

IMPROVED DENSITY ESTIMATES FOR COMPLEX SURVEYS

Cristina M. Goia¹, David R. Bellhouse and Jamie E. Stafford

ABSTRACT

The aim of this paper is to improve kernel density estimates for histogram data through a bias adjustment. We refer to a density estimate applied to a histogram as a smoothed histogram. In general, kernel density estimates are widely used and preferred to histograms because of their nice, continuous shape. Continuity enhances the visual perception of the density estimate permitting easy multiple comparisons between different density estimates by superimposition. When only histogram data is available a smoothed histogram returns these advantages to the density estimate.

However, one concern about the accuracy of a smoothed histogram is its bias. This is a problem especially when we work with binned data because we have two sources of bias: binning and subsequent smoothing. Here bias adjustments are obtained through the use of a bootstrap technique. The technique also provides a method for percentile estimation and so confidence bands for the estimated density may also be computed. These prove to be useful in comparing different density estimates.

Ultimately we apply the methods developed to the Ontario health survey data.

KEY WORDS: Kernel density estimates; Histogram.

RÉSUMÉ

Le but de cet article est d'améliorer les estimations de densité du noyau pour des données d'histogramme au moyen d'un ajustement du biais. Une estimation de densité appliquée à un histogramme sera appelée un histogramme lissé. En général, des estimations de densité du noyau sont largement utilisées et sont préférées aux histogrammes en raison de leur belle forme continue. La continuité améliore la perception visuelle des estimations de densité en permettant des comparaisons multiples et faciles de différents estimateurs de densité en les superposant. Lorsque nous ne disposons que de données d'histogramme, un histogramme lissé procure ces avantages à l'estimation de la densité.

Cependant, une préoccupation concernant l'exactitude de l'histogramme lissé est son biais. C'est particulièrement un problème quand nous travaillons avec des données coffrées parce que nous avons deux sources de biais : l'entreposage et le lissage ultérieur. Ici, des ajustements de biais sont obtenus par l'utilisation de la technique du *bootstrap*. La technique fournit également une méthode pour l'estimation de percentile de sorte que les intervalles de confiance pour la densité estimée peuvent également être calculées. Ceux-ci s'avèrent utiles pour la comparaison de différentes estimations de densité.

Finalement, nous appliquons les méthodes développées aux données de l'Enquête sur la santé de l'Ontario.

MOTS CLÉS : Estimation de densité du noyau; histogramme.

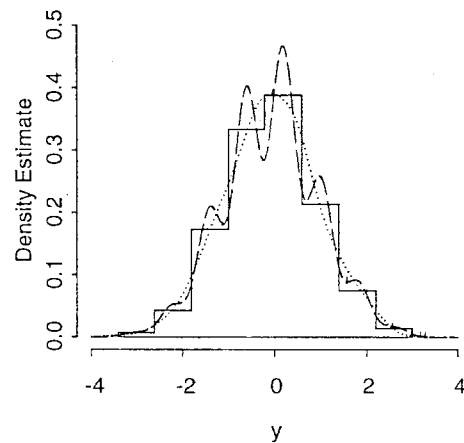
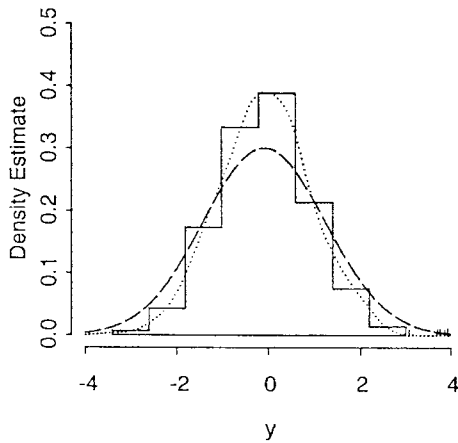
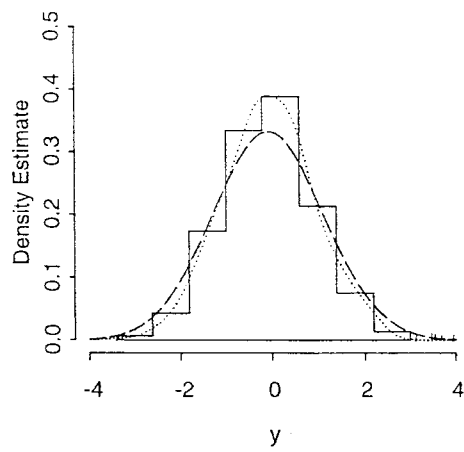
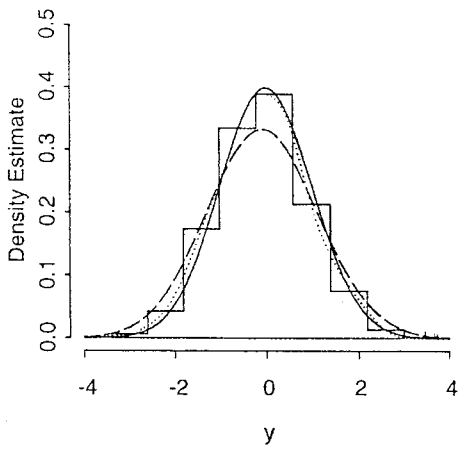
1. INTRODUCTION

1.1 Description of the problem

Very often, in the case of complex surveys the only available data is the histogram. Usually a histogram is easy to interpret but sometimes, for example when

making multiple comparisons by superimposition, we would prefer to work with continuous density estimates instead. Here we choose to compute kernel density estimates by smoothing the histogram. However, one concern about the accuracy of a smoothed histogram is its bias. This is illustrated very clear in the graphs below:

¹ Cristina M. Goia, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, N6A 5B7



The data used in obtaining the graphs is simulated data from the normal distribution. The continuous line represents the normal density function, the dotted line represents the kernel density estimate based on raw data and the dashed line the smoothed histogram. We can see that the smoothed histogram has even larger bias – obvious in the differences between the peaks of the density functions - than the kernel density estimate based on the raw data. This is natural since we have here two sources of bias: binning and subsequent smoothing. If we look to the second graph this becomes even clearer since the dotted line stays above the histogram while the dashed line goes below.

This problem cannot be fixed by controlling the window size when computing the kernel density estimates, this is what the next two graphs show. By increasing the window width we obtain what we see in the third graph: a very smooth density estimate but with a really large bias. By decreasing the window width we obtain what we see in the last graph: a density estimate that approaches somehow the original density but with a lot of variability. None of these is an acceptable solution.

Our solution is to obtain bias adjustments through the use of a bootstrap technique. Also, the technique permits the computation of percentile bands for the density estimators. Besides the variability evaluation, another benefit of the percentile bands is that we can use them to compare different densities. When trying to visualize how likely is that two density estimators come from the same distribution or not, we can superimpose the percentile bands for them. We can see then how much they overlap and get heuristic idea of how likely it is that they come or not from the same distribution.

1.2 Organization of the Paper

The methods developed here are just an example of the usual procedure used in analyzing complex survey data: a technique is developed for the IID case and then applied to the complex survey data. We started with IID data simulated from the normal distribution. Bias adjustments and percentile bands were constructed first for the kernel density estimate based on raw data, and then for the kernel density estimate

based on histogram data. Ultimately we applied our technique to the Ontario Health Survey data.

2. LITERATURE REVIEW

2.1 Histograms

For a given sample (Y_1, Y_2, \dots, Y_n) from an unknown distribution function F , the usual procedure for constructing a histogram is the following:

1. We divide an interval of the real line \mathfrak{R} , which contains (Y_1, Y_2, \dots, Y_n) , into non-overlapping intervals that are usually of equal length. We call these intervals bins and we denote the bin length by b .

2. We count the number of data points in each bin.

3. The histogram is then the step function given by $\hat{f}_H(y) = \frac{c(y)}{nb}$, where $c(y)$ is the number of data points in bin containing y .

The histogram function defined like this is a density function. Equivalently we can think about a histogram with k bins as the pair (m, p) , where $m = (m_1, m_2, \dots, m_k)$ are the midpoints of the bins and $p = (p_1, p_2, \dots, p_k)$ are the proportions of the sample contained in each bin: $p_i = \frac{c_i}{nb}$, for all $i = 1, 2, \dots, k$. We call the collection (m, p) the histogram data.

Choosing the right bin's size in constructing the histogram is a delicate issue. If b is too small the histogram shows a lot of variability wiggling up and down. If b is too large the histogram becomes smoother but the bias – evident in the gap between the true density function and the histogram – is too large.

2.2 Kernel Density Estimates

The kernel density estimate (KDE) is defined by (Silverman, 1986):

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{y - Y_i}{h}\right), \text{ for all } y.$$

The kernel $K(t)$ may be a symmetric function with $\int K(t)dt = 1$, $\int tK(t)dt = 0$ and $\int t^2 K(t)dt = k_2$.

The scale parameter h , also called the smoothing parameter or the window width, plays a similar role to the bin width b .

The Kernel that we use here is gaussian Kernel, i.e. the standard normal density function.

2.3 Smoothed histograms

When only the histogram data is available we compute the kernel density estimate by smoothing the histogram. The formula we use is:

$$\hat{f}_s(y) = \frac{1}{h} \sum_{i=1}^k p_i K\left(\frac{y - m_i}{h}\right), \text{ for all } y.$$

2.4 The sampling behavior of the density estimate

Two issues are of interest here: a percentile band and the bias of the density estimate.

By definition the α -percentile of $f(y)$ is $f_\alpha(y)$ satisfying

$$P(f(y) \leq f_\alpha(y)) = \alpha.$$

Then $(f_{0.025}, f_{0.975})$ is a 95% percentile band for f .

By definition $Bias \hat{f}(y) = E[\hat{f}(y)] - f(y)$. Approximating the expected value of the estimator

with the sample mean \bar{f} , this becomes

$Bias \hat{f}(y) = \bar{f}(y) - f(y)$. Then a density estimator adjusted for the bias is given by

$$\begin{aligned} \bar{f}(x) &= \hat{f}(x) - Bias \hat{f}(x) = \hat{f}(x) - \{\bar{f}(x) - f(x)\} \\ &= \hat{f}(x) - \bar{f}(x) + f(x). \end{aligned}$$

3. METHODS AND RESULTS

3.1 Methods developed

The kernel density estimate based on raw data

Bias adjustments and percentile bands are constructed here for the kernel density estimate \hat{f} , using only sample information obtained by simulation. We give below the algorithms for obtaining these. The method of simulation used is the bootstrap technique. We have two algorithms because when using bootstrap we have to mimic the “Real World” case as close as possible. For the “Real World” case we assume that we have the true density function f underlying the data:

"Real World" Algorithm	"Bootstrap World" Algorithm Using (Y_1, Y_2, \dots, Y_n) compute kernel density estimate $\hat{f}(y)$
1. Sample from f to get (Y_1, Y_2, \dots, Y_n)	1. Sample from \hat{f} to get $(Y_1^*, Y_2^*, \dots, Y_n^*)$
2. Using (Y_1, Y_2, \dots, Y_n) compute $\hat{f}(y)$	2. Using $(Y_1^*, Y_2^*, \dots, Y_n^*)$ compute $\hat{f}^*(y)$
3. Repeat 1&2 B times to get: $(\hat{f}_1(y), \hat{f}_2(y), \dots, \hat{f}_B(y))$	3. $(\hat{f}_1^*(y), \hat{f}_2^*(y), \dots, \hat{f}_B^*(y))$
4. Compute the percentile band $\hat{f}_\alpha(y), \hat{f}_{1-\alpha}(y)$ and the bias adjustment $Bias \hat{f}(y) = \bar{f}(y) - \hat{f}(y)$ $\bar{f}(x) = \hat{f}(x) - \bar{f}(x) + f(x)$	4. $\hat{f}_\alpha^*(y), \hat{f}_{1-\alpha}^*(y)$ $Bias \hat{f}(y) = \bar{f}^*(y) - \hat{f}(y)$ $\bar{f}(x) = 2\hat{f}(x) - \bar{f}(x)$

When evaluating the bias for the "Bootstrap World" we approximate the density function f with \hat{f} and then the two formulas for bias adjustment become:

$$Bias \hat{f}(y) = \bar{f}(y) - \hat{f}(y) \text{ and}$$

$$\bar{f}(x) = \hat{f}(x) - Bias \hat{f}(x) = \hat{f}(x) - \{\bar{f}(x) - \hat{f}(x)\}$$

$$= 2\hat{f}(x) - \bar{f}(x).$$

More useful than having a percentile band for the density estimator would be to have a percentile band for the adjusted density estimator. Then what we are

looking for is some $\bar{f}_{0.025}^*$, $\bar{f}_{0.975}^*$ for which:

$$P(\bar{f}(x) \leq \bar{f}_{0.025}^*(x)) = 0.025$$

$$P(\bar{f}(x) \leq \bar{f}_{0.975}^*(x)) = 0.975$$

But since $\bar{f}(x) = 2\hat{f}(x) - \bar{f}^*(x)$ it follows that:

$$P(2\hat{f}(x) - \bar{f}^*(x) \leq \bar{f}_{0.025}^*(x)) = 0.025$$

which is equivalent with:

$$P(\hat{f}(x) - (\bar{f}^*(x) - \hat{f}(x)) \leq \bar{f}_{0.025}^*(x)) = 0.025$$

and with:

$$P(\hat{f}(x) \leq \bar{f}_{0.025}^*(x) + \bar{f}^*(x) - \hat{f}(x)) = 0.025$$

But we know that:

$$P(\hat{f}(x) \leq \bar{f}_{0.025}^*(x)) = 0.025$$

so it follows that:

$$\bar{f}_{0.025}^*(x) + \bar{f}^*(x) - \hat{f}_s(x) = \bar{f}_{0.025}^*(x)$$

and the result is:

$$\bar{f}_{0.025}^*(x) = \hat{f}_{0.025}^*(x) - \bar{f}^*(x) + \hat{f}_s(x).$$

The percentile band is given by:

$$\bar{f}_{0.025} = \hat{f}_{0.025}^* - \bar{f}^* + \hat{f}$$

$$\bar{f}_{0.975} = \hat{f}_{0.975}^* - \bar{f}^* + \hat{f}$$

The window width used is $h_1 = 1.06n^{-1/5}$, the optimal value when working with the Gaussian Kernel according to Silverman, 1986.

Smoothed Histograms

Since this time our data is (m, p) and not (Y_1, Y_2, \dots, Y_n) , the algorithms need to be adjusted. First raw data was simulated from the normal distribution: $Y = (Y_1, Y_2, \dots, Y_n)$. Then the raw data was binned using the gaussian Kernel and the same optimal window width $h_1 = 1.06/n^{1/5}$. The binned data obtained this way: (m, p) is the data that our algorithm starts with.

The first three steps of the algorithms describe the procedure for constructing the data that we need: (m, p) . The window width when smoothing the binned data and obtaining the smoothed kernel density

estimate \hat{f}_s^* is $h_b = \frac{b}{1.25}$ where b is bin's length.

Following Jones, 1989 this is the ideal window size for \hat{f}_s , in order to make a fair comparison with the density estimator obtained from the raw data using h_1 . The rest of the algorithms are similar to the ones before.

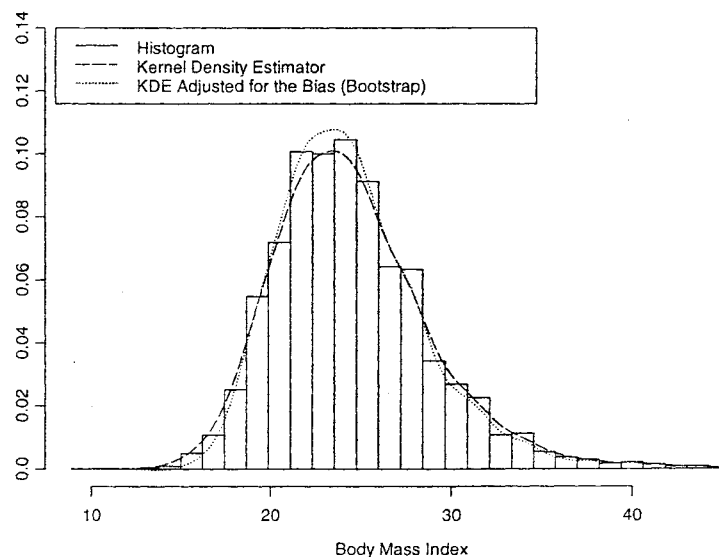
“Real World” Algorithm	“Bootstrap World” Algorithm
<p>1. Sample from f to get (Y_1, Y_2, \dots, Y_n)</p> <p>2. Using (Y_1, Y_2, \dots, Y_n) compute $\hat{f}(y)$</p> <p>3. Binning: using $m = (m_1, m_2, \dots, m_k)$ and \hat{f} get $p = (p_1, p_2, \dots, p_k)$</p> <p>4. Smoothing: using (m, p) get \hat{f}_s</p> <p>5. Repeat steps 1- 4 B times: $(\hat{f}_{s1}(y), \hat{f}_{s2}(y), \dots, \hat{f}_{sB}(y))$</p> <p>6. Compute the percentile band: $(\hat{f}_{s,\alpha}(y), \hat{f}_{s,1-\alpha}(y))$ and the bias adjustment: $\bar{f}_s = \hat{f}_s - \bar{f}_s + f$</p>	<p>Smooth the binned data (m, p) and get \hat{f}_s</p> <p>1. Sample from \hat{f}_s: $(Y_1^*, Y_2^*, \dots, Y_n^*)$</p> <p>2. \hat{f}^*</p> <p>3. (m, p^*)</p> <p>4. \hat{f}_s^*</p> <p>5. $(\hat{f}_{s1}^*(y), \hat{f}_{s2}^*(y), \dots, \hat{f}_{sB}^*(y))$</p> <p>6. $(\hat{f}_{s,\alpha}^*(y), \hat{f}_{s,1-\alpha}^*(y))$ $\bar{f}_s = 2\hat{f}_s - \bar{f}_s^*$</p>

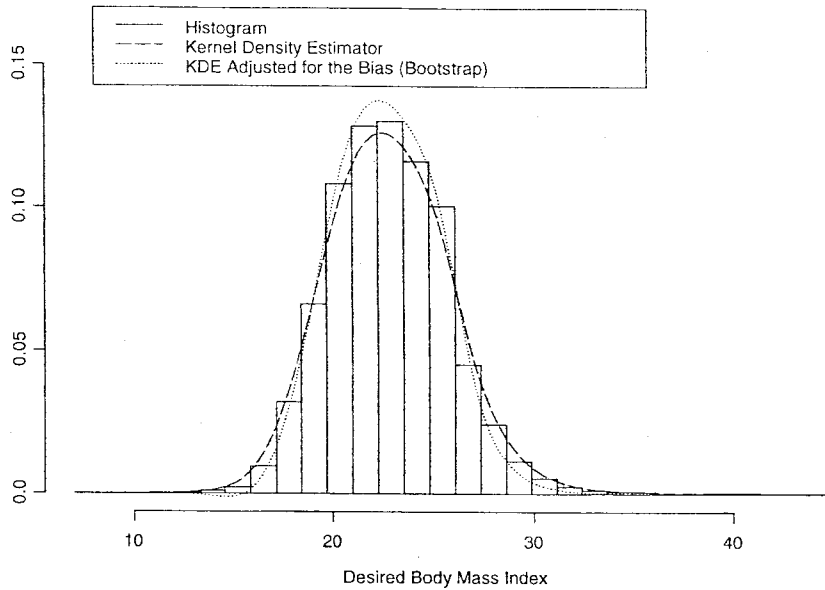
3.2 An Example

Our data comes from the Ontario Health Survey – 1990. Two measurements are of interest here: the Body Mass Index (BMI) and the Desired Body Mass Index (DBMI), i.e. the actual and desired ratio between a person’s weight (in kilograms) and height (in meters squared). For example, just to get an idea of

how this index functions, a BMI value less than 20 indicates that the person is at risk of having an eating disorder, a value greater than 27 a heart disease. The sample size is 41,939 for the BMI data and 44,457 for the DBMI data. What we use here is not the raw data but the published data, which is the histogram or binned data.

Bias adjustments

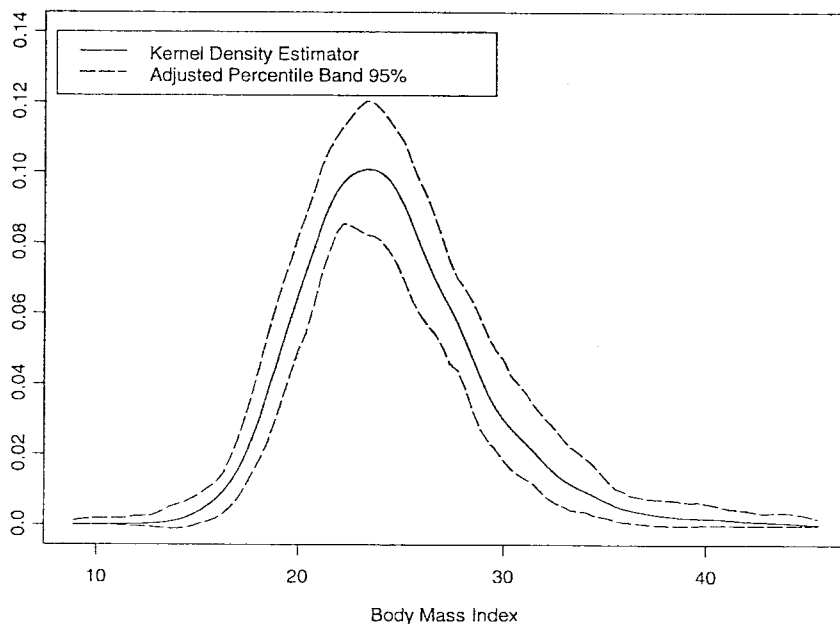


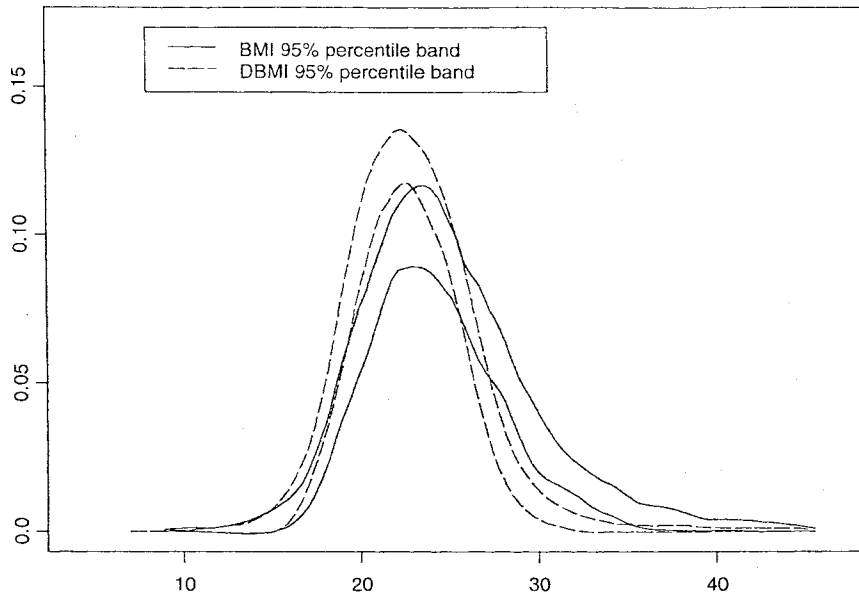


The kernel density estimate and the kernel density estimate adjusted for the bias were plotted together with the original histogram data. Note the kernel density estimate adjusted for the bias going above the histogram. Also, note the kernel density estimate adjusted for the bias going above the kernel density estimate at the peak and below on the tails. Especially this very last modification could be important because this actually means that less people than indicated by the unadjusted kernel density estimate have a critical value for BMI.

Percentile bands

All the percentile bands plotted here are actually adjusted percentile bands. The last graph is an example of the use of percentile bands as a heuristic test. We have two density estimators and for each of them we construct a percentile band. By looking at how much the two percentile bands overlap we can get an idea of how likely it is that the density estimators come from the same distribution or not. The comparison we are looking at here is between the BMI and DBMI density estimators. There is no real question here whether these two have the same underlying distribution or not, we only take the opportunity to exemplify our technique.





Acknowledgements

I wish to express sincere appreciation to Dr. James Stafford, who provided not only mathematical directions for this paper, but also enthusiasm and personal concern. Without his professional knowledge, time, patience and support this paper would have not been possible. Also, I am grateful to Dr. David Bellhouse for his help.

REFERENCES

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall: London

Jones, M.C. (1989) << Discretized and interpolated kernel density estimates >>, *Journal of the American Statistical Association* 84, 733 – 741.

Ontario Ministry of Health (1992). *Ontario Health Survey: User's Guide, Volumes I and II*. Queen's Printer for Ontario

Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, 2nd Ed. Duxbury: Belmont, California.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall: London.