让数据告诉你

陆立强　　编著

博学·数学系列

复旦大学出版社
www.fudanpress.com.cn

让数据告诉你

陆立强　编著

博学·数学系列

复旦大學出版社
www.fudanpress.com.cn

图书在版编目(CIP)数据

让数据告诉你/陆立强编著. —上海:复旦大学出版社,2008.2
(博学·数学系列)
ISBN 978-7-309-05881-9

Ⅰ. 让… Ⅱ. 陆… Ⅲ. 统计学 Ⅳ. C8

中国版本图书馆 CIP 数据核字(2007)第 204151 号

让数据告诉你

陆立强 编著

出版发行 复旦大学出版社 上海市国权路 579 号 邮编 200433
86-21-65642857(门市零售)
86-21-65100562(团体订购) 86-21-65109143(外埠邮购)
fupnet@fudanpress.com http://www.fudanpress.com

责任编辑 范仁梅
总 编 辑 高若海
出 品 人 贺圣遂

印 刷 浙江省临安市曙光印务有限公司
开 本 787×960 1/16
印 张 12.75
字 数 235 千
版 次 2008 年 2 月第一版第一次印刷
印 数 1—4 100

书 号 ISBN 978-7-309-05881-9/C·98
定 价 20.00 元

博学·数学系列

《数学分析（上下册）》
（普通高等教育“十五”国家级规划教材）
欧阳光中　姚允龙　周　渊　编著

《高等代数》
（普通高等教育“十五”、“十一五”国家级规划教材）
姚慕生　吴泉水　编著

《数学模型》
（普通高等教育“十五”、“十一五”国家级规划教材）
谭永基　蔡志杰　俞文鱿　编著

《空间解析几何》
黄宣国　编著

《概率论》
应坚刚　编著

《数理统计讲义》
（普通高等教育“十一五”国家级规划教材）
郑　明　陈子毅　汪嘉冈　编著

《常微分方程》
楼红卫　林　伟　编著

《抽象代数学》（第二版）
姚慕生　编著

《数值逼近》（第二版）
（普通高等教育“十一五”国家级规划教材）
蒋尔雄　赵风光　苏仰锋　编著

《线性代数与解析几何》
郑广平　裘祖干　陆章基　编著

《文科高等数学》
（普通高等教育“十一五”国家级规划教材）
华宣积　谭永基　徐惠平　编著

《让数据告诉你》
陆立强　编著

“博学而笃志，切问而近思。”

（《论语》）

博晓古今，可立一家之说；

学贯中西，或成经国之才。

内容提要

在五彩缤纷的现实世界中，到处充斥着数字．这些数字有时会让人看得眼花缭乱，使人心绪不宁．因此，数据的收集、处理、分析尤为重要．掌握正确的数据收集、数据处理、数据分析的方法，由表及里、去伪存真，是人们在学习、生活、工作中必不可少的．

本书用一种比较通俗的方式向大学生介绍数据分析的基础知识和基本方法，以帮助他们全面理解和正确把握数据、培养定量化的思维方式．

本书具有以下特点：叙述浅显，书中假设本书读者没有学过《高等数学》课程，所以全书没有包含任何数学公式的推导，而采用叙述的方式引入重要的概念，同时把计算公式压缩到最低的限度；案例丰富，书中大量采用案例引入主题；内容完整，本书除介绍数据采集和数据分析外，还介绍了概率和数据决策方面的内容．

前　言

数据在人们的工作和生活中不可避免，但往往也给人枯燥单调的感觉，所以大多数人对它抱着敬而远之的态度. 随着信息化时代的到来，各式各样的数据如波涛般涌入社会生活的方方面面，面对这样汹涌的数据浪潮，人们要么被它弄得晕头转向，最终被淹没；要么努力掌握其规律，为人所用，做一名数据弄潮儿.

《让数据告诉你》试图采用一种比较通俗的方式为大学生，尤其是为人文、社科专业大学生，介绍数据分析的基础知识和基本方法，帮助他们全面理解、正确把握数据，在专业学习以及今后的实际工作中习惯运用定量化的思维方式，使看似枯燥无味的数据成为探求真理、解决问题的好帮手. 基于上述思想，本书具有以下特点：

1. 叙述浅显

我们假设读者没有学过《高等数学》课程，所以全书没有包含任何数学公式的推导，采用叙述的方式引入重要的概念，同时把计算公式压缩到最低的限度.

2. 案例丰富

考虑到读者来自不同学科和专业，同时也为了说明数据分析方法的广泛应用，本书中大量采用案例引入主题，除了个别案例选自各个学科的专业杂志和书籍外，大多数案例来自发行量大、读者多的报纸和杂志.

3. 内容完整

除了数据采集和数据分析，本书还包含了概率和数据决策方面的内容，因而较全面地向读者展示数据及其应用的整体构架，激发他们进一步学习相关知识的兴趣.

本书分 4 个部分. 第一部分主要介绍正确的数据收集的过程和方法，帮助读者识别媒体报道中数据的真伪，做一个清醒的数据“消费者”，同时提醒读者在实际工作中避免因为数据本身的错误而导致的错误结论. 第二部分主要介绍数据分析的基本概念和方法，读者可以从涉及经济、政治、法律、社会、心理等各个学科的丰富的案例中了解到数据给人们的工作和生活带来的方便，为进一步学习打好基础. 第三部分通过一些案例，引入概率的基本概念，其中结合心理学来讲述主观概率，这对于帮助读者理性地面对和处理生活中出现的不确定因素会有相当大的启发. 第四部分讲述如何正确运用数据，以减少决策失误的可能性.

本书的酝酿和编写自始至终得到了复旦大学教务处的支持和鼓励，复旦大学数学科学学院邱维元教授、金路教授、楼红卫教授也非常关注本书的编写和出版，在此致以衷心的感谢. 本书在正式出版前，已经连续 3 年在复旦大学作为综合教育课程的教材试用，学生们在使用过程中提出了不少建议和意见，促进了本书内容的完善和质量的提高. 最后，限于作者的水平，书中缺点和错误在所难免，希望读者指正，联系方式：malqlu@fudan. edu. cn.

作　者

2008 年 1 月于复旦大学

目　录

第一章　统计的利弊

问题

- 有人认为：办学规模大的公立大学的毕业生中最终成为百万富翁的人数要比办学规模小的文史类学院的毕业生中产生的此类人数多．你认为如何？怎样才能证明这个结论？
- 从理论上讲，男性在静止状态下每分钟脉搏数小于女性．如何进行实际测量以证明这个理论？

1.1　统　　计

在大多数人心目中，“统计”两个字往往意味着诸如某市最近一次人口普查结果、制造业工人的平均收入等一堆枯燥乏味的数字、一个个复杂的数学符号和公式，令人望而生畏．本书旨在帮助读者从一个新的角度来理解和欣赏统计的知识和方法，使读者读完本书后就会知道：从医学成果的研制到电视节目的编播，人们生活的方方面面都受到了统计的影响，统计方法是现代社会最重要的发明成果之一．

“统计”一词实际上有两种不同的解释，其中广为人知的一种解释是：**统计就是为了某种需要而采集的一组数据**．其实统计的另一种更确切完整的定义是：**统计是在决策过程中获取和处理信息所采用的一系列步骤和准则**，具体包括：数据收集、数据分析和结论推断 3 个方面．

根据第二种定义，我们可以发现，日常生活中大多数人已经在不知不觉中应用了统计．比如，你每天读书或者上班有多种线路可以选择，为了确定哪条线路更方便，你会沿着每一条线路反复走几次，然后根据你认为的某些重要条件，如时间长短、红灯数目甚至路边的景观等，选择其中最适合的一条线路．选择路线的条件还会随着天气、季节的变化而发生改变．在这个简单的例子中，你的行为实际上就是一种统计，因为你对多条线路的情况作了采集和比较，从中获取了有用的信息并加以处理，最终帮助你做出决定．

读了本书，你将学会如果在面临比以上情况更为复杂的信息时该如何更巧

妙地改进收集和处理信息的方法;学会如何解读别人采集处理的信息;掌握面对不确定因素时的决策方法.

案例 1.1 控制人的情感的是心脏还是下丘脑?

尽管下丘脑在控制人类情感方面起着重要的作用,但是有一位心理学家注意到这样一种现象:在诗句和歌词中充斥着"爱心"、"全心"之类的词语,人们也习惯于用"衷心感谢"或者"我从心底里爱你"这样的词句表达自己的感情,但是没有人会因激动不已而脱口而出"我发自下丘脑地感谢你"或者"我发自下丘脑地爱你",于是就产生了这样的疑问:为什么人们在表述情感时会如此偏爱心脏?

带着这个问题,这位心理学家对心脏在人际关系中的作用产生了兴趣,他的研究从观察动物园中猕猴的行为开始.他发现其中一只母猴在 42 个不同场合下环抱小猴的动作中,有 40 次是抱在左胸位置.然后,他进一步观察出生不满 4 天的婴儿母亲们的行为,发现在 287 位母亲中有 237 位把婴儿抱在左胸,如果将她们按平时用手习惯分成两组,那么这些人在右撇子中占了 85%,在左撇子中占了 78%.至于为什么会习惯把孩子抱在左边,右撇子说这样可以解放右手,左撇子则认为这样对孩子更好些.也就是说,双方都把自己的行为解释成喜欢,而不是受用手习惯的影响.

于是,这位心理学家想知道,除了抱孩子使用左手以外,人们在拿其他东西的时候是否也习惯用左手.于是,他在超市门口观察购物后手里只提一个购物袋的顾客,结果发现在 438 位成年人中用左手的恰好占一半.但是当人处于紧张状态时,情况就不同了.如牙科医生在治疗时往往会要求患者手握一只橡皮球来转移注意力,这时候有一大半的人用左手拿着橡皮球.

至此,这位心理学家大胆猜测:"任何生命体所呈现的生物性倾向不是大自然的造化,而是因为生存的需要."也就是说,大多数母亲把新生儿抱在左胸实际上是因为妈妈的心跳对他们的生存是非常重要的.为了证明自己的猜测,他设计了一个测试方案并在纽约一家市立医院的婴儿室加以实施.在实验中,护士们不间断地让一组婴儿听人的心跳声,这样连续 4 天以后,测量他们的体重变化;另一组婴儿则不听心跳声,4 天后也测量体重.结果,在进食等情况相同的前提下,听心跳声音的那组婴儿的体重增加数要大于不听心跳的那一组(即使体重减少了,其减少数量也小于另外一组);进一步还发现,前者啼哭的时间也更短.据此,他得出的结论是:"成年人的正常心跳声可以安抚新生儿."这也就是新生儿的母亲会不知不觉地把孩子抱在左侧的原因.

以上案例说明了在正确的统计方法指导下,科研人员是如何在对一次自然现象的简单观察中,逐步探索发现母亲和新生儿之间的一种重要互动作用.

1.2 如何应用统计来发现规律、验证关系

生活中某些明显的差别用眼睛就可以观察到，比如说：男性的平均身高要超过女性等. 但是世界上还有许多现象和规律，单靠眼睛观察是不够的. 你能用眼睛观察到“听心跳声的婴儿长得更快”、“服用阿司匹林可以防止心脏病”这样的现象吗？如果有人告诉你“蓝色牛仔裤在某几个月比其他几个月卖得更好”，“莫扎特的音乐可以提高智商测试中与空间辨别力相关的成绩”，你会相信吗？这些关系都不是凭肉眼可以观察到的，而需要采用适当的统计方法研究以后才能加以证实.

那么，怎样才能使人相信你所发现的规律？下面我们举一个简单例子. 为了证明“在静止状态下，男性的平均脉搏数要小于女性”，读者可能会先测量自己每分钟的脉搏数，再找一个异性朋友再测量一下，最后进行比较. 问题是这样是否能足以说明上述结论的正确性？答案显然是否定的，因为一组数据根本无法代表所有的男性和女性.

上述例子告诉我们，对于未经训练的人，要求其用严格的方法来完成某项研究是不太容易做到的，但是经过简单的训练后，他们大多能够理解专家在研究中所采用的方法. 本书主要内容将围绕定量分析研究中的统计方法展开. 首先我们结合上述例子来说明其中的 3 个要点.

1. 样本要有代表性

为了体现研究成果的重要性，大多数研究人员希望将基于部分参与者的研究结果推广到更大的群体，这样的话，研究对象在大群体中是否具有代表性就十分重要. 为了便于叙述，以后我们将参与研究的对象或者人员称为**样本**(sample)，样本所属的大群体称为**总体**(population)(如何选取合适的样本将在第四章介绍). 对于心跳比较问题，将某个班级同学的脉搏数作为样本可能是一种比较便利的方法，但是如果此班级中存在影响心跳与性别关系的因素(例如，学校男子田径队队员全部都在此班上)，或者研究者希望把结论推广到和此班级同学年龄分布相差较大的年龄组，那么上述样本的代表性是有问题的. 尽管如此，还是有许多研究人员会因为这样或者那样的原因而被迫使用类似数据作为样本，这种样本通常称为“便利”样本，其含义将在后面进一步说明.

2. 样本要足够大

即使有经验的研究人员也经常会因为忽略样本个数的重要性而得出错误的

结论.还是以心跳问题为例,我们知道,将自己的脉搏数和一位异性朋友的脉搏数比较一次就去验证上述结论肯定是不行的,那么,该比较多少人才算数?2个人还是4个人?100个人够吗?这取决于研究者采集的脉搏数的差异程度.如果连续几次测量所测得的男性脉搏数都是每分钟65次,女性都是75次,那么很容易得出男女脉搏数存在差异的结论.但是,如果男性脉搏数为每分钟50次到80次之间,女性脉搏数为每分钟52次到82次之间,凭直觉我们知道需要测量更多的数据,但问题是我们究竟需要多大的样本?本书会告诉我们如何根据两组测量结果的差异确定所需的样本数.

3. 研究方法要明确

验证某个关系,一般有**观察法**(observational study)和**实验法**(experiment)两种方法.如果研究人员只是对样本的某些事项感兴趣,一般采用观察研究就可以了.比如对于心跳速度差异问题,我们只需观察(记录)样本中每个人的性别、脉搏就足够了.但是,对于"常服阿司匹林可能会防止心脏病突发"这样的问题,单纯依靠观察某个人是否常服阿司匹林以及他是否得了心脏病是不够的,因为那些关心自己健康的人在常服阿司匹林的同时得心脏病的可能会少一些,而那些不关心健康的人不常服阿司匹林同时也容易得心脏病.

为了证实因果关系,必须做实验,也就是先采用类似扔硬币的方法,把样本随机分成两组(这个过程称为**随机指派**(random assignment)),然后给其中一组服用药片,另一组则服用外观和真药一模一样的替代品.同时,为了避免实验对象受到我们期望结果的干扰,在实验结束之前所有的人员都不知道自己服用的是药还是替代品.下面我们通过案例简单介绍实验过程,其中的思想和方法将在第五章中详细讨论.

案例1.2 阿司匹林能够防止心脏病吗?

1988年,美国医生健康状况调查研究小组指导委员会公布了一项有22 071位男性医生参与的历时5年的实验结果,这些医生的年龄在40岁到84岁之间,结果的主要内容如表1.1所示.

表1.1 阿司匹林对心脏病的作用

条 件	心脏病患者数	未患心脏病者数	每千人患病数
服用阿司匹林	104	10 933	9.42
服用安慰剂	189	10 845	17.13