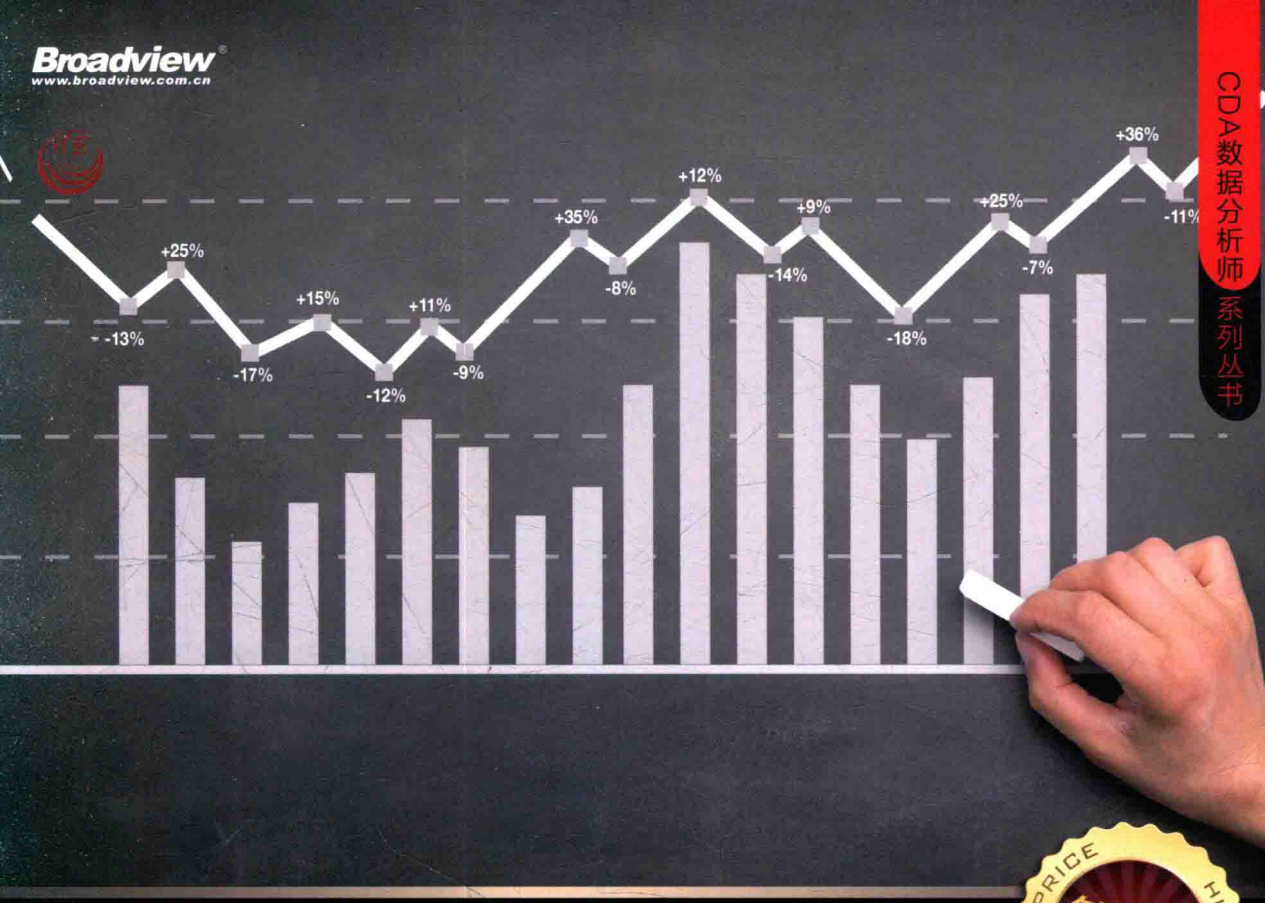
Broadview
www.broadview.com.cn

CDA数据分析师系列丛书



从零进阶!

数据分析的统计基础

经管之家 主编　曹正凤 编著

未来数据分析相关的就业岗位会有1000万人才缺口
CDA数据分析师系列丛书携你与时俱进!

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

CDA数据分析师 系列丛书

从零进阶！数据分析的统计基础

经管之家 主编　曹正凤 编著

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

内 容 简 介

《从零进阶！数据分析的统计基础（第 2 版）》共 7 章，分别讲解了数据分析的步骤和方法、描述性统计分析、数理统计基础、抽样估计、假设检验、方差分析、 相关与回归分析。本书使用简单的语言介绍了这些数据分析基本方法的核心思想和涉及的统计学、概率论等方面的理论内容，并使用图示的方法详细介绍了使用 Excel 2013 进行简单的描述性统计分析和使用 SPSS 进行相关的数据分析的过程与结果分析。

本书适合需要提升自身数据分析理论和实践能力的职场新人；在市场营销、金融、财务、人力资源管理中需要数据分析的人士，从事咨询、研究、分析等的专业人士。也可以作为数据分析师职业培训的教材，普通高等院校非统计专业数据分析的选修教材。

图书在版编目（CIP）数据

从零进阶！数据分析的统计基础 / 经管之家主编；曹正凤编著. —2 版. —北京：电子工业出版社，2016.5
（CDA 数据分析师系列丛书）
ISBN 978-7-121-28500-4

Ⅰ. ①从… Ⅱ. ①经… ②曹… Ⅲ. ①数据处理－统计分析 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 066359 号

策划编辑：张慧敏
责任编辑：王　静
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
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：787×980　1/16　印张：16　字数：397 千字
版　　次：2015 年 2 月第 1 版
　　2016 年 5 月第 2 版
印　　次：2016 年 5 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）88254881。

序言：这是一个用数据说话的时代

在 CDA（注册数据分析师）Level I 级教材付诸印刷之际，关于数据分析这个职业及其价值的报道就有很多，比如，下面两条报道就充分体现了在大数据时代下，数据分析的价值。这在以前是从来没有过的。

LinkedIn 的最新投票结果显示，“统计分析和数据挖掘”是 2014 年最大的求职法宝。LinkedIn 对全球超过 3.3 亿用户的工作经历和技能进行分析，公布 2014 年最受雇主喜欢、最炙手可热的 25 项技能，其中位列榜首的是统计分析和数据挖掘。

麦肯锡公司的一份研究预测称，到 2018 年，在“具有深入分析能力的人才”方面，美国可能面临着 14 万到 19 万人的缺口，而“可以利用大数据分析来做出有效决策的经理和分析师”缺口则会达到 150 万人。

早在 2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了一份关于管理信息的特别报告——《数据，无所不在的数据》，文中写道：“世界上有着无法想象的巨量数字信息，并以极快的速度增长……从经济界到科学界，从政府部门到艺术领域，很多地方都已感受到了这种巨量信息的影响。”2011 年，麦肯锡发布了《大数据：下一个具有创新力、竞争力与生产力的前沿领域》，使人们在这篇文章里认识到了数据的力量，于是，一夜之间，面向数据分析市场的新产品、新技术、新服务、新业态正在不断涌现。从个人、企业到国家层面，都把数据作为一种重要的战略资产，逐渐认识到了数据的价值，不同程度地渗透到每个行业领域和部门，大大提升了企业的经营利润，推动了经济的发展。

这是一个用数据说话的时代，也是一个依靠数据竞争的时代。目前世界 500 强企业中，有 90% 以上都建立了数据分析部门。IBM、微软、Google 等知名公司都积极投资数据业务，建立数据部门，培养数据分析团队。各国政府和越来越多的企业意识到数据和信息已经成为企业的智力资产和资源，数据的分析和处理能力正在成为日益倚重的技术手段。

作为一个数学和统计学的强国，数据分析、数据挖掘和大数据价值挖掘行业在我国仍属于朝阳行业，数据分析人才仍然比较稀缺。各行各业在平常工作中积累的各种各样的数据分析问题仍然没有得到及时有效地解决，有些问题，还是关乎本行业发展的至关重要的问题。数据积累越来越多，期待解决分析的数据问题也越来越多，人们逐渐习惯的使用数据作为决策的重要参考依据。据艾瑞的研究报告，未来与数据分析相关的就业岗位会在 1000 万左右，而目前来说国内合格的数据分析师不足 5 万左右，建立一个科学有效的数据分析师培训体系迫在眉睫。

在这样一个用数据说话的时代，积累了丰富的数据分析培训经验的人大经济论坛承担起使命，几番调查研究，几番反复推演论证，在 2013 年，这个大数据的“元年”，CDA 注册数据分析师应运

而生！

2003年，人大经济论坛依托中国人民大学成立，在金融、管理、统计领域已积淀11个年头，在国内享有良好声誉。

2006年，人大经济论坛数据分析培训中心设立，至今经历8个春秋，建立了大陆、台湾一线师资团队，培养人才已达3万余人。

2013年，“中国数据挖掘与数据分析俱乐部CDMC”在人大经济论坛旗下成立，2014年改名为“中国数据分析师俱乐部CDA”。来自政府、金融、电信、零售、电商、互联网、教育等行业人士加入会员，成功举办了数十场行业聚会。紧接着，积累了数据分析培训丰富经验的人大经济论坛在国内展开CDA数据分析师系统培训和认证考试，成功见证了1000余名数据分析师的成长。

2015年，人大经济论坛将提供高水平、多层次的数据分析培训服务，以在行业积累 多年的影响力，吸引更好更多的优秀师资，瞄准行业内重要的数据分析问题和难点，攻坚突破，建立更加规范的行业培训体系，引领数据分析培训行业向规范化、有效化和前瞻化方向发展，为数据分析培训做出应有的贡献。

其实，数学（含统计）和英语一样重要，都是人们不可或缺的重要技能。既然英语全民这么重视，数学及其数据分析的技能更加需求于方方面面，更应被做大做强。让我们共同期待人大经济论坛办成另一个数据的“新东方”！

覃智勇

2015年1月1日

前　言

本书第 1 版自 2015 年 2 月出版后，在市场上获得了强烈的反响，当月在当当网的新书热卖榜中排名第二，半年内销售近万册，至 2016 年 1 月已经印刷了 5 次，共发行近两万册，图书被收录进百度百科。

如此巨大的市场销量和好评，引起笔者的深思，除本书构思巧妙、内容翔实、文法流畅等主观因素外，宏观的市场环境也是不容忽视的。2015 年，中国经济由原来的爆发式增长进入到略显低迷的新常态，无论是企业还是商家都感受到了压力，钱不再像以前那样好赚了。如何实现经济增长，如何让企业存活下去，这就需要深挖企业内部的痛点和洞察外部客户的特点。深挖和洞察的过程就是数据分析的过程，数据分析时代在中国悄然到来了。

随着数据分析师的价值凸显，有越来越多先知先觉的人们纷纷转行加入到数据分析师的大军中。而统计学是数据分析师们必修的课程之一，“从零进阶！数据分析的统计基础”的本意就是让更多的人能从零基础快速进阶到数据分析领域，并且重点讲述数据分析师们必须具备的概率和统计的关键知识点。而经管之家（原人大经济论坛）适时地推出本书，使其得到了很好的市场回馈。正所谓天时地利人和，造就了一本好书。

为了和市场的发展紧密结合，以及更好地适应读者的需求，本书进行了改版。本次改版继续坚持从零进阶，强化数据分析基础理论，和市场接轨等核心理念，继续使用“三国武将”这个大家都耳熟能详的业务背景知识。根据学员的需求和市场的实际情况，作者还对本书内容进行了如下调整。

（1）进一步精练数据分析的理论基础，去除了一些不必要的数学公式。由于数据分析涉及概率论、微积分、数理统计的很多内容，但有些内容又不用全部学会，这让初学者很难找出哪些是需要学习的内容，哪些是不需要学习的内容。因此在编写本书第 1 版时，将很多数据分析师不需要知道的知识点都省略了，比如省略了统计量服从某个分布的证明过程，省略了抽样平均误差的证明过程。这样做的目的是为了让数据分析师们能更快地进入这个领域，更好地洞察数据。在编写本书的第 2 版时，继续沿用此思想，去掉了一些数据分析师不必要知道的公式，增加了更多的数据分析思想的内容。

（2）将原来的第 3 章抽样估计分解成数理统计基础和抽样估计两章，这样做的目的是考虑到原来的第 3 章涉及的理论内容太多，并且比较枯燥，将其分成两部分，一来可以在每一部分增加更多的公式解读内容，也可以补充更多的案例进来；二来降低了阅读难度，使读者能在学习知识的同时，获得更多的成就感，从而更加有兴趣学习。

（3）对试验数据进行了更多的数据分析，增加了对读者数据分析思维的培养。尤其是第 2 章的描述性数据分析过程，进行了更深入的数据分析过程剖析，主要宗旨在于让读者更快地进入到数据

分析行业的队伍中来。当然，这也使得第2版中的三国武将数据和第1版中的数据存在一些差异。

当然，仅就本书而言，读者并不会学到数据分析师所需要的全部知识，这需要几年的循序渐进学习，但我希望读者看过本书后，能快速具有数据分析师所需要的最基本的统计学知识，能快速地进入到数据分析的行业，从而具备一个数据分析师应具备的最起码的知识，在工作中能说内行话，而不是说行外话。

在本书改版之际，作者衷心感谢经管之家（原人大经济论坛）和CDA课程研发团队多年来始终不渝的关心与鼎力支持，感谢关继杰，感谢广大读者给予我的理解与感受，感谢电子工业出版社多年来的密切合作与支持。没有这一切，本书不可能取得这么好的成果，我永远感谢曾经帮助和支持过我的相识的和不相识的同志和朋友。由于作者水平有限，本书肯定会有不少缺点和不足，热切期望得到专家和读者的批评指正。

曹正凤

2016年3月于北京

目　录

第 1 章

数据分析概述

本章主要介绍数据分析的概念、步骤和方法，如何在 Excel 2013 中安装数据分析工具，以及如何安装 SPSS 数据分析软件，这是后续进行数据分析的基础。

1.1 什么是数据分析

从互联网上的词云分析中可以看到，“数据分析”这个词汇的热度很高，然而在这个喜欢炒作的年代，很多词汇和概念都是过眼云烟、昙花一现，究其原因是大多数人都没有静下心来仔细思考和踏实工作。静下来想一想“数据分析是什么”是很重要的，当人们冷静下来，再让他们解释数据分析到底是什么时，要得到一个不错的答案恐怕是很难的。

不同的人对数据分析的概念有不同的答案，比较常见的答案是，从一大堆数据中提取出想要的信息，就是数据分析。比较专业的答案是，数据分析有针对性地收集、加工、整理数据，并采用统计、挖掘技术分析和解释数据的科学与艺术。比较客观的答案是，从行业的角度看，数据分析是基于某种行业目的，有目的地收集、整理、加工和分析数据，提炼有价值信息的一个过程。

笔者认为，把数据分析看成是艺术有点过分夸张，而将其看成是过程又过于客观，但两者确实是数据分析从宏观到微观的一种很好的概括。从本质上看，要理解数据分析应从三个方面去把握：一是目标，数据分析的关键在于设立目标，专业上叫作“有针对性”，其实就是对业务需求的把握；二是方法，数据分析的方法包括描述性分析、统计分析、数据挖掘和大数据分析四种，不同的分析方法所使用的情景和功能都是不一样的，这需要在做数据分析时结合具体的情况选择使用；三是结果，数据分析最终要得出分析的结果，结果对目标解释的强弱，结果的应用效果如何。因此从这三个方面出发，我们可以给数据分析下一个概括性的定义：数据分析是指通过某种方法和技巧对准备好的数据进行探索、分析，从中发现因果关系、内部联系和业务规律等分析结果，为特定的研究或商业目的提供参考。

1.2 数据分析六部曲

概括地讲，数据分析的过程主要包括：明确分析目的和内容、数据收集、数据预处理、数据分析、数据展现和报告撰写六个步骤，如图 1.1 所示。

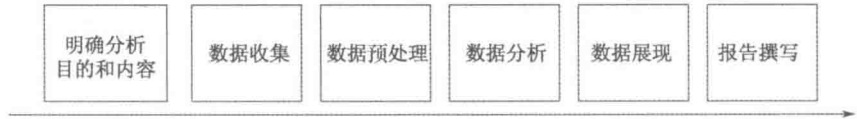


图 1.1　数据分析过程

1.2.1 明确分析目的和内容

在进行数据分析之前，数据分析师应对需要分析的项目进行详细了解，或者自己本身就对此分析项目所涉及的行业有比较深刻的了解，即使对其内部的运行规律做不到了如指掌，至少也要了解

整体框架。数据分析的对象是谁？数据分析的商业目的是什么？最后的结果要解决什么样的业务问题？数据分析师对这些问题都要了然于心。对数据分析目的的把握，是数据分析项目成败的关键。只有对数据分析的目的有深刻的理解，才能整理出完整的分析框架和分析思路，因为不同的数据分析目的所选择的数据分析方法是不同的。在企业中做数据分析时首先要明白自己想要干什么，和提出数据分析需求的部门及负责人去沟通，了解他们到底想要做什么，只有目标明确了，数据分析才能进行下去。当然，有的时候数据分析的目标不是很清晰，但肯定要有一个大致的方向，在数据分析的过程中要慢慢总结。

1.2.2　数据收集

当我们选定了数据分析的目标或大致目标之后，一个重要的问题就出现了：如何才能准确、有效地收集数据，从而客观、全面地反映所要研究的问题的真实状况。数据收集是一个按照确定的数据分析和框架内容，有目的地收集、整合相关数据的过程，它是数据分析的基础。通常数据收集的方法包括观察法、访谈法、问卷法、测验法和数据库获取法等。在商业数据分析中，数据收集一般都来源于数据库，也就是直接到数据库中获取数据，该办法需要使用到数据库工具—— SQL 语言。如今是信息化时代，任何有一定规模的企业或事业单位，都会有自己的管理信息系统，他们的商业数据都存放在数据库中，数据分析师在取得数据时，最便宜也是最方便的方法就是直接到数据库中收集数据，这就需要掌握 SQL 语言，它是数据分析中最重要的一个工具。

讲到 SQL 语言就不得不提数据库管理系统了，数据库管理系统包括两个部分，一个是数据的存储，另一个是数据的服务。数据存储一般涉及计算机领域的内容，数据分析师不用过多涉及；而对于数据的服务，数据分析师则需要了解一些基础的知识。由于数据库提供数据的服务，提供服务肯定要有服务员，而和服务员对话就需要用语言，所以 SQL 语言就是数据库提供服务的服务员所能理解的语言。这种语言有其特定的语法，学习 SQL 语言就要学习它特有的语法结构。SQL 语言的语法有很多，例如建立数据库、新建数据表、插入数据、查询数据、删除数据等，对数据分析师来讲，只需要掌握如何查询数据的语法就可以了，至于具体的查询语法这里就不叙述了，读者可查询相关书籍。也就是说，数据分析师在学习 SQL 语言时，只需要关注学习的重点，即重点学习 SQL 语言的查询语法，而无须完全掌握所有 SQL 语言的语法，即不需要成为一名优秀的数据库工程师。最后再次强调数据分析师一定要掌握 SQL 语言的查询语法，因为许多企业在招聘数据分析人才时都对这方面的技能有要求，而这也是数据收集一个非常重要的手段。

1.2.3　数据预处理

数据预处理是指对收集到的数据进行加工、整理，以便开展数据分析，它是数据分析前必不可少的阶段。数据预处理的过程概括起来包括数据审查、数据清理、数据转换和数据验证四个步骤。

第一步：数据审查

该步骤检查数据的数量（记录数）是否满足分析的最低要求，变量值的内容是否与研究目的要求一致，是否全面，包括利用描述性统计分析，检查各个变量的数据类型，变量值的最大值、最小

值、平均数、中位数等，数据个数、缺失值或空值个数等。

第二步：数据清理

该步骤针对数据审查过程中发现的明显错误值、缺失值、异常值、可疑数据，选用适当的方法进行“清理”，使“脏”数据变为“干净”数据，保证后续的数据分析得出可靠的结论。当然，数据清理还包括对重复记录进行删除。

第三步：数据转换

数据分析强调分析对象的可比性，但不同变量值由于计量单位等不同，往往造成数据不可比。对一些统计指标进行综合评价时，如果统计指标的性质、计量单位不同，则容易引起分析结果出现较大误差，再加上分析过程中其他的一些要求，需要在分析前对数据进行变换，包括无量纲化处理、线性变换、汇总和聚集、适度概化、规范化，以及属性构造等。

第四步：数据验证

该步骤的目的是初步评估和判断数据是否满足统计分析的需要，从而决定是否需要增加或减少数据量。可以利用简单的线性模型及散点图、直方图、折线图等图形进行探索性分析，利用相关分析、一致性检验等方法对数据的准确性进行验证，确保不把错误和有偏差的数据带入到数据分析模型中。

上述四个步骤是一个逐步深入、由表及里的过程。先是从表面上查找容易发现的问题（例如数据记录个数、最大值、最小值、缺失值或空值个数等），接着对发现的问题进行处理，即数据清理；然后提高数据的可比性，对数据进行一些变换，使数据在形式上满足分析的需要；最后则是进一步检测数据内容是否满足分析需要，诊断数据的真实性及数据之间的协调性等，确保优质的数据进入分析阶段。数据预处理阶段在整个数据分析过程中占据极为重要的位置，从工作量上看，它占数据数据分析全部工作量的30%~50%，因为在做数据分析时，我们根据数据分析的目标，不是一次性就能把问题解决的，而是需要反复去取数据、清洗数据，将业务逻辑转变成可被分析的量化的数据。一般的统计软件都会提供相应的功能进行数据预处理，例如SPSS软件中的数据探索功能。

1.2.4 数据分析

到了这个阶段，要想驾驭数据、分析数据，就需要选用特定的数据分析方法，熟练操作数据分析工具，实现从数据到知识的分析过程，从而解决商业问题。其一要熟悉常用的数据分析方法，最基本的是要了解例如方差、回归、因子、聚类、分类、时间序列等数据分析方法的原理、使用范围、优缺点和结果的解释；其二要熟悉“1+1”种数据分析工具，其中的一种数据分析工具是指Excel，Excel是一个最常用也是最简单的数据分析工具。现在许多公司都以Excel结合SQL做数据分析。当我们对Excel增加新的插件后，就可以进行数理统计和数据挖掘了。然而，由于Excel是一个大众化的数据分析工具，使用它进行数据分析有较多不严谨的地方，一般在学术研究中很少使用它。另一种数据分析工具是指要熟悉一个专业的分析软件，便于进行专业的数据分析、数据建模等。专业的数据分析工具主要包括SPSS、SAS、MATLAB、R等。

SPSS是世界上最早采用图形菜单驱动界面的统计软件，它最突出的特点就是操作界面极为友