Chemometrics and Intelligent Laboratory Systems, 11 (1991) 131–136
Elsevier Science Publishers B.V., Amsterdam

Bias correcting confidence intervals for a nearly common property

Daren B.H. Cline and Clifford H. Spiegelman *

Statistics Department, Texas A&M University, College Station, TX 77843 (USA)

(Received 7 October 1990; accepted 13 January 1991)

Abstract

Cline, D.B.H. and Spiegelman, C.H., 1991. Bias correcting confidence intervals for a nearly common property. *Chemometrics and Intelligent Laboratory Systems*, 11: 131–136.

Confidence intervals are an important tool. Realistic confidence intervals account for both random errors and systematic errors (bias). We improve the usual method for combining random and systematic errors. The new methods are simple and often result in increased accuracy for confidence interval levels.

INTRODUCTION

An important problem in science is to determine when individual measurements taken under different conditions are measurements of a nearly common property. When they are, the proper and common practice is to combine these measurements into a single estimate. For example, when certifying standard reference materials (SRMs) several different types of measurement devices are commonly used. Each device has its own systematic error (bias) and it is not clear whether these devices are measuring nearly the same property. We say that devices are measuring nearly the same property if the population means from these devices are within the stated bounds on systematic error. It is assumed that each bias can be quantified so that a known bound on bias error can be calculated.

This paper provides methods for testing whether these devices are measuring nearly the same property. The first method is based upon the usual bias correction to t-confidence intervals while the second, more powerful, method relies on a sophisticated bias correction to t-confidence intervals.

If we do not reject the null hypothesis that the methods measure nearly the same property then techniques found in refs. 1–3 can and should be applied to produce a single estimate as well as an uncertainty statement for the estimate of this nearly common property. We formulate the null hypothesis as stating that the means of the different measurements do not differ from a common value by more than the stated bounds on bias.

0169-7439/91/$03.50

NOTATION

We follow the notation in ref. 1. Let:

$$X_{ij} = \mu + r_i + \epsilon_{ij}, \quad i, = 1, \ldots, I$$
$$j, = 1, \ldots, n_i \geqslant 2$$

Here r_i represent the bias (or 'systematic error') for group i, μ represents the value of the common property, and the measurement variables, ϵ_{ij}, are all independent with $\mathrm{Var}(\epsilon_{ij}) = \sigma_i^2$ and mean 0. Under the additional assumption of Gaussian error distribution the best estimator of $\mu + r_i$ is $\bar{X}_i = \Sigma_{j=1}^{n_i} X_{ij}/n_i$. As in ref. 1 we assume that known bounds can be placed upon the biases r_i, $|r_i| \leqslant M_i$. These bounds are usually based upon judgement, experimental calculations, and/or theoretical calculations. (See ref. 1 for a discussion of the bounds.)

The use of bounds for systematic error is routine practice in standards laboratories [1,4,5]. The effect of bias on confidence intervals is a topic of current interest to statisticians who work on nonparametric regression and calibration (smoothing), see ref. 7 for example. The topic of this paper would be a very simple case of nonparametric regression.

We assume symmetric bounds on the biases. If the bounds are asymmetric $L_i \leqslant r_i \leqslant U_i$ then replace $\bar{X}_i$ with $\bar{X}_i - (U_i + L_i)/2$ and replace M_i with $(U_i - L_i)/2$.

This allows us to state the hypotheses to be tested as:

$H_o: |r_i| \leqslant M_i$

H_a: at least one $|r_i| > M_i$.

METHODS

The tests use a simple graph and probability calculations. We consider two cases:

The first case that we consider is when the bounds, M_i, are small. In this case we form the standard bias corrected confidence intervals for μ, see ref. 4, $\bar{X}_i \pm (t_{1-\alpha'/2, n_i - 1} S_i/\sqrt{n_i} + M_i)$. The constant $t_{1-\alpha'/2, n_i-1}$ is the $1 - \alpha'/2$ percentile of the t distribution with $n_i - 1$ degrees of freedom, and S_i^2 is the usual unbiased estimate of σ_i^2.

If any of these confidence intervals fail to overlap then the set of means are judged different at the $1 - (1 - \alpha')^k$ level, and individual (or group) estimates must be used. Alternatively Bonferonni's inequality can be used to give a conservative level of the test to be $k\alpha'$.

Our alternative approaches are useful for cases when at least one bias bound is big. One method is a procedure found in ref. 7 to shorten the length of the bias corrected confidence intervals. Assuming that r_i is within the stated bound, it follows that the minimum probability that $\bar{X}_i \pm (kS_i/\sqrt{n_i} + M_i)$ contains μ is $P(-M_i - kS_i/\sqrt{n_i} \leqslant \bar{\epsilon}_i \leqslant M_i + kS_i/\sqrt{n_i})$ which is greater than or equal to $P(-2M_i - kS_i/\sqrt{n_i} \leqslant \bar{\epsilon} \leqslant kS_i/\sqrt{n_i})$. This lower bound is attained when r_i equals either M_i or $-M_i$. Now if $S_i = \sigma_i$ (or the degrees of freedom are extremely large) then we can solve for the k such that

$$P\left(-2M_i - kS_i/\sqrt{n_i} \leqslant \bar{\epsilon}_i \leqslant kS_i/\sqrt{n_i}\right) = 1 - \alpha' \quad (1)$$

(This was done in ref. 7 where the probability was calculated from a normal distribution). Note that for any $\alpha' > 0$ and $M_i > 0$, k is less than the $1 - \alpha/2$ percentile of the relevant t or standard normal distribution. Our first method simply assumes that S_i can be used as an estimator of σ_i and that in the last probability calculation a t-distribution can be used. For this reason we have called it the 'plug-in' method. We can see from Fig. 1 that the approximation is adequate for ranges of bias and accuracy requirements encountered in practice.

An alternative procedure, which is based on an analytic approximation, requires computing the quantile function of Student's t distribution rather than solving eq. (1). We refer to it as the 'tail-approximation' method. The algorithm has three steps, executed once, and they are (for $n = n_i$):

(1) let $\hat{m} = 4\dfrac{nM_i^2}{S_i}$.

(2) let $\hat{q} = 1 - p_n(\hat{m})$,

where p_n is the χ_n^2 distribution.

(3) let $\hat{k} = t_{n-1}\left(1 - \dfrac{\alpha}{1 + \hat{q}}\right)$,

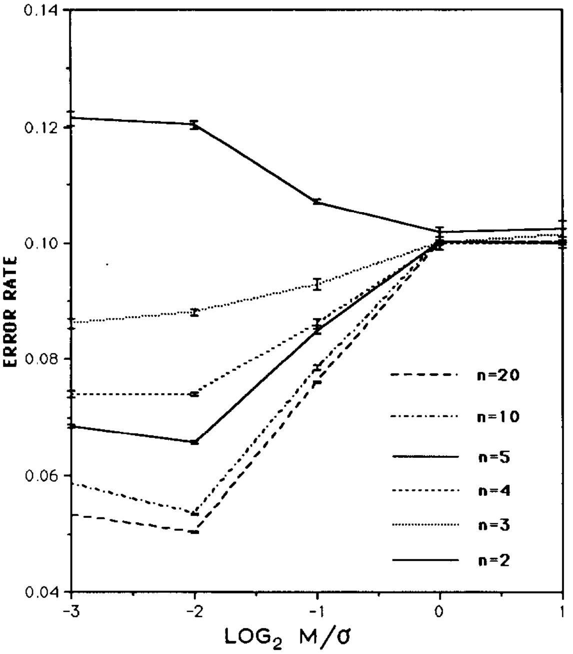

Fig. 1. Error rate for plug-in method with one standard error limit. The nominal rate is 0.10.

where t_{n-1} is the Students' $t(n-1)$ quantile function.

Then $\hat{k}$ is our new t-value.

The justification for this algorithm is as follows. We wish to solve

$$\alpha = 1 - P\left(-2M_i - \frac{k\hat{S}_i}{\sqrt{n}} \leqslant \bar{\epsilon}_i \leqslant \frac{k\hat{S}_i}{\sqrt{n}}\right) \tag{2}$$

Letting $T = \sqrt{n}\,\bar{\epsilon}_i/S_i$, $W = n\bar{\epsilon}^2/\sigma_i^2 + (n-1)S_i/\sigma_i^2$ and $m = 4nM_i^2/\sigma_i^2$, one may show that the right hand side of eq. (2) is equal to

$$P(T>k) + \left(T>k,\ W > m\frac{T^2+n-1}{(T-k)^2}\right)$$

For large k (such as the solution to the equation), the distribution of T/k, given $T>k$, is nearly Pareto and is nearly independent of k [8] (apparent from the density of T in its tails). We may define

$$q(m) = \lim_{k\to\infty} P\left(W > m\frac{T^2+n-1}{(T-k)^2}\,\middle|\, T>k\right)$$
$$\leqslant P(W>m) \tag{3}$$

Hence

$$\alpha = P(T>k)(1+q(m))$$
$$\leqslant P(T>k)(1+P(W>m))$$

After first estimating m, this is easily solved for k to give the stated algorithm. Note that for known m (and thus k), the method is guaranteed to be conservative since each term in the limit in (3) is bounded by $P(W>m)$. The approximation would be improved by actually evaluating $q(m)$ (with numerical integration) rather than by bounding it with $P(W>m)$.

EVALUATION

Our conclusions (below) are based on Monte Carlo methods of integration. We determined, for each of the proposed methods, the error rate and the expected decrease (from $t_{\alpha/2}$) in the t-value. The integrations use a simple regression technique (see ref. 6, p. 66) relying on the fact that each expectation is the integral of a function of a chi-square $(n-1)$ random variable $Y = (n-$

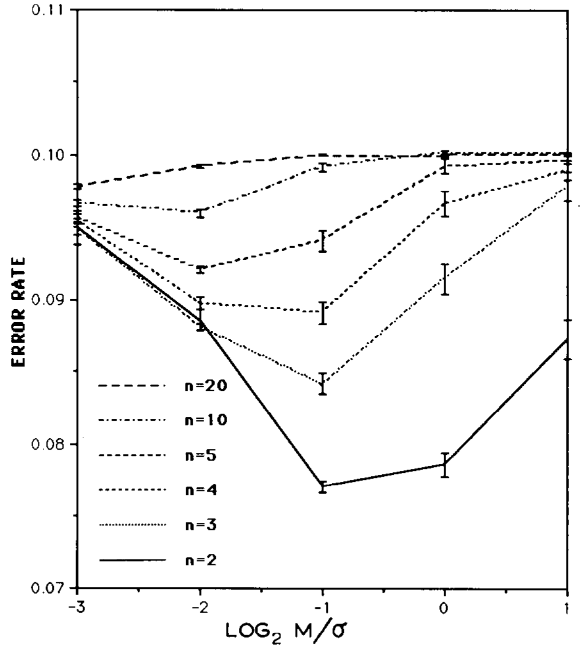

Fig. 2. Error rates for tail method with one standard error limit. The nominal rate is 0.10.

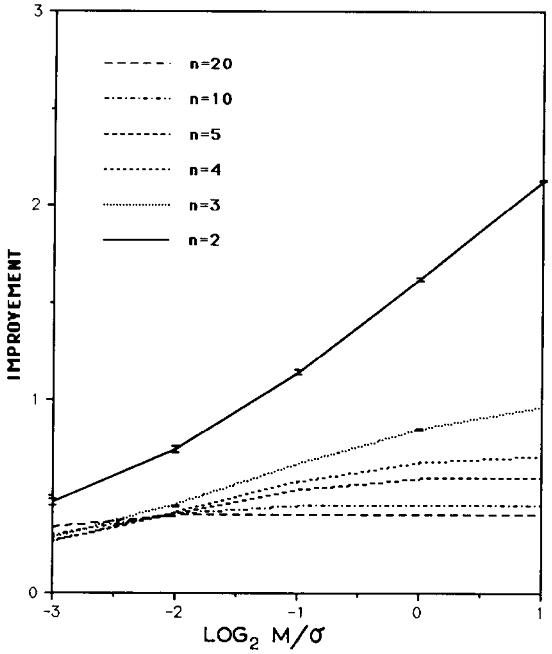

Fig. 3. Improvement in *t* values for plug-in method with one standard error limit.

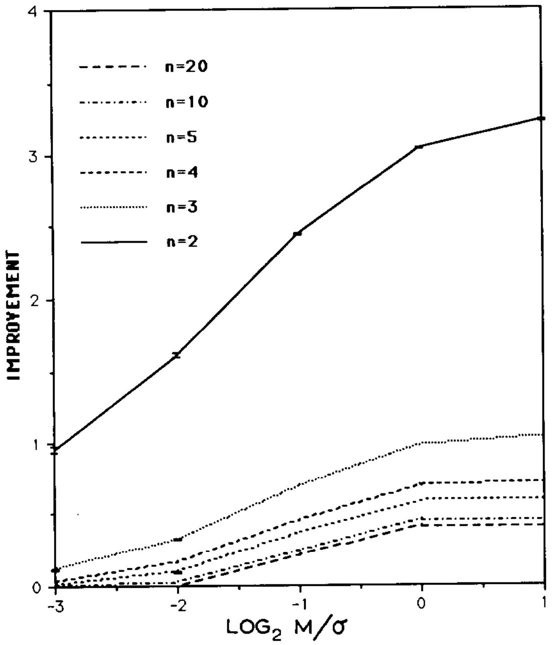

Fig. 4. Improvement in *t* values for tail approximation method with one standard error limit.

$1)S_i^2/\sigma_i^2$. The error rates were estimated by regressing the conditional probability of error, given Y, for each method, on the corresponding conditional probability using $t_{\alpha/2}$ as the t-value. The expected t-values were obtained by regressing the t values for each method on Y itself. We used 1000 random variates to generate the results.

Figs. 1–4 illustrate these estimates, as functions of sample size (n) and relative bias (M/σ_i). The relative bias was varied from 0.125 to 2.0 and the sample size was varied from 2 to 20. No substantial improvements were seen for samples of size greater than 20. Standard error bars are provided in the graphs, though in many instances they are too small to be seen.

CONCLUSIONS

The usual method of correcting for systematic error by adding the bias bound to both sides of the confidence interval is unduly conservative. It gives wider confidence intervals than desired as well as conservative coverage of the confidence intervals. We have given two methods that more accurately expand the standard textbook t-confidence intervals so as to account for systematic error. Both methods produce confidence intervals that are always shorter than the usual bias corrected confidence intervals. This can be seen in Figs. 3 and 4 where the improvement in the t-value used is as much as 3 (a 50% improvement). If the bias is large then both of our methods are more accurate than the usual method, as can be seen from Figs. 1 and 2. The standard corrections are very conservative. The values for the bias to sigma ratio were taken from extensive experience at the National Institute for Standards and Technology (formerly the NBS). Therefore, for many confidence intervals a substantial increase in accuracy and narrower confidence intervals can be obtained in practice. C. Eisenhart, a former President of the American Statistical Association, says that the power of a statistical method is its mathematical power (in this case related to the width of the confidence intervals) times the number of times that it is used. By this definition the two methods presented have the potential to be powerful statistical techniques.

Most of the time the plug-in method and the tail-approximation method are about equally accurate (see Figs. 1 and 2) and so we recommend the plug-in method except when the bias is large and the sample size is small. When the sample size is 2 or 3 (a common occurrence in one of the authors' experience) and the bias is large ($M/\sigma > 0.5$) then the tail-approximation method is more accurate and produces shorter intervals.

An important feature of the usual 90% confidence intervals is that they are conservative in the presence of outliers. When extreme observations are present the multiple of the t percentile times the estimated standard deviation generally shifts more than the sample mean. Both of our confidence interval methods also appear to be robust. The effect of an extreme observation also decreases the estimate of M/σ so that our intervals become more conservative. Small scale simulations have shown us that this heuristic argument is valid. It is possible to construct examples where the usual t-intervals are not conservative and therefore the same examples would also make out intervals non-robust. However, the user happy with t-interval performance would be happy with ours. In no case will either of our methods produce a t value smaller than the α percent point from the t-distribution with the appropriate degrees of freedom.

In addition, we have provided several figures illustrating how the calculated t-values for the two methods will depend on the observed value of M/S. These figures can be used to obtain the approximate t-values in practice. More to the point, however, they illustrate the relationship between the two methods.

For example, Fig. 5 demonstrates that when $n = 2$ the plug-in t-value is larger and more conservative than the tail-approximation. In fact, unless the bias is very small, the tail-approximation method is clearly to be preferred. This relationship evolves as n increases until it is nearly reversed when $n = 5$ (Fig. 6). For $n \geqslant 5$, the tail-approximation method appears to be much too conservative except for very large bias, when the two methods are equivalent.

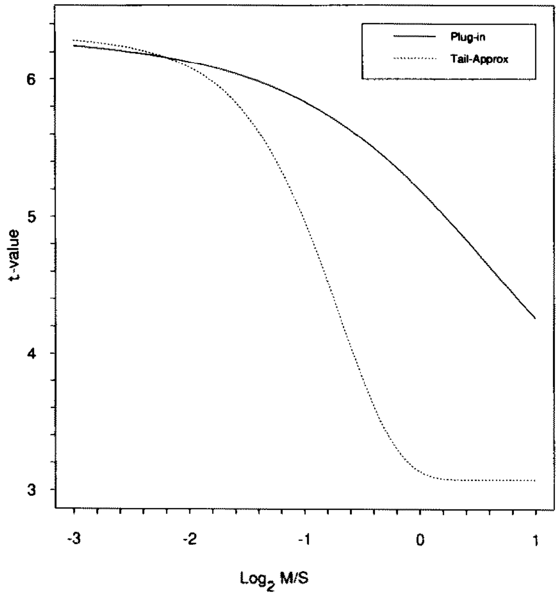

Fig. 5. t-Value vs. $\log_2 M/S$. Comparing methods when $n = 2$.

The figures thus substantiate the conclusions of our simulation: the tail-approximation method is preferred when $n = 2$ or 3, though less so in the latter case, and not otherwise.

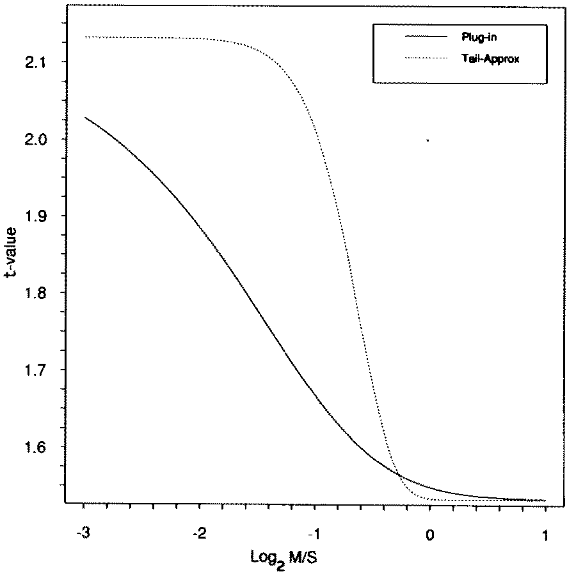

Fig. 6. t-Value vs. $\log_2 M/S$. Comparing methods when $n = 5$.

ACKNOWLEDGEMENTS

Research partially supported by National Science Foundation Grants DMS-9001011 and DMS-9000590 and Shell Development Corporation.

REFERENCES

1 K.R. Eberhardt, C.P. Reeve and C.H. Spiegelman, A minimax approach to combining means, with practical examples, *Chemometrics and Intelligent Laboratory Systems,* 5 (1988) 129–148.

2 T.E. Norwood, Jr. and K. Hinkelman, Estimating the common mean of several normal populations, *Annals of Statistics*, 5 (1977) 1047–1050.

3 P.S.R.S. Rao, Cochran's contributions to variance component models for combining estimates, in P.S.R.S. Rao and J. Sedransk (Editors), *W.G. Cochran's Impact on Statistics*, Wiley, New York, 1984, 203–221.

4 C. Eisenhart, Contributions to panel discussion on adjustments of the fundamental constants, in D.N. Langenberg and B.N. Taylor (Editors), *Precision Measurement and Fundamental Constants*, NBS Special Publication 343, Government Printing Office, Washington, DC, 1971, pp. 509–525.

5 C. Eisenhart, Expression of uncertainties of final results, in H.H. Ku (Editor), *Precision Measurement and Calibration: Statistical Concepts and Procedures*, Special Publication 300, US Department of Commerce, pp. 69–72.

6 J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, Wiley, New York, 1964.

7 G. Knafl, J. Sacks, C. Spiegelman and D. Ylvisaker, Nonparametric calibration, *Technometrics*, 26 (1984) 233–241.

8 S. Kotz and N. Johnson (Editors), *Encyclopedia of Statistical Sciences*, Vol. 6, Wiley, New York, 1985, pp. 568–574.