Journal of the KSQC
Vol. 18, No. 2, Nov. 1990

1990년 11월 韓國品質管理學會誌 제18권 제2호

Problems Occurred with Histogram and a Resolution+

Byeong Uk Park*
Hong Nae Park*
Moon Sup Song*
Jae Kee Song**

ABSTRACT

In this article, several problems inherent in histogram estimate of unknown probability density function are discussed. Those include so called sharp corners and bin edge effect. A resolution for these problems occurred with histogram is discussed. The resulting estimate is called kernel density estimate which is most widely used by data analysts. One of the most recent and reliable data-based choices of scale factor (bandwidth) of the estimate, which has been known to be most crucial, is also discussed.

1. Introduction

As a tool for exploring the unknown distributional structure of a population, histogram is

+This research was partially supported by the Basic Science Research Institute Program, Ministry of Education 1989.

* Department of Computer Science and Statistics, Seoul National University, Seoul, Korea.
**Department of Statistics, Kyungbook National University, Taegu, Korea.

well-known even to non-experts in statistics. However, histogram is known to have two serious defects. Those are called sharp corners and bin edge effect.

This article is intended to introduce for non-experts one of the existing and promising methods which does not have the above difficulties and is widely used by many data analysts. The estimate is called kernel density estimate and it is in essence a smoothed version of histogram. In section 2, this estimate is introduced as a resolution of the two problems occurred with histogram.

Kernel density estimate at a point is a weighted average of observations around the point. It is determined by two parameters. One of them determines shape of weights and the other controls amount of local averaging. Effective use of kernel density estimate is known to highly depend on the choice of the second parameter, bandwidth. Section 3 discusses this issue and introduces one of the most recent and reliable data-based methods for selecting bandwidth.

In the next section, the two problems with histogram are discussed.

2. The Problems and a Resolution

Let $X_1, \cdots, X_n$ be a random sample from a population with probability density function f. Histogram estimate $\tilde{f}$ of f is constructed as follows ;

Step 1 : Partition the real line into subintervals called bins indexed by $j=1, \cdots, m$.
Step 2 : For a point x, count the unmber of X_i's in the bin which x belongs to, call it bin frequency.
Step 3 : $\tilde{f}(x)$ = bin frequency / [$n \times$ bin width].

Sharp Corners

Suppose a point x is on the boundary of two adjacent bins. For this x, histogram estimate for $f(x)$ uses data only on one side of x and so creates bias because most data used to estimate $f(x)$ have different means from $f(x)$. This will be clear if we calculate the expected value of $\tilde{f}(x)$. Suppose the point x belongs to the j^{th} bin. Let h, B_j and f_j denote the bin width, the j^{th} bin and the frequency of the j^{th} bin, respectively. Then

$$\begin{aligned} E\tilde{f}(x) &= \frac{1}{nh} E[f_j] \\ &= \frac{1}{nh} nP[X \in B_j] \\ &= \frac{1}{h} \int_{B_j} f(t)\, dt. \end{aligned}$$

Hence the expected value of $\tilde{f}(x)$ is the average of f values over the j^{th} bin and generally this

tends to be more distant from $f(x)$ as x comes closer to the boundary.

This phenomenon of creating bias is typical for nonparametric estimators. However, the point is that the bias of histogram estimate is more severe than other nonparametric competitors, especially, than kernel density estimate given below.

Bin Edge Effect

A more important problem is concerned with grid location of histogram. The following figures show two different histograms constructed from the same set of data but with different grid locations.

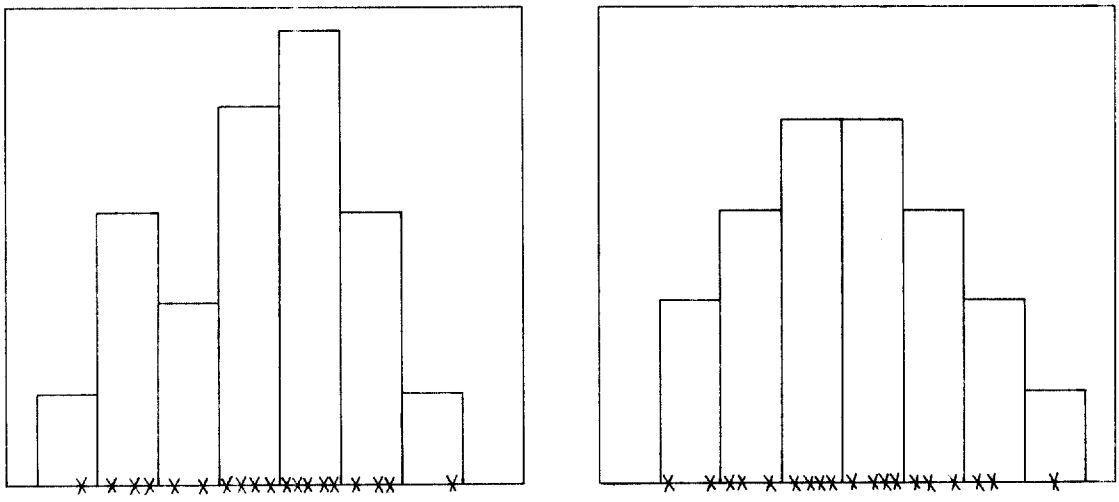

Figure 1. Histogram estimates constructed from the same set of data (x denotes a data point) but with different grid locations.

As is clearly seen in Figure 1, the first histogram catches the bimodal structure but the second does not. Thus histogram estimate may creat significantly different pictures about the same population depending on where one puts the grids. Furthermore, there is no objective rule to determine the grid location. This is the major defect which is inherent in histogram.

Kernel Density Estimate

A possible remedy for the first problem is to treat each point x as a bin center, which is introduced by Rosenblatt(1956). In other words, instead of using pre-determined bin B_j, one

uses $[x-h/2,\ x+h/2]$. With this, $\hat{f}(x)$ can be defined by

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n} I_{[x-h/2,\ x+h/2]}(X_i) \tag{2.1}$$

where

$$I_A(x)=\begin{cases}1, & \text{if } x\in A;\\ 0, & \text{otherwise.}\end{cases}$$

The estimate $\hat{f}(x)$ in (2.1) is a weighted average of observations around the point x with equal weights. It can be re-expressed as

$$\begin{aligned}\hat{f}(x)&=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}I_{[-h/2,\ h/2]}(x-X_i)\\ &=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}I_{[-1/2,\ 1/2]}\left(\frac{x-X_i}{h}\right).\end{aligned}$$

Note that $I_{[-1/2,\ 1/2]}(\cdot)$ is the probability density function of Uniform$[-1/2,\ 1/2]$.

This estimate still has one drawback, which is that it gives rough edges and this is not appropriate to estimate a smooth probability density function. One can overcome this difficulty by using a smooth probability density function K instead of $I_{[-1/2,\ 1/2]}$, which gives

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right). \tag{2.2}$$

The estimate defined in (2.2) is called kernel density estimate and it is introduced by Parzen (1962).

For the second problem, averaged shifted histogram introduced by Scott(1985) cna be used to eliminate the bin edge effect. The essential idea is to pool informations contained in several "shifted" histograms. In particular, note that histogram estimate can be written as

$$\tilde{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{\infty} I_{B_j}(x)\,I_{B_j}(X_i).$$

Let $\tilde{f}_l(x)$, $l=1,\ \cdots,\ m$, the l^{th} shifted histogram which is constructed by shifting the ordinary bins to the right by the amount lh/m. Then

$$\tilde{f}_l(x)=\frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{\infty} I_{B_j+lh/m}(x)\,I_{B_j+lh/m}(X_i).$$

Averaged shifted histogram is defined by

$$\hat{f}(x)=\frac{1}{m}\sum_{i=1}^{m}\tilde{f}_i(x).$$

It can be shown that this estimate again becomes the kernel density estimate defined in (2.2) with

$$K(x)=(1-|x|)I_{[-1,\ 1]}(x)$$

when m goes to infinity.

3. Choice of Bandwidth

Observe that the kernel density estimate defined in (2.2) is determined by two parameters, K and h. The fact that the choice of kernel function K is not so important is illustrated in Rosenblatt(1971). A common practice is to use the standard normal density for K. The more important and crucial parameter is h which is called bandwidth or smoothing parameter. Figure 2 shows how kernel density estimate depends on the choice of bandwidth.

As one can see it in Figure 2, smaller bandwidth yields more wiggling estimate and larger bandwidth yields smoother estimate. In fact, it is known that when the bandwidth is small, the kernel density estimate has small bias but has large variance and when the bandwidth is large, it has large bias but has small variance instead. Hence the optimal choice of the bandwidth h is the trade-off point between bias and variance. However, the optimal bandwidth is not available since it depends on the unknown probability density f.

There is a huge literature dealing with data-based bandwidth selection and all the methods are attempts to be close to the theoretical optimum. See Marron(1988) or Park and Marron(1990) for a survey of such methods proposed up until 1987. Recently many improved and reliable data-based bandwidth selection methods are proposed. Those include Hall, Marron and Park (1989), Hall, Sheather, Jones and Marron(1989), Sheather and Jones(1990) and Jones, Marron and Park(1990). Among these, only SJ(after Sheather and Jones) bandwidth selector is presented here because it has been found that SJ outperforms the other bandwidth selectors in the extensive simulation study conducted recently by Steve Marron. A full set of the simulation results is available from the author. A ready-to-use iterative algorithm for SJ bandwidth selector is now given and it is based on the use of the standard normal density kernel function for $\hat{f}$ in(2.2).

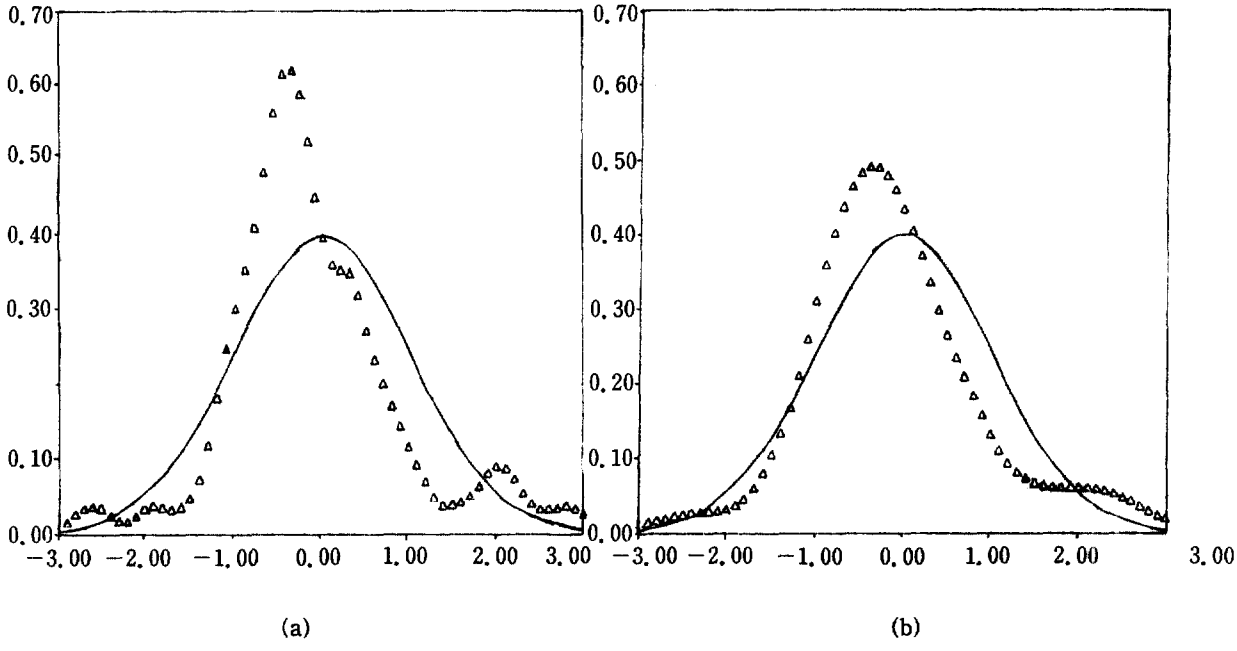

Figure 2. The solid curve denotes the true density ($N(0, 1)$). Kernel density estimates displayed ( ) are constructed using 50 data generated from $N(0, 1)$ with $K(x) = (15/8)(1-4x^2)\ I_{[-1/2,\ 1/2]}(x)$ and (a) $h=1.0$ (b) $h=2.0$.

Iterative algorithm for SJ bandwidth selector

First, compute the sample standard deviation SD of $X_1, \cdots, X_n$. Set

$$a = 1.241 SD n^{-1/7},$$
$$b = 1.231 SD n^{-1/9}.$$

Compute

$$A = [S(a)/T(b)]^{1/7}$$

where

$$S(a) = [n(n-1)^{-1}a^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{(4)}[a^{-1}(X_i - X_j)],$$

$$T(b) = -[n(n-1)^{-1}b^{-7}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{(6)}[b^{-1}(X_i - X_j)].$$

and $\phi^{(r)}$ is the r^{th} derivative of the standard normal density function.

Iteration 0 : Start with

$$h_0 = 1.102SDn^{-1/5}$$

Iteration k : Compute

$$a_k = 1.357A^{1/7}\ h_{k-1}{}^{5/7},$$

$$S_k = [n(n-1)]^{-1}a_k{}^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{(4)}\left(\frac{X_i - X_j}{a_k}\right),$$

$$h_k = [2\pi^{1/2}S_k]^{-1/5}n^{-1/5}.$$

Stop iteration when $|h_k - h_{k-1}| \leq \delta$ where δ is a pre-specified criterion value.

REFERENCE

1. Hall, P., Marron, J. S. and Park, B. U. (1989). Smoothed cross-validation. To appear.
2. Hall, P., Sheather, S., Jones, M. C. and Marron, J. S. (1989). On optimal data-based bandwidth selection in kernel density estimation. To appear.
3. Jones, M. C., Marron, J. S. and Park, B. U. (1990). A simple root n bandwidth selector. To appear.
4. Marron, J. S. (1988). "*Automatic smoothing parameter selection : A survey*", Empirical Economics, 13, 187-208.
5. Park, B. U. and Marron, J. S. (1990). "*Comparison of data-driven bandwidth selectors*", J. Amer. Statist. Assoc., 85, 66-72.
6. Parzen, E. (1962). "*On estimation of a probability density function and mode*", Ann. Math. Statist., 33, 1065-1076.
7. Rosenblatt, M. (1956). "*Remarks on some nonparametric estimates of a density function*", Ann. Math. Statist., 27, 832-837.
8. Rosenblatt, M. (1971). "*Curve estimates*", Ann. Math. Statist., 42, 1815-1842.
9. Scott, D. W. (1985). "Averaged shifted histograms : Effective nonparametric density estimators in several dimensions", Ann. Statist., 13, 1024-1040.
10. Sheather, S. and Jones, M. C. (1990). A reliable data-based bandwidth selection method for kernel density estimation. To appear.