

对数据分布特征测度的分析

陈龙禹

(吉林大学 数学学院 统计系, 长春 130012)

摘要 统计数据经过整理和显示后,我们对数据分布的类型和特点就有了一个大致的了解,但这种了解只是表面上的,还缺少代表性的数量特征值准确地描述出统计数据的分布。对统计数据分布的特征,可以从三个方面进行测度和描述:一是分布的集中趋势;二是分布的离散程度;三是分布的偏态和测度。这三个方面分别反映了数据分布的测度特征。

关键词 数据分布;特征测度;分析

中图分类号 F270 **文献标识码** A

文章编号 1005-913X(2014)08-0192-01

一、集中趋势的测度

(一)分类数据:众数

众数主要用于测度分类数据的集中趋势,同时也适用于顺序数据以及数值型数据集中趋势的测度。需要注意的是,只有当数据较多,具有明显集中趋势时,计算众数才有意义,才可以利用它来作为某种决策的参考依据。如纺织企业职工“性别”中“女性”人数最多,则“女性”为众数。再如鞋厂在制定各种尺码鞋子的生产计划时,市场上销量最多的型号是众数,也是生产厂家或经销商应该重点生产和销售的型号。

(二)数值型数据:平均数

平均数用于反映所有数值型数据的一段水平。根据计算方法的不同,有算术平均数和几何平均数之分。平均数表明所有变量值的集中趋势,受极端值的影响,它是集中趋势的最主要测度值,主要用于数值型数据集中趋势的测度。

(三)众数、中位数和平均数的比较

1.众数、中位数和均值的关系

从分布的角度看,众数始终是一组数据分布的最高峰,中位数是处于一组数据中间位置上的值,而均值则是全部数据的算术平均。因此,对同一组数据计算众数、中位数和均值,三者之间具有以下关系:在单峰分布条件下,如果数据的分布是对称的,则众数、中位数和均值必定相等,即众数=中位数=均值。如果数据是左偏分布,说明数据存在极小值,必然拉动均值向极小值一方靠,而众数和中位数由于位置是代表值,不受极值的影响,因此三者之间的关系表现为:均值<中位数<众数。如果数据是右偏分布,说明数据存在极大值,必然拉动均值向极大值一边靠,则众数<中位数<均值,如图1所示。

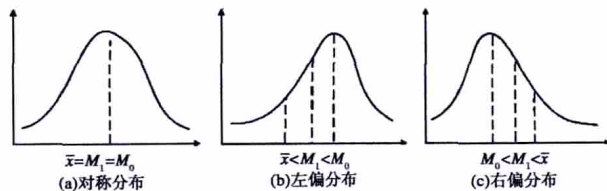


图1 众数、中位数和均值的关系图

2.众数、中位数和均值的特点与应用场合

众数、中位数和算术平均数各自具有不同的特点,掌握它们之间的关系和各自的不同特点,有助于在实际应用中选择合适的测度值来描述数据的集中趋势。

(1)虽然对于顺序数据以及数值型数据也可以计算众数,但众数主要适合于作为分类数据的集中趋势测度值,而且众数只有在数据量较多时才有意义,当数据量较少

时,不宜使用众数。

(2)中位数以及其他分位数主要适用于作为顺序数据的集中趋势测度值,虽对于顺序数据也可以使用众数,但以中位数为宜。

(3)算术平均数适合于数值型数据的集中趋势测度值,特别是当分布比较法则,不存在极端值比较偏离现象时,用算术平均数达集中趋势最合适,但平均值的主要缺点是易受数据极端值的影响,对于偏态分布的数据,平均数的代表性差。因此,当数据位偏态分布,特别是偏斜的程度较大时,可以考虑选择众数或中位数等位置代表,这时它们的代表性要比平均数好。

(4)算术平均数包含的信息是最多的、最丰富的,算术平均数具有两个重要的数学性质,即所有观测值与算术平均数的离差和等于零,所有观测值与算术平均数的离差平方和为最小。

二、离散程度的测度

集中趋势只是统计总体数据分布的特征之一。介于个体的差异性,总体中的各数据还呈现出与集中趋势的代表值和分散的离中趋势,这是数据的另一特征,它所反映的是各变量值远离其中心值的程度。因此,对统计数据的分析,除了要反映其分布的集中趋势外,还要反映统计数据的离散程度,以达到对数据变动规律的全面描述。

根据所依据数据类型不同,数据离散程度的主要指标有极差、平均差、四分位差、方差、标准差以及离散系数等。测度离散程度的主要作用如下。一是反映现象总体中变量分布的离中趋势。总体各单位的标志值存在差异,标志变动度表明总体各单位标志值的分散程度。变量值的差异越大,离散趋势也越大;反之,变量值越小,离散趋势也越小。二是衡量均值的代表性。均值作为总体数量标志的代表,其代表性取决于总体各数据的差异程度。总体中各数据的变异程度越大,均值的代表性就越小;反之,总体中各数据的变异程度越小,均值的代表性就越大。三是测定现象变动的均匀性或稳定性程度。离散程度能够表明生产过程的节奏性和其他活动的均衡性,可作为企业产品质量控制和经济管理工作的重要依据。

三、偏态与峰态的测度

(一)偏态的测度

在客观实际生活中,一些经济变量的次数分配往往是非对称型的,如收入分配、市场占有率、资源配置等等,这些经济变量经分组后,总体各单位在不同的分组变量值下分布并不均匀对称,而呈现出偏斜的分布状况,统计上将其称为偏态分布。利用众数、中位数和平均数之间的关系就可以判断分布是对称、左偏还是右偏。显然,判断偏态的方向并不困难,但要测度偏斜的程度则需要计算偏态系数。统计分折中测定偏态系数的方法很多。

(二)峰态的测度

峰态是指数据分布的尖峭状况和程度。峰态是次数分布的另一个数量特征。这个特征是“某种次数分布与正态分布相比较,是尖顶还是平顶,其尖顶或平顶的程度如何。峰度是次数分布曲线顶端的尖峭程度。峰度通常分为三种:正态峰度、尖峰度和平峰度。如果分布的形状比正态分布更高更瘦,则称为尖峰分布;如果分布的形状比正态分布更矮更胖,则称为平峰分布。峰态系数是统计中描述次数分布状态的又一个重要特征值,用以测定邻近数值周围变量值分布的集中或分散程度。

参考文献:

- [1] 陈文丽,韩立岩.λ-可加模糊测度的一般形式及分布特征[J].模糊系统与数学,2007(1).
- [2] 张红,王新生,余瑞林.基于Voronoi图的测度点状目标空间分布特征的方法[J].华中师范大学学报(自然科学版),2005(3).

[责任编辑:文筠]

收稿日期:2014-05-18

作者简介:陈龙禹(1994-),男,四川云阳人,本科学生,研究方向:数学。