Lesson 9: Using a Curve to Model a Data Distribution

Student Outcomes

- Students draw a smooth curve that could be used as a model for a given data distribution.
- Students recognize when it is reasonable and when it is not reasonable to use a normal curve as a model for a given data distribution.

Lesson Notes

This lesson introduces the concept of using a curve to model a data distribution. A smooth curve is used to model a relative frequency histogram, and the idea of an area under the curve representing the approximate proportion of data falling in a given interval is introduced. When data are approximated with a smooth curve, meaningful information can be learned about the distribution. The normal curve (a smooth curve that is bell shaped and symmetric) is introduced. Examples of data distributions that could reasonably be modeled using a normal curve and data distributions that cannot reasonably be modeled by a normal curve are both used in the lesson. In Lessons 10 and 11, students calculate the area under a normal curve and interpret the associated probabilities in context.

Classwork

Example 1 (5 minutes): Heights of Dinosaurs and the Normal Curve

All of the histograms in this lesson are relative frequency histograms. Consider reviewing the meaning of relative frequency prior to having students work on the first exercises. Relative frequency histograms were introduced in Grade 6 and Algebra I. If necessary, discuss how the height of each bar of the histogram is interpreted as the proportion of the data values that fall in the corresponding interval rather than the number of the data values (the frequency) in the interval. The relative frequency can be expressed as either a decimal or a percent.

In many of the exercises, students are asked to find an approximate percent of the data that are within one standard deviation of the mean. Students should base their estimates on the relative frequency that can be found by adding the heights of the bars within one standard deviation of the mean. When the mark for the standard deviation falls within a bar, have students round to the nearest edge of the bar.

MP.4

In several exercises, students model with mathematics when they draw a smooth curve that could be used to model the distribution. Suggest to students that if the distribution is approximately normal, the curve they draw should be bell shaped and roughly passing through the midpoints of bars and the peak in the center of the distribution. When students draw the curve, allow some leeway on the appearance of the curve. This section is the first introduction to modeling a distribution with a curve.

Scaffolding:

For students working below grade level, consider using Exercises 1–7 as an opportunity for teacher modeling or as an activity for mixed-ability groups.

Consider asking students working above grade level to develop their own plans for answering the question, "How tall was a *compy* dinosaur?" Allow them to perform calculations and create their own data displays to answer this question.

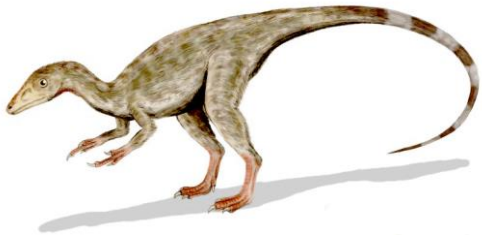
To develop motivation for the activities, consider using a discussion question, such as the following:

- Imagine you are a scientist studying dinosaurs that lived millions of years ago. What are some questions you might try to answer about these dinosaurs?
  - *Expect multiple responses such as how much they weighed, average life span, how fast they were, etc.*

In this example, the question to be answered is "How tall was a *compy* dinosaur?" Display the table of data showing the heights of 660 compy dinosaurs. Ask students what each column represents, and emphasize the meaning of relative frequency.

Example 1: Heights of Dinosaurs and the Normal Curve

A paleontologist studies prehistoric life and sometimes works with dinosaur fossils. The table below shows the distribution of heights (rounded to the nearest inch) of 660 procompsognathids, otherwise known as compys.



The heights were determined by studying the fossil remains of the compys.

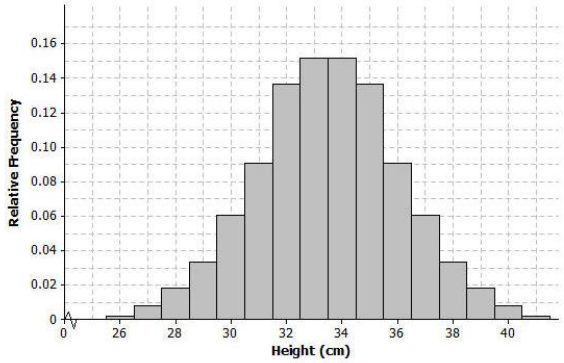
Height (cm)	Number of Compys	Relative Frequency
26	1	0.002
27	5	0.008
28	12	0.018
29	22	0.033
30	40	0.061
31	60	0.091
32	90	0.136
33	100	0.152
34	100	0.152
35	90	0.136
36	60	0.091
37	40	0.061
38	22	0.033
39	12	0.018
40	5	0.008
41	1	0.002
Total	660	1.000

Exercises 1–8 (15 minutes)

Let students work independently on Exercises 1–8. Then, discuss answers as a class. Some students may have a slightly different answer for the percent within one standard deviation. Since students are approximating an answer, results will vary. Ask students to explain how they arrived at their answers (percent). In addition, ask students to share how they drew the smooth curve for Exercise 6.

Exercises 1–8

The following is a relative frequency histogram of the compy heights:



1. **What does the relative frequency of 0.136 mean for the height of 32 cm?**

   13.6% of the 660 compys were 32 cm tall.

2. **What is the width of each bar? What does the height of the bar represent?**

   Each bar has a width of 1 cm. The height of the bar is the relative frequency for the corresponding compy height.

3. **What is the area of the bar that represents the relative frequency for compys with a height of 32 cm?**

   The area of the bar is equal to the relative frequency of 0.136.

4. **The mean of the distribution of compy heights is 33.5 cm, and the standard deviation is 2.56 cm. Interpret the mean and standard deviation in this context.**

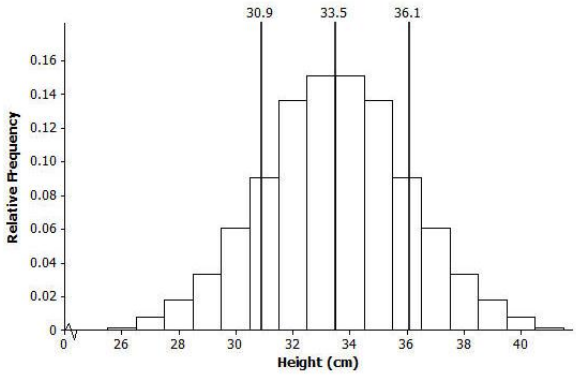
   The mean of 33.5 cm is the average height of the compys in the sample. It can be interpreted as a typical compy height.

   The standard deviation of 2.56 cm is the typical distance that a compy height is from the mean height.

5. **Mark the mean on the graph, and mark *one* deviation above and below the mean.**

   a. **Approximately what percent of the values in this data set are within one standard deviation of the mean (i.e., between $33.5 \text{ cm} - 2.56 \text{ cm} = 30.94 \text{ cm}$ and $33.5 \text{ cm} + 2.56 \text{ cm} = 36.06$ cm)?**

   $0.091 + 0.136 + 0.152 + 0.152 + 0.136 + 0.091 = 0.758 = 75.8\%$

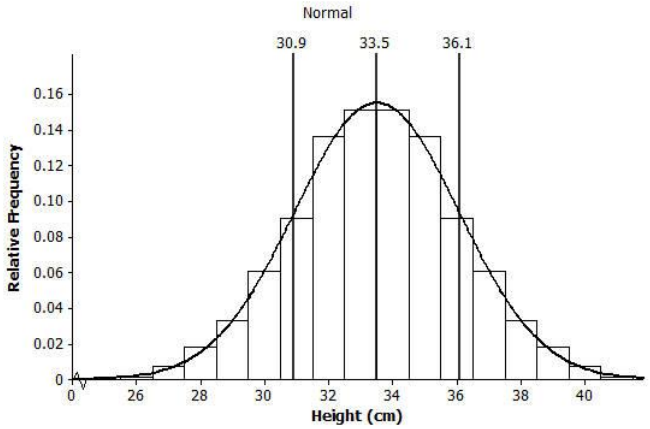


b. Approximately what percent of the values in this data set are within *two* standard deviations of the mean?

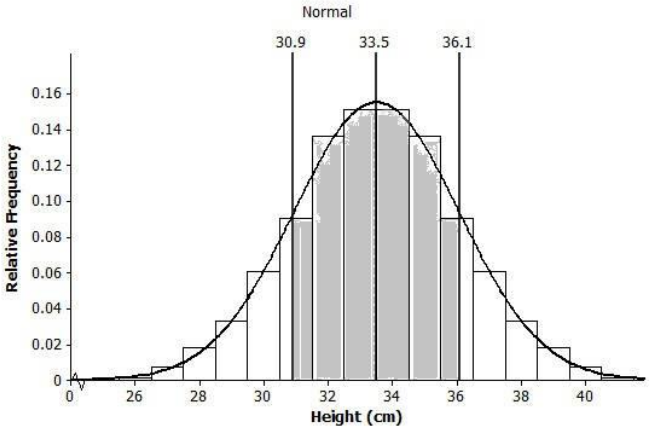
Between 29 and 38, about 94%

6. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Describe the shape of the distribution.

The curve is bell shaped and approximately symmetric and mound shaped.



7. Shade the area of the histogram that represents the proportion of heights that are within one standard deviation of the mean.



8. Based on our analysis, how would you answer the question, "How tall was a compy?"

Answers will vary. It is likely that students will say the height was between 31 cm and 36 cm.

Example 2 (5 minutes): Gas Mileage and the Normal Distribution

Read through the example as a class.

Example 2: Gas Mileage and the Normal Distribution

A normal curve is a smooth curve that is symmetric and bell shaped. Data distributions that are mound shaped are often modeled using a normal curve, and we say that such a distribution is approximately normal. One example of a distribution that is approximately normal is the distribution of compy heights from Example 1. Distributions that are approximately normal occur in many different settings. For example, a salesman kept track of the gas mileage for his car over a 25-week span.

The mileages (miles per gallon rounded to the nearest whole number) were

$23, 27, 27, 28, 25, 26, 25, 29, 26, 27, 24, 26, 26, 24, 27, 25, 28, 25, 26, 25, 29, 26, 27, 24, 26.$

Exercise 9 (10 minutes)

MP.5

Students are asked to find the mean and standard deviation using technology. Ask students to indicate how technology helps them make sense of the data. If using a graphing calculator similar to the TI-84, the mileages are entered into L1 and the frequency into L2. To find the mean and standard deviation, select 1-Var Stats and type *L1, L2*. After the 1-Var Stats entry, select Enter. Consult an appropriate manual or similar resource if using a different type of calculator or if using a statistical software package that is different from the program described above.

Remind students that when they construct the histogram, they should center the mileage in the middle of each bar.

Let students work with a partner or in a small group based on available access to technology.

Exercise 9

9. Consider the following:

a. Use technology to find the mean and standard deviation of the mileage data. How did you use technology to assist you?

***Mean* $= 26.04$ mpg**
***Standard deviation* $= 1.54$ mpg**

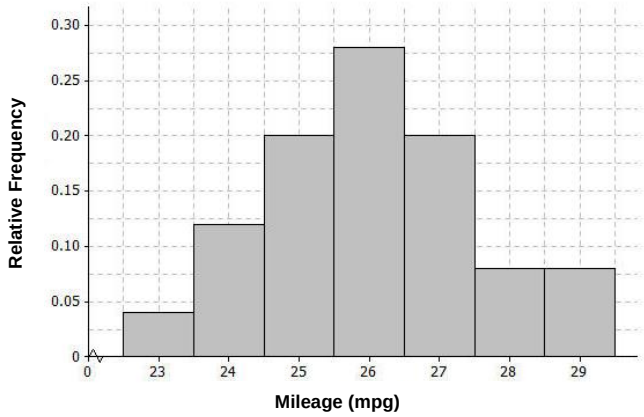
The graphing calculator does several tedious calculations for me. I entered the data into lists and was able to indicate what calculations I wanted done by writing an expression using lists. I did not have to set up the organization to find the standard deviation and perform the rather messy calculations.

b. Calculate the relative frequency of each of the mileage values. For example, the mileage of 26 mpg has a frequency of 7. To find the relative frequency, divide 7 by 25, the total number of mileages recorded. Complete the following table:

Mileage (mpg)	Frequency	Relative Frequency
23	1	0.04
24	3	0.12
25	5	0.20
26	7	0.28
27	5	0.20
28	2	0.08
29	2	0.08
Total	25	1.00

c. **Construct a relative frequency histogram using the scale below.**

Completed histogram:



d. **Describe the shape of the mileage distribution. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Is this approximately a normal curve?**

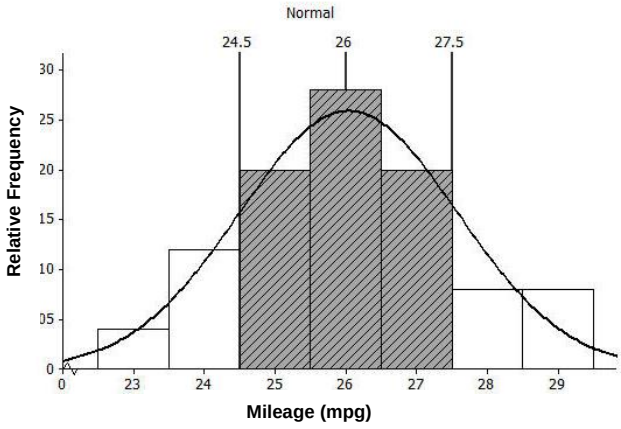
The shape is approximately normal. See the graph in part (e).

e. **Mark the mean on the histogram. Mark one standard deviation to the left and right of the mean. Shade the area of the histogram that represents the proportion of mileages that are within one standard deviation of the mean. Find the proportion of the data within one standard deviation of the mean.**

One standard deviation to the left (or below) the mean: $26.04 - 1.54 = 24.5$

One standard deviation to the right (or above) the mean: $26.04 + 1.54 = 27.58$

The proportion of the data within one standard deviation of the mean is approximately 0.68 *(which is the sum of* $0.20 + 0.28 + 0.20$*).*



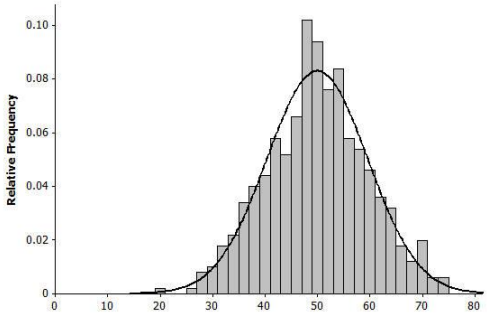
Closing (2 minutes)

Ask students to summarize the main ideas of the lesson with a neighbor or in writing. Use this as an opportunity to informally assess comprehension of the lesson. The Lesson Summary below offers some important ideas that should be included.

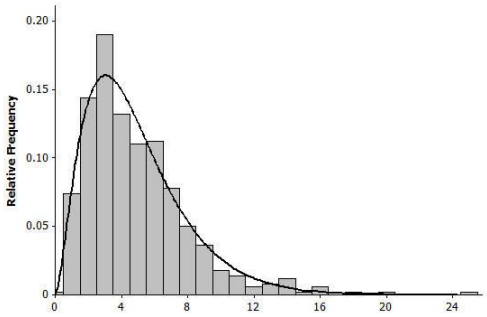
- Is the mean of a distribution that is approximately normal located near where the curve is the highest?
  - *Yes*
- Is the mean of a skewed distribution located near where the curve is the highest? Why does this happen?
  - *No. In a skewed distribution, the mean will be pulled toward the values in the tail of the distribution.*

Lesson Summary

- **A normal curve is symmetric and bell shaped. The mean of a normal distribution is located in the center of the distribution. Areas under a normal curve can be used to estimate the proportion of the data values that fall within a given interval.**



- **When a distribution is skewed, it is not appropriate to model the data distribution with a normal curve.**



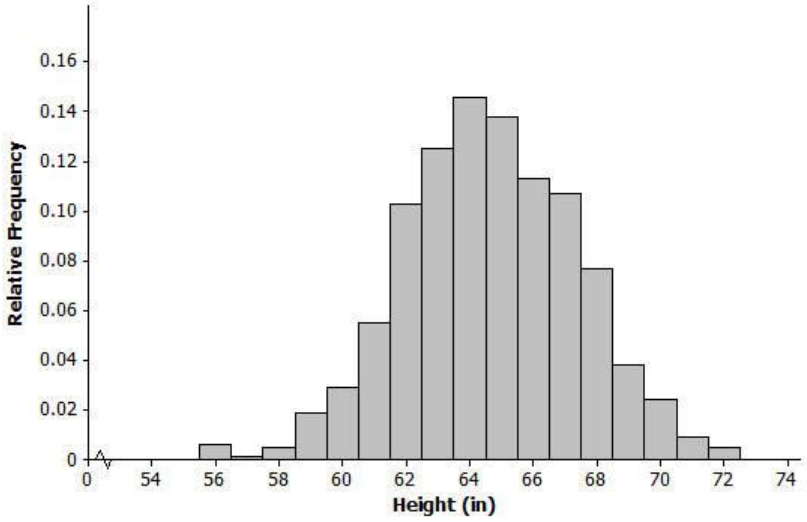
Exit Ticket (8 minutes)

Name _____ Date _____

Lesson 9: Using a Curve to Model a Data Distribution

Exit Ticket

The histogram below shows the distribution of heights (to the nearest inch) of 1,000 young women.



1. What does the width of each bar represent? What does the height of each bar represent?

2. The mean of the distribution of women's heights is 64.6 in., and the standard deviation is 2.75 in. Interpret the mean and standard deviation in this context.

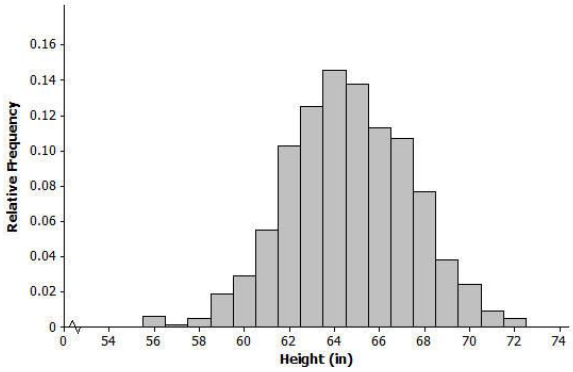
3. Mark the mean on the graph, and mark *one* deviation above and below the mean. Approximately what proportion of the values in this data set are within one standard deviation of the mean?

4. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Describe the shape of the distribution.

5. Shade the area of the histogram that represents the proportion of heights that are within one standard deviation of the mean.

Exit Ticket Sample Solutions

The histogram below shows the distribution of heights (to the nearest inch) of $1,000$ young women.



1. **What does the width of each bar represent? What does the height of each bar represent?**

   Each bar represents a 1-inch range of heights. For example, the bar above 56 represents heights between 55.50 and 56.49 inches. The height of each bar represents the proportion of the $1,000$ women in that 1-inch height range.

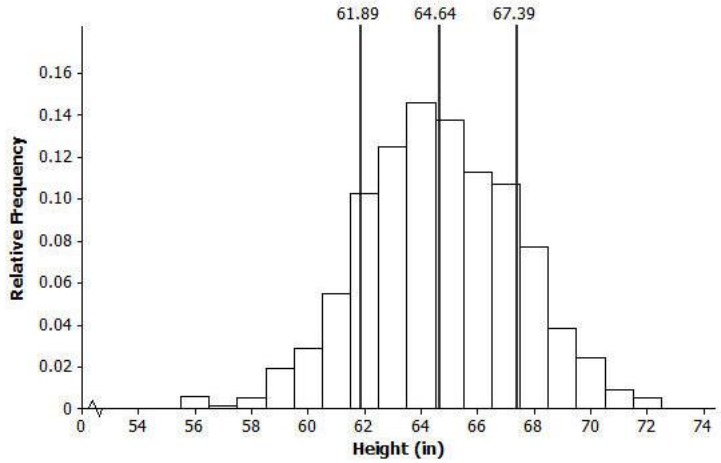
2. **The mean of the distribution of women's heights is 64.6 in., and the standard deviation is 2.75 in. Interpret the mean and standard deviation in this context.**

   The mean is the average height of the $1,000$ women, and it can be interpreted as a typical height value.

   The standard deviation is the typical number of inches that a woman's height is from the mean.

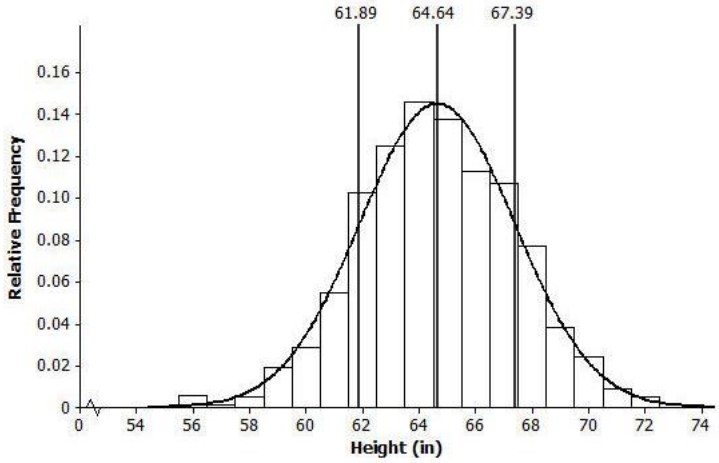
3. **Mark the mean on the graph, and mark *one* deviation above and below the mean. Approximately what proportion of the values in this data set are within one standard deviation of the mean?**

   $0.10 + 0.12 + 0.14 + 0.13 + 0.11 + 0.105 = 0.705 \approx 0.71$

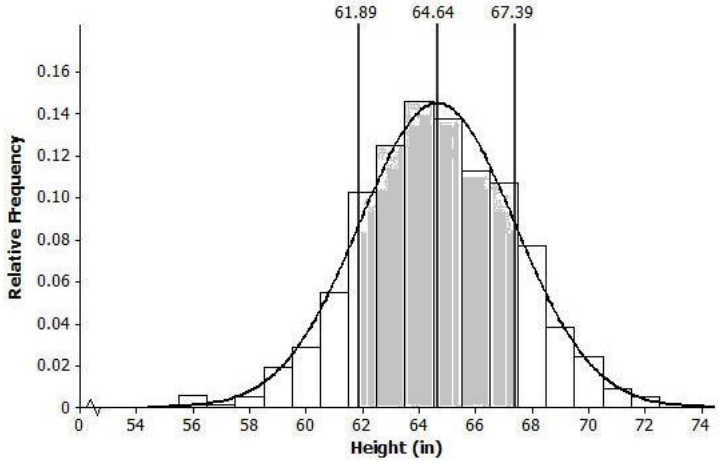


4. **Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Describe the shape of the distribution.**

Approximately normal (bell shaped and approximately symmetric)

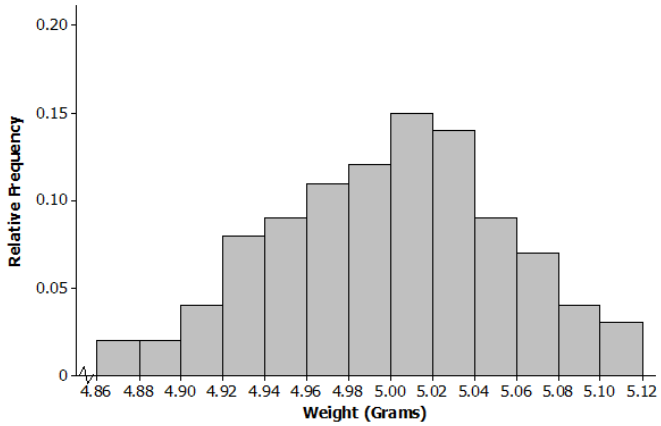


5. **Shade the area of the histogram that represents the proportion of heights that are within one standard deviation of the mean.**

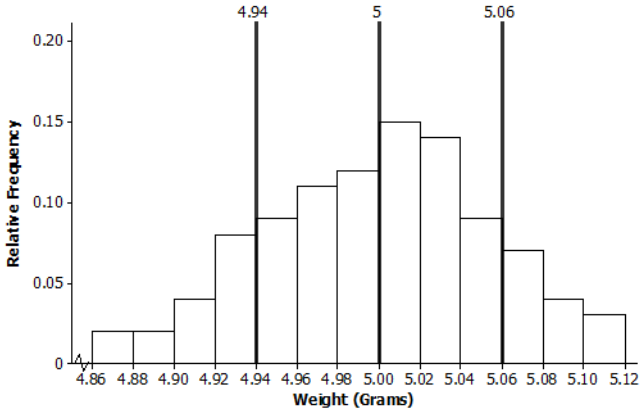


Problem Set Sample Solutions

1. **Periodically the U.S. Mint checks the weight of newly minted nickels. Below is a histogram of the weights (in grams) of a random sample of 100 new nickels.**

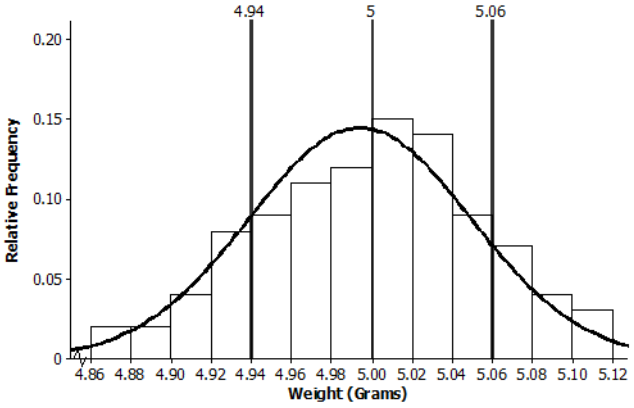


a. **The mean and standard deviation of the distribution of nickel weights are 5.00 grams and 0.06 gram, respectively. Mark the mean on the histogram. Mark one standard deviation above the mean and one standard deviation below the mean.**



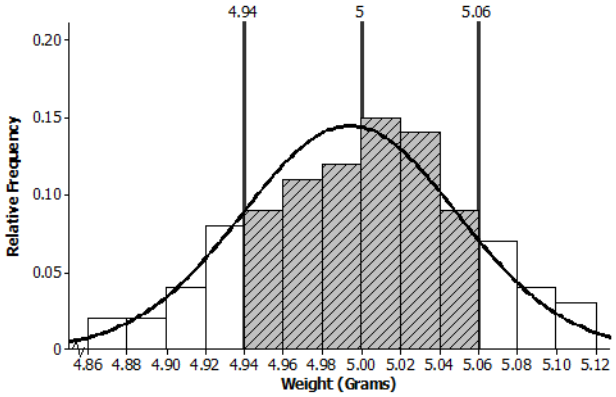
b. Describe the shape of the distribution. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Is this approximately a normal curve?

The shape is approximately normal.

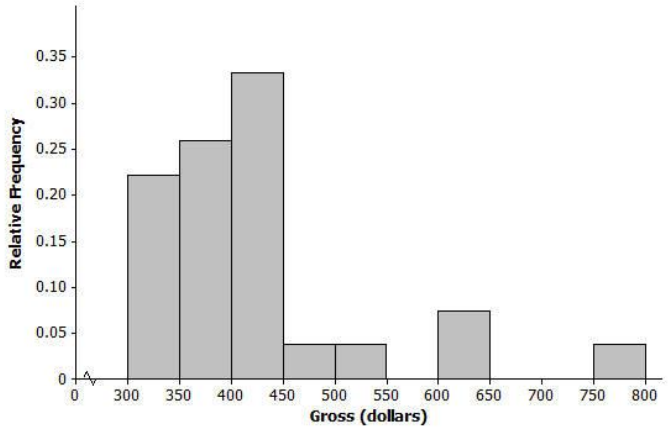


c. Shade the area of the histogram that represents the proportion of weights within one standard deviation of the mean. Find the proportion of the data within one standard deviation of the mean.

$0.09 + 0.11 + 0.12 + 0.15 + 0.14 + 0.09 = 0.70$



2. Below is a relative frequency histogram of the gross (in millions of dollars) for the all-time top-grossing American movies (as of the end of 2012). *Gross* is the total amount of money made before subtracting out expenses, like advertising costs and actors' salaries.



a. Describe the shape of the distribution of all-time top-grossing movies. Would a normal curve be the best curve to model this distribution? Explain your answer.

The shape is skewed to the right. A normal curve would not be the best curve to model the distribution.

b. Which of the following is a reasonable estimate for the mean of the distribution? Explain your choice.

i. 325 million
ii. 375 million
iii. 425 million

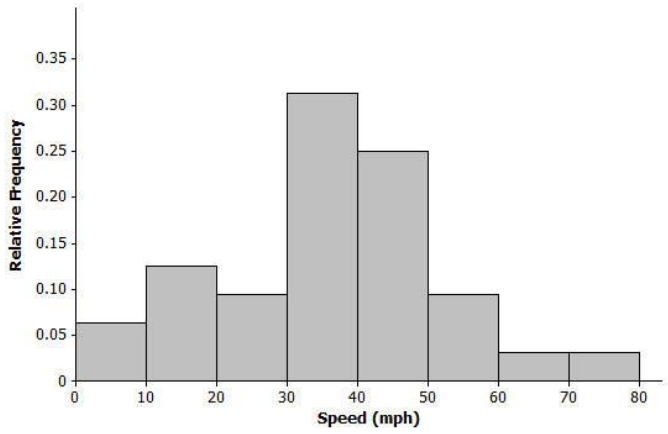
(iii) 425 million is a reasonable estimate since this is a skewed distribution, and the mean will be pulled toward the outliers.

c. Which of the following is a reasonable estimate for the sample standard deviation? Explain your choice.

i. 50 million
ii. 100 million
iii. 200 million

(ii) 100 million is a reasonable estimate because 50 million is too small, and 200 million is too large to be considered a typical deviation from the mean.

3. Below is a histogram of the top speed of different types of animals.



a. Describe the shape of the top speed distribution.

Approximately normal

b. Estimate the mean and standard deviation of this distribution. Describe how you made your estimate.

Mean is approximately 40 mph, and standard deviation is about 15 mph. Answers will vary in terms of how the estimate was made.

Note: Students could roughly estimate the mean by locating a balance point of the distribution. The standard deviation could be estimated by a typical deviation from the mean that was developed in Lesson 8.

c. Draw a smooth curve that is approximately a normal curve. The actual mean and standard deviation of this data set are 34.1 mph and 15.3 mph, respectively. Shade the area of the histogram that represents the proportion of speeds that are within one standard deviation of the mean.