A Cautionary Note on the Use of Error Bars

JOHN R. LANZANTE

NOAA/Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, New Jersey

(Manuscript received 18 May 2004, in final form 3 March 2005)

ABSTRACT

Climate studies often involve comparisons between estimates of some parameter derived from different observed and/or model-generated datasets. It is common practice to present estimates of two or more statistical quantities with error bars about each representing a confidence interval. If the error bars do not overlap, it is presumed that there is a statistically significant difference between them. In general, such a procedure is not valid and usually results in declaring statistical significance too infrequently. Simple examples that demonstrate the nature of this pitfall, along with some formulations, are presented. It is recommended that practitioners use standard hypothesis testing techniques that have been derived from statistical theory rather than the ad hoc approach involving error bars.

1. Introduction

The statistical analysis of climate data typically involves the estimation of quantities such as the mean value or the regression coefficient from a trend analysis. Such estimates are viewed as incomplete without the inclusion of an uncertainty estimate that can then be used to test a hypothesis. For example, in model intercomparison projects, an estimate bracketed by error bars may be presented for each model. In climate change studies, a trend estimate may be presented for competing datasets, or for models and observations. When the error bars for the different estimates do not overlap, it is presumed that the quantities differ in a statistically significant way. Unfortunately, as demonstrated by Schenker and Gentleman (2001, hereafter SG), this is in general an erroneous presumption.

A natural question to ask at this point is "How serious is the problem of misapplication of error bars?" To address this question, the author searched through a recent year's worth of issues from the *Journal of Climate*. A similar but less exhaustive search was made of the Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change (IPCC; Houghton et al. 2001). Instances of inappropriate use of error bars were found in both the *Journal of Climate* and the TAR.[1] Given that these are two of the most respected publications in the field of climate, some clarification on the use of error bars seems warranted.

Before beginning discussion of the problem, it is worth reviewing some of the types of error bars that are commonly presented, along with related terminology. Error bars may represent the standard deviation, the standard error, the confidence interval, or some other measure of error or uncertainty. While the standard deviation is a measure of dispersion of *individual observations* about their mean, the standard error is the standard deviation of a derived *statistic*, such as the mean, regression coefficient, correlation coefficient, etc. A confidence interval can be constructed about a sample statistic such that it contains the true population

Corresponding author address: John R. Lanzante, NOAA/Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542.
E-mail: John.Lanzante@noaa.gov

[1] In the *Journal of Climate* there were 17 articles in which error bars or related measures were presented in figures or tables and used in a two-sample setting. In only two of these articles did the authors correctly apply a two-sample test. In the majority of the other 15 articles, inferences were drawn inappropriately from the error bars; in a few cases the usage was ambiguous, perhaps leading the reader to an inappropriate inference. Although it seemed that conclusions would not change in about a third of the offending articles, for the remainder, conclusions would change in some instances; the extent of change is difficult to determine given the large number of individual cases involved. In the TAR, although proper usage was noted in quite a few cases, four cases of inappropriate use were found.

TABLE 1. Some statistical quantities from two examples involving hypothetical samples of data from observations and a GCM. Given are the sample size (n), the mean of the sample ($\overline{X}$), the std dev of the sample values (s), the standard error or std dev of the mean (SE), and the confidence interval. On the third line (difference) for each example are given the difference of the means (observations minus GCM), the standard error of the difference of the means, and a confidence interval based on a two-sample Student's t test. The intervals have been constructed using ± twice the sampling error about the mean; this is close to the value of 1.96, which in the limiting case of an infinite sample size corresponds to a 95% confidence interval. Note that while these examples have been constructed to produce "round numbers," the concepts that they illustrate are not dependent on either the particular values or the sample sizes.

		n	$\overline{X}$	s	SE	Interval
Example 1 (equal variances)						
	Observations	49	3.0	7.0	1.00	[5.00, 1.00]
	GCM	49	0.0	7.0	1.00	[2.00, −2.00]
	Difference		3.0		1.41	[5.82, 0.18]
Example 2 (disparate variances)						
	Observations	49	4.0	13.9	1.99	[7.97, 0.03]
	GCM	49	0.0	0.1	0.01	[0.03, −0.03]
	Difference		4.0		1.99	[7.97, 0.03]

statistic with a specified probability[2] and thus can be used in hypothesis testing. This note is relevant to error bars that represent confidence intervals for hypothesis testing.

2. The nature of the problem

The misperception regarding the use of error bars may arise because of a fundamental difference between one-sample and two- (or multi) sample testing. For a Gaussian-distributed variate, when only one quantity is estimated, a one-sample test (such as a Student's t test) may be performed. The null hypothesis would be that the estimated quantity is equal to some constant (e.g., that an anomaly is zero). In the one-sample case, application of a t test is equivalent to placing error bars about the quantity to see if it overlaps with the hypothesized value. However, when the interest is in comparing estimated values from two different samples, use of error bars about each estimate, looking for overlap, is not equivalent to application of a two-sample t test.

Suppose, for example, based on a single sample of data generated by a climate model, a 6-K rise in temperature is found to occur over some interval, and suppose that the standard error of this estimate is 2 K. Assuming that the temperatures are drawn from a Gaussian-distributed population, the hypotheses that the true change is zero can be assessed via a one-sample Student's t test. Employing the standard formula yields a t value of 3 that, except for a very small sample size, results in rejection of the null hypothesis of no change in temperature with a high level of confidence. Alternately, if one had preselected the same confidence level by placing error bars at 3 times the standard error about the estimated mean, it would have been found that the interval just intersects zero.

In contrast, the two-sample case is fundamentally different in that, in general, looking for overlap from two sets of error bars is not equivalent to the appropriate t test. Examples based on the data in Table 1 are used to illustrate the nature of the problem. Suppose we have finite samples of values of some quantity from both observed data and from a general circulation model (GCM). Estimates of the mean ($\overline{X}$) and standard deviation (s) of the sample values can be made along with the uncertainty [standard error (SE)] of the estimated means. As is common practice, a confidence interval about the estimated means can be constructed by taking ± twice the standard error. The intervals given in Table 1 are displayed graphically in Fig. 1 in the form of error bars.

For example 1, three sets of error bars are shown on the left side of Fig. 1 for the observations (O), GCM (G), and their difference (D). In this example, the observations and GCM have equal standard deviations. It can be seen that there is considerable overlap of the error bars from the observations and GCM. In such a case, a researcher would typically conclude erroneously that there is no statistically significant difference between their respective means. An alternate approach to the same problem is to apply a two-sample t test. Such a test has been applied and the corresponding error bars about the difference of the means (D) do not in-

[2] A confidence interval is often a scaled version of the standard error. For example, assuming a Gaussian distribution, a 95% confidence interval for the population mean is constructed by extension of ± 1.96 standard errors about the sample mean.

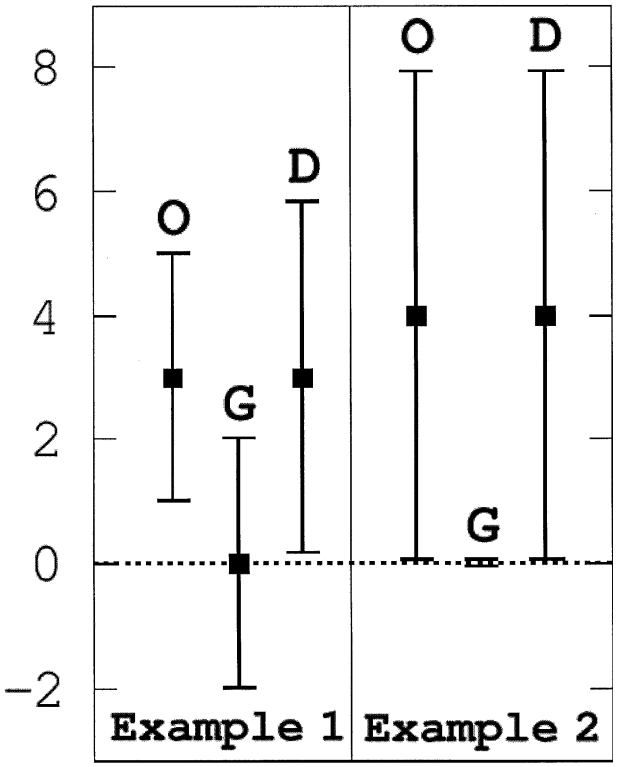

FIG. 1. Error bars corresponding to the data given in Table 1. Each triplet (O, G, and D) corresponds to observations, GCM, and differences, respectively, from Table 1. The left (right) triplet corresponds to example 1 (example 2). Each square corresponds to a mean ($\overline{X}$), and each set of error bars corresponds to a confidence interval. The units on the ordinate are arbitrary.

clude zero. Based on this test, the same researcher would conclude that there is a statistically significant difference between the means.

The reason for this apparent paradox can be understood by considering the relationship between the SE of the mean of the *individual* samples (observations and GCM) to that of the SE of the *difference* of their means. The crucial factor is that in the case of the two-sample *t* test, the SE of the difference is estimated by "pooling" the variances from the two different samples. It should be noted that while the two-sample *t* test is well founded in statistical theory, the use of overlapping error bars in the two-sample case is not.

The depiction of example 1 in Fig. 1 can be used to understand the crucial distinction between the two approaches. In order for significance to be declared using the overlapping error bars approach, one would have to move the means of the observations farther apart until the bottom whisker from the observations just touches the top whisker from the GCM. This can be expressed mathematically as

$$\overline{X}_1 - \overline{X}_2 \geq c\,\mathrm{SE}_1 + c\,\mathrm{SE}_2, \tag{1}$$

where $\overline{X}_1$ ($\overline{X}_2$) is the mean of the observations (GCM), SE_1 (SE_2) is the standard error of the observations (GCM), and c is a constant that determines the level of confidence. The two terms on the rhs of (1) represent the distances from their respective means to the end of the whisker (i.e., half the length of the confidence intervals). In example 1, $c = 2$ since the confidence intervals represent ± two standard errors. Note that throughout this paper no distinction is made between population and sample parameters; it should be understood that estimates of various population parameters from an available finite sample are being used.

Application of the two-sample *t* test to example 1 results in a different requirement for declaration of significance

$$\overline{X}_1 - \overline{X}_2 \geq c\,\mathrm{SE}_3, \tag{2}$$

where SE_3 is the standard error of the differences. The distinction between the two approaches lies in the relationship between the individual standard errors (SE_1 and SE_2) and the standard error based on pooling the individual variances (SE_3). Under some simplifying assumptions, this relationship, presented by SG along with some theoretical underpinnings and derived algebraically in the appendix, is

$$\mathrm{SE}_3 = (\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^{1/2}. \tag{3}$$

To gain insight, it is instructive to use the geometric analogy given by SG involving a right triangle. The lengths of the sides are given by SE_1 and SE_2, while the length of the hypotenuse is given by SE_3. Thus, (3) is simply an expression of the Pythagorean relationship. It can be reasoned that for the same difference in means [the left-hand sides of (1) and (2)] it will be more difficult to declare significance using (1) than (2). This is true because the rhs of (1) represents the sums of the lengths of the sides of a right triangle whereas the rhs of (2) represents the length of the hypotenuse; the latter will always be less than or equal to the former.

The amount of disparity between the two approaches depends on the ratio of the lengths of the sides of the triangle, which is equivalent to the ratio of the standard errors of the two samples. Example 1 was concocted to have equality in this regard and illustrates the case of maximum disparity. It is easy to demonstrate (SG) that when $\mathrm{SE}_1 = \mathrm{SE}_2$, the ratio of the rhs of (1) to the rhs of (2) takes on its maximum value of $\sqrt{2}$. Unfortunately,

in practice it is often the case that the two samples have comparable standard errors.

At the other extreme, when the SE from one sample is much larger than the other, (2) and (3) approach equivalency. Geometrically this occurs as one side of the right triangle approaches zero length, in which case the remaining side is also the hypotenuse. Example 2 illustrates this situation in that the variance from the observations is much larger than that from the GCM. As seen on the right side of Fig. 1, the error bars from observations and GCM just overlap and the error bars about the difference almost touch the zero line. Unfortunately, instances of large disparity in variances are typically not so common.

3. Conclusions

In summary, this note has addressed the common practice of placing error bars about the means from two distinct samples. While this practice lends itself to an appealing graphical presentation, it can often lead to an erroneous conclusion as to whether the means differ in a statistically significant manner. In particular, this approach will lead to a conservative bias in that sometimes no difference is found when it should. However, this bias is not constant, varying depending on the relative magnitudes of the sampling errors in the two samples. The maximum bias is found when the sampling variability of the two samples is comparable.

While this note has dealt with testing the difference between means, the same cautions apply to other statistical quantities. Practitioners should opt for appropriate two-sample tests suited for the parameter of choice, or for a multiple comparisons test when more than two samples are involved. While the error bar method is tempting, it is not grounded by statistical theory when more than one sample is involved.

Acknowledgments. Gabriel Lau and Keith Dixon kindly provided comments on an earlier draft of this manuscript. Neville Nicholls, editor David Stephenson, and an anonymous reviewer provided useful suggestions during the journal review.

APPENDIX

Derivation of Relationship between Standard Errors

The relationship expressed by (3), which is associated with a two-sample Student's *t* test, is derived here by invoking some simplifying assumptions as per SG. This relationship relates the individual standard errors of the means from two samples (SE_1 and SE_2) with the standard error of the difference of their means (SE_3). The derivation utilizes a number of standard statistical equations [(A1)–(A6) and (A10)–(A11)] associated with the two-sample *t* test. These are available from any of a number of introductory statistics texts, for example, Zar (1996).

Begin by defining the pooled variance (S_p^2) from the two samples

$$S_p^2 = (SS_1 + SS_2)/(\upsilon_1 + \upsilon_2), \tag{A1}$$

where SS_1 (SS_2) is the corrected sum of squares from the first (second) sample and υ_1 (υ_2) is the degrees of freedom for the first (second) sample such that

$$\upsilon_1 = n_1 - 1 \tag{A2}$$

and

$$\upsilon_2 = n_2 - 1, \tag{A3}$$

where n_1 (n_2) is the sample size of the first (second) sample and where the variances of the two samples are defined by

$$s_1^2 = SS_1/\upsilon_1 \tag{A4}$$

and

$$s_2^2 = SS_2/\upsilon_2. \tag{A5}$$

Next define the standard error of the difference of the means of the two samples:

$$SE_3 = (S_p^2/n_1 + S_p^2/n_2)^{1/2}. \tag{A6}$$

Now substituting (A1), (A4), and (A5) into (A6) and rearranging the terms yields

$$SE_3 = [(1/n_1 + 1/n_2)\,(\upsilon_1 s_1^2 + \upsilon_2 s_2^2)/(\upsilon_1 + \upsilon_2)\,]^{1/2}. \tag{A7}$$

The first simplifying assumption invoked by SG is to consider the limiting case where the sample sizes n_1 and $n_2 \to \infty$. In this limit $\upsilon_1 \to n_1$ and $\upsilon_2 \to n_2$ so that after algebraic cancellation (A7) becomes

$$SE_3 = (s_1^2/n_2 + s_2^2/n_1)^{1/2}. \tag{A8}$$

The second simplification is the assumption that the sample sizes are equal ($n_1 = n_2$) so that (A8) becomes

$$SE_3 = (s_1^2/n_1 + s_2^2/n_2)^{1/2}. \tag{A9}$$

Next note the definitions of the standard errors of the means of the two samples

$$SE_1 = (s_1^2/n_1)^{1/2} \tag{A10}$$

and

$$SE_2 = (s_2^2/n_2)^{1/2}. \quad \text{(A11)}$$

Finally, substitute (A10)–(A11) into (A9) to yield an expression equivalent to (3):

$$SE_3 = (SE_1^2 + SE_2^2)^{1/2}. \quad \text{(A12)}$$

The relationship expressed by (A12) is useful for gaining insight as to the difference between the use of individual error bars (based on SE_1 and SE_2) from the two samples and application of the two-sample *t* test (based on SE_3). Although it is based on some simplifying assumptions, yielding a less complex relationship for illustrative purposes, the principle that it expresses, namely the distinction between the two approaches, is more generally applicable.

REFERENCES

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., 2001: *Climate Change 2001: The Scientific Basis.* Cambridge University Press, 881 pp.

Schenker, N., and J. Gentleman, 2001: On judging the significance of differences by examining the overlap between confidence intervals. *Amer. Stat.,* **55,** 182–186.

Zar, J., 1996: *Biostatistical Analysis*. 3d ed. Prentice Hall, 662 pp.