统计学的世界（第8版）

作者：戴维·穆尔

目录

致读者
序言 了解统计学
第1部分 数据的生产
第1章 数据从哪里来
第2章 好样本和坏样本
第3章样本可以告诉我们什么
第4章 真实世界中的抽样调查
第5章 好实验与坏实验
第6章 现实世界中的实验
第7章 数据伦理
第8章 度量
第9章 统计数字会说谎
第1部分 内容回顾
第2部分 整合数据
第10章 好图表与坏图表
第11章 用图呈现数值变量的分布
第12章 用数字描述分布
第13章 正态分布
第14章 描述相关关系：散点图和相关系数
第15章 描述相关关系：回归、预测与因果关系
第16章 居民消费价格指数和政府统计数据
第2部分 内容回顾
第3部分 机会与概率
第17章 思考随机事件
第18章 概率模型
第19章 统计模拟
第20章 赌场的生意经：期望值
第3部分 内容回顾
第4部分 统计推断
第21章 什么是置信区间
第22章 什么是显著性检验
第23章 统计推断的滥用
第24章 双向表与卡方检验
第4部分 内容回顾
附录
版权页

致读者

这是一本写给文科类学生的统计学教科书，也就是说，它是非数学专业学生通识教育的一部分。本书的内容，以本书作者之一在普渡大学文学院为一、二年级学生授课的经验为基础发展而来。我们很高兴地发现，其他教师也认为本书对于各类学生都有用处，甚至对哲学和药学专业的学生也是如此。本书第8版修订了文字，也增加了新的内容，但是目标仍和原版相同：不是把统计学当作技术工具展示给大家，而是作为受过教育的人应有的知识素养。

统计学是人文科学的一部分

作为最不“文科”的学科，统计学的这一名声已被广泛传播，要是有人说它的好话，多半是在说它的用处。医疗保健领域的专业人员必须拥有统计学知识，才能阅读医学研究报告；经理人也要懂统计学，因为需要对大量的数字进行统计分析，才能看得出所以然。普通老百姓只有了解统计学，才能看懂民意调查结果和政府发布的统计数据，例如失业率和居民消费价格指数（CPI）。因为数据和机会无处不在，所以本书的宣传语可以这样写：每个人都用得上的统计学，读者可以从中获利多多。

这些都是事实。我甚至可以说，对于大多数学生来说，本书的通俗讲法要比一般教科书采用的以介绍方法为主的方式更能够为他们将来的统计学研究打好基础。美国统计协会和美国数学协会的联合课程委员会，就建议任何统计学入门课程都应该“强调如何进行统计思考”，而且要体现出“更多的数据和概念，少一点儿公式和推导过程”。本书就是这样做的，其风格贴近文科教育：更多的概念、更多的思考、简单的数据、很少的公式，而且没有正式的推导过程。

应该掌握统计学的另一个理由是：统计学实际上属于人文学科。人文教育强调基本的知识技能，也就是适用于各种领域的通用分析方法。传统人文学科展示了这样的方法：文学和历史研究，政治和人类社会的社会分析，认识大自然的实验科学，数学中的抽象和演绎。之所以将统计学归入人文学科，是基于这样一个事实，即用不确定的经验数据推导得出结论，也是一种通用的思考方法。这本书的两个主题—数据和机会，普遍存在于我们的日常经验中。尽管我们使用数学工具来处理数据和机会问题，但这些工具并不是纯数学的。实际上，心理学家提出了令人信服的观点，认为精通正规的数学方法对于提高我们对数据和机会进行有效推理的能力，并没有多大帮助。

本书在作者能力和读者可接受的范围内，突出了统计学是一种独立且基本的思考方法的观点。本书的重点在于统计思维，也有人称之为量化或数据分析。

本书的性质

统计学的书，有些专谈理论，有些专讲方法，而本书不谈这些。本书讲的是统计思维和统计推断，统计学与公共政策，以及包括从药学到社会学在内的人文科学的关系。我加入了许多基础的图表和数据分析技术，以使思想更生动，推理更有力。学生通过处理数据，可以学到怎样去思考数据。但是，我们没有让技巧超越概念成为主角，我们的意图是用文字而不是代数来教学，我们鼓励讨论甚至争辩，而不是只教授如何计算。当然，有些计算仍然是必要的。本书涵盖的内容要比传统的学期课程更丰富，从通识教育的精神出发，宁可涵盖更广泛也不愿过于追求细节。

虽然本书的风格并不那么正统，但仍然是一本教材。本书内容的编排是为了让学生们系统地学习，而且提供了很多练习，其中有不少需要学生们进行讨论或做出判断。即便那些喜欢从轻松阅读中获得快感的读者，也应该在读完正文后再做做练习。也请教师留意，尽管本书只用到了很简单的数学，但并不是一本轻松的教材。将重点放在观念和推理上，对于读者的要求反而要比通篇都是公式的教科书更高。

第8版新增的内容

本书适用于当下的教育环境，延续了非数学化的方式，将统计学介绍给读者，帮助大家清晰地思考在个人和职业生活中可能会碰到的问题。以下是第8版新增的内容。

•内容。在调查中互联网的使用频率增加了，但也引起了更多的争议。因此，我们在第4章加入了一个关于网络调查的讨论。为了帮助澄清可靠性这个概念，我们在第8章引入了方差，在第17章增加了贝叶斯定理。

•小结。本书每章后面的“小结”包括“本章要点”，总结了该章的主要内容，并将本章和前一章、后一章的内容联系起来。这样做的目的是帮助学生厘清各章之间的关系，以及了解统计学的整体实践。

•案例与练习。相较以前的版本，本书有超过1/3的案例和练习有所更新，以反映当下的数据和各种话题。这些案例和练习涵盖了广泛的应用领域，有助于增加学生的兴趣和亲切感。新的案例和练习包括脸谱网和评级，低脂肪食物和肥胖症，以及短信。

•设计。本书配有极具现代感的插画，以帮助学生更好地理解文字内容。文中的知识普及旨在引导学生了解“现实生活中的统计学”。网上练习现在都已经加了练习编号，可以把学生直接带到本书配套网站，了解最新的内容和资源。

•EESEE（统计案例与练习的电子百科全书）案例分析。由俄亥俄州立大学统计系开发的这些电子案例分析，为学生们提供了及时和真实的数据案例。EESEE增加了超过40个新的案例分析，有些被用作本书每章开头的案例。

•小应用程序。小应用程序是指与本书内容有关的统计应用程序，可以在本书配套网站上获得。有些程序已经做了修改，也增加了一些新的程序。

除了以上新增加的内容，本书还保留了以前各版本的优点。

•案例分析。在每章的开篇，案例分析让学生置身于与本章概念相关的真实生活场景中。每章结尾的“案例评估”会提出问题，促使学生重读开篇案例，评估他们学到的知识。

•统计学中的争议。列出了有争议的问题，并将其与该章的内容联系起来。在“关于‘统计学中的争议’的答案”中给出了后续讨论和解决建议。

•新闻中的练习。从受欢迎的新闻媒体中摘录，反映了当下发生的事件和新的数据来源。

•网上练习。每章都有，这些练习引导学生在网上浏览并认真思考统计数据和概念。

序言 了解统计学

统计学是关于数据的学问。数据就是数字，但又不仅仅是“数字”，数据是有上下文联系的数字。比如，数字10.5本身没任何意义，但是，假如我们听说一位朋友刚出生的孩子体重为10.5磅[1]，那么我们会恭喜她生了一个健康的宝宝。结合数字的上下文联系和个人知识，就可以做出判断。我们知道，体重10.5磅的新生儿个头相当大，而新生儿的体重不太会是10.5盎司[2]或10.5千克。因此，上下文联系可以使数字含义明确。

统计学通过数据获取真知灼见并且得出结论，所用的工具是图表和计算。但是，这些工具要在符合常识的思考方式下使用。让我们先快速地了解一下新闻媒体、热门政治和社会争议的数据和统计研究方面的丰富成果，迈出学习统计学旅程的第一步。在后文中，我们将对序言里列举的案例展开更为详细的研究。

数据胜过逸闻

信仰不能取代算术。

亨利·斯宾塞

一则逸闻就是一个引人入胜的故事，它之所以能够深入人心，就在于其内容能打动人。逸闻可以让事情人性化，所以新闻报道通常以逸闻开头（和结尾）。但是逸闻无法用作决策的依据，正是因其内容生动，所以也常常误导我们。我们判断一个说法可信与否的依据，是它是否有数据支持，而不是它是否来自一个动人的故事。

住得离高压电缆太近，是否会导致儿童患上白血病？美国国家癌症研究所花了5年时间和500万美元，来收集有关这个问题的数据。结果发现，在白血病和暴露在高压电缆产生的电磁场之间没有任何相关性。和这篇研究报告同时发表在《新英格兰医学期刊》上的社评大声疾呼，是时候不要在这个问题上“再浪费我们的研究资源了”。

比较一下两则报道产生的影响：一则是有关历时5年、花费500万美元的调查研究的电视新闻报道，另一则是对一位能说会道的母亲的电视采访，她患有白血病的孩子恰好住在高压电缆附近。在公众心目中，几乎每次都是逸闻胜出。但事实上，我们应该心存疑问，因为数据要比逸闻更可靠，数据可以系统地描绘出整体图景，而不是聚焦于个别事件。

我还打算补充一句，“数据胜过自封的专家。”新闻行业通常所秉持的所谓公正的立场，就是各找一名正反方“专家”发表简短的评论。我们永远无法知道是否其中一方代表的是整个学科的共识，而另一方则是蝇营狗苟于某些特殊利益的骗子。媒体这种处理冲突的做法使现在的公众认为，对应每一位专家，都有一个持相反观点的专家。如果你真的关注某个话题，就应该尝试从数据中找出结论，还要判断数据的质量是否良好。很多问题确实还未得到解决，但也有很多问题，只在那些不重视证据的人的头脑中才悬而未决。你至少可以了解一下这些“专家”的背景，还有他们引用的研究结果是否曾刊载于那些文章必须先通过外部评审之后才能发表的严谨期刊上。

数据来源非常重要

数字不会说谎，但说谎的人会算计。

查尔斯·格罗夫纳

数据是数字，而数字总显得非常可信。其实有的数字是这样，有的却并非如此。任何统计研究中最重要的事，就是数据的来源。专栏作者安·兰德斯问她的读者，如果可以重新来过，是否还会生孩子？回答的人中有70%坚定地说“不要”。但是，你对于兰德斯从泪迹斑斑的信中，引述读者哭诉他们的孩子如何像野兽一样的话语，大可一笑置之。因为兰德斯从事的是娱乐业，她邀请读者回答这个问题时，回应最

热烈的应该是那些后悔生孩子的父母。大部分的父母并不后悔生小孩。我们知道这个事实，是因为曾经有人对许多父母做过调查，而且为了规避偏见，受访的父母是随机抽取的。民意调查当然也不是没有瑕疵——这点我们后面会谈到——但它仍比邀请有一肚子不满的人来回答问题要好得多。

即使是信誉卓著的期刊，也不一定能对坏数据免疫。《美国医学会杂志》刊登过一篇文章，声称用管子将冷却的液体输送进胃里，可以缓解溃疡症状。这种治疗对病人的确有效果，但那只是因为病人通常会信服医生的权威性。也就是说，安慰剂（placebo，虚拟治疗）发生了作用。后来终于有人产生了质疑，做了有控制组的比较研究——有些病人接受了这项治疗，有些只使用了安慰剂——结果表明安慰剂组（控制组）的“表现”更好。“没有比较，就没有结论”，这是判断医学研究好坏的一个很好的依据。比如，我对于最近突然流行的“自然疗法”就心存怀疑。在这些疗法当中，极少会通过比较实验来证明它们不只是装在贴着漂亮植物图片的瓶子里售卖的安慰剂。

小心潜在变量

我的钱足够我这辈子花了，只要我不买东西。

杰基·梅森

有报道说，在美国，设有赌场的县犯罪率较高。有位大学教师说，在线上课的学生比在教室里上课的学生表现好。政府报告强调，受过良好教育的民众比受教育少的民众赚钱多。最好不要太快下结论，而要先问一句：“有没有什么是他们没告诉我的，却可以用来解释这些事？”

有赌场的县的确犯罪率较高，但是在城市或是较穷的县，犯罪率也比较高。什么样的县会开设赌场呢？这些县是否在开设赌场之前，犯罪率就已经很高了呢？在线上课的学生学得比较好，但是和在教室里上课的学生比起来，前者的年龄比较大，底子也比较好，因此他们表现好很正常。受教育程度高的人的确赚钱多，但受教育多的人和受教育少的人相比，平均来说前者的父母受的教育也较多，也比较有钱。而且，前者在更好的环境中长大，上更好的学校。这些有利条件让他们接受到更多的教育，即使他们不接受这么多教育，这些有利条件也可能会帮他们赚到很多钱。

以上这些案例，都涉及两个变量之间的相关性，并且引导我们得出结论，即其中一个变量影响了另一个变量。“赌场会增加犯罪率”和“想变富有就要多读点儿书”，是它们传达的信息。这些信息有可能是正确的，但我们看到的相关性，大部分说不定还可以由隐藏在背景中的一些变量来解释，比如开设赌场的县的本质，以及受教育程度高的人生来就具备的有利条件。好的统计研究会考虑许多的背景变量，这当然需要技巧，不过你至少可以找一找，看它们有没有这样做。

变异性无所不在

当事实改变时，我就改变主意。您呢？

约翰·梅纳德·凯恩斯

如果你的舌下温度超过37摄氏度，是不是就代表你在发烧呢？也许不是，每个人的“正常”体温都存在差异。你自己的体温在一天当中也会有变化，早上6点时稍高，下午6点时稍低。美国政府宣布上个月的失业率上升0.1个百分点，而新屋开工率下降3个百分点，导致股市波动。股市波动常常是不理性的。政府的数据是根据那些给出乐观预测的样本得出的，是好的估计，但不是百分之百的事实。同样的调查再做一遍，结果会有一些差别。和经济有关的因素原本就会上下波动，影响它的因素包括气候、罢工事件、节假日以及各式各样的其他原因。

很多人会像股市一样，对数据的一些微小变化做出过度反应，而事实上这些并不是实质性变化，根本微不足道。以下是全美最大的市场研究公司的领导人阿瑟·尼尔森的经验之谈：

太多商界人士对白纸黑字印出来的数字都深信不疑。他们认为数字就代表“事实”，让他们用“概率”的视角来看事情有点儿困难。他们不了解，数字只是一种简化的表示方法，它描述了一个范围，这个范围表

明我们在现有情况下对某个问题的实际了解有多少。

变异性无所不在。个体之间有差别，对同一个个体多测量几次，结果也会不一样，并且几乎每件事都会随着时间的变化而变化。至于一些所谓的专家对每天股市变化背后的深入原因分析，或者对一场其实是最后一秒才定输赢的球赛，却硬要归咎于输球球队的能力或战术，你尽可以对这些嗤之以鼻。

统计结论不是百分之百确定的

数学定律不能百分之百地应用于现实生活，能百分之百地用数学定律来描述的就不是现实生活。

阿尔伯特·爱因斯坦

因为变化无所不在，所以统计结论并不是绝对的。大部分中年妇女会定期做乳腺X线检查，以便能及早发现乳腺癌。但是，乳腺X线检查真的可以降低女性死于乳腺癌的风险吗？高质量的统计研究发现，对于50~64岁的妇女来说，乳腺X线检查可以减少26%的死亡率。但这个数字是该年龄阶段妇女的一个平均数字。因为总是存在着个体差异，所以对于不同的妇女来说，结果可能大不相同。有些每年做乳腺X线检查的妇女最后却死于乳腺癌，也有一辈子没做过这种检查的妇女却活到了100岁，最终死于车祸。

该报告事实上说的是，“乳腺X线检查可以将乳腺癌死亡率减少26%（置信度为95%，置信区间为17%~34%）”。而根据尼尔森的观点，26%这个数字是“某个范围的简化说法，这个范围描述了我们对现时情况的实际了解程度”。在这个案例当中，这个范围是17%~34%，而且我们对它有95%的把握。也就是说，我们相当有把握，但不是百分之百确定。当你阅读新闻报道时，可以找找看有没有诸如“95%的置信度”或“有统计学显著性”等字眼，如果有，就代表这个研究结果虽不能说百分之百确定，但有很大的把握。

数据可反映社会价值

要用统计数据骗人很容易。但如果不用统计数据，骗人就更容易了。

弗里德里克·莫斯特勒

好的数据确实胜过逸闻。比起逸闻和大声争论未来会发生什么，数据要客观得多。和其他的公开论述相比，统计数据建立在事实的基础上，而且比较科学和理性。对于存在争议的话题，统计数据应该比其他大部分证据受到更多的重视。不过，世界上没有百分之百客观这回事。社会环境会影响我们关于度量什么以及如何度量的决策，从而影响到统计数据。

以自杀率为例，它在不同国家之间有很大的差别。各国公布的自杀率之所以存在差别，似乎大部分要归因于社会观念，而不是因为自杀率真的有差别。自杀人数是根据死亡证明书计算的。填写死亡证明书（证明书上的细节在不同的国家或不同的州之间有所不同）的官员，对于诸如没有目击证人的溺死或摔死事件，可以决定其是否算作自杀。在视自杀为耻辱的地方，就有较多的自杀被报告为意外死亡。比如，在大部分人信奉天主教的国家，其公布的自杀率就比其他国家低。在日本文化中，有遇到羞辱时就光荣自杀的传统，这种传统使人们不太将自杀视为耻辱，因此日本的自杀事件被报道的比例较高。在有些国家，因为社会价值观的改变，使得自杀事件的数量在增加。一个越来越普遍的看法是，抑郁症应该被视为生理上的疾病，而不是人性的弱点，自杀是这种疾病的悲剧性结局，而不是道德瑕疵。因此，死者家属和医生愿意把自杀列为死者的死因。

在不像自杀那么敏感的问题上，社会价值观依然可能会影响统计数据。美国的失业率是每个月由美国劳工部劳工统计局用一个很大而且很专业的全国性样本计算出来的。但是，“失业”的定义是什么呢？它指的是，虽然你想工作却没有工作，而且过去两周在积极找工作。如果你过去两周没在找工作，你就不算失业人员，而只是“非劳动人口”。这样去定义失业率，反映了我们加诸工作的价值观。如果换一种定义，我们可能就会得到不一样的失业率数字。

我的意思并不是叫你不要相信失业率数字。失业率的定义已经使用了好一段时间，在各个国家之间，这

个定义也大致吻合，所以我们可以在国家间做比较。失业率是由不受政治干扰的专业人士计算出来的，是重要且有用的信息。我要强调的是，并不是每件重要的事都可以用数字表示。此外把事情简化为数字的人，会受到各种压力的影响，不管有意的还是无意的。

统计学与你

这不是一本关于统计工具的书，本书写的是统计学的观点及其对日常生活、公共政策和其他不同研究领域的影响。当然，你将会从本书中学到一些统计工具。如果你手上有一台内置了统计功能的计算器，你的生活将会容易很多，特别是这台计算器可以计算平均值、标准差和相关系数。找一台可以做“双变量统计”或“相关性分析”的计算器，如果你的电脑上有统计软件，就更好了。另一方面，你需要懂一点儿数学。如果你理解并会使用一些简单的方程式，就再好不过了。不过要记住，我们要求你学会思考，思考对于大脑的训练要远胜过数学公式。本书将从4个方面展现统计学思维方式：

•数据生产讨论了数据产生的方法，可以对特定问题给出解决方案。数据来自何处至关重要，关于如何选择样本和设计实验的基本概念，是统计学中最有影响力的内容。

•数据分析关注的是收集、组织，以及用图表、数字描述数据的方法和策略。你可以学会使用非常简单的工具聪明地分析数据。

•概率是我们用来描述机会、变化和风险的语言。由于变化无处不在，概率思维有助于将现实与背景噪声区分开。

•统计推断是指通过我们手中掌握的数据得出结论，考虑到变化无处不在，结论是不确定的。

最终数据会被用于得出结论或做出决定。本书的各个部分讨论了影响这一过程各个步骤的质量问题。读者很容易把注意力放在每章的细节上，而忽略了这些细节背后的整体图景。为了帮助你避免这种见树不见林的问题，看清楚每章与整体图景的关系，我们在每章结尾处简短地说明这一章在整体图景中的位置。你可以在每章的小结中看到这部分内容。

统计思维和工具在人们处理数据的过程中逐步显现。两个世纪前，天文学家和勘测员面对的问题是，即便把许多观察结果非常小心地拼凑起来，仍然发现它们无法很好地吻合。他们努力处理数据的变化，因而产生了最初的一些工具。当社会科学在19世纪出现时，旧的统计思维发生了转变，新思维被发现，用于描述个体和社会的变化。对遗传和生物学总体变量的研究推进了统计学的发展。20世纪前半叶诞生了建立在概率之上的用于生成数据和进行统计推断的统计设计。直到20世纪中期，人们才清楚地看到了一门新学科的诞生。由于所有研究领域都更加注重数据，而且越来越明显地意识到数据变化是无法避免的，统计学已经成为核心知识。每个受过教育的人都应该了解统计推断，阅读本书，能够让你掌握这方面的知识。

[1]1磅≈0.454千克。—编者注

[2]1盎司≈28.35克。—编者注

第1部分 数据的生产

你和你的朋友不是典型人物。比如，你喜欢听的广播节目可能就和我喜欢的不一样。当然，我和我的朋友也不是典型人物。如果想要知道整个国家（或只是大学生）的状况，我们必须认清一个事实，那就是，整体状况也许和我们自己或我们周围的状况并不相似。所以，我们需要数据。从Arbitron（一家媒体研究机构）得到的资料显示，美国最受欢迎的广播节目是成人当代音乐（2009年秋天，每周的听众总人数达到7200万人）和乡村音乐（每周的听众总人数为6400万人）。如果你喜欢当代流行打击乐（每周的听众总人数为5900万人），而我喜欢各种新闻节目（每周的听众总人数为1500万人），那么我们可能对整个唱片消费群体的音乐品位一无所知。如果我们身处唱片业，或只对通俗文化感兴趣，就必须把自己的喜好放在一边，好好地检视数据。

你可以到图书馆或者网上找数据，但我们怎么知道这些数据可不可靠呢？好的数据可以被看作像毛衣或数码影音光碟机（DVD机）之类的产品，而草率生产的数据就像草率生产的毛衣或DVD机一样令人沮丧。你买毛衣前一般都会检查一下，如果做工很差，你就不会买。数据也一样，如果质量很糟，你就不该用。本书的第一部分将会告诉你怎样分辨数据的好坏。

第1章 数据从哪里来

案例分析

你可能读了几个月的报纸，或者看了几个月的电视新闻，却没有见过一个代数公式。难怪你会认为代数看上去和日常生活完全无关。但是，你不可能一整天都接触不到统计数据和结果。比如，一则新闻提到在18~24岁年龄段的人中，有70%的人认为从互联网上下载音乐与购买二手音乐光碟或从朋友那里借来转录的做法没有区别。而对于65岁或以上的人来说，持相同观点的人只有36%。一篇更长的文章写道，得到高质量日间看护的低收入家庭的孩子，几年后在学业方面的表现更佳，与其他低收入家庭的孩子相比，未来更有可能考上大学、找到好工作。

这些数据是从哪里来的呢？为什么我们可以相信这些数据，或者不应该相信这些数据？好的数据是人类智慧劳动的成果，坏的数据则源于懒惰或者不求甚解的行为，只能误导他人。当有人向你抛来一个数字时，你应该问的第一个问题就是：它来自何处？

2008年美国总统大选期间，有一个全国性的项目叫作“大家去投票”，人们从罐中选择一粒软糖，代表他们支持的总统候选人。在佛罗里达州彭布罗克派恩斯的“咖啡苑”（Coffee Scene），顾客可以在每次买东西时挑选一粒蓝色或者红色的软糖。蓝莓软糖表示投票给民主党总统候选人巴拉克·奥巴马，红色草莓软糖表示投票给共和党总统候选人约翰·麦凯恩。

全美有34个州的超过100家餐馆参与了这项活动，每周一清点投票结果。截至2008年10月13日，投票总数超过211000张，麦凯恩在这场全国性的软糖投票中获得了53.5%的支持率。2004年美国总统大选期间的软糖投票结果，曾准确预测出小布什将赢得总统大选。约翰·麦凯恩却没有那么幸运，投票结果错误地预测麦凯恩将在2008年的美国总统大选中胜出。

对于这个投票结果，我们该如何解读呢？在本章的最后，你将了解到，针对这个软糖投票活动产生的数据，如何提出一些基础性问题。对这些提问的回答，将帮助我们评估这个投票数据的优劣，我们将在第2章深入探讨这个问题。

个体和变量

统计学是关于数据的科学。我们甚至可以说它是“数据的艺术”，因为只有通过正确的判断乃至好的品位和扎实的数学知识，才能得出好的统计数据。好的判断中的一大部分取决于你应该度量什么，以使得出的数据有助于解释你所关心的问题。我们先从一些与数据相关的基本术语入手。

个体与变量

个体是一组数据所描述的对象。个体可以是一群人，也可以是一些动物或者其他物品。变量指的是某一个体的特征，不同个体的变量可以取不同的值。

举个例子，这是统计学课程结束时，教授手中的成绩数据库中的头几行：

在这里，个体指的是选修这门课程的学生。除了学生的姓名之外，还有三个变量。第一个变量是这名学生的专业，第二个变量是该名学生所得的分数（总分为500分），第三个变量是该名学生的成绩等级。

尽管统计学是和数字打交道，但并非所有的变量都要用数字表示。有些变量代表的是“类别”，只需将个体归入不同的组别或种类即可。在这个例子的三个变量中，只有总分数是用数值表示的，专业和等级都是类别变量。统计类别变量时，我们采用的是计数或者百分比的方式。例如，我们可以给出得到等级A的学生的百分比，或者给出主修心理学的学生的百分比。

选择变量时的错误判断，可能导致在数据上浪费大量时间和金钱却徒劳无功。怎样才能做出好的判断，这可能是个富有争议性的问题。下面的例子展示了在判断应该收集哪些数据时面临的挑战。

例1 谁在做资源回收？

是谁不嫌麻烦地在做资源回收呢？在加利福尼亚州某个城市的两个社区，暂且命名为“上流社区”和“中低社区”，实验人员花费了大量的时间和金钱，将住户拿到屋外的垃圾过磅称重。因为垃圾和回收物的数量是按照住户而不是人头计算，在这里，个体就是两个社区的所有住户，变量是每周路边回收箱里垃圾的重量。

“上流社区”每周平均的垃圾重量要比“中低社区”大，我们是否可以据此说富人更严肃地对待垃圾回收这件事呢？不能。有人注意到，“上流社区”的垃圾桶里有很多沉重的玻璃酒瓶。而在“中低社区”，人们扔进垃圾箱的多是轻质塑料汽水瓶和啤酒铝罐。结论是，重量很难告诉我们人们履行资源回收承诺的实际情况。

例2 你是哪个种族？

全美人口普查时调查人员会问“你属于哪个种族”，每个家庭成员都要回答这一问题。“种族”是一个变量，美国人口普查局必须给出种族的准确定义。普查表格上列出了一个种族名单，多年来美国政治能够顺顺当当不出问题，主要归功于这个名单。

名单上应该列出多少个种族？用什么名称来代表它们？需要把混合种族列为一个独立的类别吗？亚裔人想要划分出更多的种族，比如菲律宾人和越南人，因为亚裔人口一直在增长。太平洋岛国的人希望与其他较大的亚洲族群分开。黑人领袖不希望出现混合种族类别，因为担心太多的黑人选择这个类别，将减少黑人人口的官方统计数量。

2010年的普查表（图1–1）最终列出了6个亚洲族群（加上“其他亚裔人”）和3个太平洋岛国族群（加上“其他太平洋岛国人”），表上没有出现混合种族一项，但人们可以选择多个选项。也就是说，填报混合种族的人可能会被计入多个族群，因此这个数据会比总人口大。因为无法确定如何称呼黑色人种，人口普查局用了“黑色皮肤，非裔美国人或黑人”这一说法。拉丁裔呢？这是另外一个问题，因为他们可能属于任何一个种族，而且无法找到一个能让所有人都满意的说法来代表他们。因此，只能问人们是否属于“西班牙裔、拉丁裔或祖先来自西班牙”。

在种族问题上存在的争执提醒我们，数据反映了社会。种族是一个社会概念，而不是一个生物学事实。在普查中，你可以决定自己属于哪个种族。种族在美国是一个敏感问题，出现争执也不足为奇。人口普查局的方法似乎是一种不错的折中之策。

图1–1 2010年美国人口普查表第1页

观察研究

纽约扬基队的前接球手和经理约吉·贝拉，以其幽默的话语为人所知。他曾说：“你只要观察，就能看到很多东西。”有的时候，你能做的也只有观察了。想知道黑猩猩在野外的行为，你得观察。想研究老师和小朋友在教室中的互动行为，你也得观察。若观察者知道自己应该看的是什么，就会容易一些。黑猩猩专家可能对雌猩猩和雄猩猩的相互影响有兴趣，或者想知道黑猩猩群中的某几只黑猩猩是否有统治权，以及黑猩猩会不会猎食动物。事实上，大家一直以为黑猩猩是食草动物，直到珍妮·古道尔在坦桑尼亚的贡贝国家公园仔细观察了它们的行为。现在人们已经清楚地知道，肉类是黑猩猩日常饮食的一部分。

刚开始的时候，观察者可能不知道应该记录些什么。然而迟早会出现一些模式，帮助我们决定应该衡量哪些变量。比如，黑猩猩多久猎食一次？单独行动还是群体行动？多少只一起行动？只有雄性还是雌、雄都有？在它们的食物中肉类占多大比重？系统地观察，再加上清楚地定义变量，会比只观察更有说服力。以下是一个计划完善（且很花钱）的观察研究案例。

例3 高压电缆会引发儿童白血病吗？

欢迎访问：电子书学习和下载网站（https://www.shgis.cn）
文档名称：《统计学的世界（第8版）》戴维·穆尔 著.epub
请登录 https://shgis.cn/post/1109.html 下载完整文档。
手机端请扫码查看: