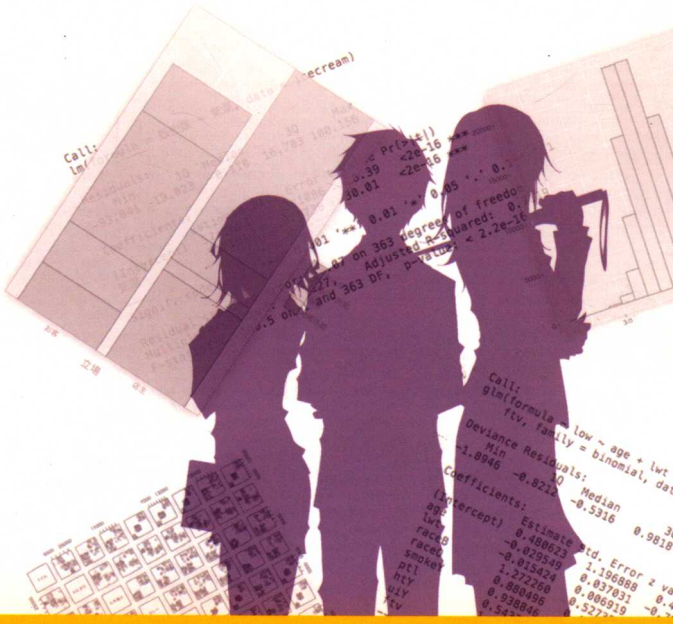
TURI
图灵程
设计丛

菜鸟侦探挑战
数据分析

[日] 石田基广◎著　支鹏浩◎译



数学菜鸟也能搞懂数据分析

从零学起→没有深奥的理论和晦涩的知识
实际体验→免费软件RStudio＋提供模拟数据

均值/直方图/*t*检验/卡方检验/回归分析
逻辑回归分析/文本挖掘……

应用R语言
轻松学统计

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

TURING
图灵程序设计丛书

菜鸟侦探挑战数据分析

[日] 石田基广◎著　支鹏浩◎译

人民邮电出版社
北京

图书在版编目（CIP）数据

菜鸟侦探挑战数据分析 /（日）石田基广著；支鹏浩译. -- 北京：人民邮电出版社，2017.1
（图灵程序设计丛书）
ISBN 978-7-115-44166-9

Ⅰ. ①菜… Ⅱ. ①石… ②支… Ⅲ. ①数据处理
Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2015）第288596号

内 容 提 要

本书以小说的形式展开，讲述了主人公俵太从大学文科专业毕业后进入征信所，从零开始学习数据分析的故事。书中以主人公就职的征信所所在的商业街为舞台，选取贴近生活的案例，将平均值、*t*检验、卡方检验、相关、回归分析、文本挖掘以及时间序列分析等数据分析的基础知识融入到了生动有趣的侦探故事中，讲解由浅入深、寓教于乐，没有深奥的理论和晦涩的术语，同时提供了大量实际数据，使用免费自由软件RStudio引领读者进一步体验数据分析，实践性非常强。本书适合所有对数据分析感兴趣但又苦于无从下手的读者阅读。

◆ 著　　[日]石田基广
译　　支鹏浩
责任编辑　傅志红
执行编辑　高宇涵　侯秀娟
责任印制　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷

◆ 开本：880×1230　1/32
印张：8.25
字数：230千字　　2017年1月第1版
印数：1-3 500册　　2017年1月北京第1次印刷

著作权合同登记号　图字：01-2016-3272号

定价：42.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第8052号

版权声明

- 本书涉及的R脚本已确认能在R（3.2.1，Windows版、Mac版）、RStudio（0.99.467，Windows版、Mac版）上正常运行，此处所用各版本均为编写脚本时的最新版本。
- 对R和RStudio的安装及基本操作不熟悉的读者请先阅读“番外篇”。
- 下述URL为介绍本书内容的网站。本书各章涉及的数据以及R脚本都可以在这个网站下载（点击“随书下载”）。另外，各位可以在网站的评论区发表对本书的感想和意见。

http://www.ituring.com.cn/book/1809

走近“数据分析”
——对于初学者而言本书具有的三大优势

相信对许多读者来说，“大数据”“数据科学”这类词并不陌生。就连电视上也曾介绍过，在改善医疗、振兴旅游业等方面，大数据被寄予了厚望。大数据已经走近了我们，而我们却难免会觉得“数据的运用确实让很多事情都方便了不少，但这么高端的东西肯定跟自己不沾边”。

实际上，数据相关的知识和技术正在融入我们的日常生活。回想20世纪90年代的数据分析，那是需要一群掌握着特殊技术的专家用造价高昂的大型计算机才能干的活。而现在，我们只需一台廉价的笔记本电脑，甚至是一部智能手机，就能完成当年专家们费时费力折腾许久才能完成的工作。

在过去，要想进行高端的数据分析，就必须先学习晦涩难懂的数学知识。再加上分析过程需要计算机来辅助计算，所以人们还必须学习编程语言并自己敲代码。如今，这类分析就连路边报刊亭大婶都能手到擒来，完全不用学那些难懂的数学了。当然，学数学还是有其好处与作用的。（笔者的一位朋友得知用PC就能进行数据分析后，一头扎进了数据分析的世界，后来还把数据分析涉及的数学和工学都重新学了一遍。其实这样的人也并不少见。）

本书是一本数据分析的入门书，面向从零学起、想掌握一定数据分析能力的读者。如果你听说过“数据分析”，有兴趣对其加深了解却又苦于无从下手，或是大学统计学上亮了红灯，想重新挑战一番，那么这本书应该会适合你。本书具有以下三大优势。

- 由浅入深地介绍数据分析的相关问题，其中包含一般入门书并不会涉及的一些深度问题

- 知识融合在故事之中，悬念迭出的剧情让人忍不住想继续读下去
- 书中内容（实际的数据分析）可在免费自由软件上执行

下面我们具体介绍这三大优势。

■关于数据分析

本书从各位耳熟能详的“平均值”开始讲起。与往常不同的是，我们将通过数据模拟来加深各位对平均值（期望值）的理解。同时，各位可以在自己的电脑上实际体验书中提及的数据模拟。

随后我们将通过具体事例为各位讲解两种分析手法，即“分析平均值是否存在差异的手法（t 检验）”和“从问卷结果中分析意见差异的手法（卡方检验）”。除此之外，书中内容还会涉及用于数据预测的回归分析、逻辑回归分析等高级分析手法。

不仅如此，本书还会介绍文本挖掘的相关知识。文本挖掘是一种对文章进行解析并从中找出有用信息的分析手法。如今市面上介绍文本挖掘的书籍寥寥无几。

举个例子，在 SNS 日益火爆的现在，许多人会把自己的意见或感想直接发表在 Facebook、Twitter 或各种博客上。比如自己买了一台空气净化器，有些人就会把使用后的感想发上来。我们只要将国内关于空气净化器的博文收集起来并加以分析，就能知道为什么某些产品广受支持，而某些产品则恶评不断了。

但说起来容易做起来难，这类评论的文本总量往往十分惊人，一条条去收集、去读根本不现实。不过，如今我们只需用市面上的一些工具，就能轻松地在网上自动收集文章并交给计算机去解析，从中找出有建设性的意见。实际上，许多生产商一直都是这样做的。他们在网上收集对自家商品的评价，调查消费者的使用感受以及偏好，从而调整新一代商品的设计定位。本书中文本挖掘的事例相对简单，但为了让各位感受到其威力，特地加入了大量说明。

■知识融入故事的形式

本书的主人公田中俵太是文科的大学应届毕业生。他自今年春天入

职征信所，职务是“侦探”。他将在天羽幸小姐的指导下从零开始学习数据分析。天羽小姐是数据分析专家，为人略显霸道，加之其人生字典里从来没有“委婉”二字，让我们的主人公俵太三天两头就要受点打击。每到这种时候，天羽小姐的助手川崎逸子小姐就会站出来，将数据分析的入门知识手把手地传授给俵太。

故事以一个小商业街为舞台，所以不会发生与“大数据”相关的案子。相反，我们将要解决的小案子都是各位身边随时可能发生的问题。随着这些案子一个个被解决，俵太的知识与技能将逐渐丰富起来。

■用免费自由软件 RStudio 进一步体验数据分析

如果只想对数据分析有个大致了解，那么通读一遍本书就足够了。不过，自己实际动手做出结果才是数据分析的最大乐趣所在。因此，我们为各章中发生的“案子”准备了实际数据（其中一部分经过了简化）。各位可以把这些数据下载下来，在自己的电脑上验证案子的来龙去脉及其解决方案。

进行数据分析需要专用软件。本书主要使用的是免费自由软件 RStudio，其使用方法的介绍已被穿插在故事的各个场景中。有兴趣的读者不妨参考本书的“番外篇”安装 RStudio。

＊＊＊

希望各位读者能通过本书感受到数据分析带来的乐趣，进而力求在数据分析方面更上一层楼。最后，对于在草稿阶段为本书整体内容提出诸多建议的石田和枝老师，从初学者角度对 RStudio 操作说明部分指出不足之处的片冈伸一老师，为奋战在枯燥的数据分析第一线的主人公们赋予了亲切外貌的插画师 shimano 老师，以及在交稿阶段不厌其烦地为笔者进行审校的 SB Creative 公司的则松直树老师，笔者在此致以衷心的感谢。

石田基广

2015 年 9 月

目录

序 故事就这样开始了

00-01 遭贼的概率

在我的印象中，侦探的工作就是凭借经验和直觉推理破案，比如查明大家族内杀人案的真凶，或者找出富商埋藏的传家宝之类的。老实说，推理类小说对我的吸引力并不大，不过侦探这职业倒是让我觉得挺帅气。正是出于这种原因，我这个到了大四下半学期仍没拿到一份Offer的学生，才会在偶然瞥见招聘网站上的征信所招聘信息时眼前一亮，还拿出了十足的诚意投了简历。结果征信所很快就回了信息，而且说是直接面试，没有笔试环节。他们真的会聘用我这么一个毫无经验的22岁毛头小子？对此半信半疑的我，怀着死马当活马医的心情去参加了面试。

面试官是一位尚算年轻的女士，脸上自始至终都没有过一丝笑意。我这边是紧张得前言不搭后语，她那边则是字字冰冷、句句带刺。总而言之，就是“唉，八成又没戏”。但出乎意料的是，几天后，沉溺在失落中的我居然接到了征信所发来的Offer。

现在回想起来有些后悔，因为当时就该觉得不对劲了。但那时的我早就被找工作的事烦得焦头烂额，根本没心思去怀疑这来之不易的Offer和用人单位。再加上大四下半学期跟毕业论文的一番苦战，等我回过神来时已经迎来了毕业典礼。至于征信所那边，自从春假时发过来一份入职指南之后就再没联系过我。

辞旧迎新的钟声早已远去，转眼已是4月1日了。我按照入职指南的要求乘车来到东京都某站，如今正向征信所所在的商业办公楼走去。

进门乘电梯上到5层，然后凭当初面试时的记忆右转，径直来到走廊尽头的征信所门前，只见门牌上写着“商业研究AMO'S事务所”。于是我站定脚步，略作深呼吸，按响了门铃。

“……”

没人应门。再按一次好了。

“……”

还是没人应门。我拿出手机看了看时间，刚过8点，或许是来得太早了些。但话又说回来了，入职指南上根本就没写几点来啊。想到这里，我又从包里翻出了录用通知书和入职指南，上面的公司名是“商业研究AMO'S事务所”，这个显然没搞错。然后把入职指南也重新确认了一遍，正文里只提到“4月1日携个人印章至公司报到”。于是只能继续按门铃，但依旧没人应门。

“???”

无奈之下，我索性直接拧了拧门把手，结果当然是打不开。门把手上有两个锁孔。我把耳朵贴到门上，然而里面一丝动静都没有。于是我又把全身都贴到门上想尽力听听声音，这时候，一个尖锐的嗓音突然从

背后传来。

“你在那干什么呢?”

“哇啊!”

我吓得赶忙回过头，发现身后站着一位年轻的女士，身材高挑，穿着一套黑色西装。当然，年轻归年轻，真论起年龄，恐怕她要比我大不少。正当我觉得这位女士有几分面熟，仔细端详思索之际，她突然把手中的提包扔在了地上。再望去时，她右手早已握着一根黑色的棒状物，我登时就傻了眼。这时，只见她左手捏住黑棒顶端向外一扯，从里面拽出了一根闪着银色光泽的金属棒。

“别别别，别冲动。我不是贼!”

“这我知道。因为实际遭贼的概率根本就不高。一年内遭贼的概率充其量也就 0.1%。”

“诶??”

“再说了，更不会有贼敢闯征信所的空门。”

“这……倒也是这么回事。”

“你小子是跟踪狂吧？你那吊儿郎当的口气我有印象。”

到这里我才回想起来。这声音，还有这眼镜，她不就是当初面试我的那个人嘛。当时她也是透过这副眼镜，冷冷地瞪着有问必结巴的我。不过现在，这位面试官正斜架着警棍慢慢向我靠近。

“我……我叫田中俵太，是今天来贵公司报到的新员工!!”

我话音刚落，她就立刻停住了。

“田中俵太？对了，今天好像是有个新人要报到来着。你是今天开始上班?”

“这里写着的。”

说着我把征信所发来的录用通知书和入职指南拿给她看。她并没有接过去，只是随便瞟了一眼，便轻轻吐了口气并捡起了地上的提包。呼……我安全了！她貌似肯收起警棍了。

“话说，你这新员工为什么要像个壁虎似的趴在门上?”

“呃，我是见按了门铃没反应，想听听里面是不是真的没人。况且门还锁着。”

“钥匙就在门垫下面。”

她指着我脚下的红色门垫说道。

00-02 两把钥匙都选对的概率

“啊？征信所居然这么随便？”

“我都说了，一般的贼根本不会把征信所当成目标。”

我弯下腰翻开门垫，眼前出现了 6 把外形和颜色完全相同的钥匙。

“这……这是？”

“这里面有一个是门把手的钥匙，还有一个是辅助锁的钥匙。不过，必须两把钥匙都选对而且同时拧，门才能打开。”

“哈哈，原来如此，拿假钥匙做障眼法啊。这个还真有想法。可是，6 把钥匙里面 2 真 4 假，应该很快就能试出来吧？”

“那你自己试试看呗。”

于是我随便捡起 2 把插进锁孔，发现拧不开。把钥匙对调了一下又插进去，还是拧不开。看来这两把里面至少有一把是假的。稍微思考之后，我决定放下其中一把钥匙，随便捡起另一把钥匙插入锁孔，两边同时拧了下，然而锁依旧不给面子。

随后又是一阵换钥匙、对调、换钥匙、对调，结果所有组合统统败下阵来。像这样两手同时去拧两把钥匙，用的是平时很少用到的肌肉，所以还是蛮辛苦的。

“……看来没我想象得那么好开啊。”

“两把钥匙都选对的概率是 1/30。”

“诶？是吗？”

“6 个中先选出一个，这有 6 种情况。然后从剩下的 5 个中再选出一个。所以选择 2 个钥匙总共有 $6 \times 5 = 30$ 种情况。然后这 30 种之中只有 1 种能打开门。”