Unit 2: Stemplots

SUMMARY OF VIDEO

Statistics is all about data. It is easy to get overwhelmed by an avalanche of numbers if we don't figure out good ways to organize it.

One of the best places to start is with a picture. You've seen charts similar to the ones in Figure 2.1 before – bar charts, pie charts, and dotplots – in this case, all ways to visualize the weight of newborn babies. Visualizing data like this can be a good first step toward organizing it and understanding it.

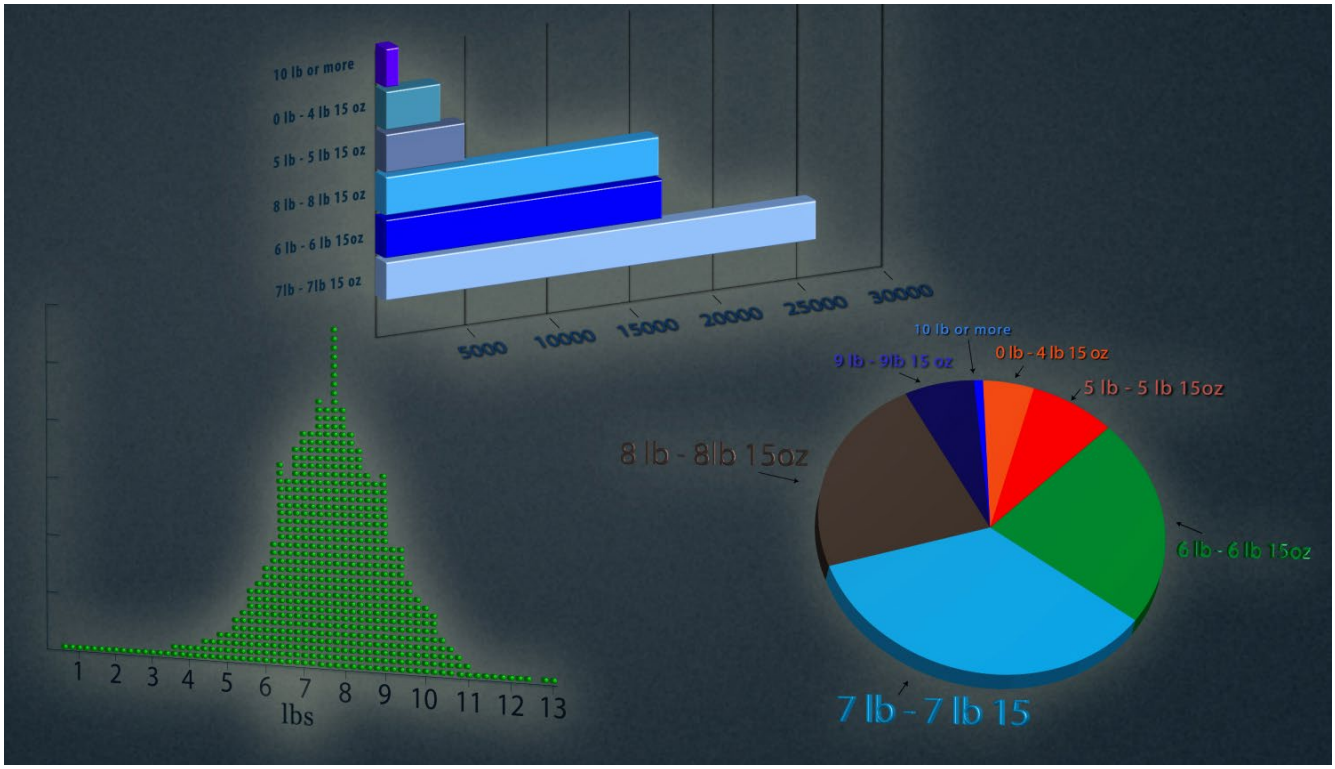

Figure 2.1. Graphic displays of weights of newborns.

In addition to the overall pattern displayed by the charts in Figure 2.1, the charts provide a framework to contextualize a particular baby's birth weight relative to the rest of the data. In other words, the charts can help us decide whether the baby was small, in the middle of the pack, or large compared to the other babies.

There are a variety of ways to visualize data and many real world datasets to work with. Let's step into the Army's boots to see the data it collected to help outfit each and every soldier with the right size uniform and gear. Soldiers' measurements have changed over the years – over time, soldiers' sizes have both increased and become more variable.

To better assess the outfitting needs of the soldiers, the Army periodically embarks on a measurement project in which many measurements – foot length, shoulder width, head size, and so forth – are taken on a large random sample of soldiers. With a better sense of the most frequently-found dimensions, the Army knows which sizes of uniforms to keep well-stocked, and which sizes are rare enough that it's cheaper to custom order them. As an illustration, here are the foot lengths (cm) of thirty soldiers:

27.2	26.9	26.6
28.0	26.8	26.1
26.2	27.3	27.6
25.7	29.0	26.5
32.8	28.8	26.9
25.0	26.7	24.6
26.3	26.8	27.0
28.0	27.3	26.5
27.4	25.0	26.6
25.8	27.0	25.9

When you see a bunch of unorganized numbers, it is hard to determine whether or not there are any important patterns. But if we organize these numbers into a stemplot, we can get a better sense of how widely foot size varied. First, using technology or a calculator, we can sort the foot sizes in order from smallest to largest. The sorted data already give us a little better sense of soldiers' foot sizes. The smallest is 24.6 centimeters and the largest is 32.8 centimeters.

Next, we separate each measurement into a stem (the first digits) and a leaf (the final digit). The stems are lined up vertically and then the leaves are filled in opposite the appropriate stems. Always include all possible stems in your data range, even those that don't have leaves to go with them. The final step is to organize the leaves in numerical order. The result is the stemplot in Figure 2.2.

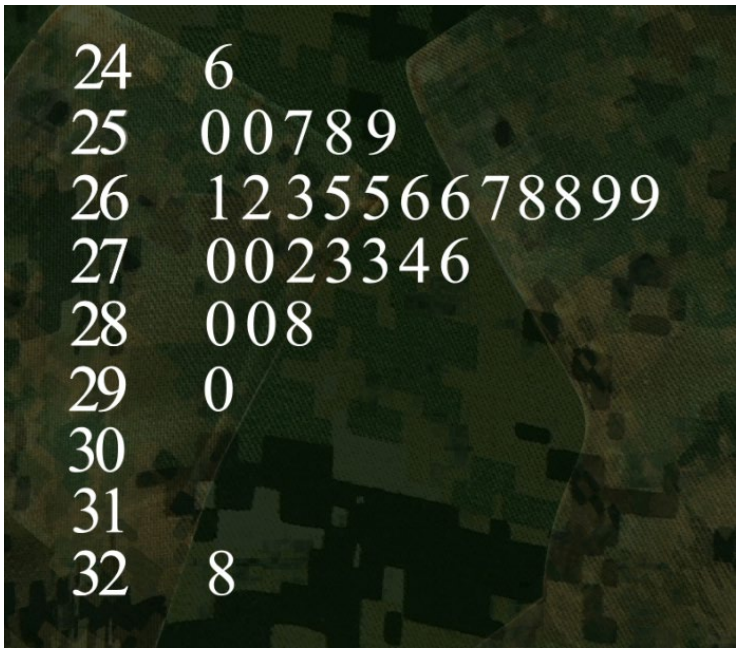

Figure 2.2. Stemplot of soldiers' foot lengths.

Displayed as a stemplot, we can see the overall pattern of our data: 26- centimeter values are the most common, and values on either side of that single peak are less common. A stemplot also lets you see at a glance how spread out the distribution is. The data points range from the smallest at 24.6 centimeters to the largest at 32.8 centimeters. Check out the overall shape – it looks pretty symmetric, except for the value of 32.8. An individual measurement like this one that falls outside the overall pattern of the data is called an outlier.

Next, let's consider another dataset where a stemplot can help us visualize the numbers – fuel economy information (city mpg) on Toyota's 2012 vehicle line. The data have been organized into the stemplot in Figure 2.3. This time, the stems have been arranged from highest at the top to lowest at the bottom. (Note: the 5|1 at the top is for 51 mpg.)

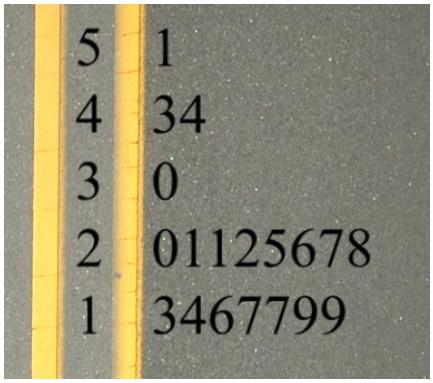

Figure 2.3 Stemplot for 2012 Toyota's city mpg.

Take a look at the overall pattern of the stemplot (Figure 2.3). Most of the mpgs are clustered at the lower end of the plot. We can expand the stem to change the resolution of the picture.

We break each stem into two, so the low digit leaves 0, 1, 2, 3, 4 are on a different stem than the high digit leaves 5, 6, 7, 8, 9. The expanded plot appears in Figure 2.4.

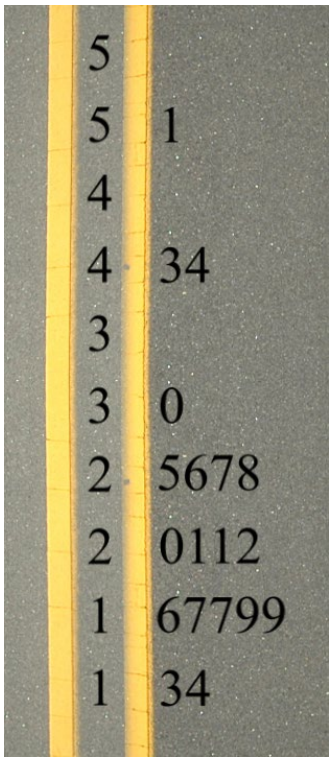

Figure 2.4. Stemplot with expanded stem.

Notice that we have outliers again, but this time an explanation is obvious. The high numbers are due to the super fuel-efficient hybrid vehicles Toyota makes.

Stemplots can be used to compare two different datasets as well. Say we wanted to compare Toyota's 2012 numbers with those from their 1984 line. We can make a back-to-back stemplot to see how mileage numbers have changed over the decades.

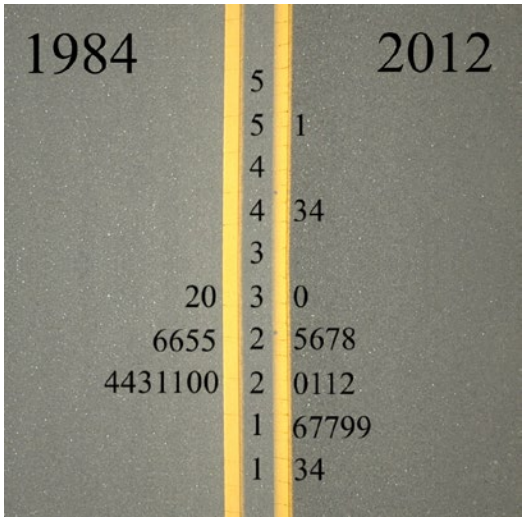

Figure 2.5. Comparing Toyota's 1984 line with its 2012 line.

What is interesting is that in 2012 Toyota had more vehicles way down at the low end, and a few more up at the high end. These extremes are easy to explain when you think about what you see on the roads – modern car buyers are interested in not-so-efficient SUVs and trucks as well as uber-efficient hybrids.

So you can see how stemplots help to tease meaning out of the disorder of raw data. They are useful for visualizing the shape of your data's distribution, and figuring out how frequently particular data classes pop up in your sea of numbers.

Later videos will show other ways to display data.

STUDENT LEARNING OBJECTIVES

A. Be able to differentiate between measurement or count data and categorical data.

B. Understand that the distribution of a variable shows what values the variable takes on and how often.

C. When presented with unorganized raw data, begin by making a graphical display of the data.

D. Be able to construct a stemplot to display the distribution of a variable for small datasets.

E. Be able to describe a graphical display such as a stemplot by first describing the overall pattern and then deviations from that pattern. Be able to identify outliers as important deviations from the overall pattern.

F. In terms of the overall shape of a distribution, recognize when it is roughly symmetric and approximate the center of the distribution.

CONTENT OVERVIEW

This unit on stemplots is the first in a series of units to focus on graphical representation of quantitative data. The emphasis in this unit is not simply to construct a stemplot but to use the plot to interpret the data's story.

The clearest picture of the distribution of values of a variable is just that – a picture. A *stemplot* (or stem-and-leaf plot) is a simple kind of graph that is constructed using the numbers themselves. Here's an example of head sizes in inches of 30 male soldiers. The head size was measured by putting a tape measure around each soldier's forehead.

23.0	22.2	21.7	22.0	22.3	22.6
22.7	21.5	22.7	24.9	20.8	23.3
24.2	23.5	23.9	23.4	20.8	21.5
23.0	24.0	22.7	22.6	23.9	21.8
23.1	21.9	21.0	22.4	23.5	22.5

To make a stemplot of these measurements, we first separate each observation into a stem, which is the first digit or digits, and a leaf, the final digit. The stems can consist of any number of digits, but the leaves generally have only a single digit. For the head circumference data, the measurements range from about 21 inches to about 25 inches and are measured to tenths of an inch. We'll take the whole inches as stems and the tenths as leaves.

First arrange the stems in order with a vertical line to their right as shown in Figure 2.6.

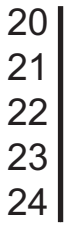

Figure 2.6. Setting up the stem of the stemplot.

Next, go through the list of observations, putting each leaf on the proper stem. The first soldier's head size was 23.0 inches, so we put leaf 0 on stem 23. The second head size is 22.2, so we put leaf 2 on stem 22. When we are finished, we have the display in Figure 2.7.

```
20 | 88
21 | 755890
22 | 2036777645
23 | 035940915
24 | 920
```

Figure 2.7. Stemplot with unordered leaves.

As a final step, arrange the leaves in order from smallest to largest. Figure 2.8 shows the completed stemplot. (This final step is unnecessary if technology is used to order the data from smallest to largest.)

```
20 | 88
21 | 055789
22 | 0234566777
23 | 001345599
24 | 029
```

Figure 2.8. The completed stemplot.

If there are too many stems with no leaves or only one leaf, it often helps to truncate the numbers and then to make a stemplot of the truncated numbers. (Truncation is faster than rounding.) If the leaves are crowded onto too few stems, expand the stem. For example, each stem can be split into two, one for leaf digits 0, 1, 2, 3, 4 and the other for leaf digits 5, 6, 7, 8, 9. (Or split each stem into five, using leaf digits 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8 and 9 for the five stems.)

Splitting stems can often reveal new information, as was the case of the fuel economy of Toyota's 2012 vehicles that was shown in Figures 2.3 and 2.4. Hence, don't be afraid to experiment with different stems or truncation to see what additional information might be learned. Finally, placing stemplots back-to-back is a good way to compare two datasets. (See Figure 2.5.)

Making the stemplot isn't the end in itself. It is a tool to help unlock the data's story. For example, the completed stemplot in Figure 2.8 gives a picture of the distribution of soldiers' head sizes. From the stemplot, we learn that the smallest head size was 20.8 and the largest was 24.9. The shape of the distribution is mound shaped (one peak). Although the two sides of the stemplot would not line up exactly if we folded the plot along the 22 stem, they come pretty close. So, we can say that this distribution is roughly symmetric. A middle value is somewhere in the 22-inch range (in other words, somewhere between 22.0 and 22.7).

The art of looking at stemplots intelligently is as important as the skill of making them. In looking at any distribution, always look first for the overall pattern of the distribution and then for any striking deviations from that pattern. In sizing up the overall pattern, look for and try to describe the following:

- center and spread
- one peak or several
- a regular shape, such as symmetric

For now, identify a center by looking at the stemplot and selecting a number that appears to best measure the middle of the distribution. (In later units, we will cover specific measures of center such as the mean and median).

KEY TERMS

A **variable** describes some characteristic of interest that can vary in value. Some variables are **categorical** (soldiers' gender – male or female). Others are **quantitative** (soldiers' head circumference or foot length).

The **distribution** of a variable describes the possible values the variable takes and how often it takes these values. Stemplots are one way to graph the distribution of a quantitative variable.

Shape, center, and spread describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as **unimodal** (single peak) or **symmetric** (one side is the mirror image of the other).

Outliers are data values that lie outside the overall pattern of the distribution. Always look for gaps in the data and outliers and try to explain them.

A **stemplot** (or **stem-and-leaf plot**) is a useful tool for conveying the shape of relatively small datasets and identifying outliers. It consists of two columns, one for the stems and the other for the leaves (often separated by a vertical line).