Section 7.1
What Is a Sampling Distribution?

Learning Objectives

After this section, you should be able to…

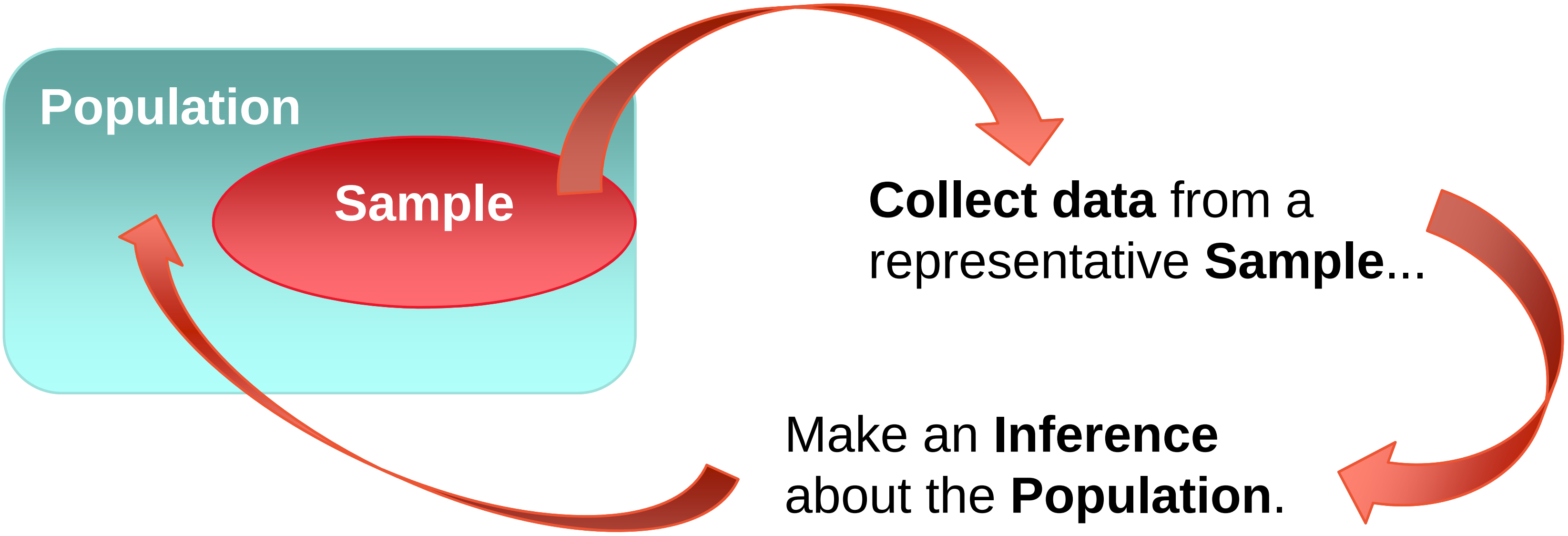
- ✓ DISTINGUISH between a parameter and a statistic
- ✓ DEFINE sampling distribution
- ✓ DISTINGUISH between population distribution, sampling distribution, and the distribution of sample data
- ✓ DETERMINE whether a statistic is an unbiased estimator of a population parameter
- ✓ DESCRIBE the relationship between sample size and the variability of an estimator

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.



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. In statistical practice, the value of a parameter is usually not known because we cannot examine the entire population.

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.

Remember **s** and **p**: **s**tatistics come from **s**amples and **p**arameters 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.

Parameters and Statistics - Try These

Identify the population, the parameter, the sample, and the statistic:

1) The Gallup Poll asked a random sample of 515 US adults whether or not they believed in ghosts. Of the respondents, 160 said "Yes."
   - Population
   - Parameter
   - Sample
   - Statistic

2) How much do gas prices vary in a large city? To find out, a reporter records the price per gallon at a random sample of 10 gas stations in the city on the same day. The range (max-min) of the sample is 25 cents.
   - Population
   - Parameter
   - Sample
   - Statistic

3) A random sample of 100 female college students has a mean of 64.5 inches; which is greater than the 63 inch mean height of all adult American women.
   - Population
   - Parameter
   - Sample
   - Statistic

Answers:

Parameters and Statistics - Try These

Identify the population, the parameter, the sample, and the statistic:

1) The Gallup Poll asked a random sample of 515 US adults whether or not they believed in ghosts. Of the respondents, 160 said “Yes.”

- **Population The population of all U.S. adults**
- **Parameter the parameter of interest is “p” - which is the proportion of all U.S. adults that believe in ghosts**
- **Sample random sample of 515 US adults**
- **Statistic phat=160/515 = .31**

$$\boxed{\hat{p} = 160/515 = .31}$$

2) How much do gas prices vary in a large city? To find out, a reporter records the price per gallon at a random sample of 10 gas stations in the city on the same day. The range (max-min) of the sample is 25 cents.

- **Population all gas stations in a large city**
- **Parameter range of gas prices at all the gas stations in the city**
- **Sample a random sample of 10 gas stations in the city**
- **Statistic sample range is 25 cents**

3) A random sample of 100 female college students has a mean of 64.5 inches; which is greater than the 63 inch mean height of all adult American women.

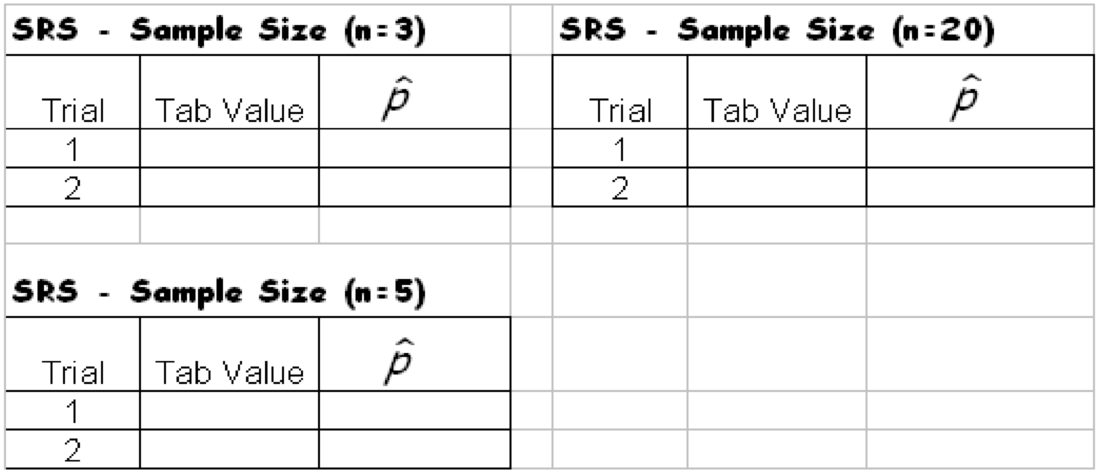
- **Population all adult American women**
- **Parameter μ=63 inch**
- **Sample random sample of 100 female college students**
- **Statistic Xbar=64.5 inches**

$$\boxed{\overline{X} = 64.5 inches}$$

Activity: Bean Counters

- **Each person should randomly select 2 samples of size 3, 2 samples of size 5, AND 2 samples of size 20**

- **PLEASE** be sure to follow the guidelines below so that your data is accurate!
- Mix the **beans** well.
- Without looking, randomly select one **bean** at a time, **bean** by **bean,** until you have a sample selected.
- Record the number of **black beans** for each sample
- Calculate the proportion ($\hat{p}$) of **black beans** for each sample
- Replace the **beans** BEFORE someone else selects a sample! Be sure to mix before someone else selects another sample.
- Each individual should select their own samples!



SRS - Sample Size (n=3)

Trial	Tab Value	$\hat{p}$
1		
2		

SRS - Sample Size (n=5)

Trial	Tab Value	$\hat{p}$
1		
2		

SRS - Sample Size (n=20)

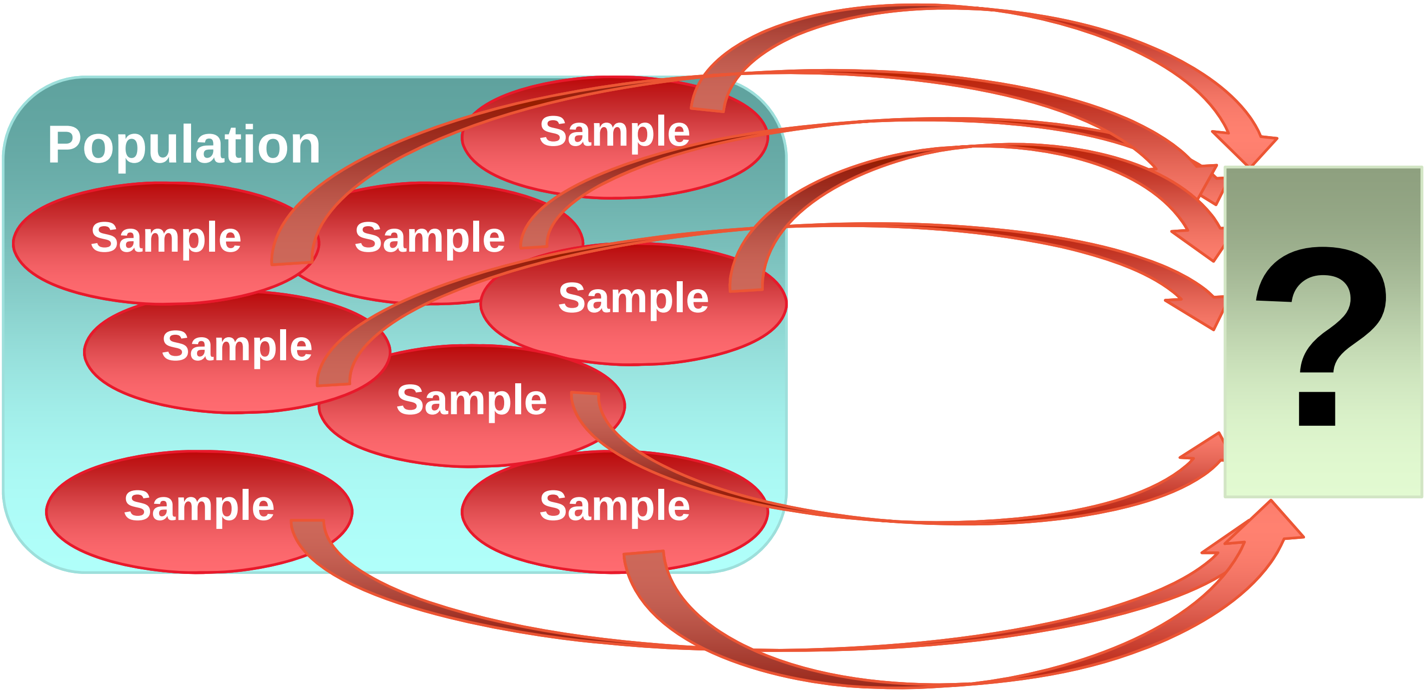
Trial	Tab Value	$\hat{p}$
1		
2		

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

In the previous activity, we took a handful of different samples of 3, 5, and 20 beans. There are many, many possible SRSs of size 3,5, 20 from a population of size 200. **How can we calculate this number?**

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

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.

Do NOT confuse with the distribution of a sample!

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.

One of the uses of probability theory in statistics is to obtain sampling distributions without simulation. We'll get to the theory later.

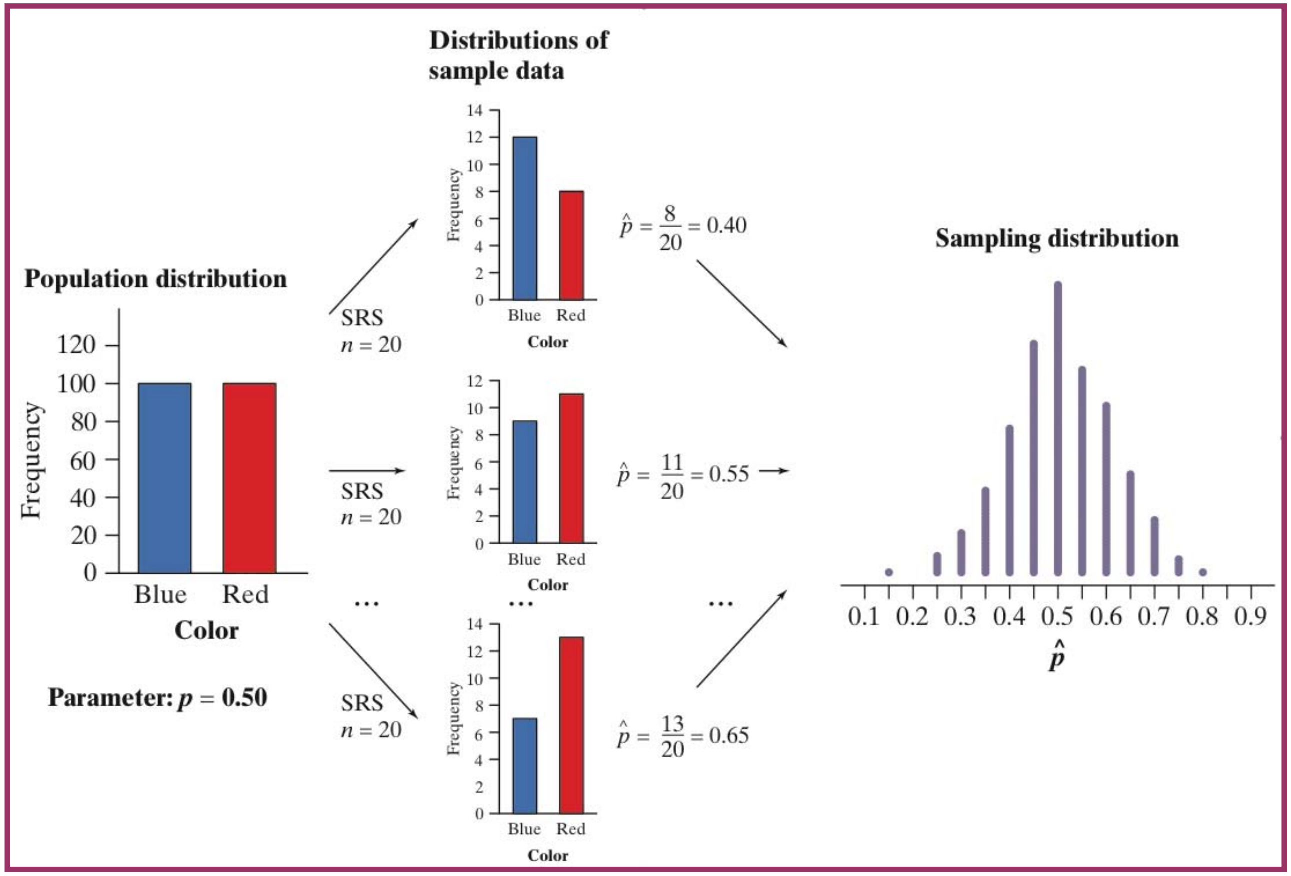
Population Distributions vs. Sampling Distributions

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.



More on Sampling Distribution…

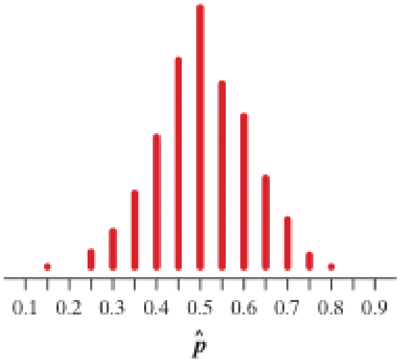
Learning Objectives

- ✓ DETERMINE whether a statistic is an unbiased estimator of a population parameter
- ✓ DESCRIBE the relationship between sample size and the variability of an estimator

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**.



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.

Describing Sampling Distributions

Center: Biased and unbiased estimators

Definition:

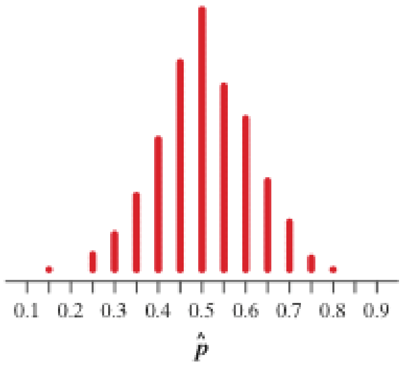
A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

In the beans example:

❑ **We collected many samples** and calculated the **sample proportion of black beans**.

❑ How well does the **sample proportion estimate (phat)** the **true proportion** of black beans, ***p* = 0.5**?

❑Therefore we can say the **sample proportion estimate** is an **unbiased estimator for the population.**



Describing Sampling Distributions

Next Spread: Low variability is better!

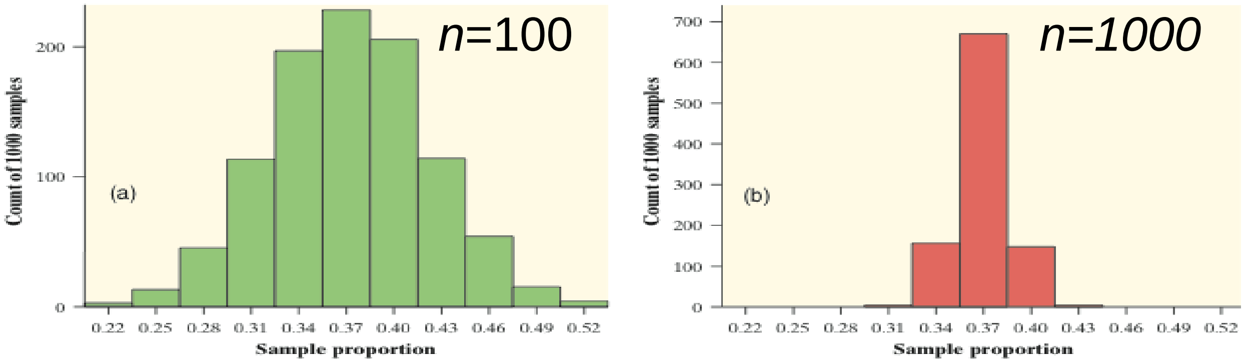
- To get a trustworthy estimate of an unknown population parameter, start by using a statistic that's an unbiased estimator.

- 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. So next we look at variability.

Spread: Low variability is better!

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.



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 at least 10 times larger than the sample.**

Mathematical Theory Versus the Real World - The “10 percent rule”

- **The “10 percent rule”** is a numerical approximation that we may apply in AP Stats.
- To see the underlying mathematical theory, check out this link: http://apcentral.collegeboard.com/apc/members/courses/teachers_corner/39161.html

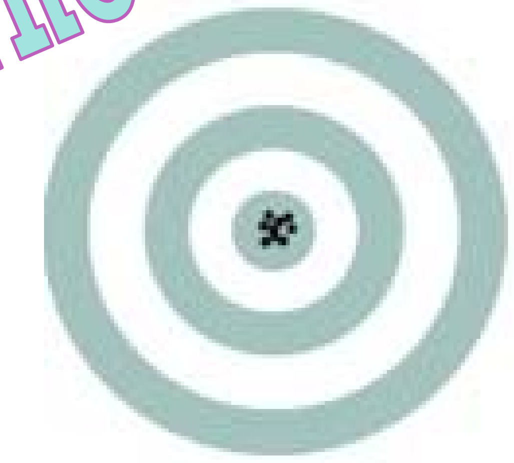
Describing Sampling Distributions

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.

The Ideal



The ideal: no bias, low variability

(d)

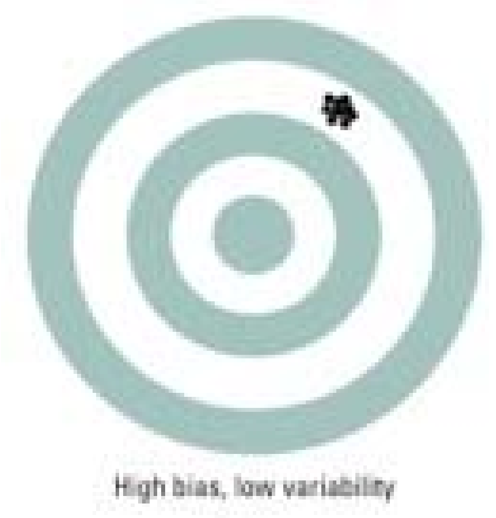
Describing Sampling Distributions

Bias, variability, and shape

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.

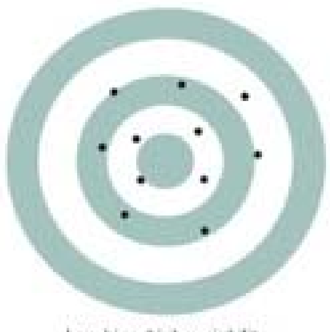


(a)

Describing Sampling Distributions

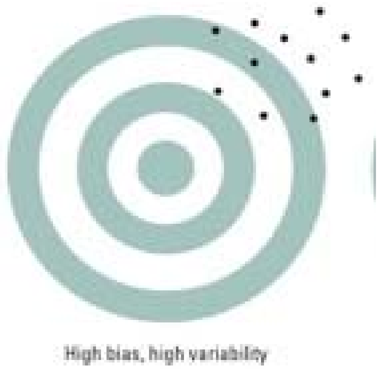
Bias, variability, and shape

Both bias and **variability** describe what happens when we take many shots at the target.



(b)

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



(c)

The lesson about center and spread is clear: **given a choice of statistics to estimate an unknown parameter, choose one with**

1. **No or low bias and**
2. **Minimum variability**.

EXAMPLE

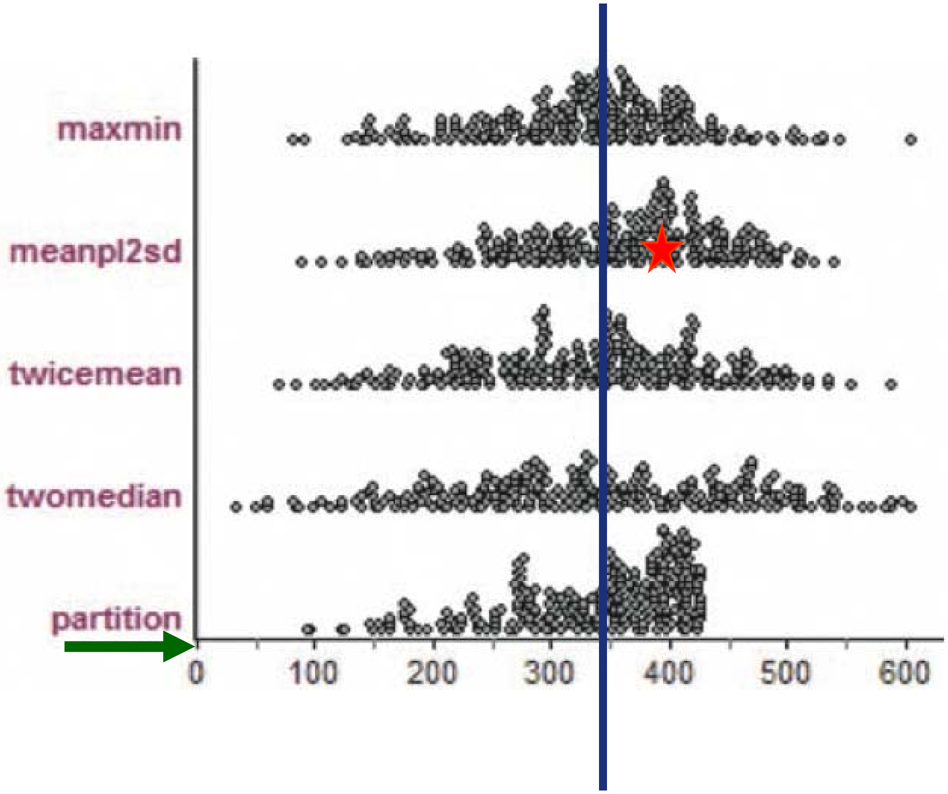
Sampling Distributions - Bias, variability, and shape

- Sampling distributions can take on many shapes.
- The same statistic can have sampling distributions with different shapes depending on the population distribution and the sample size.
- Be sure to consider the shape of the sampling distribution before doing inference.

EXAMPLE:

- Sampling distributions for different statistics used to estimate the number of German Tanks in WW II.
- The blue line represents the true number of tanks.
- The 1st four were presented by groups of students.
- The "Partition Method" was recommended by D.C. mathematicians.

1. Bias or unbias estimators?
2. Describe the different sampling distributions.
3. Which statistic gives the best estimator? Why?



ANSWERS

EXAMPLE: German tanks

Sampling Distributions - Bias, variability, and shape

Which estimators are Bias or Unbias?

- Meanpl2sd is a BIASED estimator. The center of the distribution is too high. This statistic produces consistent overestimates of the number of tanks.
- The other 4 statistics are unbias estimators.

Sampling Distribution – Describe Center, Spread (variability), Shape

- Student sampling distributions Maxmin, Twicemean and Twomedian are roughly symmetric shapes so these statistics are about equally likely to underestimate or overestimate the number of tanks.

- Among the 3 Student sampling distributions: Maxmin has the smallest variability and in general produce estimates that are closer to the actual number of tanks. Maxmin would be the best estimator from the students.

- Partition was developed by Washington DC mathematicians.
  - It is left skewed which means the statistic is more likely to overestimate than underestimate the number of tanks.
  - The math guys felt it would be better to err on the side of caution and give the military commanders an estimate that's slightly too high.

Summary

In this section, we learned that…

- ✓ A **parameter** is a number that describes a population. To estimate an unknown parameter, use a **statistic** calculated from a sample.
- ✓ The **population distribution** of a variable describes the values of the variable for all individuals in a population. The **sampling distribution** of a statistic describes the values of the statistic in all possible samples of the same size from the same population.
- ✓ A statistic can be an **unbiased estimator** or a **biased estimator** of a parameter. Bias means that the center (mean) of the sampling distribution is not equal to the true value of the parameter.
- ✓ The **variability** of a statistic is described by the spread of its sampling distribution. Larger samples give smaller spread.
- ✓ When trying to estimate a parameter, choose a statistic with low or no bias and minimum variability. Don’t forget to consider the shape of the sampling distribution before doing inference.