比较不同评价方法评价结果的两个新指标
——以《泰晤士报高等教育副刊》大学排名为例

俞立平,潘云涛,武夷山

(中国科学技术信息研究所,北京 100038)

[摘要] 利用英国《泰晤士报高等教育副刊》2007年世界大学排名的原始数据,分别采用主成分分析、因子分析、TOPSIS法、秩和比法、灰色关联法、熵权法 6种客观赋权法进行评价.然后采用首尾一致率比较各种评价方法结果的一致性程度,采用区分度评价各种评价方法结果的可靠性.发现排名靠前大学的首尾一致率要超过排名靠后大学的首尾一致率;排名靠前与靠后大学的区分度要超过排名中间大学的区分度;虽然标准差反映了评价结果的分散范围,但区分度与评价结果的方差无关;综合评价结果适用于宏观分级评价,对于微观严格排序的评价,最好采用同行评议与客观赋权评价相结合的方式.

[关键词] 科学技术教育评价,首尾一致率,区分度

[中图分类号] C32 [文献标识码] A [文章编号] 1001-4616(2008)03-0135-06

Two New Indicators to Compare Different Evaluation Methods' Effect
——Based on Times Higher-QS World University Rankings

Yu Liping, Pan Yuntao, Wu Yishan

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Adopting the data of the 2007 Times Higher-QS world university rankings, this paper analyzes world university rankings according to principle components analysis, factor analysis, technique for order preference by similarity to ideal solution, rank sum ratio, grey relational analysis, Entropy analysis, then analyses the evaluation results sameness and difference based on side sameness indicator and different ratio. The results show the top universities' side sameness is higher than that of bottom universities, the side different ratios is higher than that of middle universities. The different ratio is independent of standard error. The evaluation results is a good method to classify. Combining peer review with impersonal evaluation is suitable for microcosmic ranking.

Key words: S&T and education evaluation, side sameness, different ratio

科技教育评价是科教管理工作的重要组成部分,是推动国家科教事业持续健康发展,促进科教资源优化配置,提高科教管理水平的重要手段和保障.目前国内外综合评价方法有数十种之多,包括模糊数学方法、系统工程方法、技术经济方法等等.这些评价方法各有其特点,但大体上可分为两类,其主要区别在确定权重的方法上.一类是主观赋权,多数采取综合咨询评分确定权重,然后对无量纲的数据进行综合,如综合指数法、模糊综合评判法、层次分析法、功效系数法等;另一类是客观赋权,根据各指标间相关关系或各指标值变异程度来确定权数,如主成分分析法、因子分析法、TOPSIS法等等.存在的主要问题是,针对同一评价对象,选取相同的指标,采取同样的数据,但不同评价方法得出的评价结果不一致,这个问题已被广大学者所注意到,有必要进行进一步的深入研究.

一些学者在各种主客观赋权法比较方面进行了研究.刘占伟、邓四二、滕弘飞[1]从理论上分析了常见的几种评价方法的特点及存在的问题,包括层次分析法、模糊评价法、灰色理论法、物元分析法、聚类分析

收稿日期: 2008-03-12.
基金项目:国家十一五支撑计划项目 (2006BAH03B05)、国家自然科学基金 (70673019)资助项目.
通讯联系人:俞立平,博士后,副教授,研究方向:信息经济、科学计量. E-mail: chinayangzhou@yahoo.com.cn

法、价值工程法、神经网络法以及综合评价法. 陈衍泰,陈国宏,李美娟[2]将各学科领域的综合评价方法归纳、分类,讨论了各类方法的基本原理、优缺点及适用领域,指出目前综合评价存在着 3大问题:多方法评价结论的非一致性问题;方法针对性不强;理论研究与实际应用的脱节问题. 吴清平、张丹[3]在医疗工作质量评价中对秩和比法、层次分析法、TOPSIS法、系统聚类法的评价结果进行比较,发现秩和比法与其他方法具有一定的一致性. 总体上这方面的研究不多,实证研究更少,较少有定量方法对各种不同评价方法的结果进行比较.

本文拟用首尾一致率作为衡量各种评价方法评价结果一致性程度的指标. 用区分度作为各种评价方法评价结果的可靠性指标. 区分度本来是考试中的概念,指试卷测试题目对被测试者知识和能力水平的鉴别能力. 本文区分度是指各种评价方法评价结果对评价对象实际水平的区别能力. 比如对 5所大学进行综合评价,一种评价方法综合得分依次是 10、9、7、6、5,另一种评价方法综合得分依次是 10、9、7、5、3,很显然后一种评价方法的区分度要大于前者. 这里首尾一致率与区分度并不是一个对应的概念,各种评价方法首尾一致率只有 1个,而区分度和评价方法的数量是相等的.

为了进行实证研究,本文选取 2007年《泰晤士报高等教育副刊》世界大学排名[4]数据进行分析. 《泰晤士报高等教育副刊》每年推出世界大学排名,将同行评议与指标体系结合起来进行大学排名综合评价. 从 2007年开始,评价所利用的论文数据库已经由 Elsevier的 Scopus数据库取代了美国的 SCI数据库,这是因为 Scopus覆盖面更广,包括了许多非英语优秀科技期刊,这对发展中国家的大学而言相对公平. 《泰晤士报高等教育副刊》除了公布主观赋权评价结果外,还公布了各大学各项指标的原始数据,这就为深入分析提供了可能.

本文首先用主成分分析、因子分析、TOPSIS、秩和比法、灰色关联、熵权法 6种评价方法进行评价,将每种评价方法排序,然后计算首尾一致率与区分度,比较各种评价结果,并进行深入分析.

1　研究方法

1.1　首尾一致率

由于各种评价方法评价结果是不一致的,因此难以得到公认,但这并不是说评价就没有意义. 根据正态分布的规律,最好的与最差的评价对象总是少数,因此,最好的与最差的评价对象之间相差一般较大,区分度较好,各种不同评价方法容易取得共识,因此可以用来作为各种评价方法评价结果一致性程度的指标,其实质是首尾评价结果一致的评价对象占首尾评价对象的百分比.

针对 m个评价对象,将各种评价方法的所有评价结果全部进行降序排列,形成一张排序与评价方法的二维表,参考日常生活中的二八定律,在最好的 20% 的评价对象中,找出共同的 x个评价对象;在最差的 20% 的评价对象中,找出共同的 y个评价对象. 则首尾一致率为:

$$S = \frac{x + y}{0.4m}, \tag{1}$$

很显然 S的值在 0 ～ 1之间,S越大,说明不同评价方法结果更为一致.

之所以没有用首尾加上中段一致的评价对象占所有评价对象的百分比来计算一致率,是因为在评价结果排序的中段,由于数据较多,会出现许多少数评价方法不一致的情况. 比如共有 6种评价方法,在数据中段会出现较多 5种评价方法结果在中段,1种评价方法结果在首尾的评价对象,这些评价结果的公认程度肯定超过 4种评价方法在中段,2种评价方法在首尾的评价对象. 如何处理这些数据容易引起歧义. 而首尾数据较少,相邻两点距离大,这些情况相对会少一些,因此,首尾一致率更为稳定,容易得到公认.

1.2　区分度

对于每一种评价结果,假设有 m个评价对象,将其按分值 V高低进行降序排列,然后给每个分值编上序号 N,这里 $1 \leq i \leq m$,则函数 $V = f(N)$是单调递减函数,评价结果最好的坐标值为 $(V_1, 1)$,最差值坐标为 (V_m, m). 本文将区分度定义为

$$D = \frac{\sum_{i=1}^{m-1} \sqrt{(V_{i+1} - V_i)^2 + (N_{i+1} - N_i)^2}}{\sqrt{(V_m - V_1)^2 + (N_m - N_1)^2}}, \tag{2}$$

即评价结果相邻两点距离之和 (曲线长度) 与首尾两点距离 (极值距离) 的比值,$D \geq 1$,D越大,说明相邻

两点越分散,评价结果的区分度越好.由于评价数据可能存在误差,因此,区分度好意味着评价结果的可靠性高,评价更为稳定.

由于各种评价方法结果的极值(极大值与极小值之差)范围不一,比如 TOPSIS法结果的极值范围在 0～1之间,因子分析法结果的极值范围在 -1～1之间,而德尔菲法根据人们的习惯结果一般在 0～100之间,必须将结果标准化后才具有可比性.标准化的方法对区分度的计算也有很大影响,若标准化后的分值在 0～1之间,根据区分度的原理,势必导致相邻两点之间距离之和与首尾两点距离之比过小,导致不同评价方法区分度相差不大;若标准化后的分值根据人们的日常习惯在 1～100之间,但是由于评价对象数量不一,如此处理也不合适.本文将标准化处理后分值设定为 0～m之间,即最大值点坐标为 $(m,1)$,最小值点坐标为 $(0,m)$,中间某点的标准化值根据原值与极大值的差等比例处理.计算公式如下:

$$V_i = m \times \left(1 - \frac{|V_i' - V_1'|}{V_1' - V_m'}\right), \tag{3}$$

公式(3)中,V 为原指标值,V_i为标准化后的指标值.由于进行的是简单线性变换,因此不会改变原评价结果的分布规律,保真度较好.如果评价结果分值相同,则允许并列,实际上是两点完全重合.

标准化后,区分度的计算可以进一步简化为:

$$D = \frac{\sum_{i=1}^{m-1}\sqrt{(V_{i+1} - V_i)^2 + 1^2}}{\sqrt{(m-0)^2 + (m-1)^2}} = \frac{\sum_{i=1}^{m-1}\sqrt{(V_{i+1} - V_i)^2 + 1}}{\sqrt{2m^2 - 2m - 1}}. \tag{4}$$

1.3 几种客观赋权评价方法

主成分分析(Principle Components Analysis)是考察多个变量间相关性的一种多元统计方法,它通过线性变换,将原来的多个指标组合成相互独立的少数几个能充分反映总体信息的指标.它常被用来作为寻找判断某种事物或现象的综合指标,并且给综合指标所包含的信息以合适的解释,从而更加深刻地揭示事物的内在规律.

因子分析(Factor Analysis)可以看成是主成分分析的一种推广,因子分析的基本目的是用少数几个变量去描述多个变量间的协方差关系.其思路是将观测变量分类,将相关性较高即联系比较紧密的变量分在同一类中,每一类的变量实际上就代表了一个本质因子,从而可将原观测变量表示为新因子的线性组合.

TOPSIS的全称是逼近理想解的排序法(Technique for Order Preference by Similarity to Ideal Solution),它根据各被评估对象与理想解和负理想解之间的距离来排列对象的优劣次序.所谓理想解是设想的最好对象,它的各属性值达到所有被评对象中的最优值;而负理想解则是所设想的最差对象,它的各属性值都是所有被评对象中的最差值,用欧几里德范数作为距离测度,计算各被评对象到理想解及到负理想解的距离,距理想解愈近且距负理想解愈远的对象越优.

秩和比法(Rank Sum Ratio)是一种全新的广谱的实用数量方法,是田凤调[5]发明的一种统计学方法,该方法集中了古典参数统计和近代非参数统计各自优势,通过指标编秩来计算秩和的一个特殊平均数,进而进行综合评价.该方法在国内有较大的影响.

灰色关联分析(Grey Relational Analysis)是灰色系统分析的主要内容之一,用来分析系统中因素之间的关系密切程度,从而判断引起该系统发展的主要因素和次要因素.灰色关联分析的实质,就是比较若干数列所构成的曲线与理想数列所构成的曲线几何形状的接近程度,从而进行排序,列出评价对象的优劣次序,评价标准是灰色关联度,其值越大,评价结果越好.

熵(Entropy)概念源于热力学,后由 Shannon引入信息论.信息熵可用于反映指标的变异程度,从而可用于综合评价.设有 m 个待评对象,n项评价指标,形成原始指标数据矩阵 $X = (X_{ij})_{m\times n}$,对于某项指标 X_j,指标值 X_{ij}的差距越大,该指标提供的信息量越大,其在综合评价中所起的作用越大,相应的信息熵越小,权重越大;反之,该指标的权重也越小;如果该项指标值全部相等,则该指标在综合评价中不起作用.

以上各种评价方法各有特点,除了主成分和因子分析需要具备一定的前提条件外,其他方法并没有严格的条件限制.此外,TOPSIS法也可以进行加权,即在计算各评价对象与最优方案及最劣方案距离时,都可以赋予一定的权重,为了保证评价方法的客观性,本文不进行加权处理.

2 数据

本文采用数据为英国《泰晤士报高等教育副刊》2007年世界大学排名数据,共有 200所大学,主要指

标有 6个:同行评议、雇主评价 (他们愿意招募哪个学校的毕业生)、师生人数比、人均引文数、海外教授数、留学生人数比,这 6个指标中,同行评议与雇主评价是主观指标,每个指标数据已经进行标准化处理,各指标分值最高均为 100.《泰晤士报高等教育副刊》采取主观赋权法,6个指标权重分别赋值为 0.4、0.1、0.2、0.2、0.05、0.05,在此基础上计算指标总得分.需要说明的是,《泰晤士报高等教育副刊》做大学排名时,由于只精确到计算结果的整数位,因此出现了一些大学并列排名的现象.本研究保留了计算结果小数点后两位,因此排序结果与原《泰晤士报高等教育副刊》排序略有差异.

3 实证结果

3.1 各种评价方法的首尾一致率

利用 6个指标 200所大学数据,分别采用主成分分析、因子分析、TOPSIS法、秩和比法、灰色关联法、熵权法进行评价.在进行主成分分析与因子分析时,必须首先进行 KMO检验与 Bartlett检验. KMO是对样本充分度进行检验的指标,一般要大于 0.5.本文采用 SPSS进行数据处理,KMO值为 0.485,也就是说,不太适合进行主成分和因子分析;Bartlett值为 178.135, $P < 0.000$,也就是说,相关矩阵不是一个单位矩阵,可以进行主成分和因子分析.换句话说,大学排名采用主成分和因子分析的条件并不全部具备.前 3个主成分 (因子)的累计贡献率为 71.35%,因此采用前 3个主成分 (因子)进行评价.

表 1是所有评价结果排序后两侧各 20%大学中共有的所有大学名单. 200所大学中,前 40名大学共有的有 22所,后 40名大学共有的只有 3所.首尾一致率为 (22 + 3) /80 = 31.25%.

表 1 各种评价方法一致性情况

Table 1 The sameness of different evaluation methods

大学	主观赋权排序	主成分排序	因子排序	TOPSIS排序	秩和比排序	灰色关联排序	熵权排序
Harvard	1	1	4	1	2	1	2
Imperial College London	2	4	3	4	6	4	5
Princeton University	2	2	2	2	3	3	4
University College London	2	7	8	7	18	7	15
Massachusetts Institute of Technology	5	3	1	3	4	2	3
Columbia University	6	11	11	8	7	9	7
McGill University	7	23	25	13	1	6	1
Duke University	7	5	7	9	8	8	9
University of Pennsylvania	9	8	5	5	5	5	6
Johns Hopkins University	10	6	17	30	12	12	21
Australian National University	11	9	19	31	16	15	22
University of Tokyo	12	10	6	11	15	10	10
Stanford University	14	15	26	12	10	18	14
University of Edinburgh	16	14	9	20	24	17	18
Kyoto University	18	22	10	6	11	13	8
University of Melbourne	20	16	16	15	17	23	17
Northwestern University	22	19	23	21	31	21	29
University of Sydney	24	25	12	10	19	20	13
Tsinghua University(清华大学)	28	26	24	16	13	25	12
ETH Zurich	30	28	22	18	25	32	23
Monash University	31	29	15	23	20	14	20
Boston University	33	35	20	32	29	11	28
New York University	37	31	29	17	33	35	25
以上为前 20%大学的一致性结果,以下为后 20%大学的一致性结果							
University of Michigan	172	167	180	195	194	197	189
University of Vienna	193	190	198	197	190	199	194
University of Montreal	196	177	183	189	170	196	162

从表 1还可以看出,从微观的角度判定哪种评价方法更为科学合理是很困难的,因为每种方法的大学排序都不一样,但是前 22所大学都是超一流大学这一点应该是没有分歧的.

之所以会出现前 40名大学一致性较高,后 40名大学一致性较低的情况,原因有两个,一是世界大学为数众多,排名前 200名的大学实际上都可以归类为世界一流大学,这 200所大学的相关数据并不是随机抽样数据.二是前 40所大学 (超一流大学)一般发展相对均衡,不同指标数据相差较小;后 40所大学可能更有个性,不同指标间数据相差较大,导致不同评价方法评价结果差异较大.

3.2 各种评价结果与排序二维散点图

图 1~图 7是各种评价方法标准化处理后分值与排序散点图.所有图形都是一条递减的曲线,左下角是 200所大学中最好的,右上角是 200所大学中最差的,几乎所有图形的右下角点距和左上角点距均比较大,其中右下角点距最大,说明这 200所大学中,较好的大学与较差的一般都比较少,容易区分,最好的大学区分度更好,这和首尾一致率分析的结果是一致的.

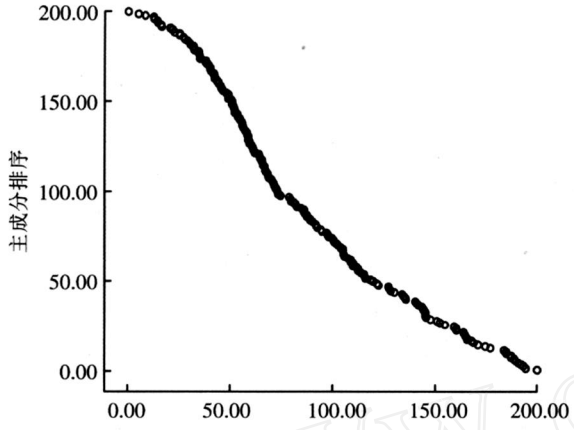

图 1 主成分分析结果与排序散点图

Fig.1 The scatter map of PCA order

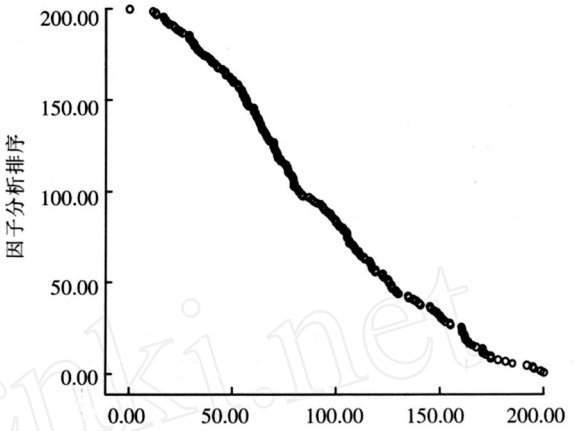

图 2 因子分析结果与排序散点图

Fig.2 The scatter map of FA order

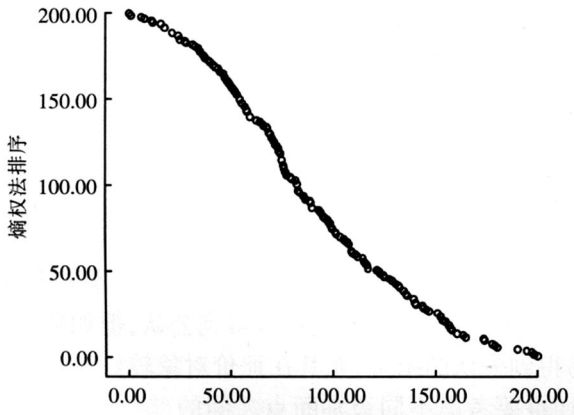

图 3 熵权法结果与排序散点图

Fig.3 The scatter map of entropy order

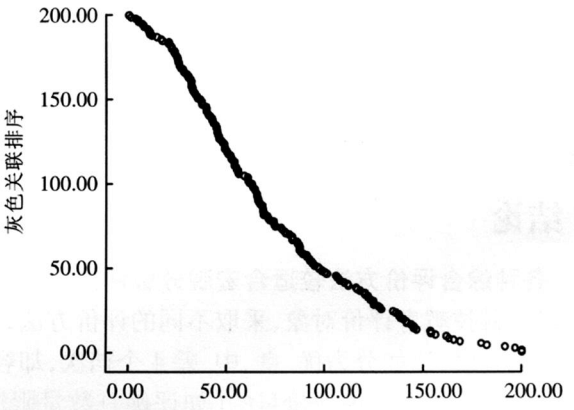

图 4 灰色关联法结果与排序散点图

Fig.4 The scatter map of GRA order

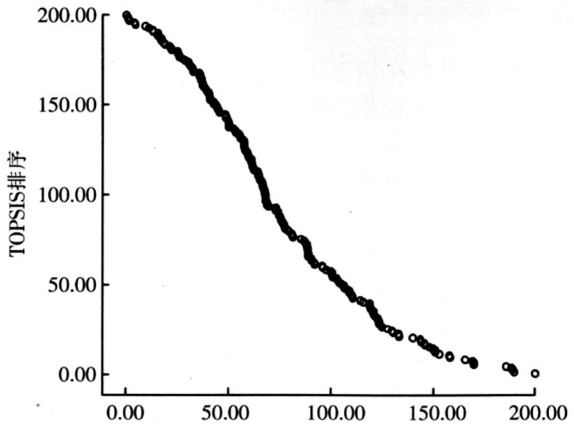

图 5 TOPSIS 法结果与排序散点图

Fig.5 The scatter map of TOPSIS order

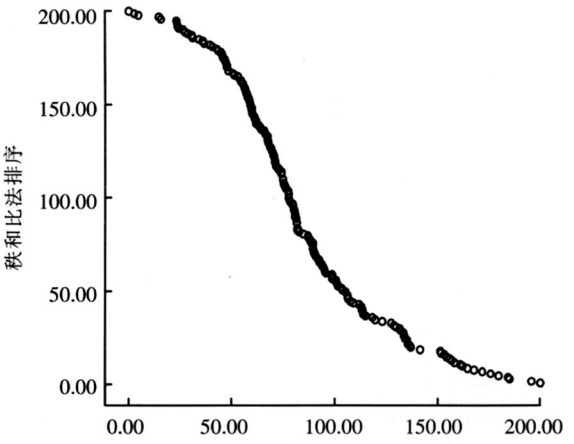

图 6 秩和比法结果与排序散点图

Fig.6 The scatter map of RSR order

有趣的是图 1~图 6均是客观赋权法,其外形非常近似,左上角上凸,右下角下凹,存在拐点;而图 7

主观赋权法总体上是下凹的,似乎客观赋权法存在某种共性规律,而主观赋权法由于是人为的,无法预知结果曲线的形状. 主客观赋权法评价二维图是否存在某种差别? 这还需要进一步进行研究.

3.3 主客观赋权法的区分度分析

区分度计算结果如表 2所示,表 2同时还给出了各种评价结果标准化后的标准差. 各种评价方法区分度由高到低依次是秩和比法、TOPSIS法、灰色关联法、主观赋权法、因子分析、主成分分析、熵权法.

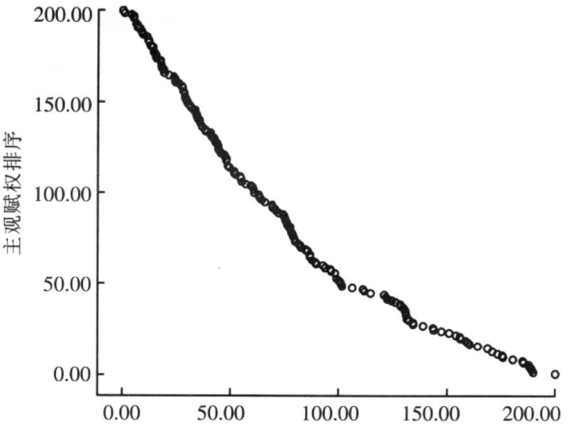

图 7　主观赋权法结果与排序散点图

Fig.7　The scatter map of Delphi

表 2　各种评价方法区分度与标准差

Table 2　The different ratio and Std Dev of different evaluation methods

统计量	主观赋权排序	主成分排序	因子排序	TOPSIS排序	秩和比排序	灰色关联排序	熵权排序
均值	72.07	87.03	91.59	76.12	84.01	70.51	88.82
最大值	200	200	200	200	200	200	200
最小值	0	0	0	0	0	0	0
标准差	51.89	49.24	47.56	43.96	39.26	45.75	45.95
标准差排序	1	2	3	6	7	5	4
区分度	1.109	1.099	1.101	1.130	1.134	1.117	1.075
区分度排序	4	6	5	2	1	3	7

由于标准差也在一定程度上反映了评价结果数据的分散程度,那么标准差能否反映区分度呢? 标准差的排序结果依次是主观赋权法、主成分分析、因子分析、熵权法、灰色关联法、TOPSIS法、稚和比法. 很显然,标准差排序与区分度没有任何关系.

4 结论

4.1 各种综合评价方法较适合宏观分级评价

对于科技教育评价对象,采取不同的评价方法,评价结果很难一致,因此难以得到公认,但如果将评价结果用来分级,比如分为优、良、中、差 4个档次,却容易得到公认的结果,尤其在评价对象较好与较差的区域. 在分级时,既要考虑评价目的 (如评优有数量限制),也要考虑不同级别断点之间的差距,断点之间差距越大,分级效果越好. 对于微观严格排序的评价,如世界 10大一流大学排名,建议采取客观赋权评价与同行评议相结合的方式. 不同评价对象首尾一致率是不一样的,首尾一致率大的评价对象说明评价对象本身数据区分度大,从而容易取得共识.

4.2 区分度是衡量不同评价方法稳定性的一个重要指标

区分度是衡量不同评价方法稳定性的一个重要指标,可以用来评价某一评价方法的可靠性. 区分度高,意味着评价结果相邻两点间的距离较大,评价排序误判的可能性较小,评价方法的灵敏度低,排序更为稳定可靠. 区分度只能用于同一评价对象不同评价方法的比较.

4.3 首尾一致率本质上也是区分度

首尾一致率本质上也是区分度,首尾一致率实际上是较优分值与较差分值的“区分度”,首尾一致率越高,在评价结果两侧的评价对象更容易区分.

[参考文献]

[1] 刘占伟,邓四二,滕弘飞. 复杂工程系统设计方案评价方法综述 [J]. 系统工程与电子技术, 2003(12): 1488-1491.

[2] 陈衍泰,陈国宏,李美娟. 综合评价方法分类及研究进展 [J]. 管理科学学报, 2004(4): 69-79.

[3] 吴清平,张丹. 秩和比法和几种常用评价方法在医疗质量评价中应用的比较 [J]. 中国医院统计, 2003(3): 3-5.

[4] Martin Ince. Ideas without borders as excellence goes global[EB/OL]. [2008-03-01]. http://www.timeshighereducation.co.uk.

[5] 田凤调. 秩和比法及其应用 [M]. 北京:中国统计出版社, 1993: 1-93.

[责任编辑:丁　蓉]