huber-data-2

Now that you have learned the basics techniques and statistical calculations used to describe a data set, the next step is figuring out how to effectively illustrate and visualize your data.

It is useful to first visualize the data before a statistician engages in a thorough analysis of the data set. In this lesson, we are going to learn useful techniques for visualizing numerical variables.

By organizing the data into a PLOT or GRAPH, a statistician is able to explore and summarize some basic properties of the data set. The discipline of quantitatively describing the main properties of a data set is known as DESCRIPTIVE STATISTICS.

The simplest type of plot is the DOT PLOT, which is used to visually convey the values of one variable. In a dot plot, there is only a horizontal x-axis, and the data points are represented as dots above this axis.

Here is a dot plot created using the variable 'price' from our 'cars93' data set (which is part of the openintro package). As you may notice, the price is reported along the x-axis in $1000s, and each point above the axis represents the price of one of the 54 cars in our data set.

When looking at this dot plot, around what price (in $1000s) does there appear to be the highest density of data points?

Since dot plots effectively display the specific numerical value of one variable for each individual in the data set, they are particularly useful when analyzing smaller data sets.

A HISTOGRAM is similar to a dot plot, but instead of showing every specific value, it partitions the values of your data into several bins, providing a more condensed representation of the data.

Here I have created a histogram using the miles per gallon data for all of our cars. As you may notice, the values of the MPG along the x-axis are partitioned into bins with a range of 5. The second bin, for example, groups together all of the cars that get 21-25 MPG in the city, and so forth. Note that the bin to the left of this contains those cars with 20 MPG since this value cannot be counted in both bins. The frequency of values in each bin, or the number of cars in each of the intervals, is reported along the y-axis.

Taller bars signify the range of values in which the majority of the data is located, whereas shorter bars represent a range of values in which only a little bit of the data is located. In other words, histograms provide a view of the DATA DENSITY.

By simply looking at this histogram, can you tell me which MPG interval has the highest frequency of values? For example, the lowest frequencies of values occur in the intervals 36-40, 41-45, and 46-50.

1. 16-20
2. 21-25
3. 26-30
4. 31-35
5. 36-40
6. 41-45
7. 46-50

16-20

How many cars get 16-20 MPG in the city?

A red line has been drawn on our histogram illustrating the previous answer.

Histograms are particularly useful in viewing and describing the shape of the distribution of the data. A distribution of data may have a left skew, a right skew, or no skew at all. SKEWNESS is a measure of the extent to which the distribution of the data 'leans' to one side or the other.

A distribution that has a left skew is one in which the left TAIL of the plot is longer. In other words, on a histogram the majority of the distribution is located to the right of the mean.

When a distribution is left-skewed, the value of the MEAN is less than that of the MEDIAN, and thus the MEAN is located further to the left of the distribution. In this plot, the green line represents the median and the blue line represents the mean.

On the other hand, a distribution that has a right skew is one in which the right tail is longer, such that the majority of the data falls to the left of the mean, when viewed on the histogram.

When a distribution is right-skewed, the value of the MEAN is greater than that of the MEDIAN, and thus the MEAN is located further to the right of the distribution. In this plot, the green line represents the median and the blue line represents the mean.

A plot that has no skew is one in which the tails on both sides of the mean balance out, and is referred to as symmetric. When a distribution is symmetric, the MEAN and MEDIAN are approximately equal in value.

In this plot, the green line represents the median and the blue line represents the mean. The green line is the only one visible since the mean and median are close to the same value.

Now, let us take a look back at the histogram we made earlier, which represents the distribution of the values for city MPG for each of the 54 cars from our 'cars93' data set.

How would you classify the shape of the distribution represented by this histogram?

1. Symmetric
2. Right-skewed
3. Left-skewed

Right-skewed

Referring to the histogram above, and keeping in mind the real shape of the distribution, would you expect the MEDIAN to be greater than, less than, or equal to the MEAN?

1. Greater than
2. Less than
3. Equal to

Less than

A special type of histogram is known as a STEM-AND-LEAF PLOT. This plot organizes numerical data in order of decimal place value. The left-hand column of the plot contains the STEMS, or the numerical values of the tens digit for each of the data points, organized vertically in increasing order. The LEAVES are located in the right-hand column of the plot and are the values of the ones digit for each data point of the corresponding stem, organized horizontally in increasing order.

In a stem-and-leaf plot, the number of leaves is equal to the number of items in the data set. The easiest way to understand a stem-and-leaf plot is to see one!

I have created a stem-and-leaf plot above using the same values for the 'mpgCity' variable as we just used for our histogram. As you can see, a stem-and-leaf plot is a useful type of histogram if you want to see the frequencies of specific values of the data. Often, there will only be one bin per tens digit, but in this case, R gives us the same bins as we saw in our histogram.

Demonstrated on this stem-and-leaf plot, how many occurences of the value '22' are there in this particular data set?

The final plot that can be used for discrete or continuous variables is known as the BOX PLOT, also called a BOX-AND-WHISKER PLOT. As you previously learned, this plot is used to summarize the main descriptive statistics of a particular data set and help illustrates the concept of variability. I have created a box-and-whisker plot so that you can be reminded of what it looks like.

A box plot is used to visually represent the MINIMUM, FIRST QUARTILE (Q1), MEDIAN, THIRD QUARTILE (Q3), and MAXIMUM of a data set. The R-command 'summary(cars93$price)' returns values for these main descriptive statistics. Try this now.

```
summary(cars93$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.40   10.95   17.25   19.99   26.25   61.90
```