

# 白话统计学

(第3版)

Statistics in Plain English (Third Edition)

蒂莫西·C·厄丹 (Timothy C. Urdan) 著  
彭志文 译

## 关于本电子书说明

本人由于一些便利条件，可以帮您提供各种中文电子图书资料，且质量均为清晰的PDF图片格式，方便阅读和携带。文学、法律、计算机、人文、经济、医学、工业、学术等方面的图书，都可以帮您找提供电子版本，500万图书馆资源收藏供你选择。

我的QQ是859109769 佳佳e图书（提供完整版）



中国人民大学出版社



# 白话统计学

(第3版)

Statistics in Plain English (Third Edition)

本书对统计学基本原理的阐释格外清晰明了、引人入胜，是我目前所见最为出色的，对于教师和学生都大有裨益。

——安德鲁·J·埃利奥特 (Andrew J. Elliot)，美国罗彻斯特大学

本书是献给学生、教师和研究者的礼物。厄丹用直白的语言清晰阐释了统计问题，从基本的统计原理到复杂的多变量技术，几乎无所不包，且易懂易学，可谓独树一帜。

——埃里克·M·安德曼 (Eric M. Anderman)，美国俄亥俄州立大学

这是一本极具亲和力的书，它十分明了地介绍了统计原理和统计术语。厄丹对初学者的困惑感同身受，用浅显的语言和恰当的例子为读者扫清了学习的障碍。

——阿维·卡普兰 (Avi Kaplan)，美国天普大学

这是迄今为止市面上最好的统计学启蒙读本，我四处向人推荐，不只是学生，还有同事。它轻松简洁，既适合不同层次的学生，也通用于不同学科。它解释概念的方式比市面上大多数教材要高明得多。

——凯瑟琳·A·罗斯特 (Catherine A. Roster)，美国新墨西哥大学

ISBN 978-7-300-18573-6



9 787300 185736 >

定价：45.00元



且学 · 且思 · 且行

# 白话统计学

(第3版)

Statistics in Plain English (Third Edition)

蒂莫西·C·厄丹 (Timothy C. Urdan) 著

彭志文 译

中国人民大学出版社

· 北 京 ·

图书在版编目 (CIP) 数据

白话统计学: 第3版/厄丹著; 彭志文译. —北京: 中国人民大学出版社, 2013.12  
(管理者终身学习)

ISBN 978-7-300-18573-6

I. ①白… II. ①厄… ②彭… III. ①统计学 IV. ①C8

中国版本图书馆 CIP 数据核字 (2013) 第 307569 号

管理者终身学习

白话统计学 (第3版)

蒂莫西·C·厄丹 著

彭志文 译

Baihua Tongjixue

---

出版发行 中国人民大学出版社

社 址 北京中关村大街 31 号

邮政编码 100080

电 话 010-62511242 (总编室)

010-62511398 (质管部)

010-82501766 (邮购部)

010-62514148 (门市部)

010-62515195 (发行公司)

010-62515275 (盗版举报)

网 址 <http://www.crup.com.cn>

<http://www.ttrnet.com>(人大教研网)

经 销 新华书店

印 刷 北京东君印刷有限公司

规 格 175 mm×250 mm 16 开本

版 次 2013 年 12 月第 1 版

印 张 15.25 插页 2

印 次 2013 年 12 月第 1 次印刷

字 数 264 000

定 价 45.00 元

---

版权所有 侵权必究

印装差错 负责调换



## 为什么使用统计学？

我是一名经常使用统计学的研究人员，也是谈话类广播节目的热心听众，我发现自己每天都会冲着收音机大喊大叫。尽管我明白这些喊叫毫无作用，但还是不能自己。电台谈话节目主持人、夸夸其谈的政治人物以及公众都知道，再没有什么比个人经验更有效果和说服力了。统计学家把这称为“奇闻轶事”的证据。我经常举一桩陈年旧事来做例子：本地国会议员办公室曾给我寄过一本小册子，上面对公共教育状况大加抨击，我致电这间办公室来表达不满。负责教育事务的工作人员接待了我。我告诉他，根据不同来源报告的统计数据来看，许多迹象表明我们的系统表现良好，高中毕业率上升，大学生人数增加，标准化测试成绩提高，所有族裔的学术能力评估测试成绩都有一定程度的改善，等等。这名工作人员告诉我，即便统计数据果真如此，她也仍然相信我们的公立高中大不如前，因为她与父亲上同一所高中，而父亲受到的教育更好。听罢，我气得挂断电话，又一次大喊大叫起来。

许多人对统计数据有着普遍的不信任感，觉得狡猾的统计学家总能“令统计数据说出他们想听的话”，或者“用统计学撒谎”。事实上，研究人员如果计算正确的话，就不能随心所欲。统计数据只会说它能说的，而且从不撒谎。但狡猾的研究人员能以不同方式解释统计数据的含义。不懂统计学的人要么对统计学家和研究人员给出的解释全盘接受，要么一概拒绝。我相信更好的选择是去了解统计学原理并用它解释自己的见闻。本书的目的就是让统计学变得更容易理解。

## 统计学的用途

“奇闻轶事”数据的一个潜在缺陷是其特殊性。议员办公室的那位工作人员告诉我，她与父亲上同一所高中，父亲接受的教育比她好，而我却毫不费力地受到了比父辈更好的教育。统计学使研究者可以从众人身上收集信息或数据，然后概

括出他们的典型经验。究竟大多数人受到的教育比他们的父母更好还是更差呢？统计学使研究者收集大量数据并把数据概括成一些数字，例如平均数。当然，把众多数据概括成一个数，难免损失了大量信息，掩盖了不同人的不同经历，所以要切记，统计学在大多数情况下不能对任何个别经验提供有用信息。但研究者一般使用统计学得出关于某个总体的一般性结论。虽然个人经验经常令人感动或引人入胜，但理解典型的或平均的经验往往更加重要。正因如此，我们才需要统计学。

统计学也用来得出有关不同分组之间整体差异的结论。例如，我家有四个孩子，两男两女，女比男高。这一个人经验可能令我得出以下结论：女人通常比男人高。当然，我们知道，平均而言，男人比女人高。我们之所以知道这一点，是因为研究者已经随机抽取了大量男女样本，并比较了他们的平均身高。研究者经常对诸如此类的比较产生兴趣：癌症病人服用一种药物是否比服用另一种药物存活的时间更长？用一种方式教孩子阅读是否比用另一种方式更有效？对某部电影的感受是否男女有别？为了回答这些问题，我们需要从随机选取的样本中收集数据，并用统计学比较这些数据。从此类比较中得出的结论通常更为可信，而非随机样本的简单观测中得出的结论则不然，例如我家男女的身高差异。

统计学也用来考察两个变量的取值是否相关并进行预测。例如，统计学能用来考察吸烟与罹患肺癌的可能性之间是否相关。长期以来，烟草公司声称吸烟与罹患癌症之间并无关系。固然有些吸烟的人患上了癌症，但也有许多吸烟的人并未患上癌症，而且吸烟的人往往干一些可能致癌的其他事情，比如食用不健康食品和缺乏锻炼。研究者借助统计学工具进行了大量的研究，最终更多有力的证据表明，吸烟与罹患癌症之间确有关系。由于统计学倾向于关注普遍模式而非个别情况，所以这种研究并不意味着所有吸烟的人都会罹患癌症。但研究表明，平均而言吸烟的人比不吸烟的人患上癌症的概率更大。

只要稍加思索，你就能想到大量有趣而重要的“关系”问题，是统计学能够帮助你解答的。自负与学术成就之间有无关系？刑事被告的相貌与定罪可能之间有无关系？能否根据各州在戒毒项目上的支出金额来预测该州的暴力犯罪率？如果我们已知父亲的身高，那么预测儿子的身高有多大把握？研究者利用统计学来确定总体之中的变量间关系，从而考察上述问题以及成千上万的其他问题。

## 如何使用本书？

本书无意让初学者在学习统计学时“毕其功于一役”。社会科学的统计学课程

如果要使用更加详细的推荐教材，不妨将本书作为补充读物。或者，如果你已经学过一两门统计学课程的话，将本书作为参考书，用来复习巩固学过的统计学概念，也将大有裨益。千万不要忘记，本书比传统教材要精炼得多！书中讨论的概念相当复杂，而表达却力求简明扼要，两相或难免有所冲突。想要更加全面地理解这些概念的话，只需参阅更传统、更详细的教科书即可。

告诉大家我的提醒之后，接下来该说说本书的潜在好处以及怎么将这些好处充分发挥了。作为一名统计学的研究者和教师，我发现统计学教科书中总是包括大量的技术性内容，这令那些非专业统计学家的人士望而生畏。虽然我刚说过这些信息是重要的，但有时候简要直白地描述一个统计量的适用条件和解释方式则更为实用。对于统计学课程的初学者、那些对“数学倾向”不感兴趣的人以及多年前学过统计学现在需要复习一下的人而言，更是如此。本书的写作目的是精炼、直白地描述和解释一些统计量，使之容易阅读和理解。

为了帮助读者以一种“各取所需”的方式使用本书，我将每一章内容分成三部分。第一部分给出统计量的简单（1~2页）描述，包括统计量的用途及其提供的信息。第二部分包括稍微多点（3~8页）的关于统计量的讨论。在这一部分中，对如何使用统计量、如何利用公式计算统计量、统计量的优缺点以及使用统计量必须满足的条件都提供了更多的信息。最后，每章的结尾都举例说明统计量的应用及解释。

开卷之前如果注意到本书的以下三个特点将有助于阅读。首先，某些章不止讨论一个统计量。例如，第2章描述了度量集中趋势的三个统计量：均值、中位数和众数。其次，某些章的内容是统计概念而不是具体的统计技术。例如，第4章讨论正态分布。也有章节讨论统计显著性和统计交互作用。最后，书中各章不一定非得循序阅读。本书组织内容的原则是越基本的统计量和统计概念越安排在靠前的章节，越复杂的概念越出现在较后位置。但也并非只有读过前一章才能理解后一章，而是每章自成一体。这样一来，读者可以根据需要选读各章。例如，如果你在统计学课上已经理解了 $t$ 检验，但对单因子方差分析不太明白，你就可以越过 $t$ 检验一章（第9章），直接跳到方差分析一章（第10章）。

## 这一版的新特点

《白话统计学》（第三版）新增和修订了不少内容。最大的改动是增加了关于数据整理组织技术、因素分析和信度分析一章（第15章）。这些在社会科学的统

计应用中十分普遍，对于那些使用调查方法的研究者更是如此。另外，第1章也新增了关于理解数据分布的一节，并增加了几幅图表以帮助理解如何使用和解释图表。本书在许多章之后都增加了“行文表述”一节，以说明在出版的论文、书籍或专著中如何表述统计量。这将有助于读者将自己的结果写入出版物或者阅读他人的著作。第三版设有配套网站 <http://www.psypress.com/statistics-in-plain-english/>，包括每章的课件汇总、多数章的互动习题集以及进一步学习统计学的有用网站链接。最重要的是，我订正了本书前一版本出现的全部错误。当然，我也可能在撰写这一版时发生了一些新错误，读者不可掉以轻心哟。

统计学是帮助人们认识有意义现象的强大工具。无论你是一名学生、一位研究者，或者只是一个有兴趣理解周围世界的公民，统计学都可以提供一种方法来帮助你弄清楚身边的环境。本书用大白话写就，更便于非统计专业人士利用统计学的功能。我希望读者发现它确实有用。

## 致 谢

首先，泰勒弗朗西斯集团下属劳特利奇出版社（Routledge/Taylor & Francis Group）的 Debra Riegert 这些年来好主意不断并且经常请客吃饭，我早就应该感谢她了。接下来，我虽有些不情愿但仍诚挚地感谢本书第三版的评论者：亚拉巴马大学的 Gregg Bell、新墨西哥大学的 Catherine A. Roster 以及一位匿名评论人。我不善于听取批评意见，但为了读者的利益，最终还是认识到忠言逆耳并采纳了大多数建议。在准备本书的各个版本时，我主要依靠几位学生提供协助。对这一版协助最大的是 Sarah Cafasso, Stacy Morris 和 Louis Hung。最后，感谢 Jeannine 使我有时间写作，感谢 Ella 和 Nathaniel 让我没把大好时光全耗在工作上。

## 第1章 导论：社会科学研究的原理和术语 1

总体和样本，统计量和参数 1

抽样问题 3

变量类型和测量尺度 4

研究设计 6

分布和图表的重要性 7

总结与展望 12

第1章的术语表 12

## 第2章 中心趋势的测度 15

中心趋势测度详解 16

例子：偏态分布的均值、中位数和众数 17

行文表述 20

总结与展望 20

第2章的术语和符号表 21

## 第3章 变异程度的测度 23

变异程度测度详解 25

例子：考察极差、方差和标准差 29

总结与展望 33

第3章的术语和符号表 33

## 第4章 正态分布 35

正态分布详解 37

例子: 非正态分布中应用正态分布概率 39

总结与展望 41

第4章的术语表 41

## 第5章 标准化与 $z$ 分数 43

标准化与 $z$ 分数详解 44

例子: 比较原始取值和 $z$ 分数 52

总结与展望 54

第5章的术语和符号表 55

## 第6章 标准误 56

标准误详解 57

例子: 样本容量和标准差对标准误的影响 66

总结与展望 67

第6章的术语和符号表 68

## 第7章 统计显著性、效应量和置信区间 69

统计显著性详解 70

效应量详解 76

置信区间详解 80

例子: 关于动机的单样本 $t$ 检验——统计显著性、置信区间和效应量 82

总结与展望 85

第7章的术语和符号表 86

推荐读物 87

## 第 8 章 相关性 88

皮尔逊相关系数详解 90

其他类型相关系数概述 98

例子：评级和考试分数之间的相关性 99

行文表述 101

总结与展望 101

第 8 章的术语与符号表 102

推荐读物 103

第 9 章  $t$  检验 104独立样本  $t$  检验详解 105配对或相依样本  $t$  检验详解 110

例子：比较男女生的学分绩点 112

例子：比较五年级与六年级的学分绩点 114

行文表述 115

总结与展望 116

第 9 章的术语和符号表 116

## 第 10 章 单因子方差分析 118

单因子方差分析详解 119

例子：5 岁、8 岁和 12 岁孩子的偏好比较 127

行文表述 131

总结与展望 131

第 10 章的术语和符号表 132

推荐读物 133



## 第 11 章 因子方差分析 134

因子方差分析详解 135

例子：表现、选择以及公开或私密评价 144

行文表述 146

总结与展望 146

第 11 章的术语表 147

推荐读物 148

## 第 12 章 复测方差分析 149

复测方差分析详解 152

例子：关于标准化测试的态度改变 158

行文表述 163

总结与展望 164

第 12 章的术语及符号表 164

推荐读物 165

## 第 13 章 回 归 166

回归详解 167

多元回归 173

例子：预测自我妨碍策略的使用 179

行文表述 181

总结与展望 182

第 13 章的术语与符号表 182

推荐读物 183

## 第 14 章 卡方独立性检验 185

卡方独立性检验详解 186

例子：世代状态与成绩水平 189

行文表述 191

总结与展望 191

第 14 章术语及符号表 191

## 第 15 章 因子分析与信度分析：数据整理技术 193

因子分析详解 194

探索性因子分析：一个更具体的例子 197

信度分析详解 203

行文表述 206

总 结 207

第 15 章的术语和符号表 207

推荐读物 209

附录 A 正态分布曲线下  $z$  两侧的面积 210附录 B  $t$  分布的临界值 213附录 C  $F$  分布的临界值 215

## 附录 D 学生化极差统计量的临界值（用于 Tukey HSD 检验） 220

## 附录 E 卡方分布的临界值 223

## 参考文献 224

## 符号表 226

## 译后记 229

# 导论：社会科学研究的原则和术语

我读研究生时经常听一位统计学教授开玩笑说，“让人不知所云，正是统计学的应有之义”。不幸的是大多数同学果然不知所云，甚至没听出玩笑的意思。与其他专业领域一样，统计学以及社会科学研究领域有其专门的术语、语言和惯例。本章讨论了一些基本的研究原理和术语，包括样本和总体的区别、抽样方法、变量类型以及推断统计和描述统计的区别。最后简要介绍了不同类型的研究设计。

## 总体和样本，统计量和参数

**总体**（population）是一个或一组对象，代表了感兴趣的特定分组或类别的所有成员。样本是从更大的总体中抽取的子集（见图 1—1）。例如，如果我想知道哈佛大学全职终身教师当前的平均收入，有两种方法能得到这个平均数。一种方法是找到哈佛大学全职终身教师的完整名单，再找出名单上每一位成员的年收入。因为这份名单包括了我所感兴趣分组的所有成员，所以它能被当作总体。如果我收集了这些数据并计算均值，则得到一个**参数**（parameter）。参数是来自总体并适用于总体的值。了解哈佛大学终身教师平均收入的另一种方法是从名单中随机选取教师的一个子集，然后计算这个子集的平均收入。这个子集就是**样本**（sample）（此例中是**随机样本**（random sample）），从样本中计算出的均值是一种**统计量**（statistic）。统计量是从样本数据中计算出的值，而参数是从总体数据中计算出并适用于总体数据的值。

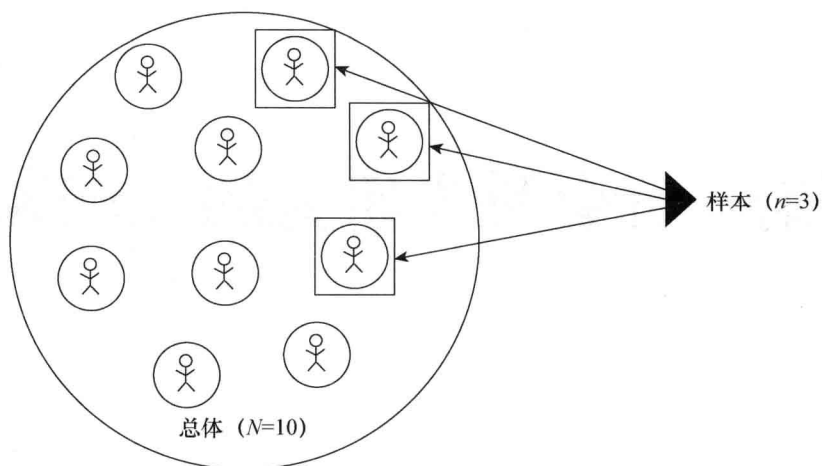


图 1—1 总体以及从总体中抽取的样本

关于样本和总体,要注意以下事项。首先,总体不一定包括大量对象。例如,如果我想知道本学期统计学课上学生的平均身高,那么课上所有的同学共同构成总体。如果我的课上只有 5 名学生,则我的总体就只包括 5 个对象。其次,总体(以及样本)所包括的对象不一定是人。例如,假设我想知道去年兽医院接诊的狗的平均年龄,则研究的总体就是由狗而不是人所组成的。同样,我可能想知道 2005 年在美国组装的福特汽车所排放的一氧化碳总量。在这个例子中,我的总体是汽车,但不是全部汽车,而是仅限于某一特定年份在某一特定国家所组装的福特汽车。

再次,研究者通常要定义总体,或者以明确的方式,或者以隐含的方式。上面的例子中,我明确地定义了(狗和汽车的)总体。然而,研究者通常不以十分明确的方式来定义总体。例如,一个研究者说他的研究目的是考察青少年抑郁症的发生率,但他的样本只包括了康涅狄格州一家心理卫生服务机构在某一年中所接待的 15 岁孩子。这就造成了潜在的问题,并且直接引入了关于样本和总体的第四个也是最后一个小的注意事项:样本不一定是抽样总体的适当代表。青少年抑郁症发生率的例子中有两个潜在的总体。一个是研究者本意要研究的,也是其研究问题内在要求的,即青少年。但青少年是一个非常大的群体,包括所有国家年龄在 13~20 岁之间的所有人。而由所选样本定义的另一个总体则要具体得多:某一特定年份里康涅狄格州一家心理卫生服务机构接待的 15 岁孩子。

### 推断统计和描述统计

搞清楚研究感兴趣的总体究竟是两者中的哪一个这个问题为什么如此重要呢?因为该项研究的使用者必须能够确定得自样本的结果究竟能在多大程度上推广(generalize)到更大总体。显而易见,康涅狄格州心理卫生服务机构接待的 15 岁

孩子中抑郁症的发生率可能与其他青少年不同。比如，平均而言，去心理卫生服务机构咨询的青少年要比那些没有寻求心理医生帮助的青少年更有可能患上抑郁症。同样，作为一个整体，康涅狄格州的青少年或许比加利福尼亚州的青少年更容易抑郁，加州的阳光和米老鼠可以令人心情变好。15岁的孩子刚上高中，还不能合法驾车，也许心里头觉得憋屈，16岁的孩子就不同了，可以自己开车兜风了，所以15岁的孩子可能比16岁的孩子更容易抑郁。总而言之，有很多理由怀疑，研究中没包括的青少年与研究中已经包括的青少年在抑郁症发生率方面可能不同。如果存在这种差异，那么从样本中得出的结果就难以应用到更大的总体中。用研究术语来说，结果不能由样本推广到总体，尤其是总体没有明确定义时。

为什么有必要推而广之？回答这一问题需要引入**描述统计**（descriptive）和**推断统计**（inferential）的区别。描述统计只能应用于从中收集数据的样本对象或总体成员。相反，推断统计是指假定样本能够代表更大的总体，从而利用样本数据得出关于总体特征的一些结论（即进行推断）。尽管有些时候研究者关心的只是对样本特征的描述，但绝大多数时候我们更关心的是从样本中得到的关于抽样总体的信息。在抑郁症的研究中，研究者对所选样本自身的抑郁水平并不是特别看重，而是想利用样本数据得出关于青少年总体抑郁水平的一些结论。研究者务必确信样本能够准确代表总体，否则就无法实现从样本数据到总体推断的“跨越”。这个过程的第一步非常重要，那就是明确定义样本所要代表的总体。

## 抽样问题

研究者选取样本的方式很多。其中最有用也最难实现的是**随机抽样**（random sampling）。“随机”一词在统计学中的意义远比日常使用中的具体。它并不意味着随意。用统计术语来说，“随机”意味着总体中的每一个对象被选入样本的概率相等。随机抽样的最大好处是样本与抽样总体之间的差异不是系统性的。在抑郁症研究的例子中，样本与总体之间存在着重要的系统性（即非随机的）差异。例如，研究者从心理卫生服务机构接待的孩子中选取样本，于是很有可能选择了比普通青少年更有可能抑郁的孩子。尽管随机选取的样本可能在很大程度上不同于更大的总体（特别是样本较小时），但这些差异是或然机会使然，而并非选择过程中的系统性偏差。

**典型抽样**（representative sampling）是选择研究对象的第二种方法。使用这种方法时，研究者有意选取在具体特征上与更大总体相匹配的对象。例如，我想

要进行一项研究，考察旧金山成年人的平均年收入，总体被定义为“旧金山的成年人”，这一总体包括很多子集（不同族裔、男女、退休成年人、残障成年人、有配偶和子女的成年人、单身成年人等）。我们可以预期不同的子集有不同的收入。要获得旧金山成年人口收入的准确信息，就得选取一个能够很好地代表总体的样本。于是，应该尽力做到样本中各组的比例与总体中各组的比例相匹配。例如，如果旧金山成年人口中有15%已经退休，我选取样本时也应该包括15%的退休人口。同样，如果旧金山成年人口中有55%是男性，样本中也应该有55%的男性。随机抽样可能得到与总体相似的样本，也可能得到与总体不相似的样本，而典型抽样则能够确保样本与总体在一些重要变量上相似。这一类抽样程序费时费力，但样本结果更可能推广至总体。

选取样本的另一种常见方法是**方便抽样**（convenience sampling）。使用方便抽样时，研究者通常根据地理距离、接触难度和参与意愿（即方便程度）来选择样本对象。例如，我想研究八年级学生的成就水平，可以从离我办公室最近的初中选取包括200名学生的样本。在这间学校里，我询问了300名八年级学生的家长，只有220名学生家长同意接受调查，发放问卷那天到校学生人数是200名，最终是从他们身上收集的数据。这就是一个方便样本。尽管这种样本选取方法比选取随机样本或典型样本要更省劲，但它不一定就是一种不好的抽样方法。如果我的方便样本与感兴趣的总体之间的差异不至于影响研究结果，那么方便抽样就不失为一种完全可接受的抽样方法。

## 变量类型和测量尺度

社会科学研究使用大量术语来描述不同类型的变量。**变量**（variable）几乎可以是能被编码的任何东西，并且具有不止一个取值（例如，收入、性别、年龄、身高、对学校教育的态度、抑郁指标的分值）。相反，**常量**（constant）具有唯一的取值。例如，如果一个样本中的每个对象都是男性，则“性别”分类就是一个常量。变量类型包括**定量**（quantitative）（或**连续**（continuous））变量和**定性**（qualitative）（或**分类**（categorical））变量。定量变量用数字或评分来赋值，表示某种数量。例如，身高是一个定量（或连续）变量，因为该变量的取值越大，表示身高越高。相反，定性变量的赋值并不意味着特定性质的多寡。如果我进行一项研究，比较缅因州、新墨西哥州和怀俄明州居民的饮食习惯，那么“所在州”变量具有三个值（即1=缅因州，2=新墨西哥州，3=怀俄明州）。注意该变量取

值为3时并不比取值为1或2时更大，只是不同而已。数字符号代表地理位置上的性质差异，并非数量差异。社会科学研究中经常使用的定性变量是**二值变量**（dichotomous variable）。这是一种具有两个不同分类的变量（如男女）。

大多数统计学教科书都描述了变量的四种不同测量尺度：定类、定序、定距和定比。**定类变量**（nominally scaled variable）利用无权重或无数值的符号以识别变量的不同水平。例如，研究者经常想要考察在一些变量（如收入）上是否存在男女差异。大多数计算机软件处理统计数据时都要求这种性别变量用数字赋值以代表分组。例如，男性用“0”表示，女性用“1”表示。此时，取值1并不意味着比取值0具有更高评价。0和1只不过是代指各组的名称或标签而已。

**定序变量**（ordinal variable）的取值则含有权重。我若想知道美国十大富豪，令最富有的人取1，第二富的人取2，依此类推，直到取10。这一赋值系统告诉我，这10个最富有的美国人各自相对于其他人的排位如何（例如，比尔·盖茨是1；奥普拉·温弗瑞是8，等等），但没有提供关于取值间距离的信息。于是，我知道首富比次富更富有，但不知道前者的财富比后者多1美元还是多10亿美元。与之不同，以**定距尺度**（interval）和**定比尺度**（ratio）来赋值的变量则包含关于相对值和距离的信息。例如，如果我知道了样本中一个对象高58英寸，一个对象高60英寸，第三个对象高66英寸，那么我就知道了样本中谁最高以及各对象比其他对象高多少或矮多少。因为高度变量用英寸测量，而所有的英寸都有相同的长度，于是高度变量使用相等间距的尺度进行测量，从而提供了关于相对位置和距离的信息。无论定距尺度还是定比尺度，所用测度的各单位间距离均相等。定比尺度还包括一个零值（例如摄氏温度）。图1—2说明了定序测量尺度与定距或定比测量尺度之间的差异。

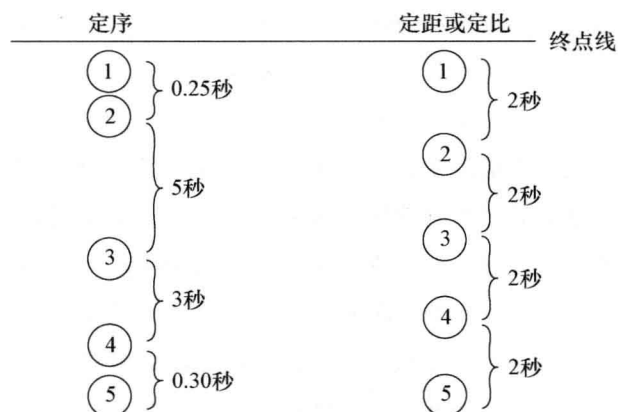


图1—2 定序测量尺度与定距或定比测量尺度之间的差异



## 研究设计

社会科学家用到很多种研究方法与设计。有时研究者使用**实验设计**（experimental design）。在这一类研究中，实验者将样本中的案例分成不同组，然后就感兴趣的一个或多个变量进行组间比较。例如，我可能想知道新的数学课程方案是否比旧方案更好。选取一个由 40 名学生组成的样本，**随机分配**（random assignment）其中 20 名使用旧课程方案，另外 20 名则使用新课程方案。然后检验各组，看哪组学到的数学概念更多。由于使用随机分配方式将学生分成两组，所以我希望两组间的任何重要差异都在两组之间平均分布，从而两组考试分数的任何差异只能归因于两种课程教学方案的效果差异。当然，事实可能并非如此。

**相关性研究设计**（correlational research designs）也是社会科学中常用的研究方法。在这类研究中，参与者通常不经过随机分组，研究者一般也不能施加实际控制。相反，研究者只能收集若干变量的数据，然后进行某些统计分析以确定不同变量之间彼此相关的强度。例如，我感兴趣的问题可能是：雇员的生产率是否与其（在家而非上班时间的）睡眠时间相关？于是，我选取了一个包括 100 名成年工人在内的样本，测量其工作生产率以及给定一周内平均每晚的睡眠时间。也许我会发现睡眠时间与生产率之间存在强相关关系。现在我想从逻辑上论证这说得通，因为工人只有休息好才能努力工作，也才更有效。尽管这一结论言之成理，但仅凭具有相关性的数据就得出如此结论则过于穿凿。相关性研究只能提供变量间是否相关的信息，而不能得出关于因果关系的结论。别忘了，还有一种可能是更有效的工作导致了在家睡眠时间更长。也许顺利完成工作可以舒缓压力，也许可以让工人早上多睡一会儿，不管哪种情况都会造成更长时间的睡眠。

实验研究设计使研究者能够将导致**因变量**（dependent variables）波动或改变的特定**自变量**（independent variables）分离出来。前面的例子中，数学课程方案是自变量，学生考试成绩是因变量，鉴于自变量在我的控制之下，从而能够合理认定所用数学课程方案类型影响学生考试分数的结论。实验设计的主要缺陷是很难在不受干扰的条件下完成，从而难以在真实世界的情形中得到一般化。例如，在前面的研究中，我很难保证数学课程方案是影响考试成绩的唯一因素，还有一些与课程方案无关的其他因素能够影响考试成绩，比如两组学生在数学学习能力上原本就有的差异，或者教师风格（思路清晰或热情投入）的差异。相关性研究设计的优势是往往比实验研究更易于实施，能够相对容易地包括多个变量，并允

许研究者同时考察多个变量。相关性研究的主要缺陷是无法施加精准控制，而精准控制却是得出关于变量之间因果联系的结论所必需的。

## 分布和图表的重要性

统计学家花费大量时间来讨论**分布** (distributions)。简言之，分布就是变量数据或取值的一个集合。通常，这些取值按照从小到大的顺序排列，并以图表形式进行展示。鉴于分布在统计学中的重要性，我将有关内容安排在本书的较前部分，举例说明若干不同类型的分布及其图形表示。本书后面部分使用大量整章篇幅来讲解统计学中最常用的几种分布，包括**正态分布** (normal distribution) (第4章和第5章)、**t分布** (*t* distribution) (第9章和第7章的一部分)、**F分布** (*F* distribution) (第10章、第11章和第12章)以及**卡方分布** (chi-square distribution) (第14章)。

先看一个简单例子。假设我正在进行一项关于选民态度的研究，选取500名选民组成的随机样本用于研究。我想知道的一则信息是样本成员的政治背景，于是询问他们是共和党、民主党还是无党派，结果发现样本中45%的成员是民主党，40%的成员是共和党，15%的成员属于无党派。政治背景是一个名义变量或者分类变量。由于名义变量是只有类别而无数值权重的变量，因此不能从高到低安排这一分布的取值。共和党员的取值不比民主党员或无党派人员的取值更多或更少，它们仅仅是不同的类别而已。所以我没有试图按照取值从低到高来组织数据，而只是将其作为不同类别加以对待，并报告样本对象中归入各类别的百分比。

用图表来表示这一分布有许多种不同的方法，包括饼图、条形图、柱状图、气泡图等。选择适当的图形表示，关键在于切记使用图形的目的是使数据易于理解。我作了两个不同的图来描述政治背景的分布。两个图都是不错的选择，因为两个图都能清楚简明地概括分布情况且易于理解。图1—3用柱状图来描述分布，而图1—4用饼图来展示数据。究竟哪个图最适合这些数据，将取决于个人偏好。如图1—3所示，*x*轴（横轴）表示政党取向，民主党、共和党 and 无党派，*y*轴（纵轴）表示样本比例。只需打量一下柱形，各组的百分比就一目了然，从中不难看出样本中哪种政党取向具有最高比例，以及各政党取向在样本比例方面的差异。图1—4中的饼图展示了相同的信息，只不过在我看来，后者更鲜明、更简洁罢了。

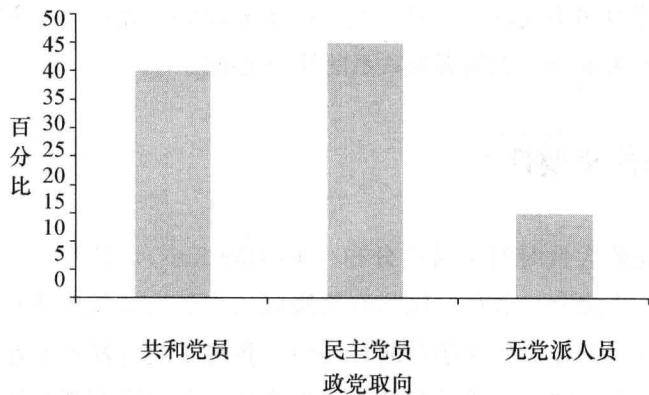


图 1—3 共和党员、民主党员和无党派人员的分布柱状图

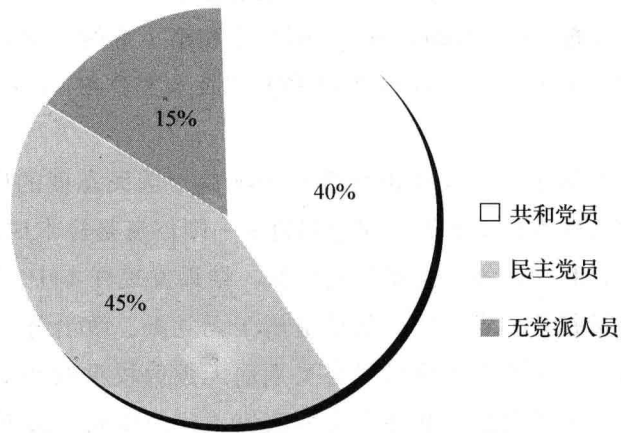


图 1—4 共和党员、民主党员和无党派人员的分布饼图

有时候研究者感兴趣的是同时考察不止一个变量的分布。例如，假设我想了解看电视的时间与做家庭作业的时间之间的联系，尤其对这种联系在不同国家的表现感兴趣。于是从几个不同国家的高中生样本中收集数据，现在得到了在 5 个不同国家（美国、墨西哥、中国、挪威和日本）中 2 个不同变量的分布。为了比较这些不同国家，我决定对每个国家的各个变量计算平均数或均值（mean）（见第 2 章），然后用柱状图来表示这些均值，如图 1—5 所示（需要注意这些数据是我编造的虚构数据）。该图清楚显示，每天平均看电视时间与平均做作业时间之差在美国最大，墨西哥次之，而在中国两者之差为零。根据虚构数据，挪威和日本的高中生做作业的时间比看电视的时间更长。如你所见，用一个简单的图形概括一个复杂的数据集合倒也并非难事。

用图形描述取值分布的另一种常用方法是折线图，如图 1—6 所示。假设我选取了一个随机样本，包括 100 名刚刚完成第一学期课程的大一新生，询问他们每

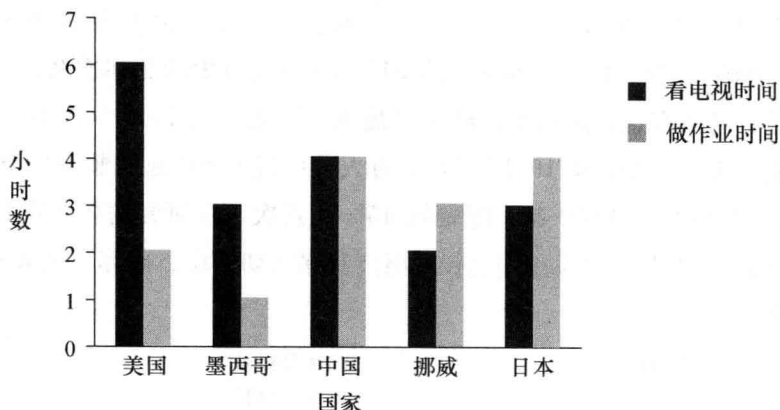


图 1—5 5个国家高中生平均看电视时间与平均做作业时间的比较

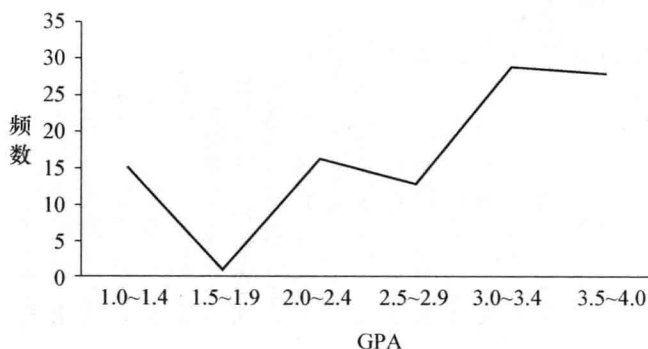


图 1—6 按不同学分绩点分组的学生频数折线图

门课的最终成绩，然后计算每人的平均学分绩点（GPA），再将平均学分绩点分成6组：1~1.4，1.5~1.9，2.0~2.4，2.5~2.9，3.0~3.4，3.5~4.0。在算出各组中包括的学生人数之后用折线图描述这些数据，结果如图1—6所示。沿 $x$ 轴列出了6个不同的学分绩点分组， $y$ 轴代表频数（frequency），通常用符号 $f$ 表示。这样，该图的 $y$ 轴显示了各个学分绩点分组包括的学生人数。从图1—6中反映出，大学第一学期学习吃力者为数不少（13人），他们只取得1.0~1.4的学分绩点。接下来学分绩点在1.5~1.9之间的一组只有1人。从这组开始，各组学生人数大致递增，学分绩点在2.0~2.9之间的有31人，在3.0~4.0之间的有55人。与此类似的折线图是发现数据趋势的便捷方法，既可以是历时趋势，也可以是跨类趋势。在这个学分绩点的例子中，除了一组为数不少的学习吃力者之外，大致趋势是学分绩点越高的分组包括的学生越多。

柱状图是另外一种明确显示数据趋势的方法。图1—7所示的叠加柱状图能够在单个图形中展示几种信息。例如，该图描述了1990—2007年间两种不同类型犯

罪(财产犯罪和暴力犯罪)的发生率。 $x$ 轴表示年份,由左及右依次为从早(1990年)到晚(2007年)。 $y$ 轴表示美国每10万人口中的犯罪次数。以这种方式展示以后,若干有意思的事实就凸显了出来。首先,从1990年到2007年的整体趋势是犯罪大幅减少。从1991年每10万人口中近6000起犯罪减少至2007年每10万人口中略低于4000起,降幅近40%。其次,在研究的各年中暴力犯罪(如谋杀、强奸、殴打)的发生率比侵犯财产犯罪(如盗窃、破坏、纵火)的发生率要低得多。

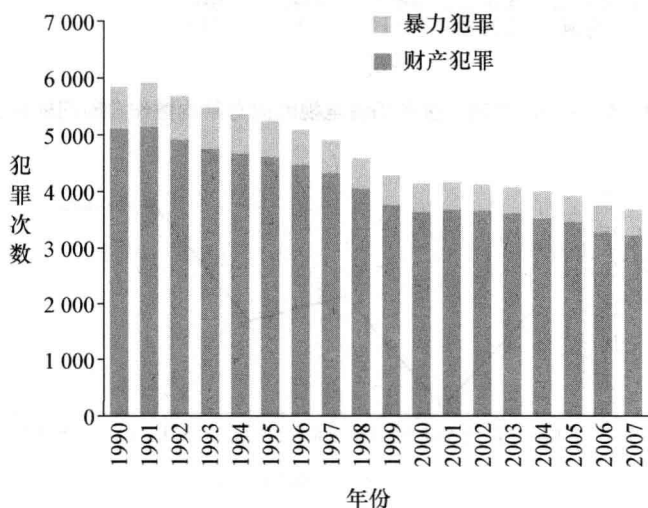


图 1—7 1990—2007 年间的犯罪率叠加柱状图

从图 1—7 中不难发现,1990—2007 年间犯罪率呈整体下降之势,但暴力犯罪率是否显著下降却不那么显而易见。这是因为暴力犯罪在全部犯罪中所占的比例远小于财产犯罪,而  $y$  轴所用的刻度太大。柱形的上面一段代表暴力犯罪,看上去很小。为了更好地理解暴力犯罪的历时趋势,我新作了一个图,如图 1—8 所示。

新图与图 1—7 叠加柱状图展示了完全一样的数据。折线图将暴力犯罪与财产犯罪完全分离开来,以便于比较两种类型犯罪率之间的差异。该图又一次清楚地显示出财产犯罪率随时间下降,但暴力犯罪率是否随时间显著下降却仍不清楚。如果仔细观察可以发现,暴力犯罪率从 1990 年每 10 万人约 800 件下降到 2007 年的每 10 万人约 500 件。这是相当可观的犯罪率下降,但如果不仔细留意的话,就很难看出来。不要忘了,使用图表的目的是让数据中包含的有意义的事实显而易见。如果必须费劲才能看明白,那么就说明这个图形不够好。

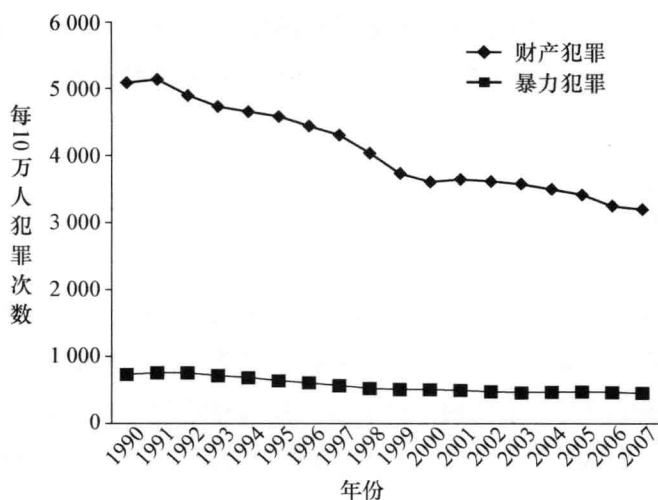


图 1—8 1990—2007 年间的犯罪率折线图

图 1—8 和图 1—7 的问题一样，y 轴的刻度太大，以至于不能清楚显示暴力犯罪率的历时趋势。解决这一问题需要变换刻度以适合暴力犯罪率数据。于是，我又作了另外一幅图（见图 1—9）来单独描述暴力犯罪数据，而没有包括财产犯罪数据。前图中的 y 轴刻度是 0~6 000 或 0~7 000，新图中的 y 轴刻度是 0~800。在新的柱状图中，1990—2007 年间暴力犯罪率同样显著下降，这一事实变得一目了然了。

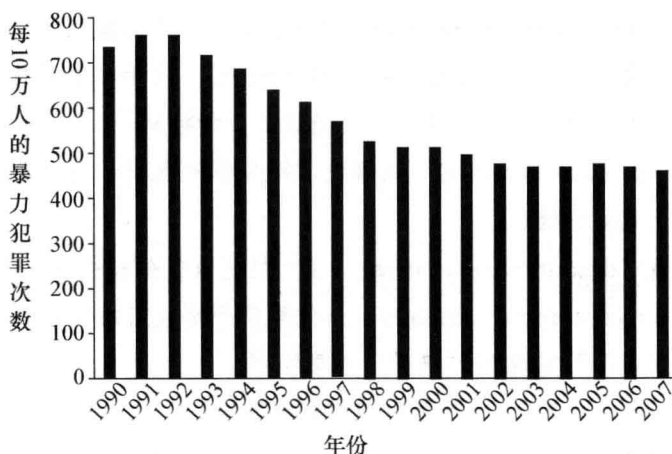


图 1—9 表示 1990—2007 年暴力犯罪率的柱状图

无论是哪种类型的变量，关于变量取值的任一集合都构成一个分布，分布可以用图表来表示。本章在这一部分中演示了几种不同类型的图表，它们各有千秋。作图的关键在于选择揭示数据最清楚的图形。读图时需要注意细节，不能局限于

图中最醒目的特征，还得关注那些不太明显的特征，比如  $x$  轴和  $y$  轴所用的刻度。正如我将在后面内容（第12章）中讨论的那样，如果忽略了细节，图表就可能造成严重的误导。

## 总结与展望

本章的目的是对社会科学研究中使用的一些基本原理和术语提供一种概览。社会科学研究涉及变量类型、实验设计和抽样方法，这些基础知识有助于理解本书后续章节所阐述的统计学应用。接下来，我们开始学习统计学。也许你对统计学还毫无头绪，在你眼里那不过是一堆希腊字母罢了，这其实倒也未必是一件坏事。



## 第1章的术语表

**卡方分布** (chi-square distributions) 与卡方统计量 ( $\chi^2$ ) 相联系的一族分布。

**常数** (constant) 只有一种取值的构造（例如，若每一个样本成员都是10岁，那么“年龄”这个构造就是一个常数）。

**方便抽样** (convenience sampling) 根据获取或可得的便利程度来选择样本。

**相关性研究设计** (correlational research design) 用于考察变量间联系的一种研究方式。在此类研究设计中，研究者对变量不施加控制。

**因变量** (dependent variable) 依假设，因变量值取决于自变量值。例如，身高在某种程度上取决于性别。

**描述统计量** (descriptive statistics) 用于描述取值分布特征的统计量。

**二值变量** (dichotomous variable) 只能取两个离散值的变量（例如，“妊娠”变量只能在“未怀孕”时取0，“怀孕”时取1值）。

**分布** (distribution) 一个变量取值的任意集合。

**实验研究设计** (experimental research design) 实验者或研究者控制研究状况的一种研究方式，通常包括控制自变量和研究对象分组。

**F分布** ( $F$  distributions) 与  $F$  统计量相联系的一族分布，常用于方差分析 (ANOVA)。

**频率** (frequency) 一种取值在一个分布中出现的频繁程度。



**一般化（或普遍性）**（generalize or generalizability）利用样本数据结果对总体特征或非样本对象下结论的能力。

**自变量**（independent variable）依假设，决定因变量取值的变量。自变量往往由研究者所控制，但并非总是如此。

**推断统计量**（inferential statistics）根据样本数据计算得来、用于对从中抽样的总体进行推断的统计量。

**定距或定比变量**（interval or ratio variable）用数值测度的变量，其中相邻两个数值之间的距离或间隔相等。（例如，2是1的两倍，4是2的两倍，1和2间的距离等于2和3间的距离。）

**均值**（mean）取值的一个分布的算术平均数。

**名义尺度变量（定类变量）**（nominally scaled variable）对各类别赋以数字标签、数值大小没有意义的变量。

**正态分布**（normal distribution）取值的频率分布呈钟形，对称，渐近，均值、中位数和众数位于分布中央。

**定序变量**（ordinal variable）用数值测度的变量，数字本身有意义（例如，2比1大），但相邻数字间的距离不固定。

**参数**（parameter）从总体数据中计算得到的一个或多个值。

**总体**（population）具有规定特征的全部对象的集合。（例如，所有健在的美国成年男性。）

**定性（或分类）变量**（qualitative or categorical variable）具有离散类别的变量。若类别用数字表示，则其含义与名词标识一样，而没有数量的意义。（例如，令1=“男性”，2=“女性”，则1不大于2，也不小于2。）

**定量（或连续）变量**（quantitative or continuous variable）用有序、有意义的数字进行赋值的变量，从而使1小于2，2小于3，诸如此类。

**随机分配**（random assignment）将样本对象随机分配到不同组（例如，实验组和控制组），或者不考虑样本对象的任何特征进行分组。

**随机样本（或随机抽样）**（random sample or random sampling）以一种方式从总体中抽取对象，以确保总体中的每一个对象都有相等的机会入选样本。

**典型抽样**（representative sampling）有目的的选择对象以获取在一些感兴趣的特征上能够代表总体的样本。（例如，选取的样本与更大的总体在各族裔比例上相同。）

**样本**（sample）从更大的总体中选取的对象集合。

**统计量** (statistic) 从样本数据中得出的一项特征或一个值。

**$t$  分布** ( $t$  distributions) 与  $t$  统计量相联系的一族分布, 常用于比较样本均值以及检验相关系数和回归斜率的统计显著性。

**变量** (variable) 研究中考察的具有不止一个取值的任一构造。

## 中心趋势的测度

当你收集数据时，最终会得到一个或多个变量的一组取值。如果将一个变量的取值从大到小依次排列，就得到了一个关于取值的**分布**（distribution）。研究者经常想了解这些取值分布的特征，例如分布形状、取值散布程度、出现最多的取值等等。研究者通常感兴趣的一个分布特征集合是中心趋势。这个集合由均值、中位数和众数组成。

**均值**（mean）可能是所有社会科学研究中最常用的统计量。均值只是一个分布中所有取值的算术平均数。研究者之所以喜欢用，是因为它提供了一个简单的数字来大致概括分布。重要的是需切记，尽管均值提供了一些有用的信息，但它并不能揭示取值分布的离散程度（即方差），也无法说明分布中有多少个取值接近均值。一个分布中可能只有很少的取值位于或接近均值位置。

**中位数**（median）是分布中排在第 50 百分位处的取值。也就是说，分布中有 50% 的取值大于中位数，50% 的取值小于中位数。研究者经常用中位数将分布中的取值分成数目相等的两组（称为**中位数分划**（median split））。中位数还可用于考察一个分布中的取值是否有偏，或者该分布的两端是否存在一些极端取值。后面对此有更详细的讨论。

**众数**（mode）是使用最少的中心趋势测度，因为它提供的信息量最少。众数只是指出分布中最常出现的取值，或者具有最高频数的取值。

## 总体和样本

你会注意到表 2—1 中用两种不同的符号来表示均值,  $\bar{X}$  和  $\mu$ 。用两种不同的符号是必要的, 因为需要将适用于**样本** (sample) 的**统计量** (statistic) 同适用于**总体** (population) 的**参数** (parameter) 区别开来。表示总体均值的符号是  $\mu$ 。统计量是从样本数据中计算出来的数值, 而参数是从总体数据中计算出来的或者适用于总体数据的数值。最重要的是要注意到, 样本是总体的代表, 样本统计量可作为总体参数的估计。均值的样本统计量用符号  $\bar{X}$  来表示。样本统计量和总体参数的区别在几章中都有所涉及 (例如, 第 1 章、第 3 章、第 5 章、第 7 章)。

表 2—1

分布均值的计算公式

$$\mu = \frac{\sum X}{N}$$

或

$$\bar{X} = \frac{\sum X}{n}$$

式中:  $\bar{X}$ ——样本均值;  
 $\mu$ ——总体均值;  
 $\sum$ ——求和运算符号;  
 $X$ ——分布中的单个取值;  
 $n$ ——样本中的取值个数;  
 $N$ ——总体中的取值个数。

## 中心趋势测度详解

中心趋势的各种测度都不难计算。利用计算器和统计软件程序可以算出这些统计量中的任意一个, 可能根本不需要手工运算。但是为了学习, 也为了在没有计算器的条件下得到这些统计值, 需要了解进一步的信息。

因为均值是一个平均数, 所以均值的计算方法是将一个分布中的所有取值加总起来除以取值的个数。如果一个分布中有 10 个取值, 那么将所有取值加起来之后再除以 10 即可。表 2—1 中列出了均值的计算公式。

对于一个简单的取值分布<sup>①</sup>而言, 中位数 ( $P_{50}$ ) 的计算甚至比均值的计算更

<sup>①</sup> 也可能计算**组频数分布** (grouped frequency distribution) 的中位数。斯巴茨 (spatz, 2007) 的著作《统计学基础: 分布的故事 (第 9 版)》中很好地描述了组频数分布中位数的计算技术。



聚，而在另一端则较少，那么这种分布称为偏态分布。偏态分布的均值、中位数和众数通常都在不同的点。

重要的是，对于偏态分布和正态分布而言，计算均值、中位数和众数的方法是一样的，只是三种中心趋势测度之间的关系不同。为了举例说明，我虚构了一个样本容量为 30 的取值分布。假设从五年级学生中随机抽取一个容量为 30 的样本，询问其是否认为学习成绩很重要，并使用 5 分制量表对重要程度进行评价，“1”表示“一点儿都不重要”，“5”表示“非常重要”。因为大多数五年级学生都倾向于认为学习成绩很重要，所以这一分布中的大多数取值都在量表评分较高的一端，而只有少数取值在评分较低的一端。将虚构取值从小到大依次排列后得到的分布如下：

1	1	1	2	2	2	3	3	3	3
4	4	4	4	4	4	4	4	5	5
5	5	5	5	5	5	5	5	5	5

如你所见，只有少数取值在分布的低端（1 和 2），更多的取值在分布的高端（4 和 5）。图 2—1 直观地展示了这一偏态分布的形态。

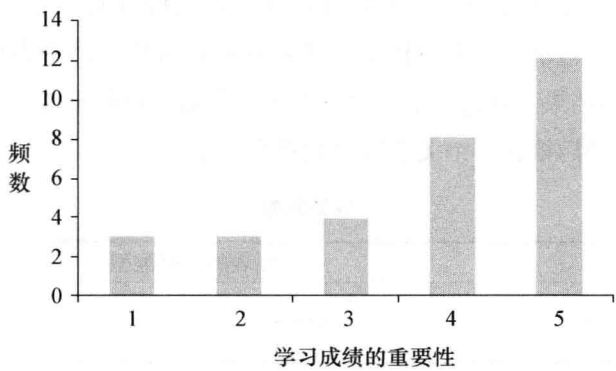


图 2—1 偏态分布

该图刻画了一些偏态分布的形态。注意到大多数取值聚集在分布的高端，少数取值向低端方向延伸出一条尾巴。由于尾部趋向低端，因此称为负偏（negatively skewed）分布。如果分布的尾部趋向高端，则称为正偏（positively skewed）分布。

扫一眼分布的取值或图形，就不难发现众数是 5，因为该分布中取值 5 的个数比其他数字都多。

只需用前面给出的公式就可以计算出均值，即将所有取值加总（ $\sum X$ ）之

后再除以分布中的取值个数 ( $n$ )。于是得到一个分数  $113/30$ ，约简后为  $3.766\bar{6}$ ，保留小数点后两位数并四舍五入，最后的均值为  $3.77$ 。

将分布中的取值从小到大依次排列后，找出居于中间位置的取值，即为中位数。这一分布中有  $30$  个取值，故居中的有两个。依次排列之后，居中的两个（第  $15$  个和第  $16$  个）取值都是  $4$ 。这两个值加起来除以  $2$  等于  $4$ ，所以中位数为  $4$ 。

前面已经提及，一个分布的均值受到所谓“异常值 (outliers)”的影响，而中位数则不受这些值的影响。异常值是指一个分布中大到或小到不同寻常的值。因为异常值位于尾部，所以偏态分布的均值通常会偏向尾部。负偏分布的均值小于中位数，因为均值被拉向尾部，而中位数则不然。前面例子中的均值 ( $3.77$ ) 就略小于中位数 ( $4$ )。正偏分布的均值则略大于中位数。

下面两个图形用以进一步说明异常值对分布均值的影响。图中给出了若干不同国家人口出生时的平均预期寿命。图 2—2 中的折线图描绘了  $13$  个国家的预期寿命，各国按预期寿命的长短依次排列，最长为日本，最短为乌干达。如你所见，从日本到土耳其的预期寿命平缓递减，但乌干达的预期寿命却大幅锐减。在这一关于国家的分布中，乌干达是一个异常值。除乌干达以外的其他所有国家，平均预期寿命为  $78.17$  岁，而图 2—2 中包括乌干达在内的全部  $13$  个国家的平均寿命则为  $76.21$  岁。增加了乌干达一个国家，就使得全部  $13$  个国家的平均预期寿命降低了近两岁。两岁听起来或许不是很多，但图 2—2 中排在前  $5$  位的国家彼此相差的也不过这个数，从中不难看出，两岁在各国人口预期寿命排名中的影响是举足轻重的。

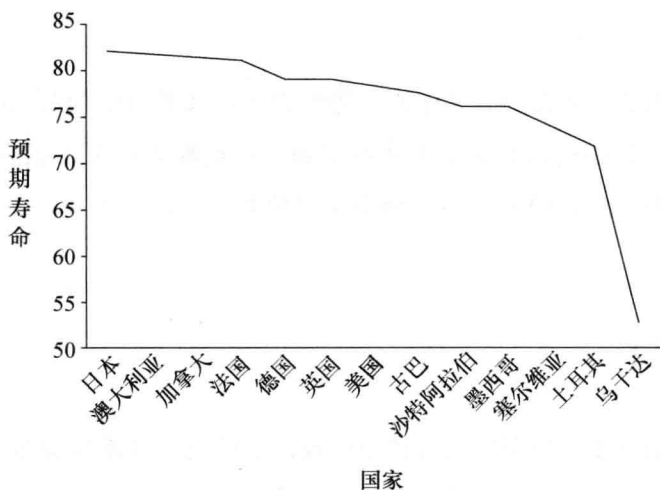


图 2—2 若干国家人口出生时的预期寿命

样本越小,异常值对均值的影响越大,因为均值是由分布中所有取值组合而成的统计量。对于较大样本而言,一个异常值不会造成非常大的影响,可对于小样本而言,一个异常值会令均值大不一样。为了说明这一效应,图2—3中分析了比图2—2中更小的一个国家子集。我们再次看到,乌干达的预期寿命(约52岁)远低于日本、英国和美国的预期寿命(都近80岁)。不包括乌干达在内,3个国家的平均预期寿命是79.75岁。而包括乌干达之后,4个国家的平均预期寿命降至72.99岁。加入一个异常值使得均值减少了近7岁。在这个小的数据集中,中位数位于英国和美国之间,约为78.5岁。这一例子解释了异常值如何将均值拉向自身方向,此时的均值便低于中位数。

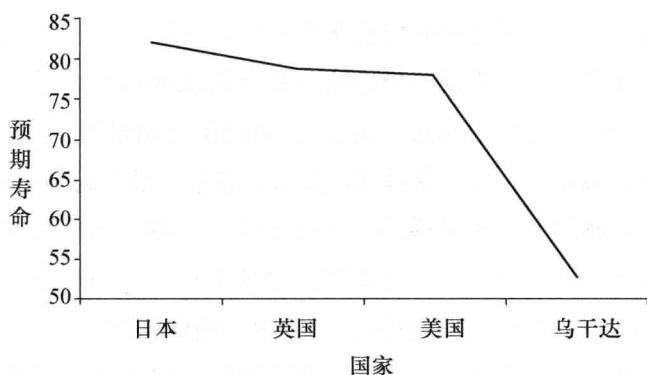


图2—3 4个国家人口出生时的预期寿命

## 行文表述

当阅读已发表论文遇到有关中心趋势的表述,或者自己撰写这些表述时,你会发现这些表述短小简洁。对于上述例子而言,正确的行文表述应该如下所示:“在这一分布中,均值( $\bar{X}=3.77$ )略小于中位数( $P_{50}=4.00$ ),表明这是一个负偏分布”。

## 总结与展望

中心趋势的测度,特别是均值和中位数,是研究者最常用也最有用的一些统计量。它们各自用一个简单数字提供了关于整个取值分布的重要信息。例如,美国男人的平均身高为5英尺9英寸,这一个数字概括了该国数百万男人的信息。



均值和中位数都很有用，但正因为如此，才容易忘记像均值这样的统计量忽略了关于分布的大量信息，包括许多分布中存在的巨大变异，这往往是危险的。如果不考虑变异程度，那么就on容易根据均值做出“一刀切”的概化，或者形成刻板印象。变异程度的测度是下一章的主题。

## 第2章的术语和符号表

**双峰 (bimodal)** 一个分布中有两个出现频率最高的值。

**分布 (distribution)** 样本中一个变量取值的集合或群组，这些值经常从小到大依次排列，但也不尽然。

**均值 (mean)** 一个分布的取值的算术平均数。

**中位数分划 (median split)** 以中位数取值为界，将一个分布中的取值分成数目相等的两组。大于中位数的取值构成大组，小于中位数的取值构成小组。

**中位数 (median)** 一个分布中位于第 50 百分位的取值。分布中有 50% 的取值大于它，50% 的取值小于它。

**众数 (mode)** 分布中出现最频繁的取值。

**多峰 (multimodal)** 一个取值分布中有两个或两个以上出现频率最高的取值。

**负偏 (negative skew)** 偏态分布中大多数取值集中在分布的高端，少数取值在分布的低端形成尾部。

**异常值 (outliers)** 距离均值超过 2 倍标准差的极端值。

**正偏 (positive skew)** 偏态分布中大多数取值集中在分布的低端，少数取值在分布的高端形成尾部。

**参数 (parameter)** 得自于总体数据的值，或由样本统计量推断出的总体的值。

**总体 (population)** 从中收集数据或抽取样本的群组。总体包括了数据可能适用的整个集合。

**样本 (sample)** 由总体中选取的、从中收集数据的个体或群组。

**偏态 (skew)** 分布中的大量取值集中在一端，相对较少的取值散布在分布的另一端，形成一个尾部。

**统计量 (statistic)** 从样本数据中得出的值。

$\Sigma$  求和。

$X$  分布中的单个取值。

$\sum X$   $X$  的和；分布中所有取值加总。

$\bar{X}$  样本均值。

$\mu$  总体均值。

$n$  样本中的对象或取值的个数。

$N$  总体中的对象或取值的个数。

$P_{50}$  中位数的符号。

## 变异程度的测度

第2章描述了关于中心趋势的测度。诸如均值和中位数之类的测度提供了有用信息，但却存在局限。单凭这些测度无法提供大量信息，认识到这一点很重要。有一句旧谚语对使用均值提出了警告：“如果你的头在冰箱、脚在烤箱，平均来说你很舒服。”考虑如下例子：假设我对100名五年级的孩子进行调查以评估他们的抑郁水平。进一步假设该抽样调查的抑郁水平平均值为10.0，中位数也是10.0。从这些信息中，我们只能了解到数据分布具有相同的均值和中位数，并且都在10.0处。现在考虑哪些是我们未知的。我们不知道是否有高分值或低分值。我们不知道样本中学生的抑郁水平是全部一样，还是彼此差别很大。我们不知道分布中最高或最低的抑郁水平是多少。简单来说，我们尚不清楚这一分布的离散程度。换言之，我们对这一分布的变异程度仍一无所知。

研究者常用来考察离散程度的三个指标是：**极差**（range）、**方差**（variance）和**标准差**（standard deviation）。当然，标准差可能是其中信息量最大、使用最广泛的。

### 极差

极差是一个分布中的最高分（**最大值**（maximum value））与最低分（**最小值**（minimum value））之差。研究者通过该统计量能够对一个分布中取值的散布程度获得初步认识，但因其颇具误导性，所以不是一个特别有用的统计量。例如，在前面所描述的抑郁水平调查中，有一名男生得1分，另有一名男生得20分，其余98名男生都得10分。在这个例子中，极差为19（ $20 - 1 = 19$ ），似乎意味着取值分布比较离散，但事实并非如此。研究者经常会扫一眼极差，以确定样本是否涵