

综合报告

直方图理论与最优直方图制作 *

张建方

(中国科学院研究生院管理学院, 北京, 100190)

王秀祥

(中国民生银行工商企业金融事业部杭州风险管理部, 杭州, 310009)

摘要

直方图是一种最为常见的密度估计和数据分析工具。在直方图理论和制作过程中, 组距的选择和边界点的确定尤为重要。然而, 许多学者对这两个参数的选择仍然采用经验的方法, 甚至现在大多数统计软件在确定直方图分组数时也是默认采用粗略的计算公式。本文主要介绍直方图理论和最优直方图制作的最新研究成果, 强调面向样本的最优直方图制作方法。

关键词: 直方图, Sturges公式, Scott公式, Cross-Validation, Histogram-Kernel Error, 误差平方和。

学科分类号: O212.7.

§1. 前 言

在非参数统计领域, 研究样本对应总体的分布, 直方图技术一直处于非常重要的地位, 扮演着经典角色。和核密度估计相比, 虽然直方图不能给出较为精确的样本密度估计, 但其以简单、直观、易懂等优点在密度估计、数据分析等过程中为大众所接受。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。直方图是用矩形的宽度和高度来表示频数分布的图形。如在直角坐标系中, 以 x 轴表示所考察的数据变量, y 轴表示频数, 再以每一组的区间为底, 该区间的频数为高作矩形, 即可得到该样本数据的频数直方图。直方图是总体密度曲线的一种近似, Chen and Zhao (1987), Zhao, et al (1990)从理论上证明了直方图估计密度函数的几乎处处收敛性。举例来说(茆诗松, 2001), 表1是上海市中心气象台发布的1884–1982年这99年来上海市年降水量数据(单位: mm)。样本数据中最小值为709.2, 最大值为1659.3。若我们设定最小分界点为620, 各组组距长度为 $h = 100$, 组数为 $k = 11$, 具体分组和各组样本频数、频率列于表2中。图1显示了上海市年降水量的直方图(利用SAS软件制作), 以及近似总体密度曲线, 从曲线的整体形状可以看出上海市年降水量分布大致服从正态分布。

*国家自然科学基金项目(70371018, 70572074)资助。

本文2007年3月26日收到。

《应用概率统计》版权所有

表1.1 上海市1884–1982年年降水量(单位mm)

1184.4	1113.4	1203.9	1170.7	975.4	1462.3	947.8	1416.0	709.2
1147.5	935.0	1016.3	1031.6	1105.7	849.9	1233.4	1008.6	1063.8
1004.9	1086.2	1022.5	1330.9	1430.4	1236.5	1008.1	1288.7	1115.8
1217.5	1320.7	1087.1	1203.4	1480.0	1269.9	1040.2	1318.4	1192.0
1016.0	1508.2	1159.6	1021.3	986.1	794.7	1318.3	1171.2	1161.7
791.2	1143.8	1602.0	951.4	1003.2	840.4	1061.4	958.0	1025.2
1265.0	1196.5	1120.7	1659.3	942.7	1123.3	910.2	1398.5	1208.6
1305.5	1242.3	1572.3	1416.9	1256.1	1285.9	984.8	1390.3	1062.2
1287.3	1477.0	1017.9	1217.7	1197.1	1143.0	1018.8	1243.7	909.3
1030.3	1124.4	811.4	820.9	1184.1	1107.5	991.4	901.7	1176.5
1113.5	1272.9	1200.3	1508.7	772.3	813.0	1392.3	1006.2	1108.8

表1.2 上海市年降水量频数、频率分布表(单位mm)

组号	区间	频数	频率
1	(620,720]	1	0.0101
2	(720,820]	5	0.0505
3	(820,920]	6	0.0606
4	(920,1020]	17	0.1717
5	(1020,1120]	18	0.1818
6	(1120,1220]	22	0.2222
7	(1220,1320]	14	0.1414
8	(1320,1420]	7	0.0707
9	(1420,1520]	6	0.0606
10	(1520,1620]	2	0.0202
11	(1620,1720]	1	0.0101

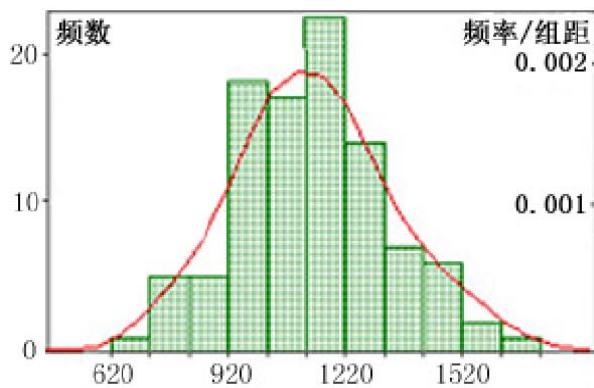


图1.1 上海市年降水量频数、频率直方图

由此, 关于直方图的制作, 我们可以概括为以下几个步骤:

- (1) 给定一组样本观测值 x_1, x_2, \dots, 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 a_0 , 满足 $a_0 \leq x_{(1)}$;
- (2) 估计组距(Bin width) h (本文主要讨论等组距情况下的直方图制作), 可得每组分界点(Bin edges) a_0, a_1, \dots, a_k , 其中, $a_{i+1} - a_i = h$, $i = 0, 1, \dots, k-1$, $x_{(n)} \leq a_k < x_{(n)} + h$;
- (3) 计算落在每组区间 $A_i = (a_i, a_{i+1}]$, $i = 0, 1, \dots, k-1$ 中的样本频数: $\gamma_1, \dots, \gamma_k$;
- (4) 以 h 为宽, $\gamma_1, \dots, \gamma_k$ 为高作矩形, 构建直方图;
- (5) 由直方图估计样本对应总体的密度分布:

$$\hat{f}_H(x) = \frac{\gamma_i}{nh}, \quad x \in (a_{i-1}, a_i], \quad i = 1, 2, \dots, k. \quad (1.1)$$

然而, 制作直方图的关键就是确定最小下界 a_0 (或其它某一分界点) 和组距 h . 关于组距的选择, 有许多方法并存在很大争议. 组距在很大程度上影响直方图的性质和总体分布特征, 不同的人采用不同的分组方式, 所得的结论会有所不同, 甚至采用不当的组距, 会使直方图“失真”. 之前, 许多文献对确定直方图组距有所讨论: 大多数学者(如, 范诗松, 2001)采用经验的方法, 认为当样本量 n 较大时, 分组数 k 取 10 到 20 之间; 当样本量 $n < 50$ 时, k 通常取 5 到 6 之间, 且每组区间中样本频数通常要求不少于 5 (两端可少一些). 谢衷洁(2004)在其著作中推荐使用了 Moore (1986) 公式: $k \approx C \cdot n^{2/5}$, $C = 1 \sim 3$. Montgomery (1996) 给出了直方图制作的三条建议: (1) 分组数一般可以近似等于样本量的平方根; (2) 各组组距相等; (3) 以比最小样本观测值稍小的值作为最小边界点. 其中, 最后一条也是确定最小下界的常用方法. 可以看出, 对直方图这两个参数的选择, 大多数学者都采用了各自不同的经验方法. Scott (1979) 和 Simonoff and Udina (1997) 分别指出, 当样本量比较大时, 组距的正确选择要比边界点的确定更为重要. 如果组距选择太大, 对应的分组数就较少, 制作的直方图会很平坦(Over-smooth), 则就不能充分显示样本信息; 反之, 如果组距选择太小, 对应分组数较多, 就会出现许多样本频数为零的区间, 制作的直方图就很粗糙(Over-rough), 同样也不能正确估计样本信息, 甚至会得出错误的结论. 因此, 正确估计组距, 就是要在这对矛盾之间选择一个平衡.

本文将以直方图组距计算为主线, 介绍直方图理论及最优直方图制作过程, 包括 Sturges 公式、Scott 公式、Cross-Validation 方法和 Histogram-Kernel Error 方法; 其后将结合实际例子讨论面向样本的最优直方图制作过程; 第四节介绍不等组距直方图和平均滑动直方图; 最后总结评价直方图制作好坏的基本标准.

§2. 最优直方图理论和制作方法

2.1 Sturges 公式

Sturges (1926) 在直方图制作方法上做了开创性的工作, 得到了分组数 k 关于样本量 n 的粗略关系式. 现在, 许多学者和一些统计软件还都是以 Sturges 公式为主要依据来确定直方图分组数. Sturges 的主要思想是用对称的二项分布 ($p = 0.5$) 来近似正态分布.

考虑理想化的直方图, 其分组数为 k . 假设每一样本观测值落在直方图第 i 个区间中的概率近似服从概率 $p = 0.5$ 的二项分布, 则第 i 组的样本频数平均为 $\gamma_i = C_{k-1}^i$, $i = 0, 1, \dots, k-1$, 其中 C_{k-1}^i 为组合数. 则当 k 很大时, 这样理想的直方图就可以近似为一个均值为 $(k-1)/2$, 方差为 $(k-1)/4$ 的正态分布. 由于总的样本量为:

$$n = C_{k-1}^0 + C_{k-1}^1 + \dots + C_{k-1}^{k-1} = 2^{k-1},$$

于是, 两边取对数即得分组数 k 关于样本量 n 的 Sturges 公式:

$$k = 1 + \log_2 n \approx 1 + 3.322 \log n. \quad (2.1)$$

由此可知, Sturges 公式是在总体分布对称且近似正态的假设下得到的, 仅考虑样本量的大小, 而忽略了样本分布的特征. 所以, Sturges 公式的运用具有很大的局限性, 制作的直方图通常过于平坦, 尤其是在样本量比较大时.

Doane (1976) 在考虑样本数据有偏的情况下, 对 Sturges 公式作了修正, 得修正表达式: $k = 1 + \log_2 n + K_e$, 其中 $K_e = \log_2(1 + \hat{\gamma}\sqrt{(n+1)(n+3)/[6(n-2)]})$, $\hat{\gamma}$ 为偏度估计值. 一般情况下, 对于非正态数据, 即有偏或者峰度较大的样本数据, 通常要求增加额外的分组, 即分组数比 Sturges 公式求得的值要大. 实践表明, Sturges 公式并不能作为计算直方图分组数的主要依据(详见第三节).

2.2 Scott 公式

为了弥补 Sturges 公式的不足, Scott (1979) 以最小平均误差平方和(Mean Integrated Square Error, 简称 MISE) 为准则, 系统地讨论了直方图理论和直方图组距确定的渐近最优公式.

定义平均误差平方和(MISE) 最小准则:

$$h^* = \min_h \{MISE\} = \min_h \left\{ E \left\{ \int [\hat{f}_H(x) - f(x)]^2 dx \right\} \right\}. \quad (2.2)$$

其中, \hat{f}_H 为直方图估计, f 为连续可微的总体密度函数. 在某一点处的直方图平均误差平方(Mean Square Error, 简称 MSE) 可以拆分成直方图估计的偏度平方与方差之和, 即

$$MSE\{\hat{f}_H(x)\} = E[\hat{f}_H(x) - f(x)]^2 = Var[\hat{f}_H(x)] + Bias^2[\hat{f}_H(x)].$$

其中 $Bias[\hat{f}_H(x)] = E[\hat{f}_H(x)] - f(x)$.

估计 MISE 可得,

$$MISE = \frac{1}{nh} + \frac{1}{12}h^2 \int f'(x)^2 dx + O(1/n + h^3). \quad (2.3)$$

所以, 其渐近平均误差平方和(Asymptotic MISE) 为:

$$AMISE = \frac{1}{nh} + \frac{1}{12}h^2 \int f'(x)^2 dx.$$

对AMISE关于 h 求导数, 最小化AMISE可得渐近最优组距和最小渐近平均误差平方和,

$$h^* = \left(6 / \left[n \cdot \int f'(x)^2 dx \right] \right)^{1/3}, \quad (2.4)$$

$$\text{AMISE}^* = \left(\frac{3}{4} \right)^{2/3} \left(\int f'(x)^2 dx \right)^{1/3} \cdot n^{-2/3}. \quad (2.5)$$

对于正态分布 $N(\mu, \sigma^2)$, 上式可简化为:

$$h^* = \left(\frac{24\sqrt{\pi}\sigma^3}{n} \right)^{1/3} \approx 3.5 \cdot \sigma \cdot n^{-1/3}. \quad (2.6)$$

从(2.4)、(2.6)式可以看出, 渐近最优组距依赖于总体密度分布. 如果总体密度未知, 则不能确定最优组距的值. Scott (1979)建议, 对于总体分布未知时的正态数据(σ 未知), 最优组距的估计值可以为:

$$\hat{h} = 3.5 \cdot \hat{\sigma} \cdot n^{-1/3}. \quad (2.7)$$

其中, $\hat{\sigma}$ 为样本标准差. Freedman and Diaconis (1981)又提出利用四分位数差值代替未知标准差 σ , 得到更稳健的表达式,

$$\hat{h} = 2 \cdot IQ \cdot n^{-1/3}. \quad (2.8)$$

其中, IQ 表示样本的四分之三分位值与四分之一分位值的差额.

对于非正态分布, 由(2.4)式考察非正态分布最优组距与正态分布最优组距的比值,

$$\frac{h_g^*}{h_N^*} = \left(\left[\int \phi'(x)^2 dx \right] / \left[\int g'(x)^2 dx \right] \right)^{1/3}.$$

其中, $\phi(x)$ 为正态密度, $g(x)$ 为非正态密度. Scott (1979)以对数正态和 t 分布作为参考对象, 分别代表有偏分布和峰度分布, 模拟了比值 h_g^*/h_N^* 关于偏度系数和峰度系数的图像, 再以(2.7)式作修正. 由此可知, Scott (1979)在估计面向样本的直方图组距时, 仅采用了基于对数正态和 t 分布的间接方法, 具有很大的不确定性.

此后, Wand (1997)对Scott (1979)的结果作了延伸, 得到了面向样本的组距选择方法, 且其构造的组距以 $n^{1/2}$ 收敛于(2.4)式. 具体表述如下:

组距估计表达式:

$$\tilde{h} = \left(- \frac{6}{\tilde{\Psi}_2(g_{21})n} \right)^{1/3}.$$

式中各参数表示为:

$$g_{21} = [2 / \{(2\pi)^{1/2} \tilde{\Psi}_4(g_{22})^{1/5} n\}]^{1/5} 2^{1/2} \hat{\sigma},$$

$$g_{22} = [2 / (5n)]^{1/7} 2^{1/2} \hat{\sigma},$$

$$\hat{\sigma} = \min\{s, IQ/1.349\}, \quad s \text{是样本标准差},$$

$$\tilde{\Psi}_r(g) = n^{-2} \sum_{j=1}^M \left(\sum_{i=1}^M c_i \kappa_{i-j}^{(r)} \right) c_j, \quad \text{取} M = 400,$$

$$c_j = \sum_{i=1}^n (1 - |\delta^{-1} X_i - j|)_+, \quad |j| = 0, \dots, M,$$

$$\delta = (G_M - G_1) / (M - 1), \quad \text{其中} G_1 = x_{(1)}, G_M = x_{(n)},$$

《应用概率统计》版权所有

$$\kappa_j^{(r)} = g^{-r-1} L^{(r)}(\delta j/g), \quad L^{(r)} \text{为正态核函数的 } r \text{ 阶导数.}$$

虽然, 模拟显示Wand (1997)方法用在小样本和中等样本时要优于Scott (1979)方法, 但其理论的复杂性和计算的繁琐很难被大众所接受.

从另一个角度分析, Terrell (1990)从(2.4)式出发, 在总体分布未知的条件下, 通过求解 $\min_f \left\{ \int f'(x)^2 dx \right\}$ 来估计 h^* 的上界, 确定最大组距. 其证明了密度函数为:

$$f(x) = \frac{15}{16\sqrt{7}\sigma} \left(1 - \frac{x^2}{7\sigma^2}\right)^2 I_{[-\sqrt{7}\sigma, \sqrt{7}\sigma]}(x).$$

所以, 有

$$h_{OS}^* \leq \left(\frac{686\sigma^3}{5\sqrt{7}n} \right)^{1/3} \approx 3.729 \cdot \sigma \cdot n^{-1/3}. \quad (2.9)$$

如用四分位数差值估计, 可得

$$h_{OS}^* \leq 2.603 \cdot IQ \cdot n^{-1/3}. \quad (2.10)$$

此前, Terrell and Scott (1985)还证明了最小分组数满足:

$$k \geq \sqrt[3]{2n} = k_{\min}. \quad (2.11)$$

2.3 Cross-Validation方法

制作直方图的主要目的是用样本估计总体分布, 但通常我们并不清楚总体的分布特征, 因此(2.4)式确定的最优组距公式是可述而不可得的. Rudemo (1982)利用Cross-Validation的思想在最小化误差平方和(Integrated Square Error)准则下构造了面向样本的最优组距估计方法. Cross-Validation方法的基本思想是: 对现有样本量为 n 的样本, 剔除其中一个样本观测值, 然后用剩余的样本单元估计参数. 回头再用估计的参数来预测被删除的样本观测值. 因为被删除的样本单元已知, 所以可以计算预测值和实际值的误差. 重复以上步骤 n 次, 对每一个样本观测值都作一次估计, 再将所有结果取平均. 可以证明, 由Cross-Validation方法计算的面向样本的组距估计值 \hat{h}_{CV} 是 h^* 的一个无偏估计.

假设样本观测值为 x_1, x_2, \dots, x_n , 定义最小误差平方和为:

$$h^* = \min_h \{ISE\} = \min_h \left\{ \int [\hat{f}_H(x) - f(x)]^2 dx \right\}. \quad (2.12)$$

展开上式括号内的积分, 得:

$$ISE = \int \hat{f}_H^2(x) dx - 2 \int \hat{f}_H(x) f(x) dx + \int f^2(x) dx. \quad (2.13)$$

第一项, $\int \hat{f}_H^2(x) dx = \sum_i \gamma_i^2 / (n^2 h)$; 第二项, $\int \hat{f}_H(x) f(x) dx = E[\hat{f}_H(x)]$; 第三项独立于样本观测值. 所以, 最小化ISE, 只需计算前两项的最小值. 为了估计 $E[\hat{f}_H(x)]$, Rudemo

(1982)在此使用了Cross-Validation方法: 对样本量为 n 的样本, 每次剔除一个样本单元, 用剩下的 $n - 1$ 个样本估计直方图 $\hat{f}_{-i}(x)$ ($-i$ 表示剔除了第 i 个样本). 重复以上步骤 n 次, 再将得到的所有估计取其平均. Rudemo (1982)同时证明了, 其平均值是 $E[\hat{f}_H(x)]$ 的无偏估计. 定义 $CV = \int \hat{f}_H^2(x)dx - 2 \int \hat{f}_H(x)f(x)dx$, 所以, 最小化(2.13)式等价于优化

$$\hat{h}_{CV} = \min_h \{CV(h)\} = \min_h \left\{ \sum_i \frac{\gamma_i^2}{n^2 h} - \frac{2}{n} \sum_{j=1}^n \hat{f}_{-j}(x_j) \right\}. \quad (2.14)$$

化简 $CV(h)$ 有,

$$CV(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_i \gamma_i^2. \quad (2.15)$$

此后, Scott and Terrell (1987)详细讨论了直方图和核密度估计的Cross-Validation方法, 并证明了直方图 $CV(h)$ 统计量的期望和方差,

$$E[CV(h)] = AMISE - \int f^2(x)dx + O(n^{-1} + h^3), \quad (2.16)$$

$$\text{Var}[CV(h)] = \frac{4}{n} \left[\int f^3(x)dx - \left(\int f^2(x)dx \right)^2 \right] + O(h^2 n^{-1} + h^{-1} n^{-2}). \quad (2.17)$$

评述用 $CV(h)$ 估计直方图的最优组距的不足: 虽然 \hat{h}_{CV} 是(2.4)式 h^* 的一个无偏估计, 但其收敛速度非常缓慢, 具体为: $\sigma_{h_{CV}}/h_{CV} = O(n^{-1/6})$; 而且, 在某些样本(如有多重观测值)条件下 $CV(h)$ 是发散的. 另外, 我们所求的 \hat{h}_{CV} 通常并非是一个全局最优解, 而是一个局部最优; 详见Scott (1992).

2.4 Histogram-Kernel Error方法

最小化平均误差平方和可求得渐近最优组距(2.4), 但当样本对应总体密度未知时, 其优化问题(2.2)不可求解, Wang and Zhang (2008)使用了替换方法. 考虑将 $f(x)$ 用分组核密度估计函数 $\hat{f}_{BK}(x)$ 代替, 定义面向样本的Histogram-Kernel Error平方和:

$$\text{ISE}_{H-BK} = \int [\hat{f}_H(x) - \hat{f}_{BK}(x)]^2 dx. \quad (2.18)$$

其中,

$$\hat{f}_{BK}(x) = \sum_i \frac{\gamma_i}{nh} K\left(\frac{x - c_i}{h}\right),$$

$K(\cdot)$ 是某一选定的核密度, c_i 是分组后各组的中点, 即 $c_i = (a_i + a_{i+1})/2$. 为了计算简单, 假设 $K(\cdot)$ 是定义在 $[-1, 1]$ 上的对称核, 且函数 $f(x)$ 连续二阶可微. Wang and Zhang (2008)证明了(2.18)等价于

$$\text{ISE}_{H-BK} = c_1 \sum_i \frac{\gamma_i^2}{n^2 h} - c_2 \sum_i \frac{\gamma_i \gamma_{i+1}}{n^2 h}. \quad (2.19)$$

其中, 常数

$$\begin{aligned} c_1 &= \int_{-1}^1 K^2(u)du - 2 \int_{-1/2}^{1/2} K(u)du + 1, \\ c_2 &= 2 \left[\int_0^{1/2} K(u-1)du + \int_{1/2}^1 K(u)du - \int_0^1 K(u) \cdot K(u-1)du \right]. \end{aligned}$$

可以看出, (2.19)只是关于样本观测值的函数. 对于给定的核密度函数 $K(u)$, 最小化(2.19)式可得 $\text{ISE}_{\text{H-BK}}$ 的最优组距 $\tilde{h}_{\text{H-BK}}$. 考虑其渐近性质, 可以证明(2.18)式的期望和方差为:

$$\text{MISE}_{\text{H-BK}} = \frac{c_1}{nh} + \frac{7c_2 - c_1}{12} h^2 \int f'(x)^2 dx + (c_1 - c_2) \int f^2(x)dx + O(n^{-1} + h^3), \quad (2.20)$$

$$\text{Var}(\text{ISE}_{\text{H-BK}}) = \frac{4}{n} (c_1 - c_2)^2 \left[\int f^3(x)dx - \left(\int f^2(x)dx \right)^2 \right] + O(h^2 n^{-1} + h^{-1} n^{-2}). \quad (2.21)$$

对 $\text{MISE}_{\text{H-BK}}$ 最小化, 得 Histogram-Kernel Error 均方和的最优组距和渐近最小平均误差平方和,

$$h_{\text{H-BK}}^* = \left[\frac{6c_1}{7c_2 - c_1} \right]^{1/3} \cdot \left[\int f'(x)^2 dx \right]^{-1/3} \cdot n^{-1/3}, \quad (2.22)$$

$$\text{AMISE}_{\text{H-BK}}^* = \left(\frac{9}{16} \right)^{1/3} c_1^{2/3} (7c_2 - c_1)^{1/3} \left[\int f'(x)^2 dx \right]^{1/3} n^{-2/3} + (c_1 - c_2) \int f^2(x)dx. \quad (2.23)$$

比较(2.22)和(2.4), 有 $h_{\text{H-BK}}^*$ 与 h^* 仅相比一个常数, 即

$$\text{Ratio}^* = \frac{h_{\text{H-BK}}^*}{h_{\text{H}}} = \left(\frac{c_1}{7c_2 - c_1} \right)^{1/3}. \quad (2.24)$$

所以, 利用(2.19)求出的最优组距 $\tilde{h}_{\text{H-BK}}$, 除以 $[c_1/(7c_2 - c_1)]^{1/3}$ 即可作为面向样本的直方图最优组距估计, 即 $\tilde{h}_H = \tilde{h}_{\text{H-BK}}/[c_1/(7c_2 - c_1)]^{1/3}$. 可以进一步证明, 该方法是可靠的(Reliable), 即

$$\mathbb{P} \left\{ \lim_{n \rightarrow \infty} \left[\frac{\tilde{h}_{\text{H-BK}}}{h_{\text{H}}} = \left(\frac{c_1}{7c_2 - c_1} \right)^{1/3} \right] \right\} = 1.$$

以 Epanechnikov 核 $K(u) = 3(1 - u^2)_+ / 4$ 为例, 可以计算 $c_1 = 9/40$, $c_2 = 17/80$, $[c_1/(7c_2 - c_1)]^{1/3} = 0.5628$, 则 $\hat{h}_H = \tilde{h}_{\text{H-BK}} / 0.5628$. 再比较(2.21)和(2.17), 可以发现, Histogram-Kernel Error 方法比 Cross-Validation 方法具有更小的误差方差, 即从渐近误差角度分析, Histogram-Kernel Error 方法比 Cross-Validation 方法更稳定.

另一方面, Wang and Zhang 利用遗传算法设计了同时求解最优边界点和最优组距的算法程序, 突出最小化 $\text{ISE}_{\text{H-BK}}$ 准则下的直方图制作通常具有全局最优性.

2.5 其它方法

制作面向样本的最优直方图, 除了上述介绍的 Sturges 公式、 Scott 公式、 Cross-Validation 方法和 Histogram-Kernel Error 方法以外, 还有其他许多统计学家也发挥了各

《应用概率统计》版权所有

自智慧, 构造了各种不同的制作方法. 如: Taylor (1987)利用Akaike信息准则求解最优组距; Daly (1988)设计了Walsh函数计算最优分组数算法; He and Meeden (1997)构造损失函数(Loss Function)求解最优分组数; Beer and Swanepoel (1999)以频数多边形(Frequency Polygon)代替未知密度函数, 利用Bootstrap方法求解最优分组数; Birgé and Rozenholc (2002)利用惩罚最大似然估计(Penalized Maximum Likelihood Estimation) 求解最优分组数; Knuth (2006)构造优化算法确定最优分组数, 使似然函数(Likelihood Function)与模型先验概率(Prior Probability)之间选择一个平衡; 其他还包括Kim and Ryzin (1975), Stone (1985), Castellan (2000), Shim (2004)等学者的工作.

§3. 应用举例

本节将进一步阐述直方图制作的具体方法, 比较各种方法的优良. 以Scott教授提供的1983年随机调查的6,973位德国公民收入数据为样本观测值(取自然对数后的值), 通过上述几种方法, 制作了面向样本的直方图估计; 如图3.1所示.

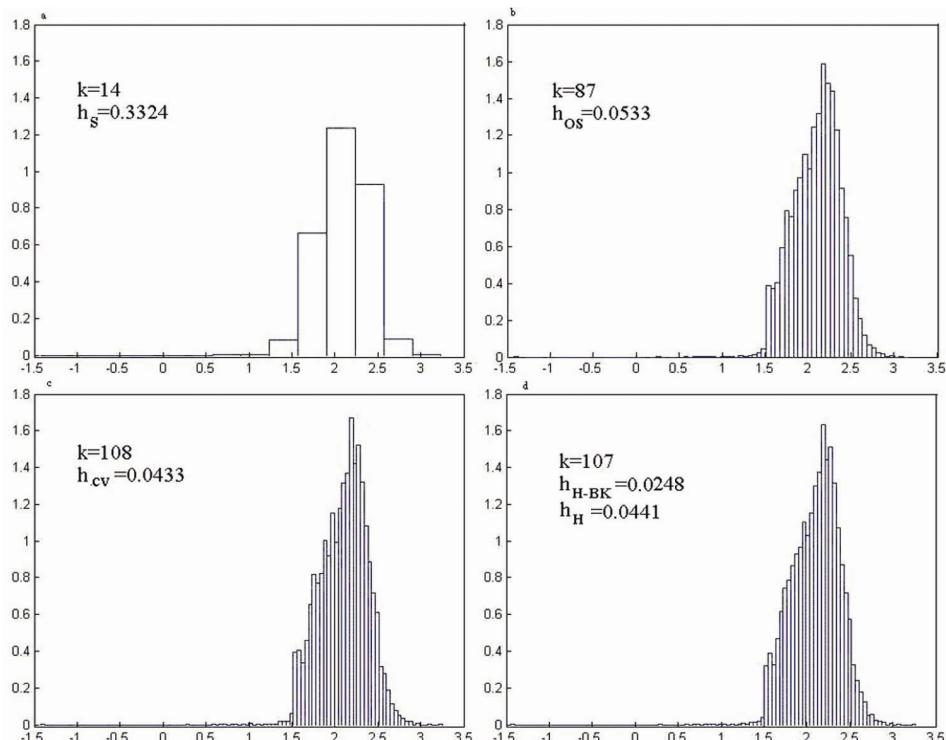


图3.1 1983年6,973位德国国民收入调查数据直方图分析

图3.1(a)是用Sturges公式制作的直方图, 其分组数为 $k_1 = 14$. 显然, 所作的直方图过于平坦, 其很大程度损失了原样本信息. 图3.1(b)是由(2.10)式估计的最大组距 $h_{OS} = 0.0533$, 对应的分组数为 $k_2 = 87$, 所作的直方图也是平坦的. 图3.1(c)是用Cross-Validation方法最小化(2.15)式得到的直方图, 其中 $\tilde{h}_{CV} = 0.0433$, $k_3 = 108$. 图3.1(d)是用Histogram-

《应用概率统计》版权所有

Kernel Error方法得到的直方图, 利用Epanechnikov核求解最小误差平方和(2.19), 得组距估计 $\tilde{h}_{H-BK} = 0.0248$, 其中有一分界点为0.3592. 由 $\hat{h}_H = \tilde{h}_{H-BK}/0.5628$, 可得 $\hat{h}_H = 0.0441$, $k_4 = 107$. 从表面看, 图3.1(c)和(d)在估计总体特征时没有显著差异, 但从(2.21)和(2.17)可知, 图3.1(d)比图3.1(c)更可取.

§4. 不等组距直方图和平均滑动直方图介绍

4.1 不等组距直方图(Adaptive Histogram)

用等组距直方图估计总体分布, 在众数附近和尾部区域具有相同的分组区间. 然而, 在众数附近由于集中的样本信息较多, 利用等组距算法制作的直方图通常在此区间估计趋于平坦; 相反, 在尾部区域由于集中的样本信息较少, 利用等组距算法制作的直方图通常在此区域估计更为粗糙. 不等组距直方图是指根据样本信息, 自动调整组距大小, 以适应不同区域的样本特征.

对给定的样本观测值 x_1, x_2, \dots, x_n , 考虑不等组距直方图理论及制作. 设第*i*个分组区间为 $A_i = (a_i, a_{i+1}]$, $a_{i+1} - a_i = h_i$. 确定 h_i^* , 以最小化 $\text{ISE}_{AH} = \int [\hat{f}_{AH}(x) - f(x)]^2 dx$. 现先估计整体方差(Integrated Error)和整体偏度平方(Integrated Squared Biased), 有

$$\text{IV}_{AH} = \frac{1}{n} \sum_i \frac{p_i(1-p_i)}{h_i}, \quad (4.1)$$

$$\text{ISB}_{AH} = \int f^2 dx - \sum_i \frac{p_i^2}{h_i}. \quad (4.2)$$

所以组距的选择由以下优化函数决定,

$$\begin{aligned} h_i^* &= \min_{h_i} \{\text{ISE}_{AH}\} = \min_{h_i} \{\text{IV}_{AH} + \text{ISB}_{AH}\} \\ &= \min_{h_i} \left\{ \frac{1}{n} \sum_i \frac{p_i(1-p_i)}{h_i} - \sum_i \frac{p_i^2}{h_i} + \int f^2 dx \right\}. \end{aligned} \quad (4.3)$$

若考虑不等组距渐近平均误差平方和, $\text{MISE}_{AH} = E \left\{ \int [\hat{f}_{AH}(x) - f(x)]^2 dx \right\}$. 则可估计在单点处的平均误差平方,

$$\text{MSE}_{AH} = \frac{f(x)}{nh} + \frac{1}{12} h^2 f'(x)^2 + O(n^{-1} + h^3). \quad (4.4)$$

最小化 MSE_{AH} , 有

$$h_{AH}^*(x) = \left[\frac{6f(x)}{nf'(x)^3} \right]^{1/3}, \quad (4.5)$$

$$\text{MSE}_{AH}^* = \left[\frac{3f(x)f'(x)}{4n} \right]^{2/3} + O(n^{-1} + h^3). \quad (4.6)$$

对(4.6)式取积分, 可得不等组距最小渐近平均误差平方和,

$$\text{AMISE}_{\text{AH}}^* = \left(\frac{3}{4}\right)^{2/3} \left(\int [f(x)f'(x)]^{2/3} dx \right) \cdot n^{-2/3}. \quad (4.7)$$

根据Hölder不等式, 有 $\int [f(x)f'(x)]^{2/3} dx \leq \left[\int f'(x)^2 dx \right]^{1/3}$. 所以(4.7)式比(2.5)式有更小的渐近平均误差平方和. Kogure (1987)证明了直方图均方误差平方和的下确界为

$$\inf \{\text{MISE}(\hat{f}_H)\} = \left(\frac{3}{4}\right)^{2/3} \left(\int [f(x)f'(x)]^{2/3} dx \right) \cdot n^{-2/3} + o(n^{-2/3}). \quad (4.8)$$

同等组距分组估计结果一样. 虽然(4.5)式可确定不等组距直方图制作的最优组距, 但该表达式很难直接应用于面向样本的不等组距直方图制作中. 为此, Kogure (1987)构造了局部等组分割法(Partitions of Locally Equisized Cells), Kanazawa (1992)还利用最小化Hellinger距离准则设计了面向样本的不等组距直方图制作算法.

4.2 平均滑动直方图(Averaged Shifted Histogram, 简称ASH)

虽然直方图组距估计要比边界点的选择更为重要, 但制作面向样本的直方图, 特别是样本量不是很大时, 边界点的选择也被看作是不可忽略的变量. 因为不同的边界点有时会影响直方图的特征和众数(Mode)个数. 为了减小边界点带来的误差, 其方法之一是由Scott (1985, 1992)提出的平均滑动直方图方法.

考察 m 个直方图, $\hat{f}_0, \hat{f}_1, \dots, \hat{f}_{m-1}$, 每一直方图组距都相同, 且为 h , 边界点选择分别包括 $a = 0, h/m, 2h/m, \dots, (m-1)h/m$ (假设在样本观测值范围之内). 平均滑动直方图定义为:

$$\hat{f}_{\text{ASH}}(x) = \frac{1}{m} \sum_{i=0}^{m-1} \hat{f}_i(x). \quad (4.9)$$

定义小区间 $I_k = [k\delta, (k+1)\delta]$, 其中 $\delta = h/m$. 假设区间 I_k 中样本数为 γ_k , 则在各个不同边界点的直方图估计为

$$\hat{f}_i(x) = \frac{1}{nh} \sum_{j=0}^{m-1} \gamma_{j+i+[(k-i)/m]/m}, \quad x \in I_k, \quad i = 0, 1, \dots, m-1. \quad (4.10)$$

其中, $[y]$ 是不大于 y 的最小整数. 将(4.10)式代入(4.9)式, 化简得,

$$\hat{f}_{\text{ASH}}(x) = \frac{1}{mn} \sum_{i=1-m}^{m-1} (m - |i|) \gamma_{k+i}, \quad x \in I_k. \quad (4.11)$$

考虑平均滑动直方图的性质. 可以证明, 当 $m \rightarrow \infty$ 时, (4.11)式收敛于核函数为 $K(u) = (1 - |u|)I_{[-1,1]}(u)$ 的核密度估计 $\hat{f}_K(x) = [1/(nh)] \sum_{i=1}^n K[(x - x_i)/h]$. 考察平均滑动直方图的平均误差平方和: $\text{MISE} = \mathbb{E} \left\{ \int [\hat{f}_{\text{ASH}}(x) - f(x)]^2 dx \right\}$, 分别计算 $\text{Var}[\hat{f}_{\text{ASH}}(x)]$,

《应用概率统计》版权所有

$\text{Bias}^2[\hat{f}_{\text{ASH}}(x)]$, 并代入, 可得

$$\begin{aligned}\text{MISE} &= \frac{2}{3nh} \left(1 + \frac{1}{2m^2}\right) - \frac{1}{n} \int f^2 + \frac{h^2}{12m^2} \int f'^2 \\ &\quad + \frac{1}{144} h^4 \left(1 - \frac{2}{m^2} + \frac{3}{5m^4}\right) \int f''^2 + O(hn^{-1} + h^5).\end{aligned}\quad (4.12)$$

可以验证, 当 $m = 1$ 时, (4.12) 式等于(2.3)式; 当 $m \rightarrow \infty$ 时, (4.12) 式收敛于三角核密度估计. 所以, 平均滑动直方图具有一般直方图计算简便的特点, 同时也具有核密度函数精确估计的优点.

§5. 结语

直方图作为一种广为人知的密度估计和数据分析工具, 在质量管理、医疗统计、经济分析等领域几乎无处不在、无处不用. 制作直方图的关键因素是正确选择边界点和估计最优组距. 本文只是用有限的篇幅简单介绍了直方图理论和最优直方图制作的最新研究成果. 需要指出的是, 制作面向样本的直方图, 首先要确定相应的最优准则. 最为常见的优化准则是最小化误差平方和(ISE), 即 L_2 准则. 除此之外, 还包括 L_1 准则(Devroye and Györfi, 1985)、 L_∞ 准则(Kim and Ryzin, 1975)、Hellinger 距离准则(Baezon, Birgé and Massart, 1999)、Kullback-Leibler 距离准则(Rodriguez and Ryzin, 1985)、Akaike 准则(Atilgan, 1990; Kanazawa, 1993)等. 在不同的准则下, 所制作的直方图不一定相同, Birgé and Rozenholc (2002) 比较了不同准则下直方图估计的优劣.

对于多维直方图的理论和制作, 可以作为一维情况下的推广. 以一般多维矩形分割为例, 在 L_2 准则下, Scott (1992) 证明了 d 维直方图的渐近最优组距,

$$h_k^* = \left[\int_{R^d} \left(\frac{\partial f(\mathbf{x})}{\partial x_k} \right)^2 d\mathbf{x} \right]^{-1/2} \cdot \left\{ 6 \prod_{i=1}^d \left[\int_{R^d} \left(\frac{\partial f(\mathbf{x})}{\partial x_i} \right)^2 d\mathbf{x} \right]^{1/2} \right\}^{1/(2+d)} \cdot n^{-1/(2+d)},$$

其中, $\mathbf{x} = (x_1, x_2, \dots, x_d)$.

作者认为, 评价一个直方图的好坏, 其主要标准包括: 1) 判断所制作的直方图是否充分利用了样本信息; 2) 是否能够充分反映出样本对应总体分布的特征. 现今关于直方图的研究, 也正是向这两个标准去努力, 以设计更加稳定、更加精良的算法.

参 考 文 献

- [1] Atilgan, T., On derivation and application of AIC as a data-based criterion for histogram, *Commun. Statist.-Theory Meth.*, **19**(1990), 885–903.
- [2] Beer, C.F. and Swanepoel, J.W.H., Simple and effective number-of-bins circumference selectors for a histogram, *Statistics and Computing*, **9**(1999), 27–35.
- [3] Birgé, L. and Rozenholc, Y., *How Many Bins Should Be Put in a Regular Histogram*, Prépublication 721, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI et VII, France, 2002.

- [4] Barron, A., Birgé, L. and Massart, P., Risk bounds for model selection via penalization, *Probability Theory and Related Fields*, **113**(1999), 301–415.
- [5] Castellan, G., Sélection d'histogrammes ou de Modèles exponentiels de polynomes par morceaux à l'aide d'un critère de type Akaike, Thèse, Mathématiques, Université de Paris-Sud, 2000.
- [6] Chen, X.R. and Zhao, L.C., Almost sure L[1]-norm convergence for data-based histogram density estimates, *Journal of multivariate analysis*, **21**(1987), 179–188.
- [7] Daly, J.E., The construction of optimal histograms, *Commun. Statist.-Theory Meth.*, **17**(1988), 2921–2931.
- [8] Devroye, L. and Györfi, L., *Nonparametric Density Estimation: The L[1] View*, Wiley, New York, 1985.
- [9] Doane, D.P., Aesthetic frequency classifications, *Amer. Statist.*, **30**(1976), 181–183.
- [10] Freedman, D. and Diaconis, P., On the histogram as a density estimation: L[2]-theory, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**(1981), 453–476.
- [11] He, K. and Meeden, G., Selecting the number of bins in a histogram: A decision theoretical approach, *Journal of Statistical Planning and Inference*, **61**(1997), 49–59.
- [12] Kanazawa, Y., An optimal variable cell histogram based on the sample spacings, *The Annals of Statistics*, **20**(1992), 291–304.
- [13] Kanazawa, Y., Hellinger distance and Akaike's information criterion for the histogram, *Statistics and Probability Letters*, **17**(1993), 293–298.
- [14] Kim, B.K. and Ryzin, J.V., Uniform consistency of a histogram density estimator and modal estimation, *Communications in Statistics*, **4**(1975), 303–315.
- [15] Knuth, K.H., *Optimal Data-Based Binning for Histograms*, eprint arXiv: physics/0605197, 2006.
- [16] Kogure, A., Asymptotically optimal cells for a histogram, *The Annals of Statistics*, **15**(1987), 1023–1030.
- [17] 范诗松主编, *统计手册*, 科学出版社, 2001.
- [18] Montgomery, D.C., *Introduction to Statistical Quality Control*, John Wiley & Sons, Inc., 1996.
- [19] Rodriguez, C.C. and Ryzin, J.V., Maximum entropy histogram, *Statistics and Probability Letters*, **3**(1985), 117–120.
- [20] Rudemo, M., Empirical choice of histogram and kernel density estimators, *Scandinavian Journal of Statistics*, **9**(1982), 65–78.
- [21] Scott, D.W., On optimal and data-based histograms, *Biometrika*, **66**(1979), 605–610.
- [22] Scott, D.W., Averaged shifted histograms: effective nonparametric density estimators in several dimensions, *The Annals of Statistics*, **13**(1985), 1024–1040.
- [23] Scott, D.W., *Multivariate Density Estimation – Theory, Practice and Visualization*, John Wiley & Sons, New York, 1992.
- [24] Scott, D.W. and Terrell, G.R., Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association*, **82**(1987), 1131–1146.
- [25] Simonoff, J.S. and Udina, F., Measuring the stability of histogram appearance when the anchor position is changed, *Computational Statistics & Data Analysis*, **23**(1997), 335–353.
- [26] Shim, K., Recent advances in histogram construction algorithms, *Lecture Notes in Computer Science*, **3129**(2004), 23.

- [27] Stone, C.J., An asymptotically optimal histogram selection rule, in: *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. II, 513–520, 1985.
- [28] Sturges, H.A., The choice of a class interval, *J. Amer. Statist. Assoc.*, **21**(1926), 65–66.
- [29] Taylor, C.C., Akaike's information criterion and the histogram, *Biometrika*, **74**(1987), 636–639.
- [30] Terrell, G.R. and Scott, D.W., Over-smoothed nonparametric density estimates, *Journal of the American Statistical Association*, **80**(1985), 209–214.
- [31] Terrell, G.R., The maximal smoothing principle in density estimation, *Journal of the American Statistical Association*, **85**(1990), 470–477.
- [32] 谢衷洁编著, 普通统计学, 北京大学出版社, 2004.
- [33] Wand, M.P., Data-based choice of histogram bin width, *The American Statistician*, **51**(1997), 59–64.
- [34] Wang, X.X. and Zhang, J.F., Histogram-kernel error and its application for bin width selection in histograms, *Acta Mathematica Applicatae Sinica* (in press), 2008.
- [35] Zhao, L.C., Krishnaiah, P.R. and Chen X.R., Almost sure $L[r]$ -norm convergence for data-based histogram density estimates, *Theory of Probability and Its Applications*, **35**(1990), 396–403.

Histogram Theories and Optimal Histogram Construction Algorithms

ZHANG JIANFANG

(College of Department, Graduate University of Chinese Academy of Sciences, Beijing, 100190)

WANG XIUXIANG

(Hangzhou SME Financial Department of China Minsheng Banking Corp., LTD., Hangzhou, 310009)

Histogram is the most widely used density estimator and data analysis tool. It is completely determined by two parameters: the bin width and one of the bin edges. However, many professional statisticians have no really definitive answers and simply give some intuitive advises when face to choose these two parameters. Even most statistical packages use the rules of thumbs for selecting the number of bins as a default. In this paper, we will present the histogram theories and optimal histogram construction algorithms that have been recently proposed. The methods of how to construct the data-based histograms are the emphasis of this paper.

Keywords: Histogram, Sturges' rule, Scott's rule, Cross-Validation, Histogram-Kernel Error, integrated square error.

AMS Subject Classification: 62G05, 62E20.