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

Douglas Curran-Everett and Dale J. Benos

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

Concepts and procedures in statistics are inherent to publications in science. Based on the incidence of standard deviations, standard errors, and confidence intervals in articles published by the American Physiological Society (APS), however, many scientists appear to misunderstand fundamental concepts in statistics (9). In addition, statisticians have documented that statistical errors are common in the scientific literature: roughly 50% of published articles have at least one error (1, 2). This misunderstanding and misuse of statistics jeopardizes the process of scientific discovery and the accumulation of scientific knowledge.

In an effort to improve the caliber of statistical information in articles they publish, most journals have policies that govern the reporting of statistical procedures and results. These were the previous guidelines for reporting statistics in the *Information for Authors* (3) provided by the APS: *1*) In the MATERIALS AND METHODS, authors were told to "describe the statistical methods that were used to evaluate the data." *2*) In the RESULTS, authors were told to "provide the experimental data and results as well as the particular statistical significance of the data." *3*) In the DISCUSSION, authors were told to "Explain your interpretation of the data. . . ." To an author unknowing about statistics, these guidelines gave almost no help.

In its 1988 revision of *Uniform Requirements* (see Ref. 13, p. 260), the International Committee of Medical Journal Editors issued these guidelines for reporting statistics:

> Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of *P* values, which fails to convey important quantitative information. . . . Give numbers of observations. . . . References for study design and statistical methods should be to standard works (with pages stated) when possible rather than to papers where designs or methods were originally reported. Specify any general-use computer programs used.

The current guidelines issued by the Committee (see Ref. 14, p. 39) are essentially identical. To an author unknowing about statistics, these *Uniform Requirements* guidelines give only slightly more help.

In this editorial, we present specific guidelines for reporting statistics.[1] These guidelines embody fundamental concepts in statistics; they are consistent with the *Uniform Requirements* (14) and with the upcoming 7th edition of *Scientific Style and Format*, the style manual written by the Council of Science Editors (6) and used by APS Publications. We have written this editorial to provide investigators with concrete steps that will help them design an experiment, analyze the data, and communicate the results. In so doing, we hope these guidelines will help improve and standardize the caliber of statistical information reported throughout journals published by the APS.

GUIDELINES

The guidelines address primarily the reporting of statistics in the MATERIALS AND METHODS, RESULTS, and DISCUSSION sections of a manuscript. *Guidelines 1* and *2* address issues of experimental design.

MATERIALS AND METHODS

Guideline 1. If in doubt, consult a statistician when you plan your study. The design of an experiment, the analysis of its data, and the communication of the results are intertwined. In fact, design drives analysis and communication. The time to consult a statistician is when you have defined the experimental problem you want to address: a statistician can help you design an experiment that is appropriate and efficient. Once you have collected the data, a statistician can help you assess whether the assumptions underlying the analysis were satisfied. When you write the manuscript, a statistician can help you ensure your conclusions are justified.

Guideline 2. Define and justify a critical significance level α appropriate to the goals of your study. For any statistical test, if the achieved significance level P is less than the critical significance level α, defined before any data are collected, then the experimental effect is likely to be real (see Ref. 9, p. 782). By tradition, most researchers define α to be 0.05: that is, 5% of the time they are willing to declare an effect exists when it does not. These examples illustrate that $\alpha = 0.05$ is sometimes inappropriate.

If you plan a study in the hopes of finding an effect that could lead to a promising scientific discovery, then $\alpha = 0.10$ is appropriate. Why? When you define α to be 0.10, you increase the probability that you find the effect if it exists.

In contrast, if you want to be especially confident of a possible scientific discovery, then $\alpha = 0.01$ is appropriate: only 1% of the time are you willing to declare an effect exists when it does not.

A statistician can help you satisfy this guideline (see *Guideline 1*).

Guideline 3. Identify your statistical methods, and cite them using textbooks or review papers. Cite separately commercial software you used to do your statistical analysis. This guideline sounds obvious, but some researchers fail to identify the statistical methods they used.[2] When you follow *Guideline 1*, you can be confident that your statistical methods were appropriate; when you follow this guideline, your reader can be confident also. It is important that you identify separately the commercial software you used to do your statistical analysis.

Guideline 4. Control for multiple comparisons. Many physiological studies examine the impact of an intervention on a set

---

Address for reprints and other correspondence: D. Curran-Everett, Division of Biostatistics, M222, National Jewish Medical and Research Center, 1400 Jackson St., Denver, CO 80206 (E-mail: EverettD@njc.org).

[1] Discussions of common statistical errors, underlying assumptions of common statistical techniques, and factors that impact the choice of a parametric or the equivalent nonparametric procedure fall outside the purview of this editorial.

[2] We include resources that may be useful for general statistics (15), regression analyses (10), and nonparametric procedures (5).

---

of related comparisons. In this situation, the probability that you reject at least one true null hypothesis in the set increases, often dramatically. A multiple comparison procedure[3] protects against this kind of mistake. The false discovery rate procedure may be the best practical solution to the problem of multiple comparisons (see Ref. 8, p. R6–R7).

Suppose you study the concurrent impact of some chemical on response variables A, B, C, D, and E. For each of these five variables are listed the achieved significance level P_i and the false discovery rate critical significance level d_i^* (see Ref. 8, p. R6–R7):

Comparison i :	Variable	P_i	d_i^*
5 :	D	0.061	0.050
4 :	E	0.045	0.040
3 :	A	0.032	0.030
2 :	B	0.017	0.020
1 :	C	0.008	0.010

If $P_i \leq d_i^*$, then the remaining i null hypotheses are rejected. Because $P_2 = 0.017 \leq d_2^* = 0.020$, null hypotheses $2 \rightarrow 1$ are rejected. In other words, after controlling for multiple comparisons using the false discovery rate procedure, only the differences in variables B and C remain statistically significant. The false discovery rate procedure is useful also in the context of pairwise comparisons (see Ref. 8, p. R7).

RESULTS

Guideline 5. Report variability using a standard deviation. Because it reflects the dispersion of individual sample observations about the sample mean, a standard deviation characterizes the variability of those observations. In contrast, because it reflects the theoretical dispersion of sample means about some population mean, a standard error of the mean characterizes uncertainty about the true value of that population mean. The overwhelming majority of original articles published by the APS report standard errors as apparent estimates of variability (9).

To see why a standard error is an inappropriate estimate of variability among observations, suppose you draw an infinite number of samples, each with n independent observations, from some normal distribution. If you treat the sample means as observations, then the standard deviation of these means is the standard error of the sample mean (Fig. 1). A standard error is useful primarily because of its role in the calculation of a confidence interval.

Most journals report a standard deviation using a $\pm$ symbol. The $\pm$ symbol is superfluous: a standard deviation is a single positive number. Report a standard deviation with notation of this form:

$$115 \text{ mmHg (SD 10)} \ .$$

As of July 2004, articles published in APS journals will use this notation in accordance with *Scientific Style and Format* (6).

---

[3] Examples of common multiple comparison procedures include the Newman-Keuls, Bonferroni, and least significant difference procedures (see Ref. 8).
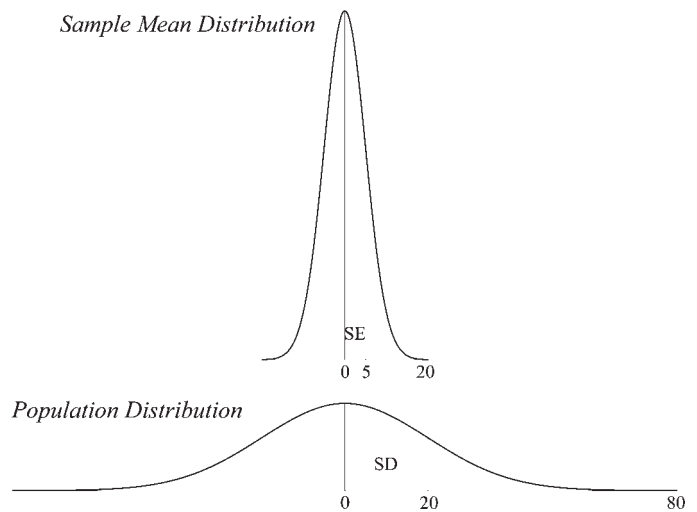

Fig. 1. The difference between standard deviation and standard error of the mean. Suppose random variable Y is distributed normally with mean $\mu = 0$ and standard deviation $\sigma = 20$ (*bottom*). If you draw from this population an infinite number of samples, each with n observations, then the sample means will be distributed normally (*top*). The average of this distribution of sample means is the population mean $\mu = 0$. If $n = 16$, then the standard deviation SD$\{\bar{y}\}$ of this distribution of sample means is SD$\{\bar{y}\} = \sigma/\sqrt{n} = 20/\sqrt{16} = 5$, known also as the standard error of the sample mean, SE$\{\bar{y}\}$. (See Ref. 9, p. 779–781.) Its dependence on sample size makes the standard error of the mean an inappropriate estimate of variability among observations.

This guideline applies also to a data graphic in which you want to depict variability: report a standard deviation, not a standard error.

Guideline 6. Report uncertainty about scientific importance using a confidence interval. A confidence interval characterizes uncertainty about the true value of a population parameter. For example, when you compute a confidence interval for a population mean, you assign bounds to the expected discrepancy between the sample mean $\bar{y}$ and the population mean μ (see Ref. 9, p. 779–781).

The level of confidence in a confidence interval is based on the concept that you draw a large number of samples, each with n observations, from some population. Suppose you measure response variable Y in 200 random samples: you will obtain 200 different sample means and 200 different sample standard deviations. As a consequence, you will calculate 200 different $100(1 - \alpha)\%$ confidence intervals; you expect about $100(1 - \alpha)\%$ of these confidence intervals to include the actual value of the population mean.

How do you interpret a single confidence interval? If you calculate a 99% confidence interval for some population mean to be $[-19, -3]$, then you can declare, with 99% confidence, that the population mean is included in the interval $[-19, -3]$.

This guideline applies also to a data graphic in which you want to depict uncertainty: report a confidence interval.

Guideline 7. Report a precise P value. A precise P value does two things: it communicates more information with the same amount of ink, and it permits each reader to assess individually a statistical result. Suppose the P values associated with the main results of your study are $P = 0.057$ and $P = 0.57$. You might be tempted to report each value as $P > 0.05$ or $P = \text{NS}$. You can communicate that the interpretations of the results differ (see *Guideline 10*) only if you report the precise P values.

Table 1. *Interpretation of P values*

P Value	Interpretation
$P \not\simeq 0.10$	Data are consistent with a true zero effect.
$0.05 \sim P \simeq 0.10$	Data suggest there may be a true effect that differs from zero.
$0.01 \not\simeq P \simeq 0.05$	Data provide good evidence that the true effect differs from zero.
$P \simeq 0.01$	Data provide strong evidence that the true effect differs from zero.

The symbol $\simeq$ means at or near, $\sim$ means near, and $\not\simeq$ means not near. Adapted from Ref. 7.

Guideline 8. Report a quantity so the number of digits is commensurate with scientific relevance. The resolution and precision of modern scientific instruments is remarkable, but it is unnecessary and distracting to report digits if they have little scientific relevance. For example, suppose you measure blood pressure to within 0.01 mmHg and your sample mean is 115.73 mmHg. How do you report the sample mean? As 115.73, as 115.7, or as 116 mmHg? Does a resolution smaller than 1 mmHg really matter? In contrast, a resolution to 0.001 units is essential for a variable like pH. This guideline is critical to the design of an effective table (11).

Guideline 9. In the Abstract, report a confidence interval and a precise P value for each main result.

DISCUSSION

Guideline 10. Interpret each main result by assessing the numerical bounds of the confidence interval and by considering the precise P value. If either bound of the confidence interval is important from a scientific perspective, then the experimental effect may be large enough to be relevant. This is true whatever the statistical result—the *P* value— of the hypothesis test. If $P < \alpha$, the critical significance level, then the experimental effect is likely to be real (see Ref. 9, p. 782).

How do you interpret a *P* value? Although *P* values have a limited role in data analysis, Table 1, adapted from Ref. 7, provides guidance. These interpretations are useful only if the power of the study was large enough to detect the experimental effect.

SUMMARY

The specific guidelines listed above can be summarized by these general ones:

- Analyze your data using the appropriate statistical procedures and identify these procedures in your manuscript: *Guidelines 2–4*.
- Report variability using a standard deviation, not a standard error: *Guideline 5*.
- Report a precise *P* value and a confidence interval when you present the result of an analysis: *Guidelines 6–10*.
- If in doubt, consult a statistician when you design your study, analyze your data, and communicate your findings: *Guideline 1*.

The mere adherence to guidelines for reporting statistics can never substitute for an understanding of concepts and procedures in statistics. Nevertheless, we hope these guidelines, when used with other resources (4, 8, 9, 11, 12, 14), will help improve the caliber of statistical information reported in articles published by the American Physiological Society.

REFERENCES

1. **Altman DG.** Statistics in medical journals: some recent trends. *Stat Med* 19: 3275–3289, 2000.
2. **Altman DG and Bland JM.** Improving doctors' understanding of statistics. *J R Stat Soc Ser A* 154: 223–267, 1991.
3. **American Physiological Society.** Manuscript sections. In: *Information for Authors: Instructions for Preparing Your Manuscript* [Online]. APS, Bethesda, MD. http://www.the-aps.org/publications/i4a/prep_manuscript. htm#manuscript_sections [March 2004].
4. **Bailar JC III and Mosteller F.** Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med* 108: 266–273, 1988.
5. **Conover WJ.** *Practical Nonparametric Statistics* (2nd ed.). New York: Wiley, 1980.
6. **Council of Science Editors, Style Manual Subcommittee.** *Scientific Style and Format: The CSE Manual for Authors, Editors, and Publishers* (7th ed.). In preparation.
7. **Cox DR.** *Planning of Experiments*. New York: Wiley, 1958, p. 159.
8. **Curran-Everett D.** Multiple comparisons: philosophies and illustrations. *Am J Physiol Regul Integr Comp Physiol* 279: R1–R8, 2000.
9. **Curran-Everett D, Taylor S, and Kafadar K.** Fundamental concepts in statistics: elucidation and illustration. *J Appl Physiol* 85: 775–786, 1998.
10. **Draper NR and Smith H.** *Applied Regression Analysis* (2nd ed.). New York: Wiley, 1981.
11. **Ehrenberg ASC.** Rudiments of numeracy. *J R Stat Soc Ser A* 140: 277–297, 1977.
12. **Holmes TH.** Ten categories of statistical errors: a guide for research in endocrinology and metabolism. *Am J Physiol Endocrinol Metab* 286: E495–E501, 2004; 10.1152/ajpendo.00484.2003.
13. **International Committee of Medical Journal Editors.** Uniform requirements for manuscripts submitted to biomedical journals. *Ann Intern Med* 108: 258–265, 1988.
14. **International Committee of Medical Journal Editors.** Uniform requirements for manuscripts submitted to biomedical journals. *Ann Intern Med* 126: 36–47, 1997.
15. **Snedecor GW and Cochran WG.** *Statistical Methods* (7th ed.). Ames, IA: Iowa State Univ. Press, 1980.

Douglas Curran-Everett
Division of Biostatistics
National Jewish Medical and Research Center
and Depts. of Preventive Medicine and Biometrics and of Physiology and Biophysics
School of Medicine
University of Colorado Health Sciences Center
Denver, CO 80262
E-mail: EverettD@njc.org

Dale J. Benos
APS Publications Committee Chair
Dept. of Physiology and Biophysics
University of Alabama at Birmingham
Birmingham, AL 35294
E-mail: benos@physiology.uab.edu