Chapter 905

Standard Deviation Estimator

Introduction

Even though it is not of primary interest, an estimate of the *standard deviation* (*SD*) is needed when calculating the power or sample size of an experiment involving one or more means. Finding such an estimate is difficult not only because the estimate is required before the data are available, but also because the interpretation of the standard deviation is vague and our experience with it may be low. How do you estimate a quantity without data and without a clear understand of what the quantity is? This section will acquaint you with the standard deviation and offer several ways to obtain a rough estimate of it before the experiment begins using the Standard Deviation Estimator.

The Standard Deviation Estimator can also be used to calculate the standard deviation of the means, a quantity used in estimating sample sizes in analysis of variance designs.

Understanding the Standard Deviation

It is difficult to understand the standard deviation solely from the standard deviation formula. There are two general interpretations that can be useful in understanding the standard deviation.

1. The standard deviation may be thought of as the average difference between an observation and the mean, ignoring the sign.
2. The standard deviation may be thought of as the average difference between any two data values, ignoring the sign.

The population standard deviation is calculated using the formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \mu\right)^2}{N}}$$

where N is the number of items in the population, X is the variable being measured, and μ is the mean of X. This formula indicates that the standard deviation is the square root of an average. This average is the average of the squared differences between each value and the mean. The differences are squared to remove the sign so that negative values will not cancel out positive values. After summing up these squared differences and dividing by N, the square root is taken to give the result in the original scale. That is, the standard deviation can be thought of as the average difference between the data values and their mean (the terms mean and average are used interchangeably).

Example

Consider the following two sets of numbers

A: 1, 5, 9

B: 4, 5, 6

Both sets have the same mean of 5. However, their standard deviations are quite different. Subtracting the mean and squaring the three items in each set results in

Set A

(1-5)(1-4) = 16

(5-5)(5-5) = 0

(9-5)(9-5) = 16

Sum = 32

$$SD_A = \sqrt{\frac{32}{3}} = 3.266$$

Set B

(4-5)(4-5) = 1

(5-5)(5-5) = 0

(6-5)(6-5) = 1

Sum = 2

$$SD_B = \sqrt{\frac{2}{3}} = 0.8165$$

The standard deviations show that the data in set A vary more than the data in set B.

Divide by N or N-1?

Note in the example above that we are dividing by *N*, not *N-1* as is usually seen in standard deviation calculations. When the standard deviation is computed using all values in the population, *N* is used as the divisor. However, when the standard deviation is calculated from a sample, *N-1* is used as the divisor. We stress that the results for a sample by using the lower-case *n* and naming the *sample standard deviation S*. The value of *S* is computed from a sample of *n* values using the formula

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

The *n-1* is used instead of *n* to correct for bias that statisticians have discovered. That is, over the long run, dividing by *n-1* provides a better estimate of the true standard deviation than does dividing by *n*. Although we divide by *n-1* rather than *n* for sample standard deviations, we recommend that for purposes of interpretation, the divisor is assumed to be *n*, so that the operation can be thought of as computing an average.

Average Absolute Deviation

If we were to devise a measure of variability with no previous experience, we might first consider the average absolute deviation (AD), sometimes called the mean absolute deviation, or MAD, which is computed by forming the deviations from the mean, taking their absolute values, and computing their average. The absolute value is applied to remove the negative signs, which, in turn, avoids the cancellation of values when the average is taken. The formula for AD is

$$AD = \frac{\sum_{i=1}^{N} \left| X_i - M \right|}{N}$$

This simple average of absolute deviations is much easier to understand, but is very difficult to work with mathematically. Opposingly, the standard deviation is more difficult to interpret directly, but it can be worked with mathematically in statistical problems. The ability to work with the standard deviation mathematically outweighs its deficiency in interpretation. Hence, we generally use the standard deviation rather than the average absolute deviation in practice.

Comparing Average Absolute Deviation and Standard Deviation

Fortunately, the average absolute deviation and the standard deviation are usually close in value. Mathematically, it can be shown that *AD* is always less than or equal to *SD*. A small simulation study is summarized below. It shows the relationship between *AD* and *SD* for data generated from various distributions.

Distribution	Percent SD > AD	Characteristics
Uniform	15%	Level
Normal	20%	Bell-Shaped
Gamma(5)	30%	Moderately Skewed Right
Gamma(5)^2	45%	Extremely Skewed Right

These distributions were selected for study because they represent a wide range of possibilities. The table shows that, for typical datasets, the standard deviation is from 15 to 30 percent larger than the average absolute deviation. And in the case of the normal distribution, the *SD* is about 20% higher than *AD*.

Hence, for planning purposes, you can think of the standard deviation as an inflated version of the average absolute deviation.

Example

In out example, we can compute *AD* for datasets A and B as follows.

Set A

$AD_A = (4+0+4)/3 = 8/3 = 2.667$. Recall that $SD_A = 3.266$.

Set B

$AD_B = (1+0+1)/3 = 2/3 = 0.667$. Recall that $SD_B = 0.8165$.

We see that the values are similar. The degree of difference is likely within the error that we would expect during the planning phase.

Standard Deviation as the Average Difference between Values

The above discussion and formula have pointed out that the standard deviation may be thought of as an average deviation from the mean. In this section, a second interpretation of the standard deviation will be given.

We can manipulate the formula for the sum of squared deviations to show that

$$\sum_{i=1}^{N}\left(X_i - M\right)^2 = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_i - X_j\right)^2}{N}$$

This formula shows that the squared deviations from the mean are proportional to the squared deviations of each observation from every other observation. Note that the mean is not involved in the expression on the right.

Using the above relationship, the standard deviation may be calculated using the formula

$$SD = \sqrt{\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(X_i - X_j\right)^2 / N}{N}}$$

Example

Consider again our simple example of sets A and B. Applying this operation to the three possible pairs of the data in set A (1, 5, 9) gives

(1-5)(1-5) = 16

(1-9)(1-9) = 64

(5-9)(5-9) = 16

The sum is 96. Dividing 96 by 3 (the number of pairs) again yields 32. Hence, the standard deviation is computed as *SD* = SQRT(96/9) = 3.266 (which matches the previous result).

Likewise, for set B (4, 5, 6), the formula results in

(4-5)(4-5) = 1

(4-6)(4-6) = 4

(5-6)(5-6) = 1

so that SD = SQRT(6/9) = 0.8165.

Estimating the Standard Deviation

Our task is to find a rough estimate of the standard deviation. Several possible methods are available in the Standard Deviation Estimator procedure which may be loaded from the PASS-Other menu. ***PASS*** provides a panel that implements each of these methods for you.

Data Tab – Standard Deviation from Data Values

One method of estimating the standard deviation is to put in a typical set of values and calculate the standard deviation.

This window is also used when you need the standard deviation of a set of hypothesized means in an analysis of variance sample size study.

Pros and Cons of This Method

This method lets you experiment with several different data values. It lets you determine the influence of different data configurations on the standard deviation. In so doing, you can come up with a likely range of SD values.

However, investigators tend to pick trial numbers that are closer to the mean and more uniform than will result in practice. This results in SD's that are underestimated. If you use this method, you should be careful that your range of possible SD values is wide enough to be accurate.

Example 1 – SD for a Set of Values

As an example, suppose that you decide that the following values represent a typical set of data that you would anticipate for one group of individuals:

10, 12, 14, 10, 11, 10, 12, 13, 9, 13, 15, 11

To calculate the appropriate standard deviation, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **Data tab**.
2. The order that the data are entered in does not matter. However, to show the use of the Counts column, we count up the number of times each value occurs. The values and their frequency counts are then entered into the **Values** and **Counts** columns. The data entry goes as follows:
   9 1
   10 3
   11 2
   12 2
   13 2
   14 1
   15 1
3. Check the **Use N-1 as divisor** box. (We use the N-1 divisor when estimating sigma from a set of data.)
4. Press the **Calculate N, Mean, SD from Values** button. The standard deviation is 1.825742. We might round this value up to 2.0 for planning purposes.

Example 2 – SD for a Set of Means

In this example, we will show you how to obtain the standard deviation of a set of hypothesized means. Care must be taken that you select the correct divisor—*N*, not *N*-1.

In this example, a researcher is studying the influence of a drug on heart rate. He estimates that the average heart rate of his group without the drug is 80. His experimental design will apply three different doses. The first dose is expected to lower the heart rate by 10%, the second by 20%, and the third by 30%. Hence, the hypothesized means for the four groups are 80, 80(0.9) = 72, 80(0.8) = 64, and 80(0.7) = 56.

To calculate the appropriate standard deviation, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **Data tab**.
2. Enter the four means into the **Values** column. The Counts column is left blank.
   80
   72
   64
   56
3. Make sure the **Use N-1 as divisor** box is not checked since we want the population standard deviation.
4. Press the **Calculate N, Mean, SD from Values** button. The standard deviation of the means is 8.944272.

Standard Error Tab – Standard Deviation from Standard Error

If the value of the standard error of the mean is available from another experiment, it may be used to estimate the standard deviation. The formula estimating the standard deviation from the standard error is

$$SD = SE\sqrt{N}$$

where N is the sample size.

Pros and Cons of This Method

This method is only useful when you have a standard error value available.

Example

To calculate the appropriate standard deviation when a previous study of 23 individuals had a standard error of the mean of 2.7984, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **Standard Error tab**.
2. Enter **23** for **N**.
3. Enter **2.7984** for **Standard Error**.
4. Press the **Calculate Standard Deviation** button. The standard deviation is 1.825742. We might round this value up to 2.0 for planning purposes.

Range Tab – Standard Deviation from Population or Sample Data Range

There are two cases in which the range may be used to estimate the standard deviation. In the first case, the sample size and data range may be available from a previous study. In the second case, a reasonable estimate of the population range may be obtainable. This window allows you to estimate SD in both of these situations.

The basic formula for estimating the standard deviation from the range is

$$SD = \frac{Range}{C}$$

where C is determined by the situation.

Determining C

If the population range can be established, it may be used to estimate sigma by dividing by an appropriate constant. To determine an appropriate value of the constant, statisticians use the fact that most of the data is contained within three standard deviations of the mean—so they set C to six. However, consultants have found that for some reasons (e.g., understated range, or non-normal population) dividing by six tends to understate the standard deviation. So they divide by five or even four. Dividing by a smaller number increases the estimated standard deviation. Our recommendation is to divide by four. To use this method, enter the divisor (4, 5, or 6) in the C (Divisor) box.

If the data range is available from a previous study, the constant C may be calculated as the median of the distribution of the range for that sample range. This distribution assumes that the data themselves are normally distributed. The median of the distribution of the range is calculated in NCSS using numerical methods, and similar calculations can be made using the Probability Calculator. To use this method, enter the data range and the sample size in the lower region of the Range tab.

Pros and Cons of This Method

The range is a poor substitute for having the standard deviation. This method should be used as a ‘last resort’.

Example 1 – Population Range 'Known'

To calculate an estimate of the standard deviation when the population range is known to be 150, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **Range tab**.
2. Enter **150** for **Range (of Population)** in the upper region of the tab.
3. Enter **4** for **C (Divisor)**.
4. Press the **Calculate Standard Deviation** button. The estimate of the standard deviation is 37.5. The value of C used is 4.

Example 2 – Previous Sample Available

To calculate an estimate of the standard deviation when a previous study of 20 animals had a minimum value of 15.3 and a maximum of 18.7, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **Range tab**.
2. Enter **20** for **N (Sample Size)** in the lower region of the tab.
3. Enter **3.4** for **Range (of Sample Data)** in the lower region of the tab. This is 18.7 minus 15.3.
4. Press the **Calculate Standard Deviation** button. The estimate of the standard deviation is 0.92243. The value of C used is 3.685916.

Percentiles Tab – Standard Deviation from Percentiles

If you are willing to assume that the population values are normally distributed (bell-shaped), you can use the values of two percentiles to estimate the standard deviation.

The basic formula for estimating the standard deviation from two percentiles is

$$SD = \left| \frac{X2 - X1}{Z\left(\frac{P2}{100}\right) - Z\left(\frac{P1}{100}\right)} \right|$$

where *X1* and *X2* are the two percentiles, *P1* and *P2* are the two percentages, and *Z(P)* is the standard normal deviate that has a tail area of *P* to the left.

Pros and Cons of This Method

This method works well if you have two accurate percentiles available and the underlying distribution of the data is normal.

Example

To calculate an estimate of the standard deviation when you know that the 25th percentile of the population is 80.25, the 75th percentile of the population is 116.38, and that the population is normally distributed, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **Percentiles tab**.
2. Enter **25** for **Percentage 1**.
3. Enter **80.25** for **Percentile Value 1**.
4. Enter **75** for **Percentage 2**.

5. Enter **116.38** for **Percentile Value 2**.
6. Press the **Calculate Standard Deviation** button. The estimate of the standard deviation is 26.78321.

COV Tab – Standard Deviation from Coefficient of Variation

The coefficient of variation (*COV*) is equal to SD divided by the mean. Hence, if you know the coefficient of variation and the mean, you can estimate SD.

Note that t-tests and the analysis of variance assume that the standard deviations are equal for all groups. If the standard deviations are proportional to their group means, you should use a data transformation (such as the square root or the logarithm) to make the standard deviations equal.

The basic formula for estimating the standard deviation from the *COV* is

$$SD = (COV)(Mean)$$

Pros and Cons of This Method

This method works well if you have estimates of the mean and COV.

Example

To calculate an estimate of the standard deviation when you know that the mean is 127 and the COV is 0.832, do the following:

1. Load the **Standard Deviation Estimator** window and click on the **COV tab**.
2. Enter **0.832** for **Coefficient of Variation (COV)**.
3. Enter **127** for **Mean**.
4. Press the **Calculate Standard Deviation** button. The estimate of the standard deviation is 105.664.

Confidence Limits Tab – Confidence Limits of the Standard Deviation

Often, you will obtain an estimate of the standard deviation from a previous study or a pilot study. Since this estimate is based on a sample, it is important to understand its precision. This can easily be calculated since the square of the sample standard deviation follows a chi-squared distribution. This confidence interval does assume that the population you are sampling from is normally distributed.

Once a confidence interval has been obtained, it would be wise to enter both values (the confidence limits) into the appropriate place in the sample size calculations to provide a range of possible sample size values (or of statistical power).

Example

Suppose a pilot study of 10 individuals yields a standard deviation of 83.21. Calculate a 95% confidence interval for the population standard deviation using these results.

1. Load the **Standard Deviation Estimator** window and click on the **Confidence Limits** tab.
2. Enter **10** for **N**.
3. Enter **83.21** for **Standard Deviation**.
4. Enter **0.05** for **Alpha**.
5. Press the **Calculate Confidence Limits** button. The confidence limits are 57.23477 and 151.909.