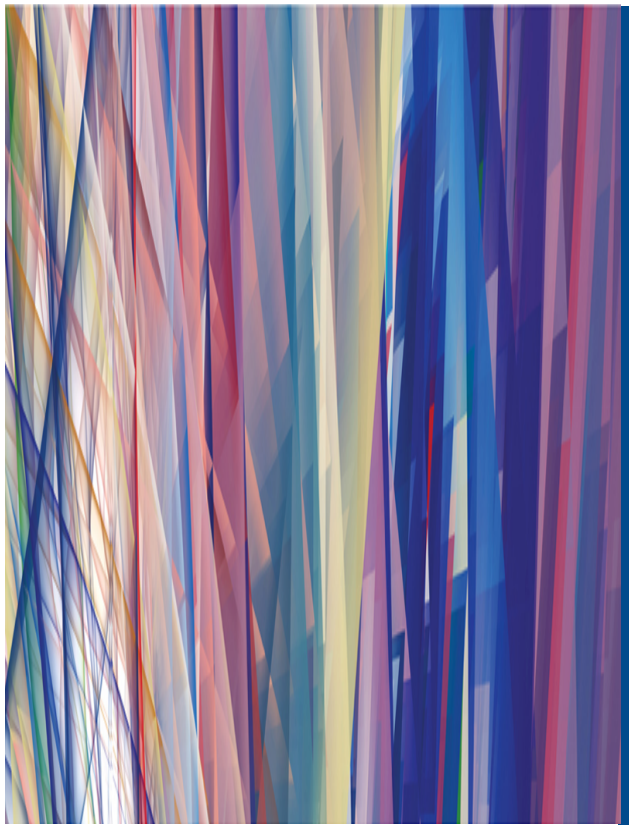


Chapter 5

Sampling Distributions

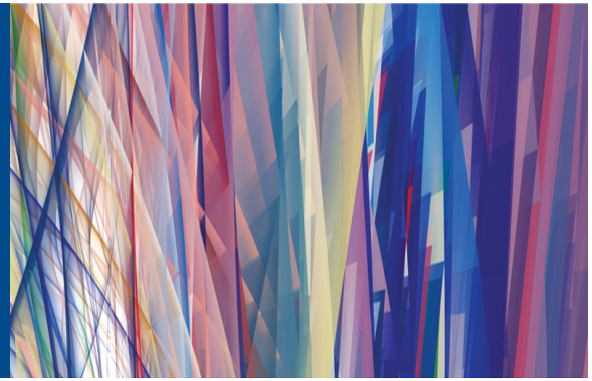
Introduction to the Practice of

STATISTICS SEVENTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

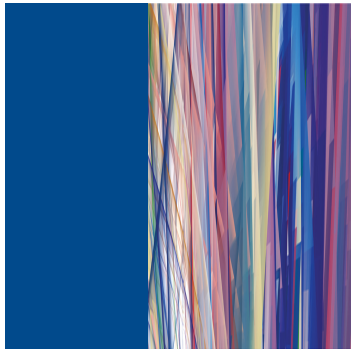
Chapter 5
Sampling Distributions



5.1 The Sampling Distribution of a Sample Mean

5.2 Sampling Distributions for Counts and Proportions

5.1 The Sampling Distribution of a Sample Mean



- Population Distribution vs. Sampling Distribution
- The Mean and Standard Deviation of the Sample Mean
- Sampling Distribution of a Sample Mean
- Central Limit Theorem

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. In statistical practice, the value of a parameter is 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.

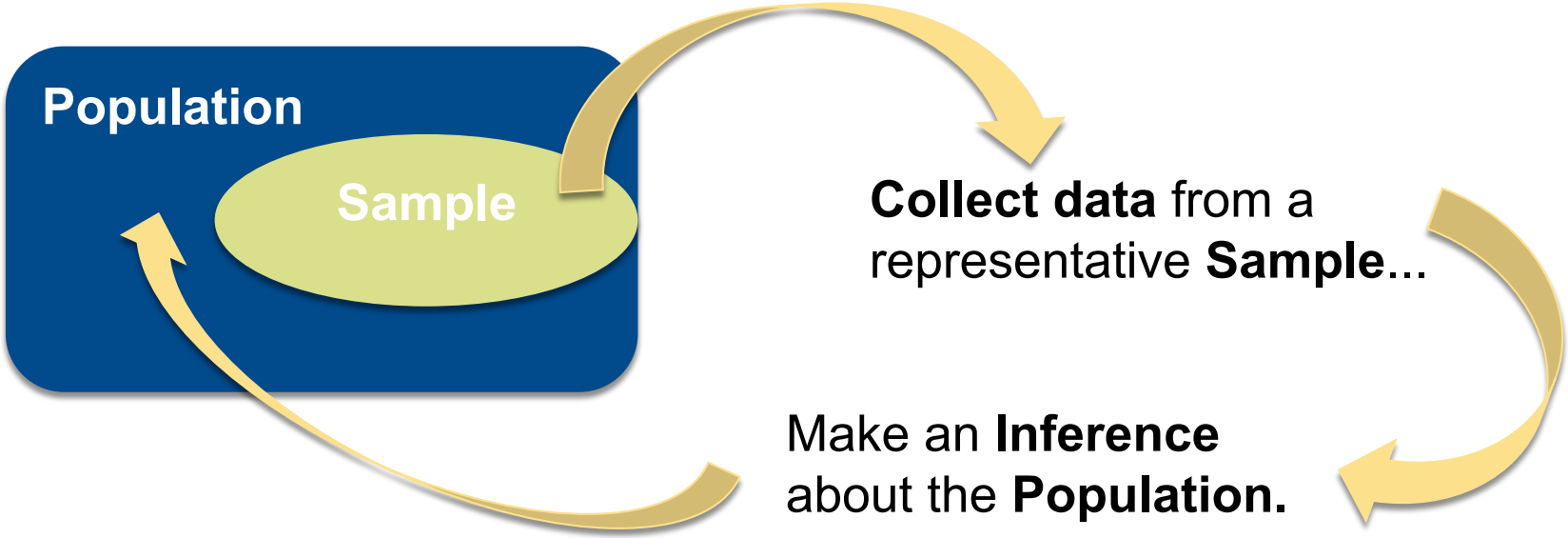
We write μ (the Greek letter mu) for the population mean and σ for the population standard deviation. We write $\bar{x}$ (x-bar) for the sample mean and s for the sample standard deviation.

Statistical Estimation

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.

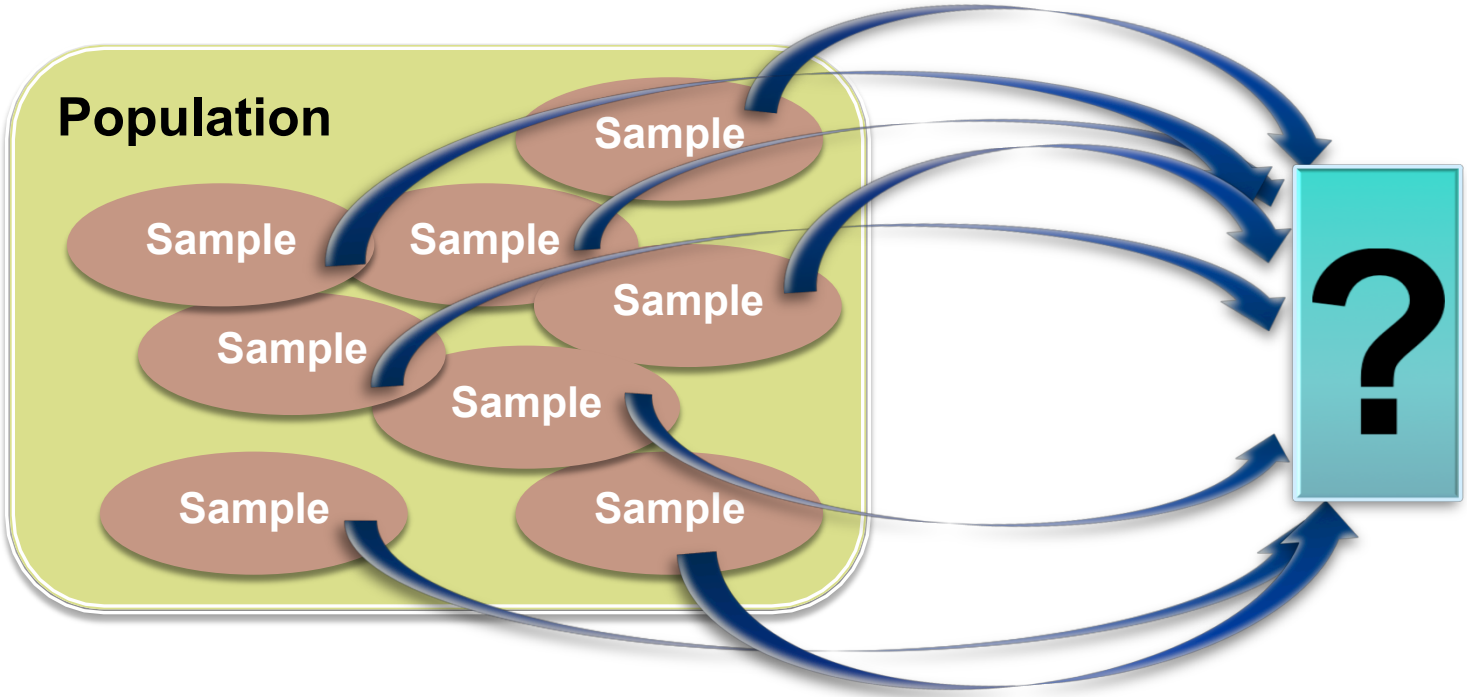
We can think of a statistic as a random variable because it takes numerical values that describe the outcomes of the random sampling process.



Sampling Variability

Different random samples yield different statistics. 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 Distributions

The law of large numbers assures us that if we measure enough subjects, the statistic x-bar will eventually get very close to the unknown parameter μ.

If we took every one of the possible samples of a certain size, calculated the sample mean for each, and graphed all of those values, we would be looking at a **sampling distribution.**

The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

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.

Mean and Standard Deviation of a Sample Mean

Mean of a sampling distribution of a sample mean
There is no tendency for a sample mean to fall systematically above or below m, even if the distribution of the raw data is skewed. Thus, the mean of the sampling distribution is an **unbiased estimate** of the population mean m.

Standard deviation of a sampling distribution of a sample mean
The standard deviation of the sampling distribution measures how much the sample mean varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$.

➔ **Averages are less variable than individual observations.**

Sample: $x_1, x_2, \ldots x_n \Rightarrow \bar{x}$

$$E(\bar{x}) = E\left\{\frac{1}{n}\Sigma x_i\right\}$$

$$= \frac{1}{n} E\{\Sigma x_i\}$$

$$= \frac{1}{n}\{E(x_1 + x_2 + \ldots x_n)\}$$

$$= \frac{1}{n}\{E(x_1) + E(x_2) + \ldots E(x_n)\}$$

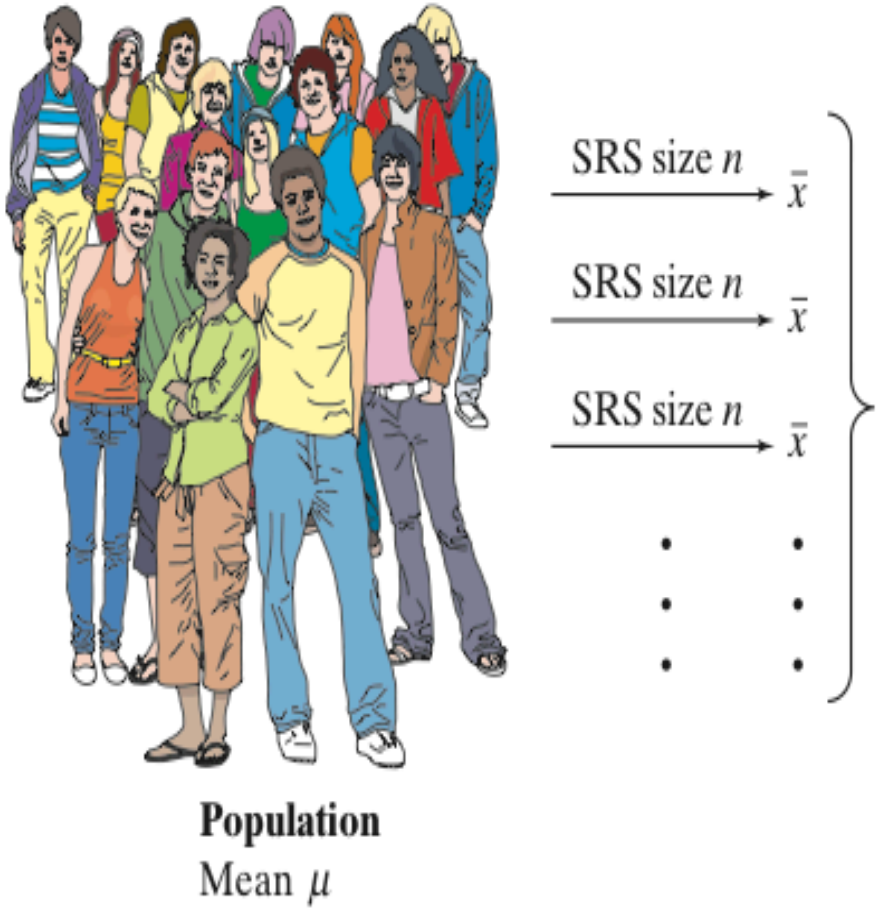
$$= \frac{1}{n}\{\mu + \mu + \ldots + \mu\}$$

$$= \frac{1}{n}\{n\mu\} = \mu$$

$\Rightarrow \bar{x}$ is an unbiased estimator of μ

The Sampling Distribution of a Sample Mean

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.



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?

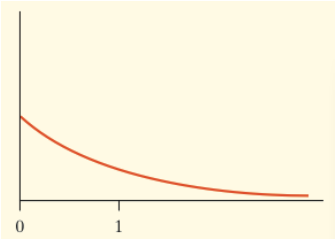
It is a remarkable fact that as the sample size increases, the distribution of sample means changes its shape: it looks less like that of the population and more like a Normal distribution!

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.

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:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Example



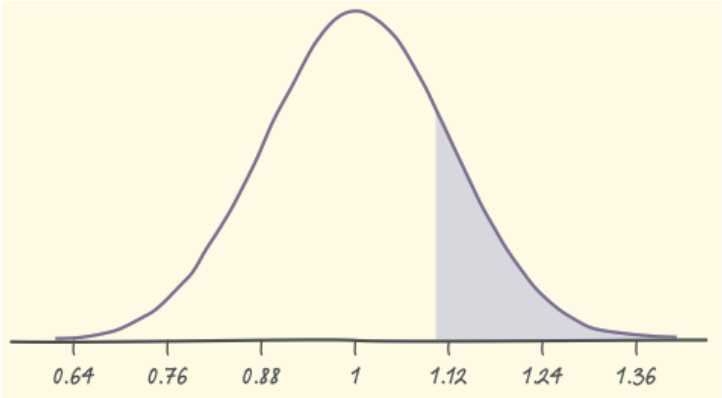
Based on service records from the past year, the time (in hours) that a technician requires to complete preventative maintenance on an air conditioner follows the distribution that is strongly right-skewed, and whose most likely outcomes are close to 0. The mean time is $\mu = 1$ hour and the standard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

The central limit theorem states that the sampling distribution of the mean time spent working on the 70 units is:

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

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



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

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

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

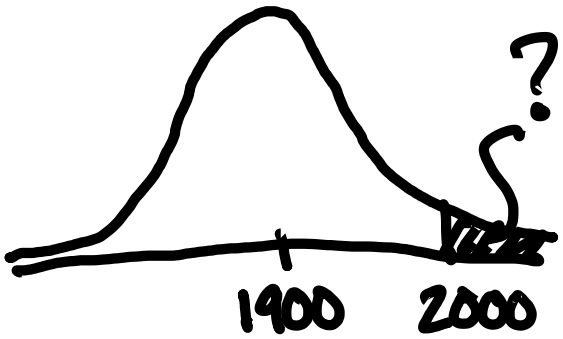
Historical records show that the costs associated with a rescue squad visit average $1900 & have a std dev of $800. Next year's budget includes $1,000,000 for an anticipated 500 visits. Find the probability that the budget is exhausted.

$\mu_X = 1900 \quad \sigma_X = 800$

Key "trick"

$$P\left(\sum X_i > 1{,}000{,}000\right) = P\left(\bar{X} > 2{,}000\right)$$

CLT $\left\{ \bar{X} \sim N\left(1900, \frac{800}{\sqrt{500}}\right) \right.$ $\rightarrow 35.7$



$$z = \frac{2000 - 1900}{35.7} = 2.80$$

$$P(z > 2.80) \approx .003$$

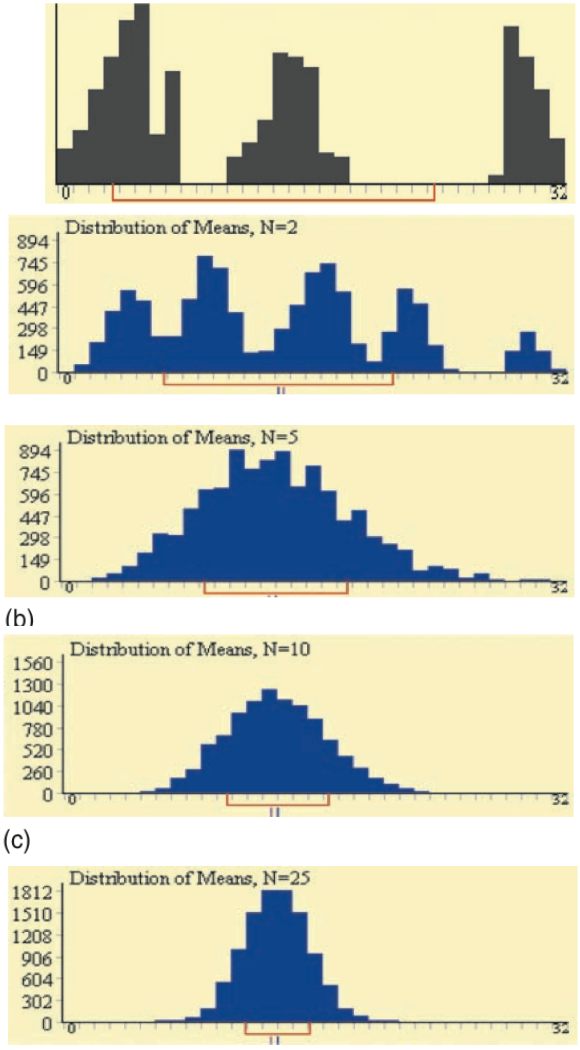
$\Rightarrow$ Budget is probably safe.

The value of coupons used by a shopper follow a Normal distn with mean 50¢ & std dev 10¢.

a) if he uses 10 coupons one week, what is the probability that he saves more than $7.00?

b) What is the probability that the average value of 1,000 used coupons is greater than 51¢?

A Few More Facts

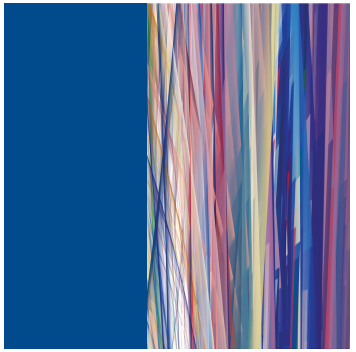


Any linear combination of independent Normal random variables is also Normally distributed.

More generally, the central limit theorem notes that the distribution of a sum or average of many small random quantities is close to Normal.

Finally, the central limit theorem also applies to discrete random variables.

5.2 Sampling Distributions for Counts and Proportions



- Binomial Distributions for Sample Counts
- Binomial Distributions in Statistical Sampling
- Finding Binomial Probabilities
- Binomial Mean and Standard Deviation
- Sample Proportions
- Normal Approximation for Counts and Proportions
- Binomial Formula

The Binomial Setting

When a random phenomenon is repeated or observed several times, we are often interested in the number of times a particular outcome occurs. Think about tossing a coin *n* times, where each toss is either a H or T.

A **binomial setting** arises when we perform several independent "trials", each with two possible outcomes: "Success" and "Failure". The four requirements for a binomial setting are:

- **B**inary? The outcomes of each trial can be labeled "Success" or "Failure."
- **I**ndependent? Trials must be **independent;** that is, knowing the result of one trial must not have any effect on the result of any other trial.
- **N**umber? The number of trials *n* must be fixed in advance (ie, n is not random).
- **S**uccess? On each trial, the probability *p* of success must be the same.

Binomial Distribution

Consider tossing a coin *n* times. Each toss gives either heads or tails. Knowing the outcome of one toss does not change the probability of an outcome on any other toss. If we define heads as a success, then *p* is the probability of a head and is 0.5 on any toss. Thus, we have a binomial setting.

Let the random variable *X* be the number of heads in those *n* tosses. The probability distribution of *X* is called a **binomial distribution.**

Binomial Distribution

The count *X* of successes in a binomial setting has the **binomial distribution with parameters n and p**, where *n* is the number of trials and *p* is the probability of a success on any one trial. The possible values of *X* are the integers from 0 to *n*. That is, S = {0, 1, 2, …, n}.

Note: Not all counts have binomial distributions; be sure to check the conditions for a binomial setting and make sure you're being asked to count the number of successes in a fixed number of trials!

Binomial Distributions in Statistical Sampling

The binomial distributions are important in statistics when we want to make inferences about the proportion *p* of successes in a population.

Suppose 10% of CDs have defective copy-protection schemes that can harm computers. A music distributor inspects an SRS of 10 CDs from a shipment of 10,000. Let *X* = number of defective CDs.

What is *P*(*X* = 0)? Note: This is not quite a binomial setting. Why?

tt

The actual probability is $P(\text{no defectives}) = \frac{9000}{10000} \cdot \frac{8999}{9999} \cdot \frac{8998}{9998} \cdot \ldots \cdot \frac{8991}{9991} = 0.3485$

Sampling Distribution of a Count

Choose an SRS of size *n* from a population with proportion *p* of successes. When the population is much larger than the sample, the count *X* of successes in the sample has approximately the binomial distribution with parameters *n* and *p*.

Using the binomial distribution, $P(X = 0) = \binom{10}{0}(0.10)^0(0.90)^{10} = 0.3487$

Binomial Mean and Standard Deviation

Mean and Standard Deviation of a Binomial Random Variable

If a count X has a binomial distribution with parameters n and p, the **mean** and **standard deviation** of X are:

$$\mu_X = np$$

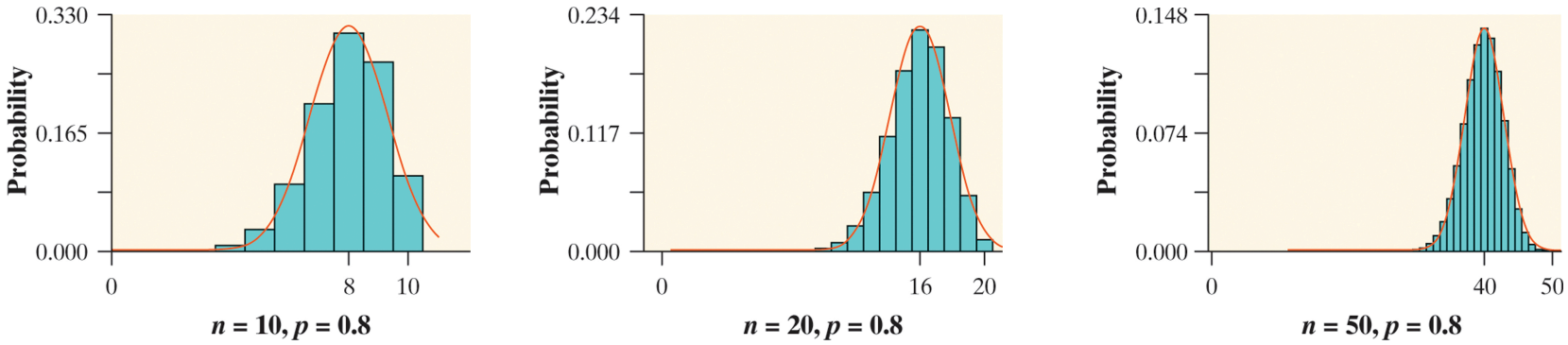
$$\sigma_X = \sqrt{np(1-p)}$$

Note: These formulas work ONLY for binomial distributions. They can't be used for other distributions!

Normal Approximation for Binomial Distributions

As *n* gets larger, something interesting happens to the shape of a binomial distribution.

$X \sim B(n, p)$ Bin b bin



Normal Approximation for Binomial Distributions

Suppose that *X* has the binomial distribution with *n* trials and success probability *p*. When *n* is large, the distribution of *X* is approximately Normal with mean and standard deviation

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

As a rule of thumb, we will use the Normal approximation when *n* is so large that $np \geq 10$ and $n(1-p) \geq 10$.

Example

Sample surveys show that fewer people enjoy shopping than in the past. A survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming." Suppose that exactly 60% of all adult U.S. residents would say "Agree" if asked the same question. Let X = the number in the sample who agree. **Estimate the probability that 1520 or more of the sample agree.**

Approximate; calculate

1) Verify that X is approximately a binomial random variable.

B: Success = agree, Failure = don't agree
I: Because the population of U.S. adults is greater than 25,000, it is reasonable to assume the sampling without replacement condition is met.
N: n = 2500 trials of the chance process.
S: The probability of selecting an adult who agrees is p = 0.60.

2) Check the conditions for using a Normal approximation.

Since $np = 2500(0.60) = 1500$ and $n(1-p) = 2500(0.40) = 1000$ are both at least 10, we may use the Normal approximation.

3) Calculate $P(X \geq 1520)$ using a Normal approximation.

$$\mu = np = 2500(0.60) = 1500$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{2500(0.60)(0.40)} = 24.49$$

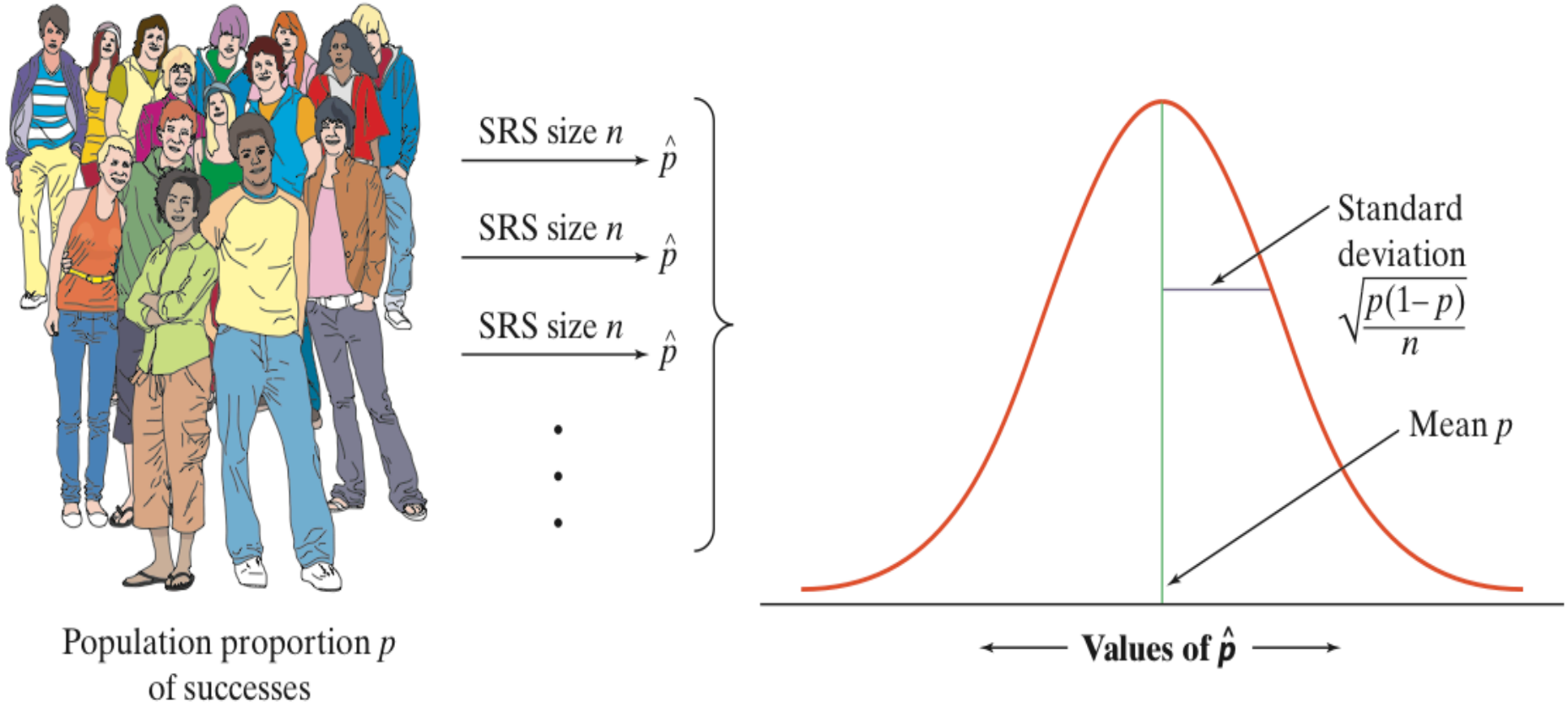
$$z = \frac{1520 - 1500}{24.49} = 0.82$$

$$P(X \geq 1520) = P(Z \geq 0.82) = 1 - 0.7939 = 0.2061$$

Sampling Distribution of a Sample Proportion

There is an important connection between the sample proportion $\hat{p}$ and the number of "successes" X in the sample.

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$



Binomial Formula

We can find a formula for the probability that a binomial random variable takes any value by adding probabilities for the different ways of getting exactly that many successes in *n* observations.

The number of ways of arranging *k* successes among *n* observations is given by the **binomial coefficient**

"n choose k"

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, 2, \ldots, n$.

Note: $n! = n(n-1)(n-2)\bullet\ldots\bullet(3)(2)(1)$

and $0! = 1$.

Binomial Probability

The binomial coefficient counts the number of different ways in which *k* successes can be arranged among *n* trials. The binomial probability $P(X = k)$ is this count multiplied by the probability of any one specific arrangement of the *k* successes.

Binomial Probability

If *X* has the binomial distribution with *n* trials and probability *p* of success on each trial, the possible values of *X* are 0, 1, 2, …, *n*. If *k* is any one of these values,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- $\binom{n}{k}$: Number of arrangements of *k* successes
- p^k: Probability of *k* successes
- $(1-p)^{n-k}$: Probability of *n-k* failures

Example

Each child of a particular pair of parents has probability 0.25 of having blood type O. Suppose the parents have five children.

(a) Find the probability that exactly three of the children have type O blood.

Let X = the number of children with type O blood. We know X has a binomial distribution with $n = 5$ and $p = 0.25$.

$$P(X=3)=\binom{5}{3}(0.25)^3(0.75)^2=10(0.25)^3(0.75)^2=0.08789$$

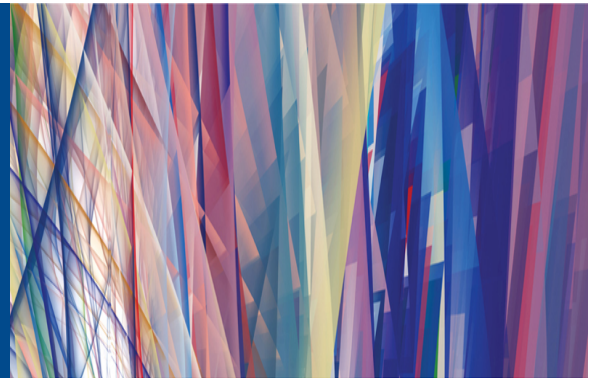
$\binom{5}{3}=\frac{5!}{3!2!}=\frac{5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)(2\cdot1)}$

(b) Should the parents be surprised if more than three of their children have type O blood?

$$P(X>3)=P(X=4)+P(X=5)$$
$$=\binom{5}{4}(0.25)^4(0.75)^1+\binom{5}{5}(0.25)^5(0.75)^0$$
$$=5(0.25)^4(0.75)^1+1(0.25)^5(0.75)^0$$
$$=0.01465+0.00098=0.01563$$

Since there is only a 1.5% chance that more than three children out of five would have Type O blood, the parents should be surprised!

Chapter 5
Sampling Distributions



5.1 The Sampling Distribution of a Sample Mean

5.2 Sampling Distributions for Counts and Proportions