Anal. Chem. **1996,** *68,* 763−770

Application of the Bootstrap to Calibration Experiments

Geoffrey Jones,[†] Monika Wortberg,[‡] Sabine B. Kreissig,[§] Bruce D. Hammock,[‡] and David M. Rocke*,[†]

Graduate School of Management and Departments of Entomology and Environmental Toxicology, University of California, Davis, California 95616, and Department of Experimental Therapy (0425), German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

In calibration experiments, a number of samples of known concentration are used to establish the relationship between a measured response and sample concentration; this relationship is then used to estimate the unknown concentration of further samples from their measured responses. In addition to the estimates themselves, it is useful to have available some measure of their precision, usually given in the form of confidence limits. The standard method of inverting prediction limits is found to work well in simple situations, but in nonlinear multivariate calibration it becomes intractable. The bootstrap offers an alternative methodology, but in the calibration framework its application is not obvious. We describe some considerations in bootstrapping calibration data and compare our methods with a previous attempt and with the standard method in linear, nonlinear, and multivariate situations. The bootstrap is found to be a useful tool in those situations where the standard method is difficult to implement.

Here we consider an assay method that delivers a response Y dependent upon the sample concentration x. Experimental error implies that Y is random, but we assume that its relationship to x has the form

$$Y = f(\theta,x) + \epsilon \tag{1}$$

where f is a function, assumed known, describing the relationship, θ is a vector of unknown parameters, and ϵ is an error term, which might be assumed to follow some known distribution. For example, in the simple linear case we might have

$$Y = a + bx + \epsilon \tag{2}$$

where a and b are the unknown intercept and slope parameters and ϵ follows a normal distribution with constant variance and zero mean.

In calibration, we first take n observations (x_i,Y_i), where the x_i are known "standard" concentrations prechosen to cover the required range of the assay. These standards are used to produce a calibration curve by estimating the parameter θ in eq 1, using some form of regression technique. Thus, the equation of the calibration curve is

$$y = f(\hat{\theta},x) \tag{3}$$

where $\hat{\theta}$ is the regression estimate of the parameters. If we now have the response Y_0 from an additional sample with unknown concentration x_0, we can invert the calibration eq 3 to get an estimate $\hat{X}_0$ for x_0. In the case of the simple linear model (eq 2), we get

$$\hat{X}_0 = (Y_0 - \hat{a})/\hat{b} \tag{4}$$

In practice, it is usual to have a small number of replicates of the unknown, yielding responses $Y_{01}...Y_{0r}$, where r denotes the number of replicates. In this case, an appropriate mean value can be used.

The standard method of producing a confidence interval is attributed to Fieller.[1] The regression procedure that produces the calibration curve can also be used to calculate so-called "prediction limits": for any concentration x, we get an interval Y_L,Y_U such that, if a future response Y is measured on a sample with concentration x, the probability that Y will be in this interval has a specified value (usually 90% or 95%). In a calibration setting, the prediction limits can be inverted to give a confidence interval: given the reponse y_0, the 90% confidence interval is those values of x whose 90% prediction interval contains y_0. The situation is illustrated graphically in Figure 1a.

Bonate[2] investigated the use of the bootstrap for producing confidence intervals in linear calibration. He found that his bootstrap intervals failed to achieve the desired coverage, particularly for small numbers of replicates, so that intervals which should have contained the true concentration 90% of the time only did so in fact about 40% of the time. By improving his method, we obtain bootstrap confidence intervals with much better coverage, even for small numbers of replicates. The methodology can be applied fairly easily, even in complex nonlinear and multivariate situations, where the standard method becomes extremely difficult to work with. Examples of such systems include nonlinear receptors and bioassays as well as immunoassays. We are especially interested in applying this method to immunoassays, particularly in multivariate analysis.

The bootstrap examines the variability of an estimate by using the existing data, together with some assumptions about how it was generated, to produce new, but plausible, "pseudo data sets". Estimates can be obtained for each of the pseudo data sets and the resulting values examined to derive approximations to the

[†] Graduate School of Management, University of California.

[‡] Departments of Entomology and Environmental Toxicology, University of California.

[§] German Cancer Research Center.

(1) Fieller, E. C. *J. R. Stat. Soc., Ser. B* **1954**, *16*, 175−185.
(2) Bonate, P. L. *Anal. Chem.* **1993**, *65*, 1367−1372.

0003-2700/96/0368-0763$12.00/0

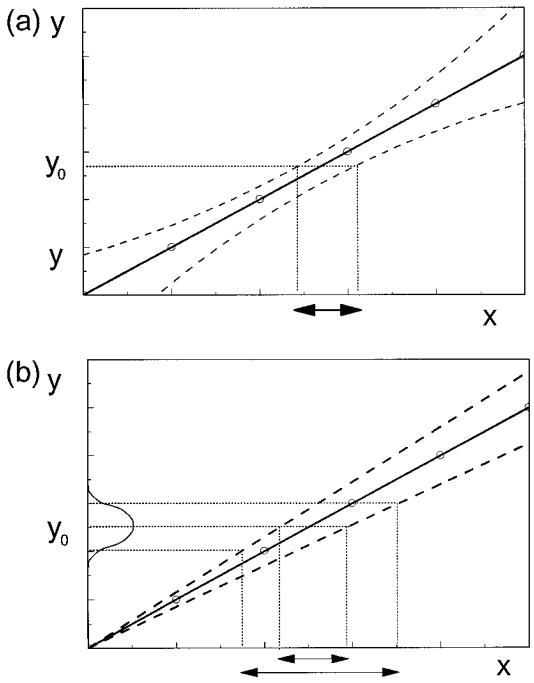

Figure 1. Two methods of producing a confidence interval given a response Y_0: (a) using the prediction limits from regression (dashed lines) and (b) using the bootstrap to simulate variability in Y_0 and the calibration line.

statistical characteristics of the original estimate. More background information and references are given in Bonate.[2] A useful introduction to the theory and implementation of the bootstrap is given by Efron and Tibshirani.[3]

In the case of calibration data, Bonate's bootstrap resamples the residuals from the calibration curve to create new calibration data (x_i^*,Y_i^*) and then uses the resulting bootstrap calibration curve and the observed response Y_0 from the unknown sample to calculate bootstrap estimates $\hat{X}^*$. This is repeated a large number of times (1000) and the distribution of $\hat{X}^*$ values used to produce a confidence interval. This ignores, however, the variability inherent in the Y_0 value, as shown in Figure 1b. We propose creating bootstrap Y_0^* values by further resampling from the residuals; thus in our bootstrap data sets, all the responses Y for both standards and unknowns are replaced by new values Y^*. This simple expedient gives a coverage probability much closer to the required level.

We also suggest adjusting the residuals as described by Efron and Tibshirani.[3] The residual variation around a sample mean or fitted curve is too small, by a known factor, to accurately reflect the variation in responses. Multiplying by the appropriate factor,

$$\sqrt{n/(n-p)} \tag{5}$$

where n is the number of points and p the number of parameters, adjusts the residuals to allow for this.

The general approach of using residuals from both standards and unknowns is shown diagrammatically in Figure 2, the details of which are explained in the examples below. We also explore the use of bootstrap-t intervals.[3,4] Other considerations concerning replication and lack-of-fit may also arise, and these too are discussed and illustrated in the simulations and examples.

First, we follow Bonate in examining the linear model with constant coefficient of variation. We compare our intervals with Bonate's and with the standard method using simulation, extending Bonate's framework to include nonnormal errors in the reponse. We then look at an example of nonlinear calibration, using both simulation and real data. Finally, we consider, using a real data set, a difficult nonlinear multivariate problem where the standard method becomes intractable.

LINEAR CALIBRATION

Following Bonate, we simulate from the model in eq 2, with errors ϵ having constant mean and standard deviation proportional to the expected value of Y, so that the coefficient of variation (cv) is constant; the calibration line is estimated using weighted least-squares regression with weights $w_i = 1/x_i^2$. Six calibration standards were used comprising triplicates of a low and a high concentration, e.g., 10, 10, 10, 1000, 1000, 1000, and the parameters a and b were chosen as described by Bonate.

The prediction limits are then

$$\hat{a} + \hat{b}x \pm ts\sqrt{\frac{x^2}{r} + \frac{1}{\sum w_i} + \frac{(x - \boldsymbol{x}_w)^2}{\mathrm{SS}x_w}} \tag{6}$$

where $\bar{x}_w$ is the weighted mean and $\mathrm{SS}x_w$ the weighted sum of squares of the standard concentrations, t is a percentage point of the appropriate t-distribution, and s is an estimate of the standard deviation of the errors. If s is taken as the square root of the mean square error from the calibration curve estimation, it has $n - 2$ degrees of freedom; a better approach is to combine estimates from this and from the r replicates of the unknown, giving $n + r - 3$ degrees of freedom. Then, for a 90% prediction interval, t is the 95th percentile of the t_{n+r-3} distribution. The standard confidence interval for an unknown with mean response $\bar{y}$ is calculated by finding the values of x that make either prediction limit equal to $\bar{y}$. On rearranging, this gives quadratic equations for the lower and upper limits that are easily solved.

To illustrate our approach to the bootstrap, we work through an example shown in Figure 2. Given the responses for the standards, an appropriate regression (here weighted least-squares) gives estimates of the parameters and n unweighted residuals $R_i = (Y_i - \hat{a} - \hat{b}x_i)/x_i$. These are adjusted as described above and placed in a residual pool. Further residuals are obtained from the responses for the unknown sample by subtracting their mean. Since our analysis uses weights dependent on x, and x_0 is unavailable for the unknown sample, the estimate $\hat{X}_0$ is used instead, so that the residuals are

$$R_i = (Y_{0i} - \bar{V}_0)/\hat{X}_0 \tag{7}$$

An exception must be made when there is only one replicate of the unknown: here, only the standards would contribute to the residual pool, although the unknown would still receive from the pool as described below. The residual from an unreplicated unknown would be zero, and the adjustment factor infinite, so its inclusion would not be possible.

(3) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman and Hall: New York, 1993.

(4) Hall, P. *The Bootstrap and Edgeworth Expansion*; Springer-Verlag: New York, 1992.

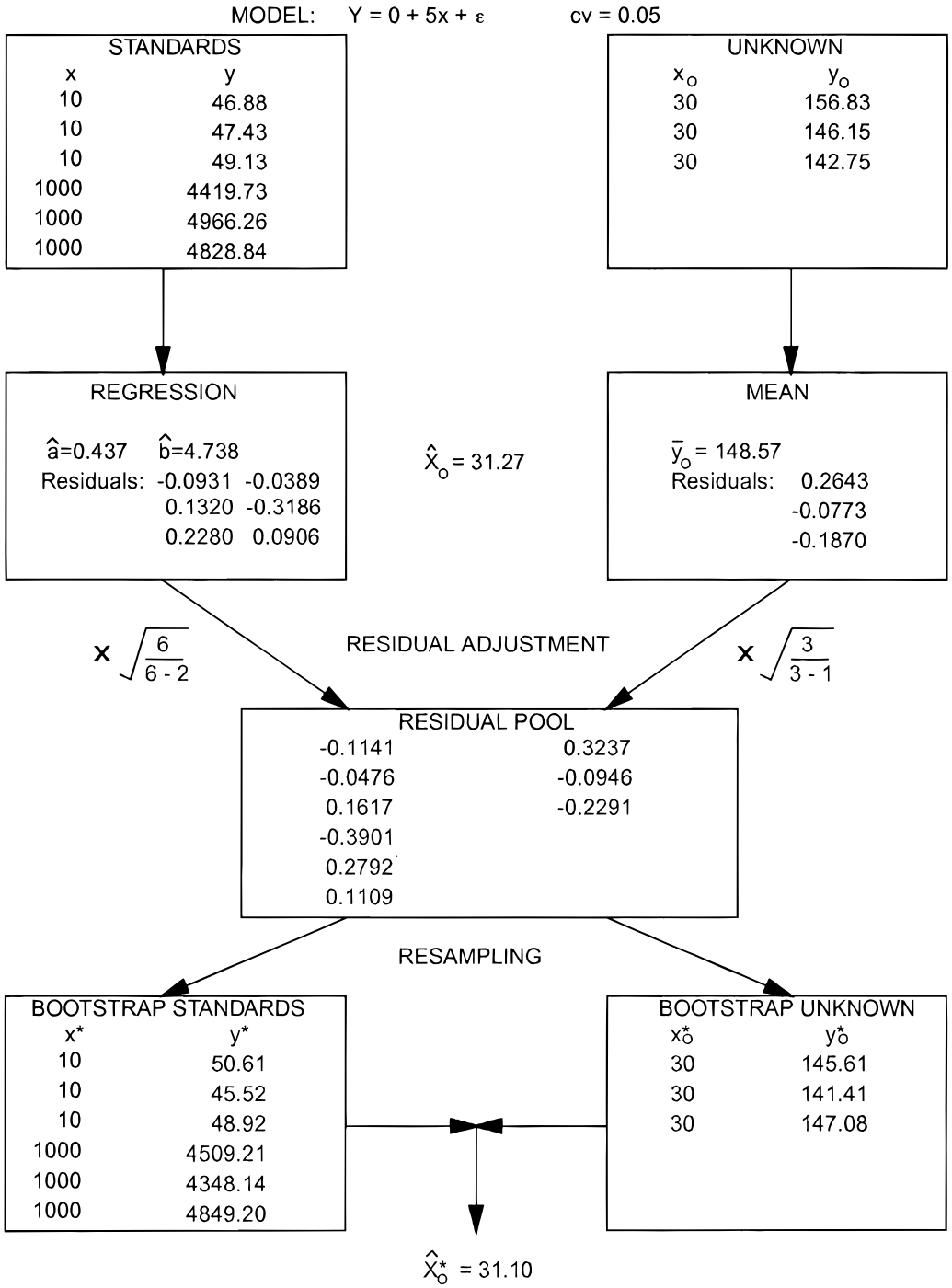

Figure 2. Example of the formation of a residual pool to produce bootstrap estimates $\hat{X}^*_0$. Figures are rounded from the computer output, so some calculations appear inexact.

Bootstrap data sets are now formed by sampling with replacement from the residual pool. The bootstrap responses are given by

$$Y^* = \hat{a} + \hat{b}x + xR^* \tag{8}$$

for the standards and

$$Y^*_0 = Y_0 + \hat{X}_0 R^* \tag{9}$$

for the unknown, where R^* represents a random drawing from the residual pool. Each bootstrap data set is used to calculate a bootstrap estimate $\hat{X}^*_0$; 1000 such values are then used to calculate the confidence interval by sorting and finding the 5th and 95th percentage points.

Figure 3 shows the resulting histogram for one of the simulated data sets when there is only one replicate of the unknown: the distribution can be bimodal. This occurs when there is an apparent gap in the residuals, so that the bootstrap responses Y^* fall into two distinct groups. This then translates into bimodality in the distribution of $\hat{X}^*_0$. For larger r, the problem disappears and the distribution becomes more symmetrical; it could be remedied for $r = 1$ by smoothing the empirical distribution of the residuals or sampling from a normal approximation. Bonate

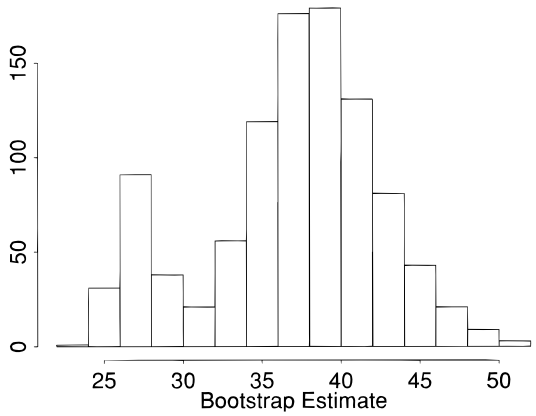

Figure 3. Histogram of 1000 bootstrap estimates with only one replicate of the unknown, showing that the distribution is sometimes bimodal.

Table 1. Effect of Bias Correction, Varying Y_0 and Residual Adjustment on the Coverage Probability (Nominally 90%) of Bootstrap Confidence Intervals (n = 10 000 Simulations)[a]

		bootstrap method			
cv		B	BY	BR	BYR
0.05	bc	0.382	0.718	0.457	0.792
	nbc	0.379	0.734	0.456	0.807
0.10	bc	0.383	0.728	0.455	0.799
	nbc	0.382	0.745	0.451	0.812
0.20	bc	0.380	0.721	0.451	0.791
	nbc	0.378	0.736	0.448	0.806

[a] Figures give the proportion of intervals containing the true concentration. bc, with bias correction; nbc, without. Methods: B, Bonate's method; BY, B + variation in Y_0; BR, B + residual adjustment; BYR, Bonate's method + variation with Y_0 + residual adjustment.

uses bias correction to counteract asymmetry. Our first simulation investigates the effect of this and our proposed improvements in the case $r = 1$ when the true concentration $x_0 = 30$.

The results are given in Table 1. It can be seen that bias correction plays an insignificant role in achieving the proper coverage. The most important correction is to allow for variability in the measured response of the unknown (although this will decline as replication increases). Adjustment of the residuals is also a significant factor. The results are not affected by the cv used, so in future simulations we keep $cv = 0.05$.

As an alternative, we also consider the bootstrap-*t*. Here we use instead of $\hat{X}_0$ the "pivotal" statistic

$$t = (\hat{X}_0 - X_0)/\mathrm{se}(\hat{X}_0) \qquad (10)$$

which is analogous to the usual *t*-statistic in normal theory statistics. Here $\mathrm{se}(\hat{X}_0)$ is the standard error, which can be approximated using the delta method[5] by

$$\mathrm{se}(\hat{X}_0) \simeq \frac{s}{\hat{b}}\sqrt{\frac{\hat{X}_0^{\,2}}{r} + \frac{1}{\sum w_i} + \frac{(\hat{X}_0 - \bar{x}_w)^2}{\mathrm{SS}x_w}} \qquad (11)$$

Bootstrap data sets are generated as for the ordinary percentile bootstrap given above, and each yields a bootstrap-*t* value t^* calculated by replacing X_0 by $\hat{X}_0$ and X_0 by $\hat{X}_0^*$ in eq 10. Similarly, $\mathrm{se}(\hat{X}_0^*)$ is found from eq 11, $\hat{X}_0$, s, and $\hat{b}$ being replaced by their bootstrap values. The 5th and 95th percentage points ($t^*_{0.05}$,$t^*_{0.95}$) are found and the confidence interval (X_L,X_U) calculated as

$$X_L = \hat{X}_0 - t^*_{0.95}\mathrm{se}(\hat{X}_0), \quad X_U = \hat{X}_0 - t^*_{0.05}\mathrm{se}(\hat{X}_0) \qquad (12)$$

Theory suggests[3,4] that these intervals should achieve greater coverage accuracy than the ordinary percentile bootstrap. Efron and Tibshirani[3] recommend the accelerated bias-corrected bootstrap, but it is difficult to see how to implement this with calibration data. Bonate also considers an alternative nonbootstrap method, derived from a naive use of the standard error given above. This seems to have little to recommend it, and Bonate's simulations suggest that its behavior is erratic. We do not consider it further.

We now compare the performance of the standard confidence interval (S), the percentile bootstrap (PB), and the bootstrap-*t* (BT) using the linear model as described above. Bonate's method (B) is also included for comparison. Since S is designed specifically for the case of normally distributed errors, it might be expected to fail when this assumption is incorrect; the bootstrap methods, however, use the observed errors, and so might be expected to work even when the errors are nonnormal. To test this hypothesis, we employed three other error structures: a lognormal distribution, a *t*-distribution with four degrees of freedom, and a mixture of normals in which one in 10 observations is an outlier. Results for 1000 simulations are given in Table 2 for one and three replicates of the unknown sample, using $x_0 = 90$. Other concentrations for the unknown were tried at various points on the calibration line: all gave substantially similar results.

Our results for B agree with Bonate's findings: the coverage is much too low. Using PB, the coverage improves to about 80% but is still short of the target 90%. BT, however, appears to achieve approximately the right coverage, as does S. There is little to choose between these two in terms of coverage and average length. The performance of each method changes little for two of the departures from normality tested; only in the case of *t*-distributed errors are the coverage probabilities seriously affected, and even here S outperforms the other methods. In short, there is no apparent advantage to using the bootstrap in this case because the standard method works adequately and is easier to calculate. Improving robustness to distributional assumptions would require changing the method of estimation,[6] not just the method of assessing precision. Use of a robust estimation method in conjunction with the bootstrap could be a viable approach to this problem, but this is beyond the scope of this paper.

Finally in this section, we repeat the experiment with a larger number of standards, $n = 15$ instead of $n = 6$. The results in Table 3 show that now PB improves and is comparable with BT and S. This is in line with theoretical predictions: the bootstrap-*t* converges more quickly to the target coverage, but for larger samples both are approximately correct.

A NONLINEAR EXAMPLE

If the calibration curve (i.e., the function $f(.)$ of eq 1) is intrinsically nonlinear,[7] exact prediction limits cannot usually be

(5) Stuart, A.; Ord, J. K. *Kendall's Advanced Theory of Statistics*, 5th ed.; Oxford University Press: Oxford, 1987; pp 323−329.

(6) Tiede, J. J.; Pagano, M. *Biometrics* **1979**, *35*, 567−74.

(7) Seber, G. A. F.; Wild, C. J. *Nonlinear Regression*; Wiley: New York, 1989; pp 4−7.

Table 2. Comparison of Confidence Interval Methods with Nine Standards When $x_0 = 90$[a]

		normal			lognormal			t4			mixture		
r	method	*p*	*m*	SD	*p*	*m*	SD	*p*	*m*	SD	*p*	*m*	SD
1	B	0.460	6.13	2.22	0.428	5.86	2.16	0.255	5.54	3.13	0.456	6.39	3.52
	PB	0.829	16.03	5.82	0.807	15.37	5.67	0.607	14.54	8.23	0.809	16.73	9.22
	BT	0.885	20.48	7.72	0.877	19.60	7.38	0.701	18.88	11.93	0.877	21.72	13.07
	S	0.905	21.30	7.86	0.893	20.37	7.64	0.717	19.25	11.51	0.891	22.30	12.84
3	B	0.582	6.00	2.09	0.568	5.88	2.29	0.413	5.59	3.27	0.574	6.03	3.38
	PB	0.817	10.20	2.96	0.802	10.05	3.25	0.721	12.20	5.62	0.804	10.35	4.84
	BT	0.911	13.04	3.72	0.901	12.85	4.05	0.864	16.47	7.54	0.889	13.11	5.84
	S	0.911	13.12	3.75	0.905	12.95	4.10	0.871	17.04	8.40	0.894	13.37	6.23

[a] *r*, number of replicates of unknown; *p*, achieved coverage; *m* mean length of interval; SD, standard deviation of interval length. Methods: B, Bonate's method; PB, percentile bootstrap; BT, bootstrap-*t*; S, standard method.

Table 3. Comparison of Confidence Interval Methods with 15 Standards When $x_0 = 90$[a]

		normal			lognormal			t4			mixture		
r	method	*p*	*m*	SD	*p*	*m*	SD	*p*	*m*	SD	*p*	*m*	SD
1	B	0.330	3.91	0.81	0.327	3.91	0.80	0.211	3.76	1.42	0.342	4.06	1.43
	PB	0.871	15.56	3.37	0.867	15.55	3.35	0.664	15.79	6.81	0.876	17.11	7.14
	BT	0.875	16.21	3.52	0.875	16.13	3.41	0.666	16.28	7.17	0.869	17.64	7.54
	S	0.883	16.42	3.37	0.884	16.43	3.35	0.671	15.76	6.08	0.887	17.03	6.05
3	B	0.493	3.92	0.79	0.491	3.92	0.78	0.289	3.83	1.41	0.518	4.06	1.37
	PB	0.899	9.34	1.76	0.894	9.34	1.75	0.716	11.12	3.99	0.898	9.84	3.11
	BT	0.917	9.95	1.88	0.916	9.95	1.87	0.718	11.42	3.90	0.911	10.41	3.25
	S	0.919	10.00	1.87	0.918	10.00	1.86	0.727	11.52	3.98	0.913	10.48	3.27

[a] *r*, number of replicates of unknown; *p*, achieved coverage; *m*, mean length of interval; SD, standard deviation of interval length. Methods: B, Bonate's method; PB, percentile bootstrap; BT, bootstrap-*t*; S, standard method.

calculated, and we have to rely on a delta method approximation. In such situations, the performance of the standard method of confidence interval construction becomes uncertain. We take as an example the determination of the herbicide atrazine in water samples by enzyme-linked immunoassay (ELISA).

ELISA is one of several forms of immunoassay, itself a version of a more general ligand−receptor interaction analysis. A dose−response curve is generated by the specific interaction of an antibody and its antigen, referred to as the analyte (in our case, atrazine). The antibody is usually immobilized on a solid surface, e.g., a well of a microtiter plate. Since the antibody−antigen binding cannot be observed directly, an enzyme-labeled analogue (tracer) is introduced. This tracer is incubated in the antibody-coated microtiter plate wells, together with the sample containing analyte molecules. According to the law of mass action, both analyte and tracer establish equilibrium binding to the limited number of solid phase antibodies, their ratio being governed by their relative affinities to the antibody. This could also be viewed as an equilibrium distribution of the two species between two phases. After the unbound molecules are washed out, an enzyme substrate is added, which is converted into a colored product. The color intensity is then measured photometrically as an optical density. The higher the initial concentration of analyte in the sample, the fewer tracer molecules are bound and the lower the optical density reading. If no analyte is present in the sample, the antibody binding sites are all occupied by a maximum number of tracer molecules, thereby generating the highest possible signal. There is usually a small amount of binding of the tracer, even in the presence of very high analyte concentrations: this is referred to as nonspecific binding. The typical dose−response curve of optical density plotted against log concentration is thus sigmoidal in shape.

A common method of fitting a calibration curve to such data is the four-parameter logistic model.[8] A detailed account of the fitting, estimation of unknown concentrations and calculation of the standard confidence interval is given by O'Connell et al.[9] Our analysis differs slightly in that we assume a constant coefficient of variation and use a log transformation of the reponses instead of estimating a variance function.[10] Thus, our model is

$$\log Y = \log\left(\frac{A - D}{1 + (x/C)^B} + D\right) + \epsilon \qquad (13)$$

where Y is the assay response, x the analyte concentration, A, B, C, and D the model parameters, and ϵ an error assumed to have a normal distribution with zero mean and variance σ^2.

A specimen calibration curve with 90% prediction limits is shown in Figure 4. The mean response $\bar{Y}_0$ for an unknown sample is used to give an estimated concentration $\hat{X}_0$ and a confidence interval (X_L, X_U), as in the linear case. Bootstrapping can also be carried out as for the linear model; one refinement not previously considered is that we now have several unknown samples for each calibration curve, so each makes a contribution to the residual pool. Starting with 96 observations, comprising 24 calibration standards and 24 unknowns in triplicates, our residual pool will contain 96 residuals, and these are resampled to construct 96 bootstrap observations; each bootstrap data set is used to calculate

(8) Rodbard, D. Mathematics and statistics of ligand assays: An illustrated guide. In *Ligand Assay: Analysis of International Developments on Isotopin and Nonisotopic Immunoassay*; Langan, J., Clapp, J. J., Eds.; Masson: New York, 1981.

(9) O'Connell, M. A.; Belanger, B. A.; Haaland, P. D. *Chemom. Intell. Lab. Syst.* **1992**, *20*, 97−114.

(10) Rocke, D. M.; Jones, G. Submitted to *Technometrics*.

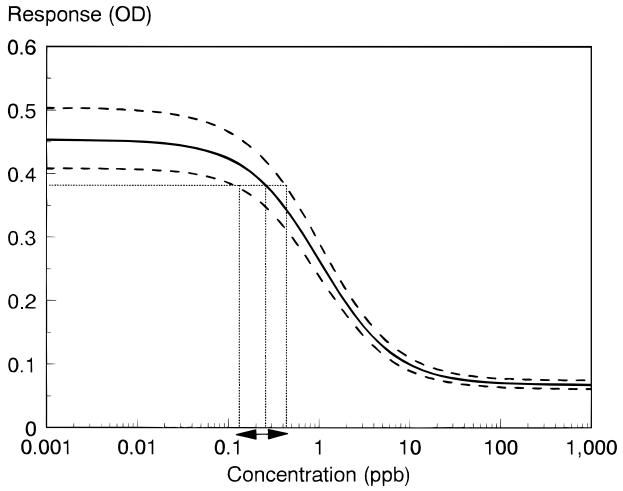

Figure 4. Typical dose−response curve for ELISA, with prediction limits (dashed lines) and confidence interval.

Table 4. Simulation Results (1000 Simulations) for Single-Analyte ELISA with $A = 0.5$, $B = 1.1$, $C = 0.86$, $D = 0.02$, and $\sigma = 0.06$ for "Unknown" Concentration x[a]

x	method	p	m	SD
0.3	S	0.900	0.166	0.030
	PB	0.894	0.159	0.019
	BT	0.896	0.165	0.030
1.0	S	0.899	0.268	0.045
	PB	0.893	0.257	0.027
	BT	0.898	0.272	0.047
	S	0.901	0.591	0.103
	PB	0.896	0.566	0.061
	BT	0.897	0.596	0.106
10.0	S	0.899	2.536	0.500
	PB	0.890	2.421	0.338
	BT	0.898	2.541	0.503

[a] p, achieved coverage; m, mean length of interval; SD, standard deviation of interval length. Methods: PB, percentile bootstrap; BT, bootstrap-t; S, standard method.

the bootstrap estimates $\hat{X}_0^*$ for all the unknown samples simultaneously.

First we consider a simulation, based on a real data set to be examined later, to compare S, PB, and BT. The parameter values $A = 0.5$, $B = 1.1$, $C = 0.86$, $D = 0.02$, and $\sigma = 0.06$ were based on the real data, as were the standard concentrations. The bootstrap-t turns out to be problematic for very small and very large concentrations, because of difficulties in approximating $\text{se}(\hat{X}_0)$, so these were avoided in the simulation. For each unknown concentration, 24 triplicates were simulated per plate for 1000 plates, thus giving 1000 calibration curves but 24 000 estimates. The results are given in Table 4. It can be seen that all three methods achieve approximately the target coverage; S and BT have very similar characteristics, but PB appears to have slightly less coverage with slightly shorter, less variable intervals.

We now turn to a real data set and find a number of new problems to be surmounted. The data were originally produced to examine experimental variation in ELISA curves: a detailed description is given by Jones et al.[11] Four sets of standards, comprising triplicates of 0, 0.1, 0.3, 1, 3, 10, 100, and 10000 ppb, were placed on each of 32 microplates, with pairs of plates being

Table 5. Number of 90% Confidence Intervals Containing the True Concentration from 96 Samples (Expected Number Should Be 86.4)[a]

	confidence interval method					
x	S	PBF	PBR	PBF*	PBR*	BT
0	92	87	85	90	87	*
0.1	90	82	82	90	85	*
0.3	85	79	74	81	81	75
1	85	69	65	83	88	75
3	75	66	63	72	86	66
10	71	58	57	70	74	60
100	69	46	44	53	61	*
10 000	91	90	90	91	92	*
total	85.7%	75.1%	72.9%	82.0%	85.2%	71.9%

[a] Methods: S, standard method; PBF, percentile bootstrap with residuals from fitted model; PBR, percentile bootstrap with residuals from within replicates; BT, bootstrap-t. An asterisk denotes adjustment to achieve compatibility with the standard method.

treated under different experimental conditions. Here we take the first set of standards for our calibration curve estimate and regard the others as "unknowns". There are 32 plates, thus 32 calibration curves, and each gives three determinations each of 0, 0.1, 0.3, 1, 3, 10, 100, and 10 000 ppb. So we have 96 confidence intervals at each concentration level, giving a total of 768 intervals that may or may not contain the true concentration they are estimating. A valid procedure should produce confidence intervals that contain the true concentration 90% of the time, so the expected number of "successes" at each concentration should be $96 \times 0.9 = 86.4$.

The results are shown in Table 5. Some adjustments were made to the bootstrap methodology, which we now describe. First, since all samples, both standards and unknowns, are replicated, this allows the possibility of obtaining all the residuals using the sample means, as opposed to using the fitted model to get residuals for the standards. This is a more symmetrical arrangement and means that the generation of the bootstrap data is model-free: we only need to assume independent errors and constant coefficient of variation. Thus we have two alternatives for the percentile bootstrap: using residuals from the fitted model (PBF—percentile bootstrap with fitted residuals) or using only residuals from within replicates (PBR—percentile bootstrap with replicate residuals). It can be seen from the table that the performance of both is poor compared to that of the standard method (S). To give a concrete example, one of the 1 pbb samples gives a point estimate of 1.36 ppb; the standard confidence interval is (0.94,1.86) which contains the true concentration, whereas the PBF and PBR intervals of (1.10,1.63) and (1.10,1.65), respectively, do not.

One explanation for this is lack-of-fit of the assumed model. It is unlikely that the true response−concentration relationship follows exactly the functional form assumed in the model (even for many supposedly "linear" relationships). This is obviously not a problem in simulations, but for real data it is an important concern. When the standards are replicated, it is possible to test for lack-of-fit by partitioning the variation around the fitted curve and comparing the lack-of-fit mean square (MSLF) to the pure error mean square (MSPE),[12] but even when this test is not

(11) Jones, G.; Wortberg, M.; Kreissig, S. B.; Gee, S. J.; Hammock, B. D.; Rocke, D. M. *Anal. Chim. Acta* **1995**, *313*, 197−207.

(12) Seber, G. A. F.; Wild, C. J. *Nonlinear Regression*; Wiley: New York, 1989; pp 30−32.

statistically significant, lack-of-fit may still exist and distort the results. The standard method as given by O'Connell et al. uses an estimate of σ that includes lack-of-fit since it comes solely from the curve-fitting process; PBR implicitly uses pure replication error excluding lack-of-fit; PBF has an intermediate position. To make a fairer comparison between the methods, we adjusted the residuals to make them compatible with method S by multiplying by the ratio of the estimated standard deviations ($\hat{\sigma}_S/\hat{\sigma}_{PB}$): these adjusted methods are denoted PBF* and PBR*. This can be seen to account for most of the disparity among the methods. The bootstrap-*t* (BT) could not be used for high and low concentrations; this is not necessarily a serious disadvantage because these concentrations were known to be beyond the limits of accurate quantitation. However, BT also performed poorly throughout, possibly because the delta method approximation to the standard error is poor.

All methods showed nonuniformity of coverage across concentrations, with PBR* perhaps the most uniform. This nonuniformity might be indicative of lack-of-fit, or it might be due to spatial effects on the plates.[13] One of the 100 ppb samples was consistently missed, and it was located in one of the corners of the plate, where measurements tend to be less reliable due to possible edge effects.

NONLINEAR MULTIVARIATE CALIBRATION

We use here as an example the analysis of mixtures of the herbicides atrazine and simazine using multianalyte ELISA (MELISA). MELISA uses a panel of antibodies to detect and quantitate mixtures of analytes which cross-react in single-antibody assays by generalizing the four-parameter logistic model.[14] In the case of binary mixtures, we use two suitably chosen antibodies, so that the responses (Y_1,Y_2) from a mixture with concentrations (x_1,x_2) are modeled by

$$\log Y_i = \log\left(\frac{A_i - D_i}{1 + [(x_1/C_{i1})^{B_{i1}/B_i^*} + (x_2/C_{i2})^{B_{i2}/B_i^*}]^{B_i^*}} + D_i\right) + \epsilon_i$$

$$i = 1, 2 \qquad (14)$$

where A_i, B_{ij}, C_{ij}, and D_i are the parameters of the calibration curve for analyte j with antibody i, and $B_i^*$is the geometric mean of B_{i1} and B_{i2}. Two microtiter plates are needed for the assay, each treated with a different antibody. Two single-analyte calibration curves are run on each plate, together with unknown samples. We assume that parameters A and D are common to both curves on the same plate. Estimates of the unknowns x_1 and x_2 for each sample are calculated by solving the system of eq 14 using the measured responses (Y_1,Y_2). Because of this complexity, the standard method of producing confidence intervals for the estimates becomes intractable; implementation of the percentile bootstrap as described above is, however, straightforward: we generate new bootstrap data for each plate separately and then calculate the bootstrap estimates (x_1,x_2).

It is known that mixtures of atrazine and simazine are hard to quantitate accurately by MELISA, since they have similar patterns of cross-reactivity: analysis of 110 samples of 1 ppb atrazine with 1 ppb simazine by Wortberg et al.[15] demonstrated strong correlation between the estimates, so that atrazine concentration might be considerably overestimated with simazine underestimated, or vice versa.

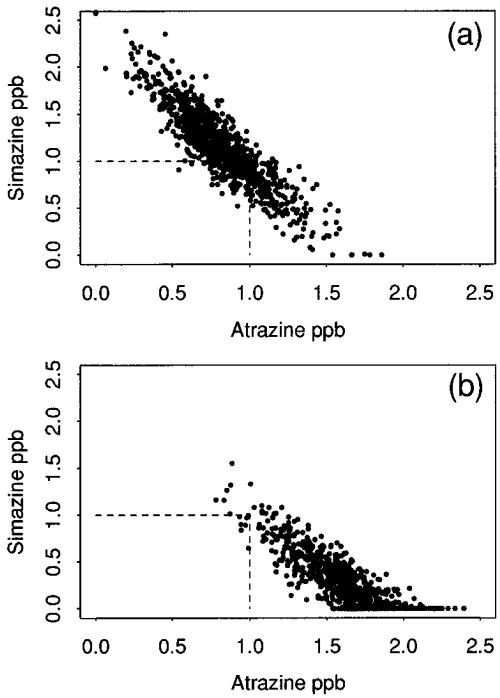

Figure 5. Bootstrap estimates from MELISA of 1 ppb atrazine with 1 ppb simazine. Dashed lines show the position of the true concentration. (a) Point estimate (0.81,1.16). (b) Point estimate (1.71,0.14).

We now show that the bootstrap can provide the same information for single unknown samples without the need for large numbers of replicates. Figure 5ab shows the results of 1000 bootstrap estimates for two of the samples assayed in Wortberg et al.; the correlation and uncertainty in the estimates can be clearly seen. In Figure 5a, the point estimate was (0.81,1.16), which is quite close to the true concentration (1.0,1.0). In Figure 5b, the point estimate is (1.71,0.14), so atrazine is overestimated at the expense of simazine, and many of the bootstrap samples indicate only atrazine. It seems that the total concentration 2 ppb is quite well-estimated, but the assay does not give precise estimation of the relative amounts. Bootstrapping thus gives an easily interpreted account of the information provided by the assay for each unknown sample.

DISCUSSION

Our examples illustrate the difficulties involved in applying statistical methods to real calibration data. In practice, errors may not be normally distributed, or even independent. The postulated model may not be quite correct, or there may be temporal or spatial effects that distort the relationship in going from standards to unknowns. The result is that theoretical "90% confidence intervals" may not contain the true concentrations 90% of the time. The analyst wants narrow intervals to be assured of the precision and accuracy of the assay, but if these intervals do not have adequate coverage probability, this assurance is false.

In particular, we have tried to adjust our procedure for lack-of-fit of the model. Our adjustment, however, was done to make

(13) Shekarchi, I. C.; Sever, J. L.; Lee, Y. J.; Castellano, G.; Madden, D. L. *J. Clin. Microbiol.* **1984**, *19*, 89−96.

(14) Jones, G.; Wortberg, M.; Kreissig, S. B.; Bunch, D. S.; Gee, S. J.; Hammock, B. D.; Rocke, D. M. *J. Immunol. Meth.* **1994**, *177*, 1−7.

(15) Wortberg, M.; Kreissig, S. B.; Jones, G.; Rocke, D. M.; Hammock, B. D. *Anal. Chim. Acta* **1995**, *304*, 339−352.

the bootstrap comparable with the standard method and was in a sense arbitrary, since it depends on the standards used to generate the calibration curve: having four concentrations each replicated five times would give a mean square error with a smaller lack-of-fit component than for 10 concentrations each replicated twice. With given standards and a given decomposition of the mean square error, different combinations of MSLF and MSPE could be used, but there is no obvious defensible choice. Lack-of-fit by its very nature depends on the sample concentration, so perhaps averaging it out across the curve is not appropriate: the resulting confidence intervals might have the correct coverage on average, but the actual coverage would vary for different sample concentrations. Nonparametric estimation of the calibration curve[16] might be the answer in some cases. Our bootstrap methodology could still be applied here without adaptation.

Some theoretical approaches[17] have tried to allow for multiple uses of the same curve. The issue here is that all calibrated values taken from a single estimated calibration curve are correlated: if that curve happens to be "bad", all the readings taken from it will be affected. These theoretical approaches tend to be complicated and conservative, producing statements such as: "at least 90% of the curves will produce confidence intervals which in the long run will contain the true concentration at least 90% of the time". It can be noted that the bootstrap methodology advocated here will reproduce, in the bootstrap estimates, the correlation between all the sample estimates taken from a single curve; this correlation can then be examined by anyone interested in doing so. Furthermore, the analysis applies to the actual number of samples being assayed at the actual estimated concentrations obtained, and not to some theoretical infinite series of unknown samples at unknown places on the curve.

(16) Knafl, G.; Spiegelman, C.; Sacks, J.; Ylvisaker, D. *Technometrics* **1984**, *26* (3), 233−241.
(17) Scheffé, H. *Ann. Stat.* **1973**, *1* (1), 1−37.

CONCLUSION

In simple calibration situations where prediction limits for the reponse are easily derived, the standard method of inverting these to get confidence intervals works well even under slight departures from normality. Bootstrapping can also work adequately but is more computationally intensive. In more complex situations where the standard method is difficult or intractable, the bootstrap is a useful tool. In applying the bootstrap to calibration data, it is important to allow the reponses to vary for both standards and unknowns and to appropriately adjust the residuals.

ACKNOWLEDGMENT

This research was funded by NIEHS Superfund 2P42-ES04699, NSF 94-06193, NSF 95-10511, U.S. EPA CR 819047, USDA Forest Service NAPIAP R8-27, Center for Ecological Health Research CR 819658, NIEHS Center for Environmental Health Sciences 1P30-ES05707, and Water Resources Center Grant No. W-840. B.D.H. is a Burroughs Wellcome Toxicology Scholar, and M.W. is a fellow of the Deutsche Forschungsgemeinschaft.

Received for review September 29, 1995. Accepted December 5, 1995.[⊗]

AC950985G

[⊗] Abstract published in *Advance ACS Abstracts,* January 15, 1996.