Math 102
Statistics Handout 2

Visually Presenting Data:

Given a data set, there are several ways to present it visually. Here are a few of the most common:

1. Frequency Tables:

   In a frequency table, the data is divided into disjoint (non-overlapping) data categories or *classes*. Then a two column table is constructed. Each class has its own row in the table. The rule for the class is put in the first column, and the number of data points in that class is put in the second column.

   Example: [Soda pop survey data]

2. Relative Frequency Tables:

   In a relative frequency table, the data is again divided into disjoint (non-overlapping) data categories or *classes* and a two column table is constructed. Each class still has its own row in the table, and the rule for the class is put in the first column. However, here, the *fraction* or *percentage* of the data points that are in a given class is given as the entry in the second column.

   Example: [Soda pop survey data]

3. Histograms (Bar Graphs):

   A histogram can be made by taking the data from either a frequency table or a relative frequency table. Instead of putting the data into a table, each class placed along the horizontal axis of a graph. Then, a bar is drawn vertically up from a given class. The height of the bar is scaled in direct proportion with the size of the data class. Both the horizontal and vertical axes of the histogram should be clearly labeled and scaled appropriately.

   Example: [Soda pop survey data]

4. Stem and Leaf Plots:

   In a stem and leaf plot, data with more than one place value is divided into classes where the largest place value is the same for all the data in a class. This place value is called the "stem". The data is then visually presented by writing the shared "stem" to the left of a vertical line, and then writing the rest of the places of each data point in the class one at a time, (in increasing order and separated by commas) to the right of the vertical line and across from the "stem". The values to the right of the vertical line are called the "leaves".

   Example: Suppose we are given a set of test scores (out of 100) on an exam:

   $\{52, 67, 72, 84, 91, 96, 85, 74, 62, 56, 77, 79, 88, 48, 80, 68, 74, 81, 78, 70\}$

Measures of Center:

Given a numerical data set, how do we go about deciding where the "middle" of the data set is. That is, for a given data set, what does it mean to be "average"?

(a) One way of computing the "average" of a numerical data set is to compute the **mean** of the data. To do this, we add up all the data values, and then divide by the number of data points. One disadvantage of using the mean is that it is sensitive to "outliers". That is, one very large or very small data value can change the mean quite a bit.

(b) Another way of computing the "average" of a numerical data set is to compute the **median** of the data. To do this, we often start by lining up all the data values in increasing order. If there are an odd number of data values, then the median of the data set is the "middle" data value. If there are an odd number of data values, then the median is the average of the "middle pair" of values. The median is less sensitive to "outliers" then the mean is, since it only cares about the order of the data and not the actual numerical values.

(c) A third way of computing the "average" of a numerical data set is to compute the **mode** of the data. The mode is the data value that occurs most often (the most "popular" value). If two data values tie, then we say that both values are the mode. If three or more values tie for the most occurrences, we usually say that there is no mode.

(d) A fourth way of computing the "average" of a numerical data set is to compute the **midrange** of the data. To find the midrange, we take the average of the largest and smallest values in the data set.

Examples:

(a) Given the data set: $\{3, 5, 12, 14, 6, 7, 10, 7, 8\}$, compute the mean, median, and mode of the data set.

(b) Given the data set: $\{3, 7, 10, 6, 4, 3, 5, 8, 6, 2\}$, compute the mean, median, and mode of the data set.

(c) Compute the mean, median, and mode for the data in the given frequency table:

Score	Frequency
5	4
7	3
9	4
10	6
12	3

The Five Number Summary and Box-and-Whisker Plot of a Data Set:

The 5 number summary of a given data set is the set of numbers: m, Q_1, Q_2, Q_3, M

Here, m is the minimum, or smallest data value, M is the maximum or largest data value, and Q_2 is the median of the data, as defined above. For the remaining numbers, Q_1 is the "first quartile" and is the median of the "lower half of the data set. That is, the median among all data values below Q_2. Q_3 is the "third quartile" and is the median of the "upper half of the data set. That is, the median among all data values above Q_2. A Box-and-Whisker plot is a picture of the 5 number summary. We will demonstrate how to construct it in the example below.

Example: Find the 5-number summary and draw a Box-and-Whisker plot for the data set:

$\{2, 3, 5, 5, 6, 8, 9, 11, 12, 16\}$