DOI: 10.13656/j.cnki.gxkx.2008.03.019
广西科学 Guangxi Sciences 2008, 15(3): 266~ 268, 273

大型竞赛的一种评卷模型*

A Grading Model of Large-scaled Competition

刘星子,林 亮,臧东冉
LIU Xing-zi, LIN Liang, ZANG Dong-ran

(桂林工学院数理系,广西桂林 541004)
(Mathematics and Physics Department, Guilin University of Technology, Guilin, Guangxi, 541004, China)

摘要:针对大型竞赛评卷过程中存在的不公平性问题,提出一种有效控制误差的评卷模型,并用该模型对2006年全国大学生数模竞赛广西赛区的阅卷实例进行验证.结果显示,新模型比传统的评分模型可以更好地消除误差,更合理地反映出参赛队伍的实际水平,具有较强的适应性.

关键词:评卷模型 一致性检验 D-检验 方差分析 T分数
中图法分类号:O213 文献标识码:A 文章编号:1005-9164(2008)03-0266-03

Abstract To solve the unfair grading problem of large-scaled competition, an effective grading model of controlling errors is presented. Then, the computation on the grading example of the CMCM-2006 contest in guangxi is illustrated to show the effectiveness of the model. The result shows that the new model may eliminate error better than traditional grading models. Moreover, it can reflect reasonably the actual level of the teams, and has strong compatibility.

Key words grading model, consistency check, D-test, analysis of variance, T mark

大型竞赛的目的不仅在于考察考生对知识的掌握程度和应用能力,而且还带有挑选、淘汰的性质.这类竞赛就必然要求做到公平公正.在现有的考试制度下,考前、施考两个环节的公平已经基本得到实现,但是在最后评卷这一环节中公平性还有待完善.特别像全国大学生数学建模(CMCM)这类大型竞赛,由于题目的灵活性和参赛学生的多样性,使得答案多种多样,评委在评卷时对评分标准的尺度也就难以把握,对考生的评分就不可避免地存在误差.产生评卷误差是客观存在的,只能控制,不能消灭.对于控制评卷误差,有关专家学者做了大量的研究,采取过很多措施,例如网上评卷模式、基于神经网络技术的评卷模式等[1~3].虽然这些措施取得了一些成效,但是由于时间、方式等因素,这些措施无法在实践中方便地实施和有效地控制误差.这个问题得不到解决,竞赛就不能在真正意义上说公平公正.

为了更好地解决这一难题,本文提出一种新的评卷模型,使各评委的评分在置信区间内趋于一致,最后分别应用SPSS、Excel软件结合新模型对全国大学生数模竞赛(CMCM)-2006广西赛区阅卷实例进行验证,并给出具体的解决方案.新模型可以比较方便地在应用中取得满意结果.

1 问题分析

1.1 问题假设

假设:(i)总共有N支参赛队伍,M位评委;(ii)评委独立工作,互不干扰而且所有评委的阅卷量相同;(iii)一般批阅一份试卷的评委不会太多,而且评委的工作能力达到一定的水准.

1.2 百分制和等级制的选取

因为大型竞赛一般难以给出具体公平的评分标准,评委阅卷中对一份试卷的界定是定性思维与定量思维相结合的过程,而且人的第一感觉往往很难改变.故一般认为,等级制比百分制简洁而且容易界定,并且前者数据处理复杂度低,是首选的做法.但是如果想将评分尽量做到公平完善时,百分制就显示出不可忽略的优势.首先百分制和等级制原本是一样的,

收稿日期:2007-11-14
修回日期:2008-01-07
作者简介:刘星子(1982-),女,硕士研究生,主要从事应用统计研究工作.
* 国家自然科学基金项目(10661006),广西"新世纪十百千人才工程"专项资金项目(2005214)及广西自然科学基金项目(桂科自0728212),桂林工学院课外学术科技作品项目资助.

只是精确度不同,百分制就是把5分或10分一等级精确到每1分一等级;其次采用百分制可将评委给分转化成标准分,通过对评分的调整使得答卷分数趋于一致,并且构造判断矩阵做一致性检验,从而进一步减少尺度误差和不公平现象;再次数据精确度是随公平性等级和尺度偏差等级数增加而增加的.故本文采用百分制.

1.3 一致性检验

大量统计资料表明[4-6],大型选拔性竞赛,考生总体成绩合理有效的分布应该呈对称正态分布或正偏态分布,即呈其他分布如双峰形分布、平坡型分布等是不合理的.故假设:(Ⅰ)任意分配若干份试卷给某评委,这位评委的给分情况服从正态分布;(Ⅱ)对于同一份试卷,不同评委批阅,其分数也应呈正态分布.所以判断各位评委所给的分数是否一致和公平应该包含两层含义:(ⅰ)对各评委所给的分数是否服从正态分布做总体正态性检验,如果某位评委的给分中既有多数偏高分又有多数偏低分,则认为该评委的评卷存在问题,是不公平的,应该将该评委的评卷分数都去掉;(ⅱ)对于同一份试卷,不同评委批阅的分数也应该具有一致性.但是因经济性及时效性约束,一般批阅一份试卷的评委不会太多.对同一份试卷不同评委批阅,只要分数差距不大,进行一致性检验就无太多意义,故一般情况所进行的一致性假设检验仅仅是针对假设(Ⅰ).

统计学将数据分布的不对称性称作偏态.判别偏态的方向并不困难,利用众数、中位数和均值之间的关系就能判断分布是左偏还是右偏,但是要测度偏斜的程度就需要计算偏态系数.偏态系数 SK 是对数据分布对称性的测度值,当 $SK=0$ 时频数分布对称;峰度是数据分布的扁平或尖峰程度,峰度系数 K 是数据分布峰度的度量值,当 $K=0$ 时为正态峰[7]. SK、K 计算公式:

$$SK=\frac{n\sum(x-\bar{x})^3}{(n-1)(n-2)\left[\sum(x-\bar{x})^2/(n-1)\right]^{3/2}},\quad(1)$$

$$K=\frac{n(n+1)\sum(x-\bar{x})^4}{(n-1)(n-2)(n-3)\left[\sum(x-\bar{x})^2/(n-1)\right]^2}-\frac{3(n-1)^2}{(n-2)(n-3)},\quad(2)$$

其中,x 为样本值,$\bar{x}$ 为均值,n 为样本数.

2 模型建立

2.1 分数总体检验

结合偏度、峰度检验法,先分析分布的偏斜方向、偏斜程度和集中趋势程度,如果分析结果不是特别理想还应该对总体进行正态检验.因为存在分组不同导致拟合结果可能不同而且还需要有足够的样本容量等问题,本文不选用检验正态总体一般采用的检验法,而采用W法和D法.W法、D法检验具体步骤[8]分别为:

W法:对称序号(i)→ a_{in} → X_{n-i+1} → X_i → $d_i=X_{n-i+1}-X_i$ → $a_{in}d_i$ → $W_j=\left[\sum a_{in}d_i\right]^2\Big/\sum(X_j-\bar{X})^2$.

D法:分数→实际得分频率 f_i → 实际频率分布累积频数 cf_i → 实际频率分布累积概率 cp_i → 得分标准化值 Z_i → 理论频率分布累积概率 ecp_i → $D_i=|cp_i-ecp_i|$ → $D'_i=|cp_{i-1}-ecp_i|$, $(cp_0=0)$ → $D=\max[\max(D_i),\max(D'_i)]$.

2.2 分数调整步骤

当确定各评委的评分分布为正态分布或近似于正态分布时,这还不足以将各评委评分汇总作为各参赛队的最后成绩.为了公平还需要对数据作调整:

步骤1 各评委评分的幅度不等,应该对各评委的评分进行标准化处理,标准化值计算公式为:

$$Z_{ji}=\frac{x_{ji}-\bar{x}_j}{S_{n_j-1}}\quad(3)$$

(x_{ji} 为第 j 位评委批阅的第 i 份试卷得分,$\bar{x}_j$ 为第 j 位评委所批阅的试卷得分均值,S_{n_j-1} 为第 j 位评委评分的标准差).

步骤2 称标准化得到的数为 Z 分数.通常 Z 分数会出现负数和小数点,使用起来很不方便,因此可以对 Z 分数进一步加以线性转换,使之成为正的数值.最典型的一种 Z 分数转换就是 T 分数.T 分数计算公式为

$$T_{ji}=10Z_{ji}+\bar{x}_j.\quad(4)$$

步骤3 由于各评委评卷尺度可能存在差异,假设5位评委中3位评委阅卷尺度基本一致,而且另外2位评委一个给分普遍偏高一个给分普遍偏低.现在2支参赛队伍存在较小差异,若较好的一支队伍由2位尺度一致的评委和1位给分普遍偏低的评委批阅,而较差的一支队伍由2位尺度一致的评委和一位给分普遍偏高的评委批阅.将这2支队伍的3个得分简单的直接相加,可能得出与实际水平相反的结论,所以应该采用方差分析对各评委的均值是否相等进行检验.如果不相等则说明存在系统误差需要进行均值调整,即将各评委的均值调整为整体均值,以此来消除系统误差.

2.3 分数汇总

经过一系列的调整之后,计算可得第 n 支参赛队伍最终得分为:

$$x_n = \frac{\text{批阅该试卷的各评委评分之和}}{\text{批阅该试卷的评委数}}. \quad (5)$$

如果情况要求对假设(Ⅱ)也要进行检验,具体检验过程和假设(Ⅰ)类似.另外本文提出的方法还能够在评委阅卷过程中,使评委的评分得到及时反馈,从而合理改进评分方案.

3 实例分析

对 CMCM-2006广西赛区的评卷进行实例验证.总共有 90支来自广西各高校的参赛小组,5位来自不同高校的评委,每位评委评阅 54份答卷,要求回避本校答卷,每支小组由 3位评委批阅.

为了便于阐述,用 M_j (j = 1, 2, 3, 4, 5) 表示这 5位评委,具体分析如下.

3.1 均值和标准差

由 SPSS[9] 软件得出各评委评分均值为: $\bar{x}_1$ = 67.952, $\bar{x}_2$ = 68.413, $\bar{x}_3$ = 78.935, $\bar{x}_4$ = 67.574, $\bar{x}_5$ = 70.667; 各评委评分标准差为: S_1 = 10.100, S_2 = 9.131, S_3 = 6.524, S_4 = 9.739, S_5 = 13.228.

3.2 偏度检验、峰度检验

将各评委的评分、均值以及阅卷的份数代入公式(1)和(2),计算得出 5为评委评分的 SK 依次为 -0.582, 0.215, -1.112, -0.143, 0.038; K 依次为 -0.533, -0.636, 0.929, -0.880, -0.920.

3.3 W法、D法检验

因为各评委的阅卷份数均大于 50,故采用 D检验.计算得出 5位评委的评分 D值依次为 0.099, 0.104, 0.128, 0.119, 0.082.当 α = 0.05, n_j = 54时,查表得到 D_{α,n_j} = 0.182.因为这 5个数均小于 D_{α,n_j},则均不拒绝 H_0,即各评委给分在置信区间 α = 0.05内服从正态分布.

3.4 T分数

首先将各评委的原始评分依次代入公式(3),求得 Z 分数,再把 Z 分数代入到公式(4),求出 T 分数.

3.5 方差分析 ANOVA

用 SPSS作方差分析如表 1.

表 1 方差分析

Table 1 ANOVA

变异来源 Origin of variation	离差平方和 Sum of squares	df	均方差 Mean square	F	$Sig.$
组间 Between groups	4880.183	4	1220.046	12.200	.000
组内 Within groups	26499.992	265	100.000		
总计 Total	31380.175	269			

表 1中 F分布的显著性概率 Sig小于系统默认的显著性概率 0.05,因此拒绝原假设,即认为这 5位评委评分均值有显著性差异,故需要进行均值调整.

3.6 消除系统误差

计算得出总体均值 $\bar{x}$ 为: 70.708,而各评委的均值 $\bar{x}_j$ 分别为: 67.952 68.413 78.935 67.574 70.667,则只需要将由各评委所评评分计算得到的 T 分数分别加上 2.756 2.295 -8.227 3.134 0.041.

3.7 分数加总

将各参赛队伍对应的各 T分数代入到公式(5),便可以计算得出各参赛队伍的最终成绩.

把经过新模型一系列检验和调整后的成绩与简单的将名次或评分相加后的成绩比较,结果见表 2.

表 2 各排序法排序结果对比

Table 2 The result contrast of each ranking method

编号 No.	标准化模型 Standardized model		简单名次加总 Simple rank addition		简单得分加总平均 Simple average of marks addition	
	T分数 T mark	名次 Rank	各名次之和 Sum of each rank	名次 Rank	均值 Mean	名次 Rank
N32	84.03	1	22	3	82.5	5
N45	83.83	2	18	1	82.83	4
N84	82.85	3	25	6	82.33	6
N74	82.71	4	22	4	85.63	1
N87	82.25	5	26	7	81.77	8

表 2中编号为 N32的参赛队,若将评分结果按照各评委评分的名次简单相加,则排名第 3;若将评分简单相加后排序,其排名第 5;而采用本文提出的方法,其排名第 1.由此可以看出本文的评分方法比传统的评分方法在一定程度上消除了评卷误差,更合理反映出参赛队伍的实际水平.

4 结束语

经过理论分析和 CMCM-2006广西赛区的阅卷工作的实例验证,可以看出,本文提出的评分模型在把握评分标准,保证阅卷稳定性和公平性方面起着核心的作用.在大型竞赛阅卷过程中,由于每位评委都有自己的主观能动性,要想评委在整个评卷过程中始终掌握评分标准的一致性和评卷的稳定性是件很难做到的事情.控制评卷误差,要注意积累评卷经验,要坚持评卷工作的科学性、公平性、准确性等原则,要组织评委认真研究评分标准和考生答卷,从而使评委评卷有依据,做到严格掌握评分标准.本文的模型还可以应用于其它一些竞赛性评卷工作中,同样具有较强的适用性.

(下转第 273页 Continue on page 273)

$$\psi(u)\leqslant e^{-R^{*}u}, \qquad (2.39)$$

where $R^{*}=R_1$ and $u\in\{0,1,2,\cdots\}$.

Proof We first prove that $\{e^{-R^{*}U_k}, k\in N^{+}\}$ corresponds to a supermartingale.

Letting $\bar{Y}_{k-1}=(Y_1=y_1,\cdots,Y_{k-1}=y_{k-1})$, it follows that $\bar{Y}_{k-1}$ summarize all relevant information about the surplus process during the $k-1$ first periods. We have

$E(e^{-R^{*}U_k}\mid I_0=i,\bar{Y}_{k-1})=E(e^{-R^{*}(U_{k-1}+Z_k-Y_k)}\mid I_0=i,$
$\bar{Y}_{k-1})=e^{-R^{*}U_{k-1}}E(e^{-R^{*}(Z_k-Y_k)}\mid I_0=i,\bar{Y}_{k-1})=$
$e^{-R^{*}U_{k-1}}[\delta E(e^{-r(1-Y_k)}\mid I_0=i,\bar{Y}_{k-1})+$
$(1-\delta)E(e^{rY_k}\mid I_0=i,\bar{Y}_{k-1})]=e^{-R^{*}U_{k-1}}[\delta E(e^{-r(1-Y_k)}\mid$
$I_{k-1}=i_{k-1})+(1-\delta)E(e^{rY_k}\mid I_{k-1}=i_{k-1})]$.

Since $R^{*}=R_1$, we have

$[\delta E(e^{-R^{*}(1-Y_k)}\mid I_{k-1}=1)+(1-\delta)E(e^{R^{*}Y_k}\mid I_{k-1}=$
$1)]=1$,

and from formulae(2.36), we obtain that

$\delta E(e^{-R^{*}(1-Y_k)}\mid I_{k-1}=0)+(1-\delta)E(e^{R^{*}Y_k}\mid I_{k-1}=$
$0)\leqslant\delta E(e^{-R^{*}(1-Y_k)}\mid I_{k-1}=1)+(1-\delta)E(e^{R^{*}Y_k}\mid I_{k-1}=$
$1)=1$.

Then, it follows that

$E(e^{-R^{*}U_k}\mid I_0=i,\bar{Y}_{k-1})=e^{-R^{*}U_{k-1}}[\delta E(e^{-r(1-Y_k)}\mid$
$I_{k-1}=i_{k-1})+(1-\delta)E(e^{rY_k}\mid I_{k-1}=i_{k-1})]\leqslant e^{-R^{*}U_{k-1}}$.

From the Kolmogorov's inequality for positive supermartingales, one can find that

$P(\max_{k\in\{0,1,\cdots\}}\{e^{-R^{*}U_k}\geqslant 1\mid I_0=i_0)\leqslant e^{-R^{*}u}, u\in\{0,1,$
$\cdots\}$.

Since

$\psi(u\mid i_0)=P(\min_{u\in\{0,1,\cdots\}}\{U_k\}<0\mid I_0=i_0)\leqslant$
$P(\max_{k\in\{0,1,\cdots\}}\{e^{-R^{*}U_k}\geqslant 1\}\mid I_0=i_0)\leqslant e^{-R^{*}u}$,

we obtain

$$\psi(u\mid i_0)\leqslant e^{-R^{*}u}, u\in N, \qquad (2.40)$$

From formulae(2.40), the nonconditional ruin probability must satisfy the following inequality

$$\psi(u)=(1-q)\psi(u\mid 0)+q\psi(u\mid 1)\leqslant e^{-R^{*}u}.$$

References

[1] Gerber H U. Mathematical fun with the compound binomial process[J]. ASTIN Bulletin, 1998, 18: 161-168.

[2] Willmot G E. Ruin probabitilies in the compound binomial model [J]. Insurance: Mathematics and Economics12, 1993, 12: 133-142.

[3] Dickson D C M. Some comments on the compound binomial model[J]. Astin Bulletin, 1994, 24: 33-45.

[4] De Vylder F, Marceau E. Classical numerical ruin probabilities[J]. Scandination Actuarial Journal, 1996, 2: 109-123.

[5] Cheng S, Zhu R. The asymptotic formulas and Lundberg upper bound in fully discrete risk model[J]. Applied Mathematics: A Journal of Chinese Universities Series A, 2001, 16(3): 348-358.

[6] Pavlova K P, Willmot G E. The discrete stationary renewal risk model and the Gerber-Shiu discounted penalty function [J]. Insurance: Mathematics and Economics, 2004, 35: 267-277.

[7] Yuen K C, Guo J Y. Ruin probabilities for time-correlated claims in the compound binomial model[J]. Insurance: Mathematics and Economics, 2001, 35: 47-57.

[8] Xiao Y T, Guo J Y. The compound binomial risk model with time-correlated claims[J]. Insurance: Mathematics and Economics, 2007, 41: 124-133.

[9] Cossette H, Landriault D, Marceau E. Ruin probabilities in the compound Markov biomial model[J]. Scandination Actuarial Journal, 2003, 4: 301-323.

[10] Cossette H, Landriault D, Marceau E. Exact expressions and upper bound for ruin probabilities in the compound Markov biomial model[J]. Insurance: Mathematics and Economics, 2004, 34: 449-466.

(责任编辑: 尹 闯)

(上接第268页 Continue from page 268)

参考文献:

[1] 张昌应.网上评卷误差控制的方法与实施[J].高教探索, 2003(3): 77-79.

[2] 丁文,杨卫东,刘继来.基于神经网络技术的评卷误差控制模型及其应用[J].浙江工业大学学报, 2003, 31(4): 419-423.

[3] 高爱国.利用数学建模评阅竞赛试卷[J].高师理科学刊, 2004, 24(1): 8-11.

[4] 李志学.建立公平绩效评价的分值转换模型研究[J].中国管理科学, 2005, 13(10): 126-130.

[5] 张厚粲,刘昕.考试改革与标准参照测验[M].辽宁:辽宁教育出版社, 1992.

[6] 刘应成.考试系统中成绩正态分布检验的设计与实现[J].重庆工学院学报, 2004, 18(6): 188-191.

[7] 袁卫,庞皓,曾五一,等.统计学[M].北京:高等教育出版社, 2004.

[8] 魏宗舒.概率论与数理统计教程[M].北京:高等教育出版社, 2004.

[9] 卢纹岱. SPSS for Windows统计分析[M].北京:电子工业出版社, 2005.

(责任编辑: 尹 闯)