数据分析技术丛书
HZ BOOKS 华章IT
[PACKT] PUBLISHING

R Statistical Application Development by Example Beginner's Guide

R统计应用开发实战

[印度] Prabhanjan Narayanachar Tattar 著
程豪 译

系统讲解R应用开发的统计学基础，并针对不同问题给出具体的R实现代码

机械工业出版社
China Machine Press

R Statistical Application Development by Example Beginner's Guide

R统计应用开发实战

[印度] Prabhanjan Narayanachar Tattar 著

程豪 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 统计应用开发实战 /（印）塔特尔（Tattart, P. N.）著；程豪译 . —北京：机械工业出版社，2015.3
（数据分析技术丛书）
书名原文：R Statistical Application Development by Example Beginner's Guide

ISBN 978-7-111-49347-1

I. R… II. ① 塔… ② 程… III. 统计分析－统计程序 IV. C819

中国版本图书馆 CIP 数据核字（2015）第 029659 号

本书版权登记号：图字：01-2013-7569

R 统计应用开发实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：陈佳媛　　责任校对：董纪丽
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2015 年 4 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：15.75
书　　号：ISBN 978-7-111-49347-1　　定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

The Translator's Words 译 者 序

随着 R 在各个数据分析领域的广泛应用，学习运用 R 语言处理和分析问题越来越受到人们的关注。然而，国内关于 R 语言应用方面的著作并未详尽地介绍其在统计学领域的使用和开发。而且对于非统计学专业，或并未接受过编程训练的朋友，在使用 R 语句进行数据分析时，很容易出现不知如何处理，以及得出结果后不知如何进行解释等问题。本书从应用的角度对有分析需求或有学习兴趣的朋友给予一定的指导，为读者详细展示了数据获取—数据分析—统计建模—解释说明整个过程。

在翻译本书时，我竭尽全力注重每个细节，希望可以尽己之所能，还原作者的研究成果，并清晰简洁地呈现给读者。但是译文也可能会存在一些问题，还请大家见谅并予以指正。

在此还要感谢我的导师易丹辉教授对我孜孜不倦的教诲，让我用踏实认真的态度完成整本书的翻译和审校。感谢我的家人，谢谢他们一路走来对我的陪伴、包容和理解。最后，我要特别感谢我的至爱刘钰洁，大学同学李倩、刘双和翟树芬，研究生同学蔡丽、鲁韶菲和宋丹，感谢他们利用自己宝贵的时间和精力参与本书翻译和审校工作，并对他们所付出的辛苦和汗水表示诚挚的敬意。

希望本书可以对广大有需求的读者有所帮助。

前　言 Preface

开放源代码的R软件正在迅速成为统计领域优先考虑的应用软件，而且在它已适用的丰富的科学网络中，R软件也被用到机器学习、数据挖掘等领域。对整个社会来讲，数学理论和统计应用相结合确实具有标志性，R软件在其中起着举足轻重的作用。本书在用R语言建立统计模型方面算是一次简洁易懂的尝试，即使是那些对统计和R不太熟悉的读者，也可参考阅读。在不同背景的同事和朋友使用这个软件的过程中，我发现许多人对学习这方面的知识都很感兴趣，而且会将其应用在他们的研究领域，帮助他们在涉及不确定性的分析中做出适当的决策。如果是十年前，我的朋友可能会满足于有人推荐一本有用的参考书。仅此而已！几乎在所有的领域，该项工作均已通过计算机完成，而且他们可以以电子表、数据库形式获得数据，有时也可以获得普通的文本格式数据。对于一个合适的统计学模型，会有这样一个不变的问题："使用什么软件？"我可以仅仅用一个字母来回答："R！"为什么要使用R？因为它确实是一个简单的决定，并且在过去的7年间成为我学术研究时的伙伴。在本书里，这些经历已经转化为各章的具体内容，以及较清晰的R语言建模分析。

与我的那些有志于统计建模的同事和朋友交流的一个意外收获是，我了解到他们对这一主题的学习曲线。克服困难的第一步是介绍绝大部分初学者都会熟悉的基本概念，比如数据。只是在细微处有些不同，而且我坚信介绍自身领域的内容会激励读者在他们的路上走得更远。针对大多数统计软件，R会提供模块和使用包，而且使用包几乎覆盖了多数最近发明的统计学方法。本书的前5章以基础知识和R软件为主体，因此包括了R基础知识、数据可视化、探索性数据分析和统计推断。

基础知识部分会使用有趣的例子来加以说明，并且为后5章建立框架。这部分首先介绍了回归模型、线性和logistic回归，这些是应用部分最重要的研究热点。这种讨论实质上更普遍，而且这种方法也很容易应用于不同的领域。最后两章受到了布雷曼（Breiman）学校的

启发，因此详细介绍了分类和回归树的现代方法，并且用实际的数据集进行了说明。

本书的主要内容

第 1 章通过问卷和数据集介绍不同类型的数据。在一些有趣的背景下，该章详细说明了统计模型的需求。然后简要介绍了 R 软件安装和相关使用包。通过介绍 R 程序讨论了离散型和连续型随机变量。

第 2 章首先简单介绍了 R 语言基础知识。该章通过清晰、简洁的例子讨论了数据帧、向量、矩阵和列表。然后详细介绍了如何从外面导入 csv、xls 和其他格式的数据。该章还涉及了为其他软件从 R 写入数据 / 对象，并且在最后介绍了包含在 R 会话管理上的一个对话。

第 3 章分别针对分类型和数值型数据集讨论了有效的图表展示方法。对于分类型数据，可以用来作条形图、散点图、样条和镶嵌图，以及四折图；而对于连续型 / 数值型数据，可以作直方图、箱线图和散点图。同时该章也简单介绍了 ggplot2 的使用方法。

第 4 章涵盖了初级数据分析所用到的很直观的技术和方法。作为初级分析的步骤，探索性数据分析（EDA）的可视化技术（如茎叶图，字母值，以及对耐抗线、平滑数据和中位数平滑的建模方法）会给出很独到的见解。

第 5 章首先强调了似然函数和极大似然估计。通过用一些针对具体问题而定义的函数来研究参数的置信区间。本章还介绍了一些重要的统计检验，包括比较均值的 Z- 检验和 t- 检验、卡方检验和比较方差的 F- 检验。

第 6 章的线性回归分析建立了结果变量和解释变量集之间的线性关系。这个线性回归模型使用了很多潜在的假设，而且这些假设都可以用一些验证方法进行证明。一个模型可能受到一个简单的观测、一个简单的因变量取值或一个解释变量的影响。该章深入讨论了统计度量，帮助去除一个或更多的异常情况。给定很多协变量，利用模型选择方法可以发现一个有效的模型。

第 7 章讨论的是，当因变量是一个二分类变量时，logistic 回归模型会被用作一个分类模型。通过模型的残差，对模型进行统计诊断和有效性验证，可以实现优化。受试者工作特征曲线也会被用来识别一个更好的分类模型。

第 8 章就前两章提出的模型的过度拟合问题进行讨论。岭回归明显减少了一个模型过度拟合的问题，而且样条模型也为下一章讨论的模型奠定了基础。

第 9 章提出了一个基于树的回归模型。这些树最初是使用 R 函数进行建模的，最终的树也通过基本的代码重新进行调整，而这些代码有助于理解分类回归机制。

第 10 章用 bagging 算法和随机森林比较分类回归的两处改进。该章通过一个数据集帮助读者巩固从第 6 章 ~ 第 10 章介绍的所有模型。

第 1 章 ~ 第 5 章是 R 软件和统计的基本知识，而第 6 章 ~ 第 10 章详细讨论了应用和现代回归模型。

在本书的最后给出了参考文献。

阅读本书前的准备工作

R 是学习本书时唯一需要的软件，下载网址为：http://www.cran.r-project.org/。在一个 R 语言工作环境中完成任务时，将会用到 R 软件包。在 R 软件包 RSADBE 中可以获得本书中的数据集，RSADBE 是本书英文书名的缩写，详见网址：http://www.cran.r-project.org/web/packages/RSADBE/index.html。

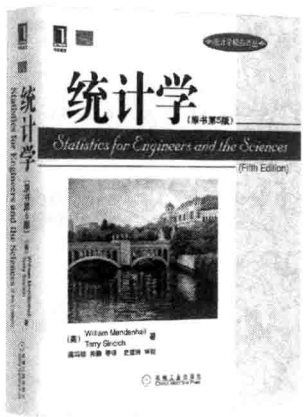
本书的读者对象

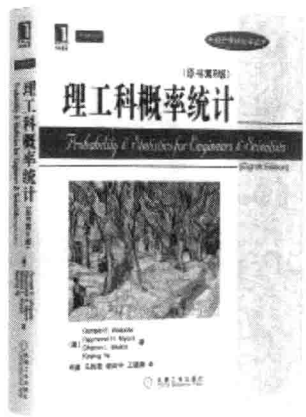
本书适合有一定统计学基础而且在他们的领域需要统计学应用的读者阅读。前 7 章对于统计学的任何硕士研究生均是有用的，如果想进一步学习也可以很容易完成本书其他章节的阅读并获得分类回归的应用知识。

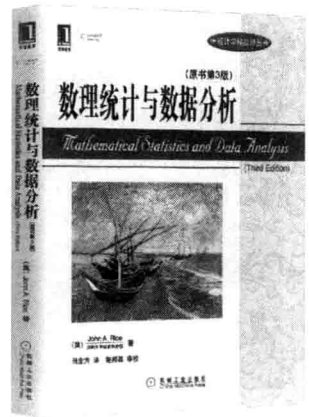
勘误与相关资源

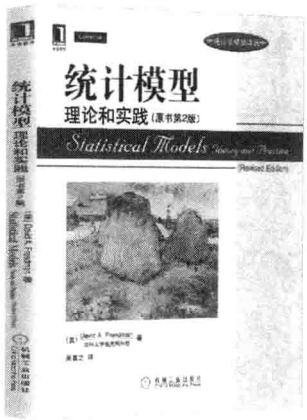
如果你发现任何错误，请访问下面的网址进行反馈 http://www.packtpub.com/submit-errata。选择相应图书，点击勘误表提交表单的链接，并且录入勘误的细节。一旦你的反馈被证实，你的提交表将会被接收而且会上传到我们的网站，或者被添加到已经存在的勘误表中。本书的相关资源可登录华章网站（www.hzbook.com）本书页面下载。

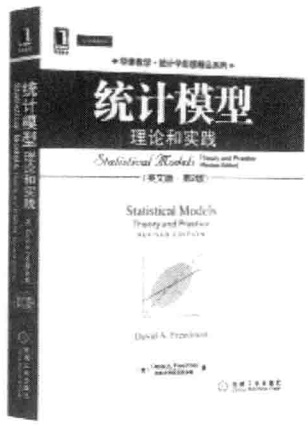
推荐阅读











统计学（原书第5版）

作者：William Mendenhall 等 ISBN：978-7-111-26437-8 定价：128.00元

理工科概率统计（原书第8版）

作者：Ronald E Walpole 等 ISBN：978-7-111-27708-8 定价：98.00元

数据统计与数据分析（原书第3版）

作者：John A RIce ISBN：978-7-111-33646-4 定价：85.00元

统计模型：理论和实践（原书第2版）

作者：David A Freedman ISBN：978-7-111-30989-5 定价：45.00元

统计模型：理论和实践（英文版·第2版）

作者：David A Freedman ISBN：978-7-111-341797-5 定价：38.00元

多元数据分析（英文版·第7版）

作者：Joseph F Hair Jr 等 ISBN：978-7-111-34198-7 定价：109.00元

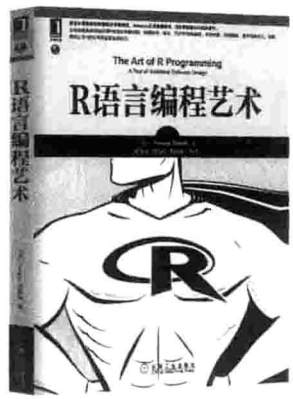
推荐阅读

数据挖掘与R语言

作者：Luis Torgo ISBN：978-7-111-40700-3 定价：49.00元

R语言经典实例

作者：Paul Teetor ISBN：978-7-111-42021-7 定价：79.00元



R语言编程艺术

作者：Norman Matloff ISBN：978-7-111-42314-0 定价：69.00元

R语言与数据挖掘最佳实践和经典案例

作者：Yanchang Zhao ISBN：978-7-111-47541-5 定价：49.00元

R语言与网站分析

作者：李明 ISBN：978-7-111-45971-2 定价：79.00元

R的极客理想——工具篇

作者：张丹 ISBN：978-7-111-47507-1 定价：59.00元

Contents 目　录

译者序

前言

第 1 章　数据特征 …… 1

1.1　问卷调查及其组成部分 …… 1

1.2　在计算机科学中的不确定性研究 …… 5

1.3　R 安装 …… 6

1.3.1　使用 R 包 …… 7

1.3.2　RSADBE——本书的 R 包 …… 8

1.3.3　离散分布 …… 9

1.3.4　离散均匀分布 …… 10

1.3.5　二项分布 …… 11

1.3.6　超几何分布 …… 13

1.3.7　负二项分布 …… 14

1.3.8　泊松分布 …… 15

1.4　连续分布 …… 16

1.4.1　均匀分布 …… 16

1.4.2　指数分布 …… 17

1.4.3　正态分布 …… 18

1.5　本章小结 …… 20

第 2 章　数据导入和导出 …… 21

2.1　data.frame 和其他格式数据 …… 21

2.1.1　常数、向量和矩阵 …… 21

2.1.2　列表对象 …… 28

2.1.3　data.frame 对象 …… 30

2.1.4　表对象 …… 33

2.2　函数 read.csv、read.xls 以及外来程序包 …… 35

2.3　导出数据 / 图表 …… 41

2.3.1　导出 R 对象 …… 41

2.3.2　导出图表 …… 41

2.4　管理一个 R 会话 …… 43

2.5　本章小结 …… 45

第 3 章　数据可视化 …… 46

3.1　分类数据的可视化技术 …… 47

3.1.1　条形图 …… 47

3.1.2　点图 …… 52

3.1.3　脊柱图、马赛克图 …… 54

3.1.4　饼图和四折图 …… 58

3.2　连续型变量数据的可视化 …… 59

3.2.1 箱线图 …… 60
3.2.2 直方图 …… 62
3.2.3 散点图 …… 66
3.2.4 帕累托图 …… 70
3.3 ggplot 概述 …… 71
3.4 本章小结 …… 73

第 4 章 探索性分析 …… 75

4.1 基本汇总统计量 …… 75
4.1.1 百分位数、四分位数和中位数 …… 76
4.1.2 折页数 …… 76
4.1.3 四分位极差 …… 77
4.2 茎叶图 …… 80
4.3 字母值 …… 83
4.4 数据变换 …… 84
4.5 袋状图：二元箱线图 …… 86
4.6 耐抗线 …… 88
4.7 平滑数据 …… 90
4.8 中位数平滑 …… 93
4.9 本章小结 …… 95

第 5 章 统计推断 …… 97

5.1 极大似然估计 …… 98
5.1.1 可视化似然函数 …… 98
5.1.2 寻找极大似然估计 …… 101
5.1.3 使用 fitdistr 函数 …… 103
5.2 置信区间 …… 105
5.3 假设检验 …… 108
5.3.1 二项式检验 …… 109
5.3.2 比例检验和卡方检验 …… 111
5.3.3 基于正态分布检验：单样本 …… 113
5.3.4 基于正态分布检验：两样本 …… 118
5.4 本章小结 …… 121

第 6 章 线性回归分析 …… 122

6.1 简单线性回归模型 …… 123
6.1.1 随意选择参数会发生什么 …… 123
6.1.2 建立一个简单线性回归模型 …… 126
6.1.3 ANOVA 及置信区间 …… 128
6.1.4 模型验证 …… 129
6.2 多元线性回归模型 …… 133
6.2.1 平均 K 个简单线性回归模型或建立一个多元回归模型 …… 134
6.2.2 建立一个多元线性回归模型 …… 136
6.2.3 多元线性回归模型的 ANOVA 和置信区间 …… 137
6.2.4 有用的残差图 …… 139
6.3 回归诊断 …… 141
6.3.1 杠杆点 …… 142
6.3.2 影响点 …… 142
6.3.3 DFFITS 和 DFBETAS …… 143
6.4 多重共线性问题 …… 143
6.5 选择模型 …… 145
6.5.1 逐步选择 …… 145
6.5.2 基于准则的方法 …… 146
6.6 本章小结 …… 150

第 7 章　logistic 回归模型 …… 151

7.1　二元回归问题 …… 151
7.2　probit 回归模型 …… 153
7.3　logistic 回归模型 …… 155
7.4　模型验证和诊断 …… 160
7.4.1　广义线性模型的残差图 …… 160
7.4.2　广义线性模型的影响点和控制点 …… 163
7.5　接收操作曲线 …… 166
7.6　德国的信用甄别数据集的 logistic 回归 …… 168
7.7　本章小结 …… 171

第 8 章　正规化回归模型 …… 172

8.1　过度拟合问题 …… 172
8.2　回归样条 …… 176
8.2.1　基函数 …… 176
8.2.2　分段线性回归模型 …… 176
8.2.3　自然三次样条函数和一般的 B 样条曲线 …… 179
8.3　线性模型的岭回归 …… 183
8.4　logistic 回归模型的岭回归 …… 187
8.5　再看模型评估 …… 188
8.6　本章小结 …… 193

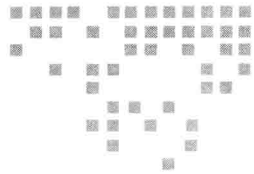
第 9 章　分类与回归树 …… 194

9.1　递归划分法 …… 194
9.1.1　划分数据 …… 196
9.1.2　第一个树 …… 197
9.2　构造回归树 …… 200
9.3　构造分类树 …… 209
9.4　德国信用数据集的分类树 …… 215
9.5　树的修剪和完善 …… 218
9.6　本章小结 …… 220

第 10 章　分类与回归树及其他 …… 222

10.1　分类与回归树的改进 …… 222
10.2　Bagging …… 225
10.2.1　bootstrap 算法 …… 225
10.2.2　bagging 算法 …… 227
10.3　随机森林 …… 230
10.4　整合 …… 233
10.5　本章小结 …… 238

参考文献 …… 239



第1章 Chapter 1

数据特征

数据是由不同类型变量的观察值组成的，任何数据分析师在接触统计分析的最初阶段，都必须理解这些错综复杂的事物。这一章阐述了数据的重要性，首先介绍一个虚拟的问卷调查模板，然后深入探讨该专题，而后解释了问卷调查在计算机科学领域具有的不确定性，最后介绍了离散型和连续型随机变量。

本章主要内容：

- 主要变量类型的辨别，例如定类变量（或名义变量）、分类变量和连续变量
- 许多实际的实验中所产生的不确定性
- R软件安装和软件包
- 离散型和连续型随机变量的数学形式及其应用

1.1 问卷调查及其组成部分

本节的目的是尽可能多地介绍一些变量类型。通常，入门课程是首先介绍概率论知识，然后介绍生成随机变量的必要条件。但是本书并没有按这种方式组织，而是直接从数据开始。之所以选择这种方式，是因为这种方式建立在读者已经熟悉的内容上，然后将其与R语言基础模块结合起来。

用户可能对问卷调查很熟悉。在一个婴儿出生后可能会做一个问卷调查，帮助医院了解母亲的经历、婴儿的健康状况以及新生儿监护人所关注的问题。连锁商店在卖出商品后可能会马上让消费者填一份问卷来了解消费者满意度。消费者的满意度取决于商店的服务（后面

会详细举例说明），这些可以通过一些问题得到反馈。问卷可以通过电子邮件、电话、短信等方式进行。例如，可能会收到一条短信要求以“是否”的形式来回答问题。Outlook收件箱可能收到一封邮件，要求对选项“将出席会议”、“不能出席会议”或者“还没有决定”进行投票。

假设多品牌的汽车中心的老板想得到他的消费者的满意度百分比。消费者由于各种各样的原因将他们的车开到服务中心，老板想要了解满意度水平并且找到可以提升消费者满意度的地方来改进服务。众所周知，满意度水平越高，消费者的回头率越高。根据这些设计一份问卷并收集消费者的数据。部分问卷可能如图1-1所示，消费者给出的数据显示出不同的数据特征。消费者编号和问卷编号变量可能是按顺序排列的数字或者随机产生的唯一数字。这些变量是人们回复的唯一标识。也可能会有后续问卷，这时回复者的消费者编号可能相同，而需要改变问卷编号则来标识后续部分。这种变量的一般优点不适用于分析问卷。

Customer ID: **Questionnaire ID**:

1. Full Name (in caps): _____
2. Gender: Male/ Female
3. Age in Years: _____
4. Car Model: _____
5. Car Manufacture (MM/YY): _____/_____
6. Did the workshop fix all your minor problems? Yes No
7. Did the workshop fix all your major problems? Yes No
8. What is the mileage (km/liter) of car? _____
9. Odometer: _____
10. Please give an overall rating of your satisfaction for the work done.
    a. Very Poor
    b. Poor
    c. Average
    d. Good
    e. Very Good

一个假设的调查问卷

在这项调查中的**全名**（Full Name）信息和应答是打破僵局的起点。在非常特殊的情况下，名字可能对分析目的有用。就我们的目的而言，名字将是一个不用于分析目的的简单的**文本变量**。询问**性别**（Gender）是因为在相当多的情况下，性别可能是解释调查的主要特性的重要因素之一，在这种情况下，它是唯一的。**性别**是**分类变量**的一个例子。

年龄（Age in Years）变量是数值型数据，在本质上是一个连续变量的例子。第4个和第5个问题能帮助多品牌经销商识别汽车模型及其年代。在这里，第一个问题是关于汽车模型的类型。客户的车型可能会有所不同（如大众甲壳虫、福特奋进、丰田花冠、本田思域、塔塔纳米），如后面的屏幕截图所示。虽然型号名称实质是一个名词，但是我们在一定意义上对调查问卷的第一个问题做一个区分，前者是一个文本变量，而后者是一个分类变量。其次，

根据汽车模型可以很容易确定汽车的类型，如掀背车、轿车、旅行车或多用途车辆，并且这种分类变量可以像汽车整体尺寸一样作为一个**有序变量**。自出厂之日算起，汽车开了多久可以用来解释里程表中的计数。

第6个和第7个问题简单地询问顾客，这些汽车上的大小问题是否被完全解决了。这是一个二选一的问题，只能选择是或否。小凹痕、电动车窗故障、车里琐碎的声音、音乐的扬声器声音低，以及其他类似的问题，这些问题不会影响车辆的正常行驶，可被视为需修理的次要问题。盘式刹车问题、定位问题、转向时发出咔嗒咔嗒的声音等类似的问题，让车主和其他交通参与者处于危险中而受到高度关注，由于它们影响到汽车的功能，可被视为主要问题。所有用户都希望能够在一次汽车服务中完全解决他们的问题。调查的主要目的就是看服务中心在处理汽车的主要问题和次要问题时的效率如何。用+1和-1或者其他比较简便的标签来标记是或否。

第8个问题“汽车的行驶里程（km/L）是多少?”给出了汽油或柴油平均消耗量的度量。在大多数情况下，这个数据可能只是由顾客估计回答的，在5km/L到25km/L的数据范围内。在里程数较低的情况下/对于不经常开车的客户，他们可能对发动机、车轮定位等部件的精细度要求更高。一般来说，如果里程接近销售公司或者一些机构（如**印度汽车研究协会（ARAI）**）的保证里程，顾客会更开心。一个重要的变量是汽车到达服务点的总里程数。车辆在5000km、10000km、20000km、50000km、100000km范围内进行一定的保养。此变量也可能与车辆的年龄有关。

现在让我们看最后一个问题。这里，询问客户对汽车服务整个体验的评价。客户的反馈可能通过完成汽车服务后马上进行的一个小测试获得，也可以通过发送给客户邮箱ID的问卷获得。非常差的评分表明车间对客户的服务不好，而非常好的评分传达出客户对汽车维修服务完全满意。注意到客户回答的某种规律，因为我们能够把评分按照“非常差<差<一般<好<非常好”的顺序排列。这意味着当我们分析研究数据时，必须参考评分的结构。在下一节中，我们将通过一个假设的数据集来阐述这些概念。

理解在R环境中的数据特点

下图给出一段R会话。这里只涉及调查中的R会话和前文表中的样本数据。此处只简单地感受或者体验R，但是不一定遵循R代码。R安装过程在安装说明中有详细介绍。此处，用户在会话中加载R的数据对象SQ（SQ代表抽样问卷）。对象SQ的本质是一个存储多种其他对象的data.frame。对于data.frame函数的详细技术细节，请参阅第2章。对象data.frame的名称可以使用函数variable.names提取。R中的class函数帮助我们鉴别R对象的本质。当我们有一个变量列表时，使用函数sapply找到它们是非常有用的。上面提到的步骤详见下图。

Customer_ID	Questionnaire_ID	Name	Gender	Age	Car_Model	Car Manufacture Year	Minor Problems	Minor Problems	Mileage	Odometer	Satisfaction Rating
C601FAKNQXM	QC601FAKNQXM	J. Ram	Male	57	Beetle	Apr-11	Yes	Yes	23	18892	Good
C5HZ8CP1NFB	QC5HZ8CP1NFB	Sanjeev Joshi	Male	53	Camry	Feb-09	Yes	Yes	17	22624	Average
CY72H4J0V1X	QCY72H4J0V1X	John D	Male	20	Corolla	Dec-10	Yes	No	21	25207	Good
CH1NZ05VCD8	QCH1NZ05VCD8	Pranathi PT	Female	20	Nano	Apr-10	Yes	Yes	24	42008	Good
CV1Y10SFW7N	QCV1Y10SFW7N	Pallavi M Daksh	Female	54	Civic	Oct-11	Yes	Yes	23	32556	Average
CXO04WUYQAJ	QCXO04WUYQAJ	Mohammed Khan	Male	53	Civic	Mar-12	Yes	No	14	41449	Good
CJQZAYMI59Z	QCJQZAYMI59Z	Anand N T	Male	65	Endeavor	Aug-11	Yes	Yes	23	28555	Good
CIZTA35PW19	QCIZTA35PW19	Arun Kumar T	Male	50	Beetle	Mar-09	Yes	No	19	36841	Very Poor
C12XU9J0OAT	QC12XU9J0OAT	Prakash Prabhak	Male	22	Nano	Mar-11	Yes	No	23	1755	Very Good
CXWBT0V17G	QCXWBT0V17G	Pramod R.K.	Male	49	Nano	Apr-11	No	No	17	2007	Good
C5YOUIZ7PLC	QC5YOUIZ7PLC	Mithun Y.	Male	37	Beetle	Jul-11	Yes	No	14	28265	Poor
CYF269HVUO	QCYF269HVUO	S.P. Bala	Male	42	Nano	Dec-09	Yes	Yes	23	27997	Poor
CAIE3Z0SYK9	QCAIE3Z0SYK9	Swamy J	Male	47	Camry	Jan-12	Yes	Yes	7	27491	Good
CE09UZHDP63	QCE09UZHDP63	Julfikar	Male	31	Endeavor	May-12	Yes	Yes	25	29527	Very Poor
CDWJ6ESYPZR	QCDWJ6ESYPZR	Chris John	Male	24	Fortuner	Aug-09	Yes	Yes	17	2702	Good
CH7XRZ6W9JQ	QCH7XRZ6W9JQ	Naveed Khan	Female	47	Civic	Oct-11	No	No	21	6903	Good
CGXATR9DQEK	QCGXATR9DQEK	Prem Kashmiri	Male	54	Camry	Mar-10	No	Yes	6	40873	Poor
CYQO5RFIPK1	QCYQO5RFIPK1	Sujana Rao	Female	32	Civic	Mar-12	Yes	No	8	48172	Very Good
CG1SZ8IDURP	QCG1SZ8IDURP	Josh K	Male	39	Endeavor	Jul-11	Yes	Yes	8	15274	Poor
CTUSRQDX396	QCTUSRQDX396	Aravind	Male	61	Fiesta	May-10	Yes	Yes	22	9934	Average

一个问卷调查的假设性数据集

```
> data(SQ)
> class(SQ)
[1] "data.frame"
> variable.names(SQ)
 [1] "Customer_ID"           "Questionnaire_ID"      "Name"
 [4] "Gender"                "Age"                   "Car_Model"
 [7] "Car_Manufacture_Year"  "Minor_Problems"        "Major_Problems"
[10] "Mileage"               "Odometer"              "Satisfaction_Rating"
> sapply(SQ,class)
$Customer_ID
[1] "character"

$Questionnaire_ID
[1] "character"

$Name
[1] "character"

$Gender
[1] "factor"

$Age
[1] "numeric"

$Car_Model
[1] "character"

$Car_Manufacture_Year
[1] "Date"

$Minor_Problems
[1] "factor"

$Major_Problems
[1] "factor"

$Mileage
[1] "integer"

$Odometer
[1] "integer"

$Satisfaction_Rating
[1] "ordered" "factor"
```

了解一个R对象的变量类型

变量类型和我们理解的一致，正如它们本来的含义，可以看到，顾客编号（Customer_ID）、问卷编号（Questionnaire_ID）和名字（Name）变量是字符型变量；性别（Gender）、汽车模型（Car_Model）、次要问题（Minor_Problems）和主要问题（Major_Problems）是因子变量；出售日期（DoB）和汽车生产时间（Car_Manufacture_Year）是日期变量；里程数（Mileage）和里程表（Odometer）是数值型变量；最后的满意度变量（Satisfaction_Rating）是一个有序和因子变量。

在本章的余下部分，我们深入探究不同数据类型的属性的更多细节。从更规范的统计学角度来讲，一个变量称为一个随机变量（random variable），在本书的其他部分缩写为RV。这里需要说明的是，在这本书中我们不对概率理论做过多讲解。按照Freund（2003）或Ross（2001）提出的说法来假设读者熟悉概率。随机变量是从概率（样本）空间Ω映射到实数的一个函数。在前面的例子中，我们有里程表（Odometer）和满意度（Satisfaction_Rating）作为随机变量的两个示例。在形式化语言中，随机变量通常使用字母X，Y，...表示。这里需要说明的是，我们所观察的实际值是随机变量的真实值。一般情况下，真实值用小写字母x，y，...表示。让我们更详细地解释一下。

假设我们用X表示随机变量Satisfaction_Rating，这里样本空间Ω包含非常差、差、一般、好、非常好5个元素。为方便起见，我们分别用O_1、O_2、O_3、O_4和O_5表示这些元素。随机变量X取对应概率，P_1，...，P_5的，O_1，...，O_5中的一个值。如果概率已知，那么我们不用担心统计分析。简单地说，如果我们知道随机变量Satisfaction_Rating的概率，那么可以简单地用它来断定客户是否给出更多的“非常好”的评价。然而，我们的调查数据包含每位使用汽车服务的顾客，因此我们有代表性的概率，而不是实际的概率。现在，我们已经看到了在R会话中的20个观察值，并对应于各行有Satisfaction_Rating列下的一些值。对于这20个观察值，我们用符号X_1，...，X_{20}表示满意度。在我们收集数据之前，随机变量X_1，...，X_{20}可以假定为Ω中的任何值。通过收集的数据，我们可以看到，第1个客户给的评定为好（即O_4），第2个为一般（O_3），依此类推，直到第20个客户将服务评定为一般（再次为O_3）。按照惯例，数据表中的观察值X_1，...，X_{20}实际上是随机变量X_1，...，X_{20}的真实值。

1.2 在计算机科学中的不确定性研究

20世纪的人们对机会/随机性持怀疑态度，归因于缺乏精确的仪器，许多变量的这些信息无法得到。对于是否需要当前的随机性建模人们也有所怀疑，人们觉得现在工具太准确，消除了不确定性多变量的信息。但是，事实并非如此。我们来看一些例子。在前面的章节中，我们处理了关于汽车经销商的服务水平问卷调查的数据。人们很自然地接受不同的人以不同的方式做出反应，进一步来说，汽车十分复杂且由不同部件组装，所以在几乎