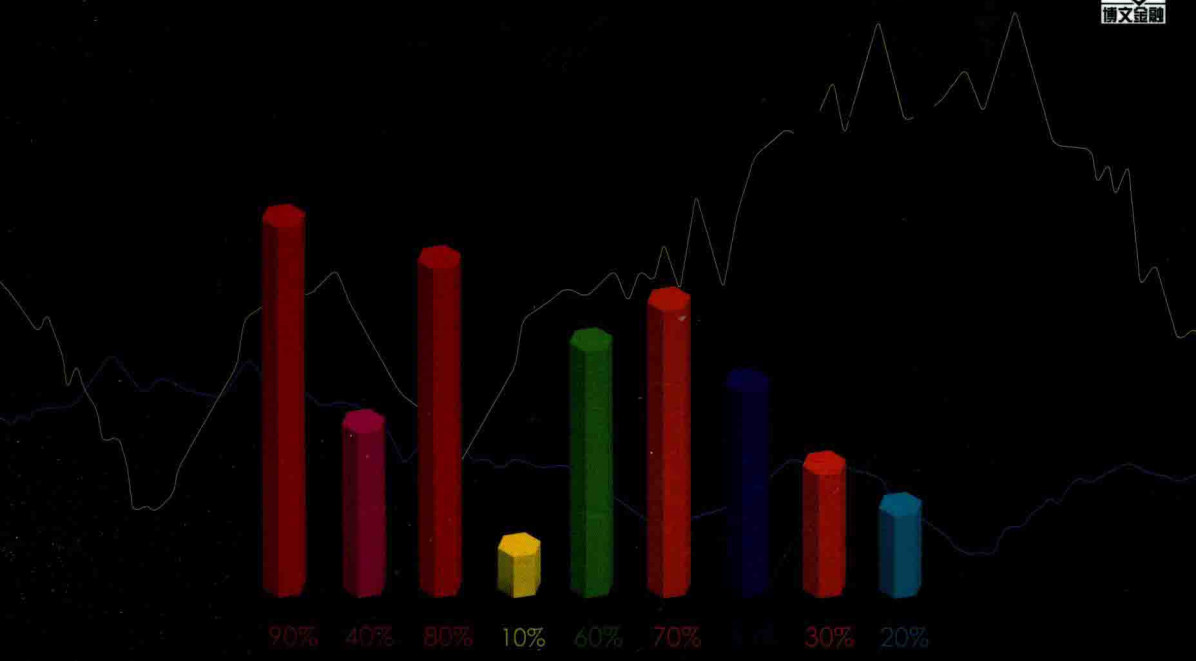
大数据及人工智能产教融合系列丛书

博文金融



大数据时代下的统计学

（第2版）

杨轶莘◎编著

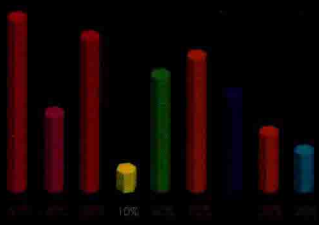
博学·慎思·明辨·笃行

五大统计学专业方向 / 62个统计学知识点 / 47个经典的统计学案例

教会你如何说服别人相信数据的力量

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



名家力荐

市场上有关统计学的图书多为教科书似的理论书，充斥着枯燥的公式。当然，也有幽默风趣、深入浅出的外版入门书籍，但有些在语言上很难适应国内读者的习惯。本书很适合作为我国高校学子学习统计学的入门读物。

财会类畅销书作者 宋娟

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的人士会了解统计学、应用统计学，人们也必然需要更多的统计学科普读物。它不需要很难理解，也不需要涉及很多理论知识，最重要的是把问题讲清楚，让大家领会精神。

深圳悦策数据科技有限公司 技术总监 马超

统计学是一门客观的、偏理科的学科，本书作者却融入了很多个人色彩在里面，十分亲切，有点小清新，透着浓浓的人情味儿，看得出作者是一个有想法、有情怀的学者。

清华大学 教授 刘春鹏

本书有两个优点：第一，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；第二，洞察性——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。

鹏博士云数据中心 运营总监 孙伟辉

上架建议：统计学

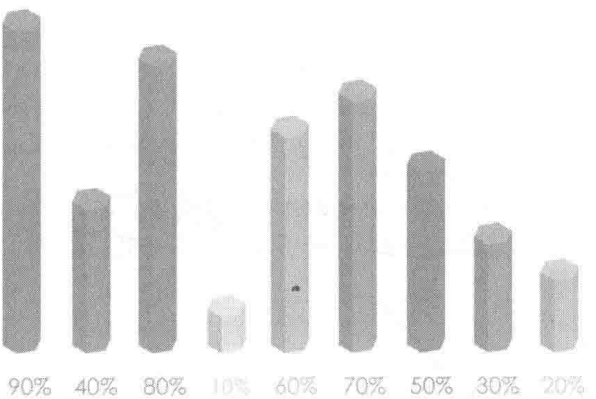
ISBN 978-7-121-37087-8

9 787121 370878

定价：59.00元

责任编辑：高洪霞
封面设计：侯士卿

大数据及人工智能产教融合系列丛书



大数据时代下的统计学

（第2版）

杨轶莘◎编著

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

内 容 简 介

本书从大数据切入，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。

全书分为 8 章，第 1 章概述了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、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 章探讨了在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、用户及整个社会带来的价值。

本书不仅可以使读者感受到数字的美感和哲学的智慧，还可以使读者获得思辨的洞察力。更重要的是，拥有本书就相当于拥有了一种武器，其中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. —2 版. —北京：电子工业出版社, 2019.9
（大数据及人工智能产教融合系列丛书）
ISBN 978-7-121-37087-8

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（2019）第 144532 号

责任编辑：高洪霞
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
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：245 千字
版　　次：2015 年 9 月第 1 版
　　2019 年 9 月第 2 版
印　　次：2019 年 9 月第 1 次印刷
定　　价：59.00 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）88254888，88258888。

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。

本书咨询联系方式：（010）51260888-819，faq@phei.com.cn。

前　言

在不知不觉中，人们已经进入了一个数据为王的时代。大数据以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。不管人们是否愿意，都要拥抱这个“大数据时代”。同时，大数据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，你就可以直击事物本质的规律并能取得事半功倍的效果。

笔者很想把这门学科分享给对它感兴趣的人。这就是编写本书的初衷。

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的人想去了解和应用统计学，也必然需要更多的统计学读物。

据笔者观察，市场上有关统计学的书大多都像教科书，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，在语言和写作方式上很难符合东方人的阅读习惯。

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，那么统计学是要面临江河日下被取代的危机，还是要迎来一个破茧而出的春天？本书抛砖引玉，试着给出一部分答案。

对于统计学来说，大数据时代是最好的时代，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。而如果大数据没有了统计学思维的辅助、修正和补充，当热潮退去，那么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“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数据处理的重要依据。
- 洞察性——站在统计学和哲学的思想高度，对时下热门话题进行分析思考。
- 前瞻性——任何行业，未来都是数据生意。

本书有两大特色：

- 将统计学和大数据结合在一起，探讨两者的差异和相关性。
- 大部分章节都是按照【案例】+【知识点】的结构进行讲述的，清晰明了。本书应用的案例也都和人们的生活息息相关，更具代入感和认同感，语言也更符合读者的阅读习惯。

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，强调在实践中两者结合使用的重要性。

第 3 章：描述数据，告诉读者当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”，隆重推出了统计学中最经典、最重要、最具代表性的一个分布——正态分布，详细介绍了正态分布的理论、应用及相关知识点。

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，尽管它不可能完全准确，但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，从大数据思维的一个角度切入，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，以及其统计学

的内涵、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，还强调了如何从大数据中获取洞察力和决策力。

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 …… 2
【知识点】统计学的定义 …… 2
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 …… 4
【知识点 1】随机性 …… 5
【知识点 2】概率 …… 5
1.3　小概率事件≠必然不会发生的事件 …… 7
【案例】挑战者号航天飞机失事 …… 7
【知识点】“必然会发生”的事件和“必然不会发生”的事件 …… 7
1.4　你真的了解数据吗？ …… 8
【案例】淘宝的客户评价体系 …… 9
【知识点】数据的类型 …… 10
1.5　数据来自哪里？ …… 11
【案例】大数据，大偏差——谷歌的流感预测模型真的靠谱吗？ …… 12
【知识点 1】二手数据 …… 13
【知识点 2】相关关系和因果关系 …… 13

第2章 样本魅影 …… 15
2.1 样本——窥一斑而见全豹，观滴水而知沧海 …… 16
【案例1】客户满意度调查 …… 16
【案例2】救护车垄断业务调查 …… 17
【知识点】随机样本、方便样本和自愿回应样本 …… 18
2.2 抽样——尝一勺锅里的靓汤 …… 20
【案例1】红豆和绿豆 …… 20
【案例2】“捉放法”估算鱼苗成活率 …… 21
【案例3】被解雇的市场调研部员工 …… 22
【知识点1】简单随机抽样 …… 23
【知识点2】抽样中存在的错误风险 …… 24
【知识点3】访问员 …… 25
2.3 不回应误差——沉默不是金 …… 26
【案例】“不回应”的影响有多大 …… 26
【知识点1】不回应 …… 27
【知识点2】如何降低不回应率 …… 27
2.4 措辞的艺术——僧推/敲月下门 …… 29
【案例】几字之差对民众支持率的影响 …… 29
【知识点1】响应误差 …… 30
【知识点2】有效性和可靠性 …… 30
2.5 大数据时代，当“样本”已成往事 …… 32
【案例】Farecast，美国创业梦 …… 32
【知识点】大数据的4V特征 …… 33
第3章 描述数据 …… 36
3.1 均值——可能会说谎的天平 …… 37
【案例1】中关村创业者平均年龄39岁 …… 37
【案例2】令人“啼笑皆非”的统计局数据 …… 38
【知识点】均值计算 …… 38
3.2 寻找中位数 …… 39
【案例1】腾讯笔试题：大数据量寻找中位数 …… 39

【案例 2】淘宝卖家评分体系 …… 40
【知识点 1】求取中位数 …… 42
【知识点 2】四分位数 …… 42
3.3 标准差、标准误，傻傻分不清楚 …… 45
【案例 1】均值-方差证券资产组合理论 …… 45
【案例 2】语文成绩调研 …… 45
【知识点 1】标准差 …… 46
【知识点 2】标准误 …… 47
3.4 数据可视化——“云想衣裳花想容” …… 49
【知识点 1】什么是数据可视化？ …… 50
【知识点 2】数据可视化的主要应用 …… 50
【知识点 3】数据可视化的工具 …… 51
第 4 章 正态“女神” …… 53
4.1 期望——量化你的预期 …… 54
【案例 1】掷骰子和伯努利试验 …… 54
【案例 2】赌场就是概率场 …… 55
【知识点 1】概率分布 …… 56
【知识点 2】期望 …… 57
【知识点 3】方差 …… 59
4.2 大数定律——为什么十赌九输 …… 60
【案例 1】澳门风云 …… 60
【案例 2】谁会是被骗的人 …… 61
【知识点】大数定律 …… 62
4.3 正态分布——大道至简，大美天成 …… 63
【案例】高尔顿钉板 …… 63
【知识点】正态分布 …… 64
4.4 中心极限定理 …… 66
【案例】肯德基和麦当劳的博弈 …… 66
【知识点】中心极限定理 …… 67

第5章 统计推断 …… 70
5.1 点估计——统计学家比间谍干得漂亮 …… 71
【案例1】第二次世界大战中的德军坦克数 …… 71
【案例2】首家新鲜咖啡速递服务企业 …… 72
【知识点1】样本统计量和总体参数 …… 73
【知识点2】点估计 …… 74
5.2 置信区间——责善切戒尽言 …… 75
【案例】美国盖洛普公司的民意调查 …… 75
【知识点1】置信水平 …… 76
【知识点2】置信区间 …… 76
5.3 两类错误：有罪被判无罪和无罪被判有罪哪个更严重 …… 78
【案例1】法律中的人文精神 …… 78
【案例2】抗击埃博拉要避免两类错误 …… 79
【知识点1】零假设和备择假设 …… 80
【知识点2】两类错误 …… 81
5.4 假设检验——“凑巧”可以拒绝吗？ …… 82
【案例1】奶茶情缘 …… 82
【案例2】咖啡新鲜吗？ …… 84
【知识点1】显著性水平 …… 85
【知识点2】p 值 …… 85
【知识点3】统计显著 …… 86
【知识点4】统计显著对比实际显著 …… 86
【知识点5】假设检验对比置信区间 …… 87
【知识点6】单侧检验对比双侧检验 …… 87
5.5 p 值——打开“潘多拉魔盒”的钥匙 …… 89
【案例】p 值变了，结果就变了 …… 90
【知识点1】p 值的历史和思想 …… 91
【知识点2】p 值误用 …… 92

第 6 章　变量间的关系 …… 94
6.1　卡方分析——细腻的眼神里岂容得半粒沙 …… 94
【案例 1】仙道迟到事件发生率分析 …… 94
【案例 2】性别和文化程度是相互独立的吗？ …… 95
【知识点 1】卡方分布 …… 96
【知识点 2】卡方检验 …… 97
6.2　相关性分析——早起的鸟儿有虫吃 …… 100
【案例 1】早起的鸟儿有虫吃 …… 100
【案例 2】化妆品销售额与广告费的关系分析 …… 101
【知识点 1】相关关系 …… 102
【知识点 2】相关分析 …… 103
【知识点 3】相关表、相关图和相关系数 …… 104
【知识点 4】*t* 统计量 …… 105
6.3　ANOVA——地域，我们没有什么不同 …… 105
【案例】“地域歧视”问题 …… 105
【知识点 1】方差分析 …… 106
【知识点 2】方差分析统计模型 …… 107
【知识点 3】离差平方和及其分解 …… 109
【知识点 4】均方 …… 110
【知识点 5】*F* 统计量 …… 111
【知识点 6】方差分析表 …… 112
6.4　回归分析——对不起，其实我也想长高 …… 116
【案例 1】子女身高的遗传发现 …… 116
【案例 2】身高的地区差异分析 …… 117
【知识点 1】回归分析 …… 118
【知识点 2】随机误差项 …… 119
【知识点 3】最小二乘法 …… 119
【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 …… 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】……135
【Svensson Method】……135
【工作环境和员工满意度】……137
7.3　条件概率……139
【生男生女的问题】……140
【门后的世界：到底是谁错了】……141
7.4　极大似然估计——看起来最像……144
【白狐，iPhone 6 Plus 和房价】……144
7.5　统计软件……146
【名门闺秀 SAS】……147
【国民初恋 SPSS】……148
【小家碧玉 Stata，Minitab，Excel】……148
【清新萝莉 R】……150
7.6　贝叶斯……151
【起源】……152
【核心思想】……153
【自拍杆和蓝牙耳机】……155
7.7　来自星星的统计陷阱……157
【问卷调查的潜在陷阱】……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】外滩踩踏事件 ……168
【案例 3】大数据和途牛网 ……170
8.2 大数据驱动运营 ……171
【案例】DataEye，数据驱动手游运营 ……176
8.3 商业智能——决策者的锦囊 ……178
【案例】广告业的商业智能 ……179
8.4 市场智能——商业智能的衍生智慧 ……180
8.5 消费智能——当数据成为一种服务 ……183

第1章

大数据时代下的统计学

不知不觉中，“大数据”的概念“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，各个行业也都希望能搭上这辆顺风车。大数据的核心是数据。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加。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。

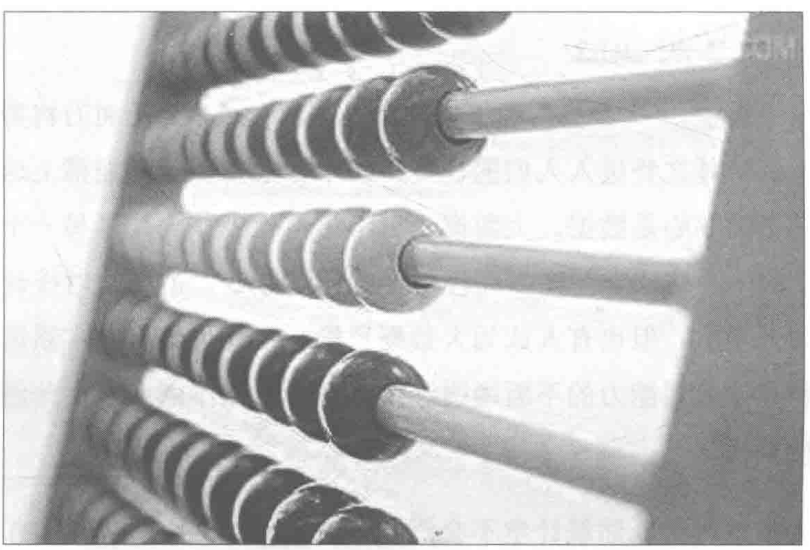
在大数据时代既然统计学不会消亡，反而会起到举足轻重的作用，那么统计方法就不应该只是少数学者所掌握的工具，而应该走向生活、走向大众，成为一种像读书看报一样的普通技能。

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》中将统计学定义如下：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·吐温却说世界上只有三种谎言：谎言、该死的谎言和统计学。一正一反，两种评价大相径庭。