统计学经典译丛

STATS: DATA AND MODELS

统计学

数据与模型（第3版）

理查德·D·德沃（Richard D. De Veaux）
保罗·F·威勒曼（Paul F. Velleman）　戴维·E·博克（David E. Bock）　著
耿修林　译

STATISTICS CLASSICS

Pearson
中国人民大学出版社

STATS: DATA AND MODELS

统计学

数据与模型

（第3版）

理查德·D·德沃（Richard D. De Veaux）
保罗·F·威勒曼（Paul F. Velleman） 著
戴维·E·博克（David E. Bock）
耿修林 译

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

统计学：数据与模型：第3版/理查德·D·德沃等著；耿修林译. —北京：中国人民大学出版社，2016.7
（统计学经典译丛）
ISBN 978-7-300-22938-6

Ⅰ. ①统… Ⅱ. ①理…②耿… Ⅲ. ①统计学-高等学校-教材 Ⅳ. ①C8

中国版本图书馆CIP数据核字（2016）第119431号

统计学经典译丛
统计学：数据与模型（第3版）
理查德·D·德沃
保罗·F·威勒曼 著
戴维·E·博克
耿修林 译
Tongjixue：Shuju yu Moxing

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）　　010－62511770（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 三河市汇鑫印务有限公司
规　　格 185 mm×260 mm　16开本　　**版　　次** 2016年7月第1版
印　　张 27.5插页1　　**印　　次** 2016年7月第1次印刷
字　　数 690 000　　**定　　价** 69.00元

译者序

统计学这门学科与其他学科不一样。对于两个显然不同的数值，我们却经常这样表述：它们在统计意义上没有差别。不明就里的人一定会认为是睁着眼睛说瞎话，然而这恰恰是统计学的典型特征。对统计学抱有畏难情绪的人，可能认为它与数学有关，也可能认为它与数据计算有关，其实从应用的角度看这些都是次要的，关键是如何把握统计科学的思考方式。运用统计方法认识事物，既不是思辨式的演绎逻辑，也不是找通项公式的数学归纳逻辑或枚举式的语文归纳逻辑，不太恰当地讲，它是那种“尝一勺水知四海味”的归纳推断逻辑，由此就引发了统计学的基本概念、方法原理、分析结论认识与其他学科的差异。单纯地讲清楚这个道理可能比较容易，但把这样的认识逻辑有机地贯穿和渗透到整个知识体系中，并让初学者能实实在在地领会可能就不那么简单了。

本人从事初等统计学教学工作多年，虽然能驾轻就熟，但偶尔难免也会漫不经心乃至生出惰性。比如，不再愿意花精力琢磨如何讲清楚基本概念，不再精心构思每一堂课教学内容的安排，不再想方设法让一个初学者掌握统计学思考问题的方式，不愿意再改进以让一门枯燥的课程更有趣味性。深知这样的教学心态不好，所以一直想花点工夫系统地了解统计教学新的传授理念。这次得到中国人民大学出版社王伟娟女士的垂顾，约请我翻译《统计学：数据与模型》（第3版），十分愉快地接受了任务。俗语有云：人到中年万事休。我希望借机历练自己，借用当下流行的话说就是再励志一把。

《统计学：数据与模型》（第3版）的三位作者理查德·D·德沃、保罗·F·威勒曼、戴维·E·博克，长期从事统计学的教学工作，拥有丰富的教学经验，凭着在统计学教学中的突出业绩获得过多个奖项。德沃先后任教于宾夕法尼亚大学沃顿商学院、普林斯顿大学工程学院和威廉姆斯学院，是美国统计学会会员，拥有斯坦福大学统计学博士学位，研究领域是科技产业大数据分析和挖掘。威勒曼是美国统计学会会员和美国科学促进协会会员、康奈尔大学统计学教授，在多媒体统计教学界有一定的影响，拥有普林斯顿大学统计学硕士和博士学位，曾师从著名的统计学家图基，研究领域主要是探索性数据分析。博克长期在伊萨卡预科学校、伊萨卡学院、康奈尔大学教授统计学课程，是统计学预修考试的主考官和试卷的主审人。《统计学：数据与模型》一经问世便受到欢迎，与几位作者良好的教育背景和丰富的教学经验积累密不可分。

在翻译过程中，本人认为《统计学：数据与模型》（第3版）的突出特点是：

第一，内容安排紧紧相扣，层次推进十分自然。本土的统计学教科书大多从数值资料统计分析开始，然后把比例类统计推断分析当作特例来对待，该书却反其道而行之，这样做的一个明显好处是，能够巧妙地利用比例与比例抽样分布标准差的关系，很自然地过渡到统计区间估计和假设检验的讨论，比如两总体比例差假设检验，就根本不需要事先强调等方差的假定条件。在介绍相关与回归分析时，该书把有关内容拆分成几块，按照不同的顺序予以讲解。比如，在介绍了统计特征数字后，以简单回归

为对象，把简单回归方程的求解转化成几个特征数字来说明；在介绍完区间估计和假设检验后，讨论回归模型的推断分析问题；在学过方差分析之后，再来讲解多元回归的代表性分析和建模问题等。

第二，大量使用图表形式展现和描述数据。该书除了安排专门的章节介绍统计资料的图表描述方法外，在相关统计分析工具假定条件的讨论中，特别推崇运用图形进行考察。比如，利用正态概率图、直方图检查样本观察资料的单峰对称性，运用箱线图分析是否存在异常点，通过各类残差图诊断线性性假定、等方差假定等。

第三，特别注重检查方法应用的假定条件是否满足。该书不太强调计算过程，但重视统计方法应用条件符合性的检查。比如，在计算观察资料特征数字时，特别提醒需要考察资料的对称性，对分布对称的资料，可以直接使用一般的均值和标准差，对分布不对称的资料，需要采用位置特征数、分位数差等指标描述。在进行回归分析时，强调要注意检查线性性、独立性、等方差性、渐近正态性等条件。

第四，引入问题的背景介绍生动有趣，体现了统计方法应用的强大功能。全书绝大部分章节都从一个真实的实例说起，引出相应的统计问题，然后结合该章的内容逐步进行统计分析方法的应用讲解。这些实例和例题来源广泛，有的属于生态环境保护领域，有的是体育运动方面的，有的是社会现象和社会管理方面的，有的是工商经营活动方面的，让人们感受到日常生活中统计无处不在。

第五，提出了数据分析的基本策略。这一点几乎体现在该书每一章的内容中，在此我们仅举例说明。比如，在有异常值的情况下，最好能同时报告带有异常值和不带有异常值的分析结果，并做出适当的解释说明；先考察数据资料分布图，然后决定采用什么样的数据处理方式；注意检查异常值，并对其进行适当分析。

《统计学：数据与模型》（第3版）的篇幅很大，一共有31章。考虑到国内教学的实际情况，翻译过程中，在保持主要特色的前提下，对全书做了相应的删减和调整处理。主要是：原书第1章和第2章的内容相对简略，在译著中做了合并处理；把“统计资料获取”这一单元的3章（随机性、样本调查、实验设计与观察研究）的主要内容融入译著的第1章；删除了随机性和概率论这一单元的4章，即随机性与概率、概率运算准则、随机变量、随机变量的概率分布，这主要是考虑到国内大学开设的数学课程都讲授过概率论与数理统计。

由于本人能力和水平有限，译著中可能存在不准确甚或错误的地方，恳请读者批评指正。

耿修林

前　言

《统计学：数据与模型》自问世以来，不断接到来自教师和学生的反馈意见，让我们深受鼓舞。如果要说这本书有什么与众不同的特点，那就是学生们竟然读得懂它。根据我们掌握的情况，无论是大学预科生还是大学生，都认为这本书写得明白晓畅、通俗易懂。功夫不负有心人，这表明我们孜孜不倦、循循善诱的编撰理念收到了应有的效果。

和其他同类的初级读物不同，《统计学：数据与模型》重在告诉大家如何理解统计数据处理的结果，这样的定位贯穿全书，包括问题导入形式的设计，以及观察数据的广泛使用等。其中最为重要的一点是，本书致力于教会大家如何运用统计思考方式而不是过分关注统计计算。因此，为激发读者的学习兴趣，本书安排了大量实例，根本意图不在于“怎样寻找到答案”，而在于“如何认识数据分析的结论”。

《统计学：数据与模型》此前已经出过两版，目前推出的是第 3 版。在如何开展统计学教学和如何学会统计学思维上，第 3 版延续了前两版的成功做法。在此基础上，为了更清晰、更有趣味，本书对部分章节进行了修订，除了引进一些最新的案例外，还对涉及的统计学概念做了梳理以使前后更连贯。总体来说，《统计学：数据与模型》（第 3 版）新的改动主要体现在以下各个方面：

- 例题设计。几乎每一章都补充了一些新的例题，这些例题旨在帮助读者学会如何利用相关的统计学概念和统计方法，各章例题不是割裂开来分别选取的，而是按照相应章节的主要内容，从问题的背景介绍出发，然后根据内容的演进不断地运用同一个例题介绍概念和方法的使用。
- 复习思考题。新增了大量复习思考题，题目的编写先从概念类练习开始，然后逐步过渡到综合性的练习。
- 实例应用讲解。全书超过 1/3 的章节都新增或替换了实例应用介绍，从陈述所要讨论的问题开始，到观察资料的描述显示、方法应用、假定条件检查、模型构建、计算分析及其结论，一步一步讲解方法的应用过程。
- 数据来源。全书绝大部分例题和复习思考题中用到的数据资料，都是从近年来的新闻报道、研究论文等中搜集来的，基本上都是真实的。
- 生动的背景介绍。每一章都从一个真实的背景事例开始讲解，然后结合数据进行分析，在第 3 版中，我们更新和补充了很多这样的背景事例，比如：地震、飓风行进路线预报、企鹅潜水时的心律变化、是否相信鬼神的调查、骑摩托车是否戴头盔的对照检查、男女司机开车系安全带的差别、奥林匹克滑冰比赛名次等。

一本教科书再好，如果没有人阅读也会失去价值。为了让《统计学：数据与模型》（第 3 版）更符合读者的口味，我们做了许多改进，使之具有以下特点：

- 可读性。在编撰本书时，我们力求做到讲解细致、深入浅出，通过运用丰富的资料，使大家能够读得更明白。
- 变通性。这里所讲的变通性，并不意味着书中的内容浅显或不够专业，相反，我们试图准确地对一些概念进行介绍，并且尽一切可能进行更深入的解释和论证。
- 紧凑性。每章只重点讲解一个方面的知识点，以使学习起来内容更加集中。

● 连贯性。本书尽力避免“说的是一套做的是另一套”，总是先从数据的图像展示和假定条件检查入手，然后过渡到模型分析。

● 趣味性。那些只想匆匆浏览本书的同学，可能对我们的写作方式感到不爽，因为对于重要的概念、定义和求解示范，我们没有采用特别的格式加以突出。这是一本需要通读的教科书，因此我们尽力做到让阅读的过程更愉快。

教科书的选材通常是有讲究的，对于哪些应该包含进来，哪些不应包含，需要认真筹划。在本书的选材及其顺序安排中，我们一直遵循几项基本的原则。首先，我们要确保选择的每项内容能递进地融入整个统计认识架构；其次，符合美国统计协会的统计教学指南；具体的要求是：加强训练统计知识和统计思维，提高广泛利用实际数据的能力，强调对计算机处理结果的理解，培养学生的自学能力，让学生学会应用计算机软件处理和分析数据，注重学习过程中的实效。

学习统计少不了数学基础，数学在统计学中的作用表现在三个方面：能够对统计学中的一些概念用简洁明了的语言进行表述，能够说明数据计算方法，能够对一些基本结论提供证明。对于数学的这些作用，我们首推第一个。在了解统计学的概念、概率论和推断分析的过程中，如果能借助数学手段，可以使讨论变得更为清晰简练。我们深知使用数学表达式会让一些人感到沮丧，为此我们也采用语言文字并辅以数值分析的例子，来说明概念和推断分析的结果。本书不太关心统计学定理的证明。尽管有些统计学定理是非常值得关注的，而且大部分也很重要，但是这些定理的证明对学习初等统计学的学生来说没有太大的帮助，甚至有可能会干扰他们对相关概念的理解。不过，在我们认为利用数学工具有助于清晰表达并且不会带来阅读困难时，我们是不会刻意回避的。对于数据资料的统计计算不要太在意，尽管有些统计计算不是太复杂，但计算总是让人感到心烦意乱，更重要的一点是，把心思花在统计计算上往往意义不大。如今计算机已经得到普及，大量的统计计算都可以运用计算机软件处理，所以让学生用手工方法进行统计计算确实有点事倍功半。有鉴于此，我们使用的函数表达式都是经过精心安排的，主要还是侧重于如何更好地理解统计概念和方法。

理查德·D·德沃
保罗·F·威勒曼
戴维·E·博克

目　录

第1章 统计与数据

统计学向来不太受尊重，用数字语言表述问题，似乎被看成是故弄玄虚，“你可以运用统计证明一切”更是佐证。在学校的教学活动中，统计学课程也没有获得应有的地位，学生们并不是因为统计学的趣味性而心甘情愿地选修统计课。然而，统计学真的很有趣，绝不像人们普遍认为的那样。统计学告诉我们怎样应用数据认识世界，分析实际问题时如果具有统计意识，可以帮助我们获得更清晰、更精确的结论。

1.1 什么是统计

稍加留意就可发现，不论我们在百货店购物还是在网上冲浪，总有人在收集与我们购物或上网有关的资料。对接运的每一单货物，联合包裹服务公司（United Parcel Service，UPS）会跟踪记录包裹的运送情况，并将之存储在所建立的庞大的业务数据库中，如果你是UPS公司的客户，你可以随时进行查询。据说UPS公司的数据库接近17TB——相当于美国国会图书馆的数据库的规模，该数据库涵盖了馆藏的每一本图书。有了这些数据，我们总希望能做点什么。

在认识我们今天所处的纷繁复杂的世界时，统计学自有其一席之地。对列入美国国家食品药品管理局（Food and Drug Administration，FDA）监控名单的转基因食品或新药，统计学家可以参与评估它们可能存在的风险。利用统计工具，可以预测某个地区新增艾滋病例数，可以分析顾客对某种商品销售的反应。自然科学家、社会科学家与统计学家合作，能更好地厘清失业与环境控制之间的联系。像学前教育是否会影响孩子们上学后的学习表现、维生素C是否有助于预防疾病等问题，都可以运用统计方法进行分析。总之，无论何时何地，只要拥有数据，统计就能派上用场。我们编撰这样一本书的主要目的是，希望帮助大家在解决问题时形成个人见解，告诉大家如何运用统计工具展示数据资料的信息，掌握数据分析的一些技巧。

讲到现在，什么是统计呢？“统计”一词有两个含义：一是统计学，它是一门有关推断的科学，通过一系列独特的方法和技术帮助人们认识事物；二是统计资料，它是从数据资料中产生的具体结果。

用一个关键词概括一门科学的内容，是一件很有挑战性的事。如果说经济学研究财富、心理学研究思维、生物学研究生命、历史学研究史实、哲学带有思辨色彩、工程学研究如何制造、会计学关注财务信息，那么统计学研究的就是变异，这是统计学与其他学科不一样的独特之处。

1.2 数据及其种类

1.2.1 数据的意义

若干年以前，小地方的大多数商店对它们的顾客都比较了解。当你走进一家五金店，店主人可能不等你开口就会跟你说，一款新的配件已经到货了；裁缝知道你爸爸衣服的尺寸；理发师了解你妈妈喜欢什么样的发型。即使到了今天，诸如此类的商店仍然存在，但那些可以通过电话或网络订购物品的大商店日益增多。即使是这些大商店，如果你拨打800电话准备买一双新式跑鞋，客户服务经理也许能叫出你的名字，甚至会询问你对6周前买的一双袜子有什么评价。在冬季来临之前，你可能会莫名其妙地接到一封电子邮件，告知你一款适合冬季跑步用的保护头套已经上市。顾客成千上万，我

们不禁要问：这些销售经理是怎么知道你，又是怎么知道你住在哪儿，以前买过什么东西的呢？

答案就是数据。通过收集客户、交易、销售方面的数据，能让商家按照货物清单开展促销活动，帮助商家预测客户的消费偏好。通过这些数据，商家能够事先估计消费者在未来可能买什么东西、买多少，以便及早准备好商品。利用数据和从数据资料中获得的信息，能够改善客户服务，将半个世纪前店主对顾客面对面的了解再现出来。

亚马逊网站（Amazon. com）于 1995 年 7 月开通，号称“全球最大的网上书店”。到 1997 年底，亚马逊已建立起多达 250 万册的图书目录，向全球 150 多个国家或地区的超过 150 万人销售图书。2007 年，亚马逊的营业额达到 148 亿美元。现在，亚马逊除了经营传统的图书业务外，还进一步拓宽了业务范围，经销从价值 40 万美元的项链到西藏牦牛奶酪等各种精选出来的商品。为了更好地做好销售服务和获得更好的销售业绩，亚马逊不断维护和升级改造公司的网站。在决定怎样改进网站性能之前，亚马逊的工程师一般会收集和分析数据，并从中找到最好的切入点。如果你有空浏览亚马逊的网站，能查到各种各样的图书或者商品的介绍和报价。亚马逊的统计人员能掌握你是否登录推荐的链接、是否会购买推介的商品、花多少时间访问公司的网站等信息。正如亚马逊前数据挖掘中心主管考哈威（Ronny Kohavi）所说，“在亚马逊，数据才是真正的王，点击量和销售数据是王冠上的明珠，有了这些数据，我们才能为消费者提供个性化体验。总之，直觉必须让位于数据，我们必须亲身体验网上购物环境，通过倾听顾客的意见改进我们自身的工作”。

1.2.2　什么是数据

对数据我们或多或少有所了解，但数据的确切含义是什么？数据是否等同于数字？你过去购物的金额是数据，进一步讲是数值数据，对亚马逊公司的数据库来说，品名和商品的其他标记也是数据，只不过不是数值数据。有的情况下，数据用数值的形式表现出来，但实际上是标签，这一点容易弄混，比如 0321692551 是数字值，但它是亚马逊某本图书的编码。

不管是什么类型的数据，如果不把它与具体的背景联系起来，那么此时的数据毫无用处。新闻记者都知道，叙述新闻事件一开始就要交代清楚“是谁”（who）、“是什么”（what）、“什么时间”（when）、“什么地点”（where），如有可能的话还需要说明“是什么原因”（why），这就是新闻报道中的“5W”。我们认为，最好还要加上“是怎样解决的”（how），才显得更完整。将 5W 和 1H 说清楚，数据就有了背景，也就有了价值。在 5W 和 1H 中，who、what 是最基本的，如果不能说清楚这两者，便没有所谓的数据，即使有了这样的数据也没有信息价值。

假如给定如表 1—1 所示的数据：

表 1—1　　**亚马逊收集的数据**

B000001OAA	10.99	G. 切里斯	902	15783947	15.89	堪萨斯乐队	伊利诺伊州	波士顿乐队
加拿大	P. 萨缪尔森	橘郡男孩	无	B000068ZVQ	Bad Blood 专辑	纳什维尔之声	H. 凯瑟琳	无
Mammals 乐队	10783489	俄亥俄州	无	芝加哥乐队	12837593	11.99	马萨诸塞州	16.99
312	D. 莫妮卡	10675489	413	B00000I5Y6	440	B000002BK9	展翅高飞	有

对于表1—1，我们可能不明就里。为什么会是这样？原因在于这些数据没有把背景交代清楚。如果我们不明白这些数据表明的是什么意思，不清楚它们记录的是什么内容，怎么会知道这些数据的意义呢？假如把表1—1换成表1—2：

表1—2　　亚马逊收集的数据

序号	购物者姓名	目的地	单价	区号	购买过的CD	赠品	产品编号	乐队
10675489	H. 凯萨琳	俄亥俄州	10.99	440	纳什维尔之声	无	B00000I5Y6	堪萨斯乐队
10783489	P. 萨缪尔森	伊利诺伊州	16.99	312	橘郡男孩	有	B000002BK9	波士顿乐队
12837593	G. 切里斯	马萨诸塞州	15.98	413	Bad Blood专辑	无	B000068ZVQ	芝加哥乐队
15783947	D. 莫妮卡	加拿大	11.99	902	展翅高飞	无	B000001OAA	Mammals乐队

由表1—2可以看出，这是从亚马逊订购CD的4条销售记录。表1—2中的列标题说明了记录的是什么，行标题标明了是谁。注意：仔细体会一下“who”的确切所指。这里所说的“who”，不是我们习惯意义上的理解，即使销售记录中真的涉及某个人，他也不是数据背景中所指的人，“who”实际上是购买序号（不是购买东西的某个人）。拿表1—2来说，“who”就是最左边的那一列。至于“when”、“where”等，需要从亚马逊公司数据库管理员那里获取（在数据库管理中，这类信息被称为元数据）。

一般而言，数据表中的每一行都对应着一个个体，记录着这些个体的一些特征，主要是按照具体情况对这些个体特征赋予不同的内容。所观察的所有个体的集合称为调查对象，所观察的人称为主体或参与者，动物、工厂、网站以及其他无生命的主体，一般称为观察单位。数据库中，数据表的行叫做记录，看看表1—2就清楚了。个体是一个被广泛使用的术语，表1—2中的个体就是一个个CD订单。人们经常把数据值当作观察值，并没有清楚地讲明是谁，这时候应该确保你能知道数据指向的是谁，否则你就不知道数据所表明的内容。

一系列个体构成样本，样本中的个体是从我们想要认识的更大的调查对象（总体）中筛选出来的。毫无疑问，亚马逊公司关心的是它的客户，与此同时，该公司也想知道如何吸引那些从未通过亚马逊网站购物的网络用户，所以亚马逊公司的所有客户甚至潜在用户都可被看成是总体。样本是总体的缩影，为了能将样本信息上升到对总体的认识，需要关注样本的代表性。

例1—1

2009年3月，《消费者》（*Consumer Reports*）公布了来自不同制造商的116款大屏幕高清电视机的检测报告，据此谈谈总体、样本和“who”。

答：该问题的总体是所有近期在市场上销售的大屏幕高清电视机，被检测的116款大屏幕高清电视机构成了研究样本，这里的“who”指的是每一台被检测的电视机。

数据表中的列通常用来反映个体特征，因此我们把列名叫做变量。变量这个词看似简单，但要真正了解它的含义，也许不像你想象的那样，而是取决于我们想了解什么。表1—2中的区号是数字，我们能不能把它们当作一般数字来用呢？能说610是305的两倍吗？单纯从数字的角度来看，确实是这么回事，但是能说宾夕法尼亚州的安伦顿的区号（为610），是佛罗里达州的基韦斯特的区号（为305）的两倍吗？变量

能发挥重要的作用，不能漫不经心地对待它。

1.2.3　数据的类型

一些变量仅能告诉我们某个个体属于哪一组或哪一类，比如你是男性还是女性，打过耳洞还是没有打过耳洞，等等。从诸如此类的事例中，我们能学到什么呢？很自然的想法就是数数，即首先要搞清楚每个类别包含了多少个体。据此，我们就能比较每个类别的大小。

有些变量需要通过测量才能知道它的数值，并带有相应的计量单位。计量单位是测量的尺度，像日元、肘尺（古代长度单位，自人手肘部至中指末端，长度大约为 45.72～55.88 厘米）、克拉、埃（波长单位）、纳秒、英里/小时、摄氏度等，都是计量单位。根据这些计量单位，人们能知道物体有多少，以及相互之间的差距。如果没有这些计量单位，变量的测量结果就没有意义。举个例子，如果你不知道是用欧元、美元还是日元结算，承诺给你增加年薪 5 000，你会答应吗？与分类变量相比，我们可以基于测量的变量做更多的事情。利用测量变量的数据，可以观察现象的状态和趋势，比如，过去看场电影花多少钱，不同电影院的票价差异，过去 20 年间电影票价的变动情况等。

分类的、能将不同个体划分到不同类别中的变量，称为分类变量或属性变量，也可叫做定性变量。能用度量衡单位测量，并且测量的结果是数值化的，这样的变量称为定量变量。变量类型划分的价值不在于变量分类本身，重要的是如何针对不同类型的变量加深认识，以及采用什么样的数据处理方法。

假如某个变量的值只能用文字而非数字表示，可以肯定该变量一定是属性变量。然而，有些变量可能有两种不同的表示形式，比如年龄，我们可以用周岁值表示，也可以用诸如小孩、青少年、成年、老年之类的词来表示。在某门课程的评价调查中，就“该课程的教学对你是否有价值”这个测项拟定 5 个备选答案：毫无价值（用 1 表示），无价值（用 2 表示），说不上有价值无价值（用 3 表示），有些价值（用 4 表示），非常有价值（用 5 表示）。如果将课程价值作为变量，该变量是属性变量还是定量变量呢？调查人员也许把课程价值当作属性变量处理，仅仅统计选每个答案的学生人数。换个角度，如果调查的目的是了解该课程的改善情况，这时调查人员有可能用能产生直观感受的代码来表示，其结果就是把课程价值当作数量变量。问题是：代码只是代码，不是测量的结果。关于课程价值的回答，我们确信存在某种顺序，显然选择大代码的学生所表明的他们对课程的好感，要大于选择小代码的学生。平均得分 4.5 的课程比平均得分 2 的课程，带来的价值要大得多。尽管对课程评价的代码计算了平均值，但我们仍然需要注意，课程价值这个变量与真正的定量变量还是有一定差别的。要想把课程价值当成定量变量处理，我们必须设想存在一个“课程价值单位”，或者人为地默认具有这个类似的概念。课程价值变量虽然没有度量衡计算单位，但其回答结果带有一定的顺序，我们将诸如此类的变量称为顺序变量。这并不意味着就可以轻松区分变量的类别了，一个变量究竟是属性变量还是定量变量，还要依据我们的研究目的来确定。

例 1—2

背景材料见例 1—1。假定 116 款电视机的检测报告中，公布了每款电视机的制

造商、生产成本、屏幕尺寸、类型（液晶、等离子、背投），以及综合评分（0～100）。问：哪些是属性变量，哪些是定量变量，相应的计量单位分别是什么，该调查的“why”指的是什么？

答：例1—2中涉及5个变量，分别是：制造商、生产成本、屏幕尺寸、类型、综合评分。其中：制造商是属性变量，只能计数，没有度量衡单位；生产成本是定量变量，计量单位可以是美元；屏幕尺寸是定量变量，计量单位是英寸；类型是属性变量，没有度量衡单位；综合评分是定量变量，计量单位是分。进行该项调查，旨在帮助消费者更好地挑选电视机。

在统计活动中，数数是一件司空见惯的事。亚马逊公司的销售人员将商品运给消费者时，首先需要清点多少商品走陆路，多少商品采用一级航空运送，多少商品采用二级航空运送。对运输方式这样的属性变量，数数是汇总各类运送商品数量的一种常见方法。那么，可数数的变量是否就一定是属性变量呢？回答是否定的，我们经常也用数数的方法对待数量性质的现象，比如，数字音乐播放器中有多少首歌曲，这个学期你选修了多少门课程，回答这些问题只能数数。歌曲数、课程数本身是数量性质的变量，它们的计量单位是“首”或“门”，只不过为简单起见，我们都把它们叫做“数”。因此，数数存在两种不同的用法。对属性变量，我们数每个类别的个体数，此时各个类别的标识是我们所说的“what”，被计数的个体是数据中的“who”。数数本身不是数据，只是汇总数据的过程。为了做好寄运工作（why），亚马逊公司的销售人员需要计算每类运输方式中寄运商品的数量，那么寄运的方式就是我们说的“what”，寄运的商品就是“who”。有的情形下，我们关注的是现象的数量，这时也可能需要数数。比如亚马逊公司的销售人员为跟踪观察青少年消费群体的增长和预测CD的未来销售情况（why），他们会按月汇总访问过亚马逊网站的青少年人数，就此来看，“what”指网上购物的青少年，“who”指每个月，计量单位是青少年购物者人数。青少年是类别变量，但可以把它当作年龄对待，如此一来青少年人数也是数量变量，可以通过计数得到这一消费群体的人数。

学生的学号是用数字表示的。学号是不是定量变量呢？不是，因为它没有计量单位。那它是不是属性变量呢？是的，只不过是一种较为特殊的属性变量。就学生的学号这个例子而言，属性数与学生数是一样多的，也就是说有多少学生就有多少类别，并且每个类别只有一个人。当然，每个类别只有一个个体这种现象比较少见，在非常特殊的情形下才会有，比如亚马逊公司不想把你与其他客户弄混了，网站管理人员会给你分配一个特别的账号，这样当你每次登录公司网站时，管理人员就知道是你了。如果每个类别只有一个个体，称这样的变量为辨识变量，除学生的学号外，像UPS的包裹编号、社保账号、亚马逊的图书编号等，都属于辨识变量范畴。辨识变量只有区分类别的价值，不可能有其他用途。不过随着时代的发展，在大数据背景下上述辨识变量的意义也许是非凡的，因为通过这些辨识变量，不仅可以将不同来源的数据彼此联系起来，而且能提供隐私保护和其他特殊的服务。从统计的角度讲，如果你打算去了解辨识变量的作用，那么将同时失去分析这些数据的可能性，比如，没有人会把某个班级今年学号的平均数与上年学号的平均数进行比较，因为这样做毫无意义。

在分析数据之前，需要了解数据背后的“who”、“what”、“why”，要是对这些不清楚，我们就不知道从哪儿下手。当然除此之外，我们了解数据的背景情况越多，