B.L. Davis, S.L. Heisler, J.J. Shah, P.K. Hopke, and D.L. Johnson, Atm Environ **18** (1984) 1555.
[13] Novak, J.H., and D.B. Turner, J. Air Polut. Control Assoc., **26** (1976) 570.
[14] Cheng, M-D., and P.K. Hopke, An Intercomparison of Linear Programming Procedures for Aerosol Mass Apportionment, Air Pollution Control Association Paper No. 85-21.8 (1985).
[15] Frank, I.E., and B.R. Kowalski, Statistical Receptor Models Solved by Partial Least Squares, ACS Divn of Environ Chem Sympos Abstracts (Philadelphia, Aug 1984) p. 202.
[16] Henry, R.C., Proc Air Pollut Control Assoc Spec Conf, **SP-48** (1982) 141.
[17] Lowenthal, D.H.; R.C. Hanumara, K.A. Rahn, and L.A. Currie, Estimates and Uncertainties in Chemical Mass Balance Apportionments: Quail Roost II Revisited, prepared for submission to Atmospheric Environment (1985).
[18] Stevens, R.K., and T.G. Pace, Atmos. Environ. **18** (1984) 1499.

DISCUSSION

of the L.A. Currie paper, The Limitations of Models and Measurements as Revealed Through Chemometric Intercomparison

Leon Jay Gleser
Department of Statistics
Purdue University

The construction and use of Simulation Test Data (STD) to help evaluate alternative chemometric methodologies is a highly welcome contribution to the field. Dr. Currie, and the agencies and colleagues whom he credits, are to be congratulated for an approach which has the potential to promote improvements in the art of quantitative chemical analysis.

What follows is a brief discussion of some previous use of standard data sets in statistical research, along with some warnings about the possible pitfalls connected with the use of such approaches. In particular, the parallel that Dr. Currie draws between the use of standard data sets and interlaboratory comparisons using common reference materials cannot be pushed too far. Many interacting factors lead to bias in modeling and analysis of complex data sets; the contributions of these factors would be confounded in typical interlaboratory comparison designs. One factor, scientific judgment, cannot even be identified in standard frequentist reports of statistical data analysis. This suggests that subjective scientific judgments need to be given more explicit mention in reports of statistical analyses, perhaps through the use of the Bayesian approach to inference.

To make standard data sets more closely resemble real-world data, the use of the "bootstrap" is suggested. The "bootstrap" can also help in providing the estimates of statistical precision that Dr. Currie notes were lacking in the two studies conducted to date.

Standard Data Sets in Statistics

Statisticians have long recognized the usefulness of having common data sets on which new methodologies can be tried out, and their relative merits assessed. For example,

new methodologies for classification and discriminant analysis are often applied to the iris data of E. Anderson [1][1], which was featured in a famous paper by R. A. Fisher [2]. (To Anderson's undoubted frustration, these data are usually referred to as "Fisher's iris data.")

Another famous data set is Longley's [3] econometric linear regression data. In these data, the independent variables in the regression are highly interrelated (multicolinear). Longley ran the data through several computer software packages designed to do least squares analysis. In theory, all of these programs solve the same set of linear equations to estimate the regression slopes. However, the solutions obtained by the various algorithms differed, in some cases even by sign! What had happened was that the multicollinearity in the data made the answers obtained highly sensitive to roundoff and truncation of the data, and the algorithms differed by where and by how much roundoffs were done. Longley's paper had the very beneficial consequence that software developers now pay careful attention to numerical analysis in designing statistical algorithms. Further, it stimulated study of the resistance of statistical methodology to data perturbations (robustness).

However, Longley's paper (and particularly his data) may also have had a less salutory effect on software development. Software developers now know that consumers will test out their programs on Longley's data. [See, for example, Lachenbruch's review [4] of STAN, Version II.0 by David Allen.] This may lead them to overcompensate for multicollinearity problems, and consequently overlook or neglect other potential problems or sacrifice desirable features to include subroutines necessary to accurately process multicollinear data.

This last comment points out a real danger in the use of standard data sets, namely that their existence can bias the direction which development of methodology and software takes. The best guard against such bias is the creation of standard data sets of many types.

An *artificial* standard data set (simulated according to a known model for the distribution of errors) can lead to a particularly serious bias. Chemometricians who know that their work will be evaluated by such data sets will tend to use a methodology which is known to be efficient for the given statistical model. Such a methodology, however, may not do well against *real* data, for which the given statistical model is not necessarily a good approximation. Alternatively, chemometricians may object to evaluations on the basis of such data, arguing (with considerable merit) that such data do not reflect their practical experience.

The reason, of course, for using artificial data is that the "truth" or "signal" underlying the "noise" (error) in the data is known. This allows us to separate *bias* (lack of validity) from *precision* (reliability, repeatability). In this respect there is an obvious parallel, which Dr. Currie correctly points out, with the use of common reference materials in interlaboratory comparisons. The goal of such studies is to eliminate bias (which is usually reflected in interlaboratory variation), and to estimate precision (intralaboratory variation). However, whereas common reference materials are "real" (although they may be ideal examples of materials analyzed in practice), this is not clearly the case with data simulated from specified statistical populations (e.g., Gaussian populations). Real populations may have "heavy tails" and/or other funny features (e.g., several modes) which are not modeled by standard distributions.

One obvious solution is to vary the distributional assumptions which generate the errors in artificial data. This approach is widely used in statistics to study the *robustness* properties of statistical methodologies.

Another possible solution is to use the ideas underlying the "bootstrap" (Efron [5], Diaconis and Efron [6], Freedman and Peters [7,8]) to simulate data which have "real world" error distributions.

The Bootstrap

In using the "bootstrap," we start by assuming that the observed data y_i are related to unknown parameters θ and errors e_i by a model

$$y_i = G(\theta, e_i), \qquad i = 1,2,\ldots,n, \tag{1}$$

where $G(\cdot,\cdot)$ is known. Given a value for θ (which may be a vector), we assume that the eq (1) can be inverted to obtain the errors e_i. That is,

$$e_i = H_i(\theta, y_1,\ldots,y_n), \qquad i = 1,\ldots,n. \tag{2}$$

Given "real" data $y_1, y_2,\ldots,y_n$, with n sufficiently large to give us some hope of accurately estimating θ by

$$\hat{\theta} = \hat{\theta}(y_1,\ldots,y_n),$$

we now construct the residuals (estimated errors)

$$\hat{e}_i = \mathrm{H_i}(\hat{\theta}, y_1,\ldots,y_n), \qquad i = 1,2,\ldots,n. \tag{3}$$

The resulting finite population $\{\hat{e}_1,\ldots,\hat{e}_n\}$ of residuals is the statistical population from which we can randomly sample new errors $\tilde{e}_i$, $i = 1,2,\ldots,m$, to create standard data sets

$$\tilde{y}_i = G(\theta^*, \tilde{e}_i), \quad i = 1,2,\ldots,m,$$

where θ^* can be chosen to have any desired value.

[1]Figures in brackets indicate literature references.

The data sets so simulated are not entirely "real." The model (1) relating observations y_i to errors e_i must still be specified, and need not be correct. However, such models can be specified (and criticized) taking account of chemical and physical theory, without also imposing statistical assumptions about distributions of errors. As such, these models are "prestatistical."

A population of residuals $\hat{e}_i$ that is too small can misrepresent statistical variation. Thus, attempts should be made to constantly enlarge this population with new residuals obtained from real data obtained in contexts described by the model (1). Since the "bootstrap" is a fairly recent statistical development, new insights into problems and advantages connected with the method are constantly being published. Consequently, the input of specialists in "bootstrap" methodology should be sought when applying this method to the generation of standard data sets. In particular, changes in instrumentation, personnel, or experimental design may change the error population over time. Careful attention should be paid to detect such shifts in distribution.

Not all measurement contexts lend themselves to the "bootstrap," since the transformation (2) from observations to errors may not exist, or may not be well defined. (This may be the case, for example, with the Gamma Ray Spectrum Analysis example discussed by Dr. Currie.) However, when the "bootstrap" does apply, it can be used both to create standard data sets, and also to provide nonparametric estimates of precision [5,7,8].

Estimates of Precision

Although I share Dr. Currie's concern that the laboratories in his two examples either failed to provide estimates of precision, or gave incorrect estimates, I must point out that in Dr. Currie's two examples, it is not clear what measures of precision are appropriate. In all of the analyses, multiple decisions are made. For example, in the Gamma-Ray examples, the locations and amplitudes of several peaks had to be determined *simultaneously*. Although individual standard errors can be given, these do not directly provide measures of simultaneous accuracy [9]. Further in the detection spectrum and precision spectra sets, even the *number* of peaks was unknown. This produces a highly complicated estimation problem for which only large-sample approximations to precision are available. There is some evidence in the literature that such large-sample approximations have considerable bias in moderate samples (see, e.g., [6]).

Similar problems arise in the three data sets for the NBS-EPA Source Apportionment study, particularly in the case of Data Set I, where the number of sources is left unspecified. Both ridge regression and factor analysis are exploratory methodologies, requiring iteration and judgment that are difficult to describe analytically. The only available measures of precison for such techniques are large-sample approximations which refer to analytical formulas for the estimators not directly related to the way such estimators are actually obtained. For example, I know of no way to specify the precision of estimates of slope obtained by the ridge trace method. Published formulas for the precisions of ridge regression estimators refer to those estimators in which the ridge factor k is a specified function of the data, rather than being obtained by inspection of the ridge trace.

Given the complex natures of the estimation problems that Dr. Currie describes, and the fact that statistical theory has not yet provided reasonable estimates of precision for some of the methodologies used in these problems, it is not surprising that the laboratories either failed to provide measures of precision, or gave estimates that were off the mark. Clearly, there is much theoretical statistical work yet to be done.

In the meantime, it should be mentioned again that the "bootstrap" can provide estimates of precison in cases where the assumptions (1), (2) underlying the bootstrap are applicable.

Analogy to CMP Interlaboratory Comparisons

As already noted, Dr. Currie makes an analogy between the use of standard data sets in the two examples he discusses, and traditional interlaboratory comparisons using common reference materials. However, this analogy cannot be carried too far, since there are some important differences in context.

In traditional interlaboratory comparisons, differences between laboratories are usually assumed to be due to variations in the calibration (adjustment) of the instruments, or to differences in instrumentation or technique. Consequently, a one factor (additive) components of variance ANOVA model can reasonably be employed to assign variability between inter- and intra-laboratory sources.

In the standard data set context described by Dr. Currie, however, there are at least three factors which can describe variability between laboratories: 1) different models or assumptions used, 2) different statistical methodologies employed, and 3) different numerical algorithms. Further, the "levels" of these factors (particularly factor 1) appropriate to describe a given laboratory's analysis are not always apparent. (Not all assumptions made are clearly stated). For example, "outliers" can be discarded, parameter values can be truncated (e.g., negative estimated amplitudes reported as zero), or several different analyses may be run but only one (the one that the laboratory thinks is "right") reported. Consequently, it will be difficult to separate sources of interlaboratory variation.

Even worse, even if all factor levels can be accurately identified (or set in advance), there is the clear possibility of

interaction among factors, making interpretation of results difficult. For example, different methodologies work "best" in the context of different models, and different models and methodologies lead to different algorithms and thus to different reasons for numerical instability when such algorithms are applied to data.

How do we then interpret Dr. Currie's examples? First, and most important, we see that in certain far more precisely defined contexts (in terms of model) than would be met in practice there are wide variations in conclusions between laboratories, and also wide variations from the correct answer. Second, the divergences between laboratories cannot be assigned to sampling variation because (and this is the beauty of the standard data sets) the data are fixed. However, the divergence of the centroid of the laboratory conclusions from the truth may be due to sampling variation (the sample did not represent the population), or to poor laboratory conclusion-making processes, or to both. We cannot partition this last variability in terms of possible causes, because no accurate measures of precision over sampling variation are provided by the laboratories, or (in some cases) known.

To better understand the sources of interlaboratory variation, we need to start with pilot studies that control the levels of each of the factors (1-3) listed above. To establish the contribution of algorithms to interlaboratory variability, we need to ask numerical analysts to study the possible numerical errors that can occur in algorithms, describe the situations that produce these errors, and suggest remedies to reduce such errors. (Here, our "pilot sample" design fixes all factors but the "algorithm" factor.) To establish the contribution of methodology (or rather methodology - model interactions) to such variability, chemometricians (particularly statisticians) need to use mathematical analysis and simulaton to identify formulas for the precisions (sampling variability) that can be assigned to the various methodologies in the contexts of various models (Here, we assume a fixed, perfectly accurate algorithm, and vary combinations of method and model.) Finally, we need to study and assess the variability due to choice of model, and also to the other "scientific judgments" made by a laboratory in choosing methodology and algorithms and in announcing measures of precision. It is particularly in this last type of study that standard data sets, both real and artificial, can be most useful.

Scientific Judgment and Bayesian Inference

The question of how to analyze the biases introduced by "scientific judgment" has a direct relationship to a long-standing controversy between classical (frequentist) statisticians and those statisticians who advocate a Bayesian approach. Scientific decision-making involves subjective judgments about both models and types of permissible conclusions. When such judgments are unstated, we have seen that this can obscure our understanding of how decisions are reached, and thus prevent us from finding sources of "bias" or error.

Bayesian statisticians, who try to mathematically model their subjective judgments in terms of prior probabilities over unknown parameters (and models), are often accused by frequentist statisticians of proposing analyses that lack "scientific objectivity." Clearly the contrary is true. The scientist who claims to base conclusions only on the "objective" evidence provided by observed frequencies is nevertheless often guilty of imposing unstated judgments on such evidence. The Bayesian, at least, tries to bring these judgments into the open, where they can be assessed along with the data. Even if we doubt that probability models can ever serve as adequate models of subjective belief, we can still applaud the Bayesian's efforts to expose the methods by which this belief interacts with the evidence in the data to produce new judgments or belief. Rather than criticize the Bayesian for being "subjective" or "biased", the frequentists need to find ways of making their own decision-making processes available for objective study, so that we can gain the opportunity to learn how to improve scientific judgment.

References

[1] Anderson, E., The Irises of the Gaspe Peninsula. *Bull. Amer. Iris Soc.* **59**, 2-5 (1935).

[2] Fisher, R.A, The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179-188 (1936).

[3] Longley, J.W., An appraisal of least squares programs for the electronic computer from the point of view of use. *Journ. Amer. Statist. Assoc.* 62, 819-841 (1967).

[4] Lachenbruch, Peter A., Review of "STAN, Version II.0." *The American Statistician* 39, 146-148 (1985).

[5] Efron, B., *The Jackknife, the Bootstrap and Other Resampling Plans*, Soc. Indust. and Appl. Math.: Philadelphia, PA (1982).

[6] Diaconis, P., and B. Efron, Computer intensive methods in statistics. *Scientific American*, May, 116-130 (1983).

[7] Freedman, D.A., and S.C. Peters, Using the bootstrap to evaluate forecasting equations. Tech. Report, Dept. Statistics, U. Calif. Berkeley (1983).

[8] Freedman, D.A., and S.C. Peters, Bootstrapping a regression equation: Some empirical results. *Journ. Amer. Statist. Assoc.* 79, 97-106 (1984).

[9] Miller, R.G. Jr., *Simultaneous Statistical Inference*, 2nd edit. McGraw-Hill: NY (1981).