3.1 Graphical Displays of Information

- With all the data available today, it's important to be able to filter out the unnecessary data and to present the useful data in an accessible format

- One of the best formats for presenting data is a graph or other visual display

- We examined many different graphical displays in Chapter 1

  Bar Graphs, Histograms, Frequency Tables, Pictographs, Box-and-Whisker Plots, etc.

- We will examine the display of data further in this chapter along with methods for analyzing data involving one variable.
Histograms

A histogram is a frequency distribution where the horizontal access is divided into equal class intervals in to which data have been divided. The heights of the rectangles (that have no spaces between them) represent the frequencies associated with the corresponding intervals. It is important that each interval have the same width. Histograms are typically used for continuous data.

For example: height of students, speed of a car, weight of a person, time to wake up in the morning.

Continuous Variable: a variable or data that can have an infinite number of possible values in a given interval. A measure of quantity will always be continuous.

How to Make a Histogram

1. Choose the number of intervals (if the question doesn’t specify, choose between 5 and 10)
2. Calculate the range of your data (largest data point – smallest data point)
3. Round your range UP to a number that is easily divided by the number of intervals you chose.
4. Calculate your bin width \( \text{bin width} = \frac{\text{range}}{\text{number of intervals}} \)
5. Determine the first value for your first interval. \( \text{lowest value} - \frac{(\text{rounded range} - \text{actual range})}{2} \)
6. If any data points fall on the border of any of the intervals, add a decimal place to ensure that this doesn’t happen.
7. Make a frequency table using the intervals you have determined.
8. Draw the histogram (no spaces between bars)
Example 1

The heights of two high school classes were measured in centimetres. The results were as follows:


a) Determine the range of the data

\[ \text{Range} = \text{highest value} - \text{lowest value} \]

\[ = 193 - 120 \]

\[ = 73 \]

b) Determine an appropriate bin (interval) width that will divide the data into 8 intervals.

\[ \text{bin width} = \frac{\text{range}}{\text{number of intervals}} \]

\[ = \frac{80}{8} \]

\[ = 10 \]

Note:
Round your range UP to a value that can be divided easily.
c) Determine the first value of your first interval

We added \( \frac{7}{3} \) to 73 when we rounded our range, therefore we should subtract \( 3.5 \) from our smallest value 120; which makes our starting point 116.5.

or just use the formula:

\[
\text{Initial value} = \left[ \text{lowest value} - \frac{\text{rounded range} - \text{actual range}}{2} \right]
\]

\[
= \left[ 120 - \frac{80 - 73}{2} \right]
\]

\[= 116.5\]

Note:
1. If you have rounded your range up you should subtract half of the amount you rounded from the smallest value to evenly distribute the 'excess of your range'.
2. Make sure no data points lie on the border of two intervals. (Do this by subtracting .5 from a whole number, .05 from data with one decimal point, .005 from data with two decimal points and so on)

d) Create a frequency table using your intervals

<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>116.5-126.5</td>
<td>3</td>
</tr>
<tr>
<td>126.5-136.5</td>
<td>3</td>
</tr>
<tr>
<td>136.5-146.5</td>
<td>1</td>
</tr>
<tr>
<td>146.5-156.5</td>
<td>7</td>
</tr>
<tr>
<td>156.5-166.5</td>
<td>7</td>
</tr>
<tr>
<td>166.5-176.5</td>
<td>9</td>
</tr>
<tr>
<td>176.5-186.5</td>
<td>15</td>
</tr>
<tr>
<td>186.5-196.5</td>
<td>5</td>
</tr>
</tbody>
</table>

Notice that the number one interval ends with, the next interval starts with the same number. This is because the data for a histogram is continuous!!!
Example 2

Earthquakes are measured on a scale known as the Richter Scale. There data are a sample of earthquake magnitudes in Canada between 1960 and 1965.

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>5.0</td>
<td>6.0</td>
</tr>
<tr>
<td>5.6</td>
<td>6.5</td>
</tr>
<tr>
<td>6.5</td>
<td>6.5</td>
</tr>
<tr>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>6.4</td>
<td>7.2</td>
</tr>
<tr>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>5.2</td>
<td>4.6</td>
</tr>
<tr>
<td>6.3</td>
<td>7.2</td>
</tr>
<tr>
<td>6.0</td>
<td>5.4</td>
</tr>
<tr>
<td>6.0</td>
<td>5.6</td>
</tr>
<tr>
<td>5.6</td>
<td>6.2</td>
</tr>
</tbody>
</table>
What is wrong with this histogram of the data?

The width of each bar (bin width) is too small. Too small of a bin width results in a histogram that does not effectively summarize the distribution (too many small bars). Between 5 and 10 intervals will usually produce a good bin width of the data.

What is wrong with this histogram of the data?

Pieces of data fall on the border of two intervals. This makes it ambiguous as to where the data actually falls. Make sure to take the necessary steps when creating your intervals to make sure that this doesn't happen.
Let's make an effective histogram for the data:

**a) Determine the range of the data**

\[ \text{range} = \text{highest value} - \text{lowest value} \]
\[ = 7.2 - 4.6 \]
\[ = 2.6 \]

**b) Determine an appropriate bin (interval) width that will divide the data into 6 intervals.**

\[ \text{bin width} = \frac{\text{range}}{\text{number of intervals}} \]
\[ = \frac{3.0}{6} \]
\[ = 0.5 \]

**c) Determine the first value of your first interval**

We added 0.4 to 2.6 when we rounded our range, therefore we should subtract 0.2 from our smallest value 4.6; which makes our starting point 4.4.

However, some data will still fall on the border of the intervals, so we should add a decimal place by subtracting .05 from our starting point.
d) Create a frequency table using your intervals

Notice that the number one interval ends with, the next interval starts with the same number. This is because the data for a histogram is continuous!!

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.35 - 4.85</td>
<td>1</td>
</tr>
<tr>
<td>4.85 - 5.35</td>
<td>13</td>
</tr>
<tr>
<td>5.35 - 5.85</td>
<td>12</td>
</tr>
<tr>
<td>5.85 - 6.35</td>
<td>6</td>
</tr>
<tr>
<td>6.35 - 6.85</td>
<td>5</td>
</tr>
<tr>
<td>6.85 - 7.35</td>
<td>3</td>
</tr>
</tbody>
</table>

e) Create a histogram of the data

Magnitude of Canadian Earthquakes (1960-1965)
Distributions of Data

A frequency distribution is categorized by the general shape of the corresponding histogram.

Typically is it described one of four ways:

1. U-shaped
2. Uniform
3. Mound-shaped
4. Skewed

Symmetric distributions

Frequency Distribution: A set of values of a variable, together with the frequency of each value.

U-Shaped Distribution

The scores from the game of spider solitaire form this type of distribution

- a U-shaped distribution occurs when there are peaks at either end of the range
- it may also be described as a bimodal distribution

bimodal: a distribution that has two peaks
What is another example of a frequency distribution that would be U-shaped (bimodal)?

Marks on a test where half the class just "didn't get it".

**Uniform Distribution**

This is the distribution you would expect from an experiment such as rolling a die.

- When each outcome has a similar frequency, it is called a uniform distribution.
- The height of each bar is roughly equal.
Mound Shaped Distribution

Rolling a pair of dice and recording the sum results in this type of distribution.

- In this distribution, there is an interval with the greatest frequency, and the frequencies of all other intervals decrease on either side of that.
- The frequency distribution takes on a mound (or bell) shape.

What is similar about these distributions?

**symmetric distributions:**
when the data show a mirror symmetry about the centre.
Skewed Distributions

Scores from a game of solitaire produce a distribution that is skewed to the **right**.

In a skewed distribution, the interval or group of intervals that contains the greatest frequencies is near one end of the histogram. As a result, these distributions seem to tail off to the left or right.

<table>
<thead>
<tr>
<th>Score</th>
<th>0-19</th>
<th>20-39</th>
<th>40-59</th>
<th>60-79</th>
<th>80-99</th>
<th>100-119</th>
<th>120-139</th>
<th>140-159</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>20</td>
<td>15</td>
<td>12</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Right Skewed**

The mean is skewed to the right

mode $<$ median $<$ mean

**Left Skewed**

The mean is skewed to the left

mean $<$ median $<$ mode