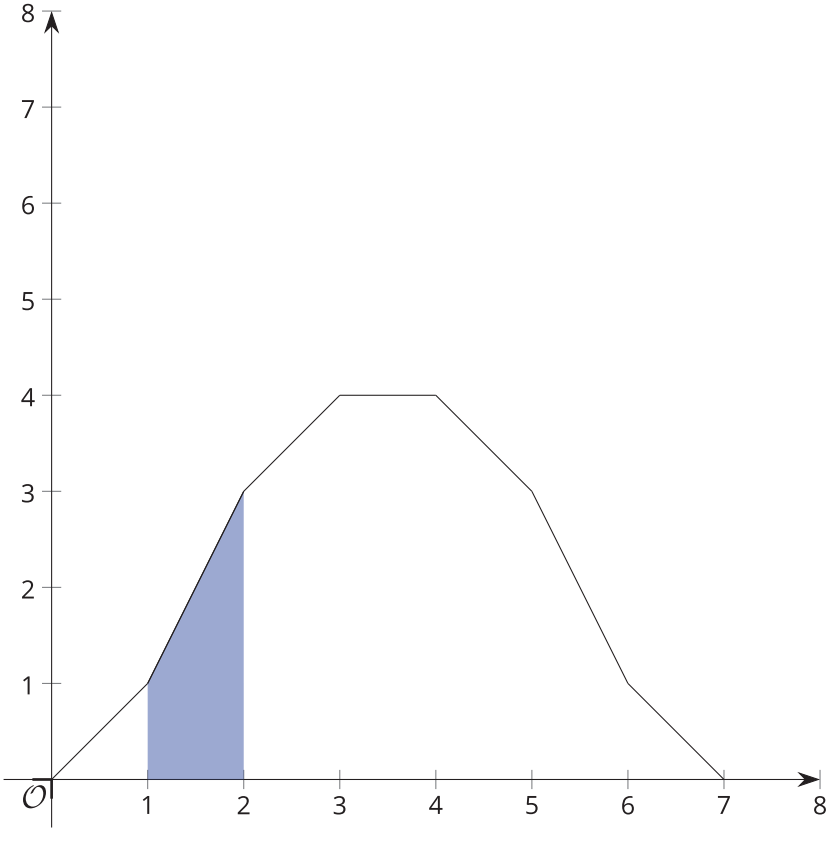
Lesson 6: Areas in Histograms

- Let's find proportions of data in certain intervals.

6.1: Find the Area



1. Find the shaded area between the function, the x-axis, and the boundaries $x = 1$ and $x = 2$. Explain or show your reasoning.

2. What proportion of the area between the function, the x-axis, and the boundaries $x = 0$ and $x = 7$ is shaded? Explain or show your reasoning.

6.2: Story Submissions

A publisher takes submissions for short stories to include in a book. 200 stories are submitted, but the publisher needs to be aware of how long each story is. The way the publisher will put together the collection of stories, a page typically contains 200 words. The mean number of words for each story is 2,600 and the standard deviation is 400 words.

1. If a histogram is created using intervals of 200 words, what would be the area of the bar representing the number of stories that contain between 2,000 and 2,200 words? Explain or show your reasoning.

2. What proportion of the total area is represented by the bar for stories that contain between 2,000 and 2,200 words? Explain or show your reasoning.

3. What proportion of stories in this group contain between 2,000 and 2,200 words? Explain or show your reasoning.

4. How does the proportion of the area you calculated relate to the proportion of stories in the group that contain between 2,000 and 2,200 words?

5. What proportion of stories in this group are within 1 standard deviation of the mean number of words?

6. What proportion of stories in this group are within 2 standard deviations of the mean number of words?

7. What proportion of stories in this group are within 1 standard deviation of 2,400 words?

Are you ready for more?

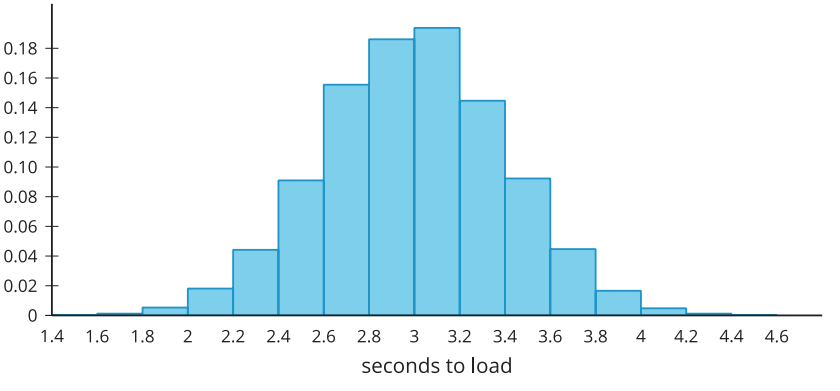
Prove more generally that the proportion of total area taken up by a bar in a histogram is equal to the proportion of all data values that are contained in the interval represented by bar. To begin, let n represent the number of data values in an interval given by one bar, M represent the number of data values in the entire set, and w be the width of the interval in each bar of the histogram. Prove that the proportion of area taken up by the bar is $\frac{n}{M}$.

6.3: Website Load Times

A company collects data from 10,000 websites about how long it takes to load the site. The number of seconds it takes to fully load the website is summarized in the relative frequency table.

seconds to load	relative frequency
1.4–1.6	0.0003
1.6–1.8	0.0012
1.8–2.0	0.0053
2.0–2.2	0.0181
2.2–2.4	0.0442
2.4–2.6	0.0910
2.6–2.8	0.1555
2.8–3.0	0.1861
3.0–3.2	0.1938
3.2–3.4	0.1447
3.4–3.6	0.0923
3.6–3.8	0.0447
3.8–4.0	0.0166
4.0–4.2	0.0048
4.2–4.4	0.0012
4.4–4.6	0.0002

The relative frequency histogram summarizes the same data.



The mean time to load a website is 3 seconds and the standard deviation is 0.4 seconds.

1. Would a normal distribution be a good model for this distribution? Explain your reasoning.

2. What proportion of websites loaded within 1 standard deviation of the mean?

3. What proportion of websites loaded within 2 standard deviations of the mean?

4. What proportion of websites loaded within 1 standard deviation of 2.8 seconds?

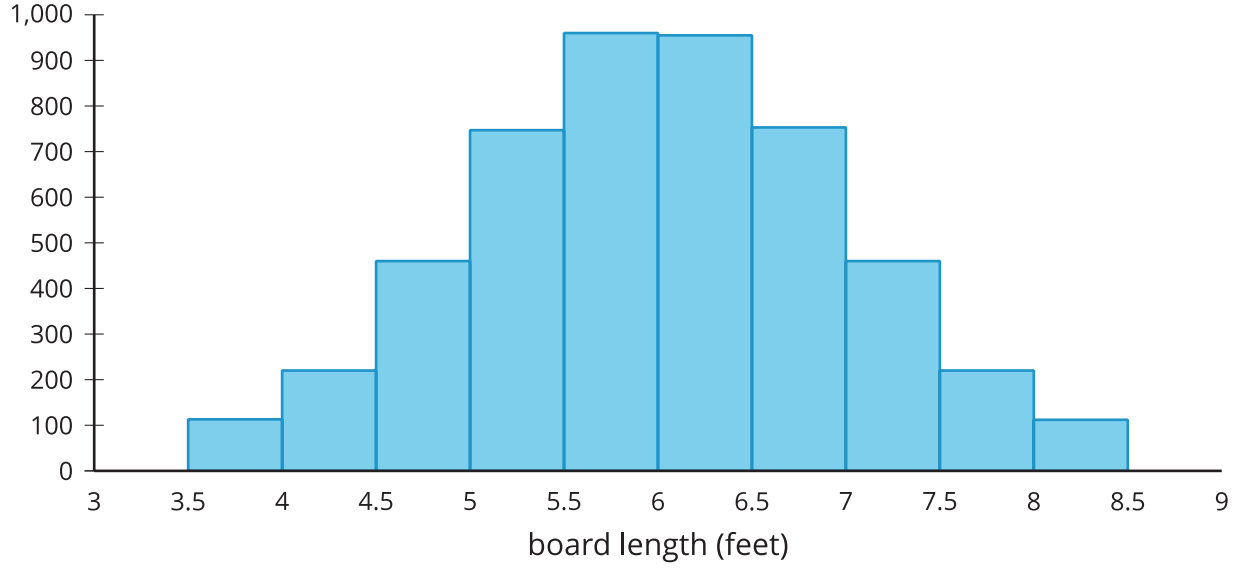
5. Compare the proportion of websites within 1 standard deviation of the mean to the proportion of stories in the submissions that are within 1 standard deviation of the mean number of words from the previous task. Do the same for the proportion within 2 standard deviations.

Lesson 6 Summary

There is an important connection between areas in histograms and the data represented by the histogram. In particular, the proportion of the total area in the histogram that is represented by a single bar in the histogram is equivalent to the proportion of all the data that is included in that interval. This is made more interesting by the fact that, for normally distributed data, the proportion of values in an interval whose endpoints are described by the mean and standard deviation is always the same.

For example, a woodshop produces boards of various lengths. During a certain week, 5,000 boards are produced and measured. The mean length is 6 feet, and the standard deviation length is 1 foot. The table and histogram show a summary of the board lengths.

board length	3.5–4	4–4.5	4.5–5	5–5.5	5.5–6	6–6.5	6.5–7	7–7.5	7.5–8	8–8.5
frequency	113	220	460	747	960	955	753	460	220	112



The total area of all the rectangles in the histogram is 2,500 since we could stack all the bars on top of one another and have a rectangle that is 5,000 tall and 0.5 wide. If we look at just the rectangles representing boards between 5.5 and 6 feet wide, the area is 480, which is 19.2% of the total area since $\frac{480}{2{,}500} = 0.192$. Similarly, we can see from the data that 19.2% of the data is in this same interval since $\frac{960}{5{,}000} = 0.192$. It is not a coincidence that these values are the same! The proportion of the total area that is in one of the rectangles is always equivalent to the proportion of all the data values that are in the same interval.

When the data is normally distributed, the proportions of certain regions are always the same. For example, there is always about 68% of the data within one standard deviation of the mean. Since the boards produced by the woodshop are approximately normal, we can test this information.

The boards within one standard deviation of the mean are between 5 and 7 feet long. Using the table, we can see that 3,415 boards are in this range ($747 + 960 + 955 + 753 = 3{,}415$) and those represent 68.3% ($\frac{3{,}415}{5{,}000} = 0.683$) of the boards produced in the woodshop.

Let's say that, another week, the woodshop produces 5,000 boards again, but this time, the mean is 6.5 feet and the standard deviation is 0.75 feet. As long as the board lengths continue to be approximately normal, we can expect about 68% of the boards to be within 1 standard deviation of the mean. For that week, it means that about 68% of the boards will be between 5.75 and 7.25 feet long.

In fact, as long as the interval can be described using only the mean and standard deviation and the data is normally distributed, the proportion of data values in the interval can be found. In general, about 68% of the data is within 1 standard deviation of the mean, about 95% of the data is within 2 standard deviations of the mean, and more than 99% of the data is within 3 standard deviations of the mean.