6.1: The Sampling Distribution of Sample Proportions

Recall the statistical analysis process involves four steps:

1. Ask a question that can be answered by collecting data.
2. Decide what to measure and collect the data.
3. Summarize the data and analyze the data.
4. Draw a conclusion and communicate the results to your audience.

In this unit, we turn our focus to the fourth step: using sample data to draw inferences about a population. In particular, we will focus on categorical/qualitative variables. With categorical variables, individuals in the population fall into some category. We summarize categorical data in a sample by calculating a sample proportion. We will then use sample proportions to draw conclusions about population proportions, which is a proportion (portion, percentage, rate, etc.) for the entire population. This process of drawing conclusions about population parameters from what we observe in a sample (sample statistics) is called statistical inference. In upcoming sections, we will use sample statistics to estimate population parameters, and we will use sample statistics to test claims about population parameters.

What Proportion of the Earth is Covered by Water?

Suppose we want to know the proportion of the Earth's surface which is covered in water. If we collected all points on the surface of the Earth, what proportion of them would be on water? Every point on the surface of the Earth can be described by a pair of coordinates called latitudes and longitudes. Latitudes are horizontal lines that are parallel to the equator. They take on values between -90 and 90 degrees. Latitudes below the equator (0 degrees) are considered negative. Longitudes are vertical lines that are perpendicular to latitudes and take on values between -180 and 180 degrees. Longitudes left of the prime meridian at 0 degrees are considered negative. Each point on the surface of the Earth has an associated latitude and longitude pair.

1. Go to https://www.random.org/geographic-coordinates/ and select a random pair of coordinates by clicking the "Pick Random Coordinates" button. Write the latitude and longitude rounded to the nearest whole degree you found below:

   Latitude:

   Longitude:

   Is this random point on water?

For each point on Earth, the variable is whether or not the point is over water. If a single point is over water, we call it success, and if not, we call it a failure. (Recall we saw this language when computing binomial probability). The proportion of all points that are over water is the total number of points that are over water divided by the total number of points. This population proportion is an example of a parameter. A population proportion is denoted by the symbol p.

We calculate the proportion of points that are over water in a sample by dividing the number of points that are over water in the sample by the number of points in the sample (the sample size). The proportion of points on Earth over water in a sample is an example of a statistic, and is denoted by the symbol $\hat{p}$, pronounced p-hat.

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

2. Let's try it! Create a sample of 10 random points and record latitude and longitude rounded to a whole degree, and whether said point is over water in the table below:

	1	2	3	4	5	6	7	8	9	10
latitude/longitude pair										
Over water? (Y/N)										

   a. What is the number of observed successes (count of the number of Yes's from the table)?

   b. What is the number of failures?

   c. Compute the sample proportion.

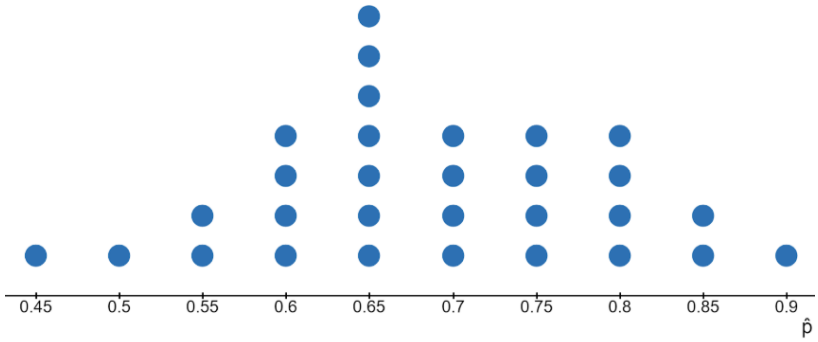
Population	Sample
Collection of all points on Earth	A random sample of 10 points on Earth

Parameter	Statistic
The proportion of points that are over water (p)	Proportion of points that are over water in the sample ($\hat{p}$)

It is important to recognize that there are many samples of points on Earth, each with their own proportion of points over water, but there is only one population proportion. Sample proportions vary from sample to sample, but the population proportion is a single number.

3. Next, we will examine multiple samples of 20 points on Earth. We will calculate a sample proportion $\hat{p}$ for each sample. These proportions are 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.

   The dotplot below shows the results of 30 samples. Each dot represents a sample proportion from a random sample of 20 points on Earth. Use the dotplot to answer the following questions.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

a. What does a dot on the graph represent?

b. Describe the shape of the dotplot.

c. Estimate the center of the distribution.

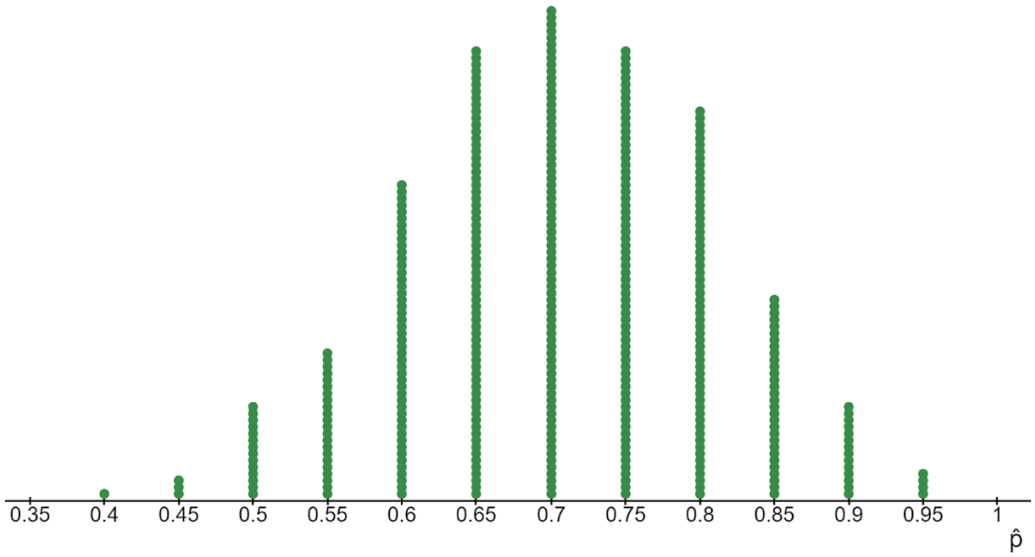
d. What is the range of possible sample proportions?

e. Did each random sample of 20 points on the Earth yield the same sample proportion?

f. Using the dotplot, what is the best estimate of the population proportion of all points on Earth that are over water?

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

We use technology to further simulate part of the sampling distribution of sample proportions of points on Earth over water. Counting out points on a map is time consuming and inefficient, but technology can be used to better simulate a distribution of sample proportions. The dotplot below displays 400 sample proportions. Each sample proportion is based on a random sample of 20 points on Earth.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

4. Use the above distribution to answer the following questions:
   a. What is the mean of the distribution (visually estimated)?

   b. What is the center, shape, and spread of the distribution?

It turns out that sampling distributions of sample proportions become more normal as the sample size increases. A sampling distribution of sample proportions is the distribution of all possible sample proportions from samples of a given size. If the sample size is large enough, this distribution is approximately normal.

Mean and Standard Error of Sampling Distributions

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

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

A sample proportion is an estimate of the population proportion. When a sample proportion deviates from the population proportion, the deviation is an error in the estimate. Because of this, 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:

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

5. 71% of points on the surface of Earth are on water. So, the population proportion of points on Earth covered over water is 0.71. Use the formulas above to compute the mean and standard error of all sample proportions of points on Earth over water when the sample size is 25. Round the standard error to three decimal places.

   Mean of sample proportions =

   Standard error of sample proportions =

Criteria for Approximate Normality

Statisticians have learned that sampling distributions of sample proportions are approximately normal whenever $np \geq 10$ and $n(1-p) \geq 10$. Since p is the proportion of successes, and n is the sample size, np is the expected, or mean, number of successes in a sample of size n. That is, on average, samples of size n will have np successes. Similarly, $1-p$ is the proportion of failures, and $n(1-p)$ is the expected number of failures. Thus, the sampling distribution of sample proportions is approximately normal if it meets the criteria for approximate normality, which requires there are at least 10 expected successes and 10 expected failures in the sample.

As an example, suppose we sample 50 points on Earth and assume that 71% of all points on Earth are over water. The expected number of successes in a sample is $np = 35.5$ and the expected number of failures is $n(1-p) = 50(0.29) = 14.5$ which could also be calculated as the sample size minus the number of failures.

This means that, on average, samples of 50 points on Earth contain 35.5 points over water and 14.5 points over land. Of course, the number of points over water will vary in individual samples. But since both of the numbers are greater than 10, the normal distribution is a good approximation for the sampling distribution of sample proportions of points on Earth over water in samples of size 50.

You try!

6. When a sampling distribution of sample proportions satisfies the normality criteria we can use the normal distribution properties to find probabilities corresponding to sample proportions.

   The Gallup organization conducts surveys in countries throughout the world to obtain categorical and quantitative data on people and their views about important issues. Gallup surveyed people in the United States in March 2019 to obtain information on U.S. adults' views regarding global warming. They found that 51% of U.S. adults are "concerned believers" who take global warming seriously and believe it poses a serious threat within their lifetime.

   a. For this problem, let's assume that Gallup's result (0.51) is the proportion of all U.S. adults who take global warming seriously. Suppose we sample 120 U.S. adults and determine the proportion who take global warming seriously. We can apply the Central Limit Theorem to describe the sample proportions that are likely to occur given the sample size and assumed population proportion.

      i. What is the sample size (n) and population proportion (p)?

      ii. Sample size, $n =$ _____

      iii. Population proportion, $p =$ _____

   b. Let's explore if the normality criteria are met. First, find the following values. Round answers to one decimal place.

      i. $np =$ _____

      ii. $n(1-p) =$ _____

   c. Are the normality criteria met? Explain.

   d. Find the mean and compute the standard error of the sampling distribution of sample proportions (use three decimal places for the standard error).

      i. Mean $= \mu_{\hat{p}} = p =$ _____

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

   e. Which statement below best describes the standard error of the sampling distribution of sample proportions? Select an answer that is incorrect and explain why.

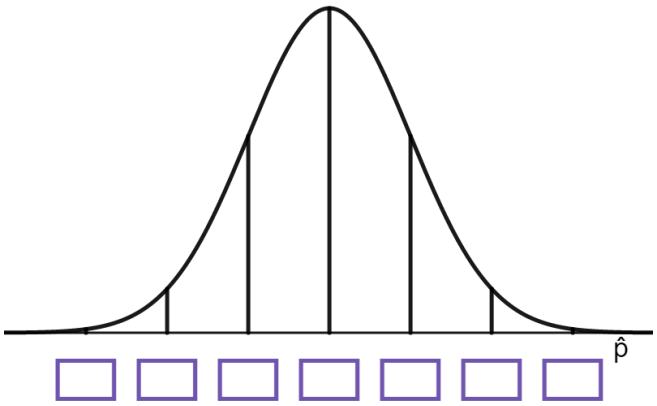
      i. Sample proportions vary, and the difference between the lowest possible sample proportion and highest possible sample proportion is 0.046.

      ii. 0.046 is the typical distance which sample proportions are from the population proportion.

      iii. All sample proportions are within 0.044 from the population proportion.

      iv. No sample proportions equal the population proportion.

f. The boxes under the normal distribution below are one standard error apart, with the center box at the mean. Use the mean and standard error above to enter the correct values into the boxes.



Images are created with the graphing calculator, used with permission from Desmos Studio PBC.

g. Suppose that in a random sample of 120 U.S. adults 48 respond that they take global warming seriously. Compute the sample proportion, $\hat{p}$. Write your answer as a decimal rounded to two decimal places.

h. What is the Z-score of the sample proportion from part g? Use the mean and standard error from part d.

i. Does this Z-score indicate that this sample proportion is unusual? Explain how you know.

j. Use desmos to find the probability of observing a sample proportion that is less than or equal to the value from part g. Round your answer to four places after the decimal. Write the function you used in desmos to find the result.

Summary

- The Central Limit Theorem for sample proportions states that a sampling distribution of sample proportions is approximately normal if the sample size is large enough. Our criteria for determining this are: $np \geq 10$ and $n(1-p) \geq 10$.
- When the criteria for approximate normality are satisfied, the normal distribution may be used to determine probabilities about sample proportions.
- The mean and standard error (or standard deviation) for the sampling distribution of sample proportions are given by:

$$\text{Mean } = \mu_{\hat{p}} = p$$

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