Building histograms from data sets

Now that we understand the basics of histograms, let's dive into the details about how we can actually build a histogram if we're only given the data set. Let's start by defining some important terms related to histograms.

The first one is **class interval**, or **class**, or **bin**, as we mentioned in the previous lesson. We always divide our data set into class intervals with equal **class width**. The class width is usually the difference between either the upper limits of two consecutive classes or between the lower limits of two consecutive classes.

For example, if the first class in our histogram is $5 - 9$ and the second class is $10 - 14$, then the class width is given by the difference between the upper limits, $14 - 9 = 5$, or by the difference between the lower limits, $10 - 5 = 5$.

A **class midpoint** is the value that's at the center of a particular class. The class midpoint is half the sum of the lower and upper limits of the class. For example, for the class $100 - 104$, the class midpoint is

$$\frac{100 + 104}{2} = 102$$

If at all possible, it's nice to choose a class width that's odd (like a width of 5, 13, 27, etc.) because it'll make the class midpoint an even number, instead of a decimal.

Let's look at the steps that we need to use to turn raw data into a histogram.

1. Put the data set in ascending order, then find the range as the difference between the largest and smallest values.

2. Determine the number of bins, or classes, that we want to have in our histogram. As a rule of thumb, it's best to use $5 - 6$ classes for most of the data we'll work with during our statistical studies. However, we might want to use up to 20 classes when we deal with larger data sets. It all depends on how large our data set is and the number of classes that would best represent the data.

3. Divide the range by the number of classes, then round up the result to get the class width.

4. Build a table, putting each class in a separate row.

5. Find the frequency for each class by counting the data points that fall into each one. It's worth mentioning that we can choose either overlapping intervals or non-overlapping intervals in the previous step. However, for overlapping intervals like $0 - 4$ and $4 - 8$, the data point 4 should be included in the second interval. In other words, the interval $0 - 4$ actually contains data from 0 to $3.999...$, and the interval $4 - 8$ contains data from 4 to $7.999...$ Overlapping intervals are particularly useful when the data set contains decimals.

6. Graph the histogram by placing the classes along the horizontal axis and their frequencies along the vertical axis, such that the height of each bar is the frequency of each class.

Let's work through an example so that we can see these steps in detail.

Example

20 students recorded the number of hours they spent doing homework last week. Construct a histogram for the number of homework hours spent by the group.

$$3, 2, 6, 13, 7, 5, 12, 1, 8, 4, 5, 9, 6, 15, 4, 3, 10, 14, 5, 11$$

First, put the values in ascending order. This makes it easier to see the maximum and minimum values of the data set.

$$1, 2, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15$$

The range is $15 - 1 = 14$. We'll divide the data set into 5 classes, which means the class width will be

$$\frac{14}{5} = 2.8 \approx 3$$

Remember that it's better to round the result up, which is why we rounded 2.8 to 3. Since 1 is the minimum value, we can use it as a starting point, or the lower limit of the first class. So our table for the classes and their frequencies will be

Classes	Frequency
1 - 3	4
4 - 6	7
7 - 9	3
10 - 12	3
13 - 15	3

We don't need it for this particular problem, but we could also determine the midpoint of each class.

Classes	Frequency	Class midpoint
1 - 3	4	2
4 - 6	7	5
7 - 9	3	8
10 - 12	3	11
13 - 15	3	14

Now we can use the table to draw the histogram.