普通高等教育经管类专业系列教材

因果推断计量经济学

徐小君　蓝嘉俊 ◎ 主编

清華大學出版社
北　京

内容简介

本书主要介绍因果推断计量经济学的基本概念、基本原理和基本运算，以及因果推断实证研究中常用的设计方法。全书共分 12 章，具体内容包括：经济计量简介与基础知识、一元线性回归模型、多元线性回归分析、线性回归模型的拓展与应用、因果关系与因果图、工具变量估计方法、样本选择模型、潜在结果分析框架、断点回归设计概述、面板数据回归分析、时间序列基础知识和时间序列模型应用，等等。

本书既适用于高等院校经济与管理类专业的本科生，也可作为研究生和其他研究者学习社会科学计量方法的入门参考书目。

图书在版编目(CIP)数据

因果推断计量经济学 / 徐小君, 蓝嘉俊主编.
北京 : 清华大学出版社, 2025. 6. -- (普通高等教育经管类专业系列教材). -- ISBN 978-7-302-69328-4

Ⅰ. F224.0

中国国家版本馆 CIP 数据核字第 2025CG2979 号

责任编辑：高　岫
封面设计：马筱琨
版式设计：思创景点
责任校对：马遥遥
责任印制：沈　露

出版发行：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：三河市天利华印刷装订有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张：**17　　**字　　数：**435 千字
版　　次：2025 年 7 月第 1 版　　**印　　次：**2025 年 7 月第 1 次印刷
定　　价：69.00 元

产品编号：105676-01

前　言

百年大计，教育为本。改革开放以来，我国的教材建设已经为社会各类教育和研究工作做出了重要贡献。党的十八大以来，习近平总书记对教育事业，特别是培养社会主义建设者和接班人工作高度重视，围绕培养什么人、怎样培养人、为谁培养人这一教育的根本问题，就教育改革发展提出一系列新理念、新思想、新观点，突出强调要坚持党对教育事业的全面领导，坚持把立德树人作为根本任务，坚持优先发展教育事业，坚持社会主义办学方向，坚持扎根中国大地办教育，坚持以人民为中心发展教育，坚持深化教育改革创新，坚持把服务中华民族伟大复兴作为教育的重要使命。新时代我们要继续坚持守正创新，深刻认识到当今教育所面临问题的复杂程度和解决问题的艰难程度，在党的坚强领导下，立足基本国情，遵循教育规律，坚持改革创新，坚定不移走中国特色社会主义教育发展道路，发展具有中国特色、世界水平的现代教育，为社会主义现代化建设提供基础性、战略性支撑。

党的二十大报告首次提出“加强教材建设和管理”，把教材建设作为立德树人与培养社会主义建设者和接班人的重要举措，凸显了教材工作的重要地位。习近平总书记强调，培养出好的哲学社会科学有用之才，就要有好的教材。要高度重视教材建设，应以本教材的编写为契机，总结经验，明确方向，加大力度，推动全校打造更多更高水平的教材，推动人才培养和一流大学、一流学科建设达到更高水平。加快构建中国特色哲学社会科学，归根结底是建构中国自主的知识体系，教材要坚持以习近平新时代中国特色社会主义思想为指导，立足中国实践，坚持问题导向，要大力推动习近平新时代中国特色社会主义经济思想进教材、进课堂、进学生头脑，加快形成中国特色高质量教材体系。

高校经济学类教材要立足于新时代我国经济发展与经济改革，揭示中国特色社会主义经济的本质和发展规律，要实现思想性、系统性、实践性与学理性的统一。教材要构建新的框架体系，探讨新的观点，阐释新的理论学说，尝试新的研究方法，运用新的话语体系。教材要坚持观点正确，理论阐释深入，体系结构新颖，达到以教材为载体，创新和传播中国理论，讲好中国故事，为落实立德树人根本任务提供支撑。经济学理论创新离不开对世界各地经济实践的深刻认识，这既需要对实践中的问题进行更好的提炼，也需要避免极端化、相互排斥或者非此即彼。中国经济学应该受启发于现代经济学基本原理，并根据中国的历史背景和发展实践，创新提炼出具有一般意义的经济理论，教材要坚持马克思主义经济学的指导，借鉴西方经济学知识体系中适用的理论、概念、话语、方法，借鉴中国古代知识体系中优秀的思想，推动中国经济学理论逻辑、历史逻辑和实践逻辑等诸多方面的有机结合，从概念范畴、理论体系、研究方法等方面提升中国经济学的逻辑自洽性、理论阐释的科学性和指导现实的精准性。教材要围绕和关注中国实际问题，坚持理论前沿与基础知识相统一的原则，在提供基础框架的同时，尽力为学者呈现学科前沿的最新进展，并将中国的经济发展经验及案例呈现在教材之中。

本教材是华侨大学教材建设资助项目。华侨大学将学习与宣传贯彻党的二十大精神和习近平总书记重要指示精神作为贯穿各项工作的主线，以建设成为国内一流、国际上声誉良好的大学为目标引领，坚持“顶天、立地、树人”的发展思路，落实价值引领、能力培养、知识

传授相结合的育人理念，以学生为中心深化教育教学改革，传承与弘扬“爱国奉献、勇于担当、开放包容、追求卓越”的华侨大学精神，不断提高人才培养质量和办学水平。华侨大学教材建设资助项目坚持以习近平新时代中国特色社会主义思想为指导，是落实和推进华侨大学“侨校+名校”发展战略的具体举措，努力为全面建成社会主义现代化强国、实现第二个百年奋斗目标添砖加瓦。

近年来因果推断方法已在医学、心理学和流行病学等各个学科中广泛应用。2019 年和 2021 年诺贝尔经济学奖分别颁给了实验计量经济学和因果推断计量经济学这两个研究领域。吸收和借鉴各个学科中最新的科学方法，并将这些方法加以发展和应用，解决经济学中的各种问题，是经济学科研和教学的重要任务。编写本教材，一方面要吸收和借鉴各个学科中关于因果推断、人工智能和大数据等的相关技术，另一方面要更大范围地推广和普及计量经济学中最新的知识、技术和应用，从而提高计量经济学的教学质量。

随着计量经济学中因果推断方法的发展和广泛应用，相关的专著和教材也随之出现，如《基本无害的计量经济学》和《精通计量学》。当前，已有的相关教材一般都具备以下两个特征。一是现有涉及因果推断的专著或教材一般比较专业并且难度较大，适合研究生学习，而不太适合本科生。二是各类教材大多没有给出系统介绍经典计量基础知识和因果推断方法的统一框架。有些教材只介绍经典计量基础知识，而有些教材只侧重介绍微观计量经济学或因果推断方法。本教材既介绍经典计量基础知识，又引入因果推断方法，并将两者有机地结合在一起，将新的研究成果引入初级计量经济学的教学中，推动本科计量经济学教育高质量发展。

本教材主要介绍因果推断计量经济学的基本概念、基本原理和基本运算，主要内容包括经济计量简介与基础知识、一元线性回归模型、多元线性回归分析、线性回归模型的拓展与应用、因果关系与因果图、工具变量估计方法、样本选择模型、潜在结果分析框架、断点回归设计概述、面板数据回归分析、时间序列基础知识和时间序列模型应用，等等。

本教材的特色和优势如下。

(1) 经典计量理论与最新因果推断方法相结合，体现了计量经济学研究和发展的科学性、系统性和前沿性。传统计量经济学一般介绍违背经典回归模型的异方差、自相关和多重共线性等问题，而将因果识别和因果推断方法排除在外，从而使得学生学完课程之后，仍然对经济学中相关关系和因果关系没有形成准确认知。本教材建立了科学的分析框架，将相关关系和因果关系、观测数据和实验数据、真实历史事件和虚拟的反事实等加以区别和比较，并将传统计量经济学和新近发展的因果推断方法系统整合，为学生进一步学习打下基础。

(2) 简单易学、实用性强，有助于本科生学习计量经济学和因果推断方法。本教材借助计量经济学研究中的各种案例分析，训练和提高学生的实践和应用能力。

本书免费提供多种教学资源，包括但不限于教学课件与教学资料、教案与教学大纲、习题及答案，便于教师教学和学生自学，可通过扫描右侧二维码获取。

教学资源

感谢学校和学院领导对本教材编写的支持。本教材在编写过程中，还得到了很多同事和学生的帮助，在此表示特别感谢。经济学院同事蓝乐琴和严圣艳老师，阅读了本书的部分内容并提出了修改建议。我的研究生们几乎都参与了本教材的编写和资料收集与整理过程。他们是：张婷婷、周纪宇、谢紫珊、李雨辰、林钰琪、顾妍婕、梁菁仪、张继鹏、谢奔、覃雅婷、雷智聪、陈函希。感谢上述老师和同学对本书编写的帮助和支持。本教材在编写过程中参考了国内外诸多相关的优秀书籍和资料，在此也对同行表示感谢。由于水平有限，书中可能存在错误之处，欢迎读者批评指正。

编　者

2025 年 6 月

目　录

第1章 经济计量简介与基础知识

计量经济学(Econometrics)是经济学的一个分支，它结合经济理论、数学和统计学方法，用来分析、解释和预测经济现象，并通过实际数据验证经济理论。本章首先介绍计量经济学的基本概念及其主要应用，随后介绍基本的概率和统计学知识，为学生后续学习打好基础。

1.1 计量经济学简介

1.1.1 什么是计量经济学

计量经济学是以一定的经济理论和数据资料为基础，运用数学、统计、实验方法与计算技术，以建立经济计量模型为主要手段，从而定量分析和研究具有随机性特性的经济变量关系和规律的一门学科。其主要内容包括理论计量经济学和应用经济计量学。理论计量经济学以介绍和研究计量经济学的理论与方法为主要内容，侧重于计量理论与方法的发现、证明与数学推导，与经济理论、数理统计和数学联系极为密切。应用计量经济学以理论计量经济学为基础，为了解决实际经济活动中的数量问题建立与应用计量经济学模型，侧重于对实际问题的探索和处理。

“计量经济学”(econometrics)一词，是挪威经济学家弗里希(Ragnar Frisch)在 1926 年仿照“生物计量学”(biometrics)一词提出的。1930 年国际计量经济学学会成立了，其在 1933 年创办了《计量经济学》杂志，这可以作为计量经济学学科诞生的标志。

计量经济学的建立和发展过程是综合应用经济理论、实验、统计、数学方法的过程。经济学为其提供理论基础，实验和统计为其提供经验和数据资料，数学为其提供研究方法。计量理论模型的设定、样本数据的收集，以经济理论为依据，建立在对所研究经济现象的透彻认识基础上。而模型参数的估计和模型有效性的检验，则是统计学和数学方法在经济研究中的具体应用。没有理论模型和样本数据，实验、统计和数学方法就没有发挥作用的“对象”和“原料”。反之，如果没有实验、统计和数学方法所提供的“原料”，将无法形成“产品”。没有理论指导的实践，以及没有经验事实为支撑的理论研究，一定程度上都缺乏意义。因此，计量经济学的发展和应用，广泛涉及经济学、实验、统计、数学等各个学科的理论、应用和实践。计量经济学是一门交叉融合发展的经济学科。

计量经济学研究的一般步骤如下：先根据实际需要确定研究任务和目的，并结合经济理论、经济活动及经验事实，设定理论模型和计量模型，然后从实验或经济活动中收集样本数据，利用统计和数学的计算技术，通过样本数据对模型中的参数进行估计，并对模型

进行检验。如果检验通过，则可以应用模型；反之，则需要对模型进行修订。图 1.1 给出了计量经济学的研究步骤。

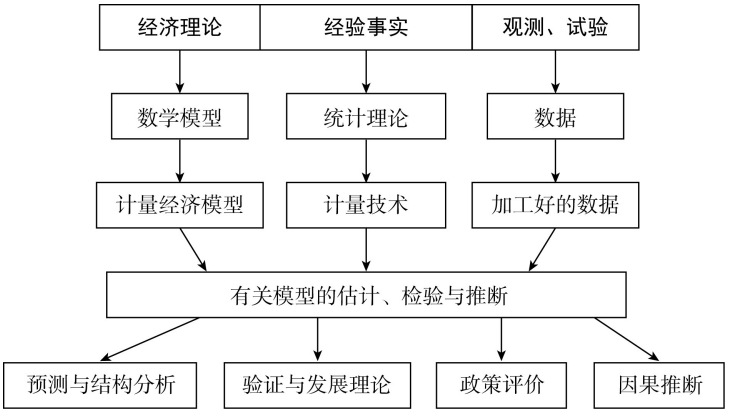

图 1.1 计量经济学的研究步骤

计量经济学的应用一般包括以下 4 个方面。

(1) 经济预测和结构性分析。经济预测，即运用计量模型对经济现象进行描述性分析和相关性分析。其基本原理是拟合和模拟历史，从已经发生的经济活动中找出变化趋势和规律，利用趋势和规律对未来进行预测。经济结构分析，即研究一个或几个经济变量发生变化，以及结构参数的变动对其他变量以致整个经济系统产生何种影响。其一般方法是弹性分析、乘数分析、比较静态分析及比较动态分析。

(2) 检验与发展经济理论，即利用实验数据或统计资料，运用计量经济学模型分析某一个理论假说正确与否。正确的经济理论应该能很好地解释过去，也能够预测未来。如果按照某种经济理论建立的计量经济模型，能够很好地拟合实际观察数据，则意味着该理论是符合客观事实的；反之，则表明该理论不能解释客观事实。基于经济理论的计量经济学模型有如下两个功能：①按照经济理论建立模型，能够较好地拟合历史数据和预测未来，这就检验了经济理论的正确性。②基于现有的经验事实和数据，经济学家通过探索和尝试，建立不同的理论和模型，以模拟、解释现实和预测未来。如果经济学家找到的理论及其模型，能够很好地反映经济运行规律，则这一过程可称为经济学家利用计量经济学发现和发展了经济理论。

(3) 政策评价，也称为项目评估(program evaluation)，即利用计量经济模型定量分析政策对经济系统或实验对象的影响和效果。计量经济学政策评估主要包括两个方面：①从宏观角度，对诸如国家政策的影响和效果的评估与评价，如评价中国某一个时期的货币政策或财政政策的效果。②从微观角度，评估某一项措施的效果。比如：企业采用一项新的管理制度产生的影响和效果；学校采用一种新教学方法产生的影响和效果；地方政府提供的免费职业培训对工人就业和工资的影响效果；等等。

(4) 因果推断，主要是利用实验或观察数据，对事物或变量之间的因果关系情况进行判断和分析。在定性分析和研究因果关系的基础上，再计量分析变量之间因果关系的作用机制和影响程度。计量经济学的因果推断，可以验证或发现和发展经济理论。

计量经济学作为一门学科已有近百年的发展历史。在中国，计量经济学的发展已有近 40 年，极大地推动了中国经济学教育与研究的学术化、规范化、国际化，成为经济学研究理论联系实际的主要方法与工具。在很长一段时间内，中国经济学研究以定性分析为主，

缺少对现实经济的定量分析和实证研究，计量经济学的引进与广泛应用，使中国经济学研究水平得到很大提升，并且在国际经济学术界初显其学术影响力。

早期的实证经济学研究缺乏解决内生性问题的有效办法，导致实证研究的结论缺乏可信性。20 世纪 80 年代，经济学家们开始对传统的经验和实证研究进行反思，认为研究者利用观测数据和传统计量分析方法进行研究，特别是因果推断时缺乏严谨的态度，对分析结果特别是因果关系，缺乏可靠性分析和稳健性检验。之后，经济学界对经验研究的可信性进行了越来越多的探讨，传统计量经济学正在经历一场可信性革命，而因果推断正是这场可信性革命的核心。传统计量经济学课程的内容多以“可信性革命”之前的经典计量经济学为主，强调经典线性回归模型下(普通最小二乘)估计量的性质，以及经典假定违背之后估计量存在的问题及修正方法，较少强调变量之间的因果关系。因果推断是“可信性革命”的核心内容，强调运用实验和准实验数据，寻找特定的外生性冲击，挖掘研究情境及事情变迁的过程，从而解决核心解释变量的内生性问题，准确识别其对被解释变量的因果效应。当前，计量经济学借助准实验设计方法对因果关系进行评估，显著提高了从观察数据中进行因果关系评估的可信性。因果推断已经成为国内外计量经济学研究的主流。

1.1.2　因果推断简介

探求事物的原因，是人类永恒的精神活动之一。从古希腊的哲学到中国先秦的诗歌，都充满了对原因的追问和对因果关系的思考。比如，亚里士多德就在《物理学》(*Physics*)和《形而上学》(*Metaphysics*)两书中反复强调，我们只有知道了事物的原因，才能算真正理解这个事物。又如屈原在《天问》的开篇就追问日月星辰运行的原因。

长期以来，人们一方面好奇地追问原因和结果的关系，另一方面又苦于这些概念的模糊性。于是，这些话题在很长一段时间都仅仅局限在哲学和文学的范围内。精确地描述因果关系，尤其是用数学的语言来描述因果关系，则是近代的事情了。这一思想飞跃，得益于现代统计学的发展。统计学家称之为“因果推断”(causal inference)。虽然因果推断在现代统计学的萌芽阶段就已经产生，但是它的发展并非一帆风顺：它长期被主流忽视、怀疑，甚至攻击。例如哲学家大卫·休谟(David Hume)甚至认为人们凭借经验根本无法认识因果关系。直至最近四十年，尤其是最近二十年，因果推断才得到了广泛认可，成为当今主流的研究方向之一。当今世界，很多年轻的学者加入了因果推断的研究队伍，相关研究来自统计学、经济学、社会学、政治科学、教育学、流行病学、计算机科学、哲学等领域。

人们常常思考关于原因和结果的问题。例如，某人死于肺癌，是不是因为他常常吸烟？感冒症状减轻了，是不是因为服用了维生素 C 片？大学教育能否提高收入水平？类似的问题，充斥着我们的日常生活。但是，这些看似简单的问题，却不容易获得准确答案。例如，有人吸烟，却没有得肺癌；而有人不吸烟，却得了肺癌。又如，可能仅仅喝白开水，感冒症状也会减轻。再如，有人没有上大学，却靠做生意发了大财。当然，有点概率论常识的人，都会很容易意识到，这些事件带有随机性。在生活中，我们可以观察到吸烟的人更可能得肺癌；服用维生素 C 的人，平均来说，感冒恢复得更快；上过大学的人平均收入更高。但是，这些统计上的“相关关系”是否就是“因果关系”呢？

1923 年耶日·奈曼(Jerzy Neyman，1894—1981)在他的博士毕业论文《概率论在农业实验中的应用》中，提出了用于因果推断的“潜在结果”(potential outcomes)的数学模型，并将它和统计推断结合起来。以农业实验为例，实验者想检测两种肥料对产量的影响。其

基本思想是，样本分组随机化实验保证了两个组的各种影响因素是相似的，那么它们之间结果的区别就可以归因于干预因素的作用。内曼关于因果推断的相关论文，引发了包括罗纳德·费希尔(Ronald Fisher)在内的统计学家的激烈争论。同时期，费希尔对随机化实验进行了深入的研究，虽然他没有使用内曼潜在结果的记号，但是因果推断始终是他思考的对象。随后的几十年，随机对照实验(randomized controlled trial, RCT)成为美国食品药品监督管理局批准新药的黄金标准。最近三十年，大量的随机化实验出现在社会科学中，用来研究复杂社会问题中的因果关系。比如，麻省理工学院和哈佛大学的三位经济学家阿比吉特·班纳吉(Abhijit Banerjee)、埃斯特·迪弗洛(Esther Duflo)和迈克尔·克雷默(Michael Kremer)，采用实验的方法研究发展经济学，获得了 2019 年的诺贝尔经济学奖。2021 年大卫·卡德(David Card)、乔舒亚·安格里斯特(Joshua Angrist)和吉多·因本斯(Guido Imbens)三位经济学家，因他们近三十年来在经济学中应用因果推断方法的开创性工作，被授予了诺贝尔经济学奖。

随机对照实验虽然可以用来进行因果推断，但是很多数据并非源自随机化实验。这类研究通常被称为观察性研究(observational study)。比如，要研究吸烟和肺癌的因果关系，基本的伦理不允许我们随机地让一部分人抽烟、让一部分人不抽烟。再如，要研究接受大学教育对收入的影响，我们不能随机地让一部分人上大学、让一部分人不上大学。对于很多流行病学和社会科学的问题，应进行观察性研究。当然，人们也迫切地想从这些观察性研究中获得关于因果关系的知识。

计量经济学重要任务之一是研究变量间的因果联系。因果推断也是当前国际学术界最热门的研究领域之一，因果推断以经验和理论为指导，通过实验、观测和相关技术分析，研究事物和变量之间的因果关系，主要研究内容包括：因果发现(causal discovery)分析，即探索和发现事物和变量之间的作用机制和影响关系，探寻事物和变量演变的顺序和逻辑关系；因果效应(causal effect)计量，即在得到因果关系方向之后，评估原因变量发生的各种变化，导致结果变量发生变化的程度。经济学中常用的因果推断方法有 5 种：随机试验方法、回归方法(匹配法)、双重差分方法、断点回归方法及工具变量(IV)方法。因果关系研究是可以在随机控制实验中得到完全展示的，但是随机控制实验是人为设计实验，很多社会生活中的问题无法使用随机控制实验来研究。20 世纪 90 年代开始，大卫·卡德、乔舒亚·安格里斯特和吉多·因本斯开始发表有关工具变量法、双重差分方法、断点回归法等自然实验的方法，利用观察性数据，也可以研究因果关系。这让经济学研究的“可信度”大大提高，并且在政策方面产生了更直接和深远的影响。

虽然因果推断已经取得了一些进展，但是这些还不足以回应现实世界向其发出的挑战。理论上，目前的研究范式还不能完美地应对复杂的实际工作需要。一些学者考虑了因果推断和微分方程的关系，但是这方面的研究还在初始阶段。不管是鲁宾还是珀尔的范式，对于有反馈的因果系统，都有致命的缺陷，这也是值得思考的问题。另外，现有的工作大多数都是在评估某个给定的原因对某个给定的结果的作用，而科学研究的本质是探索未知的原因。虽然因果图的结构学习对探索原因有帮助，但是这方面的理论还不够丰富。

我们正处于大数据时代，如何从海量数据中挖掘因果关系，也是一个非常有挑战性但是引人入胜的话题。大数据(big data)是指在较短的时间范围内，用传统和常规技术工具难以进行收集、管理和应用的巨型数据集合。对大数据的研究和应用，需要有新的处理模式和技术才能使其呈现更好和更有效的结果。虽然大数据技术和因果推断两种方法在过去是

独立发展的，但是在未来会相互交织而产生新的成果。从应用的角度来看，因果推断一直与很多学科有紧密联系。这些跨学科的研究常常超越现有的因果推断理论，成为新理论和方法研究的源头活水。

1.1.3　数据的类型和结构

计量经济学研究中主要应用的数据包括实验数据和观测数据。

1. 实验数据

实验数据(experimental data)一般是在科学实验环境下取得的数据。在实验中，实验环境是受到严格控制的，得到的数据一定是某一约束条件下的结果。在自然科学研究中，实验的方法应用得非常普遍，因此，自然科学研究中所用的数据多为实验数据。例如，新开发药物的疗效测试、农作物品种试验等。这些实验数据主要用于考察变量之间的因果关系。在实验中，研究人员要控制某一情形的所有相关方面，操纵少数感兴趣的变量，然后观察实验的结果。实验是检验变量间因果关系的一种常用方法。

2. 观测数据

观测数据(observational data)是对客观现象进行实地调查、观测和统计所取得的数据，在数据取得的过程中一般没有人为控制和条件约束。在社会经济问题研究中，观测是取得数据最主要的方法之一。很多社会经济问题不适合应用实验的方法，只能通过实际调查得到数据，用各种调查方法得到的数据都属于观测数据。例如，2020 年我国的 GDP、年末人口数据等。社会经济领域的统计数据大部分是观测数据。观测数据常见的数据结构类型包括横截面数据、时间序列数据及面板数据。横截面数据是在一个固定时间上，对不同个体进行观测而收集的数据。时间序列数据是对个体在多个时间点上重复观测得到的数据。面板数据是对多个个体在多个时间点进行观测得到的数据。

(1) 横截面数据(cross-sectional data)是假定从总体中随机抽样得到的样本，在横截面分析中，也可以忽略样本数据搜集中细小的时间差别，有时所有单位的数据并非完全对应于同一时间段。如果一系列家庭都是在同年度中的不同时间被调查，得到的数据仍被认为是横截面数据。随机抽样得到的横截面数据集，其重要特征是数据的排序不影响计量分析。

(2) 时间序列数据(time series data)是由人们对一个或多个变量在不同时间进行观测，进而获得的观测值所构成的。这类数据反映了某一事物、现象等随时间的变化所呈现的状态或程度。与横截面数据不同，时间序列对观测值按时间先后排序，时间顺序是潜在的重要信息。时间序列数据的计量分析用到了许多和横截面分析相同的工具。由于许多时间序列数据具有明显的周期性、趋势性和持续性，对其研究和分析可能需要更加复杂的技术。时间序列变量的数据频率一般是固定的，数据是根据相同的规则定期出现的。常见的时间频率是天、周、月、季度、年，如股票数据一般以天为单位进行记录，消费者价格指数一般是以月度为单位，国内生产总值一般是以季度为单位，人口统计数据一般以年为单位。

(3) 面板数据(panel data)由数据集中每个横截面单位的一个时间序列组成，也叫“平行数据”，包括时间序列和截面两个维度。在时间序列上取多个截面，并在这些截面上同时收集样本观测值，从而构成样本数据。这些数据集是从不同观测对象在不同时间段或时点上收集来的，旨在描述这些观测对象随着时间变化而变化的情况。面板数据包括横截面和

时间序列两个维度上的数据：个体维度(i=1,2,···,N)和时间维度(t=1,2,···,T)。个体可以是个人、企业、行业或国家；时间维度可以是年、季、月或日。例如，对于 2000 至 2020 年全国各省份的 GDP，如果只考虑某一时间段或时点，它就是截面数据；如果只考虑某一观测对象，它就是时间序列数据。

1.2 概率论复习

经济变量大多为随机变量，经济变量之间的关系为随机关系。经济变量特征的刻画和经济变量关系的研究都需要以概率论和统计学的知识为基础。本章后续部分对概率论和统计学的内容进行简要回顾，为相关概念和理论知识的学习打好基础。

事件发生的可能性称为概率(probability)，用 $P(A)$表示事件 A 的概率。一个随机现象的所有结果发生的可能性大小构成其概率分布。概率分布是概率论研究的核心内容之一。

1.2.1 随机变量及分布

为便于采用数学工具研究随机现象，常用随机变量来表示随机现象。随机现象的可能结果常表现为数值，如一次考试的分数、明天股票的收盘价、明年的 GDP 等，即便不是数值，也可以用数值表示，例如，是否考上大学可以用 1 和 0 表示，抛掷一枚硬币出现的正反面也可以用 0 和 1 表示。因此，可以用变量表示随机现象。当变量取某个值或在某个范围内取值时，对应着随机现象的一个结果，这样的变量称为随机变量。随机变量分为离散随机变量和连续随机变量。如果随机现象的可能结果为有限个(如 100 个鸡蛋孵出的小鸡个数)，或者无限个但可以罗列，则可以用取离散值(比如 0,1,2,···)的变量表示该随机现象，得出的随机变量为离散随机变量。如果随机现象的可能结果对应连续的实数区间(例如一次考试的分数在[0, 100]内取值)，则必须用连续取值的随机变量表示，这样的随机变量为连续随机变量。常常用英文字母或者希腊字母(如 x, y, ξ, η 等)表示随机变量。离散随机变量的概率分布常用列表的形式或者概率函数的形式给出，连续型随机变量则用概率密度刻画概率分布。

1. 离散随机变量的概率分布

以本书用到的二项分布为例介绍离散随机变量的有关概念。

1) 二项分布(binomial distribution)

独立进行 n 次试验(例如用 n 个鸡蛋孵化小鸡)，每次试验中事件 A (例如“孵出小鸡”)发生的概率为 p。设 ξ 表示 n 次试验中 A 发生的次数(例如孵出小鸡的个数)，ξ 为随机变量，其概率分布用概率函数表示为

$$P(\xi = i) = C_n^i p^i (1-p)^{n-i} \quad (i = 0,1,2,\cdots,n) \tag{1.2.1}$$

称 ξ 为服从参数(n, p)的二项分布。以随机变量的取值为横坐标、对应的概率为纵坐标，画出概率分布图，如图 1.2 所示。

从图 1.2 中可以看出：①二项分布在某个取值处概率达到最大；②随着 n 的增大，二项分布逐渐接近正态分布。二项分布随机变量 ξ 的数学期望和方差分别为 $E\xi = np$ 和 $\mathrm{Var}(\xi) = np(1-p)$ 。

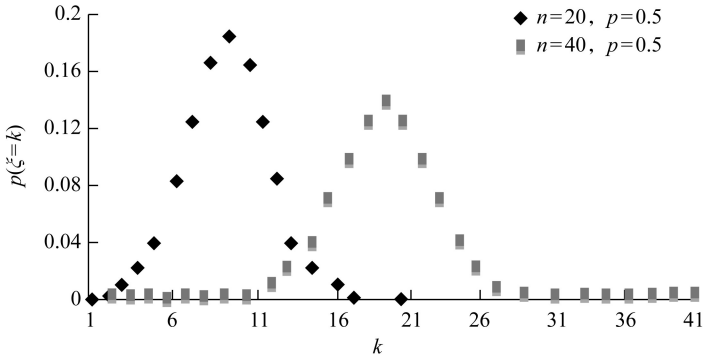

图 1.2　二项分布概率分布图

2) 两点分布——二项分布的特例

n=1 时的二项分布称为两点分布，此时随机变量 ξ 取 0 和 1 两个值，因此也称为 0-1 分布。由于 $C_1^0=1$、$C_1^1=1$，两点分布的概率函数为

$$P(\xi=i)=p^{\xi}(1-p)^{1-\xi}=p^{i}(1-p)^{1-i}\quad(\xi=i=0,1) \tag{1.2.2}$$

两点分布常用来描述有两个结果的随机现象。例如，公司的财务状况分为正常和危机；高校本科毕业生的选择为就业和继续读书；银行贷款客户在贷款到期时的行为分为履约(还款)和违约；等等。

2. 连续随机变量的概率分布

连续随机变量的取值为实数区间，不能通过列举的方法表示取值和对应的概率。此外，连续随机变量取特定值的概率均为 0，其分布特点需要通过在不同范围内取值的概率来刻画。连续随机变量采用概率密度函数描述其概率分布的规律性。设 $\varphi(x)$ 为连续随机变量 ξ 的概率密度函数，A=[a, b]为某个区间，其概率函数为

$$P(A)=P(\xi\in[a,b])=\int_a^b\varphi(x)\mathrm{d}x \tag{1.2.3}$$

从式(1.2.3)可以看出，概率密度函数满足两个条件：①取非负值，$\varphi(x)\geqslant 0$；②在 $(-\infty,+\infty)$ 上的积分等于 1。随机变量在某区间内取值的概率等于概率密度函数图形围出的曲边梯形的面积。

连续随机变量的取值统一规定为全部实数 $(-\infty,+\infty)$，如果实际取值为实数的一部分，例如正数 $(0,+\infty)$，只需要将取其他值时的概率密度设为 0 即可。因此，定义连续随机变量时，只需要给出概率密度函数即可。本书中提到的连续随机变量为正态分布随机变量，以及由正态分布衍生出的 χ^2 分布、t 分布和 F 分布随机变量。

1) 正态分布(normal distribution)

如果随机变量 ξ 的密度函数为

$$\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\quad(-\infty<x<\infty) \tag{1.2.4}$$

称 ξ 服从标准正态分布(standard normal distribution)，记为 $\xi\sim N(0,1)$，称 ξ 为正态随机变量。如果 ξ 的密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty<x<\infty,\ \sigma>0) \tag{1.2.5}$$

称 ξ 服从参数为 (μ,σ^2) 的正态分布，记为 $\xi\sim N(\mu,\sigma^2)$。参数为 (μ,σ^2) 正态分布的密度函数可以用标准正态分布密度函数表示为 $f(x)=\sigma^{-1}\phi\left[(x-\mu)/\sigma\right]$。正态分布的概率密度函数关于 $x=\mu$ 对称，最大值为 $1/\sqrt{2\pi}\sigma$。不同参数对应的正态分布的概率密度函数图形，如图 1.3 所示。

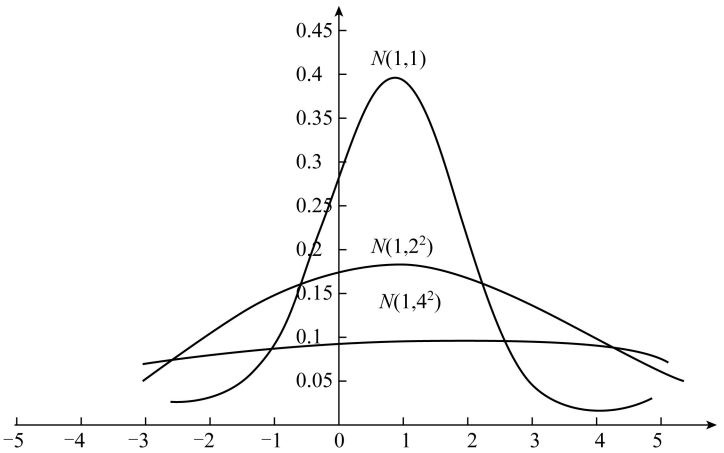

图 1.3　正态分布的概率密度函数图形

从图 1.3 中可以看出，参数 μ 决定密度函数的位置(中心)，σ^2 决定图形的陡峭程度。σ^2 越大，图形越扁平，随机变量在离开中心($x=\mu$)较远处区间内取值的概率越大；σ^2 越小，图形越陡峭，随机变量在离开中心($x=\mu$)较远处区间内取值的概率越小，在中心区域取值的概率越大。

标准正态分布的分布函数用 $\Phi(x)$ 表示，是标准正态概率密度函数的变上限定积分：

$$\Phi(x)=\int_{-\infty}^{x}\varphi(t)\mathrm{d}t=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2\sigma^2}}\mathrm{d}t$$

$\Phi(x)$ 的值可以通过查表获得。一般正态分布随机变量 $\xi\sim N(\mu,\sigma^2)$ 可以标准化为标准正态分布 $\eta=(\xi-\mu)/\sigma\sim N(0,1)$，对应的概率计算可以转化为标准正态分布：

$$\begin{aligned}P\{a<\xi\leqslant b\}&=P\{(a-\mu)\sigma<(\xi-\mu)/\sigma\leqslant(b-\mu)/\sigma\}\\&=P\{(a-\mu)\sigma<\eta\leqslant(b-\mu)/\sigma\}\\&=\Phi\left[(b-\mu)/\sigma\right]-\Phi\left[(a-\mu)/\sigma\right]\end{aligned}$$

标准正态分布概率表只给出 $x\geqslant 0$ 的 $\Phi(x)$ 值，当 $x<0$ 时，根据正态分布对称性得出的关系式 $\Phi(x)=1-\Phi(-x)$ 进行概率计算。

除了数学期望和方差外，计算中常常用到正态分布的三阶矩和四阶矩。设 $\xi\sim N(\mu,\sigma^2)$，$E(\xi-\mu)^3=0$，$E(\xi-\mu)^4=3\sigma^2$。

正态分布具有可加性：联合正态随机变量的线性组合仍然服从正态分布。

2) 正态分布的衍生分布

χ^2 分布(chi-square distribution)

标准正态随机变量的平方和服从的分布称为 χ^2 分布：如果 $\xi \sim N(0,1)$，则 $\eta = \xi^2$ 服从的分布称为自由度为 1 的 χ^2 分布，记为 $\eta \sim \chi^2(1)$；如果 $\xi_1 \sim N(0,1)$，$\xi_2 \sim N(0,1)$，ξ_1 和 ξ_2 相互独立，则 $\zeta = \xi_1^2 + \xi_2^2$ 服从的分布称为自由度为 2 的 χ^2 分布，记为 $\eta \sim \chi^2(2)$。以此类推可以得出：n 个相互独立的标准正态分布随机变量 $\xi_i \sim N(0,1)$，$i = 1,2,\cdots,n$ 的平方和 $\eta = \xi_1^2 + \xi_2^2 + \cdots + \xi_n^2$ 服从的分布称为自由度为 n 的 χ^2 分布，记为 $\eta \sim \chi^2(n)$。χ^2 分布的有关概率可以通过查表获得。

χ^2 分布具有可加性：独立的 χ^2 分布随机变量和仍然服从 χ^2 分布，分布的自由度等于求和 χ^2 分布自由度的和。

χ^2 分布为非对称分布，只在正数部分有非 0 的概率密度。

不同自由度 χ^2 分布的概率密度曲线，如图 1.4 所示。

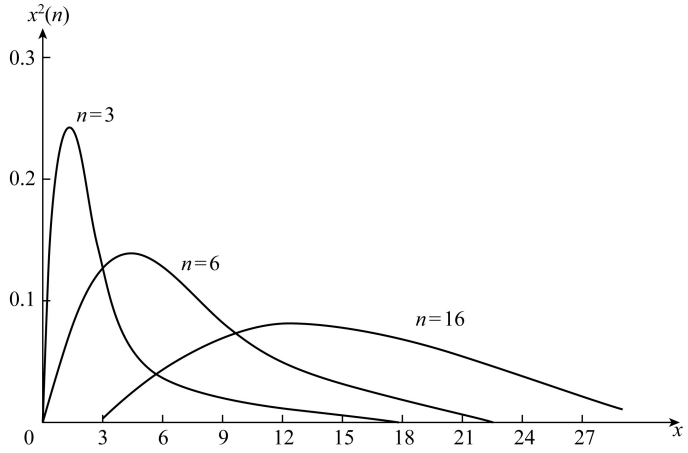

图 1.4　χ^2 分布的概率密度曲线

χ^2 分布的自由度是其概率密度函数中的唯一参数，自由度决定了 χ^2 分布的数学期望和方差：如果 $\xi \sim \chi^2(n)$，则 $E\xi = n$，$\mathrm{Var}(\xi) = 2n$。

t 分布(t- distribution)

t 分布(也称学生 t 分布，英文为 student's t-distribution)是由标准正态随机变量和 χ^2 分布随机变量通过运算形成的。设 $\xi \sim N(0,1)$，$\eta \sim \chi^2(n)$ 相互独立，定义随机变量为

$$t = \frac{\xi}{\sqrt{\eta/n}} \left(\text{也可形象地写为} = \frac{N(0,1)}{\sqrt{\chi^2(n)/n}}\right) \tag{1.2.6}$$

称 t 服从自由度为 n 的 t 分布，记为 $t \sim t(n)$。同样可以通过随机变量函数的分布由标准正态的密度函数和 χ^2 分布的密度函数求出 t 分布的密度函数。

t 分布为对称分布。根据 $\chi^2(n)$ 的定义，式(1.2.6)中分母根号下的 $\chi^2(n)/n$ 为 n 个独立随机分布变量 x_i^2，$x_i \sim N(0,1)$ $(i = 1,2,\cdots,n)$的平均值。根据大数定律，当 $n \to \infty$ 时，以概率 1 收敛到(标准正态分布的方差) $E(x_i^2) = 1$，因此当自由度 n 增大时，t 分布收敛趋于标准正态分布。不同自由度的 t 分布和标准正态分布的概率密度曲线如图 1.5 所示。

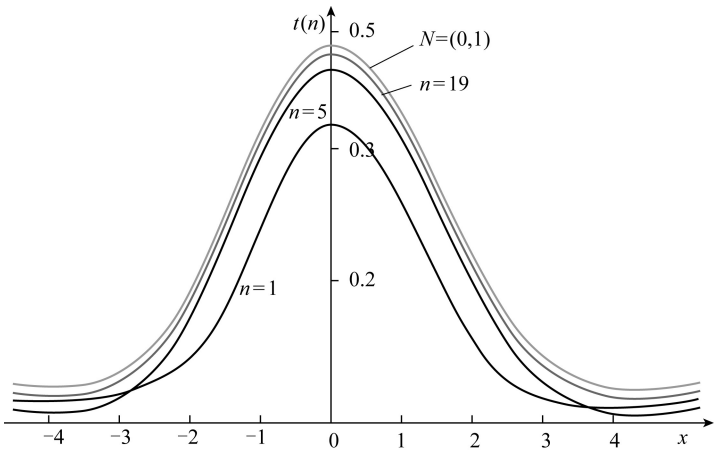

图 1.5　t 分布和标准正态分布的概率密度曲线

从图 1.5 中看出，当 n=19 时，t 分布和标准正态分布十分接近，从密度函数图像上已很难看出区别。

t 分布的自由度是其概率密度函数中的唯一参数，自由度决定了矩：t 分布只存在阶数低于自由度的矩，如果 $\xi \sim \chi^2(n)$，则对 $r < n$，如果 r 为奇数，$E\xi = 0$；如果 r 为偶数，$E\xi$ 的表达式较为复杂。t 分布的数学期望为 0，$t(2)$分布不存在方差和更高阶的矩，当 $n > 3$ 时，$t(n)$的方差为 $\mathrm{Var}(\xi) = E\xi^2 = n/(n-2)$。$t(5)$分布存在 4 阶矩 $E\xi^4 = 25$。

t 分布的有关概率可以通过查表获得。

***F* 分布(*F*- distribution)**

F 分布是由两个独立的 χ^2 分布随机变量经运算形成的。设 $\xi \sim \chi^2(n_1)$，$\eta \sim \chi^2(n_2)$ 为两个相互独立的随机变量，定义随机变量为

$$F = \frac{\xi/n_1}{\eta/n_2} \quad \left(\text{也可形象地写为} = \frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}\right) \tag{1.2.7}$$

则称随机变量 F 服从第一自由度(也称分子自由度 d. f. n，英文为 degree of freedom in numerator)为 n_1、第二自由度(也称分母自由度 d. f. d，英文为 degree of freedom in denominator)为 n_2 的 F 分布，记为 $F \sim F(n_1, n_2)$。

F 分布为非对称分布，只在正数部分有非零的概率密度。从 F 分布的定义得出，如果 $F \sim F(n_1, n_2)$，则 $1/F \sim F(n_2, n_1)$，而对于 $\alpha > 0$，$P(F < \alpha) = P((1/F) > 1/\alpha)$。因此，$F$ 分布概率表只给出形如 $F > \alpha$ 事件的概率，其他类型事件的概率可以通过如上的关系式求出。

3. 概率分布的统一表示——分布函数

可以用同一工具描述离散随机变量和连续随机变量的概率分布，这个工具就是分布函数(也称累积分布函数 CDF，英文为 cumulative distribution function)。设 ξ 为随机变量，x 为任意实数，事件 $(\xi \leqslant x)$ 的概率 $P(\xi \leqslant x)$ 构成以 x 为自变量的函数，称为 ξ 的分布函数，用 $F(x)$表示分布函数，即 $F(x) \equiv P(\xi \leqslant x)$。

从定义可以直接得出分布函数的性质：①单调递增；②右连续；③ $0 \leqslant F(x) \leqslant 1$；④ $\lim\limits_{x \to -\infty} F(x) = 0$，$\lim\limits_{x \to +\infty} F(x) = 1$。离散随机变量的分布函数为阶梯函数，在随机变量取值处发生跳跃，连续随机变量的分布函数连续可导，导函数为概率密度函数。

1.2.2　随机变量数字特征

概率分布给出了随机变量在任何范围内取值的概率，完全描述了随机变量(随机现象)的规律性，但人们仍然希望得到描述随机变量总体特征的综合信息，例如，随机变量平均值、随机变量的随机性、随机变量的分布特征(如对称性)等。

1. 数学期望(mean, expectation)

随机变量取值的概率平均值称为数学期望。离散随机变量的数学期望等于随机变量的各个取值与对应概率乘积之和，连续随机变量的数学期望等于概率密度函数乘以自变量 $(x\varphi(x))$ 在 $(-\infty,+\infty)$ 上的积分。

数学期望为随机变量以概率为权重的平均值，是随机变量分布的“中心”，称为位置(location)参数。例如，正态分布随机变量 $\xi \sim N(\mu,\sigma^2)$ 的数学期望为 μ，μ 是分布的中心，决定了概率密度曲线的位置。当人们对随机现象(例如投资的未来收益)的结果进行预测时，综合各种可能情况得出平均结果，即数学期望，这便是“期望”(expectation)的含义。

2. 方差(variance)和标准差(standard deviation)

方差用于衡量随机变量的随机性或者不确定性(uncertainty)。随机变量 ξ 的方差定义为

$$\mathrm{Var}(\xi) = E[\xi - E(\xi)]^2$$

由于 $E[(\xi - E(\xi))E\xi] = 0$，方差的计算公式也可以表示为

$$\mathrm{Var}(\xi) = E[\xi - E(\xi)\xi] = E\xi^2 - (E\xi)^2$$

由此得

$$E\xi^2 = \mathrm{Var}(\xi) + (E\xi)^2$$

如果 ξ 的数学期望等于 0，即 $E\xi = 0$，则 $\mathrm{Var}(\xi) = E\xi^2$。

方差为 0 的随机变量以概率 1 等于数学期望，可以看作常数。方差的平方根称为标准差。标准差和数学期望具有相同的单位，便于比较。

3. 矩(moment)

设 ξ 为随机变量，k 为正整数，数学期望 $E\xi^k$ 称为 ξ 的 k 阶矩(k-th order moment)，$E(\xi - E(\xi))^k$ 称为 ξ 的 k 阶中心矩。中心矩可以表示矩的函数，例如二阶中心矩(方差)可以表示为二阶矩和一阶矩平方的差 $\mathrm{Var}(\xi) = E\xi^2 - (E\xi)^2$。反之，二阶矩可以表示为二阶中心矩与一阶矩平方的和。

除了一阶矩(数学期望)和二阶中心矩(方差)外，常用三阶、四阶矩更为细致地刻画随机变量的分布特征。设随机变量 ξ 存在四阶矩，称

$$\lambda = \frac{E(\xi - E\xi)^3}{[\mathrm{Var}(\xi)]^{3/2}} \tag{1.2.8}$$

为 ξ 的偏度(skewness)。

偏度衡量了随机变量概率分布的对称性，$\lambda = 0$ 表明随机变量的概率分布关于数学期望值对称，$\lambda > 0$ ($\lambda < 0$)表明取值大于数学期望值的那部分概率分布大于(小于)取值小于数学期望值的部分概率分布，概率分布向右(左)偏斜[skew to right (left)]。偏度等于零，表明分

布关于数学期望值对称。例如正态分布为对称分布，偏度等于 0。

$$\kappa = \frac{E(\xi - E\xi)^4}{[\mathrm{Var}(\xi)]^2} \tag{1.2.9}$$

其中，κ 为 ξ 的峰度。

峰度用来刻画概率分布的尾部(两端)特征，即随机变量离中心较远取值的概率，衡量了随机现象的极端情况发生概率的大小。峰度越大，分布在两端的概率越大，极端情况发生的概率越大，此时称对应的分布为厚尾的(fat-tailed)。常常与正态分布比较来判断一个分布是否具有厚尾性。正态分布的峰度为 3，如果 $\kappa > 3$，称对应的分布具有厚尾性。在自由度较小时，t 分布的峰度大于 3，具有厚尾特征。例如 $t(5)$分布的峰度为 9。图 1.6 给出了自由度为 5 的 t 分布和标准正态分布的密度函数图。

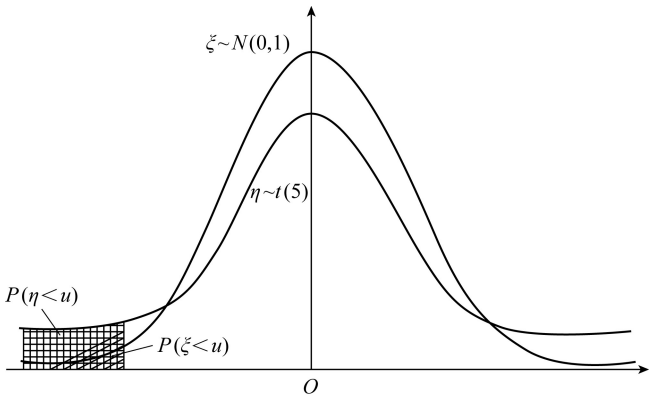

图 1.6　$t(5)$分布的厚尾性

从图 1.6 中看出，$t(5)$分布具有厚尾性：当 x 大于或者小于某个值之后，$t(5)$分布的概率密度曲线位于标准正态分布之上。

从峰度的含义可以看出选择概率分布的重要性。以投资收益为例，设 ξ 表示投资(例如购买股票)的年收益率，收益率均值为 25%，方差为 1。投资者能够承受不超过 40%的投资损失，即 $\xi > -0.4$。在对投资进行评估时，需要计算($\xi < -0.4$)(投资者不能承受损失)的概率。如果将 ξ 的分布选为正态分布，概率计算结果为 $P(\xi < -0.4) = \Phi(-0.4)$。如果 ξ 的实际分布不是正态分布，而是比正态分布尾部更厚的分布，按正态分布计算的概率会低估风险事件发生的可能性，以此制定的预防措施不能有效防范未来的风险。

1.2.3　随机向量

经济变量为随机变量，随机变量之间的相关关系是研究经济变量关系的基础。多个随机变量形成的向量称为随机向量。随机向量研究的重点在于各随机变量之间的相关关系。

1. 联合分布(joint distribution)

将随机变量 ξ 和 η 形成的随机向量记为 $\boldsymbol{X} = (\xi, \eta)'$。$X$ 的概率分布称为联合分布，而 ξ 的概率分布和 η 的概率分布称为 $\boldsymbol{X}$ 的边际分布(marginal distribution)。由概率定义的二元函数 $F(x, y) \equiv P(\xi \leqslant x, \eta \leqslant y)$ 称为 $\boldsymbol{X}$ 的联合分布函数。根据取值可以将随机向量分为离散随机向量和连续随机向量。连续随机向量的概率分布由联合概率密度函数确定。设 $\phi(x, y)$ 为

非负的二元函数，在 $(-\infty,+\infty)\times(-\infty,+\infty)$ 的二重积分等于 1，如果对平面上任何矩形区域 $[a,b]\times[c,d]$，$a<b$，$c<d$，概率 $P\{a<\xi\leqslant b,\ c<\eta\leqslant d\}$ 都可以用 $\phi(x,y)$ 在 $[a,b]\times[c,d]$ 的积分表示，表达式为

$$P\{a<\xi\leqslant b,c<\eta\leqslant d\}=\int_a^b\int_c^d\phi(x,y)\mathrm{d}x\mathrm{d}y$$

则称 $\phi(x,y)$ 为 $\boldsymbol{X}=(\xi,\eta)'$ 的联合概率密度。ξ 的概率密度 $\phi_\xi(x)$ 和 η 的概率密度 $\phi_\xi(y)$ 称为 $\boldsymbol{X}$ 的边际概率密度。

边际概率密度可以根据联合概率密度求出，由联合概率密度唯一确定，但反过来不成立：一般情况下不能根据边际概率密度求出联合概率密度，边际概率密度不能唯一确定联合概率密度。这很容易理解：联合概率密度不仅包括了随机向量中各随机变量的概率分布信息，还包括随机变量之间相互影响的概率分布信息。后一种信息是随机向量区别于随机变量的地方，是研究重点。给定边际概率密度，可以形成无限个联合概率密度，形成的方式决定了随机变量间的相关关联方式和关联的紧密程度。

2. 协方差和相关性(covariance and correlation)

协方差是随机向量的重要数字特征，衡量了两个随机变量之间的线性相关性。随机变量 ξ 和 η 的协方差定义为

$$\mathrm{Cov}(\xi,\eta)=E[(\xi-E\xi)(\eta-E\eta)]$$

另一种表达形式为

$$\mathrm{Cov}(\xi,\eta)=E(\xi\eta)-E(\xi)E(\eta)$$

当 $E\xi=0$ 或者 $E\eta=0$ 时，协方差的表达式简化为

$$\mathrm{Cov}(\xi,\eta)=E(\xi\eta)\tag{1.2.10}$$

$\mathrm{Cov}(\xi,\eta)=0$ 时，称 ξ 和 η 不相关。不相关只表明 ξ 和 η 的线性函数 $a+b\xi$ 和 $c+d\eta$ 没有相关关系，并不表明 ξ 和 η 的其他函数形成的随机变量不相关。例如 $\mathrm{Cov}(\xi,\eta)=0$ 并不意味着 $\mathrm{Cov}(\xi^2,\eta^2)=0$。将协方差标准化得出相关系数，表达式为

$$\rho(\xi,\eta)=\mathrm{Cov}(\xi,\eta)\Big/\sqrt{\mathrm{Var}(\xi)\mathrm{Var}(\eta)}$$

相关系数衡量了随机变量的相关关联方式和关联的紧密程度，大于 0 表示正相关，小于 0 表示负相关，等于 0 表示不相关，(绝对值)等于 1 表示完全相关。

随机变量之间的相关性，是变量间推断的基础，相关性越强(弱)，通过一个变量得出的另一个变量的信息越多(少)。当 ξ 和 η 不相关时，无法根据 ξ 的取值和概率分布得出 η 的信息，而当 ξ 和 η 完全相关时，二者具有确定的相关关系，根据 ξ 可以确定 η。

3. 条件分布(conditional distribution)

研究变量之间关系的另一个工具是条件分布。给定随机变量 ξ 的值 x 时，随机变量 η 的分布称为条件分布。条件概率定义的分布函数 $F_{\eta|\xi}(y\,|\,\xi=x)\equiv P(\eta\leqslant y\,|\,\xi=x)$ 称为给定 $\xi=x$ 的条件下 η 的条件分布函数，而给定 $\xi=x$ 时 η 的条件概率分布密度函数 $\phi_\eta(y\,|\,\xi=x)$ 可以用 (ξ,η) 的联合概率密度和 ξ 的密度计算，表达式为

$$\phi_\eta(y\mid\xi=x)=\frac{\phi(x,y)}{\phi_\xi(x)}$$

从中可以得出条件概率密度、边际概率密度和联合概率密度之间的关系为

$$\phi(x,y)=\phi_\eta(y\mid\xi=x)\times\phi_\xi(x) \tag{1.2.11}$$

即联合概率密度等于条件概率密度与边际概率密度的乘积。这一法则称为乘法法则。联合分布函数、条件分布函数和边际分布函数之间也满足这一法则。

如果ξ的概率分布对η的概率分布没有影响，则ξ和η独立。显然，相互独立随机变量的条件分布等于无条件分布(边际分布)：$F_\eta(y\mid\xi=x)=F_\eta(y)$，$\phi_\eta(y\mid\xi=x)=\phi_\eta(y)$。根据乘法法则，独立随机变量的联合分布等于边际分布的乘积。因此，只有在独立的情况下边际分布才能确定联合分布。在独立的情况下，分别研究ξ和η的分布已经足够，其联合分布中没有任何额外信息。

不相关和独立都是用来说明ξ和η之间不会相互影响，但独立要比不相关强得多：不相关只表明ξ和η之间不会相互影响，但不排除ξ和η的函数(例如ξ^2和η^2)之间相互影响；独立则意味着ξ和η的任何函数形式之间不会相互影响。因此，如果ξ和η独立，则一定不相关；反之，则不成立。只有在两个随机变量的联合分布为正态分布的情况下，独立和不相关才是等价的。

可以证明，如果ξ和η的任何函数形式不相关，ξ和η独立。

4. 条件期望和条件方差

采用条件概率分布计算的数学期望和方差分别称为条件期望(conditional expectation)和条件方差(conditional variance)。给定$\xi=x$，η的条件期望$E(\eta\mid\xi=x)$和条件方差$\mathrm{Var}(\eta\mid\xi=x)$分别定义为

$$E(\eta\mid\xi=x)=\int_{-\infty}^{+\infty}x\phi_\eta(y\mid\xi=x)\mathrm{d}y$$

和

$$\mathrm{Var}(\eta\mid\xi=x)=E(\eta^2\mid\xi=x)-[E(\eta\mid\xi=x)]^2$$

如上定义的条件数学期望$E(\eta\mid\xi=x)$和条件方差$\mathrm{Var}(\eta\mid\xi=x)$是以$\xi$取特定值为条件的，$\xi$为随机变量，取值可以变化，并服从一定的概率分布。条件期望和条件方差是ξ的函数，因此是随机变量，常表示为$E(\eta\mid\xi)$和$\mathrm{Var}(\eta\mid\xi)$。

1) 条件数学期望的重要性质

下面介绍条件数学期望的三个重要性质。

(i) $$E[E(\eta\mid\xi)]=E\eta \tag{1.2.12}$$

(ii) 设$g(\xi)$为ξ的函数形成的随机变量，则

$$E(\eta g(\xi)\mid\xi)=g(\xi)E(\eta\mid\xi) \tag{1.2.13}$$

性质(i)也称为数学期望迭代律(iterated law)，可叙述为对随机变量先取条件期望再取无条件期望，其结果等同于该随机变量的无条件期望。性质(ii)可以叙述为条件期望中被条件确定的量可以移到条件期望的外面。在条件期望$E(\eta g(\xi)\mid\xi)$中，$g(\xi)$由ξ完全确定，当ξ

给定时，$g(\xi)$ 可以看作常数直接移到条件期望之外。

由性质(ii)还可以得到条件期望的另一个经常用到的重要性质：

(iii) 如果 $E(\eta|\xi)=0$，则对 ξ 函数形成的随机变量 $g(\xi)$，有

$$E[\eta g(\xi)]=0 \tag{1.2.14}$$

性质(iii)可由性质(i)和性质(ii)直接推出

$$E[\eta g(\xi)]=E[E(\eta g(\xi)|\xi)]=E[g(\xi)E(\eta|\xi)]=0$$

条件数学期望的性质在模型设定和参数估计中有重要应用。

2) 不相关、$E(\eta|\xi)=0$ 和独立

(i) $E(\eta|\xi)=0$ 给出的变量 η 和 ξ 之间的关系，强于 η 和 ξ 的不相关关系。

为方便，不妨设 $E\eta=0$，此时 $\mathrm{Cov}[\eta,g(\xi)]=E[\eta g(\xi)]=0$，表明 η 与 ξ 的任何函数形成的随机变量(包括 ξ 本身)不相关，强于 η 和 ξ 的不相关关系。如果把条件期望 $E(\eta|\xi)$ 看作 ξ 对 η 的推测的话，$E(\eta|\xi)=0$ 表明 ξ 的任何函数形式提供的信息对推测 η 都没有作用。

(ii) $E(\eta|\xi)=0$ 给出的变量 η 和 ξ 之间的关系，弱于 η 和 ξ 的独立关系。

$E(\eta|\xi)=0$ 只能得出 η 本身与 ξ 的任何函数形式 $g(\xi)$ 不相关，但不能保证 η 的其他函数形式(例如 η^2)与 $g(\xi)$ 不相关。η 和 ξ 独立则表明：η 的任何函数形式 $f(\eta)$ 和 ξ 的任何函数形式 $g(\xi)$ 不相关。

1.2.4　极限定理

极限定理讨论一组随机变量的平均值当随机变量个数趋于无穷大时的极限，用于研究样本量无限增加时样本均值的性质。大数定律讨论随机变量均值的概率极限，中心极限定理则讨论分布极限。

1. 大数定律(law of large number，LLN)

大数定律的直观背景：对随机变量进行平均能够减小随机性。以测量误差为例进行说明：设某一物体的长度 μ 未知，对其进行 n 次测量，x_i 为第 i 次测量结果，$i=1,2,\cdots,n$。由于存在测量误差，x_i 为随机变量。设 x_i 独立同分布，$Ex_i=\mu$，$\mathrm{Var}(x_i)=\sigma^2$，方差 σ^2 越小，测量中的随机性越小，测量越准确。将 n 次测量结果进行平均能够减少方差：平均值 $\bar{x}=(x_1+x_2+\cdots+x_n)/n$ 的方差为 $\mathrm{Var}(\bar{x})=\sigma^2/n$，随 n 的增加迅速减少。由于测量是独立的，每次测量都有新的信息，$\bar{x}$ 采用了 n 次测量的信息，对物体长度的估计更为准确。如果无限增加测量次数 $(n\to\infty)$，可用的信息无限多时，能否将测量中的不确定性全部去除呢？这便是大数定律回答的问题：对随机变量施加合适的条件，使得其均值趋于一个常数而没有任何随机性。随机变量的极限涉及概率，用依概率收敛表示。

结论 1

独立同分布大数定律：设 $x_1,x_2,\cdots,x_n$ 为 n 个独立同分布随机变量，数学期望和方差分别为 $E(x_i)=\mu$ 和 $\mathrm{Var}(x_i)=\sigma^2$，$i=1,2,\cdots,n$。$\bar{x}=(x_1+x_2+\cdots+x_n)/n$，则

$$\bar{x}\xrightarrow{\ p\ }\mu(\text{或者 }\bar{x}-\mu\xrightarrow{\ p\ }0)$$

其中，p 表示依概率收敛。

当随机变量序列独立但不同分布时，只要随机变量方差具有有限上界，大数定律仍然成立。

结论 2

独立但不同分布大数定律：设 $x_1,x_2,\cdots,x_n$ 为 n 个独立同分布随机变量，数学期望和方差分别为 $E(x_i)=\mu_i$ 和 $\mathrm{Var}(x_i)=\sigma_i^2$，$i=1,2,\cdots,n$，方差有界：存在 $c>0$ 使得 $\sigma_i^2<c$，则

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_i)\xrightarrow{p}0 \tag{1.2.15}$$

证明：由切比雪夫(Chebyshev)不等式可知，对任意的 $\varepsilon>0$，有

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_i)-0\right|>\varepsilon\right]<\frac{\mathrm{Var}(\bar{x})}{\varepsilon^2}$$

只要 $\lim_{n\to\infty}\mathrm{Var}(\bar{x})\to 0$，则式(1.2.15)成立。由于 $x_1,x_2,\cdots,x_n$ 独立，则

$$\mathrm{Var}(\bar{x})=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(x_i)=\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\leqslant\frac{c}{n}\xrightarrow{n\to\infty}0$$

结论 3

既不独立也不同分布的大数定律：当随机变量 $x_1,x_2,\cdots,x_n$ 不独立时，只要 $x_i(i=1,2,\cdots,n)$ 的数学期望和方差存在，$\lim_{n\to\infty}\mathrm{Var}(\bar{x})\to 0$，且满足一些其他条件，则式(1.2.15)成立。

尽管在三种条件下大数定律都成立，但对 $x_1,x_2,\cdots,x_n$ 的分布要求的条件越宽松，$x_1,x_2,\cdots,x_n$ 中的信息积累速度越慢，随机变量均值以概率收敛的速度也越慢。

为便于应用，常将大数定律(1.2.15)表示为

$$n^{-1}\sum_{i=1}^{n}x_i-n^{-1}\sum_{i=1}^{n}E(x_i)\xrightarrow{p}0$$

如果极限 $\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}E(x_i)$ 存在，则式(1.2.15)可表示为

$$p\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}x_i=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}E(x_i) \tag{1.2.16}$$

在极限推导中，可以将 $n^{-1}\sum_{i=1}^{n}E(x_i)$ 和 $n^{-1}\sum_{i=1}^{n}x_i$ 作为渐近等价量相互替换。

2. 中心极限定理(central limit theory，CLT)

仍以物体长度测量为例说明中心极限定理。测量值 x_i 为随机变量，n 次测量结果的平均值 $\bar{x}$ 也是随机变量。在进行统计分析(如假设检验)时，常常需要知道 $\bar{x}$ 的分布。但 x_i 服从什么分布常常并不知道，因此 $\bar{x}$ 的分布是未知的，即使知道 x_i 的分布，除了具有可加性的几种分布(如正态分布、χ^2 分布等)外，一般情况下并不能由 x_i 的分布求出 $\bar{x}$ 的分布，这给统计分析带来困难。中心极限定理表明，只要满足一定条件，不管 x_i 服从什么分布，在 $n\to\infty$ 时，$\bar{x}$ 的分布接近正态分布。只要 n 足够大，可以用正态分布作为 $\bar{x}$ 的分布进行统计分析。

中心极限定理的条件较为复杂，这里仅叙述结果，详细内容可以参考有关文献。

结论 4

独立同分布中心极限定理：设 $x_1,x_2,\cdots,x_n$ 为 n 个独立同分布随机变量，数学期望和方

差分别为 $E(x_i)=\mu$ 和 $\mathrm{Var}(x_i)=\sigma^2$ ($i=1,2,\cdots,n$)。设 $\bar{x}=(x_1+x_2+\cdots+x_n)/n$，则在一定条件下，有

$$\sqrt{n}(\bar{x}-\mu)\xrightarrow{F}N(0,\sigma^2) \tag{1.2.17}$$

其中，F 表示依分布收敛，记为 $\sqrt{n}(\bar{x}-\mu)\sim_{(a)}N(0,\sigma^2)$；$\sim_{(a)}$ 表示渐进分布(asymptotic distribution)。

二项分布、$\chi^2(n)$ 分布当 $n\to\infty$ 时接近正态分布都是中心极限定理应用的例子。根据二项分布的可加性，服从参数为(n,p)的二项分布随机变量，等于 n 个独立的 0-1 分布随机变量的和，而 $\chi^2(n)$ 等于 n 个独立的标准正态分布随机变量的平方和，除以 n 后为独立同分布随机变量的平均值，可以应用中心极限定律。从图 1.2 和图 1.4 可以看出，二项分布的概率函数散点图和 χ^2 分布的密度函数曲线随着 n 的增大接近正态分布密度曲线(见图 1.3)。

结论 5

不独立、不同分布的中心极限定理：$x_1,x_2,\cdots,x_n$ 为随机变量序列，如果 $\lim\limits_{n\to\infty}\mathrm{Var}(\sqrt{n}\bar{x})=\sigma^2>0$，并满足其他相关条件，则式(1.2.17)成立。

本书中在用到大数定律和中心极限定理时，不再逐一验证需要的条件，而默认这些条件满足，直接应用大数定律和中心极限定理的结论。

1.3　统计学复习

以概率论为基础，统计学(statistics)研究数据处理的方法和技术。要研究某个对象，首先需要从中获取数据。从研究对象获取数据的过程称为抽样，获取的数据称为样本。

1.3.1　样本和统计量

1. 样本

总体(population)是指研究对象的某个数量指标。例如(本章其余内容均以该例对有关概念和原理进行说明)，考查一名大学生的英语水平，用 100 分考卷得分衡量英语水平，总体就是该同学的英语分数。由于随机性(每次考出不同的分数)，英语分数为随机变量，用 X 表示。X 的概率分布刻画了该同学的英语水平。例如，数学期望 $E(X)$=85，表明该同学英语平均成绩(可以看作真实成绩)为 85 分，考分的随机性使得每次考试的具体分数围绕 85 上下波动，如果 $\mathrm{Var}(X)$=0.5，由切比雪夫不等式得出 $P(|X-85|>3)\leqslant 0.5/3^2\approx 0.056$，表明该同学英语考分十分稳定，每次考分上下波动超过 3 分的概率不超过 5.6%。

为了研究 X，需要从中进行抽样，抽取的样本(sample)为 $X_1,X_2,\cdots,X_n$。这里可以将抽样理解为该同学的 n 次英语考试，$X_1,X_2,\cdots,X_n$ 为考试分数。在统计学中，$X_1,X_2,\cdots,X_n$ 为随机变量而不是具体的数值，由此得出的统计量也是随机变量，这一点十分重要。如何理解样本是随机变量呢？统计方法研究着眼于事前(ex-ante)研究，即在实际抽样之前就要给出数据处理的方法——统计量，而实际抽样之前的样本是理论上的，可以是任何一个结果，因此是随机变量。进行实际抽样得到样本值之后，只需要将数值代入事先给出的统计量之中就可以得出数据处理结果。以英语水平的估计为例，$X_1,X_2,\cdots,X_n$ 是假设的 n 次英语考试分数，统计分析的任务是在得到具体的分数之前给出估计英语水平的方法。统计分析表

明，如果每次英语考试互不影响、试题难度相同、考生状态相同(相当于假定 $X_1, X_2, \cdots, X_n$ 独立同分布)，用 n 次考试分数的平均值估计 $\bar{X} = (X_1 + X + \cdots + X_n)/n$ 评价该同学的英语水平最为精确(在各种标准下最优)，统计分析到此结束。当得到一组具体考试成绩，比如 10 次考试成绩 84.5、85.2、83.1、…、86，代入 $\bar{X}$ 计算出英语水平估计值为 85.2，只是进行了算术运算，与统计分析无关。

2. 统计量及其分布

样本的函数称为统计量(statistic)。统计量中不能包含未知参数，给出样本值必定能够算出统计量的值。常用的统计量有样本均值和样本方差。设 X 为总体，数学期望 $EX = \mu$ 和方差 $\mathrm{Var}(X) = \sigma^2$ 未知。$X_1, X_2, \cdots, X_n$ 为从 X 中抽取的独立同分布样本，样本均值 $\bar{X}$ 和样本方差 S^2 定义为

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.3.1}$$

统计分析中常常需要知道统计量的分布。为了方便，假定总体服从正态分布 $X \sim N(\mu, \sigma^2)$。由正态分布的可加性和 χ^2 分布的定义经推导可以得出 $\bar{X}$ 和 S^2 的分布为

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \tag{1.3.2}$$

需要注意的是，即使不知道总体的分布，当样本量 n 很大时，由中心极限定律可知，$\sqrt{n}(\bar{X} - \mu) \sim_{(a)} N(0, \sigma^2)$，仍然可以用正态分布近似 $\bar{X}$ 的分布进行统计分析。对 S^2 的分布可以得出类似的结果：当 n 很大时，S^2 近似服从正态分布。

1.3.2 参数估计

参数估计是指通过样本提供的信息对总体 X 的概率分布中包含的未知参数进行估计的过程。

设总体 X 服从分布 $F(x, \theta)$，分布中包含未知参数 θ，$X_1, X_2, \cdots, X_n$ 为从总体中抽取的独立同分布样本，用 $X_1, X_2, \cdots, X_n$ 的函数 $h(X_1, X_2, \cdots, X_n)$ 估计参数 θ，函数 h 称为 θ 的估计量(estimator)，记为 $\hat{\theta} \equiv h(X_1, X_2, \cdots, X_n)$。$\hat{\theta}$ 是样本的函数，是随机变量。

1. 估计量的评价标准

估计量应尽量利用样本中包含的信息，使估计结果达到最大精确度。估计量的评价标准有三个：无偏性、一致性和有效性。

- 无偏性：如果 $E(\hat{\theta}) = \theta$，称 $\hat{\theta}$ 为 θ 的无偏估计。无偏性表明，作为一个随机变量，$\hat{\theta}$ 以 θ 为中心，围绕 θ 变动。
- 一致性：如果 $p\lim_{n\to\infty}\hat{\theta} = \theta$，称 $\hat{\theta}$ 为 θ 的一致估计(consistent estimator)。一致性表明，作为一个随机变量序列，当样本量 n 逐渐增加时，$\hat{\theta}$ 在概率收敛意义上无限接近未知参数 θ。
- 有效性：设 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 都是参数 θ 的无偏估计，如果 $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$，称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更为有效。

一致性是参数估计的重要性质，它表明，只要样本量足够大，估计结果可以无限接近未知参数。从极限状态来看，具有一致性的估计量在样本量无限大、样本信息无限多时，

不再有任何随机性，完全等于被估计参数，估计量之间没有区别。但实际上样本是有限的，有限样本下(无偏)估计量的优劣主要根据其有效性程度。因此，有效性是估计量的有限样本性质。

2. 估计量的求法

根据对总体的了解程度将参数估计方法分为矩估计方法和极大似然估计方法，仅采用总体 X 矩的有关信息对未知参数估计的方法称为矩估计法，采用总体的概率分布估计未知参数的方法称为极大似然估计。

1) 矩估计法

矩估计法(method of moment)是一种最古典的参数估计方法，其理论基础是大数定律。设总体分布为 $X \sim F(x,\theta)$，其中，θ 为未知参数(可以是包含多个未知参数的向量)。

结论 6

如果存在函数 $h(X,\theta)$ 使得

$$Eh(X,\theta)=0 \tag{1.3.3}$$

则称式(1.3.3)为总体矩条件(population moment conditions)。

例如，如果 $X \sim N(\mu,\sigma^2)$，则 $EX=\mu$，$E(X^2)=\sigma^2+\mu^2$，由此得出 $E(X-\mu)=0$，$E(X^2-\sigma^2-\mu^2)=0$，设 $h_1(X,\mu)=X-\mu$，$h_2(X,\mu)=X^2-\sigma^2-\mu^2$，得出总体的两个矩条件为

$$Eh_1(X,\mu)=0$$
$$Eh_2(X,\mu)=0$$

设 $X_1,X_2,\cdots,X_n$ 为从总体中抽取的独立同分布样本，$h(X_1,\theta),\cdots,h(X_n,\theta)$ 也是独立同分布的随机变量。$h(X_i,\theta)$，$i=1,2,\cdots,n$ 的均值为

$$\bar{h}=(X_1,X_2,\cdots,X_n,\theta)\equiv\frac{1}{n}\sum_{i=1}^{n}h(X_i,\theta)$$

$\bar{h}$ 称为总体均值 $Eh(X,\theta)$ 的对应量(counterpart)。由大数定律，当 $n\to\infty$ 时，样本均值依概率收敛于对应的总体均值为

$$\frac{1}{n}\sum_{i=1}^{n}h(X_i,\theta)\xrightarrow{\ p\ }Eh(X,\theta)=0 \tag{1.3.4}$$

用 $\bar{h}=(X_1,X_2,\cdots,X_n,\theta)$ 作为 $Eh(X,\theta)=0$ 的一致估计，称

$$\frac{1}{n}\sum_{i=1}^{n}h(X_i,\hat{\theta}_{MM})=0 \tag{1.3.5}$$

为样本矩条件(sample moment conditions)。显然，样本矩条件是将总体矩条件中的总体矩用对应的样本矩代替，并将参数 θ 用其估计 $\hat{\theta}_{MM}$ 代替得到的，这种方法称为类推原则(analogy principle)。从样本矩条件中解出 $\hat{\theta}_{MM}$，称为 θ 的矩估计量。从估计的原理看出，样本量趋于无限大时，矩估计依概率收敛于被估计参数，因此具有一致性。在满足一些条件时，矩估计量渐近服从正态分布。

矩估计法的关键在于寻找合适的总体矩条件。

例子 1　正态总体参数估计——矩估计法

设总体 X 的数学期望 μ 和方差 σ^2 为未知参数，总体矩条件为

$$E(X-\mu)=0,\quad E[X^2-(\sigma^2+\mu^2)]=0$$

$X_1,X_2,\cdots,X_n$ 为从 X 中抽取的样本，样本矩条件为

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})=0,\quad \frac{1}{n}\sum_{i=1}^{n}[X_i^2-(\hat{\sigma}^2+\hat{\mu}^2)]=0$$

从中解开未知参数的矩估计为

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X},\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i^2-\bar{X})^2$$

其中，$\bar{X}$ 为样本均值。

计算表明，$E(\hat{\sigma}^2)=[(n-1)/n]\sigma^2$，因此 $\hat{\sigma}^2$ 不是 σ^2 的无偏估计。为了得到无偏估计，对 $\hat{\sigma}^2$ 进行调整，得出 $n/(n-1)\hat{\sigma}^2$，调整得出的统计量便是样本方差 $S^2=n/(n-1)\hat{\sigma}^2$。

采用矩估计法，从公式(1.2.8)和公式(1.2.9)可以直接得出偏度和峰度的计量为

$$\hat{\lambda}=\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^3}{(S^2)^{3/2}},\quad \hat{\kappa}=\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^4}{(S^2)^2} \tag{1.3.6}$$

由于对总体的分布函数不做要求，矩估计法得出的估计量 $\hat{\theta}_{MM}$ 的精确分布未知。对参数进行假设检验，需要 $\hat{\theta}_{MM}$ 的分布。

结论 7

由中心极限定理可以得出

$$\sqrt{n}(\hat{\theta}_{MM}-\theta)\xrightarrow{F}N(0,\ \sigma^2) \tag{1.3.7}$$

其中，$\sigma^2=\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\hat{\theta}_{MM})$。

由此得出

$$\hat{\theta}_{MM}\sim_{(a)}N(\theta,\ \hat{\sigma}^2/n) \tag{1.3.8}$$

其中，$\hat{\sigma}^2$ 为 σ^2 的估计量；$\sim_{(a)}$ 表示近似或渐近服从。

在假设检验中，常采用参数估计的近似分布确定检验临界值。

矩估计法的最大优点是不需要知道总体的分布。矩估计法以总体矩条件为基础，如果矩条件不正确，由此得出的估计量不具有一致性。

2) 极大似然估计法

设总体服从的分布类型已知，分布包含未知参数 θ(可以是多个参数形成的向量)。对于离散型随机变量，设其概率函数为 $P(\xi=x_l)$，$l=1,2,\cdots,n$，设从总体中抽取容量为 n 的样本 $\xi_1,\xi_2,\cdots,\xi_n$，样本的联合概率函数 $L(\xi_1,\xi_2,\cdots,\xi_n;\theta)=\prod_{i=1}^{n}P(\xi_i=\xi)$ 称为似然函数。似然函数是未知参数 θ 的函数，使似然函数取得极大值的 θ 值称为极大似然(maximum likelihood，ML)估计，记为 $\hat{\theta}_{\mathrm{ML}}$。为使问题简化，对似然函数进行对数变换(对数变换为严

格单调变换，不改变函数的极值点)，变换后的函数称为对数似然函数，记为

$$l(\xi_1,\xi_2,\cdots,\xi_n;\theta)=\ln L(\xi_1,\xi_2,\cdots,\xi_n;\theta)=\sum_{i=1}^{n}\ln\left[P(\xi_i=\xi)\right] \tag{1.3.9}$$

当对数似然函数关于 θ 可导时，可以通过求导并令导数等于 0 得出极大似然估计 $\hat{\theta}_{\mathrm{ML}}$。

对于连续型随机变量，设其概率密度函数为 $f(x,\ \theta)$，样本似然函数为 $x_1,x_2,\cdots,x_n$ 的联合密度函数 $L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}f(x_i,\theta)$，对数似然函数为

$$l(x_1,x_2,\cdots,x_n;\theta)=\ln L(x_1,x_2,\cdots,x_n;\theta)=\sum_{ni=1}^{n}\ln[f(x_i,\theta)] \tag{1.3.10}$$

通过极大化对数似然函数可得出参数的极大似然估计 $\hat{\theta}_{\mathrm{ML}}$。

例子 1(续)　正态总体参数估计——极大似然法

设总体服从正态分布 $X\sim N(\mu,\sigma^2)$，μ 和 σ^2 为需要估计的未知参数。$X_1,X_2,\cdots,X_n$ 为从 X 中抽取的样本，似然函数(样本的联合密度函数)为

$$L(x_1,x_2,\cdots,x_n;\mu,\ \sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

对数似然函数为

$$l(x_1,x_2,\cdots,x_n;\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

对数似然函数对未知参数求导，令导数等于 0，得出

$$\frac{\partial l}{\partial\mu}=-\frac{1}{2\sigma^2}\times2\times(-1)\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\frac{\partial l}{\partial\sigma^2}=-\frac{n}{2\sigma^2}-\frac{1}{2(\sigma^2)^2}\times(-1)\times\sum_{i=1}^{n}(x_i-\mu)^2=0$$

从中解出参数的极大似然估计

$$\hat{\mu}=\bar{X},\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i^2-\bar{X})^2$$

从估计结果看出，对于服从正态分布的随机变量，极大似然估计法得出的结果与矩估计法得出的结果相同。

结论 8

与矩估计得出的估计量一样，极大似然估计量 $\hat{\theta}_{\mathrm{ML}}$ 也具有一致性和渐近正态性，即

$$\hat{\theta}_{\mathrm{ML}}\xrightarrow{\ p\ }\theta,\quad \sqrt{n}(\hat{\theta}_{\mathrm{ML}}-\theta)\xrightarrow{\ F\ }N(0,\ \sigma^2) \tag{1.3.11}$$

其中，$\sigma^2=\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\hat{\theta}_{\mathrm{ML}})$。

与矩方法估计量相比，极大似然估计更为有效：极大似然估计量的方差小于或者等于矩方法估计量的方差。极大似然估计方法的不足在于依赖总体的分布：需要事先假设总体

所服从的分布，如果假设的分布与真实分布不同，则估计量不具有一致性。

1.3.3 假设检验

参数估计是随机变量，当以参数估计值代替参数值进行推断时，必须考虑估计带来的随机成分。假设检验给出了以估计值为基础的推断方法。

仍以大学生英语水平测试为例。设原先的英语分数服从正态分布 $X \sim N(85, 0.5)$。为提高英语成绩，该同学参加了英语辅导班。辅导后进行 10 次英语考试，考试的平均分为 88 分。需要推断的问题是：辅导是否提高了英语成绩？之所以有这样的疑问，是因为 10 次的平均考分是随机变量，88 分只是随机变量的一次取值。据此，对英语成绩进行推断需要考虑随机因素：比 85 分提高了 3 分是考试的随机性带来的，还是真实水平提高所致？为此需要检验。检验之前需要给出原假设(null hypothesis)和备择假设(alternative hypothesis)。为简单起见，假定辅导后英语成绩仍服从正态分布，方差不变，即 $N(\mu, 0.5)$，需要检验 μ 是否仍然为 85。由此给出的待检验假设为

$$H_0 : \mu = 85; \quad H_1 : \mu > 85$$

将备择假设选为 $\mu > 85$ 而不是 $\mu \neq 85$，是考虑到辅导不会导致成绩下降。这样的检验称为单边备择假设(one-sided alternative hypothesis)检验。如果备择假设为 $\mu \neq 85$，则称为双边备择假设(two-sided alternative hypothesis)。选择单边备择假设还是双边备择假设直接影响检验结果，需要根据实际情况谨慎选择。

检验的基本思想是：如果辅导无效，该生的英语成绩无变化，仍然为 $X \sim N(85, 0.5)$ (即 H_0 成立)，那么辅导后 10 次考试的平均分大于等于 88 的概率有多大呢？设 $\bar{X}$ 为 10 次平均成绩，如果假设 H_0 成立，则 $\bar{X} \sim N(85, 0.5/10)$，标准化后得出 $\sqrt{10}(\bar{X} - 85)/\sqrt{0.5} \sim N(0, 1)$，查标准正态分布概率表得出 $P(\sqrt{10}(\bar{X} - 85)/\sqrt{0.5} > 1.645) = 0.05$，由此推出 $P(\bar{X} > 85 + 0.3678) = 0.05$。这表明，如果英语水平没有提高，该生 10 次考试平均分数高于 85.37 的概率仅为 5%，是一个小概率事件。根据小概率事件原理可知，小概率事件在一次具体抽样中是不会发生的。而实际的平均分数为 88 分表明，认为该生参加辅导后英语水平仍保持不变(即 H_0)是错误的。因此得出结论：辅导后该生英语水平提高了(接受 H_1)。临界值 85.37 将实数轴分为两部分：$(-\infty, 85.37)$ 和 $(85.37, \infty)$，当 $\bar{X} \in (85.37, \infty)$ 时，拒绝原假设；否则不能拒绝。因此，称 $(85.37, \infty)$ 为原假设的拒绝域。

如果将问题一般化：$X \sim N(85, \sigma^2)$，考试次数为 n 次，则推出的小概率事件为 $P(\bar{X} > 85 + 1.645\sigma/\sqrt{n}) = 0.05$。显然，在显著水平给定($\alpha = 0.05$)后，辅导后 10 次考试的平均分大于多少才能认为英语水平提高与两个因素有关。第一个因素是英语成绩的标准差 σ，标准差越小，表明成绩越稳定，随机成分越少，考出的平均分稍大于 85 分就会使我们相信是英语水平提高(而不是偶然因素影响)造成的，从而拒绝原假设；相反，如果标准差很大，表明成绩不稳定，随机成分很多，只有当考出的成绩远远大于 85 分才会使人相信是英语水平提高了。第二个因素是考试次数 n，n 越大，平均分的方差越小，随机成分越小，越能代表真实英语水平。

结论 9：假设检验的步骤

第一步：根据检验的问题提出原假设和备择假设。在给出备择假设时，要根据实际情况谨慎选择单边或者双边备择假设。

第二步：根据检验的需要设计检验统计量，并推导出在原假设下检验统计量的分布。

第三步：给定检验显著水平，求出检验的拒绝域。

第四步：根据样本值计算检验统计量的值，如果检验统计量的值落入拒绝域，拒绝原假设，否则不能拒绝原假设。

例子 2　正态总体假设检验——方差

设某学生的英语成绩服从正态分布 $X \sim N(85,2^2)$。为了提高英语成绩的稳定性，该生参加了英语辅导班。在参加完辅导班后参加了 10 次英语考试，考试成绩为 $X_1,\cdots,X_{10}$。经计算，平均成绩为 88 分，成绩的样本方差为 $S^2=[(X_1-88)^2+\cdots+(X_{10}-88)^2]/9=1.3$。设辅导后的英语成绩仍然服从正态分布 $N(\mu,\sigma^2)$，需要检验的问题是：辅导是否提高了该生英语成绩的稳定性？

根据问题提出的原假设和备择假设为

$$H_0:\sigma^2=2^2,\quad H_1:\sigma^2<2^2$$

在原假设下[即 $X_1,\cdots,X_{10}$ 来自总体 $X\sim N(\mu,2^2)$]，表达式为

$$\frac{(n-1)S^2}{2^2}=\frac{9}{4}S^2\sim\chi^2(n-1)=\chi^2(9)$$

给定显著水平 $\alpha=0.05$，查 χ^2 分布概率表得出 $P(2.25S^2<3.325)=0.05$，即 $P(S^2<1.478)=0.05$。因此，以 S^2 为检验统计量的拒绝域为(0, 1.478)。用样本值计算的 S^2 值为 1.3，位于拒绝域内，拒绝原假设，认为辅导提高了该生的英语成绩稳定性。

例 2 中的备择假设是 $\sigma^2<2^2$，样本方差 S^2 是 σ^2 的无偏估计量，对 σ^2 的推断要根据 S^2 的取值大小来完成，S^2 越小，则 σ^2 小的可能性越大。拒绝原假设意味着辅导后方差变小，因此拒绝域是 $\{S^2<1.478\}$。由此看出，假设检验的拒绝域与备择假设的选择有关。如果将例 2 中的备择假设改为双边假设 $\sigma^2\neq2^2$，那么检验的是辅导后英语成绩的稳定性是否发生明显变化(显著下降或者显著提高)，则显著大于 2^2 的 S^2 和小于 2^2 的 S^2 均能否定原假设，此时的拒绝域为两部分：$\{S^2<u_{\alpha/2}^{(L)}\}\cup\{S^2>u_{\alpha/2}^{(U)}\}$，下临界值 $u_{\alpha/2}^{(L)}$ 和上临界值 $u_{\alpha/2}^{(U)}$ 分别满足 $P\left(S^2<u_{\alpha/2}^{(L)}\right)=P\left(2.25\chi^2(9)<u_{\alpha/2}^{(L)}\right)=\alpha/2$ 和 $P\left(S^2>u_{\alpha/2}^{(U)}\right)=P\left(2.25\chi^2(9)>u_{\alpha/2}^{(U)}\right)=\alpha/2$。

例如，取显著水平 $\alpha=0.05$ 时，查 χ^2 分布概率表得 $u_{0.05/2}^{(L)}=1.22$ 和 $u_{0.05/2}^{(U)}=8.45$，得出拒绝域为 $\{S^2<1.22\}\cup\{S^2>8.45\}$。

按照采用的检验统计量将假设检验分类，常用的假设检验有采用 χ^2 分布统计量的 χ^2 检验，采用 t 分布统计量的 t 检验，等等。

本章小结

1. 计量经济学将经典计量与因果推断、大数据科学相结合，目的是在模型研究中充分利用“数据资料”，使其所揭示和描述的“经济如何运行的信息”与现实的经济运行实际更加吻合。它倡导“经济理论、数学、统计学结合”的本质，具有坚实的概率论基础，注重“利用现有的数据资料以提取关于经济如何运行的信息”，遵循“实践观察、行为分析和实验研究→经济理论(理论与模型)→建立总体回归模型→获取样本观测数据→估计模型→检验模型→应用模型”的研究步骤。

2. 随机变量分为离散和连续两种，离散随机变量用概率函数或者分布律描述其概率分布，连续随机变量则用概率密度函数描述其分布。概率函数和概率密度函数中包含参数。

3. 随机变量的数学期望是随机变量的一阶矩，方差是二阶中心矩。矩是随机变量分布的数字特征，是概率分布中参数的函数，决定了概率分布的位置和形状。由概率分布可以求出矩，但矩不能确定概率分布，概率分布对随机变量的描述比矩更为详细。

4. 随机向量的联合分布，不仅体现了各个随机变量的概率分布，还体现了随机变量之间的相互作用和影响。协方差和相关系数刻画了随机变量本身之间的相互影响，并不涉及随机变量函数之间的相互影响。条件期望和条件方差能够更深入地刻画随机变量之间的关系。$E(\eta|\xi)=0$ 给出的 ξ 和 η 之间的关系强于不相关，但弱于独立。

5. 条件期望满足迭代律，即对随机变量先取条件期望再取无条件期望，其结果等同于该随机变量的无条件期望。条件期望的另一条重要性质可以叙述为被条件确定的量可以移到条件期望的外面。

6. 极限定理是研究一组随机变量的平均值的极限性质。大数定律表明，随机变量的平均值 $\bar{X}$ 中的随机成分(用方差衡量)随随机变量的个数增加而减少，直至趋于 0，$\bar{X}$ 以概率趋于常数。中心极限定理则表明，不管参与平均的随机变量服从什么分布，随机变量个数足够大时，$\bar{X}$ 近似服从正态分布：$\sqrt{n}(\bar{X}-\mu)\sim_{(a)} N(0,\sigma^2)$，其中 $\sigma^2=\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\bar{X})$。$n$ 越大，近似效果越好，渐近正态分布称为 $\bar{X}$ 的大样本分布。

7. 参数估计分为矩估计法和极大似然估计法。矩估计法需要知道总体的矩条件 $Eh(X,\theta)=0$，在得到样本后采用类推原则得出样本矩条件 $n^{-1}\sum_{i=1}^{n}h(X_i,\hat{\theta})=0$，并从中解出参数估计量 $\hat{\theta}$。矩估计法的基础是大数定律。矩方法估计量具有一致性和渐近正态性。

8. 极大似然估计是将似然函数看作未知参数的函数、极大化似然函数得出的。极大似然估计具有一致性和渐近正态性，比矩方法估计量更有效，但极大似然估计需要设定总体的分布，分布设置不当会导致估计量的不一致。

9. 假设检验是以小概率事件原理为基础的对未知参数的推断。对应同样的原假设，备择假设可以是单边的，也可以是双边的，不同备择假设对应不同的拒绝域。假设检验的核心是根据待检验假设构造合适的检验统计量，并求出原假设下检验统计量的分布或者渐近分布。