MAS 170 Elementary Statistics Fall 2021
Descriptive Statistics, Ch 3--6
revised: 4 June 2021

The purpose of this note is to supplement the textbook's account of basic descriptive statistics.

A ***histogram*** for a list of numbers X is a drawing with a collection of rectangles. The base of each rectangle is an interval, called a ***class interval***, of a horizontal axis whose units are the same as X (that is, if the numbers in X are dollars, or years, then the units of the horizontal axis are dollars or years, etc). The area of a rectangle whose base is the class interval from A to B is the fraction of the numbers in the list X that lie between A and B. For example, if 40% of the numbers in the list X are in the range 10 to 30, then a rectangle in a histogram for X whose base is the class interval from 10 to 30 has area 0.4 = 40%, and therefore has a height of .4/(30-10) = .02 = 2 percent per X unit. The vertical axis in a histogram has units called ***density*** units. Rectangles in a histogram must be non-overlapping, and the total area of all the rectangles must be 100%.

The ***average*** of a list of numbers is the sum of all the numbers divided by the number of entries in the list. We write AVE(X) to denote the average of a list X. The average of the list X turns out to be the balance point of a histogram for X, in the following sense. Imagine the rectangles of the histogram are made of uniformly thick clay. Glue the rectangles together along the edges where they meet. This solid clay histogram balances on a fulcrum placed at the average point on the base. *[Technical assumption: the numbers that fall within each rectangle base interval must be close to evenly distributed within that interval for this balance point interpretation to be valid.]*

Now suppose we have a histogram for a list X that has no rectangles of height zero. Suppose that L is the location, along the horizontal axis, of the left edge of the left-most rectangle in the histogram, and R is the location of the right edge of the right-most rectangle. A number A in the range

```
  L <= A <= R
```

is said to have p-th ***percentile*** rank if the area of the histogram on the interval from L to A is p percent. The ***median*** of the list X is the number with 50th percentile rank. The ***interquartile range*** of a list X is the interval from A to B, where A is the 25th percentile and B is the 75th percentile. [Note: these definitions of percentile and median are dependent on the histogram that is chosen.]

Operations on Lists
===================

Given lists X,Y of the same length, and given constants a,b, we can form new lists X^2, aX+b, and XY by doing the obvious operations on the entries of the list(s). For example, if X = -2,1,3 and Y = 1,4,0, then we have the following lists.

```
       X = -2,1,3
       Y = 1,4,0
     X^2 = (-2)^2,1^2,3^3 = 4,1,9
  2X - 3 = 2(-2)-3, 2(1)-3, 2(3)-3 = -7,-1,3
      XY = (-2)(1),(1)(4),(3)(0) = -2,4,0
```

The ***root-mean-square*** of a list X, denoted rms(X), is the square root of the average of the list X^2. Here is rms(X) in symbols.

```
  rms(X) = SQRT(AVE(X^2))
```

The standard deviation of a list X, denoted SD(X), is the rms of the list X-AVE(X). The list X-AVE(X) is sometimes called the list of deviations from average. Here is SD(X) in symbols.

```
  SD(X) = rms(X-AVE(X))
```

The standard deviation is interpreted as a the (absolute) size of a typical error, that is, distance from average, in a game of chance where you draw entries from the list X at random.

The list X in ***standard units*** is the list (X-AVE(X))/SD(X).

The Standard Normal Distribution
================================

A list X is said to follow a ***normal distribution*** if the percent of entries between pairs of numbers

$$-(A-AVE(X))/SD(X),\ (A-AVE(X))/SD(X)$$

in the list (X-AVE(X))/SD(X) is close to the numbers given in the standard normal table in the Appendix of the textbook, for every entry A in the list X.