Explorations in statistics: standard deviations and standard errors

Douglas Curran-Everett
Division of Biostatistics and Bioinformatics, National Jewish Health, and Department of Preventive Medicine and Biometrics and Department of Physiology and Biophysics, School of Medicine, University of Colorado Denver, Denver, Colorado

Submitted 21 March 2008; accepted in final form 6 May 2008

In 2004, Dale Benos and I proposed concise guidelines for reporting statistics (12). We included a brief explanation or example to clarify each guideline, and we provided additional resources readers could use in concert with the guidelines. Nevertheless, we recognized that the guidelines—even with explanations and examples—were inadequate for readers who wanted to learn about fundamental concepts in statistics. Why are fundamental concepts in statistics important? They form the cornerstone of scientific inquiry. If we fail to understand the essential distinction between standard deviation and standard error: a standard deviation estimates the variability among sample observations whereas a standard error of the mean provides an estimate of the uncertainty of the true value of the mean. If we fail to report the standard deviation, then we fail to fully report our data. Because it incorporates information about sample size, the standard error of the mean is a misguided estimate of variability among observations. Instead, the standard error of the mean provides an estimate of the uncertainty of the true value of the population mean.

R: Software to Explore Concepts

In my statistics course, I use the freeware package R (34). R is a system—a language and an environment—for statistical analysis and data graphics. The R environment is a command-line environment in which > represents the command line. You can submit an R command in two ways: type the command in the interactive R Console, or submit the command from a script.2 Because this series relies on R merely to explore fundamental concepts in statistics, I provide a script of the requisite R commands.

Regardless of whether you use a Mac or a PC, there are three preliminary steps to perform:
1. On your Desktop, create a folder called Advances.
2. Download the script Advances_Statistics_Code.R to your Advances folder.3
3. Download and install R.

Installation and basic operations. If you use a Mac, download R from http://cran.us.r-project.org/bin/macosx/.

After you have installed R, double-click on Advances_Statistics_Code.R to open it. To submit particular commands in Advances_Statistics_Code.R, highlight the commands you want to submit and then press ⌘ (command key+enter). In the R Console, default colors are blue for command and black for result. Data graphics are shown in the Quartz device window. If you use a PC, download R from http://cran.us.r-project.org/bin/windows/base/.

After you have installed R, a shortcut for R will exist on your Desktop. To simplify the process of starting R from within your Advances folder, move this shortcut into your Advances folder.3 This file is available through the Supplemental Material link for this article on the Advances in Physiology Education website.

Address for reprint requests and other correspondence: D. Curran-Everett, Div. of Biostatistics and Bioinformatics, M222, National Jewish Health, 1400 Jackson St., Denver, CO 80206 (e-mail: EverettD@njj.org).

1 These guidelines can be accessed through the American Physiological Society "Information for Authors" (9).

2 A script is a file composed of R commands.

3 This file is available through the Supplemental Material link for this article on the Advances in Physiology Education website.

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
folder, right-click on the shortcut, and then click on Properties. Paste the full address (path) of your Advances folder

C:\Documents and Settings\...\Desktop\Advances

into the Start in: location (Fig. 1) and then click OK. Now double-click on the R shortcut to open R. To open Advances_Statistics_Code.R, click File/Open script...4 or click the Open script icon \(\text{[2]}\), select the script filename, and then click Open. Advances_Statistics_Code.R will open in the R Editor.

To submit particular commands in Advances_Statistics_Code.R, highlight the commands you want to submit, right-click, and then click Run line or selection. Or, if you highlight the commands you want to submit, you can simply type Ctrl+R. In the R Console, default colors are red for command and blue for result. Data graphics are shown in the R Graphics device window.

Basic syntax. The script Advances_Statistics_Code.R contains comments in addition to commands. Comments define sections of the script and explain many of the commands. The character # denotes a comment: all text after the first # on a line is a comment. For example, lines 5–8 of Advances_Statistics_Code.R are

\[
\# \quad \text{--Define population parameters and sample numbers------}
\]

\[
\#
\]

\[
\text{PopMean} <- 0 \quad \text{# population mean}
\]

\[
\text{PopSD} \quad <- 1 \quad \text{# population standard deviation}
\]

The two commands in these lines of code, PopMean \(<- 0\) and PopSD \(<- 1\), assign values to the variables PopMean and PopSD, the population mean and standard deviation.

If you highlight and then submit these lines of code, this is what you see in the R Console:

\[
> \# \quad \text{--Define population parameters and sample numbers------}
> \#
> \text{PopMean} \quad <- 0 \quad \text{# population mean}
> \text{PopSD} \quad <- 1 \quad \text{# population standard deviation}
> \\
\]

It is not obvious the commands have done a thing. If you type and then submit each variable name5 in the R Console, however, you see that the commands have assigned the values of 0 and 1 to PopMean and PopSD:

\[
> \text{PopMean}
\]

\[
\text{[1]} \quad 0
\]

\[
> \text{PopSD}
\]

\[
\text{[1]} \quad 1
\]

The Simulation: Observations and Sample Statistics

If we want to explore the distinction between standard deviation and standard error, we need some data. When I teach a class on regression, I introduce a data set in this way:

\[
\sigma
\]

Fig. 2. The population. We can generate a standard normal distribution by transforming some random variable \(Y\) to \(z\) by the relationship \(z = (Y - \mu)/\sigma\), where \(z\) represents the number of standard deviations \(Y\) is from the mean \(\mu\.)

<table>
<thead>
<tr>
<th>Column</th>
<th>Heading</th>
<th>Sample Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sample</td>
<td>Sample number</td>
</tr>
<tr>
<td>2</td>
<td>Mean</td>
<td>Mean (\bar{y})</td>
</tr>
<tr>
<td>3</td>
<td>SD</td>
<td>Standard deviation (s)</td>
</tr>
<tr>
<td>4</td>
<td>SE</td>
<td>Standard error of the mean (SE { \bar{y} } = s/\sqrt{n})</td>
</tr>
<tr>
<td>5</td>
<td>(t)</td>
<td>Observed value of (t = \bar{y}/SE { \bar{y} })</td>
</tr>
<tr>
<td>6</td>
<td>LCI</td>
<td>Lower confidence interval bound</td>
</tr>
<tr>
<td>7</td>
<td>UCI</td>
<td>Upper confidence interval bound</td>
</tr>
</tbody>
</table>

Table 1. Sample statistics calculated for each random sample

\[
\begin{align*}
\text{PopSD} \quad & <- 1
\end{align*}
\]

To do this, click within the R Console, type the variable name, and then press Enter.

---

4 The notation \textit{click A then B} means \textit{click A, then click B}.

5 To do this, click within the R Console, type the variable name, and then press Enter.
It is difficult to choose an example that is relevant to everyone. So instead, I want to use an example that is relevant to no one: cement.

I then proceed to discuss a 1932 study that examined the impact of the composition of cement on the heat released by the cement as it hardened (15).

Suppose the random variable $Y$ represents not the heat from cement but the physiological thing you study: L-ascorbic acid transport, differential gene expression, TNF-$\alpha/H_9267/H_11005$ or venous capacitance in trout. Assume that your $Y$ is distributed normally with mean $\mu/H_9262/H_11005$ and standard deviation $\sigma/H_9268/H_11005$. We now have an example that is relevant to everyone. Unfortunately, we now also have a problem: different responses have different means and standard deviations. We can circumvent this problem if we consider the distribution of each response to be a standard normal distribution with mean $\mu/H_9262/H_11005$ and standard deviation $\sigma/H_9268/H_11005$ (Fig. 2). This standard normal distribution is the population from which we will obtain our simulated sample observations – our data.

As the statistical cornerstone for our explorations, suppose we want to estimate $\mu/H_9262/H_11005$ and $\sigma/H_9268/H_11005$, the mean and standard deviation of our population (see Ref. 14). To do this, we draw at random a sample of $n$ observations from the population. For simplicity, suppose we limit the sample to nine observations. This is the R command (Advances_Statistics_Code.R, line 36) that generates the sample and rounds each value to three decimal places:

```r
TheData <- round(rnorm(nObs, mean = PopMean, sd = PopSD), 3)
```

The sample size is defined by the command nObs <- 9 (Advances_Statistics_Code.R, line 10).

Because we had so much fun taking 1 random sample, we repeat the process until we have drawn a total of 1000 random samples, each with 9 observations, from our population. Mercifully, the command for (i in 1:nSamples) in line 35 of

Advances_Statistics_Code.R does this for us. These are the observations—the data—for samples 1, 2, and 1000:

```
> ¥ Sample Observations
[1] 0.422 1.103 1.006 1.034 0.285 -0.647 1.235 0.912 1.825
[2] 0.154 -0.654 -0.147 1.715 0.720 0.804 0.256 1.155 0.646
[1000] 0.560 -1.138 0.485 -0.864 -0.277 2.198 0.050 0.500 0.587
```

Your sample observations will differ.
We have our data, but if we want to really understand the distinction between standard deviation and standard error, we also need some sample statistics. So each time we draw a sample of 9 observations, we calculate the sample statistics listed in Table 1. These are the statistics for samples 1, 2, and 1000:

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>t</th>
<th>LCI</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.797</td>
<td>0.702</td>
<td>0.234</td>
<td>3.407</td>
<td>0.362</td>
<td>1.232</td>
</tr>
<tr>
<td>2</td>
<td>0.517</td>
<td>0.707</td>
<td>0.236</td>
<td>2.193</td>
<td>0.079</td>
<td>0.955</td>
</tr>
<tr>
<td>1000</td>
<td>0.333</td>
<td>0.975</td>
<td>0.325</td>
<td>0.718</td>
<td>−0.371</td>
<td>0.838</td>
</tr>
</tbody>
</table>


With these 1000 sets of sample observations and statistics, we are ready to explore the essential distinction between standard deviation and standard error.

### Standard Deviation

In each of our 1000 samples, the 9 observations differ because the underlying population (see Fig. 2) is distributed over a range of possible values. The typical measure of the variability among experimental measurements is the sample standard deviation $s$:

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}},$$

where $n$ is the number of observations in the sample, $y_i$ is an individual observation, and $\bar{y}$ is the sample mean. The sample standard deviation characterizes the dispersion of observations about the sample mean and estimates the population standard deviation $\sigma$. For example, the standard deviation of the observations in sample 1, $0.422, 1.103, \ldots, 1.825$, is $s = 0.702$, which estimates $\sigma = 1$. The empirical distribution of the 1000 sample standard deviations is centered at 0.966, slightly less than the actual value of 1 (Fig. 3). The command in line 104 of Advances_Statistics_Code.R returns this value. Your value will differ slightly.

A larger standard deviation means greater dispersion: more variability (Fig. 4).

### Standard Error of the Mean

In words, what is the standard error of the mean SE $\{\bar{y}\}$? I ask this question of my students on the first day of class. Often students can explain in words how to calculate the standard error: divide the standard deviation by the square root of the sample size. Seldom can a student explain the concept behind the standard error: if I repeat an experiment a whole bunch of times—and each time I calculate a sample mean—the standard deviation of those sample means will be the standard error of the mean (12–14). The standard error of the mean answers a theoretical question: if I repeat an experiment a whole bunch of times, by how much will a typical sample mean differ from the population mean?

By virtue of our simulation, we have 1000 sample means (Fig. 5). If we treat these 1000 sample means as observations, we can calculate their average and standard deviation:

$$\text{Ave} \{\bar{y}\} = \mu \quad \text{and} \quad \text{SD} \{\bar{y}\} = \sigma/\sqrt{n}.\quad (8)$$

If the sample size $n$ increases, then the standard deviation of the theoretical distribution of the sample mean will decrease: the more sample observations we have, the more certain we will be that the sample mean $\bar{y}$ is near the actual population mean $\mu$. (Fig. 6).

---

6 A statistic is a quantity calculated from the sample observations.

7 We will use the statistics in columns 4–7 in subsequent explorations.

8 I derive these results for the mean and standard deviation in Ref. 14. The Central Limit Theorem states that the theoretical distribution of the sample mean will be approximately normal regardless of the distribution of the original observations. If the distribution of the original observations happens to be normal, then the theoretical distribution of the sample mean will be exactly normal.
The standard deviation of the distribution of the sample mean is the standard error of the sample mean SE \( \{ \bar{y} \} \).

**Summary**

As this exploration has demonstrated, the standard deviation and standard error of the mean estimate quite different things: a standard deviation estimates the variability among individual observations in a sample—it can also estimate the variability in the underlying population—but a standard error of the mean provides an estimate of the uncertainty of the true value of the population mean.

In a sample, the observations—the data—differ because the population from which they were drawn is distributed over a range of possible values. The standard deviation describes the dispersion of these sample observations about the sample mean. If we fail to report the standard deviation, then we fail to fully report our data. Because it incorporates information about sample size, the standard error of the mean is a misguided estimate of variability among observations. Instead, the standard error of the mean provides an estimate of the uncertainty of the true value of the population mean.

In the next installment of this series, we will explore some of the concepts behind hypothesis testing: test statistics and \( P \) values.

**APPENDIX**

The probability density function—the theoretical distribution of possible values—of the sample standard deviation \( s \) is

\[
f(s) = 2 \left( \frac{n^*}{\sigma} \right) s^{n^* - 2} e^{-s^2/\sigma^2} \frac{\Gamma(n^*)}{\Gamma(n^* + 1/2)} e^{-s^2/\sigma^2} , \quad \text{for} \quad 0 < s < \infty , \quad (A1)
\]

where \( n^* = (n - 1)/2 \) and the gamma-function \( \Gamma(n^*) \) is

\[
\Gamma(n^*) = \int_0^\infty w^{n^* - 1/2} e^{-w} \, dw .
\]

If \( n^* \) is a positive integer greater than 1, then \( \Gamma(n^*) = (n^* - 1)! \). For example, \( \Gamma(4) = 3! \cdot 2 = 6 \).

Figure 7 depicts the probability density function of the sample standard deviation for 5, 10, 20, 30, 40, 50, and 100 observations.

**REFERENCES**