Last Word on Perspectives “Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel”

Douglas Curran-Everett\textsuperscript{1,2,3} and Dale J. Benos\textsuperscript{4}

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

WE APPRECIATE ALL COMMENTS, past (13, 15) and present (2, 14, 16, 17), on our guidelines (6) for reporting statistics in journals published by the American Physiological Society (APS). In 2004 we wrote that the guidelines embodied fundamental concepts in statistics (8) and that they were consistent with the Uniform Requirements (12), used by roughly 650 biomedical journals, and with Scientific Style and Format (4), used by APS publications. In 2007 we will go farther: the guidelines reflect mainstream statistical concepts and accepted statistical practices.

Although we published guidelines—not dictums—for reporting statistics, Clayton (2) and Rangachari (17) described our wording of the guidelines as dictatorial and prescriptive. These reactions raise a question not about statistics but about composition: if you want to write an effective guideline, just how do you do it? We followed three principles of composition from The Elements of Style (19):

15. Put statements in positive form.
16. Use definite, specific, concrete language.
17. Omit needless words.

Because it was impossible for each guideline, so written, to fully explain itself, we included a brief explanation or example to elucidate subtleties the guideline itself could not possibly fully explain itself, we included a brief explanation or example. Because it was impossible for each guideline, so written, to fully explain itself, we included a brief explanation or example to elucidate subtleties the guideline itself could not possibly accommodate.\textsuperscript{1} In addition, we provided other resources (1, 3, 5, 8–10, 12, 18) interested readers could use in concert with the guidelines.

Specific Guidelines

In this section, we address comments made about specific guidelines.

Guideline 2. Define and justify a critical significance level \(\alpha\) appropriate to the goals of your study. This guideline reinforces the notion that a statistical benchmark of 0.05, that is, \(\alpha = 0.05\), is not always the optimum choice. As Clayton (2) points out, there are different statistical philosophies about the interpretation and use of the critical significance level \(\alpha\) and the observed significance level \(P\). The explanation of this guideline reflects not a specific philosophy but the reality of use: in science, the statistical philosophies about \(\alpha\) and \(P\) have been melded together.

Guideline 4. Control for multiple comparisons. Clayton (2) is concerned that we mentioned the Newman-Keuls procedure in a footnote to the exposition of this guideline (6). In that footnote, we listed also the Bonferroni and least significant difference procedures as examples of common multiple comparison procedures. By the mere mention of these procedures, we did not mean to endorse them: a review (5) cited in the same footnote illustrates that each of these three procedures is of limited practical value.

Guideline 5. Report variability using a standard deviation. This guideline reinforces the essential difference between a standard deviation and a standard error (see Refs. 6–8). Clayton (2) expresses concern about the logic of this guideline but then promptly reinforces it. Variability among sample observations is a basic scientific and statistical characteristic. Sir Ronald Fisher (11) argued its value:

[Populations] always display variation ... The variation itself was not an object of study, but was recognised rather as a troublesome circumstance which detracted from the value of the average ... [The] study of the causes of variation of any variable phenomenon ... should be begun by the examination and measurement of the variation which presents itself.

Guideline 6. Report uncertainty about scientific importance using a confidence interval and Guideline 10. Interpret each main result by assessing the numerical bounds of the confidence interval and by considering the precise \(P\) value. Clayton (2) argues these guidelines harbor errors of interpretation and give poor advice. We disagree.

When we wrote this guideline, we synthesized a lot of information into the phrase scientific importance, and we packed a lot of information into the explanation.

If either bound of the confidence interval is important from a scientific perspective, then the experimental effect may be large enough to be relevant.

The examples below illustrate interpretations and advantages of confidence intervals.

Suppose you study the impact of new drugs on systemic hypertension. You find three studies in the American Journal of Physiology-Heart and Circulatory Physiology that investigated independently the effect of three different drugs on systemic hypertension. Each study involved a sample of 25 subjects, and—because they wanted to be especially confident of their findings—each reported a 99% confidence interval for the mean change in systemic blood pressure. For each drug, these are
the sample mean $\bar{y}$, sample standard deviation $s$, $P$ value, and 99% confidence interval:

<table>
<thead>
<tr>
<th></th>
<th>Sample Mean $\bar{y}$</th>
<th>Sample Standard Deviation $s$</th>
<th>$P$ Value</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>-20</td>
<td>18</td>
<td>&lt;0.001</td>
<td>-30 to -10</td>
</tr>
<tr>
<td>Drug B</td>
<td>-0.2</td>
<td>18</td>
<td>&lt;0.001</td>
<td>-0.3 to -0.1</td>
</tr>
<tr>
<td>Drug C</td>
<td>-20</td>
<td>18</td>
<td>0.07</td>
<td>-50 to +10</td>
</tr>
</tbody>
</table>

How do you interpret these results?

**Drug A** decreased blood pressure by 20 mmHg, a change that differed convincingly from 0 ($P < 0.001$). The confidence interval suggests the true mean impact of drug A is likely to be between a 10- and 30-mmHg decrease in blood pressure, a change that is scientifically meaningful and reasonably precise. **Drug A** produced a convincing change of scientific importance.

**Drug B** decreased blood pressure by 0.2 mmHg, a change that also differed convincingly from 0 ($P < 0.001$). The confidence interval suggests the true mean impact of drug B is likely to be between a 0.1- and 0.3-mmHg decrease in blood pressure, a change that is scientifically trivial but quite precise. **Drug B** produced a convincing change of no scientific importance.

**Drug C** decreased blood pressure by 20 mmHg, a change consistent with 0 ($P = 0.07$). The confidence interval suggests the true impact of drug C could range from a 10-mmHg increase to a 50-mmHg decrease in blood pressure. If the true impact of drug C is a 10-mmHg increase in blood pressure, then drug C is not a viable drug with which to decrease blood pressure. In contrast, if the true impact of drug C is a 50-mmHg decrease in blood pressure, then drug C does decrease blood pressure. Because it is relatively long, the confidence interval for drug C is an imprecise estimate of the true impact of drug C on blood pressure. **Drug C** bears further study using a larger sample size.

Note that the scientific importance of the upper and lower bounds of a confidence interval depends on scientific context. Although a standard error characterizes uncertainty about the true value of some population characteristic (for example, the mean), a confidence interval is a more useful estimate (6–8, 14).

**Summary**

Clayton (2) reminds us that statistics, like science, continues to evolve. Of course it does. But fundamental concepts of statistics—concepts of statistical significance, scientific importance, variability, uncertainty, multiple testing—remain unchanged. The guidelines we published in 2004 (6) embody those fundamental concepts. We continue to believe the guidelines offer a concise, accurate framework that we hope will help improve the caliber of statistical information reported in articles published by the American Physiological Society.

**REFERENCES**