

Nonparametric Density Estimation for Stratified Samples

Robert Breunig*
Australian National University

26 November, 2007

Abstract

We consider a weighted, non-parametric density estimator for stratified samples. We derive the optimal bandwidth using information on within-stratum variances and means. We provide a plug-in bandwidth when all strata are normally distributed. We show that the optimal sampling scheme is stratified sampling proportional to size, irrespective of the stratum-specific densities.

Keywords: *nonparametric density estimation, bandwidth selection, stratified sampling, optimal sampling*

1 Introduction

This paper considers the properties of the non-parametric kernel density estimator when the data are drawn using a stratified sampling scheme. We have two objectives. The first is to derive the optimal window width under stratified sampling in the case where we use all the data and a single window width to estimate the density. The second is to derive the optimal sampling scheme which minimizes the integrated mean squared error of the kernel density estimator. We also provide a plug-in value for the optimal window width for the case where all strata are normally distributed.

In the case where large samples are available from each stratum, the natural approach would be to estimate stratum-specific densities and sum those using population proportions. We assume that it is not possible to do this in what follows. We assume that the distributions in the individual strata are not of independent interest. What is of interest is an estimate of the population density that uses the sampling information.

We assume that the analyst knows the stratum-specific sample sizes, the population fraction in each stratum, and stratum-specific means and variances for the variable of interest. While the analyst needs to know the sampling weights, we do not assume that the analyst knows the stratum to which an observation belongs. This set of assumptions is akin to

*2036 H.C. Coombs Building (#9), Economics Program, Research School of Social Sciences, The Australian National University, Canberra, ACT 0200 Australia; E-mail: Robert.Breunig@anu.edu.au

the standard textbook assumption of variances being known or available when deriving the optimal sampling scheme.

There are practical cases where separate estimation within each strata may be impossible. This will arise where stratum populations are small and confidentiality considerations prevent stratum identifiers from being included in the data. It may also be the case that some strata are represented by few observations and it is not possible to efficiently estimate a separate density for those strata. For example, we may have a nationally representative, stratified sample of household income from Australia where stratification is often conducted along state/territory, aboriginal/non-aboriginal, and rural/remote/very remote categories. Some strata and sub-strata end up with only a handful of observations. Special surveys (for example, focused on remote aboriginal communities) provide the necessary information about the means and variances of these strata, even though in the analysis sample we do not have enough information to estimate stratum-specific densities.

The single-bandwidth approach developed here will generally not be superior to estimating separate stratum-specific densities with a different window width optimally selected for each strata, although, as we show below, it appears to do better in certain limited circumstances. The results derived under our simplifying assumptions provide some insight into the ramifications of using a single window width on a stratified sample of data when it is not possible to estimate stratum-specific densities. We also demonstrate that using a single bandwidth which ignores the sampling information comes at a large cost in mean squared error.

2 Density Estimation and Stratified Sampling

Consider the following population model where the population are divided into M strata,

$$Y_{ij}, \quad i = 1, \dots, M \quad j = 1, \dots, N_i.$$

The total number of elements in the population is $\sum N_i = N$ and the proportion of elements in each stratum, i , is $\theta_i = \frac{N_i}{N}$. We treat the finite population within each stratum as large enough to be well approximated by a continuous density g_i , with mean μ_i and variance σ_i^2 . We assume that the first two moments exist and are finite for each stratum.

The distribution of interest is

$$f(Y) = \sum_{i=1}^M \theta_i g_i. \tag{1}$$

¹Alternatively, consider each stratum as an i.i.d. draw from a different superpopulation where characteristics of the finite population are of interest for analysis. In this case, combining the super-population parameters using the stratum population proportions produces an overall density estimate of interest.

A sample of n_i elements are independently drawn by simple random sampling with replacement from each stratum. The total sample size is $\sum n_i = n$. Sample inclusion probabilities may not be equal for all elements in the sample, however, they will be equal for all elements in the same stratum. The probability that the j -th element in the i -th stratum is included in the sample is $\pi_{ij} = \pi_i = \frac{n_i}{N_i}$.

Rosenblatt's (1956) kernel estimator for the density² in the i th stratum, based on the sample of size n_i may be written as

$$\hat{g}_i(y) = \frac{1}{hn_i} \sum_{j=1}^{n_i} K_j \quad (2)$$

where h is the window width and $K_j = K\left(\frac{y_{ij}-y}{h}\right)$ is a symmetric, kernel function which satisfies:

$$(A1) \quad (i) \int K(\psi)d\psi = 1 \quad (ii) \int \psi K(\psi)d\psi = 0 \quad (iii) \int \psi^2 K(\psi)d\psi = \gamma_2 < \infty.$$

In what follows, we assume that it is not possible or practical to generate a separate estimate of the distribution in each stratum. Using the sample data from all strata, the usual estimator for the density at a point y is

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^M \sum_{j=1}^{n_i} K\left(\frac{y_{ij}-y}{h}\right) = \sum_{i=1}^M \frac{n_i}{n} \hat{g}_i(y). \quad (3)$$

$\hat{f}(y)$ is a sample-weighted average of the stratum-specific density estimates which will not be unbiased for the parameter of interest.

$$E\hat{f}(y) = \sum_{i=1}^M \frac{n_i}{n} g_i(y) + \sum_{i=1}^M \frac{n_i}{n} bias_i(h) \quad (4)$$

where $bias_i(h)$ represent stratum-specific bias terms which will depend upon h . Usually we choose h such that $h \rightarrow 0$ as $n_i \rightarrow \infty$, therefore the $bias_i(h)$ terms will become small as the n_i become large. Even then, however, we have additional bias arising from the fact that we are implicitly weighting the stratum-specific densities by the sample proportions. It is thus clear that the density estimate, $\hat{f}(y)$, will only be asymptotically unbiased for (1) when either $\frac{n_i}{N_i} = \frac{n}{N}$ or $g_i = g$. These conditions are unlikely to be met in most surveys. The first condition is violated when sampling is disproportionate. This is often a desired trait when particular populations of interest are sampled more heavily relative to the rest of the population. Though we are interested in an overall estimate of the density, it is problematic to assume that variables of interest will be identically distributed in different strata ($g_i = g$). Ignoring dis-proportionality in the survey design or differences between strata will lead to biased estimation, even in the simple case of non-parametric kernel density estimation.

²Rosenblatt (1956) and Parzen (1962) initiated the literature on nonparametric density estimation. For a summary and subsequent developments in the case where data is independently and identically distributed (i.i.d.), see Pagan and Ullah (1999).

The obvious solution is a weighted estimator with weights proportional to the inverse of the selection probabilities ($w_{ij} \propto \pi_{ij}^{-1}$).³

$$\hat{f}_w(y) = \frac{1}{h \sum \sum w_{ij}} \sum_{i=1}^M \sum_{j=1}^{n_i} w_{ij} K\left(\frac{y_{ij} - y}{h}\right) = \sum_{i=1}^M \theta_i \hat{g}_i(y) \quad (5)$$

As noted above, however, this is not unbiased for (1) since $\hat{g}_i(y)$ is not unbiased for g_i . The bias, which depends upon the window width h , is

$$bias\left(\hat{f}_w(y)\right) = \sum_{i=1}^M \theta_i bias_i = \sum_{i=1}^M \theta_i \frac{h^2}{2} g_i''(y) \gamma_2 + o(h^2) \quad (6)$$

where the typical bias term upto $O(h^2)$ will depend on the second derivative (g_i'') of the true underlying density

Assuming that the sampling is independent between strata (which is usually the case), it is straightforward to show upto $O(\frac{1}{nh})$ that

$$\int Var\left(\hat{f}_w(y)\right) dy = \frac{1}{h} \left[\int (K(\psi))^2 d\psi \right] \sum_{i=1}^M \frac{\theta_i^2}{n_i}. \quad (7)$$

Silverman (1986) provides details of the non-stratified case for sampling with replacement. If we consider each stratum as such a sample, it is then straightforward to work out (4) through (7).

Proposition 1: If the densities of strata 1 through M are given as g_1 through g_M , the population density $f(y)$ is estimated using a kernel satisfying (A1), and a stratified sample of data is drawn independently in each stratum, then the window width which minimizes the mean-squared error of $\hat{f}_w(y)$ will be

$$h_{st} = (\gamma_2^2)^{-\frac{1}{5}} \left(\left[\int_{\psi} (K(\psi))^2 d\psi \right] \right)^{\frac{1}{5}} \left(\sum_{i=1}^M \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}} \left(\int_y \left[\sum_{i=1}^M \theta_i (g_i'') \right]^2 dy \right)^{-\frac{1}{5}}. \quad (8)$$

Proof: using (6) and (7), write the integrated mean squared error of $\hat{f}_w(y)$ as

$$\int_y \left\{ Var\left(\hat{f}_w(y)\right) + \left(bias\left(\hat{f}_w(y)\right) \right)^2 \right\} dy. \quad (9)$$

We minimize this expression with respect to h to get the result in Proposition 1. \diamond

Implementing this result in practice requires knowledge of the second derivative of the true stratum-specific densities, which are unknown. One solution to this problem is to specify a family of distributions which will allow a value to be assigned to the term $\int_y \left[\sum_{i=1}^M \theta_i (g_i'') \right]^2 dy$ in (8). For the i.i.d. case when y is normally distributed, the optimal window width is $h^* = 1.06 \sigma n^{-\frac{1}{5}}$ where σ is the standard deviation of y . h^* is commonly employed in statistics

³The second equality in (5) follows from imposing $\pi_{ij} = \pi_i = \frac{n_i}{N_i}$ and requiring $\sum \sum w_{ij} = 1$.

packages and is a frequently used starting point for other bandwidth selection techniques such as cross-validation.

Here it is natural to ask whether a similar reference window width can be derived based upon underlying normal distributions in all of the strata.

Corollary 1:⁴ If g_1 through g_M are normally distributed with mean μ_i and variance σ_i^2 and the density is estimated using a standard normal kernel, then the optimal window width (in the mean squared error sense) will be

$$h_{st} = 0.87 \left(\sum_{i=1}^M \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}} (\lambda_1 + \lambda_2)^{-\frac{1}{5}} \quad (10)$$

where λ_1 is a weighted sum of stratum-specific standard deviations

$$\lambda_1 = \frac{3}{8} \sum_{i=1}^M \theta_i^2 \sigma_i^{-5}$$

and λ_2 is a weighted sum of a function of the distance between stratum means

$$\lambda_2 = \sum_{i=1}^M \sum_{l \neq i}^M \theta_i \theta_l \frac{(\sigma_i^2 + \sigma_l^2)^{-\frac{5}{2}}}{\sqrt{2}} \left\{ 3 - 6 \frac{(\mu_i - \mu_l)^2}{(\sigma_i^2 + \sigma_l^2)} + \frac{(\mu_i - \mu_l)^4}{(\sigma_i^2 + \sigma_l^2)^2} \right\} e^{-\frac{1}{2} \frac{(\mu_i - \mu_l)^2}{(\sigma_i^2 + \sigma_l^2)}}$$

The optimal window width is inversely proportional to a weighted sum of the sample sizes, n_i . In the case where $n_i = \frac{n}{M}$ and $\theta_i = \frac{1}{M}$, then $\sum_{i=1}^M \frac{\theta_i^2}{n_i} = n$ and the window width will be proportional to $n^{-\frac{1}{5}}$ as in the i.i.d. case, but the proportionality constant will differ from 1.06σ . Only when strata share common means and variances and the population and sample proportions are equal in all strata, will this result collapse to the i.i.d. case, $h^* = 1.06\sigma n^{-\frac{1}{5}}$.

When strata share common means and variances, $h_{st} = 1.06\sigma \left(\sum_{i=1}^M \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}}$. Thus even in the case of homogeneous populations in all strata, the optimal window width differs from h^* unless $\theta_i = \frac{n_i}{n}$. This is analogous to the case of estimation of the mean, where even when all strata have identical means, the variance of the estimator \bar{y} is different for a stratified sample than for a simple random sample.

Numerical examples and simulation results⁵ show that use of the weighted estimator combined with the correct window width gives large improvements in mean squared error relative to using h^* . Large bias arises if dis-proportionately sampled data is treated as being a simple random sample. Even when data is proportionally sampled from each stratum, large efficiency gains can be realized by taking into account the sampling information.

⁴Proof available from the authors.

⁵A summary of these is given in section 3. Full details are available from the authors.

Optimal allocation

If we have some information about stratum-specific means and standard deviations, can we use that information to construct an optimal sampling allocation to minimize the integrated mean squared error of the estimator of $f(y)$? In the case of stratified sampling for mean estimation, over-sampling strata with higher variance can give a more precise estimate of the mean. Does a similar result hold here?

Interestingly, it turns out that proportional sampling is the optimal allocation, provided that we are optimally choosing h_{st} .

Proposition 2: Under the assumptions of Proposition 1 and using h_{st} of (8), the sampling allocation which minimizes (9) is sampling proportional to stratum size,

$$n_i = n\theta_i.$$

Proof: The integrated mean squared error of $\hat{f}_w(y)$, using h_{st} is

$$IMSE(\hat{f}_w(y)) = \frac{5}{4}\gamma_2^{\frac{2}{5}} \left[\int (K(\psi))^2 d\psi \right]^{\frac{4}{5}} \left(\int_y \left[\sum_{i=1}^M \theta_i g_i'' \right]^2 dy \right)^{\frac{1}{5}} \left(\sum_{i=1}^M \frac{\theta_i^2}{n_i} \right)^{\frac{4}{5}} \quad (11)$$

If we minimize this quantity with respect to n_1, \dots, n_M constrained by $\sum n_i = n$,

$$\frac{\partial IMSE}{\partial n_j} = -\frac{4}{5}k^* \left(\sum_{i=1}^M \frac{\theta_i^2}{n_i} \right)^{-\frac{1}{5}} \frac{\theta_j^2}{n_j^2} + \lambda = 0$$

where λ is the Lagrange multiplier and k^* is defined from (11). To solve for λ , multiply both sides of the equation by n_j^2 and take the square root of both sides of the equation. This provides $\sqrt{\lambda} \sum n_j = \sqrt{\lambda}n = \left(\frac{4}{5}k^*\right)^{\frac{1}{2}} \left(\sum_{i=1}^M \frac{\theta_i^2}{n_i}\right)^{-\frac{1}{10}}$, giving an expression for λ . Replacing λ in the above equation with this expression then provides $\frac{\theta_j^2}{n_j^2} = \frac{1}{n^2}$ and $n_j = n\theta_j$. \diamond

This is perhaps a surprising result given the intuition from the mean estimation problem. However, in this case, we are not estimating any single parameter from each stratum, but instead the entire distribution. Even from a stratum whose distribution has a small variance we need a sample size sufficiently large to estimate the contribution of that stratum to the overall population density.

3 Numerical Illustration

In this section, we briefly compare the use of a weighted estimator combined with the proposed window width of (10) to weighted estimation using a naively chosen window width and to separate estimation of stratum-specific densities weighted by the appropriate mixing proportions with optimally chosen window widths for each stratum. For the naive case, we

set $h^* = 1.06\sigma n^{-\frac{1}{5}}$ based upon the full sample size, n , and full sample standard deviation, σ . This choice ignores all information about the sampling scheme. For the separate stratum, we use $h_i^* = 1.06\sigma_i n_i^{-\frac{1}{5}}$ in each stratum. We use a simple case with two normally distributed strata with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 .

We compare the integrated mean squared error under these three scenarios. This exercise is meant to be an illustration of the trade-offs involved in these options rather than a compelling practical example. In practice, the more interesting case is that of many strata with very small sample sizes in each stratum.

Table 1 presents, in the top panel, the values of h^* and h_{st} generated by setting $\sigma_1 = \sigma_2 = 1$ and varying the means of the two strata. The second to last column of Table 1 gives the ratio of the approximate IMSE (upto $O(\frac{1}{nh})$) of $\hat{f}_w(y)$ from equation (5) when a single window width is used. The last column of Table 1 compares the ‘ideal’ approach of using stratum-specific window widths and a weighted combination of stratum-specific densities as in (5). Standard normal kernels are used in all cases. The approximate integrated mean squared error is numerically calculated. The bottom panel of Table 1 presents the same information, holding means constant at zero and allowing the standard deviations of the two strata to become increasingly different.

Given the use of the weighted estimator for the density, using the proper window width gives large improvements in mean squared error over the standard reference window width. This is true even when the sampling is proportional (to stratum size). As the two strata become increasingly different (either in mean or in standard deviation) the gains in integrated mean squared error become quite large. When we compare the optimal (under the constraint of using only one parameter) window width h_{st} with the ‘ideal’ approach of separate strata-specific density estimation, we see that there is little loss in mean squared error from using one window width in the case where the different strata have identical standard deviations. As the distance between the strata increases (where $\mu_2 - \mu_1$ is between 1 and 3.5), using h_{st} actually provides superior integrated mean squared error, but the gains are small. The improvement in variance from a larger window width under h_{st} dominates any bias increase in these cases. Examining the point-wise data, the larger window width of h_{st} performs better at estimating the density between the two modes where there is mixing of the two distributions. In the case where stratum-specific standard deviations differ, however, the constraint of using only one window width, even when chosen optimally, comes at fairly high penalty in terms of integrated mean squared error. This is intuitively obvious—using a small window width in the stratum with small variance and a large window width in the stratum with large variance will out-perform a compromise window width halfway between the two.

Table 1: Comparison of window widths and integrated mean squared errors (IMSE) from weighted and unweighted estimation

<u>Identical Standard Deviations</u>						
$\mu_2 - \mu_1$	h^*	h_{st}	$IMSE(h^*)$	$IMSE(h_{st})$	$\frac{IMSE(h_{st})}{IMSE(h^*)}$	$\frac{IMSE(h_{st})}{IMSE(h_1^*, h_2^*)}$
Proportional Sampling: $n_2 = n_1$						
0.0	0.42199	0.42168	0.00836	0.00836	1.00000	1.14870
1.0	0.47180	0.47833	0.00737	0.00737	0.99963	1.01265
2.0	0.59679	0.58571	0.00602	0.00602	0.99928	0.82701
3.0	0.76076	0.49827	0.01140	0.00708	0.62081	0.97214
4.0	0.94361	0.47387	0.02639	0.00744	0.28199	1.02219
5.0	1.13625	0.47930	0.04896	0.00736	0.15028	1.01061
Non-proportional Sampling: $n_2 = 2 * n_1$						
0.0	0.38912	0.39811	0.00665	0.00664	0.99898	1.15927
1.0	0.43505	0.45159	0.00587	0.00586	0.99732	1.02197
2.0	0.55030	0.55296	0.00478	0.00478	0.99995	0.83462
3.0	0.70150	0.47041	0.00858	0.00562	0.65550	0.98108
4.0	0.87011	0.44738	0.01935	0.00591	0.30553	1.03160
5.0	1.04775	0.45250	0.03562	0.00584	0.16409	1.01991
<u>Identical Means</u>						
σ_2/σ_1	h^*	h_{st}	$IMSE(h^*)$	$IMSE(h_{st})$	$\frac{IMSE(h_{st})}{IMSE(h^*)}$	$\frac{IMSE(h_{st})}{IMSE(h_1^*, h_2^*)}$
Proportional Sampling: $n_2 = n_1$						
1.0	0.42199	0.42168	0.00836	0.00836	1.00000	1.14870
2.0	0.66723	0.53353	0.00746	0.00661	0.88582	1.21051
3.0	0.94361	0.55208	0.01389	0.00639	0.45980	1.31608
4.0	1.23031	0.55526	0.03291	0.00635	0.19298	1.39578
5.0	1.52152	0.55602	0.07298	0.00634	0.08690	1.45196
6.0	1.81506	0.55625	0.14528	0.00634	0.04363	1.49281
Non-proportional Sampling: $n_2 = 2 * n_1$						
1.0	0.38912	0.39811	0.00665	0.00664	0.99898	1.15927
2.0	0.61526	0.50370	0.00578	0.00525	0.90896	1.12066
3.0	0.87011	0.52121	0.01031	0.00507	0.49199	1.17002
4.0	1.13448	0.52421	0.02400	0.00504	0.21022	1.21202
5.0	1.40301	0.52493	0.05293	0.00504	0.09519	1.24154
6.0	1.67368	0.52515	0.10518	0.00504	0.04788	1.26270

$IMSE(h_1^*, h_2^*) = .00728$ for first six rows of table.

$IMSE(h_1^*, h_2^*) = .00573$ for the next six rows in the table.

$IMSE(h_1^*, h_2^*) = .00728$ when $\sigma_2/\sigma_1 = 1$ and decreases to .00425 when $\sigma_2/\sigma_1 = 6$ and sampling is proportional.

$IMSE(h_1^*, h_2^*) = .00573$ when $\sigma_2/\sigma_1 = 1$ and decreases to .00399 when $\sigma_2/\sigma_1 = 6$ and sampling is non-proportional.

4 Concluding Remarks

This paper is part of a growing literature which attempts to begin unifying survey design and nonparametric density estimation.⁶ As such we begin by analyzing the plug-in window width for normal data, the point of departure for most theoretical considerations of nonparametric density estimation. This is also of use as a starting point for other data-driven bandwidth selection techniques.

⁶ Breunig (2001) considers the case of clustered data. Chambers et al. (2003) discuss nonparametric regression with stratified samples.

The framework here is general and does not depend upon any minimal strata sample sizes. The technique presented in this paper, to choose one bandwidth for all the data which takes into account the strata differences, will work even when there are many strata with few observations per strata. Of course, knowledge of stratum-specific means and variances (or access to reasonable estimates thereof) is necessary. This is a problem which is frequently faced by survey statisticians in designing an optimal allocation. Using pretests, previous survey samples, or simple aggregation rules to combine similar strata are all ways around this problem, though all are imperfect. The technique does not relieve the researcher of the need to make intelligent choices according to the particular application.

Acknowledgements: The suggestions of two anonymous referees have greatly helped me in improving the paper. I am grateful to Aman Ullah for his suggestions and comments on this paper. I have also benefitted from conversations with Chris Skeels, Nilanjana Roy, and Andrew Weiss and the comments of participants in seminars at Georgia State University, the U.S. Bureau of Labor Statistics, and University of California, Riverside.

References

- [1] Breunig, R. (2001), Kernel density estimation for clustered data, *Econometric Reviews*, 20(3), 353-67.
- [2] Pagan, A. and A. Ullah (1999) *Nonparametric Econometrics* (Cambridge University Press, New York).
- [3] Chambers, R., A. Dorfman and M. Sverchkov (2003), Nonparametric Regression with Complex Survey Data, in: R. Chambers and C. Skinner, eds. *Analysis of Survey Data* (John Wiley & Sons, Ltd., New York).
- [4] Parzen, E. (1962) On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics* 33, 1065-1076.
- [5] Rosenblatt, M. (1956) Remarks on Some Nonparametric Estimates of Density Function, *Annals of Mathematical Statistics* 27, 832-837.
- [6] Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London).