# **Chapter 7: Sampling Distributions**

Section 7.1 What is a Sampling Distribution?



### Introduction

The process of *statistical inference* involves using information from a sample to draw conclusions about a wider population.

Different random samples yield different statistics. We need to be able to describe the *sampling distribution* of possible statistic values in order to perform statistical inference.

We can think of a statistic as a random variable because it takes numerical values that describe the outcomes of the random sampling process. Therefore, we can examine its probability distribution using what we learned in Chapter 6.



**Collect data** from a representative **Sample**...

Make an **Inference** about the **Population**.

## **Parameters and Statistics**

As we begin to use sample data to draw conclusions about a wider population, we must be clear about whether a number describes a sample or a population.

## **Definition:**

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

K

 $\mathbf{F}$ 

\*In statistical practice, the value of a parameter is usually not known because we cannot examine the entire population. (Parameters are denoted with the symbols:  $\mu$ , p,  $\sigma$ )

A statistic is a number that describes some characteristic of a sample.

\*The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter. (Statistics are denoted with the symbols:  $\overline{x}$ ,  $\hat{p}$ , s)

Remember s and p: statistics come from samples and parameters come from populations **Example 1:** Identify the population, the parameter, the sample, and the statistic in each of the following settings.

a) The Gallup Poll asked a random sample of 515 U.S. adults whether or not they believe in ghosts. Of the respondents, 160 said "Yes."

\*The population is all U.S. adults. \*The parameter of interest is p, the proportion of all U.S. adults who believe in ghosts. \*The sample is the 515 people who were interviewed in this Gallup Poll. \*The statistic is  $\hat{p} = \frac{160}{515} = 0.31$ , the proportion of the sample who say they believe in ghosts. b) During the winter months, the temperatures outside the Starnes' cabin in Colorado can stay well below freezing (32°F, or 0°C) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50°F. She wants to know how low the temperature actually gets in the cabin. A digital thermometer records the indoor temperature at 20 km randomly chosen times during a given day. The minimum reading is 38°F.

\*The population is all times during the day in question. \*The parameter of interest is the minimum temperature in the cabin that day. \*The sample consists of the 20 temperature readings at randomly selected times. \*The statistic is the sample minimum, 38°F.

## **Sampling Variability**

How can  $\overline{x}$  be an accurate estimate of  $\mu$ ? After all, different random sample would produce different values of  $\overline{x}$ .

This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?"

\*Take a large number of samples from the same population.

\*Calculate the statistic (like the sample mean or sample proportion ) for each sample.

\*Make a graph of the values of the statistic.

\*Examine the distribution displayed in the graph for shape, center, and spread, a well as outliers or other unusual features.



## **Definition:**

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

In practice, it's difficult to take all possible samples of size *n* to obtain the  $\Re$  actual sampling distribution of a statistic. Instead, we can use simulation to imitate the process of taking many, many samples.

**Example 2:** We used Fathom software to simulate choosing 500 SRSs of size n = 20 from a population of 200 chips, 100 red and 100 blue. The figure below is a dotplot of the values of the sample proportion of red chips,  $\hat{p}$  from these 500 samples.

a) There is one dot on the graph at 0.15. Explain what this value represents.

In one SRS of 20 chips, there were 3 red chips. So  $\hat{p} = 3/20 = 0.15$  for this sample.



b) Describe the distribution. Are there any obvious outliers?

Shape: Symmetric and somewhat bell-shaped. Center: Around 0.5. Spread: The values of  $\hat{p}$  fall mostly between 0.25 and 0.75. Outliers: One sample with  $\hat{p} = 0.15$  stands out. c) Would it be surprising to get a sample proportion of 0.85 or higher in an SRS of size 20 when p = 0.5? Justify your answer.

It is very unlikely to obtain an SRS of 20 chips in which  $\hat{p} = 0.85$  from a population in which p = 0.5. A value of  $\hat{p}$  this large or larger never occurred in 500 simulated samples.

d) Suppose your teacher prepares a bag with 200 chips and claims that half of them are red. A classmate takes an SRS of 20 chips; 17 of them are red. What would you conclude about your teacher's claim? Explain.

This student's result gives strong evidence against the teacher's claim. As noted in part (c), it is very unlikely to get a sample proportion of 0.85 or higher when p = 0.5.

#### **Population Distributions vs. Sampling Distributions**

There are actually three distinct distributions involved when we sample repeatedly and measure a variable of interest.

1. The **population distribution** gives the values of the variable for all the individuals in the population.

2. The **distribution of sample data** shows the values of the variable for all the individuals in the sample.

3. The **sampling distribution** shows the statistic values from all the possible samples of the same size from the population.



## **Describing Sampling Distributions**

The fact that statistics from random samples have definite sampling distributions allows us to answer the question, "How trustworthy is a statistic as an estimator of the parameter?" To get a complete answer, we consider the center, spread, and shape.

## **Center: Biased and unbiased estimators**

In the chips example, we collected many samples of size 20 and calculated the sample proportion of red chips. How well does the sample proportion estimate the true proportion of red chips, p = 0.5?



Note that the center of the approximate sampling distribution is close to 0.5. In fact, if we took ALL possible samples of size 20 and found the mean of those sample proportions, we'd get *exactly* 0.5.

## **Definition:**

A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

×,

## Spread: Low variability is better!

**Example 3:** Television executives and companies who advertise on TV are interested in how many viewers watch particular shows. According to Nielsen ratings, Survivor was one of the most-watched television shows in the United States during every week that it aired. Suppose that the true proportion of U.S. adults who have watched Survivor is p = 0.37.

The top dotplot in the figure on the next slide shows the results of drawing 400 SRSs of size n = 100 from a population with p = 0.37. We see that a sample of 100 people often gave a  $\hat{p}$  quite far from the population parameter. That is why a Gallup Poll asked not 100, but 1000 people whether they had watched *Survivor*. Let's repeat our simulation, this time taking 400 SRSs of size n = 1000 from a population with proportion p = 0.37 who have watched *Survivor*. The bottom dotplot in the figure below displays the distribution of the 400 values of  $\hat{p}$  from these new samples. Both graphs are drawn on the same horizontal scale to make comparison easy.



We can see that the spread of the top dotplot in the figure above is much greater than the spread of the bottom dotplot. With samples of size 100, the values of  $\hat{p}$  vary from 0.25 to 0.54. The standard deviation of these  $\hat{p}$ -values is about 0.05. Using SRSs of size 1000, the values of  $\hat{p}$ -values is about 0.05. Using SRSs of size 1000, the values of  $\hat{p}$  only vary from 0.328 to 0.412. The standard deviation of these  $\hat{p}$ -values is about 0.015, so most random samples of 1000 people give a p that is within 0.03 of the actual population parameter, p = 0.37.

Larger samples have a clear advantage over smaller samples. They are much more likely to produce an estimate close to the true value of the parameter.

The **variability of a statistic i**s described by the spread of its sampling distribution. This spread is determined mainly by the size of the random sample. Larger samples give smaller spreads. The spread of the sampling distribution does not depend much on the size of the population, as long as the population is at least 10 times larger than the sample.

Taking a larger sample doesn't fix bias. Remember that even a very large voluntary response sample or convenience sample is worthless because of bias.



## Bias, variability, and shape

We can think of the true value of the population parameter as the bull's- eye on a target and of the sample statistic as an arrow fired at the target. Both bias and variability describe what happens when we take many shots at the target.



**Bias** means that our aim is off and we consistently miss the bull's-eye in the same direction. Our sample values do not center on the population value.

High **variability** means that repeated shots are widely scattered on the target. Repeated samples do not give very similar results.

The lesson about center and spread is clear: given a choice of statistics to estimate an unknown parameter, choose one with no or low bias and minimum variability. **Example 3:** Refer to the Activity we did at the beginning of this chapter. Mrs. Friedman's student teams came up with four different methods for estimating the number of tanks in the bag: (1) maxmin = maximum + minimum, (2) meanpl2sd =  $\overline{x} + 2s_x$ , (3) twicemean =  $2\overline{x}$  and (4) twomedian = 2M. She added one more method, called "partition." The figure below shows the results of taking 250 SRSs of 4 tanks and recording the value of the five statistics for each sample. The vertical line marks the actual value of the population parameter N; there were 342 tanks in the bag.



a) Which of the four statistics proposed by the student teams is the best estimator? Justify your answer.

Meanpl2sd is a biased estimator: the center of its sampling distribution is too high. This statistic produces consistent overestimates of the number of tanks. The other three statistics proposed by the students appear to be unbiased estimators. All three sampling distributions have roughly symmetric shapes, so these statistics are about equally likely to underestimate or overestimate the number of tanks. Because maxmin has the smallest variability among the three, it would generally produce estimates that are closer to the actual number of tanks. Among the students' proposed statistics, maxmin would be the best estimator. b) Why was the partition method, which uses the statistic  $(5/4) \cdot \text{maximum}$ , recommended by the mathematicians in Washington, D.C.?

The partition method uses a statistic  $(5/4 \cdot \text{maximum})$  that is an unbiased estimator and that has much less variability than any of the student teams' statistics. Its sampling distribution is left-skewed, so the mean of the distribution is less than its median. Because more than half of the dots in the graph are to the right of the mean, the statistic is more likely to overestimate than underestimate the number of tanks. The mathematicians believed that it would be better to err on the side of caution and give the military commanders an estimate that is slightly too high.

**AP EXAM TIP:** Terminology matters. Don't say "sample distribution" when you mean sampling distribution. You will lose credit on free-response questions for misusing statistical terms.