文章编号:1000-5862(2014)04-0437-04

异常值对义务教育均衡发展程度评估的影响

钟　君,许志勇
(天津市教育招生考试院,天津　300387)

摘要:差异系数和基尼系数是义务教育均衡发展程度评估 2 种常用的测算方法,通过向正态分布数据中加入双侧异常值,以及向截尾正态分布数据中加入单侧异常值进行模拟研究. 研究结果表明:异常值对差异系数的影响程度明显高于基尼系数. 因此,在义务教育均衡发展程度评估中,应根据评估目的综合考量,选取合适的测算方法.

关键词:义务教育;均衡发展;异常值;差异系数;基尼系数
中图分类号:B 841.7;TP 301.6　　**文献标志码**:A

0　引言

在国家大力推进义务教育均衡发展过程中,如何科学检验义务教育均衡建设成果、评估义务教育均衡发展成效是一个现实而重要的问题. 这包含 2 个层面的内容:(i)建立全面反映均衡发展内涵的指标体系;(ii)确立科学测算均衡发展程度的方法. 国内学者对此也开展了专门研究,如翟博等[1-3]从受教育机会、教育资源配置、教育过程和教育结果 4 个方面建立了评估指标,并主要采用了差异系数来测算教育均衡发展程度;傅禄建等[4]从资源配置、教育过程和办学质量 3 个方面构建了评估指标,并专门采用了基尼系数来研究教育均衡发展程度. 此外,国家《县域义务教育均衡发展督导评估暂行办法》也采用了差异系数来评估县域内中小学校际均衡情况.

评估义务教育均衡发展程度,无论是采用差异系数法还是基尼系数法,均面临统计分析中一个无法回避的问题就是异常值的影响. 这里的异常值不是指由于人工记录、数据录入等过失性错误导致的数据错误,而是指正确采集到的少数过大或过小的极端值. 在我国现阶段义务教育校际差异、城乡差异和区域差异还比较明显的情况下,异常值的存在无法避免. 但是在“办好每一所学校,教好每一个学生”的教育发展理念下,又必须关注每一所学校的数据,不应按照常规的统计分析程序清理掉异常值[5]. 因此,在对义务教育发展具有重要导向作用的均衡发展程度评估中,不同的统计测算方法对异常值的敏感程度如何、异常值对评估结果有何影响,都是值得思考,并应审慎对待的问题.

1　义务教育均衡发展程度的测算方法

1.1　差异系数

差异系数(coefficient of variation)是一种描述数据离散程度的相对差异量数,指一组数据标准差与平均值的百分比,通常用 CV 表示. 假设一组数据为 $X_1, X_2, \cdots, X_n$,其平均值记为 $\bar{X}$,标准差记为 S,则这组数据差异系数的计算公式为

$$CV = S/\bar{X} \times 100\%, \qquad (1)$$

其中 $\bar{X} = \sum_{i=1}^{n} X_i$, $S = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)}$.

由于差异系数本身经过“均值化”的无量纲化处理[6],不以原始数据的单位为单位,常用于 2 种测量单位不同或测量单位虽然相同但平均值相差很大的数据资料的差异情况的比较. 差异系数大,表示数据分散范围广,参差不齐,差异较大;差异系数小,表示数据较为集中,变动范围小,差异较小. 此外,差异系数的无量纲特性,特别适合多指标综合评价,利用多个单项指标评价结果合成综合评价结果.

正因为差异系数既能通过度量数据的离散程度来评价单项指标的均衡程度,又适合将多个单项指标评价结果合成综合评价结果,因此在义务教育均

收稿日期:2014-05-13
基金项目:天津市教育科学“十二五”规划课题(CEYP 6015)资助项目.
作者简介:钟　君(1981-),男,四川广安人,助理研究员,硕士,主要从事教育统计与测量方面的研究.

衡发展程度评估中得以采用.

1.2 基尼系数

基尼系数(gini coefficient)是意大利经济学家基尼(Corrado Gini)于1922年以洛伦兹曲线(Lorenz curve)为基础提出的,用于定量测定收入分配的差异程度. 后来逐渐有学者将基尼系数从经济领域引入教育领域,开始用基尼系数分析研究教育问题,如张长征等[7]和孙百才[8]分别用基尼系数测算方法研究了我国改革开放后近二三十年的教育公平程度及教育平等问题.

洛伦兹曲线是经济学中用以描述社会收入分配状况的一种曲线,由累计的一定人口数量占总人口的百分比与这部分人口所获得的收入占总收入的百分比的对应关系来表示,如图1所示.

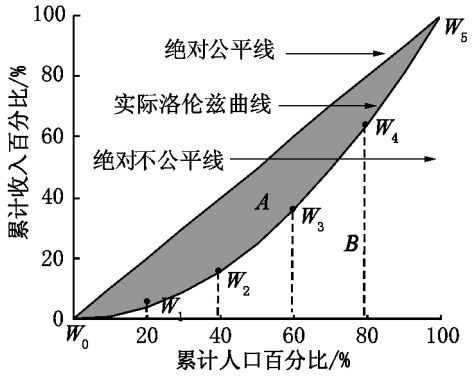

图1 洛伦兹曲线与收入不平等

基尼系数由图1中的绝对公平线和实际洛伦兹曲线围成的面积 A 与绝对公平线和绝对不公平线围成的面积 $A+B$ 之比来测度. 基尼系数通常用 G 表示,用公式表示为

$$G = A/(A+B).$$

基尼系数取值范围为0 ~ 1,越接近0就表明收入分配越趋向平等,反之,收入分配越趋向不平等. 基尼系数的计算方法很多,张建华[9]提出了一种简便易用的方法. 假定全部人口平均分为 n 组,每组人口占全部人口的比例即为 $1/n$,并假定以 W_i 表示从第1组直到第 i 组人口累计收入占全部人口总收入的百分比,则基尼系数的计算公式为

$$G = 1 - \frac{1}{n}\left(2\sum_{i=1}^{n-1} W_i + 1\right). \quad (2)$$

国际上通行的基尼系数计算,一般采用"五等份分组"法,即把全部人口平均分成5组,如图1所示. 此时,基尼系数的计算公式为

$$G = 1 - \frac{1}{5}\left(2\sum_{i=1}^{4} W_i + 1\right). \quad (3)$$

2 异常值对义务教育均衡发展程度测算的影响

2.1 异常值的诊断

假设一组数据 $X_1, X_2, \cdots, X_n$,其平均值为 $\bar{X}$,残余误差为 $V_i = X_i - \bar{X}$,标准差为 S,根据拉依达(PaǔTa)准则[10-11],异常值的判别依据为:若 $|V_i| = |X_i - \bar{X}| > 3S$,则 X_i 为异常值,应予舍弃;若 $|V_i| = |X_i - \bar{X}| \leqslant 3S$,则 X_i 为正常值,应予保留. 利用拉依达准则剔除异常数据的步骤为:(i)计算一组数据的平均值 $\bar{X}$、标准差 S、残余误差 V_i;(ii)剔除这组数据中 $|V_i| > 3S$ 对应的全部 X_i;(iii)利用剩下的数据,重复上述2个步骤,直至不再有需要剔除的数据为止.

以《义务教育均衡发展程度测评:综合教育基尼系数方法》中,全部39所小学生均计算机台数为例[12],利用拉依达准则,重复5遍剔除异常值的过程如表1所示.

表1 生均计算机台数异常值剔除过程

序号	样本数	最小值	最大值	平均值	标准差	差异系数	基尼系数
1	39	0.04	0.78	0.149 23	0.152 93	1.02	0.40
2	38	0.04	0.53	0.132 63	0.113 94	0.86	0.35
3	37	0.04	0.47	0.121 89	0.094 01	0.77	0.33
4	36	0.04	0.34	0.112 22	0.074 38	0.66	0.31
5	35	0.04	0.30	0.105 71	0.064 23	0.61	0.31
6	34	0.04	0.22	0.100 00	0.055 43	0.55	0.26

由表1可以看出,随着异常值的逐个剔除,差异系数和基尼系数也逐渐减小,尤其是差异系数的减小程度明显高于基尼系数. 采用拉依达准则剔除全部异常值后,差异系数比最初减小了约46%,基尼系数比最初减小了约35%. 由此猜想,异常值对差异系数的影响程度高于基尼系数.

2.2　异常值对差异系数和基尼系数影响程度的模拟研究

在我国现实教育环境下,无论从教育资源配置,还是从办学质量来看,总体而言还是中等学校居多,待提高学校和优质学校相对少一些. 鉴于此,为简便起见,这里假定评估指标数据服从正态分布,并在此基础上开展异常值对差异系数和基尼系数影响的模拟研究.

模拟研究的基本思路是: 产生正态分布随机数,向其中掺加异常值,然后计算并分析差异系数和基尼系数的变化情况.

2.2.1　正态分布随机数的产生　假定正态分布的均值为 μ,标准差为 σ. 当差异系数大于 1/3 时,即 $\sigma/\mu > 1/3$,此时 $\mu - 3\sigma < 0$,产生随机数中出现负数的可能性就会增大,这与现实中各评估指标一般不会出现负值的情形不符. 因此,在模拟研究中考虑 2 种情况: (i) 当差异系数小于 1/3 时,直接产生正态分布随机数,这里取 $\mu = 10, \sigma = 2$; (ii) 当差异系数大于 1/3 时,产生大于零的截尾正态分布随机数,这里取 $\mu = 10, \sigma = 5$. 鉴于现实中一个区域内学校数量的有限性,这里分别考虑 20 所学校、50 所学校、100 所学校和 200 所学校的情况,并模拟产生相应数量的随机数.

2.2.2　异常值的产生　由于现实中异常值仅是少数,这里只在各种模拟情形下,逐步添加 1 ~ 4 个异常值进行模拟研究. 但为避免异常值对原始随机样本的过大影响,控制异常值的数量不超过随机样本总量的 10%,即在 20 所学校的情况,最多添加 2 个异常值.

(i) 在 $\mu = 10, \sigma = 2$ 产生正态分布随机数的条件下,根据拉依达准则在 $(\mu - 5\sigma, \mu - 3\sigma)$ 区间随机产生 2 个异常值,在 $(\mu + 3\sigma, \mu + 5\sigma)$ 区间随机产生 2 个异常值. 按照 $(\mu - 5\sigma, \mu - 3\sigma)$ 区间异常值产生的先后顺序及 $(\mu + 3\sigma, \mu + 5\sigma)$ 区间异常值产生的先后顺序,将 2 个区间中的异常值逐个交替添加到正态分布随机数中,即研究添加双侧异常值对差异系数和基尼系数的影响.

(ii) 在 $\mu = 10, \sigma = 5$ 产生截尾正态分布随机数的条件下,根据拉依达准则,只在 $(\mu + 3\sigma, \mu + 5\sigma)$ 区间随机产生 4 个异常值,以避免在 $(\mu - 5\sigma, \mu - 3\sigma)$ 区间上产生负值,并将 4 个异常值按产生的先后顺序逐个添加到截尾正态分布随机数中,即研究添加单侧异常值对差异系数和基尼系数的影响.

每种情形经过 20 次模拟运算,计算得到差异系数和基尼系数的结果如表 2 所示.

表 2　模拟计算结果

	异常值个数	20 所学校		50 所学校		100 所学校		200 所学校	
		差异系数	基尼系数	差异系数	基尼系数	差异系数	基尼系数	差异系数	基尼系数
正态分布 $\mu=10$ $\sigma=2$	0	0.21	0.11	0.20	0.11	0.20	0.11	0.20	0.11
	1	0.28	0.10	0.23	0.10	0.22	0.11	0.21	0.11
	2	0.32	0.10	0.25	0.10	0.23	0.11	0.22	0.11
	3	—	—	0.28	0.11	0.24	0.11	0.22	0.11
	4	—	—	0.29	0.13	0.26	0.12	0.23	0.11
截尾正态分布 $\mu=10$ $\sigma=5$	0	0.48	0.25	0.44	0.23	0.45	0.24	0.45	0.24
	1	0.59	0.25	0.49	0.23	0.48	0.24	0.46	0.24
	2	0.63	0.25	0.53	0.24	0.50	0.24	0.47	0.24
	3	—	—	0.56	0.25	0.53	0.25	0.49	0.24
	4	—	—	0.58	0.26	0.54	0.26	0.50	0.25

3　讨论

差异系数和基尼系数是义务教育均衡发展程度评估 2 种常用的测算方法,本文通过向正态分布数据加入双侧异常值,以及向截尾正态分布数据加入单侧异常值进行模拟,研究结果表明: 异常值对差异系数的影响程度明显高于基尼系数. 但由于统计分布的多样性,若数据服从其他分布时,异常值对差异系数和基尼系数的影响如何,可参照此方法进一步研究.

此外,不同阶段我国义务教育均衡发展建设的重点可能不同,应根据不同测算方法的特点,选择适合评估目的的测算方法,以有利于达到促进义务教育高水平均衡发展的目的.

4 参考文献

[1] 翟博. 教育均衡发展:理论、指标及测算方法 [J]. 教育研究,2006(3):16-28.

[2] 翟博. 中国基础教育均衡发展实证分析 [J]. 教育研究,2007(7):22-30.

[3] 翟博,孙百才. 中国基础教育均衡发展实证研究报告 [J]. 教育研究,2012(5):22-30.

[4] 傅禄建,汤林春. 义务教育均衡发展程度测评:综合教育基尼系数方法 [M]. 上海:华东师范大学出版社,2013:56-74.

[5] 程开明. 统计数据预处理的理论与方法述评 [J]. 统计与信息论坛,2007,22(6):98-103.

[6] 张卫华,赵铭军. 指标无量纲化方法对综合评价结果可靠性的影响及其实证分析 [J]. 统计与信息论坛,2005,20(3):33-36.

[7] 张长征,邬志坚,李怀祖. 中国教育公平程度实证研究:1978—2004——基于教育基尼系数的测算与分析 [J]. 清华大学教育研究,2006,27(2):10-14.

[8] 孙百才. 测度中国改革开放 30 年来的教育平等:基于教育基尼系数的实证分析 [J]. 教育研究,2009(1):12-18.

[9] 张建华. 一种简便易用的基尼系数计算方法 [J]. 山西农业大学学报:社会科学版,2007,6(3):275-283.

[10] 何平. 剔除测量数据中异常值的若干方法 [J]. 航空计测技术,1995,15(1):19-22.

[11] 张敏,袁辉. 拉依达(PaǔTa)准则与异常值剔除 [J]. 郑州工业大学学报,1997,18(1):84-88.

[12] 傅禄建,汤林春. 义务教育均衡发展程度测评:综合教育基尼系数方法 [M]. 上海:华东师范大学出版社,2013:87-89.

The Impact of Outliers on the Balanced Development of Compulsory Evaluation

ZHONG Jun, XU Zhi-yong

(Tianjin Municipal Educational Admission and Examinations Authority, Tianjin 300387, China)

Abstract: The coefficient of variation and the Gini coefficient are two commonly used methods in the balanced development of compulsory evaluation. While adding the outliers to two-tail of normal data and adding the outliers to one-tail of truncated normal data by simulation, the impact of outliers on the coefficient of variation was significantly higher than the Gini coefficient. Therefore, choosing the appropriate method in the balanced development of compulsory evaluation should be based on the purpose of evaluation.

Key words: compulsory evaluation; balanced development; outliers; coefficient of variation; Gini coefficient

(责任编辑:冉小晓)