7.1 Lecture Notes & Examples

Section 7.1 - What is a Sampling Distribution? (pp. 414-428)

1. Parameters and Statistics

Hyena Problem - In the hyena problem, we took repeated samples from a population to try to determine the proportion of male hyenas in the population. The proportions you found in the samples are called **statistics** because they describe **samples**. The actual proportion in the population is called a **parameter** because it is from the **population**.

Parameter - a number that describes some characteristic of the *population*. In statistical practice, the value of a parameter is usually not known because we cannot examine the entire population.

Statistic - 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 statistics to estimate an unknown parameter.

It is essential from this point on that we always distinguish parameters and statistics (p=P, s=S). For example,

(μ = mu) μ is the population mean so it is a parameter
while $\bar{x}$ is a statistic because it is the sample mean.

For proportions, p is the population proportion (a parameter)
while $\hat{p}$ is the sample proportion (a statistic).

$\hat{P}$ is read "p hat"

Parameters	Statistics
μ - population mean	$\bar{x}$ – sample mean
σ – population standard deviation	s – sample standard deviation
p = proportion of a population (Ex: What proportion of HFII Students own iPhone?)	$\hat{p}$ = sample proportion (ex: Ask 100 HFII Students if they own an iPhone? What is the proportion of those 100 who own an iPhone?

Example - A pediatrician wants to know the 75th percentile for the distribution of heights of 10-year-old boys, so she takes a sample of 50 patients and calculates Q_3 = 56 inches.

What is the population? All 10 year old boys

What is the parameter? (a statistic of the population) 75th percentile

What is the sample? 50 10 year old patients who are boys

What is the statistic? $Q_3 = 56$ inches

CHECK YOUR UNDERSTANDING

Each boldface number in Questions 1 and 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.

Parameter is $\mu = 20$ oz of iced tea

Statistic is $\bar{x} = 19.6$ oz of iced tea

2. On a New York–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.

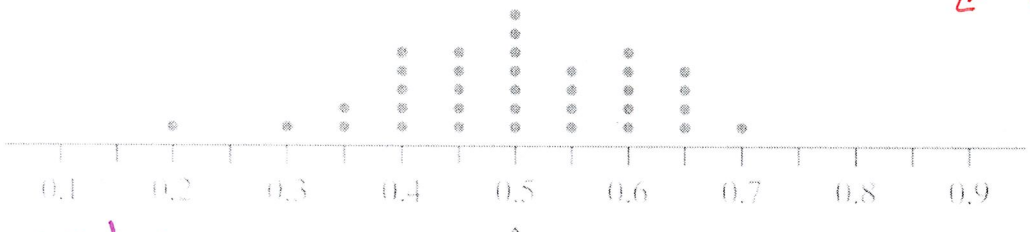
Parameter is $p = .10$ or 10% of passengers

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

2. Sampling Variability - As we saw in the hyena problem, the various samples for any given population produced different sample proportions. This basic fact is called **sampling variability**. In the hyena experiment, we:

- Took repeated samples from the same population,
- Calculated the sample proportion $\hat{p}$ for each sample,
- Made a graph of the values of the statistic, (dot plots)
- Examined the distribution displayed in the graph for shape, center, and spread, as well as outliers and other deviations. (SOCS)

Suppose one group in the hyena experiment took 35 samples of size 20 and their results are shown in the dotplot below.



This is not an exact sampling distribution since the group did not take all possible samples of size 20.

This is an approximate sampling distribution.

$\hat{p}$ = sample proportion males

Shape: roughly symmetric, roughly bell shaped unimodal with peak at 0.5

Center: Mean = 0.499 (Balance Point)

Spread: Stand. Dev. = 0.112: On average, values of $\hat{p}$ (sample proportion) are 0.112 from mean of 0.499. The values vary from 0.2 to 0.7

Outliers: No outliers

Of course, this group only took 35 different simple random samples of 20 hyenas. There are many, many possible SRSs of size 20 from a population of 100. If we took every one of those samples, calculated $\hat{p}$ for each, and graphed all those $\hat{p}$-values, we would have a ***sampling distribution***.

all possible samples of 20 from 100 is $\binom{100}{20} \approx 5.36 \times 10^{20}$

Sampling Distribution - 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.

The sampling distribution is an ideal pattern that would emerge if we looked at *all* possible samples from a given population.

(Theory)

As the figure below shows, there are actually three distinct distributions involved when we sample repeatedly and measure a variable of interest. The **population distribution gives the values of the variable for all the individuals in the population**. In this case, the individuals are the 200 chips and the variable we're recording is color. Our parameter of interest is the proportion of red chips in the population, p = 0.50. Each random sample that we take consists of 20 chips.

There are actually *three* different distributions involved when we sample repeatedly and measure a variable of interest:

(1) The **population distribution** describes individuals

(2) The **distributions of sample data** describes individuals

(3) The **sampling distribution** describes how a statistic varies in many samples from the population

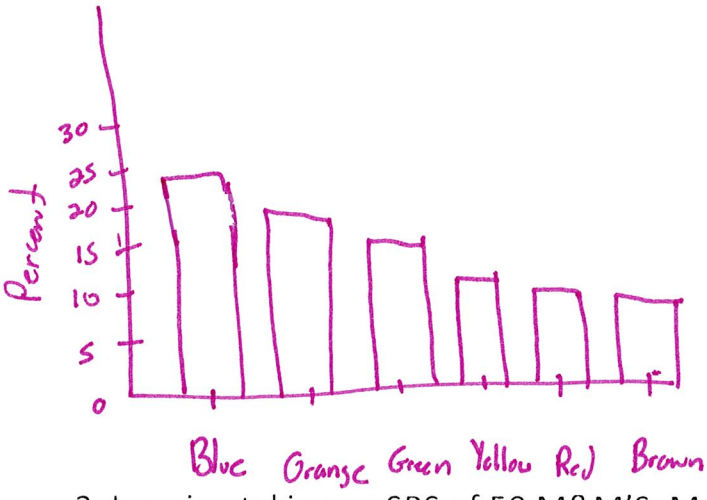
It is imperative that you can keep these three distributions straight. The population distribution and the distributions of sample data describe *individuals*. The sampling distribution describes how a statistic varies in many samples from the population.

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.

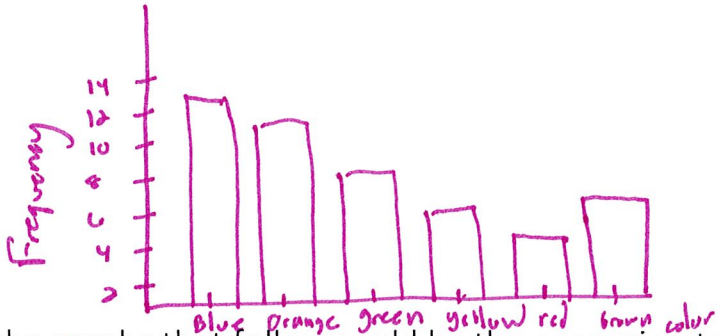
Individuals: M+M's
Variable: Color
Parameter of Interest: proportion of orange m+m's



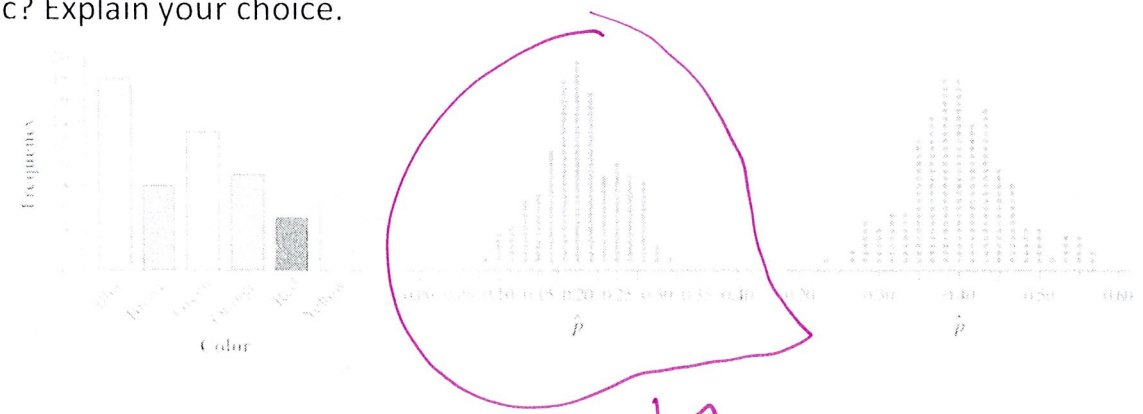
2. Imagine taking an SRS of 50 M&M'S. Make a graph showing a possible distribution of the sample data. Give the value of the appropriate statistic for this sample. Answers will vary:

For this sample, there are 11 orange m+m's so $\hat{p} = \frac{11}{50} = .22$ $\hat{p} = \frac{n}{50} = \frac{\text{\#}}{50}$

13 blue
11 orange
8 green
6 yellow
5 red
7 brown



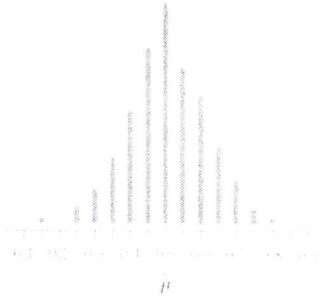
3. Which of the graphs that follow could be the approximate sampling distribution of the statistic? Explain your choice.



centered around $\hat{p} = .20$

3. Describing Sampling Distributions

a. Center: Biased and unbiased estimators



How well does the sample proportion of hyenas estimate the population proportion of hyenas? The dotplot shows the approximate sampling distribution of $\hat{p}$. We noted earlier that the center of this distribution is very close to 0.5, the parameter value. In fact, if we took all possible samples of 20 hyenas from the population, calculated $\hat{p}$ for each sample, and then found the mean of those $\hat{p}$-values, we would get exactly 0.5. For this reason, we say $\hat{p}$ is an **unbiased estimator** of p.

* mean of a sampling distribution will always equal the mean of the population for any sample size *

Unbiased Estimator - 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.

Note: unbiased does not mean perfect. An unbiased estimator will almost always provide an estimate that is not equal to the population parameter. It is called unbiased because in repeated samples, the estimates will not consistently be too high or too low. Consistent

Biased Estimator - A statistic used to estimate a parameter is a biased estimator if the mean of its sampling distribution not equal to the true value of the parameter being estimated.

b. Spread: Low variability is better! - To get a trustworthy estimate of an unknown population parameter, start by using a statistic that is an unbiased estimator. This ensures that you do not get an over or underestimate. However, this does not guarantee that the value of the statistic from your sample will be close to the actual parameter value.

The key to success is larger samples. The size of the random sample drives the variability of the sampling distribution. The variability is not a function of the size of the population.

Variability of a Statistic

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined primarily by the *size of the random sample*. Larger samples give smaller spread. The spread of the sampling distribution does not depend on the size of the population, **as long as the population is 10 times larger than the sample** (10% rule).

* *

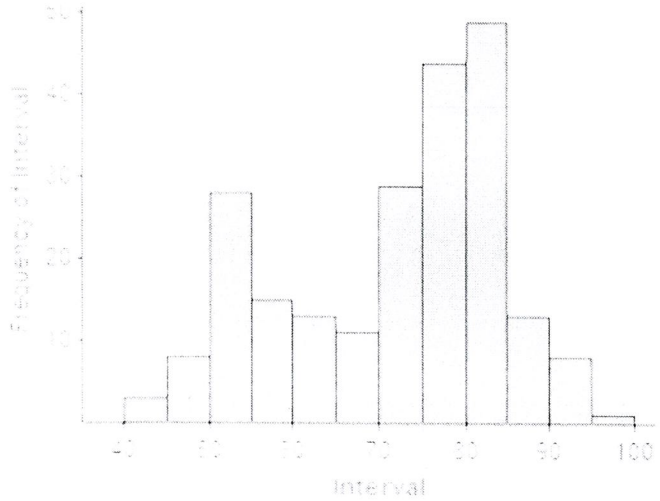
Spread affected by sample size, not population size (as long as 10% rule applies)

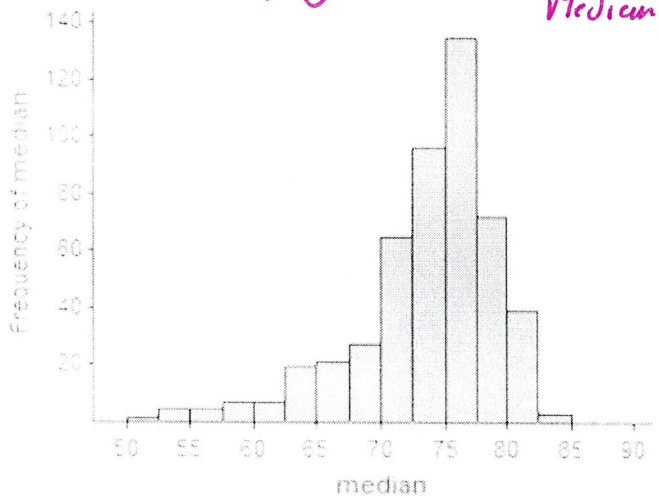
10(sample) ≤ population

Larger sample size results in smaller spread of variability (* but doesn't eliminate bias)

CHECK YOUR UNDERSTANDING

Sampling Distribution of Medians





The histogram above left shows the intervals (in minutes) between eruptions of the 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 the 500 sample median values is 73.5.

1. Is the sample median an unbiased estimator of the population median? Justify your answer.

Median does not appear to be an unbiased estimator of the population median. The mean of the 500 sample medians is 73.5, whereas the median of the population is 75.

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. Explain what this means in terms of overestimating or underestimating the population median

Skewed left, which means that, in general, underestimates of population median will be greater than overestimates

c. Bias, variability and shape - The true value of a population parameter can be thought of as the bull's eye on a target and the sample statistic as the bullet fired at the target. Both bias and variability describe what happens when we take many shots at the target.

unbias = precision

low variability = accuracy

Bias means 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 are not giving very similar results.

High Bias, Low Variability

Low Bias High variability

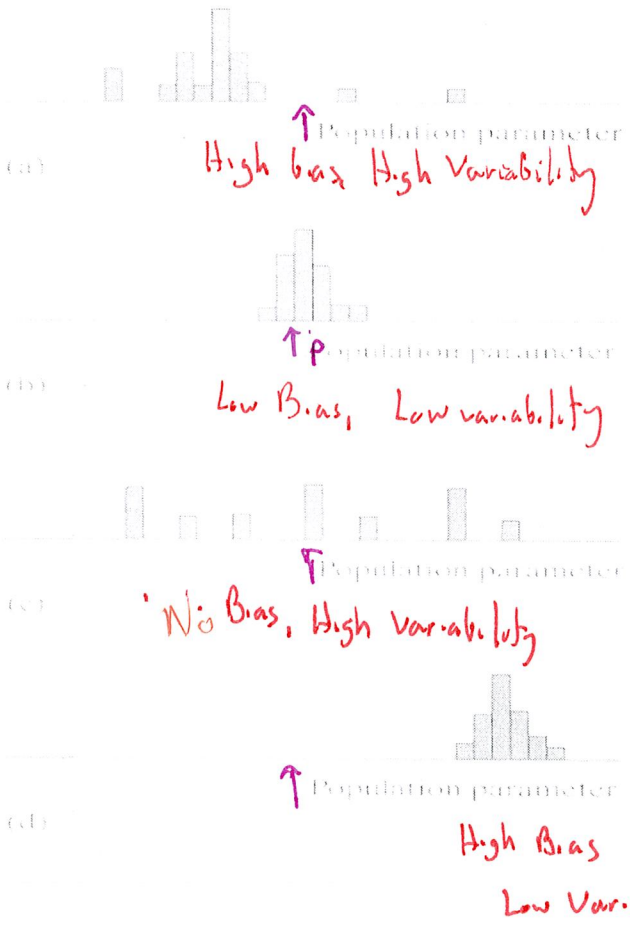
Notice that low variability can accompany high bias and low or no bias can accompany high variability.

Ideally, we would like our estimates to be *accurate* (unbiased) and *precise* (have low variability).

The Ideal No Bias, low variability

High Bias, High Variability

The bottom line is choose a statistic that has low or no bias and minimum variability.



High bias, High Variability

Low Bias, Low variability

No Bias, High Variability

High Bias Low Variability

Application - The figure to the left shows histograms of four sampling distributions of different statistics intended to estimate the same parameter.

(a) Which statistics are unbiased estimators? Why?

Graph C shows an unbiased estimator because the mean of the distribution is very close to the population parameter. (Also could say Graph B since bias is so small)

(b) Which statistic does the best job of estimating the parameter? Why?

Graph B. It has almost no bias and has little variability.