HZ BOOKS
华章教育

21世纪经济学系列

Modern Applied Statistics: Big Data Analysis Base

现代应用统计学

大数据分析基础

王建军 宋香荣 编著

机械工业出版社
China Machine Press

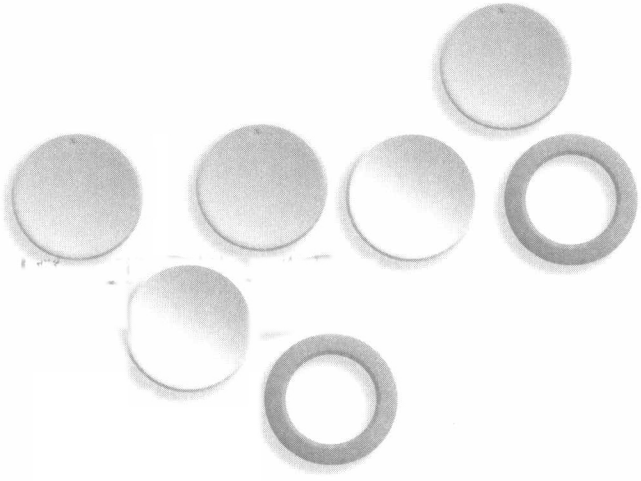
21世纪经济学系列

Modern Applied Statistics: Big Data Analysis Base

现代应用统计学

大数据分析基础

王建军 宋香荣 编著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

现代应用统计学：大数据分析基础 / 王建军，宋香荣编著．—北京：机械工业出版社，2016.8
（21 世纪经济学系列）

ISBN 978-7-111-53962-9

I. 现… II. ①王… ②宋… III. ①应用统计学 ②统计数据－统计分析 IV. ①C8 ②O212.1

中国版本图书馆 CIP 数据核字（2016）第 122228 号

本书将常用软件 Excel 与统计专业软件 R 语言有机结合为一体，突出应用统计特点和当代数据分析的要求，体系上具有最新统计理念和实用统计方法的特点。

本书将统计学定义为分析数据信息的科学，用专门章节阐述了统计学中最重要的两个概念——数据与变量，详尽描述数据特点与分类，为以后各章统计分析方法作为铺垫。本书的作者极为重视变量，认为统计学是从变量开始确定搜集数据，选择数据，而目前各类统计学教材忽视变量，主要原因是从数学的角度看统计学的变量，变量则是可以抽象为某个符号，而符号恰恰又是极不重要的，可以用任何字母代替的；数学将变量分为统计学认为并不那么重要的连续变量与离散变量、确定性变量与随机性变量等；这类变量的分类法对统计分析方法的选择无任何实际意义。而应用统计学解决问题，首先是从变量入手，先有变量，然后才有数据。即使是先有数据，统计学也是从变量的需要选择数据。

目前多数统计学的课程体系是从数学的逻辑体系排列的，本书打破传统，突出统计方法的实用，将统计方法分为三类：探索性统计分析、验证性统计分析和统计预测。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：方　琳　　责任校对：殷　虹
印　　刷：北京瑞德印刷有限公司　　版　　次：2016 年 9 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：15.5
书　　号：ISBN 978-7-111-53962-9　　定　　价：35.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010) 68995261　88361066　　投稿热线：(010) 88379007
购书热线：(010) 68326294　88379649　68995259　　读者信箱：hzjg@hzbook.com

前 言

现代应用统计学为适应大数据时代的要求将统计学最常用的方法分为探索性统计分析、验证性统计分析和预测分析三类。本书将统计学中传统的参数统计与现代非参数统计方法有机结合为一体，将常用软件 Excel 与统计专业软件 R 语言有机结合为一体，突出应用统计特点和当代数据分析的要求，体系上具有最新统计理念和实用统计方法的特点。

目前多数统计学的课程体系是根据数理统计逻辑体系排列的，本书打破传统，将统计体系按可能遇到的问题排列，形成问题导向型的体系，突出统计方法的实用性，分为统计学基础概念、探索性统计分析、验证性统计分析和预测分析。

本书将统计学定义为分析数据信息的科学，用专门章节阐述了统计学中最重要、最基本的两个概念——数据与变量，详尽描述数据特点与分类，为以后各章统计分析方法的选择作铺垫。本书作者极为重视变量，认为统计学是从变量开始确定研究现象的量化，然后才搜集数据、选择数据，而目前多数统计学教材忽视变量，主要原因是从数学的角度看统计学的变量。变量既可以抽象为某个符号（而符号恰恰又极不重要），又可以用任何字母代替。数学将变量分为统计学认为并不那么重要的连续变量与离散变量、确定性变量与随机性变量等。这种变量的分类法对统计分析方法的选择并无多少实际意义。本书作者认为应用统计学解决问题，首先是从变量入手，先有变量，然后才有数据。即使是先有数据，统计学也是根据变量的需要选择数据。

本书按现实中的问题类型将探索性统计分析分为一个定性变量探索、一个定量变量探索、两个定性变量与两个定量变量关系、多个数值变量关系的探索分析。验证性统计分析分为一个变量的分布验证分析、两个变量关系的验证分析、多个变量关系的统计分析。将统计预测方法单独列出，显示其极为重要。

经典统计源自数学，置信区间理论显得重要，95% 的置信区间的准确含义是“抽 100 次样本，有 95 次样本构造的置信区间可能包含总体参数”。可是实际上只有一次抽样，总体参数是否包含在此次置信区间内并不知道。实际应用最多的是点估计，用样本均值作为总体均值的点估计，样本比例作为总体比例 P 的点估计是经过数学极大似然法

证明的。其实，区间估计也是以样本均值为中心构造置信区间的，为了提高置信度，置信区间的半径估计过大从而失去实用的参考价值，置信区间过大对数学并无任何影响，对结论应用的统计学就显得无所适从了。

验证性统计分析对用归纳法得出的理论假设进行检验，强调的是现实数据是否支持理论假设，是否得到与理论一致的结论。验证方法分为经典参数统计与非参数统计，本书将其合并使用，验证性统计方法与计量经济理论检验是一致的，所以本书中融入了一些计量经济学的基本理念，强调统计检验的理论分析，并先分析为什么会有影响，再用统计数据验证。这样避免了盲目套用模型，也免得出现极为荒谬的伪回归模型。

本书极为重视预测分析，将回归模型预测单独列出，与时间序列预测合并为一篇。截面数据回归模型预测插入值，时间序列预测外推值，各有应用范围和价值。

统计学中的数学原理讲到多深才够？通常一般的教材为了体系上的逻辑需要从概率论讲起，而本教材认为大学的高等数学部分已经讲过，就不必简单而无用地重复，同时验证性统计分析在应用统计学中只用到结论就行了，不必再强调数学体系的完整。另外，统计调查中抽样调查的理论需要讲多少才够呢？其实很多统计学教材会讲到抽样的各类方法，如简单随机抽样、分层抽样、整群抽样、等距抽样，等等。然而这些方法非一两节课能讲清楚，主要是因为估计的方差理论不是简单推导可得的，所以作者认为花费大量篇幅讲这些看似重要，但只有专业统计人士才有机会用到抽样调查理论知识，大可不必。因此在本教材中一概略去，特别是在统计调查方法部分。

统计学应用主要是将数学原理与实际问题相结合，其中复杂的计算部分已由专业统计软件完成。本教材在众多的统计软件中，选择的是 R 软件与 Excel 相结合，其出发点是 Excel 办公软件普及率高，探索性统计分析计算功能较强，可完成大部分统计计算。同时复杂的统计计算由 R 软件完成，主要是因为 R 软件除了具有免费的优势而且功能强大，普及已成趋势，为了赶上时代发展，所以本书也采用 R 软件。全书更加体现统计算法为主的现代统计学理念。

统计学是一门科学与艺术的学问，其中艺术性是指需要统计分析的经验。本书的一个重要特点就是强调实践的经验部分，突出实践中容易错误的地方，放弃了一些似乎有用、其实无用的统计概念与方法，如离散系数、峰度与偏度、置信区间，同时吸取了一些现代统计知识，如稳健性的方法、截尾平均、MAD、中位数回归等。值得一提的是，在面临多种统计方法供选择时，作者也进行了比较说明。

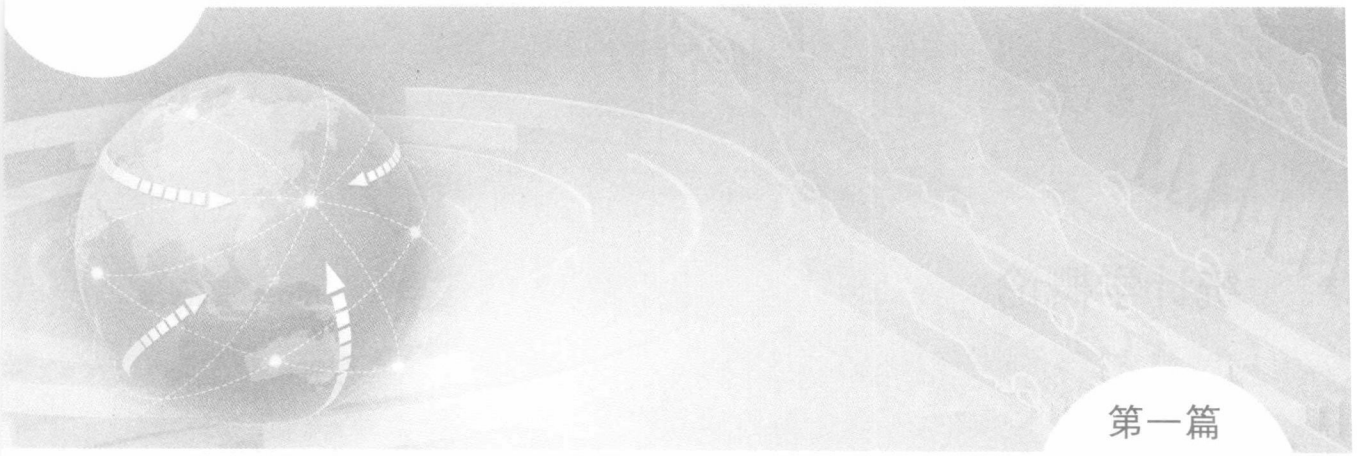
本书创新性地以大数据分析为基础，难免观点不妥，恳请广大读者不吝赐教，以便及时更正。书中数据可通过邮箱 xjcdtjx@sina.com 与作者联系获得。

本书获得“新疆高等学校地方特色和民文教材项目”的推荐与资助。

2016 年 1 月

目 录

第二篇 探索性统计分析与验证性统计分析



第一篇

统计学基础概念

第1章

统计学概论

1.1 统计学的定义

统计学有多种定义，比较各种定义更能准确理解统计学。

1. 按统计的任务目标定义

统计学（statistics）是分析数据信息的科学，是从数据中得到信息的方法论科学。

这个定义强调了统计学的研究对象是数据，统计学的目的是分析数据得到信息，最终目标是为了得到有用的信息，有信息才能体现统计学的价值，这就是“有为才有位”，有价值就有市场需要，分析数据是统计的核心工作，得到有价值的信息是统计的目标任务。

2. 按统计的工作过程定义

统计学是搜集、整理、分析、解释数据并从数据中得出结论的科学。

这个定义目前最流行，主要是从统计工作的全过程来定义的，这个定义着重强调了统计工作的搜集、整理、分析、解释数据过程。

政府统计部门主要工作仍是准确、及时地搜集数据。搜集的数据要有代表性，结论才有意义；分析数据要依据数据的特点选择方法，充分挖掘数据所蕴含的未知信息；要正确理解并解释结论的含义。从数据中得出结论，即要得到有实际意义的信息。这个定义全面、准确，但是并行罗列全过程，似乎各个过程一样重要，其实统计学最重要的环节是分析数据，故有忽视重点分析数据的弊端。

3. 按统计研究对象的特征定义

统计学是一门处理数据中变异性的科学和艺术。

统计学是对不确定性问题做决策的科学。

统计学就是研究数据及其存在规律的科学。

统计学是用数据说话的学问。

这些定义别具特色，也抓住了统计学是分析数据变异性、随机性、不确定性，即研

究数据变化的本质。统计学确实是分析数据的变化的，分析变化的程度、变化的方向、变化的原因。特别强调了统计学是科学和艺术，科学是指统计学有数学理论基础，有严谨的逻辑推理和理论体系，统计学的艺术性指的是统计学方法应用需要根据具体问题具体对待，需要积累经验，总结客观规律，需要根据具体的任务，根据掌握数据的特征、实践经验和专业知识，选择正确的统计分析方法，才能得到较为可靠的结论。统计学有章法可依，而无定法可循，贵在得法，不能照搬书本。

4. 理解统计学的定义要点

（1）统计学是关于数据的科学。有些偏重于搜集数据，有些偏重于分析数据得到信息。

（2）统计学强调实践性。统计学是一门应用学科，既有理论的科学性，也有实践应用的主观经验和艺术性。

（3）研究统计学的视角不同，重点不同，定义就不同。统计是处理数据的一门科学，它所提供的是一套有关数据搜集、处理、分析、解释并从数据中得出结论的方法，其目的是探索数据的内在数量规律性，以达到对客观事物的科学认识。统计学应用就是实现目的主观性、对象客观性与方法科学性三者的统一。

（4）统计的核心是分析数据。要以统计数据为基础，要用统计分析方法为工具，找出信息，分析变化的原因和趋势。统计学是用概率的语言表述现象，通过利用概率论建立数学模型，搜集所观察的数据，进行量化的分析，并进而进行推断和预测，为相关决策提供依据和参考。

（5）统计学也称为数据科学（data science）。数据科学的三个功能：分析数据来预测未来（predictive analytics）；分析数据找出特征的描述分析（descriptive analytics）；分析数据找出最佳措施和最优化的结果（prescriptive analytics ）。

1.2　统计学的作用

1. 统计学用数据表述思想

统计学是用数据说话的学问。

例如，2013 年我国 GDP 总量是 92 403 亿美元，位居世界第二位，很明显中国整体经济实力强大，但从人均 GDP 来看，我国仅为 6807 美元，低于全球平均数 10 513 美元，我国仍属于发展中国家。用不同数据就会产生不同的效果，也说明统计学是用数据说话的学科。（国际统计年鉴 2014）

数据说明中国的人均支出横向看是呈现增长趋势，纵向比较属于低水平（见表 1-1）。

表 1-1 数据比较中国经济水平

居民人均消费支出（单位：美元）				
年份	2000 年	2005 年	2010 年	2013 年
世界	3 247	3 514	3 625	4 578
高收入国家	15 419	16 785	16 931	19 048
中等收入国家	783	928	1 169	1 545
低收入国家	201	226	263	314
中国	439	582	879	1 307

表 1-2 的数据说明西部地区的人均收入与支出明显低于全国平均水平，与东部地区相比较经济发展差距很大，说明西北五省经济发展落后。

表 1-2 2013 年人均收支指标比较

地区	人均可支配收入（元）	人均消费支出（元）	地区	人均可支配收入（元）	人均消费支出（元）
全国	18 310.8	13 220.4	陕西	14 371.5	11 217.3
江苏	24 775.5	17 925.8	甘肃	10 954.4	8 943.4
浙江	29 775.0	20 610.1	青海	12 947.8	11 576.5
福建	21 217.9	16 176.6	宁夏	14 565.8	11 292.0
广东	23 420.7	17 421.0	新疆	13 669.6	11 391.8

资料来源：《中国统计年鉴 2014》。

2. 统计学用数据描述现状

要想做到心中有“数”，只有通过数据才能准确了解真实情况，才能分析掌握动态规律，才能弄清各种关系，才能对症下药、从容处之。

例：中国 2007 年每公顷化肥施用量 306 kg。据世界粮农组织（FAO）统计分析，目前世界平均每公顷耕地化肥施用量约为 120 kg，美国为 110 kg，德国为 212 kg，日本为 270 kg，英国为 290 kg，荷兰为 623 kg。我国化肥总用量和单位面积用量已经处于世界较高水平。

3. 统计学用数据分析规律

① 研究变量间关系的变化规律

主要用统计模型描述规律，特别是用计量经济学的模型描述经济规律。

如：研究家庭收入如何决定消费支出的数学规律。

如：研究资本存量 K 和劳动力数量 L 如何影响生产总值 GDP 的规律。

② 研究趋势变化规律。

如利用表 1-3 中的数据研究中国城镇化的变化规律，可以通过图形展示趋势变化。

表 1-3 中国城镇化率趋势 （%）

年份	1978	1990	2000	2005	2010	2013
城镇	17.92	26.41	36.22	42.99	49.95	53.73
乡村	82.08	73.59	63.78	57.01	50.05	46.27

4. 统计学用数据辅助决策

（1）用数据选择决策方案。

在决策中人们希望收入高，成本少，风险小，就需要用统计数据在各类方案中选择最优、次优的方案，统计分析结果就起到了辅助决策的作用。在目前的大数据时代，现场数据通过物联网、互联网瞬时反馈到云计算中心，数据信息辅助决策系统及时提醒显示，为决策者采取正确措施提供了数据依据。

（2）用数据划分分类标准。

如贫困线是划分贫困人口的标准，低于贫困线的就是贫困人口，2012 年的贫困线标准为年人均收入 2300 元。

（3）用数据制定规划目标。

如用数据表示“十三五”的发展目标。

如用数据制定和谐社会的衡量标准，用数据衡量小康基本标准。

全面建设小康社会统计监测指标体系

监测指标	单位	标准值（2020 年）
人均 GDP	元	≥ 31 400
R&D 经费支出占 GDP 比重	%	≥ 2.5
第三产业增加值占 GDP 比重	%	≥ 50
城镇人口比重	%	≥ 60
失业率（城镇）	%	≤ 6
基尼系数	—	≤ 0.4
城乡居民收入比	以乡村居民收入为 1	≤ 2.8
基本社会保险覆盖率	%	≥ 90
居民人均可支配收入	元	≥ 15 000
恩格尔系数	%	≤ 40
人均住房使用面积	平方米	≥ 27
5 岁以下儿童死亡率	‰	≤ 12
平均预期寿命	岁	≥ 75
居民文教娱乐服务支出占家庭消费支出比重	%	≥ 16
平均受教育年限	年	≥ 10.5

5. 统计学用数据预测未来

统计得到的数据均是过时的、曾经的数据，但人们关心的、想要的是关于未来的预测数据，这就需要用到统计预测方法。时间序列就是研究预测的重要方法（参照第 10 章）。

6. 统计学用样本推断全体

要想得到真实的数据只有两个方法：一个是普查，另一个是随机抽样调查。抽样调查就是用少数随机抽到的样本代表总体，用科学的推断方法得到真实的总体数据。

抽样调查的优点是：① 时效性高。短期可以完成。② 调查费用少。抽样调查是针对总体中的一部分单位进行的，抽样调查可以大大减少调查费用，提高调查效率。③ 调查的完整性强。抽样调查可以减少调查的工作量，调查内容可以求多、求全或求专，可以保证调查对象的完整性。④ 可以从数量上以部分推算总体，利用概率论和数理统计原理，以一定的概率保证推算结果的可靠程度，起到全面调查认识总体的功能，保证调查的精度。

7. 统计学用数据验证假设

现实中经常会遇到下列问题，需要检验某药品的治疗效果是否有效，检验某项新工艺是否提高了劳动生产率，等等，这类问题属于用统计数据验证假设问题。只有用统计数据的变化才能验证假设是否成立。

例如，研究某减肥方法是否有效，研究某药品治疗高血压是否有效，就需要比较减肥（用药）前后的数据变化是否显著。

例如，研究工资待遇方面是否存在性别歧视，需要比较男女收入数据，当差距很大时才能说明性别歧视存在。

8. 统计学用数据发现理论假设

例如，通过大量观察数据发现，人们提出下列理论假设“对司机的安全保护提升了，路人的安全性就会降低”。规定汽车必须装安全带的制度是为了减少车祸伤亡，但在安全带的保护下，司机将车开得更快，事故反而增加了。调查数据显示司机有安全带保护，自身伤亡减少了，而路人伤亡增加了。

9. 用数据确定正常值范围

识别数据正常波动还是异常，是进行统计分析的前提，正常波动一般不需要进行统计分析，只有出现异常值或异常波动，才要统计分析，因此，确定正常值范围很重要。医学领域需要确定正常值范围，只有通过大量的统计数据才能确定正常值范围，如血压正常值范围、血糖正常值范围。工业企业生产线的产品质量是否正常，需要确定一个正常值范围。

1.3 统计学的分类

1.3.1 理论统计学与应用统计学

1. 理论统计学与应用统计学的定义

统计学可以分为理论统计学（theorem statistics）与应用统计学（applied statistics）。

- **理论统计学** 它偏向理论的推论过程和结果，主要指的是创新出一些统计的定理或公式，或对于现有的统计的定理或公式进行推广或新解释，如数理统计学、概

率论等。

- **应用统计学**　它是从理论统计学中抽取出来的，为解决实际问题而经常需要用到的统计学方法的集成。要想正确使用理论统计学者所创造出来的统计定理或公式，达到解决某一现实问题或评估某一事件的目的，只要了解在何种状况下该用哪一个统计方法、数学公式或定理，以及该如何解释所解出的结果。应用统计学偏向于解决实际问题的分析过程和结果。现代应用统计学是将多个统计分支的方法进行有机地综合，如商务统计学、经济统计学、社会统计学等。应用统计学从方法上分为探索统计与验证统计。

2. 描述统计和推断统计

统计学一般又分为描述统计（descriptive statistics）和推断统计（inference statistics）。

- **描述统计**　包括了探索统计，并考虑了数据的调查搜集与整理，是统计学中描绘或总结观察量的基本情况的统计总称。研究者可以透过对数据资料的汇总处理，将原始调查资料进行摘要变为图表，再进行图像化处理，以直观方式了解整体资料分布的情况，如测算平均值、比例等。描述统计的常用方法是分组、汇总与数据概括。描述统计的成果是统计图表。描述统计是数据的基础处理方法，占统计分析的 80% 以上工作量。
- **推断统计**　用样本推断总体，主要推断总体的平均与比例。推断统计是从随机抽取的样本数据中提取信息，推断整个总体的统计方法体系。推断统计主要是利用参数估计和假设检验来推断总体。它用样本推断总体的参数，推断总体变量间的关系。一般在推断统计前也要先进行描述统计分析。描述统计与推断统计之间的关系如图 1-1 所示。

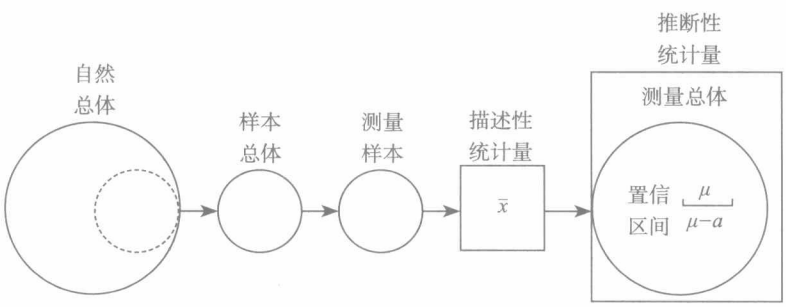


图 1-1　描述统计与推断统计之间的关系

描述统计如同对数据进行“体检”，正常人也要体检，推断统计如同“诊断治疗”，只有数据出现异常情况才可以进行推断统计。

1.3.2 探索性统计分析与验证性统计分析

应用统计学主要又分为探索性统计分析（exploratory statistical analysis）和验证性统计分析（confirmatory statistical analysis）。

（1）探索性统计分析（exploratory data analysis，EDA）给定一组数据，用统计方法进行数据概括分析来描述这些数据，或用图形展示所给数据。探索性统计分析是一种数据分析视角，也是一套统计分析方法。由数据的类型决定分析方法的选择，探索性研究要以问题为导向，也包括数据处理（data manipulation）。

数据概括分析是研究者通过分析数据资料，了解变量观察值的中心位置与分散的情况。运用的统计方法有数据中心位置测量（平均数、中位数、众数、几何平均数）和数据变异程度测量（标准差、中位数绝对离差 MAD、全距（极差）、四分位距、频数分布表）。

数据可视化（data visualization）图形展示通常使用的方法有多边图、直方图、饼图、散点图、箱线图、茎叶图、条形图等。

（2）验证性统计分析（confirmatory data analysis，CDA），也称验证性数据分析，利用数据来检验研究假设和变量的变化程度是否有统计意义。

探索性统计分析的角色是侦探，搜集线索和证据；验证性统计分析的角色是法官，根据数据信息对假设进行宣判（数据支持假设，还是目前得到的数据不支持研究假设）。研究者依据数据的形态建立一个用以解释其随机性和不确定性的数学模型，以之来推论研究中的假设。

探索性统计分析与验证性统计分析都被称作应用统计学，如图 1-2 所示。

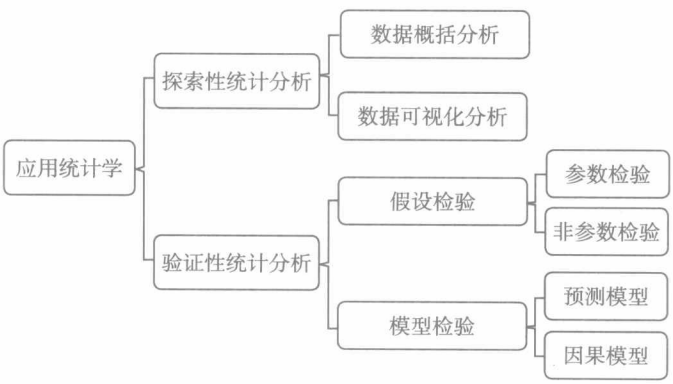


图 1-2 应用统计学体系

探索性统计分析中包括了总体参数的点估计。假设检验是变量分布的检验，模型检验是变量间关系的检验，模型检验包含了预测模型和因果模型。本书第 9 章强调要区分两类模型的应用。统计检验方法中目前流行贝叶斯统计检验法，即先验信息（经验信息）+

样本信息 = 后验信息。通常用后验信息进行假设检验。

1.4　统计学与数学和计算机的关系

1.4.1　统计学与计算机的关系

计算机对于统计学是计算工具，在没有计算机以前，统计学基本上是数学的一个分支，主要研究统计学公式与定理，如数理统计主要研究随机变量的分布与性质，因为实际应用统计时计算量是巨大的，现实问题极其复杂而无法手工计算。

计算机解决了统计数据的存储问题，网络解决了统计数据的传输问题，统计软件解决了统计的复杂计算问题。所以现在的统计学可以成为一个独立的一级学科，与数学并行，不再是隶属关系。

计算机的应用使统计学的学习方式发生了根本性变化。统计公式推导过程尽量压缩，主要要求学生掌握统计的思想、统计方法的选择、统计软件计算结果的解释。计算机的应用增加了统计实验和操作应用环节，体现了统计作为应用和方法论学科的特点。图 1-3 显示了统计与计算机的关系。

图 1-3　统计与计算机关系

没有计算机就没有应用统计学！计算机是非常重要的统计学应用工具，计算机和统计的发展相辅相成，统计学因为有计算机而应用广泛，统计学因为有计算机而学习容易，统计学因为有计算机而飞速发展。统计是分析数据的，计算机与互联网的飞速发展，促使大数据时代到来，统计学前途辉煌。

1.4.2　常用统计软件

统计软件是将各类统计功能编写程序称为可以调用的模块，模块越多，统计功能越强大。常用统计软件如下：

Excel　数据表格软件，必然有一定的统计计算功能。计算机都装有办公软件 Office 及 Excel。有统计计算和画统计图功能。Excel 使用方便，可以进行常用统计简单分析。本书介绍 Excel 的主要原因是可以方便地计算简单的统计问题。

R 软件　主要特点是功能强大、免费。本书介绍它用于较复杂的统计计算。R 语言编程很方便。有各个专业方向统计学家编写的统计软件包。从网络上可以搜索下载不断更新和增加的软件包和程序。目前 R 软件是发展最快的软件，受到世界上统计专业人员及师生的欢迎，是网上程序资源更新及时、方法最齐全的软件，是用户量增加最快的统