

Why Divide By n-1?

by Keith M. Bower

When computing the sample standard deviation, Six Sigma practitioners frequently ask why the denominator uses n-1 instead of the sample size, n. Though a relatively intuitive mathematical answer is available, the question leads to a far more important discussion with direct relevance to understanding the practice of statistics in quality improvement.

The standard deviation

Karl Pearson developed the standard deviation in 1893, showing that when an entire population is available for assessment, the population standard deviation (σ) takes the form in equation (1).¹

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

In practice, obtaining measurements for all N values in the population may not be feasible, so a sample of size n is taken from the population. These n observations may then be used to compute s (an estimate of σ) using equation (2).²

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

The use of n-1 in the denominator causes confusion for many Six Sigma practitioners. Why not simply use the number of observations, as in equation (1)? The answer lies in considering the amount of information required to compute s.

Example

Consider a random sample of size $n = 3$ having been drawn from a population of size N where $n < N$. The three datapoints are 2, 4, and 6. Compute the sample mean (\bar{x}) by adding the three observations, then dividing by the sample size of 3, reaching the answer $\bar{x} = 4$.

Of course, all three datapoints in the sample are required to compute \bar{x} ; hence, 3 is used in the denominator.

To understand how to compute s , consider the three columns in Table 1.

Table 1

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-2	4
4	0	0
6	2	4

Using equation (2), we find that $S = \sqrt{\frac{4+0+4}{2}} = \sqrt{4} = 2$

It is important to note that column two in Table 1 sums to zero. This will always be the case, as proven in the Appendix. Therefore, not all three rows of data are required.

For example, based on the results from the first two rows in column 2 (i.e., -2 and 0), the final row must be 2 for that column to sum to zero. This is one explanation as to why a single *degree of freedom* (a measure of the amount of information available for estimating the population variance) is lost when computing s ; hence, $n-1$ is used in the denominator in equation (2).

With large sample sizes the practical difference between using n or $n-1$ is trivial. Of far more important consideration, however, is whether σ would ever be used in practice, or is s always to be used, even when 100% inspection occurs?

Types of statistical studies

A hypothetical example illustrates three types of statistical studies: descriptive, enumerative, and analytic.³ Consider a large barrel containing several varieties of fish, including goldfish. Each of the studies offers a method for computing the proportion of all fish in the barrel that are goldfish.

Scenario one—a descriptive study

If it is possible to classify all fish in the barrel and the assessment is isolated to just one barrel, the study will be *descriptive*. There is no wider inference, as the barrel contains the entire population of fish. The proportion of goldfish thereby obtained will be the population proportion, and no statistical inference procedures are necessary, nor would be sensible.

Scenario two—an enumerative study

If it is not possible to assess all fish in the barrel, a smaller, representative sample may be drawn from it. The more manageable sample would provide a point estimate of the true proportion of goldfish. This would be an *enumerative* study; many statistical procedures are based on this type of sampling scenario.

Scenario three—an analytic study

If the analyst is concerned with the *process* that places these fish in barrels, then the population to be assessed not only includes this barrel, but all barrels in the future, as well as those from the past. This would be considered an *analytic* study. Even if all fish could be assessed, as they were in Scenario One, the barrel must still be regarded as a sample from the wider population.

Six Sigma practitioners are concerned with improving processes. They therefore perform analytic studies to understand how a process behaves. For many statistical procedures in widespread use to have validity, the process requires evidence of stability. When a process exhibits stability, statistical procedures based on enumerative principles may have legitimacy.

From a practical perspective, even if 100% inspection occurs on a process, the use of $n-1$ in the estimate of σ (by using s) is appropriate when performing an analytic study.

Summary

Six Sigma practitioners must understand not only the statistical tools that may be useful for process improvement, but also potential restrictions on their use. By considering the distinctions between descriptive, enumerative, and analytic studies, practitioners can consider the validity of results from the statistical procedures they use.

References

1. Helen M. Walker, *Studies in the History of Statistical Method* (Baltimore: Williams & Wilkins, 1929), 188.
2. Robert V. Hogg and Johannes Ledolter, *Applied Statistics for Engineers and Physical Scientists*, 2nd ed. (New York: Macmillan, 1992), 23.
3. For more information on descriptive, enumerative and analytic studies, see Ralph L. Liberatore, "Teaching the Role of SPC in Industrial Statistics," *Quality Progress*, July 2001, 89-94.

Appendix

Proof that $\sum_{i=1}^n (x_i - \bar{x}) = 0$

$$(i) \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i) - \sum_{i=1}^n (\bar{x})$$

$$(ii) \sum_{i=1}^n (x_i) - \sum_{i=1}^n (\bar{x}) = \sum_{i=1}^n (x_i) - \sum_{i=1}^n \left\{ \frac{\sum_{i=1}^n (x_i)}{n} \right\}$$

$$(iii) \sum_{i=1}^n (x_i) - \sum_{i=1}^n \left\{ \frac{\sum_{i=1}^n (x_i)}{n} \right\} = \sum_{i=1}^n (x_i) - n \times \left\{ \frac{\sum_{i=1}^n (x_i)}{n} \right\}$$

$$(iv) \sum_{i=1}^n (x_i) - n \times \left\{ \frac{\sum_{i=1}^n (x_i)}{n} \right\} = \sum_{i=1}^n (x_i) - \sum_{i=1}^n (x_i) = 0$$

About the author

Keith M. Bower is a statistician and webmaster for www.KeithBower.com, a site devoted to providing access to online learning materials for quality improvement using statistical methods. He received a bachelor's degree in mathematics with economics from Strathclyde University in Great Britain and a master's degree in quality management and productivity from the University of Iowa in Iowa City, USA. He is a member of ASQ and the Six Sigma Forum.

Copyright © 2005 American Society for Quality. All rights reserved.