ACTUARIAL RESEARCH CLEARING HOUSE
1993 VOL. 1

# ESTIMATING DENSITIES OF FUNCTIONS OF OBSERVATIONS

by

Edward W. Frees

University of Wisconsin - Madison

School of Business and Department of Statistics
1155 Observatory Drive
Madison, Wisconsin 53706

September, 1992

Keywords and Phrases: Spatial statistics, insurance statistics, U-statistics, nonparametric estimation.

# ESTIMATING DENSITIES OF FUNCTIONS OF OBSERVATIONS

## ABSTRACT

Density estimates, such as histograms and more sophisticated versions, are important in applied and theoretical statistics. In applied statistics, a density estimate provides the data analyst with a graphical overview of the shape of the distribution. This overview allows the data analyst to arrive immediately at a qualitative impression of the location, scale and various aspects of the extremes of the distribution. In theoretical statistics, the shape of the density allows the researcher to link the data to families of curves, perhaps indexed parametrically. By estimating a density nonparametrically, certain aspects of the data can be viewed without the limitation of *a priori* imposing limitations of a class of parametric curves.

In this paper, we introduce density estimation for functions of observations. To motivate the study, one type of function that is used is the interpoint distance between observations arising in spatial statistics from the fields of biometry and regional science. A second type of function consider are the sums of observations as might occur in claims models in insurance. The nonparametric density estimates are introduced and certain computational issues are discussed. A central limit theorem for the estimator is provided. What is interesting about this asymptotic result is that, under certain mild conditions, the density estimate enjoys a rate of convergence similar to parametric estimates. This rate of convergence is much faster than the usual rate of convergence in nonparametric density estimation.

# 1. INTRODUCTION AND BACKGROUND

## 1.1 Introduction

In this paper, problems from two different areas, spatial statistics and insurance, are discussed. Specifically, in the spatial statistics area, we consider estimating the density of the interpoint distance between pairs of objects. For example, in biometry, the objects might be distances between trees. In the regional sciences, the objects might be distances between population centroids. In the insurance area, we consider estimating the density of a sum of insurance claims. The distribution of sums of observations are called *convolutions* and their applications has received a great deal of attention in the insurance, as well as reliability, areas.

The common theme of these applications is that we are interesting in estimating the density of a *function* of observations. In the spatial statistics application, the function is some distance function between pairs of vector-valued observations. In the insurance application, the function is a fixed sum of univariate observations.

It is always possible to assume that the distribution of the basic observations comes from a parametric family of distributions. If this is the case, it is straight-forward to calculate the distribution of functions of observations, either analytically or using numerical integration. In this paper we consider instead *nonparametric* estimation. Either the data analyst may not have a convenient parametric family available or may wish to corroborate results from a parametric analysis with an analysis that does not use the assumption that the observations come from a specific parametric family. To this end, in the following subsection a brief background on nonparametric estimation is provided.

## 1.2 Background on Nonparametric Estimation

In 1948, Hoeffding introduced a class of statistics, called *U-statistics*, which have turned out to be one of the cornerstones of classical nonparametric statistics. A related work that appeared about the same time is due to Halmos (1946). A U-statistic is defined to be a sample average of a function of one or more observations. Hoeffding called this function a *kernel*. The *U* prefix stands for the fact that each statistic is an *unbiased* estimator of a parameter of interest. Over the intervening 40-plus years, this class has received a substantial amount of attention for at least two important reasons. First, it has been demonstrated that many important statistics can be demonstrated to be members of this class; for example, see Randles and Wolfe (1979). Second, an even greater number of statistics are closely related to this class of statistics such as the Von Mises (1947) *V-statistics*; for example, see Serfling (1980). By studying this broad of class of statistics, properties for many important statistics can be established as special cases.

Lately, a number of researchers have studied not just the average but the entire distribution function of an evaluation of the kernel. Specifically, let $X_1, X_2, \ldots, X_n$ be a random sample and consider a kernel g that depends on $m \geq 1$ variables. For convenience, the function g is assumed to be real-valued and symmetric in its arguments. The associated kernel distribution function (KDF) is defined by

$$H(t) = P(g(X_1, \ldots, X_m) \leq t), \tag{1.1}$$

where $t \in R$. Using the random sample $X_1, \ldots, X_n$, the KDF can be estimated by the associated empirical kernel distribution function (EKDF),

$$H_n(t) = \binom{n}{m}^{-1} \Sigma_c I(g(X_{i_1}, \ldots, X_{i_m}) \leq t). \tag{1.2}$$

Here I(.) is the indicator function of a set, $\{i_1, \ldots, i_m\}$ is an ordered subset of $\{1, \ldots, n\}$ and $\Sigma_c$ means sum over all $\binom{n}{m}$ such subsets. For a fixed t, $H_n(t)$ in (1.2) is a U-statistic with a bounded kernel and has several well-known properties, cf., Sen (1980, Chapter 3) and Serfling (1981, Chapter 5). For a fixed n, $H_n$ is a stochastic process called an empirical kernel process (also called a U-statistic empirical process by Shorack and Wellner, 1986, Chapter 23.4). This process was first discussed by Silverman (1976) in the context of exchangeable random variables. Silverman established weak convergence of $H_n$, when suitably normalized, to a Gaussian process. Independently, Sen (1983) discussed $H_n$ from the viewpoint of U-statistic theory, established weak convergence to a Gaussian process, and established strong uniform consistency. Convergence using yet stronger metrics was established by Silverman (1983) and strong approximations for the empirical kernel process are also available, see Csörgõ, Horvath and Serfling (1983) and extensions by Dehling, Denker and Philipp (1987). This process has been used in the study of interpoint distances, a problem in spatial statistics. For further discussion of this aspect, see Silverman and Brown (1978), Weber (1983), Jammalamadaka and Janson (1986) and Section 2 below.

Much of the statistical motivation for examining the properties of an EKDF is similar to the motivation for studying the usual empirical distribution function. Many interesting statistics can be expressed as functionals of an EKDF. In this case, properties of the EKDF can be used to analyze the statistic of interest. This method of analysis can be attributed to Serfling's (1984) introduction of *generalized L-statistics*. A generalized L-statistic is a linear combination of evaluations of a kernel g, there being $\binom{n}{m}$ evaluations of the form $g(X_{i_1}, \ldots, X_{i_m})$. Additional statistical motivation for exploring properties of EKDF's was provided by Sen (1983) who gave a special case of a generalized M-statistic.

(However, this terminology was used by Serfling, 1984.) Extensions to spread functionals were given by Janssen, Serfling and Veraverbeke (1984) and to simple linear regression by Frees (1991). On a separate note, Akritas (1986) developed a "V-empirical process," like the empirical kernel process but using V-statistics in lieu of U-statistics, primarily for statistical inference in random censoring models.

## 1.3 Densities of Functions of Observations

A function closely related to the KDF is the kernel density function, defined to be $h(t) = H'(t)$, when it exists. In this paper we investigate the estimation of $h(t)$. The motivation is the same as Hoeffding's: to establish a broad class of statistics that handle several important examples as special cases and that are closely related to other important examples. The nonparametric estimator introduced here is the kernel estimate

$$h_n(t) = \frac{1}{\left(\binom{n}{m} b_n\right)} \sum\nolimits_c w\left(\frac{t-g(X_{i_1}, \ldots, X_{i_m})}{b_n}\right) \tag{1.3}$$

Here, $\{b_n\}$ is the so-called "bandwidth", or window width, and w(.) is a kernel function of the type popularized by Rosenblatt (1971). It is regrettable that the adjective "kernel" is used in two such distinct fashions. The first usage is in connection with U-statistic theory as a function of observations and the second usage is in connection with density estimation as a weight indicating the closeness of an observation to a fixed point. However, the usage of the word seems to be permanently imbedded in each literature and we will attempt to clarify each usage as it appears in the paper.

The form of the density function in (1.3) is complex in appearance but is really quite straightforward to compute. For many applications, one only needs to compute the $\binom{n}{m}$ evaluations of the kernel and then apply any standard kernel density routine to these evaluations, as if they were data. For applications in which it is impractical to compute all $\binom{n}{m}$ evaluations, subsamples or random resamples may be used; see Section 2.4 below. One of the most important properties of the density estimator is it's fast rate of convergence. In the case of $m=1$, $h_n$ is the usual kernel density function has a rate of convergence which depends on the smoothness of the density and is strictly slower than $n^{1/2}$. See Silverman (1986) for an introduction or Izenman (1991) for a recent review of nonparametric density estimation that corresponds to the case $m=1$. Asymptotic properties of $h_n$ in the case $m > 1$ are radically different. It turns out, because of the additional averaging in (1.3), that the sampling bias in using $h_n(t)$ to estimate h(t) is of a lower order of magnitude than the usual case (corresponding to $m = 1$). This reduction of bias allows us to establish rates of convergence similar to unbiased estimators. For example, it turns out, under mild conditions, that the simple histogram estimator approximates the kernel density well up to terms of order $n^{1/2}$. This is a much closer order of approximation than the usual kernel density estimator in the case $m = 1$. By "simple histogram estimator" we mean the choice $w(x) = I(-1 < x \leq 1)/2$ which yields $h_n(t) = (H_n(t+b_n) - H_n(t-b_n))/(2\ b_n)$ in (1.3).

In Section 2, applications of the kernel estimate are illustrated using an example from spatial statistics and from insurance. In Section 3, asymptotic properties of the kernel estimate in (1.3) are studied. Readers interested in the general statistical methodology may wish to skip directly to Section 3. Here, it is noted that this problem also arises in estimating a component of the asymptotic variance in rank regression. We close in Section 4 with some concluding remarks. The proofs of all properties can be found in the Appendix.

## 2. MOTIVATING EXAMPLES AND RESAMPLE ESTIMATORS

### 2.1 Redwood Locations

Diggle (1983) provides the location of 62 redwood seedlings in a unit square. The data are originally from Strauss (1975). A graph of the data, which appears in Figure 1, shows that the data do not appear to be randomly dispersed throughout the square. To measure the degree of spatial randomness, Diggle proposes looking at the distribution function of distances between points. In the notation of this paper, let $g(x_1, x_2) = |\ x_1 - x_2\ |$ be the Euclidean distance between the two-dimensional points $x_1$ and $x_2$. There are 1,891 ( = ( $\binom{62}{2}$ ) ) interpoint distances. Thus, from equation (1.3), the nonparametric density estimator is

$$h_{1n}(t) = \frac{1}{\left( \binom{n}{2} b_n \right)} \sum_{i<j} w\left( \frac{t - |X_i - X_j|}{b_n} \right) \tag{2.1}$$

where $\Sigma_{i<j}$ means sum over all observations such that $i<j$. In Figure 2, a histogram of these distances are given with the kernel density estimate superimposed. Here, as for all the examples in this paper, the kernel used is the quadratic kernel due to Epanechnikov (see, for example, Silverman, 1986, p. 42). Is this the pattern that one would expect of a spatially random process? As a reference distribution, Bartlett (1964) has shown, for observations distributed uniformly on the unit square, that

$$h_0(t) = \begin{cases} (2t)\,(\pi - 4\,t + t^2) & 0 \le t \le 1 \\ (2t)\,(\,-2 - t^2 + 4(t^2-1)^{1/2} + 2\,\sin^{-1}(2t^{-2} - 1)\,) & 1 < t \le \sqrt{2}. \end{cases}$$

A graph of this reference distribution appears in Figure 3. Comparing Figures 2 and 3, it is easy to see that the empirical interpoint density is larger than the reference density for small values of interpoint distances. This point was noted by Strauss and Diggle who attribute it to the underlying clustering method associated with the generation of seedlings. Now, as is the typical case in density estimation, the above graphs make it straightforward to detect differences. In contrast, Diggle advocates graphing the distribution functions. Here, the empirical interpoint distribution function is $H_{1n}(t) = \binom{n}{2}^{-1} \Sigma_{i<j}\, I(\,|X_i - X_j| \le t)$. While providing the same information, it is more difficult to detect differences when considering distribution functions that tend to accumulate effects.

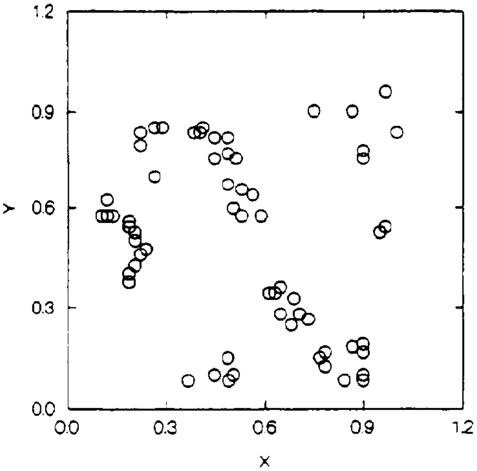

Figure 1. Scatterplot of the Locations of 62 Redwood Seedlings. Data has been rescaled to a unit square.

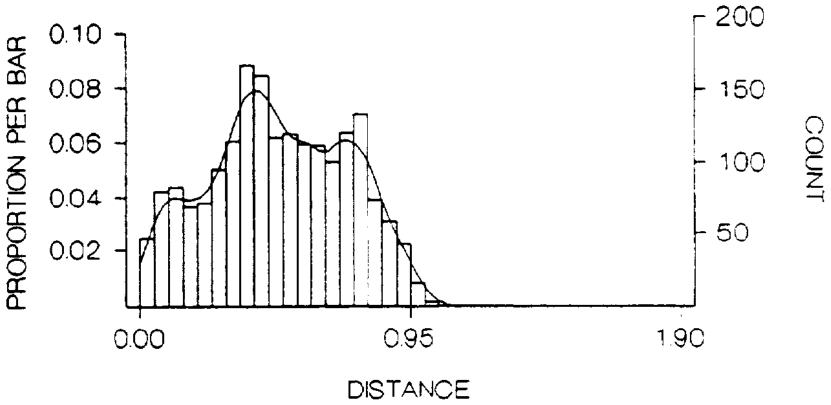

**Figure 2. Histogram of 1,891 Interpoint Distances. The kernel density estimator is superimposed on the raw histogram.**

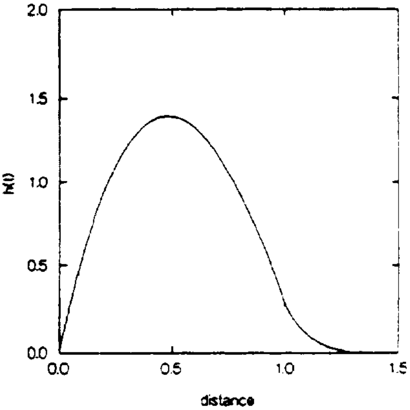

**Figure 3. Bartlett's reference distribution. Theoretical interpoint density for observations distributed uniformly on the unit square.**

## 2.2 Distances Between Population Centroids

In the regional sciences, researchers are concerned with quantifying the separation between population regions. One way of measuring this separation is by measuring the distance between population centroids. For example, in Figure 4 is a histogram of the distances between state population centroids. Since there are 51 "states" in the U.S. (including the District of Columbia), there are $\binom{51}{2}$ = 2,550 distances. Although the data is observational, researchers are interested in the deviation from the uniform distribution. Further, the distance variable is an important factor in quantifying population movements. For example, a classical "gravity" model of migration from the o$^{th}$ to the d$^{th}$ state might be expressed as

$$M_{od} = c\,\frac{P_o P_d}{D_{od}^a}\left(\frac{I_d}{I_o}\right)^b\left(\frac{E_d}{E_o}\right)^f e_{od}$$

Here, $P$ is state population, $I$ is state income, $E$ is state (un)employment, $D$ is distance between population centroids, $a$, $b$, $c$ and $f$ are parameters to be estimated, and $e_{od}$ is the multiplicative error term. Gravity models are discussed in the survey by Greenwood (1975) and the monograph by Haynes and Fotheringham (1984). A more recent discussion of the relationship between migration rates and migration distance is in Long, Tucker, and Urton (1988). This model can be easily converted to the linear model via the logarithmic transform. Thus, the gravity model theory suggests using logarithms of the distance between population centroids in a regression model. For data analysts who use the Box-Tidwell (1962) approach of symmetrizing regressors, the skewed distribution in Figure 4 also suggests a logarithmic transformation.

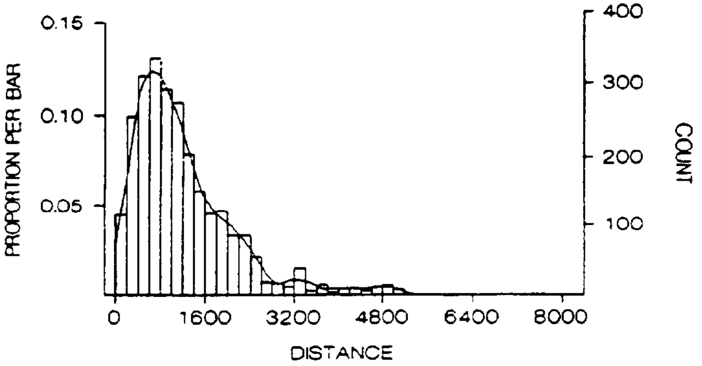

Figure 4. Histogram of 2,550 distances between population centroids. The kernel density estimator is superimposed on the raw histogram.

## 2.3 Convolutions of Insurance Claims

The application which originally motivated this study was the estimation of convolution distribution functions and densities. These are of interest in insurance (Panjer, 1980 and Hogg and Klugman, 1984) and reliability (Bagchi et al, 1986) applications. To illustrate, we assume that the insurer has available $X_1, \ldots, X_n$, a random sample of insurance claims from a particular line of business. As an example, we consider here the 1989 Total Charges for 33 patients at a Wisconsin Hospital. Each patient was female, aged 30-49 and admitted to the hospital for circulatory disorders. The data appears in the Appendix in Table 1 and is summarized in the histogram in Figure 5. From an insurer's standpoint (or a risk manager associated with the hospital), what is of interest is the distribution of the sum of claims, $X_1+X_2+ \ldots + X_m$. Interpret 'm' to be the expected number of claims in a specified financial period, for example, a month or quarter. Of course, this analysis assumes that there are no time trends in the particular line of business being analyzed. To define an estimator of the density of the convolution, use $g(x_1,\ldots,x_m) = x_1+\ldots+x_m$. From (1.1), the m-fold convolution of F is $F^{(*m)}(t) = P(X_1+\ldots+X_m \leq t)$. Here, use F to denote the distribution function and $f=F'$ to be the corresponding density, when it exists. An estimate of the convolution density, $f^{(*m)}(t) = (\partial/\partial t)\, F^{(*m)}(t)$, from (1.3) is

$$f_n^{(*m)}(t) = \left( \binom{n}{m} b_n \right)^{-1} \Sigma_c\, w( (t - (X_{i_1} + \ldots + X_{i_m})) / b_n ), \tag{2.2}$$

with this choice of g.

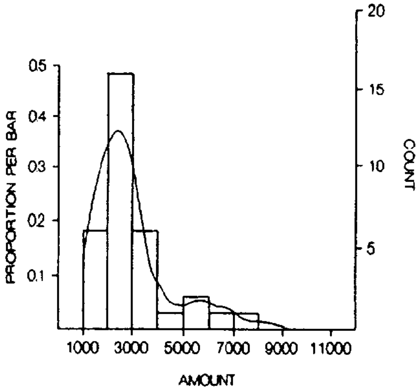

Figure 5. 1989 Charges for 33 patients at a Wisconsin Hospital. Each patient was female, aged 30-49 and admitted for circulatory disorders.

To illustrate the effect of an additional expected claim on the distribution of claims, in Figures 6 through 9 are histograms of the sum of claims for m=2,3,4, and 5, respectively. For reference, for the original 33 claims, the average claim is $\bar{x}$ = 2955 with standard deviation $s_x$ = 1481. As m increases, we see that the bimodal nature of the empirical density flattens out and that the curve becomes more symmetric. By allowing the risk manager to choose his or her best approximation of the expected number of claims, the histograms allow the data analyst to provide a graphical summary of the distribution of the sum of claims.

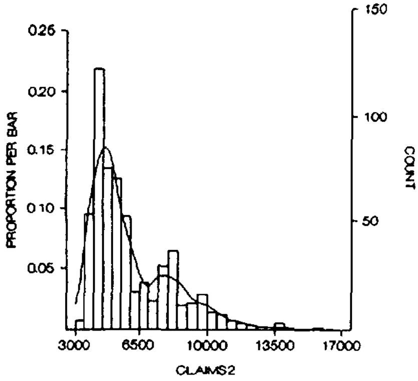

Figure 6. Distribution of all 561 pairs of 1989 Total Hospital Charges.

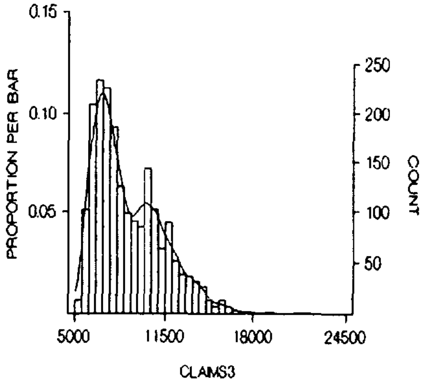

Figure 7. Distribution of the sum of 3 claims. The resample density estimate is calculated using R=2000 resamples.

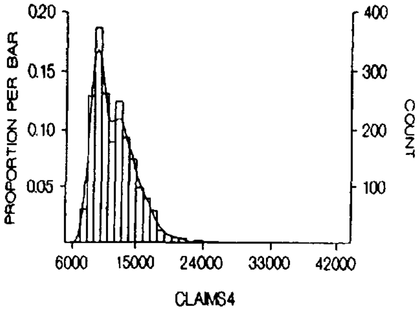

Figure 8. Distribution of the sum of 4 claims. The resample density estimate is calculated using R=2000 resamples.

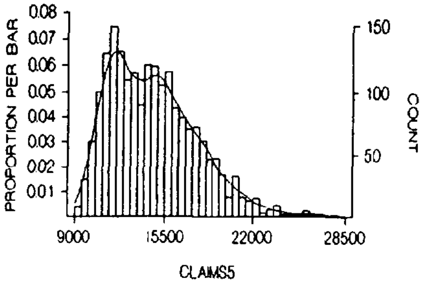

Figure 9. Distribution of the sum of 5 claims. The resample density estimate is calculated using R=2000 resamples.

## 2.4 Resample Density Estimators

In principle, to compute the density estimate when $n=33$ and $m=5$, there are $\binom{33}{5} = 237{,}336$ evaluations of a kernel required. This many evaluations, combined with the usual kernel density estimation routine, would require a prohibitive amount of calculations for all but the most powerful of computing environments currently available. However, by choosing subsets, a much smaller number of evaluations can be used to compute an effective evaluation of an average. This idea is originally due to Blom (1976) in the U-statistics literature; see Frees (1989) for a recent review. As in Frees (1989), in this paper we use the technique of so-called "random resamples."

To define the resampling estimator, let $R=R(n)$ be a positive integer depending on n such that $R\to\infty$ as $n\to\infty$. Based on the observed sample, we draw R independent (conditional on the sample, $\{X_1, \ldots, X_n\}$) realizations, as follows. For $r=1, \ldots, R$, make m draws without replacement from $\{X_1, \ldots, X_n\}$ to get $\{X_1^{*r}, \ldots, X_m^{*r}\}$. Define the resample estimator as

$$h_R(t) = \frac{1}{R\, b_n} \sum_{r=1}^{R} w\left(\frac{t-g(X_1^{*r}, \ldots, X_m^{*r})}{b_n}\right). \tag{2.3}$$

It is important to emphasize the fact that the resample estimator is only one of several ways of evaluating the average. It is computationally simple and has intuitive appeal due to the recent increased popularity of using simulation techniques for resampling statistics. The Figures 7 through 9 were created using the resampling estimator in (2.3) with $R= 2{,}000$ resamples.

## 3. PROPERTIES OF THE KERNEL ESTIMATE

We begin this section with a lemma that shows how to bound the deviation of the estimate from the true kernel density in terms of the bias and the bandwidth $\{b_n\}$. This establishes some basic results on rates of convergence, including the pointwise convergence of the estimator. Further, it demonstrates the usual trade-off in density estimation involving the choice of the window width $b_n$. For the case $m > 1$, we then show that $n^{1/2}(h_n(t) - h(t))$ is asymptotically normal, the main result of this section.

To state the lemma, we use the maximal deviation of $H_n$ from H,

$$M_n = \sup_t \quad | H_n(t) - H(t) | \tag{3.1}$$

and the bias term,

$$B_n(t) = E\, h_n(t) - h(t). \tag{3.2}$$

We use following mild assumptions on the kernel w(.).

*Assumption* K1. Assume $\lim_{|y| \to \infty} w(y) = 0$, $\lim\sup_{|y| \to \infty} | w(y) |$ is bounded, $\int w(y)\, dy = 1$, and that $\int \ | dw(y) |$ is finite.

In the case w(.) is differentiable, then $\int \ | dw(y) | = \int \ | w'(y) | \ dy$. With these minimal assumptions, we now have the following

Lemma. Assume K1. Then,

$$h_n(t) - h(t) = O(M_n/b_n) + B_n(t) \qquad \textit{almost surely.} \tag{3.3}$$

Remarks: An interesting aspect of Lemma 1 is that the maximal deviation of the density from its estimate can be bounded in terms of the maximal deviation of the kernel distribution function from the EKDF. From Silverman (1976) or Sen (1983), we have that $M_n = O_p(n^{-1/2})$. From Dehling, Denker and Philipp (1987, Corollary 2), this latter bound can be written as

$$M_n = O(n^{-1/2} (\log \log n)^{1/2}). \tag{3.4}$$

In (3.4), the big oh notation means almost surely after possibly enlarging the underlying probability space. Thus, as an immediate corollary of Lemma 1, if $b_n \to 0$ such that $B_n(t) \to 0$, then $h_n(t)$ is a consistent estimator of h(t). One may use weak consistency or strong consistency in the sense of (3.4). Analogous to density estimation, rates of convergence of the bias term $B_n(t)$ may be framed in terms of the smoothness of h(.). For example, from (3.2), a change of variables, and a Taylor-series expansion, we have

$$\begin{aligned} B_n(t) &= b_n^{-1} \int w((t-u)/b_n)\, h(u)\, du \; - \; h(t) \\ &= \int w(y)\, h(t-y\, b_n)\, dy \; - \; h(t) \\ &= \int w(y)\, \Sigma_{j=1}^{r} \; (-y\, b_n)^j / j! \;\; h^{(j)}(t) + y^r\, O(b_n^r))\, dy = O(b_n^r). \end{aligned} \tag{3.5}$$

Here, we have assumed $\int y^j w(y)\, dy = 0$, $j=1,\ldots,r-1$, $\int \; | y^r w(y) | \; dy < \infty$ and that the $r^{th}$ derivative of h is bounded in a neighborhood of t. The choice of r=2 is important. This corresponds to the situation where we wish to use a nonnegative kernel so that w(.) may be interpreted as a probability density function. From the Lemma and (3.5), using $b_n = O(n^{-1/(2r+2)})$ yields $h_n(t) - h(t) = O_p(n^{-r/(2r+2)})$, thus establishing a rate of convergence. Further, if the $r^{th}$ derivative of h is uniformly bounded, then the bound in (3.5) is uniform in t. That is, with the Lemma, we have

$$\sup_t \;\; | h_n(t) - h(t) | \;\; = \;\; O_p(n^{-r/(2r+2)}). \tag{3.6}$$

The rate of convergence in (3.6) is the same as the usual density estimate and, for the case m=1, is in some sense optimal. However, in the case m>1, this rate can be improved as follows.

Define the distribution function $H_1(x,t) = P(g(x, X_2, \ldots, X_m) \le t)$ and suppose that the corresponding density, $h_1(x,t) = (\partial/\partial t)\, H_1(x,t)$, exists.

**Theorem.** Assume K1, m>1, n $b_n \to \infty$,

$$n^{1/2}\, B_n(t) \to B, \tag{3.7}$$

that $h_1$ exists, and, for $\delta>0$, E $| h_1(X,t) |^{2+\delta} < \infty$. Then,

$$n^{1/2}\, (h_n(t) - h(t)) \to_D \; N(B, \sigma^2), \tag{3.8}$$

where $\sigma^2 = m^2\, (E\, h_1(X_1,t)^2 - (h(t))^2)$.

*Remarks on the Assumptions*: The condition in (3.7) is mild and is satisfied, for example, with (3.5) and using $b_n = o(n^{-1/2r})$ so that $B=0$. Although easily satisfied in a number of cases, the requirement that $h_1$ exist is important. As an example where it is not satisfied, consider the kernel $g(x_1, x_2) = \max(x_1, x_2)$. With this choice, we have $H_1(x,t) = I(x \leq t)F(t)$ which is not differentiable at $t=x$. Suppose, for the moment, that the assumption is not necessary and that (3.8) holds. Now, with this choice of g, it is easy to check that $h(t) = 2f(t)F(t)$. Thus, if (3.8) were to hold, then $h_n(t)/ (2\, F_n(t)\,)$ would be a nonparametric root-n estimator of the probability density function in the sense that $h_n(t)/(2\, F_n(t)\,) - f(t) = O_p(n^{-1/2})$. It is well-known that this is faster than the optimal nonparametric rate, see Stone (1980). This demonstrates the necessity of the assumption.

*Illustration of the Theorem - Convolution Density Estimator:* As an illustration of the theorem, consider the convolution density estimator introduced in Subsection 2.3. With $g(x_1,\ldots,x_m) = x_1+\ldots+x_m$, we have $H_1(x,t) = P(x+X_2+\ldots+X_m \leq t) = F^{(*m-1)}(t-x)$. Thus, assuming differentiability, we have $h_1(x,t) = f^{(*m-1)}(t-x)$. Assume that there are sufficient conditions on the density so that (3.7) holds. A sufficient condition so that $E\ |\ h_1(X,t)\ |^{\,2+\delta} < \infty$ holds is the requirement that $f^{(*m-1)}(t)$ is bounded in 't'. Thus, with K1, we have

$$n^{1/2}\,(f_n^{(*m)}(t) - f^{(*m)}(t)) \to_D N(B, \sigma_1^2)$$

where $f_n^{(*m)}(t)$ is given in (2.2) and $\sigma_1^2 = m^2\,(\,E\,(f^{(*m-1)}(t-X))^2 - (f^{(*m)}(t))^2)$.

*Remarks on Confidence Intervals:* An important application of a limit result, such as in the theorem above, is that a confidence interval for the parameter can be easily generated. In practice, the only thing that is required is knowledge of, or a consistent estimate of, the asymptotic variance parameter $\sigma^2$ in (3.8). To develop a consistent estimator of $\sigma^2$, we advocate the use of the *jackknife* technique, pioneered in the U-statistic area by Sen (1960). To this end, define

$$h_{n,i}(t) = \frac{1}{\left(\binom{n-1}{m-1} b_n\right)} \sum\nolimits_{c(i)} w\left(\frac{t-g(X_i,X_{i_2}, \ldots,X_{i_m})}{b_n}\right)$$

where $\{i_2,\ldots,i_m\}$ is an ordered subset of $\{1,\ldots, i-1, i+1,\ldots, n\}$ and $\Sigma_{c(i)}$ means sum over all such subsets. Define the jackknife estimator of $\sigma^2$ as

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (h_{n,i}(t) - h_n(t))^2 .$$

It is straightforward to check the consistency of this estimator using standard techniques such as in Sen (1981, pages 80-81) and the triangular array arguments of this paper.

*Properties of the Resampling Estimator:* To illustrate the large sample properties of the resampling estimator, we present the following

**Corollary.** Assume the conditions of the Theorem and $R / (n / b_n^2) \to \infty$ hold. Then,

$$n^{1/2} (h_R(t) - h(t)) \to_D N(B, \sigma^2).$$

Thus, the resampling estimator enjoys the same first order asymptotic properties as the density estimate. The only requirement is that the resample size, R, grow fast enough relative to the sample size, n, and the bandwidth, $b_n$. Rates for other modes of convergence can be checked using, for example, methods in Frees (1989).

To consider other aspects of the properties of the kernel density estimator, we now consider another special case.

*Example - Scale Parameter in Rank Regression:* As another example, assume that the observations are univariate with distribution function F and use $m=2$ and $g(x_1,x_2) = | x_2 - x_1 |$. From (1.1), we have $H(t) = E\,(F(X+t) - F(X-t))$ and thus $h(t) = \int (f(u+t) + f(u-t))\, f(u)\, du$. Here, we use $f = F'$ and assume that the interchange of differentiation and integration is valid. In particular, kernel estimates of $h(0)/2 = \int f^2(u)\, du$ were introduced by Schuster (1969). Properties of these estimates were investigated by Schweder (1975), who established asymptotic normality and Cheng and Serfling (1981), who considered certain score variations. See Hettmansperger (1984, p. 249) for an introduction.

*Remarks on the Selection of the Kernel and Bandwidth:* Much of the traditional literature in density estimation developed to provide guidelines for selection of the kernel and bandwidth have been driven by the goal of balancing the trade-off between minimizing bias and variance. The trade-off is that smaller values of the bandwidth decrease the bias but increase the variance (see, for example, Silverman, 1986, Section 3.3.1). In contrast, in the case when $m > 1$, from (3.8), the asymptotic mean square error for the kernel density estimate is $B^2 + \sigma^2$. Interestingly, the kernel and bandwidth appear only in the bias term B, not in the asymptotic variance nor in the rate of convergence. Further, in problems where the smoothness of h is known, it is easy to choose a kernel sufficiently smooth so that $B_n(t)$ is small enough so that $B=0$. For example, if we use a nonnegative kernel such as the normal curve or the Epanechnikov kernel, we can use $r=2$. In this case, to minimize the asymptotic mean square error from the Theorem, we only need to choose the bandwidth such that $b_n$ is small enough so that $b_n = o(n^{-1/4})$ but large enough so that $n\, b_n \rightarrow \infty$.

A complete solution to the problem of selection is not currently available, although there has been some promising work done on some special cases. Sheather and Hettmansperger (1985) have investigated the selection of the kernel and bandwidth for the example of scale parameter estimation in rank regression. A generalization of this parameter is the integrated squared density derivative, an estimation problem that has been recently investigated by several authors. To see the connection with the current paper, under strong enough smoothness conditions, the $m^{th}$ derivative of h(t) evaluated at 0 is $h^{(m)}(0)$ $= \int (f^{(m)}(u) + (-1)^m f^{(m)}(u))\, f(u)\, du$. Taking m to be even, we have $h^{(m)}(0)/2 = \int f^{(m)}(u)\, f(u)\, du = (-1)^{m/2} \int (f^{(m/2)}(u))^2\, du$. The last equality was pointed out by Hall and Marron (1987) who considered $(-1)^{m/2}\, h^{(m)}(0)/2$, among other estimators. Recent works such as Bickel and Ritov (1988), Jones and Sheather (1991) and Park and Marron (1992) have also considered the estimation of the integrated square density derivative, $\int (f^{(m)}(u))^2\, du$, since this is a key parameter in selecting the bandwidth for the usual kernel density estimate ($m=1$).

*V-Statistic Type Alternative Estimator:* In classical nonparametric statistics, an alternative to a U-statistic is the corresponding V-statistic. From a finite sample perspective, U-statistics enjoy the unbiasedness property while V-statistics are nonparametric maximum likelihood estimates. From a large sample perspective, the two versions are close to one another and have similar large sample properties. For more details, see, for example, Serfling (1980, Chapters 5 and 6). For this estimator, the V-statistic alternative is

$$h_{V,n}(t) = \frac{1}{n^m\, b_n} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} w\left( \frac{t-g(X_{i_1}, \ldots, X_{i_m})}{b_n} \right)$$

$$= \frac{1}{b_n} \int \ldots \int w\left(\frac{t-g(x_1,\ldots,x_m)}{b_n}\right) dF_n(x_1) \ldots dF_n(x_m)$$

where $\{F_n\}$ is the usual empirical distribution function. The choice between the two alternatives will probably depend upon the application at hand. For example, in the special case of the integrated squared density derivative, Jones and Sheather (1991) argue for the V-statistic version (in the language of this paper). For the spatial statistics problems, it can be argued that the U-statistic version is more intuitively appealing.

## 4. SUMMARY AND CONCLUDING REMARKS

The purpose of this paper is to introduce a nonparametric kernel density of a function of observations. A series of examples has been provided to motivate the usefulness of the estimator. Basic properties of the estimate have been established, including consistency and asymptotic normality. An important aspect of the asymptotics is the fast rate of convergence due to the additional averaging of a function of more than one observations.

As with the usual kernel density estimator of an observation, there are many possible alternative ways of estimating densities of functions of observations. See Silverman (1986) or Izenman (1991) for a discussion of these alternatives. This paper has focussed on the kernel method due to its intuitive appeal and widespread popularity.

Other extensions of the estimator similar to those found studied in the usual kernel estimator may be useful in applications. Perhaps a finite sample, or second order, study of the bias and variance would provide insights into the problem of selecting a kernel and bandwidth, a problem discussed in Section 3. Also in Section 3, we alluded to using the approach of this paper to study derivatives of densities of functions of observations. Other extensions could include the study of adaptive kernel estimates or functional central limit theorems. We cite these as areas for potential future study.

## ACKNOWLEDGEMENT

This work was partially supported by National Science Foundation grant SES-0022200.

## REFERENCES

Akritas, M. (1986), "Empirical processes associated with V-statistics and a class of estimators under random censoring," *Annals of Statistics* 14, 619-637.

Aubuchon, J. C. and Hettmansperger, T. P. (1989), "Rank-based inference for linear models: asymmetric errors," *Statistics & Probability Letters* 8, 97-107.

Bagchi, U., Ord, J. K. and Sullivan, R. S. (1986), "Parameter estimation for convolutions and compound distributions," *Operations Research Letters* 6, 301-308.

Bartlett, M. S. (1964), "Spectral analysis of two-dimensional point processes," *Biometrika* 44, 299-311.

Bickel, P. and Ritov, Y. (1988), "Estimating integrated squared density derivatives," *Sankyha Series A* 50, 381-393.

Blom, G. (1976), "Some properties of incomplete U-statistics," *Biometrika* 63, 573-580.

Box, G. and Tidwell, P. (1962), "Transformations of the independent variables," *Technometrics* 5, 317-325.

Cheng, K. F. and Serfling, R. J. (1981), "On estimation of a class of efficacy-related parameters," *Scandinavian Actuarial Journal*, 83-92.

Csörgõ, S., Horvath, L. and Serfling, R. (1983), "An approximation for the empirical process of U-statistic structure," Technical Report, Johns Hopkins University.

Dehling, H., Denker, M. and Philipp, W. (1987), "The almost sure invariance principle for the empirical process of U-statistic structure," *Annales de l'Institute Henri Pioncare* 23, 121-134.

Diggle, P. J. (1983), *Statistical Analysis of Spatial Point Patterns*, Academic Press, New York.

Frees, E. W. (1989), "Infinite order U-statistics," *Scandinavian Journal of Statistics* 16, 29-45.

Frees, E. W. (1991), "Trimmed slope estimates for simple linear regression," *Journal of Statistical Planning and Inference* 27, 203-221.

Greenwood, M. (1975), "Research on internal migration in the United States: A survey," *Journal of Economic Literature* 13, 397-433.

Hall, P. and Marron, J. (1987), "Estimation of integrated squared density derivatives," *Statistics and Probability Letters* 6, 109-115.

Haynes, K. E. and Fotheringham, A. S. (1984), *Gravity and Spatial Interaction Models*, Sage Publications, Beverly Hills, CA.

Hettmansperger, T. (1984), *Statistical Inference Based on Ranks*, Wiley, New York.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution," *Annals of Mathematical Statistics* 19, 193-225.

Hogg, R. V. and Klugman, S. A. (1984), *Loss Distributions*, Wiley, New York.

Izenman, A. J. (1991), " Recent developments in nonparametric density estimation," *Journal of the American Statistical Association* 86, 205-224.

Jammalamadaka, S. R. and Janson, S. (1986), "Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances," *Annals of Probability* 4, 1347-1358.

Janssen, P., Serfling, R. and Veraverbeke, N. (1984), "Asymptotic normality for a general class of statistical functions and applications to measures of spread," *Annals of Statistics* 12, 1369-1379.

Jones, M. and Sheather, S. (1991), "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives," *Statistics & Probability Letters* 11, 511-514.

Long, L, Tucker, C. and Urton, W. (1988), "Migration distances: An international comparison," *Demography* 25, 633-640.

Panjer, H. H. (1980), "The aggregate claims distribution and stop-loss reinsurance," *Transaction of the Society of Actuaries* XXXII, 523-545.

Park, B. and Marron, S. (1992), "On the use of pilot estimators in bandwidth selection," *Journal of Nonparametric Statistics* 1, 231-240.

Randles, R. and Wolfe, D. (1979), *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York.

Rosenblatt, M. (1971), "Curve estimates," *Annals of Mathematical Statistics* 42, 1815-1842.

Schweder, T. (1975), "Window estimation of the asymptotic variance of rank estimators of location," *Scandinavian Journal of Statistics* 2, 113-126.

Sen, P. (1960), "On some convergence properties of U-statistics," *Calcutta Statistical Association Bulletin* 10, 1-18.

Sen P. (1981), *Sequential Nonparametrics: Invariance Principles and Statistical Inference*, Wiley, New York.

Sen, P. (1983), "On the limiting behavior of the empirical kernel distribution function," *Calcutta Statistical Association Bulletin* 32, 1-8.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Sheather, S. J. and Hettmansperger, T. P. (1985), "A data based algorithm for choosing the window width when estimating the integral of $f^2(x)$," *Technical Report* Number 52, March. Department of Statistics, Pennsylvania State University, University Park, PA.

Shorack, G. and Wellner, J. (1986), *Empirical Processes with Applications to Statistics*, Wiley, New York.

Silverman, B. (1976), "Limit theorems for dissociated random variables," *Advances in Applied Probability* 8, 806-819.

Silverman, B. (1983), "Convergence of a class of empirical distribution functions of dependent random variables," *Annals of Probability* 11, 745-751.

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

Silverman, B. and Brown, T. (1978), "Short distances, flat triangles and Poisson limits," *Journal of Applied Probability* 15, 815-825.

Singh, R. S. (1981), "On the exact asymptotic behavior of estimators of a density and its derivatives," *Annals of Statistics* 9, 453-456.

Stone, C. J. (1980), "Optimal rates of convergence for nonparametric estimators," *Annals of Statistics* 8, 1348-1360.

Strauss, D. J. (1975), "A model for clustering," *Biometrika* 62, 467-475.

Weber, N. C. (1983), "Central limit theorems for a class of symmetric statistics," *Mathematical Proceedings of the Cambridge Philosophical Society* 94, 307-313.

## APPENDIX A. DATA

TABLE 1. 1989 TOTAL HOSPITAL CHARGES
FOR 33 FEMALES AGED 30-49 HOSPITALIZED FOR CIRCULATORY DISORDERS
FROM A WISCONSIN HOSPITAL

| | | | |
|---|---|---|---|
| 2337 | 2179 | 2348 | 4765 |
| 2088 | 2872 | 1924 | 2294 |
| 2182 | 2138 | 1765 | 2467 |
| 3609 | 2141 | 1850 | 3191 |
| 3020 | 2473 | 1898 | 7787 |
| 6169 | 1802 | 2011 | 2270 |
| 3425 | 3558 | 2315 | 1642 |
| 5878 | 2101 | 2242 | 5746 |
| 3041 | | | |

## APPENDIX B. PROOFS OF SECTION 3 PROPERTIES

An important device in the analysis is to write $h_n(t)$ as a functional of the EKDF $H_n$ and to use established properties of $H_n$. For example, from (1.3) we have

$$h_n(t) = b_n^{-1} \int w((t-u)/b_n)\, dH_n(u). \tag{A.1}$$

PROOF OF LEMMA: From (A.1), Assumption K1 and using integration by parts and a change of variables, we have

$$h_n(t) = b_n^{-1} \int (1 - H_n(t-y\, b_n))\, dw(y). \tag{A.2}$$

Taking expectations and applying Fubini's Theorem yields,

$$E\, h_n(t) = b_n^{-1} \int (1 - H(t-y\, b_n))\, dw(y). \tag{A.3}$$

By adding and subtracting $E\, h_n(t)$, (A.3), and the triangle inequality, we have

$$\begin{aligned} | h_n(t) - h(t) | &\le b_n^{-1} \left| \int (H_n(t-y\, b_n) - H(t-y\, b_n))\, dw(y) \right| + B_n(t) \\ &\le b_n^{-1} M_n \int | dw(y) | + B_n(t). \end{aligned}$$

Thus, Assumption K1 is sufficient for the result. ■

Define the projections, $W_{1n}(x,t) = b_n^{-1} E\, w( (t-g(x, X_2,\dots,X_m))/b_n ) - b_n^{-1} E\, w( (t-g(X_1,\dots,X_m))/b_n )$, and $\hat{h}_n(t) = n^{-1} \Sigma_{i=1}^{n} W_{1n}(X_i,t)$. In the following result, the estimate $\hat{h}_n(t)$ is decomposed into the parameter of interest, the projection, the bias and the remainder term. The proof of Theorem 1 is a straightforward application of Theorem A.1 and a triangular array central limit theorem and thus is omitted.

**Theorem A.1** Assume K1, $m > 1$, $n\, b_n \to \infty$, that $h_1$ exists, and, for $\delta > 0$, $E\, | h_1(X,t) |^{2+\delta} < \infty$. Define the remainder term, $R_n(t)$, by

$$\begin{aligned} h_n(t) &= E\, h_n(t) + \hat{h}_n(t) + R_n(t) \\ &= h(t) + \hat{h}_n(t) + B_n(t) + R_n(t). \end{aligned} \tag{A.4}$$

Then, $R_n(t) = O_p(b_n^{-1/2}\, n^{-1})$.

PROOF OF THEOREM A.1:

Define the degenerate kernel

$$K_n(x_1,\ldots,x_m) = b_n^{-1}\, w(\, (t-g(x_1,\ldots,x_m))/b_n\,) - b_n^{-1}\, E\, w(\, (t-g(X_1,\ldots,X_m))/b_n\,) - m^{-1}\, \Sigma_{j=1}^{m}\, W_{1n}(x_j,t).$$

From (A.4), we have

$$R_n(t) = \left(\begin{smallmatrix} n \\ m \end{smallmatrix}\right)^{-1} \Sigma_c\, K_n(X_{i_1}, \ldots, X_{i_m}).$$

Let $\{j_1, \ldots, j_m\}$ be another subset of $\{1, 2, \ldots, n\}$. Now,

$$\mathrm{Var}(R_n(t)) = \left(\left(\begin{smallmatrix} n \\ m \end{smallmatrix}\right)\right)^{-2} \Sigma_{c(i)}\, \Sigma_{c(j)}\, E\, K_n(X_{i_1}, \ldots, X_{i_m})\, K_n(X_{j_1}, \ldots, X_{j_m}).$$

Now, when the subsets $\{i_1, \ldots, i_m\}$ and $\{j_1, \ldots, j_m\}$ have 0 and 1 elements in common, the product $E\, K_n(X_{i_1}, \ldots, X_{i_m})\, K_n(X_{j_1}, \ldots, X_{j_m})$ is 0, by independence and the degeneracy of $K_n$. When there is more than 1 element in common, then by Chebyshev's theorem, the product is bounded by

$$\mathrm{Var}\, K_n \le b_n^{-2}\, E\, w(\, (t-g(X_1,\ldots,X_m))/b_n\,)^2 + E\, (W_{1n}(X_1,t))^2 = O(b_n^{-1}).$$

The proportion of evaluations having *c* elements in common is $\left(\begin{smallmatrix} m \\ c \end{smallmatrix}\right)\left(\begin{smallmatrix} n-m \\ m-c \end{smallmatrix}\right)/\left(\begin{smallmatrix} n \\ m \end{smallmatrix}\right)$, $c=0, 1, \ldots, m$. See, for example, Serfling (1980, p. 183). Thus, using $c=0$ and $c=1$,

$$\mathrm{Var}(R_n(t)) = (1 - \frac{1}{\binom{n}{m}}\, ( \binom{n-m}{m} + m\binom{n-m}{m-1} )\, )\, O(b_n^{-1}) = O(b_n^{-1} n^{-2})$$

which is sufficient for the result. ■

PROOF OF COROLLARY: By the Theorem, one only needs to show that $n\, E\, (h_R(t) - h_n(t))^2 \rightarrow 0$. This is immediate using the conditional independence of $\{g(X_1^{*r}, \ldots, X_m^{*r})\}$. ■