DOI:10.13656/j.cnki.gxkx.2012.01.019

广西科学 Guangxi Sciences 2012,19(1):28～30,34

频率插值密度估计的渐近无偏性与相合性*

Asymptotic Unbiasedness and Consistency of Frequency Interpolation Density Estimation

吴果林,王彦辉

WU Guo-lin,WANG Yan-hui

(桂林航天工业高等专科学校信息工程系,广西桂林 541004)

(Department of Information Engineering,Guilin College of Aerospace Techology,Guilin,Guangxi,541004,China)

摘要:从频率插值的定义出发,证明单变量频率插值函数是总体分布密度的一个相合估计且渐近无偏.

关键词:密度估计 频率插值 直方图 渐近无偏性 相合性

中图法分类号:O212.7 **文献标识码**:A **文章编号**:1005-9164(2012)01-0028-03

Abstract:Based on the definition of the frequency interpolation,which is more reasonable and converge faster than histogram estimation,a single variable frequency interpolation function is proved to be the consistent and asymptotic unbiased estimator.

Key words:density estimation,frequency interpolation,histogram,asymptotic unbiase,consistency

在非参数统计领域研究样本对应总体的分布时,直方图技术一直处于非常重要的地位,扮演着经典角色.该技术以简单、直观、易懂等优点在密度估计、数据分析等领域中为大众所接受,而且随着样本量的增加,直方图同样也能很好地估计出总体分布特征.Chen 等[1]和 Zhao 等[2]从理论上证明了直方图估计密度函数的几乎处处收敛性.然而,直方图估计 $\hat{f}(x)$ 有两个明显的缺点:其一,$\hat{f}(x)$ 在区间段内为一个常数,是不连续函数,这与实际的密度函数 $f(x)$ 不相符;其二,$\hat{f}(x)$ 的收敛速度太慢,在相同的积分均方误差(MISE)下,直方图估计 $\hat{f}(x)$ 所需的样本很大[3].鉴于直方图估计的上述缺点,Scott,D. W[4]在直方图技术的基础上,提出了频率插值密度估计,其构造如下:

给定一组样本观测值 $x_1,x_2,\cdots,x_n$,对此进行排序,并设 $x_{(1)}$ 和 $x_{(n)}$ 为最小和最大样本观测值,确定最小下界 $t_0 \leqslant x_{(1)}$;

估计组距 h,可得每组分界点 $t_0,t_1,t_2,\cdots,t_k$,其中,$t_{i+1}-t_i=h,i=0,1,\cdots,k-1,x_{(n)} \leqslant t_k < x_{(n)}+h$;利用直方图制作技术,构建样本的直方图,得

$$\hat{f}_i=\frac{v_i}{nh},x \in (t_{i-1},t_i],\ i=0,1,\cdots,k-1, \tag{0.1}$$

其中 v_i 表示样本落在区间 $(t_{i-1},t_i]$ 的频数;

记 $(t_{i-1},t_i]$ 的中点为 a_i,以点 $(t_0,\hat{f}_1),(a_1,\hat{f}_1),(a_2,\hat{f}_2),\cdots,(a_k,\hat{f}_k),(t_k,\hat{f}_k)$ 作为插值节点,作分段线性插值函数;

由样本构建的上述插值函数 $\hat{f}_{FP}(x)$ 作为总体分布的密度估计,其形式为

$$\hat{f}_{FP}(x)=\begin{cases}\dfrac{v_1}{nh},x \in [t_0,a_1],\\ \dfrac{x-a_{i+1}}{a_i-a_{i+1}}\hat{f}_i+\dfrac{x-a_i}{a_{i+1}-a_i}\hat{f}_{i+1},\\ \qquad x \in (a_i,a_{i+1}],i=1,\cdots,k-1,\\ \dfrac{v_k}{nh},x \in (a_k,t_k].\end{cases} \tag{0.2}$$

显然,(0.2)式所表示的频率插值函数 $\hat{f}_{FP}(x)$ 是区间 $[t_0,t_k]$ 连续函数,且易证

$$\int_{t_0}^{t_k}\hat{f}_{FP}(x)\mathrm{d}x=1.$$

与制作直方图一样,构建频率插值的关键是确定

收稿日期:2011-04-07
修回日期:2011-11-15

作者简介:吴果林(1977-),男,讲师,硕士,主要从事非参数统计、数值计算研究。

*2011 年度广西教育厅科研项目(201106LX054)资助。

组距 h，组距 h 的选择在很大程度上影响频率插值的收敛效果和收敛速度. Scott，D. W[4] 利用积分均方误差（MISE）准则，在假定 $f''(x)$ 绝对连续及 $\int_{-\infty}^{+\infty} f''(x)\mathrm{d}x < \infty$ 的情况下，得出最优组距 $h^* = O(n^{-\frac{1}{5}})$，渐近积分均方误差（AMISE）的收敛速度为 $O(n^{-\frac{4}{5}})$. 之后，文献[5，6]又从文献[4]构建最优组距的公式出发，讨论了频率插值最优组距的上界问题. 易见，频率插值函数是分段一次线性函数，具有一致收敛性，但光滑性较差，而大多数连续型随机变量的分布密度函数（如正态、指数、对数正态分布等）都具有较高的光滑度，显然这与实际情形不相符. 为此，文献[7]提出用频率直方图的中点作为插值节点，通过这些节点构建一个三次自然样条函数，作为总体分布密度的估计——频率样条插值密度估计。该估计具有二阶光滑度，弥补了频率多边形函数光滑性较差的缺点，且收敛速度与频率多边形一样，也是 $O(n^{-\frac{4}{5}})$.

然而，频率插值 $\hat{f}_{FP}(x)$ 作为总体分布密度 $f(x)$ 的一个估计，相合性是一个最基本的要求，即对于任意一个 $\varepsilon > 0$ 及给定点 x，需满足

$$\lim_{n\to\infty} P(|\hat{f}_{FP}(x) - f(x)| \geqslant \varepsilon) = 0. \tag{0.3}$$

本文从频率插值的定义出发，先证明频率插值 $\hat{f}_{FP}(x)$ 是密度函数 $f(x)$ 渐近无偏估计，再在此基础上证明 $\hat{f}_{FP}(x)$ 为 $f(x)$ 的相合估计，即有（0.3）式成立.

1 相关引理

引理 1[3] 设密度函数 $f(x)$ 在区间 $[t_0, t_k]$ 上连续，由（0.1）式定义的 $\hat{f}_i = \frac{v_i}{nh}$ 是密度函数 $f(x)$ 在区间 $(t_{i-1}, t_i]$ 上的直方图估计，则

(1) $v_i \sim B(n, p_i)$，其中 $p_i = \int_{t_{i-1}}^{t_i} f(x)\mathrm{d}x$，

(1.1)

(2) $E(\hat{f}_i) = \frac{p_i}{h}$，$\mathrm{var}(\hat{f}_i) = \frac{p_i(1-p_i)}{nh^2}$，$\mathrm{Cov}(\hat{f}_i, \hat{f}_{i+1}) = \frac{-p_i p_{i+1}}{nh^2}$. (1.2)

引理 2[3] 设密度函数 $f(x)$ 在区间 $[t_0, t_k]$ 上满足利普希茨（Lipschitz）条件，由（0.1）式定义的 $\hat{f}_i = \frac{v_i}{nh}$ 是密度函数 $f(x)$ 在区间 $(t_{i-1}, t_i]$ 上的直方图估计，且当 $n \to \infty$ 时，有 $h \to 0$ 和 $nh \to \infty$，则对任意 $x \in (t_{i-1}, t_i]$，有

$$\lim_{n\to\infty} \mathrm{MSE}\{\hat{f}_i\} = 0.$$

其中 $\mathrm{MSE}\{\hat{f}_i\} = \mathrm{Var}(\hat{f}_i) + \mathrm{Bias}^2(\hat{f}_i)$，这里 $\mathrm{Bias}(\hat{f}_i) = E(\hat{f}_i) - f(x)$.

引理 3 设密度函数 $f(x)$ 在区间 $[t_0, t_k]$ 上连续，且当 $n \to \infty$ 时，有 $h \to 0$ 和 $nh \to \infty$，则由（0.2）式定义的频率插值 $\hat{f}_{FP}(x)$，对任意 $x \in [t_0, t_k]$，有

$$\lim_{n\to\infty} \mathrm{Var}[\hat{f}_{FP}(x)] = 0. \tag{1.3}$$

证明 根据频率插值 $\hat{f}_{FP}(x)$ 的定义，由引理 2 可知，只需证明当 $x \in [a_i, a_{i+1}]$，$i = 1, \cdots, (k-1)$ 时，（1.3）式成立.

对任意 $x \in [a_i, a_{i+1}]$，有

$$\mathrm{Var}[\hat{f}_{FP}(x)] = \mathrm{Var}[\frac{x-a_{i+1}}{a_i - a_{i+1}}\hat{f}_i + \frac{x-a_i}{a_{i+1}-a_i}\hat{f}_{i+1}] = \frac{(x-a_{i+1})^2}{h^2}\mathrm{Var}(\hat{f}_i) + \frac{(x-a_i)^2}{h^2}\mathrm{Var}(\hat{f}_{i+1}) - 2\frac{(x-a_{i+1})(x-a_i)}{h^2}\mathrm{Cov}(\hat{f}_i, \hat{f}_{i+1}). \tag{1.4}$$

由引理 1（1.2）式得

$$\mathrm{Var}(\hat{f}_i) = \frac{p_i(1-p_i)}{nh^2}, \tag{1.5}$$

$$\mathrm{Var}(\hat{f}_{i+1}) = \frac{p_{i+1}(1-p_{i+1})}{nh^2}, \tag{1.6}$$

$$\mathrm{Cov}(\hat{f}_i, \hat{f}_{i+1}) = \frac{-p_i p_{i+1}}{nh^2}. \tag{1.7}$$

将（1.5）式、（1.6）式、（1.7）式代入（1.4）式，整理得

$$\mathrm{Var}[\hat{f}_{FP}(x)] = \frac{(x-a_{i+1})^2}{nh^4}p_i + \frac{(x-a_i)^2}{nh^4}p_{i+1} - \frac{[p_i(x-a_{i+1}) - p_{i+1}(x-a_i)]^2}{nh^4} \leqslant \frac{(x-a_{i+1})^2}{nh^4}p_i + \frac{(x-a_i)^2}{nh^4}p_{i+1} \leqslant \frac{p_i}{nh^2} + \frac{p_{i+1}}{nh^2}. \tag{1.8}$$

结合引理 1（1.1）式，且 $f(x)$ 在区间 $[t_{i-1}, t_{i+1}]$ 上连续，故（1.8）式可化为

$$\frac{p_i}{nh^2} + \frac{p_{i+1}}{nh^2} = \frac{\int_{t_{i-1}}^{t_i} f(x)\mathrm{d}x}{nh^2} + \frac{\int_{t_i}^{t_{i+1}} f(x)\mathrm{d}x}{nh^2} = \frac{\int_{t_{i-1}}^{t_{i+1}} f(x)\mathrm{d}x}{nh^2}. \tag{1.9}$$

由积分中值定理得 $\int_{t_{i-1}}^{t_{i+1}} f(x)\mathrm{d}x = 2hf(\theta)$，其中 $\theta \in [t_{i-1}, t_{i+1}]$，则

$$\frac{p_i}{nh^2} + \frac{p_{i+1}}{nh^2} = \frac{2hf(\theta)}{nh^2} = \frac{2f(\theta)}{nh}. \tag{1.10}$$

将（1.10）式代入（1.8）式，得

$$\mathrm{Var}[\hat{f}_{FP}(x)] \leqslant \frac{2f(\theta)}{nh} \leqslant \frac{2M}{nh},$$

其中 $M = \max\{f(x), x \in [t_0, t_k]\}$. 因此当 $n \to \infty$

时，$h \to 0$ 和 $nh \to \infty$ 时 (1.3)式成立.

2 主要结果

定理 1 设密度函数 $f(x)$ 在区间 $[t_0, t_k]$ 上满足利普希茨(Lipschitz)条件，具有连续二阶导数，且当 $n \to \infty$ 时，$h \to 0$，则由(0.2)式定义的频率插值 $\hat{f}_{FP}(x)$ 对任意 $x \in [t_0, t_k]$，有

$$\lim_{n\to\infty} E[\hat{f}_{FP}(x)] = f(x). \tag{2.1}$$

证明 首先证明当 $x \in [t_0, a_1]$ 时 (2.1)式成立.

当 $x \in [t_0, a_1]$ 时，由(0.2)式有 $\hat{f}_{FP}(x) = \frac{v_1}{nh}$，则由引理 1 可得

$$E[\hat{f}_{FP}(x)] - f(x) = \frac{p_1}{h} - f(x).$$

由积分中值定理得

$$p_1 = \int_{t_0}^{t_1} f(x)\mathrm{d}x = hf(\xi_1)，\text{其中 } \xi_1 \in [t_0, t_1].$$

又因为 $f(x)$ 满足利普希茨(Lipschitz)条件，于是有

$$|E[\hat{f}_{FP}(x)] - f(x)| = |f(\xi_1) - f(x)| \leqslant \gamma_1 |\xi_1 - x| \leqslant \gamma_1 h, \tag{2.2}$$

其中 γ_1 为一正常数. 因此当 $n \to \infty, h \to 0$ 时(2.1)式成立.

同理可证当 $x \in (a_k, t_k]$，有(2.1)式成立.

其次证明当 $x \in [a_i, a_{i+1}]$ 时(2.1)式成立.

当 $x \in [a_i, a_{i+1}]$ 时，由(0.2)式有

$$\hat{f}_{FP}(x) = \frac{x - a_{i+1}}{a_i - a_{i+1}}\hat{f}_i + \frac{x - a_i}{a_{i+1} - a_i}\hat{f}_{i+1}.$$

于是，由引理 1(1.2)式可得

$$E[\hat{f}_{FP}(x)] - f(x) = \frac{x - a_{i+1}}{a_i - a_{i+1}}\frac{p_i}{h} + \frac{x - a_i}{a_{i+1} - a_i} \cdot \frac{p_{i+1}}{h} - f(x) = \frac{x - a_i}{h^2}p_{i+1} + \frac{a_{i+1} - x}{h^2}p_i - f(x). \tag{2.3}$$

由于 $f(x)$ 在区间 $[a_i, a_{i+1}]$ 上具有连续二阶导数，对 $f(x)$ 在点 $x = t_i$ 处进行泰勒展开得

$$f(x) = f(t_i) + f'(t_i)(x - t_i) + \frac{f''(\xi_i)}{2!}(x - t_i)^2, \tag{2.4}$$

其中 ξ_i 介于 x 与 t_i 之间. 于是

$$p_i = \int_{t_{i-1}}^{t_i} f(x)\mathrm{d}x = \int_{t_{i-1}}^{t_i} [f(t_i) + f'(t_i)(x - t_i) + \frac{f''(\xi_i)}{2!}(x - t_i)^2]\mathrm{d}x = f(t_i)h - \frac{f'(t_i)}{2}h^2 + \frac{f''(\xi_i)}{6}h^3 = f(t_i)h - \frac{f'(t_i)}{2}h^2 + O(h^3). \tag{2.5}$$

同理可得

$$p_{i+1} = \int_{t_i}^{t_{i+1}} f(x)\mathrm{d}x = f(t_i)h + \frac{f'(t_i)}{2}h^2 + O(h^3). \tag{2.6}$$

将(2.5)式和(2.6)式代入(2.3)式，考虑到 $f(x)$ 满足利普希茨(Lipschitz)条件，有

$$|E[\hat{f}_{FP}(x)] - f(x)| = \left|\frac{x - a_i}{h^2}p_{i+1} + \frac{a_{i+1} - x}{h^2}p_i - f(x)\right| = \left|\left[\frac{x - a_i}{h}f(t_i) + \frac{a_{i+1} - x}{h}f(t_i)\right] + \frac{f'(t_i)}{2}(2x - a_i - a_{i+1}) + O(h^2) - f(x)\right| \leqslant |f(t_i) - f(x)| + \left|\frac{f'(t_i)}{2}(2x - a_i - a_{i+1})\right| + |O(h^2)| \leqslant \gamma_i |t_i - x| + \left|\frac{f'(t_t)}{2}\right|(|x - a_i| + |x - a_{i+1}|) + |O(h^2)| \leqslant \gamma_i h + |f'(t_i)|h + |O(h^2)|,$$

其中 γ_i 为一正常数. 因此当 $n \to \infty, h \to 0$ 时(2.1)式成立.

定理 2 设密度函数 $f(x)$ 在区间 $[t_0, t_k]$ 上满足利普希茨(Lipschitz)条件，具有连续二阶导数，且当 $n \to \infty$ 时，有 $h \to 0$ 和 $nh \to \infty$，则由(0.2)式定义的频率插值 $\hat{f}_{FP}(x)$ 对任意 $x \in [t_0, t_k]$，有

$$|\hat{f}_{FP}(x) - f(x)| \xrightarrow{P} 0.$$

证明 任取一点 $x \in [t_0, t_k]$，对任意的 $\varepsilon > 0$，由切比雪夫不等式有

$$P(|\hat{f}_{FP}(x) - E[\hat{f}_{FP}(x)]| \geqslant \frac{\varepsilon}{2}) \leqslant \frac{4}{\varepsilon^2}\mathrm{Var}(\hat{f}_{FP}(x)).$$

另一方面，由定理 1 可知 $\lim_{n\to\infty} E[\hat{f}_{FP}(x)] = f(x)$，故当 n 充分大时有

$$|E[\hat{f}_{FP}(x)] - f(x)| < \frac{\varepsilon}{2}.$$

注意，此时如果 $|\hat{f}_{FP}(x) - E[\hat{f}_{FP}(x)]| < \frac{\varepsilon}{2}$，就有

$$|\hat{f}_{FP}(x) - f(x)| \leqslant |\hat{f}_{FP}(x) - E[\hat{f}_{FP}(x)]| + |E[\hat{f}_{FP}(x)] - f(x)| < \varepsilon,$$

故

$$\{|\hat{f}_{FP}(x) - E[\hat{f}_{FP}(x)]| < \frac{\varepsilon}{2}\} \subset \{|\hat{f}_{FP}(x) - f(x)| < \varepsilon\}.$$

等价地

$$\{|\hat{f}_{FP}(x) - E[\hat{f}_{FP}(x)]| \geqslant \frac{\varepsilon}{2}\} \supset \{|\hat{f}_{FP}(x) - f(x)| \geqslant \varepsilon\},$$

由此即有

(下转第 34 页 Continue on page 34)

lag,1996:1-11.

[3] Kemnitz A,Marangio M,Mihók P. [r,s,t]-chromatic numbers and hereditary properties of graphs[J]. Discrete Mathematics,2007,307:916-922.

[4] 龚劬,张新军. 二部图的[r,s,t]-着色[J]. 重庆大学学报:自然科学版,2007,30(12):95-97.

[5] Dekar L,Effantin B,Kheddouci H. [r,s,t]-coloring of trees and bipartite graphs[J]. Discrete Mathematics,2008,311:1521-1533.

[6] 俞竺君,左连翠. 含点不交偶圈的图的[r,s,t]-着色[J]. 天津师范大学学报:自然科学版,2010,33(2):18-22.

[7] Bondy J A,Murty U S R. Graph theory[M]. Berlin:Springer,2008.

[8] 李光海,李武装. 关于几类图的邻点可区别全染色[J]. 河南师范大学学报:自然科学版,2006,35(1):139-140.

(责任编辑:陈小玲)

(上接第 30 页 Continue from page 30)

$$P(|\hat{f}_{FP}(x)-f(x)|\geqslant\varepsilon)\leqslant P(|\hat{f}_{FP}(x)-E[\hat{f}_{FP}(x)]|\geqslant\varepsilon/2)\leqslant\frac{4}{\varepsilon^2}\mathrm{Var}(\hat{f}_{FP}(x)).$$

由引理 3 可知,当 $n\to+\infty$,$\mathrm{Var}(\hat{f}_{FP}(x))\to 0$. 定理 2 证明完毕.

参考文献:

[1] Chen X R,Zhao L C. Almost sure $L[1]$-norm convergence for data-based histogram density estimates[J]. Journal of Multivariate Analysis,1987,21:179-188.

[2] Zhao L C,Krishnaiah P R,Chen X R. Almost sure $L[r]$-norm convergence for data-based histogram density estimates[J]. Theory of Probability and Its Applications,1990,35:396-403.

[3] Scott D W. Multivariate density estimation-theory,practice and visualization[M]. New York:John Wiley & Sons,1992.

[4] Scott D W. Frequency polygons[J]. J Amer Statist Assoc,1985,80:348-354.

[5] Scott D W,Terrell G R. Biased and unbiased cross-validation in density estimation[J]. J Amer Statist Assoc,1987,82:1131-1146.

[6] Terrell G R. The maximal smoothing principle in density estimation[J]. J Amer Statist Assoc,1990,85:470-477.

[7] 吴果林,张德全. 频率样条插值密度估计[J]. 桂林航天工业高等专科学校学报,2010,58(2):256-258.

(责任编辑:尹 闯)