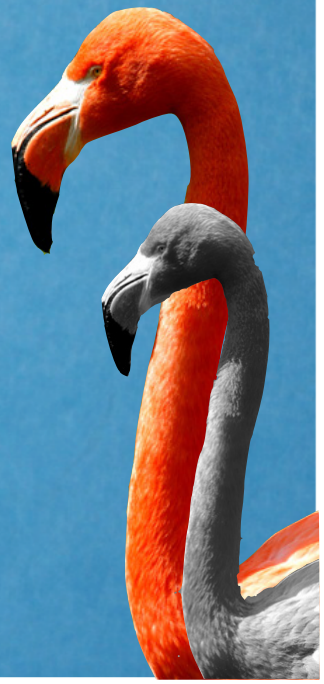
CHAPTER 7
Sampling Distributions

Data to gather: Write your height on a card and on the list on the board

7.1
What Is A Sampling Distribution?



The Practice of Statistics, 5th Edition
Starnes, Tabor, Yates, Moore

Bedford Freeman Worth Publishers

What Is A Sampling Distribution?

Learning Objectives

After this section, you should be able to:

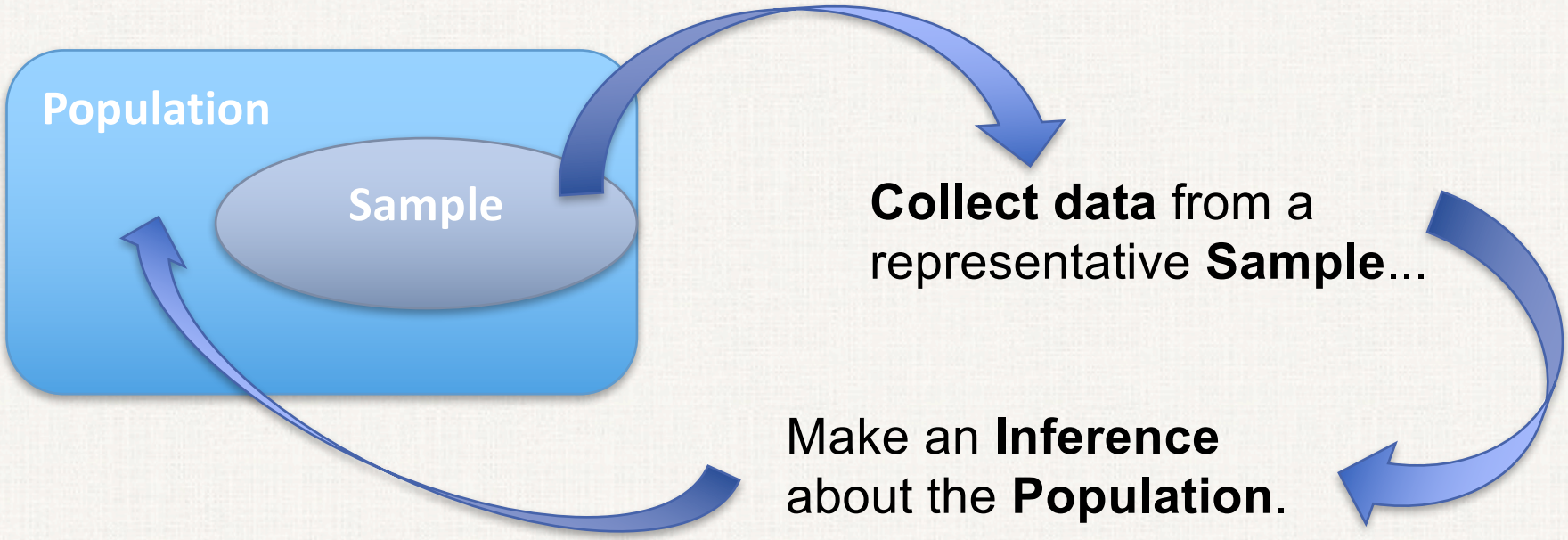
- ✓ DISTINGUISH between a parameter and a statistic.
- ✓ USE the sampling distribution of a statistic to EVALUATE a claim about a parameter.
- ✓ DISTINGUISH among the distribution of a population, the distribution of a sample, and the sampling distribution of a statistic.
- ✓ DETERMINE whether or not a statistic is an unbiased estimator of a population parameter.
- ✓ DESCRIBE the relationship between sample size and the variability of a statistic.

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.



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.

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

Remember **s** and **p**:
statistics come from **s**amples and
parameters come from **p**opulations

We write μ (the Greek letter mu) for the population mean and $\bar{x}$ ("x-bar") for the sample mean. We use p to represent a population proportion. The sample proportion $\hat{p}$ ("p-hat") is used to estimate the unknown parameter p.

Heights and cell phones

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

(a) 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 that the 75th percentile in the sample is 56 inches.

(b) A Pew Research Center Poll asked 1102 12- to 17-year-olds in the United States if they have a cell phone. Of the respondents, 71% said Yes.

Solution:

(a) The population is all 10-year-old boys;

the parameter of interest is the 75th percentile for all 10-year-old boys.

The sample is the 50 10-year-old boys included in the sample;

the statistic is 56 inches, the 75th percentile of the heights in the sample.

(b) The population is all 12- to 17-year-olds in the United States;

the parameter is *p*, the proportion of all 12- to 17-year olds with cell phones.

The sample is the 1102 12- to 17-year-olds in the sample;

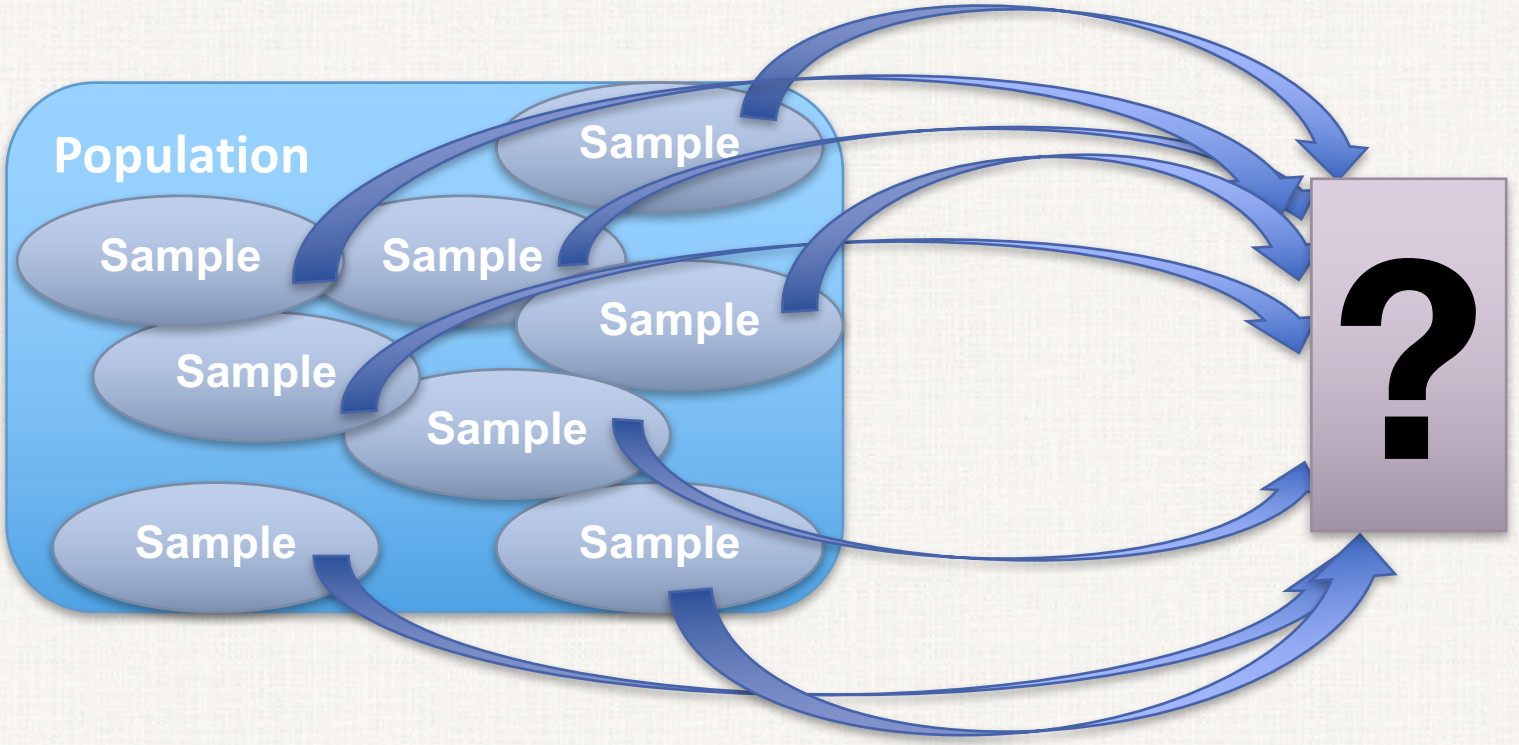
the statistic is the sample proportion with a cell phone, $\hat{p} = 71\%$.

Sampling Variability

How can $\bar{x}$ be an accurate estimate of μ? After all, different random samples would produce different values of $\bar{x}$.

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?"



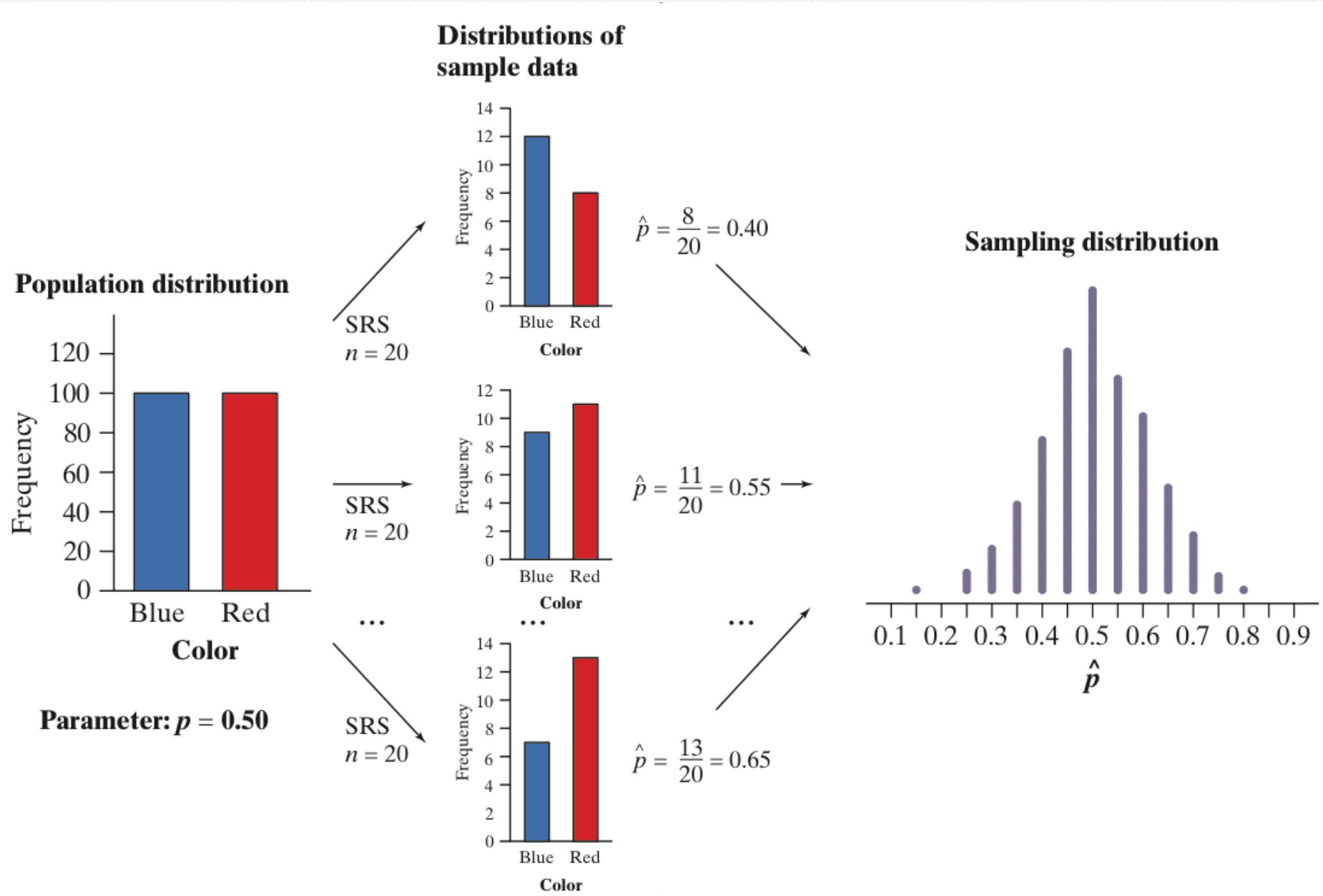
Sampling Distribution

If we took every one of the possible samples of size *n* from a population, calculated the sample proportion for each, and graphed all of those values, we'd have a **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.

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

Sampling Distribution vs. Population Distribution

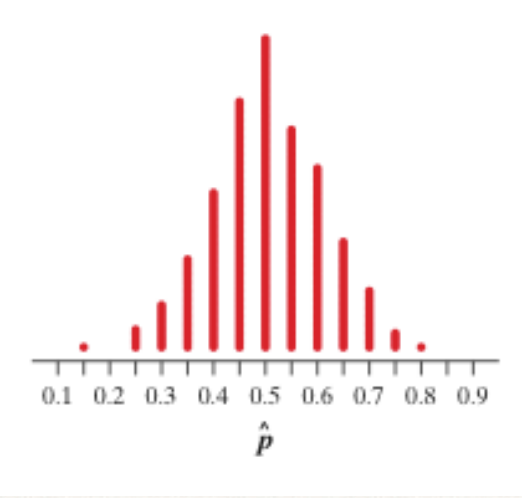


Describing Sampling Distributions

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.

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.

Activity - Sampling heights

1. Write your height on a card and place it in the box
2. Cards should be mixed thoroughly. Each student will take two samples of 4 cards (separately).
3. For your SRS of four students, calculate the sample mean $\bar{x}$ and the sample range (maximum – minimum) of the heights.

   example:

Height	Sample mean ($\bar{x}$)	Sample range (max – min)
62, 75, 68, 63	67	75 – 62 = 13

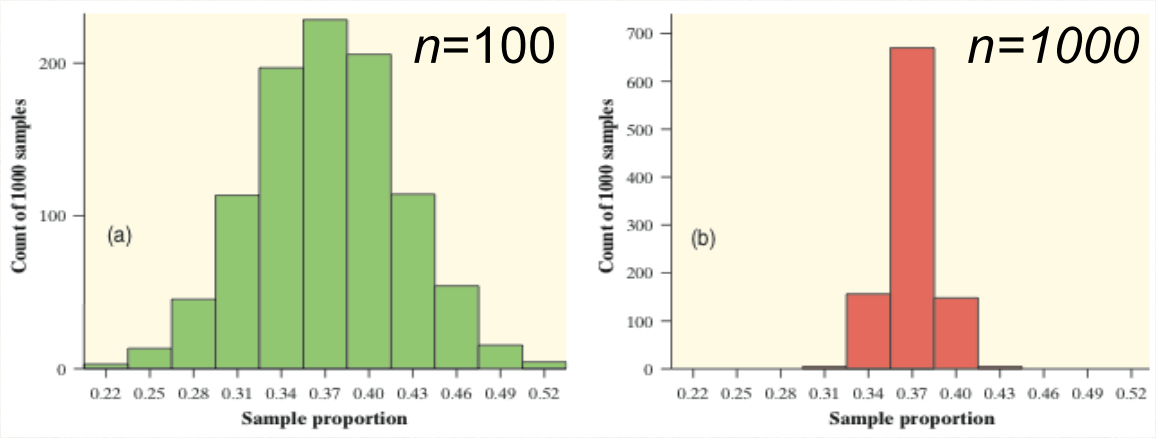
4. Plot the values of your two sample means and sample ranges on the class dotplots.
5. Using the height list on the board, find the population mean μ and the population range.
6. Based on our approximate sampling distributions of $\bar{x}$ and the sample range, which statistic appears to be an unbiased estimator? Which appears to be a biased estimator?

Describing Sampling Distributions

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 that you won't tend to overestimate or underestimate.

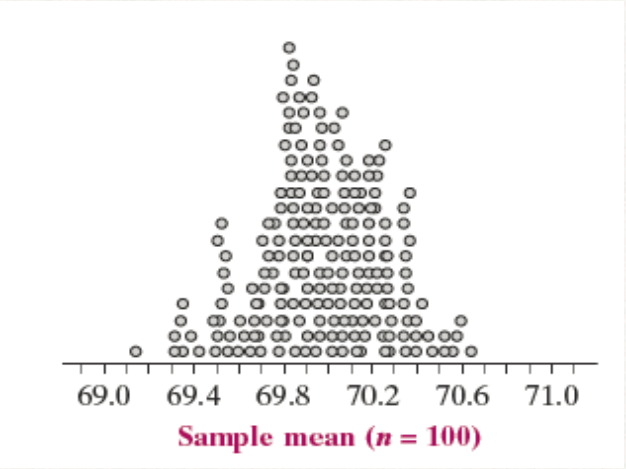
Unfortunately, using an unbiased estimator doesn't guarantee that the value of your statistic will be close to the actual parameter value.

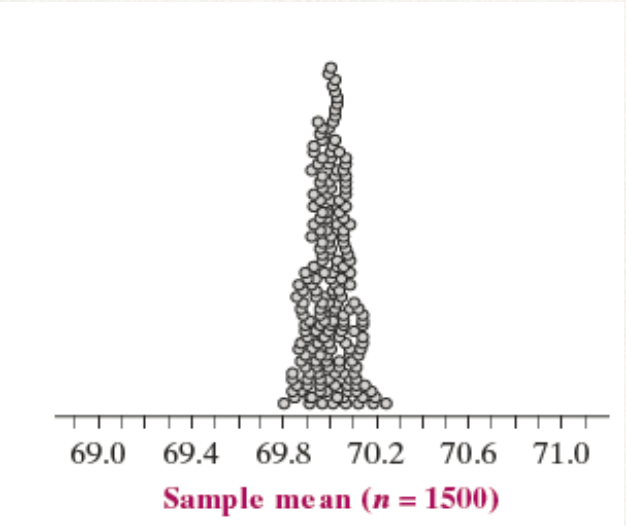


Larger samples have a clear advantage over smaller samples. They are much more likely to produce an estimate close to the true value of the parameter.

Sampling heights

Suppose that the heights of adult males are approximately Normally distributed with a mean of 70 inches and a standard deviation of 3 inches. To see why sample size matters, we took 200 SRSs of size 100 and calculated the sample mean height and then took 200 SRSs of size 1500 and calculated the sample mean height. Here are the results, graphed on the same scale for easy comparisons.





As you can see, the spread of the approximate sampling distributions is very different. When the sample size was larger, the distribution of the sample mean was much less variable. In other words, when the sample size is larger, the sample mean will typically be closer to the true mean.

Describing Sampling Distributions

There are general rules for describing how the spread of the sampling distribution of a statistic decreases as the sample size increases. One important and surprising fact is that the variability of a statistic in repeated sampling does not depend very much on 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 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.

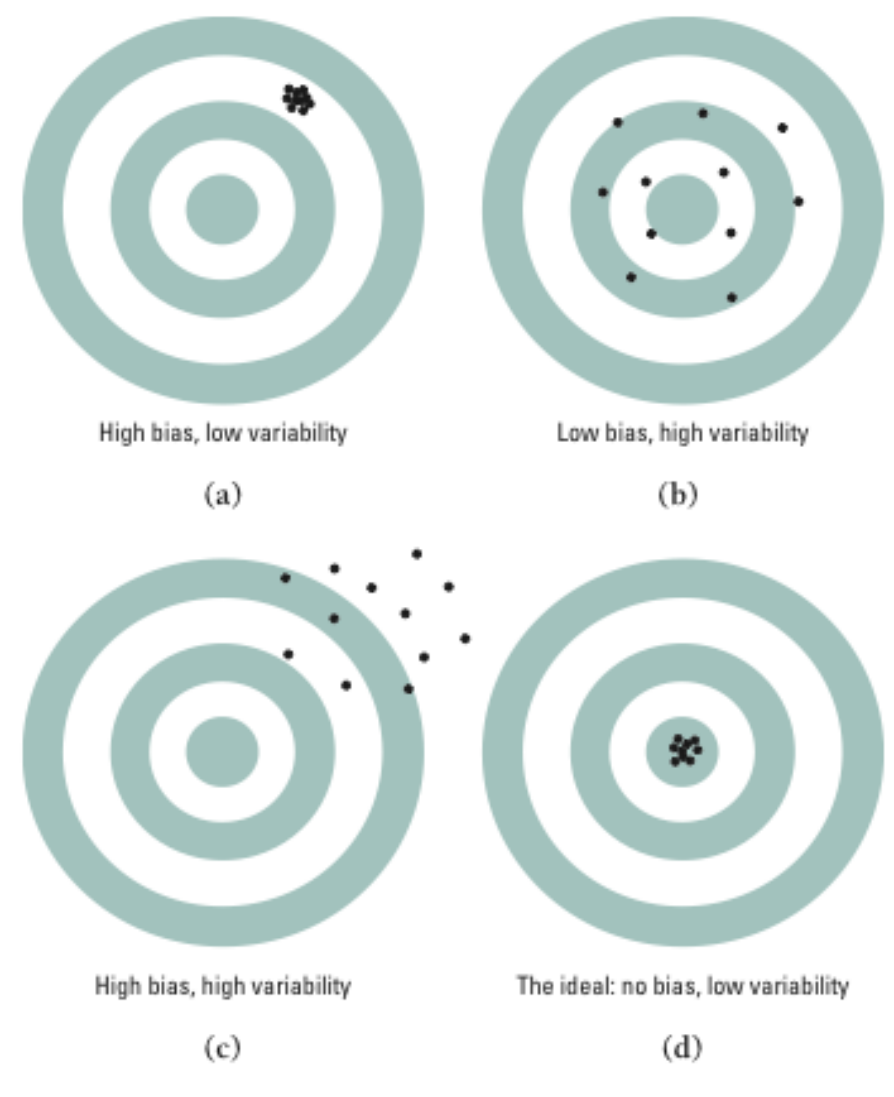
Bias, Variability, and Shape

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.



More tanks

Here are four additional methods for estimating the total number of tanks. The partition method that was chosen by the mathematicians is also included for comparison.

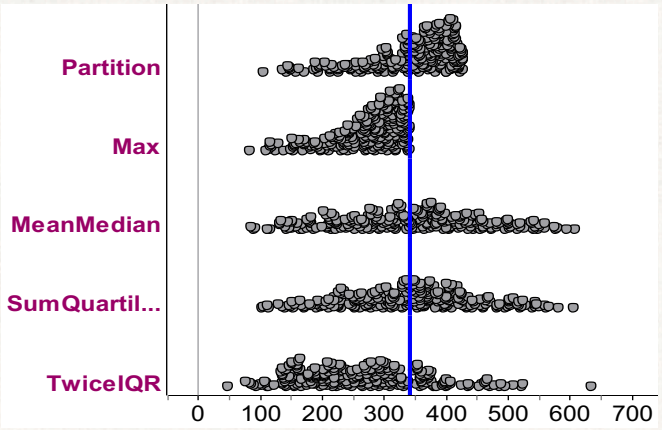
(1) Partition = (5/4)maximum

(2) Max = maximum

(3) MeanMedian = mean + median

(4) SumQuartiles = $Q_1 + Q_3$

(5) TwiceIQR = $2IQR$



The graph shows the approximate sampling distribution for each of these statistics when taking 250 samples of size 4 from a population of 342 tanks.

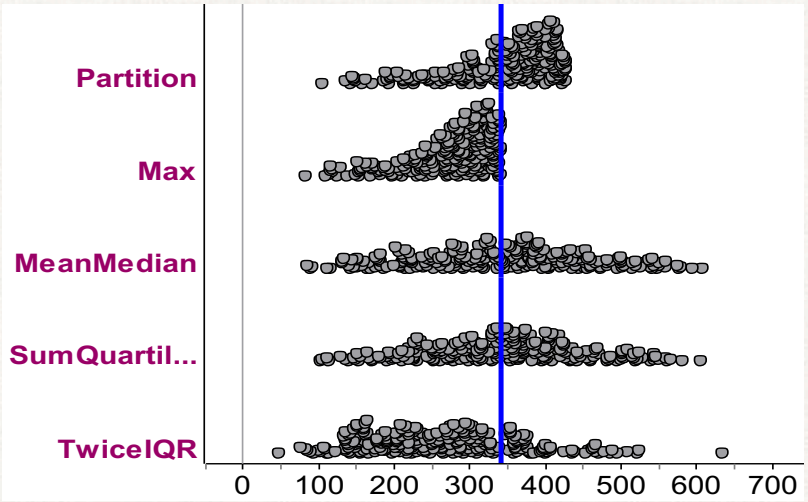
(a) Which of these statistics appear to be biased estimators? Explain.

(b) Of the unbiased estimators, which is best? Explain.

(c) Explain why a biased estimator might be preferred over an unbiased estimator.

More tanks solution

(a) The statistics Max and TwiceIQR appear to be biased estimators because they are consistently too low. That is, the centers of their sampling distributions appear to be below the correct value of 342.



(b) Of the three unbiased statistics, Partition is best since it has the least variability.

(c) *Explain why a biased estimator might be preferred over an unbiased estimator.*

Even though Max is a biased estimator, it often produces estimates very close to the truth. MeanMedian, although unbiased, is quite variable and not close to the true value as often. For example, in 120 of the 250 SRSs, Max produced an estimate within 50 of the true value. However, MeanMedian was this close in only 79 of the 250 SRSs.

What Is A Sampling Distribution?

Section Summary

In this section, we learned how to…

- ✓ DISTINGUISH between a parameter and a statistic.
- ✓ USE the sampling distribution of a statistic to EVALUATE a claim about a parameter.
- ✓ DISTINGUISH among the distribution of a population, the distribution of a sample, and the sampling distribution of a statistic.
- ✓ DETERMINE whether or not a statistic is an unbiased estimator of a population parameter.
- ✓ DESCRIBE the relationship between sample size and the variability of a statistic.
- ✓ Read p. 422-435 ccc 1, 3, 5, 7, 9, 11, 13, 15, 17, 19