Ignorability and Coarse Data

Daniel F. Heitjan; Donald B. Rubin

Annals of Statistics, Volume 19, Issue 4 (Dec., 1991), 2244-2253.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199112%2919%3A4%3C2244%3AIACD%3E2.0.CO%3B2-J

Annals of Statistics

The Annals of Statistics
1991, Vol. 19, No. 4, 2244–2253

IGNORABILITY AND COARSE DATA

BY DANIEL F. HEITJAN[1] AND DONALD B. RUBIN[2]

Pennsylvania State University and Harvard University

We present a general statistical model for data coarsening, which includes as special cases rounded, heaped, censored, partially categorized and missing data. Formally, with coarse data, observations are made not in the sample space of the random variable of interest, but rather in its power set. Grouping is a special case in which the degree of coarsening is known and nonstochastic. We establish simple conditions under which the possible stochastic nature of the coarsening mechanism can be ignored when drawing Bayesian and likelihood inferences and thus the data can be validly treated as grouped data. The conditions are that the data be coarsened at random, a generalization of the condition missing at random, and that the parameters of the data and the coarsening process be distinct. Applications of the general model and the ignorability condition are illustrated in a numerical example and described briefly in a variety of special cases.

1. Introduction. Recent years have seen a growing interest in statistical methods that properly account for incomplete data [Little and Rubin (1987)]. The type of incompleteness that has been studied most thoroughly is missing data, in which each data value is either perfectly known or entirely unknown. Rubin (1976) states a general model for missing data that explicitly incorporates randomness in the missing data indicators and presents conditions under which inferences from data subject to missing values must take account of the mechanisms that give rise to the missing values.

In a number of common situations, however, data are neither entirely missing nor perfectly present. Instead, we observe only a subset of the complete-data sample space in which the true, unobservable data lie; we refer to this kind of incomplete data as coarse data. Coarse data arise in various ways. Perhaps the most elementary form of coarseness is rounding, which occurs when data values are observed or reported only to the nearest integer. A related problem is that of heaping, which includes the phenomenon known as digit preference. A data set is said to be heaped if it includes items reported with various levels of coarseness. For example, histograms of age often exhibit heaps at common ages such as integral multiples of ten years with adults, or integral multiples of six or twelve months with children. Another common source of coarse data is censoring, which arises in studies of failure time data when an item has not failed by the time observation of it ends. In this case the

Received February 1990; revised November 1990.
[1]Research supported by USPHS Grant HL-33407.
[2]Research supported by NSF Grant SES-88-05433.
AMS 1980 *subject classifications*. Primary 62F99; secondary 62A10, 62A15.
Key words and phrases. Censored data, coarsened at random, grouped data, heaped data, interval-censored data, missing at random, missing data, rounded data.

value for that failure time is known only to lie beyond the last point at which it was observed. Interval censoring, a close relative of grouping common in studies of cancer, occurs when units are observed at endpoints of intervals that are perhaps themselves random quantities, and failure, if it occurs, is only known to lie within those limits.

Our purpose is to present a general model that covers the principal kinds of coarsening observed in practice, and to extend Rubin's results on ignorability of the missing data mechanism to this setting. In the coarsening model that we propose, observations are not made in the sample space of the random variable of interest, but rather in its power set. Within this framework, we state conditions under which it is appropriate to ignore the possible stochastic nature of the coarsening and base inferences on the model for the underlying data and the observed coarse data, as is appropriate with grouping. In this respect, our principal result involves a generalization of the concept of missing at random [MAR, Rubin (1976)] to the case of coarse data.

2. The general theory.

2.1. *Data grouping and the resultant likelihood function.* A particularly straightforward type of data coarsening is grouping, and so we introduce notation and basic ideas in this context. Suppose a vector random variable X with sample space Ξ is distributed according to density $f(x;\theta)$ with respect to some measure, the statistician's goal being to draw inferences about θ. Also suppose that instead of observing X directly, one only observes $Y = Y(X)$, a coarse version of X that defines the subspace of Ξ into which X has fallen, without revealing the precise value of X. The sample space of Y is then 2^{Ξ}, the set of all subsets of Ξ, the power set of Ξ. We refer to this situation as one of data grouping. In practice all data are observed as coarse, that is, as falling in one of a countable number of subsets; therefore we restrict the sample space of Y to consist of the ensemble Ω of sets in the power set that have positive probability.

Under grouping as we have defined it, the random variable Y is a function of X; therefore the conditional distribution $r(y;x,\theta)$ of Y given $X = x$ is degenerate:

$$r(y;x,\theta) = \begin{cases} 1, & \text{if } y = Y(x), \\ 0, & \text{if } y \neq Y(x). \end{cases} \tag{2.1}$$

In words, the observed y is the subspace of Ξ in which X lies (i.e., $y = \{x \in \Xi: Y(x) = y\}$), $Y(x) = y$ for all x in y and $r(y;x,\theta)$ is the indicator function for the set y. Hence the likelihood arising from y, the observed value of Y, under grouping is

$$L_G(\theta;y) = \int_{\Xi} r(y;x,\theta)\, f(x;\theta)\, dx \tag{2.2}$$

or

$$L_G(\theta; y) = \int_y f(x; \theta)\, dx, \tag{2.3}$$

where dx represents integration with respect to the underlying dominating measure (typically Lebesgue or counting measure).

EXAMPLE 1. Suppose the real n-vector X is the result of i.i.d. sampling from a univariate exponential distribution with mean $1/\theta$:

$$f(x; \theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i). \tag{2.4}$$

Suppose further that $Y = Y(X)$ is the rounding set that arises when the data are reported truncated to the next smallest integer. Letting $\lfloor \cdot \rfloor$ be the greatest integer or floor function, then $Y = (Y_1, \ldots, Y_n)$, where

$$Y_i = [\lfloor X_i \rfloor, \lfloor X_i \rfloor + 1). \tag{2.5}$$

The grouped-data likelihood with observed grouped data y is (2.3) with $f(x; \theta)$ given by (2.4) and y defined from (2.5). Letting $c(y_i) = \int_{y_i} u\, du / \int_{y_i} du$ be the center of the grouping set for X_i, in this case the grouped-data likelihood can be written as

$$L_G(\theta; y) = \prod_{i=1}^{n} \left(\exp\{-\theta[c(y_i) - \tfrac{1}{2}]\} - \exp\{-\theta[c(y_i) + \tfrac{1}{2}]\}\right). \tag{2.6}$$

Likelihoods of the form L_G in (2.3) can be analytically unpleasant because of the required integration. A substitute likelihood, which entirely ignores the coarsening and thus avoids the integration, treats the interval centers $c(y) = (c(y_1), \ldots, c(y_n))$ as if they were the observed values:

$$L_A(\theta; y) = f(c(y); \theta), \tag{2.7}$$

where the subscript A indicates approximate. In Example 1, $L_A(\theta; y) = \prod_i \theta \exp(-\theta c(y_i))$. As in Example 1, L_A and L_G are generally not proportional, and thus inferences that are based on ignoring the grouping and substituting centers are not in general correct. In this sense the missing data mechanism is nonignorable [cf. Rubin (1976) and Little and Rubin (1987), Chapter 11].

The fact that it is in general incorrect to substitute interval midpoints for grouped data has long been recognized. Much early research [Sheppard (1898), Fisher (1922)] and also some later research efforts [Lindley (1950), Dempster and Rubin (1983)] have been devoted to ways of adjusting inferences based on the simple likelihood (2.7) to make them more like inferences based on the correct likelihood (2.3), at least when the grouping is not too coarse. Other efforts have considered theoretical and computational aspects of basing infer-

ences directly on (2.3) [Gjeddebaek (1949), Kulldorff (1961)]. Heitjan (1989) gives a detailed, recent review of this area. In contrast to this traditional interest, our interest is in cases where L_G may be incorrect because the grouping process is itself stochastic.

2.2. *Data coarsening.* Consider now a more general form of grouping in which the precision of reporting is a function of a random variable G with sample space Γ, whose distribution, conditional on $X = x$ and a parameter γ, is given by $h(g; x, \gamma)$. The variable G determines the precision of reporting in the sense that the value of G determines which of a collection of possible mappings $X \mapsto Y$ to use in coarsening X, where y, the observed value of $Y = Y(X, G)$, is the coarsened data. Hence, the conditional distribution of Y given $(X, G) = (x, g)$ and the parameters is degenerate:

$$r(y; x, g, \theta, \gamma) = r(y; x, g) = \begin{cases} 1, & \text{if } y = Y(x, g), \\ 0, & \text{if } y \neq Y(x, g). \end{cases} \tag{2.8}$$

We assume that the random variable G is not directly observed, but can at best only be inferred from the observed value y. In words, if $x \in y$ and y is consistent with g (which happens with probability 1), then $y = Y(x, g)$ will be observed. Because G is never directly observed, formally, neither it nor the degenerate density (2.8) is needed, but only the implied distribution of y given x indexed by γ:

$$k(y; x, \gamma) = \int_\Gamma r(y; x, g) h(g; x, \gamma)\, dg. \tag{2.9}$$

The inclusion of G is highly useful, however, for modeling the coarsening process, as illustrated in the following and subsequent examples.

EXAMPLE 2. To continue to illustrate with exponential data, suppose that the pairs (X_i, G_i), $i = 1, \ldots, n$, represent i.i.d. draws from a bivariate distribution where, as in Example 1, X_i is exponential with mean $1/\theta$ so $f(x; \theta)$ is given by (2.4). Assume G_i is binary (0-1), where the probability of being 0 is a function of x_i and a parameter $\gamma = (\gamma_1, \gamma_2)$; for concreteness, we choose this function to be $\Phi[\gamma_1 - \gamma_2 x_i]$, where $\Phi[\cdot]$ is the standard normal integral. The conditional distribution of G given $X = x$ is therefore

$$h(g; x, \gamma) = \prod_{i=1}^{n} \{\Phi[\gamma_1 - \gamma_2 x_i]\}^{1-g_i} \{1 - \Phi[\gamma_1 - \gamma_2 x_i]\}^{g_i}. \tag{2.10}$$

Suppose the coarsening function for unit i is

$$Y_i = \begin{cases} [\lfloor X_i \rfloor, \lfloor X_i \rfloor + 1), & \text{if } G_i = 0, \\ [20\lfloor X_i/20 \rfloor, 20\lfloor X_i/20 \rfloor + 20), & \text{if } G_i = 1, \end{cases} \tag{2.11}$$

which defines $r(y; x, g)$ from (2.8). In this model, g is known from y.

Although at first reading this model may seem peculiar, it can be motivated by considering the important problem of heaping in epidemiologic studies of populations of smokers. Marginal distributions of cigarettes smoked per day tend to have large heaps at integral multiples of twenty, particularly at the higher numbers of cigarettes [e.g., see Moolgavkar, Dewanji and Luebeck (1989)]. One possible explanation is that the true number of cigarettes smoked follows some discrete distribution and that those who smoke only a few cigarettes are likely to report the exact number they smoke, but that the more cigarettes one smokes, the more likely one is to report not the exact number of cigarettes but the number of cigarettes in complete packs (multiples of twenty cigarettes). This behavior is reflected in our model (2.10) and (2.11), where $G_i = 0$ indicates reporting cigarettes smoked per day and $G_i = 1$ indicates reporting cigarettes smoked per day in multiples of twenty.

2.3. *Inference with coarse data.* For coarsened data, three candidate likelihoods for inference come immediately to mind. The first ignores the coarsening altogether and uses the approximate likelihood L_A in (2.7). The second treats the coarsened data as if they were grouped data and uses likelihood L_G in (2.3). This likelihood is appealing because, although it ignores the stochastic nature of the coarsening, it does account for the grouping. In particular, using L_G allows the drawing of inferences for θ using only the density of interest $f(x;\theta)$ and the observed data y.

The third likelihood correctly accounts for the coarsening of X and the fact that the degree of coarsening is stochastic:

$$L_C(\theta,\gamma;y) = \int_y f(x;\theta)k(y;x,\gamma)\,dx, \tag{2.12}$$

where the subscript C implies both coarsened and correct. Likelihoods of type L_A in (2.7) are clearly not correct except in trivial cases, although they can be approximately correct. The question of primary interest to us concerns what types of coarsening models make the generally incorrect grouped-data likelihood L_G in (2.3) proportional to the correct likelihood L_C in (2.12), and thus a fully proper basis for inference.

An answer is provided by the concept of coarsened at random, a generalization of the notation of missing at random proposed in Rubin (1976).

DEFINITION 1. The data y are coarsened at random (CAR) if, for the fixed, observed value of y and for each value of γ, $k(y;x,\gamma)$ takes the same value for all $x \in y$, that is, for all values of x that are consistent with the observed coarse data y.

With Definition 1, and the definition of distinctness of parameters [θ and γ are a priori independent for Bayesian inference and lie in disjoint parameter spaces for likelihood-based inference; cf. Rubin (1976)], the following theorem is immediate.

THEOREM 1. *Suppose the data y are coarsened at random and the parameters θ and γ are distinct. Then*:

(i) *The likelihood ratio for θ based on the grouped-data likelihood L_G, $L_G(\theta_1; y)/L_G(\theta_2; y)$, equals the correct likelihood ratio based on L_C, $L_C(\theta_1, \gamma; y)/L_C(\theta_2, \gamma; y)$, for all γ such that $L_C(\theta_2, \gamma; y) > 0$ and all θ_1, θ_2 in the parameter space of θ.*

(ii) *The posterior distribution of θ based on L_G equals the correct posterior distribution based on L_C, and θ and γ are a posteriori independent.*

COROLLARY 1. *If g is known from y, then y is CAR if and only if $h(g; x, \gamma)$ takes the same value for all $x \in y$.*

PROOF. If g is known from the observed y, then at that y, $k = r \times h$, where $r = 1$ for all $x \in y$. □

EXAMPLE 3. In our running example with exponential data, g is known from y, and $h(g; x, \gamma)$ is given by (2.10), which for any fixed y depends on x except when $\gamma_2 = 0$, in which case it is free of x. Hence, for the general model with nonzero values of γ_2 in the parameter space, the data are never CAR, and thus likelihood or Bayesian inferences that ignore the coarsening are generally incorrect. Specifically, the incorrect likelihood L_G that accounts for the grouping but not for the stochastic coarsening can be written as

$$L_G = \prod_{i=1}^{n} \left(\exp\{-\theta[c(y_i) - 10]\} - \exp\{-\theta[c(y_i) + 10]\}\right)^{g_i} \times \left(\exp\{-\theta[c(y_i) - \tfrac{1}{2}]\} - \exp\{-\theta[c(y_i) + \tfrac{1}{2}]\}\right)^{1-g_i}, \tag{2.13}$$

since g is known from y. The correct likelihood L_C for this model is

$$L_C = \prod_{i=1}^{n} \left\{\int_{c(y_i)-10}^{c(y_i)+10} \theta \exp(-\theta x)[1 - \Phi(\gamma_1 - \gamma_2 x)]\, dx\right\}^{g_i} \times \left\{\int_{c(y_i)-1/2}^{c(y_i)+1/2} \theta \exp(-\theta x)\Phi(\gamma_1 - \gamma_2 x)\, dx\right\}^{1-g_i}. \tag{2.14}$$

Note that L_G is not generally proportional to L_C unless $\gamma_2 = 0$.

3. A numerical illustration. To explore the magnitude of the effect of using an incorrect grouped-data likelihood with coarse data, we generated one sample of $n = 1000$ observations under the model of Example 3 where $1/\theta = 5$, $\gamma_2 = 0, 0.2, 0.4$ and $\gamma_1 = \gamma_2 5 \ln 2$. For the purpose of maximizing likelihoods, we assumed the relationship $\gamma_1 = \gamma_2 5 \ln 2$ was known, and so dealt with only two unknown parameters, θ and $\gamma = \gamma_2$; this form was chosen to create plausible values for the smoking illustration described at the end of Example 2.

Likelihoods of the forms L_A and L_G were maximized over θ and L_C was maximized over (θ, γ); standard errors were based on a finite-differences

TABLE 1
ML estimates of $1/\theta$ $(= 5)$ *and* γ *using* L_A, L_G *and* L_C

Likelihood	True γ	$1/\theta$			γ		
		MLE	SE	Z	MLE	SE	Z
L_A	0	7.48	0.24	10.50	—	—	—
	0.2	6.88	0.22	8.66	—	—	—
	0.4	6.59	0.21	7.63	—	—	—
L_G	0	4.83	0.19	−0.87	—	—	—
	0.2	3.05	0.14	−14.06	—	—	—
	0.4	2.53	0.12	−21.09	—	—	—
L_C	0	4.75	0.21	−1.15	−0.008	0.011	−0.74
	0.2	4.72	0.23	−1.21	0.205	0.020	0.22
	0.4	4.76	0.22	−1.11	0.367	0.033	−1.00

estimate of the Hessian of the likelihood. Results are summarized in Table 1. The estimates based on L_A are in error by 7–11 standard errors and are far from the estimates based on L_C. Estimates and standard errors from L_G and L_C are similar when $\gamma = 0$, that is, when the data are generated under a model for which both L_G and L_C are valid. The estimates diverge, however, as γ increases away from 0, so that by $\gamma = 0.4$, the MLE of the mean, $1/\theta$, under L_G is over 21 standard errors from the true value 5, whereas the MLE under L_C is only about 1.1 standard errors away. Standard errors based on L_G also diverge from those under L_C as γ increases; by $\gamma = 0.4$, the standard error under L_G is about half the standard error under the correct model.

4. Discussion and applications. The theoretical structure described in Section 3 can be applied to a variety of situations. Here we simply indicate a few examples, some of which are pursued in detail in Heitjan (1991).

4.1. *Missing data.* Consider the special case with binary G_i, where $G_i = 1$ implies $Y_i = \{X_i\}$ and $G_i = 0$ implies Y_i is the full sample space of X_i. That is, if $G_i = 1$, X_i is observed, whereas if $G_i = 0$, X_i is missing. Missing data can be viewed as a special case of coarse data, and Rubin's (1976) condition missing at random can be viewed as a special case of coarsened at random. Specifically, g is known from y, and $h(g; x, \gamma)$ is the missing data mechanism, which is MAR if it takes the same value for all $x \in y$; by Corollary 1, this is equivalent to CAR.

4.2. *Partially classified count data.* Suppose i.i.d. X_i take the values $1, 2, 3$ with positive probabilities θ_1, θ_2 and $1 - \theta_1 - \theta_2$, respectively, and the G_i are i.i.d. and binary, with $G_i = 1$ indicating $y_i = \{X_i\}$ and $G_i = 0$ indicating $Y_i = \{1, 2\}$ if $X_i = 1$ or 2 and $Y_i = \{3\}$ if $X_i = 3$. The G_i given X_i are also independent with $h(g; x, \gamma) = \prod_{i=1}^{n} h_*(g_i; x_i, \gamma)$, where $h_*(g_i; x_i, \gamma)$ is given in Table 2, and g is known from y. Clearly, the data are CAR if either

TABLE 2
Partially classified counted data: Values of Y_i and h_* for nonzero values of $r_i \times h_*$

g_i	x_i	$y_i = Y_i(x_i, g_i)$	$h_*(g_i; x_i, \gamma)$
0	1	$\{1, 2\}$	γ_1
	2	$\{1, 2\}$	γ_2
1	1	$\{1\}$	$1 - \gamma_1$
	2	$\{2\}$	$1 - \gamma_2$
	3	$\{3\}$	1

(a) $\gamma_1 = \gamma_2$ by assumption or (b) the value $y_i = \{1, 2\}$ does not appear in the sample. Blumenthal (1968) and Hocking and Oxspring (1971, 1974) have analyzed models of this kind, although the latter authors consider only the CAR model with $\gamma_1 = \gamma_2$. Blumenthal noted that if $\gamma_1 \neq \gamma_2$, MLEs do not exist without further restrictions and that the MLE of θ_1 under the assumption $\gamma_1 = \gamma_2$ is inconsistent if $\gamma_1 \neq \gamma_2$.

4.3. *Censored data.* A variety of cases of censored data can be addressed using our structure. Two common examples are right-censored failure data with type I censoring, where the censoring times are fixed in advance, and type II censoring, where the study is stopped after the kth failure, k fixed in advance. Both types lead to CAR data sets, so that likelihood and Bayes inferences may ignore the stochastic nature of the coarsening. Kalbfleisch and Prentice (1980) reach the same conclusion, although they employ specialized arguments rather than the general concept of CAR; see also Lagakos (1979) and Heitjan (1991).

Types of censoring that may or may not lead to CAR data include censoring due to competing risks [Kalbfleisch and Prentice (1980); Cox and Oakes (1984)] and interval censoring of failure time data [Finkelstein (1986); Rücker and Messerer (1988); Self and Grossman (1986)]. Heitjan (1991) presents a detailed discussion of these situations in biomedical contexts using the general concept of CAR.

4.4. *Age heaping.* Age heaping occurs because people may report ages rounded to the nearest year (e.g., 6 years) or half-year (e.g., $4\frac{1}{2}$ years) or the the nearest month (e.g., 19 months). This situation is interesting because when the reported ages are recorded only as the number of months, the resultant observations y do not necessarily imply known values of the grouping indicator g. Let $G_i = 0$ indicate that reported age is true age truncated to the next lowest month, $G_i = 1$ indicate that reported age is true age truncated to the next lowest half-year and $G_i = 2$ indicate that reported age is true age truncated to the next lowest year. In this case, a recorded age of 24 months implies $y_i = [24, 36)$ with associated possible values of G_i in $\{0, 1, 2\}$, a recorded age of 30 months implies $y_i = [30, 36)$ with possible values of G_i in $\{0, 1\}$ and a recorded age of 25 months implies $y_i = [25, 26)$ with 0 the only possible value

of G_i. Specific models for the analysis of such data are presented in Heitjan and Rubin (1986, 1990); an analysis of heaped height data is presented in Wachter and Trussell (1982).

4.5. *Conclusions.* We have introduced a general model for describing various kinds of data coarsening and defined the concept of coarsened at random (CAR), which generalizes the notion of missing at random to more complicated incomplete data problems. The theorem states that if the data are CAR, likelihood and Bayesian inferences can ignore the stochastic nature of the coarsening and treat the observed data as if they were simple grouped data. Our example illustrates the danger in doing this when the data are not CAR. The model and concept of CAR appear to be applicable to a wide range of problems.

Acknowledgments. We thank Professor Roderick Little and two referees for many valuable suggestions.

REFERENCES

BLUMENTHAL, S. (1968). Multinomial sampling with partially categorized data. *J. Amer. Statist. Assoc.* **63** 542–551.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

DEMPSTER, A. P. and RUBIN, D. B. (1983). Rounding error in regression: The appropriateness of Sheppard's corrections. *J. Roy. Statist. Soc. Ser. B* **45** 51–59.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42** 845–854.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

GJEDDEBÆK, N. F. (1949). Contributions to the study of grouped observations. Application of the method of maximum likelihood in case of normally distributed observations. *Scand. Actuar. J.* **32** 135–159.

HEITJAN, D. F. (1989). Inference from grouped continuous data: A review (with discussion). *Statist. Sci.* **4** 164–183.

HEITJAN, D. F. (1991). Ignorability and coarse data: Some biomedical examples. Unpublished manuscript.

HEITJAN, D. F. and RUBIN, D. B. (1986). Inferences from coarse data using multiple imputation. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface* (T. J. Boardman, ed.) 138–143. Amer. Statist. Assoc., Washington, D.C.

HEITJAN, D. F. and RUBIN, D. B. (1990). Inferences from coarse data *via* multiple imputation: Age heaping in a Third World nutrition study. *J. Amer. Statist. Assoc.* **85** 304–314.

HOCKING, R. R. and OXSPRING, H. H. (1971). Maximum likelihood estimation with incomplete multinomial data. *J. Amer. Statist. Assoc.* **66** 65–70.

HOCKING, R. R. and OXSPRING, H. H. (1974). The analysis of partially categorized contingency data. *Biometrics* **30** 469–483.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KULLDORFF, G. (1961). *Contributions to the Theory of Estimation From Grouped and Partially Grouped Samples*. Almquist and Wiksell, Stockholm.

LAGAKOS, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics* **35** 139–156.

LINDLEY, D. V. (1950). Grouping corrections and maximum likelihood equations. *Proc. Cambridge Phil. Soc.* **46** 106–110.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

MOOLGAVKAR, S. H., DEWANJI, A. and LUEBECK, G. (1989). Cigarette smoking and lung cancer: Reanalysis of the British doctors' data. *Journal of the National Cancer Institute* **81** 415–420.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

RÜCKER, G. and MESSERER, D. (1988). Remission duration: An example of interval-censored observations. *Statistics in Medicine* **7** 1139–1145.

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. (3)* **29** 353–380.

WACHTER, K. W. and TRUSSELL, J. (1982). Estimating historical heights (with discussion). *J. Amer. Statist. Assoc.* **77** 279–303.

CENTER FOR BIOSTATISTICS AND EPIDEMIOLOGY
PENNSYLVANIA STATE UNIVERSITY
COLLEGE OF MEDICINE
HERSHEY, PENNSYLVANIA 17033

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138