



jiaoyu tongji de jiben lilun
yu

SPSS caozuo jishu

● 王秀玲 刘兰英 编著

教育统计的基本理论 与 spss操作技术



杭州出版社

目 录

序	1
前 言	3
第一章 频数分布和频率分布	5
第一节 频数分布表和频率分布表	5
第二节 频数分布图和频率分布图	16
第二章 频数分布的特征量	26
第一节 集中趋势指标	26
第二节 离散程度指标	35
第三章 地位量数	44
第一节 百分位数	44
第二节 标准分数	49
第四章 总体平均数的估计与显著性检验	65
第一节 样本平均数的抽样分布	65
第二节 总体平均数的区间估计	72
第三节 单个总体平均数的显著性检验	77
第五章 两个总体平均数差异的显著性检验	92
第一节 两个独立正态总体平均数差异的显著性检验	92

第二节	两个相关总体平均数差异的显著性检验	101
第六章	多个总体平均数差异的显著性检验	109
第一节	多个独立总体平均数差异的显著性检验	109
第二节	多个相关总体平均数差异的显著性检验	127
第七章	相关分析	138
第一节	两变量的相关分析	138
第二节	偏相关分析操作技术简介	158
第八章	χ^2 检验	166
第一节	计数数据单向分类的 χ^2 检验	166
第二节	计数数据双向分类的 χ^2 检验	175
附 录		194
附 表		220
主要参考文献		246

序

改革开放以后,我国各高校教育系和心理系相继恢复了教育和心理统计学的教学,至今已有二十余年。这期间,出版了众多教材或著作,这对提高我国教育和心理科学的研究水平起到了很好的促进作用。就已出版的教材来说,我以为大多数处在同一层次的水平上,还难以适应从事基础教育工作的广大教师 and 实际研究工作者的需要。而且,大多数教材的分量都超过规定的教学时数,教育系的学生常常反映这是一门最难学的课程。从 20 世纪 90 年代开始,信息技术进入高速发展的时期,统计学作为一门方法学或工具学科,越发使人感到现有的一些教育、心理统计学的编著远远滞后于形势发展的要求。所以,编写多样化的、充分运用现代信息技术的现代统计学教材,是进一步提高教育科学研究水平的一项重要工作。

前年,我在美国参观了上、中、下三种水平的五所大学。在斯坦福大学书店里,仅统计学的样书就有满满三大书架(店堂里只陈列一二本样书,工作人员每天早上补充售出的样书)。品种之多,内容之丰富多彩,装帧、印刷之精美,令我欣喜若狂。这里有理论研究的专著,更多的是应用性的工具书,还有操作性的实用手册。即使是应用性的工具书,内容也有深浅之分,各具特色,而不是简单重复。这与我们的现状相比,差距太大了。当然,那里的书价之高,也让人望而却步。

现在,杭州师范学院王秀玲副教授和上海师范大学刘兰英老师撰写了《教育统计的基本理论与 SPSS 操作技术》一书,定位明确,面向广大中小学老师。我读了以后,颇有新鲜感。学习和掌握统计分析方法,关键是要在实际使用中学习。正是在这一点上,两位老师作了很好的尝试,她们把学习和运用基本统计分析方法与实地使用社会科学统计软件包 SPSS 密切地结合起来。从最简单的基础内容——频数和频率分布表与图的制作,到最重要的基本统计推断方法—— t 检验、 χ^2 检验等,直至方差分析,除了举例介绍常规的计算方法外,把重点放在如何运用 SPSS10.0 软件在计算机上一步步地操作,从而展示了分析的全过程,直至得出最后的分析结果,使初学者免除了硬记计算公式和进行繁琐计算的烦恼,有利于培养读者学习和应用统计分析方法的兴趣。附录中 II 还简明扼要地介绍了 SPSS10.0 软件的安装、启动、运行和退出的方法,缺乏计算机操作基础的读者也能慢慢学会。其次,学习和掌握统计分析方法,特别需要循序渐进,逐步扩大知识范围。离开了应用,片面追求系统、完整,往往会事倍功半,欲速则不达。所以本书只选取了教育、心理统计学中最基本、最常用的内容,理论部分也限于概念的定义和统计意义的解释,以及方法的举例示范等,对原理不作探究讨论。我想,这会减少初学者的困难。首先学会应用,在应用中增进理解,这样或许有助于读者产生继续学习的兴趣和要求。

浙江大学教育系

王 权

2002 年 5 月 15 日

前 言

走进新世纪,我国教育事业呈现出蓬勃发展的新局面。在教育发展和改革的浪潮中,人们关注着教育规模、数量的扩张,更注重教育质量的提高。教育质量的提高,关键在于教师的教学水平和开展教育科研的能力。中小学教师的教育教学研究水平如何,不仅影响着教育质量,而且对教育的可持续发展也至关重要。

作者长期在师范院校工作,承担《教育统计》等课程教学任务。在教学过程中,深感教育统计的基本理论与实际应用存在一定程度的脱节。学习者能掌握教育统计的一般原理、完成作业并较顺利地通过考试,可一旦要对实际问题进行统计处理,则常常感到力不从心。近年来作者以“理论与应用相结合”的思路探索教育统计课程的教学,以尽可能浅显易懂的方式阐述教育统计的基本理论,把课堂教学与实际教育科研活动相结合,把统计方法的选择与SPSS(Statistics Package for Social Science)软件的具体操作相结合,以最经济有效的方法培训学习者掌握基本的数据分析与处理技术。几年的尝试证明,这不仅有利于提高学习者运用现代计算技术对教育问题进行数量化分析的能力,对提高中小学教师的科学研究水平也将起到重要的作用。在总结多年教学经验的基础上,我们编写出版本书,旨在为中小学教师、大专院校的学生(尤其是非教育专业的学生)提供简明的教材和自学参考资料。

本书的特点是:

1. 提要导引,指导学习。每章的内容提要为读者点明学习目标及把握学习的重点,以期起到纲举目张的作用。

2. 体系完整,浅显易懂。以实例引入,在保持基本理论体系完整的前提下,侧重于基本原理的理解和基本方法的选择,对一些数理统计的基本原理只作叙述或引用而不加以数学的严格证明,尽量降低数学方面的要求。依据教育统计的基本原理选择处理实际问题的方法,并对各种结果作出科学的解释。

3. 重视操作,强调应用。各章中的每一节都提供了教育教学实践和研究的例子,既注重典型范例的分析和解决,更重视内容和方法的应用,强调用 SPSS 程序处理数据的操作过程。把教育统计的基本理论与教育实际问题相结合,将“量化”处理方法与 SPSS 具体操作相结合。

全书共八章。王秀玲拟定本书大纲,撰写第二、四、五、六章及附录,并完成全书的定稿、校稿工作;第一、三、七、八章由刘兰英撰写。

在本书撰写直至出版的过程中,我们的导师、浙江大学教育系王权教授给了我们很多指导,对此,我们深表感谢。我们还得到了家人的支持,尤其是王秀玲爱人徐宪民先生给予我们诚挚的关心与鼓励,爱女浙江大学硕士研究生徐璿帮助录入部分书稿。杭州师范学院继续教育学院的领导为本书的写作和出版发行给予很大支持。撰写过程中,我们参阅了国内外专家的许多著作,有关中小学为我们的教育实验研究提供了机会,使我们获得了珍贵的第一手数据资料。在此,我们一并深致谢意。由于作者水平所限,书中若有不妥甚至错误之处,谨请读者不吝赐教。

作 者

2002年初夏于杭州

第一章 频数分布和频率分布

- ◇ 频数分布与频率分布
- ◇ 频数分布表与频率分布表的编制
- ◇ 生成频数与频率分布表的 SPSS 技术
- ◇ 频数分布图与频率分布图的编制
- ◇ 生成频数与频率分布图的 SPSS 技术

第一节 频数分布表和频率分布表

【实例 1-1】

用一份测验同时测查五(1)和五(2)两个班的学生,其中五(1)班 51 人,五(2)班 48 人,两个班的成绩如下所示。凭观察判断两个班学生成绩各自在哪个分数段的人数最多或比率最高。

五(1)班的成绩:

75 60 57 75 55 64 68 72 56 61 90 86 73 72 95
81 71 56 80 78 74 72 65 84 88 87 88 64 72 79
67 76 75 94 45 54 55 80 72 92 70 65 75 63 89
52 65 91 59 64 94

五(2)班的成绩:

65 54 66 77 76 62 72 71 78 61 82 75 95 60 58

99 70 64 84 75 90 70 75 86 69 84 62 73 74 79
 68 80 86 56 71 94 83 67 60 86 74 79 68 90 70
 65 48 62

【统计方法】

采用频数分布表或频率分布表。因为上述所列数据都是原始的测验成绩,显得杂乱无章,很难一下子看出两个班成绩各自的分布情况及其相互间的差异。通常,我们会将每个班学生的测验成绩从高分到低分依次排序后再进行比较。总的看来,这种做法,虽然可以了解到诸如最高分与最低分,所有分数的分布范围,不同分数各自重复出现的次数,以及大多数学生的分数分布在什么范围等信息,但在总体上还是难以一目了然地反映事物的全貌及其特性,达到清晰、概括地统计数据的目的。因而,这里我们采用频数或频率分布表,借以直观反映众多无序数据所蕴含的信息。

【基本理论】

一、频数分布与频率分布的概念

频数是指某一个随机事件在 N 次试验中实际出现的次数。所谓频率,又称相对次数,指某一随机事件在 N 次试验中实际出现的次数占总试验次数的比。如果用 f 表示频数, N 表示总次数, R_i 表示频率,那么:

$$R_i = \frac{f}{N} \quad \text{公式(1-1-1)}$$

譬如,在抛硬币游戏中,如果总共抛了 10 次,当硬币落在地上时,3 次正面朝上,7 次反面朝上,那么,我们可以认为,正面朝上与反面朝上的频数分别为 3 与 7,而正面朝上与反面朝上的频率分别为 $3/10$ 与 $7/10$ 。

由各随机事件出现的频数或频率构成的分布称为频数或频率分布。用以描述频数或频率分布的表即为频数或频率分布表。

二、频数分布表与频率分布表的编制

(一) 离散型随机变量频数与频率分布表的编制

例如,某班 45 名学生在限定时间内进行五次投篮命中比赛。其中,1 只都没有投进的有 3 人,投进 1 只球的有 6 人,投进 2、3、4 只球的分别有 7、16、9 人,全投进的有 4 人。命中的球数是离散型随机变量,因此这 45 名学生投球命中情况的频数与频率分布表为:

表 1-1-1 45 名学生投球命中频数与频率分布表

命中球数	0	1	2	3	4	5	总和
频数	3	6	7	16	9	4	45
频率	0.07	0.13	0.16	0.35	0.20	0.09	1.00

该表是离散型随机变量频数与频率分布表中较为简单的一种。还有较为复杂的,如在一张表中反映两组或两组以上数据资料分布情况的频数与频率分布表,这种表我们称之为复合表,如表 1-1-2 所示。

表 1-1-2 某年级男女生投球命中频数与频率分布表

命中球数		0	1	2	3	4	5	总和
男	频数	3	6	7	16	9	4	45
	频率	0.07	0.13	0.16	0.35	0.20	0.09	1.00
女	频数	5	8	5	12	7	4	41
	频率	0.12	0.20	0.12	0.29	0.17	0.10	1.00

离散型随机变量频数与频率分布表的编制较为简单,在此略去编制步骤。

(二) 连续型随机变量频数与频率分布表的编制

1. 频数分布表的编制。

以实例 1-1 中五(1)班的成绩为例,简述编制连续型随机变量频数分布表的主要步骤:

(1) 求全距。所谓全距,指的是一批数据中最大值与最小值的差,通常用符号 R 表示,即 $R = X_{\max} - X_{\min}$ 。本例中的全距 $R = 95 - 45 = 50$ 。

(2) 定组距与组数。所谓组距,就是组与组之间间隔的距离(常用符号 i 表示);组数,就是分好组后总共有几组(常用符号 k 表示)。全距、组距和组数三者的关系为 $k = \frac{R}{i}$ 。常用的组距为 1、2、4、5、10 个单位等。如果全距和组距确定了,那么组数也就随之而定了。

(3) 定组限。所谓组限,就是每组的起止范围。一批数据中的最小值为最小一组的下限,自然就有最小值 $X_{\min}$ 包含在最小的组中。本例中最小值为 45,所以最小一组的下限就是 45。在列表时无须写出各组的上限,一般情况下数值较大一组的下限就是数值较小一组的上限。如表 1-1-3 中第(1)列所示。在将数据归组时,如遇有数据正好等于某组的上限,则可将它归入数据较大的一组。如可将 50 归入第二组,将 55 归入第三组。一批数据中最大值 $X_{\max}$ 应包含在最大组中。

为书写方便,也可用组中值(常用符号 X_c 表示)代表各组的组限。所谓组中值,就是每组上下限的平均数,即 $X_c = (\text{上限} + \text{下限})$

÷ 2。如表 1-1-3 中第(2)列所示。

(4) 归类划记,列出频数。分好组后,依次将每个数据准确地划归所属的组别,并以“正”或“𠄎”等符号记录。如表中第(3)列所示(一般不列入表中,这里只是为了说明编制步骤)。然后,根据划记结果,合计各组频数,如表中第(4)列所示。

将表 1-1-3 中的第(1)(2)(4)三列或第(1)(4)列拼在一起,就构成本例所指的“51 名学生测验成绩的频数分布表”。如果根据各区间(或组)的人数求得其频率,如表中第(6)列所示,那么由第(1)(4)(6)列构成频率分布表。

这里,我们把更多的内容项目统计在同一张表中,只是为便于说明其他形式分布表的编制步骤。

2. 累积频数分布表的编制。

所谓累积频数分布表,就是指用累积频数(常用符号 cf 表示)描述数据分布特征的分布表。其编制步骤大致与简单频数分布表相同,其不同之处在于,登记累积频数时从最小的一组开始,每往下一组,必须把该组频数与以上各组的频数进行累加。如表中第(5)列所示。本例中数值最小一组的频数为 1,它的累积频数也为 1;往下第二组的频数为 2,则该组累积频数为 $2+1=3$;第三组的频数为 6,则该组累积频数为 $6+3$ (或 $6+2+1$) = 9,其余依次类推。最高一组的累积频数应等于总频数。若将表中的第(1)(4)(5)列拼在一起,就构成了累积频数分布表。类似于累积频数分布表的编制,容易得到累积频率分布表。表 1-1-3 中的第(1)(4)(6)和(7)列一起构成本例的累积频率分布表。而每一组的累积频率分别乘以 100,又可以将之转换成各组的累积百分比,如表 1-1-3 中的第(1)(4)(6)(7)(8)列一起构成了累积百分比表。

表 1-1-3 五(1)班 51 名学生测验成绩分布表

组限 (1)	组中值 (X_c) (2)	划记 (3)	频数 (f) (4)	累积频数 (cf) (5)	频率 (R_i) (6)	累积频率 (CR_i) (7)	累积百分比 (%) (8)
45—	47.5	—	1	1	0.02	0.02	2
50—	52.5	┐	2	3	0.04	0.06	6
55—	57.5	正 一	6	9	0.12	0.18	18
60—	62.5	正 一	6	15	0.12	0.30	30
65—	67.5	正	5	20	0.10	0.40	40
70—	72.5	正 正	9	29	0.18	0.58	58
75—	77.5	正 正	7	36	0.14	0.72	72
80—	82.5	正	4	40	0.08	0.80	80
85—	87.5	正	5	45	0.09	0.89	89
90—	92.5	正	5	50	0.09	0.98	98
95—	97.5	—	1	51	0.02	1.00	100
总和			51		1.00		

观察表 1-1-3 可知,五(1)班学生的测验成绩在“70—”范围的人数最多,55 分以下只占 4%,95 分以上占 2%。如果将五(2)班的成绩也编制成如表 1-1-3 的分布表(表略),同样可以较快地获取这方面的信息,并可对两个班的成绩作出比较。

三、分布表的主要用途

编制分布表是对一批杂乱无章的原始数据进行分类整理的重要手段。我们可以从分布表上一目了然地获得各种数据的频数(或频率)各是多少,数据总体分布的状况如何等方面信息。在实际应用时,频数分布表、频率分布表、累积频数分布表、累积频率分布表四者既有联系又各有侧重,既可单独使用又可联合使用。具

体区别表现为:

频数分布表主要反映各组数据出现的实际频数。当我们主要对各组的实际频数感兴趣时,则可编制频数分布表。譬如,我们在制定基础教育发展规划时,需要对一个时期内的人口按年龄组进行统计归类和发展趋势作预测,那么可以将某地区人口按年龄组统计频数编制成频数分布表,从中发现人口的年龄结构状况,以此了解小学入学人数的高峰期等信息。

累积频数分布表是对频数进行累积的结果,所以,当我们需要利用分布表获知位于某个数值以下(或以上)的数据总个数时,可采用累积频数分布表。

频率分布表主要反映各组数据的相对频数或其百分比,所以当我们侧重关心各组频数的相对比例结构时,通常要编制频率分布表。譬如,在研究普教系统的教师资格考试合格率问题时,可能仅对某几门课程的考试成绩的合格率作统计或对其进行比较,但由于参加各课程考试的人数可能不一致,此时如果再按分数段进行统计归类以编制频数分布表,意义不大。而如果采用频率分布表,则可非常直观地看出在各门课程考试成绩的相应比率,有利于比较研究。

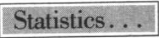
累积频率分布表,我们可以将它作为利用频率分布表进行各类情况相互比较研究时的延续应用。

【SPSS 操作技术与结果解释】

一、操作步骤

以五(1)班测验成绩这批数据为例,说明运用 SPSS 系统生成频数分布表与频率分布表的具体操作步骤。

建立或打开数据文件。[具体操作参考附录 II] → 单击主菜单中的 Analyze(统计分析) → 单击 Descriptive Statistics(描述性统计) → 鼠标箭头指向 Frequencies(频数)即出现图 1-1-1 → 单击 Fre-

quencies, 打开 Frequencies 对话框, 如图 1-1-2 所示。→ 在 Frequencies 对话框中, 选定左侧框中需要生成频数分布表的变量“成绩” → 点击  按钮, 使其进入右侧 Variable[s] 框中 → 选中 Display frequency tables(显示频数分布表) →  打开统计量选择对话框(如图 1-1-3 所示) → 在左下方 Dispersion 矩形框

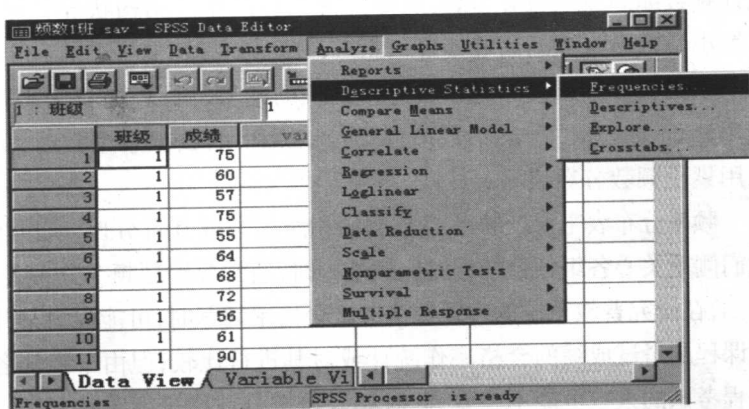


图 1-1-1 频数分布统计表菜单

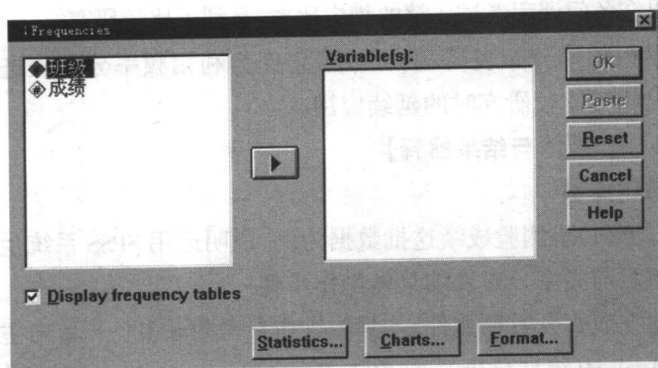


图 1-1-2 Frequencies 主对话框

中选择 Range(全距)、Maximum(最大值)、Minimum(最小值)(其他统计量的选择参见第二章) → 单击 **Continue** 按钮,返回主对话框 → 单击 **Format...** 打开频数分布表输出格式选项对话框(如图 1-1-4 所示)。其中,Order by:排序组。在该组中选择频数表中变量排列顺序,包含:Ascending values:按变量值的升序排列,这是系统默认方式。Descending values:按变量值的降序排列。Ascending counts:按频数的升序排列。Descending counts:按频数的降序排列。Multiple variables:多个变量的处理方式。如果统计分析中包含有多个变量,通过该选项组来选择对多个变量的处理方式:Compare variables:所有变量均对比显示在同一个图表中。Organize output by variables:每个变量分别显示在独立图表中。Suppress tables with more than [] categories:设置控制频数表输出的范围,默认值为 10。在本实例中,均采用系统默认值 → 单击 **Continue** 按钮,返回主对话框 → 单击 **OK** 按钮提交系统运行。

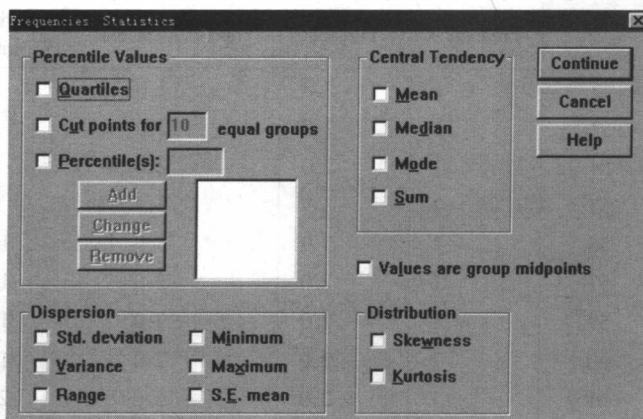


图 1-1-3 Statistics 统计量选择对话框

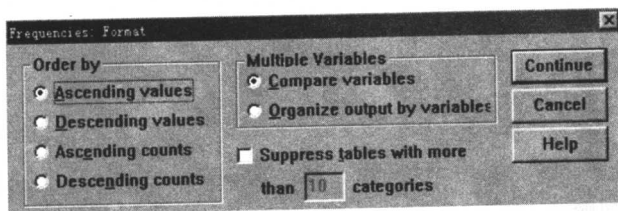


图 1-1-4 频数分布表输出格式选项对话框

二、运行结果

表 1-1-4 Statistics

成绩		
N	Valid	51
	Missing	0
Range		50
Minimum		45
Maximum		95

表 1-1-5 成绩 Frequency Table

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 45	1	2.0	2.0	2.0
52	1	2.0	2.0	3.9
54	1	2.0	2.0	5.9
55	2	3.9	3.9	9.8
56	2	3.9	3.9	13.7
57	1	2.0	2.0	15.7
59	1	2.0	2.0	17.6
60	1	2.0	2.0	19.6
61	1	2.0	2.0	21.6
63	1	2.0	2.0	23.5