Class #4 - Wednesday February 10
Measures of Variation (or "Dispersion")

Textbook readings:

- Ross, Sec 3.5: Sample Variance and Sample Standard Deviation
- Phillips, Chapter 4: Measures of Variability

Introduction: The following data sets have the same mean (compute them!), but clearly C is much more spread out (more "dispersed") than B, and B is much more spread out than A

$$A = \{4, 4, 4, 4, 4\}, \quad B = \{1, 2, 5, 6, 6\}, \quad C = \{-40, 0, 5, 20, 35\}$$

We can see this from the frequency histograms of these datasets. But how can we numerically measure the greater variation in C as compared to B as compared to A?

We will define a statistic called the **sample variance**. The variance, and its square root, which is called the **standard deviation**, are the two most common measures of variation when considering data sets.

Formulas/Definitions:

- the deviation of an individual data value x_i is $x_i - \bar{x}$ (i.e., the difference between x_i and the mean; see Ross Sec 3.2.1, p78)

- square each of the individual deviations and add them up to get the "sum of squared deviations" SS_x:
$$SS_x = \Sigma_{i=1}^n (x_i - \bar{x})^2$$
(understand why we square the deviations! Read pp99-100 of Ross)

- the **sample variance** is the "average" of the squared deviations, but for technical reasons we divide by $n - 1$ instead of n:
$$\text{sample variance ("s squared"): } s^2 = \frac{SS_x}{n-1} = \frac{\Sigma_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- the **standard deviation** is just the square root of the variance:
$$\text{sample standard deviation: } s = \sqrt{s^2} = \sqrt{\frac{SS_x}{n-1}} = \sqrt{\frac{\Sigma_{i=1}^n (x - \bar{x}_i)^2}{n-1}}$$

- an advantage of using the standard deviation instead of the variance is that the standard deviation is in the same units as the original data

Spreadsheet Functions

- `=var(data)` and `=stdev(data)` compute the sample variance and sample standard deviation
- there are also functions `=varp(data)` and `=stdevp(data)` which compute the *population* variance and standard deviation
- the difference is that for the population statistics you divide by the size of the data set n instead of $n - 1$