Measures of Shape

Skewness and Kurtosis

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of the data includes skewness and kurtosis.

Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Computing

The **moment coefficient of skewness** of a data set is

$$\text{skewness: } g_1 = m_3 / m_2^{3/2}$$

where

$$m_3 = \sum(x-\bar{x})^3 / n \quad \text{and} \quad m_2 = \sum(x-\bar{x})^2 / n$$

$\bar{x}$ is the mean and n is the sample size, as usual. m_3 is called the **third moment** of the data set. m_2 is the **variance**, the square of the standard deviation.

sample skewness:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

http://www.tc3.edu/instruct/sbrown/stat/shape.htm

Computing

The **moment coefficient of skewness** of a data set is

$$\text{skewness: } g_1 = m_3 / m_2^{3/2} \tag{1}$$

where

$$m_3 = \sum(x-\bar{x})^3 / n \quad \text{and} \quad m_2 = \sum(x-\bar{x})^2 / n$$

$\bar{x}$ is the mean and n is the sample size, as usual. m_3 is called the **third moment** of the data set. m_2 is the **variance**, the square of the standard deviation.

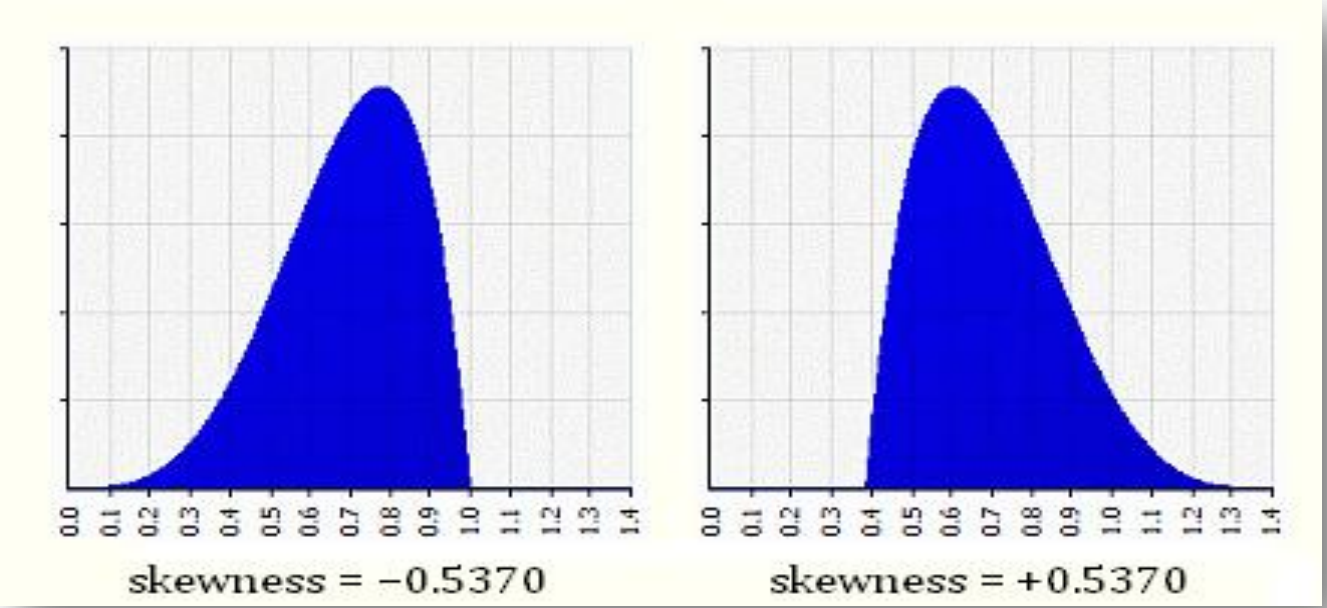
You'll remember that you have to choose one of two different measures of standard deviation, depending on whether you have data for the whole population or just a sample. The same is true of skewness. If you have the whole population, then g_1 above is the measure of skewness. But **if you have just a sample**, you need the **sample skewness**:

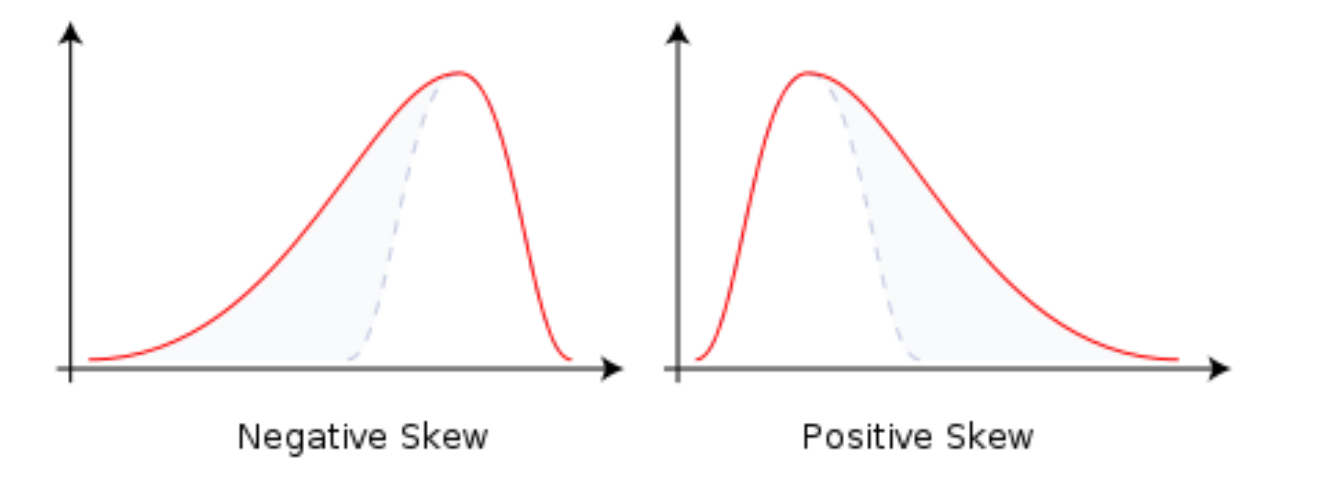
$$\text{sample skewness: } G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \tag{2}$$

source: D. N. Joanes and C. A. Gill. "Comparing Measures of Sample Skewness and Kurtosis". *The Statistician* 47(1):183–189.

Excel doesn't concern itself with whether you have a sample or a population: its measure of skewness is always G_1.

Visualizing





Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

Computing

The **moment coefficient of kurtosis** of a data set is computed almost the same way as the coefficient of skewness: just change the exponent 3 to 4 in the formulas:

$$\text{kurtosis: } a_4 = m_4 / m_2^2$$

where

$$m_4 = \sum(x-\bar{x})^4 / n \quad \text{and} \quad m_2 = \sum(x-\bar{x})^2 / n$$

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

Therefore, kurtosis: $a_4 = m_4 / m_2^2$ and excess kurtosis: $g_2 = a_4 - 3$

Note: The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

Computing

The **moment coefficient of kurtosis** of a data set is computed almost the same way as the coefficient of skewness: just change the exponent 3 to 4 in the formulas:

$$\text{kurtosis: } a_4 = m_4 / m_2^2 \quad \text{and} \quad \text{excess kurtosis: } g_2 = a_4 - 3 \tag{5}$$

where

$$m_4 = \sum(x-\bar{x})^4 / n \quad \text{and} \quad m_2 = \sum(x-\bar{x})^2 / n$$

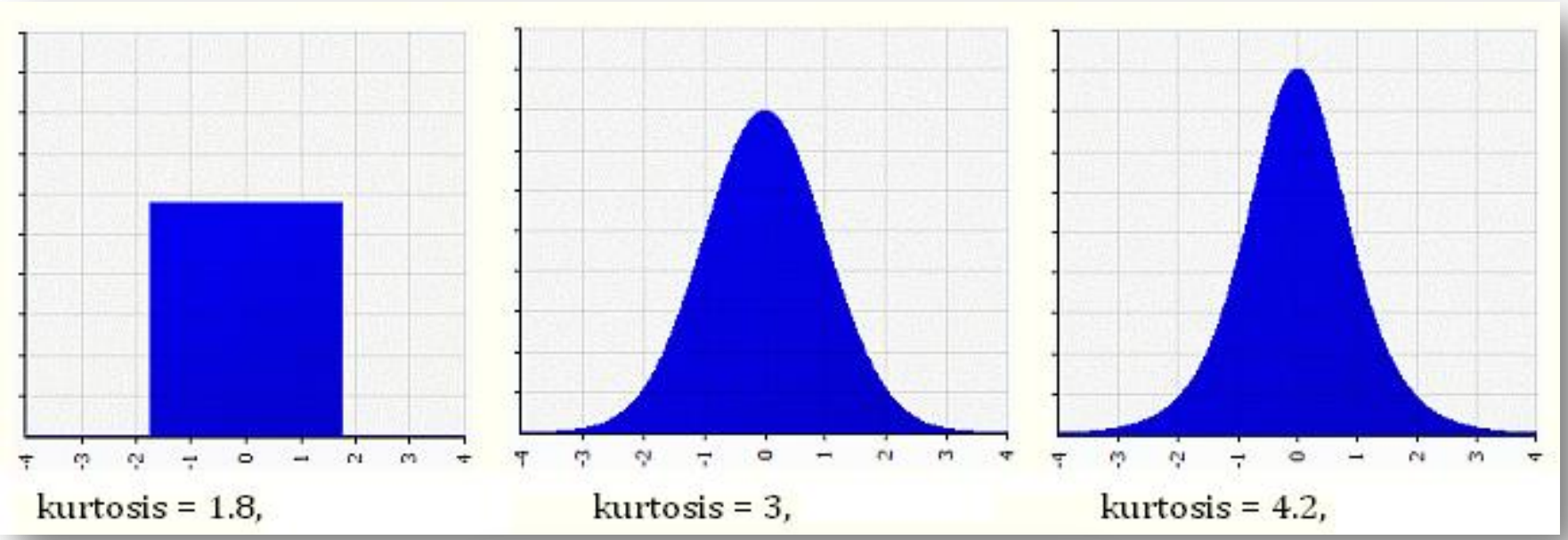
Again, the excess kurtosis is generally used because the excess kurtosis of a normal distribution is 0. $\bar{x}$ is the mean and n is the sample size, as usual. m_4 is called the **fourth moment** of the data set. m_2 is the **variance**, the square of the standard deviation.

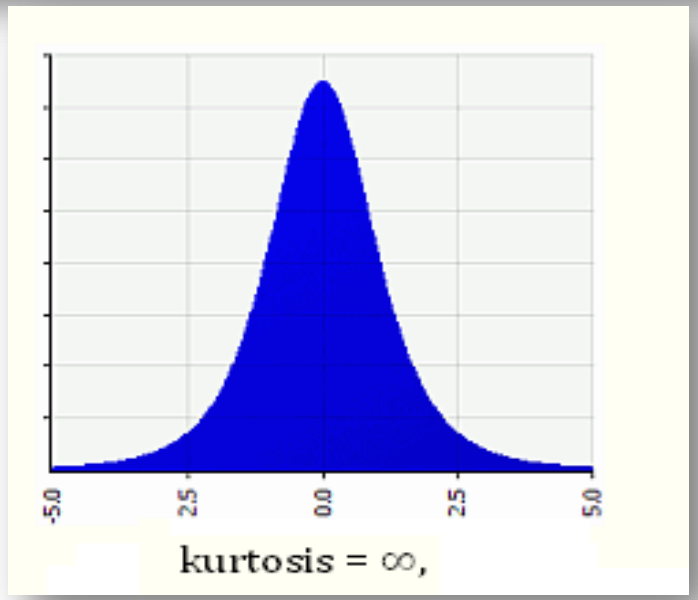
Just as with variance, standard deviation, and kurtosis, the above is the final computation if you have data for the whole population. But **if you have data for only a sample**, you have to compute the sample excess kurtosis using this formula, which comes from Joanes and Gill:

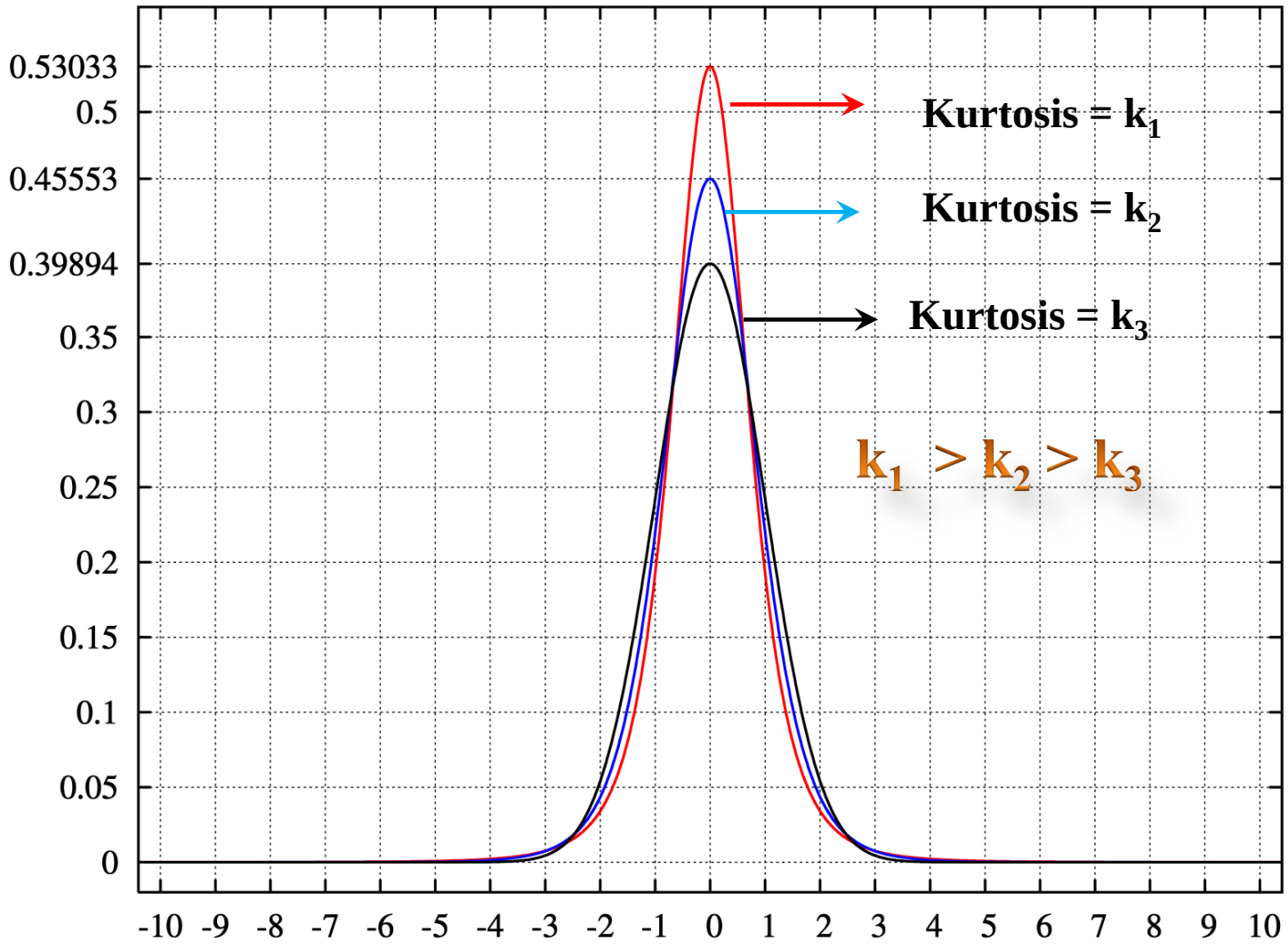
$$\text{sample excess kurtosis: } G_2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6] \tag{6}$$

Excel doesn't concern itself with whether you have a sample or a population: its measure of kurtosis is always G_2.

Visualizing






0.53033
0.5
0.45553
0.39894
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
Kurtosis = k1
Kurtosis = k2
Kurtosis = k3
k1 > k2 > k3

Example: Let's continue with the example of the college men's heights, and compute the kurtosis of the data set. n = 100, $\bar{x}$ = 67.45 inches, and the variance m_2 = 8.5275 in^2 were computed earlier.

Class Mark, x	Frequency, f	$x-\bar{x}$	$(x-\bar{x})^4 f$
61	5	-6.45	8653.84
64	18	-3.45	2550.05
67	42	-0.45	1.72
70	27	2.55	1141.63
73	8	5.55	7590.35
	Σ	n/a	19937.60
	m_4	n/a	199.3760

Finally, the kurtosis is

$$a_4 = m_4 / m_2^2 = 199.3760/8.5275^2 = 2.7418$$

and the excess kurtosis is

$$g_2 = 2.7418-3 = -0.2582$$