TURING 图灵程序设计丛书

Think Stats

统计思维

程序员数学之概率统计

[美] *Allen B. Downey* 著

张建锋 陈钢 译

O'REILLY®

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

统计思维
程序员数学之概率统计

Think Stats

[美] Allen B. Downey 著
张建锋 陈钢 译

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo
O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

统计思维 ：程序员数学之概率统计 / （美）唐尼(Downey,A.B.）著 ；张建锋，陈钢译. -- 北京 ：人民邮电出版社，2013.6
（图灵程序设计丛书）
书名原文：Think stats
ISBN 978-7-115-31737-7

Ⅰ. ①统… Ⅱ. ①唐… ②张… ③陈… Ⅲ. ①概率统计 Ⅳ. ①O211

中国版本图书馆CIP数据核字(2013)第086989号

内 容 提 要

《统计思维：程序员数学之概率统计》是一本以全新视角讲解概率统计的入门图书。抛开经典的数学分析，Downey 手把手教你用编程理解统计学。概率、分布、假设检验、贝叶斯估计、相关性等，每个主题都充满趣味性，经编程解释后变得更为清晰易懂。

本书研究数据主要来源于美国全国家庭成长调查（NSFG）与行为风险因素监测系统（BRFSS），数据源及解决方案的相关代码全部开放，具体章节列出了大量学习和进阶资料，方便读者参考。

本书面向广大程序员和计算机专业的学生。

图灵程序设计丛书
统计思维：程序员数学之概率统计

◆ 著 [美] Allen B. Downey
译 张建锋 陈 钢
责任编辑 刘美英
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京天宇星印刷厂印刷

◆ 开本：880×1230 1/32
印张：5
字数：125千字 2013年6月第1版
印数：1－4 000册 2013年6月北京第1次印刷
著作权合同登记号 图字：01-2012-8793号
ISBN 978-7-115-31737-7

定价：29.00元

读者服务热线：(010)51095186转604 印装质量热线：(010)67129223
反盗版热线：(010)67171154

目录

第1章

程序员的统计思维

本书讨论如何将数据转换为知识。数据是廉价的（至少相对而言如此），但知识却异常宝贵。

我会介绍以下三门相互关联的学科。

- 概率论
  主要研究随机事件。人们对某些事件发生的可能性高低一般都有直观的认识，所以未经特殊训练就会使用“可能”、“不可能”之类的词汇。但本书会介绍如何量化这种可能性。

- 统计学
  统计学旨在根据数据样本推测总情况。大部分统计分析都基于概率，所以这两方面的内容通常兼而有之。

- 计算
  量化分析的最佳工具。计算机是处理统计量的常用工具。此外，计算实验还有助于理解概率论和统计学中的概念。

本书的主要目的就是要让懂编程的人通过编程来理解概率论和统计学。人们通常是从数学角度讲解概率论和统计学，而且很多人也因此学会了概率论和统计学。但在概率论和统计学中，有很多概念从

数学角度很难理解，但如果用计算方法就比较容易。

记得我妻子怀上我们第一个孩子时，我听到过这样一个问题：第一胎多在预产期后出生吗？本章接下来介绍的例子就源自这个问题。

1.1 第一个孩子出生晚吗

如果在 Google 上搜索这个问题，你会看到大量的相关讨论。有些人说确实如此，也有人说这没根据，还有人持完全相反的观点：第一个孩子会在预产期之前出生。

在这类讨论中，人们会用各种数据来证明自己的说法，常见的例子如下。

> “我有两个朋友最近都刚生了第一个孩子，两个宝宝的出生时间都比预产期晚了差不多两周。”
>
> “我的第一个孩子晚了两周才出生，我想我的第二个孩子会提前两周。”
>
> “我觉得这没道理，因为我姐姐是我妈妈的第一个孩子，她就提前出生了，我的几个表姐也一样。”

诸如此类的传闻称为经验之谈（anecdotal evidence），因为它们基于非公开发表的数据，而且通常是个人感受。在非正式场合，这类说辞没问题，所以这里并不是说上述观点不对。问题在于，我们需要更有说服力的证据和更可靠的结论。但这些经验之谈显然做不到这一点，原因如下。

- *观察的数量太少*
  第一胎婴儿的妊娠期比较长，但这种差异可能在自然波动范围内。这种情况下，我们需要比较大量孕妇的妊娠期数据才能判断这种差异是否真的存在。

- *选择偏差*
  第一胎婴儿出生比较晚的父母会更有兴趣加入这样的讨论。这种对数据进行选择的过程就会导致结果不准确。

- 确认偏差
  相信这种说法的人会提供支持示例，而怀疑这种说法的人则会引用反例。

- 不准确
  传闻通常都是个人的经历，在记忆、表述和复述等方面都会不准确。

那么，更好的做法是什么呢？

1.2 统计方法

为了解决上述经验之谈的种种不足，我们会运用以下统计学手段。

- 收集数据
  使用大型全国性调查的数据，这些数据是为得出美国人口方面可靠的统计推断而专门收集的。

- 描述性统计
  计算能总结数据的统计量，并评测各种数据可视化的方法。

- 探索性数据分析
  寻找模式、差异和其他能解答我们问题的特征。同时，我们会检查不一致性，并确认其局限性。

- 假设检验
  在发现明显的影响时（比如两个族群间的差异），我们需要评判这种影响是否真实，也就是说是否是因为随机因素造成的。

- 估计
  我们会用样本数据推断全部人口的特征。

通过这些步骤，绕过各种陷阱，我们就能得到更加合理也更可能正确的结论。

1.3 全国家庭成长调查

美国疾病控制与预防中心（CDC）从1973年开始推行全国家庭成长调查（NSFG），目的是收集（美国）“家庭的生活、婚姻状况、生育、避孕和男女健康信息。调查的结果用于……制定健康服务和健康教育计划，以及对家庭、生育和健康的统计研究”。[1]

我们会利用调查收集的数据来研究诸如“第一个小孩是否出生得较晚”之类的问题。为了有效使用这些数据，我们需要理解这个调查是怎么设计的。

NSFG是一个*横断面研究*（cross-sectional study），意思就是它的数据是一群人在某个时间点的情况。另一种常见方法是*纵贯研究*（longitudinal study），就是在一段时间内反复观察同一群人。

NSFG已经进行了7次，每次称为一个*周期*（cycle）。我们会使用来自Cycle 6的数据，这些数据是在2002年1月到2003年3月间收集的。

NSFG的目的是得到关于*人口情况*的一些结论，调查对象是15到44岁的美国人。

参与调查的人称为*被调查者*（respondent），一组被调查者就称为*队列*（cohort）。通常，横断面研究意味着具有*代表性*，即目标人群中的每一个人都有同等的几率参与调查。当然，实际很难实现这种理想状况，但执行调查的人会尽可能地做到这一点。

NSFG不具有代表性，而是有意进行了*过采样*（oversample）。设计者所调查的西班牙裔、非裔美国人和青少年的比例都高于他们在美国人口中的比例。过采样这些人群是为了确保其中的被调查者数量够大，从而得到有效的统计推断。

当然，过采样增大了根据调查结果推断全体人口结论的难度。稍候我们会继续讨论这一点。

注1：参见 http://cdc.gov/nchs/nsfg.htm。

习题1-1

尽管NSFG已经进行了7次，但它并不是纵贯研究。阅读维基百科页面 http://wikipedia.org/wiki/Cross-sectional_study 和 http://wikipedia.org/wiki/Longitudinal_study 可以弄清楚原因。

习题1-2

这个练习需要从NSFG下载数据，本书接下来会用到这些数据。

1. 打开 http://thinkstats.com/nsfg.html，阅读数据的使用协议，然后点击“I accept these terms”（假设你确实同意）。

2. 下载 2002FemResp.dat.gz 和 2002FemPreg.dat.gz 两个文件。前者是被调查者文件，每一行代表一个被调查者，总共7643个女性被调查者。后者是各个被调查者的怀孕情况。

3. 调查的在线资料地址：http://www.icpsr.umich.edu/nsfg6 。浏览左侧导航栏中调查的各部分，大致了解一下其中的内容。还可以在 http://cdc.gov/nchs/data/nsfg/nsfg_2002_questionnaires.htm 上阅读调查问卷的内容。

4. 本书的配套网站提供了处理NSFG数据文件的代码。从 http://thinkstats.com/survey.py 下载，然后在放置数据文件的目录中运行。程序会读取数据文件，然后会显示每个文件的行数：

```
Number of respondents 7643
Number of pregnancies 13593
```

5. 浏览一下代码，大致了解一下其功能。下一节会详细介绍。

1.4 表和记录

诗人、哲学家 Steve Martin 曾说：

> “Oeuf”就是egg，“chapeau”就是hat。好像所有的东西在法语中都跟在英语中的叫法不一样。

跟法语一样，数据库程序员的语言也跟我们的日常语言稍有不同。因为我们要谈到数据库，所以有必要学习一些专业术语。

被调查者文件中的每一行都表示一个被调查者。这行信息称为一条记录（record），组成记录的变量称为字段（field），若干记录的集合就组成了一个表（table）。

看一下 survey.py 中的代码，就会看到 Record 和 Table 这两个类的定义，前者是代表记录的对象，后者则是表示表的对象。

Record 有两个子类，分别是 Respondent 和 Pregnancy，两者分别是被调查者和怀孕的记录。目前这些类暂时还是空的，其中还没有用于初始化其属性的 init 方法。我们会用 Table.MakeRecord 方法将一行文本转换成一个 Record 对象。

Table 也有两个子类 Respondents 和 Pregnancies。这两个类的 init 方法设置了数据文件的默认名称和要创建的记录的类型。每个 Table 对象都有一个 records 属性，是一个 Record 对象的列表。

每个 Table 的 GetFields 方法返回一个指定记录字段的元组（tuple）列表，这些字段就是 Record 对象的属性。

例如，下面是 Pregnancies.GetFields：

```
def GetFields(self):
        return [
            ('caseid', 1, 12, int),
            ('prglength', 275, 276, int),
            ('outcome', 277, 277, int),
            ('birthord', 278, 279, int),
            ('finalwgt', 423, 440, float),
            ]
```

第一个元组的意思从第 1 列到第 12 列是 caseid 字段，且类型为整数。每个元组包含如下信息。

- field
  保存该字段的属性的名称。大部分情况下，我使用 NSFG 编码手册中的名称，全部用小写。

- start
  该字段的起始列编号。例如，caseid的起始编号是1。可以在NSFG编码手册中查询这些编号：http://www.icpsr.umich.edu/nsfg6/。

- end
  该字段的结束列编号。例如，caseid的结束列编号是12。跟Python中不一样，这里的结束列也是该字段的一部分。

- *转换函数*
  将字符串转换成其他类型的函数。可以用内置的函数，比如int和float，也可以使用用户自定义的函数。如果转换失败，属性的值就会是字符串'NA'。如果某个字段不需要转换，可以使用identity函数或是str函数。

从pregnancy记录中可以得到以下变量。

- caseid
  被调查者的整数ID。

- prglength
  怀孕周期，单位是周。

- outcome
  怀孕结果的整数代码。代码1表示活婴。

- birthord
  正常出生的婴儿的顺序。例如，第一胎婴儿的编号是1。如果没有正常出生，该字段为空。

- finalwgt
  被调查者的统计权重。这是一个浮点值，表示这名被调查者所代表的人群在美国总人口中的比例。过采样人群的权重偏低。

如果你仔细阅读编码手册，就会发现这些变量大部分都经过了*重编码*(recode)，也就是说这并不是调查所采集的*原始数据*，而是根据原始数据计算出来的。

例如，第一胎活婴的 prglength 在原始数据中有变量 wksgest（妊娠周数）时就等于该变量的值，否则就会用 mosgest * 4.33（妊娠月数乘以每个月的平均周数）估计出来。

重编码通常遵循数据一致性和准确性原则。除非有特别原因一定要使用原始数据，否则就应该直接使用重编码后的数据。

你可能还发现了 Pregnancies 有 Recode 方法，用来做一些其他的检查和重编码工作。

习题1-3

在这个练习中，我们会编写一个程序来看看 Pregnancies 表中的数据。

1. 在 survey.py 和数据文件的目录中创建一个 first.py 文件，然后将下面的代码输入或复制到文件中：

```
import survey
table = survey.Pregnancies()
table.ReadRecords()
print 'Number of pregnancies', len(table.records)
```

   结果应该是 13 593 条怀孕记录。

2. 编写一个循环遍历表（table），计算其中活婴的数量。查阅临床结果（outcome）的文档，确认你的结果跟文档中的总结一致。

3. 修改这个循环，将活婴的记录分成两组：一组是第一胎出生；另一组是其他情况。再看一些出生顺序（birthord）的文档，看看你的结果跟文档中的结果是否一致。

   在处理新的数据集时，这种检查对于发现数据中的错误和不一致性、检查程序中的错误以及检验对字段编码方式的理解是否正确等都是很有用的。

4. 分别计算第一胎婴儿和其他婴儿的平均怀孕周期（单位是周）。两组之间有差异吗？差异有多大？

从 http://thinkstats.com/first.py 可下载这个练习的答案。

1.5 显著性

在前面的练习中，我们比较了第一胎婴儿和其他婴儿的妊娠期。如果一切顺利，读者会发现第一胎婴儿的出生时间比其他婴儿的出生时间平均晚 13 个小时。

类似这样的差异称为*直观效应*（apparent effect），意思就是似乎发生了有意思的事情，但还不确定。我们还需要考虑以下问题。

- 如果两组的均值不一样，其他汇总统计量如何，比如中位数和方差？我们能更精确地描述它们之间的差异吗？
- 有没有可能这两组实际上是一样的，而我们所观察到的这种差异只是随机产生的？如果是，那这个结论就不是*统计显著*的。
- 这种直观效应有没有可能是因为选择偏差或是实验设置中的错误导致的？如果是，那么这种直观效应就是*人为的*，也就是我们意外创造的，而并非发现了事实。

本书接下来的大部分内容都是为了回答这些问题。

习题1-4

学习统计学的最好方法就是从一个自己感兴趣的项目开始。有没有“第一胎婴儿出生较晚”这类吸引你的问题来研究？

思考自己感兴趣的问题，例如传统观念、有争议的话题或是有社会影响的问题，看看你能否将这些问题转换成统计学问题。

寻找能解决该问题的数据。国外政府是很好的数据来源，因为公共研究的数据通常都是免费的[2]。另一个查找数据的好去处是 Wolfram Alpha，其中收集了很多经过验证的高质量的数据集，网址是 http://wolframalpha.com。Wolfram Alpha 的搜索结果是有版权限制的，在使用之前应该阅读一下协议。

注 2：在撰写这段内容的时候，英国某法院规定“信息自由法案”（Freedom of Information Act）也适用于科学研究数据。

Google 和其他的一些搜索引擎也能帮你寻找数据，但网络上各种资源的质量高低不一，判断起来不容易。

如果发现已经有人回答了你的问题，要仔细看看回答是否合理。数据和分析中的缺陷可能会导致结论不可靠。如果是这样，你应该采用不同的方法来分析数据，或者是寻找其他更好的数据来源。

如果已发表的论文回答了你的问题，那就应该能弄到原始数据，很多作者都会在网上提供。但如果数据涉及个人隐私，最好联系一下作者，告诉他你要如何使用数据，或是接受特定的使用协议。坚持到底！

1.6 术语

- *经验之谈*（anecdotal evidence）
  个人随意收集的证据，而不是通过精心设计并经过研究得到的。
- *直观效应*（apparent effect）
  表示发生了某种有意思的事情的度量或汇总统计量。
- *人为*（artifact）
  由于偏差、测量错误或其他错误导致的直观效应。
- *队列*（cohort）
  一组被调查者。
- *横断面研究*（cross-sectional study）
  收集群体在特定时间点的数据的研究。
- *字段*（field）
  数据库中组成记录的变量名称。
- *纵贯研究*（longitudinal study）
  跟踪群体，随着时间推移对同一组人反复采集数据的研究。
- *过采样*（oversampling）
  为了避免样本量过少，而增加某个子群体代表的数量。