Chapter 6
Random Variables

- **6.1 Discrete and Continuous Random Variables**
- **6.2 Transforming and Combining Random Variables**
- **6.3 Binomial and Geometric Random Variables**

Random Variable and Probability Distribution

A **probability model** describes the possible outcomes of a chance process and the likelihood that those outcomes will occur.

A numerical variable that describes the outcomes of a chance process is called a **random variable**. The probability model for a random variable is its probability distribution

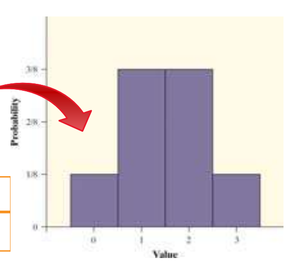
Definition:

A **random variable** takes numerical values that describe the outcomes of some chance process. The **probability distribution** of a random variable gives its possible values and their probabilities.

Example: Consider tossing a fair coin 3 times.
Define X = the number of heads obtained

X = 0: TTT
X = 1: HTT THT TTH
X = 2: HHT HTH THH
X = 3: HHH

Value	0	1	2	3
Probability	1/8	3/8	3/8	1/8



Discrete Random Variables

There are two main types of random variables: *discrete* and *continuous*. If we can find a way to list all possible outcomes for a random variable and assign probabilities to each one, we have a **discrete random variable**.

Discrete Random Variables and Their Probability Distributions

A **discrete random variable** X takes a fixed set of possible values with gaps between. The probability distribution of a discrete random variable X lists the values x_i and their probabilities p_i:

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

The probabilities p_i must satisfy two requirements:

1. Every probability p_i is a number between 0 and 1.
2. The sum of the probabilities is 1.

To find the probability of any event, add the probabilities p_i of the particular values x_i that make up the event.

Example: Babies' Health at Birth

Read the example on page 343.

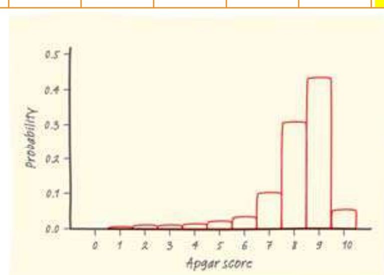
(a) **Show that the probability distribution for X is legitimate.**

(b) **Make a histogram of the probability distribution. Describe what you see.**

(c) **Apgar scores of 7 or higher indicate a healthy baby. What is $P(X \geq 7)$?**

Value:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

(a) All probabilities are between 0 and 1 and they add up to 1. This is a legitimate probability distribution.



(c) $P(X \geq 7) = .908$
We'd have a 91 % chance of randomly choosing a healthy baby.

(b) The left-skewed shape of the distribution suggests a randomly selected newborn will have an Apgar score at the high end of the scale. There is a small chance of getting a baby with a score of 5 or lower.

Mean of a Discrete Random Variable

When analyzing discrete random variables, we'll follow the same strategy we used with quantitative data – describe the shape, center, and spread, and identify any outliers.

The mean of any discrete random variable is an average of the possible outcomes, with each outcome weighted by its probability.

Definition:

Suppose that *X* is a discrete random variable whose probability distribution is

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

To find the **mean (expected value)** of *X*, multiply each possible value by its probability, then add all the products:

$$\mu_x = E(X) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots$$
$$= \sum x_i p_i$$

Example: Apgar Scores – What's Typical?

Consider the random variable *X* = Apgar Score

Compute the mean of the random variable *X* and interpret it in context.

Value:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

$$\mu_x = E(X) = \sum x_i p_i$$
$$= (0)(0.001) + (1)(0.006) + (2)(0.007) + \ldots + (10)(0.053)$$
$$= 8.128$$

The mean Apgar score of a randomly selected newborn is 8.128. This is the long-term average Agar score of many, many randomly chosen babies.

Note: The expected value does not need to be a possible value of *X* or an integer! It is a long-term average over many repetitions.

Standard Deviation of a Discrete Random Variable

Since we use the mean as the measure of center for a discrete random variable, we'll use the standard deviation as our measure of spread. The definition of the **variance of a random variable** is similar to the definition of the variance for a set of quantitative data.

Definition:

Suppose that X is a discrete random variable whose probability distribution is

Value:	x_1	x_2	x_3	…
Probability:	p_1	p_2	p_3	…

and that μ_X is the mean of X. The **variance** of X is

$$Var(X) = \sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \ldots$$
$$= \sum (x_i - \mu_X)^2 p_i$$

To get the **standard deviation of a random variable**, take the square root of the variance.

Example: Apgar Scores – How Variable Are They?

Consider the random variable X = Apgar Score

Compute the standard deviation of the random variable X and interpret it in context.

Value:	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Probability:	0.001	0.006	0.007	0.008	0.012	0.020	0.038	0.099	0.319	0.437	0.053

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$
$$= (0-8.128)^2(0.001) + (1-8.128)^2(0.006) + \ldots + (10-8.128)^2(0.053)$$
$$= 2.066$$

Variance

$$\sigma_X = \sqrt{2.066} = 1.437$$

The standard deviation of X is 1.437. On average, a randomly selected baby's Apgar score will differ from the mean 8.128 by about 1.4 units.

Continuous Random Variables

Discrete random variables commonly arise from situations that involve counting something. Situations that involve measuring something often result in a **continuous random variable**.

Definition:

A **continuous random variable** *X* takes on all values in an interval of numbers. The probability distribution of *X* is described by a **density curve**. The probability of any event is the area under the density curve and above the values of *X* that make up the event.

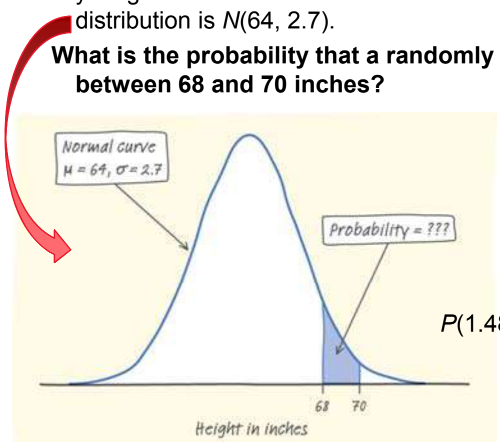
The probability model of a discrete random variable *X* assigns a probability between 0 and 1 to each possible value of *X*.

A continuous random variable *Y* has *infinitely many* possible values. All continuous probability models assign probability 0 to every individual outcome. Only *intervals* of values have positive probability.

Example: Young Women's Heights

Read the example on page 351. Define *Y* as the height of a randomly chosen young woman. *Y* is a continuous random variable whose probability distribution is *N*(64, 2.7).

What is the probability that a randomly chosen young woman has height between 68 and 70 inches?



$P(68 \le Y \le 70)$ = ???

$$z = \frac{68-64}{2.7} = 1.48 \qquad z = \frac{70-64}{2.7} = 2.22$$

$P(1.48 \le Z \le 2.22) = P(Z \le 2.22) - P(Z \le 1.48)$

$= 0.9868 - 0.9306$

$= 0.0562$

There is about a 5.6% chance that a randomly chosen young woman has a height between 68 and 70 inches.

11

Section 6.1
Discrete and Continuous Random Variables

Summary

In this section, we learned that…

- ✓ A **random variable** is a variable taking numerical values determined by the outcome of a chance process. The **probability distribution** of a random variable *X* tells us what the possible values of *X* are and how probabilities are assigned to those values.
- ✓ A **discrete random variable** has a fixed set of possible values with gaps between them. The probability distribution assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1.
- ✓ A **continuous random variable** takes all values in some interval of numbers. A density curve describes the probability distribution of a continuous random variable.

12

Section 6.1
Discrete and Continuous Random Variables

Summary

In this section, we learned that…

- ✓ The **mean of a random variable** is the long-run average value of the variable after many repetitions of the chance process. It is also known as the **expected value** of the random variable.
- ✓ The expected value of a discrete random variable *X* is

$$\mu_x = \sum x_i p_i = x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots$$

- ✓ The **variance of a random variable** is the average squared deviation of the values of the variable from their mean. The **standard deviation** is the square root of the variance. For a discrete random variable *X*,

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + (x_3 - \mu_X)^2 p_3 + \ldots$$

Section 6.2
Transforming and Combining Random Variables

Learning Objectives

After this section, you should be able to…

- ✓ DESCRIBE the effect of performing a linear transformation on a random variable
- ✓ COMBINE random variables and CALCULATE the resulting mean and standard deviation
- ✓ CALCULATE and INTERPRET probabilities involving combinations of Normal random variables

Linear Transformations

In Section 6.1, we learned that the mean and standard deviation give us important information about a random variable. In this section, we'll learn how the mean and standard deviation are affected by transformations on random variables.

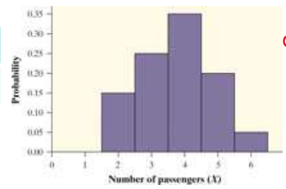
In Chapter 2, we studied the effects of linear transformations on the shape, center, and spread of a distribution of data. Recall:

1. ***Adding (or subtracting) a constant, a, to each observation:***
   - Adds *a* to measures of center and location.
   - Does not change the shape or measures of spread.
2. ***Multiplying (or dividing) each observation by a constant, b:***
   - Multiplies (divides) measures of center and location by *b*.
   - Multiplies (divides) measures of spread by |b|.
   - Does not change the shape of the distribution.

Linear Transformations

Pete's Jeep Tours offers a popular half-day trip in a tourist area. There must be at least 2 passengers for the trip to run, and the vehicle will hold up to 6 passengers. Define *X* as the number of passengers on a randomly selected day.

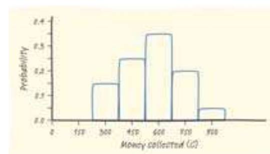
Passengers x_i	2	3	4	5	6
Probability p_i	0.15	0.25	0.35	0.20	0.05



The mean of *X* is 3.75 and the standard deviation is 1.090.

Pete charges $150 per passenger. The random variable *C* describes the amount Pete collects on a randomly selected day.

Collected c_i	300	450	600	750	900
Probability p_i	0.15	0.25	0.35	0.20	0.05



The mean of *C* is $562.50 and the standard deviation is $163.50.

Compare the shape, center, and spread of the two probability distributions.

Linear Transformations

How does multiplying or dividing by a constant affect a random variable?

Effect on a Random Variable of Multiplying (Dividing) by a Constant

Multiplying (or dividing) each value of a random variable by a number *b*:

- Multiplies (divides) measures of center and location (mean, median, quartiles, percentiles) by *b.*
- Multiplies (divides) measures of spread (range, *IQR*, standard deviation) by |*b*|.
- Does not change the shape of the distribution.

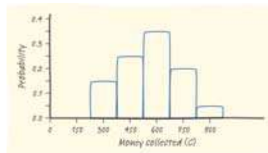
Note: Multiplying a random variable by a constant *b* multiplies the variance by b^2.

Linear Transformations

Consider Pete's Jeep Tours again. We defined C as the amount of money Pete collects on a randomly selected day.

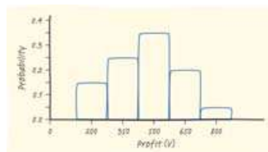
Collected c_i	300	450	600	750	900
Probability p_i	0.15	0.25	0.35	0.20	0.05

The mean of C is \$562.50 and the standard deviation is \$163.50.



It costs Pete \$100 per trip to buy permits, gas, and a ferry pass. The random variable V describes the profit Pete makes on a randomly selected day.

Profit v_i	200	350	500	650	800
Probability p_i	0.15	0.25	0.35	0.20	0.05



The mean of V is \$462.50 and the standard deviation is \$163.50.

Compare the shape, center, and spread of the two probability distributions.

Transforming and Combining Random Variables

Linear Transformations

How does adding or subtracting a constant affect a random variable?

Effect on a Random Variable of Adding (or Subtracting) a Constant

Adding the same number a (which could be negative) to each value of a random variable:

- Adds a to measures of center and location (mean, median, quartiles, percentiles).
- Does not change measures of spread (range, *IQR*, standard deviation).
- Does not change the shape of the distribution.

Transforming and Combining Random Variables

Linear Transformations

Whether we are dealing with data or random variables, the effects of a linear transformation are the same.

Effect on a Linear Transformation on the Mean and Standard Deviation

If $Y = a + bX$ is a linear transformation of the random variable X, then

- The probability distribution of Y has the same shape as the probability distribution of X.
- $\mu_Y = a + b\mu_X$.
- $\sigma_Y = |b|\sigma_X$ (since b could be a negative number).

Transforming and Combining Random Variables

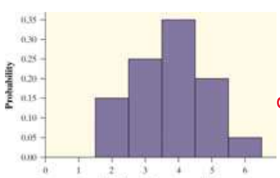
Combining Random Variables

So far, we have looked at settings that involve a single random variable. Many interesting statistics problems require us to examine two or more random variables.

Let's investigate the result of adding and subtracting random variables. Let X = the number of passengers on a randomly selected trip with Pete's Jeep Tours. Y = the number of passengers on a randomly selected trip with Erin's Adventures. Define $T = X + Y$. What are the mean and variance of T?

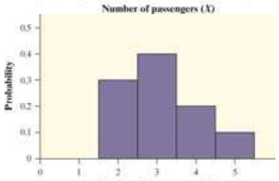
Passengers x_i	2	3	4	5	6
Probability p_i	0.15	0.25	0.35	0.20	0.05

Mean $\mu_X = 3.75$ Standard Deviation $\sigma_X = 1.090$



Passengers y_i	**2**	**3**	**4**	**5**
Probability p_i	0.3	0.4	0.2	0.1

Mean $\mu_Y = 3.10$ Standard Deviation $\sigma_Y = 0.943$



Transforming and Combining Random Variables

Combining Random Variables

Transforming and Combining Random Variables

How many total passengers can Pete and Erin expect on a randomly selected day?

Since Pete expects $\mu_X = 3.75$ and Erin expects $\mu_Y = 3.10$, they will average a total of 3.75 + 3.10 = 6.85 passengers per trip. We can generalize this result as follows:

Mean of the Sum of Random Variables

For any two random variables *X* and *Y*, if *T* = *X* + *Y*, then the expected value of *T* is

$$E(T) = \mu_T = \mu_X + \mu_Y$$

In general, the mean of the sum of several random variables is the sum of their means.

How much variability is there in the total number of passengers who go on Pete's and Erin's tours on a randomly selected day? To determine this, we need to find the probability distribution of *T*.

Combining Random Variables

Transforming and Combining Random Variables

The only way to determine the probability for any value of *T* is if *X* and *Y* are **independent random variables.**

Definition:

If knowing whether any event involving *X* alone has occurred tells us nothing about the occurrence of any event involving *Y* alone, and vice versa, then *X* and *Y* are **independent random variables.**

Probability models often assume independence when the random variables describe outcomes that appear unrelated to each other.

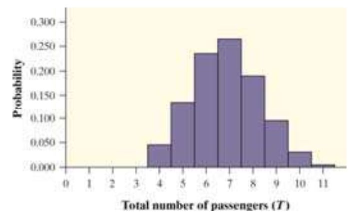
You should always ask whether the assumption of independence seems reasonable.

In our investigation, it is reasonable to assume *X* and *Y* are independent since the siblings operate their tours in different parts of the country.

Combining Random Variables

Let $T = X + Y$. Consider all possible combinations of the values of X and Y.

x_i	p_i	y_i	p_i	$t_i = x_i + y_i$	p_i
2	0.15	2	0.3	4	(0.15)(0.3) = 0.045
2	0.15	3	0.4	5	(0.15)(0.4) = 0.060
2	0.15	4	0.2	6	(0.15)(0.2) = 0.030
2	0.15	5	0.1	7	(0.15)(0.1) = 0.015
3	0.25	2	0.3	5	(0.25)(0.3) = 0.075
3	0.25	3	0.4	6	(0.25)(0.4) = 0.100
3	0.25	4	0.2	7	(0.25)(0.2) = 0.050
3	0.25	5	0.1	8	(0.25)(0.1) = 0.025
4	0.35	2	0.3	6	(0.35)(0.3) = 0.105
4	0.35	3	0.4	7	(0.35)(0.4) = 0.140
4	0.35	4	0.2	8	(0.35)(0.2) = 0.070
4	0.35	5	0.1	9	(0.35)(0.1) = 0.035
5	0.20	2	0.3	7	(0.20)(0.3) = 0.060
5	0.20	3	0.4	8	(0.20)(0.4) = 0.080
5	0.20	4	0.2	9	(0.20)(0.2) = 0.040
5	0.20	5	0.1	10	(0.20)(0.1) = 0.020
6	0.05	2	0.3	8	(0.05)(0.3) = 0.015
6	0.05	3	0.4	9	(0.05)(0.4) = 0.020
6	0.05	4	0.2	10	(0.05)(0.2) = 0.010
6	0.05	5	0.1	11	(0.05)(0.1) = 0.005



Recall: $\mu_T = \mu_X + \mu_Y = 6.85$

$$\sigma_T^2 = \sum (t_i - \mu_T)^2 p_i$$

$= (4 - 6.85)^2(0.045) + \ldots + (11 - 6.85)^2(0.005) = 2.0775$

Note: $\sigma_X^2 = 1.1875$ and $\sigma_Y^2 = 0.89$

What do you notice about the variance of T?

Combining Random Variables

As the preceding example illustrates, when we add two *independent* random variables, their variances add. *Standard deviations do not add*.

Variance of the Sum of Random Variables

For any two *independent* random variables X and Y, if $T = X + Y$, then the variance of T is

$$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$$

In general, the variance of the sum of several independent random variables is the sum of their variances.

Remember that you can add variances only if the two random variables are independent, and that you can NEVER add standard deviations!

Transforming and Combining Random Variables

Combining Random Variables

Transforming and Combining Random Variables

We can perform a similar investigation to determine what happens when we define a random variable as the difference of two random variables. In summary, we find the following:

Mean of the Difference of Random Variables

For any two random variables *X* and *Y*, if *D* = *X* - *Y*, then the expected value of *D* is

$$E(D) = \mu_D = \mu_X - \mu_Y$$

In general, the mean of the difference of several random variables is the difference of their means. *The order of subtraction is important!*

Variance of the Difference of Random Variables

For any two *independent* random variables *X* and *Y*, if *D* = *X* - *Y*, then the variance of *D* is

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$$

In general, the variance of the difference of two independent random variables is the sum of their variances.

Combining Normal Random Variables

Transforming and Combining Random Variables

So far, we have concentrated on finding rules for means and variances of random variables. If a random variable is Normally distributed, we can use its mean and standard deviation to compute probabilities.

An important fact about Normal random variables is that *any sum or difference of independent Normal random variables is also Normally distributed.*

Example

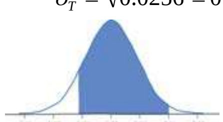
Mr. Starnes likes between 8.5 and 9 grams of sugar in his hot tea. Suppose the amount of sugar in a randomly selected packet follows a Normal distribution with mean 2.17 g and standard deviation 0.08 g. If Mr. Starnes selects 4 packets at random, what is the probability his tea will taste right?

Let *X* = the amount of sugar in a randomly selected packet.
Then, $T = X_1 + X_2 + X_3 + X_4$. We want to find $P(8.5 \le T \le 9)$.

$$\mu_T = \mu_{X1} + \mu_{X2} + \mu_{X3} + \mu_{X4} = 2.17 + 2.17 + 2.17 + 2.17 = 8.68$$

$$\sigma_T^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \sigma_{X_3}^2 + \sigma_{X_4}^2 = (0.08)^2 + (0.08)^2 + (0.08)^2 + (0.08)^2 = 0.0256$$

$$\sigma_T = \sqrt{0.0256} = 0.16$$



$$z = \frac{8.5 - 8.68}{0.16} = -1.13 \quad \text{and} \quad z = \frac{9 - 8.68}{0.16} = 2.00$$

P(-1.13 ≤ *Z* ≤ 2.00) = 0.9772 – 0.1292 = 0.8480
There is about an 85% chance Mr. Starnes's tea will taste right.

Section 6.2
Transforming and Combining Random Variables

Summary

In this section, we learned that…

- ✓ Adding a constant *a* (which could be negative) to a random variable increases (or decreases) the mean of the random variable by *a* but does not affect its standard deviation or the shape of its probability distribution.
- ✓ Multiplying a random variable by a constant *b* (which could be negative) multiplies the mean of the random variable by *b* and the standard deviation by |*b*| but does not change the shape of its probability distribution.
- ✓ A **linear transformation** of a random variable involves adding a constant *a*, multiplying by a constant *b*, or both. If we write the linear transformation of *X* in the form $Y = a + bX$, the following about are true about *Y*:
  - ✓ **Shape:** same as the probability distribution of *X*.
  - ✓ **Center:** $\mu_Y = a + b\mu_X$
  - ✓ **Spread:** $\sigma_Y = |b|\sigma_X$

Section 6.2
Transforming and Combining Random Variables

Summary

In this section, we learned that…

- ✓ If *X* and *Y* are any two random variables,

$$\mu_{X \pm Y} = \mu_X \pm \mu_Y$$

- ✓ If *X* and *Y* are **independent random variables**

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$$

- ✓ The sum or difference of independent Normal random variables follows a Normal distribution.

Section 6.3
Binomial and Geometric Random Variables

Learning Objectives

After this section, you should be able to…

- ✓ DETERMINE whether the conditions for a binomial setting are met
- ✓ COMPUTE and INTERPRET probabilities involving binomial random variables
- ✓ CALCULATE the mean and standard deviation of a binomial random variable and INTERPRET these values in context
- ✓ CALCULATE probabilities involving geometric random variables

Binomial Settings

Binomial and Geometric Random Variables

When the same chance process is repeated several times, we are often interested in whether a particular outcome does or doesn't happen on each repetition. In some cases, the number of repeated trials is fixed in advance and we are interested in the number of times a particular event (called a "success") occurs. If the trials in these cases are *independent* and each success has an equal chance of occurring, we have a **binomial setting.**

Definition:

A **binomial setting** arises when we perform several independent trials of the same chance process and record the number of times that a particular outcome occurs. The four conditions for a binomial setting are

B • **B**inary? The possible outcomes of each trial can be classified as "success" or "failure."

I • **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 • **N**umber? The number of trials n of the chance process must be fixed in advance.

S • **S**uccess? On each trial, the probability p of success must be the same.

Binomial Random Variable

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.

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

Definition:

The count X of successes in a binomial setting is a **binomial random variable.** The probability distribution of X is a **binomial distribution** with parameters n and p, where n is the number of trials of the chance process and p is the probability of a success on any one trial. The possible values of X are the whole numbers from 0 to n.

***Note*: When checking the Binomial condition, be sure to check the BINS and make sure you're being asked to count the number of successes in a certain number of trials!**

Binomial and Geometric Random Variables

Binomial Probabilities

In a binomial setting, we can define a random variable (say, X) as the number of successes in n independent trials. We are interested in finding the probability distribution of X.

Example

Each child of a particular pair of parents has probability 0.25 of having type O blood. Genetics says that children receive genes from each of their parents independently. If these parents have 5 children, the count X of children with type O blood is a binomial random variable with n = 5 trials and probability p = 0.25 of a success on each trial. In this setting, a child with type O blood is a "success" (S) and a child with another blood type is a "failure" (F). What's $P(X = 2)$?

P(SSFFF) = (0.25)(0.25)(0.75)(0.75)(0.75) = $(0.25)^2(0.75)^3$ = 0.02637

However, there are a number of different arrangements in which 2 out of the 5 children have type O blood:

SSFFF	SFSFF	SFFSF	SFFFS	FSSFF
FSFSF	FSFFS	FFSSF	FFSFS	FFFSS

Verify that in each arrangement, $P(X = 2) = (0.25)^2(0.75)^3 = 0.02637$

Therefore, $P(X = 2) = 10(0.25)^2(0.75)^3 = 0.2637$

Binomial and Geometric Random Variables

Binomial Coefficient

Note, in the previous example, *any one arrangement* of 2 S's and 3 F's had the same probability. This is true because no matter what arrangement, we'd multiply together 0.25 twice and 0.75 three times.

We can generalize this for any setting in which we are interested in k successes in n trials. That is,

$$P(X = k) = P(\text{exactly } k \text{ successes in } n \text{ trials})$$
$$= \text{number of arrangements} \cdot p^k (1-p)^{n-k}$$

Definition:

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

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

for k = 0, 1, 2, …, n where

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

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,

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

Number of arrangements of k successes

Probability of k successes

Probability of $n-k$ failures

Binomial and Geometric Random Variables

Example: Inheriting Blood Type

Each child of a particular pair of parents has probability 0.25 of having blood type O. Suppose the parents have 5 children

(a) Find the probability that exactly 3 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.

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

(b) Should the parents be surprised if more than 3 of their children have type O blood?

To answer this, we need to find $P(X > 3)$.

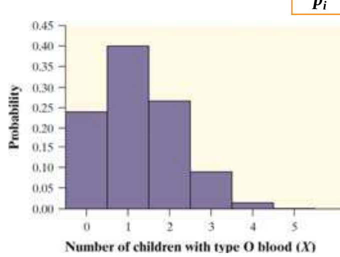
$$P(X>3)=P(X=4)+P(X=5)$$
$$=\binom{5}{4}(0.25)^4(0.75)^1+\binom{5}{5}(0.25)^5(0.75)^0$$
$$=5(0.25)^4(0.75)^1+1(0.25)^5(0.75)^0$$
$$=0.01465+0.00098=0.01563$$

Since there is only a 1.5% chance that more than 3 children out of 5 would have Type O blood, the parents should be surprised!

Mean and Standard Deviation of a Binomial Distribution

We describe the probability distribution of a binomial random variable just like any other distribution – by looking at the shape, center, and spread. Consider the probability distribution of X = number of children with type O blood in a family with 5 children.

x_i	0	1	2	3	4	5
p_i	0.2373	0.3955	0.2637	0.0879	0.0147	0.00098



Shape: The probability distribution of X is skewed to the right. It is more likely to have 0, 1, or 2 children with type O blood than a larger value.

Center: The median number of children with type O blood is 1. Based on our formula for the mean:

$$\mu_X=\sum x_i p_i=(0)(0.2373)+1(0.39551)+...+(5)(0.00098)$$
$$=1.25$$

Spread: The variance of X is $\sigma_X^2=\sum(x_i-\mu_X)^2 p_i=(0-1.25)^2(0.2373)+(1-1.25)^2(0.3955)+...+(5-1.25)^2(0.00098)=0.9375$

The standard deviation of X is $\sigma_X=\sqrt{0.9375}=0.968$

Binomial and Geometric Random Variables

Mean and Standard Deviation of a Binomial Distribution

Binomial and Geometric Random Variables

Notice, the mean μ_X = 1.25 can be found another way. Since each child has a 0.25 chance of inheriting type O blood, we'd expect one-fourth of the 5 children to have this blood type. That is, μ_X = 5(0.25) = 1.25. This method can be used to find the mean of any binomial random variable with parameters *n* and *p*.

Mean and Standard Deviation of a Binomial Random Variable

If a count *X* has the binomial distribution with number of trials *n* and probability of success *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!**

Example: Bottled Water versus Tap Water

Mr. Bullard's 21 AP Statistics students did the Activity on page 340. If we assume the students in his class *cannot* tell tap water from bottled water, then each has a 1/3 chance of correctly identifying the different type of water by guessing. Let *X* = the number of students who correctly identify the cup containing the different type of water.

Find the mean and standard deviation of *X*.

Since *X* is a binomial random variable with parameters *n* = 21 and *p* = 1/3, we can use the formulas for the mean and standard deviation of a binomial random variable.

$$\mu_X = np = 21(1/3) = 7$$

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{21(1/3)(2/3)} = 2.16$$

We'd expect about one-third of his 21 students, about 7, to guess correctly.

If the activity were repeated many times with groups of 21 students who were just guessing, the number of correct identifications would differ from 7 by an average of 2.16.

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?*

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

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

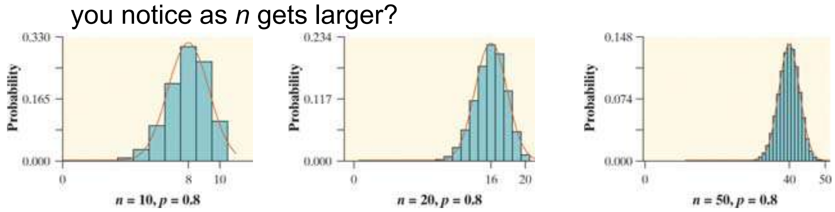
In practice, the binomial distribution gives a good approximation as long as we don't sample more than 10% of the population.

Sampling Without Replacement Condition

When taking an SRS of size *n* from a population of size *N*, we can use a binomial distribution to model the count of successes in the sample as long as $n \le \frac{1}{10}N$

Normal Approximation for Binomial Distributions

As *n* gets larger, something interesting happens to the shape of a binomial distribution. The figures below show histograms of binomial distributions for different values of *n* and *p*. What do you notice as *n* gets larger?



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 $np \ge 10$ and $n(1-p) \ge 10$. That is, the expected number of successes and failures are both at least 10.

Example: Attitudes Toward Shopping

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 US 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.**

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.

$$\mu = np = 2500(0.60) = 1500$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{2500(0.60)(0.40)} = 24.49$$

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

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

Geometric Settings

In a binomial setting, the number of trials n is fixed and the binomial random variable X counts the number of successes. In other situations, the goal is to repeat a chance behavior *until a success occurs*. These situations are called **geometric settings.**

Definition:
A **geometric setting** arises when we perform independent trials of the same chance process and record the number of trials until a particular outcome occurs. The four conditions for a geometric setting are

B • **B**inary? The possible outcomes of each trial can be classified as "success" or "failure."

I • **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.

T • **T**rials? The goal is to count the number of trials until the first success occurs.

S • **S**uccess? On each trial, the probability p of success must be the same.

Geometric Random Variable

Binomial and Geometric Random Variables

In a geometric setting, if we define the random variable *Y* to be the number of trials needed to get the first success, then *Y* is called a **geometric random variable**. The probability distribution of *Y* is called a **geometric distribution.**

Definition:

The number of trials *Y* that it takes to get a success in a geometric setting is a **geometric random variable.** The probability distribution of *Y* is a **geometric distribution** with parameter *p*, the probability of a success on any trial. The possible values of *Y* are 1, 2, 3, ….

***Note*: Like binomial random variables, it is important to be able to distinguish situations in which the geometric distribution does and doesn't apply!**

Example: The Birthday Game

Read the activity on page 398. The random variable of interest in this game is *Y* = the number of guesses it takes to correctly identify the birth day of one of your teacher's friends. What is the probability the first student guesses correctly? The second? Third? What is the probability the k^{th} student guesses corrrectly?

Verify that *Y* is a geometric random variable.

B: Success = correct guess, Failure = incorrect guess
I: The result of one student's guess has no effect on the result of any other guess.
T: We're counting the number of guesses up to and including the first correct guess.
S: On each trial, the probability of a correct guess is 1/7.

Calculate *P*(*Y* = 1), *P*(*Y* = 2), *P*(*Y* = 3), and *P*(*Y* = k)

$$P(Y=1)=1/7$$
$$P(Y=2)=(6/7)(1/7)=0.1224$$
$$P(Y=3)=(6/7)(6/7)(1/7)=0.1050$$

Notice the pattern?

Geometric Probability

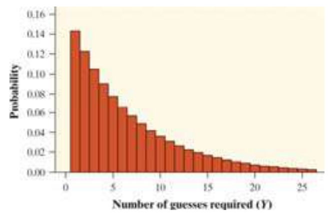
If *Y* has the geometric distribution with probability *p* of success on each trial, the possible values of *Y* are 1, 2, 3, … . If *k* is any one of these values,

$$P(Y=k)=(1-p)^{k-1}p$$

Mean of a Geometric Distribution

Binomial and Geometric Random Variables

The table below shows part of the probability distribution of *Y*. We can't show the entire distribution because the number of trials it takes to get the first success could be an incredibly large number.



y_i	1	2	3	4	5	6	…
p_i	0.143	0.122	0.105	0.090	0.077	0.066	

Shape: The heavily right-skewed shape is characteristic of any geometric distribution. That's because the most likely value is 1.

Center: The mean of *Y* is $\mu_Y = 7$. We'd expect it to take 7 guesses to get our first success.

Spread: The standard deviation of *Y* is $\sigma_Y = 6.48$. If the class played the Birth Day game many times, the number of homework problems the students receive would differ from 7 by an average of 6.48.

Mean (Expected Value) of Geometric Random Variable

If *Y* is a geometric random variable with probability *p* of success on each trial, then its **mean** (expected value) is $E(Y) = \mu_Y = 1/p$.

Section 6.3
Binomial and Geometric Random Variables

Summary

In this section, we learned that…

- ✓ A **binomial setting** consists of *n* independent trials of the same chance process, each resulting in a success or a failure, with probability of success *p* on each trial. The count *X* of successes is a **binomial random variable**. Its probability distribution is a **binomial distribution**.
- ✓ The **binomial coefficient** counts the number of ways *k* successes can be arranged among *n* trials.
- ✓ If *X* has the binomial distribution with parameters *n* and *p*, the possible values of *X* are the whole numbers 0, 1, 2, . . . , *n*. The binomial probability of observing *k* successes in *n* trials is

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

Section 6.3
Binomial and Geometric Random Variables

Summary

In this section, we learned that…

- ✓ The **mean** and **standard deviation** of a binomial random variable *X* are

$$\mu_X = np$$

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

- ✓ The **Normal approximation** to the binomial distribution says that if *X* is a count having the binomial distribution with parameters *n* and *p*, then when *n* is large, *X* is approximately Normally distributed. We will use this approximation when $np \geq 10$ and $n(1 - p) \geq 10$.

Section 6.3
Binomial and Geometric Random Variables

Summary

In this section, we learned that…

- ✓ A **geometric setting** consists of repeated trials of the same chance process in which each trial results in a success or a failure; trials are independent; each trial has the same probability *p* of success; and the goal is to count the number of trials until the first success occurs. If *Y* = the number of trials required to obtain the first success, then *Y* is a **geometric random variable**. Its probability distribution is called a **geometric distribution**.

- ✓ If *Y* has the geometric distribution with probability of success *p*, the possible values of *Y* are the positive integers 1, 2, 3, ... The **geometric probability** that *Y* takes any value is

$$P(Y = k) = (1-p)^{k-1}p$$

- ✓ The **mean** (expected value) of a geometric random variable *Y* is 1/p.