Organizing and Picturing Data

Often, the first step in organizing a set of measurements consists of creating a **frequency table** that places the data into several (usually 5 to 20) **classes**.

The number of classes you choose will depend on what you are looking for and how many measurements you are trying to analyze. For instance, if you have only 25 measurements then organizing them into 20 classes is hardly better than simply listing the measurements in ascending order. Using 5 or 6 classes in such a case would be more helpful in getting a feel for what "shape" the data have.

Some definitions are now called for:

Lower class limits are the smallest numbers that can actually belong to the different classes.

Upper class limits are the largest numbers that can actually belong to the different classes.

Class boundaries are the numbers that separate classes *without* the gaps created by the class limits. The boundary between two classes should be *exactly* half way between the upper class limit of the lower class and the lower class limit of the next class i.e. the class boundaries are the midpoints between adjacent class limits. For example, if the upper class limit of a class is 72.1 and the lower class limit of the next class is 72.2 then the class boundary is $\frac{72.1+72.2}{2}=71.25$.

Class marks are the midpoints of the classes (and can be found, for example, by averaging the upper and lower class limits of each class).

Class widths are the differences between consecutive lower (or upper) class limits. **N.B.** the class width is *not* (upper class limit) – (lower class limit).

A **frequency table** is then constructed using the following steps:

1. Decide on the number of classes – should be between 5 and 20.

2. Determine the class width by *class width = round up of* $\left(\frac{range}{number\_of\_classes}\right)$. If the number of classes divides evenly into the range then add another class to ensure that all the data is included.

3. Select the first lower class limit as either the lowest score or a slightly lower ***but nicer*** value. This value is the starting point.

4. Obtain the remaining class limits by adding the class width to the starting point and continuing. The upper class limits should now be obvious!

5. Use a table of *ranked scores* to determine the frequencies for each class and whatever other information you may want in the table (relative frequencies, cumulative frequencies, relative cumulative frequencies, etc.)

6. **Be Neat.**

The other definitions you will need are the following:

- $Relative\ frequency = \dfrac{class\_frequency}{sum\_of\_all\_frequencies}$.

- The *cumulative frequency* for a class is the sum of the frequencies for that class and all previous classes.

- **Histogram.** Bar graph with a *continuous* horizontal scale (so there are *no gaps* between bars i.e. the bars are contiguous) consisting of either the class *boundaries* or the *class marks* and frequencies on the vertical scale. The highest value on the vertical scale should be either the highest class frequency or a *nice number* slightly higher in order to produce an 'intelligent' scale. All histograms must be *appropriately labeled.*

- **Relative frequency histogram.** Same as above except with relative frequencies. Vertical scale should be chosen to produce an easily readable scale.

- **Frequency polygon.** Like a histogram – same frequencies on the vertical scale. Class marks on the horizontal scale with one extra class mark at either end. Above each *class mark* plot the point corresponding to the class frequency and then connect the points with *straight lines*. The extra class marks are to ensure that the graph begins and ends with a frequency of zero.

- **Ogive (cumulative frequency polygon).** (Brownie points to whoever discovers where the heck "ogive" comes from!) This is a *line chart.* The vertical scale is from zero to the sum of all the frequencies (i.e. the size of the data set). The horizontal scale consists of the *class boundaries.* Directly above each *class boundary* plot a point corresponding to the cumulative frequency for that value. Starting with the first (lower) class boundary (which has a *cumulative* frequency of zero) connect the points with straight lines. An ogive can also consist of cumulative *relative* frequencies.