

Sampling errors and confidence intervals for order statistics: Implementing the family support act and welfare reform*

William C. Horrace

*University of Arizona, USA. Tel.: +1 520 621 6230; Fax: +1 520 621 8450;
E-mail: whorrace@u.arizona.edu*

Peter Schmidt

Michigan State University, USA

Ann Dryden Witte

Florida International University, Wellesley College and NBER, USA

Federal regulations recommend that states reimburse the child care costs of recipients of cash assistance and other low-income families, up to the 75th percentile of the market price of care in the relevant local market. The regulations recommend that states carry out surveys to estimate these 75th percentiles. This estimation problem raises two major statistical issues: (1) picking a sample design that will allow one to estimate the 75th percentiles cheaply, efficiently and equitably; and (2) assessing the sampling variability of the estimates obtained.

For the state of Massachusetts, we developed a sampling design that equalized the standard errors of the estimated percentiles across 65 distinct local markets. This design was selected because state administrators felt the public, child care providers and child advocates would find it equitable, thus limiting costly appeals. Estimation of standard errors for the sample 75th percentiles requires estimation of the density of the population at the 75th percentile. We implement and compare a number of parametric and nonparametric methods of density estimation. A kernel estimator appeared to provide the most reasonable estimates. On the basis of the mean integrated squared error criterion, we selected the Epanechnikov kernel and the Sheather-Jones automatic bandwidth selection procedure. Because some of our sample sizes were too small to rely on asymptotics, we also constructed nonparametric confidence intervals using the hypergeometric distribution. For most of our samples, these confidence intervals were similar to those based on the asymptotic standard errors.

Substantively, we find wide variation in the price of child care, depending on age of the child, type of care and geographic location. For full-time care, the estimated 75th percentiles ranged from \$242 per week for infants in child care centers in Boston to \$85 per week for family child care in western Massachusetts.

*The study on which this paper is based was commissioned by Massachusetts' Executive Offices of Health and Human Services. We wish to thank Janet George, Undersecretary of Health and Human Services, and Nancy Wilber, Research Director of Massachusetts' Office for Children, for their support and guidance. We would also like to thank Marie Sweeney of Workplace Solutions who managed the study and oversaw data collection. We thank Marcellus Andrews, Wellesley College, for valuable advice on solutions for simultaneous equations and Tasneem Chipty, Ohio State University, for suggestions for nonparametric confidence intervals.

1. Introduction

The Family Support Act (FSA) required states to provide child care to employed welfare recipients and individuals enrolled in federally subsidized employment and training programs. The Act allowed states to receive federal matching funding if the state's reimbursement rate for child care expenditures did not exceed the "applicable local market rate." Regulations implementing the Act set the applicable local market rate at the 75th percentile of prices for designated types of care in local areas. Under the FSA, states were required to conduct annual "market-rate surveys" to determine the 75th percentile for each child care market in the state.

On October 1, 1996, the FSA was superseded by the Personal Responsibility and Work Opportunity Act (PRWOA), popularly known as welfare reform. Regulations implementing PRWOA continue to recommend that reimbursement rates be set on the basis of market-rate surveys, although surveys are only required every two years. Under current regulations, there is a supposition that reimbursement rates that allow payments up to the 75th percentile are consistent with providing equal access to child care for families receiving child care subsidies.

These requirements raise a number of interesting statistical issues. First, they pose an interesting sample design question: *What sample design will allow one to estimate the 75th percentiles cheaply, efficiently and equitably?* Those administering the Act are interested in administrative as well as statistical efficiency. To limit appeals, reimbursement rates must be determined in a way that will be generally accepted as fair and reasonable.

Given a sample, it is simple to find its 75th percentile. However, unless we sample the entire population of child care providers, the sample 75th percentile is only an estimate of the population 75th percentile, and it is subject to sampling error like any other estimate. This raises the second relevant statistical question: *How do we assess and report the sampling variability (uncertainty) in our estimated 75th percentiles?* Here we follow and compare two different approaches from the statistical literature. First, we can adopt a nonparametric approach to construct confidence intervals for the population 75th percentile. Second, we can rely on asymptotic results that express the standard error of the sample percentile in terms of the density of the population at the percentile. This leads us to the problem of density estimation, to which we apply traditional kernel estimators. This approach gives sensible answers for samples of size (roughly) 50 or more.

These two statistical problems are obviously related, because the choice of a sampling design is motivated primarily by the desire to manage sampling variability in the estimates. We relied on preliminary data to estimate the density at the 75th percentile for each of the relevant populations (markets), so that we could express the standard error of the estimated percentiles in terms of the sample sizes chosen for each market. We then picked sample sizes, subject to a constraint on the overall number of observations and some other administratively imposed restrictions, to equalize the standard errors of the estimated 75th percentiles across different markets. This

procedure does not lead to allocating the same number of observations to each market. Rather, it allocates more observations to populations in which the 75th percentile is harder to estimate; these are markets that are larger or that have more price variability.

We addressed these statistical issues in the course of conducting the market-rate survey for the state of Massachusetts in the six-month period from December, 1993 to June, 1994. Subsequent to the completion of the survey, we explored and compared the results of using various parametric and nonparametric methods of constructing confidence intervals. We conducted a rather thorough investigation of the effects of the choice of kernel and bandwidth, and we constructed nonparametric confidence intervals using the hypergeometric distribution as well as the binomial.

The outline of the paper follows. In the next section, we describe the way in which the samples for the study were selected. In Section 3, we describe the data collection process. In Section 4, we use kernel methods of density estimation to calculate standard errors for the 75th percentiles of our samples, and in Section 5 we construct confidence intervals using these standard errors and also using a direct nonparametric approach. The final section contains our conclusions.

2. Sampling

Sampling theory most often considers the selection of samples to estimate the mean of the distribution of some variable for some well-defined population. The FSA and PRWOA imply a different sampling problem since they require estimation of a percentile rather than a mean, and since they do not clearly define the relevant populations.

Massachusetts requested that we estimate the 75th percentile of market price for thirteen different areas of the state. Each area contained a child care “resource and referral” agency (CCRA) that provides information on child care to families and other purchasers of child care in the area. Since there is a one-to-one correspondence between areas and CCRA’s we will refer interchangeably to geographical areas and CCRA’s. The state also asked that, for each area, we estimate the price for five different types of care (group care for infants, group care for toddlers, group care for preschoolers, care for school-age children, and family child care). These requirements defined 65 distinct markets (5 types of care in each of 13 different areas of the state). For convenience, we will refer to the 65 distinct markets as “cells”. Each of these 65 markets or cells defines a distinct population of interest, and we therefore needed to provide 65 different estimates of the 75th percentile of the distribution of price.

2.1. *The sampling plan*

The first statistical decision for the project was the type of sample to select. After discussion with state administrators and examination of pricing information

available from the resource and referral agencies, we decided to sample providers of child care (rather than purchasers of child care), and to choose sample sizes and sampling fractions that were unequal across markets, with random sampling within each market.

The decision to sample providers of child care rather than purchasers is motivated largely by the prevalence of child care subsidies and by considerations of convenience and cost. Subsidies are pervasive in child care and come in a bewildering variety of forms (e.g., tax deductions, direct subsidies to providers, in-kind subsidies to providers, direct subsidies to parents). The extent and nature of these subsidies makes obtaining the actual price paid by each purchaser very difficult. An attempt to elicit such detailed information would have made the questionnaire much longer and more difficult to complete, and would undoubtedly have significantly raised our nonresponse rate. Further, a sample of providers can be completed more quickly and at lower cost than a sample of purchasers. Providers' identities and locations are known by the state licensing authority (the Office for Children) whereas the universe of purchasers is unknown, and the location of purchasers requires the use of screening questions.

The decision to take a random sample of providers within each market implicitly defines the population of interest as the set of prices charged in a market (one for each provider), rather than as the set of prices paid (one for each purchaser). This is a question of interpretation of the regulations governing the federal/state Child Care Subsidy program and was decided by state administrators. The distinction between the population of prices charged and the population of prices paid will matter if providers are of different size and if size correlates with price. Compared to a random sample of purchasers, a random sample of providers clearly puts less weight on large providers and more weight on small providers. If the population of interest were defined as the set of prices paid, one for each purchaser, we could still utilize a sample of providers, but we would want to sample them with probabilities proportional to their number of customers. This would not have been easy in the case of the Massachusetts survey, however, because a provider's number of customers is not known very precisely by the state until they are sampled.¹

From the point of view of the state of Massachusetts as a whole, our sample is a stratified random sample. However, our reasons for deciding on this type of sample were not the traditional ones of minimizing the sampling errors of estimates for the state as a whole. (For excellent discussions of the usual reason for stratification and unequal sampling fractions, see Kish [12] or Kendall and Stuart [10, pp. 182–187].) For the purposes of determining reimbursement rates for the federal/state Child Care Subsidy program, the distribution of the price of child care in Massachusetts as a

¹As pointed out by a referee, we could have asked for a list of prices charged and the approximate number of children at the center who pay each price. We did not do so because we were concerned about the effect on the response rate of asking such a complicated question and about the accuracy of the information we would obtain.

whole is not of interest. Rather, the distributions of prices in local markets are relevant. Stratification was dictated by the desire to provide reasonably accurate estimates for each of our 65 cells. Some of the cells contained very few providers (sometimes less than 20) while others contained very many (more than 1500). A simple random sample over the entire state would be unlikely to contain an adequate number of providers from the smaller cells. Conversely, sampling the same number of providers within each cell would result in more precise estimates in the smaller cells than in the larger ones. Our sampling plan sought to equalize the standard error of the estimates across the 65 cells, and generally resulted in smaller cells having smaller numbers of observations but larger sampling fractions than larger cells.

The dispersion of the price of child care within cells is also relevant to the accuracy of estimation of the 75th percentile, and therefore to the sampling plan. The standard deviation of the price for a week of child care ranged from under \$10 for the part-time (less than 35 hours a week) care of preschoolers in Western Massachusetts (e.g., Pittsfield) to over \$59 for full time care of infants in group facilities (i.e., day care centers) in the western suburbs of Boston (e.g., Cambridge, Newton and Wellesley). Obviously it is harder to estimate the 75th percentile precisely when the data are more variable. Therefore, our desire to equalize the precision of our estimates across cells implied that more observations were allocated to cells with high price variability than to cells with low price variability.

The decision to allocate observations to cells in such a way as to equalize precision of estimation was made basically for administrative and political reasons. Child care providers, particularly the owners and managers of day care centers, are very well organized and well informed in Massachusetts. Large differences in the accuracy with which prices were estimated might well be perceived as unfair and inequitable. The estimated prices will be used to determine the rate at which the state will reimburse providers for care. These rates have major financial implications for providers, and a perception that rates are set unfairly would be likely to lead to large numbers of appeals, which are costly both administratively and politically.

To discuss more precisely the process by which we sought to equalize precision of estimation across cells, we define the following notation. Suppose that the size of population i is N_i , and we have data on a simple random sample (drawn without replacement) of size $n_i < N_i$, for cell $i = 1, 2, \dots, 65$. A standard formula gives the approximate variance of the estimated 75th percentile as:

$$\sigma^2(75\%ile)_i = [p(1 - p)/(n_i f^2(75\%ile)_i)][1 - n_i/N_i], \tag{1}$$

where $p = .75$, $f(75\%ile)_i$ is the population density function evaluated at the 75th percentile for cell i , and the approximation is valid for n_i sufficiently large. See, for example, Kendall and Stuart [11, pp. 251–254]; however, we multiply their formula by the finite population correction $(1 - n_i/N_i)$.

Evaluation of the variances in equation (1) requires the population density $f(75\%ile)_i$. We estimated these densities using preliminary price data provided

by the areas' resource and referral agencies (CCRA's). These data are not really representative of the population, but they provide reasonably reliable information on the variability of prices within cells. We originally attempted to estimate the densities using a nonparametric (nearest neighbors) approach, but were unable to achieve results that seemed plausible, probably because the preliminary data were not sufficiently numerous to support nonparametric density estimation. Because of time constraints (there were now less than two weeks remaining before interviews were to begin), we adopted the simple strategy of assuming normality of the populations for the purpose of density estimation. This simply required estimating the sample variance, say s_i^2 , from the preliminary data for cell i . The density of a normal population with variance σ^2 , evaluated at the 75th percentile, is given by

$$f(75\%ile) = (2\pi)^{-1/2} \exp(-.228)/\sigma. \quad (2)$$

Thus our estimate of $f(75\%ile)_i$ is proportional to $1/s_i$. The normality assumption is not really supported by our data, but later calculations based on the final data set indicate that density estimates under normality are at least moderately reliable in these data.

With estimates of $f(75\%ile)_i$ in hand, we can evaluate the variances $\sigma^2(75\%ile)_i$ as a function of the number of observations n_i in cell i , using the expression in (1) above. The state of Massachusetts initially approved funding to interview 1600 providers. We sought to allocate these to the 65 cells so as to equalize the sampling variances. This requires the solution of the set of simultaneous equations:

$$\sigma^2(75\%ile)_1 = \sigma^2(75\%ile)_2 = \dots = \sigma^2(75\%ile)_{65} \quad (3)$$

with respect to the sample sizes n_1, n_2, \dots, n_{65} , subject to the adding-up constraint $\sum_{i=1}^{65} n_i = 1600$. Of course, given that the sample sizes n_i must be integers, the equalities in (3) cannot hold exactly. We solved this problem using MATHCAD 4.0, ignoring the integer nature of the sample sizes, and then rounded the optimal sample sizes to the nearest integer. The ensuing results satisfied the adding-up constraint and the equality of variances in (3) nearly held, so we judged that we had adequately solved this problem. In particular, a more complicated integer-programming approach (minimizing a function of the deviations from equality with respect to integer sample sizes) seemed unnecessary as a practical matter.

Unsurprisingly, the sampling plan that equalized sampling variability across cells led to considerable variation in sample sizes and sampling fractions. With less than a week before field work began, the state found the degree of variability in sample sizes and sampling fractions unacceptable and requested that we impose certain restrictions on them. After some negotiation, we arrived at the final sampling plan, which required that no sample contain less than 7 providers and that sampling fractions be no less than 5 percent and no greater than 50 percent. At the last minute, the state also authorized the inclusion of another 200 providers, raising the total sample size to 1800. Imposition of the upper bound of 50 percent on the sampling

fractions led to some large values for the standard error $\sigma(75\%ile)$, and 30 of the additional 200 observations were used to allow this restriction to be relaxed for certain cells. The state asked that we increase the precision of estimates for preschoolers and we allocated 67 of the additional 200 observations to achieve this. The state wanted more precise estimates for this category because most state subsidies are for the care of preschoolers. The net result is that sampling variability was equalized only approximately across cells. In general the value of the (ex ante) standard error of the estimated 75th percentile was equalized at \$5.17 for group care for preschoolers and \$5.51 for other groups. However, the standard error was lower (as low as \$2.94) for cells where the constraints of at least 7 observations or sampling fraction of at least 5 percent were binding, and it was higher (as high as \$6.30) for cells where the sampling fractions were constrained to be only 50 percent or slightly higher. See Witte [24] for details regarding development of the sampling plan and tables that give sample sizes, sampling fractions and ex ante standard errors.

The sampling plan just described did not take into account the effects of non-response on the sample sizes and therefore on the standard errors. Since this was Massachusetts' first market rate survey, we really had no basis for estimating response rates across cells. As will be discussed in Section 3, we ultimately had a nonresponse rate of about 9%, which is high enough to affect the standard errors of the estimated 75th percentiles in a non-trivial way. If the nonresponse rate can be predicted in advance, it obviously should be taken into account in choosing the ex ante sample sizes. If the objective is equalization of the ex ante standard errors across cells, the relevant question is whether we can predict which cells are likely to have higher or lower nonresponse rates. In subsequent work done in conducting the market-rate survey for the state of Delaware, we did successfully anticipate and control for differential nonresponse.

2.2. *Selection of members of the sample*

The sampling frame for group care was Massachusetts' Office for Children's licensing list as of November 1, 1993. For family child care, the sampling frame was the licensing list as of December 20, 1993. Since the turnover rate for family child care providers is quite high, it was important to have the most recent licensing list possible. For school-age programs, the sampling frame contained both providers on the licensing lists on January 5, 1994 and license-exempt, school-based programs.

Licensing lists may not provide ideal sampling frames. However, they provide a reasonably good sampling frame in Massachusetts because both group and family child care providers are legally required to be licensed. Further, Massachusetts provides many benefits to licensed providers (e.g., courses, financial subsidies, accounting help) which makes it beneficial for the majority of providers to comply with the licensing requirements. Previous work indicates that virtually all group care providers and over 90 percent of family child care providers are licensed in Massachusetts. Unlicensed family child care providers care for a few children who

are generally related to the provider or live in the neighborhood. They often charge below market rates for child care (Diner et al. [5]) and correspondingly it is not clear that they should be included in a study such as this even if they could be sampled. A more serious problem is that school-based programs for school-age children are typically license-exempt. The state provided us with a list of license-exempt school-age providers from which to sample, and we also obtained data on licensed school-age providers. However, for school-age programs we had no reliable knowledge of the population size (N), and this affected our subsequent analysis in ways that will be described below.

The main problem we encountered with the licensing lists was obsolescence. Approximately 10 percent of group care providers and over a third of family child care providers on the licensing lists were no longer providing care. Overall, 19 percent of the originally selected sample was found to be no longer providing care. To deal with this obsolescence, we selected backup as well as regular samples. When a provider selected for the regular sample was no longer providing care, the provider was replaced by a randomly selected provider from the backup sample.

From a statistical point of view, the licensing lists provide an over inclusive sampling frame. As noted above, the licensing lists contain virtually all market-oriented child care providers in Massachusetts.² However, they contain many providers who are no longer in the market and therefore should not have appeared in the sampling frame. We could not identify these providers until we contacted them. When we learned that a provider was no longer providing care, we discarded that provider from the sample and pulled another provider at random from the backup sample, as described above. This procedure provides us with a random sample of providers within each cell. However, the obsolescence in the licensing lists means that the sampling fractions are not as originally planned. Actual sampling fractions will be relatively higher for cells containing more providers who have left the market (e.g., family child care providers) and relatively lower for cells containing few providers who have left the market (e.g., group providers). The issue is a potentially important one, but did not significantly impact the precision of the estimates for Massachusetts. Armed with the results from Massachusetts, future work could adjust population sizes by expected obsolescence rates.

3. Questionnaires and data collection

Separate questionnaires were developed for family child care and group care providers since there are marked differences between the two types of providers in size (group care providers typically care for far more children than family providers, who are generally limited to 6 or fewer children), methods of operation (family child

²Providers must obtain a license prior to providing care.

care is provided in a home while group care is provided in nonhome settings) and pricing policies (family child care providers typically quote a daily rate while group care providers typically quote weekly rates). Group care providers were asked for their basic rate for children in each age group (infants, toddlers, preschoolers and school age children). Family child care providers were asked the prices charged for up to 6 children currently in care. See Appendix G of Sweeney and Witte [20] for copies of the questionnaires.

Providers were allowed to quote rates exactly as they would to prospective clients (e.g., on an hourly, daily, weekly, monthly or annual basis). They were asked to quote both rates for full-time care (35 or more hours per week) and rates for part-time care if they provided both types of care. We believe that requesting rates in the same form as quoted to prospective clients lowered both response and nonresponse bias. It also provided valuable information regarding the pricing policies of child care providers. However, it required that we obtain information on hours and weeks of operation in order to be able to convert all prices obtained to a common time period.

Interviews were conducted between January and April, 1994 by a professional social science interview firm, the New England Research Institute, with many years of experience. Interviewers received extensive training and often had advanced degrees and/or had worked in child care or related fields. A child care specialist was available to resolve difficult issues and answer questions. Responses were recorded using Computer-Assisted Telephone Interviewing (CATI). Interviewers were prepared to provide interviews in English, Spanish and Portuguese.

Group care providers were called up to seven times and family child care providers up to five times prior to classifying a case as "no response." This number of callbacks exceeds that for the majority of studies cited by Potthoff et al. [14]. To decrease nonresponse, providers were called at varying times: days, evenings and weekends. If reached at an inconvenient time, the provider was asked to select a more convenient time and was called at that time. The overall response rates for providers in the market was 91 percent. Most nonresponse was due to failure to contact. The refusal rate was only 2 percent. There was virtually no item nonresponse.

It is common for group care providers to provide care for several of our different categories of care (infant, toddler, etc.). Once a provider has been identified and reached, the extra cost of asking additional questions is relatively low. In order to increase our sample sizes, we decided to ask any provider who was sampled to give us prices for all types of care that they provide. Thus, for example, a provider who was chosen randomly from the set of providers of group care for infants might also end up providing an observation in the samples of providers of group care for toddlers, preschoolers or school age children. This resulted in some moderate increases in our sample sizes, compared to the original sampling plan. For example, the original sampling plan called for a total of 415 observations for group care for preschoolers, and we ended up with 458 observations, even though some of the original 415 observations did not respond.

An obvious disadvantage of this approach is that the sample is no longer a simple random sample from the relevant population; we have oversampled providers of multiple categories of care. The sample within each cell can now be regarded as a stratified random sample, where the strata are providers of certain categories of care (e.g., group care for infants and toddlers, but not preschoolers or school age children). Our estimates of the population 75th percentiles should therefore in principle be calculated from observations weighted to reflect the sampling probabilities for the different strata. We did not perform such a weighting, for two reasons. First, we do not have the complete list of types of care provided by a non-sampled provider, and therefore we cannot accurately calculate the relevant sampling probabilities. Second, although we did no formal statistical tests, it did not appear that the distribution of prices charged (for a given category of care) differed according to the number of other categories of care provided.

A second problem with this approach is that these extra observations distorted our initial allocation of observations across cells, which was intended to equalize sampling variability. These extra observations could have been anticipated at the time of the original sample design, if we had had a list of the types of care provided by each provider (whether ultimately sampled or not). In subsequent work done in connection with the market-rate survey for the state of Delaware, we have in fact obtained such a list and sought to equalize ex-ante sampling variability while anticipating the multiple observations provided by providers of multiple types of care.

Since providers were allowed to quote the rate actually charged clients, the data available contained prices for diverse periods. The most common period of charge for group care providers was weekly. Family child care providers most frequently quoted daily rates. Since more children are cared for in group than in family settings, we decided to convert all prices to a weekly basis. As noted earlier, we collected sufficient information on the questionnaire regarding such things as hours and weeks of operation that these conversions could be made accurately. See Sweeney and Witte [20] for more detail.

4. Estimation of standard errors for 75th percentiles

Recall that our sampling units are “cells” defined by categories of care and geographic areas. Given the sample for a specific cell, the 75th percentile of the sample is an obvious estimator of the 75th percentile of the population, and it is the efficient estimator, in the absence of additional information, for the case of a simple random sample. The literature on quantile estimation in finite samples, which is not terribly extensive, does provide some alternative estimators. For the case of more complex designs, such as stratified random samples, more efficient estimates are available if the relevant sampling probabilities are known. Relevant references include Sedransk and Meyer [16], Francisco and Fuller [7], and the survey by Sedransk and

Smith [17]. There is also a literature on improved estimation when auxiliary information is available. References include Chambers and Dunstan [3] and Rao, Kovar and Mantel [15]. The auxiliary information takes the form of additional variables that are linearly related to the variable whose quantile we are estimating, and that are available for the entire population (sampled or non-sampled). Information on the complete list of types of care provided by each provider would be a possible example of such auxiliary information, since a linear specification of price as a function of categories of care offered might be plausible.

Table 1 gives summary statistics for the distribution of price in each of our 65 cells. Specifically, it gives the sample size (n), the population size (N), the sample mean and variance, and the sample 75th percentile. (It also gives standard errors and confidence intervals for the 75th percentiles, as will be discussed below. Also, as noted above, the population size is considered unknown and is therefore not reported for the cells corresponding to care for school age children.) A more detailed description of these results is given in Appendix D of Sweeney and Witte [20]. In general, prices of care are lower for older children than for younger children; for family child care than for group care; and for locations in the western and southeastern parts of the state than for the western suburbs of Boston. To give some idea of the numerical magnitudes involved, consider group care for preschoolers, which is generally less expensive than group care for toddlers or infants, but more expensive than family child care or care for school age children. Among the 13 geographical areas (CCRA's) considered, we find 75th percentiles of prices as low as \$101.78 and as high as \$178.00 (per week).

We now turn to the statistical issue of how to calculate standard errors and/or confidence intervals for these percentiles. In this section we will address the calculation of standard errors, with confidence intervals to be discussed in the next section. Formula (1) above gives the variance of the estimated 75th percentile, $\sigma^2(75\%ile)$, as a function of the sample size (n), the population size (N), and the density of the population evaluated at the 75th percentile, which we denoted $f(75\%ile)$. We observe the sample size and the population size, but we need to estimate $f(75\%ile)$ in order to implement the calculation in (1). (For cells corresponding to care for school-age children, with N unknown, we omit the finite population correction term $(1 - n/N)$, which is obviously equivalent to treating N as infinite.)

Density estimation is, in the words of Wegman [21, pp. 309], "an extremely popular though somewhat controversial subject." At the outset, there is a fundamental distinction between parametric and nonparametric approaches. A parametric approach assumes a particular distributional form, such as normal, for the data. This implies a specific functional form $f(x, \theta)$ for the density, where θ is the set of parameters that characterize the distribution (e.g., μ and σ^2 under normality) and x is the point of evaluation. One can then estimate the parameters θ and evaluate the density at any desired point (x), such as the 75th percentile of the distribution. We followed this procedure in Section 2 above to estimate $f(75\%ile)$ from our preliminary data, assuming normality. In general, the parametric approach is very straightforward, but there is the considerable question of how to choose the distribution. Our assumption

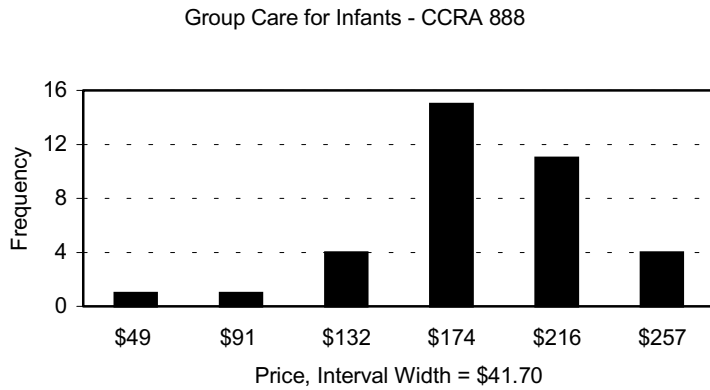


Fig. 1.

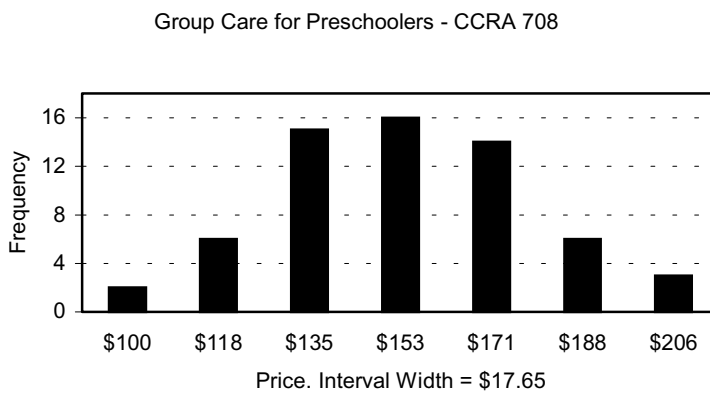


Fig. 2.

of normality was an assumption of convenience, and it was not really supported by the data. Indeed, we have 65 different data sets (cells) and it seems unlikely that any one distribution would be adequate for all or most of them. For example, prices for infant care tend to be negatively skewed; prices for preschooler care and toddler care are approximately normally distributed in some areas; and prices for family child care are often positively skewed, sometimes quite markedly. Figures 1–4 are typical histograms for the four types of care. The horizontal axis gives the prices upon which the frequency intervals are centered, and the interval widths, while the vertical axis gives frequency. We have a negatively skewed distribution in Fig. 1, approximate symmetry in Figs 2 and 3, and a positively skewed distribution in Fig. 4. It did not seem reasonable to select a single distribution for the prices observed for various types of child care in various areas of Massachusetts, nor did it seem reasonable to try to find an appropriate distribution for each of our 65 cells.

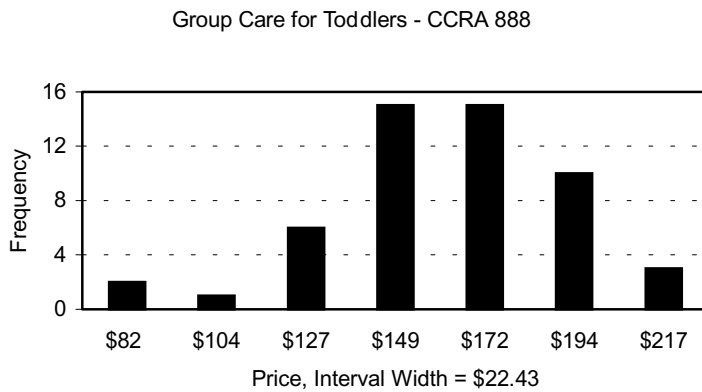


Fig. 3.

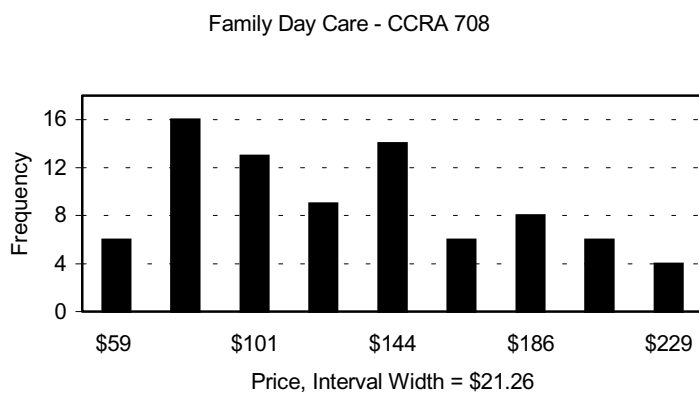


Fig. 4.

A nonparametric approach to density estimation does not make distributional assumptions, though other more or less arbitrary decisions must still be made. The most common and perhaps standard type of nonparametric density estimator is a kernel estimator. Suppose that we have a random sample X_1, X_2, \dots, X_n from a distribution with density f , and that x represents the point at which we wish to estimate f ; in our case x is the 75th percentile. Then a kernel estimator of $f(x)$ takes the general form:

$$\hat{f}(x) = n^{-1} \sum_{i=1}^n h^{-1} K[(x - X_i)/h] \quad (4)$$

where K is a kernel function satisfying $\int_{-\infty}^{\infty} K(u) du = 1$, and where h is a bandwidth parameter the choice of which will be discussed shortly. Our final results were

Table 2

Estimated densities and sampling errors for weekly prices infant care, south shore (CCRA 717)

Method	$\hat{f}(75\%ile)$	$\hat{\sigma}(75\%ile)$
Assume Normal Distribution*	0.0133	\$4.97
Gaussian Kernel, $h = 10$	0.0187	\$3.52
Gaussian Kernel, $h = 25$	0.0117	\$5.63
Gaussian Kernel, $h = 50$	0.0070	\$9.38
Gaussian Kernel, $h_{HSM} = 11.4$	0.0179	\$3.69
Gaussian Kernel, $h_{SJ} = 10.7$	0.0183	\$3.60
Epanechnikov Kernel, $h = 10$	0.0214	\$3.08
Epanechnikov Kernel, $h = 25$	0.0175	\$3.77
Epanechnikov Kernel, $h = 50$	0.0116	\$5.71
Epanechnikov Kernel, $h_{HSM} = 24.6$	0.0177	\$3.73
Epanechnikov Kernel, $h_{SJ} = 46.4$	0.0122	\$5.40

*Sample mean = \$169.76; sample standard deviation = \$24.89.

obtained using the Epanechnikov kernel:

$$K(u) = 0.75(1 - u^2) \text{ for } |u| \leq 1; K(u) = 0 \text{ for } |u| > 1. \tag{5}$$

However, we also used the Gaussian (normal) kernel:

$$K(u) = (2\pi)^{-1/2} \exp(-0.5u^2) \tag{6}$$

There are no clear rules for the choice of which kernel to use, though Epanechnikov [6] shows that his kernel is optimal in a particular sense. Ultimately, we chose the Epanechnikov kernel simply because it seemed to give results that were more sensible than those given by the Gaussian kernel.

No matter which kernel is used, the bandwidth parameter h must be chosen. Statistical theory gives only limited guidance on this point. The kernel estimate $\hat{f}(x)$ is consistent for the density $f(x)$ provided that h is chosen in such a way that $h \rightarrow 0$ but $nh \rightarrow \infty$ as $n \rightarrow \infty$. For a given sample size, these rules are not literally useful. However, it is still true that we must balance our desire for a small bandwidth, so that only observations close to x have a strong influence on $\hat{f}(x)$, with our desire for a larger bandwidth, so that our estimate of $f(x)$ is based on a large enough number of observations.

The effect of the choice of kernel and bandwidth can be seen in Table 2, in which we present estimates of the density $f(75\%ile)$ and the standard error $\sigma(75\%ile)$ of prices for infant care for the South Shore area (CCRA 717). The number of observations is $n = 21$. We use bandwidths $h = 10, 25$ and 50 with both a Gaussian and an Epanechnikov kernel. For comparison, we also present the estimates that assume a normal distribution for prices. These results make clear that the choice of kernel and bandwidth can have strong effects. For this data set, the Epanechnikov kernel results in larger estimates of $f(75\%ile)$, and therefore smaller estimates of $\sigma(75\%ile)$, than the Gaussian kernel. This is not a general result, however. For this

data set, larger bandwidths result in smaller density estimates and larger estimates of the standard error of the 75th percentile; again, this is not a general result. The prices of full-time infant care in this CCRA ranged from \$104 to \$200, with a mean of \$169.76 and 75th percentile of \$185, and were negatively skewed. At least for the Epanechnikov kernel, we can discuss the choice of bandwidth in terms of how many and which observations will enter into the estimation of the density, since the weight given to an observation X_i is nonzero if and only if its distance from x is less than h . Given that the largest observation is only \$15 larger than the point at which we wish to evaluate the density, a bandwidth much larger than \$15 would seem to be clearly too large. It would lead to the density being estimated using predominantly observations that lie below the point of evaluation, rather than symmetrically distributed around it, which is not sensible. However, choosing $h = 10$ would result in the estimate of $f(75\%ile)$ being based on only nine observations. The choice of $h = 25$ is arguably reasonable for this data set, though a slightly smaller value like $h = 15$ or $h = 20$ would also be reasonable. With $h = 25$ and the Epanechnikov kernel, we estimate $f(75\%ile)$ as 0.0175 and $\sigma(75\%ile)$ as \$3.77. Using a slightly smaller bandwidth would make the estimate of $\sigma(75\%ile)$ slightly smaller. In either case the estimate is less than the estimate of $\sigma(75\%ile)$ based on the normal distribution, which is \$4.97. The latter estimate is clearly unreliable in this case given the obvious non-normality of the data.

In the remainder of the paper we will report estimates using the Epanechnikov kernel and the all-purpose bandwidth choice $h = 25$. An attractive alternative to such an all-purpose choice, or to the use of judgement for each data set, is the use of “automatic,” data-based bandwidth selection procedures. See, e.g., Jones, Marron and Sheather [9] for a readable survey. These are procedures that yield a selection for the kernel bandwidth, h , based on the data and the optimization of some criterion function. According to Andrews [1], there are basically two types of techniques: cross-validation and plug-in methods. We will focus on plug-in methods, because they are generally regarded as superior (e.g. Park and Marron [13]). For information on cross-validation techniques, the reader is referred to Silverman [19].

The plug-in techniques we use are derived by the minimization of the mean integrated squared error (MISE) for a kernel-based estimate of an unknown density. In the case of density estimation, the MISE criterion is preferred to the usual mean squared error (MSE), because it ensures *global* accuracy of the density estimate (Silverman [19]). Additionally, the MISE is highly tractable relative to other measures of estimation discrepancy. Using the same notation as before, the asymptotic representation of the MISE as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow \infty$ is given by:

$$M(h) = n^{-1}h^{-1}R(K) + h^4\mu_2^2I_2/4 - h^6\mu_2\mu_4I_3/24 \quad (7)$$

where, for $j = 0, 1, 2, \dots$,

$$R(K) = \int \{K(x)\}^2 dx, \mu_j = \int x^j K(x) dx, I_j = \int \{f^{(j)}(x)\}^2 dx. \quad (8)$$

For the Gaussian kernel, $R(K) = 0.2821, \mu_2 = 1$ and $\mu_4 = 3$, while for the Epanechnikov kernel, $R(K) = 0.6, \mu_2 = 0.2$ and $\mu_4 = 0.0857$. The I_j unfortunately are functions of the unknown density. However, by substituting or “plugging-in” kernel-based estimates of the I_j , equation (7) can be minimized with respect to h to yield an asymptotically optimal bandwidth estimate.

We consider two such estimates of the optimal bandwidth. The first, h_{SJ} , is due to Sheather and Jones [18], while the second, h_{HSJM} , is due to Hall, Sheather, Jones and Marron [8]. These estimates differ in their rates of convergence to the optimal bandwidth: h_{SJ} converges at a rate of $n^{-5/14}$, while h_{HSJM} has a faster $n^{-1/2}$ rate. They also differ in terms of their kernel-based estimates of the I_j in equation (7) and in terms of their methods of calculation. We now summarize the calculation of each of these bandwidth estimates.

The bandwidth h_{SJ} ignores the third term of equation (7), so we only need a kernel-based estimate of I_2 . Sheather and Jones recommend a formulation based on higher order derivatives of the Gaussian kernel. The kernel used to estimate the I_j need not be the same as the kernel, $K(x)$, used for the actual density estimation. They obtain the bandwidth, h_{SJ} , that minimizes $M(h)$ as the solution to the equation:

$$[R(K)/\{\mu_2^2 S_D(\hat{\alpha}_2(h))\}]^{1/5} n^{-1/5} - h = 0 \tag{9}$$

where

$$S_D(\alpha) = \{n(n-1)\}^{-1} \alpha^{-5} \sum_{i=1}^n \sum_{j=1}^n \phi^{(4)}\{(X_i - X_j)/\alpha\}, \tag{10A}$$

$$\hat{\alpha}_2(h) = 1.357\{S_D(a)/T_D(b)\}^{1/7} h^{5/7}, \tag{10B}$$

$$T_D(b) = -\{n(n-1)\}^{-1} b^{-7} \sum_{i=1}^n \sum_{j=1}^n \phi^{(6)}\{(X_i - X_j)/b\}, \tag{10C}$$

$$a = 0.920\lambda n^{-1/7} \text{ and } b = 0.912\lambda n^{-1/9}. \tag{10D}$$

Here $\phi^{(j)}$ is the j^{th} derivative of the Gaussian kernel, λ is the sample interquartile range (the difference between the 25th and the 75th percentile prices) and $S_D(\alpha_2(h))$ is the kernel-based estimate of I_2 .

Our second estimate of the optimal bandwidth, h_{HSJM} , includes the third term of equation (7) necessitating estimation of both I_2 and I_3 . The bandwidth, h_{HSJM} , is calculated using an asymptotically equivalent expression to produce a closed-form solution:

$$h_{HSJM} = (\hat{J}_1/n)^{1/5} + \hat{J}_2(\hat{J}_1/n)^{3/5} \tag{11}$$

where

$$\hat{J}_1 = R(K)/(\mu_2^2 \hat{I}_2), \tag{12A}$$

$$\hat{J}_2 = \mu_4 I_3 / (20 \mu_2 \hat{I}_2), \tag{12B}$$

$$\hat{I}_2 = \{n(n-1)\}^{-1} \alpha^{-5} \sum_{i=1}^n \sum_{j=1}^n L^{(4)}\{(X_i - X_j)/\alpha\}, \tag{12C}$$

$$\hat{I}_3 = -\{n(n-1)\}^{-1} \beta^{-7} \sum_{i=1}^n \sum_{j=1}^n \phi^{(6)}\{(X_i - X_j)/\beta\}. \tag{12D}$$

$$\alpha = 4.29 \lambda n^{-1/11} \text{ and } \beta = 0.91 \lambda n^{-1/9}. \tag{12E}$$

Here the \hat{I}_j estimate the I_j and are not functions of h , so solution of h_{HSJM} does not require numerical analysis. Also, $L^{(4)}$ is the fourth derivative of an optimal pilot kernel selected by the authors, which is supported on the interval $(-1, 1)$:

$$L^{(4)}(x) = 135135(-184756x^{10} + 504900x^8 - 491400x^6 + 200200x^4 + 29700x^2 + 756)/16384 \tag{13}$$

In Table 2, we can see the results for these bandwidth selection procedures for the data set corresponding to infant care in the South Shore area (CCRA 717). For the Gaussian kernel, we find $h_{HSJM} = 11.4$ and $h_{SJ} = 10.7$, which imply estimated densities of approximately 0.018 and estimated standard deviations for the 75th percentile of approximately \$3.60. For the Epanechnikov kernel there is more difference across methods. We obtain $h_{HSJM} = 24.6$ and $h_{SJ} = 46.4$. These lead to density estimates of 0.0177 and 0.0122, respectively, and standard deviations of \$3.73 and \$5.40. It is not surprising or a matter of concern that the optimal bandwidth choice by a given method (e.g., h_{SJ}) should be different for different kernels (Gaussian versus Epanechnikov), but it is disappointing that different bandwidth choices should lead to substantively different results for a given kernel, as they do here for the Epanechnikov kernel.

One way to choose among kernels and bandwidths is to calculate the integrated mean square error $M(h)$ in equation (7). For the data set considered in Table 2, we obtain the following results: for the Gaussian kernel, $M(h_{HJSM}) = 0.0179$ and $M(h_{SJ}) = 0.0183$; for the Epanechnikov kernel, $M(h_{HJSM}) = 0.0177$ and $M(h_{SJ}) = 0.0077$. This would indicate the choice of the Epanechnikov kernel and h_{SJ} . Unfortunately, for the 65 data sets considered, no one procedure-kernel combination consistently resulted in the lowest value of $M(h)$. However, the combination of the Epanechnikov kernel and the Sheather-Jones bandwidth h_{SJ} most often resulted in the lowest value of $M(h)$, so that is the procedure-kernel combination for

which we will report results in the rest of the paper. The bandwidth h_{SJ} is reported in Table 1 for each of our 65 data sets. We can see that h_{SJ} is generally larger than our “all-purpose” choice $h = 25$. The use of h_{SJ} generally (but not uniformly) results in larger estimates of $\sigma(75\%ile)$ than the choice $h = 25$, though the differences in the estimates are usually not substantial.

In most cases the estimated standard error is in the range of \$3 to \$6, which is at or slightly lower than the level predicted ex ante in setting up the sampling plan. (See Section 2.1 above.) In fact, the similarity between these ex post standard errors and the ex ante standard errors calculated in formulating the sampling plan is both remarkable and comforting.

5. Construction of confidence intervals

It is a standard result that sample quantiles are distributed normally asymptotically (as $n \rightarrow \infty$). Furthermore, subject to some restrictions on the way the bandwidth is chosen, the kernel estimate $\hat{f}(75\%ile)$ is a consistent estimate of the density $f(75\%ile)$, and the variance formula (1) is correct for large n , so $\hat{\sigma}(75\%ile)$ is a consistent estimate of $\sigma(75\%ile)$. Therefore, for n large enough to rely on asymptotic results, a 95% confidence interval for the population 75th percentile is just the sample 75th percentile plus or minus 1.96 times $\hat{\sigma}(75\%ile)$. Table 1 reports these 95% confidence intervals, under the heading “Conf. Int. (kernel),” based on the Epanechnikov kernel, for both the Sheather-Jones bandwidth h_{SJ} and our all-purpose bandwidth $h = 25$. For example, for infant care in CCRA 717, the 95% confidence interval for $h = 25$ is $\$185 \pm (1.96)(\$3.77) = [\$177.61, \$192.39]$.

As is usual, the degree to which one can rely on asymptotic properties of statistical procedures depends both on the sample size and on distributional characteristics of the data. Our sample sizes and data are such that we next consider methods of constructing confidence intervals which do not rely on distributional assumptions or on asymptotics.

We first discuss a standard method based on the binomial distribution; see, e.g., Conover [4, pp. 111–116]. Let θ be the population 75th percentile, which is the parameter of interest. Given a random sample X_1, \dots, X_n taken *with replacement* from a population with this percentile, the binomial distribution yields probability statements about the number of observations that are greater than or less than θ . These probability statements can be inverted to yield confidence statements about θ , given the data.

To be more explicit, let the sample be arranged in increasing order, so that $X_1 \leq X_2 \leq \dots \leq X_n$. Let S be defined as the number of observations that are less than or equal to θ . Then the distribution of S is binomial, with number of trials n and probability of “success” of 0.75:

$$P(S = s) \equiv B(s|n, .75) = [n!/s!(n - s)!](.75)^s (.25)^{n-s}. \tag{14}$$

To find a 95% confidence interval for θ , we find integers k and l such that

$$P(k \leq S \leq l) = \sum_{j=k}^l B(j|n, .75) \geq 0.95; \quad (15)$$

then a 95% confidence interval for θ is $[X_k, X_l]$.

There will generally be different values of k and l that satisfy equation (15). One method that is often recommended is to pick k and l so as to minimize the range $(l-k)$. We choose a slightly different method. We define k to be the largest integer such that $P(S < k) \leq 0.025$ and l to be the smallest integer such that $P(S > l) \leq 0.025$. This ensures not only that $[X_k, X_l]$ is a 95% confidence interval for θ , but that X_k is a lower bound for θ with confidence level 97.5% and that X_l is an upper bound for θ with confidence level 97.5%. This seems important in the present application, since policy makers may be interested in either upper or lower bounds; someone who is primarily interested in the lower bound wants the tightest lower bound for a given confidence level, not the shortest interval.

A disadvantage of this procedure is that it fails to give an upper bound for the confidence interval when $n \leq 12$. This is so because $P(\text{all } n \text{ observations} \leq \theta) > 0.025$ for $n \leq 12$. The lower bound always exists provided $n \geq 3$. We believe that the failure to give an upper bound provides a valuable warning – avoid very small sample sizes.

Future market rate surveys may well want to constrain sample size to some minimum (e.g. 13) or more explicitly incorporate considerations regarding confidence intervals. For our Delaware work, we constrained sample sizes so that upper bounds for confidence intervals were always available.

Confidence intervals calculated using the binomial distribution are based on the assumption of sampling with replacement from an infinite size population. For the market rate survey and in many other applications, population sizes are finite and sampling is without replacement. In such circumstances, the hypergeometric distribution rather than the binomial is theoretically appropriate. The method to calculate confidence intervals using the hypergeometric distribution is a slight adaptation of Wilks [23, pp. 333].

Consider a population of size N , where J is the number of population elements less than or equal to θ , so that the number of elements greater than θ is $N - J$. The probability that $n < N$ trials without replacement from this population result in exactly S observations being less than or equal to θ is given by the hypergeometric probability:

$$\begin{aligned} P(S = s) &= H(s|n, N, J) \\ &= [J!/s!(J-s)!][(N-J)!/(n-s)!(N-J-n+s)!] \times \\ &\quad [N!/n!(N-n)]^{-1}, \end{aligned} \quad (16)$$

for $s \leq J$ and $n - s \leq N - J$. For this particular application J must be chosen so that θ is the 75th percentile, and so we let $J = \text{INT}(0.75N)$, where $\text{INT}(x)$ is the largest integer less than or equal to x .

Just as in the binomial case, to find a 95% confidence interval for θ , we find integers k and l such that

$$P(k \leq S \leq l) = \sum_{j=k}^l H(j|n, N, J) \geq 0.95; \tag{17}$$

then a 95% confidence interval for θ is $[X_k, X_l]$. Selection of k and l is as defined earlier. This procedure like the binomial procedure yields no upper bound if the sample size is too small. However, for the hypergeometric distribution, the sample size lower limit is a function of the population size. For instance, in a population of size $N = 25$ we must have $n \geq 10$ to ensure that an upper bound can be found, and for a population of size $N = 10$ we must have $n \geq 7$. Sample size requirements for the existence of an upper bound are tabulated below, for various population sizes.

Population size (N)	Minimum sample size (n)
104- ∞	13
44-103	12
28-43	11
20-27	10
15-19	9
12-14	8
10-11	7
8-9	6
7	5
6	4

Notice that $N = \infty$ requires $n \geq 13$ which coincides with the analogous binomial result.

Our confidence intervals for the 75th percentile based on the binomial and hypergeometric distributions are given in the last two columns of Table 1. The hypergeometric confidence intervals must always be at least as narrow as the binomial. However, they are usually not very different, and in most cases they are relatively similar to the corresponding confidence intervals based on the kernel density estimates. For example, returning to the case of infant care in CCRA 717, the confidence intervals based on the kernel estimates are [\$174.42, \$195.58] and [\$177.61, \$192.39], for $h_{S,J} = 46.4$ and $h = 25$ respectively, while the confidence interval is [\$175, \$190] based on the binomial distribution and [\$180, \$190] based on the hypergeometric distribution. This case (cell) is more or less typical in terms of the degree of similarity between the different confidence intervals. However, in some cases the confidence intervals based on different methods are rather different. Most of these fit one of two patterns. First, there are cases in which the sample size is rather small (e.g., group care for infants, CCRA 701, $n = 7$). With very small sample sizes, we should

expect differences in technique to have large effects, and this is an argument for larger sample sizes. Second, there are cases in which the sample size is close to the population size (e.g., group care for toddlers, CCRA 702, $n = 32$, $N = 39$). Here the finite population correction in the kernel standard errors makes a large difference, and so does the use of the hypergeometric distribution rather than the binomial. In such cases the binomial distribution should clearly not be used. In fact, for our application, the only cases in which the use of the binomial distribution makes sense are those for care for school-age children, for which the hypergeometric distribution is not available, because the population size is unknown.

However, there are a few other cells where choice of technique matters considerably and that do not fit either of these patterns. For example, in family child care, CCRA 888, we have $n = 129$, $N = 816$, and yet rather different confidence intervals for the various techniques ([\\$87.74, \\$112.26] versus [\$94.26, \$105.74] versus [\$90, \$125]). This data set has a large number of observations clustered at or just below \$100, and then a gap, with only one observation between \$100 and \$125. The choice of technique will naturally matter more when the underlying distribution of prices is very unusual.

As a general rule, we prefer the nonparametric (hypergeometric) confidence intervals for small to moderate size samples (say $n < 50$), but parametric intervals appear quite reasonable for larger sample sizes. For our application, the nonparametric confidence intervals for moderately sized samples (say $15 \leq n \leq 50$) are typically asymmetric around the 75th percentile, having a wider upper than lower interval. This seems reasonable since when estimating the upper bound for the 75th percentile one uses information for the sparse upper tail of the data.

6. Concluding remarks

We believe that our work in connection with Massachusetts' market-rate survey provides useful insights for future market-rate surveys, and more generally for applied work estimating order statistics in similar settings. We have focused on the problem of the estimation of standard errors and construction of confidence intervals for the 75th percentile of market prices, because the 75th percentile is legally relevant, and because the assessment of its sampling variability is important. This is true for at least two reasons. First, some idea of the size of the likely sampling errors is necessary to derive a sensible sampling plan. We showed how to use preliminary estimates of the standard errors of the sample percentiles to develop a sampling plan that would equalize these standard errors across cells, given a constraint on the overall sample size. Second, as in virtually any estimation problem, it is hard to know how much credence to put on point estimates unless the sampling variability is known. Decision makers should want to know not just that our estimated 75th percentile of price is \$185, but also how sure we are that the population percentile lies within some reasonable range about this estimate. Massachusetts' administrators did not request

confidence intervals, and indeed had neither budget nor time to explore the issue of confidence intervals during the six months allowed for the completion of the market survey project. We subsequently used the Massachusetts data to construct confidence intervals, and successfully used our results to recommend to the state of Delaware that the construction of confidence intervals be an integral part of their market rate survey. We considered standard errors based on the asymptotic variance formula (1), with the density estimated under a parametric assumption (such as normality), and also using kernel methods. We do not recommend a parametric approach for problems similar to ours, essentially because of the problem of choosing an appropriate distribution. The kernel estimates depend on the choice of kernel and bandwidth, and an important practical question is how strong this dependence is. Our estimates were reasonably robust to the choice of kernel and bandwidth, except for the smallest cells. Both automatic bandwidth choice procedures and more judgemental methods of choosing the bandwidth seemed to work adequately; naturally, the automatic methods are easier to apply.

The real point of estimating the standard errors of the 75th percentiles was to construct confidence intervals. The confidence intervals constructed from these standard errors rely on asymptotics because the sample 75th percentile is only asymptotically normal, because the variance formula (1) is only asymptotically valid, and because we rely on the consistency of the kernel estimates of the density that appears in (1). The reliance on asymptotics is problematic in some of our smaller cells. We also considered nonparametric confidence intervals based on the hypergeometric distribution, which do not rely on asymptotics. These agreed reasonably well with the confidence intervals based on standard errors estimated using kernels, except for the smallest cells. In retrospect, in designing the sampling scheme, we should have allocated more observations to the smallest cells. Our sampling scheme allocated enough observations to each cell to allow precise estimation of the relevant standard error, but not necessarily enough to allow credible inference. We probably should have imposed a minimum sample size of 15 or 20, even for cells for which the population size was very small.

Our work also suggests that larger sample sizes or fewer cells would be beneficial in future market rate surveys. When deciding on the size of the sample to use, administrators will have to trade off the budget available for the survey and the precision of estimates that can be obtained. Similarly, when deciding on the number of distinct prices to estimate, administrators must balance their need to set reimbursement rates that differ by type of care, age of child and geographic area against the precision with which the distinct prices are estimated. Recall that the estimated ex ante standard errors for 75th percentile prices were in the range of \$5.50, and the ex post standard errors varied somewhat but were still in this ballpark. A 95% confidence interval will be roughly plus or minus two standard errors, for a width of about \$22. Since the percentile being estimated ranged from \$85 to \$242, this is typically more than plus or minus 5% error. We can reduce this error with more observations per cell,

which obviously can be done by using more observations in total, or by reducing the number of cells.

If budgetary and administrative considerations make it impossible to increase sample size or reduce the number of cells, it would still be possible to increase the precision of estimates by using regression based approaches. One might follow Tukey and use an approach that would allow us to “borrow strength” from other cells to improve precision within a given cell. We had 13 geographical region and five types of care, for a total of 65 cells, and we treated the cells independently, estimating the 75th percentile separately for each. If we had been estimating the mean price, we might have done the same (a separate sample mean for each cell), which would be equivalent to a regression of price on intercept and 64 dummy variables representing cells. We might also formulate a regression of price on intercept, 12 dummy variables representing region, and four dummy variables representing type of care. This regression model contains only 17 parameters, not 65, and would generate more precise estimates than the unrestricted model. Its results would be valid under its maintained assumption of no interactions between the effects of region and type of care. A similar formulation would be possible in our case, using a quantile regression model (e.g., Basset and Koenker [2]) to express the 75th percentile as a function of dummy variables for region and type of care, plus some interactions. We could expect improved precision, relative to separate estimation for each cell, so long as the complete set of interactions is not used. We hope to investigate this approach in future research.

We believe that we were successful in constructing reasonably precise estimates of the population 75th percentile of child care prices, and reliable measures of the precision of these estimates. As with many successful applied statistical exercises, the essential requirements were a thorough understanding of the substantive problem and the ability to find and apply a variety of different, relevant statistical techniques. None of the techniques we used involved any significant methodological innovation, but this was not a “cookbook” exercise. Techniques were combined and applied in a novel way, based on the nature of the problem.

References

- [1] Andrews, D.W., Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation, *Econometrica* **59** (1991), 817–858.
- [2] Basset, G. and R. Koenker, An Empirical Quantile Function for Linear Models with iid Errors, *Journal of the American Statistical Association* **77** (1982), 407–415.
- [3] Chambers, R.L. and R. Dunstan, Estimating Distribution Functions from Survey Data, *Biometrika* **73** (1986), 597–604.
- [4] Conover, W., *Practical Nonparametric Statistics*, 2nd ed., New York: John Wiley, 1980.
- [5] Diner, H., E. Mauser and A. Witte, *Who's Licensed in Massachusetts*, report prepared for the Massachusetts' Office for Children and the United Way of Massachusetts Bay, 1988.
- [6] Epanechnikov, V.A., *Theory of Probability and its Applications* **14** (1969), 153–158.
- [7] Francisco, C.A. and W.A. Fuller, Quantile Estimation with a Complex Survey Design, *Annals of Statistics* **19** (1991), 454–469.

- [8] Hall, P., S.J. Sheather, M.C. Jones and J.S. Marron, On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation, *Biometrika* **78** (1991), 263–269.
- [9] Jones, M.C., J.S. Marron and S.J. Sheather, A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association* forthcoming, 1995.
- [10] Kendall, M. and A. Stuart, *The Advanced Theory of Statistics*, vol. 3, London: Charles Griffin & Company, 1976.
- [11] Kendall, M. and A. Stuart, *The Advanced Theory of Statistics*, vol. 1, London: Charles Griffin & Company, 1977.
- [12] Kish, L., *Survey Sampling*, New York: John Wiley, 1965.
- [13] Park, B.U. and J.S. Marron, Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistical Association* **85** (1990), 66–72.
- [14] Potthoff, R., K. Manton and M. Woodbury, Correcting for Nonavailability Bias in Surveys by Weighting Based on Number of Callbacks, *Journal of the American Statistical Association* **88** (1993), 1197–1218.
- [15] Rao, J.N.K., J.G. Kovar and H.J. Mantel, On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information, *Biometrika* **77** (1990), 356–375.
- [16] Sedransk, J. and J. Meyer, Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling, *Journal of the Royal Statistical Society, Series B* **40** (1978), 239–252.
- [17] Sedransk, J. and P.J. Smith, Inference for Finite Population Quantiles, in: *Handbook of Statistics*, Vol. 6, P.R. Krishnaiah and C.R. Rao, eds., Elsevier Science Publishers B.V., 1988, pp. 267–289.
- [18] Sheather, S.J. and M.C. Jones, A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation, *Journal of the Royal Statistical Society* **53** (1991), 683–690.
- [19] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall, 1986.
- [20] Sweeney, M. and A. Witte, *Massachusetts Market-Rate Survey*, report to the Massachusetts’ Executive Office of Health and Human Services, 1994.
- [21] Wegman, E., Density Estimation, in: *Encyclopedia of Statistical Sciences*, vol. 2, S. Kotz and N. Johnson, eds., New York: John Wiley, 1982, pp. 308–315.
- [22] Welsh, A.H., Asymptotically Efficient Estimation of the Sparsity Function at a Point, *Statistics and Probability Letters* **6** (1988), 427–432.
- [23] Wilks, S.S., *Mathematical Statistics*, New York: John Wiley, 1962.
- [24] Witte, A., *Sample Size and Sample Selection*, Report to the Massachusetts’ Executive Offices of Health and Human Services and Massachusetts’ Office for Children, 1993.

Table 1
Percentiles, standard errors and confidence intervals

Group care for infants										
CCRA	<i>n</i>	<i>N</i>	mean	•	75%ile	Bandwidth <i>h_{SJ}</i>	Conf. Int. (kernel, <i>h_{SJ}</i>)	Conf. Int. (kernel, <i>h</i> = 25)	Conf. Int. (binomial)	Conf. Int. (hypergeometric)
701	7	21	142.72	25.78	166.44	79.4	[132.64, 200.24]	[145.72, 187.15]	[125.00, * * * *]	[125.00, * * * *]
702	23	47	162.32	29.52	176.25	60.4	[162.81, 189.69]	[165.46, 187.04]	[169.00, 200.00]	[169.00, 190.00]
704	33	64	187.52	13.42	197.75	26.2	[192.41, 203.09]	[192.51, 202.99]	[190.00, 205.00]	[190.00, 205.00]
708	27	46	214.77	59.23	241.25	66.4	[227.65, 254.85]	[232.50, 250.00]	[234.00, 262.00]	[234.00, 251.27]
709	6	21	145.30	36.15	163.50	107.1	[115.77, 211.23]	[148.48, 178.52]	[144.80, * * * *]	[144.80, * * * *]
710	28	81	194.74	30.88	215.94	49.8	[203.86, 228.02]	[207.15, 224.73]	[205.00, 225.00]	[207.00, 224.00]
714	9	14	148.79	28.44	164.88	75.4	[144.49, 185.26]	[153.47, 176.28]	[140.62, * * * *]	[160.00, 173.21]
715	11	21	143.68	19.45	153.75	51.9	[139.11, 168.39]	[142.07, 165.43]	[135.00, * * * *]	[140.00, 168.45]
716	5	14	144.00	32.09	137.50	58.7	[106.78, 168.22]	[122.22, 152.78]	[130.00, * * * *]	[130.00, * * * *]
717	21	41	169.76	24.89	185.00	46.4	[174.42, 195.58]	[177.61, 192.39]	[175.00, 190.00]	[180.00, 190.00]
720	14	18	154.13	27.30	160.00	61.3	[149.17, 170.83]	[152.01, 167.99]	[150.00, 202.00]	[147.00, 173.00]
721	23	37	172.06	28.55	185.00	53.7	[174.34, 195.66]	[179.24, 190.76]	[179.00, 198.65]	[180.00, 190.00]
888	36	66	209.46	46.78	242.49	80.4	[227.56, 257.42]	[233.37, 251.61]	[221.00, 254.27]	[225.00, 251.27]
Group care for toddlers										
CCRA	<i>n</i>	<i>N</i>	mean	•	75%ile	Bandwidth <i>h_{SJ}</i>	Conf. Int. (kernel, <i>h_{SJ}</i>)	Conf. Int. (kernel, <i>h</i> = 25)	Conf. Int. (binomial)	Conf. Int. (hypergeometric)
701	15	30	131.36	26.85	150.00	76.9	[130.81, 169.19]	[136.85, 163.15]	[130.00, 168.00]	[140.00, 155.65]
702	32	39	144.72	30.85	160.00	50.1	[153.19, 166.81]	[152.05, 167.95]	[140.00, 189.00]	[150.00, 168.13]
704	41	78	160.06	17.52	172.25	28.8	[166.96, 177.54]	[167.20, 177.30]	[165.00, 180.60]	[168.00, 178.98]
708	45	86	193.92	35.35	212.82	54.9	[203.32, 222.32]	[204.45, 221.19]	[202.54, 230.25]	[209.01, 224.48]
709	8	26	128.79	28.23	140.00	24.7	[130.18, 149.82]	[130.07, 149.93]	[135.00, * * * *]	[137.50, * * * *]
710	36	121	176.10	29.20	195.00	49.2	[183.61, 206.39]	[187.25, 202.75]	[186.00, 203.23]	[186.00, 200.00]
714	17	22	136.64	23.63	158.63	62.2	[148.22, 169.03]	[151.13, 166.12]	[137.50, 162.00]	[145.03, 160.00]
715	12	15	126.31	19.68	130.00	45.9	[121.87, 138.13]	[123.70, 136.30]	[125.00, * * * *]	[130.00, 140.00]
716	9	15	125.92	12.92	129.25	32.7	[120.08, 138.42]	[121.31, 137.19]	[120.00, * * * *]	[125.00, 131.25]
717	25	48	149.29	24.35	164.50	44.8	[154.71, 174.29]	[157.72, 171.28]	[160.00, 172.80]	[160.00, 172.00]
720	16	19	142.52	23.81	150.00	47.2	[143.15, 156.85]	[144.67, 155.33]	[140.00, 170.00]	[145.00, 150.00]
721	34	45	150.63	26.62	161.83	44.1	[155.42, 168.25]	[156.33, 167.33]	[155.00, 180.00]	[157.00, 165.00]
888	52	84	172.28	31.74	194.00	48.9	[185.91, 202.09]	[187.58, 200.42]	[185.00, 206.70]	[191.22, 206.24]

Table 1 continued
 Percentiles, standard errors and confidence intervals

Group care for preschoolers										
CCRA	<i>n</i>	<i>N</i>	mean	•	75%ile	Bandwidth <i>h_{SJ}</i>	Conf. Int. (kernel, <i>h_{SJ}</i>)	Conf. Int. (kernel, <i>h</i> = 25)	Conf. Int. (binomial)	Conf. Int. (hypergeometric)
701	22	126	100.77	15.77	111.00	34.5	[101.24, 120.76]	[102.46, 119.54]	[105.00, 115.00]	[105.00, 115.00]
702	41	162	114.90	26.95	128.04	44.0	[117.33, 138.75]	[118.00, 138.08]	[115.00, 150.00]	[121.00, 137.00]
704	46	288	124.90	13.88	134.50	23.0	[129.24, 139.76]	[129.04, 139.96]	[130.00, 140.00]	[130.00, 140.00]
708	62	233	162.42	26.49	178.00	41.6	[169.31, 186.69]	[170.27, 185.73]	[171.00, 187.85]	[171.00, 185.77]
709	14	113	88.68	17.50	101.78	46.7	[85.41, 118.41]	[89.97, 113.58]	[90.00, 110.00]	[90.00, 110.00]
710	54	281	139.21	30.46	156.60	50.2	[144.95, 168.25]	[145.32, 167.88]	[147.00, 175.00]	[149.00, 172.06]
714	25	86	115.14	19.07	126.67	40.7	[116.29, 137.01]	[118.17, 135.18]	[116.67, 138.80]	[120.00, 132.75]
715	21	71	105.46	19.91	112.81	38.3	[101.58, 124.05]	[102.86, 122.76]	[100.00, 138.57]	[100.00, 138.57]
716	11	54	109.39	14.61	115.00	36.3	[101.66, 128.34]	[103.79, 126.21]	[105.00, * * * *]	[105.00, 118.00]
717	29	150	123.97	20.44	139.25	39.7	[128.18, 150.32]	[129.82, 148.68]	[130.00, 145.73]	[130.00, 145.73]
720	19	90	127.79	20.68	130.00	27.3	[120.74, 139.26]	[120.97, 139.03]	[120.00, 160.00]	[125.00, 150.00]
721	46	159	119.51	24.79	131.00	33.5	[122.75, 139.25]	[123.27, 138.73]	[124.25, 135.90]	[124.25, 135.00]
888	68	189	134.52	30.68	152.31	37.6	[144.89, 159.73]	[145.04, 159.58]	[144.00, 160.00]	[145.00, 158.08]
Care for school age children										
CCRA	<i>n</i>	<i>N</i>	mean	•	75%ile	Bandwidth <i>h_{SJ}</i>	Conf. Int. (kernel, <i>h_{SJ}</i>)	Conf. Int. (kernel, <i>h</i> = 25)	Conf. Int. (binomial)	Conf. Int. (hypergeometric)
701	11	**	52.76	14.05	60.60	32.9	[46.17, 75.03]	[46.82, 74.38]	[45.00, * * * *]	**
702	21	**	55.81	25.86	60.00	35.6	[47.27, 72.73]	[49.01, 70.99]	[50.00, 75.10]	**
704	33	**	58.14	19.39	70.00	35.0	[59.73, 80.27]	[60.38, 79.62]	[65.63, 72.45]	**
708	36	**	59.36	21.21	73.00	41.4	[61.41, 84.59]	[62.29, 83.71]	[66.97, 85.50]	**
709	13	**	45.50	15.69	56.58	40.1	[41.11, 72.04]	[45.16, 67.99]	[50.00, 60.00]	**
710	40	**	56.53	21.32	70.00	35.4	[59.18, 80.82]	[59.36, 80.64]	[59.58, 75.00]	**
714	22	**	37.23	14.72	50.70	35.9	[38.69, 62.71]	[40.42, 60.98]	[39.26, 55.00]	**
715	18	**	45.67	15.60	45.00	19.8	[37.33, 52.67]	[36.26, 53.74]	[41.25, 80.00]	**
716	10	**	38.48	9.72	42.25	25.4	[31.51, 52.99]	[31.61, 52.89]	[37.50, * * * *]	**
717	24	**	51.63	30.70	60.00	42.6	[45.15, 74.85]	[48.74, 71.26]	[54.00, 90.00]	**
720	19	**	53.30	31.11	66.36	62.0	[43.42, 89.30]	[35.41, 97.32]	[40.00, 120.00]	**
721	41	**	53.86	15.84	64.00	27.4	[56.80, 71.20]	[57.03, 70.97]	[57.50, 70.00]	**
888	26	**	58.82	30.60	70.00	53.0	[52.61, 87.39]	[60.29, 79.71]	[65.00, 75.10]	**

Table 1 continued
 Percentiles, standard errors and confidence intervals

CCRA	n	N	mean	•	75%ile	Family child care				
						Bandwidth <i>h_{SJ}</i>	Conf. Int. (kernel, <i>h_{SJ}</i>)	Conf. Int. (kernel, <i>h</i> = 25)	Conf. Int. (binomial)	Conf. Int. (hypergeometric)
701	56	858	97.70	14.40	100.00	20.3	[94.73, 105.27]	[94.44, 105.56]	[100.00, 101.25]	[100.00, 101.25]
702	77	1228	117.83	30.32	135.00	38.0	[126.01, 143.99]	[127.07, 142.93]	[125.00, 140.00]	[125.00, 140.00]
704	100	1624	112.80	41.13	140.00	45.9	[128.19, 151.81]	[129.99, 150.01]	[125.00, 145.00]	[125.00, 145.00]
708	82	760	142.32	49.37	179.06	71.4	[160.08, 198.05]	[154.10, 204.03]	[155.00, 200.00]	[155.00, 200.00]
709	44	515	76.98	21.04	85.00	16.1	[80.73, 89.27]	[79.44, 90.56]	[84.00, 90.00]	[84.00, 90.00]
710	121	1836	124.62	37.98	150.00	46.9	[141.03, 158.97]	[142.31, 157.69]	[146.25, 157.50]	[148.75, 157.50]
714	21	645	121.36	14.23	128.75	24.8	[119.70, 137.80]	[119.70, 137.80]	[118.75, 140.00]	[118.75, 140.00]
715	48	613	110.80	34.54	122.50	41.0	[110.55, 134.45]	[111.51, 133.49]	[112.50, 130.00]	[112.50, 125.00]
716	15	287	85.65	20.08	90.00	17.0	[82.90, 97.10]	[80.90, 99.10]	[87.45, 105.00]	[87.45, 105.00]
717	43	629	128.16	31.26	150.00	56.3	[133.99, 166.01]	[132.31, 167.69]	[135.00, 170.00]	[135.00, 150.00]
720	26	391	89.01	24.23	105.56	53.8	[89.44, 121.67]	[89.98, 121.13]	[85.00, 120.00]	[85.00, 120.00]
721	62	1140	123.57	28.46	149.38	49.8	[137.79, 160.96]	[139.23, 159.52]	[137.50, 157.50]	[137.50, 157.50]
888	129	816	99.69	35.26	100.00	10.3	[87.74, 112.26]	[94.26, 105.74]	[90.00, 125.00]	[90.00, 125.00]

** Population size is unknown for providers of care for school age children. No "finite population size" corrections were made, and confidence intervals based on the hypergeometric distribution are not available.

*** 95% confidence intervals based on the binomial distribution have no upper limit when $n < 13$. Similarly, 95% confidence intervals based on the hypergeometric distribution have no upper bound when n is sufficiently small.