POINTS OF SIGNIFICANCE

Error bars

The meaning of error bars is often misinterpreted, as is the statistical significance of their overlap.

Last month in Points of Significance, we showed how samples are used to estimate population statistics. We emphasized that, because of chance, our estimates had an uncertainty. This month we focus on how uncertainty is represented in scientific publications and reveal several ways in which it is frequently misinterpreted.

The uncertainty in estimates is customarily represented using error bars. Although most researchers have seen and used error bars, misconceptions persist about how error bars relate to statistical significance. When asked to estimate the required separation between two points with error bars for a difference at significance $P = 0.05$, only 22% of respondents were within a factor of 2 (ref. 1). In light of the fact that error bars are meant to help us assess the significance of the difference between two values, this observation is disheartening and worrisome.

Here we illustrate error bar differences with examples based on a simplified situation in which the values are means of independent (unrelated) samples of the same size and drawn from normal populations with the same spread. We calculate the significance of the difference in the sample means using the two-sample *t*-test and report it as the familiar *P* value. Although reporting the exact *P* value is preferred, conventionally, significance is often assessed at a $P = 0.05$ threshold. We will discuss *P* values and the *t*-test in more detail in a subsequent column.

The importance of distinguishing the error bar type is illustrated in **Figure 1**, in which the three common types of error bars—standard deviation (s.d.), standard error of the mean (s.e.m.) and confidence interval (CI)—show the spread in values of two samples of size $n = 10$ together with the *P* value of the difference in sample means. In **Figure 1a**, we simulated the samples so that each error bar type has the same length, chosen to make them exactly abut. Although these three data pairs and their error bars are visually identical, each represents a different data scenario with a different *P* value. In **Figure 1b**, we fixed the *P* value to $P = 0.05$ and show the length of each type of bar for this level of significance. In this latter scenario, each of the three pairs of points represents the same pair of samples, but the bars have different lengths because they indicate different statistical properties of the same data. And because each bar is a different length, you are likely to interpret each one quite differently. In general, a gap between bars

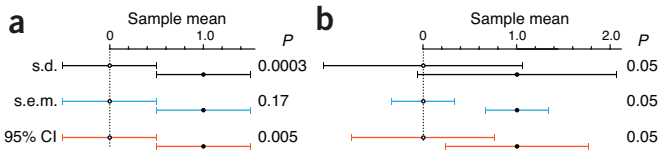

Figure 1 | Error bar width and interpretation of spacing depends on the error bar type. (**a**,**b**) Example graphs are based on sample means of 0 and 1 ($n = 10$). (**a**) When bars are scaled to the same size and abut, *P* values span a wide range. When s.e.m. bars touch, *P* is large ($P = 0.17$). (**b**) Bar size and relative position vary greatly at the conventional *P* value significance cutoff of 0.05, at which bars may overlap or have a gap.

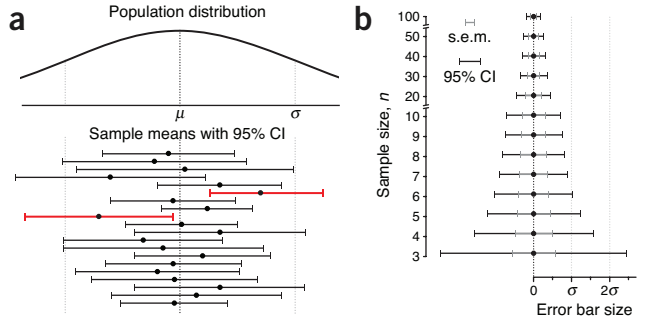

Figure 2 | The size and position of confidence intervals depend on the sample. On average, CI% of intervals are expected to span the mean—about 19 in 20 times for 95% CI. (**a**) Means and 95% CIs of 20 samples ($n = 10$) drawn from a normal population with mean μ and s.d. σ. By chance, two of the intervals (red) do not capture the mean. (**b**) Relationship between s.e.m. and 95% CI error bars with increasing *n*.

does not ensure significance, nor does overlap rule it out—it depends on the type of bar. Chances are you were surprised to learn this unintuitive result.

The first step in avoiding misinterpretation is to be clear about which measure of uncertainty is being represented by the error bar. In 2012, error bars appeared in *Nature Methods* in about two-thirds of the figure panels in which they could be expected (scatter and bar plots). The type of error bars was nearly evenly split between s.d. and s.e.m. bars (45% versus 49%, respectively). In 5% of cases the error bar type was not specified in the legend. Only one figure[2] used bars based on the 95% CI. CIs are a more intuitive measure of uncertainty and are popular in the medical literature.

Error bars based on s.d. inform us about the spread of the population and are therefore useful as predictors of the range of new samples. They can also be used to draw attention to very large or small population spreads. Because s.d. bars only indirectly support visual assessment of differences in values, if you use them, be ready to help your reader understand that the s.d. bars reflect the variation of the data and not the error in your measurement. What should a reader conclude from the very large and overlapping s.d. error bars for $P = 0.05$ in **Figure 1b**? That although the means differ, and this can be detected with a sufficiently large sample size, there is considerable overlap in the data from the two populations.

Unlike s.d. bars, error bars based on the s.e.m. reflect the uncertainty in the mean and its dependency on the sample size, *n* (s.e.m. = s.d./$\sqrt{n}$). Intuitively, s.e.m. bars shrink as we perform more measurements. Unfortunately, the commonly held view that "if the s.e.m. bars do not overlap, the difference between the values is statistically significant" is incorrect. For example, when $n = 10$ and s.e.m. bars just touch, $P = 0.17$ (**Fig. 1a**). Conversely, to reach $P = 0.05$, s.e.m. bars for these data need to be about 0.86 arm lengths apart (**Fig. 1b**). We cannot overstate the importance of recognizing the difference between s.d. and s.e.m.

The third type of error bar you are likely to encounter is that based on the CI. This is an interval estimate that indicates the reliability of a measurement[3]. When scaled to a specific confidence level (CI%)—the 95% CI being common—the bar captures the population mean CI% of the time (**Fig. 2a**). The size of the s.e.m. is compared to the 95% CI in **Figure 2b**. The two are related by the *t*-statistic, and in large samples the s.e.m. bar can be interpreted as a CI with a confidence level of 67%. The size of the CI depends on *n*; two useful approximations for the CI are 95% CI ≈ 4 × s.e.m ($n = 3$) and 95% CI ≈ 2 × s.e.m. ($n > 15$).

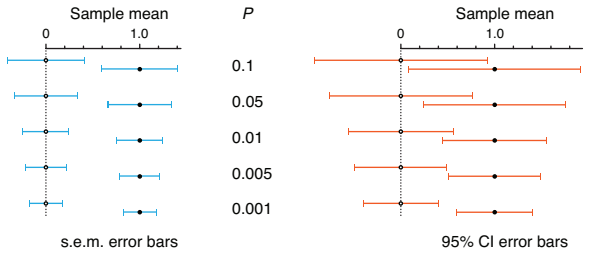

Figure 3 | Size and position of s.e.m. and 95% CI error bars for common *P* values. Examples are based on sample means of 0 and 1 ($n = 10$).

A common misconception about CIs is an expectation that a CI captures the mean of a second sample drawn from the same population with a CI% chance. Because CI position and size vary with each sample, this chance is actually lower.

This variety in bars can be overwhelming, and visually relating their relative position to a measure of significance is challenging. We provide a reference of error bar spacing for common *P* values in **Figure 3**. Notice that $P = 0.05$ is not reached until s.e.m. bars are separated by about 1 s.e.m, whereas 95% CI bars are more generous and can overlap by as much as 50% and still indicate a significant difference. If 95% CI bars just touch, the result is highly significant ($P = 0.005$). All the figures can be reproduced using the spreadsheet available in **Supplementary Table 1**, with which you can explore the relationship between error bar size, gap and *P* value.

Be wary of error bars for small sample sizes—they are not robust, as illustrated by the sharp decrease in size of CI bars in that regime (**Fig. 2b**). In these cases (e.g., $n = 3$), it is better to show individual data values. Furthermore, when dealing with samples that are related (e.g., paired, such as before and after treatment), other types of error bars are needed, which we will discuss in a future column.

It would seem, therefore, that none of the error bar types is intuitive. An alternative is to select a value of CI% for which the bars touch at a desired *P* value (e.g., 83% CI bars touch at $P = 0.05$). Unfortunately, owing to the weight of existing convention, all three types of bars will continue to be used. With our tips, we hope you'll be more confident in interpreting them.

Martin Krzywinski & Naomi Altman

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2659).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Belia, S.F., Fidler, F., Williams, J. & Cumming, G. *Psychol. Methods* **10**, 389–396 (2005).
2. Frøkjær-Jensen, C., Davis, M.W., Ailion, M. & Jorgensen, E.M. *Nat. Methods* **9**, 117–118 (2012).
3. Cumming, G., Fidler, F. & Vaux, D.L. *J. Cell. Biol.* **177**, 7–11 (2007).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.