

深入浅出的统计理论基础 · 63个鲜活的生活实例

帮助读者建立统计理论知识框架，形成数据分析思维逻辑，

学会数据分析方法

Broadview®
www.broadview.com.cn



人人都会 数据分析

从生活实例学统计

谢运恩 李安富 著

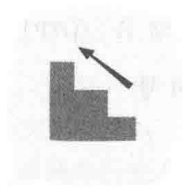
 中国工信出版集团

 电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

人人都会 数据分析

从生活实例学统计

谢运恩 李安富 著



电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

数据分析已经成为数据时代各行各业突破各自行业发展瓶颈的最有效手段,无论是公司职员还是个体商户或大公司管理者,都需要有数据分析的能力。

本书系统地介绍了数据分析的统计理论基础内容,共5章。第1章阐述了数据分析在当今生活中的重要性,以及人们成为各自领域的数据分析师的必要性和学习路径;第2章从数据描述的三个维度展开,详细介绍了如何从集中趋势、离散程度和分布形态对数据进行描述,从而使分析者充分了解自己手头的数据;第3章介绍了推断性数据统计分析的内容,介绍了如何通过样本数据特性推断出总体数据特征;第4章是关于预测分析的,介绍了变量之间的相关分析,以及如何使用容易获取的数据信息预测难以获取的数据信息,用过去的历史数据信息预测未来可能出现的数据信息;第5章介绍了数据结果可视化的内容,包括各种统计图形的功能及使用场景。

本书以数据分析的统计理论基础为主题,大多数知识点都列举了生活中的实用案例,适合高等院校学生、公司职员、个体商户和企业管理者学习参考。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

人人都会数据分析:从生活实例学统计 / 谢运恩, 李安富著. —北京: 电子工业出版社, 2017.12

ISBN 978-7-121-32966-1

I. ①人… II. ①谢… ②李… III. ①数据处理 IV. ①TP274

中国版本图书馆 CIP 数据核字(2017)第 263301 号

策划编辑: 王 静

责任编辑: 李利健

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: 18 字数: 418.2 千字

版 次: 2017 年 12 月第 1 版

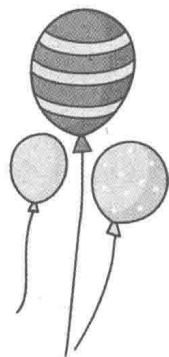
印 次: 2017 年 12 月第 1 次印刷

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 zltz@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

人人都应该会数据分析

数据分析已经成为数据时代各行各业突破各自行业发展瓶颈的最有效手段，无论是公司职员还是个体商户或大公司管理者，都需要有数据分析的能力。很多人认为数据分析能力就是对数据进行描述和做出漂亮的统计图形的能力，这是狭隘的理解。数据分析能力的基础在于数据分析的逻辑思维能力，而逻辑思维能力的培养必须建立在统计理论基础之上。

统计学是数据分析的基础

在移动互联网时代，信息的获取成本越来越低，这也导致拿来主义和实用主义盛行。数据分析科学是一门需要紧密联系生活实际的科学。数据分析方法非常多，这是因为不同的数据分析环境需要选择不同的数据分析方法来处理。因此，在数据和数据分析方法之间，天然地存在鸿沟，拿来主义和实用主义经常行不通，而磨平鸿沟的方法就是学习统计理论基础。

笔者接触过很多希望学习数据分析或正在学习数据分析的朋友，发现他们身上都有一个通病，就是将学习各种分析软件作为学习重点，觉得学习数据分析就是学习分析软件操作。他们在进行具



体的数据分析时，通常的做法是套用别人的数据分析逻辑和方法，如果分析结果与参考结果一致，那么没有问题，但是通常的情况是不一致的，这时他们就需要浪费大量的时间和精力查找原因。而不一致的原因通常是数据情况不符合导致的，解决的办法是根据实际的数据分析环境进行数据调整或选择其他合适的数据分析方法，而要做到这些，归根结底需要的还是扎实的统计理论基础。

学习数据分析应该以统计理论基础为主线，分析软件仅仅是各种分析方法的实现工具。因此，在开始学习时，应该辅以一到两种操作简便的分析软件，例如 SPSS。这样就能够比较快速地建立统计理论框架，并掌握一款分析软件。当你的统计理论基础夯实以后，会发现学习各种分析软件都能够游刃有余，甚至像 R 和 Python 这种编程语言也能在比较短的时间内掌握。

从生活实例中开始学统计学

标准版的统计教材往往偏重于理论推导，描述语言专业严谨，这不可避免地带来了晦涩难懂的问题。本书的写作初衷就是为了解决没有任何统计基础的各行各业的数据分析学习者，他们学习和构建自己统计理论基础的需求，大部分知识点都列举了生活中的应用案例，能够帮助学习者更好地理解和掌握各个知识点的分析逻辑，并内化为自己的统计理论基础的知识框架。

除此之外，本书在章节设计上做了精心的安排，共包括 5 章内容。

第 1 章阐述了数据分析在当今生活中的重要性，以及人们成为各自领域的数据分析师的必要性和学习路径。

第 2 章从数据描述的三个维度展开，详细介绍了如何从集中趋势、离散程度和分布形态对数据进行描述，从而使分析者充分了解自己手头的数据。

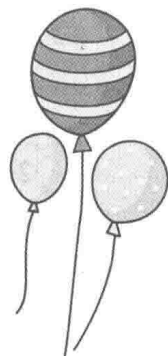
第 3 章为推断性数据统计分析的内容，介绍了如何通过样本数据特性推断出总体数据特征。

第 4 章为预测分析内容，介绍了变量之间的相关分析，以及如何使用容易获取的数据信息预测难以获取的数据信息，用过去的历史数据信息预测未来可能出现的数据信息。

第 5 章是数据结果可视化的内容，包括各种统计图形的功能及使用场景。

“不积跬步，无以至千里”，夯实基础的过程总是比较乏味的，希望本书通俗的描述语言和鲜活的生活应用案例能够帮助数据分析学习者顺利建立统计理论的知识框架，形成数据分析思维逻辑，成为各自领域的佼佼者。

作者



目录

第 1 章 生活在数据时代.....	1
1.1 数据分析无处不在	2
1.1.1 常用的国家统计指标.....	2
1.1.2 制造业的数据分析应用	9
1.1.3 营销领域的数据分析应用	13
1.1.4 医疗行业的数据分析应用	15
1.2 人人都能成为数据分析师	16
1.2.1 数据分析过程	17
1.2.2 数据分析工具	21
1.2.3 数据分析师的成长之路	26
第 2 章 耳熟能详的数据你真的了解吗	29
2.1 数据的类型	30
2.1.1 数据的结构属性分类	30

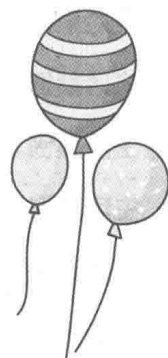
2.1.2	数据的连续性特征分类	31
2.1.3	数据的测量尺度分类	33
2.2	数据描述的三个维度	35
2.3	数据的集中趋势描述	36
2.3.1	算术平均值	37
2.3.2	几何平均值	39
2.3.3	众数	40
2.3.4	中位数	41
2.4	数据的离散程度描述	42
2.4.1	极差	42
2.4.2	平均偏差	43
2.4.3	方差和标准差	44
2.4.4	变异系数	48
2.4.5	四分位极差	49
2.5	数据的分布形态描述	50
2.5.1	概率	50
2.5.2	概率分布	53
2.5.3	离散型概率分布：二项分布	54
2.5.4	离散型概率分布：多项分布	56
2.5.5	离散型概率分布：超几何分布	57
2.5.6	离散型概率分布：泊松分布	59
2.5.7	连续型概率分布：指数分布	62
2.5.8	连续型概率分布：均匀分布	65
2.5.9	连续型概率分布：正态分布	66
2.5.10	正态分布作为二项分布近似	73
2.5.11	正态分布作为泊松分布近似	76
第3章	数据分析的“内核”：推断分析	79
3.1	见微知著的抽样	80
3.1.1	抽样的意义	80
3.1.2	抽样方法	81
3.1.3	样本推断的理论基础	84
3.2	数据的处理	86



3.2.1	数据处理不良案例	86
3.2.2	正确的数据存储形式	87
3.3	样本到总体的桥梁: 抽样分布	88
3.3.1	抽样分布的定义	88
3.3.2	Z分布	90
3.3.3	T分布	95
3.3.4	切比雪夫定理	98
3.3.5	卡方 (χ^2) 分布	99
3.3.6	F分布	100
3.4	数据分析的第一板“斧”: 参数估计	102
3.4.1	参数估计的类型	102
3.4.2	Z分布与总体均值的区间估计	104
3.4.3	T分布与总体均值的区间估计	110
3.4.4	切比雪夫定理与总体均值的区间估计	113
3.4.5	卡方 (χ^2) 分布与总体方差的区间估计	115
3.4.6	F分布与两个总体方差比的区间估计	119
3.4.7	两个总体均值差的区间估计	121
3.4.8	总体比率的区间估计	133
3.4.9	样本容量的确定	135
3.5	数据分析的第二板“斧”: 假设检验	142
3.5.1	假设检验的理论基础	142
3.5.2	单样本的假设检验	150
3.5.3	两样本的假设检验	159
3.5.4	多样本的假设检验与方差分析	166
3.6	数据分析的第三板“斧”: 非参数检验	184
3.6.1	非参数检验	185
3.6.2	卡方检验	187
3.6.3	Wilcoxon 符号秩检验	191
第4章	数据分析的终极目的: “为我所用”	195
4.1	“相关”是继续分析的前提	196
4.1.1	相关关系	196
4.1.2	相关分析	198



4.2	“回归”是相关分析的归宿	209
4.2.1	回归分析综述	209
4.2.2	简单线性回归分析	211
4.2.3	多元线性回归分析	226
4.3	发现事物随时间变化的规律	235
4.3.1	时间序列分析综述	235
4.3.2	长期趋势分析	238
4.3.3	季节变动趋势分析	248
4.3.4	循环变动和不规则变动	252
4.3.5	时间序列分析应用	254
第 5 章	给数据披上靓丽“外衣”：数据可视化	259
5.1	数据的可视化	260
5.1.1	数据可视化工具	260
5.1.2	常用的统计图	261
5.2	条形图、直方图和饼图	262
5.2.1	条形图	262
5.2.2	直方图	264
5.2.3	饼图	266
5.3	线图与面积图	267
5.3.1	线图	267
5.3.2	面积图	267
5.4	散点图	268
5.5	高低图与箱线图	270
5.5.1	高低图	270
5.5.2	箱线图	271
5.6	时间序列图	272
5.7	帕累托图	274
5.8	P-P 概率图和 Q-Q 概率图	275
5.8.1	P-P 概率图	275
5.8.2	Q-Q 概率图	276



案例目录

生活案例：使用算术平均值了解公司的待遇水平	37
生活案例：使用几何平均数识别生产线上的隐形损耗	39
生活案例：使用众数制订服装企业的生产计划	40
生活案例：使用中位数识破招聘启事的工资陷阱	41
生活案例：使用极差描述气温的变化幅度	42
生活案例：使用平均偏差评价生产线的稳定性	43
生活案例：使用标准差（方差）提高巧克力生产线的工艺水平	47
生活案例：使用变异系数客观评定员工绩效	49
生活案例：概率是赌场老板制定游戏规则的法宝	52
生活案例：产品抽检的二项分布应用	55
生活案例：预测足球比赛结果的多项分布应用	56

生活案例：超市抽奖活动的超几何分布概率	58
生活案例：使用泊松分布进行杂货店的库存管理	60
生活案例：使用二项分布和泊松分布分析准备的试卷数量是否足够	62
生活案例：指数分布告诉你，为什么电子产品只保修三年	64
生活案例：使用均匀分布分析家具物流的送货时间	66
生活案例：使用正态分布分析研究生毕业论文的完成时间	69
生活案例：使用峰度与偏度检验政府精准扶贫效果	72
生活案例：使用正态分布和二项分布分析足球比赛的猜边概率	75
生活案例：使用正态分布和泊松分布分析民营电缆厂的良品率	76
生活案例：估算西部儿童看电视时间的置信区间	107
生活案例：估算大学生每日体育锻炼时间的置信区间	108
生活案例：购买保险公司产品的消费者年龄置信区间	109
生活案例：酱料厂罐装产品重量的置信区间	112
生活案例：心脏搭桥手术后，患者住院时间的置信区间	114
生活案例：使用 T 分布和 Z 分布分析食品包装上的营养成分表	117
生活案例：比较不同超市的电子秤精度	120
生活案例：比较水稻新品种在不同地区的产量	124
生活案例：比较不同睡眠时间对记忆力的影响	127
生活案例：研究新药对睡眠质量的影响	128
生活案例：比较两家航空公司的起飞延误时间	129
生活案例：蒲公英的降脂作用	132
生活案例：移动支付的普及程度调查	135
生活案例：管理制药厂的供应商	137
生活案例：出口苹果的重量控制	139



生活案例：电视节目的收视率调查	140
生活案例：使用单样本 Z 检验分析笔记本电池的续航能力改进效果	154
生活案例：使用单样本 T 检验分析新型麻醉剂的麻醉效果	156
生活案例：使用卡方检验管理饮料工厂的产品质量	158
生活案例：机场对不同航空公司航班的起飞时间管理	161
生活案例：新药的临床实验	162
生活案例：新药的副作用研究	164
生活案例：自动化生产线的工艺调整	165
生活案例：使用单因素方差分析研究新水稻品种的性状	171
生活案例：使用无交互两因素方差分析研究员工的绩效管理方式	175
生活案例：使用有交互两因素方差分析研究员工绩效的管理方式	181
生活案例：使用卡方检验分析讲座的到场人数是否存在异常波动	188
生活案例：使用卡方检验分析不同年龄人群对汽车颜色的喜好	190
生活案例：使用 Wilcoxon 符号秩检验分析职业影评人的打分差异	192
生活案例：使用相关分析理解“总理指数”	201
生活案例：使用偏相关分析研究上网时间是否真的影响学习成绩	205
生活案例：使用距离相关分析研究汽车的性价比	206
生活案例：简单线性回归分析用于货运总量预测国内生产总值	216
生活案例：货运总量预测国内生产总值的置信区间	221
生活案例：货运总量与国内生产总值回归模型的效果检验	224
生活案例：使用多元线性回归分析优化快递业务员的路线	229
生活案例：使用移动平均法预测玩具公司的销售量	239
生活案例：使用加权移动平均法预测玩具公司的销售量	240
生活案例：使用指数平滑法预测玩具公司的销售量	242

生活案例：使用指数平滑法预测玩具公司的销售量	244
生活案例：使用回归方程法预测创业公司的销售额	246
生活案例：创业公司销售额的季节变动趋势分析	250
生活案例：公司销售量的循环变动分析	252

读者服务

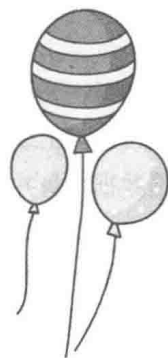
轻松注册成为博文视点社区用户 (www.broadview.com.cn), 扫码直达本书页面。

★ **提交勘误**: 您对书中内容的修改意见可在 [提交勘误](#) 处提交, 若被采纳, 将获赠博文视点社区积分 (在您购买电子书时, 积分可用来抵扣相应金额)。

★ **交流互动**: 在页面下方 [读者评论](#) 处留下您的疑问或观点, 与我们和其他读者一同学习交流。

页面入口: <http://www.broadview.com.cn/32966>





第 1 章

生活在数据时代

二百多年前，英国批判现实主义小说家狄更斯曾经说过：“这是一个最好的时代，这也是一个最坏的时代。”这句话放在今天，依然适用。随着科技的进步，计算机技术的成熟以及移动互联网的普及，我们已经步入大数据时代。在大数据时代，我们每个人的信息都是透明的，个人偏好、行为习惯以及社会关系都可以被捕捉，这促进和发展了广告精准推送以及服务量身定制等商业模式。这些新的商业模式让用户获得更好的个人体验的同时，也直接戳向了传统商业模式的软肋，使传统企业叫苦不迭，以至于银行都抱怨自己是弱势群体。由此可见，这个时代对于固守传统商业模式和经营理念的行业及厂商来说，无疑是最坏的时代。大数据时代的到来并非是一个非好即坏的结局，如果选择故步自封，那么必然会被这个飞速变化的时代所抛弃；如果选择审时度势，顺应时代发展的需求并主动求变，享受大数据时代带来的普惠成果，那么就有很大的机会进一步跨越发展。

1.1 数据分析无处不在

生活在大数据时代，能够被记录和分析的数据无处不在。特别是随着智能手机以及各种可穿戴智能设备的出现，每个人的行为、位置、偏好，甚至各项生理指标都成了可被记录和分析的数据。但是数据本身并不能产生价值，只有经过有效的数据分析过程，隐藏在大量数据背后的富含价值的信息才会展现在我们面前，从而帮助我们做出正确的行为决策，带给我们巨大的价值回报。现如今，各行各业都逐渐意识到数据统计及分析的重要性，数据分析的应用已经渗透到人们生产和生活的各个领域。无论是政府机构、机关企事业单位、私营工厂，还是家庭、个人，数据分析都扮演着越来越重要的角色，发挥着重要的作用。

1.1.1 常用的国家统计指标

政府对国家的管理需要关注到社会生活的方方面面。从人们的衣食住行到国家的内政外交，政府对于每项政策的制定和决策的实施，都要依赖于大量的统计调研和数据分析。国家统计局是主管全国统计工作的政府机构，对国民经济和社会发展情况进行统计分析、统计预测和统计监督，为政府对国家的有效管理提供统计数据支持。除此之外，国家统计局还会定期发布全国国民经济和社会发展情况的统计数据报告，人们可以通过报告中的统计指标对整个国家的运行情况进行了解。经济和人口指标是很重要的国家运行情况指标，它们可以反映整个国家的经济发展水平和人民的生活状况。下面我们就介绍几种常用的国家统计指标，帮助大家更好地理解我们国家的运行情况。

1. 国内生产总值

国内生产总值（Gross Domestic Product）的简称是我们熟悉的GDP。国内生产总值是指在一定时期内，一个国家或地区的经济中所生产出来的全部最终产品和劳务的价值，它是衡量国家经济状况最常用的指标之一。国内生产总值有三种表现形态：价值形态、收入形态和产品形态。从价值形态看，它是所有常住单位的价值增加值之和；从收入形态看，它是所有常住单位在一定时期内创造并分配给常住单位和非常住单位的初次分配收入之和；从产品形态看，它是所有常住单位在一定时期内最终使用的产品价值与净出口的产品价值的总和。图 1-1 显示的是我国从 1995 年到 2014 年间的国内生产总值及其增长率的变化情况。从图中可知，从 1999 年到 2007 年，我国的国内生产总值增长率都是逐年递增的，并在 2007 年达到峰值，年增长率达到 14.2%。此外，国内生产总值也是逐年增加，从 1995 年的 61339.89 亿元增长到 2014 年的 643974 亿元，20 年间增长了 10.5 倍，成为仅次于美国的世界第二大经济体。我们都切身体会到了祖国快速发展给我们的日常生活带来的极大变化。

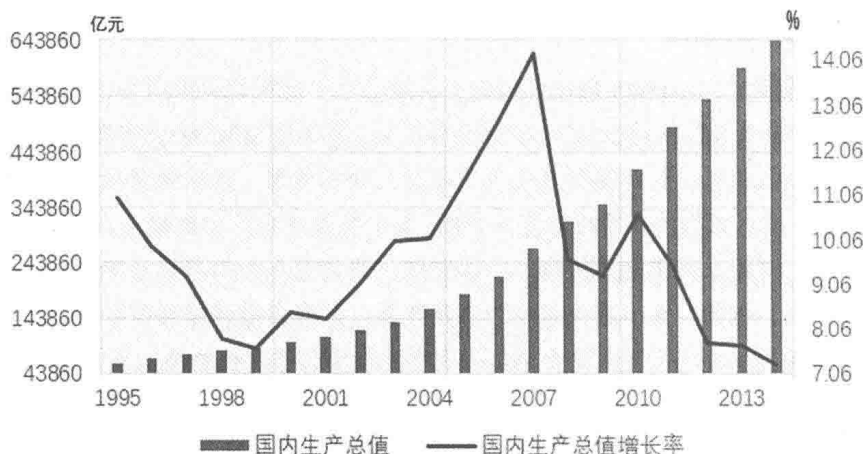


图 1-1 国内生产总值趋势变化图

2. 国民总收入

国民总收入（Gross National Income，简称为GNI）。国民总收入是在1993年由国民生产总值（Gross National Product，GNP）改称而来的。国民总收入是指一个国家或地区的国民在一定时期内，在国内和国外所生产的最终产品及劳务价值的总和。例如，我国有很多国民在海外工作并取得报酬，这部分收入属于国民总收入。国民总收入常用于反映一个地区或国家居民的富足情况。图1-2显示的是我国从1995年到2014年间国内生产总值和国民总收入的变化情况，可以发现，国内生产总值和国民总收入之间是高度相关的，数值相差很小，由此可以推断，我国的国内生产总值大部分是由我国公民自己创造的。中国人民是世界上最勤劳的人民，改革开放以后，中国人民创造了令世界瞩目的发展奇迹。

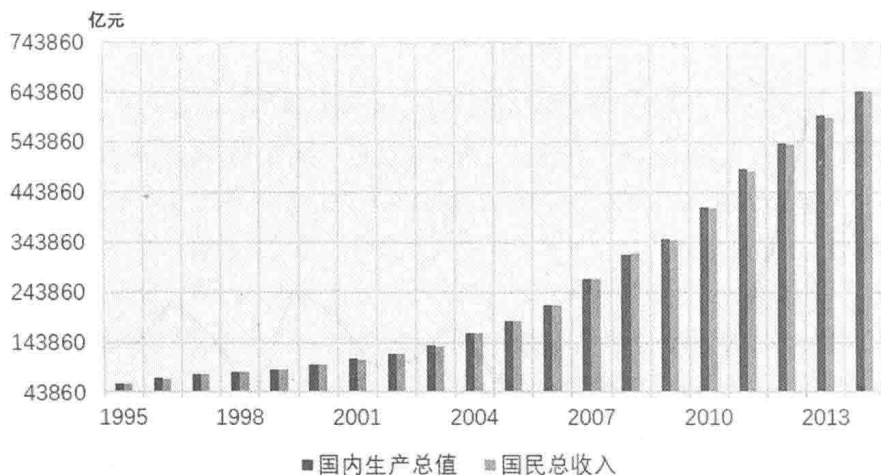


图 1-2 国民总收入及国内生产总值柱形图



3. 居民消费价格指数

居民消费价格指数 (Consumer Price Index) 就是这几年经常提及的CPI, 它是反映市场物价变化情况的基本术语。每个国家统计CPI的方法都略有区别, 我国的CPI指数是按食品、烟酒及日用品、服装、家庭设备用品及服务、医疗保健及个人用品、交通和通信、娱乐教育文化用品及服务 and 居住这八大类来计算的。这八大类的权重相加定为 100, 其中食品所占的比重最大, 包括粮食、肉禽及其制品、蛋、水产品、鲜菜、鲜果等具体商品。国家统计局会在八大类消费品中分别选出一种或几种最具有代表性的商品, 例如, 大多数人是吃米还是吃面, 是穿皮鞋还是穿布鞋等, 把这些代表性商品的物价按每月、每季、每年折算成物价指数, 定期向社会公布, 这就是人们常说的CPI指数。

CPI是反映城乡居民消费水平和消费品价格变动情况的重要指标, 也是观察通货膨胀水平的重要指标。如果CPI在过去的一年中上升了 6.8%, 就表示当下的生活成本比一年前平均要高出 6.8%, 这无疑是不受欢迎的。随着生活成本的提高, 人们口袋里钱的价值便会随之降低, 也就是说, 去年的 100 元纸币到今年就只能购买到价值 93.2 元的商品或服务。CPI涨幅过大表明货币贬值幅度过大, 是通货膨胀的表现, 而通货膨胀可能成为经济不稳定的因素。政府能够通过CPI了解国家的物价情况, 如果CPI涨幅过大, 政府可以通过实施紧缩的货币政策和财政政策来抑制通货膨胀, 保证国家经济的稳定。

经济的高速增长不可避免地会拉高CPI, 因此, 需要在保持经济增长的前提下将CPI控制在合理的范围内, 国际上一般认为CPI在 2%到 3%属于可接受范围。图 1-3 是我国自 1995 年以来的居民消费价格指数变化图, 从图中可以看到, 1995 年和 1996 年, 2007 年和 2008 年, 以及 2011 年的CPI都超过或接近 5%, 处于较高水平。离现在比较近的 2007 年和 2008 年, 我们能够很明显地感觉到物价的快速上涨, 例如, 方便面从原来的 1.5 元/包涨到了 2 元/包; 原来 500 元/月的大学生生活费已经不够用了。

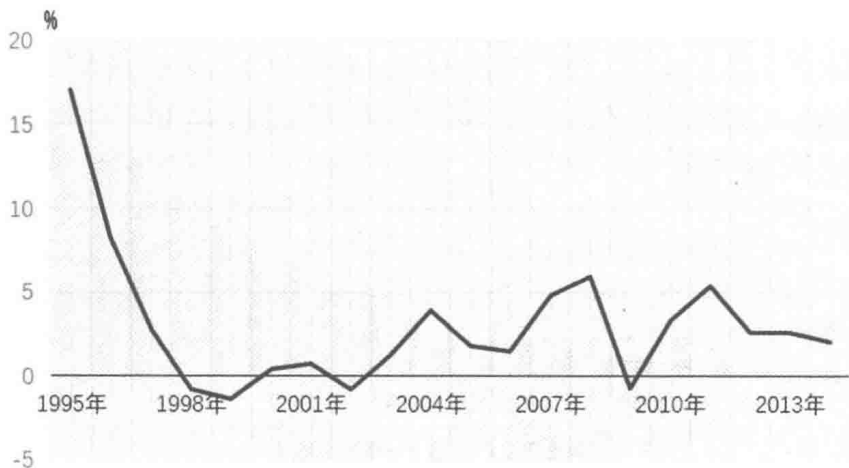


图 1-3 居民消费价格指数变化图