众数计算问题之我见

——数值型数据集中趋势测度探讨

吴海建

ABSTRACT

In the formula of the current Mode, there is some blemish. The Mode does not present the concentrate trend of all the frequencies. It not only has no a consideration of the group interval, but also the further groups' frequencies. In this paper, the author bring upped a new calculation formula of Mode.

关键词：众数；集中趋势；组距分组；计算公式

一、问题提出的背景

众数(Mode)是一组数据中出现次数最多的变量值，它通常是作为统计分组整理后数据集中趋势的测度值，用 M_0 表示。众数不仅可以用来测度定类数据的集中趋势，而且也可以作为定序数据、数值型数据(即定距数据和定比数据)集中趋势的测度值。

测度定类数据和定序数据的集中趋势，由于它们的变量值均为离散型，所以根据定义很容易确定众数。而对数值型数据或定量数据，则由于这些数据可能是离散型，也可能是连续型，所以数值型数据集中趋势的测度就较为复杂。

目前，在数值型数据集中趋势测度中，众数的计算分为两种情况：

其一，当数据是单变量值分组时，众数的确定方法与定类数据、定序数据等情况类似，即查找出现次数最多的变量值并确定为众数值。这种情况的众数计算非常简单，本文不再赘叙。

其二，当数据是组距分组时，目前都是根据公式来计算众数。但是，大多数文献中所使用的众数计算公式，其推导证明依据并不充分。本文针对此问题提出一点自己的见解，并提出一种较为严格的组距数列资料的众数计算公式。

二、目前众数计算公式的缺陷

1. 目前众数的计算公式

在组距分组的条件下，设众数组的频数为 f_m，众数组前、后一组的频数分别为 f_{m-1}、f_{m+1}，从众数组直方图的两个顶角向相邻两组直方图的两个顶角引直线，再由交叉点向横轴引垂线，垂足即为众数，如图 1 所示。

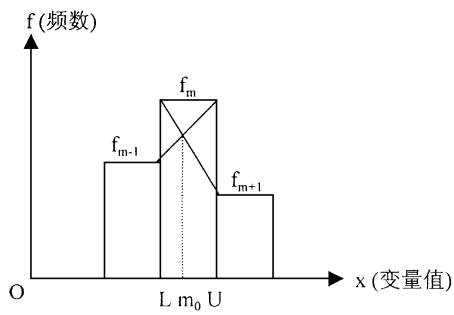

图 1 众数与相邻两组的关系示意图

根据相似三角形的关系，推导出组距分组时众数的计算公式如下：

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times I_m \text{（下限公式）} \quad \text{或} \quad M_0 = U - \frac{\Delta_2}{\Delta_1 + \Delta_2} \times I_m \text{（上限公式）}$$

式中，L、U 分别为众数组的下限值、上限值，I_m 为众数组的组距，Δ_1、Δ_2 分别为众数组的频数与其前一组、后一组的频数之差，即 $\Delta_1 = f_m - f_{m-1}$、$\Delta_2 = f_m - f_{m+1}$。

2. 公式中存在的缺陷

上述公式的几何证明过程很容易，但是，无论其下限公式还是上限公式，它们在推导证明过程中，都至少忽略了下列应该考虑的因素：

(1)组距大小对众数的影响；

(2)远离众数组的其它组的频数对众数的影响。

三、改进的众数计算公式

1. 众数概念的考察

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

从众数的含义考察, 它有三个基本要点:

第一, 是出现次数最多的变量值;

第二, 是数据集中趋势的测度值;

第三, 通常是依据统计分组整理后的组距分组资料计算。

2. 众数计算因素的考察

众数计算的影响因素, 主要是频数大小和频数分布。但在组距数列资料条件下, 众数的计算应当考虑到分组资料频数分布的特点。

(1)频数大小。各组频数大小的分布的确是反映现象的集中趋势, 但各组频数的大小还受其各组组距大小的影响。在组距分组条件下, 由于可能存在着等距分组和异距分组两种情况, 所以, 出现次数最多的变量值, 它与频数最大的组不一定完全相同。在等距分组情况下, 两者是相同的; 而在异距分组情况下, 两者通常是不一致的。

(2)频数分布。众数作为集中趋势的测度值, 它应当反映所有频数的集中趋势。如果仅用众数所在组的前、后两组的频数来确定众数值, 而不考虑其它各组频数对它的影响, 那么, 不论是在哪一种分组数列条件下, 它所反映的都只是众数组及其前、后组频数的集中趋势, 而不是全部各组所有频数的集中趋势。

故, 根据组距分组资料计算众数, 一方面, 单纯依据某组频数出现的最多来确定其为众数所在组, 这是不合适的; 另一方面, 只考虑众数组前、后两组的频数来确定众数值的具体位置, 也是不全面的。

3. 改进的众数计算公式

根据上述的分析, 众数的计算公式应为:

$$M_0^* = L + \frac{S_{m+1}}{S_{m-1} + S_{m+1}} \times I_m$$

或 $M_0^* = U - \frac{S_{m-1}}{S_{m-1} + S_{m+1}} \times I_m$

式中, L、U 分别为众数组的下限、上限, I_m 为众数组的组距, $S_{m-1} = \sum_{i=1}^{m-1} f_i$, $S_{m+1} = \sum_{i=m+1}^{K} f_i$, S_{m-1}、S_{m+1} 分别表示众数所在组前、后各组的频数累计。众数组为第 m 组, K 为总的组数, I_i 为各组的组距, f_i 为各组的频数, $i = 1, 2, \cdots, m-1, m, m+1, \cdots, K-1, K$。$\sum_{i=1}^{K} f_i = N$, N 为总频数。

其推导过程如图 2 所示。

如图 2 所示, 由于各矩形面积= 频数密度× 组距= $\frac{f_i}{I_i} \times I_i = f_i$, 所以图中 M_0^* 所在组左、右两侧各矩形的面积之和就是 S_{m-1}、S_{m+1}, 即众数组前、后各组的频数累计。如果资料是给定的, 则总频数 $N = S_{m-1} + f_m + S_{m+1}$ 为确定的值, 所以图 2 中所有矩形面积之和也为确定的值。此时, 若 S_{m+1} 减小, 则 s_{m-1} 必然增大, M_0^* 的位置向左侧偏移,

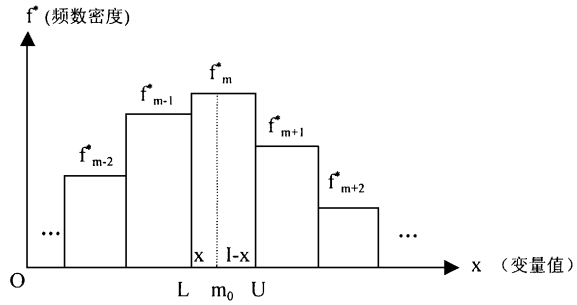

图 2 众数推导过程

(图中, 频数密度= 频数÷组距, 即单位组距上分布的频数。)

x 的值减小; 反之, M_0^* 的位置将向右侧偏移, x 的值增大。因此, 面积 xf_m^*、$(I_m - x)f^*m$ 与 S_{m-1}、S_{m+1} 之间存在如下比例关系:

$$\frac{S_{m-1}}{S_{m+1}} = \frac{(I_m - x)f_m^*}{xf_m^*}$$

即 $\frac{S_{m-1}}{S_{m+1}} = \frac{I_m}{x} - 1$

也就是 $\frac{S_{m-1} + S_{m+1}}{S_{m+1}} = \frac{I_m}{x}$

所以 $x = \frac{S_{m+1}}{S_{m-1} + S_{m+1}} \times I_m$

故 $M_0^* = L + X = L + \frac{S_{m+1}}{S_{m-1} + S_{m+1}} \times I_m$

同理: $\because I_m - x = I_m - \frac{S_{m+1}}{S_{m-1} + S_{m+1}} \times I_m$

$$= (1 - \frac{S_{m+1}}{S_{m-1} + S_{m+1}}) \times I_m$$

$$= \frac{S_{m-1}}{S_{m-1} + S_{m+1}} \times I_m$$

$$\therefore M_0^* = U - (I_m - x) = U - \frac{S_{m-1}}{S_{m-1} + S_{m+1}} \times I_m$$

4. 改进公式的使用

使用改进的众数计算公式 M_0^*, 都要先对数据计算频数密度, 并根据频数密度来确定众数所在组, 然后利用公式计算众数值。等距数列和异距数列的众数计算公式形式完全一致。

具体使用要注意, 在等距数列中, 由于各组的组距相等, 组距不影响频数分布与频数密度分布的一致性, 所以可以用频数密度, 也可以直接用各组的频数查找众数组, 并利用上述公式计算众数值。而对异距数列, 则一定要先计算各组的频数密度, 并依据频数密度确定众数组, 再利用上述公式计算众数值。

四、新、旧众数计算公式比较实例

因为众数是测度集中趋势的, 它在一定程度上反映

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

总体的一般水平。而算术平均数是测度一般水平的，它在一定程度上反映总体的集中趋势。两者的差别在于，从数是用频数测度，算术平均数是用标志值测度。故，检验新、旧众数两种计算公式的优劣，可以通过比较它们各自与算术平均数的差别，一般来说，差别小者较好。

用 M_0^*、M_0 分别表示新、旧众数，$\bar{x}$ 表示算术平均数，根据资料计算比较如下：

表 1　某车间 50 名工人日加工零件数分组表

按零件数分组(个) x	组距 I	组中值	频数 f (人)	频数密度	频数累计 S 向上	频数累计 S 向下
(1)	(2)	(3)	(4)	(5)=(4)÷(2)	(6)	(7)
110~115	5	112.5	6	1.2	6	50
115~120	5	117.5	11	2.2	17	44
120~125	5	122.5	15	3.0	32	33
125~130	5	127.5	8	1.6	40	18
130~135	5	132.5	6	1.2	46	10
135~140	5	137.5	4	0.8	50	4
合计		—	50	—	—	—

$$\bar{x}=\frac{\sum xf}{\sum f}=123.40$$

$$M_0^*=L+\frac{S_{m+1}}{S_{m-1}+S_{m+1}}\times I_m=120+\frac{18}{18+17}\times 5=120+2.57=122.57$$

$$M_0=L+\frac{\Delta_1}{\Delta_1+\Delta_2}\times I_m=120+\frac{4}{4+7}\times 5=120+1.82=121.82$$

2. 在异距数列情况下，如表 2 资料：

表 2　某车间 50 名工人日加工零件数分组表

按零件数分组(个) x	组距 I	组中值	频数 f	频数密度	频数累计 S 向上	频数累计 S 向下
(1)	(2)	(3)	(4)	(5)=(4)÷(2)	(6)	(7)
110~120	10	115	17	1.7	17	50
120~125	5	122.5	15	3.0	32	33
125~130	5	127.5	8	1.6	40	18
130~135	5	132.5	6	1.2	46	10
135~140	5	137.5	4	0.8	50	4
合计	—	—	50	—	—	—

$$\bar{x}=\frac{\sum xf}{\sum f}=123.15$$

$$M_0^*=L+\frac{S_{m+1}}{S_{m-1}+S_{m+1}}\times I_m=120+\frac{18}{18+17}\times 5=122.57$$

M_0 用频数计算失效。若类似地用频数密度替代频数进行计算，即 Δ_1、Δ_2 分别为众数组的频数密度与其前、后组频数密度之差，则有

$$M_0=L+\frac{\Delta_1}{\Delta_1+\Delta_2}\times I_m=120+\frac{1.3}{1.3+1.4}\times 5=122.41$$

3. 将表 1、表 2 资料的计算结果列表比较如表 3：

表 3　新、旧众数计算结果比较分析表

资料来源	数列种类	$\bar{x}$	M_0^* (新)	M_0 (旧)	算术平均数与众数之差 $\bar{x}-M_0^*$	算术平均数与众数之差 $\bar{x}-M_0$
表 1	等距	123.40	122.57	121.82	0.83	1.58
表 2	异距	123.15	122.57	122.41	0.58	0.74

在表 3 中，表 1 和表 2 的资料是同一批数据资料，其差别仅在于分组时选择的组距不同，即等距与异距。

从计算结果来看，可有如下结论：

第一，在等距数列和异距数列两种情况下，M_0^* 与 $\bar{x}$ 的差别，均小于 M_0 与 $\bar{x}$ 的差别。

第二，对同一批数据，新的众数公式计算结果不受数列的组距变化影响。

故，新的众数计算公式具有集中趋势测度的稳定性、方法应用的广泛性、数值计算准确性及与平均水平一致性等特性，是一种较理想的众数计算公式。

作者简介：吴海建，男，41 岁，1984 年毕业于中国人民大学，现任教于河南财经学院统计学系。

（责任编辑：石庆焱）

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net