文章编号: 1006-3080(2010)01-0130-04

离散样本数据的多峰分布测度及其应用

韩志国, 陈智高, 王基铭
(华东理工大学商学院, 上海 200237)

摘要: 在描述样本数据多峰分布现象的基础上, 定义了用样本数峰值之间的凹陷区域面积测度表征离散样本数据多峰分布程度的统计指标, 给出了多峰分布度的计算方法。提出了在多样本组中识别具有多峰分布形态的样本组和测度该样本组多峰分布度的步骤。以一个具有多峰分布形态的离散样本组为实例, 阐述了多峰分布度在数理统计中的应用, 分析了造成多峰分布的原因和解决办法, 提高了样本的统计指标性能。

关键词: 样本数据分布; 统计指标; 多峰分布度
中图分类号: O213 文献标志码: A

Multi-peak Distribution Measurement of Discrete Sample Data and Its Application

HAN Zhi-guo, CHEN Zhi-gao, WANG Ji-ming
(*School of Business, East China Unwiersity of Science and Technology, Shanghai* 200237, *China*)

Abstract: By analyzing the multi-peak distribution of discrete sample data, a statistical index of multi-peak distribution degrees is defined and the computing formula is given in this paper. The multi-peak distribution degree is measured by means of the concave area between the peaks' values of those samples. Then, the procedure is presented to identify the samples with multi-peak distribution and to measure the distribution degree. Finally, the proposed method is illustrated by using an application example, in which the causes of multi-peak distribution are analyzed and the solving method is given.

Key words: sample data distribution; statistical index; multi-peak distribution degree

在基于抽样的统计分析中, 单变量样本数据分布形态是最基本的分析对象, 也是多变量多因素统计分析的基础。单变量样本数据的集中和离散趋势分析已有比较成熟的统计指标可采用, 比较常见的是用平均值反映数据的集中趋势, 用标准差反映数据的离散趋势[1]。这两个指标对于大多数单变量样本观测值的分布分析是简便和有效的。此外, 在数理统计中, 还有 k 阶原点距和 k 阶中心距等统计指标, 用来更细微地描述样本观测值的分布特征。这些统计指标再与标准差等相结合又构成了一些新的统计分析指标, 其中比较重要的有三阶中心矩与标准差的三次方之比的偏度系数、四阶中心矩与标准差的四次方之比的峰度系数, 前者反映分布的偏斜程度, 后者反映分布的陡峭程度。

这些统计指标从某个角度描述样本数据的分布特征, 而且综合性的偏度系数和峰度系数都是针对单峰分布形态。由于样本数据的统计分析指标相对成熟并已广泛应用, 目前相关的研究已不多见。桂文林[2]研究了标准差和平均差的内在关系, 认为在一般分布的情况下, 平均差描述了离散空间的最大

收稿日期: 2009-04-13

作者简介: 韩志国(1971-), 男, 北京人, 博士生, 主要研究方向: 工程管理。E-mail: hanzg@sinopec.com.cn

垂直距离, 方差指标度量了离散空间的面积; 平均差在数量上不会超过标准差, 方差超过平均差平方的部分是离差的方差, 反映的是离差间的变异程度。王学民[3] 举例说明了一些文献将偏度描述为反映分布在众数两边的对称偏斜性的一个量是欠妥当的, 将峰度描述为反映分布在众数附近"峰"的尖峭程度的一个量是错误的。除少量概念上的基础研究外, 单变量样本数据分布特性的研究主要集中在应用领域。例如, 构建均值-方差-峰度资产组合优化模型, 用于分散风险并取得适当投资收益的组合投资方式的决策分析[4]; 将偏度和峰度引入到收入分配公平性的研究, 从 4 种分布形态分析贫富均匀问题[5] 等。

已有的单变量样本数据分布研究都是针对单峰形态的, 但实际中经常出现多峰分布的情况。如果一组样本数据的多峰分布现象比较严重, 表明存在需要引起关注的样本数据歧义问题, 而平均值、标准差、偏度和峰度等指标难以识别这种现象。具有多峰分布形态的异常样本数据会降低统计分析结果的可信度, 识别这类样本的方法目前还很鲜见。本文应用数理统计方法, 定义表征离散样本数据多峰分布程度的统计指标及其计算方法, 提出识别多峰分布形态样本组及其多峰分布度测度的步骤, 以期为样本数据的清理和预处理、不合理样本数据的原因分析以及问题的消解提供理论依据和可行方法。

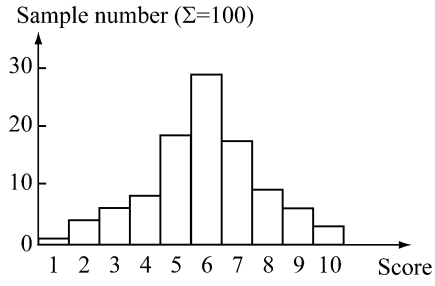
1　离散样本数据的多峰分布形态

多峰分布形态指一组样本数据出现多个峰值, 其他数据分布在这些峰值周围的一种分布趋势。相同的平均值和标准差可能具有不同的样本数据多峰分布形态。表 1 所示的 A、B 两组各含 100 个离散样本数据的例子可以说明这种现象。示例样本为一个评价指标权重的专家打分结果, 打分分值为[0, 10] 内的整数, 在各个分值上分布着不同数量的样本。A、B 两组的平均值 $\bar{x}$ 和标准差 σ 非常接近, 但分布形态差异很大, 如图 1 所示。

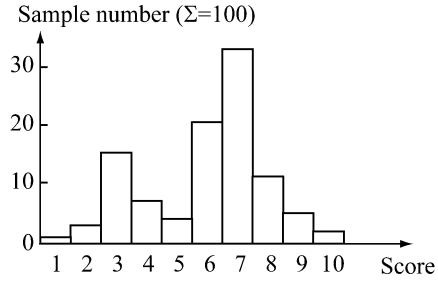
表 1　多峰分布比较示例的样本数据(样本数$\sum$= 100)

Table 1　Sample data of multi-peak distribution examples for comparing

	Scores of sample											$\bar{x}$	σ
	0	1	2	3	4	5	6	7	8	9	10		
Sample group A	1	4	6	8	18	28	17	9	6	3	0	4.92	1.86
Sample group B	1	3	15	7	4	20	32	12	4	2	0	4.96	1.97



(a) Distribution of sample group A



(b) Distribution of sample group B

图 1　两组样本数据分布形态差异程度的比较示例

Fig. 1　An example of two sample groups with different distribution

尽管两组样本数据的平均值和标准差基本一致, 但 A 组样本数据呈现出比较理想的正态分布形态, 而 B 组样本数据则表现为一大一小的两个峰值。对于这种多峰分布特征有必要加以识别和测度, 为分析产生的原因提供线索。

2　多峰分布度的定义与测度

从多峰分布形态入手, 可以发现有无多峰现象的主要特征是在样本数据取值区间内是否形成样本数的凹陷。如果存在凹陷, 即有多峰, 否则就没有多峰; 凹陷程度越重, 意味着多峰程度越显著, 两者呈正比关系。由此, 本文以样本数据取值区间内凹陷区域面积的大小来测度多峰分布的程度。

定义 1　对于一组观测值分布在[x_1, x_2] 值域内的离散样本数据, 其中至少有一个观测值 j^* 具有最多的样本数目, 在 x_1, x_2 的两端至 j^* 的区间内如果存在 k 个观测值 j_k ($0 < k < x_2 - x_1$), j_k 的样本数目大于其左右相邻观测值的样本数目, 即称这组离散样本数据的分布形态为多峰分布形态。

定义 2　根据定义 1, 用 j_k 与 j^* 之间出现的凹陷区域面积的大小来测度一组样本数据的多峰分布的显著程度, 称为多峰分布度, 记为 M。

M 以[0, 1] 值域的无量纲数值表示。考虑 M 的两个极端情况, 当样本数据分布没有凹陷时, M = 0;

当凹陷最为严重时，$M=1$。设一个样本数量为 n 的离散值样本组，样本数据限定在某区间[x_1，x_2]内，其中 $n/2$ 的样本观测值为 x_1，$n/2$ 的样本观测值为 x_2。显然，该样本数据的分布具有最大的凹陷区域面积，多峰分布度计算式如下：

$$M=\frac{\sum_k m_k}{\frac{n}{2}(x_2-x_1-1)}=\frac{2\sum_k m_k}{n(x_2-x_1-1)}$$

$$(k=1,2,\cdots,K;x_2>x_1) \qquad (1)$$

其中：K 为样本组中凹陷区域的数目；m^k 为样本组中第 k 个凹陷区域的面积。

对于表1中的 B 组样本数据，$j^*=6$，$j_1=2$ 上有一个峰值，它与 j^* 之间有一个凹陷区域，根据式(1)，相应的多峰分布度为 $M=0.42$。

3 多峰分布形态识别和测度的步骤

测度离散样本数据多峰分布度，首先要找出样本数据的凹陷区域。在实际的统计分析中，当样本组比较多时，凹陷区域的识别和凹陷面积的计算很费时。利用计算机编程的优势，能快速地找出具有凹陷区域的样本组，并计算出这些样本组的多峰分布度。

本文给出了在 I 个样本组中识别离散样本数据具有凹陷区域的样本组，并测度其多峰分布度的步骤如图2所示。此步骤可以在计算机上编程实现。

4 多峰分布度应用实例

识别出多峰分布形态，并测度了多峰分布度，也就知道了哪些样本组的样本数据存在歧义，进而能够分析其形成的原因和提出解决办法。以一个石油化工工程建设项目管理绩效评价研究课题的调查数据统计分析为例，说明多峰分布度的应用。

该课题设计的项目管理绩效评价参数近600项，一个参数项对应一个样本组。这些参数的权重设定采用问卷调查和平均值统计指标的方法。结合判断抽样和配额抽样，在领域专家中确定了300多个样本。因为专家可以不回答自己认为不熟悉的评价参数，每个参数项的实际样本数在100～300之间，这样构成了近600个样本组，每个样本组有100～300个样本的离散样本数据集合。

按照图2所示步骤，通过多峰分布样本组的识别和多峰分布度的测度，发现约有5%的样本组具有比较明显的多峰分布形态。对问卷设计、参数项界定、参数项理解、样本类别等因素的综合分析，归纳出3类造成样本数据多峰分布形态的原因。其中主要的一类原因是参数项界定与参数项理解上的误差，这种误差使得不同类别的专家给出了不同的样本数据，形成了明显的样本数据的多峰分布形态。

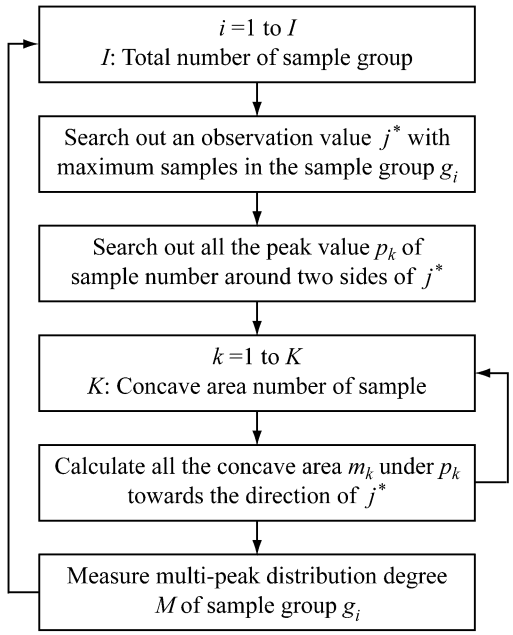


图2 多样本组中识别和测度多峰分布度 M 的步骤

Fig. 2 Procedure to identify the sample group with multi-peak distribution and to measure the degree

图3为参数项界定与参数项理解误差引发多峰分布形态的一个典型样本组，它的200个离散样本数据见表2。

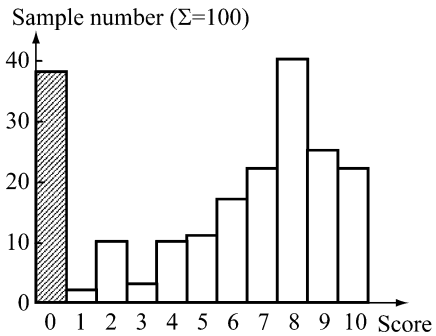


图3 典型多峰分布样本组实例的分布形态

Fig. 3 Typical distribution example of sample group with multi-peak

该样本组的平均值 $\bar{x}=5.74$，标准差 $\sigma=3.46$，样本数最多的峰值对应的 $j^*=8$。如图3所示，有38位专家将不熟悉而不打分的规则误为不相关的0分，在0分与8分之间形成了一个很大的凹陷区域。该区域内的2分又有一个样本数为10的小峰值，由于它小于0分的峰值，仅以一个大凹陷区域计算多

表 2　典型多峰分布样本组的样本数据实例(样本数 n= 200)

Table 2　Typical data example of sample group with multi-peak distribution(n= 200)

	Scores of sample											$\bar{x}$	σ
	0	1	2	3	4	5	6	7	8	9	10		
Distribution of sample	38	2	10	3	10	11	17	22	40	25	22	5.74	3.46

峰分布度。根据式(1)计算得 M=0.2122。

如上分析, 舍去误差造成的 0 分的样本是合理的, 减去 38 个样本后的 162 个样本数也能够满足样本规模要求。如此处理后的样本数据呈现明显的正态分布趋势, 平均值 $\bar{x}$= 7.08, 标准差 σ= 2.30, 多峰分布度 M=0.009 6。显然, 样本组的统计指标有了显著的改进。

5　结束语

样本数据的多峰分布表明具有歧义的数据趋向、平均值和标准差、以及原点距和中心距等统计指标难以识别和区分这种多峰分布形态。应用本文定义的以多峰之间凹陷区域面积来测度的多峰分布度能够反映离散样本数据的多峰分布显著程度, 在多样本组中识别多峰分布样本组和测度多峰分布度, 可以采用本文给出的步骤在计算机上编程实现。

离散样本数据的多峰分布度统计指标对于样本数据歧义程度的测度具有显著的实际意义, 能为样本数据的清理和预处理, 造成不合理样本数据的原因分析以及问题的消解提供数理统计上的理论依据。

参考文献:

[1] 李怀祖. 管理研究方法论(第 2 版)[M]. 西安: 西安交通大学出版社, 2004.

[2] 桂文林, 伍超标. 标准差和平均差的内在关系及应用研究[J]. 数理统计与管理, 2005, **24**(2): 50-54.

[3] 王学民. 偏度和峰度概念的认识误区[J]. 统计与决策, 2008(12): 145-146.

[4] 张萍. 均值-方差-峰度资产组合优化模型[J]. 科学技术与工程, 2008, **8**(1), 16-20.

[5] 邵建平, 邓兆卉. 分配公平性的分布偏度与峰度描述研究[J]. 统计与决策, 2008(3): 144-147.

第 **13** 届亚洲化学大会论文摘要

A Kinetic Model with Catalyst Deactivation for Propane Dehydrogenation over Pt-Sn / Al_2O_3 Catalyst

LI Qing,　*SUI Zhi-jun*,　*ZHOU Xing-gui* *

(*State Key Laboratory of Chemical Engineering*, *East China University of Science and Technology*, *Shanghai* 200237, *China*)

Abstract: Propene as the important industrial material for the production of polypropylene, acrylonitrile etc., is needed more than ever now and alternative ways instead of steam cracking of hydrocarbon feed stocks and refinery conversion processes are urgently needed to satisfy the continuously rising demand. As a promising approach, the dehydrogenation of propane has been developed commercially since 1980s. Although kinetic study of the system is helpful for reactor design and process optimization and then the application of industrial production of propene, not so many related studies have been conducted until now. In this communication, the kinetics of propane dehydrogenation on a Pt-Sn/Al_2O_3 catalyst has been studied over the temperature range of 550—600 ℃ at atmospheric pressure. The kinetic experiments were carried out in the fixed bed tubular reactor with an inner diameter of 6 mm varying both the feed composition and reaction temperature. A Langmuir-Hinshelwood mechanism assuming weak adsorption of propane was employed to describe the propane dehydrogenation process, while catalyst activity that would decline with the reaction time was described by a deactivation model, which relates the activity to the on-stream time after regeneration, the reaction temperature and the gas composition. The parameters of the kinetic models of dehydrogenation and deactivation were determined separately by fitting the experiments. The results showed that both the kinetic model for dehydrogenation and the model for catalyst deactivation were reasonable, which fit the experimental data in an acceptable accuracy.