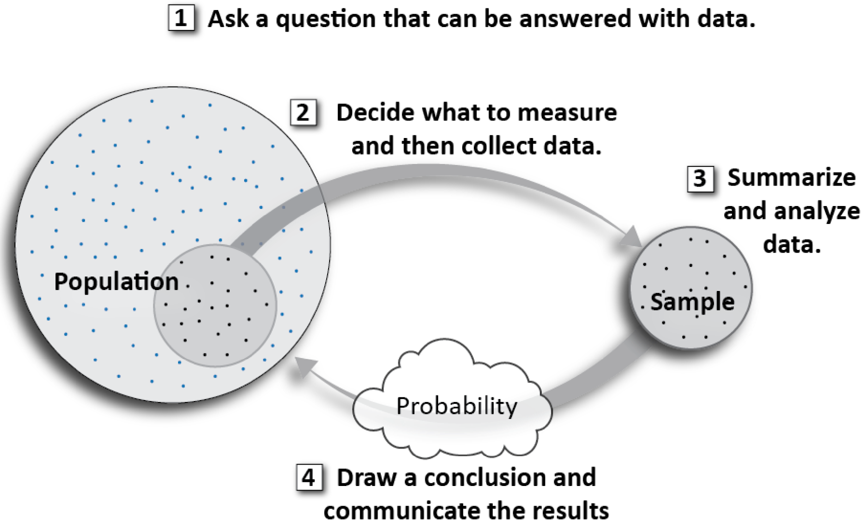
Carnegie Foundation
for the Advancement of Teaching

7.1.1

Sampling Distributions Lesson 1

THE BIG PICTURE

We have described statistical analysis as a process having four steps. These are represented below.


1 Ask a question that can be answered with data.
2 Decide what to measure and then collect data.
3 Summarize and analyze data.
Population
Sample
Probability
4 Draw a conclusion and communicate the results

In Module 7 we turn our attention to the fourth step. We will focus on categorical data. Recall that we use proportions and percentages to summarize categorical data. With categorical variables, we note whether each individual in the sample or population falls into some category. We summarize the data by calculating a proportion, which means that we divide the total number who fall into the category by the total sample or population size. We will use categorical data from samples to draw conclusions about **population proportions**. A population proportion is the proportion for the entire population.

When we use sample data to draw a conclusion about a population, we say we are making a **statistical inference** about the population. In this module we learn about two main types of inference: (1) using sample data to estimate a population proportion (Lessons 7.1.1 through 7.2.2) and (2) using sample data to test a claim about a population proportion (Lessons 7.3.1 through 7.3.4).

Before we learn about the two types of inference, we need to explore how proportions vary as we take different samples (Lesson 7.1.1). First let's look at the types of questions we'll try to answer using statistical inference.

TRY THESE

1 What proportion of all M&M's candies is blue? Which type of inference will help us answer this question (estimate a population proportion, or test a claim about the population proportion)?

2 A student in class claims that less than 1/3 of all M&M's are blue. We want to answer the question: "Is she correct?" Which type of inference will help us answer this question (estimate a population proportion, or test a claim about the population proportion)?

3 If you need to learn the population proportion of M&M's that is blue (and the manufacturer refuses to tell you), what should you do?

4 To investigate the proportion of M&M's that is blue we need to gather data. Will the data we collect be quantitative or categorical?

5 How would you summarize your data (with sample means, proportions, dotplots, bar graphs, etc.)?

6 You also want to learn the average age of people who eat M&M's. You decide to gather ages from randomly selected M&M's eaters. Are these data values quantitative or categorical?

7 How would you summarize your data (with sample means, proportions, dotplots, bar graphs, etc.)?

In this module, we use sample data and probability to investigate questions like the ones above about population proportions.

INTRODUCTION

Suppose we want to know the proportion of all M&M's that is blue. For each M&M candy, the variable is whether or not the M&M is blue. When we consider a single M&M, if it is blue we call it a **success**. If it is any other color than blue, we call it a **failure**. Note that success doesn't mean good and failure doesn't mean bad. A success is just the outcome we are interested in, and failure is anything else. The proportion of all M&M's that is blue is the total number of blue M&M's divided by the total number of M&M's in the world. This *population proportion* is an example of a **parameter**. A population proportion is denoted by the symbol, p.

Language Tip
***Parameters* are numerical summaries of populations. *Statistics* are numerical summaries of samples.**

We calculate the proportion of M&M's that is blue in a sample by dividing the number of blue M&M's in the sample by the number of M&M's in the sample (the sample size). The proportion of M&M's that is blue in a sample is an example of a **statistic**. A **sample proportion** is denoted by the symbol, $\widehat{p}$, pronounced *p-hat*.

$$\widehat{p} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

It is important to recognize that there are many samples of M&M's, each with their own proportion of blue candies, but *there is only one population proportion*! Sample proportions vary from sample to sample, but the population proportion does not vary.

We use **sample statistics** to either estimate a population parameter or test a claim made about a population parameter.

In this activity, we will gather multiple samples of 25 M&M's from a population. We will calculate a sample proportion ($\widehat{p}$) for each sample. Our many different samples will produce many different sample proportions. These proportions are just a small part of the collection of *all* sample proportions. The collection of all sample proportions forms a distribution of values called the **sampling distribution of sample proportions**.

Population	**Sample**
Collection of all M&M's	One set of 25 M&M's

Parameter	**Statistic**
Proportion of blue candy in the population (p)	Proportion of blue candy in sample ($\widehat{p}$)

TRY THESE

8 Your instructor has given you a cup that contains a *random sample* of 25 M&M's.

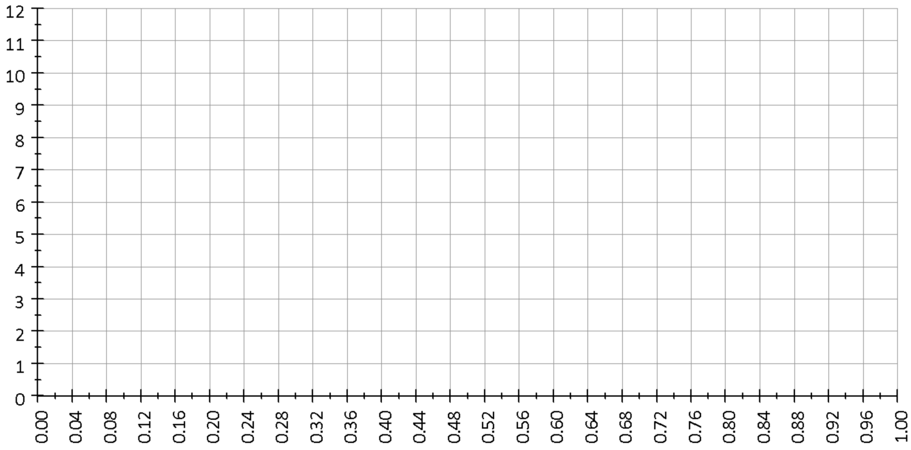
A Count the colors of candies in your sample and fill in the chart below:

	Blue	Brown	Green	Orange	Red	Yellow	Total
Number of Candies							
Proportion of Candies							

B Record the proportion of blue candies in your sample in the table on the board. Do you believe that your sample proportion will be the same as the population proportion? Why?

C As students add their sample proportions to the table, plot them all on the dotplot below. Did everyone have the same sample proportion?

Dotplot for the Sample Proportions of Blue M&M's



D Think about the dotplot created by you and your class. Each dot represents a sample proportion of M&M's that were blue. Give your best estimate of the population proportion of *all* M&M's that is blue.

The dotplot that you and your classmates constructed is part of the *sampling distribution of sample proportions*.

YOU NEED TO KNOW

A **sampling distribution of sample proportions** is the distribution of *all* possible sample proportions from samples of a given size.

Language Tip

A *sampling distribution* is a description of all possible values of a statistic from random samples of a given size.

9 Think about the distribution of sample proportions on the board.

A Describe the shape of the distribution of sample proportions.

B Estimate the mean of the sample proportions in the dotplot.

C In Question 8D you estimated the proportion of all M&M's that is blue. In Question 9B you estimated the mean of all sample proportions. What is the connection between these values?

In statistics we use samples to make inferences about populations. In Module 7 we use sample proportions to make inferences about population proportions. We summarize these ideas below.

YOU NEED TO KNOW

In this lesson we have discussed the following terms.

Population: All individuals or objects of interest to a researcher.

Sample: Part of the population used to represent the entire population.

Parameter: A numerical value that summarizes the population.

Statistic: A numerical value that summarizes the sample.

Success/Failure: The outcome we are/are not interested in for a categorical variable.

Distribution of Sample Proportions: All possible values of the sample proportion and how often it takes on each value. Each sample has the same size.

10 Identify each of the following elements in the M&M activity.

A Success:

B Failure:

C Population:

D Parameter:

E Sample:

F Statistics:

G Sampling Distribution:

NEXT STEPS

In the next part of this lesson, we use a computer to further simulate part of the sampling distribution of sample proportions of blue M&M's. Counting out candies is time consuming and inefficient, but technology can be used to better simulate a distribution of sample proportions. We now use a computer to simulate additional sample proportions.

Download and open the *Blue M&M's* simulation at

http://bit.ly/SWdownloads

(*Note*: Be sure to enable editing or macros if prompted to do so.) The input fields for the spreadsheet in the simulation are as follows: p is the population proportion, and n is the sample size. The simulation creates and plots 1000 sample proportions, from 1000 different samples. Once you enter values for the population proportion or sample size, the simulation automatically creates and plots the updated distribution of sample proportions.

TRY THESE

Simulating a Distribution of Sample Proportions

11 Let's set p to 0.24 and n to 25. Here we assume that 24% of all M&M's are blue and that we are taking samples of 25, like the samples we had in our cups.

A With n = 25 and p = 0.24, describe the shape of the distribution of sample proportions.

B Visually estimate the mean of the distribution of simulated sample proportions. How does your estimate compare to the population proportion of M&M's that is blue?

C How does this distribution compare to the one our class constructed on the board in terms of shape, center, and spread?

D The population proportion of M&M's that is blue is 0.24. Your sample proportion (from Question 8A) probably deviates from this. If your sample proportion is used to estimate the population proportion, *the deviation is an error*. What is the error in your sample proportion?

E Most of the sample proportions deviate from the population proportion. Look at the dotplot, and think about the average error in the sample proportions plotted there. Give your best guess about the value of this average error.

F The simulation displays the standard deviation of the simulated sample proportions. This can be used to describe the *spread* of their distribution. What is this value?

NEXT STEPS

The proportion of all M&M's that is blue is 0.24. In the dotplot of sample proportions from the computer simulation, the mean should be very close to 0.24.

We denote the mean of sample proportions as $\mu_{\hat{p}}$. We have seen that this mean is equal to the population proportion (p).

Mean of sample proportions: $\mu_{\hat{p}} = p$

Language Tip

Different samples will result in different sample proportions. We measure this variability with the standard deviation. Since the sample proportion is an estimate for the population proportion, we can consider the difference between the two to be an error in the estimation. For this reason we call the standard deviation of the sample proportions the *standard error*.

The standard deviation of sample proportions is called the **standard error** of sample proportions. We denote the standard error of sample proportions as $\sigma_{\hat{p}}$.

The formula for the standard error is:

Standard error of sample proportions: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

12 The population proportion of M&M's that is blue is 0.24.

A What is the mean of all sample proportions (from samples of size 25) of M&M's that are blue?

Mean of all sample proportions = $\mu_{\hat{p}}$ = ______________

B What is the standard error of sample proportions (from samples of size 25) of M&M's that are blue?

Standard error of all sample proportions = $\sigma_{\hat{p}}$ = ______________

C How do the mean and standard error you computed in Questions 12A and 12B compare to those you estimated in Questions 11B and 11F?

D Look at the dotplot from the simulation and think about the sample proportions that you and your classmates calculated. What sample proportions would you say are unlikely or unusual? Why?

NEXT STEPS

In the following questions we investigate the role that sample size has upon the distribution of sample proportions.

TRY THESE

The Role of Sample Size

13 We now investigate what happens to the distribution of sample statistics if we change the sample size. Recall that the sample size is the number of M&M's in each sample.

A If we increase the sample size, do you think the distribution of sample proportions will be narrower or wider?

B Change the sample size in the simulation to the values in the table below. Write in the standard deviation of the newly simulated sample proportions for each sample size. Below each standard deviation, compute the standard error for the appropriate sample size. Evaluate if they are similar.

Sample Size	**10**	**100**	**200**
Standard dev. of sample proportions:			
Standard error: $\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}}$			

The standard deviations in the second row are from the 1000 simulated sample proportions. The standard errors in the bottom row are the theoretical standard errors of all sample proportions from samples of the corresponding size. This means that if we could collect every possible sample of 10 M&M's and calculate the standard deviation of all the sample proportions, the answer would be 0.135.

C How do the standard deviations from the simulated sample proportions compare to the standard errors?

D As we increase the sample size, what happens to the spread of the distribution of sample proportions? Was your answer to Question 13A correct?

E As we increase the sample size, what happens to the standard error of sample proportions?

STUDENT NAME ____DATE ______________________

TAKE IT HOME

For your homework open the *Orange M&M's* simulation located at

http://bit.ly/SWdownloads

This spreadsheet simulates distributions of sample proportions of M&M's that are orange, for a particular sample size.

Suppose that 20% of the population of all M&M's is orange. This means the population proportion is p = 0.20.

1 In the spreadsheet, enter the assumed proportion, p = 0.20, of M&M's that are orange. Assume that we sample from samples of size 20 and enter n = 20 into the spreadsheet.

A Estimate the mean of the sample proportions from the dotplot.

B Recall that the mean of the sampling distribution of sample proportions is $\mu_{\hat{p}} = p$, the population proportion. Give the mean of all sample proportions. How does this compare to your answer in Question 1A?

Mean of all sample proportions = $\mu_{\hat{p}}$ = ______________

C The standard deviation of the 1000 sample proportions is an estimate of the standard error of all sample proportions. Give the estimate that is computed in the simulation.

D Use the formula below to compute the standard error of all sample proportions from samples of size n = 20. How does this value compare to your answer to Question 1C?

Standard error of all sample proportions = $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ = ______________

E What will happen to the *standard error* if we increase the sample size?

2 Now we consider how changing the sample size will affect the distribution of sample proportions.

A How would increasing the sample size affect the *mean* of sample proportions?

B For each sample size below, give the values of the standard deviation of the sample proportions from the simulation.

Sample Size (*n*)	**Standard Error**
25	
50	
1000	

C Do the standard deviations of sample proportions support your prediction in Question 1E?

D Using the sample sizes in the table, compute the *standard error*, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, of the sample proportions of orange M&M's in the simulated distribution of sample proportions. Compare these to your estimates in Question 2B.

Sample Size (*n*)	**Computed Standard Error**
25	
50	
1000	

E Describe the dotplots of the sample proportions of orange M&M's in terms of shape and symmetry.

F What effect does increasing the sample size have on the shape, symmetry, and spread of the distribution of sample proportions?