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.

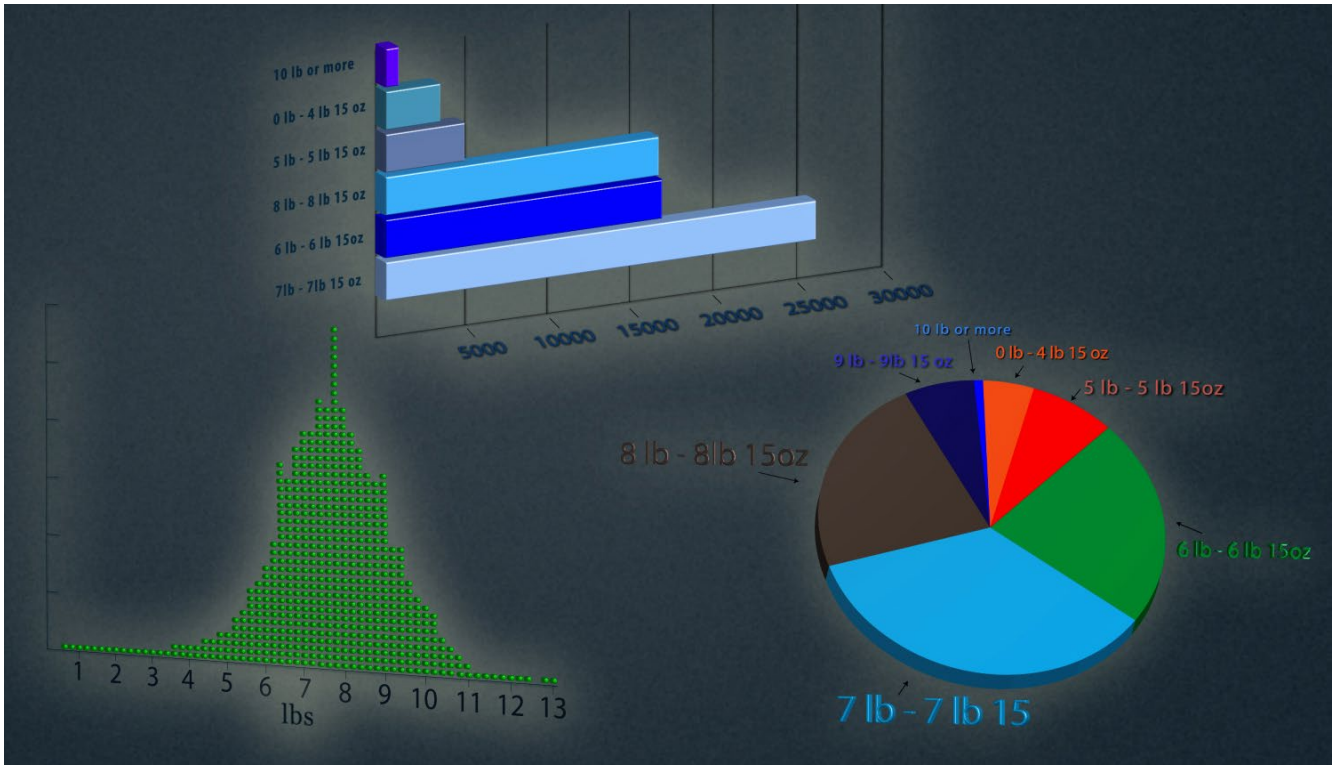

Figure 2.1. Graphic displays of weights of newborns.

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.

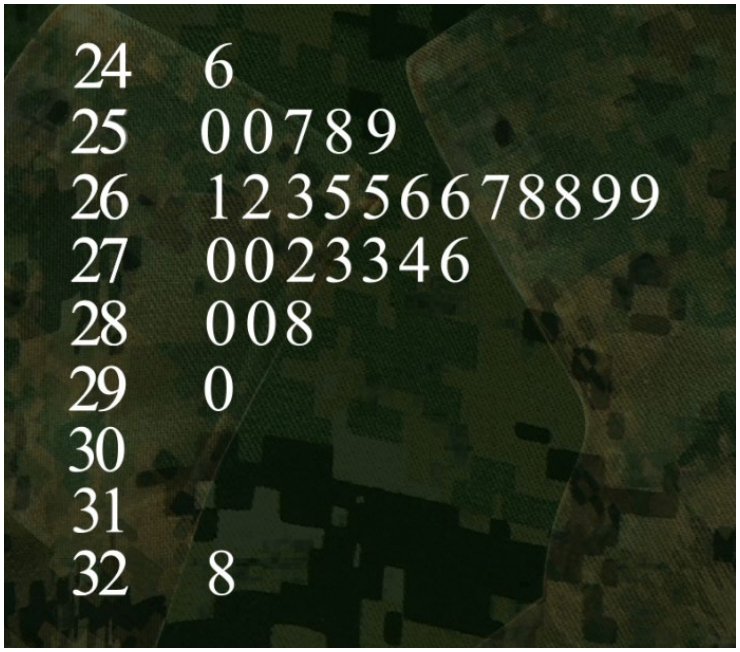

Figure 2.2. Stemplot of soldiers' foot lengths.

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.)

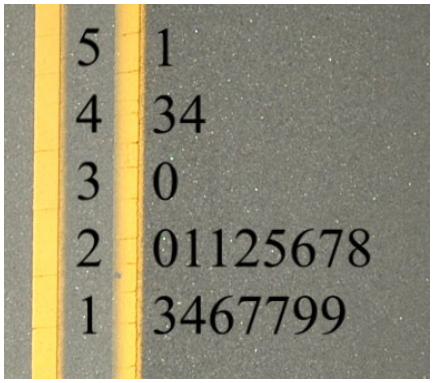

Figure 2.3 Stemplot for 2012 Toyota's city mpg.

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.

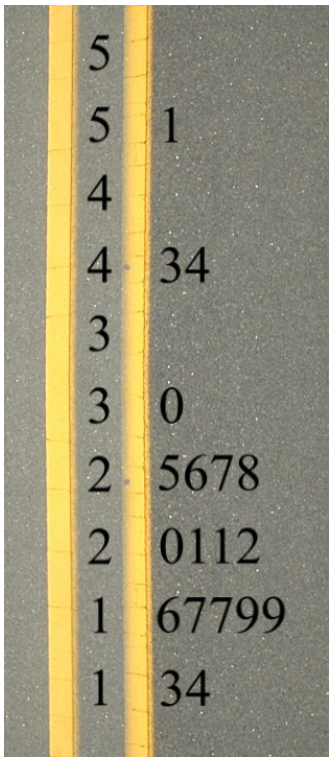

Figure 2.4. Stemplot with expanded stem.

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.

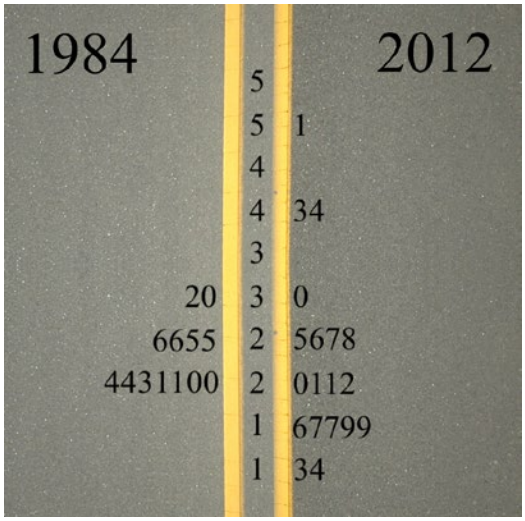

Figure 2.5. Comparing Toyota's 1984 line with its 2012 line.

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.