文章编号: 1000- 2243(2010)03- 0358- 05

关于评委评分的评价模型

吕书龙, 梁飞豹, 刘文丽

(福州大学数学与计算机科学学院, 福建 福州 350108)

摘要: 通过对评委评分数据的分析和研究, 构建了一套评价评委评分的数学模型. 该模型较准确地反映了各评委对评分标准的把握程度, 继而讨论了评委评分的合理性和有效性.

关键词: 评委评分; 数学模型; 评价

中图分类号: O212 文献标识码: A

An evaluation model of rater score

LÜ Shu- long, LIANG Fei- bao, LIU Wen- li

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350108)

Abstract By the research and analysis of data for the rater score, the mathematical models of evaluation for the rater bias are investigated in this paper. Our results indicate that the moderate scale for the rater, and the reasonability and effectiveness of rater score are discussed later.

Keywords rater score; mathematical model; evaluate

在各种面试、比赛场合通常都有专门评委或专家, 依据专门题集、固定流程, 对每个参评或参赛选手进行评分, 然后将所有评委的评分, 采用“截头去尾”法(即去掉一个最高分和一个最低分后)得到平均分, 作为该选手的最后得分. 这种评分机制具有应用广泛、简单实用及操作性强等特点. 众所周知, 由于各种原因, 各评委的评分存在偏差, 这些偏差主要体现在评委评分的宽严程度差异和评委本身评分的一致性上[1]. 常用的分析理论和模型有: 经典测量理论(CCT)和项目反映理论中的 Rasch模型[2]等, 应用上述理论和模型需要面试或比赛过程的细节: 如题目的难度、选手的能力参数、分数等级等[1]. 目前许多学者是从评委角度提出对评分的控制[2-4], 本研究从最终评分数据出发, 对评委评分偏差导致的各种问题进行分析, 这将更有现实意义, 也更能体现面试或比赛过程的公正、公平和公信.

1 问题的提出

表 1为所设计的评委评分数据, 从表 1的评分数据矩阵出发, 将评分中可能出现的问题, 归结为以下几个子命题:

命题 **1** 某个评委是否领会评分标准, 并进行客观的评分;

命题 **2** 某个评委的评分是否有偏差, 即是否公平合理;

命题 **3** 某个评委的评分是否具有区分性, 即是否不作为, 充当”老好人”.

表 **1** ***m*** 个评委对 ***n*** 个选手的评分及总评

Tab. 1 Rater scores and general scores by *m*- raters to *n*- candidates

选手	评委 1	评委 2	…	评委 *m*	总评
1	x_{11}	x_{12}	…	x_{1m}	$\bar{x}_1$
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nm}	$\bar{x}_n$

收稿日期: 2009- 08- 28

通讯作者: 吕书龙(1977-), 讲师; E- mail: wujispace@fzu.edu.cn

基金项目: 福建省教育厅科研资助项目(JB08027, JB08026); 福州大学研究生教育研究资助项目(09AY09)

为便于讨论, 先作出如下的假设和预处理:

假设 **1**　假设某次面试或比赛中, 有 m 个评委, n 个选手, 每个评委必须给每个选手评分, 设第 i 个评委给第 j 个选手的评分为 x_{ij}, $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots, n$; $x_{ij} \in [0, 100]$. 然后采用"截头去尾"法, 得到每个选手的最后总评分, 即表 1数据中的最后一列.

假设 **2**　各评委均经过相关的培训和测试, 且熟悉评分细则和标准, 能够做到独立、客观、公正的评分.

2　基本概念、定义及数学模型

着重针对表 1中的数据进行分析, 包括横向分析、纵向分析及交叉分析.

横向分析: 即希望对每个选手而言, 各评委的评分能够比较集中, 偏差较小, 这样能较集中地反映该选手的真实面试情况.

纵向分析: 即希望对每个评委而言, 他对每个选手的评分能够有效地区分开, 体现出较明显的高低优劣之分.

交叉分析: 即将横向分析和纵向分析有机结合, 得到一个统一的指标表征评委的评分水平.

依据表 1, 首先定义即将用到的一些概念和指标.

定义 **1**　称 $y_{ij}^2 = (x_{ij} - \bar{x}_i)^2$ 为第 j 个评委在第 i 个选手评分上的偏差平方, 而称 $y_{ij} = x_{ij} - \bar{x}_i$ 为第 j 个评委在第 i 个选手评分上的偏差.

定义 **2**　称 $W_j = \sum_{i=1}^{n} y_{ij}^2 \omega_i$ 为第 j 个评委的偏差系数, 该值越小则吻合程度越好. 其中 ω_i 为第 i 个选手的评分吻合权重, 可取 $\omega_i = \dfrac{\bar{x}_i}{\sum_{i=1}^{n} \bar{x}_i}$, 因为总评分的高低决定了选手胜出的可能性大小; 更一般的情况下, 可取权重 $\omega_i = \dfrac{1}{n}$, 则第 j 个评委的偏差系数就变为平均偏差平方和.

定义 **3**　称 $S_j^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{\cdot j})^2$ 为第 j 个评委评分的样本方差, 其中 $\bar{x}_{\cdot j} = \dfrac{1}{n} \sum_{i=1}^{n} x_{ij}$, 该样本方差越大说明该评委的评分越容易区分; 称 $SS_j = S_j^2 / W_j$ 为第 j 个评委评分的区分度, 该值的大小体现了区分性的高低. 该指标值越大说明该评委在吻合程度高的前提下具有高度的区分性; 该值越小则区分性越低.

定义 **4**　称 $S_i^2 = \dfrac{1}{m} \{ \sum_{j=1}^{m} (x_{ij} - \bar{x}_i)^2 - (\max_j \{x_{ij}\} - \bar{x}_i)^2 - (\min_j \{x_{ij}\} - \bar{x}_i)^2 \}$ 为选手 i 的有效方差; 称 $S^2 = \sum_{i=1}^{n} S_i^2 \omega_i$ 为系统评分方差, 即每个选手有效方差的加权之和.

2 1　模型 **1**— 评分控制模型

假设: 对于每个选手而言, 各个评委对其所作的评分分值近似服从正态分布 $N(\mu, \sigma^2)$. 其中 μ 对应每个选手的总评分, σ^2 对应系统评分方差. 以第 i 个选手为例, 可得如下分布: x_{ij} ($j = 1, 2, \cdots, m$) 独立同分布, 均服从 $N(\mu_i, \sigma^2)$, 其中 $\mu_i \triangleq \bar{x}_i$, $\sigma^2 \triangleq S^2$.

在上述假设前提下, 对每个选手判断每个评委评分是否在 $\mu \pm 2\sigma$ 范围内; 为直观起见, 绘制 $\mu \pm 2\sigma$ 控制图 (图 1, 其概率达到 0 9544). 图 1中, 打"×"的点对应的评委评分为异常评分.

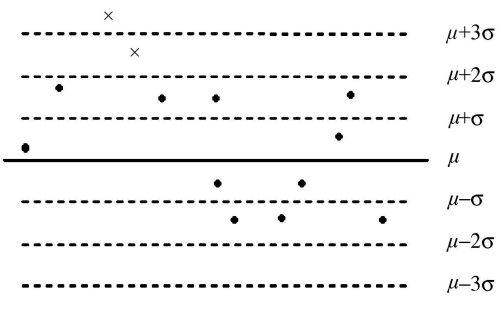

图 1　$\mu \pm 2\sigma$ 散点控制图

Fig 1　$\mu \pm 2\sigma$ Scatter control chart

对应表 1数据容易得到评分异常矩阵:

$$A = (A_{ij})_{n \times m} = \begin{cases} 0 & x_{ij} \in [\mu_i - 2\sigma, \ \mu_i + 2\sigma] \\ 1 & \text{else} \end{cases}$$

$$(i = 1, 2, \cdots, n; \ j = 1, 2, \cdots, m) \qquad (1)$$

由"截头去尾"法可知, 对于每个选手的评分数据而言,

被剔除的数据量共 2个, 比例为$\frac{2}{m}$; 假设每个评委给出会被"截去"的异常评分的概率为 p, 一般情况下有 $p \in (0, \frac{2}{m}]$. 而发生异常评分的次数 C_j可假设服从二项分布 $b(n, p)$. 由于正态分布随机变量落在 $\mu \pm 2\sigma$ 范围内的概率已大于 0.95, 所以取 $\mu \pm 2\sigma$ 为正常区域, 则得到如下的假设检验模型: H_0: 第 j个评委领会评委标准, 评分客观; H_1: 第 j个评委尚未领会评委标准 (给定显著性水平 α).

构造统计量:

$$C_j = \sum_{i=1}^{n} A_{ij} \sim b(n, p) \tag{2}$$

由 $P(C_j > b_\alpha) \leqslant \alpha$, 其中 b_α 为满足该不等式的最小的正整数, 易得接受域为 $[0, b_\alpha]$.

在假设 H_0 成立的条件下, 只要 C_j 落在上述接受域中, 即认定第 j个评委领会评分标准, 评分比较客观; 否则认定该评委对评分标准把握不够, 尚无法作出客观的评分, 建议该评委对评分标准作进一步的培训和测试. 使用该模型可以解决命题 1. 本模型可将每个评委的异常评分情况详细罗列, 对于处在临界边缘的评委, 也需要督促他们加强评分标准的熟悉和培训.

诚然, 统计每个评委的无效评分 (处在最高分或最低分) 次数也可以作为一种参考指标, 因为无效评分次数的增多必然导致异常评分次数的增多, 但是还不能得出无效评分次数多的评委就是评分标准把握不好的, 因为"截头去尾"的评分规则, 必然要去掉两个评分, 而不管其评分吻合与否.

2.2 模型 2— 偏差吻合模型

对于每个评委而言, 其对 n个选手的评分数据 x_{ij}, $i = 1, 2, \cdots, n$ 和最终评分数据 $\bar{x}_i$, $i = 1, 2, \cdots, n$ 正好为成对数据; 在假设 2的前提下, x_{ij}, $i = 1, 2, \cdots, n$ 应该是以 $\bar{x}_i$, $i = 1, 2, \cdots, n$ 为中心作随机波动. 对于标准把握较好的、评分客观公正的评委而言, 通常有 $y_{ij} \to 0$, 同时偏差系数 $W_j \to 0$. 因此对于偏差数据 $y_{ij} = x_{ij} - \bar{x}_i$, $i = 1, 2, \cdots, n$, 不妨假设偏差所服从的分布为:

$$y_{ij} \sim N(0, \sigma^2), \quad i = 1, 2, \cdots, n; \text{ 其中 } y_{ij}, \ i = 1, 2, \cdots, n \text{ 独立同分布, } \sigma^2 \text{ 未知} \tag{3}$$

因此可构造 σ^2 未知的关于期望为 0的假设检验模型[5].

为统一处理, 依据表 1数据和偏差定义, 得到偏差矩阵:

$$B = (B_{ij})_{n\times m} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{mn} \end{pmatrix} \tag{4}$$

作如下的假设检验模型: H_0 为第 j个评委无偏差; H_1 为第 j个评委偏差显著 (给定显著性水平 α).

构造统计量:

$$t_j = \frac{\bar{y}_j - 0}{\hat{S}_{yj}/\sqrt{n}} \sim t(n-1), \text{ 其中 } \bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \ \hat{S}_{yj} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2} \tag{5}$$

查 t分布表, 得 $t_{1-\alpha/2}(n-1)$ 和 $t_{\alpha/2}(n-1)$, 代入矩阵 (4) 中的数据, 可得 t_j的值. 当 $t_j < t_{1-\alpha/2}(n-1)$ 或 $t_j > t_{\alpha/2}(n-1)$ 时拒绝 H_0, 认为该评委的评分存在较大偏差; 否则接受 H_0, 认为该评委的评分比较合理、客观. 使用该模型可以解决命题 1.

如果 $t_j < t_{1-\alpha/2}(n-1)$ 成立, 则可判定该评委有给低分的倾向; 反之, 若 $t_j > t_{\alpha/2}(n-1)$ 成立, 则可判定该评委有给高分的倾向, 这些都反映了评委对评分标准的领会程度不够, 从而导致宽严不一, 得出的评分产生较大的偏差. 利用该模型可基本解决命题 2. 对于评分偏差较大、整体偏低或整体偏高的评委, 有待重新熟悉评分标准; 对于处于临界边缘的评委有待进一步熟悉评分标准. 对检验值的使用主要集中在处于临界值附近和拒绝域中.

有个问题需要特别说明一下, 是否检验值 $t_j \to 0$就能说明评委无偏差呢?答案是否定的. 这里有两种情况:

1) 如果评委 j的评分落在总评分的较小领域中, 则有 $\bar{y}_j \to 0$, 从而 $t_j \to 0$, 这是所希望的.

2)如果评委 j的评分较均匀地落在总评分两侧较远处, 正负相互抵消后也会导致 $\bar{y}_j \to 0$, 从而 $t_j \to 0$, 而这是不希望看到的.

显然单纯地依赖 t_j做出评委偏差程度大小的评价是不严谨也是不合适的. 为此, 需要计算每个评委的偏差系数 W_j, $j = 1, 2, \cdots, m$, 以观察评委评分的整体偏差情况. 由偏差系数的定义及假设 2可知, 偏差系数在正常情况下应该较小, 出现较大偏差的可能性很小. 如果评委 j出现较大的偏差系数则可判定该评委对评分标准的把握程度肯定存在较大偏差. 所以, 将 t_j和 W_j综合起来, 就可以较全面地分析评委把握标准的程度, 才能对评委的偏差程度做出合理的解释. 具体的分析将在模拟试验中说明.

2.3　模型 3—区分度模型

研究每个评委的自身一致性问题[1], 即该评委所作评分除了满足吻合度高外, 同时也应该体现出高区分度, 避免出现"老好人"的评分现象. 当然, 如果所有选手的水平相当, 则出现高区分度的可能性将会大大降低. 由区分度的定义:

$$SS_j = S_j^2 / W_j \quad (j = 1, 2, \cdots, m) \tag{6}$$

其中: S_j^2体现评委 j自身评分的离散程度; 而 W_j体现评委 j评分的偏差程度. 这样 SS_j就可以在满足吻合程度的前提下, 体现评委 j的区分能力. 如果某个评委的评分具有"老好人"现象, 即他对所有选手的评分比较集中在某个值附近, 则必然导致 S_j^2偏小, 而 W_j偏大, 最终导致区分度 SS_j的值将相对特别小. 通过区分度值大小的排序, 可以对评委评分进行排队, 确定哪些评委在评分上有不作为行为. 这样就可以解决命题 3. 具体的分析将在模拟试验中说明.

3　模拟试验

为检验模型的鉴别能力, 采用随机模拟的方法进行验证. 取评委个数 $m = 9$, 选手个数 $n = 10$, 基于正态分布随机模拟产生了 90个评分数据. 另外, 为了测试模型的鉴别能力, 将评委 1、2、3、4的数据作了异常处理(评委 1偏低, 评委 2微调偏高, 评委 3微调有高有低, 评委 4评分较集中, 改动的数据用粗斜体表示)如表 2所示.

表 2　模拟评委评分数据(保留 1位小数)

Tab 2　Simulation data of the rater score (reservations a decimal)

选手	评委									总评
	1	2	3	4	5	6	7	8	9	
1	***74.9***	***83.0***	***82.9***	81.3	78.4	80.7	78.8	80.5	79.5	80.3
2	72.3	***75.2***	72.8	***80.5***	70.2	72.3	73.0	71.4	74.1	73.0
3	***88.3***	94.6	***89.8***	***84.9***	96.2	94.1	92.9	95.6	95.1	92.9
4	79.9	79.3	80.7	81.6	81.9	81.5	80.9	79.2	81.8	80.8
5	70.7	***78.5***	***79.1***	***83.2***	72.2	74.6	75.2	73.5	72.5	75.1
6	73.8	74.2	75.8	81.3	75.1	73.9	73.8	74.4	74.8	74.6
7	78.3	76.7	***73.8***	78.7	77.5	76.6	78.3	77.0	77.3	77.4
8	***85.6***	93.0	93.6	***81.2***	93.0	93.8	93.2	93.7	91.7	92.0
9	90.8	89.5	90.0	***81.7***	91.5	88.1	89.9	89.6	89.3	89.6
10	***80.3***	***86.2***	83.1	81.3	82.6	83.2	83.1	84.9	83.0	83.0

表 3　模型 1计算得到的无效评分次数和异常评分次数, 取 $p = 2/m$

Tab. 3　Invalid times and abnormal times of score be calculated in model 1($p = 2/m$)

评分次数	评委								
	1	2	3	4	5	6	7	8	9
无效评分次数	4	2	1	7	4	1	1	1	0
异常评分次数	4	2	3	6	1	0	0	0	0

注: 选手 6的 1评和 7评都是最低分, 所以计算无效评分次数时都增加 1次.

由模型 1计算得到表 3, 且其接受域是[0, 4]. 由表 3可知, 评委 4的评分标准掌握不好; 而评委 1

处于临界值上, 需要进一步加强评分标准的熟悉. 由于评委 2和 3只做了微调, 虽然未出现异常超标, 但也需要加强培训.

表 4 模型 2计算得到的 T 检验值和偏差系数

Tab 4 T – test value and the deviation coefficient be calculated in model 2

计算结果	评委								
	1	2	3	4	5	6	7	8	9
T 检验值 t_j	– 2. 79	2. 09	0. 40	– 0. 14	– 0. 01	0. 04	0. 17	0. 23	0. 10
偏差系数 W_j	12. 78	3. 96	4. 92	42. 94	3. 73	1. 01	0. 54	2. 25	1. 44

由模型 2计算得到表 4, 且其接受域是 (– 2. 262 157, 2. 262 157). 由表 4知, 评委 1不在接受域, 左偏较大, 说明给分太严了; 评委 2虽然在接受域中, 但 T 检验值相对偏大, 说明给分较宽松.

从偏差系数看, 评委 1和评委 4的偏差很大, 明显看出他们对标准把握不够, 给分不客观.

本模型将两个指标综合起来看更能说明问题: 评委 1给分较严同时偏差较大; 评委 2给分偏松, 偏差稍高; 评委 3给分略松, 偏差稍高; 而评委 4虽然检验值很小, 但其偏差系数极大, 正好符合上述分析中的情况 2. 因此对于评委 1, 2, 3, 4都需要进一步加强评委标准的熟悉和理解, 而且最好是采取一些有针对性的模拟培训, 特别是评委 4.

表 5 模型 3计算得到的区分度数据

Tab 5 Differentiation data be calculated in model 3

计算结果	评委								
	1	2	3	4	5	6	7	8	9
$S^2_{\cdot j}$	47. 50	55. 44	51. 29	2. 61	81. 02	63. 30	58. 60	73. 94	61. 42
W_j	12. 78	3. 96	4. 92	42. 94	3. 73	1. 01	0. 54	2. 25	1. 44
SS_j	3. 72	14. 0	10. 42	0. 06	21. 72	62. 67	108. 52	32. 86	42. 65

由模型 3计算得到表 5, 从表 5中可以看出, 评委 4的评分确实出现“老好人”的现象, 评分老是集中在某个值附近, 没有体现出应有的区分度; 导致评委自身评分的方差很小, 而偏差系数很大, 所以比值极小.

对 SS_j 值从小到大排序可知, 评委 4, 1、3, 2都是靠前的, 有“老好人”之嫌, 这和前面两个模型的诊断结果也是吻合的.

4 结语

1)将评委评分可能存在的问题命题化, 而且仅针对最终的一个评分数据矩阵, 采用参数和非参数方法并结合假设检验理论, 构造了评分控制模型、偏差吻合模型和区分度模型, 并对模拟的数据进行有效的计算和验证, 从计算结果看出, 模型能较好地解决文中所列的前 3个命题. 所构建模型很容易转化成计算机程序, 可操作性强.

2)模型 2和模型 3构建的指标, 能够诊断评委评分的部分问题. 若能制定出一个更精确的标准, 将更加有效.

3)模型仅针对最终的评分数据矩阵进行分析, 如果有评分细则及相关评委的评分细节, 模型可以做得更加精细, 评价会更加精确. 这将是进一步研究的方向.

参考文献:

[1] 孙晓敏, 张厚粲. 国家公务员结构化面试中评委偏差的 IRT分析 [J]. 心理学报, 2006, 38(4): 614– 625.
[2] 孙晓敏, 薛 刚. 多面 Rasch模型在结构化面试中的应用 [J]. 心理学报, 2008, 40 (9): 1 030– 1 040.
[3] 陈 媛, 樊治平. 综合性面试中的面试官分组方法 [J]. 系统工程, 2008, 26(1): 96– 101.
[4] 元继学. 专家主观评分比赛中群决策机制的研究 [J]. 中国软科学, 2009 (2): 173– 176.
[5] 梁飞豹, 徐荣聪, 刘文丽. 概率论与数理统计 [M]. 北京: 北京大学出版社, 2005.

(责任编辑: 郑美莺)