Steve Sawin
Statistics

Numerical Descriptive Measures

Having seen the shape of a distribution by looking at the histogram, the two most obvious questions to ask about the specific distribution is where is the data clumped and how spread out is it? Both of these are numerical, quantitative questions. They are sufficiently vague, however, that each has several reasonable answers depending on the situation.

Measures of Center: Mean and Median

Any measure of the center of a distribution can be called the "average," though in practice we usually use that term to mean the *mean.*

Definitions: The *mean* of a set of numbers is the sum of all the numbers divided by how many there are. We can write it as a formula as follows. Suppose $x_1, x_2, \ldots, x_n$ are n numbers. The mean of these n numbers is

$$(x_1 + x_2 + \cdots + x_n)/n,$$

Which we write more compactly as

$$\left(\sum_{i=1}^{n} x_i\right)/n.$$

This funny notation $\sum_{i=1}^{n}$, called *sigma notation* because $\sum$ is the Greek letter sigma, means run through each number from 1 to n, and for each number, substitute that number in for i in the formula that follows, and then add them up. We will use sigma notation frequently in the future. A nice geometric way to think about the mean is if you put a weight on the number line at each of the n values, the mean is the place where the whole ensemble would balance on the head of a pin.

The *median* of the data is (roughly) the number such that half of the data points are less than it and half are above it. We say roughly because we have to be careful when several observations have the same value. The precise way to say it is put the n observations in order from smallest to largest. Then, if n is odd pick the $(n+1)/2$th value (that is, count $(n+1)/2$ along the sequence and pick that value. If n is even, it is the average of the $n/2$th value and the $n/2+1$th value.

The book speaks of the *mode,* which is the peak of the histogram, but we will not care about that at all.

Names: The tradition is to use Greek letters for parameters, and Roman letters for statistics. When you calculate the mean of a sample (a statistic) you write it as $\overline{x}$. When you calculate the mean of a population (parameter) you write it as μ (the Greek letter "mu"). There is not really a standard terminology for medians, but the median of a sample is usually called m, and the median of a population is sometimes called M.

Properties: The mean is sensitive to outliers, meaning a few extreme values tend to pull the mean towards them.

Calculating: Both are easy to calculate by hand in small examples, both are best to do by calculator or computer for large data sets. In Excel, both show up in the "Descriptive Statistics" choice in the Data Analysis Toolkit. Both can also be calculated directly in Excel. For example, typing =MEAN(A1:A20) into a cell will put the mean of the numbers from A1 to A20 in that cell, and =MEDIAN(A1:A20) will put the median of the numbers from A1 to A20 into it. There are some other commands which are variations on these which treat blank or text values differently, which you can play around with. For large data sets the median is actually much harder to calculate than the mean, though that is rarely an issue with a computer.

Rules of Thumb: The mean is the point at which the histogram would balance, which usually makes it pretty easy to estimate from a histogram. The median is the point where half the data is below it and half above, which is just a little harder to estimate, but still not bad. For symmetric distributions they should be equal, for right skewed distributions the mean will generally be higher (and both will generally be above the peak) while for left skewed the mean is lower (and both are lower than the peak).

Use: In most situations the mean and median are very close and both fit our sense of the middle or typical value pretty well. In highly skewed data they can be very different, and it is generally less clear what the middle of the data should be. Generally, in highly skewed data the median is closer to our sense of what the middle should be, and people tend to use medians for skewed distributions like income and prices. However, the mean is much nicer mathematically, and this makes it a more practical quantity to deal with most of the time in inferential statistics.

An important point that occasionally comes up is that because the median depends only on the ordering of the values, it makes sense even for ranked data, while the mean, which involves adding and dividing, does not. You should never use the mean to summarize ranked data, though people often do.

Measures of Variation

Variation is an extremely important notion. Quality Control and most of the modern applications of statistics to business and engineering focuses on reducing variation, the enemy of planning. In science, variation is error, the thing that stands between your measurements and the true answer. We will consider five measures of variation. One the *range,* is of little significance, though very simple. Two are nearly the same thing, the *populations standard deviation* and the *sample standard deviation* have almost the same definition and, in any case where they are at all useful, practically the same value. The last two, *population variance* and *sample variance,* are the squares of the two standard deviations, so they are just a repackaging of the same information.

Definitions: The *range* is the difference between the largest and the smallest value. Since this only tells you about the largest and smallest values, it is generally not very useful.

The *population variance* is defined by the formula

$$\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)/n = \left((x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2\right)/n$$

where $x_1, x_2, \ldots, x_n$ are your observations and μ is the mean of these numbers. In other words, for each number we take the difference between it and the mean (this can be seen as its distance from the center), square it (so that it is positive whether the value is smaller or larger than the mean), and then average these quantities. It is easy to see that this number gets bigger as the data gets more spread out, does not change if you add a constant to all of them (that is, shift the whole histogram to the left or right without changing shape) and does not change if you add lots of more numbers that are equally spread out. So it is a good measure of variation.

One problem with the population variance is that it does not *scale* properly. If you double all your numbers, you do not double the variance (in fact you multiply by four). Another way to say this is that if your data has units like inches (e.g., if it represents heights) then the variance would have units of square inches (like an area). To solve this we take the square root. This is called the *population standard deviation:*

$$\sqrt{\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)/n.}$$

It is generally considered a better way to report the variation in a population.

For technical reasons we shall ignore for now, one uses slightly different formulas for computing the standard deviation and variance for samples. The *sample variance* is given by

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

and the *sample standard deviation* is given by

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

Notice the only difference is you are calling the mean $\overline{x}$ rather than μ and are dividing by $n-1$ rather than n. Since the standard deviation and variance really don't tell you much unless you have a lot of data, these two quantities are generally extremely close. If someone asks for "the standard deviation" without specifying which, they mean the sample standard deviation.

Names: The population standard deviation is called σ (The Greek lower case sigma, you already met its upper case cousin) and so of course the population variance is called σ^2. The sample standard deviation is called s, and the sample variance is then s^2.

Properties: Standard deviation and variance are both sensitive to outliers, and adding an outlier will generally bump these numbers up significantly. The standard deviation can generally be interpreted as the typical distance from the mean of a random point. For what it is worth, if you spun the histogram around the mean (its balancing point remember) the standard deviation tells you how hard a push you would need to get it spinning.

Calculating: To calculate standard deviation or variance by hand, first calculate the mean. then make a column of the observations, then make a column of each observation minus the mean, then make a column of the squares of these differences. Then just add these numbers in the last column up, divide by n or $n-1$, and take the square root or not. Except in the simplest case though, you are better off using Excel. The Descriptive Statistics option will give you (sample) standard deviation and sample variance (and range too). If you want the population standard deviation of cells A1 through A20, use =STDEVP(A1:A20). Population variance is =VARP(A1:A20). You can also do the sample s.d. and variance directly with =STDEV(A1:A20) and =VAR(A1:A20).

Rules of Thumb: The best way to think of the standard deviation is as a sort of typical distance from the mean. In particular, if your data is bell-shaped, then roughly 68% of the data will be "within one standard deviation of the mean." That is, 68% of the data will be greater than $\mu-\sigma$, the point a distance σ below the mean μ, and μ_σ, the point a distance σ above the mean μ. Similarly, 95% of the data will fall within two standard deviations of the mean (between $\mu - 2\sigma$ and $\mu + 2\sigma$) and over 99% will fall within three standard deviations of the mean. This means the standard deviation and the mean are a sort of universal measure of how unusual and observation is. Someone whose height is one standard deviation above the mean is tall, but not surprising. Someone two standard deviations above the mean is strikingly tall, you would look twice. If you meet someone who is three standard deviations above the mean, you will stare, and then go home and tell your roommate about it.

This rule of thumb is technically only true for population standard deviation, but it works well enough for both. Even of the data is far from bell shaped, at least 3/4 of the data falls within two standard deviations of the mean, so you can get some idea of the standard deviation by looking at the histogram.

Use: The standard deviation is most useful when your data is roughly bell shaped. When it is skew or otherwise far from bell-shaped, it is more difficult to interpret. Generally we will be interested in looking at the distribution "scaled by the mean and s.d. that is to say, we will talk about points one standard deviation above the mean, or two standard deviations below the mean, or whatever.

In fact, the best universal measure of where a data point fits in a bell-shaped distribution is its *z-score.* If x is a number in a distribution of mean μ and standard deviation σ, its z-score is

$$z = \frac{x-\mu}{\sigma}.$$

A z-score of 2 means the data point is two-standard deviations above the mean.

The variance will occasionally be useful to talk about, but it will really be a kind of helpful understudy to the standard deviation.