HZ BOOKS
华章科技

CAMBRIDGE

硅谷创业教父　Steve Blank
微软前执行副总裁　沈向洋
脸书前首席技术官　Adam D'Angelo
谷歌高级院士　Jeff Dean
等37位专家　联/袂/推/荐

微软、亚马逊、谷歌和领英等公司
互联网产品成功的秘诀！

关键

可信赖的线上对照实验

迭代

罗恩·科哈维（Ron Kohavi）
[美] 黛安·唐（Diane Tang）　著
许亚（Ya Xu）
韩玮　胡鹃娟　等译

TRUSTWORTHY ONLINE CONTROLLED EXPERIMENTS

A Practical Guide to A/B Testing

机械工业出版社
China Machine Press

关键迭代
可信赖的线上对照实验

Trustworthy Online Controlled Experiments
A Practical Guide to A/B Testing

罗恩·科哈维（Ron Kohavi）
［美］黛安·唐（Diane Tang） 著
许亚（Ya Xu）

韩 玮 胡鹃娟 段玮韬 胡泽浩
廖一正 王 璐 赵振宇 钟 婧 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

关键迭代：可信赖的线上对照实验 /（美）罗恩・科哈维（Ron Kohavi），（美）黛安・唐（Diane Tang），（美）许亚（Ya Xu）著；韩玮等译 . -- 北京：机械工业出版社，2021.4
（数据分析与决策技术丛书）
书名原文：Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing
ISBN 978-7-111-67880-9

I. ① 关… II. ① 罗… ② 黛… ③ 许… ④ 韩… III. ①数据处理 IV. ① TP181

中国版本图书馆 CIP 数据核字（2021）第 055899 号

本书版权登记号：图字 01-2020-7582

This is a Simplified-Chinese edition of the following title published by Cambridge University Press: Ron Kohavi, Diane Tang, Ya Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*（978-1-108-72426-5）.

© Ron Kohavi, Diane Tang, and Ya Xu 2020.

This Simplified-Chinese edition for the People's Republic of China (excluding Hong Kong, Macau and Taiwan) is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press and China Machine Press in 2021.

This Simplified-Chinese edition is authorized for sale in the People's Republic of China (excluding Hong Kong, Macau and Taiwan) only. Unauthorized export of this simplified Chinese is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of Cambridge University Press and China Machine Press.

Copies of this book sold without a Cambridge University Press sticker on the cover are unauthorized and illegal.

本书原版由剑桥大学出版社出版。

本书简体字中文版由剑桥大学出版社与机械工业出版社合作出版。未经出版者预先书面许可，不得以任何方式复制或抄袭本书的任何部分。

此版本仅限在中华人民共和国境内（不包括香港、澳门特别行政区及台湾地区）销售。

本书封面贴有 Cambridge Unirversity Press 防伪标签，无标签者不得销售。

关键迭代：可信赖的线上对照实验

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：王春华 柯敬贤　　责任校对：殷 虹

印 刷：大厂回族自治县益利印刷有限公司　　版 次：2021 年 4 月第 1 版第 1 次印刷

开 本：170mm×230mm 1/16　　印 张：18.5

书 号：ISBN 978-7-111-67880-9　　定 价：99.00 元

客服电话：（010）88361066 88379833 68326294　　投稿热线：（010）88379604

华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

版权所有・侵权必究

封底无防伪标均为盗版

本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

本书赞誉

“精益方法的核心是科学的方法：创建假设、运行实验、收集数据、提取洞察以及对假设进行验证或修改。A/B 测试是创建可验证且可重复的实验的黄金标准，而本书正是这方面的权威书籍。”

——Steve Blank，斯坦福大学兼职教授，硅谷创业教父，*The Startup Owner's Manual*[⊖]以及 *The Four Steps to the Epiphany* 的作者

“对于希望通过线上对照实验来优化产品功能和提高项目效率或营收的管理人员、领导者、研究人员或工程师，本书是很好的资源。我十分了解 Kohavi 的工作对微软和必应产生的影响。很高兴看到这些知识现在可以分享给更广泛的读者。”

——沈向洋，微软人工智能研究部门前执行副总裁[⊜]

“一本既严谨又通俗易懂的好书。读者可以学习到如何将革新了互联网产品开发的可信赖的对照实验引入自己的机构。”

——Adam D'Angelo，Quora 联合创始人兼首席执行官，脸书前首席技术官

“本书很好地概述了几家公司如何通过线上实验和 A/B 测试来改善其产品。经验丰富的 Kohavi、Tang 和 Xu 在本书中分享了出色的建议。本书提供了许多现实世界的实用示例，以及作者多年积累的关于大规模应用这些技术的经验和教训。”

——Jeff Dean，谷歌高级院士兼谷歌搜索高级副总裁

⊖ 本书中文版《创业者手册》已由机械工业出版社出版，ISBN 978-7-111-40530-6。——编辑注

⊜ 沈向洋已于 2019 年 11 月 14 日辞去微软公司执行副总裁的职务，2020 年 3 月 5 日受聘为清华大学高等研究院双聘教授。——译者注

“你是否希望你的组织持续做出更好的决策？本书是关于如何在数字时代基于数据做决策的新‘圣经’。阅读本书就像身处亚马逊、谷歌、领英或微软的内部会议一样。作者首次披露了世界上最成功的公司如何做决策。除了普通商业书籍的忠告和逸事之外，本书还展示了该做什么以及如何做好。这是数字世界中关于决策的手册，其中有专门针对业务主管、工程师和数据分析师的内容。”

——Scott Cook，Intuit 联合创始人兼执行委员会主席

“线上对照实验是十分强大的工具，理解它的工作原理、优势以及如何优化它，可以使专家和更广泛的读者受益。本书是罕见的技术权威，有机地将良好的阅读体验和重要课题的探讨结合起来。”

——John P. A. Ioannidis，斯坦福大学医学、
健康研究与政策、生物医学数据科学和统计学教授

“哪个线上选项会更好？我们经常需要做出这样的选择。要确定哪种方法更有效，我们需要严格的对照实验，也就是 A/B 测试。微软、谷歌和领英的专家撰写的这本精彩且生动的书介绍了 A/B 测试的理论和最佳实践，是所有从事线上相关业务的人的必读书籍！”

——Gregory Piatetsky-Shapiro 博士，KDnuggets 总裁，
SIGKDD 联合创始人，数据科学与分析领域“领英最强音”得主

“ Ron Kohavi、Diane Tang 和 Ya Xu 是线上实验方面的顶级专家。多年来我一直在使用他们的工作成果，很高兴他们现在联手编写了这本终极指南。我向我的所有学生以及所有参与线上产品和服务的人推荐这本书。”

——Erik Brynjolfsson，麻省理工学院教授，*The Second Machine Age* 的合著者

“如果没有线上对照实验，靠软件支撑的现代商业就会缺乏竞争力。本书由该领域最有经验的三位先驱撰写，书中介绍了该领域的基本原理，并通过令人信服的示例对其进行了说明，同时深入探讨并提供了大量实用建议。这是一本必读书籍！”

——Foster Provost，纽约大学斯特恩商学院教授，
畅销书 *Data Science for Business* 的合著者

“在过去的 20 年里，科技行业也逐渐认识到对照实验是理解复杂现象和解决极具挑战性问题的绝佳工具，这一点科学家们早在几个世纪以前就了解了。设计对照实验、大规模运用它们并解释其结果的能力是现代高科技企业经营的基础。本书作者设计并实现了多个世界上最强大的实验平台。阅读本书是从他们的经验中学习如何使用这些工具和技术的绝好机会。”

——Kevin Scott，微软执行副总裁兼首席技术官

“线上实验推动了亚马逊、微软、领英以及其他领先的数字化企业的成功。这本实用的图书为读者了解这些公司数十年的实验经验提供了难得的机会。本书应该放在每个数据科学家、软件工程师和产品经理的书架上。”

——Stefan Thomke，哈佛商学院 William Barcley Harding 教授，

Experimentation Works: The Surprising Power of Business Experiments 的作者

“线上业务成功的秘诀在于实验，这已不再是秘密。三位大师在本书中详解了 A/B 测试的基本组成元素，你可以据此不断改进你的线上服务。”

——Hal Varian，谷歌首席经济学家，

Intermediate Microeconomics: A Modern Approach 的作者

“实验是面向线上产品和服务的最佳工具之一。本书包含了从微软、谷歌和领英的多年成功测试中获得的实践知识。通过真实的示例、陷阱及其特点、解决方案来分享洞察和最佳实践。我强烈推荐这本书！”

——Preston McAfee，微软前首席经济学家兼副总裁

“实验是数字策略的未来，这本书将成为实验领域的‘圣经’。Kohavi、Tang 和 Xu 是当今实验领域最著名的三位专家，他们的书提供了数字实验的实用路线图，非常有用。本书将他们数十年来在微软、亚马逊、谷歌和领英进行的案例研究整理成通俗易懂、有深度且清晰的实用方法，任何数字业务的经理都应阅读这本书。”

——Sinan Aral，麻省理工学院 David Austin 管理学教授，*The Hype Machine* 的作者

“对于任何严肃的实验从业者而言，这本书都是必不可少的。它非常实用，同时又很深入，这是我之前从未见过的。阅读它能让你感觉自己拥有了超能力。从统计的细

微差别到评估结果，再到衡量长期影响，这本书全都涵盖了。必读！”

——Peep Laja，顶级转化率专家，CXL 的创始人兼技术负责人

“线上实验对于改变微软的文化至关重要。当 Satya[⊖]谈论‘成长心态’时，实验是尝试新想法并从中学习的最好方法。学习如何快速迭代对照实验增强了必应的盈利能力，并通过 Office、Windows 和 Azure 业务线迅速在微软推广。”

——Eric Boyd，微软人工智能平台企业副总裁

“作为一名企业家、科学家和高管，我艰难地学到：一盎司的数据顶得上一磅的直觉[⊜]。但是如何获得好的数据？本书将作者在亚马逊、谷歌、领英和微软的数十年经验汇编成易于学习且井井有条的指南。这本书是线上实验的‘圣经’。”

——Oren Etzioni，Allen Institude of AI 的首席执行官兼华盛顿大学计算机科学教授

“互联网公司以前所未有的规模、速度和复杂度运行着实验。本书作者在这些发展中发挥了关键作用，能够从他们的经验中学习，读者很幸运。”

——Dean Eckles，麻省理工学院通信与技术专业 KDD 职业发展教授，脸书前科学家

“本书为一个关键但未受到足够重视的领域提供了丰富的参考资源。每章的实例研究展示了成功业务的内在运作和经验。重视开发和优化‘综合评估标准’是特别重要的一课。”

——Jeremy Howard，奇点大学，fast.ai 的创始人，Kaggle 的前总裁兼首席科学家

“关于 A/B 测试的指南有很多，但很少有像本书这么正统的。我已经关注 Ronny Kohavi 18 年了，发现他的建议聚焦于实践，经过了经验的‘打磨’并在实际环境中进行过检验。Diane Tang 和 Ya Xu 的加入使得理解的广度变得无与伦比。我建议你将这本著作与其他任何书进行比较——当然是以对照方式。”

——Jim Sterne，市场分析峰会创始人，数据分析协会名誉理事

“这是一本关于运行线上实验的极其有用的方法书。书中结合了复杂的分析方法、

⊖ 微软首席执行官。——译者注

⊜ 一盎司约等于 28.35 克，一磅约等于 450 克。这里意指数据的重要性远远高于直觉。——编辑注

简洁的论述以及来之不易的实践经验。”

——Jim Manzi，Foundry.ai 的创始人，
Applied Predictive Techonologies 的创始人和前首席执行官兼董事长，
Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society 的作者

“每当实验被设计用于新领域的时候，如农业、化学、医药以及现在的电子商务等，都可以帮助该领域取得进步。本书由三位业内顶尖专家所著，涵盖了丰富的实践建议，以及如何和为何运行线上实验并避开陷阱。实验是有成本的，不懂得哪些方法可行会增加更多的成本。”

——Art Owen，斯坦福大学统计系教授

“这是一本商业主管和运营经理的必读书。就像运营、金融、审计和战略组成了现今商务的基础一样，在这个人工智能的时代，理解和实践线上对照实验将成为必备的知识点。Kohavi、Tang 和 Xu 在这本书里罗列了这个知识领域切实可行的核心内容。”

——Karim R. Lakhani，哈佛大学科技创新实验室教授及总监，
Mozilla Corp. 董事会成员

“真正的‘数据驱动’型组织深知仅有数据分析是不够的，还必须致力于实验。这本书是影响力极大的实验设计的手册和宣言，简明易懂且出类拔萃。我认为本书的实用主义很值得借鉴。最关键的是，本书阐明了企业文化是与技术实力旗鼓相当的重要成功因素。”

——Michael Schrage，麻省理工学院数字经济项目科研院士，
The Innovator's Hypothesis: How Cheap Experiments Are Worth More than Good Ideas 的作者

“这本关于实验的重要图书融汇了三个来自世界顶尖科技公司的优秀领导者的智慧。如果你是试图在你的公司实践数据驱动文化的软件工程师、数据科学家或者产品经理，那么这是为你准备的优秀且实用的书。”

——Daniel Tunkelang，Endeca 首席科学家和领英前数据科学与工程总监

“随着每一个领域的数字化和数据驱动化，执行并利用线上对照实验成了必备技能。Kohavi、Tang 与 Xu 为数据从业者和公司主管提供了一个全面且研究充分的必读指导。”

——Evangelos Simoudis，Synapse Partners 联合创始人与执行总裁
The Big Data Opportunity in Our Driverless Future 的作者

“在这本行业内目前最有战略意义的书中，三位作者提供了他们十余年艰苦奋战的实验经验。”

——Colin McFarland，Netflix 实验平台总监

“这本 A/B 测试的实用指南将实验界三位顶尖专家的经验融合成了通俗易懂并易于实践的模块。每章都带你梳理运行实验时最重要的考虑因素——从实验指标的选择到机构的经验传承的重要性。如果你正在寻找一个可以平衡理论与实践的实验导师，那么这本书绝对适合你。”

——Dylan Lewis，Intuit 实验平台负责人

“唯一比没有实验更糟糕的是具有误导性的实验，因为它会给你带来错误的自信！本书根据一些世界上最大的运行实验的机构的见解，详细介绍了实验的相关技术。不管你以任何身份参与线上实验，请立即阅读本书以避免错误并获得对结果的信心。”

——Chris Goward，*You Should Test That*！的作者，
WiderFunnel 的创始人兼首席执行官

“这是一本现象级的图书。作者汲取了丰富的经验，并提供了既易于阅读又全面详尽的参考资料。强烈建议任何想运行严格的线上实验的人阅读。”

——Pete Koomen，Optimizely 的联合创始人

“作者们是线上实验的先驱。他们建立的实验平台以及在平台上运行的实验对许多互联网大品牌是一种革新。他们的研究和演讲启发了整个行业。这是一本业界期待已久的权威且实用的图书。”

——Adil Aijaz，Split Software 的联合创始人兼首席执行官

“ A/B 测试以及数据驱动的决策是互联网与大数据时代基础方法论的重要基石。本书总结了三位互联网行业领导者多年来在实验领域的经验心得，值得每一个想要了解如何用数据驱动决策、加速创新的从业者阅读。”

——连乔，快手副总裁

“互联网时代的产品日新月异，做产品需要有好的增长思维，探索并测试不同的想法来增长用户规模、收入或利润。在线实验是互联网产品测试新想法，并基于数据做科学决策的重要方法。在实践中，我们需要追求大规模、高效率、低成本、低风险、科学做实验的系统能力，但建设这样的实验能力往往需要多管齐下，包括平台、方法论、数据、流程、文化等多方面的建设，其中的挑战和复杂度不容小觑。本书从多个层面，结合近些年的理论研究和硅谷大公司的实战经验，为实验能力的建设和应用提供一些思路和指导，是在线实验领域一本具有标杆意义的参考书。”

——蒋锡茸，腾讯副总经理

“ A/B 实验是数据驱动的核心引擎，今天互联网的大部分业务都非常习惯和依赖于 A/B 实验进行产品和运营决策，很多公司每年进行几万次到几十万次的实验，这些决策直接关系到产品和业务的成败。因此，如何做好 A/B 实验是数据科学的重要问题。本书的作者都是实验行业的领导者，译者同样是在头部互联网企业和实验行业有丰富经验和影响力的数据科学家。非常推荐这本书给数据科学的专业从业者作为必读书籍。对于每个希望能够深刻理解 A/B 实验的朋友，本书同样值得一读。”

——郭飞，腾讯 PCG 公共数据科学部负责人

“本书是 A/B 实验领域非常重要的一本书。它的重要性不仅体现在权威的理论叙述，还因为它集 A/B 实验领域三位领军人物的多年最佳实践于一体。真正的数据驱动的标志之一是通过因果推断对增量价值进行衡量。A/B 实验是因果推断的黄金准则，可信的 A/B 实验是可信的因果推断的前提。任何相信数据驱动的精细化运营的人都应该仔细阅读本书。”

——谢辉志，阿里巴巴数据科学总监

“在互联网经济高速发展的今天，Ron Kohavi、Diane Tang 和 Ya Xu 的书非常及时。它是一本业界科学家写给从业者的书，叙述严谨且非常实用。每个需要运行

线上算法实验的团队，不管是算法工程师还是领导者，都应把此书作为自己文献库的必备参考书籍”。

——秦志伟（Tony），滴滴 AI Labs 首席研究员

“由于互联网时代的用户需求和市场环境快速变化，企业必须采用实验迭代和数据驱动的决策体系和创新模式。可信赖的对照实验（A/B 测试）是其中最核心的一步，也是领英、脸书、谷歌、微软等顶尖企业不断创新和增长的秘密和源泉。Ron、Diane 和 Ya 是业界最具影响力的实验领导者，也是出色的教育者。我曾多次听他们分享并与之交流讨论实验迭代，很高兴他们将最前沿的的经验和研究总结成书。他们结合自己在硅谷最顶尖企业设计实验、驱动产品和改进业务的体会，总结出一套方法、工具和实战案例，这值得所有数据科学家、产品经理和管理者学习和借鉴。”

——孙天澍，南加州大学 Robert Dockson 讲席教授

译者序

从日常运营、产品开发上线到长期目标决策，线上对照实验贯穿于互联网公司的方方面面。在实践中，最基本的问题是实验的可信赖度：如何科学严谨地设计和运行实验，从而保证结果是准确、可信赖的？尤其在机构开始初步建立实验体系并尝试借助实验来做决策的阶段，由于基础设施、数据、流程、方法的原因，实验的质量可能会出问题。同时，对于如何检测实验的质量和如何运行可信赖的实验，大家往往缺乏知识积累和经验指导。无法保证质量的实验结论会误导决策，也会影响机构对基于实验做决策的方法论的信心。

在实践中，我们会尝试摸索并总结实验的技术细节与方法论，但不易察觉的陷阱仍然常常出现。大家一直在期待一本兼顾系统性的方法论与基于实战的经验法则的书籍。本书填补了这一巨大空白，甚至被实验从业者私下称为线上实验领域的“圣经”。本书的三位作者都是美国一线互联网公司的领军人物，他们领导构建了各自公司的实验平台和实验文化，每个平台每年运行上万个实验，对帮助公司进行数据驱动的决策起到了至关重要的作用。本书基于近些年实验领域的研究成果和实践经验，对实验的方法和应用做了很好的全景式描述，包括业界经典的实验案例、实践中的常见陷阱及规避方案、如何建设实验平台和文化、如何设计指标、实验的经验积累、实验的道德规范、实验背后的统计学知识，以及高阶方法的探讨和应用。本书既有助于对实验感兴趣的读者了解实验的应用场景和业务价值，也是实验从业者可以放在手边的参考书。目前，本书在亚马逊的平均评分是 4.7/5，在 Goodreads 的平均评分是 4.5/5。

本书翻译团队的成员分别来自硅谷和国内的一线互联网公司，多位成员参与或主导了各自公司的实验平台和实验文化的建设、研发及应用，并在日常工作和研究中经常与作者交流。作为本书英文版的第一批读者，能够获得翻译本书的机会，我们很荣幸。在阅读和翻译的过程中，我们有了更多的收获和更深的理解。感谢 Ron、Diane 和 Ya 分享他们在实验领域十多年的经验教训！感谢刘锋编辑对翻译工作的支持！希望这本有价值的书可以在中文读者中传播，并使对实验感兴趣的读者和实验从业者从中受益！

以下是翻译团队各成员的介绍：

韩玮：爱彼迎数据科学资深专家，专注于搜索算法和实验领域，之前在沃尔玛实验室负责相关工作。于宾夕法尼亚大学获得应用数学博士学位和统计学硕士学位，本科毕业于中国科学技术大学数学系。

胡鹃娟：现任爱彼迎数据科学家，拥有四年的 A/B 实验分析经验，此前在领英任资深数据科学家。于加州大学戴维斯分校获得统计学硕士学位、香港中文大学获得金融硕士学位，本科毕业于中国科学技术大学 00 班统计专业。

段玮韬：领英资深应用研究专家，现负责领英实验科学团队。他与许亚一起在实验领域紧密合作长达五年之久，书中的很多材料和结论都提炼自他和许亚的工作经验。

胡泽浩：优步数据科学经理，优步人工智能和增长平台两个数据科学团队的负责人，拥有五年用数据及实验驱动产品开发的经验。于宾夕法尼亚大学获得经济学博士学位，本科毕业于香港大学经济系。

廖一正：爱彼迎资深数据科学家，负责爱彼迎中国区搜索引擎算法开发，领导着实验分析委员会。于斯坦福大学获得土木与环境工程博士学位，研究领域为应用机器学习和统计。

王璐：雪花（Snowflake）计算数据科学家，拥有将近七年的实验设计与分析、统计建模以及产品分析经验。曾任爱彼迎数据科学家以及吉利德科学生物统计师。于加州大学洛杉矶分校获得生物统计博士学位，本科毕业于浙江大学

生物信息系。

赵振宇：腾讯数据科学总监。此前先后在雅虎和优步负责实验系统、因果推断、机器学习应用研究和平台建设，以及开源项目研发工作。于美国西北大学获得统计学博士学位，本科毕业于中国科学技术大学。

钟婧：苹果公司 Siri 部门资深数据科学家，此前先后在微软必应部门及脸书公司从事机器学习建模和 A/B 实验、用户和产品数据分析、产品战略分析等方向的研究工作。于密歇根大学获得博士学位，本科毕业于清华大学电子工程系。

前言——如何阅读本书

如果我们有数据，那就看数据。

如果我们只有观点，那就按我的观点来。

——Jim Barksdale，网景前首席执行官

本书旨在分享多年来 Ron 在亚马逊和微软、Diane 在谷歌以及 Ya 在微软和领英大规模运行线上对照实验积累的实践经验。虽然我们不是代表谷歌、领英或微软官方，而是以个人身份写作此书，但书中凝聚了我们工作多年积累的关键经验教训和遇到的常见陷阱，并提供了软件平台的搭建以及公司文化的培养方面的指导：如何利用线上对照实验建立数据驱动文化而不是依赖 HiPPO（Highest Paid Person's Opinion，最高薪者的意见）（R. Kohavi, HiPPO FAQ 2019）。我们相信书中的很多经验适用于各种线上环境，不论是大大小小的公司，还是具体到公司内部的团队或组织。书中强调了评估实验结果可信赖度的必要性。我们相信特威曼定律蕴含的怀疑论：任何看起来有趣或与众不同的数字通常都是错的。我们鼓励读者对实验结果，尤其是有突破性的正面结果做二次检查，以及做验证性测试。获得数据很简单，但获得你能信任的数据很难！

本书第一部分适合所有读者，由四章组成。第 1 章概述运行线上对照实验的好处，并介绍实验相关术语。第 2 章用一个例子剖析运行实验的全过程。第 3 章描述常见的陷阱以及如何建立实验的可信赖度。第 4 章概述如何搭建实验平台并规模化线上实验。

第二部分到第五部分针对一些特定的读者群体，当然也欢迎其他读者按需阅读。第二部分的五章内容介绍实验的基础原理，比如机构指标。我们推荐所有人阅读这一部分，尤其是领导者和高管。第三部分的两章内容介绍线上对照实验的补充技法，可以帮助管理层、数据科学家、工程师、分析师、产品经理等进行资源和时间的投资。第四部分专注于实验平台的搭建，面向工程师群体。最后，第五部分深入讨论进阶的实验分析专题，面向数据科学家。

本书的配套网站为 https://experimentguide.com，它囊括了更多的材料和勘误，并提供了开放性讨论的空间。本书作者的所有收益将捐献给慈善机构。

致谢

我们想感谢这些年来和我们一起并肩作战的同事们。虽然没有办法一一列出名字，但这本书是以和大家共同工作，以及业界和其他领域的同人对线上对照实验的研究和实践为基础的。我们从你们所有人身上学到了很多，感谢你们！

特别感谢编辑 Lauren Cowles 在本书的整个写作过程中与我们紧密协作。Cherie Woodward 提供了逐行编辑和写作风格上的指导，以帮助调和我们三个作者的风格。Stephanie Grey 和我们一起制作了图表，并持续进行改进。Kim Vernon 进行了最终版本的编辑和参考文献的检查。

最重要的是，我们对家人怀有深深的感恩，因为写作本书时错过了很多和他们相处的时间。谢谢 Ron 的家人 Yael、Oren、Ittai 和 Noga；谢谢 Diane 的家人 Ben、Emma 和 Leah；谢谢 Ya 的家人 Thomas、Leray 和 Tavis。没有你们的支持和热情，我们不可能完成本书！

感谢谷歌以下人员：Hal Varian、Dan Russell、Carrie Grimes、Niall Cardin、Deirdre O'Brien、Henning Hohnhold、Mukund Sundararajan、Amir Najmi、Patrick Riley、Eric Tassone、Jen Gennai、Shannon Vallor、Eric Miraglia、David Price、Crystal Dahlen、Tammy Jih Murray、Lanah Donnelly 以及所有在谷歌从事实验工作的伙伴。

感谢领英以下人员：Stephen Lynch、Yav Bojinov、Jiada Liu、Weitao Duan、

Nanyu Chen、Guillaume Saint-Jacques、Elaine Call、Min Liu、Arun Swami、Kiran Prasad、Igor Perisic 以及整个实验平台团队。

感谢微软以下人员：Omar Alonso、Benjamin Arai、Jordan Atlas、Richa Bhayani、Eric Boyd、Johnny Chan、Alex Deng、Andy Drake、Aleksander Fabijan、Brian Frasca、Scott Gude、Somit Gupta、Adam Gustafson、Tommy Guy、Randy Henne、Edward Jezierski、Jing Jin、Dongwoo Kim、Waldo Kuipers、Jonathan Litz、Sophia Liu、Jiannan Lu、Qi Lu、Daniel Miller、Carl Mitchell、Nils Pohlmann、Wen Qin、Thomas Schreiter、Harry Shum、Dan Sommerfield、Garnet Vaz、Toby Walker、Michele Zunker 以及分析和实验平台团队。

特别感谢 Maria Stone 和 Marcus Persson 对整本书的反馈，以及 Michelle N. Meyer 对第 9 章给予的专家意见。

其他对本书提供过意见的包括：Adil Aijaz、Jonas Alves、Alon Amit、Kevin Anderson、Joel Barajas、Houman Bedayat、Beau Bender、Bahador Biglari、Stuart Buck、Jike Chong、Jed Chou、Pavel Dmitriev、Yurong Fan、Georgi Georgiev、Ilias Gerostathopoulos、Matt Gershoff、William Grosso、Aditya Gupta、Rajesh Gupta、Shilpa Gupta、Kris Jack、Jacob Jarnvall、Dave Karow、Slawek Kierner、Pete Koomen、Dylan Lewis、Bryan Liu、David Manheim、Colin McFarland、Tanapol Nearunchron、Dheeraj Ravindranath、Aaditya Ramdas、Andre Richter、Jianhong Shen、Gang Su、Anthony Tang、Lukas Vermeer、Rowel Willems、Yu Yang 和 Yufeng Wang。

感谢很多帮助过我们但没有被提到名字的人们。

目录

本书赞誉
译者序
前言——如何阅读本书
致谢

第一部分　线上对照实验概览

第 1 章　概述和写作动机　003
1.1　线上对照实验的术语　005
1.2　为什么进行实验？相关性、因果关系和可信赖度　008
1.3　有效运行对照实验的必要元素　010
1.4　宗旨　011
1.5　随时间推移的改进　013
1.6　有趣的线上对照实验实例　015
1.7　战略、战术及它们和实验的关系　020
1.8　补充阅读　023

第 2 章　运行和分析实验——一个全程剖析的案例　025
2.1　设立实验　025
2.2　假设检验：确立统计显著性　028
2.3　设计实验　030

2.4 运行实验并获得数据 032
2.5 分析结果 033
2.6 从结果到决策 034

第 3 章 特威曼定律与实验的可信赖度 037
3.1 曲解统计结果 038
3.2 置信区间 041
3.3 对内部有效性的威胁 041
3.4 对外部有效性的威胁 046
3.5 细分群的差异 049
3.6 辛普森悖论 052
3.7 鼓励健康的怀疑态度 054

第 4 章 实验平台和文化 055
4.1 实验成熟度模型 055
4.2 基础设施和工具 062

第二部分 基础原理

第 5 章 速度很重要：一个全程案例剖析 075
5.1 关键假设：局部线性近似 077
5.2 如何测量网站的性能 078
5.3 减速实验的设计 080
5.4 对不同页面元素的影响是不同的 081
5.5 极端结果 083

第 6 章 机构指标 085
6.1 指标的分类 086
6.2 指标的制定：原则和技术 089
6.3 指标的评估 091
6.4 指标的演变 092

6.5 更多的资源 093
6.6 补充材料：护栏指标 093
6.7 补充材料：可操纵性 095

第 7 章 实验指标和综合评估标准 097
7.1 从业务指标到适用于实验的指标 098
7.2 将关键指标组合成一个 OEC 099
7.3 案例：亚马逊电子邮件的 OEC 101
7.4 案例：必应搜索引擎的 OEC 103
7.5 Goodhart 法则、Campbell 法则以及 Lucas 批判 104

第 8 章 机构的经验传承与统合分析 107
8.1 什么是机构的经验传承 107
8.2 为什么机构的经验传承有用 108

第 9 章 对照实验中的伦理 111
9.1 背景 111
9.2 数据收集 116
9.3 文化与流程 117
9.4 补充材料：用户标识符 117

第三部分 补充及替代技法

第 10 章 补充技法 121
10.1 补充技法的空间 121
10.2 基于日志的分析 122
10.3 人工评估 124
10.4 用户体验调研 125
10.5 焦点小组 125
10.6 问卷调查 126
10.7 外部数据 127

10.8 总结 129

第 11 章 观察性因果研究 131

11.1 对照实验不可行的情况 131
11.2 观察性因果研究的设计 133
11.3 陷阱 138
11.4 补充材料：被驳斥的观察性因果研究 141

第四部分 实验平台搭建

第 12 章 客户端实验 145

12.1 服务器端和客户端的差异 145
12.2 对实验的潜在影响 148
12.3 结论 152

第 13 章 工具化日志记录 153

13.1 客户端与服务器端的工具化日志记录 153
13.2 处理多源的日志 155
13.3 工具化日志记录的文化 156

第 14 章 选择随机化单元 157

14.1 随机化单元和分析单元 159
14.2 用户级别的随机化 160

第 15 章 实验放量：权衡速度、质量与风险 163

15.1 什么是放量 163
15.2 SQR 放量框架 164
15.3 四个放量阶段 165
15.4 最终放量之后 168

第 16 章 规模化实验分析 169

16.1 数据处理 169

16.2 数据计算 170
16.3 结果汇总和可视化 172

第五部分 实验分析

第 17 章 线上对照实验中的统计学知识 177
17.1 双样本 t 检验 177
17.2 *p* 值和置信区间 178
17.3 正态性假设 179
17.4 第一 / 二型错误和统计功效 181
17.5 偏差 183
17.6 多重检验 183
17.7 费舍尔统合分析 184

第 18 章 方差估计和提高灵敏度：陷阱及解决方法 185
18.1 常见陷阱 186
18.2 提高灵敏度 189
18.3 其他统计量的方差 190

第 19 章 A/A 测试 193
19.1 为什么运行 A/A 测试 193
19.2 如何运行 A/A 测试 198
19.3 A/A 测试失败时 199

第 20 章 以触发来提高实验灵敏度 201
20.1 触发示例 201
20.2 数值示例 204
20.3 最佳的和保守的触发 205
20.4 总体实验效应 206
20.5 可信赖的触发 207
20.6 常见的陷阱 207

20.7 开放性问题 209

第 21 章 样本比率不匹配与其他可信度相关的护栏指标 211

21.1 样本比率不匹配 212
21.2 调试 SRM 214

第 22 章 实验变体之间的泄露和干扰 219

22.1 示例 220
22.2 一些实际的解决方案 223
22.3 检测和监控干扰 227

第 23 章 测量实验的长期效应 229

23.1 什么是长期效应 229
23.2 短期效应和长期效应可能不同的原因 230
23.3 为什么要测量长期效应 232
23.4 长期运行的实验 233
23.5 长期运行实验的替代方法 235

参考文献 241
索引 261

第一部分

线上对照实验概览

1

第 1 章
概述和写作动机

一个精确的测量胜过一千个专家的意见。

——海军准将格雷斯·霍珀

2012 年，必应（微软的搜索引擎）的一名员工提出了关于改进搜索页广告标题陈列方式的一个想法（Kohavi and Thomke 2017）：将标题下方的第一行文字移至标题同一行，以使标题变长，如图 1.1 所示。

在成百上千的产品建议中，没有人预料到这样一个简单的改动竟然成了必应历史上最成功的实现营收增长的想法！

起初，这个产品建议的优先级很低，被埋没在待办列表中超过半年。直到有一天，一个软件工程师决定试一试这个从编程角度来说非常简单的改动。他实现了该想法，并通过真实的用户反馈来评估它：随机给一部分用户显示新的标题陈列方式，而对另一部分用户依旧显示老版本。用户在网站上的行为，包括广告点击以及产生的营收都被一一记录。这就是一个 A/B 测试的例子：一种简单的用于比较 A 和 B 两组变体的对照实验。A 和 B 也分别称为对照组和实验组。

该测试开始后的几个小时，“营收过高”的警报被触发，提示实验有异常。实验组，也就是新的标题陈列方式，产生了过高的广告营收。这种“好到难以置信”的警报非常有用，它们通常能提示严重的漏洞，如营收被重复上报（双重

计费）或者因网页出错而导致只能看到广告。

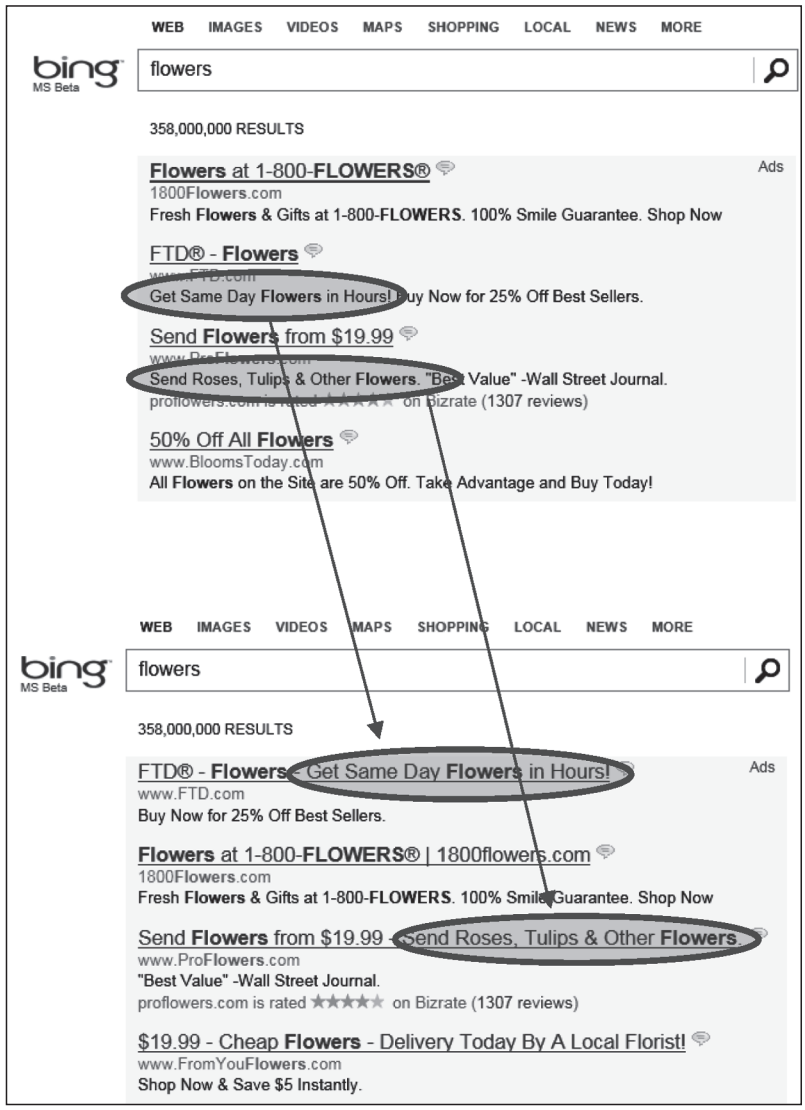


图 1.1　改进必应上广告陈列方式的实验

然而就这个实验而言，营收增长是真实有效的。在没有显著损害其他关键用户体验指标的情况下，必应的营收增长高达 12%，这意味着仅在美国，当年的营收增长就将超过 1 亿美金。这一实验在后来很长一段时间里被多次重复验证。

1
~
3

这个例子体现了关于线上对照实验的几个关键主题：

- 一个想法的价值很难被预估。在这个案例中，一个价值超过每年 1 亿美金的简单的产品改动被耽搁了好几个月。
- 小改动可以有大影响。一个工程师几天的工作就能带来每年 1 亿美金的回报。当然这样极端的投资回报率（return-on-investment, ROI）也很罕见。 4
- 有很大影响的实验是少见的。必应每年运行上万个实验，但这种小改动实现大增长的案例几年才出一个。
- 运行实验的启动成本要低。必应的工程师可以使用微软的实验平台 ExP，来便利地科学评估产品改动。
- 综合评估标准（overall evaluation criterion，OEC，本章将详细描述）必须清晰。在这个案例中，营收是 OEC 的一个关键组成，但仅营收本身不足以成为一个 OEC。以营收为唯一指标可能导致网站满是广告而伤害用户体验。必应使用的 OEC 权衡了营收指标和用户体验指标，包括人均会话数（用户是否放弃使用或者活跃度增加）和其他一些成分。关键宗旨是即使营收大幅增长，用户体验指标也不能显著下降。

接下来的一小节介绍对照实验的术语。

1.1　线上对照实验的术语

对照实验有一段长而有趣的历史，我们的网站有相关分享（Kohavi, Tang and Xu 2019）。 对照实验有时也称为 A/B 测试、A/B/*n* 测试（强调多变体测试）、实地实验、随机对照实验、分拆测试、分桶测试和平行飞行测试。不论实验中有几个变体，我们在本书中都将使用对照实验和 A/B 测试这两个互通的术语。

很多公司广泛使用线上对照实验，例如爱彼迎（Airbnb）、亚马逊（Amazon）、缤客（Booking.com）、易贝（eBay）、脸书（Facebook）、谷歌（Google）、领英（LinkedIn）、来福车（Lyft）、微软（Microsoft）、奈飞（Netflix）、推特（Twitter）、优步（Uber）、Yahoo!/Oath 和 Yandex（Gupta et al. 2019）。这些公司每年运行成千上万个实验，实验有时涉及百万量级的用户，测试内容更是涵盖各个方面，包括用户界面（User Interface, UI）的改动、关联算法（搜索、广告、个性化、

推荐等）、延迟 / 性能、内容管理系统、客户支持系统等。实验可运行于多种平台或渠道：网站、桌面应用程序、移动端应用程序和邮件。

最常见的线上对照实验把用户随机分配到各变体，且这种分配遵循一以贯之的原则（一个多次访问的用户始终会被分配至同一变体）。在开篇必应的例子中，对照组是原本的广告标题陈列方式，实验组是长标题陈列方式。用户在必应网站上的互动被以日志的形式记录，即监测和上报。根据上报的数据计算得到的各项指标可以帮助我们评估两个变体之间的区别。

5

最简单的对照实验有两个变体，如图 1.2 所示：对照组（A）和实验组（B）。

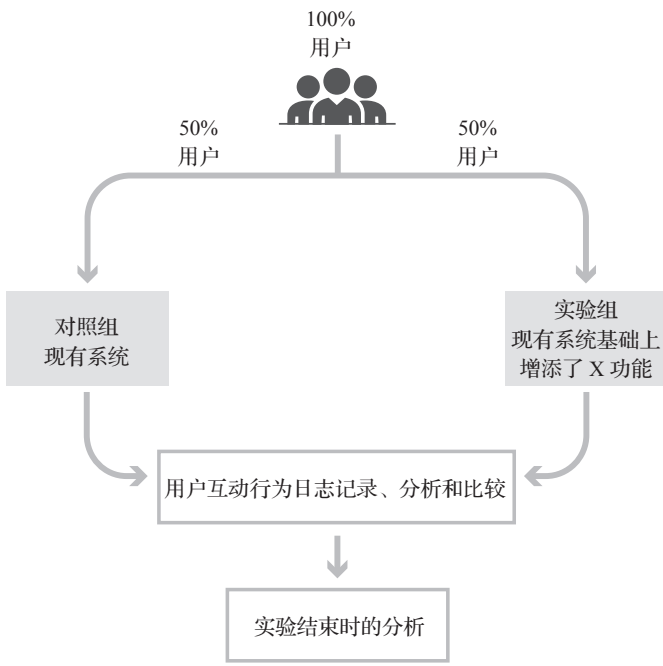


图 1.2　一个简单的对照实验：A/B 测试

以下，我们将遵循 Kohavi 和 Longbottom 等人（Kohavi and Longbottom（2017）, Kohavi, Longbottom et al.（2009））使用的术语进行介绍，并提供其他领域的相关术语。更多关于实验和 A/B 测试的资源可以在本章结尾的补充阅读部分中找到。

综合评估标准（Overall Evaluation Criterion, OEC）：实验目标的定量测量。

例如，你的 OEC 可能是人均活跃天数，指示实验期间用户有几天是活跃的（即有访问并有其他行动）。OEC 的增长意味着用户更频繁地访问了网站，这是好的结果。OEC 需要在短期内（实验期间）可测量，同时要对长期战略目标有因果关系的驱动作用（见 1.7 节，以及第 7 章）。在搜索引擎的例子中，OEC 可以是使用量（如人均会话数）、关联（如成功的会话、成功需时）以及广告营收的综合考量（有些搜索引擎不会用到所有这些指标，有些则会用到更多种类的指标）。

在统计学中，OEC 也常称为*响应变量*或*因变量*（Mason, Gunst and Hess 1989, Box, Hunter and Hunter 2005）。其他的同义词还有*结果*、*评估和适应度函数*（Quarto-vonTivadar 2006）。虽然选择单一指标（可能是一个对多重目标进行加权组合的指标）常常是必须的和高度推荐的（Roy 2001, 50, 405-429），但实验可能有多重目标，且分析也可以采用平衡的分析看板的方法（Kaplan and Norton 1996）。

我们将在第 7 章进一步讨论如何决定一个实验的 OEC。

参数：对照实验中被认为会影响 OEC 或其他我们感兴趣的指标的变量。参数有时也称为*因素*或*变量*。参数的*赋值*也称为*因子水平*。一个简单的 A/B 测试通常只有一个参数，两个赋值。对于线上实验，单变量多赋值（如 A/B/C/D）的设计非常普遍。多变量测试，也称*多元检验*，则可用于同时评估多个参数（变量），比如字体颜色和字体大小。多变量测试可以帮助实验者在参数间有交叉影响时找到全局最优值。

变体：被测试的用户体验，一般通过给参数赋值实现。对于简单的 A/B 测试，A 和 B 就是两个变体，通常被称为对照组和实验组。在某些文献中，变体只指代实验组。而我们把对照组也看作一种特殊的变体——用于进行对比的原始版本。比如，实验中出现漏洞时，你需要中止这个实验，并确保所有用户被分配到对照组这个变体。

随机化单元：以伪随机化（如哈希）过程将单元（如用户或页面）映射至不同变体。正确的随机分配过程非常重要，它可以确保不同变体的群体在统计意义上的相似性，从而高概率地确立因果关系。映射时需遵循一以贯之和独立的原则（即如果以用户为随机化单元，那么同一个用户应该自始至终有一致的体验，并且一个用户被分配到某一变体的信息不会透露任何其他用户的分配信息）。运行线上对照实验时，非常普遍且我们也强烈推荐的是以*用户*为随机化单

元。有些实验设计会选择其他的随机化单元，例如页面、会话或用户日（即同一
7 用户在由服务器决定的每个 24 小时的窗口内体验不变）。详见第 14 章。

正确的随机分配是至关重要的！如果实验设计为各个变体获得相同比例的用户，那么每个用户被分配到任何一个变体的概率应该是一样的。千万不要轻视随机分配。下面的例子解释了正确进行随机分配的挑战和重要性。

- 20 世纪 40 年代，RAND 公司需要为蒙特卡罗方法寻找随机数，为此，他们制作了一份由脉冲机器生成的百万乱数表。然后由于硬件偏移，原表被发现有严重的偏差，导致需要为新版重新生成随机数（RAND 1995）。
- 对照实验起初应用于医药领域。美国退伍军人事务部曾做过一个用于结核的链霉素的药物试验，由于医师在甄选程序中出现了偏差，这一试验最终宣告失败（Mark 1997）。英国有一项类似的试验以盲态程序甄选并获得了成功，成为对照试验领域的分水岭时刻（Doll 1998）。

任何因素都不应影响变体的分配。用户（随机化单元）不能被随意地分配（Weiss 1997）。值得注意的是，随机不代表“随意或无计划，而是一种基于概率的慎重选择”（Mosteller, Gilbert and Mcpeek 1983）。Senn（2012）探讨了更多关于随机分配的迷思。

1.2 为什么进行实验？相关性、因果关系和可信赖度

假设你在一家提供订阅服务的公司（比如奈飞）工作，公司每个月有 *X*% 的用户流失（取消订阅）。你决定引入一个新功能，观察到使用这个新功能的用户的流失率仅为一半：*X*%/2。你可能据此推断出因果关系：该新功能使得流失率减半。由此得出结论：如果我们能让更多的用户发现这一功能并使用它，订阅数将会激增。错了！根据这个数据，我们无法得出该功能降低或增加用户流失率的结论，两个方向皆有可能。

同样提供订阅服务的微软 Office 365 有一个例子表明了这种逻辑的谬误。使用 Office 365 时看到错误信息并遭遇系统崩溃的用户有较低的流失率，但这
8 并不代表 Office 365 应该显示更多的错误信息或者降低代码质量使得系统频繁崩溃。这三个事件都有一个共同的因素：使用率。产品的重度用户看到较多的

错误信息，经历较多的系统崩溃，其流失率也较低。相关性并不意味着因果关系，过度依赖观察结果往往导致做出错误的决策。

1995 年，Guyatt et al.（1995）引入了证据可信度等级来为医学文献做出推荐评级，Greenhalgh 在之后关于循证医学的实践讨论中进一步扩展了这个模型（1997, 2014）。图 1.3 展示了一个翻译成我们的术语的基础版证据可信度等级（Bailar 1983, 1）。随机对照实验是确立因果关系的黄金准则。对随机对照实验的系统性检阅（即统合分析）则有更强的实证性和普适性。

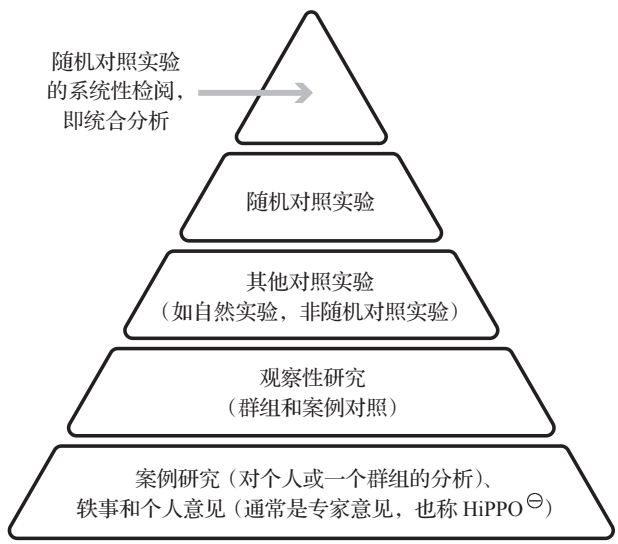


图 1.3　用于评估实验设计质量的证据可信度等级（Greenhalgh 2014）

这一领域还有更多更复杂的模型，比如牛津循证医学中心提出的证据分级（Level of Evidence）（2009）。

谷歌、领英和微软的实验平台每年可以运行成千上万个线上对照实验，并提供可信赖的实验结果。我们相信线上对照实验有以下特性：

- 它是以高概率确立因果关系的最佳科学方法。
- 能够检测其他技术难以检测到的微小变动，比如随时间的变化（灵敏度）。 9
- 能够检测到意想不到的变动。虽然常被低估，但很多实验发掘了一些对

⊖ HiPPO 是 Highest Paid Person's Opinion 的缩写，可以翻译成“最高薪者的意见”。——译者注

其他指标出乎意料的影响，比如性能的降低、系统崩溃和出错的增加或是对其他模块的点击的吞噬。

本书的一个重点是强调实验中可能出现的陷阱，并给出能让实验结果更可信赖的方法。线上对照实验有其独一无二的线上收集大量可靠数据、随机分配和避免或检测陷阱的能力（见第 11 章）。当线上对照实验不可行的时候，我们才推荐使用其他可信度较低的方法，如观察性研究。

1.3 有效运行对照实验的必要元素

科学严谨的对照实验并不能用于所有的决策。比如，你无法在一个投资并购（M&A）场景中运行对照实验，因为我们无法让投资并购和它的虚拟事实（没有该投资并购）同时发生。接下来我们将梳理有效运行对照实验的必要元素（Kohavi，Crook and Longbotham 2009），并提出机构的宗旨。我们会在第 4 章讨论实验成熟度模型。

1）存在可以互不干扰（或干扰很小）地被分配至不同变体的实验单元，比如实验组的用户不会影响对照组的用户（见第 22 章）。

2）有足够的实验单元（如用户）。为了对照实验的有效性，我们推荐实验应包含上千个实验单元：数目越多，能检测到的效应越小。好消息是，即使是小型的软件初创公司通常也能很快地累积足够的用户，从检测较大的效应开始运行实验。随着业务的增长，检测较小变动的能力会变得越来越重要（例如，大型网站必须有能力检测出用户体验关键指标和营收百分比的微小变动），而实验灵敏度也会随着用户基数的增长而提高。

3）关键指标（最好是 OEC）是经过一致同意的，且可以在实践中被评估。如果目标难以测量，那么应对使用的代理指标达成一致（见第 7 章）。可靠的数据最好能以低成本被广泛地收集到。在软件领域，记录系统事件和用户行为通常比较简单（见第 13 章）。

4）改动容易实现。软件的改动一般比硬件的要简单。然而即使是软件的改动，有些领域也需要一定级别的质量控制。推荐算法的改动很容易实现和评估，但美国飞机的飞行控制系统软件的改动则需经过美国联邦航空管理局一整套不同的批准流程。服务器端软件比客户端软件要容易改动得多（见第 12 章），这就

是为什么从客户端软件请求服务越来越普遍，从而使服务的升级和改动可以更快实现并运行对照实验。

大部分复杂的线上服务都有或者可以有这些必要组成部分，来运行基于对照实验的敏捷开发流程。很多“软件 + 服务”的实现也能相对容易地达到要求。Thomke 指出机构可以通过实验与“创新系统”的结合实现利益最大化（Thomke 2003 ）。敏捷软件开发就是这样的创新系统。

对照实验不可行的时候，也可以用建模或其他的实验技术（见第 10 章）。关键是，如果可以运行对照实验，那么它将提供评估改动的最可靠且最灵敏的机制。

1.4　宗旨

对有意运行线上对照实验的机构来说，有三个关键的宗旨（ Kohavi et al. 2013 ）：

1. 该机构想要进行数据驱动决策，且有正式的 OEC。

2. 该机构愿意为运行对照实验投资基础设施和测试，并确保实验结果是可信赖的。

3. 该机构意识到其无法很好地评估想法的价值。

1.4.1　宗旨 1: 该机构想要进行数据驱动决策，且有正式的 OEC

很少听到一个机构的领导层说他们不追求数据驱动（乔布斯领导的苹果公司是个著名的例外，Ken Segall 宣称：“我们不测试任何广告。印刷品、电视、广告牌、网页、零售或者任何形式的广告都不测试。”(Segall 2012, 42))。然而测量新功能给用户带来的额外好处是有成本的，且目标测量通常会显示进展不如预期设想的顺利。很多机构不会将资源用于定义和测量进展。更简单的做法是生成一个计划并实施它，然后宣告成功。关键指标则设立为“计划实施的百分比”，而忽视新功能对关键指标是否有任何正面影响。 11

要做到数据驱动，机构应当定义一个在较短期间内（例如一、两星期内）能够便利测量的 OEC。大型机构也许有多个 OEC 或者多个关键指标，经细化后为不同领域所共用。难点在于找到这样的指标：短期内可测量、足够灵敏以检测

到差别，并且可以以此预测长期目标。例如，“利润”不是一个好的 OEC，一个短期的操作（如提高价格）可以增加短期利润，但可能会降低长期利润。客户终生价值则是战略上很有力的 OEC（Kohavi, Longbottom et al. 2009）。对 OEC 达成一致且以 OEC 来协调整个组织的重要性怎么强调都不过分（见第 6 章）。

有时，人们也用“数据启示”（data-informed）或“数据感知”（data-aware）来避免对仅依赖单一数据来源（如对照实验）“驱动”决策的误解（King, Churchill and Tan 2017, Knapp et al. 2006）。本书将数据驱动和数据启示作为同义词使用。最终，决策的形成应基于多方数据来源，包括对照实验、用户调研、新代码维护成本的预估等。数据驱动或数据启示的机构会收集各种相关的数据来驱动决策，并告知 HiPPO，而不是依靠直觉（Kohavi 2019）。

1.4.2 宗旨 2: 该机构愿意为运行对照实验投资基础设施和测试，并确保实验结果是可信赖的

在通过软件工程师的工作就能满足对照实验运行的必要条件的线上软件领域（网站、移动端、桌面应用和线上服务）：能够可靠地随机分配用户；能够收集数据；能够相对容易地对软件进行改动，比如增加新的功能（见第 4 章）。即便是规模较小的网站也有足够的用户数量以运行必要的统计测试（Kohavi,
12 Crook and Longbotham 2009）。

正如 Eric Ries 在 *The Lean Startup*（Ries 2011）一书中所推崇的，对照实验在与敏捷软件开发（Martin 2008, K. S. Rubin 2012）、客户开发流程（Blank 2005）以及最简可行产品（Minimum Viable Product, MVP）相结合时格外有效。

在某些其他领域，可靠地运行对照实验可能比较困难甚至完全不可行。对照实验所需的某些干预在医药领域可能是不道德的或不合法的。某些硬件设备的生产交付时间可能较长且难以修改，因此极少在新的硬件设备（例如新手机）上运行基于用户的对照实验。在这些情况下，我们可以用其他的技术，比如*补充技法*（见第 10 章）。

假如可以运行对照实验，那么就要确保实验结果的可信赖性。从实验中获得数据是简单的，但是获得可信赖的数据是困难的。第 3 章将致力于介绍如何获得可信赖的实验结果。

1.4.3　宗旨 3: 该机构意识到其无法很好地评估想法的价值

团队认为一个产品功能有用才会去开发它，然而，很多领域的大部分想法都无法提升关键指标。在微软，被测试过的想法里仅有三分之一能改善目标指标（Kohavi, Crook and Longbotham 2009）。在已经高度优化的产品领域，如必应和谷歌，想要成功就更难了，测试的成功率仅在 10% ～ 20% 之间（Manzi 2012）。

Slack 的产品生命周期总监 Fareed Mosavat 曾发推特称，在 Slack 只有 30% 的商业化实验显示正向结果。“如果在一个实验驱动的团队工作，你得习惯这一点，你的工作至少有 70% 可能会被弃置。要据此建立你的工作流程。”（Mosavat 2019）

Avinash Kaushik 在“ Experimentation and Testing primer”（Kaushik 2006）一文中写道：“80% 的时间我们都搞错了客户需要什么。”Mike Moran（Moran 2007, 240）曾称奈飞认为他们尝试过的东西有 90% 是错的。Quicken Loans 的 Regis Hadiaris 称：“在我做实验的五年里，我猜对结果的正确率和联盟棒球手击中球的概率差不多。是的，我做了五年，只能‘猜’对 33% 的实验结果！”（Moran 2008）。Etsy 的 Dan McKinley（McKinley 2013）称“几乎所有都是失败的”，而对产品功能，他说：“我们谦虚地认识到初次尝试就成功的功能是很少见的。我强烈认为这一经验是普遍的，只是没有被普遍地意识到或承认。”最后，Colin McFarland 在 *Experiment!*（McFarland 2012, 20）一书中写道：“无论你认为多么不费吹灰之力，做了多少研究，或者有多少竞争者正在做，有时候想法就是会失败，并且失败的频率比你预期的更高。”

13

不是所有领域的数据都是这么差，但是大多数运行对照实验的面向客户的网站和应用程序都认识到这个无法改变的现实：我们不擅长预判想法的价值。

1.5　随时间推移的改进

在实践中，关键指标的改进是由很多 0.1% ～ 2% 的小改动积累而来的。很多实验仅影响一部分用户，所以你需要将一个作用于 10% 用户的 5% 的影响稀释，从而得到一个小很多的影响（如果实验的触发群体和其他用户类似，那么

稀释后的影响就是 0.5%），详情见第 3 章。正如阿尔·帕西诺在电影《挑战星期天》里所说："……成功是一寸一寸实现的。"

1.5.1 谷歌广告的例子

在经历超过一年的开发和一系列的实验后，谷歌在 2011 年发布了一套经过改进的广告排序系统（Google 2011）。工程师们在已有的广告排序系统里开发并测试了新改进的模型，以测量广告的质量分数，也开发并测试了对广告竞价系统的改动。他们开展了上百个对照实验和多次迭代，有些实验横跨多个市场，有些则在某一市场长期运行，以深度了解广告商受到的影响。这一大型的后端改动——以及运行对照实验——最终验证了规划和叠加多个改动可以通过提供更高质量的广告改善用户体验，并且高质量广告的更低价格也提升了广告商的体验。

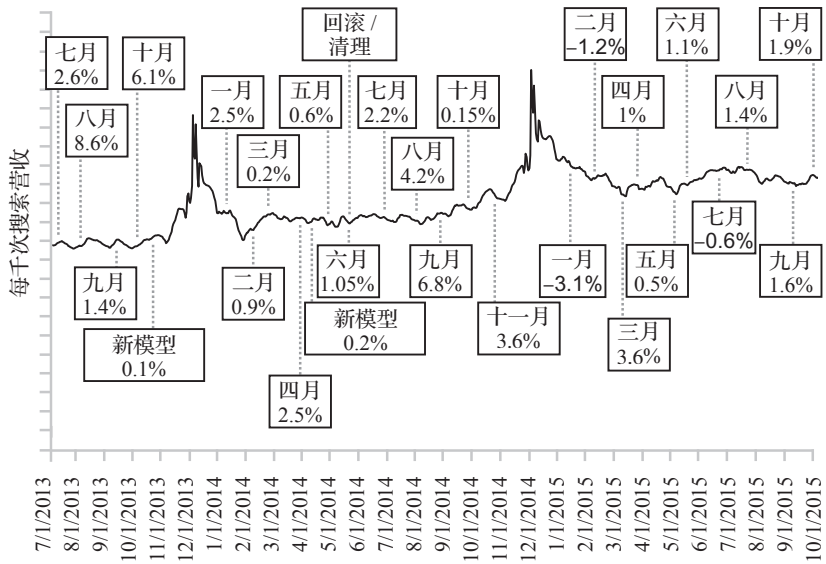
1.5.2 必应关联算法的例子

必应几百人的关联算法团队每年的任务就是将单个 OEC 指标提高 2%。这
14 2% 由当年发布给用户的所有对照实验的实验效应（即 OEC 的增量）相加而成（假设可简单相加）。由于该团队每年运行上千个实验，有些实验会因为偶然性出现正向结果（Lee and Shen 2018），2% 的贡献取决于一种重复实验：一旦某个想法的实现在经过多次迭代和优化之后成功了，一个单对照组的认证实验就会开始运行。这个认证实验的实验效应决定了对 2% 这一目标的贡献。最近的研究建议缩减实验效应以提高精确度（Coey and Cunningham 2019）。

1.5.3 必应广告的例子

必应的广告团队每年持续地给营收带来 15% ～ 25% 的增长（eMarketer 2016），但是大部分的提高都是一点一点完成的。每个月会发布一个打包了多项实验结果的"包裹"，如图 1.4 所示。多数都是小的改进，有些甚至因为空间限
15 制或法律法规要求而给营收带来负增长。

十二月左右的季节性激增很有信息量。在十二月，随着用户的购物意图显著提高，广告空间也相应增加，搜索带来的营收也随之提高。



（*）显然，数字是经过扰动处理的

图 1.4　必应广告营收随时间的变化（y 轴代表每年增长大约 20%），具体数字不重要

1.6　有趣的线上对照实验实例

有趣的实验是指那些预期结果和实际结果相差很远的实验。如果你认为一件事情会发生，然后它真的发生了，那么你不会学到什么。如果你认为一件事情会发生，但没有发生，那么你会学到一些重要的东西。如果你原本认为一件事情不值一提，但它带来了惊人的或突破性的结果，那么你会学到一些非常有价值的东西。

本章开头必应的例子和本小节介绍的其他实例都有不同寻常且出乎预料的正面结果。必应和社交网络（如脸书和推特）相结合的尝试，是一个没有达到预期效果的例子——这一尝试在历经了长达两年的多个结果显示无价值的实验后告弃。

虽然经久的进步来自持续的实验和很多小的改进（如必应广告的例子所示），但这里介绍的几个有惊人效果的例子表明了我们有多不擅长预估想法的价值。

1.6.1 UI 实例：41 阶蓝

谷歌和微软的很多例子都一致显示：小的界面设计决策也可能有重大的影响。谷歌在搜索结果页面上测试过 41 个阶度的蓝色（Holson 2009），这让当时的视觉设计负责人很受挫。然而，谷歌对配色方案的调整给用户活跃度带来了实质性的提高（谷歌没有报告单一改动的结果），并促成了之后设计团队和实验团队的高度协作关系。微软的必应也做过类似的配色改动，帮助用户更成功地完成任务，改善了任务成功需时，将美国市场的年营收提高了超过 1 千万美金（Kohavi et al. 2014, Kohavi and Thomke 2017）。

这些都是微小改动带来巨大影响的很好的例子，但由于配色方案已经被广泛地测试，在更多实验中“玩”配色已不太可能带来更加显著的改进。

16

1.6.2 在正确的时机显示推广

2004 年，亚马逊在主页上放置了信用卡推广，该推广带来了可观的利润，但是点击率（Click-Through Rate, CTR）很低。团队进而运行了把该推广移至购物车页面的实验，用户把商品加入购物车后，可以看到一个简单的计算，从而对该信用卡带来的优惠额度一目了然，如图 1.5 所示（Kohavi et al. 2014）。

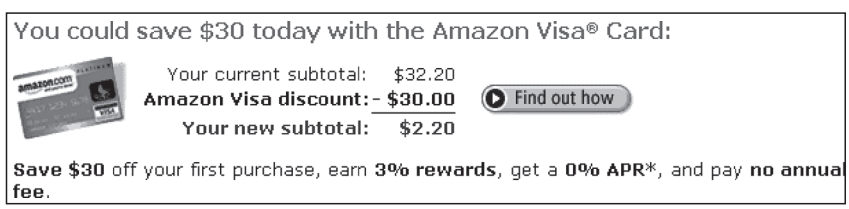


图 1.5　亚马逊购物车的信用卡推广

因为已经在购物车添加了商品的用户有非常明确的购物意图，所以这一推广出现在了正确的时机。对照实验显示这一简单的改动帮助亚马逊增加了数千万美金的年利润。

1.6.3 个性化推荐

亚马逊的 Greg Linden 做过一个产品雏形，根据用户购物车里的商品显示个性化的推荐（Linden 2006, Kohavi, Longbottom et al. 2009）。当你添加一件商

品时，会看到一些推荐，再添加另一件商品，会看到新的推荐。Grey Linden 指出，虽然这个产品雏形看起来颇有前景，但一个市场营销高级副总裁极力反对，称它会分散用户注意力而影响结账。Greg Linden 被禁止继续这项工作。尽管如此，他还是运行了一个对照实验，并且这一新功能大幅获胜，放弃它将使亚马逊蒙受可观的损失。带着新的紧迫性，购物车推荐很快被发布了。如今，很多网站都在使用购物车推荐模型。

1.6.4　速度非常关键

2012 年，微软必应的一个工程师对 JavaScript 的生成方式做了改动，大大缩短了 HTML 到达客户端的时间，从而提高了性能。相应的对照实验显示了多项指标令人惊叹的改进。他们接着做了跟进实验来测量对服务器性能的影响。 17
结果显示性能提高的同时也显著提高了关键用户指标，例如成功率和首任务成功需时，且每 10 毫秒的性能提升（眨眼速度的 1/30）足以支付雇佣一个全职工程师一年的成本（Kohavi et al. 2013）。

2015 年，随着必应性能的提高，人们产生了疑问：服务器结果返回时间的 95 百分位数（即 95% 的搜索请求）已经在 1 秒以下，进一步提高性能是否仍有价值。必应的团队开展了跟进研究，发现关键用户指标仍旧得到了显著提高。虽然对营收的相对影响有一定程度的减弱，但由于必应的营收在那几年增长如此之多，以至于性能上每 1 毫秒的提升都比过去更有价值，每 4 毫秒的改进就能雇佣一个工程师一年！我们将在第 5 章深度剖析相关实验以及性能的重要性。

许多公司都做过性能实验，结果都显示性能多么关键。在亚马逊，一个 100 毫秒的减速实验使销售额减少了 1%（Linden 2006b, 10）。必应和谷歌的一个联合讲演（Schurman and Brutlag 2009）展示了性能对关键指标的显著影响，包括去重搜索词条、营收、点击率、用户满意度和首点击需时。

1.6.5　减少恶意软件

广告是一项有利可图的生意，用户安装的“免费软件”经常包含插入广告污染页面的恶意软件。图 1.6 显示了一个含有恶意软件的必应搜索结果页面。注意有多个广告（用方框圈出）被插入这个页面（Kohavi et al. 2014）。

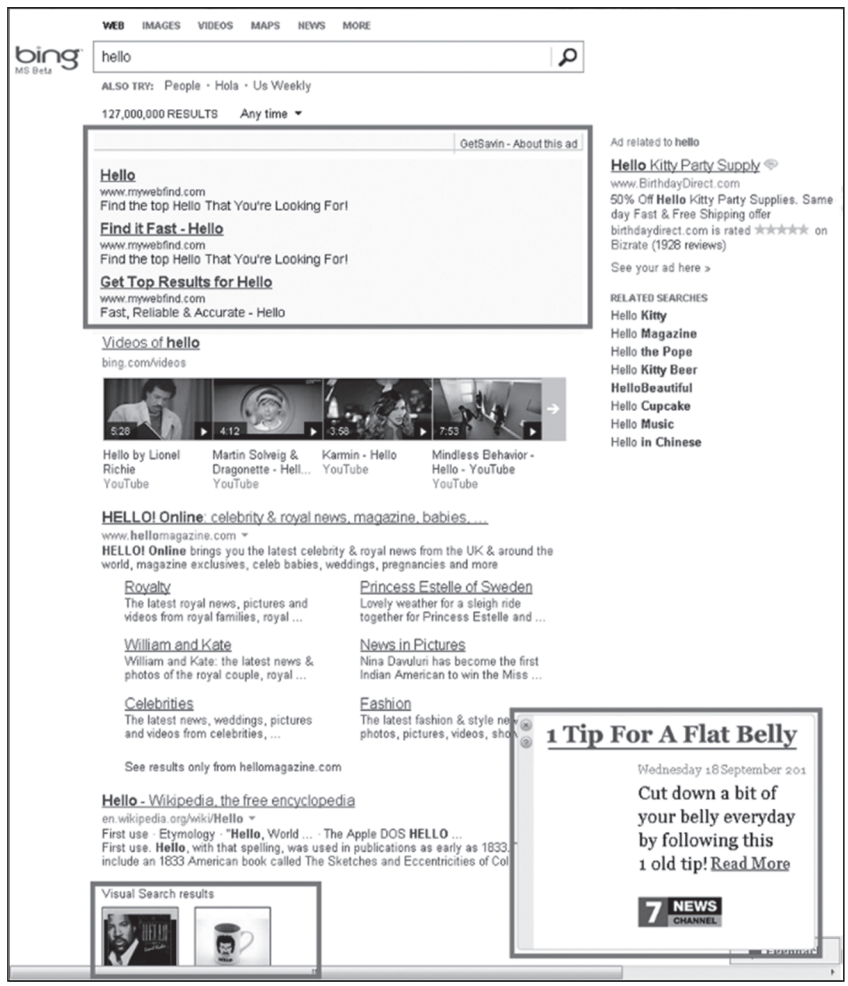


图 1.6　含恶意软件的必应页面显示了多个广告

这些不相关的低质量广告不仅移除了必应自己的广告，从而侵占了微软的营收，也给用户带来了糟糕的体验，用户可能都没有意识到为什么他们会看到这么多广告。

微软对 380 万潜在受影响的用户运行了一个对照实验：修改文档对象模型（Document Object Model, DOM）的基本路径被覆写，只允许少数可靠来源的修改（Kohavi et al. 2014）。结果显示必应所有的关键指标（包括人均会话数）都有所提高，这意味着用户访问的增加或用户流失的减少。除此之外，用户的搜索

也更加成功，能更快地点击到有用的链接，必应的年营收也增加了数百万美金。前文讨论过的关键性能指标，比如页面加载需时，在受影响页面上也提升了几百毫秒。 18

1.6.6　后端改动

后端算法的改动是在运用对照实验时常常被忽视的领域（Kohavi, Longbottom et al. 2009），但它可能带来重大影响。我们在前文中描述过谷歌、领英和微软的团队如何一点点做改动，这里我们讨论一个来自亚马逊的实例。

回到 2004 年，当时亚马逊已经有了基于两个数据集的很好的推荐算法。其 19
标志性功能本来是“买了 *X* 的用户也买了 *Y*”，但后来被延伸为“**浏览**了 *X* 的用户也买了 *Y*”和“浏览了 *X* 的用户也**浏览**了 *Y*”。有人提出了一个方案，使用同样的算法推荐“**搜索**了 *X* 的用户也买了 *Y*”。这个算法的支持者给出了含义不明的搜索的例子，比如“24”，多数人会联想到 Kiefer Sutherland 主演的电视剧。亚马逊的原算法对“24”返回的结果比较糟糕（图 1.7 左），有会 24 首意大利歌曲的 CD、24 月龄婴儿穿的衣服、24 英寸的毛巾杆等。而新的算法表现比较出色（图 1.7 右），根据用户搜索“24”后实际购买的项，返回了相关电视剧的 DVD

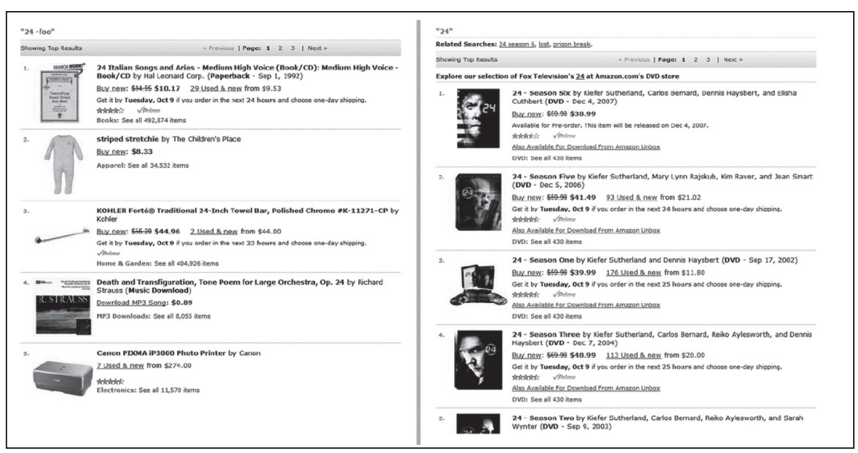


图 1.7　有 BBS 和没有 BBS 时在亚马逊上搜索“24”的结果⊖

⊖ BBS 是“基于用户行为的搜索”（Behavior-Based Search）的缩写。在搜索词中加入“-foo”，亚马逊显示的是没有 BBS 的结果。——译者注

和书籍的搜索结果。该算法的一个不足是返回的某些商品并没有包含搜索关键词。亚马逊运行了一个对照实验，即使有上述不足，这一改动也将亚马逊的营收提高了 3%——数百万美金。

1.7 战略、战术及它们和实验的关系

当运行线上对照实验的必要元素都具备时，我们坚信应该用它来启示机构从战略到战术的各个级别的决策。

20 战略（Porter 1996, 1998）和对照实验有协同作用。David Collis 在 *Lean Strategy* 中写道："与其压制创业型行为，不如用有效的策略鼓励它——通过在确定好的边界范围内鼓励创新和实验。"（Collis 2016）他定义了一套精益战略流程（lean strategy process），通过严谨的规划和不受限的实验来看守边界。

精心运作的实验结合适当的指标可以帮助机构变得更加数据驱动，完备业务战略和产品设计，提高运营效率。通过将战略融入 OEC，对照实验可以为战略提供极佳的反馈回路。经实验评估的产品创意是否充分地改进了 OEC？或者，出乎意料的实验结果可能启示了其他的机会，从而改变战略方向（Ries 2011）。产品设计决策对同调性很重要，尝试多个设计版本可以给设计师提供有用的反馈回路。最后，很多战术的改变能够提高运营效率，如 Porter 所定义，"执行与竞争对手类似的活动，但比他们执行得更好"（Porter 1996）。

我们现在来看两种关键情境。

1.7.1 情境 1: 你已有业务战略并且产品有足够的用户用于实验

对于这种情境，实验可以帮助你从现有的战略和产品出发，爬升至局部最优点。

- 实验能帮忙找到高 ROI 的领域：那些相对于成本提升 OEC 最多的领域。在投入大量资源之前，在不同的领域测试 MVP 有助于快速探索广阔领域。
- 实验能在对设计师来说不明显但可能有重要影响的优化方面提供帮助（例如颜色、间距和性能）。
- 实验能帮助网站持续迭代至更好的设计，而不是让团队直接重新设计整个网站。用户的初始效应（用户习惯了旧的功能，即习惯了现在的网站）

常常使网站的重新设计无法成功达到预期目标，甚至无法在关键指标上达到现在网站的水平（Goward 2015, slides 22 ～ 24, Rawat 2018, Wolf 2018, Laja 2019）。

- 实验在优化后端算法（比如推荐和排序算法）和基础设施方面有关键作用。

有既定的业务战略对运行实验非常关键：这个战略决定了 OEC 的选择。一旦确定好，对照实验通过助力团队优化和改进 OEC 来加速创新。我们见过一些因选择了不恰当的 OEC 而误用实验的例子。选择的指标应当具备一些关键特征并且具有不可操纵性（见第 7 章）。 21

在我们的公司里，不仅有团队专注于如何正确地运行实验，也有团队专注于指标：选择指标、验证指标并随时间演化指标。指标的演化既发生在战略随时间变化后，也发生在认识到现有指标的不足之处后，例如 CTR 因可操纵而需要演化。由于实验的运行时间通常很短，负责指标的团队也要确定哪些指标在短期内可测量的同时能驱动长期目标。Hauser and Katz（1998）写道，“公司必须确立这样的指标：短期内可以被影响，并最终影响公司长期目标”（见第 7 章）。

将战略和 OEC 绑定也创造了战略廉正（Strategic Integrity）（Sinofsky and Iansiti 2009）。作者指出：“战略廉正不是指制定出杰出的战略或拥有完美的组织，它是关于机构统一地执行正确的战略且清楚如何执行。”OEC 是可以使战略变得明晰且明确什么功能会伴随这个战略上线的完美机制。

没有好的 OEC，你最终会浪费资源。想象一下在一艘正在下沉的邮轮上进行改善食物或照明的实验。乘客的安全应该在这些实验的 OEC 里占极高的权重——实际上，权重高到我们一点都不愿意降级。这一点可以通过 OEC 里的高权重来体现，或者把乘客的安全视为护栏指标（见第 21 章）。类比软件行业，游轮乘客的安全相当于软件的崩溃：如果一个新功能增加了产品的崩溃次数，该功能会被视为非常糟糕的体验，相比之下其他考虑因素都很苍白。

定义实验的护栏指标对于明确机构不愿意改变什么很重要，因为一个战略也“要求你在权衡取舍时——选择不要做什么”（Porter 1996）。1972 年，美国东方航空的 401 航班不幸坠毁的原因是，飞行员的注意力集中在出故障的起落架指示灯上，而没有注意到飞机偶然脱离了自动巡航，飞行高度这一关键的护栏指标逐渐降低，飞机最终坠毁在佛罗里达大沼泽地，导致 101 人遇难（Wikipedia contributors, Eastern Air Lines Flight 401 2019）。

运营效率的提升可以打造长期的差异化优势，如 Porter 在 *Japanese Companies*
22 *Rarely have Strategies*（1996）和 Varian 在他关于持续改善的文章（2007）中指出的。

1.7.2 情境 2: 你有产品和战略，但结果显示你需要考虑调整

对于情境 1，对照实验是帮助“爬坡”的好工具。想象一个多维度的想法空间，如果正在被优化的 OEC 是“高度”，那么你也许可以逐步靠近山顶。然而有时候，基于内部变化率的数据或者外部增长率的数据或是其他一些基准，你需要考虑做出调整：跳到这个空间里的另一处，也许是一座更高的山峰，或者改变战略和 OEC（从而改变地势）。

一般来说，我们推荐构建一个想法的组合：大部分想法应该是在当前位置“附近”尝试局部优化，但应当有一小部分是试图跨越到另一座更高山峰的大胆想法。我们的经验是多数大跨越的尝试都会失败（如大型网站的重新设计），然而这里有一个风险 / 回报的权衡：极少数的成功可能有极大的回报，能够补偿多数的失败。

测试大胆的想法时，运行和评估实验的方式也有所变化。具体来说，你需要考虑：

- 实验的运行时长。例如，测试重大的 UI 设计时，短期内测量到的实验变化可能被初始效应或改版厌恶所影响。实验组和对照组的直接比较可能无法测量真实的长期效应。对于双边市场，除非是非常大的改动，否则测试一个改动可能无法在市场上诱发效应。一个类比是低温房间里的冰块：房间温度的小幅提升不会引起变化，而一旦达到熔点（如 32 华氏度[⊖]），冰块就会开始融化。更长时间或更大规模的实验，或者别样的实验设计（比如前文中谷歌广告质量的例子是以国家为单位的实验设计），在这样的情境中也许是必需的（见第 23 章）。
- 被测试的创意的数量。你也许需要很多不同的实验，因为每个实验只测试一个具体的战术，该战术是总体战略的一个组成部分。个别实验没能成功改进 OEC 可能是因为这个特定战术的失败，不代表总体战略有问
23 题。实验通过设计来测试特定的假设，然而战略是更宽泛的。即便如此，

⊖ 32 华氏度等于 0 摄氏度。——编辑注

对照实验可以帮助精细化战略，或者彰显其无效性以鼓励战略调整（Ries 2011）。如果经对照实验评估后的很多战术都以失败告终，那么也许是时候想想温斯顿·丘吉尔的话了："无论战略制定得多漂亮，你都应该时不时看看结果。"在大约两年的时间里，必应有一个战略是和社交媒体（特别是脸书和推特）相结合：展开第三个面板显示社交搜索结果。在花费了超过两千五百万美金但仍然没有显著提升关键指标后，这一战略最终被放弃（Kohavi and Thomke 2017）。放弃大的赌注是很难的，但是经济学理论告诉我们，已经失败的赌注是沉没成本，我们应该基于从更多实验中收集的可用数据来做前瞻性的决策。

Eric Ries 用"成功实现的失败"来描述公司成功、忠实并严谨地执行一个计划，结果发现这个计划是极其有缺陷的（Ries 2011）。他建议：

> "精益创业方法将初创公司的努力设想为评估战略的实验，看看哪部分是优秀的，哪部分是疯狂的。真正的实验遵循科学的方法。它开始于一个清晰的假设并预测会发生什么。然后经验性地测试这些预测。"

因为用实验来评估战略耗时且有挑战性，如 Sinofsky and Iansiti（2009）写道：

> "……产品开发过程充满风险和不确定性。这是两个完全不同的概念…… 我们不能减少不确定性——你不知道的就是不知道。"

我们不同意：运行对照实验能够通过尝试最简可行产品大大降低不确定性（Ries 2011），获得数据并进行迭代。即便如此，不是所有人都有几年的时间来投资测试一个新战略，如果是这种情况，那么你可能需要面对不确定性做决策。

一个值得记住的有用的概念是 Douglas Hubbard（2014）提出的"信息的预期价值"（Expected Value of Information, EVI），它体现了额外的信息如何帮助你做决策。

1.8　补充阅读

与线上实验和 A/B 测试直接相关的书有好几本（Siroker and Koomen 2013, Goward 2012, Schrage 2014, McFarland 2012, King et al. 2017）。它们大多数有好的励志故事，但在统计方面并不准确。Georgi Georgiev 最近的一本书包含了

24 全面的统计解释（Georgiev 2019）。

关于对照实验的文献有很多（Mason et al. 1989, Box et al. 2005, Keppel, Saufley and Tokunaga 1992, Rossi, Lipsey and Freeman 2004, Imbens and Rubin 2015, Pearl 2009, Angrist and Pischke 2014, Gerber and Green 2012）。

关于运行线上对照实验的入门书也有一些（Peterson 2004, 76-78, Eisenberg 2005, 283-286, Chatham, Temkin and Amato 2004, Eisenberg 2005, Eisenberg 2004）;（Peterson 2005, 248-253, Tyler and Ledford 2006, 213-219, Sterne 2002, 116-119, Kaushik 2006）。

多臂老虎机是一种随实验进程动态分配流量的实验（Li et al. 2010, Scott 2010）。比如，我们可以每小时查看一次实验进程，观察每个变体的表现，并调整各变体的流量比例。增加表现好的变体的流量，同时减少表现差的变体的流量。

基于多臂老虎机的实验往往比“经典”A/B 实验更有效率，因为它们将流量逐步地分配至表现更好的变体，而不用等到实验结束。虽然这类实验适用于解决一大类问题（Bakshy, Balandal and Kashin 2019），但也有一些重要的限制条件：评估目标需要是单一的 OEC（例如，多个指标的权衡可以简单地公式化为单一指标），并且这个 OEC 在流量重新分配时能够被很好地测量，比如点击率相对于会话。此外，将原本处于“差”变体中的用户按不同比例分配到其他表现较好的多个变体可能会造成潜在的偏差。

2018 年 12 月，本书的三位作者组织了第一届实用线上对照实验峰会。来自包括爱彼迎、亚马逊、缤客、脸书、谷歌、领英、来福车、微软、奈飞、推特、优步、Yandex 和斯坦福大学在内的 13 个机构的 34 个专家在峰会的分组讨论中介绍了实验的概况和挑战（Gupta et al. 2019）。对实验中所面临的挑战感兴
25 趣的读者可以从该峰会的相关文章中获益。