Bootstrapping in Controlled Calibration Experiments

Geoffrey Jones and David M. Rocke

Center for Image Processing and Integrated Computing
University of California
Davis, CA 95616
(*g.jones@mussey.ac.nz*)
(*dmrocke@ucdavis.edu*)

We consider the determination of an unknown quantity—for example, the concentration of a particular chemical in a given sample or samples—using controlled calibration. Here several samples are prepared with concentrations chosen to cover a required range, and these are used to establish the relationship between concentration and the measured response to an assay method. This relationship is then used to estimate the concentration in the unknown samples from their measured responses. Confidence intervals for the estimated concentrations can usually be calculated by inverting a prediction interval, but in some situations this method becomes intractable. We explore the use of the bootstrap as an alternative in linear, nonlinear, and multivariate controlled calibration, using both simulation and real datasets from the field of immunoassay. We also discuss the alternatives afforded by replication of the design points. The bootstrap is found to be comparable to the standard method in simple situations and is easy to apply even in complex situations in which standard approaches perform poorly or are intractable.

KEY WORDS: Confidence intervals; ELISA; Immunoassay; Inverse estimation; Nonlinear multivariate regression; Replication.

We are concerned with the application of an assay system that, given a sample with concentration x, produces a response Y whose relationship to x has the form

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

where f is a function, assumed known, describing the relationship (perhaps a scientific or empirical law), θ is a vector of unknown parameters, and ε is a term representing experimental error, which might be assumed to follow some known distribution. This includes, for example, the simple linear model

$$Y = a + bx + \varepsilon, \tag{2}$$

in which a and b are the unknown intercept and slope parameters and ε is assumed to follow a normal distribution with constant variance and zero mean. In Section 4 we discuss a multivariate extension in which both x and Y are vectors and $f(.,.)$ is nonlinear.

The data for such an experiment consist of two parts, "standards" and "unknowns." The standards are prepared samples having known concentrations carefully chosen by the experimenter to cover a required range of x values. We assume here that the preparations are without error—that is, that these x values are exact [but see Racine-Poon, Weihs, and Smith (1991) for the case of dilution errors]. Application of the assay procedure now yields data (x_i, Y_i) for $i = 1, \dots, n$. Often the standards are replicated so that some of the x_i are equal, to enable examination of the appropriateness of the chosen $f(.,.)$.

The data for the unknowns consist of observed Y values only, from which we attempt to estimate their unknown concentrations. We consider first the case of one unknown sample with concentration x_0, giving a response Y_0. In many cases this would also be replicated and we would have several responses $Y_{01}, \dots, Y_{0r}$.

The classical method of estimating x_0 is to first use the standards to estimate the parameter θ by applying an appropriate regression technique. This gives the so-called calibration curve

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

where $\hat{\theta}$ is the regression estimate of the parameters. Given the response Y_0 from a sample with unknown concentration x_0 (or an appropriate mean response if replication is used), then, provided that $f(.,.)$ is monotonic and Y_0 lies in the range of $f(.,.)$, we can always invert the calibration curve to produce an estimate $\hat{X}_0$ for x_0. For example, in the simple linear model [Eq. (2)] we get

$$\hat{X}_0 = \frac{Y_0 - \hat{a}}{\hat{b}}. \tag{4}$$

There are other ways of obtaining an estimate. Krutchkoff (1967) advocated regressing x on Y to get a prediction equation for x_0; Brown (1982) suggested that this method is justifiable in the linear case even though x is not random, because the resulting estimator is Bayesian with respect to a particular prior distribution on x. This approach may not be appropriate for nonlinear calibration, however, especially when, as is the case in our examples, $f(.,.)$ has horizontal asymptotes. Another alternative is the maximum likelihood estimator (MLE) of x_0, possibly derived through the profile likelihood as in the work of Brown and Sundberg (1987). The MLE is identical to the classical estimator except in the case of multivariate calibration when the dimension of the vector response y is greater than that of the unknown x. True maximum likelihood estimation then becomes problematic if many unknowns are calibrated from a single set of standards, as is the case in our examples, because all unknowns must be estimated simultaneously. We are concerned here mainly with methods of producing bootstrap datasets in the controlled calibration setting: Once these datasets have been generated, any chosen estimator can be applied to them. Thus, we use the classical estimator $\hat{X}_0$, as given previously, in our investigation.

The standard method of producing a confidence interval for x_0 is due to Fieller (1954). The regression procedure that produces the calibration curve can also be used to calculate prediction limits $[Y_L(x), Y_U(x)]$, which when inverted give the required confidence region $\{x : y_0 \in [Y_L, Y_U]\}$. This method will work in the univariate case provided that the slope of the calibration line is sufficiently large relative to statistical uncertainty, a condition that has implications for our simulation study as noted in Section 2. The situ-

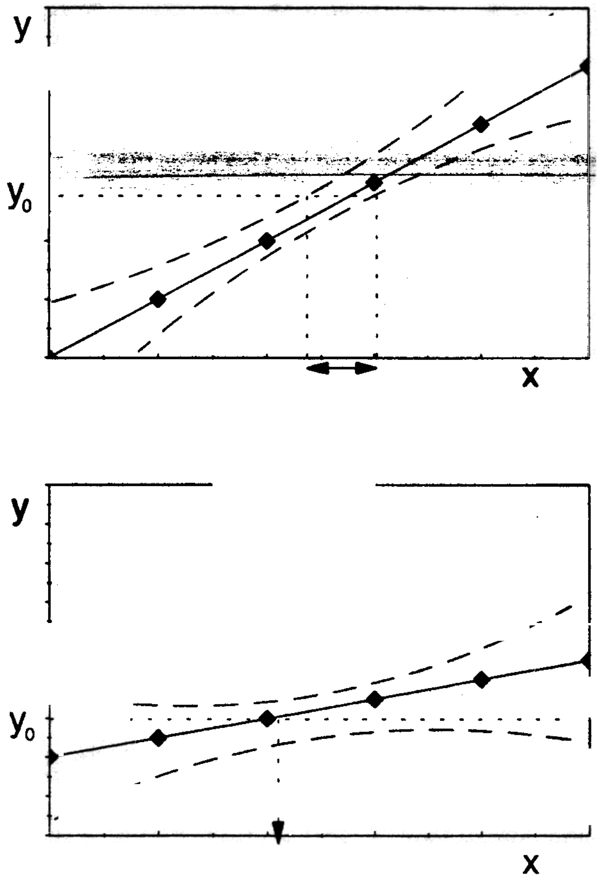

Figure 1. The Standard Method of Producing a Confidence Interval Given a Response Y_0, Using the Prediction Limits From Regression (dashed lines). In (a), the slope of the calibration curve is sufficient to produce an interval. The method fails in (b) when the slope of the calibration curve is too shallow relative to the statistical uncertainty.

from this pool to create the bootstrap datasets. Details are given in Section 2.

The second aspect to consider is the way in which the bootstrap data are used. The simplest, the percentile bootstrap, involves simply calculating the estimate $\hat{X}_0^*$ for each bootstrap dataset and then taking the appropriate percentiles of this distribution. Hall (1992) suggested that this is equivalent to "looking up the wrong table backwards" (p. 36) and that a better approach is to use some form of pivotal statistic, a bootstrap t. Gruet and Jolivet (1993) examined two methods of doing this, the usual pivot based on the asymptotic distribution of $\hat{X}_0^*$ and the predictive pivot based on that of $Y_0 - f(x_0, \hat{\theta})$. The two are easily seen to be exactly equivalent in univariate linear calibration. For the univariate nonlinear case, we investigate both alternatives, together with the naive percentile method. We find that pivoting is indeed superior for small datasets with few unknowns and is probably better than the standard, nonbootstrap method when the error distribution is nonnormal. With a moderate number of replicated unknowns, however, the percentile bootstrap can be surprisingly competitive, provided that the residuals are adjusted before resampling in a manner described later. This result is particularly useful in nonlinear multivariate calibration in which the asymptotic approaches do not appear to be very reliable and pivoting becomes mathematically or computationally intractable.

We first discuss the method of generating of bootstrap datasets in controlled calibration. Then we examine the use of our proposed methods in the production of confidence intervals for calibration in the simple linear case [Eq. (2)], using simulation to compare with the standard intervals as given previously. Next we look at an example of nonlinear calibration, using both simulation and real data. Finally we consider, using a real dataset, a difficult nonlinear multivariate problem in which the simple percentile bootstrap outperforms the asymptotic methods.

1. BOOTSTRAP DATASETS

Efron (1979) illustrated the use of the bootstrap method for setting confidence limits. The bootstrap examines the variability of an estimate by using the existing data, together with some assumptions about how they were generated, to produce new, but plausible, "pseudodatasets" by the process of resampling. In controlled calibration the structure of the data allows several different methods of resampling, as noted previously. Our proposal is to use both parts of the data, standards and unknowns, to create a "residual pool," then to use resampling from this pool to create our bootstrap datasets. In the case of the standards, the regression used to estimate the calibration curve also provides residuals $Y_i - f(x_i, \hat{\theta})$, which are placed in the pool. For an unknown, provided that more than one replicate exists, we take as residuals the deviation $Y_{0i} - \bar{Y}_0$ of each replicate from the mean response. If the unknown is not replicated, it will not contribute to the pool, although it will receive from the pool when the bootstrap data are produced. We discuss later a technical issue concerning this case.

Bootstrap responses, Y^* for a standard and Y_0^* for an unknown, are then given by

$$Y^* = f(x, \hat{\theta}) + R^* \tag{5}$$

and

$$Y_0^* = \bar{Y}_0 + R^*, \tag{6}$$

where R^* represents random drawings with replacement from the residual pool.

We also suggest adjusting the residuals as described, for example, by Efron and Tibshirani (1993, p. 122). 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,

$$\sqrt{\frac{n}{n-p}} \tag{7}$$

where n is the number of points and p the number of parameters, adjusts the residuals to allow for this known undervariability. This adjustment factor may differ between standards and unknowns and between unknowns having different numbers of replicates; not to use it would create imbalance in the resampling plan. One could go further and adjust each regression residual by its standard error, but this would greatly increase the complexity, therefore the time, of the analysis, and simulations suggest that it does not significantly change the properties of the bootstrap intervals. In the case of nonlinear regression, the residuals may not add to 0: In this case it is necessary (Freedman 1981) to center the regression residuals before proceeding.

This method assumes that the mean response is modeled correctly by $f(x, \theta)$ and that the variance of the error terms is constant. If we require a transformation, known a priori, to achieve constant variance, we simply work with the transformed data and model; if on the other hand the variance is a known function of x, as in Bonate's (1993) example, then the residuals can be adjusted appropriately using the known x values for the standards and the estimated $\hat{X}_0$ for the unknown. For example if the standard error of the response Y is thought to be proportional to the concentration x we would use weights $1/x^2$ in the regression and obtain residuals

$$R_i = \frac{Y_i - \hat{Y}_i}{x} \tag{8}$$

from the standards and

$$R_j = \frac{Y_{0j} - \bar{Y}_0}{\hat{X}_0} \tag{9}$$

from the unknowns. A more complicated situation arises if the variance-stabilizing transformation is estimated from the data (as in the Box–Cox approach) or if the variance function contains parameters to be estimated. The effect of estimating these transformation or weighting parameters on prediction intervals was noted by Carroll and Ruppert (1991). They suggested a bootstrap adjustment to the usual intervals. The effect on calibration confidence intervals was shown by Zeng and Davidian (1997), with a similar proposal for bootstrap adjustment. An approach to the bootstrapping of heteroscedastic data, using the "wild bootstrap," that does not require estimation of a variance function was given by Härdle and Mammen (1991). For further details, see Mammen (1992, pp. 13–17).

If all the standard concentrations are replicated, we also have the possibility of using residuals from each replicate set, calculated as for the unknowns, instead of the regression residuals. New responses for the standards are then obtained by adding a resampled residual onto the mean of the replicates, as was done for the unknowns. This simpler, more symmetrical arrangement does not make the assumption that the model used in the analysis is in fact correct: We need only to assume independence and a known variance structure for the errors. An additional advantage is that the residuals are automatically centered at 0.

2. LINEAR CALIBRATION

We now investigate the characteristics of 90% confidence intervals derived from our proposed bootstrapping method in the case of the simple linear model. To this end, we simulated from the model in Equation (2) with $a = 0, b = 5$ and with errors ε having zero mean and standard deviation $\sigma = 1$. Note that fixing the values of a and $b(\neq 0)$ is without loss of generality because of equivariance, but the value of b/σ affects the accuracy of the calibration. Nine calibration standards were used comprising concentrations of 1, 1, 1, 3, 3, 3, 5, 5, 5. The calibration line $\hat{a} + \hat{b}x$ was estimated using ordinary least squares regression. Because the standard method fails if the estimated slope is too shallow, it would be necessary to reject any datasets for which $\hat{b}$ was not significantly nonzero (5% two-tailed test). The chosen design and parameter values ensure that this is extremely unlikely.

The prediction limits are given by

$$\hat{a} + \hat{b}x \pm ts\sqrt{\frac{1}{r} + \frac{1}{n} + \frac{(x - \bar{x})^2}{\mathrm{SS}x}}, \tag{10}$$

where $\bar{x}$ is the mean and SSx the sum of squares of the n standard concentrations, r is the number of replicates of the unknown, 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 squared error from the calibration curve estimation, it has $n - 2$ df; a better approach is to combine estimates from this and from the replicates of the unknown, giving $n + r - 3$ df. 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 $\bar{y}_0$ is calculated by finding the values of x that make either prediction limit equal to $\bar{y}_0$. On rearranging, this gives quadratic equations for the lower and upper limits, which are easily solved.

For the bootstrap, ordinary least squares regression of the standards data gives nine regression residuals $Y_i - \hat{a} - \hat{b}x_i$, which are multiplied by the adjustment factor (here $\sqrt{9/7}$) and placed in the residual pool. The unknown has three replicates and so contributes three residuals $Y_{0j} - \bar{Y}_0$, each of which is multiplied by $\sqrt{3/2}$. We then create our bootstrap dataset of responses,

$$Y_I^* = \hat{a} + \hat{b}x_i + R_i^*, \qquad i = 1, \ldots, 9, \tag{11}$$

for the standards and

$$Y_{0j}^* = \bar{Y}_0 + R_j^*, \qquad j = 1, \ldots, 3, \tag{12}$$

for the unknown, where R_i^* and R_j^* represent random drawings with replacement from the residual pool. An alternative procedure, given that here the standard concentrations are themselves in triplicates, would be to treat the standards in the same way as the unknowns, using only their respective means to get the residuals and to calculate their bootstrap responses. This second method of using only residuals from within replicates is "model-free" in the sense that it does not use the assumption of linearity or the parameters of the fitted model.

Once a bootstrap dataset has been produced, we estimate the bootstrap calibration line parameters $\hat{a}^*$ and $\hat{b}^*$ from the bootstrap standards and hence the bootstrap estimate $X_0^* = (\bar{Y}_0^* - \hat{a}^*)/\hat{b}^*$. For the ordinary percentile bootstrap, 1,000 such values can be used to produce a 90% confidence interval for x_0 by sorting and finding the 5th and 95th percentage points. As an alternative, we also consider the bootstrap t. Here we use, instead of $\hat{X}_0$, the asymptotic "pivotal" statistic

$$t = \frac{\hat{X}_0 - x_0}{\mathrm{se}(\hat{X}_0)}, \tag{13}$$

which is analogous to the usual t statistic in normal theory statistics. We require se$(\hat{X}_0)$, the standard error of $\hat{X}_0$, which unfortunately in the present case does not exist because $\hat{X}_0$ has infinite variance. It does, however, have a finite asymptotic variance given by the delta method (Stuart and Ord 1987, pp. 323–329) as

$$\widehat{\mathrm{se}}(\hat{X}_0) \simeq \frac{s}{\hat{b}}\sqrt{\frac{1}{r} + \frac{1}{n} + \frac{(\hat{X}_0 - \bar{x})^2}{SSx}} \tag{14}$$

Bootstrap datasets are generated as previously, each yielding a bootstrap-t value t^*. The 5th and 95th percentage points $(t^*_{.05}, t^*_{.95})$ are found and the confidence interval (X_L, X_U) calculated as

$$X_L = \hat{X}_0 - t^*_{.95}\widehat{\mathrm{se}}(\hat{X}_0), \quad X_U = \hat{X}_0 - t^*_{.05}\widehat{\mathrm{se}}(\hat{X}_0). \tag{15}$$

Theory suggests (Hall 1992) that these intervals should achieve greater coverage accuracy than the ordinary percentile bootstrap, provided that the scale parameter se$(\hat{X}_0)$ can be well estimated.

The predictive pivot is based on the statistic

$$t_p = \frac{\bar{Y}_0^* - f(x_0, \hat{\theta})}{\widehat{\mathrm{se}}(\bar{Y}_0^* - f(x_0))}. \tag{16}$$

The presence of x_0 in the denominator makes the preceding form very difficult to use. Gruet and Jolivet (1993) suggested substituting the estimate $\hat{X}_0$, in which case

$$t_p = \frac{\bar{Y}_0 - \hat{a} - \hat{b}x_0}{s\sqrt{\frac{1}{r} + \frac{1}{n} + (\hat{X}_0 - \bar{x})^2/SSx}} = t, \tag{17}$$

so that in the linear case the predictive pivot gives the same result as the usual pivot.

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 previously. Results for 10,000 simulations are given in Table 1 for one and three replicates of a single unknown sample, for various values of the true concentration x_0. The estimated coverage probability (p) is the proportion of intervals containing the true concentration. If the actual coverage is approximately equal to the nominal coverage (here 90%), then the estimates of p have a standard error of approximately .003. PB can be seen to produce confidence intervals with inadequate coverage, the actual coverage being about 85%–86%. BT, however, appears to perform much better, with coverage a little lower than S but slightly shorter intervals.

If we adopt the alternative method of producing bootstrap datasets, using only residuals from within replicates, the percentile bootstrap (PBR) and bootstrap t (BTR) change little in performance. Thus, this simpler procedure seems a reasonable alternative to the use of residuals from the calibration curve in the case in which standards are replicated.

To investigate the effect of adjusting the residuals as described, we repeated the simulation without residual adjustment. The coverage of the percentile bootstrap fell to about .80 for $r = 1$ and .81 for $r = 3$, showing that residual adjustment can significantly improve the performance of the percentile method. The effect on the bootstrap t, however, was negligible. Any scale factor appears in both the numerator and denominator of Equation (13); hence we could double all the residuals and the calculated t would be unchanged. Residual adjustment in this case produces only a balancing out of the contributions of the standards and unknowns.

When we repeated the experiment with a larger number of standards, $n = 15$ instead of $n = 9$, we found that now BT is comparable to S, but although the coverage of PB improves, it is still short of the target. The results are summarized in Figure 2. This is in line with theoretical predictions: The bootstrap t converges more quickly to the target coverage, but for larger samples both will be approximately correct.

Table 1. Comparison of Coverage Probability and Interval Length for Four Methods and Different Error Distributions (n = 10,000 simulations)

		$r = 1$			$r = 3$		
x_0	Method	p	m	sd	p	m	sd
.5	S	.899	.864	.240	.901	.603	.151
	PB	.858	.757	.212	.866	.541	.137
	BT	.889	.846	.240	.898	.599	.150
	PBR	.849	.749	.226	.861	.539	.144
	BTR	.888	.840	.236	.899	.599	.150
1.5	S	.904	.811	.224	.901	.530	.130
	PB	.853	.709	.199	.867	.476	.118
	BT	.884	.782	.220	.899	.526	.129
	PBR	.848	.703	.214	.859	.472	.123
	BTR	.883	.777	.219	.901	.523	.128
2.5	S	.899	.776	.216	.897	.484	.120
		.845	.676	.193	.862	.435	.107
		.872	.740	.211	.894	.480	.117
		.842	.669	.205	.865	.432	.114
		.872	.737	.210	.900	.480	.119

NOTE: p = achieved coverage; m = mean length of interval; sd = standard deviation of interval length; S = standard method; PB = percentile bootstrap; BT = bootstrap t; PBR = PB with replication residuals; BTR = BT with replication residuals.

Because 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, so they might be expected to outperform the standard method when the errors are nonnormal. To test this, we investigated two other error structures, a lognormal distribution (achieved by exponentiating and recentering a standard normal) and a t distribution with 6 df. Results are summarized in Figure 3. With t_6 errors, the coverage is too low even for S when there is only one replicate of the unknown; for two or more replicates, the situation reverts to that of normal errors. The lognormal distribution causes different problems: The coverage becomes too high, but with BT less seriously affected than S. There might thus be some advantage in using the bootstrap t in place of the standard method.

We compared our resampling schemes, in which all residuals are pooled, with the two plans investigated by Gruet and Jolivet (1993)—namely, resampling from the regression residuals only or resampling separately for the regression data and the unknowns. We found that for one unknown with a small number of replicates (3) there was very little difference in performance between the methods. In many applications, however, there are a moderate number of replicated unknowns. With six unknowns, each replicated three times, our pooling methods performed slightly better for the bootstrap t and considerably better for the percentile bootstrap, the coverage being about .88 with pooling and .85 without.

To summarize, the bootstrap t would seem to be a useful method of obtaining confidence intervals in linear regression. There is no need to adjust the residuals, and any of the resampling schemes considered previously would be satisfactory provided there are a reasonable number of standards (e.g., the nine standards used in the simulation). The simpler percentile bootstrap can also be made to perform reasonably well if there are many replicated unknowns, but residual adjustment and pooling should be used.

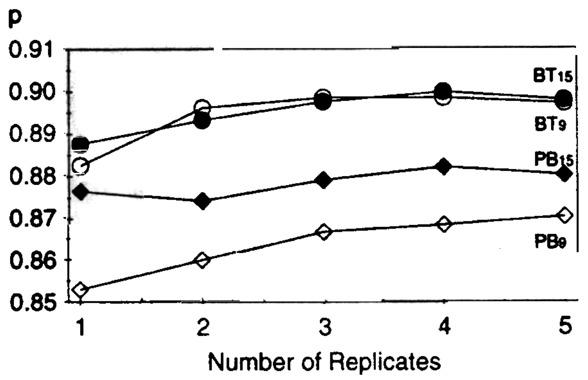

Figure 2. Comparison of Coverage Probabilities Using 9 and 15 Calibration Standards. p = coverage in 10,000 simulations, averaged over six concentrations: PB = percentile bootstrap; BT = bootstrap t.

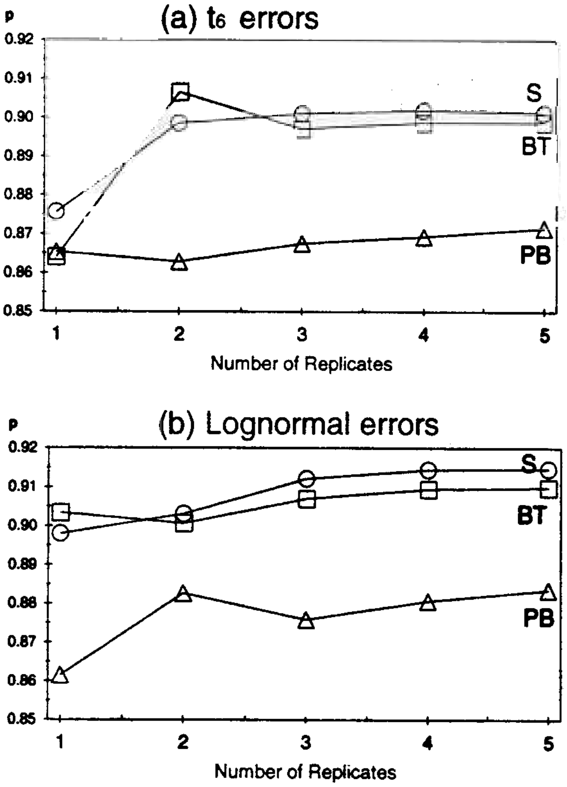

Figure 3. Coverage Probabilities for Nonnormal Errors With Nine Calibration Standards, Using the Standard Method (S), Percentile Bootstrap (PB), and Bootstrap t (BT).

Finally in this section we note a curious phenomenon that occurs when there is only one replicate of the unknown: The distribution of the bootstrap estimates $\hat{X}_0^*$ and t^* can be bimodal. This occurs when there is a "gap" in the residuals derived from the standards so that the bootstrap unknown responses Y_0^* divide into two distinct groups. The problem disappears when there are two or more replicates of the unknown; it can be remedied in the case of one unknown by smoothing the empirical distribution of the residuals or by switching to a parametric approach using the appropriate normal distribution.

3. A NONLINEAR EXAMPLE

If the calibration curve [i.e., the function $f(.,.)$ of Eq. (1)] is intrinsically nonlinear (Seber and Wild 1989, pp. 4–7), exact prediction limits cannot usually be calculated and we have to rely on a delta-method approximation. In such situations, the standard method of confidence-interval construction becomes approximate in nature. We now investigate the usefulness of the bootstrap in nonlinear calibration, taking as an example the determination of the herbicide atrazine in water samples by enzyme-linked immunosorbent assay (ELISA).

ELISA is a form of chemical analysis that uses the specific reaction of an antibody to a chosen analyte to produce a response, here an optical density, which depends on the analyte concentration. Often both standards and unknowns are processed together on a 96-well microtiter plate (as in Fig. 4). These plates are blocks of plastic with 96 small depressions formed into them in which the reactions occur. All processing of the 96 observations on a plate occurs simultaneously. The dose-response curve is typically sigmoidal with unknown horizontal asymptotes that have to be estimated. The responses typically show marked heteroscedasticity, the variance increasing with the mean response, although this aspect is ignored in some commercial software packages.

A common method of fitting a calibration curve to such data is the four-parameter logistic model (Rodbard 1981). A detailed account of the fitting, estimation of unknown concentrations, and calculation of the standard confidence interval was given by O'Connell, Belanger, and Haaland (1992). They used pseudolikelihood to estimate the variance function: Our analysis differs slightly in that we assume a constant coefficient of variation and use a log transformation of the responses [Jones et al. (1995) and Rocke and Jones (1997) showed this to be reasonable for these data]. Thus, our model is

$$\log Y = \log\left(\frac{A-D}{1+\left(\frac{x}{C}\right)^B}+D\right)+\varepsilon, \qquad (18)$$

contributes three further residuals to the pool; alternatively we could use only within-replicate residuals, treating the

Standards	Unknowns		
zero	1	9	17
0.1	2	10	18
0.3	3	11	19
1.0	4	12	20
3.0	5	13	21
10	6	14	22
100	7	15	23
blank	8	16	24

Figure 4. Typical ELISA Template, With 24 Unknown Samples in Triplicates.

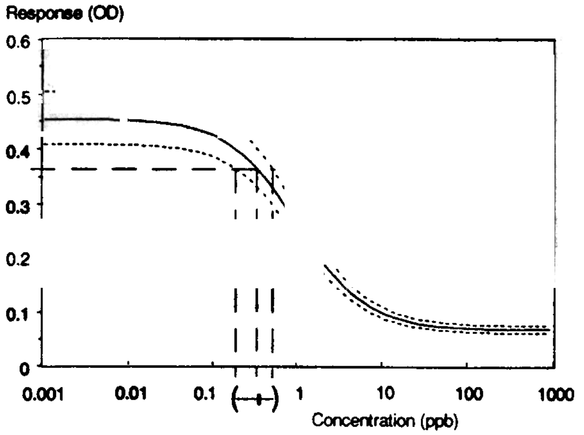

Figure 5. A Typical Dose-Response Curve for ELISA, With Prediction Limits (dashed lines) and Confidence Interval.

standards in the same way as the unknowns. There are now 96 residuals, and these are resampled to construct the 96 bootstrap observations. Analysis of the bootstrap datasets now yields a bootstrap estimate $\hat{X}_0^*$, a bootstrap-t value t^*, or a predictive pivot value t_p^* for each of the 24 unknowns.

First, we consider a simulation, based on a real dataset, to be examined later, to compare the standard method (S), the ordinary percentile bootstrap (PB or PBR if within-replicate-only residuals), the bootstrap t (BT or BTR), and the predictive pivot bootstrap (PP or PPR). The parameter values $A = .5, B = 1.1, C = .86, D = .02$, and $\sigma = .06$ used in the simulation were based on the real data, as were the standard concentrations. Responses for standards and unknowns were simulated using Equation (16). The concentrations used for the unknowns 0, .1, .3, 1, 3, 10, 100, and 10,000 ppb (parts per billion). We expected problems with the very small and very large concentrations because $\text{se}(X_0^*)$ is unbounded as the concentration increases and its asymptotic approximation breaks down as the concentration tends to 0. Furthermore, $x_0 = 0$ is at the boundary of the parameter space, so the usual asymptotic theory is not assured. We find that all methods overcover at $x_0 = 0$; perhaps one-sided intervals would be more appropriate and better behaved here. Two important points should be noted: First, it is important to include 0 and very high concentrations in the calibration dataset because these are what determine the asymptote parameters A and D; second, concentrations near 0 and at high levels are very poorly determined anyway and would not ordinarily be used in a quantitative way.

The results are given in Table 2. Use of within-replicate-only residuals did not significantly affect the performance of any of the bootstrap methods, so these results are not shown. It can be seen that, for the middle range of concentrations .3–10 ppb, all four methods achieve approximately the target coverage of 90% but with the bootstrap intervals slightly shorter and much less variable in length than the standard method. Interestingly, the percentile bootstrap outperformed both pivotal methods, perhaps because of the inaccuracy of the delta-method approximations used. It would seem that the expected advantage of pivoting can be lost by not pivoting well. Gruet and Jolivet (1993) used Edgeworth expansions in powers of r to suggest the superiority of the predictive pivot over the ordinary bootstrap that this is not borne out in practice here implies that suc expansions are of limited use when r is small because tl size of the function derivatives also comes into play. F(the large and small concentrations, however, the ordinai bootstrap t fails when either $\hat{X}_0$ or $\hat{X}_0^*$ becomes 0 or inf nite (or nearly so), whereas the predictive pivot can cop with these values. The preference for one or the other (these pivotal methods then depends on the characteristic of the particular $f(.,.)$ used.

We now examine the real dataset on which the simulatio was based and find several new problems to be surmounte(The data were originally produced to examine experimer tal variation in ELISA curves: A detailed description wa given by Jones et al. (1995). Thirty-two microtiter plates us ing the same template of atrazine concentrations were run a various times under different experimental conditions. Th design of the templates consisted of four sets of standard in triplicates at concentrations of 0, .1, .3, 1, 3, 10, 100, an 10,000 ppb. Here we use one set of standards to estimate calibration curve and regard the others as unknowns, giv ing three separate determinations of each of eight unknow concentrations per plate. Thus we have 32 × 3 = 96 confi dence intervals at each concentration level, giving a total o 768 intervals, which may or may not contain the true con

Table 2. Simulation Results (1,000 simulations) for Single-Analyte ELISA With A = .5, B = 1.1, C = .86, D = .02, and σ = .06 for "Unknown" Concentration x

	Method	p	m	
.0	S	.947	.076	.06
	PB	.947	.076	.04
	BT	—	—	—
	PP	.934	.103	.02
.1	S	.897	.145	.03
	PB	.896	.140	.02
	BT	—	—	—
	PP	.911	.137	.02
.3	S	.900	.166	.03
	PB	.895	.160	.01
	BT	.887	.156	.01
	PP	.886	.155	.01
1.0	S	.899	.268	.04
	PB	.894	.257	.02
	BT	.888	.254	.02
	PP	.887	.253	.02
3.0	S	.901	.591	.10
	PB	.896	.566	.06
	BT	.889	.557	.06
	PP	.888	.555	.06
10.0	S	.899	2.536	.50
	PB	.890	2.421	.33
	BT	.883	2.378	.33
	PP	.883	2.382	.34
100.0	S	.904	inf	—
	PB	.897	inf	—
	BT	—	—	—
	PP	.892	inf	—
10,000.0	S	.868	inf	—
	PB	.898	inf	—
	BT	—	—	—
	PP	.889	inf	—

NOTE: p = achieved coverage; m = mean length of interval; sd = standard deviation of interva length; S = standard method; PB = percentile bootstrap; BT = bootstrap t; PP = predictive pivot bootstrap.

centration they are estimating. A valid procedure should produce confidence intervals that contain the true concentration 90% of the time, so the number of "successes" at each concentration should have a mean of 96 × .9 = 86.4.

The original results, using the template shown in Figure 4, gave very poor coverage (about 75%) for the standard and the bootstrap methods. All methods showed nonuniformity of coverage across concentrations, which might be an indication of lack-of-fit of the model. An alternative explanation is the presence of spatial effects on the plates. It is well known to practitioners that spatial effects can sometimes develop, especially in certain locations such as the edges or corners of a microtiter plate. They are variously attributed to temperature gradients, dilution errors, inhomogeneity of the plate material, bias in the plate-reading device, and others. Because the same template was used for each plate, some of the samples could be expected to be affected more than others. One of the 100 ppb samples was consistently missed, and it was located in one of the corners of each plate. To investigate the effect of spatial correlation, we rearranged the template by randomly grouping the 12 replicates at each concentration level into a calibration triplicate and three triplicated unknowns. The effect on coverage performance was dramatic, as shown in Table 3. Most concentrations now achieved greater than nominal coverage, with the shortfall at the 100 and 10,000 ppb concentrations being due apparently to their corner positions. The coverage across all concentrations was essentially at the nominal, with slightly lower coverage at 100 and 10,000 ppb and slightly higher compensating coverage at the other concentrations. The bootstrap t (BT*) could again 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. Both pivotal methods, however, failed to outperform the simpler percentile methods, as in the simulations. Again the use of within-replicate-only residuals made little difference; only PBR is shown here for comparison.

The poor coverage without randomization shows the danger of spatial correlation in microplate data. It is perhaps impractical to expect a technician to pipette each sample replicate in a random position, but not to do so means that the real errors in the estimated concentrations may be very different from what they are often assumed to be. This very serious problem is beyond the scope of the present article and remains for further work.

Table 3. Number of 90% Confidence Intervals Containing the True Concentration from 96 Samples (expected number should be 86.4)

	Confidence interval method				
x	*S*	*PB*	*PBR*	*BT*	*PP*
0	95	94	94	—	94
.1	94	94	93	—	94
.3	96	96	95	96	95
1	91	91	91	91	91
3	93	92	92	91	92
10	93	93	93	91	93
100	68	60	60	—	60
10,000	77	77	77	—	77
Total	92.1%	90.8%	90.!	96.1%	90.6%

NOTE: Methods are as in Tables 1 and 2.

4. NONLINEAR MULTIVARIATE CALIBRATION

Most approaches to multivariate calibration have considered only the linear case or simple extensions that are still linear in the model parameters. Brown (1982) extended Fieller's approach to give confidence regions in multivariate linear calibration. The coverage is exact, but when the dimension of **x** is less than that of **Y** the region may be empty. Several proposals have been made to overcome this (Brown and Sundberg 1987; Oman 1988; Mathew and Kasala 1994; Mathew and Zha 1996). These proposals tend to be very difficult to implement even in the linear cases considered by the authors. Clarke (1992) considered a nonlinear model with multivariate **Y** but univariate x; he needed simulation to derive the distribution of his suggested statistic. Bootstrapping is thus an attractive possibility here; we shall demonstrate that it can be done fairly easily even with complex nonlinear models with multivariate **x** and **Y**. For comparison purposes we also investigate a nonlinear multivariate extension of Fieller's approach and the likelihood ratio statistic suggested by Brown and Sundberg (1987).

We use here as an example the analysis of mixtures of the herbicides atrazine and terbutryn 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 (see Jones et al. 1994; Wortberg, Jones, Kreissis, Rocke, and Hammock 1995). 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

$$\begin{aligned}\log Y_i &= f(x, \theta_i) + \varepsilon_i \\ &= \log\left(\frac{A_i - D_i}{1 + \left(\left(\frac{x_1}{C_{i1}}\right)^{B_{i1}/B_i^*} + \left(\frac{x_2}{C_{i2}}\right)^{B_{i2}/B_i^*}\right)^{B_i^*}} + D_i\right) \\ &\quad + \varepsilon_i, \qquad i = 1, 2, \end{aligned} \tag{19}$$

where A_i, B_{ij}, C_{ij}, 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. The two plates are analyzed separately, so ε_1 and ε_2 are independent. 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 equations (17) using the measured responses (Y_1, Y_2). Because of this complexity the standard asymptotic methods of producing confidence intervals for the estimates are mathematically and computationally difficult; implementation of the percentile bootstrap as described previously is, however, straightforward: We generate new bootstrap data for each plate separately and then calculate the bootstrap estimates (x_1, x_2).

To produce a confidence region from many bootstrap point estimates is more problematic now than in the one-dimensional case. One alternative is to assume multivariate normality of the estimator, estimate the mean and covariance matrix, and draw the appropriate elliptical contour. An alternative, nonparametric approach is to estimate the multivariate density. We used ASH, or average shifted histograms (Scott 1992), to estimate the bivariate density, then drew a contour at a level such that the integrated ASH estimate inside the contour was 90%. Figure 6 shows the results of 1,000 bootstrap estimates for one of the unknowns together with the ASH-derived confidence region (B).

The Fieller approach may be regarded as an adjustment of the squared distance function $\|\mathbf{Y}_0 - \mathbf{f}(\mathbf{x}_0, \hat{\boldsymbol{\theta}})\|^2$ to account for uncertainty in the parameter estimate $\hat{\boldsymbol{\theta}}$. It was argued by Jones (1996) that approximately

$$D(\mathbf{x}_0) \equiv \sum_{i=1}^{2} \frac{(\log Y_{0i} - f(\mathbf{x}_0, \hat{\boldsymbol{\theta}}))^2}{\sigma_i^2(1+v_i)} \sim \chi_2^2, \qquad (20)$$

where σ^2 is the variance of ε_i and $\sigma_i^2 v_i$ the asymptotic variance of $f(\mathbf{x}_0, \hat{\boldsymbol{\theta}})$, obtained via the delta method. Because $\mathbf{x}$ and $\mathbf{Y}$ have the same dimension here, then, provided that we are away from the boundary (x_{01} and x_{02} both nonzero), $\hat{x}_0$ solves $D(\mathbf{x}_0) = 0$. A confidence region can then be obtained by evaluating $D(\boldsymbol{x}_0)$ over a grid and contouring at the appropriate level. Such a region is shown in Figure 6, marked F.

The likelihood ratio approach is in principle straightforward, provided that we deal with one unknown at a time. For a trial value x_0, the point (x_0, Y_{0i}) (or points, if replicated) is added to the ith calibration set; estimation of the curve parameters by least squares regression leads to an error sum of squares SS_i. Assuming normality, the log-likelihood, to within an added constant, is given by

$$l(\mathbf{x}_0) = -\frac{n_1}{2}\log\text{SS}_1 - \frac{n_2}{2}\log\text{SS}_2,$$

where n_1 and n_2 give the total number of data points i each augmented calibration set. We now find the minimiz $\check{\mathbf{x}}_0$ of $l(\mathbf{x}_0)$ and assume the usual asymptotic result

$$2(l(x_0) - l(\check{\mathbf{x}}_0)) \sim \chi_2^2 \qquad (2$$

to get a confidence region. In practice, however, there a considerable computational difficulties both in minimizin $l(\mathbf{x}_0)$ and in evaluating it over a grid where the convergenc of the curve-fitting algorithm may break down for some $\mathbf{x}$ The resulting confidence region is marked L in Figure 6.

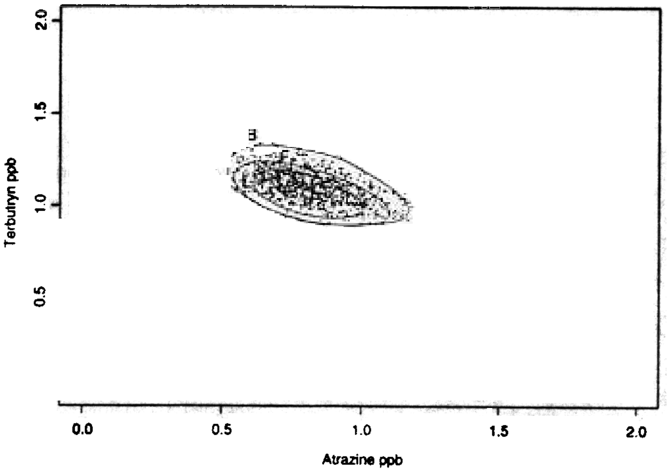

Figure 6. Bootstrap Estimates and Estimated 90% Confidence Regions From MELISA of 1 ppb Atrazine With 1 ppb Terbutryn: B = Bootstrap Method; F = Fieller Method; L = Likelihood Ratio Method.

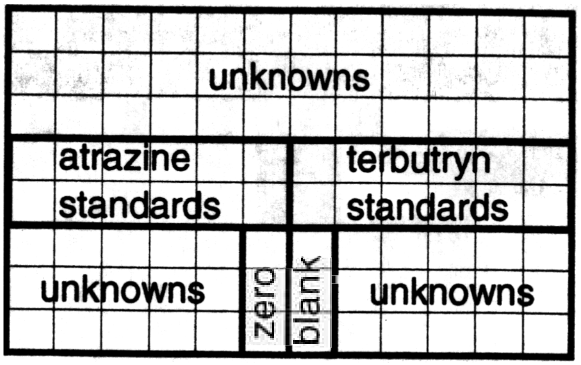

Figure 7. Template for MELISA With Two Analytes. Two such plat are used, each treated with a different antibody.

Our full dataset for this example consisted of six pai of plates, each containing duplicated standards for bot atrazine and terbutryn, triplicated zeros and blanks, and 2 triplicated "unknown" samples of 1 ppb atrazine with 1 pp terbutryn (see Fig. 7). This gave a total of 132 determina tions of the 1 + 1 mixture, although results from the sam pair of plates are not independent because they use the sam estimated standard curves. To reduce the problem of spa tial effects noted earlier, we randomized the positions of th unknowns, separately for each plate. The number of con fidence regions containing the true concentrations for eac pair of plates by each method is shown in Table 4. Th two asymptotic methods (F and L) do not perform wel the bootstrap confidence regions, however, would seem t provide a reasonable summary of the uncertainty in eac estimate.

Table 4. Coverage of Confidence Regions in Multianalyte ELISA Experiment

	Confidence interval method		
Plates	Fieller	L. ratio	Bootstrap
	14		
	8		
	7		
	14		
	20		
	19		
	74.5%		

NOTE: Figures show the number of regions (out of 22 samples on each pair of plates) containir the true concentration. The expected number at the nominal 90% coverage is 19.8.

5. CONCLUSION AND DISCUSSION

Our results suggest that bootstrapping can be made to work reasonably well in controlled calibration experiments even when the sample size is not large. Pivoting may lead to better coverage properties in small datasets, but for larger datasets even the simple percentile bootstrap, with residual adjustment, can approximate the correct coverage; the advantage of pivoting may be lost in nonlinear models in which the standard error must be approximated.

If both standards and unknowns are replicated, the use of within-replicate residuals is simpler and gives results comparable with those obtained using regression residuals. These different approaches correspond to different levels of assumptions made in constructing the bootstrap datasets. The use of regression residuals assumes that the mean response is modeled correctly and that the correct variance function is used: Within-replicate residuals assume only the correct variance function. If the mean response is modeled incorrectly, this will bias the calibration estimates and the calculated confidence intervals from either method will be misleading. In some cases it might be better to estimate the calibration curve nonparametrically (Knafl, Speigelman, Sacks, and Ylvisaker 1984). Our bootstrap methodology could be applied in this case without adaptation.

Another possible failure of assumptions concerns nonindependence of the errors. There may be spatial or temporal effects that cannot be eliminated by randomization of the design, resulting again in inadequate coverage of both standard and bootstrap intervals. In general, false assumptions will tend to give misleadingly reassuring intervals, by whatever method they are produced. The advantage of our bootstrap methodology is that it is easy to apply even in quite complex ~~situations, and it gives results comparable to~~ the standard method in simple ones.

ACKNOWLEDGMENTS

We are grateful to a referee and an associate editor for some important references and suggestions that led to significant improvements in this work. We thank Monika Wortberg for providing the ELISA data and David Scott for providing software for multivariate density estimation. This research was funded by NIEHS Superfund 2P42-ES04699, NSF 93-01344, NSF 94-06193, and NSF 95-10511.

[Received February 1996. Revised March 1999.]

REFERENCES

Bonate, P. L. (1993), "Approximate Confidence Intervals in Calibration Using the Bootstrap," *Analytical Chemistry*, 65, 1367–1372.

Brown, P. J. (1982), "Multivariate Calibration," *Journal of the Royal Statistical Society*, Ser. B, 44, 287–321.

Brown, P. J., and Sundberg, R. (1987), "Confidence and Conflict in Multivariate Calibration," *Journal of the Royal Statistical Society*, Ser. B, 49, 46–57.

Carroll, R. J., and Ruppert, D. (1991), "Prediction and Tolerance Intervals With Transformation and/or Weighting," *Technometrics*, 33, 197–210.

Clarke, G. P. Y. (1992), "Inverse Estimates From a Multiresponse Model," *Biometrics*, 48, 1081–1094.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

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

Fieller, E. C. (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society*, Ser. B, 16, 175–185.

Freedman, D. A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218–1228.

Gruet, M.-A., and Jolivet, E. (1993), "Calibration With a Nonlinear Standard Curve: How to Do It?" *Computational Statistics*, 9, 249–276.

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

Härdle, W., and Mammen, E. (1991), "Bootstrap Methods for Nonparametric Regression," in *Nonparametric Functional Estimation and Related Topics* (Series C: Mathematical and Physical Sciences, Vol. 335), ed. G. Roussas, Amsterdam: Kluwer, pp. 111–124.

Jones, G. (1996), "The Statistics of Multiple Immunoassay," unpublished Ph.D. thesis, University of California at Davis, Division of Statistics.

Jones, G., Wortberg, M., Kreissig, S. B., Bunch, D. S., Gee, S. J., Hammock, B. D., and Rocke, D. M. (1994), "Extension of the Four-Parameter Logistic Model for ELISA to Multianalyte Analysis," *Journal of Immunology Methods*, 177, 1–7.

Jones, G., Wortberg, M., Kreissig, S. B., Gee, S. J., Hammock, B. D., and Rocke, D. M. (1995), "Sources of Experimental Variation in Calibration Curves for Enzyme-Linked Immunosorbent Assay," *Analytica Chimica Acta*, 313, 197–207.

Jones, G., Wortberg, M., Kreissig, S. B., Hammock, B. D., and Rocke, D. M. (1996), "On the Application of the Bootstrap to Calibration Experiments," *Analytical Chemistry*, 68, 763–770.

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

Krutchkoff, R. G. (1967), "Classical and Inverse Regression Methods of Calibration," *Technometrics*, 9, 425–439.

Mammen, E. (1992), *When Does Bootstrap Work? Asymptotic Results and Simulations*, New York: Springer-Verlag.

Mathew, T., and Kasala, S. (1994), "An Exact Confidence Region in Multivariate Calibration," *The Annals of Statistics*, 22, 94–105.

Mathew, T., and Zha, W. (1996), "Conservative Confidence Regions in Multivariate Calibration," *The Annals of Statistics*, 24, 707–725.

Mee, R. W., Eberhardt, K. R., and Reeve, C. P. (1991), "Calibration and Simultaneous Tolerance Intervals for Regression," *Technometrics*, 33, 211–219.

O'Connell, M. A., Belanger, B. A., and Haaland, P. D. (1992), "Calibration and Assay Development Using the Four-Parameter Logistic Model," *Chemometrics and Intelligent Laboratory Systems*, 20, 97–114.

Oman, S. D. (1988), "Confidence Regions in Multivariate Calibration," *The Annals of Statistics*, 16, 174–187.

Racine-Poon, A., Weihs, C., and Smith, A. F. M. (1991), "Estimation of Relative Potency With Sequential Dilution Errors in Radioimmunoassay," *Biometrics*, 47, 1235–1246.

Rocke, D. M., and Jones, G. (1997), "Optimal Design for ELISA and other Forms of Immunoassay," *Technometrics*, 39, 162–170.

Rodbard, D. (1981), "Mathematics and Statistics of Ligand Assays: An Illustrated Guide," in *Ligand Assay: Analysis of International Developments on Isotopin and Nonisotopic Immunoassay*, eds. J. Langan and J. J. Clapp, New York: Masson.

Rosen, O., and Cohen, A. (1995), "Constructing a Bootstrap Confidence Interval for the Unknown Concentration in Radioimmunoassay," *Statistics in Medicine*, 14, 935–952.

Scheffé, H. (1973), "A Statistical Theory of Calibration," *The Annals of Statistics*, 1, 1–37.

Schwenke, J. R., and Milliken, G. A. (1991), "On the Calibration Problem Extended to Nonlinear Models," *Biometrics*, 47, 563–574.

Scott, D. W. (1992), *Multivariate Density Estimation*, New York: Wiley.

Seber, G. A. F., and Wild, C. J. (1989), *Nonlinear Regression*, New York: Wiley.

Stuart, A., and Ord, J. K. (1987), *Kendall's Advanced Theory of Statistics* (5th ed.), Oxford, U.K.: Oxford University Press.

Wortberg, M., Jones, G., Kreissig, S. B., Rocke, D. M., and Hammock, B. D. (1995), "An Immunoarray for the Simultaneous Determination of Multiple Triazine Herbicides," *Analytica Chimica Acta*, 304, 339–352.

Zeng, Q., and Davidian, M. (1997), "Bootstrap Adjusted Calibration Confidence Intervals for Immunoassay," *Journal of the American Statistical Association*, 92, 278–290.