# 抽样调查计算机数据文件的形成过程和质量管理

## ——兼谈社科研究人员与计算机软件人员的协调

周　基　玉

抽样调查计算机数据处理的质量管理，直接涉及到抽样调查统计结果的信度及课题调研的分析结论。为此，需要科研人员与软件人员共同努力、协调配合。本文从数据文件检查的要点、形成数据文件过程的要点、科研人员应提供的支持条件及科研人员与软件人员之间的协调等四个方面，探讨了运用计算机的研究技术对确保研究质量的重要性及具体要求。

作者：周基玉，女，1953年生，上海人口情报中心计算机房计算机软件工程师。

抽样调查是社会科学研究的基本手段之一，对调查获取的大量原始数据加以处理、分析，可以在一定程度上揭示研究对象的内在联系，并对理论假设进行检验。抽样调查课题的数据处理现在一般在计算机上完成。

数据处理分为两个部分，一是将书面原始编码表加工处理成计算机磁盘或磁带上的数据文件（即原始数据的集合），二是用数据文件进行各种统计运算。数据文件既是前期工作（调查、编码、录入）的结果，又是后期统计处理的基础。数据文件的质量直接影响到统计分析的正确性乃至整个研究课题的质量。社科研究人员和软件人员对从原始编码表到形成数据文件的过程往往重视不够，这个过程处理得不好，会费时费力，造成大量返工，甚至积重难返，影响数据文件的质量，降低统计结果的可信度，以致贻误整个课题。笔者曾多次承接抽样调查数据的计算机处理工作，深感确保和提高数据文件质量的重要性，这需要科研人员与软件人员共同努力、协调配合。本文讨论数据文件的形成过程及其质量管理，供软件人员借鉴以更好地管理质量，供社会科学研究人员借鉴以了解数据文件的形成过程和应与软件人员协调配合的地方。

## 一、数据文件检查的要点

未经检查或检查不完全的数据是没有意义的，用不知出错率为多少的数据作出的统计结果没有任何说服力。

对数据文件的检查分为整体检查、数据范围检查、数据逻辑关系检查和录入人员随机出错检查四个部分。

1. 整体检查。整体检查有两项内容，一项内容是数据文件的整体记录数应与调查表有效样本数相等，且全部有效样本的录入应无遗漏、无重复，数据文件中无空白记录。执行这项检查可用高级语言编程序，以调查表的编号（调查表中一般都有供管理用的编号，且每份调查表的编号是唯一的）作检查，即按编码的要求检查编号是否重复、是否连续、是否超出编码表编号范围等。

整体检查的另一项内容是检查数据文件中每条记录的每项数据是否有遗漏。为使此项检查便于进行、数据文件又便于阅读，在制定编码规则时，一般规定数值型数据的前导零（如“0012”中的“00”）不可省略，不能以空格作为编码值。这样，此项检查的要求就变为：数据文件中应无一格（byte）空白。

2. 数据范围检查。数据文件中的每一数据项应符合编码规则规定的数据范围，包括数值型数据值的范围和字符型数据字符的范围。数据范围的检查率应为100%，即对每条记录的每个数据项都要检查。

3. 数据逻辑关系检查。各数据项之间常符合一定的逻辑关系，科研人员应对所有要检查的逻辑关系加以详细描述。逻辑关系的检查率也应为100%，即对每条记录都要作全部逻辑关系的检查。

数据范围和逻辑关系的检查很重要。在最后形成的数据文件中，这两项内容的出错率应为零，否则统计时可能会出现不可思议的结果，例如文化程度为大学的人，培训要求却为中专，等等。

4. 录入员随机出错检查。经过上述检查的数据文件从含义上来说已没有错误，但在符合上述要求的范围内，仍有可能出现随机错误。例如“年龄”数据项的规定范围是15至50，录入员虽未使数据超出范围，却把数据“34”错输入为“35”了。

执行这项检查有两种方法。一种是将全部数据重复录入两遍甚至三遍，然后用程序对两次的数据作比较，前后对照一致的才允许进入数据文件；另一种是对已录入的数据抽一部分与编码表对照。具体用哪种方法要看科研人员要求的检查率，而科研人员规定检查率时，不仅要看课题允许的时间和工作量，而且主要应参照课题前期调查、编码的检查率，高于或低于它都是不恰当、不科学的。

经过上述检查的数据文件，若出错率在科研人员规定的范围内，就可交付使用。

还应指出一点，由于检查是对整个数据文件进行的，因此上述前三项检查不仅检查了录入过程中产生的错误，而且检查了调查过程出错和编码过程出错。尤其是当每条记录的数据量大、逻辑关系又复杂时，手工对调查、编码过程的数据逻辑关系检查是极其费时费力的，即使做了，也很难保证出错率为零，而用计算机检查则很容易做到这一点。

## 二、形成数据文件过程的要点

1. 软件人员根据科研人员提出的总体要求（包括时间、数据量、数据逻辑关系量、出错率等）作系统分析，排定时间表，确定应用软件。时间表中应包括：系统分析、设计和编程、程序测试、录入、检错改错、数据文件合成等所需要的时间。应用软件应考虑面向程序员，录入员两个方面，选择便于设计录入屏幕、检错、改错、文件合成的应用软件，一般还需自己编一些接口程序。我们是用在HPP3000小型机上运行的“VPLUS”应用软件来完成

录入和检查，再用“SORT/MERCE”软件和自己编一些用于整体检查、改错的程序来完成整个过程。

2. 根据上一步结果，设计确定输入格式和方法、文件合成方法及每一细节之间的接口等问题。在数据量（记录长度、记录数）较大、数据复杂（数据类型多、逻辑关系繁复）的情况下，应考虑所设计的每一步的可行性。

其中有几点要注意：（1）受软硬件功能限制或受编码表格式等因素的影响，有时将文件按记录或按字段分割开录入，分文件完成后再合并成总文件。系统设计时，应尽量避免按字段分割文件，因为这样的文件在合成时，两部分文件各自都要有用以连接的标识字段，才能保证正确连接，这就增加了标识字段的录入、检错、改错的工作量；合并按字段分割的文件时还要编程，其中要比较判断标识字段是否相同等等。

（2）需按字段分割文件时，要注意数据项逻辑关系检查的问题，若一条逻辑关系涉及的数据项分别在两个分文件里，这些逻辑关系就要单列出来，待文件合并后再检查，这对检错改错都不方便，应避免这种情况。这不仅是软件人员如何选择分割点的问题，亦是科研人员在设计调查表、编码表时就应予以注意和综合考虑的，应在编码表上将有逻辑关系的数据项适当编排，提供比较方便的分割点。

（3）对数据文件的四项检查分别应放在哪一步执行？（I）整体检查中的数据项空白检查、数据范围和逻辑关系检查应在每录入一条记录时就实时（录入同时检查——编者注）检查，以便实时修改（录入同时修改——编者注），若受应用软件功能限制无法每录完一条记录就检查，则应尽量缩短检查周期，比如半天或一天录入完后就检查，否则积累过多，再翻编码表、原始调查表就非常麻烦，甚至不可能，改错亦不方便。（II）整体检查中的编号检查一般放在全部数据录入完后进行，并可顺便产生数据文件编号的列表。（III）录入员随机出错的抽查，应在每个录入员每录完一定数量的数据后就抽查一次。

3. 根据系统设计的要求，进行每一步中应用软件的数据定义、程序编制，然后对每一步进行调试，直到符合要求。

4. 录入、检错改错过程。这个过程工作量大，且十分繁琐，为使这个过程简便易行，要求系统分析高质量、系统设计高质量，并要求编码表高质量。对计算机录入来说编码表的质量不仅在于书写清晰无误，而且在于编码表上字段的编排情况如何。应将一条记录的全部字段放在一张编码表上，否则可能造成需按字段分割文件的麻烦，并且在大量、反复的录入、改错过程中，对编码表翻页、查看逻辑关系都会造成很大麻烦。

5. 若是分为几个分文件录入的话，至此要进行文件合成。按记录分的文件的合成只须用操作系统下的拷贝命令就可合成，按字段分的文件的合成要用高级语言编程来完成。

6. 数据文件完成后，要交付使用时应附两个书面报告：数据说明报告、检查情况报告。

数据说明报告对数据文件存在的物理介质（包括盘片或磁带规格、格式化版本等）、文件名、文件中各字段的逻辑位置、字段中各编码值的含义等进行说明，使接受数据文件的人能够顺利使用它。这份数据说明报告是数据文件的重要档案。

检查情况报告的内容应包括：

（1）该数据文件的总数据量、总出错率；

（2）说明每一项检查内容的检查方法、检查率、出错率、改错情况；

（3）该数据文件的编号范围列表、数据值范围列表、校对通过的逻辑关系列表（在计算机上最后实现的这些内容与科研人员一开始提出的可能不尽相同）。

对软件人员来说，其工作的科学性、有效性，不仅在于其自身，而且在于前期工作提供的条件。综合前述的内容，可以看到，软件人员希望得到这样的前期结果：

（1）每份调查表有唯一的编号；

（2）编码规则中应规定数值型数据的前导零不可省略，不能以空格作为编码值；

（3）质量较好的编码表，即书写清晰，一条记录的所有字段在一张编码表上，有合适的分割点；

（4）明确的数据范围列表和数据逻辑关系列表。

## 三、科研人员应提供的支持条件

上述过程由软件人员具体管理，科研人员仅在一开始提出总体要求和数据检查要求（数据范围检查、逻辑关系检查）、抽查率、出错率等，和结束后得到数据文件及报告。然而科研人员若对软件人员的工作给予较好的支持条件（如上述），将极大地有助于这个过程的处理。

**其中数据检查的具体要求是至关重要的，它影响到形成数据文件的每个细节和整体过程，**若科研人员提出的数据范围、逻辑关系等描述得不完备的话，会引起经常性的、大的返工。从前面讲的软件人员工作步骤中我们可以看到，数据逻辑关系的修改增删将引起从编程调试到录入、检错、改错的返工，甚至要从系统分析、系统设计上进行修改，一次修改的过程中还可能有反复，可谓牵一发而动全身。若在数据文件中没有解决上述问题，待统计结果出来就更为麻烦，统计结果可能互相矛盾、无法解释，修改返工的工作量就更大，使初始时比较容易办到的事变得十分困难、甚至于不可能办到。科研人员应认真对待这一环节。科研人员以书面形式向软件人员提供的数据值范围、数据项间逻辑关系应符合如下要求：

（1）包括的内容要完整，要考虑到所有可能出现的情况。如数据项“出生年”定为2个字符时，就应考虑到有可能出现“18××”年的值，当18××年与19××年后2个字符相同时怎么区分，这种情况下“参加工作年”应大于“出生年”的逻辑关系怎么实现等等。

（2）描述的语言要严格准确，因果关系要清楚。如描述某个逻辑关系时，某几个字段“不可全为零”和“全不可为零”之间有很大差别，若描述为“不可为零”则含义不清，如此等等。

一份清晰明确的数据检查要求将给软件人员带来极大便利，使软件人员能够按既定方案一步步顺利进行，用最少的时间产生高质量的数据文件。

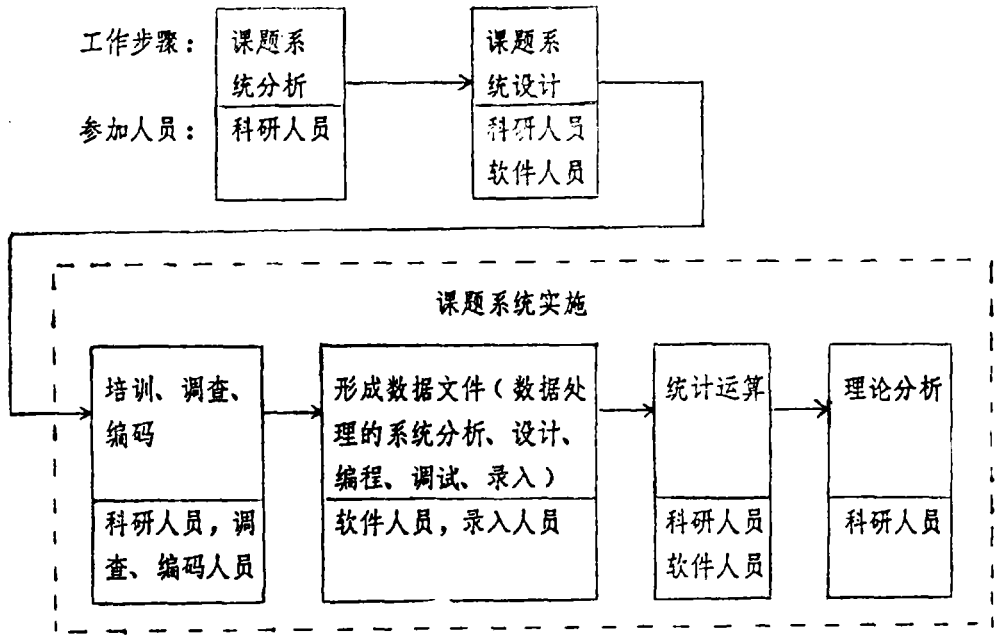
## 四、科研人员与软件人员之间的协调

除上述软件人员的工作所需要的支持条件外，还有一个问题需要协调。

软件人员往往对课题的前期工作未参与、不了解，到课题需录入数据时才得到一大堆前期结果。尽管计算机对各种要求的工作总是可以完成，程序编得长些短些最后总能达到目

的，但应知道，有些事手工做很难，在计算机上会很容易，如对数据的逻辑检查、重新排序等；有些手工很容易的事，计算机上却会很麻烦，如将两个文件连接成记录长度增加、记录数不增加的一个文件。所以前期调查、编码时对编码规则、编码格式等甚至更细小的问题的处理方法，都会影响到数据文件的形成，处理不当会造成很多不便。如我们曾做过一个课题，科研人员的设计将每条记录的二百多个字段分在十张编码表上，每张表上是二十多条记录的部分字段，这就给以后的计算机录入、逻辑校对等造成许多困难，因为软件人员若将录入屏幕设计为按字段分页录入，则逻辑校对、文件连接大为不便，若设计为按记录跨页录入，则在录入、校对、改错时手工频繁地翻页十分烦琐，而且不论用哪种方法，都需为了标识和检查每个分记录所属的总记录而多次重复录入记录编号，所增加的录入量占总录入数据量的六分之一，再加上校对、改错，浪费的工作量就更大，且出错概率也增大，而这些重复录入的记录编号最终是不纳入数据文件的，仅仅是为检错、连接文件的一时之需。若科研人员与软件人员事先稍有联系，将编码表设计为一条记录的所有字段在一张表上，是轻而易举的，不仅有助于形成数据文件，就是对登录编码的人员来说也更为便利。

由此可见，软件人员对课题的参与与科研人员应有适当的协调。图示如下：




整个课题过程中的每一步都对以后产生影响，都与最终结果有内在联系。

本文仅就数据处理中数据文件的形成这一部分做了详细讨论，我们还应认识到数据处理本身是研究过程的一部分，而不是外在于研究过程的对计算机这一工具的使用。理论研究和数据处理是互相依赖的，而保证课题数据处理质量的第一要素就是研究内容与研究技术的紧密配合，在这一意义上可以说，没有科学的数据处理就没有科学的理论研究。

责任编辑：张宛丽