# Video: “L9Histogram-BinNumber” (3:21)

(0:00):
Unfortunately there is not standardized way to pick the number of bins so you’re going to have to develop some intuition about good bin sizes. Most tools have an automatic way of picking the right number of bins and then expect you to adjust as appropriate. For example, MATLAB always picks 10 bins and excel uses the square root of n where n is the number of data points. What should you do? I would take a preliminary look using the defaults and then adjust the bins if you need to.

(0:32):
Your choice of bin size can make a huge difference in the appearance of your histogram, in other words size matters. The histogram for this data uses the MATLAB default of 10 bins looks like this. Excel's default choice of the square root of 751 is about 27, so the histogram with 25 bins looks like this. Both histograms are good but I think the histogram with 25 bins shows the data features a little better. The 25 bin histogram has a smoother change in silhouette while the 10 bin histogram has a larger jump in bar height. Notice that the vertical scales are quite different for these two cases. If you pick fewer bins, say five, the histogram looks like this. When you don’t pick enough bins most of the data falls into a few bins and you can’t tell much about the data. If we pick more bins, say 100, the histogram takes on a more jagged appearance. As the bins become too small each will have too few data points to show the data features. For most data sets, trying out a couple of likely number of bins and picking the one that reveals the best data features is all you need to do.

(1:48):
What can possibly go wrong with this? Let’s take the same example of 751 beak sizes and 25 bins. Notice that all the data falls within 6 and 14 millimeters. Now let’s add a single data point with the value 80 millimeters to the data set and still use 25 bins. The histogram looks like this because the data range is evenly divided between 6 and 80 millimeters most of the data falls within 3 bins. Histograms mainly go wrong because the data sets are not even in some data sets. Either they have outliers or the density of the points in certain regions changes a lot.

(2:31):
If the problem is due to a few outliers we want to separate those outliers from the data and examine them carefully. In our example the 80 millimeter outlier is likely a mistake in data entry. Since the rest of the data is within 6 and 14 millimeters a brown finch with an 80 millimeter beak size is unimaginable. Be careful about throwing out the outliers unless you really isolated the cause of the problem. Some great scientific discoveries have been missed by scientists who just threw out the points that didn’t fit. Another approach is to only bin the high density regions and lump everything outside the region into bins at the end. For multiple high density regions you may need to create multiple histograms. Another approach is to use unequal bin sizes. They select the bins so that each bin hold 10% of the data.