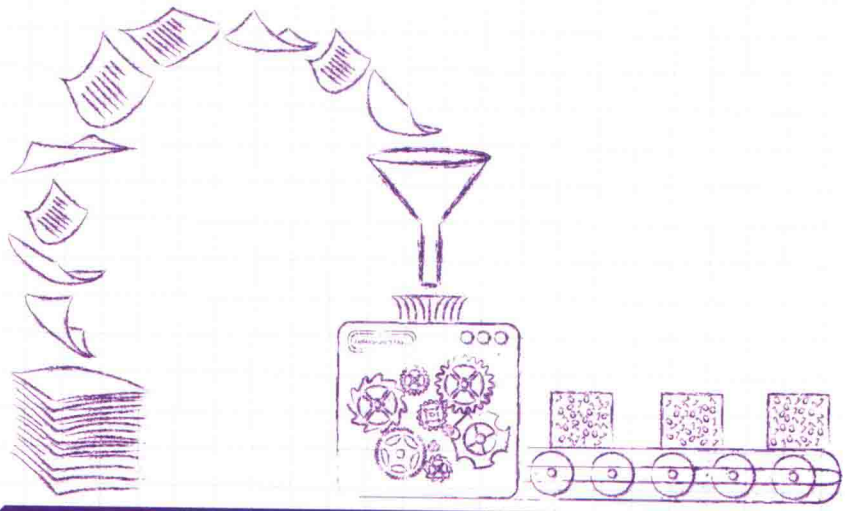
大数据时代的统计学思维

让你从众多数据中找到真相

刘强 编著



大数据时代，各种资源林林总总，你是否感觉到眼花缭乱？
劳心费力找到的信息，是否让你一叶障目，不见泰山？

- **经典** 精选经典、常用的10多种分析方法
- **易懂** 书中的30多个实例，全部源自生活
- **好学** 尽量避开公式，用常见实例辅助学习
- **有用** 带你避开数字陷阱，直达事实真相

中国水利水电出版社
www.waterpub.com.cn

大数据时代的统计学思维

让你从众多数据中发现规律

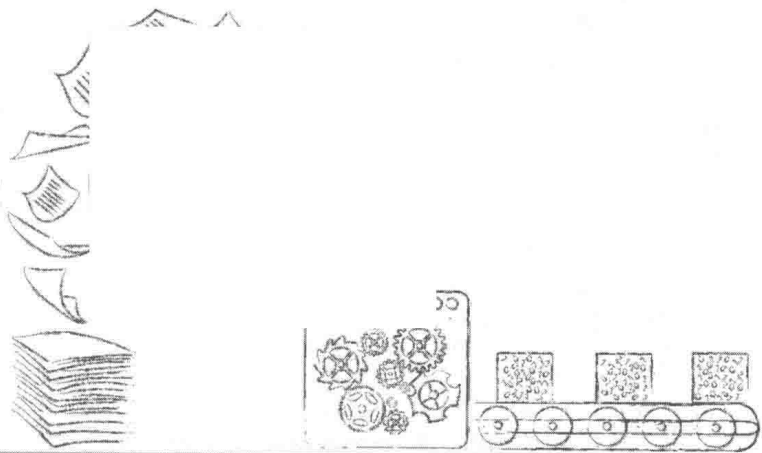
[redacted] 编著

[redacted]

大数据时代的统计学思维

让你从众多数据中找到真相

刘强 编著



中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

大数据时代，对数据进行统计、分析和学习变得尤为重要，并被应用在各方各面，如无人驾驶汽车、AlphaGo、机器学习和人工智能等，而统计思维也成为大数据时代的基本思维。不仅这些高科技以统计学为基础理论，大数据时代的每个人都应该懂点统计学，学会读懂并分析数据，学会让数据说话，让数据为自己服务。《大数据时代的统计学思维：让你从众多数据中找到真相》就是大数据时代统计学思维的科普书籍，全书共 10 章，第 1 章用几个有趣好玩的例子引导读者进入统计学的世界，并调动读者学习统计学的兴趣。第 2~10 章结合生活和工作中的例子全面介绍统计学原理和方法，涵盖统计学中的数据收集、数据处理和统计推断等内容，既有抽样调查、概率、相关性分析、回归分析等实用统计方法，也有大数定律和中心极限定理等基本统计学原理。用实例引导理论，通俗易懂，不知不觉中将统计思想和统计学知识传输给读者。

《大数据时代的统计学思维：让你从众多数据中找到真相》在科普统计学方法和原理的同时，又保持了实用性和趣味性，确保读者能有所收获，感受统计学在大数据时代的魅力。

《大数据时代的统计学思维：让你从众多数据中找到真相》是一本统计学和统计思维科普书籍，适合统计学入门读者、对大数据感兴趣的读者以及任何想学习统计方法的读者学习和参考，也适合大数据时代下每一位不想与时代脱节、想从众多数据中获取真相的读者学习。

图书在版编目（CIP）数据

大数据时代的统计学思维 ：让你从众多数据中找到真相 / 刘强编著. -- 北京 ：中国水利水电出版社，2018.5

ISBN 978-7-5170-6239-4

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①统计学－普及读物 Ⅳ. ①C8-49

中国版本图书馆CIP数据核字(2017)第327473号

书　　名	大数据时代的统计学思维：让你从众多数据中找到真相 DASHUJU SHIDAI DE TONGJIXUE SIWEI:RANG NI CONG ZHONGDUO SHUJU ZHONG ZHAODAO ZHENXIANG
作　　者	刘强　编著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座　100038） 网址：www.waterpub.com.cn E-mail：zhiboshangshu@163.com 电话：（010）62572966-2205/2266/2201（营销中心）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	北京智博尚书文化传媒有限公司
印　　刷	三河市龙大印装有限公司
规　　格	170mm×230mm　16 开本　13.75 印张　161 千字
版　　次	2018 年 5 月第 1 版　2018 年 5 月第 1 次印刷
印　　数	0001—5000 册
定　　价	69.80 元

前言

Preface

统计学，其实是一门五味杂陈的学科。因为它既有自己的独立性，可以自成体系；又能够和其他学科紧密联系，从而形成新的交叉学科，如社会统计学、经济统计学和人口统计学等。在大数据时代的潮流下，人们对于数据越来越重视，研究和分析数据的理论和方法越来越丰富，统计学越来越发挥出它的作用，在各个领域都体现出应用价值。无论是人工智能技术下的超级 AI，还是海量数据中的深度学习技术，或者是以假乱真的虚拟现实，统计学都是它们赖以存在的理论基础。

大数据时代，大家都在谈论“大数据”。可是，到底大数据时代的本质是什么？大数据时代必须掌握哪些知识？许多人并没有一个清醒的认识。在大数据时代下，一切都离不开数据，而所有数据都离不开统计学。在统计学的作用下，大数据发挥出了它巨大的威力，是统计学让数据有了实实在在的说服力。

如今，似乎我们一切的衣食住行都是在用数据说话，无论到哪里都离不开数据的支持。例如，地图上的导航系统，就是实时分析路面数据得到的，数据分析需要统计学；购物网站上的商品推荐，就是由推荐系统根据我们的购物记录得到的，商品推荐需要统计学；医院诊断某一疾病的病因，就是统计分析各种指标得到的，诊断疾病也需要统计学……这样的例子还有许多，无一不说明了统计学在现代社会的作用和价值。

然而，大多数人在享受着地图导航、商品推荐、疾病诊断这些便利，身处于统计学的应用海洋之中却浑然不知。当然，对于大多数普通人来说，不可能都去学习新时代下那些最时髦的数据分析技术，更不可能人人都成为统计学方面的专家。但是，对于每一个人来说，有必要了解一下统计学的基础理论，理解一些统计学的基本原理，能够用统计学的基本方法来分析简单的数据问题，这些都不是一件困难的事，也不是一件令人乏味的事。这么做不为别的，只是为了跟上大数据时代的思想潮流，让统计学的思想深入人心。

统计思想，已然成为大数据时代不可或缺的基本思想，从统计学的热度就可见一斑。越来越多的学校开设了统计学专业，越来越多的家长建议自己的孩子学习统计学，越来越多的书籍、文章在介绍统计学的知识等，这些都体现了统计学的重要性和受欢迎的程度。这些并不是没有道理的？我们可以看到，如果你是一位营销人员，要是你不懂统计学，不会对数据进行分析，那么，你就不可能取得好的营销效果；如果你是一位产品经理，要是你不懂统计学，不会处理用户数据，那么，你就不可能对产品进行改进；如果你是一位老板助理，要是你不懂统计学，不会做数据图表，那么，你就不可能赢得老板的认可……

可能有许多人认识到了统计学的重要性，但是，看到那一堆数据

报表和复杂的计算公式就望而却步了。其实，统计学不仅仅只有那些，没有复杂的计算公式，不用那些高深的理论，照样可以学习到统计学的一些基本原理和方法，毕竟，不是每个人都需要弄清楚统计学的每一个方面。有了这个觉悟之后，对于那些想学点统计学的读者，本书将会是一个极好的选择。既有最基本的统计学方法和原理，又有丰富具体的实际应用案例，手把手地带领读者轻松走进统计学的大门。

本书内容及体系结构

本书共分为 10 章，第 1 章是引导读者进入统计学的世界，用几个好玩的例子调动读者学习统计学的兴趣，同时让读者能够感受到统计学的魅力。从第 2 章开始，每一章都是介绍统计学的某个原理或者方法，其中，每一章都有 5 个左右的应用实战案例。对于每一个案例，都从统计学的角度来看，进行系统详细的分析讲解，力求做到深入浅出，让读者既容易理解又能学到东西。

本书特色

1．选择基础统计理论、讲解通俗易懂

本书不走寻常路，抛开了统计学中那些深奥的理论知识，并不是就给读者讲述一大堆空洞而乏味的理论知识。对于统计学的理论，力求去粗取精，绝不贪多求全。本书将那些实用的、基础的、易懂的统计方法和原理，完整地展示给读者，力求用通俗易懂的讲解，给读者带来感悟和启发，让读者能够完全理解和吸收，体会到统计学的魅力。

鉴于本书只是一本入门级别的科普类书籍，且本书的一大特色是计算量比较小，没有生涩难懂的学术概念。因此，对于绝大多数读者来说，阅读起来毫无压力，大家可以轻松踏入统计学的大门。

2．精选典型应用案例、分析深入浅出

对于统计学的基本方法和原理，如果没有实际的应用案例，那么，学习起来的效果将会大打折扣。本书正是意识到了这一点，故在介绍统计学方法和原理的同时，都会尽可能地引入一些经典的统计学实战案例。这些例子涉及范围广，有些贴近生活，有些又是历史上的趣事，读者不会觉得枯燥乏味。

对于每个案例，本书从统计学的角度深入浅出地进行了分析，思路清晰，逻辑严谨，让读者在这些案例的解析中，了解统计学的核心思想，掌握实用的统计学方法。

3．精挑细选案例、应用场景丰富

本书一共介绍了超过 30 个应用案例，每个应用案例具体来说都是一种应用场景，这些案例涵盖了九大领域，既包含了物流运输、生产销售这样的经济领域，也囊括了人际交往、家庭生活这样的生活领域。如此丰富的应用场景，保证读者能够和自己的实际生活结合起来，从而能够让读者迅速将统计学的思想和方法应用到日常的每一件事情上。

本书读者对象

- 统计学入门的读者
- 对大数据感兴趣的读者
- 想学习实用统计方法的读者

本书由刘强组织编写，同时参与编写的还有张昆、张友、赵桂芹、陈冠军、魏春、张燕、项宇峰、晁楠、高彩琴、郭现杰、刘琳、王凯迪、王晓燕、吴金燕、尹继平、张宏霞、张晶、姚志娟、马翠翠、范陈琼、孟春燕、王晓玲、肖磊鑫、薛楠、杨丽娜，在此一并表示感谢！

因水平和成书时间有限，书中难免存有疏漏和不当之处，敬请指正。

目录

Contents

第1章

人人都要会统计

大数据时代到了，掀起了一波“统计热”。统计学，在大数据时代越来越受到重视，并被应用在越来越多的方面，如无人驾驶汽车、AlphaGo、机器学习和人工智能等，都是以统计学为基础理论。大数据时代，人人都应该会点统计学，学会读懂数据，学会用数据说话。

1.1 从啤酒和尿布说起

啤酒和尿布，本来是毫无联系的两件物品，可是，当时有一家超市对销售数据进行统计分析，发现了它们之间的联系，并将这两种风马牛不相及的商品摆在一起，就是这样看起来荒谬的举措，让尿布和啤酒的销量大幅增加，刺激了整个超市营业额的增长。这并不是在开玩笑，而是一个用好统计学的真实案例，至于这家超市，正是如今风靡全球的沃尔玛。啤酒和尿布的故事也就一直流传至今，仍然被人们津津乐道，已然被视为利用统计方法挖掘数据关系的经典之作。

原来，在当时的美国，对于大部分家庭来说，一般都是母亲留在家里照顾婴儿，而父亲去超市购买尿布。很多时候，这些年轻的父亲们在购买尿布的同时也会有点自己的小心思——顺便为自己买点啤酒。正因如此，啤酒与尿布这两件看上去毫不相干的商品，就经常会出现在同一个购物篮里。当时，沃尔玛通过分析啤酒和尿布的销售数据，发现了它们之间这种隐蔽而又合乎情理的关系，就决定改变商品的摆放位置，将啤酒和尿布摆放在相同的区域。

对于家里有婴儿的父亲们，如果去某个商店买尿布的时候，商店里没有啤酒，或者啤酒的位置离尿布太远了，那么，他们心里的小算盘就会很快被打消，或者下次就会到别的商店去看看，这就无形中造成了营业额的降低和顾客的流失。而沃尔玛将尿布和啤酒摆到了一起，看似只是尝试一次小小的位置调整，却让那些父亲们能够很方便地同时购买两件商品，这不仅仅提高了这些顾客的购物体验，由于这些父亲们在买尿布的同时顺便又买了啤酒，还使啤酒和尿布的销量增加了，这样也就让超市的营业额增加了。

从上面的故事可以看出，看似一次漫不经心的位置调整，却能够为超市吸引更多的顾客，并带来更多的销量和利润，这给我们带来了什么启示呢？试想一下，如果沃尔玛超市没有认真对各种商品的销售数据进行分析，又怎么能够发现啤酒和尿布之间的联系，更不会有接下来的调整了。

从统计学的角度来看，沃尔玛通过分析啤酒和尿布的销量，发现了这两个变量之间的相关关系。在如今的大数据时代，人们对于数据越来越重视，像网购时的浏览点击、用户在网页上的逗留时间、用户的 GPS 定位信息等都已经被当做了数据。收集到的这些数据可以用来进行相关性分析或者回归分析，就可以发现隐藏在这些纷繁复杂的数据背后的相关关系或者因果联系，即从数据中得到的规律。

根据分析数据得到的这些规律或者某种联系，不仅网站能够用来设计精准的推荐系统，为我们购物、听歌、看电影时提供方便；还有医院也能够用来找到疾病的病因，为我们的身体健康保驾护航等。可以这样来说，我们目前衣食住行等各方面的大多数便利，都是建立在分析数据的基础上完成的。统计学，对于现在这个快速高效的互联网社会，起着功不可没的作用。

其实，统计学并非最近几年才被重视，也并非只是在最近才体现出巨大的作用。在学术方面，早在这门学科成熟之前，统计学就已经和经济学、社会学等结下了很深的缘分。将统计学应用于经济学、社会学等其他学科之中，从而形成了许多新的交叉学科，让人类知识的触角更进一层。统计学里面的各种方法和原理，早就已经成为科研领域的必修课，实验室的人们根据它来分析实验数据。

在工作方面，统计学兼具用数据说话的缜密严谨和用图表示人的形象生动。几乎所有的白领岗位，都要学会基本的数据分析并将数据可视化，也就是将分析数据的成果用统计图表的形式进行展示。总