IVT NETWORK
INSTITUTE OF VALIDATION TECHNOLOGY | an informa business
Unlimited Validation Knowledge Awaits...
Published on *IVT Network* (http://www.ivtnetwork.com)

SR #2. Data Variance, Central Tendency, and Measurement Calculations

By **Alan M Golden**

Dec 23, 2021 8:00 am EST

Peer Reviewed

Statistics Roundtable (SR) is an IVT feature that provides readers an opportunity to discuss information about statistics and demonstrate the application of statistics to pharmaceutical problems. The use of statistics is fundamental in regulated industries; the statistical basis for decision-making is an expectation. Any effort to increase the understanding and application of statistics in daily work life will be useful to readers.

The potential scope of SR content is extensive; statistics applications in validation and QA activities are numerous. Our goal in SR is to provide basic understanding of statistics concepts specific to pharmaceutical applications. We will discuss statistics concepts in validation and QA in simple language, and then demonstrate their use in example problems. Readers have opined their preference for case studies describing practical applications of theory; we intend to emphasize representative problem applications.

Comments from readers are needed to help us fulfill our objective for this column. SR will be most successful when validation, quality, and statistics communities participate in this endeavor. Suggestions for future discussion topics are invited. Readers are also asked to contribute manuscripts for publication – please share your successful experiences with others. Please contact column coordinator Jeremy Ebersole via the comments section below with questions, suggestions, or topics for discussion.

INTRODUCTION

This paper examines basic concepts in statistics and their application to validation. It explores the definition of statistics, why statistics are important in validation, and the regulatory requirements for statistics in processes. It further addresses the concept and measurement of variance, data distribution, and the relationship of variance to validation. Ways to measure variance, and when different types of measurements are appropriate are described.

Statistics Definition and Objective

Statistics is the practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample. What this means practically is that statistics enables use of a limited amount of data to predict the outcome of an event, the result of a process,

or a way to describe an entire population – we observe a relatively small sample of the population and make judgments applicable to the larger whole.

In validation, we are asked to predict the future behavior of a system from a limited look at that system. How do we know with any degree of certainty that the system will continue to function as we want (or need) as we continue to use the system? We need to show with a high degree of certainty that our product or process is going to meet its requirements. In cleaning, we need to be confident that every time we clean, the cleaning process has produced the desired results. For test methods, we need to have confidence in the results of our tests. In processes or manufacturing validation, we need to have confidence that every time we run our process, we will deliver acceptable product within specification. Our customers demand it, regulatory agencies require it, and our ethics make it imperative. By validation of our processes and test methods, we take a big step toward meeting these requirements. The proper use and application of statistics in validation is a powerful tool to ensure we meet our obligations to produce safe and reliable products.

Regulatory Requirements

There are regulatory requirements to use proper statistical techniques in validation. Some of the most pertinent FDA guidance and regulations are outlined below.

Process Validation: General Principles and Practices

- The sampling plan must result in statistical confidence (§ 211.165(c) and (d)); and the batch must meet its predetermined specifications (§ 211.165(a)).
- The second principle in this regulation further requires that in-process specifications "…shall be derived from previous acceptable process average and process variability estimates where possible and determined by the application of suitable statistical procedures where appropriate." This requirement, in part, establishes the need for manufacturers to analyze process performance and control batch-to-batch variability.

Bioanalytical Method Validation

Response function is determined by appropriate statistical tests based on the actual standard points during each run in the validation.

21CFR Part 820, Subpart O--Statistical Techniques, §820.250

- (a) Where appropriate, each manufacturer shall establish and maintain procedures for identifying valid statistical techniques required for establishing, controlling, and verifying the acceptability of process capability and product characteristics.
- (b) Sampling plans, when used, shall be written, and based on a valid statistical rationale. Each manufacturer shall establish and maintain procedures to ensure that sampling methods are adequate for their intended use and to ensure that when changes occur the sampling plans are reviewed. These activities shall be documented.

VARIANCE

In validation, our most important goal is the understanding and control of variance. The definition of variance is the fact or quality of being different, divergent, or inconsistent. Variance is inherent in all processes or test methods. In other words, it is what happens when testing the same sample one hundred times or cleaning the same surface one hundred times and getting one hundred different answers or results. The questions for validation are how to measure the difference and more importantly, does the inherent variation in the system exceed what my product or process can tolerate. How much variation it tolerable is determined by the overall risk profile of our process or product. What are we trying to do and why? Does our process need to be sterile or just particulate clean? What are we measuring and what test are we using? The higher the risk, the less variability our process or testing can tolerate.

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of data are spread out from their average value.

Sources of Variation

There are many sources of variance. The following list is a short example of some of the more common sources of variance encountered in validation. Understanding the variance in your system is key to successful validation.

- **Equipment.**Even if the equipment used is from the same manufacturer and same model number, due to production variation at the equipment manufacturer, there will be enough variation in individual pieces of equipment to yield different results even when reading the same sample. This is part of the reason equipment needs to have a complete installation qualification, operational qualification, and if needed a process qualification (IQ/OP/PQ) before use. This will give the expected variation in any given piece of equipment and a determination if it is suitable for use in your process.
- **Location.**The location of your equipment or process can play a big part in the amount of variance you observe. Changes in temperature, humidity, atmospheric pressure, vibration can all affect the results of process or test. This is why revalidation (IQ/OQ) is recommended after a piece of equipment or process is moved or transferred to a new location.
- **People.**Let's face it, people are "buggy." If there are manual operations or steps in a test or process, human variability will be a factor. People cannot be calibrated and unlike machines are susceptible to fatigue, distraction and any number of influences which may cause them to vary from one step to another. Manual steps such as pipetting, swabbing, and scrubbing can all introduce variation. If there are manual operations, "person to person" variation should be factored into the validation plan.
- **Measurement System.**There is inherent variability in any measurement system. That inherent variability can come from any of the factors mentioned here.
- **Materials.**A significant contributor to the overall variation in a system comes from materials. Variance in incoming materials, tubing, purchased standards, molded parts, and essentially anything purchased for a system can contribute to the overall variance. If possible, multiple lots of materials should be tested during validation to determine the effect this variation on the validation.

CENTRAL TENDENCY

A brief discussion of the concept of central tendency is required. A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The arithmetic mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

Arithmetic Mean

The mean (or average) is the most well-known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have*n* values in a data set and they have values*x1, x2, ..., xn*, the sample mean, usually denoted by

Image not found
///C:/Users/BRUZZE~1/AppData/Local/Temp/msohtmlclip1/01/clip_image002.png
(pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots x_n)}{n} \quad \text{or} \quad \bar{x} = \frac{\sum x}{n}$$

For the data set: 1,2,3,4,5, the mean would be calculated as:

1+2+3+4+5 = 15, divided by 5 = 3

The mean is the most common measure of central tendency used in validation. It should be noted that it is the only measure of central tendency where further statistics can be applied. The downside of using the mean is that it can be affected by the influence of outliers. One or two points at an extreme value can shift the mean far off what the "true" process mean is. Outliers can have many anomalous causes. There may have been an error in data transmission or transcription. Outliers can arise due to changes in system behavior, human error, instrument error or simply through natural deviations in populations. A sample may have been contaminated with elements from outside the population being examined. Alternatively, an outlier could be the result of a flaw in the assumed theory, calling for further investigation. If your system or test method is prone to outliers, it can be useful to incorporate an outlier determination routine to determine if data can be excluded from the calculation. Please note however, these routines have to be established prior to validation runs in the validation protocol. If your system is prone to outliers, it is always better to investigate and try to eliminate the issue. Outliers present in validation will most likely be present during production or testing.

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median and the mode are the only measures of central tendency that can be used forordinal data, in which values are ranked relative to each other but are not measured absolutely.

The median is less affected by outliers and skewed data. However, no further measures of variance can be calculated from the median.

Mode

The mode is the most frequent score in a data set. This is the only central tendency measure that can be used withnominal data which have purely qualitative category assignments. No further statistics can be calculated from the mode, and it is not frequently used in validation as there are no numeric values assigned to nominal data.

Central Tendency Calculation

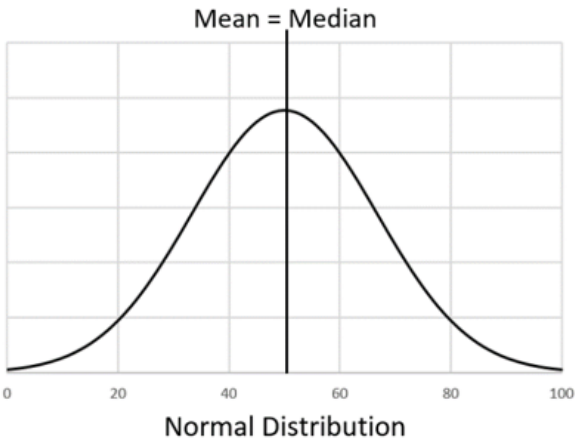
The following table shows a basic example of central tendency calculations. In the second data set where an outlier has been introduced, please notice that the mean is shifted but the median and mode remain the same.

	Data Set 1	Data Set 2
	1	1
	2	2
	3	3
	4	4
	4	4
	5	5
	6	6
	7	7
	8	8
	9	50
n =	10	10
Sum =	49	90
Mean =	4.9	9
Median =	4.5	4.5
Mode =	4	4

Table 1

DISTRIBUTION OF VALUES AROUND THE CENTRAL TENDENCY

Values in a data set can be distributed around the central tendency in two broad categories, normal distribution, and non-normal distribution. Figure A below shows data in a normal distribution. The data is evenly distributed around the central tendency and the mean = the median. There is no bias or shift in the data distribution.



Non-normal data distributions show a bias toward one side or the other of the data spread. The mean ? the median. See Figures B and C below.

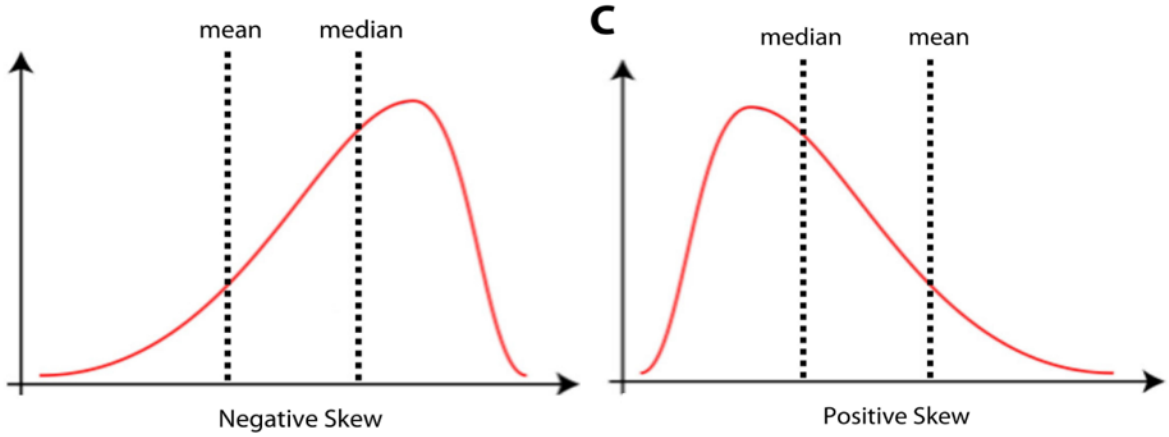


Figure B **Figure C**

Data collected during validation should follow a normal distribution. There should be no bias in the data set, and the variance in values should be spread equally above and below the mean value. This is sometimes referred to as ”common cause variation.” Skewed data may be the result of what is termed “special cause variation,” a true bias in the process or measurement system and should be investigated. Data can be tested to ensure it is normally distributed by a variety of methods (D'Agostino's K-squared test, Jarque–Bera test, Anderson–Darling test, Shapiro–Wilk test, and others). Many of these tests are available within common statistics software packages such as Minitab® Statistical Software, JMP® Software, or can be calculated in programs such as Excel® . If the data set is not normally distributed, there are mathematical transformations that can be used to transform the data to a normal distribution; but use of these transformations has to be justified. If the data set is not normally distributed, it is better to determine why and correct the system or test method. The following discussion on measuring variance assumes normally distributed data.

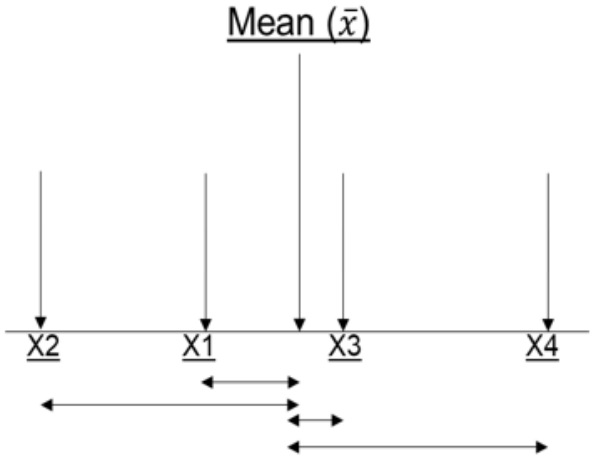
MEASURING VARIANCE

There are many ways to measure variance. For the purposes of this paper, we will discuss the three principal ways of measuring variance in validation:

- Standard Deviation of the Mean
- Coefficient of Variation
- Capability Index (Process Potential) and Process Performance

Standard Deviation

Standard deviation is a measure of the spread of the data around the mean of the data set. It is literally the mean of the distance each data point is from the mean of the data.



X2 =	2
X1 =	5
X3 =	8
X4 =	12
Mean=	6.75

Figure D

The standard deviation is calculated using the following formula:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \bar{X})^2}{n-1}}$$

- $\sum_{i=1}^{n}(Xi - \bar{X})^2$

  is the sum of each of the (data values – the mean value) squared. The result is squared to remove negative numbers that result from data points less than the mean.

- $n - 1$

  is the number of data points in the data set – 1 to indicate the degrees of freedom in the data set. In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. Most of the time the sample variance has n?1 degrees of freedom, since it is computed from n random scores minus the only 1 parameter estimated as intermediate step, which is the sample mean.

- *SD*

  Is the sample standard deviation. This is used if you have a sample from a larger population. If you measure the entire population, you can use the population standard deviation, which is calculated using n as opposed to n-1 in the denominator.

Standard Deviation Calculation

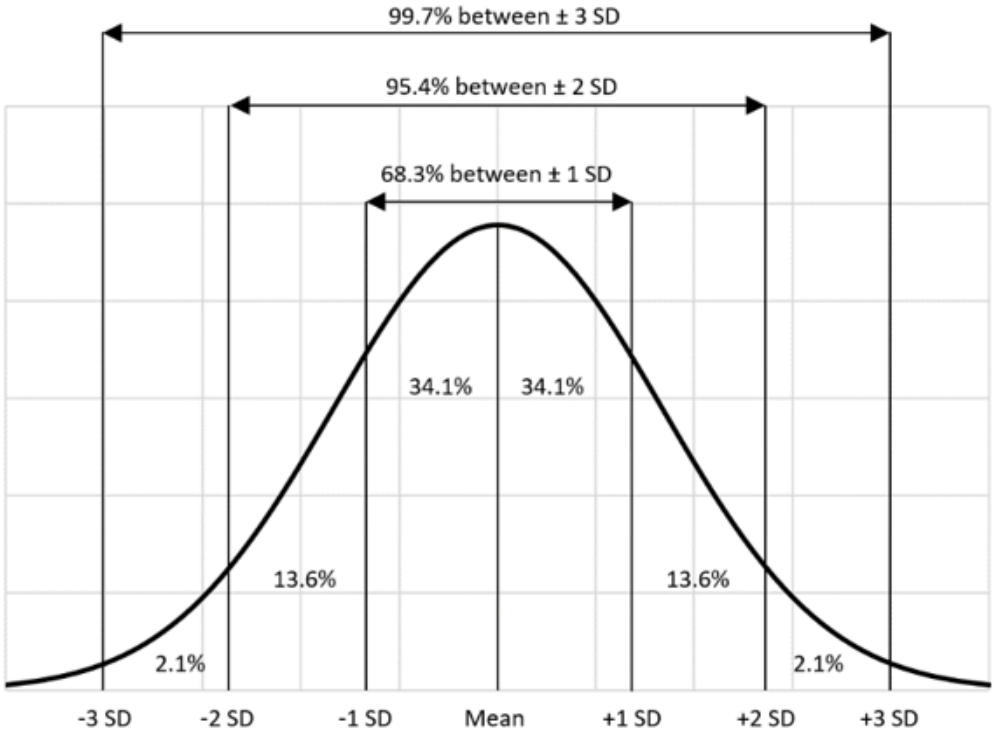
The following table shows a very simple example of a calculation of SD from the data set in Figure D. The overall variance in the data set is the square of the SD, as the SD is the square root of the variance.

Data (X)	Deviation From Mean	Squared Deviation
2	-4.75	22.56
5	-1.75	3.06
8	1.25	1.56
12	5.25	27.56
Mean = 6.75		Sum of Squared Dev = 54.75
Observations = n = 4		Variance (S2) = $\frac{SS}{n-1} = 18.25$
		$SD = \sqrt{S^2} = 4.27$

Table 2

The Meaning of Standard Deviation

What does standard deviation tell us about our data? By looking at a graph of normal data and corresponding standard deviations, it is easy to see how much information is gained from understanding the variance in our data set. For a normal distribution, the data is completely described by only two variables, the mean, and the standard deviation. Understanding the variance in your data (or validation) allows a determination if the process or test method is capable based on the needs of the product. See Figure E.



In a normally distributed data set, only 3 points in 1000 will lie outside the ± 3SD limits.

Figure E

The following table shows the probability of a data point in subsequent data sets falling in a range based on the variance in the system. This is a useful predictor of future performance. Please note that not much is gained by going beyond +/- three SD.

SD Range		Probability (%)
-1.00 Sigma	+1.00 Sigma	68.26

-1.96	1.96	95
-2	2	95.44
-2.58	2.58	99
-3	3	99.73
-3.29	3.29	99.9
-6	6	99.99

Table 3

FUTURE CONTENT

The next discussion will continue to examine statistics used in validation, including Coefficient of Variation, Process Capability and Acceptance Sampling.

REFERENCES

1. Berger, R. W., Benbow, D. W., Elshennawy, A. K., Walker, H.F. (2007). The Certified Quality Engineer Handbook, Second Edition. Milwaukee, Wisconsin: ASQ Quality Press
2. Boyles, R., (1991). "The Taguchi Capability Index". *Journal of Quality Technology*. 23 (1). Milwaukee, Wisconsin: American Society for Quality Control.
3. David, Stirzaker (2007). Elementary probability. Boston, Mass. Cambridge University Press
4. Ghasemi, A., Zahediasl, S., (2012) Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. Int J Endocrinol Metab. 10(2): 486–489. Accessed online January 30, 2019 at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/
5. Grubbs, F. E. (February 1969). "Procedures for detecting outlying observations in samples". Technometrics. 11 (1): 1–21.
6. Montgomery, D., (2004). Introduction to Statistical Quality Control. New York, New York: John Wiley & Sons, Inc.
7. Process Capability (Cp, Cpk) and Process Performance (Pp, Ppk) – What Is the Difference? https://www.isixsigma.com/tools-templates/capability-indices-process-capability/process-capability-cp-cpk-and-process-performance-pp-ppk-what-difference/ (accessed on Jan 22, 2019)
8. Razali, Nornadiah; Wah, Yap Bee (2011). "Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests" Journal of Statistical Modeling and Analytics. 2 (1): 21–33.
9. Thode, H. (2002). Testing for Normality. New York: Marcel Dekker, Inc.
