Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel

Douglas Curran-Everett¹,²,³ and Dale J. Benos⁴

¹Division of Biostatistics, National Jewish Medical and Research Center, and Departments of ²Preventive Medicine and Biometrics and ³Physiology and Biophysics, School of Medicine, University of Colorado at Denver and Health Sciences Center, Denver, Colorado; and ⁴Department of Physiology and Biophysics, University of Alabama, Birmingham, Alabama

WE SCIENTISTS rely on statistics. In part, this is because we use statistics to report our own science and to interpret the published science of others. For some of us, reporting and interpreting statistics can be like reading an unfamiliar language: it is awkward to do, and it is easy to misinterpret meaning. To facilitate these tasks, in 2004, we wrote an editorial (8) in which we proposed specific guidelines to help investigators analyze data and communicate results. These guidelines for reporting statistics can be accessed through the American Physiological Society (APS) Information for Authors (4).

In this followup editorial, we report initial reactions to the guidelines and the subsequent impact of the guidelines on reporting practices. We also revisit three guidelines. In 2004, we hoped the guidelines would improve and standardize the caliber of statistical information reported throughout journals published by the APS. We still do.

Initial Reactions

Initial reactions to the guidelines were mixed. What we heard, however, was quite positive.

This is really very helpful indeed. I wish all journals would adopt this as standard.

Physiologist¹

We were delighted that many of the people who congratulated—even thanked—us were statisticians.

I have just read the guidelines published in Physiological Genomics and wish to congratulate you on a nice job!

[About] 6 years ago I reviewed a paper for an APS journal . . . I offered several trivial biological suggestions and then asked [the authors] to report the [standard deviation] rather than [standard error] to represent the variability about their sample mean. The authors adopted all of the biological suggestions but rejected the one statistical critique stating that it was standard policy to report the [standard error] and their colleagues all expected it.

I am quite certain that these guidelines will prompt much grousing from the biologists about being too theoretical and unnecessary, as well as from the statisticians that you left out the rules nearest and dearest to their hearts.

Statistician

If statisticians groused about the guidelines, they never groused to us.

Good guidelines . . . Quite reasonable without being overly fussy. The interpretation of P values is much more sensible than one often gets with medical journals, where $P < 0.05$ is all they care about.

Statistician

Every now and then, a biologist did grouch:

I do not agree with the edict about presenting data as [standard deviations] rather than [standard errors of the mean]. These presentations are for visual effect only . . . To me, this edict is silly, particularly since showing [standard deviations rather than standard errors of the mean] is a cosmetic issue only.

Physiologist

One biologist was moved to write a Letter to the Editor (15). For the most part, when someone did complain about the guidelines, it was


about which they complained.

Subsequent Impact

Within a year, the Editor of Clinical and Experimental Pharmacology and Physiology had solicited a critique of the guidelines (18). The critique set the stage for the journal to revise its own guidelines for reporting statistics (J. Ludbrook, personal communications).²

In May 2006, we received an email from Australia that hinted the guidelines had little immediate impact on reporting practices:

Many of us were pleased to see this article come out, particularly in a major, trend-setting publication. Early optimism faded as authors and editors did not appear to want to reinforce your message . . .

Most of my colleagues [report standard errors and $P < 0.05$] and, when questioned, tell me that they are using the appropriate method: that is what I will find in the journals, that is what their predecessors used, they know what it means (so why don’t I?), and in any case [the result is obvious].

Physiologist

To estimate the actual impact of the guidelines on reporting practices, we sampled original articles published from August

¹ Because these comments reflect unsolicited personal correspondence, we have elected to withhold the names.

2003 through July 2004, the year before the guidelines, and we reviewed all original articles published from August 2005 through July 2006, the second year after the guidelines. If the guidelines affected the reporting of statistics, we expected the incidence of standard errors to decrease and the incidence of standard deviations, confidence intervals, and precise P values to increase.

What did our literature review reveal? That the guidelines had virtually no impact on the occurrence of standard errors, standard deviations, confidence intervals, and precise P values (Table 1). There were two exceptions: in one journal, the use of standard errors decreased from 88% to 81%; in another journal, the use of precise P values increased from 4% to 17%.

### Table 1. American Physiological Society journal manuscripts in 1996, 2003, and 2006: reporting of statistics

<table>
<thead>
<tr>
<th>Journal</th>
<th>n</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
<th>Confidence Interval</th>
<th>Precise P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am J Physiol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Physiol</td>
<td>43</td>
<td>30</td>
<td>21</td>
<td>88</td>
<td>0</td>
</tr>
<tr>
<td>Endocrinol Metab</td>
<td>28</td>
<td>30</td>
<td>18</td>
<td>86</td>
<td>0</td>
</tr>
<tr>
<td>Gastrointest Liver Physiol</td>
<td>26</td>
<td>28</td>
<td>8</td>
<td>92</td>
<td>0</td>
</tr>
<tr>
<td>Heart Circ Physiol</td>
<td>60</td>
<td>62</td>
<td>17</td>
<td>87</td>
<td>0</td>
</tr>
<tr>
<td>Lung Cell Mol Physiol</td>
<td>25</td>
<td>26</td>
<td>20</td>
<td>84</td>
<td>0</td>
</tr>
<tr>
<td>Renal Physiol</td>
<td>27</td>
<td>25</td>
<td>15</td>
<td>93</td>
<td>0</td>
</tr>
<tr>
<td>J Appl Physiol</td>
<td>62</td>
<td>57</td>
<td>24</td>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>J Neurophysiol</td>
<td>58</td>
<td>61</td>
<td>36</td>
<td>69</td>
<td>2</td>
</tr>
</tbody>
</table>

Values are percentages of journals that reported numbers of manuscripts reviewed (n), standard deviation, standard error of the mean, and precise P value; for example, P = 0.03 (rather than P < 0.05) or P = 0.12 (rather than P > 0.05). In 1996, these journals published a total of 3693 original articles; the number of articles reviewed represents a 10% sample (systematic random sampling, fixed start) of the original articles published by each journal (9). From August 2003 through July 2004, these journals published a total of roughly 3500 original articles; the number of articles reviewed represents a 10% sample (systematic random sampling, fixed start) of the original articles published by each journal. From August 2005 through July 2006, these journals published a total of 3675 original articles; the number of articles reviewed represents a complete survey of the original articles published by each journal. *Italicized pairs indicate values that differ statistically from 2003 to 2006 (0.05 ≤ P ≤ 0.10, exact binomial probability, 1-tailed) and from 1996 to 2003 (0.001 ≤ P ≤ 0.05, Fisher’s exact test, 2-tailed).

### The 2004 Guidelines Revisited

These guidelines addressed the reporting of statistics in the RESULTS section of a manuscript:


Our 2004 editorial (8) summarized the theoretic rationale for each of these guidelines. The subsequent publication of Scientific Style and Format (6) reinforced their application. Because

---

Fig. 1. The distinction between standard deviation and standard error of the mean. Suppose random variable \( Y \) is distributed normally with mean \( \mu = 0 \) and standard deviation \( \sigma = 1 \) (inset). If you draw from this population random samples with sizes that range from 5 to 10,000 observations, for each sample you can estimate the mean, the standard deviation, and the standard error of the mean (see Eq. 1).

As sample size increases, the sample standard deviation fluctuates more tightly about the population standard deviation \( \sigma \). If sample size is infinite—–if the entire population is sampled—–then the sample standard deviation \( \bar{y} \) will equal the population standard deviation \( \sigma \). As sample size increases, the standard error of the mean decreases progressively because of its dependence on sample size. If sample size is infinite, then the standard error of the mean will be 0, and the sample mean \( \bar{y} \) will equal the population mean \( \mu \); there is no uncertainty about the true value of the population mean.
they offer the greatest benefit to authors and readers of RESULTS sections, we revisit each of these guidelines.

Guideline 5. Report variability using a standard deviation. The distinction between standard deviation and standard error of the mean is far more than cosmetic: it is an essential one. These statistics estimate different things: a standard deviation estimates the variability among individual observations in a sample, but a standard error of the mean estimates the theoretical variability among sample means (8, 9).

Individual observations in a sample differ because the population from which they were drawn is distributed over a range of possible values. The study of this intrinsic variability is important: it may reveal something novel about underlying scientific processes (12). The standard deviation describes the variability among the observations we investigators measure; it characterizes the dispersion of sample observations about the sample mean.

In contrast, the standard error of the mean provides an answer to a theoretical question: If I repeat my experiment an infinite number of times, by how much will the possible sample means vary about the population mean?

The fundamental difference between standard deviation and standard error of the mean is reflected further in how these statistics are defined. The standard deviation $s$ and the standard error of the mean $\text{SE}\{\bar{y}\}$ are

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

where $n$ is the number of observations in the sample, $y_i$ is an individual observation, and $\bar{y}$ is the sample mean. Because it incorporates information about sample size, the standard error of the mean is a misguided estimate of variability among observations (Fig. 1 and Ref. 8). By itself, the standard error of the mean has no particular value (9). Even with a large sample size ($n = 35$), the interval

$$ [\bar{y} - \text{SE}\{\bar{y}\}, \bar{y} + \text{SE}\{\bar{y}\}] \quad (2) $$

is just a 68% confidence interval. In other words, we can declare, with modest 68% confidence, that the population mean is included in the interval $[\bar{y} - \text{SE}\{\bar{y}\}, \bar{y} + \text{SE}\{\bar{y}\}]$.

To summarize, in most experiments, it is essential that an investigator reports on the variability among the actual individual measurements. A sample standard deviation does this: it describes the variability among the actual experimental measurements. On the other hand, if an investigator were to repeat an experiment many times, and each time calculate a sample mean, the average of those sample means will be the population mean; the standard deviation of those sample means will be the standard error of the mean (9). A standard error is simply the standard deviation of a statistic: here, the sample mean. In nearly all experiments, however, a single sample mean is computed. Therefore, it is inappropriate to report a standard error of the mean—a theoretical estimate of the variability of possible values of a sample mean about a population mean—as an estimate of the variability among actual experimental measurements.

Guideline 6. Report uncertainty about scientific importance using a confidence interval. A confidence interval focuses attention on the magnitude and uncertainty of an experimental result. In essence, a confidence interval helps answer the question, is the experimental effect big enough to be relevant? A confidence interval is a strong tool for inference: it provides the same statistical information as the $P$ value from a hypothesis test, it circumvents the drawbacks inherent to a hypothesis test, and it provides information about scientific importance (9).

In biomedical research, the routine use of confidence intervals is recommended (1–3, 8, 9, 13), and, in clinical medicine, the use of confidence intervals is indeed widespread (2, 3). In journals published by APS, however, the incidence of confidence intervals was as rare in 2006 as it was in 1996 (see Table 1). Guideline 6. Report a precise $P$ value. In 2004 we wrote that a precise $P$ value communicates more information with the same amount of ink, and it permits readers to assess a statistical result according to their own criteria (8). You would think most authors would report a precise $P$ value. Most do not (see Table 1).

On occasion, an author who reported a precise $P$ value did so with unnecessary precision:

$$ P = 0.030711 \quad \text{or} \quad P = 10^{-31}. $$

The guidelines for rounding $P$ values to sufficient precision are listed in Table 2.

Table 2. Guidelines for rounding $P$ values to sufficient precision

<table>
<thead>
<tr>
<th>$P$ Value Range</th>
<th>Rounding Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.01 \leq P \leq 1.00$</td>
<td>Round to 2 decimal places: round $P = 0.031$ to $P = 0.03$</td>
</tr>
<tr>
<td>$0.001 \leq P \leq 0.009$</td>
<td>Round to 3 decimal places: round $P = 0.0066$ to $P = 0.007$</td>
</tr>
<tr>
<td>$P &lt; 0.001$</td>
<td>Report as $P &lt; 0.001$; more precision is unnecessary</td>
</tr>
</tbody>
</table>

Guideline 7. Report a precise $P$ value. In 2004 we wrote that a precise $P$ value communicates more information with the same amount of ink, and it permits readers to assess a statistical result according to their own criteria (8). You would think most authors would report a precise $P$ value. Most do not (see Table 1).

On occasion, an author who reported a precise $P$ value did so with unnecessary precision:

$$ P = 0.030711 \quad \text{or} \quad P = 10^{-31}. $$

The guidelines for rounding $P$ values to sufficient precision are listed in Table 2.

The Next Step

All of us–authors, reviewers, editors–gravitate toward the familiar. For most of us, reporting standard deviations, confidence intervals, and precise $P$ values is quite unfamiliar. To make matters worse, these reporting practices have been unfamiliar for decades. There is considerable inertia to overcome before standard deviations, confidence intervals, and precise $P$ values become commonplace in journals published by the APS.

Reform is difficult (5, 7, 10, 11, 14, 16). The question is, how can we help it along? In part, the answer is that all of us–authors, reviewers, editors, Publications Committee–must make a concerted effort to use and report statistics in ways that are consistent with best practices. With these guidelines (8), we summarized the best practices of statistics.

In 2004, we hoped that the guidelines would improve and standardize the caliber of statistical information reported throughout journals published by the APS. It is clear that the

4 This guideline is mirrored in Scientific Style and Format Section 12.5.2.2 of Ref. 6.
5 This guideline is mirrored in Scientific Style and Format Section 12.5.1.2 of Ref. 6.
mere publication of the guidelines failed to impact reporting practices. We still have an opportunity.

ACKNOWLEDGMENTS

We thank Margaret Reich (Director of Publications and Executive Editor, American Physiological Society) for a 1-yr print subscription to APS journals, and we thank Matthew Strand (National Jewish Medical and Research Center, Denver, CO) and the Editors of the American journals for comments and suggestions.

REFERENCES