

# Sampling Distribution

CHAPTER 18

# Modeling the Distribution of Sample Proportions

- ▶ Rather than showing real repeated samples, *imagine* what would happen if we were to actually draw many samples.
- ▶ Now imagine what would happen if we looked at the sample proportions for these samples. What would the histogram of all the sample proportions look like?

## Modeling the Distribution of Sample Proportions (cont.)

- ▶ We would expect the histogram of the sample proportions to center at the true proportion,  $p$ , in the population.
- ▶ As far as the shape of the histogram goes, we can simulate a bunch of random samples that we didn't really draw.

## Modeling the Distribution of Sample Proportions (cont.)

- ▶ It turns out that the histogram is unimodal, symmetric, and centered at  $p$ .
- ▶ More specifically, it's an amazing and fortunate fact that a Normal model is just the right one for the histogram of sample proportions.
- ▶ To use a Normal model, we need to specify its mean and standard deviation. The mean of this particular Normal is at  $p$ .

# Modeling the Distribution of Sample Proportions (cont.)

- ▶ When working with proportions, knowing the mean automatically gives us the standard deviation as well—the standard deviation we will use is

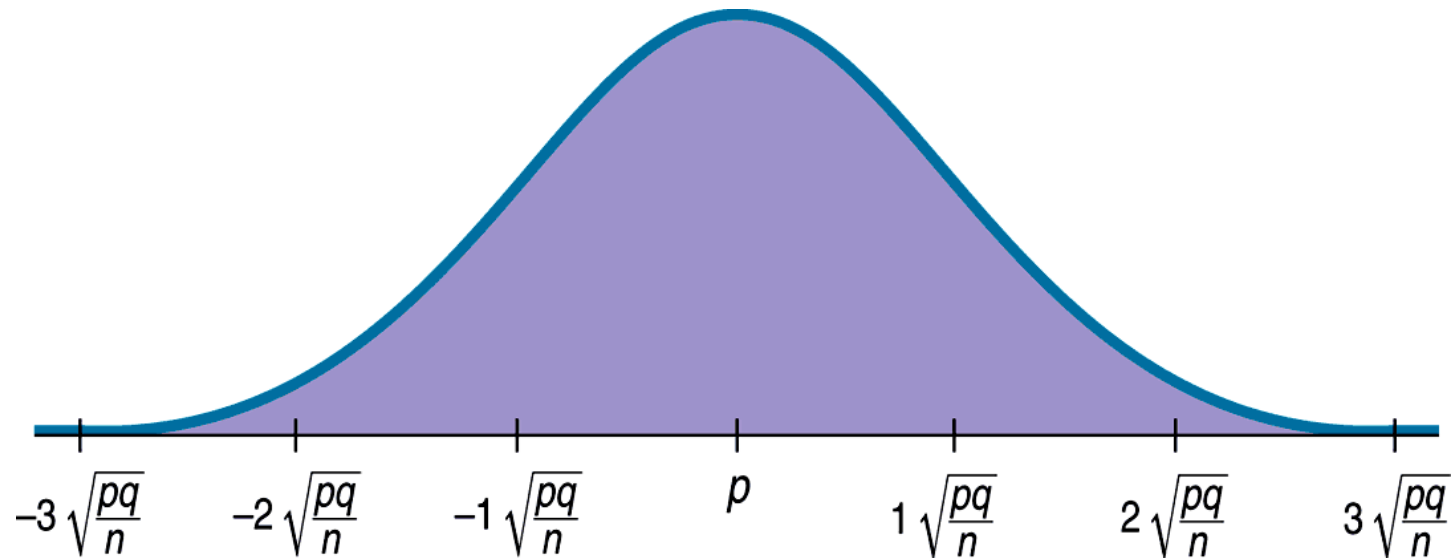
$$\sqrt{\frac{pq}{n}}$$

- ▶ So, the distribution of the sample proportions is modeled with a probability model that is

$$N\left(p, \sqrt{\frac{pq}{n}}\right)$$

# Modeling the Distribution of Sample Proportions (cont.)

- ▶ A picture of what we just discussed is as follows:



# How Good Is the Normal Model?

- ▶ The Normal model gets better as a good model for the distribution of sample proportions as the sample size gets bigger.
- ▶ Just how big of a sample do we need? This will soon be revealed...

# Assumptions and Conditions

- ▶ Most models are useful only when specific assumptions are true.
- ▶ There are two assumptions in the case of the model for the distribution of sample proportions:
  1. The sampled values must be independent of each other.
  2. The sample size,  $n$ , must be large enough.



## Assumptions and Conditions (cont.)

- ▶ Assumptions are hard—often impossible—to check. That's why we *assume* them.
- ▶ Still, we need to check whether the assumptions are reasonable by checking *conditions* that provide information about the assumptions.
- ▶ The corresponding conditions to check before using the Normal to model the distribution of sample proportions are the 10% Condition and the Success/Failure Condition.

## Assumptions and Conditions (cont.)

1. 10% condition: If sampling has not been made with replacement, then the sample size,  $n$ , must be no larger than 10% of the population.
2. Success/failure condition: The sample size has to be big enough so that both  $np$  and  $nq$  are greater than 10.

So, we need a large enough sample that is not too large.

# Example:

A candy company claims that 25% of the jelly beans in its spring mix are pink. Suppose that the candies are packaged at random in small bags containing about 300 jelly beans. A class of students opens several bags, counts the various colors of jelly beans, and calculates the proportion that are pink in each bag. Is it appropriate to use a Normal model to describe the distribution of the proportion of pink jelly beans?

Yes – Check the Conditions and Assumptions

Less than 10% - yes

10 success and 10 failures – Yes

Independent – Yes

# A Sampling Distribution Model for a Proportion

- ▶ A proportion is no longer just a computation from a set of data.
  - ▶ It is now a random quantity that has a distribution.
  - ▶ This distribution is called the **sampling distribution model** for proportions.
- ▶ Even though we depend on sampling distribution models, we never actually get to see them.
  - ▶ We never actually take repeated samples from the same population and make a histogram. We only imagine or simulate them.

## A Sampling Distribution Model for a Proportion (cont.)

- ▶ Still, sampling distribution models are important because
  - ▶ they act as a bridge from the real world of data to the imaginary world of the statistic and
  - ▶ enable us to say something about the population when all we have is data from the real world.

# The Sampling Distribution Model for a Proportion (cont.)

- ▶ Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of  $p$  is modeled by a Normal model with

- ▶ Mean:  $\mu(p) = p$

- ▶ Standard deviation:  $SD(p) = \sqrt{\frac{pq}{n}}$

# Example

- ▶ Assume that 25% of students at a university wear contact lenses. We randomly pick 200 students. What is the probability that more than 28% of this sample wear contact lenses?
- ▶ Mean .25
- ▶  $SD = \sqrt{\frac{.75(.25)}{200}} = .0306$
- ▶  $(.28 - .25)/.0306 = .9804$
- ▶ 16.34%

# What about Quantitative Data??

- ▶ Proportions summarize **categorical data**.
- ▶ Means summarize **quantitative data**.



# Categorical or Quantitative

Example: Of all the cars on the interstate, 80% exceed the speed limit. What is the probability that a random sample of 50 cars traveling on the interstate contains less than 32 exceeding the speed limit?

Are the data categorical or quantitative?

So are we dealing with means or proportions?

# Quantitative Data

- ▶ What if we're dealing with means rather than proportions?
- ▶ How does the sampling distribution of  $\bar{x}$  behave?

## Central Limit theorem

(Your text calls it the fundamental theorem of statistics)

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The sampling distribution, or distribution of  $\bar{x}$ , is approximately normal if the sample is large enough or if the population is known to be approximately normal.

Rule of thumb:  $n \geq 30$

~~This is WRONG!!!!~~ But how large is large enough? Even if the population is skewed, the sampling distribution is approximately normal if the sample is large enough.

# Central Limit Theorem

When we draw a simple random sample of size  $n$  from **ANY** shaped population with mean and finite standard deviation, that when the sample is large, the sampling distribution of the **sample mean** is approximately normal. This is true even if the original population is not normal!

# Check for Understanding

Jasmyn and Matthew are studying together for their upcoming statistics test. Matthew is attempting to explain the Central Limit Theorem to Jasmyn and says, “When you take larger and larger samples from a population, the dot plot of the sample values looks more and more normal.” Did Matthew give a good explanation of the Central Limit Theorem? Why or why not?

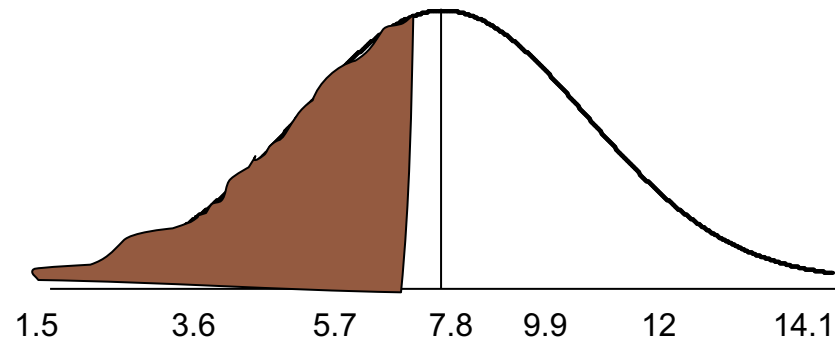
# CLT summary

- ▶ 1. The mean of the population (what we want to find) will be the same as the mean of all your many samples.
- ▶ 2. The Standard Deviation of all your many samples will be the population standard deviation divided by  $\sqrt{n}$  (your sample size)
- ▶ 3. The histogram of the samples will appear normal (bell shaped).
- ▶ 4. The larger the sample size ( $n$ ), the smaller the standard deviation will be and the more constricted the graph will be.

Example: At birth, babies average 7.8 pounds, with a standard deviation of 2.1 pounds. A random sample of 34 babies born to mothers living near a large factory that may be polluting the air and water show a mean birth weight of only 7.2 pounds.

a) What is the probability of a mother having a baby weighing 7.2 pounds or less if the distribution of baby weights is approximately normal?

$$z = \frac{7.2 - 7.8}{2.1} = -.286$$



$$P(x \leq 7.2) = .388$$

Example: At birth, babies average 7.8 pounds, with a standard deviation of 2.1 pounds. A random sample of 34 babies born to mothers living near a large factory that may be polluting the air and water show a mean birth weight of only 7.2 pounds.

b) The distribution of the weights of babies is actually skewed left. Are the weights of the 34 babies born to the mothers near the factory unusually low? Should the community be concerned?



Example: At birth, babies average 7.8 pounds, with a standard deviation of 2.1 pounds. A random sample of 34 babies born to mothers living near a large factory that may be polluting the air and water show a mean birth weight of only 7.2 pounds.

b) The distribution of the weights of babies is actually skewed left. Are the weights of the 34 babies born to the mothers near the factory unusually low? Should the community be concerned?

Check the conditions:

Random sample of babies given.

There's potentially more than  $10(34)$  or 340 babies that could be born in the community....over time.

Large enough sample?  $n=34 > 30$

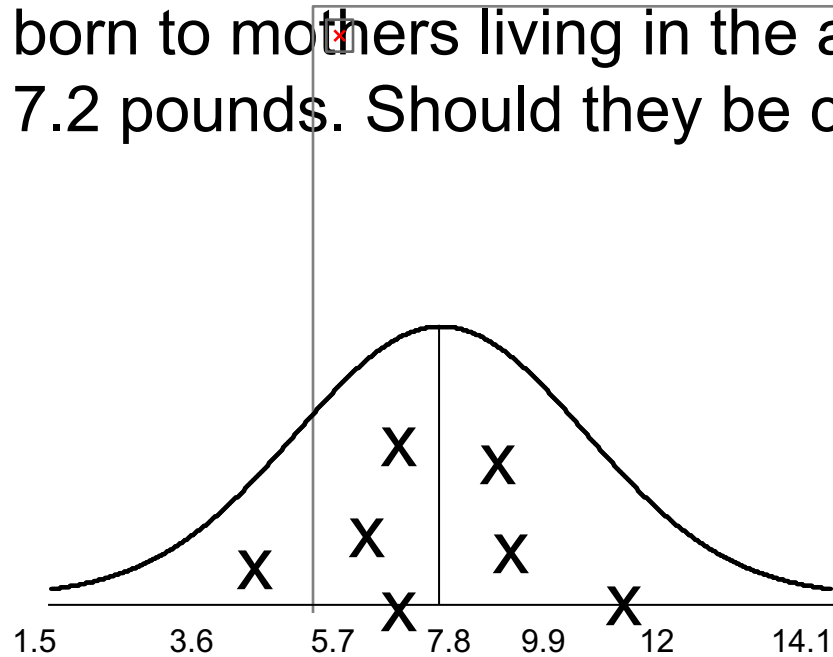
Example: At birth, babies average 7.8 pounds, with a standard deviation of 2.1 pounds. A random sample of 34 babies born to mothers living near a large factory that may be polluting the air and water show a mean birth weight of only 7.2 pounds.

b) The distribution of the weights of babies is actually skewed left. Are the weights of the 34 babies born to the mothers near the factory unusually low? Should the community be concerned?

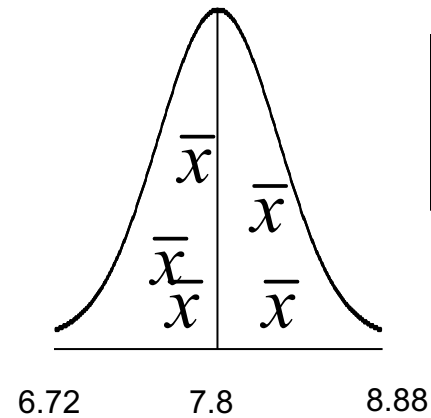
Find the mean and standard deviation of the sampling distribution.

$$\mu_{\bar{x}} = 7.8 \quad \sigma_{\bar{x}} = \frac{2.1}{\sqrt{34}} = .36$$

Weights of newborn humans have a mean of 7.8 pounds with a standard deviation of 2.1 pounds. Suppose a group of citizens living near a chemical factory find a random sample of 34 babies born to mothers living in the area have a mean birth weight of only 7.2 pounds. Should they be concerned?



distribution of baby weights



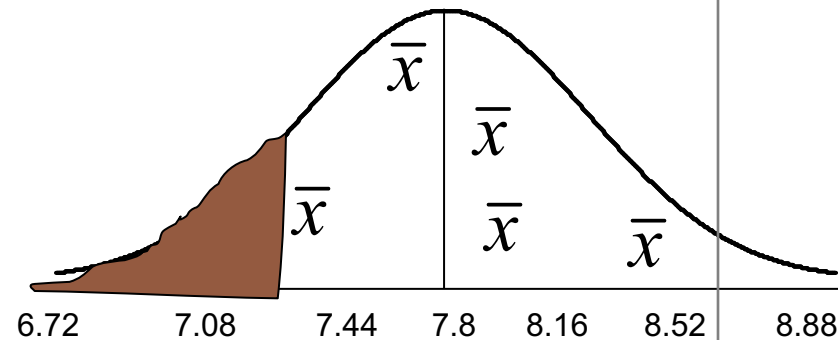
sampling distribution of baby weights

Note the scale

Weights of newborn humans have a mean of 7.8 pounds with a standard deviation of 2.1 pounds. Suppose a group of citizens living near a chemical factory find a random sample of 34 babies born to mothers living in the area have a mean birth weight of only 7.2 pounds. Should they be concerned?

$$\mu = 7.8 \quad \sigma = \frac{2.1}{\sqrt{34}} = .36$$

$$z = \frac{7.2 - 7.8}{\frac{2.1}{\sqrt{34}}} = -1.67$$



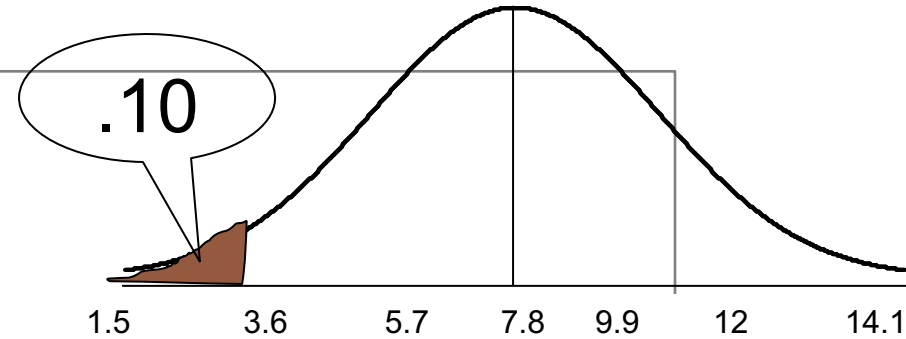
$$P(\bar{x} \leq 7.2) = .0478$$

What is the cut off for the bottom 10% of babies born?

First find  $z$  for the bottom 10%.

$$-1.28 = \frac{x - 7.8}{2.1}$$

$$x = 5.109 \text{ pounds}$$



In the library on a university campus, there is a sign in the elevator that indicates a limit of 16 persons.

Furthermore, there is a weight limit of 2500 pounds.

Assume the average weight of students, faculty, and staff on campus is 150 pounds, the standard deviation is 27 pounds, and that the distribution of weights of individuals on campus is approximately normal. If a random sample of 16 persons from the campus is taken:

a) Describe the sampling distribution.

$$\mu_{\bar{x}} = 150 \quad \sigma_{\bar{x}} = \frac{27}{\sqrt{16}} = 6.75$$

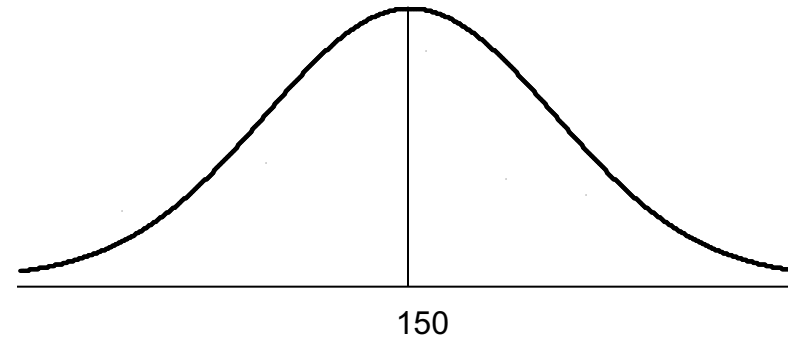
Shape is approximately normal since the population is approximately normal

b) What is the chance that a random sample of 16 persons on the elevator will exceed the weight limit?

$$\mu_{\bar{x}} = 150 \quad \sigma_{\bar{x}} = \frac{27}{\sqrt{16}} = 6.75$$

Check out the conditions:  $n$  is NOT  $> 30$   
But we know that the population is approximately normal so the sampling distribution is approximately normal. Phew!

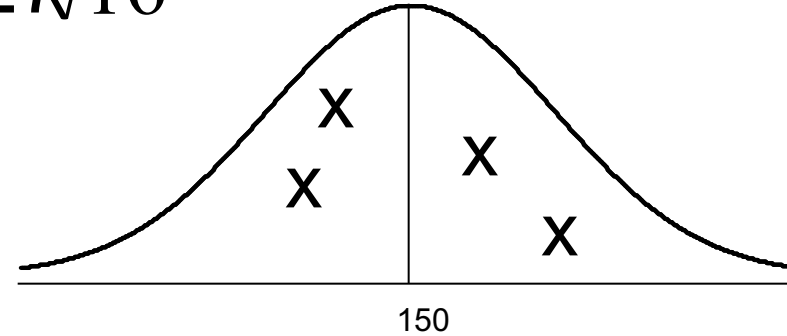
$$P(\bar{x} > 156.25) = .177$$



What is the probability of randomly selecting a person on campus that weighs more than 156 pounds?

$$\begin{aligned} \mu_{\bar{x}} &= 150 & \sigma_{\bar{x}} &= \frac{27}{\sqrt{16}} = 6.75 \\ \mu &= 150 & \sigma &= 27 \end{aligned}$$

$$P(x > 156.25) = .412$$





The Harvard College Alcohol Study finds that 67% of college students sampled nationally (n=15000) support the efforts to “crack down on underage drinking.” Upon reading the study, a local college takes a random sample of 100 students and finds that 62% of it’s students support a crackdown on underage drinking. The campus newspaper then writes an article claiming that “support at our campus is lower at our school than nationally.” Does the survey support that claim?

$$\mu_p = p = .67 \quad \sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.67)(.33)}{100}} = .0470$$

$$z = \frac{.62 - .67}{.0470} = -1.06$$

$$p(p \leq .62) = .1446 \text{ (chart) or } .1438 \text{ (calc)}$$

# Parameter

- ▶ A **Parameter** is a number that describes the population.
  - ▶ A parameter always exists but in practice we rarely know it's value b/c of the difficulty in creating a census.
  - ▶ We use Greek letters to describe them (like  $\mu$  or  $\sigma$ ). If we are talking about a percentage parameter, we use rho ( $\rho$ )
    - ▶ Ex: If we wanted to compare the IQ's of all American and Asian males it would be impossible, but it's important to realize that  $\mu_{\text{Americans}}$  and  $\mu_{\text{males}}$  exist.
    - ▶ Ex: If we were interested in whether there is a greater percentage of women who eat broccoli than men, we want to know whether  $\rho_{\text{women}} > \rho_{\text{men}}$

# Statistic

- ▶ **A statistic is a number that describes a sample.** The value of a statistic can always be found when we take a sample . It's important to realize that a statistic can change from sample to sample.
- ▶ Statistics use variables like  $\bar{x}$ ,  $s$ , and  $\hat{\theta}$  (non greek).
- ▶ We often use statistics to estimate an unknown parameter.
  - ▶ Ex: I take a random sample of 500 American males and find their IQ's. We find that  $\bar{x} = 103.2$ .
  - ▶ I take a random sample of 200 women and find that 40 like broccoli. Then  $\hat{p}_w = .2$
- ▶ **IMPORTANT! A POPULATION NEEDS TO BE AT LEAST 10 TIMES AS BIG AS A SAMPLE TAKEN FROM IT. IF NOT, YOU NEED A SMALLER SAMPLE**

# Bias

- ▶ We say something is biased if it is a poor predictor

## *Unbiased Statistic/Unbiased Estimator*

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

# Variability

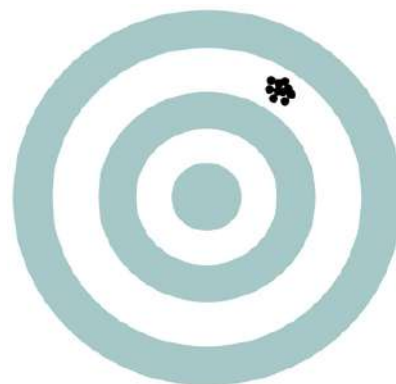
## ***Variability of a Statistic***

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling design and the size of the sample. Larger samples give smaller spread. As long as the population is much larger than the sample (say, at least 10 times as large), the spread of the sampling distribution is approximately the same for any population size.

\*Variability of population doesn't change- (scoop example)  
size of scoop matters

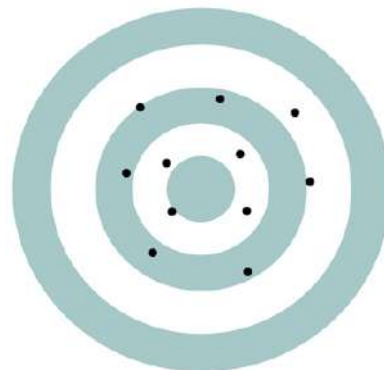
# How can we use samples to find parameters if they give us different results?

- ▶ Imagine an archer shooting many arrows at a target: 4 situations can occur



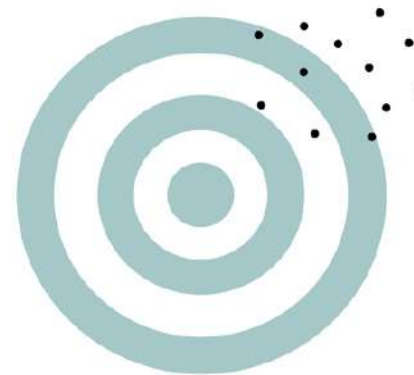
High bias, low variability

(a)



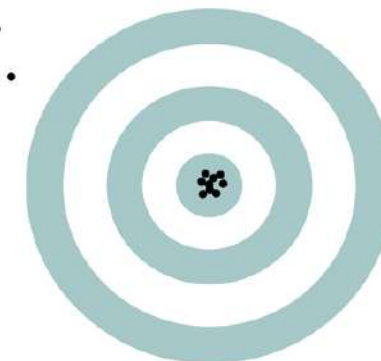
Low bias, high variability

(b)



High bias, high variability

(c)



The ideal: low bias, low variability

(d)

- ▶ a) consistent but off target. b) all over the place. tends to average a bulls eye but each result is far from center. c) worse than a as the archer is consistently missing high and to the right but nearly as consistently as situation a. d) is ideal- low bias and low variability.