经济和社会事务

人口和住房普查

编辑手册

第一次修订本

联合国

ST/ESA/STAT/SER.F/82/Rev.1

经济和社会事务部
统计司

方法研究　　F辑　第82/Rev.1号

人口和住房普查编辑手册

第一次修订本

联　合　国
纽约，2011年

经济和社会事务部

联合国秘书处经济和社会事务部在经济、社会和环境领域的全球政策与国家行动之间发挥重要的桥梁作用。该部在三个相互关联的主要领域中开展工作：（一）汇编、制作和分析范围广泛的经济、社会和环境数据及信息，供联合国会员国在审查共同问题和评估政策抉择时使用；（二）便利会员国在许多政府间机构里就采取什么联合行动应对现有的或新出现的全球挑战进行谈判；（三）向有兴趣的政府提供咨询，帮助它们采取适当方式方法把联合国历次主要会议和首脑会议制订的政策框架转化为国家方案，并且通过技术援助协助它们建立国家能力。

说　　明

本出版物中使用的名称以及材料的编写方式，并不意味着联合国秘书处对任何国家、领土、城市或地区或其当局的法律地位，或者对其边界或界限的划分表示任何意见。

本出版物中使用的“国家”一词在适当情况下亦指领土或地区。

使用“发达地区”和“发展中地区”等名称是为了统计上的方便，未必表示对某一国家或地区在发展进程中所处发展阶段做出判断。

联合国文件都用英文大写字母附加数字编号。

ST/ESA/STAT/SER.F/82/Rev.1

联合国出版物
出售品编号：C.09.XVII.11

ISBN 978-92-1-730226-8

由联合国印刷，纽约

前　　言

多年来，为了协助各国制定计划和执行经过改进且讲究成本效益的人口与住房普查，联合国出版了一系列的手册和技术报告。这些手册和报告不时地进行检查和重复，以反映新的发展动态和普查过程中出现的新问题。本出版物是为帮助各国准备2000年及未来轮次的普查而编写的系列手册的组成部分。其他手册包括：

(a) 《人口与住房普查管理手册》，F辑第83号（联合国出版物，出售品编号：E. 00. XVII. 15）；

(b) 《地理信息系统与数字制图手册》F辑第79号（联合国出版物，出售品编号：E. 00. XVII. 12）；

(c) 《人口普查中收集经济特征指南》（即将出版）。

《人口和住房普查的原则和建议》（第二次修订本）（联合国，2008年）审查了对整个普查工作至关重要的普查初期阶段质量控制与改进系统的优缺点，并且审查了应当作为整个普查方案组成部分并与其他普查计划和程序相结合的编辑计划的重要意义。《人口与住房普查编辑手册》的用户会发现，参考《原则和建议》特别有益，因为它在第三、第四和第五章提供了相当多的背景资料。

本出版物的目的在于就普查和调查数据的编辑方法向各国提供一个总的看法，同时向有关的官员提供关于如何使用各种不同普查编辑方法的信息。还想要鼓励各国保管一份关于本国编辑经历的历史，促进各种主题专家与数据处理专家之间的沟通，并且做好有关当前普查或调查活动的文献资料工作，以避免下次普查或调查期间重复劳动。

本《手册》评论了手工编辑和计算机辅助编辑的长处与短处。在大规模普查中，手工校正很难在经济上切实可行。一般通过特别设计的计算机程序，根据相关个人或住户或其他个人或住户的其他信息进行查错和缺失数值填补。本《手册》主要处理自动化数据纠错方面的问题。

计算机编辑在探查和纠正误差方面发挥重要作用。在计算机编辑阶段，可以通过与主题专家协商进行详尽的一致性检查。查出来的错误可以参照原明细表纠正，亦可自动纠正。虽然自动编辑可以加速数据处理，但须对收录的数据加以仔细控制。

除前言部分外，本出版物分为五章。前言介绍了普查程序和通常在普查中发生的各种误差类型。第一章介绍了普查编辑工作的基本原则。第二至第

五章介绍了在各个不同处理阶段编辑普查数据的程序和方法。附件介绍了技术方面需要考虑的因素，特别是与编程有关的技术方面。

虽然本《手册》侧重于人口与住房普查的编辑工作，但其中许多概念和方法也适用于调查作业。

对于哈佛人口与发展研究中心的Michael J. Levin为起草本《手册》所做的贡献，谨致谢忱。同时感谢加拿大国家统计局的Michael Bankier、Wesley Benjamin、Marcel Bureau、Sean Crowe、Sylvain Delisle和Darryl Janes，他们参与了草稿审查，并就本出版物的定稿提出了宝贵意见。

目　　录

表

图

页次

方　框

附件图片

缩 略 语

CD	已亡子女数
CEB	平均生育数
CLA	非身边子女数
CLH	身边子女数
CS	存活子女数
EA	查点区
GIS	地理信息系统
ILO	国际劳工组织
ICIDH	《缺陷、残疾和障碍的国际分类》
NIM	最近邻插补法
OCR	光学特征判读
OMR	光学标记判读
PES	查点后调查
SAS	统计分析系统
SNA	《国民账户体系》
SPSS	社会科学应用统计组合程序
UNESCO	联合国教育、科学及文化组织

导　　言

A.　本《手册》的宗旨

1.　经过精心设计且在最终产品中误差极少的普查或调查[1]是一个国家的宝贵资源。为了获得准确的普查或调查结果，数据必须尽可能地避免出现误差和不一致，尤其要避免在数据处理阶段过后出现这些情况。

在收集和录入数据期间及以后以及在调整个别项目的时候，探查数据记录之内和之间存在的误差的程序即所谓人口与住房普查编辑

2.　没有任何普查或调查是十全十美的。各国早就认识到普查和调查资料是有问题的，因而采取了各种方法来处理数据差距和回答不一致的问题。可是由于普查间隔时间较长，有关数据编辑程序的文件记载往往不太适当。因此，各国不得不为新的普查或调查重新制订早些时候数据收集活动中使用过的程序。

3.　《人口与住房普查编辑手册》的目的在于填补普查和调查数据编辑方法中存在的这种知识差距，并且向有关的官员提供有关各种普查编辑方法的信息。它还想要鼓励各国保管本国编辑经历的历史，促进主题专家与数据处理专家之间的沟通，并且用文件记载当前普查或调查期间开展的各项活动，以避免将来重复劳动。

4.　本《手册》可供参与制订普查或调查编辑规范和方案团队工作的主题专家、[2]数据处理专家或方法学家参考。它采取“烹饪书”的办法，让各国选用最适合本国目前统计现状的编辑方法。本出版物也适用于促进这些专家在制订和实施自己的编辑方案时加强彼此间的沟通。

5.　本导言分别介绍普查过程、普查中发生的各种误差类型，以及普查编辑工作的基本原则。随后几章提出了在数据处理的各个不同阶段的程序和方法。虽然本《手册》集中在人口与住房普查的编辑方面，但其中的许多概念和方法也适用于调查业务。

B.　普查过程

6.　一次人口和/或住房普查是收集、汇编、评价、分析和发表人口调查数据和/或与所有人及其生活区有关的住房、经济与社会数据的全过程（联合国，2007年）。传统上，普查是一个国家在明确规定的期限内在全国或其中明确划定的部分地区进行的。最近有些国家开始执行意在覆盖全国的连续调查的“长”表，以期在一定时期内提供完全覆盖。无论在哪一种方案中，普查都能提供特定时间点上人口与住房状况的快照。

[1] 普查是进行全面清点。调查通常只清点总人口的较小部分。这里所讲的编辑工作对这两种活动都适用。

[2] 按照本《手册》中的定义，主题专家包括人口学家、社会科学家、经济学家，以及工作在人口、住房和其他相关领域的其他专家。

7. 普查的基本目的是提供有关一个国家人口规模、分布和特征的信息。普查数据用于制订政策、规划和行政工作，也用于对教育、劳动力、计划生育、住房建设、卫生医疗、交通运输和农村发展等方案进行管理和评价。行政方面的一项基本用途是划定选区和分配施政机构的代表比例。普查还是一种宝贵的研究资源，它为有关人口构成与分布的科学分析和预报未来人口增长趋势的统计模式提供必要的数据。普查为工商业提供评估住房、学校、装修材料设备、食品、服装、娱乐设施、医疗用品，以及其他商品和服务等方面需求所需的基本数据。

8. 所有普查和调查都有某些有共性的主要特征，其中包括：(a) 准备工作；(b) 调查或收集数据；(c) 数据处理，包括数据录入（键入或扫描）、编辑和制表；(d) 建构数据库和发表普查结果；(e) 成果评价；以及(f) 成果分析。

9. 准备工作包括多种成分，比如确定普查的法律依据；编制预算；编拟日程表；行政组织工作；制图；编列居住单元清单；拟订制表方案；编写问卷；以及制订计划和培训调查、预检、数据处理和传播等方面的工作人员。

10. 查点过程取决于选择的查点方法、查点阶段的时机选择和期间、监督的层级，以及是否使用和如何使用抽样法。收集了数据之后，必须予以编码、录入、编辑和制表。数据处理既产生微观数据库又产生宏观数据库。国家普查办公室或统计局使用这些数据库进行制表、时间序列分析、绘图和制图，并且把地理信息系统（GIS）用于专题制图和其他传播方法。使用各种不同的方法，从内容和覆盖范围两方面对普查结果进行评价，其中包括人口统计分析和查点后调查。最后，通过各种方式进行成果分析，其中包括描述性成果概要、侧重于政策研究的普查成果分析，以及对有关国家人口统计的一个或更多方面及社会状况进行详细的分析研究。

C. 普查过程中的误差

11. 普查中出现的误差来自许多方面，大体上可分为覆盖范围误差和内容误差这两种类别。

1. 覆盖范围误差

12. 覆盖范围误差是在普查中对个人或居住单元的漏查或重复查点造成的。造成覆盖范围误差的原因包括：查点区域的地图或清单不完整或不准确；普查员未能查点其指定区域内的所有居住单元；重复计数；对不愿意被查点的个人忽略不计；对某些类别的个人处理不当（比如访问者或非居民外国人）；以及查点后个别普查记录丢失或损毁。覆盖范围误差应尽可能在现场得到解决。办公室的编辑过程只能消除确实重复的记录。可是一定要仔细认定这些是不是重复的个人或住户。比如孪生子女可能有相同的信息，只是序号不一致。因此，适用于这一过程的编辑规则就要确定：对于似乎重复的信息，什么情况下应采纳，什么情况下应丢弃，又在什么情况下应通过插补予以更改。

13. 第三章介绍的结构编辑检查住户人数记录正确与否、次序正确与否、以及有无重复的个人。

2. 内容误差

14. 内容误差是个人、住户和居住单元特征的报告或记录错误造成的。产生内容误差的原因有：问题设计不当或提问次序欠妥，或者普查对象和普查员之间的交流有问题，此外还有编码和数据录入错误、手工和计算机编辑错误，以及结果的制表错误。务必妥善编制编辑索引（或称查账索引）并将其储存在普查过程的每个阶段，以确保不丢失数据。现将上述误差逐一解释如下：

(a) 问卷设计失误

15. 提问或提示措词不当是产生内容误差的原因之一。最值得仔细考虑的是问卷的类别、格式，以及确切的措辞和问卷项目安排，因为设计不当的问卷存在的瑕疵是不能在查点期间或过后得到弥补的。应采取预检的做法，以尽量减少可能因为问卷设计问题而产生的误差。譬如讲，如果跳转模式不清楚或设置的位置不妥当，普查员就可能错误地跳转问卷部分，因而收集不到相关的信息。

(b) 普查员失误

16. 普查员与普查对象之间是互动关系，除非采用自填式问卷。普查员可能在提问时出错：要么缩减或改变了问题的措词，要么没有向普查对象完整解释有关问题的含义。普查员也可能会在记录回答的时候增添误差。普查员的素质和普查员培训是确保收集数据质量的重要因素。必须从普查程序的各个方面对普查员进行适当培训。要让他们懂得自己在普查过程中扮演角色的重要性，以及如何使其计数与普查的其他阶段相适应。另外，由于普查员的背景各不相同，受教育的水平也有高有低，所以制定的培训方案要有把握使普查员明了如何提问才能获得适当的回答。

(c) 普查对象的错误

17. 如果普查对象误解了特定问卷项目的话，也会给数据带来误差。由于故意答错或者由代理人回答问题（即由信息所涉个人以外的其他人代为回答问卷），也会产生误差。通过普查的宣传活动和对调查人员的培训来说明普查的目的以及各种提问的理由，可以改进个人回答问题的质量。有些国家使用自填式的问卷，所以不存在普查员与普查对象之间的互动。对于自管方式来说，在普查对象误解问卷提出的问题或提示的情况下，会出现误差。

18. 属于普查对象和普查员的错误最好在查点阶段进行处理，因为这时候调查表、普查对象和普查员都能找得到。监督员也要能够在查点期间定期检查普查员收集的数据，以确保普查员不要把系统性的偏差带入数据。监督员应在问卷被送往地区或中央办公室之前在现场处理与普查员和普查对象有关的误差。

(d) 编码错误

19. 编码过程中可能出现误差，因为编码员可能对信息错误编码。数据录入过程中的打字错误也会给数据带来误差。一般来讲，如果在这一阶段疏于监督和检验，就会延迟发表数据，因为以后的查错和纠错就困难多了。通常在编码作业之前进行手工编辑。

(e) 数据录入错误

20. 可将查点范围检查和某些基本一致性检查功能编入数据录入软件，以防无效录入。有一种智能数据录入系统可以确保每个域或数据项目都在该项目的允许值范围以内。该系统增加了数据录入操作员键入合理数据的机会，并且在一定程度上减轻了日后数据准备阶段的数据编辑负担。不过，这些检查功能可能会减缓数据录入速度。因此必须在数据录入过程中的一致性检查量和保持适当的数据录入速度之间进行仔细权衡。需要事先确立这种平衡关系，以免数据录入人员在这些努力方面花费过多的时间。对键入操作的检验肯定能提高数据质量。键入的表格可以通过重新键入同样的信息来检验，通常采取抽样的方法进行检验。附件二中讨论了问卷格式与键盘输入之间的关系；附件三则讨论了扫描与键入方面的考虑。

(f) 计算机编辑错误

21. 编辑工作是普查数据处理中的关键步骤之一。在编辑过程中通过插补缺失的数据或者替换与似乎可信数据不一致的信息，来更改或纠正无效和不一致的数据。而自相矛盾的是，任何这种编辑操作都有可能带来新的误差。

(g) 制表误差

22. 制表阶段可能因为数据处理错误或使用“未知”（未供给）的信息而产生误差。这一阶段的误差很难在不产生新误差的情况下予以纠正。由于缺少交叉制表检查，加之打印错误，所以在出版阶段会出现误差。不要试图自行纠正表格，而有必要维持数据处理系统，以便在表格中出现不一致情况的时候追加编辑。如果有关误差经过全过程的所有阶段直至被载入出版过程，错误就会显露出来，而结果就是产生可疑数值。如果在制表阶段“纠错”，譬如讲如果发现少量的杂项未知数并将其放进“总计”而未进入分配的话，这些表格就不能被其他分析者复制，其总体价值就比较小。较为明智的做法是把普查看作一个反馈系统，以便在编辑阶段而不是在制表阶段更改数据。在发布表格以前一定要进行彻底检查，以确保为所有预定的地域单位编制了全部计划内的表格。虽然在编辑阶段引进的范围检查和一致性检查可以减少大多数误差，但在制表以后进行一次总量检查（有时叫做“宏观编辑”）也是不可或缺的。要由训练有素和有经验的人通检一次各种不同的表格，以核查各个栏目中报告的数字是否与当地的已知情况相符。在限定数量的情况下，快速参照一下普查计划可以查出编码错误。对选定的比率和增长率进行计算并与先前的普查数字或其他已发表的抽样调查数字作比较，也很管用。不过，只有在所用概念可比的情

况下方可与其他基于调查的数据或行政资料数据作比较。如果在最后制表过程中发现了错误，首先应对微观数据集进行纠错，其部分理由是为了让国家统计/普查机构的其他数据处理人员能够制作可比的表格。另外，鉴于国家统计/普查机构有时候向公共和私营部门的研究工作者和其他用户释放一些微观数据文件，所以需要使表格能够复制。

23. 如上所述，普查过程涉及到许多相互关联的序列作业，而每一项作业过程中都可能产生误差。要记住，计算机编辑是反馈系统的组成部分，而计算机编辑不仅向前方给制表输送数据，而且向后方给收集资料和现场数据处理工作输送数据。国家统计/普查机构预防计算机编辑出问题的最好办法就是最大限度地发挥现场编辑的作用。国家统计/普查机构也需要确保编码和数据录入准确无误，并且在包括录入、编辑和制表在内的所有作业中间进行持续不断的反馈。

D. 本《手册》的结构

24. 第一章考察编辑工作在普查和调查中发挥的作用。其他几章涵盖各种专题。第二章提出了编辑程序的具体应用，其中包括各种缺失数据插补方法。第三章介绍结构编辑，即同时进行住房和人口项目的编辑，以及旨在帮助其余编辑工作的某些程序，比如查明每个住户是否有一个、而且只有一个户主。第四章概述了人口编辑，而第五章涵盖了住房编辑。最后由一系列附件来逐一审查与人口和住房普查的编辑和插补程序有关的特定问题。

第 一 章
普查和调查中的编辑工作

A. 编辑工作的历史回顾

25. 在计算机到来以前，普查工作大多雇用大量半熟练的办事员来逐份编辑表格。可是由于各种项目间的关系太复杂（即便只有为数不多的若干项目），要想涵盖数据中所有可能存在的不一致问题，就连简单的检查都难以启动。不同的办事员对规则有不同的解释法，甚至同一个办事员自己都不一致。

26. 随着计算机的引进，普查的编辑工作发生了显著变化。计算机能够比手工编辑探查到更多的不一致误差。编辑规范变得越来越精细，越来越复杂。自动插补已成为可能，同时有了与之并存的编辑规则（Nordbotten，1963年；Naus，1975年）。与此同时，这一进程为联系越来越多的普查对象创造了条件，或者说至少有越来越多的普查对象填写问卷。许多编辑团队开始感到，编辑工作做得越多越好，而编辑越精细，结果就越准确。软件程序产生了数以千计的误差信息，要求人工审查原始表格，或者对一些调查而言，需要重新采访普查对象。[3]

27. 有了计算机，数据集的更改就变得越来越容易了。有时候，这些更改纠正了记录或项目。许多记录多次经过计算机，而每次经过都有不同的个人检查过其中的误差和不一致之处（Boucher，1991年；Granquist，1997年）。

28. 在这整个过程中产生了几种通用普查编辑软件包，其中有些沿用至今。起初，这些软件包是为主计算机开发的；其中有些后来经修改可用于个人电脑。在这一时期，费勒吉和奥尔特（Fellegi和Holt，1976年）开发了一种通用编辑和插补新方法。这种方法虽然没有立即被投入实际应用，但是目前正在越来越广泛地被各国统计/普查机构用于日趋精细的编辑工作。

29. 1980年代普查编辑工作取得重大进步，各国统计/普查机构开始使用个人计算机进行数据录入、编辑和制表。突然间，数据处理员可以在数据录入阶段或紧接其后进行联机编辑了。对于调查和小国的普查来说，工作人员可以制订方案，以便在收集数据的过程中或在把数据直接输入计算机的同时捕捉到误差。计算机编辑让普查员能够更多地连续接触普查对象，以便于解决在编辑过程中遇到的问题（Pierzchala，1995年）。

30. 早些年，对普查和调查数据进行越来越精细而彻底的检查过程似乎做得非常成功。编辑团队创建了日趋复杂的编辑规范，数据处理专家花费数月时间来开发流程图或决策图及程序代码。分析家很少评价这些软件包。似乎编

[3] 附件五载有关于各种插补方法的讨论；附件五、附件六讨论了各种计算机编辑软件程序。

辑过程可以纠正此前数据收集、编码和录入各阶段所产生的任何问题。可是许多分析家也看出来了，在许多情况下，所有这些额外编辑对数据是有损害的，或者说至少延迟了公布结果或导致结果中出现了偏差。有时候，程序通过数据的次数太多——先纠正一个项目，再纠正另一个项目……，以至于最终结果远远偏离了最初未经编辑的数据。

31. 对于许多普查和大型调查来说，这种广泛的编辑在很大程度上拖延了普查或调查进程。办事员花费很多时间手工查找表格；数据处理专家仍在继续开发用于搜索少数情况的应用程序。Granquist（1997年）注意到，许多研究表明，对于大部分这种额外工作来讲，“质量上的改进微不足道，或者没有什么改进，甚或起反作用；有许多类严重系统性误差通过这种编辑是找不出来的”。

32. 随着各国统计机构不断发展普查和调查工作，扩大计算机编辑范围是可能的，甚至是大有可能成功的。因此，每个国家统计/普查机构都必须面临的问题是：为了达到它的目的，何种水平的计算机编辑比较合适。

B. 编辑团队

33. 在国家主管统计的机构准备普查的时候，需要考虑改进其工作质量的各种潜在途径。其中之一就是建立一个编辑团队。编辑过程应该是编辑团队的责任，其中包括普查管理人、主题专家和数据处理员。普查准备工作一开始，就应成立编辑团队，最好在起草问卷期间成立。编辑团队从一开始就很重要，而且整个编辑过程一直如此。精心组织团队，仔细制订和实施编辑与插补规则，即可保证快速有效地开展普查工作。

方 框 1

普查编辑应做哪些工作？

普查编辑要做到：

- ☞ 向用户提供高质量的普查数据；
- ☞ 查出误差类别和来源；
- ☞ 提供经过调整的普查结果。

34. 普查官员与用户界开会讨论制表和其他数据产品事宜，可以深入了解所需进行的编辑工作。用户往往要求提供某种特殊表格或表格类型，这就需要进行额外的编辑工作，以消除潜在的不一致性。编辑团队应在初期编辑阶段规划实施这些表格，而不要等到普查数据处理搞完之后在专用表格中去处理。在预检或总排演阶段制订编辑规则和计算机程序，就有可能对程序本身进行检验，并加快编辑和插补的各个不同部分之间的转换时间。然后编辑团队要摸清这些不同过程的影响并采取必要的补救措施。

35. 主题专家和数据处理专家要共同制订编辑和插补规则。编辑团队应早在普查准备阶段就制订查错和编辑计划。普查或调查编辑团队创建保持一致性和纠错的成套书面规则。

36. 除了制订编辑和插补规则之外，主题专家和数据处理专家还必须在包括分析阶段在内的普查和调查的所有阶段通力合作。编辑工作过多的风险跟编辑工作太少而在数据集中存有未经编辑的信息或假信息的风险一样大。因此，这两组专家都要负责适当维持自己的元数据库。编辑团队还必须有效利用现有行政资源和调查登记表来改进日后的普查或调查业务。

37. 在国家主管统计/普查的机构使用主计算机的情况下，主题专家和数据处理专家之间的沟通交流是有限的。在微机到来以后此种分工持续了一段时间，但是计算机程序包正在变得越来越方便用户，现在已有许多专题人员可以实际开发和检验自己的制表方案和编辑。虽然主题专家通常不处理数据，但是他们一般都懂得数据处理专家处理数据所采取的步骤。

C. 编辑实践：经过编辑的数据与未经编辑的数据

38. 各国进行普查编辑是为了改进数据和提供数据的方法。《手册》的这一节着重概括国家统计/普查机构在发布未经编辑的数据的时候所面临的问题。下面使用一套假设的数据来说明这些问题。

39. 一个虚构的国家，其全国统计/普查机构面临着设法满足多种用户需要的难题。有些用户可能想要普查数据中包含一些未知的分项信息用于分析研究，另一些用户为了规划或政策目的则可能需要最少干扰（即可能的误差）的数据。如果国家统计/普查机构发布了未经编辑的数据表，比如表1左侧的那些数据，那么无论分析者还是决策者都不得不在使用这些数据的时候作些假设。表1只用了很少的人数来说明这个问题。它表明，这个国家有23人未报告性别，[4] 有15人未报告年龄。这些漏报的原因可能是无回答，也可能是键入错误。还有两例，既漏报了性别又漏报了年龄。

[4] 在本出版物中表示性别的两个英文单词“sex”和“gender”可以互换使用。

表1
按15岁年龄组和性别分列的抽样人口的编辑和未编辑数据

	未经编辑的数据				经过编辑的数据		
年龄组	总计	男	女	未报告	总计	男	女
总　　计	4 147	2 033	2 091	23	4 147	2 045	2 102
15岁以下	1 639	799	825	15	1 743	855	888
15岁至29岁	1 256	612	643	1	1 217	603	614
30岁至44岁	727	356	369	2	695	338	357
45岁至59岁	360	194	166	0	341	182	159
60岁至74岁	116	54	59	3	114	53	61
75岁及以上	34	12	22	0	37	14	23
未报告	15	6	7	2			

40. 大多数国家都将自行决定采取何种办法来处理这种未知数。一个合乎逻辑但可能很天真的做法就是按照同已知数值一样的比例来分配未知数。如果国家统计/普查机构决定推算插补未知数，编辑团队可以决定采纳12名男性和11名女性，这是个差不多对等的比例数，但是偏离了实际，因为普查结果是女性居多。而表1右侧所显示的经过编辑的数据与实际结果比较吻合。

41. 还有其他一些处理未知数的备选方法。譬如讲，编辑团队可以决定仅根据性别分布来进行缺失值插补，而忽略其他可用的信息，比如配偶间的关系，一个不知性别的人是否报告为另一人的母亲，或者一个不知性别的人是否有已生子女人数的实际录入。另一种备选的插补方法是把上述其他变量考虑在内。

42. 可供国家统计/普查机构选择的另一种方法就是根据年龄分布进行插补。对于表1中说明的抽样人口而言，总共有15例未报年龄。也可以按照和已知数据一样的比例来分配这些缺失数据，这同样是一种合乎逻辑的插补方法。编辑团队或许还可以通过考虑其他变量和组合获得较好的普查结果，比如夫妻间的相对年龄、父母和子女间的或者祖父母和孙儿孙女间的相对年龄，或者学龄儿童、退休者和劳动者的存在。

43. 在表1中，右侧经过编辑的数据就“比较干净”，因为那些未知数已被压制了（见“经过编辑的数据”列）。表内这一侧没有未知数，因为程序把它们分配给了其他应答者。可是，传统上许多人口统计者和其他主题专家都想要在表格中显示出这些未知数，就如同表1中未经编辑的数据那样。他们相信，采取这种做法可以对普查/调查数据进行各种评价，以计量普查程序的有效性或者帮助规划未来的普查和调查。通过不带未知数的制表和带有未知数的制表，上述两项目标都可以实现：一种是供实体用户使用的经过编辑的表格；另一种是供评价用的未经编辑的表格。

44. 统计机构要尽一切努力保管好所收集的原始数据。要将一套完整的原始录入数据存档，不仅作为一部分历史记录，而且一旦工作人员决定从头开始重新编辑数据集的任何部分，亦可借助原始数据作参考。不过，某些关键项目值（比如年龄、性别和生育力）应保存在每份记录上，以便于人口统计者和其他人分析编辑结果。

45. 在公布的表格中与使用未知数有关的另一个问题，就是这些未知数可能影响到趋势分析。新技术的应用使得这种分析比以往容易多了。例如表2根据两次连续的普查结果列出年龄分布情况。对这个小国家而言，未知数减少了：报告的人数从2000年的217人或大约占应答者的6.5%减至2010年的只有15人或占不到1%。

46. 这里，国家统计/普查机构必须摸清不一致的未知数对特定普查及普查间的变化影响有多大。例如，由于2000年普查的未知数占6.5%，就很难在这两次普查之间对15岁年龄组的分布情况加以比较。在这10年当中，似乎15-29岁的人所占的百分比从27%增长到30%，但是未知数的分布可能会改变分析结果。

表2

按15岁年龄组分列2000年和2010年有未知数的人口及人口变化

	人数				百分比	
年龄组	2010年	2000年	人数变化	百分比变化	2010年	2000年
总计	4 147	3 319	828	24.9	100.0	100.0
15岁以下	1 639	1 348	291	21.6	39.5	40.6
15-29岁	1 256	902	354	39.2	30.3	27.2
30-44岁	727	538	189	35.1	17.5	16.2
45-59岁	360	200	160	80.0	8.7	6.0
60-74岁	116	89	27	30.3	2.8	2.7
75岁及以上	34	25	9	36.0	0.8	0.8
未报	15	217	-202	-93.1	0.4	6.5

47. 经过修订的表3表明未知数已经按比例或通过某种插补方法被分配了。这里就可以很容易地看出两次普查间的人数和百分比变化以及年龄组的分布情况。当然，为了获得准确可靠的结果，编辑团队要确保两次普查和/或调查之间及其内部的一致性。“未报”行被删除了。

表3

按15岁年龄组分列2000年和2010年没有未知数的人口及人口变化

	人数				百分比	
年龄组	2010年	2000年	人数变化	百分比变化	2010年	2000年
总计	4 147	3 319	828	24.9	100.0	100.0
15岁以下	1 743	1 408	335	23.8	42.0	42.4
15-29岁	1 217	952	265	27.8	29.3	28.7
30-44岁	695	578	117	20.2	16.8	17.4
45-59岁	341	230	111	48.3	8.2	6.9
60-74岁	114	109	5	4.6	2.7	3.3
75岁及以上	37	42	-5	-11.9	0.9	1.3

D. 编辑的基本原则

48. 编辑工作包括系统检查无效和不一致的回答，以及随后根据预定的规则采取手动或自动方式纠正错误（使用“未知数”或动态插补方法）。其他编辑作业涉及到使用计算机进行电子纠错。如果国家统计/普查机构不对普查或调查结果进行编辑的话，普查出版物中就可能包含一定数量的无价值数据。通过编辑过程可以减少歪曲的估计数，方便数据处理，并且提高用户的信心。另外，据Pullum、Harpham和Ozsever（1986年）认为，“编辑或整理的基本目的是检查各种回答是否彼此一致，并且符合调查文书规定的基本格式”。

49. 普查的原始数据文件中包含着各种各样的误差。数据处理把这些误差分为两大类：一是可能阻碍进一步处理的误差；二是产生无效或不一致结果但不会中断继续处理作业的逻辑流程的误差。正如在《人口和住房普查的原则

和建议》（第二次修订本）（联合国，2008年，第1.311段）中着重指出的那样，所有第一类误差都必须予以纠正；第二类误差要尽量多地加以纠正。因此，在数据处理阶段普查编辑的基本目的就是尽可能多地查出误差并修改数据集，以使数据项目有效、一致。可是处理过程不可能纠正所有普查误差，其中包括对问卷的一些回答虽然内部一致，但实际上却是普查对象的误报或者普查员的记录错误。

50. 有两类编辑：(1) 确信编辑，即确信发现错误；(2) 疑问编辑，即针对可疑数据项目进行的校订（Granquist和Kovar，1997年，第420页）。确信编辑旨在查出肯定有误的数据项目，而疑问编辑则针对有可能是无效的或不一致的数据。确信的误差，即通过确信编辑查出来的误差，包括无效值或缺失值以及属于不一致的错误。对比之下，疑问编辑鉴别大多在主要编辑范围以外的数据项目，即与同一问卷上的其他数据相比相对较高或较低的项目以及其他可疑的记录。为了维护对普查的信心，尤其在国家统计/普查机构决定公布微观数据的时候，必须通过编辑过程查出并处理确信的误差。疑问编辑的纠错比较困难，与查找和纠正确信的误差相比而言效益较少，而整个编辑过程的附加成本较高。

51. 鉴于在普查中包括的所有项目都特别是因为计划者和决策者需要它们，所以相对来讲，在普查的编辑和插补过程中要比在调查过程中解决较多的疑问编辑问题。然而，在确定普查的最终编辑结果的时候，专题人员要对试点普查的编辑和数据处理期间的编辑进行调查研究，以确保特定的编辑达到预期的成本效益。这些调研要成为普查评价的组成部分。Granquist和Kovar（1997年，第422页）指出，“在编辑过程的评价或研究中很少报告有关达标率即对原始数据修改标记数所占份额”的数据。

52. 另有一套技巧和术语与微观编辑和宏观编辑有关。如前所述，普查和调查编辑工作探查数据记录中的和记录之间的误差。本《手册》描述微观编辑，其中涉及到确保特定数据记录的有效性和一致性的方法以及特定住户的各项记录之间的关系。另一种方法——宏观编辑方法——则旨在检查综合数据，以确保其合理性。在这种方法中，表格依靠经过编辑的数据运转，并且根据预测的频率和公差限度进行检查，以鉴别数据中存在的各种问题；如果发现了“误差”，宏观编辑可以对综合数据进行总体更改，发回一个单位记录以便重新处理，或者追加新的微观编辑以更正错误。譬如讲，一个国家可能有很高比例的人未报年龄。经过年龄插补获得一个完整数据集，然后在宏观或总体层面进行检查，即可确信，有选择的老年人报告缺失不至于歪曲插补值。编辑团队可以根据这种分析结果采取措施，以减少偏差风险。不论宏观编辑还是微观编辑都需要在实施前进行彻底检验。

53. 如前所述，编辑过程要尽可能保持原始数据不变。编辑团队需要有未经改动的高质量数据，但是也需要维护本组织现场收集的数据。在计算机处理的所有阶段都需要保存原始数据，以防备编辑团队需要重新审查编辑过程。有时候编辑团队发现编辑过程中发生了系统性误差，就需要重新访问原始数

据。有时候发现部分数据集丢失或重复以至于不得不重新构建或重新编辑数据集，这种情况下也需要进行复查。

54. 有的时候误差根源是在数据处理办公室以外。Banister（1980年，第2页）指出，如果“某个子群体没有回答普查（问卷）中的某个特定问题的比例很高，这可能意味着他们没有看懂问题，或者他们拒绝回答这个问题，或者对配合普查漠不关心”。因此她极力主张在普查的储存媒体上和公布的表格中要包括相关子群体的不响应率。现在国家统计/普查机构一般都把这些数据存入光盘或其他载体，以供研究者使用。

55. 有越来越多的证据表明，不论计算机编辑量有多大都顶不上高质量的普查数据收集工作。国家统计/普查机构知道，在一定程度上计算机编辑不但是有限的，而且对数据生产会起反作用：自动纠错给数据集增添的错误比它纠正的错误还多。更改普查项目不同于纠正项目。所以编辑团队必须共同确定编辑过程的初期、中期和末期。

56. 不管编辑和插补程序能否改进数据质量，一个未经改动的数据集都会极大地方便分析和使用。这一过程从设计普查问卷开始。通常由人口统计学家和其他主题专家确定问卷内容，一般需要经过与用户群体协商。但是归根到底，普查数据不是“主要为人口统计纯粹主义者生产的，而是为范围广大得多的学者、决策者和普通百姓受众生产的”（Banister，1980年，第17页）。可是，当普查的可信度和国家统计/普查机构面临受质疑的危险的时候，获得不带有无效和不一致录入数据的普查资料是至关重要的。正如Banister所揭示的那样，“普查组织可能会援引报纸撰稿人写幽默文章的实例，或者公民愤怒谴责普查官员公布的表格内竟然出现了年仅三岁的爷爷和月票公交乘客乘坐根本没有的火车之类的怪事”。

57. 问题在于要确定距离获得优质数据集的差距有多大。如前所述，随着计算机、第一主计算机、接着是微机的陆续到来，编辑过程基本上实现了完全自动化。在许多国家统计/普查机构，主题专家实际上变成了编辑工作的积极推动者。因此各机构现正在从事以往很难进行的一致性方面的诸多检测，尤其是涉及到记录间和住户间的检查的检验。遗憾的是，微机的这一特征也带来了不少问题，而最大的问题就是过度编辑。

1. 过度编辑的弊端

58. 过度编辑在好几个方面对编辑过程有负面影响，其中包括影响发布数据的及时性、增加成本和歪曲实值。它也会使人对数据质量缺乏安全感。

(a) 及时性

59. 国家统计/普查机构的编辑工作做得越多，整个过程就越费时间。主要问题是确定增加的时间使普查产品增值多少。每个编辑团队都要对整个普查产品追加的时间和资源所产生的净效益进行编辑过程中的评估和事后的评估。

从投入的时间来看，收益是如此之低，以至于与其让用户不能及时获得信息还不如让数据有点缺陷。

(b) 财务方面

60. 同样，随着时间的延长，普查过程的费用也增加了。每个国家统计/普查机构都要搞清楚：在其增加了编辑工作量和复杂程度的时候，对于增加的工作量而言增加费用是否值得，以及它是否能够承受这些增加的费用。

(c) 歪曲实值

61. 虽然编辑过程旨在对数据质量产生积极影响，但是校订数量和复杂程度的增加也可能会产生负面影响。有时候，编辑团队由于种种原因会错误地更改项目，其中包括：主题专家和数据处理专家之间的沟通有误；在一个非常复杂、精密设计的方案中发生了错误；或者在一次编辑中多次处理一个普查项目。国家统计/普查机构想要尽可能避免此类问题。例如Granquist和Kovar（1997年）指出，使用既定的夫妻间年龄差推算插补一对夫妇的年龄可能是有益的，但是如果这种情况太多，就有可能导致数据出现偏差。

(d) 虚假安全感

62. 过度编辑可能会给国家统计/普查机构的工作人员一种虚假的安全感，尤其在有关机构没有实施并且用文件记载质量措施的情况下。另外，不管团队做多少编辑工作都免不了会在普查制表过程中出现一些不合规范的结果，所以有必要提醒用户或许会发生一些小误差。在许多国家公布抽样微观数据的情况下尤其是这样。国家统计/普查机构都不希望释放有碍计划工作的数据，所以必须格外用心确保所有关键变量都经过妥善编辑并可供规划之用。举例来讲，任何国家统计/普查机构都不想发布带有性别或年龄未知数的微观数据或表格。另一方面，像残疾或识字之类的统计变量少做一些编辑也无妨。由于国家统计/普查机构不能编辑所有成对的变量，所以在交叉分组列表中难免会存在一些不一致之处，但是编辑团队应该检查最重要的组合。当编辑团队发现有不一致的时候，要有矫正程序可用。

2. 未知数的处理

63. 编辑团队要在普查规划初期就确定如何处理“未报”或未知的情况。如前所述，表格中未知数所在的列数和行数既不提供信息也没有什么用，所以大多数计划人员倾向于推算插补这些数据。在未知数未经处理的情况下，许多用户在作为结果产生的表格中按照已知数据的比例分配未知数，即所谓事后插补未知数。此外，有些标示信息的无效值不能作为空白处理，而必须给予说明。譬如讲，“南美洲”作为出生地必须指明是哪个国家（比如秘鲁）。编辑团队需要确定如何系统处理未知数。

3. 假改变

64. 在国家统计/普查机构制订其编辑规则的时候通常不使用模型工作，尽管可以这么做。编辑团队制订的规则要适合实际人口或住房特征。所有数据都要达到规则要求。譬如讲，一套规则可能要求户主的子女至少要比户主小15岁。可是户主的一个子女实际上可能不是亲生的：他（她）或许是户主配偶的亲生孩子。因此年龄差别也许就不到15岁。大多数国家的计划者都不在计划中区分“子女”和“继子女”；有鉴于此，在上述情况下如果编辑规则改变子女的年龄，就可能在诸如教育程度、劳动参与率等领域产生不一致。因此，在充分实施此项规则之前应当先进行试验，看其结果如何。

4. 确定允许误差

65. 编辑团队必须制订每个项目——有时是项目组合——的“允许误差”。允许误差指明编辑团队在采取补救行动之前所能允许的无效及不一致回答的数量。比如在一次普查中，对于大多数项目而言，不论因为何种缘故未能提供“可接受”回答的普查对象所占百分比都不会太大。对于某些项目来说，比如年龄和性别，允许误差可能相当低，因为它们与太多的其他项目组合用于制订计划。在回答项目缺失或不一致的比例很低（不到1%或2%）的情况下，任何适当的编辑规则都不大可能影响到对这些数据的使用。而如果这种比例很高（5%到10%，甚至更高，视情况而定）的话，简单的（甚或复杂的）插补就可能扭曲普查结果。

66. 为了最大限度减少缺失的回答，国家统计/普查机构应确保普查工作人员尽一切努力在现场获取信息。如果某个特定国家决定有的情况下——比如对识字水平或残疾——之类的项目它不需要抬高的精确度，那么这些项目的允许误差或许就高多了。有的时候，编辑团队可以通过指派普查员返回现场、或者进行电话再访问、或者运用他们所掌握的某个方面的知识来纠正错误过多的项目。不过，鉴于重返现场或采取其他后续行动成本太高，国家统计/普查机构往往可以决定：要么不用这个项目，要么仅在附有谨慎说明的情况下才用它。

67. 问题在于由谁来确定特定项目的允许误差。或许得由包括主题专家和数据处理专家在内的编辑团队来决定相关的允许误差。主题专家肯定要在一定时期内使用有关项目，因此从专业角度来讲能否获得最高质量的数据与他们利益攸关。而数据处理专家则可能发现，实际上他们不能制订适当的编辑方案来把允许误差降至可接受的水平，或者数据本身就不允许任何方案顺利保留在允许误差之内。

5. 在编辑过程中学习

68. 在进行数据编辑的同时，需要详细分析正反两方面的反馈信息，以便改进当前普查或调查及未来普查和调查的质量。编辑团队要不断地总结哪些方面做得比较成功，哪些方面没有产生积极效果。他们还必须确定那些正常运

行的编辑程序是否可以进一步改进和精简，以便用户能够更快地获得数据。在普查过程中，国家统计/普查机构越早发现错误，其纠正错误的可能性越大。

6. 质量保证

69. 质量保证在普查工作的所有阶段都十分重要。因此，当然应该建立正规质量保证机制来监督计算机编辑和插补阶段的进展情况。审计索引、绩效计量和诊断策略都是进行编辑质量分析和加快编辑进程所不可或缺的（Granquist和Kovar，1997年；加拿大国家统计局，1998年）。

7. 编辑成本

70. 本《手册》可以帮助各国降低在完成普查或调查数据编辑和插补所花费的时间与资源方面涉及的高成本。正如Granquist和Kovar（1997年，第418页）所述，甚至“1990年代的编辑成本跟1970年代一样高，尽管经过持续不断的技术开发利用编辑过程有了很大改进”。对大多数国家来说，编辑活动在花费时间和经费方面的成本高得不成比例。所以每个国家都必须摸清其投资回报率。据这两位作者估计，1990年代初期，住户调查的编辑成本约占全世界普查总预算额的20%。

71. 过度编辑可能拖延普查出成果。虽然国家普查/调查工作人员对这方面的普查经验已经有所耳闻，但是Pullum、Harpham和Ozsever（1986年）的研究发现，世界生育力调查的机器编辑使得发表成果的时间拖后了大约一年。或许国家统计/普查机构最好将其投资重点放在提高普查或调查的查点方面。

8. 插补

72. 插补是解决编辑过程中查出的回答数据缺失、无效或不一致问题的手段。插补就是通过改变正在编辑的一份或多份记录中的一个或多个回答或缺失值来解决上述问题，以确保获得似乎可信的、内部一致的记录成果。接触普查对象或对问卷进行人工分析可在编辑过程的早期阶段消除一些问题。然而一般来说，由于涉及到应答负担、经费和任务期限等制约因素，所以不可能在早期阶段解决一切问题。于是就通过插补手段来处理其余的编辑失误，因为生产包含插补数据的完整而一致的文件是可取的办法。有条件存取微观数据并掌握良好辅助信息的编辑团队成员最适合进行插补工作。

(a) 经过插补的记录要非常接近有失误的编辑记录。通常，最佳做法是尽可能少地插补变数，以便尽可能多地维持普查对象的原始数据。基本设想是：普查对象的回答只有一两处、而不是更多处错误的可能性较大（实际情况未必如此）；

(b) 插补的记录要能满足所有编辑需要；

(c) 编辑团队应设置插补值的标记，并且清楚地说明插补的方法和数据来源；

(d) 编辑团队应保管各记录领域中未经插补的和经过插补的数值，以备评价插补的程度和效果。

9. 存档

73. 用文件记录普查或调查工作的全过程，然后将这些文件存档，是质量保障程序的组成部分。国家统计/普查机构既需要保管经过编辑的数据文件又需要保管未经编辑的数据文件，以备日后分析之用。有些程序，比如诸多扫描格式，会自动保持原图像。同样，在键入批量数据之后要立即并置链接和保存数据，以供潜在的分析使用。但是无论使用哪一种程序，都有必要将未经编辑加工的原始文件副本存档。实际上，未经编辑的数据副本应分别保管在国家统计局的几个地方，以及国内其他地方，同时还要在国外保管。

74. 文件记录要完整，以便于普查或调查的计划者日后能够重建同样的程序来保证与考虑中的普查或调查的可比性。程序和结果必须可复制。最后，未经编辑的数据和经过编辑的数据都要储存在几个地方，并且采取适当措施以保证其在一定时期内可供实际使用。

75. 其他章节提到，部分文献工作涉及到两类编辑报告。第一类报告提供概要统计，其中包括误差数量和百分比（基于适当的分母，比如居住单元总数、总人口、工作年龄人口、成年女性人口，等等）。第二类报告包含至少一个“案例”结构样本，其中包括未经编辑的住户或住房记录、按居住单元或单元内个人分列的误差列表及其解决情况，以及经过编辑的居住单元或住户记录。

76. 应按逻辑地理层级（当然按主要行政区划）提供两套误差列表；另一方面，按照地理层级的较低层面提供误差列表有助于处理在普查员培训和质量管理方面遇到的问题以及与查点有关的其他问题。

第 二 章
编辑应用程序

77. 本章综述编辑和插补程序的应用软件。它提供了一个从扫描或键入原始数据、结构编辑和内容编辑、直至提供经过编辑的数据集的整个普查或调查编辑工作综合流程框架。[5]选取了一些实例来说明未经编辑的数据可能会向用户提出什么类型的问题，以及为何经过编辑的数据较为有用。本章考虑了与初步编辑过程有关的键入和编码问题。另外还考虑了计算机编辑中的一般问题并且就一些话题（比如检查有效性和一致性）提出指导原则。对计算机编辑的两种方法——静态插补（冷卡）法和动态插补（热卡）法——进行了详细评述。

78. 普查数据集是扫描还是键入，应遵循某种通用的流程。普查编辑团队从未经编辑的数据入手。在大多数情况下，所有数据都被普查员或办公人员预先编了码，因此数据集已经准备好可以进行结构编辑了。有的情况下需要通过操作把经过扫描的数据转换成另外一种可以用计算机进行编辑处理的形式，这要取决于使用何种编辑软件包。有的情况下，扫描的数据也需要进行第二次自动编码作业，以便将诸如出生地、产业和职业之类的项目填入。

79. 无论在哪种情况下，未经编辑的数据都要以能够用计算进行结构编辑的形式出现（详见第三章）。结构编辑需要检查落实所有主要行政区划都按地理或数字顺序显示，而在每个主要行政区范围内都按地理或数字顺序显示各个次级行政区划。然后，在每个次级行政区划内都要按地理或数字顺序显示各个地方。这一程序一直进行到最低地理层级。正如下一章所述，必须制定适当的程序以确保每个居住单元在数据集中出现一次，而且仅出现一次。

80. 结构编辑还必须确保所有记录类别都在适当时出现，而任何记录类别都不应在不该出现的时候重复出现。因此，对于一次人口与住房普查来说，人口记录或者住房记录都可以先出现，但是整个数据集要自始至终遵守习惯。大多数情况下应只呈现住房记录，因此要处理掉多余的记录；程序员还必须向没有住房记录的住户提供住房记录。同样地，凡有人住的居住单元（通常按照住房记录中的定义）都必须有人口记录；没人住的单元就不能有人口记录。

81. 关于结构特质，有必要提请注意：由于在普查的各个过程中出现差错，最终还将在内容编辑期间（乃至此后）重新运行结构编辑。这在普查中属于正常情况，应当有所预料。因此应将相应的时间、人力和设备要求预置于整个系统之中。

[5] 在原本为2000年的普查工作编写本《手册》的时候，大多数国家都是键入本国的数据。现在，大多数国家都采用了扫描的办法，有时候辅以键入数据的后续行动。甚至在编写本《手册》的过程中，新技术也在不断涌现，于是人们得以使用个人数字助理（PDA）和互联网来进行数据收集和交互式编辑工作（例如见Ireback (2000年)）。正如2000年代初期技术欠发达国家在扫描方面有困难一样，目前有许多国家发现其个人数字助理的应用尚待改进。

82. 然后就可以着手进行内容编辑了。每个人口和住房项目都必须单独考虑，通常也综合考虑，以确定每个有关项目的有效性和各种项目当中最适合的项目。第四章和第五章涉及到《联合国人口与住房普查原则和建议》（第二次修订本）中的各种人口和住房项目。

83. 完成了内容编辑之后，要建立经过完全编辑的数据集。未经编辑的数据要分别保存在几个安全的地方，而重要的未编辑项目（或者所有未编辑项目）也应出现在各类记录的末尾。这里有必要再次指出，编制完表格以后可能需要重新运行内容编辑，以解决特定交叉列表所产生的具体问题。

84. 普查和调查编辑工作目的在于找出数据记录中的缺漏和不一致之处。而插补程序就是用来纠正这些缺失和不一致数据的。编辑工作确立了各种专用程序来处理缺失数据和各种不可接受的录入数据。通过插补更正了输入的无效数据，并可解决数据集内存在的不一致问题。其产品是一套编辑之后用于制表的微观数据档案，其中载有可接受并且总体保持一致的每个居住单元和个人的所有应用项目的计数值。

85. 有必要再度强调，不论做多少编辑工作都取代不了高质量的查点活动。在使用各种插补方法处理随机缺漏和不一致数据的情况下，编辑过程能够发挥有效作用。可是，如果在收集数据阶段发生了系统性的误差，不管编辑工作程序多么精密复杂，都不可能实际改进数据质量。调查话题的选择对于确保获取数据的质量具有核心重要意义。要让普查对象在接受采访的时候愿意并且能够提供适足的信息。因此或许有必要避免涉及可能会引起人们恐惧、地方偏见或迷信的话题，以及过分复杂以至于在人口普查的背景下普通受访者难以回答的问题。对需要回答的每个问题精准措辞以获得最可靠的回答，将不可避免地取决于国情，并且应在普查之前进行充分检验。因此，最为重要的是国家统计/普查机构要划拨足够的资源，以获得最高质量的普查数据。

86. 为了实施编辑过程的计算机编辑阶段，编辑团队要编写书面编辑指南或说明书、决策表、流程图和伪代码。如图8所示，伪代码是一套书面编辑指南或说明。

87. 流程图有助于主题专家了解各种变量之间的联系，并便于撰写编辑指南。附件四中给出了流程图范例。主题专家与计算机专家合作撰写编辑指南，说明每个数据项目需要采取的行动。编辑指南要简单明了，不要模棱两可，因为它们是编辑程序包的基础。

88. 整个普查编辑团队——不论主题专家还是数据处理专家——都要广泛参与人口统计数据处理和分析工作。因为不合格的人员可能会无意中给普查增添误差和偏差。

A. 编码方面的考虑

89. 如前所述，在二十世纪后半叶的大部分时间里各国都采用键入数据的办法。虽然现在各国大多采用扫描办法来处理普查数据，但它们依然继续键

入数据。甚至在扫描表格的时候，仍然需要把某些变量转换成数码。产生可计算机处理的数码和字母数字的过程叫做编码。

90. 有些编辑软件包可以很容易地接受和运作字母数字的数据，但是在包含非数值数据的情况下，大多数软件包都会在分类、求和，以及导出百分比、中位数等方面遇到问题。

91. 要尽一切可能避免使用完全由字母组成的或者由字母与数字组合而成的代码（叫做字母数字）。在扫描表格的时候，字母数字不会构成大问题；另一方面，对于许多计算机软件包来说，它们的使用需要进行大量操作，或者至少需要占用大量空间。许多编辑程序只有在将字母置于括号中间或者用某种其他方式使之变得特异的情况下才能处理字母。

92. 在制订编码方案的时候，普查和调查工作人员要考虑到投入时间、精力和资金的回报问题。对于小国家或小型调查来说，编码没有什么重要意义，因为需要的处理工作量要比普查小多了。被扫描的数据受到与增加信息栏数有关问题的影响也不太大。

93. 另一方面，譬如讲，如果普查或调查为关系项目使用两栏而不是一栏的话，扫描就会带来误差，而在单一信息栏的情况下就不会有这种误差；就是说，对代码组1至9，扫描员可以取一个字母、或一个空白符、或一个转换为可读字符的杂散标记。不过，正如本出版物进一步所述，这些问题很容易在校订过程中得到解决。

94. 可是如果有两个分栏，比如代码组1至10，那就有引进一系列新误差的风险。现在，法定值不是1至9，而是有即将输入的新数值，其范围在0至99的任何位置，另外还有前面提到的字母、空白符和杂散标记。当编辑人员收到一个数值13的时候，必须着手对如何处理该数值进行战略决策：它的意思是不是3，因而1是错误的？它是否指10，因而3是错误的？在大多数情况下，主题专家都对项目提供编辑说明，但是这些数值的存在自动增加了编辑的时间和复杂性，并可能降低最终数据集的质量。

95. 最常见的问题之一，也是本出版物将在后面讨论的一个问题，是在生育力系列的分栏中产生的。现在许多国家收集有关住户中的儿童、别处的儿童和死亡儿童的信息，有时也收集这些儿童的合计信息，并且按照儿童性别分列。这样，各国就可能有多达12个项目。这里的问题是每个这种项目应该使用多少位数。在使用两个分栏的情况下，住户中的男孩数可在0至99范围内的任何位置上；在只有一个分栏的情况下，这个数字则只能介于0到9之间。可是鉴于在住户中一位女性不大可能生育9个以上的男孩，使用两位数就很有可能动用杂散标记或扫描误读——譬如把0误读为9，结果就成了总共有91个子女而不是01个子女。

96. 因此，对于住户中目前住在别处或者已经死亡的男孩和女孩来说，选用单栏大概是最适合的。可是对于住户中的子女总数、别处的子女总数、死亡子女总数和全算在内的子女总数来说，选用两个分栏或许比较合适。很大程

度上这要取决于一个国家的生育率水平。偶尔，一个不同寻常的住户会实际拥有九口人以上，均属于特殊类别；但是在普查工作中，统计机构总得尽力在误差和有用数据之间取得平衡。

97. 对于序数变量，可以考虑下列关系代码组系列：

1 户主
2 配偶
3 子女
4 同胞兄弟姊妹
5 父母亲
6 孙儿孙女
7 其他亲属
8 非亲属

对大多数国家来说，这一套标准代码组涵盖了大部分住户成员关系。有的国家给户主增加了一个“0”码，这样即可给住户的其他成员增加第10个类别了。

98. 这些代码组可用以获取住户构成信息（见附件一中的导出变量）。不过有许多国家，尤其是流行艾滋病毒/艾滋病的国家，需要提供比这些代码组多得多的详细信息。这些国家可能需要有关继子女、继父母、祖父母、侄女侄儿等住户成员的信息。在这种情况下，统计机构将需要更多两位数代码组才能履行其职能。

99. 在一个国家决定采用多栏编码的情况下，它也需要确定如何使用这些分栏。在上面举出的例子中，假定成员关系代码组是按顺序排列的。可是一旦决定使用双栏，负责这个项目的主题专家可能决定授予各个分栏本身的重要性。比如：

10 户主
11 配偶
12 同胞兄弟姊妹
13 同胞兄弟姊妹的配偶
21 子女
22 领养子女
23 继子女
24 侄儿/侄女
31 父母亲
32 岳父母
33 伯父/伯母、叔父/婶母
41 孙儿孙女
77 其他亲属
88 非亲属
90 集体户人口

100. 该方案在第一栏给出了世代编码：1代表户主这一代人；2代表其下一代；3代表其上一代；4代表下两代，等等，然后用数字表示每一类别中的亲属种类。虽然这些数值可以帮助重新构建家庭，但是办公人员和某些普通用户会发现它们使用起来非常不方便。

101. 不过，此类编码可以考虑用于某些社会-经济变量。举例来讲，关于种族隶属，第一个数字代表主要部落或族群；第二个数字代表较小的部落或族群。如果有10个以上较小族群存在，那么第一栏就显然需要使用两位数。

102. 同样，对于需要三、四个数字的项目来说，比如职业或产业，第一个数字代表主要职业或产业；第二个数字代表次要职业或产业；第三个数字代表特殊职业或产业。鉴于大多数国际编码方案都有用代码组表示的层级，统计机构就无需做更多的工作了。

103. 在国家统计/普查机构为编辑程序和此后的制表编制代码组清单的时候，可能希望为某些项目建立通用代码组。举例来讲，在许多国家，地点代码组（出生地、父母出生地、原籍、工作地点等）、语言、种族隶属/人种，以及公民身份等都非常相似。一个通用“地点”编码方案或许可以编为三位数代码组，其中第一位代表洲，第二位代表地区，第三位代表特定国家。国家统计/普查机构也可以使用由国际组织（比如联合国统计司）编制的国家数字代码组（联合国，1999年）。为密切相关的变量编制的一套通用代码组可以减少编码错误，并且有助于数据处理员进行编辑。通用代码组还可以在适当时让数据处理员能够使用出自一个项目的录入数据来判断另一项目的录入数据。

104. 编码结构能够方便编码过程以及日后在编辑、制表和分析中进行加工处理。对于拥有众多移民或族群的大国来讲，基于洲、地区和国家并各赋予其不同代码或数字的代码组要比简单列表更可取。

105. 图1提供了诸如出生地、公民身份、语言和种族隶属等项目的代码组实例。对菲律宾来讲，讲伊诺卡洛语和他加禄语的菲律宾人的代码与菲律宾语言的普通代码是不同的。依特定国情而定，这些代码组本身也是彼此有别的。虽然英语的代码是单一的，但是有不止一个族群讲英语。因此，加拿大与美国的出生地、公民身份和种族隶属的代码组相差悬殊。对于出生在法国的人来说，只要有法国公民身份、讲法语并且属于法国人种，就使用同样的代码组。因此，如果这些项目当中有一项缺失而且编辑团队认为合适的话，数据处理员可以将其他人的录入代码移过来。

106. 如果一份问卷上的一组项目不是相互独立的，那么普查/调查工作人员或许就不必把它们全都问到。编辑团队必须根据具体情况决定什么时候直接使用其他项目赋值，什么时候使用其他可用的变量。

图1
某些项目的通用代码组实例

项目类别	出生地	公民身份	语言	种族隶属
法国/法国人、法语	10	10	10	10
西班牙/西班牙人、西班牙语	20	20	20	20
拉丁美洲	25	25	20	25
菲律宾/菲律宾人、菲律宾语	30	30	30	
伊诺卡洛语			32	
他加禄语			32	
英国/英格兰人、英语	40	40	40	40
加拿大	50	50	40	50
美国	52	52	40	52

107. 如果不同普查之间（或者普查与调查之间）对譬如讲工作或种族划分的定义不同的话，就会发生另一个问题。国家统计/普查机构必须决定如何把当前编辑数据的和以往普查数据集的这些变更考虑在内才能表明趋势。如果备有未经编辑的原始数据，数据处理员即可修改适当的校订并全部重新运行。

108. 举例来讲，如果只鉴别少数情况的话，一个欧洲国家可以使用单一代码来代表所有南亚国家的原籍国。可是由于迁移方式发生了转变，下次调查或普查可能就需要在数据处理的全过程使用单独的代码组来分别代表印度、孟加拉国、巴基斯坦、斯里兰卡，以及其他南亚国家。

B. 手工校正与自动校正

109. 普查数据的手工编辑工作可能要花费数月乃至数年，而且会带来许许多多的人为误差。相对于计算机编辑而言手工编辑是下策，部分原因是不可能创造或再造一个手工校正过程的编辑轨迹。计算机（或自动）编辑减少了所需花费的时间，同时降低了发生人为误差的几率。计算机编辑和手工编辑都通过查找可接受数值来检查输入项的有效性，但是计算机程序还对照相关的输入项来检查输入数据的一致性。最后也是最为重要的一点，就是自动化的编辑可以产生一条编辑轨迹，因而是可以复制的，而手工编辑不可复制。

110. 在计算机输入的早期阶段，对输入项进行编辑是不可能的事；也就是说，所有校正要么作为编码和校验的办公业务的组成部分，要么作为键入数据之后的计算机操作的组成部分，不得不完全靠手工操作。较新型的软件包具有内置编辑功能，所以能够确保不输入无效数据——除非打字员强行输入；同时能够标出不一致之处，以便打字员手动纠正，或由计算机编程员来纠正。随着扫描技术日益普及，也重复发生了上述进化过程：在扫描技术发展的早期阶段，不能在录入期间进行编辑；但是近年来已将有效性编辑以及数据转换和重新编码等功能嵌入扫描系统。

111. 在普查和调查工作收集海量数据的情况下，工作人员不一定能够查询原始文件来纠正错误。即便能找到原始问卷，上面录入的数据有时也是错

误的或不一致的。计算机编辑和插补系统能够立即纠正或更改错误数据并报告所有发现的误差和所有已完成的更改。计算机编辑要仔细策划以节省工作人员的时间，用于其他数据处理活动。虽然通过计算机系统处理大量数据也很费时间，但是不像手工校正那样耗费时间。

112. 有好几种手工校正形式。举一个有关性别回答中的误差例子：监督员检查一位普查员的工作时发现一个明显错误，比如把一个名叫“玛莉”的人指定为“男性”。在把性别改成“女性”的时候，监督员进行了手工操作。如果监督员不改正问卷而将其发送给现场办公室，办公人员可能会发现问题并手工校正。中央办公室在编码时，编码员可能会发现这种性别与性别之间的错误搭配，然后进行手工纠错。或者，编码员可能没有发现问题，但是在打字员录入问卷数据时，可能会注意到姓名和性别不匹配问题，并且在键入之前先进行手工纠错。

113. 然而，倘若没有注意到错误，而打字员输入的编码是“女性”，那么在这一点上接下来就是一系列不同的规程。对于像生育力信息块这样的与性别有关的项目，编辑程序可能会标记这样的事实，即：这是一个有生育力信息的男性，并且在打字员输入数据时在屏幕上显示这一信息。然后打字员即可查找问卷，发现果真是个女性并随即予以手工改正。或者，如果国家统计/普查机构使用一种独立于键入过程的编辑程序，计算机程序就会标出这是个有生育信息的男性。然后，通过利用地理信息，办公室工作人员可以在储藏箱中找到原始问卷，将其抽出并确认该普查对象，即名叫“玛莉”的人被错误地报告为“男性”。在这一点上，办公人员可以将此信息发回到打字员；打字员可以找到记录并予以改正。

114. 这个例子说明了手工编辑的利弊。在上述任何步骤上普查工作人员都可能注意到这个错误——即姓名和性别不匹配的问题——并予以纠正。不过，采用手工编辑的国家统计/普查机构大概在每个阶段都有工作人员负责校验此种关系。在这方面的工作上花费了巨大能量，而结果却收效甚微：尤其从总体来看，与指示工作人员不进行手工编辑的结果相比没有什么两样。

115. 本来，在数据集中进行校正的唯一办法就是手工改动。有些国家仍然对使用自动校正法感到不放心，所以在上面讲的某一阶段采取手工校正的做法。如果数据集很小，或校正时机不大要紧，或工作队是劳动力密集型的，那么在许多情况下手工校正也行。好处是，如果问卷上提供的信息既完整又准确，而且不一致的问题实际上看一看表格就能解决，那么普查或调查质量或许会略有改进（比如编辑团队要假定“玛莉”不是“加利”；如果出现生育力信息的话，实际上假定这是为此人收集的信息——收集的信息没错）。事实上，编辑和插补程序很少能改进数据收集的工作质量。此类程序只能改变某些要素。

116. 有的时候，通过查看问卷进行手工校正没有什么效果。无论原因何在，信息不在那里。有的时候，一个人不愿意说出他（她）的年龄，所以问卷上的这个项目就留下空白。在这种情况下，靠审查问卷解决不了这个问题。于

是，编辑团队必须决定如何处理这种情况。为了手工校正，国家统计/普查机构必须要么给“未知数”赋值，要么使用一套数值来赋予年龄项以数值。

117. 除非与普查对象沟通，否则人工校正会不可避免地降低质量和一致性。它既耗费时间，又增加成本。计算机不知疲倦，且运行速度快多了；它们不会发生可能干扰维护质量或一致性的人为问题；而且在大多数情况下会使数据处理更加便宜。现在大多数国家都采用某种自动校正方法。

118. 缺失和不一致的回答会降低数据质量，并且难以提供易于理解的普查表格。有些用户倾向于在制表中把缺失和不一致的回答归入“未报”类，而其他用户则倾向于把这些情况按比例分配到已报告的相一致的输入项目当中去。还有的用户提议制订插补规则，以便用“可能的”回答来替代缺失或不一致的回答。通过使用计算机，就能根据问卷中的其他信息或者根据有类似特征的其他个人或居住单元报告的相关信息行之有效地插补缺失或不一致的回答。

119. 鉴于计算机能够查找许多特征，编辑过程中应当利用这一特点。这样，涉及诸多相关特征的编辑规程就可能导致插补比简单编辑所能产生的更适当的回答了。另一方面，设计水平不高的编辑程序可能导致产生质量低劣的普查数据。编辑团队要由来自相关学科的富有经验的主题专家和数据处理人员组成。编辑团队成员要精心选择用以检验一致性的变量，以确定对编辑和插补程序的技术要求。程序输出数据应包括更改或插补应答的百分比。然后分析家就能更好地判断数据质量，譬如讲，如果插补的百分比很高，就意味着警告须谨慎使用数据。

120. 编辑（或审计）轨迹可以显示对每个变量所做的修改。这种轨迹用来跟踪收到数据以后贯穿编辑和插补全过程的应变记录。

C. 数据校正的指导方针

121. 不论手工编辑还是自动编辑，都要通过消除缺漏和无效输入项以及更改不一致的项目，使数据尽可能接近于代表现实生活。

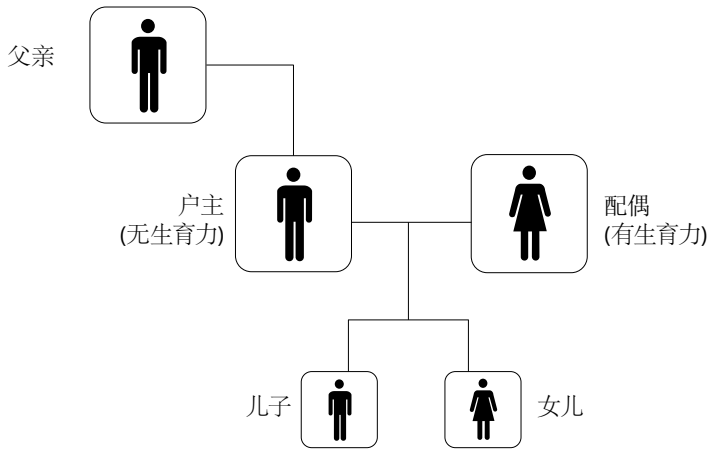
方 框 2

数据校正的主要指导方针

- 根据既定的编辑规程，切记有关数据校正的下列建议或许是有益的：
- 最大限度地减少对原始记录数据做必要的改动；
- 消除输入项目之间的明显不一致；
- 依照明确规定的程序，通过使用相关居住单元、个人或该住户或类似群体中其他个人的其他输入项目为引导，为错误或缺失项目提供录入数据。有些情况下，“未报”类别对某些项目是适用的。

122. 现以下图（图2）为例来说明特定住户。该图显示了一个有相容关系和性别输入项目的住户。户主为男性，没有生育力信息；配偶为女性，有适当的生育力信息。

图2
一个包含了成员关系、性别和生育力信息的典型假设住户



123. 可是在许多情况下，信息是不一致的。于是便提出了如下问题：对于有输入项目不一致的住户应如何编辑？如果户主和配偶都被报告为男性（像图3那样），编辑团队如何处理？以往，典型的编辑规则会假定，夫妻当中第一人为男性（尤其是如果此人是户主的话），而第二人或配偶为女性。

124. 如果户主恰好是妻子而非丈夫，那么适用的规则就是错误的，而国家统计/普查机构最终将会产生四种错误：

(a) 户主的性别是错的；

(b) 配偶的性别是错的；

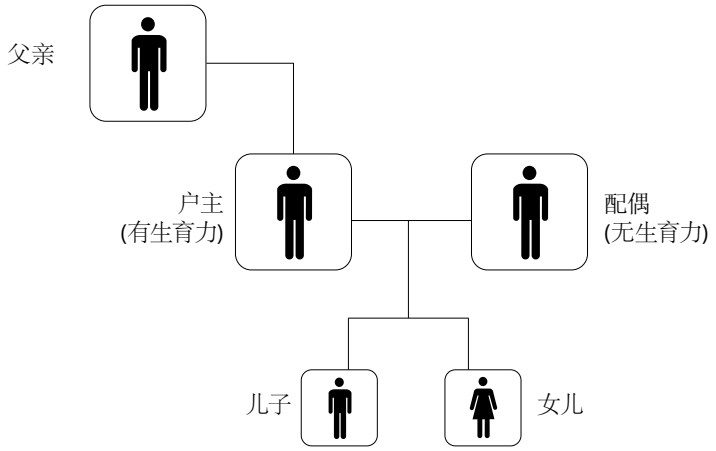
(c) 户主的生育力信息被删除；

(d) 男性配偶被错误地赋予了生育力。

这显然不是好的编辑程序。

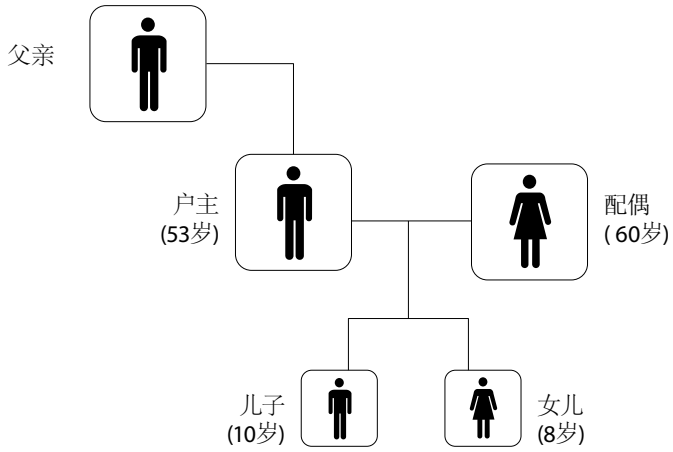
125. 对比之下，如果好的编辑程序发现户主和配偶属于同一性别，它会查验这两个人的生育力信息。只要户主有生育力信息，户主就是女性。于是这两个项目就符合编辑规则了。

图3
户主和配偶属于同一性别的例子



126. 图4中的另一个例子也说明了这个问题。大多数国家都把生育年龄定为15岁至49岁。假定一位妇女报告在52岁时生了一个孩子，根据通过指明孩子母亲的行数得到的直接证据或计算的年龄差（母亲与其亲生子女之间的年龄差大概不能超过50岁，不过对于领养子女而言，年龄差可能会大些）。编辑团队必须决定特定年龄差是可以接受的呢，还是必须更改，通过编辑来取代这个或那个年龄。如果编辑扩大了可接受的育龄范围，而其他妇女报告在更老的年龄生过孩子，那就是年龄本身报告有误，数据集中就会录入更多的异常。在这种情况下，编辑团队也必须决定，对于特定变量来说报告的年龄是否适当。

图4
有某些成员年龄的住户例子

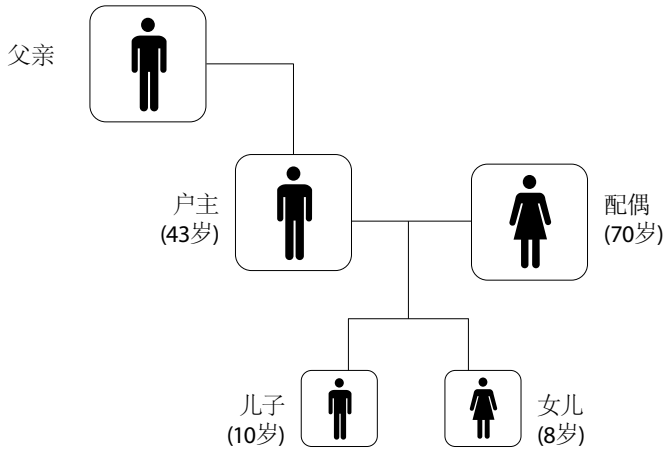


127. 图5提出了另外一种情况。假设编辑中发现一位70岁的妇女有一个10岁的儿子和一个8岁的女儿。这种情况是可能的，因为这对儿女或许是丈夫

与前妻所生。在这种情况下，子女与户主有关，而不是与配偶有关；不过更有可能打字员错把“40”打成了“70”。不管什么原因，在母亲和子女年龄相差50岁以上的情况下，假定主题专家要求数据处理员要么更改母亲的年龄，要么更改孩子的年龄。这个要求导致进行另一项更为复杂的编辑。鉴于该妇女为70岁，而长子10岁，编辑团队必须决定需要改变谁的年龄。编辑团队可以决定把长子的年龄改为20岁，这就解决了长子的问题；或者可以更改配偶的年龄。在第一种情况下，第二个孩子的问题依然有待解决，这也需要编辑。

128. 如果只考虑母亲和一个孩子的年龄，就采用一种插补方法随机指定年龄。可是，如果编辑还要查看丈夫年龄的话，编辑团队就有较大可能根据这一新补充的信息改变配偶的年龄。这一改变会使整个家庭的年龄更协调。

图5
有可能报告年龄不一致的住户例子



D. 有效性和一致性检查

129. 编辑过程的主要要求之一，就是不要有任何项目包含无效值。此外，对一切有关项目的回答在特定记录范围内以及记录之间必须保持一致。无效的输入项目无论从技术上还是从观感上都是不可接受的。譬如讲，在性别上允许使用只对男性和只对女性的代码组。任何其他值都是不可接受的，并且需要将其改成“未知”或改为可接受的两性之一；因为大多数国家都是在许多变量的性别基础上制订计划和决策的，数据中如果存在未知数，就会使获得工作所需的单一数值的努力变得复杂化。同样，如果在制表中出现诸如“茅草墙壁和水泥屋顶”、“13岁的女孩有20个子女”、“3岁获博士学位”之类的输入项，就说明统计机构无能，尽管少数此类情况并不影响一个国家的实际计划工作。

130. 插补作业必须尽最大可能同时考虑到与相关的变量有关、但未必是作为相关变量的结果而产生的一切信息。不过，有些情况下，编辑工作需要在

确定某个输入项的有效性之前进行一致性检查。如果根据一致性检查，通过插补程序填充了一个数值，必须将此数值与原始录入数据作比较，以查明是否有了实际的改变。如果没有改变，那就维持原始录入不动。

1. 自上而下的编辑法

131. 该程序从准备编辑的第一个项目（即所谓的“上”）开始，通常这是问卷上的第一个变量；然后依次逐项编辑，直到完成所有项目的编辑过程。一般做法是首先考虑应答率和各个项目的相对重要性。编辑工作通常都是从性别和年龄入手，因为这两项十分重要，尤其在动态插补方面。虽然自上而下的编辑法不完全保持各个数据项目之间的关系，但是它的确提供了一个完成编辑工作的适当架构。

132. 在编辑过程中，有些校订不止一次地改变一个项目值。这一过程有可能给数据集带来一两个误差。插补的值也许与其他数据不一致。甚至在依次处理各变量的时候，某个特定变量也应尽可能同时对照所有其他变量进行编辑。举例来讲，根据母亲的年龄插补的一个子女的年龄可能与该子女的报告的学龄或在该区居住的年头不相符。在这种情况下，将对年龄进行重新插补，直至达到一致。在最终赋值以前，插补的年龄是一个中间变量。在创建编辑指令的时候，直到最终赋值以前不要将插补的中间变量记录为变更值。

133. 虽然对少数项目来说在个别条件下编辑方案或可接受一个空白或“未报”输入项，但是相关的信息能够为大多数留有空白或有错误录入数据的项目提供输入项。以此种方式提供的输入项就个体而言也许正确也许不正确，但是凭借计算机在对各种不同储存值进行比较方面所具有的广泛能力和速度，我们可以确定能适当说明情况的置换输入项。与基于未经编辑的记录或通过插补已将所有不可接受的输入项转换为“未报”项目记录的制表相比而言，在大多数情况下作为这种结果产生的制表有时会较为相容。

134. 编辑程序还必须进行结构检查（见第三章）。编辑要检查人口项目（见第四章）和住房项目（见第五章）。此外，正如附件一着重指出的那样，编辑规程或许还应创立关于为制表所需的个别记录的一个或几个记录变量。

135. 极为重要的是，要避免循环编辑——即修改一个或几个项目，接着又在后来的某个时刻再把它们改回到原来的样子。本《手册》在其他地方指出，工作人员必须多次运行，以核实他们确已完全编辑了所有项目。有可能创造这样的编辑准则：在首次运行期间改变数据，但是在第二次运行中把它应用于业经改变的数据的时候，又将数据改回到原先的配置。这种操作步骤可在多次运行中连续进行。编辑团队要避免将此种准则引进编辑过程。

2. 多变量编辑法

136. 上面第1部分介绍的普查和调查中自上而下的编辑法不一定总能产生最佳结果——即最接近实际变量分布的结果。如前所述，如果在应用这种自上而下的方法时不够谨慎的话，往往会给编辑造成麻烦。

137. 另一种方法是以费勒吉-奥尔特系统为基础的多变量编辑法。这种方法需要较多的计算专家经验和计算机能力，但是大概能够获得更加接近“实际”的结果。在题为“插补方法”的附件五中介绍了各种不同的多变量编辑方法。在多变量编辑系统中，有必要确定一套关于检验各种变量间关系的肯定语句。然后该编辑程序结合住户数据来逐一检验各个语句，看其是否切合实际。对于任何不真实的语句，编辑都逐一跟踪无效输入项或不相容之处。在完成全部检测之后，编辑和插补系统必须评估如何最好地更改记录才能使其通过编辑的全过程。编辑团队通常奉行的准则是坚持最小化的变更和尽可能少地修改变量，以获得可接受的记录。

138. 图6中列举的11条说明语句提供了一个适用于某些人口特征的多变量编辑规则实例。其中规定，户主年龄必须在15岁或以上。对于一般化的编辑，如果国家规定最低年龄为X岁，那就最好用“X”岁这个标准。实例中的各语句，比如“关系”、“性别”、“年龄”、“婚姻状况”和“生育力”等，集中在其他重要基本变量上。这些变量是密切相关的，因此编辑团队应通盘查看它们，以便最有效地进行数据编辑。这里应该指出的是，虽然所有变量都很重要，但是有些变量要比其他变量对提供数据更加至关重要。

139. 图6显示了一种简单情况，由于某种原因户主和配偶性别一样——发现两人都是男性，而其中一人是有生育力的男性。很明显，这里的性别有误（如底部总计所示），有生育力的男性应改为女性。

图6
某些人口特征的多变量编辑规则实例

序号	规　　则	关系	性别	年龄	婚姻状况	生育力
1	户主年龄应在15岁或以上					
2	配偶年龄应在15岁或以上					
3	配偶应已婚					
4	如果有配偶，户主应为已婚					
5	若果有配偶，户主和配偶应互为异性	1	1			
6	15岁以下的人应该没有结过婚					
7	男性应该没有生育力		1			1
8	15岁以下的女性应该没有生育力					
9	对于15岁或以上的女性，生育力项目不得空白					
10	子女年龄要小于户主					
11	父母亲年龄应比户主大					
	总　　计	1	2			1

注： 数字“1”表示有两个以上的项目不相容。以项目5为例：鉴于户主和配偶性别一样，足见编辑在关系和性别方面有误，因此在对应的方格中出现了1。

140. 在图7的例子中，两个配偶都属于图6所示的同一人口。两个人都被报告为男性。这里的编辑程序简单而直截了当。记录误差数量最多的变量是首先被编辑的变量。在图7中，编辑程序实施“性别”插补，因为根据图6，该变量误差最多，表现在：(1) 关系与性别方面和(2) 生育力与性别方面。当编辑

程序检查生育力并发现户主有生育力信息但配偶没有这种信息的时候，就通过插补指定户主为“女性”。最后，当编辑团队检查计数系列而所有肯定语句均属实之时，就无需进一步编辑了。

图7

在未经编辑的数据集中户主和配偶性别一样及其解决的例子

个　人	关系	性别	生育子女人数
未经编辑的数据			
1	户主	男	03
2	配偶	男	空白
编辑后的性别数据			
1	户主	女	03
2	配偶	男	空白

141. 此项编辑的编辑规范可按图8所示写出。如果生育力信息是完整的，编辑就产生效果。可是显然编辑是不完整的，因为它只照管户主和配偶的生育力信息完整而准确的情况。

图8

用伪代码写出的纠正性别变量的编辑规范实例

```
如果户主的性别 = 配偶的性别
    如果户主的生育力项目不是空白
        如果配偶的生育力项目是空白
            （如果户主的性别并非已然是女性）使性别 = 女性结束条件
            （如果配偶的性别并非已然是男性）使性别 = 男性结束条件
        否则    另想别的办法，因为二人性别一样，且都有生育力！！！
                [可采取的“别的办法”可以是采用先前户主的性别；或者户主的性
                别，或者采用适当回答的所有户主性别的比率，等等。]
            结束条件
        结束条件
    否则    户主生育力为空白的情况
        如果配偶的生育力不是空白
            （如果户主的性别并非已然是男性）使性别 = 男性结束条件
            （如果配偶的性别并非已然是女性）使性别 = 女性结束条件
        否则    另想别的办法，因为二人均无生育力！！！
                [可采取的“别的办法”可以是采用先前户主的性别；或者户主的性
                别，或者采用适当回答的所有户主性别的比率，等等。]
        结束条件
    结束条件
结束条件
```

142. 下面图9提供了这样一个例子：编辑程序根据主要信息考虑了一个年仅13岁就有三个子女的寡妇户主。当通过编辑规则运行程序的时候，产生了下述结果：

图9

对一个非常年少的有三个子女的寡妇进行多变量编辑分析的实例

序号	规　　则	关系	性别	年龄	婚姻状况	生育力
1	户主年龄应在15岁或以上	1		1		
2	配偶年龄应在15岁或以上					
3	“配偶”应已婚					
4	如果有配偶，户主应已婚					
5	如果有配偶，户主和配偶应互为异性					
6	不到15岁的人应未结过婚			1	1	
7	男子不应有生育力信息					
8	不到15岁的女性不应有生育力信息		1	1		1
9	对于15岁或以上的女性，生育力项目不应空白					
10	“子女”年龄须小于户主					
11	“父亲或母亲”年龄须大于户主					
	总　　计	**1**	**1**	**3**	**1**	**1**

143. 现在我们再来考虑一个有三个子女的13岁寡妇户主。根据第一项规则——户主年龄须在15岁以上——查出第一项编辑失误，因为户主还不到15岁。鉴于她只有13岁所以对“关系”和“年龄”两栏都作了标记，因为这两个变量不一致。她不是配偶，所以不论规则2还是规则3都没有被触发。由于同样的原因，规则4和规则5也没有被触发——它们仅适用于配偶。可是根据规则6，不到15岁的人（在本案例中是13岁）应未结过婚，而我们的13岁的寡妇却“成为寡妇”，因此与规则不符。由于规则7仅适用于男性，所以未被触发。根据规则8，不到15岁的女性不应有生育力信息；既然此人确有生育力信息，这就与规则不符。规则9、10和11对此人不适用。

144. 根据肯定语句系列，与年龄变量有关的误差最多，所以首先要更改这个变量。在我们改正了年龄之后，重新运行检验；如果这一更正解决了所有问题，编辑程序即告完成。否则，程序将编辑不一致数量多的变量。

E. 数据的校正和插补方法

145. 如前所述，任何普查和调查都会因为存在“未报”、“未知”或信息缺失等缘故而在数据记录中留下空白。由于普查对象、普查员或数据录入的错误，也会出现无效输入项目。纠正的方法不尽相同，这得取决于项目。在大多数情况下，可以给数据项目指定有效代码组，并可以通过使用有关个人或住户记录中的其他数据项目的回答或对其他住户或个人记录的回答来适当确保这些代码组的正确性。

146. 本《手册》提出两种纠正有错误数据的方法。一种是静态插补或“冷卡”方法，主要用在有数据缺失或未知数的项目方面。另一种是动态插补或“热卡”方法，可用于有数据缺失以及不一致或无效项目方面。使用各种各样方法的不同计算机软件包以及这些软件包中的不同程序都以不同的方式利用冷卡或热卡，这些情况在相关附件中均有说明。

1. 静态插补或“冷卡”法

147. 在静态插补（或冷卡）中，编辑程序从一个预定的数据集中给缺失项目指定一个特定的回答，或者根据有效回答的分布按比例插补回答。在冷卡插补法中，程序并不更新原始变量组。在处理了第一、第二、第十或任何其他人的纪录之后，各数值没有改变初始静态矩阵中的数据。原始数值可为任何缺失数据提供插补。

148. 静态插补跟动态插补一样，是一种随机方法，但在一定时期内数值不变。附件五对这种方法作了说明。

149. 有时候静态插补使用一种比例推算法，根据预定的比例指定回答。作为按比例回答分布的一个例证，假定用有效数据（即出自完成填报的项目而非缺漏项目的数据）对从事农业的33岁男子的每周工作时间进行制表，结果表明：25%的人每周工作50小时；40%的人每周工作60小时；35%的人每周工作70小时。关于从事农业的33岁男子每周工作时间的回答缺漏或无效的项目，将有25%的人用50小时取代，40%的人用60小时取代，35%的人用70小时取代。可是除非有以往普查、调查或其他来源的可靠数据可用，否则这种方法就需要对当前普查所获的有效回答进行预先制表，而这在经济上或操作上未必可行。

2. 动态插补或“热卡”法

150. 从数据中清除未知数的另一种方法就是动态或热卡插补法，用它来给缺失的、未知的、不正确的或不一致的输入项分配数值。这种方法是美国普查局原创的，但是后来其他机构对此进行了改进。当数据集中出现一个未知数（或者有时出现几个未知数）的时候，动态插补便用一个或多个变量来估计可能给出的回答。动态插补在普查编辑中越来越受欢迎，因为它用法简便，并且能产生清洁、可复制的结果。此外，通过清除未知数更容易掌握不同轮次普查和调查之间的趋势，因为分析人员不必在个案基础上处理未知数了。

151. 对于动态插补而言，在有类似特征的个人当中，如果其中某个人的某方面（或某些方面）的信息属于未知数，即可根据这些个人的已知数据来确定可用以插补的最适当信息。这些特征包括性别、年龄、与户主的关系、经济状况和教育程度。插补矩阵本身是一组数值，就好像一副扑克牌。这些矩阵储存信息，并且在遇到未知数的时候提供信息。这副牌伴随着更新和/或逻辑“洗牌”不断发生变化，因而在数据处理过程中不断地变换对回答的插补：即所谓“热卡”插补。

152. 热卡中储存的数值代表了与有着类似信息的“最近邻”有关的信息。请注意，最近邻通常是指上一个最近邻，尤其在其他地方所描述的自上而下方法中，各居住单元和单元中的人只考虑一次，然后程序继续运行。因此，比如在一个村子里，在一个人的产妇孤儿状态不明的情况下，热卡会载有关于最近遇到的有同样性别、年龄特征和有效产妇孤儿状态的个人的信息。在移徙流量相对较大或艾滋病毒/艾滋病或引发不寻常统计活动的其他现象的发生率较高的国家，这种方法特别重要。同样，同一村或一组村庄范围内的住房特征要比在一个村或一组村庄与国内其他地方之间相比更有可能相似。

153. 举一个简单例子，来说明单一数值可以作为“卡片组”储存。譬如讲，如果由于某种原因一个人的性别无效，那就给该卡片组任意指定一个初始值（男性或女性），于是便确定了初始值。这个种子值遂成为所遇到的第一个性别不明者的性别。不过，如果第一人的性别有效，就用第一人的性别取代种子值。如果第二人的性别不明，那就通过插补矩阵来为其指定储存的性别。在这种情况下，插补的性别就是第一人的性别。本质上，当编辑发现一个可接受的项目值的时候，就把它纳入插补矩阵。当发现一个不可接受的项目值的时候，就通过插补程序用出自插补矩阵的有效值来取代它。

154. 这里描述的动态插补（热卡）法的问题之一就是：如果两个不同的项目有未知的数值，那就不一定用同一个“供体”的个人信息来指定有效回答了。每个数值都可能来自一个“真”人，但这些可能是不同的个人。一个比较好的方法就是同时从同一个人的信息指定两个变量。不过，给这些复杂的矩阵编程也许会遇到一些困难。

155. 下面图10包含的数据涉及到一个由10个人组成的住户群体。标有“X”和“XX”的空白指明了缺失数据。通常用9和99来指明缺失信息，在本案例中分别代表性别（一个9代表一位数）和年龄（99代表两位数）。可是有时候需要用9这个数字来代表另一个实值，比如在一些数量有限的关系代码组中的实值；因此，这些数值应当非常保守地使用。实际上，如果另有一个值（比如“X”、“.”或“..”）可以使用的话，那大概就应该用它了。请注意，虽然其他变量也可以用于插补（比如教育程度和职业），但它们没有包括在这个简单例证中。

图10

样本住户作为动态插补输入值的实例

身份编号	关系	性别	年龄
1	1	1	39
2	2	2	35
3	3	1	13
4	3	X	10
5	4	2	40
6	4	1	XX
7	4	2	13
8	5	X	XX
9	5	1	44
10	5	2	36

注： X和XX = 缺失信息。

156. 如果制作了一个叫做性别数组（代码 = 1）的插补矩阵，那么这个插补矩阵似乎就像：性别 = 1。

157. 处理了个人1之后，这个数值将仍为1。不过，在处理完第二个人之后，该数值将变为2。现在，这个变量似乎就像：性别数组 = 2。

158. 对于每个业经处理的个人的有效性别输入项来说，这个人的性别代码即可取代性别插补矩阵值。在第三个人处理完之后，就再一次通过插补将该值变为1，或男性。

159. 第四个人的性别不明，所以编辑查找插补矩阵值，在本案例中该值为男性，并且用插补矩阵值取代这个未知的值。第五个人为女性，所以它取代了插补矩阵中源自第三人（男性）的上一个值。这一过程一直持续到第八人。

160. 编辑再度使用插补，第八人成为女性，因为从第七人获得的插补矩阵值是女性。编辑先后两次使用插补矩阵获取数值：一次获得了男性值，另一次获得了女性值。鉴于性别出现频率大致均等，所以从长远来说插补使用每种性别的时间大约是各占一半。在处理了所有10个人的数据之后，变量似乎是：性别数组 = 2。

161. 虽然插补矩阵是以这种方式指定性别，但是还有使用该程序的其他较为复杂的方式。譬如讲，编辑程序可以使用与户主的关系和性别来帮助确定一个人的年龄。请考虑下列部分关系代码组清单：

1 = 户主

2 = 配偶

3 = 子女

4 = 其他亲属

5 = 非亲属

162. 数据处理员可以创造初始年龄值，该值或可逼近按性别确定关系的实际情况。这些数值不太重要，因为几乎可以肯定地说，它们在使用之前就通过编辑程序被取代了。同时，编辑也需要插补许多数值，所以很少有初始值影响到最终制表。这些数值或可如图11所示。

图11

基于性别和关系的初始静态年龄矩阵

	关　系				
	户主 (1)	配偶 (2)	儿子/女儿 (3)	其他亲属 (4)	非亲属 (5)
男性(1)	35	35	12	40	40
女性(2)	32	32	12	37	37

163. 请再次考虑图10中引进的10个人。在我们的例证中由于第一人被列为户主（代码 = 1）而且他是男性（代码 = 1），所以他的年龄（39岁）便

在插补过程中取代了第一要素（坐标1, 1）。然后卡片组就包含图12所示的各项数值。

图12
变更一个项目之后的动态插补矩阵例证

	关　系				
	户主	配偶	儿子/女儿	其他亲属	非亲属
	(1)	(2)	(3)	(4)	(5)
男性(1)	39	35	12	40	40
女性(2)	32	32	12	37	37

164. 第二人是配偶（代码 = 2），女性（代码 = 2），所以她的年龄（35岁）便取代了第二行第二栏中的数值，从而将卡片组更改为包含这些值。该住户其他人的年龄也同样取代了直到第五人的各个插补矩阵值。

165. 请注意，先前的性别插补程序将性别1指定给第四人。由于编辑要求的是插补性别值，所以编辑不用此人的年龄更新数组。编辑将只用出自性别和关系原本就正确无误的记录数值进行更新。可是当编辑运行到第六人时，发现年龄不明。此人为男性，是户主的一位“其他亲属”。因此，编辑使用属于“其他亲属”关系组（第一行第四栏）的男性插补矩阵要素，并且指定了这一类人的年龄值（即“男性其他亲属”——在本案例中40岁）。

166. 第八人既无性别报告又无年龄报告。编辑插补的性别为女性，然后根据这个分配的性别和关系代码（5）分配了年龄。在本案例中，年龄为37岁。

167. 虽然编辑根据已知的关系插补了年龄值，但是对其他变量使用了一个先前分配的性别值。这里，使用已分配的数值进行进一步的插补是一个不够高明的编辑方法实例（见下文的第3(d) 节）。最好查看其他已知的数据项目，比如婚姻状况，以便用于插补。

168. 在编辑完第十个人的数据之后，各项插补矩阵数值如图13所示。在本例中，两次插补都使用了初始静态矩阵。通常只有很少数（如果有的话）初始值用于插补。大多数情况下都使用根据人口查点指定的数值。

图13
变更多个项目之后的动态插补矩阵例证

	关　系				
	户主 (1)	配偶 (2)	儿子/女儿 (3)	其他亲属 (4)	非亲属 (5)
男性(1)	39	35	13	40	44
女性(2)	32	35	12	13	36

3. 与动态插补（热卡）法有关的问题

(a) 地理方面的考虑

169. 如果编辑程序使用动态插补法来插补缺失值的话，应力图使用尽可能小的地理限定区域内储存的数据。这一程序应能增加获得正确回答的概率，因为生活在同一个小地区的人们通常在人口统计、住房和其他特征方面同质性较强。如果人口缺乏同质性，就不存在相关性，因此编辑团队须逐项查看各种变量。另外，下面将要讨论到，有些变量在某些情况下不适合（比如在非常暖和的地区不宜集中供暖），编辑要把这种情况考虑在内。

(b) 相关项目的使用

170. 在使用动态插补获取缺失值的时候，要努力使用相关的项目来指定可能适合的数值。举例来讲，如果一个人的婚姻状况不明，编辑程序就要确定此人在该住户是否有配偶。如果是这样的话，编辑就将为其指定一个已婚代码而无需使用插补矩阵了。可是如果没有这方面证据的话，编辑程序就得依靠插补矩阵值。

(c) 变量次序如何影响矩阵

171. 使用插补矩阵的国家统计/普查机构，在拟订变量次序的时候要考虑到需要什么变量。对于人口项目，这些机构从一开始就要求编辑性别和年龄，以便能够把这些项目用于其他插补矩阵。总体编辑不得在插补矩阵中使用未经编辑的变量，尽管大多数计算机软件包都接受“未知”行或栏。应答率和各种变量中的属性分布有助于确定最佳变量以及这些变量中最有用的属性，用以扩展热卡。随后的插补矩阵可在编辑后使用这些数据项目。不过，统计机构应当考虑尽可能从插补矩阵中排除经过编辑的数据。

172. 举例来讲，如果编辑根据性别和关系插补年龄的话，在要么性别、要么关系是插补的情况下，不要更新这种插补矩阵数组（性别和关系）中的存储单元。作为一条规则，只有当年龄、性别和关系都有效并且相一致的时候，编辑软件包才把年龄输入适当的性别和关系存储单元。可是，有时候由于其他因素不可避免地要使用经过编辑的数据。有必要指出的是，大多数国家忽视这一建议，而根据先前插补的数值进行插补。一个可能的解决办法就是坚持使用插补标记，以提醒人们不要用插补的数据来使供体匹配一个缺失单位。

(d) 插补矩阵的复杂性

173. 国家统计/普查机构通过细化插补矩阵来提高获得相容且“正确”的插补矩阵值的几率。譬如讲，仅用关系这一项即可插补婚姻状况。可是守寡和离婚的几率伴随年龄而增长。因此，通过年龄和关系来插补婚姻状况是有道理的。编辑程序使用当前个人的年龄和关系，从存储于插补矩阵中的有着同样特征的上一个最近邻个人的有效记录中获取婚姻状况值。

174. 但是上述程序可能会带来新的问题。国家统计/普查机构通常按固定的顺序来编辑问卷项目，在自上而下的方法中，先编辑婚姻状况，后编辑年龄。如果是这样的话，在记录中婚姻状况和年龄都缺失的情况下，就不可能从最近邻的上一个记录中的同样年龄和关系值中获取婚姻状况值。[6]结果，编辑程序可能就不能确定该记录的年龄类别。另一个解决办法就是让插补数组有一行或一栏“未报”项目。使用有着相同关系和年龄“未报”栏的上一个最近邻记录中的婚姻状况类别来指定一个婚姻状况值。不过，有两个因素质疑此种做法。一是因为在同样组合中的“未报”情况太少，以致很难更新缺失项目的插补数组。二是基本上不可能获得适当的冷卡，亦即热卡所需的这些“未知”数值组合的初始值，因为它们在“真实”世界根本不存在。

[6]最好的编辑做法不需要在热卡中使用经过编辑的数值。有时候由于受到出成果的期限或计算机编程困难的制约，这种做法难以执行。在这种情况下，就需要对几个变量之一进行估算，然后将其数值纳入热卡，以便日后用以插补缺失变量。

175. 上述问题的解决办法增加了数据处理员的工作量，但是结果产生了比较清洁的产品。编辑程序首先通过查验确定是否存在有效代码组。如果当前个人的记录中没有对该项目的有效代码组，那么插补矩阵就不把该项目用于该记录。数据处理员可以通过创造一个较简单的插补数组来助推编辑过程。让我们来继续说明前面举出的例子。如果因为数据缺失而使得编辑程序必须插补婚姻状况的话，一般情况下插补数组将有两个维数：即年龄和关系。如果在查验后发现没有年龄方面的有效代码，那就只用关系来插补婚姻状况。由于关系方面的编辑先于婚姻状况，所以关系代码会是有效的。编辑程序对所有动态插补程序都采取同样的原则。

(e) 插补矩阵的开发

176. 主题工作人员要和数据处理员协作准备适当的插补矩阵。（有些编辑团队使用多重插补矩阵。）只能用有效的回答来更新插补矩阵；编辑团队不使用配给的或插补的数值。主题专家和数据处理员必须检查编辑规范和热卡，以确保一致性和完整性。

177. 要把很大的时间和精力集中在开发插补矩阵方面，其中包括深入研究行政案卷和以往普查或调查成果尤其是冷卡数据的使用。甚至在研究与发展阶段结束以后编辑人员也不应随机应用插补矩阵。在各种插补矩阵尚未实现内部相容的情况下，需要下大气力促使其协调一致。如果插补矩阵不使用标准协约，那么工作人员就必须单独考虑个案。

178. 虽然对本《手册》的各种实例来说插补矩阵中的每个存储单元都存有一个数值，但是一些编辑团队为每个单元都保有不止一种可能性。你可以把这设想为一个二维矩阵，它有个第三维度，就如同回归到黑板里面一样。这些存储单元提供了一个额外的维度。举例说明：如果在一个有四个孩子的家庭，所有子女的年龄都是未知数，那么计算机就不会把同样的值分配四次，那就成为四胞胎了。实际上，将指定四个不同的年龄。然而即便如此，也会不止一次地指定某一个年龄，这要取决于矩阵内的储存数据如何了。

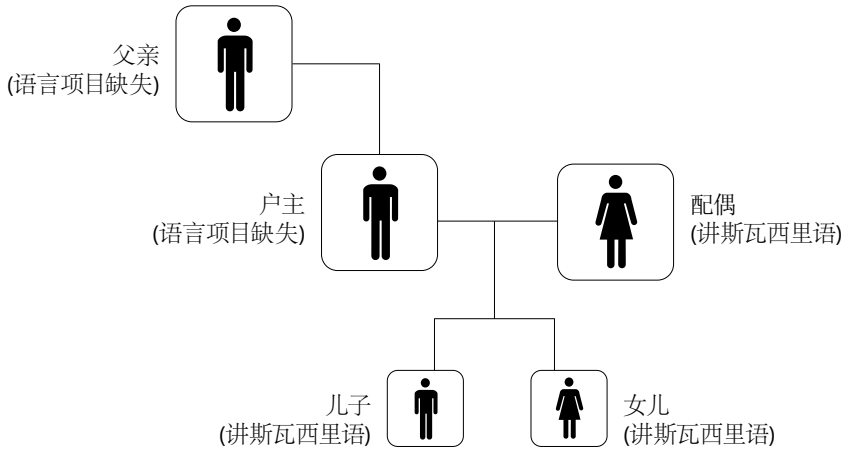
(f) 标准化的插补矩阵

179. 标准化的插补矩阵可以简化编辑过程。具有标准维数的插补矩阵可用于社会和经济变量，比如年龄组和性别等，且一经检验，即可很快付诸应用。

180. 举例来讲，国家统计/普查机构可能想要开发一种插补矩阵，以便给从未编码的特定语言确定代码。编辑程序需要查找据报告讲某种特定语言的另外一个人，而首先要查的场所几乎肯定是住户。如果在有关住户没有查到，编辑程序可以选择属于同一性别和年龄的前面一个人所讲的语言（在所有三项都有效的情况下已经更新了插补矩阵）。通过这一步骤有可能指定适当的语言，因为通常讲同一或类似语言的人们所在的地域彼此都很接近。

181. 在图14中，户主的“语言”变项情况不明。不论因为何种缘故，扫描员或打字员也许未能找到语言输入项或代码，或者出了什么别的差错。不过，鉴于配偶和子女全都讲斯瓦希里语，所以可将此种语言指定给户主以及户主的父亲（后者也缺失语言输入项目）。请注意，图14中的户主是女性。

图14
户主及其父亲未被指定语言的实例



182. 如果住户成员全都没有报告的语言项目，那么编辑程序就得另想办法了。首先，编辑查看其他变量，以便对所用语言给出一个间接的估计。有时候，民族、种族划分或出生地可以表示用以插补的适当语言。如果有这种标识符可用，那么编辑团队或可选择它来确定户主的语言。如果没有，则编辑可用年龄和性别来进行插补。插补矩阵大概如图15所示。

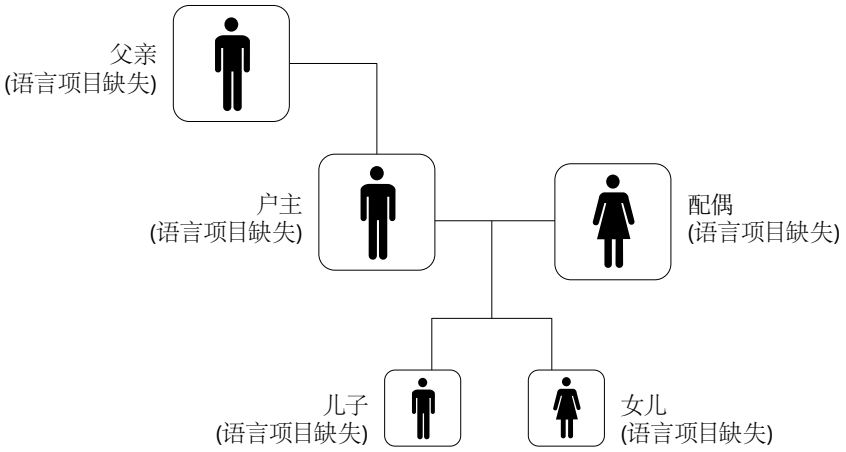
图15
语言项目动态插补矩阵的初始值

	年　龄					
性别	15岁以下	15-29岁	30-44岁	45-59岁	60-74岁	75岁及以上
男	语言1	语言1	语言1	语言1	语言1	语言2
女	语言1	语言1	语言1	语言1	语言1	语言2

183. 如果决定插补的话，编辑程序根据年龄组和性别来指定户主所用的语言。在这种情况下，插补矩阵中的各输入项仅适用于以前住户的户主，因为特定住户中的所有其他人都获得同户主一样的语言代码。

184. 在这个问题上，如果住户仍然没有任何人报告讲某种特定语言的话，那么编辑程序就根据户主的年龄和性别，用插补矩阵来给户主指定一种语言。这种指定的语言是在数据文件中另一位同龄同性别的户主最近使用的语言。鉴于插补矩阵随着遇到的可接受的案例增加而不断更新，这种指定的语言很可能是一般社会所讲的一种语言。

图16
没有指定语言的住户成员例证



185. 从运行编辑程序一开始，就会遇到编辑规则的例外情况。在工作人员从一个地方转移到另一地方的时候，必须细心注意语言的变化。有些国家一定要关注讲话人的本地化混合。然而，尽管存在此种情况，除非有选择地忽略某些语言的报告，否则插补所产生的分配值与未分配值的百分比应大体相当。

186. 另一编辑或许要查看宗教项目。对宗教项目的回答也可以根据年龄和性别进行插补。如果资料齐全，编辑程序就将继续更新，并从插补矩阵提取回答，用于插补未知信息。该插补矩阵看似语言插补矩阵，但存储单元中所储存的是宗教信息而不是语言信息。

187. 这种扩展采取自上而下依次进行的方式。编辑团队使用精密方法——比如费勒吉-奥尔特方法和最近邻插补法（NIM）（见附件五）——同时应用所有相关的编辑软件。当前的程序还假定存在某种适当的编辑次序。

188. 有许多经济特征（诸如劳动力参与率、上一周或数周的工作时间和去年的工作时间等）都可以通过使用类似特征得到插补。通过使用类似的插补矩阵，编辑程序可以快速检查变量特征值，而总体编辑过程应当运行更快。

189. 有时很难为系列插补矩阵中的第一个矩阵恰如其分地获得经过编辑的特征。通常，统计机构不想把未经编辑的项目纳入插补矩阵的维数；无论性别还是年龄，在其经过编辑以前都不会被编辑程序用作插补矩阵的维数。因此，头几个插补矩阵将使用无需编辑或不会改变数值的不同变量。对于人口项目的第一个插补矩阵，编辑或许使用居住单元中的个人号码，其中包括一个代表空单元的“0”数码。

190. 对于一般住房编辑来说，第一个插补矩阵或许使用居住单元中的个人号码作为初始维数，但是编辑团队或许修改住房项目功能来说明空单元。举例来讲，如果第一项住房编辑针对“外墙的建筑材料”或“墙壁类型”，那么初始值或许基于居住单元中的个人号码，其中包括一个说明该单元是何时空置的数值。

191. 如果居住单元是空的但“墙壁类型”有效，编辑就更新有外墙类型的第一存储单元。在已知墙壁类型的情况下，对于一个有人住的单元来说，编辑更新对应于该居住单元个人号码的存储单元。可是在其他墙壁的建筑材料不明的情况下，插补矩阵将根据该单元个人号码提供外墙建材的值。

192. 在初步使用该插补矩阵之后，编辑团队或许会转到其他住房特征方面，比如“住房类型”或“房屋保有方式”。不管选择了什么，都必须明确区分不同的居住单元并提供足够的多样性，以避免同一属性被反复选用。重复选择同样的属性可能会提供准冷卡值而非动态插补（热卡）值。举例来讲，在一个军营的“集体宿舍区”使用动态插补法，如果只选择了年龄和性别特征的话，就可能导致同一数值被反复使用。在这种情况下，或许所有居民都是男性，并且在一个有限的年龄范围。因此，这一特殊矩阵未必能够提供最佳结果。如果“房屋保有方式”有足够的多样性，有足够的房主与租户的百分比，这个变量就会有效。否则，有关国家就可以使用不同的住房类型了。

193. 一般来说，许多编辑团队发现，使用可比的插补矩阵维数，他们的检查工作量就会少些，就会较快地出成果，而且或许会获得较为准确的成果。

(g) 什么情况下不使用动态插补

194. 如果编辑团队选择根本不用动态插补法，编辑的次序仍然很重要。譬如讲，年龄跟许多项目有关联，其中包括与户主的关系、学校教育程度、劳动就业和（女性）生育力等。让我们来考虑图17中鉴别的住户成员状况：

图17

户主和子女而子女年龄值缺失的例证

个人	关系	年龄	学龄	工作	职业	生育子女人数
1	1	40	12	1	33	空白
3	3	X	7	空白	空白	空白

注： X = 年龄值缺失。
空白 = 不适用。

195. 关于第3个人的记录，有第3种关系(子女) 项目，但是没有报告年龄。为了查到年龄，编辑团队可以利用户主与子女间的年龄差（要么使用冷卡数值，要么使用插补法从上一单元获得数值）。譬如讲，如果年龄差为25岁，那么该子女的年龄即为15岁（户主的年龄40岁减去年龄差25岁）。

196. 学校教育年级也是已知的，在本案例为七年级。15岁的年龄可能与此学龄比较一致。鉴于对特定年龄而言适当的学龄范围小于就户主与子女间的年龄差而言的年龄范围，所以最好先检查学龄水平是否适当。如果报告了学龄水平，那就可以通过使用年龄差来提供适当的年龄；而这个年龄差可以要么通过静态（冷卡）插补法来确定，要么通过动态插补（热卡）法来确定。如果不知道学龄，那就可以借助户主与子女间的年龄差来指定年龄。

197. 不过，甚至年龄差都有可能缺失。实际上，在大多数国家，学龄要比年龄更有可能缺失。下面的事例说明，在年龄和学龄都缺失的情况下编辑团队所可能采取的步骤。

198. 在图18中，年龄和学龄都缺失，但是有其他信息存在。由此可以推测：第三人还不到就业年龄，而且年轻到尚未生育子女（或者他是个男子）。使用就业信息，通过一套冷卡值可以获得年龄值，但这个年龄将低于可接受的最低工作年龄。或者，如果编辑团队使用动态插补法，一个插补矩阵值会给出年龄值。这个选定的年龄大概应该用户主的年龄作为保持一致性的变量之一。举例来讲，如果户主年龄为20岁而非40岁的话，那就显然不适合把第3个人的年龄指定为14岁。年龄定下来之后，也就可以确定学龄了。而后者应该是与年龄和工作状况相一致的。

图18

户主和子女而子女的年龄和学龄缺失的例证

个人	关系	年龄	学龄	工作	职业	生育子女人数
1	1	40	12	1	33	空白
3	3	X	X	空白	空白	空白

199. 如果编辑团队决定插补所有或大部分项目，那就应制订一项旨在以合乎逻辑的方式建构编辑的策略。对于人口项目来说，编辑工作应从考虑一切可能含有未知数的项目入手。编辑团队要利用从各种调查和行政案卷、以往的普查和本次拟议普查的试点项目获得的资料以及现有的其他信息，来帮助确定把各个项目纳入第一个及随后的各插补矩阵。虽然必须结合每个国家的具体情况来开发插补矩阵的细节，但所有国家统计/普查机构都有可能掌握一些可用

于这种目的的数据资料。通过对热卡中的各种变量组进行检验，将会有助于为特定国家获得最佳变量组。

200. 有许多编辑软件包可以在运行中跟踪居住单元的个人号码。比如，一个用于不明性别的插补矩阵可以根据有关居住单元居住者的个人号码指定男性或女性。因此，拟为一口人住所的个人选定的未知或无效性别初始值或许是男性；为两口人住所的个人选定的性别初始值或许是女性；对于三口人的住所，该值可能是男性，如此类推。只有在所有一致性校订项目（诸如户主和配偶的性别以及生育力信息的存在）均已检验并解决之后，才作为最后的补救办法使用该矩阵。

(h) 插补矩阵要多大为好？

201. 大多数计算机软件报都能接受多维插补矩阵。在建构插补矩阵之前要考虑到下述要点：

(一) 插补矩阵过大带来的问题

202. 在主题事务团队与数据处理专家合作的时候，一些国家统计/普查机构面临的最大问题之一，就是编辑们过于勤奋了。在开发编辑软件包的过程中人们很容易沉迷于编程，并在这方面耗费过多的时间，以至于延缓了普查或调查进程。举例来讲，编辑团队也许会断定，为了确定年龄项目，除了参考“性别”、“教育程度”和“劳动力参与率”之外，还有必要把妇女的“生育子女人数”考虑在内。补充了总共生育“子女人数”项目之后也许会略微改进年龄的估计，但是因此而增加程序的复杂程度可能也是不划算的。编辑团队不得不做出判断：就精确度和效率而言，插补矩阵需要多少维数才能产生最佳结果。插补矩阵过大（承载存储单元过多）就不能彻底更新，因而可能不适当地转而使用冷卡数值。

(二) 认识插补矩阵的功能

203. 除了插补矩阵过大的问题之外，还可能存在路径混乱的问题。有必要切实保证主题人员以及数据处理员能够跟踪所有路径。他们必须共同确保插补矩阵执行预期的任务。主题人员和数据处理员还必须共同核实插补矩阵的每个变量或维数都得到妥善实施。另外，他们还必须确保所有变量或维数组合都能正常发挥作用。

(三) 插补矩阵过小带来的问题

204. 如果插补矩阵的维数太少，或者如果由于组合的缘故（比如年龄组合或教育层次太少），致使同一插补矩阵值被反复使用而得不到更新的话，查补矩阵就会过小。举例来讲，在年龄不明的情况下，假如在年龄数组中没有一个性别维数的话，很可能特定家庭中的所有子女都被指定为同一年龄。主题工作人员应与数据处理员合作，检验插补矩阵中的所有不同组合，并确保没有任何组合被频繁使用。

(四) 难以用于插补矩阵的项目

205. 实践证明，像“职业”和“产业”之类的项目非常难以编辑。虽然单独的职业和产业插补矩阵可能会产生不一致的结果，但是若试图核查所有成对的职业和产业输入项的话，那也可能代价高昂且难度很大。举例来讲，如果发现理发师或美发师在水产加工场工作，那或许就需要使用某个别的编辑程序了。此外，职业和产业类别太多，也会使得动态插补的难度极大。对于某些项目来说，编辑团队也许就断定编辑会起反作用，因而宁愿填写“不明”或“未报”。要不然，使用某种静态插补（冷卡）方法就行了。

4. 插补矩阵的查验

206. 在一个编辑软件包中，插补矩阵的基本结构大概如图19所示。编辑规范必须鉴别查补所用的数组，并且在初始数值组中使用冷卡值。

(a) 建立初始静态矩阵

207. 每当发现一个人的所有三个项目——在本案例中就是“关系”、“性别”和“年龄”——都报告了有效值的时候，就采取下述步骤来更新插补矩阵。不过，如果编辑程序发现性别项目无效（或空白），插补矩阵就根据有效的关系和性别代码组（业经编辑的变量）来选定一个数值。

图19
一个冷卡数组的数值组合插补代码示例

插补矩阵的初始值A01-年龄-女性-性别关系(2, 6)

户主	配偶	子女	其他亲属	父母亲	未报	性别
40	40	10	20	65	20	男
40	40	10	20	65	20	女

.
.
插补代码
若年龄 = 0:98
　　设A01-年龄-女性-性别关系（即性别+关系） = 年龄
否则
　　信息“年龄未知，故插补”年龄
　　写入“年龄未知，故插补，年龄 = ”年龄
　　插补年龄 = A01-年龄-女性-性别关系（性别，关系）
　　信息“现在年龄已知”年龄
结束条件
.
.
.

(b) 误差信息

208. 编辑软件包应提供几种方法，以确保妥善执行编辑和插补作业。下面评审其中的两个特征：信息指令和书写指令。

209. 一个信息源是如上文图19所示的信息显示。这种指令产生各级地理区域（诸如普查区、次级普查区划和主要普查区划）以及每份问卷的具体信息和累计数（发生信息的总次数）。所有问卷的一份摘要报告或可如图20所示：

图20
每个误差的插补数摘要报告示例

累计数	误差序号	信　息	行　号
–	14-1	每个妇女的子女数过多	2629
–	14-2	每个妇女的子女数过多	2645
2	14-3	现有的男孩数未报	2669
2	14-4	现有的女孩数未报	2678
33	14-5	最后一次生育月份不明	2723
7	15-6	无平均生育数；母亲和子女间的年龄差尚可	2892

注： 这里“14”仅指特定系列中的项目14；误差按顺序排列。

210. 按问卷组织的报告（图21）或可给出问卷号，其中包括所有特定地域代码组。然后报告可以按项目（本案例中之年龄）和按软件程序中的行号（如图中右侧所示）列出在方案中发现的误差。在本例中，年龄为空白，但是插补矩阵根据此人的关系和性别提供了年龄为48岁。对于本案例，具体年龄不明，但若有需要，信息指令亦可写出此项信息。

图21
问卷中的误差报告示例

	问卷标识码：0101017	行号
年龄(1) =	年龄不明，需插补	#46
年龄(1) = 48	现在知道了年龄	

211. 当然，虽然逐一列举抽样检测或有选择的小数据集的所有误差有一定的意义，但是在流水线生产中的产出规模将极为庞大而笨重（并且很快就变得毫无用处）。为了避免这种结局，应关闭问卷中存在的全部或部分问题。当然，概括统计量依旧不变。

(c) 定制的误差列表

212. 该软件还可以提供另一种指令，以便进行更加详细的编辑规范和编辑流程分析。这种指令可用于在进行改变之前显示信息，然后在所有更改之后显示信息。最后，它可以再现更改后的这个或这些记录。通过这种方式，分析家可以确信所有编辑路径都运行正常。图22可以显示这些结果。产出的第一行

给出各项变量（比如省、关系、性别、年龄等）。然后显示输入的数据；接下来一行是误差（本案例中无年龄值）；最后一行显示更改后的数据。

图22
包括多种变量在内的辅助性问卷误差列表示例

	省	区	户主	关系	性别	年龄
输入的数据	01	01	17	1	1	
误差	年龄不明，需要插补，故年龄 = 空白					
编辑后的误差	01	01	17	1	1	48

213. 该程序可以帮助编辑团队确定编辑路径是否恰当。

214. 检验是普查和调查编辑工作的重要组成部分。下面的方法是检验编辑程序的一个可能的方式。程序一开始，先让专家通过创造一个“理想的”住户进行系统分析。理想的住户是一个完整的住户——有户主、配偶、子女、其他亲属和非亲属——及其全部特征。理想住户必须不出任何差错地通过编辑的全部过程。然后，该单元在同一个文件中一遍又一遍地进行复制。该程序按下述方式持续进行：

(a) 数据处理员按照与编辑规范和编辑程序相对应的顺序，给每个住户依次引进一个单一的误差；

(b) 然后分析人员在编辑之初检查所有路径；

(c) 一俟编辑过程正常走完全部路径，数据处理员便运行整个数据集的一个样本，以便在实际数据集中查找特异反应并作必要的修改；

(d) 最后，数据处理员操作整个数据集。

215. 在各项检验信息工作正常并且进行了适当修改的情况下，数据处理员即可决定关闭这些检验信息，转而检查较低层次（比如查每份问卷）。如果大国要运行其整个数据集而留下各个问卷的信息陈述的话，如此产生的行号和文件将禁止使用。可是这些信息的汇总报告应仍旧继续，因为它提供了各种地理层级的有用信息。其产出大体如图22所示。

216. 计算机编辑通常都有保护程序。编辑轨迹显示出所有数据改变，并且标记变更案例和被取代的数值。通过查阅编辑轨迹，确定修改数是否低到可以接受整组记录的程度。

217. 如果个别项目错误过多，该项目也许没有经过适当的预先检验（自检或结合其他项目检验），说明普查员或普查对象没有搞懂这个项目。比如说，有时候普查员糊涂了，只从男性成年人获取有关生育力的信息而不从妇女收集这种信息。如果此类数据收集系统化，编辑团队或许让编程员将生育力数据从一对夫妇的男方移至女方；否则，编辑团队在这一阶段改正这个错误。

218. 通常，编辑程序需要查看几个不同的文件，以覆盖所有状况。另外，数据处理员也需要修改句法或逻辑上的毛病。就连最有经验的数据处理专家偶尔也会误将“小于”符号键入“大于”符号，而且直到检查了好几遍之后

才发现错误，因为特定问题也许不能直接就看出来。同样，逻辑上的小毛病可能起初并不明显。在这方面也需要主题专家和数据处理专家合作，以便在编辑过程中尽早解决此类问题。

(d) 编辑多少遍为好？

219. 如前所述，一俟问卷定下来就应着手制定和试验编辑规范说明书及编辑程序。个别项目应采取自上向下方法单独拟订，但即使要同时编辑好几个变量，个别项目的编辑也需要对整个数据集的小部分进行试验。应由主题专家制订编辑规范说明书，然后把个别编辑程序交由编程员付诸实施。接下来即可构建总体编辑，并且在越来越大的部分数据集上运行，同时不断加以改进。

220. 一般来说，不论对部分程序还是对整体程序，一个编辑程序最好运行三次。现解释如下：

221. 第一轮编辑应用实值插补矩阵，而不用初始静态矩阵中产生的值。一些国家使用其他来源的数据——要么出自以往的普查或调查，要么出自行政案卷——来为数组提供冷卡值。数据处理员操作整个数据集或其中一大部分，来为插补矩阵提供数值。出自实际数据集的冷卡值可能比较准确且通用。编辑工作仅用这种初始静态矩阵数值的2%，其余全是动态插补值。

222. 第二轮编辑进行实际编辑。这一轮编辑需要反复运行好几遍，以覆盖所有状况。这时，数据处理员需要进行修改，纠正句法或逻辑错误。此外，甚至最有经验的数据处理专家也会犯错误，而且，有些特定问题也许不会立即显现出来，所以直到运行好几遍才发现错误。同样，逻辑上的小差错起初也不一定看出来。

223. 第三轮编辑的目的是要确信(1) 数据集中没有错误了；(2) 编辑程序没有带来新的错误。在数据处理员最后一次运行编辑程序的时候，错误列表上不应出现任何错误。如果依然有错误，那大概就是编辑在逻辑上出差错了，所以数据处理员需要进行修改。另外，这一轮通常通过编辑逻辑告诉数据处理员是否编辑过程偶然间引进了新的错误。

5. 插补标记

224. 插补标记是用以维护未经编辑的数据信息的一种方法。如前所述，许多编辑团队都担心会在改变未经编辑的回答时丢失潜在的信息。在因为不一致而改变一个数值的时候，编辑团队也许希望保存原始值，以便在普查后进行进一步的人口统计或误差分析。主题专家和编程人员都想要分析有关数据缺失、无效或不一致问题的方方面面。编辑团队成员需要确信插补的和未经插补的分布是一致的，并且摸清编辑和插补方案中是否存在任何系统性的误差。譬如讲，有时候数据处理科学家会偶尔仅使用冷卡数值，因为编辑程序忽视了插补矩阵的更新。如果国家进行普查预检，编辑团队就需要在预检之后调查某些变量之间的关系，以便问卷最后定稿。在有大容量硬盘的微型计算机普及以前，许多统计机构在普查工作中其磁带和其他存储媒介没有足够的空间保管额

外的数据。可是近年来，对大多数国家来讲，保管未经编辑的数据已经不成问题了。

225. 一些国家选择为每个项目维持一个简单的二进制会计变量作为标记。这种方法很简单，且每个变量仅占用一个字节。例如，美国普查局在住房和人口普查的每个记录末尾给每个变量设置了插补标记。举例来讲，每个住房变量的初始标记变量为“0”，但是如果原始项目发生了任何变化，该变量就变成为“1”。编辑程序不保留原始值，尽管统计机构有时候也为每个记录或在总计中汇编这些变量。

226. 还有其他一些用于保存未经编辑的回答资料的方法。在图23的示例中，国家统计/普查机构使用插补矩阵把配偶的年龄从70岁改成40岁。国家统计/普查机构可以轻易地将插补以前的值（在本案例中为70岁）安置在为插补标记保留的空格，同时保留用于出版制表的分配变量（即本案例中的40岁）。为了审查数据集中的更改，统计机构可以制作分配值和未分配值的频率分布表和交叉分组列表。如果在这种数据集编辑效果分析之后基于编辑程序的制表呈现可疑或异常情况，编辑团队可能想要考虑总体或部分改变编辑流程；而由于近年来硬盘容量有了极大增加，所以可将全部初始值存入记录，以备将来使用。各机构大概想要至少保管两套文件，因为只保管业经编辑数据的文件可能制表速度较快。

图23
带有插补值标记的人口记录示例

个人	性别	年龄	平均生育数(CEB)	性别标记	年龄标记	CEB标记
1	1	40	空白			1
2	2	40	7		70	

227. 图24说明了一位13岁的女性被记录为生过一个孩子（平均生育数为1）。不过，编辑团队已经决定生第一胎的最低年龄为14岁；14岁以下的女性生育多半可能是搞错了。跟以往一样，这就再次提出了这样一个问题：此种情况是代表数据集中的误差呢，还是算作实值。

图24
一位生育力空白并加了标记的年轻女性的标记示例

个人	性别	年龄	平均生育数(CEB)	性别标记	年龄标记	CEB标记
生育力空白						
4	2	13	1			
生育力空白但加了标记						
4	2	13	空白			1

228. 根据编辑规则，插补平均生育数的“空白”信息。请注意，平均生育数标记有点复杂，因为它不但要录入数字，而且要说明是经过插补的空白。假定主题工作人员想要研究据报告有一个子女的13岁个人号码及特征。数据处

理员可以在留作标记用的记录空格记录初始信息，通常在记录末尾。然后，发表的那套表格将排除有关该女性平均生育数的信息，但是该信息仍然可用于将来的研究。日后，尤其在制订后续调查或下次普查计划的时候，编辑团队可以利用关于13岁女子生孩子的信息决定是否需要降低纳入的年龄门槛。

229. 在使用插补标记方面的一个问题，就是上述程序要占用计算机的很大内存空间。当标记重复每个变量的时候，经过编辑的数据集文件大小差不多相当于未经编辑数据集的两倍。对许多国家来说，这样的长期储存是难以承受的。然而，可以把原始数据和校订数据储存起来，供日后重建使用。

230. 人口众多的国家或可在抽样研究的基础上使用插补标记。譬如讲，一个国家或许想要每100个居住单元创建一个数据集。然后编辑程序在这个较小的数据集上运行插补标记，以评估编辑如何影响数据质量，并判断未经编辑的数据和经过编辑的数据有何差异。

F. 其他编辑系统

231. 本《手册》的大部分都描述使用自上向下方法进行普查和调查的计算机编辑。有些国家采用另外一种较为复杂的计算机编辑程序，叫做多变量编辑法（见上文D. 2节）。费勒吉和奥尔特（1976年）率先开发了这些程序，它们通常应用于普查和调查中的一些最重要的变量，即：年龄、性别、关系和婚姻状况。不过，它们可以运用到任何变量组，或用于普查或交叉问卷上的所有变量。在这种方法中，编辑程序同步查看一个人或一个住户中的所有人的这些项目，以找出缺失或不一致的回答。当发现未知（空白）、无效或不一致的输入项目的时候，就通过一系列检测确定在选定的项目当中哪个误差最大，而这个误差就要首先纠正。然后重复进行检测，以确定没有无效或不一致的项目了；如果还有问题，那就通过编辑来纠正由大多数有剩余问题的项目。程序反复运行，直到不存在误差为止。

232. 加拿大统计局发展了费勒吉-奥尔特方法，将其用于1976年至1991年的普查。在1996年的加拿大普查中，该方法经过改良叫做“新插补法”(NIM)。首次可以借助此法“同时对大量的[编辑和插补]问题的数字及定性变项最少改变插补”（Bankier、Houle和Luc，未注明出版日期）。

233. 如果使用传统的动态插补或热卡方法进行编辑的话，一系列问卷项目的插补信息可能来自许多不同的个人，这要取决于用以更新插补矩阵的信息如何了。举例来讲，如果个人A的性别、关系和婚姻状况正确，就用这些值来更新适当的插补矩阵。如果A的年龄缺失或无效，当然不会用来更新插补矩阵。实际上，将用其他项目来更新该值。于是，如果下一个人有不一致的性别而“性别”经过了插补的话，那么个人A就贡献性别。如果年龄也不知道，那么编辑程序就使用另外一个人的年龄。

234. 新插补法适用项目供体，希望所有缺失或不一致的信息都可以出自同一个或几个供体。为了从单一供体获得全部或大部信息，必须把总体数据记

录储存在计算机存储器中。然后，如果年龄和性别未知或无效的话，储存的同一变量就会为这两个项目提供数值。

235. 热卡插补方法应有下述目标：

(a) 经过插补的住户要非常近似于编辑失误的住户；

(b) 住户的插补数据要尽可能出自单一供体而不是两个或更多的供体。另外，插补的住户应非常近似于那个单一供体；

(c) 基于现有供体的同样良好的插补操作应有同样的被选用机会，以避免错误地造成人数虽少但是重要的群体规模膨胀（Bankier、Houle和Luc，未注明出版日期）。

236. 按照最近邻插补法，实现这些目标首先要确定刚编辑过的住户要尽可能近似于编辑失误的住户。这就是说，两个住户要有尽可能多的质量变项相匹配，在数字变量之间仅有很小的差异。有这些特征的住户叫做“最近邻”。下一步是确定每个最近邻不匹配变量的最小子集（包含数字和质量两种变项），如果是插补的话，这些子集可以使住户通过编辑。然后在这些插补操作中随机选择一个通过了编辑并且既近似于编辑失误住户又近似于已通过编辑的住户的插补（Bankier、Houle和Luc，未注明出版日期）。

237. 本章讨论了一般编辑和指标程序。第三章涉及到结构编辑问题，其中包括计算机编辑的第一项也是最重要的任务，因为它规定，必须让每个居住单元在其国家等级制度中的适当位置上出现，并且仅出现一次。

第 三 章
结构编辑

238. 结构编辑检查覆盖范围和测定各种记录的内部一致性如何。结构编辑必须确保：(a) 普查区内的所有住户和集体住所都有记录，并且有适当的顺序；(b) 凡是有人住的居住单元都有个人记录，但是空置单元没有个人记录；(c) 各住户不得有重复的个人记录，也不得缺漏个人记录；以及(d) 各普查区内不得有重复的或缺漏的住房记录。因此，结构编辑检查的目的在于确信问卷总体是完整的。

239. 鉴于用以判定和纠正结构误差的技术变化如此之快，所以普查或调查所使用的特定结构编辑程序可能需要在一定时期内有所改变。因此本章来考

方 框 3

结构编辑准则

结构编辑应设法完成下列任务：

- 确保每一批调查区（EA）记录都有区域代码（省、地区、调查区，等等），并且这些批量记录要有通用名称；
- 确保每个居住单元都包括在内；而调查区内的的所有住户均已录入；
- 将住户归并到适当的调查区，并将调查区归并到适当的上一地理层级；
- 根据人口规模和问卷布局，帮助判定问卷册内或册外的个人页和住户页；
- 给每个人的记录指定有效的记录类别；
- 在居住单元以外单独处理群体住所或集体宿舍的记录；
- 确保各类记录之间的一致性：比如，空置单元没有人，有人居住单元至少有一人。确保每个住户的各种人数记录与住房记录的住户总人数相一致。在对单一住户使用多项文件的情况下要确保问卷数正确，并且彼此间有适当联系；
- 清除住户内部（重复个人）和住户之间（重复住户，或住户部分重复）的重复记录，以避免过分覆盖；
- 在一类记录范围内处理空白记录；
- 处理缺失的居住单元。

察有关项目有效性和记录之间和记录内部各种项目关系的比较一般性问题。第四章和第五章处理与特定人口和住房项目有关的问题。

A. 地域编辑

1. 住所定位（地域）

240. 根据《人口和住房普查的原则和建议》（第二次修订本）（联合国，2008年，第2.78段），“地域”的定义是“一个清晰可辨的人口群集……其中的居民生活在相互毗邻的住所群，并且该人口群集有一个名称或有当地认可的地位”。在该《原则和建议》第2.78-2.88段关于“地域”和“城市与乡村”的定义项下可以发现与住所定位有关的更多信息。对于参与进行住房普查的人来说，有关这方面信息的研究必不可少，因为在进行住房普查的时候，用于描述住所定位的地理概念不论对执行普查还是对日后普查结果的制表都极为重要（联合国，2008年，第2.455段）。

241. 在编辑地域的时候，地域代码组必须绝对准确。为数据处理获得完整而准确的地域层级代码组，是整个普查工作最困难的任务之一。如果地域被错误编码，数据录入操作员就会把相关的居住单元指定给国内的其他地方，此类错误往往纠正起来十分困难。

2. 城乡居民

242. 一国之内，传统的城乡地区差别系建立在这样一种臆断基础上，即：不论其定义如何，城市地区提供了不同的生活方式，通常生活水准高于农村地区。在许多工业化国家，这种差别已经变得模糊了，城乡之间的主要生活环境差异一般表现在人口密度方面。虽然发展中国家城乡之间在生活方式和生活水准方面的差距依然很大，但在这些国家，城市化的迅速发展产生了对有关城市地区不同规模信息的巨大需求（联合国，2008年，第2.82段）。

243. 大多数国家在普查之前确定了哪些地域属于“城市”地区，那些地域属于“农村”地区，而在收集了普查数据以后，再作必要的调整。如果国家给城市和农村居民分配代码组（比如城市为1，农村为2），那么根据编辑团队预定的规范，这些代码组可以在打字的时候录入或者可在编辑时确定。如果编辑团队提供了一份清单，表明哪些地域单位是城市的，哪些地域单位是农村的，数据处理员即可轻易地给住房记录指定适当的代码组。

244. 要努力确保人口特征总体上与普查区相一致。举例来讲，在一些国家，除了医生、教师和类似职业工作者之外，基本上不应在农村地区发现专业人员，也不应在城市地区发现农业工人。编辑团队要通过检查确保地理区域的分类是正确的。

B. 覆盖范围检查

1. 事实上的查点和法律上的查点

245. 常住居所定义。一般来讲，为了普查目的“常住居所”的定义是在普查时个人居住的地方，而他（她）已在此生活了一段时间或打算居住一段时间。一般情况下所查点的大多数个人都有一段时间没有搬迁了，因此可以明确界定其常住居所地域。而对其他人来说，上述定义的应用则可能会产生多种解释，尤其是如果有关个人经常搬迁的话。建议国家在考虑常住居所的时候规定一个12个月的门槛，据此，长住居所要么(a) 是此人在既往12个月内大部分时间（亦即至少六个月零一天）连续居住的地方，其中不包括临时外出度假或因公出差的时间，或者打算在此至少再住六个月；要么(b) 此人已在此至少居住了12个月，其中不包括临时外出度假或因公出差的时间，或者打算在此至少再住12个月（联合国，2008年，第1.461-463段）。

246. 有了上述常住居所定义，国家统计/普查机构一般收集事实上的普查资料（即普查期间发现有关个人经常夜间居住的地方）或法律上的普查资料（即通常发现他们所在的地方）。检查住房记录之间关系的编辑工作，尤其是检查住所中的人数与个人记录之间关系的编辑工作，必须考虑到普查的类别。有时候，国家既收集事实上的信息又收集法律上的信息。每个人的项目都可以表明他（她）是否(1) 常住居民；(2) 临时到访，但在别处有常住的家；或者(3) 经查住在该住户，但有时候人不在。按照事实上的信息制表，如果所有三种情况都有，仅使用(1)和(2)两项；按照法律上的信息制表，如果所有三种情况都有，则仅用(1)和(3)两项。

247. 实施这些不同编辑方法的国家统计/普查机构，在每个阶段——不论在数据收集阶段、数据处理阶段，还是在数据传播阶段或分析阶段——使用它们时都必须十分小心谨慎。特别是数据集的制表，应当在考虑到预期的人口类型的基础上进行。这三种类别的用户要熟悉选定的人口，因为对所有数据集的分析将导致把某些个人纳入两次。如果需要的是一个事实上的人口，那么制表就必须包括第三种人——即临时外出的人；如果需要的是法律上的人口，那么制表就必须包括第二种人。在制表初期，即为打印报告和辅助媒介指标的时候，编辑团队或许选择制作总数据集的一个子集以供处理。对于后来的制表，文档文件则应明确陈述如何处理各种可能性。或许采用多重文件的方法比较合适。

248. 编辑程序应该能够确保，在所有三类记录都有的情况下，要有适当的相应性。如果事实上的记录中回答者甚少，这可能说明他们实际上是身在外地的居民，或者有另一个需要特殊处理的查点问题。

2. 住户和居住单元的层次

249. 第五章将考察住户、居住单元和住所之间的关系。这些概念的实施由各国统计/普查机构决定。不过，在着手进行个别住房编辑之前，编辑团队必须制定检查方法，以确保在数据收集和录入过程中遵守层次。

3. 问卷的零碎信息

250. 在逐项编辑以前，作为结构编辑工作的组成部分要通过计算机程序对有效记录、确实记录和重复行号进行检查。还必须核实当下编辑的记录是不是住集体宿舍者的记录。数据录入操作员可能会在键入记录时出错；偶尔他们也会忘记删除零碎信息（部分记录）。初步编辑的作用时检查文件中的零碎记录，以便将其清除。最常见的情况是记录上只有地域代码组，但是没有人口或住房项目。

C. 住房记录的结构

251. 可能被纳入全国住房普查或调查收集信息范围的话题之一，就是一栋建筑物中有多少住所。在这种情况下，查点单元是一栋建筑物，而收集的信息是常规和基本住所的数量（见联合国，2007年，第2.524段）。

252. “总体编辑”是指为确保作为建筑物组成部分的居住单元数与住房记录中的居住单元总数相吻合所做的工作。如果有关建筑物按五个居住单元编码而单个居住单元的实际数是十四个的话，编辑团队就必须决定采取何种措施加以调整：要么(a) 根据个体记录的计数更改第一个数字（多数情况下这个数字比较可接受）；要么(b) 使用有关现有记录的信息引进另一记录数字（应避免采取这种办法）。

D. 住房和人口记录的一致性

253. 如果普查或调查既包括住房记录又包括人口记录，结构编辑需要确保这两类记录相一致。

1. 空置房和有人住的房屋

254. 空置居住单元不应有人口记录，但是有人住的居住单元必须要有人口记录。如果有人口记录而住房被登记为空置，空置状况应该为有人住。有的时候在同一个项目中把“空置”和“保有”两种状况登在了一起，所以在作判断时也要把这种情况考虑在内。另外，如果回答有房主自用单元的单元数或租户租用单元的“付租金”信息，那么编辑程序就使用这一信息作判断；否则就可能需要一个插补矩阵。

255. 如果一个被认为有人居住的单元缺失人口记录，那么编辑团队必须决定是把它算作空置单元呢，还是用另一单元的人来替代。如果单元是空置的，通过查补可以轻易改变空置状况的变项。可是如果单元有人居住，那么编辑团队就必须决定是否和如何用同样的人数来指定有尽可能类似特征的另一单元的人。由于不可能知道缺失个人的特征，所以这种方法即使要用也只有在编辑团队确定别无其他可替代方法的情况下才能使用。下面概述三种可能的替代方法：

(a) 选择让一个居住单元空置

256. 在这种情况下编辑团队决定，在现场输入的空置居住单元记录应让它继续空置，所以不予插补任何数值。第五章描述了空置单元的住房插补问题。

(b) 反复重访有问题的居住单元以图填完问卷

257. 国家统计/普查机构可以选择让普查员反复核实空置单元数据的做法，直到他们确信这些单元要么空置，要么有人居住，并且直到普查员收集到了至少最低限度的特征。在这种情况下，编辑团队应制定编辑程序，以检查单元是空置的或者有足够的特征被认为“有人住”。依编辑团队定义的“最低限度”信息而定，来应用第四章中描述的常规编辑方法，或者如前所述使用来自供体记录的数据来插补“缺漏的”个人。

(c) 用另一居住单元的人替补缺失的个人

258. 关于替补整个住户或缺漏个人的规程，本章另有描述。这些规程需要假定缺漏的个人有跟被替补的人一样的特征，这几乎肯定是异乎寻常的，而且程序本身非常难以操作。可是若没有这些程序，人数的查点以及按特征分列的个人都可能会减少。

2. 重复的住户和居住单元

259. 居住单元发生重复有多种多样的原因。有时候个别数据录入操作员会把同一个居住单元输入两次。由于在国家统计/普查机构内部缺乏质量保障措施，有时候不同的数据录入员会偶尔重新输入同一个居住单元甚或整个普查区。第三，普查员记录的某个居住单元的地域代码不正确，由此产生重复信息，把同样的地域标识指定给另一个居住单元。

260. 如果办公室的监督人键入批量数据的话，大概就不会发生重复了。不过，应该开发一种编辑程序，它可以确保不会因为数据录入员二次键入同一个或多个住户而发生住户重复的问题。各国在结构检查结束并消除了重复记录问题以前，不要进行数据分类整理。分类之前，工作人员可以手动纠正批量数据；分类以后，工作人员就不能发现重复问题了。数据分类后，编辑可以检查重复的住户，并且使用插补法消除此后的重复输入项目。

3. 缺漏的住户和居住单元

261. 同样，分类整理之后，缺漏的住户也许就显露出来了。举例来讲，编辑程序设想了一个最低层地域内的住户顺序，比如1、2、3、4，但是只收到了1、2、4的记录。于是就必须决定要么重新给各单元编号，要么找一个“可接受的”方法，用另一单元来替补单元3。如果很明显事实上确有住户缺漏并需要补充的话，有好几种方法可以插补住户的缺失值。一种方法是简单重复上一住户。可是如果已知该住户人数的话（此种情况很常见，即便其成员特征不

明），就有可能倒回去复制有同样人数的上家居住单元。同样，如果你知道该住户成员的年龄和性别，这方面的信息也有助于找到一个替补住所。最好不要试图使用热卡插补法来创造有关住户成员的信息，因为这种方法往往产生互不相容的变量。

4. 居住人数和居住人总数的一致性

262. 在住房记录中记录的居住人数应当恰好等于住在该房屋的总人数。编辑程序合计了人数，然后将此数与住房记录上的居住人数作比较。如果与居住人数值有出入，那就必须要么把居住人数调整到等于总人数，要么调整个别输入项目。第五章阐述了居住人数的编辑问题。

(a) 当居住人数多于居住人总数的时候

263. 如果住房记录上的“居住人数”特定变量值大于个人记录的总和，那么编辑团队就遇到了实质问题。没有人会知道缺漏人口的特征。因此对编辑团队来说，是用特征值来插补缺失个人特征呢，还是用出自类似住户的替补个人来补缺？这是个两难问题。缺漏的个人不应被替补。可是，如果接受了居住人数，这个替代办法就会减小查点人口的规模。编辑团队必须通盘分析，然后确定一个合适的途径。

264. 现有好几种办法可以查找和替代缺失的记录，但是其中没有一种是完全令人满意的。整户人家可以用不同的重要特征来存储。如果发现一个有人但是人口不齐全的住户，可以在文件中搜索一个所有或大部分已知特征相匹配的住户，然后根据这个供体住户的其他人来调整缺失个人的数据。不过，这种操作程序十分复杂，所以采用此法的国家统计/普查机构应当早早提前为此做计划。

265. 另一个可替代的方法就是先将有缺失记录的所有住户作标记，然后继续进行其余的编辑。在编辑过程的末尾，当所有个人输入项目都已矫正之时，编辑团队可以决定由数据处理专家通盘检查文件，以便用完全编辑的数据集来进行补充修改。通过使用自上向下方法，编辑团队可能会发现适当的供体。

(b) 按性别检查人数

266. 有时候在住房记录上按性别报告居住人数。在这种情况下编辑必须单独合计每一种性别的人数。同样，如果合计人数与居住人数有出入的话，也必须根据具体情况调整其中的一个值。通常情况是调整住房记录上的总人数，而不是追加“缺失”的记录或删除载有有用信息的记录，因为普查员有可能在住所表上出差错。

(c) 顺序编号

267. 人口记录应按顺序排列，即依次编号。这些号码应作为一个变数出现在问卷上，比如行号或序号。另外，序号应出现于数字顺序编号。可能会发

生这样的错误：有时候因为普查员未按正确顺序汇编信息而搞乱了问卷或个人表格的顺序；或者他们有可能跳页，从而无意间在数据集中留下空白页。虽然缺乏排序通常不会影响编辑或指标，但是不少国家统计/普查机构还是选择按适当顺序重新进行个人排序。因此，编辑程序要能够找出失序的个人并重新安排他们的顺序。鉴于重新排序有时会影响到与户主的关系，所以有必要在编辑规范中把这个问题考虑在内。重新排序肯定对诸如母亲的行号或丈夫的行号之类的变量有影响。

5. 居住人与建筑物/住户类型之间的一致性

268. 住户成员之间关系的类别应与居住单元类别相一致。有时候，住户成员出现在一个据称为集体住所的房屋内；反之亦然。在这些情况下，关系类别或居住单元类别必须考虑到房屋大小和其他变项。

E. 重复记录

269. 在光学读取或其他扫描问卷中不大可能产生重复行号。对于需要键入的表格来说，国家统计/普查机构也许选择检查住户清单与准备手动键入的住户行号之间的一致性。这种手工检查可能会提高键入数据的质量，尤其在用(1) 出现在住户所有成员名单那一页的个人姓名与(2) 个人栏、行或页上的数据作比较的情况下，更能提高数据质量。两个人起初似乎记录重复了，但在参考其姓名时，可能发现他（她）们其实是双胞胎。

270. 如果妥善确定了数据筛选和跳转模式程序的话，键入的表格不应有重复行号。当今的大多数软件包作为数据录入程序都可以自动创造序号。在工作人员输入某个人的重复记录的时候可能引进错误，或者一个错误行号可能会造成一项重复记录。在处理每项记录的时候，编辑程序就拿它跟本居住单元的前面人口记录作比较。编辑程序要确保正确捕捉到了每一行号。重复的行号是错误，必须更改。

271. 各国可能决定制定自己的键入模式，而不用现成软件包。这样的话，编辑团队就须确定可接受的误差水平。有许多方法可用于此类决策。一种方法或许是依照下述准则行事：

(a) 如果两个不同记录的行号一样而不一样的特征数是2或以下，那么编辑就删除其中一个记录，因为他可能是重复的；

(b) 如果有三个或以上的特征不相同，那就改变行号。[7]

[7] 传统上，重复记录是由手工追查纠正的。但近年来这些工作至少已经部分地实现自动化了。最近有一篇论文（Winkler，2006年）开始探讨结构编辑和内容编辑一起实现自动化的问题。

F. 特殊人口

1. 集体户的个人

272. 结构编辑应区别对待生活在集体户（如慈善机构、兵营或疗养院）里的人和生活在普通居住单元的人。鉴于集体户通常没有户主，所以各国必须确定如何最好地区分不同类别的居住单元。一种方法就是要有一个不同的集体

户记录类别。另一种方法是指定一个特殊的关系代码，代表“群体”或“集体”住所。

(a) 如果集体户是一种不同的记录类别

273. 如果国家统计/普查机构选择为此单独使用一种记录类别，那么编辑团队就不会在判断哪些记录是集体户的或哪些是集体的记录方面感到难办了。集体户的制表只需直接指明这些记录即可，所以做起来很简便。集体户所特有的一些变项（比如集体类别）可以单独进行编辑和插补。被集体户记录排除在外的变项可以很容易地进行检查，以确信其实为空白。可是会由此产生一个庞大的文件，因为这些记录可能比一般人口记录短，但是却要占用像在一个矩形文件中一样大的空间。另外，在编辑和插补过程中，有时候某些程序可能必须同时检查人口记录和集体户的记录。

(b) 如果用一个变项来区分集体户记录和其他记录

274. 如果使用一个单独的变项而不是单独的记录类别的话，编辑团队可能会更加难以判断哪些记录是集体户的或集体的记录。在这些情况下，仍然可以很容易地进行集体户的制表，只需参考变项本身，这些变项指明哪些记录是集体户个人的。属于集体户所特有的变项，比如集体户类别，依旧可以单独进行编辑和插补。对于被集体记录轻易排除在外的变项可以通过参考集体户代码进行检查，以确信其实为空白。由此产生一种较为紧凑的文件，因为集体户个人的补充记录并非实际需要，而只不过作为一种有着不同住户/集体户变项代码的人口记录被纳入其中。在编辑和插补过程中，有时候程序只需检查人口记录，而不用同时检查人口记录和集体记录。

(c) 如果“集体户类别”代码缺失

275. 表明集体户的代码可能会缺失或无效，或者集体代码和关系代码之间不匹配。如果集体户代码缺失，但是关系代码表明是集体户的话，建议的解决办法是相应地改变集体户代码。如果有集体代码，但是关系代码缺失，或许可以根据集体户的类别确定关系代码。

(d) 如果有集体户代码，但其中所有人都彼此有关系

276. 如果有集体户代码，但是从关系代码组可以看出该居住单元中的所有人都彼此有关系，那么应把集体户代码改成居住单元代码。另一方面，如果该单元按住户编码，但是单元中没有哪两个人之间是有亲属关系的，那或许就有必要将该单元改为群体或集体住所。一个住户中可以有五、六个人互相没有关系，但仍然不是集体户。正如前面强调指出的那样，编辑团队成员之间也许有必要商量解决特殊的罕见情况。

(e) 各类集体户的区分

277. 大多数国家都区分各种不同类别的集体户。它们往往把信息进一步细分成各种特定类别的集体住所。这种信息要么可以作为一个“集体住所类别”项目单独编码，要么可以作为多种可能性纳入住户关系代码组。

2. 难以查点的群体

(a) 季节移民

278. 在一些有季节性移徙现象的国家，采访员需要了解由于参考的时间不同一个居住单元是空置的还是有人住的。因此，即便一个住户有完整的信息，该住户也有可能在别的地方被计数（查点）了。当然，反过来也是如此。如果不加小心的话，那些在不同的地方有两套住所的住户（这些居民有时叫做“雪鸟”，因为他们在一年中的不同季节生活在自己所偏爱的不同地区）就有可能在普查中完全被漏掉。

279. 有时候这种迁移非常有规律：整个住户在一年之内的部分时间生活在一个地方，在剩下的时间里又到另一地方去生活。国家统计/普查机构和编辑团队必须决定如何处理各种不同类型的情况。譬如讲，有些人在每年的部分时间里生活在另外一个住所，比如一些人在温热季节生活在他们国家的较冷地方，而在寒冷季节生活在他们国家的暖和地方。另一种情况就是游牧民，他们在一年当中有部分时间流动放牧，但是在另一部分时间过定居生活——也许国家就选择在一年中的这部分时间里进行普查。

(b) 无家可归者

280. 按照定义，在无家可归者的记录中不应该有住房信息。可是，通过创造一个“虚设的”记录（即起初包含某些变量空白值的一份新记录），可以比较容易地进行结构编辑，并使得记录与其他居住单元的结构相一致。编辑团队不得不拿定主意是否创建这种旨在帮助进行数据处理和制表的虚设住房记录。

(c) 游牧民和生活在边远地区的人

281. 跟无家可归者一样，这些人在结构编辑方面也很困难。有些国家会收集一些“住房”信息，以便用这种信息帮助进行“居住单元”的结构编辑。因此，对这些人的住房编辑有别于在标准单元中使用的住房编辑。人口信息的收集应当跟生活在标准居住单元中的个人一样，并且依照下述准则以正常方式进行编辑。

(d) 临时出国的普通百姓居民

282. 在法律上的普查中，临时出国但其户口所在的住户可以报告他们的居民百姓应当纳入标准人口编辑。而在事实上的普查中，一些指标应表明有关个人暂时不在家，以便摸清法律上的人口和事实上的人口。住房编辑不会因为

有不在住所的人而影响记录。不过很明显，这些人不会纳入事实上的普查；因此他们不会出现在人口编辑中。

(e) 包括无证件者或普查时在港口船舶上的过境者在内，临时住在这个国家但日常不穿越国境的外国平民

283. 由于事实上的普查应包括凡在普查时在国内生活的所有人，所以这些人也应被包括在内。个人应包括在普查时他们居住的地方，并且用人口项目的标准编辑方法进行编辑。如果一个集体户或其他非标准居住单元没有收集住房记录，那么也不对这些个人进行住房编辑。如果港口停泊的船只被视为居住单元，那就应该描述住房特征，并且使用其他船舶的热卡信息进行编辑。

284. 临时住在国内的外国人大概不包括在法律上的普查之内。无证件的人或可包括在内，尤其在那些普查中不单独区分有证件者和无证件者的国家（尽管如此区分一般会产生较好的普查结果）。过境者在编辑后不包括在法律上的普查之中，除非他们过境本地区但平常仍居住在这个国家。如果一艘船通常在这个国家的港口停泊，那么船上的人大概就会像普通居民一样包括在内并且同样进行编辑。

(f) 难民

285. 难民可能住在临时住所并可能需要用某种特定变项、单独记录类别或虚设的住房记录来说明他们的状况。编辑团队需要制定和实施适当的程序。通常，住房和人口项目会使用标准编辑程序，而在热卡中包括一个“难民住房”标识。

(g) 驻在国外的军事、海军和外交人员及其家属以及住在该国的外国军事、海军和外交人员

286. 对于法律上的普查，通常把国内外的军事、海军和外交人员及其家属包括在内。在许多国家，有关军事方面的信息不是通过普查来获得，而该国的统计机构只能进行简单查点，或者尽量少查其他信息。在信息有限的情况下，将难以使用热卡，并且有可能给数据集带来误差；因此，通常倾向于不在普查中以此种方式把军事住户纳入报告范围。外交人员可能有类似的问题。不过，如果使用标准问卷和程序的话，一国之内的查点可能会产生不错的结果；因此，应把这些居住单元纳入正常编辑程序，不过应有某种指标来指明特殊居住单元的状况。鉴于不一定以标准方式对境外的居住单元进行查点，所以在评估把这些单元纳入编辑范围的时候，是否需要小心从事？（在其中一些列表中还是可以将其纳入的。）

287. 对于事实上的普查，通常只把境内的居住单元包括在内。生活在境外的军事、海军和外交住户一般不在普查范围之内。这些人员的住房情况通常由派遣国内住在他们单元的人来报告；当前居民人口要包括在内。

(h) 每日跨越国境来工作的外国平民

288. 无论在事实上的普查还是在法律上的普查中，每日跨越国境来工作的外国平民都不包括在内，因为他们在参考日期不住在这个国家，他们平常也不住在这个国家。无论在事实上的普查还是在法律上的普查中，他们通常都由其派出国来报告。

(i) 每日跨越国境到另一国家去工作的居民百姓

289. 每日跨越国境到另一国家去工作的居民百姓是进行普查的国家的居民，无论法律上的普查还是事实上的普查都要把他们包括在内。他们的住房项目和人口项目均须按照标准方式进行编辑。

(j) 作为这个国家的居民但在普查时人在海上的商业海员和渔民（包括以船为唯一住所的水上人家）

290. 商业海员在纯粹法律上的普查中查点，并且在经过修正的事实上的普查（即经过调整包括没有其他住所者的普查）中查点，但是不在事实上的普查中查点。在纳入普查的情况下，住房编辑需要包括指明特殊的居住类别，但是应有可能在国家标准问卷适用于船舶的情况下，对人口项目进行标准编辑。

G. 确定户主和配偶

1. 编辑户主的变项

291. 在鉴定住户成员方面，传统做法是首先确定户主或参考人，然后根据与户主或参考人的关系依次确定住户的其他成员。户主的定义是在住户中被其他成员确认为户主的人。各国可以使用它们最合适的词语来鉴定这个人（比如户主、住户参考人等），只要这个人是唯一被用于确定住户成员之间的关系就行。建议各国在其公布的报告中提出有关这方面的概念和定义（联合国，2007年，第2.114段）。

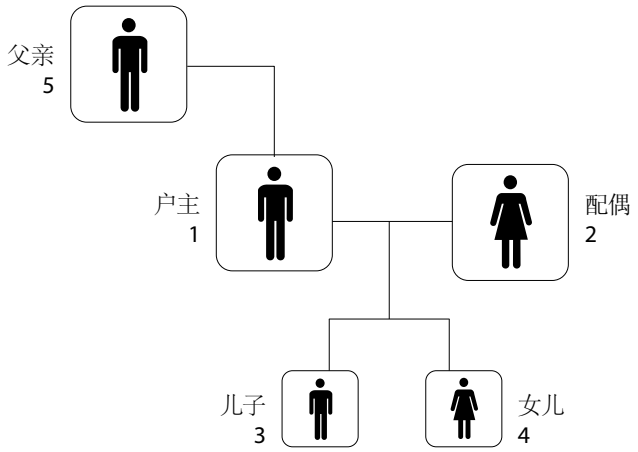
(a) 关系的顺序

292. 单元中的成员关系顺序对编辑有影响，因为许多编辑都假定户主是第一人（个人1），因而他（她）的数据要第一个进行编辑。举例来讲，像语言、种族划分和宗教之类的变项在编辑中都是先检查户主。如果户主有这些变项中的有效信息，该信息即可在发生信息缺失、编码错误或键入错误的情况下用于对本住户中任何其他人的数据插补（见第四章）。之所以需要首先编辑户主，是因为他（她）的特征可用于指定或插补其他成员的值。

(b) 如果户主不是第一人（个人1）

293. 普查员根据他们在现场遇到的与指定户主有关的各种不同情况采取的行动，会影响到编辑过程。为了更好地理解这个问题，让我们先来考虑图25中的解释。

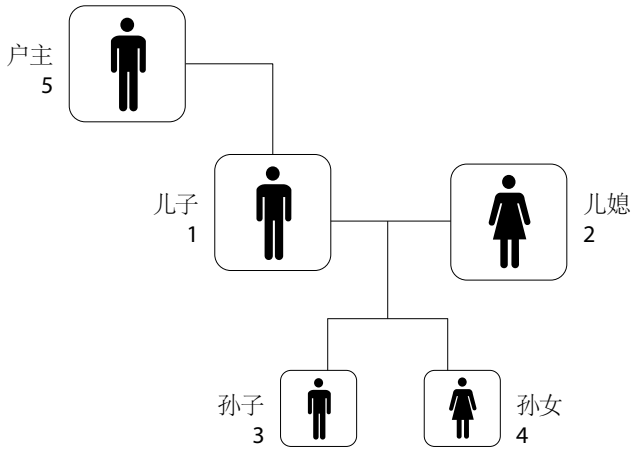
图25
户主被列为个人1的住户示例



294. 这个住户显示了在现场遇到的一种典型状况：户主和配偶、他们的子女、以及户主的父亲。如果普查员按这种方式收集信息，那么基于户主为住户个人1的编辑工作即可顺利进行。

295. 可是，如果将爷爷指定为户主，就得像图26提供的第二种描述那样，按照重新组合的成员关系进行查点。如果一位普查员进入一个房屋，发现一个由丈夫、妻子和两个子女组成的核心家庭，而正在访查过程中，户主的父亲进了屋，并声称他是户主。经这位推定户主的同意，第5人就会变成户主，第1人成为儿子，第2人成为儿媳，等等。

图26
户主被列为第五人的住户示例



296. 从这两个住户示例可以明显看出，在指定不同的人为户主的基础上就会有不同的编辑路径。对于其余的编辑和制表来说，确定实际户主有三种不同的可能性：(a) 可以用一个指示符来标明哪个人是户主，而该指示符的使用可以贯穿整个编辑和制表过程；(b) 如果户主没有被列为个人1，他（她）可移至第一的位置，而名单上位置较高的人可以各自往下移一个位置；或者(c) 可以改变关系代码组，以便让个人1作户主，而不管其他关系如何。

(一) 给户主的记录分配一个指示符

297. 在有关户主的编辑程序中，用一个指示符来确定户主在该单元中的行号。如果户主依旧在收集数据时的位置，就可以把指示符定在那个位置上，而在特定编辑或制表所需的任何时候都可以轻易找到户主。可以给户主的行号确定一个“户主指示符”变项，并且在编辑过程中用它来分配或插补单元中其他人的缺失值或无效特征。如果户主是该住户中的个人1，“户主指示符”的变量值就是“1”。

(二) 让个人1作户主

298. 编辑团队可能选择将户主移至该住户的第一位置。这方面的编程多少要比上面(一)的要求更复杂一些。数据处理专家必须制订一个把户主移至名单第一位置的程序，接下来是原先排在第1的人，然后是本来排名第2的人，依此类推，直至达到紧挨户主前面的那个人。于是，如果户主是在排名第5的位置，那么个人排名顺序就会从1、2、3、4、5变成5、1、2、3、4。经过这一改变，户主将排在第一位，这会使剩下的编辑更容易些，因为户主将永远保持在这个位置。虽然如此，但是如果这么做的话，就可能对数据集的完整性造成一定的“损害”。因为个人的顺序业经改变，分析者可能难以确定从现场收集的个人实际顺序和这种顺序对解析结果的潜在影响。

(三) 重新指定关系代码组，以便让个人1作户主

299. 如果编辑团队决定把名单上的个人1作户主，那就需要在编辑中遵循下述程序(a)和(b)：

(a) 给个人1分配户主的值；

(b) 实施例程，给该住户的所有其他人重新配值，以调整住户。

300. 举例来讲，在图26中，开始时父（母）亲是户主。当个人1被指定为户主的时候，个人2就需要被指定为“配偶”，个人3和4将被指定为“子女”，而个人5将被指定为“父（母）亲”（图25所示）。子例程将需要包含一个矩阵，以保持初始的和已改变的值。

301. 这一程序甚至在更大程度上影响到数据集的完整性。没有像前例中那样改变个人顺序，分析者不难确定从现场收集的个人实际顺序。可是，所有关系将会改变，分析者将不知道哪个人起初被选定为户主。而且，如果在普查或调查中收集了母亲的个人号、父亲的个人号或配偶的个人号，必须在任何再编码方案中考虑到这些号码。另一方面，把户主置于第一的位置，也许名义上

制表比较容易。与前例所不同的是，对于这一程序编程员无须实际上把记录移来移去。

(c) 不止一个户主

302. 如果发现了不止一个户主，编辑团队必须决定指定谁为户主。必须根据由主题专家和编辑流程确定的特征进行编辑。然后编辑程序要重新指定原本被认定为户主的其他人的关系。

303. 在一些国家存在允许“双户主”的特殊现象，究其原因，要么是社会经济状况决定的（比如男性户主频繁离家从事采矿或其他活动，所以由配偶作户主），要么是因为普查对象坚持主张“平等”。传统上，为了编辑的目的，有必要指定一个且只有一个户主，并且在这种情况下保持记录中的原始数据。然而，联合国（2008年，第2.117段）吸纳了一个关于联合户主的条款。如果一个国家选择包括“双户主”，那就必须在编辑中保留双户主；可是本《手册》此后建议的许多编辑方法也得修改：比如在双户主信仰不同宗教、属于不同种族、或具有不同的其他人口统计和社会特征的情况下，一个人就不再可能被用于插补程序了。

(d) 没有户主

304. 同样，如果发现一个住户没有户主，那么编辑就必须决定指定谁为户主。在这种情况下，有可能需要通过编辑调整该住户其他成员之间的关系。在以这种方式确定户主的时候，要考虑到年龄、教育程度和经济活动等变项，以确定最适合的户主。附件四提供了一个编辑流程示例。

2. 编辑配偶

(a) 在一夫一妻制的社会发现恰好只有一个配偶的情况

305. 如果发现恰好只有一个配偶，“配偶指示符”变项就在后面的编辑中跟踪配偶的行号。这些编辑或许包括查找相反性别的户主和配偶，或适当的年龄差异，或者其他相关特征。（在有同性配偶的国家，编辑可能需要适应当地情况。）

(b) 在一夫一妻制的社会发现不止一个配偶的情况

306. 在一夫一妻制的社会，如果在数据集中发现住户有不止一个配偶的情况，那么编辑必须确定谁是配偶，并且重新指定原本被认定为配偶的其他人的关系。在这种情况下，主题专家也必须确定编辑的特征和流程。

(c) 在一夫多妻或一妻多夫的社会发现多个配偶的情况

307. 如果在一夫多妻或一妻多夫的社会发现住户有不止一个配偶的情况，编辑团队也许想要对此信息听其自然，或进行一些一致性检查。譬如讲，

最起码每个一夫多妻或一妻多夫配偶应该是户主的异性。如果发现有同性配偶，应适用先前的同性配偶编辑原则。

(d) 户主和配偶的其他特征

308. 在总体编辑鉴定户主和配偶部分的时候，好的编辑做法需要插补户主和配偶的其他重要项目。这些项目包括户主和配偶的年龄及婚姻状况，日后在插补文件时以及为了别的编辑目的可能会需要这方面的信息。在开始时获得有关户主的宗教、种族划分和语言等方面的“社会”项目也是个好主意，尤其在户主没有被列为第一人的情况下。鉴于大多数编辑软件包都从第一人入手，然后依次往下进行，所以有必要在编辑单元中其他人之前先搞定户主的信息。附件四提供了关于配偶编辑的一份示范流程图。

H. 年龄和出生日期

1. 如果有出生日期但是没有年龄信息

309. 在只有出生日期而年龄不明的情况下，可以通过用普查或调查日期减去出生日期的办法来获得年龄信息。有些国家统计/普查机构选择只根据普查年份和出生年份来获得年龄信息，用这种方法得出的数值有可能产生偏差。如果使用年份和月份，年龄会比较准确，但使用年-月-日的结果准确度最高。

2. 如果年龄和出生日期不一致

310. 在普查或调查获取的数据既有年龄又有出生日期的情况下，“计算的”年龄是通过用参考日期减去出生日期的办法获得的。如果该值比报告年龄相差一年以上的话，编辑团队或许就要采取补救行动了。正常情况下，出生日期比报告年龄更重要；而计算的年龄要被报告年龄所取代。

I. 无效输入项的计数

311. 一些编辑团队可能选择在实际动手编辑之前清点主要变项（比如年龄和性别）或所有变项的无效和不一致输入项目数的程序。如果编辑团队事先有所准备或者使用同样这些项目进行过阶段性调查，他们可能有好几种不同的动态插补数组可供使用。要是现有的无效或不一致的输入项目比例很小，编辑团队也许决定仅使用几种现有的变数进行插补。如果误差比例较大，编辑团队就可能需要使用较多的变数来说明所需的大量插补。

312. 通常，较小的插补矩阵较为可取，因为在开发编辑和插补程序的时候这种矩阵比较易于检查，而且在实际编辑过程中也比较容易使用。可是如果数值反复使用的话，就需要有较大的多样化插补矩阵。

第 四 章
人口项目编辑

313. 第四章是关于人口项目编辑，其中包括与人口统计、人口迁移、社会和经济特征等有关的项目。这些编辑的详细说明考虑到个别项目的有效性和各人口项目之间以及人口项目与住房项目之间的一致性。对各种项目间的关系有了一定的了解之后就可以设计已执行的编辑程序，以确保高质量的制表数据。譬如讲，人口记录不应该出现15岁的女性有10个子女或7岁的子女上大学的记录。

314. 在指定人口项目值的时候，编辑团队必须决定是否指定“未报”值；一个未知的静态插补（冷卡）值或其他值；或者基于其他个人或居住单元特征的一个动态插补（热卡）值。

315. 在许多情况下动态插补法比较可取，因为它排除了制表阶段的编辑过程，而在此阶段只有表内的信息本身才能用来对未知数做出抉择。在没有已作有效回答的其他相关项目的情况下，插补矩阵可为空白、无效输入项、或已找出的不一致之处提供输入项。一些国家在人口特征方面全国有多样性，但是就大部分地区而言差异很小。在其他一些国家，各地区之间可能有相当大的差异，尤其是城乡居住方面的差别很大。在设计插补矩阵的时候，尤其在确定冷卡值的时候，一定要考虑到上述差异。编辑团队要具体指明在什么情况下应为空白提供输入项。这种输入项应取自有类似特征的上一个居住单元。

316. 所有人口记录都要有序列号，以助数据处理。第三章描述了结构编辑检查序号与序号顺序之间的一致性。

317. 编辑团队应当只编辑每份人口记录中的适用项目。编辑的项目可以因城市/农村、气候和/或其他条件而异。最好根据这些条件有选择性地进行编辑，但在实际当中，很少有国家有足够的时间和技能制定多重数组，用以改变缺失或不一致的数据。实际实施这种被采纳的程序的国家就更少了。

318. 就问卷收集的信息有时也仅适用于部分人群。譬如讲，生育力的问题只向妇女提问；经济活动问题只向成年人提问。

319. 有的时候编辑团队应该允许某些项目输入“未报”字样。编辑团队也许对插补一些特征的回答缺乏可靠依据。要对照生产适当的列表特征以供制订计划和政策使用的要求，对是否让项目回答保留“未报”状态的决择权衡利弊。只要“未报”的情况与报告的情况的分布持平，那么在规划者需要有选择的信息的时候分配“未报”项目就不会出问题。不过，如果“未报”的情况以

某种方式呈现扭曲，汇编后的插补就可能成问题，尤其对小地区或特殊类的状况更是如此。举例来讲，如果十几岁的女性普查对象拒绝说出其生育力情况而没有收集到这方面信息的话，编辑程序就没有能力帮助获得这种信息。

320. 人口编辑通常要比住房编辑复杂，因为一般情况下交叉分组列表要复杂多了。大多数国家只用各种不同的地理层次来汇编个体住房特征，但在人口项目方面则可能有许多层次的交叉分组列表。如上图所示，决定不使用动态插补的国家应确定某种指示符来指明“未知的”数值，以便在发生无效或不一致回答时使用。

321. 对于使用动态插补的国家，编辑团队要设计带有区分人口特征的维度的简单插补矩阵。对于大多数国家来说，年龄组和性别是动态插补的最佳基本变项，应首先对它们进行编辑。采用多变量编辑方法的国家统计/普查机构要同时编辑年龄、性别和其他变项（比如关系和婚姻状况）。可能在动态插补中有用的其他项目包括受教育程度和就业状况。

322. 编辑团队必须十分谨慎，不要在插补过程中扭曲数据。团队不要设想插补的和未插补的数据需要有同样的分布。未知的数据本身就是扭曲的。譬如讲，老年人不大可能像年轻人那样报告自己的年龄。

A. 人口统计特征

323. 关于每个人的关系、性别、年龄和婚姻状况的数据是任何普查的基本数据，或许应该一起编辑。人口或分组人口的年龄和性别结构几乎对所有基于人口普查的计划工作都是不可或缺的。这些项目对于制作有用的表格也至关重要，因为基本上所有其他分析都以按年龄和性别分列的其他变量的交叉分组制表为基础。

324. 本《手册》第二章介绍了人口和住房数据的多变量(费勒吉-奥尔特)编辑方法。鉴于各种人口统计变项是构成所有普查规划整体所不可或缺的，在时间和专门技术条件允许的情况下要尽量使用这种方法。几乎可以肯定，优先检查年龄和性别及其他选定的变项以确定误差或不一致问题的编辑重点，有益于确保数据集的总体质量。首先编辑误差率最高的项目，其次是误差或不一致问题较少的项目。

1. 关系

325. 关系项目用于帮助确定住户和家庭结构。它一般出现在普查和调查问卷开头的部分，这有助于确保居住单元的每个成员都能查点到。普查员和普查对象使用有关住户成员之间关系的信息来确保没有人被漏掉。关系项目还有助于检查住户成员之间性别和年龄上的一致性。结构编辑中包括确定每户只有一位户主和没有多于一个的配偶（在一夫一妻制社会）。

(a) 关系编辑

326. 鉴于关系方面的统计数据变得日益重要，在为各类制表目的开发有关家庭和亚家庭构成的编辑程序时要小心行事。首先，设计适当的关系代码组显然对这方面的工作是有益的（见附件一中有关“家庭类型”以及亚家庭号码和亚家庭关系重新编码部分）。

327. 在不能指定关系因而不能使用动态插补的情况下，必须指定无效或不一致的回答为“未知”项。在使用动态插补的时候，可以根据按年龄和性别及其他适当特征分列的插补矩阵指定关系。插补矩阵不应插补与住户中既定关系相抵触的关系。举例来讲，即使在一夫多妻或一妻多夫的住户中也不要插补第二和第三配偶，除非编辑团队决定实施此种编辑。

(b) 如果户主必须排在第一位

328. 如果户主没有作为第一人出现，第三章介绍的结构编辑指出可以使用一个指示符来跟踪户主的位置。如果编辑团队想要让户主作第一人，那就可以像结构编辑那一章指出的那样，要么通过重新安排个人顺序，要么通过保持住户不动而重新安排关系来把户主排在第一位。前一种方法需要有相当高的编程技能，而后一种方法，若不极为小心的话，就可能给数据集带来损害。

(c) 如果关系编码颠倒

329. 有时候，普查员收集的关系“颠倒”了：他们不是收集每个住户成员与户主的关系，而是收集户主与每个成员的关系。因此，作为第三人的关系是“父（母）亲”而不是“子女”。结果，住户可能有四个父（母）亲而不是有四个子女。当编辑团队发现此类系统性问题的时候，就必须设计一个不会过分损伤住户的解决方案。

330. 关系换向程序通常涉及到运行一个包含“原有关系”和“反向关系”的“查找”文件，其中考虑到普查对象的性别。

(d) 如果存在一夫多妻或一妻多夫的配偶

331. 如果按第三章指示进行的话，结构编辑就已经检查了是否有“一个且只有一个”户主和一夫一妻制的住户“没有多于一个的配偶”。对于一夫一妻制的住户，编辑团队应决定何时允许有多配偶关系，何时不允许有多配偶关系。有的时候，住户似乎有多配偶关系，但实际上是个错误。

332. 举例来讲，一个住户或许确定有一个户主和一个配偶，但是又报告了另外一对互为“配偶”的人，于是就总共有了三个配偶。编辑应检查核实第二对配偶关系不是实际上的父亲和母亲、儿子和儿媳、姐姐和姐夫（或妹妹和妹夫）、或其他某种组合。有时候能够在一定程度上有把握地确定这些关系，可有时候不能。在给上述明细关系编码的时候，编辑团队要预见到会有适当插补。在一夫多妻或一妻多夫的住户，当有额外配偶为真配偶的时候，编辑要检查性别，或许还有年龄。

(e) 如果出现多个父母亲

333. 住户报告的“父母亲”不应超过两个以上，而且父母亲应该互为异性。如果出现了多于两个的父母亲，那么多出来的父母亲大概就应算作“其他亲属”了。有时候，普查或调查有一个“父（母）亲”或“岳父（岳母）”的代码，这样就可以有四个而不是两个“父母亲”了，而每种性别的父母亲不应多于两个。

(f) 如果普查收集限定性别的关系

334. 有的普查或调查收集限定性别的关系：分别为“丈夫”和“妻子”，而不是“配偶”；“儿子”和“女儿”，而不是“子女”；诸如此类。如果这些回答不加编辑，表格中就可能包含“男性”女儿或“女性”丈夫之类的数据。编辑团队必须确定编辑的优先次序——是先编辑关系，还是先编辑性别。某些情况（比如丈夫和妻子）要比其他情况（比如年幼子女）的编辑更重要。请注意，最好不要使用限定性别的亲属关系，因为冗余非但不会澄清反而会掩盖关系的性质，并因此而需要做更多的编辑工作。

(g) 如果亲属关系和婚姻状况不相匹配

335. 当亲属关系和婚姻状况重叠时应彼此一致，即：报告“配偶”关系的人在婚姻状况项目中应该是“已婚”。如果项目不一致，编辑团队就需要决定修改哪个变项。有时候关系模糊不清，所以在制订编辑规范的时候要加小心。譬如讲，在许多国家，英文brother-in-law既可能是配偶的兄弟（称内兄或内弟，且不一定已婚），也可能是同胞姊妹的配偶（称姐夫或妹夫，且必定已婚）。

336. 目前在关系报告方面还出现了其他几个更现代的问题。如果两个异性未婚者婚外同居，其关系代码或许是“未婚伴侣”或者也叫“配偶”。如果普查或调查中有未婚伴侣的代码，那么适当的婚姻状况就不应该是“已婚”了，除非此人已经和与之同居者以外的另一人结婚。

337. 同样，现在也有同性恋者以情人或非情人的关系同居。有非情人关系者或可编码为“同居者”或“非亲属”。对于有情人关系者，在一些国家或许适合纳入“未婚伴侣”类。然后，编辑团队还必须确定对应的适当婚姻状况。普查不能区分柏拉图式的（纯友谊）关系和情人关系。

2. 性别

338. 性别是最容易收集的特征之一，但是在有关的编辑方面却需要一些思考。它是最重要的变项之一，因为大凡人口特征都是在性别基础上进行分析的。性别插补需要与其他变项作些比较。有些情况下，性别应以相关个人间的性别差异为基础，通常在户主和配偶之间，但也在父母亲和岳父岳母之间，大概不应让性别栏填上“无效”或“未知”字样，因为它是个如此重要的变项。所以应考虑如何最有效地获得有关国家实际情况的适当结果。附件四载有一个户主和配偶性别编辑流程图示例。

339. 如果一个人不是户主或户主的配偶，就没有其他人可参考了；因此应检查此人记录中的其他项目。如果有适当的生育力项目发生，即应指定女性代码。可是，如果此人性别缺失或无效，但是有一配偶，其性别业已指明，那么编辑即可为此人插补性别。

(a) 如果性别代码无效，但户主和配偶为同性

340. 在明显自相矛盾的情况下应改变性别代码，即便存在有效代码。举例来讲，记录表明，在已经有一户主和配偶的住户或者在有已婚夫妇的亚家庭又有第二对夫妇。如果这第二对夫妻二个人报告了同一性别，就可以用有关生育的信息和其他项目来判断谁是男的谁是女的。然后即可纠正错误记录。

(b) 如果一名男子有生育信息或一名成年女子没有生育信息

341. 编辑过程中可能查出一名男子有生育信息并且/或者在本住户有子女，这个错误可能是母亲的个人序号或某个类似变项造成的。如果没有配偶记录，可将性别改为女性，而不是删除生育信息。同样，如果一名成年女子没有生育信息并且没有伴随子女，那么，在编辑团队认定的某些情况下即可将其改为男性。

(c) 如果性别代码无效而有配偶在

342. 如果性别输入项空白或无效，编辑程序应使用与户主的关系和配偶性别（如果配偶性别有效的话）的输入项目来判断正确代码。如果与户主的关系是“户主”，那么编辑程序就检查有无配偶（即检查该住户中报告关系为配偶的另一人）。通过确定配偶的性别代码，即可给户主指定相反的性别代码。

(d) 如果配偶的性别代码无效

343. 如果这个人与户主的关系是“配偶”，而户主的性别已定，那么编辑程序即可将与户主性别相反的性别代码指定给此人。

(e) 如果性别代码无效而有女性信息

344. 问卷中有无计其数的迹象可以表明普查对象是不是女性。如果编辑程序尚未确定某个人的性别而存在任何女性指示符的话，这个人就应被指定为女性。举例来讲，如果被编辑者有足够的生育力项目，那就可以将其性别指定为女性。生育力项目包括既往平均生育数、生活在该住户的子女人数、在别的地方生活的子女人数、已亡子女人数，以及过去12个月内出生并健在的子女人数。另外一种可能性就是此人还可能是该住户中另一个人的母亲，因而此人的行号等于该住户另一人的母亲的行号。

(f) 如果性别代码无效而此人是配偶的丈夫

345. 如果这个人是该住户的另一人的丈夫，那么根据现时这位丈夫行号的项目，应指定其性别输入为男性。

(g) 如果性别代码无效而又没有足够的信息判断性别

346. 如果编辑团队根本不使用动态插补，那就必须指定一个未知性别值。不幸的是，这就意味着所有表格都要为不知性别的人额外载有一个分栏或一行或额外的成套分栏或行数。由于性别是一个二元变项，可以交替指定数值，无论先从哪种性别开始第二个无效输入项就用相反的性别，如此循环往复进行下去。

(h) 关于性别插补比率的说明

347. 在使用冷卡插补法的情况下，指定女性的概率较高。只有成年女性才有生育力方面的输入项目，因此对她们的随机选择就容易被扭曲。由于这个缘故，如果缺乏足够的可用信息的话，没有信息的人就多半是男性而不是女性。所以有必要考虑设计估计到两性总体比例的插补矩阵。

3. 出生日期和年龄

348. 年龄是最难收集和编辑的特征之一。不过，它大概也是最重要的变项，因为几乎所有人口特征都是在年龄基础上进行分析的。年龄的编辑需要同其他变项以及住户中其他人进行广泛比较。大多数情况下，插补的年龄应该以储存的相关个人之间的年龄差为基础，如果不能在此基础上进行年龄插补，那就应使用有关个人记录中的其他特征。编辑程序大概需要一系列插补矩阵，其中包括年龄与性别、婚姻状况、亲属关系和受教育程度等；母亲与子女之间的年龄差；丈夫与妻子之间的年龄差；以及户主和配偶之间的年龄差。

(a) 年龄和出生日期

349. 结构编辑根据出生日期计算年龄。不过，首先有必要审视年龄与出生日期之间的区别。正如《人口和住房普查的原则和建议》第二次修订本（联合国，2008年，第2.136段）所述，年龄信息可以要么通过了解出生日期（年-月-日）、要么通过直接询问此人上一次生日时的年龄来获得。

350. 出生日期产生了较为精确的信息，只要情况允许，要尽量使用这种信息。如果出生的确切日期和月份都不知道，可代之以指明出生那一年的季节。在人们知道其生日的情况下，提问出生日期是适当的，既可以按阳历确定也可以按阴历确定，抑或用年份数码来表示，或者按照传统的民间习俗用正圆圈内的名称来辨识。

351. 不过极为重要的是，普查员和普查对象之间要就基于哪种日历达成明确共识。如果可能有些普查对象得靠不同于其他普查对象的另一种日历来表明自己的生日的话，必须规定在问卷中注明所用的日历制度。普查员最好不要试图从一种日历日期转换为另一种日历日期。所需要的转换最好作为计算机编辑工作的组成部分来进行（联合国，2008年，第2.137段）。

352. 由于多种原因，直接提问年龄有可能得到不太准确的回答。即便所有回答都基于同样的年龄认知方法，普查对象也不一定懂得所需要的是其上一

次生日的年龄，还是下一次生日的年龄，抑或是最近生日的年龄。此外还可能发生其他问题：可能按照以0或5结尾的四舍五入方法确定年龄；估计数可能不这么确定；而故意谎报相对而言可能更容易（联合国，2008年，第2.138段）。

方 框 4

年龄估算和插补

年龄的估算和插补应该：

- (a) 在年龄空白处指定年龄；
- (b) 检查已婚者的最小年龄；
- (c) 检查户主的最小年龄；
- (d) 检查父母亲的最小年龄；以及
- (e) 进行国家规定的其他检查。

353. 许多国家统计/普查机构要么收集出生日期，要么收集年龄数据，但不是两项数据都收集。《人口和住房普查的原则和建议》第二次修订本（联合国，2007年）指出，用整年头表示的年龄非常重要：它被用于许多编辑程序，并且是许多插补矩阵的一个维度。更重要的是，不少国家政策都以年龄为基础，所以必须努力获得最高质量的年龄报告。然而，即便在理想情况下也会有些人未报告年龄。因此必须努力确保适当插补年龄，并且使之与住户成员的其他回答相一致。

(b) 出生日期与年龄之间的关系

354. 在结构编辑过程中，如果没有单独收集年龄信息的话，应根据出生日期计算年龄。在个人编辑过程中的年龄编辑将是对各种记录内部和记录之间一致性的彻底检验，但是第一步要首先根据出生日期和普查日期计算年龄。有必要检验根据出生日期计算的年龄，以确保它是在普查日期的界限范围内。

355. 在普查当年、但在普查日期以后出生的子女，其年龄将被计算为——1岁，而且必须经过核实。在普查日期以后查点的婴儿大概应该从普查中扣除。可是，如果经审查发现由于查点或数据处理的缘故出生日期是错误的，要使用其他变量获得较好的年龄估计数。

(c) 如果计算的年龄高于上限

356. 对于2000年及其后的普查来讲，大多数国家将选择记录完全四位数的出生年份。对于那些在2010年前后的出生年份来说，可接受的范围是在1900年代或2000年代直至普查当年。虽然对计算机工作来说三位数就够用了，但是若使用三位数的年份，可能会让普查员和办公室人员搞混了。有时候计算的年龄会高于普查界定年龄的上限而需要调整。如果普查是在2010年进行的，而某人报告在1860年出生，那么150岁的计算年龄可能就超出了可接受的范围，所以需要更改。

(d) 年龄编辑

357. 编辑程序应检查个人报告年龄与其母亲、父亲或子女的报告年龄的一致性。编辑规则应规定父母亲年龄与其子女年龄之间的最低限度差异。在进行年龄插补的时候，要对诸如在本区生活的年头（居住期间）和完成学业的最高学历（教育程度）之类的输入项目进行一致性检查。所有这些检查都应在改变年龄之前或在指定插补年龄之前进行。

358. 编辑应从检查有效性入手。如果年龄有效，专家或许想要查看此人年龄是否与其母亲的年龄相一致（如果在住户中发现其母亲的话）并且与其子女的年龄相一致（如果这个人是妇女并且在本住户有其子女）。如果年龄不一致，此人的年龄应予注明，以便在日后更正。

(e) 有户主和配偶存在的年龄编辑

359. 下一步编辑是确定配偶是否存在。如果配偶在，要检查配偶年龄的有效性（根据国家规定的最低结婚年龄，至少X岁）。如果年龄不一致并且使用动态插补法，这个人就将使用一个根据丈夫和妻子之间的年龄差推导出来的特殊插补值。年龄差的变化小于年龄本身，所以程序中的插补矩阵将储存夫妻间的年龄差（根据以往的记录）。该值追加到此人配偶的年龄或从中扣除，以形成一个计算年龄。

360. 为了确保这个计算年龄与其他特征相一致，插补矩阵还应包括婚姻状况、居住期间和达到的最高学历。如果不包括这些变项，就可能导致计算的年龄少于此人在此地居住的年头，或者少于接受学校教育的年头。譬如讲，插补矩阵可能给出8岁的年龄，但是此人的记录也许表明他（她）已在此居住了10年。假如没有别的变项可供参考，那么当编辑程序进行在此地居住年头编辑的时候，另一插补矩阵就会把居住年头从正确值改成错误值。

(f) 如果户主配偶不在但子女在，户主年龄的编辑

361. 如果不能通过与户主配偶的年龄作比较来确定户主的年龄，那么编辑程序可以检查亲属关系。如果关系是“户主”，编辑程序可以检查该住户中已知年龄正确的其他人——比如儿子和女儿——的记录（如果有的话）。程序检查儿子或女儿的年龄，并使用类似于上述夫妻“年龄差”的动态插补法计算此人年龄。跟前面一样，计算的年龄考虑到居住期间和最高教育程度。然后填完的年龄将与这些变量相一致，并且避免通过纳入作为插补矩阵组成部分填写的在本区居住年头和最高学历而产生明显的错误。

(g) 如果户主的父（母）亲在，户主年龄的编辑

362. 如果一个人不属于上面讲的任何一类，编辑程序可以搜寻这个人在该住户的父母亲。如果发现了此人的父（母）亲，那就可以使用带有年龄差的插补矩阵来计算年龄。子女和父母亲之间的年龄差一般要比夫妻间的年龄差大的多。因此，编辑程序只在夫妻年龄差的方法失败的情况下才应用这种编辑

法。计算的年龄要考虑到包括最高学历在内的教育特征、在本区居住的年头、婚姻状况、生育力以及经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的最低年龄应该是多大。

(h) 如果户主有孙子孙女，户主年龄的编辑

363. 如果一个人不属于上述任何一类，编辑程序可以搜索此人在该住户的孙儿孙女。如果发现了此人的孙儿孙女，即可用一个带有年龄差的插补矩阵来计算年龄。户主与孙儿孙女之间的年龄差要比夫妻间的年龄差或户主与子女之间的年龄差大多了。因此，程序只在夫妻间和户主与子女之间年龄差编辑失败后才应用这一编辑。计算的年龄应考虑到包括最高学历在内的教育特征、在该区生活的年头、婚姻状况、生育力和经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的最低年龄应该是多大。

(i) 如果没有其他人的年龄可参考，户主年龄的编辑

364. 如果一个人不属于上述类别之一，那么编辑程序可以搜索户主的一个其他亲属或非亲属。如果发现了这么一个人，而这个人报告了年龄，那么编辑团队必须决定是否使用现有的带有年龄差的插补矩阵的任何信息。不过，由于户主与其他亲属或非亲属之间的年龄差太大，编辑团队也许会决定干脆放弃努力，而只使用其他变项进行户主年龄的动态插补。不论在哪种情况下，只有在夫妻间的、户主与子女之间的、户主与父母之间的、以及户主与孙儿孙女之间的年龄差的方法均已失败之后，才应用这种编辑方法。不管计算的年龄是如何确定的，它都要考虑到包括最高学历在内的教育特征、在该区生活的年头、婚姻状况、生育力和经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的最低年龄应该是多大。

(j) 如果户主年龄已定，配偶年龄的编辑

365. 配偶年龄的编辑通常与户主年龄的编辑同时进行，因为联合编辑需要来自这两个人的信息。可是如果分开进行编辑的话，在配偶的年龄无效或者与其他变项不一致的情况下，应该使用带有户主与其他变项年龄差的动态插补矩阵来确定配偶年龄的最佳估计数。跟前面一样，计算的年龄应考虑到包括最高学历在内的教育特征、在该区生活的年头、婚姻状况、生育力和经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的最低年龄应该是多大。

(k) 如果已知住户成员之一的年龄，其他已婚夫妇年龄的编辑

366. 编辑应首先确定这是不是一个已婚者的记录。如果是，程序即可在该住户的其他记录中搜索这个人的配偶。如果没有发现配偶，程序就进入下一部分编辑。如果发现一位配偶，要检查配偶年龄的有效性（根据国家规定的最低结婚年龄至少应该X岁）。如果年龄不一致并且要使用动态插补法的话，现

在程序将要使用一个根据丈夫与妻子之间年龄差推导出来的一个特殊插补值。年龄差要小于年龄本身，所以程序中的一个插补矩阵会储存这种夫妻间的年龄差（根据以往记录）。该值将追加到此人配偶的年龄上或者从中扣除以形成一个计算年龄。

367. 为了确保这个计算年龄与其他特征相一致，插补矩阵还应包括婚姻状况、居住期间和最高教育程度。若将这些变项排除在外，计算的年龄就有可能少于此人在此地生活的年头，或少于其接受学校教育的年头。

(l) 如果户主年龄已定，子女年龄的编辑

368. 如果这个人是户主的儿子或女儿，可以使用户主的年龄、年龄差、居住期间和受教育程度推导出一个计算的年龄。其计算的年龄也应考虑到包括最高学历在内的教育特征、在该区生活的年头、婚姻状况、生育力和经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的最低年龄应该是多大。

(m) 如果户主年龄已定，其父母亲年龄的编辑

369. 如果这个人是户主的父（母）亲，可以使用户主的年龄、年龄差、居住期间和受教育程度推导出一个计算的年龄。这个计算的年龄应考虑到包括最高学历在内的教育特征、在该区生活的年头、婚姻状况、生育力和经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的最低年龄应该是多大。

(n) 如果户主的年龄已定，其孙子孙女年龄的编辑

370. 如果这个人是户主的孙子或孙女，可以使用户主的年龄、年龄差、居住期间和受教育程度推导出一个计算的年龄。这个计算的年龄也应考虑到包括最高学历在内的教育特征、在该区生活的年头、婚姻状况、生育力和经济活动等。编辑程序应假定要是一个人结了婚、有了子女或报告了所从事的任何经济活动的话，他（她）的年龄至少应该是12岁。

(o) 所有其他人年龄的编辑

371. 编辑团队应确定适用于该住户其他亲属或非亲属的插补矩阵。相关的指导原则取决于特定的普查或调查以及国家的社会和经济特征。举例来讲，一个结了婚、有了孩子或参加经济活动的人可能至少在国家规定的最低年龄以上。根据这一信息，如果使用动态插补法的话，从插补矩阵得到的值就不应低于最低年龄。同样，如果一个人正在上学、曾经上过学或有一定读写能力、但不是户主、未曾结过婚、也没有任何经济活动的话，那么这个人就应该被置于一个年龄小于成年人最低年龄但是大于或等于最低入学年龄的群体之中。然后即可发现适用于小于最低入学年龄者的插补矩阵值。这种方法虽然不太完美，但是它限定了插补矩阵可取的数值范围。

4. 婚姻状况

372. 在《人口和住房普查的原则和建议》第二次修订本（第2.144–2.151段）中，婚姻状况的定义是指每个人涉及国家婚姻法律或惯例的个人状况。需要鉴定的婚姻状况类别包括但不限于：(a) 单身（从未结婚）；(b) 已婚；(c) 丧偶未再婚；(d) 离异未再婚；以及(e) 已婚但分居。在一些国家，可能(b)类需要再细分一个亚类，即契约上已婚但夫妻双方尚未在一起生活的婚姻状况。在所有国家，(e)类均应包括法律上分居和事实上的分居，如果需要的话可以用分居亚类来表示。尽管夫妻分居仍然可以被视为已婚（因为双方都不能自由再婚），但是(e)亚类的任何一方都不得纳入(b)类。在一些国家有必要考虑到习惯婚姻（这是合法婚姻，根据习惯法有约束力）和超越法律权限的婚姻，后者往往叫做事实上的（两愿的）婚姻。

(a) 婚姻状况编辑

373. 编辑团队必须为普查或调查确定适当的初婚年龄。一个国家的不同地方或不同族群可以有不同的最低初婚年龄（某个年龄X）。譬如讲，农村人口一般比城市人口结婚早，编辑规则应考虑到这一事实。通常国家统计/普查机构在查点之前就确定最低结婚年龄，以便只向高于这个年龄的人提问婚姻状况问题。低于这个年龄的人便自动归入“从未结婚”的一类。不过，如果每个人都要被问到婚姻状况的话，编辑团队就必须开发一种适用于整个人口的编辑程序。

(b) 在不使用动态插补的情况下指定婚姻状况

374. 虽然只有年龄在X及以上的人才填写婚姻状况项目（因为表格中规定最小初婚年龄为X），但编辑团队必须决定是否编辑和在多大程度上编辑这个项目。如果国家仅用“未报”或“未知”的代码来填充无效或不一致回答的话，那么当发现无效或不一致输入项目的时候，就该用“未报”代码来取代不恰当的回答。如果X岁以下的人在回答中漏报“从未结婚”，即应予以插补；鉴于各国调查机构都要向公众散发问卷数据样板，所以像“婚姻状况”之类的项目一定要有输入内容。

(c) 在使用动态插补的情况下指定婚姻状况

375. 如果使用动态插补，婚姻状况的编辑就要：(a) 如果输入项目超范围就予以插补；(b) 检查填报的婚姻状况是否与关系和年龄相一致。

(d) 配偶应该是已婚的

376. 凡是关系类别代码为“配偶”的人，均应被编码为“已婚”。

(e) 一对已婚夫妇的配偶

377. 如果个人A的配偶（个人B）的行号是一个变量，那么个人B就应有个人A作配偶；而且，A和B均应已婚并互为异性。

(f) 如果配偶在，户主应当已婚

378. 如果婚姻状况一栏没有输入项，但是与户主的关系一栏输入了“户主”，那么程序就要查看是否有配偶存在（通过检查该住户其他成员的关系）。如果有配偶，程序就指定户主的婚姻状况为“已婚”。

(g) 户主无配偶亦无子女

379. 如果没有配偶，而此人是男性、有子女的话，那么程序就按照有子女的年龄插补婚姻状况。如果此人没有子女，程序或许就按无子女的年龄插补婚姻状况。如果一位男性户主在户内没有妻子，那他的婚姻状况很可能是离异、分居或者丧偶。

(h) 如果所有其他方法均告失败，那就插补婚姻状况

380. 对于一个有超范围代码因而不能根据上述检测予以指定代码的人，下一步就应检查年龄。如果年龄有一个不满X岁的输入项，那就应该为其指定“从未结婚”。在所有其他情况下，均应使用插补矩阵指定输入项。应通过下述项目建立插补矩阵：性别和年龄（二维）；性别、年龄和关系（三维）；或者性别、年龄、关系和平均生育数（四维）。这里，编辑团队也应确定编辑顺序，所以在涉及插补矩阵的时候有必要记住哪些项目已经编辑过，哪些尚未编辑过。如果在婚姻状况之前只编辑过性别和关系，那么插补矩阵就必须允许其他项目填充“未报”代码。

(i) 年轻人的年龄与婚姻状况的关系

381. 对于已经报告了除“从未结婚”之外的有效婚姻状况的所有人来说，都应该进行一致性检查。所有结过婚的人年龄都必须在X岁或以上；这里，X是一个人所可能结婚的最低年龄。如果年龄不到X岁或这里是空白的话，还要根据其他相关变项（比如平均生育数或从事的经济活动）进一步进行一致性检查。如果这些项目均无效，应给婚姻状况指定“从未结婚”；在所有其他情况下均应更改婚姻状况。

5. 初婚的年龄

382. 根据《人口和住房普查的原则和建议》第二次修订本（第2.192段），“初婚日期”系由初次结婚的年、月、日组成。在一些难以获得初婚日期的国家，可以收集关于结婚年龄或者多少年前结婚（婚姻存续时间）的信息。不仅包括契约上的初婚和事实上的婚姻，而且包括习惯婚姻和宗教婚姻。对于在普查时已丧夫、分居或离异的妇女来讲，应填报其“终止初婚的日期”或“自终止初婚以来的年头”。关于终止初婚的信息（如果恰当）可以提供为计算作为数据处理阶段产生的一个话题的“初婚延续时间”所需的数据。在一些报告的婚姻存续时间比年龄还要可靠的国家，按婚姻存续时间估计的平均生育数列表，能够产生比基于妇女年龄分类的存活子女人数资料更好的生育力估

计数。关于婚姻存续时间的数据可以通过从当前年龄减去结婚年龄获得，或者直接从结婚以来度过的年头获得。

383. 每个已婚者都要输入初婚日期。编辑程序要检查如下一致性：从未结婚的人不应有这方面的信息，但是结过婚的人要报告有效的初婚年、月、日。编辑团队需要决定是否日子和月份一定要有效，就是说：不使用动态插补的国家可以指定月份和日子“不详”；而使用动态插补可以插补缺失的月、日数值。

(a) 从未结婚者的结婚年龄应为空白

384. 从来没有结过婚的人不应报告初婚年龄。如果出现了一个从未结婚者的初婚年龄输入数据，编辑团队必须决定是修改此人的婚姻状况还是将其结婚年龄留作空白。如果要更改婚姻状况，仅使用“未报”字样的国家将使用这个代码。使用动态插补的国家或许应使用年龄和性别项目获得对婚姻状况的适当回答。

(b) 结过婚的人应有输入项

385. 对于初婚年份，不使用动态插补的国家可以指定“未报”或“未知”代码。使用动态插补的国家可以用其他变项——比如配偶年龄或配偶之间的年龄差、子女数和最近一年出生的子女等项目——来确定有关个人的适当初婚年份。

6. 生育力：平均生育数和存活子女数

386. “平均生育数”是存活生育子女总数，因此不包括死产、流产和堕胎。有时候，人口统计学家使用“存活平均生育数”这种表述方式，但本《手册》使用“平均生育数”或“已生子女”。

387. 本节包含的每个主题所应收集的数据全域系由15岁及以上（或其他可接受的某个最低年龄）的妇女组成，而不论其婚姻状况或特定类别（比如结过婚的妇女）。在不收集50岁及以上妇女数据或用这种数据制表的国家，应集中力量只收集15岁至50岁妇女的数据；在近期生育力调查中，一些国家或许可将年龄限制降低几岁（联合国，2007年，第2.170段）。

(a) 收集的生育力项目

388. 在《人口和住房普查的原则和建议》中，联合国（2007年）建议收集三个生育力项目的信息，即：平均生育数、最后一个存活子女年龄、以及母亲生第一个成活子女时的年龄。对有关年龄、日期或婚姻存续时间项目的回答可以改进基于平均生育数的生育力估计数。另外也有许多国家仍在收集有关存活子女的信息，这种信息特别有助于进行追溯的生育力分析。

389. 普查和调查使用国家规定的最低年龄（有时也用最高年龄）来收集所有妇女的生育力信息。

(b) 生育力编辑的一般规则

390. 对小于指定最低生育年龄的妇女和所有男子都要进行检查，并删除其任何生育力信息。

391. 生育力编辑的目的是确保输入数据彼此一致并与年龄相一致：

(a) 存活生育子女总数不得大于有关个人年龄加国家规定的最低年龄乘以一个因子的积数。如果允许女子每年生育一次的话，该因子为1；若生育间隔为一年半的话，该因子即为1.5；依此类推。关于目的为确定母亲与存活最大子女间最小年龄差的编辑，见下面“头胎生育年龄”部分；

(b) 生育子女总数不得多于生活在该居住单元中的、在外地的和已亡子女人数之和。如果总数多于各个部分之和，编辑团队必须确定哪一部分排在前面，以便加以调整；

(c) 如果收集的子女人数既有存活的也有已故的，那么这些子女的总数不得超过生育子女总数；

(d) 平均生育数不得少于“以往12个月生育子女”项目的输入值；

(e) 依不同的国家和实际平均生育数及存活子女数而定，或可使用一个适用于“以往12个月生育子女”的插补矩阵，按年龄和平均生育数分配一个回答。不过，如果出现空白的话，在给“以往12个月生育子女”指定数值的时候一定要非常小心谨慎。对大多数国家来说，这个项目的空白意味着没有生育子女。如果分配数值的话，就可能扭曲数据；

(f) 有时候，各国按性别分列收集平均生育数、存活子女数和其他生育力项目。在这些情况下，这方面的编辑一般都是检查总计数值，但是国家可能想要进行一些补充检查，以说明可用的附加信息。这些补充检查包括确保生育的男孩数是存活男孩和已故男孩的总和，而生育的女孩数是存活女孩和已故女孩的总和。至于未按性别区分的子女的编辑，当总数与各部分之和不相等的时候，就需要采取适当行动了。

(c) 平均生育数和存活子女数之间的关系

392. 关于平均生育数和存活子女数的数据用于间接估计生育率和死亡率状况。普查或调查结果按女性单一年龄和5岁年龄组分列。采用各种不同的算法获取常量和变量死亡率估计数。然而，为了得到最佳结果，编辑团队必须仔细确定对现有数据的适当编辑程序。

393. 与设计一般编辑程序有关的部分问题是不同的国家要求获取不同类别的信息。举例来讲，不同的国家分别收集下列信息组：

(a) 仅平均生育数；

(b) 平均生育数和存活子女数（两性综合或两性分列信息）；

(c) 平均生育数、存活子女数和已亡子女数（两性综合或两性分列信息）；

(d) 平均生育数、身边子女数、非身边子女数和已亡子女数（两性综合或两性分列信息）。

(d) 在只报告了平均生育数的情况下如何编辑

394. 如果国家不用动态插补，“平均生育数”项目的一个无效值或缺失值应被指定为“未知”。在使用动态插补的国家，专家必须决定要不要对所有项目使用动态插补。如果专家使用这种方法，即可根据这个妇女的单一年龄和至少另一特征获得平均生育数。还有可能仅使用母亲单一年龄的单一维度数组。其他特征或许是比如教育程度或宗教信仰之类的项目，因为据了解，在许多国家不同的教育程度或宗教身份会有不同的生育率。

(e) 在既有平均生育数又有存活子女数的情况下如何编辑

395. 如果既回答了“平均生育数”又回答了“存活子女数”，程序就需要确定：

(a) 这些项目是否内部一致（平均生育数是否等于或大于存活子女数）；

(b) 是否每个项目都与该女子的年龄相符；

(c) “平均生育数”是否与“最后一年生的（最近一胎）子女”相符（如果收集了此项目的话）。

396. 人口统计学家使用“平均生育数”和“存活子女数”这两个项目，可以间接获得死亡率估计数。因此，编辑程序必须维持这两个项目之间的关系。有时候这两个项目只报告了一个，而另一个不明。简单的编辑做法就是假定平均生育数中没有已故数，从而使得两个项目一样。可是，使两个项目变得一样的结果是，间接死亡率估计数就不考虑可能在产后死亡的婴儿，因而会低估死亡率而高估预期寿命。如果普查或调查中很少出现此种情况，它不会造成太大的损害。但如果此种情况的发生频率达到一定程度——就像在使用间接估计方法的国家所预期的那样，后果可能会相当严重。图27给出了一个示例。

图27
有生育力信息的住户示例

个人	关系	性别	年龄	平均生育数	存活子女数
1	户主	1	60		
2	配偶	2	60	5	99
3	女儿	2	40	3	3
4	孙女	2	20	1	1
5	孙女	2	18	0	0
6	孙女	2	1		

注： 99 = 数据缺失或无效。

397. 这里，配偶报告总共生育5个子女，但不管因为何种缘故没有得到存活子女数的记录。可能是普查对象或普查员没有报告项目值，也可能是数据

录入员键入了错误信息。许多国家设计了一种编辑程序，可根据平均生育数给子女调查指定数值“5”。可是这种做法会造成数据偏差。

398. 实际上，根本无须改变数值。那些不使用动态插补的国家可以选择用“未知”字样填充。当然，这一抉择也会造成偏差，因为这种编辑决定了对于表格来讲“未知”和“已知”的分布是一样的。如果国家要求填报平均生育数和存活子女数以间接确定死亡率估计数的话，那也许是因为这个国家在数据填报方面有困难。在这种情况下，在数据中保留未知项目可能会造成最终分析误差。无论平均生育数未知还是存活子女数未知的妇女数据都不能用于确定死亡率估计数，因为不能确定平均生育数与存活子女数的差异。

399. 使用动态插补的国家最起码应考虑根据有关女子的其他生育力项目和年龄确定其缺失的信息。在握有关于女性年龄、平均生育数和存活子女数的有效信息的情况下可以更新插补矩阵，并可在该项目缺失的情况下用它来进行插补。当“平均生育数”缺失的时候，插补矩阵有女性年龄和存活子女数。当存活子女数缺失的时候，插补矩阵有女性年龄和平均生育数。

400. 另外，在设计插补矩阵的时候一定要记住：平均生育数和存活子女数必须符合母亲与长子（女）之间的年龄差（如果有这方面的信息的话），并且与特定年龄的母亲生育子女的总数相一致。

401. 譬如讲，插补的平均生育数与母亲年龄之间的差异应该不得小于12。那么，一个使用女性5岁年龄分组的插补矩阵就几乎肯定可以在某些情况下插补可估算的信息。

402. 图28中附随的插补矩阵显示某女子的年龄（横跨上栏）和平均生育数（侧面向下）。各输入项是存活子女的插补值。有时候回答是恰当的，但有时不恰当。如果程序遇到一个19岁的女子生育了5个子女，那么，5个存活子女这个值大概就应通过年龄差的标准了（根据存活子女数和报告的年龄，年龄差为15岁）。然而，对于一个15岁的女子来讲，无论其生育子女的数值5（年龄差为10岁）还是其存活子女的数值4（年龄差为11岁）都是不可接受的。

图28

在年龄和平均生育数值均有效的情况下用于确定存活子女数的初始值

平均生育数	年龄												
	15	16	17	18	19	20	21	22	23	24	25-29	30-34	35+
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0
2		2	2	2	2	2	2	2	2	2	1	1	1
3			3	3	3	3	3	3	3	3	2	2	2
4					4	4	4	4	4	4	3	3	3
5					5	5	5	5	5	5	4	4	4

403. 在年轻女子适用单一年龄的情况下，插补矩阵比较好。那么，只有对特定年龄的年龄差的有效回答才能被纳入插补矩阵，而只有有效的回答才能从插补矩阵中提取出来。

(f) 在平均生育数、存活子女数和已亡子女数都报告的情况下如何编辑

404. “平均生育数”是“存活子女数”和“已亡子女数”之和。任何不一致的问题都可以通过下述方式得到解决。

(一) 如果所有三项都已报告

405. 如果三项信息全有，程序需要确定：

(a) 这三个项目（即平均生育数、存活子女数和已亡子女数）是否内部一致；

(b) 这三项是否各与该女子年龄相一致；

(c) 平均生育数是否与最后一年生育数（或最近一胎）相一致（如果收集了这种信息的话）。

406. 如果所有这些都一致，编辑就算完成了。不过，如果有任何不一致的地方，编辑必须予以解决。这三个项目也许会内部不一致，譬如讲，一个女子可能生了5个子女，但是只有两个子女存活，并有两个子女已亡。编辑团队应确定哪个变量比其他变量先发生。在许多情况下，女性可能记得她生育的所有子女，尽管它可能忘记死去的儿女排行老几。那么，编辑团队就可以选择接受平均生育数和存活子女数，然后通过两数相减获得一个新的与已亡子女相符的数值。

(二) 如果报告了两项

407. 既然平均生育数（CEB）是存活子女数（CS）加上已亡子女数（CD）之和，那么，只要知道了这三项信息中的任何两项，计算机程序就可以确定第三个变量：

(a) 如果已知CEB和CS，CD = CEB - CS；

(b) 如果已知CS和CD，CEB = CS + CD；

(c) 如果已知CEB和CD，CS = CEB - CD。

通常都是首先运行这些检验。一俟程序确定所有三项信息全都有效且一致，编辑即告结束。

(三) 如果只报告了一项

408. 在国家不使用动态插补的情况下，如果三个项目当中只报告了一项，另外两项应填报“未知”。在国家使用动态插补的情况下，编辑团队需要确定用什么方法至少再获得一个项目，然后通过减法或追加来获得第三个项目。可以使用一个两维度的矩阵，根据第一个项目和女子的单一年龄来获得第二个生育力值。举例来讲，如果已知平均生育数，即可从上述插补矩阵来获得存活子女数，并通过减法获得已亡子女数。同样，如果已知存活子女数，即可从该女子单一年龄和存活子女数的插补矩阵获得平均生育数，并通过减法获得已亡子女数。

(四) 如果一项都没报告

409. 如果三个项目当中一项都没报告，编辑团队必须决定下一步怎么办。在不使用动态插补的国家，所有项目都应成为“未知”项，并且不得用于死亡率或生育率之类的间接方法。在使用动态插补的国家，专家必须决定要不要对所有项目使用动态插补。

410. 如果专家决定使用动态插补，即可根据女子的单一年龄和至少一个其他特征获得其平均生育数。还有可能仅使用母亲单一年龄的单一维度数组。其他特征或许是教育程度或宗教信仰。

411. 一俟确定了第一个项目，即可按照上面讲的只报告了一个项目的编辑步骤获得第二个生育力项目。然后可以从前两个项目获得第三个项目。这三个项目应该是兼容的，因为只有在所有项目都兼容的情况下插补矩阵方可更新。所获得的生育力值也应与该地区的其他妇女的数据相一致，因为从那些妇女获得的信息被用来更新插补矩阵。

(g) 在平均生育数、身边子女数、非身边子女数和已亡子女数都有报告的情况下如何编辑

(一) 如果所有信息都已报告

412. 如果所有四项信息都存在，编辑程序需要确定：

(a) 这四个项目是否内部一致，以至于平均生育数等于身边子女、在外子女和已亡子女之和；

(b) 这四个项目是否各自与该女子年龄相一致；

(c) 平均生育数是否与最近一年生的（或最近一胎）子女“相一致”（如果有这方面信息的话）。

413. 如果所有这些项目都一致，编辑就算完成了。不过，如果有任何不一致的地方，编辑程序就需要予以解决。跟上面讲的报告了三个项目的情况一样，四个项目也不一定全都内部一致。在这种情况下，编辑团队也要确定先解决哪个项目的问题。在许多情况下女性普查对象可能会记得生育的所有子女，尽管她或许忘记了哪几个子女去了外地或说不出确切的已亡子女数。那么，编辑团队就可以选择接受平均生育数和存活子女数（即非身边子女数与身边子女数之和），并且用减法获得新的相一致的其他变量值。编辑团队可能需要制定各种变量组合的算法。

(二) 如果四个项目报告了三项

414. 平均生育数（CEB）是身边子女数（CLH）、非身边子女数（CLA）和已亡子女数（CD）之和。如果四项信息中报告了任何三项，计算机程序都可以确定第四个变量：

如果已知CEB、CLH和CLA，CD = CEB - CLH - CLA。

如果已知CLH、CLA和CD，CEB = CLH + CLA + CD。

如果已知CEB、CLH和CD，CLA = CEB - CLH - CD。

如果已知CEB、CLA和CD，CLH = CEB - CLA - CD。

(三) 如果四个项目只报告了两项

415. 如果只有两个项目是已知的，那么编辑团队就必须决定下一步怎么办。譬如讲，在许多国家妇女不报告已亡子女数。最容易忽略的另一个项目就是不在本居住单元住的子女信息，这种信息也不能直接得到。因此，为了确保获得生育力项目的最佳数据，必须在设计问卷、实施查点和数据处理方面谨慎从事。

416. 通过计算居住单元内的子女总数可以获得住在本单元的子女数（CLH）。只要单元内的一个女性有适当的关系，就应对单元内的子女数作简单计数。如果不止一个女性有这种关系，编辑程序仍可适用，其假定条件就是在收集数据期间子女数据直接随母亲而定。如果所有这些其他方法都不管用，使用动态插补的国家可以根据母亲年龄和另一个已知变量来插补住在本单元的子女数。（见下面关于根据其他项目和母亲年龄插补个别生育力项目的一般规则。）有必要尽可能使用女子的单一年龄以及单一平均生育数、住在本单元的子女数、住在外地的子女数和已亡子女数。

417. 举例来讲，平均生育数和已亡子女数可能是有效输入项，但是住在本住户的子女数和住在外地的子女数也许是无效的。在这种情况下，可以通过总计与母亲有适当关系的子女数据来确定住在家里的子女数（假定母亲是户主）。于是，四个项目中可以获得三个项目值了，而第四个项目——即住在外地的子女数据——可以用减法确定：CLA = CEB - CLH - CD。

418. 可是，如果只有两个项目已知的话，那就多半需要对平均生育数和身边子女数进行重新编码了。妇女通常愿意报告平均生育数，并且通常可以通过观察或通过与普查对象在查点中合作获得关于身边子女数的信息，但是这些解决办法对在外子女或已亡子女不适用。于是，编辑就可以使用一个有女子年龄和平均生育数（CEB）的插补矩阵，或者再好一些，一个有女子年龄、平均生育数（CEB）和身边子女数（CLH）的插补矩阵。这些变项将从一个具有同样特征的类似妇女那里获非身边子女数（CLA）的信息。

419. 使用只有女子年龄和平均生育数（CEB）两维度矩阵而不包括身边子女数（CLH）这个第三维度的国家冒险获取与另外两个项目不相容的非身边子女数（CLA）。举例来讲，如果女子年龄为25岁而平均生育数CEB是5个，那就可能从插补矩阵获得非身边子女数为3的值。如果身边子女数是2，编辑就没问题。已亡子女数应为0，这样，生育力的各项目即为：CEB = 5；CLH = 2；CLA = 3；CD = 0。

420. 然而，实际身边子女数或许是4，只有女子年龄和平均生育数可用来确定非身边子女数。这样的话非身边子女数值3就会与其他项目不相容。平均生育数（5个）就会少于存活子女总数（4个在家，3个在外，总共7个）。因

此，应使用一个三维度矩阵：对于平均生育数CEB为5、身边子女数CLH为4而言，插补矩阵中的非身边子女数值或许为1（而已亡子女数值应该用减法确定为0）。或者，插补矩阵中的非身边子女数值应为0（而已亡子女数值应该用减法确定为1）。需要为其他对已知信息涉及类似的插补矩阵（如图29所示）。

图29
拟为各种成对已知信息设计的插补矩阵示例

如果这些为已知项目……		使用其中之一的动态插补（然后减去或加上）	
平均生育数	身边子女数	非身边子女数	已亡子女数
平均生育数	非身边子女数	身边子女数	已亡子女数
平均生育数	已亡子女数	身边子女数	非身边子女数
身边子女数	非身边子女数	平均生育数	已亡子女数
身边子女数	已亡子女数	平均生育数	非身边子女数
非身边子女数	已亡子女数	平均生育数	身边子女数

421. 在每一种情况下，四个项目当中各有两项已知信息。第三个项目通过动态插补获得，而第四个项目则通过加减法获得。编辑团队必须根据文化环境确定哪种办法最好。

（四） 如果只报告了一个项目

422. 如果四个项目当中只有一项是已知的，情况就更加棘手了。在这种有多项信息缺失的情况下，各国必须决定如何继续编辑下去。如果使用动态插补，如前所述，第一个插补矩阵应当使用一个譬如女子单一年龄之类的项目和一个已知的项目来创建一个二维插补矩阵，用以插补任何一个其他项目。一旦确定了两个项目，按定义另外两项依然是未知的。因此，继续使用第三个项目的动态插补应该不会产生与其他项目不相容的问题，因为它们是未知的。使用上面讨论的两个已知项目和两个未知项目的方案来获取第三个项目。然后用减法获得第四个项目。所有这四个项目均应相互兼容。

（五） 如果一个项目都没报告

423. 如果这四个项目当中一项都没有报告，编辑团队必须决定如何在没有任何已知项目的情况下继续进行编辑。如果国家不使用动态插补，所有项目都应成为“未知”项，并且不得用于间接估计死亡率或生育率的方法。在使用动态插补的国家，专家则必须决定要不要对所有项目使用插补法。

424. 如果专家决定使用动态插补，即可根据女子的单一年龄和至少另一特征来获得生育子女的数值。还有可能使用仅有母亲单一年龄的单一维度数组。其他可用的特征或许有教育程度或宗教信仰之类的项目，因为据悉在许多国家由于教育程度不同或宗教信仰不同会有不同的生育率。

425. 一旦确定了第一个项目，即可使用上面讲的只有一个已知项目的方法来获得第二个生育力项目值。然后就可以根据前两个项目获得第三个项目

值，而第四个项目值可以用减法获得。这四个项目只应该是相互兼容的，因为只有在所有项目相一致的情况下才能更新插补矩阵。获得的生育力数据也应与该地区的其他妇女相一致，因为来自这些妇女的信息被用来更新插补矩阵。

(h) 五种或以上项目的特殊情况

426. 随着国际移徙在一些较小国家变得更加重要，收集信息的内容又追加了非身边子女数这一项。当“在外子女”这一变项被进一步细分为“在外但未出国子女”和“出国在外子女”的时候，四个变项——即“在家”、“在外”、“已故”和“总数”——的程序必须扩大到顾及这一补充信息。另外，如前所述，让每个有完整生育力信息输入、所有项目都有效、内部一致且与年龄一致的妇女有个单一年龄数组线，这是个好主意；那么，在生育力信息不一致（包括与年龄不一致）的情况下，就可以使用这条完整而适当的数组线进行插补了。

(i) 所有生育力项目共享单一供体来源的重要性

427. 因此，要是可能的话，非常有必要在一无所知的情况下插补所有项目。为了确保所有信息出自同一个女性来源，可能需要设计一个使用所有生育力信息的插补矩阵。在这种情况下，只有当编辑程序确认所有生育力项目都已知的时候才更新插补矩阵。正如上文第426段指出的那样，最好不要按项目插补，而应该在好几个项目值缺失的情况下使用另一位妇女的总体信息进行插补。

(j) 亲生子女与同住户子女和存活子女的关系

428. 如果国家使用亲生子女方法来帮助检查设计和实施生育力编辑程序，出自同住户子女的信息和母子矩阵的信息可以帮助评估编辑结果的可靠性。由于很少有国家使用这种方法来帮助编辑，所以至今它仍处于试验阶段；不过，现有成果看来是有希望的。

7. 生育力：最后存活子女的出生日期和普查前12个月内出生的子女

429. 最后出生子女的信息有助于提供普查或调查之前的现时生育率估计数。一种做法就是获得最后出生子女的出生日期（年、月、日）及其性别，并且确定该子女是否存活。第二种做法就是收集普查前12个月内的出生信息；对于普查员和普查对象来说采用第二种做法比较容易，因为只需回答“是”或“否”即可，而无需回答确切日期。

430. 在数据处理期间，普查日期以前12个月内出生并存活子女的估计数可以从最后出生子女的生日推导出来（然后作为记录保存）。对于限定年龄的现时生育率和其他生育力量度来说，这种方法提供的数据要比基于一个妇女在普查前12个月内平均生育数的信息更准确（联合国，2008年，第2.188-2.191段）。

431. 应当指出的是，关于最后出生活着的子女生日的信息并不产生关于过去12个月内子女总数的数据。即便在报告最后一个活产子女中没有错误，这个项目也只能确定在过去12个月内至少有一个活产子女的妇女人数，而不是出生婴儿数，因为有很小一部分妇女一年之内生育不止一个子女（联合国，2007年，第2.189段）。

432. 只需收集年龄在15岁至50岁之间据报一生中至少生过一个活产子女的妇女的信息。此外还应该对按性别报告平均生育数据的妇女收集有关各种婚姻状况类别的信息。如果收集了一个妇女样本的平均生育数据，那么还应收集同一样本的现时生育率信息（联合国，2007年，第2.190段）。

433. 应将下述校订纳入编辑程序。介于国家规定的最低年龄和国家规定的最高年龄之间的所有妇女均应输入最近一胎子女的出生日期。程序应检查一致性。譬如讲，对于不在选定年龄组范围之内的男子和妇女来说，其记录中不应出现相关的信息。另外，凡在年龄组范围内的产次大于零的妇女均应有一项最近一胎子女出生年月日的有效数据（或者说明是否在最近12个月内生育过——如果用到这个问题的话）。

434. 编辑团队需要确定是否“月-日”必须有效：使用动态插补法的编辑团队可以插补缺失的月份和日期；不用动态插补法的编辑团队会给月-日指定“未知”。如果主题专家（通常为人口统计学家）想要了解母亲在生孩子时的实际年龄作为生育力分析重新编码的话，那么最起码应该插补缺失的最近一胎出生月份。于是即可获得这个再编码。

435. 同样，有些人口统计学家想要分析最近一次生育以来过了几个月。通过编辑最近一胎的年份和月份，可以为获得自最近一次生育以来跨越的完整月数提供必要信息。如果最近一胎的生日也收集到了，就可以用它来确定自上一胎以来经过月数的再编码(关于获得自最近一次生育以来经过月数的方法，见附件一)。

436. 如果关于最后子女的出生年份的信息缺失或无效，不使用动态插补的国家可以指定“未报”或“未知”。使用动态插补的国家可以利用年龄和平均生育数之类的其他变量来获得最后生育子女的日期。

437. 鉴于最近一胎的生日对于计量国家、地区和地方最近生育率的重要用途，应当考虑追加检查。有益的编辑需要在住户范围内检查一个或多个不满周岁的子女，并且利用母亲与这个或这些子女的关系（或母亲对该子女的个人号——如果收集的话）来确定该子女是否被报告为母亲的最近一胎孩子。检查应采取两种方式：不满周岁的子女应参照母亲来检查；最近一胎应参照住户列表来检查。

438. 同时也收集在普查或调查前一年死亡人口信息的国家可能决定，在最近一胎项目填报为“已故”或“已亡”的情况下，包括对照最近一胎检查在普查前一年不满周岁的儿童死亡数据。如果因为母亲去世或迁移不在该住户，或者由于某种原因子女未报，这种检查就做不了。但尽管如此，用这种方式还是能够检查一定比例的婴儿死亡率的。

8. 生育力：生第一胎时的年龄

439. 母亲生第一胎子女时的年龄可用于根据头胎间接估计生育率，并且提供关于开始生小孩的信息。如果普查中包含这个题目，就应收集每一位至少生过一个孩子的妇女的相关信息（联合国，2008年，第2.193段）。

440. 生第一胎时的年龄要么可以通过一个明确的项目“生第一胎时的年龄”直接获得，要么通过母亲现在的年龄与长子（女）之间的年龄差获得（如果后者年龄已知的话）。这里，国家规定的生小孩的最低年龄不是生物学上的最低年龄。举例来讲，如果一个国家规定可接受的生第一胎的最低年龄是13岁，那么普查对象可能报告或普查员可能记录一个人生（第一胎）孩子的年龄是11岁或12岁。于是，编辑团队就必须决定：是改变最低年龄的规定呢，还是删除这一生育年龄，抑或要么改变母亲年龄，要么改变她在生第一胎时的年龄（要么使用子女的年龄，要么使用母亲年龄，根据确定年龄差所使用的变量而定）。同样，编辑团队还必须确定生第一胎时的“最高年龄”。虽然有的妇女到50多岁还能生孩子，但这种事情并不常见，而为了纠正错误，编辑程序必须判断这种离群值是否真实。

441. 有必要牢记，生第一胎的最低年龄和最高年龄（以及母亲与其住在家里的最大子女间的年龄差）必须符合国家习俗和传统。主题专家必须决定什么情况下一个填报的生第一胎的年龄属于误差，而不是法定值。在规则已定的情况下，专家必须决定如何纠正此类问题。如果不使用动态插补，程序应将此项目指定为“未知”。在使用动态插补的情况下，可以根据差不多年龄的其他妇女的数据和类似的平均生育数来确定生第一胎时的年龄。确定插补矩阵的专家可能还想要兼顾诸如城/乡居民（如果两地区生育率有别的话）、妇女劳动力参与率（尽管现时的劳动力状况未必与其生投胎时的状况一样）以及教育程度等因素。

9. 死亡率

442. 在缺乏令人满意的基于民事登记的连续死亡统计数据的国家，过去12个月死亡数被用来估计按性别和年龄分列的死亡率和死亡方式。为了确保从该项目产生可靠的估计数，必须尽可能完整而准确地按性别和年龄报告过去12个月的死亡数据。以往数十年来，在普查问卷中广泛包含死亡率的问题，因而导致在使用间接估计方法估计成年人死亡率方面有所改进（联合国，2007年，第2.194段）。

443. 理想上，应当谋求对每个住户进行关于普查之前12个月内死亡总人数的死亡率调查。如果不能获得关于12个月内死亡人数的信息，起码也要收集关于不满周岁的子女死亡数据。同时还要收集所报告的每个逝者的姓名、年龄、性别和死亡日期（年、月、日）。对每位普查对象，要注意清楚地讲明参考期，以免由于错误解释而导致误报。譬如讲，可以采用各个国家的某个节日或历史纪念日来作为确切的起算日（联合国，2007年，第2.195段）。

(a) 逝者的年龄和性别

444. 《人口和住房普查的原则和建议》第二次修订本（联合国，2007年）建议收集普查前一年内死亡者的姓名、年龄和性别以及死亡年月日。不使用动态插补的国家可以将这种变量的无效项目指定为“未知”项。使用动态插补的国家或可使用年龄（按年龄组）、性别和死亡年份作为其他变量插补矩阵的维度。实际插补矩阵大概需要依具体国家而定，且编辑团队要共同协作开发适当的插补矩阵。国家或各级行政地理层次的人口结构可用于设计最适当的编辑程序。

(b) 死亡原因

445. 现在有些国家收集关于普查前12个月内死亡者的死亡原因的信息。由于这个问题较为敏感，还因为难以在现场得到回答，所以有的国家这样提问：是否因为事故或暴力而死亡？以期从特定年龄组获得有关艾滋病毒/艾滋病的间接信息。此项编辑通常需要设定，如果收集不到信息或信息无效，项目值一般可指定为“未知”。如果国家选择使用插补法，那么使用性别加0岁、1-4岁然后是5岁年龄组的热卡比较合适。

(c) 产妇死亡率

446. 在本轮普查中，越来越多的国家还问到死亡者是不是女性，以及死时是否怀孕。这个项目帮助确定国家和地区级的产妇死亡率。对于无效或空白输入，该项目编辑可能需要填报“未知”。可是如果国家选择进行插补的话，则显然仅对女性应用热卡，且只适用于有可能怀孕的年龄——大约在12岁至54岁之间，而年龄大概要按单一年份算，而不按5岁年龄组算。

(d) 婴儿死亡率

447. 最后，《人口和住房普查的原则和建议》第二次修订本（第2.191段）建议收集“过去12个月内”出生后死亡婴儿的信息。通常，这个问题只结合普查前12个月出生婴儿项目提问。如果使用另一个当前生育力项目——即最近一胎的出生日期，那么本项目大概就不该使用。

448. 关于普查前一年内新生儿死亡人数的数据可以帮助那些有良好死亡登记的国家检查其全国和地区级的婴儿死亡率。那些缺乏良好死亡登记的国家可以利用这种信息获得婴儿死亡率估计数。在普查前一年内死亡的最近一胎的编辑检查与当年不满周岁的儿童死亡人数相对照，也能提供有关婴儿死亡率的有用信息。

449. 这个项目的编辑需要费一些脑子，并且依特定国情而异。理想上，关于生育及存活子女数的信息可用于检查报告的信息；如果住户中只有一个成年女性，检查起来就相对比较容易。如果该单元有好几个妇女，那就得当心，以确保恰当的子女跟恰当的妇女相联系。

10. 母亲遗孤或父亲遗孤与母亲的行号

450. 关于收集遗孤信息，可以直接提问两个问题：(a) 在普查的时候住户中被查点者的亲生母亲是否还活着；(b) 在普查的时候住户中被查点者的亲生父亲是否还活着。调查应确保获得有关亲生父母亲的信息。因此要留心将养父母排除在外。鉴于通常有不止一个存活子女回答孤儿状况的问题，所以需要设计一些问题来克服在同胞兄弟姊妹报告父母亲方面的重复回答的问题。为此目的，应补充提出两个问题，即：(c) 普查对象是不是其母亲的存活长子（女）；(d) 普查对象是不是其父亲的存活长子（女）。制表的时候只应参考长子（女）的数据（联合国，2007年，第2.199段）。

451. 关于“母亲是否活着”和“母亲的行号”两项的编辑是相互关联的，因此应一起进行。对于不用“是”来回答“母亲是否活着”项目的个人来说，应检查母亲的行号以确保有效的输入；如果呈现有效输入，应给“母亲是否活着”项目指定代码“是”。对于不用“是”来报告母亲是否活着的个人来说，应检查母亲的行号，以核实它是不是“00”或者它是否等于一位年龄大于或等于12岁的妇女的行号。如果其中任何一种情况属实，编辑程序就认定此人有一位母亲，并且给母亲的生死状况项目指定“是”。如果母亲的行号输入无效且“母亲是否活着”项被赋予代码“否”或“不知道”，则母亲行号的输入值应予删除。在其他所有情况下，均应对“母亲是否活着”项目指定代码“不知道”，而关于“行号”的任何输入值均应予以删除。

452. 对于母亲是否活着报告说“否”或“不知道”的个人，国家或许选择不编辑其母亲的行号。在所有其他情况下，应检查行号的一致性，或者使用此人的关系以及被报告为母亲的某个人的行号、性别、关系和年龄予以指定行号。在存在不一致问题或母亲状况不能确定的情况下，或可指定“住在别处”代码。应该指出的是，在结构编辑中，如果户主不是第一人且尔后被移至第一人的位置，那么母亲对一个或多个人的行号可能需要调整。

B. 移民特征

453. 由于人口自然增长（生育率和死亡率）及净移民的结果，一个国家的人口特征随着时间推移在不断变化。移民可以是长期移居（从一出生）或者短期移居，按以往的居住地和居留期间或在一个以往的明确时间点上计数。鉴于这些项目往往是相互关联的，所以有些国家可能适合进行类似于基本人口统计变量的那种综合编辑。如果采取自顶向下的编辑方法，编辑顺序就变得十分重要，因为有些项目必须先于其他项目进行编辑。

454. 与其他项目相比，移民项目往往需要较详细的代码组，因为较小的地域单位对于规划和政策用途是必要的。小区域的详尽信息可能是进行新学校或医疗设施所需要的。此外，国内外各地可能需要各种不同的编码方案和不同的编辑程序。

455. 传统上，大多数国家都没有经历过太多的国际移民，所以重点放在国内移居方面，这依然是首要关切之所在。然而，在一个日益全球化的世界上，也越来越多地注重国际移民问题了。

456. 对于国内移居（内部移民），应检查关于国内出生地和在本区生活年头，看这些数据是否一致，因为这两个项目之间显然是有关联的。另外，住户的不同成员回答问题之间也存在合理的关系。举例来讲，如果没有回答某个子女在本区生活的年头，可以根据母亲的回答进行插补，而编辑程序会通过检查确保这项插补值不超过该子女的年龄。

457. 对国际移民来说，令人关切的是出生国家和入境时间。

1. 出生地

458. 出生地首先是有关个人的出生国。应该指出的是，出生国未必与公民身份有关，这是另外一个问题（见联合国，2007年，第2.92-2.102段和本《手册》关于出生国一节）。对于在正在进行普查的国家出生的人（当地人）来讲，出生地的概念还包括在有关个人出生时其母亲居住的国家明确规定的地理单位类型。可是在一些国家，当地人的出生地被定义为实际出生地所在的地理单位。每个国家都应说明它在普查中使用哪种定义（联合国，2007年，第2.57段）。

(a) 出生国和在本区居住的年头两个输入项之间的关系

459. 可以检查出生地和居住期间的输入值的一致性，因为这两个项目之间关系密切。住户的不同成员之间也相互关联，并且可以根据其他家庭成员设想有关个人是否已经移居。

(b) 给无效出生地输入项指定“未知”代码

460. 如果一个国家选择不使用动态插补，那么关于出生地的任何无效回答都可以变成“未知”项。通常，除非编码出差错，否则国家不应编辑家庭成员之间的或地域之间的不一致。

(c) 使用静态插补法插补出生地

461. 只有在出生国的输入超范围的情况下才更改这个项目。如果在本区居住年头的代码是“一直”，那么这个国家的代码应该予以指定。如果输入值不是“一直”，那么前一个人的信息可以用上。举例来讲，如果前一个人是母亲，那么母亲在本区居住的年头即可同这个人的年龄作比较。如果母亲的输入值大于或等于这个人的年龄，即应指定“这个国家”代码；否则，就指定“母亲的出生国”。如果不能根据母亲的输入值指定出生国，可以用同样的方式使用其他相关个人的输入值。如果在这些检测之后还不能指定输入值，可将出生国指定为“未知”项。

462. 由于现在各国将其数据样本供公众使用，所以有必要在编辑过程中为空白输入项提供特定代码，这种空白是执行跳转模式的普查员跳转留下的。就是说，调查问卷往往告诉普查员如果有关个人一直在那个地方居住的话可将出生地的问题跳过。在编辑过程中，为了帮助用户应指定这种特定地方的代码，这样，他们日后在制作交叉列表的时候就无需查看两个地方了。

(d) 使用动态插补法插补出生地

463. 跟前面一样，只有在超范围的情况下才更改出生国。如果在本区居住年头输入值是“一直”，那就应将“这个国家”代码指定给出生国项目。如果输入值不是“一直”，则应从该住户的其他成员搜寻出生国的线索。

(e) 指定一个有母亲在的人的出生地

464. 如果出生国项目是空白或无效，而且在此地居住期间也不是“一直”，那就可以搜索此人母亲的信息。如果在该住户发现有母亲同住，就审查母亲的居住期间。如果她在本区居住的年头是“一直”，即可将此人的出生国指定为“这个国家”。如果此人的母亲不是一直在本区居住，但是此人的年龄小于或等于其母亲在本区居住的年头，那么程序也可以将出生国指定为“这个国家”。如果此人的年龄大于母亲在本区居住的年头，而且母亲的出生国有效，那就将此人的出生国指定为与其母亲出生地一样的国家。

(f) 指定户主子女的出生地

465. 如果此人的母亲不在该住户，但此人是户主的儿子或女儿，那就可以根据户主的记录进行几项检查。如果户主在本区居住年头是“一直”，程序应指定“这个国家”为此人记录上的出生国。如果户主在本区居住年头不是“一直”，但此人的年龄小于或等于户主在本区居住的年头，那么程序也应该指定“这个国家”为出生国。不过，如果此人的年龄大于户主在本区居住的年头，程序则应指定户主的出生国——如果其出生国编码有效的话。

(g) 指定子女的出生地，但不是户主的子女

466. 根据有关个人是高于还是低于国家编辑团队规定的年龄（X岁），可以对其进行截然不同的插补。如果此人年龄低于X，应根据相应性别的年龄低于X岁的之前第一人的记录来插补其出生国。

(h) 指定有丈夫的成年妇女的出生地

467. 如果此人年龄为X岁或以上并且是女性，编辑程序应查看她是否在本住户有丈夫。如果该妇女有一位丈夫，而且他有一个有效的出生国代码，程序应将其丈夫的出生国指定给她的记录。如果其丈夫没有有效的出生国代码，应查看他在本区居住年头的输入值。如果丈夫在本区居住的年头编码为“一直”，这位妇女的出生国应被指定为“这个国家”。如果丈夫在本区居住的年头代码不是“一直”，那么这位妇女的出生国应按年龄和性别予以插补。

(i) 指定没有丈夫的成年妇女的出生地

468. 虽然一位高于编辑团队确定的最低年龄的妇女在本住户没有丈夫，但是她可能是住户内一个子女的母亲。在这种情况下，程序应搜索其长子（女）的记录。如果找不到子女记录，程序可按年龄和性别插补出生国。如果其子女有一个有效的出生国代码而报告的母亲在本区居住年头多于子女的年

龄，那么程序应按年龄和性别插补出生国。但是如果母亲在本区居住的年头少于或等于子女的年龄，程序则应指定其子女的出生国。

(j) 指定男子的出生地

469. 为了获得一个男子的出生地，编辑程序可以尝试找到他的妻子的记录，或者，如果她是户主，程序应尝试找到其子女的记录。首先，程序试图找到该男子的妻子的记录。如果找到了她的记录，而丈夫在本区居住的年头少于或等于妻子的居住年头，那么就可以将妻子的出生国指定给该男子的记录。如果该男子在本区居住的年头多于妻子的年头，那就应该使用一个插补矩阵按年龄和性别插补其出生国。如果该男子是一家之长，在户内有一儿子或女儿，并且在本区居住的年头等于或少于其子女的年龄，那么，程序就应为该男子指定同子女一样的出生国。如果她在本区居住的年头多于其子女的年龄，程序则应按年龄为其插补出生国。

2. 公民身份

470. 要收集公民身份的信息，以便把人口分为三类：(a) 由出生地决定的公民；(b) 由入籍决定的公民，而不论是通过申报、选择权、婚姻途径还是通过其他途径入籍的；以及(c) 外国人。此外还应收集有关外国人的公民权国家的信息。有必要照直记录公民权国家而无须使用一个形容词来指明公民身份，因为有些形容词跟指定族群的形容词是一样的。

471. 关于公民权国家的信息编码要足够详细，以便鉴别国内的外国人口中有代表性的所有公民权国家的个人。为了编码的目的，建议各国使用《国家或地区统计用途标准代码组》（联合国，1999年）中提出的数字编码方法。使用标准代码组按公民权国家对外国人口进行分类，将扩大这种数据的用途并促进各国间关于外国人口的国际信息交流。如果一个国家决定联合各公民权国家组成广泛的群体，则建议采用上述出版物中鉴定的区域和次级区域分类方法（联合国，2008年，第2.97段）。

(a) 公民身份的编辑

472. 公民身份取决于每个国家的相关定义。在大多数国家（但不是所有国家），在一个国家出生的个人便自动成为由出生地决定的该国公民。因此，编辑应查看出生地与公民身份的关系，可能还需要给在该国出生的人指定“由出生地决定的公民”代码。

(b) 民族/种族与公民身份的关系

473. 有些国家还收集“民族”或“种族”数据，这方面的补充信息可用于确定公民身份，尤其在所收集的回答无效的情况下。对许多国家来说，第一代和第二代移民在其民族血统与其公民身份之间应有几乎完全的一致性。对于有长期国际移民历史的国家来讲，这一特征也许没有太大价值，但仍然可以结合其他变项予以考虑。

(c) 入籍与公民身份的关系

474. 在有入籍情况发生的国家，普查项目也许涉及到或涉及不到对入籍的要求。譬如讲，如果需要有一定的居留期的话，可以使用一个“居留期间”项目来检验是否达到了入籍期要求。于是，如果一个人生在国外，但是对公民身份的回答无效或不一致的话，编辑团队即可选择将其公民身份指定为“入籍”。对于没有在居留期间方面达到入籍要求的其他人，则使用冷卡插补法将其指定为“外国人”。

(d) 居留期间与公民身份的关系

475. 也许在问卷上没有“居留期间”这个项目，或者在确定公民身份方面要求比较模糊，或者也许编辑团队选择不用它。那末，如果公民身份值无效或与出生地不一致的话，即可在不使用动态插补的情况下指定“未知”。选择对无效值使用动态插补的国家至少应使用两个特征（其中之一大概是出生地），以便从本地区的类似个人获得“已知”的信息。

3. 居留期间

476. “居留期间”是指截至普查之日用整年表示的一个时段，在此期间每个人生活在：(a) 普查时作为其常住居所的地域；和(b) 这个地域所在的主要或较小普查区划（联合国，2008年，第2.64段）。

(a) 居留期间的编辑

477. 跟出生国家一样，居留期间在汇编人口流动性统计数据中也十分重要。有些情况下，某个分组人口的流动性可能要比整个人口大得多。这个项目的编辑考虑到个人出生地和住户其他成员的回答。“居留期间”要结合“前一居住地”“规定的以往特定日期的居住地”进行编辑。

(b) 事实上/法律上的居留与居留期间

478. 普查是*事实*上的还是*法律*上的，这对编辑可能会有影响。因为法律上的普查是在常住居所收集信息；而在事实上的普查中，个人是在普查那一晚上所居住的地域接受查点的，这两种普查中对“居留期间”的回答不一定引出同样的信息。另外，代码组和编辑程序必须考虑到有关个人要么是“一直”住在此地，要么是“从未离开过”。对这些人，编辑程序应跳过一致性和其他校订。

(c) 年龄与居留期间的关系

479. 第一部分编辑应检查年龄与出生地之间的一致性，并查看在当地或普查区中居住年头的输入是否有效。个人在一个地域或普查区居住的年头不能大于这个人的年龄。另外，一个在国外出生的人不可能一直住在这个地域或普查区。如果在该地域或普查区居住的年头多于年龄而出生国是这个国家，那么编辑程序应指定在该地域或普查区居住的年头为“一直”。如果在该地域或普

查区居住的年头大于年龄但出生国不是这个国家，则应将此人的年龄指定为在该地域或普查区居住的年头。在这种情况下假定虽然这个人生在国外，但是在其不满周岁的时候迁移到这个地域或普查区的。

(d) 出生地与居留期间的关系

480. 在输入超范围的情况下，应进行跟出生地一样的检验。应搜索相关的上一人（母亲、户主、丈夫、子女等）。应当根据查到的信息进行插补。不过，在指定数值以前必须确保其与所编辑个人记录中的年龄和出生地相一致。

(e) 一直住在此地者

481. 如果一个人住在该地域或普查区的年头是“一直”，但出生国不是“这个国家”的话，编辑团队或许想要将此人的年龄指定到在该地域或普查区居住期间。专家将假定，虽然这个人生在国外，但是在其不满周岁的时候迁移到这个地域或普查区的。下一部分编辑将检查在该地域或普查区居住年头的输入值。鉴于一个人在该地域或普查区居住的年头不可能多过这个人的年龄，所以将把年龄指定给在该地域或普查区居住的年头。

(f) 根据母亲居留期间插补个人居留期间

482. 如果类别没有一个有效的代码，编辑程序可以通过搜索此人在住户中的母亲来进行一种记录内部的检查。如果找到了，母亲的记录可以提供对指定缺失数据有用的信息。如果此人的母亲一直住在该地域或普查区，而且她的出生国是“这个国家”（正如所预期的那样），那么编辑程序即可将此人在该地域或普查区居住年头的类别代码指定为“一直”。如果母亲的出生国不是“这个国家”，那么即便她在该地域或普查区居住的年头是“一直”，这也表明母亲的类别出了差错。然后程序将忽略母亲的出生国，而用年龄指定在当地或普查区居住期间。如果母亲在该地域或普查区居住年头的输入值不是“一直”但却是一个有效的代码，而此人的年龄小于其母亲在该地域或普查区居住年头的话，编辑程序将回过头来检查母亲的出生国。如果母亲的出生国是“这个国家”，程序就用这个人的年龄来指定其在该地域或普查区居住的年头。不过，如果此人的年龄等于或大于母亲在该地域或普查区居住年头的话，程序就将指定“母亲在该地域或普查区居住年头”为此人在该地域或普查区居住的年头。

(g) 根据子女的居住期间插补个人的居住期间

483. 如果有关个人是子女（儿子或女儿），编辑程序应检查户主的记录，以便获取可能的信息来帮助指定关于居留期间的缺失值。如果户主出生在“这个国家”并且一直住在这个地域或普查区，那么程序就指定子女在该地域或普查区居住年头为“一直”。如果户主一直住在该地域或普查区、但不是出生在“这个国家”，那就把子女的年龄指定为在该地域或普查区居住的年

头。如果户主在该地域或普查区居住年头的输入值不是“一直”但却是一个有效代码，那么此项信息即可使用，条件是它须与正在编辑的子女记录年龄相一致。如果子女的年龄等于或大于户主在该地域或普查区居住的年头，程序就把户主在该地域或普查区居住的年头指定为儿子或女儿在该地域或普查区居住的年头。如果子女的年龄小于户主在该地域或普查区居住的年头，程序就指定一个取决于户主出生国家的值：如果户主出生在“这个国家”，该值就是“一直”；如果户主不是出生在“这个国家”，程序就将其儿子或女儿的年龄指定为在该地域或普查区居住的年头。

(h) 在没有其他可参考信息的情况下个人的居住期间

484. 如果以上所有努力都没有产生一个有效的值，程序可以为此给在该地域或普查区居住年头项目指定“未报”或“未知”。如果该值仍然无效，在使用动态插补的情况下应指定“未知”。选用动态插补法来处理无效值的国家，应至少使用两个特征来从本地区的类似个人获取“已知”的信息。

4. 先前居住地

485. 先前居住地是紧接在有关个人迁移到他（她）目前常住居所之前曾经居住过的主要或次级普查区，或外国（联合国，2007年，第2.67段）。

(a) 先前居住地的编辑

486. “先前居住地”项目应结合“居留期间”进行编辑。如果有关个人在此地（国家、地域或普查区，依普查项目而定）出生且从未迁移过，要么将此项目留作空白，要么明确指定“从未迁移”。不过，空白可能会给制表带来问题，所以编辑团队需要决定哪种处理方法最好。

(b) 改变了边界后的先前居住地

487. 国家边界有可能随着时间而改变，所以要留心，确保在编码方案上体现相应的改变。此外，代码组的确定方式应有利于逻辑分组。举例来讲，如前所述，在一个三位数代码组中，第一位代表居住地所在的洲，第二位代表该洲范围内的地区，第三位代表该地区内的特定国家。

(c) 如果个人自出生以来从未迁移过

488. 数据处理员就某些单个项目制表。因此专家应确保除了其他地方代码组之外还要使用“此地生人”的专用代码组。这样，编辑程序就能区分在某地出生的人和在一个地方出生但是迁移到同一地理区域的另一地方的人。

(d) 使用居住单元中其他人的信息

489. 如果“先前居住地”是无效的或者是不一致的，通常适用类似于对“居留期间”进行编辑的规则。如果母亲住在该居住单元的话，编辑程序可

以检查母亲的先前居住地。然后，程序可以查看户主的先前居住地，其信息可用于子女；在成年人不经常迁移的国家亦可用于成年人。

(e) 如果先前居住地没有其他人的适当信息可用

490. 如果上述各种方法都不能产生有效的值，那么程序可以给此人在先前居住地居住年头指定“未报”或“未知”代码。如果输入值依然无效，在不使用动态插补的情况下应指定“未知”代码。选择使用动态插补处理无效值的国家应至少使用两个特征来从本地区的类似个人获取“已知”的信息。

5. 以往特定日期的居住地

491. 以往特定日期的居住地是有关个人在普查以前的特定日期曾经居住过的主要或次级普查区，或外国。所选择的参考日期应当是对国家宗旨最有益的日期。在大多数情况下这个日期被认为是在普查前一年或五年。前一个参考日期提供单一年份的现时移民状况；后者可能比较适合收集用于分析国际移民状况的数据，但不太适合分析当前的国内移民状况。在选择参考日期的时候还应考虑到个人能否准确回忆其在普查之日前一年或五年时的常住居所。对于每五年搞一次普查的国家来说，大多数人都能很容易地把五年前的日期跟上一次普查联系起来。在其他情况下，一年前的回忆可能要比五年的回忆更准确。

492. 不过，有些国家可能需要使用不同于一年或五年的时间参考点，因为这两个间隔回忆起来都有困难。根据国情也许有必要确定一个可以跟大多数人都记着的某个重大事件联系起来的日子。另外，对国际移民来讲，与到达这个国家的年份有关的信息可能是有益的（联合国，2008年，第2.69段）。

493. 关于“以往特定日期居住地”的编辑类似于先前居住地的编辑。通常，各国会提问“居住地”和“先前居住地”，或者干脆问“以往特定时间居住地”。如果有关个人出生在调查地点（国家、地域或普查区，依普查项目而定）且从未迁移过，这个项目要么留作空白，要么可以注明从未迁移。如前所述，空白可能会给制表带来问题。然后适用前三段描述的关于先前居住地的相同程序。

6. 到达年份

494. 《人口和住房普查的原则和建议》（第二次修订本）将移民变项分为国内移徙和国际移徙。到达年份通常是指从国外某地进入这个国家的年份。因此，到达年份这个项目通常是结合其配对项目——即入境前的居住地——一起提问的（联合国，2008年，第2.103段）。

(a) 年龄与到达年份的关系

495. 编辑的第一部分应检查年龄与出生地之间的一致性以及到达该地域或普查区的年份输入值的有效性。一个人在特定地域或普查区居住的年头不能多于这个人的年龄。另外，一个在国外出生的人不可能一直住在该地域或普查

区。如果在该地域或普查区居住年头大于年龄而且出生国就是这个国家，那么编辑程序应将到达该地域或普查区的年份项目指定为“一直”。如果在该地域或普查区居住年头大于年龄但出生国不是这个国家，处理这种情况的一个办法把这个人的年龄指定为在该地域或普查区居住的年头。在这种情况下假定虽然此人生在国外，但在其不满周岁时迁入该地域或普查区。

496. 为了帮助用户使用公用样本，统计机构应为此项目提供“不到一年”和“一直”的代码组。“一直”的代码通常应指现居住地的实际地点，以帮助直接制表。“不到一年”代码可以使用户确信他们已经在其交叉分组列表中查看了人口中的每一个人。

(b) 出生地与到达年份的关系

497. 在输入项目超范围的情况下，应使用同出生地一样的检验。应搜索前面的相关者（母亲、户主、丈夫、子女等）。要根据查到的信息进行插补。不过在指定一个数值以前必须确保它与所编辑的个人记录中的年龄和出生地相一致。

(c) 一直在此地居住者

498. 如果对一个人自到达以来在该地域或普查区居住年头的回答是“一直住在此”但是出生国不是“这个国家”，编辑团队或许要使用此人的年龄来指定其到达该地域或普查区的年份。专家会假定虽然此人生在国外，但是在其不满周岁时迁入了该地域或普查区。编辑的下一部分将检查到达该地域或普查区的年份输入值的有效性。既然一个人在该地域或普查区居住的延续期间不能大于这个人的年龄，所以对这种情况，将把年龄指定为构成在该地域或普查区居住的年头。

(d) 根据母亲的到达年份确定个人到达年份

499. 如果该类别没有一个有效的代码组，编辑程序可以通过搜索这个人在本住户中的母亲记录进行内部检查。如果找到了她，她的记录就能提供对指定缺失值有用的信息。如果此人的母亲一直住在这个地域或普查区，而且她的出生国是“这个国家”（正如所预期的那样），程序即可给此人在该地域或普查区居住年头的类别指定“一直”。如果母亲的出生国不是“这个国家”，尽管她在该地域或普查区居住年头是“一直”，这也表明母亲的类别有差错。然后，程序将忽略母亲的出生国，而根据到达该地域或普查区的年份指定年龄。如果母亲到达该地域或普查区年份的输入值不是“一直”但却是一个有效的值，而此人的年龄小于其母亲在该地域或普查区居住的年头，那么编辑就回过头来检查母亲的出生国。如果母亲的出生国是“这个国家”，程序就把这个人的年龄指定为在该地域或普查区居住的年头。可是，如果一个人的年龄等于或大于其母亲在该地域或普查区居住的年头的话，程序就把“母亲到达该地域或普查区的年份”指定给此人到达该地域或普查区的年份。

(e) 根据户主的到达年份确定子女到达年份

500. 如果有关个人是子女（儿子或女儿），编辑程序应检查户主的记录，以便获取可能的信息来帮助指定关于到达年份的缺失值。如果户主出生在“这个国家”并且一直住在这个地域或普查区，那么程序就指定子女在该地域或普查区居住年头为“一直”。如果户主一直住在该地域或普查区、但不是出生在“这个国家”，那就把子女的年龄指定为在该地域或普查区居住的年头。如果户主在该地域或普查区居住年头的输入值不是“一直”但却是一个有效代码，那么此项信息即可使用，条件是它须与正在编辑的子女记录年龄相一致。如果子女的年龄等于或大于户主在该地域或普查区居住的年头，程序就把户主到达该地域或普查区的年份指定为儿子或女儿到达该地域或普查区居住的年份。如果子女的年龄小于户主在该地域或普查区居住的年头，程序就指定一个取决于户主出生国家的值：如果户主出生在“这个国家”，该值就是“一直”；如果户主不是出生在“这个国家”，程序就将其儿子或女儿的年龄指定为在该地域或普查区居住的年头。

(f) 在没有其他信息可用的情况下个人的到达年份

501. 如果以上所有努力都没有产生一个有效的值，程序可以为此给到达该地域或普查区的年份指定“未报”或“未知”。如果该值仍然无效，在使用动态插补的情况下应指定“未知”。选用动态插补法来处理无效值的国家，应使用适当数量的特征来从本地区的类似个人获取“已知”的信息。

7. 居留期间与到达年份之间的关系

502. 有必要指出的是，有些国家集中于国内移民并且包括关于居留期间的项目（往往连同先前居住地项目）。其他国家则集中在国际移民方面，其中包括到达年份（往往包括迁移前的居住地）。大多数国家要么有相当可观的国内迁移而鲜有国际移居，要么有相当可观的移居入境而鲜有国内迁移。不过有些国家国内国际移民都有，因此包括这两个项目。

503. 如果包括两个项目，统计局的工作人员必须非常仔细地开发编辑程序，以免造成内部不一致。这就是说，诸如年龄、居住期间和到达年份等变量必须通盘考虑，以确保居住期间和自到达以来的时间总和不比年龄大。因此，编程员在设计程序的时候需要同时考虑所有这三项变量。

504. 在使用动态插补的情况下，统计工作人员可能需要使用一种热卡，其中包括多维度数组，用以说明各种年龄和年头。另外，当居住期间和迁入年份是单一年头的时候，热卡也必须使用单一年头，否则，譬如讲，5岁年龄组的信息更新就可能在进行插补时造成抵触。

505. 同时，还需要在进行此项检查的过程中收集居住期间或迁入年份的分组数据或同时收集这两种数据的时候，以及在设计和实施热卡的时候，要特别小心。分组数据有可能造成重复问题。各国可能会认定在这种情况下提供一个“未知数”或许是最好办法。

8. 常住居所

506. 一般来讲，为了普查目的“常住居所”的定义是在普查时个人居住的地方，而他（她）已在此生活了一段时间或打算居住一段时间（联合国，2008年，第1.461-1.463段）。建议国家在考虑常住居所的时候规定一个12个月的门槛：

(a) 此人在既往12个月内大部分时间（亦即至少六个月零一天）连续居住的地方，其中不包括临时外出度假或因公出差的时间，或者打算在此至少再住六个月；

(b) 此人已在此至少居住了12个月，其中不包括临时外出度假或因公出差的时间，或者打算在此至少再住12个月。

507. 不过，进行事实上的普查的国家可以包括一个关于“常住居所”的补充项目，以便既获取事实上的信息又获取法律上的信息。对这个项目的编辑方法因特定国情而异。对于从未迁移过的人而言，常住居所等同于现住地；因此缺失的信息可以直接插补。

508. 可是，如果数据表明曾经迁移过，情况就变得较为复杂了。通常，如果该项目留下空白的话，国家假定常住居所和现住地是一样的，因而普查员和/或普查对象把信息省略掉了。

509. 不过，如果关于“居住期间”或“到达年份”的数据显示有证据或有些迹象说明改变过居住地的话，那么统计工作人员可能要试图开发一些方法，以便获得关于特定地理区域或全国的最恰当推测。虽然具体编辑程序取决于特定国情，但是作为最后一招大概不得不使用“未知”类别了。

510. 如果指示普查员在常住居所与普查地点相同的情况下可将此项目留作空白，而在编辑过程中，应在常住居所项目中填写普查地点代码。另一变量应表明编辑们已经做了这一更改。有一个完整的代码组会有助于公用样本用户对其数据进行完整制表。

C. 社会特征

511. 社会特征因国而异，但一般是指用来描述一国社会文化各个方面的项目。识字、就学和受教育程度，以及专业和学历等教育项目，可根据联合国教育、科学及文化组织（教科文组织）《国际教育标准分类》（ISCED）1997年修订本的类别进行分类（联合国，2008年，第2.202-2.230段）。

1. 读写能力（识字状况）

512. 有关识字状况的数据应向10岁及以上的所有人员收集。但在许多国家中，某些10-14岁的人也许即将入学识字，这一年龄组的识字率可能会误导。因此，在对识字状况进行国际比较时，应针对15岁及以上的所有人员，来编制识字状况数据表。如果国家面向年龄更小的人员收集数据，则有关识字状况的制表应至少区分15岁以下人员和15岁及以上人员(联合国，2008年，第2.202段)。

513. 每个国家都必须为识字状况制表确定最低年龄；同样，编辑小组必须为识字资料的编辑确定最低年龄。因为可能需要补充制表供内部使用。在编制调查表的过程中，编辑小组应确定据以收集数据的最低年龄，以及应在哪一级教育水平终止提问。因此，如果调查对象的学校教育已达到某种水平，则查点员不必询问有关识字状况的问题。但在编辑期间应填上该项目，以便为那些使用公用数据的研究人员和其他人员提供帮助。

514. 识字状况的编辑，首先要检查已学完的最高年级；如果根据设定，所学完的最高年级构成“识字”，则应分配“是”的代码。处于规定学校教育水平的人员应视为识字。如果发现识字状况的代码无效，则应分配一个值。登入项应为“未报”，或者利用一个基于特定变量（如，最高年级和性别）的插补矩阵来确定。“最高水平”取决于特定国家对“识字状况”的定义。

2. 就学

515. 从原则上讲，应向所有年龄的人员收集就学资料。就学与正式学龄人口尤其相关，正式学龄一般介于5-29岁之间，但可能会因国家而异，具体取决于国家的教育结构。如果扩大数据收集范围，以使就学数据纳入学前教育和/或生产和服务企业、社区组织和其他非教育机构为成人组织的其他系统化教育和培训计划，则可对年龄范围进行适当调整（联合国，2008年，第2.209段）。

(a) 就学项目的编辑

516. 每个国家的编辑小组都必须决定什么年龄适合作为收集就学数据的依据。大多数国家还将学校教育分为若干级，如果这些级别需按年龄编辑，那么专业人员还必须确定适合于各级教育的年龄组。有关其他所有人的登录都必须进行修改。如果编辑程序对这一类别产生的答案不一致，则必须更改年龄或就学情况。通常情况下，在进行这种编辑时，年龄是确定的，因此改变的是就学情况。在适合于特定国家的情况下，对于超过预定年龄的人员，应要求查点员略去就学方面的提问。如果学员到了中年继续接受中等或高等学校教育，则可能不适合确定就学的年龄上限。答复和答复组合一般会在普查前通过预备调查进行测试，因此，可在实际普查前，做出这些决定。

(b) 全日制或非全日制入学

517. 有些国家可能希望获得全日制或非全日制就学资料。该项目的列入，可能需要作为就学项目编辑的一部分，或者进行单独的编辑。

(c) 就学和经济活动之间的一致性

518. 应首先进行与其他主要项目（如，主要经济活动）之间的一致性编辑。如果就学是主要经济活动的登入项之一，而一个人将上学作为其主要活动来填报，那么应对就学分配“是”的代码，而主要经济活动应为“学生”。即：答复应一致。在所有其他情况下，任何有效的答复均应接受。

(d) 无效或不一致就学登入项的分配

519. 就学登入项如果超出了范围，而所完成最高年级的登入项有效，则应利用一个基于年龄、性别和最高年级的插补矩阵，来为其分配一个登入项。如果最高年级无有效代码，则应采用识字登入项为其分配。如果识字状况无有效代码，则只根据年龄和性别为其分配登入项。

520. 插补矩阵可能需要反映按性别和年龄开列的不同就学格局（有时按照1年跨度的年龄组或小跨度年龄组开列）。

3. 受教育程度（学完的最高年级或教育级别）

(a) 受教育程度的编辑

521. 受教育程度（最高年级或教育级别）的编辑应包括：(a) 有效登入项与年龄之间的一致性检查，(b) 在原登入项超出范围的情况下，插补登入项。如上所述，在不采用动态插补的国家中，值应为“未报”。而在采用动态插补的国家中，可能需要对年幼人员按性别和单年跨度年龄组分类，而对略大儿童则按性别和小跨度年龄组分类。如果一个国家的数据同时涵盖了最高年级和最高教育级别，则可能需要多重插补矩阵（联合国，2008年，第2.215段）。根据就业和所上最高年级，对“目前就读年级”进行二次编码时，应参照附件一的建议。

(b) 受教育程度的最低年龄

522. 每个国家的编辑小组都必须确定入学的最低年龄。如果确定了最低年龄，则所学完的最高级别一般不应超过一个人的年龄加上某个常数（相当于入学的最低年龄）。同样，有必要对儿童采用单年跨度年龄组，因为如果年龄组跨度太大，则在更新插补矩阵时，会出现误差。

(c) 年龄与受教育程度的关系

523. 编辑小组还必须确定数据集中将允许多少噪声。通常情况下，最好对年龄和受教育程度相矛盾的少数特例进行改变，而不是接受大量实际上不一致的答复。因此，如果原有登录超出了范围或者与年龄不一致，则可分配一个登入项。在不采用动态插补的国家中，应登录“未报”。而在采用动态插补的国家中，可能需要根据年龄（包括对学龄人口按单年跨度年龄组分类）、性别和就学来确定登入项。联合国教科文组织承认将识字状况与受教育程度分开，因此，“读写能力”很可能不应作为插补矩阵中的一个值。

4. 专业和学历

524. 人员按教育水平和专业开列的资料，对调查劳动力市场上特定专业领域合格人力的供求匹配情况具有重要意义，对于规划和监管各级、各类和各分支教育机构和培训计划的产能也同样具有重要意义（联合国，2008年，第2.223段）。

525. 未满15岁（或其他预定年龄）的人员，不应有专业和/或学历方面的信息。15岁及以上的人员，受教育程度与专业和/或学历之间应存在某种关系。在每种情况下，如果出现无效登入项，则不采用动态插补的国家，应登录“未知”。而采用动态插补的国家则可考虑利用年龄、性别、受教育程度或者职业，来分配专业和/或学历。

5. 宗教

526. 为普查之目的，可将宗教定义为：(a) 宗教或精神信仰偏好，而不管这种信仰是否由一个有组织的团体为代表，或者(b) 隶属于一个信奉特定宗教或精神信条的有组织团体。在普查中调查宗教的每个国家都应采用最适合其需求的定义，并应在普查出版物中阐明所采用的定义（联合国，2008年，第2.152段）。

(a) 宗教项目的编辑

527. 宗教是适合第二章所述标准编辑例子的变量之一。与其他项目不同的是，宗教的“未回答”情况很突出，需要加以考虑：有些人可能不愿填报其宗教情况。一个人的有效值（包括“未回答”），要么直接从另一住户成员（如果有可用的值）获取，要么从具有类似特征的另一户主获取。编辑小组应确定用于其他社会变量的合理编辑方案。对于户主，不管其是否为单元内填报的第一人，都应首先指定和进行编辑。如果一个人对宗教的填报无效或者未知，而这个人又是户主，则应采取以下步骤：

(b) 户主没填宗教，但单元内其他人填了宗教

528. 首先要确定住房单元内是否有其他任何人填了有效的宗教，然后按所填的第一个有效宗教来分配。

(c) 户主或单元内任何其他人都没填宗教

529. 如果住户内没有任何人填报宗教，则要么分配“未知”（前提是该国不采用动态插补），要么根据具有类似特征的最近户主进行插补，所依据的特征包括：年龄和性别，以及语言、出生地和该情况下的其他有关变量。

(d) 非户主，未填宗教

530. 如果此人不是户主，且未填报宗教，则编辑小组可按户主的宗教情况进行分配。

6. 语言

531. 普查可收集三类语言数据（联合国，2008年，第2.156段），即：

- 母语：指个人幼年时期在家中通常说的语言；

- 日常语言，指个人在其现住家中目前所说的语言，或者最常说的语言；
- 说一种或多种指定语言的能力。

(a) 语言项目的编辑

532. 在可能列在调查表上的三种不同语言尺度中，母语和通常语言这两种语言是相互关联的。当两者都出现在调查表上时，编辑小组应考虑将它们合在一起编辑。如果其中一个无效，则可利用另一个提供登入项。

(b) 语言项目的编辑：户主

533. 语言是适合第二章所述例子的另一变量。编辑小组应确定用于其他社会变量的合理编辑方案。如果一个人对语言（母语或日常语言）的填报无效或者未知，而这个人又是户主，则首先要确定住房单元内是否有其他任何人填了有效的语言，然后按所填的第一种有效语言来分配。如果没有任何人填报，则要么分配“未知”（前提是不采用动态插补），要么根据具有类似特征的最近户主进行语言插补，所依据的特征包括：年龄和性别，以及其他语言变量、出生地和这些情况下的其他有关变量。

(c) 语言项目的编辑：非户主

534. 如果此人不是户主，且所填语言无效，则可按户主的语言进行分配。

(d) 语言项目的编辑：利用原属种族或出生地

535. 语言和原属种族密切相关，有时还与出生地密切相关，就有些国家而言，可以将它们合在一起编辑。另外，编辑小组应考虑通过对代码的组织，来反映这些变量之间的关系。视代码的位数以及一国语言和族群的分布而定，可确定各种对应关系，帮助分配未知或不一致的答复。

(e) 语言项目的编辑：母语

536. 如果母语未知，但此人是菲律宾人并生于菲律宾，则可分配与之相当的语言——他家禄语、伊洛卡诺语或菲律宾其他语言。通常情况下，只有对户主按这种方式分配语言，而且该语言的代码也分配给该住户的其他成员，但各国编辑小组需要考虑特定情况，包括地理（如，城市或农村居住地），年龄或其他项目。

(f) 语言项目的编辑：说一种指定语言的能力

537. 说一种指定语言的能力是适合第二章所述例子的第三个变量。同样，应首先对户主进行编辑。如果户主语言的值无效或者未知，则首先要看住房单元内是否有任何其他人填报了有效的语言能力，然后按所填的第一种有效语言能力来分配。如果没有这种人，则要么分配“未知”（前提是该国不采用动态插补），要么根据具有类似特征的最近户主进行语言能力插补，例如，所依据的特

征包括：年龄和性别，以及出生地和这些情况下的其他有关变量。如果此人不是户主，且所填的指定语言能力无效，则可按户主的语言能力进行分配。

7. 种族和原住民

538. 是否要在国家普查中收集和发布各种族或民族团体的信息，取决于诸多考虑因素和本国情况，例如：包括国家对这类数据的需求、在普查中提出种族问题的适当性和敏感性。在移民、融合和政策对少数民族产生影响的情况下，确定一国人口的种族文化已日显重要（联合国，2008年，第2.160段）。

539. 《人口和住房普查的原则和建议》第二次修订本（第2.163段）建议在确定原住民（通常为种族项目的一部分）时，应尤其慎重。在编制代码清单时，必须认真确保原住性的识别具有唯一性，从而开发出适当的编辑规则和数据表，帮助进行原住民方面的规划和政策制定。例如，对于同一族群，如果对其作为游牧民族时与其定居在居民区时进行比较，则可能需要单独的代码。可开发特别的编辑规则（这可部分通过查找特定原住民的文件进行），确保它们在后续表格中得到适当和全面的确定。可为这些人群指定特别的插补办法，或者采用现有热卡内的补充类别。

(a) 种族项目的编辑

540. 其他一些变量，如果收集的话，可在填报无效或未知的情况下，帮助“确定”种族。在很多国家中，国内和国外出生地与种族之间都存在着某种关系。同样，在很多国家中，“母语”通常可作为种族的良好指标，因此，代码即使不同，也会类似。

(b) 种族项目的编辑：户主

541. 原属种族也适合第二章所述的例子。编辑小组应考虑采用为其他社会变量介绍的方案。应首先对户主进行编辑。如果一个人对原属种族的填报无效或者未知，而这个人又是户主，则首先要看住房单元内是否有任何其他人填报了有效的种族，然后按所填的第一个有效种族来分配。如果没有这种人，则要么分配“未知”（前提是该国不采用动态插补），要么根据具有类似特征的最近户主进行种族插补，所依据的特征包括：年龄和性别，以及语言、出生地和这些情况下的其他有关变量。

(c) 种族项目的编辑：非户主

542. 如果此人不是户主，且所填的原属种族无效，则可按户主的原属种族进行分配。

(d) 种族项目的编辑：利用语言和出生地

543. 原属种族和语言密切相关，有时还与出生地密切相关，就有些国家而言，可以将它们合在一起编辑。另外，编辑小组应考虑通过对代码的组织，

来反映这些变量之间的关系。视代码的位数以及一国族群和语言的分布而定，可确定各种对应关系，帮助分配未知或不一致答复。

544. 例如，如果原属种族未知，但此人说一种菲律宾语言并生于菲律宾，也许可以分配与之相当的原属种族。通常情况下，只有对户主按这种方式分配种族，而且其代码也分配给其他成员，但各国编辑小组需要考虑特定情况，包括地理（如，城市或农村居住地），年龄或其他项目。

8. 残疾

545. 残疾状况将人口分为有残疾和无残疾。残疾人指在履行特殊任务或参加与角色有关的活动方面，要比普通人口受到更大限制的人员。联合国建议在评估残疾状况时，从四个方面考虑：(1) 行走，(2) 视觉，(3) 听觉，(4) 认知（联合国，2008年，第2.351-2.352段和第2.367-2.371段）。

546. 用来识别残疾人的问题应列出主要残疾类别，供每人逐一核对是否有其中的残疾。残疾可根据《缺陷、残疾和障碍的国际分类》（ICIDH）的以下方面进行监测：(1) 能和残疾，包括身体功能和身体结构（残损）以及活动（受限）和参与（局限），(2) 背景因素，包括环境因素和个人因素（联合国，2008年，第2.354段）。

(a) 残疾普查问题

547. 建议在设计普查问题时，特别注意那些用以衡量残疾的问题。普查问题的措辞和构成会大大影响其识别残疾人的精确度。每个方面都必须与一个单独的问题相关。[8]所用语言应简单、明了、不含糊。应始终避免负面用语。残疾问题应向住户成员逐个提出，避免用笼统问题询问住户是否有残疾人。如果有必要，可找一个代答者为家庭的失能成员回答。最重要的是要逐个考虑每个家庭成员，而不是采用笼统的问题来调查。分级答复选项也可改善残疾的填报情况（联合国，2008年，第2.373段）。

[8] 在将各方面结合起来提问时，通常会让调查对象糊涂，例如，提出一个有关视觉或听觉的问题时，他们以为需要在这两方面都有困难时，才可回答“是”。此外，对于受到特定限制的人员，获得相关数据对于内部规划和跨国比较都有用。

(b) 残疾项目的编辑

548. 在某人不回答所提的残疾问题时，难以断定该项目空白是因为无残疾，还是因为调查对象出于各种理由不愿意回答。一国的编辑小组必须确定其是否想按照常规方式来编辑这个项目，即：在不采用动态插补时，分配未知；而在采用动态插补时，采用其他个人的答复。另外，专家可做出决定，答复只有在列明存在残疾的情况下，才应接受，而任何无效的答复应为“无残疾”。在后者情况下，不采用动态插补。

(c) 多种残疾

549. 收集多种残疾信息的国家需要修改编辑规则。编辑程序需要跟踪共有多少人可能是残疾的，以及这些残疾人的重复和分布情况。如前所述，大多

数国家都不适宜利用他人数据来按“残疾”分配，因此，在无效的情况下，可能需要分配“未知”，甚至“是否残疾未知”。

(d) 残疾原因的编辑

550. 一国的编辑小组必须确定其是否按照常规方式来编辑这个项目，即：在不采用动态插补时，分配未知；而在采用动态插补时，采用其他个人的答复。另外，专家可做出决定，答复只有在列明残疾原因的情况下，才应接受，不采用插补矩阵。

D. 经济特征

551. 经济活动状况信息原则上应涵盖整个人口，但在实践中，要根据各国规定的最低年龄，涵盖那些处于最低年龄或最低年龄以上的人员。不应将最低离校年龄自动作为收集活动状况信息的年龄下限。一个国家如果通常有很多儿童从事农业或其他类型的经济活动（例如，采矿、编织和小买卖），那么相对于儿童就业不普遍的国家而言，该国需要选择一个更低的年龄下限。

552. 经济特征制表应至少将15岁以下人员与15岁及以上人员区分开来；最低离校年龄高于15岁以及经济活动儿童年龄低于这一年龄的国家，应努力获得这些儿童的经济特征数据，以至少对15岁及以上人员的数据实现国际可比性。对于超过正常退休年龄的老年男性和女性，其对经济活动的参与也通常被忽视。在衡量经济活动人口时，需要注意这一点。在衡量经济活动人口时，通常不应采用最高年龄限制，因为有相当多超过退休年龄的老人可能会经常或者偶尔从事经济活动（联合国，2008年，第2.241段）。

553. 每个国家都必须确定参与经济活动的最低年龄。希望收集童工数据的国家可能需要选择一个较低的年龄下限，但必须记住，对于不在劳动力之列的儿童，如果在查点时，将他们错误地列入劳动力范畴，则可能会带来某些干扰。在确定最低年龄后，将针对X岁或以上的人员进行经济活动项目的编辑和制表；因此，只有在需要确定所有登入项都为空白时，才对X岁以下儿童的数据进行编辑。为方便所有制表，为X岁以下儿童登录的所有答复都应剔除。

1. 活动状况

554. 经济活动状况由若干经济变量组成，其中某些变量将在下文介绍。这些变量可满足数据收集需要，但可能需要再分类，才可进行数据处理和分析。

555. “现时活动状况”指，一个人在较短基准期内（如，一周，或一天）与经济活动之间的关系。如果一国人口的经济活动不会因为那些引起年内变化的季节因素或其他因素而大受影响，则最适合采用现时活动。一周或一天的基准期可以是最近某个固定的周、最近一个完整的日历周，或者查点前的最后7天（联合国，2008年，第2.248段）。

556. 根据联合国（2008年，第2.253段）的规定，就业者包括在规定年龄以上，并在较短基准期内（一周或一天）属于以下情况的所有人：(a) 为获取现金或实物报酬、利润或家庭利得而从事某种工作的人；或者(b) 已经从事某项工作，并且与该工作之间具有正式隶属关系，但暂时工作缺勤的人员；或者暂时不从事个体活动的人员（如，商业企业活动、农场或服务活动）。

557. 非现时活动人口，或者与之含义相当的非劳动力人口，由在较短基准期内（用以衡量现时活动）既未就业过，也未失业过的所有人组成，包括未满经济活动人口最低年龄的人员（联合国，2008年，第2.278-2.279段）。

(a) 与活动状况有关的类别

(一) 失业人口

558. 根据联合国（1998年，第2.271段）的规定，失业人口包括所有超过规定年龄，并在基准期内符合以下条件的人：

(a) 无工作：未从事过有酬工作或自营工作；

(b) 目前能工作：在基准期内能从事有酬工作或自营工作；

(c) 找工作：已在规定的最近期间内，采取具体措施以寻找有偿工作或自营工作。具体措施可包括：在公营或私营职业介绍所登记；向用人单位递交申请；查看工作场所、农场、工厂大门、市场或其他集散地；在报纸上刊登广告，或应聘广告职位；找亲戚朋友帮忙；寻找用地、建筑物、机器或设备以建立自己的企业；安排资金；申请批文和执照等。在对失业者进行分类时，将首次求职者与其他求职者区分开来会有用处。

559. 一般而言，列为失业者，一个人必须符合上述三项标准。但如果常规求职手段的相关性有限，劳动力市场主要是无组织的或者范围有限，同时劳动力吸收不足，或者劳动力主要从事自营工作，那么在采用失业的标准定义时，可放宽“找工作”这一标准，这种做法主要针对发展中国家，在这些国家中，该标准并不能完全反映失业程度。在极端情况下，可放宽到完全取消该标准，从而使“无工作”和“目前能工作”成为可以适用的两项基本标准（联合国，2008年，第2.272段）。

560. 失业项目——“下岗”、“找工作”（不论此人是否可以获取一份工作）和“最近一次工作的年份”——应合在一起编辑。另外，失业项目还需要与经济活动的答复一致，而且在大多数情况下，如果工时、行业、职业、工作者级别和工作场所等项目已经填报，则不应填报它们。如果主题专家确定，在答复空白或无效时，需要有“下岗”登入项，那么可采用基于人员年龄、性别或者受教育程度的插补矩阵。

(二) 找工作

561. “找工作”应与“下岗”和“不找工作的原因”一起编辑。主题人员应利用这些项目的登入项，确定编辑规则以推算其他项目。编辑应考虑本地和区域情况以及普查或调查变量。

(三) 非现时活动人口

562. “非现时活动”人口，或“非劳动力”人口由在较短基准期内（用以衡量现时活动）既未就业过，也未失业过的所有人组成（联合国，2008年，第2.278段）。根据其作为非“现时活动”人口的原因，可将他们归为以下任何一组：

(a) 在教育机构学习；

(b) 做家务；

(c) 领取养老金或资本收益；

(d) 其他原因。

“现时非活动”的编辑已纳入上述经济活动的编辑。

(四) 不找工作的原因

563. 该项目的编辑只针对那些被记录为“未找工作”的人员；所有其他人都应有一个空白登入项。另外，如果职业、行业和就业身份等出现了有效登入项，则应登入“有工作但没去工作”的代码。该代码代表那些在基准期内就业但没去工作的经济活动人口。在所有其他情况下，如果不采用动态插补，则可分配“未知”。就采用动态插补的国家而言，可利用年龄、性别和主要活动分配一个登入项。

(b) 经济活动状况的编辑

564. 经济活动一般包括以下类别：

经济活动人口：

(1) 就业者；

(2) 失业者。

非经济活动人口：

(1) 学生；

(2) 操持家务者；

(3) 退休金或资本收益领取者；

(4) 其他。

(一) 就业人员

565. 如果选择了第一类(“就业者”)，那么工时、职业、行业、经济活动状况和工作场所等变量应填报，如果没有填报，则应进行编辑和填补，要么

作为“未知”，要么采用冷卡值或热卡值。如果选择了第一类，则下岗、找工作和最近一次工作年份等变量应为空白，如果填报了，则应改为空白。

(二) 失业人员的经济活动

566. 如果选择了第二类(“失业者”)，那么下岗、找工作和最近一次工作年份等变量应填报，如果没有填报有效的登入项，则应进行编辑和填补，要么作为“未知”，要么采用冷卡值或热卡值。如果选择了第二至第六类，则工时、职业、行业、经济活动状况和工作场所等变量应为空白，如果填报了，则应改为空白。

(三) 学生和退休人员的经济活动

567. 如果选择了第三类“学生”，则主题人员需要确定，就学变量的登入项是否必须为“是，在学校”。如果选择了第五项“退休金领取者”，则主题人员需要确定有关人员是否必须达到某个年龄才退休。

(四) 经济活动无效，同时填报了就业变量

568. 经济活动的登入项无效，同时填报了工时、职业、行业和工作场所中的某些变量，则调查对象的经济活动应编码为1——“就业”。很可能需要一个插补矩阵来选择适当的答复。

(五) 经济活动无效，同时填报了失业变量

569. 如果填报了“下岗”、“找工作”和“最近一次工作年份”等变量中的任何一个，则经济活动的登入项应采用2-6中的一个值进行编码。如果此人正在上学，则该值很可能为3。如果此人为老人，则该值很可能为5。否则，主题专家可做出决定，利用插补矩阵来分配一个适当的答复。

(六) 经济活动无效，同时没有填报任何经济变量

570. 如果任何经济活动项目都没有答复，则主题专家很可能希望采用插补矩阵，来确定最适合的答复，然后插补其他经济项目。

2. 工时

571. 工时指在普查的经济活动所涉期间，在正常工作时间内和加班期间生产货物和服务实际花费的总时间。如果所涉期间较短，例如普查前一周，则建议按照小时计量工时。在这种情况下，为计量工作时间，可要求为一周的每天提供单独信息。如果所涉期较长，例如普查前12个月，则应以周为单位计算工时，或在可能的情况下，以日为单位，或者采用更长的时间间隔。有些活动尽管不会直接导致货物或服务的生产，但按照定义，仍然属于该项工作的任务和责任，如：准备、修理或维护工作场所或工具，在这类活动上所花费的时间也应列入工时。实践中，工时还包括从事这些活动过程中的非活动时间，如：等候时间、待命时间或其他短休时间。但不包括较长的午饭时，以及因为休假、假日、病假或产业纠纷而没有工作的时间(联合国，2008年，第2.323段)。

572. 该项目的编辑应只针对那些在经济活动答复中填报“就业，在工作”，或者“自营，在工作”的人员。就某些国家而言，还应对操持家务者列入工时。编辑小组预定的类别应予以接受。如果不采用动态插补，则空白、零或非数字代码应改为“未报”，如果填报的时数等于零，主题专家不妨将经济活动变量改为“不在工作”。

573. 如果采用动态插补，则插补矩阵的变量应至少包括年龄组和性别，但也可采用其他变量，如：受教育程度、职业或行业大类。

3. 职业

574. 职业是指，就业人员在一项工作中所从事的工作类型（或失业人员以前所从事的工作类型），而不管此人属于行业或就业身份中的哪一类别（联合国，2008年，第2.301段）。

575. 该项目的编辑应只针对那些在经济活动答复中填报“就业”的人员。如果不采用动态插补，则空白、零或无效答复应改为“未报”。

576. 在职业代码中，往往由不同位数表示主要职业和次要职业代码。职业项目的自填情况几乎不可避免，这给编码增加了负担。

577. 如果采用动态插补，则插补矩阵的变量应至少包括年龄组和性别，但也可采用其他变量，如：受教育程度或行业大类。

4. 行业

578. 根据《人口和住房普查的原则和建议》第二次修订本（第2.306段）的定义，“行业”指在为经济特征数据所确定的所涉期内，就业人员工作所在（或失业人员最近一次工作所在）基础单位的活动。至于选择待分类工作/活动的指导，见该修订本的第2.307段。

579. 该项目的编辑应只针对那些在经济活动答复中填报“就业”的人员。如果不采用动态插补，则空白、零或无效答复应改为“未报”。

580. 在行业代码中，往往由不同位数表示主要和次要行业代码。行业项目的自填情况几乎不可避免，这给编码增加了负担。

581. 如果采用动态插补，则插补矩阵的变量应至少包括年龄组和性别，但也可采用其他变量，如：受教育程度或行业大类。

5. 就业身份

582. 就业身份是指经济活动人口在就业中的身份，即在其工作中与其他人或组织机构之间的明示或默示雇用合同类型。用来界定分类各个组别的基本标准是经济风险类型，其中一个要素是一个人与其工作之间的关联程度，以及这个人在这个工作中对基层单位和其他工作者有或将有什么样的权力。应慎重确保经济活动人口在按就业身份分类时，所依据的工作同于按“职业”、“行

业”与“部门”对人员进行分类时所依据的工作（联合国，2008年，第2.310段）。

583. 应按就业身份，将经济活动人口划分为（联合国，2008年，第2.311段）：

(a) 雇员：可区别其中的固定合同雇员（包括正式雇员）和其他雇员；

(b) 雇主；

(c) 自营工作者；

(d) 家属帮工；

(e) 生产合作社成员；

(f) 不能按身份划分的人员。

584. 法人企业的“业主-经理”通常归入雇员一类，但为便于做出某种说明和分析,可能会将其归为雇主一类，在这种情况下，应单独列出。

585. 该项目的编辑只针对那些在经济活动答复中填报“就业”的人员。如果不采用动态插补，则空白、零或无效答复应改为“未报”。如果采用动态插补，则插补矩阵的变量应至少包括年龄组和性别，但也可采用其他变量，如：受教育程度或行业大类。

6. 收入

586. 《人口和住房普查的原则和建议》中与人口经济特征有关的普查细目重点关注了国际劳工组织（劳工组织）建议中所界定的经济活动人口，该建议中的经济生产概念是相对于《国民账户体系》而确定的（联合国，2008年，第2.331段）。经济活动人口包括在特定所涉期内，为经济货物和服务的生产（见《国民账户体系》定义），提供或等待提供劳动力的所有男性和女性（联合国，2008年，第2.237段）。

587. 在这个框架内，收入可按以下方面定义：(a) 每个活动人员从其所从事的工作中获取的月现金和/或实物收入，或者(b) 住户从各种来源获得年现金和/或实物总收入。在一般的实地调查中，特别是人口普查，收集可靠的收入数据，尤其是自营职业收入和财产收入，将相当困难。列入非现金收入又进一步加大了难度。在人口普查中收集收入数据，即使限于现金收入，也会带来工作负担、答复误差和其他方面的特殊问题。因此，该细目，包括更广义的收入，一般更适合用于抽样调查。视本国的要求而定，各国可能希望获得有限的现金收入信息。按照这种定义，所收集的信息除了可服务于普查的直接目的外，还可为住户收入分配、消费和积累方面的统计提供一定的帮助。

588. 《人口和住房普查的原则和建议》第二次修订本确定了两类收入：个人收入和住户收入。两个项目均需进行类似的编辑。就个人收入而言，如果不采用动态插补，则应为无效收入答复分配“未报”或“未知”。如果采用动

态插补，可采用年龄、性别、受教育程度、行业、职业和其他参数来形成收入的插补矩阵。

589. 住户收入是住户赚得的所有收入之和，登在住房记录上。采用动态插补的编辑规则大致相同，不过采用的是户主而非每个人的年龄、性别和受教育程度。有关住户和家庭收入二次编码的更深入讨论，见附件一。

7. 机构部门

590. 就业的机构部门与工作所属企业的法律组织、主要职能、行为和目标有关（联合国，2008年，第2.335段）。

591. 某些潜在行业和职业与就业的机构部门（公司、政府、非营利部门、住户或其他部门）之间存在着某种关系。有些国家可能决定检查变量之间的这些关系，以便在对这些变量进行交叉列表时，确保列表不会出现相互矛盾的问题。

592. 就编辑而言，不采用动态插补的国家将需要在情况未知的情况下，给机构部门分配答复——“未知”。采用动态插补的国家应考虑采用地理区域内类似人员的年龄和性别，也许还需要采用主要行业或职业。

8. 非正规部门的就业

593. 在非正规部门活动对创造就业和创收起着重要作用的情况下，有些国家采用有关活动状况的项目和其他经济项目，来确定非正规部门。其他国家则通过具体问题询问非正规部门的参与情况（联合国，2008年，第2.337段）。

594. 非正规部门的编辑应简单明了。如果非正规部门的参与情况和正规部门的参与情况无关，那么对于空白或无效登入项，可分配“未知”，或者采用基于年龄和性别的热卡。如果非正规部门的参与情况和正规部门的参与情况有关，则可在热卡矩阵中补充变量，以显示该人是否同时还在正规部门工作（联合国，2008年，第2.343段）。

9. 工作地点

595. “工作地点”是现时就业者从事工作的地点，也是一个通常就业者主要工作的地点——该主要工作用以确定该就业者的其他经济特征，如职业、行业和就业身份。尽管有关工作地点的信息可用来建立就业劳动力的地区档案（相对于按居住地建立的人口档案），但主要目的是将工作地点信息与居住地信息联系起来（联合国，2008年，第2.346段）。

596. “工作地点”用于通勤方面的统计，因此填报信息的任何变动都需反映所考虑的具体地理区域。这样，国家的编辑小组不妨考虑对无效情况分配“未知”，而且只对“已知”情况进行分析。

597. 如果接受自填项目，并必须加以编码，则该项目的编码活动将会使时间和复杂程度增加。如果确定数位的等级，例如：第一位数表示省，第二位数表示区等，则很可能会提高编码活动的效率和精确度。

598. 就插补矩阵而言，数据处理人员需要确保分配给矩阵的只是那些可能性较大的地理位点。明智的做法可能是，为每个行政单位或其他地理区域采用新的冷卡，以确保不会选择以前的值。就插补矩阵本身而言，可列入年龄和性别，也许还可列入修订的主要职业或行业大类。另外，国内工作和国外工作可能需要不同的插补矩阵。

599. 本章着眼于《原则和建议》中建议的人口变量。没有任何一个国家应采用所有这些变量，选定变量及其与其他变量之间的空间关系应在温室和普查前调查的条件下，进行全面测试。人口项目与住房变量不同，通常按很多不同的组合进行交叉制表，因此需要进行全面测试。

第　五　章
住房项目编辑

600.　住房项目编辑规范需考虑各项目的有效性以及各项目间的一致性。了解特定国家各项目之间的具体关系，可制定一致性编辑计划，以便为制表提供较高质量的数据。例如，在墙壁由竹子建筑的情况下，住房单元不应有水泥屋顶。同样，如果建筑物内有抽水马桶、浴缸或者淋浴设施，那么单元的房间内就应有自来水。

601.　与人口项目一样，对于缺失的无效项目，编辑小组必须确定，究竟是分配“未报”——“未知”或其他值的静态插补（冷卡）值，还是采用基于其他住房单元特征的动态插补（热卡）值。如前所述，在很多情况下，都倾向于动态插补，因为无需制表阶段所需要的那种插补——在制表阶段，只有制表本身的信息，才可用来对未知项目做出决定。在没有任何其他具有有效答复的相关项目时，这样建立的插补矩阵可为空白、无效登入项或者已确定的不一致情况提供登入项。在有些国家中，全国的住房特征可能有所不同，但在大部分居民区内却差异很小。在其他国家中，各居民区之间，尤其是城市和农村地区之间，特定项目可能有很大差异。在建立插补矩阵时，必须考虑这种差异，对于初始冷卡值，尤其如此。编辑小组可能希望具体说明，在什么情况下，应根据具有其他类似特征的以前住房单元，为空白项提供一个登入项。

602.　除了一国缺乏集体（团体）住所方面的住房信息外，应为每个序号分配一个（而且只有一个）住房记录（见第三章对一系列质量保证办法的简单介绍）。根据编辑小组的决定，编辑程序可在住房记录缺失时生成一个。同样，在出现两个或多个记录时，该程序可以去掉一个或多个记录。

603.　每个住房记录的编辑最好有选择地只针对所适用的项目。编辑的项目可有所不同，具体取决于城市/农村、气候和其他条件。但在实践中，很少有国家有时间或专门知识去开发和实施多重阵列，以修改缺失或不一致的数据。实际进行选择性编辑的国家甚至更少。

604.　不过，出于美观而非技术的原因，随着编辑工作的日益复杂和详细，目前更注重确保选定地理区域只含“适当的”答复，对于住房项目尤其如此。例如，如果一国的某些地理区域没有电，则不应有空调、电冰箱或电炉。可以编写某种编辑规则来解决某些地区的这类问题，以确保最终数据集不会出现任何异常情况。很可能最好采取“散弹”法，去除实际上可能无关的情况。例如，尽管在有些情况下，一个地区的富人会在没有其他方式获取供电的情况下，购买气体发生器使用，但编辑小组仍可决定不将这些情况列入数据集。

605. 调查表收集的资料还取决于住所类型（住房单元或集体住所）以及住房单元是空置还是住用。对于集体或集团住所而言，可将编辑只限于那些在集体住所收集的项目，或者同时在集体住所和其他住房单元收集的项目。

606. 根据定义，无家可归者通常没有住房记录。如果因为一国决定为他们提供标识，而存在这些记录，则该国处理这些记录的方式与处理集体住所记录的方式相同，或者，要求进行完全不同的编辑或不进行任何编辑。

607. 有时，应允许对特定项目采用“未报告”登入项。当国家的编辑小组没有可靠的依据来插补特定特征的答复时，可能会发生这种情况。在决定分配“未报告”答复时必须考虑到，需要为制表提供适当的特征，用于规划和政策。在规划者需要选定信息时，只要“未报告”情况与已报告情况的分布相同，分配“未报告”情况就不应带来问题。但是，如果“未报告”情况有某种偏差，则汇编后的插补就可能会有问题，对于小区域或特定类型的条件，尤其如此。例如，答卷人如果居住在国家界定的“不达标”住房内，则可能不愿透露某些住房特征。如果查点员不予以报告，则规划者可能无法采取补救程序，来减轻不达标状况。

608. 住房项目编辑往往比人口项目编辑简单，因为交叉指标一般没有那么复杂。大多数国家只按照不同的地理级别编辑个别住房特征。如上所述，不采用动态插补的国家应为“未知”确定一个标识符，以便在出现无效或不一致答复时使用。

609. 对于采用动态插补的国家，编辑小组应利用能够区分住房特征的元素，编制简单的插补矩阵。对于大多数国家来说，不管是住房单元还是集体住所——包括这些类别内的单元类型，有关“住所类型”的变量是动态插补的最佳基本变量。

610. 对于某些国家来说，地理区域可作为这些插补矩阵的一个元素，还可采用权属。例如，如果一个国家的住房单元约有半数租赁，半数自有，则适合将权属作为插补矩阵的一个元素。但如果只有5%的单元为租赁，那么可能更适合采用其他某些特征。在插补矩阵中，权属通常是一个有用的变量，在权属大类占有较大比例的国家中，尤其如此。可考虑的其他特征包括：墙壁类型和是否有电。

611. 对于每个国家来说，作为插补矩阵元素列入的特定变量必须与数据集中的变量对应，住房项目即是如此，必须认真确保个别项目以及项目组合能对特征进行区分。

A. 核心细目和补充细目

612. 住房普查的查点单位为(a) 建筑物；(b) 住所；(c) 住所居住者。联合国对人们普遍感兴趣和具有普遍价值的基本编辑细目清单，这些细目对进行全面的国家统计比较具有重要意义。为方便用户，下文将给出这些细目的建议代码和若干补充细目。细目按查点单位类型显示。

1. 住所——类型（核心细目）

613. 下文列出的分类为联合国指定的三位数代码体系（2008年，第2.412-2.454段），该体系按大类将具有类似结构特征的住房单元和集体住所归组。居住者（人口）在不同人群中的分布可提供普查时住房膳宿方面的宝贵信息。分类还为抽样调查分层提供了有用的基础。住所可分为以下类别：

1 住房单元
- 1.1 常规寓所
  - 1.1.1 备有全部基本设施
  - 1.1.2 没有全部基本设施
- 1.2 其他住房单元
  - 1.2.1 半永久性住房单元
  - 1.2.2 活动住房单元
  - 1.2.3 临时住房单元
  - 1.2.4 在不打算供人居住的永久性建筑物中的住房单元
  - 1.2.5 不打算供人居住的其他房舍

2 集体住所
- 2.1 旅馆、寄宿舍和其他寄宿公寓
- 2.2 机构
  - 2.2.1 医院
  - 2.2.2 惩教所（监狱、教养所）
  - 2.2.3 军事机构
  - 2.2.4 宗教机构（男修道院、女修道院等）
  - 2.2.5 退休所、养老院
  - 2.2.6 学生宿舍及类似住所
  - 2.2.7 职工宿舍（如，宿舍和护士之家）
  - 2.2.8 孤儿院
  - 2.2.9 其他
- 2.3 营地和工作者住所
  - 2.3.1 军营
  - 2.3.2 工人营（工棚）
  - 2.3.3 难民营
  - 2.3.4 境内流离失所者营地
  - 2.3.5 其他
- 2.4 其他

614. 编辑小组开发的编辑规则应能确保所有集体住所和住房单元都具有内部一致的信息。如果住所类型的值未知或者无效，则编辑小组不妨开发一个利用其他变量的编辑规则，来分配住所类型。否则，在值为无效，并且不采用动态插补的情况下，应分配“未知”。对无效值采用动态插补的国家统计/普查机构，应至少采用两个特征，如：住房类型、权属、房间数、房屋面积或空置状况，以从地理区域内类似住房单元获取“已知”信息。

2. 住所位置（核心细目）

615. 住所位置属于地理变量，与结构编辑一起在第三章中进行了介绍。

3. 住用状况（核心细目）

616. 对于居住者暂时不住或者暂时“住用”的住房单元，是应将其记录为“住用”，还是记录为“空置”，需要根据所进行的普查是常住人口普查，还是现住人口普查来确定。不论在哪种情况下，都最好尽可能将用作基本住所的住房单元与用作第二住所的住房单元区分开来。在第二住所与基本住所的特点有明显区别时，这一点非常重要。属于这种情况的如：在一年中的某个季节，农业住户从其在乡村中的永久住所迁移到农业生产经营单位的简易房中（联合国，2008年，第2.466段）。对常规住房住用状况的建议分类如下：

1 住用
2 空置
- 2.1 季节性空置
  - 2.1.1 度假房
  - 2.1.2 季节工人住所
  - 2.1.3 其他
- 2.2 非季节性空置
  - 2.2.1 第二住所
  - 2.2.2 出租
  - 2.2.3 出售
  - 2.2.4 待拆
  - 2.2.5 其他

617. 如果住房单元被住用，则居住者数量和人口计数方面的记录不得为零。如果没有记录任何人，则要么单元为空置，要么人员缺漏。如前面在结构编辑中所述，专家必须制定用以确定单元是否空置的程序。如果被列为住用，但实际上却空置，则必须制定相应的办法来确定空置类型，即：将其列为“未知”，或者采用动态插补。如果单元被列为空置，但实际上却被住用——因为存在关于居住者数量或人口计数的记录，那么，必须将住用状况改为“住用”。

618. 如果该项目的值为无效，居住者数量的值为零，而且没有人口记录，那么在不采用动态插补，应分配“未知空置”。如果该项目的值为无效，但居住者数量不为零或者有人口记录，则应分配“占用”。对无效值采用动态插补（以插补空置类型）的国家，应至少采用两个特征，以从地理区域内类似的住房单元获取“已知”信息，或者分配“未知空置”。

4. 所有权——类型（核心细目）

619. 该细目指住房单元本身的所有权类型，而不是它们所在土地的所有权类型（联合国，2008年，第2.467段）。不应将所有权与权属混淆。应收集有关资料说明下列情况：住房单元是由公共部门（中央政府、地方政府、公营公司）拥有，还是属于私有（由住户、私营公司、合作社、住房协会等所拥有）。这个问题有时需要加以充分阐述，说明在以分期付款或按揭贷款购买的情况下，款额是否全部付清。住房单元可按所有权分类如下：

1 房主自住

2 非房主自住

2.1 共有

2.2 私有

2.3 社区拥有

2.4 合作社拥有

2.5 其他

620. 如果所有权与权属有关，那么在开发编辑规则时，应考虑到这一点；如果无关，则所有权类型很可能独立于其他住房变量。如果“所有权类型”的值无效，那么应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（可采用墙壁建筑材料、权属、住房单元类型和房间数），以从地理区域内类似住房单元获取“已知”信息。

5. 房间——数量（核心细目）

621. 根据定义，房间是指位于住房单元或其他住所中的，四周由从地面到天花板或房顶的墙，或由至少2米高的墙所围住的、一块能放置一张成人床（即：至少有4平方米面积）的空间。因此，各类房间总数包括：卧室、餐厅、起居室、书房、可居住的阁楼、佣人房、厨房、用于专业或商业目的的房间，以及能够满足墙或建筑面积标准、用于或准备用于居住的其他独立空间。通道、阳台、大厅、卫生间和厕所等即使符合标准，也不能算作房间。对于面积小于4平方米，但在其他方面符合“房间”定义的空间，如果需要了解其数量，可以为国家统计之目的，单独收集这方面的资料（联合国，2008年，第2.472段）。

622. 房间数可能独立于其他住房变量，因此，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国

家，应至少采用两个特征（如，住房单元类型、墙壁建筑材料、权属和空置状况），以从地理区域内类似住房单元获取“已知”信息。

6. 卧室——数量（补充细目）

623. 除了查点房间数目外，很多国家普查还收集一个住房单元中的卧室数量，这一细目的查点单位是住房单元。根据定义，卧室是指内有一张床、供晚间休息的房间（联合国，2008年，第2.475段）。

624. 有时，查点员对卧室数报告的值大于房间数的值。[9]在这种情况下，如果国家只对无效或不一致的答复采用“未报”，那么卧室数量一项应采用“未报”。如果采用动态插补，则应根据一个将房间数作为元素之一的插补矩阵，来对卧室进行“估计”。这样，卧室数将不会大于房间数，因为卧室的值只有在房间和卧室的值一致时，才予以更新。最简单的情况是线性阵列，以房间数作为单元，卧室的值处于单元内。更复杂的插补矩阵可能包括住房单元的人数和建筑类型。

[9] 如果房间和卧室都有，则应一起编辑。卧室数不应超过房间数。卧室数为“补充”细目，因此只有当两者都有时，才进行该项目的编辑。

625. 否则，如果卧室的值无效，那么应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（其中一个为房间数），以从地理区域内类似住房单元获取“已知”信息。

7. 使用面积（补充细目）

626. 该细目是指住房单元中可用的房屋面积，即根据住房单元外墙内侧测量而得的房屋面积，不包括不适于居住的地下室或阁楼。在多住房的建筑物中，所有共用空间都不包括在内。对住房单元和集体住所应采取不同的办法（联合国，2008年，第2.476段）。

627. 房屋面积可能与房间数和/或卧室数有关，因此国家的编辑小组在开发编辑规则时，可能希望将此列入考虑。动态插补的其他有用项目包括：居住者数量和每房间的居住者数量。在大多数情况下，房屋面积都独立于其他住房编辑项目。可能需要指定平方米之类的度量单位。如果该项目的值无效，那么应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（包括住房单元类型、墙壁的建筑材料、权属和空置），以从地理区域内类似住房单元获取“已知”信息。

8. 供水系统（核心细目）[10]

628. 根据联合国（2008年，第2.479段）的规定，普查需要取得的基本供水系统资料包括：住房单元内是否安装了自来水，该资料将显示，是否经由整个社区系统的管道，或诸如高压箱、抽水机等私人供水装置，向住房单元供水。这一细目的查点单位是住房单元。还需要列明在该单元内是否有水笼头，如果没有，水笼头与门之间的距离是否在某个范围内。建议的距离是200米，因为一般认为，在这一距离内取水家用时，不会让住房单元内的居住者过于费

[10] 关于以下变量，细目单位实际是指住房单元的供水系统、厕所和排污设施、洗浴设施、烹饪设施、照明和固体废物处置。

力。除了水笼头的位置外，还可能特别希望获得供水来源的资料。因此，建议按供水系统将住房单元分类如下：

1 单元内有自来水

1.1. 由社区系统供水

1.2. 由私人来源供水

2 在单元外，但在200米范围内有自来水

2.1. 由社区系统供水

2.1.1. 专用

2.1.2. 合用

2.2 由私人来源供水

2.2.1. 专用

2.2.2. 合用

3 其他

629. 社区供水系统要受公共机构的检查和管制。这种系统一般由公共机构经营，但在有些情况下，由合作社或私人企业经营。

630. 与水有关的设施项目——供水系统、饮用水、厕所和排污设施、洗浴设施和热水供应——很可能应该一起编辑。这些项目密切相关，因此在一个项目缺失或无效的情况下，可采用其他项目生成相应的值。在没有自来水的地理区域，专家可能需要对这些单位采用专门的编辑规则。另外，该地区的其他单位很可能有类似特征，建议在采用动态插补时，将这些项目用于动态插补。

631. 如果水系统的值无效，那么应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（这一般包括住房单元类型，其次包括厕所和排污设施、洗浴设施），以从地理区域内类似住房单元获取“已知”信息。

9. 饮用水——主要来源（核心细目）

632. 饮用水应与水系统一起编辑，上述的很多标准在此也适用。由于瓶装水和非传统的其他饮用水来源通常会列入调查表，所以也必须列入编辑（联合国，2008年，第2.483段）。

633. 如果饮用水的值无效，那么应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（这一般包括住房单元类型，其次包括水系统、厕所和排污设施、洗浴设施），以从地理区域内类似住房单元获取“已知”信息。

10. 厕所——类型（核心细目）以及

11. 污水处理（核心细目）

634. 厕所设施和污水处理项目应与其他管道设备变量一起编辑，以取得最一致的结果。尽管《人口和住房普查的原则和建议》第二次修订本将这两个变量结合在一起，但2010年则将它们分开。不过，这些项目应一起编辑，并且在可能的情况下，应采用同样的动态插补矩阵。

635. 有些国家认为，最好将非抽水马桶的分类加以扩充，以区别某些被广泛使用的类型，并说明一定程度的卫生情况（联合国，2008年，第2.487段）。建议按厕所设施将住房单元分类如下：

1 住房单元内有厕所
- 1.1 抽水马桶/手动冲水马桶
- 1.2 其他

2 在住房单元外有厕所
- 2.1 专用
  - 2.1.1 抽水马桶/手动冲水马桶
  - 2.1.2 通风的坑式改良厕所
  - 2.1.3 不通风但有盖的坑式厕所
  - 2.1.4 临时盖上或无棚的洞或挖坑
  - 2.1.5 其他
- 2.2 合用
  - 2.2.1 抽水马桶/手动冲水马桶
  - 2.2.2 通风的坑式改良厕所
  - 2.2.3 不通风但有盖的坑式厕所
  - 2.2.4 临时盖上或无棚的洞或挖坑
  - 2.2.5 其他

3 无厕所
- 3.1 便桶（人工清除排泄物）
- 3.2 利用自然环境，如：灌木丛、河、溪等

636. 厕所设施和污水处理是与水有关的其他住房项目，应与其他与水有关的项目一起编辑。与“私人”、“合用”、“专用”等有关的值可用来确定各值是否一致，如果不一致，可用来确定采取何种编辑路径来解决问题。在存在一个或多个与水有关的其他变量时，无需采用“未知”或动态插补，便可为未知或不一致信息做出估计。但如果这不能提供一个有效的值，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采

用两个特征（一般包括住房单元类型，还包括供水、墙壁建筑材料、权属和空置状况），以从地理区域内类似住房单元获取“已知”信息。

12. 洗浴设施（核心细目）

637. 根据联合国（2008年，第2.490段）的建议，应取得关于每套住房单元地基范围内是否安装固定浴缸或淋浴装置的资料。这一细目的查点单位也是住房单元。可收集补充资料，说明其装置是否为住所居住者专用，以及是否有供洗浴用的热水或是只有冷水。在世界上的有些地区，上述提议的区分也许不是满足国家需求的最佳选择。例如，也许需要区别是否在住所内有单独的浴室、在建筑物内有单独的浴室、在建筑物内有开放浴室、公共澡堂等。建议按有无洗浴装置及其类型将住房单元分类如下：

1 在住房单元内有固定浴缸或淋浴装置

2 在住房单元内无固定浴缸或淋浴装置

2.1 在住房单元外有固定浴缸或淋浴装置

2.1.1 专用

2.1.2 合用

2.2 无固定浴缸或淋浴装置

638. 洗浴设施类型应与其他与水有关的项目一起编辑。与“私人”、“合用”、“专用”等有关的值可用来确定各值是否一致，如果不一致，可用来确定采取何种编辑途径来解决问题。在存在一个或多个与水有关的其他变量时，无需采用“未知”或动态插补，便可为未知或不一致信息做出估计。但在其他所有办法都不成功时，如果该值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（一般包括住房单元类型，其次包括供水、墙壁建筑材料、权属或空置状况），以从地理区域内类似住房单元获取“已知”信息。

13. 有无厨房（核心细目）

639. 《人口和住房普查的原则和建议》第二次修订本（联合国，2008年，第2.494段）指出，收集关于是否有厨房这一方面的资料便于收集有关烹调设备种类（如，炉子、烤盘或明火）的资料，以及是否有厨房洗涤槽和防止食物腐坏的食物贮存间等资料。建议按照有无厨房或专作烹饪用的其他空间，将住房单元分类如下：

1 住房单元内有厨房

1.1 专用

1.2 合用

2 住房单元内有作烹饪用的其他空间，如小厨房

2.1 专用

2.2 合用

3 住房单元内无厨房或作烹饪用的其他空间

3.1 在住房单元外有厨房或作烹饪用的其他空间

3.1.1 专用

3.1.2 合用

3.2 无厨房或作烹饪用的其他空间

640. 烹饪设施项目的编辑利用与“私人”、“合用”、“专用”等有关的值，来确定各值是否一致，如果不一致，可用来确定采取何种编辑路径来解决问题。在存在一个或两个烹饪变量时，无需采用“未知”或动态插补，便可为未知或不一致信息做出估计。但如果该值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（一般包括住房单元类型，其次包括供水、墙壁建筑材料、权属或空置状况），以从地理区域内类似住房单元获取“已知”信息。

14. 烹饪燃料（核心细目）

641. 在需要对自然资源使用进行密切监督的背景下，很多国家都将烹饪燃料这一细目列入其住房普查中。查点单位为住房单元。“烹饪燃料”是指准备正餐时主要使用的燃料。如果使用了两种燃料（如，电和燃气），则应统计最常用的燃料。烹饪燃料的分类需视国情而定，可能涉及电、燃气、油类、煤、柴禾、动物粪便等。针对集体住所收集这方面的资料，也有用，如果一个国家的集体住所数目比较客观，则尤应如此（联合国，2008年，第2.496段）。

642. 烹饪燃料类型的答复应与烹饪设施的答复一起编辑。编辑小组确定这两个变量之间的关系，并制定相应的编辑规则，来检查两者之间的一致性。与“私人”、“合用”、“专用”等有关的值，很可能用来确定各值是否一致，如果不一致，可用来确定采取何种编辑路径来解决问题。在存在一个或两个烹饪变量时，无需采用“未知”或动态插补，便可为未知或不一致信息做出估计。但如果该值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（包括烹饪设施，建筑类型、墙壁建筑材料、权属或空置状况），以获取类似于地理区域内住房单元的信息。

15. 照明和/或供电——类型（核心细目）

643. 应收集有关住房单元内照明类别的资料，例如电、燃气、油灯或其他某些来源提供的照明。如果以电照明，有些国家可能希望收集资料，说明其电力是主要由社区、发电厂供应，还是通过其他某些来源（如工业设备）供应。除照明类别外，各国可对电力是否用于照明以外的目的（例如烹调、热水、暖气等）进行评估。如果根据一国的住房条件，可以从有关照明种类的资料获取该信息，则无需进行进一步调查（联合国，2008年，第2.497段）。

644. 如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（一般包括住房单元类型、墙壁建筑材料、权属或空置状况），以从地理区域内类似住房单元获取“已知”信息。

16. 固体废料处理——主要类型（核心细目）

645. 根据《人口和住房普查的原则和建议》第二次修订本（第2.500段）的定义，该细目是指对住房单元中居住者所产生的固体废料/垃圾进行的收集和处理。该细目的查点单位为住房单元。建议依照以下准则，按照固体废料的处理类型将住房单元分类：

1 由指定废料收集者定期收集的固体废料
2 由指定废料收集者不定期收集的固体废料
3 由自行指定的废料收集者收集的固体废料
4 由居住者将固体废料倒入当局监督的当地垃圾堆场
5 由居住者将固体废料倒入非当局监督的当地垃圾堆场
6 由居住者燃烧固体废料
7 由居住者填埋固体废料
8 居住者将固体废料倒入河/海/溪/池塘中
9 居住者将固体废料制成堆肥
10 其他安排

646. 固体废料独立于其他住房变量。如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（一般包括住房单元类型、墙壁建筑材料、权属、空置状况或厨房设施），以从地理区域内类似住房单元获取“已知”信息。

17. 供暖——类型及所用能源（补充细目）

647. 该细目是指住房单元的供暖类型及所用能源。查点单位为所有住房单元。该细目对有些国家无关，因为地理位置和气候的缘故，这些国家的住所不需要暖气。供暖类型是指为绝大部分空间提供暖气的供暖系统类型：可能是所有住所或一套住所的中央供暖系统，也可能不是中央供暖系统，而是在住所内分别以炉、壁炉或其他取暖装置提供暖气。“供暖所用能源”，与供暖类型密切相关，是指主要的能源，例如固体燃料（煤、褐煤及其制品、木材）、石油、气体燃料（天然气或液化气）、电等（联合国，2008年，第2.501段）。

648. 供暖类型和供暖所用能源相互关联，还与热水供应和住房单元内的其他公用设施（如，电和管道燃气）有关。在为供暖类型和供暖所用能源制定编辑规范时，编辑小组应考虑这些项目的可用性。供暖类型可能独立于其他住房项目，因此可能必须单独编辑。但在“供暖所用能源”为未知或不一致时，

程序可以检查照明所用的能源类型。最后，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，住房单元类型、墙壁建筑材料、权属和空置状况），以从地理区域内类似住房单元获取“已知”信息。

18. 有无热水（补充细目）

649. 这一细目是指住房单元内是否有热水供应。热水是指加热至一定温度以管道提供给用户使用的水。收集的资料可说明在住所内或住所外是否有专用或合用的热水（联合国，2008年，第2.502段）。

650. 有无热水可能与水的加热方法有关——尽管使用太阳能给水加热可能与其他住房项目无关。编辑小组必须根据其他住房项目和地理位置的情况，确定适当的编辑规则。最后，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，自来水的特征），以从地理区域内类似住房单元获取“已知”信息。

19. 有无管道燃气（补充细目）

651. 该细目是指住房单元内是否有管道燃气。管道燃气通常指通过管道输送，并对消耗量加以记录的天然或人工燃气。这一细目也许同某些国家无关，因当地缺乏天然气或未发展管道系统（联合国，2008年，第2.503段）。

652. 管道燃气与照明类型和烹饪燃料以外的其他住房项目无关。编辑小组必须确定适当的编辑路径以及如何检查一致性。如果该项目的值无效或不一致，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，加热所用能源、建筑类型、住房单元类型、墙壁建筑材料、权属和空置状况），以从地理区域内类似住房单元获取“已知”信息。

20. 住房单元的利用（补充细目）

653. “住房单元的利用”是指该住房单元是否完全用作居住（住宅）目的。住房单元可用作居住，并可用作商业或制造业目的，或其他目的（联合国，2008年，第2.504段）。

654. “住房单元的利用”独立于其他住房项目。如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，住房单元类型、墙壁建筑材料、权属和所有权），以从地理区域内类似住房单元获取“已知”信息。

21. 一个或多个住户居住（核心细目）

655. “多个住户居住”独立于其他住房项目。如果该项目的值无效，一国应计算户主数量并加以采用。重要的是要注意，必须首先通过结构编辑确定户主，然后再进行这种编辑。

22. 居住人数（核心细目）

656. 通常居住在一住房单元或一套集体住所内的每个人应算作一名居住者。因此，该细目的查点单位是住所。然而，由于住房普查通常与人口普查同时进行，这一定义是否适用取决于在人口普查中收集和记录的个人资料是否需要指出当事人在普查日所在的地点，或是否需要指出其惯常住所。应注意区别将活动单元（例如，船、篷车和拖车）作为住所的人员和利用这些单元作为交通工具的人员（联合国，2008年，第2.510段）。

657. “居住人数”与人口数记录有关，两者应一致。如果不一致，必须采取措施纠正居住人数项目或人口数记录。通常情况下，居住人数在经过调整后等于单元内的人数。该项目不应为“未知”，也不应插补。

23. 建筑——类型（核心细目）

658. 对于有些空间被用作居住目的的建筑物，联合国（2008年，第2.514段）建议按建筑类型进行以下分类。

1 包含一个住房单元的建筑物
- 1.1 独栋
- 1.2 联栋

2 包含多个住房单元的建筑物
- 2.1 一至二层
- 2.2 三至四层
- 2.3 五至十层
- 2.4 十一层或以上

3 由机构住户成员居住的建筑物

4 其他所有建筑物

659. 如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（这可包括外墙建筑材料、建造期间和/或建筑物内住房单元类型），以从地理区域内类似住房单元获取“已知”信息。

24. 建造年份或时期（补充细目）

660. 建造年份或时期指住所所在建筑物的年龄。对于在上一个普查间隔期间修建的建筑物，如果未超过10年，建议调查这类建筑物的准确修建年份。如果普查间隔期间超过10年，或者以前从未进行过普查，则应调查在过去10年里所建建筑物的准确年份。对于修建期间超过10年的建筑物来说，收集资料的期间应有助于评价现有房屋存量的年龄。在收集该细目的数据时可能会遇到困难，因为在有些情况下，居住者可能不知道建造日期（联合国，2007年，第2.519段）。

661. 有些国家，甚至采用动态插补的国家，对关于建筑年份或时期的项目，接受“未知”答复。在这种情况下，这类国家可能不会对该项目采用动态插补，即使对其他变量采用了插补矩阵，也是如此。对无效值采用动态插补的国家，应至少采用两个特征（包括建筑类型、外墙建筑材料和/或建筑物内住房单元的类型），以从地理区域内类似住房单元获取“已知”信息。

25. 建筑物内住所——数量（补充细目）

662. 建筑物内住房单元数目的编辑已在第三章作为结构编辑的一部分进行了解释。

26. 外墙建筑材料（核心细目）

663. 该细目指住所所在建筑物外部（外层）墙壁的建筑材料。如果这些墙壁由多种材料修建，则应报告所用的主要材料。所区分的类型（砖、混凝土、木材、土坯等）取决于有关国家最常使用的材料，以及这些材料对建筑耐久性或评估持久性的重要性（联合国，2008年，第2.525段）。

664. 如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如建造时期和/或建筑物内住房单元类型），以从地理区域内类似住房单元获取“已知”信息。

27. 地面、屋顶的建筑材料（补充细目）

665. 在有些情况下，也许特别希望了解建筑屋顶和地板的材料，以进一步评估建筑物内住所的质量。这一细目是指用于屋顶和/或地板的材料（但也许指建筑物的其他部分，例如框架或地基，具体情况要视各国的特定需求而定）。一般只列举主要材料，就屋顶而言，可能是瓦片、水泥、金属板、棕榈、禾秆、竹或类似的植物材料、泥、塑料板或其他某些材料（联合国，2008年，第2.528段）。

666. 有时，有关外墙建筑材料的答复与屋顶建筑材料的答复不一致；例如，如果所列的墙壁建筑材料不足以支撑屋顶时，会发生这种情况。如上所述，在这种情况下，专家必须确定是改变两个变量中的一个，还是采用“未知”作为答复。如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如建筑类型、外墙建筑材料、住房单元类型、墙壁建筑材料、权属和空置状况），以从地理区域内类似住房单元获取“已知”信息。

667. 填报的地面建筑材料与屋顶和墙壁的建筑材料可能一致也可能不一致。如果国家的编辑小组发现了不一致或无效组合，则必须确定是分配“未知”，还是采用插补矩阵以改变一个或多个答复。如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如建筑类型、外墙建筑材料、住房单元类型、权属和空置状况），以从地理区域内类似住房单元获取“已知”信息。

28. 有无电梯（补充细目）

668. 该细目是指在多层建筑物内是否有电梯（一个可上下运送人和货物的封闭平台）。需要收集资料，查明电梯是否在大部分时间都运行（即：除了定期维修外，多数时间都在运行的电梯）（联合国，2008年，第2.529段）。

669. 如果建筑物只有一层或者是一个单一的独立单元，则不应有电梯。如果有电梯，编辑小组必须确定以哪个优先：是层数，还是电梯。如果电梯优先，则必须改变层数——要么使值为“未知”，要么采用动态插补取得另一值。如果层数有限，而建筑物只有一层，则“有电梯”的答复必须改为“无”。

670. 在有电梯的情况下，如果需要电，则应该查明建筑物内是否有电。

671. 最后，如果电梯的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如建筑类型和外墙建筑材料），以从地理区域内类似住房单元获取“已知”信息。

29. 农场建筑（补充细目）

672. 有些国家的普查可能需要具体说明被查点建筑物是否为农场建筑物。农场建筑物为农业生产经营单位的一部分，用于农业和/或住房之目的（联合国，2008年，第2.531段）。

673. 农场建筑独立于其他住房项目。各国可决定检查该项目与职业和行业人口项目之间的一致性。另外，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征，以从地理区域内类似住房单元获取“已知”信息。

30. 修缮状况（补充细目）

674. 该细目是指建筑物是否需要修缮以及需要修缮的种类。查点单位为建筑物。可根据以下修缮情况对建筑物分类：“不需修缮”、“需要小修”、“需要适度修缮”、“需要大修”、“无法修缮”。小修多半是指建筑物及其组成部分的定期维修，例如修缮破裂的窗子等。适度修缮是指修补一般性的问题，例如屋顶檐沟缺失、大片灰泥破裂、楼梯无安全扶手等。在建筑物有严重的结构问题时需要大修，例如屋顶板或瓦片缺失、外部墙面破裂和穿孔、楼梯缺损等。“不能修缮”一词是指无法修缮的建筑物，换言之，由于严重的结构问题，这类建筑物应予拆除而非修缮；该词通常是指那些仅剩下框架，没有完整外墙和/或屋顶的建筑（联合国，2008年，第2.532段）。

675. 建筑物修缮状况独立于其他住房变量。因此，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，建筑物类型、外墙建筑材料和住房单元类型），以从地理区域内类似住房单元获取“已知”信息。

31. 户主和住户其他基准成员的特征（核心细目）

676. 户主特征通常从人口记录中取得，可帮助提供用于规划和分析的交叉指标信息。这些项目，包括年龄、性别、原属种族、宗教或收入有助于确定不同的社会状况或需求。由于已针对人口项目编辑了这些特征，因此无需进一步编辑（联合国，2008年，第2.533段）。

32. 权属（核心细目）

677. 根据联合国（2007年，第2.536段）的定义，权属是关于住户居住整个或部分住房单元的安排。查点单位是居住一住房单元的住户。可按权属将住户分类如下：

1 住户成员拥有住房单元

2 住户成员租用整个或部分住房单元

2.1 住户成员作为二房东租用整个或部分住房单元

2.2 住户成员作为三房客租用整个或部分住房单元

3 免租金占用住房单元

4 其他安排

678. 权属可能与所有权类型有关，因此，编辑小组在拟定编辑规则时，可能需要考虑这两个项目之间的关系。另外，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，住房单元类型、租赁和空置状况），以从地理区域内类似住房单元获取“已知”信息。

33. 租金和房主自住成本（补充细目）

679. 显然，租金费用应只在租赁单元下发生，而房主费用应只在房主自住单元下发生，除此之外，“租金和房主自住成本”项目将独立于其他住房变量。编辑小组必须查看每种情况，然后确定这些变量之间的最适当关系。如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征，以从地理区域内类似住房单元获取“已知”信息（联合国，2008年，第2.540段）。

34. 配备家具或不配备家具（补充细目）

680. “单元是否配备家具”是一个新项目。编辑小组应考虑对该项目进行测试，如果需要列入，那么在采用动态插补解决无效值或不一致问题的情况下，应确定最适合动态插补的项目（联合国，2008年，第2.542段）。

35. 有无信息和通信技术设备（核心细目）

681. 在当代社会，是否配备信息与通信技术设备正日显重要。这些设备所提供的服务正在改变着主要社会经济现象的结构和格局。住房普查为了解住

户利用这些设备的情况提供了一个极好机会。所选择的细目应足以了解信息和通信技术在住户中的地位，并且能够用于政府和私人部门的规划活动，以扩大和改善所提供的服务，评估它们对社会的影响。建议的分类如下：

1 拥有收音机的住户

2 拥有电视机的住户

3 拥有固定电话的住户

4 拥有蜂窝式移动电话的住户

5 拥有个人电脑的用户

6 家中接入互联网的住户

7 从家以外的其他地方接入互联网的用户

8 没有接入互联网的住户

682. 信息和通信技术设备是新项目。需要用电的项目应只在有电的地方才有。但随着太阳能、风能和其他“可再生能源”的使用更加频繁，在为该项目拟定编辑规则时，必须考虑该因素。国家编辑小组应在进行普查或调查前，全面测试该项目及其插补矩阵。可用于热卡插补的项目包括住户的社会水平（例如，可根据财富指数确定）和户主年龄（联合国，2008年，第2.543段）。

683. 这些细目指住房单元内相应项目的可用性。例如，电话表示电话线而非电话机本身，因为一根电话线可以连接多部电话（联合国，2008年，第2.547-2.548段）。电话与编辑期间使用的其他住房项目无关。但如果某些地理区域没有电话，则编辑小组应在拟定编辑规则时考虑到这一点。如果“电话”项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，住房单元类型、墙壁建筑材料和权属），以从地理区域内类似住房单元获取“已知”信息。

36. 汽车数量（补充细目）

684. “汽车数量”是指住户成员通常可用的汽车和货车数目。“通常可用”一词指居住者拥有的、或根据其他某些或长或短协议（如，租约等）居住者可使用的汽车和货车，另外还包括由雇主提供但住户可以使用的汽车和货车，但不包括专门用于运货的货车（联合国，2008年，第2.551段）。

685. 汽车数量独立于其他住房变量。如果一国的有些地区没有任何车辆，那么专家可能希望对特定地理区域采取特别的编辑规则。另外，如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，住房单元类型、墙壁建筑材料、权属和所有权，或在该特例的情况下，采用成人居住者数量），以从地理区域内类似住房单元获取“已知”信息。

37. 有无耐用家电（补充细目）

686. 可根据国情收集相关资料，以说明住户是否具备诸如洗衣机、洗碗机、冰箱、冷藏箱之类的耐用家电（联合国，2008年，第2.552段）。

687. 对于大多数家电来说，必须在单元内有电，才可使用。在出现这些项目时，编辑小组应考虑采用某种编辑规则来检查电的情况（以煤气作为动力的冰箱或“冰柜”可能例外）。此外，如果在具体国家中，洗衣机或洗碗机需要自来水，那么编辑规则还需要将此列入考虑。可利用编辑规则确定特定项目是否存在，具体要看电和水的供应情况而定，应在出现不一致情况时，采取适当的措施。同时，一国的特定地区可能没有电或自来水，专家在拟定编辑规则时，可能需要承认这一点。如果该项目的值无效或不一致，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，住房单元类型、电、墙壁建筑材料和权属），以从地理区域内类似住房单元获取“已知”信息（因为住户的社会水平应类似）。

38. 有无户外空间（补充细目）

688. 该细目指是否具备可供住房单元内住户成员进行休闲活动的户外空间。分类可包括：作为住房单元一部分（如，独立住房的场院）的户外空间、与建筑物毗连的户外空间（如，与公寓房屋毗连的场院和操场）、距住房单元10分钟步行距离作为公共休闲区一部分的户外空间（如，公园、体育中心和类似场地等）或在10分钟步行距离内没有户外空间（联合国，2007年，第2.553段）。

689. 可供住户使用的户外空间数量独立于其他住房项目。但在某些地理区域或某些类型的建筑物中，可能没有任何户外空间。编辑小组在拟定编辑规则时，可能需要考虑到具体的情况。如果该项目的值无效，则应在不采用动态插补的情况下，分配“未知”。对无效值采用动态插补的国家，应至少采用两个特征（如，建筑物类型和住房单元类型），以从地理区域内类似住房单元获取“已知”信息。

B. 住用和空置的住房单元

690. 上述编辑规则针对的是被住用的住房单元。但空置住房单元和住用的住房单元有时具有不同的特征，因而不采用相同的编辑规则。如果对于空置住房单元并不收集所有住房项目（通常都会如此），那么国家统计/普查机构的编辑小组需要为每类单元拟定不同的编辑规则。编辑小组需要特别注意插补矩阵变量，因为这些很可能会有不同。

691. 本章考察了《人口和住房普查的原则和建议》第二次修订本所建议的住房变量。没有任何一个国家应该采用所有这些变量，为确保答复的全面性和可靠性，应在温室和普查前调查中全面测试所选定的变量及其与其他变量之间的空间关系。住房变量本身就很重要，可用来编制财富指数，对一国全部或部分地区的福利情况进行评估。

附　件　一
衍生变量

1.　为最有效地利用普查或调查数据，各国通常需要那些源自其他变量组合或变体的变量。例如，有关经济活动状况（见第四章第D.1部分）就属于所收集的若干普查变量的组合。不必在国家统计/普查机构每次想要进行特殊制表时，都编写一个程序对有关信息二次编码，可由数据处理专家编写一个程序一次性进行二次编码，在个人记录上存储二次编码资料，然后在以后的制表中使用。在确定是否生产和存储二次编码资料时，国家统计/普查机构需要确定制表中使用二次编码的频率以及特定二次编码的相关性。重要的是要记住，二次编码也会占用个人记录的空间。人口规模越大，使用的空间也越多。

2.　很多变量都可以这样生成。例如，如果填报了出生日期，那么为填报年龄，可以用普查基准日减去出生日，一次性确定年龄，然后将该资料存在记录上。同样，可以将每个人的收入相加，来获得住户收入，然后将该数据列入住房记录，供以后使用。

3.　有时衍生变量得自单个记录中一个或若干登入项的组合，或这有时来自若干记录。例如，“非经济活动人口——上学”这一类别可能需要查看多达四个项目的答复。在制作表格格式或规划补充表时，采用衍生变量不仅使编程更容易和更有效率，而且还有助于使各个时期的数据具有可比性。下文将介绍一些具有可能性的衍生记录的例子。

A.　住房记录的衍生变量

1.　住户收入

4.　住户收入的衍生变量等于住户内所有人员在所有收入类别中的收入之和。收入类别信息可包括：工资、自营收入、利息和股息、社会保障和退休收入、汇款、特许权使用费和租金。如果还收集总收入，编辑期间应通过各类别的相加，来检查每个人的总收入。然后根据所记录的总收入来核对该总收入。如果合计收入不等于填报的总收入，则编辑小组必须拟定纠正计划。要么改变总数，以反映各部分之和，要么改变一项或多项个人收入类别。在确定了住户所有个人的总收入时，可通过加总个人收入，得出住户收入变量的值。

5.　编辑小组必须考虑到，住户内一人或多人因业务失败或其他原因而有负收入的情况。在这种情况下，住户总收入将减少，而不是增加，减少或增加的金额相当于该人的收入。

2. 家庭收入

6. 家庭收入的衍生变量等于家庭内所有人员在所有收入类别中的收入之和。与住户不同的是，家庭通常只包括彼此有亲缘关系的个人，但该定义将取决于特定国家的情况。对于某些国家来说，住户与家庭是一回事，因此，无需家庭收入的衍生变量。家庭收入类别信息可包括：工资、自营收入、利息和股息、社会保障和退休收入、汇款、特许权使用费和租金。如果还收集总收入，编辑期间应通过各类别的相加，来检查每个人的总收入。然后根据所记录的总收入来核对该总收入。如果合计收入不等于所填报的总收入，则编辑小组必须拟定纠正计划。要么改变总数，以反映各部分之和，要么改变一项或多项个人收入类别。在确定了所有个人的总收入时，可通过加总家庭内的个人收入，得出家庭收入。

7. 编辑小组必须考虑到，家庭内一人或多人因业务失败或其他原因而有负收入的情况。在这种情况下，家庭总收入将减少，而不是增加，减少或增加的金额相当于该人的收入。

3. 家庭核心

8. 对于住户构成，《人口和住房普查的原则和建议》第二次修订本为家庭核心（由下文中的一种类型组成）确定了代码，括号内为二次编码建议：

1 已婚夫妇（或基于自愿结合而生活在一起的夫妇），无子女（住户和配偶，或者联合户主，或者自愿结合而生活在一起的夫妇）

2 已婚夫妇（或基于自愿结合而生活在一起的夫妇），有一个或多个未婚子女（如上，但通过住户检索，或者对住户单元内未婚子女数的二次编码，至少有一名未婚子女）

3 一位父亲，有一个或多个未婚子女（男性户主，无妻子，如上所述，至少有一名未婚子女）

4 一位母亲，有一个或多个未婚子女（女性户主，无丈夫，如上所述，至少有一名未婚子女）

4. 住户类型

9. 《人口和住房普查的原则和建议》第二次修订本列入了各类住户的一般特征，以帮助确定住户构成的二次编码。各国可决定进行一个单一的二次编码，或一系列的二次编码，具体取决于数据的潜在用户。

10. 首个二次编码可确定住户类型（由以下项目表示），包括定义。下一节将列出所建议的二次编码。

1 一人住户

2 核心住户：单一家庭核心，即：有子女或无子女（一个或多个子女）的已婚配偶家庭或者自愿结合伴侣，或者有（一个或多个子女）的单亲父母

3 大家庭：单一家庭核心加上与户主有亲缘关系的其他人，两个或多个家庭核心，或者彼此有亲缘关系的两个或多个人员但不属于家庭核心

4 混合住户（其他类型的住户）

5. 住户构成

11. *单人住户*是住户但不是家庭，因此，在住户构成编码中，应作为单独类别列入。

12. *核心家庭住户*。核心家庭住户可分为（并且可接收以下类型的个别代码）：(1) 有子女的已婚家庭，(2) 无子女的已婚家庭，(3) 有子女的自愿结合伴侣，(4) 无子女的自愿结合伴侣，(5) 有子女的父亲，(6) 有子女的母亲。为确定适当的代码，首先要确定户主性别，然后检索住户的配偶和资料。核心住户类型可以为两位数代码，两位数的头一位为2（代码1用于单人住户）；因此，代码21将代表有子女的已婚夫妇家庭。

13. *大家庭住户*。大家庭也可分类，（根据以上代码分配）将包括：（31） 单一家庭核心加上与核心有亲缘关系的其他人，(32) 彼此有着亲缘关系的两个或多个家庭核心，无其他人，(33) 彼此有亲缘关系的两个或多个家庭核心加上与核心有亲缘关系的其他人，(34) 彼此有亲缘关系的两个或多个人员，但任何一个都不属于某个家庭核心。实际代码的确定，将需要检索住户的核心数量以及住户内人员之间的关系。如果住户已按核心编码，则不执行该程序。

14. *混合住户*。其他所有住户为混合住户。采用前面的方案，可得出：(41) 单一家庭核心加上其他人，其中某些人与核心有关，有些则无关，(42) 单一家庭核心加上其他人，其他人中的任何人都与核心无关，(43) 彼此有着亲缘关系的两个或多个家庭核心加上其他人，其中有些人至少与一个核心有着亲缘关系，有些人则与任何核心都无亲缘关系，(44) 彼此有着亲缘关系的两个或多个家庭核心加上其他人，其中的任何人都与任何核心无关，(45) 彼此没有亲缘关系的两个或多个家庭核心，有或没有任何其他人，(46) 彼此有亲缘关系的两个或多个人员（其中任何一个都不属于某个家庭核心），加上彼此没有亲缘关系的其他人，以及(47) 没有任何亲缘关系的人。同样，通过一系列的检索和概括，可为每类住户确定适当的代码。

6. 家庭构成

15. 家庭是住户的子集；因此，家庭构成的二次编码将包括适合于上述家庭的类别。一人住户不构成家庭，因此，将不列入家庭构成的二次编码中。同样，混合住户是住户但不是家庭，因此，也不列入。个别国家将需要确定其

是希望为所有家庭（为核心家庭和大家庭）列入一个单一的二次编码，还是为家庭核心和大家庭列入单独的二次编码，前提是这些二次编码不会重叠（不过有理由对所有家庭，将核心家庭住户与大家庭列在一起）。

7. 住户和家庭身份

16. 住户和家庭身份表示一个人与其他住户或家庭成员之间的亲缘关系。确定住户和家庭身份的方法不同于划分住户成员的传统方法，传统上只根据住户成员与户主和基准人的关系来进行划分。

17. 《人口和住房普查的原则和建议》第二次修订本为住户身份建议了一些编码方案。第一套代码指，在一个至少有一个家庭核心（即：同时也是家庭的住户）的住户中的人，根据建议，二次编码的决定因素有：

1.1 丈夫（男户主或男配偶）

1.2 妻子（女户主或女配偶）

1.3 自愿结合（同居）关系中的伙伴（根据关系代码（如果有）加以确定；或者结合关系代码和婚姻状况加以确定）

1.4 单亲母亲（根据女性没有丈夫，但有孩子予以确定）

1.5 单亲父亲（根据男性没有妻子，但有孩子予以确定）

1.6 与双亲住在一起的子女（户主的子女，且家中有双亲）

1.7 与单亲母亲住在一起的子女（户主有子女，但子女的父亲不在）

1.8 与单亲父亲住在一起的子女（户主有子女，但子女的母亲不在）

1.9 不属于家庭核心一员的人（任何其他亲属）。《人口和住房普查的原则和建议》第二次修订本将该项目分为两组：(1) 与亲属住在一起和(2) 与非亲属住在一起

18. 二次编码的第二组针对住户没有任何家庭核心的住户人员——独居；与其他亲属和/或非亲属住在一起，不包括户主的配偶或子女。类别如下：

2.1 独居的人员（单人住户）

2.2 与其他人住在一起的人员（在一个住户单元中生活的人员，没有户主的配偶或子女）。该类别又进一步分为(1) 与兄弟姐妹生活在一起的人员，(2) 与其他非兄弟姐妹亲属生活在一起的人员，或者(3) 与非亲属生活在一起的人员

19. 从这些类别中确定的应是一个单变量，因为它们是相互排斥的。变量为两位数。有些统计机构可能希望第一位数独立于第二位数；在这种情况

下，第一位数将说明住户是否为家庭核心，第二位数将确定一个人的住户身份类型。

20. 《人口和住房普查的原则和建议》第二次修订本还列入了按家庭身份对一个人的分类，包括（1） 一对“户主-配偶”中的男性或女性，有或没有子女，（2） 单亲——按性别划分，（3） 户主的子女，已婚夫妇的子女或者单亲父母的子女——按父母的性别划分，（4） 非家庭核心的成员（有亲属关系或无亲属关系，如果有亲属关系，列出属于什么关系）。以上用以确定住户身份的因素可用来确定家庭身份。

8. 艾滋病毒/艾滋病对住户结构的影响

21. 鉴于艾滋病毒/艾滋病传播对很多国家住户结构的影响，二次编码有助于描述各种住房单元。例如，通过二次编码描述缺代住户（只有祖父母和孙子）、户主年龄不足18岁的住户、户主丧夫的住户等，可对这种传播的社会经济影响进行间接评估。劳动力队伍之内和之外的儿童、这些住户内劳动力的结构等，可帮助政府全面描述艾滋病毒/艾滋病的影响。

9. 亲属

22. 亲属指那些与户主有着某种亲缘关系的人，其衍生变量是与户主有亲缘关系的所有人之和。在大量没有亲缘关系的人一起生活在住房单元内时，该值尤其重要。当很多没有亲缘关系的人员以这种方式生活在一起时，常常将他们归类为住在“集体住所”或“集团住所”。

23. 在开发数据集时，国家统计机构通常为那些按年龄分列的不同亲属类别确定衍生变量。例如，可以为有亲属关系的0-5岁、5-17岁、6-17岁、0-17岁的儿童，以及65岁及以上的亲属、75岁及以上的亲属确定衍生变量。

24. 例如，家庭内“有亲属关系的儿童”包括：户主自己的子女，以及住户内与户主有亲属关系、在18岁以下、不管婚姻状况如何的其他人，不包括户主的配偶。有亲属关系的儿童可包括或不包括收养的儿童，因为他们与户主没有亲属关系，但这种决定要取决于国家的情况。

10. 家庭中的工作者

25. 有时，国家希望按照工作者的人数来比较住户变量，例如按住户大小和每个被赡养者的工作者数量开列的收入分布情况。国家可为家庭内的工作者数量确定衍生变量，即：对那些在所涉期内——如，一周或一年（一个日历年或最近12个月）——至少工作了1小时的人员相加。国家可采用“上周”工作的人员数量，前提是只对该期间收集数据。

11. 全套管道设施

26. 普查调查表上的有些项目用来获取管道设施方面的数据。这些项目通常与有无自来水、抽水马桶、浴缸或淋浴有关，通常在被住用和空置住房

单元中获得。全套管道设施的衍生变量可有助于比较不同地区或群体之间在某个时间点或不同时期的社会经济条件。例如，在单元内或单元所在建筑物外存在三种设施时——自来水（热水或冷水）、抽水马桶以及浴缸或淋浴，可以为全套管道设施获取衍生变量。编辑小组需要为全套管道设施确定最合适的一套变量。

27. 在本例中，如果三个项目分别设问，可获取衍生变量，在编辑期间，将对存在的所有三个项目求和。如果住房单元有自来水、一个抽水马桶和一个浴缸或淋浴设施，则“具有全套管道设施”。如果不同时具有三个项目，则“缺乏全套管道设施”。

12. 全套厨房设备

28. 通过普查，可从调查表有关烹饪设备、冰箱和水槽的项目中，获得厨房设备数据。这些项目同时针对被住用住房单元和空置住房单元收集。如果在查点住所时，同一建筑物内存在烹饪设备（电、煤油或燃气炉、微波炉和非便携式烤箱或烹调用炉）、一个冰箱和一个带有自来水的水槽，则可以认为该单元具有“全套厨房设备”。这些设备不必位于同一房间。

29. 在分别提问上述三个项目时，可获取衍生变量，在编辑期间，将对存在的所有三个项目求和。“缺乏全套厨房设备”包括：在所列三项厨房设备全都具备时，设备位于另一建筑物内的情况；只有部分而非全套设备；在查点住所时，同一建筑物内所列三套厨房设备都不具备。

13. 毛租金

30. 国家可能会收集有关现金或合约租金的数据。现金租金通常不包括公用设施的费用。有时，国家还需要关于毛租金的资料。毛租金等于现金或合约租金，加上由出租人支付的公用设施（水、电和燃气）和燃料（包括石油、煤、煤油和木材）月均估计费用。在将公用设施和燃料列入租金费用方面，存在着不同做法，毛租金可消除由此产生的差异。对于不支付现金租金而住用的出租单元，可在制表时，作为“非现金租金”单列。

31. 如果所付租金和公用设施费是单独收取的，则毛租金的衍生变量等于这两者之和。

14. 财富指数

32. 财富指数是计量一国或一国部分地区福利状况的尺度。在大多数情况下，指数采用住户资产构建。通常情况下，要通过因素分析来取得最佳组合的项目及其变体。通常为项目赋予二进制值——1为“有”0为“无”，然后对值加总。值越大，财富越多。例如，有一台电视机的编码为“1”，没有时则为“0”。另一方面，对于厕所，“户外”厕所的编码为1，“重力冲水式”厕所为2，“冲水式”厕所为3（涉及三组二进制变量）。在加总时，可对各项目进行加权。

33. 然后，取财富指数各值分布的每五分之一，便可得出五分位数。最小的五分位数包括最穷的住户，最大的五分位数将包括最富有的住户。

B. 人口记录的衍生变量

1. 经济活动状况

34. 经济活动状况的衍生变量对于制表非常有用，但需要数个变量的资料。在《人口和住房普查的原则和建议》第二次修订本的以下类别中，需要重新配置若干变量。衍生变量可包括两个类别，并总共细分为六个子类：

1 经济活动人口

1.1 就业者

1.2 失业者

2 非经济活动人口

2.1 学生

2.2 操持家务者

2.3 养老金或资本收入领取者

2.4 其他

35. 在很多相关表中，要采用有关经济活动的各种分类，因此，编辑小组应考虑在数据记录中插入衍生变量，而不是由数据处理人员在制表期间对经济状况进行再分类。制表期间的再分类可能会引入误差，因为不同数据处理人员可能会按略微不同的方式进行再分类；及时采用一个程序，也可能会有不同的再分类，具体取决于编辑或制表的特定要求。经济特征方面的专家应对衍生变量进行设定。

2. 亲生子女

36. 有时，国家希望取得关于“亲生子女”的资料，即：户主和/或配偶在生物学意义的子女。表中所列的“亲生子女”可进一步分为与双亲生活在一起的子女，或与单亲父母生活在一起的子女。

37. “亲生子女”的衍生变量可按照编辑小组选择的定义，通过加总特定人员（通常为女性）的亲生子女数来获得。有时，用户对“亲生子女”项目提供按年龄划分的更详细资料。例如在美国，针对那些不足6岁，以及介于6-17岁之间的亲生子女数确定衍生变量。这些值列在所有女性的记录中。有关资料专门用来确定那些有子女的女性劳动力的特征。

3. 同住父母

38. 这些数据目的是为了考察单亲家庭子女在与那些与双亲同住的子女相比时，所具有的特征。在编辑时，首先利用关系代码确定特定人员与多少个

父母同住，然后获得衍生变量。该程序首先查看每个子女的关系代码，然后将该资料与小家庭资料结合在一起，以确定有多少父母住在住房单元内。

4. 目前在校年级

39. 有些国家询问两个有关教育的问题：

(a) 有关人员目前是否上学；

(b) 最高受教育程度。

40. 在这些国家中，编辑小组通常发现，如果一个人在查点时实际上在上学，那么上述两个项目之间会有不匹配的情况。有时，一个人的最高受教育程度可能会比其目前在校年级低一年。该人如果处于一系列年级或教育级别的中间阶段，统计数据将不会受到影响。但是如果正在特定教育级别系列中读一年级，则可能无法与其他来源的数据进行匹配。例如，一个读一年级的人员将被记录为在校，但在受教育程度项目中却没有记录。同样，一个即将升入中学的人员将记录为在校，但其受教育程度却记录为小学的最高年级（或教育级别）。

41. 可为这两个项目的组合，确定一个被称为“目前在校年级”的衍生变量。一个人目前如果没有上学，则代码将与最高受教育程度的代码相同，如果正在上学，则在编辑时，对受教育程度的年级（教育级别）加1，然后分配到“目前在校年级”中。

42. 有些国家询问三个有关教育的问题，即：上述两个项目，以及第三个关于是否完成最高年级的项目。如果也获得了该信息，则应同时用来确定“目前在校年级”。

5. 上次分娩以来的月数

43. 如果要收集有关上次分娩日期的资料，可通过二次编码，来间接估计按年龄划分的逐年总生育率。二次编码采用查点日期（通常为月份和年份）和上次分娩日期，将两者日期全部换算为月份，然后相减，得出上次分娩以来的月数。该数据存在妇女记录中，以帮助确定逐年生育率估计数据。

附　件　二

调查表格式与键入的关系

1.　在普查或调查中，人口项目的两个最常见调查表格式是个人页和住户页。

2.　个人页包括一、两张人口信息对页，每人单有一页。这种方法很有用，因为一个人的所有信息都出现在一页上，可方便收集。另外，这种格式还可为查点期间进行内部一致性检查提供方便。个人页可装订成一个小册子，以便于现场处理（见图A. II. 1）。

图A.II.1

调查表个人页示例

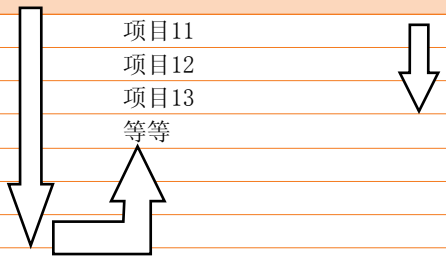
第X个人的个人页		第X+1个人的个人页	
项目1	项目10	项目1	项目10
项目2	项目11	项目2	项目11
.	.	.	.
.	.	.	.
.	.	.	.

3.　个人页上各项目的编码和键盘输入基本上为机械操作，不能指望编码员/数据输入员对所提供信息的有效性进行评价，但可要求他们找准代码或按键。图A. II. 2说明了为特定个人记录在单页上的信息流。为该人键入列在单页上的数据，要比键入列在若干页上的数据更容易。有效性将在随后的计算机编辑阶段进行检查。

图A.II.2

调查表个人页内信息流示例

个人页	
项目1	项目11
项目2	项目12
项目3	项目13
等等	等等
•	
•	
•	



4.　住户页在一页上列出一个住户的所有信息，如果可能，则在若干页上列出一个住户的所有信息，其中在每一页上列出所有住户成员。按照这种方式

列出住户成员很有用，因为不必为每个人打印调查表项目，从而节省了空间。此外，查点员可以在收集数据时，对各住户成员之间的填补内容进行比较。

图A.II.3
将所有人列在同一页上的调查表住户页示例

住　户　页					
人员编号	项目1	项目2	项目3	项目4	等等
1					
2					
3					
4					
5					
.					
.					
.					

5. 第三种方法是对每个人采用单独的表格，由查点员在查点期间或查点后装订成活页小册子。这种方法的有效性在于，查点员只按住户所需的准确数目收集数据表（页）。缺点是，在转移或进行其他处理时，这些表可能是分开的，如果普查部门不能将同一住户的表重新汇集在一起，则可能会在编辑和覆盖面方面带来很多问题。

6. 调查表页面大小也是需要考虑的一个因素，不仅是为了查点，而且也是为了键入。在编码和键入期间，文件必须平放在工作台上，编码员或数据输入员必须能方便地找到表内项目的位置，然后进行处理。

7. 同样，当所有信息都列在一页上时，工作人员键入住户页上的数据更容易，也显然更快，因为数据输入员不必翻页。图A. II. 4显示的是住户页上的信息流。

图A.II.4
每页列有多个人的调查表住户页内信息流示例

住　户　页				
人员编号	项目1	项目2	项目3	等等
1	⇨			⇩
2	⇨			⇩
3	⇨			⇩
4				
5				
•				
•				
•				

8. 如果键入的人口或住房信息超过1页，则可能会出现问题。为解决这类问题，国家统计机构很可能会采取下述两种方法之一。

9. 一次输入一个人的信息。数据输入员可键入若干页中第一页上一个人的信息行，然后转到第二页和随后各页。在输完第一个人的各页之后，数据输入员再返回该住户所在住户页的第一页，键入第二个人、第三个人……的信息。只要数据输入员在整个键入过程中不串行，这种键入方法就可行。在将一个人的项目错误键入另一人的信息行时，可以通过计算机编辑程序加以解决，但这种程序本身很难编写。

10. 可以一次输入一页数据。数据输入员可键入整页的信息，然后转到下页。在这种情况下，数据输入员键入第一页上的所有信息，而不管所涉及的人有多少，然后翻到下页，键入有关所有人的下一部分信息。可以包括或不包括跳转模式，具体取决于键入类型（带计算机编辑或不带有计算机编辑）。在很多情况下，在计算机编辑期间，来自各组键入数据的记录都必须组合在一起，个人编号的任何错误键入都必须加以处理。

11. 在以下例子（图A. II. 5）中，住户的人口信息没有带来不同寻常的键盘输入问题，因为普查为所有人的所有项目获取了答复。

图A.II.5

涉及多人的住户页示例，无键盘输入问题

住 户 页				
人员编号	关系	性别	年龄	等等
1	户主	男	40	
2	配偶	女	35	
3	孩子	女	18	
4	孩子	男	12	
5	兄弟姐妹	男	35	
6	配偶的兄弟姐妹	女	30	
7	兄弟姐妹的孩子	男	5	
8	兄弟姐妹的孩子	女	3	
等等				

12. 但同一住户的第二页（图A. II. 6）可能会带来某些键盘输入问题。例如，如果决定只对5岁及以上人员收集语言使用数据，那么第八个人（年龄为3岁）的记录将为空白。数据输入员应为这个孩子留出空白单元，否则就得在以后设法通过计算机编辑来纠正。

13. 同样，其他项目也应空白，如：低于最低劳动力参与年龄的人员、低于最低生育年龄的人员以及所有男性的生育率项目。在图A. II. 6中，数据输入员可能会将编号为6的人员关于平均生育数的信息（本例中，为4个子女）错

误键入到编号为5的人员项下。之后，计算机编辑将会删除该男性的生育率项目，然后为女性插补该女性的生育率，但插补的值可能不正确。

图A.II.6

涉及多人的住户页示例，可能有键盘输入问题

住户页 2				
人员编号	语言	劳动力	平均生育数	等等
1	语言1	是		
2	语言1	否	3	
3	语言1	否	0	
4	语言1			
5	语言1	是		
6	语言1	否	4	
7	语言1			
8				
等等				

14. 由于费用或空间的限制，一国必须多次使用住户表。但在人口较少时，或者一国有能力负担额外费用时，个人页表格由于键盘输入引起的匹配错误很可能要少于住户表的错误。

附　件　三
扫描与键盘输入

1.　很多国家正在使用扫描设备——光学标记识别（OMR）或光学字符识别。在作业顺畅有效且费用不高时，这些都比键盘输入有优势。但很多国家，甚至一些努力采用扫描的国家，可能无法负担最初的启动费用，或查点期间和查点后的持续维护费用。从积极的方面来说，很多国家把为普查准备的扫描仪用于许多目的，包括其他调查，以及诸如输入和输出表之类的行政记录。各国还可以将扫描工作外包，或者在普查需要时租用扫描仪。

2.　键盘输入的优点在于，键入过程中学习的技能可用于国家统计/普查机构和其他政府机构的其他活动中。在普查培养了专门的数据输入员之后，这些数据输入员可以为各类后续调查键入数据。这些调查可包括：事后质量抽样调查和其他调查，如：生育率或住户收入和支出调查。工作人员还可键入行政记录，如：生命记录以及有关贸易、迁入/迁出移民和海关方面的记录。

A.　输入数据

1.　扫描

3.　尽管修改可能要依靠嵌入系统内的跳转模式，但采用光学或其他扫描设备捕获数据的国家，通常不能在捕获过程中对数据进行校正。但决定通过键盘输入其数据的国家，可有几项选择，具体取决于它们需要键入数据的速度有多快和需要多少人工检查。每项选择取决于编辑小组的要求、数据输入员的技能和编辑程序的先进程度。

4.　所需数据输入设备的数量和类型取决于所选择的数据捕获方法、可用于该普查阶段的时间、国家大小、数据捕获活动的分散程度以及其他因素。就键盘数据输入而言，平均输入速度通常为每小时键击5 000-10 000次。有些操作员的速度远低于这一水平，而其他操作员则大大超过这一水平。影响操作员速度的因素有：(a) 辅助软件和程序；(b) 操作员任务的复杂程度；(c) 设备的工效特点、可靠性和速度；(d) 是否一直有工作的问题；(e) 受聘工作人员的培训情况和悟性；(f) 工作人员的积极性（联合国，2007年，第1.193段）。

2.　埋头键入

5.　埋头键入有两种形式。第一种是，遇到什么数据项目就键入什么数据项目，无任何跳转模式。在这种情况下，键入速度更快，因为数据输入员在遇到无效不一致信息时，不必停顿。这样还可以更准确，因为键入员的工作更机

械。第二种埋头键入形式需要停下来检查问卷的无效或不一致结果，这样速度会更慢，同时还需要键入人员具有更多的专业知识。必须认真考虑速度方面的较高代价。让人费解的是，如果数据输入员发现数据记录正确但却编码错误，那么这种方法还可提高准确度。有时错误键入本身也可能会立即受到质疑，因为编辑包可提供自动检查。

(a) 无跳转模式的埋头键入

6. 在所有登入项目都被键入或通过人工跳过时，可保持一定的节律，某些跳转模式将不会回避那些有效但暂时不一致的信息。例如，如果一个人被记录为男性，多数编辑小组将会要求跳过有关生育率的整个板块。在这种情况下，数据输入员将击键跳过（利用空格键或箭头跳过男性或年幼女性的记录），因为所有字段将是空的。但这需要时间，空格可能不是完全正确。例如，数据输入员输入的空格可能太多或太少，其他项目可能因为不能对齐而可能被错误键入。如果所有字段都按这种方式键入，那么可在没有跳转模式时，键入该信息。例如，当数据输入员遇到育龄成年女性时（已对其收集和编码“平均生育数”、“存活子女”或上年“生育子女”等项目的女性），键入所有项目。如果键入生育信息，则计算机编辑程序可确定哪个或哪组项目是有效的，以及哪些是必须修改的。当编辑程序确定该人为成年女性，但生育信息为空白，则必须采用动态插补或其他适当的手段，来获取用于制表的生育信息。如果因为跳转模式丢失了实际信息，则编辑小组必须确定这种丢失与提高的效率和速度相比，是否值得。如果有跳转模式，数据输入员仍可将屏幕退回到适当的位置进行纠正。尽管数据输入员在利用空格键跳过无需键入的项目时，会浪费时间，但采用这种数据输入形式，可在编辑期间而非键入时处理性别、年龄和生育项目之间的不一致问题。

(b) 有跳转模式的埋头键入

7. 第二种埋头键入方法要采用有跳转模式的键入。同样，如果编辑小组要求跳转模式（通常是为了反映查点员收集数据的方式），而跳转模式容易采用且数据输入员能很快学会键入模式，则会使键入更加容易和快捷。如果跳转模式非常复杂，那么数据输入员会被弄糊涂，总是键入错误的位置。如果采用的有限模式涵盖了被键入的大部分记录，那么带有跳转模式的键入方式将最有效。

8. 编辑小组需要为其国家的普查或调查确定适当的跳转模式。例如，对于儿童，也就是低于国家所定潜在就业年龄的人员，跳过所有就业项目是有道理的。通常情况下，这些要占人口项目的一半，因此对儿童跳过这些项目，可提高效率，但有些特殊情况除外，例如年龄处于边界线的儿童，或者国家对童工问题感兴趣。

9. 编辑小组将按项目逐个确定各年龄组所需列入的项目。工作人员可将项目编组，以方便跳转模式的管理。

10. 为跳转模式做出明确决定，并不总是很容易。例如，试考虑以下顺序：

1 此人的公民身份如何?
- 在本国出生(跳到项目3)
- 入籍公民
- 非公民

2 此人是哪年入境的?

3 下一个项目

11. 对于在本国出生的人，可以通过跳转模式，从1跳到3，即：跳过“入境年份”。但在有时，由于查点员或编码员犯了错误，或者键盘输入出错，数据输入员会违反跳转模式。这涉及很多因素，其中包括数据输入员的技能水平、文化背景、调查表布局和屏幕布局。编辑小组通常一起讨论，确定这种情况下的跳转模式是否合理。

3. 交互式键入

12. 交互式键入可在普查输入阶段使用，但更适合于调查，尤其是项目分配会影响调查结果的小型调查。交互式键入可能需要手动或自动纠错，具体取决于可用来修改或纠正的信息而定。

13. 试分析小型调查的情况。对于小型调查来说，每个答复都很重要。例如，如果一国进行1%的抽样调查，每个答复将代表100个人员、住房单元或农业生产经营单位。少数的无效或不一致情况可能会给调查结果带来很大的影响。在这种情况下，人口统计学家和其他社会科学家通常希望对处理工作实施大量的控制。

14. 可按几种方式进行控制。人口统计学家和其他专家可自己键入数据，并在键入过程中，利用数据收集表中记录的信息，检查不相关的、无效的或不一致的答复。通常情况下，通过直接查看所收集的信息，他们可立即解决相互矛盾的问题、错码或其他不一致的问题。有时，可决定将不完整或无效调查表返回现场。这种交互式键入能产生最佳结果，因为人口统计学家同时作为数据输入员，但到目前为止，这也是最昂贵的方式，没有多少国家能够负担得起。

15. 编辑小组可指定非常详细的编辑规则，确定键入期间出现特殊情况时，数据输入员必须怎么做。他们可以决定，数据输入员对于每个未解决的无效代码，应键入什么代码。编辑小组可解决详细规则没有涉及的情况，并可修改规则（这可能会使键入的第一部分和第二部分出现不一致的问题）。

16. 在埋头键入中发挥重要作用的跳转模式，在这种情况下也很重要。如埋头键入一样，数据输入员必须了解和学习所用的任何跳转模式。如上所

述，跳转模式可提高键入速度，但通常会带来一些质量损失。对于交互式键入，一种经验性常规是，跳转越少，质量越高。

17. 在拟定键入指令后，不管是否采用埋头键入法，国家统计/普查机构都必须让实际的数据输入员测试键入指令，然后再进行实际操作。在测试键入指令时，可以排除系统中的缺陷，从而获得最佳键入效果。

B. 核实

18. 国家统计/普查机构还必须决定适合采用哪一级的核实。对于键入的数据，很多专家建议进行100%的核实。在这种情况下，所有项目都重新键入（在现有信息基础上键入），以确保所收集的数据就是输入机器供计算机处理的数据。但在通常情况下，全面核实不切合实际，因为国家没有时间、财力或人力重新键入所有数据。对键盘输入新手，所核实的样本比例应较大，而对于有经验的老手，则应较小。另外，如果测出的键入出错率很低，数据输入员的出错非常少，则很可能没有必要进行全面核实。

19. 在任何核实活动中，首先都需要确定需要什么信息。国家是希望对个别键盘输入员进行跟踪，还是对一组输入员进行跟踪？是否希望确定，是采用现学输入技能的做法，还是采用保持已有技能的做法？为确定工作流量和获得的技能，控制单位也很重要，包括日报告、周报告或月报告。

20. 最后，非常重要的是，核实应具有独立性，同时要由不同组的输入员，或者至少由同一组的不同成员，从数据录入中进行核实。利用不同组的输入员能确保活动的独立性，从而获得更好的结果。

21. 对于扫描数据，有些国家也进行核实，以确保扫描的全面和完整。由于扫描技术仍然很新，所以即使在利用试点调查或预备调查数据对系统进行了全面测试时，纸张质量变化、实际表格在不同地方的印刷、存储等问题也可能会带来问题，这些问题需要通过核实加以解决。

22. 如果错误是有规律的，那么可以通过编辑程序删除，键入员和核实员不应对纠正问题做出判断，但应负责发现错误。这些错误可能与扫描设备的测试不当有关，这种不当会给某些项目或项目组合带来规律性错误，会混淆某些数位的读取（例如，互换2s和3s，或者8s和9s），还会误读连续的勾选框，等等。

23. 误读连续勾选框一直是最近几年中的一个问题，有时只能在编辑期间加以解决。如果表格不是相邻的，将需要采取其他办法，很可能需要在结构编辑期间对问题进行解决。如前所述，首先需要创建一个很好的结构化文件，然后才能开始内容编辑。

1. 依附式核实

24. 核实方法要么为依附式，要么为独立式。在依附式核实中，数据输入员要在其他工作人员以前键入的数据之上，来键入数据。如果键击不同，软

件包会提示数据输入员，而数据输入员将根据所用程序，要么推翻以前的数据，要么对不一致情况加以注明。键入数据来自原始调查表，因此数据输入员通常可以自己就原有键入是否有错，做出适当的决定。

2. 独立式核实

25. 在独立式核实中，数据输入员从零开始重新键入数据；他们可利用原始调查表，建立一个完全独立的键入数据文件。然后使用计算机程序对原有键入数据集和核实数据集这两个文件进行比较，以测试不一致情况。可能需要一些人工操作，来纠正无效和不一致的键击问题。

C. 扫描数据编辑的考虑因素

26. 目前，越来越多的国家对数据进行扫描。2000年代初，其中很多国家吃惊地发现，扫描带来的错误类型不同于键入错误。编辑扫描数据的部分问题是，在扫描过程中缺乏质量控制。在2000年代之初，该技术很新，很多统计部门都不具备必要的背景或设施，来为所有项目开发适当的质量控制程序。没有开发出适当质量控制程序的很多国家，最终都是对所有项目不开发该程序；因此，问题末尾的某些项目——尤其是生育项目，结果成了无效或不一致的项目。

27. 当然，在键入数据中发现的很多不一致问题，也存在于扫描数据中。但讨论扫描数据使用中出现的某些特殊问题，会有一定的用处。可以扫描的调查表要求做标记的人员为机器阅读提供帮助，因此，项目的列示方式可能会在数据收集期间给查点员和调查对象带来各种问题。与这些项目有关的问题必须有系统地加以解决。在有关项目与其他项目密切相关时（如，宗教与种族密切相关），可采用文中介绍的常规编辑做法。

28. 但当用于规划和政策的项目可能会带来问题时，必须谨慎。通常情况下，“性别”项目不会带来问题，因为只有两种可能。但如上所述，尽管键入员通常只需要键入1或2时（或“未知”代码时），但性别栏可能会出现任何值——如由于出现其他位数、字母字符或其他字符。因此，除了对键入数据进行的编辑外，还需要补充某些类型的编辑，以考虑到这些杂项值。

29. 关系代码可很好地说明这一问题。如果关系代码为一位数（如文中所述），则通常不会出现问题。但如果采用两位数，那么当第一位数编码错误或由扫描仪错误读取时，有时会在扫描时发生问题。通常情况下，如果采用1-12的代码，键入员只限于键入这些代码，输入程序报将会发觉所输入的非法代码。在扫描的情况下（尽管可通过扫描程序包发现问题），但几乎一切都会被接受。在这种情况下，错误代码必须在编辑期间进行修改；否则，将会在制表阶段带来各种问题。

30. “年龄”有时也是一个问题，在采用三栏（允许人们的年龄大于100岁）时，尤其如此；因此，为达到适当编辑的目的，可能需要按位数逐个

分析——需要对一位数、十位数和百位数进行单独查看。一旦确定年龄捕获正确，则可采用常规编辑。

31. 但如果“年龄”和“出生日”同时存在，同时一个项目优先于其他项目时，则误导性信息可能会带来问题。通常情况下，主题专家喜欢采用出生日期和普查/调查参照日，来（用减法）计算确切的年龄，以便与填报年龄进行比较。在缺失一个或数个位数时，必须认真确保剩下的所有位数得到适当采用，以便对用于比较的年龄做出最佳估计。如果扫描没有读取相应的位数（如，一个单一的位数），那么编辑时应考虑到这一点，以便对原有的位数做出最佳推测。键入时，通常不会发生这类问题。

32. 2000年代初，扫描中问题最大的项目是“生育”——包括平均生育数、子女存活数、头年或头几年平均生育数。就大多数国家而言，主要问题是在扫描期间缺乏质量控制，这使数据捕获中会出现奇怪的项目。例如，在给定国家的死亡女童项目值为17、18或19时，数据如果不经过编辑，将不能用于规划之目的。

33. 在扫描数据中，“死亡率”信息也会带来问题。对于键入数据，如果在普查前的年份中，有一系列关于死亡的项目（如，死者的性别和年龄，不管此人是自然死亡，还是产妇死亡），那么甚至可以对擦除和勾销项目进行键入。但对于扫描数据，通常不能读取擦除项目，扫描仪会将这种项目视为空白，然后继续捕获。编辑程序必须将该信息转移到适当的空间中，以用于制表和后续分析。还应注意的是，较新的扫描作业可以在捕获期间和之后进行这些转移。

D. 结论

34. 遗憾的是，每个国家的问题取决于个别扫描仪的具体程序和功能，因而难以制定出通用的指导方针。但到目前为止，在处理的所有情况中，扫描问题一直是有规可循，即：当工作人员确定了用以减轻问题的运算法则时，可获得全面编辑的数据集。

附　件　四
流程图示例

1.　编辑小组的任务之一是为编辑过程中使用的变量确定关系结构。流程图可方便确定变量之间的各种联系，有助于进行明确简练的编辑设定。对相关联系的这些设定有助于主题和数据处理专家使编辑过程形象化，方便两个群体之间的交流。

2.　以下各页列出了三个流程图示例：

(a)　用来确定户主的流程图；

(b)　用来确定住户内是否有配偶的流程图；

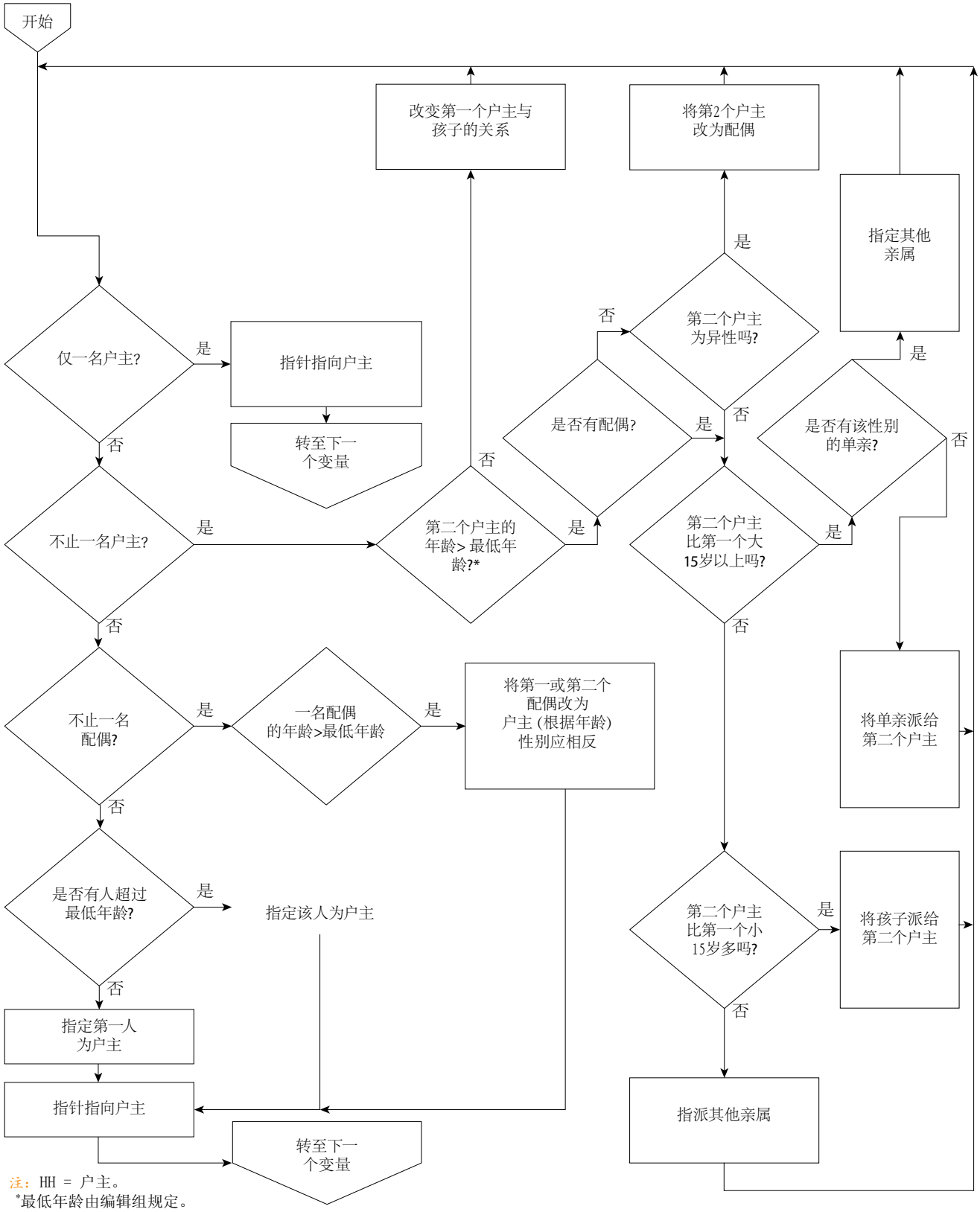
(c)　用来编辑户主和配偶性别变量的流程图。

这些流程图示例仅用于说明，因而应按说明对待。编辑小组可在必要时，根据国情对示例作进一步的修改。

3.　应为普查的每个变量确定编辑流程图。流程图应由编辑小组共同编制，数据处理专家应将它们与编辑规范一起使用，以开发用以编辑普查数据的计算机程序。流程图和编辑规范应适当制成文档，以便用于未来的普查和调查数据处理。

图A.IV.1

用来确定户主的流程图示例

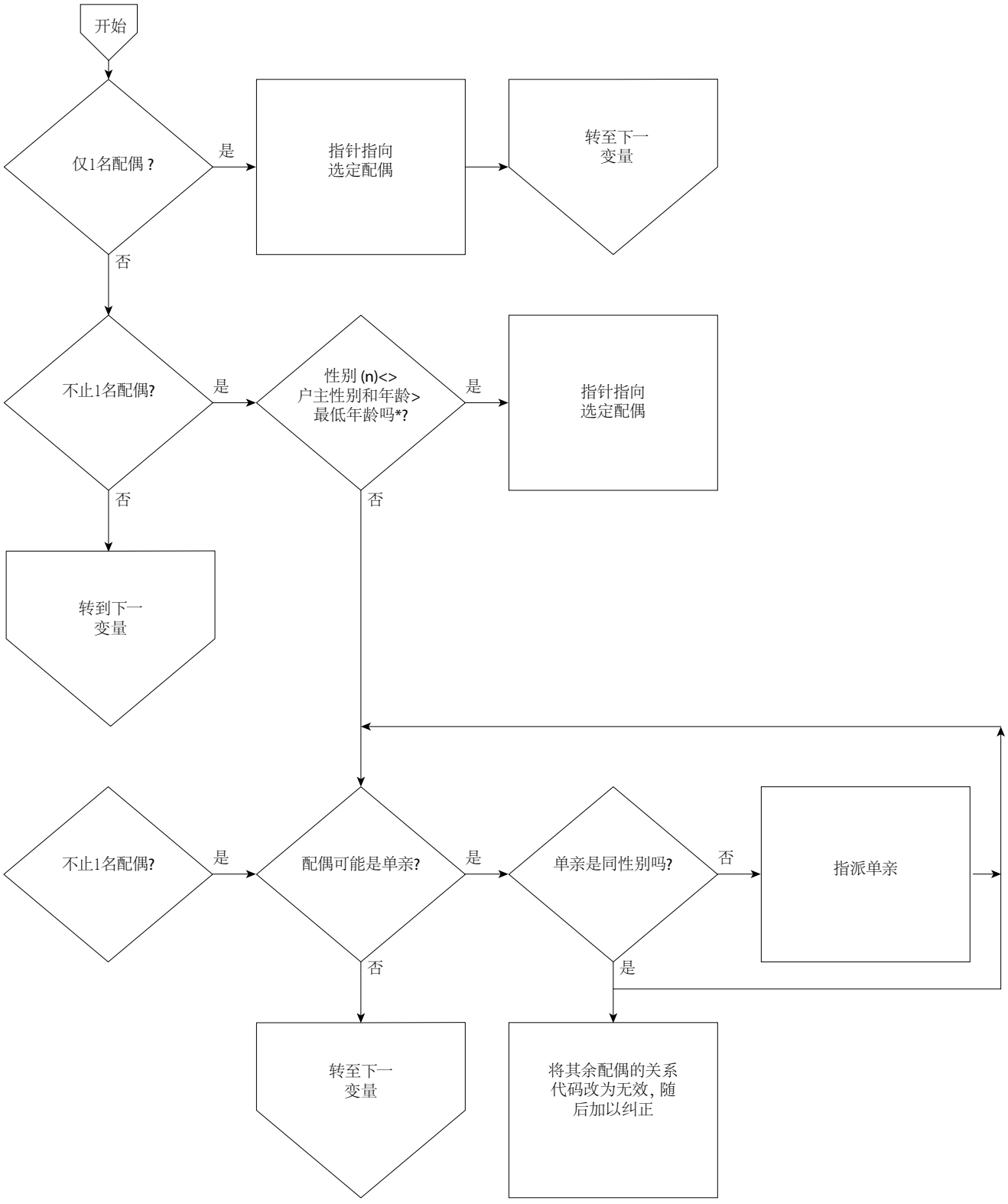


注：HH = 户主。

*最低年龄由编辑组规定。

图A.IV.2

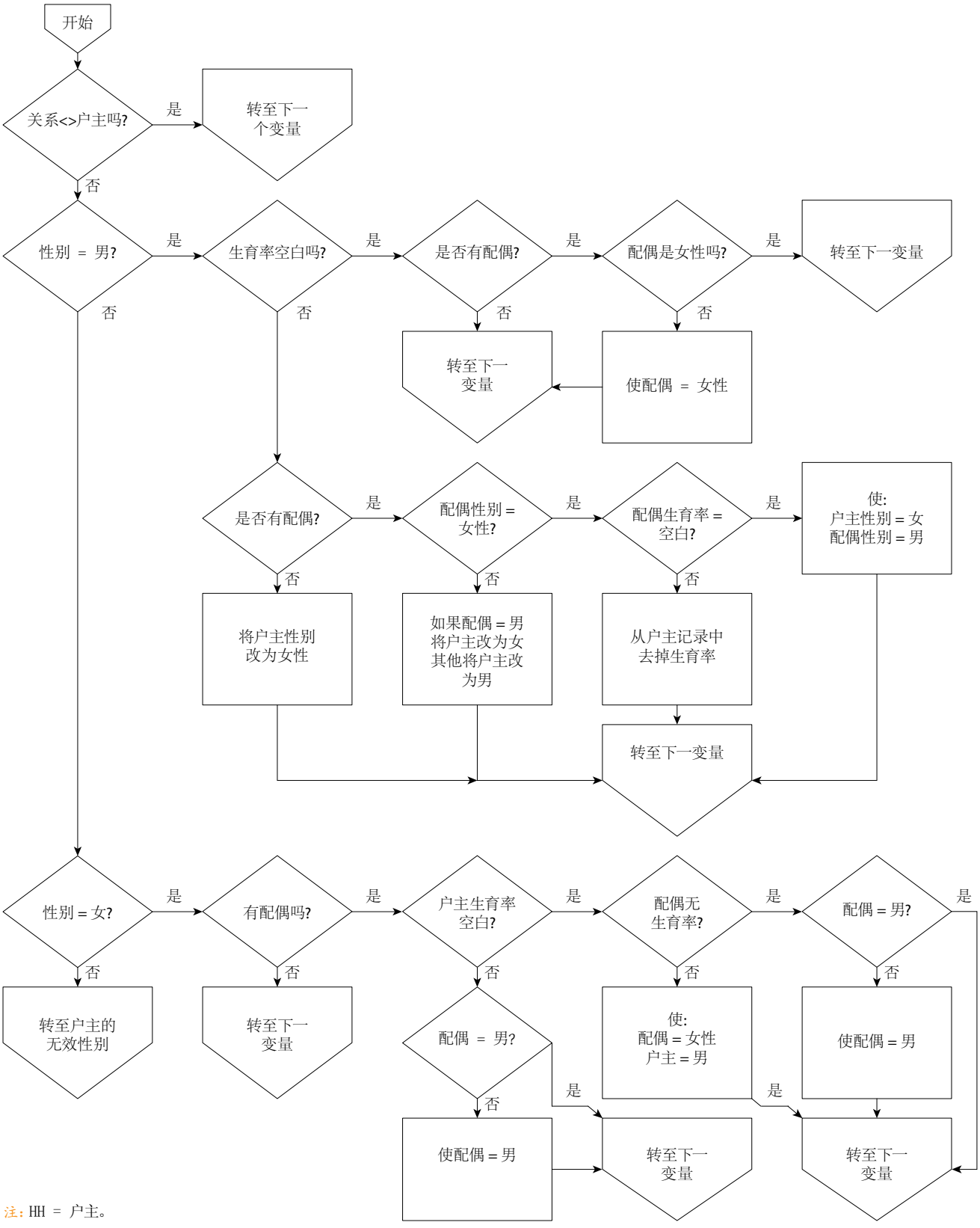
用来确定住户内是否有配偶的流程图示例



注：HH = 户主。
* 最低年龄由编辑组规定。

图A.IV.3

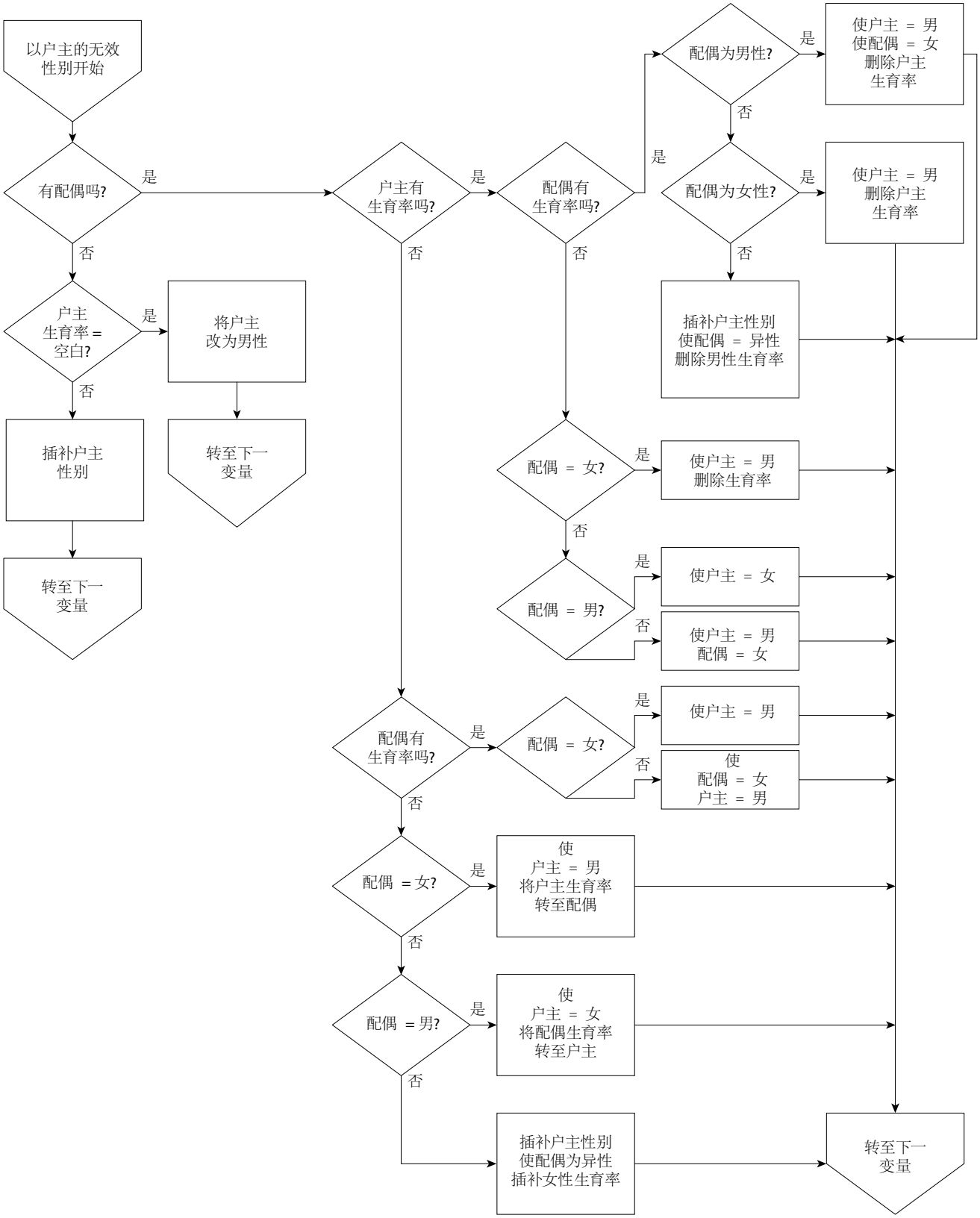
用来编辑户主和配偶性别变量的流程图示例



注：HH = 户主。

图A.IV.3 (续)

用来编辑户主和配偶性别变量的流程图示例



附　件　五
插补方法

1.　推算方法有很多。下文介绍的多数方法曾在Kalton和Kasprzyk（1982年，1986年）；Sande（1982年）；以及Särndal、Swensson和Wretman（1992年）的论文中有过论述。

2.　插补方法可分为随机插补或确然插补，具体取决于插补数据中的随机程度。

3.　**确然插补法**包括：演绎插补；基于模型的插补法，如：均值插补和回归插补；以及（适当情况下的）最近邻插补。

4.　**演绎插补法**允许缺失值或不一致值的推算具有确定性。通常情况下，这要以调查表中其他项目的答复模式为依据。

5.　更一般而言，插补法必须将一个不确定是否为真值的值替换掉。以下几段将简单介绍某些常见的插补办法。

6.　除了单一供体动态插补法外，下述几种方法将一次插补一个项目。这样，在每个插补类别中，记录中的项目按顺序逐个加以考虑。通常情况下，需要考虑的只是那些与所涉项目相关或与少数密切关联变量相关的显性编辑规则。由于可能有显性或隐性编辑规则，来将所涉项目与那些将在后续过程中加以考虑的其他项目联系起来，所以该方法可能会产生这样一种插补值：一方面能够通过目前所用编辑规则的检验；另一方面，又不能通过那些将在后续过程中加以考虑的其他编辑规则的检验。只有当一整套编辑规则，包括所有隐性编辑规则，都得到考虑时，才能确保插补值能够通过所有编辑规则的检验。隐性编辑规则是指，从逻辑上将两个或多个显性编辑规则结合起来时，而推导出的编辑规则。

7.　在以下说明中，“已通过编辑规则检验的记录”指，在利用与所涉项目有关的所有编辑规则进行检查时，都获得通过的记录。“未通过编辑规则检验的记录”则指，在利用与所涉项目有关的编辑规则进行检查时，至少未通过其中一项检查的记录。

8.　**整体均值插补**采用“已通过编辑规则检验的记录”的项目均值，来填充所有“未通过编辑规则检验的记录”的缺失项目或不一致项目。这种方法可产生合理的点估计数据，但是如果利用标准方差估计式来计算方差估计数据，该方法就不那么具有吸引力了。除非插补率很低，或者修改后的方差估计式对所用的插补进行了考虑，否则方差估计数据可能会被严重低估。

9. **分类均值插补**采用界定的插补类别，以产生具有相似性的记录组别。在每一类别内，采用“已通过编辑规则检验的记录”的项目均值，来填充所有“未通过编辑规则检验的记录”的缺失项目或不一致项目。这与整体均值插补非常类似，但其对分布的影响，以及方差估计的问题很可能不那么严重。

10. **回归插补**，或者更普遍的**基于模型的插补**采用“已通过编辑规则检验的记录”中的数据，根据一组预测变量，来对需要插补的变量进行回归。回归中的预测变量可以是调查表的项目或者辅助变量。然后利用回归方程，来为缺失或不一致项目的值进行插补。在基于模型的插补中，这是一个特例。该方法通常用于商业调查中的连续变量，在商业调查应用中，以前场合的数据通常能较好地预测目前场合下的值，并能取得令人满意的结果。

11. **最近邻插补**或**距离函数匹配**采用“最邻近的”已通过编辑规则检验的记录，来填补“未通过编辑规则检验的记录”的项目值，其中“最邻近”按照其他已知变量的距离函数来界定。该方法可在插补类别内采用，而且通常适用于连续变量，但也适用于非数字变量。

12. **随机插补法**包括增加了随机残差的回归法或任何其他确然插补法，以及热卡法或冷卡法。

13. 每一个确然插补法都对应一个随机法。为实现这一目的，可为源自确然插补法的插补值，增加一个源自适当分布的随机残差。该法有助于更好保存数据文件的频率结构。Kalton和Kasprzyk（1986年）对该方法的使用进行了评论。

14. **热卡和冷卡插补**旨在使插补值中的变异性比确然插补法更切合实际。在热卡插补法下，要从已通过编辑规则检验的调查或普查记录中（随机）选取各种值，然后用这些值来取代缺失或不一致的值。冷卡插补法则根据其他来源进行插补——通常采用历史数据，如：相同调查或普查在早些时候的数据。热卡和冷卡插补的形式各种各样。

15. **随机整体插补**是最简单形式的热卡插补。对于每个没有通过编辑规则检验的记录，要从所有通过编辑检查的记录组中随机选择一项记录，然后用该记录关于所涉项目的填报值，来为“未通过编辑检查的记录”进行插补。

16. **组内随机插补**同样采用插补类别，来将供体记录的随机选择限于一组与待插补记录之间具有某种相似性的记录组。

17. **顺序热卡插补**也采用插补类别，其优点是，一次性通过数据文件，便足以完成插补程序。插补程序从每个插补类别的冷卡值开始，然后依次考虑数据文件中的记录。当检测到一个已通过编辑规则检验的记录时，将采用该记录关于所涉项目的值，来取代插补类别的存储值。在检测到一项未通过编辑规则检验的记录时，则用存储值来取代其缺失或不一致的值。插补类别数目不宜过多，因为要确保每个推算类别中都有供体。如果数据文件中记录的顺序是随机的，那么这种方法将类似于组内随机插补。其缺点是，通常导致供体的多次使用，因而可能会对项目分布和方差估计产生不利影响。

18. **分层热卡插补**为增强型顺序热卡插补，要采用很多插补类别。在无法从初始插补类别中发现供体时，类别将逐层坍塌，直至找到供体为止。

19. **单一供体热卡插补**算法的目的是，从单一供体，来为未通过编辑规则检验的记录插补数据。因此，对于被编辑规则确认为有问题的记录，这种方法有助于对这类记录的所有项目值进行联合插补。实际上，这种方法通常是为了对记录中变量密切相关的每个部分，采用单一供体。这种方法的明显优势是，不仅可以像上述热卡插补法那样，更好地保持边际分布，而且还可以保持联合频率分布。其另一优点是，可以减少插补值问题——这种问题是：插补值将无法通过那些将在随后各变量部分加以考虑的其他编辑规则的检验。在单一供体热卡插补法下，通过编辑规则检验的记录是指，已通过有关部分所有编辑规则检验的记录。未通过编辑规则检验的记录则指，至少未通过其中一项编辑规则检验的记录。

20. **Fellegi-Holt编辑和插补法**（Fellegi和Holt，1976年）同时考虑所有编辑规则。这种方法的关键特点是，插补规则从相应的编辑规则中推导，无明确规范。对于没有通过编辑规则检验的每项记录，将首先执行一个找错步骤，确定插补的最小变量集及其可接受的值域，然后进行插补。在大多数的实施过程中，要根据编辑规则所涉及的但无需插补的其他变量进行匹配，以便从那些通过了编辑规则检验的记录中，选择一个单一的供体。这种方法寻找的是一个单一的完全匹配对象，不能扩大用于编辑规则没有明确涉及的其他变量。偶尔的情况下，可能无法找到合适的供体，这时必须采用默认插补法。

21. **最近邻插补法**（NIM）（Bankier等，1996年；Bankier、Lachance和Poirier，1999年）类似于Fellegi-Holt法，也是同时考虑所有编辑规则，不明确规定插补行动，并从一个单一供体中进行插补。这种方法根据可用的潜在供体，为每个没有通过编辑规则检验的记录，确定变化最小的插补行动。这有助于确保有供体可用。与Fellegi-Holt法不同的是，这种方法首先寻找供体，然后确定变化最小的插补行动，在寻找供体时，要根据编辑规则所涉及的所有变量（包括可能会进行插补的变量）进行匹配。邻近的数字变量匹配项加上大多数（不一定是全部）其他变量的匹配项，可满足这种方法的要求。要确定基于每个潜在供体的插补行动，并确认那些变化最小的插补行动。这种方法还考虑了那些接近最小变化的插补行动；有时，这些会带来看似更加合理的插补记录。最后，要在变化最小和接近最小变化的插补行动中，随机选择一个，以进行插补。

22. Fellegi-Holt法和最近邻插补法对计算的要求较高，但通过现有的高效运算法则，可以利用现代计算机来进行实施和运用。对于最近邻插补法来说，尤其如此，与Fellegi-Holt法相比，这种方法可以方便地处理更大一些的编辑和插补问题。

23. 上述所有插补法都为每个缺失或不一致的值产生一个单一的插补值，都会在某种程度上使所涉项目值的通常分布失真，而在采用标准方差估计

式时，可能会导致不适当的方差估计数据。失真度的差异很大，具体取决于插补数量和所用的方法。

24. **多重插补**由Rubin（1987年）提出，为解决该问题，该方法为需要插补的每个值进行数次插补（m）。然后从已完成的数据集m中，产生所涉项目的估计数据。由此产生一个单一的合并估计数据和一个表明插补值不确定性的合并方差估计数据。

25. 大多数插补系统都混合采用各种插补法；通常要在可能的情况下采用演绎插补法，其次采用一种或几种其他办法。大多数国家统计机构都在普查编辑和插补中，采用某种形式的动态插补法。目前，顺序热卡插补和Fellegi-Holt法的使用最为普遍。在目前采用Fellegi-Holt法的国家统计机构中，其中一个已转用最近邻插补法，其他不少国家统计机构正在考虑采用这种方法。然而，考虑到预计的主要读者群，本《手册》重点介绍了顺序热卡插补。

附　件　六
计算机编辑软件包

1. 有了价格相对低廉的微机，各国应能对普查和调查数据进行全面和及时的编辑。[a]直到最近，每个国家还不得不自行编写编辑程序，这需要付出高昂的代价，因为需要时间进行调试和处理。但标准计算机编辑软件包的问世，可在更大程度上满足一国的编辑需求，对数据处理专业知识的要求也较少。

2. 利用计算机编辑软件包的优点之一是，如果使用得当，可保持数据的清洁和一致性，从而可以更及时地进行制表。很多计算机软件包（如，SAS和SPSS）或其他高级语言都可用来编写编辑程序。一国还可采用为编辑普查和调查数据而专门编写的计算机软件包。就大多数国家而言，一般编辑如果采用软件包，其速度要比采用专门编写的程序来得快，因为与专门编写的程序相比，软件包将不需要那么多的数据处理知识。

3. 良好的计算机编辑软件包会为主题专家和程序设计员之间的交流提供空间，并应有助于将叙述性或伪代码安排在编程代码附近——除非编程代码本身能使主题专家一目了然。人口学家或其他专家应能够逐行检查程序，并准确理解程序在做什么。

4. 一国考虑采用的任何现有计算机编辑软件包都必须能够为普查数据编辑所需的各种检查、测试和插补执行和产生各种报告。即使在数据处理专家专门编写编辑程序时，这些要求也适用。软件包必须满足以下要求：

有能力键入和/或核实登入项数据，应为跳转模式的增加提供空间。例如，编辑小组可决定遇到男性时，必须跳过生育率信息；

进行结构编辑，据此可以确定应存在的记录类型实际上是否存在，例如，每个序号是否存在住房记录；

(a) 为缺失部分生成相应的记录和/或为现有记录增加权数；

(b) 确定每个变量都有一个有效值；

(c) 存储已编辑的所有或部分记录；

(d) 测试同一记录内和不同记录之间两个或多个特征的一致性。其中一个部分是测试住户内的一致性，将各答复与前面住户成员的答复进行对照检查。如果国家采用动态插补，则采用热卡技术来插补各值；

[a] 值得注意的是，本《手册》对普查编辑进行了重点介绍。由于调查具有较少的答卷人和较多的问题（通常如此），所以通常要准备更详细的编辑规则。本附件讨论的某些软件包是为人口大国的调查而设计，但也适用于进行普查的较小国家。在文件较大时，将利用某些更容易的统计方法，如回归分析和多元分析。另外，相对于每个被抽样个体“代表”了很多个体的调查而言，在有了一个完备文档的情况下，未回答的影响将较小。因此，每个国家统计局都必须对各种软件包进行测试，以确定一个最适合其需求的软件包。

(e) 利用一个记录内或来自多项记录的数个值建立一个衍生变量，然后将衍生变量插入适当的记录中；

(f) 找出并消除重复记录；

(g) 按小区域，生成错误和变动日记。

5. 软件包或程序通常一次编辑一项记录，但现有软件包也允许进行记录间检查，尤其是住房单元内的检查。

6. 如文中所述，在引进Fellegi-Holt法（1976年）及其后续方法之前，几乎所有编辑都采用自上而下法；换言之，按顺序编辑项目——通常（但不总是）按照其收集时的顺序。例如，由于有待编辑的第一个人口项目通常为“关系”，“性别”将根据该项目进行编辑，然后根据性别和关系对“年龄”进行编辑，以此类推。

7. 在过去几个年代中，已根据Fellegi-Holt法开发了一些最小变化插补系统。其中包括加拿大统计局的CANEDIT及通用编辑和插补系统（GEIS）；加拿大普查编辑和插补系统（CANCEIS）（Bankier，2005年；陈，2007年）。[b] 美国普查局开发了DISCRETE（陈等人，2000年；Winkler，1997年a；Winkler，1997年b；Winkler和陈，2002年）以及“结构化经济编辑和查询方案”（SPEER）（Greenberg与Surdi，1984年；Winkler和Draper，1997年）。Kovar和Winkler（2000年）对加拿大和美国系统进行了较详细的比较。Fellegi-Holt衍生的其他编辑包包括CHERRYPI，这是由荷兰统计局根据Fellegi-Holt开发的编辑包（De Waal和Van de Pol，1997年）。

[b] 在对加拿大普查编辑和插补系统（CANCEIS）的评述中，Banquier、Lachance和Poirier（2000年，第10页）总结如下：当最小变化供体插补适当时，CANCEIS以其高效的编辑和插补运算法则，在解决非常普遍的插补问题方面显示了巨大的前景，这些普遍问题涉及大量的编辑规则与大量的定性和定量变量。但如果没有足够的供体，或者如果更适合采用另一方法进行插补，那么Fellegi/Holt最小变化编辑和插补运算法则应该仍然成为较小插补问题的一种选择。

8. 1996年加拿大普查局（和其他普查局）采用了另一种称为最近邻插补法的方法。1996年版本同时对同一住房内的所有人，插补年龄、性别、婚姻状况和关系的答复（Bankier，1999年）。2001年加拿大普查和后续统计活动对该方法进行了改进和扩充（Bankier、Lachance和Poirier，2000年；Banquier，2001年）。

9. 最近邻法首先检索最近邻供体，然后根据这些供体确定最小变化插补。一方面，Fellegi-Holt法需要插补尽可能少的变量，保存分组人口的完整性；另一方面，最近邻插补法则颠倒操作顺序——首先检索供体，然后改变最小数量的变量，这种方法具有计算上的优势。但最近邻插补法可进行只采用供体的插补，而Fellegi-Holt可与其他方法（如，自上而下法）一起使用。加拿大统计局在2001年和2006年普查中，将最近邻插补法列入了加拿大普查编辑和插补系统中。

10. 2000年后的一系列会议将欧洲的众多统计人员召集在一起，以从各个方面考虑普查和调查的编辑和插补问题。讨论围绕“编辑和插补新方法的开发和评价”（EUREDIT）这一主题进行。加拿大统计局还开发了计量和减少插补可变性的方法。他们为调查设计的SIMPVAR系统旨在处理四个主要插补法（比值插补、均值插补、热卡插补和最近邻插补）（Rancourt等人，1997年）。国家统计局（ISTAT）采用DIESIS（数据插补和编辑系统——意大利软件）和

其他方法（Di Zio，2002年；Bianci等人，2005年），对2001年意大利普查的编辑和插补系统进行了介绍。

11. 除了采用实际个案的方法外，还有用来插补未知项的其他方法。有时，采用平均尺度。有些国家采用回归模型（俄罗斯联邦，国家统计委员会，2000年）。对于2000年美国短表上的年龄，还采用了回归法进行插补（Williams，1998年）。

12. 由于交互式键入的进步，有些系统将编码、键入和编辑并入一个单一的系统中，对于调查来说，尤其如此。其中包括巴西的CRIPTAX系统（Hanono和Barbosa，未注明出版日期），该系统要求在登录时进行编辑。其他系统，包括普查和调查处理系统（CSPro）具有交互编辑的特点。如其他地方所述，各国统计局必须确定机器、软件、人员、时间等方面的投资回报。

13. 至于自上而下法，美国普查局为1980-2000年的普查开发了一体化微机处理系统（IMPS）。这个软件包放在DOS系统中，包括数据录入、编辑和制表以及其他功能，很多国家仍在使用该软件包。在1990年代晚期和21世纪，普查局开发了Windows版本，即上述的CSPro，该系统执行的很多任务都相同，具有Windows兼容的特点。[c]CSPro对小国家的所有调查和普查都运转良好；但处理速度很慢，编辑时间要比IMPS更长。不过，编辑从零开始的国家应采用Windows版本。在Fellegi-Holt法下，IMPS和CSPro的应用软件都可以开发；最近邻法需要进行更多的工作。

[c]CSPro是进行普查和调查数据录入、编辑、制表和发布的软件包，它在Windows环境下，将一体化微机处理系统(IMPS)和一体化调查分析系统(ISSA)结合起来。

14. CSPro允许用户从单一的一体化开发环境中，创造、修改和运行有关数据录入、批编辑和制表的应用软件，它在个案（一个或多个问卷）基础上处理数据，其中个案可包括一项或多项数据记录。数据储存在由数据字典描述的美国信息互换标准代码（ASCII）文本文件中。CSPro含有强大的共同程序语言，以实施数据录入控制和编辑规则。

15. 具体而言，CSPro的批编辑功能可发现并报告问卷数据中的结构、值和一致性错误。该软件包可根据简单或复杂的方法，改变（插补）数据值。可生成简要或详细的错误和纠正数据报告、访问多重查找文件、从第二文件中读取或写入第二份文件中。

16. CSPro提供的各种工具，可用来查看数据和其他文本文件，查看CSPro生成的表格和专题地图，将IMPS和ISSA数据字典转换到CSPro和从CSPro转换IMPS和ISSA，将环境系统研究所（ESRI）形状文件（地图）转为CSPro地图文件。该系统由美国普查局、Macro国际公司和Serpro公司联合开发，并由美国国际开发署提供主要资金，它属于公众域，可免费获得和自由发布，可从www.census.gov/ipc/www/cspro下载。

术 语 表

阵列	一组数值。有时称为矩阵，可用来存储重复性数值数据。
审计跟踪	跟踪字段中各值变化以及各变化原因和来源的一种方法。审计跟踪一般在初始访问结束后开始。
自动更正	在没有人工干预的情况下，由计算机更正数据错误，是自动数据编辑的一个方面（Pierzchala，1995年）。
记录间编辑	在涉及多份调查记录的字段上所进行的编辑。统计编辑是记录间编辑的一个例子，因为分布是根据调查中所有记录的各组字段产生（Pierzchala，1995年）。
分类均值插补	采用所定义的插补类别，来建立具有一定程度类似性的记录组。
清洁记录	没有缺失值并通过了所有编辑规则检验的记录（Pierzchala，1995年）。
代码表	关于数据项所有认可（许可）值的清单。
冷卡	初始静态矩阵；一种在更正前就给定各元素，并且在更正期间各元素不改变的一种更正基础。例如，更正基础可以是上年的数据。一个经过修正的冷卡可根据（可能汇总的）当前信息调整冷卡值。
整套编辑规则/关系式	显性编辑规则加上隐性编辑规则（在希望插补能够满足编辑规则要求的情况下），用来为插补产生可行的区域（Pierzchala，1995年）。
一致性编辑	检查行列式关系，如：分项相加后等于总和，或者“收获面积”始终小于“种植面积”（Pierzchala，1995年）。
数据捕获	使所收集数据变为机器可读形式的过程。基本编辑检查通常在数据捕获软件的子模块中进行。
演绎插补	一种可用来确定地推断缺失值或不一致值的方法，通常要以调查表上其他项目的答复模式为基础。
确然编辑	一种如果违反则以1的概率指向数据中某项错误的编辑。如：年龄 = 5岁而身份 = 母亲。与随机编辑形成对照（Pierzchala，1995年）。

确然插补	一个字段中只有一个值会使记录满足所有编辑规则要求时，将发生这种情况。在某些情况下（如分项相加后不等于总和），要进行这种插补。它是调查数据自动编辑和插补时，需要检查的第一个解决方案（Pierzchala，1995年）。
距离函数	对于数字数据而言，一种同时根据受体（候选）和供体记录的匹配变量进行界定，并用来对相似性概念进行量化的函数。在热卡插补中，可用来寻找匹配记录（Pierzchala，1995年）。
距离函数匹配	从“最邻近”已通过编辑规则检验的记录中，为未通过编辑规则检验的记录分配一个项目值，其中“最近邻”按照其他已知变量的距离函数来界定。
供体插补	将需要插补的每项记录（受体或候选记录）与源自某个固定供者总体的某项记录进行配对的一种方法，例如热卡插补中的情况（Pierzchala，1995年）。
编辑规则/关系式（定义1）	对每个变量能够假定的值所进行的逻辑限制（Pierzchala，1995年）。
编辑规则/关系式（定义2）	检查禁答组合的规则（Pierzchala，1995年）。
编辑跟踪	见“审计跟踪”。
显性编辑规则/关系式	由主题专家明确编写的编辑规则（与隐性编辑规则相对）（Pierzchala，1995年）。
未通过编辑规则检验的记录	在编辑和插补中利用与所涉项目有关的编辑规则进行检查时，至少未通过其中一项检查的记录。
Fellegi-Holt自动更正法	在这种自动更正法中，要尽量减少数据项的变动数目，并利用Fellegi-Holt模型来确定插补项目的可接受值或值域。可适用基于冷卡或热卡法的顺序插补或同时插补。
Fellegi-Holt系统	指Fellegi和Holt于1976年在《美国统计协会杂志》上发表的论文中所提出的假设以及编辑和插补目标。Fellegi-Holt模型的关键特点是，它证明需要有隐性编辑规则，来确保数据字段中不插补的一组值能始终使最终（插补）记录符合所有（个案）。
标记	是一种变量，用来注明有关另一个或另一些变量的有用信息。例如，如果将一个项目从无效更改为有效，则可用标记注明原始信息，或注明该项目值已更改过。
流程图	必须完成的各项职能的图解说明。

手工编辑 由人在数据输入计算机前进行的编辑（见“人工编辑”）（Pierzchala，1995年）。

埋头数据输入 一种数据输入方式，指数据输入时，数据输入机器不检测其中的错误，有助于操作员快速输入数据。

抬头数据输入 一种数据输入方式，指数据输入时，数据输入机器检测其中的错误，有助于操作员当场纠正错误（Pierzchala，1995年）。见“交互式键入”。

热卡插补 一种插补方法，指供体记录取自当前抽样数据卡片（与之相对的冷卡则指，供体记录源自昔日调查数据的一种插补法）（Pierzchala，1995年）。

隐性编辑规则/关系式 一种未阐明的编辑规则，根据主题专家编写的显性编辑规则，从逻辑上推导得出（Pierzchala，1995年）。

插补 向某个字段分配一个值，以作为未答项的值，或者在某项记录值不符合某组编辑规则的情况下，用来取代该记录的值（Pierzchala，1995年）。

交互式键入 一种数据输入方式，指数据输入时，数据输入机器检测其中的错误，有助于操作员当场纠正错误。见“抬头数据输入”。

内部一致性 该用语与给定样本单位各变量之间的关系有关，是多数调查程序中进行编辑的原因（Ford，1983年；Pierzchala，1995年）。

宏观编辑 通过（1）汇总数据的检查，或者（2）整个记录的检查，来对各个错误进行检查。检查要以估计数据为基础（Granquist，1987年；Pierzchala, 1995年）。

人工编辑 由人在数据输入计算机前进行的编辑（见“手工编辑”）（Pierzchala，1995年）。

匹配 在热卡插补法中，将供体记录与受体（候选）记录进行匹配的行为（Pierzchala，1995年）。

匹配变量 用来找出受体（候选）记录和供体记录之间某个匹配项的变量（Pierzchala，1995年）。

微观编辑 对数据记录进行的传统编辑。在逻辑上与宏观编辑相对（Pierzchala，1995年）。

微观-宏观编辑 一种编辑方法，指兼用微观编辑和宏观/统计编辑，以取代详细的微观编辑。两种编辑并用情况下的微观编辑不像单纯的微观编辑那么详细。最好“在开发调查编辑规则时，抱着一种‘影响估计数据’的想法，而非‘捕获所有不一致数据’的想法”（Granquist，不同日期；Pierzchala，1995年）。

最小集	为确保通过所有编辑规则的检验，而要求插补的最小字段集（Pierzchala，1995年）。
基于模型的插补	采用“已通过编辑规则检验的记录”中的数据，根据一组预测变量，来对需要插补的变量进行回归。
多重插补	为要插补的每个值进行数次插补，然后产生所涉项目的估计数据。
多元编辑	一种统计编辑方式，指利用多元分布来评价数据和寻找界外值（Pierzchala，1995年）。
最近邻插补	采用“最邻近的”已通过编辑规则检验的记录，来填补“未通过编辑规则检验的记录”的项目值，其中“最邻近”按照其他已知变量的距离函数来界定。
最近邻插补法	类似于Fellegi-Holt法，也是同时考虑所有编辑规则，不明确规定插补行动，并从一个单一供体中进行插补。这种方法根据可用的潜在供体，着眼于每个没有通过编辑规则检验的记录，以确定最小变化插补行动。
未回答	不完整的调查表或有缺漏的调查表（Pierzchala， 1995年）。
界外值	根据边界的某种确定，而位于某个边界外的项目值（Pierzchala，1995年）。
整体均值插补	采用“已通过编辑规则检验的记录”的项目均值，来填充所有“未通过编辑规则检验的记录”的缺失项目或不一致项目。
已通过编辑规则检验的记录	在编辑和插补期间利用与所涉项目有关的所有编辑规则进行检查时，都获得通过的记录。
指针	指针是一个变量，用来为一个项目或其他变量作标记以备日后查阅。例如，可用来注明“户主”和“配偶”的行号，以便日后确保配偶双方性别相反且都已结婚。
生产运行	在从编辑或制表程序中消除初始“缺陷”后，处理大量数据的行为。
伪代码	书面编辑指令或说明。
质量错误	可能会扭曲数据质量的错误：例如，导致偏差的规律性错误（Granquist，1984年；Pierzchala，1995年）。
定量编辑规则	对按照连续比例计量的字段所采用的编辑规则（Pierzchala，1995年）。
组内随机插补	采用插补类别，来将供体记录的随机选择限于一组与待插补记录之间具有某种相似性的记录组。

随机整体插补 对于每个没有通过编辑规则检验的记录，要从所有通过编辑检查的记录组中随机选择一项记录，然后用该记录关于所涉项目的填报值，来为“未通过编辑检查的记录”进行插补。

记录 磁存储的计算机可读调查数据。通常情况下，每份调查表有一份记录，不过可以将一份调查表的数据细分为多项记录，如：人口和住房（Pierzchala，1995年）。

回归插补 采用“已通过编辑规则检验的记录”中的数据，根据一组预测变量，来对需要插补的变量进行回归。

检索 在热卡插补法中，检索供体记录的行为（Pierzchala，1995年）。

顺序热卡插补 按顺序对一系列变量进行编辑时，发生的插补，只将编辑过的值用作后续热卡变量。

相似性 在数字数据中，指基于规定匹配变量的两项记录之间的接近度。该概念要根据某种标准，利用距离函数来量化（Pierzchala，1995年）。

单一供体热卡插补 从单一供体，来为未通过编辑规则检验的记录插补数据，对于被编辑规则确认为有问题的记录，这种方法有助于对这类记录的所有项目值进行联合插补。

统计编辑 根据答复数据的统计分析而进行的一组检查：例如，根据对假定有效填报人的比值所进行的统计分析，两个字段的比值位于基于这种分析的限度之间（Greenberg和Surdi，1984年；Pierzchala，1995年）。

统计插补 统计插补的一个例子是，利用回归模型，在这个模型中，要对因变量进行插补，而自变量的系数则从假定的有效答复中推导（Pierzchala，1995年）。

（热卡中的）统计匹配 根据某种统计标准，将供体记录与受体（候选）记录进行匹配的行为，目的是将供体数据转移至受体（Pierzchala，1995年）。

随机编辑 一种如果违反则以小于1的概率指向数据中某项错误的编辑（Pierzchala，1995年）。

结构编辑 根据两个或多个已编辑字段之间的逻辑关系所进行的检查。例如，总和必须等于分项之和；或者，由于调查表内固有的跳转模式，位于不相交路径上的两个变量不能都为非零。结构编辑可确保数据记录能够保持调查表的结构（Pierzchala，1995年）。

结构插补 在数个变量之间存在结构关系时，采用结构插补。例如，总和必须等于分项之和：因此，对于母亲来说，已生育子女必须等于存活子女加上死亡子女（Pierzchala，1995年）。

审核编辑 对特定记录中各字段之间所作的编辑项目检查。包括：对每项记录每个字段进行检查，以弄清其所含登入项是否有效；对某种预定字段组合中的登入项进行检查，以弄清各登入项之间是否一致（Pierzchala，1995年）。

权数 在Fellegi-Holt学派的编辑和插补中，根据可靠性对字段分配权数（在其他所有条件相同的情况下），权数越大，对字段进行插补的可能性就越大。还可以为编辑检查分配权数（Pierzchala，1995年）。

记录内编辑 审核编辑的另一名称（Pierzchala，1995年）。

参考文献

Banister, J.（1980年）。“普查编辑和插补的使用和滥用”。《亚太普查论坛》第6卷第3期第1-20页。

Bankier, M.（1999年）。“加拿大1996年普查中新插补法的使用经验以及用于日后普查的延伸插补法”，数据编辑研习会记录，联合国欧洲经济委员会，意大利（罗马）。

_____（2005年）。“加拿大2006年普查的编辑和插补”。统计数据编辑工作会议上宣读的论文，渥太华，2005年5月16-18日。

_____，A. M. Houle和M. Luc（无日期）。“加拿大普查人口变量插补”。手稿。

Bankier, M.、M. Lachance和P. Poirier（1999年）。“新插补法的一般使用”。载于《调查研究方法科的议事录》。弗吉尼亚州亚历山德里亚：美国统计协会，即将出版。

_____（2000年）。“加拿大2001年普查最小变化供体插补法”。联合国欧洲经济委员会统计数据编辑工作会议上宣读的论文，英国加的夫，2000年10月18-20日。

Bankier，M.、P. Poirie和M. Lachance（2001年）。“加拿大普查编辑和插补系统（ANCEIS）内的有效方法”。美国统计协会年会记录，2001年8月5-9日。

Bankier, M.等人（1996年）。“同时插补数字和定性普查变量”。载于《调查研究方法科的议事录》。弗吉尼亚州亚历山德里亚：美国统计协会，第287-292页。

Bianchi, G等人（2005年）。“人口变量编辑和插补新法”。统计数据编辑工作会议上宣读的论文，渥太华，2005年5月16-18日。

Boucher, L.（1991年）。“制造商年薪的微观编辑：增加值是什么？”载于《年度研究会议记录》。哥伦比亚特区华盛顿：美国普查局，第765-781页。

Chambers, Ray（2000年）。“EUREDIT中编辑和插补的评价标准”。联合国欧洲经济委员会统计数据编辑工作会议上宣读的论文，英国加的夫，2000年10月18-20日。

Chen, Bor-Chung（2007年）。“2006年普查测试数据的编辑和插补中对加拿大

普查编辑和插补系统（CANCEIS）的尝试”。《统计研究部研究》丛书，第2007-1期。哥伦比亚特区华盛顿：美国普查局。

_____及其他人（2000年）。“对ACS调查采用编辑DISCRETE系统”。统计研究部统计研究报告丛书，第RR2000/03期。哥伦比亚特区华盛顿：美国普查局。

De Waal, Tom以及Frank van de Pol（1997年）。“编辑过程中应用CHERRYPI的诀窍”。统计数据编辑工作会议上宣读的论文，布拉格，1997年10月14-17日。

Di Zio, M（2002年）。“评价编辑和插补法：意大利经验”。联合国欧洲经济委员会统计数据编辑工作会议，赫尔辛基，2002年5月27-29日。

Fellegi, I. P.和D. Holt（1976年）。“自动编辑和插补的系统性方法”。载于《美国统计协会杂志》第71卷第353期（3月），第17-35页。

Ford, Barry L.(1983年)。“热卡法概述”。载于《抽样调查的不完整数据》，第2卷《理论与参考书目》，William G. Madow、Ingram Olkin和Donald B. Rubin编辑。

Granquist, L.（1984年）。“数据编辑及其对统计数据进一步处理的影响”。在统计计算研习班上的发言，布达佩斯，1984年11月12-17日。

_____（1987年）。“瑞典统计局计算机辅助编辑的短期开发方案”。在数据编辑联合小组会议上的报告，马德里，1987年4月22-24日。斯德哥尔摩：瑞典统计局。

_____（1997年）。“关于编辑的新观点”。载于《国际统计评论》第65卷第3期，纽约：学术出版社，第381-387页。

_____以及J. G., Kovar（1997年）。“调查数据的编辑，多少才算够？”，载于《调查计量和过程质量》，Lyberg等人编辑。纽约：Wiley and Sons出版社，第415-435页。

Greenberg、Brian和Rita Surdi（1984年）。“比值编辑的交互式灵活编辑和插补系统”。载于《美国统计协会调查研究方法科的议事录》。弗吉尼亚州亚历山德里亚：美国统计协会，第421-426页。

Hanono、Reina Marta和Dulce Maria Rocha Barbosa（未注明日期）。“统计调查数据捕获、编辑和编码应用开发的通用环境”。里约热内卢：巴西地理和统计研究所（IBGE）。

Ireback, H.（2000年）。“新信息技术对瑞典统计局收集数据的影响”。联合国欧洲经济委员会统计数据编辑工作会议上宣读的论文，英国加的夫，2000年10月18-20日。

Kalton, G.和D. Kasprzyk（1982年）。“缺失调查答复的插补”。载于《调查研究方法科的议事录》，美国统计协会，第23-31页。

_____（1986年）。“缺失调查数据的处理”。载于《调查方法》，第12卷第1-16页。

Kovar, J.和W. Winkler（2000年）。“编辑经济数据时‘通用编辑和插补系统’(GEIS)和‘结构化经济编辑和查询方案’（SPEER）的比较”。《普查局统计研究部统计研究报告丛书》，第RR2000/04期。哥伦比亚特区华盛顿：美国普查局。

Naus, J. I.（1975年）。《数据质量控制和编辑》。纽约：Marcel Dekker出版社。

Nordbotten, S.（1963年）。“各统计观察值的自动编辑”。《欧洲统计学家会议统计标准和研究》，第2期。纽约：联合国。

Pierzchala, M.（1995年）。“编辑系统和软件”。载于《企业调查方法》，B. G. Cox等人编辑。纽约：John Wiley and Sons出版社，第425-411页。

Poirier, C.（2000年）。“数据编辑和插补的原型知识基础”。联合国欧洲经济委员会统计数据编辑工作会议上宣读的论文，英国加的夫，2000年10月18-20日。

Pullum, T. W.、T. Harpham与N. Ozsever（1986年）。“大样本调查的机器编辑：世界生育调查经验”，载于《国际统计评论》，第54卷，第311-326页。

Rancourt, E.等人（1997年）。“在存在插补情况下的方差估计”，《1997年研讨会记录：调查和普查的新方向》。渥太华：加拿大统计局，第273-279页。

Rubin, D. B.（1987年）。《调查中未回答的多重插补》。纽约：Wiley出版社。

俄罗斯联邦，国家统计委员会（2000年）。“根据具有熵变的回归模型所进行的数据插补”。联合国欧洲经济委员会统计数据编辑工作会议上宣读的论文，英国加的夫，2000年10月18-20日。

Sande, I. G.（1982年）。“调查中的插补：应对现实”。《美国统计学家》第36卷，第145-152页。

Särndal, C. E.、B. Swensson和J. Wretman（1992年）。《模型辅助的调查抽样》。纽约：Springer-Verlag出版社。

加拿大统计局（1998年）。《加拿大统计局质量指导方针》第三版。渥太华：加拿大统计局。

联合国（1992年a）。《人口和住房普查手册，第一部分：人口和住房普查的规划、组织与管理》。方法研究，F辑，第54号，出售品编号：E.92.XVII.8。

_____（1992年b）。《人口和住房普查手册，第二部分：人口和社会特征》。方法研究，F辑，第54号，出售品编号：E.91.XVII.9。

_____（1999年）。《用于统计的标准国家或地区代码》，统计文件，M辑，第49/Rev.4号，出售品编号：M.98.XVII.9。

_____（2008年）。《人口和住房普查的原则和建议》第二次修订本。统计文件，M辑，第67/Rev.2号，出售品编号：E.07.XVII.8。

联合国统计委员会和欧洲经济委员会（1994年）。《统计数据编辑，第1卷第1期：方法和技术》。欧洲统计学家会议，统计标准和研究系列，第44号，出售品编号：94.II.E.36。

_____（1997年）。《统计数据编辑，第1卷第2期：方法和技术》。欧洲统计学家会议，统计标准和研究系列，第48号，出售品编号：96.II.E.30。

Williams, Todd R.（1998年）。“2000年普查短表人员年龄的插补：基于模型的方法”。哥伦比亚特区华盛顿：普查局统计研究部，统计研究报告系列，第RR98/07号。

Winkler, W. E.（1997年a）。“美国十年一次普查的编辑/插补系统”。统计数据编辑工作会议上宣读的论文，布拉格，1997年10月14-17日。

_____（1997年b）。“集覆盖和编辑离散数据”。技术报告。哥伦比亚特区华盛顿：美国普查局。

_____（2006年）。“数据质量：自动编辑/插补和记录链接”。《美国普查局统计研究部统计研究报告丛书》，第2006-7期。哥伦比亚特区华盛顿：美国普查局。

_____，以及B. C. Chen（2002年）。“扩充统计数据编辑的Fellegi-Holt模型”。《美国普查局统计研究部统计研究报告丛书》，第2002-02期。

Winkler, W. E.和L. R. Draper（1997年）。结构化经济编辑和查询方案”（SPEER）编辑系统。载于《联合国统计委员会和欧洲经济委员会统计数据编辑第2卷：方法和技巧》。欧洲统计学家会议统计标准和研究，第48期。出售品编号：E.96.II.E.30，第56-62页。