Measures of Variation

We have looked at measures of central tendency, which describe how the data clusters around the center. However, with the mean and the median we do not know how far each observation is from the center. Statistics that describe the "spread" or dispersion of data are called measures of variation.

Let's use a simple example of two classes that took the same quiz in Statistics. Their grades on the quiz are:

Class 1	66	68	70	72	73	77	80	82	86	86	90	90	93	97	100
Class 2	75	75	76	77	78	79	80	82	84	85	87	87	87	88	90

Using our knowledge of measures of central tendency, we can calculate the mean and the median for each of the classes:

Class 1

$\overline{X}$ = 1230/15 = 82
Position of Q_2: (15 + 1) / 2 = 8
Q_2 = 82

Class 2

$\overline{X}$ = 1230/15 = 82
Position of Q_2: (15 + 1) / 2 = 8
Q_2 = 82

It turns out that the mean and the median for the two classes are identical. However, you can see that students in two classes did not get the same grades: in Class 2, each observation is closer to the mean and the median than in Class 1.

Range

The range is the difference between the largest and the smallest observation in a set of data. The very first thing you see after looking at the two classes' grades is that the grades have different ranges:

Range for Class 1 = Largest observation – Smallest observation = 100 – 66 = 34
Range for Class 2 = 90 – 75 = 15

Range is the simplest measure of variation, but it doesn't show precisely how close or how far the data is from the center.

Variance

Variance is the measure of variation that shows how far each observation is from the mean. The variance is denoted by S^2 (sample variance) and is calculated using the formula:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

The expression in parentheses means that you have to find the difference between each observation and the mean and then square each difference. The summation sign Σ (sigma) means that you have to add all squared differences. Next, divide the result by (n-1) (number of observations minus one). Here is the detailed variance calculation for Class 1:

	X_i (individual grades)	$\overline{X}$ (mean)	$X_i - \overline{X}$ (difference)	$(X_i - \overline{X})^2$ (squared difference)
	66	82	-16	256
	68	82	-14	196
	70	82	-12	144
	72	82	-10	100
	73	82	-9	81
	77	82	-5	25
	80	82	-2	4
	82	82	0	0
	86	82	4	16
	86	82	4	16
	90	82	8	64
	90	82	8	64
	93	82	11	121
	97	82	15	225
	100	82	18	324
Sum (Σ):	1230	1230	0	**1636**

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{1636}{15-1} \approx \mathbf{116.86}$$

As you see from column 3, the difference between each observation and the mean can be positive or negative. The sign indicates the position of the observation relative to the mean, for example, the negative sign of the difference shows that that particular observation is below, or smaller than, the mean. Our interest, however, is focused on how

far the observations are from the center. Squaring the difference helps to remove the negative sign.

However, because we squared the difference, the variance is expressed in squared units. In our case, the variance is in squared points. In the same way, you will get squared dollars or squared gallons when calculating the variance of gas prices or gas mileage. Squared units are inconvenient for interpretations or statistical analysis, and the standard deviation eliminates this problem.

Standard Deviation

Standard deviation of the sample (denoted by S) is found by taking a square root of variance, or

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = \sqrt{116.86} \approx 10.81$$

Standard deviation shows – in standard units – how far the observations are from the center. By calculating the standard deviation of two sets of data, you can compare variation in both sets. See if you can calculate the variance and standard deviation for Class 2:

$$S^2 = \frac{(75-82)^2 + (75-82)^2 + (76-82)^2 + \ldots + (87-82)^2 + (88-82)^2 + (90-82)^2}{15-1} \approx 26.86$$

$$S = \sqrt{S^2} = \sqrt{26.86} = 5.18$$

Standard deviation of Class 1 is greater than the standard deviation of Class 2, which means that the data (or grades) in Class 1 are more dispersed around the mean than the data (grades) in Class 2.

Later in the course, you will learn more about the uses of standard deviation, but so far a very important thing to know is the so-called **empirical rule**: the majority of data (68%) clusters within one standard deviation of the mean. This means that 68% of the data lies between $\overline{X}$ - S and $\overline{X}$ + S. In our example, in Class 1, 68% of the students got grades between 71 and 93:

$$\overline{X} - S = 82 - 10.81 \approx 71 \text{ and } \overline{X} + S = 82 + 10.81 \approx 93$$

Now, using the empirical rule, you can calculate the range of grades that the majority (68%) of students in Class 2 got:

$$\overline{X} - S = 82 - 5.18 \approx 77 \text{ and } \overline{X} + S = 82 + 5.18 \approx 87$$