STATISTICAL ANALYSES FROM SUMMARY DATA AND THEIR IMPACT ON THE ISSUE OF CONFIDENTIALITY

Joseph L. Gastwirth, George Washington University *

Abba M. Krieger, University of Pennsylvania

Donald B. Rubin, Educational Testing Service **

This paper discusses statistical issues relevant to the problem of confidentiality. A general structure is presented and statistical methods for estimating parameters from grouped, rounded, or intentionally contaminated data are surveyed.

1. Introduction

There is growing concern that information provided by an individual to a specific source remains confidential to that source. The information collected, however, can be valuable for policy making decisions or research. For example, patients provide doctors with information necessary for the treatment of a disease, and these data are valuable to a researcher for studying the causes and cures for that disease. The information collected by the Internal Revenue Service is another example, because it might be important for making economic policy decisions. The objective is then to find appropriate forms for making these data available in such a way as to maintain the confidentiality of an individual's record and to provide information from which accurate inferences can be drawn. The paper treats the problem from the perspective of the protection of confidentiality. The issue of individual privacy, that is whether an individual has the right not to participate in a research study unrelated to the original purpose for which the data were collected, cannot be resolved by statistical methodology.

Before we discuss statistical methodologies that can contribute to solving this problem, it is helpful to develop a framework that can be used to categorize the subproblems, and identify major areas where almost no techniques currently exist. We define the agent to whom the data are given as the "collector" (C), and the agent who wants the data for policy decisions or research purposes as the "analyst" (A). In general, C obtains an identifier (i.e., label) and m variables for each of N individuals. We will denote this Nx(1+M) data set by $(\underline{I},\underline{\underline{X}})$. The analyst A wishes to have the value of the statistic $S(\underline{\underline{X}})$. However, C does not pass $\underline{\underline{X}}$ to A. We will denote the transformed version of $(\underline{I},\underline{\underline{X}})$ which is passed on to A by $(\underline{I}_T,\underline{\underline{X}}_T)$. Our objective is to study ways of creating $\underline{I}_T$ and $\underline{\underline{X}}_T$ so that it is related to $(\underline{I},\underline{\underline{X}})$ through the following two general conditions:

CC: (Confidentiality Condition)

$$(\underline{I}_T,\underline{\underline{X}}_T) \not\Longrightarrow (\underline{I},\underline{\underline{X}})$$

where

$\not\Longrightarrow$ means that confidentiality is maintained. Confidentiality is violated if too much information about one or more of the individuals is recoverable from $(\underline{I}_T,\underline{\underline{X}}_T)$.

AC: (Analyst Condition)

For each statistic $S(\underline{\underline{X}})$ of interest (e.g. the Gini index of one of the variables in $\underline{\underline{X}}$ or the correlation between two of the variables in $\underline{\underline{X}}$) there exists an estimating procedure S_T on $(\underline{I}_T,\underline{\underline{X}}_T)$ such that $S_T(\underline{I}_T,\underline{\underline{X}}_T)$ is close to $S(\underline{\underline{X}})$.

These conditions clearly identify the tradeoff. By its nature, CC demands that $(\underline{I}_T,\underline{\underline{X}}_T)$ be "far" from $(\underline{I},\underline{\underline{X}})$ while AC demands that these data sets be "close". The (N,K) dominance rule formalizes CC in context of categorical data. (c.f.,6A,7A, 8A, 9A). Our overview puts more emphasis on techniques currently available to A.

Section 2 is devoted to the problem where there is only one collector and one variable (i.e., m=1). We extend the problem in Section 3 to the case with multivariate $\underline{\underline{X}}$. Finally, in Section 4 we treat the problem of data merging where the analyst requires information from more than one collector. Section 2 treats the least complex problem; hence, there is a vast literature on statistical methods which can be adapted for the purpose of solving this problem. For this reason, a disproportionate share of the paper is devoted to the second section. This is not to imply that this problem is the most important. The fact that relatively little work has been done on the problems discussed in the third and fourth sections indicate the possibility of interesting research topics in these areas.

2. Statistical Techniques for Univariate Data

In this section we assume that C transfers information on only one variable to A. Since $S(\underline{\underline{X}})$ does not depend on $\underline{I}$, the identifiers in themselves are of no use when $\underline{\underline{X}}$ is given. We also assume that $\underline{I}$ and $\underline{\underline{X}}_T$ are such that $\underline{I}$ is unimportant when $\underline{\underline{X}}_T$ is given, so that $S_T(\underline{I}_T,\underline{\underline{X}}_T) = S_T(\underline{\underline{X}}_T)$. In this case we may let $\underline{I}_T=\underline{I}_\phi$ ($\underline{I}_\phi$ denotes the case where no information about the identifiers is available). Setting $\underline{I}_T=\underline{I}_\phi$ will aid in satisfying CC at no expense to AC. The issue is then to explore ways of creating $\underline{\underline{X}}_T$. If $\underline{\underline{X}}$ is such that $(\underline{I}_\phi,\underline{\underline{X}}) \not\Longrightarrow (\underline{I},\underline{\underline{X}})$ then there is no need to transform $\underline{\underline{X}}$ in order to preserve CC. The statistically interesting problems arise when $\underline{\underline{X}}$ needs to be transformed.

One method of transforming $\underline{\underline{X}}$ that is well-suited for solving this problem is to group the data: $\underline{\underline{X}}_T$ will then consist of the following information:

1) boundaries for the grouping intervals $a_0 < a_1 < \ldots < a_k$

2) n_i = number of observations between $a_{i-1}, a_i)$, i=1,...,k, and perhaps

3) $\bar{x}_i$ = means of the n_i observations falling into interval i, or even additionally,

4) s_i^2 = variance of the n_i observations falling into interval i.

If the boundary points are chosen wisely so that none of the n_i are small, then CC will be satisfied. The senses in which AC is satisfied are explored in subsections 2.1 and 2.2 Other methods for transforming $\underset{=}{X}$ are briefly considered in subsection 2.3.

2.1 Grouped Data with the Parametric Form of the Density Known

It is often convenient to model the data set $\underset{=}{X}$ by asserting that the variable of interest follows some density function f(x). We first assume that f(x) is thought to belong to a specific parametric family (e.g., normal). Although in order for AC to be satisfied all the information needed to be given in $\underset{=}{X}_T$ is the values for the sufficient statistics for the assumed parametric family, it is important to guard against the possibility that the model is mis-specified[1]. Including grouped data in $\underset{=}{X}_T$ provides a richer database for exploratory data analysis.

We first focus on estimating parameters from grouped data. Although methods for finding maximum likelihood estimates from grouped data exist, these approaches are typically computationally cumbersome unless appropriate summary statistics in each interval are included in $\underset{=}{X}_T$ (e.g., $\bar{X}_i$, s_i^2 for the normal). Since the boundary a_i is approximately the $(\sum_{j=1}^{i} n_j)$ order statistic, another approach is to consider estimations of the form $S_T = \sum_{i=1}^{k-1} c_i a_i$ (i.e. best linear combinations of those order statistics).

Ogawa [26] has determined the optimum value of the weights, c_i, for the location and scale parameters for a symmetric density of the form $1/\sigma\ f((x-u)/\sigma)$ and the approach is readily adaptable to other shapes. It is important to note that the c_i depend on f(x). His results have been used and generalized by many authors. Table 1 summarizes some of the vast literature reporting the efficiency of estimates based or the optimum spacings relative to the maximum likelihood estimate for various distributions.

A quick study of the results in Table 1 indicates that with 5 to 10 properly selected intervals, relatively efficient estimates are readily derived. Since many large data sets are reported with at least 15 intervals[2], these results suggest that the parameters of assumed density functions can be well estimated from linear combinations of order statistics. Moreover, robust linear combinations of a few percentiles have been derived for certain families of densities, e.g. in order to estimate μ for symmetric densities, [13], and promising research has begun on estimating the scale parameter. Therefore, in the univariate case with large samples, techniques from grouped data appear to be valuable as the database is rich enough to satisfy C2, even against mis-specification of the parametric family, and still preserve CC.

When $\underset{=}{X}_T$ is extended to include the group means, $\bar{x}_i$, i=1,..,k, even better estimates can be obtained. Smith, [32] studied estimators which are linear combinations of the group means, (i.e., $\sum_{i=1}^{k} c_i \bar{x}_i$) as well as estimators combining the group means with the percentiles (boundary points). In the parametric cases he studied, the efficiency of the optimal linear combination estimators of location for double-exponential and logistic data were greater than 93% when 5 percentiles were used. For the scale parameter, an efficiency of over 95% relative to the best estimator for each distribution was obtained for the normal, double-exponential, log-normal and log-logistic when only three equally spaced fractiles were used. Indeed Smith showed that optimal linear combinations of group means are more efficient than optimal linear combinations of order statistics based on the boundaries. Naturally, there may be a tradeoff between maximizing the efficiency of an estimator relative to the MLE for the complete sample and the robustness of the estimation to the mis-specification of the underlying density (c.f. [32]). Although suppliers of data could also report goodness of fit tests for the underlying data, we shall see that accurate bounds for the variance and higher moments can be obtained from grouped data just by utilizing the general shape of the underlying density.

Although the results in this section are based on optimal spacing, other choices of percentiles, such as equi-probable or those based on expected values of order statistics do nearly as well. Mosteller demonstrated this when f(x) is normal. Eisenberger and Posner [6] derived optimum estimators for both μ and σ with f(x) normal, by minimizing several criteria of the form $V(\hat{\mu}) + bV(\hat{\sigma})$ (b=1,2,3,). In all cases the efficiency of the estimators of both $\hat{\mu}$ and $\hat{\sigma}$ were greater than 90% and 98% for 10 and 20 percentiles respectively.

2.2 Grouped Data with Assumptions on the General Shape of the Density

In some applications one may know that the density function decreases (e.g. income and survival data in the tail) or is unimodal (e.g. educational test data).

For decreasing density funtions where $\bar{x}_i$ are given, Gastwirth and Krieger [1A] have derived upper and lower bounds from grouped data for $\int_{a_k}^{h} h(x)\, dF(x)$ assuming h(x) is convext (e.g. all moments) and a_k is finite. The effectiveness of these results is illustrated by looking at two examples:

Example 1 $F(x) = (1-e^{-x})/(1-e^{-2})$ for $0 \leq x \leq 2$

Example 2 $F(x) = 125(x^3-1)/124x^3$ for $1 \leq x \leq 5$.

In order to indicate the value of the information contained in the group means, Krieger [17] obtained the bounds with and without the $\bar{x}_i$. Table 2 presents these results. For these cases, we can see that it is preferable to have the group means rather than reporting information for more intervals.

Krieger [18] obtained analagous results assuming the underlying density function is unimodal. For example, for the standard normal density and 8 groups, bonds on the variance are given by the interval (0.9683, 1.0309). Similar bounds for percentiles (i.e. $F^{-1}(p)$) with known and unknown group means are illustrated for the standard normal in Table 3.

Even if assumptions about the shape of f(x) over the entire domain are difficult to justify, the analyst might still be willing to make assumptions about the shape of the density function over more restricted regions. The results described above can be used for these restricted regions and then pieced together for the entire domain. Finally, if the analyst is not willing to make any assumptions about the shape of f(x) it is possible to obtain results in certain cases (c.f. bounds on the Gini index and other measures of variation as shown in [8]).

2.3 Rounded Data and Contaminated Data

One possible approach for transforming $\underline{X}$ is to round the values of $\underline{X}$. In order for CC to hold, this rounding must be coarse. Mathematically, rounding can be viewed as a special case of grouping where the data is grouped into intervals of width h centered at c-ih, i=0, ±1, ±2,.... We treat rounded data separately, however, as specialized techniques have been developed for this form.

Of particular note are Sheppard's corrections which relate moments of rounded and unrounded distributions under certain regularity conditions in the limiting case of small rounding errors. For discussion, see Kendall and Stuart [2A], and Wold [3A]. For the normal, Sheppard's corrections give the MLE in the limiting case [4A]. Using Fourier techniques, McNeil [24] developed a general formula for a consistent estimator of a parameter and gave several illustrations. For estimating the mean of a normal density function with known σ he obtained quite efficient estimates (74%) relative to the MLE even when $h=2\sigma$. When $h \leq \sigma$, the efficiency is greater than 90%. For estimating the scale parameter of the exponential distribution, as long as h was less than the mean of the data his estimates had efficiency > .92. Unfortunately, for multi-parameter families the algebra involved becomes rather complex. David and Mishriky [4] showed that the grouping has only a moderate effect on the expected value of order statistics supporting the contention "that estimates appropriate in the ungrouped case will continue to give good results in the grouped case."

Another possible approach is to intentionally contaminate $\underline{X}$. In our framework we could write $\underline{X}_T = \underline{X}+\underline{e}$ where for example $e_1,\ldots,e_N$ are independent and identically distributed from a suitably chosen distribution with enough variance to insure CC holds. One idea which will aid in satisfying C2 is to let

$$d_i=(e_i-\bar{e})\left[\frac{-2\sum_{i=1}^{N} X_i\,(e_i-\bar{e})}{\sum_{i=1}^{N}(e_i-\bar{e})^2}\right]$$

and then let $\underline{X}_T=\underline{X}+\underline{d}$. The mean and variance of the transformed values agree with the mean and variance of the actual data. This method can be extended to make sure that the first ℓ moments of $\underline{X}$ and $\underline{X}_T$ are the same. Of course, $\ell << N$ is needed for CC.

3. Statistical Techniques for Multivariate Data with One Collector

We now assume that C transfers to A information on more than one variable. As in the case with one variable, we assume that there is no need to transfer the labels (i.e. we let $\underline{I}_T = \underline{I}_\phi$). We first consider grouped data and then summarize other possible approaches for creating $\underline{X}_T$. All of the literature appears to be devoted to the multiple regression problem.

3.1 Grouped Data

One approach using grouped data for the bivariate case is due to Mosteller [25]. He uses the number of observations falling into four corners of the plane determined by the lines $x=\mu_x \pm k\sigma_x$, and $y=\mu_y$. The ARE of Mosteller's approach for estimating ρ assuming normality is about 55% for the optimum choice of k. In order to improve upon this method, more groups must be used. This may be difficult to achieve and still maintain enough observations in each cell to insure CC.

Another approach has been developed by Gastwirth and Spruill. They divide the x data into k ordered groups and then obtain the mean and the variance of the y points for each of these groups. They have studied estimators which only use $(\bar{x}_i, s_{x_i}, \bar{y}_i, s_{y_i})$, i= 1,...,k. The simplest one is to fit a linear regression through $(\bar{x}_i,\bar{y}_i)$ and then estimate ρ by $\hat{\beta}s_x/s_y$. The theoretical properties of the procedure will appear elsewhere [33], but the following summary of their Monte-Carlo results reported in Table 4 shows that little information is lost when this estimate of ρ with ten groups is compared to the true correlation coefficient or the estimate based on complete data.

The book by Haitovsky [11] studies the field of estimation of multiple regression from grouped data. Some of the problems he encountered will be alleviated if the group means and variances are reported as above. Since this area is relatively unstudied, a few other techniques that have been proposed should be noted. One interesting suggestion is due to T. Jabine.

If we denote the data of interest by $(\underline{X},\underline{Y})$ then $\underline{X}_T=\underline{X}$ and $\underline{Y}_T$ is created by grouping $\underline{X}$, looking at all Y values in a particular X group, and then assigning these Y values to an X value in that group at random. Ordinary regression methods can now be used, and if there is a relation between X and Y, again the ordering of the X's should allow it to be estimated.

3.2 Rounded Data

Recent work by Dempster and Rubin [5A] suggests that in regression problem with rounded independent variables, estimates of regression coefficients can be very sensitive to the behavior of the distribution of the independent variables if they are highly colinear. In the limiting case of large samples and small rounding errors, Sheppard's corrections are appropriate, but the importance of this result with coarsely rounded data is not clear. These results suggest that some multivariate estimation problems may be quite difficult using $\underline{X}_T$ if $\underline{X}_T$ is sufficiently rounded to satisfy CC. With coarsely rounded data and colineas $\underline{X}$, supplementary information about the shape of the distribution of $\underline{X}$ may be needed.

3.3 Contaminated Data

One way for the data to be intentionally contaminated is similar to the method used for the univariate case. We can let $\underline{X}_T=\underline{X}+\underline{d}$ and $\underline{Y}_T=\underline{Y}+\underline{d}'$ where $\underline{d}$ and $\underline{d}'$ are determined as above. If it is deemed desirable, we can carry this idea one step further by restricting $\underline{d}$ and $\underline{d}'$ to satisfy the condition that the correlation between $\underline{X}$ and $\underline{Y}$ equals the correlation between $\underline{X}_T$ and $\underline{Y}_T$.

Although when there is only one collector C, A can request the regression equation from C, this is often not a practical solution to the problem because A might want to perform more extensive analyses. To expect C to be able to or be willing to provide this information might be unreasonable. Furthermore, there are instances where the data analytic techniques are only determined after some preliminary investigations are performed. Another reason for describing approaches assuming one collector is to see which methods are adaptable to more than one collector. With two collectors, the regression equation cannot be derived by C.

4. Two Collectors

We now suppose that there are two collectors CC and CA with data sets $(\underline{I},\underline{X})$ and $(\underline{J},\underline{Y})$ where $\underline{I}=\underline{J}$ but $\underline{X}\neq\underline{Y}$; that is, the data sets have identical individuals, but different variables. The analyst's objective is to estimate $S(\underline{X},\underline{Y})$ which needs the merged data set in the sense that $S(\underline{X},\underline{Y})$ cannot be calculated from $(\underline{I}_\phi,\underline{X})$ and $(\underline{J}_\phi,\underline{Y})$ because A needs to relate an individual's X value to his Y value. By analogy with our previous discussions, we seek $(\underline{I}_T,\underline{X}_T)$ and $(\underline{J}_T,\underline{Y}_T)$ such that CC is satisfied for $(\underline{I},\underline{X})$ and $(\underline{J},\underline{Y})$ (i.e. $(\underline{I}_T,\underline{X}_T)\ \eta\ (\underline{I}_T,\underline{X}_T) \not\Leftrightarrow (\underline{I},\underline{X})\ \mu\ (\underline{J},\underline{Y})$)

This problem is considerably more complicated than the single collector problem because it is necessary to provide A with some information about $\underline{I}$ and $\underline{J}$ so that the data sets can be merged. The two collector problem has received little attention in the literature.

We are primarily interested in cases such that knowing values of $\underline{X}$ and $\underline{Y}$ is sufficient to violate confidentiality. Otherwise, $(\underline{X},\underline{Y}) \not\Leftrightarrow (\underline{J},\underline{X})\ \mu\ (\underline{J},\underline{Y})$ and confidentiality can be preserved with appropriate communication between collectors simply by linking the data sets using non-informative identifiers (e.g. translate both $\underline{I}$ and $\underline{J}$ by the same random amount).

The statistically interesting case has the requirement that $(\underline{I}_T,\underline{X}_T)$ and $(\underline{J}_T,\underline{Y}_T) \not\Leftrightarrow (\underline{X},\underline{Y})$. There are many ways in which this can be accomplished. If we look at the four statements $\underline{I}_T\neq\underline{I}$, $\underline{J}_T\neq\underline{J}$, $\underline{X}_T\neq\underline{X}$, and $\underline{Y}_T\neq\underline{Y}$, at least two of them must hold in order to preserve confidentiality. There are methods that transform identifiers, methods that transform data, and methods that transform both identifiers and data.

One possible method for transforming the identifiers is to divide $\underline{I}$ into small groups. $\underline{I}_T$ and $\underline{J}_T$ can be generated by each collector randomly permuting the identifiers for each subgroup of $\underline{I}$. We can then merge the data sets as if $\underline{I}_T$ and $\underline{J}_T$ were the true identifiers. This method has many variants because the size of the subgroups and the method for creating permutations need to be specified. Sometimes the approach described in section 3.1 based on ordering one file and only merging the group means and variances can be adapted to the present case.

Insofar as the $\underline{X}$ and $\underline{Y}$ values are concerned, we still may have to create $\underline{X}_T$ and $\underline{Y}_T$ unless $\underline{X} \not\Leftrightarrow (\underline{I},\underline{X})$ and $\underline{Y} \not\Leftrightarrow (\underline{J},\underline{Y})$. Otherwise, CC would be violated for the $\underline{X}$ or $\underline{Y}$ data. One possible approach would be to create $\underline{X}_T$ and $\underline{Y}_T$ contaminating $\underline{X}$ and $\underline{Y}$ as mentioned in the one collector case. Further work is needed to address these interesting issues.

5. Conclusions

Although many results are available to allow A to study a univariate data set from one collector, there are few methods available when the data set is multivariate. The problem of estimation from multivariate grouped data, for example, has not been extensively treated in the literature. The multivariate problem becomes more complicated when we are constrained by confidentiality considerations. Another area where there are only a few results is data merging. Designing methodologies to obtain information from more than one collector is even more complicated because it is necessary to combine results for multivariate data with data merging techniques and still preserve confidentiality.

REFERENCES

1. Bloch, D. "A note on the estimation of the location parameter of the Cauchy distribution." J. Amer. Statist. Ass. 61, 852-5.

2. Chan, L.K. "Linear Estimation of the Location and Scale Parameters of the Cauchy Distribution Based on Sample Quantiles." JASA 65, 851-859.

3. Chernoff, H., Gastwirth, J.L., and Johns, M.V. Jr. "Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation." Ann. Math. Statist. 38, 52-72.

4. David, H.A. and Mishriky, R.S. (1965). "Order statistics for discrete populations and for grouped samples." J. Amer. Statist. Ass. 63, 1390-8.

5. Dyer, D.D. "Estimation of the Scale Parameter of the Chi Distribution Based on Sample Quantiles." Technometrics 15, 489-495.

6. Eisenberger, I. and Posner, E.C."Systematic statistics used for data compression in space telemetry." J. Amer. Statist. Ass. 60,97-133.

7. Eisenhart, C., Hastay, M.W., and Wallis, W.A. (Eds.) "Selected Techniques of Statistical Analysis" McGraw-Hill, New York.

8. Gastwirth, J.L. "The Estimation of the Lorenz Curve and Gini Index" Review of Economics and Statistics 52, 306-316.

9. Gastwirth, J.L. and Glauberman, N. "The Interpolation of The Lorenz Curve and Gini Index from Grouped Data." Econometrica 44, 479-483.

10. Gupta, S.S. and Gnanadesikan, M. "Estimation of the parameters of the logistic distribution" Biometrika 53, 565-70.

11. Haitovsky, "Regression Estimation from Grouped Observations." Griffin Statistical Monographs and Courses.

12. Hammersley, J.M. and Morton, K.W. "The estimation of location and scale parameters from grouped data." Biometrika 41, 296-301.

13. Hogg, R.V. "Some observations on robust estimation." J. Amer. Statist. Ass. 62, 1179-86.

14. Huber, P.J. "Robust Estimation of a Location Parameter" Ann. Math. Statist. 35, 73-101.

15. Huber, P.J. "Robust estimation."In: Selected Statistical Papers 2, Mathematical Centre Tracts 27, Mathematisch Centrum Amsterdam.

16. Kakwani, N.C. and Podder, N. "Efficient Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations Econometrica 44, 137-148.

17. Krieger, A.M. "Bounds on Moments and Percentiles from Grouped Data." Ph.D Dissertation, Dept. of Statistics, Harvard University 1974.

18. Krieger, A.M., "Bounding Moments, the Gini Index and Lorenz Curve from Grouped Data for Unimodal Density Functions." submitted to JASA.

19. Kulldorf, G. "Estimation of One or Two Parameters of the Exponential Distribution on the Basis of Suitably Chosen Order Statistics." Annals of Mathematical Statist. 34, 1419-1431.

20. Kulldorff, G. and Vännman, K. "Estimation of the Location and Scale Parameters of a Pareto Distribution by Linear Functions of Order Statistics." J. Amer. Statist. Ass. 68, 218-227.

21. Lindley, D.V. "Grouping Corrections and Maximum Likelihood Equations." Proceedings of the Cambridge Philosophical Society 46, Pt. 1.

22. Lloyd, E.H. "Least-squares Estimation of Location and Scale Parameters Using Order Statistics." Biometrika 39, 88-95.

23. Mann, Nancy R. "Point and Interval Estimation Procedures for the two Parameter Weibull and Extreme-Value Distributions." Technometrics 10, 231-56.

24. McNeil, D.R. "Consistent Statistics for Estimating and Testing Hypotheses from Grouped Samples." Biometrika 53, 545-557.

25. Mosteller, F. "On Some Useful Inefficient Statistics." Ann. Math. Statist. 17, 377-408.

26. Ogawa, J. "Contributions to the Theory of Systematic Statistics, I." Osaka Math. J.3,175-213.

27. Saleh, A.K.M.E. "Determination of the Exact Optimum Order Statistics for Estimating the Parameters of the Exponential Distribution from Censored Samples." Technometrics 9, 279-92.

28. Saleh, A.K.M.E. and Ali, M.M. "Asymptotic Optimum Quantiles for the Estimation of the Parameters of the Negative Exponential Distribution." Ann. Math. Statist. 37, 143-51.

29. Salem. A.B.Z. and Mount, T.C. "A Convenient Descriptive Model of Income Distribution: The Gamma Density." Econometrica 42, 1115-1127.

30. Sarhan, A.E. and Greenberg, B.G. (Eds.) "Contributions to Order Statistics." Wiley, New York.

31. Sarhan, A.E. and Greenberg, B.G. "Simplified Estimates for the Exponential Distribution." Annals of Mathematical Statistics, 34, 102-116.

32. Smith, J.T. "Some Statistical Methods for Data Grouped by Quantiles" Unpublished Thesis, John Hopkins University, Baltimore, Md., 1972.

33. Spruill, N., Personal Communications, 1978.

34. United States Bureau of the Census: Current Population Reports (Series P-60). Wash., D.C.: Government Printing Office.

35. Wilk, M.B., Gnanadesikan, R., and Huyett, Marilyn J. "Estimation of parameters of the gamma distribution using order statistics." Biometrika 49, 525-45.

1A Gastwirth, J.L., and Krieger, A.M., "On Bounding Moments from Grouped Data," J. Amer. Statist. Ass. 70, 468-471.

2A Kendall, M.G., Stuart, A., "The Advance Theory of Statistics" Vol. 1, 3rd Ed., Hafner Publishing Co., N.Y., 1969.

3A Wold, H., "Sheppard's Correction Formulae in Several Variables". Skand Aktuarietidskrist, 17, 248-255.

4A Fisher, R.A., "On the Mathematical Foundations of Theoretical Statistics." Phil. Trans. A., 309-368.

5A Dempster, A.P. Rubin, D.B., Personal Communications, 1978.

6A Report on Statistical Disclosure and Disclosure Avoidance Techniques, Statistical Policy Working Paper 2, U.S. Dept. of Commerce, 1978.

7A Sande, G., "Towards Automated Disclosure Analysis for Establishment Based Statistics" Statistics Canada Report, 1977.

8A Sande, G., "Confidentiality and Polyhedra." to appear in the 1978 Proceedings of the American Statistical Association.

9A Cox, L.H., "Automated Statistical Disclosure Control." to appear in the 1978 Proceedings of the American Statistical Association.

10A Harter, H.L., "Estimating the Parameters of Negative Exponential Populations from One or Two Order Statistics." Ann. Math. Statist., 32, 1078.

11A Harter, H.L., and Moore, A.H., "Local Maximum Likelihood Estimation of the Parameters of Three-Parameter Lognormal Populations from Complete and Censored Samples." Journal of the American Statist. Ass. 61, 842.

FOOTNOTES

1. It is necessary for the number of sufficient statistics to be small so that CC is satisfied.

2. This assumes that the a_i are properly chosen. In cases where the data are reported at different points in time, the interval boundaries must be appropriately updated (c.f. [34]).

* This research was supported by an NSF grant MCS77-13882 to George Washington University.

** This research was facilitated by a Guggenheim Fellowship.

Table 1: Efficiencies of Linear Combinations of Order Statistics

Distribution	Parameter	Number of Percentiles	ARE	Ref.
Normal	μ	4	.934	[30]
Normal	μ	10	.981	[30]
Normal	σ (μ known)	6	.893	[30]
Normal	μ and σ	6	.727	[30]
Logistic	μ (σ known)	1	.750	[10]
Logistic	μ (σ known)	3	.938	[10]
Logistic	σ (μ unknown)	2	.684	[10]
Pareto	β (scale); $.5\leq\gamma\leq 5$	3	$\geq .90$	[20]
Pareto	β (scale); $.5\leq\gamma\leq 5$	10	$\geq .98$	[20]
Chi	σ (d.f. known, 1 to 30)	4	$\geq .92$	[5]
Exponential	σ (scale)	4	.927	[30]
Exponential	σ (scale)	6	.960	[31]
Cauchy	μ	5	.952	[1]
Cauchy	σ	9	.978	[2]
Cauchy	μ,σ	10	$\geq .94$	[2]

Table 2: Bounds on Variance

k	Example 1: Variance = .7479 No Means	Example 1: Variance = .7479 Means	Example 2: Variance = 2.4194 No Means	Example 2: Variance = 2.4194 Means
2	(.6051, .8712)	(.7319, .7706)	(.5862, 4.6846)	(2.3313, 2.6229)
4	(.7239, .7771)	(.7549, .7505)	(1.1494, 3.0422)	(2.3962, 2.4646)

Table 3: Bounds on Percentiles with 8 Groups

p	Lower Bound without $\bar{x}_i$	Lower Bound with $\bar{x}_i$	Actual Value	Upper Bound with $\bar{x}_i$	Upper Bound without $\bar{x}_i$
0.2000	-0.8165	-0.7997	-0.7938	-0.7856	-0.7537
0.5000	-0.0206	-0.0088	-0.0000	0.0078	0.0119
0.8000	0.7580	0.7856	0.7938	0.7994	0.8165

Table 4: Estimating ρ for α b variate Normal: Monte Carlo Results (50 replications) for 10 groups, 100 operations

	ρ .9	ρ .75	ρ .5	ρ .25
ungrouped $\hat{\rho}$ ± S.E.	.8986 ±.0064	.7503 ±.0147	.4975 ±.0252	.2351 ±.0228
ground $\hat{\beta}$ $\hat{\sigma}x/\hat{\sigma}y$±S.E.	.8984 ±.0069	.7504 ±.0153	.4958 ±.0268	.2465 ±.0244

Source: N. Spruill, Dissertation to be submitted to Department of Statistics, George Washington University