A Closer Look at Confidence Intervals

Kristin L. Sainani, PhD

Confidence intervals give more information than *P* values. They indicate statistical significance, provide a plausible range of values for the true effect, and reveal the precision of the estimate. Compared with *P* values, they also are easier to interpret and are less likely to be misleading. However, most readers have only a vague understanding of how confidence intervals are derived, which leads to some confusion about their meaning. This article aims to provide a deeper understanding of confidence intervals (and related statistical concepts, such as standard errors) and dispel common misconceptions.

HYPOTHETICAL CASE STUDY

For the purposes of illustration, I generated a mock dataset that contains data on cognitive function and serum vitamin D levels for a random sample of 100 European men ages 40-79 years (data are loosely based on Lee et al. [1]). Cognitive function was measured by the digit symbol substitution test (DSST).

Pearson correlation coefficient: A measure of linear correlation. The correlation coefficient is unitless and ranges from −1.0 (perfect negative correlation) to +1.0 (perfect positive correlation). A value of 0 indicates no correlation.

We can use statistics from the sample data to make inferences about European men ages 40-79 years in general. For example, the mean vitamin D level in the sample is 63 nmol/L, and the correlation coefficient between vitamin D and the DSST score is 0.15 (indicating a weak correlation). Thus we can infer that the true mean vitamin D for all European men ages 40-79 years is around 63 nmol/L and that the true correlation between vitamin D and the DSST score is around 0.15. However, we need to put margins of error around these estimates to reflect the uncertainty that comes with taking a single sample. These margins of error are called confidence intervals. To understand exactly how confidence intervals are constructed, one first must understand the concept of a distribution.

Statistic: Any summary measure calculated from data—for example, a mean, difference in means, correlation coefficient, odds ratio, or regression coefficient.

DISTRIBUTIONS

A distribution describes the frequency with which different values (or ranges of values) occur. Distributions are characterized by their shape and by numerical summaries, such as means and standard deviations. Histogram plots are used to display this information visually. One must distinguish between the distribution of a trait and the distribution of a statistic.

Distribution: Describes the frequency with which different values (or ranges of values) occur.

The Distribution of a Trait

It is straightforward to understand the distribution of a trait, such as vitamin D levels or a DSST score. For example, a DSST score follows a normal distribution, which means that the distribution has a bell-like shape (Figure 1A). The mean is approximately 28 and the

K.L.S. Department of Health Research and Policy, Division of Epidemiology, Stanford University, HRP Redwood Building, Stanford, CA 94305. Address correspondence to K.L.S.; e-mail: kcobb@stanford.edu

Disclosure: nothing to disclose

Submitted for publication October 10, 2011; accepted October 10, 2011.

Vol. 3, 1134-1141, December 2011
DOI: 10.1016/j.pmrj.2011.10.005

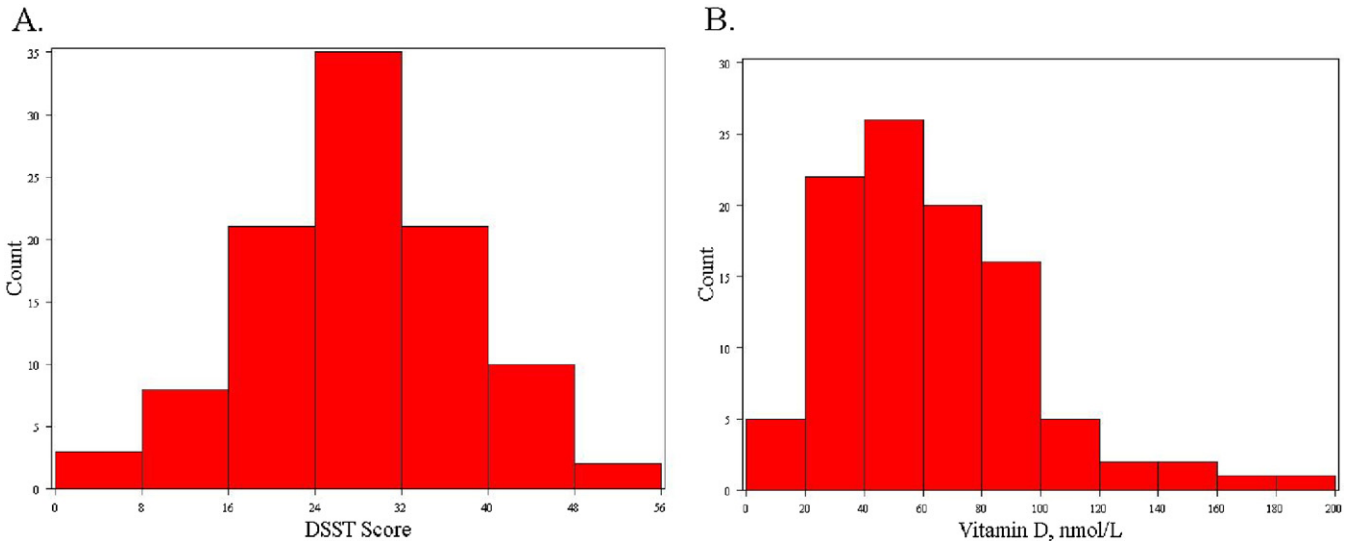

Figure 1. Histograms displaying the distributions of digit symbol substitution test (DSST) scores (A) and serum vitamin D levels (B) in the sample. The y axis gives the number of men in each range of values. For example, 3 men had DSST scores between 0 and 8, and 35 had DSST scores between 24 and 32.

standard deviation (the average scatter around the mean) is approximately 10.

Normal distribution: A distribution with a symmetric, curved, bell-like shape.

Standard deviation: A measure of variability; roughly, the average distance from the mean.

Many biological traits, such as brain function or intelligence quotient, follow a normal distribution. Most people cluster within a certain distance from the mean, but a few people have abnormally low or high values. Normal distributions have useful mathematical properties; for example, when a variable is normally distributed, approximately 68% of observations will fall within 1 standard deviation of the mean (for DSST, 18 to 38); approximately 95% will fall within 2 standard deviations of the mean (for DSST, 8 to 48); and about 99.7% will fall within 3 standard deviations of the mean (for DSST, −2 to 58). This phenomenon is called the 68-95-99.7 rule.

68-95-99.7 rule: On a normal distribution, 68% of observations fall within 1 standard deviation of the mean, 95% fall within 2 standard deviations of the mean, and 99.7% fall within 3 standard deviations of the mean.

Vitamin D has a more "right-skewed" distribution—that is, a string of abnormally high values forms a "tail" to the right (Figure 1B). The mean is 63 nmol/L and the standard deviation is 33 nmol/L. Right-skewed distributions are common in medicine; because many biological measures cannot be lower than zero, the range is restricted on the left but not on the right.

The Distribution of a Statistic

Like traits, statistics follow distributions. However, the distribution of a statistic is a trickier concept. A statistic is one number. For example, the mean vitamin D in our hypothetical sample was 63 nmol/L. How can a single number have a distribution?

The key is that the distribution of a statistic is a theoretical construct. Statisticians ask a thought experiment: What would happen if one could repeat a particular study over and over again with different samples of the same size? How much would the value of the statistic (eg, the sample mean or correlation coefficient) fluctuate from sample to sample? By answering this question, statisticians are able to pinpoint exactly how much uncertainty is associated with a given statistic.

Of course, it is not possible to actually repeat the same study over and over, but statisticians can simulate this process. In computer simulation, researchers repeat the same virtual study many times and directly observe the results. (Statisticians also can predict the behavior of statistics by using mathematical theory, but computer simulation is more intuitive.)

For example, I performed a computer simulation to determine the distribution of the sample mean for vitamin D. I set up a virtual population of European men ages 40-79 years, took repeated random samples of 100 men from this population, and calculated the mean vitamin D for each sample. The steps of the simulation are as follows:

1. Specify the underlying distribution of vitamin D in all European men ages 40-79 years (eg, right-skewed, standard deviation = 33 nmol/L, mean = 62 nmol/L).

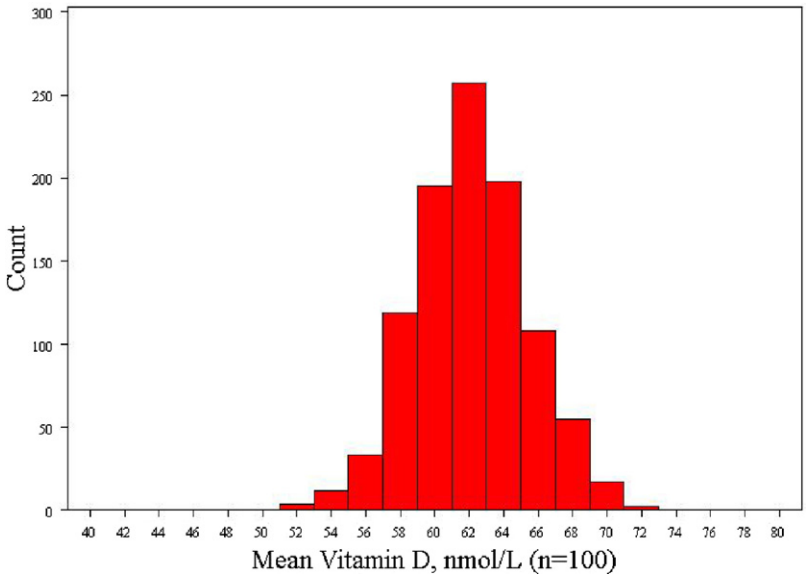

Figure 2. The distribution of mean vitamin D levels in 1000 samples of 100 men (computer simulation). The true mean value was set at 62 nmol/L, but this value is arbitrary—the mean does not affect the variability of the statistic.

2. Select a random sample of 100 virtual men from the population.
3. Calculate the mean vitamin D for the sample.
4. Repeat steps (2) and (3) a large number of times, say 1000 times. (Note: The number of repetitions in a simulation is arbitrary—it just has to be a "large number"; we will get similar results whether we use 500, 1000, or 10,000 repetitions.)
5. Explore the distribution of the 1000 means.

Figure 2 shows a histogram plot for the 1000 means. Surprisingly, the means follow a (near) normal distribution! Vitamin D itself (the trait) follows a right-skewed distribution; however, mean vitamin D (the statistic) follows a normal distribution. Why would this be? Only a few men in each sample have extreme values, and these extreme values are drowned out in the average; thus the right tail disappears. High and low values also tend to balance each other out; thus the means cluster close to the true mean (which I arbitrarily set at 62 nmol/L). The variability is also reduced: The range is 51-74 nmol/L (compare this range with the range for the trait: 0-200 nmol/L), and the standard deviation is just 3.3 nmol/L (compare this standard deviation with the standard deviation of the trait: 33 nmol/L).

Standard error: A measure of the variability of a statistic.

The standard deviation of a statistic is also called a standard error (to distinguish it from the standard deviation of a trait). The standard error of a mean is affected by the sample size and by the variability of the underlying trait (Figure 3). For example, I performed a simulation in which I took samples of 400 men rather than 100 men (Figure 3A). The standard error shrunk to 1.7 nmol/L, reflecting the fact that there is less variation between bigger samples. I also performed a simulation in which I increased the standard deviation of vitamin D from 33 nmol/L to 40 nmol/L; this change increased the standard error to 4.0 nmol/L (Figure 3B).

Statisticians have worked out the distributions of most statistics. For example, sample means follow a normal distribution for large samples and a t-distribution for small samples ($n < 100$). (A t-distribution is just a normal distribution with slightly fatter tails.) The standard error of a sample mean is $\frac{SD}{\sqrt{n}}$, where SD is the standard deviation of the trait (eg, vitamin D) and n is the sample size. For example, for our mock dataset, the standard error for mean vitamin D is:

$$\frac{SD}{\sqrt{n}} = \frac{33}{\sqrt{100}} = 3.3 \text{ nmol/L}$$

Correlation coefficients also follow a normal distribution (or a t-distribution for small samples), with a standard error of approximately $\frac{1 - r^2}{\sqrt{n}}$, where r is the true correlation coefficient. For example, if the true correlation coefficient is 0.15, the standard error is:

$$\frac{1 - 0.15^2}{\sqrt{100}} = 0.1$$

To illustrate, I performed a computer simulation for the correlation coefficient. I had the computer take 15,000 samples of 100 men, assuming a true correlation between vitamin D and a DSST score of 0.15. The computer simulation reveals that the correlation coefficient indeed follows a normal distribution with a standard error of 0.1 (Figure 4).

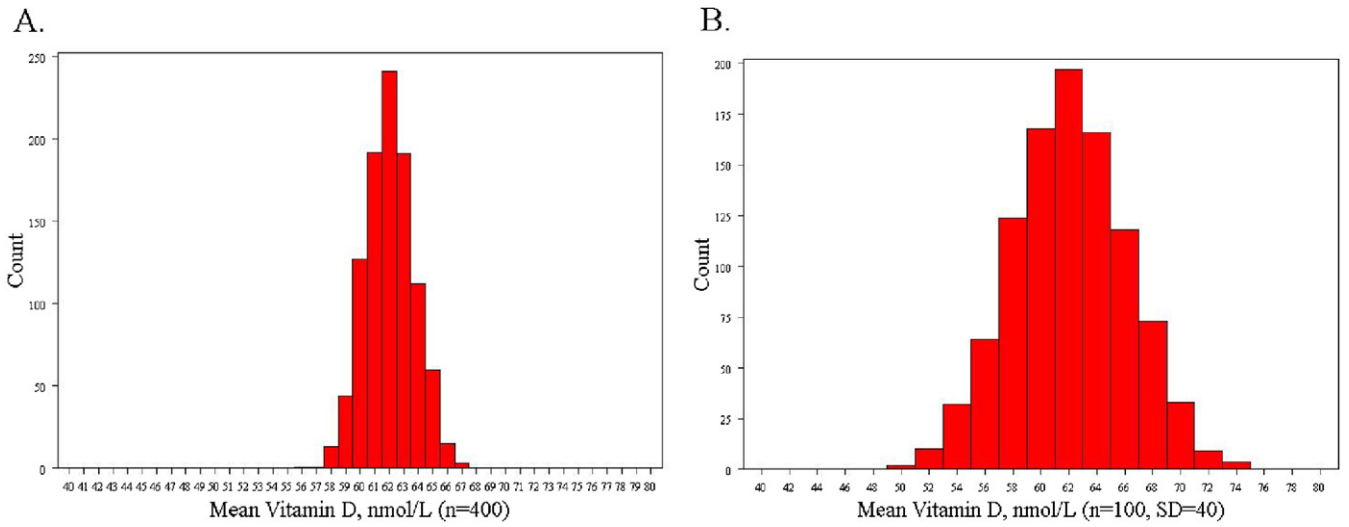

Figure 3. The distribution of mean vitamin D levels in 1000 samples when the sample size is increased to 400 (A) and when the standard deviation (SD) of vitamin D is increased to 40 nmol/L (B).

Many common statistics follow normal distributions (or t-distributions), including proportions, differences in proportions, differences in means, and regression coefficients. This is extremely useful because the normal distribution is predictable.

BUILDING A CONFIDENCE INTERVAL

Once we know the distribution of a given statistic (including the shape and standard error), building a confidence interval is fairly straightforward. The goal of the confidence interval is to capture the true effect (eg, the true mean or the true correlation coefficient) most of the time. For example, a 95% confidence interval should include the true effect approximately 95% of the time, and a 99% confidence interval should include the true effect approximately 99% of the time.

Mean vitamin D follows a normal distribution (Figure 2). Consequently, 95% of studies will yield a sample mean that falls within 2 standard errors of the true mean (because of the 68-95-99.7 rule for normal distributions, reviewed previously). Thus, to ensure that we cross the true mean 95% of the time, we just need to add a margin of ±2 standard errors to the sample mean.

For example, if the true mean is 62 nmol/L and the standard error is 3.3 nmol/L, then there is a 95% chance that

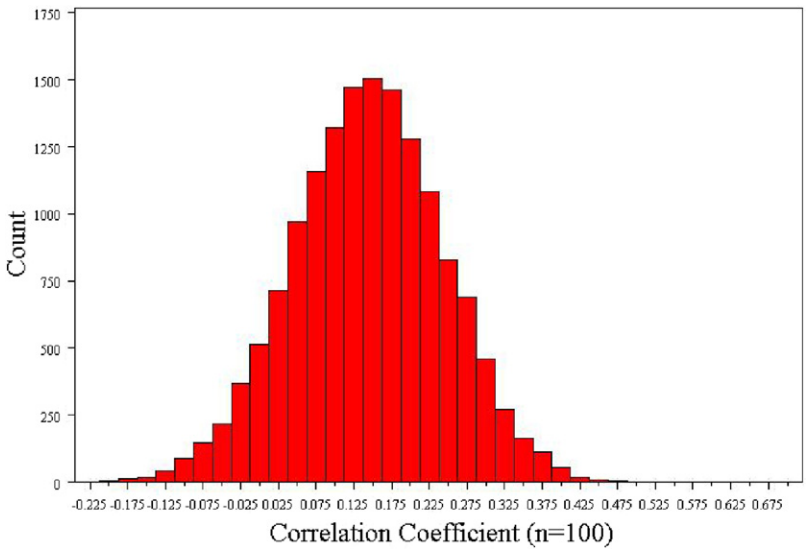

Figure 4. The distribution of a correlation coefficient in 15,000 samples of 100 men (computer simulation). The true correlation coefficient was set at a value of 0.15.

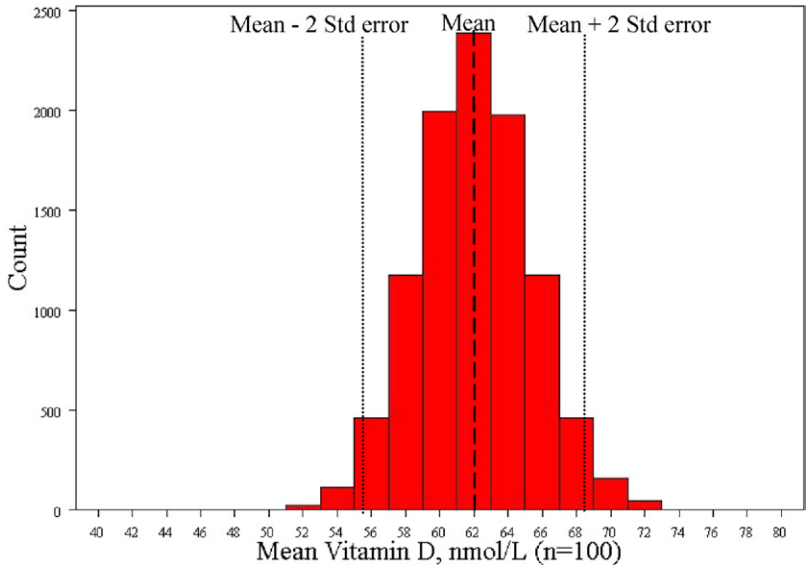

Figure 5. The distribution of mean vitamin D levels in 1000 samples of 100 men (computer simulation). The true mean is 62 nmol/L and the standard error is 3.3 nmol/L; 95% of sample means fall within 2 standard errors of the mean. Std = standard.

the sample mean will fall between 55.4 nmol/L and 68.6 nmol/L (Figure 5). If we construct an interval of ±2 standard errors for any value in this range, the interval will include the true mean. For example, if the sample mean is 68.6 nmol/L, the interval is 62.0 to 75.2 (68.6 ± 6.6), which just hits 62.

Thus to build a 95% confidence interval for any statistic that follows a normal distribution, we use the following formula:

$$\text{Sample value} \pm 2\ x\ (\text{standard error})$$

For example, the 95% confidence interval for mean vitamin D in our original mock dataset is:

$$63\ \text{nmol/L} \pm 2\ x\ (3.3) = 56.4{-}69.6\ \text{nmol/L}$$

and the 95% confidence interval for the correlation coefficient is

$$0.15 \pm 2\ x\ (0.1) = -0.05{-}0.35$$

To demonstrate that this approach works, I performed a simulation of 20 studies of 100 men. For each study, I calculated the 95% confidence interval for the mean by using the aforementioned formula. I then plotted these 20 confidence intervals (Figure 6). As expected, 1 of 20 confidence intervals (5%) missed the true mean, and the remainder (95%) hit it.

If we want to capture the true effect more than 95% of the time, we need to make the confidence interval wider. On a normal distribution, 99% of observations fall within about 2.6 standard errors of the mean, so the 99% confidence interval is:

$$\text{Sample value} \pm 2.6\ x\ (\text{standard error})$$

For example, the 99% confidence interval for mean vitamin D in our mock dataset is as follows:

$$63\ \text{nmol/L} \pm 2.6\ x\ (3.3) = 54.4{-}71.6$$

A standard normal chart provides the values for other confidence levels (such as 90% confidence).

INTERPRETING CONFIDENCE INTERVALS

Table 1 provides examples of correct interpretations of confidence intervals. The confidence interval gives a plausible range of values for the true effect. For example, we can be 95% certain that the true mean vitamin D level in European men ages 40-79 years is between 56.4 and 69.6 nmol/L, and we can be 99% certain that it is between 54.4 and 71.6 nmol/L. We also can be 95% sure that the true correlation between vitamin D and the DSST score in this population is between −0.05 and +0.35. The fact that the interval is wide tells us that we have estimated the effect only imprecisely.

Confidence intervals also provide information about statistical significance. The 95% confidence interval for the correlation coefficient includes 0, the null value (no correlation). Thus the association between vitamin D and DSST score is not significant at the 0.05 significance level (ie, $P > .05$). A 95% confidence interval that excluded 0 would indicate a statistically significant association between vitamin D and the DSST score ($P < .05$).

MISCONCEPTIONS ABOUT CONFIDENCE INTERVALS

Although confidence intervals are easier to understand and interpret than *P* values, some misconceptions still surround them. Table 1 gives examples of some incorrect interpreta-

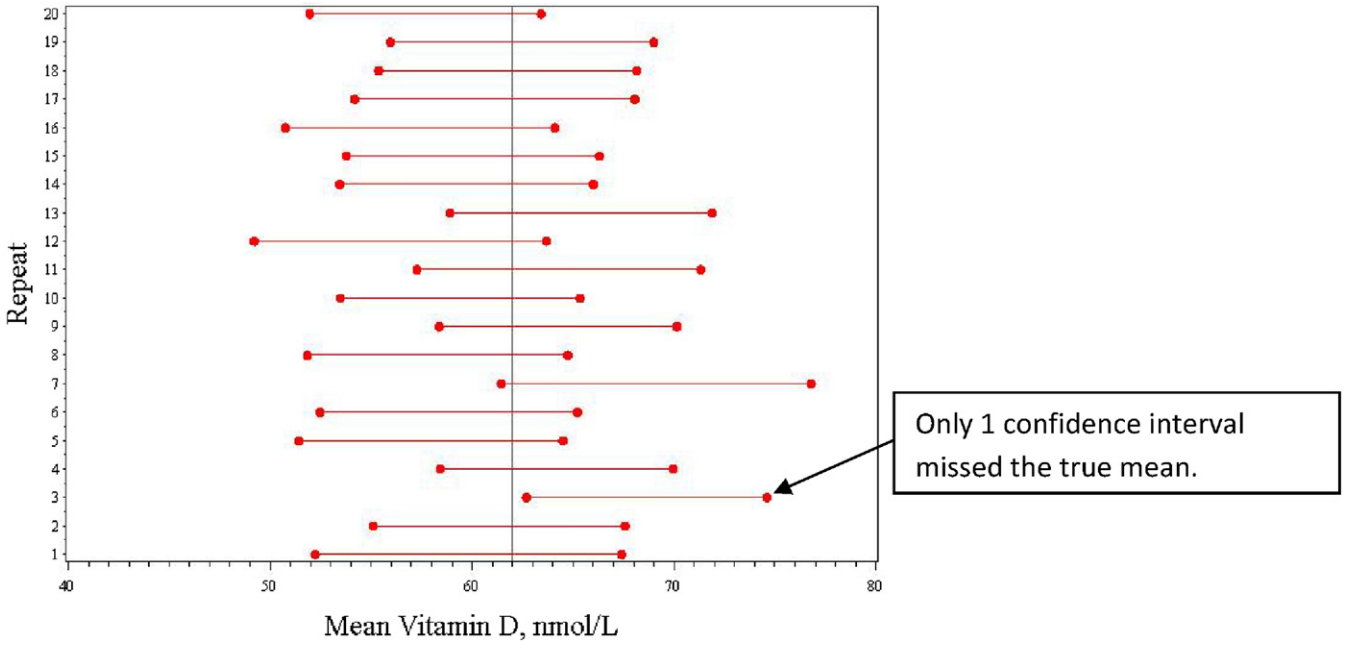

Figure 6. The 95% confidence intervals from 20 studies of 100 men (computer simulation). The solid vertical line at 62 is the true mean. Only 1 confidence interval (5%) missed the true mean.

tions of confidence intervals that arise from these misconceptions.

Confidence Intervals Are About Statistics, Not Individuals

A common misconception is that confidence intervals provide a reference range for the trait. For example, one might conclude that 95% of European men ages 40-79 years have a vitamin D level between 56.4 and 69.6 nmol/L. However, confidence intervals are about statistics, not individual values. Just consider Figure 1B. Only a small proportion of men in the sample have vitamin D levels between 56.4 and 69.6 nmol/L.

Table 1. *Examples of correct and incorrect interpretations of the 95% confidence level: 56.4 to 69.6 nmol/L*

Correct interpretations
The plausible range of values for the true mean vitamin D level is between 56.4 and 69.6 nmol/L.
We can be 95% confident that the true mean vitamin D level is between 56.4 and 69.6 nmol/L.
The mean vitamin D level is significantly lower than 70 nmol/L ($P < .05$).
Incorrect interpretations:
Ninety-five percent of persons in the population have vitamin D values between 56.4 and 69.6 nmol/L.
There is a 95% chance that a randomly selected person has a vitamin D value between 56.4 and 69.6 nmol/L.
There is a 95% chance that the true mean is between 56.4 and 69.6 nmol/L.

The confusion likely arises because the formula for a reference range sometimes resembles the formula for a confidence interval. If a trait is normally distributed (such as DSST), the 95% reference range is the mean ±2 standard deviations. However, a reference range is based on standard deviations, whereas a confidence interval is based on standard errors, and thus they are quite different entities.

Confidence Intervals Are Uncertain; Effects Are Not

Authors and readers should be careful about how confidence intervals are described in words. It is not correct to say, "There is a 95% chance that the true mean is between 56.4 nmol/L and 69.6 nmol/L." This is a subtle point, but the problem with this statement is that the true mean is fixed, that is, it is either in the interval or it is not in the interval, so the "chance" of it being in the interval is either 100% or 0%. The uncertainty is a property of the confidence interval, not the mean (see Figure 6; notice how the confidence intervals shift, not the mean). Thus it is more appropriate to say, "There is a 95% chance that the confidence interval hits the mean," rather than, "There's a 95% chance that the mean is in the confidence interval." Thinking carefully about the subtle distinction between these two statements may confer some deeper philosophical insight into the meaning of a confidence interval.

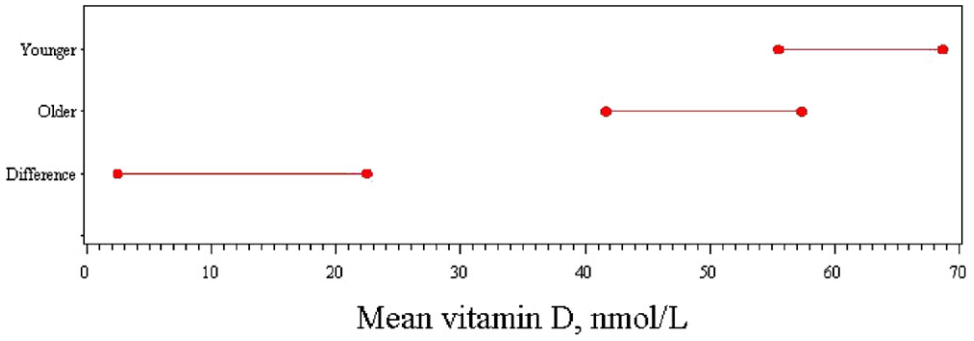

Figure 7. The 95% confidence intervals for the mean vitamin D levels for younger and older men overlap, but the means are significantly different ($P < .05$), as illustrated by the fact that the 95% confidence interval for the difference in means excludes 0.

Effects with Overlapping Confidence Intervals May Still Be Significantly Different

When comparing confidence intervals for two effects (for example, two means), many people use a simple rule: if the confidence intervals don't overlap, the effects are statistically different; otherwise, they are statistically indistinguishable. Surprisingly, however, this rule is erroneous; whereas the former statement is correct, the latter statement is not always true. For example, when the confidence intervals for the means of two independent groups just touch, the *P* value for their difference is actually approximately 0.006, not 0.05 [2,3].

To illustrate, I divided the mock dataset into older men (≥60 years) and younger men (<60 years) and estimated the mean vitamin D level in these 2 groups. The 95% confidence interval for the mean for younger men was 55.5 to 68.7 nmol/L, and for older men it was 41.7 to 57.4 nmol/L (Figure 7). These two intervals overlap from 55.5 to 57.4. However, mean vitamin D level is significantly different between the two groups ($P < .05$). The 95% confidence interval for the difference in means is 2.5 to 22.5, which excludes 0 (Figure 7).

This result may seem counterintuitive. If 56 nmol/L is a plausible value for both means, then isn't it plausible that the means are equal? The issue is that statistics for between-group comparisons follow a different distribution than statistics for single groups. The between-group comparison must account for how 2 means are changing simultaneously. In essence, 56 nmol/L might be a plausible value for both means, but it may not be plausible that 56 nmol/L will occur in both groups at the same time. See the In-Depth box for a more mathematical proof.

IN-DEPTH: Effects With Overlapping Confidence Intervals May Still Be Significantly Different

Assume that there are two independent groups (group A and group B).

The mean for group A = $\overline{X_A}$

The mean for group B = $\overline{X_B}$

Assume that the standard deviation of the trait (SD) and the sample size (n) are equal in the 2 groups and that $\overline{X_A} > \overline{X_B}$.

The standard error for a single mean is: $\frac{SD}{\sqrt{n}}$

The standard error of the difference in means $= \sqrt{\frac{SD}{n}+\frac{SD}{n}} = \sqrt{2}x\frac{SD}{\sqrt{n}}$.

The 95% confidence intervals for the individual means ($\overline{X_A}$ and $\overline{X_B}$) will overlap if the difference in means is less than 4 standard errors (because the error bar for each confidence interval equals 2 standard errors):

$\overline{X_A} - \overline{X_B} < 4\ x\ \frac{SD}{\sqrt{n}}$

However, the means will be statistically distinct ($P < .05$) if their difference exceeds 2 standard errors *of the difference in means* (because the confidence interval for the difference will exclude 0):

$\overline{X_A} - \overline{X_B} < 2\sqrt{2}x\frac{SD}{\sqrt{n}} = 2.82x\frac{SD}{\sqrt{n}}$

Thus in cases in which $2.82x\frac{SD}{\sqrt{n}} < \overline{X_A} - \overline{X_B} < 4x\frac{SD}{\sqrt{n}}$, the difference in means is statistically significant but the individual confidence intervals overlap.

CONCLUSIONS

Confidence intervals give a plausible range of values for the true population value. The confidence interval is based on the distribution of a statistic, which is a theoretical construct; 95% confidence intervals include the true effect about 95% of the time and miss it about 5% of the time (whereas 99% confidence intervals span the true effect approximately 99% of the time and miss it approximately 1% of the time.) Confidence intervals should not be confused with reference ranges, which are about individuals, not about statistics. Statistically distinct effects sometimes can have overlapping confidence intervals.

REFERENCES

1. Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry 2009;80:722-729.
2. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. Am Stat 2001;55,182-186.
3. Belia S, Fidler F, Williams J, Cumming G. Researchers misunderstand confidence intervals and standard error bars. Psychol Methods 2005;10:389-396.