[10-09-20-T12]

Center and dispersion

We desire to characterize a collection of data by assigning to it a number, or at most a few numbers.

Let us assume a set of data $S = \{x_1, x_2, x_3, \ldots, x_n\}$.

■ Measures of central tendency

There are several measures of central tendency: arithmetic mean (there are other means such as the geometric mean and the harmonic mean), median, mode. Here we focus on the arithmetic mean, often called the "mean" or the "average". We will use the symbol $\overline{x}$ to represent the arithmetic mean. The average is supposed to be typical of the data. Since means tend to be centrally located within a set of data ordered from least to greatest, they are called *measures of central tendency*. When we say "average" here, we intend the arithmetic mean.

The familiar average is easily computed,

$$\overline{x} = \frac{\sum x}{n}.$$

Note that all sums are understood to be from $i = 1$ to $i = n$. That is, $\sum x$ means $\sum_{i=1}^{n} x_i$.

While we will not discuss the weighted average, it is important. It is often used. For example, a teacher might weight quizzes, take-home assignments, exams, projects, presentations differently. The greater the weight of an item, the greater its influence on the computed average. If item x_1 has weight w_1, x_2 weight w_2, etc, then the formula for a weighted mean is,

$$\overline{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + w_3 + \cdots w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{\sum w x}{\sum w}.$$

The last expression following our agreement on notation.

The familiar average may be typical of the data set and it may be located near the center of the data set, but it is mute on several interesting issues. For example, it does not indicate how the data are distributed about the mean. Perhaps there are two classes each with an average grade of "B", but one class might have scores clustered around "B" while the other class has mostly "A"s and "D"s. These are clearly two different kinds of classes.

■ Measures of dispersion (variation)

How might we compute a number that measures how the data are dispersed about the mean?

We begin by computing the difference of each data point x_i from the mean $\overline{x}$. Thus,

$$x_1 - \overline{x}$$
$$x_2 - \overline{x}$$

$$\begin{matrix} x_3 & - & \bar{x} \\ \vdots & \vdots & \vdots \\ x_n & - & \bar{x} \end{matrix}$$

We might expect that the average of these differences would be a number characterizing the dispersion of the data. Such an average would be useless, because it must always exactly equal zero. The sum of the negative differences (when $x_i < \bar{x}$) will exactly equal the sum of the positive differences ($x_i > \bar{x}$).

■ Mean Absolute Deviation (MD)

The Mean Absolute Deviation (MD) produces a non-zero average,

$$MD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}.$$

But, the MD is not often used.

■ Standard Deviation

Suppose we square the differences, $(x_i - \bar{x})^2$. That will guarantee a positive sum. We find the average of that sum. Thus,

$$sample\ variance = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

The number s^2 is non-zero, in fact it is always positive, and it is does assign a single measure to the dispersion of the data. The number s^2 is called the "sample variance". If we take the square root of s^2, then the units, if any, (feet, miles, kilograms, hours, hogsheads) of this average will match those of the data and the magnitude of the number will be similar to the magnitude of the data. This square root, s, we call the "sample standard deviation". It provides a measure of the dispersion of the data about the mean.

$$sample\ standard\ deviation = \sqrt{s^s} = s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}.$$

We have been speaking of a *sample* and the numbers we compute are therefore called *statistics*. When we speak of a *population*, we write μ, σ^2, σ for $\bar{x}$, s^2, s, respectively. The numbers μ, σ^2, σ are called *parameters*. Thus,

$$population\ standard\ deviation = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}.$$

It should be understood that a sample is a subset of the population.

■ Estimating the population standard deviation and variance

Quite frequently it is inconvenient or impossible to know the population parameters. It turns out that if we multiply s by $k = \sqrt{\frac{n}{n-1}}$, the result is a good approximation of σ. That is,

$$\sigma \approx s\sqrt{\frac{n}{n-1}} = \sqrt{\frac{n}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}.$$

And the expression inside the square root,

$$\sqrt{\frac{n}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2}$$

is called the *biased corrected variance.*

Notice that as n increases, $\frac{n}{n-1}$ becomes close to 1. Thus, for large n, say $n \geq 30$, this approximation of σ is very good.