

## 学习目标

调查是一个收集信息的系统，这些信息来自人群或者与人群有关，用以描述、比较或者解释人们的知识、态度和行为。调查系统包含为完成收集有效数据而设计的多种活动，如确定信息收集的目标；设计研究方案；准备可靠而且有效的调查手段；实施调查；管理和分析调查数据；报告调查结果。

调查研究人员可以通过询问人们问题而直接收集信息；或者通过检查有关人们思想和行为的书面、口头以及视觉记录，间接地收集信息。调查者还可以通过观察在自然状态或者实验状态中的人群而获取信息。

本书有两个主要目的：教你如何管理调查数据；帮助你更好地使用统计和定性调查的信息。本书介绍了数据管理和统计学的基本词汇，以及选择和解读通常用于分析调查数据的统计和定性方法背后的原则和逻辑。本书不能教你成为一位调查统计学家。要想成为调查统计学家，你需要进行正规的学习。如果在选择或者应用统计方法方面你

需要帮助，你必须得到统计学家的建议。如果此书能够如愿以偿，你将能够确切地告诉统计学顾问或者定性研究人员你的需要，并且你将能够解读提供给你的数据。

这些特定目标能够使你：

✓组织和用于分析的数据

- 起草分析计划
- 定义和格式化数据文件
- 创建编码手册
- 确定编码的可靠性
- 确认处理不完整数据或者缺失数据和异常值，以及重新编码的技术

✓确定把数据准确录入电子数据表、数据库管理程序和统计程序的方法

✓学会使用以下分析术语：

- 分布
- 临界值
- 偏度
- 转换
- 集中趋势度量
- 离散趋势
- 变异
- 统计显著性
- 实际显著性

## 学习目标

- $p$  值
- $\alpha$  值
- 直线性
- 曲线性
- 散点图
- 无效假设

✓列出选择合适分析方法的步骤

✓辨别名义等级、有序等级、数值等级和数据等概念

以便做到：

- 分清自变量和因变量
- 正确使用均值、中位数和众数
- 正确使用极差、标准差、百分位数和四分位数间距
- 理解相关和回归的逻辑和使用
- 学会操作和解读假设检验的步骤
- 比较和对照假设检验与置信区间的使用
- 计算比值比和风险比值
- 理解卡方分布和卡方检验的逻辑和使用
- 理解  $t$  检验的逻辑和使用
- 理解方差的逻辑和使用
- 阅读和解释计算机输出结果

✓理解定性数据内容分析的基本步骤

✓进行开放式问题的分析，这些开放式问题询问被访

者最喜欢什么和最不喜欢什么

## 第 1 章 数据管理

数据管理包括调查人员组织调查数据使之便于分析的活动。数据管理以制定分析计划为起点，以数据分析的开始为终点。数据管理行为包括：

- 起草分析计划
- 创建编码手册
- 确定可靠的编码
- 检查调查数据的缺失值
- 录入数据并且证实数据录入的准确性
- 清理数据

如果你计划做复杂的分析，你将会发现自己在不断地管理和修正更新数据，以便确信数据是干净的、完整的和适合于你要做的分析的。如果你的调查相对简短，涉及大约 50 人以及 5 个变量以下，你仍然必须确保数据的干净和完整，只是用比较少的时间管理文档。

有些调查者估计：数据管理所用时间相当于典型的数

据分析过程时间的 20% ~ 50%。对于你将用于数据管理的时间抱有现实的态度，并且确保你拥有资源（人员、时间、财力）对你来说是重要的。下面，将依次讨论每一种数据管理行为。

## 一、起草分析计划

分析计划描述你计划进行分析的主要调查目的、假设或研究问题。例 1.1 展示了一个分析计划样本的一部分。

### 例 1.1 儿童受暴力伤害调查的分析计划的一部分

- 调查目的 1：比较男孩和女孩中至少一星期一次目击和没有目击暴力事件的人数

- 假设：目击暴力事件的女孩人数比男孩多

- 变量：性别（女孩和男孩）；目击暴力行为（是或否——每星期一次或更多）

预计的分析：卡方检验，用来检验至少一星期一次目击和没有目击暴力事件的男孩和女孩的数量差异。

- 调查目的 2：比较男孩和女孩恐惧量度的得分

- 研究问题：男孩和女孩恐惧量度的得分是否有区别

- 变量：性别（女孩和男孩）；平均得分（连续从 1~50）
- 预计的分析： $t$  检验，用来检验男孩和女孩恐惧量度平均得分的差异

调查人员通常在最终选定调查项目和确认调查方法之前草拟初步的分析计划。决定你想分析什么甚至会预示出你的调查内容。基金机构（如国家卫生研究院）和论文委员会要求你列出你的调查假设或者研究问题，描述你将如何分析数据来检验假设及回答问题。

不管你的分析计划如何好，抽样和数据收集过程中遇到的实际问题也许会迫使你修改你的计划。假设你是调查者，你的分析计划如例 1.1 中所示（调查目的 2）。你决定除了检验“恐惧得分”的平均分差异，还要比较“恐惧”得分高于 25 分及低于 24 分的男孩和女孩的数量。因而你只好修改你的计划，使之既包括计划中的  $t$  检验又包括卡方检验。你应该预计到最初的计划不得不做修改，特别是在涉及很多数据的大规模的调查中。

## 二、创建编码手册

编码是计算机程序用来识别变量的单位或符号。假设

1 000 个男孩和女孩完成了关于学校中暴力的调查问卷。你所关心的问题之一是有多少男孩报告在学校中经常受到威胁。要找出关于这个主题的内容，你必须使计算机程序知道要寻找哪些变量。在此案例中，变量是性别和暴力威胁。

看一下例 1.2 中所引用的有关学校中暴力的调查。小方框右边用来记录被访者答案的数字就是编码。你可以要求统计软件程序告诉你有多少人第 1 题回答 1 并且第 8 题回答 4。要做到这一点，你必须告诉程序你所感兴趣的变量的名称（问题 1 为 SEX，问题 8 为 SCHLHURT）以及它们的值（1 = 男孩，4 = 几乎总是）。借助于许多统计程序，你还能够指定在数据中何处能找到变量。参见下面有关录入数据的部分。

### 例 1.2 学校暴力调查摘录

1. 你是男孩还是女孩？ [SEX]

男孩  1

女孩  2

8. 去年在学校里有人告诉你他们将伤害你的频率是多少？ [SCHLHURT]

从来没有  1

有时有  2

很多次 3

几乎总是 4

在例 1.2 中，方括号中的 SEX 和 SCHLHURT 对应于问题所代表的变量。所有的编码手册包含对问题的描述、编码和与调查相关的变量。一个好的编码手册包含充足的信息，使得将来的研究人员能够复制这个调查、研究方法和研究成果。例 1.3 展示了一个编码手册的目录。

### 例 1.3 编码手册目录

#### I. 调查团队描述

这一部分包括对负责实施调查的机构及个人的特点和经验的描述。

#### II. 方法

##### A. 抽样

1. 抽样设计（包括合格入选标准，如 12 岁或以下、上个月读过至少 3 本书）
2. 抽样方法（如分层随机抽样、方便抽样）
3. 样本量
4. 登记

5. 抽样统计（包括加权和抽样误差计算）

B. 研究对象：知情同意

C. 研究设计——被访者如何分组；调查实施的次数及进度安排

Ⅲ. 调查

A. 调查的副本以及每个回答是如何编码的；与每一个问题相关的变量名

B. 培训数据收集人员；质量控制

C. 可靠性和有效性的信息

Ⅳ. 数据文件描述

变量名（如 EDUC），变量标签（教育），变量值及变量值标签（1 = 高于 12 年级，2 = 低于 11 年级，9 = 无数据）

V. 调查手段

问题及评分方案

你可以在调查结束后对编码手册进行整理。此时，你

还应该整理用于分析的数据文件（有些人视之与编码手册等同），参见例 1.4 数据文件和编码手册。所有的变量被分解成不连续的单位，也称为“值”；“值”对应于变量的编码。例如，在学校、邻里及其他地方受暴力威胁的频率有 4 个值：0 = 从来没有，1 = 有时，2 = 很多次，3 = 几乎天天。编码是 0、1、2、3。一个一位数的编码被用来代表没有提供信息，此例中为 9（你也可以用其他的一位数的编码）。同样，13 个值或编码代表被访者出生地的 13 个国家，两位数字的编码被用来代表缺失值（99）。通常，编码手册按照变量在调查中出现的顺序列出变量。数据文件可以照此顺序也可以不照此顺序。

尽管统计软件程序在术语上有所变化，它们都要求变量名用大写字母（如 POOL 或 EDUC）并且避免使用特殊字符，如逗号或分号。有些程序限制一个变量名中字符的个数（通常最多约为 8 个）。变量标签是变量的实际的名字（如“生活质量感知”是变量名为 POOL 的变量标签），为了理解你的数据，统计程序需要知道你的每一个变量的变量名（如 COMM）、变量标签（社区）、变量值标签和变量值（1 = 都市，2 = 乡村）。请注意，你的程序可能会用稍微不同的术语，但是概念完全一样。

在大规模调查项目中，编码手册是项目的正式记录。它包含调查手段（如果有关，包括评分系统）；变量名、变量标签、变量值及编码；编码在数据文件中的位置；调

查方法、特点及结果；调查团队所有成员的名字。如果你打算用以往项目的数据（做另外的分析或者用不同的样本重复调查），你会发现原始编码手册是必需的。你可以用最喜欢的搜索引擎，输入关键字“调查”和“编码手册”，从国际互联网上找到调查编码手册的整个范例。

#### 例 1.4 编码手册的一部分

变量名	变量标签	变量值：标签和编码
PROJID	项目编码	7 位识别号码
AGE	年龄	学生年龄；99 = 无数据
GRADE	年级	学校中年级：1~6；9 = 无数据
GENDER	性别	被访者性别：1 = 男性，2 = 女性；9 = 无数据
CNTRY	出生国家	孩子出生的国家： 01 = 阿根廷，02 = 玻利维亚， 03 = 智利，04 = 哥伦比亚， 05 = 古巴，06 = 厄瓜多尔， 07 = 萨尔瓦多，08 = 危地马拉， 09 = 洪都拉斯，10 = 墨西哥， 11 = 尼加拉瓜，12 = 秘鲁， 13 = 波多黎各；99 = 无数据
SPSPAN	西班牙语讲得好	1 = 一点也不会，2 = 会一点， 3 = 很好，4 = 非常好；9 = 无数据
SPENGLIS	英语讲得好	1 = 一点也不会，2 = 会一点， 3 = 很好，4 = 非常好；9 = 无数据
SCHLHURT	在学校受伤害	在学校里，有人告诉孩子他（她） 将伤害孩子的频率： 0 = 从来没有，1 = 有时， 2 = 很多次，3 = 几乎每天； 9 = 无数据

续表

变量名	变量标签	变量值：标签和编码
NEIBHURT	在邻里受伤害	在邻里，有人告诉孩子他（她） 将伤害孩子的频率： 0=从来没有，1=有时， 2=很多次，3=几乎每天； 9=无数据
ANYWHERE	在任何地方受伤害	在任何地方，有人告诉孩子他 （她）将伤害孩子的频率： 0=从来没有，1=有时， 2=很多次，3=几乎每天； 9=无数据

### 三、确定可靠的编码

如果调查规模小，只有一个人做编码，可以通过对所有或部分数据重新编码的方式来检查其一致性，从而确认其可靠性。第二次编码应该在第一次结束后的一星期，这段时间足以使编码人员忘记第一套编码，使他们不是机械地复制。数据经过再次编码以后，应该比较两套编码的一致性。可以让第二个人对两套编码进行判断并解决不一致的问题。

大规模调查中，应该让第二个人独立地对数据的一部分进行编码。为了确保编码人员之间的可靠性，必须为编码人员提供正规培训，以及所有术语的清晰定义。

尽管你做出最大努力建立了一个高质量的编码手册和数据管理系统，但是编码人员之间仍然会有不同意见。为了发现他们一致性的程度——编码人员或者访问员之间的可靠性——你可以计算一个统计量，叫作  $\kappa$  (kappa) ，它度量了两个编码人员之间的一致性是否是由于偶然性造成的。

度量两位编码人员之间的一致性： $\kappa$  统计量

假设两位访问员被要求独立检查 100 个对有两个或更多孩子的离婚男性的访谈。访谈是关于这些男性最近的单身生活质量。访问员要研究访谈记录，以发现其中有多少男性在访谈过程中提起过他们的孩子。如果一个男性一次也没有提起他的孩子，编码 1 为“否”，如果他提起过他的孩子，编码 2 为“是”。以下是访问员的编码：

		访问员 1		
访问员 2	否	是		
否	20 <sup>C</sup>	15		35 <sup>B</sup>
是	10	55 <sup>D</sup>		65
	30 <sup>A</sup>	70		

访问员 1 说访谈的 100 人中 30(<sup>A</sup>) 人没有提到他们的孩子，而访问员 2 却说 35(<sup>B</sup>) 人没有提到他们的孩子。两位访问员意见一致的是：20(<sup>C</sup>) 人没有提到他们的孩子。

描述访问员之间一致性程度的最佳方法是什么？说

100 人中的 20 人或者 20%<sup>(c)</sup> )可能是不准确的，因为访问员一致同意访谈中 55%<sup>(d)</sup> )的男性提到他们的孩子。合计的一致性，55% + 20%，这是个过高的估计，因为对于仅仅两个分类（是和否），有些一致可能是由偶然性所导致的。这一点体现在下面的公式中，其中， $O$  是观测到的一致性， $C$  是偶然性所致的一致性：

$$k = \frac{O - C \text{ (非偶然性所致的一致性)}}{1 - C \text{ (可能的非偶然性所致的一致性)}}$$

下面是此公式如何用于以上案例：

1. 计算访问员间的一致由于偶然性没有提及男性的孩子的访谈次数。因为有 100 次访谈，你通过把两个访问员所记的“否”的次数相乘，再除以 100 而得到此值： $(30 \times 35)/100 = 10.5$ 。

2. 计算访问员间的一致由于偶然性提及男性的孩子的访谈次数。通过把两个访问员所记的“是”的次数相乘，再除以 100 而得到此值： $(70 \times 65)/100 = 45.5$ 。

3. 把步骤 1 和步骤 2 所得数字相加，再除以 100，得到偶然性所致一致性的比例： $(10.5 + 45.5)/100 = 0.56$ 。

观测到的一致性是：20% + 55% = 75% ,或 0.75。因此，非偶然性所致的一致性是： $0.75 - 0.56 = 0.19$ ，这是分子。可能的非偶然性所致的一致性 100% 减去偶然性所致的一致性 56% ,或  $1 - 0.56 = 0.44$ ，这是分母。因而：

$$k = \frac{0.19}{0.44}$$

$$k = 0.43$$

“高的” kappa 值意味着什么？有些专家给 kappa 值赋予下列定性术语：0~0.2 = 微小的，0.2~0.4 = 较小的，0.4~0.6 = 中等的，0.6~0.8 = 充足的，0.8~1.0 = 几乎完美。

你如何达到访问员之间充足的或几乎完美的一致性，即可靠性？你必须确保所有的访问员收集和记录完全一样主题的资料，并且他们事先对每一个重要变量的含义达成共识。上例所得访问员问的较小 kappa 值 0.43 可归因于访问员和调查者之间的定义差别，在使用的定义方面缺乏培训以及编码的错误。

## 四、检查调查数据的缺失值





缺失值源于调查中未回答的问题以及遗漏掉的调查。在录入数据之前，你应该一收到第一批完成的调查结果就立刻开始检查，看看是否有缺失值问题。在自填式问卷调查中，尤其是邮寄问卷，你会发现未回答的问题。被访者不回答问题有多种原因：他们可能不想回答某类问题，也可能只是没有看见某些问题，或者，可能不理解某些问题。由于阅读量太大或者所需要的阅读水平太高，他们可

能不明白对完成调查的说明，或者不明白问题本身。他们可能对自己应该如何回答某些问题不确定（把框涂满、圈出正确答案等），也就是说他们可能发现问题格式难以理解，见例 1.5。

### 例 1.5 可能令人迷惑的问题格式

请用 X 标出最能描述你的选择

请回答每一个问题

	0	1	2	3
				
32. 去年你有多少次告诉他人你要伤害他们？	从来没有	有时	很多次	几乎每天
33. 去年你有多少次在 <u>别人打你之前</u> ， <u>掌击、拳击或撞击此人</u> ？	从来没有	<del>有时</del>	很多次	几乎每天
34. 去年你有多少次在 <u>别人打你之后</u> ， <u>掌击、拳击或撞击此人</u> ？	从来没有	有时	很多次	几乎每天
35. 去年你有多少次把别人打成重伤？	从来没有	有时	很多次	几乎每天
36. 去年你有多少次用刀攻击或刺伤别人？	从来没有	有时	很多次	几乎每天



你是否对每一个问题标上一个答案，即使你的答案是“从来没有”？

例 1.5 提供了一个例证，说明问题的格式有可能使被访者迷惑不解。此例中，尽管要求和提醒被访者回答所有问题，但他们只回答了一个问题。其他项缺少答案说明什么？被访者拒绝回答这些问题是因为他们从未有过问题中所描述的行为吗？如果是这样，为什么被访者不选择答案“从来没有”事实上对被访者来说他们只回答认为与他们有关的问题是很常见的事。事实是例 1.5 中所引用的问卷的调查者可能是为了节省空间、避免重复罗列答案选择，而选用这种列表格式。但是，被访者可能发现这种格式令人迷惑不解，他们更习惯于看例 1.6 所示的问卷。

#### 例 1.6 相对容易理解的问卷格式

去年你有多少次用刀攻击或刺伤别人？

从来没有

有时

很多次

几乎每天

令人迷惑的问题格式导致缺失值，因为它们使被访者不知道应该如何回答问题。你可以通过事先进行广泛的认知预测试和试访，避免某些引起被访者误解的调查问题。认知预测试是对潜在被访者进行的访问，其中，你要求他们基于你的调查方法解读每一个问题和回答选项。试访是