Int. J. Contemp. Math. Sciences, Vol. 2, 2007, no. 29, 1451 - 1454

On the Minimized Error of Histogram Estimator on Location and Scale Density Function Family

E. Baloui Jamkhaneh[1] and M. Eshaghi Gordji[2]

[1] Azad University, Ghahemshahr Branch, Ghahemshahr, Iran
[2] Department of Mathematics, Semnan University, Semnan, Iran
e_baloui@yahoo.com & madjideg@walla.com

Abstract

In this article would like to introduce histogram estimators with fixed interval length and variable interval length. We show that if interval length is function of x, then histogram estimator is more efficiency. Finally we prove that in location and scale density function family minimum error with scale parameter have inverse relationship and relative efficiency with respect to any values of location and scale parameters is constant.

Mathematics Subject Classification: 62G

Keywords: Scale density function, Histogram estimator

Introduction

Let $f(x)$ be a density function and $X_1, X_2, \ldots, X_n$ a random sample of size n from a population whose underlying density is $f(x)$. Let $v(x)$ be the number of sample points having values less than or equal to x. Then a natural approximation to the cumulative distribution function F is

$$\hat{F}_n(x) = \frac{v(x)}{n}.$$

Similarly, a natural approximation to the probability density function f is

$$\hat{f}_k(x_k) = \frac{v(x_k) - v(x_{k-1})}{n(x_k - x_{k-1})}$$

where the x_k are points defined by a mesh on the real line and where $v_k = v(x_k) - v(x_{k-1})$ is the number of sample point falling in the kth interval

$[x_{k-1}, x_k]$. A histogram estimate is a step function $\hat{f}(x)$ of height $\hat{f}_k$ along each such interval.

The integrated mean squared error was used where as a global measure of error. Huesemann and Terrell [1] shown the integrated mean squared error (IMSE) of $\hat{f}(x)$ is

$$IMSE(\hat{f}(x)) = \int_{-\infty}^{+\infty} var(\hat{f}(x))dx + \int_{-\infty}^{+\infty} Bias^2(\hat{f}(x))dx.$$

Let $f'(x)$ be square Riemann integrable, with $f'(x_k) \neq 0$, and defined over the entire real line. Let h be the length of each interval so that $h(n) \longrightarrow 0$ as $n \longrightarrow \infty$ and $nh(n) \longrightarrow \infty$ as $n \longrightarrow \infty$; then the asymptotically optimal fixed interval length can be shown to be [2]

$$h^* = [\frac{6}{n\int_{-\infty}^{+\infty}(f'(x))^2dx}]^{\frac{1}{3}}. \tag{1}$$

When the optimal constant interval length h^* is used, the following minimal integrated mean squared error is obtained:

$$IMSE_f^* = \frac{3}{2}6^{\frac{1}{3}}n^{-\frac{2}{3}}[\int_{-\infty}^{+\infty}(f'(x))^2dx]^{\frac{1}{3}}. \tag{2}$$

Terrell and Scott [3] and Kogure [4] addressed this issue for probability densities of one random variable. If h becomes a function of x, then the asymptotically optimal interval length at one point becomes

$$h^*(x) = \left\{\frac{6f(x)}{n(f'(x))^2}\right\}^{\frac{1}{3}}, \tag{3}$$

and the corresponding intergrated mean squared error becomes

$$IMSE^* = \frac{3}{2}6^{-\frac{1}{3}}n^{-\frac{2}{3}}\int_{-\infty}^{+\infty}(f(x)f'(x))^{\frac{2}{3}}dx, \tag{4}$$

we obtain:

$$\begin{aligned}\frac{IMSE_f^*}{IMSE^*} &= \frac{\left\{\int_{-\infty}^{+\infty}(f'(x))dx\right\}^{\frac{1}{3}}}{\int_{-\infty}^{+\infty}\{f(x)f'(x)\}^{\frac{2}{3}}dx} = \frac{\left\{\int_{-\infty}^{+\infty}(f'(x))^2dx\right\}^{\frac{1}{3}}\left\{\int_{-\infty}^{+\infty}f(x)dx\right\}}{\int_{-\infty}^{+\infty}\{f(x)f'(x)\}^{\frac{2}{3}}dx} \\ &= \frac{\left\{\int_{-\infty}^{+\infty}((f'(x))^{\frac{2}{3}})^3dx\right\}^{\frac{1}{3}}\left\{\int_{-\infty}^{+\infty}((f(x))^{\frac{2}{3}})^{\frac{3}{2}}dx\right\}}{\int_{-\infty}^{+\infty}(f(x))^{\frac{2}{3}}(f'(x))^{\frac{2}{3}}dx}.\end{aligned}$$

Now by the inequality of Hölder, above ratio is always being greater/or equal to one; that is $IMSE^* \leq IMSE_f^*$ which mean if h is function of x therefore estimator is more efficiency and amount of the efficiency can be find for different density.

Example. Let $f(x)$ is standard normal density function; then

$$\left\{\int_{-\infty}^{+\infty}(f'(x))^2dx\right\}^{\frac{1}{3}}=\left\{\int_{-\infty}^{+\infty}\frac{x^2}{2\pi}e^{-x^2}dx\right\}^{\frac{1}{3}}=0.52,$$

if the h is constant then with (1) we will have $h^*=3.49n^{-\frac{1}{3}}$ and when h is function of x, when with (3) we have:

$$h^*=2.468|x|^{-\frac{2}{3}}e^{\frac{x^2}{6}}n^{-\frac{1}{3}},$$

and

$$\begin{aligned}\int_{-\infty}^{+\infty}(f(x)f'(x))^{\frac{2}{3}}dx &= \int_{-\infty}^{+\infty}\frac{x^{\frac{2}{3}}}{(2\pi)^{\frac{2}{3}}}e^{-\frac{2}{3}x^2}dx=\frac{1}{(2\pi)^{\frac{2}{3}}}\int_0^{\infty}u^{-\frac{1}{6}}e^{-\frac{2}{3}u}du\\ &= \frac{1}{(2\pi)^{\frac{2}{3}}}\Gamma(\frac{5}{6})(\frac{3}{2})^{\frac{5}{6}}=0.464.\end{aligned}$$

Therefore by using (5) we have:

$$\frac{IMSE_f^*}{IMSE^*}=\frac{0.52}{0.464}=1.12.$$

1 Minimal IMSI this location and scale density function family

Let $f(X)$ be a density function from location and scale density function family, that is

$$f(x)=\frac{1}{\sigma}g(\frac{x-\mu}{\sigma});\quad \sigma>0;\mu\in\mathbb{R},$$

(μ is called the location parameter and σ is called the scale parameter and $g(.)$ is a density function) then

$$\left\{\int_x(f'(x))^2\right\}^{\frac{1}{3}}=\left\{\int_x\frac{1}{\sigma^4}\left(g'(\frac{x-\mu}{\sigma})\right)^2dx\right\}^{\frac{1}{3}}=\frac{1}{\sigma}\int_u(g'(u))^2du, \qquad (5)$$

with (2) we see $IMSE_f^*$ with scale parameter have inverse relationship. Whenever if h be function of x, then

$$\int_x(f'(x)f(x))^{\frac{2}{3}}dx = \int_x\left\{\frac{1}{\sigma^3}g'(\frac{x-\mu}{\sigma})g(\frac{x-\mu}{\sigma})\right\}^{\frac{2}{3}}dx=\frac{1}{\sigma}\int_u(g'(u)g(u))^{\frac{2}{3}}du.$$

It mean that in this case; with considering (4) $IMSE^*$ with scale parameter have inverse relationship. Noticing to the variance in the density with scale parameter have relationship, then the more variance of variable in the population, then the more variance of variable in the population, the less minimal integrated mean squared error. Although histogram estimator for the variable interval case have more efficiency, but the relative efficiency of this two kinds of estimators with respect to any values of parameters is constant.

Example. Let $f(x)$ is normal density function with mean μ and variance σ^2, then

$$\begin{aligned}\left\{\int_{-\infty}^{+\infty}(f'(x))^2dx\right\}^{\frac{1}{3}} &= \left\{\int_{-\infty}^{+\infty}\frac{(x-\mu)^2}{2\pi\sigma^2}\exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\}dx\right\}^{\frac{1}{3}}\\ &= \left\{\int_0^{\infty}\frac{1}{2\pi\sigma^3}u^{\frac{1}{2}}e^{-u}du\right\}^{\frac{1}{3}} = \frac{0.52}{\sigma}.\end{aligned}$$

Then

$$h^* = 3.49n^{-\frac{1}{3}}\sigma, \quad IMSE_f^* = 0.429n^{-\frac{2}{3}}\sigma^{-1}.$$

When h is function of x, then $IMSE^* = 0.383n^{-\frac{2}{3}}\sigma^{-1}$. It has been proven on either condition if the variance random variable is greater then $IMSE^*$ would be less. But the ratio $\frac{IMSE_f^*}{IMSE}$ in population with normal distribution with mean μ and variance σ^2 is equal 1.12. Which mean the efficiency of two methods about the different values of parameters is constant.

Result which we got in some density of this family shown in following table:

Density function		$IMSE^*$	$IMSE_f^*$	Ratio
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$0.383^{-\frac{2}{3}}\sigma^{-1}$	$0.429n^{-\frac{2}{3}}\sigma^{-1}$	1.12
Laplace	$\frac{1}{\beta}\exp\left\{-\frac{\lvert x-\alpha\rvert}{\beta}\right\}$	$0.5119n^{-\frac{2}{3}}\beta^{-1}$	$0.491n^{-\frac{2}{3}}\beta^{-1}$	1.0582
Exponential	$\frac{1}{\lambda}\exp\left\{-\frac{(x-\mu)}{\lambda}\right\}$	$0.61905n^{-\frac{2}{3}}\lambda^{-1}$	$0.6551n^{-\frac{2}{3}}\lambda^{-1}$	1.0582
Pareto	$\frac{\theta}{x^2}$	$0.5615n^{-\frac{2}{3}}\theta^{-1}$	$0.766n^{-\frac{2}{3}}\theta^{-1}$	1.3645

References

[1] J. A. Huesmann and G. R. Terrell (1991). Optimal parameter choice for error minimization in Bivariate histograms. Journal of multivariate analysis 37, 85-103.

[2] D. W. Scott (1979). On optimal and eta-based histogram. Biometrika 66 606-610.

[3] G. R. Terrell and D. W. Scott (1983). Variable window density estimates. Presented to the statistical joint meetings, Toronto, Canada, Angust.

[4] A. Kigure (1987). Asymptotically optimal cells for a histogram. Ann. Statist. 15 1023-1030.

Received: September 9, 2006