Journal of Official Statistics, Vol. 13, No. 4, 1997, pp. 385–399

Estimating Interpolated Percentiles from Grouped Data with Large Samples

Edward L. Korn[1], *Douglas Midthune*[2], *and Barry I. Graubard*[3]

The possible values for the percentiles of a discrete distribution are the same as the possible values of the distribution itself. When data are discrete due to grouping or rounding, there is frequently interest in the percentiles of the underlying continuous distribution. Simple interpolation methods are discussed that smooth the empirical cumulative distribution function to obtain estimates of these underlying percentiles. Some limited simulations are used to examine the properties of these methods. Modifications are discussed for survey data that are obtained with complex sampling designs. Three examples are presented from the second National Health and Nutrition Examination Survey and the 1987 National Health Interview Survey.

Key words: Histograms; quantiles; smoothing; survey data analysis.

1. Introduction

Percentiles (or quantiles) offer an easy-to-interpret description of the location, spread, and tails of a distribution. The sample percentile, which is based on one or two appropriately chosen order statistics, is easy to calculate and is nonparametric, but will be less efficient than an estimator derived from a (correctly-specified) parametric model. Harrel and Davis (1982) and Kaigh and Lachenbruch (1982) suggested estimators utilizing more of the order statistics to improve the efficiency whereas Azzalini (1981) suggested estimating percentiles from a distribution function derived from a kernel density estimator; see Keating and Tripathi (1986) for a review and further references.

The problem addressed in this article is the potential bias of the sample percentile calculated from a large data set of grouped or rounded observations. For example, consider the selected sample percentiles of the blood lead distributions for boys and girls aged five years or less displayed in Table 1. These sample percentiles are estimated from data from the second National Health and Nutrition Examination Survey (NHANES II) conducted in 1976–1980 in the United States. The sample percentiles are weighted by the sample weights of the survey so that the estimates represent the U.S. population of boys and girls

[1] Head of the Clinical Trials Section, Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.
[2] Statistical Programmer with Information Management Services, Inc., Silver Spring, MD 20904, U.S.A.
[3] Mathematical Statistician with the Biometry Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.
Acknowledgments: The authors thank Karen Kafadar, David Hoaglin, two referees and an Associate Editor for their comments.

Table 1. Weighted sample and interpolated percentiles (μg/dl) of blood lead distributions for boys and girls aged ≤5 years based on data from NHANES II

Percentile	Boys (n = 1,283)		Girls (n = 1,162)	
	Sample	Interpolated	Sample	Interpolated
10	9	9.4	9	9.0
25 (lower quartile)	12	11.9	11	11.2
50 (median)	15	15.2	15	14.9
75 (upper quartile)	20	19.9	20	19.6
90	25	25.3	25	24.7

aged five years or less; see Section 4 below. Note that the sample percentiles are integers. This is because the data for blood lead levels on the public use data tape (Hematology and Biochemistry, catalog number 5411, Version 2) are recorded as integers (of *μg/dl*). Although this level of precision may be appropriate for individual values, one feels that estimates based on a thousand values as in Table 1 should have more precision. For comparison purposes, the detailed description of the blood lead levels from this survey (National Center for Health Statistics et al. 1984) displays means to the nearest 0.1 *μg/dl*. Table 1 also displays estimated interpolated percentiles that will be defined below in Section 2. These interpolated percentiles, which have been rounded to the nearest 0.1 *μg/dl*, show that the distribution of boys' lead values is shifted to the right compared to the distribution of the girls' values.

As a second example, consider Figure 1, which displays the (sample-weighted) histograms of the family income distributions of individuals living in the Northeastern versus the Southern regions of the United States. These data are also taken from NHANES II, in which sampled individuals were asked to classify their family income into one of twelve categories: under \$1000, \$1000–\$1999, \$2000–\$2999, ..., \$6000–\$6999, \$7000–\$9999, \$10,000–\$14,999, \$15,000–\$19,999, \$20,000–\$24,999, and \$25,000 and over. (In Figure 1 and in what follows, the category "\$25,000 and over" has been arbitrarily given an upper endpoint of \$35,000.) From Figure 1, we see that the distribution of incomes for the Northeastern region (sample size = 4,219) is shifted to the right compared to the distribution of incomes for the Southern region (sample size = 5,283). In contrast, the sample percentiles that utilize the midpoint of a category to represent individuals in that category show only a difference at the upper quartile (Table 2). The interpolated percentiles, which have been rounded to the nearest \$10 in Table 2, give a better description of the populations displayed in Figure 1.

In the next section we describe two methods of interpolating the sample cumulative distribution function to estimate interpolated percentiles. These are compared with each other and with the sample percentiles via some limited computer simulations. Section 3 discusses constructing confidence intervals and standard errors for interpolated percentiles. Since many large data sets are derived from surveys with complex designs, in Section 4 we discuss the added complications in using such data. We return to the examples described above in Section 4 and present a more complex example involving data from a food frequency questionnaire administered in the 1987 National Health Interview Survey. We end with a discussion of some of the inherent limitations of the proposed methods.

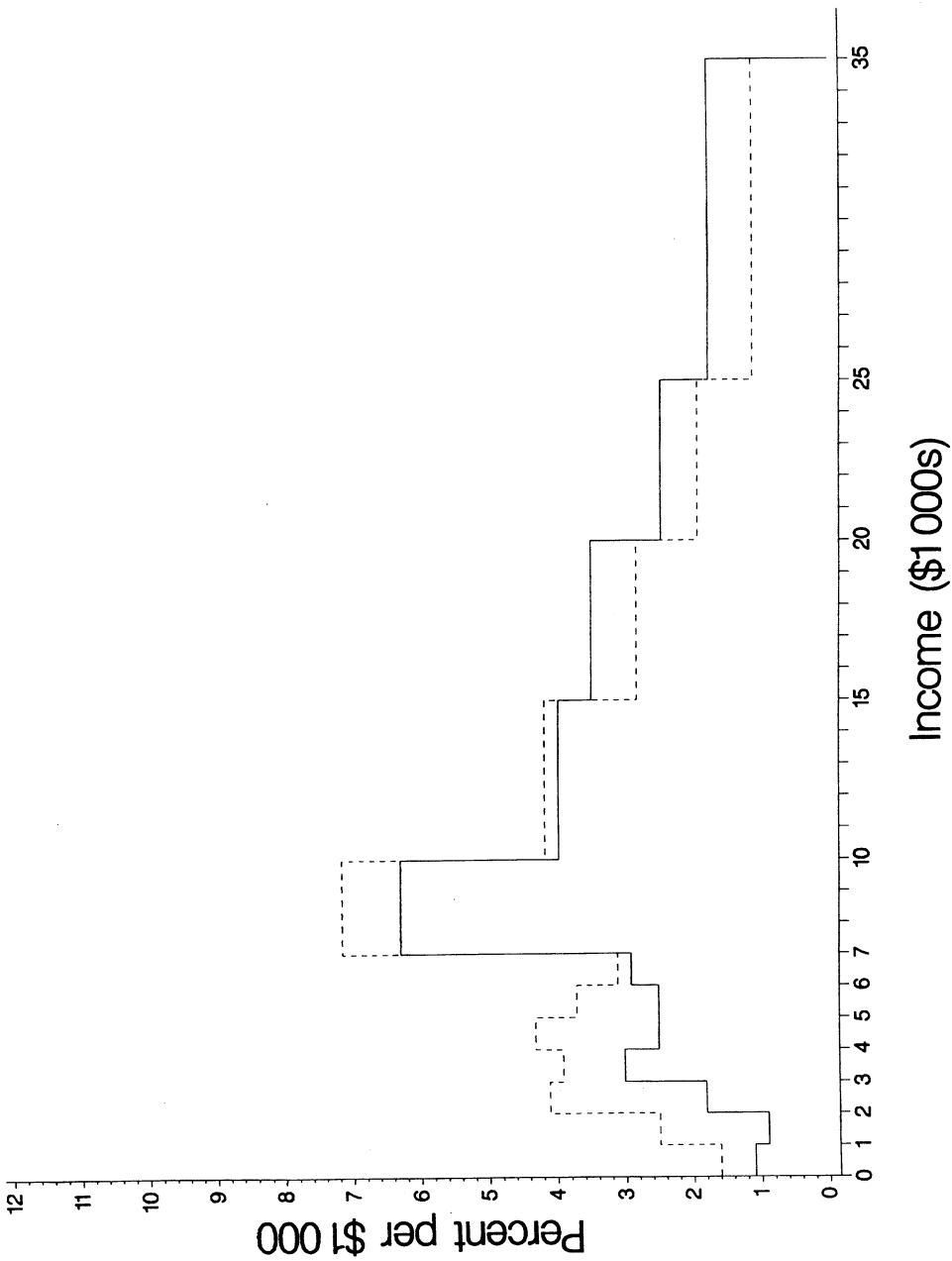

Fig. 1. Family income distributions for individuals living in the Northeastern region (solid line histogram) and the Southern region (dotted line histogram) of the United States based on data from NHANES II

Table 2. Weighted sample and interpolated percentiles (U.S. dollars) of family income distributions for the Northeastern and Southern regions of the United States based on data from NHANES II

Percentile	Northeast ($n = 4{,}219$)		South ($n = 5{,}284$)	
	Sample	Interpolated	Sample	Interpolated
25 (lower quartile)	8,500	8,740	8,500	7,300
50 (median)	12,500	14,020	12,500	11,040
75 (upper quartile)	22,500	21,680	17,500	18,180

2. Estimation

We assume that the underlying latent variable Y has a continuous cumulative distribution function (CDF), $F(y)$, but that the observed variable X can equal only the discrete values $X_1 < X_2 < \ldots < X_K$. We additionally assume that the grouping endpoints $-\infty \leq a_0 < a_1 < \ldots < a_K \leq \infty$ are given, such that if $Y \in [a_{i-1}, a_i)$ then the observation $X = X_i$. We will refer to X_i as the "midpoint" of the interval $[a_{i-1}, a_i)$, even though it may not equal $(a_{i-1} + a_i)/2$; in fact, its value can be chosen by the analyst. The target parameter is the sth percentile of the Y distribution, $F^{-1}(s)$. The observed data are $x_1, x_2, \ldots, x_n$, with associated order statistics denoted by $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. For ungrouped data, the sample sth percentile is usually defined as some linear combination of $x_{(j^*)}$ and $x_{(j^*+1)}$, where j^* is the greatest integer less than or equal to ns (SAS Institute Inc. 1990). In the present context of discrete data and large sample sizes, $x_{(j^*)}$ and $x_{(j^*+1)}$ will usually be equal, so that the particular linear combination chosen is not important. To be definite, we will define the sample sth percentile as $(x_{(j^*)} + x_{(j^*+1)})/2$ if ns is an integer, and $x_{(j^*+1)}$ otherwise; see also Schmeiser and Deutsch (1977).

Chaddock (1921) and Woodruff (1952) discussed a method for estimating a percentile with discrete data based on a linear interpolation of the empirical CDF of X as an estimator of $F(y)$. Let $p = (p_1, p_2, \ldots p_K)$ be the observed proportions of the data in the different grouping intervals. The sth percentile is estimated by $\hat{F}_L^{-1}(s; p)$, where

$$\hat{F}_L(y; p) = p_1 + p_2 + \ldots + p_{i(y)-1} + \frac{y - a_{i(y)-1}}{a_{i(y)} - a_{i(y)-1}} p_{i(y)} \tag{2.1}$$

and $i(y)$ is defined as the i such that $y \in [a_{i-1}, a_i)$. Figure 2 gives a graphical example of the linear interpolation. This interpolation can be interpreted as a distribution uniformly of the mass observed at z_i over the interval $[a_{i-1}, a_i)$. If $a_0 = -\infty$ or $a_K = \infty$, then a_0 and a_K are set to arbitrary specified finite values $a_0^* < a_1$ and $a_K^* > a_{K-1}$ in the definition (2.1).

To potentially improve upon the linear interpolation method, we consider an average quadratic interpolation. The idea is (1) to use a quadratic interpolation of the CDF for the interval containing the sample percentile and the interval to its immediate left, (2) to do the same using the interval containing the sample percentile and its immediate right-hand neighbor, (3) to average the two CDF estimates, and (4) to estimate the percentile from this average curve. A graphical example is given in Figure 3. In the real data example shown in Figure 3, the fitted quadratic curves do not differ substantially over the range considered; thus, averaging has only a minor effect on estimated percentiles in this context. Formally, we estimate the sth percentile by $\hat{F}_Q^{-1}(s; p)$ where

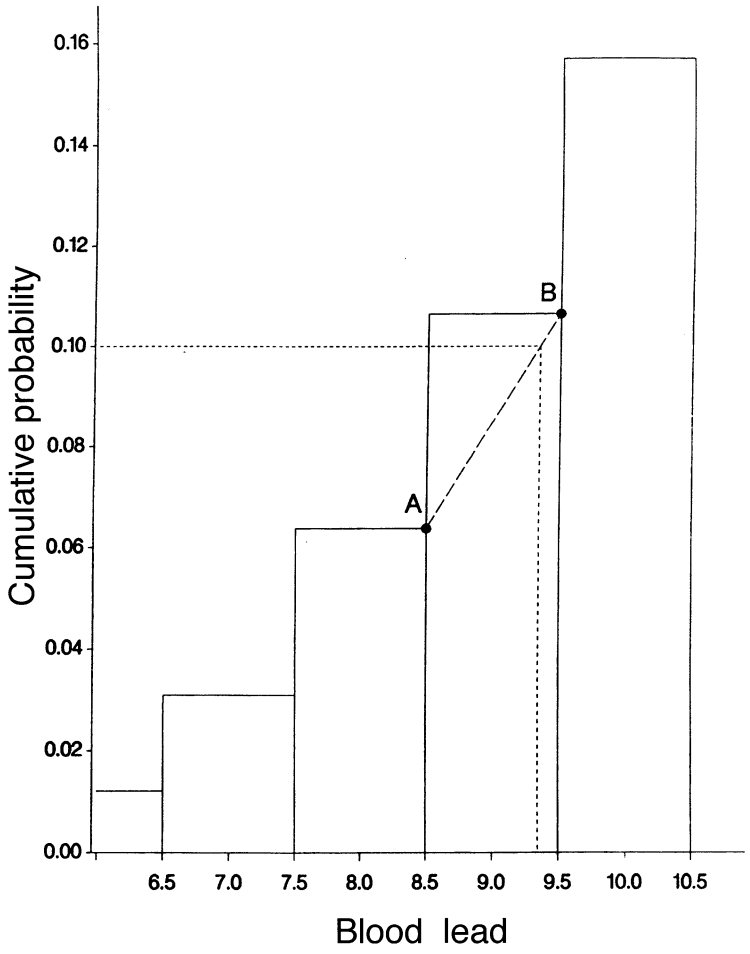

Fig. 2. Demonstration of linear interpolation for the cumulative distribution function around the 10th percentile. The straight dashed line between the points A and B is used for the interpolation. (Data are from boys sampled in NHANES II)

$\hat{F}_Q(y;p) = [\hat{F}_{Q1}(y;p) + \hat{F}_{Q2}(y;p)]/2$ and for $h = 1, 2$,

$$\hat{F}_{Qh}(y;p) = p_1 + p_2 + \ldots + p_{i(y)-1} + \int_{a_{i(y)-1}}^{y} (\alpha_{i(y)-1+h} + \beta_{i(y)-1+h}t)\,dt$$

$$= p_1 + p_2 + \ldots + p_{i(y)-1} + \alpha_{i(y)-1+h}(y - a_{i(y)-1}) + \beta_{i(y)-1+h}(y^2/2 - a^2_{i(y)-1}/2)$$

where α_j and β_j are given by the solution to

$$p_{j-1} = \int_{a_{j-2}}^{a_{j-1}} (\alpha + \beta t)\,dt = \alpha(a_{j-1} - a_{j-2}) + \beta(a^2_{j-1}/2 - a^2_{j-2}/2)$$

$$p_j = \int_{a_{j-1}}^{a_j} (\alpha + \beta t)\,dt = \alpha(a_j - a_{j-1}) + \beta(a^2_j/2 - a^2_{j-1}/2)$$

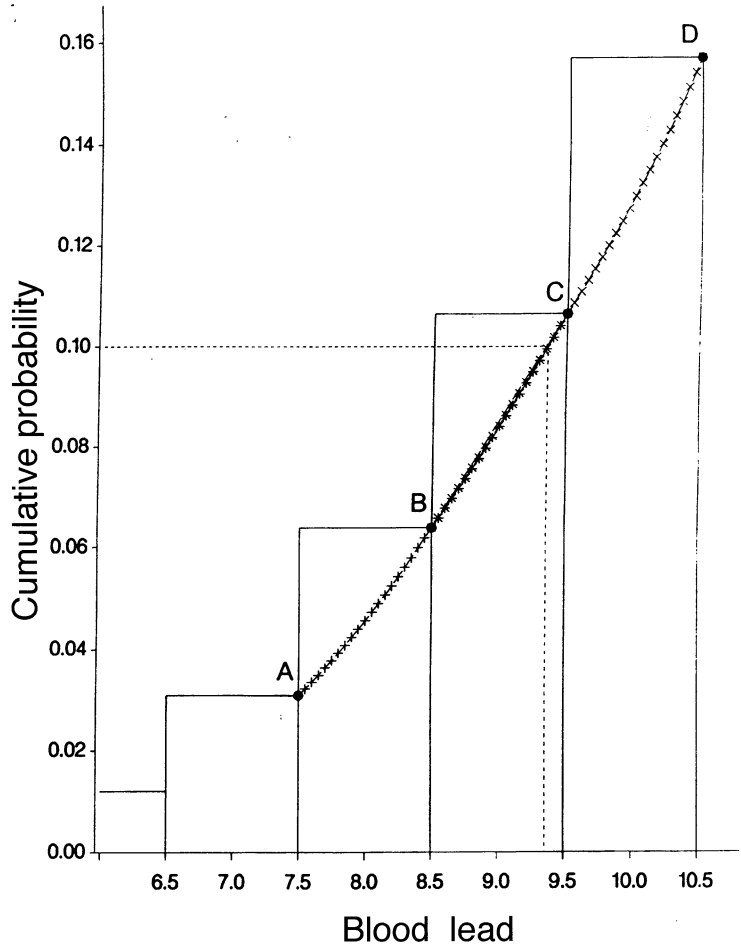

Fig. 3. Demonstration of average quadratic interpolation for the cumulative distribution function around the 10th percentile. The average of the quadratic fit through the points A, B, and C (+'s) and the quadratic fit between the points B, C, and D (x's) are used for the interpolation. (Data are from boys sampled in NHANES II)

which is

$$\beta_j = 2[p_j/(a_j - a_{j-1}) - p_{j-1}/(a_{j-1} - a_{j-2})]/(a_j - a_{j-2})$$

$$\alpha_j = p_{j-1}/(a_{j-1} - a_{j-2}) - (a_{j-2} + a_{j-1})\beta_j/2$$

When $i(y) = 1$, $\hat{F}_Q$ is defined to be $\hat{F}_{Q2}$; when $i(y) = K$ (the last interval), $\hat{F}_Q$ is defined to be $\hat{F}_{Q1}$. Note that the values of the midpoints X_i are not used for either the linear or average quadratic estimators. An alternative locally quadratic estimator of the CDF could be obtained by integrating a frequency polygon (Scott 1985). However, this would require equal-length grouping intervals, and is therefore not pursued here. An estimator based on a cubic interpolation was also considered, but did not offer an improvement over the average quadratic estimator.

To compare the behavior of the sample, linear and average quadratic percentile estimators, simulations were done with three symmetric distributions (normal, Cauchy, and uniform) and one asymmetric distribution (lognormal) for Y. The scales of the distributions

were chosen so that the difference between the 90th and 10th percentiles was always equal to 2.56 (corresponding to standard normal values), except for the uniform distribution where it was set equal to 4.8. Since the simulation results would be expected to be sensitive to the proximity of the percentile to a midpoint, a range of medians was chosen for each distribution. For the symmetric distributions, the medians were chosen to be -0.5 to 0.5 by 0.1; for the lognormal, 1.5 to 2.5 by 0.1. For each distribution, the results presented below are averaged over the eleven choices of median, with 10,000 data sets simulated for each choice of median.

Table 3 presents the simulation results for data sets of size 1,000 and grouping intervals typically of length 1.0; see the footnote to the table for the exact interval and midpoint definitions. As expected, the sample percentiles behave poorly in terms of mean squared error, and will not be discussed further. For the median, the average quadratic estimator has somewhat lower mean squared error than the linear estimator for the Cauchy and lognormal distributions, and slightly higher mean squared error for the uniform distribution. The superiority of the linear estimator for the uniform distribution is expected since this estimator is assuming the density is locally uniform. For the 10th and 90th percentiles, the average quadratic estimator is substantially better than the linear estimator for the non-uniform distributions, and somewhat worse for the uniform distribution. The slightly different results for the 10th and 90th percentiles for the symmetric distributions reflect simulation error.

Table 4 presents the corresponding results for grouping intervals typically of length 0.5. Except for the 10th percentile of the lognormal distribution, for which the average

Table 3. Simulated square root of average mean squared error of three percentile estimators for estimating median, and 10th and 90th percentiles with four underlying distributions based on data sets with $n = 1{,}000$ grouped into intervals of length 1.0[a] (see text)

	Normal	Cauchy	Uniform	Lognormal
Percentile estimator			Median	
Sample	.314	.314	.324	.315
Linear	.039	.068	.090	.056
Average quadratic	.037	.057	.093	.046
			10th percentile	
Sample	.285	.305	.280	.275
Linear	.127	.178	.057	.168
Average quadratic	.057	.140	.067	.080
			90th percentile	
Sample	.285	.305	.280	.313
Linear	.128	.178	.057	.125
Average quadratic	.057	.139	.067	.092

[a]For the symmetric distributions the grouping intervals were taken to be $(-\infty, -2.5)$, $[-2.5, -1.5)$, $[-1.5, -0.5)$, $[-0.5, 0.5)$, $[0.5, 1.5)$, $[1.5, 2.5)$, $[2.5, \infty)$, with midpoints -3, -2, -1, 0, 1, 2, and 3. When it was necessary to interpolate in the first or last interval, the infinity endpoints were taken to be -3.5 and 3.5, respectively. For the lognormal distribution, the grouping intervals were taken to be $[0, .5)$, $[.5, 1.5), \ldots, [4.5, 5.5)$, $[5.5, \infty)$ with midpoints 0.25, 1, 2, ..., 5, 6. When it was necessary to interpolate in the last interval, the infinity endpoint was taken to be 6.5.

Table 4. Simulated square root of average mean squared error of three percentile estimators for estimating median, and 10th and 90th percentiles with four underlying distributions based on data sets with n = 1,000 grouped into intervals of length 0.5[a] (see text)

	Normal	Cauchy	Uniform	Lognormal
Percentile estimator			Median	
Sample	.161	.160	.179	.161
Linear	.038	.025	.092	.037
Average quadratic	.038	.023	.094	.036
			10th percentile	
Sample	.145	.186	.153	.146
Linear	.058	.130	.053	.048
Average quadratic	.052	.132	.054	.028
			90th percentile	
Sample	.145	.186	.153	.173
Linear	.058	.130	.053	.092
Average quadratic	.052	.132	.054	.093

[a]For the symmetric distributions the grouping intervals were taken to be $(-\infty, -3.0)$, $[-3.0, -2.5), \ldots, [2.5, 3.0)$, $[3.0, \infty)$ with midpoints -3.25, $-2{,}75, \ldots$, 2.75, 3.25. When it was necessary to interpolate in the first or last interval, the infinity endpoints were taken to be -3.5 and 3.5, respectively. For the lognormal distribution, the grouping intervals were taken to be $[0, .5)$, $[.5, 1.0), \ldots, [5.5, 6.0)$, $[6.0, \infty)$ with midpoints 0.25, 0.75, 1.25, ..., 5.75, 6.25. When it was necessary to interpolate in the last interval, the infinity endpoint was taken to be 6.5.

quadratic estimator is substantially better, the differences are slight. Table 5 presents the results for the median based on grouping intervals that are of length 0.5 to the left of the median, and of length 1.0 to the right of the median. The average quadratic estimator is somewhat better for the Cauchy and lognormal distributions, with the other differences slight.

Based on these simulation results, we recommend using some form of interpolated percentile estimator. The average quadratic estimator appears generally superior to the linear estimator, and is essentially as easy to calculate.

Table 5. Simulated square root of average mean squared error of three percentile estimators for estimating median with four underlying distributions based on data sets with n = 1,000 grouped into intervals of length 0.5 to the left of 0.0 for the symmetric distributions, and to the left of 2.0 for the lognormal distribution, and length 1.0 to the right of 0.0 or 2.0, respectively[a] (see text)

	Normal	Cauchy	Uniform	Lognormal
Percentile estimator			Median	
Sample	.227	.226	.239	.227
Linear	.038	.050	.091	.049
Average quadratic	.038	.041	.093	.038

[a]For the symmetric distributions the grouping intervals were taken to be $(-\infty, -3.5)$, $[-3.5, -3.0), \ldots, [-0.5, 0)$, $[0, 1.0)$, $[2.0, 3.0)$, $[3.0, \infty)$ with midpoints -3.75, $-3.25, \ldots -0.25$, 0.5, 1.5, 2.5, 3.5. When it was necessary to interpolate in the first or last interval, the infinity endpoints were taken to be -4.0 and 4.0, respectively. For the lognormal distribution, the grouping intervals were taken to be $[0, .5)$, $[.5, 1.0)$, $[1.0, 1.5)$, $[1.5, 2.0)$, $[2.0, 3.0), \ldots, [5.0, 6.0)$, $[6.0, \infty)$ with midpoints 0.25, 0.75, 1.25, 1.75, 2.5, 3.5, ..., 5.5, 6.5. When it was necessary to interpolate in the last interval, the infinity endpoint was taken to be 7.0.

3. Confidence Intervals and Standard Errors

We describe a method of confidence interval construction following Woodruff (1952). Let $\hat{F}(y;p)$ be a continuous interpolated estimate of the distribution function $F(y)$ based on p. In particular, the linear interpolation $[\hat{F}_L(y;p)]$ or average quadratic interpolation $[\hat{F}_Q(y;p)]$ of the last section could be used. Assume that we are interested in the sth percentile, and let $\theta = F^{-1}(s)$ and $\hat{\theta} = \hat{F}^{-1}(s;p)$ be the true and estimated percentile. Let $\hat{var}[\hat{F}(y;p)]$ be an estimator of the variance of $\hat{F}(y;p)$. If $\hat{F}(\theta;p)$ were an approximately unbiased estimator of $F(\theta)$ with an approximate normal distribution, then with approximately $1-\alpha$ probability

$$F(\theta) \in \hat{F}(\theta;p) \pm z_{1-\alpha/2}\sqrt{\hat{var}[\hat{F}(\theta;p)]}$$

and therefore

$$\theta \in F^{-1}\left\{\hat{F}(\theta;p) \pm z_{1-\alpha/2}\sqrt{\hat{var}[\hat{F}(\theta;p)]}\right\}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution. An approximate $1-\alpha$ level confidence interval for θ could then be given by substituting $\hat{\theta}$ for θ and $\hat{F}^{-1}$ for F^{-1} in the above expression to yield

$$\theta \in \hat{F}^{-1}\left\{s \pm z_{1-\alpha/2}\sqrt{\hat{var}[\hat{F}(\theta;p)]_{\theta=\hat{\theta}}}\right\} \tag{3.1}$$

since $s = \hat{F}(\hat{\theta};p)$. Since both the linear and average quadratic interpolated distribution functions are linear combinations of the multinomial cell proportions p, the estimated variances $\hat{var}[\hat{F}_L(y;p)]$ and $\hat{var}[\hat{F}_Q(y;p)]$ are easy to calculate. For example, when $\hat{\theta} \in [a_2, a_3)$

$$\hat{var}\,\hat{F}_L(\theta,p)\big|_{\theta=\hat{\theta}} = \hat{var}(p_1) + \hat{var}(p_2) + 2\,\hat{cov}(p_1,p_2) + \left(\frac{\hat{\theta}-a_2}{a_3-a_2}\right)^2 \hat{var}(p_3)$$

$$+ 2\left(\frac{\hat{\theta}-a_2}{a_3-a_2}\right)[\hat{cov}(p_1,p_3) + \hat{cov}(p_2,p_3)] \tag{3.2}$$

and since p has a multinomial distribution, $\hat{var}(p_1) = p_1(1-p_1)/n$, $\hat{cov}(p_1,p_2) = -p_1p_2/n$, etc. If a standard error for $\hat{\theta}$ is required, one can use the half-width of a .6826 level confidence interval, since for a standard normal distribution this is the proportion of mass that falls within one standard deviation of the mean.

One problem with the above approach is that $\hat{F}(\theta;p)$ may not be a good estimator of $F(\theta)$ for large sample sizes. In fact, except for special $F(y)$ (e.g., a uniform distribution with linear interpolation), the interpolated distribution functions will be asymptotically biased. This is not surprising since there is no sample information about $F(\theta)$ in between grouping endpoints. An implication of the asymptotic bias is that the coverage probability of the intervals (3.1) will go to zero with increasing sample size. For example, Table 6 displays simulated coverage probabilities for the intervals (3.1) based on the linear and average quadratic interpolations using the same simulated data sets as the ones used for Table 3. For this sample size of 1,000, the coverage probabilities can be dramatically lower than the nominal value of 90 per cent.

Table 6. Simulated coverage probabilities for nominal 90% confidence intervals for the median, and 10th and 90th percentiles with four underlying distributions based on data sets with n = 1,000 grouped into intervals of length 1.0[a] (see text)

	Normal	Cauchy	Uniform	Lognormal
Percentile estimator			Median	
Linear	.88	.36	.90	.66
Average quadratic	.90	.45	.90	.77
			10th percentile	
Linear	.31	.64	.87	.14
Average quadratic	.84	.77	.82	.30
			90th percentile	
Linear	.31	.64	.87	.69
Average quadratic	.84	.77	.82	.90

[a]See the footnote to Table 3.

However, even in this setting we believe that confidence intervals still have a useful role in describing the variability of the percentile estimators. To see how well the intervals (3.1) are accomplishing this task, we examine the probability that the intervals cover $\theta_I = \hat{F}^{-1}(s; \pi)$ where $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ are the probabilities (based on $F(y)$) that an observation falls into the different grouping intervals. The parameter θ_I can be thought of as the "true" interpolated percentile; its value depends on $F(y)$ and the interpolation method. Using the same simulated intervals as described in Table 6, the simulated coverage probabilities of the true interpolated percentiles are .90 to two significant digits for all combinations listed. Thus, the confidence intervals do accurately reflect the variability of the estimator about the quantity it is estimating; we return to this point in the Discussion.

The intervals (3.1) are an example of what is referred to as a "reflected" confidence interval; other possible constructions (Slud, Byar, and Green 1984) are not pursued here. It should be noted that if one were interested in a confidence interval for the percentile of the grouped data distribution (π) instead of $F(y)$, then one could modify a standard nonparametric interval based on the inversion of a nonparametric test by closing the interval (Noether 1972).

4. Survey Data

Large data sets are frequently acquired from surveys using multistage designs involving unequal probabilities of selection and stratification. When a survey samples individuals in the population with unequal probabilities of selection, the sample weights effectively represent the number of individuals in the population that each sampled individual represents. The sample weight associated with an individual is the inverse of that individual's probability of being included in the sample, adjusted, if necessary, for non-response. There is often an additional poststratification to ensure that the sum of the sample weights equals known population values for various subgroups, e.g., age/race/sex subgroups. Weighted estimators, which are weighted by the sample weights, are

approximately unbiased for their corresponding population quantity (Kish and Frankel 1974), whereas unweighted estimators that ignore the sampling design can be badly biased. In the present application, the weighted proportions of observations in the different grouping intervals, $p_w = (p_{w1}, p_{w2}, \ldots, p_{wK})$, estimate the population proportions, and can be substituted for p in the estimators described in Section 2.

The sample design also has implications for the calculation of confidence intervals for the percentiles. Complex multistage designs can induce a correlation structure among the observations. Treating the observations as if they were from a simple random sample can therefore lead to incorrect confidence intervals. Fortunately, there are variance estimation techniques to correctly account for the sample design (Wolter 1985). In the present application for either the linear or average quadratic estimators, all that is required for the confidence interval construction of Section 3 is the estimated covariance matrix of p_w, which is readily available from standard survey computer software, e.g., Shah et al. (1991). The estimated $\hat{var}(p_{w1})$, $\hat{cov}(p_{w1}, p_{w2})$, etc., can be substituted into (3.2) to estimate the variance of $\hat{F}(\theta; p_w)$, which can be substituted into (3.1). An alternative approach is to estimate the variability of $\hat{F}(\theta; p_w)$ directly using a resampling method such as the jackknife, balanced repeated replication, or the bootstrap (Kovar, Rao, and Wu 1988). One additional modification of (3.1) is sometimes advisable. Even though the sample sizes are large, the degrees of freedom for survey variance estimators are limited by the number of primary sampling units. Therefore, in (3.1) we advise using the $1 - \alpha/2$ quantile of a t-distribution with d degrees of freedom instead of $z_{1-\alpha/2}$, where d equals the number of sampled primary sampling units (PSUs) minus the number of strata from which they are sampled (Korn and Graubard 1990). When considering estimation for a subset of the population, we recommend letting d equal the number of sampled PSUs that contain sampled observations in the subset minus the number of strata that contain observations in the subset.

We now return to the two examples presented in the Introduction. NHANES II sampled persons aged six months through 74 years. The sampling design can be approximated by the sampling of 64 PSUs from 32 strata, where the PSUs are counties or small groups of contiguous counties; see McDowell et al. (1981) for details. People living in poverty areas were oversampled, as were individuals five years or younger, or 60 years or over. The base sample weights, which were the inverses of the selection probabilities, were adjusted for nonresponse based on income and age groupings, geographic region, and whether or not the individual was within a standard metropolitan statistical area. These adjusted weights were poststratified by 76 age/race/sex categories to yield the final sample weights (National Center for Health Statistics et al. 1984). The sample percentiles presented in Tables 1 and 2 were based on weighted data using these final sample weights; the interpolated percentiles used the average quadratic interpolation based on the weighted proportions of observations in the grouping intervals.

To see the effect of the sample weighting, Table 7 displays the percentiles for blood lead for boys from Table 1 along with unweighted estimators that ignore the sample design. The confidence intervals for the interpolated percentiles are also presented using the sample design as described above, and ignoring the sample design. The estimators ignoring the sample design, i.e., the unweighted estimators, are shifted to the right. Additionally, the confidence intervals that ignore the sample design are much narrower

Table 7. Average quadratic interpolated percentiles (μg/dl) with 90% confidence intervals of blood lead distributions for boys aged ≤ 5 years (n = 1,283) based on data from NHANES II, using and ignoring the sample design

Percentile	Using the sample design		Ignoring the sample design	
	Estimate	90% con. int.	Estimate	90% con. int.
10	9.4	(8.5, 10.1)	9.7	(9.4, 10.0)
25	11.9	(11.1, 12.6)	12.2	(11.9, 12.5)
50	15.2	(14.4, 16.1)	15.7	(15.3, 16.0)
75	19.9	(18.9, 21.0)	20.3	(19.9, 20.9)
90	25.3	(23.8, 27.5)	27.0	(25.8, 28.0)

than the ones that appropriately incorporate it into the variance estimation. This is primarily due not to the sample weighting but rather to the intraclass correlation of the blood lead values at the PSU level (data not shown). Confidence intervals that ignore the sample design should not be used for this example.

We now present a more complex example that involves interpolation of CDFs on several scales. The food frequency questionnaire of the 1987 National Health Interview Survey (1987 NHIS) was administered to 22,080 persons ages 18 years or older (Block and Subar 1992). The design of the 1987 NHIS sampled 198 PSUs with oversampling of black and Hispanic respondents within certain PSUs. There was a household non-response adjustment to the base sampling weight, as well as poststratification to 60 age/race/sex categories; see National Center for Health Statistics et al. (1988) for details of the survey and a copy of the food frequency questionnaire. Individuals administered the food frequency questionnaire were asked to state how often they ate a variety of foods (during the past year) as the number of times per day, per week, per month, or per year. There was also a category "less than 6 a year or never."

To calculate (weighted) sample percentiles without interpolation, each person's answers were first converted to a weekly basis. For example, "one time per day" equals seven, "two times per week" equals two, "three times per month" equals .75, etc. Individuals reporting less consumption than six times per year or the category "less than 6 a year or never" were assigned the value 0. (Some authors, e.g., Patterson et al. (1995), calculate percentiles only for those individuals reporting more consumption than six times a year.) The sample percentiles were then calculated from the cumulative distribution function using these converted values weighted by the sample weights. These are displayed for "Potatoes, baked, boiled, or mashed" in Table 8 for the 16,841 White respondents (Hispanics excluded) with non-missing potato consumption data. The discreteness of the sample percentiles is not surprising since 72 per cent of the individuals reported their potato consumption on a "per week" basis.

To calculate interpolated percentiles, we first grouped the individuals by which time scale they use to report their consumption. Then we used average quadratic interpolation to estimate the CDF for that time scale. In doing this interpolation for the 5,812 men who reported on a "per week" basis, for example, intervals endpoints of (1.5 times per week, 2.5 times per week) were taken for a reported value of two times per week, etc. The sample weights were used in estimating each CDF. An overall interpolated CDF for men was calculated by taking the weighted average of the interpolated CDFs for each of the

Table 8. Weighted sample and interpolated percentiles (number of servings per week) of consumption of potatoes (baked, boiled, or mashed) for men and women based on data from the 1987 NHIS (Whites, Hispanics excluded)

Percentile	Men (n = 7,159)		Women (n = 9,682)	
	Sample	Interpolated (90% con. int.)	Sample	Interpolated (90% con. int.)
25 (lower quartile)	1.0	0.98 (.95, 1.02)	1.0	0.89 (.86, .92)
50 (median)	2.0	2.01 (1.96, 2.06)	2.0	1.92 (1.87, 1.97)
75 (upper quartile)	3.0	3.32 (3.26, 3.39)	3.0	3.20 (3.14, 3.26)

time scales. (A step function at zero was used for the CDF for men reporting consumption less than six times a year or never.) The weights for this weighted average were the sum of the sample weights of the individuals reporting on that time scale. For example, the sum of the sample weights of the men reporting on a "per week" basis was 73 per cent of the sum of the sample weights of all the men with usable data. The interpolated percentiles for potatoes displayed in Table 8 were obtained from the overall interpolated CDFs for men and women, and suggest that men consume more servings of potatoes than women. This latter statement is supported by the fact that men consume on average .14 more servings per week than women (p = .0017, test of means using SUDAAN).

5. Discussion

Without some assumption on the form of the underlying distribution function, there is an obvious identifiability problem in estimating percentiles from grouped data. The sample percentiles, which estimate the percentiles of the discrete distribution accumulated from the underlying continuous distribution, can be thought of as the right answer to the wrong problem. The interpolated percentiles offer an approximate answer to the right problem. To quote Tukey (1962, pp. 13–14): "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise." If one knows a parametric family containing the underlying distribution function, however, then a parametric estimation of the percentile will be asymptotically unbiased (unlike the interpolated percentiles) and more efficient. For example, suppose in the simulations in the first column of Table 3 that one knew that the underlying distribution was normal (with unknown mean and variance). One could then estimate the mean and variance using maximum likelihood for the grouped data, and use these parameter estimates to estimate the percentiles. When this was done, the simulated square root of the average mean squared error was .033 for the median and .045 for the 10th and 90th percentiles, compared to .037 and .057 using the average quadratic interpolation. The obvious risk in using a parametric family is that one might assume the wrong family. For example, if one incorrectly assumed an underlying normal distribution when it was truly uniform, the simulated errors were .109 for the 10th and 90th percentiles, compared to .067 using the average quadratic interpolation.

Even if one agrees that interpolated percentiles offer a better description of an underlying distribution than sample percentiles, one could argue with our calculation of confidence intervals. After all, the coverage probability for the underlying percentile θ may be far from the nominal coverage, even though the coverage probability for the interpolated percentile θ_I is approximately correct. However, what is the choice? We are reluctant to present percentile estimators with no measure of variability. All in all, we believe that the presentation of the confidence intervals is beneficial if one is careful in that they are interpreted as a measure of variability of a possibly biased estimator.

Although interpolated percentiles and their confidence intervals may offer a good nonparametric description of distributions, they are unneeded for certain inferential tasks. For example, consider testing the null hypothesis that the underlying distributions for two different groups of individuals are equal. If the null hypothesis were true, then the discrete grouped-data distributions would be equal also. Therefore, one could test directly the equality of the grouped-data distributions. As usual, the choice of the test statistic would depend upon the alternative hypothesis of interest. For example, one could test the equality of the blood lead distributions for boys and girls displayed in Table 1 by testing the equality of the means for a location shift alternative, or the equality of the means of the log lead values for a scale shift alternative. In either case, there would be no need for interpolation. Standard survey software could then provide a p-value as well as a standard error of the difference in means. Of course, the interpolated percentiles would still provide a useful description of the underlying distributions.

6. References

Azzalini, A. (1981). A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method. Biometrika, 68, 326–328.

Block, G. and Subar, A.F. (1992). Estimates of Nutrient Intake from a Food Frequency Questionnaire: the 1987 National Health Interview Survey. Journal of the American Dietetic Association, 92, 969–977.

Chaddock, R.E. (1921). The Graphic Representation of a Frequency Distribution. Journal of the American Statistical Association, 17, 769–775.

Harrel, F.E. and Davis, C.E. (1982). A New Distribution-Free Quantile Estimator. Biometrika, 69, 635–640.

Kaigh, W.D. and Lachenbruch, P.A. (1982). A Generalized Quantile Estimator. Communications in Statistics – Theory and Methods, 11, 2217–2238.

Keating, J.P. and Tripathi, R.C. (1985). Percentiles, Estimation of. In Encyclopedia of Statistical Sciences, Volume 6, S. Kotz and N.L. Johnson (eds.). New York: Wiley.

Kish, L. and Frankel, M.R. (1974). Inference from Complex Samples (with discussion). Journal of the Royal Statistical Society, Series B, 36, 1–37.

Korn, E.L. and Graubard, B.I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data: Use of Bonferroni t-statistics. The American Statistician, 44, 270–276.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. Canadian Journal of Statistics, 16 (supplement), 25–45.

McDowell, A., Engel, A., Massey, J.T., and Maurer, K. (1981). Plan and Operation of the

Second National Health and Nutrition Examination Survey, 1976–80. Vital and Health Statistics, Series 11, No. 15, Washington: National Center for Health Statistics.

National Center for Health Statistics, Annest, J.L., and Mahaffey, K. (1984). Blood Lead Levels for Persons Ages 6 months–74 years, United States, 1976–80. Vital and Health Statistics Series 11, No. 223 (DHHS Pub. No. PHS 84-1683). Washington: U.S. Government Printing Office.

National Center for Health Statistics, Schoenborn, C.A., and Marano, M. (1988). Current Estimates from the National Health Interview Survey: United States, 1987. Vital and Health Statistics, Series 10, No. 166 (DHHS Pub. No. PHS 88-1594). Washington: U.S. Government Printing Office.

Noether, G.E. (1972). Distribution-free Confidence Intervals. The American Statistician, 26, 39–41.

Patterson, B.H., Harlan, L.C., Block, G., and Kahle, L. (1995). Food Choices of Whites, Blacks and Hispanics: Data from the 1987 National Health Interview Survey. Nutrition and Cancer, 23, 105–119.

SAS Institute Inc. (1990). SAS Procedures Guide, Version 6, Third Edition. Cary, NC: SAS Institute Inc., 625–626.

Schmeiser, B.W. and Deutsch, S.J. (1977). Quantile Estimation from Grouped Data: The Cell Midpoint. Communications in Statistics – Simula. Computa., B6, 221–234.

Scott, D.W. (1985). Frequency Polygons: Theory and Application. Journal of the American Statistical Association, 80, 348–354.

Shah, B.V., Barnwell, B.G., Hunt, P.N., and LaVange, L.M. (1991). SUDAAN User's Manual, Research Triangle Park, NC: Research Triangle Institute.

Slud, E.V., Byar, D.P., and Green, S.B. (1984). A Comparison of Reflected versus Test-Based Confidence Intervals for the Median Survival Time, Based on Censored Data. Biometrics, 40, 587–600.

Tukey, J.W. (1962). The Future of Data Analysis. Annals of Mathematical Statistics, 33, 1–67.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer Verlag.

Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. Journal of the American Statistical Association, 47, 635–646.

Received January 1996
Revised November 1996