# 集中趋势测度中的几个问题探析*

● 孙小素

(山东工商学院 统计学院,山东 烟台 264005)

摘 要:集中趋势的测度是描述统计学的重要内容。为了测定数据的集中趋势,统计学家设计了许多方法,主要有:不同的变量类型和次数分布类型,决定了不同的集中趋势测度方法;个体量与总量的不同对应关系,决定了数值平均数的不同计算方法;研究目的不同,研究对象的特点不同,决定了平均发展速度测度时的不同设计方法。充分体现了统计的科学性和艺术性。

关键词:集中趋势;测度;变量类型;分布类型
中图分类号:C812 文献标识码:A 文章编号:1004-5465(2009)04-107-05

## Exploration on the Measurement of Central Tendency

SUN Xiao-su

(Dept. of Statistics, Shandong Institute of Business and Technology, Yantai 264005, China)

Abstract: Measurement of central tendency is the key content in statistics. Statisticians put forward many methods to measure the central tendency in data which are scientific and artistic. First, methods to measure the central tendency in data are decided by the different types of variables and their distributions. Second, they are decided by the relation between individual value and population value. Third, they are designed according to different rasearch purpose in the measure of mean development speed.

Key words: central tendency; measurement; types of variables; types of distribution

集中趋势的测度是描述统计学的重要内容。为了测定数据的集中趋势,统计学家针对不同的研究对象和研究目的,设计了许多指标和方法,体现了统计的科学性和艺术性。然而,人们在实际应用中,总是不能选择出最恰当的指标反映数据的集中趋势。究其原因,主要是现有的教科书对这部分内容的介绍不够透彻,甚至在部分内容的介绍上缺乏严谨性。因此,有必要结合这些指标和方法的特点,揭示隐藏其中的科学性和艺术性,让人们在理解的基础上正确选择集中趋势的测度

* 收稿日期:2009-06-10
作者简介:孙小素(1965-),女,河南焦作人,山东工商学院统计系副教授,厦门大学计划统计系博士生,研究方向:质量管理、统计理论与方法。

指标。

## 一、集中趋势测度中涉及的研究对象

### （一）三种不同的变量类型

在统计学中，涉及到三种不同类型的变量：分类变量、顺序变量和数值型变量。统计学家针对不同的变量类型，设计出了不同的集中趋势测度方法。

### （二）三种不同的次数分布形状

集中趋势既可依据未进行任何整理的原始数据，也可根据整理后的数据加以确定，只是适用的指标和方法不同。这里，整理是指排序、编制次数分布数列等活动。并且，次数分布的形状不同，适用的集中趋势指标也不同。

由于社会经济现象性质不同，各种统计总体各有不同的次数分布，形成各种不同类型的分布特征。概括起来，各种不同性质的社会经济现象的次数分布类型，大致有三种：钟型分布、U型分布和 J型分布。

1. 钟型分布。钟型分布的特征是“两头小，中间大”，即靠近中间的变量值分布的次数多，靠近两边的变量值分布的次数少，其曲线图宛如一口古钟，如图 1所示。

钟型分布可分为对称分布和偏态分布（非对称分布）。对称分布是以某变量值为对称轴，左右两侧对称，两侧变量值分布的次数随着与中间变量值距离的增大而渐次减少，如图 1（Ⅰ）所示。偏态分布为非对称的钟型分布，它们各有不同方向的偏态，如图 1中的（Ⅱ）、（Ⅲ）。图（Ⅱ），曲线是正偏分布（或称右偏分布），图（Ⅲ），曲线是负偏分布（或称左偏分布）。

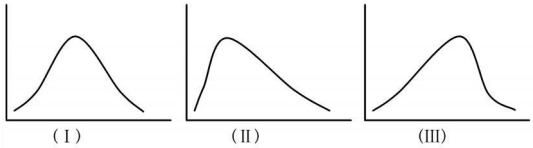


图 1　钟型分布示意图

2. U型分布。U型分布的形状与钟型分布相反，靠近中间的变量值分布次数少，靠近两端的变量值分布次数多，形成“两头大，中间小”的 U型分布。例如人口死亡率按年龄的分布便是如此。人口总体中，幼儿和老年人死亡率高，而中青年死亡率低。图 2是 U型分布图。

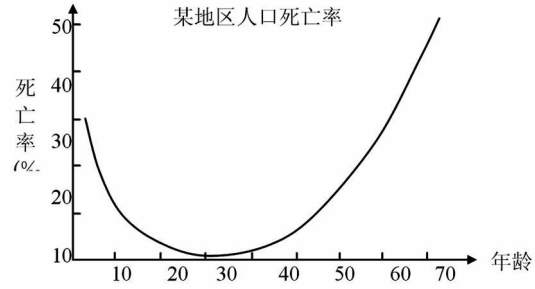


图 2　U型分布示意图

3. J型分布。J型分布有两种类型，一种是次数随着变量的增大而增多，呈正 J型，如图 3（Ⅰ）所示。例如心脑血管发病率按年龄的分布便是如此。另一种呈反 J型分布，即次数随着变量增大而减少，如图 3（Ⅱ）所示。例如肺炎发病率按年龄的分布便是如此。

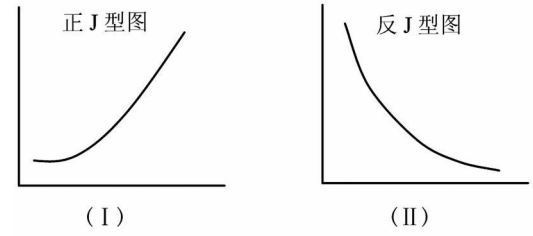


图 3　J型分布示意图

## 二、集中趋势测度中需辨析的几个问题

### （一）测度集中趋势，谁更名副其实？

集中趋势是指一组数据向其某值靠拢的倾向和程度。统计学家设计的集中趋势测度指标主要有三个：众数、中位数和数值平均数，但哪一个指标更名副其实呢？

1. 众数。变量数列中出现次数最多的变量值称为众数。从众数的定义中不难发现，用众数反映变量数列的集中趋势最直接，最名副其实。不仅如此，众数的适用范围也最广。不管是何种变

量（分类变量，顺序变量，数值型变量），也不管这些数据呈现何种分布（钟形分布，包括对称的和非对称的；U形分布；J形分布），只要进行了整理，编制成次数分布数列，都可以用众数反映其集中趋势。

2. 中位数。将变量从小到大（或从大到小）排序后，处于中间位置上的那个数就是中位数。和众数不同的是，用中位数反映变量数列的集中趋势是有条件的。首先，只有顺序变量和数值型变量才可能用中位数反映其集中趋势。因为这两类变量能够排出顺序，而分类变量无法排出数据的大小顺序（即使排出顺序，也没有任何意义），因此，分类数据不能用中位数表示其集中趋势。其次，顺序变量和数值型变量的分布必须是钟形的。因为，对于U形分布，中位数不仅不是变量数列的集中趋势，恰恰相反，它是变量数列分布最稀疏的地方（见图2）；而对于J形分布，中位数所表示的密集程度只相当于数列最集中程度的一半左右（见图3）。因此，对于形状为U形分布和J形分布的顺序变量和数值型变量，其集中趋势不能用中位数来测度，只有钟形分布方可。

总之，要想用中位数反映变量数列的集中趋势，必须满足两个条件：第一，变量是顺序变量或数值型变量；第二，这两类变量的分布必须是钟形的。

3. 数值平均数。在集中趋势的测度指标中，数值平均数是非常重要的一个。不管数据是否加以整理，都可计算数值平均数。之所以用数值平均数来测度变量数列的集中趋势，是因为客观实际中，许多社会现象统计总体的分布都趋于钟形分布，例如，农作物的单位面积产量的分布、零件公差的分布、商品市场价格的分布等。而钟形分布最大的特点是，变量的平均水平基本上是数列分布最密集的地方。这就是说，能不能用数值平均数表示变量分布的集中趋势也是有条件的。其一，变量必须是数值型变量，因为只有数值型变量计算其数值平均数才有意义；其二，数值型变量的分布必须是完全对称的钟形分布（见图1（Ⅰ））。之所以强调必须是完全对称的钟形分布，是因为对于右偏的钟形分布（见图1（Ⅱ）），数值平均数会比众数偏大；对于左偏的钟形分布（见图1（Ⅲ）），数值平均数又会比众数偏小，因此数值平均数只是数据集中趋势的一个近似值，近似的程度随钟形分布的偏斜程度而不同，偏斜程度越小，用数值平均数反映数据的集中趋势就越精确；反之，则越不精确。

总之，只有满足上述两个条件，变量的数值平均数才能等价于其集中趋势，把数值平均数作为集中趋势的测度指标才会名副其实。上面讨论的结论可用表1简洁地表示出来。

| 集中趋势测度指标 | 适用的数据类型 | 适用的分布形状 | 数据是否整理 |
|---|---|---|---|
| 众数 | 分类变量、顺序变量、数值型变量 | 钟形分布、U形分布、J形分布 | 是 |
| 中位数 | 顺序变量、数值型变量 | 钟形分布 | 是 |
| 数值平均数 | 数值型变量 | 钟形分布 | 是否整理均可 |

（二）位置平均数的说法，什么时候用更合适？

在一般教科书中，众数和中位数又被称为位置平均数，用来反映变量数列的一般水平。我们认为，这也是有条件的。其一，变量必须是数值型变量，因为只有数值型变量计算其平均水平才有意义；其二，数值型变量的分布必须是钟形的。首先来看U形分布的众数和中位数。对于U形分布（见图2），变量有两个众数，一个是变量的最小值，另一个是变量的最大值，不管是哪一个，众数都不能表示变量的一般水平。U形分布的中位数在其底部中央，虽不能表示数据的集中趋势，但却是数据的一般水平，只是这个一般水平无法合理解释。在统计中，对于无法合理解释的指标，人们便也不再计算。也就是说，U形分布的中位数不能被称为位置平均数。同理，也可解释J形分布的众数和中位数不能作为数据一般水平代表的原因。只有满足这两个条件，变量的集中趋势才能等价于其一般水平，位置平均数的称呼才会名副其实。

（三）中位数的位置究竟该如何确定？

就笔者查阅到的文献看，在介绍中位数的计

算公式时，一般都给出了两套计算公式：

其一，对于未分组或已分组且为单项式数列的资料，中位数按下式确定：

$$M_e=\begin{cases} x_{\frac{n+1}{2}} & n\text{为奇数} \\ \frac{1}{2}\left\{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right\} & n\text{为偶数} \end{cases} \quad (1)$$

或

$$M_e=\begin{cases} x_{\frac{\sum f+1}{2}} & \sum f\text{为奇数} \\ \frac{1}{2}\left\{x_{\frac{\sum f}{2}}+x_{\frac{\sum f}{2}+1}\right\} & \sum f\text{为偶数} \end{cases} \quad (2)$$

式中，$M_e$表示中位数，$x_{(i)}$为一组数据$x_1$，$x_2$，…，$x_n$从小到大排序后处于第$i$个位置上的数，$n$或$\sum f$为总次数。其中，(1)式用于确定未分组资料的中位数，(2)式用于确定已分组且为单项式数列资料的中位数。

其二，对于已分组且为组距式数列的资料，中位数按下式确定：

$$M_e=L_M+\frac{\sum f/2-S_{M_{e-1}}}{f_M}d_M\text{(下限公式)}$$

$$=U_M+\frac{\sum f/2-S_{M_{e+1}}}{f_M}d_M\text{(上限公式)} \quad (3)$$

式中，$L_M$、$U_M$分别为中位数所在组的下限和上限；$S_{M_{e-1}}$是到中位数组前面一组为止的向上累计频数，$S_{M_{e+1}}$是到中位数组后面一组为止的向下累计频数；$d_M=U_M-L_M$为中位数组的组距；其他符号的含义同前。

上面介绍的两套确定中位数的方法，除了适用的资料类型、公式的具体表达方法等形式上的不同外，其最本质的差异在于确定中位数所在位置上的不同。第一套公式确定出的中位数位置为$(n+1)/2$(或$(\sum f+1)/2$)，而第二套公式确定出的中位数位置为$\sum f/2$，显然这是不严谨的。相比之下，第一套公式确定出的中位数位置符合中位数的有关定义，更合理，也更容易理解。因此应该修订第二套公式，使之与第一套公式相一致。建议将第二套公式修订为：

$$M_e=L_M+\frac{(\sum f+1)/2-S_{M_{e-1}}}{f_M}d_M\text{(下限公式)}$$

$$=U_M+\frac{(\sum f+1)/2-S_{M_{e+1}}}{f_M}d_M\text{(上限公式)} \quad (3)$$

(四)数值平均数为什么会有不同的计算方法？

一般教科书都会介绍数值平均数的两种算法——算术平均数和几何平均数，也有教科书将调和平均数作为算术平均数的一种变形加以介绍，并指出算术平均数(包括调和平均数)用于计算静态的单位标志平均数，而几何平均数用于计算时间上相互衔接的比率的平均数。至于其中的原因，却没有任何介绍。

笔者认为，之所以会存在不同的数值平均数计算方法，根本原因在于个体量与总量的不同对应关系。这里个体量是指，对于我们研究的变量，总体中每个个体的取值；总量是指总体的总量。

研究对象的特点不同，个体量与总量的基本对应关系也不同。一般说来，个体量与总量之间存在两种完全不同的对应关系。

其一，总量$=\sum$个体量。这是我们最熟知、也最常见的个体量与总量的关系。如一个班同学的总成绩是每个同学的成绩之和，一个国家的GDP是其所有经营单位的增加值之和等等。对于这种数量对应关系，数值平均数的计算必须采用算术平均方法(或其变形，如调和平均方法)。

其二，总量$=\prod$个体量。总量与个体量之间的这种对应关系也很常见。如“十一五”期间经济的总发展速度，就等于期间各年的经济发展速度连乘积；连续几道工序加工出来的产品的合格率，就等于这几道工序各自合格率的连乘积。对于这种数量对应关系，数值平均数的计算就必须采用几何平均方法，而不能使用算术平均方法。

(五)平均发展速度测度方法如何因研究目的的不同而不同？

研究目的不同，研究对象的特点不同，决定了平均发展速度测度时的不同设计方法。根据不同的研究目的及研究对象的不同特点，统计学家设计了平均发展速度的两种计算方法：几何平均法($\bar{b}=\sqrt[n]{\prod_{i=1}^{n}b_i}=\sqrt[n]{y_n/y_0}$，$b_i$表示第$i$期的环比发展

速度，$\bar{b}$表示平均发展速度，$y_n$表示时间序列第 n 期的发展水平，$y_0$表示时间序列的最初发展水平）与高次方程法（$\bar{b}+\bar{b}^2+\cdots+\bar{b}^n=\sum_{i=1}^{n} y_n / y_0$，各符号的含义同几何平均法，通过解此方程即可得到$\bar{b}$）。这两种算法中，几何平均法侧重于考察现象最末期的发展水平，该方法计算的定基发展速度$\bar{b}^n$与实际资料最末期的定基发展速度相一致，因此，如果我们不太关心现象的发展过程，主要关心发展的最终结果，如 GDP的发展目标是否实现，就可以选择此方法。而高次方程法侧重于考察现象的整个发展过程，该方法计算的各期定基发展速度的总和（$\bar{b}+\bar{b}^2+\cdots+\bar{b}^n$）与实际资料各期定基发展速度的总和相一致，因此，如果我们关心过程的发展质量，如投资额、绿化面积，就可以选择此方法。另外，由于两种方法的设计原理不同，适用的时间序列也不相同。几何平均法既适用于时点序列，也适用于时期序列；而高次方程法只适用于时期序列。

综上所述，数据有无整理，变量属于何种类型，变量的具体分布形状，个体量与总量之间的不同对应关系，研究目的是什么，这些都会影响到集中趋势测度方法的选择。统计学家针对种种不同的情形设计出了相应的方法，充分体现了统计的科学性。细细品味，你或许还能体会到其中的艺术性。

## 参 考 文 献

[1] 黄良文，曾五一. 统计学原理[M]. 北京：中国统计出版社，2000. 52～75.

[2] 袁卫，庞皓，曾五一. 统计学[M]. 北京：高等教育出版社，2000. 50～70.

[3] 黄良文，杨灿. 统计学[M]. 成都：四川人民出版社，2006. 53～60.

[4] 曾五一，肖红叶. 统计学导论[M]. 北京：科学出版社，2006. 40～55.

[5] 吴喜之. 统计学：从数据到结论[M]. 北京：中国统计出版社，2004. 38.

[6] 贾俊平. 统计学[M]. 北京：清华大学出版社，2004. 73～85.

[7] GUDMUND R. IVERSEN, MARY GERGEN, 吴喜之，等译. 统计学：基本概念和方法[M]. 北京：高等教育出版社/施普林顿出版社，2000. 103～108.

[8] 吴凤庆，王艳明. 统计学[M]. 北京：科学出版社，2008. 75～85.