Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 479–484

Comment

Gordon Sande[1]

It is a pleasure to provide comments on an article developing an idea whose time has finally come. The standard comment is that this is a fine beginning, and *if only* the authors had done it differently. However, the authors are to be commended for doing it at all. But first, a review of some related topics.

Randomized response was introduced by Warner (1965). There was the usual round of enhancements and improvements by others. Warner (1971) provided his own enhancements. Rather than just further develop a clever algebraic trick, he clearly described a viewpoint in which there is a distribution about which we would like to make inferences and a distribution from which we can make observations. They are different distributions but related in a well-controlled statistical fashion so that we can make the inferences even though we cannot make observations directly of the distribution of interest. The relationship is that we have a mixture of distributions. We do not know from which distribution in the mixture an observation has been taken. The mixing scheme is randomized response. As an example of his method Warner discusses the problem of a central database holder who is trying to release a useful sample of observations that can be used for analysis while protecting the confidentiality of the data in the database. The observation space is more a convolution of distributions than a mixture, as the proposed scheme is to add data records rather than anything we would now tend to think of as randomized response. The notion that randomized response can be used when the analyst *interviews* the respondents through *the computer* to obtain the released data is clear.

The basic notion underlying a public use sample is that it is a sample from the *same distribution* as the original data. Actually we only want to make inferences about the distribution of the original data, so we are within the viewpoint that Warner described. The operational effect of being a sample from the same distribution is that we can use the *same methods*, and even the *same software*, as we would use for the original data. For training purposes this is an overriding concern. For analysis purposes that is a considerable convenience. But we also wish to deidentify, sample and perturb the public use sample to protect the confidentiality of the respondents. There is a fundamental conflict between the objective of perturbing the data and having it be a sample from the same distribution as the original data. It should be no surprise that producing a public use sample is difficult. When it is perturbed enough it is no longer a sample from the same distribution as the original data and becomes unacceptable for analysis (Marsh et al. 1994).

The alternate question plays a central role in randomized response. The intent is that the alternate question should have a distribution similar to the real question so that the response would not indicate which question is being answered. If the real question is

[1] Sande and Associates, Inc., 600 Sanderling Court, Secaucus, NJ 07094, U.S.A.

how often you have not fully declared all your purchases to customs in the last year, with an expected answer of zero through five, then an alternate question of the score of your favourite soccer team, with an expected answer of zero through five, might be plausible but the score of your favourite basketball team, with an expected answer of 60 to 110, would not be plausible. How to do this sensibly in the field is the practitioner's art. It is trivial after the fact that if one is *interviewing* the respondents through *the computer* to provide the response they would have given if they had been interviewed using randomized response. The pragmatic advice would be to use a random variable from the observed empirical distribution as the alternate question and it will not be apparent when the alternate question has been asked. The marginal distribution after randomized response will be unchanged.

The use of randomized response to release data while protecting confidentiality has been rediscovered several times. Dalenius (1977) develops several topics including non-reversible privacy transformations and notes both that Warner (1971) has a short discussion of randomized response used for this purpose and that they can be used to provide public use samples. This author (Cox and Sande 1979) suggested it even with the use of the observed empirical distribution and some cautions about inconsistent records. Fox and Tracy's (1986) monograph on randomized response methods has a section titled *Disclosure Control*. Adam and Wortmann's (1989) survey on query protection methods for statistical databases includes using randomized response to provide a randomized database from which it is safe to provide queries. Särndal et al.'s (1992) sampling text includes comments on the use of randomized response for confidentiality protection. An extensive literature review would surely provide additional rediscoveries of the use of randomized response for confidentiality protection.

The difficulty of analysis makes its presence felt in the query protection methods as there is a discussion of how to determine the related query that is required to determine the randomized response estimators. The query protection needs both entries, or equivalently one entry and the total, in a contingency table with two entries to do the estimation for the randomized response procedure used for the *yes* or *no* variables. The related query being determined provides the total. There is no development of the technique to more elaborate queries. A possible impression from this example is the common misconception that randomized response can only be applied to a single variable. An example of the bivariate use of randomized response is Chen (1978). Rather than treating randomized response as a clever algebraic trick, Chen (1979) analyzes randomized response with the theory of misclassified observations. The results are, of course, identical. His viewpoint that this is just misclassification with a known mechanism makes much already developed theory available. The description of randomized response as purposive misclassification is a bridge to existing practice (Kuha and Skinner 1997).

None of this will be very surprising to experimental physical scientists. Particles of light, photons, have a tendency to be misclassified as they often arrive at the *wrong* detector. We do not know which photon was misclassified but we often know the mechanism. The optical design problems of the Hubble space telescope are a well-known example of this type of problem. The misclassification mechanism is called the point spread function in image restoration. There are several solution techniques available. Directly deblurring images using a point spread function is prone to yield negative intensities. This is often called Weiner filtering after its signal processing analog. The negative intensities are not considered acceptable and truncating the estimator to be positive is not considered to be an improvement. Rather than just solving with the point spread function, the best positive estimate can be found using non-negative least squares (Lawson and Hanson 1974). This is not commonly used in image restoration as there are other techniques which use the special structure of the problem. The E-M algorithm is used, except it is called the Richardson-Lucy method (Richardson 1972, Lucy 1974) as it was developed before the E-M algorithm name was proposed, to provide an iterative technique of moderate iteration cost with slow convergence. The physical science experiments differ from survey practice in their much larger population of photons, their large number of classifications, or detectors, and the insights from the highly structured circumstances of the experiments. Photons are generally cooperative respondents. There is no need to be concerned with the confidentiality of photons.

It is often instructive to consider a method in extreme settings. An example which presented itself recently was the problem of providing a microdata release for secondary analysts of a substance abuse survey (Sande 1996). The need for confidentiality in such a survey is greater than usual, if such judgments can be made. There is much interest in the data. The respondents are likely to be fairly recognizable and one might expect some attempts to be made to reidentify the sample. Randomized response is already a standard interview technique for substance abuse, so the notion of using randomized response for confidentiality protection of a microdata release is very natural. Under moderate rates for the alternate question there will be some fraction of the records which have no modification to either their identification, often called key, or data fields and possibly even both. This is an uncomfortable outcome, so we might choose to require that some of both the identification and data fields must always be modified. This is a restriction on the randomization which chooses the fields to be modified. The next step is to require that a fixed number of fields be subject to modification. The control can be extended both across fields and over records so there will be the right number of randomly chosen modifications in both directions (Cox 1987). Since we want to be sure of the protection provided it is readily suggested that a third or a half of the fields be modified. Perhaps some additional information might be provided when half the fields are being modified by splitting the record into two complementary randomized responses in the style of a Monte Carlo method variance reducing swindle.

The ability to do effective record linkage is reduced when the matches may be spurious because of the purposive misclassification. If our objective is to frustrate record linkage attempts, we can take the additional step of having the alternate question never return the correct response. A readily available alternate question is the *anything but the actual response* distribution given by the empirical distribution with a zero at (i.e., conditional on) the actual response. This alternate question never provides the correct response and looks like the true responses in the population. The marginal distributions will no longer be exactly preserved. (The marginal distribution can be exactly preserved in some cases by use of other conditional distributions. We may not like these conditional distributions when they exist. We must be prepared for the general case so there may be little gained from exactly preserving the marginal distribution when it is possible.) The *anything but* alternate is very close to Warner's (1965) original suggestion of inverting a *yes or no* question as the alternate question. It fails when there are only two responses and there is a high rate of use of the alternate. This scheme of controlling the randomization and conditioning on the actual response would never be attempted for field interviews as it is too elaborate but is simple after the fact in the computer. The analysis would require corresponding care to deal with both the controlled randomization and the known mechanism misclassification. We can now guarantee that the microdata release is a third or a half carefully controlled statistical noise, fit only for statistical consumption and use in making inferences. One might even expect some rather wry comments on the nature of such data.

When we are in such an extreme setting as a substance abuse survey we may be willing to tolerate the increase in variability, and the need for special analysis techniques, if that is the only way to have access to a microdata release. *Safe data* has a high price which may be justified. The provision of a *safe setting* for the data will also be attractive but may be difficult to arrange. When the source of data for the microdata release is a larger survey or census, the variability can be lowered by using a higher sampling rate. The higher security of the controlled and conditioned randomized response release will permit the higher sampling rate. If we increase the sampling rate to include the whole source file, we will have produced a randomized database from which we can permit all possible queries. When the permitted queries must be built up from the usual classification-based queries, the problem of producing the randomized response estimators should be tractable.

Enough review. The problem of inconsistent data is raised by the authors. Their example of maternal parity is a nice example of a single attribute which would be subject to randomization being represented by several fields in the data records. The opposite form is where several attributes, town within region is their example, are represented in a single field of the data record. Another version of this problem is where the same attribute is represented in several fields, as would occur if one had both the birth year and age of a respondent. Redundancy is often deliberately sought to aid in the editing and checking of the data. Randomization should not be applied to an attribute in a way which would produce inconsistent results in data fields. All these examples illustrate that the number of attributes and the number of data fields are not the same. These examples are simple to recognize but the problems introduced by longitudinal data are much more difficult.

Randomized response is more than just a set of clever algebraic tricks, notwithstanding the impression one gets from its literature. The ability to do randomized response after the fact in the computer on the behalf of the respondent means that the problems of operational simplicity and respondent biases are eliminated. It is simple to exactly preserve the marginal distributions although we may choose to do so only approximately. There are estimation techniques other than the proposed one available. The solution method proposed by the authors is the direct analog of Weiner filtering with truncation to positive values. Physical scientists choose not to use it for positive quantities and that is undoubtedly good advice. The connections to conventional statistics come through the notions of mixture distributions and of known mechanism misclassification. These connections should be used.

The authors are to be commended for actually trying a method that others have only talked about. With the method in use, we can think about it in more constructive terms. One immediate benefit is that it helps us recognize the nature of the *assumption* that a public use sample is a sample from the same distribution as the original data. It has become a definition of a public use sample. This need not be the case. We can now discuss the consequences of the assumption and possible alternative microdata release strategies. The use of mixture distributions to bypass the issue of disclosure risk in a microdata release would seem to be worthwhile. It does require that we not be timid in the use of randomization. It is clear that record linkage techniques have become sufficiently capable that preparing conventional public use samples will not be possible much longer (Fellegi 1998, Winkler 1998). The use of highly attribute-coarsened public use samples for training purposes may still be possible. For other uses we will need either *safe settings* or techniques to provide *safe data*. The use of mixture distributions can frustrate the record linkage methods when used with both controlled and conditioned randomization. The cost will be increased variability for the fixed sample sizes of special purpose surveys or increased sample size for microdata releases from censuses. We will need to gain experience with the tradeoffs involved. The need for special analysis techniques is evident when we use the mixture distributions viewpoint. We see that the special analysis is not greatly different than what we already do when we use the viewpoint of randomized response as purposive misclassification, although not all analyses have dealt with misclassification.

References

Adam, N.R. and Wortmann, J.C. (1989). Security-Control Methods for Statistical Databases: A Comparative Study. ACM Computing Surveys, 21, 515–556.

Chen, T.T. (1977). Analysis of Randomized Response as Purposively Misclassified Data. Proceedings of the Section on Survey Research Methods, the American Statistical Association, 765–770.

Chen, T.T. (1978). Log-Linear Models for the Categorical Data Obtained from Randomized Response Techniques. Proceedings of the Social Statistics Section, the American Statistical Association, 284–288.

Cox, L.H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. Journal of the American Statistical Association, 82, 520–524.

Cox, L.H. and Sande, G. (1979). Techniques for Preserving Statistical Confidentiality. Bulletin of the International Statistical Institute, 42, 499–512.

Dalenius, T. (1977). Privacy Transformations for Statistical Information Systems. Journal of Statistical Planning and Inference, 1, 73–86.

Fellegi, I.P. (1997). Record Linkage and Public Policy – A Dynamic Evolution. Record Linkage Techniques – 1997, Washington DC: Office of Management and Budget.

Fox, J.A. and Tracy, P.E. (1986). Randomized Response: A Method for Sensitive Surveys. Beverly Hills, CA: Sage.

Kuha, J. and Skinner, C. (1997). Categorical Data Analysis and Misclassification. In L. Lyberg et al. (eds.). Survey Measurement and Process Quality, Ch. 28, New York: Wiley.

Lawson, C.L. and Hanson, R.J. (1974). Solving Least Squares Problems, Englewood Cliffs, NJ: Prentice-Hall.

Lucy, L.B. (1974). An Iterative Technique for the Rectification of Observed Distributions. Astronomical Journal, 79, 745–754.

Marsh, C., Dale, A., and Skinner, C. (1994). Safe Data versus Safe Settings: Access to Microdata from the British Census. International Statistical Review, 62, 35–53.

Richardson, W.H. (1972). Bayesian-Based Iterative Method of Image Restoration. Journal of the Optical Society of America, 62, 55–59.

Sande, G. (1996). Mixture Distributions Microdata Releases. Research Proposal. Secaucus NJ: Sande and Associates.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling, New York: Springer Verlag.

Warner, S.L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, 63–69.

Warner, S.L. (1971). The Linear Randomized Response Model. Journal of the American Statistical Association, 66, 997–1001.

Winkler, W.E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. In J. Domingo-Ferrer, (ed.), Statistical Data Protection '98. Conference Proceedings 25–27 March, Lisbon, Portugal.