Standard Deviation, Standard Error

Which 'Standard' Should We Use?

George W. Brown, MD

• **Standard deviation (SD) and standard error (SE) are quietly but extensively used in biomedical publications. These terms and notations are used as *descriptive statistics* (summarizing numerical data), and they are used as *inferential statistics* (estimating population parameters from samples). I review the use and misuse of SD and SE in several authoritative medical journals and make suggestions to help clarify the usage and meaning of SD and SE in biomedical reports.**

(*Am J Dis Child* 1982;136:937-941)

Standard deviation (SD) and standard error (SE) have surface similarities; yet, they are conceptually so different that we must wonder why they are used almost interchangeably in the medical literature. Both are usually preceded by a plus-minus symbol (±), suggesting that they define a symmetric interval or range of some sort. They both appear almost always with a mean (average) of a set of measurements or counts of something. The medical literature is replete with statements like, "The serum cholesterol measurements were distributed with a mean of 180 ± 30 mg/dL (SD)."

In the same journal, perhaps in the same article, a different statement may appear: "The weight gains of the subjects averaged 720 (mean) ± 32 g/mo (SE)." Sometimes, as discussed further, the summary data are presented as the "mean of 120 mg/dL ± 12" without the "12" being defined as SD or SE, or as some other index of dispersion. Eisenhart[1] warned against this "peril of shorthand expression" in 1968; Feinstein[2] later again warned about the fatuity and confusion contained in any $a \pm b$ statements where b is not defined. Warnings notwithstanding, a glance through almost any medical journal will show examples of this usage.

Medical journals seldom state why SD or SE is selected to summarize data in a given report. A search of the three major pediatric journals for 1981 (*American Journal of Diseases of Children*, *Journal of Pediatrics*, and *Pediatrics*) failed to turn up a single article in which the selection of SD or SE was explained. There seems to be no uniformity in the use of SD or SE in these journals or in *The Journal of the American Medical Association* (*JAMA*), the *New England Journal of Medicine*, or *Science*. The use of SD and SE in the journals will be discussed further.

If these respected, well-edited journals do not demand consistent use of either SD or SE, are there really any important differences between them? Yes, they are remarkably different, despite their superficial similarities. They are so different in fact that some authorities have recommended that SE should rarely or never be used to summarize medical research data. Feinstein[2] noted the following:

> A standard error has nothing to do with standards, with errors, or with the communication of scientific data. The concept is an abstract idea, spawned by the imaginary world of statistical inference and pertinent only when certain operations of that imaginary world are met in scientific reality.[2(p336)]

Glantz[3] also has made the following recommendation:

> Most medical investigators summarize their data with the standard error because it is always smaller than the standard deviation. It makes their data look better . . . data should never be summarized with the standard error of the mean.[3(pp25-26)]

A closer look at the source and meaning of SD and SE may clarify why medical investigators, journal reviewers, and editors should scrutinize their usage with considerable care.

DISPERSION

An essential function of "descriptive statistics" is the presentation of condensed, shorthand symbols that epitomize the important features of a collection of data. The idea of a *central value* is intuitively satisfactory to anyone who needs to summarize a group of measurements or counts. The traditional indicators of a central tendency are the *mode* (the most frequent value), the *median* (the value midway between the lowest and the highest value), and the *mean* (the average). Each has its special uses, but the mean has great convenience and flexibility for many purposes.

The dispersion of a collection of values can be shown in several ways; some are simple and concise, and others are complex and esoteric. The *range* is a simple, direct way to indicate the spread of a collection of values, but it does not tell how the values are distributed. Knowledge of the mean adds considerably to the information carried by the range.

Another index of dispersion is provided by the differences (deviations) of each value from the mean of the values. The trouble with this approach is that some deviations will be positive, and some will be negative, and their sum will be zero. We could ignore the sign of each deviation, ie, use the "absolute mean deviation," but mathematicians tell us that working with absolute numbers is extremely difficult and fraught with technical disadvantages.

A neglected method for summarizing the dispersion of data is the calculation of percentiles (or deciles, or quartiles). Percentiles are used more frequently in pediatrics than in other branches of medicine, usually in growth charts or in other data arrays that are clearly not symmetric or bell shaped. In the general medical literature, percentiles are sparsely used, apparently because of a common, but erroneous, assumption that the mean ± SD or SE is satisfactory for summarizing central tendency and dispersion of all sorts of data.

From the Los Lunas Hospital and Training School, New Mexico, and the Department of Pediatrics, University of New Mexico School of Medicine, Albuquerque.

Reprint requests to Los Lunas Hospital and Training School, Box 1269, Los Lunas, NM 87031 (Dr Brown).

STANDARD DEVIATION

The generally accepted answer to the need for a concise expression for the dispersion of data is to square the difference of each value from the group mean, giving all positive values. When these squared deviations are added up and then divided by the number of values in the group, the result is the *variance.*

The variance is always a positive number, but it is in different units than the mean. The way around this inconvenience is to use the square root of the variance, which is the population standard deviation (σ), which for convenience will be called SD. Thus, the SD is the square root of the averaged squared deviations from the mean. The SD is sometimes called by the shorthand term, "root-mean-square."

The SD, calculated in this way, is in the same units as the original values and the mean. The SD has additional properties that make it attractive for summarizing dispersion, especially if the data are distributed symmetrically in the revered bell-shaped, gaussian curve. Although there are an infinite number of gaussian curves, the one for the data at hand is described completely by the mean and SD. For example, the mean + 1.96 SD will enclose 95% of the values; the mean ± 2.58 SD will enclose 99% of the values. It is this symmetry and elegance that contribute to our admiration of the gaussian curve.

The bad news, especially for biologic data, is that many collections of measurements or counts are not symmetric or bell shaped. Biologic data tend to be skewed or double humped, J shaped, U shaped, or flat on top. Regardless of the shape of the distribution, it is still possible by rote arithmetic to calculate an SD although it may be inappropriate and misleading.

For example, one can imagine throwing a six-sided die several hundred times and recording the score at each throw. This would generate a flattopped, ie, rectangular, distribution, with about the same number of counts for each score, 1 through 6. The mean of the scores would be 3.5 and the SD would be about 1.7. The trouble is that the collection of scores is not bell shaped, so the SD is not a good summary statement of the true form of the data. (It is mildly upsetting to some that no matter how many times the die is thrown, it will never show its average score of 3.5.)

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} \qquad SD = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

SD of Population

μ = Mean of Population

N = Number in Population

Estimate of Population SD From Sample

$\bar{X}$ = Mean of Sample

n = Number in Sample

Fig 1.—Standard deviation (SD) of population is shown at left. Estimate of population SD derived from sample is shown at right.

$$SD_{\bar{x}} = \frac{SD}{\sqrt{n}} = SEM \qquad SE_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$$

SEM

SD = Estimate of Population SD

n = Sample Size

SE of Proportion

p = Proportion Estimated From Sample

q = (1 − p)

n = Sample Size

Fig 2.—Standard error of mean (SEM) is shown at left. Note that SD is estimate of population SD (not σ, actual SD of population). Sample size used to calculate SEM is n. Standard error of proportion is shown at right.

The SD wears two hats. So far, we have looked at its role as a *descriptive statistic* for measurements or counts that are representative only of themselves, ie, the data being summarized are not a sample representing a larger (and itself unmeasurable) universe or population.

The second hat involves the use of SD from a random sample as an *estimate* of the population standard deviation (σ). The formal statistical language says that the sample *statistic,* SD, is an unbiased estimate of a population *parameter,* the population standard deviation, σ.

This "estimator SD" is calculated differently than the SD used to describe data that represent only themselves. When a sample is used to make estimates about the population standard deviation, the calculations require two changes, one in concept and the other in arithmetic. First, the mean used to determine the deviations is conceptualized as an estimate of the mean, $\bar{x}$, rather than as a true and exact population mean (μ). Both means are calculated in the same way, but a population mean, μ, stands for itself and is a parameter; a sample mean, $\bar{x}$, is an estimate of the mean of a larger population and is a statistic.

The second change in calculation is in the arithmetic: the sum of the squared deviations from the (estimated) mean is divided by n − 1, rather than by N. (This makes sense intuitively when we recall that a sample would not show as great a spread of values as the source population. Reducing the denominator [by one] produces an estimate slightly larger than the sample SD. This "correction" has more impact when the sample is small than when n is large.)

Formulas for the two versions of SD are shown in Fig 1. The formulas follow the customary use of Greek letters for population parameters and English letters for sample statistics. The number in a sample is indicated by the lowercase

"n," and the number in a population is indicated by the capital "N."

The two-faced nature of the SD has caused tension between medical investigators on the one hand and statisticians on the other. The investigator may believe that the subjects or measurements he is summarizing are self-contained and unique and cannot be thought of as a random sample. Therefore, he may decide to use the SD as a descriptive statement about dispersion of his data. On the other hand, the biostatistician has a tendency, because of his training and widespread statistical practice, to conceive of the SD as an estimator of a parameter of a population. The statistician may hold the view that any small collection of data is a stepping-stone to higher things.

The pervasive influence of statisticians is demonstrated in the program for calculating the SD that is put into many handheld calculators; they usually calculate the estimator SD rather than the "descriptor SD."

In essence, the investigator and his statistical advisor, the journal reviewers, and the editors all confront a critical decision whenever they face the term "standard deviation." Is it a descriptive statistic about a collection of (preferably gaussian) data that stand free and independent of sampling constraints, ie, is it a straightforward indication of dispersion? Or, is the SD being used as an estimate of a population parameter? Although the SD is commonly used to summarize medical information, it is rare that the reports indicate which version of the SD is being used.

STANDARD ERROR

In some ways, standard error is simpler than the SD, but in other ways, it is much more complex. First, the simplicities will be discussed. The SE is always smaller than the SD. This may account for its frequent use in medical publications; it makes the data look "tighter" than does the SD. In the previously cited quotation by Glantz,[3] the implication is that the SE might be used in a conscious attempt at distortion or indirection. A more charitable view is that many researchers and clinicians simply are not aware of the important differences between SD and SE. At first glance, the SE looks like a measure of dispersion, just as the SD does. The trouble is that the dispersion implied by the SE is different in nature than that implied by the SD.

The SE is always an estimator of a population characteristic; it is not a descriptive statistic—it is an inferential statistic. The SE is an estimate of the interval into which a population parameter will probably fall. The SE also enables the investigator to choose the probability that the parameter will fall within the estimated interval, usually called the "confidence interval."

Here is a statement containing the SE: The mean of the sample was 73 mg/dL, with an SE of the mean of 3 mg/dL. This implies that the mean of the population from which the sample was randomly taken will fall, with 95% probability, in the interval of $73 \pm (1.96 \times 3)$, which is from 67.12 to 78.88. Technically the statement should be: 95 out of 100 confidence intervals calculated in this manner will include the population parameter. If 99% probability is desired, the confidence interval is $73 \pm (2.58 \times 3)$, which is from 65.26 to 80.74.

As Feinstein[2] notes, the SE has nothing to do with standards or with errors; it has to do with predicting confidence intervals from samples. Up to this point, I have used SE as though it meant only the SE of the mean (SEM). The SE should not be used without indicating what parameter interval is being estimated. (I broke that rule for the sake of clarity in the introduction of the contrast between SD and SE.)

Every sample statistic can be used to estimate an SE; there is an SE for the mean, for the difference between the means of two samples, for the slope of a regression line, and for a correlation coefficient. Whenever the SE is used, it should be accompanied by a symbol that indicates which of the several SEs it represents, eg, SEM for SE of the mean.

Figure 2 shows the formula for calculating the SEM from the sample; the formula requires the estimator SD, ie, the SD calculated using $n-1$, not N. It is apparent from the formula for the SEM that the larger the sample size, the smaller the SEM and, therefore, the narrower the confidence interval. Stated differently, if the estimate of a population mean is from a large sample, the interval that probably brackets the population mean is narrower for the same level of confidence (probability). To reduce the confidence interval by half, it is necessary to increase the sample size by a multiple of four. For readers who know that the SD is preferred over the SEM as an index for describing dispersion of gaussian data, the formula for the SEM can be used (in reverse, so to speak) to calculate the SD, if sample size is known.

The theoretical meaning of the SEM is quite engaging, as an example will show. One can imagine a population that is too large for every element to be measured. A sample is selected randomly, and its mean is calculated, then the elements are replaced. The selection and measuring are repeated several times, each time with replacement. The collection of means of the samples will have a distribution, with a mean and an SD. The mean of the sample means will be a good estimate of the population mean, and the SD of the means will be the SEM. Figure 2 uses the symbol $SD_{\bar{x}}$ to show that a collection of sample means ($\bar{x}$) has a SD, and it is the SEM. The interpretation is that the true population mean (μ) will fall, with 95% probability, within ± 1.96 SEM of the mean of the means.

Here, we see the charm and attractiveness of the SEM. It enables the investigator to estimate from a sample, at whatever level of confidence (probability) desired, the interval within which the population mean will fall. If the user wishes to be very confident in his interval, he can set the brackets at ± 3.5 SEM, which would "capture" the mean with 99.96% probability.

Standard errors in general have other seductive properties. Even when the sample comes from a population that is skewed, U shaped, or flat on top, most SEs are estimators of nearly gaussian distributions for the statistic of interest. For example, for samples of size 30 or larger, the SEM and the sample mean, $\bar{x}$, define a nearly gaussian distribution (of sam-

ple means), regardless of the shape of the population distribution.

These elegant features of the SEM are embodied in a statistical principle called the Central Limit Theorem, which says, among other things:

> The mean of the collection of many sample means is a good estimate of the mean of the population, and the distribution of the sample means (if n = 30 or larger) will be nearly gaussian regardless of the distribution of the population from which the samples are taken.

The theorem also says that the collection of sample means from large samples will be better in estimating the population mean than means from small samples.

Given the symmetry and usefulness of SEs in inferential statistics, it is no wonder that some form of the SE, especially the SEM, is used so frequently in technical publications. A flaw occurs, however, when a confidence interval based on the SEM is used to replace the SD as a descriptive statistic; if a description of data spread is needed, the SD should be used. As Feinstein[2] has observed, the reader of a research report may be interested in the span or range of the data, but the author of the report instead displays an estimated zone of the mean (SEM).

An absolute prohibition against the use of the SEM in medical reports is not desirable. There are situations in which the investigator is using a truly random sample for estimation purposes. Random samples of children have been used, for example, to estimate population parameters of growth. The essential element is that the investigator (and editor) recognize when descriptive statistics should be used, and when inferential (estimation) statistics are required.

SE OF PROPORTION

As mentioned previously, every sample statistic has its SE. With every statistic, there is a confidence interval that can be estimated. Despite the widespread use of SE (unspecified) and of SEM in medical journals and books, there is a noticeable neglect of one important SE, the SE of the proportion.

The discussion so far has dealt with measurement data or counts of elements. Equally important are data reported in proportions or percentages, such as, "Six of the ten patients with zymurgy syndrome had so-and-so." From this, it is an easy step to say, "Sixty percent of our patients with zymurgy syndrome had so-and-so." The implication of such a statement may be that the author wishes to alert other clinicians, who may encounter samples from the universe of patients with zymurgy syndrome that they may see so-and-so in about 60% of them.

The proportion—six of ten—has an SE of the proportion. As shown in Fig 2, the SE_p in this situation is the square root of (0.6×0.4) divided by ten, which equals 0.155. The true proportion of so-and-so in the universe of patients with zymurgy syndrome is in the confidence interval that falls symmetrically on both sides of six of ten. To estimate the interval, we start with 0.6 or 60% as the midpoint of the interval. At the 95% level of confidence, the interval is $0.6 \pm 1.96\ SE_p$, which is $0.6 \pm (1.96 \times 0.155)$, or from 0.3 to 0.9.

If the sample shows six of ten, the 95% confidence interval is between 30% (three of ten) and 90% (nine of ten). This is not a very narrow interval. The expanse of the interval may explain the almost total absence of the SE_p in medical reports, even in journals where the SEM and SD are used abundantly. Investigators may be dismayed by the dimensions of the confidence interval when the SE_p is calculated from the small samples available in clinical situations.

Of course, as in the measurement of self-contained data, the investigator may not think of his clinical material as a sample from a larger universe. But often, it is clear that the purpose of publication is to suggest to other investigators or clinicians that, when they see patients of a certain type, they might expect to encounter certain characteristics in some estimated proportion of such patients.

JOURNAL USE OF SD AND SE

To get empiric information about pediatric journal standards on descriptive statistics, especially the use of SD and SE, I examined every issue of the three major pediatric journals published in 1981: *American Journal of Diseases of Children*, *Journal of Pediatrics*, and *Pediatrics*. In a less systematic way, I perused several issues of *JAMA*, the *New England Journal of Medicine*, and *Science*.

Every issue of the three pediatric journals had articles, reports, or letters in which SD was mentioned, without specification of whether it was the descriptive SD or the estimate SD. Every issue of the *Journal of Pediatrics* contained articles using SE (unspecified) and articles using SEM. *Pediatrics* used SEM in every issue and the SE in every issue except one. Eight of the 12 issues of the *American Journal of Diseases of Children* used SE or SEM or both. All the journals used SE as if SE and SEM were synonymous.

Every issue of the three journals contained articles that stated the mean and range, without other indication of dispersion. Every journal contained reports with a number ± (another number), with no explanation of what the number after the plus-minus symbol represented.

Every issue of the pediatric journals presented proportions of what might be thought of as samples without indicating that the SE_p (standard error of the proportion) might be informative.

In several reports, SE or SEM is used in one place, but SD is used in another place in the same article, sometimes in the same paragraph, with no explanation of the reason for each use. The use of percentiles to describe nongaussian distributions was infrequent. Similar examples of stylistic inconsistency were seen in the haphazard survey of *JAMA*, the *New England Journal of Medicine*, and *Science*.

A peculiar graphic device (seen in several journals) is the use, in illustrations that summarize data, of a point and vertical bars, with no indication of what the length of the bars signifies.

A prevalent and unsettling practice is the use of the mean ± SD for data that are clearly not gaussian or not symmetric. Whenever data are reported with the SD as large or larger than the mean, the inference must be that several values are zero or negative. The mean ± 2 SDs should embrace about 95% of the values in a gaussian distribution. If the SD is as large as the mean, then the lower tail of the bell-shaped curve will go below zero. For many

biologic data, there can be no negative values; blood chemicals, serum enzymes, and cellular elements cannot exist in negative amounts.

An article by Fletcher and Fletcher[4] entitled "Clinical Research in General Medical Journals" in a leading publication demonstrates the problem of ± SD in real life. The article states that in 1976 certain medical articles had an average of 4.9 authors ± 7.3 (SD)! If the authorship distribution is gaussian, which is necessary for ± SD to make sense, this statement means that 95% of the articles had $4.9 \pm (1.96 \times 7.3)$ authors, or from −9.4 to +19.2. Or stated another way, more than 25% of the articles had zero or fewer authors.

In such a situation, the SD is not good as a descriptive statistic. A mean and range would be better; percentiles would be logical and meaningful.

Deinard et al[5] summarized some mental measurement scores using the mean ± SD and the range. They vividly showed two dispersions for the same data. For example, one set of values was 120.8 ± 15.2 (SD); the range was 63 to 140. The SD implies gaussian data, so 99% of the values should be within ± 2.58 SDs of the mean or between 81.6 and 160. Which dispersion should we believe, 63 to 140 or 81.6 to 160?

ADVICE OF AUTHORITIES

There may be a ground swell of interest among research authorities to help improve statistical use in the medical literature. Friedman and Phillips[6] pointed out the embarrassing uncertainty that pediatric residents have with *P* values and correlation coefficients. Berwick and colleagues,[7] using a questionnaire, reported considerable vagueness about statistical concepts among many physicians in training, in academic medicine, and in practice. However, in neither of these reports is attention given to the interesting but confusing properties of SD and SE.

In several reports,[8-10] the authors urge that we be wary when comparative trials are reported as not statistically significant. Comparisons are vulnerable to the error of rejecting results that look negative, especially with small samples, but may not be. These authorities remind us of the error of failing to detect a real difference, eg, between controls and treated subjects, when such a difference exists. This failure is called the "error of the second kind," the Type II error, or the beta error. In laboratory language, this error is called the false-negative result, in which the test result says "normal" but nature reveals "abnormal" or "disease present." (The Type I error, the alpha error, is a more familiar one; it is the error of saying that two groups differ in some important way when they do not. The Type I error is like a false-positive laboratory test in that the test suggests that the subject is abnormal, when in truth he is normal.)

In comparative trials, calculation of the Type II error requires knowledge of the SEs, whether the comparisons are of group means (requiring SEM) or comparisons of group proportions (requiring SE_p).

At the outset, I mentioned that we are advised[2,3] to describe clinical data using means and the SD (for bell-shaped distributions) and to eschew use of the SE. On the other hand, we are urged to examine clinical data for interesting confidence intervals,[11,12] searching for latent scientific value and avoiding a too hasty pronouncement of not significant. To avoid this hasty fall into the Type II error (the false-negative decision), we must increase sample sizes; in this way, a worthwhile treatment or intervention may be sustained rather than wrongly discarded.

It may be puzzling that some authorities seem to be urging that the SE should rarely be used, but others are urging that more attention be paid to confidence intervals, which depend on the SE. This polarity is more apparent than real. If the investigator's aim is description of data, he should avoid the use of the SE; if his aim is to estimate population parameters or to test hypotheses, ie, inferential statistics, then some version of the SE is required.

WHO IS RESPONSIBLE?

It is not clear who should be held responsible for data displays and summary methods in medical reports. Does the responsibility lie at the door of the investigator-author and his statistical advisors, with the journal referees and reviewers, or with the editors? When I ask authors about their statistical style, the reply often is, "The editors made me do it."

An articulate defender of good statistical practice and usage is Feinstein,[2] who has regularly and effectively urged the appropriate application of biostatistics, including SD and SE. In his book, *Clinical Biostatistics*, he devotes an entire chapter (chap 23, pp 335-352) to "problems in the summary and display of statistical data." He offers some advice to readers who wish to improve the statistics seen in medical publications: "And the best person to help re-orient the editors is you, dear reader, you. Make yourself a one-person vigilante committee."[2(p349)]

Either the vigilantes are busy in other enterprises or the editors are not listening, because we continue to see the kind of inconsistent and confusing statistical practices that Eisenhart[1] and Feinstein[2] have been warning about for many years. I can only echo what others have said: When one sees medical publications with inappropriate, confusing, or wrong statistical presentation, one should write to the editors. Editors are, after all, the assigned defenders of the elegance and accuracy of our medical archives.

References

1. Eisenhart C: Expression of the uncertainties of final results. *Science* 1968;160:1201-1204.
2. Feinstein AR: *Clinical Biostatistics.* St Louis, CV Mosby Co, 1977.
3. Glantz SA: *Primer of Biostatistics.* New York, McGraw-Hill Book Co, 1981.
4. Fletcher R, Fletcher S: Clinical research in general medical journals: A 30-year perspective. *N Engl J Med* 1979;301:180-183.
5. Deinard A, Gilbert A, Dodd M, et al: Iron deficiency and behavioral deficits. *Pediatrics* 1981;68:828-833.
6. Friedman SB, Phillips S: What's the difference?: Pediatric residents and their inaccurate concepts regarding statistics. *Pediatrics* 1981;68:644-646.
7. Berwick DM, Fineberg HV, Weinstein MC: When doctors meet numbers. *Am J Med* 1981;71:991-998.
8. Freiman JA, Chalmers TC, Smith H Jr, et al: The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 'negative' trials. *N Engl J Med* 1978;299:690-694.
9. Berwick DM: Experimental power: The other side of the coin. *Pediatrics* 1980; 65:1043-1045.
10. Pascoe JM: Was it a Type II error? *Pediatrics* 1981;68:149-150.
11. Rothman KJ: A show of confidence. *N Engl J Med* 1978;299:1362-1363.
12. Guess H: Lack of predictive indices in kernicterus—or lack of power? *Pediatrics* 1982; 69:383.