Section 3.1

Sampling Distributions
Outline

- Statistical inference
- Statistic versus parameter
- Sampling distribution
- Variability of the statistic: standard error
- Importance of sample size
- Importance of random sampling
The Big Picture

Population

Sampling

Sample

Statistical Inference
**Statistical Inference**

*Statistical inference* is the process of drawing conclusions about the entire population based on information in a sample.
Statistic and Parameter

A **parameter** is a number that describes some aspect of a population.

A **statistic** is a number that is computed from data in a sample.

- We usually have a sample statistic and want to use it to make inferences about the population parameter.
The Big Picture

Population
PARAMETERS

Sampling

Sample
STATISTICS

Statistical Inference
# Parameter versus Statistic

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu$ (mu)</td>
</tr>
<tr>
<td>Proportion</td>
<td>$p$</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>$\sigma$ (sigma)</td>
</tr>
<tr>
<td>Correlation</td>
<td>$\rho$ (rho)</td>
</tr>
</tbody>
</table>
Parameter versus Statistic

- For each of the following, state whether the quantity described is a parameter or a statistic, and give the correct notation.
  - Average household income for all houses in the US, using data from the US census
  - Correlation between height and weight for players on the 2010 Brazil world cup team, using data from all 23 players on the roster
  - Proportion of people who use an electric toothbrush, using data from a sample of 300 adults
Election Polling

• Before the 2012 presidential election, 1000 registered voters were asked who they plan to vote for in the 2012 presidential election

• What proportion of voters planned to vote for Obama?

\[ \hat{p} = 0.50 \quad \quad p = ??? \]

http://www.politico.com/p/2012-election/polls/president
**Point Estimate**

We use the statistic from a sample as a *point estimate* for a population parameter.

- Point estimates will not match population parameters exactly, but they are our best guess, given the data.
Election Polls

- Actually, several polls were conducted over this time frame (9/7/12 – 9/9/12):

  - **National '12 President General Election**
    - Washington Post-ABC News
    - 09/07/2012-09/09/2012
    - 710 likely voters
    - D Barack Obama: 49%
    - R Mitt Romney: 48%
    - No opinion: 3%

  - **National '12 President General Election**
    - Public Policy Polling/SIEU/Daily Kos
    - 09/07/2012-09/09/2012
    - 1000 registered voters
    - D Barack Obama: 50%
    - R Mitt Romney: 44%
    - Undecided: 6%

  - **National '12 President General Election**
    - CNN/ORC International
    - 09/07/2012-09/09/2012
    - 709 likely voters
    - D Barack Obama: 52%
    - R Mitt Romney: 46%
    - Neither: 2%
    - No opinion: 1%

http://www.politico.com/p/2012-election/polls/president
IMPORTANT POINTS

• Sample statistics *vary* from sample to sample. (they will not match the parameter exactly)

• **KEY QUESTION**: For a given sample statistic, what are plausible values for the population parameter? How much uncertainty surrounds the sample statistic?

• **KEY ANSWER**: It depends on how much the statistic varies from sample to sample!
Many Samples

• To see how statistics vary from sample to sample, let’s take many samples and compute many statistics!
Reese’s Pieces

• What proportion of Reese’s pieces are orange?

• Take a random sample of 10 Reese’s pieces

• What is your sample proportion? ⇒ dotplot

• Give a range of plausible values for the population proportion

• You just made your first sampling distribution!
Sampling Distribution

A *sampling distribution* 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
Center and Shape

**Center:** If samples are randomly selected, the sampling distribution will be centered around the population parameter.

**Shape:** For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.

- StatKey
Sampling Caution

• If you take *random samples*, the sampling distribution will be centered around the true population parameter.

• If sampling bias exists (if you do not take random samples), your sampling distribution may give you bad information about the true parameter.
Lincoln’s Gettysburg Address
Sampling Distribution

• We’ve learned about center and shape, but remember what we really care about is **variability** of the sampling distribution

• Remember our key question and answer: to assess uncertainty of a statistic, we need to know how much the statistic varies from sample to sample!

• The variability of the sample statistic is so important that it gets its own name...
The **standard error** of a statistic, SE, is the standard deviation of the sample statistic.

- The standard error measures how much the statistic varies from sample to sample.
- The standard error can be calculated as the standard deviation of the sampling distribution.
Sample Size Matters!

As the sample size increases, the variability of the sample statistics tends to decrease and the sample statistics tend to be closer to the true value of the population parameter.

- For larger sample sizes, you get less variability in the statistics, so less uncertainty in your estimates.
Reese’s Pieces

- **StatKey**

**StatKey** Sampling Distribution for a Proportion

Choose samples of size \( n = 10 \)

Sampling Dotplot of Proportion

- **Data Tables**
  - **Original Population**
    - Proportion: 0.45
  - **Sample**
    - Count: 4
    - Sample Size: 10
    - Proportion: 0.400

Statistics: Unlocking the Power of Data
Summary

- *Statistical inference* is drawing conclusions about a population based on a sample.
- We use a *sample statistic* to estimate a *population parameter*.
- To assess the uncertainty of a statistic, we need to know how much it varies from sample to sample.
- To create a *sampling distribution*, take many samples of the same size from the population, and compute the statistic for each.
- *Standard error* is the standard deviation of a statistic.