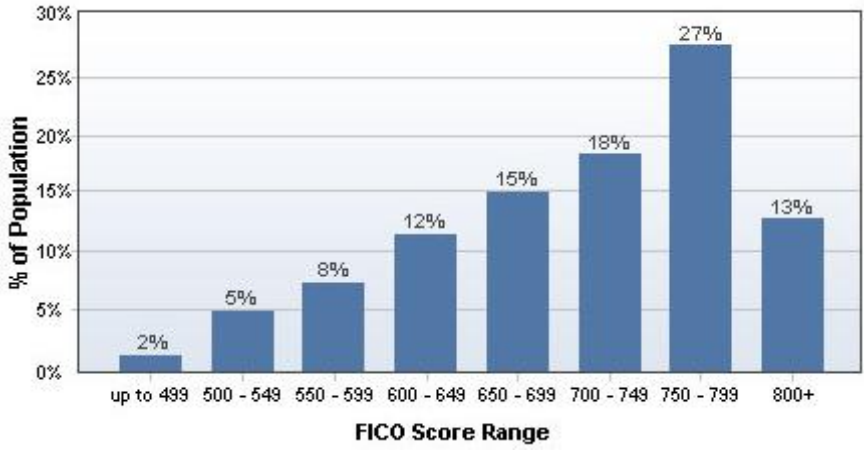
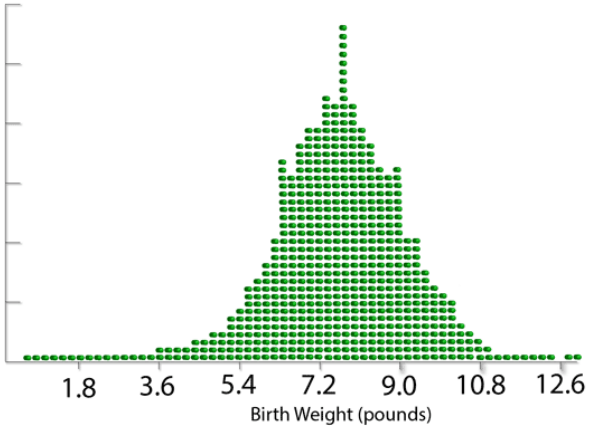
David Youngberg
BSAD 210—Montgomery College

LECTURE 05: OF DATA AND DISPLAYS I

I. Data Displays
   a. *Histogram*—a histogram divides data into groups and displays the number of observations per group
      i. Advantage: Easily organizes lots of data, especially when there are many possible divisions (e.g. income or other continuous variable)



   b. *Dot Plot*—like a histogram, a dot plot shows the number of observations, but doesn't divide them into groups
      i. Advantage: Don't lose information through generalizing

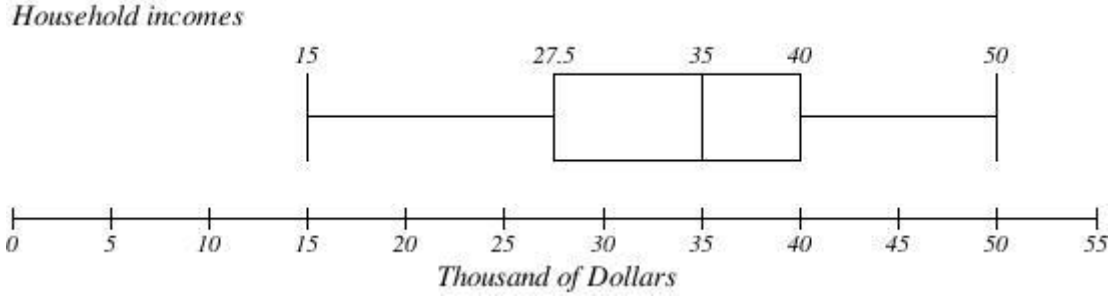


   c. *Stem-and-Leaf Display*—a table displaying data with their "tens+" digit on the left and the last digit, in numerical order, listed on the right.
      i. Example of stem-and-leaf showing data values of 15, 10, 9, 22, 10, 16, 21

0	9
1	0056
2	12

ii. Advantage: Gives us all the specific values of the data while showing a distribution, though better for smaller data sets

d. *Box Plot*—a display which shows where quartiles of data are
   i. A quartile is a part of a data set with one-fourth of the total observations. The 1st quartile is a data value which indicates when, from the minimum to that value, are the first fourth of the observations are
   ii. Note you can also divide the data into other segments such as in five equal parts (quintiles), ten equal parts (deciles), one hundred equal parts (percentiles), etc.
   iii. The lines on either side of the box show the range between the maximum and 3rd quartile and between the minimum and 1st quartile
   iv. The box is between the 1st and 3rd quartile with a line (the median, or 2nd quartile); the box is the *interquartile range*.
   v. The larger the distance between these points, the more disperse the observations. The shorter the distance, the more concentrated
   vi. Advantage: Like the steam-and-leaf diagram, it illustrates dispersion but it is able to handle virtually any number of observations. All you need to make a box plot are five numbers: maximum, minimum, 1st quartile, 3rd quartile, and median (2nd quartile).



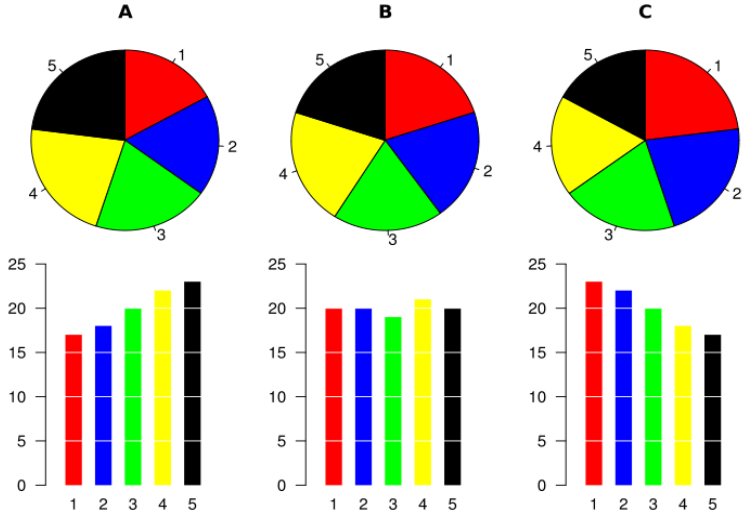
II. Categorical Data

a. All of these previous types of displays help us organize data given as a continuous variable, such as a number. But sometimes you want to organize *categorical data*, where there are several groups and the data consists of how many observations are in each group.

b. *Pie Chart*—a circular chart divided into sections, or wedges, describing a percent of total each group is. Bigger wedges mean a bigger percent. This is one of the most widely used charts out there but it's not perfect (as I will show you).
   i. Advantage: It is widely used and easy to understand.

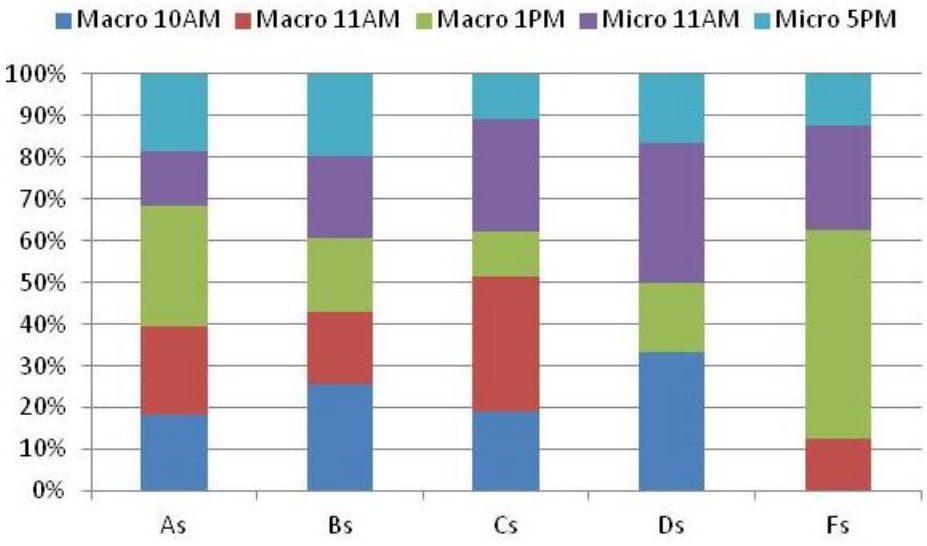
c. *Bar Chart*—like a histogram, but each bar represents a category rather than a range of a distribution.
    i. Advantage: It is also widely used and easy to understand. It typically has an advantage over bar charts in showing each group's size relative to the other.

d. In B, is red or blue larger?

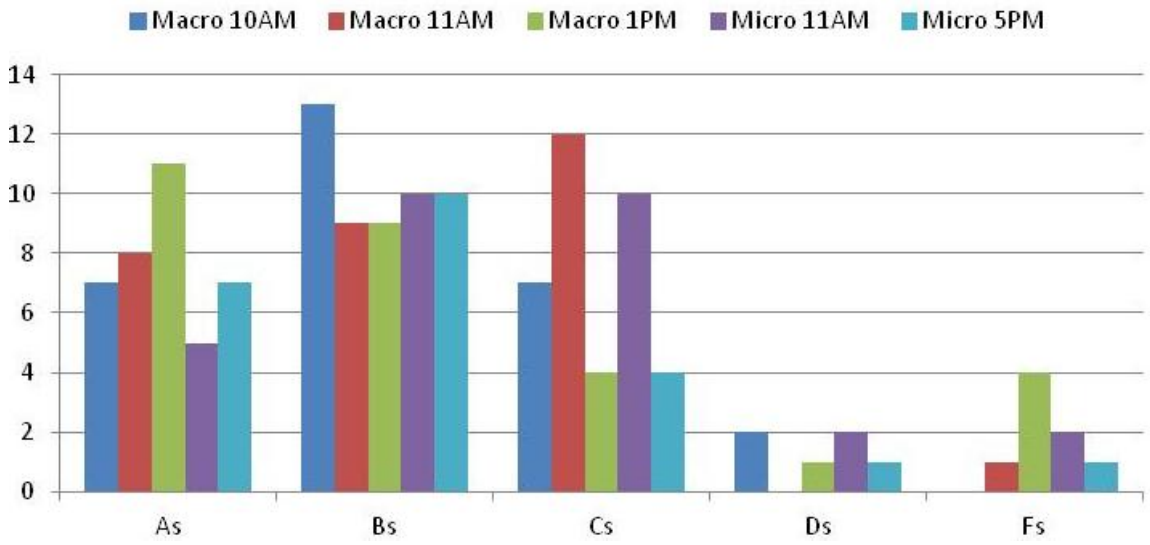


e. *100% Stacked Bar*—this chart sets each group as a bar representing not a value, but 100%. Each group is then divided based on a different category, with the vertical distance determined by the percent.
    i. Advantage: It's great for comparing groups in the second-level category and displays in a small amount of space.
    ii. Stacked bars can be a little deceptive, though. Consider this grade data from my classes in the spring of 2014.

f. Which class received the most Fs? Which class did the worst over all?



i. The first question is easy to answer, but the second one is trickier. Consider the bar graph, using the same data:



g. Bar charts are, in general, your best option—it's clear which class got the most Fs and it's also clear Fs overall were unusual—but note that this graph takes a bit more room and looks a bit cluttered. Taking up too much space with too much going on can be a problem, too.