

Statistical Analysis

Note: The word “data” is plural. “Datum” is the singular form. “The data show that...” is correct, NOT “The data shows that...”

Is the central tendency of data the only thing to describe?

No, we might want to describe the variance of the data – how much they deviate from the center point. What is the shape of this graph? Is it a gradual slope that is spread wide around the mean, or does the shape of the graph rise steeply, with most of the data clustered around the mean?

The two graphs show the same mean but with data spread out differently. The way we measure spread is with the number called the “standard deviation,” which tells you how far your data points have spread away from the mean. It’s sometimes called the “root-mean-square”: the square root of the mean squared deviation of a quantity from a given baseline (usually the mean of your data). Yeah, right. Okay, let’s try again:

1. First, find how much all your data points are different from the mean.
(Take each of your numbers and subtract the mean from them, one at a time, so you have a bunch of differences from the mean.)
 2. Second, square all these differences. (SQUARE)
(Why? Because the ones below the mean are negative, and above the mean are positive, and they would just cancel out if you added them together and that would eliminate any hope of finding how far the numbers are from the mean.)
 3. Now, add up all the differences and divide by the number of data points you had. (MEAN)
 4. Lastly, take the square root of that number. (ROOT) This way you get back a reasonable-sized number that works in your data, and is not too big like the squares would have been.
- Putting it all together, you have just found the ROOT-MEAN-SQUARE, or standard deviation.

The formula:

$$S = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

where n = number of data points, x_i is any one particular datum,
and \bar{x} is the mean of all the data.

Please be aware that the standard deviation can be found differently for different situations. We are giving this formula for the case when you have all the data for a certain population. On Microsoft Excel, the formula command is “stdevp(cell:cell)”, the “p” standing for “population”. The formula is slightly different if you are taking a sample from a larger population and estimating the standard deviation from that sample.

Special notes:

Standard deviation should only be used when the mean is chosen as the measure of center. Outliers in the data will strongly influence the standard deviation. Do not use this if your data is very skewed because the spread will be very different on left and right.

One helpful use of standard deviation is that for a “normal” distribution (meaning a bell curve), approximately two-thirds of the data will lie in between the standard deviation above the mean (\bar{x}) and the standard deviation below the mean.

$(\bar{x} + s) - (\bar{x} - s) = \frac{2}{3}$ of the data.

Analysis Chart

Basic analysis of numerical data includes finding the mean, median, or mode (whichever measure of central tendency is most appropriate for your data), the upper and lower extremes, and computing the percentage difference between the mean or medians of variables being compared. More advanced would be to include the standard deviation and a description of the spread, or general shape, of the data when graphed.

A good way to show both the central tendency and the spread of the data is to create a chart showing:

<u>Minimum</u>	<u>Q1</u>	<u>Median (Q2)</u>	<u>Q3</u>	<u>Maximum</u>	<u>St. Dev.</u>
----------------	-----------	--------------------	-----------	----------------	-----------------

Graphs

Hopefully the graph will be able to relate the independent variable (X – what you changed, on the horizontal axis) and the dependent variable (Y – what you measured as an effect, on the vertical axis). Note that sometimes you can make a nice connection between your hypothesis and the graph: “If I change X, then I can measure a trend in Y.” A viewer of your display should be able to make a quick connection between your hypothesis and your graph.

When you are displaying your data, you might wish to use a stem plot and a box and whisker plot. Stem plots, like histograms, quickly show if your data clusters around a set of numbers. You can use stem plots to show two sets of data like this:

Box and whisker plots are especially helpful for quickly giving the viewer a comparison of two or more groups of data. For example, if you were comparing the math test scores of different groups of people, each group of people would be its own box and whisker plot on a common scale for all:

Note that in box and whisker plots, exactly one fourth of all the data points are in each section. Therefore, box and whisker plots quickly tell you where half the data lie. (In

the interquartile range, the “box”.) They also tell you the median and whether the data is evenly distributed around the median or if it is skewed higher or lower. This can be valuable for quick visual comparison. You can also describe the way the data is clustered in your written analysis.

Box and whisker plots also enable you to define whether you have an outlier in your data or not. You can tell if one of the extremes is an outlier by multiplying the interquartile range (the third quartile, Q3, minus the first quartile, Q1) by 1.5. Let’s call that number “a”. If your datum in question is $< Q1 - a$, or is $> Q3 + a$, then it is an outlier and could be legitimately thrown out of the data for calculation of the mean.

.....
What if the data is not numbers, but just a record of how many responses are in particular groups? For example, “10 people like dogs while 14 people like cats” cannot be analyzed with mean, median, standard deviation, etc. In this case you would present the data in a bar chart and a pie chart, and calculate the percent differences between the quantities.