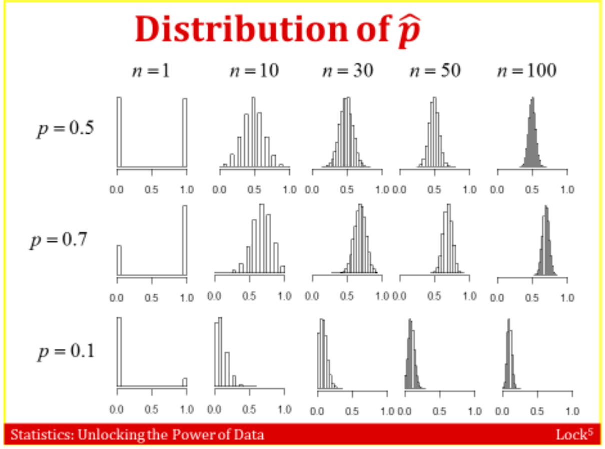
$$\sqrt{\frac{p(1-p)}{n}}$$

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

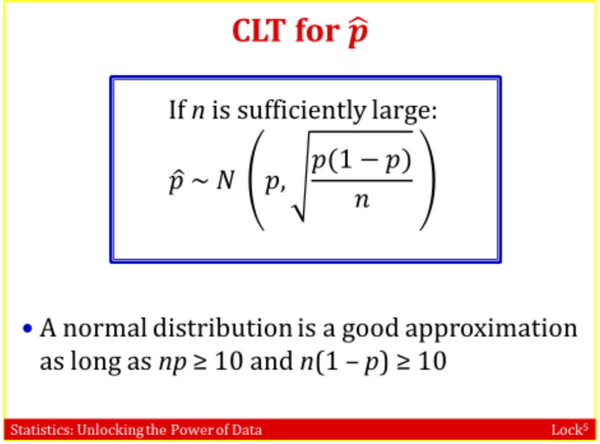
- **The Sampling Distribution of a Proportion** $\hat{p}$ ( $\hat{p} = \frac{x}{n}$):

  The sampling distribution of the sample proportion may be considered approximately normal only if $\boldsymbol{np} \geq 10$ and $\boldsymbol{n}(1-\boldsymbol{p}) \geq 10$

1. Mean of the sampling distribution of $\hat{p}$ : $\mu_{\hat{p}} = P$
2. Standard error of the sampling distribution of $\hat{p}$ : $SE = \sqrt{\frac{p(1-p)}{n}}$


Distribution of $\hat{p}$

Statistics: Unlocking the Power of Data Lock5


CLT for $\hat{p}$

If $n$ is sufficiently large:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- A normal distribution is a good approximation as long as $np \geq 10$ and $n(1-p) \geq 10$

Statistics: Unlocking the Power of Data Lock5

***Example 1: Proportion Speaking a Language other than English in Oregon***

From the 2010 US Census, we learn that 14.6% of the residents of Oregon speak a language other than English at home. If we take random samples of size *n = 100* and calculate the proportion of the sample that speaks a language other than English at home, describe the shape, mean, and standard error of the distribution of sample proportions.

$\boldsymbol{np} \geq 10$ and $\boldsymbol{n}(1-\boldsymbol{p}) \geq 10$.

*The distribution will be bell-shaped with a mean of 0.146 and a standard error of*

$$SE = \sqrt{\frac{0.146(1-0.146)}{100}} = 0.035.$$

Describe the shape, mean, and standard error of the distribution of sample proportions if we instead take random samples of size *n = 500*. What is the effect of the larger sample size on the shape, mean and standard error?

*The distribution will be bell-shaped with a mean of 0.146 and a standard error of*

$$SE = \sqrt{\frac{0.146(1-0.146)}{500}} = 0.016.$$

*The larger sample size does not affect the shape or the center, but it reduces the standard error.*

***Quick Self-Quiz: Distribution of a Sample Proportion***

From the 2010 US Census, we learn that 71.8% of the residents of Missouri are 21 years old or over. If we take random samples of size *n = 200* and calculate the proportion of the sample that is 21 years old or over, describe the shape, mean, and standard error of the distribution of sample proportions.

## 6.1-CI: **Confidence Interval for a Proportion**

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

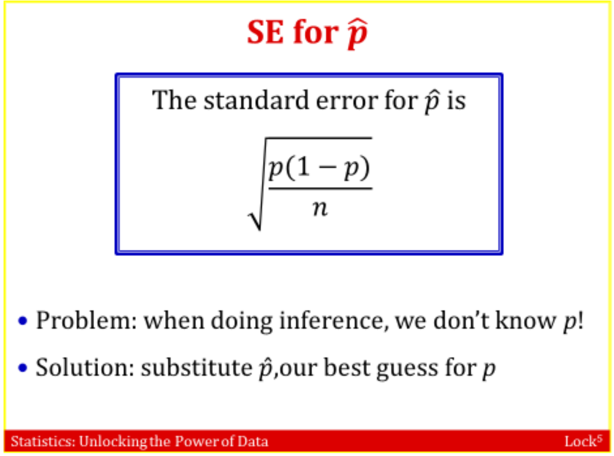
**Statistic ± Critical Value · SE**

Margin of error

We can change the generic formula to describe a confidence interval for a population proportion: $\hat{p} \pm z^* \cdot SE$

Notice that we need the standard error. The formula for the standard error when describing a confidence interval for a population proportion is …

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



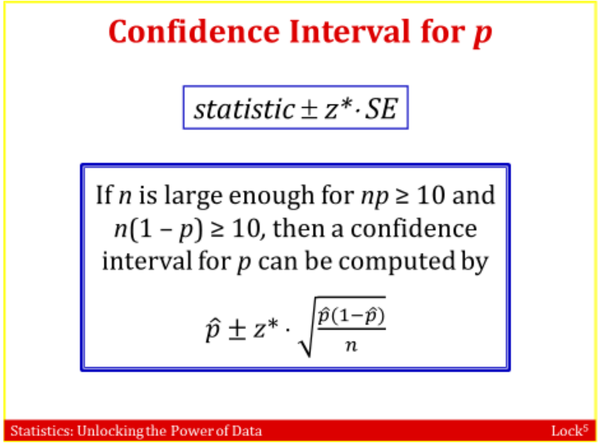
**SE for $\hat{p}$**

The standard error for $\hat{p}$ is

$$\sqrt{\frac{p(1-p)}{n}}$$

- Problem: when doing inference, we don't know $p$!
- Solution: substitute $\hat{p}$,our best guess for $p$

Statistics: Unlocking the Power of Data Lock$^5$



**Confidence Interval for $p$**

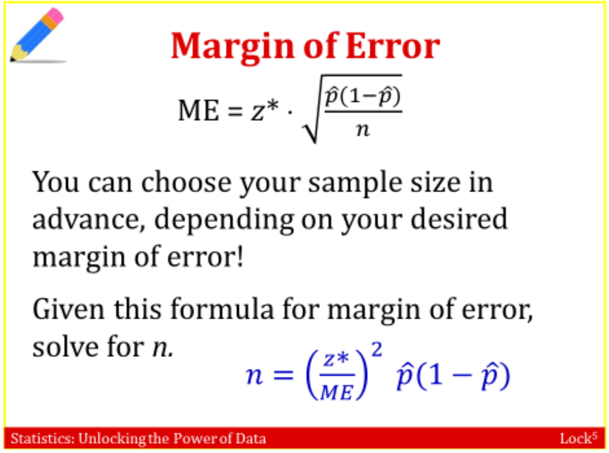
$$statistic \pm z^* \cdot SE$$

If $n$ is large enough for $np \geq 10$ and $n(1-p) \geq 10$, then a confidence interval for $p$ can be computed by

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Statistics: Unlocking the Power of Data Lock$^5$

- The **level of confidence** represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained. The level of confidence is denoted $(1 - \alpha)\cdot 100\% = C$
- For example, a 95% level of confidence ($\alpha = 0.05$) implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, we will expect 95 of the intervals to contain the parameter and 5 not to include the parameter.
- The level of confidence is always expressed as a percent
- The level of confidence is $c= (1 - \alpha) \bullet 100\%$
  - When $\alpha = .05$, then $c= (1 - \alpha) = .95$, and we have a 95% level of confidence
  - When $\alpha = .01$, then $c= (1 - \alpha) = .99$, and we have a 99% level of confidence
- The quantity $Z_{\alpha/2}$ is the $(1- \alpha/ 2)$ 100th percentile.
- Frequently used confidence levels, and their critical values(Z value), are
  - C=90% corresponds to $\alpha$ =.10, $z^*= Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$
  - C=95% corresponds to $\alpha$ =.05, $z^*== Z_{\frac{\alpha}{2}} = Z_{0.025} =1.96$
  - C=98% corresponds to $\alpha$ =.02, $z^*= Z_{\frac{\alpha}{2}} = Z_{0.01} =2.33$
  - C=99% corresponds to $\alpha$ =.01, $z^*= Z_{\frac{\alpha}{2}} = Z_{0.005} =2.575$



## Margin of Error

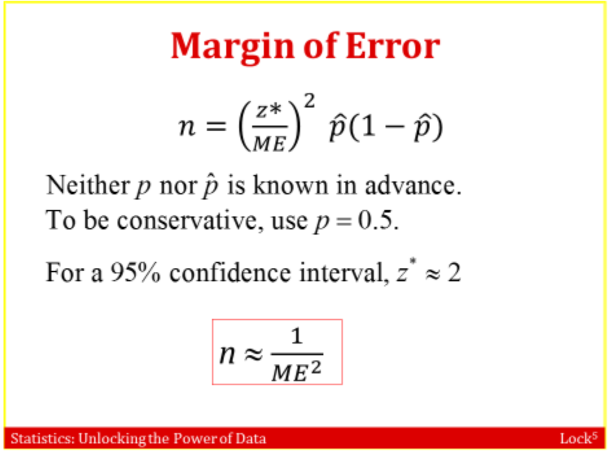
$$\text{ME} = z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

You can choose your sample size in advance, depending on your desired margin of error!

Given this formula for margin of error, solve for *n*.

$$n = \left(\frac{z*}{ME}\right)^2 \hat{p}(1-\hat{p})$$

Statistics: Unlocking the Power of Data Lock[5]



## Margin of Error

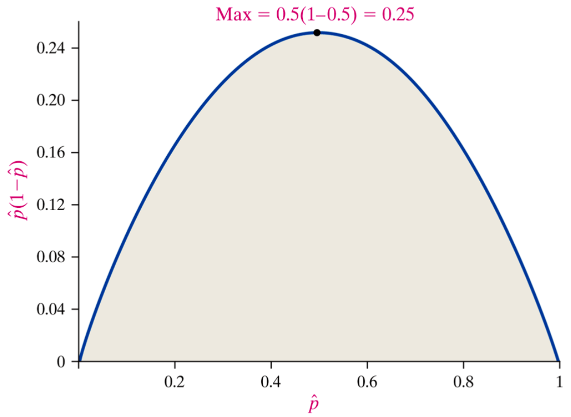
$$n = \left(\frac{z*}{ME}\right)^2 \hat{p}(1-\hat{p})$$

Neither $p$ nor $\hat{p}$ is known in advance.
To be conservative, use $p = 0.5$.

For a 95% confidence interval, $z^* \approx 2$

$$\boxed{n \approx \frac{1}{ME^2}}$$

Statistics: Unlocking the Power of Data Lock[5]



***Example 1: Movie Goers are More Likely to Watch at Home***

In a random sample of 500 movie goers in January 2013, 320 of them said they are more likely to wait and watch a new movie in the comfort of their own home. Find and interpret a 95% confidence interval for the proportion of movie goers who are more likely to watch a new movie from home.

*We see that* $\hat{p} = \frac{320}{500} = 0.64$. *it is important to use the sample proportion in decimal form rather than percent form. The confidence interval is given by:*

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

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.64 \pm 1.96 \cdot \sqrt{\frac{0.64(1-0.64)}{500}}$$

$$0.64 \pm 0.042$$

$$0.598 \quad to \quad 0.682$$

*We are 95% sure that the proportion of all movie goers who are more likely to wait and watch a new movie at home is between 0.598 and 0.682.*

***Example 2: Sample Size and Margin of Error for Movie Goers***

(a) What is the margin of error for the confidence interval found in Example 1?

Margin of error is 0.042.

(b) What sample size is needed if we want a margin of error within ±2%? (Use the sample proportion from the original sample.)

$$n = \left(\frac{z^*}{ME}\right)^2 \tilde{p}(1-\tilde{p}) = \left(\frac{1.96}{0.02}\right)^2 0.64(1-0.64) = 2212.76$$

We need a sample size of at least n = 2,213 to have a margin of error this small. This is substantially more than the sample size of 500 used in the actual survey.

(c) What sample size is needed if we want a margin of error within ±2%, and if we use the conservative estimate of p = 0.5?

$$n = \left(\frac{z^*}{ME}\right)^2 \tilde{p}(1-\tilde{p}) = \left(\frac{1.96}{0.02}\right)^2 0.5(1-0.5) = 2401.$$

We need a sample size of at least n = 2,401 to have a margin of error this small

Notice that if we have less knowledge of the actual proportion, we need a larger sample size to arrive at the same margin of error.

***Quick Self-Quiz: What Percent of US Adults are Thriving?***

In a random sample of 1500 US adults, 780 of them are "Thriving" which is defined as rating their current life as 7 or higher on a 10 point scale and rating their future life as 8 or higher on a 10 point scale. Find and interpret a 99% confidence interval for the proportion of all US adults who are thriving (based on this definition.)

$$\hat{p} =$$

*The confidence interval is given by:*

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

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

*We are 99% sure that the proportion of all US adults who are thriving is between and*

## 6.1-HT: **Hypothesis Test for a Proportion**

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 single proportion:

$$Z = \frac{\hat{p} - p_0}{SE}$$

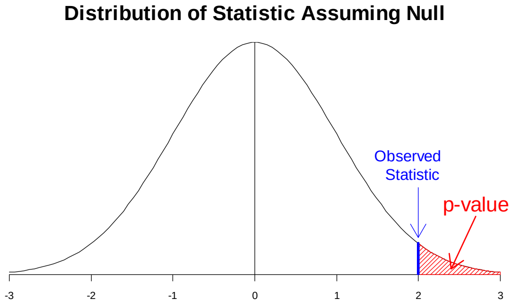
Notice that we need the standard error. The formula for the standard error when describing a test statistic for a single proportion is …

$$\sqrt{\frac{p_0\,(1-p_0)}{n}}$$

Therefore, the formula to describe a test statistic for a single proportion:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

Once we calculate the test statistic we can compare the test statistic to a standard normal curve to find the p-value and conclude the hypothesis test.



- Explain the Logic of Hypothesis Testing:
  When observed results are unlikely under the assumption that the null hypothesis is true, we say the result is **statistically significant**. When results are found to be statistically significant, we reject the null hypothesis.

- **Steps of a Significance Test about a Population Proportion**

- ***Step 1: Assumptions***
  - The data are obtained using randomization
  - The sample size is sufficiently large that the sampling distribution of the sample proportion is approximately normal: np≥ 10 and n(1-p) ≥ 10

  The sampling distribution of $\hat{p}$ is approximately normal, with mean $\mu_{\hat{p}} = p$ and

  Standard error $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

- ***Step 2: Hypotheses***

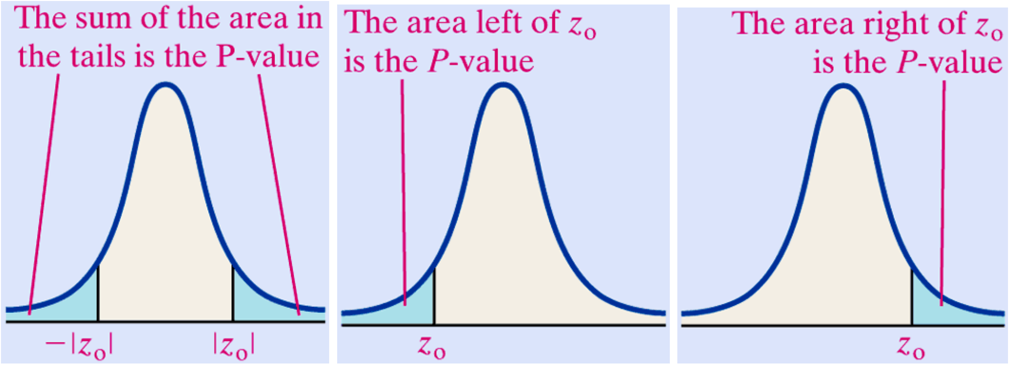
| **Form** | **Null and alternative hypotheses** |
|---|---|
| Right-tailed test, one-tailed test | $H_0$: $p \le p_0$ versus $H_a$: $p > p_0$ |
| Left-tailed test, one-tailed test | $H_0$: $p \ge p_0$ versus $H_a$: $p < p_0$ |
| Two-tailed test | $H_0$: $p = p_0$ versus $H_a$: $p \ne p_0$ |

- ***Step 3: Test Statistic***
  The test statistic measures how far the sample proportion falls from the null hypothesis value, $p_0$, relative to what we'd expect if $H_0$ were true
  The test statistic is: (when we test a hypothesis, the null hypothesis is always assumed true)

$$Z = \frac{Statistic - Null}{SE} = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

- ***Step 4: P-value***
  The P-value summarizes the evidence
  It describes how unusual the observed data would be if $H_0$ were true.

  **The smaller the *P*-value, the stronger the evidence the data provide against the null hypothesis.**
  **The significance level is a number such that we reject $H_0$ if the P-value is less than or equal to that number**


The sum of the area in the tails is the P-value
$-|z_0|$
$|z_0|$
The area left of $z_0$ is the $P$-value
$z_0$
The area right of $z_0$ is the $P$-value
$z_0$

Two-Tailed — p-value=2normalcdf( , )

Left-Tailed — p-value=normalcdf(-100 , z)

Right-Tailed — p-value=normalcdf(z,100)

- ***Step 5: Conclusion***
  **We summarize the test by reporting and interpreting the P-value**

  Reject $H_0$ when the p-value < $\alpha$.

  Do not reject the null hypothesis if the $P$-value > $\alpha$

- Interpreting the Conclusion
  If you reject $H_0$, the interpretation is "There is evidence that [whatever $H_a$ says]."

  If you do not reject $H_0$, the interpretation is "There is insufficient evidence that [whatever $H_a$ says]".

***Example 1: NFL Overtime***

At the start of overtime in a National Football League game, a coin is flipped to determine which team will kick off and which will receive. The question of interest is how much advantage (if any) is given to the team that wins the coin flip at the start of the sudden death overtime period. In the overtime games played between 1974 and 2009, the winner of the coin flip won the game in 240 of the 428 games in which a winner was determined in overtime. Assume that the overtime games played during this time period can be viewed as a sample of all possible NFL overtime games. Do the data provide sufficient evidence to conclude that the team winning the coin flip has an advantage in overtime games? Show all details of the test.

*We are testing $H_0$: $p = 0.5$ vs $H_a$: $p > 0.5$ where p represents the proportion of times that the team winning the coin flip wins the game. The sample proportion is* $\hat{p} = \frac{240}{428} = 0.56$ *and the sample size is n = 428.*

*The test statistic is* $z = \frac{Statistic - Null}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.56-0.5}{\sqrt{\frac{0.5(1-0.5)}{428}}} = 2.483.$

*This is a right-tail test, and we see that the area to the right of 2.483 in a normal distribution is 0.0065, so the p-value is 0.0065.*

*We reject $H_0$ and conclude that there is evidence that the winner of the coin flip has an advantage in overtime games in the NFL.*