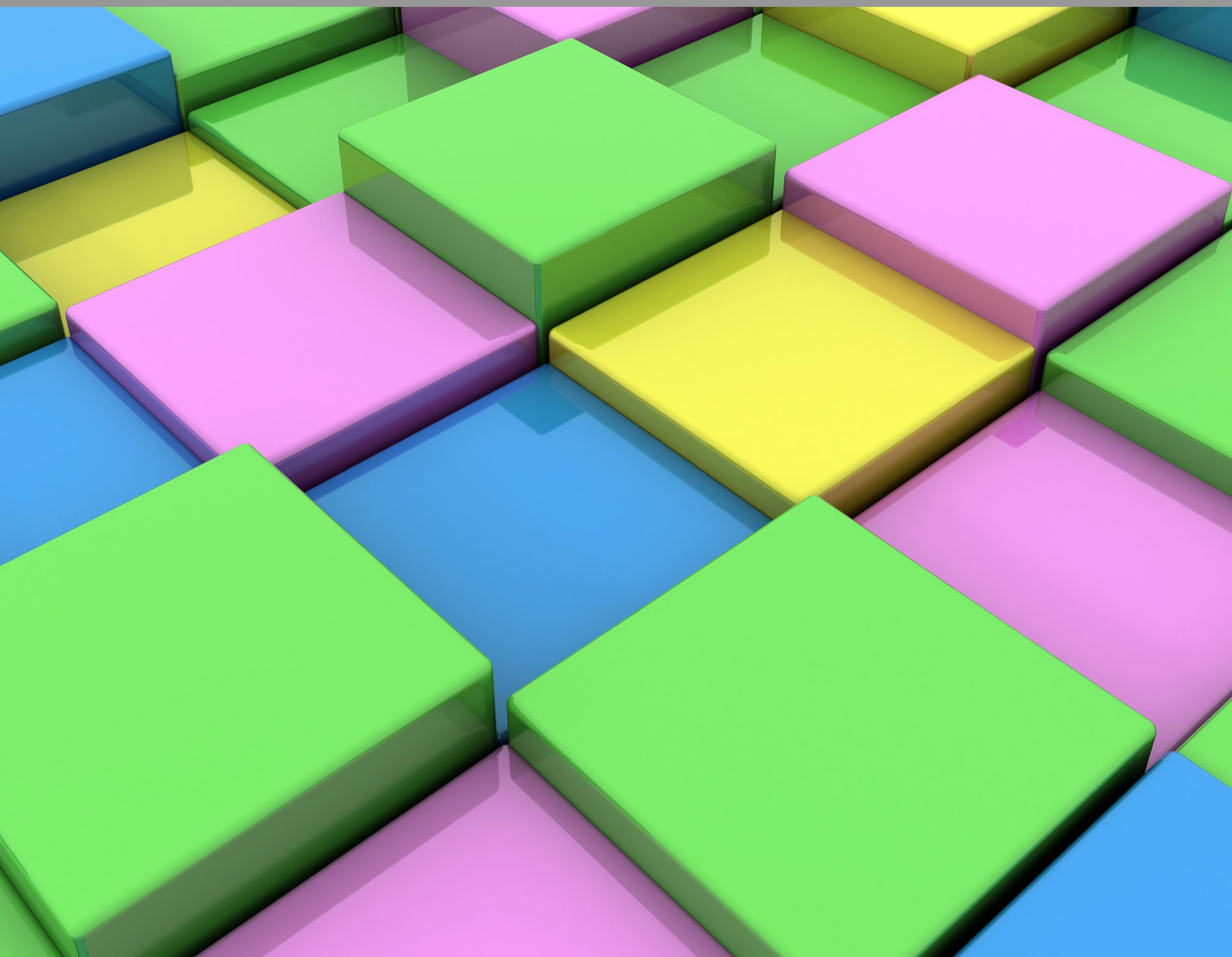


ck-12

flexbook  
next generation textbooks

# CK-12 Basic Probability and Statistics

## A Full Course



# Basic Probability and Statistics-A Full Course

---

Brenda Meery, (BrendaM)

**Say Thanks to the Authors**

Click <http://www.ck12.org/saythanks>

*(No sign in required)*



To access a customizable version of this book, as well as other interactive content, visit [www.ck12.org](http://www.ck12.org)

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Copyright © 2012 CK-12 Foundation, [www.ck12.org](http://www.ck12.org)

The names “CK-12” and “CK12” and associated logos and the terms “**FlexBook®**” and “**FlexBook Platform®**” (collectively “CK-12 Marks”) are trademarks and service marks of CK-12 Foundation and are protected by federal, state, and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections must include the referral attribution link <http://www.ck12.org/saythanks> (placed in a visible location) in addition to the following terms.

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the Creative Commons Attribution/Non-Commercial/Share Alike 3.0 Unported (CC BY-NC-SA) License (<http://creativecommons.org/licenses/by-nc-sa/3.0/>), as amended and updated by Creative Commons from time to time (the “CC License”), which is incorporated herein by this reference.

Complete terms can be found at <http://www.ck12.org/terms>.

Printed: June 22, 2012

**flexbook**  
next generation textbooks



## AUTHORS

Brenda Meery, (BrendaM)

# Contents

<b>1</b>	<b>Independent and Dependent Event</b>	<b>1</b>
1.1	Independent Events . . . . .	2
1.2	Dependent Events . . . . .	8
1.3	Mutually Inclusive and Mutually Exclusive Events . . . . .	10
1.4	Review Questions . . . . .	17
<b>2</b>	<b>The Next Step ... Conditional Probability</b>	<b>19</b>
2.1	What are Tree Diagrams? . . . . .	20
2.2	Order and Probability . . . . .	25
2.3	Conditional Probability . . . . .	34
2.4	Review Questions . . . . .	40
<b>3</b>	<b>Introduction to Discrete Random Variables</b>	<b>43</b>
3.1	What are Variables? . . . . .	44
3.2	The Probability Distribution . . . . .	46
3.3	A Glimpse at Binomial and Multinomial Distributions . . . . .	49
3.4	Using Technology to Find Probability Distributions . . . . .	55
3.5	Review Questions . . . . .	67
<b>4</b>	<b>Probability Distributions</b>	<b>70</b>
4.1	Normal Distributions . . . . .	71
4.2	Binomial Distributions . . . . .	76
4.3	Exponential Distributions . . . . .	86
4.4	Review Questions . . . . .	95
<b>5</b>	<b>Measures of Central Tendency</b>	<b>99</b>
5.1	The Mean . . . . .	100
5.2	The Median . . . . .	114
5.3	The Mode . . . . .	123
5.4	Review Questions . . . . .	126
<b>6</b>	<b>The Shape, Center and Spread of a Normal Distribution</b>	<b>136</b>
6.1	Estimating the Mean and Standard Deviation of a Normal Distribution . . . . .	138
6.2	Calculating the Standard Deviation . . . . .	141
6.3	Connecting the Standard Deviation and Normal Distribution . . . . .	153
6.4	Review Questions . . . . .	156
<b>7</b>	<b>Organizing and Displaying Distributions of Data</b>	<b>161</b>
7.1	Line Graphs and Scatter Plots . . . . .	162
7.2	Circle Graphs, Bar Graphs, Histograms, and Stem-and-Leaf Plots . . . . .	174
7.3	Box-and-Whisker Plots . . . . .	194
7.4	Review Questions . . . . .	204



<b>8</b>	<b>Organizing and Displaying Data for Comparison</b>	<b>217</b>
8.1	Review . . . . .	218
8.2	Double Line Graphs . . . . .	222
8.3	Two-sided Stem-and-Leaf Plots . . . . .	228
8.4	Double Bar Graphs . . . . .	231
8.5	Double Box-and-Whisker Plots . . . . .	234
8.6	Review Questions . . . . .	238

---

# CHAPTER 1 Independent and Dependent Event

## Chapter Outline

---

- 1.1 INDEPENDENT EVENTS
  - 1.2 DEPENDENT EVENTS
  - 1.3 MUTUALLY INCLUSIVE AND MUTUALLY EXCLUSIVE EVENTS
  - 1.4 REVIEW QUESTIONS
- 

### Introduction

Probability is present in many parts of our everyday lives. When you say it is raining and your numbers came up in the lottery, you are talking about 2 independent events. The fact that it is raining is not dependent on the fact that your numbers came up in the lottery, and vice versa. If you say that you have the flu and you are taking medicine, you are talking about dependent events. The terms independent and dependent in mathematics are the same as those found in the English language. In order to determine probabilities mathematically, we need to understand the differences between the definitions of independent and dependent. Independent events are those where the outcome of one event is not affected by the other, and dependent events are events where the outcome of one is affected by the other. Other terms, such as mutually inclusive and mutually exclusive, are also important. With mutually inclusive events, the concept of double counting is taken into account in the calculation of probabilities when using the Addition Principle.

# 1.1 Independent Events

## Learning Objectives

- Know the definition and the notion for independent events.
- Use the rules for addition, multiplication, and complementation to solve for probabilities of particular events in finite sample spaces.

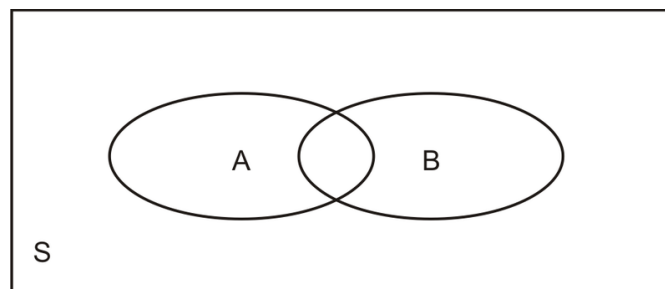
## What's in a Word?

The words *dependent* and *independent* are used by students and teachers on a daily basis. In fact, they are probably used quite frequently. You may tell your parent or guardian that you are independent enough to go to the movies on your own with your friends. You could say that when you bake a cake or make a cup of hot chocolate, the taste of these are dependent on what ingredients you use. In the English language, the term dependent means to be unable to do without, whereas independent means to be free from any outside influence.

What about in mathematics? What do the terms dependent and independent actually mean? This lesson will explore the mathematics of independence and dependence.

## What are Venn Diagrams and Why are They Used?

In **probability**, a **Venn diagram** is a graphic organizer that shows a visual representation for all possible **outcomes** of an experiment and the events of the experiment in ovals. Normally, in probability, the Venn diagram will be a box with overlapping ovals inside. Look at the diagram below:



The  $S$  represents all of the possible outcomes of an experiment. It is called the **sample space**. The ovals  $A$  and  $B$  represent the outcomes of the events that occur in the sample space. Let's look at an example.

Let's say our sample space is the numbers from 1 to 10. Event  $A$  will be the odd numbers from 1 to 10, and event  $B$  will be all the prime numbers. Remember that a prime number is a number where the only factors are 1 and itself. Now let's draw the Venn diagram to represent this example.

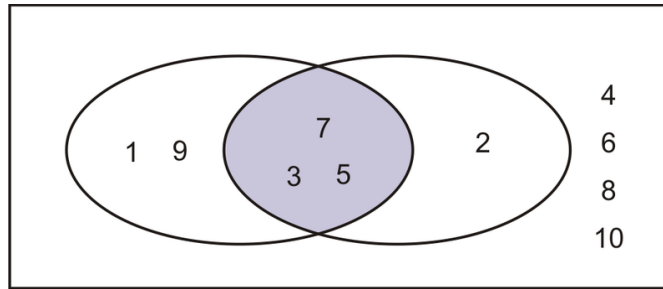
We know that:

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

$$A = \{1, 3, 5, 7, 9\}$$

$$B = \{1, 3, 5, 7\}$$

### 1.1. Independent Events



Notice that the prime numbers are part of both sets and are, therefore, in the overlapping part of the Venn diagram. The numbers 2, 4, 6, 8, and 10 are the numbers not part of  $A$  or  $B$ , but they are still members of the sample space. Now you try.

**Example 1**

2 coins are tossed. Event  $A$  consists of the outcomes when tossing heads on the first toss. Event  $B$  consists of the outcomes when tossing heads on the second toss. Draw a Venn diagram to represent this example.

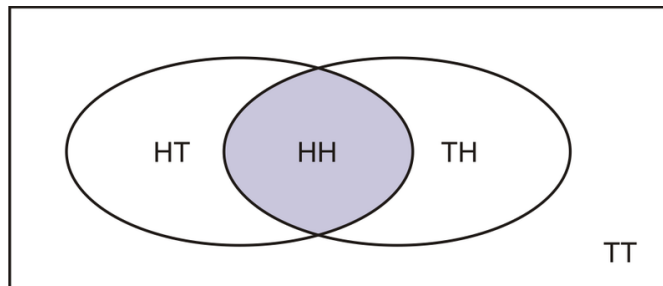
**Solution:**

We know that:

$$S = \{HH, HT, TH, TT\}$$

$$A = \{HH, HT\}$$

$$B = \{HH, TH\}$$



Notice that event  $A$  and event  $B$  share the Heads + Heads outcome and that the sample space contains Tails + Tails, which is neither in event  $A$  nor event  $B$ .

**Example 2**

In  $ABC$  High School, 30 percent of the students have a part-time job, and 25 percent of the students from the high school are on the honor roll. Event  $A$  represents the students holding a part-time job. Event  $B$  represents the students on the honor roll. Draw a Venn diagram to represent this example.

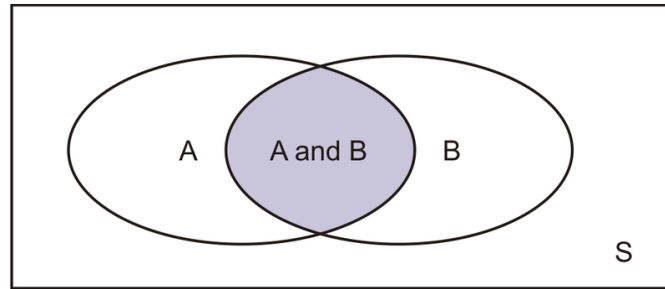
**Solution:**

We know that:

$$S = \text{students in } ABC \text{ High School}$$

$$A = \text{students holding a part-time job}$$

$$B = \text{students on the honor roll}$$



Notice that the overlapping oval for  $A$  and  $B$  represents the students who have a part-time job and are on the honor roll. The sample space,  $S$ , outside the ovals represents students neither holding a part-time job nor on the honor roll.

In a Venn diagram, when events  $A$  and  $B$  occur, the symbol used is  $\cap$ . Therefore,  $A \cap B$  is the intersection of events  $A$  and  $B$  and can be used to find the probability of both events occurring. If, in a Venn diagram, either  $A$  or  $B$  occurs, the symbol is  $\cup$ . This symbol would represent the union of events  $A$  and  $B$ , where the outcome would be in either  $A$  or  $B$ .

### Example 3

You are asked to roll a die. Event  $A$  is the event of rolling a 1, 2, or a 3. Event  $B$  is the event of rolling a 3, 4, or a 5. Draw a Venn diagram to represent this example. What is  $A \cap B$ ? What is  $A \cup B$ ?

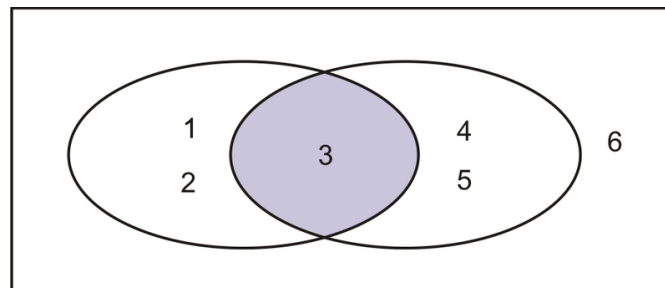
#### Solution:

We know that:

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$A = \{1, 2, 3\}$$

$$B = \{3, 4, 5\}$$



$$A \cap B = \{3\}$$

$$A \cup B = \{1, 2, 3, 4, 5\}$$

### Independent Events

In mathematics, the term independent means to have one event not dependent on the other. It is similar to the English definition. Suppose you are trying to convince your parent/guardian to let you go to the movies on your own. Your parent/guardian is thinking that if you go, you will not have time to finish your homework. For this reason, you have to convince him/her that you are independent enough to go to the movies *and* finish your homework. Therefore, you are trying to convince your parent/guardian that the 2 events, going to the movies and finishing your homework, are independent. This is similar to the mathematical definition. Say you were asked to pick a particular card from a deck of cards and roll a 6 on a die. It does not matter if you choose the card first and roll a 6 second, or vice versa. The probability of rolling the 6 would remain the same, as would the probability of choosing the card.

Going back to our Venn diagrams, **independent events** are represented as those events that occur in both sets. If we look just at Example 2, event  $A$  is a student holding a part-time job, and event  $B$  is the student being on the honor roll. These 2 events are independent of each other. In other words, whether you hold a part-time job is not dependent on

#### 1.1. Independent Events

your being on the honor roll, or vice versa. The outcome of one event is not dependent on the outcome of the second event. To calculate the probabilities, you would look at the overlapping part of the diagram.  $A$  and  $B$  represent the probability of both events occurring. Let's look at the probability calculation:

$$P(A) = 30\% \text{ or } 0.30$$

$$P(B) = 25\% \text{ or } 0.25$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$P(A \text{ and } B) = 0.30 \times 0.25$$

$$P(A \text{ and } B) = 0.075$$



In other words, 7.5% of the students of ABC high school are both on the honor roll and have a part-time job.

In Example 1, 2 coins are tossed. Remember that event  $A$  consists of the outcomes when getting heads on the first toss, and event  $B$  consists of the outcomes when getting heads on the second toss. What would be the probability of tossing the coins and getting a head on both the first coin and the second coin? We know that the probability of getting a head on a coin toss is  $\frac{1}{2}$ , or 50%. In other words, we have a 50% chance of getting a head on a toss of a fair coin and a 50% chance of getting a tail.

$$P(A) = 50\% \text{ or } 0.50$$

$$P(B) = 50\% \text{ or } 0.50$$

$$P(A \text{ and } B) = P(A) \times P(B)$$

$$P(A \text{ and } B) = 0.50 \times 0.50$$

$$P(A \text{ and } B) = 0.25$$

Therefore, there is a 25% chance of getting 2 heads when tossing 2 fair coins.

#### **Example 4**

2 cards are chosen from a deck of cards. The first card is replaced before choosing the second card. What is the probability that they both will be sevens?



**Solution:**

Let  $A = 1^{\text{st}}$  seven chosen.

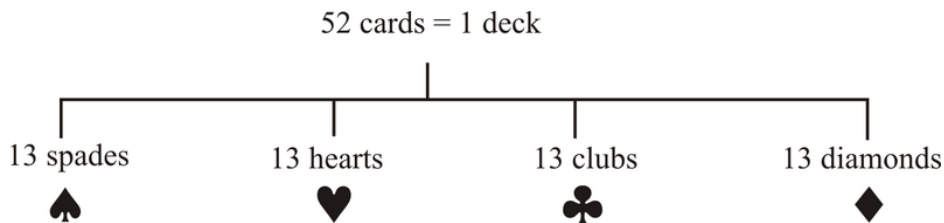
Let  $B = 2^{\text{nd}}$  seven chosen.

**A little note about a deck of cards**

A deck of cards consists of 52 cards.

Each deck has 4 parts (suits) with 13 cards in them.

Each suit has 3 face cards.



4 suits      1 seven per suit



The total number of sevens in the deck =  $4 \times 1 = 4$ .

Since the card *was* replaced, these events are independent:

$$P(A) = \frac{4}{52}$$

Note: The total number of cards is

$$P(B) = \frac{4}{52} \checkmark 52 \text{ after choosing the first card,}$$

because the first card is replaced.

$$P(A \text{ and } B) = \frac{4}{52} \times \frac{4}{52} \text{ or } P(A \cap B) = \frac{4}{52} \times \frac{4}{52}$$

$$P(A \cap B) = \frac{16}{2704}$$

$$P(A \cap B) = \frac{1}{169}$$

**Example 5**

The following table represents data collected from a grade 12 class in DEF High School.

**TABLE 1.1: Plans after High School**

Gender	University	Community College	Total
Males	28	56	84
Females	43	37	80
Total	71	93	164

Suppose 1 student was chosen at **random** from the grade 12 class.

- (a) What is the probability that the student is female?  
 (b) What is the probability that the student is going to university?

Now suppose 2 people both randomly chose 1 student from the grade 12 class. Assume that it's possible for them to choose the same student.

- (c) What is the probability that the first person chooses a student who is female and the second person chooses a student who is going to university?

**Solution:**

$$\begin{aligned} \text{Probabilities: } P(\text{female}) &= \frac{80}{164} \swarrow \boxed{164 \text{ total students}} \\ P(\text{female}) &= \frac{20}{41} \\ P(\text{going to university}) &= \frac{71}{164} \end{aligned}$$

$$\begin{aligned} P(\text{female}) \times P(\text{going to university}) &= \frac{20}{41} \times \frac{71}{164} \\ &= \frac{1420}{6724} \\ &= \frac{355}{1681} \\ &= 0.211 \end{aligned}$$

Therefore, there is a 21.1% probability that the first person chooses a student who is female and the second person chooses a student who is going to university.



## 1.2 Dependent Events

For 2 events to be dependent, the probability of the second event is dependent on the probability of the first event. In English, remember, the term dependent means to be unable to do without. This is similar to the mathematical definition of **dependent events**, where the second event is unable to happen without the first event occurring.

### Example 6

Remember in Example 4 when you were asked to determine the probability of drawing 2 sevens from a standard deck of cards? What would happen if 1 card is chosen and not replaced? In this case, the probability of drawing a seven on the second draw is dependent on drawing a seven on the first draw. Now let's calculate the probability of the 2 cards being drawn *without* replacement. This can be done with the **Multiplication Rule**.

### Solution:

Let  $A = 1^{\text{st}}$  seven chosen.

Let  $B = 2^{\text{nd}}$  seven chosen.

4 suits      1 seven per suit



The total number of sevens in the deck  $= 4 \times 1 = 4$ .

$$P(A) = \frac{4}{52}$$

Note: The total number of cards is

$$P(B) = \frac{3}{51} \quad \checkmark \text{ 51 after choosing the first card if it is not replaced.}$$

$$P(A \text{ and } B) = \frac{4}{52} \times \frac{3}{51} \text{ or } P(A \cap B) = \frac{4}{52} \times \frac{3}{51}$$

$$P(A \cap B) = \frac{12}{2652}$$

$$P(A \cap B) = \frac{1}{221}$$

Notice in this example that the numerator and denominator decreased from  $P(A)$  to  $P(B)$ . Once we picked the first card, the number of cards available from the deck dropped from 52 to 51. The number of sevens also decreased from 4 to 3. Again, the explanation given in the example was that the first card chosen was kept in your hand and not replaced into the deck before the second card was chosen.

### Example 7

#### 1.2. Dependent Events

A box contains 5 red marbles and 5 purple marbles. What is the probability of drawing 2 purple marbles and 1 red marble in succession *without replacement*?

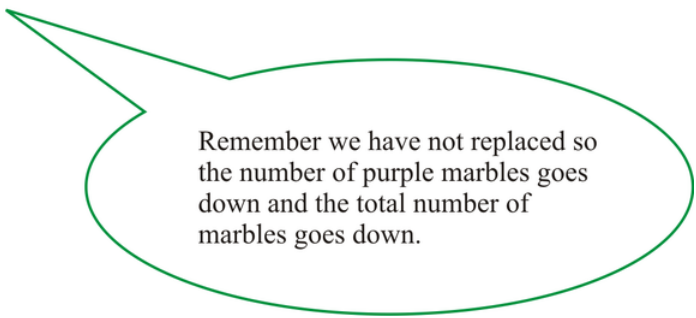
**Solution:**

On the first draw, the probability of drawing a purple marble is:

$$P_1(\text{purple}) = \frac{5}{10}$$

On the second draw the probability of drawing a purple marble is:

$$P_2(\text{purple}) = \frac{4}{9}$$



Remember we have not replaced so the number of purple marbles goes down and the total number of marbles goes down.

On the third draw, the probability of drawing a red marble is:

$$P(\text{red}) = \frac{5}{8}$$

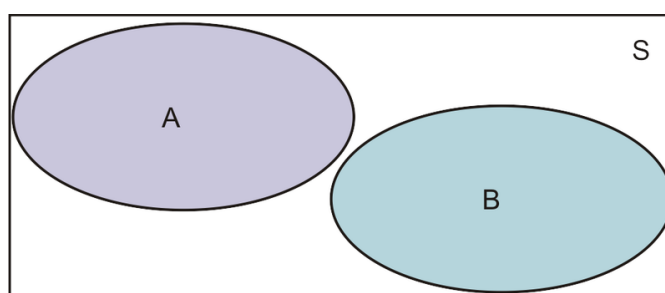
Therefore, the probability of drawing 2 purple marbles and 1 red marble is:

$$\begin{aligned} P(1 \text{ purple and 1 purple and 1 red}) &= P(1P \cap 1P \cap 1R) \\ &= P_1(\text{purple}) \times P_2(\text{purple}) \times P(\text{red}) \\ &= \frac{5}{10} \times \frac{4}{9} \times \frac{5}{8} \\ &= \frac{100}{720} \\ &= \frac{5}{36} \end{aligned}$$

## 1.3 Mutually Inclusive and Mutually Exclusive Events

When determining the probabilities of events, we must also look at 2 additional terms. These terms are mutually inclusive and mutually exclusive. When we add probability calculations of events described by these terms, we can apply the words *and* ( $\cap$ ) and *or* ( $\cup$ ) using the Addition Rule. Let's take another look at the Venn diagrams when defining these terms.

2 events  $A$  and  $B$  that cannot occur at the same time are **mutually exclusive events**. They have no common outcomes. See in the diagram below that  $P(A \text{ and } B) = 0$ . Notice that there is no intersection between the possible outcomes in event  $A$  and the possible outcomes in event  $B$ . For example, if you were asked to pick a number between 1 and 10, you cannot pick a number that is both even and odd. These events are mutually exclusive.



$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ and } B) = 0$$

To calculate the probability of picking a number from 1 to 10 that is even or picking a number from 1 to 10 that is odd, you would follow the steps below:

$$A = \{2, 4, 6, 8, 10\}$$

$$P(A) = \frac{5}{10}$$

$$B = \{1, 3, 5, 7, 9\}$$

$$P(B) = \frac{5}{10}$$

$$P(A \text{ or } B) = \frac{5}{10} + \frac{5}{10}$$

$$P(A \text{ or } B) = \frac{10}{10}$$

$$P(A \text{ or } B) = 1$$

The probability of picking a number from 1 to 10 that is even and picking a number from 1 to 10 that is odd would just be 0, since these are mutually exclusive events. In other words,  $P(A \text{ and } B) = 0$ .

If events  $A$  and  $B$  share some overlap in the Venn diagram, they may be considered not mutually exclusive events, but **mutually inclusive events**. Look at the diagrams below to see how these events can occur. Mutually inclusive events can occur at the same time. Say, for example, in the problem above, you wanted to pick a number from 1 to 10 that is less than 4 and pick an even number.

$$A = \{1, 2, 3\}$$

$$P(A) = \frac{3}{10}$$

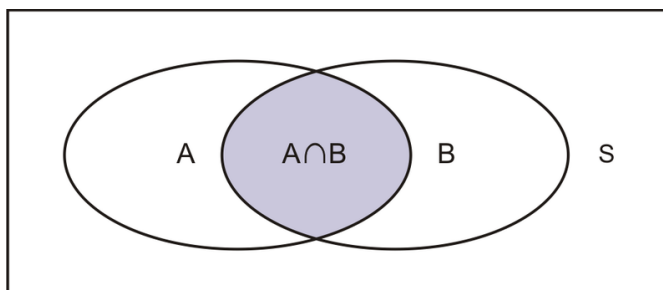
$$B = \{2, 4, 6, 8, 10\}$$

$$P(B) = \frac{5}{10}$$

$$P(A \text{ and } B) = \frac{1}{10}$$

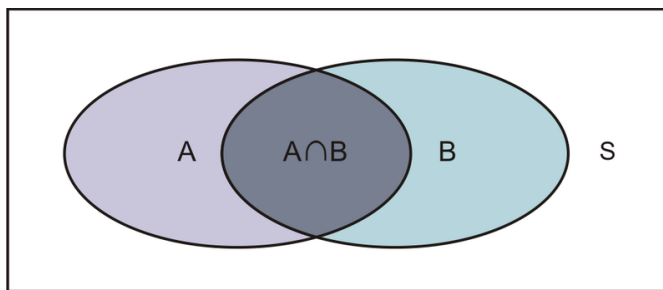
The reason why  $P(A \text{ and } B) = \frac{1}{10}$  is because there is only 1 number from 1 to 10 that is both less than 4 and even, and that number is 2.

When representing this on the Venn diagram, we would see something like the following:



A and B  
 $A \cap B$

Mutually exclusive events, remember, cannot occur at the same time. Mutually inclusive events can. Look at the Venn diagram below. What do you think we need to do in order to calculate the probability of  $A \cup B$  just from looking at this diagram?



A or B  
 $A \cup B$

If you look at the diagram, you see that the calculation involves not only  $P(A)$  and  $P(B)$ , but also  $P(A \cap B)$ . However, the items in  $A \cap B$  are also part of event  $A$  and event  $B$ . To represent the probability of  $A$  or  $B$ , we need to subtract the  $P(A \cap B)$ ; otherwise, we are double counting. In other words:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

or

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where  $\cap$  represents *and* and  $\cup$  represents *or*.

This is known as the **Addition Principle (Rule)**.

Addition Principle

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(A \cap B) = 0$  for mutually exclusive events.

Think about the idea of rolling a die. Event  $A$  is the sample set where the number rolled on the die is odd. Event  $B$  is the sample set where the number rolled is greater than 2.

$$\text{Event } A = \{1, 3, 5\}$$

$$\text{Event } B = \{3, 4, 5, 6\}$$

Notice that the events have 2 elements in common. Therefore, they are not mutually exclusive, but mutually inclusive.

What if we said that we are choosing a card from a deck of cards? Event  $A$  is the sample set where the card chosen is an 8. Event  $B$  is the sample set where the card chosen is an Ace ( $A$ ).

$$\text{Event } A = \{8\heartsuit, 8\diamonds, 8\clubsuit, 8\spadesuit\}$$

$$\text{Event } B = \{A\heartsuit, A\diamonds, A\clubsuit, A\spadesuit\}$$

Notice that the events have no elements in common. Therefore, they are mutually exclusive.

Look at the example below to understand the concept of double counting.

### **Example 8**

What is the probability of choosing a card from a deck of cards that is a club or a ten?

**Solution:**

$P(A)$  = probability of selecting a club

$$P(A) = \frac{13}{52}$$

$P(B)$  = probability of selecting a ten

$$P(B) = \frac{4}{52}$$

$$P(A \cap B) = \frac{1}{52}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$

$$P(A \cup B) = \frac{16}{52}$$

$$P(A \cup B) = \frac{4}{13}$$

### **Example 9**

#### 1.3. Mutually Inclusive and Mutually Exclusive Events

What is the probability of choosing a number from 1 to 10 that is less than 5 or odd?

**Solution:**

$$A = \{1, 2, 3, 4\}$$

$P(A)$  = probability of selecting a number less than 5

$$P(A) = \frac{4}{10}$$

$$P(A) = \frac{2}{5}$$

$$B = \{1, 3, 5, 7, 9\}$$

$P(B)$  = probability of selecting a number that is odd

$$P(B) = \frac{5}{10}$$

$$P(B) = \frac{1}{2}$$

$$P(A \cap B) = \frac{2}{10}$$

$$P(A \cap B) = \frac{1}{5}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = \frac{2}{5} + \frac{1}{2} - \frac{1}{5}$$

$$P(A \cup B) = \frac{4}{10} + \frac{5}{10} - \frac{2}{10}$$

$$P(A \cup B) = \frac{7}{10}$$

Notice in the previous 2 examples how the concept of double counting was incorporated into the calculation by subtracting the  $P(A \cap B)$ . Let's try a different example where you have 2 events happening.

**Example 10**

2 fair dice are rolled. What is the probability of getting a sum less than 7 or a sum equal to 10?

**Solution:**

$P(A)$  = probability of obtaining a sum less than 7

$$P(A) = \frac{15}{36}$$

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$P(B)$  = probability of obtaining a sum equal to 10

$$P(B) = \frac{1}{36}$$

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

There are no elements that are common, so the events are mutually exclusive.

$$P(A \text{ and } B) = 0$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = \frac{15}{36} + \frac{1}{36} - 0$$

$$P(A \cup B) = \frac{16}{36}$$

$$P(A \cup B) = \frac{4}{9}$$

### Example 11

2 fair dice are rolled. What is the probability of getting a sum less than 7 or a sum less than 4?

**Solution:**

$P(A)$  = probability of obtaining a sum less than 7

$$P(A) = \frac{15}{36}$$

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

$P(B)$  = probability of obtaining a sum less than 4

$$P(B) = \frac{3}{36}$$

### 1.3. Mutually Inclusive and Mutually Exclusive Events

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Notice that there are 3 elements in common. Therefore, the events are not mutually exclusive, and we must account for the double counting.

$$P(A \text{ and } B) = \frac{3}{36}$$

$$P(A \cap B) = \frac{3}{36}$$

$$P(A \cap B) = \frac{1}{12}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = \frac{15}{36} + \frac{3}{36} - \frac{1}{12}$$

$$P(A \cup B) = \frac{15}{36} + \frac{3}{36} - \frac{3}{36}$$

$$P(A \cup B) = \frac{15}{36}$$

$$P(A \cup B) = \frac{5}{12}$$

### Points to Consider

- What is the difference between the probabilities calculated with the Multiplication Rule versus the Addition Rule?
- Can mutually exclusive events be independent? Can they be dependent?

### Vocabulary

**Addition Principle (Rule)** With 2 events, the probability of one event occurring or another is given by:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

**Dependent events** 2 or more events whose outcomes affect each other. The probability of occurrence of one event depends on the occurrence of the other.

**Independent events** 2 or more events whose outcomes do not affect each other.

**Multiplication Rule** For 2 independent events ( $A$  and  $B$ ) where the outcome of  $A$  does not change the probability of  $B$ , the probability of  $A$  and  $B$  is given by:  $P(A \text{ and } B) = P(A) \times P(B)$ .



**Mutually exclusive events** 2 outcomes or events are mutually exclusive when they cannot both occur simultaneously.

**Mutually inclusive events** 2 outcomes or events are mutually inclusive when they can both occur simultaneously.

**Outcomes** The possible results of 1 trial of a probability experiment.

**Probability** The chance that something will happen.

**Random** When everyone or everything in a population has an equal chance of being selected.

**Sample space** The set of all possible outcomes of an event or group of events.

**Venn diagram** A diagram of overlapping circles that shows the relationships among members of different sets.

The union of 2 events, where the sample space contains 2 events,  $A$  and  $B$ , and each member of the set belongs to  $A$  or  $B$ .

The intersection of 2 events, where the sample space contains 2 events,  $A$  and  $B$ , and each member of the set belongs to  $A$  and  $B$ .

## 1.4 Review Questions

Answer the following questions and show all work (including diagrams) to create a complete answer.

- Determine which of the following are examples of independent events.
  - Rolling a 5 on one die and rolling a 5 on a second die.
  - Choosing a cookie from the cookie jar and choosing a jack from a deck of cards.
  - Winning a hockey game and scoring a goal.
- Determine which of the following are examples of independent events.
  - Choosing an 8 from a deck of cards, replacing it, and choosing a face card.
  - Going to the beach and bringing an umbrella.
  - Getting gasoline for your car and getting diesel fuel for your car.
- Determine which of the following are examples of dependent events.
  - Selecting a marble from a container and selecting a jack from a deck of cards.
  - Rolling a number less than 4 on a die and rolling a number that is even.
  - Choosing a jack from a deck of cards and choosing a heart.
- Determine which of the following are examples of dependent events.
  - Selecting a book from the library and selecting a book that is a mystery novel.
  - Rolling a 2 on a die and flipping a coin to get tails.
  - Being lunchtime and eating a sandwich.
- 2 dice are tossed. What is the probability of obtaining a sum equal to 6?
- 2 dice are tossed. What is the probability of obtaining a sum less than 6?
- 2 dice are tossed. What is the probability of obtaining a sum greater than 6?
- 2 dice are tossed. What is the probability of obtaining a sum of at least 6?
- A coin and a die are tossed. Calculate the probability of getting tails and a 5.
- ABC* High School is debating whether or not to write a policy where all students must have uniforms and wear them during school hours. In a survey, 45% of the students wanted uniforms, 35% did not, and 10% said they did not mind a uniform and did not care if there was no uniform. Represent this information in a Venn diagram.
- ABC* High School is debating whether or not to write a policy where all students must have uniforms and wear them during school hours. In a survey, 45% of the students wanted uniforms, and 55% did not. Represent this information in a Venn diagram.
- For question 11, calculate the probability that a person selected at random from *ABC* High School will want the school to have uniforms.
- Consider a sample set as  $S = \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$ . Event *A* is the multiples of 4, while event *B* is the multiples of 5. What is the probability that a number chosen at random will be from both *A* and *B*?
- For question 13, what is the probability that a number chosen at random will be from either *A* or *B*?
- Thomas bought a bag of jelly beans that contained 10 red jelly beans, 15 blue jelly beans, and 12 green jelly beans. What is the probability of Thomas reaching into the bag and pulling out a blue or green jelly bean?
- Thomas bought a bag of jelly beans that contained 10 red jelly beans, 15 blue jelly beans, and 12 green jelly beans. What is the probability of Thomas reaching into the bag and pulling out a blue or green jelly bean and then reaching in again and pulling out a red jelly bean?
- Jack is a student in Bluenose High School. He noticed that a lot of the students in his math class were also in his chemistry class. In fact, of the 60 students in his grade, 32 students were in his math class and 28 students were in his chemistry class. He decided to do a survey to find out what the probability was of selecting a

student at random from his grade who is taking either chemistry or math, but not both. Draw a Venn diagram and help Jack with his calculation.

18. Brenda did a survey of the students in her classes about whether they liked to get a candy bar or a new math pencil as their reward for positive behavior. She asked all 75 students she taught, and 32 said they would like a candy bar, 25 said they wanted a new pencil, and 4 said they wanted both. If Brenda were to select a student at random from her classes, what is the probability that the student chosen would want:
  - a. a candy bar or a pencil?
  - b. neither a candy bar nor a pencil?
19. A card is chosen at random from a standard deck of cards. What is the probability that the card chosen is a heart and spade? Are these events mutually exclusive?
20. A card is chosen at random from a standard deck of cards. What is the probability that the card chosen is a heart and a face card? Are these events mutually exclusive?

# CHAPTER 2 The Next Step ... Conditional Probability

## Chapter Outline

---

- 2.1 WHAT ARE TREE DIAGRAMS?
  - 2.2 ORDER AND PROBABILITY
  - 2.3 CONDITIONAL PROBABILITY
  - 2.4 REVIEW QUESTIONS
- 

### Introduction



This chapter builds on the concepts learned in the previous chapter on probability. Starting with tree diagrams as a means of displaying outcomes for various trials, we will learn how to read the diagrams and find probabilities. We will also find that order does not matter unless working with permutations. Permutations, such as the combination of your lock at the gym, require their own special formula. When outcomes for permutations have repetitions, these repetitions must also be included in the calculations in order to account for the multiple entries. Combinations, like permutations, also have their own special formula. Combinations, such as the number of teams of 4 that can be arranged in a class of 15 students, are different from permutations, because the order in combinations is insignificant. In this chapter, we will also learn about conditional probability. Conditional probability comes into play when the probability of the second event occurring is dependent on the probability of the first event.

## 2.1 What are Tree Diagrams?

### Learning Objectives

- Know the definition of conditional probability.
- Use conditional probability to solve for probabilities in finite sample spaces.

In the last chapter, we studied independent and dependent events, as well as mutually inclusive and mutually exclusive events. We used the Addition Rule for dependent events, as well as mutually inclusive and mutually exclusive events. The Addition Rule, or Addition Principle, is used to find  $P(A \text{ or } B)$ , while the Multiplication Rule is used for independent events.

**Addition Rule** – For 2 events,  $A$  and  $B$ , the probability of selecting one event or another is given by:  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ .

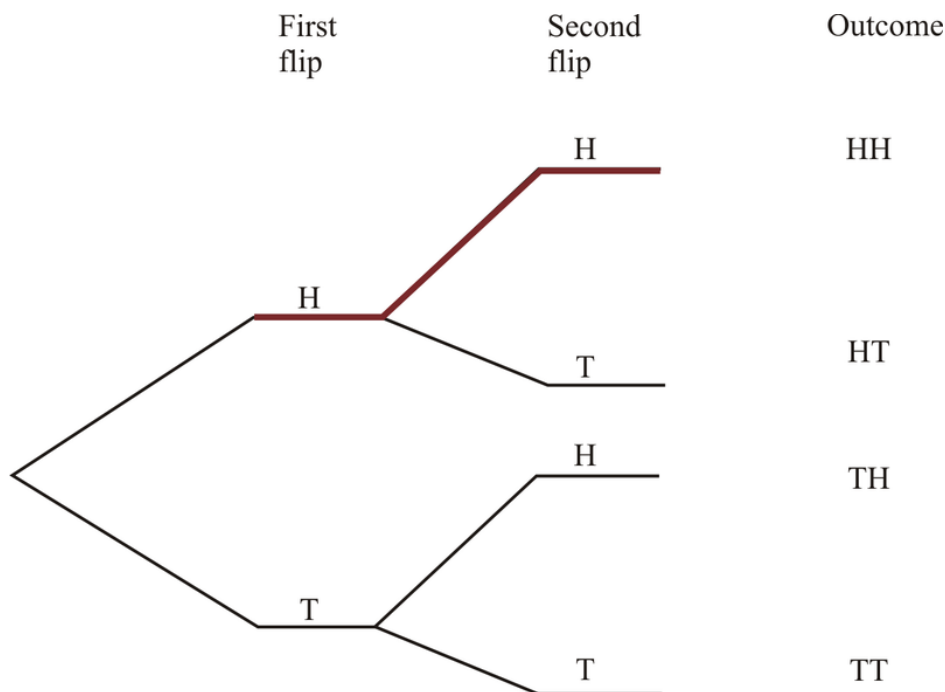
**Multiplication Rule** – For 2 independent events,  $A$  and  $B$ , where the outcome of  $A$  does not change the probability of  $B$ , the probability of  $A$  and  $B$  is given by:  $P(A \text{ and } B) = P(A) \times P(B)$ .



**Tree diagrams** are another way to show the outcomes of simple probability events. In a tree diagram, each outcome is represented as a branch on a tree.

Let's say you were going to toss a coin 2 times and wanted to find the probability of getting 2 heads. This is an example of independent events, because the outcome of one event does not affect the outcome of the second event. What does this mean? Well, when you flip the coin once, you have an equal chance of getting a head (H) or a tail (T). On the second flip, you also have an equal chance of getting a head or a tail. In other words, whether the first flip was heads or tails, the second flip could just as likely be heads as tails. You can represent the outcomes of these events on a tree diagram.

### 2.1. What are Tree Diagrams?



From the tree diagram, you can see that the probability of getting a head on the first flip is  $\frac{1}{2}$ . Starting with heads, the probability of getting a second head will again be  $\frac{1}{2}$ . But how do we calculate the probability of getting 2 heads? These are independent events, since the outcome of tossing the first coin in no way affects the outcome of tossing the second coin. Therefore, we can calculate the probability as follows:

$$P(A \text{ and } B) = \frac{1}{2} \times \frac{1}{2}$$

$$P(A \text{ and } B) = \frac{1}{4}$$

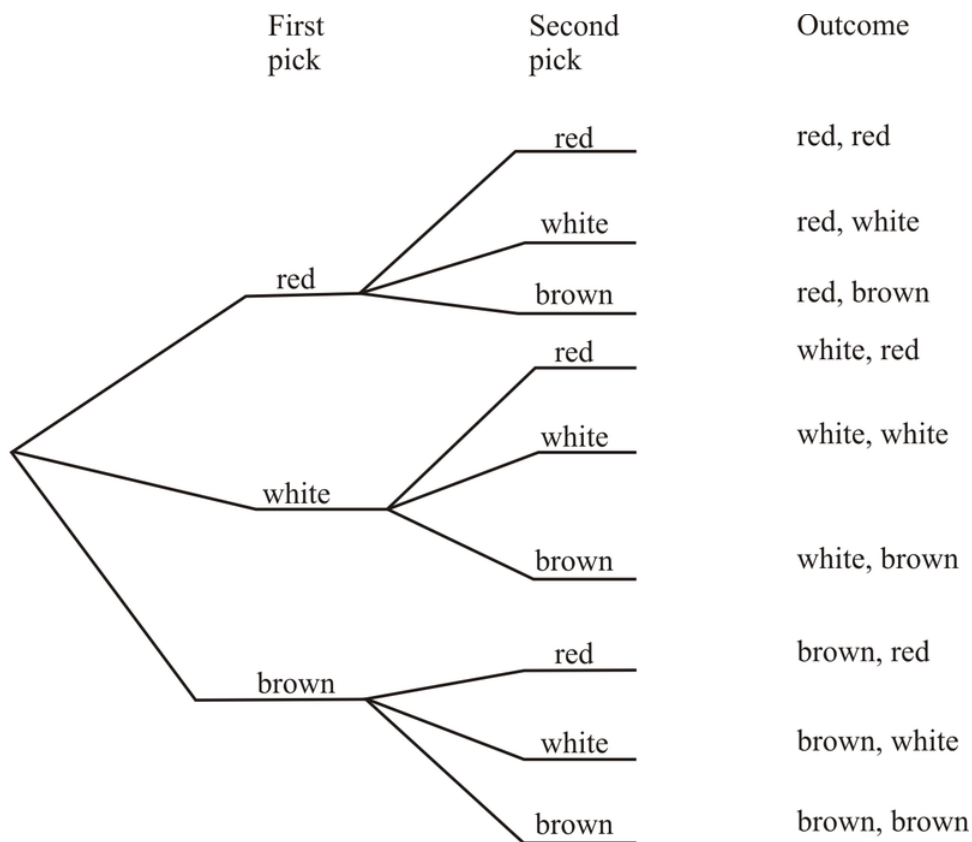
Therefore, we can conclude that the probability of getting 2 heads when tossing a coin twice is 25%, or  $\frac{1}{4}$ . Let's try an example that is a little more challenging.

### **Example 1**

Irvin opens up his sock drawer to get a pair socks to wear to school. He looks in the sock drawer and sees 4 red socks, 8 white socks, and 6 brown socks. Irvin reaches in the drawer and pulls out a red sock. He is wearing blue shorts, so he replaces it. He then draws out a white sock. What is the probability that Irvin pulls out a red sock, replaces it, and then pulls out a white sock?

### **Solution:**

First let's draw a tree diagram.



There are 18 socks in Irvin's sock drawer. The probability of getting a red sock when he pulls out the first sock is:

$$P(\text{red}) = \frac{4}{18}$$

$$P(\text{red}) = \frac{2}{9}$$

Irvin puts the sock back in the drawer and pulls out the second sock. The probability of getting a white sock on the second draw is:

$$P(\text{white}) = \frac{6}{18}$$

$$P(\text{white}) = \frac{1}{3}$$

Therefore, the probability of getting a red sock and then a white sock when the first sock is *replaced* is:

$$P(\text{red and white}) = \frac{2}{9} \times \frac{1}{3}$$

$$P(\text{red and white}) = \frac{2}{27}$$

One important part of these types of problems is that order is not important.

Let's say Irvin picked out a white sock, replaced it, and then picked out a red sock. Calculate this probability.

### 2.1. What are Tree Diagrams?

$$P(\text{white and red}) = \frac{1}{3} \times \frac{2}{9}$$

$$P(\text{white and red}) = \frac{2}{27}$$

So regardless of the order in which he takes the socks out, the probability is the same. In other words,  $P(\text{red and white}) = P(\text{white and red})$ .

### Example 2

In Example 1, what happens if the first sock is *not replaced*?

#### Solution:

The probability that the first sock is red is:

$$P(\text{red}) = \frac{4}{18}$$

$$P(\text{red}) = \frac{2}{9}$$

The probability of picking a white sock on the second pick is now:

$$P(\text{white}) = \frac{6}{17}$$

Notice the denominator decreased by 1. We now have **17** remaining socks in the drawer.

So now, the probability of selecting a red sock and then a white sock, without replacement, is:

$$P(\text{red and white}) = \frac{2}{9} \times \frac{6}{17}$$

$$P(\text{red and white}) = \frac{12}{153}$$

$$P(\text{red and white}) = \frac{4}{51}$$


If the first sock is white, will  $P(\text{red and white}) = P(\text{white and red})$  as we found in Example 1? Let's find out.

$$P(\text{white}) = \frac{6}{18}$$

$$P(\text{white}) = \frac{1}{3}$$

The probability of picking a red sock on the second pick is now:



$$P(\text{red}) = \frac{4}{17}$$


Notice the denominator decreased by 1. We now have **17** remaining socks in the drawer.

$$P(\text{white and red}) = \frac{1}{3} \times \frac{4}{17}$$

$$P(\text{white and red}) = \frac{4}{51}$$

As with the last example,  $P(\text{red and white}) = P(\text{white and red})$ . So when does order *really* matter?

## 2.2 Order and Probability

Permutations and combinations are the next step in the learning of probability. It is by using permutations and combinations that we can find the probabilities of various events occurring at the same time, such as choosing 3 boys and 3 girls from a class of grade 12 math students.

In mathematics, we use more precise language:

If the order doesn't matter, it is a combination.

If the order does matter, it is a permutation.

Say, for example, you are making a salad. You throw in some lettuce, carrots, cucumbers, and green peppers. The order in which you throw in these vegetables doesn't really matter. Here we are talking about a combination. For combinations, you are merely selecting. Say, though, that Jack went to the ATM to get out some money and that he has to put in his PIN number. Here the order of the digits in the PIN number is quite important. In this case, we are talking about a permutation. For permutations, you are ordering objects in a specific manner.

### Permutations

The **Fundamental Counting Principle** states that if an event can be chosen in  $p$  different ways and another independent event can be chosen in  $q$  different ways, the number of different ways the 2 events can occur is  $p \times q$ . In other words, the Fundamental Counting Principle tells you how many ways you can arrange items. **Permutations** are the number of possible arrangements in an ordered set of objects.

#### Example 3

How many ways can you arrange the letters in the word MATH?

#### Solution:

You have 4 letters in the word, and you are going to choose 1 letter at a time. When you choose the first letter, you have 4 possibilities ('M', 'A', 'T', or 'H'). Your second choice will have 3 possibilities, your third choice will have 2 possibilities, and your last choice will have only 1 possibility.

Therefore, the number of arrangements is:  $4 \times 3 \times 2 \times 1 = 24$  possible arrangements.

The notation for a permutation is:  ${}_n P_r$ ,

where:

$n$  is the *total* number of objects.

$r$  is the number of objects chosen.

For simplifying calculations, when  $n = r$ , then  ${}_n P_r = n!$ .

The **factorial function (!)** requires us to multiply a series of descending natural numbers.

Examples:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$$

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

$$1! = 1$$

Note: It is a general rule that  $0! = 1$ .

With TI calculators, you can find the factorial function using:

MATH ▶▶▶ (PRB) ▼▼▼ (4)

#### Example 4

Solve for  ${}_4P_4$ .

#### Solution:

$${}_4P_4 = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

This represents the number of ways to arrange 4 objects that are chosen from a set of 4 different objects.

#### Example 5

Solve for  ${}_6P_3$ .

#### Solution:

Starting with 6, multiply the first 3 numbers of the factorial:

$${}_6P_3 = 6 \cdot 5 \cdot 4 = 120$$

This represents the number of ways to arrange 3 objects that are chosen from a set of 6 different objects.

The formula to solve permutations like these is:

$${}_nP_r = \frac{n!}{(n-r)!}$$

Look at Example 5 above. In this example, the total number of objects ( $n$ ) is 6, while the number of objects chosen ( $r$ ) is 3. We can use these 2 numbers to calculate the number of possible permutations (or the number of arrangements) of 6 objects chosen 3 at a time.

$$\begin{aligned} {}nP_r &= \frac{n!}{(n-r)!} \\ {}_6P_3 &= \frac{6!}{(6-3)!} \\ {}_6P_3 &= \frac{6!}{3!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{\cancel{3 \times 2 \times 1}} \\ {}_6P_3 &= \frac{120}{1} \\ {}_6P_3 &= 120 \end{aligned}$$

#### Example 6

What is the total number of possible 4-letter arrangements of the letters 's', 'n', 'o', and 'w' if each letter is used only once in each arrangement?

#### Solution:

In this problem, there are 4 letters to choose from, so  $n = 4$ . We want 4-letter arrangements; therefore, we are choosing 4 objects at a time. In this example,  $r = 4$ .

## 2.2. Order and Probability

$${}_n P_r = \frac{n!}{(n-r)!}$$

$${}_4 P_4 = \frac{4!}{(4-4)!}$$

$${}_4 P_4 = \frac{4!}{0!} = \frac{4 \times 3 \times 2 \times 1}{1}$$

$${}_4 P_4 = \frac{24}{1}$$

$${}_4 P_4 = 24$$



Notice the rule  $0! = 1$

### Example 7

A committee is formed with a president, a vice president, and a treasurer. If there are 10 people to select from, how many committees are possible?

#### Solution:

In this problem, there are 10 committee members to choose from, so  $n = 10$ . We want to choose 3 members to be president, vice-president, and treasurer; therefore, we are choosing 3 objects at a time. In this example,  $r = 3$ .

$$\begin{aligned} {}_n P_r &= \frac{n!}{(n-r)!} \\ {}_{10} P_3 &= \frac{10!}{(10-3)!} \\ {}_{10} P_3 &= \frac{10!}{7!} = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \\ {}_{10} P_3 &= \frac{720}{1} \\ {}_{10} P_3 &= 720 \end{aligned}$$

### Permutations with Repetition

There is a subset of permutations that takes into account that there are double objects or repetitions in a permutation problem. In general, repetitions are taken care of by dividing the permutation by the factorial of the number of objects that are identical.

If you look at the word TOOTH, there are 2 O's in the word. Both O's are identical, and it does not matter in which order we write these 2 O's, since they are the same. In other words, if we exchange 'O' for 'O', we still spell TOOTH. The same is true for the T's, since there are 2 T's in the word TOOTH as well.

If we were to ask the question, "In how many ways can we arrange the letters in the word TOOTH?" we must account for the fact that these 2 O's are identical and that the 2 T's are identical. We do this using the formula:

$\frac{{}_n P_r}{x_1! x_2!}$ , where  $x$  is the number of times a letter is repeated.

$$\begin{aligned}\frac{{}_n P_r}{{x_1!x_2!}} &= \frac{{}_5 P_5}{2!2!} \\ \frac{{}_5 P_5}{2!2!} &= \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} \\ \frac{{}_5 P_5}{2!2!} &= \frac{120}{4} \\ \frac{{}_5 P_5}{2!2!} &= 30\end{aligned}$$

### Tech Tip: Calculating Permutations on the Calculator

#### Permutations ( $nPr$ )

Enter the  $n$  value. Press **MATH**. You should see modes across the top of the screen. You want the fourth menu: PRB (arrow right 3 times). You will see several options.  $nPr$  is the second. Press **2**. Enter the  $r$  value. Press **ENTER**.

#### Example 8

Compute  ${}_9 P_5$ .

#### Solution:

**9** **MATH** **▶▶▶** (PRB) **2** (nPr) **5** **ENTER**

$${}_9 P_5 = 15,120$$

#### Example 9

How many different 5-letter arrangements can be formed from the word APPLE?

#### Solution:

There are 5 letters in the word APPLE, so  $n = 5$ . We want 5-letter arrangements; therefore, we are choosing 5 objects at a time. In this example,  $r = 5$ , and we are using a word with letters that repeat. In the word APPLE, there are 2 P's, so  $x_1 = 2$ .

$$\frac{{}_n P_r}{{x_1!}} = \frac{{}_5 P_5}{2!}$$

There are 5 letters,  $n = 5$ , and you are choosing all 5 digits,  $r = 5$

$$\frac{{}_5 P_5}{2!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = \frac{120}{2}$$

There are 2 letters repeating (P's), therefore divide by 2!

$$\frac{{}_5 P_5}{2!} = 60 \text{ arrangements}$$

#### Example 10

How many different 6-digit numerals can be written using all of the following 7 digits?

3, 3, 4, 4, 4, 5, 6.

#### Solution:

There are 7 digits, so  $n = 7$ . We want 6-digit arrangements; therefore, we are choosing 6 objects at a time. In this

example,  $r = 6$ , and we are using a group of digits with numbers that repeat. In the group of 7 digits (3, 3, 4, 4, 4, 5, 6), there are two 3's and three 4's, so  $x_1 = 2$  and  $x_2 = 3$ .

$$\frac{{}_n P_r}{x_1! x_2!} = \frac{{}_7 P_6}{2! 3!}$$

There are 7 numbers,  $n = 7$ , and you are choosing 6 digits,  $r = 6$

$$\frac{{}_7 P_6}{2! 3!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2}{2 \times 1 \times 3 \times 2 \times 1} = \frac{5040}{12}$$

There are two threes therefore divide by  $2!$ , and there are three fours, therefore divide by  $3!$

$$\frac{{}_7 P_6}{2! 3!} = 420$$

When there are no repetitions, remember that we use the standard permutation formula:

$${}_n P_r = \frac{n!}{(n-r)!}$$

### Example 11

In how many ways can first and second place be awarded to 10 people?

#### Solution:

There are 10 people ( $n = 10$ ), and there are 2 prize winners ( $r = 2$ ).

$${}_n P_r = \frac{n!}{(n-r)!}$$

$${}_{10} P_2 = \frac{10!}{(10-2)!}$$

$${}_{10} P_2 = \frac{10!}{8!}$$

$${}_{10} P_2 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{\cancel{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}}$$

$${}_{10} P_2 = 90$$

**Example 12**

In how many ways can 3 favorite desserts be listed in order from a menu of 10 (i.e., permutations without repetition)?

**Solution:**

There are 10 menu items ( $n = 10$ ), and you are choosing 3 favorite desserts ( $r = 3$ ) in order.

$${}_{10}P_3 = \frac{10!}{(10 - 3)!}$$

$${}_{10}P_3 = \frac{10!}{7!}$$

$${}_{10}P_3 = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}$$

$${}_{10}P_3 = 720$$

**Combinations**

If you think about the lottery, you choose a group of lucky numbers in hopes of winning millions of dollars. When the numbers are drawn, the order in which they are drawn does not have to be the same order as on your lottery ticket. The numbers drawn simply have to be on your lottery ticket in order for you to win. You can imagine how many possible combinations of numbers exist, which is why your odds of winning are so small!



**Combinations** are arrangements of objects *without* regard to order and without repetition, selected from a distinct number of objects. A combination of  $n$  objects taken  $r$  at a time ( ${}_nC_r$ ) can be calculated using the formula:

$${}_nC_r = \frac{n!}{r!(n - r)!}$$

**Example 13**

Evaluate:  ${}_7C_2$ .

**Solution:**

$$\begin{aligned}
{}_7C_2 &= \frac{7!}{2!(7-2)!} \\
{}_7C_2 &= \frac{7!}{2!(5)!} \\
{}_7C_2 &= \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(5 \times 4 \times 3 \times 2 \times 1)} \\
{}_7C_2 &= \frac{5040}{(2)(120)} \\
{}_7C_2 &= \frac{5040}{240} \\
{}_7C_2 &= 21
\end{aligned}$$

**Example 14**

In how many ways can 3 desserts be chosen in any order from a menu of 10?

**Solution:**

There are 10 menu items ( $n = 10$ ), and you are choosing 3 desserts in any order ( $r = 3$ ).

$$\begin{aligned}
{}_{10}C_3 &= \frac{10!}{3!(10-3)!} \\
{}_{10}C_3 &= \frac{10!}{3!(7)!} \\
{}_{10}C_3 &= 120
\end{aligned}$$

**Example 15**

There are 12 boys and 14 girls in Mrs. Cameron's math class. Find the number of ways Mrs. Cameron can select a team of 3 students from the class to work on a group project. The team must consist of 2 girls and 1 boy.

**Solution:**

There are groups of both boys and girls to consider. From the 14 girls ( $n = 14$ ) in the class, we are choosing 2 ( $r = 2$ ).

Girls:

$$\begin{aligned}
{}_{14}C_2 &= \frac{14!}{2!(14-2)!} \\
{}_{14}C_2 &= \frac{14!}{2!(12)!} \\
{}_{14}C_2 &= \frac{87178291200}{2(479001600)} \\
{}_{14}C_2 &= \frac{87178291200}{958003200} \\
{}_{14}C_2 &= 91
\end{aligned}$$

From the 12 boys ( $n = 12$ ) in the class, we are choosing 1 ( $r = 1$ ).



Boys:

$${}_{12}C_1 = \frac{12!}{1!(12-1)!}$$

$${}_{12}C_1 = \frac{12!}{1!(11)!}$$

$${}_{12}C_1 = \frac{479001600}{1(39916800)}$$

$${}_{12}C_1 = \frac{479001600}{39916800}$$

$${}_{12}C_1 = 12$$

Therefore, the number of ways Mrs. Cameron can select a team of 3 students (2 girls and 1 boy) from the class of 26 students to work on a group project is:

$$\text{Total combinations} = {}_{14}C_2 \times {}_{12}C_1 = 91 \times 12 = 1092$$

### Example 16

If there are 20 rock songs and 20 rap songs to choose from, in how many different ways can you select 12 rock songs and 7 rap songs for a mix CD?

**Solution:**

As in Example 13, we have multiple groups from which we are required to select, so we have to calculate the possible combinations for each group (rock songs and rap songs in this example) separately and then multiply together.

Using TI technology: for  ${}_nC_r$ , type the  $n$  value (the total number of items), and then press **MATH** **▶** **▶** **▶** **(PRB)** **▼** **▼** (to number 3) **ENTER**. Then type the  $r$  value (the number of items your want to choose), and finally, press **ENTER**.

Rock:

$${}_{20}C_{12} = \frac{20!}{12!(20-12)!}$$

$${}_{20}C_{12} = \frac{20!}{12!(8)!}$$

$${}_{20}C_{12} = 125970$$

Rap:

$${}_{20}C_7 = \frac{20!}{7!(20-7)!}$$

$${}_{20}C_7 = \frac{20!}{7!(13)!}$$

$${}_{20}C_7 = 77520$$

Therefore, the number of possible combinations is:

$$\begin{array}{cc} \text{Rock} & \text{Rap} \\ {}_{20}C_{12} \times {}_{20}C_7 & = 125,970 \times 77,520 = 9.765 \times 10^9 \text{ possible combinations} \end{array}$$

## 2.2. Order and Probability

**Tech Tip: Calculating Combinations on the Calculator****Combinations** ( $nCr$ )

Enter the  $n$  value. Press  $\boxed{\text{MATH}}$ . You should see modes across the top of the screen. You want the fourth menu: PRB (arrow right 3 times). You will see several options.  $nCr$  is the third. Press  $\boxed{3}$ . Enter the  $r$  value. Press  $\boxed{\text{ENTER}}$ .

**Example 17**

Compute  ${}_{10}C_6$ .

**Solution:**

$${}_{10}C_6 \quad \boxed{1} \boxed{0} \boxed{\text{MATH}} \boxed{\blacktriangleright} \boxed{\blacktriangleright} \boxed{\blacktriangleright} \boxed{(\text{PRB})} \boxed{3} \boxed{(\text{nCr})} \boxed{6} \boxed{\text{ENTER}}$$
$${}_{10}C_6 = 210$$

## 2.3 Conditional Probability

What if the probability of a second event is affected by the probability of the first event? This type of probability calculation is known as **conditional probability**.

When working with events that are conditionally probable, you are working with 2 events, where the probability of the second event is conditional on the first event occurring. Say, for example, that you want to know the probability of drawing 2 kings from a deck of cards. As we have previously learned, here is how you would calculate this:

$$\begin{aligned}
 P(\text{first king}) &= \frac{1}{13} \\
 P(\text{second king}) &= \frac{3}{51} \\
 P(2 \text{ kings}) &= \frac{1}{13} \times \frac{3}{51} \\
 P(2 \text{ kings}) &= \frac{3}{663} \\
 P(2 \text{ kings}) &= \frac{1}{221}
 \end{aligned}$$

Now let's assume you are playing a game where you need to draw 2 kings to win. You draw the first card and get a king. What is the probability of getting a king on the second card? The probability of getting a king on the second card can be thought of as a conditional probability. The formula for calculating conditional probability is given as:

$$\begin{aligned}
 P(B|A) &= \frac{P(A \cap B)}{P(A)} \\
 P(A \cap B) &= P(A) \times P(B|A)
 \end{aligned}$$

Another way to look at the conditional probability formula is as follows. Assuming the first event has occurred, the probability of the second event occurring is:

$$P(\text{second event}|\text{first event}) = \frac{P(\text{first event and second event})}{P(\text{first event})}$$

Let's work through a few problems using the formula for conditional probability.

### Example 18

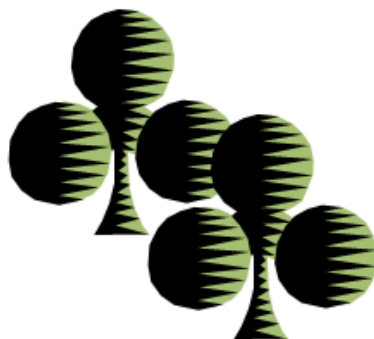
You are playing a game of cards where the winner is determined when a player gets 2 cards of the same suit. You draw a card and get a club ( $\clubsuit$ ). What is the probability that the second card will be a club?

**Solution:**

**Step 1:** List what you know.

First event = drawing the first club

Second event = drawing the second club



$$P(\text{first club}) = \frac{13}{52}$$

$$P(\text{second club}) = \frac{12}{51}$$

$$P(\text{club and club}) = \frac{13}{52} \times \frac{12}{51}$$

$$P(\text{club and club}) = \frac{156}{2652}$$

$$P(\text{club and club}) = \frac{1}{17}$$

**Step 2:** Calculate the probability of choosing a club as the second card when a club is chosen as the first card.

$$\text{Probability of drawing the second club} = \frac{P(\text{club and club})}{P(\text{first club})}$$

$$P(\text{club}|\text{club}) = \frac{\frac{1}{17}}{\frac{13}{52}}$$

$$P(\text{club}|\text{club}) = \frac{1}{17} \times \frac{52}{13}$$

$$P(\text{club}|\text{club}) = \frac{52}{221}$$

$$P(\text{club}|\text{club}) = \frac{4}{17}$$

**Step 3:** Write your conclusion.

Therefore, the probability of selecting a club as the second card when a club is chosen as the first card is 24%.

### **Example 19**

In the next round of the game, the first person to be dealt a black ace wins the game. You get your first card, and it is a queen. What is the probability of obtaining a black ace?

**Solution:**

**Step 1:** List what you know.

First event = being dealt the queen

Second event = being dealt the black ace



$$P(\text{queen}) = \frac{4}{52}$$

$$P(\text{black ace}) = \frac{2}{51}$$

$$P(\text{black ace and queen}) = \frac{4}{52} \times \frac{2}{51}$$

$$P(\text{black ace and queen}) = \frac{8}{2652}$$

$$P(\text{black ace and queen}) = \frac{4}{663}$$

**Step 2:** Calculate the probability of choosing black ace as a second card when a queen is chosen as a first card.

$$P(\text{black ace}|\text{queen}) = \frac{P(\text{black ace and queen})}{P(\text{queen})}$$

$$P(\text{black ace}|\text{queen}) = \frac{\frac{4}{663}}{\frac{4}{52}}$$

$$P(\text{black ace}|\text{queen}) = \frac{4}{663} \times \frac{52}{4}$$

$$P(\text{black ace}|\text{queen}) = \frac{52}{663}$$

$$P(\text{black ace}|\text{queen}) = \frac{4}{51}$$

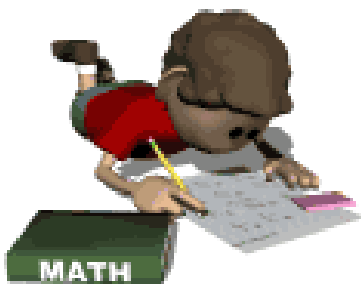
**Step 3:** Write your conclusion.

Therefore, the probability of selecting a black ace as the second card when a queen is chosen as the first card is 7.7%.

### Example 20

At Bluenose High School, 90% of the students take physics and 35% of the students take both physics and statistics. What is the probability that a student from Bluenose High School who is taking physics is also taking statistics?

### 2.3. Conditional Probability



**Solution:**

**Step 1:** List what you know.

$$P(\text{physics}) = 0.90$$

$$P(\text{physics and statistics}) = 0.35$$

**Step 2:** Calculate the probability of choosing statistics as a second course when physics is chosen as a first course.

$$P(\text{statistics}|\text{physics}) = \frac{P(\text{physics and statistics})}{P(\text{physics})}$$

$$P(\text{statistics}|\text{physics}) = \frac{0.35}{0.90}$$

$$P(\text{statistics}|\text{physics}) = 0.388$$

$$P(\text{statistics}|\text{physics}) = 39\%$$

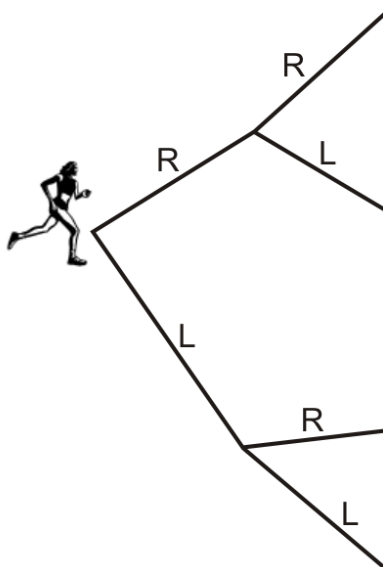
**Step 3:** Write your conclusion.

Therefore, the probability that a student from Bluenose High School who is taking physics is also taking statistics is 39%.

### **Example 21**

Sandra went out for her daily run. She goes on a path that has alternate routes to give her a variety of choices to make her run more enjoyable. The path has 3 turns where she can go left or right at each turn. The probability of turning right the first time is  $\frac{1}{2}$ . Based on past runs, the probability of turning right the second time is  $\frac{2}{3}$ . Draw a tree diagram to represent the path. What is the probability that she will turn left the second time after turning right the first time?

**Solution:**



**Step 1:** List what you know.

$$P(\text{right the first time}) = \frac{1}{2}$$

$$P(\text{right the second time}) = \frac{2}{3}$$

$$P(\text{left the second time}) = 1 - \frac{2}{3} = \frac{1}{3}$$

$$P(\text{right the first time and left the second time}) = \frac{1}{2} \times \frac{1}{3}$$

$$P(\text{right the first time and left the second time}) = \frac{1}{6}$$

**Step 2:** Calculate the probability of choosing left as the second turn when right is chosen as the first turn.

$$P(\text{left the second time} | \text{right the first time}) = \frac{P(\text{right the first time and left the second time})}{P(\text{right the first time})}$$

$$P(\text{left the second time} | \text{right the first time}) = \frac{\frac{1}{6}}{\frac{1}{2}}$$

$$P(\text{left the second time} | \text{right the first time}) = \frac{1}{6} \times \frac{2}{1}$$

$$P(\text{left the second time} | \text{right the first time}) = \frac{2}{6}$$

$$P(\text{left the second time} | \text{right the first time}) = \frac{1}{3}$$

$$P(\text{left the second time} | \text{right the first time}) = 0.33\bar{3}$$

$$P(\text{left the second time} | \text{right the first time}) = 33\%$$

**Step 3:** Write your conclusion.

Therefore, the probability of choosing left as the second turn when right was chosen as the first turn is 33%.

### Points to Consider

#### 2.3. Conditional Probability

- How does a permutation differ from a combination?
- How are tree diagrams helpful for determining probabilities?

### Vocabulary

**Combinations** The number of possible arrangements  $({}_nC_r)$  of objects ( $r$ ) without regard to order and without repetition selected from a distinct number of objects ( $n$ ).

**Conditional probability** The probability of a particular dependent event, given the outcome of the event on which it depends.

**Factorial function (!)** To multiply a series of consecutive descending natural numbers.

**Fundamental Counting Principle** If an event can be chosen in  $p$  different ways and another independent event can be chosen in  $q$  different ways, the probability of the 2 events occurring is  $p \times q$ .

**Permutations** The number of possible arrangements  $({}_nP_r)$  in an ordered set of objects, where  $n$  = the number of objects and  $r$  = the number of objects selected.

**Tree diagrams** A way to show the outcomes of simple probability events, where each outcome is represented as a branch on a tree.



## 2.4 Review Questions

Answer the following questions and show all work (including diagrams) to create a complete answer.

- A bag contains 3 red balls and 4 blue balls. Thomas reaches in the bag and picks a ball out at random from the bag. He places it back into the bag. Thomas then reaches in the bag and picks another ball at random.
  - Draw a tree diagram to represent this problem.
  - What is the probability that Thomas picks:
    - 2 red balls
    - a red ball in his second draw
- A teacher has a prize box on her front desk for when students do exceptional work in math class. Inside the box there are 20 math pencils and 10 very cool erasers. Janet completed a challenge problem for Ms. Cameron, and Ms. Cameron rewarded Janet's innovative problem-solving approach with a trip to the prize box. Janet reaches into the box and picks out a prize and then drops it back in. Then she reaches in again and picks out a prize a second time.
  - Draw a tree diagram to represent this problem.
  - What is the probability that Janet reaches into the box and picks out an eraser on the second pick?

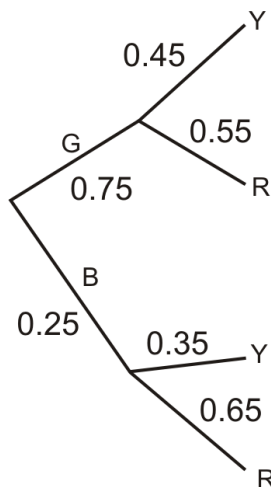


- Determine whether the following situations would require calculating a permutation or a combination:
  - Selecting 3 students to attend a conference in Washington, D.C.
  - Selecting a lead and an understudy for a school play
  - Assigning students to their seats on the first day of school
- Solve for  ${}_7P_5$ .
- Evaluate  ${}_4P_2 \times {}_5P_3$ .
- How many different 4-digit numerals can be made from the digits of 56987 if a digit can appear just once in a numeral?
- In how many ways can the letters of the word REFERENCE be arranged?
- In how many ways can the letters of the word MISSISSIPPI be arranged?
- In how many ways can the letters of the word MATHEMATICS be arranged?
- If there are 4 chocolate chip, 2 oatmeal, and 2 double chocolate cookies in a box, in how many different orders is it possible to eat all of these cookies?
- A math test is made up of 15 multiple choice questions. 5 questions have the answer A, 4 have the answer B, 3 have the answer C, 2 have the answer D, and 1 has the answer E. How many answer sheets are possible?
- In how many ways can you select 17 songs from a mix CD of a possible 38 songs?
- If an ice cream dessert can have 2 toppings, and there are 9 available, how many different selections can you make?

14. If there are 17 randomly placed dots on a circle, how many lines can be formed using any 2 dots?
15. A committee of 4 is to be formed from a group of 13 people. How many different committees can be formed?
16. There are 4 kinds of meat and 10 veggies available to make wraps at the school cafeteria. How many possible wraps have 1 kind of meat and 3 veggies?
17. There are 15 freshmen and 30 seniors in the Senior Math Club. The club is to send 4 representatives to the State Math Championships.
  - a. How many different ways are there to select a group of 4 students to attend the State Math Championships?
  - b. If the members of the club decide to send 2 freshmen and 2 seniors, how many different groupings are possible?



18. Students in BDF High School were asked about their preference regarding the new school colors. They were given a choice between green and blue as the primary color and red and yellow as the secondary color. The results of the survey are shown in the tree diagram below. You can see that 75% of the students choose green as the primary color. Of this 75%, 45% chose yellow as the secondary color. What is the probability that a student in BDF High School selected red as the secondary color if he or she chose blue as the primary color?



19. 2 fair dice are rolled. What is the probability that the sum is even given that the first die that is rolled is a 2?
20. 2 fair dice are rolled. What is the probability that the sum is even given that the first die rolled is a 5?
21. 2 fair dice are rolled. What is the probability that the sum is odd given that the first die rolled is a 5?
22. Steve and Scott are playing a game of cards with a standard deck of playing cards. Steve deals Scott a king. What is the probability that Scott's second card will be a red card?
23. Sandra and Karen are playing a game of cards with a standard deck of playing cards. Sandra deals Karen a seven. What is the probability that Karen's second card will be a black card?



24. Donna discusses with her parents the idea that she should get an allowance. She says that in her class, 55% of her classmates receive an allowance for doing chores, and 25% get an allowance for doing chores and are good to their parents. Her mom asks Donna what the probability is that a classmate will be good to his or her parents given that he or she receives an allowance for doing chores. What should Donna's answer be?
25. At a local high school, the probability that a student speaks English and French is 15%. The probability that a student speaks French is 45%. What is the probability that a student speaks English, given that the student speaks French?
26. At a local high school, the probability that a student takes statistics and art is 10%. The probability that a student takes art is 60%. What is the probability that a student takes statistics, given that the student takes art?



# CHAPTER 3

# Introduction to Discrete Random Variables

## Chapter Outline

- 3.1 WHAT ARE VARIABLES?
- 3.2 THE PROBABILITY DISTRIBUTION
- 3.3 A GLIMPSE AT BINOMIAL AND MULTINOMIAL DISTRIBUTIONS
- 3.4 USING TECHNOLOGY TO FIND PROBABILITY DISTRIBUTIONS
- 3.5 REVIEW QUESTIONS

### Introduction

In this chapter, you will learn about discrete random variables. **Random variables** are simply quantities that take on different values depending on chance, or probability. Discrete random variables can take on a finite number of values in an interval, or as many values as there are positive integers. In other words, a discrete random variable can take on an infinite number of values, but not all the values in an interval. When you find the probabilities of these values, you are able to show the probability distribution. A probability distribution consists of all the values of the random variable, along with the probability of the variable taking on each of these values. Each probability must be between 0 and 1, and the probabilities must sum to 1.

You will also be introduced to the concept of a binomial distribution. This will be discussed in depth in the next chapter, but in this chapter, you will use a binomial distribution when talking about the number of successful events or the value of a random variable. A binomial distribution is only used when there are 2 possible outcomes. For example, you will use the binomial distribution formula for coin tosses (heads or tails). Other examples include yes/no responses, true or false questions, and voting Democrat or Republican. When the number of possible outcomes goes beyond 2, you use a multinomial distribution. Rolling a die is a common example of a multinomial distribution problem.



In addition, you will use factorials again for solving these problems. Factorials were introduced in Chapter 2 for permutations and combinations, but they are also used in many other probability problems. Finally, you will use a graphing calculator to show the difference between theoretical and experimental probability. The calculator is an effective and efficient tool for illustrating the difference between these 2 probabilities, and also for determining the experimental probability when the number of trials is large.

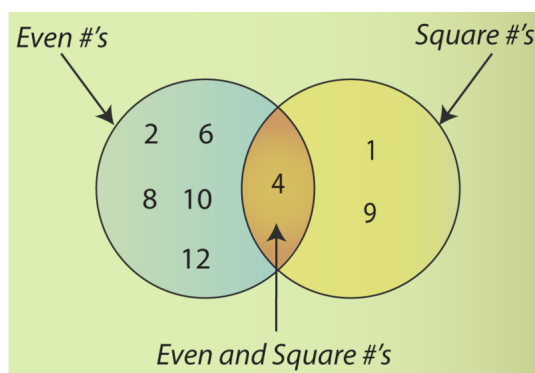


## 3.1 What are Variables?

### Learning Objectives

- Demonstrate an understanding of the notion of discrete random variables by using them to solve for the probabilities of outcomes, such as the probability of the occurrence of 5 heads in 14 coin tosses.

In the previous 2 chapters, we looked at the mathematics involved in probability events. We looked at examples of event  $A$  occurring if event  $B$  had occurred (conditional events), of event  $B$  being affected by the outcome of event  $A$  (dependent events), and of event  $A$  and event  $B$  not being affected by each other (independent events). We also looked at examples where events cannot occur at the same time (mutually exclusive events), or when events were not mutually exclusive and there was some overlap, so that we had to account for the double counting (mutually inclusive events). If you recall, we used Venn Diagrams (below), tree diagrams, and even tables to help organize information in order to simplify the mathematics for the probability calculations.



Our examination of probability, however, began with a look at the English language. Although there are a number of differences in what terms mean in mathematics and English, there are a lot of similarities as well. We saw this with the terms independent and dependent. In this chapter, we are going to learn about variables. In particular, we are going to look at discrete random variables. When you see the sequence of words *discrete random variables*, it may, at first, send a shiver down your spine, but let's look at the words individually and see if we can "simplify" the sequence!

The term discrete, in English, means to constitute a separate thing or to be related to unconnected parts. In mathematics, we use the term discrete when we are talking about pieces of data that are not connected. Random, in English, means to lack any plan or to be without any prearranged order. In mathematics, the definition is the same. Random events are fair, meaning that there is no way to tell what outcome will occur. In the English language, the term variable means to be likely to change or subject to variation. In mathematics, the term variable means to have no fixed quantitative value.

Now that we have seen the 3 terms separately, let's combine them and see if we can come up with a definition of a discrete random variable. We can say that discrete variables have values that are unconnected to each other and have variations within the values. Think about the last time you went to the mall. Suppose you were walking through the parking lot and were recording how many cars were made by Ford. The variable is the number of Ford cars you see. Therefore, since each car is either a Ford or it is not, the variable is discrete. If you randomly selected 20 cars from the parking lot and determined whether or not each was manufactured by Ford, you would then have a discrete random variable.

### 3.1. What are Variables?



Now let's define discrete random variables. **Discrete random variables** represent the number of distinct values that can be counted of an event. For example, when Robert was randomly chosen from all the students in his classroom and asked how many siblings there are in his family, he said that he has 6 sisters. Joanne picked a random bag of jelly beans at the store, and only 15 of 250 jelly beans were green. When randomly selecting from the most popular movies, Jillian found that *Iron Man 2* grossed 3.5 million dollars in sales on its opening weekend. Jack, walking with his mom through the parking lot, randomly selected 10 cars on his way up to the mall entrance and found that only 2 were Ford vehicles.

## 3.2 The Probability Distribution

When we talk about the probability of discrete random variables, we normally talk about a probability distribution. In a **probability distribution**, you may have a table, a graph, or a chart that shows you all the possible values of  $X$  (your variable), and the probability associated with each of these values ( $P(X)$ ).

It is important to remember that the values of a discrete random variable are mutually exclusive. Think back to our car example with Jack and his mom. Jack could not, realistically, find a car that is both a Ford and a Mercedes (assuming he did not see a home-built car). He would either see a Ford or not see a Ford as he went from his car to the mall doors. Therefore, the values for the variable are mutually exclusive. Now let's look at an example.

### Example 1

Say you are going to toss 2 coins. Show the probability distribution for this toss.

### Solution:

Let the variable be the number of times your coin lands on tails. The table below lists all of the possible events that can occur from the tosses.

TABLE 3.1:

Toss	First Coin	Second Coin	$X$
1	H	H	0
2	H	T	1
3	T	T	2
4	T	H	1

We can add a fifth column to the table above to show the probability of each of these events (the tossing of the 2 coins).

TABLE 3.2:

Toss	First Coin	Second Coin	$X$	$P(X)$
1	H	H	0	$\frac{1}{4}$
2	H	T	1	$\frac{1}{4}$
3	T	T	2	$\frac{1}{4}$
4	T	H	1	$\frac{1}{4}$

As you can see in the table, each event has an equally likely chance of occurring. You can see this by looking at the column  $P(X)$ . From here, we can find the probability distribution. In the  $X$  column, we have 3 possible discrete values for this variable: 0, 1, and 2.

### 3.2. The Probability Distribution

$$P(0) = \text{toss } 1 = \frac{1}{4}$$

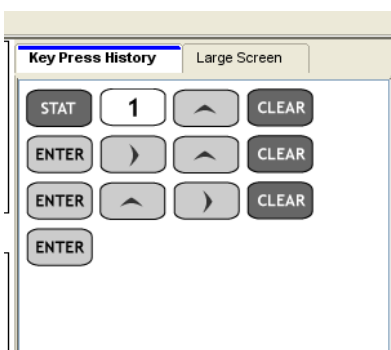
$$P(1) = \text{toss } 2 + \text{toss } 4$$

$$= \frac{1}{4} + \frac{1}{4}$$

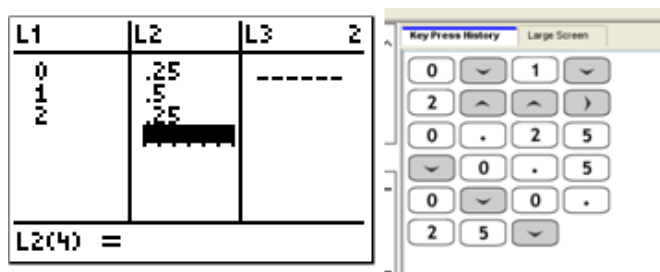
$$= \frac{1}{2}$$

$$P(2) = \text{toss } 3 = \frac{1}{4}$$

Now we can represent the probability distribution with a graph, called a histogram. A **histogram** is a graph that uses bars vertically arranged to display data. Using the TI-84 PLUS calculator, we can draw the histogram to represent the data above. Let's start by first adding the data into our lists. Below you will find the key sequence to perform this task. We will use this sequence frequently throughout the rest of this book, so make sure you follow along with your calculator.

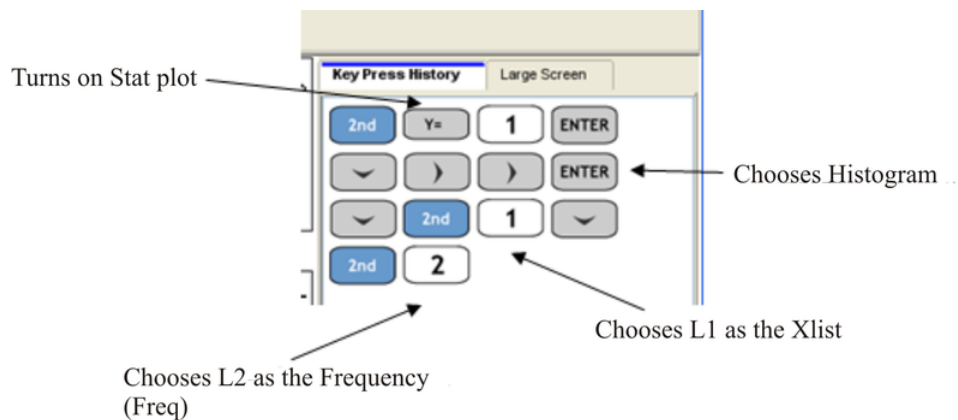


This key sequence allows you to erase any data that may be entered into the lists already. Now let's enter our data.

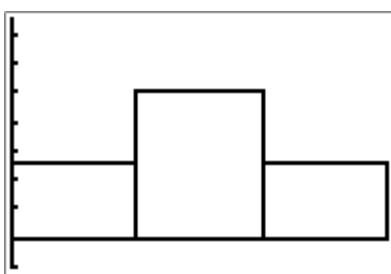


Now we can draw our histogram from the data we just entered.

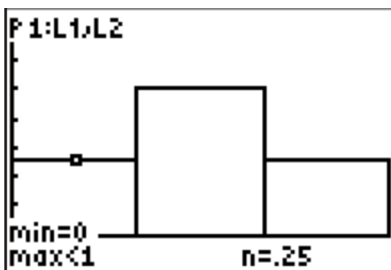




The result is as follows:



We can see the values of  $P(X)$  if we press **TRACE**. Look at the screenshot below. You can see the value of  $P(X) = 0.25$  for  $X = 0$ .



It's clear that the histogram shows the probability distribution for the discrete random variable. In other words,  $P(0)$  gives the probability that the discrete random variable is equal to 0,  $P(1)$  gives the probability that the discrete random variable is equal to 1, and  $P(2)$  gives the probability that the discrete random variable is equal to 2. Notice that the probabilities add up to 1. One of the rules for probability is that the sum of the probabilities of all the possible values of a discrete random variable must be equal to 1.

### Example 2

Does the following table represent the probability distribution for a discrete random variable?

$X$	0	1	2	3
$P(X)$	0.1	0.2	0.3	0.4

### Solution:

Yes, it does, since  $\sum P(X) = 0.1 + 0.2 + 0.3 + 0.4$ , or  $\sum P(X) = 1.0$ .

### 3.2. The Probability Distribution

## 3.3 A Glimpse at Binomial and Multinomial Distributions

In Chapter 4, we will learn more about binomial and multinomial distributions. However, we are now talking about probability distributions, and as such, we should at least see how the problems change for these distributions. We will briefly introduce the concepts and their formulas here, and then we will get into more detail in Chapter 4. Let's start with a problem involving a binomial distribution.

### Example 3

The probability of scoring above 75% on a math test is 40%. What is the probability of scoring below 75%?

#### Solution:

$$P(\text{scoring above } 75\%) = 0.40$$

$$\text{Therefore, } P(\text{scoring below } 75\%) = 1 - 0.40 = 0.60.$$

The randomness of an individual outcome occurs when we take 1 event and repeat it over and over again. One example is if you were to flip a coin multiple times. In order to calculate the probability of this type of event, we need to look at one more formula.

The probability of getting  $x$  successes in  $n$  trials is given by:

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$

where:

$a$  is the number of successes from the trials.

$p$  is the probability of the event occurring.

$q$  is the probability of the event not occurring.

Now, remember that in Chapter 2, you learned about the formula  ${}_n C_r$ . The formula is shown below:

$${}_n C_r = \frac{n!}{r!(n-r)!}$$

Also, recall that the symbol **!** means factorial. As a review, the **factorial function (!)** just means to multiply a series of consecutive descending natural numbers.

Examples:

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

$$1! = 1$$

Note: it is generally agreed that  $0! = 1$ .

**Technology Tip:** You can find the factorial function using:

MATH ►►► (PRB) ▼▼▼ (4)

Now let's try a few problems with the binomial distribution formula.

**Example 4**

A fair die is rolled 10 times. Let  $X$  be the number of rolls in which we see a 2.

- (a) What is the probability of seeing a 2 in any one of the rolls?  
 (b) What is the probability of seeing a 2 exactly once in the 10 rolls?

**Solution:**

(a)  $P(X) = \frac{1}{6} = 0.167$

(b)  $p = 0.167$

$q = 1 - 0.167 = 0.833$

$n = 10$

$a = 1$

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$

$$P(X = 1) = {}_{10} C_1 \times p^1 \times q^{(10-1)}$$

$$P(X = 1) = {}_{10} C_1 \times (0.167)^1 \times (0.833)^{(10-1)}$$

$$P(X = 1) = 10 \times 0.167 \times 0.193$$

$$P(X = 1) = 0.322$$

Therefore, the probability of seeing a 2 exactly once when a die is rolled 10 times is 32.2%.

Interestingly, it was Blaise Pascal (pictured below) with Pierre de Fermat who provided the world with the basics of probability. These 2 mathematicians studied many different theories in mathematics, one of which was odds and probability. To learn more about Pascal, go to [http://en.wikipedia.org/wiki/Blaise\\_Pascal](http://en.wikipedia.org/wiki/Blaise_Pascal). To learn more about Fermat, go to <http://en.wikipedia.org/wiki/Fermat>. These 2 mathematicians have contributed greatly to the world of mathematics.



**Example 5**

A fair die is rolled 15 times. What is the probability of rolling two 2's?

- (a) What is the probability of seeing a 2 in any one of the rolls?  
 (b) What is the probability of seeing a 2 exactly twice in the 15 rolls?

**Solution:**

(c)  $P(X) = \frac{1}{6} = 0.167$

3.3. A Glimpse at Binomial and Multinomial Distributions

$$\begin{aligned} \text{(d) } p &= 0.167 \\ q &= 1 - 0.167 = 0.833 \\ n &= 15 \\ a &= 2 \end{aligned}$$

$$\begin{aligned} P(X = a) &= {}_n C_a \times p^a \times q^{(n-a)} \\ P(X = 2) &= {}_{15} C_2 \times p^2 \times q^{(15-2)} \\ P(X = 2) &= {}_{15} C_2 \times (0.167)^2 \times (0.833)^{(15-2)} \\ P(X = 2) &= 105 \times 0.0279 \times 0.0930 \\ P(X = 2) &= 0.272 \end{aligned}$$

Therefore, the probability of seeing a 2 exactly twice when a die is rolled 15 times is 27.2%.

### Example 6

A pair of fair dice is rolled 10 times. Let  $X$  be the number of rolls in which we see at least one 2.

- (a) What is the probability of seeing at least one 2 in any one roll of the pair of dice?  
 (b) What is the probability that in exactly half of the 10 rolls, we see at least one 2?

### Solution:

If we look at the chart below, we can see the number of times a 2 shows up when rolling 2 dice.

+	1	2	3	4	5	6
1	1,1	2,1	3,1	4,1	5,1	6,1
2	1,2	2,2	3,2	4,2	5,2	6,2
3	1,3	2,3	3,3	4,3	5,3	6,3
4	1,4	2,4	3,4	4,4	5,4	6,4
5	1,5	2,5	3,5	4,5	5,5	6,5
6	1,6	2,6	3,6	4,6	5,6	6,6

- (a) The probability of seeing at least one 2 in any one roll of the pair of dice is:

$$P(X) = \frac{11}{36} = 0.306$$

- (b) The probability of seeing at least one 2 in exactly 5 of the 10 rolls is calculated as follows:

$$\begin{aligned} p &= 0.306 \\ q &= 1 - 0.306 = 0.694 \\ n &= 10 \\ a &= 5 \end{aligned}$$

$$\begin{aligned}
 P(X = a) &= {}_n C_a \times p^a \times q^{(n-a)} \\
 P(X = 5) &= {}_{10} C_5 \times p^5 \times q^{(10-5)} \\
 P(X = 5) &= {}_{10} C_5 \times (0.306)^5 \times (0.694)^{(10-5)} \\
 P(X = 5) &= 252 \times 0.00268 \times 0.161 \\
 P(X = 5) &= 0.109
 \end{aligned}$$

Therefore, the probability of rolling at least one 2 exactly 5 times when 2 dice are rolled 10 times is 10.9%.

It should be noted here that the previous 2 examples are examples of binomial experiments. We will be learning more about binomial experiments and distributions in Chapter 4. For now, we can visualize a **binomial distribution** experiment as one that has a fixed number of trials, with each trial being independent of the others. In other words, rolling a die twice to see if a 2 appears is a binomial experiment, because there is a fixed number of trials (2), and each roll is independent of the others. Also, for binomial experiments, there are only 2 possible outcomes (a successful event and a non-successful event). For our rolling of the die, a successful event is seeing a 2, and a non-successful event is not seeing a 2.

### Example 7

You are given a bag of marbles. Inside the bag are 5 red marbles, 4 white marbles, and 3 blue marbles. Calculate the probability that with 6 trials, you choose 3 marbles that are red, 1 marble that is white, and 2 marbles that are blue, replacing each marble after it is chosen.

### Solution:

Notice that this is not a binomial experiment, since there are more than 2 possible outcomes. For binomial experiments,  $k = 2$  (2 outcomes). Therefore, we use the binomial experiment formula for problems involving heads or tails, yes or no, or success or failure. In this problem, there are 3 possible outcomes: red, white, or blue. This type of experiment produces what we call a **multinomial distribution**. In order to solve this problem, we need to use one more formula:

$$P = \frac{n!}{n_1! n_2! n_3! \dots n_k!} \times (p_1^{n_1} \times p_2^{n_2} \times p_3^{n_3} \dots p_k^{n_k})$$

where:

$n$  is the number of trials.

$p$  is the probability for each possible outcome.

$k$  is the number of possible outcomes.

Notice that in this example,  $k$  equals 3. If we had only red marbles and white marbles,  $k$  would be equal to 2, and we would have a binomial distribution.

The probability of choosing 3 red marbles, 1 white marble, and 2 blue marbles in exactly 6 picks is calculated as follows:

### 3.3. A Glimpse at Binomial and Multinomial Distributions

$$n = 6 \text{ (6 picks)}$$

$$p_1 = \frac{5}{12} = 0.416 \text{ (probability of choosing a red marble)}$$

$$p_2 = \frac{4}{12} = 0.333 \text{ (probability of choosing a white marble)}$$

$$p_3 = \frac{3}{12} = 0.25 \text{ (probability of choosing a blue marble)}$$

$$n_1 = 3 \text{ (3 red marbles chosen)}$$

$$n_2 = 1 \text{ (1 white marble chosen)}$$

$$n_3 = 2 \text{ (2 blue marbles chosen)}$$

$$k = 3 \text{ (3 possibilities)}$$

$$P(x = 6) = \frac{n!}{n_1!n_2!n_3!\dots n_k!} \times (p_1^{n_1} \times p_2^{n_2} \times p_3^{n_3} \dots p_k^{n_k})$$

$$P(x = 6) = \frac{6!}{3!1!2!} \times (0.416^3 \times 0.333^1 \times 0.25^2)$$

$$P(x = 6) = 60 \times 0.0720 \times 0.333 \times 0.0625$$

$$P(x = 6) = 0.0899$$

Therefore, the probability of choosing 3 red marbles, 1 white marble, and 2 blue marbles is 8.99%.

### Example 8

You are randomly drawing cards from an ordinary deck of cards. Every time you pick one, you place it back in the deck. You do this 5 times. What is the probability of drawing 1 heart, 1 spade, 1 club, and 2 diamonds?

### Solution:

$$n = 5 \text{ (5 trials)}$$

$$p_1 = \frac{13}{52} = 0.25 \text{ (probability of drawing a heart)}$$

$$p_2 = \frac{13}{52} = 0.25 \text{ (probability of drawing a spade)}$$

$$p_3 = \frac{13}{52} = 0.25 \text{ (probability of drawing a club)}$$

$$p_4 = \frac{13}{52} = 0.25 \text{ (probability of drawing a diamond)}$$

$$n_1 = 1 \text{ (1 heart)}$$

$$n_2 = 1 \text{ (1 spade)}$$

$$n_3 = 1 \text{ (1 club)}$$

$$n_4 = 2 \text{ (2 diamonds)}$$

$$k = 1 \text{ (for each trial, there is only 1 possible outcome)}$$

$$P(x = 5) = \frac{n!}{n_1!n_2!n_3!\dots n_k!} \times (p_1^{n_1} \times p_2^{n_2} \times p_3^{n_3} \dots p_k^{n_k})$$

$$P(x = 5) = \frac{5!}{1!1!1!2!} \times (0.25^1 \times 0.25^1 \times 0.25^1 \times 0.25^2)$$

$$P(x = 5) = 60 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

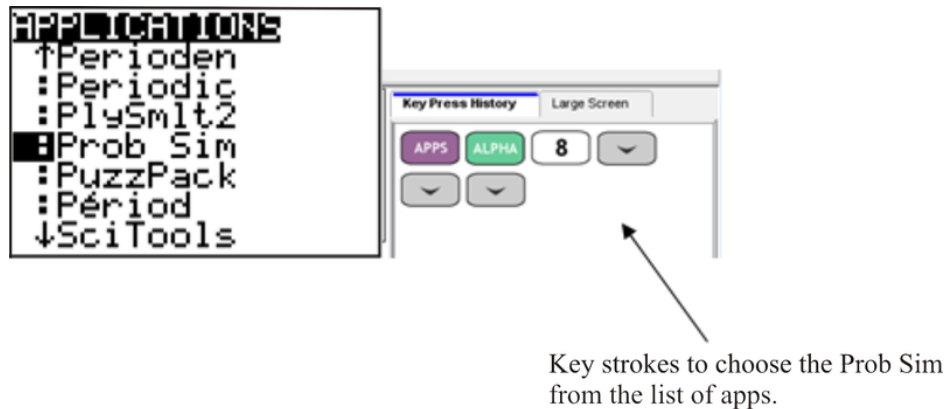
$$P(x = 5) = 0.0586$$

Therefore, the probability of choosing 1 heart, 1 spade, 1 club, and 2 diamonds is 5.86%.

## 3.4 Using Technology to Find Probability Distributions

If we look back at Example 1, we were tossing 2 coins. If you were to repeat this experiment 100 times, or if you were going to toss 10 coins 50 times, these experiments would be very tiring and take a great deal of time. On the TI-84 calculator, there are applications built in to determine the probability of such experiments. In this section, we will look at how you can use your graphing calculator to calculate probabilities for larger trials and draw the corresponding histograms.

On the TI-84 calculator, there are a number of possible simulations you can do. You can do a coin toss, spin a spinner, roll dice, pick marbles from a bag, or even draw cards from a deck.



After pressing **ENTER**, you will have the following screen appear.



Let's try a spinner problem. Choose Spin Spinner.



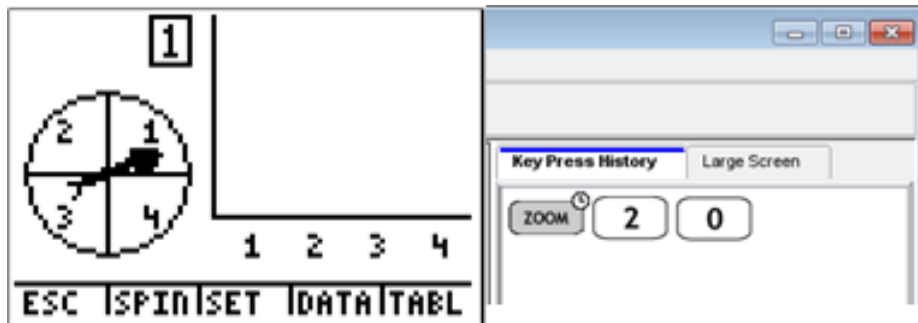
### Example 9

You are spinning a spinner like the one shown below 20 times. How many times does it land on blue?



**Solution:**

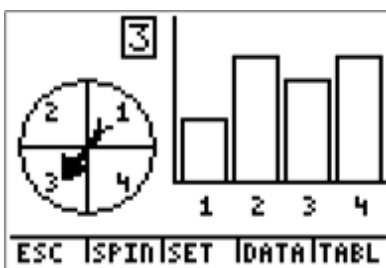
In Spin Spinner, a wheel with 4 possible outcomes is shown. You can adjust the number of spins, graph the frequency of each number, and use a table to see the number of spins for each number. Let's try this problem. We want to set this spinner to spin 20 times. Look at the keystrokes below and see how this is done.



In order to match our color spinner with the one found in the calculator, you will see that we have added numbers to our spinner. This is not necessary, but it may help in the beginning to remember that 1 = blue (for this example).

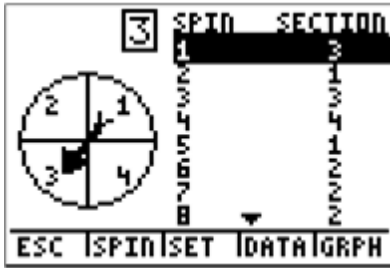


Now that the spinner is set up for 20 trials, choose SPIN by pressing `WINDOW`.



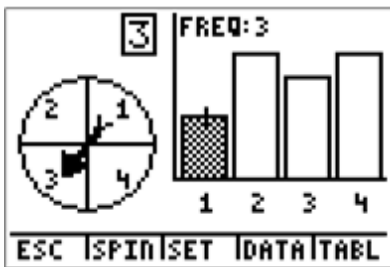
### 3.4. Using Technology to Find Probability Distributions

We can see the result of each trial by choosing TABL, or pressing **GRAPH**.



And we see the graph of the resulting table, or go back to the first screen, simply by choosing GRPH, or pressing **GRAPH** again.

Now, the question asks how many times we landed on blue (number 1). We can actually see how many times we landed on blue for these 20 spins. If you press the right arrow (**▶**), the frequency label will show you how many of the times the spinner landed on blue (number 1).



To go back to the question, how many times does the spinner land on blue if it is spun 20 times? The answer is 3. To calculate the probability of landing on blue, we have to divide by the total number of spins.

$$P(\text{blue}) = \frac{3}{20} = 0.15$$

Therefore, for this experiment, the probability of landing on blue with 20 spins is 15%.

The above example introduces us to a new concept. We know that the spinner has 4 equal parts (blue, purple, green, and red). In a single trial, we can assume that:

$$P(\text{blue}) = \frac{1}{4} = 0.25$$

However, we know that we did the experiment and found that the probability of landing on blue, if the spinner is spun 20 times, is 0.15. Why the difference?

The difference between these 2 numbers has to do with the difference between theoretical and experimental probability. **Theoretical probability** is defined as the number of desired outcomes divided by the total number of outcomes.

### Theoretical Probability

$$P(\text{desired}) = \frac{\text{number of desired outcomes}}{\text{total number of outcomes}}$$

For our spinner example, the theoretical probability of landing on blue is 0.25. Finding the theoretical probability requires no collection of data.

In the case of the experiment of spinning the spinner 20 times, the probability of 0.15, found by counting the number of times the spinner landed on blue, is called the experimental probability. **Experimental probability** is, just as the name suggests, dependent on some form of data collection. To calculate the experimental probability, divide the number of times the desired outcome has occurred by the total number of trials.

### Experimental Probability

$$P(\text{desired}) = \frac{\text{number of times desired outcome occurs}}{\text{total number of trials}}$$

You can try a lot of examples and trials yourself using the NCTM Illuminations page found at <http://illuminations.nctm.org/activitydetail.aspx?ID=79>.

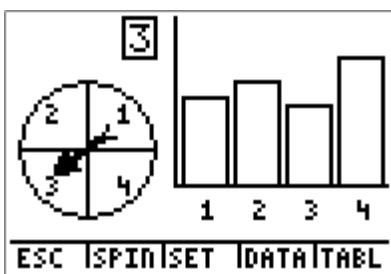
What is interesting about theoretical and experimental probabilities is that the more trials you do, the closer the experimental probability gets to the theoretical probability. To show this, try spinning the spinner for the next example.

#### Example 10

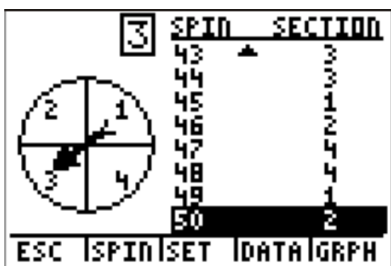
You are spinning a spinner like the one shown to the right 50 times. How many times does it land on blue?

#### Solution:

Set the spinner to spin 50 times and choose SPIN by pressing **WINDOW**.



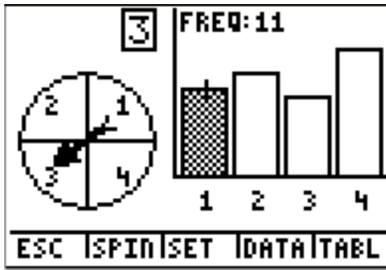
You can see the result of each trial by choosing TABL, or pressing **GRAPH**.



Again, we can see the graph of the resulting table, or go back to the first screen, simply by choosing GRPH, or pressing **GRAPH** again.

The question asks how many times we landed on blue (number 1) for the 50 spins. Press the right arrow (**▶**), and the frequency label will show you how many of the times the spinner landed on blue (number 1).

### 3.4. Using Technology to Find Probability Distributions

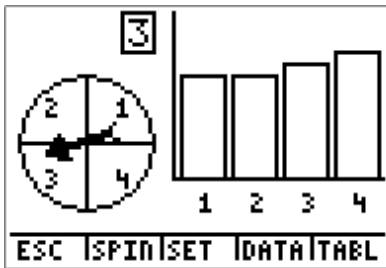


Now go back to the question. How many times does the spinner land on blue if it is spun 50 times? The answer is 11. To calculate the probability of landing on blue, we have to divide by the total number of spins.

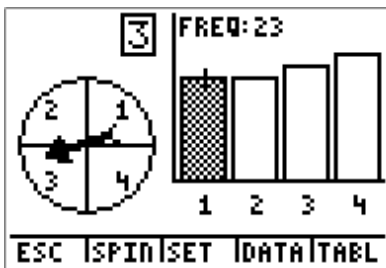
$$P(\text{blue}) = \frac{11}{50} = 0.22$$

Therefore, for this experiment, the probability of landing on blue with 50 spins is 22%.

If we tried 100 trials, we would see something like the following:



In this case, we see that the frequency of 1 is 23.



So how many times does the spinner land on blue if it is spun 100 times? The answer is 23. To calculate the probability of landing on blue in this case, we again have to divide by the total number of spins.

$$P(\text{blue}) = \frac{23}{100} = 0.23$$

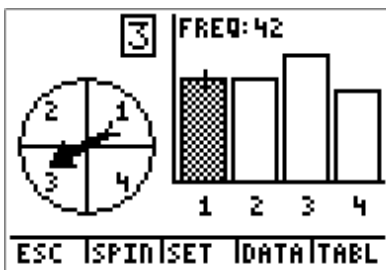
Therefore, for this experiment, the probability of landing on blue with 100 spins is 23%. You can see that as we perform more trials, we get closer to 25%, which is the theoretical probability.

### Example 11

How many times do you predict we would have to spin the spinner in Example 10 to have the experimental probability equal the theoretical probability?

**Solution:**

With 170 spins, we get a frequency of 42 for blue.



The experimental probability in this case can be calculated as follows:

$$P(\text{blue}) = \frac{42}{170} = 0.247$$

Therefore, the experimental probability is 24.7%, which is even closer to the theoretical probability of 25%. While we're getting closer to the theoretical probability, there is no number of trials that will guarantee that the experimental probability will exactly equal the theoretical probability.

Let's try an example using the coin toss simulation.

**Example 12**

A fair coin is tossed 50 times. What is the theoretical probability and the experimental probability of tossing tails on the fair coin?

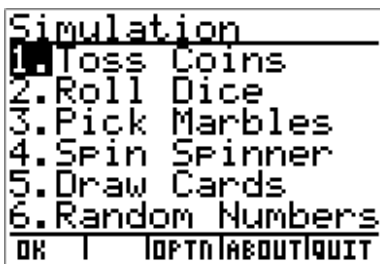
**Solution:**

To calculate the theoretical probability, we need to remember that the probability of getting tails is  $\frac{1}{2}$ , or:

$$P(\text{tails}) = \frac{1}{2} = 0.50$$

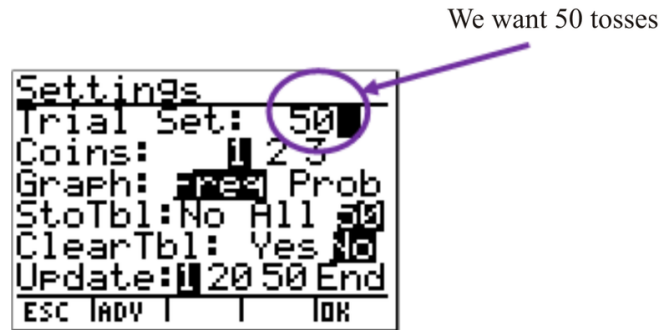
To find the experimental probability, we need to run the coin toss simulation in the probability simulator. We could also actually take a coin and flip it 50 times, each time recording if we get heads or tails.

If we follow the same keystrokes to get into the Prob Sim app, we get to the main screen.

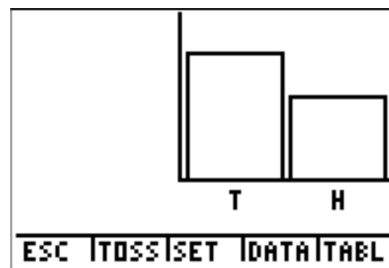


Choose 1. Toss Coins and then press **ZOOM**.

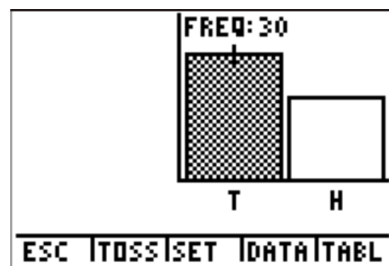
### 3.4. Using Technology to Find Probability Distributions



Choose OK by pressing **GRAPH** and go back to the main screen. Then choose TOSS by pressing **WINDOW**.



To find the frequency, we need to press the **▶** arrow to view the frequency for the tossing experiment.



We see the frequency of tails is 30. Now we can calculate the experimental probability.

$$P(\text{tails}) = \frac{30}{50} = 0.60$$

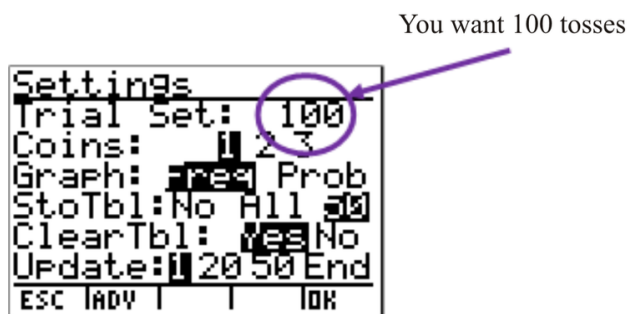
### Example 13

What if the fair coin is tossed 100 times? What is the experimental probability? Is the experimental probability getting closer to the theoretical probability?

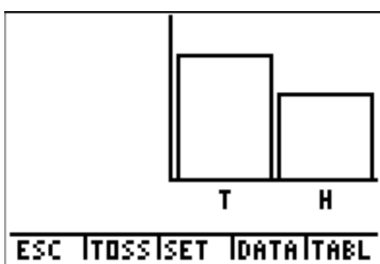
#### Solution:

To find the experimental probability for this example, we need to run the coin toss in the probability simulator again. You could also, like in Example 12, actually take a coin and flip it 100 times, each time recording if you get heads or tails. You can see how the technology is going to make this experiment take a lot less time.

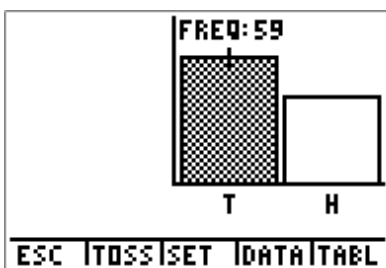
Choose 1. Toss Coins and then press **ZOOM**.



Choose OK by pressing **GRAPH** and go back to the main screen. Then choose TOSS by pressing **WINDOW**.



To find the frequency, we need to press the **▸** arrow to view the frequency for the tossing experiment.



Notice that the frequency of tails is 59. Now you can calculate the experimental probability.

$$P(\text{tails}) = \frac{59}{100} = 0.59$$

With 50 tosses, the experimental probability of tails was 60%, and with 100 tosses, the experimental probability of tails was 59%. This means that the experimental probability is getting closer to the theoretical probability of 50%.

You can also use this same program to toss 2 coins or 5 coins. Actually, you can use this simulation to toss any number of coins any number of times.

#### Example 14

2 fair coins are tossed 10 times. What is the theoretical probability of both coins landing on heads? What is the experimental probability of both coins landing on heads?

#### Solution:

The theoretical probability of getting heads on the first coin is  $\frac{1}{2}$ . Flipping the second coin, the theoretical probability of getting heads is again  $\frac{1}{2}$ . The overall theoretical probability is  $(\frac{1}{2})^2$  for 2 coins, or:

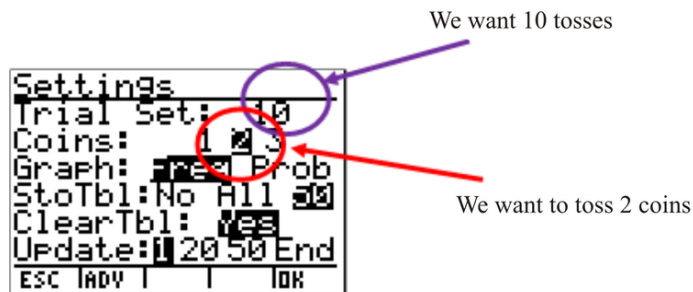
#### 3.4. Using Technology to Find Probability Distributions

$$P(2H) = \frac{1}{2} \times \frac{1}{2}$$

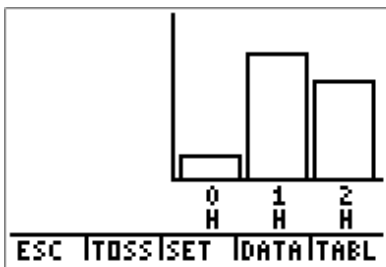
$$P(2H) = \left(\frac{1}{2}\right)^2$$

$$P(2H) = \frac{1}{4}$$

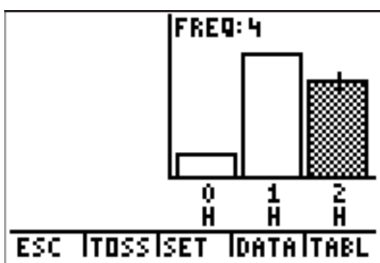
To determine the experimental probability, let's go to the probability simulator. Again, you can also do this experiment manually by taking 2 coins, tossing them 10 times, and recording your observations.



Now let's toss the coins.



Find the frequency of getting 2 heads ( $2H$ ).



The frequency is equal to 4. Therefore, for 2 coins tossed 10 times, there were 4 times that both coins landed on heads. You can now calculate the experimental probability.

$$P(2H) = \frac{4}{10}$$

$$P(2H) = 0.40 \text{ or } 40\%$$

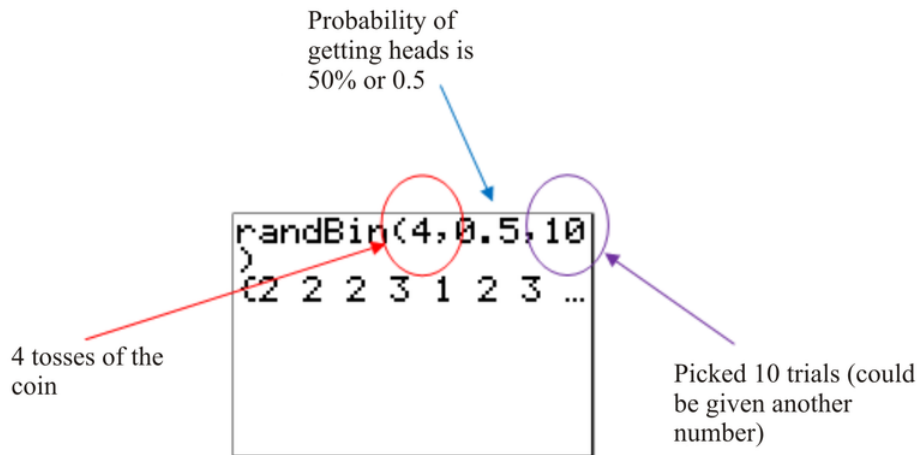


To try another type of probability simulation, you can use the Texas Instruments Activities Exchange. Look up simple probability simulations on <http://education.ti.com/educationportal/activityexchange/Activity.do?cid=US&aId=9327>.

You can also use the randBin function on your calculator to simulate the tossing of a coin. The randBin function is used to produce experimental values for discrete random variables. You can find the randBin function using:

**MATH**  $\blacktriangleright$   $\blacktriangleright$   $\blacktriangleright$  (PRB)  $\blacktriangledown$   $\blacktriangledown$   $\blacktriangledown$   $\blacktriangledown$   $\blacktriangledown$   $\blacktriangledown$  (7)

If you wanted to toss 4 coins 10 times, you would enter the command below:



The list that is produced contains the count of heads resulting from each set of 4 coin tosses. If you use the right arrow ( $\blacktriangleright$ ), you can see how many times from the 10 trials you actually had 4 heads.

### Example 15

You are in math class. Your teacher asks what the probability is of obtaining 5 heads if you were to toss 15 coins.

- Determine the theoretical probability for the teacher.
- Use the TI calculator to determine the actual probability for a trial experiment of 10 trials.

### Solution:

(a) Let's calculate the theoretical probability of getting 5 heads in the 15 tosses. In order to do this type of calculation, let's bring back the concept of factorial from an earlier lesson.

### Numerator (Top)

In the example, you want to have 5 H's and 10 T's. Our favorable outcomes would be HHHHHTTTTTTTTTT, with the H's and T's coming in any order. The number of favorable outcomes would be:

$$\begin{aligned} \text{number of favorable outcomes} &= \frac{\text{number of tosses!}}{\text{number of heads!} \times \text{number of tails!}} \\ \text{number of favorable outcomes} &= \frac{15!}{5! \times 10!} \\ \text{number of favorable outcomes} &= \frac{15 \times 14 \times 13 \times 12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(5 \times 4 \times 3 \times 2 \times 1) \times (10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)} \\ \text{number of favorable outcomes} &= \frac{1.31 \times 10^{12}}{120 \times 3628800} \\ \text{number of favorable outcomes} &= 3003 \end{aligned}$$

### 3.4. Using Technology to Find Probability Distributions

*Denominator (Bottom)*

The number of possible outcomes =  $2^{15}$

The number of possible outcomes = 32,768

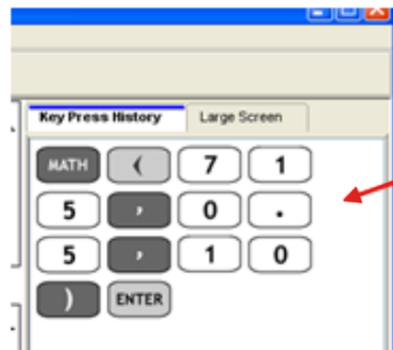
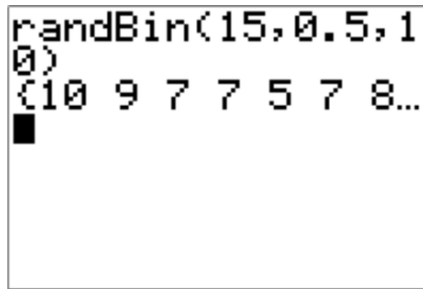
Now you just divide the numerator by the denominator:

$$P(5 \text{ heads}) = \frac{3003}{32768}$$

$$P(5 \text{ heads}) = 0.0916$$

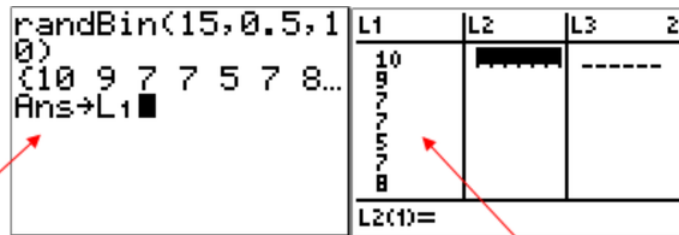
Therefore, the theoretical probability would be 9.16% of getting 5 heads when tossing 15 coins.

b) To calculate the experimental probability, let's use the randBin function on the TI-84 calculator.



Key Strokes for this example

From the list, you can see that you only have 5 heads 1 time in the 10 trials.



Storing the information into List 1 (L<sub>1</sub>)

L<sub>1</sub>

Therefore, the experimental probability can be calculated as follows:

$$P(5 \text{ heads}) = \frac{1}{10} = 10\%$$

## Points to Consider

- How is the calculator a useful tool for calculating probabilities in discrete random variable experiments?
- How are these experimental probabilities different from what you would expect the theoretical probabilities to be? When can the 2 types of probability possibly be equal?

## Vocabulary

**Binomial distribution** A distribution produced by an experiment where there is a fixed number of successes in  $X$  (random variable) trials, and each trial is independent of the others.

**Discrete random variables** Only have a specific (or finite) number of numerical values within a certain range.

**Experimental probability** The actual probability of an event resulting from an experiment.

**Factorial function (!)** The function of multiplying a series of consecutive descending natural numbers.

**Histogram** A graph that uses vertically arranged bars to display data.

**Multinomial distribution** A distribution produced by an experiment where the number of possible outcomes is greater than 2 and where each outcome has a specific probability.

**Probability distribution** A table, a graph, or a chart that shows you all the possible values of  $X$  (your variable), and the probability associated with each of these values ( $P(X)$ ).

**Random variables** Variables that take on numerical values governed by a chance experiment.

**Theoretical probability** A probability that is the ratio of the number of different ways an event can occur to the total number of equally likely possible outcomes. The numerical measure of the likelihood that an event,  $E$ , will happen.

$$P(E) = \frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}}$$

## 3.5 Review Questions

Answer the following questions and show all work (including diagrams) to create a complete answer.

- Match the following statements from the first column with the probability values in the second column.

**TABLE 3.3:**

Probability Statement	$P(X)$
a. The probability of this event will never occur.	___ $P(X) = 1.0$
b. The probability of this event is highly likely.	___ $P(X) = 0.33$
c. The probability of this event is very likely.	___ $P(X) = 0.67$
d. The probability of this event is somewhat likely.	___ $P(X) = 0.00$
e. The probability of this event is certain.	___ $P(X) = 0.95$

- Match the following statements from the first column with the probability values in the second column.

**TABLE 3.4:**

Probability Statement	$P(X)$
a. I bought a ticket for the State Lottery. The probability of a successful event (winning) is likely to be:	___ $P(X) = 0.80$
b. I have a bag of equal numbers of red and green jelly beans. The probability of reaching into the bag and picking out a red jelly bean is likely to be:	___ $P(X) = 0.50$
c. My dad teaches math, and my mom teaches chemistry. The probability that I will be expected to study science or math is likely to be:	___ $P(X) = 0.67$
d. Our class has the highest test scores in the State Math Exams. The probability that I have scored a great mark is likely to be:	___ $P(X) = 1.0$
e. The Chicago baseball team has won every game this season. The probability that the team will make it to the playoffs is likely to be:	___ $P(X) = 0.01$

- Read each of the following statements and match the following words to each statement. You can put your answers directly into the table. Here is the list of terms you can add:

- certain or sure
- impossible
- likely or probable
- unlikely or improbable
- maybe
- uncertain or unsure

TABLE 3.5:

Statement	Probability Term
Tomorrow is Friday.	
I will be in New York on Friday.	
It will be dark tonight.	
It is snowing in August!	
China is cold in January.	

4. Read each of the following statements and match the following words to each statement. You can put your answers directly into the table. Here is the list of terms you can add:

- certain or sure
- impossible
- likely or probable
- unlikely or improbable
- maybe
- uncertain or unsure

TABLE 3.6:

Statement	Probability Term
I am having a sandwich for lunch.	
I have school tomorrow.	
I will go to the movies tonight.	
January is warm in New York.	
My dog will bark.	

5. The probability of scoring above 80% on a math test is 20%. What is the probability of scoring below 80%?
6. The probability of getting a job after university is 85%. What is the probability of not getting a job after university?
7. Does the following table represent the probability distribution for a discrete random variable?

$X$	2	4	6	8
$P(X)$	0.2	0.4	0.6	0.8

8. Does the following table represent the probability distribution for a discrete random variable?

$X$	1	2	3	4	5
$P(X)$	0.202	0.174	0.096	0.078	0.055

9. Does the following table represent the probability distribution for a discrete random variable?

$X$	1	2	3	4	5	6
$P(X)$	0.302	0.251	0.174	0.109	0.097	0.067

10. A fair die is rolled 10 times. Let  $X$  be the number of rolls in which we see a 6.
- a. What is the probability of seeing a 6 in any one of the rolls?
  - b. What is the probability that we will see a 6 exactly once in the 10 rolls?
11. A fair die is rolled 15 times. Let  $X$  be the number of rolls in which we see a 6.
- a. What is the probability of seeing a 6 in any one of the rolls?

- b. What is the probability that we will see a 6 exactly once in the 15 rolls?
12. A fair die is rolled 15 times. Let  $X$  be the number of rolls in which we see a 5.
- What is the probability of seeing a 5 in any one of the rolls?
  - What is the probability that we will see a 5 exactly 7 times in the 15 rolls?
13. A pair of fair dice is rolled 10 times. Let  $X$  be the number of rolls in which we see at least one 5.
- What is the probability of seeing at least one 5 in any one roll of the pair of dice?
  - What is the probability that in exactly half of the 10 rolls, we see at least one 5?
14. A pair of fair dice is rolled 15 times. Let  $X$  be the number of rolls in which we see at least one 5.
- What is the probability of seeing at least one 5 in any one roll of the pair of dice?
  - What is the probability that in exactly 8 of the 15 rolls, we see at least one 5?
15. You are randomly drawing cards from an ordinary deck of cards. Every time you pick one, you place it back in the deck. You do this 7 times. What is the probability of drawing 2 hearts, 2 spades, 2 clubs, and 3 diamonds?
16. A telephone survey measured the percentage of students in ABC town who watch channels NBX, FIX, MMA, and TSA. After the survey, analysis showed that 35 percent watch channel NBX, 40 percent watch channel FIX, 10 percent watch channel MMA, and 15 percent watch channel TSA. What is the probability that from 7 randomly selected students, 1 will be watching channel NBX, 2 will be watching channel FIX, 3 will be watching channel MMA, and 2 will be watching channel TSA?
17. Based on what you know about probabilities, write definitions for *theoretical* and *experimental* probability.
- What is the difference between theoretical and experimental probability?
  - As you add more data, do your experimental probabilities get closer to or further away from your theoretical probabilities?
  - Is tossing 1 coin 100 times the same as tossing 100 coins 1 time? Why or why not?
18. Use the randBin function on your calculator to simulate 5 tosses of a coin 25 times to determine the probability of getting 2 tails.
19. Use the randBin function on your calculator to simulate 10 tosses of a coin 50 times to determine the probability of getting 4 heads.
20. Calculate the theoretical probability of getting 3 heads in 10 tosses of a coin.
21. Find the experimental probability using technology of getting 3 heads in 10 tosses of a coin.
22. Calculate the theoretical probability of getting 8 heads in 12 tosses of a coin.
23. Calculate the theoretical probability of getting 7 heads in 14 tosses of a coin.

**CHAPTER 4**

# Probability Distributions

## Chapter Outline

---

- 4.1**    **NORMAL DISTRIBUTIONS**
  - 4.2**    **BINOMIAL DISTRIBUTIONS**
  - 4.3**    **EXPONENTIAL DISTRIBUTIONS**
  - 4.4**    **REVIEW QUESTIONS**
- 

### Introduction

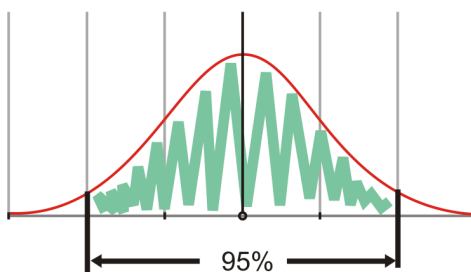
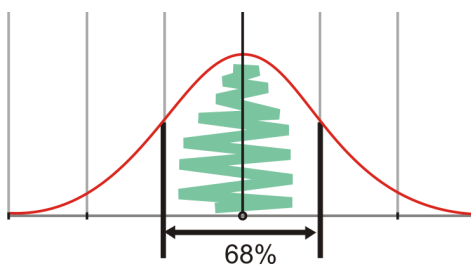
For a standard normal distribution, the data presented is continuous. In addition, the data is centered at the mean and is symmetrically distributed on either side of that mean. This means that the resulting data forms a shape similar to a bell and is, therefore, called a bell curve. Binomial experiments are discrete probability experiments that involve a fixed number of independent trials, where there are only 2 outcomes. As a rule of thumb, these trials result in successes and failures, and the probability of success for one trial is the same as for the next trial (i.e., independent events). As the sample size increases for a binomial distribution, the resulting histogram approaches the appearance of a normal distribution curve. With this increase in sample size, the accuracy of the distribution also increases. An exponential distribution is a distribution of continuous data, and the general equation is in the form  $y = ab^x$ . The closer the correlation coefficient is to 1, the more likely the equation for the exponential distribution is accurate.

## 4.1 Normal Distributions

### Learning Objectives

- Be familiar with the standard distributions (normal, binomial, and exponential).
- Use standard distributions to solve for events in problems in which the distribution belongs to one of these families.

In Chapter 3, you spent some time learning about probability distributions. A **distribution**, itself, is simply a description of the possible values of a random variable and the possible occurrences of these values. Remember that probability distributions show you all the possible values of your variable ( $X$ ), and the probability associated with each of these values ( $P(X)$ ). You were also introduced to the concept of binomial distributions, or distributions of experiments where there are a fixed number of successes in  $X$  (random variable) trials, and each trial is independent of the other. In addition, you were introduced to binomial distributions in order to compare them with multinomial distributions. Remember that multinomial distributions involve experiments where the number of possible outcomes is greater than 2, and the probability is calculated for each outcome for each trial.



In this first lesson on probability distributions, you are going to begin by learning about normal distributions. A **normal distribution curve** can be easily recognized by its shape. The first 2 diagrams above show examples of normal distributions. What shape do they look like? Do they look like a bell to you? Compare the first 2 diagrams above to the third diagram. A normal distribution is called a *bell curve* because its shape is comparable to a bell. It

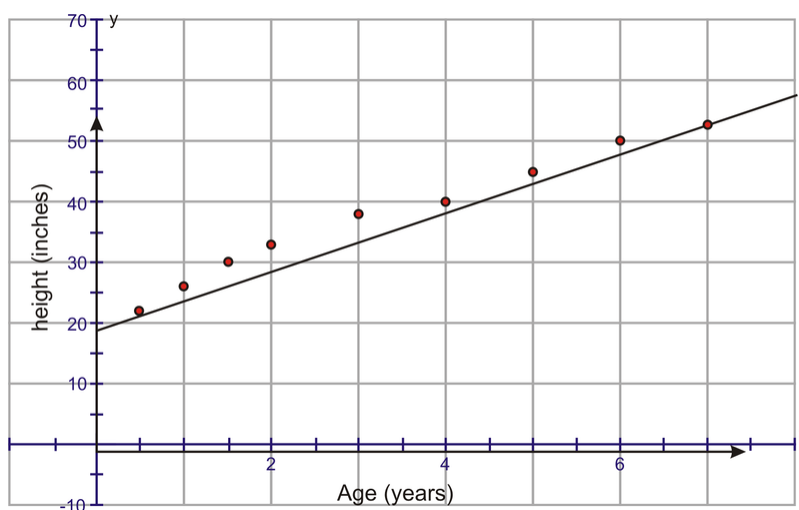


has this shape because the majority of the data is concentrated at the middle and slowly decreases symmetrically on either side. This gives it a shape similar to a bell.

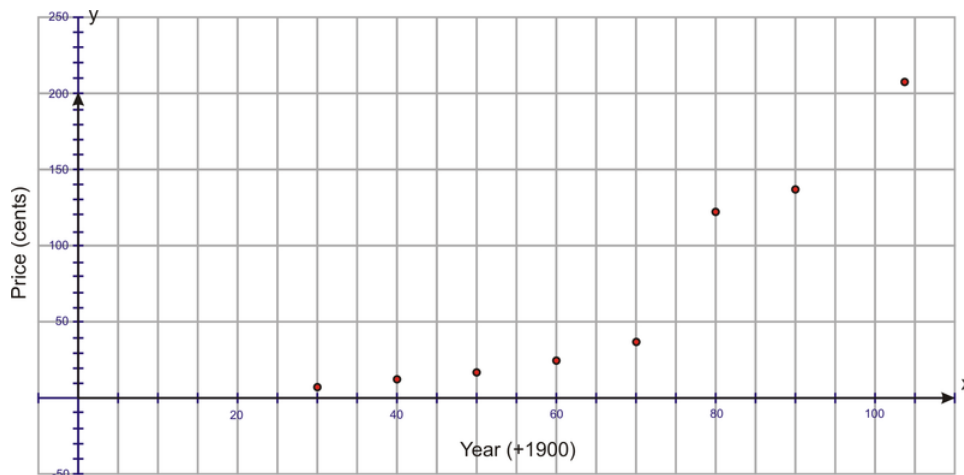


Actually, the normal distribution curve was first called a Gaussian curve after a very famous mathematician, Carl Friedrich Gauss. He lived between 1777 and 1855 in Germany. Gauss studied many aspects of mathematics. One of these was probability distributions, and in particular, the bell curve. It is interesting to note that Gauss also spoke about global warming and postulated the eventual finding of Ceres, the planet residing between Mars and Jupiter. A neat fact about Gauss is that he was also known to have beautiful handwriting. If you want to read more about Carl Friedrich Gauss, look at [http://en.wikipedia.org/wiki/Carl\\_Friedrich\\_Gauss](http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss).

In Chapter 3, you also learned about discrete random variables. Remember that discrete random variables are ones that have a finite number of values within a certain range. In other words, a discrete random variable cannot take on all values within an interval. For example, say you were counting from 1 to 10. You would count 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. These are **discrete values**. 3.5 would not count as a discrete value within the limits of 1 to 10. For a normal distribution, however, you are working with continuous variables. **Continuous variables**, unlike discrete variables, can take on any value within the limits of the variable. So, for example, if you were counting from 1 to 10, 3.5 would count as a value for the continuous variable. Lengths, temperatures, ages, and heights are all examples of continuous variables. Compare these to discrete variables, such as the number of people attending your class, the number of correct answers on a test, or the number of tails on a coin flip. You can see how a continuous variable would take on an infinite number of values, whereas a discrete variable would take on a finite number of values. As you may know, you can actually see this when you graph discrete and continuous data. Look at the 2 graphs below. The first graph is a graph of the height of a child as he or she ages. The second graph is the cost of a gallon of gasoline as the years progress.



#### 4.1. Normal Distributions

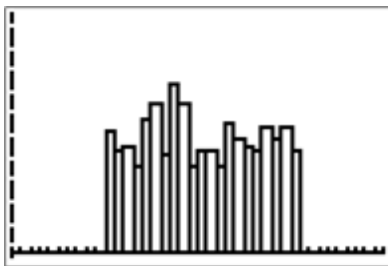


If you look at the first graph, the data points are joined, because as the child ages from birth to age 1, for example, his height also increases. As he continues to age, he continues to grow. The data is said to be continuous and, therefore, you can connect the points on the graph. For the second graph, the price of a gallon of gas at the end of each year is recorded. In 1930, a gallon of gas cost 10¢. You would not have gone in and paid 10.2¢ or 9.75¢. The data is, therefore, discrete, and the data points cannot be connected.

Let’s look at a few problems to show how histograms approximate normal distribution curves.

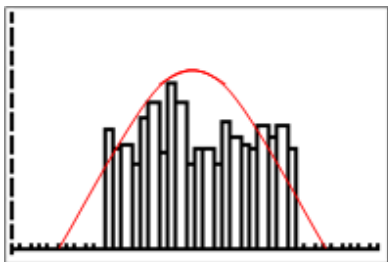
**Example 1**

Jillian takes a survey of the heights of all of the students in her high school. There are 50 students in her school. She prepares a histogram of her results. Is the data normally distributed?

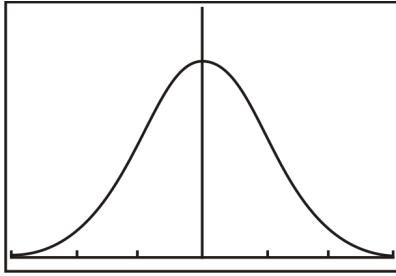


**Solution:**

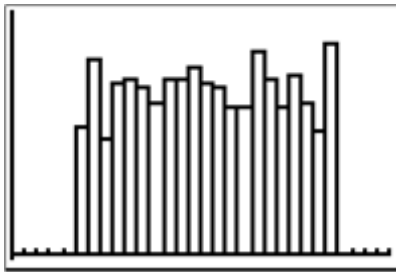
If you take a normal distribution curve and place it over Jillian’s histogram, you can see that her data does not represent a normal distribution.



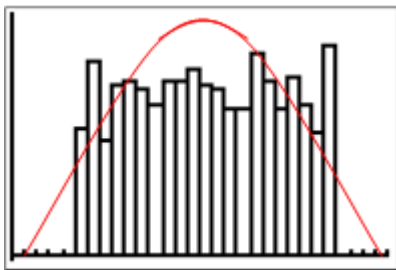
If the histogram were actually shaped like a normal distribution, it would have a shape like the curve below:

**Example 2**

Thomas did a survey similar to Jillian's in his school. His high school had 100 students. Is his data normally distributed?

**Solution:**

If you take a normal distribution curve and place it over Thomas's histogram, you can see that his data also does not represent a normal distribution.

**Example 3**

Joanne posted a problem to her friends on FaceBook. She told her friends that her grade 12 math project was to measure the lifetimes of the batteries used in different toys. She surveyed people in her neighborhood and asked them, on average, how many hours their typical battery lasts. Her results are shown below:

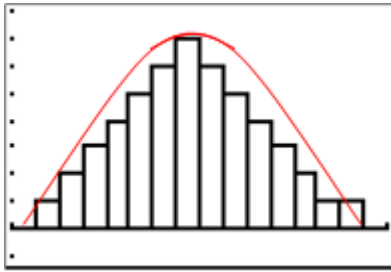
98	108	107	79	100	112	97	79	41	127
135	100	92	80	66	62	119	118	56	112
99	83	86	62	127	155	107	140	144	122
110	116	134	102	133	157	100	96	55	132
126	171	169	146	122	74	70	82	84	93

Is her data normally distributed?

**4.1. Normal Distributions**

**Solution:**

If you take a normal distribution curve and place it over Joanne's histogram, you can see that her data appears to come from a normal distribution.



This means that the data fits a normal distribution with a mean around 105. Using the TI-84 calculator, you can actually find the mean of this data to be 105.7.



What Joanne's data does tell us is that the mean score (105.7) is at the center of the distribution, and the data from all of the other scores (times) are spread from that mean. You will be learning much more about standard normal distributions in a later chapter. But for now, remember the 2 key points about a standard normal distribution. The first key point is that the data represented is continuous. The second key point is that the data is centered at the mean and is symmetrically distributed on either side of that mean.

Standard normal distributions are special kinds of distributions and differ from the binomial distributions you learned about in the last chapter. Let's now take a more detailed look at binomial distributions and see how they differ dramatically from the standard normal distribution.

## 4.2 Binomial Distributions

In the last chapter, you found that **binomial experiments** are ones that involve only 2 choices. Each observation from the experiment, therefore, falls into the category of a success or a failure. For example, if you tossed a coin to see if a 6 appears, it would be a binomial experiment. A successful event is the 6 appearing. Every other roll (1, 2, 3, 4, or 5) would be a failure. Asking your classmates if they watched *American Idol* last evening is an example of a binomial experiment. There are only 2 choices (yes and no), and you can deem a yes answer to be a success and a no answer to be a failure. Coin tossing is another example of a binomial experiment, as you saw in Chapter 3. There are only 2 possible outcomes (heads and tails), and you can say that heads are successes if you are looking to count how many heads can be obtained. Tails would then be failures. You should also note that the observations are independent of each other. In other words, whether or not Alisha watched *American Idol* does not affect whether or not Jack watched the show. In fact, knowing that Alisha watched *American Idol* does not tell you anything about any of your other classmates. Notice, then, that the probability for success for each trial is the same.

The distribution of the observations in a binomial experiment is known as a **binomial distribution**. For binomial experiments, there is also a fixed number of trials. As the number of trials increases, the binomial distribution becomes closer to a normal distribution. You should also remember that for normal distributions, the random variable is continuous, and for binomial distributions, the random variable is discrete. This is because binomial experiments have 2 outcomes (successes and failures), and the counts of both are discrete.

Of course, as the sample size increases, the accuracy of a binomial distribution also increases. Let's look at an example.

### Example 4

Keith took a poll of the students in his school to see if they agreed with the new “no cell phones” policy. He found the following results.

TABLE 4.1:

Age	Number Responding <i>No</i>
14	56
15	65
16	90
17	95
18	60

### Solution:

Keith plotted the data and found that it was distributed as follows:



Clearly, Keith's data is not normally distributed.

**Example 5**

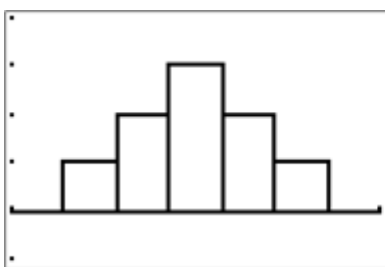
Jake sent out the same survey as Keith, except he sent it to every junior high school and high school in the district. These schools had students from grade 6 (age 12) to grade 12 (up to age 20). He found the following data:

**TABLE 4.2:**

Age	Number Responding <i>No</i>
12	274
13	261
14	259
15	289
16	233
17	225
18	253
19	245
20	216

**Solution:**

Jake plotted his data and found that the data was distributed as follows:



Because Jake expanded the survey to include more people (and a wider age range), his data turned out to fit more closely to a normal distribution.

Notice that in Jake's sample, the number of people responding no on the survey was 2,225. This number has a huge advantage over the number of students who responded no on Keith's survey, which was 366. You can make much more accurate conclusions with Jake's data. For example, Jake could say that the biggest opposition to the new "no cell phones" policy came from the middle or junior high schools, where students were 12–15 years old. This is not the same conclusion Keith would have made based on his small sample.

In Chapter 3, you did a little work with the formula used to calculate probability for binomial experiments. Here is the general formula for finding the probability of a binomial experiment from Chapter 3.

The probability of getting  $X$  successes in  $n$  trials is given by:

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$

where:

$a$  is the number of successes from the trials.

$p$  is the probability of success.

$q$  is the probability of failure.

Let's look at a simple example of a binomial probability distribution just to recap what you learned in Chapter 3.

**Example 6**

A coin is tossed 3 times. Find the probability of getting exactly 2 tails.

**Solution:**

There are 3 trials, so  $n = 3$ .

A success is getting a tail. We are interested in exactly 2 successes. Therefore,  $a = 2$ .

The probability of a success is  $\frac{1}{2}$ , and, thus,  $p = \frac{1}{2}$ .

Therefore, the probability of a failure is  $1 - \frac{1}{2}$ , or  $\frac{1}{2}$ . From this, you know that  $q = \frac{1}{2}$ .

$$\begin{aligned} P(X = a) &= {}_n C_a \times p^a \times q^{(n-a)} \\ P(2 \text{ tails}) &= {}_3 C_2 \times p^2 \times q^1 \\ P(2 \text{ tails}) &= {}_3 C_2 \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^1 \\ P(2 \text{ tails}) &= 3 \times \frac{1}{4} \times \frac{1}{2} \\ P(2 \text{ tails}) &= \frac{3}{8} \end{aligned}$$

Therefore, the probability of seeing exactly 2 tails in 3 tosses is  $\frac{3}{8}$ , or 37.5%.

**Example 7**

A local food chain has determined that 40% of the people who shop in the store use an incentive card, such as air miles. If 10 people walk into the store, what is the probability that half of them will be using an incentive card?

**Solution:**

There are 10 trials, so  $n = 10$ .

A success is a person using a card. You are interested in 5 successes. Therefore,  $a = 5$ .

The probability of a success is 40%, or 0.40, and, thus,  $p = 0.40$ .

Therefore, the probability of a failure is  $1 - 0.40$ , or 0.60. From this, you know that  $q = 0.60$ .

$$\begin{aligned} P(X = a) &= {}_n C_a \times p^a \times q^{(n-a)} \\ P(5 \text{ people using a card}) &= {}_{10} C_5 \times p^5 \times q^5 \\ P(5 \text{ people using a card}) &= {}_{10} C_5 \times (0.40)^5 \times (0.60)^5 \\ P(5 \text{ people using a card}) &= 252 \times 0.01024 \times 0.07776 \\ P(5 \text{ people using a card}) &= 0.201 \end{aligned}$$

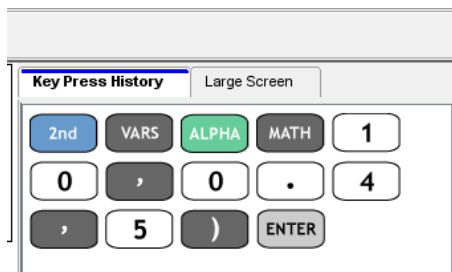
Therefore, the probability of seeing 5 people using a card in a random set of 10 people is 20.1%.

**Technology Note**

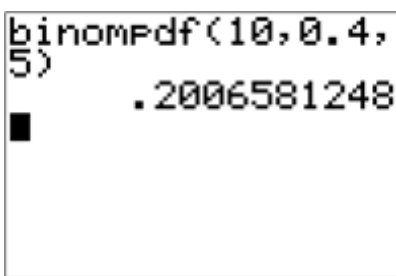
You could have used technology to solve this problem, rather than pencil and paper calculations. However, with technology, it is often very helpful to check our answers using pencil and paper as well. With Example 7, you could have used the binompdf function on the TI-84 calculator to solve this problem.

The key sequence for using the binompdf function is as follows:

**4.2. Binomial Distributions**



If you used the data from Example 7, you would find the following:



Notice that you typed in  $\text{binompdf}(n, p, a)$ . What if you changed the problem to include the phrase *at most*? In reality, probability problems are normally ones that use *at least*, *more than*, *less than*, or *at most*. You do not always have a probability problem using the word *exactly*. Take a look at Example 8, a problem similar to Example 7.

### Example 8

A local food chain has determined that 40% of the people who shop in the store use an incentive card, such as air miles. If 10 people walk into the store, what is the probability that *at most* half of these will be using an incentive card?

#### Solution:

There are 10 trials, so  $n = 10$ .

A success is using a card. We are interested in at most 5 people using a card. That is, we are interested in 0, 1, 2, 3, 4, or 5 people using a card. Therefore,  $a = 5, 4, 3, 2, 1$ , and 0.

The probability of a success is 40%, or 0.40, and, thus,  $p = 0.40$ .

Therefore, the probability of a failure is  $1 - 0.40$ , or 0.60. From this, you know that  $q = 0.60$ .

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$

$$P(5 \text{ people using a card}) = {}_{10} C_5 \times p^5 \times q^5$$

$$P(5 \text{ people using a card}) = {}_{10} C_5 \times (0.40)^5 \times (0.60)^5$$

$$P(5 \text{ people using a card}) = 252 \times 0.01024 \times 0.07776$$

$$P(5 \text{ people using a card}) = 0.201$$

$$P(4 \text{ people using a card}) = {}_{10} C_4 \times p^4 \times q^6$$

$$P(4 \text{ people using a card}) = {}_{10} C_4 \times (0.40)^4 \times (0.60)^6$$

$$P(4 \text{ people using a card}) = 210 \times 0.0256 \times 0.04666$$

$$P(4 \text{ people using a card}) = 0.251$$



$$P(3 \text{ people using a card}) = {}_{10}C_3 \times p^3 \times q^7$$

$$P(3 \text{ people using a card}) = {}_{10}C_3 \times (0.40)^3 \times (0.60)^7$$

$$P(3 \text{ people using a card}) = 120 \times 0.064 \times 0.02799$$

$$P(3 \text{ people using a card}) = 0.215$$

$$P(2 \text{ people using a card}) = {}_{10}C_2 \times p^2 \times q^8$$

$$P(2 \text{ people using a card}) = {}_{10}C_2 \times (0.40)^2 \times (0.60)^8$$

$$P(2 \text{ people using a card}) = 45 \times 0.16 \times 0.01680$$

$$P(2 \text{ people using a card}) = 0.121$$

$$P(1 \text{ person using a card}) = {}_{10}C_1 \times p^1 \times q^9$$

$$P(1 \text{ person using a card}) = {}_{10}C_1 \times (0.40)^1 \times (0.60)^9$$

$$P(1 \text{ person using a card}) = 10 \times 0.40 \times 0.01008$$

$$P(1 \text{ person using a card}) = 0.0403$$

$$P(0 \text{ people using a card}) = {}_{10}C_0 \times p^0 \times q^{10}$$

$$P(0 \text{ people using a card}) = {}_{10}C_0 \times (0.40)^0 \times (0.60)^{10}$$

$$P(0 \text{ people using a card}) = 1 \times 1 \times 0.00605$$

$$P(0 \text{ people using a card}) = 0.00605$$

The total probability for this example is calculated as follows:

$$P(X \leq 5) = 0.201 + 0.251 + 0.215 + 0.121 + 0.0403 + 0.0605$$

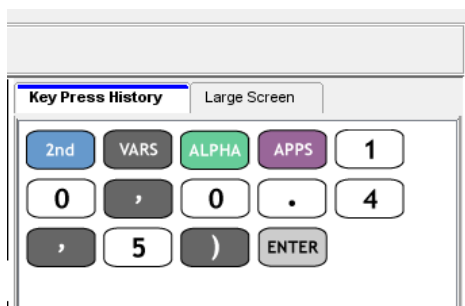
$$P(X \leq 5) = 0.834$$

Therefore, the probability of seeing at most 5 people using a card in a random set of 10 people is 82.8%.

### Technology Note

You can see now that the use of the TI-84 calculator can save a great deal of time when solving problems involving the phrases *at least*, *more than*, *less than*, or *at most*. This is due to the fact that the calculations become much more cumbersome. You could have used the binomcdf function on the TI-84 calculator to solve Example 8. Binomcdf stands for binomial cumulative probability. Binompdf simply stands for binomial probability.

The key sequence for using the binompdf function is as follows:



If you used the data from Example 8, you would find the following:

```
binomcdf(10,0.4,
5)
.8337613824
```

You can see how using the binomcdf function is a lot easier than actually calculating 6 probabilities and adding them up. If you were to round 0.8337613824 to 3 decimal places, you would get 0.834, which is our calculated value found in Example 8.

### Example 9

Karen and Danny want to have 5 children after they get married. What is the probability that they will have exactly 3 girls?

#### Solution:

There are 5 trials, so  $n = 5$ .

A success is when a girl is born, and we are interested in 3 girls. Therefore,  $a = 3$ .

The probability of a success is 50%, or 0.50, and thus,  $p = 0.50$ .

Therefore, the probability of a failure is  $1 - 0.50$ , or 0.50. From this, you know that  $q = 0.50$ .

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$

$$P(3 \text{ girls}) = {}_5 C_3 \times p^3 \times q^2$$

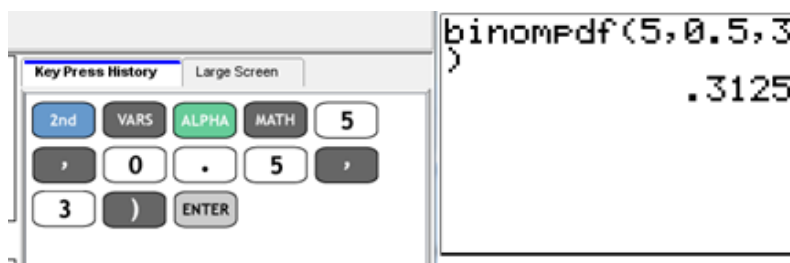
$$P(3 \text{ girls}) = {}_5 C_3 \times (0.50)^3 \times (0.50)^2$$

$$P(3 \text{ girls}) = 10 \times 0.125 \times 0.25$$

$$P(3 \text{ girls}) = 0.3125$$

Therefore, the probability of having *exactly* 3 girls from the 5 children is 31.3%.

When using technology, you will select the binompdf function, because you are looking for the probability of *exactly* 3 girls from the 5 children.



Using the TI-84 calculator gave us the same result as our calculation (and was a great deal quicker).

### Example 10

Karen and Danny want to have 5 children after they get married. What is the probability that they will have *less than* 3 girls?

**Solution:**

There are 5 trials, so  $n = 5$ .

A success is when a girl is born, and we are interested in *less than* 3 girls. Therefore,  $a = 2, 1$ , and 0.

The probability of a success is 50%, or 0.50, and, thus,  $p = 0.50$ .

Therefore, the probability of a failure is  $1 - 0.50$ , or 0.50. From this, you know that  $q = 0.50$ .

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$

$$P(2 \text{ girls}) = {}_5 C_2 \times p^2 \times q^3$$

$$P(2 \text{ girls}) = {}_5 C_2 \times (0.50)^2 \times (0.50)^3$$

$$P(2 \text{ girls}) = 10 \times 0.25 \times 0.125$$

$$P(2 \text{ girls}) = 0.3125$$

$$P(1 \text{ girl}) = {}_5 C_1 \times p^1 \times q^4$$

$$P(1 \text{ girl}) = {}_5 C_1 \times (0.50)^1 \times (0.50)^4$$

$$P(1 \text{ girl}) = 5 \times 0.50 \times 0.0625$$

$$P(1 \text{ girl}) = 0.1563$$

$$P(0 \text{ girls}) = {}_5 C_0 \times p^0 \times q^5$$

$$P(0 \text{ girls}) = {}_5 C_0 \times (0.50)^0 \times (0.50)^5$$

$$P(0 \text{ girls}) = 1 \times 1 \times 0.03125$$

$$P(0 \text{ girls}) = 0.03125$$

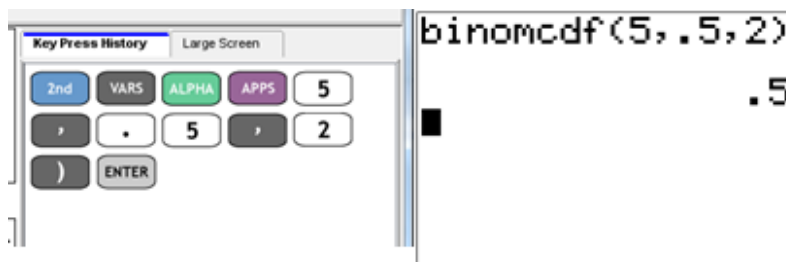
The total probability for this example is calculated as follows:

$$P(X < 3) = 0.3125 + 0.1563 + 0.03125$$

$$P(X < 3) = 0.500$$

Therefore, the probability of having *less than* 3 girls in 5 children is 50.0%.

When using technology, you will select the binomcdf function, because you are looking for the probability of *less than* 3 girls from the 5 children.



**Example 11**

A fair coin is tossed 50 times. What is the probability that you will get heads in 30 of these tosses?

**Solution:**

There are 50 trials, so  $n = 50$ .

A success is getting a head, and we are interested in *exactly* 30 heads. Therefore,  $a = 30$ .

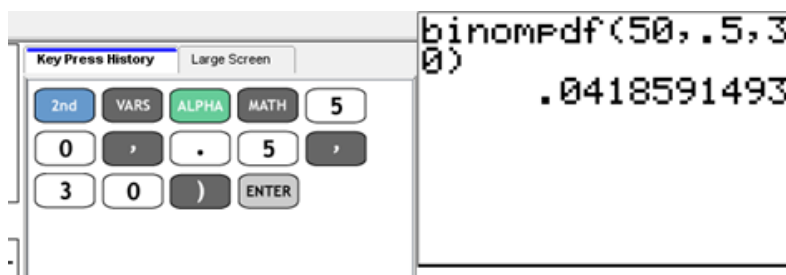
The probability of a success is 50%, or 0.50, and, thus,  $p = 0.50$ .

Therefore, the probability of a failure is  $1 - 0.50$ , or 0.50. From this, you know that  $q = 0.50$ .

$$\begin{aligned}
 P(X = a) &= {}_n C_a \times p^a \times q^{(n-a)} \\
 P(30 \text{ heads}) &= {}_{50} C_{30} \times p^{30} \times q^{20} \\
 P(30 \text{ heads}) &= {}_{50} C_{30} \times (0.50)^{30} \times (0.50)^{20} \\
 P(30 \text{ heads}) &= (4.713 \times 10^{13}) \times (9.313 \times 10^{-10}) \times (9.537 \times 10^{-7}) \\
 P(30 \text{ heads}) &= 0.0419
 \end{aligned}$$

Therefore, the probability of getting *exactly* 30 heads from 50 tosses of a fair coin is 4.2%.

Using technology to check, you get the following:

**Example 12**

A fair coin is tossed 50 times. What is the probability that you will get heads in *at most* 30 of these tosses?

**Solution:**

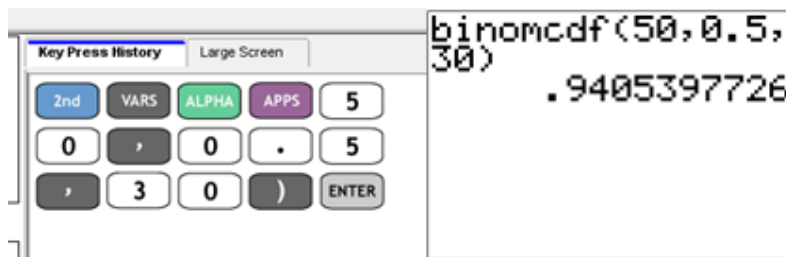
There are 50 trials, so  $n = 50$ .

A success is getting a head, and we are interested in *at most* 30 heads. Therefore,  $a = 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18,$  and, 0.

The probability of a success is 50%, or 0.50, and, thus,  $p = 0.50$ .

Therefore, the probability of a failure is  $1 - 0.50$ , or 0.50. From this, you know that  $q = 0.50$ .

Obviously, you will be using technology to solve this problem, as it would take us a long time to calculate all of the individual probabilities. The binomcdf function can be used as follows:



Therefore, the probability of having *at most* 30 heads from 50 tosses of a fair coin is 94.1%.

### Example 13

A fair coin is tossed 50 times. What is the probability that you will get heads in *at least* 30 of these tosses?

#### Solution:

There are 50 trials, so  $n = 50$ .

A success is getting a head, and we are interested in *at least* 30 heads. Therefore,  $a = 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38,$  and 30.

The probability of a success is 50%, or 0.50, and, thus,  $p = 0.50$ .

Therefore, the probability of a failure is  $1 - 0.50$ , or 0.50. From this, you know that  $q = 0.50$ .

Again, you will obviously be using technology to solve this problem, as it would take us a long time to calculate all of the individual probabilities. The binomcdf function can be used as follows:



Notice that when you use the phrase *at least*, you used the numbers 50, 0.5, 29. In other words, you would type in  $1 - \text{binomcdf}(n, p, a - 1)$ . Since  $a = 30$ , at least  $a$  would be anything greater than 29. Therefore, the probability of having *at least* 30 heads from 50 tosses of a fair coin is 10.1%.

### Example 14

You have a summer job at a jelly bean factory as a quality control clerk. Your job is to ensure that the jelly beans coming through the line are the right size and shape. If 90% of the jelly beans you see are the right size and shape, you give your thumbs up, and the shipment goes through to processing and on to the next phase to shipment. A normal day at the jelly bean factory means 15 shipments are produced. What is the probability that exactly 10 will pass inspection?

#### Solution:

There are 15 shipments, so  $n = 15$ .

A success is a shipment passing inspection, and we are interested in *exactly* 10 passing inspection.

Therefore,  $a = 10$ .

The probability of a success is 90%, or 0.90, and, thus,  $p = 0.90$ .

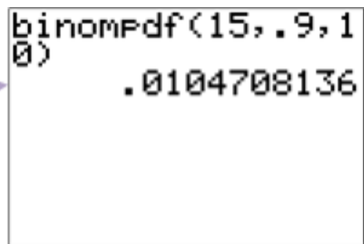
Therefore, the probability of a failure is  $1 - 0.90$ , or 0.10. From this, you know that  $q = 0.10$ .

## 4.2. Binomial Distributions

$$P(X = a) = {}_n C_a \times p^a \times q^{(n-a)}$$
$$P(10 \text{ shipments passing}) = {}_{15} C_{10} \times p^{10} \times q^5$$
$$P(10 \text{ shipments passing}) = {}_{15} C_{10} \times (0.90)^{10} \times (0.10)^5$$
$$P(10 \text{ shipments passing}) = 3003 \times 0.3487 \times (1.00 \times 10^{-5})$$
$$P(10 \text{ shipments passing}) = 0.0105$$

Therefore, the probability that *exactly* 10 of the 15 shipments will pass inspection is 1.05%.

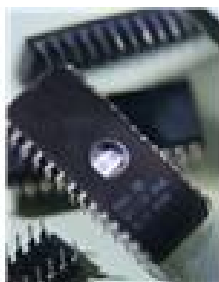
Checking our answer on  
the TI-84 calculator



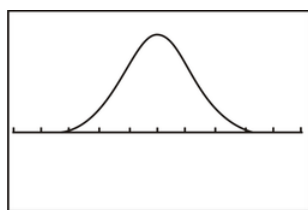
```
binompdf(15,.9,10)
.0104708136
```

## 4.3 Exponential Distributions

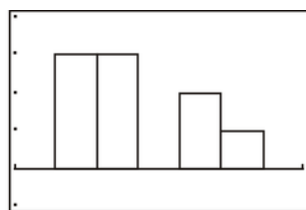
A third type of probability distribution is an **exponential distribution**. When we discussed normal distributions, or **standard distributions**, we talked about the fact that these distributions used **continuous data**, so you could use standard distributions when talking about heights, ages, lengths, temperatures, and the like. The same types of data are used when discussing exponential distributions. Exponential distributions, contrary to standard distributions, deal more with rates or changes over time. For example, the length of time the battery in your car will last is an exponential distribution. The length of time is a continuous random variable. A **continuous random variable** is one that can form an infinite number of groupings. So time, for example, can be broken down into hours, minutes, seconds, milliseconds, and so on. Another example of an exponential distribution is the lifetime of a computer part. Different computer parts have different life spans, depending on their use (and abuse). The rate of decay of the computer part determines the shape of the exponential distribution.



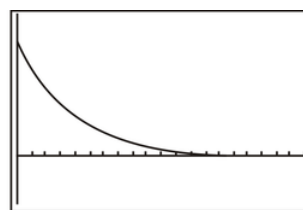
Let's look at the differences between the normal distribution curve, a binomial distribution histogram, and an exponential distribution. List some of the similarities and differences that you see in the figures below.



Standard distribution curve



Binomial distribution histogram



Exponential distribution curve

Notice that with the standard distribution and the exponential distribution curves, the data represents continuous variables. The data in the binomial distribution histogram, on the other hand, is discrete. Also, the curve for the

standard distribution is symmetrical about the mean. In other words, if you draw a horizontal line through the center of the curve, the 2 halves of the standard distribution curve would be mirror images of each other. This symmetry does not exist for the exponential distribution curve (nor for the binomial distribution). Did you notice anything else?

Let's look at some examples where the resulting graphs would show you an exponential distribution.

**Example 15**

ABC Computer Company is doing a quality control check on their newest core chip. They randomly chose 25 chips from a batch of 200 to test and examined them to see how long they would continuously run before failing. The following results were obtained:

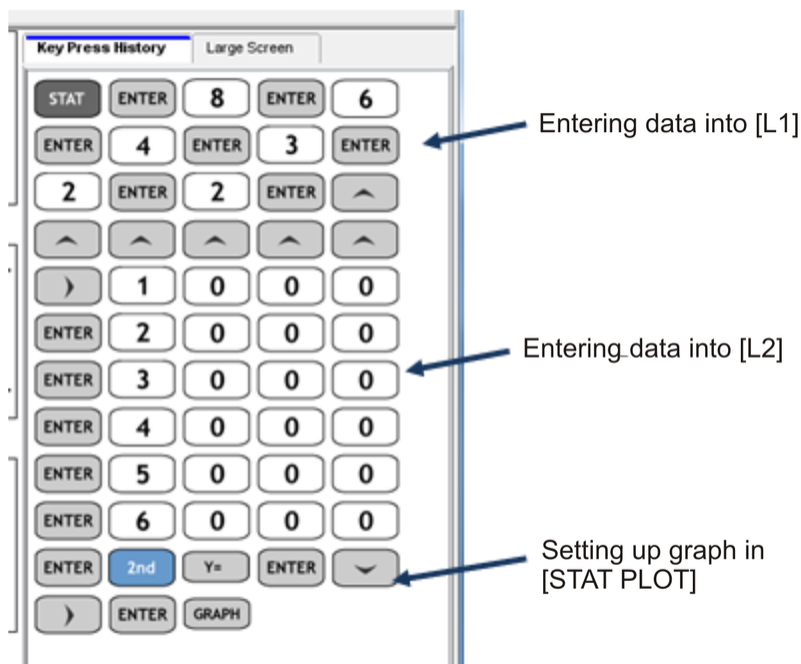
**TABLE 4.3:**

Number of Chips	Hours to Failure
8	1,000
6	2,000
4	3,000
3	4,000
2	5,000
2	6,000

What kind of data is represented in the table?

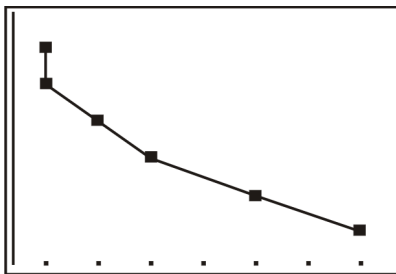
**Solution:**

In order to solve this problem, you need to graph it to see what it looks like. You can use graph paper or your calculator. Entering the data into the TI-84 involves the following keystrokes. There are a number of them, because you have to enter the data into L1 and L2, and then plot the lists using STAT PLOT.

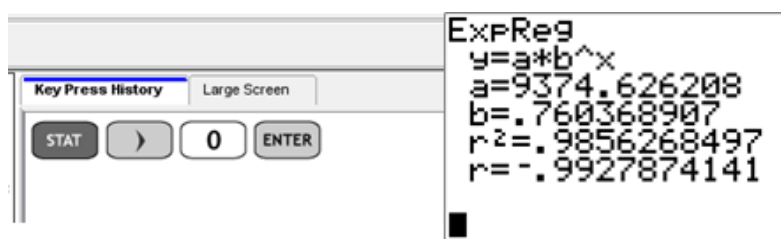


After you press **GRAPH**, you get the following curve.





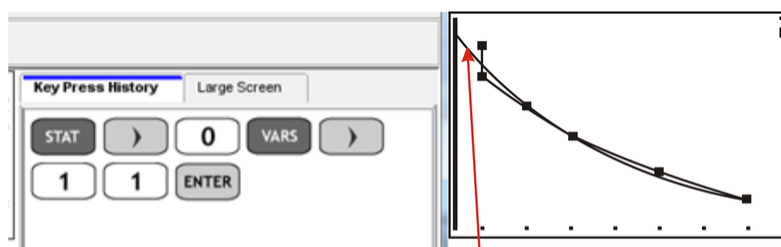
This curve looks somewhat like an exponential distribution curve, but let's test it out. You can do this on the TI-84 by pressing **[STAT]**, going to the CALC menu, pressing **[0]**, and pressing **[ENTER]**.



Notice that the  $r^2$  value is close to 1. This value indicates that an exponential curve is a good fit for this data and that the data, therefore, represents an exponential distribution.

You use regression to determine a rule that best explains the data you are observing. There is a standard quantitative measure of this best fit, known as the **coefficient of determination** ( $r^2$ ). The value of  $r^2$  can be from 0 to 1, and the closer the value is to 1, the better the fit. In our data above, the  $r^2$  value is 0.9856 for the exponential regression. If we had done a quadratic regression instead of an exponential regression, our  $r^2$  value would have been 0.9622. The data is not linear, but if we thought it might be, the  $r^2$  value would have been 0.9161. Remember, the higher the  $r^2$  value, the better the fit.

You can even go 1 step further and graph the exponential regression curve on top of our plotted points. Follow the keystrokes below and test it out.



Our curve based on the exponential regression

Note: It was not indicated that the data was in L1 and L2 when finding the exponential regression. This is because it is the default of the calculator. If you had used L2 and L3, you would have had to add this to your keystrokes.

### 4.3. Exponential Distributions



Take a look at the formula that you used with the exponential regression calculation (ExpReg) above. The general formula was  $y = ab^x$ . This is the characteristic formula for an exponential distribution curve. Siméon Poisson was one of the first to study exponential distributions with his work in applied mathematics. The Poisson distribution, as it is known, is a form of an exponential distribution. He received little credit for his discovery during his lifetime, as it only found application in the early part of the 20<sup>th</sup> century, almost 70 years after Poisson had died. To read more about Siméon Poisson, go to [http://en.wikipedia.org/wiki/Sim%C3%A9on\\_Denis\\_Poisson](http://en.wikipedia.org/wiki/Sim%C3%A9on_Denis_Poisson).

### Example 16

Radioactive substances are measured using a Geiger-Müller counter (or a Geiger counter for short). Robert was working in his lab measuring the count rate of a radioactive particle. He obtained the following data:



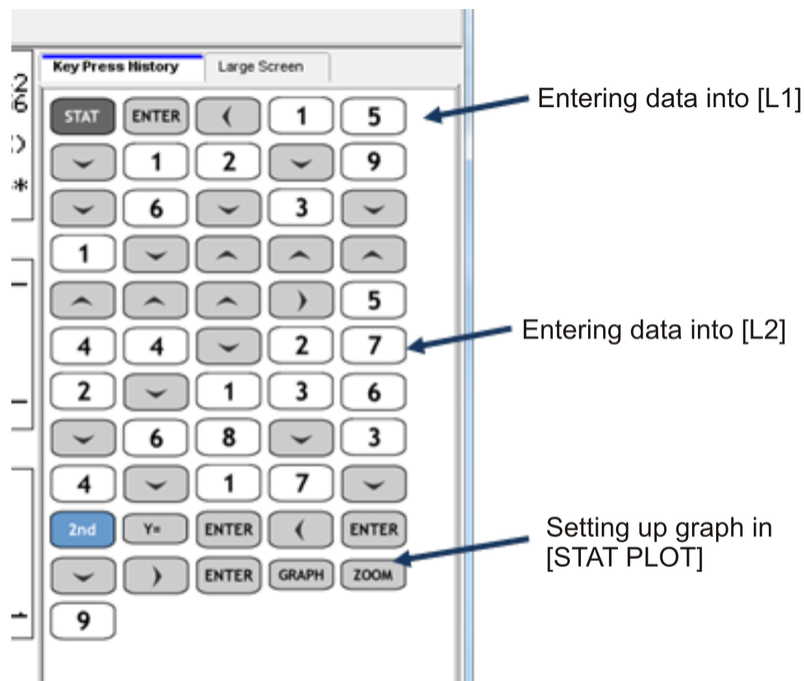
**TABLE 4.4:**

Time (hr)	Count (atoms)
15	544
12	272
9	136
6	68
3	34
1	17

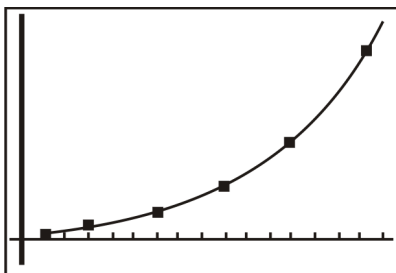
Is this data representative of an exponential distribution? If so, find the equation. What would be the count at 7.5 hours?

**Solution:**

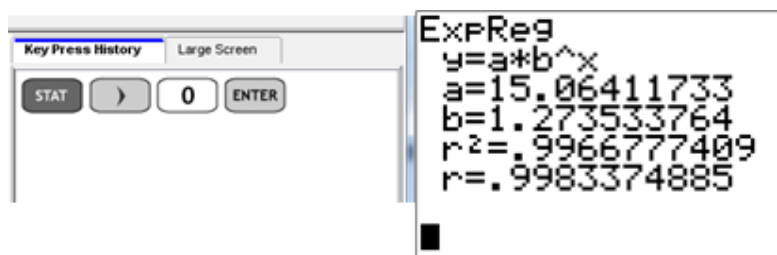
Remember, we can plot this data using pencil and paper, or we can use a graphing calculator. We will use a graphing calculator here.



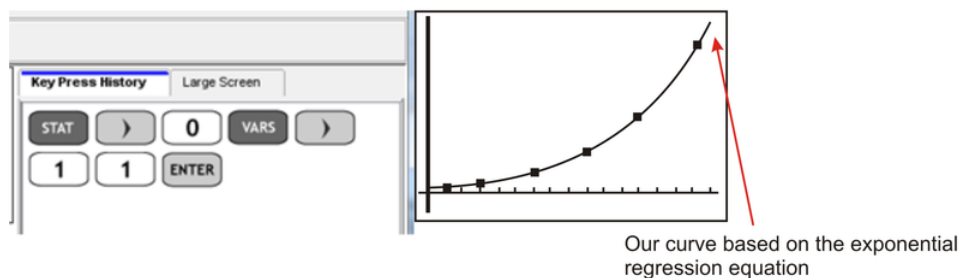
The resulting graph appears as follows:



At a glance, it does look like an exponential curve, but we really have to take a closer look by doing the exponential regression.



In the analysis of the exponential regression, we see that the  $r^2$  value is close to 1, and, therefore, the curve is indeed an exponential curve. We should go 1 step further and graph this exponential equation onto our coordinate grid and see how close a match it is.



It is a very good match, so the equation representing our data is, therefore,  $y = 15.06(1.274^x)$ .

The last part of our problem asked us to determine what the count was after 7.5 hours. In other words, what is  $y$  when  $x = 7.5$ ? This question can be answered as shown below:

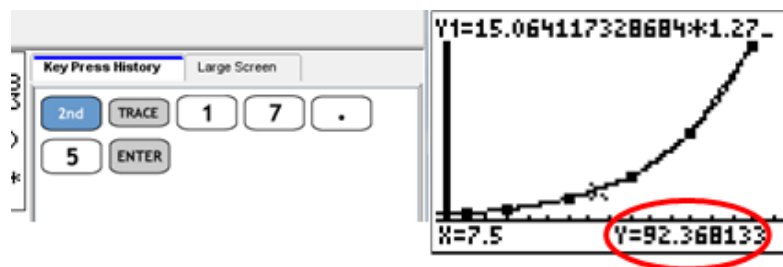
$$y = 15.06(1.274^x)$$

$$y = 15.06(1.274^{7.5})$$

$$y = 15.06(6.149)$$

$$y = 92.6 \text{ atoms}$$

We can check this on our calculator as follows:



Our calculation is a bit over, because we rounded the values for  $a$  and  $b$  in the equation  $y = ab^x$ , whereas the calculator did not.

### Example 17

Jack believes that the concentration of gold decreases exponentially as you move further and further away from the main body of ore. He collects the following data to test out his theory:



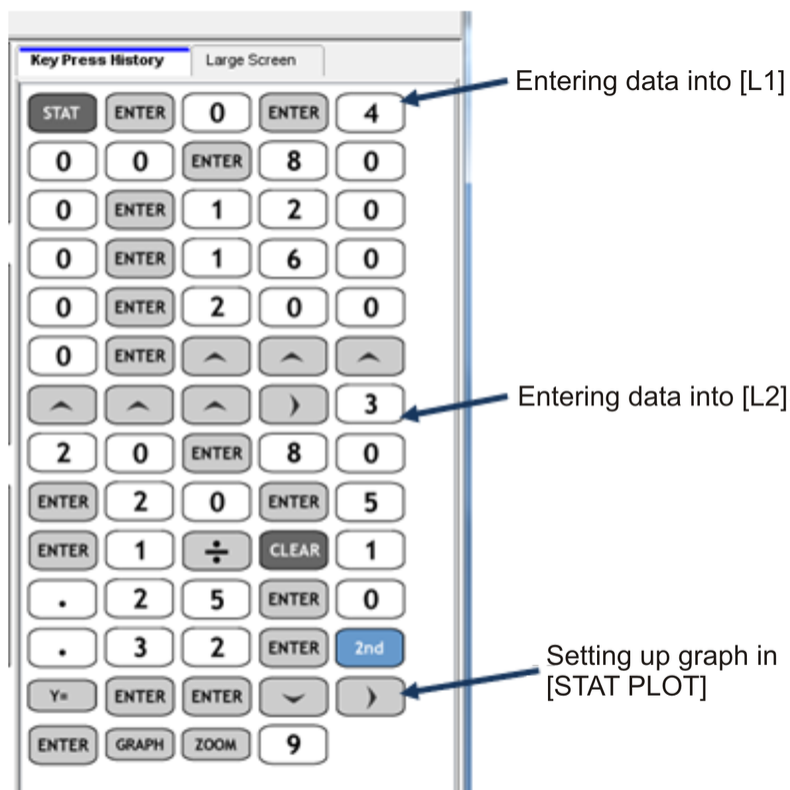
TABLE 4.5:

Distance (m)	Concentration (g/t)
0	320
400	80
800	20
1,200	5
1,600	1.25
2,000	0.32

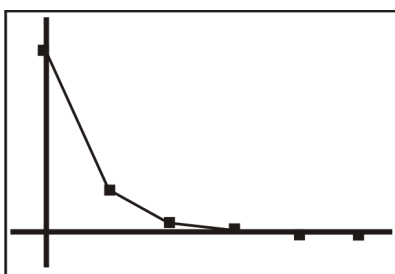
Is this data representative of an exponential distribution? If so, find the equation. What is the concentration at 1,000 m?

**Solution:**

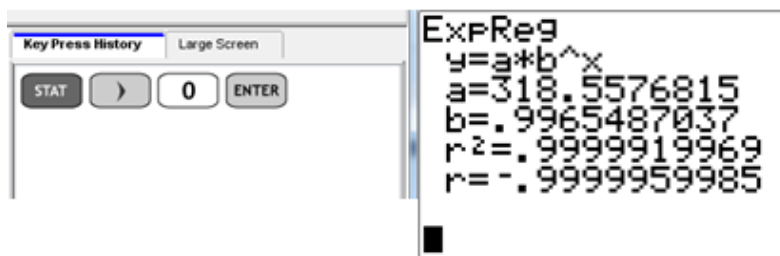
Again, we can plot this data using pencil and paper, or we can use a graphing calculator. As with Example 16, we will use a graphing calculator here.



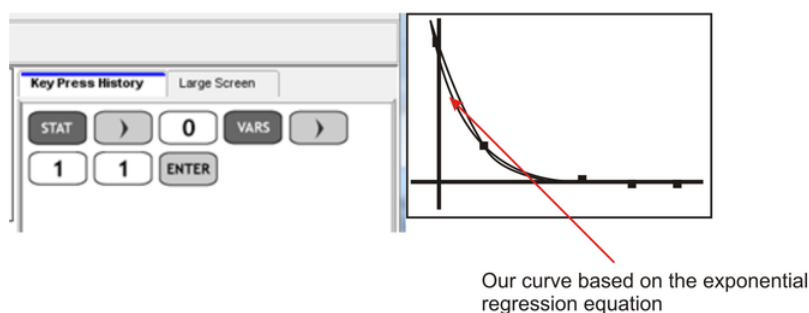
The resulting graph appears as follows:



At a glance, it does look like an exponential curve, but we really have to take a closer look by doing the exponential regression.



In the analysis of the exponential regression, we see that the  $r^2$  value is close to 1, and, therefore, the curve is indeed an exponential curve. We will go 1 step further and graph this exponential equation onto our coordinate grid and see how close a match it is.



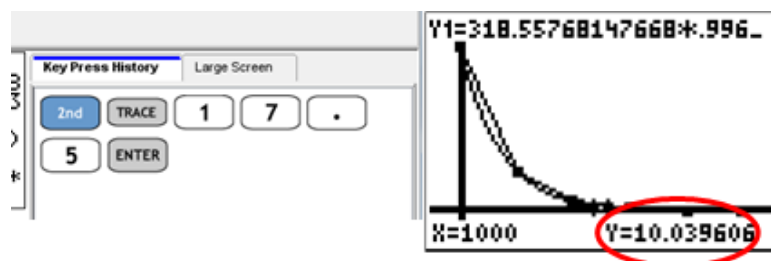
It is a very good match, so the equation representing our data is, therefore,  $y = 318.56(0.9965^x)$ .

The problem asks, "What is the concentration at 1,000 m?" This question can be answered as shown below:

$$\begin{aligned}
 y &= 318.56(0.9965^x) \\
 y &= 318.56(0.9965^{1000}) \\
 y &= 318.56(0.03001) \\
 y &= 9.56 \text{ g/t}
 \end{aligned}$$

Therefore, the concentration of gold is 9.56 grams of gold per ton of rock.

We can check this on our calculator as follows:



Our calculation is a bit under, because we rounded the values for  $a$  and  $b$  in the equation  $y = ab^x$ , whereas the calculator did not.

### Points to Consider

- Why is a normal distribution considered to be a continuous probability distribution, whereas a binomial distribution is considered to be a discrete probability distribution?
- How can you tell if a curve is truly an exponential distribution curve?

## Vocabulary

**Binomial experiments** Experiments that include only 2 choices, with distributions that involve a discrete number of trials of these 2 possible outcomes.

**Binomial distribution** A probability distribution of the successful trials of a binomial experiment.

**Continuous random variable** A variable that can form an infinite number of groupings.

**Continuous variables** Variables that take on any value within the limits of the variable.

**Continuous data** Data where an infinite number of values exist between any 2 other values. Data points are joined on a graph.

**Coefficient of determination** A standard quantitative measure of best fit. Has values from 0 to 1, and the closer the value is to 1, the better the fit.

**Discrete values** Data where a finite number of values exist between any 2 other values. Data points are not joined on a graph.

**Distribution** The description of the possible values of a random variable and the possible occurrences of these values.

**Exponential distribution** A probability distribution showing a relation in the form  $y = ab^x$ .

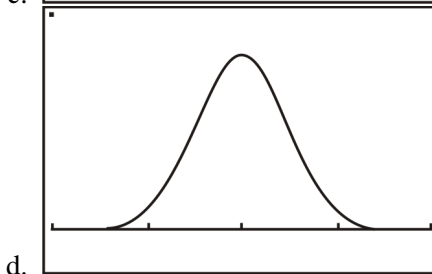
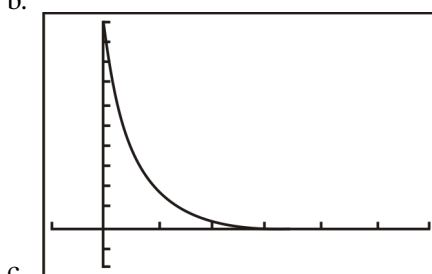
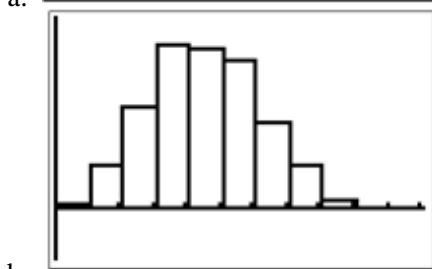
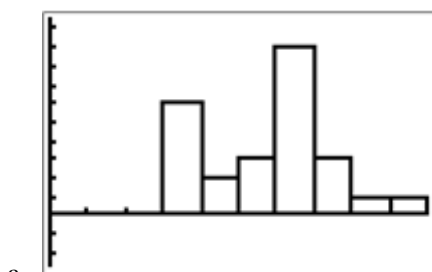
**Normal distribution curve** A symmetrical curve that shows the highest frequency in the center (i.e., at the mean of the values in the distribution) with an identical curve on either side of that center.

**Standard distributions** Normal distributions, which are often referred to as bell curves.

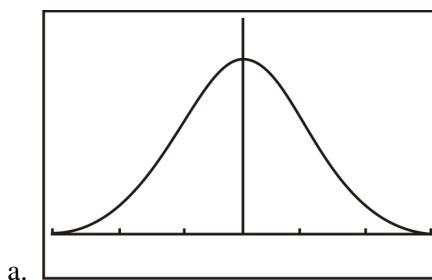
## 4.4 Review Questions

Answer the following questions and show all work (including diagrams) to create a complete answer.

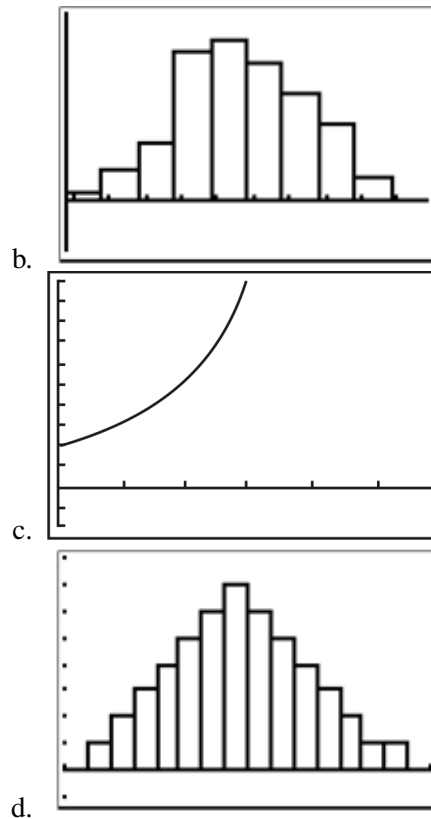
- Look at the following graphs and indicate whether they are binomial distributions, normal (standard) distributions, or exponential distributions. Explain how you know.



- Look at the following graphs and indicate whether they are binomial distributions, normal (standard) distributions, or exponential distributions. Explain how you know.







- It is determined that because of a particular genetic trend in a family, the probability of having a boy is 60%. Janet and David decide to have 4 children. What is the probability that they will have *exactly* 2 boys?
- For question 3, what is the probability that Janet and David will have *at least* 2 boys?



- For question 3, what is the probability that Janet and David will have *at most* 2 boys?
- The following data was collected on a recent 25-point math quiz. Does the data represent a normal distribution? Can you determine anything from the data?

20	17	22	23	25
14	15	14	17	9
18	2	11	18	19
14	21	19	20	18
16	13	14	10	12

- A recent blockbuster movie was rated PG, with an additional violence warning. The manager of a movie theater did a survey of moviegoers to see what ages were attending the movie in an attempt to see if people were adhering to the warnings. Is his data normally distributed? Do moviegoers at the theater regularly adhere

#### 4.4. Review Questions

to warnings?

17	9	20	27	16
15	14	24	19	14
19	7	21	18	12
5	10	15	23	14
17	13	13	12	14

8. The heights of coniferous trees were measured in a local park in a regular inspection. Is the data normally distributed? Are there areas of the park that seem to be in danger? The measurements are all in feet.



22.8	9.7	23.2	21.2	23.5
18.2	7.0	8.8	25.7	19.4
25.0	8.8	23.0	23.2	20.1
23.1	18.5	21.7	21.7	9.1
4.3	7.8	3.4	20.0	8.5

9. Thomas is studying for his AP Biology final. In order to complete his course, he must do a self-directed project. He decides to swab a tabletop in the student lounge and test for bacteria growing on the surface. Every hour, he looks in his Petri dish and makes an estimate of the number of bacteria present. The following results were recorded.



**TABLE 4.6:**

<b>Time (hr)</b>	<b>Bacteria Count</b>
0	1
1	6
2	40
3	215
4	1,300
5	7,800

Is this data representative of an exponential distribution? If so, find the equation. What is the count after 1 day?

10. If you watch a grasshopper jump, you will notice the following trend:



**TABLE 4.7:**

<b>Jump Number</b>	<b>Distance (m)</b>
1	4
2	2
3	1.1
4	0.51
5	0.25
6	0.13

---

Is this data representative of an exponential distribution? If so, find the equation. Why do you think the grasshopper's distance decreased with each jump?

# CHAPTER 5

# Measures of Central Tendency

## Chapter Outline

---

- 5.1 THE MEAN
  - 5.2 THE MEDIAN
  - 5.3 THE MODE
  - 5.4 REVIEW QUESTIONS
- 

### Introduction

Here's an activity that will involve all the students in your class and will also serve as a learning tool to enhance your understanding of the **measures of central tendency**, which are mean, median and mode. Prior to the beginning of class, fill a pail with single, plastic interlocking blocks similar to those shown below. You and your classmates will each use only 1 hand to gather a handful of blocks from the pail.



Before you and your classmates begin to pick your handfuls of blocks, have a brain-storming discussion to reveal your knowledge of the measures of central tendency. Record the various responses and refer to these as the lessons progress.

You and your classmates can now each proceed to the pail to collect a handful of blocks. Once you have had some time to compare your handful with those of your classmates, record each of your numbers of blocks on post-it notes. The post-it notes for you and your classmates can now be placed in order on a large sheet of grid paper. The grid paper allows for repeated numbers to be posted in the same column.

What do you think you would be finding if you were to determine the mean number of blocks that had been picked from the pail? Now share your blocks with your classmates, and have your classmates do the same. so that you each have a similar number of blocks. From this sharing process, it is very likely that 2 groups of students will be created. One group will have stacks of one number of blocks, and another group will have stacks of another number of blocks. You and your classmates may come to realize that further sharing will not create stacks of the same size for each of you. Is it clearer to you now what we are talking about when we use the term mean?

Place your stacks of blocks in a safe place, for they will be used again in the discovery of the mode and the median. The numbers that were placed on the grid paper can also be used for mathematical calculations of the mean, median, and mode of your data.

## 5.1 The Mean

### Learning Objectives

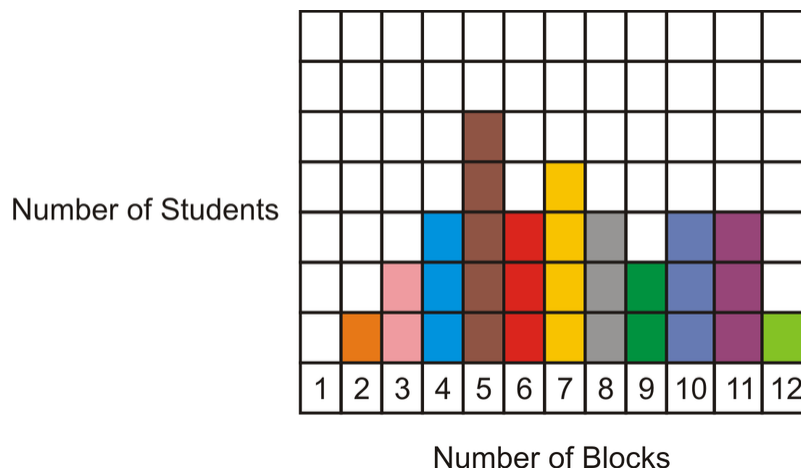
- Understand the mean of a set of numerical data.
- Compute the mean of a given set of data.
- Understand the affect of an outlier on the mean of a set of data.
- Understand the mean of a set of data as it applies to real world situations.

Now that you have had some fun discovering what you are finding when you are looking for the mean of a set of data, it is time to actually calculate the mean of your handfuls of blocks.

The term *central tendency* refers to the middle, or typical, value of a set of data, which is most commonly measured by using the 3 m's—mean, median and mode. In this lesson, we will explore the mean, and then we will move on to the median and the mode in the following lessons.

The **mean**, often called the *average* of a numerical set of data, is simply the sum of the data values divided by the number of values. This is also referred to as the arithmetic mean. The mean is the balance point of a distribution.

To calculate the actual mean of your handfuls of blocks, you can use the numbers that were posted on your grid paper. These posted numbers represent the number of blocks that were picked by each student in your class. Therefore, you are calculating the mean of a population. A population is a collection of all elements whose characteristics are being studied. You are not calculating the mean number of some of the blocks, but you are calculating the mean number of all of the blocks. We will use the example below for our calculations:



From the grid paper, you can see that there were 30 students who posted their numbers of blocks. The total number of blocks picked by all the students can be calculated as follows:

$$1 \times 2 + 2 \times 3 + 3 \times 4 + 5 \times 5 + 3 \times 6 + 4 \times 7 + 3 \times 8 + 2 \times 9 + 3 \times 10 + 3 \times 11 + 1 \times 12$$

$$2 + 6 + 12 + 25 + 18 + 28 + 24 + 18 + 30 + 33 + 12 = 208$$

The sum of all the blocks is 208, and the mean is the number you get when you divide the sum by the number of students who placed a post-it-note on the grid paper. The mean number of blocks is, therefore,  $\frac{208}{30} \approx 6.93$ . This means that, on average, each student picked 7 blocks from the pail.

### 5.1. The Mean

When calculations are done in mathematics, formulas are often used to represent the steps that are being applied. The symbol  $\Sigma$  means “the sum of” and is used to represent the addition of numbers. The numbers in every question are different, so the variable  $x$  is used to represent the numbers. To make sure that all the numbers are included, a subscript is often used to name the numbers. Therefore, the first number in the example can be represented as  $x_1$ . The number of data values for a population is written as  $N$ . The mean of the population is denoted by the symbol  $\mu$ , which is pronounced “mu.” The following formula represents the steps that are involved in calculating the mean of a set of data:

$$\text{Mean} = \frac{\text{sum of the values}}{\text{the number of values}}$$

This formula can also be written using symbols:

$$\mu = \frac{\Sigma x_1 + x_2 + x_3 + \dots + x_n}{N}$$

You can now use the formula to calculate the mean number of blocks per student:

$$\begin{aligned}\mu &= \frac{\Sigma x_1 + x_2 + x_3 + \dots + x_n}{N} \\ \mu &= \frac{2 + 6 + 12 + 25 + 18 + 28 + 24 + 18 + 30 + 33 + 12}{30} \\ \mu &= \frac{208}{30} \\ \mu &\approx 6.93\end{aligned}$$

This means that, on average, each student picked 7 blocks from the pail.

### **Example 1**

Stephen has been working at Wendy’s for 15 months. The following numbers are the number of hours that Stephen worked at Wendy’s for each of the past 7 months:

$$24, 24, 31, 50, 53, 66, 78$$

What is the mean number of hours that Stephen worked each month?

**Solution:**

**Step 1:** Add the numbers to determine the total number of hours he worked.

$$24 + 25 + 33 + 50 + 53 + 66 + 78 = 329$$

**Step 2:** Divide the total by the number of months.

$$\frac{329}{7} = 47$$

The mean number of hours that Stephen worked each month was 47.

Stephen has worked at Wendy’s for 15 months, but the numbers given above are for 7 months. Therefore, this set of data represents a sample, which is a portion of the population. The formula that was used to calculate the mean of

the blocks must be changed slightly to represent a sample. The mean of a sample is denoted by  $\bar{x}$ , which is called “ $\bar{x}$  bar.”

The number of data values for a sample is written as  $n$ . The following formula represents the steps that are involved in calculating the mean of a sample:

$$\text{Mean} = \frac{\text{sum of the values}}{\text{the number of values}}$$

This formula can now be written using symbols:

$$\bar{x} = \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{n}$$

You can now use the formula to calculate the mean number of hours that Stephen worked each month:

$$\begin{aligned}\bar{x} &= \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{n} \\ \bar{x} &= \frac{24 + 25 + 33 + 50 + 53 + 66 + 78}{7} \\ \bar{x} &= \frac{329}{7} \\ \bar{x} &= 47\end{aligned}$$

The mean number of hours that Stephen worked each month was 47.

The formulas only differ in the symbol used for the mean and the case of the variable used for the number of data values ( $N$  or  $n$ ). The calculations are done the same way for both a population and a sample. However, the mean of a population is constant, while the mean of a sample changes from sample to sample.

### Example 2

Mark operates a shuttle service that employs 8 people. Find the mean age of these workers if the ages of the 8 employees are as follows:

55 63 34 59 29 46 51 41

### Solution:

Since the data set includes the ages of all 8 employees, it represents a population. The mean age of the employees can be calculated as shown below:

$$\begin{aligned}\mu &= \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{N} \\ \mu &= \frac{55 + 63 + 34 + 59 + 29 + 46 + 51 + 41}{8} \\ \mu &= \frac{378}{8} \\ \mu &= 47.25\end{aligned}$$

The mean age of all 8 employees is 47.25 years, or 47 years and 3 months.

If you were to take a sample of 3 employees from the group of 8 and calculate the mean age for those 3 workers, would the result change? Let's use the ages 55, 29, and 46 for one sample of 3, and the ages 34, 41, and 59 for another sample of 3:

$$\begin{aligned}\bar{x} &= \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{n} \\ \bar{x} &= \frac{55 + 29 + 46}{3} \\ \bar{x} &= \frac{130}{3} \\ \bar{x} &= 43.33\end{aligned}$$

$$\begin{aligned}\bar{x} &= \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{n} \\ \bar{x} &= \frac{34 + 41 + 59}{3} \\ \bar{x} &= \frac{134}{3} \\ \bar{x} &= 44.66\end{aligned}$$

The mean age of the first 3 employees is 43.33 years.

The mean age of the second group of 3 employees is 44.66 years.

The mean age for a sample of a population depends upon what values of the population are included in the sample. From this example, you can see that the mean of a population and that of a sample from the population are not necessarily the same.

### Example 3

The selling prices of the last 10 houses sold in a small town are listed below:

\$125,000	\$142,000	\$129,500	\$89,500	\$105,000
\$144,000	\$168,300	\$96,000	\$182,300	\$212,000

Calculate the mean selling price of the last 10 homes that were sold.

### Solution:

The prices are those of a sample, so the mean of the prices can be calculated as follows:

$$\begin{aligned}\bar{x} &= \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{n} \\ \bar{x} &= \frac{125,000 + 142,000 + 129,500 + 89,500 + 105,000 + 144,000 + 168,300 + 96,000 + 182,300 + 212,000}{10} \\ \bar{x} &= \frac{\$1,393,600}{10} \\ \bar{x} &= \$139,360\end{aligned}$$

The mean selling price of the last 10 homes that were sold was \$139,360.

The mean value is one of the 3 m's and is a measure of central tendency. It is a summary statistic that gives you a description of the entire data set and is especially useful with large data sets, where you might not have the time to examine every single value. You can also use the mean to calculate further descriptive statistics, such as the variance and standard deviation. These topics will be explored in a future lesson. The mean assists you in understanding and making sense of your data, since it uses all of the values in the data set in its calculation.

When a data set is large, a **frequency distribution table** is often used to display the data in an organized way. A frequency distribution table lists the data values, as well as the number of times each value appears in the data set. A frequency distribution table is easy to both read and interpret.



The numbers in a frequency distribution table do not have to be put in order. To make it easier to enter the values in the table, a tally column is often inserted. Inserting a tally column allows you to account for every value in the data set, without having to continually scan the numbers to find them in the list. A slash (/) is used to represent the presence of a value in the list, and the total number of slashes will be the frequency. If a tally column is inserted, the table will consist of 3 columns, and if no tally column is inserted, the table will consist of 2 columns. Let's examine this concept with an actual problem and data.

**Example 4**

60 students were asked how many books they had read over the past 12 months. The results are listed in the frequency distribution table below. Calculate the mean number of books read by each student.



**TABLE 5.1:**

<b>Number of Books</b>	<b>Number of Students (Frequency)</b>
0	1
1	6
2	8
3	10
4	13
5	8
6	5
7	6
8	3

---



**Solution:**

To determine the total number of books that were read by the students, each number of books must be multiplied by the number of students who read that particular number of books. Then all the products must be added to determine the total number of books read. This total number divided by 60 will tell you the mean number of books read by each student. The formula that was written to determine the mean,  $\bar{x} = \frac{\sum x_1 + x_2 + x_3 + \dots + x_n}{n}$ , does not show any multiplication of the numbers by their frequencies. However, this can be easily inserted into this formula as shown below:

$$\bar{x} = \frac{\sum x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_n f_n}{f_1 + f_2 + f_3 + \dots + f_n}$$

This formula will now be used to calculate the mean number of books read by each student.

$$\begin{aligned}\bar{x} &= \frac{\sum x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_n f_n}{f_1 + f_2 + f_3 + \dots + f_n} \\ \bar{x} &= \frac{\sum (0)(1) + (1)(6) + (2)(8) + (3)(10) + (4)(13) + (5)(8) + (6)(5) + (7)(6) + (8)(3)}{1 + 6 + 8 + 10 + 13 + 8 + 5 + 6 + 3} \\ \bar{x} &= \frac{\sum 0 + 6 + 16 + 30 + 52 + 40 + 30 + 42 + 24}{60} \\ \bar{x} &= \frac{240}{60} \\ \bar{x} &= 4\end{aligned}$$

The mean number of books read by each student was 4 books.

Suppose the numbers of books read by each student were randomly listed, and it was your job to determine the mean of the numbers.

0 5 1 4 4 6 7 2 4 3 7 2 6 4 2  
8 5 8 3 4 3 6 4 5 6 1 1 3 5 4  
1 5 4 1 7 3 5 4 3 8 7 2 4 7 2  
1 4 6 3 2 3 5 3 2 4 7 2 5 4 3

An alternative to entering all the numbers into a calculator would be to create a frequency distribution table like the one shown below:

**TABLE 5.2:**

Number of Books	Tally	Number of Students (Frequency)
0		1
1		6
2		8
3		10
4		13
5		8
6		5
7		6
8		3



Now that the data has been organized, the numbers of books read and the numbers of students who read the books are evident. The mean can be calculated as it was above.

**Example 5**

The following data shows the heights in centimeters of a group of grade 10 students:

183 171 158 171 182 158 164 183  
 179 170 182 183 170 171 167 176  
 176 164 176 179 183 176 170 183  
 183 167 167 176 171 182 179 170

Organize the data in a frequency distribution table and calculate the mean height of the students.

**Solution:**

**TABLE 5.3:**

Height of Students(cm)	Tally	Number of Students (Frequency)
171		4
158		2
176		5
182		3
164		2
179		3
170		4
183		6
167		3

$$\bar{x} = \frac{\sum x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_n f_n}{f_1 + f_2 + f_3 + \dots + f_n}$$

$$\bar{x} = \frac{\sum (171)(4) + (158)(2) + (176)(5) + (182)(3) + (164)(2) + (179)(3) + (170)(4) + (183)(6) + (167)(3)}{4 + 2 + 5 + 3 + 2 + 3 + 4 + 6 + 3}$$

$$\bar{x} = \frac{\sum 684 + 316 + 880 + 546 + 328 + 537 + 680 + 1098 + 501}{32}$$

$$\bar{x} = \frac{5570}{32} \approx 174.1 \text{ cm}$$

The mean height of the students is approximately 174.1 cm.

The mean is often used as a summary statistic. However, it is affected by extreme values, or **outliers**. This means that when there are extreme values at one end of a data set, the mean is not a very good summary statistic. For example, if you were employed by a company that paid all of its employees a salary between \$60,000 and \$70,000, you could probably estimate the mean salary to be about \$65,000. However, if you had to add in the \$150,000 salary of the CEO when calculating the mean, then the value of the mean would increase greatly. It would, in fact, be the mean of the employees' salaries, but it probably would not be a good measure of the central tendency of the salaries.

Technology is a major tool that is available for you to use when doing mathematical calculations, and its use goes beyond entering numbers to perform simple arithmetic operations. For example, the TI-83 calculator can be used to determine the mean of a set of given data values. You will first learn to calculate the mean by simply entering the data values into a list and determining the mean. The second method that you will learn about utilizes the frequency table feature of the TI-83.


### Example 6

Using technology, determine the mean of the following set of numbers:

24, 25, 25, 25, 26, 26, 27, 27, 28, 28, 31, 32



**Solution:**

**Step 1:**

Stat → Enter →  → Enter → Put the data in L<sub>1</sub>

L1	L2	L3	1
27			
27			
28			
28			
31			
32			
L1(13) =			

**Step 2:**

Stat → CALC → EDIT  TESTS → Enter → 1-Var Stats L1   
 1: 1-Var Stats  
 2: 2-Var Stats  
 3: Med-Med  
 4: LinReg(ax+b)  
 5: QuadReg  
 6: CubicReg  
 7: QuartReg

To enter  $L_1$ , press  $2^{nd}$  1

Enter → 1-Var Stats  
 $\bar{x}=27$   
 $\Sigma x=324$   
 $\Sigma x^2=8814$   
 $Sx=2.449489743$   
 $\sigma x=2.34520788$   
 $\downarrow n=12$


Notice that the sum of the data values is 324 ( $\Sigma x = 324$ ).

Notice that the number of data values is 12 ( $n = 12$ ).

Notice the mean of the data values is 27 ( $\bar{x} = 27$ ).

Now we will use the same data values and use the TI-83 to create a frequency table.


### Step 1:

Stat → Enter →  CALC TESTS → Enter → Put the data in  $L_1$ , but enter each number  
 1: Edit...  
 2: SortA(  
 3: SortD(  
 4: ClrList  
 5: SetUpEditor

only once.

L1	L2	L3	1
25			
26			
27			
28			
31			
32			
L1(B)=			

### Step 2:

Stat → Enter →  CALC TESTS → Enter → Put the frequency in  $L_2$  →

L1	L2	L3	2
25			
26			
27			
28			
31			
32			
L2(B) =			

### Step 3:

#### 5.1. The Mean

Stat → Enter →

L1	L2	3
25	3	---
26	2	
27	2	
28	2	
31	1	
32	1	
---	---	

L3 = L1 \* L2

→ Enter →

L1	L2	L3	3
25	3	75	
26	2	52	
27	2	54	
28	2	56	
31	1	31	
32	1	32	
---	---	---	

L3(B) =

**Step 4:**

Press **2ND** **0** to obtain the CATALOG menu of the calculator. Scroll down to the sum function and enter  $L_3$  →

sum(L3) 324

You can repeat this step to determine the sum of  $L_2$  →

sum(L2) 12

Now the mean of the data can be calculated as follows:

$$\bar{x} = \frac{324}{12} = 27$$

Note that not all the data values and frequencies are visible in the screenshots, but rest assured that they were all entered into the calculator.

After entering the data into  $L_1$  and the frequencies into  $L_2$ , another way to solve this problem with the calculator would have been to press **2ND** **STAT**, go to the MATH menu, choose option 3, and enter  $L_1$  and  $L_2$  so that you have mean( $L_1$ ,  $L_2$ ). Then press **ENTER** to get the answer. This way, the calculator will do all the calculations for you.

In addition to calculating the mean for a given set of data values, you can also apply your understanding of the mean to determine other information that may be asked for in everyday problems.

**Example 7**

During his final season with the Cadillac Selects, Joe Sure Shot played 14 regular season basketball games and had an average of 24.5 points per game. In the first 2 playoff games, Joe scored 18 and 26 points, respectively. Determine his new average for the season.

**Solution:**

**Step 1:** Multiply the given average by 14 to determine the total number of points he had scored before the playoff games.

$$24.5 \times 14 = 343$$

**Step 2:** Add the points from the 2 playoff games to this total.

$$343 + 18 + 26 = 387$$

**Step 3:** Divide this new total by 16 to determine the new average.

$$\bar{x} = \frac{387}{16} \approx 24.19$$

All of the values for the means that you have calculated so far have been for ungrouped, or listed, data. A mean can also be determined for data that is grouped, or placed in intervals. Unlike listed data, the individual values for grouped data are not available, and you are not able to calculate their sum. To calculate the mean of grouped data, the first step is to determine the midpoint of each interval, or class. These midpoints must then be multiplied by the frequencies of the corresponding classes. The sum of the products divided by the total number of values will be the value of the mean. The following example will show how the mean value for grouped data can be calculated.

**Example 8**

In Tim's school, there are 25 teachers. Each teacher travels to school every morning in his or her own car. The distribution of the driving times (in minutes) from home to school for the teachers is shown in the table below:

**TABLE 5.4:**

Driving Times (minutes)	Number of Teachers
0 to less than 10	3
10 to less than 20	10
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

The driving times are given for all 25 teachers, so the data is for a population. Calculate the mean of the driving times.

**Solution:**

**Step 1:** Determine the midpoint for each interval.

For 0 to less than 10, the midpoint is 5.

For 10 to less than 20, the midpoint is 15.

For 20 to less than 30, the midpoint is 25.

For 30 to less than 40, the midpoint is 35.

For 40 to less than 50, the midpoint is 45.

**Step 2:** Multiply each midpoint by the frequency for the class.

For 0 to less than 10,  $(5)(3) = 15$

For 10 to less than 20,  $(15)(10) = 150$

For 20 to less than 30,  $(25)(6) = 150$

For 30 to less than 40,  $(35)(4) = 140$

For 40 to less than 50,  $(45)(2) = 90$

**Step 3:** Add the results from Step 2 and divide the sum by 25.

$$15 + 150 + 150 + 140 + 90 = 545$$

$$\mu = \frac{545}{25} = 21.8$$

Each teacher spends a mean time of 21.8 minutes driving from home to school each morning.

To better represent the problem and its solution, a table can be drawn as follows:

**TABLE 5.5:**

Driving Times (minutes)	Number of Teachers $f$	Midpoint Of Class $m$	Product $mf$
0 to less than 10	3	5	15
10 to less than 20	10	15	150
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90

For the population,  $N = 25$  and  $\sum mf = 545$ , where  $m$  is the midpoint of the class and  $f$  is the frequency. The mean for the population was found by dividing  $\sum mf$  by  $N$ . As a result, the formula  $\mu = \frac{\sum mf}{N}$  can be written to summarize the steps used to determine the value of the mean for a set of grouped data. If the set of data represented a sample instead of a population, the process would remain the same, and the formula would be written as  $\bar{x} = \frac{\sum mf}{n}$ .

**Example 9**

The following table shows the frequency distribution of the number of hours spent per week texting messages on a cell phone by 60 grade 10 students at a local high school.



**TABLE 5.6:**

Time Per Week (Hours)	Number of Students
0 to less than 5	8
5 to less than 10	11
10 to less than 15	15
15 to less than 20	12
20 to less than 25	9
25 to less than 30	5





Calculate the mean number of hours per week spent by each student texting messages on a cell phone. Hint: A table may be useful.

**Solution:**

**TABLE 5.7:**

Time Per Week (Hours)	Number of Students $f$	Midpoint of Class $m$	Product $mf$
0 to less than 5	8	2.5	20.0
5 to less than 10	11	7.5	82.5
10 to less than 15	15	12.5	187.5
15 to less than 20	12	17.5	210.0
20 to less than 25	9	22.5	202.5
25 to less than 30	5	27.5	137.5

$$\begin{aligned}\bar{x} &= \frac{\sum mf}{n} \\ \bar{x} &= \frac{20.0 + 82.5 + 187.5 + 210.0 + 202.5 + 137.5}{60} \\ \bar{x} &= \frac{840}{60} \\ \bar{x} &= 14\end{aligned}$$

The mean time spent per week by each student texting messages on a cell phone is 14 hours.

Now that you have created several distribution tables for grouped data, it's time to point out that the first column of the table can be represented in another way. As an alternative to writing the interval, or class, in words, the words can be expressed as [# - #), where the front square bracket closes the class, so the first number is included in the designated interval, but the open bracket at the end does not close the class, so the last number is not included in the designated interval. Keeping this in mind, the table in Example 9 can be presented as follows:

**TABLE 5.8:**

Time Per Week (Hours)	Number of Students $f$	Midpoint of Class $m$	Product $mf$
[0 - 5)	8	2.5	20.0
[5 - 10)	11	7.5	82.5
[10 - 15)	15	12.5	187.5
[15 - 20)	12	17.5	210.0
[20 - 25)	9	22.5	202.5
[25 - 30)	5	27.5	137.5

**Lesson Summary**

You have learned the significance of the mean as it applies to a set of numerical data. You have also learned how to calculate the mean using appropriate formulas for the given data for both a population and a sample. When the data was presented as a list of numbers, you learned how to represent the values in a frequency distribution table, and when the data was grouped, you learned how to represent the data in a distribution table with appropriate intervals, as well as how to calculate the mean of this data. The use of technology in calculating the mean was also demonstrated in this lesson.

**Points to Consider**

- Is the mean only used as a measure of central tendency, or is it applied to other representations of data?
- If the mean is applied to other representations of data, can its value be calculated or estimated from this representation?
- What other measures of central tendency can be used as a statistical summary when the mean is not the best measure to use?

## 5.2 The Median

### Learning Objectives

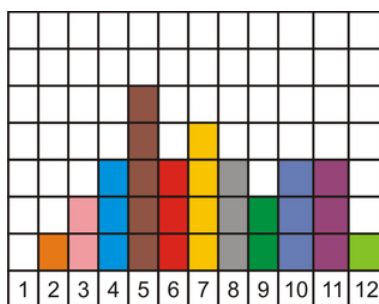
- Understand the median of a set of data as being an important measure of central tendency.
- Determine the median of a set of numerical data when there is an odd number of values and an even number of values.
- Understand the application of the median to real-world problems.

Before class begins, bring out the blocks that you and your classmates chose from the pail for the lesson on mean. In addition, have the grid paper on display where each student in your class posted his or her number of blocks.

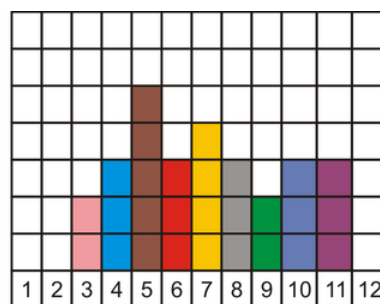
To begin the class, refer to the comments on the measures of central tendency that were recorded from the previous lesson, when the brainstorming session occurred. Highlight the comments that were made with regard to the median of a set of data and discuss this measure of central tendency with your classmates. Once the discussion has been completed, choose a handful of blocks like you did before, with your classmates doing the same.

Next, form a line with your classmates as a representation of the data from smallest to largest. In other words, those with the largest number of blocks should be at one end of the line, and those with the fewest number of blocks should be at the other end of the line. Have those at the ends of the line move out from the line 2 at a time, with 1 from each end leaving the line at the same time. This movement from the ends should enforce the concept of what is meant by *center*, and as you and your classmates move away, you will see that the last 1 or 2 people remaining in the line actually represent the center of the data.

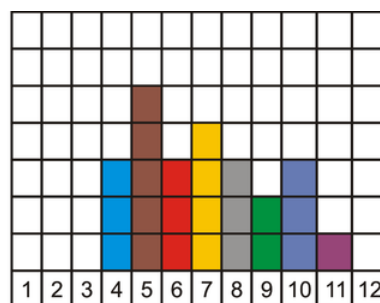
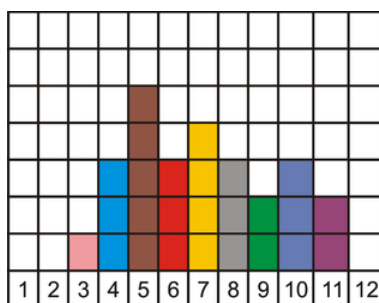
A similar activity can be done by using the grid paper chart. Instead of you and your classmates moving from a line, you could simply remove your post-it notes the same way.

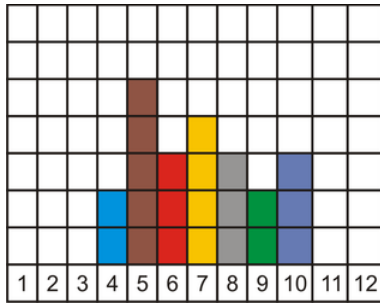


One from each end

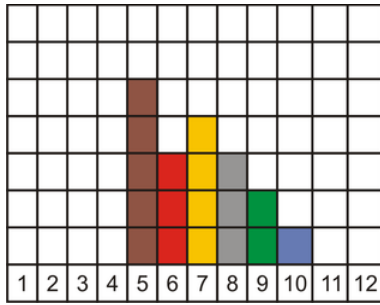
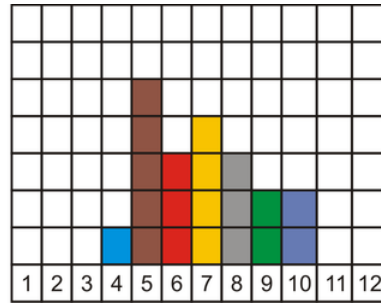


One from each end

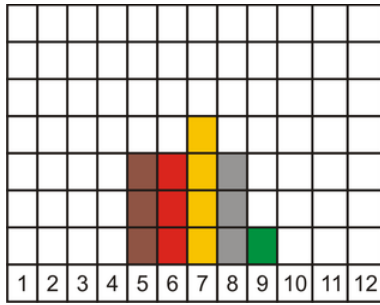
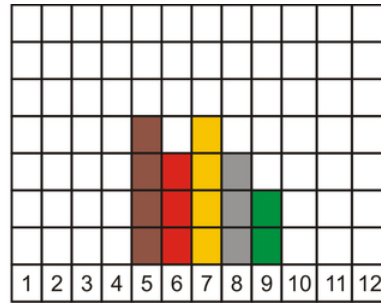




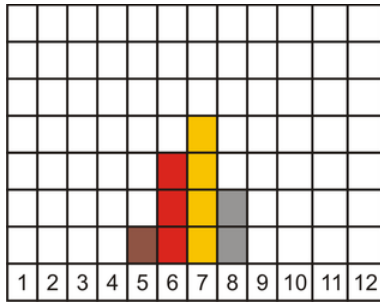
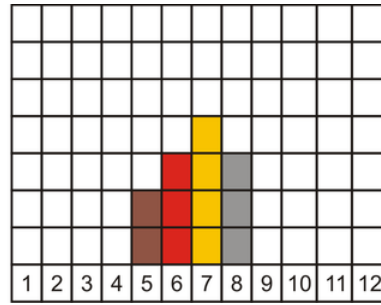
One from each end



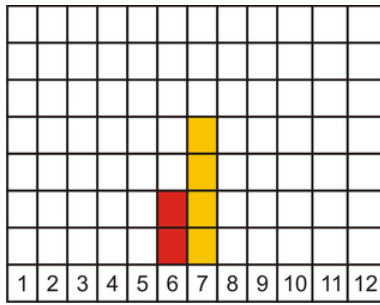
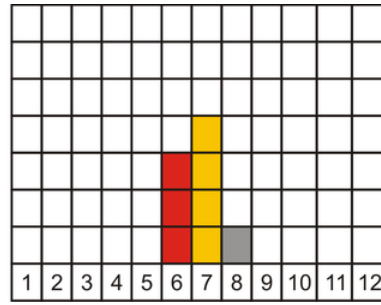
One from each end



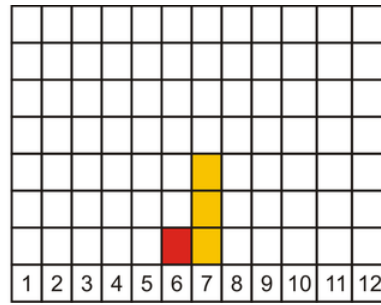
One from each end

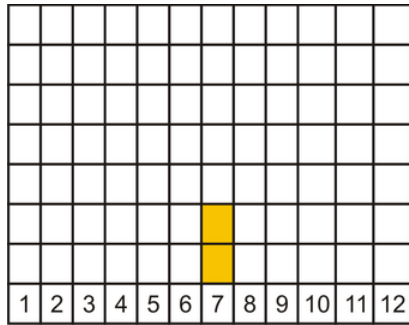


One from each end



One from each end





There are two students left – each with 7 blocks. The column is the center of the grid paper. The concept of center is visible.

Another simple activity to reinforce the concept of center is to place 5 desks in a single row and to have one of your classmates sit in the middle desk.



Your classmate should be sitting in desk number 3. The other students in your class will quickly notice that there are 2 desks in front of your classmate and 2 desks behind your classmate. Therefore, your classmate is sitting in the desk in the middle position, which is the median of the desks.

From the discussion, the activity with the blocks, and the activity with the desks, you and your classmates should have an understanding of the meaning of the median with respect to a set of data. Here is another example. The test scores for 7 students were 25, 55, 58, 64, 66, 68 and 70. The mean mark is 48.6, which is lower than all but 1 of the student's marks. The one very low mark of 25 has caused the mean to be skewed. A better measure of the average performance of the 7 students would be the middle mark of 64. The **median** is the number in the middle position once the data has been organized. Organized data is simply the numbers arranged from smallest to largest or from largest to smallest. 64 is the only number for which there are as many values above it as below it in the set of organized data, so it is the median. The median for an odd number of data values is the value that divides the data into 2 halves. If  $n$  represents the number of data values and  $n$  is an odd number, then the median will be found in the  $\frac{n+1}{2}$  position.

### Example 10

Find the median of the following data:

- 12, 2, 16, 8, 14, 10, 6
- 7, 9, 3, 4, 11, 1, 8, 6, 1, 4

### Solution:

- The first step is to organize the data, or arrange the numbers from smallest to largest.

$$12, 2, 16, 8, 14, 10, 6 \quad \rightarrow \quad 2, 6, 8, 10, 12, 14, 16$$

The number of data values is 7, which is an odd number. Therefore, the median will be found in the  $\frac{n+1}{2}$  position.

$$\frac{n+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4$$

## 5.2. The Median

In this case, the median is the value that is found in the 4<sup>th</sup> position of the organized data.

$$2, 6, 8, \boxed{10}, 12, 14, 16$$

This means that the median is 10.

b) The first step is to organize the data, or arrange the numbers from smallest to largest.

$$7, 9, 3, 4, 11, 1, 8, 6, 1, 4 \rightarrow 1, 1, 3, 4, 4, 6, 7, 8, 9, 11$$

The number of data values is 10, which is an even number. Therefore, the median will be the mean of the number found before the  $\frac{n+1}{2}$  position and the number found after the  $\frac{n+1}{2}$  position.

$$\frac{n+1}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5$$

The number found before the 5.5 position is 4, and the number found after the 5.5 position is 6.

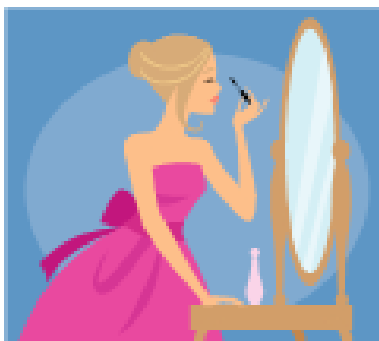
$$1, 1, 3, 4, \boxed{4, 6}, 7, 8, 9, 11$$

This means that the median is  $\frac{4+6}{2} = \frac{10}{2} = 5$ .

### **Example 11**

The amount of money spent by each of 15 high school girls for a prom dress is shown below:

\$250	\$175	\$325	\$195	\$450	\$300	\$275	\$350	\$425
\$150	\$375	\$300	\$400	\$225	\$360			



What is the median price spent on a prom dress?

**Solution:**

\$150	\$175	\$195	\$225	\$250	\$275	\$300	$\boxed{\$300}$	\$325	\$350	\$360	\$375
\$400	\$425	\$450									

The prices have been organized from least to greatest, and the number of prices is an odd number. Therefore, the median will be in the  $\frac{n+1}{2}$  position:  $\frac{n+1}{2} = \frac{15+1}{2} = \frac{16}{2} = 8$ .

The median price is \$300, which is the 8<sup>th</sup> position.

### Example 12

The students from a local high school volunteered to clean up the playground as an act of community service. The numbers of pop cans collected by 20 of the students are shown in the following table:



16	22	10	8	14
12	36	18	12	10
34	26	44	6	20
31	25	15	9	13

What is the median number of pop cans collected by a student?

**Solution:**

6	8	9	10	10
12	12	13	14	15
16	18	20	22	25
26	31	34	36	44

There is an even number of data values in the table, so the median will be the mean of the number before the  $\frac{n+1}{2}$  position and the number after the  $\frac{n+1}{2}$  position:  $\frac{n+1}{2} = \frac{20+1}{2} = \frac{21}{2} = 10.5$ .

The number before the 10.5 position is 15, and the number after the 10.5 position is 16. Therefore, the median is  $\frac{15+16}{2} = \frac{31}{2} = 15.5$ .

The median number of pop cans collected by a student is 15.5.

Often, the number of data values is quite large, and the task of organizing the data can take a great deal of time. To help organize data, the TI-83 calculator can be used. The following example will show you how to use the calculator to organize data and find the median for the data values.

### Example 13

The local police department spent the holiday weekend ticketing drivers who were speeding. 50 locations within the state were targeted as being ideal spots for drivers to exceed the posted speed limit. The number of tickets issued during the weekend in each of the locations is shown in the following table. What is the median number of speeding tickets issued?



32 12 15 8 16 42 9 18 11 10  
 24 18 6 17 21 41 3 5 35 27  
 13 26 16 28 31 3 7 37 10 19  
 23 33 7 25 36 40 15 21 38 46  
 17 37 9 2 33 41 23 29 19 40

**Solution:**

Using the TI-83 calculator



**Step 1:**

```
STAT → 2nd CALC TESTS ENTER →
```

L1	L2	L3	1
32	---	---	
12			
15			
8			
16			
42			
9			
L1()=32			

```

1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
    
```

**Step 2:**

```
STAT → 2nd CALC TESTS ENTER 2nd 1 → SortA(L1
```

```

1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
    
```

Done



The numbers that you entered into L1 are now sorted from smallest to largest.

### Step 3:

STAT → **2ND** CALC TESTS ENTER →

L1	L2	L3	1
1	---	---	
2	---	---	
3	---	---	
4	---	---	
5	---	---	

L1(1)=2

You can now scroll down the list to reveal the ordered numbers.

There are 50 data values in the table. The median will be the mean of the number before the  $\frac{n+1}{2}$  position and the number after the  $\frac{n+1}{2}$  position:  $\frac{n+1}{2} = \frac{50+1}{2} = \frac{51}{2} = 25.5$ . The number before the 25.5 position is 19, and the number after the 25.5 position is 21. This means that the median is  $\frac{19+21}{2} = \frac{40}{2} = 20$ .

2	3	3	5	6	7	7	8	9	9
10	10	11	12	13	15	15	16	16	17
17	18	18	19	19	21	21	23	23	24
25	26	27	28	29	31	32	33	33	35
36	37	37	38	40	40	41	41	42	46

The median number of speeding tickets is 20.

The median of the data can also be determined without even sorting the data. The following are 2 ways that you can use the TI-83 to determine the median of the values without sorting the data.

#### Method One:

All of the data values have been entered into L1. The median of the data values can now be determined by using the TI-83 as follows:

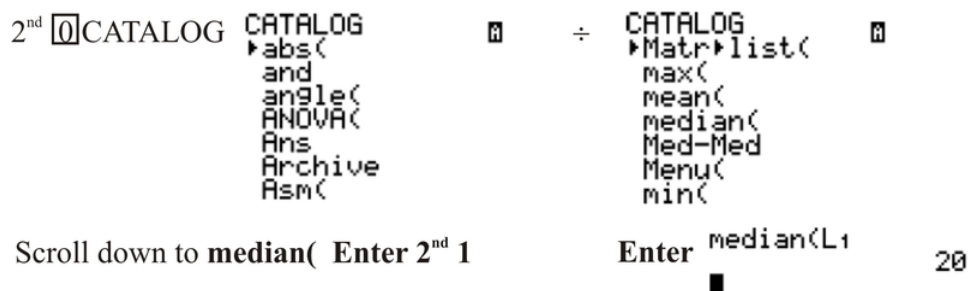
STAT → EDIT **2ND** TESTS Enter 1-Var Stats (L1) Enter →

1-Var Stats	1-Var Stats	↑Sx=12.36624768
2-Var Stats	2-Var Stats	σx=12.24196063
3:Med-Med	(L1) → 2 <sup>nd</sup> 1	n=50
4:LinReg(ax+b)		minX=2
5:QuadReg		Q1=11
6:CubicReg		↓Med=20
7↓QuartReg		

Med = 20 indicates that the median of the data values in L1 is 20.

#### Method Two:

Above the 0 key is the word CATALOG, and this function acts like the yellow pages of a telephone book. When you press **2ND** **0** to access the CATALOG menu, an alphabetical list of terms appears. You can either scroll down to the word median (this will take a long time) or press the blue  $\div$  to access all terms beginning with the letter 'm'.



Again, the median of the values in L1 is 20.

Whichever method you use, the result will be same. Using technology will save you time when you are determining the median of a set of data values. When a set of data values is given in the form of a frequency table, technology is often used to determine the median.

#### Example 14

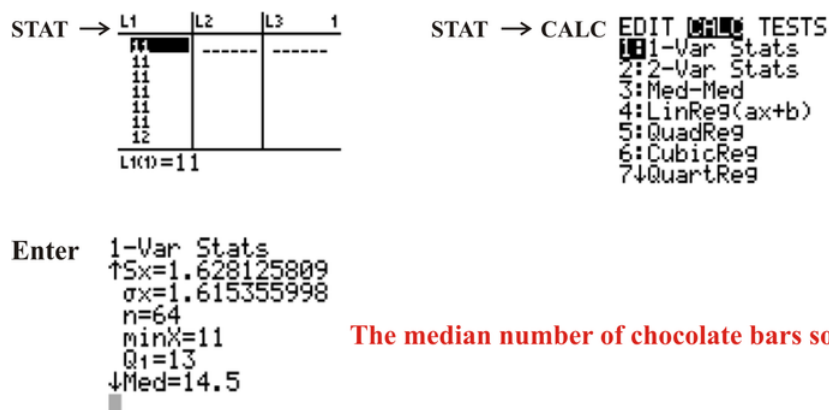
The following data values represent the numbers of chocolate bars sold by students during a recent fundraising campaign. What was the median number of chocolate bars sold?

**TABLE 5.9:**

Number of Bars	Number of Students
11	6
12	8
13	5
14	13
15	17
16	15

#### Solution:

Using the TI-83 calculator:



When data is entered into a frequency table, a column that displays the **cumulative frequency** is often included. This column is simply the sum of the frequencies up to and including that frequency. The median can be determined by using the information that is presented in the cumulative frequency column.

#### Example 15

The following table shows the number of goals scored by Ashton during each of 26 hockey games. What is the median number of goals scored by Ashton during a game?

Goals	Frequency	Cumulative Frequency	
1	6	6 ←	6 scores of 1
2	9	15 ←	15 scores of 1 or 2
3	4	19 ←	19 scores of 1, 2 or 3
4	7	26 ←	26 scores of 1, 2, 3 or 4

**Solution:**

There are 26 data values, which is an even number. The middle position is  $\frac{n+1}{2} = \frac{26+1}{2} = \frac{27}{2} = 13.5$ , and the median is the sum of the numbers above and below position 13.5 divided by 2. According to the table, the numbers in the 13<sup>th</sup> and 14<sup>th</sup> positions are 2's. Therefore, the median is  $\frac{2+2}{2} = \frac{4}{2} = 2$  goals.

**Lesson Summary**

You have learned the significance of the median as it applies to a set of numerical data. You have also learned how to calculate the median of a set of data values, whether the number of values is an odd number or an even number. In addition, you learned that if an outlier affects the mean of a set of data values, then the median is the better measure of central tendency to use. The use of technology in calculating the median was also demonstrated in this lesson.

**Points to Consider**

- The median of a set of data values cuts the data in half. Is only the median of an entire set of data a useful value?
- Is the median of a set of data useful in any other aspect of statistics?
- Other than as a numerical value, can the median be used to represent data in any other way?

## 5.3 The Mode

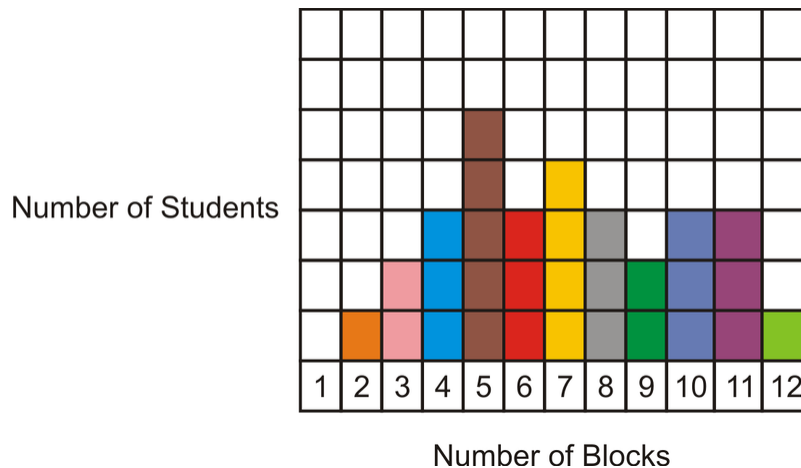
### Learning Objectives

- Understand the concept of mode.
- Identify the mode or modes of a data set for both quantitative and qualitative data.
- Describe a distribution of data as being modal, bimodal, or multimodal.
- Identify the mode of a set of data given in different representations.

Before class begins, bring out the blocks that you and your classmates chose from the pail for the lesson on mean. In addition, have the grid paper on display where each student in your class posted his or her number of blocks.

To begin the class, refer to the comments on the measures of central tendency that were recorded from the lesson on mean, when the brainstorming session occurred. Highlight the comments that were made with regard to the mode of a set of data and discuss this measure of central tendency with your classmates. Once the discussion has been completed, choose a handful of blocks like you did before, with your classmates doing the same.

The mode of the set of blocks can be given as a quantitative value or as a qualitative value. You and your classmates can tell from the grid paper which number of blocks was picked most. The chart below shows that 5 students each picked 5 blocks from the pail. This is a mode for quantitative data, since the answer is in the form of a number.



To extend the mode to include qualitative data, you and your classmates should each now determine the color or colors of block(s) that appear most often in each of your handfuls. To do this, group your blocks according to color and count them, and have your classmates do the same. There may be more than 1 color that occurs with the same highest frequency. The color(s) that appear most often for each handful of blocks is the mode for that particular handful.

The **mode** of a set of data is simply the value that appears most frequently in the set. If 2 or more values appear with the same frequency, each is a mode. The downside to using the mode as a measure of central tendency is that a set of data may have no mode or may have more than 1 mode. However, the same set of data will have only 1 mean and only 1 median. The word *modal* is often used when referring to the mode of a data set. If a data set has only 1 value that occurs most often, the set is called **unimodal**. Likewise, a data set that has 2 values that occur with the greatest frequency is referred to as **bimodal**. Finally, when a set of data has more than 2 values that occur with the same greatest frequency, the set is called **multimodal**. When determining the mode of a data set, calculations are not required, but keen observation is a must. The mode is a measure of central tendency that is simple to locate, but it is not used much in practical applications.

**Example 16**

The posted speed limit along a busy highway is 65 miles per hour. The following values represent the speeds (in miles per hour) of 10 cars that were stopped for violating the speed limit.

76    81    79    80    78    83    77    79    82    75

What is the mode?

**Solution:**

There is no need to organize the data, unless you think that it would be easier to locate the mode if the numbers were arranged from least to greatest. In the above data set, the number 79 appears twice, but all the other numbers appear only once. Since 79 appears with the greatest frequency, it is the mode of the data values.

Mode = 79 miles per hour

**Example 17**

The weekly wages of 7 randomly selected employees of Wendy's were \$98.00, \$125.00, \$75.00, \$120.00, \$86.00, \$92.00, and \$110.00. What is the mode of these wages?

**Solution:**

Each value in the above data set occurs only once. Therefore, this data has no mode.

**Example 18**

The ages of 12 randomly selected customers at a local coffee shop are listed below:

23, 21, 29, 24, 31, 21, 27, 23, 24, 32, 33, 19

What is the mode of the above ages?

**Solution:**

The above data set has 3 values that each occur with a frequency of 2. These values are 21, 23, and 24. All other values occur only once. Therefore, this set of data has 3 modes.

Modes = 21, 23, and 24

Remember that the mode can be determined for qualitative data as well as quantitative data, but the mean and the median can only be determined for quantitative data.

**Example 19**

6 students attending a local swimming competition were asked what color bathing suit they were wearing. The responses were red, blue, black, pink, green, and blue.

What is the mode of these responses?

**Solution:**

The color blue was the only response that occurred more than once and is, therefore, the mode of this data set.

Mode = blue

When data is arranged in a frequency table, the mode is simply the value that has the highest frequency.

**Example 20**

The following table represents the number of times that 100 randomly selected students ate at the school cafeteria during the first month of school.

**5.3. The Mode**

Number of Times Eating in the Cafeteria	2	3	4	5	6	7	8
Number of Students	3	8	22	29	20	8	10

What is the mode of the numbers of times that a student ate at the cafeteria?

**Solution:**

The table shows that 29 students ate 5 times in the cafeteria. Therefore, 5 is the mode of the data set.

Mode = 5 times

**Lesson Summary**

You have learned that the mode of a data set is simply the value that occurs with the highest frequency. You have also learned that it is possible for a set of data to have no mode, 1 mode, 2 modes, or more than 2 modes. Observation is required to determine the mode of a data set, and this mode can be for either quantitative or qualitative data.

**Points to Consider**

- Is reference made to the mode in any other branch of statistics?
- Can the mode be useful when presenting graphical representations of data?

**Vocabulary**

**Bimodal** The term used to describe the distribution of a data set that has 2 modes.

**Cumulative frequency** The sum of the frequencies up to and including that frequency.

**Frequency distribution table** A table that lists a group of data values, as well as the number of times each value appears in the data set.

**Measures of central tendency** Values that describe the center of a distribution. The mean, median, and mode are 3 measures of central tendency.

**Mean** A measure of central tendency that is determined by dividing the sum of all values in a data set by the number of values.

**Median** The value of the middle term in a set of organized data. For a set of data with an odd number of values, it is the value that has an equal number of data values before and after it, or the middle value. For a set of data with an even number of values, the median is the average of the 2 values in the middle positions.

**Mode** The value or values that occur with the greatest frequency in a data set.

**Multimodal** The term used to describe the distribution of a data set that has more than 2 modes.

**Outliers** Extreme values in a data set.

**Unimodal** The term used to describe the distribution of a data set that has only 1 mode.

## 5.4 Review Questions

Show all work necessary to answer each question. Be sure to include any formulas that are needed.

### The Mean

**Section A** – All the review questions in this section are selected response.

1. What is the mean of the following numbers?

10, 39, 71, 42, 39, 76, 38, 25

- a. 42
  - b. 39
  - c. 42.5
  - d. 35.5
2. What name is given to a value in a data set that is much lower or much higher than the other values?
    - a. A sample
    - b. An outlier
    - c. A population
    - d. A tally
  3. What symbol is used to denote the mean of a population?
    - a.  $\Sigma$
    - b.  $\bar{x}$
    - c.  $x_n$
    - d.  $\mu$
  4. What measure of central tendency is calculated by adding all the values and dividing the sum by the number of values?
    - a. median
    - b. mean
    - c. mode
    - d. typical value
  5. The mean of 4 numbers is 71.5. If 3 of the numbers are 58, 76, and 88, what is the value of the 4<sup>th</sup> number?
    - a. 64
    - b. 60
    - c. 76
    - d. 82

**Section B** – All questions in this section require you to show all the work necessary to arrive at a correct solution.

1. Determine the means of the following sets of numbers:
  - a. 20, 14, 54, 16, 38, 64
  - b. 22, 51, 64, 76, 29, 22, 48
  - c. 40, 61, 95, 79, 9, 50, 80, 63, 109, 42
2. The mean weight of 5 men is 167.2 pounds. The weights of 4 of the men are 158.4 pounds, 162.8 pounds, 165 pounds, and 178.2 pounds. What is the weight of the 5<sup>th</sup> man?

3. The mean height of 12 boys is 5.1 feet. The mean height of 8 girls is 4.8 feet.
  - a. What is the total height of the boys?
  - b. What is the total height of the girls?
  - c. What is the mean height of the 20 boys and girls?
4. The following data represents the number of advertisements received by 10 families during the past month. Calculate the mean number of advertisements received by each family during the month.

43    37    35    30    41    23    33    31    16    21

5. The following table of grouped data represents the weight (in pounds) of all 100 babies born at a local hospital last year. Calculate the mean weight for a baby.

**TABLE 5.10:**

Weight (pounds)	Number of Babies
[3 – 5)	8
[5 – 7)	25
[7 – 9)	45
[9 – 11)	18
[11 – 13)	4

6. A group of grade 6 students each earned a mark on an in-class assignment. The marks for the boys were 90, 50, 70, 80, and 70. The marks for the girls were 60, 20, 30, 80, 90, and 20.
  - a. Find the mean mark for the boys.
  - b. Find the mean mark for the girls.
  - c. Find the mean mark for all the students.
7. The mean of 4 numbers is 31. (a) What is the sum of the 4 numbers? The mean of 6 other numbers is 28. (b) Calculate the mean of all 10 numbers.
8. The following numbers represent the weights (in pounds) of 9 dogs:

22    19    26    18    29    33    20    16    30

- a. What is the mean weight of the dogs?
  - b. If the heaviest and the lightest dogs are removed from the group, find the mean weight of the remaining dogs.
9. To demonstrate her understanding of the concept of mean, Melanie recorded the daily temperature in degrees Celsius for her hometown at the same time each day for a period of 1 week. She then calculated the mean daily temperature.

**TABLE 5.11:**

Day	Sun	Mon	Tues	Wed	Thurs	Fri	Sat
Temperature (°C)	–7°C	0°C	–1°C	1°C	–4°C	–6°C	3°C



$$\bar{x} = \frac{-7 + -1 + 1 + -4 + -6 + 3}{6}$$

$$\bar{x} = \frac{-14}{6}$$

$$\bar{x} = -2.3^{\circ}\text{C}$$

Melanie reported the mean daily temperature to be  $-2.3^{\circ}\text{C}$ .

- (a) Is Melanie correct? Justify your answer.  
 (b) If you do not agree with Melanie's answer, can you tell Melanie what mistake she made?

10. Below are the points scored by 2 basketball teams during the regular season for their first 12 games:

Honest Hoopers	93	78	84	106	116	93	90	75	104	100	123	57
Bouncy Baskets	110	89	91	121	84	79	114	66	50	101	106	114

Which team had the higher mean score?

**Section C** – Match the words in the left column with the correct symbol from the right column.

<b>1. Sample mean</b>	<b>A. <math>mf</math></b>
<b>2. The sum of</b>	<b>B. <math>N</math></b>
<b>3. Population mean</b>	<b>C. <math>\bar{x}</math></b>
<b>4. Number of data for a sample</b>	<b>D. <math>\mu</math></b>
<b>5. Product of the midpoint and the frequency</b>	<b>E. <math>n</math></b>
<b>6. Number of data for a population</b>	<b>F. <math>\Sigma</math></b>

### The Median

**Section A** – All the review questions in this section are selected response.

1. What is the median of the following numbers?

10, 39, 71, 42, 39, 76, 38, 25

- a. 42.5  
 b. 39  
 c. 42

- d. 35.5
2. The front row in a movie theatre has 23 seats. If you were asked to sit in the seat that occupied the median position, in what number seat would you have to sit?
- 1
  - 11
  - 23
  - 12
3. What is the median mark achieved by a student who recorded the following marks on 10 math quizzes?

68, 55, 70, 62, 71, 58, 81, 82, 63, 79

- 68
  - 71
  - 69
  - 79
4. A set of 4 numbers that begins with the number 32 is arranged from smallest to largest. If the median is 35, which of these could possibly be the set of numbers?
- 32, 34, 36, 38
  - 32, 35, 38, 41
  - 32, 33, 34, 35
  - 32, 36, 40, 44
5. The number of chocolate bars sold by each of 30 students is as follows:

32, 6, 21, 10, 8, 11, 12, 36, 17, 16, 15, 18, 40, 24, 21, 23, 24, 24, 29, 16, 32, 31, 10, 30, 35, 32, 18, 39, 12, 20

What is the median number of chocolate bars sold by the 30 students?

- 18
- 21
- 24
- 32

**Section B** – All questions in this section require you to show all the work necessary to arrive at a correct solution.

1. The following table lists the retail price and the dealer's costs for 10 cars at a local car lot this past year: **Deals on Wheels**



TABLE 5.12:

Car Model	Retail Price	Dealer's Cost
Nissan Sentra	\$24,500	\$18,750
Ford Fusion	\$26,450	\$21,300
Hyundai Elantra	\$22,660	\$19,900
Chevrolet Malibu	\$25,200	\$22,100
Pontiac Sunfire	\$16,725	\$14,225
Mazda 5	\$27,600	\$22,150
Toyota Corolla	\$14,280	\$13,000
Honda Accord	\$28,500	\$25,370
Volkswagen Jetta	\$29,700	\$27,350
Subaru Outback	\$32,450	\$28,775

- (a) Calculate the median for the data on the retail prices for the above cars.
- (b) Calculate the median for the data on the dealer's costs for the above cars.
2. Due to high winds, a small island in the Atlantic suffers frequent power outages. The following numbers represent the number of outages each month during the past year:

4 5 3 4 2 1 0 3 2 7 2 3

What is the median number of monthly power outages?

3. The Canadian Coast Guard has provided all of its auxiliary members with a list of 14 safety items (flares, fire extinguishers, life jackets, fire buckets, etc.) that must be aboard each boat at all times. During a recent check of 15 boats, the number of safety items that were aboard each boat was recorded as follows:

7 14 10 5 11 2 8 6 9 7 13 4 12 8 3

What is the median number of safety items aboard the boats that were checked?

4. A teacher's assistant who has been substituting has been recording her biweekly wages for the past 13 pay periods. Her biweekly wages during this time were the following:

\$700    \$550    \$760    \$670    \$500    \$925    \$600  
 \$480    \$390    \$800    \$850    \$365    \$525

What is her median biweekly wage?

5. A minor hockey league has 50 active players who range in age from 10 years to 15 years. The following table shows the ages of the players:

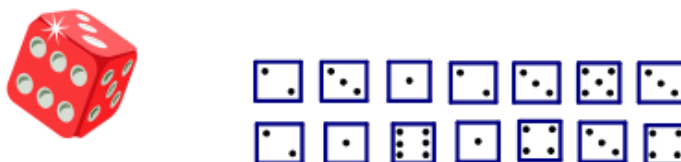


**TABLE 5.13:**

Age of Players(yrs)	Number of Players
10	6
11	8
12	7
13	11
14	10
15	8

Use the TI-83 calculator to determine the median age of the players.

6. A die was thrown 14 times, and the results of each throw are shown:



What is the median score for the 14 rolls of the die?

7. At a local golf club, 100 players competed in a 1-day tournament. The fifth hole of the course is a par 6. The scores of each player on this hole were recorded, and the results are shown below:



Score	2	3	4	5	6	7	8
Number of Players	3	8	22	29	20	8	10

What is the median score for the players?

8. A math teacher at a local high school begins every class with a warm-up quiz based on the work presented the previous day. Each test consists of 6 questions valued at 1 point for each correct answer. The results of Monday’s quiz for the 40 students in the class are shown below:



Mark	0	1	2	3	4	5	6
Number of Students	2	6	7	10	5	3	7

What is the median score for Monday's quiz?

9. In Canada, with the loonie and the toonie, you could have a lot of coins in your pocket. A number of high school students were asked how many coins they had in their pocket, and the results are shown in the following table:



Number of Coins	0	1	2	3	4	5	6	7
Number of Students	2	5	8	9	6	4	9	7

What is the median number of coins that a student had in his or her pocket?

10. A group of 12 students participated in a local dirt bike race that required them to cover a 1-mile course in the fastest time possible. The times, in minutes, of the 12 participants are shown below:



3.5 min   4.2 min   3.1 min   5.3 min   6.2 min   4.6 min  
 5.1 min   6.7 min   5.4 min   4.4 min   3.9 min   5.0 min

What is the median time of the participants in the race?

#### 5.4. Review Questions

## The Mode

**Section A** – All the review questions in this section are selected response.

- Which of the following measures can be determined for quantitative data only?
  - mean
  - median
  - mode
  - none of these
- Which of the following measures can be calculated for qualitative data only?
  - mean
  - median
  - mode
  - all of these
- What is the term used to describe the distribution of a data set that has 1 mode?
  - multimodal
  - unimodal
  - nonmodal
  - bimodal
- What is the mode of the following numbers, which represent the ages of 8 hockey players? 12, 11, 14, 10, 8, 13, 11, 9
  - 11
  - 10
  - 14
  - 8
- Which of the following measures can have more than 1 value for a set of data?
  - median
  - mode
  - mean
  - none of these

**Section B** – All questions in this section require your answer to be a complete sentence.

- What are the modes of the following sets of numbers?
  - 3, 13, 6, 8, 10, 5 6
  - 12, 0, 15, 15, 13, 19, 16, 13, 16, 16
- A student recorded her scores on weekly English quizzes that were marked out of a possible 10 points. Her scores were as follows:

8, 5, 8, 5, 7, 6, 7, 7, 5, 7, 5, 5, 6, 6, 9, 8, 9, 7, 9, 9, 6, 8, 6, 6, 7

What is the mode of her scores on the weekly English quizzes?

- The following table represents the number of minutes that students spent studying for a math test:

**TABLE 5.14:**

Studying Time (minutes)	Number of Students
[0 – 10)	2
[10 – 20)	10

**TABLE 5.14:** (continued)

Studying Time (minutes)	Number of Students
[20 – 30)	6
[30 – 40)	4
[40 – 50)	3

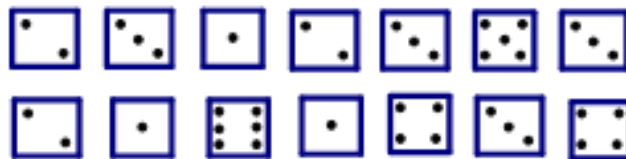
What is the mode of the amounts of time spent studying for the math test?

4. A pre-test for students entering high school mathematics was given to 48 students. The following table shows the number of questions attempted out of 50 by each of the students taking the test:

43 39 40 40 41 44  
 42 41 41 41 42 44  
 43 41 40 41 42 41  
 42 41 42 42 41 44  
 41 42 43 40 42 39  
 42 40 39 42 43 42  
 42 39 41 41 42 40  
 43 44 40 42 44 39

What number of questions was attempted the most by the students?

5. A die was tossed 14 times. What is the mode of the numbers that were rolled?



6. The following table represents the number of times that 24 students attended school basketball games during the year:

Number of Games	5	6	7	8	9
Number of Students	1	5	3	9	6

What is the mode of the numbers of games that students attended?

7. The newly-formed high school soccer team is playing its first season. The following table shows the number of goals it scored during each of its matches:

Number of Goals	1	2	3
Number of Matches	8	8	$m$

If the mean number of goals scored is 2.04, what is the smallest possible value of  $m$  if the mode of the numbers of goals scored is 3?

#### 5.4. Review Questions

8. What is the mode of the following numbers, and what word can be used to describe the distribution of the data set?

5, 4, 10, 3, 3, 4, 7, 4, 6, 5, 11, 9, 5, 7

9. List 3 examples of how mode could be useful in everyday life?
10. The temperature in °F on 20 days during the month of June was as follows:

70°F, 76°F, 76°F, 74°F, 70°F, 70°F, 72°F, 74°F, 78°F, 80°F

74°F, 74°F, 78°F, 76°F, 78°F, 76°F, 74°F, 78°F, 80°F, 76°F

What is the mode of the temperatures for the month of June?



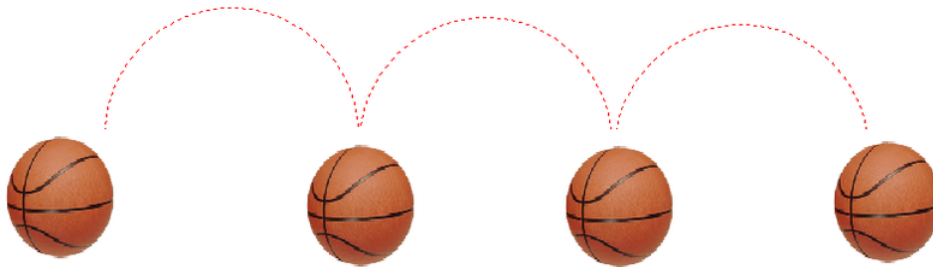
**CHAPTER 6**

# The Shape, Center and Spread of a Normal Distribution

## Chapter Outline

- 6.1 ESTIMATING THE MEAN AND STANDARD DEVIATION OF A NORMAL DISTRIBUTION
- 6.2 CALCULATING THE STANDARD DEVIATION
- 6.3 CONNECTING THE STANDARD DEVIATION AND NORMAL DISTRIBUTION
- 6.4 REVIEW QUESTIONS

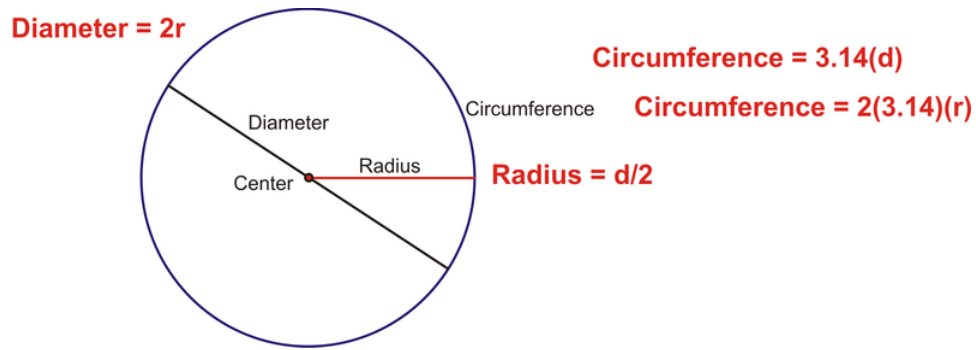
### Introduction



What a way to start math class! Imagine your math teacher allowing you to bounce a basketball around the classroom.



Prior to the beginning of class, go to the physical education department of your school and borrow some basketballs. These will be used in this chapter to introduce the concept of a normal distribution. You may find your class to be somewhat noisy, but you're sure to enjoy the activity. Begin the class by reviewing the concept of a circle and the terms associated with the measurements of a circle. You and your classmates should be able to provide these facts based on your previous learning. To ensure that everyone understands the concepts of center, radius, diameter, and circumference, work in small groups to create posters to demonstrate your understanding. These posters can then be displayed around the classroom. The following is a sample of the type of poster you may create:

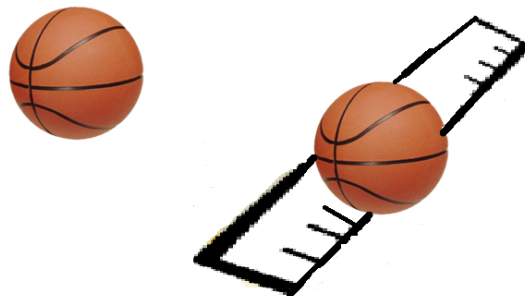


Once you and your classmates have completed this and you are confident that you all understand the measurements associated with a circle, distribute the basketballs to everyone in the class. While your classmates are playing with the balls, set up a table with tape, rulers, string, scissors, markers, and any other materials that you think you may find useful to answer the following questions:

- What is the circumference of the basketball?
- What is the diameter of the basketball?
- What is the radius of the basketball?
- How did you determine these measurements?
- What tools did you use to find your answers?



When playtime is over, use the tools provided to answer all of the above questions. It is the job of you and your classmates to determine a method of calculating these measurements. For the questions that require numerical measurements as answers, take 2 measurements for each. You must plot your results for the diameter of the basketball, so remember to record your data.



Oops! Don't forget that the ruler cannot go through the basketball!

When you have answered the above questions, plot your 2 results for the diameter of the basketball on a large sheet of grid paper, and have your classmates do the same.

## 6.1 Estimating the Mean and Standard Deviation of a Normal Distribution

### Learning Objectives

- Understand the meaning of normal distribution and bell-shape.
- Estimate the mean and the standard deviation of a normal distribution.

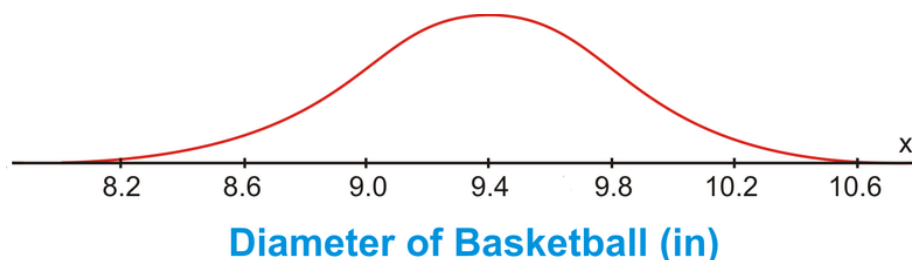
Now that you have created your plot on a large sheet of grid paper, can you describe the shape of the plot? Do the dots seem to be clustered around 1 spot (value) on the chart? Do some dots seem to be far away from the clustered dots? After you have made all the necessary observations to answer these questions, pick 2 numbers from the chart to complete this statement:

“The typical measurement of the diameter is approximately \_\_\_\_\_ inches, give or take \_\_\_\_\_ inches.”

We will complete this statement later in the lesson.

### Normal Distribution

The shape below should be similar to the shape that has been created with the dot plot.



When you made the observations regarding the measurements of the diameter of the basketball, you must have noticed that they were not all the same. In spite of the different measurements, you should have seen that the majority of the measurements clustered around the value of 9.4 inches. This value represents the approximate diameter of a basketball. Also, you should have noticed that a few measurements were to the right of this value, and a few measurements were to the left of this value. The resulting shape looks like a bell, and this is the shape that represents a **normal distribution** of data.

In the real world, no examples match this smooth curve perfectly. However, many data plots, like the one you made, will approximate this smooth curve. For this reason, you will notice that the term *assume* is often used when referring to data that deals with normal distributions. When a normal distribution is assumed, the resulting bell-shaped curve is symmetric. That is, the right side is a mirror image of the left side. In the figure below, if the blue line is the mirror (the line of symmetry), you can see that the pink section to the left of the line of symmetry is the mirror image of the yellow section to the right of the line of symmetry. The line of symmetry also goes through the  $x$ -axis.



If you knew all of the measurements that were plotted for the diameter of the basketball, you could calculate the mean (average) diameter by adding the measurements and dividing the sum by the total number of values. It is at

this value that the line of symmetry intersects the  $x$ -axis. In other words, the mean of a normal distribution is the center, or balance point, of the distribution.

You can see that the 2 colors form a peak at the top of the line of symmetry and then spread out to the left and to the right from the line of symmetry. The shape of the bell flattens out the further it moves away from the line of symmetry. In other words, the data spreads out in both directions away from the mean. This spread of the data is measured by the **standard deviation**, and it describes exactly how the data moves away from the mean. You will learn more about standard deviation in the next lesson. For now, that is all you have to know about standard deviation—it is a measure of the spread of the data away from the mean.

Now you should be able to complete the statement that was presented earlier in this lesson.

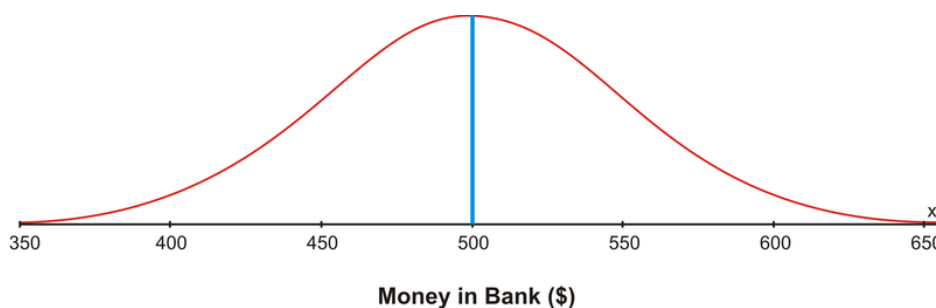
“The typical measurement of the diameter is approximately 9.4 inches, give or take 0.2 inches.”

This statement assumes that the mean of the measurements was 9.4 inches and the standard deviation of the measurements was 0.2 inches.

### Example 1

For each of the following graphs, complete the statement. Fill in the first blank in each statement with the mean and the second blank in each statement with the standard deviation. Assume that the standard deviation is the difference between the mean and the first tick mark to the left of the mean.

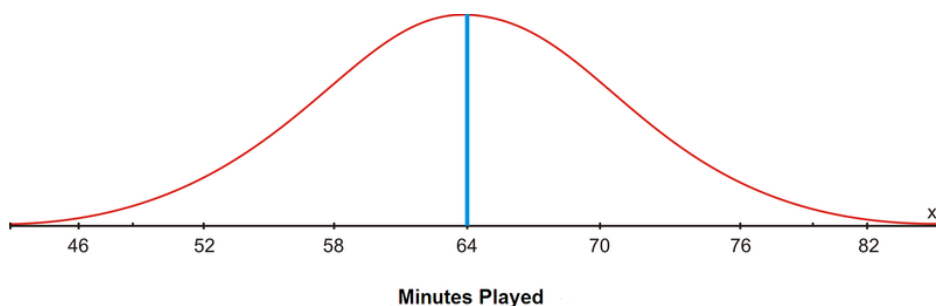
a) “The typical measurement is approximately \_\_\_\_\_ in the bank, give or take \_\_\_\_\_.”



### Solution:

“The typical measurement is approximately \$500 in the bank, give or take \$50 .”

b) “The typical measurement is approximately \_\_\_\_\_ goals scored, give or take \_\_\_\_\_ goals.”



### Solution:

“The typical measurement is approximately 64 goals scored, give or take 6 goals.”

### Lesson Summary

In this lesson, you learned what was meant by normal distribution. You also learned about the smooth bell curve that is used to represent a data set that is normally distributed. In addition, you learned that when data is plotted on a

#### 6.1. Estimating the Mean and Standard Deviation of a Normal Distribution

bell curve, you can estimate the mean by using the value where the line of symmetry crosses the  $x$ -axis. Finally, the spread of data in a normal distribution was represented by using a give or take statement.

**Points to Consider**

- Is there a way to determine the actual values for a give or take statement?
- Can a give or take statement go beyond a single give or take?
- Can all the actual values be represented on a bell curve?

## 6.2 Calculating the Standard Deviation

### Learning Objectives

- Understand the meaning of standard deviation.
- Understanding the percents associated with standard deviation.
- Calculate the standard deviation for a normally distributed random variable.

This semester you decided to join the school's bowling league for the first time. Having never bowled previously, you are very anxious to find out what your bowling average is for the semester. If your average is comparable to that of the other students, you will join the league again next semester. Your coach has told you that your mean score is 70, and now you want to find out how your results compare to that of the other members of the league.

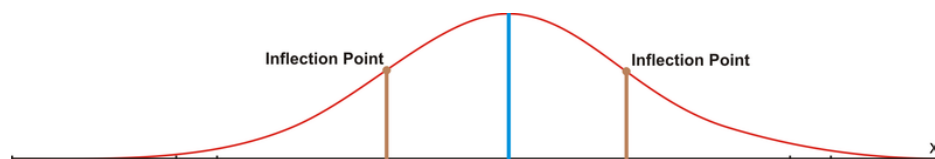
Your coach has decided to let you figure this out for yourself. He tells you that the scores were normally distributed and provides you with a list of the other mean scores. These average scores are in no particular order. In other words, they are random.

54	88	49	44	96	72	46	58	79
92	44	50	102	80	72	66	64	61
60	56	48	52	54	60	64	72	68
64	60	56	52	55	60	62	64	68

We will discover how your mean bowling score compares to that of the other bowlers later in the lesson.

### Standard Deviation

In the previous lesson, you learned that standard deviation is a measure of the spread of a set of data away from the mean of the data.

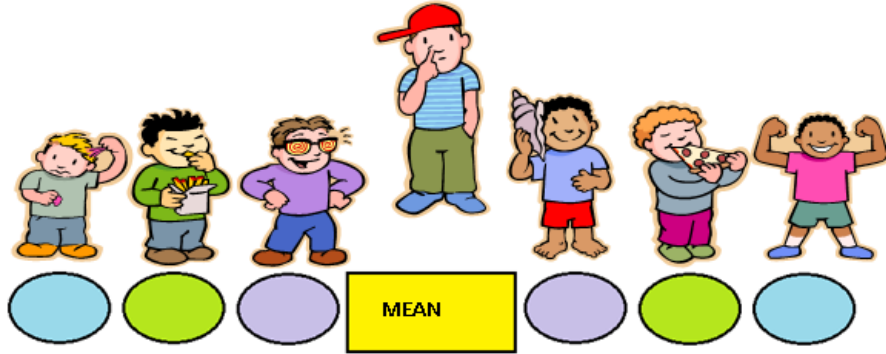


In a normal distribution, on either side of the line of symmetry, the curve appears to change its shape from being concave down (looking like an upside-down bowl) to being concave up (looking like a right-side-up bowl). Where this happens is called an inflection point of the curve. If a vertical line is drawn from an inflection point to the  $x$ -axis, the difference between where the line of symmetry goes through the  $x$ -axis and where this line goes through the  $x$ -axis represents 1 standard deviation away from the mean. Approximately 68% of all the data is located within 1 standard deviation of the mean.

To emphasize this fact and the fact that the mean is the middle of the distribution, let's play a game of Simon Says. Using color paper and 2 types of shapes, arrange the pattern of the shapes on the floor as shown below. Randomly select 7 students from your class to play the game. You will be Simon, and you are to give orders to the selected students. Only when *Simon Says* are the students to obey the given order. The orders can be given in many ways, but 1 suggestion is to deliver the following orders:

- "Simon Says for Frank to stand on the rectangle."

- “Simon Says for Joey to stand on the closest oval to the right of Frank.”
- “Simon Says for Liam to stand on the closest oval to the left of Frank.”
- “Simon Says for Mark to stand on the farthest oval to the right of Frank.”
- “Simon Says for Juan to stand on the farthest oval to the left of Frank.”
- “Simon Says for Jacob to stand on the middle oval to the right of Frank.”
- “Simon Says for Sean to stand on the middle oval to the left of Frank.”

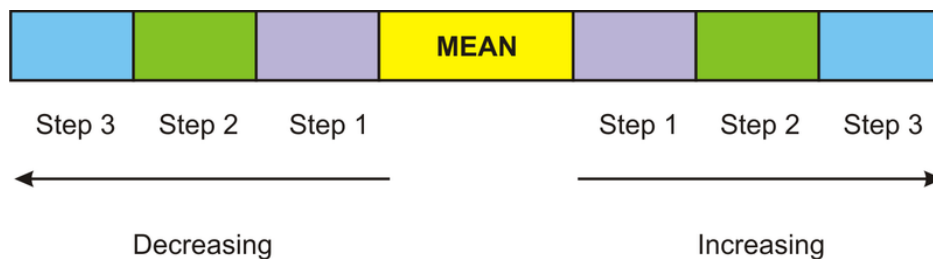


Once the students are standing in the correct places, pose questions about their positions with respect to Frank. The members of the class who are not playing the game should be asked to respond to these questions about the position of their classmates. Some questions that should be asked are the following:

- “Which 2 students are standing closest to Frank?”
- “Are Joey and Liam both the same distance away from Frank?”
- “Which 2 students are furthest away from Frank?”
- “Are Mark and Juan both the same distance away from Frank?”

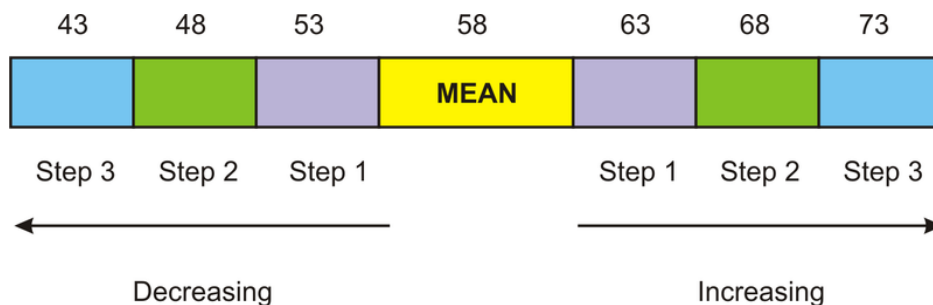
When the students have completed playing Simon Says, they should have an understanding of the concept that the mean is the middle of the distribution and the remainder of the distribution is evenly spread out on either side of the mean.

The picture below is a simplified form of the game you have just played. The yellow rectangle is the mean, and the remaining rectangles represent 3 steps to the right of the mean and 3 steps to the left of the mean.

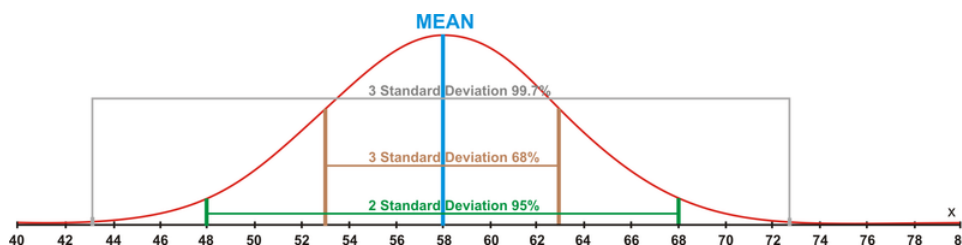


If we consider the spread of the data away from the mean, which is measured using standard deviation, as being a stepping process, then 1 step to the right or 1 step to the left is considered 1 standard deviation away from the mean. 2 steps to the left or 2 steps to the right are considered 2 standard deviations away from the mean. Likewise, 3 steps to the left or 3 steps to the right are considered 3 standard deviations away from the mean. The standard deviation of a data set is simply a value, and in relation to the stepping process, this value would represent the size of your footstep as you move away from the mean. Once the value of the standard deviation has been calculated, it is added to the mean for moving to the right and subtracted from the mean for moving to the left. If the value of the yellow mean tile was 58, and the value of the standard deviation was 5, then you could put the resulting sums and differences on the appropriate tiles.





For a normal distribution, 68% of the data values would be located within 1 standard deviation of the mean, which is between 53 and 63. Also, 95% of the data values would be located within 2 standard deviations of the mean, which is between 48 and 68. Finally, 99.7% of the data values would be located within 3 standard deviations of the mean, which is between 43 and 73. The percentages mentioned here make up what statisticians refer to as the **68-95-99.7 Rule**. These percentages remain the same for all data that can be assumed to be normally distributed. The following diagram represents the location of these values on a normal distribution curve.



Now that you understand the distribution of the data and exactly how it moves away from the mean, you are ready to calculate the standard deviation of a data set. For the calculation steps to be organized, a table is used to record the results for each step. The table will consist of 3 columns. The first column will contain the data and will be labeled  $x$ . The second column will contain the differences between the data values and the mean of the data set. This column will be labeled  $(x - \bar{x})$ . The final column will be labeled  $(x - \bar{x})^2$ , and it will contain the square of each of the values recorded in the second column.

### Example 2

Calculate the standard deviation of the following numbers:

$$2, 7, 5, 6, 4, 2, 6, 3, 6, 9$$

### Solution:

**Step 1:** It is not necessary to organize the data. Create a table and label each of the columns appropriately. Write the data values in column  $x$ .

**Step 2:** Calculate the mean of the data values.

$$\mu = \frac{2 + 7 + 5 + 6 + 4 + 2 + 6 + 3 + 6 + 9}{10} = \frac{50}{10} = 5.0$$

**Step 3:** Calculate the differences between the data values and the mean. Enter the results in the second column.

## 6.2. Calculating the Standard Deviation

TABLE 6.1:

$x$	$(x - \mu)$
2	-3
7	2
5	0
6	1
4	-1
2	-3
6	1
3	-2
6	1
9	4

---

**Step 4:** Calculate the values for column 3 by squaring each result in the second column.

**Step 5:** Calculate the mean of the third column and then take the square root of the answer. This value is the standard deviation ( $\sigma$ ) of the data set.

TABLE 6.2:

$(x - \mu)^2$
9
4
0
1
1
9
1
4
1
16

---

$$\sigma^2 = \frac{9 + 4 + 0 + 1 + 1 + 9 + 1 + 4 + 1 + 16}{10} = \frac{46}{10} = 4.6$$

$$\sigma = \sqrt{4.6} \approx 2.1$$

Step 5 can be written using the formula  $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$ .

The standard deviation of the data set is approximately 2.1.

Now that you have completed all the steps, here is the table that was used to record the results. The table was separated as the steps were completed. Now that you know the process involved in calculating the standard deviation, there is no need to work with individual columns—work with an entire table.

TABLE 6.3:

$x$	$(x - \mu)$	$(x - \mu)^2$
2	-3	9
7	2	4
5	0	0

**TABLE 6.3:** (continued)

$x$	$(x - \mu)$	$(x - \mu)^2$
6	1	1
4	-1	1
2	-3	9
6	1	1
3	-2	4
6	1	1
9	4	16

**Example 3**

A company wants to test its exterior house paint to determine how long it will retain its original color before fading. The company mixes 2 brands of paint by adding different chemicals to each brand. 6 one-gallon cans are made for each paint brand, and the results are recorded for every gallon of each brand of paint. The following are the results obtained in the laboratory:

**TABLE 6.4:**

Brand A (Time in months)	Brand B (Time in months)
15	40
65	50
55	35
35	40
45	45
25	30

Calculate the standard deviation for each brand of paint.

**Solution:**

Brand A:

**TABLE 6.5:**

$x$	$(x - \mu)$	$(x - \mu)^2$
15	-25	625
65	25	625
55	15	225
35	-5	25
45	5	25
25	-15	225

$$\mu = \frac{15 + 65 + 55 + 35 + 45 + 25}{6} = \frac{240}{6} = 40$$

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{625 + 625 + 225 + 25 + 25 + 225}{6}}$$

$$\sigma = \sqrt{\frac{1750}{6}} \approx \sqrt{291.66} \approx 17.1$$

The standard deviation for Brand A is approximately 17.1.

Brand B:

**TABLE 6.6:**

$x$	$(x - \mu)$	$(x - \mu)^2$
40	0	0
50	10	100
35	-5	25
40	0	0
45	5	25
30	-10	100

$$\mu = \frac{40 + 50 + 35 + 40 + 45 + 30}{6} = \frac{240}{6} = 40$$

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{0 + 100 + 25 + 0 + 25 + 100}{6}}$$

$$\sigma = \sqrt{\frac{250}{6}} \approx \sqrt{41.66} \approx 6.5$$

The standard deviation for Brand B is approximately 6.5.

Note: The standard deviation for Brand A (17.1) was much larger than that for Brand B (6.5). However, the means of both brands were the same. When the means are equal, the larger the standard deviation is, the more variable are the data.

To find the standard deviation, you subtract the mean from each data value to determine how much each data value varies from the mean. The result is a positive value when the data value is greater than the mean, a negative value when the data value is less than the mean, and 0 when the data value is equal to the mean.

If we were to add the variations found in the second column of the table, the total would be 0. This result of 0 implies that there is no variation between the data value and the mean. In other words, if we were conducting a survey of the number of hours that students use a cell phone in 1 day, and we relied upon the sum of the variations to give us some pertinent information, the only thing that we would learn is that all the students who participated in the survey use a cell phone for the exact same number of hours each day. We know that this is not true, because the survey does not show all the responses as being the same. In order to ensure that these variations do not lose their significance when added, the variation values are squared prior to calculating their sum.

What we need for a normal distribution is a measure of spread that is proportional to the scatter of the data, independent of the number of values in the data set and independent of the mean. The spread will be small when the data values are consistent, but large when the data values are inconsistent. The reason that the measure of spread should be independent of the mean is because we are not interested in this measure of central tendency, but rather, only in the spread of the data. For a normal distribution, both the variance and the standard deviation fit the above profile for an appropriate measure of spread, and both values can be calculated for the set of data.

To calculate the **variance** ( $\sigma^2$ ) for a population of normally distributed data:

**Step 1:** Determine the mean of the data values.

**Step 2:** Subtract the mean of the data from each value in the data set to determine the difference between the data value and the mean:  $(x - \mu)$ .

**Step 3:** Square each of these differences and determine the total of these positive, squared results.

**Step 4:** Divide this sum by the number of values in the data set.

These steps for calculating the variance of a data set for a population can be summarized in the following formula:

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n}$$

where:

$x$  is a data value.

$\mu$  is the population mean.

$n$  is number of data values (population size).

These steps for calculating the variance of a data set for a sample can be summarized in the following formula:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

where:

$x$  is a data value.

$\bar{x}$  is the sample mean.

$n$  is number of data values (sample size).

The only difference in the formulas is the number by which the sum is divided. For a population, it is divided by  $n$ , and for a sample, it is divided by  $n - 1$ .

#### **Example 4**

Calculate the variance of the 2 brands of paint in Example 3. These are both small populations.

**TABLE 6.7:**

<b>Brand A (Time in months)</b>	<b>Brand B (Time in months)</b>
15	40
65	50
55	35
35	40
45	45
25	30

#### **Solution:**

Brand A

**TABLE 6.8:**

$x$	$(x - \mu)$	$(x - \mu)^2$
15	-25	625
65	25	625
55	15	225
35	-5	25

TABLE 6.8: (continued)

$x$	$(x - \mu)$	$(x - \mu)^2$
45	5	25
25	-15	225

$$\mu = \frac{15 + 65 + 55 + 35 + 45 + 25}{6} = \frac{240}{6} = 40$$

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n}$$

$$\sigma^2 = \frac{625 + 625 + 225 + 25 + 25 + 225}{6} = \frac{1750}{6} \approx 291.\overline{66}$$

Brand B

TABLE 6.9:

$x$	$(x - \mu)$	$(x - \mu)^2$
40	0	0
50	10	100
35	-5	25
40	0	0
45	5	25
30	-10	100

$$\mu = \frac{40 + 50 + 35 + 40 + 45 + 30}{6} = \frac{240}{6} = 40$$

$$\sigma^2 = \frac{\sum(x - \mu)^2}{n}$$

$$\sigma^2 = \frac{0 + 100 + 25 + 0 + 25 + 100}{6} = \frac{250}{6} \approx 41.\overline{66}$$

From the calculations done in Example 3 and in Example 4, you should have noticed that the square root of the variance is the standard deviation, and the square of the standard deviation is the variance. Taking the square root of the variance will put the standard deviation in the same units as the given data. The variance is simply the average of the squares of the distance of each data value from the mean. If these data values are close to the value of the mean, the variance will be small. This was the case for Brand B. If these data values are far from the mean, the variance will be large, as was the case for Brand A.

The variance and the standard deviation of a data set are always positive values.

**Example 5**

The following data represents the morning temperatures ( $^{\circ}\text{C}$ ) and the monthly rainfall (mm) in July for all the Canadian cities east of Toronto:

Temperature ( $^{\circ}\text{C}$ )

11.7 13.7 10.5 14.2 13.9 14.2 10.4 16.1 16.4  
4.8 15.2 13.0 14.4 12.7 8.6 12.9 11.5 14.6

Precipitation (mm)

18.6 37.1 70.9 102 59.9 58.0 73.0 77.6 89.1  
86.6 40.3 119.5 36.2 85.5 59.2 97.8 122.2 82.6

Which data set is more variable? Calculate the standard deviation for each data set.

**Solution:**

Temperature ( $^{\circ}\text{C}$ )

**TABLE 6.10:**

$x$	$(x - \mu)$	$(x - \mu)^2$
11.7	-1	1
13.7	1	1
10.5	-2.2	4.84
14.2	1.5	2.25
13.9	1.2	1.44
14.2	1.5	2.25
10.4	-2.3	5.29
16.1	3.4	11.56
16.4	3.7	13.69
4.8	-7.9	62.41
15.2	2.5	6.25
13.0	0.3	0.09
14.4	1.7	2.89
12.7	0	0
8.6	-4.1	16.81
12.9	0.2	0.04
11.5	-1.2	1.44
14.6	1.9	3.61

$$\mu = \frac{\sum x}{n} = \frac{228.6}{18} \approx 12.7$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

$$\sigma^2 = \frac{136.86}{18} \approx 7.6$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

$$\sigma = \sqrt{\frac{136.86}{18}} \approx 2.8$$

The variance of the data set is approximately  $7.6^{\circ}\text{C}$ , and the standard deviation of the data set is approximately  $2.8^{\circ}\text{C}$ .

Precipitation (mm)

**TABLE 6.11:**

$x$	$(x - \mu)$	$(x - \mu)^2$
18.6	-54.5	2970.3
37.1	-36.0	1296

TABLE 6.11: (continued)

$x$	$(x - \mu)$	$(x - \mu)^2$
70.9	-2.2	4.84
102.0	28.9	835.21
59.9	-13.2	174.24
58.0	-15.1	228.01
73.0	-0.1	0.01
77.6	4.5	20.25
89.1	16.0	256
86.6	13.5	182.25
40.3	-32.8	1075.8
119.5	46.4	2153
36.2	-36.9	1361.6
85.5	12.4	153.76
59.2	-13.9	193.21
97.8	24.7	610.09
122.2	49.1	2410.8
82.6	9.5	90.25

$$\mu = \frac{\sum x}{n} = \frac{1316.1}{18} \approx 73.1$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

$$\sigma^2 = \frac{14016}{18} \approx 778.\overline{66}$$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{14016}{18}} \approx 27.9$$

The variance of the data set is approximately 778.66 mm, and the standard deviation of the data set is approximately 27.9 mm.

Therefore, the data values for the precipitation are more variable. This is indicated by the large variance of the data set.

### Example 6

Now that you know how to calculate the variance and the standard deviation of a set of data, let's apply this to a normal distribution by determining how your bowling average compared to those of the other bowlers in your league. This time technology will be used to determine both the variance and the standard deviation of the data.

54 88 49 44 96 72 46 58 79  
 92 44 50 102 80 72 66 64 61  
 60 56 48 52 54 60 64 72 68  
 64 60 56 52 55 60 62 64 68

**Solution:**



Stat → Enter → L1 | L2 | L3 | 1 Stat → Calc → EDIT TESTS

L1	L2	L3	1
54	-----	-----	
88			
49			
44			
96			
72			
46			
L1()=54			

1:1-Var Stats  
2:2-Var Stats  
3:Med-Med  
4:LinReg(ax+b)  
5:QuadReg  
6:CubicReg  
7↓QuartReg

Enter → 1-Var Stats L1 → Enter → 1-Var Stats

$\bar{x}=63.66666667$   
 $\Sigma x=2292$   
 $\Sigma x^2=153068$   
 $Sx=14.2868571$   
 $\sigma x=14.08703107$   
 $\downarrow n=36$

From the list, you can see that the mean of the bowling averages is approximately 63.7 and that the standard deviation is approximately 14.1.

To use technology to calculate the variance involves naming the lists according to the operations that you need to do in order to determine the correct values. In addition, you can use the CATALOG menu of the calculator to determine the sum of the squared variations. All of the same steps used to calculate the standard deviation of the data are applied to give the mean of the data set. You could also use the CATALOG menu to find the mean of the data, but since you are now familiar with 1-Var Stats, you can use this method.

Stat → Enter → L1 | L2 | L3 | 1 Stat → Calc → EDIT TESTS

L1	L2	L3	1
54	-----	-----	
88			
49			
44			
96			
72			
46			
L1()=54			

1:1-Var Stats  
2:2-Var Stats  
3:Med-Med  
4:LinReg(ax+b)  
5:QuadReg  
6:CubicReg  
7↓QuartReg

Enter → 1-Var Stats L1 → Enter → 1-Var Stats

$\bar{x}=63.66666667$   
 $\Sigma x=2292$   
 $\Sigma x^2=153068$   
 $Sx=14.2868571$   
 $\sigma x=14.08703107$   
 $\downarrow n=36$

The mean of the data is approximately 63.7. L2 will now be renamed L1 - 63.7 to compute the values for  $(x - \bar{x})$ .

Likewise, L3 will be renamed L2<sup>2</sup>.

## 6.2. Calculating the Standard Deviation

Stat → Enter →

L1	$\bar{x}$	L3	2
54			
88			
49			
44			
96			
72			
46			

L2 = L1 - 63.7

→ Enter →

L1	L2	L3	2
54	-9.7		
88	24.3		
49	-14.7		
44	-19.7		
96	32.3		
72	8.3		
46	-17.7		

L2(1) = -9.7

Stat → Enter →

L1	L2	$\bar{x}^2$	3
54	-9.7		
88	24.3		
49	-14.7		
44	-19.7		
96	32.3		
72	8.3		
46	-17.7		

L3 = L2<sup>2</sup>

→ Enter →

L1	L2	L3	3
54	-9.7	94.09	
88	24.3	590.49	
49	-14.7	216.09	
44	-19.7	388.09	
96	32.3	1043.3	
72	8.3	68.89	
46	-17.7	313.29	

L3(1) = 94.09

2<sup>nd</sup> 0 ( Catalogue) → Ln ( S) → CATALOG and scroll down to sum( → Enter

- ▶ 2-SampFTest
- ▶ 2-SampTInt
- ▶ 2-SampTTest
- ▶ 2-SampZInt(
- ▶ 2-SampZTest(
- Scatter
- Sci

Here we type in 2<sup>nd</sup> 3 → L<sub>3</sub> → Enter sum(L<sub>3</sub>)

7144.04

Ans/36

198.4455556

The sum of the values in L3 divided by the number of data values (36) is the variance of the bowling averages.

### Lesson Summary

In this lesson, you learned that the standard deviation of a set of data is a value that represents a measure of the spread of the data from the mean. You also learned that the variance of the data from the mean is the square of the standard deviation. Calculating the standard deviation manually and calculating it by using technology were additional topics you learned in this lesson.

### Points to Consider

- Does the value of standard deviation stand alone, or can it be displayed with a normal distribution?
- Are there defined increments for how data spreads away from the mean?
- Can the standard deviation of a set of data be applied to real-world problems?

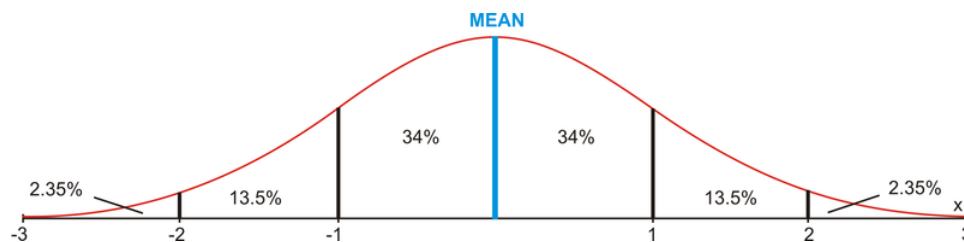
## 6.3 Connecting the Standard Deviation and Normal Distribution

### Learning Objectives

- Represent the standard deviation of a normal distribution on a bell curve.
- Use the percentages associated with normal distributions to solve problems.

In the problem presented in the first lesson regarding your bowling average, your teacher told you that the bowling averages were normally distributed. In the previous lesson, you calculated the standard deviation of the averages by using the TI-83 calculator. Later in this lesson, you will be able to represent the value of the standard deviation in a normal distribution.

You have already learned that 68% of the data lies within 1 standard deviation of the mean, 95% of the data lies within 2 standard deviations of the mean, and 99.7% of the data lies within 3 standard deviations of the mean. To accommodate these percentages, there are defined values in each of the regions to the left and to the right of the mean.



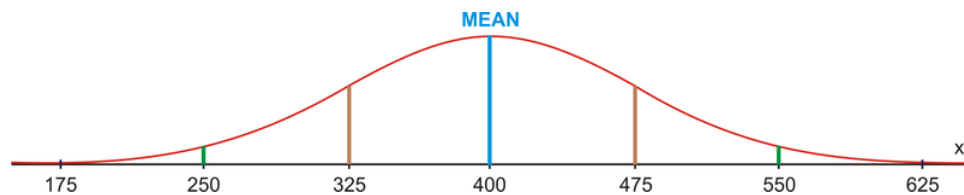
These percentages are used to answer real-world problems when both the mean and the standard deviation of a data set are known.

### Example 7

The lifetimes of a certain type of light bulb are normally distributed. The mean life is 400 hours, and the standard deviation is 75 hours. For a group of 5,000 light bulbs, how many are expected to last each of the following times?

- between 325 hours and 475 hours
- more than 250 hours
- less than 250 hours

### Solution:



- 68% of the light bulbs are expected to last between 325 hours and 475 hours. This means that  $5,000 \times 0.68 = 3,400$  light bulbs are expected to last between 325 and 475 hours.
- $95\% + 2.35\% = 97.35\%$  of the light bulbs are expected to last more than 250 hours. This means that  $5000 \times 0.9735 = 4867.5 \approx 4868$  of the light bulbs are expected to last more than 250 hours.

### 6.3. Connecting the Standard Deviation and Normal Distribution

c) Only 2.35% of the light bulbs are expected to last less than 250 hours. This means that  $5000 \times 0.0235 = 117.5 \approx 118$  of the light bulbs are expected to last less than 250 hours.

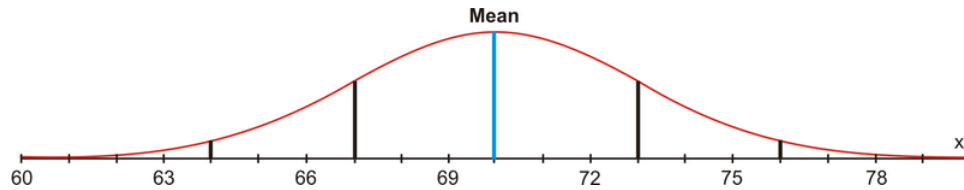
### Example 8

A bag of chips has a mean mass of 70 g, with a standard deviation of 3 g. Assuming a normal distribution, create a normal curve, including all necessary values.

a) If 1,250 bags of chips are processed each day, how many bags will have a mass between 67 g and 73 g?

b) What percentage of the bags of chips will have a mass greater than 64 g?

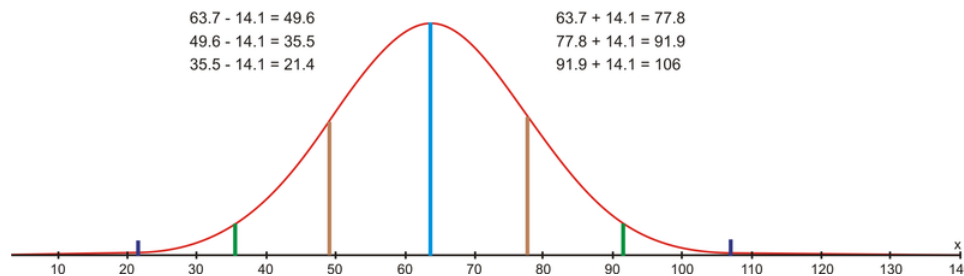
**Solution:**



a) Between 67 g and 73 g lies 68% of the data. If 1,250 bags of chips are processed, 850 bags will have a mass between 67 g and 73 g.

b) 97.35% of the bags of chips will have a mass greater than 64 grams.

Now you can represent the data that your teacher gave to you for the bowling averages of the players in your league on a normal distribution curve. The mean bowling score was 63.7, and the standard deviation was 14.1.



From the normal distribution curve, you can see that your average bowling score of 70 is within 1 standard deviation of the mean. You can also see that 68% of all the data is within 1 standard deviation of the mean, so you did very well bowling this semester. You should definitely return to the league next semester.

### Lesson Summary

In this lesson, you have learned what is meant by a set of data being normally distributed and the significance of standard deviation. You are now able to represent data on a bell-curve and to interpret a given normal distribution curve. In addition, you can calculate the standard deviation of a given data set both manually and by using technology. All of this knowledge can be applied to real-world problems, which you are now able to answer.

### Points to Consider

- Is the normal distribution curve the only way to represent data?
- The normal distribution curve shows the spread of the data, but it does not show the actual data values. Do other representations of data show the actual data values?

### Vocabulary

**Normal distribution** A symmetric bell-shaped curve with tails that extend infinitely in both directions from the mean of a data set.

**Standard deviation** A measure of spread of a data set equal to the square root of the sum of the squared variances divided by the number of data values.

**Variance** A measure of spread of a data set equal to the mean of the squared variations of each data value from the mean of the data set.

**68-95-99.7 Rule** The rule that includes the percentages of data that are within 1, 2, and 3 standard deviations of the mean of a set of data.

## 6.4 Review Questions

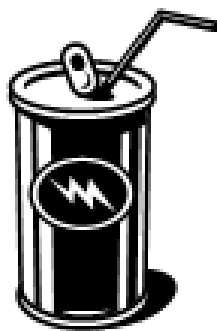


**Part A.** For each question, circle the most appropriate answer.

- If the standard deviation of a population is 6, the population variance is:
  - 2.44
  - 3
  - 6
  - 36
- What is the sample standard deviation of the following data values? 3.3    2.9    8.5    11.5
  - 3.61
  - 4.17
  - 6.55
  - 13.52
- Suppose data are normally distributed, with a mean of 100 and a standard deviation of 20. Between what 2 values will approximately 68% of the data fall?
  - 60 and 140
  - 80 and 120
  - 20 and 100
  - 100 and 125
- The sum of all of the deviations about the mean of a set of data is always going to be equal to:
  - positive
  - the mode
  - the standard deviation total
  - 0
- What is the population variance of the following data values? 40    38    42    47    35
  - 4.03
  - 4.51
  - 15.24
  - 20.34
- Suppose data are normally distributed, with a mean of 50 and a standard deviation of 10. Between what 2 values will approximately 95% of the data fall?
  - 40 and 60

- b. 30 and 70
  - c. 20 and 80
  - d. 10 and 95
7. Suppose data are normally distributed, with a mean of 50 and a standard deviation of 10. What would be the variance?
- a. 10
  - b. 40
  - c. 50
  - d. 100
8. If data are normally distributed, what percentage of the data should lie within the range of  $\mu \pm 3\sigma$ ?
- a. 34%
  - b. 68%
  - c. 95%
  - d. 99.7%
9. If a normally distributed population has a mean of 75 and a standard deviation of 15, what proportion of the values would be expected to lie between 45 and 105?
- a. 34%
  - b. 68%
  - c. 95%
  - d. 99.7%
10. If a normally distributed population has a mean of 25 and a standard deviation of 5.5, what proportion of the values would be expected to lie between 19.5 and 30.5?
- a. 34%
  - b. 68%
  - c. 95%
  - d. 99.7%

**Part B.** Answer the following questions and show all work (including diagrams) to create a complete answer.



11. In the United States, cola can normally be bought in 8 oz cans. A survey was conducted where 250 cans of cola were taken from a manufacturing warehouse and the volumes were measured. It was found that the mean volume was 7.5 oz, and the standard deviation was 0.1 oz. Draw a normal distribution curve to represent this data and then answer the following questions. (b) 68% of the volumes can be found between \_\_\_\_\_ and \_\_\_\_\_. (c) 95% of the volumes can be found between \_\_\_\_\_ and \_\_\_\_\_. (d) 99.7% of the volumes can be found between \_\_\_\_\_ and \_\_\_\_\_.
12. The mean height of the fourth graders in a local elementary school was found to be 4'8", or 56". The standard deviation was found to be 5". Draw a normal distribution curve to represent this data and then answer the

following questions. (b) 68% of the heights can be found between \_\_\_\_\_ and \_\_\_\_\_. (c) 95% of the heights can be found between \_\_\_\_\_ and \_\_\_\_\_. (d) 99.7% of the heights can be found between \_\_\_\_\_ and \_\_\_\_\_.

13. The following data was collected: 5    8    9    10    4    3    7    5 Fill in the chart below and calculate the standard deviation and the variance.

**TABLE 6.12:**

<b>Data (<math>x</math>)</b>	<b>Mean (<math>\mu</math>)</b>	<b>Mean – Data (<math>\mu - x</math>)</b>	<b>Square of Mean – Data (<math>(\mu - x)^2</math>)</b>
------------------------------	--------------------------------	---	---

$\Sigma$

---

14. The following data was collected. 11    15    16    12    19    17    14    18    15    10 Fill in the chart below and calculate the standard deviation and the variance.

**TABLE 6.13:**

<b>Data (<math>x</math>)</b>	<b>Mean (<math>\mu</math>)</b>	<b>Mean – Data (<math>\mu - x</math>)</b>	<b>Square of Mean – Data (<math>(\mu - x)^2</math>)</b>
------------------------------	--------------------------------	---	---

$\Sigma$

---

15. Mrs. Meery has recorded her exam results for the current mathematics exam. The results are shown below:





64 98 78 76 56 48 89 78 69 90 89  
97 67 58 59 50 78 89 68 83 72 91

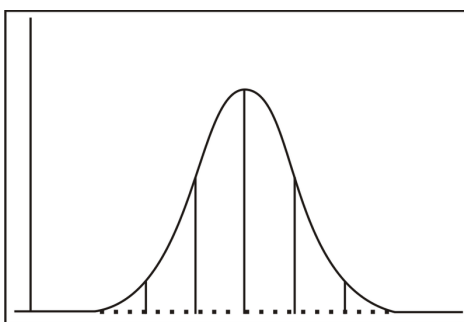
(b) Determine the mean for this data. (c) Determine the standard deviation for this data. (d) Determine the variance for this data. (e) Draw a normal distribution curve to represent the data Mrs. Meery found in her class.

16. Mrs. Landry decided to do the same analysis as Mrs. Meery for her math class. She has recorded her exam results for the current mathematics exam. The results are shown below:

89 87 81 84 76 72 67 49 55 38 67 90 59  
87 89 69 92 90 79 84 69 93 85 70 87 80

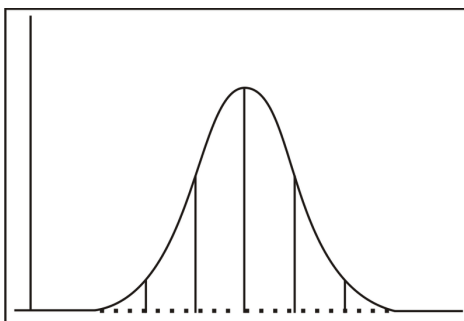
(b) Determine the mean for this data. (c) Determine the standard deviation for this data. (d) Determine the variance for this data. (e) Draw a normal distribution curve to represent the data Mrs. Landry found in her class.

17. 200 senior high students were asked how long they had to wait in the cafeteria line for lunch. Their responses were found to be normally distributed, with a mean of 15 minutes and a standard deviation of 3.5 minutes. Copy the following bell curve onto your paper and answer the questions below.



(b) How many students would you expect to wait more than 11.5 minutes? (c) How many students would you expect to wait more than 18.5 minutes? (d) How many students would you expect to wait between 11.5 and 18.5 minutes?

18. 350 babies were born at Neo Hospital in the past 6 months. The average weight for the babies was found to be 6.8 lbs, with a standard deviation of 0.5 lbs. Copy the following bell curve onto your paper and answer the questions below.



(b) How many babies would you expect to weigh more than 7.3 lbs? (c) How many babies would you expect to weigh more than 7.8 lbs? (d) How many babies would you expect to weigh between 6.3 and 7.8 lbs?

19. Sudoku is a very popular logic game of number combinations. It originated in the late 1800's by the French press, *Le Siècle*. The average times (in minutes) it takes those in a senior math class to complete a Sudoku puzzle are found below. Draw a normal distribution curve to represent this data. Determine what time a student must complete a Sudoku puzzle in to be in the top 0.13%.

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

20 15 21 24 7 19 10 17 15 22 31 19 20 21  
 21 9 12 26 24 28 19 16 24 11 17 31 25 13  
 16 18 22 32 9 15 19 27 14 25 32 29

20. Sheldon has planted seedlings in his garden in the back yard. After 30 days, he measures the heights of the seedlings to determine how much they have grown. The differences in heights can be seen in the table below. The heights are measured in inches. Draw a normal distribution curve to represent the data. Determine what the differences in heights of the seedlings are for 68% of the data.



<http://en.wikipedia.org/wiki/Seedling>

10 3 8 4 7 12 8 5 4 9 3 8  
 6 10 7 10 11 8 12 9 10 7 8 11

**CHAPTER 7**

# Organizing and Displaying Distributions of Data

## Chapter Outline

---

- 7.1 LINE GRAPHS AND SCATTER PLOTS**
  - 7.2 CIRCLE GRAPHS, BAR GRAPHS, HISTOGRAMS, AND STEM-AND-LEAF PLOTS**
  - 7.3 BOX-AND-WHISKER PLOTS**
  - 7.4 REVIEW QUESTIONS**
- 

### Introduction

The local arena is trying to attract as many participants as possible to attend the community's "Skate for Scoliosis" event. Participants pay a fee of \$10.00 for registering, and, in addition, the arena will donate \$3.00 for each hour a participant skates, up to a maximum of 6 hours. Create a table of values and draw a graph to represent a participant who skates for the entire 6 hours. How much money can a participant raise for the community if he/she skates for the maximum length of time?

This problem will be revisited later in the chapter.

When data is collected from surveys or experiments, it is often displayed in charts, tables, or graphs in order to produce a visual image that is helpful in interpreting the results. From a graph or table, an observer is able to detect any patterns or trends that may exist. The most common graphs that are used in statistics are line graphs, scatter plots, bar graphs, histograms, frequency polygons, circle graphs, and box-and-whisker plots.

## 7.1 Line Graphs and Scatter Plots

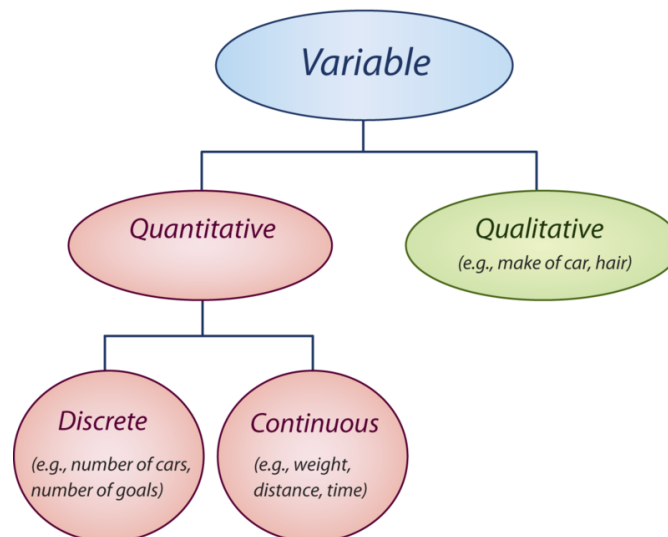
### Learning Objectives

- Represent data that has a linear pattern on a graph.
- Represent data using a broken-line graph.
- Understand the difference between continuous data and discrete data as it applies to a line graph.
- Represent data that has no definite pattern as a scatter plot.
- Draw a line of best fit on a scatter plot.
- Use technology to create both line graphs and scatter plots.

Before you continue to explore the concept of representing data graphically, it is very important to understand the meaning of some basic terms that will often be used in this lesson. The first such definition is that of a **variable**. In statistics, a variable is simply a characteristic that is being studied. This characteristic assumes different values for different elements, or members, of the population, whether it is the entire population or a sample. The value of the variable is referred to as an observation, or a measurement. A collection of these observations of the variable is a **data set**.

Variables can be quantitative or qualitative. A **quantitative variable** is one that can be measured numerically. Some examples of a quantitative variable are wages, prices, weights, numbers of vehicles, and numbers of goals. All of these examples can be expressed numerically. A quantitative variable can be classified as discrete or continuous. A **discrete variable** is one whose values are all countable and does not include any values between 2 consecutive values of a data set. An example of a discrete variable is the number of goals scored by a team during a hockey game. A **continuous variable** is one that can assume any countable value, as well as all the values between 2 consecutive numbers of a data set. An example of a continuous variable is the number of gallons of gasoline used during a trip to the beach.

A **qualitative variable** is one that cannot be measured numerically but can be placed in a category. Some examples of a qualitative variable are months of the year, hair color, color of cars, a person's status, and favorite vacation spots. The following flow chart should help you to better understand the above terms.



### Example 1

Select the best descriptions for the following variables and indicate your selections by marking an 'x' in the appropriate boxes.

**TABLE 7.1:**

Variable	Quantitative	Qualitative	Discrete	Continuous
Number of members in a family				
A person's marital status				
Length of a person's arm				
Color of cars				
Number of errors on a math test				

**Solution:**

**TABLE 7.2:**

Variable	Quantitative	Qualitative	Discrete	Continuous
Number of members in a family	x		x	
A person's marital status		x		
Length of a person's arm	x			x
Color of cars		x		
Number of errors on a math test	x			x

Variables can also be classified as dependent or independent. When there is a linear relationship between 2 variables, the values of one variable depend upon the values of the other variable. In a linear relation, the values of  $y$  depend upon the values of  $x$ . Therefore, the **dependent variable** is represented by the values that are plotted on the  $y$ -axis, and the **independent variable** is represented by the values that are plotted on the  $x$ -axis.

**Example 2**

Sally works at the local ballpark stadium selling lemonade. She is paid \$15.00 each time she works, plus \$0.75 for each glass of lemonade she sells. Create a table of values to represent Sally's earnings if she sells 8 glasses of lemonade. Use this table of values to represent her earnings on a graph.

**Solution:**

The first step is to write an equation to represent her earnings and then to use this equation to create a table of values.

$y = 0.75x + 15$ , where  $y$  represents her earnings and  $x$  represents the number of glasses of lemonade she sells.

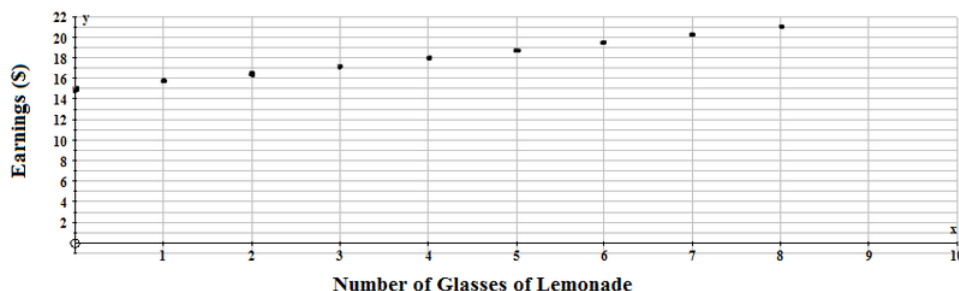


TABLE 7.3:

Number of Glasses of Lemonade	Earnings
0	\$15.00
1	\$15.75
2	\$16.50
3	\$17.25
4	\$18.00
5	\$18.75
6	\$19.50
7	\$20.25
8	\$21.00

The dependent variable is the money earned, and the independent variable is the number of glasses of lemonade sold. Therefore, money is on the  $y$ -axis, and the number of glasses of lemonade is on the  $x$ -axis.

From the table of values, Sally will earn \$21.00 if she sells 8 glasses of lemonade.



Now that the points have been plotted, the decision has to be made as to whether or not to join them. Between every 2 points plotted on the graph are an infinite number of values. If these values are meaningful to the problem, then the plotted points can be joined. This type of data is called **continuous data**. If the values between the 2 plotted points are not meaningful to the problem, then the points should not be joined. This type of data is called **discrete data**. Since glasses of lemonade are represented by whole numbers, and since fractions or decimals are not appropriate values, the points between 2 consecutive values are not meaningful in this problem. Therefore, the points should not be joined. The data is discrete.

Now it is time to revisit the problem presented in the introduction.

The local arena is trying to attract as many participants as possible to attend the community's "Skate for Scoliosis" event. Participants pay a fee of \$10.00 for registering, and, in addition, the arena will donate \$3.00 for each hour a participant skates, up to a maximum of 6 hours. Create a table of values and draw a graph to represent a participant who skates for the entire 6 hours. How much money can a participant raise for the community if he/she skates for the maximum length of time?

**Solution:**

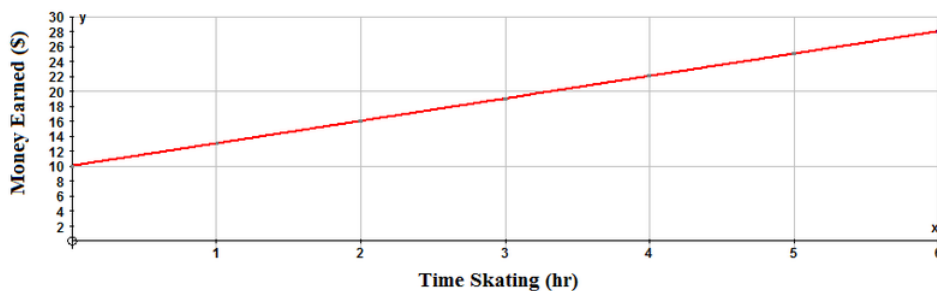
$y = 3x + 10$ , where  $y$  represents the money made by the participant, and  $x$  represents the number of hours the participant skates.



**TABLE 7.4:**

Numbers of Hours Skating	Money Earned
0	\$10.00
1	\$13.00
2	\$16.00
3	\$19.00
4	\$22.00
5	\$25.00
6	\$28.00

The dependent variable is the money made, and the independent variable is the number of hours the participant skated. Therefore, money is on the  $y$ -axis, and time is on the  $x$ -axis as shown below:



A participant who skates for the entire 6 hours can make \$28.00 for the "Skate for Scoliosis" event. The points are joined, because the fractions and decimals between 2 consecutive points are meaningful for this problem. A participant could skate for 30 minutes, and the arena would pay that skater \$1.50 for the time skating. The data is continuous.

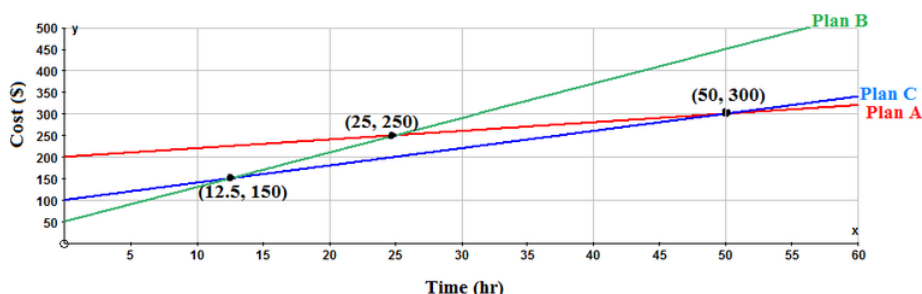
Linear graphs are important in statistics when several data sets are used to represent information about a single topic. An example would be data sets that represent different plans available for cell phone users. These data sets can be plotted on the same grid. The resulting graph will show intersection points for the plans. These intersection points indicate a coordinate where 2 plans are equal. An observer can easily interpret the graph to decide which plan is best, and when. If the observer is trying to choose a plan to use, the choice can be made easier by seeing a graphical representation of the data.

### **Example 3**

#### **7.1. Line Graphs and Scatter Plots**



The following graph represents 3 plans that are available to customers interested in hiring a maintenance company to tend to their lawn. Using the graph, explain when it would be best to use each plan for lawn maintenance.



### **Solution:**

From the graph, the base fee that is charged for each plan is obvious. These values are found on the y-axis. Plan A charges a base fee of \$200.00, Plan C charges a base fee of \$100.00, and Plan B charges a base fee of \$50.00. The cost per hour can be calculated by using the values of the intersection points and the base fee in the equation  $y = mx + b$  and solving for  $m$ . Plan B is the best plan to choose if the lawn maintenance takes less than 12.5 hours. At 12.5 hours, Plan B and Plan C both cost \$150.00 for lawn maintenance. After 12.5 hours, Plan C is the best deal, until 50 hours of lawn maintenance is needed. At 50 hours, Plan A and Plan C both cost \$300.00 for lawn maintenance. For more than 50 hours of lawn maintenance, Plan A is the best plan. All of the above information was obvious from the graph and would enhance the decision-making process for any interested client.

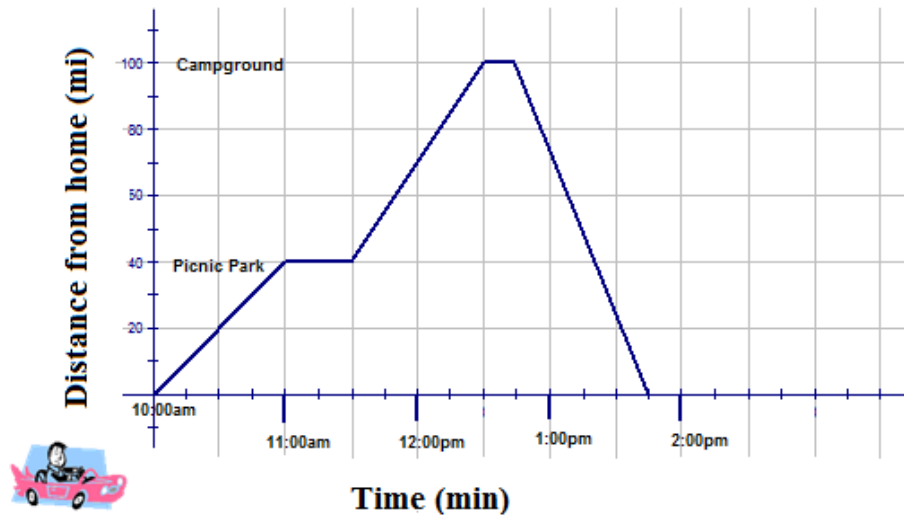
The above graphs represent linear functions, and are called linear (line) graphs. Each of these graphs has a defined slope that remains constant when the line is plotted. A variation of this graph is a **broken-line graph**. This type of line graph is used when it is necessary to show change over time. A line is used to join the values, but the line has no defined slope. However, the points are meaningful, and they all represent an important part of the graph. Usually a broken-line graph is given to you, and you must interpret the given information from the graph.

### **Example 4**



The following graph is an example of a broken-line graph, and it represents the time of a round-trip journey, driving from home to a popular campground and back.





- How far is it from home to the picnic park?
- How far is it from the picnic park to the campground?
- At what 2 places did the car stop?
- How long was the car stopped at the campground?
- When does the car arrive at the picnic park?
- How long did it take for the return trip?
- What was the speed of the car from home to the picnic park?
- What was the speed of the car from the campground to home?

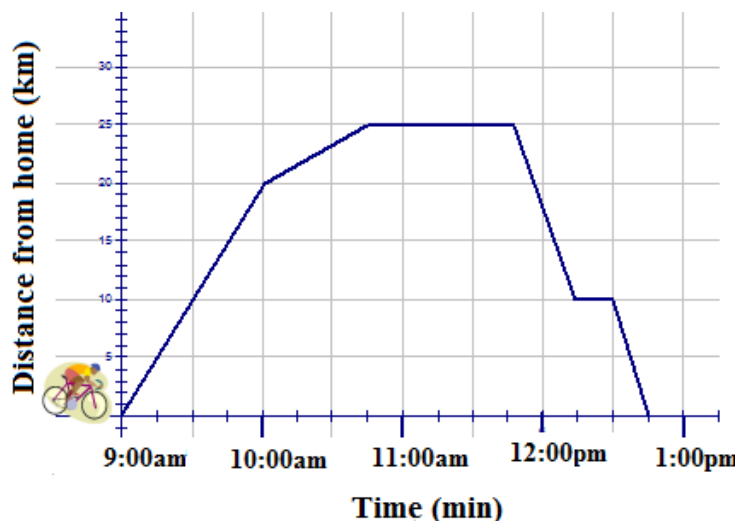
**Solution:**

- It is 40 miles from home to the picnic park.
- It is 60 miles from the picnic park to the campground.
- The car stopped at the picnic park and at the campground.
- The car was stopped at the campground for 15 minutes.
- The car arrived at the picnic park at 11:00 am.
- The return trip took 3 hours and 45 minutes.
- The speed of the car from home to the picnic park was 40 mi/h.
- The speed of the car from the campground to home was 100 mi/h.

**Example 5**

Sam decides to spend some time with his friend Aaron. He hops on his bike and starts off to Aaron's house, but on his way, he gets a flat tire and must walk the remaining distance. Once he arrives at Aaron's house, they repair the flat tire, play some poker, and then Sam returns home. On his way home, Sam decides to stop at the mall to buy a book on how to play poker. The following graph represents Sam's adventure:

7.1. Line Graphs and Scatter Plots



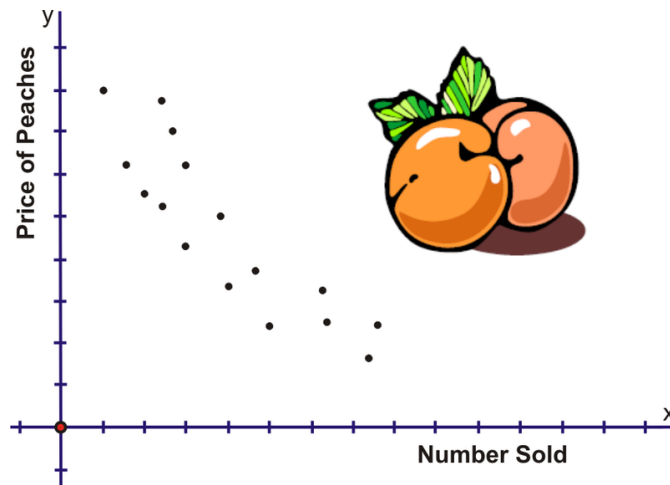
- How far is it from Sam's house to Aaron's house?
- How far is it from Aaron's house to the mall?
- At what time did Sam have a flat tire?
- How long did Sam stay at Aaron's house?
- At what speed did Sam travel from Aaron's house to the mall and then from the mall to home?

**Solution:**

- It is 25 km from Sam's house to Aaron's house.
- It is 15 km from Aaron's house to the mall.
- Sam had a flat tire at 10:00 am.
- Sam stayed at Aaron's house for 1 hour.
- Sam traveled at a speed of 30 km/h from Aaron's house to the mall and then at a speed of 40 km/h from the mall to home.

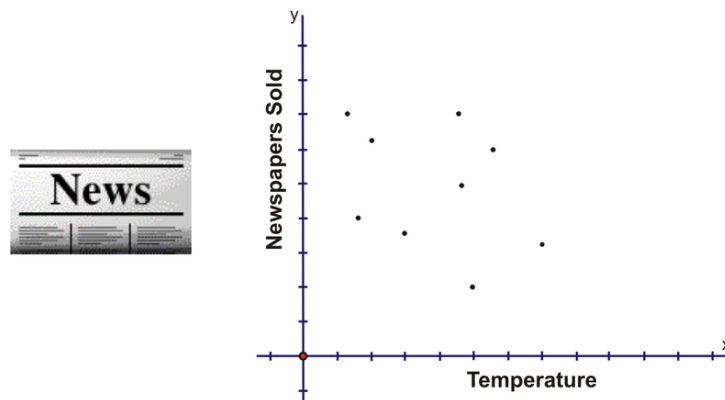
Often, when real-world data is plotted, the result is a linear pattern. The general direction of the data can be seen, but the data points do not all fall on a line. This type of graph is called a scatter plot. A **scatter plot** is often used to investigate whether or not there is a relationship or connection between 2 sets of data. The data is plotted on a graph such that one quantity is plotted on the  $x$ -axis and one quantity is plotted on the  $y$ -axis. The quantity that is plotted on the  $x$ -axis is the independent variable, and the quantity that is plotted on the  $y$ -axis is the dependent variable. If a relationship does exist between the 2 sets of data, it will be easy to see if the data is plotted on a scatter plot.

The following scatter plot shows the price of peaches and the number sold:



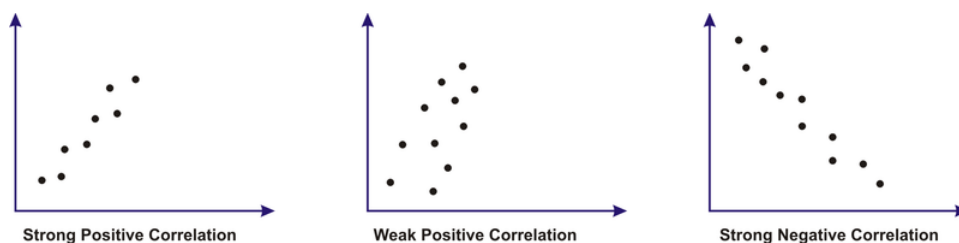
The connection is obvious—when the price of peaches was high, the sales were low, but when the price was low, the sales were high.

The following scatter plot shows the sales of a weekly newspaper and the temperature:



There is no connection between the number of newspapers sold and the temperature.

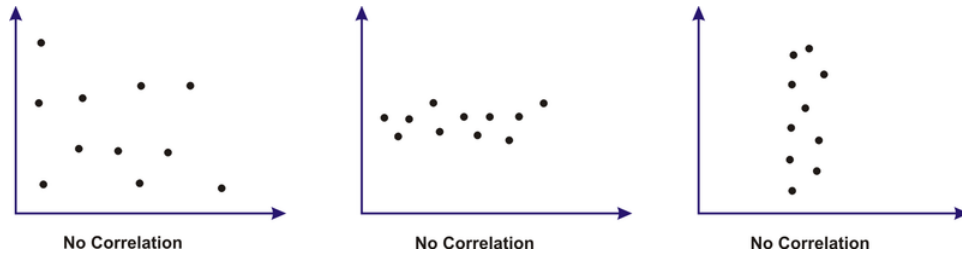
Another term used to describe 2 sets of data that have a connection or a relationship is **correlation**. The correlation between 2 sets of data can be positive or negative, and it can be strong or weak. The following scatter plots will help to enhance this concept.



If you look at the 2 sketches that represent a positive correlation, you will notice that the points are around a line that slopes upward to the right. When the correlation is negative, the line slopes downward to the right. The 2 sketches that show a strong correlation have points that are bunched together and appear to be close to a line that is in the middle of the points. When the correlation is weak, the points are more scattered and not as concentrated.

In the sales of newspapers and the temperature, there was no connection between the 2 data sets. The following sketches represent some other possible outcomes when there is no correlation between data sets:

### 7.1. Line Graphs and Scatter Plots

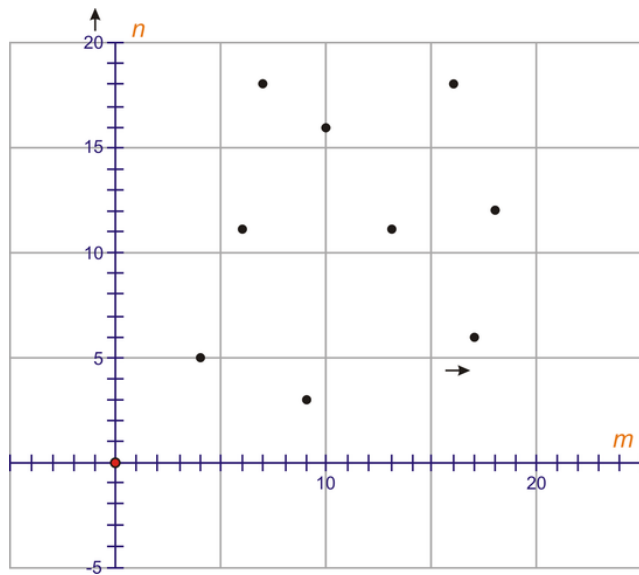


**Example 6**

Plot the following points on a scatter plot, with  $m$  as the independent variable and  $n$  as the dependent variable. Number both axes from 0 to 20. If a correlation exists between the values of  $m$  and  $n$ , describe the correlation (strong negative, weak positive, etc.).

$m$	4	9	13	16	17	6	7	18	10
$n$	5	3	11	18	6	11	18	12	16

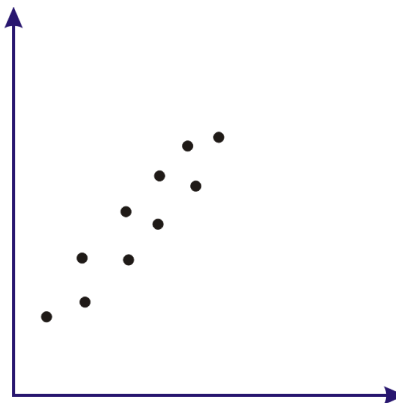
**Solution:**



**This is no correlation.**

**Example 7**

Describe the correlation, if any, in the following scatter plot:



**Solution:**

In the above scatter plot, there is a strong positive correlation.

You now know that a scatter plot can have either a positive or a negative correlation. When this exists on a scatter plot, a line of best fit can be drawn on the graph. The **line of best fit** must be drawn so that the sums of the distances to the points on either side of the line are approximately equal and such that there are an equal number of points above and below the line. Using a clear plastic ruler makes it easier to meet all of these conditions when drawing the line. Another useful tool is a stick of spaghetti, since it can be easily rolled and moved on the graph until you are satisfied with its location. The edge of the spaghetti can be traced to produce the line of best fit. A line of best fit can be used to make estimations from the graph, but you must remember that the line of best fit is simply a sketch of where the line should appear on the graph. As a result, any values that you choose from this line are not very accurate—the values are more of a ballpark figure.

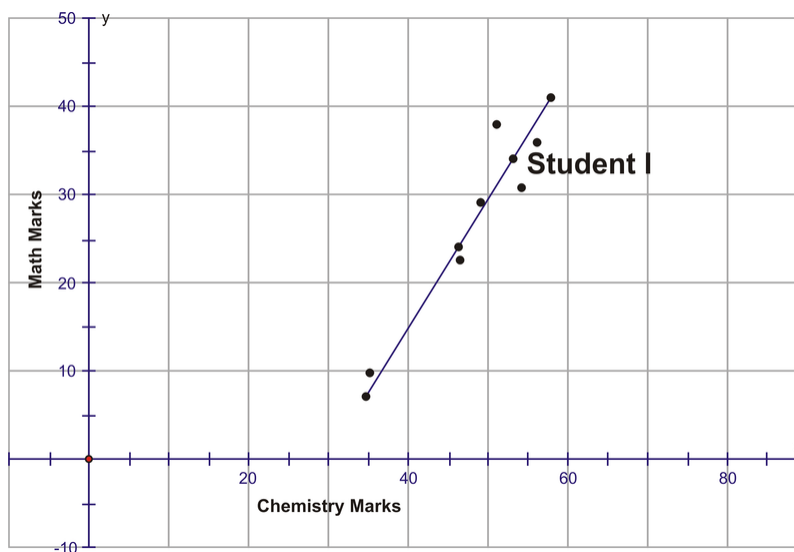
**Example 8**

The following table consists of the marks achieved by 9 students on chemistry and math tests:

**TABLE 7.5:**

Student	A	B	C	D	E	F	G	H	I
Chemistry Marks	49	46	35	58	51	56	54	46	53
Math Marks	29	23	10	41	38	36	31	24	?

Plot the above marks on scatter plot, with the chemistry marks on the  $x$ -axis and the math marks on the  $y$ -axis. Draw a line of best fit, and use this line to estimate the mark that Student I would have made in math had he or she taken the test.

**Solution:**

If Student I had taken the math test, his or her mark would have been between 32 and 37.

Scatter plots and lines of best fit can also be drawn by using technology. The TI-83 is capable of graphing both a scatter plot and of inserting the line of best fit onto the scatter plot.

**Example 9**

Using the data from Example 8, create a scatter plot and draw a line of best fit with the TI-83.

**TABLE 7.6:**

Student	A	B	C	D	E	F	G	H	I
Chemistry Marks	49	46	35	58	51	56	54	46	53
Math Marks	29	23	10	41	38	36	31	24	?

**Solution:**

[STAT] → 2nd [CALC] TESTS → [ENTER]

```

1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
    
```

L1	L2	L3	1
49	29		
46	23		
35	10		
58	41		
51	38		
56	36		
54	31		

L1(1)=

[2nd][Y=]

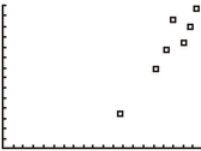
```

1:Plot1...On
  L1 L2
2:Plot2...Off
  L1 L2
3:Plot3...Off
  L1 L2
4:PlotsOff
    
```

[ENTER] Scroll to the correct functions and press enter after each selection.

```

Plot1 Plot2 Plot3
Off Off Off
Type: [ ] [ ] [ ]
Xlist:L1
Ylist:L2
Mark: [ ] +
    
```



The [TRACE] function will give the coordinates of the points.

The calculator can now be used to determine a linear regression equation for the given values. The equation can be entered into the calculator, and the line will be plotted on the scatter plot.

```

EDIT [2nd][TESTS]
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
    
```

[ENTER][2nd][1][,][2nd][2][ENTER]

```

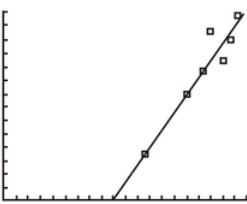
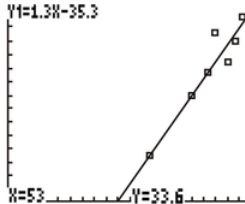
LinReg
y=ax+b
a=1.301512605
b=-35.26218487
    
```

[Y=]

```

1:1.3X-35.3
V1=
V2=
V3=
V4=
V5=
V6=
V7=
    
```

[GRAPH]

From the line of best fit, the calculated value for Student I’s math test mark was 33.6. Remember that the mark that you estimated was between 32 and 37.

**Lesson Summary**

In this lesson, you learned how to represent data by graphing a straight line of the form  $y = mx + b$ , and also by using a scatter plot and a line of best fit. Interpreting a broken-line graph was also presented in this lesson. You learned about correlation as it applies to a scatter plot and how to describe the correlation of a scatter plot. You also learned how to draw a line of best fit on a scatter plot and to use this line to make estimates from the graph. The final topic

that was demonstrated in the lesson was how to use the TI-83 calculator to produce a scatter plot and how to graph a line of best fit by using linear regression.

**Points to Consider**

- Can any of these graphs be used for comparing data?
- Can the equation for the line of best fit be used to calculate values?
- Is any other graphical representation of data used for estimations?

## 7.2 Circle Graphs, Bar Graphs, Histograms, and Stem-and-Leaf Plots

### Learning Objectives

- Construct a stem-and leaf plot.
- Understand the importance of a stem-and-leaf plot in statistics.
- Construct and interpret a circle graph.
- Construct and interpret a bar graph.
- Create a frequency distribution chart.
- Construct and interpret a histogram.
- Use technology to create graphical representations of data.

What is the puppet doing? She can't be cutting a pizza, because the pieces are all different colors and sizes. It seems like she is drawing some type of a display to show different amounts of a whole circle. The colors must represent different parts of the whole. As you proceed through this lesson, refer back to this picture so that you will be able to create a meaningful and detailed answer to the question, "What is the puppet doing?"



### Circle Graphs

Circle graphs, or pie charts, are used extensively in statistics. These graphs appear often in newspapers and magazines. A **pie chart** shows the relationship of the parts to the whole by visually comparing the sizes of the sections (slices). Pie charts can be constructed by using a hundreds disk or by using a circle. The hundreds disk is built on the concept that the whole of anything is 100%, while the circle is built on the concept that  $360^\circ$  is the whole of anything. Both methods of creating a pie chart are acceptable, and both will produce the same result. The sections have different colors to enable an observer to clearly see the differences in the sizes of the sections. The following example will first be done by using a hundreds disk and then by using a circle.

#### Example 10

The Red Cross Blood Donor Clinic had a very successful morning collecting blood donations. Within 3 hours, people had made donations, and the following is a table showing the blood types of the donations:

**TABLE 7.7:**

Blood Type	A	B	O	AB
Number of donors	7	5	9	4

Construct a pie graph to represent the data.



**Solution:**

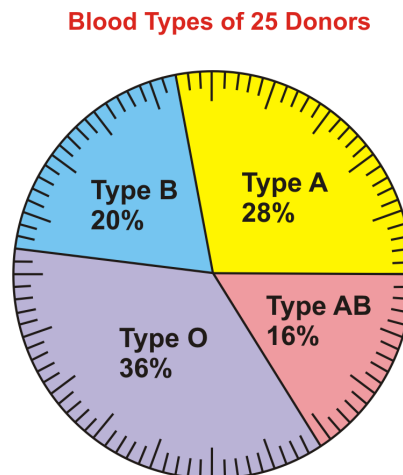
**Step 1:** Determine the total number of donors:  $7 + 5 + 9 + 4 = 25$ .

**Step 2:** Express each donor number as a percent of the whole by using the formula  $\text{Percent} = \frac{f}{n} \cdot 100\%$ , where  $f$  is the frequency and  $n$  is the total number.

$$\frac{7}{25} \cdot 100\% = 28\% \quad \frac{5}{25} \cdot 100\% = 20\% \quad \frac{9}{25} \cdot 100\% = 36\% \quad \frac{4}{25} \cdot 100\% = 16\%$$

**Step 3:** Use a hundreds disk and simply count the correct number for each blood type (1 line = 1 percent).

**Step 4:** Graph each section. Write the name and correct percentage inside the section. Color each section a different color.



The above pie chart was created by using a hundreds disk, which is a circle with 100 divisions in groups of 5. Each division (line) represents 1 percent. From the graph, you can see that more donations were of Type O than any other type. The fewest number of donations of blood collected was of Type AB. If the percentages had not been entered in each section, these same conclusions could have been made based simply on the size of each section.

**Solution:**

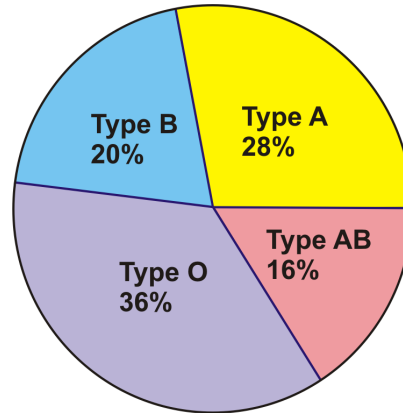
**Step 1:** Determine the total number of donors:  $7 + 5 + 9 + 4 = 25$ .

**Step 2:** Express each donor number as the number of degrees of a circle that it represents by using the formula  $\text{Degrees} = \frac{f}{n} \cdot 360^\circ$ , where  $f$  is the frequency and  $n$  is the total number.

$$\frac{7}{25} \cdot 360^\circ = 100.8^\circ \quad \frac{5}{25} \cdot 360^\circ = 72^\circ \quad \frac{9}{25} \cdot 360^\circ = 129.6^\circ \quad \frac{4}{25} \cdot 360^\circ = 57.6^\circ$$

**Step 3:** Using a protractor, graph each section of the circle.

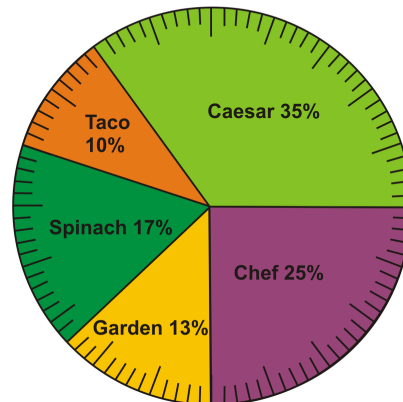
**Step 4:** Write the name and correct percentage inside each section. Color each section a different color.

**Blood Types of 25 Donors**

The above pie chart was created by using a protractor and graphing each section of the circle according to the number of degrees needed. From the graph, you can see that more donations were of Type O than any other type. The fewest number of donations of blood collected was of Type AB. Notice that the percentages have been entered in each section of the graph and not the numbers of degrees. This is because degrees would not be meaningful to an observer trying to interpret the graph. In order to create a pie chart by using a circle, it is necessary to use the formula to calculate the number of degrees for each section, and in order to create a pie chart by using a hundreds disk, it is necessary to use the formula to determine the percentage for each section. In the end, however, both methods result in identical graphs.

**Example 11**

A new restaurant is opening in town, and the owner is trying very hard to complete the menu. He wants to include a choice of 5 salads and has presented his partner with the following circle graph to represent the results of a recent survey that he conducted of the town's people. The survey asked the question, "What is your favorite kind of salad?"

**Salad Choices**

Use the graph to answer the following questions:



- Which salad was the most popular choice?
- Which salad was the least popular choice?
- If 300 people were surveyed, how many people chose each type of salad?
- What is the difference between the number of people who chose the spinach salad and the number of people who chose the garden salad?

**Solution:**

1. The most popular salad was the caesar salad.

2. The least popular salad was the taco salad.

3. Caesar salad:  $35\% = \frac{35}{100} = 0.35$

$$(300)(0.35) = 105 \text{ people}$$

Taco salad:  $10\% = \frac{10}{100} = 0.10$

$$(300)(0.10) = 30 \text{ people}$$

Spinach salad:  $17\% = \frac{17}{100} = 0.17$

$$(300)(0.17) = 51 \text{ people}$$

Garden salad:  $13\% = \frac{13}{100} = 0.13$

$$(300)(0.13) = 39 \text{ people}$$

Chef salad:  $25\% = \frac{25}{100} = 0.25$

$$(300)(0.25) = 75 \text{ people}$$

4. The difference between the number of people who chose the spinach salad and the number of people who chose the garden salad is  $51 - 39 = 12$  people.

If we revisit the puppet who was introduced at the beginning of the lesson, you should now be able to create a story that details what she is doing. An example would be that she is in charge of the student body and is presenting to the students the results of a questionnaire regarding student activities for the first semester. Of the 5 activities, the one that is orange in color is the most popular. The students have decided that they want to have a winter carnival week more than any other activity.

**Stem-and-Leaf Plots**

In statistics, data is represented in tables, charts, and graphs. One disadvantage of representing data in these ways is that the actual data values are often not retained. One way to ensure that the data values are kept intact is to graph the values in a stem-and-leaf plot. A **stem-and-leaf plot** is a method of organizing the data that includes sorting the data and graphing it at the same time. This type of graph uses a stem as the leading part of a data value and a leaf as the remaining part of the value. The result is a graph that displays the sorted data in groups, or classes. A stem-and-leaf plot is used most when the number of data values is large.

**Example 12**

At a local veterinarian school, the number of animals treated each day over a period of 20 days was recorded. Construct a stem-and-leaf plot for the data set, which is as follows:

**7.2. Circle Graphs, Bar Graphs, Histograms, and Stem-and-Leaf Plots**



28 34 23 35 16  
 17 47 05 60 26  
 39 35 47 35 38  
 35 55 47 54 48

**Solution:**

**Step 1:** Some people prefer to arrange the data in order before the stems and leaves are created. This will ensure that the values of the leaves are in order. However, this is not necessary and can take a great deal of time if the data set is large. We will first create the stem-and-leaf plot, and then we will organize the values of the leaves.

Stem	Leaf
0	5
1	6, 7
2	8, 3, 6
3	4, 5, 9, 5, 5, 8, 5
4	7, 7, 7, 8
5	5, 4
6	0

The leading digit of a data value is used as the stem, and the trailing digit is used as the leaf. The numbers in the stem column should be consecutive numbers that begin with the smallest class and continue to the largest class. If there are no values in a class, do not enter a value in the leaf—just leave it blank.

**Step 2:** Organize the values in each leaf row.

Stem	Leaf
0	5
1	6, 7
2	3, 6, 8
3	4, 5, 5, 5, 5, 8, 9
4	7, 7, 7, 8
5	4, 5
6	0

Now that the graph has been constructed, there is a great deal of information that can be learned from it.

The number of values in the leaf column should equal the number of data values that were given in the table. The value that appears the most often in the same leaf row is the trailing digit of the mode of the data set. The mode of this data set is 35. For 7 of the 20 days, the number of animals receiving treatment was between 34 and 39. The veterinarian school treated a minimum of 5 animals and a maximum of 60 animals on any one day. The median of the data can be quickly calculated by using the values in the leaf column to locate the value in the middle position. In this stem and leaf plot, the median is the mean of the sum of the numbers represented by the 10<sup>th</sup> and the 11<sup>th</sup> leaves:  $\frac{35+35}{2} = \frac{70}{2} = 35$ .

### Example 13

The following numbers represent the growth (in centimeters) of some plants after 25 days.

Construct a stem-and-leaf plot to represent the data, and list 3 facts that you know about the growth of the plants.

18 10 37 36 61  
 39 41 49 50 52  
 57 53 51 57 39  
 48 56 33 36 19  
 30 41 51 38 60

**Solution:**

Stem	Leaf	Stem	Leaf
1	8, 0, 9	1	0, 8, 9
2		2	
3	7, 6, 9, 9, 3, 6, 0, 8	3	0, 3, 6, 6, 7, 8, 9, 9
4	1, 9, 8, 1	4	1, 1, 8, 9
5	0, 2, 7, 3, 1, 7, 6, 1	5	0, 1, 1, 2, 3, 6, 7, 7
6	1, 0	6	0, 1

Answers will vary, but the following are some possible responses:

- From the stem-and-leaf plot, the growth of the plants ranged from a minimum of 10 cm to a maximum of 61 cm.
- The median of the data set is the value in the 13<sup>th</sup> position, which is 41 cm.
- There was no growth recorded in the class of 20 cm, so there is no number in the leaf row.
- The data set is multimodal.

## Bar Graphs

The different types of graphs that you have seen so far are plots to use with quantitative variables. A qualitative variable can be plotted using a bar graph. A **bar graph** is a plot made of bars whose heights (vertical bars) or lengths (horizontal bars) represent the frequencies of each category. There is 1 bar for each category, with space between each bar, and the data that is plotted is discrete data. Each category is represented by intervals of the same width. When constructing a bar graph, the category is usually placed on the horizontal axis, and the frequency is usually placed on the vertical axis. These values can be reversed if the bar graph has horizontal bars.

### Example 14

Construct a bar graph to represent the depth of the Great Lakes:

Lake Superior – 1,333 ft.

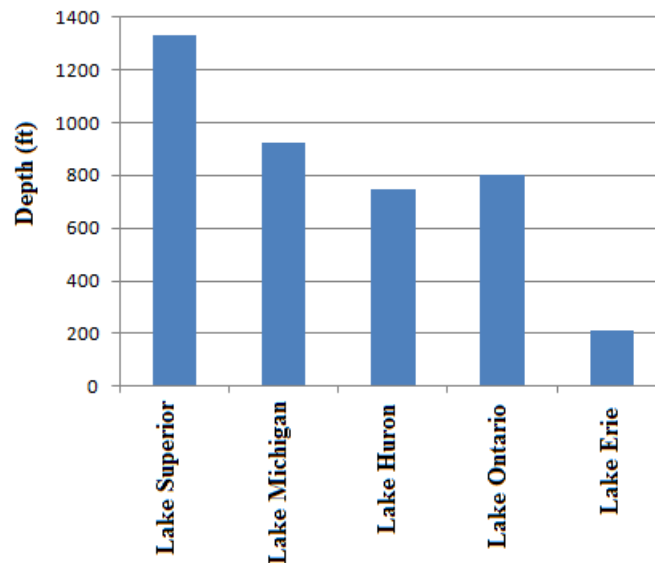
Lake Michigan – 923 ft.

Lake Huron – 750 ft.

Lake Ontario – 802 ft.

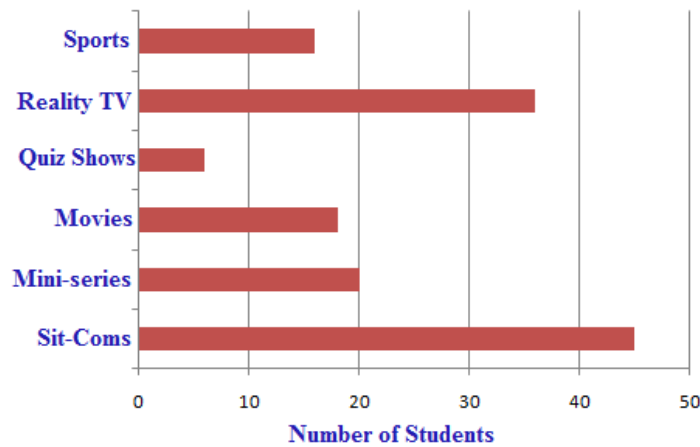
Lake Erie – 210 ft.

**Solution:**



### Example 15

The following bar graph represents the results of a survey to determine the type of TV shows watched by high school students:



Use the bar graph to answer the following questions:

- What type of show is watched the most?
- What type of show is watched the least?
- Approximately how many students participated in the survey?
- Does the graph show the differences between the preferences of males and females?

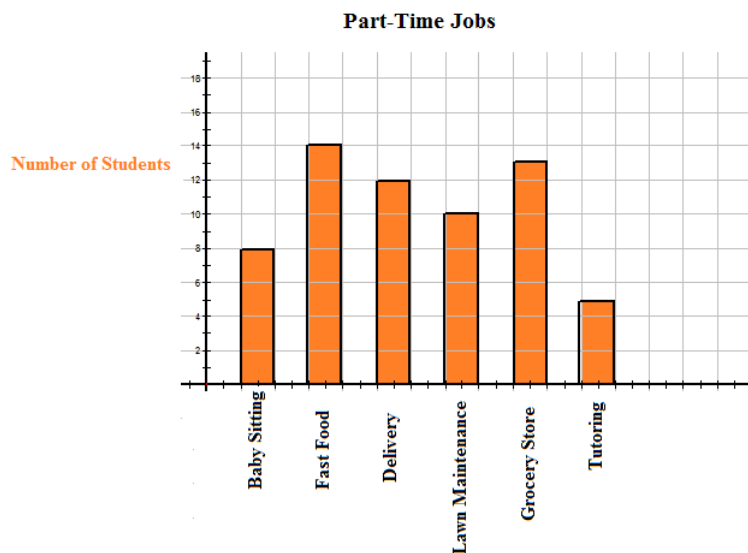
**Solution:**

- Sit-coms are watched the most.
- Quiz shows are watched the least.
- Approximately  $45 + 20 + 18 + 6 + 35 + 16 = 140$  students participated in the survey.
- No, the graph does not show the differences between the preferences of males and females.

If bar graphs are constructed on grid paper, it is very easy to keep the intervals the same size and to keep the bars evenly spaced. In addition to helping in the appearance of the graph, grid paper also enables you to more accurately determine the frequency of each class.

### Example 16

The following bar graph represents the part-time jobs held by a group of grade 10 students:



Using the above bar graph, answer the following questions:

- What was the most popular part-time job?
- What was the part-time job held by the least number of students?
- Which part-time jobs employed 10 or more of the students?
- Is it possible to create a table of values for the bar graph? If so, construct the table of values.
- What percentage of the students worked as a delivery person?

**Solution:**

- The most popular part-time job was in the fast food industry.
- The part-time job of tutoring was the one held by the least number of students.
- The part-time jobs that employed 10 or more students were in the fast food, delivery, lawn maintenance, and grocery store businesses.
- Yes, it's possible to create a table of values for the bar graph.

**TABLE 7.8:**

Part-Time Job	Baby Sitting	Fast Food	Delivery	Lawn Care	Grocery Store	Tutoring
Number of Students	8	14	12	10	13	5

- The percentage of the students who worked as a delivery person was approximately 19.4%.

$$\begin{array}{r}
 8+14+22+18 \\
 \phantom{8+14+22+18} \quad 62 \\
 12/62 \\
 \text{Ans} * 100 \\
 \phantom{Ans} \quad 19.35483871
 \end{array}$$

## Histograms

An extension of the bar graph is the histogram. A **histogram** is a type of vertical bar graph in which the bars represent grouped continuous data. While there are similarities between a bar graph and a histogram, such as each bar being the same width, a histogram has no spaces between the bars. The quantitative data is grouped according to a determined bin size, or interval. The bin size refers to the width of each bar, and the data is placed in the appropriate bin.

The **bins**, or groups of data, are plotted on the  $x$ -axis, and the frequencies of the bins are plotted on the  $y$ -axis. A grouped **frequency distribution** is constructed for the numerical data, and this table is used to create the histogram. In most cases, the grouped frequency distribution is designed so there are no breaks in the intervals. The last value of one bin is actually the first value counted in the next bin. This means that if you had groups of data with a bin size of 10, the bins would be represented by the notation [0-10), [10-20), [20-30), etc. Each bin appears to contain 11 values, which is 1 more than the desired bin size of 10. Therefore, the last digit of each bin is counted as the first digit of the following bin.

The first bin includes the values 0 through 9, and the next bin includes the values 9 through 19. This makes the bins the proper size. Bin sizes are written in this manner to simplify the process of grouping the data. The first bin can begin with the smallest number of the data set and end with the value determined by adding the bin width to this value, or the bin can begin with a reasonable value that is smaller than the smallest data value.

### Example 17



Construct a frequency distribution table with a bin size of 10 for the following data, which represents the ages of thirty lottery winners:

38 41 29 33 40 74 66 45 60 55  
 25 52 54 61 46 51 59 57 66 62  
 32 47 65 50 39 22 35 72 77 49

**Solution:**

**Step 1:** Determine the range of the data by subtracting the smallest value from the largest value.

$$\text{Range: } 77 - 22 = 55$$

**Step 2:** Divide the range by the bin size to ensure that you have at least 5 groups of data. A histogram should have from 5 to 10 bins to make it meaningful:  $\frac{55}{10} = 5.5 \approx 6$ . Since you cannot have 0.5 of a bin, the result indicates that you will have at least 6 bins.

**Step 3:** Construct the table.

**TABLE 7.9:**

<b>Bin</b>	<b>Frequency</b>
[20 – 30)	3
[30 – 40)	5
[40 – 50)	6
[50 – 60)	8
[60 – 70)	5
[70 – 80)	3

**Step 4:** Determine the sum of the frequency column to ensure that all the data has been grouped.

$$3 + 5 + 6 + 8 + 5 + 3 = 30$$

When data is grouped in a frequency distribution table, the actual data values are lost. The table indicates how many values are in each group, but it doesn't show the actual values.

There are many different ways to create a distribution table and many different distribution tables that can be created. However, for the purpose of constructing a histogram, the method shown works very well, and it is not difficult to complete. When the number of data values is very large, another column is often inserted in the distribution table. This column is a tally column, and it is used to account for the number of values within a bin. A tally column facilitates the creation of the distribution table and usually allows the task to be completed more quickly.

### **Example 18**

The numbers of years of service for 75 teachers in a small town are listed below:

1, 6, 11, 26, 21, 18, 2, 5, 27, 33, 7, 15, 22, 30, 8  
 31, 5, 25, 20, 19, 4, 9, 19, 34, 3, 16, 23, 31, 10, 4  
 2, 31, 26, 19, 3, 12, 14, 28, 32, 1, 17, 24, 34, 16, 1,  
 18, 29, 10, 12, 30, 13, 7, 8, 27, 3, 11, 26, 33, 29, 20  
 7, 21, 11, 19, 35, 16, 5, 2, 19, 24, 13, 14, 28, 10, 31

Using the above data, construct a frequency distribution table with a bin size of 5.

**Solution:**

$$\text{Range: } 35 - 1 = 34$$

$$\frac{34}{5} = 6.8 \approx 7$$

You will have 7 bins.

For each value that is in a bin, draw a stroke in the Tally column. To make counting the strokes easier, draw 4 strokes and cross them out with the fifth stroke. This process bundles the strokes in groups of 5, and the frequency can be readily determined.



**TABLE 7.10:**

Bin	Tally	Frequency
[0 – 5)	/	11
[5 – 10)		9
[10 – 15)	/	12
[15 – 20)	/	14
[20 – 25)		7
[25 – 30)	/	10
[30 – 35)	/	12

$$11 + 9 + 12 + 14 + 7 + 10 + 12 = 75$$

Now that you have constructed the frequency table, the grouped data can be used to draw a histogram. Like a bar graph, a histogram requires a title and properly labeled  $x$ - and  $y$ -axes.

### Example 19

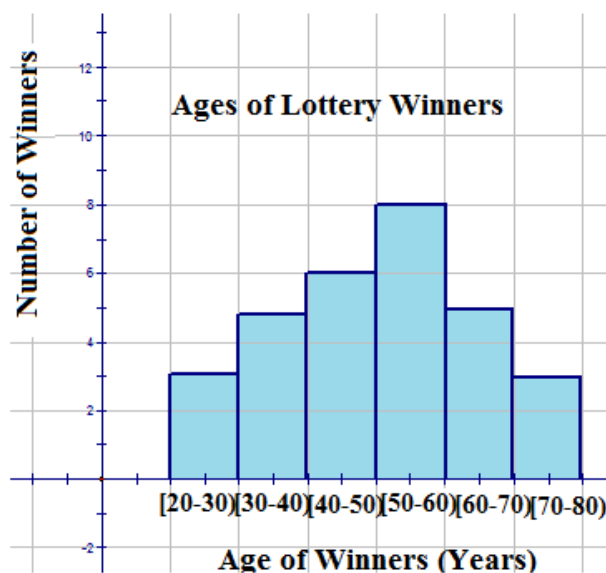
Use the data from Example 17 that displays the ages of the lottery winners to construct a histogram. The data is shown again below:

**TABLE 7.11:**

Bin	Frequency
[20 – 30)	3
[30 – 40)	5
[40 – 50)	6
[50 – 60)	8
[60 – 70)	5
[70 – 80)	3

### Solution:

Use the data as it is represented in the distribution table to construct the histogram.



From looking at the tops of the bars, you can see how many winners were in each category, and by adding these numbers, you can determine the total number of winners. You can also determine how many winners were within a specific category. For example, you can see that 8 winners were 60 years of age or older. The graph can also be used to determine percentages. For example, it can answer the question, “What percentage of the winners were 50 years of age or older?” as follows:

$$\frac{16}{30} = 0.53\overline{3} \quad (0.53\overline{3})(100\%) \approx 5.3\%.$$

### Example 20

a) Use the data and the distribution table that represent the ages of teachers from Example 18 to construct a histogram to display the data. The distribution table is shown again below:

TABLE 7.12:

Bin	Tally	Frequency
[0 – 5)		11
[5 – 10)		9
[10 – 15)		12
[15 – 20)		14
[20 – 25)		7
[25 – 30)		10
[30 – 35)		12

b) Now use the histogram to answer the following questions.

i) How many teachers teach in this small town?

ii) How many teachers have worked for less than 5 years?

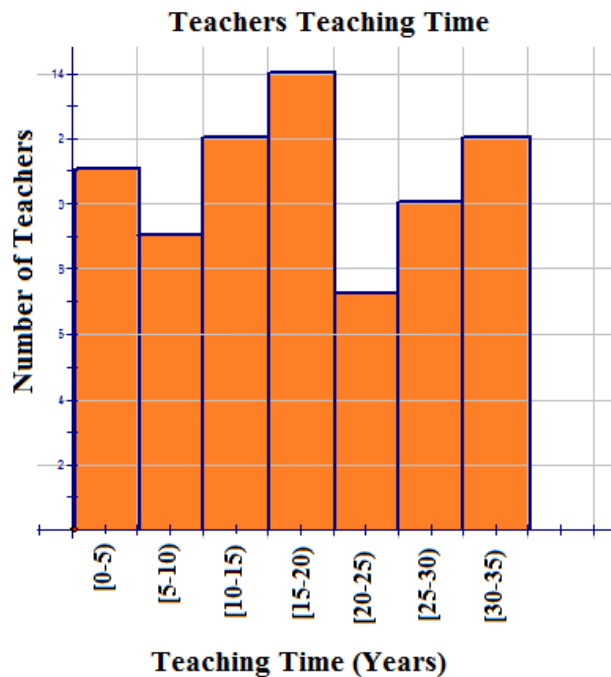
iii) If teachers are able to retire when they have taught for 30 years or more, how many are eligible to retire?

iv) What percentage of the teachers still have to teach for 10 years or fewer before they are eligible to retire?

v) Do you think that the majority of the teachers are young or old? Justify your answer.

**Solution:**

a)



b) i)  $11 + 9 + 12 + 14 + 7 + 10 + 12 = 75$

In this small town, 75 teachers are teaching.

ii) 11 teachers have taught for less than 5 years.

iii) 12 teachers are eligible to retire.

iv)  $\frac{17}{75} = 0.22\overline{66}$        $(0.2266)(100\%) \approx 2.3\%$

Approximately 2.3% of the teachers must teach for 10 years or fewer before they are eligible to retire.

v) The majority of the teachers are young, because 46 have taught for less than 20 years.

Technology can also be used to plot a histogram. The TI-83 can be used to create a histogram by using STAT and STAT PLOT on the calculator.

### Example 21

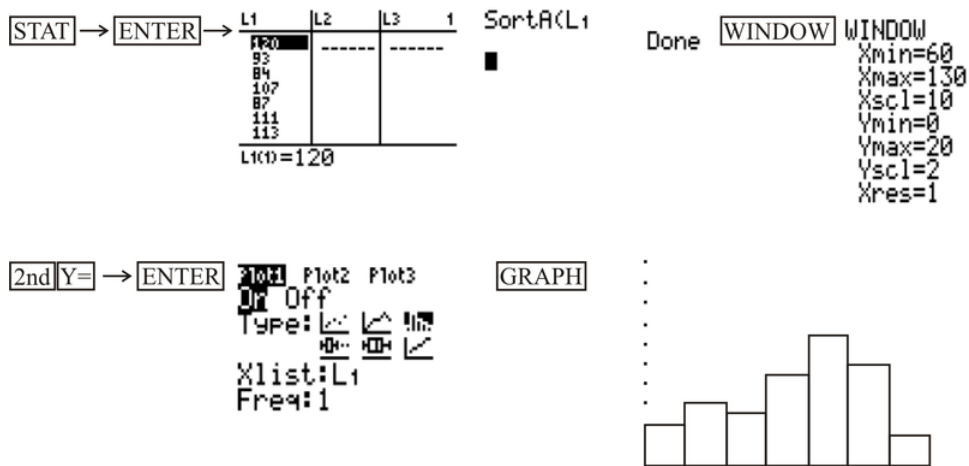
Scientists have invented a new dietary supplement that is supposed to increase the weight of a piglet within its first 3 months of growth. Farmer John fed this supplement to his stock of piglets, and at the end of 3 months, he recorded the weights of 50 randomly selected piglets.

The following table is the recorded weights (in pounds) of the 50 selected piglets:

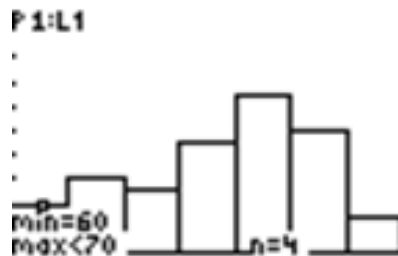
120	111	65	110	114	72	116	105	119	114
93	113	99	118	108	97	107	95	113	75
84	120	102	104	84	97	121	69	100	101
107	118	77	105	109	78	89	68	74	103
87	67	79	90	109	94	106	96	92	88

Using the above data set and the TI-83, construct a histogram to represent the data.

**Solution:**

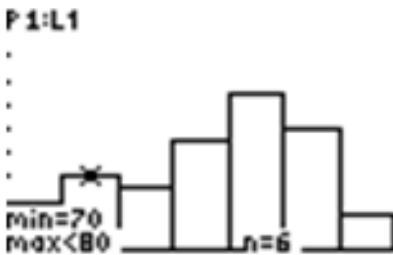


Using the TRACE feature will give you information about the data in each bar of the histogram.



The TRACE feature tells you that in the first bin, which is [60-70), there are 4 values.

## 7.2. Circle Graphs, Bar Graphs, Histograms, and Stem-and-Leaf Plots

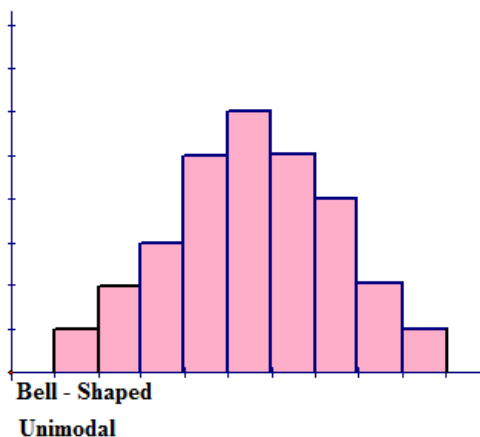


The TRACE feature tells you that in the second bin, which is [70-80), there are 6 values.

To advance to the next bin, or bar, of the histogram, use the cursor and move to the right. The information obtained by using the TRACE feature will enable you to create a frequency table and to draw the histogram on paper.

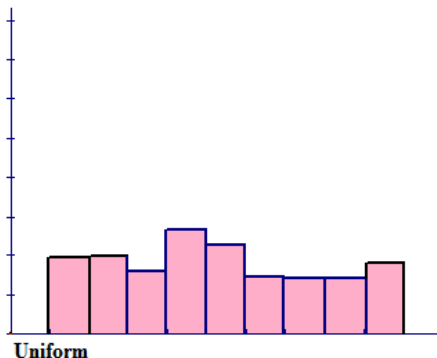
The shape of a histogram can tell you a lot about the distribution of the data, as well as provide you with information about the mean, median, and mode of the data set. The following are some typical histograms, with a caption below each one explaining the distribution of the data, as well as the characteristics of the mean, median, and mode. Distributions can have other shapes besides the ones shown below, but these represent the most common ones that you will see when analyzing data.

a)



For a **symmetric histogram**, the values of the mean, median, and mode are all the same and are all located at the center of the distribution.

b)

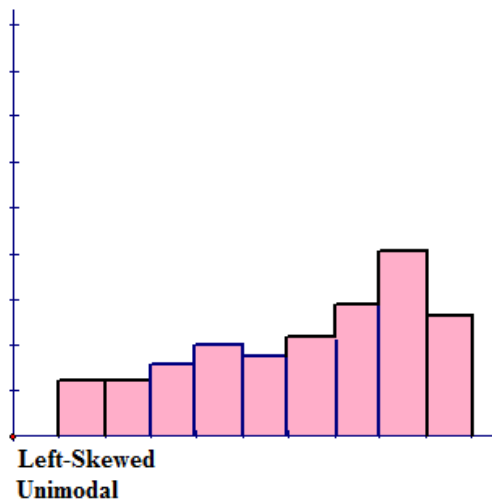


c)



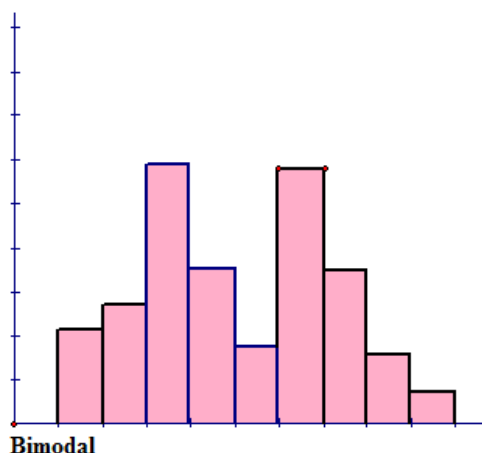
For a histogram that is skewed to the right, the mean is located to the right on the distribution and is the largest value of the measures of central tendency. The mean has the largest value because it is strongly affected by the outliers on the right tail that pull the mean to the right. The mode is the smallest value, and it is located to the left on the distribution. The mode always occurs at the highest point of the peak. The median is located between the mode and the mean.

d)



For a histogram that is skewed to the left, the mean is located to the left on the distribution and is the smallest value of the measures of central tendency. The mean has the smallest value because it is strongly affected by the outliers on the left tail that pull the mean to the left. The median is located between the mode and the mean.

e)



In each of the above graphs, the distributions are not perfectly shaped, but are shaped enough to identify an overall pattern. Figure a represents a bell-shaped distribution, which has a single peak and tapers off to both the left and to the right of the peak. The shape appears to be symmetric about the center of the histogram. The single peak indicates that the distribution is unimodal. The highest peak of the histogram represents the location of the mode of the data set. The mode is the data value that occurs the most often in a data set.

Figure b represents a distribution that is approximately uniform and forms a rectangular, flat shape. The frequency of each class is approximately the same.

Figure c represents a **right-skewed distribution**, which has a peak to the left of the distribution and data values that taper off to the right. This distribution has a single peak and is also unimodal.

Figure d represents a **left-skewed distribution**, which has a peak to the right of the distribution and data values that taper off to the left. This distribution has a single peak and is also unimodal.

Figure e has no shape that can be defined. The only defining characteristic about this distribution is that it has 2 peaks of the same height. This means that the distribution is bimodal.

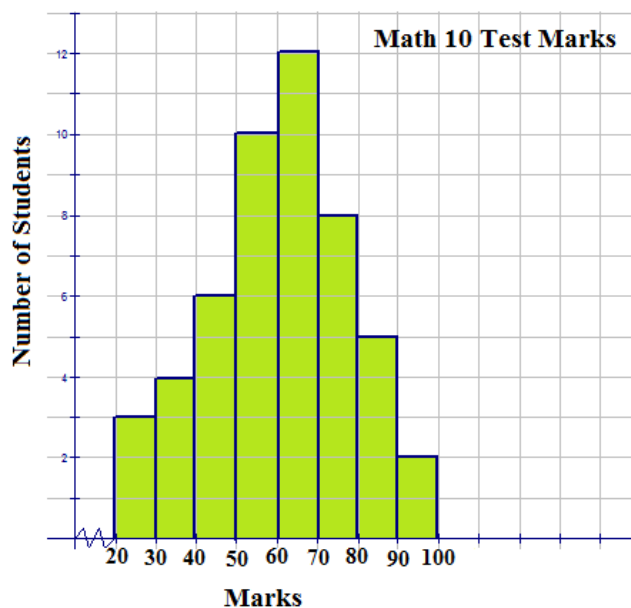
Another type of graph that can be drawn to represent the same set of data as a histogram represents is a frequency polygon. A **frequency polygon** is a graph constructed by using lines to join the midpoints of each interval, or bin. The heights of the points represent the frequencies. A frequency polygon can be created from the histogram or by calculating the midpoints of the bins from the frequency distribution table. The **midpoint** of a bin is calculated by adding the upper and lower boundary values of the bin and dividing the sum by 2.

### **Example 22**

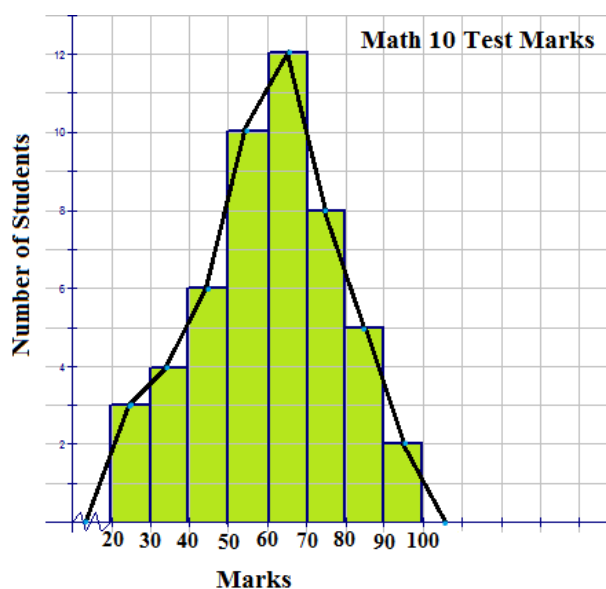
The following histogram represents the marks made by 40 students on a math 10 test.

Use the histogram to construct a frequency polygon to represent the data.





*Solution:*



There is no data value greater than 0 and less than 20. The jagged line that is inserted on the  $x$ -axis is used to represent this fact. The area under the frequency polygon is the same as the area under the histogram and is, therefore, equal to the frequency values that would be displayed in a distribution table. The frequency polygon also shows the shape of the distribution of the data, and in this case, it resembles a bell curve.

### Example 23

The following distribution table represents the number of miles run by 20 randomly selected runners during a recent road race:

**TABLE 7.13:**

Bin	Frequency
[5.5 – 10.5)	1
[10.5 – 15.5)	3

TABLE 7.13: (continued)

Bin	Frequency
[15.5 – 20.5)	2
[20.5 – 25.5)	4
[25.5 – 30.5)	5
[30.5 – 35.5)	3
[35.5 – 40.5)	2



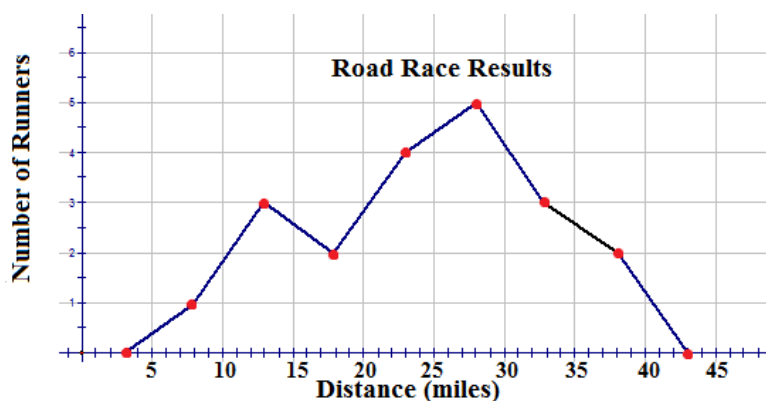
Using this table, construct a frequency polygon.

**Solution:**

**Step 1:** Calculate the midpoint of each bin by adding the 2 numbers of the interval and dividing the sum by 2.

$$\begin{array}{l} \text{Midpoints: } \frac{5.5 + 10.5}{2} = \frac{16}{2} = 8 \\ \frac{10.5 + 15.5}{2} = \frac{26}{2} = 13 \\ \frac{15.5 + 20.5}{2} = \frac{36}{2} = 18 \\ \frac{20.5 + 25.5}{2} = \frac{46}{2} = 23 \\ \frac{25.5 + 30.5}{2} = \frac{56}{2} = 28 \\ \frac{30.5 + 35.5}{2} = \frac{66}{2} = 33 \\ \frac{35.5 + 40.5}{2} = \frac{76}{2} = 38 \end{array}$$

**Step 2:** Plot the midpoints on a grid, making sure to number the  $x$ -axis with a scale that will include the bin sizes. Join the plotted midpoints with lines.



A frequency polygon usually extends 1 unit below the smallest bin value and 1 unit beyond the greatest bin value. This extension gives the frequency polygon an appearance of having a starting point and an ending point, which provides a view of the distribution of data. If the data set were very large so that the number of bins had to be increased and the bin size decreased, the frequency polygon would appear as a smooth curve.

**Lesson Summary**

In this lesson, you learned how to represent data that was presented in various forms. Data that could be represented as percentages was displayed in a circle graph, or pie chart. Discrete data that was qualitative was displayed on a bar graph. Finally, continuous data that was grouped was graphed on a histogram or on a frequency polygon. You also learned to detect characteristics of a distribution by simply observing the shape of a histogram. Once again, technology was shown to be an asset when constructing a histogram.

**Points to Consider**

- Can any of these graphs be used for comparing data?
- Can these graphs be used to display solutions to problems in everyday life?
- How do these graphs compare to ones presented in previous lessons?

## 7.3 Box-and-Whisker Plots

### Learning Objectives

- Construct a box-and-whisker plot.
- Construct and interpret a box-and-whisker plot.
- Use technology to create box-and-whisker plots.

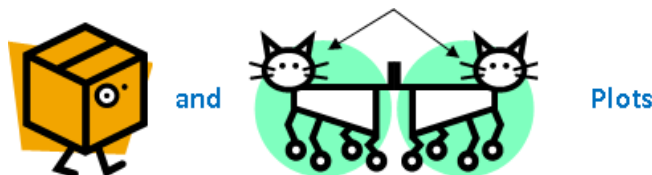
Bob, a financial planner, recorded the number of clients he saw each day over an 11-day period. The numbers of clients he saw are shown below:

31, 33, 29, 40, 51, 27, 30, 43, 38, 23, 42

After Bob reviewed the data, he was satisfied with the results and thought that he had met with a sufficient number of clients. Display the set of data in order to explain whether the claim made by Bob is true or false. Use the display to justify your answer.

We will revisit this problem later in the lesson to explain whether or not Bob's claim was true or false.

### Box-and Whisker Plots

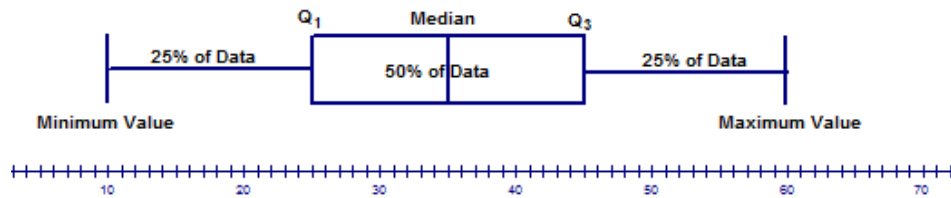


In traditional statistics, data is organized by using a frequency distribution. The results of the frequency distribution can then be used to create various graphs, such as a histogram or a frequency polygon, which indicate the shape or nature of the distribution. The shape of the distribution will allow you to confirm various conjectures about the nature of the data.

To examine data in order to identify patterns, trends, or relationships, exploratory data analysis is used. In exploratory data analysis, organized data is displayed in order to make decisions or suggestions regarding further actions. A **box-and-whisker plot** (often called a box plot) can be used to graphically represent the data set, and the graph involves plotting 5 specific values. The 5 specific values are often referred to as a **five-number summary** of the organized data set. The five-number summary consists of the following:

- a. The lowest number in the data set (minimum value)
- b. The median of the lower quartile:  $Q_1$  (median of the first half of the data set)
- c. The median of the entire data set (median)
- d. The median of the upper quartile:  $Q_3$  (median of the second half of the data set)
- e. The highest number in the data set (maximum value)

The display of the five-number summary produces a box-and-whisker plot as shown below:



The above model of a box-and-whisker plot shows 2 horizontal lines (the whiskers) that each contain 25% of the data and are of the same length. In addition, it shows that the median of the data set is in the middle of the box, which contains 50% of the data. The lengths of the whiskers and the location of the median with respect to the center of the box are used to describe the distribution of the data. It's important to note that this is just an example. Not all box-and-whisker plots have the median in the middle of the box and whiskers of the same size.

Information about the data set that can be determined from the box-and-whisker plot with respect to the location of the median includes the following:

- If the median is located in the center or near the center of the box, the distribution is approximately symmetric.
- If the median is located to the left of the center of the box, the distribution is positively skewed.
- If the median is located to the right of the center of the box, the distribution is negatively skewed.

Information about the data set that can be determined from the box-and-whisker plot with respect to the length of the whiskers includes the following:

- If the whiskers are the same or almost the same length, the distribution is approximately symmetric.
- If the right whisker is longer than the left whisker, the distribution is positively skewed.
- If the left whisker is longer than the right whisker, the distribution is negatively skewed.

The length of the whiskers also gives you information about how spread out the data is.

A box-and-whisker plot is often used when the number of data values is large. The center of the distribution, the nature of the distribution, and the range of the data are very obvious from the graph. The five-number summary divides the data into quarters by use of the medians of the upper and lower halves of the data. Remember that, unlike the mean, the median of the entire data set is not affected by outliers, so it is the measure of central tendency that is most often used in exploratory data analysis.

### **Example 24**

For the following data sets, determine the five-number summaries:

- 12, 16, 36, 10, 31, 23, 58
- 144, 240, 153, 629, 540, 300

### **Solution:**

- The first step is to organize the values in the data set as shown below:

12, 16, 36, 10, 31, 23, 58  
10, 12, 16, 23, 31, 36, 58

(10), 12, 16, 23, 31, 36, (58)

Now complete the following list:

### 7.3. Box-and-Whisker Plots

Minimum value  $\rightarrow$  10

$Q_1 \rightarrow$  12

Median  $\rightarrow$  23

$Q_3 \rightarrow$  36

Maximum value  $\rightarrow$  58

b) The first step is to organize the values in the data set as shown below:

144, 240, 153, 629, 540, 300

144, 153, 240, 300, 540, 629

144, 153, 240, 300, 540, 629

$$\frac{240 + 300}{2} = \frac{540}{2} = 270$$

144, 153, 240 | 300, 540, 629

Now complete the following list:

Minimum value  $\rightarrow$  144

$Q_1 \rightarrow$  153

Median  $\rightarrow$  270

$Q_3 \rightarrow$  540

Maximum value  $\rightarrow$  629

### **Example 25**

Use the data set for Example 1 part a) and the five-number summary to construct a box-and-whisker plot to model the data set.

### **Solution:**

The five-number summary can now be used to construct a box-and-whisker plot. Be sure to provide a scale on the number line that includes the range from the minimum value to the maximum value.

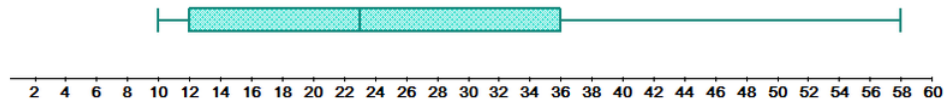
a) Minimum value  $\rightarrow$  10

$Q_1 \rightarrow$  12

Median  $\rightarrow$  23

$Q_3 \rightarrow$  36

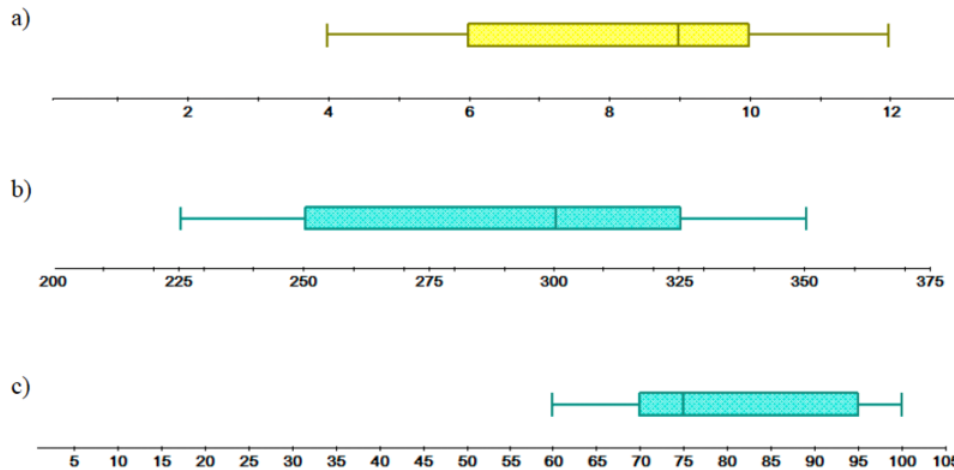
Maximum value  $\rightarrow$  58



It is very visible that the right whisker is much longer than the left whisker. This indicates that the distribution is positively skewed.

### Example 26

For each box-and-whisker plot, list the five-number summary and describe the distribution based on the location of the median.



### Solution:

a) Minimum value  $\rightarrow$  4

$Q_1 \rightarrow$  6

Median  $\rightarrow$  9

$Q_3 \rightarrow$  10

Maximum value  $\rightarrow$  12

The median of the data set is located to the right of the center of the box, which indicates that the distribution is negatively skewed.

b) Minimum value  $\rightarrow$  225

$Q_1 \rightarrow$  250

Median  $\rightarrow$  300

$Q_3 \rightarrow$  325

Maximum value  $\rightarrow$  350

The median of the data set is located to the right of the center of the box, which indicates that the distribution is negatively skewed.

c) Minimum value  $\rightarrow$  60

$Q_1 \rightarrow$  70

Median  $\rightarrow$  75

### 7.3. Box-and-Whisker Plots

$$Q_3 \rightarrow 95$$

Maximum value  $\rightarrow 100$

The median of the data set is located to the left of the center of the box, which indicates that the distribution is positively skewed.

### Example 27

The numbers of square feet (in 100s) of 10 of the largest museums in the world are shown below:

650, 547, 204, 213, 343, 288, 222, 250, 287, 269

Construct a box-and-whisker plot for the above data set and describe the distribution.

### Solution:

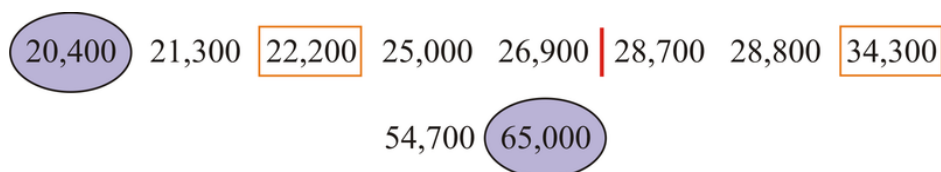
The first step is to organize the data values as follows:

20,400 21,300 22,200 25,000 26,900 28,700 28,800 34,300 54,700 65,000

Now calculate the median,  $Q_1$ , and  $Q_3$ .

20,400 21,300 22,200 25,000 26,900 28,700 28,800 34,300 54,700 65,000

$$\text{Median} \rightarrow \frac{26,900 + 28,700}{2} = \frac{55,600}{2} = 27,800$$



$$Q_1 = 22,200$$

$$Q_3 = 34,300$$

Next, complete the following list:

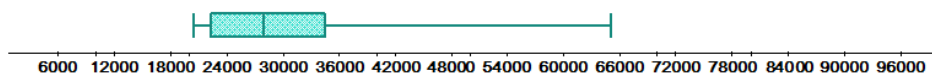
Minimum value  $\rightarrow 20,400$

$$Q_1 \rightarrow 22,200$$

Median  $\rightarrow 27,800$

$$Q_3 \rightarrow 34,300$$

Maximum value  $\rightarrow 65,000$



The right whisker is longer than the left whisker, which indicates that the distribution is positively skewed.



The TI-83 or TI-84 can also be used to create a box-and whisker plot. In the following examples, the TI-83 is used. In the next chapter, key strokes using the TI-84 will be presented to you. The five-number summary values can be determined by using the TRACE feature of the calculator or by using CALC and 1-Var Stats.

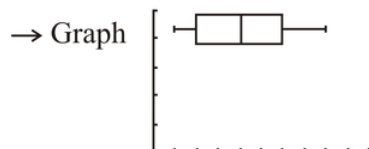
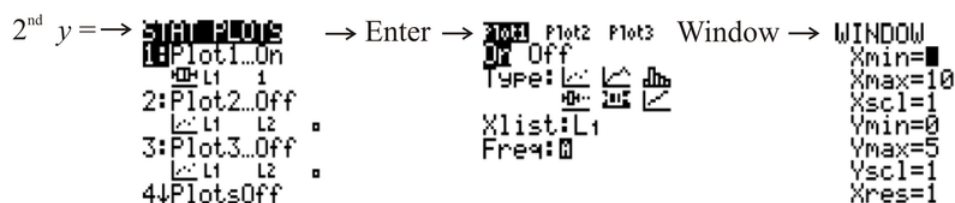
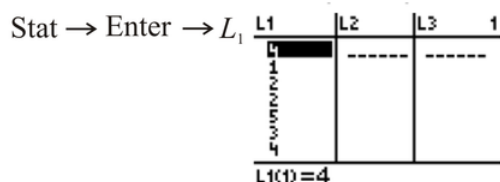
### Example 28

The following numbers represent the number of siblings in each family for 15 randomly selected students:

4, 1, 2, 2, 5, 3, 4, 2, 6, 4, 6, 1, 7, 8, 4

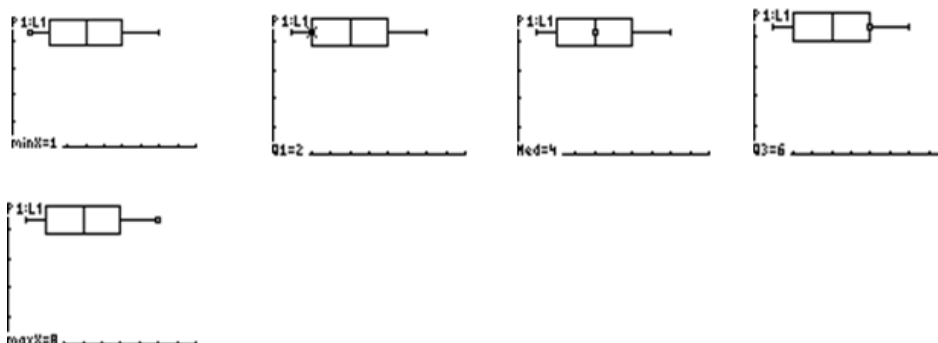
Use technology to construct a box-and-whisker plot to display the data. List the five-number summary values.

**Solution:**



The five-number summary can be obtained from the calculator in 2 ways.

1. The following results are obtained by simply using the TRACE feature.



The values at the bottom of each screen are the five-number summary.

2. The second method involves using CALC and 1-Var Stats for L1.

### 7.3. Box-and-Whisker Plots

```

EDIT  [ ] TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg

1-Var Stats
n=15
minX=1
Q1=2
Med=4
Q3=6
maxX=8

```

Many data sets contain values that are either extremely high values or extremely low values compared to the rest of the data values. These values are called **outliers**. There are several reasons why a data set may contain an outlier. Some of these are listed below:

- The value may be the result of an error made in measurement or in observation. The researcher may have measured the variable incorrectly.
- The value may simply be an error made by the researcher in recording the value. The value may have been written or typed incorrectly.
- The value could be a result obtained from a subject not within the defined population. A researcher recording marks from a math 12 examination may have recorded a mark by a student in grade 11 who was taking math 12.
- The value could be one that is legitimate but is extreme compared to the other values in the data set. (This rarely occurs, but it is a possibility.)

If an outlier is present because of an error in measurement, observation, or recording, then either the error should be corrected, or the outlier should be omitted from the data set. If the outlier is a legitimate value, then the statistician must make a decision as to whether or not to include it in the set of data values. There is no rule that tells you what to do with an outlier in this case.

One method for checking a data set for the presence of an outlier is to follow the procedure below:

- Organize the given data set and determine the values of  $Q_1$  and  $Q_3$ .
- Calculate the difference between  $Q_1$  and  $Q_3$ . This difference is called the **interquartile range (IQR)**:  $IQR = Q_3 - Q_1$ .
- Multiply the difference by 1.5, subtract this result from  $Q_1$ , and add it to  $Q_3$ .
- The results from Step 3 will be the range into which all values of the data set should fit. Any values that are below or above this range are considered outliers.

### Example 29

Using the procedure outlined above, check the following data sets for outliers:

- 18, 20, 24, 21, 5, 23, 19, 22
- 13, 15, 19, 14, 26, 17, 12, 42, 18

### Solution:

- Organize the given data set as follows:

18, 20, 24, 21, 5, 23, 19, 22  
5, 18, 19, 20, 21, 22, 23, 24

Determine the values for  $Q_1$  and  $Q_3$ .

5, 18, 19, 20, 21, 22, 23, 24

$$Q_1 = \frac{18+19}{2} = \frac{37}{2} = 18.5 \quad Q_3 = \frac{22+23}{2} = \frac{45}{2} = 22.5$$

Calculate the difference between  $Q_1$  and  $Q_3$ :  $Q_3 - Q_1 = 22.5 - 18.5 = 4.0$ .

Multiply this difference by 1.5:  $(4.0)(1.5) = 6.0$ .

Finally, compute the range.

$$Q_1 - 6.0 = 18.5 - 6.0 = 12.5$$

$$Q_3 + 6.0 = 22.5 + 6.0 = 28.5$$

Are there any data values below 12.5? Yes, the value of 5 is below 12.5 and is, therefore, an outlier.

Are there any values above 28.5? No, there are no values above 28.5.

b) Organize the given data set as follows:

13, 15, 19, 14, 26, 17, 12, 42, 18

12, 13, 14, 15, 17, 18, 19, 26, 42

Determine the values for  $Q_1$  and  $Q_3$ .

12, 13, 14, 15, 17, 18, 19, 26, 42

$$Q_1 = \frac{13+14}{2} = \frac{27}{2} = 13.5 \quad Q_3 = \frac{19+26}{2} = \frac{45}{2} = 22.5$$

Calculate the difference between  $Q_1$  and  $Q_3$ :  $Q_3 - Q_1 = 22.5 - 13.5 = 9.0$ .

Multiply this difference by 1.5:  $(9.0)(1.5) = 13.5$ .

Finally, compute the range.

$$Q_1 - 13.5 = 13.5 - 13.5 = 0$$

$$Q_3 + 13.5 = 22.5 + 13.5 = 36.0$$

Are there any data values below 0? No, there are no values below 0.

Are there any values above 36.0? Yes, the value of 42 is above 36.0 and is, therefore, an outlier.

### 7.3. Box-and-Whisker Plots

## Lesson Summary

You have learned the significance of the median as it applies to dividing a set of data values into quartiles. You have also learned how to apply these values to the five-number summary needed to construct a box-and-whisker plot. In addition, you have learned how to construct a box-and-whisker plot and how to obtain the five-number summary by using technology. The last topic that you learned about in this lesson was the meaning of the term outlier. Some reasons why an outlier might exist in a data set and the procedure for determining whether or not a data set contains an outlier were also discussed.

## Points to Consider

- Are there still other ways to represent data graphically?
- Are there other uses for a box-and-whisker plot?
- Can box-and-whisker plots be used for comparing data sets?

## Vocabulary

**Bar graph** A plot made of bars whose heights (vertical bars) or lengths (horizontal bars) represent the frequencies of each category.

**Bins** Quantitative or qualitative categories. Bins are also known as classes.

**Box-and-whisker plot** A graph of a data set in which the five-number summary is plotted. 50 percent of the data values are in the box, and the remaining 50 percent are divided equally on the whiskers.

**Broken-line graph** A graph that is used when it is necessary to show change over time. A line is used to join the values, but the line has no defined slope.

**Continuous data** Data for which the plotted points can be joined.

**Continuous variable** A variable that can assume all values between 2 consecutive values of a data set.

**Correlation** A statistical method used to determine whether or not there is a linear relationship between 2 variables.

**Data set** A collection of observations of a variable.

**Dependent variable** The variable represented by the values that are plotted on the y-axis.

**Discrete data** Data for which the plotted points cannot be joined.

**Discrete variable** A variable that can only assume values that can be counted.

**Five-number summary** 5 values for a data set that include the smallest value, the lower quartile, the median, the upper quartile, and the largest value.

**Frequency distribution** A table that lists all of the classes and the number of data values that belong to each of the classes.

**Frequency polygon** A graph that uses lines to join the midpoints of the tops of the bars of a histogram or to join the midpoints of the classes.

**Histogram** A graph in which the classes, or bins, are on the horizontal axis and the frequencies are plotted on the vertical axis. The frequencies are represented by vertical bars that are drawn adjacent to each other.

**Independent variable** The variable represented by the values that are plotted on the  $x$ -axis.

**Interquartile range (IQR)** The difference between the third quartile and the first quartile.

**Left-skewed distribution** A distribution in which most of the data values are located to the right of the mean.

**Line of best fit** A straight line drawn on a scatter plot such that the sums of the distances to points on either side of the line are approximately equal and such that there are an equal number of points above and below the line.

**Midpoint** The value obtained by adding the lower and upper limits of a class and dividing the sum by 2.

**Pie chart** A circle that is divided into sections (slices) according to the percentage of the frequencies in each class.

**Qualitative variable** A variable that can be placed into specific categories according to some defined characteristic.

**Quantitative variable** A variable that is numerical in nature and that can be ordered.

**Right-skewed distribution** A distribution in which most of the data values are located to the left of the mean.

**Scatter plot** A graph used to investigate whether or not there is a relationship between 2 sets of data. The data is plotted on a graph such that one quantity is plotted on the  $x$ -axis and one quantity is plotted on the  $y$ -axis.

**Stem-and-leaf plot** A method of organizing data that includes sorting the data and graphing it at the same time. This type of graph uses the stem as the leading part of the data value and the leaf as the remaining part of the value.

**Symmetric histogram** A histogram for which the values of the mean, median, and mode are all the same and are all located at the center of the distribution.

**Variable** A characteristic that is being studied.

## 7.4 Review Questions

### Line Graphs and Scatter Plots

Show all work necessary to answer each question.

**Section A** – All questions in this section are selected response.

1. What term is used to describe a data set in which all points between 2 consecutive points are meaningful?
  - a. discrete data
  - b. continuous data
  - c. random data
  - d. fractional data
2. What is the correlation of a scatter plot that has few points that are not bunched together?
  - a. strong
  - b. no correlation
  - c. weak
  - d. negative
3. Which of the following calculations will create the line of best fit on the TI-83?
  - a. quadratic regression
  - b. cubic regression
  - c. exponential regression
  - d. linear regression ( $ax + b$ )
4. What type of variable is represented by the number of pets owned by families?
  - a. qualitative
  - b. quantitative
  - c. independent
  - d. continuous
5. What term is used to define the connection between 2 data sets?
  - a. relationship
  - b. scatter plot
  - c. correlation
  - d. discrete
6. What type of data, when plotted on a graph, does not have the points joined?
  - a. discrete data
  - b. continuous data
  - c. random data
  - d. independent data
7. What name is given to a graph that shows change over time, with points that are joined but have no defined slope?
  - a. linear graph
  - b. broken-line graph
  - c. scatter plot
  - d. line of best fit

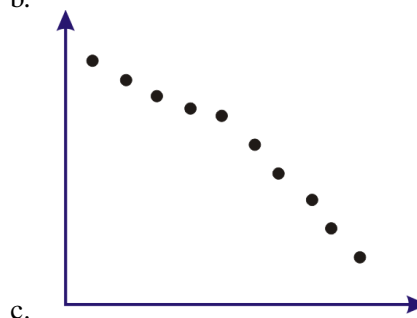
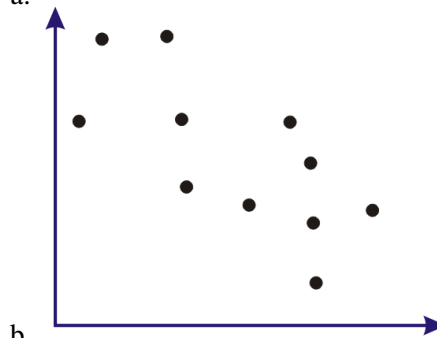
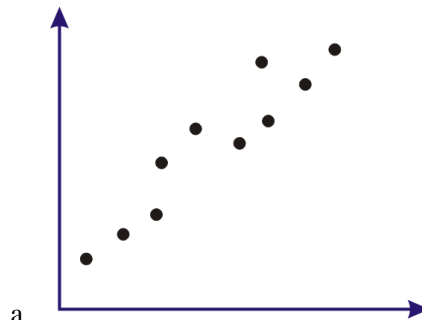
**Section B** – All questions in this section are long answer questions. Be sure to show all of the work necessary to arrive at the correct answer.

- Select the best descriptions for the following variables and indicate your selections by marking an 'x' in the appropriate boxes.

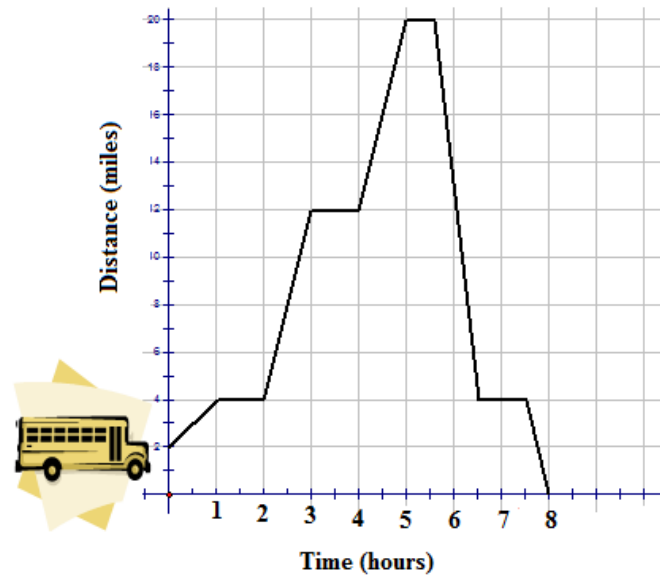
**TABLE 7.14:**

Variable	Quantitative	Qualitative	Discrete	Continuous
Men's favorite TV shows				
Salaries of baseball players				
Number of children in a family				
Favorite color of cars				
Number of hours worked weekly				

- Describe the correlation of each of the following graphs:



- Answer the questions below for the following broken-line graph, which shows the distance, over time, of a bus from the bus depot.



- What was the fastest speed of the bus?
  - How many times did the bus stop on its trip? (Do not count the beginning and the end of the trip.)
  - What was the initial distance of the bus from the bus depot?
  - What was the total distance traveled by the bus?
4. The following table represents the sales of Volkswagen Beetles in Iowa between 1994 and 2003:

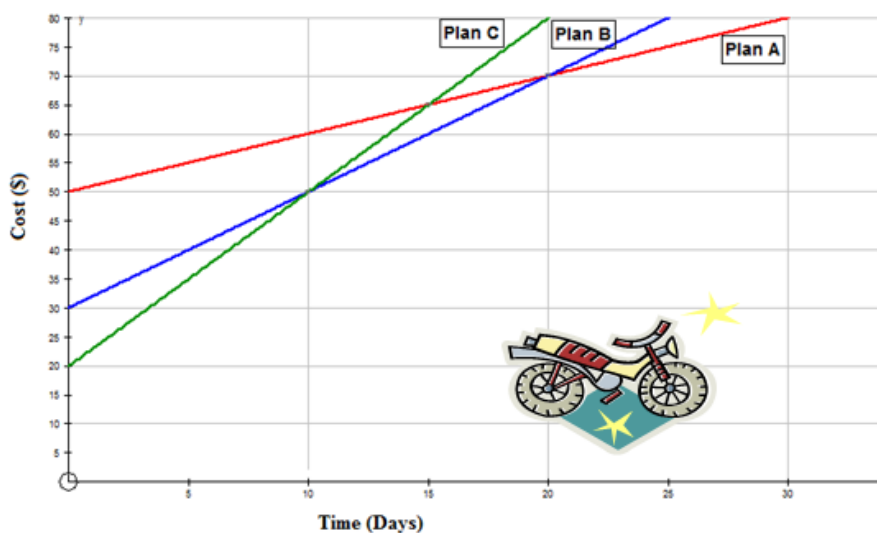
**TABLE 7.15:**

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Beetles Sold	50	60	55	50	70	65	75	65	80	90

- Create a scatter plot and draw the line of best fit for the data. Hint: Let 0 = 1994, 1 = 1995, etc.
- Use the graph to predict the number of Beetles that will be sold in Iowa in the year 2007.
- Describe the correlation for the above graph.

5. You are selling your motorcycle, and you decide to advertise it on the Internet on Walton's Web Ads. He has 3 plans from which you may choose. The plans are shown on the following graph. Use the graph and explain when it is best to use each plan.



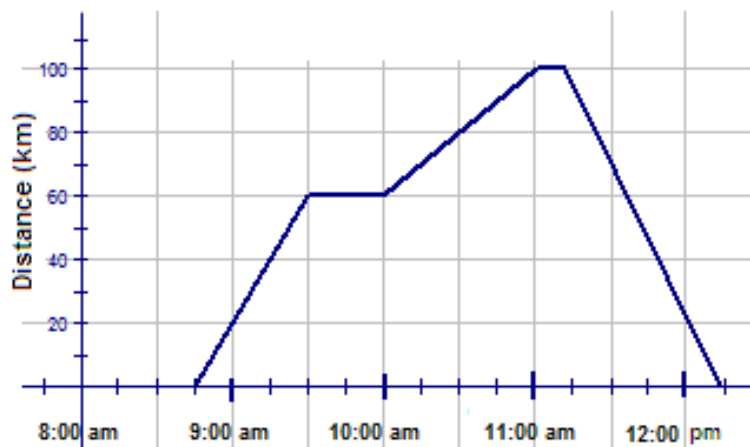


6. The data below gives the gasoline used by cars with the same-sized engines when driven at various speeds.



Speed (m/h)	32	64	77	42	82	57	72
Gasoline Used (m/gal)	40	27	24	37	22	36	28

- Draw a scatter plot and a line of best fit. (You may use technology.)
  - If a car were traveling at a speed of 47 m/h, estimate the amount of gasoline that would be used.
  - If a car uses 29 m/gal of gasoline, estimate the speed of the car.
7. For the following broken-line graph, write a story to accompany the graph, and provide a detailed description of the events that are occurring.



8. Plot the following points on a scatter graph, with  $m$  as the independent variable and  $n$  as the dependent variable. Number both axes from 0 to 20. If a correlation exists between the values of  $m$  and  $n$ , describe the correlation (strong negative, weak positive, etc.).

- a.  $m$  5 14 2 10 16 4 18 2 8 11  
 $n$  6 13 4 10 15 7 16 5 8 12
- b.  $m$  13 3 18 9 20 15 6 10 21 4  
 $n$  7 14 9 16 7 13 10 13 3 19

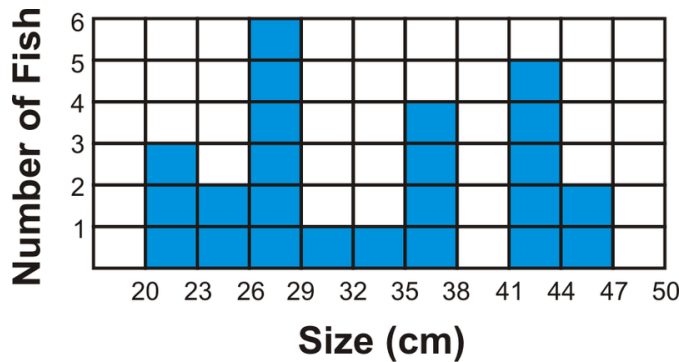
**Circle Graphs, Bar Graphs, Histograms, and Stem-and-Leaf Plots**

**Section A** – All questions in this section are selected response. Circle the correct answer.

1. In the following stem-and-leaf plot that represents the ages of 23 people waiting in line at Tim Horton’s, how many people were older than 32?

Stem	Leaf
0	4
1	0, 7, 8
2	3, 3, 4, 7, 8
3	2, 2, 2, 3, 5, 7, 7
4	0, 0, 1, 1, 3
5	6, 7

- a. 4
- b. 12
- c. 14
- d. 11



2.

The above histogram shows data collected during a recent fishing derby. The number of fish caught is being compared to the size of the fish caught. How many fish caught were between 20 cm and 29 cm in length?

- (a) 3
- (b) 11
- (c) 25
- (d) 6

3. What name is given to a distribution that has 2 peaks of the same height?

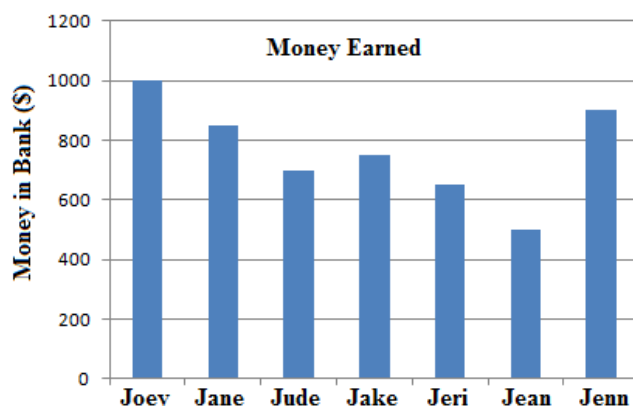
- a. uniform
  - b. unimodal
  - c. bimodal
  - d. discrete
4. What is the midpoint of the bin [14.5-23.5)?
- a. 19
  - b. 4.5
  - c. 18.5
  - d. 38
5. The following stem-and-leaf plot shows the cholesterol levels of a random number of students. These values range from 2.0 to 9.0. What percentage of the students have levels between 5.0 and 7.0, inclusive?

Stem	Leaf
2	3, 4, 6
3	2, 4, 6, 7, 8
4	2, 3, 3, 4, 6, 7, 9
5	0, 3, 6, 7
6	0, 3
7	1
8	2, 7, 9

- a. 6 %
  - b. 20%
  - c. 24%
  - d. 28%
6. What is the dependent variable in the following relationship? The time it takes to run the 100 yard dash and the fitness level of the runner.
- a. fitness level
  - b. time
  - c. length of the track
  - d. age of the runner
7. What name is given to the graph that uses lines to join the midpoints of the classes?
- a. bar graph
  - b. stem-and-leaf
  - c. histogram
  - d. frequency polygon
8. What is the mode of the following data set displayed in a stem-and-leaf plot?

Stem	Leaf
0	4
1	0, 7, 8
2	3, 3, 4, 7, 8
3	2, 2, 2, 3, 5, 7, 7
4	0, 0, 1, 1, 3
5	6, 7

- 32
- 41
- 7
- 23



How many of the above people have less than \$850 in the bank?

- 2
- 4
- 3
- 6

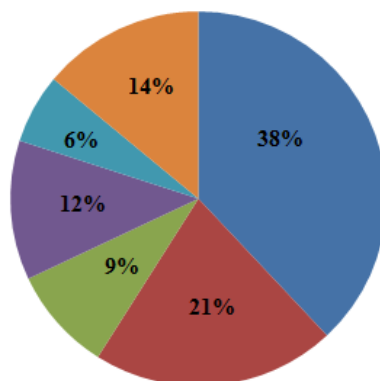
- On a recent math test, 9 students out of the 25 who took the test scored above 85. What percentage of the students scored above 85?
  - 0.36 %
  - 3.6 %
  - 360 %
  - 36 %

**Section B** – All questions in this section are long-answer questions. Be sure to show all of the work necessary to arrive at the correct answer.

- Construct a stem-and-leaf plot for the following data values:

20 12 39 38 18 58 49 59 66 50  
 23 32 43 53 67 35 29 13 42 55  
 37 19 38 22 46 71 9 65 15 38

2. The following pie chart represents the time spent doing various activities during one day. Using the chart, supply a possible activity that may be represented by each of the percentages shown in the graph.



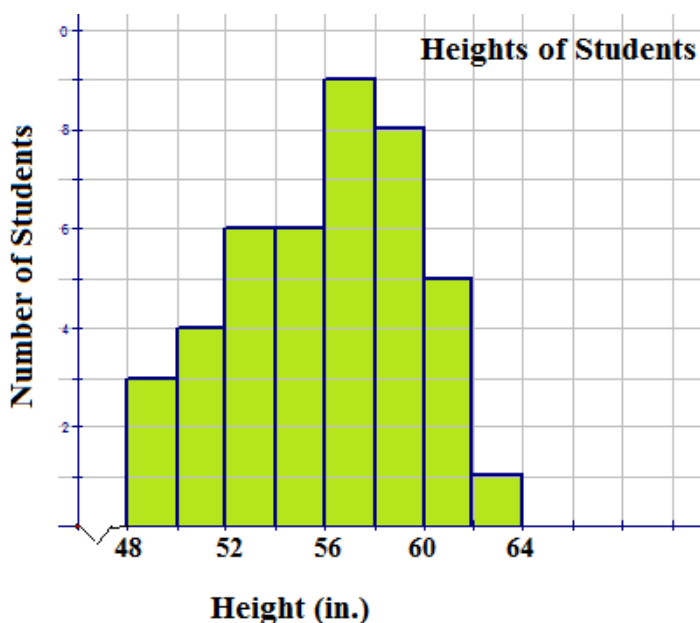
3. Just like Presidents of the United States, Canadian Prime Ministers must be sworn into office. The following data represents the ages of 22 Canadian Prime Ministers when they were sworn into office. Construct a stem-and-leaf plot to represent the ages, and list 4 facts that you know from the graph.

52 74 60 39 65 46 55 66 54 51 70 47 69 47 57 46  
48 66 61 59 46 45

4. A questionnaire on the makes of people's vehicles showed the following responses from 30 participants. Construct a frequency distribution and a bar graph to represent the data. (*F* = Ford, *H* = Honda, *V* = Volkswagen, *M* = Mazda)

*F M M M V M F M F V H H F V F*  
*H H F M M V H M V V F V H M F*

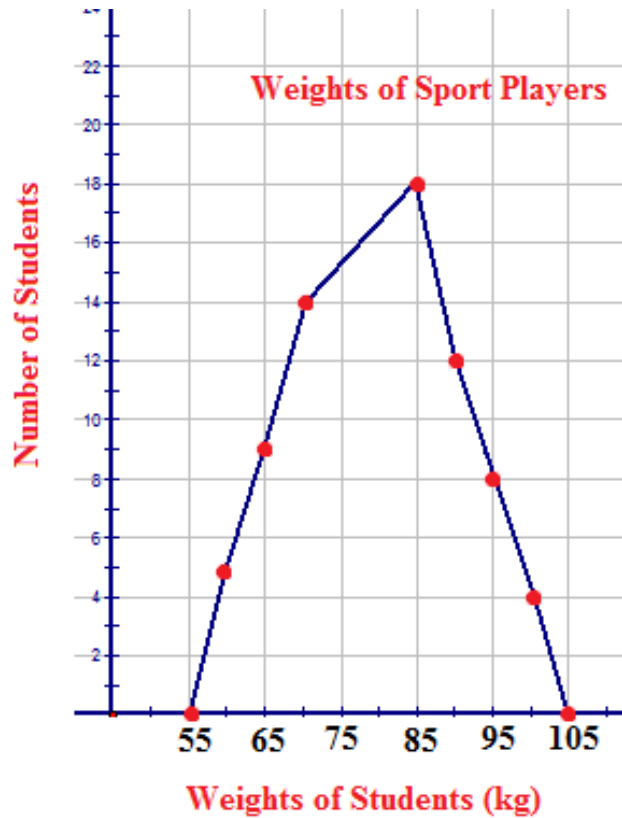
5. The following histogram displays the heights of students in a classroom:



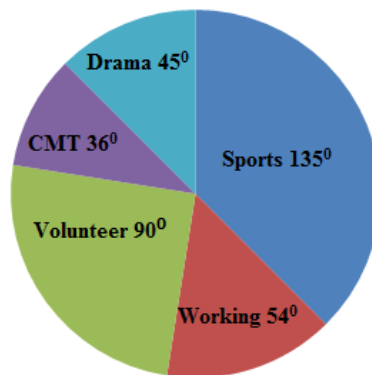
Use the information represented in the histogram to answer the following questions:

- How many students were in the class?
- How many students were over 60 inches in height?

- c. How many students had a height between 54 in and 62 in?
  - d. Is the distribution unimodal or bimodal? How do you know?
6. The following frequency polygon represents the weights of players who all participated in the same sport. Use the polygon to answer the following questions:



- a. How many players played the sport?
  - b. What was the most common weight for the players?
  - c. What sport do you think the players may have been playing?
  - d. What do the weights of 55 kg and 105 kg represent?
  - e. What 2 weights have no recorded players weighing those amounts?
7. The following circle graph, which is incomplete, shows the extracurricular activities for 200 high school students:



Use the circle graph to answer the following questions:

- a. What makes the above circle graph incomplete?
- b. How many students participated in sports?

- c. How many students do CMT?
  - d. How many students participate in volunteer activities after school?
  - e. Construct the circle graph so that it shows percentages and not degrees.
8. The following data represents the results of a test taken by a group of students:

95 56 70 83 59 66 88 52 50 77 69 80  
54 75 68 78 51 64 55 67 74 57 73 53

Construct a frequency distribution table using a bin size of 10 and display the results in a properly labeled histogram.

9. Using the data for question 8, use technology to construct the histogram.
10. In a few sentences, explain the type of graph that you find most helpful for interpreting data.

### Box-and-Whisker Plots

**Section A** – All questions in this section are selected response.

1. Which of the following is not a part of the five-number summary?
  - a.  $Q_1$  and  $Q_3$
  - b. the mean
  - c. the median
  - d. minimum and maximum values
2. What percent of the data is contained in the box of a box-and-whisker plot?
  - a. 25%
  - b. 100%
  - c. 50%
  - d. 75%
3. What name is given to the horizontal lines to the left and right of the box of a box-and-whisker plot?
  - a. axis
  - b. whisker
  - c. range
  - d. plane
4. What term describes the distribution of a data set if the median of the data set is located to the left of the center of the box in a box-and-whisker plot?
  - a. positively skewed
  - b. negatively skewed
  - c. approximately symmetric
  - d. not skewed
5. What 2 values of the five-number summary are connected with 2 horizontal lines on a box-and-whisker plot?
  - a. Minimum value and the median
  - b. Maximum value and the median
  - c. Minimum and maximum values
  - d.  $Q_1$  and  $Q_3$

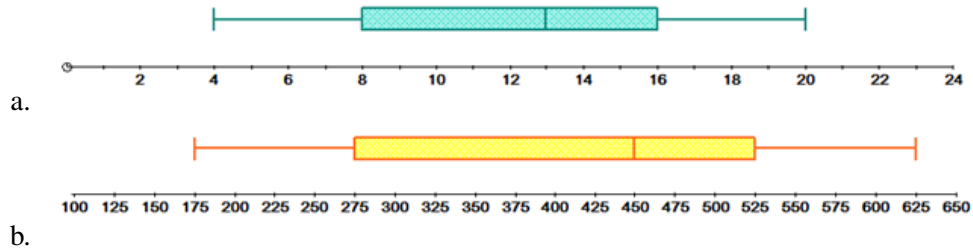
**Section B** - Show all work necessary to answer each question.

1. For the following data sets, determine the five-number summaries:
  - a. 74, 69, 83, 79, 60, 75, 67, 71

### 7.4. Review Questions

b. 6, 9, 3, 12, 11, 9, 15, 5, 7

2. For each of the following box-and-whisker plots, list the five-number summary and comment on the distribution of the data:



3. The following data represents the number of coins that 12 randomly selected people had in their piggy banks:

35 58 29 44 104 39 72 34 50 41 64 54

Construct a box-and-whisker plot for the above data.

4. The following data represent the time (in minutes) that each of 20 people waited in line at a local book store to purchase the latest Harry Potter book:

15 8 5 10 14 17 21 23 6 19 31 34 30 31  
3 22 17 25 5 16

Construct a box-and-whisker plot for the above data. Are the data skewed in any direction?

5. Firman’s Fitness Factory is a new gym that offers reasonably-priced family packages. The following table represents the number of family packages sold during the opening month:

24	21	31	28	29
27	22	27	30	32
26	35	24	22	34
30	28	24	32	27
32	28	27	32	23
20	32	28	32	34

Construct a box-and-whisker plot for the data. Are the data symmetric or skewed?

6. The following data represents the number of flat-screen televisions assembled at a local electronics company for a sample of 28 days:

48	55	51	44	59	49	47
45	51	56	50	57	53	55
47	49	51	54	56	54	47
50	53	52	55	51	59	48

Using technology, construct a box-and-whisker plot for the data. What are the values for the five-number summary?

7. Construct a box-and-whisker plot to represent the average number of sick days used by 9 employees of a large industrial plant. The numbers of sick days are as follows:

39 31 18 34 25 22 32 23 22

8. Shown below is the number of new stage shows that appeared in Las Vegas for each of the past several years. Construct a box-and-whisker plot for the data and comment of the shape of the distribution.

31 29 34 30 38 40 36 38 32 39 35



9. The following data represent the average snowfall (in centimeters) for 18 Canadian cities for the month of January. Construct a box-and-whisker plot to model the data. Is the data skewed? Justify your answer.

**TABLE 7.16:**

<b>Name of City</b>	<b>Amount of Snow(cm)</b>
Calgary	123.4
Charlottetown	74.5
Edmonton	80.6
Fredericton	73.8
Halifax	64.0
Labrador City	110.4
Moncton	82.4
Montreal	63.6
Ottawa	48.9
Quebec City	53.8
Regina	35.9
Saskatoon	25.4
St. John's	97.5
Sydney	44.2
Toronto	21.8
Vancouver	12.8
Victoria	8.3
Winnipeg	76.2

---

10. Using the procedure outlined in this chapter, check the following data sets for outliers:
- 25, 33, 55, 32, 17, 19, 15, 18, 21
  - 149, 123, 126, 122, 129, 120

# CHAPTER 8

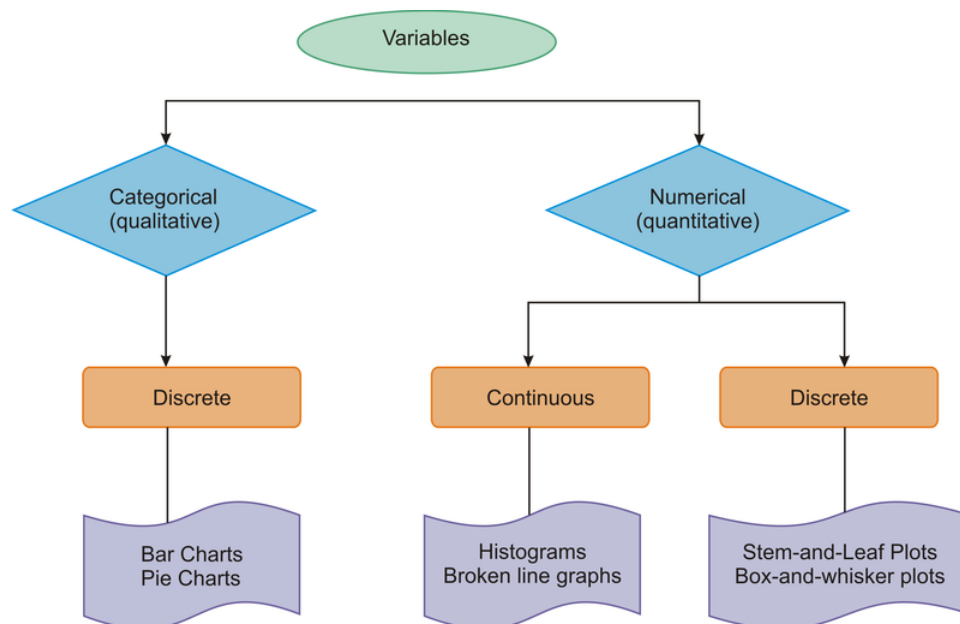
## Organizing and Displaying Data for Comparison

### Chapter Outline

- 8.1 REVIEW
- 8.2 DOUBLE LINE GRAPHS
- 8.3 TWO-SIDED STEM-AND-LEAF PLOTS
- 8.4 DOUBLE BAR GRAPHS
- 8.5 DOUBLE BOX-AND-WHISKER PLOTS
- 8.6 REVIEW QUESTIONS

### Introduction

Throughout this book, you have learned about variables. You have learned about random variables, discrete variables, continuous variables, numerical (or quantitative) variables, and categorical (or qualitative) variables. The various forms of graphical representations you have learned about in the previous chapters can be added to your learning of variables. The graphic below may help to summarize what you have learned.



Broken-line graphs, histograms, pie charts, stem-and-leaf plots, and box-and-whisker plots all represent useful (often very useful) tools in determining trends. Broken-line graphs, for example, allow you to show situations such as the distance traveled in specific time spans. Histograms use continuous grouped data to show the frequency trend in the data. Bar charts are a little different from histograms in that they use grouped discrete data, as do stem-and-leaf plots. Bar graphs, as you know, have gaps between the columns, while histograms do not. Stem-and-leaf plots are excellent for giving you a quick visual representation of data. Used for only smaller sets of data, stem-and-leaf plots are a good example of representations of grouped discrete data. Box-and-whisker plots are a final visual way of representing grouped data that you have learned about in the previous chapters. In a box-and-whisker plot, you are able to find the five-number summary to describe the spread of the data.

## 8.1 Review

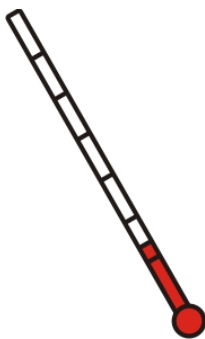
### Learning Objectives

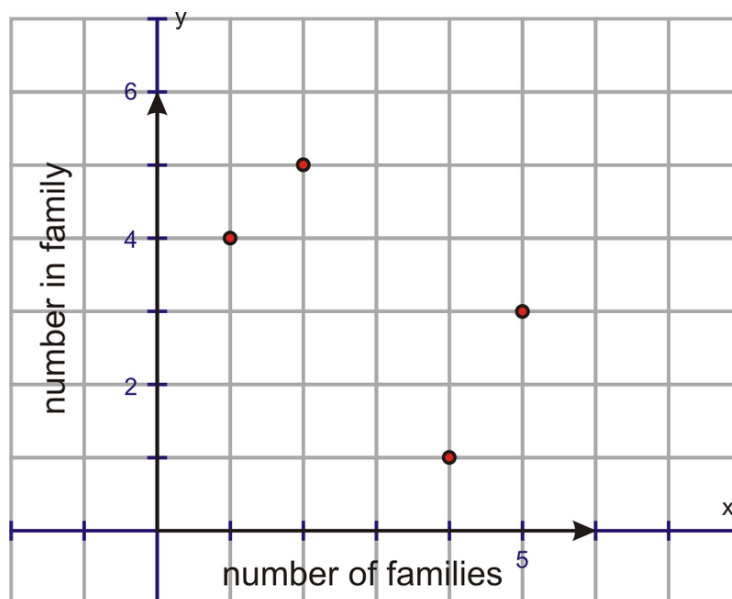
- Organize and describe distributions of data by using a number of different methods, including frequency tables, histograms, standard line and bar graphs, stem-and-leaf displays, scatter plots, and box-and-whisker plots.

In previous chapters, you learned about discrete and continuous data and were introduced to categorical and numerical forms of displaying data. In this final chapter, you will learn how to display discrete and continuous data in both categorical and numerical displays, but in a way that allows you to compare sets of data.

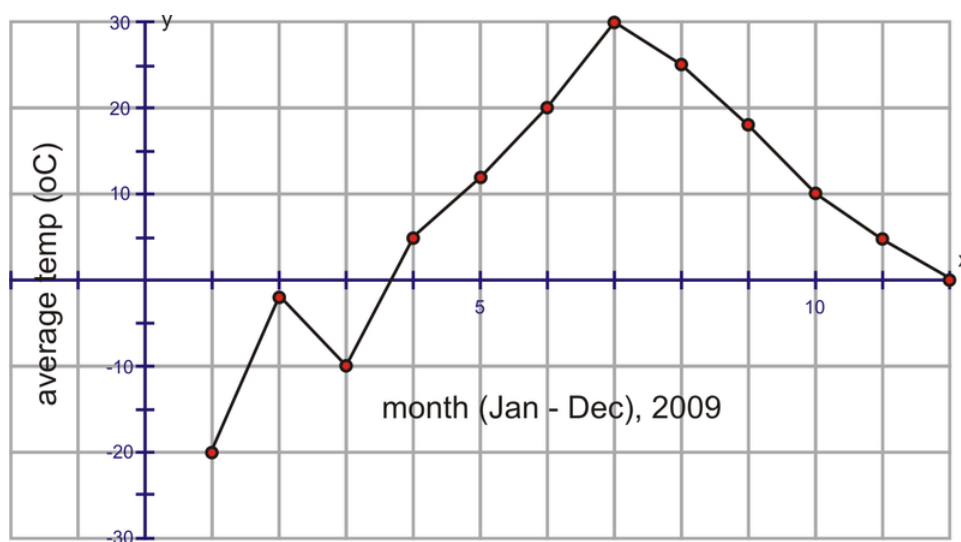


Remember that discrete data is represented by exact values that result from counting, as in the number of people in the households in your neighborhood. Continuous data is represented by a range of data that results from measuring. For example, taking the average temperatures for each month during a year is an example of continuous data. Also remember from an earlier chapter how you distinguished between these types of data when you graphed them.



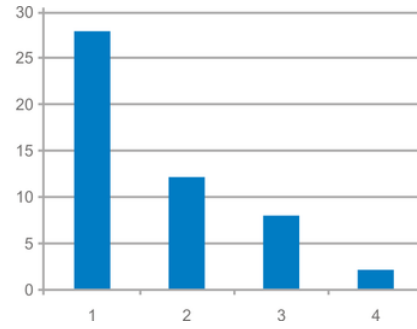
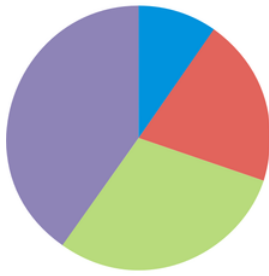


The graph above shows discrete data. Remember that you know this because the data points are not joined. The graph below represents the average temperatures during the months in 2009. This data is continuous. You can easily tell this by looking at the graph and seeing the data points connected together.

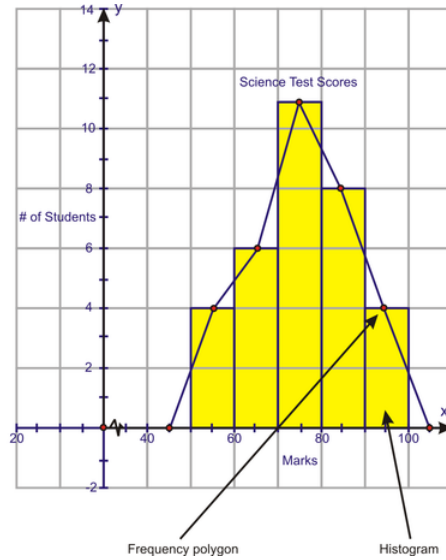
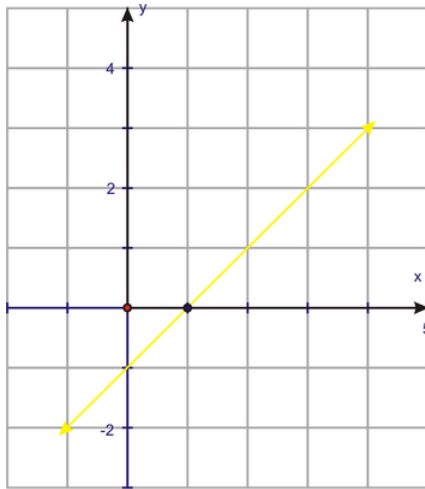


2 newer terms used are the categorical and numerical data forms. **Categorical data** forms are just what the term suggests. These are data forms that are in categories and describe characteristics, or qualities, of a category. These data forms are more **qualitative data** and, therefore, are less numerical than they are descriptive. Graphs such as pie charts and bar charts show descriptive data, or qualitative data. Below are 2 examples of categorical data represented in these types of graphs:

### 8.1. Review

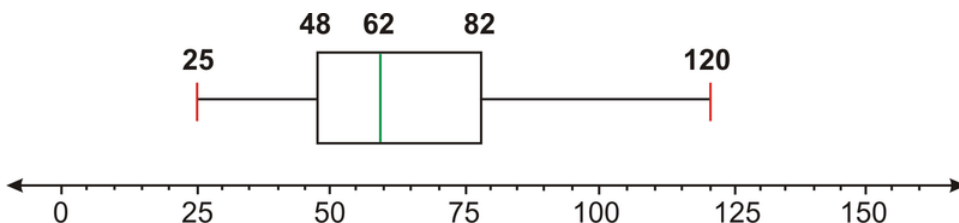


**Numerical data is quantitative data.** Numerical data involves measuring or counting a numerical value. Therefore, when you talk about discrete and continuous data, you are talking about numerical data. Line graphs, frequency polygons, histograms, and stem-and-leaf plots all involve numerical data, or quantitative data, as is shown below:



Stem	Leaf
2	5, 8
3	4, 5, 5, 5
4	0, 0, 2, 7, 9
5	0, 0, 0, 0, 5, 5, 8

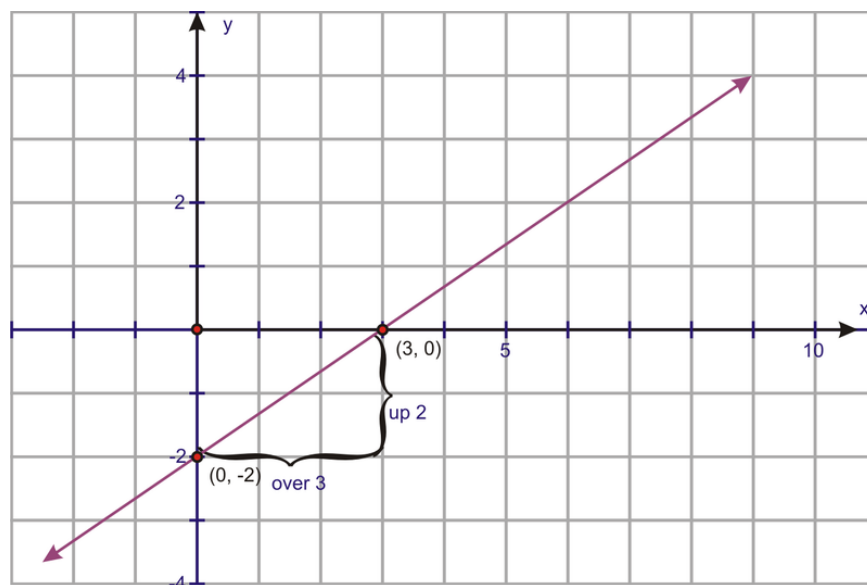
Box-and-whisker plots are also considered numerical displays of data, as they are based on quantitative data (the mean and median), as well as the maximum (upper) and minimum (lower) values found in the data. The figure below is a typical box-and-whisker plot:



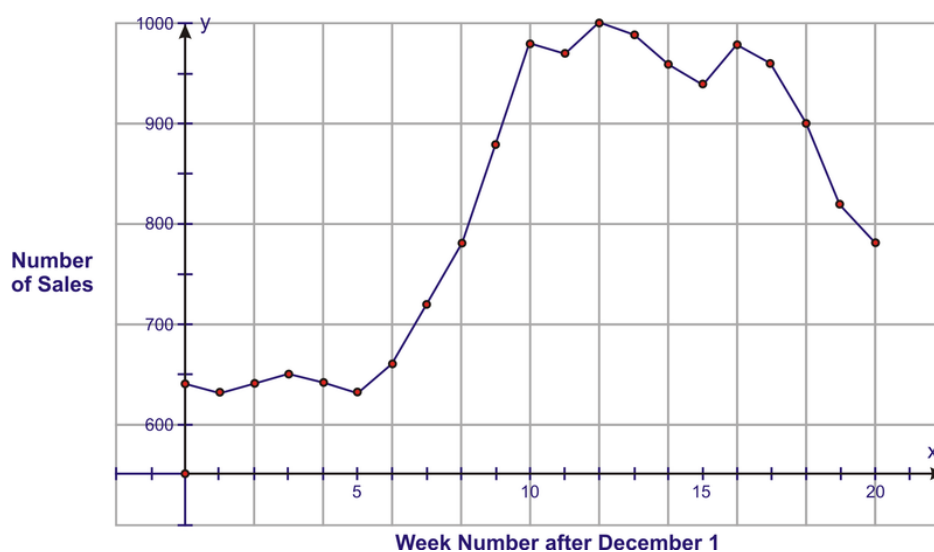
You will spend the remainder of the chapter learning about how to compare sets of categorical and numerical data.

## 8.2 Double Line Graphs

Remember a line graph, by definition, can be the result of a linear function or can simply be a graph of plotted points, where the points are joined together by line segments. Line graphs that are linear functions are normally in the form  $y = mx + b$ , where  $m$  is the slope and  $b$  is the  $y$ -intercept. The graph below is an example of a linear equation with a slope of  $\frac{2}{3}$  and a  $y$ -intercept of  $-2$ :



The second type of line graph is known as a broken-line graph. In a broken-line graph, the slope represents the rate of change, and the  $y$ -intercept is actually the starting point. The graph below is a broken-line graph:



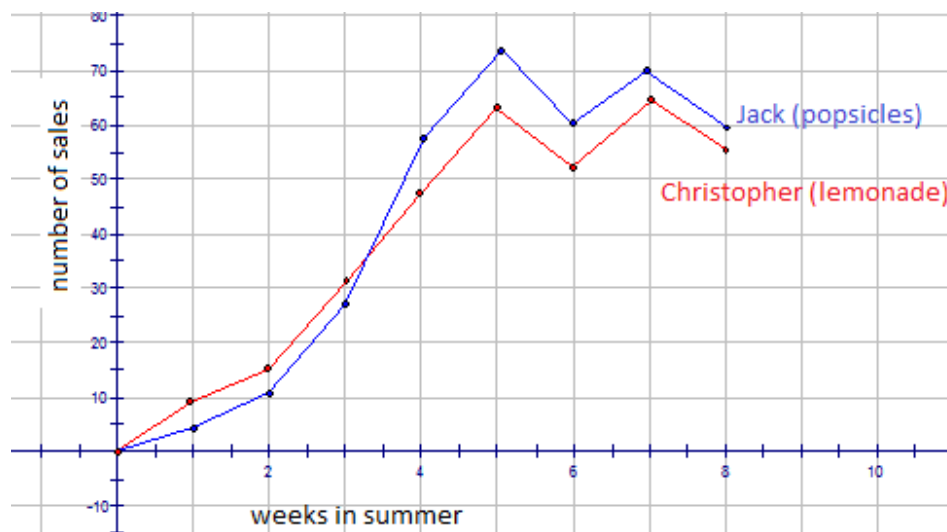
When the measurements began, the number of sales (the  $y$ -intercept) was 645. The graph shows a significant increase in the number of sales from weeks 5 through 10 and a significant reduction in the number of sales from weeks 16 through 20.



In this lesson, you will be learning about comparing 2 line graphs that each contain data points. In statistics, when line graphs are in the form of broken-line graphs, they are of more use. Linear functions (i.e.,  $y = mx + b$ ) are more for algebraic reasoning. **Double line graphs**, as with any double graphs, are often called parallel graphs, due to the fact that they allow for the quick comparison of 2 sets of data. In this chapter, you will see them referred to only as double graphs.

### Example 1

Christopher and Jack are each opening businesses in their neighborhoods for the summer. Christopher is going to sell lemonade for 50¢ per glass. Jack is going to sell popsicles for \$1.00 each. The following graph represents the sales for each boy for the 8 weeks in the summer.



- Explain the slopes of the line segments for Christopher's graph.
- Explain the slopes of the line segments for Jack's graph.
- Are there any negative slopes? What does this mean?
- Where is the highest point on Christopher's graph? What does this tell you?
- Where is the highest point on Jack's graph? What does this tell you?
- Can you provide some reasons for the shape of Jack's graph?
- Can you provide some reasons for the shape of Christopher's graph?

### Solution:

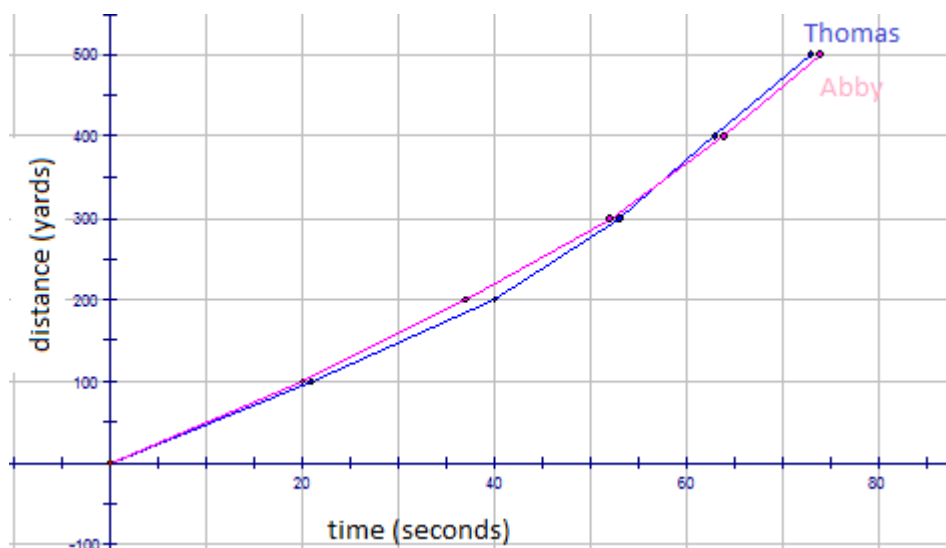
- The slope of the line segments for Christopher's graph (red) are positive for the first 5 weeks, meaning he was increasing his sales each week. This is also true from weeks 6 to 7. From weeks 5 to 6 and weeks 7 to 8, the slopes were decreasing, meaning there was a decrease in sales.
- The same trend that is seen for Christopher's graph (red) is also seen for Jack's graph (blue). The slope of the line segments for Jack's graph (blue) are positive for the first 5 weeks, meaning he was increasing his sales each week. This is also true from weeks 6 to 7. From weeks 5 to 6 and weeks 7 to 8, the slopes were decreasing, meaning there was a decrease in sales.
- Negative sales from weeks 5 to 6 and weeks 7 to 8 (for both boys) mean there was a decrease in sales during these 2-week periods.
- The highest point on Christopher's graph occurred in week 7, when he sold 65 glasses of lemonade. This must have been a very good week—nice and hot!

## 8.2. Double Line Graphs

- e. The highest point on Jack's graph occurred in week 5, when he sold 74 popsicles. This must have been a very hot week as well!
- f. Popsicles are a great food when you are warm and want a light snack. You can see how as the summer became hotter, the sales increased. Even in the weeks where it looks like Jack had a decrease in sales (maybe a few rainy days occurred, or it was not as hot), his sales still remained at a good level.
- g. Lemonade is a very refreshing drink when you are warm. You can see how as the summer became hotter, the sales increased. Even in the weeks where it looks like Christopher had a decrease in sales (maybe a few rainy days occurred, or it was not as hot), his sales still remained at a good level, just as Jack's sales did.

### Example 2

Thomas and Abby are training for the cross country meet at their school. Both students are in the 100 yard dash. The coach asks them to race 500 yards and time each 100 yard interval. The following graph represents the times for both Thomas and Abby for each of the five 100 yard intervals.



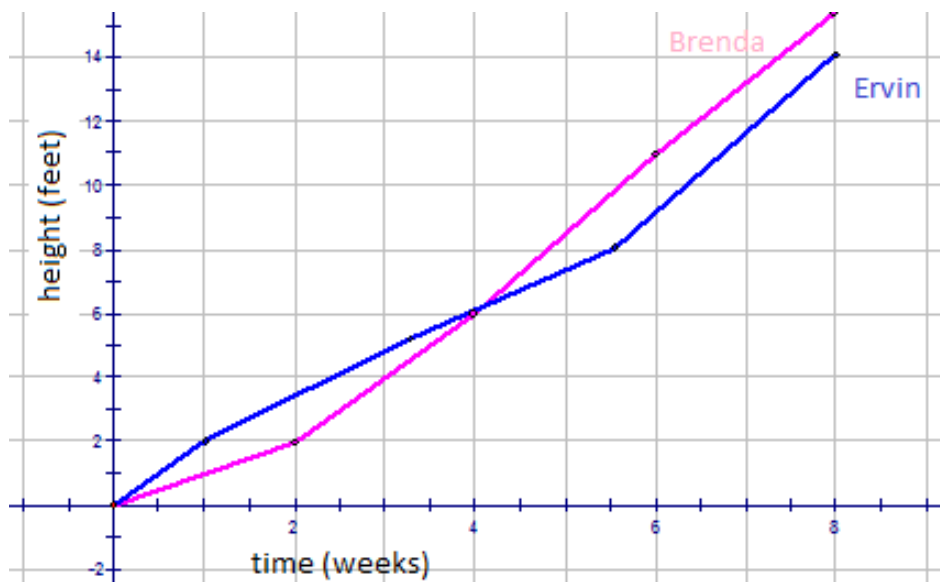
- Who won the race? How do you know?
- Between what times did Thomas (blue) appear to slow down? How do you know?
- Between what times was Abby (red) ahead of Thomas? How do you know?
- At what time did Thomas pass Abby? How do you know?

### Solution:

- Thomas (blue) won the race, because he finished the 500 yards in the least amount of time.
- Between 20 and 40 seconds, Thomas (blue) seems to slow down, because the slope of the graph is less steep.
- Between 0 and 57 seconds, Abby (pink) is ahead of Thomas (blue). You can see this, because the pink line is above the blue line.
- At 57 seconds, Thomas (blue) passes Abby (pink). From this point onward, the blue line is above the pink line, meaning Thomas is running faster 100 yard intervals.

### Example 3

Brenda and Ervin are each planting corn in a section of garden in their back yard. Brenda says that they need to put fertilizer on the plants 3 to 5 times per week. Ervin contradicts Brenda, saying that they need to fertilize only 1 to 2 times per week. Each gardener plants his or her garden of corn and measures the heights of their plants. The graph for the growth of their corn is found below:



- Who was right? How do you know?
- Between what times did Brenda's (pink) garden appear to grow more? How do you know?
- Between what times were Ervin's (blue) heights ahead of Brenda's? How do you know?

**Solution:**

- Brenda (pink) is correct, because her plants grew more in the same amount of time.
- Between 4 and 8 weeks, Brenda's plants seemed to grow faster (taller) than Ervin's plants. You can tell this, because the pink line is above the blue line after the 4-week mark.
- From 0 and 4 weeks, Ervin's plants seemed to grow faster (taller) than Brenda's plants. You can tell this, because the blue line is above the pink line before the 4-week mark.

**Example 4**

Nicholas and Jordan went on holidays with their families. They decided to monitor the mileage they traveled by keeping track of the time and the distance they were on the road. The boys collected the following data:

Nicholas

Time (hr)	1	2	3	4	5	6
Distance (miles)	60	110	175	235	280	320

Jordan

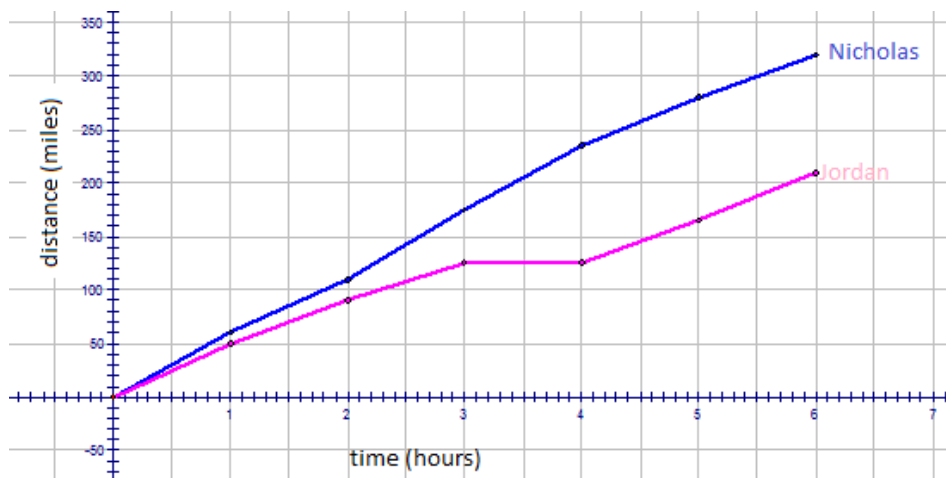
Time (hr)	1	2	3	4	5	6
Distance (miles)	50	90	125	125	165	210

- Draw a graph to show the trip for each boy.
- What conclusions could you draw by looking at the graphs?

**Solution:**

8.2. Double Line Graphs

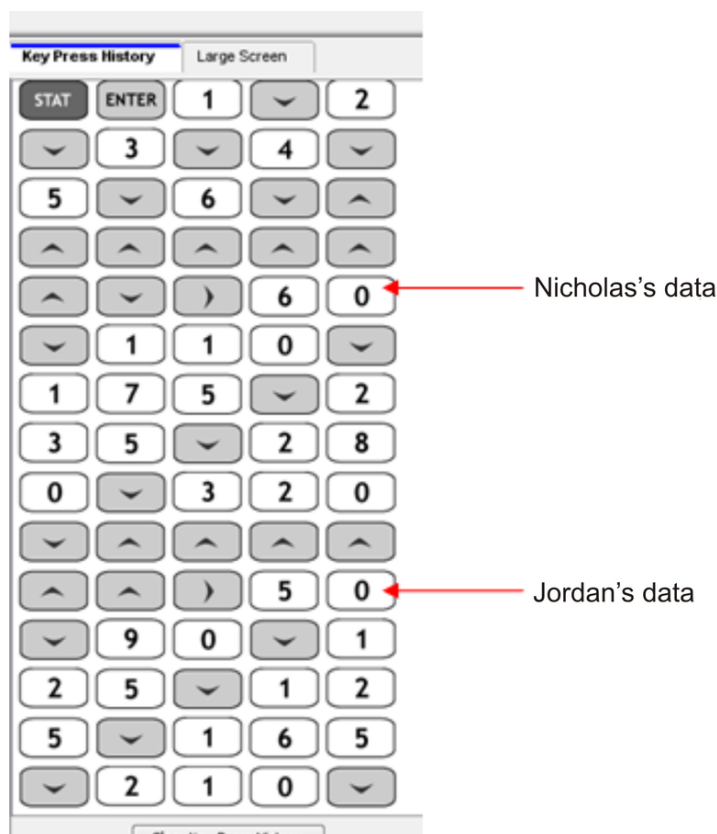
a.



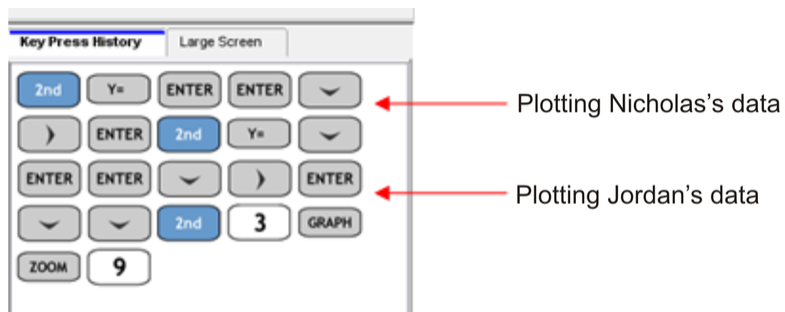
b. Looking at the speed of Nicholas's family vehicle and the shape of the graph, it could be concluded that Nicholas's family was traveling on the highway going toward their family vacation destination. The family did not stop and continued on at a pretty steady speed until they reached where they were going.

Jordan's trip was more relaxed. The speed indicates they were probably not on a highway, but more on country-type roads, and that they were traveling through a scenic route. In fact, from hours 3 to 4, the family stopped for some reason (maybe lunch), and then they continued on their way.

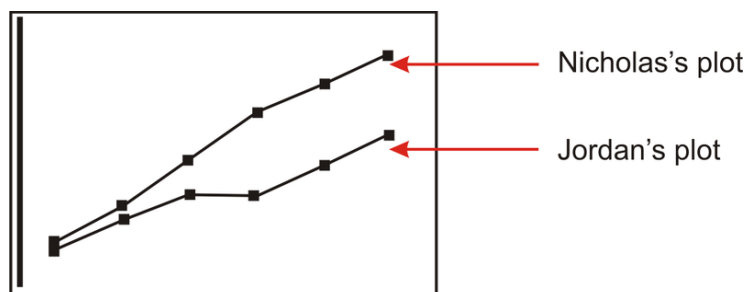
You can also use TI technology to graph this data. First, you need to enter in all of the data Nicholas and Jordan collected.



Now you need to graph the 2 sets of data.



The resulting graph looks like the following:



## 8.3 Two-sided Stem-and-Leaf Plots

As you have learned in an earlier chapter, stem-and-leaf plots are an excellent tool for organizing data. Remember that stem-and-leaf plots are a visual representation of grouped discrete data, but they can also be referred to as a modal representation. This is because by looking at a stem-and-leaf plot, we can determine the mode by quick visual inspection. In the last chapter, you learned about single-sided stem-and-leaf plots. In this lesson, you will learn about **two-sided stem-and-leaf plots**, which are also often called back-to-back stem-and-leaf plots.

### Example 5

The girls and boys in one of BDF High School's AP English classes are having a contest. They want to see which group can read the most number of books. Mrs. Stubbard, their English teacher, says that the class will tally the number of books each group has read, and the highest mode will be the winner. The following data was collected for the first semester of AP English:

Girls	11	12	12	17	18	23	23	23	24	33	34	35	44	45	47	50	51	51
Boys	15	18	22	22	23	26	34	35	35	35	40	40	42	47	49	50	50	51

- Draw a two-sided stem-and-leaf plot for the data.
- Determine the mode for each group.
- Help Mrs. Stubbard decide which group won the contest.

### Solution:

a.

Girls		Boys
7, 8, 2, 2, 1	1	5, 8
3, 3, 3, 2	2	2, 2, 3, 6
5, 4, 3	3	4, 5, 5, 5
7, 5, 4	4	0, 0, 2, 7, 9
1, 1, 0	5	0, 0, 1

- The mode for the girls is 23 books. It is the number in the girls column that appears most often. The mode for the boys is 35 books. It is the number in the boys column that appears most often.
- Mrs. Stubbard should decide that the boys group has won the contest.

### Example 6

Mrs. Cameron teaches AP Statistics at GHI High School. She recently wrote down the class marks for her current grade 12 class and compared it to the previous grade 12 class. The data can be found below. Construct a two-sided stem-and-leaf plot for the data and compare the distributions.

2010 class	70	70	70	71	72	74	74	74	74	75	76	76	77	78	79	80	81
	82	82	82	83	84	85	85	86	87	93	98	100					
2009 class	76	76	76	76	77	78	78	78	79	80	80	82	82	83	83	83	85
	85	88	91	95													

**Solution:**

2009 class		2010 class
9, 8, 8, 8, 7, 6, 6, 6, 6	7	0, 0, 0, 1, 2, 4, 4, 4, 4, 5, 6, 6, 7, 8, 9
8, 5, 5, 3, 3, 3, 2, 2, 0, 0	8	0, 1, 2, 2, 2, 3, 4, 5, 5, 6, 7
5, 1	9	3, 8
	10	0

There is a wide variation in the marks for both years in Mrs. Cameron's AP Statistics Class. In 2009, her class had marks anywhere from 76 to 95. In 2010, the class marks ranged from 70 to 100. The mode for the 2009 class was 76, but for the 2010 class, it was 74. It would seem that the 2009 class had, indeed, done slightly better than Mrs. Cameron's current class.

### Example 7

The following data was collected in a survey done by Connor and Scott for their statistics project. The data represents the ages of people who entered into a new hardware store within its first half hour of opening on its opening weekend. The M's in the data represent males, and the F's represent females.

12M	18F	15F	15M	10M	21F	25M	21M
26F	29F	29F	31M	33M	35M	35M	35M
41F	42F	42M	45M	46F	48F	51M	51M
55F	56M	58M	59M	60M	60F	61F	65M
65M	66M	70M	70M	71M	71M	72M	72F

Construct a back-to-back stem-and-leaf plot showing the ages of male customers and the ages of female customers. Compare the distributions.

**Solution:**

### 8.3. Two-sided Stem-and-Leaf Plots

Male		Female
5, 2, 0	1	5, 8
5, 1	2	1, 6, 9, 9
5, 5, 5, 3, 1	3	
5, 2	4	1, 2, 6, 8
9, 8, 6, 1, 1	5	5
6, 5, 5, 0	6	0, 1
2, 1, 1, 0, 0	7	2

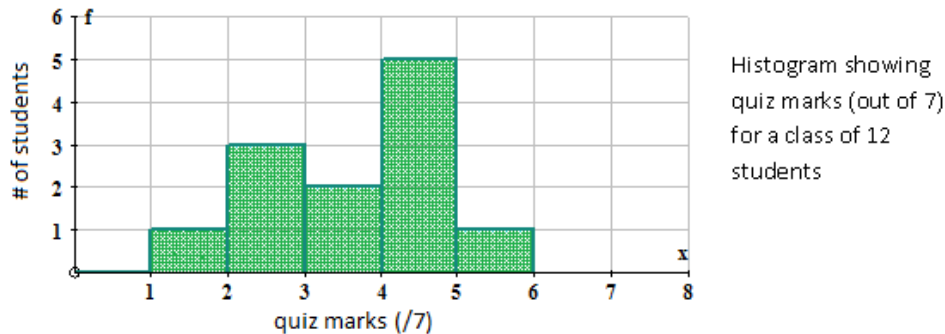
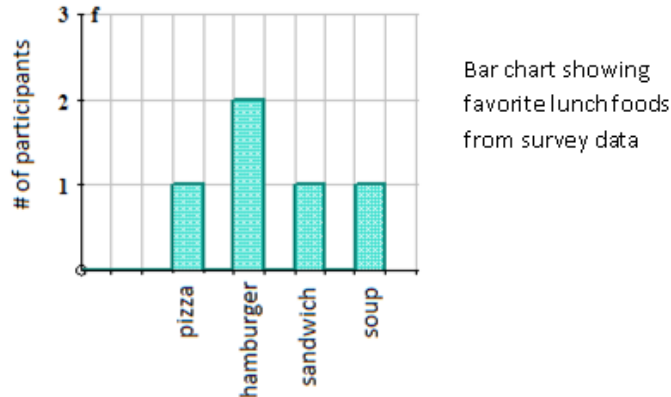
For the male customers, the ages ranged from 10 to 72. The ages for the male customers were spread out throughout this range, with the mode being age 35. In other words, for the males found to be at the store in the first half hour of opening day, there was no real age category where a concentration of males could be found.

For the female customers, the ages ranged from 15 to 72, but they were concentrated between 21 and 48. The mode for the ages of the female customers was 29 years of age.



## 8.4 Double Bar Graphs

In Chapter 7, you studied both histograms and bar graphs. Remember that histograms have measurements on the horizontal axis ( $x$ ) and frequencies on the vertical axis ( $y$ ). A bar chart, on the other hand, displays categories on the horizontal ( $x$ ) axis and frequencies on the vertical ( $y$ ) axis. This means that bar charts are more qualitative, and, therefore, display categorical data. The figure below shows 1 bar chart (on the top) and 1 histogram (on the bottom):



Let's look at an example of **double bar graphs**.

### Example 8

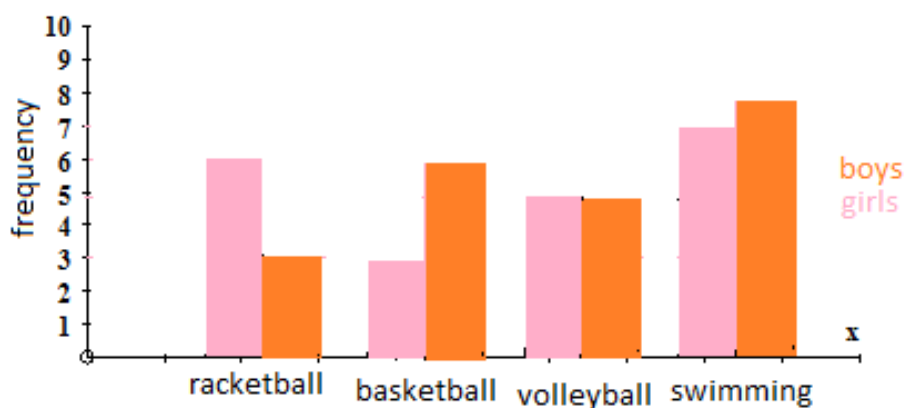
Kerry-Sue is surveying a random sample of students to determine which sports they would like to have set up at the end-of-year Safe Grad event. She collects the following data:

**TABLE 8.1:**

Sports	Girls	Boys
Racquetball	6	3
Basketball	3	6
Volleyball	5	5
Swimming	7	8

Draw a double bar graph and help Kerry-Sue determine which 2 sports would be most equally-liked by both boys and girls at the end-of-year Safe Grad event.

**Solution:**



According to the double bar graph, volleyball and swimming seem to be almost equally liked by both the girls and the boys. Therefore, these 2 sports would be the ones Kerry-Sue should choose to set up at the end-of-year Safe Grad event.

### Example 9

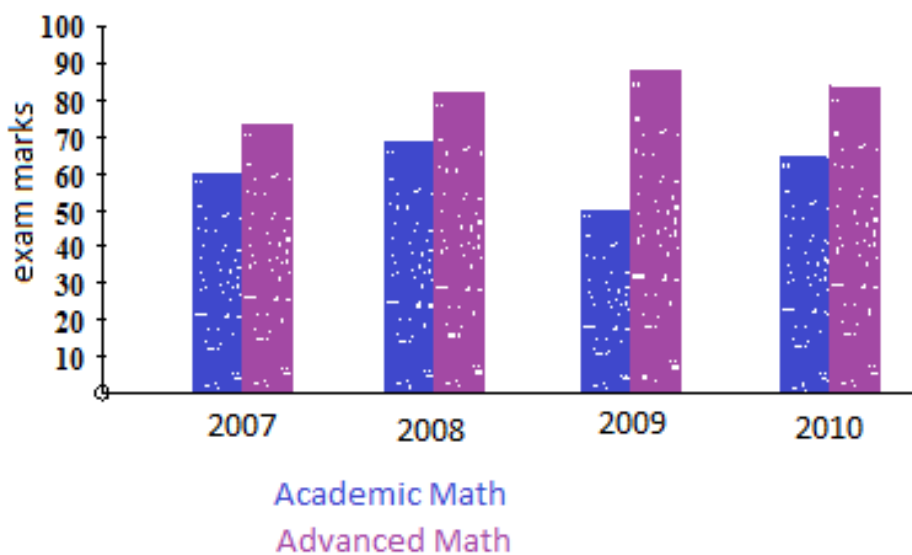
Mrs. Smith teaches both academic and advanced math. She has been teaching these 2 courses for the past 4 years. She decided she wanted to compare her grades to see how each class was doing over the past few years and see if she has improved her class instruction at all. Her data can be found below:

**TABLE 8.2:**

Marks	Academic Math	Advanced Math
2007	61.3	74.7
2008	67.9	80.3
2009	50.9	86.8
2010	63.7	81.5

Draw a double bar graph and help Mrs. Smith determine if her class instruction has improved over the past 4 years.

**Solution:**



Both the academic and advanced math marks went up and down over the past 4 years. If Mrs. Smith looks at the difference between 2007 and 2010, she can see that there is an improvement in the final grades for her students. Although there are many factors that can affect these grades, she can say that the change in her instruction is making some difference in the results for her students. Other factors might have contributed to the huge decline in grades for the academic math students in 2009. You can make this conclusion for Mrs. Smith, as there was a marked improvement in her advanced math course. In 2009, it seemed her instruction methods were working well with the advanced students, but other factors were affecting the academic students.

## 8.5 Double Box-and-Whisker Plots

**Double box-and-whisker plots** give you a quick visual comparison of 2 sets of data, as was also found with other double graph forms you learned about earlier in this chapter. The difference with double box-and-whisker plots is that you are also able to quickly visually compare the means, the medians, the maximums (upper range), and the minimums (lower range) of the data.

### Example 10

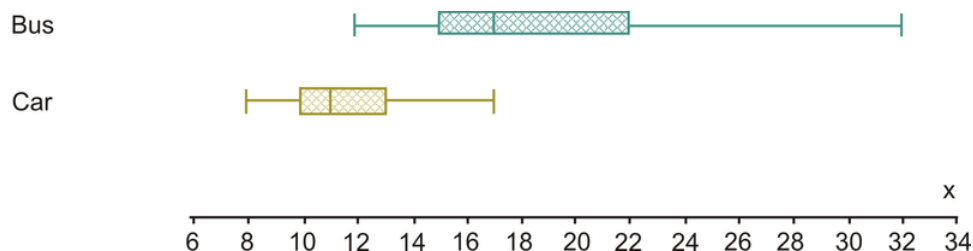
Emma and Daniel are surveying the times it takes students to arrive at school from home. There are 2 main groups of commuters who were in the survey. There were those who drove their own cars to school, and there were those who took the school bus. They collected the following data:

Bus times (min)	14	18	16	22	25	12	32	16	15	18
Car times (min)	12	10	13	14	9	17	11	10	8	11

Draw a box-and-whisker plot for both sets of data on the same number line. Use the double box-and-whisker plots to compare the times it takes for students to arrive at school either by car or by bus.

### Solution:

When plotted, the box-and-whisker plots look like the following:



Using the medians, 50% of the cars arrive at school in 11 minutes or less, whereas 50% of the students arrive by bus in 17 minutes or less. The range for the car times is  $17 - 8 = 9$  minutes. For the bus times, the range is  $32 - 12 = 20$  minutes. Since the range for the driving times is smaller, it means the times to arrive by car are less spread out. This would, therefore, mean that the times are more predictable and reliable.

### Example 11

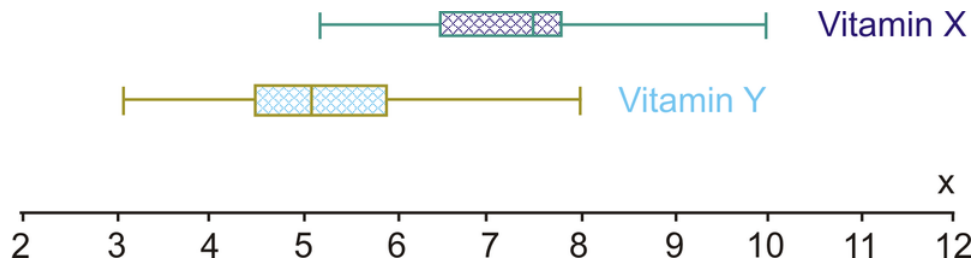
A new drug study was conducted by a drug company in Medical Town. In the study, 15 people were chosen at random to take Vitamin X for 2 months and then have their cholesterol levels checked. In addition, 15 different people were randomly chosen to take Vitamin Y for 2 months and then have their cholesterol levels checked. All 30 people had cholesterol levels between 8 and 10 before taking one of the vitamins. The drug company wanted to see which of the 2 vitamins had the greatest impact on lowering people's cholesterol. The following data was collected:

Vitamin X	7.2	7.5	5.2	6.5	7.7	10	6.4	7.6	7.7	7.8	8.1	8.3	7.2	7.1	6.5
Vitamin Y	4.8	4.4	4.5	5.1	6.5	8	3.1	4.6	5.2	6.1	5.5	4.2	4.5	5.9	5.2

Draw a box-and-whisker plot for both sets of data on the same number line. Use the double box-and-whisker plots to compare the 2 vitamins and provide a conclusion for the drug company.

**Solution:**

When plotted, the box-and-whisker plots look like the following:



Using the medians, 50% of the people in the study had cholesterol levels of 7.5 or lower after being on Vitamin X for 2 months. Also 50% of the people in the study had cholesterol levels of 5.1 or lower after being on Vitamin Y for 2 months. Knowing that the participants of the survey had cholesterol levels between 8 and 10 before beginning the study, it appears that Vitamin Y had a bigger impact on lowering the cholesterol levels. The range for the cholesterol levels for people taking Vitamin X was  $10 - 5.2 = 4.8$  points. The range for the cholesterol levels for people taking Vitamin Y was  $8 - 3.1 = 4.9$  points. Therefore, the range is not useful in making any conclusions.

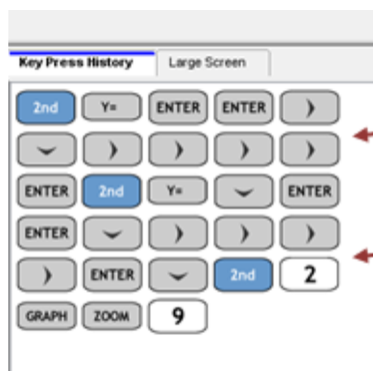
**Drawing Double Box-and-Whisker Plots Using TI Technology**

The above double box-and-whisker plots were drawn using a program called Autograph. You can also draw double box-and-whisker plots by hand using pencil and paper or by using your TI-84 calculator. Follow the key sequence below to draw double box-and-whisker plots.



The first thing you have to do is enter the data in for Vitamin X into [L1] and the data from the Vitamin Y study into [L2]. A portion of this key sequence is here but it is the same as we have used throughout this book.

After entering the data into L1 and L2, the next step is to graph the data by using STAT PLOT.

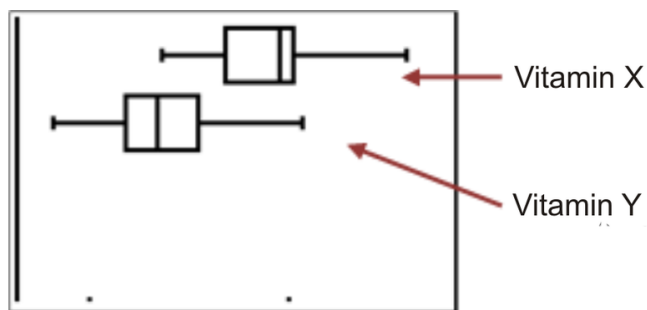


Setting up [STAT PLOT] 1 for [L1]

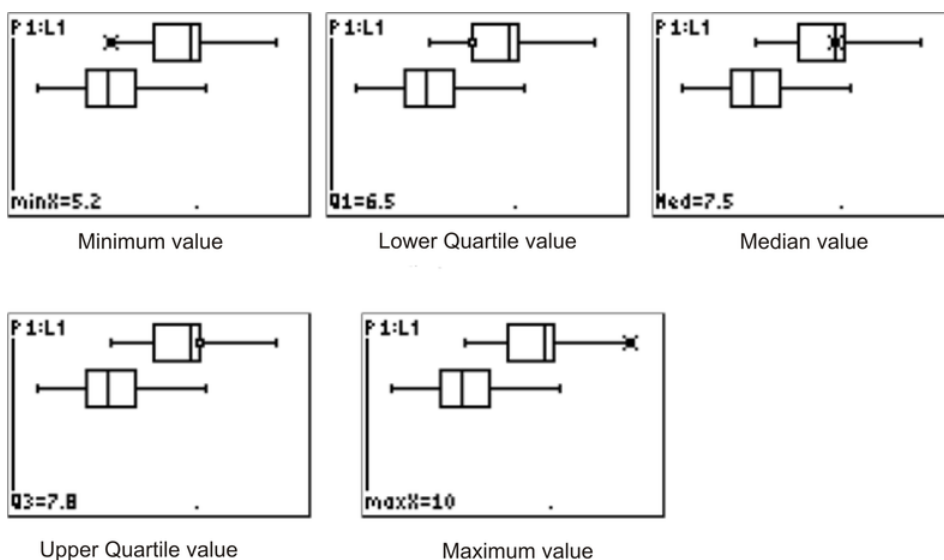
Setting up [STAT PLOT] 1 for [L2]

The resulting graph looks like the following:

### 8.5. Double Box-and-Whisker Plots



You can then press **TRACE** and find the five-number summary. The five-number summary is shown below for Vitamin X. By pressing the **▼** button, you can get to the second box-and-whisker plot (for Vitamin Y) and collect the five-number summary for this box-and-whisker plot.



### Example 12

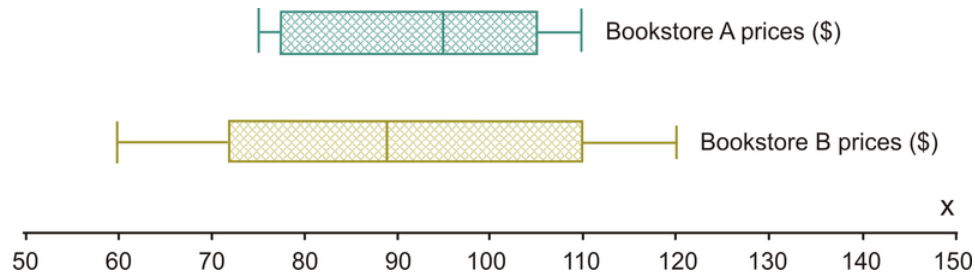
2 campus bookstores are having a price war on the prices of their first-year math books. James, a first-year math major, is going into each store to try to find the cheapest books he can find. He looks at 5 randomly chosen first-year books for first-year math courses in each store to determine where he should buy the 5 textbooks he needs for his courses this coming year. He collects the following data:

Bookstore A prices(\$)	95	75	110	100	80
Bookstore B prices(\$)	120	60	89	84	100

Draw a box-and-whisker plot for both sets of data on the same number line. Use the double box-and-whisker plots to compare the 2 bookstores' prices, and provide a conclusion for James as to where to buy his books for his first-year math courses.

### Solution:

The box-and-whisker plots are plotted and look like the following:



Using the medians, 50% of the books at Bookstore A are likely to be in the price range of \$95 or less, whereas at Bookstore B, 50% of the books are likely to be around \$89 or less. At first glance, you would probably recommend to James that he go to Bookstore B. Let's look at the range to see the spread of data. For Bookstore A, the range is  $\$110 - \$75 = \$35$ . For Bookstore B, the range is  $\$120 - \$60 = \$60$ . With the spread of the data much greater at Bookstore B than at Bookstore A, (i.e., the range for Bookstore B is greater than that for Bookstore A), to say that it would be cheaper to buy James's books at Bookstore A would be more predictable and reliable. You would, therefore, suggest to James that he is probably better off going to Bookstore A.

### Points to Consider

- What is the difference between categorical and numerical data, and how does this relate to qualitative and quantitative data?
- How is comparing double graphs (pie charts, broken-line graphs, box-and-whisker plots, etc.) useful when doing statistics?

### Vocabulary

**Categorical data** Data that are in categories and describe characteristics, or qualities, of a category.

**Double bar graphs** 2 bar graphs that are graphed side-by-side.

**Double box-and-whisker plots** 2 box-and-whisker plots that are plotted on the same number line.

**Double line graphs** 2 line graphs that are graphed on the same coordinate grid. Double line graphs are often called parallel graphs.

**Quantitative data** Numerical data, or data that is in the form of numbers.

**Qualitative data** Descriptive data, or data that describes categories.

**Numerical data** Data that involves measuring or counting a numerical value.

**Two-sided stem-and-leaf plots** 2 stem-and-leaf plots that are plotted side-by-side. Two-sided stem-and-leaf plots are also called back-to-back stem-and-leaf plots.

## 8.6 Review Questions

Answer the following questions and show all work (including diagrams) to create a complete answer.

- In the table below, match the following types of graphs with the types of variables used to create the graphs.

**TABLE 8.3:**

Type of Graph	Type of Variable
a. Histogram	_____ discrete
b. Stem-and-leaf plot	_____ discrete
c. Broken-line graph	_____ discrete
d. Bar chart	_____ continuous
e. Pie chart	_____ continuous

- In the table below, match the following types of graphs with the types of variables used to create the graphs.

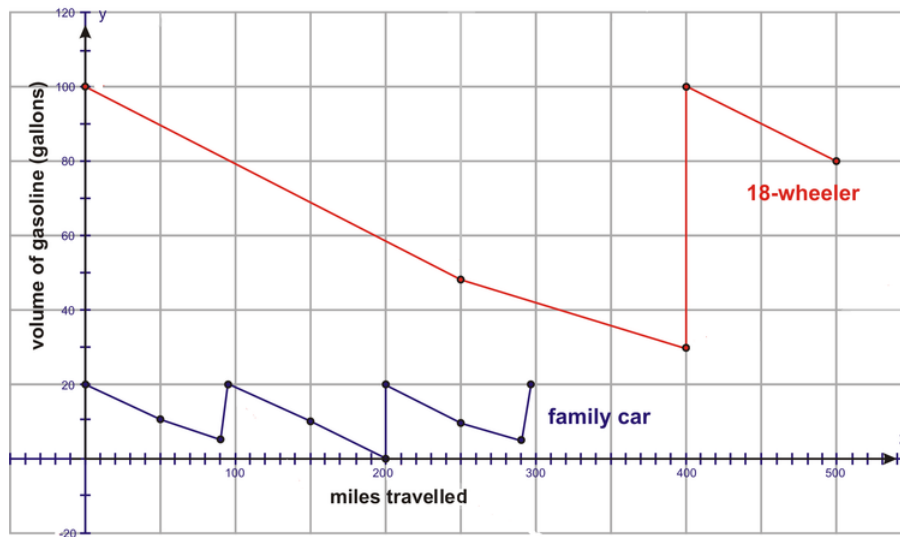
**TABLE 8.4:**

Type of Graph	Type of Variable
a. Broken-line graph	_____ qualitative
b. Bar chart	_____ numerical
c. Pie chart	_____ categorical
d. Stem-and-leaf plot	_____ quantitative
e. Histogram	_____ numerical

- Jack takes a pot of water at room temperature ( $22^{\circ}\text{C}$ ) and puts it on the stove to boil ( $100^{\circ}\text{C}$ ), which takes about 5 minutes. He then takes a cup of this water, adds a package of hot chocolate, and mixes it up. He places the cup on the counter to cool for 10 minutes to  $40^{\circ}\text{C}$  before having his first sip. After 30 minutes, the hot chocolate is now at room temperature. Thomas is making chocolate chip cookies. He mixes all of the ingredients together at room temperature, which takes him about 5 minutes, and then places the cookies in the oven at  $350^{\circ}\text{C}$  for 8 minutes. After cooking, he takes them off the pan and places them on a cooling rack. After 15 minutes, the cookies are still warm (about  $30^{\circ}\text{C}$ ), but he samples them for taste. After 30 minutes, the cookies are at room temperature and ready to be served. Draw a broken-line graph for each set of data. Label the graphs to show what is happening.
- Scott is asked to track his daily video game playing. He gets up at 7 A.M. and plays for 1 hour. He then eats his breakfast and gets ready for school. He runs to catch the bus at 8:25 A.M. On the bus ride (about 35 minutes), he plays his IPOD until arriving for school. He is not allowed games at school, so he waits for the bus ride home at 3:25 P.M. When he gets home, he does homework for 1 hour and plays games for 1 hour until dinner. There are no games in the evening. Michael gets up at 7:15 A.M., eats breakfast, and gets ready for school. It takes him 30 minutes to get ready. He then plays games until he goes to meet the bus with Scott. Michael is in Scott's class, but he has a free period from 11:00 A.M. until 11:45 A.M., when he goes outside to play a game. He goes home and plays his 1 hour of games immediately, and he then works on his homework until dinner. He, like Scott, is not allowed to play games in the evening. Draw a broken-line graph for each set of data. Label the graphs to show what is happening.
- The following graph shows the gasoline remaining in a car during a family trip east. Also found on the graph



is the gasoline remaining in a truck traveling west to deliver goods. Describe what is happening for each graph. What other conclusions may you draw?



6. Mr. Dugas, the senior high physical education teacher, is doing fitness testing this week in gym class. After each test, students are required to take their pulse rate and record it on the chart in the front of the gym. At the end of the week, Mr. Dugas looks at the data in order to analyze it. The data is shown below:

Girls	70	88	80	76	76	77	89	72	72	76	72	75	77	80	76	68	68
	82	78	60	64	64	65	81	84	84	79	78	70					
Boys	76	88	87	86	85	70	76	70	70	79	80	82	82	82	83	84	85
	85	78	81	85													

Construct a two-sided stem-and-leaf plot for the data and compare the distributions.

7. Starbucks prides itself on its low line-up times in order to be served. A new coffee house in town has also boasted that it will have your order in your hands and have you on your way quicker than the competition. The following data was collected for the line-up times (in minutes) for both coffee houses:

Starbucks	20	26	26	27	19	12	12	16	12	15	17	20	8	8	18
Just Us Coffee	17	16	15	10	16	10	10	29	20	22	22	12	13	24	15

Construct a two-sided stem-and-leaf plot for the data. Determine the median and mode using the two-sided stem-and-leaf plot. What can you conclude from the distributions?

8. The boys and girls basketball teams had their heights measured at practice. The following data was recorded for their heights (in inches):

Girls	171	170	176	176	177	179	162	172	160	157	155
	168	178	174	170	155	155	154	164	145	171	161
Boys	168	170	162	153	176	167	158	180	181	176	172
	168	167	165	159	185	184	173	177	167	169	177

Construct a two-sided stem-and-leaf plot for the data. Determine the median and mode using the two-sided stem-and-leaf plot. What can you conclude from the distributions?

9. The grade 12 biology class did a survey to see what color eyes their classmates had and if there was a connection between eye color and sex. The following data was recorded:

**TABLE 8.5:**

Eye color	Males	Females
blue	5	5
green	6	8
brown	3	4
hazel	4	3

Draw a double bar chart to represent the data, and draw any conclusions that you can from the resulting chart.

10. Robbie is in charge of the student organization for new food selections in the cafeteria. He designed a survey to determine if 4 new food options would be good to put on the menu. The results are shown below:

**TABLE 8.6:**

Food Option	Yes votes	No votes
Fish burgers	10	5
Vegetarian pizza	7	18
Brown rice	23	9
Carrot soup	20	20

Draw a double bar chart to represent the data, and draw any conclusions that you can from the resulting chart.

11. The guidance counselor at USA High School wanted to know what future plans the graduating class had. She took a survey to determine the intended plans for both boys and girls in the school's graduating class. The following data was recorded:

**TABLE 8.7:**

Future Plans	Boys	Girls
University	35	40
College	27	22
Military	23	9
Employment	10	5
Other/unsure	5	10

Draw a double bar chart to represent the data, and draw any conclusions that you can from the resulting chart.

12. International Baccalaureate has 2 levels of courses, which are standard level (SL) and higher level (HL). Students say that study times are the same for both the standard level exams and the higher level exams. The following data represents the results of a survey conducted to determine how many hours a random sample of students studied for their final exams at each level:

HL Exams	15	16	16	17	19	10	5	6	5	5	8	10	8	12	17
SL Exams	10	6	6	7	9	12	2	6	2	5	7	20	18	8	18

Draw a box-and-whisker plot for both sets of data on the same number line. Use the double box-and-whisker

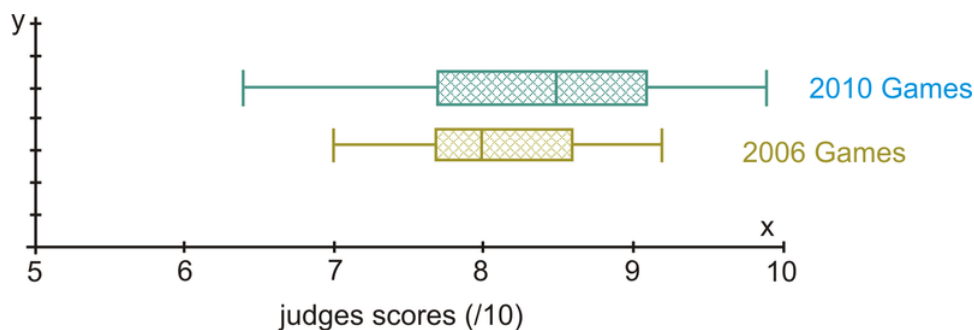
plots to determine the five-number summary for both sets of data. Compare the times students prepare for each level of exam.

13. Students in the AP math class at BCU High School took their SATs for university entrance. The following scores were obtained for the math and verbal sections:

Math	529	533	544	562	513	519	560	575	568	537	561	522	563	571
Verbal	499	509	524	530	550	499	545	560	579	524	478	487	482	570

Draw a box-and-whisker plot for both sets of data on the same number line. Use the double box-and-whisker plots to determine the five-number summary for both sets of data. Compare the data for the 2 sections of the SAT using the five-number summary data.

14. The following box-and-whisker plots were drawn to analyze the data collected in a survey of scores for the doubles performances in the figure skating competitions at 2 Winter Olympic games. The box-and-whisker plot on the top represents the scores obtained at the 2010 winter games in Whistler, BC. The box-and-whisker plot on the bottom represents the scores obtained at the 2006 winter games in Torino, Italy.



15. Use the double box-and-whisker plots to determine the five-number summary for both sets of data. Compare the scores obtained at each of the Winter Olympic games.

