

统计思维

統計学が最強の学問である
「実践編」

「目」
西内启

李晨——著
译

大数据时代
瞬间洞察因果的
关键技能

所谓“大数据”，骨子里其实就是统计学
前作畅销 37 万册的日本人气教授西内启持续发力，
专为职场人士量身打造

后浪出版公司

统计思维总有一天会像读与写一样，
成为有效率公民的一种必备能力。

—— H.G. 威尔斯

浙江人民出版社

统计思维

后浪出版公司

{ 大数据时代
瞬间洞察因果的
关键技能 }

西内启

李晨
译

著

浙江人民出版社

图书在版编目 (CIP) 数据

统计思维 / (日) 西内启著 ; 李晨译 . -- 杭州 :
浙江人民出版社 , 2017.7
ISBN 978-7-213-08338-9

浙江省版权局
著作权合同登记号
图字 : 11-2017-213

I . ①统… II . ①西… ②李… III . ①概率统计
IV . ① O211

中国版本图书馆 CIP 数据核字 (2017) 第 193683 号

TOKEIGAKU GA SAIKYO NO GAKUMON DEARU [JISSEN HEN]

by HIROMU NISHIUCHI

Copyright © 2014 HIROMU NISHIUCHI

Chinese (in Simplified character only) translation copyright © 2017 by Ginkgo(Beijing)Book Co., Ltd.

All rights reserved.

Chinese (in Simplified character only) translation rights arranged with Diamond, Inc.
through Bardou-Chinese Media Agency.

统计思维

[日]西内启 著 李 晨 译

出版发行 : 浙江人民出版社 (杭州市体育场路 347 号 邮编 310006)

责任编辑 : 潘海林

责任校对 : 徐永明

特约编辑 : 李 峥

封面设计 : Cyali

印 刷 : 北京盛通印刷股份有限公司

开 本 : 690 毫米 × 960 毫米 1/16

印 张 : 23

字 数 : 280 千

版 次 : 2017 年 12 月第 1 版

印 次 : 2017 年 12 月第 1 次印刷

书 号 : ISBN 978-7-213-08338-9

定 价 : 52.00 元

后浪出版咨询(北京)有限责任公司 常年法律顾问 : 北京大成律师事务所

周天晖 copyright@hinabook.com

未经许可, 不得以任何方式复制或抄袭本书部分或全部内容

版权所有, 侵权必究

本书若有质量问题, 请与本公司图书销售中心联系调换。电话 : 010-64010019

目 录

序 章

商务与统计学之间千丝万缕的联系	001
-----------------	-----

01 商务与统计学之间的鸿沟到底因何存在	003
----------------------	-----

02 “把握”“预测”与“洞察”的统计学	007
----------------------	-----

第 1 章

当我们谈论统计学时，我们在谈些什么	015
-------------------	-----

03 “洞察”的统计学必须掌握的三大基本功	017
-----------------------	-----

04 深奥的“平均值”	027
-------------	-----

05 为何平均值可以捕捉真实情况？	036
-------------------	-----

06 了解商务数据的切入点——标准差	050
--------------------	-----

第 2 章

统计学为何是“最强”的商务武器	063
-----------------	-----

07 介于冒失鬼和糊涂虫之间的“最强”思维方式	065
-------------------------	-----

08 “误差范围”与数据数量的关系	071
-------------------	-----

- 09 假设检验的思维方式 082
- 10 用z检验来劝告冒失鬼 089
- 11 专为少量数据设计的屠龙刀 104
- 12 检验的多重性及其处方 116

第3章

堪称洞察之王道的各种分析工具

127

- 13 统计学的王道——“回归分析” 129
- 14 如何求回归直线? 138
- 15 从错综复杂中抽丝剥茧的多元回归分析 148
- 16 Logistic回归与对数比 163
- 17 商务应用中回归模型总结 177
- 18 商业实战中回归模型的使用方法——输入篇 187
- 19 商业实战中回归模型的使用方法——输出篇 192

第4章

挖掘数据背后隐藏的宝库

201

- 20 心理学家开发的因子分析 203
- 21 笔试得分高,商务礼仪也会相应优秀? 210
- 22 聚类分析基本思想 221
- 23 素未谋面的100人如何实现高效沟通? 232

24 本书的总结 247

25 商务实践中的分析顺序 252

26 本书无法学到的3点知识 259

谢 辞

269

数学附录

271

附录1 偏差的绝对值与中位数 271

附录2 偏差的平方与平均值 275

附录3 平均值与比例的标准误差 277

附录4 方差与无偏方差 282

附录5 正态分布的数学性质 284

附录6 中心极限定理 292

附录7 切比雪夫不等式 297

附录8 关于平均值与比例之差的 z 检验 299

附录9 χ^2 (卡方) 分布与 t 分布的关系 302

附录10 费希尔确切概率检验 308

附录11 z 检验与 χ^2 (卡方) 检验 311

附录12 邦费罗尼校正 318

附录13 一元回归分析 321

附录14 一元回归分析与 t 检验的关系 328

附录15 多元回归分析 331

附录16 比值比 337

附录17 统计功效与样本量设计 341

参考文献	348
索引	352
出版后记	357

序 章

商务与统计学之间千丝万缕的联系

01 商务与统计学之间的鸿沟到底因何存在

《统计学是最强学问》是本怎样的书

市面上有许多统计学入门书，其写作方式却鲜有变化。首先是概率论，然后介绍正态分布之类的概率分布，接下来讲解估计、检验、相关系数和回归分析，一般就是以这样的顺序从数学的角度理解上述各种概念。

有幸售出35万册的畅销书《统计学是最强学问》（以下简称“前作”），则是统计学“入门的入门”。它从实用的角度说明了统计学在现代社会中发挥着多大的作用，以及现今普遍使用的统计方法，是经过了怎样的历史、因何人的思考而诞生的。可以说，前作是将涉足统计学领域所需的基本常识凝缩在了一本书中。

因此，在看到“即使读了这本书也无法应用统计学”这样的感想或批判时，我毫不意外，因为事实确实如此。

企业对大数据赞不绝口，可引进了昂贵的系统却只用来画一张漂亮的饼状图。前作的目的仅在于填补统计学与社会之间的鸿沟，并改变上

述状况。

若能吸引更多人关注统计学，我的目的便达到了。接下来，读者只要在琳琅满目的入门书中选出适合自己的来学习，日本人的统计能力自然也就会上升……这就是我当时的想法。

续作（本书）的理由

听到了同侪的反馈，我才知道这种想法似乎有些过于乐观。这就是写作本书的理由。

现将他们所认为的现有统计学入门书不适合自己的理由总结如下：

- 出现公式就读不下去
- 乍一接触到统计工具，不知道它的含义
- 不知道各种方法对自己的工作有何助益
- 不知道自己工作适用哪些统计学知识

他们曾问我是否有书能满足上述需求，但我确实尚未见过这样的书。

前作中也曾提到，统计学是有力且广泛通用的工具，在诸多学术领域都有应用。各学科的目的、思维方式、研究对象的性质不同，同样的统计方法会有不同的应用方式，更有许多专门用于某一学术领域的统计方法。正是因此，经济学和心理学本科生的统计学教材内容相差很大。那些不想提及这些差距而仅介绍共同部分的统计学入门书，便只能使用抽象的公式，因而就会枯燥无味。

也就是说，大多数教科书与商务人士对统计学的需求并不相符，只是因为这些书原本就不是为了用于商业而写，这类书的作者和出版机构

并没有任何过错。

为什么找不到好的统计学教科书？

理想情况下，最适合为商务人士写作统计学教科书的，也许是在商业各领域积累了定量分析实务经验的管理专家。

然而，管理专家中很多都在做历史与个案的定性研究，即使是进行定量分析的研究者，研究主题也不一定是用统计来解析如何提高某项业务收益率。举例来说，很多管理专家是从企业战略和收益关系等宏观视角进行统计分析。在分析某项业务的客户时，不同的专家也有着迥异的指导思想和方法。另外，也有人善于使用本质与统计学完全不同的、名为管理会计财务分析的定量分析方法。

这一点也同样适用于最近广受瞩目的“数据科学家”。将这些研究数理统计学或机器学习^①方法的研究员放到商业活动的第一线，他们也不一定能立刻派上用场。即使有个别人能够立刻适应环境，那也只能说是出于运气，或是因为他本身算是商业直觉很好的年轻人罢了。

理解分析方法本身或是了解众多特殊的分析方法，与思考在现实中如何活用这些方法、创造何种价值，本身就是不同的事情。即使你所在的企业聘用的年轻数据科学家没有立刻适应工作，错也不在他们。与上文类似，只是因为他们并不是为了商业目的而学习罢了。

写到这里，大概会有人想问我到底是哪一种人。我与统计相关的职业生涯，是我易于将知识活用于商业活动的基础。

^① 是作为人工智能核心的多领域交叉学科，研究计算机如何模拟和实现人的学习行为以获取新的知识或技能，重组已有的知识结构使之不断改善自身的性能。——译者注

我早先致力于研究公共卫生学。对于这门学问,只要目标是人类健康,采用何种研究方法都可以。因此我在美国求学时所属的公共卫生学研究生院 (School of Public Health) 不仅有医学院毕业后拿到医生执照的老师,还有经济学家、法学家、教育学家、社会学家、信息技术从业人员、曾在广告代理机构任职的营销专家……来自各个领域的专业人士聚集在那里,以人类健康为目标活用各种知识,形成新想法。

物理学是自然科学的王者,经济学是社会科学的女王,公共卫生学则可以说是科学的综合格斗术。尽管在进行“格斗”时可以选择任何方法,但最后一定会归结于“证据”,也就是统计学的分析结果。有时即使对手是女王,也要用证据将其击倒,而最近女王也正在学会用证据来反击。

我从年轻时开始直到现在,每日运用统计学技术分析横跨各领域(从基因到公共卫生政策)的实证数据并将其活用,这大概是我侥幸获得的最大财富。若是没有这笔财富,我也许根本不会想在前作中介绍不同领域的统计学的思维方式以及其中的差异吧。

02 “把握”“预测”与“洞察”的统计学

商务活动需要的是“洞察”人的统计学

除了能让人理解许多领域中统计学的使用方法及其差别，公共卫生学所使用的统计学还在另一点上有助于商务活动。那就是它与几乎所有的商务活动相同，以人类的行为和社会状态的变化作为研究对象。

让我们和其他的学术领域比较一下。工学和农学的研究对象分别是非生物和人类以外的生物，我们可以控制大部分条件对它们的量和质进行实验，如在超低温下停止分子的震动来测定数据，或是收集实验生物所有的基因，只要技术和预算允许，实验想做就可以做。

与之相对，大多数人只能从观察和预测的角度来看计量经济学所研究的股票价格和经济状况等宏观变动。

研究声音、图像和自然语言的机器学习领域，也有其特殊的专业技术，如共振峰频率、特征点和修饰结构等。它以声音、图像和句子知识为基础，赋予0/1二进制记录的数据以意义，通过算法准确快速地处理对人类而

言理所当然的认知行为。

需要事先声明的是，如果你所需要的是这方面的统计学知识，这本书对于你来说可能并没有用。实际上已经有了许多更加适合你的书。这本书所涉及的，全部是为了洞察个人，并对其行动和状态进行部分改善的统计学。

人类个体多样，决策机制复杂，用自己的行为来改变别人很难，或者说强制改变别人的做法在伦理上面临困难。因此经济学家研究合理性，政治学家研究权利，社会学家研究人与社群的相互作用，心理学家研究认知和感情，教育学家研究知识和能力，营销专家研究需求和欲望，就像这样，不同的领域尝试从不同的切入点去理解人类。然而，如果被问到什么对改善人的行动来说最重要，我一定会回答“全部都很重要”。

我在商务领域从事分析工作时，无论身处哪个行业、哪种职位，都需要用数据来洞察个人和集团的行动，进而寻找改变这种行动的方法。在这几年中，我在调查和数据分析上都没有遇到困难，应该是因为商务活动所需要的，就正是洞察人类行为的统计学吧。

统计学应用于商务的三大利器

洞察人类的统计学具体做的，就是“洞察”人类行为的“因果关系”。此外，统计学还被用于“把握现状”和“预测未来”，但本书（基本上）没有涉及它们。

所谓“把握现状”，举例来说，就是通过市场调查来估算现在有多少人正在使用某种产品。本书并不涉及这一类的统计学，原因在于这种方法已经被使用在商务第一线，且已十分完善。

已经有许多商务人士从调查数据中推算出平均值和比例，并将之总

结成图表。调查公司的发展也使得几千人规模的调查不再那么困难了。

只要收集了相应的数据，即使不理解标准误差的概念而没有将其写在报告中，只要误差并不大，这些数据使用起来也没有什么问题。明显偏向某些性别和年龄的数据可以推测整个日本的平均值吗？部分项目未被回答的问卷要如何处理呢？现在有着各种高级的统计方法来处理这类问题，所以很少有商务人士会因为执着于此而获利。实际上，在为其提供建议的时候，我也只会说：“有关误差和调查对象分布偏差的问题，调查公司能够给出准确的答案，交给他们去做比较好吧。”

另外，制造业的质量管理方面，试验品的规格平均值是多少，其浮动有何特点，也属于“把握现状”的统计学方法，但这也不在本书的讨论范围之内。制造业利用统计学来进行质量管理可称得上是战后日本的看家本领，无论是门外汉还是专家，公司内部和业界的培训与资料内容都极其丰富。那些在我出生之前就能在商业中活用统计学的人，已经把我要说的都说完了。如果一定要解释，以往的统计学的应用，都是以“物体”和“能以物理方法测量的量与质”为基础，而我想要提出的，则是将统计学应用于“人”和“无法以物理方式测量的量与质”的领域。如果你对这点感兴趣，本书可能会对你有所助益。

“预测未来”，举例来说，就是利用统计学来尝试准确地预测未来股价或原材料价格是否会上升，或者库存会如何变化。此外，机器学习的图像和声音识别当中，仅从数据来准确预测“如果是人类会如何认识”，以此来模仿人类的认知方式，也属于“预测未来”。

本书不涉及这一点的原因是，在没有弄清分析方法之前，根据复杂的状况做出准确的预测是非常困难的。经济学家写作的统计学入门书中有时会有利用时间序列分析来预测股票价格的例子，但是如果真的可以

通过充分预测来获利，这些方法也就根本不会写在书上了（即使所写的方法果真能够获利，模仿者的增加就会改变市场环境，其结果，或是利润因介入者过多而减少，或是加大经济泡沫，受到强烈批判）。

熟知各种统计方法的专业投资家的收益率，其实不会比随机购买上市公司全体股票这种猴子也能做到的投资方法收益率高。如果想要使用统计学来增加资产，我认为还不如重视上述实证数据（《漫步华尔街》，伯顿·马尔基尔著）。

关于预测的困难，纳特·西尔弗所著的《信号与噪声》中有详细介绍，对统计学在预测上的使用方法更加有兴趣的读者不妨参考阅读。

如何将“洞察”的统计学转化为商业优势？

在供应与采购部门工作的人，最关心的可能是预测采购价格和出货量的变化，然后制定相应对策。然而对于市场部门来说，洞察常常比预测更重要。举例来说，比起准确预测“这种制成品可以卖出多少”“进行何种推广活动可以将商品卖出去”“做出何种商品才能大卖”之类的洞察才能成为利益的源泉。也就是说，追溯客户购买背后的原因的这种“因果关系”才是重要的。

医学和公共卫生学领域也是这样。即使在统计学上明确了人的生活习惯与死亡率之间的关系，大部分医学相关人士对于准确预测某个人会在几岁时死亡的兴趣也并不大。统计学被应用于医学的目的，不是在于探索人会在几岁时死亡这一结果，而是在于如何让人更加健康长寿。“再这样下去你就只剩下多少年的寿命了”，大概有读者在体检时被医生这么威胁过。前作中提及（本书后面也会出现）的弗雷明汉研究项目中，诞生了根据性别、年龄、血压、吸烟史等项目计算心脏病发病概率的“弗雷明汉危