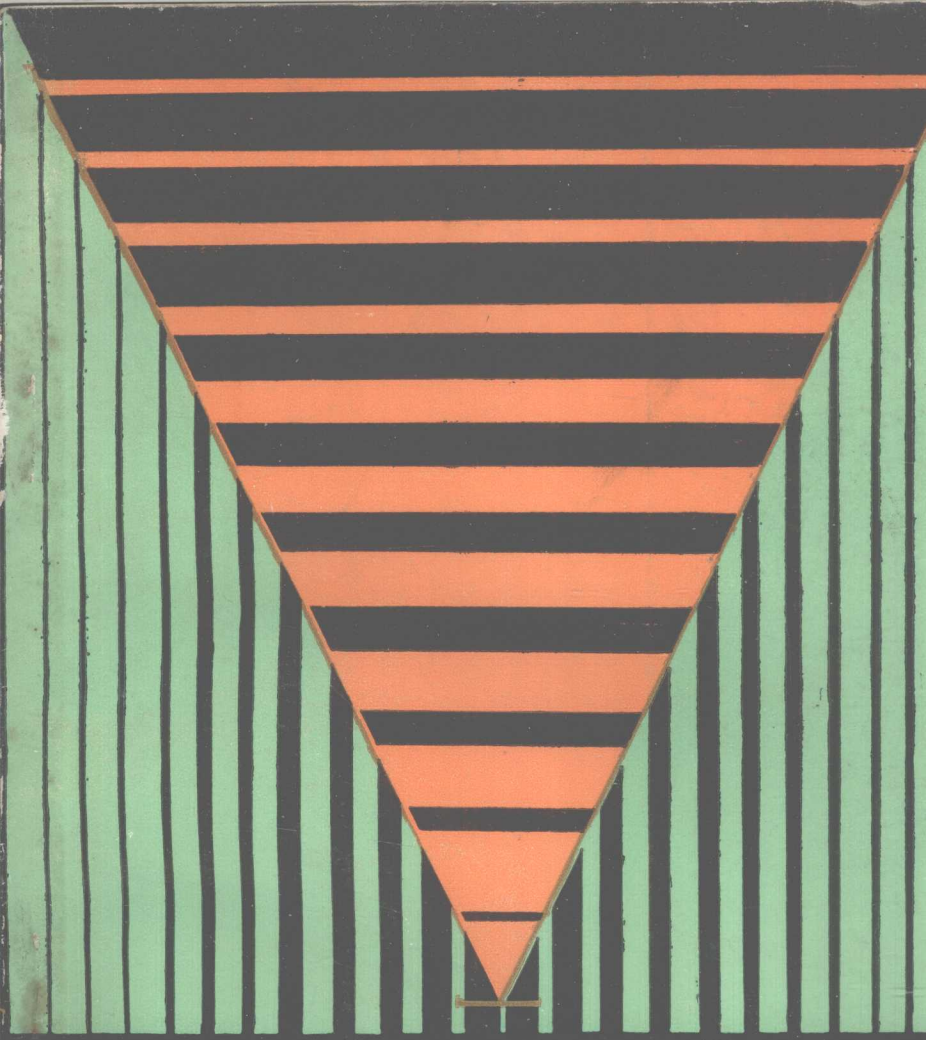

标准化考试常识

曾桂兴

四川教育出版社

标准化考试常识

曾桂兴

四川教育出版社出版

一九八七年·成都

编 者 的 话

教育测量学是教育学科中的一门重要学科。近年来，在国务院有关部门的重视和支持下，这门学科在停开了近三十年后又逐渐恢复和发展起来。目前，我国正在以教育测量学理论为指导，对高考和其它传统考试方法进行改革，试行标准化考试方法。笔者由于教学的关系，在报刊上发表过一些绍标准化考试方法的文章。四川教育出版社的同志提议将这些文章综合成一本较为系统的普及性读物，以便更具体地向读者介绍这一科学的考试方法。《标准化考试常识》就是这样成书的。

本书限于介绍成就测验（Achievement tests）方面的标准化考试方法，而未涉及品格、兴趣、态度方面的测验；其中又以介绍常模参照性的标准化考试为主，而对目标性的标准化考试只在专章（第四章）中介绍。在编写中，力求贯彻通俗性和应用性的原则；在阐述基本理论的同时，尽量列举具体事例及笔者的研究成果以作说明；最后本书附有电子计算机的程序，目的是使读者易于掌握及能在实际中应用。此外，在编写中还参考了一些同志的研究成果，在此谨向原著者致谢！

由于时间仓促，加上笔者能力水平有限，书中出现一些

错误在所难免，欢迎读者批评指正。

编者

于广东省教育科学研究所

1985年11月

目　　录

第一章　预备知识

一、标准化考试与教育测量

标准化考试是教育测量学这门学科里的一个分支。它包括命题标准化、实测标准化、评分计分标准化、分数解释标准化等内容。为了弄清楚这种考试方法，需要简单了解一下教育测量学的有关知识。

教育测量学是在二、三十年代兴起，五十年代定型，最近一、二十年迅速发展起来的一门教育学科。它是以现代教育学、心理学和统计学为基础，运用各种测试手段和方法，对学业成就、教育效果、教育对象的品格、兴趣、态度、潜力及造就方向等一系列教育问题，进行科学地测量和评价的一门学科。

自古以来，有教育活动便有检查教育活动成效的方法。《学记》上说："比年入学、中年考校。一年视离经办志，三年视敬业乐群，五年视博习亲师，七年视论学取友，谓之小成；九年知类通达，强立而不反，谓之大成。"这在当时已经是一套有系统、有目的的考评方法了。国外许多教育测量学者都认为：教育测量实际起源于中国古代的科举制度。隋唐以来，我国实行的科举制，通过考试选拔人才。这种做法对教育测量作为一门学科的出现，产生过重大的影响；这种做法也曾对法国大革命时期的启蒙思想家产生过影响。著名的伏尔泰曾赞叹说："人类的精神，肯定想象不出比这样

政府更好的政府。在这个政府里，重要的衙门彼此统属，任何事情都在那里决定，而其成员，都是经过几场严格考试的。”后来随着我国封建制度日渐腐朽，科举制也就成了统制思想和摧残人才的东西；也由于我国封建制度的束缚，未能使这一学科在我国得到应有的发展。自十九世纪末西方心理学发达以来，特别是有关心理测验的发展，有力地推动了教育测量作为一门学科的出现，使对教育效果的考评方法逐渐趋于客观和科学化。在辛亥革命后，这门学科随西方科学技术一起被引入我国。当时我国的大学教育系和中师学校相继开设教育测量学的课程，不少学者从事这门课的教学和研究。解放初期因精简课程而停开此课，迄今近三十年，致使我国不少教育科研人员、教育行政干部、教育工作者不懂得教育测量学的基本理论和方法，不知道还有一门用测量教育效果来评价教育成就的学科；致使各种不科学的、陈旧的考试方法正在起着评价教育效果、安置、选择和评价人才的重要作用。开展教育测量学的学习和研究，对于改变这种落后状态，实现考评方法的科学化和现代化，培养和造就具有真才实学的各种人才，促进四化建设，都有重要的现实意义。

跟物质（如物理、化学方面）的直接测量不同的是，教育测量是一种间接性的测量，因而它比物质测量困难得多，复杂得多。在教育上的测量是通过编制各种“量表”来对教育对象的知识、能力、品格、兴趣及学习倾向等进行推估和评判。显然，这些“量表”编造得怎么样，测量手段是否科学，评判依据是否合理等，都会直接关系到测量的实际效果。

尽管教育测量存在不少的困难和障碍，但这一课题始终吸引着大批有才华的教育学、心理学和统计学等方面的学

者，正在为它日臻完善而勤奋地探索着。这是因为教育测量无论对从事教育的实际工作者，还是教育科学研究者或是各行各业的行政领导，关系都很密切。就学校教育而言，学生成绩的考查与评定，教师水平的考核、学校管理的评价、各级各类学校的招生与分配，人才的甄选，以及教学方法、教材改革研究等，无不与教育测量密切相关。近年来教育部（现为国家教委）决定恢复和发展这门学科，现已把它列为培养教育科学研究生的专业之一。我们相信，在国家教委的领导和教育科研人员的努力下，这门学科将以崭新的姿态出现在我们中国，它将吸引着大批有志于研究我国教育科学现代化的科研工作者，去开发这一领域。

二、标准分数

在标准化考试中是运用标准分数来衡量考生的成绩的。为了说清楚标准化考试的概念，下面先介绍标准分数及有关的统计学知识。

1. 分数的意义

从孤立观点来看，分数是毫无意义的。如考生某科得到50分，这个分数的价值是什么？是100分中的50分，还是120分（如一九八五年高考中语文科的满分）中的50分？或是50分（如一九八五年高考中生物科的满分）中的50分呢？另外，在某一科的全体考生中，得50分的有多少人？如果绝大部分的考生都能得到50分，那么在高考这类筛选性的考试中，这个50分的价值是不大的。因为它既不是“优”，也无法进行“择优”。如果只有少数人能获得50分，那么这个分数是很有价值的好成绩。可见，考察一个分数的意义或价

值，必须与考察这个分数在特定总体（指全体考生的分数）中的地位与作用密切结合起来。

2. 分数的比较

由于各科考试题目的难易程度和评分标准不同，因而在不同科中所获得的成绩是不能直接比较的。不但不同的分数不能比较，就是相同的分数也不能直接比较。就上述50分而言，在一九八四年理科考生中，如果在政治科（平均分为72.4）或语文科（平均分为66.9）获得这个分数，就不是一个好成绩，因为50分都在这两科的平均分数之下。但如果在数学科（平均分31.5）或生物科（平均分32.2）中获得50分，那就是一个较好或很好的成绩，因为它都高于两科的平均水平；就在这种情况下，生物科中得50分的价值却远优于数学科中的50分。可见，比较不同科分数的价值，也必须与分数在总体中的平均数紧密联系起来。

教育统计学知识告诉我们：平均数是反映一群分数的集中（或称重心）位置，其一般计算公式为

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (1.1)$$

式中，N为考生数（亦称为容量），X_i为考生在考试中的得分，Σ为求和号，$\sum_{i=1}^{N} X_i$表示对这N个考生的分数求和。而标准差则是反映这群分数的离散或波动程度（分散的长度），其一般计算公式为

$$S = \sqrt{\frac{1}{N}\Sigma(X-\overline{X})^2} = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

$$= \frac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2} \qquad (1.2)$$

式中，$\overline{X}$为平均数，由式（1.1）计出；$X - \overline{X}$表示考生分数与平均数之差，统计上称为离差；ΣX^2表示对N个分数的平方后再求和；其余符号同式（1.1）。

标准差的平方，就称为方差。由式（1.2）两边平方即得

$$S^2 = \frac{1}{N}\Sigma (X - \overline{X})^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 \qquad (1.3)$$

试比较下列两组分数：

Ⅰ组：8、11、16、4、6

Ⅱ组：8、12、10、9、6

依式（1.1）可计得它们的平均数均为9，依式（1.2）可计得第一组分数的标准差$S_1 = 4.3$，方差$S_1^2 = 18.6$；第二组分数的标准差$S_2 = 2$，方差$S_2^2 = 4$。可见，尽管两组分数的平均数相同，但两组分数的离散或波动程度是不同的：前者较分散，后者较集中。从齐整角度上看，方差小者（即第二组）的分数较好；若从区分或筛选考生的角度上看，方差大者（即第一组）的分数好些。本例说明，不同科（组）之间分数的比较，除了看平均数之外，还要看标准差。

3. 原始分数和标准分数

我们习惯使用的分数，是未经加工或转化处理的分数，一般称为原始分数。它是直接从试卷上获得的分数。

现在我们所介绍的标准分数，是对原始分数进行标准化处理或转换的分数，其转换分式为：

$$Z = \frac{X - \bar{x}}{S} \tag{1.4}$$

式中，X为原始分数，$\bar{x}$为原始分数的平均数，S为该科分数的标准差，Z就是原始分数X所对应的标准分数了。

显然，在公式（1.4）中所表示的标准分数，是一种与平均数$\bar{X}$和标准差S联合起来考虑的分数。$X - \bar{X}$是定性地表示原始分数X离开平均数（中心位置）的长度和方向：$X - \bar{X} > 0$，表示原始分数高于平均数；$X - \bar{X} < 0$，表示原始分数低于平均数。至于高于或低于平均数的程度，则以S为量度单位。可见，标准分数Z能具体反映出原始分数高于或低于平均数的方向及远近数值。

如例，一九八四年广东省高考化学科中已知$\bar{x} = 60.1$，$S = 19.3$。设某考生在该科考试中得83分，则据式（1.4）可计得它所对应的标准分数为

$$Z = \frac{83 - 60.1}{19.3} = 1.18$$

这就是说，原始分数83分所对应的标准分数是1.18；也表明83分高于平均数约1.18个标准差。

4. 原始分数的弊病和标准分数的优点

原始分数的第一个弊病，是无法反映某个分数在特定考试集体（亦称总体）中的位置和优劣程度。而标准分数却能克服这一缺陷，比较准确地反映考生在特定总体中的相对位置。

例如，一九八四年广东省高考化学科中，原始分数83分在 体中所处的地位和价值是不得而知的。但经过式（1.4）转化为标准分数后，我们不但知道这个分数高于平均数约

1.18个标准差，而且通过查正态分布表（见附表Ⅰ）可知，这个标准分数所对应的概率P＝0.1190。这就告诉我们，如果考生成绩从高至低排列，则大约有11.9%的考生在此成绩之上。显然，这是少数人才有的好成绩（注意：从高分至低分排列时，某成绩所处百分率愈少，则该成绩愈好）。设另一位考生在这门化学科中得分50，则

$$Z=\frac{50-60.1}{19.3}=-0.523$$

这表明，原始分数50分不但处于平均数之下约0.523个标准差，而且查正态分布表得 P＝.3015。注意：当Z＞0 时，可直接查附表中的正态分布表；当Z＜0 时，需用 1 减去所查得表中的概率值，才表示从高分至低分排列顺序的概率值。本例就是说，约有（1 －.3015）＝.6985，即约70%的考生在此成绩之上。显然，自高分至低分的筛选性考试中，它不是一个好成绩。

刚才所说的正态分布是一种两头小、中间大而隆起形似“钟形”的特定分布（见图1.1）。在筛选性考试中，考生成绩分布一般遵从正态分布或近似正态分布。

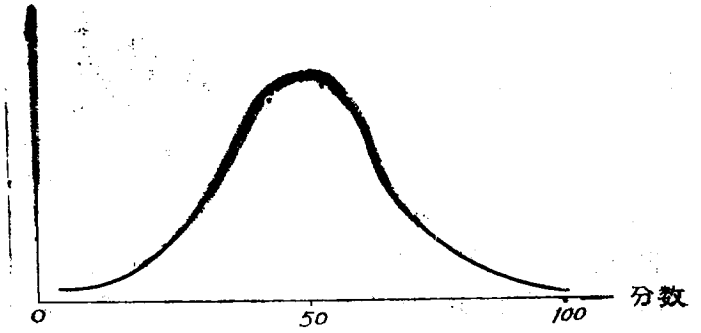

图1.1正态分布曲线

原始分数的第二个弊病是，多科考试的成绩之间一般不具有可加性。这是由于各科分数的性质不同（即平均数不同）、测量单位不同（如标准差不同）而造成各科分数之间不等值的缘故。具体说来，就象一九八四年广东省理科考生中化学科的50分（性质在平均数之下）不能与数学科的50分（性质在平均数之上）直接相加。这跟人民币二元与港币一元或人民币一元与港币二元不能直接相加的道理一样，尽管两者总数都是三元，但各自的价值是不同的。如果硬性相加，结果必然出错。在高考中考生成绩分布一般是正态或接近正态的。这时经过标准化转换的标准分数，无论在那一科中都具有平均数为0、标准差为1的共同特征；这就是说，由于各科的标准分数之中心位置相同，反映分数离散程度的标准差也相同，因而使标准分数具有等值和可加的性质，当然也就具有可比性了。

此外，使用标准分数在命题时可以不受各科固定满分（如50分、100分、或120分）的限制。因为标准分数只是反映考生在该科总体中的相对成绩，而不管原来各科的满分值是多少。使用标准分数还可以帮助录取单位结合自己专业的特殊需要，有效地估计考生的相关科分数在特定总体中的优良程度（采用前述第4点的计算及查表方法）而决定取舍。

三、相关系数

在标准化考试中常用相关系数来评估考试的某些质量，那么什么是相关系数？如何计算和评价相关系数呢？

相关系数是表示两种分数之间联系程度的测度指标，其数值的大小就反映了这两种分数之间联系程度的强弱。例如，在高考中，人们从经验中会感觉到，数学科与物理科，

语文科与外语科等有某种内在的联系。即从多数来说，那些数学成绩好的考生，其物理成绩也会好些；语文科与外语科、同类考生的中学等级成绩与大学等级成绩之间亦有类似的内在关系，这种内在关系就用相关系数来描述。相关系数主要有：

1.积差相关系数

积差相关系数习惯上就称为相关系数。这种相关系数所用的数据是测量数据（如百分制记分的分数）而不是等级数据（如评为优、良、中、可、差）。

〔例1.1〕一九八四年高考中某考室考生（22人）的数学与物理成绩如下，试评估这两种成绩之间的相关程度。

考生号	1	2	3	4	5	6	7	8	9	10	11	12	13
数　学	65	68	51	25	31	30	70	17	41	49	54	39	28
物　理	67	55	28	43	26	22	48	24	52	29	32	47	19

考生号	14	15	16	17	18	19	20	21	22
数　学	37	54	13	43	39	12	32	55	39
物　理	21	58	33	37	37	13	42	64	23

令数学成绩为X，物理成绩为y。我们先用几何描点法来看看这两种分数之间的联系情况（见图1.2）。

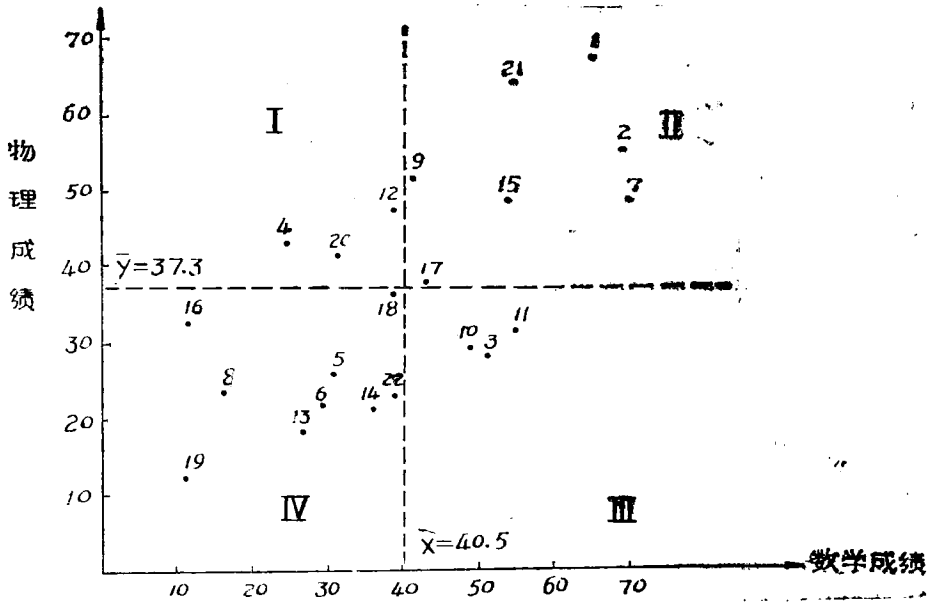

图1.2　某考室考生数学与物理成绩的散点图

在上图中，点2表示第2个考生的数学与物理成绩的交叉点（68，55）；点19表示第19个考生的数学与物理成绩的交叉点（12，13）；余类推。这些数对点的垂直方向对应于X轴所示的数学成绩，其水平方向对应于Y轴所示的物理成绩。由这些相叉点所组成的图，就称为散点图。

分别在$\overline{x}$和$\overline{y}$处画二条互相垂直的虚些线后，就将这些散点分成四个象限。这样就能清楚地看到：（1）那些高于数学平均成绩的点（即在垂直虚线以右），其相应的物理成绩的多数也高于物理平均成绩（在水平虚线之上）；（2）那些低于数学平均成绩的点，其相应的物理成绩的多数也低于物理平均成绩；（3）多数点分布在Ⅱ、Ⅳ象限上，它们形似一“带状”。这表明两科成绩之间有明显的联系。通过绘制散点图，可使我们作出以下的估计：（1）凡多数点在Ⅳ、Ⅳ象限上并呈一明显“带状”者，表明两种分数之间有较强的正相关；（2）凡多数点在Ⅰ、Ⅲ象限上并呈一明显“带状”者，则表明两种分数之间有较强的负相关；（3）“带状”愈明显，相关程度愈强；“带状”愈不明显，即分布在四个象限上的点都差不多（此时呈一种“布满污渍状”），则表明两种分数之间的弱联系或不存在线性相关关系。

相关系数的具体计算公式为

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2 \cdot \sum(Y-\overline{Y})^2}}$$