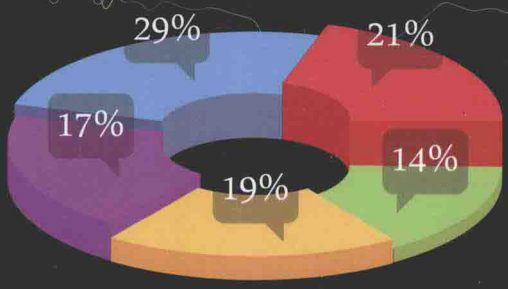
互联网+时代，最后都是数据的生意
而统计学能让数据插上价值的翅膀

博文金融

STATISTICS



大数据时代下的统计学

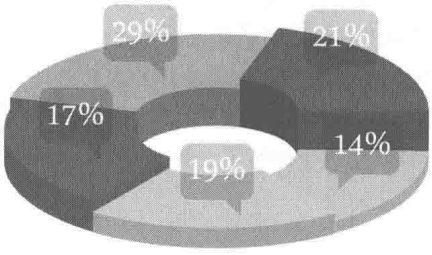
杨轶莘◎编著

博学 · 慎思 · 明辨 · 笃行

5大统计学专业方向 / 65个统计学知识点 / 50个经典的统计学案例
教会你如何说服别人Believe in the power of data（相信数据的力量）

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



大数据时代下的统计学

杨轶莘◎编著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

这是一本讨论时下热门话题——大数据的书，生动活泼地阐释了晦涩艰深的统计学原理，条理清晰地告诉读者如何从数据中获取智慧。

全书分为 8 章，第 1 章概述了大数据时代统计学面临的机遇和挑战。第 2、3 章讲述了统计学在思想方法及数据表述上和大数据处理方法的异同；第 4 章介绍了对统计学影响深远的正态分布；第 5 章探讨了大数据时代统计推断是否失效；第 6 章重点从统计学视角讲述了大数据时代最热门的变量间的“相关性”问题；第 7 章以一种比较开放的态度讨论统计学中一些有意思又实用的话题；第 8 章探讨大数据能够给企业、用户及整个产业和社会带来什么价值。

拥有本书，不仅可以使读者感受数字的美感和哲学的智慧，还能够使读者掌握思辨的洞察力。更重要的是，拥有本书就相当于拥有了一种武器，数据驱动的思维模式将会使读者在生活、工作中受益匪浅。

图书在版编目（CIP）数据

大数据时代下的统计学 / 杨轶莘编著. —北京：电子工业出版社，2015.9

ISBN 978-7-121-26936-3

Ⅰ. ①大… Ⅱ. ①杨… Ⅲ. ①统计学 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字（2015）第 189068 号

责任编辑：徐津平
特约编辑：赵树刚
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编 100036
开　　本：720×1000　1/16　　印张：12.5　　字数：320 千字
版　　次：2015 年 9 月第 1 版
印　　次：2015 年 9 月第 1 次印刷
定　　价：39.80 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

不知不觉中，人们进入了一个数据为王的时代。大数据的字眼以一种迅雷不及掩耳之势进入人们的视野，更加强调了数据在这个时代的重要性。不管人们愿意或者不愿意，都在诚惶诚恐地拥抱着这个所谓的大数据时代。大数据的火热也带火了另外一个看上去有点神秘、有点距离感的学科：统计学。

为什么编写本书

笔者作为一个在校园里学了 11 年统计学的资深学院派，深深地被这门学科打动：它有着数学的美感，充满了哲学的智慧，并且透露出思辨的洞察力。你可以把它看作一种工具，或者一种武器。有了它，你可以事半功倍地直击事物本质的规律。

笔者很想把这门学科分享给有兴趣的人。这就是编写这本书的初衷。

统计学本身就是大数据时代的一门重要学科。随着大数据逐渐走进公众的视野，统计学也必然会迎来更多的关注。这就意味着，越来越多的非统计学专业人士会了解统计学、应用统计学。人们也必然需要更多的统计学读物。

据笔者观察，市场统计学的教材大多像教科书，充斥着枯燥的公式和深奥的理论。当然，也有一些幽默风趣、深入浅出的入门书籍，如查尔斯·韦兰的《赤裸裸的统计学》（*Naked Statistics*），但也因为是外国作品编译的问题，在语言和写作方式上很难符合东方人的阅读习惯。

这本书讨论大数据，讨论统计学，更讨论二者之间千丝万缕的联系。大数据时代将迎来技术的变革，以及工作方式和思维模式的变革。大数据时代也挑战着传统统计学的思维和研究模式。统计学这门学科是将要面临江河日下的被取代的危机，还是迎来一个破茧而出的春天？本书试着抛砖引玉地给出一部分答案。

大数据时代，对于统计学来说，是最好的时代，也是最坏的时代。统计学，必须与时俱进，勇敢地接受大数据时代的挑战和变革，才会走得更长远。而大数据，没有了统计学思维的辅助、修正和补充，当热潮退去，也只能在这个浮躁的时代中渐渐被人们遗忘。

本书特点

本书从当下热门话题大数据切入，引入与之息息相关的统计学。深入浅出地讲述了在“数据为王”的时代下，统计学作为分析、解读数据的学科，如何为商业、社会、生活等领域提供决策支持。

- 热门性——业界和学术界热议的词“大数据”对大多数人来说仍是“犹抱琵琶半遮面”。
- 经典性——久经时间考验的统计学理论仍是实践中数据处理的重要依据。
- 洞察性——站在统计学哲学的思想高度对时下热门话题进行分析思考。
- 前瞻性——下一个时代是数据的时代。无论什么行业，未来都是数据生意。

本书和市面上很多书籍相比，有两点最具特色：

- 本书将统计学和大数据结合在一起，探讨两者的差异和相关性。
- 本书行文按照【案例】+【知识点】+【分析】的结构，清晰明了。应用的案例也都和人们的生活息息相关，更符合国人阅读习惯，更具代入感和认同感。

本书内容

本书共分为 8 章，各章内容如下。

第 1 章　大数据时代下的统计学，讲解了统计学的基本原理、应用领域及数据的获取方法。

第 2 章　样本魅影，重点介绍了统计学最核心的思想，即用样本信息推论总体，并和大数据的推论思想进行比较，强调二者在实践中结合使用的重要性。

第 3 章　描述数据，告诉读者面临大量数据的时候，如何迅速提炼出有用信息，以一种直接、感性的方式勾勒出隐藏在冷冰冰的数据背后的内涵。

第 4 章　正态女神，隆重推出了统计学最经典、最重要、最具代表性的一个分布——正态分布，详细介绍了关于正态分布的理论、应用和相关的知识点。

第 5 章　统计推断，讲述了统计推断是用样本来估计总体的，是一种具有科学依据的合理猜测，尽管它不可能百分百准确，却对人们认知事物有着不可估量的作用。

第 6 章　变量间的关系，从大数据思维的其中一个角度切入，即强调事物的相关关系而非因果关系，重点讲述了究竟什么是相关关系，它的统计学内涵、方法及应用。

第 7 章　统计杂谈，以一种漫谈的方式，深入浅出地讲解了统计学一些热门应用的理论。特别强调了这些理论在实践中的误用，并告诉读者正确的使用方法和解读方法。

第 8 章　大数据，在水一方，探讨了大数据巨大的商业价值，除此之外还强调如何从大数据中获取洞察力和决策力。

关于作者

本书由杨轶莘主笔编写，其中第 6 章由王辉撰写。

杨轶莘：瑞典厄勒布鲁大学商学院统计学博士毕业，北京诺贝伦思教育咨询有限公司高级咨询师，旗下商学院 CN 网站联合创始人和网站知识分享类微

信节目《杨博夜话》制作人和主持人。

王辉：北京大学汇丰商学院金融学（数量金融方向）研究生。善于统计综合评价方法的应用、金融计量学、经济计量分析领域的研究。2013—2014 年，主持项目《社区养老现状和需求研究》，获第四届全国大学生市场调查分析大赛一等奖和第三届海峡两岸市场调查分析大赛二等奖。2014—2015 年，参与朱喜安教授的国家社科基金课题《综合评价方法的优良标准研究》。

目　录

第 1 章　大数据时代下的统计学 1

1.1　统计学——天使还是恶魔 1

【知识点】统计学的定义 1

1.2　概率——上帝的指引 3

【案例 1】硬币的指引 3

【案例 2】赌徒的错觉 3

【知识点 1】随机性 4

【知识点 2】概率 4

1.3　小概率事件≠必然不会发生的事件 6

【案例】挑战者号航天飞机（STS Challenger）失事 6

【知识点】“必然会发生”和“必然不会发生”的事件 6

1.4　你真的了解数据吗 7

【案例】淘宝的客户评价体系 8

【知识点】数据的类型 8

1.5　数据来自哪里 10

【案例】大数据，大偏差——谷歌的流感预测模型真的靠谱吗 10

【知识点 1】二手数据 11

【知识点 2】相关关系和因果关系 11

第 2 章　样本魅影 14

2.1　样本——窥一斑而见全豹，观滴水而知沧海 15

【案例 1】客户满意度调查 15

【案例 2】救护车垄断业务调查 16

【知识点】随机样本，方便样本和自愿回应样本 17

2.2 抽样——尝一勺锅里的靓汤 18
【案例 1】红豆和绿豆 18
【案例 2】“捉放法”估算鱼苗成活率 19
【案例 3】被解雇的市场调研部员工 20
【知识点 1】简单随机抽样 21
【知识点 2】抽样中存在的错误风险 22
【知识点 3】访问员 23
2.3 不回应误差——沉默不是金 24
【案例】不回应的影响有多大 24
【知识点 1】不回应（Nonresponse） 24
【知识点 2】如何降低不回应率 25
2.4 措辞的艺术——僧推/敲月下门 26
【案例 1】娱乐圈话题：锋菲恋 26
【案例 2】几字之差对于民众支持率的影响 27
【案例 3】双重否定的疑惑 28
【知识点 1】响应误差（Response Error） 29
【知识点 2】有效性（Validity）和可靠性（Reliability） 29
2.5 大数据时代，当“样本”已成往事 31
【案例】Farecast，美国创业梦 31
【知识点】大数据的 4V 特征 32

第 3 章 描述数据 34

3.1 均值——可能会说谎的天平 34
【案例 1】中关村创业者平均 39 岁 34
【案例 2】令人啼笑皆非的统计局数据 35
【知识点】均值计算 36
3.2 寻找中位数——排序，数到中间 37
【案例 1】腾讯笔试题：大数据量寻找中位数 37
【案例 2】淘宝卖家评分体系 38

【知识点 1】求取中位数 39
【知识点 2】四分位数 40
3.3 标准差、标准误，傻傻分不清楚 42
【案例 1】均值-方差证券资产组合理论 42
【案例 2】语文成绩调研 42
【知识点 1】标准差（Standard Deviation） 43
【知识点 2】标准误（Standard Error） 43
3.4 图形替数据说话——“剩女”和相亲市场 46
【案例】“剩女”和潜力巨大的相亲市场 46
【知识点 1】饼状图（Pie Chart） 48
【知识点 2】条状图（Bar Chart） 49
【知识点 3】散点图（Scatter Plot） 50
3.5 数据可视化——“云想衣裳花想容” 51
【案例】谁在开网店 51
【知识点 1】什么是数据可视化 54
【知识点 2】数据可视化主要应用领域 55
【知识点 3】数据可视化的工具 55
第 4 章 正态女神 57
4.1 期望——量化你的预期 58
【案例 1】掷骰子和伯努利试验 58
【案例 2】赌场就是概率场 59
【知识点 1】概率分布 60
【知识点 2】期望（Expectation） 61
【知识点 3】方差 62
4.2 大数定律——为什么十赌九输 63
【案例 1】澳门风云 63
【案例 2】谁会是被骗的大傻瓜 64
【知识点】大数定律 65

4.3 正态分布——大道至简，大美天成 65
【案例 1】高尔顿钉板 65
【案例 2】女博士嫁人难，谁之过 67
【知识点】正态分布 68
4.4 中心极限定理 70
【案例】肯家和麦家的博弈 70
【知识点】中心极限定理 70

第 5 章 统计推断 74

5.1 点估计——统计学家比间谍干得漂亮 75
【案例 1】二战中的德军坦克数 75
【案例 2】首家新鲜咖啡速递服务企业 76
【知识点 1】样本统计量和总体参数 77
【知识点 2】点估计 77
5.2 置信区间——责善切戒尽言 79
【案例】美国盖洛普公司的民意调查 79
【知识点 1】置信水平 79
【知识点 2】置信区间 80
5.3 两类错误：有罪被判无罪和无罪被判有罪哪个更严重 81
【案例 1】法律中的人文精神 81
【案例 2】抗击埃博拉要避免两类错误 82
【知识点 1】零假设和备择假设 84
【知识点 2】两类错误 84
5.4 假设检验——“凑巧”可以拒绝吗 85
【案例 1】奶茶情缘 85
【案例 2】咖啡新鲜吗 87
【知识点 1】显著性水平 88
【知识点 2】p 值 88
【知识点 3】统计显著 88

【知识点 4】统计显著 vs. 实际显著 89
【知识点 5】假设检验 vs. 置信区间 89
【知识点 6】单侧检验 vs. 双侧检验 90
5.5 *p* 值——打开潘多拉魔盒的钥匙 92
【案例】金榜题名无望、少年得志梦断 92
【知识点 1】*p* 值的历史和思想 93
【知识点 2】*p* 值误用 94
第 6 章 变量间的关系 96
6.1 卡方分析——细腻的眼神里岂容得半粒沙 97
【案例 1】仙道迟到事件发生率分析 97
【案例 2】性别和文化程度是相互独立的吗 98
【知识点 1】卡方分布 99
【知识点 2】卡方检验 100
6.2 相关性分析——早起的鸟儿有虫吃 102
【案例 1】早起的鸟儿有虫吃 102
【案例 2】化妆品销售额与广告费的关系分析 103
【知识点 1】相关关系 104
【知识点 2】相关分析 105
【知识点 3】相关表、相关图和相关系数 106
【知识点 4】相关系数 *t* 统计量 107
6.3 ANOVA——地域，我们没有什么不同 107
【案例】地域歧视问题 107
【知识点 1】方差分析 108
【知识点 2】方差分析统计模型 109
【知识点 3】离差平方和及其分解 110
【知识点 4】均方 111
【知识点 5】AMOVA F 统计量 112
【知识点 6】方差分析表 113

6.4 回归分析——对不起，其实我也想长高 117
【案例 1】子女身高遗传学的发现 117
【案例 2】身高地区差异分析 117
【知识点 1】回归分析 119
【知识点 2】随机误差项 119
【知识点 3】最小二乘法 120
【知识点 4】回归分析 T 检验 121
【知识点 5】回归分析 F 检验 122
【知识点 6】拟合优度 R^2 123

第 7 章 统计杂谈 124

7.1 为什么对回归情有独钟 124
【回归和电影】 126
【回归和手游】 128
7.2 调查问卷中的分类变量 132
【疼痛】 133
【Rank-Invariant】 134
【Svensson Method】 135
【工作环境和员工满意度】 136
7.3 条件概率和更多的信息 138
【生男生女的问题】 139
【门后的世界：到底是谁错了】 140
7.4 极大似然估计——看起来最像 142
【白狐，iphone 6 plus 和房价】 143
7.5 R you happy 145
【名门闺秀 SAS】 145
【国民初恋 SPSS】 146
【小家碧玉 Stata、Minitab、Excel】 147
【清新萝莉 R】 148

7.6 贝叶斯 149
【起源】 150
【定义】 150
【自拍杆和蓝牙耳机】 152
7.7 来自星星的统计陷阱 155
【被黑的统计机构】 155
【统计局的无奈】 157
【王老吉状告加多宝】 158

第 8 章 大数据，在水一方 161

8.1 洛阳纸贵——大数据思维 161
【案例 1】罩杯和败家程度 166
【案例 2】外滩踩踏悲剧 167
【案例 3】大数据和途牛网 169
8.2 大数据驱动运营 171
【案例】DataEye，数据驱动手游运营 175
8.3 商业智能——决策者的锦囊 177
【案例】广告业的商业智能 178
8.4 市场智能——商业智能的衍生智慧 179
8.5 消费智能——当数据成为一种服务 182

第 1 章

大数据时代下的统计学

不知不觉中，“大数据”这个概念“忽如一夜春风来，千树万树梨花开”，以迅雷不及掩耳之势进入人们的视野。各行各业恨不得搭上这辆顺风车。大数据的语义核心是数据。大数据火了，也带火了另一个和数据相关的学科——统计学。许多高校增设统计学专业，市场上对统计人才的需求也大大增长。但也有人认为大数据思维和统计思维有着本质区别，随着获取和存储数据能力的不断增强，大数据方法的不断成熟，传统的统计学必将被取代。

既然在大数据时代统计学不会消亡，反而会起到举足轻重的作用，那么统计方法就不应该只是少数学者所掌握的工具，而应该走向生活、走向大众，使应用统计学方法转化成一种像读书看报一样的普通技能。

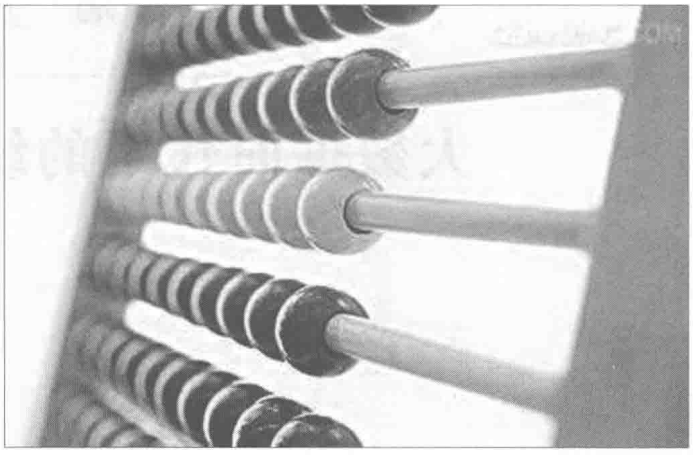
1.1 统计学——天使还是恶魔

【知识点】统计学的定义

在《不列颠百科全书》中统计学定义如下：收集、分析、表述和解释数据

的艺术和科学。这个定义被科学界普遍认可。

那么统计学究竟是一门怎样的学科呢？



白衣天使南丁格尔也说：“若想了解上帝在想什么，我们就必须学统计，因为统计学就是在量测他的旨意。”不过，犀利的大文豪马克·吐温却说世界只有三种谎言：谎言、该死的谎言和统计学。一正一反，两种评价大相径庭。

其实，统计学是一门基于数据的学科。数据是严谨的、枯燥的、冷冰冰的。同时，数据又是丰富的、客观的、忠实的、从不会欺骗人的。数据是数字，但不只是数据。统计学还是一门关于数据的艺术。当然，数据收集不是目的。如何高效、准确地分析所得数据，并把它转化成比数据本身更有用的知识才是统计学的目的。

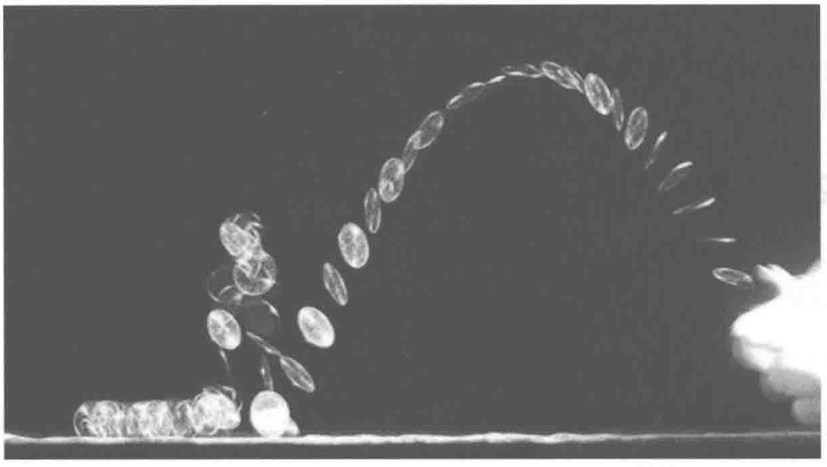
世间的一切，貌似杂乱却又暗自遵循着某种规律，就像毕德哥拉斯学派形容的那样，万物皆是数，“在理性的基础上，所有的判断都是统计学”。在不知不觉中，国家、企业和个人已经成为一个个“运行于数字之上的国家、企业和个人”。

统计学是“万金油”，它在金融、经济、医学等领域的应用最广、知名度最高，但也不乏一切令人意想不到的领域的人也在使用着统计学知识。不管是大数据时代，还是小数据时代，统计学都不是万能的，可没有统计学却是万万不能的。

1.2 概率——上帝的指引

【案例1】硬币的指引

我是一个有选择恐惧症的人，遇到难以决断的事，就会抛硬币来决定，认为这样做更接近“上帝的指引”。比如，我会在心里默念：我的统计学会不会挂科？然后，告诉自己，如果数字的那面朝上就会挂科。接着，我把一枚硬币抛向天空，忐忑地等待它落下。结果令人沮丧：菊花朝上！



如果我继续不厌其烦地抛那枚硬币，抛了1 000次，我会惊讶地发现数字和菊花出现的次数都大约为500次。这就意味着，上天给我的指引其实是十分中立的：你挂科或者不挂科的可能性各占一半。原来这就是随机性中暗含的规律性。而这种规律性就量化地体现为概率。

【案例2】赌徒的错觉

统计学和赌博自古至今一直“血脉相连”。庄家常胜，胜在深谙统计学，把