CSC 239 Concept: Constructing a histogram

The construction of a histogram is an important first step in any statistical analysis. The visual appearance of the histogram can be used to determine if there are outliers and what the appropriate statistics are for reporting center and spread. For more complex statistical analysis, the shape of the histogram can be used to determine what analyses are appropriate and which are not. However, the creation of a histogram is not a simple process; it is iterative and benefits greatly from automation. The iterative process for constructing a histogram is an example of computational thinking that belongs in the Evaluation category.

Learning goal – students can use a statistics package to generate multiple histograms from a single data set and can choose the most meaningful visualization of the symmetry or skew of the data.

Discussion: students are shown the algorithm for constructing a histogram:

1) Determine the minimum and maximum of the data
2) Create a number of bins for aggregating data
    a. Bin range must equal or exceed the data range
    b. Bins must not be too few or too many
3) Create a distribution table for the data based on the bins
4) Graph the distribution table
    a. If the resulting histogram is too spiky, return to step 2 and create fewer bins
    b. If the resulting histogram is too shapeless, return to step 2 and create more bins

There are computer programs that completely automate the construction of a histogram. However, since the histogram is a visual statistic, it is best constructed using human automation; in particular the choice of the number of bins is critical in generating a histogram that best provides visual information about the distribution of the underlying data. In CSC239, students are shown that the construction of a histogram is critical in determining whether the mean and standard deviation or the mode and quartiles are better for summarizing the center and spread of the data.

Assessment – Students are given a data set and are asked to:

- generate histogram bins that include all the data
    - students will need to find the min and max of the data set and construct data bins that include both min and max.
- generate histogram bins that are all of equal width
    - the generated bins are all of equal width; especially, there is no “more” bin at the top of the range, or “less” at the bottom of the range
- generate the “best” number of bins
    - students generate multiple histograms and choose the one with the number of bins that best visualizes the data (i.e. patterns in the data set are made explicit and reasonable hypotheses can be made about the data set).

As this activity is done for many problem sets, the students have many opportunities to mean the objectives, and also to see different visualizations for different data sets. (E.g. small data sets versus large ones; symmetric distributions versus skewed ones.)