CHAPTER 7: SAMPLING DISTRIBUTIONS

Part 1: What is a Sampling Distribution?

Introduce with German Tank Problem pg 422

[redacted]

Parameter:

Statistic:

Parameter = a number that describes some characteristic of the population; mu for mean, p for population proportion

Statistic = number that describes some characteristic of a sample; xbar for mean, phat for sample proportion

**Statistic comes from Samples; Parameter comes from Population

[redacted]

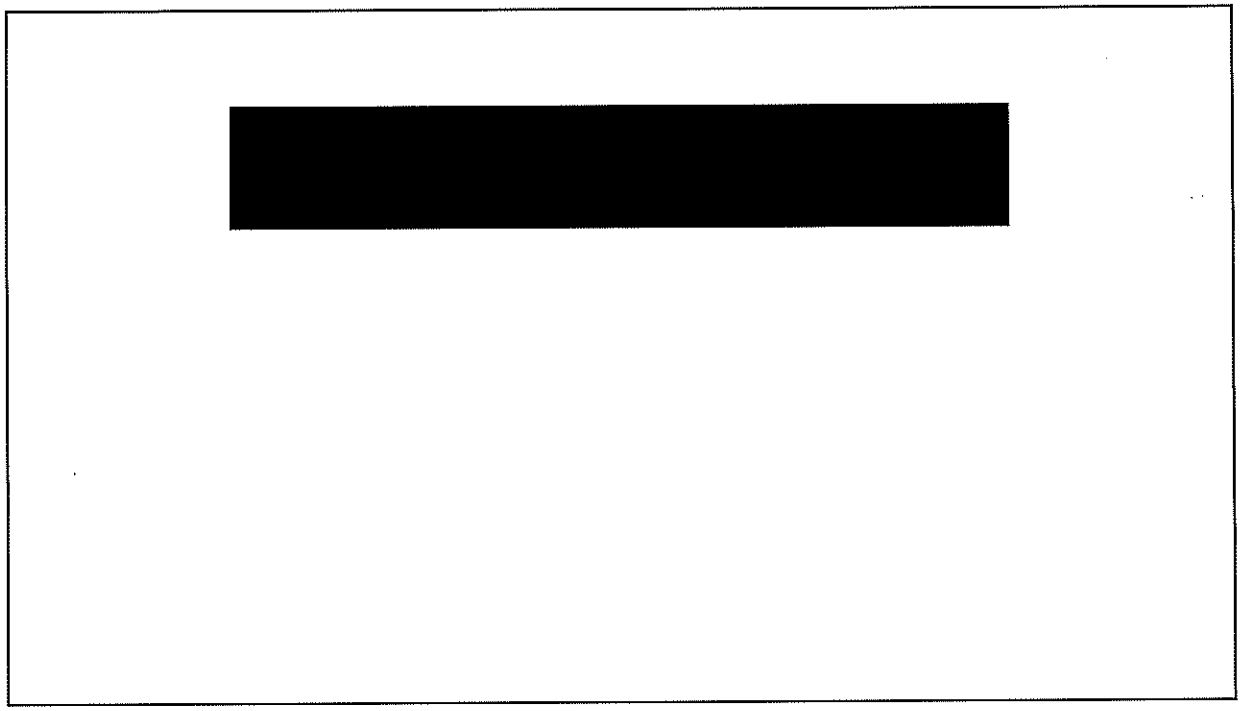
Identify the population, the parameter, the sample, and the statistic in each of the following settings:

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."

b) During the winter months, the temperature outside a cabin in Colorado can stay well below freezing for weeks at a time. To prevent the pipes from freezing, Mrs. Starnes sets the thermostat at 50 degrees Fahrenheit. 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 Fahrenheit.

a) Population = all US adults, parameter of interest p is the proportion of all US adults who believe in ghosts. The sample is the 515 people who were intervewed. The statistic phat = 160/515 = 0.31, the porportion of the sample who say they believe in ghosts.

b) The population is all times during the day in question; parameter of interest is the true minimum temperature in the cabin that day. Sample consists of 20 randomly selected timed temperatures, and the statistic is the sample minimum = 38 degrees



Variability: "The idea that different samples render different results."

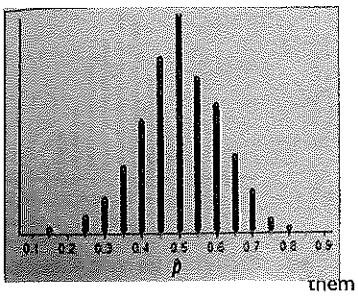
To make sense of sampling variability:

1. Take a large number of samples from the same population
2. Calculate the statistic (whatever that may be) for each sample
3. Make a graph of the values of the statistic
4. Examine the distribution displayed in the graph for SOCS

Distribution: the distribution of values taken by the statistic in all possible samples of the same size from the same population
usually hard to do – that's why we have simulation!
the ideal pattern that emerges if we look at all possible samples
Try Activity on pg 426

[redacted]

Fathom software (statistics software) was used to simulate choosing 500 SRSs of size n=20 from a population of 200 poker chips, 100 red and 100 blue. The figure is a dotplot 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.

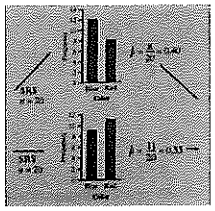
b) Describe the distribution.

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.

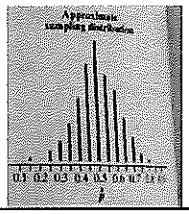
d) Suppose your teacher prepares a bag with 200 chips and claims that half of them are red. A classmate takes an SRS of 20 chips; 17 of them are red. What would you conclude about your teacher's claim? Explain.

[redacted]

- Population distribution:



- Distribution of sample data:



- Sampling Distribution:

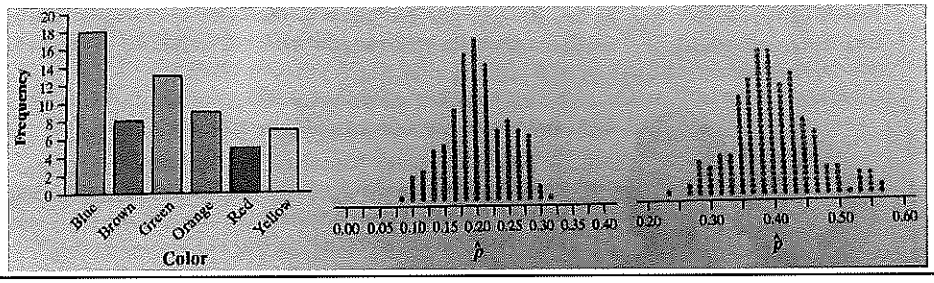
Population = gives the values of the variable for all individuals in the population
Distribution of sample = shows the values of the variable for the individuals in a particular sample
Sampling = describes how a statistic varies in many samples from a population

***TERMINOLOGY MATTERS!!! You can't just assume that AP readers know what you mean - if you misspeak they will count off!

[redacted]

Mars, Inc. 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.
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.
3. Which of the following graphs could be the approximate sampling distribution of the statistic? Explain your choice.



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

Center: Biased and unbiased Estimators

Returning to the poker chips example, we know the population proportion is $p = 0.5$. How well does the sample proportion estimate the population proportion? Review the sample distribution dotplot from the previous slide and note the center of that distribution (it should be close to 0.5). What's even better, if we averaged the centers of all possible distributions of sample data, we would get exactly 0.5. For this reason, we say that $\hat{p}$ is an unbiased estimator.

Unbiased estimator:

Biased estimator:

Unbiased estimator: a statistic used to estimate a parameter is an unbiased estimator if the mean of its sampling distribution is equal to the value of the parameter being estimated

Biased: the center of the sampling distribution for the statistic does not equal the value of the parameter being estimated

**you can use estimators in any capacity – mean, SD, range, etc.

ACTIVITY: Sampling heights pg 430

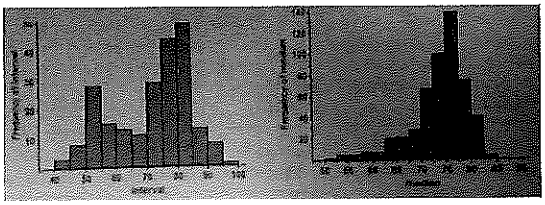
Spread: Low variability is better!
To get a trustworthy estimate of an unknown population parameter, start by using a statistic that's an unbiased estimator. This ensures you won't consistently overestimate or underestimate the parameter.

The sample proportion $\hat{p}$ from a random sample of any size is an unbiased estimator of the parameter p. Larger random samples, though, have an advantage because they give us more precise estimates than smaller random samples.

Variability of a Statistic:

****Taking a larger sample doesn't fix bias though!! Remember that even a very large voluntary resonse sample or convenience sample is worthless because of bias

Variability = described by the spread of its sampling distribution; this spread is determined by mainly 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 (10% condition)



The histogram above left shows the 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 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.
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.
3. Describe the shape of the sampling distribution.

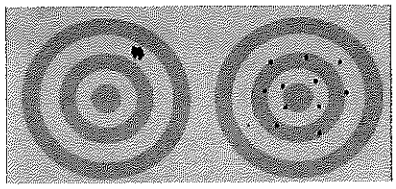
[redacted]

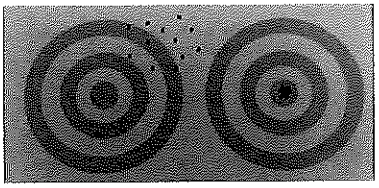
Bias, variability, and shape

We can think of the true value of a population parameter as the bull's eye on a target and 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 = aim is off; our sample values do not center on the population value

High Variability = repeated shots are widely scattered; repeated samples do not give similar results





[redacted]

The German Tank Problem

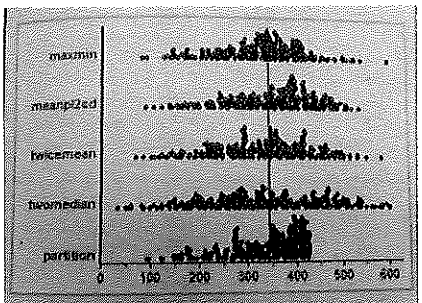


FIGURE 7.9 Results from a Fathom simulation of 250 SRSs of 4 tanks. The approximate sampling distributions of five different statistics are shown.

Refer back to the Activity we did on the German tanks. Another teacher, Mrs. Friedman, had her student teams come up with four different methods for estimating the number of tanks in the bag: (1) "maxmin" = max + min, (2) "meanpl2sd" = $\bar{x} + 2s_x$, (3) "twicemean" = $2\bar{x}$, and (4) "twomedian" = 2(median). She added one more method, called "partition." The figure 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.

Problem: Use the information in the figure to help answer these questions:

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

b) Why was the partition method, which uses the statistic (5/4)•maximum, recommended by the mathematicians in Washington, DC?