Section 5.2
Frequency Tables, Histograms, and Frequency Polygons

Learning targets:

1. Demonstrate understanding of new terminology.
2. Organize raw data into a frequency table.
3. Use data from a frequency table to create a histogram.
4. Use a histogram to create a frequency polygon.

Frequency Distributions

- A frequency distribution is a way of taking a large set of data and breaking it down into smaller groups, or intervals, of equal width.
- Data values that fall within each interval are then counted, giving us the frequency at which the data occurs in each interval.
- Frequency distributions can be in table format, or may come in a graphical representation.

More Terminology:

- **Frequency table:** a frequency distribution for a set of data displayed in a table format.

- **Histogram:** a frequency distribution for a set of data displayed in a special type of bar graph format.

- **Frequency polygon**: a line graph which is produced by joining the midpoints of the intervals of a histogram using straight lines.

Guidelines: In a frequency distribution,

1. You should have somewhere between 5 and 12 intervals.

2. How wide each interval should be will vary with the different data sets.

   If you know how many intervals you want , you can estimate the width of each interval as follows:

   range ÷ how many intervals you want

   (You will sometimes have to round this value up to make it easier to work with and to make sure all your data will fit in the number of intervals you have chosen).

   Alternatively, if you know what interval width you'd like to work with, you can calculate the number of intervals you'd need as follows:

   range ÷ interval width

   (You always have to round this value up to the next whole number to ensure all of your data will fit).

4. Your first interval should start a value slightly smaller than the smallest value in the data set.

5. Your last interval doesn't necessarily have to end with the largest value in the data set; it can end with a larger value.

Range : $3.3 - 1.1 = 2.2$

$2.2 \div 6 = .3636$

$\approx .4$

Example 1:

A Macintosh apple orchard has 40 trees with these heights, given in metres:

1.1	1.3	1.4	1.2
1.5	1.7	1.6	1.3
1.5	2.0	2.1	1.8
1.9	2.3	2.2	2.1
1.7	2.0	2.2	2.5
2.3	2.4	1.9	1.8
3.1	3.2	3.3	2.7
2.8	2.6	2.5	2.3
3.0	2.4	2.7	2.4
2.6	2.8	2.2	2.1

Organize the raw data into a frequency table.

Range = max – min = 3.3 – 1.1 = 2.2

Let's use 6 intervals.

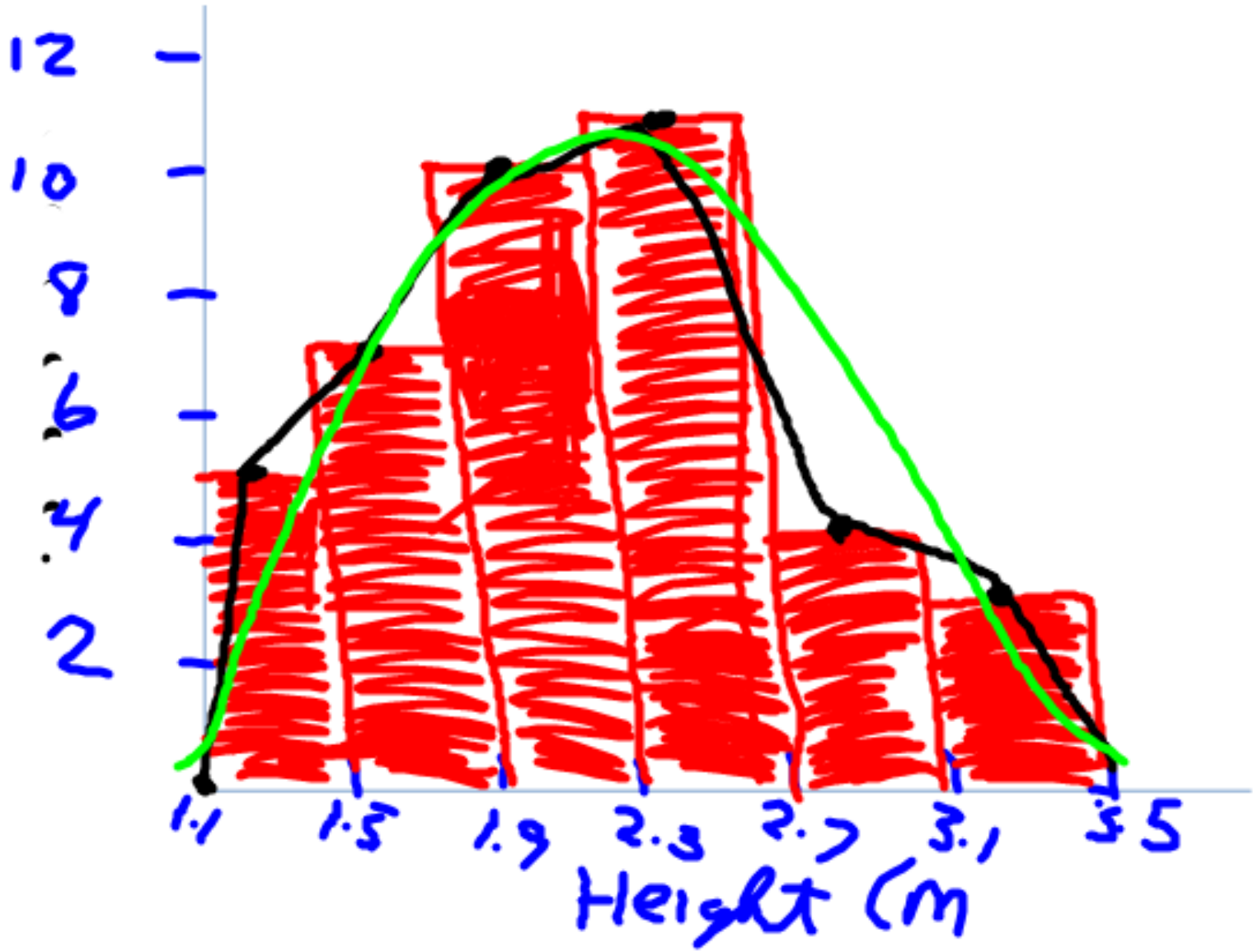
2.2 ÷ 6 = 0.366666... We can round this up to 0.4

Height of Trees (m)	Tally	Frequency
1.1 – 1.5	𝍸	5
1.5 – 1.9	𝍸 ||	7
1.9 – 2.3	𝍸 𝍸	10
2.3 – 2.7	𝍸 𝍸 |	11
2.7 – 3.1	||||	4
3.1 – 3.5	|||	3

-
-
-

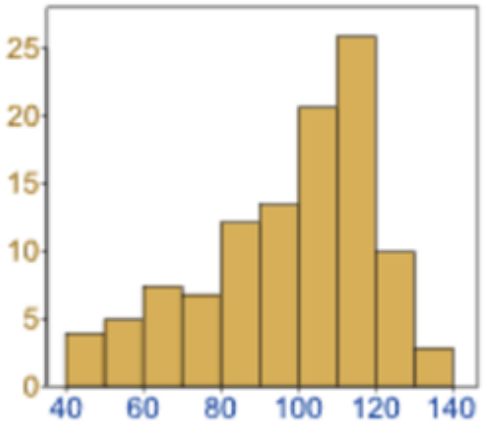
- *After you fill in the intervals in column one, double check to make sure that every value from the data set has an interval to be assigned to. If not, you will have to either adjust the width or add another interval.*

- Use the frequency table to create a histogram.
- Draw the frequency polygon.

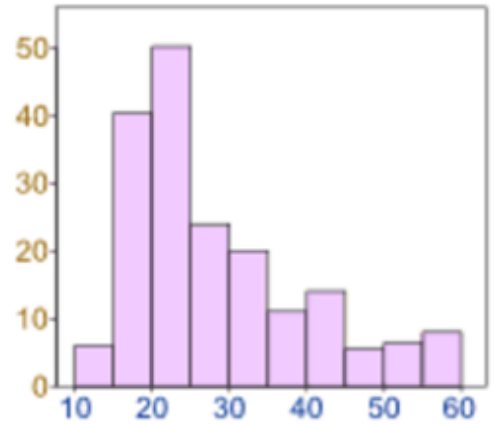


Data can be "distributed" (spread out) in different ways.

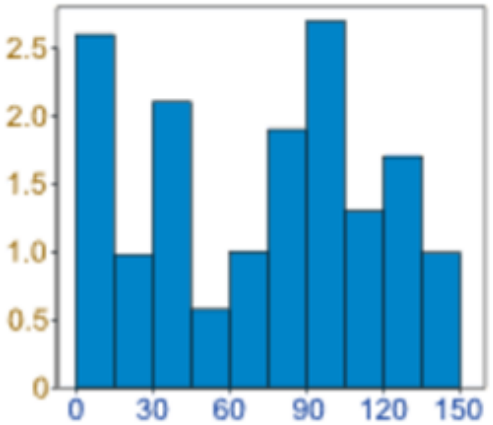
It can be spread out more on the left



Or more on the right



Or it can be all jumbled up

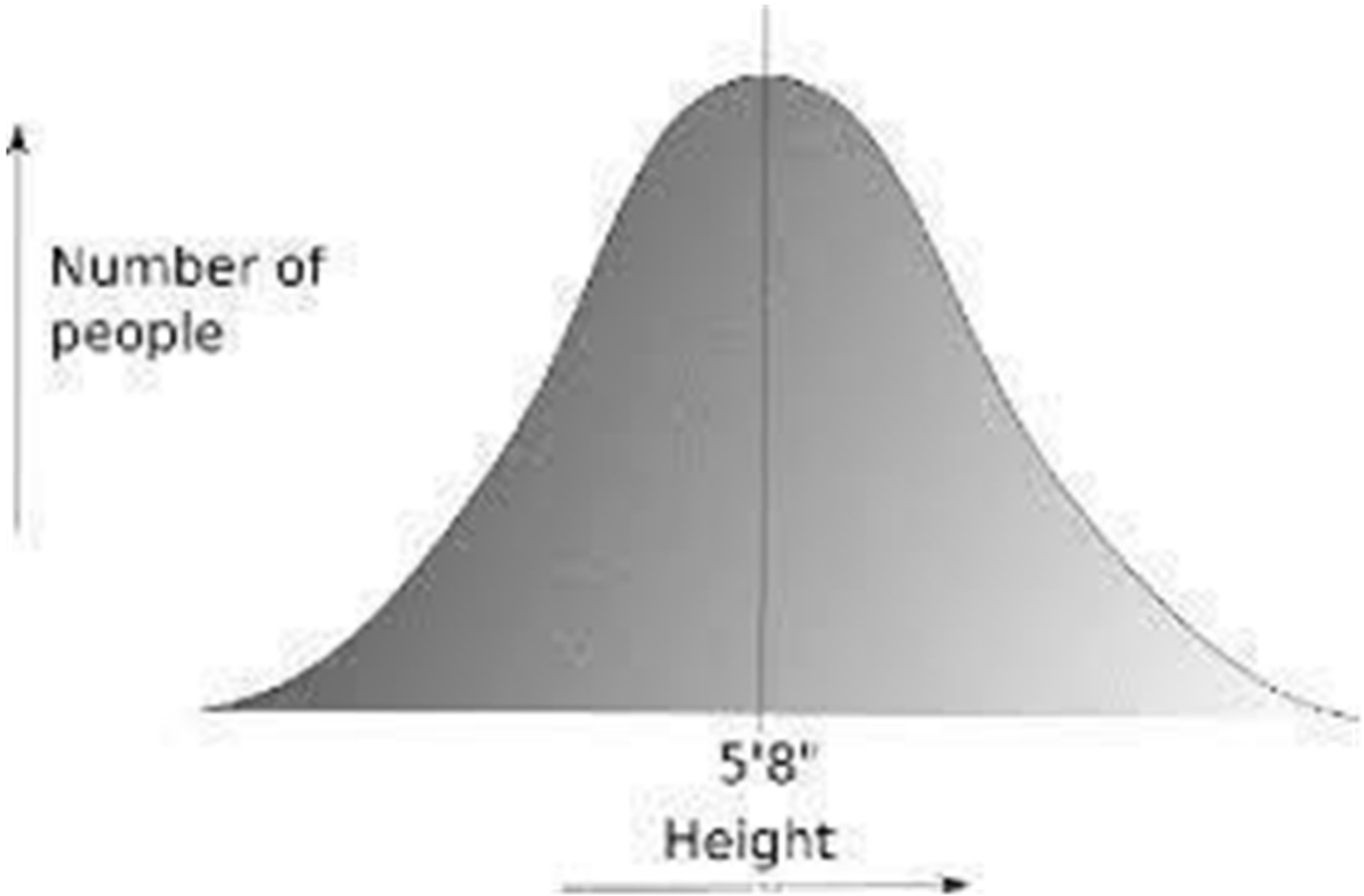


Imagine you measured the heights of 10,000 people. You can easily get the average height by adding all the heights together and dividing by 10,000. Say this came out as 5'8".

--> What proportion of the group were exactly 5'8" tall? How many were 5'7" or 6'2"?

So you divide all the heights into a histogram with intervals of 1 inch.

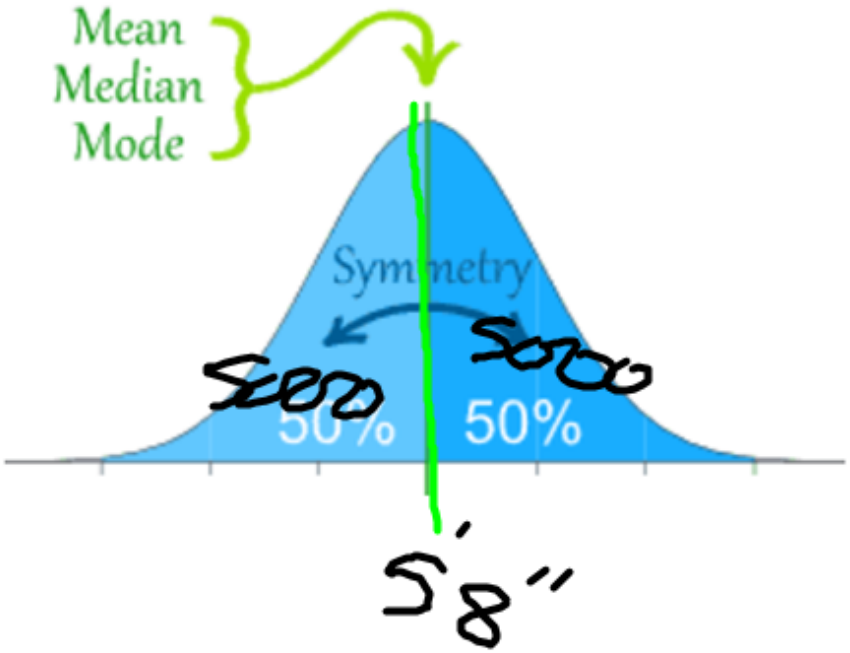
--> The histogram would look something like this:


Number of people
5'8"
Height

This shape is called a **NORMAL DISTRIBUTION**.

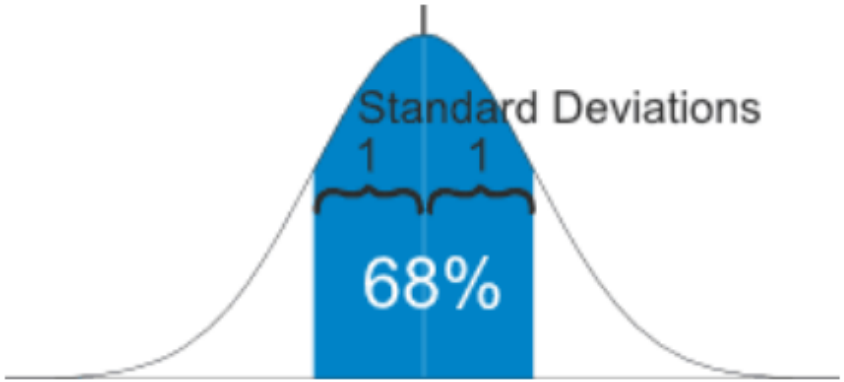
It also gets called the "**Bell Curve**" because it looks like the cross-section of a bell.

We say the data is "normally distributed":



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean



68% of values are within
1 standard deviation of the mean

95% of values are within
2 standard deviations of the mean

99.7% of values are within
3 standard deviations of the mean

Based on this we can now say that any value in a normally distributed set of data is:

- **likely** to be within 1 standard deviation (68 out of 100 should be)
- **very likely** to be within 2 standard deviations (95 out of 100 should be)
- **almost certainly** within 3 standard deviations (997 out of 1000 should be)

ASSIGNMENT:

Page 222
#'s 3,4