PIERCE, J. A. (1943). Correction formulas for moments of a grouped-distribution of discrete variates. *J. Amer. Statist. Assoc.* **38** 57–62.

PRATT, J. W. (1981). Concavity of the log-likelihood. *J. Amer. Statist. Assoc.* **76** 103–106.

PREECE, D. A. (1981). Distribution of final digits in data. *The Statistician* **30** 31–60.

PREKOPA, A. (1973). On logarithmic concave measures and functions. *Acta Sci. Math. (Szeged)* **34** 335–343.

PRENTICE, R. L. and GLOECKLER, L. A. (1978). Regression analysis of grouped survival data with applications to breast cancer data. *Biometrics* **34** 57–67.

RUBIN, D. B. (1978). Multiple imputation in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proc. Survey Res. Methods Sec.* Amer. Statist. Assoc., Washington.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Sample Surveys and Censuses.* Wiley, New York.

SANDON, F. (1924). Note on the simplification of the calculation of abruptness coefficients to correct crude moments. *Biometrika* **16** 193–195.

SCHADER, M. and SCHMID, F. (1984). Computation of maximum likelihood estimates for μ and σ from a grouped sample of a normal population. A comparison of algorithms. *Statist. Hefte* **25** 245–258.

SELF, S. G. and GROSSMAN, E. A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics* **42** 521–530.

SHEPPARD, W. F. (1898). On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proc. London Math. Soc.* **29** 353–380.

STEVENS, W. L. (1948). Control by gauging (with discussion). *J. Roy. Statist. Soc. Ser. B* **10** 54–98.

STIRLING, W. D. (1984). Iteratively reweighted least squares for models with a linear part. *Appl. Statist.* **33** 7–17.

STOER, J. and BULIRSCH, R. (1980). *Introduction to Numerical Analysis.* Springer, New York.

STUDENT (1908). The probable error of a mean. *Biometrika* **6** 1–25.

SWAN, A. V. (1969). Algorithm AS 16. Maximum likelihood estimation from grouped and censored normal data. *Appl. Statist.* **18** 110–114.

TALLIS, G. M. (1967). Approximate maximum likelihood from grouped data. *Technometrics* **9** 599–606.

TALLIS, G. M. and YOUNG, S. S. (1962). Maximum likelihood estimation of parameters of the normal, log-normal, truncated normal and bivariate normal distributions from grouped data. *Austral. J. Statist.* **4** 49–54.

TANNER, M. A. and WONG, W. H. (1987a). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics* **29** 23–32.

TANNER, M. A. and WONG, W. H. (1987b). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–540.

THOMPSON, W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics* **33** 463–470.

TOCHER, K. D. (1949). A note on the analysis of grouped probit data. *Biometrika* **36** 9–17.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69** 169–173.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295.

WACHTER, K. W. and TRUSSELL, J. (1982). Estimating historical heights. *J. Amer. Statist. Assoc.* **77** 279–293.

WOLD, H. (1934). Sheppard's correction formulae in several variables. *Skandinavisk Aktuarietidskrift* **17** 248–255.

WOLYNETZ, M. S. (1979a). Algorithm AS 138. Maximum likelihood estimation from confined and censored samples. *Appl. Statist.* **28** 185–195.

WOLYNETZ, M. S. (1979b). Algorithm AS 139. Maximum likelihood estimation in a linear model from confined and censored normal data. *Appl. Statist.* **28** 195–206.

YONEDA, K. and UCHIYAMA, M. (1956). Some estimations in the case of relatively large class intervals. *Yokohama Math. J.* **4** 99–118.

Comment

James Burridge

Heitjan's paper is a useful and interesting survey of the current state of the art regarding "grouped data." Grouping is, as Heitjan says, "ubiquitous." Yet all of us have been brought up on statistical theory and methods intended to deal with "continuous" data—data that none of us will ever see! Justifications for such a perverse situation are of course that it is usually convenient to treat the data as if they were continuous and, often, that the grouping is fine enough for any necessary corrections to be ignorable. There remains of course the grey area where it is not clear whether or not adjustments ought to be used. It is irritating in practice to have, on occasion, to worry about such things. Perhaps, in the near future hopefully, authors of statistical packages will enable us to analyze grouped data as a matter of routine. Certainly the continuing advances in computer processor power are making it increasingly feasible, if not desirable, to analyze the data that are actually observed. However, much of the conventional elegant theory of mathematical statistics may seem less compelling if we routinely adopt such a view: I wonder, for example, whether many results associated with sufficiency may ultimately be seen as mathematical curiosities or, at best, as approximations.

James Burridge is a Lecturer in the Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, England.

One consequence of the emphasis on continuous data is that grouped data are generally regarded as tiresome, steeped in ambiguities or just computationally inconvenient. The elegance of the classical results for continuous data can lead us to forget that models, i.e., Gamma, normal, etc. are almost invariably tentative and, at best, approximations to the "real" world. So we should not be unduly worried about grouped, i.e., uncertain data. For example, many statisticians will cheerfully employ a Weibull distribution to describe a lifetime distribution yet will feel strangely ill at ease when told that the lifetime of interest started in January 1973 and ended in July 1987. For some reason the grouping represented by the uncertain start date seems to cause even more distress than the uncertain end date, despite the obvious symmetry in many applications. I would like to comment briefly on this "doubly grouped" case because it frequently arises in medical and socioeconomic applications.

If the lifetime is known to start at time u and to finish between times a and b, and if $F(y)$ is the distribution function of lifetime, then few of us will hesitate to write the likelihood as a product of terms such as

$$F(b-u) - F(a-u) \tag{1}$$

(assuming independence of course). If u is also unknown an extra layer of uncertainty creeps in. However, we can usually proceed as follows. We may be prepared to assume that start times and lifetimes are all mutually independent. Thus, if the lifetime starts between 0 and τ, then the likelihood will be a product of terms such as

$$P(a < T < b) = \int_{u=0}^{\tau} \{F(b-u) - F(a-u)\} g(u)\, du \tag{2}$$

where g is a density on $[0, \tau]$ and $T = U + Y$. Usually g will be uniform although other choices, possibly involving unknown parameters, may sometimes be desirable. Personally I would prefer to use (2) rather than an approximation based on a Sheppard's type correction. If routines are to be developed to evaluate (1) for use with grouped data then it seems not much harder to provide routines for evaluating (2). (Actually a purist might object that a, b and τ are also subject to error. However, I suppose we have to stop somewhere. Grouping is indeed ubiquitous, if not pernicious!) An intriguing technical point raised by (2) is the effect it has on unimodality properties with respect to the unknown parameters of the resulting likelihoods. Likelihoods based on (1) often have unimodality properties similar to those of the corresponding continuous data likelihoods (see Burridge, 1981a, 1982; Silvapulle and Burridge, 1986).

The situation for the doubly grouped case (2) is less clear. Suppose we consider a linear model for lifetime Y of the form

$$Y = \mu + \sigma E \tag{3}$$

where E represents an "error" with log-concave density f_E. If U also has a log-concave density, as in the uniform case, then standard unimodality properties apply (e.g., see Prekopa, 1973; Pratt, 1981; Burridge, 1982). Let $\alpha = \mu/\sigma$, $\phi = 1/\sigma$. Then $f_T(t)$ is log-concave in t, f_T is log-concave in (α, ϕ), and $P(a < T < b)$ is log-concave in α (or μ) when ϕ is given. I give an outline of the proof of the latter result. Recall that $T = U + Y$. Then, for the case $\tau \le a$,

$$\begin{aligned} P(a \le T \le b) &= P(a - U \le Y \le b - U) \\ &= P(\phi(a-U) - \alpha \le E \le \phi(b-U) - \alpha) \\ &= P(A(\alpha, \phi)) = p(\alpha, \phi), \quad \text{say.} \end{aligned}$$

The event A is a convex set in (U, E) space. Let $A_i = A(\alpha_i, \phi_i)$ and write, for given $0 < \lambda < 1$,

$$\begin{aligned} A_\lambda &= A(\alpha_\lambda, \phi_\lambda) \\ &= A(\lambda\alpha_1 + (1-\lambda)\alpha_2,\ \lambda\phi_1 + (1-\lambda)\phi_2) \end{aligned}$$

and

$$\begin{aligned} A'_\lambda = \{(u, e):\ & u = \lambda u_1 + (1-\lambda)u_2, \\ & e = \lambda e_1 + (1-\lambda)e_2 \\ & \text{for some } (u_i, e_i) \in A_i,\ i = 1, 2\}. \end{aligned}$$

Prekopa's result tells us that

$$\lambda \log P(A_1) + (1-\lambda)\log P(A_2) \le \log P(A'_\lambda).$$

It is an easy matter to show that $A'_\lambda \subset A_\lambda$ for the special case $\phi_1 = \phi_2$ and hence that p is a log-concave function of α (or μ). Although interesting, the above result is not too useful because it is more usual to use a linear model for $\log Y$ and the effect of the convolution $T = Y + U$ is less clear. It is easy to show that f_T is again log-concave in (α, ϕ) but unfortunately the quantity $p(\alpha, \phi)$ involves an integration over a nonconvex region so an immediate application of Prekopa's theorem does not appear possible. I look forward to seeing a clarification of this issue.

Further questions and uncertainties arise in regression contexts. We all know what to do if the response (or dependent) variable is grouped or censored, but we are somewhat at a loss when the explanatory (or regressor) variables are also grouped. Heitjan has given us a useful summary of relevant methods in this latter case—most methods appear to involve some variant of Sheppard's corrections, justifiably I think given the current state of integration routines. The appropriate elaboration comparable to (2) would be,

well, elaborate! Also, unimodality properties of the resulting likelihoods are far from clear at present. The lack of symmetry in our view of grouping is perhaps understandable in the regression context, but less so in the lifetime example cited earlier. The difficulty, I am sure, is primarily psychological, or a result of our training perhaps, rather than an intrinsic feature of the real problem. Bayesians, perhaps, are less troubled by such uncertainties—if you can't observe it, specify a (prior) distribution and then integrate to get rid of it. What could be simpler! Unfortunately computers seem to take a dim view of integration.

The main obstacle to the "correct" routine analysis of grouped continuous data has of course been computational and it is tempting to think that grouped data are always computationally more awkward to analyze than ungrouped data. That this is not always the case is illustrated by two rather different types of problem. The first arises in the context of the so-called three parameter Weibull, lognormal or gamma distributions, i.e., models involving an unknown shift or cutoff value as, for example, in the Weibull case

$$P(X > x) = \begin{cases} 1, & x \leq \gamma, \\ \exp[-\{(x-\gamma)/\alpha\}^{\beta}], & x > \gamma. \end{cases}$$

This type of problem has been studied extensively by Cheng and Amin (1983), Smith (1985) and Cheng and Iles (1987). If α, β and γ are all unknown then the conventional continuous data likelihood can result in an inconsistent global MLE. If, however, the data are analyzed as grouped data, the problem apparently disappears, in large samples at least (e.g., see Titterington, 1985). Cheng and Amin (1983) suggested the maximum product of spacings (MPS) as an alternative and it is interesting to note that their "likelihood" has the functional form of a grouped data likelihood. Presumably the difficulties encountered by conventional MLE discussed by Cheng and Amin can arise in certain multivariate situations. However, their MPS method does not seem to extend easily to the multivariate case. A thorough-going grouped data approach seems, by comparison, straightforward—conceptually at least.

Another situation where a grouped data analysis appears to help is in the empirical Bayes (or "type 2 likelihood") analysis of survival data using Dirichlet, gamma or related processes (e.g., see Kalbfleisch, 1978; Burridge, 1981b). In such an analysis one aim is to specify a prior distribution on the space of all possible distribution functions. Difficulties arise when we attempt to calculate a marginal likelihood by integrating out the Dirichlet or gamma process. The corresponding joint marginal distribution of survival times is not absolutely continuous everywhere and so a joint marginal density is not satisfactorily defined for certain values of the lifetime variables—in particular there is a positive probability of ties occurring. Typically, different results are obtained if we assume that two tied observations are merely "close" rather than "exactly equal." An arbitrary breaking of the ties, while tempting, can cause difficulties as indicated in Burridge (1981b). A grouped data formulation of the problem appears to circumvent, or at least reduce, the problem—essentially because we no longer have to make an arbitrary decision about whether points coincide or are just near to each other. The main obstacle to the grouped data analysis is the usual computational one of integration.

Finally, I would like to endorse Heitjan's comment that Bayesian methods do not encounter special difficulties when dealing with grouped data—the main practical difficulty with both is the apparent frequent need to evaluate high dimensional integrals. When we have efficient and reliable routines for the latter I am sure we will see a rapid development of packages offering "exact" Bayesian and grouped data analyses. In the meantime we will often have to resort to Sheppard's type corrections of the sort reviewed in Heitjan's paper.

ADDITIONAL REFERENCES

BURRIDGE, J. (1981b). Empirical Bayes analysis of survival time data. *J. Roy. Statist. Soc. Ser. B* **43** 65–75.

CHENG, R. C. H. and AMIN, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. Roy. Statist. Soc. Ser. B* **45** 394–403.

CHENG, R. C. H. and ILES, T. C. (1987). Corrected maximum likelihood in non-regular problems. *J. Roy. Statist. Soc. Ser. B* **49** 95–101.

KALBFLEISCH, J. (1978). Nonparametric Bayesian analysis of survival time data. *J. Roy. Statist. Soc. Ser. B* **40** 214–221.

SILVAPULLE, M. J. and BURRIDGE, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser. B* **48** 100–106.

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90.

TITTERINGTON, D. M. (1985). Comment on "Estimating parameters in continuous univariate distributions." *J. Roy. Statist. Soc. Ser. B* **47** 115–116.