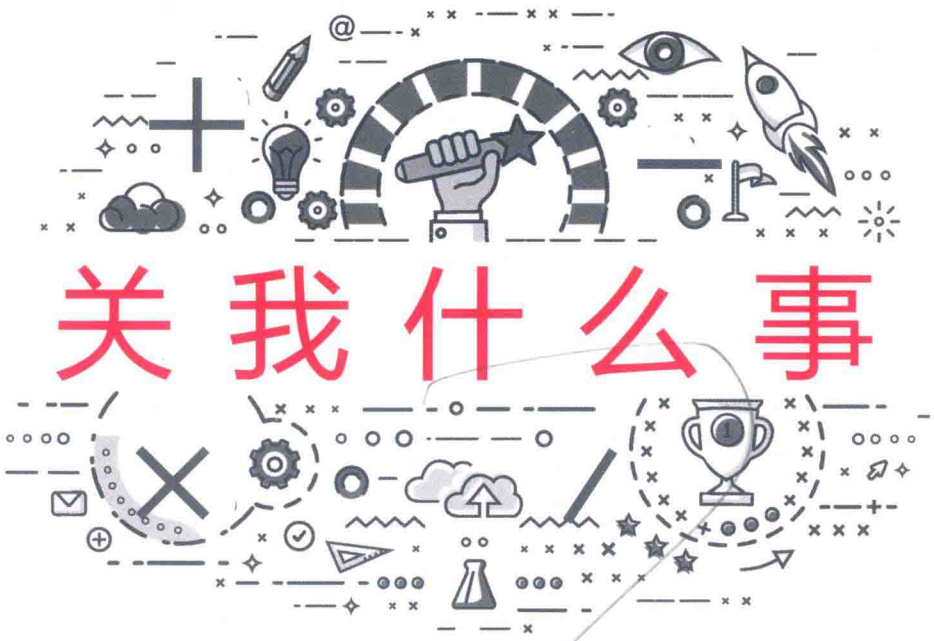
统计学关我什么事



生活中的极简统计学

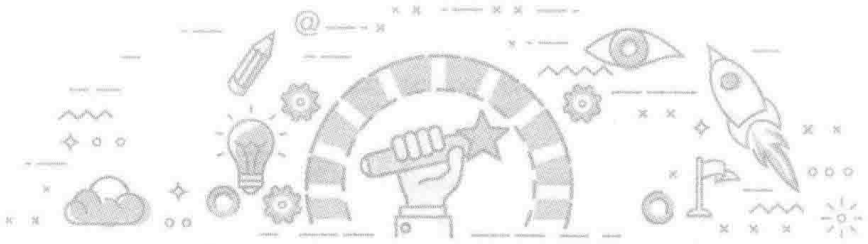
[日] **小岛宽之** 著　**罗梦迪** 译

画画“面积图”就能学，会加减乘除就能学

让数据说话，用统计思维做出好决策

从二胎性别概率、潜在顾客分析、中奖概率大小
到垃圾邮件筛选、微软的帮助系统、谷歌的自动翻译工具……统计学与生活、商业息息相关。

北京时代华文书局



统计学关我什么事

生活中的极简统计学

[日] **小岛宽之** 著　**罗梦迪** 译

北京时代华文书局

图书在版编目（CIP）数据

统计学关我什么事：生活中的极简统计学 /（日）小岛宽之著；罗梦迪译 .
-- 北京：北京时代华文书局，2018.4
ISBN 978-7-5699-2304-9

Ⅰ . ①统… Ⅱ . ①小… ②罗… Ⅲ . ①统计学－通俗读物 Ⅳ . ① C8-49

中国版本图书馆 CIP 数据核字 (2018) 第 055781 号

统计学关我什么事：生活中的极简统计学

TONGJIXUE GUANWO SHENMESHI: SHENGHUO ZHONG DE JIJIAN TONGJIXUE

作　　者 | （日）小岛宽之
译　　者 | 罗梦迪

出 版 人 | 王训海
选题策划 | 胡俊生
责任编辑 | 张超峰
装帧设计 | 红杉林文化　嘉承设计
责任印制 | 刘　银

出版发行 | 北京时代华文书局 http://www.bjsdsj.com.cn
北京市东城区安定门外大街 138 号皇城国际大厦 A 座 8 楼
邮编：100011　　电话：010-64267955　64267677
印　　刷 | 三河市祥达印刷包装有限公司　0316-3656589
（如发现印装质量问题，请与印刷厂联系调换）
开　　本 | 880mm×1230mm　1/32　印　　张 | 8.25　字　　数 | 180 千字
版　　次 | 2018 年 6 月第 1 版　印　　次 | 2018 年 6 月第 1 次印刷
书　　号 | 978-7-5699-2304-9
定　　价 | 48.00 元

目录

第0讲

只要会做四则运算，便可掌握贝叶斯统计学

本书的特点

0-1 从零基础达到应用水平

本书是“贝叶斯统计学”的**超级入门书**。“超级”的含义：

- 从零基础开始学习
- 抛开烦琐的符号和计算过程，学习运用贝叶斯统计
- 不只是随便说说，而是毫无保留地传授知识

对贝叶斯统计学感兴趣的人不在少数。然而此前的教科书，导入部分编写浅显，中途却难度骤增，这使很多读者大受挫折。这是因为在尚未理解贝叶斯统计的本质时，就被灌输了一大堆概率符号，使得理解起来更为困难。

为了不再重蹈这样的覆辙，本书编写之时做了一些功课，具体会在下节进行说明：

0-2 仅使用面积图和简单算术

贝叶斯统计的基础是概率公式——“贝叶斯公式”，它立足于“条件概率”的发展事项。“贝叶斯公式”是高等数学中很难理解的一个概念，原因有二：第一，公式复杂而不够直观；第二，条件概率在某种程度上属于“不可靠的”概念，对于思维缜密的人来说总觉得“哪里有些奇怪”。

事实上，上述第二点在贝叶斯统计中是至关重要的。因为正是这份“不可靠”，才是贝叶斯统计的本质，它与便利性息息相关。后面我们会讲到，贝叶斯统计在 20 世纪初曾因为其“不可靠”而遭到批判，一度被斥于统计学之外。但由于贝叶斯统计的“不可靠”与“便利性”为表里一致的关系，“正因为不可靠才得以运用”。在一部分学者对于这种“便利性”的关注下，贝叶斯统计于 20 世纪后半期恢复了其应有的地位。在 21 世纪的今天，贝叶斯统计已经成为统计学的主流。

笔者着重考虑了这两点，在编写过程中也有所侧重，并做了如下功课。

功课 1 将不出现“贝叶斯公式”（极少一部分除外）的方针贯彻到底

以“通过面积图进行图解”的方针作为贝叶斯公式的替代。从本质上来讲，二者是相同的，然而对于大多数读者而言，图解的方式更加直观且易于理解。同时，通过“面积图”可以更清晰地看出“贝叶斯公式”的“不可靠”和“便利性”究竟体现在哪里。

功课 2 只需简单算术的计算水平即可

这意味着，只需要会做四则运算就可以掌握了，连开方和文字式计

算都不需要。而且这其中的四则运算，即使是不擅长手算的人也可以借助计算器轻而易举地完成。

当然，在本书末尾会出现“贝塔分布”“正态分布”这些有难度的概念。因为如果不介绍这些概念，是无法达到前文所述“毫无保留的传授”程度的。全面理解这些概念，需要用到大学的微分积分知识，这对于许多读者来说是很大的负担。因此在本书中也只能作一些相对简单的解说。

这也就是说，本书的方针——向读者灌输仅通过四则运算就能掌握的公式。这也是本书编写时所做的功课之一。在这个意义上，本书并非“充分齐全”的教材。然而如果想要“充分理解”贝叶斯统计学的人，不妨在读过本书之后再试着挑战一下专业书籍。本书的目的是抛开烦琐的数学概念，将“贝叶斯统计学隐藏的本质”剖析呈现出来。

0-3 比尔·盖茨也在关注它！贝叶斯统计在商业活动中的应用

随着因特网的普及和同步技术的发展，贝叶斯统计开始运用于商业领域。通过互联网可以实现自动收集顾客的购买和检索记录，从而推测顾客的类别。在这一点上，贝叶斯统计学完胜传统意义上的统计学。

如今，**许多互联网企业都在实际应用贝叶斯统计**。其中，微软由于很早就开始在商业活动中运用贝叶斯统计学而闻名。Windows 的操作系统帮助功能中就导入了贝叶斯统计。此外，在网上搜索“小孩病症”的时候，优先显示可靠结果的软件也已经开发出来。微软的前董事长比

尔·盖茨在 1996 年曾在报纸内容中称，微软之所以在激烈的市场竞争中胜出，正是由于采用了贝叶斯统计。比尔·盖茨还在 2001 年关于基本方针的演讲中称，微软的 21 世纪战略正是贝叶斯统计战略，公开表示，已经在全世界范围内挖到了许多贝叶斯统计研究人才。该发言引起了很大关注。

谷歌搜索引擎的自动翻译系统中也引入了贝叶斯统计技术。

当然，贝叶斯统计技术在 IT 企业之外的各个领域也有着广泛应用。例如，消除传真图像中的杂音就运用了贝叶斯统计技术。此外，医疗领域的“自动诊断系统”等也需要用到贝叶斯统计。

通过阅读本书可以得知，**贝叶斯统计的优势在于，“在数据少的情况下也可以进行推测，数据越多，推测结果越准确”，以及“对所获的信息可做出瞬时反应，自动升级推测”的学习功能**。了解了这一点之后，就完全可以理解为什么贝叶斯统计是非常适合应用于高端商业的技术了。

从事商业活动的人，如果能够熟练使用贝叶斯统计，那是再好不过的。本书中的案例和解说，为这一类人群提供了很好的参考。

0-4 贝叶斯统计依存于人的心理

在 0-2 节中有提到，“贝叶斯统计在某种程度上是不可靠的”。究其原因，是由于**贝叶斯统计中所涉及的概率是“主观的”**。换言之，通过贝叶斯统计得到的概率并非客观的数值，而是依存于人的心理的主观

数值。在从这个意义上讲，贝叶斯统计具备了一定的“思想”。也正是因此，注重客观性的科学界为贝叶斯统计打上了“假冒伪劣”的烙印，并导致它一度消亡。

然而，遗憾的是，关于贝叶斯统计学的绝大部分书籍中，并未对这一问题进行记载。也许是作者们不愿将其公之于众，抑或只是因为他们对此不甚了解罢了。实际上，**几乎没有一本教科书对于这个问题正面进行过阐述。**然而，所谓的“主观性”和“思想性”，才正是贝叶斯统计学的本质和它具有便利性的根本原因所在。因此，在解说贝叶斯统计学的时候，如果忽视掉这一点，是难以将贝叶斯统计学的本质传达给读者的。

本书不刻意避开贝叶斯统计的“主观性”和“思想性”，而是将这些特点展现出来进行解说，特别是对于贝叶斯统计学与传统的统计学之间的差异进行详解。希望众多读者能够为贝叶斯统计学的神奇和有趣拍手称赞。

0-5 附带简单的填空练习题，适合自学

本书沿袭之前出版的《完全自学 统计学入门》（钻石社）的编写方法，用最详尽的语言解释说明，并在每一讲之后设置简单的填空练习题。学习数学的最佳方法是做一些简单的练习题。本书中收录的练习题并非应用题的形式，而是用来对讲义内容进行巩固的，希望各位读者认真练习，加深理解。

读完这本书，您一定会产生这样的想法：

“咦？明明没有经过登山训练，却不知不觉到了山顶呢！”

那么，就让我们向着山顶，出发吧。

第1部

快速学习！理解贝叶斯统计学的精髓

在第1部中，将为您解说关于“贝叶斯统计的推算应该用何种方法来思考，具有什么样的性质”的问题。解说中采用了我们身边的许多事例，如“这位顾客是来买东西，还是随便逛逛”“收到的是真命巧克力？还是义理巧克力”对于读者来说，这些例子应当是很容易想象和理解的。另一方面，本书内容涉及贝叶斯统计学与“序贯理性”“内曼－皮尔逊统计学”的区别，这对于贝叶斯统计学的特征，已经探讨得相当深入了。

第1讲

信息增加导致概率变化
“贝叶斯推理”的基本方法

1-1 通过贝叶斯推理来辨别“买东西的人”和“随便逛逛的人”

本讲将通过一个商业案例，为大家介绍经典的贝叶斯推理方法。

商店里的售货员最关心的问题莫过于**“这位顾客究竟是来买东西的，还是随便逛逛而已”**。真正来买东西的顾客，一般而言，比起四处逛逛看看，更倾向于在最短时间内找到自己需要的商品。另一类顾客则是这样的：一时不急着买，而是先随便问问价格，为以后购买做个参考。对待前者，作为售货员，理应为其介绍需要的商品并让其买下；而对待后者，如果同样花费时间为其推荐商品，顾客不但不会购买，反而会感到厌烦，结果适得其反。

所以对于店员来说，通过顾客的行为来揣测他们的真实想法，是一项重要的本领。很多店员可以做到：通过直觉来判断顾客属于哪一类，而这正是身为一名店员的重要工作技巧。在此，我们将这种“基于直觉的判断”数值化，从而使它可以通过计算获得。把方法编成手册，教给新店员，这就像在互联网上能够实现自动判断的 AI（人工智能）一样，

是一项意义非凡的工作。

下文将具体介绍“将店员的判断方法数值化”的方法，该方法恰巧适用贝叶斯统计学。进而言之，通过该事例，我们也可以弄懂贝叶斯统计学的概念。下文将分节进行解说。

1-2 第一步：通过经验设定“先验概率”

假设一个场景：面前有一位顾客，此时你需要做的是，推测该顾客究竟是“来买东西的人”，还是“随便逛逛的人”。只有做出正确的判断，才能采取正确的接待方法。

推算的第一步：将两种顾客（来买东西的顾客、随便逛逛的顾客）的比例进行数值分配。这句话的意思是：假设面前的这位顾客一定属于两种中的一种，以此为前提，该顾客为第一种或第二种的可能性分别为多少？将这个可能性用数值表示出来。

在贝叶斯统计学中，这种**“某种类别的概率（比例）”**有一个专有名词，叫作**“先验概率”**。“事前”的含义是：在获得某项信息之前。此

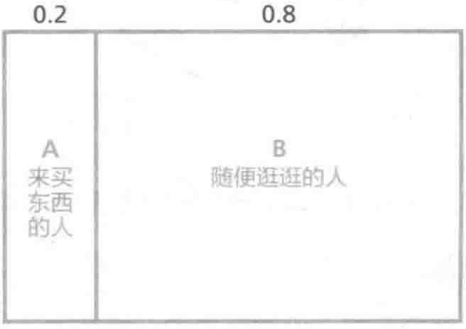
处的**“信息”**是指：**附加的状况**，比如顾客忽然过来询问。通过“过来询问”这一信息，可以对顾客类别的**推算进行修改**，而**“先验概率”**是指，在“过来询问”或“不过来询问”的情况发生之前进行的概率判断。

通常，**“先验概率”**可通过经验来判断。在特殊情况下，即使没有类似经验，也可以进行判断，这部分特殊事例将在第 3 讲进行解说，此处暂且不做讨论。

根据自己的经验，每 5 位顾客中就有 1 位是“来买东西的”，也就是说，这一部分顾客占全体的 20%（0.2），那么剩下“随便逛逛”部分的比例便为 80%（0.8）。这两个数字，便是两类顾客的**“先验概率”**。

在这个事例中，在观察面前顾客的行为之前，判断“该顾客是属于概率 0.2 的买东西的人，还是概率 0.8 的随便逛逛的人”，这个过程被称为**“某一类别的先验分布”**，如**图表 1-1** 所示。

图表 1-1　先验分布：分割长方形



图表 1-1 中的大长方形被分割为两部分，两部分的面积所占比例分别为 0.2 和 0.8，这正是分割时的诀窍。本书将在后面逐渐阐明：**“面积”的概念在贝叶斯概率的计算中，起着重要的作用。**