统计数据质量评估方法述评

■卢二坡

近年来我国经济持续、快速增长，引起了世界的关注，作为衡量经济发展规模和水平的宏观经济统计数据及其质量也成为国内外相关机构和研究者关注的焦点问题之一。同时，准确而充分的统计信息是决策与科学研究的基础，统计数据质量问题关系到宏观经济决策的科学性，也直接影响到科学研究特别是社会科学研究能否产生正确的研究结果。因此对统计数据质量作出科学的评估，具有重要现实意义和应用价值。

一、统计数据质量的逻辑性评估方法

（一）基于规则的逻辑性评估方法

基于规则的逻辑性评估方法主要是将经过各种专业审核的有关统计数据资料集中在一起，从总体上检验数据相互间是否平衡和是否符合逻辑。叶长法、岑国荣（1998）归纳了 4 种基本的逻辑平衡审核评估的方法，（1）差额平衡法，即观察各种增减关系的数据，看运算结果是否相互平衡；（2）同项相等的方法，指观察同一项指标在不同表种上出现的数据是否相等一致；（3）相关平衡方法，指对于那些存在必定大于（或小于、等于）关系的指标进行检查，如果出现反常，则数据必定有错；（4）运用生产和使用的平衡关系进行评估的方法，指对于那些存在明显内在联系的指标，特别是在生产和使用之间存在平衡关系的指标，看其偏差是否在合理范围之内。

基于规则的逻辑性评估方法既可用于原始调查资料，也可用于汇总数据，但是这种方法只适用于存在有逻辑平衡关系的数据，虽然能运用计算机对数据间的各种逻辑上错误进行检验和纠正，但是面对大量混杂在原始数据中的非逻辑性平衡异常数据却难以作较准确的判断。

（二）基于相关性的逻辑性评估方法

许多社会经济现象在数量上存在着相互依存关系，一个社会经济现象发生变化时，影响着另一个社会经济现象也发生数量上的变化，并且在一定的生产技术条件下，反映现象的各个指标之间存在的关系相对稳定。基于相关性的逻辑性评估方法就是在相关性较高的指标中，根据指标之间的这种关系，从已知正确的指标来对被评估指标作出评估意见，即如果指标间关系出现较大的波动，则可以初步判定被评估指标存在一定的质量问题。这种方法主要是根据指标间的比例关系、部分指标与总体指标间结构关系以及相关指标的弹性系数等方面进行判断，也可以运用回归分析、主成分分析等计量方法。

在用该方法对统计数据质量进行评估时，需要注意以下几个方面：（1）相关指标间的关系并非永远稳定；（2）与被评估指标相关联的统计数据也必须是可靠的；（3）与被评估指标相关联的指标往往不止一个，根据不同的关联指标进行判断的结果应该一致，如果用不同指标变动来观察和判断被评估数据的质量结论完相反，由此产生的争论也没太大意义。

二、从异常值的角度对数据质量进行评估

异常值又称离群值、野值、极端值等，是指在数据集中与众不同的数据，使人怀疑这些数据并非随机偏差，而是产生于完全不同的机制（总体或分布）。异常数据的产生主要有两方面的原因，一是由客观的因素造成，如总体条件突然变化或人们未知的某个因素的突然出现等等；二是由主观的因素造成，即人为的因素如被调查人员虚报、瞒报数据，调查人员算错或抄错数据等等，由这种原因产生的异常点是有质量问题的统计数据。因此从异常值角度对数据质量进行评估，不但要识别出异常值，还要结合异常值产生的背景判断其是否产生于统计数据质量问题。对异常值检验的统计方法可分为基于统计分布的方法、基于探索性数据分析的方法和基于时间序列分析的方法等。

（一）基于统计分布的异常值检验

这种方法是假定给定的统计数据服从一个随机分布（如正态分布、Γ分布等），并用不一致性测试来识别异常点。

成邦文等（2001）的研究表明，反映现象规模大小的“社会经济规模指标”如产量、产值、人员等，近似服从对数正态分布，基于此，他们提出了统计数据质量检查和异常点识别的对数正态分布检验法，并将这种方法用于我国研究与开发机构年报调查（成邦文等，2000）。成邦文等（2000）的分析思路是，采用 K—S 检验法、检验法或其它方法对作对数处理后的数据进行正态分布检验；如果数据不符合正态分布，则用法识别出异常点；最后将识别出的异常点与上年对比，如果数据没有剧烈变化，则认为该数据是正常的，否则，则认为是不正常的。成邦文等（2001b）还证明了反映社会经济规模大小的多维指标也近似服从多维正态分布，并基于此提出了对这类数据质量及其异常点进行检查和识别的多维对数正态分布检验法。

基于统计分布的异常值检验法存在两个问题，一是在许多情况下，数据使用者并不知道这个数据的分布，而且现实数据也往往不符合任何一种理想状态的数学分布；二是即使在低维（一维或二维）时的数据分布已知，在高维的情况下估计数据点的分布也是极其困难的。因此，必须事先知道数据的分布特征这就限制了它的应用范围。

（二）基于探索性数据分析的异常值检验

数据分析技术的整个操作步骤大体可以划分成两大阶段：探索阶段和证实阶段。探索性数据分析提供了丰富多彩的详细考察一组数据的方法，分离出数据的模式和特点，把它们清晰地显示给分析者（David C. Hoaglin 等，1998）。探索性数据分析能够在不毁坏原始数据中其他数据的前提下而突出异常数据或没有用处的数据，从而为判断数据质量提供依据。适合于此类目的的探索性数据分析方法主要有茎叶图法、字母值法、箱线图法、编码表、悬浮式直方图等等。

傅德印（2001）以我国“开发区高新技术企业主要经济指标”数据为例，就探索性数据分析方法特别是茎叶图法用于统计汇总数据质量控制进行了探讨。其分析思路是首先用茎叶图法找出汇总数据中的极端值，然后从横向、纵向两方面判断极端值是否为数据质量问题，以横向为例，如果每个指标上都表现为异常值，列为质量较正常组，如果该开发区数

据的某些指标正常，而个别指标表现极端，则列为有质量问题组，需要进一步调查和核实。

探索性数据分析方法具有不受极端值影响，展示数据具有包含信息量大，且能简单、直观的显示出极端值，以及不需要过多数学计算，易于理解，易于为基层人员接受的特点，因此特别适合于汇总数据的质量探制，特别是对统计技术水平要求比较低的情况，当然这种方法对于宏观统计数据质量评估也是适用的。

（三）基于时间序列的异常值分析

与探索性数据分析中的不同，时间序列分析中的异常点是以多种形式出现的，并且只有在一个描述性模型中才能对其进行定义和识别。目前对于时间序列异常值的分类标准并不统一，最为常用的分类还是加性的异常值（Additive outlier，简称 AO）和更新的异常值（Innovation outlier，简称 IO）。AO 仅影响一个观察值，它在序列中或者偏大或者偏小，经过这一点后，时间序列又恢复到正常的路径；IO 则会连续影响若干个数据点。如果某个时刻发生异常以后导致时间序列的永久性变化，这个异常值被称为均值漂移异常值（Level shift outlier，简称 LS）；介于 LS 与 IO 和 AO 之间的是暂时性变更异常值（Temporary change outlier，简称 TC），这种异常在某时刻发生以后，干扰的效应会随时间而递减消失。时间序列中的异常值既可能是由于统计数据质量问题，也可能与各种历史因素及外部冲击有关。

在对时间序列进行异常分析时，最困难的是如何决定异常值的类型以及异常值出现的确切时间。Chung chen 和 Lon-Mu liu(1993)提出的多个异常点识别的联合估计诊断方法具有良好的统计性能及抗干扰性并且能够在大多数情况下实现对异常点的正确识别。李子奈、周健（2005）主要采用这种方法对我国 36 个宏观经济时间序列的结构变化进行了全面的分析，研究表明我国的宏观经济统计数据中存在 10%以上的异常点；大部分异常点或多或少都是以聚集成堆的形式出现的，它们之间存在着深刻的联系，孤立的异常点不是我国宏观经济时间序列的主要特征。并且他们通过对异常数据出现的背景进行分析，发现总的看来我国经济时间序列异常点的出现大多数与各种历史因素以及外部冲击有关，据此他们认为这些异常数据基本上都是真实的。

应用基于时间序列的异常值检验方法的前提条件是历史数据不存在系统性偏误，其缺点是序列异常的概念并没有得到普遍的认同，运用这种方法会遗漏不少的异常数据，并且由于这种方法对历史数据要求较高，运用的诊断方法比较复杂，所以其主要适合于科研上对统计数据质量以及结构变化的诊断。

三、从误差的角度对数据质量进行评估

统计数据质量问题本质上是误差问题，即所提供的统计数据与客观的社会经济现象实际的数量特征之间的差距问题（杨清，2000）。在实践中绝对准确的数据是不存在的，我们通常强调其精确度。在精确度上足以达到我们认识社会经济现象数量特征和数量规律的需要的统计数据，就可以认为是准确的统计数据，而精确度的高低又取决于统计误差（即统计数据与客观实际值之差）的大小，因此从误差的角度对数据质量进行的评估更能体现“统计数据能充分描述经济现实的思想”。

统计调查有两大误差来源—抽样误差和非抽样误差，抽样误差是由样本推断总体过程中不可避免的误差，它本身并不是错误的结果，并且对抽样误差的研究已经非常成熟，只要能设计出样本估计量，就能给出相应的估计量的误差公式。除了抽样误差以外的所有其他误差都是非抽样误差，测定非抽样误差的方法有两种思路，一种思路是试图对估计值建立总误差模型并测算出非抽样误差的具体数值及其在总误差中所占份额的大小，然而这种思路无论是在理论上还是在实际操作上都是非常昂贵和很复杂的；杨清（2000）提出了另一种思路，即首先直接判断原始资料中是否存在失真资料，然后设法找出这些失真资料，再对其进行修正或剔除，消除这些误差的影响，进而得到一个较好的估计，从而保证统计数据的质量。屈耀辉、曾五一（2004）针对农产量抽样调查中的计量误差，以具体例子介绍了基于两种不同统计思想的定量甄别方法的具体运用，一种方法是效应比较甄别法，该方法首先进行效应比较，即运用方差分析的原理分析是否存在偏差效应，如果存在，再采 Tukey 检验法或 Scheffe 检验法判定哪位调查员偏差效应最明显。另一种方法是贝叶斯估计 3σ图甄别法，该方法首先找出获得亩产“真值”的贝叶斯估计 μ′，然后画出 μ′-3σ甄别图，并根据调查值离 μ′ 的远近判定出有无计量误差及其大小。

从误差的角度评估统计数据质量的方法，主要适合于对原始调查数据质量的控制和检验，并且随着抽样调查技术在我国应用的发展，应用这种方法对原始数据质量进行评估，显得非常重要，但是如何对各种非抽样误差，尤其是各种各样原因引起的计量误差进行检测和度量，目前对这方面的研究还很不够，这也是当前需要进一步研究的课题。

四、从核算的角度进行的评估

与前面各种统计数据质量评估方法不同，从核算角度对数据进行评估，是根据被评估指标所要求的核算方法，通过探究指标核算中存在的问题，分析其存在的原因，最大限度的挖掘现有资料，重新对其进行估算，并依据估算的结果对官方估计值进行检验。

从核算的角度重新对统计数据进行估算，也存在不少问题，例如，采用不同的估计方法对同一数据的估计结果可能相差很大，例如 Wu (1993) 和 SSB and Hitosubashi(1997)的研究，而如果没有必要的信息，则难以解释这种差距；由于所需要的基础数据有关信息难以获得，因而估算时不得不建立很强的假设（如上述 Wu(2002)的研究），这使得估算的结果会出现偏差或可靠性受到怀疑。尽管如此，通过严格规范的方法重新对有关指标进行估算，是对官方统计数据强有力的检验，同时也提供了有关该指标的可供参考的统计数据，因此这种方法和其它方法比较起来是对宏观统计数据质量评估的更加规范的方法。当然，由于宏观统计数据特别是国民经济核算数据的估算是非常复杂的系统工程，且数据的收集异常艰难，因此，这种方法的应用也仅限于专门的机构和研究人员。

五、结束语

统计数据质量受到许多方面因素的影响，并且可能产生在数据生产的每一个环节，因此统计数据质量评估是一个很复杂的问题。本文对统计数据准确性的评估方法作了归纳，并对各种方法的特点及应用场合作了分析，除了这些基本的评估方法外，还有事后预测及反常结果判断的方法、判别分析法等，不同的统计数据质量评估方法有不同的特点、应用前提以及适用场合，因此在实际对统计数据质量进行评估时，应该根据统计数据的类型，统计数据的使用者，以及统计数据所处的阶段等选取合适的方法，同时这些评估方法必须结合定性分析和调查研究才能更好的对统计数据质量作出正确的判断。

（作者单位/厦门大学经济学院）

（责任编辑/浩　天）