AP Statistics
Unit 05 – Distributions
Day 01 Notes – Sampling Distributions (7.1)

Name Key
Period _____

In this section, we will be introduced to the "big ideas" of sampling distributions. The **first big idea** is the difference between a statistic and a parameter – we briefly went over this in unit 01. A parameter is a number that describes some characteristic of a population. A statistic estimates the value of a parameter using a sample from the population. Although it may seem simple when introduced, the concept is a crucial distinction that people often fail to make when they perform statistical inference.

The **second big idea** is that statistics vary. For example, the mean weight in a sample of high school students is a variable that will change from sample to sample. This means that statistics have distributions, whereas parameters do not.

The **third big idea** is the difference between the distribution of the population, the distribution of the sample, and the sampling distribution of a statistic.

The **fourth big idea** is how to describe a sampling distribution. To adequately describe a sampling distribution, we need to address its shape, center, and spread. If the mean of the sampling distribution is the same as the value of the parameter being estimated, then the statistic is called an unbiased estimator. Ideally, the spread of a sampling distribution will be very small, meaning that the statistic is very precise. Larger sample sizes result in sampling distributions with smaller spreads.

**Parameter:** a number that describes some characteristic of the population.

Parameter = population

common parameters: $\mu$ (mean) — also known as the "true mean"
$\sigma$ (standard deviation)
$p$ (proportion) — also known as the "true proportion"

**Statistic:** a number that describes some characteristic of a sample.

Statistic = sample

common statistics: $\bar{x}$ (sample mean)
$\hat{p}$ (sample proportion)

**Example: From Ghosts to Cold Cabins**

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 US adults whether or not they believe in ghosts. Of the respondents, 160 said "Yes."

population: all US adults
parameter: the true proportion of all US adults who believe in ghosts.
Sample: 515 US adults interviewed in this Gallup Poll.
Statistic: $\hat{p} = 160/515 = 0.31$ the proportion of the sample who say they believe in ghosts.

b. During the winter months, the temperature outside the Starneses' cabin in Colorado can stay well below freezing (32 degrees F or 0 degrees C) for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50 degrees F. She wants to know how low the temperature actually gets in the cabin. A digital thermometer records the indoor temperature at 20 randomly chosen times during a given day. The minimum reading is 38 degrees F.

population: all times during the given day.
parameter: true minimum temperature in the cabin that day.
Sample: 20 temperature readings at randomly selected times.
Statistic: the sample minimum: 38°F

CHECK YOUR UNDERSTANDING:

Each boldface number in Questions 1 & 2 is the value of either a **parameter** or a **statistic**. In each case, state which it is and use appropriate notation to describe the number.

1. On Tuesday, the bottles of Arizona Iced Tea filled in a plant were supposed to contain an average of **20** ounces of iced tea. Quality control inspectors sampled 50 bottles at random from the day's production. These bottles contained an average of **19.6** ounces of iced tea.

$\mu = 20$ ounces.

$\bar{x} = 19.6$ ounces.

2. On a NY to Denver flight, **8%** of the 125 passengers were selected for random security screening before boarding. According to the Transportation Security Administration, **10%** of passengers at this airport are chosen for random screening.

$p = 0.10$ or 10% of passengers.

$\hat{p} = 0.08$ or 8% of the sample of passengers.

**Sampling Variability:** the value of a statistic varies in repeated random sampling. It is different from sample to sample.

*Activity: Reaching for Chips*

**Sampling Distribution:** the distribution of values taken by a statistic in all possible samples of the same size from the same population.

**Example: Reaching for Chips**

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 dot plot of the values of $\hat{p}$, the sample proportion of red chips, 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, unimodal, and somewhat bell-shaped.

Center: around 0.5.

Spread: the values of $\hat{p}$ fall mostly between 0.25 and 0.75.

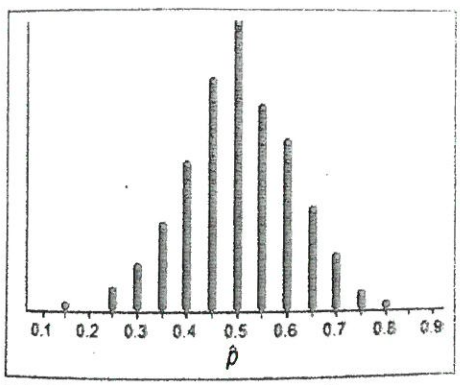



FIGURE 7.2 Dotplot of the sample proportion $\hat{p}$ of red chips in 500 simulated SRSs, created by Fathom software.

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. A value of $\hat{p}$ this large or larger never occurred in 500 simulated samples.

d. Suppose Mrs. De Marre 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 Mrs. DeMarre'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 Distribution:** gives the values of the variable for all individuals in the population.

**Distribution of Sample Data:** gives the values of the variable for all individuals in the sample.

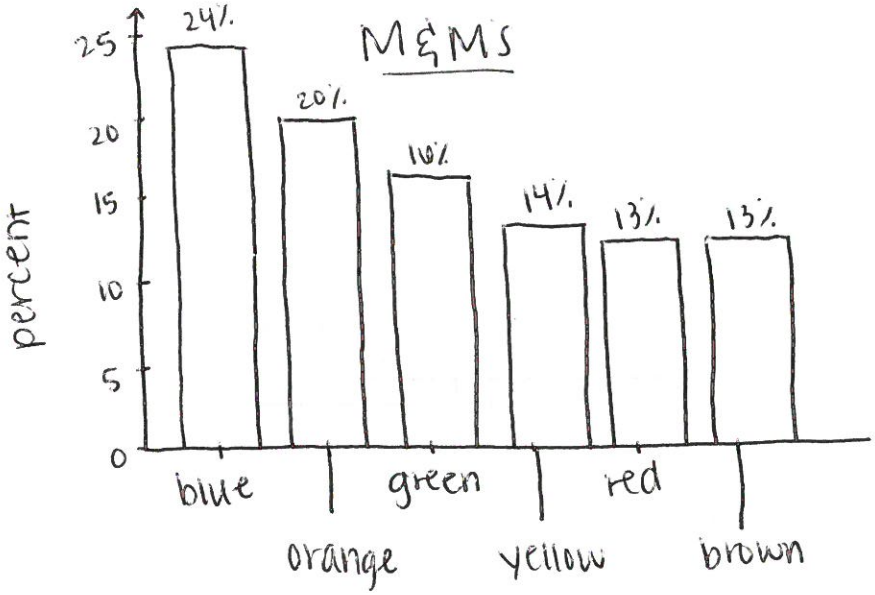
* *Be careful: The population distribution and the distribution of sample data describe individuals. A sampling distribution describes how a statistic varies in many samples from the population.* AP Tip: Sampling distribution ≠ sample distribution

**CHECK YOUR UNDERSTANDING:**

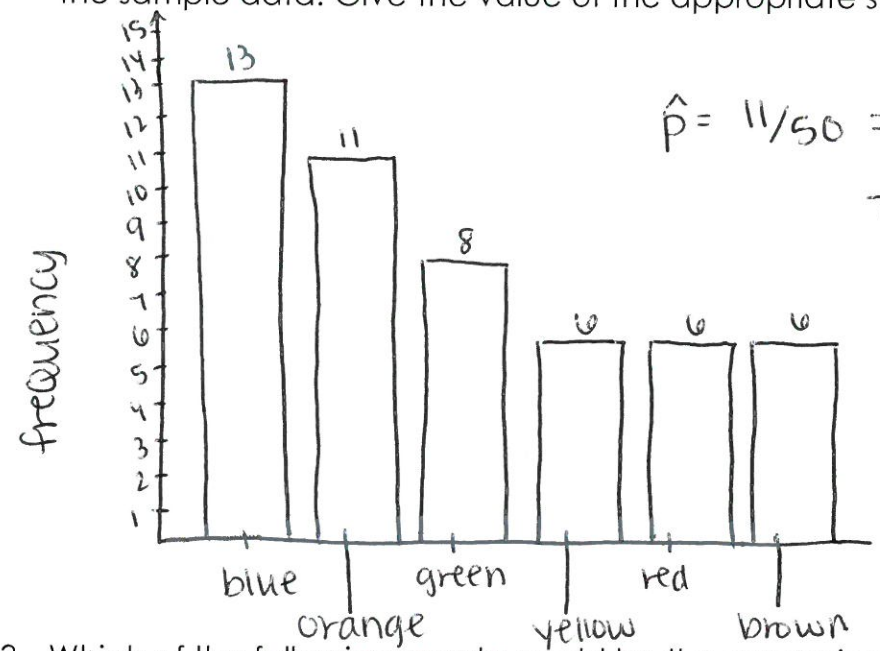
Mars, Incorporated, says that the mix of colors in its M&M'S Milk Chocolate Candies is 24% blue, 20% orange, 16% green, 14% yellow, 13% red, and 13% brown. Assume that the company's claim is true. We want to examine the proportion of orange M&M'S in repeated random samples of 50 candies.

1. Graph the population distribution. Identify the individuals, the variable, and the parameter of interest.

Individuals: M&Ms Milk chocolate candies
Variable: color
Parameter of interest: proportion of orange M&Ms



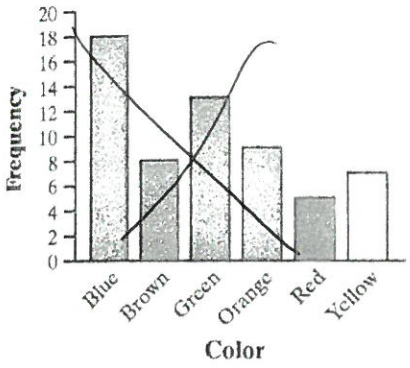
2. Imagine taking an SRS of 50 M&M'S. Make a graph showing a possible distribution of the sample data. Give the value of the appropriate statistic for this sample.



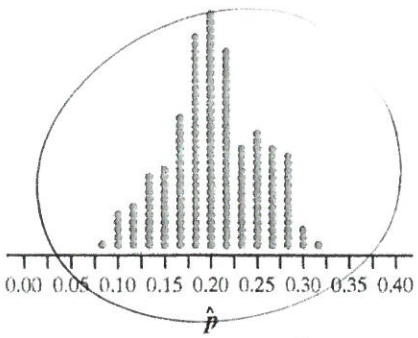

$\hat{p} = 11/50 = 0.22$

The sample proportion of orange M&Ms is 0.22 for this sample.

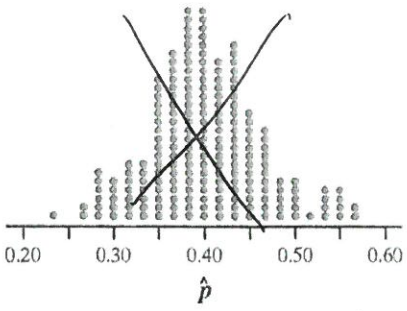
3. Which of the following graphs could be the approximate sampling distribution of the statistic? Explain.




this graph shows the distribution of M&M colors, not the distribution of the sample proportion of orange M&Ms.




centered @ 0.2




this graph is centered at about 0.4 when it should be centered at about 0.2.

**Unbiased estimator:** (explanation on page 431)

A statistic used to estimate a parameter is an unbiased estimator if the mean of it's sampling distribution is equal to the value of the parameter being estimated.

$\hat{p}$ $\bar{x}$ var

(when we divide by n-1)

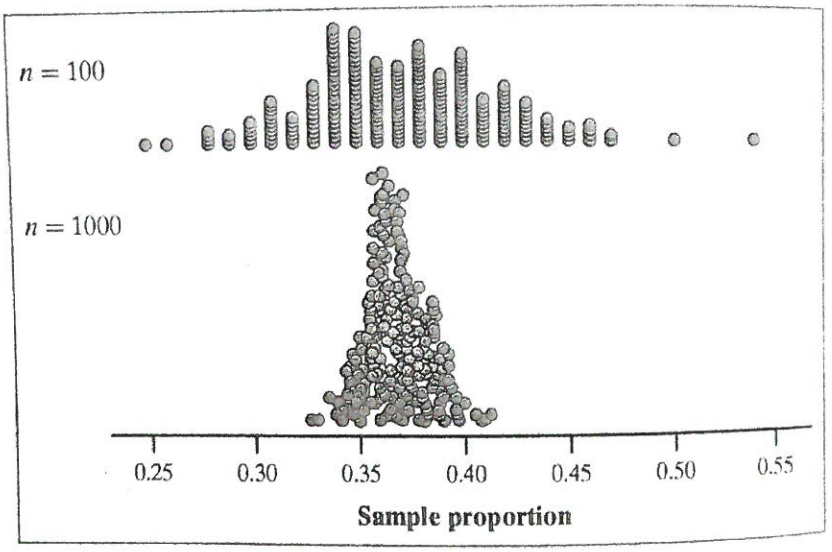
Spread: Low variability is better!

**Example: Who Watches Survivor?**

*Why sample size matters*

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 US during every week that it aired. Suppose that the true proportion of US adults who have watched *Survivor* is p = 0.37.

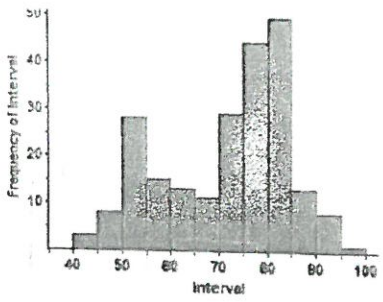
The top dotplot in the figure below shows the results of drawing 400 SRSs of size n = 100 from a population with p = 0.37. We see that a sample of 100 people often gave a $\hat{p}$ quite far from the population parameter. That is why a Gallup Poll asked not 100, but 1000 people whether they had watched *Survivor*. The bottom dotplot 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.

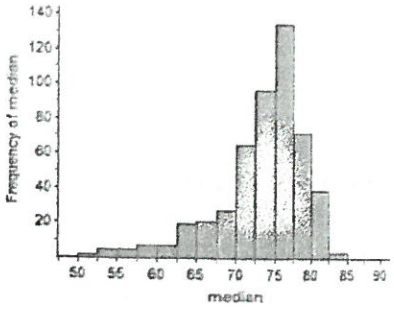



We can see that the spread of the top dotplot is much greater than the spread of the bottom dotplot. With samples of size 100, the values of $\hat{p}$ vary from 0.25 to 0.54. The standard deviation of these $\hat{p}$-values is about 0.05. Using SRSs of size 1000, the values of $\hat{p}$ only vary from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

**Variability of a Statistic:** is 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.

**CHECK YOUR UNDERSTANDING:**



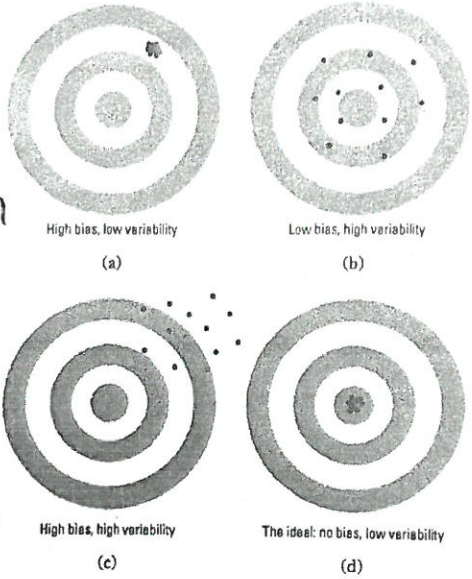




The histogram above left shows that intervals (in minutes) between eruptions of Old Faithful geyser for all 222 recorded eruptions during a particular month. For this population, the median is 75 minutes. We used Fathom software to take 500 SRSs of size 10 from the population. The 500 values of the sample median are displayed in the histogram above right. The mean of these 500 values is 73.5.

1. Is the sample median an unbiased estimator of the population median? Justify your answer. NO. The mean of the approximate sampling distribution of the sample median (73.5) is not equal to the median of the population.

2. Suppose we had taken samples of size 20 instead of size 10. Would the spread of the sampling distribution be larger, smaller, or about the same? Justify your answer.
   Smaller. Larger samples provide more precise estimates because larger samples include more information about the population distribution.
3. Describe the shape of the sampling distribution.
   Unimodal and skewed to the left.

**Bias, Variability, and Shape:**




Bias: our aim is off and we consistently miss the bull's-eye in the same direction

High Variability: repeated shots are widely scattered on the target. Repeated samples do not give very similar results.

Bull's-eye = true population parameter
arrow fired = sample statistic

We want our estimates to be accurate (unbiased) and precise (have low variability).