Never *P* alone: The value of estimates and confidence intervals

Tom Lang
Tom Lang Communications and Training International, Kirkland, WA, USA

Correspondence to:
Tom Lang
10003 NE 115th Lane
Kirkland, WA 98933
USA
tomlangcom@aol.com
+1 425 242 1370

Abstract
How results are reported influences how they are interpreted. Although *P* values have been granted great importance, they have no clinical interpretation. Rather, they are a measure of chance as an explanation for the results. Their either-or interpretation takes attention away from the results themselves–the difference between groups or the effect size–which are more important. Effect sizes are also estimates. Estimates are only useful if they are accompanied by a measure of precision. In medicine, this measure is usually the 95% confidence interval (CI). This article explains the concepts underlying CIs and illustrates how they are more useful than *P* values in reporting research. As such, journals are increasingly asking for CIs, instead of, or at least in addition to, *P* values.

Introduction
Statistics can be divided into two broad areas: **descriptive statistics**, in which data are summarised in a few numbers to make them more manageable, such as percentages and medians, and **inferential statistics**, in which measurements of a sample are generalised to the population from which the sample was drawn. This article is concerned with inferential statistics; in particular, the reporting of estimates and confidence intervals.

Most medical research is done on samples, but the findings are actually estimates of what we would expect if the treatment were to be given to the population from which the sample was drawn. For example, we can't study all patients with, say, epilepsy, we can only study a sample of such patients. When we're done, we hope that what we have learned from the sample will also be true for all patients who have epilepsy.

However, the sample is almost always only a tiny fraction of the population, so we need to know how good our estimate is. In medicine, this measure of precision is most often expressed as a confidence interval (CI), usually a 95% CI, although the "confidence coefficient" (the 95%) may be 90% for smaller samples and theoretically can be any number. Thus, understanding estimates and confidence intervals is important to understanding the medical literature.

In this article, I illustrate the concepts underlying estimates and confidence intervals with a hypothetical example. Hypothetical, because the concepts involved differ from the actual research methods and mathematics used to compute the confidence intervals. After giving the example, I'll explain how the confidence interval is actually determined. Interested readers are

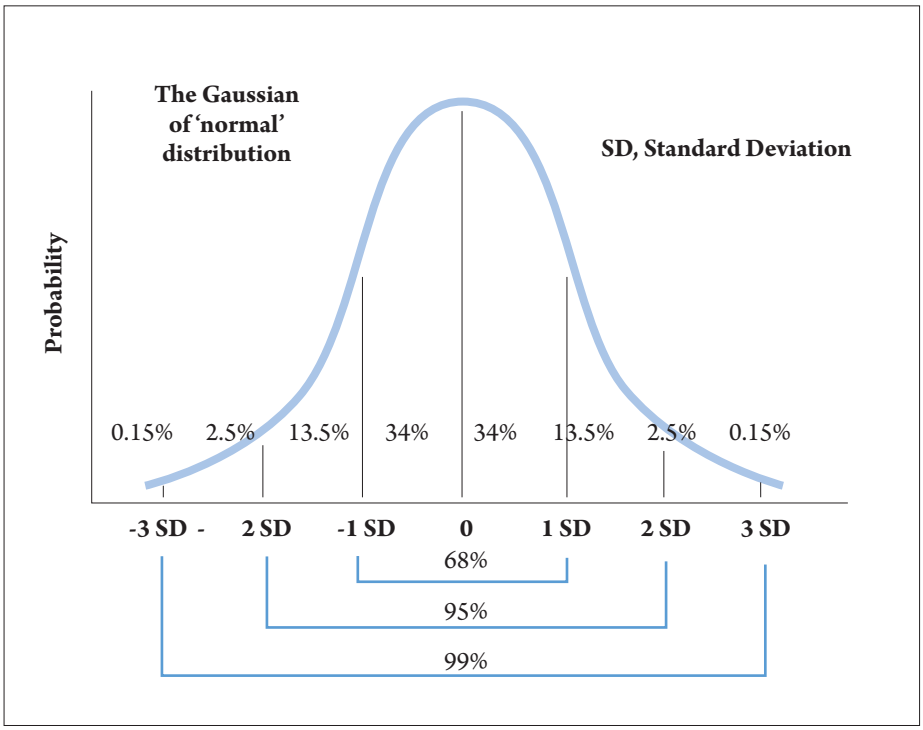

Figure 1. The relationship between the standard deviation and the area under the normal curve holds for all normal distributions, no matter how flat or peaked.

In a distribution of data, the SD is the preferred "measure of dispersion," or spread of the data. Other normal distributions have an SD, but the name changes to connect it with the distribution. In a distribution of all possible sample means, as described below, the SD is called the standard error of the mean (SE). It has the same mathematical properties as the SD, it's just associated with a different distribution.

invited to read *Statistics without Tears*, by Rowntree[1] for a fuller description of the approach taken here and *How to Report Statistics in Medicine*, by Lang[2] for more information about reporting estimates and confidence intervals.

Background information

Before we can talk about estimates and confidence intervals, we have to review some basic concepts of probability. In particular, we need to review the properties of the "normal distribution" or the gaussian or bell-shaped curve.

In any normal distribution, the mean value equals the median value equals the modal value, and the curve is symmetrical about the mean. It also has two "inflection points" where the curve changes direction to give it its bell shape. Most importantly, the area under the curve can be described in units of standard deviation (SD), and this relationship holds for any normal distribution *(Figure 1)*. Importantly, this relationship allows us to compare values on any normal distribution with those of any other, no matter how peaked or flattened the curves.

Suppose we wanted to compare patient survival in two groups of different sizes. It wouldn't be fair to compare the raw numbers of survivors between groups because one group is larger than the other. Instead, we convert the raw numbers into a common unit – percentages – to accommodate the difference in group size and then compare the percentages.

Now, suppose we want to compare two different normal distributions. Linda took the final exam in her law class, and Bill took his in economics. We want to compare their scores to determine who is the better student *(Figure 2)*. We can't compare Bill's score of 90 to Linda's score of 80 because each test has a different distribution of values; one test had more questions than the other, which changes the range of possible scores, or maybe one class had more variability than the other because more people did well and more people did poorly on the test.

As we did with percentages, however, we can compare scores from different distributions if we can express the values in a common measure. We do this by converting raw scores into units of SD (a "standard score," or z-score), which we can then compare on a common distribution. A score equal to the mean value of the common distribution (or "standard normal distribution") has an SD of zero; half the values are less than the score and half are greater. A score 1 SD above the mean is greater than about 84% of the values (50% to the left of the median or center value plus 34%) and less than about 16%, whereas a score of -1 SD below the mean is greater than about 16% of the values and less than about 84% (*Figure 1*).

Getting back to Bill and Linda, if we now express the two scores in terms of SDs, we see that Bill's score of 90 was 2 SD above the mean in his class, and Linda's was 3 SD above the mean in her's. So, Bill did better than about 97.5% of his classmates, but Linda did better than about 99.9% of hers. Linda did relatively better, even though her raw score was less than Bill's.

It is important to remember that the SD indicates these proportions only for normal distributions. So, normal distributions can be appropriately summarised with means and SDs, but distributions of other shapes should be summarized with different descriptive statistics.

Estimating a population value

An **estimate** is a *probable* value for a population that is inferred from a *measured* value of a sample. In medicine, we sometimes want to estimate the value of a physical trait in a population, such as average birth weight. We might also want to estimate the response to an intervention, such as differences between groups ("between-group comparisons") or in the same group before and after treatment ("with-in group comparisons").

Here's the hypothetical example. Imagine a gnome, a mythical being that guards the earth's underground treasures. Gnomes have

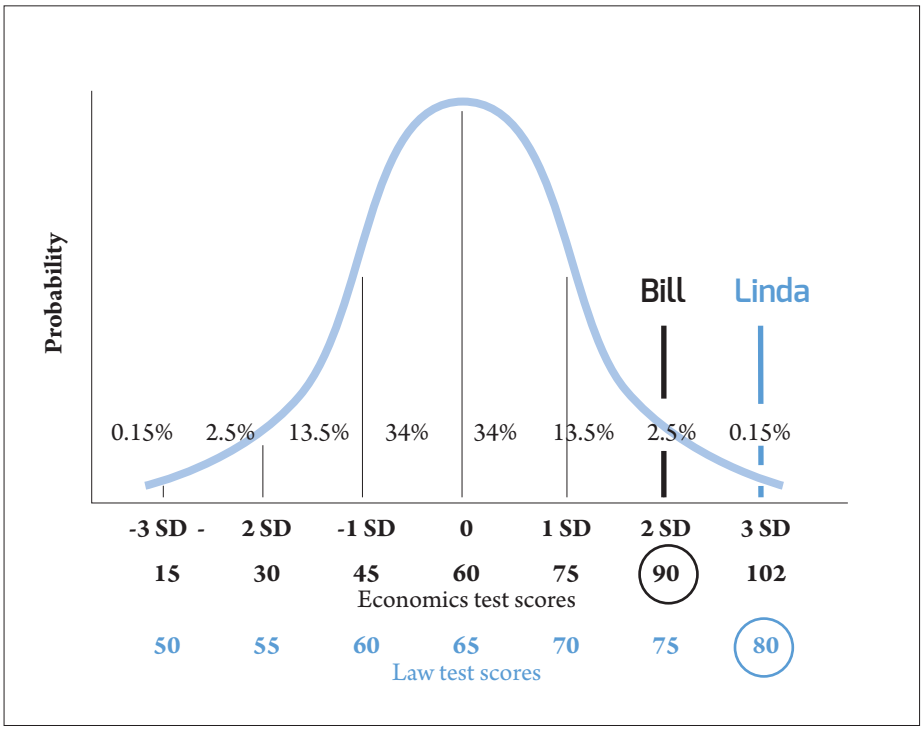

Figure 2. Comparing two distributions of different proportions with the standard deviation. The distributions of scores on the law and economics tests are shown below the standard normal distribution. Linda's raw score of 80 is 3 SD above the mean in her class, and Bill's is 2 SD above the mean in his. Clearly, Linda did relatively better than Bill on her test.

only been seen in small groups, however, so no one knows how tall the average gnome is. Thus, our research question is "How do we estimate the average height of all gnomes if we can only measure a few of them?"

Suppose that a gnome magically appears on your desk. You measure him and find that he is exactly 10 cm tall *(Figure 3)*. What's our best guess about the average height of all gnomes? The answer is 10 cm, because it's all the information we have in a sample size of 1.

Now suppose a second gnome appears beside the first one. This gnome could be 10 cm, but probably he will be a little bigger or a little smaller. Supposed he is 11 cm tall. Now, what is our best guess about the average height of all gnomes? The answer is: 10.5 cm, because it's all the information we have. That is, we average the heights of our sample of two, which is 10.5 cm. In fact, **the sample mean is the best estimate of the population mean** because it uses all the available data.

We could repeat this process if, say, 10 gnomes were to appear: measure the height of each gnome and then calculate their mean height. This sample mean will, again, be the best estimate of the mean height of the gnome population. The same is true for other characteristics of the sample as well: for medians, ranges, and standard deviations, for example.

Figure 3. A gnome 10 cm tall. If this gnome is the only one we've measured, our best estimate of the average height of all gnomes is 10 cm because that is all the information we have. The mean of the sample is best estimate of the mean of the population.

Notice that our sample was small: 10 gnomes out of a population of several thousand gnomes (or so I've been told). With so many gnomes, how likely is it that our estimate, based on 10 gnomes, is accurate? If we happened to get a single sample containing the smallest gnomes, we would underestimate the average height in the population, and if we happened to get a sample containing the largest gnomes, we would overestimate it. What we need is a way to determine how precise our estimate might be. This measure is the confidence interval.

The hypothetical example illustrating confidence intervals

Suppose we have unlimited resources and unlimited cooperation of all the gnomes, such that we can take all possible random samples of, say, 10 gnomes. In other words, we draw a sample of 10 gnomes, measure the height of each, calculate the sample mean, graph the mean, and then return the gnomes to the population. We then draw another sample of 10 gnomes and repeat the process: measure each one, calculate the sample mean, graph the mean, and return the gnomes to the population. We repeat this process until we have taken samples of every possible combination of 10 gnomes *(Table overleaf)*. (You can see why the example is fictitious: agencies funding research into gnomes won't pay for this kind of sampling.)

We repeat this process until we have taken samples of every possible combination of 10 gnomes. (You can see why the example is fictitious: agencies funding research into gnomes won't pay for this kind of sampling.)

When we graph the means of all our samples *(Figure 4)*, we find that they are normally distributed. (This

Sample	Height of each of 10 gnomes in the sample, cm										Sample
No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**means**
1	16	11	5	14	7	13	12	13	15	20	12.6
2	16	12	12	2	4	5	14	7	11	8	9.1
3	1	9	2	6	8	10	4	7	2	10	5.9
4	2	8	3	19	13	9	6	6	14	5	8.5
5	14	4	18	13	12	5	19	11	8	8	11.2
6	14	11	2	2	9	17	11	10	8	16	10
7	5	3	13	11	1	14	13	3	8	7	7.8
8	6	15	13	11	9	13	6	7	15	2	9.7
9	18	14	3	8	14	9	12	7	2	17	10.4
10	3	5	5	2	20	7	14	4	7	7	7.4

Table. The Heights of 100 Gnomes as Collected in 10 Samples of 10 Gnomes.
The overall mean (SD) of the 10 sample means is 9.3 (2.0) cm, which is the best estimate of the mean height (and SD) of the gnome population. The SE equals the standard deviation of the sample (2.0) divided by the square root of the sample size of 10 (3.2), or 0.63. Twice the SE is 1.3, so the mean -2 SE = 8 and the mean +2 SE = 10.6 cm, giving us an estimated mean of 9.3 cm (95% CI, 8 to 10.6 cm). See text for details.

result is explained by what is called "the central limit theorem," which I won't address here.) Remember that the "area under the curve" can be expressed in units of standard deviation. More importantly, the mean of this graph of sample means is, again, our best estimate of the population mean. Now, however, instead of a single sample mean, we have a *distribution* of sample means. When we had a sample of data, we called the measure of dispersion the standard deviation (SD). Now we have a distribution of sample means, so we are going to call the standard deviation the "standard error of the mean (SE)."

The SD and SE represent the same concept and have the same mathematical properties: both can be used to indicate the area under a normal curve. The only difference is that the standard deviation is a descriptive statistic that indicates the variability of a distribution of data, whereas the standard error of the mean is an inferential statistic that indicates the variability of an estimate; that is, the variability of the distribution of the means of all possible samples of the same size.

Remember that about 68% of the data will be included in the range defined by -1 SD below the mean to +1 SD above the mean, and that about 95% will be included between -2 SD and +2 SD. These relationships are the same for the SE: about 68% of the sample means will be included in the range defined by -1 SE below the mean of the sample means to +1 SE above the mean, and about 95% will be included between -2 SE and +2 SE *(Figure 5)*.

The mean of this distribution of sample means is the best estimate of the population mean, and the range given by plus or minus 2 SEs is a 95% CI. In other words, we measured only samples of gnomes, and the estimate varied from sample to sample. However, the mean in 95 of 100 samples of 10 will probably fall within the range defined by 2 SEs above and below the mean of our distribution of sample means.

Calculating confidence intervals

In reality, we generally measure only a single sample. The (measured) sample mean is the best estimate of the population mean, and the 95% CI is calculated from the SE with the simple formula:

$$\text{SE} = \frac{\textbf{Standard deviation of the sample}}{\textbf{Square root of the sample size}}$$

One SE on either side of our mean of sample means is about a 68% CI. To get the

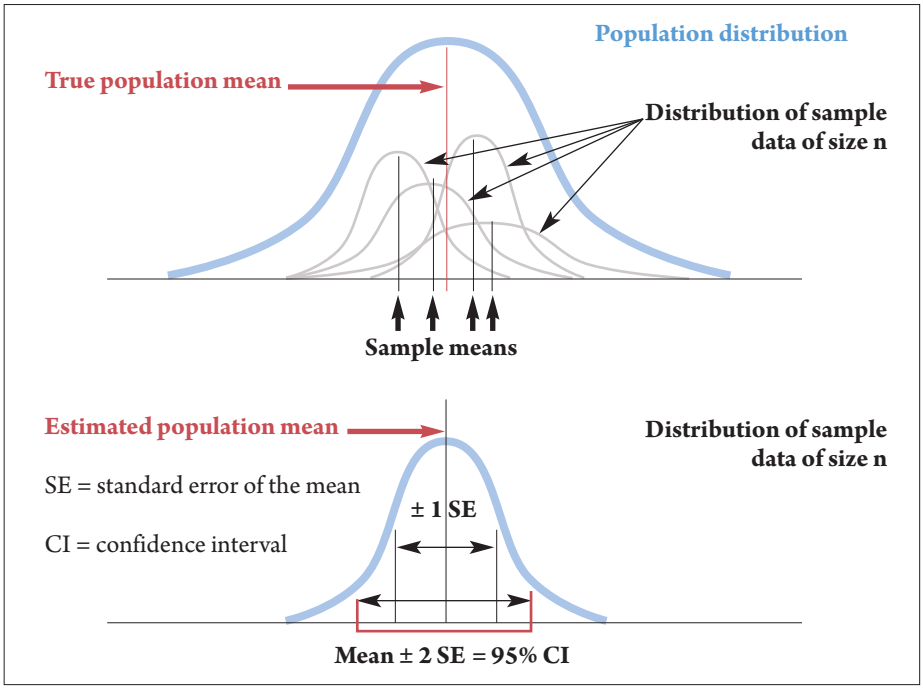

Figure 4. The conceptual process of creating a 95% confidence interval around the estimated mean height.
Upper panel: we take all possible samples of the same size from the population of interest, compute the mean height of each sample, and graph the means. Lower panel: the new distribution of means will be normally distributed, so 95% of the samples we drew had means that ranged between two SEs above and below the overall mean of the new distribution. The overall mean is the estimated height, and the range between the mean plus and minus 2 SEs is the 95% interval for the estimate.

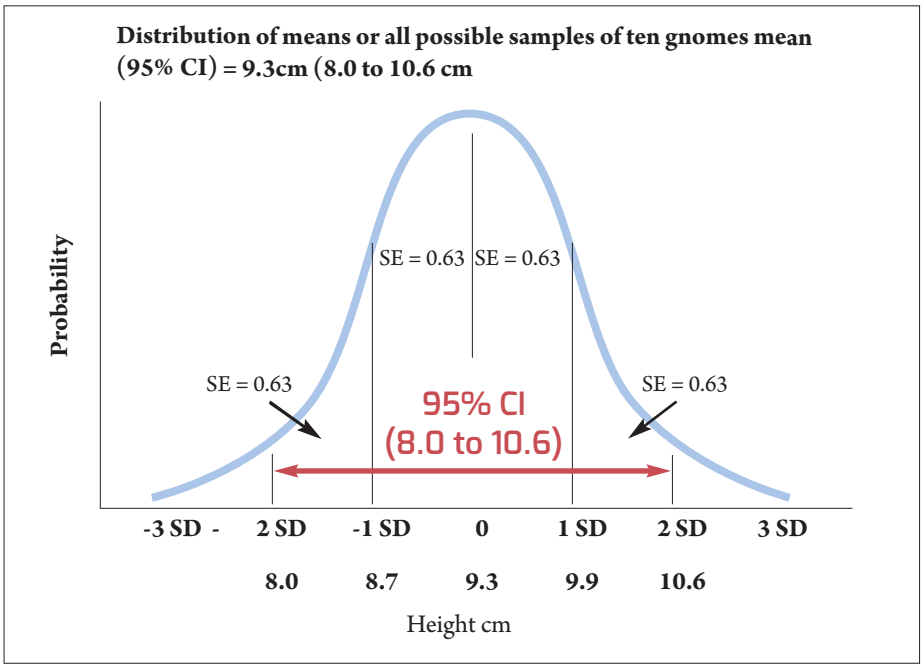

Figure 5. The distribution of sample means in the example (summarised in the Table) of estimating the average height of gnomes.

95% CI, we essentially double the SE, which gives the range of values in which we expect the mean height to fall in 95 of 100 similar samples.

Using data from the example in the table, the mean of the distribution of all possible samples of the same size (although only 10 are shown here) is 9.26 cm. The SE is 1.96, and 2 SEs equal about 3.8. Adding and subtracting the 3.8 to the mean of 9.26 gives us an estimated height of 9.26 cm with a 95% CI of 6.2 to 13.8 cm.

The value of confidence intervals

Confidence intervals have enormous value in reporting the results of medical research. The results of most biomedical studies (that is, the "effect size") are actually estimates and so should be accompanied by CIs. In addition, CIs are increasingly preferred to *P* values when reporting results. The *P* value is a mathematical measure of chance as an explanation and has no biological interpretation. On the other hand, CIs keep the interpretation focused on the biological implications of the effect size.

Here's an example of the value of confident intervals. Consider this sentence:

"The drug reduced diastolic blood pressure (DBP) by a mean of 15 mm Hg (95% CI = 3.5 to 26.5 mm Hg; *P* = 0.01)."

In this particular study, the effect size was a reduction in DPB of 15 mm Hg, and the reduction was statistically significant. That is, if the drug did nothing (the assumption of the null hypothesis), we would expect to get a reduction in DBP of 15 mm Hg or higher *by chance* in only 1 of 100 similar studies. Given that low probability, we decide that the drug was probably responsible for the reduction (we "reject the null hypothesis").

Let's assume that the 15-mm Hg reduction in DBP is clinically important. Although this result is statistically significant and clinically important in this particular study, the 95% CI tells us that the reduction in DBP would probably range from 3.5 to 26.5 mm Hg in 95 of 100 similar studies. A drop of 26.5 mm Hg is clinically important, but a drop of only 3.5 mm Hg probably is not. That is, the confidence is "heterogeneous": it contains both clinically important and clinically unimportant values. So, we can't really say for sure that the drug is effective in 95 of 100 trials; our 15-mm Hg estimate is not precise enough. We need to do the study again, probably with a larger sample, to improve (narrow) the precision of the estimate. When all the values in the CI are clinically important (or when all are not clinically important)–that is, when the CI is "homogenous"–we have a more definitive answer to our question about the efficacy of the drug.

The misuse of the standard error of the mean

The SE is often used incorrectly as a descriptive statistic. Especially in the basic life sciences, measurements are routinely reported as means and SEs. This practice is established and poses no problem to those who are used to seeing measurements presented this way. However, because the SE is always smaller than the SD, it makes measurements look more precise than they would look if they were reported with SDs, so this distortion needs to be kept in mind when interpreting the SE. My research, which mostly concerns statistical reporting in clinical medicine, indicates that the SE is appropriately reported in only a few circumstances, such as in tables reporting regression analysis. The SD is preferred to describe a distribution of data, and the 95% CI is the preferred measure of precision for an estimate.

References

1. Rowntree D. Statistics without tears: an introduction for non-mathematicians. London: Penguin Books; 2000.
2. Lang TA, Secic M. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. 2nd ed. Philadelphia: American College of Physicians; 2006.

Author information

Tom is an international consultant and educator in medical writing/editing, critical appraisal of clinical research, and scientific publications. He is past President of the Council of Science Editors and current Treasurer of the World Association of Medical Editors. He also served on the CONSORT and PRISMA groups and wrote *How to Write, Publish, and Present in the Health Sciences.*