CrossMark

Measures of variability and precision in statistics: appreciating, untangling and applying concepts

Adam Ciarleglio

SUMMARY

This reflection presents a discussion of some common measures of variability and how they are appropriately used in descriptive and inferential statistical analyses. We argue that confidence intervals (CIs), which incorporate these measures, serve as tools to assess both clinical and statistical significance.

KEYWORDS

Standard deviation; standard error; confidence interval; descriptive statistics; inference.

In this reflection, we will discuss: (1) the importance of considering measures of variation when describing the distribution of a variable; (2) the difference between two commonly confused measures of variation, the standard deviation (s.d.) and standard error (s.e.); and (3) confidence intervals (CIs) as indicators of precision and tools for inference.

Centre and spread: qualities of a distribution

Consider the study information and statistics shown in Box 1. For the moment, we will focus on the group who converted to psychosis ('converters') and assume that, even though the size is small, the sample is representative of the population of all converters.

If our goal is to learn about the distribution of striatal glutamate values among converters, for either descriptive or inferential purposes, we might begin by asking two questions. The first question is: What is a typical striatal glutamate value for converters?

The sample mean gives one possible answer to this question. Using the formula provided in Box 2, the value of the sample mean is 30.25 IU. This value serves as an estimate for the mean striatal glutamate for the population of all converters. On its own, the sample mean provides important but limited information about the distribution of the striatal glutamate values. To get a more complete

characterisation of the distribution, we also need to know how variable the individual values are.

This brings us to the second question: How spread out are individual striatal glutamate values among the converters? The standard deviation offers one possible answer to this question. Before defining the standard deviation, we first need to discuss variance. The formula for the sample variance is shown in Box 2. Like the sample mean, the sample variance is an estimate for a population parameter, namely the population variance. The formula shows that the sample variance is the sum of all the squared deviations (differences between each individual value and the sample mean) divided by 1 less than the sample size. The sample variance tells us how large a typical squared deviation is for values from the distribution. The sample standard deviation (s.d.) is the square root of the sample variance and is a more natural measure of variation to report, since its units are always on the same scale as the individual observations. In our example, the sample s.d. of 4.59 IU means that a typical striatal glutamate observation for converters will fall within 4.59 IU of the sample mean of 30.25 IU. Since the sample s.d. for the non-converters is smaller than the sample s.d. for the converters, we

Adam Ciarleglio, PhD, is an Assistant Professor of Biostatistics in the Department of Biostatistics and Bioinformatics at the Milken Institute School of Public Health, George Washington University, Washington, DC, USA. His research focuses on developing and applying statistical methods to address mental health research questions.
Correspondence Adam Ciarleglio. Email: aciarleglio@gwu.edu

BOX 1 Example study

Kegeles et al (2019) collected data on a small sample (*n* = 19) of individuals at clinical high risk for converting to psychosis. The individuals were followed up to see who converted within the 2-year study period. At baseline, investigators collected striatal glutamate proton magnetic resonance spectroscopy (^{1}H MRS) data (measured in international units (IU) (Cecil 2013)), among other measures. The sample statistics for the striatal glutamate values for those who converted and those who did not are as follows:

converters (*n* = 7): mean 30.25 IU, s.d. = 4.49 IU

non-converters (*n* = 12): mean 25.86 IU, s.d. = 3.84 IU

know that the values for the non-converters are more tightly clustered around their sample mean.

Standard deviation and standard error: what's the difference?

Another measure of variation that is sometimes confused with the standard deviation is the standard error (s.e.). The standard error of the sample mean (s.e.m.) is a measure of how much variation there is in the sample mean itself. To understand this, we need to appreciate that the value of the sample mean depends on the sample from which it is computed and therefore can differ from one sample to the next. If we could perform repeated sampling by randomly selecting samples of the same size from the population, compute the sample mean for each sample and find the standard deviation of those sample means, we would have the s.e.m.. This notion of repeated sampling from the population underlies many important concepts in statistics and we will see it again below. To be clear, the s.e.m. is a standard deviation, but it is a standard deviation of the sample mean values rather than of the individual values.

Fortunately, there is a formula to compute the s.e.m. and it is given in Box 2. For our example, the s.e.m. for the converters is $4.59/\sqrt{7} = 1.73$ IU. Notice that the formula for the s.e.m. depends on two quantities: the standard deviation for the individual sample values and the sample size. It is clear that a larger sample size would have resulted in a smaller s.e.m., holding the sample s.d. fixed at 4.59. For example, had the sample size been 70 rather than 7 then the s.e.m. would have been $4.59/\sqrt{70} = 0.55$. Although the s.e.m. formula may not be intuitive, the effect of a larger sample on the s.e.m. should be: the mean from a larger sample is less variable than the mean from a smaller sample.

It can be mathematically proven that the mean of all possible sample means from those repeated samples is equal to the population mean that we

want to estimate. So, if the s.e.m. is small, then the distribution of the sample means is tightly clustered around the population mean. If our sample mean comes from a distribution with a small s.e.m., then we should be optimistic that our sample mean is close to the true value in the population.

Although the s.e.m. is a type of standard deviation, it provides very different information from the standard deviation. The sample s.d. is useful as a descriptive statistic, providing an estimate for the variability in the individual values. In contrast, the s.e.m. should be thought of as a measure of the precision of the sample mean. Accordingly, the sample s.d. should be reported along with the sample mean when the goal is to describe the distribution of a variable of interest. The s.e.m. is not particularly helpful in summarising the distribution of the data and can be misleading if presented in place of the s.d., since the s.e.m. will always be smaller than the s.d. The s.e.m. does play an important role in making inferences about the population mean, as we will see below.

Confidence intervals do double duty

Suppose our goal is to draw inference about the mean striatal glutamate in the population of converters. We know that the sample mean provides a single-value estimate of the parameter of interest, i.e. the population mean striatal glutamate value among converters. We can also construct a confidence interval (CI), which provides a range of plausible values for the population mean.

The formula for the confidence interval for a mean is given in Box 2. The 95% CI based on our sample of 7 converters is approximately $30.25 \pm 2.45 \times 1.73 = [26.01–34.48]$. On the basis of this interval, we can make the following inferential statement: we are 95% confident that the mean striatal glutamate in the population of converters could be a value as low as 26.01 IU to a value as high as 34.48 IU. The width of the interval clearly depends on two quantities: the value of C and the s.e.m.. Holding C fixed (i.e. fixing the confidence level) we can see that a small s.e.m. yields a narrow interval whereas a large s.e.m. yields a wide one. Since a smaller s.e.m. denotes higher precision, a narrower interval denotes a more precise range of values for the population mean.

There is an important and often underappreciated connection between confidence intervals and hypothesis tests. The connection is as follows: for a two-sided hypothesis test, conducted at a significance level of α, the rejection region for the test is made up of all values outside of the $(100 - \alpha)$% CI. For example, suppose that the null hypothesis states that the population mean striatal glutamate

for converters is 25 IU. Conducting a two-sided test at the $\alpha = 5\%$ significance level is equivalent to checking whether the null value of 25 IU is outside of the 95% CI. Since 25 is outside of the interval, then we necessarily reject the null hypothesis. In other words, 25 is not a plausible value for the population mean striatal glutamate since the 95% CI does not include it – so the hypothesis claiming that 25 is the true value should be rejected.

Up to this point we have focused on the standard error and confidence intervals for a single mean. Fortunately, these concepts can be extended to conduct inference for other population quantities (such as differences in means or proportions, correlation coefficients, regression coefficients and odds ratios).

For example, suppose that we want to know whether the mean striatal glutamate in the population of converters is different from that in the population of non-converters. The 95% CI for the difference in the two means is [0.26–8.51] (see the Supplementary appendix available at https://doi.org/10.1192/bja.2020.41 for how this is computed). Since the value of zero is not in the interval, we know that we would have rejected a null hypothesis stating that the true difference is 0, at the 5% significance level. But we learn even more from the confidence interval, namely that, even though the null value of 0 is not contained in the interval, the difference could be as small as 0.25 IU, which is very close to 0. The wide range of the confidence interval points to relatively low precision in the estimate of the difference between the sample means from the converter and non-converter groups. Reporting the range of plausible values, rather than just the *P*-value from the hypothesis test alone, allows the reader to assess both the clinical relevance of the results and their statistical significance. Accordingly, it makes sense to either default

to reporting the confidence interval in lieu of the hypothesis test results or in addition to them whenever possible.

Conclusions

We have briefly discussed the standard deviation (s.d.) as a measure of variability in individual values and contrasted that with the standard error (s.e.), which is a measure of variability of a statistic. We specifically focused on the standard error of the mean (s.e.m.) and viewed it as a measure of precision of the sample mean. We also saw how confidence intervals incorporate this precision measure to serve the dual role of aiding in the assessment of both clinical and statistical significance.

Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bja.2020.41.

Declaration of interest

None.

An ICMJE form is in the supplementary material, available online at https://doi.org/10.1192/bja.2020.41.

References

Cecil K (2013) Proton magnetic resonance spectroscopy: technique for the neuroradiologist. *Neuroimaging Clinics of North America*, **23**: 381–92.

Kegeles LS, Ciarleglio A, León-Ortiz P, et al (2019) An imaging-based risk calculator for prediction of conversion to psychosis in clinical high-risk individuals using glutamate ^{1}H MRS. *Schizophrenia Research* [Epub ahead of print] 12 Sep. Available from: https://doi.org/10.1016/j.schres.2019.09.004.