Summarizing Data: One Variable

The Shape of the Data

In this lesson we will introduce histograms, which are another way to display quantitative data. We will also learn some vocabulary that is used to describe the features of the data that we might see in the histogram.

The pattern of values of data, showing their frequency of occurrence relative to each other, is called the **distribution** of the data. A **histogram** is useful for visualising distributions.

Example 1
Let's start with a histogram of the life expectancies for the 197 countries and territories in our dataset.

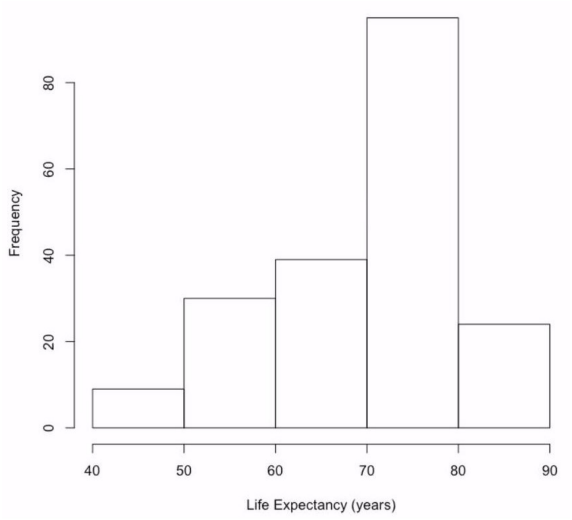

Figure 1: Histogram of the life expectancy data

The first step in constructing a histogram is to divide the data values into intervals or **bins** that are mutually exclusive. Since the life expectancies range from about 48 to 83 years we've defined our bins to capture every decade from 40 to 90. There are five bins in the

histogram and they are defined as:

> 40 and ≤ 50 years
> 50 and ≤ 60 years
> 60 and ≤ 70 years
> 70 and ≤ 80 years
> 80 and ≤ 90 years

The **cutpoints** are the values that define the beginning and the end of the bins. In this case, the cutpoints are 40, 50, 60, 70, 80, and 90. The vertical axis in our histogram is the **frequency** or count of the number of data values in each bin. For example, from our first bin, we know that there are nine countries or territories with life expectancies greater than 40 years and at most 50 years.

The width and number of bins in a histogram can be any convenient value. However, it is possible to dramatically change the appearance of a histogram with the choice of bin size, particularly when the number of data values is small. As an example, if we change the bins to cover only a two-year span, the life expectancy histogram appears to be more noisy, with more variation in the frequencies among bins.

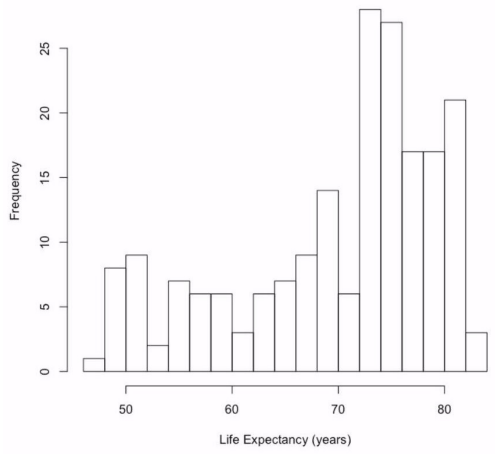

Figure 2: Histogram of the life expectancy data - small bin size

The **tails** of a histogram are the bars on the far left and right where the extremes of the data values are. In the histogram we see that the left tail of the life expectancy distribution is longer.

Example 2
Below are a few examples of the range of shapes of distributions we see in histograms. From a histogram, we can get lots of information about the data and we should check for features

such as the presence or absence of a peak. If there is one peak, the data value where this occurs is called the **mode**. A distribution can be **unimodal** (one peak), **bimodal** (two peaks), or **multimodal** (multiple peaks). If all data values occur about an equal number of times, the distribution does not have a mode and is said to be **uniform**.

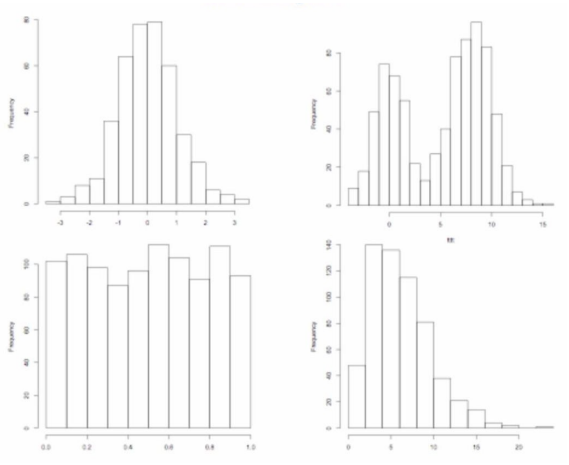

Figure 3: Various histogram shapes - unimodal, bimodal, uniform, right-skewed

From a histogram, we can also get an idea of the extent of spread in the data.

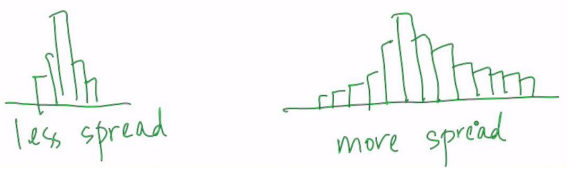

Figure 4: Histograms display the spread of the data

We can also see the extent of symmetry in the distribution of the data. Distributions can be symmetric or skewed. A **left-skewed** distribution has a long left tail. A **right-skewed** distribution has a long right tail.

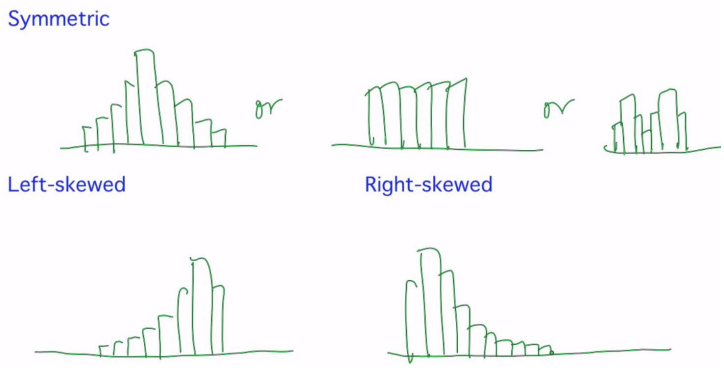

Figure 5: Histograms display the shape of the data

Histograms allows us to notice gaps in the data and **outliers** which are data values that are much larger or smaller than the rest of the data.

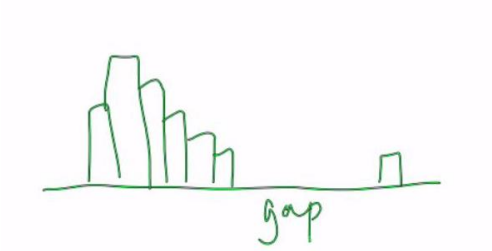

Figure 6: Histograms display the gaps in the data

Using the new terminology, we can now say that the distribution of the life expectancy data in Figure 1 is unimodal (mode between 70-80), left-skewed and without outliers. Some of these features can also be seen in the boxplot of life expectancies in Figure 7. Note that the median is in the right half of the box, and the left whisker is longer than the right whisker. This is typical of left-skewed distributions.

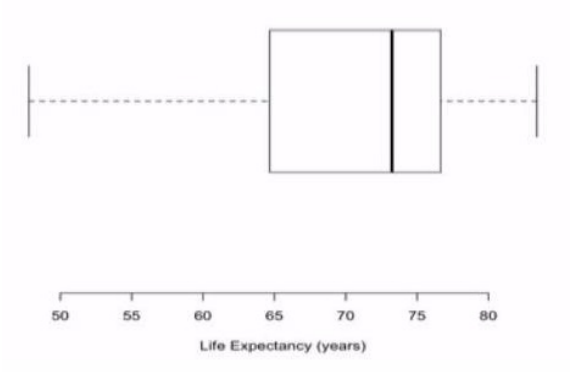

Figure 7: Boxplot of the data

We can also see the features of the life expectancy data in the five-number summary:

Minimum	1st Quartile	Median	3rd Quartile	Maximum
47.8	64.7	73.2	76.7	83.4

Table 1: Five number summary for life expectancy data

The difference between the third quartile and the median ($76.7 - 73.2 = 3.5$) is smaller than the difference between the median and the first quartile ($73.2 - 64.7 = 8.5$). The difference between the median and the maximum ($83.4 - 73.2 = 10.2$) is also smaller than the distance from the median to the minimum (73.2 - 47.8 = 25.4). Another feature typical of many left-skewed distributions is that mean is less than the median ($69.9 < 73.2$). This happens because the mean is pulled down by the values in the long left tail.

Example 3
Let's look at the shape for another set of data, the differences between the estimated ages using the Di Gangi method and the actual ages of death for our 400 skeletons. The histogram is shown in Figure 8. The distribution of these data is unimodal (mode is between -20 and -10), symmetric, and without outliers.

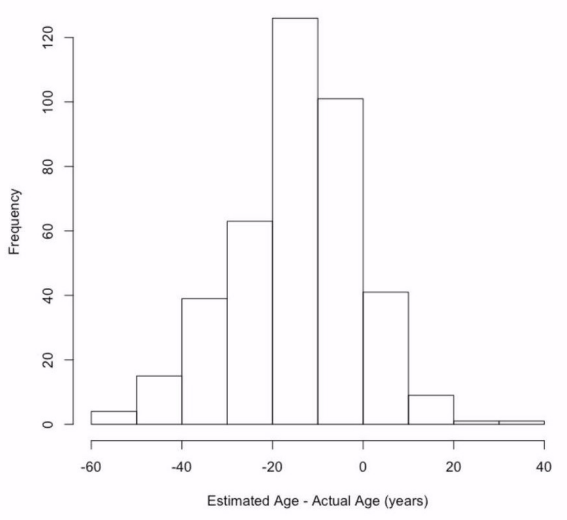

Figure 8: Histogram of estimated age − actual age for the skeleton data

Comparing the histogram with the modified boxplot in Figure 9, we see that the boxplot is also quite symmetric with the median near the middle of the box, and the whiskers of similar length. The data values indicated outside the fences are not outliers here, since they are not separated from the rest of the data.

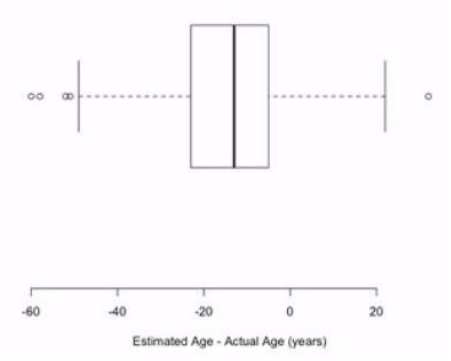

Figure 9: Boxplot of estimated age − actual age for the skeleton data

This symmetry can also be seen in the five-number summary in Table 2. The distance from the median to the first quartile $(-13-(-23)=10)$ is similar to the distance between the median and the third quartile $(-5-(-13)=8)$. Also, the distance from the median to lower fence $(-3-(-60)=47)$ is similar to the distance between the median and the upper fence $(32-(-13)=45)$. In addition, the mean (-14.2) and median (-13) are quite close to each other.

Minimum	1st Quartile	Median	3rd Quartile	Maximum
-60	-23	-13	-5	32

Table 2: Five number summary for skeleton data (Di Gangi estimated age - actual age)

Example 4
We can also look at the histogram for the 2012 salaries of the New York Red Bulls soccer team in Figure 10. This distribution is unimodal (mode is less than $100,000), right-skewed and includes 2 large outliers. The key feature of these data is the large gap between the salaries of the two highest-earning players and the 23 other players on the team.

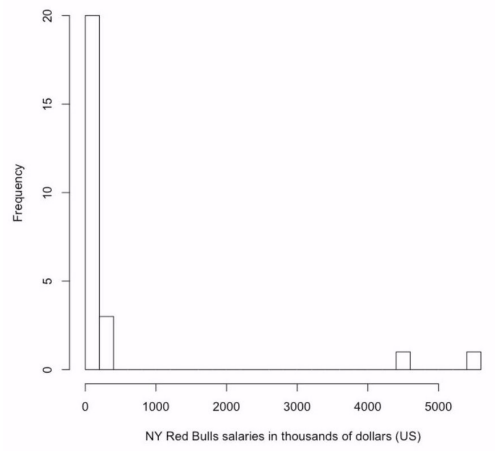

Figure 10: Histogram of the 2012 salaries of the New York Red Bulls

If we remove the two players with the largest salaries we can examine the distribution of salaries for the remaining 23 players (Figure 11).

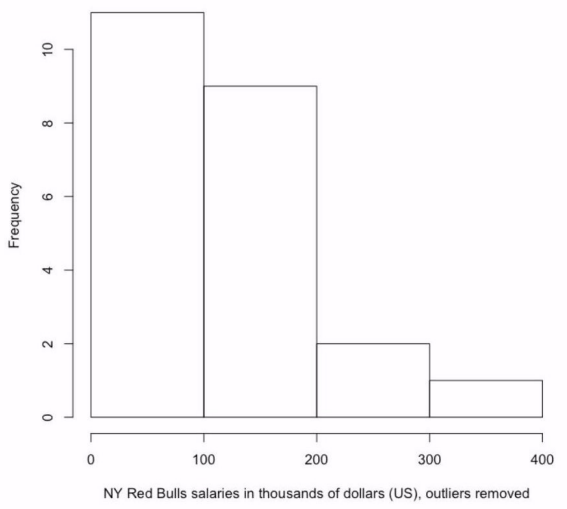

Figure 11: Histogram of the 2012 salaries of the New York Red Bulls (2 outliers removed)

The boxplot with the outliers excluded (Figure 12) shows the skew in the data. The median is slightly to the left of center of the box, and the right whisker is much longer than the left whisker.

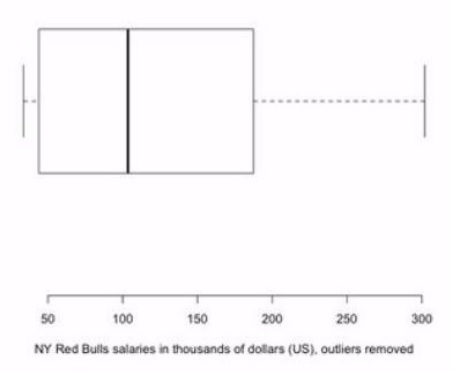

Figure 12: Boxplot of the 2012 salaries of the New York Red Bulls (2 outliers removed)

We can also see right-skew in the five number summary in Table 3. The distance from the median to the first quartile ($103500 - 44000 = 59500$) is less than the distance between the median and the third quartile ($187500 - 103500 = 84000$). Also, the distance from the median to lower fence ($103500 - 33750 = 69750$) is less than the distance between the median and the upper fence ($301999 - 103500 = 198499$). And the mean (119904) is larger than the median (103500).

Minimum	1st Quartile	Median	3rd Quartile	Maximum
33750	44000	103500	187500	30199

Table 3: Five number summary for NY Red Bulls 2012 salaries (2 outliers removed)

Summary of the relative location of the mean, median and mode in a unimodal distribution

For most distributions, if the distribution is

symmetric: the mean, median and mode are approximately the same,

left-skewed: mean $<$ median $<$ mode,

right-skewed: mode $<$ median $<$ mean.

As shown in Figure 13, for symmetric and skewed distributions we can see the overall features of the shape of the data in the histogram or boxplot. For a bimodal distribution, the boxplot fails to capture the two peaks and a histogram can be more informative.

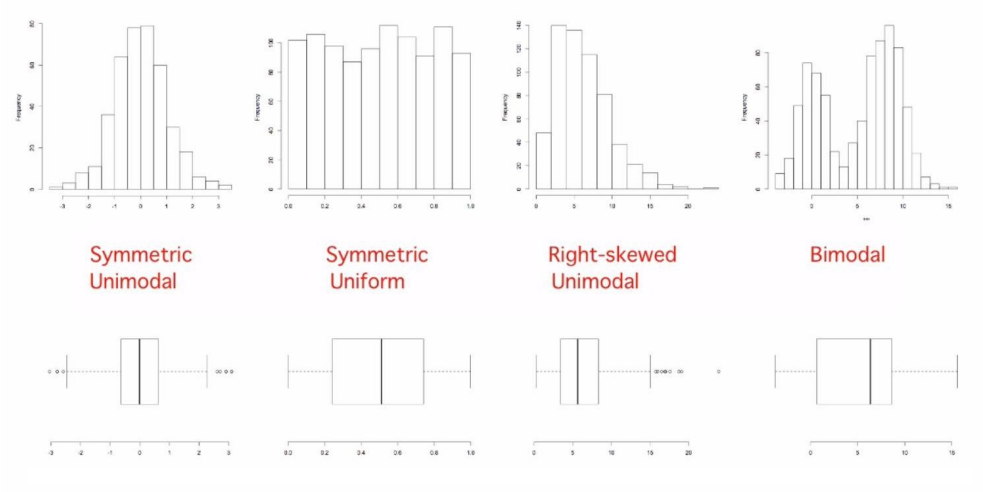

Figure 13: Comparison of histograms and boxplots for various shapes of data

Unimodal and symmetric distributions of data with histograms that have roughly a bell shape are very common. For data with this shape, the standard deviation is an important measure of spread or variability. The **empirical rule** tells us approximately how the frequency of data values is related to the standard deviation.

Empirical Rule:
68% of data values are within [mean − SD, mean + SD]
95% of data values are within [mean − 2×SD, mean + 2×SD]
99.7% of data values are within [mean − 3×SD, mean + 3×SD]

Example 5
As an example of how well the empirical rule works, we can look at the skeleton data. As was shown in Figure 8, the error in age estimation has a unimodal and symmetric distribution. For these data, the mean ($\bar{x}$) is -14.2 and the standard deviation (SD) is 14.1.

$$\bar{x} - 1 \times SD = -14.2 - 14.1 \sim -28 \qquad \bar{x} + 1 \times SD = -14.2 + 14.1 \sim 0$$

273 of the 400 data values or 68.3% are within the range $[-28, 0]$

$$\bar{x} - 2 \times SD = -14.2 - 2 \times 14.1 \sim -42 \qquad \bar{x} + 2 \times SD = -14.2 + 2 \times 14.1 \sim 14$$

380 of the 400 data values, or 95% are within the range $[-42, 14]$

$$\bar{x} - 3 \times SD = -14.2 - 3 \times 14.1 \sim -57 \qquad \bar{x} + 3 \times SD = -14.2 + 3 \times 14.1 \sim 28$$

397 of the 400 data values, or 99.3% are within the range $[-57, 28]$

We can see that the distribution of the skeleton data is very close to the 68, 95, 99.7% that we expected from the empirical rule.

Example 6
Although the Empirical Rule was derived from properties of symmetric and unimodal distributions, it works surprisingly well in other situations. Let's see how it works for the left-skewed life expectancy data (Figure 1). The mean for these data is 69.9 and the standard deviation is 9.7. It can be shown that

$128/197 = 65\%$ of data values are within 1 standard deviation of the mean
$186/197 = 94.4\%$ of data values are within 2 standard deviations of the mean
$197/197 = 100\%$ of data values are within 3 standard deviations of the mean

This example demonstrates that even for skewed distributions, we are very close to the 68, 95, 99.7% of the empirical rule.