4.4 Measures of Variability: Range, Variance, and Standard Deviation

While mean and median tell you about the center of your observations, it says nothing about the 'spread' of the numbers.

Example: Suppose two machines produce nails which are on average 10 inches long. A sample of 11 nails is selected from each machine.

- Machine A: 6, 8, 8, 10, 10, 10, 10, 10, 12, 12, 14
- Machine B: 6, 6, 6, 8, 8, 10, 12, 12, 14, 14, 14

To verify, let's compute the mean:

- mean for machine A: 110 / 11 = 10
- mean for machine B: 110 / 11 = 10

In both cases, the mean is 10, indeed. However, the first machine seems to be the better one, since most nails are close to 10 inches. Therefore:

We must find additional numbers indicating the 'spread' of the data.

The Range

The easiest measure of the data spread is the range. It is simply the highest data value minus the lowest data value (we have seen the range before). In the above example, the range is the same for both data, namely 14 - 6 = 8. The range is, while useful, too crude a measure of variability.

The Variance

We want to find out how much the data points are spread around the mean. To do that, we could find the difference between each data point and the mean, and average these differences. However, we want to measure the differences to the mean regardless of the sign (positive or negative difference). Therefore, we could find the absolute value of the difference between each data point and average that. But for theoretical reasons an absolute value function is not easy to deal with, so that one chooses a square function instead (which also neutralizes signs). Finally, for yet other theoretical reasons we shall use not the sample size n to compute an average, but instead n - 1.

Hence, we will use this formula to compute the data spread, or variance:

Variance = add up the squares of (Data points - mean), then divide that sum by (n - 1)

There are two symbols for the variance, just as for the mean:

- σ^2 is the variance for a population
- s^2 is the variance for a sample

In other words, the variance is computed according to the formulas:

- $\sigma^2 = \frac{\sum (x - \mu)^2}{n - 1}$ (for the population variance)

- $s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$ (for the sample variance)

We had to use two formulas because one involves the population mean, the other the sample mean. Practically, however, the formula is the same. It is useful to compute the variance at least once "by hand" before we show how to use Excel to accomplish the same feat quickly and easily.

How to find the variance "by hand":

1. Make a table of all x values
2. Find the mean of the data
3. Include a column with the difference to the mean
4. Include a column with the square of difference to the mean
5. Add the last column and divide the sum by (n - 1).

Here is the table that this procedure produces for the above sample of nails from machine A and B:

Machine A:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
6	4	16
8	2	4
8	2	4
10	0	0
10	0	0
10	0	0
10	0	0
10	0	0
12	-2	4
12	-2	4
14	4	16

Therefore, the variance for machine A is: (16 + 4 + 4 + 0 + 0 + 0 + 0 + 0 + 4 + 4 + 16) / 10 = 48 / 10 = 4.8

Machine B:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
6	4	16
6	4	16
6	4	16
8	2	4
8	2	4
10	0	0
12	-2	4
12	-2	4
14	-4	16

14	-4	16
14	-4	16

Therefore, the variance for machine B is: (16 + 16 + 16 + 4 + 4 + 0 + 4 + 4 + 16 + 16 + 16) / 10 = 112 / 10 = 11.2

In other words, the variance - or spread around the mean, for machine A is 4.8 while machine B has a variance (spread) of 11.2. That means that machine A seems to produce nails that, as a rule, produces nails that stick pretty close to the average nail length. Machine B, on the other hand, produces nails with more variability that machine A. Therefore, Machine A would be much preferred over machine B.

Note: The unit of the variance is the square of the original unit; hence, it is not the best number (considering units). Therefore, one introduces an additional number, called the standard deviation:

The Standard Deviation

The standard deviation is the square root of the variance.

As with the mean, there are two letters for variance and standard deviation:

- σ^2 is the variance for a population and $\sigma = \sqrt{\sigma^2}$ is the population standard deviation
- s^2 is the variance for a sample and $s = \sqrt{s^2}$ is the sample standard deviation

Example: Consider the sample data 6, 7, 5, 3, 4. Compute the standard deviation for that data.

To compute the standard deviation, we must first compute the mean, then the variance, and finally we can take the square root to obtain the standard deviation. In this case we do not need to create a table since there are so few numbers:

- Computing the mean: $\mu = (6+7+5+3+4)/5 = 5$
- Computing the variance: $\sigma^2 = \left[(6-5)^2 + (7-5)^2 + (5-5)^2 + (3-5)^2 + (4-5)^2\right]/4 = 2.5$
- Standard deviation: $\sigma = \sqrt{2.5} = 1.58$

Short-Cut for Variance

There is a nice short-cut to compute the variance that can be proved as an exercise:

$$\sigma^2 = \frac{1}{n-1}\sum(x-\mu)^2 = \frac{1}{n-1}\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)$$

At first the second formula looks much more complicated, but it is actually easier since it does not involve computing the mean first. In other words, using the second formula we can compute the variance (and therefore the standard deviation) without first having to compute the mean.

In our above example of machine B we would compute the variance using this shortcut as follows:

x	x^2

6	36
6	36
6	36
8	64
8	64
10	100
12	144
12	144
14	196
14	196
14	196
sum(x) = 110	**sum(x^2) = 1212**

Therefore the variance is:

$$1/ (11 - 1) * (1212 - 110^2/11) = 0.1 * (1212 - 1100) = 11.2$$

which of course is the same number as before, but a little easier to arrive at. However, Excel - as usual - provides built-in function to compute the range, the variance, and the standard deviation.

Using Excel to compute Range, Variance, and Standard Deviation

Excel provides simple formulas to compute the range, the variance, and the standard deviation:

- the Excel formula to compute the range is "=max(RANGE) - min(RANGE)"
- the Excel formula to compute the variance is "=var(RANGE)"
- the Excel formula to compute the standard deviation is "=stdev(RANGE)"

Example: Use the above formulas to compute the mean, the range, the variance, and the standard deviation of the salaries of graduates for the University of Florida. The data set (in Excel format) can be obtained by using the University of Florida Salary Levels data set we utilized beore.

All that is involved here is adding the appropriate formulas to the Excel worksheet. The results (including the formulas) are displayed below:

	A	B	C	D	E	F	G	H
1	**Gender**	**College**	**Salary**	**Graduation Date**				
2	1	7	$28,900	1.00				
3	1	7	$28,000	1.00				
5	1	7	$30,300	1.00		**Mean:**	$26,064	=AVERAGE(C2:C1101)
6	1	1	$18,000	1.00		**Range:**	$58,300	=MAX(C2:C1101)-MIN(C2:C1101)
7	0	7	$31,700	1.00		**Variance:**	$48,552,772	=VAR(C2:C1101)
8	1	3	$26,000	1.00		**Std. Dev.:**	$6,968	=STDEV(C2:C1101)
9	1	7	$25,000	1.00				

u. florida graduation salaries

Note: The variance is displayed as dollars, even though that is not correct. The correct unit for the variance, of course, is "square dollars" which does not make much sense. The standard deviation, on the other hand, has indeed dollars as unit.