●张训苏

次数分布偏度测定的新方法：DM—ADM模型及其原理与应用

正态分布已被广泛地应用到社会经济统计分析与推断中。但在社会经济统计实践中变量的分布往往并不是服从正态分布，而是服从偏态分布。研究偏态的程度——偏度的测定不仅有助于给出偏度测定的科学方法，而且有助于深入地认识与应用正态分布。为此，生物统计学派代表人物、著名英国统计学家卡尔·皮尔逊（Karl·Pearson）曾提出过皮尔逊法，他人也曾提出过第三动差法和四分位数等，香港胡孝绳先生编著、木屋书社1976年印行的《统计学》，以及中国统计出版社1984版权威教材《社会经济统计学原理教科书》等书刊均将这些方法作为经典理论编入其中。但现行方法存在着明显的不足①。笔者在多年来研究的基础上，提出新的方法：DM—ADM模型。该模型具有灵敏度高、依据充分、可行性强等优点。

一、DM—ADM模型的原理与判断法则

DM—ADM模型的设计思路是：将两个均以众数M_0为参照系的标志变异指标加以比较，形成一个不受数列或变量值本身平均水平与计量单位影响、具有确定取值区间的结构相对数，并能够根据该数的大小与正负科学而准确地判断和比较次数分布的偏度。其基本公式是：

$$SK_m = DM/ADM \cdots\cdots (1)$$

或$$SK_m = (\bar{x} - M_0)/ADM \cdots\cdots (2)$$

式中，SK_m为DM—ADM模型系数，DM是众数平均差，ADM为众数绝对平均差即各个标志值与众数离差绝对值的算术平均数，是其英文名词mean of absolute deviation from the mode的缩写（也是为了区别以算术平均数为参照系所建立的平均差A.D.），$\bar{X}$、M_0分别为算术平均数和众数。

在此基础上，可以导出两个具体的实用公式。即对于已分组的观察数据或离散型变量，由于

$$DM = \bar{x} - M_0 = \Sigma (X - M_0) f/\Sigma f$$

$$ADM = \Sigma |X - M_0| f/\Sigma f$$

故

$$SK_m = \Sigma (X - M_0) f/\Sigma |X - M_0| f \cdots\cdots (3)$$

同理，对于连续型变量x及其分布密度函数f(x)，有

$$SK_m = \int_a^b (x - M_0) f(x)\,dx / \int_a^b |x - M_0| f(x)\,dx \cdots\cdots (4)$$

其判断法则是：（1）$SK_m = 0$，表明$\bar{x}$与M_0相等，即在分布曲线上的两点重合，这时分布呈对称分布（若峰度值等于或接近于正态分布值，可以判定分布服从或近似服从正态

分布）；（2）$SK_m<0$，表明$\bar{X}<M_0$，在分布曲线图中$\bar{X}$在M_0的左边，此时分布负偏或左偏；（3）$SK_m>0$，分布右偏或正偏；（4）SK_m的绝对值大小与偏度成正比。

同时，依据笔者对偏度状态的划分与研究，以及$|SK_m|\leqslant 1$[2]，在参照统计理论中对相关系数给出的经验判断[3]，笔者给出DM—ADM模型判定偏度的具体尺度（见表1）。这样一来，在一般情况下均可精确地测定一个次数分布的偏度，和比较不同分布偏度的强弱。当然，在必要时要结合拟合优度检验，来进一步判定次数分布的状态。

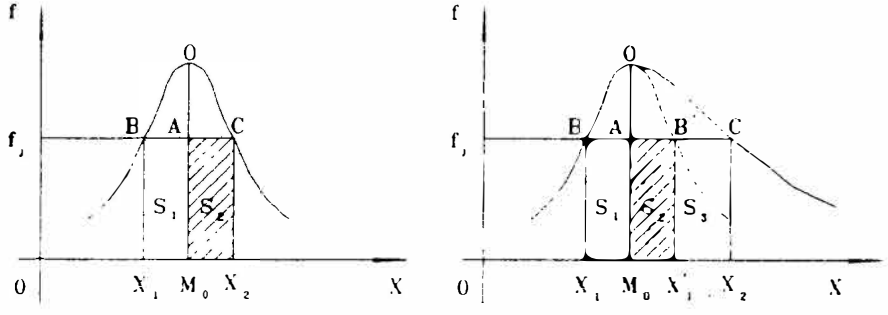
表1　　DM—ADM模型判断细则

$\|SK_m\|$的值	0	(0,0.2)	(0.2, 0.4)	(0.4, 0.6)	(0.6, 0.8)	(0.8,1)	1
$SK_m\geqslant 0$	对称分布	轻微右偏	弱右偏	中度右偏	显著右偏	高度右偏	极右偏
$SK_m\leqslant 0$	对称分布	轻微左偏	弱左偏	中度左偏	显著左偏	高度左偏	极左偏

二、DM—ADM模型的几何导源和依据

1．对次数分布对称状况的几何剖析

对于常见的钟形分布（如图1），经过中线OM_0上任意一点A（0点为分布最高点、M_0为众数位置点），作平行于横轴的直线交纵轴和曲线分别于点f_j、B和C，过B、C分别作垂直于横轴直线交X_1和X_2点。若分布对称（见图1（a）），则恒有$|AB|=|AC|=M_0-X_1=X_2-M_0$，以及长方形面积$S_1=S_2$，即$(M_0-X_1)f_j=(X_2-M_0)f_j$。



（a）对称分布（$S_3=0$）　　（b）右偏分布（$S_3\neq 0$，直线为对称部分）

图1

在非对称情况下（图1（b），以右偏为例），因不对称而导致线段AC比线段AB多出线段B′C（B′为B以OM_0为轴的对称点），面积S_{AC}x$_{2}$$_{M_0}$（记为$S_2+S_3$）比面积$S_{AB}x_{1}$$_{M_0}$多出面积$S_{B'C}X_1X_2'$（记为$S_3$，$X_1'$为$X_1$的对称点）。

而　　$S_3=(S_2+S_3)-S_1$　　（由于$S_2=S_1$）

$=(X_2-M_0)f_j-(M_0-X_1)f_j$

$=(X_2-M_0)f_j+(X_1-M_0)f_j$

由几何图形不难看出，在f_j不变的情况下，若右偏程度越强，则线段B′C就越长，面积S_3就

越大，即S_3的大小与偏度强弱成正比（左偏时，非对称部分为$-S_3$，偏度与$-S_3$成反比，但与面积S_3仍成正比）。

2．DM—ADM模型的几何导源及其理论依据（仍以右偏为例）

据上述分析，笔者便设想将所有的S_3加总与其总面积（$S_1+S_2+S_3$）的加总之比所产生的结构相对数，即$\Sigma_j S_3/\Sigma_j(S_1+S_2+S_3)$来测定偏度。为此，以分组资料为例来推导。由于

$$\Sigma_j S_3=\Sigma_j[(X_1-M_0)f_j+(X_2-M_0)f_j]$$
$$=\Sigma_j[(X_1-M_0)f_{j1}+(X_2-M_0)f_{j2}]$$
$$=\Sigma_j(X-M_0)f \quad (令 f_{j1}=f_{j2}=f_j)$$

同理$\Sigma_j(S_1+S_2+S_3)=\Sigma_j[|X_1-M_0|f_j+|X_2-M_0|f_j]$

$$=\Sigma_j|X-M_0|f$$

令X、f分别为分组资料中的组中值和频数，则有

$$\Sigma_j S_3/\Sigma_j(S_1+S_2+S_3)=\Sigma(X-M_0)f/\Sigma|X-M_0|f=SK_m$$

这正是DM－ADM模型的变形公式（3）。

若将上式对S_3求导数，得

$$SK_m'=\frac{\Sigma_j\Sigma_j(S_1+S_2)}{(\Sigma_j(S_1+S_2+S_3))^2}>0$$

，这表明SK_m与S_3成正比，而S_3与右偏程度成正比，从而证明了SK_m与右偏程度成正比。而结合左偏，就可得出$|SK_m|$与偏度成正比这个判断法则中的重要内容。这正是DM－ADM模型判定法则确立的主要依据所在。

三、DM－ADM模型的应用及其与皮尔逊法的对比

1．DM－ADM模型的应用

设现有690株豌豆，因某种原因导致株高分布如表2所示，由$\overline{X}=55.3$，ADM＝5.0，$M_0=59.5$，则由公式（2）得$SK_m=-0.83$，据其判断法则知，这690株豌豆株高呈高度负（左）偏状态。再看一个已知分布密度函数的次数分布，即连续型随机变量x的分布密度函数为$f(x)=e^{-x}$（$x\geqslant 0$），由众数定义知$M_0=0$，DM＝ADM＝1，从而$SK_m=1$，由判断法则知其呈极右偏分布，这与由其分布图（图2）可直观得其呈完全不对称的右偏分布状态相吻合，且易知$f(x)=e^{-x}$的偏度强于690株豌豆的偏度。

表2　690株豌豆株高分布图　　单位：cm

组　别	组中值x	次数 f	$(X-M_0)f$	$\lvert X-M_0\rvert f$
5—15	10	4	－198	198
15—25	20	12	－474	474
25—35	30	20	－590	590
35—45	40	50	－975	975
45—55	50	100	－950	950
55—65	60	500	250	250
65—75	70	4	42	42
合　计	－	690	－2895	3479

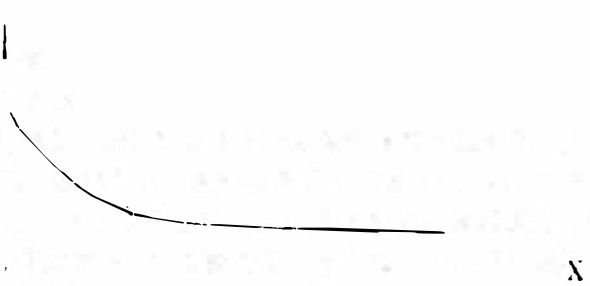


图2　$f(x)=e^{-x}$（X、0）的分布曲线

2．与皮尔逊法的对比分析

对于皮尔逊法，即由皮尔逊系数

$SK=(\overline{X}-M_0)/\sigma$（$\sigma$为均方差）

的正负与大小分别判定偏态方向与程度，我们认为其有一定的合理性和适应性，尤其是通过$\overline{X}$—M_0的大小比较来判断偏态方向是很科学的、值得借鉴的（因为$\sigma>0$），但它与DM－ADM模型比较却存在明显的不足。（1）皮尔逊法难以达到比较不同数列或分布偏度这一目的。因为σ只反映各个标志值与$\overline{X}$离差平方平均数的算术平方根大小，并不能代表变量的平均水平，根据因素分析原理，用$(\overline{X}-M_0)$除以σ不能消除因变量平均水平差异所造成的不可比因素，进而不便于科学地比较不同数列或分布的偏度。由DM－ADM模型及其设计思路知，DM－ADM模型克服了这一不可比因素，可以横向比较。（2）皮尔逊法分子$(\overline{X}-M_0)$即$\Sigma(X-M_0)f/\Sigma f$实质上是以M_0为参照系反映各个标志值的变异程度和非对称程度，而分母σ是以$\overline{X}$为参照系反映各个标志值的变异程度指标，两者在反映变异程度方面缺乏可比性，用两者之比来测定偏度就难以达到目的。而DM－ADM模型分子与分母均以M_0为参照系，不存在此缺陷。（3）皮尔逊法适应范围小、灵敏度差。对于690株豌豆，若用皮尔逊法，则$SK=-0.04$（其中$\sigma=9.5$），系数值的绝对值很小，远没有接近极值－3或3④，由皮尔逊法，无法判断出其呈高度负偏状态，而其呈高度负偏状态几乎可从分布表直观得出。对于分布密度函数$f(x)=e^{-x}$（$x\geqslant 0$），皮尔逊法同样无法判断或者说只得出很不准确的判断结论。（4）尚没有充分的依据说明、至少说没有见到过证明皮尔逊系数绝对值大小一定与偏度成正比，以及为什么SK的取值区间一般在－3与3之间。也就是说皮尔逊法理论依据不够充分。而DM－ADM模型却有较充分的依据。（5）皮尔逊法因其系数取值区间难以界定而难以给出很具体的判断法则，而DM－ADM法却可以给出具体的判断法则，从而更具有可操作性。

注：①见拙作《皮尔逊偏态测定法质疑》、《再谈皮尔逊偏态测定法》，分别刊于《河南统计》1988年第8期和1989年第3期。

②证明（以分组资料为例）：对于非负且不恒为零的次数f，有

$-|X-M_0|f\leqslant(X-M_0)f\leqslant|X-M_0|f$

$-\Sigma|X-M_0|f\leqslant\Sigma(X-M_0)f\leqslant\Sigma|X-M_0|f$　　同除以正值$\Sigma|X-M_0|f$，得

$-1\leqslant\dfrac{\Sigma(X-M_0)f}{\Sigma|X-M_0|f}\leqslant 1$，即$|SKm|\leqslant 1$。

③见杨曾武等《社会经济统计学原理讲义》，中国统计出版社1984年版。

④杨曾武等主编《社会经济统计学原理教科书》第217页指出，SK的值一般在－3与3之间。此书由中国统计出版社1984年版。