NONPARAMETRIC DENSITY ESTIMATION USING COMPLEX SURVEY DATA

Trent D. Buskirk, Arizona State University
Department of Mathematics, Tempe, AZ 85287-1804

Key Words: Kernel Density Estimation, Superpopulation Models, Design Based Inference.

1. Introduction

In survey sampling the finite population from which the data are collected using a complex survey design can be treated as one realization from an infinite superpopulation which itself is generated according to a particular distribution (or density) function. In this case, it is generally assumed that the underlying distribution function is continuous. Because the finite population (and hence any data collected from it) will be at best discrete, and because interest is given to estimating a continuous density function, interpolation methods which can smooth the data are desired.

Traditional data smoothing techniques often involve the use of kernel functions (Wand and Jones 1995). These techniques commonly regard the data as an *iid* realization of some superpopulation distribution. Bickel and Freedman (1984) made the usual superpopulation assumption but modified it slightly by allowing for possibly differing distributions in each of several strata. Francisco and Fuller (1991) introduced techniques for quantile estimation which incorporated a survey design approach for data collected using stratified cluster sampling. Recently, Korn and Graubard (1998) suggested multiple techniques for scatter and box plots of data which account for the survey design. Korn and Graubard also proposed a method involving the use of kernel smoothing to estimate the conditional mean of bivariate data obtained using a complex survey design. However, in their work, there is no mention of incorporating the sample design to obtain a univariate density estimate of the variable of interest.

2. The Sample Weighted Kernel Density Estimator

For estimating a continuous, superpopulation density function, f, using data obtained in a complex survey design framework, we propose the Sample Weighted Kernel Density Estimator (SWKDE) given by:

$$\hat{f}_{S_{Nn}}(x) = \frac{1}{wh} \sum_{i \in S_{Nn}} \mathcal{K}\left(\frac{x - y_i}{h}\right) w_i$$

where S_{Nn} is any sample of size n taken from a universe, $\{y_1, y_2, \ldots, y_N\}$, of finite size, N, and $w_i = \pi_i^{-1}$ with $\pi_i = P_D\{i \in S_{Nn}\}$ and $w = \sum_{i \in S_{Nn}} w_i$. $\mathcal{K}$ is any kernel function having compact support which is symmetric about the origin and h, the bandwidth, is any positive quantity. For each $i = 1, 2, \ldots, N$, w_i is called the unit's *sampling weight*. The sampling design is incorporated into the SWKDE through these sampling weights.

3. What Does the SWKDE Estimate?

Under a superpopulation model which assumes that the finite universe is realized from a collection of N random variables, identically and independently distributed according to a density, f, we define the (finite) *population kernel density estimator*, PKDE by:

$$\hat{f}_N(x) = \frac{1}{Nh} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - y_i}{h}\right)$$

for any $x \in (-\infty, \infty)$. Under the auspices of this superpopulation model the PKDE is actually a standard, non-parametric kernel density estimator of f based on a random "sample" of N observations, Wand and Jones (1995). The SWKDE estimates the superpopulation density, f, through this PKDE.

4. Stratified Random Sampling

Suppose that the finite population of size N can be subdivided into L distinct subpopulations called strata with respective sizes $N_1, N_2, \ldots, N_L$. Assume that the random variables generating these subpopulations are identically distributed according to a density, f_k, for each $k = 1, \ldots, L$. Suppose also that these random variables are independent within and across strata. Within this framework it follows that for $y \in (-\infty, \infty)$, the function:

$$f(y) = \sum_{k=1}^{L} W_k f_k(y)$$

where $W_k = \frac{N_k}{N}$ for each $k = 1, 2, \ldots, L$, is indeed a population density function which can be estimated by the SWKDE under a stratified random sampling design (without replacement).

Let S_k, $k = 1, \ldots, L$, denote a simple random sample of size n_k taken from subpopulation k, where $n = \sum_{k=1}^{L} n_k$. Under this design the SWKDE becomes:

$$\hat{f}_{S_{Nn}}(x) = \frac{1}{Nh} \sum_{k=1}^{L} \left[\frac{N_k}{n_k} \sum_{i \in S_k} \mathcal{K}\left(\frac{x - y_{ki}}{h} \right) \right]$$

Using the stratum weights defined earlier, the SWKDE under a stratified random sampling design takes the form:

$$\hat{f}_{S_{Nn}}(x) = \sum_{k=1}^{L} W_k \hat{f}_{S_k}(x)$$

where $\hat{f}_{S_k}(x)$ is the SWKDE under a simple random sampling (without replacement) design for subpopulation k. This version of the SWKDE estimates the superpopulation density function, $f(y)$, by estimating the finite population quantity, $\hat{f}_N(x)$, given by:

$$\hat{f}_N(x) = \sum_{k=1}^{L} W_k \hat{f}_{N_k}(x)$$

where $\hat{f}_{N_k}(x)$ is the PKDE for subpopulation k.

5. Model & Design Based Properties of the SWKDE

▷ Under the stratified random sampling design:

- $E_D \left[\hat{f}_{S_{Nn}}(x) \right] = E_D \left[\sum_{k=1}^{L} W_k \hat{f}_{S_k}(x) \right] = \hat{f}_N(x)$
- $V_D \left[\hat{f}_{S_{Nn}}(x) \right] = \frac{1}{N} \sum_{k=1}^{L} W_k \left(\frac{w_k - 1}{N_k - 1} \right) \times \left[\sum_{i=1}^{N_k} \left(\frac{1}{h} \mathcal{K}\left(\frac{x - y_{ki}}{h} \right) - \hat{f}_{N_k}(x) \right)^2 \right]$

where $w_k = \frac{N_k}{n_k}$ is the common sampling weight for elements sampled from the kth stratum.

▷ Under the superpopulation model generating the stratified finite population:

- $E_M \left[\hat{f}_{S_{Nn}}(x) \right] = \frac{1}{h} \int \mathcal{K}_h(x - y) f(y) \, dy$
- $E_M \left[\hat{f}_{S_{Nn}}(x) - \hat{f}_N(x) \right] = 0$
- $V_M \left(\hat{f}_{S_{Nn}}(x) \right) = \frac{1}{N} \sum_{k=1}^{L} W_k w_k \left[\left(\mathcal{K}_h^2 * f_k \right)(x) - \left(\mathcal{K}_h * f_k \right)^2(x) \right]$
- $V_M \left(\hat{f}_{S_{Nn}}(x) \right) \leq \frac{1}{N} \left[\left(\sum_{k=1}^{L} w_k \right) \left(\mathcal{K}_h^2 * f \right)(x) - \frac{1}{L} \left(\mathcal{K}_h * f \right)^2(x) \right]$

where $\mathcal{K}_h(t) = \frac{1}{h} \mathcal{K}(\frac{t}{h})$ and * indicates the convolution operator.

▷ As defined by Isaki and Fuller (1982), the *anticipated variance* of the SWKDE under the stratified superpopulation model and the stratified random sampling scheme is:

- $AV\{\hat{f}_{S_{Nn}}(x) - \hat{f}_N(x)\} = E_M \left(V_D \left[\hat{f}_{S_{Nn}}(x) \right] \right) = \frac{1}{N} \sum_{k=1}^{L} W_k (w_k - 1) \left[\left(\mathcal{K}_h^2 * f_k \right)(x) - \left(\mathcal{K}_h * f_k \right)^2(x) \right]$

6. The National Crime Victimization Survey

To illustrate the use of the SWKDE, we consider data collected from the National Crime Victimization Survey (NCVS) during the 1992-1994 collection period. The NCVS is a stratified, multistage cluster sample yielding nationally representative samples. Information contained in the personal weighting and age variables for women were extracted from the 1994 data tapes of the NCVS (ICPSR, 1997), and used to estimate the age distribution of women experiencing:

▷ Sexual crimes involving rape, attempted rape, and coerced and unwanted sexual activity;

▷ Non-sexual assault crimes;

▷ Other personal crimes involving neither sex nor assault.

Altogether, there were 14,966 women surveyed by the NCVS during 1994, 237 of whom experienced sexual crimes and 1,057 of whom experienced non-sexual assault crimes.

Two density estimates, using the Epanechnikov kernel function, (with confidence bands computed using jackknife variance estimation, not including bias, Lohr 1999) are displayed for bandwidths of $1\frac{1}{2}$ and $2\frac{1}{2}$ years for female victims of assault and sexual

crimes in Figures 1 and 2, respectively. A density estimate for the ages of females encountering neither sexual nor assault crimes based on the Tri-Weight kernel function using a bandwidth of 3 years is displayed in Figure 3.

7. Conclusion

Nonparametric kernel density estimation has become a well established technique for density estimation under a theorized parametric model. The uses of nonparametric density estimation within the framework of an established survey design have been less widespread. A method which seeks to combine both the parametric, superpopulation model and the probability model induced by the sampling scheme has been introduced for the simple random sampling and the stratified random sampling designs. These results can be extended further for the case of a more general, stratified, multi-stage sample design, such as the National Crime Victimization Survey. The large-sample properties of the SWKDE will appear in a subsequent paper.

8. Selected References

Bickel, P.J. and Freedman, D.A. (1984),"Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics,* 12, 470-482.

Francisco, C.A. and Fuller, W.A. (1991), "Quantile estimation with a complex survey design," *The Annals of Statistics,* 19, 454-469.

Isaki, C.T. and Fuller, W.A. (1982), "Survey design under the regression superpopulation model," *Journal of the American Statistical Association,* 77, 89-96.

Korn, E.L. and Graubard, B.I. (1998). "Scatterplots with survey data." *The American Statistician.* **52** 58-69.

Lohr, S. (1999), *Sampling: Design and Analysis.* Pacific Grove, CA: Brooks/Cole.

U.S. Dept. of Justice, Bureau of Justice Statistics, NATIONAL CRIME VICTIMIZATION SURVEY, 1992-1994 [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. 3rd ICPSR ed., 1997.

Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing (Monographs on Statistics and Applied Probability, no. 60),* New York: Chapman & Hall.

Figure 1: SWKDE using the Epanechnikov kernel function with a 1.5 year bandwidth.

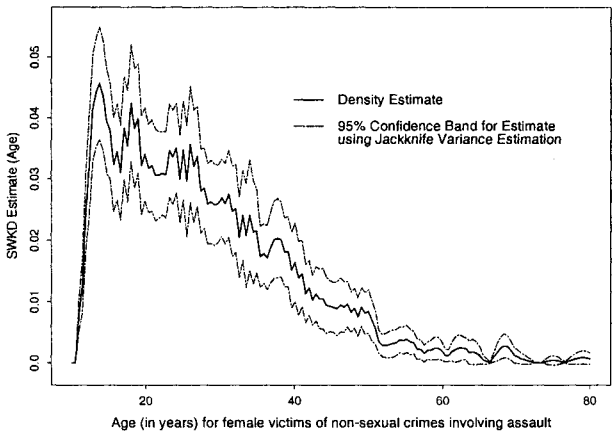

Figure 2: SWKDE using the Epanechnikov kernel function with a 2.5 year bandwidth.

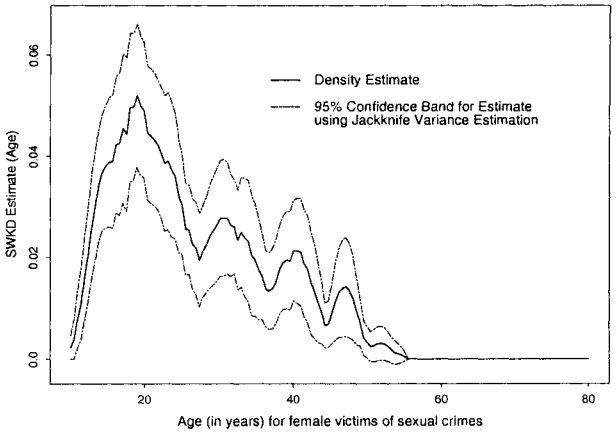

Figure 3: SWKDE using the Tri-weight kernel function with a 3 year bandwidth.