Maintaining Standards: Differences between the Standard Deviation and Standard Error, and When to Use Each

David L Streiner, PhD[1]

Many people confuse the standard deviation (SD) and the standard error of the mean (SE) and are unsure which, if either, to use in presenting data in graphical or tabular form. The SD is an index of the variability of the original data points and should be reported in all studies. The SE reflects the variability of the mean values, as if the study were repeated a large number of times. By itself, the SE is not particularly useful; however, it is used in constructing 95% and 99% confidence intervals (CIs), which indicate a range of values within which the "true" value lies. The CI shows the reader how accurate the estimates of the population values actually are. If graphs are used, error bars equal to plus and minus 2 SEs (which show the 95% CI) should be drawn around mean values. Both statistical significance testing and CIs are useful because they assist the reader in determining the meaning of the findings.

(Can J Psychiatry 1996;41:498–502)

Key Words: *statistics, standard deviation, standard error, confidence intervals, graphing*

Imagine that you've just discovered a new brain protein that causes otherwise rational people to continuously mutter words like "reengineer," "operational visioning," and "mission statements." You suspect that this new chemical, which you call LDE for Language Destroying Enzyme, would be found in higher concentrations in the cerebrospinal fluid (CSF) of administrators than that of other people. Difficult as it is to find volunteers, you eventually get samples from 25 administrators and an equal number of controls and find the results shown in Table I. Because you feel that these data would be more compelling if you showed them visually, you prepare your paper using a bar graph. Just before you mail it off, though, you vaguely remember something about error bars, but can't quite recall what they are; you check with a few of your colleagues. The first one tells you to draw a line above and below the top of each bar so that each part is equal to the standard deviation. The second person disagrees, saying that the lines should reflect the standard errors, while the third person has yet another opinion—the lines should be plus and minus 2 standard errors, that is, 2 standard errors above and 2 below the mean. As you can see in Figure 1, these methods result in very different pictures of what's going on. So, now you have 2 problems: first, what is the difference between the standard error and the standard deviation, and second, which should you draw?

Manuscript received March 1996.
This article is the eleventh in the series on Research Methods in Psychiatry. For previous articles please see Can J Psychiatry 1990;35:616–20, 1991; 36:357–62, 1993;38:9–13, 1993;38:140–8, 1994;39:135–40, 1994; 39:191–6, 1995;40:60–6, 1995;40:439–44, 1996;41:137–43, and 1996; 41:491–7.
[1]Professor, Departments of Clinical Epidemiology and Biostatistics and of Psychiatry, McMaster University, Hamilton, Ontario.
Address reprint requests to: Dr David L Streiner, Department of Clinical Epidemiology and Biostatistics, McMaster University, 1200 Main Street West, Hamilton, ON L8N 3Z5
e-mail: streiner@fhs.csu.mcmaster.ca

Standard Deviation

The standard deviation, which is abbreviated variously as S.D., SD, or *s* (just to confuse people), is an index of how closely the individual data points cluster around the mean. If we call each point "X_i," so that "X_1" indicates the first value, "X_2" the second value, and so on, and call the mean "M," then it may seem that an index of the dispersion of the points would be simply $\Sigma(X_i - M)$, which means to sum (that's what the Σ indicates) how much each value of X deviates from M; in other words, an index of dispersion would be the *Sum of (Individual Data Points – Mean of the Data Points)*.

Logical as this may seem, it has 2 drawbacks. The first difficulty is that the answer will be zero—not just in this situation, but in every case. By definition, the sum of the values above the mean is always equal to the sum of the values

Table I
Levels of LDE in the CSF of Administrators and Controls

Group	Number	Mean	SD
Administrators	25	25.83	5.72
Controls	25	17.25	4.36

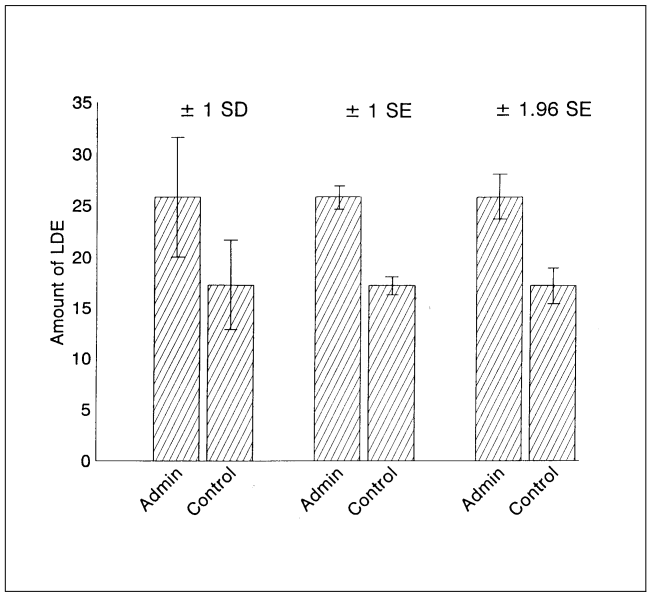

Figure 1. Data from Table I, plotted with different types of error bars.

below it, and thus they'll cancel each other out. We can get around this problem by taking the absolute value of each difference (that is, we can ignore the sign whenever it's negative), but for a number of arcane reasons, statisticians don't like to use absolute numbers. Another way to eliminate negative values is to square them, since the square of any number—negative or positive—is always positive. So, what we now have is $\Sigma(X_i - M)^2$.

The second problem is that the result of this equation will increase as we add more subjects. Let's imagine that we have a sample of 25 values, with an SD of 10. If we now add another 25 subjects who look exactly the same, it makes intuitive sense that the dispersion of these 50 points should stay the same. Yet the formula as it now reads can result only in a larger sum as we add more data points. We can compensate for this by dividing by the number of subjects, N, so that the equation now reads $\Sigma(X_i - M)^2/N$.

In the true spirit of Murphy's Law, what we've done in solving these 2 difficulties is to create 2 new ones. The first (or should we say third, so we can keep track of our problems) is that now we are expressing the deviation in squared units; that is, if we were measuring IQs in children with autism, for instance, we may find that their mean IQ is 75 and their dispersion is 100 squared IQ points. But what in heaven's name is a squared IQ point? At least this problem is easy to cure: we simply take the square root of the answer, and we'll end up with a number that is in the original units of measurement, so in this example, the dispersion will be 10 IQ points, which is much easier to understand.

The last problem (yes, it really is the last one) is that the results of the formula as it exists so far produce a *biased* estimate, that is, one that is consistently either higher or (as in this case) lower than the "true" value. The explanation of this is a bit more complicated and requires somewhat of a detour. Most of the time when we do research, we are not interested so much in the samples we study as in the populations they come from. That is, if we look at the level of expressed emotion (EE) in the families of young schizophrenic males, our interest is in the families of all people who meet the criteria (the population), not just those in our study. What we do is *estimate* the population mean and SD from our sample. Because all we are studying is a sample, however, these estimates will deviate by some unknown amount from the population values. In calculating the SD, we would ideally see how much each person's score deviates from the population mean, but all we have available to us is the sample mean. By definition, scores deviate less from their own mean than from any other number. So, when we do the calculation and subtract each score from the sample mean, the result will be smaller than if we subtracted each score from the population mean (which we don't know); hence, the result is biased downwards. To correct for this, we divide by N – 1 instead of N. Putting all of this together, we finally arrive at the formula for the standard deviation, which is:

$$SD = \sqrt{\frac{\Sigma(X_i - M)^2}{N-1}}$$

(By the way, don't use this equation if, for whatever bizarre reason, you want to calculate the SD by hand, because it leads to too much rounding error. There is another formula, mathematically equivalent and found in any statistics book, which yields a more precise figure.)

Now that we've gone through all this work, what does it all mean? If we assume that the data are normally distributed, then knowing the mean and SD tells us everything we need to know about the distribution of scores. In any normal distribution, roughly two-thirds (actually, 68.2%) of the scores fall between –1 and +1 SD, and 95.4% between –2 and +2 SD. For example, most of the tests used for admission to graduate or professional schools (the GRE, MCAT, LSAT,

and other instruments of torture) were originally designed to have a mean of 500 and an SD of 100. That means that 68% of people get scores between 400 and 600, and just over 95% between 300 and 700. Using a table of the normal curve (found in most statistics books), we can figure out exactly what proportion of people get scores above or below any given value. Conversely, if we want to fail the lowest 5% of test takers (as is done with the LMCCs), then knowing the mean and SD of this year's class and armed with the table, we can work out what the cut-off point should be.

So, to summarize, the SD tells us the distribution of *individual scores* around the mean. Now, let's turn our attention to the standard error.

Standard Error

I mentioned previously that the purpose of most studies is to estimate some population parameter, such as the mean, the SD, a correlation, or a proportion. Once we have that estimate, another question then arises: How accurate is our estimate? This may seem an unanswerable question; if we don't know what the population value is, how can we evaluate how close we are to it? Mere logic, however, has never stopped statisticians in the past, and it won't stop us now. What we can do is resort to probabilities: What is the probability (*P*) that the true (population) mean falls within a certain range of values? (To cite one of our mottos, "Statistics means you never have to say you're certain.")

One way to answer the question is to repeat the study a few hundred times, which will give us many estimates of the mean. We can then take the mean of these means, as well as figure out what the distribution of means is; that is, we can get the standard deviation of the mean values. Then, using the same table of the normal curve that we used previously, we can estimate what range of values would encompass 90% or 95% of the means. If each sample had been drawn from the population at random, we would be fairly safe in concluding that the true mean also falls within this range 90% or 95% of the time. We assign a new name to the standard deviation of the means: we call it the *standard error of the mean* (abbreviated as SEM, or, if there is no ambiguity that we're talking about the mean, SE).

But first, let's deal with one slight problem—replicating the study a few hundred times. Nowadays, it's hard enough to get money to do a study once, much less replicate it this many times (even assuming you would actually want to spend the rest of your life doing the same study over and over). Ever helpful, statisticians have figured out a way to determine the SE based on the results of a single study. Let's approach this first from an intuitive standpoint: What would make us more or less confident that our estimate of the population mean, based on our study, is accurate? One obvious thing would be the size of the study; the larger the sample size, N, the less chance that one or two aberrant values are distorting the results and the more likely it is that our estimate is close to the true value. So, some index of N should be in the denominator of SE, since the larger N is, the smaller SE would become. Second, and for similar reasons, the smaller the variability in the data, the more confident we are that one value (the mean) accurately reflects them. Thus, the SD should be in the numerator: the larger it is, the larger SE will be, and we end up with the equation:

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

(Why does the denominator read $\sqrt{N}$ instead of just N? Because we are really dividing the variance, which is SD^2, by N, but we end up again with squared units, so we take the square root of everything. Aren't you sorry you asked?)

So, the SD reflects the variability of *individual data points*, and the SE is the variability of *means*.

Confidence Intervals

In the previous section, on the SE, we spoke of a range of values in which we were 95% or 99% confident that the true value of the mean fell. Not surprisingly, this range is called the confidence interval, or CI. Let's see how it's calculated. If we turn again to our table of the normal curve, we'll find that 95% of the area falls between −1.96 and +1.96 SDs. Going back to our example of GREs and MCATs, which have a mean of 500 and an SD of 100, 95% of scores fall between 304 and 696. How did we get those figures? First, we multiplied the SD by 1.96, subtracted it from the mean to find the lower bound, and added it to the mean for the upper bound. The CI is calculated in the same way, except that we use the SE instead of the SD. So, the 95% CI is:

$$95\%\ CI = M \pm (1.96 \times SE)$$

For the 90% CI, we would use the value 1.65 instead of 1.96, and for the 99% CI, 2.58. Using the data from Table I, the SE for administrators is $5.72 / \sqrt{25}$, or 1.14, and thus the 95% CI would be $25.83 \pm (1.96 \times 1.14)$, or 23.59 to 28.07. We would interpret this to mean that we are 95% confident that the value of LDE in the population of administrators is somewhere within this interval. If we wanted to be more confident, we would multiply 1.14 by 2.58; the penalty we

pay for our increased confidence is a wider CI, so that we are less sure of the exact value.

The Choice of Units

Now we have the SD, the SE, and any one of a number of CIs, and the question becomes, which do we use, and when? Obviously, when we are describing the results of any study we've done, it is imperative that we report the SD. Just as obviously, armed with this and the sample size, it is a simple matter for the reader to figure out the SE and any CI. Do we gain anything by adding them? The answer, as usual, is yes and no.

Essentially, we want to convey to the reader that there will always be sample-to-sample variation and that the answers we get from one study wouldn't be exactly the same if the study were replicated. What we would like to show is how much of a difference in findings we can expect: just a few points either way, but not enough to substantially alter our conclusions, or so much that the next study is as likely to show results going in the opposite direction as to replicate the findings. To some degree, this is what significance testing does—the lower the *P* level, the less likely the results are due simply to chance and the greater the probability that they will be repeated the next time around. Significance tests, however, are usually interpreted in an all-or-nothing manner: either the result was statistically significant or it wasn't, and a difference between group means that just barely squeaked under the $P < 0.05$ wire is often given as much credence as one that is highly unlikely to be due to chance.

If we used CIs, either in a table or a graph, it would be much easier for the reader to determine how much variation in results to expect from sample to sample. But which CI should we use? We could draw the error bars on a graph or show in a table a CI that is equal to exactly one SE. This has the advantages that we don't have to choose between the SE or the CI (they're identical) and that not much calculation is involved. Unfortunately, this choice of an interval conveys very little useful information. An error bar of plus and minus one SE is the same as the 68% CI; we would be 68% sure that the true mean (or difference between 2 means) fell within this range. The trouble is, we're more used to being 95% or 99% sure, not 68%. So, to begin with, let's forget about showing the SE: it tells us little that is useful, and its sole purpose is in calculating CIs.

What about the advice to use plus and minus 2 SEs in the graph? This makes more sense; 2 is a good approximation of 1.96, at least to the degree that graphics programs can display the value and our eyes discern it. The advantages are twofold. First, this method shows the 95% CI, which is more meaningful than 68%. Second, it allows us to do an "eyeball" test of significance, at least in the 2-group situation. If the top of the lower bar (the controls in Figure 1) and the bottom of the higher bar (the administrators) do not overlap, then the difference between the groups is significant at the 5% level or better. Thus we would say that, in this example, the 2 groups were significantly different from one another. If we actually did a *t* test, we would find this to be true: $t(48) = 2.668$, $P < 0.05$. This doesn't work too accurately if there are more than 2 groups, since we have the issue of multiple tests to deal with (for example, Group 1 versus Group 2, Group 2 versus 3, and Group 1 versus 3), but it gives a rough indication of where the differences lie. Needless to say, when presenting the CI in a table, you should give the exact values (multiply by 1.96, not 2).

Wrapping Up

The SD indicates the dispersion of individual data values around their mean, and should be given any time we report data. The SE is an index of the variability of the means that would be expected if the study were exactly replicated a large number of times. By itself, this measure doesn't convey much useful information. Its main function is to help construct 95% and 99% CIs, which can supplement statistical significance testing and indicate the range within which the true mean or difference between means may be found. Some journals have dropped significance testing entirely and replaced it with the reporting of CIs; this is probably going too far, since both have advantages, and both can be misused to equal degrees. For example, a study using a small sample size may report that the difference between the control and experimental group is significant at the 0.05 level. Had the study indicated the CIs, however, it would be more apparent to the reader that the CI is very wide and the estimate of the difference is crude, at best. By contrast, the much-touted figure of the number of people affected by second-hand smoke is actually not the estimate of the mean. The best estimate of the mean is zero, and it has a very broad CI; what is reported is the upper end of that CI.

To sum up, SDs, significance testing, and 95% or 99% CIs should be reported to help the reader; all are informative and complement, rather than replace, each other. Conversely, "naked" SEs don't tell us much by themselves, and more or less just take up space in a report. Conducting our studies with these guidelines in mind may help us to maintain the standards in psychiatric research.

Résumé

Beaucoup de gens confondent l'écart-type et l'erreur-type de la moyenne et ne savent pas lequel utiliser pour présenter les données sous forme graphique ou tabulaire. L'écart-type indique la variabilité des données originales et devrait être mentionné pour toutes les études. L'erreur-type montre la variabilité des valeurs moyennes, comme si l'étude avait été reprise de nombreuses fois. En soi, l'erreur-type n'a pas d'utilité particulière; toutefois, on s'en sert pour créer les intervalles de confiance à 95 % et à 99 % utilisés pour établir la fourchette de valeurs dans laquelle se situe la valeur «réelle». Les intervalles de confiance signalent au lecteur la précision des estimations des valeurs démographiques. Lorsqu'on se sert de graphiques, la barre d'erreur représente un intervalle de plus à moins 2 écarts-types (ce qui correspond à l'intervalle de confiance de 95 %). Elle devrait entourer la valeur moyenne. Les épreuves de signification statistique et les intervalles de confiance présentent une grande utilité, car ils aident le lecteur à établir l'importance des constatations.