浅 议 众 数

徐明生
(厦门大学 计划统计系,福建 厦门 361005)

摘要:在统计学文献中,众数扮演很小的角色。然而,在许多应用中,众数被证明是个有用的统计参数。本文介绍了众数在实际中的一些应用,并提出了三种概率密度 f(x)的众数估计方法。

关键词:众数;概率密度;点估计
中图分类号:O212 文献标识码:A 文章编号:1005-5762(2001)02-0020-02

一、引言

众数通常被定义为"频数分布中,出现次数最多的标志值。"翻开统计学教材和有关的期刊杂志,有关众数的介绍和研究很少。与算术平均数相比,大家似乎对它不大感兴趣。本文中,笔者介绍了一些众数的实际应用例子,并特别对概率密度 f(x) 的众数估计提出了三种方法,希望本文能起到抛砖引玉作用,引起读者对众数的兴趣,并对它进行更深入的研究。

二、众数的再解释

众数的重要性体现在它与算术平均数相比,适用的数据类型更广,从下面几点就可看出:

1. 假定 f(x) 是随机变量 x 的概率密度函数,如果对于一些 x=M,若有 f(M)≥f(x),等号只有 x=M 时才成立,则 f(x)的众数定义为 M。

很显然,不是所有的概率密度都会有一个且只有一个众数,当总体分布没有明显的集中趋势,而趋于均匀分布时,则没有众数。如果变量数列具有两个标志值出现次数最多,称为双众数,双众数的出现可能是由于总体单位不具有同质性。

2. 用 X 表示未分组连续变量值的一个可数集合(如样本),如果集合中的有两个或两个以上值相等,则出现次数最多的标志值即为众数。

3. 假定 X_g(g=1,2,…,k)表示已分组值的一个可数集合,N_g 是第 g 组中标志值的频数。如果对于一些 g=m,若有 $N_m \geq N_g$,等号只在 g=m 时才成立,则第 m 组称为众数组,从中找到一个合适的值就可当作众数。

4. 一个可数集的单位被分成 L 类(严格说,这其实也是分组,为了强调,在此单独列出),这种分类标准可按单位的公制特征,也可按单位的非公制特征。例如,工业产品可分为"合格品"与"不合格品",而我们知道划分合格的标准是不同的。第 C 类的单位的频数为 P_c(c=1,2,…,L),如果对于一些 c=m,若有 $P_m \geq P_c$,同样等号只在 c=m 时才成立,则第 m 类称为众数类。

三、众数的应用

在应用统计学中,众数有许多不同的应用,主要有如下三大方面:

1. 在多数情况下,众数仅作为描述位置的测度参数,主要是常见的位置平均数

作为这种测度参数,众数有个优点是它不受可能出现的极端值的影响。这个特性也是众数常被用作处理高度偏斜分布的位置测度的部分原因。在社会经济统计学中可找到这类例子,特别是在处理收入分配的统计中常用到众数。

与此相关,众数有时还被用来描述分布偏斜的程度,我们知道在适度倾斜时,下述公式:

算术平均数-众数=3(算术平均数-中位数)

近似成立。

在这方面,英国的统计学家卡尔·皮尔逊(Karl Pearson)还提出了用参数$\frac{\text{算术平均数}-\text{众数}}{\text{标准差}}$来衡量分布的偏斜程度。

当然,作为位置测度工具,众数也有一些明显的缺陷。如它与算术平均数相比,没有精确的计算公式,这也是造成统计人员对它不大感兴趣的一个原因。

2. 众数的应用还常出现在不确定情况下理性决策中

例如,成衣、鞋袜、帽子、自行车等商品的生产和销售,厂家或商店为了了解消费者的需求,所关心的不是这些商品的号码、尺寸或型号的平均数,而是这些商品

的号码、尺寸、型号的众数。又如，在竞选中，候选人为了获胜，一般是选择在选民最感兴趣话题上作文章。

3. 众数还常用来表示最可能的结果

这方面的应用常与预测有关。从下两个例子就可看出。

在天气预报中，在预测未来一段时间内的降雨量就常用到概率密度 $f(x)$，如果气象台在预报时必须给出一个降雨量数值，而大家知道气象台准确预报的记录是关系到它的声誉的，那么最好的策略是选择众数。

又如在 PERT(计划评审技术) 等随机网络规划中，完成某项任务所需的时间常被假设为贝他分布，计算完成某项任务所需的时间期望值 t 常用如下公式：

$$t = \frac{a + 4m + b}{6}$$

其中，a 表示所需时间的最乐观估计，b 表示最悲观估计，而 m 表示贝他分布的众数值估计，即最可能的所需时间估计。

四、概率密度的众数估计

概率密度的众数估计是应用众数时常遇到的问题，它可以采用如下方法解决。我们可以用 $f^*(x)$ 估计 $f(x)$，这样可以选取使 $f^*(x)$ 最大的 x 值作为众数 M 的估计值。可以采取一种变换的方法，其基本思路很简单。如果 $f(x)$ 有一个明显的众数值 M，那么我们可以希望一个样本容量 n 适度大的随机抽取样本会显示在 M 附近集聚趋势。这种集聚程度可以用许多不同方法来衡量。在此，笔者给出两种直观方法。

1. 如果我们对样本值从小到大排序，即

$$x_1 < x_2 < \cdots < x_j < \cdots < x_{j+k} < \cdots < x_n,$$

并取一些整数 $k < n$，计算 $Z(j,k) = x_{j+k} - x_j (j = 1, \cdots, n-k)$，则我们可以从这些 z 值样本中得到众数位置的信息。例如，对 z 值排序

$$Z_1 < Z_2 < \cdots < Z_{n-k},$$

则 $z_j = z(j,k)$ 值越大，众数落入其对应的区间 (x_j, x_{j+k}) 的可能性 $L(z_j)$ 越小，即

$$L(Z_1) > L(Z_2) > \cdots > L(Z_{n-k})。$$

2. 另一种方法，我们用 $I(x) = (x-h, X+h)$ 表示区间长度 $2h$ 固定而中点 x 可变的开区间。$n(x)$ 表示样本值落入 $I(x)$ 的频数。通过改变属于区间 (x_1, x_n) 的 x 值而计算得到的这些时 $n(x)$，可以提供众数位置的一些思路。$n(x)$ 越大，众数落入其对应的区间 $I(x)$ 的可能性也越大。

在找到众数大致位置后，下面给出三种众数的点估计方法，分别用 M_1，M_2 和 M_3 表示：

(1) 象上述第一种方法，用 $z(j,k) = x_{j+k} - x_j (j = 1, \cdots, n-k$ 和几个整数 $k < n)$ 表示 z 值的一个样本。

在估计 M 时，我们可选取最短区间 (x_j, x_{j+k})，即使 $z(j,k)$ 取最小值的区间。很显然，这不是点估计，只能算是一种‘最大可信度区间估计’。为了得到点估计，一种简单的方法就是取最短区间的中点作为 M 的点估计，用 M_1 表示。

(2) 象上述第二种方法，计算 $I(x)$ 和 $n(x)$。在估计众数 M 时，可选取 $n(x)$ 最大时对应的 $I(x)$，同样这也不是点估计，我们可选取 $I(x)$ 的中点作为众数 M 的点估计，用 M_2 表示。

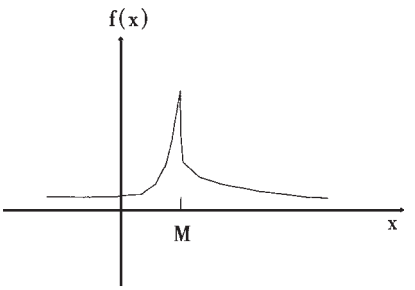
(3) 这种方法比较复杂，计算公式如下：

$$M_3 = \frac{\sum_{j=1}^{n-k} \frac{1}{2}(x_{j+k} + x_j) \Big/ (x_{j+k} - x_j)^p}{\sum_{j=1}^{n-k} 1 \Big/ (x_{j+k} - x_j)^p}$$

其中，$0 < P < k$，$\frac{1}{2}(x_{j+k} + x_j)$ 是区间 (x_j, x_{j+k}) 的中点，M_3 构造的思路是对这些中点采用加权平均方法。与前两种方法相比，此法的特点是使用了所有的样本值。

上述的三种方法是笔者在直觉的基础上构造的，毫无疑问，这三种方法估计效果的好坏与概率密度 $f(x)$ 本身的特性密切相关。更具体地，让我们考虑下面两种不希望的情形：

如果 $f(x)$ 显示一个很尖的峰（如下图），那么若样本容量不很大的话，三种方法，特别是 M_1 和 M_2，其估计效果可能会很差。



尖峰概率密度图

如果 $f(x)$ 是皮尔逊型的指数形式，那么我们得到的点估计值就一定是有偏的。