Sampling Distributions

A Sampling Distribution

We are moving from descriptive statistics to inferential statistics.

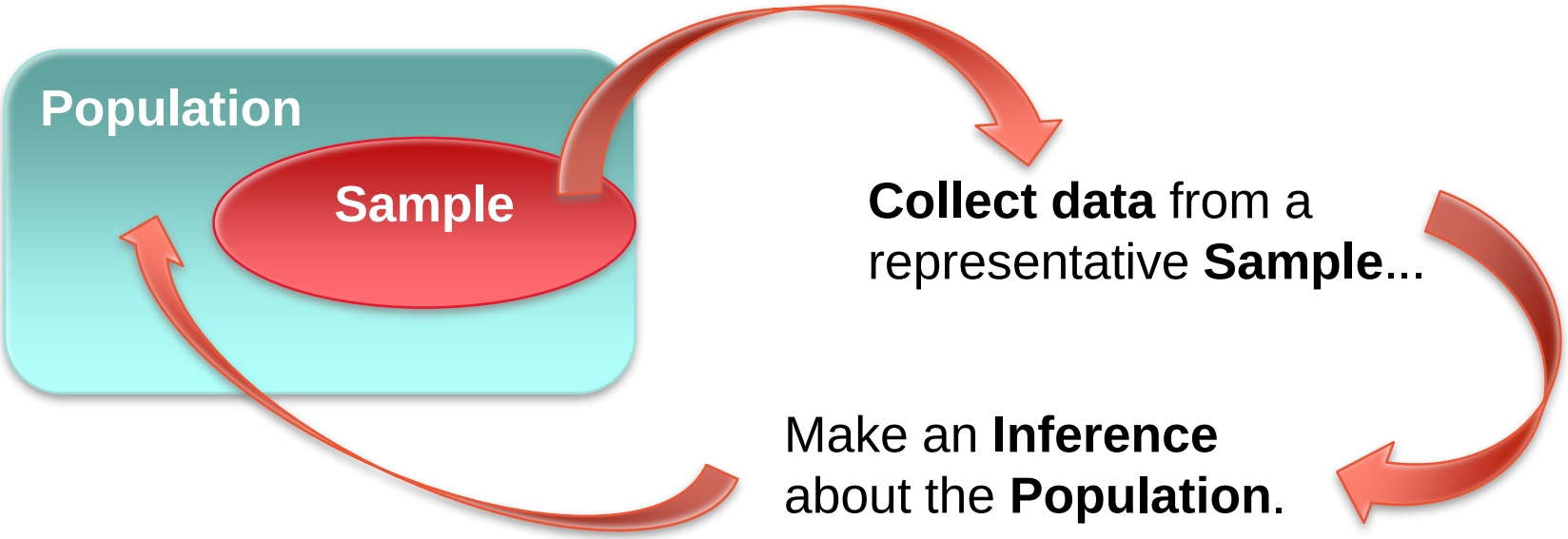
Inferential statistics allow the researcher to come to conclusions about a population on the basis of descriptive statistics about a sample.

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.



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.

A Sampling Distribution

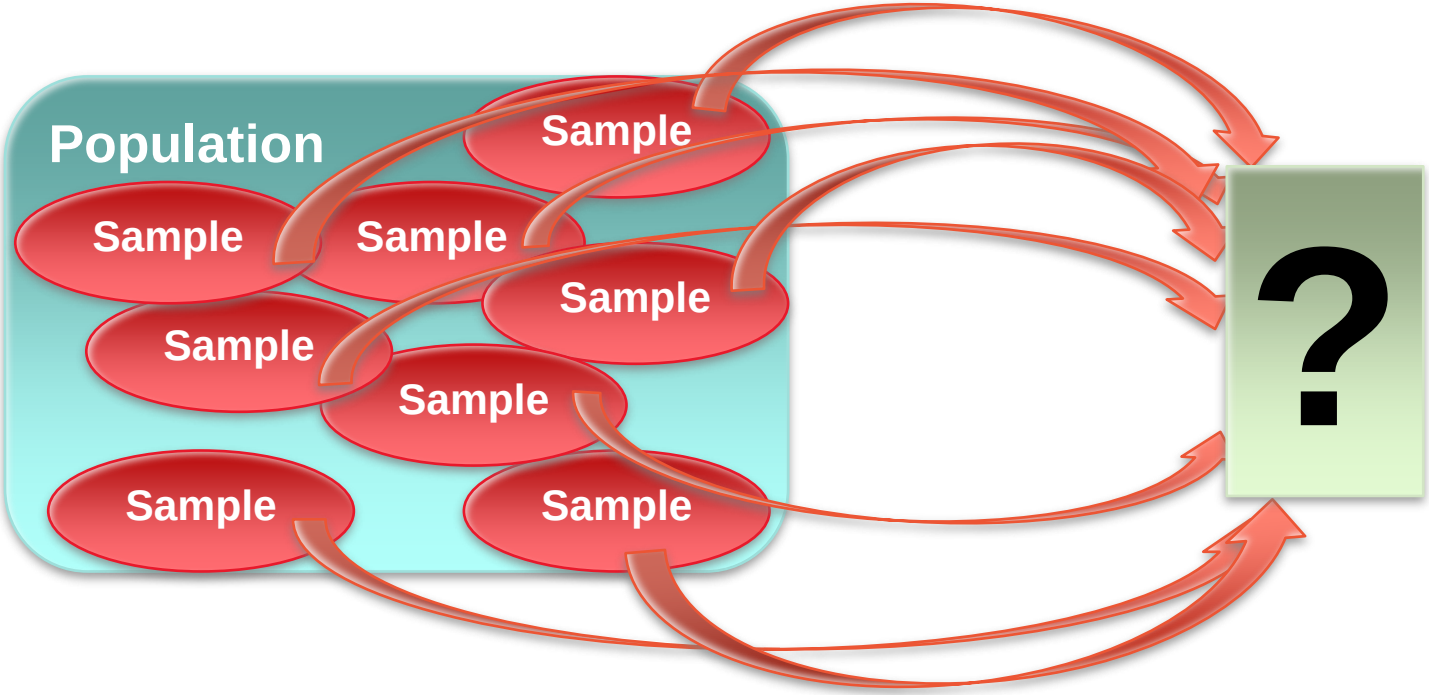
- A theoretical frequency distribution of the scores for or values of a statistic, such as a mean. Any statistic that can be computed for a sample has a sampling distribution.
- A sampling distribution is the distribution of statistics that *would be* produced in repeated random sampling (with replacement) from the same population.
- It is all possible values of a statistic and their probabilities of occurring for a sample of a particular size.
- Sampling distributions are used to calculate the probability that sample statistics could have occurred by chance and thus to decide whether something that is true of a sample statistic is also likely to be true of a population parameter.

Sampling Variability

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

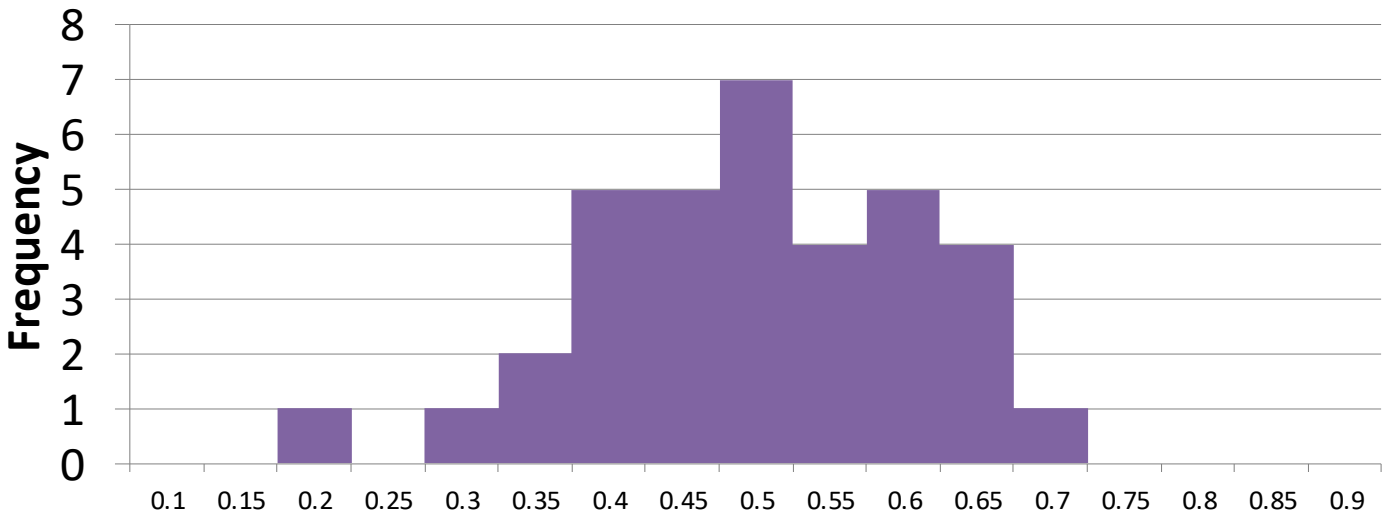
This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?"



Example: Reaching for Chips

- **Take a sample of 20 chips, record the sample proportion of chips, and return all chips to the bag.**



Sampling Distribution

In the previous activity, we took a handful of different samples of 20 chips. There are many, many possible SRSs of size 20 from a population of size 200. 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.**

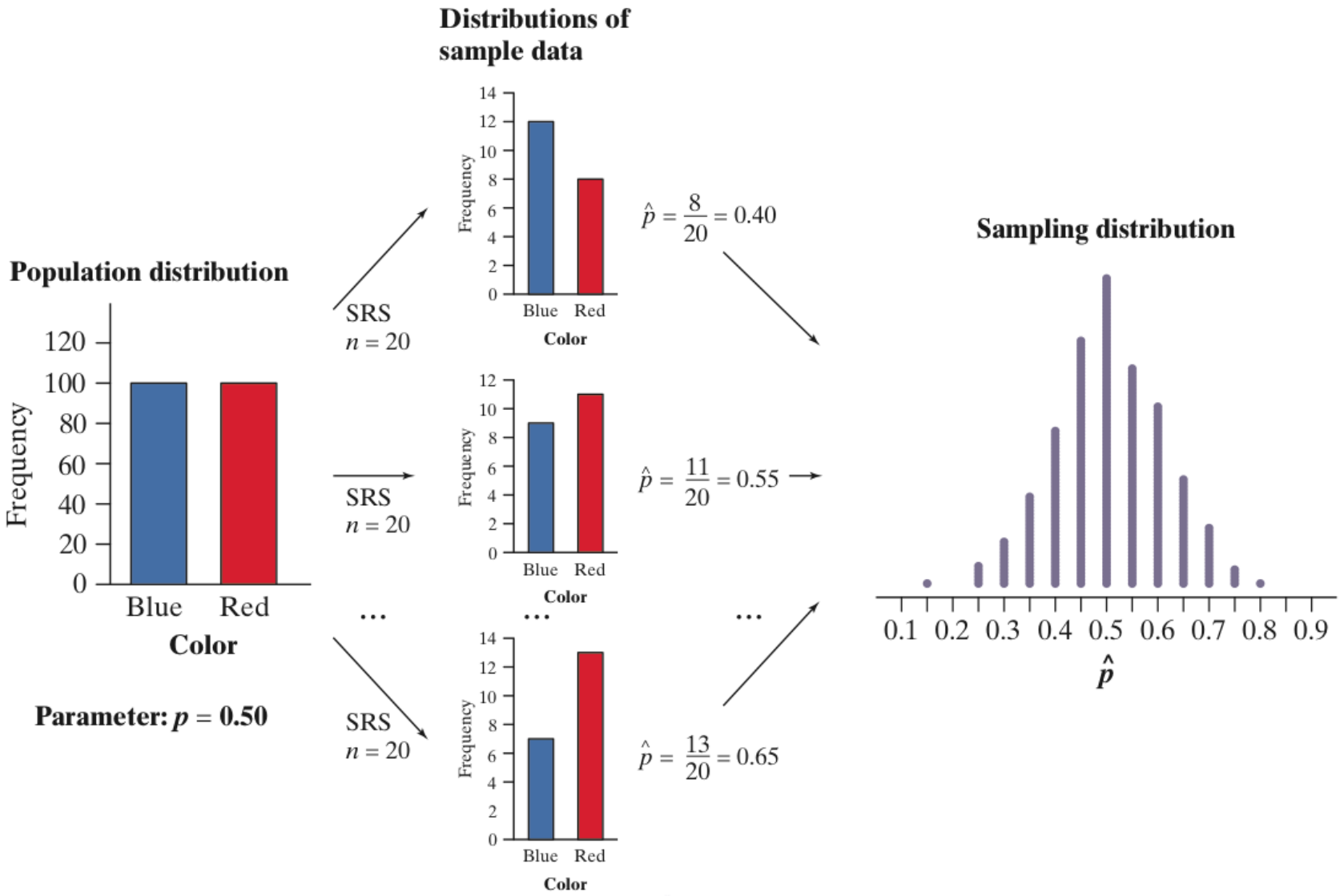
Definition:

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.

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



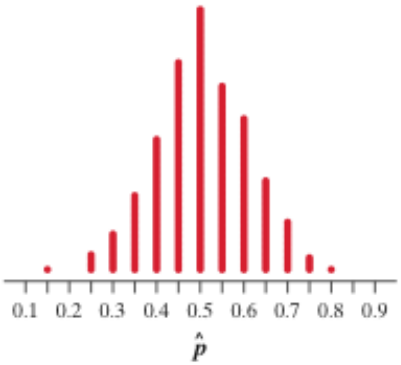
What Is a Sampling Distribution?

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.

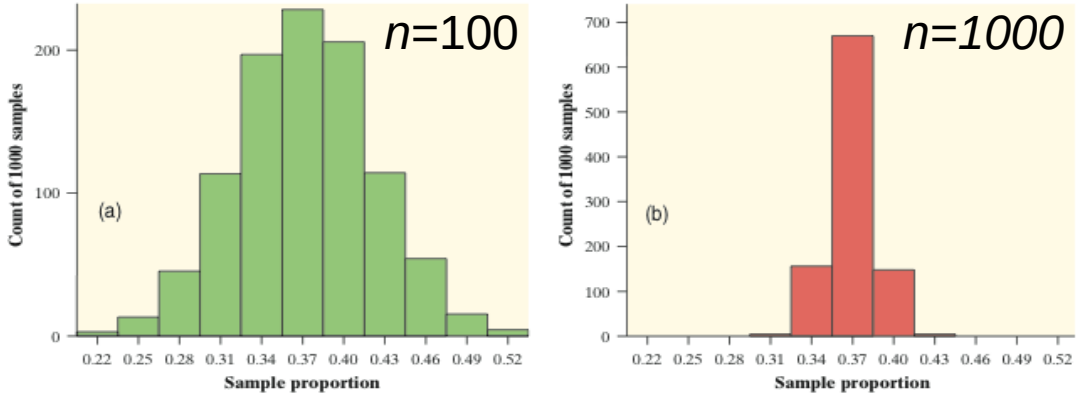
Definition:

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.

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.

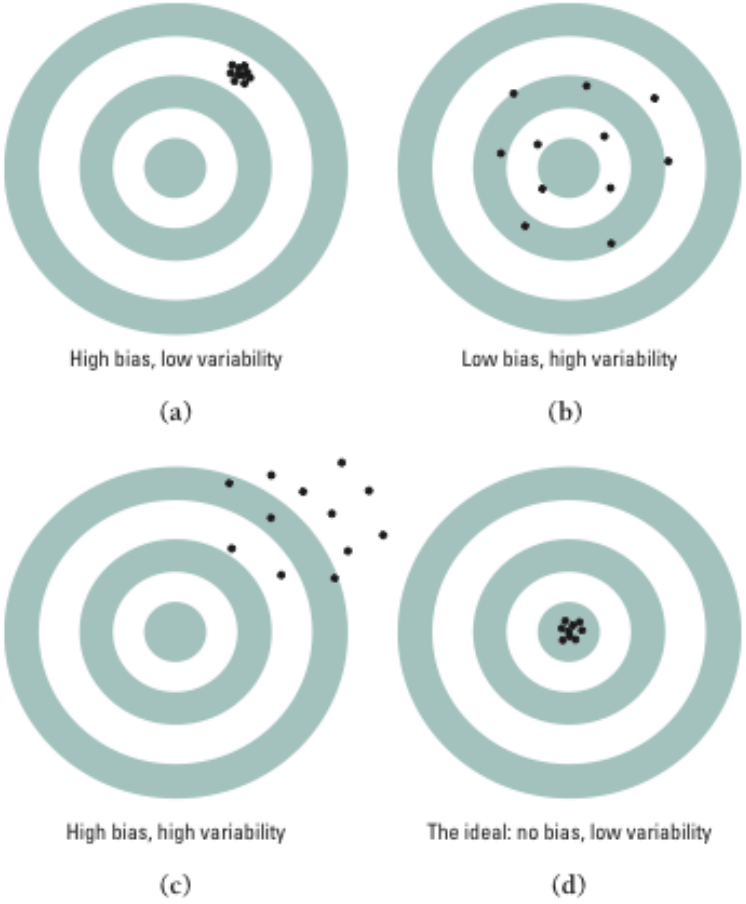
Variability of a Statistic

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.

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. 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.

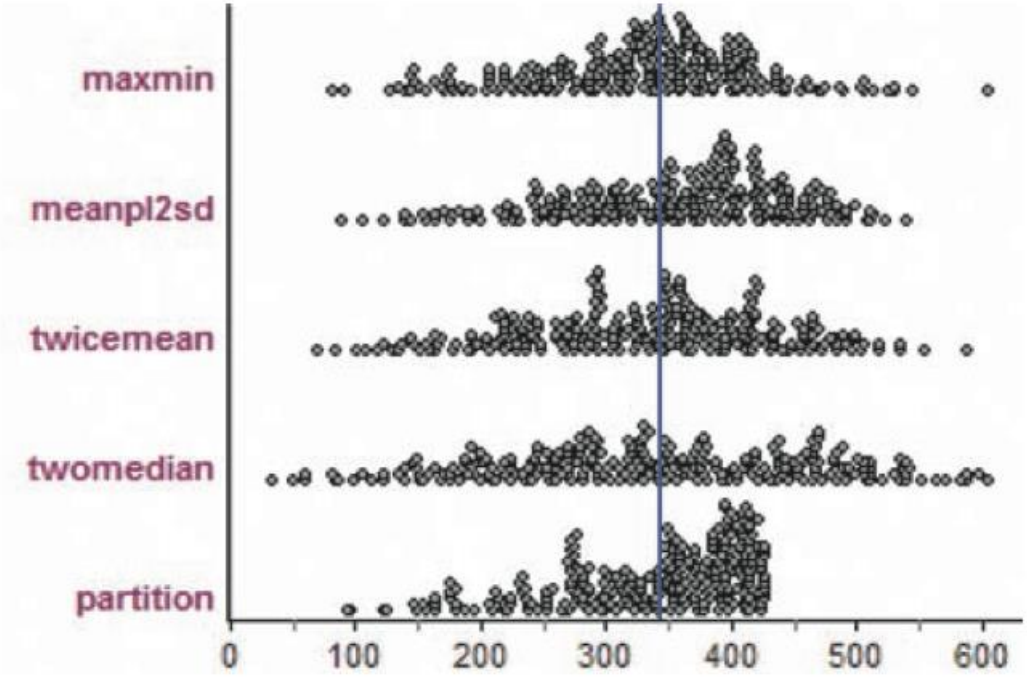
High **variability** means that repeated shots are widely scattered on the target. Repeated samples do not give very similar results.

The lesson about center and spread is clear: **given a choice of statistics to estimate an unknown parameter, choose one with no or low bias and minimum variability**.

Describing 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.



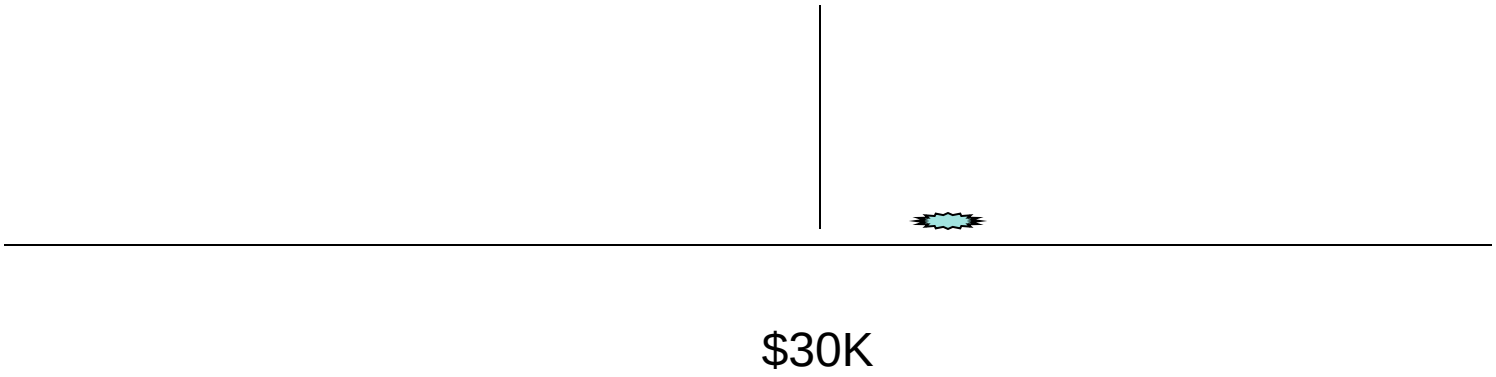
Sampling distributions for different statistics used to estimate the number of tanks in the German Tank problem. The blue line represents the true number of tanks.

Note the different shapes. Which statistic gives the best estimator? Why?

A Sampling Distribution

Let's create a sampling distribution of means…

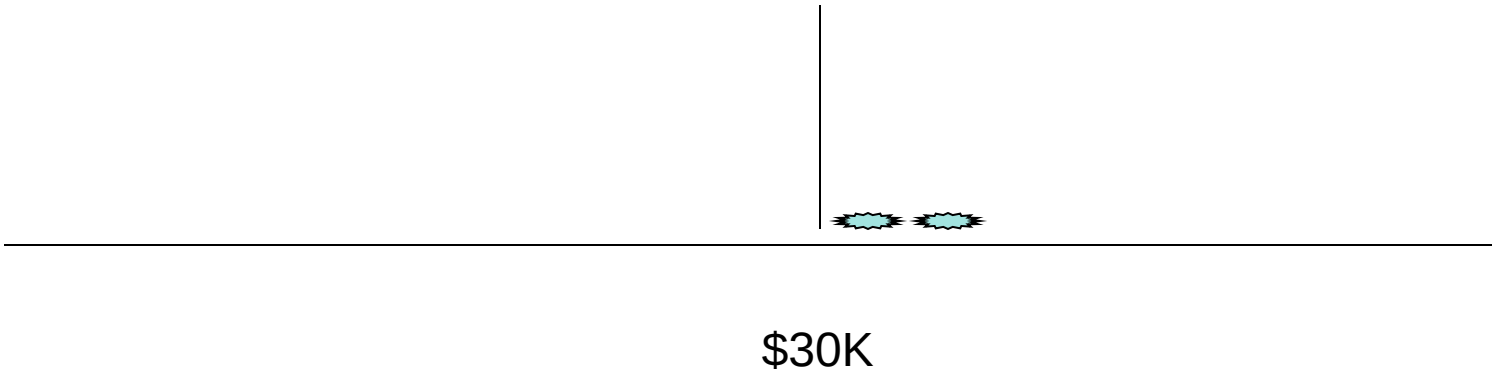
Take a sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

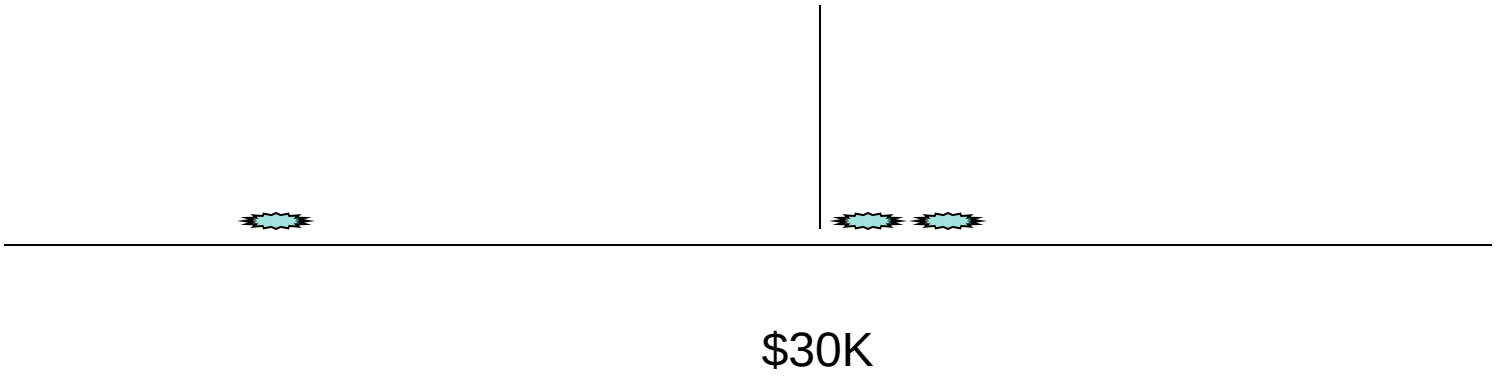
Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

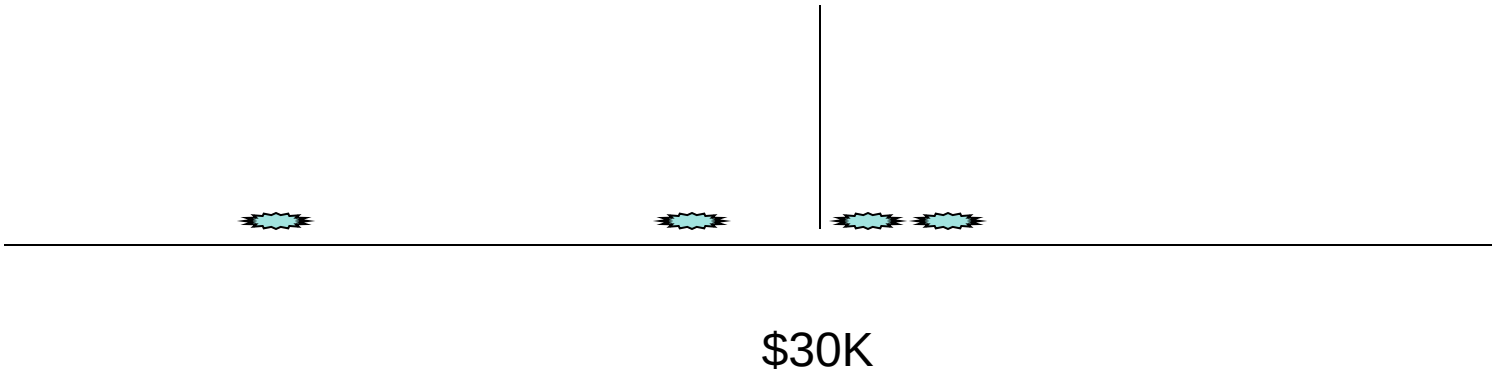
Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

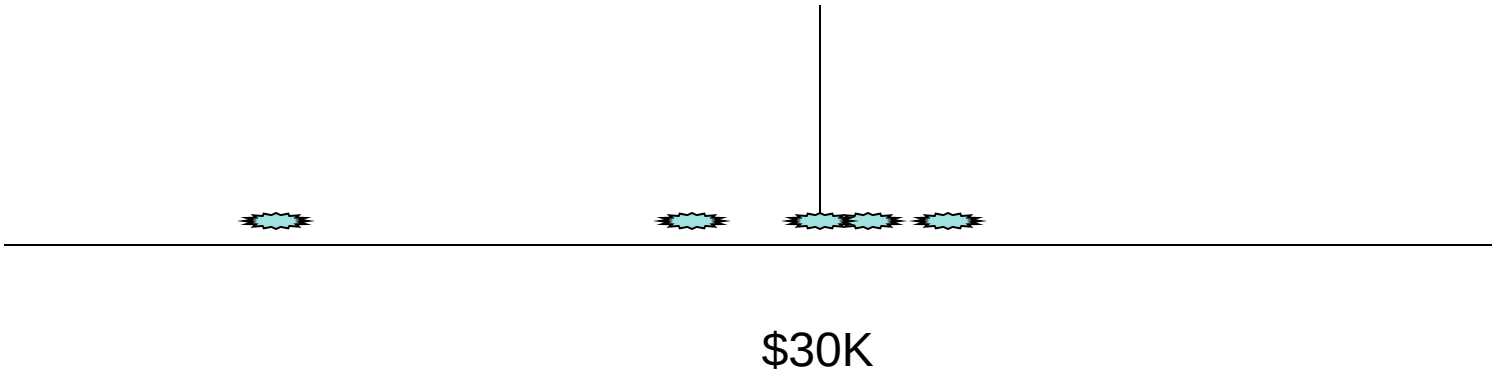
Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

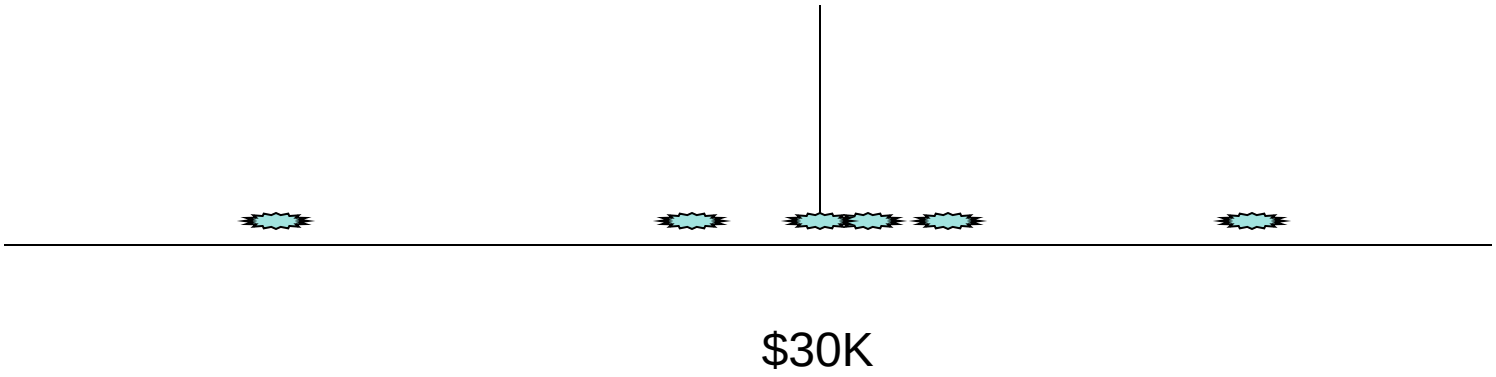
Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

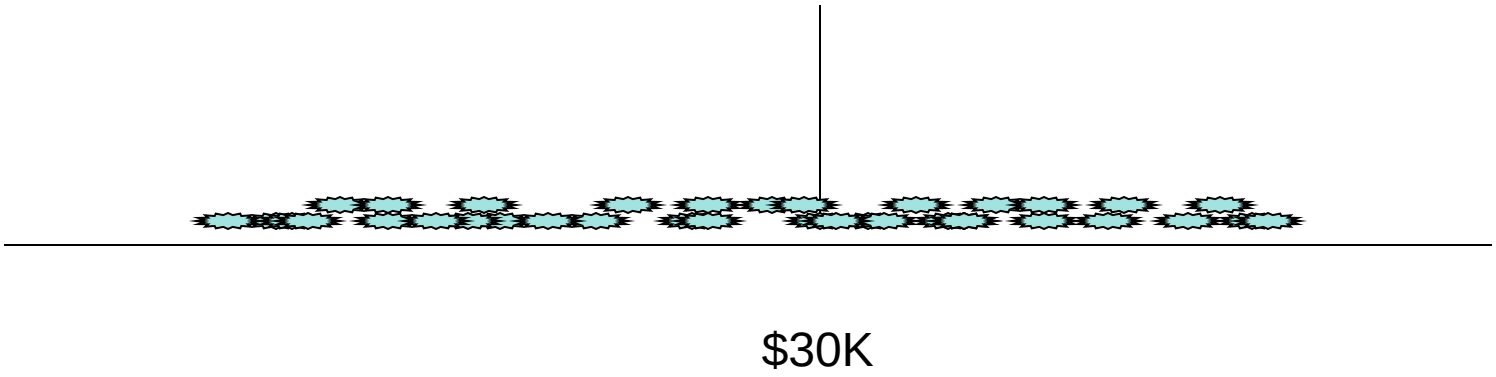
Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

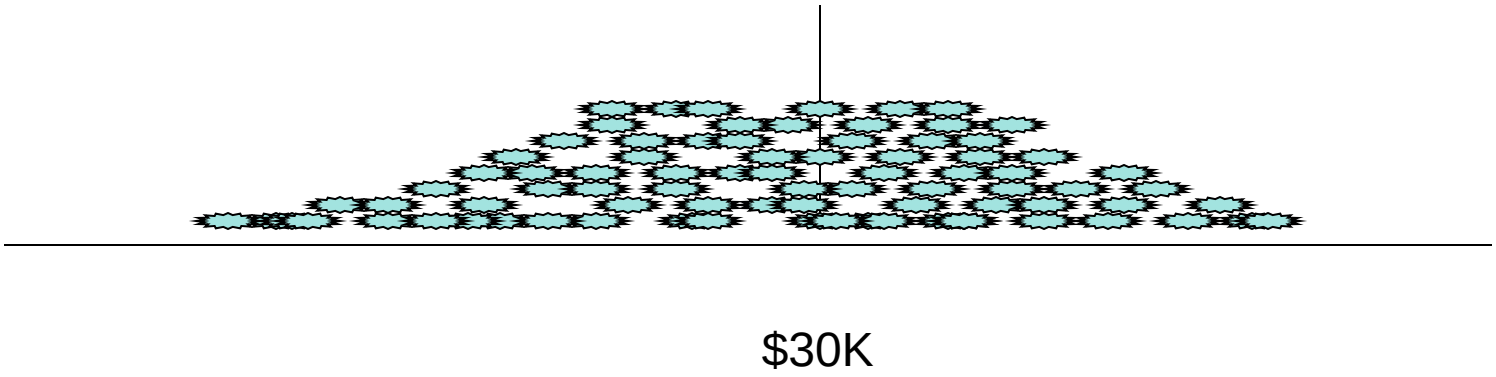
Let's repeat sampling of sizes 1,500 from the US. Record the mean incomes. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

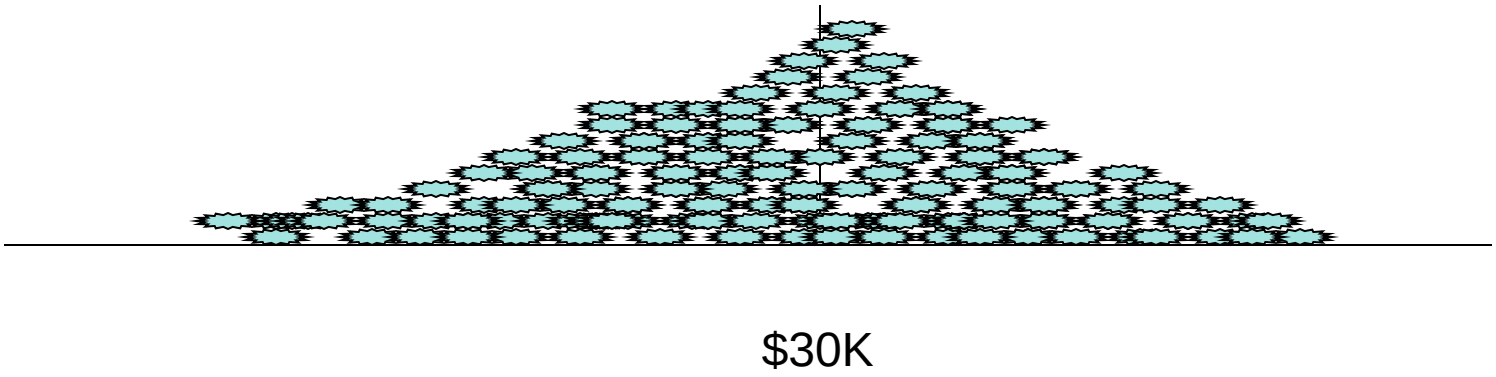
Let's repeat sampling of sizes 1,500 from the US. Record the mean incomes. Our census said the mean is $30K.



A Sampling Distribution

Let's create a sampling distribution of means…

Let's repeat sampling of sizes 1,500 from the US. Record the mean incomes. Our census said the mean is $30K.



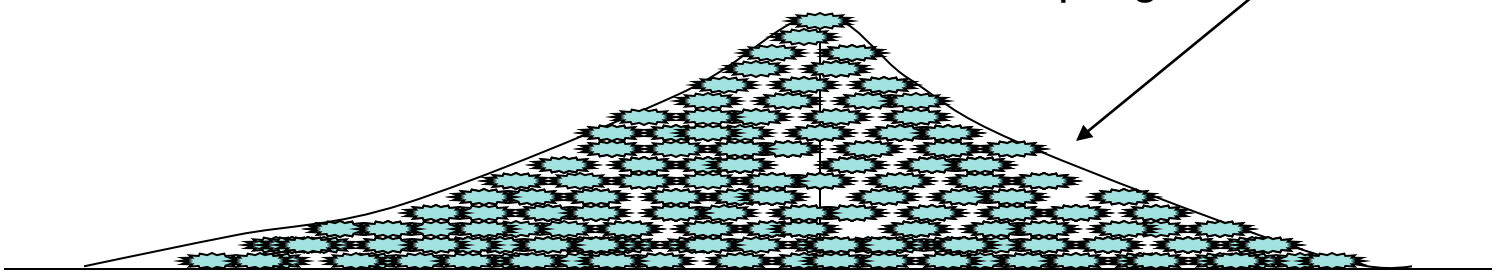
$30K

A Sampling Distribution

Let's create a sampling distribution of means…

Let's repeat sampling of sizes 1,500 from the US. Record the mean incomes. Our census said the mean is $30K.

The sample means would stack up in a normal curve. A normal sampling distribution.

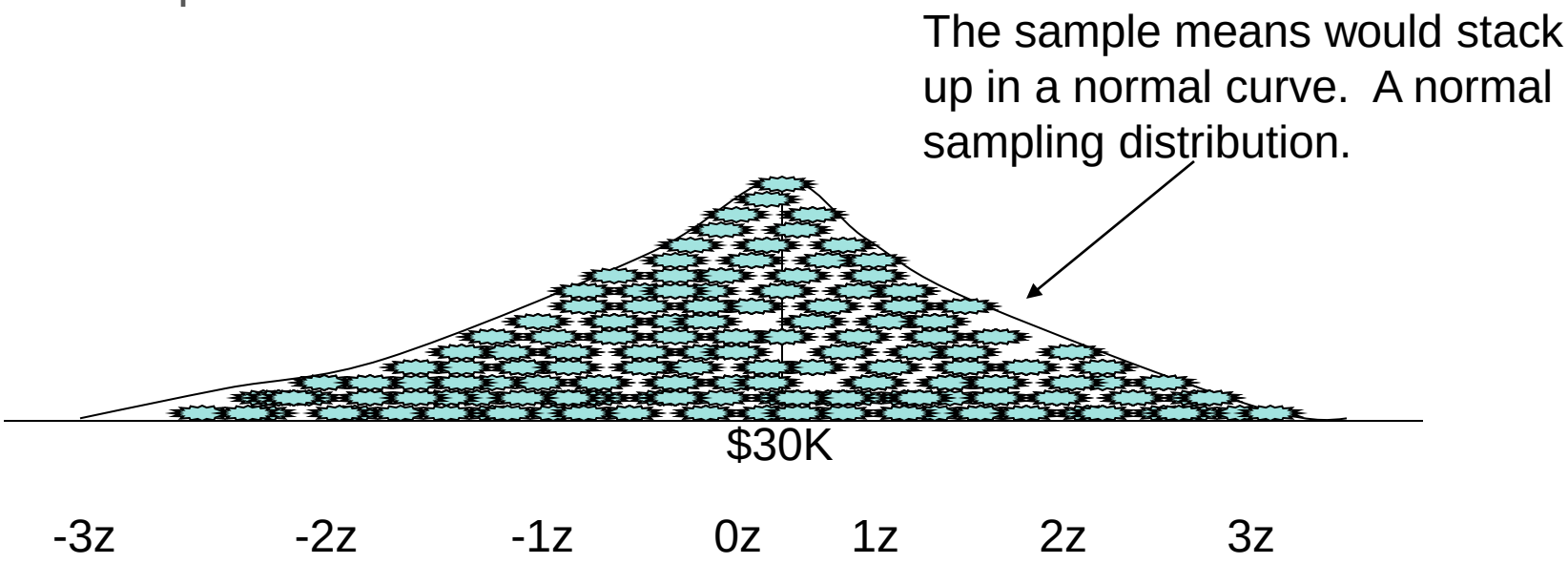


$30K

A Sampling Distribution

Say that the standard deviation of this distribution is $10K.

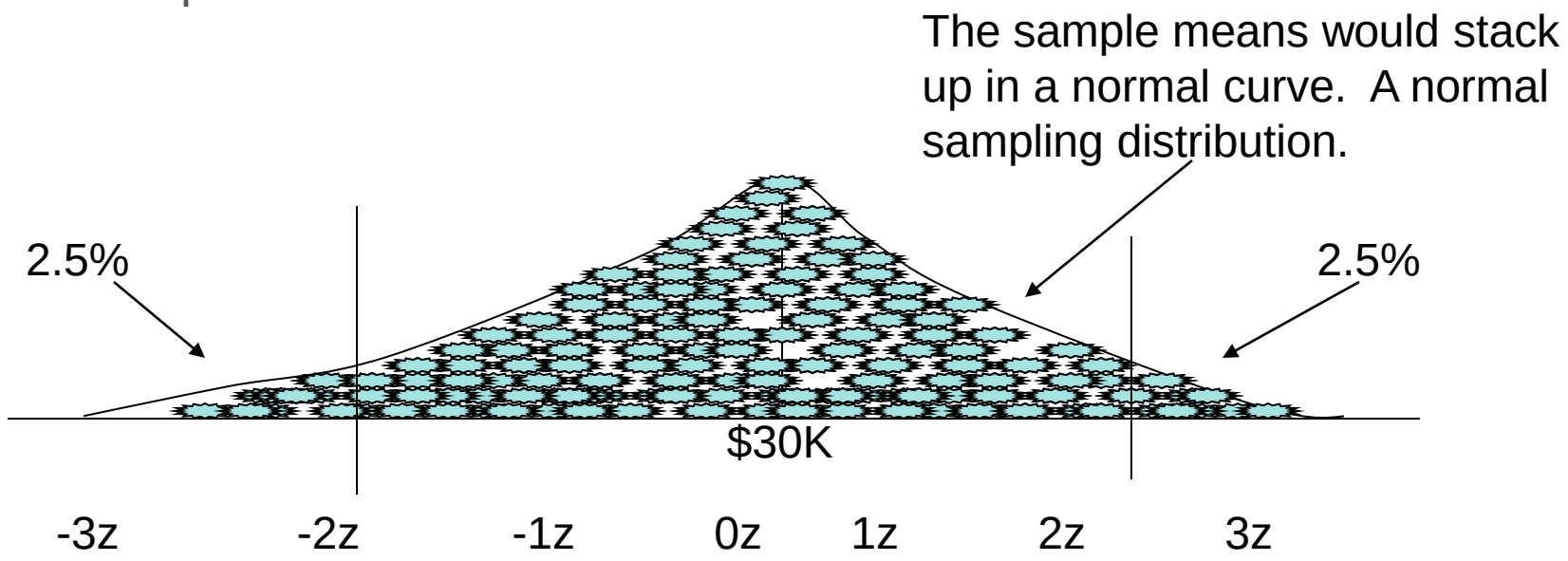
Think back to the empirical rule. What are the odds you would get a sample mean that is more than $20K off.



A Sampling Distribution

Say that the standard deviation of this distribution is $10K.

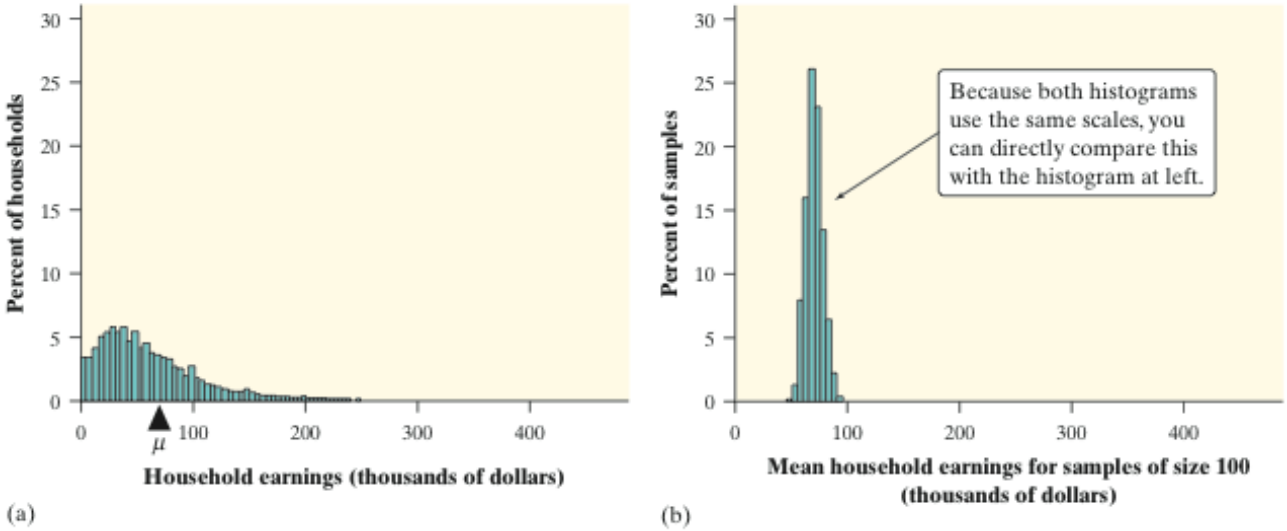
Think back to the empirical rule. What are the odds you would get a sample mean that is more than $20K off.



Sample Means

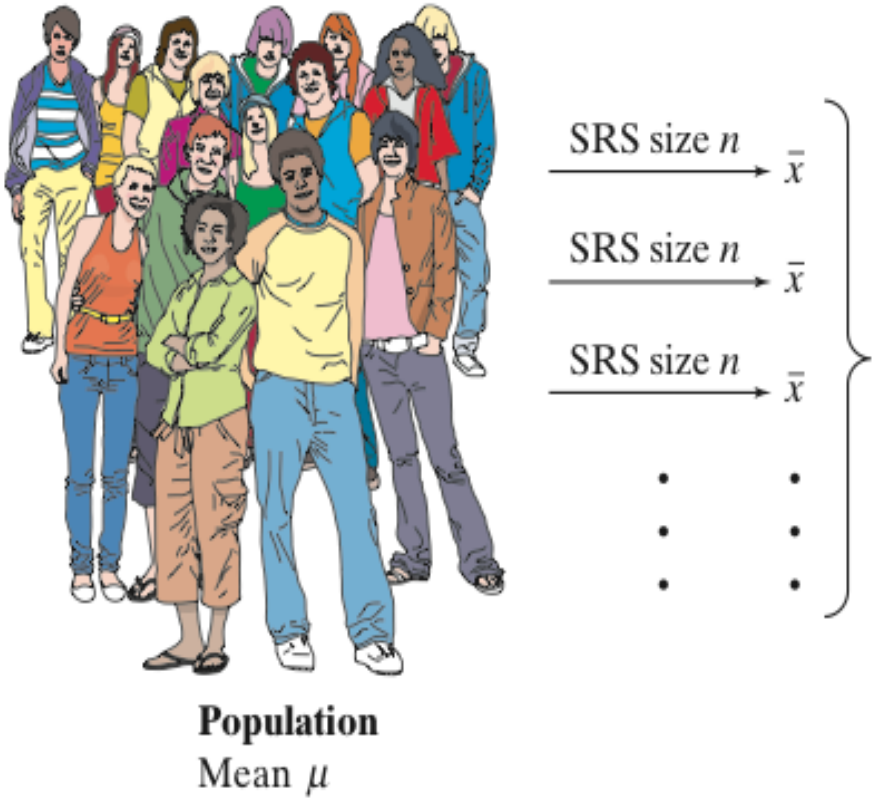
Sample proportions arise most often when we are interested in categorical variables. When we record quantitative variables we are interested in other statistics such as the median or mean or standard deviation of the variable. Sample means are among the most common statistics.

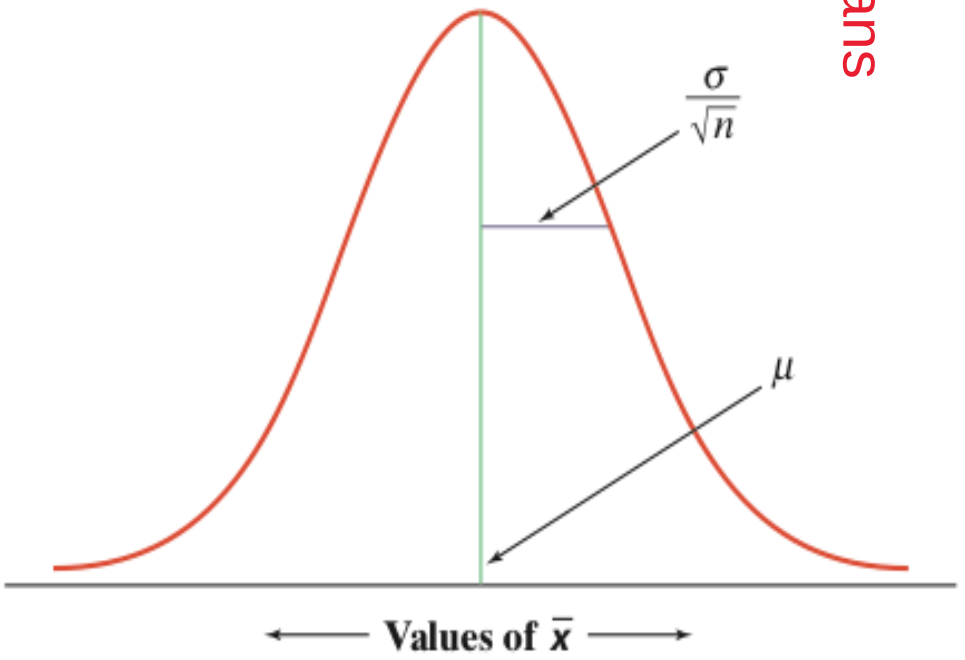
Consider the mean household earnings for samples of size 100. Compare the population distribution on the left with the sampling distribution on the right. What do you notice about the shape, center, and spread of each?



The Sampling Distribution of $\bar{x}$

When we choose many SRSs from a population, the sampling distribution of the sample mean is centered at the population mean μ and is less spread out than the population distribution. Here are the facts.





Sampling from a Normal Population

We have described the mean and standard deviation of the sampling distribution of the sample mean $\bar{x}$ but not its shape. That's because the shape of the distribution of $\bar{x}$ depends on the shape of the population distribution.

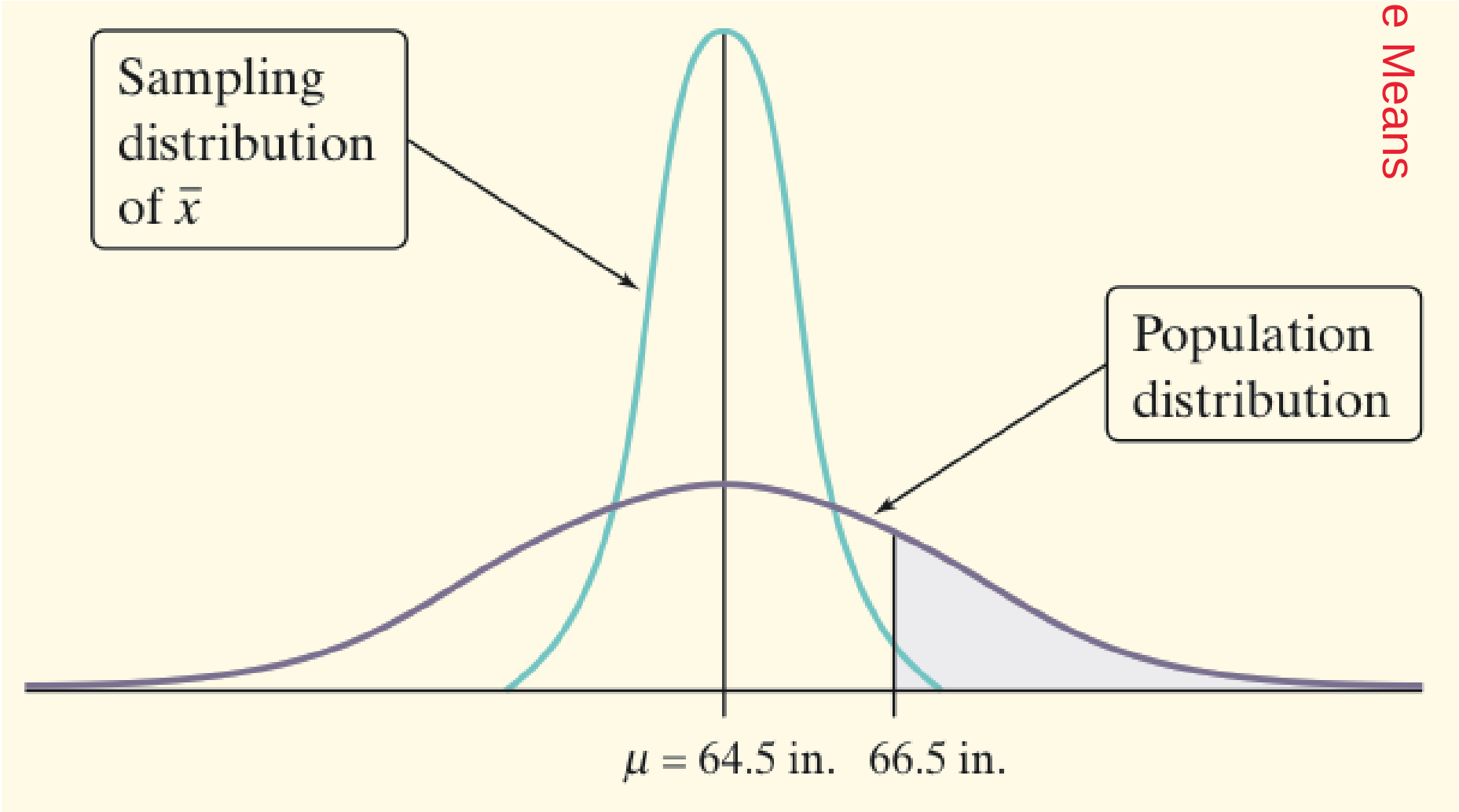
In one important case, there is a simple relationship between the two distributions. If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$. *This is true no matter what the sample size is*.

Sampling Distribution of a Sample Mean from a Normal Population

Suppose that a population is Normally distributed with mean μ and standard deviation σ. Then the sampling distribution of $\bar{x}$ has the Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, provided that the 10% condition is met.

Example: Young Women's Heights

The height of young women follows a Normal distribution with mean μ = 64.5 inches and standard deviation σ = 2.5 inches.



The Central Limit Theorem

Most population distributions are not Normal. What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?

It is a remarkable fact that as the sample size increases, the distribution of sample means changes its shape: it looks less like that of the population and more like a Normal distribution! When the sample is large enough, the distribution of sample means is very close to Normal, *no matter what shape the population distribution has*, as long as the population has a finite standard deviation.

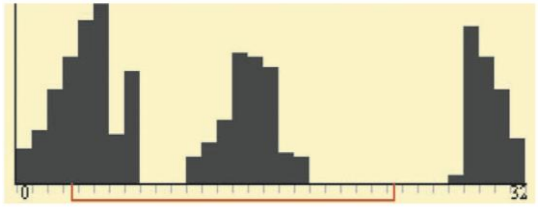
Definition:

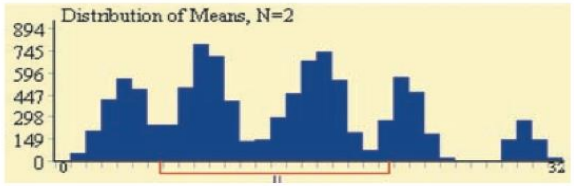
Draw an SRS of size n from any population with mean μ and finite standard deviation σ. The **central limit theorem (CLT)** says that when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal.

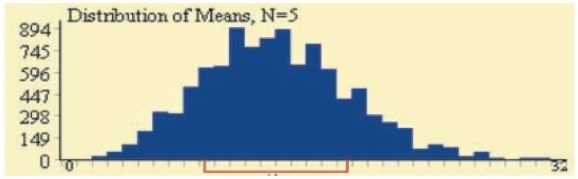
Note: How large a sample size *n* is needed for the sampling distribution to be close to Normal depends on the shape of the population distribution. More observations are required if the population distribution is far from Normal.

The Central Limit Theorem

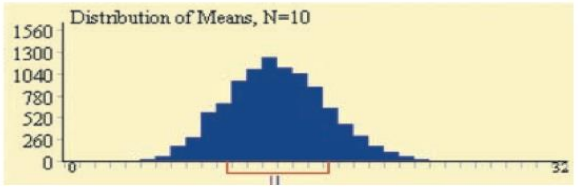
Consider the strange population distribution from the Rice University sampling distribution applet.



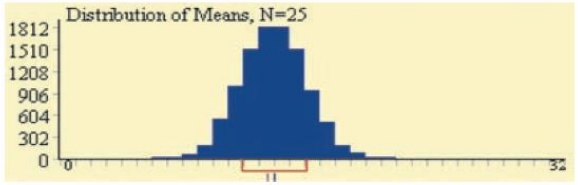




(b)



(c)



(d)

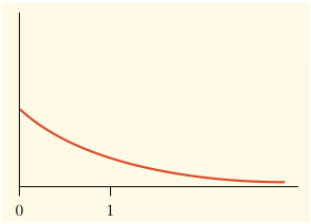
Describe the shape of the sampling distributions as *n* increases. What do you notice?

Normal Condition for Sample Means

If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$. This is true no matter what the sample size n is.

If the population distribution is not Normal, the central limit theorem tells us that the sampling distribution of $\bar{x}$ will be approximately Normal in most cases if $n \geq 30$.

Example: Servicing Air Conditioners



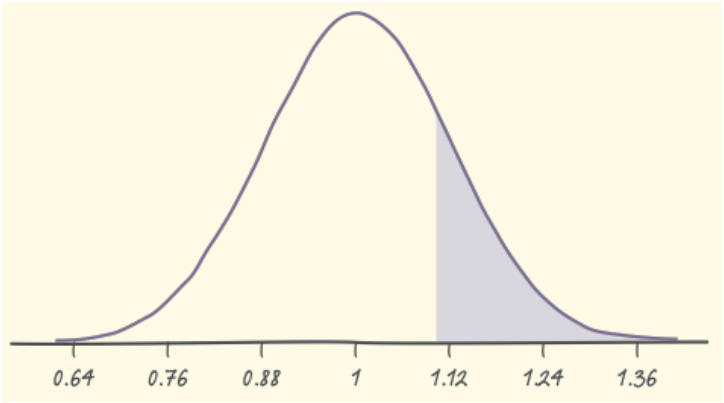
Based on service records from the past year, the time (in hours) that a technician requires to complete preventative maintenance on an air conditioner follows the distribution that is strongly right-skewed, and whose most likely outcomes are close to 0. The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$

Your company will service an SRS of 70 air conditioners. You have budgeted 1.1 hours per unit. Will this be enough?

Since the 10% condition is met (there are more than 10(70)=700 air conditioners in the population), the sampling distribution of the mean time spent working on the 70 units has

$$\mu_{\bar{x}} = \mu = 1 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{70}} = 0.12$$

The sampling distribution of the mean time spent working is approximately $N(1, 0.12)$ since $n = 70 \geq 30$.



We need to find P(mean time > 1.1 hours)

$$z = \frac{1.1 - 1}{0.12} = 0.83$$

$$P(\bar{x} > 1.1) = P(Z > 0.83)$$
$$= 1 - 0.7967 = 0.2033$$

If you budget 1.1 hours per unit, there is a 20% chance the technicians will not complete the work within the budgeted time.

Sample Means

Summary

In this section, we learned that…

- ✓ When we want information about the population mean μ for some variable, we often take an SRS and use the sample mean $\bar{x}$ to estimate the unknown parameter μ. The **sampling distribution** of $\bar{x}$ describes how the statistic varies in all possible samples of the same size from the population.

- ✓ The mean of the sampling distribution is μ, so that $\bar{x}$ is an unbiased estimator of μ.

- ✓ The standard deviation of the sampling distribution of $\bar{x}$ is $\sigma/\sqrt{n}$ for an SRS of size n if the population has standard deviation σ. This formula can be used if the population is at least 10 times as large as the sample (10% condition).

Sample Means

Summary

In this section, we learned that…

- ✓ Choose an SRS of size n from a population with mean μ and standard deviation σ. If the population is Normal, then so is the sampling distribution of the sample mean $\bar{x}$. If the population distribtution is not Norma the **central limit theorem (CLT)** states that when n is large, the sampling distribution of $\bar{x}$ is approximately Normal.

- ✓ We can use a Normal distribution to calculate approximate probabilities for events involving $\bar{x}$ whenever the Normal condition is met

  If the population distribution is Normal, so is the sampling distribution of $\bar{x}$.

  If $n \geq 30$, the CLT tells us that the sampling distribution of $\bar{x}$ will be approximately Normal in most cases.

What Is a Sampling Distribution?

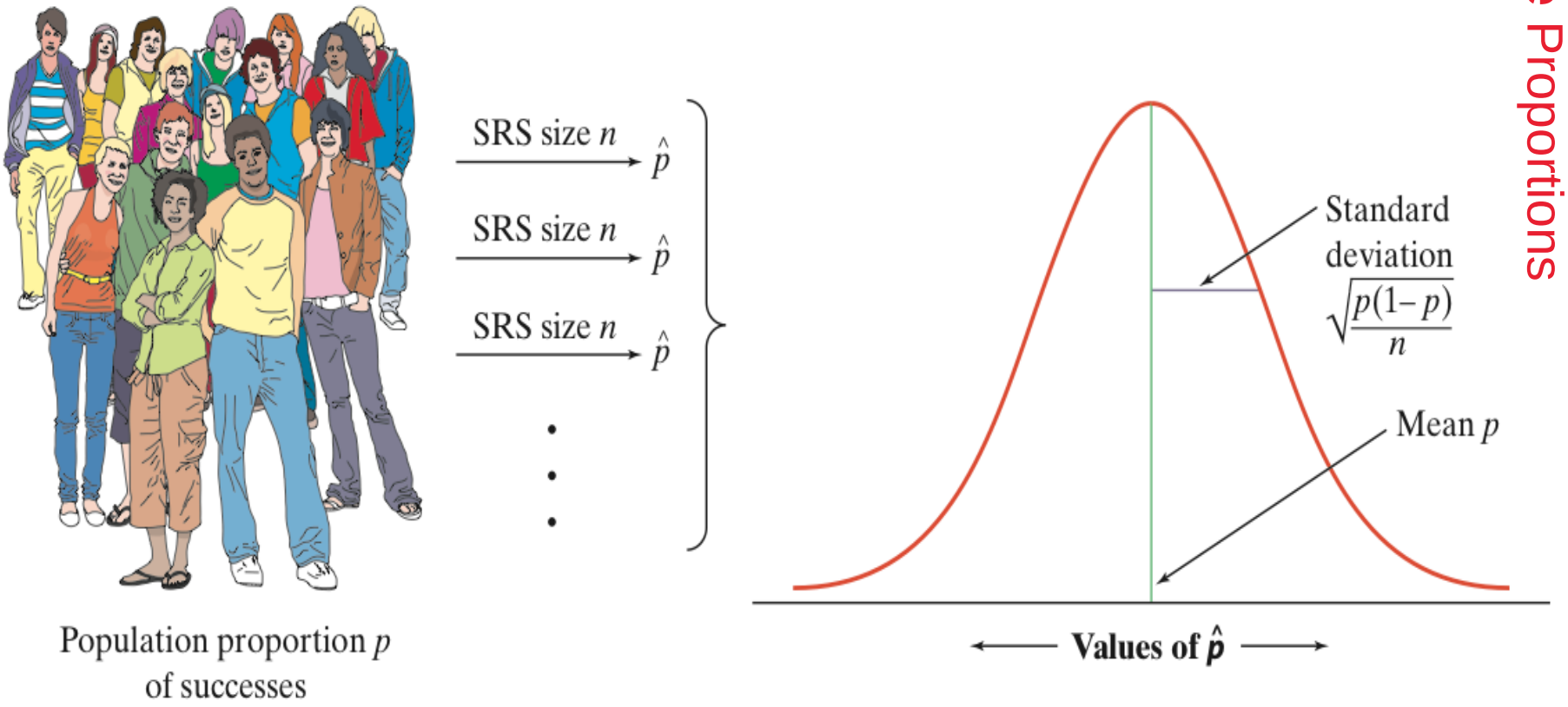
Summary

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.

The Sampling Distribution of $\hat{p}$

We can summarize the facts about the sampling distribution of $\hat{p}$ as follows:



Using the Normal Approximation for $\hat{p}$

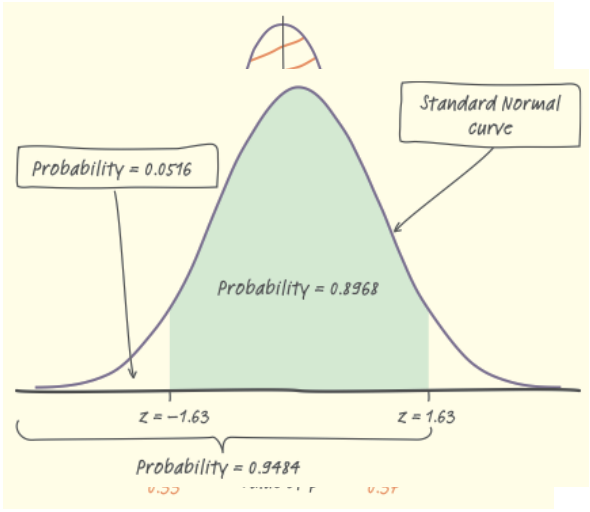
Inference about a population proportion p is based on the sampling distribution of $\hat{p}$. When the sample size is large enough for np and $n(1-p)$ to both be at least 10 (the Normal condition), the sampling distribution of $\hat{p}$ is approximately Normal.

STEP 4

A polling organization asks an SRS of 1500 first-year college students how far away their home is. Suppose that 35% of all first-year students actually attend college within 50 miles of home. What is the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value?

STATE: We want to find the probability that the sample proportion falls between 0.33 and 0.37 (within 2 percentage points, or 0.02, of 0.35).

PLAN: We have an SRS of size n = 1500 drawn from a population in which the proportion p = 0.35 attend college within 50 miles of home.



$$\mu_{\hat{p}} = 0.35 \qquad \sigma_{\hat{p}} = \sqrt{\frac{(0.35)(0.65)}{1500}} = 0.0123$$

DO: Since np = 1500(0.35) = 525 and $n(1-p)$ = 1500(0.65)=975 are both greater than 10, we'll standardize and then use Table A to find the desired probability.

$$z = \frac{0.33 - 0.35}{0.123} = -1.63 \qquad z = \frac{0.37 - 0.35}{0.123} = 1.63$$

$$P(0.33 \le \hat{p} \le 0.37) = P(-1.63 \le Z \le 1.63) = 0.9484 - 0.0516 = 0.8968$$

CONCLUDE: About 90% of all SRSs of size 1500 will give a result within 2 percentage points of the truth about the population.

Sample Proportions

Summary

In this section, we learned that…

- When we want information about the population proportion p of successes, we often take an SRS and use the sample proportion $\hat{p}$ to estimate the unknown parameter p. The **sampling distribution** of $\hat{p}$ describes how the statistic varies in all possible samples from the population.
- The **mean** of the sampling distribution of $\hat{p}$ is equal to the population proportion p. That is, $\hat{p}$ is an unbiased estimator of p.
- The **standard deviation** of the sampling distribution of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ for an SRS of size n. This formula can be used if the population is at least 10 times as large as the sample (the 10% condition). The standard deviation of $\hat{p}$ gets smaller as the sample size n gets larger.
- When the sample size n is larger, the sampling distribution of $\hat{p}$ is close to a Normal distribution with mean p and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.
- In practice, use this Normal approximation when both $np \geq 10$ and $n(1-p) \geq 10$ (the Normal condition).