ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research, volume 177
International Conference on Education Reform and Management Science (ERMS 2018)

An Example of Application of Statistical Analysis in Teaching Quality Evaluation

Shi-dong LI[1] and Zhi-fang CAI[2]

[1]Shenzhen Polytechnic, Xi Li Lake, Nanshan, Shenzhen, China 518055

[2]Nanshan Second Experimental School, Oversea Chinese Town, Nanshan, Shenzhen, China 518053

Keywords: Exam results, Teaching, Quality, Evaluation, Correlation, Variance, Statistics.

Abstract. A widely used teaching quality evaluation method is based on comparison of the mean values of the student's exam results. However, if the students with great diversity are not grouped randomly, this method may lead to unfair and incorrect results. As an example, this paper studied data from a primary school to discuss the issue and a set of new teaching quality evaluation methods based on statistical analysis was presented. And conclusions have been made that most of the mean values of exam results among different classes inside the school have no significant statistic difference, and even if some difference exits, the main contribution to that difference is mainly from the non-random distribution of students rather than the difference of teaching quality among teachers. The analysis method used in this paper can be applied in all similar teaching evaluation cases.

数理统计分析在教学质量评估中的应用实例

李时东[1], 蔡芝芳[2]

[1]深圳职业技术学院，西丽湖，深圳，中国

[2]南山第二实验学校，华侨城，深圳，中国

关键词：考试成绩；教学质量；测评；评价；评估；相关分析；方差分析；统计

摘要：学生考试成绩平均分数排名往往是人们进行教学评价的主要数据，但如果差异很大的生源，其分配（抽样）不是随机的，可能导致不正确的教学评价结果。本文以某小学一年级的成绩为例，探讨了如何用相关分析和方差分析方法来科学分析考试成绩，并评估了差异化生源对教学评价的影响程度。文章经分析得出结论，班级间的平均成绩差别，在该案例中并不总是有统计学意义的显著性，因而不应简单以此判断教师或学校优劣。文章同时指出，该案例中，即便我们认为平均分差异存在，造成此差异的影响因素中，生源差异因素要远远大于教师教学差异因素的影响。本文所采用的评估方法，具有通用性，也可方便地推广应该用到所有其他情形。

1. 引言

评价学校或教师时，考试成绩虽不能作为教学评价唯一指标，但作为一个检验教学效果的定量评判依据，依然有其它方法不可比拟的客观、严谨的优势。其中平均成绩横向排名，作为一个最容易获得、最大众化的统计指标，成为相当一部分家长和学生，选择学校或选择班级时，最看重的指标，甚至也成为部分学校内部考核教师绩效时的一个重要指标。

但问题是以平均成绩的横向排名，来评价甚至选择学校和教师，真的是很公平很科学吗？

显然，直觉上我们知道，对学校或对教师的教学评价指标，如果以学生考试平均分数排名作为主要数据，对生源相对较差的非重点学校或非重点班级教师是不公平的，即学生平均

成绩的简单排名其实并不能反映教师或学校教学水平和能力的排名。但这样的直觉，有没有或能不能以数理统计的分析结果，给出一个定量的，令人信服的证明？差异化较大的生源若因分班或择校而非随机分布，多大程度上会影响教学评估的公平性与准确性？

虽然许多文献，比如文献[1][2][3][4][5]给出了各种考试成绩的分析，或教师教学评价的各种方法，但都没有给出一个简单可行的，能排除生源差异影响的考试成绩分析方法，也没有拿出有力的统计学依据来证明用平均分来评判学校或教师，可能造成不公平不科学这一事实。文献[5]和文献[2]介绍了国外流行的“增值测量法”，非常有价值，但依然没有评估生源差异对教学评价会产生多大程度的影响。

通过给出的实例，本文介绍了一套非常简易的方法，运用了多种常用的统计分析工具，而且还给出直观的效果图示，评估了生源差异对教师及教学评价影响的程度。教师及各级教学部门，无须具备复杂的数学知识，即可模仿本文，以数理统计方法来对成绩进行分析，进而完成教学评价，并使教学评价结果更加公正和科学。

2．原始数据的背景与现有评价方法存在的问题

我们获得了某小学一年级的语文数学成绩，由小学所在的区教育局统一命题考试，实行教学、考试、批卷分离。该年级7个班级有数学教师四名，分别教一二班、三四、五六班、七班，而语文教师则各班配一名不同的教师。按照现有方法可以排定班级名次（见表1）。在成绩公布后，语文数学平均分靠后的任课教师，立刻就感受到了某种压力。

人们通常将此排名结果就当成对教师进行绩效考核的重要依据了，但这样是否科学合理？首先，从表1直观看到，语文成绩排名在前的，数学成绩也排名前列。在语文数学不是同一个教师的情况下，有理由猜测生源因素对成绩的影响存在，但也不可否认教师水平因素的存在，问题是那一个是造成班级间平均分差异的主要因素？其次，班级间的平均分高低确有不同，这仅仅是普通的随机误差，还是有显著的统计学差异？其问题的实质，就在于生源有差异的情况下如何进行教学评价，并以此指导教学工作？下面我们将围绕这些问题展开讨论。

表1 各班平均成绩（原始数据汇总）

	1班(51人)		2班(49人)		3班(52人)		4班(50人)		5班(50人)		6班(50人)		7班(52人)		总体(354人)	
	语文	数学	语文	数学	语文	数学	语文	数学	语文	数学	语文	数学	语文	数学	语文	数学
平均分	86.31	80.76	87.73	84.84	81.80	79.43	80.46	78.58	78.27	77.67	82.14	78.55	72.42	72.95	81.25	78.93
标准差	7.79	13.05	8.46	11.73	12.84	10.34	11.86	12.49	14.68	12.25	13.17	14.72	16.22	19.52	13.27	14.03
名次	2	2	1	1	4	3	5	4	6	6	3	5	7	7		

3．分析方法与分析结果的解读

3.1 班级内部和班级之间的语文数学成绩相关分析

事实上，不但各班平均成绩的语文名次和数学名次有明显的相关性，班级内部学生成绩，更是有如此规律。图1为5、2、7班语文数学成绩对照图，横坐标为学号，纵坐标为成绩，仅从视觉上就可以直观看出同一班内，语文、数学成绩随学号有惊人的相似性。

能将此视觉直觉数量化的，是一个统计参数——相关系数γ和相关指数的平方γ^2。分别取学生语文为随机变量Y、数学成绩为随机变量X，则Y、X的相关系数γ定义为：

$$\gamma = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

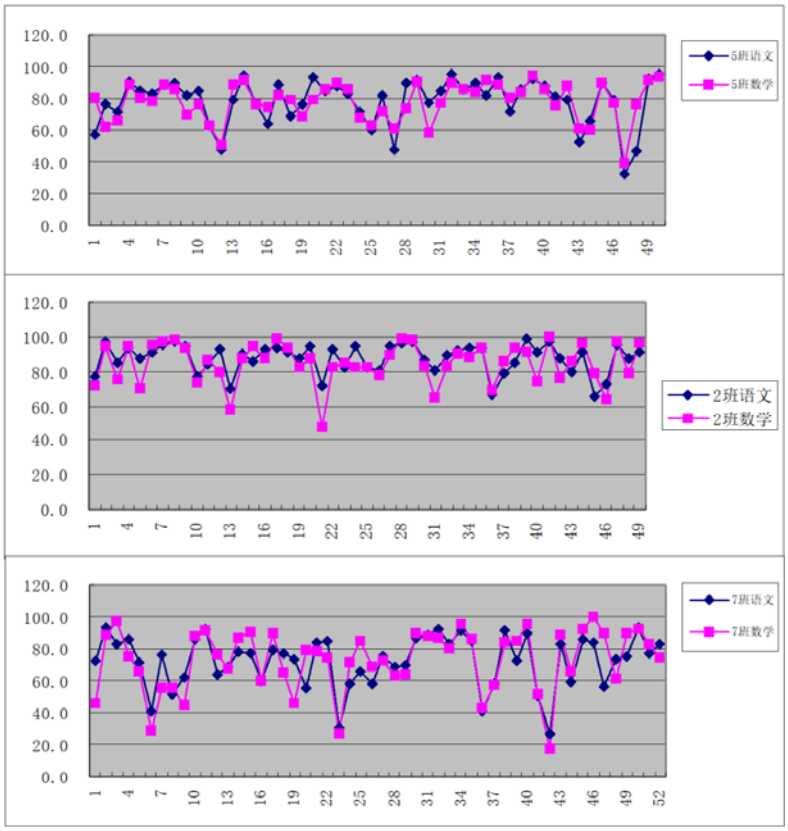

图1 同一班级内部语文数学成绩的相关性

各班级内部以及年级总体语文数学成绩相关系数和决定系数计算结果见表2。相关系数只能在-1~+1范围之间变动，γ=0代表不相关，γ越接近+1，说明正相关性越大。正相关性大也就是同一个学生语文成绩差数学成绩也以更大的概率显得比较差，反之亦然。

表2 各班级内部和年级总体语文数学成绩相关系数（本文所有相关系数均通过了统计显著性检验，过程略）

	各班班级内部语文成绩 Y 与数学成绩 X 之间的相关系数							
	1 班	2 班	3 班	4 班	5 班	6 班	7 班	总体
相关系数的 γ1	0.59	0.75	0.59	0.70	0.79	0.80	0.79	0.73
决定系数或相关指数 γ_1^2	0.35	0.56	0.34	0.49	0.62	0.64	0.63	0.54

表2和图1表明，所有班级内部语文数学成绩均呈现很强正相关性，就年级总体而言，生源因素（含学生、家长因素）对学习成绩离散性产生的贡献率[6]在 γ^2=54%左右，而所有其他因素（包括教师水平、教材差异、随机误差等因素），总共才占46%的影响[6]，因此学生本身差异对班级间平均成绩及其排名的影响不仅不可忽略，而且还可能是第一大贡献因素。但要注意，相关性强并不说明数学、语文成绩互为因果，而更可能是那些影响学生语文、数学成绩背后的共同因素（智商、情感、家庭等的生源差异因素）可能较强，甚至占主导因素，这一点不能混淆。

比如7班的学号为42的学生，语文数学成绩分别为26.5和17分，即不可归因于教师。小学一年级数学只是简单的加减法，非常接近智商测验中的运算能力测验，再差的教师都不至于教出此17分的数学成绩，而更可能是因为学生本身因素（年龄、智商、非智力因素、考试时生病等），与教师水平明显无关，且如果其数学成绩仅17分，则其语文26分就同样不可归因于语文教师。

所以分班为非随机的情况下，抽样亦非随机，故不能简单地以学生成绩均值及其排名来考核教师教学绩效。

3.2 方差分析与多重比较检验[7]

为进一步定量分析学生自身因素与教师因素之间谁是主要因素，我们用双因素方差分析法。因为数学教师是每两个班级为同一教师，不妨以数学成绩为参照系来考察语文成绩，该分析模型在考察语文成绩差异时，不但考虑了班级差异的影响，还考虑了生源差异的影响。

我们将各班所有学生的语文成绩作为观察值，先按班级、再按学号排序，记为因变向量Y，各班所有学生的数学成绩先按同样方法排序，记为自变向量X，我们将数据按表3格式整理后输入Matlab的aoctool协方差统计工具中[7]，可同时得到“同均值”、“分离均值”、“同直线”、“平行直线”、“分离直线”5种数学模型的试算结果[7]。这里，最适合模型为分离直线模型：

$$y=(\alpha+\alpha_i)+(\beta+\beta_i)x+\varepsilon \tag{2}$$

因为：①此模型使来源于误差项的离差平方和为最小，②ANOVA方差表（表4）中“分班*数学成绩”项的概率p-值为0.0020541，即7条拟合线的斜率明显有交叉，所以宜采用此模型[7]。

表3 整理后准备输入Matlab的aoctool协方差分析工具箱的数据的格式

数学成绩 X 向量	75.0	56.5	……	71.5	95	……	39	……	17	……	74.0
语文成绩 Y 向量	71.0	67.5	……	77.5	97	……	32	……	26.5	……	82.5
分班因素 g	1 班	1 班	……	2 班	2 班	……	5 班	……	7 班	……	7 班

我们解释表4的含义：来源于某个因素的离差平方和越大，说明该因素对观测值语文成绩的离散程度（差异性）贡献越大。我们发现虽然班级因素、学生自身数学成绩、两者交互因素对语文成绩都有显著影响，但其中来源于数学成绩的平方和为28201，远高于来自班级因素的2552和交叉项1532之和，如按均方差则差别更悬殊，说明造成语文成绩内部差异的来源中，占压倒性第一因素的差异源来自表征学生本身因素的数学成绩差异（但勿误会成因果关系），其次才是包含教师影响的分班因素。

表4 经Matlab运算后，得到的输出结果ANOVA表（分离直线数学模型）

Source（差异源）	d.f.（自由度）	Sum Sq（平方和）	Mean Sq（均方差）	统计值 F	显著性概率 Prob>F
分班	6	2551.7915	425.2986	5.8972	7.1698e-006
数学成绩	1	28201.0941	28201.0941	391.0387	0
分班*数学成绩	6	1532.0405	255.3401	3.5406	0.0020541
Error	340	24520.2639	72.1184		

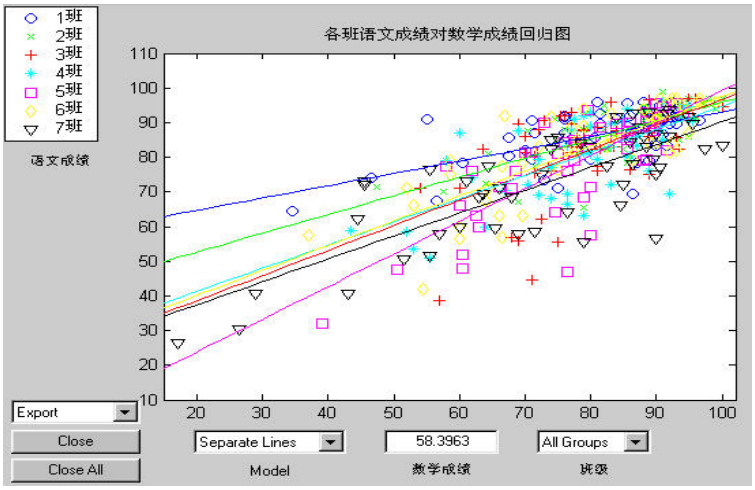

图2 Matlab的Aoctool协方差统计工具箱分析界面

分班因素与数学成绩交互作用显著，表示各班间语文对数学成绩的拟合回归线有交叉，见图2。我们直观地看到，拟合回归线在高分数区域交叉而在低分数区域分叉并散开，提示各班之间主要差异来源是在“低分段”（见后文）。

为比较各班级成绩之间的平均分之差是否存在统计学意义上的“显著差异”，我们进一步将aoctool协方差分析的结果输入Matlab的多重比较函数Multcompare中[7]，得到结果见图3。

我们还可计算在剔除了“粗大误差”成绩的影响后[8]的“调整语文成绩平均分”，具体剔除方法见本文3.3节。剔除后结果数值同时反映在表5和图3上。它反映出按文献[8]方法剔除“粗大误差”成绩学生后，语文成绩差异比想象的要小。

从表5我们看到，具有相同数学教师的1、2班之间，3、4班之间，5、6班之间，语文成绩无统计学显著差异，起码证实这三个对照组内部的两班级间语文教师两两之间无显著差异。

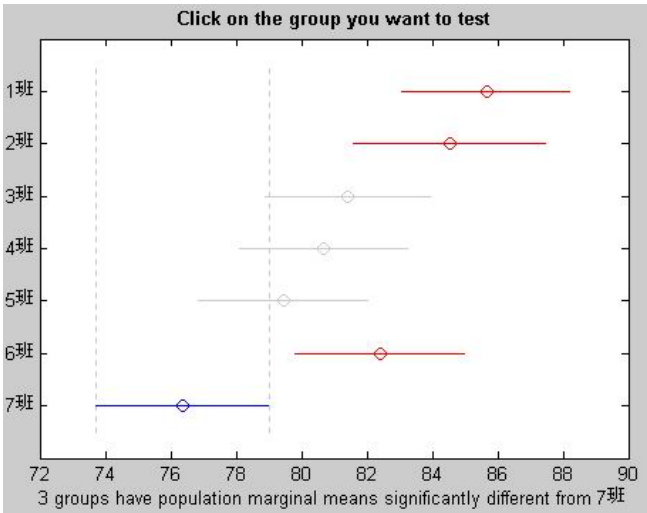

图3 Matlab的Multcompare多重比较分析界面

要特别注意，图3、表5说明经调整后7班语文成绩仍然与1、2、6三个班级有显著差异，但我们依然不能以此判定7班语文教师必定水平很差，而要具体分析。首先，造成语文成绩离差平方和的第一来源仍然是源自自身数学成绩的差异，其次，方差分析是以数学成绩为参照协变量，而恰恰7班数学教师是单独的，所以也不排除7班数学教师特别优秀的原因。

表5 经协方差调整后的语文平均分及排名、以及两两比较是否有显著性（有显著差异则标示*）

组	计数	求和	平均	方差	排名	差异显著性	y1	y2	y3	y4	y5	y6	y7
y1	51	4340.56	85.11	54.61	1	y1					*		*
y2	49	4109.17	83.86	33.28	2	y2							*
y3	52	4236.26	81.47	108.96	4	y3							
y4	50	4034.37	80.69	71.11	5	y4							
y5	50	3954.70	79.09	94.45	6	y5	*						
y6	50	4119.36	82.39	62.96	3	y6							*
y7	52	3969.68	76.34	97.44	7	y7	*	*				*	

3.3 离散性特大的“异常数据”中隐含的有用信息

如图2，我们发现，低分数段个别数据离散程度明显偏离了正常的分布区域，这些数据集中在低分段，且多集中在7班、5班、3班。比如前面提到的某学生，语文成绩26.5分数学成绩为17分，可能无法对应基本日常生活。我们采用“T检验准则剔除异常数据”[8]的鉴别方法，

将语文成绩在53.65以下，数学成绩在49.75以下定为“异常数据”。单科成绩落入此区域的学生为17人，双科成绩落入此区域的有5人，共23人，定义为“成绩异常学生”。

显然教学水平再差的教师都不至于使学生考试成绩如此之低，但恰恰正是这些“异常数据”对班级平均成绩排名造成决定性影响。因此在对其中的原因进行深入分析后，若证实确非教师因素造成，就可以大胆剔除这些“异常数据”，以确保教学评价结果具有准确公平性和实际指导意义。

为此，我们有必要对这些考试成绩中的所谓“异常数据”做进一步研究：

1）纯粹的偶然因素。比如确实有学生考试当天感冒发烧之类。

2）入学年龄太低或试卷难度偏大。深圳小学生入学年龄为6岁，是否对部分发育迟缓的孩子就太小呢？若是由于此原因，其成绩差，必是暂时或一过性的。如果真是这个因素，学校首要的任务是维护这些学生的自尊心，采取耐心和宽容的心态，尤其是在对教师的考核上要松绑，以免对教师的压力传递到这部分学生身上，形成拔苗助长的倾向。小学低年级考试成绩有如此大的差异，使我们怀疑是否课程的难度过大？一年级的主要教学目标不应是知识的学习和掌握，而是在于学习兴趣的培养、学习习惯的养成上。试卷难度过大，使儿童产生畏难厌学情绪，反而不利于学生的潜能开发，使本来暂时的一过性的问题，演化成儿童的幼年挫折感。

3）智商异常。从任课教师中了解到，部分“成绩异常学生”恐怕不能归咎于发育迟缓了，而是要进行深入的病理学研究，包括脑功能、出生过程、用药史、甚至听力测验。比如某“成绩异常学生”，在家长带去儿童医院测得韦氏智商仅为70分，其家长竟沾沾自喜地回来告诉老师智商“及格”了。这既说明孩子智商的确有问题，更反映出家长知识水平可能较低。

4）感统失调或儿童多动症。教师反映部分“成绩异常学生”的确存在多动动作笨拙等现象。

5）家长对孩子教育持放任自流的态度。这在大部分“成绩异常学生”中普遍存在。

总之，对“成绩异常学生”进行原因和预后分析，不但对准确的教学评价至关重要，更对学生本人非常有意义。基础教育原则是机会均等，因此是否关注这些学生，有没有办法提高这部分学生的知识能力，而非歧视和施压，本身就是比平均分排名更重要的一个教学评价指标。特别要注意的是，剔除这部分“成绩异常学生”后，再进行教学结果评价，是确保教学测评公平和科学的关键。

4．今后进一步研究的思路与方向

本研究是以统计学方法，从期末考试的成绩报表中，挖掘出相关的信息，目的在于指出目前流行的，也就是仅凭平均成绩排名，来衡量教师教学成效的缺陷，并给出具体修正方法。

但是，如果能收集更多数据，比如学生入学时测试成绩，以及几年累积的成绩记录，则可以从动态而非静态的角度考虑问题，再配合本文介绍的方法，那么，对教学测评的准确性可以更大的提高。

其次，本文采用数学成绩作为参照协变量，是在没有条件进行智商、心理素质普查的情况下，说明学生差异成因的一种不得已的简化的分析方法。实际中要准确进行教学评价，如果可以对学生的智商、心理素质、体质等进行科学而全面的测试，形成一个初始综合评价指标，以此来做为教学评价的生源参照系，则采用本文同样的分析方法，可以更客观准确。

此外，无论对学校还是教师而言，其教学成果，并不局限于静态的成绩，甚至也不仅仅是动态的成绩提高这一方面。对于学生而言，性格品德的培养，学习习惯和思维习惯的培养，合作精神的培养，体育锻炼带来的健康的身体，有时会影响知识教学的学时投入，也就是，素质教育与知识教育的目标并不总是一致的。但是，显然对于学生的长久发展潜能来看，有时分数稍低所换来的好处，可能远远大于考试成绩的提高。

5．结论与对今后教学评价工作的建议

本研究最主要的意义，在于说明，无论对教师或学校教学评价不应只简单地以算术平均分数来比较，而应该考虑生源差异的影响，也应兼顾教师对各类学生的培养能力上，还要进一步考察学校或教师使学生低进高出的能力上。

本文基于实例所演示介绍的几种数理统计分析法，在很大程度上可以估测甚至排除生源差异的对教学质量评估的干扰性影响。

此外对所谓离散性特大的“粗大误差异常数据”，我们要注意其中隐含的有价值信息，需要进一步具体分析，以便进一步指导教学工作。

References

[1] DING Xiaojie, KONG Xiangyong, WANG Yinglin, WANG Zhuien, XU Wenjuan, DONG Zhengping, Analysis on Influence Factors of Neijing Selected Readings Test Scores of Traditional Chinese Medicine Undergraduate Students, Chinese Medicine Modern Distance Education of China 2018.02.

[2] Wang Bin-hua, Teacher Evaluation: by Value – Added Method, Theory and Practice of Education Vol 25 2005 No.12 P20~P23.

[3] Weichen WANG, Reform And Its Suggestions of Teachers Evaluation in US In Context of Value-Added System, Elementary & Secondary Schooling Abroad, 2011.10.

[4] TANG Xia, The Reform of Learning-outcome-oriented Evaluation System of American Primary and Secondary School Teachers: A Case Study of NYC’s Teacher Development and Evaluation System, TEACHER DEVELOPMENT RESEARCH, 2018.03 Vol 2 No.1.

[5] Chunrong REN, Value-Added Method——a Scientifical Way to Fairly Evaluate The Effectiveness of Schools Base on Data of Exam Results, China Examinations, 2007.04 P12.

[6] Depei ZHANG, Yunling LUO, Applied Probability and Statistics, High Education Press, Sept 2000.

[7] Guiming CHEN, Hongyu QI, Wei PAN, MATLAB Mathematical Statistics, Science Press, Mar 2002.

[8] Shangxu HU, Dezhao CHEN, Observed Data Analysis and Process, Zhejiang University Press, Mar 1996.