

中学教师  
继续教育丛书

# 教育统计 与测量

JIAOYU TONGJI YU CELIANG

上海教育出版社

00107

中学教师继续教育丛书

# 教育统计与测量

本书编写组



290713764

上海教育出版社

(沪)新登字107号

中学教师继续教育丛书

**教育统计与测量**

本书编写组

上海教育出版社出版发行

(上海永福路123号)

各地新华书店经销 江苏太仓印刷厂印刷

开本 787×1092 1/32 印张 7 字数 150,000

1995年2月第1版 1995年2月第1次印刷

印数 1—5,300 本

ISBN 7-5320-4165-4/G·4090 定价: 4.80 元

00107

## 说 明

这套《中学教师继续教育丛书》由上海市教育局师范教育处组织编写，供中学教师职务培训使用。

教师职务培训是根据教师的专业技术职务的任职要求，以提高教师政治思想、职业道德素质、教育教学能力、教育科研能力为主要目标而进行的培训。它包括教师修养、专业知识与技能、教育理论、教育教学实践研究、教育科学研究五个方面的内容。这套丛书的编写，力求紧密结合中学教育实际，能帮助教师适应课程、教材改革的要求。我们要求职务培训应具有针对性、适用性、实践性、科学性和先进性。为了体现这些要求，各册教材在编写的过程中，都先编写了讲义，组织了两遍以上的试教，并在试教前后进行了初期、中期、终结三个阶段的论证，反复听取了学员和专家的意见，然后正式编写。

由于编写这类丛书是全新的工作，不当之处在所难免，希望广大读者和专家给予批评、指正。

上海市教育局师范教育处

1993年8月

资料室

051	.....	总论	第二章
101	.....	测验的编制	第三章
108	.....	测验的编制	第四章
071	.....	<b>目 录</b>	
071	.....	附录	

<b>第一章 测量、测验、测题</b> .....	1
<b>第一节 测量</b> .....	1
<b>第二节 测验</b> .....	7
<b>第三节 编制测验的基本步骤</b> .....	16
<b>第四节 测题的编制</b> .....	21
<b>第二章 集中量、差异量、相关量</b> .....	31
<b>第一节 数据处理</b> .....	31
<b>第二节 集中量</b> .....	39
<b>第三节 差异量</b> .....	47
<b>第四节 相关量</b> .....	54
<b>第三章 分数解释</b> .....	65
<b>第一节 正态分布</b> .....	65
<b>第二节 测验分数的解释</b> .....	77
<b>第四章 统计推断</b> .....	99
<b>第一节 统计推断中几个概念</b> .....	99
<b>第二节 总体平均数的推断</b> .....	102
<b>第三节 平均数差异的显著性检验</b> .....	111
<b>第四节 方差分析</b> .....	121
<b>第五节 <math>\chi^2</math> 检验</b> .....	127
<b>第六节 相关系数的假设检验</b> .....	136
<b>第五章 信度、效度、难度、区分度</b> .....	145
<b>第一节 信度</b> .....	146

第二节	效度 .....	156
第三节	测题的难度 .....	164
第四节	测题的区分度 .....	168
附录	.....	176
一、	相关系数一览表 .....	176
二、	计算器使用 .....	178
三、	1990年某校高一入学考试成绩表 .....	182
四、	1990年某校中预班入学考试成绩表 .....	188
五、	某校1989学年第一学期高一期中考试考试成绩统计表 .....	192
六、	附表 .....	193
1S	.....	.....
18	.....	.....
18	.....	.....
08	.....	.....
YF	.....	.....
18	.....	.....
80	.....	.....
80	.....	.....
YF	.....	.....
00	.....	.....
00	.....	.....
101	.....	.....
111	.....	.....
181	.....	.....
781	.....	.....
801	.....	.....
181	.....	.....
081	.....	.....

# 第一章 测量、测验、测题

## 第一节 测 量

### 一、教育测量

#### (一)教育测量的概念

教育测量就是对教育领域内的事物和现象依照一定的法则表示为数字的过程。如学生的思想品德和学习成绩的测量,教师教学效果的测量,学校的经费和行政管理效率的测量等。本书所介绍的教育测量,主要是指对学生某学科经过学习和训练后所获得的知识 and 技能方面的测量。

#### (二)教育测量的可能性

对学生所学得的知识 and 技能的测量是一种心理属性的测量。由于人的心理属性是抽象的、不易捉摸的,实现这方面的测量与物理测量相比是比较困难的。但是,正如美国心理学家桑代克与教育测量学家麦柯尔曾经指出的,“凡是存在的,必有其数量”,“既有数量,即可测量”。虽然心理属性看不见,摸不着,但它是客观存在的,必定会反映在某些行为之中,因而实现教育测量完全是可能的。只要在技术和工具上狠下功夫,教育测量必将更完善、更可靠。

#### (三)教育测量的性质

##### 1. 教育测量的间接性

当今,我们还无法直接测量人的心理属性,只能测量人的外显行为,即我们只能通过一个人对测验题目的反应,来推论

出他的心理特质(即个体特有的、稳定的、可辨别的特征)。教育测量是通过测验题目的行为反应来推论他的学业成绩,所以说教育测量是间接的。

## 2. 教育测量的相对性

我们在判断某人的行为时,并没有绝对的标准、永恒的标准,而是他的行为和别人的行为进行比较时才能作出判断。如测量一个人的智力的高低、兴趣的大小等,都是与所在团体的大多数人的行为或某种人为确定的标准相比较而言的。所以说教育测量是相对的。

## 3. 教育测量的客观性

由于人的行为是客观存在的,具有内在一致性。只要测量的量具,也就是测验标准化,评分、计分标准化,分数转换及解释标准化,测量的值也是稳定的,对结果的推论较为可靠,也就是说能比较客观地反映人的行为。这就是教育测量的客观性。

## 二、测量的方法和量表

### (一) 测量方法

根据测量能否直接测到所要测的属性,可以将测量分为直接测量和间接测量两类。直接测量可以用工具直接测得事物的属性。如天平称物体的重量,用尺子量物体的长度。间接测量是根据测量的结果去推测事物的属性。也就是说,测量结果本身并不是所要测量的目标,只有通过推理分析,才能对目标进行推测。如根据水银温度计中水银柱的长短来推测温度的高低。在智力测验中,根据被试人的分数来推测其智力的高低。

教育测量属于间接测量。例如,我们要测量学生代数方



面的知识和技能,就要让学生来完成这方面的一套试题,根据学生做对试题的多少来推出该学生所掌握的代数知识和技能的程度。这种程度通常用分数来表示。

## (二)量表

量表是依据事物的特性和设定的法则,使一组数字或数能够用来达到描写事物所拥有的特性的程度。量表的水平高低与计量单位、参照点有关。

### 1. 计量单位

计量单位是计量事物的标准量的名称。如果没有计量单位,则数量的多少和大小就无法表示。作为计量单位必须满足两个条件:

(1) 具有确定的意义。同一单位在所有人的心目中都有同一意义,不允许有不同意义。

(2) 单位的距离等值。如30千克、40千克、80千克、90千克的四个物体A、B、C、D,则A、B之间的重量差和C、D之间的重量差是相等的。

### 2. 参照点

参照点就是计算的起点。参照点也可称为零点。若参照点不同就无法进行比较。

参照点可分为两种:

(1) 绝对零点。如长度、重量都是有绝对零点的。

(2) 相对零点。如地平面以海拔 $\times\times$ 米表示,是以海平面为相对零点。摄氏温度计是以水的冰点为相对零点。

## (三)四种测量量表

量表的种类根据测量的精确程度与数的区分性、等级性、等距性、等比性等特点,可以将量表分成四类,将它们从低级到高级依次排列为类别量表、等级量表、等距量表和等比量表。

### 1. 类别量表(或名称量表)

类别量表就是根据某一特点,对两个和两个以上的对象进行分类并用数字加以表示。它是量表中最简单的形式。如人口统计中规定男性为“1”,女性为“0”。又如学生的学号等。这里用来描述各类事物的数字仅仅是事物的名称,它只具有表示事物相同与不同的特性,没有参照点和单位,没有数量的大小、多少、位次和倍数关系,即它只具有数的同一性和区分性,而不具有等级性、等距性、等比性和可加性。因此,不能将它进行加减乘除运算。

### 2. 等级量表(或顺序量表)

等级量表就是根据某一特点,将事物分成等级,并且用数字表示。如工厂产品分成一等品、二等品、三等品。又如,学生的思想品德评定,根据学生的表现,分别用5、4、3、2、1五个数字表示优、良、中、及格、不及格。百米赛跑中,第一名、第二名、第三名。这些数字不仅具有数的区分性,而且还具有数的等级性(顺序性)。可以表示事物大小、好坏、先后的位次、顺序关系,但没有单位和参照点,不具有数的等距性、等比性和可加性。因此,也不能将它们进行加减乘除运算。

### 3. 等距量表

等距量表是具有相等单位和人定参照点的量表。如摄氏温度计测量的温度,每相差一度的距离是相等的。但只有参照点 $0^{\circ}\text{C}$ ,而不是绝对零点。钟表计时也是等距量表。这种量表上的数字不仅具有数的区分性和等级性,而且具有等距性,但没有绝对零点,没有等比性。因此,量表上的数值只能作加减运算,不能作乘除运算。

### 4. 比率量表(或等比量表)

比率量表是有相等单位和绝对零点的量表。如长度、重

量。这种量表上的数值不仅具有数的区分性、等级性、等距性、等比性，还具有绝对零点。因此，量表上的数值可以进行加、减、乘、除四则运算。例如，甲重是60千克，乙重是30千克，可以说甲比乙重30千克（60千克-30千克=30千克），也可以说甲重是乙的2倍（60÷30=2）。

我们了解各类量表的不同特征，对测量结果的解释具有重要意义。为此，把四种测量量表的比较列成表1-1。

表1-1 四种测量量表的比较

量表类别		类别量表	等级量表	等距量表	比率量表
特征	零点	无	无	有(相对零点)	有(绝对零点)
	单位	无	无	有	有
	区分性	有	有	有	有
	等级性	无	有	有	有
	等距性	无	无	有	有
功能	等比性	无	无	无	有
	类别	有	有	有	有
	等级	无	有	有	有
适用运算	比较大小	无	无	有	有
	比较倍数	无	无	无	有
适用运算	加、减	无	无	有	有
	乘、除	无	无	无	有

#### (四)教育测量的测量水平

量表的特征越多，功能也越多，适用的运算就越多，它的测量水平也就越高。所以比率量表优于等距量表，等距量表优于等级量表，等级量表优于类别量表。我们尽量使量表达达到较高水平。

在教育测量中，对学生的知识、技能的测量所获得的测验

分数都应属于等级量表。因为测验分数之间只能表明哪个大、哪个小，不能表明大多少。在一次测验中，50分与60分之间的差同90分与100分之间的差，虽然都相差10分，但是它们的差异是不相等的，相等单位是很难获得的，绝对零点也难以确定，这表明测验分数是不等距的。所以，严格地说，教育测量所用的百分制是等级量表，但等级量表在统计方法上将受到很多限制。为了提高测量水平，我们把测验分数的等级量表提高到等距量表。所采用的办法是：假设测验分数是呈正态分布，统计时将原始分数转化为标准分数（见第二章），使其成为等距量表。如果在编制测验时小心谨慎，尤其对测验结果两极端分数的微小差异可能表明着巨大差别的现象加以留意的话，就能使误差减到最低程度。如果测验的编制程序能使测验分数接近于等距量表，而且把测验分数当作等距量表处理时所得到的结果也确实有意义，那么，以测验分数当作等距量表来处理也是可行的。

### 三、评价

通过测量，我们对事物的属性可以作数量化描写，但这并不是我们对问题研究的最终目的，还应该进一步对事物进行评价（或称作评估）。

测量后，对测量的结果作出价值判定即为评价。例如，某学生的数学成绩是64分。这分数是他成绩量化的表示，不能说明价值意义。用这分数与其他学生的成绩进行比较，是好还是差；或者和一个确定的标准（例如某一单元的教学目标）进行比较，说明它是否达到目标，相差多少；或者与他自己以往成绩进行比较，是进步了，还是退步。这些都是对该生成绩所作的价值判定，即评价。

测量和评价是紧密相连的两个问题。测量是评价的基础，若没有测量，就没有可进行评价的素材。评价是对测量结果的进一步判定，若仅有测量，而不进行评价，则无法确定该事物属性的价值如何，是优是劣。因此，也可以把评价的过程看作是一个总结的过程。当然，测量和评价之间是有界限的。

## 第二节 测 验

上一节提到，对学生知识、技能的测量一般不能采用直接测量，只能采用间接测量。即让学生完成一套有关的试题，根据学生答对试题的多少来推出该学生知识技能水平。这也就是通常说的测验。所以，测验是教育测量的重要工具。

### 一、测验的概念

美国心理与教育测量学家布朗给测验下了定义：测验是测量一个行为样本的系统程序。

这个定义有下面三层意思：

首先，测验所测量到的是人的行为，就是被试者对测验题目所作的反应，根据反应推测所要测量的属性。在数学测验中，“行为”可理解为：(1) 背出、默出、指出。如默写三角形全等的定义，或指出三线八角中哪些是同位角。(2) 举例说明。如举例说明方程与等式的联系与区别。(3) 求、解、计算、证明、作图。如，求方程的解，解三角形等。根据学生的这些行为反应来测量学生在数学方面的知识和技能。如果被试者在测验中所表现出来的行为能够正确地反映所要测量的属性，那么这个测验将为我们提供有效的信息。

其次，一般来说，一个测验只是全部可能题目的一个样

本。它一般不可能包含所有题目。因此，用于测验的题目就必须是所有同类题目所组成的总体中一个有代表性的样本，并且，当从同一总体中抽取不同样本时应该得到相近分数，分数越相近，测验越可靠。

第三，测验是一个系统程序。也就是说，在测验的编制、实施、记分、分数的解释等方面都是按照统一的标准和严格的规定进行的。测验的题目是根据测验的目标，经过系统的分析和选择而组合起来的。在实施测验时，对所有的被试者都在相同的条件下施测相同的题目。在评分方面，也是根据事先确定的规则，保证在评分者之间取得一致的意见。测验的分数必须与统一的标准进行比较，才能确定其优劣。这样，才能得到测验的真实结果，使被试的测验分数具有可比性。这也就是通常所称的标准化。

## 二、测验的误差、信度和效度

测量的有效性和可靠性是通过效度和信度这两个数量指标来描述的。为了说明效度和信度的含义，就得从误差分析谈起。

### (一) 测验中的误差

误差是在测量中与目的无关的变因所产生的不准确或不一致的效应。

我们通常把教育测量中的误差分为两种：系统误差和非系统误差。系统误差是由与测量目的无关的变因所引起的一种稳定而有规律的效应。例如，在数学能力的测验中，对学生的阅读能力的要求超过了学生的实际的语文能力，就会影响学生的成绩，由这种变因所引起的误差就是系统误差，又称条件误差。非系统误差，又称随机误差或测验误差，它是由与测

量目的无关的偶然因素引起的而又不易控制的误差。例如，同一份试卷在不同时间进行测验，考生所得分数的差异就可以归诸偶然因素。

系统误差、随机误差与获得分数、真分数密切相关。所谓获得分数是学生在一次测验中所得到的卷面分数。所谓真分数就是在假设测量没有误差时所得到的分数。可见真分数只是一个理论上的概念，在实际测量中是得不到的。真分数也可以理解为经过无限次测量所得分数的平均数。

由于误差存在，一个学生获得的分数  $x_t$  与他的真分数  $x_\infty$  之间总是存在差异的，获得的分数与真分数之间有下列的关系式：

$$x_t = x_\infty + x_e, \quad (1-1)$$

其中， $x_t$ ——获得分数， $x_\infty$ ——真分数， $x_e$ ——随机误差。

在公式(1-1)中未见系统误差，其实系统误差是稳定的，只要不改变测验条件，在测验的多次施测过程中，它的出现是稳定的，可见它必定包含在所谓真分数之中。这样，真分数  $x_\infty$  又可分为两个分量之和，即

$$x_\infty = x_v + x_s, \quad (1-2)$$

其中， $x_\infty$ ——真分数， $x_v$ ——有效分数， $x_s$ ——与测验目的无关的无效分数，即系统误差。

因此，获得分数便可表示为：

$$x_t = x_v + x_s + x_e. \quad (1-3)$$

其中，系统误差只影响测值的准确性，而随机误差既影响准确性，又影响一致性。

以上诸式是对被试个人而言的。但是，我们将要讨论的有效性、与可靠性(一致性)是整个测验的性质。关于这些问题，我们将在第五章中进行讨论。

### 三、测验的分类

测验的种类,可以从不同角度加以划分。例如,按测验对象、测验性质、测验材料、测验功能等可作不同分类。这里介绍四种分类方法。

#### (一)根据测验的标准化程度分类

##### 1. 标准化测验

标准化测验是按照系统的科学程序组织、具有统一的标准、并对误差作了严格控制的测验。标准化测验在每个环节上都要求标准化,具体包括试题编制的标准化,施测过程的标准化,评分记分的标准化,分数合成与解释的标准化等。

标准化测验的试题是由有关专家根据一定的教育目标集体编制的。这些专家不但要精通本学科的知识,还要受过心理与教育测量学方面的训练。在编题前要制定编题计划,以保证题目对知识和能力两个维度均具有代表性;所有题目都要经过预测和统计分析,取得难度、区分度等有关资料;在拼配试卷时,题目的难易和排列顺序要得当,以符合学生的心理特点。标准化测验的题目一般都有多套等值的复本。

标准化测验有很严格的实施手续和施测条件。测验手册中对考场设置、收发试卷手续、测验说明、主试者的态度、受试者注意事项以及如何计时、意外事件如何处置等均有明确规定,任何人不得随意改变。总的要求是保证实施手续与施测条件客观化,避免环境与各种偶然因素对测验成绩的影响。

标准化测验对评分记分、分数的合成方法和解释方法都有明确的规定,尽可能采用最有效的方法。从测验中直接得到的原始分数要转化到一个有确定参照点(相对零点)和单位的量表上去才有意义。这种转化后的分数叫导出分数。导出分数有两种:一种叫常模参照分数;另一种叫目标参照分数。



前者是把个人的分数与其他人进行比较,以所在团体的平均分数(即常模)作参照点,根据个人分数距平均分数的远近确定个人在团体中所处的位置。后者是把个人分数与教育目标比较,以某种可接受的最低标准作参照点,看一个人的成绩是否达到标准或对某一个指定范围的内容或技能掌握了多少。标准化测验在向当事人报告分数时都采用导出分数,而且把每个人的分数看作是一个区间,而不是一个确定的点,并指出落在某一个区间的可能有多大。在解释分数时,还考虑到对当事人可能产生的心理影响。

## 2. 非标准化测验

标准化测验之外的其他测验就是非标准化测验,如学校常用的教师自编测验。

教师自编测验的试卷一般由教师个人或集体编制,施测由教师个人或学校组织,评分记分由教师本人进行。因此,这种测验的测量误差比标准化测验要大,客观性也较差。但是,这种测验与教师日常教学工作息息相关,其内容与教材内容、教学目标、教学进度一致,难易程度适合于学生水平,有较强的针对性,便于随时了解学生的学习状况,及时地改进教学。这种测验是学校内和校际间经常使用的一种测验。如果在编制这种测验时,采用标准化测验的某些编制原理与方法,就会使测量误差得到较好的控制,测验的效度和信度也可以得到提高。

### (二)按测验目的分类

#### 1. 描述性测验

描述性测验的目的在于对个人或团体的能力、性格、兴趣、知识水平进行描述。如现行的智力测验和教育测验。

#### 2. 诊断性测验