经济、金融计量学中的非参数估计技术

李竹渝 鲁万波 龚金国 编著

科学出版社
www.sciencep.com

经济、金融计量学中的非参数估计技术

李竹渝　鲁万波　龚金国 编著

科学出版社
北京

内 容 简 介

本书利用非参数估计技术处理含有不确定性的、与实际现象密切联系的经济、金融模型. 主要内容包括：非参数核密度估计方法及其在金融资产收益率分布估计、资产组合相依结构 Copula 研究上的应用；常用非参数回归估计方法、基本统计性质及其在计量经济模型中的应用以及非参数估计技术在金融时间序列分析中的应用.

本书适合应用数学专业，特别是经济、管理和统计专业的高年级本科生、研究生及青年教师阅读. 可作为经济、管理类研究生学位课、选修课教材或参考书，也适合于实际从事经济管理、计量金融类的专业人员和有兴趣了解现代非参数估计技术的广大读者.

图书在版编目(CIP)数据

经济、金融计量学中的非参数估计技术/李竹渝，鲁万波，龚金国编著. —北京：科学出版社，2007

ISBN 978-7-03-019029-1

Ⅰ. 经… Ⅱ. ①李…②鲁…③龚… Ⅲ. 非参数统计-应用-计量经济学 Ⅳ. F224.0

中国版本图书馆 CIP 数据核字(2007)第 076542 号

责任编辑：陈玉琢 莫单玉 / 责任校对：包志虹

责任印制：赵德静 / 封面设计：王 浩

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2007 年 6 月第 一 版 开本：B5(720×1000)

2007 年 6 月第一次印刷 印张：10 1/2

印数：1—3 000 字数：192 000

定价：28.00 元

（如有印装质量问题，我社负责调换〈明辉〉）

序

从20世纪二三十年代起，计量经济学作为经济学的一个独立分支学科出现，对现代经济学的发展起了重大的推动作用. 而使用非参数模型的理论和方法研究计量经济，特别是研究宏观计量经济，则始于20世纪的60年代，其作为一种新的模型化方法使计量经济的研究获得新的动力.

《经济、金融计量学中的非参数估计技术》是作为国家社科基金项目的研究成果，用非参数模型和方法研究计量经济领域中若干重要问题的专著. 笔者有幸在该书的形成期间，阅读了初稿. 与同类著作相比，该书有面目一新的感觉，是一本内容充实、富于创新的好书.

使用Copula作为描述数据分布相依性的工具，并运用到计量经济领域诸多实际数据的统计分析，是该书最重要的特点；其次，该书十分重视模型的适用性，从而进行相应的模型诊断技术的研究，其中模型的误差分布的推断是该书的另一个重要特点；再次，该书拥有丰富的数据实例，详实的实证分析，这对于书中原理的理解和方法的运用都十分有益. 我相信该书的出版将会积极推动非参数方法在计量经济领域的使用和发展.

柴根象

2005年11月于上海同济新村

常用的简写符号

df (degree of freedom)

i. i. d. (independent and indentically distributed)

pdf (probability density function)

AMISE (asymptotic MISE)

AMSE (asymptotic MSE)

APLM (additive partial linear model)

ASE (averaged squared error)

$$\mathrm{ASE} \triangleq d_A(h) = n^{-1}\sum_{i=1}^{n}[\hat{m}_n(X_i) - m(X_i)]^2 w(X_i)$$

BCV (biased cross validation)

CV (cross validation)

DPI (direct Plug-in)

iff (if and only if)

ISE (integrated squared error) 可积平方误差

$$\mathrm{ISE} \triangleq d_I(h) = \int [\hat{m}_n(x) - m(x)]^2 f(x) w(x) \mathrm{d}x$$

MISE (mean integrated square error) 可积均方误差

$$\mathrm{MISE} \triangleq d_M(h) = E\int [\hat{m}_n(x) - m(x)]^2 w(x) \mathrm{d}x$$

ML (maximum likelihood)

MLE (maximum likelihood estimation)

MSE (mean squared error) 均方误差

$$\mathrm{MSE} \triangleq E[(\hat{m}_n(x) - m(x)]^2$$

NN (nearest neighbor)

PLM (partial linear model)

RSS (residual sum of squares)

S. D. (standard deviation)

S. E. (standard error)

S. T. (subject to)

STE (solved the equation)

SCV (smoothed cross validation)

Var (variance)

常用的数学、统计学的符号和记号

$[x]$　不超过 x 的整数；

$\hat{f}_n(x)$　$f(x)$的估计量；

$m(x)=E(Y|X=x)$　Y 在已知条件 $X=x$ 上的回归函数曲线；

$f\left(\frac{y}{x}\right)=\frac{f(x,y)}{f_x(x)}$　给定 $X=x$ 处 y 的条件密度；

$f^{(r)}(x)=\frac{\mathrm{d}f^{(r)}(x)}{\mathrm{d}x^r}$　密度函数 $f(x)$的 $r(r>0)$阶导函数；

$\sigma^2(x)=E(Y^2|X=x)-m^2(x)$　给定 $X=x$ 处 Y 的条件方差；

$x=\operatorname{argmax}g(u)$　iff$g(\cdot)$在 x 处有唯一最大值；

$x=\operatorname{argmin}g(u)$　iff$g(\cdot)$在 x 处有唯一最小值；

$F_1^n\triangleq\sigma((X_1,Y_1),\cdots,(X_n,Y_n))$　由$\{(X_i,Y_i)\}_{i=1}^n$生成的 σ 代数；

$F_n^\infty\triangleq\sigma((X_n,Y_n),\cdots)$　由$\{(X_n,Y_n),\cdots\}$生成的 σ 代数；

实数序列 a_n,b_n 有如下关系：

$$a_n=O(b_n)\quad \text{iff}\ \frac{a_n}{b_n}\to\text{常数}(n\to\infty)；$$

$$a_n=o(b_n)\quad \text{iff}\ \frac{a_n}{b}\to 0(n\to\infty)；$$

$$a_n\sim b_n\quad \text{iff}\ \frac{a_n}{b}\to 1(n\to\infty)；$$

随机变量序列 A_n，实数 A 有如下关系：

$$A_n\longrightarrow A\quad \text{a. s.}\quad \text{iff}\ P\{\lim_{n\to\infty}A_n=A\}=1,$$

$$A_n\xrightarrow{p}A\quad \text{iff}\ \lim_{n\to\infty}P\{|A_n-A|>\varepsilon\}=0；$$

可等价表为：$A_n-A=o_p(1)$，　$n\to\infty$.

目　录

第1章 引 言

1.1 计量经济模型简述

如果以1930年12月29日世界计量经济学会成立及其学术刊物《Econometrica》1933年正式出版作为计量经济学正式以经济学的一个独立分支诞生的标记，那么计量经济学已经有70多年的历史了.70年来，计量经济学的迅速发展与现代统计学和计算科学等学科的急速发展密不可分，无数从事计量经济学研究的工作者利用统计学、数学和计算科学等知识，展开对社会经济、金融发展现象的科学研究，取得了令人瞩目的成绩，从超过半数以上纪念诺贝尔经济学奖获得者主要从事数量经济学的研究就可以说明计量分析在经济、金融的发展中有不可撼动的强势地位.这里我们用数量经济学(Quantity Economics)一词，其也包含了通常意义上的数理经济学(Mathematical Economics)，数理经济学可以简单理解为列方程和解方程，它是将数学概念和方法应用于经济学.首先，用数学公式来描述经济系统中的基本现象，如效用函数、需求供给变量、收入消费储蓄函数等，利用数学联立方程组来描述这些基本变量之间的因果关系；然后再进一步解这个联立方程组，按要求讨论解的存在性、稳定性、合理性等.如果没有进一步的模型估量，数理经济学始终停留在理论阶段.当人们使用计量经济学一词时，其专指英语单词是Econometrics，国内也有人译作经济计量学，这是由1969年获得第一次诺贝尔经济学奖得主之一、挪威的R.弗里希(Ragnar Frish)教授1926年仿照英语的生物计量学Biometrics最先使用的，弗里希教授被尊称为计量经济学的奠基人.在首期《Econometrica》杂志上，作为编辑给出的出版说明中，弗里希教授指出，计量经济学的“主要目的应该是促进有助于理论数量方法与实证数量方法相统一的研究”，“计量经济学不应看作是数学在经济学中应用的同义词.经验表明，统计学、经济理论和数学对理解现代经济生活的定量关系都是必须的，但其中任何一种都是不够的，三者的结合才是强有力的，且正是这三者的结合构成了计量经济学”.国际著名计量经济学家W.Green的专著《Econometric Analysis》开篇就明确地提出：“计量经济学是经济学的一个领域，它运用数理统计和统计推断工具对经济理论所假定的关系进行实证研究”.

所以，计量经济学运用数理统计知识分析推断经济数据，对构建于数理经济学基础之上的数学模型提供了实证支持.计量经济学若无数理经济阶段强有力的推导过程和结论，所涉及的经济函数的建立、经济行为方程等将缺乏一定的理论基

础. 科学地讲，数量经济学的发展依赖于数学(形式科学)、统计学(工具性学科)、计算科学(工具性学科)与经济学的交叉发展. 从 1969 年第一次诺贝尔经济学颁奖以来，绝大多数获奖成果均属于经济学与数学、统计学交叉学科的研究. 计量经济学是专门研究经济现象的数量特征、数量关系和数量变动的经济学，也是一门研究经济问题的方法论学科，它受宠于诺贝尔经济学奖的主要原因应该说是计量经济学通过把经济理论具体化、经济现象数量化，使人们能够更深刻地理解和掌握经济规律、把握经济现象的内涵和本质，更好地按客观经济规律行动和决策.

如果说计量经济学首先是要用数学方法把经济关系表述为数量模式(模型体系)，再运用数理统计方法进行逻辑推演和统计推断，然后将所得结果赋予经济意义的解释，那么在运用数学方法把经济关系、或经济理论表述为模型体系时就不得不给出一定的假设条件. 现实经济现象错综复杂，人们要阐明某个特定的经济问题时势必要进行去繁化简、将要分析的对象变为可以理解且能够把握处理的问题，而所谓假设条件实际上是对特定经济问题进行模型分析的一些前提条件. 例如考虑一个多元线性计量经济模型

$$y = \chi\beta + \varepsilon \tag{1.1}$$

这里我们已经设定它是针对某个特定经济问题的合理模型，其中 y 是响应变量或被解释变量，可观测；$\chi = (x_0, x_1, \cdots, x_k)$ 是解释变量向量的矩阵，x_j 可以是随机的，也可以是预先设计的或可观察的固定变量，$j = 0, 1, \cdots, k$；ε 为不可观测的随机误差，通常也称为模型的扰动项；$\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^{\mathrm{T}}$ 表示 $k+1$ 维未知参数的列向量. 通过估计 β，我们可以找出解释变量与响应变量之间的数量关系，或者说可以通过 β 的估计值和 χ 的观测值来得到的响应变量 y 的估计或预测. 在计量经济模型的实际应用中，如果我们已经得到 (χ, y) 的 n 个观测值，则第 i 个模型可表为

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1, \cdots, n$$

对这样一个模型，有很多途径可以求得未知参数 β 的估计，如常用的最小二乘法、极大似然估计法. 这些方法，特别是后者，其实都依赖于对模型的基本假设. 概括地讲，为了使预先设定的模型有意义且能很好地拟合所要描述的特定的经济现象，同时也为了能较为有效地做出统计判断，我们需要对模型(1.1)预先给出一组假设条件. 这些假设条件基本上是由模型自身数学和统计学运用的背景所要求的，但也适当考虑实际背景. 原始假设条件大致分为三大类.

Ⅰ. 首先针对解释变量：计量经济模型中一般假设解释变量构成矩阵 $X = (x_{ij})$ 为可观测的，满足 $R(X) = k+1$，即 X 为满秩阵，各解释变量之间不相关，或通常所称无多重共线性. 使用最小二乘法时，对 X 还需假定 $(X^{\mathrm{T}}X)^{-1}$ 的存在.

Ⅱ. 其次是针对模型的扰动项的假设，即假设误差项满足零均值 $E\varepsilon_i = 0, i = 1, 2, \cdots, n$，互不相关，或者在时间域观测响应变量和解释变量时，扰动项非自相关 $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$，方差有限且相同，即满足 $E\varepsilon_i^2 = \sigma^2, i = 1, 2, \cdots, n$，一般 σ^2 未知.

如使用极大似然估计法，我们通常还需假设随机误差服从正态分布，即

$$\varepsilon_i \sim N(0,\sigma^2), \qquad i=1,2,\cdots,n \tag{1.2}$$

这也是后面进行统计推断与模型假设检验的必要条件.

Ⅲ. 最后一类假设是针对解释变量与扰动项之间的相关关系，即一般需假设模型的解释变量与扰动项之间不相关：$\mathrm{Cov}(x_{ji},\varepsilon_i)=0, i=1,2,\cdots,n, j=1,\cdots,k.$

众所周知，在这些假设下，利用各种数理统计方法，可以得到模型(1.1)中未知参数 β 的较好的估计，即 β 的估计可以具有统计上一些基本的优良性质，如无偏性、有效性、一致性等. 注意，仅在误差分布已知或确定的假设下，可以进行模型的显著性检验，进而还可以给出估计参数和计量模型的置信区间. 一个好的经典计量经济模型，是指它在满足这些假设条件下与实际观测到的经济数据拟合的误差最小的模型.

但是，这里给出的前提是我们已经设定(1.1)是针对某个特定经济问题的合理模型！显然，如果这个设定不正确，那么无论选择什么样的统计估计方法，无论如何改进统计技术，所得到的结论与实际经济现象都会产生较大的偏差. 所以，一直以来，模型的设定问题都是计量经济学家、特别是微观计量经济学家处理实际模型中比较棘手的问题，也是实际计量经济建模研究中遭受质疑最多的问题，有人甚至说模型所依赖的设定和假设条件是微观计量经济学家的“天敌”. 20 世纪七八十年代以来，人们对计量经济学模型的研究已经开始聚焦于自由度更大的非稳定、非线性、非参数统计方法上，其本质就是努力让观测到的经济数据能直接与计量模型“对话”，放宽对计量经济模型的约束条件，减少人为的主观性或从先验的经济理论建模的偏差，建立更紧密接近经济实际的计量模型. 那么，什么是非参数估计方法？它在经济、金融计量模型的分析中如何运用呢？下面先简要介绍非参数估计，然后逐章开始讨论非参数密度估计、非参数回归估计和金融时间序列的非参数估计，以及在计量经济模型中的应用.

1.2 非参数估计方法简介

什么是非参数估计方法？要想用一两句话来回答这个问题不是一件容易的事，这里从两个方面讨论. 首先是非参数密度估计. 计量经济模型的研究中，涉及的研究对象指标作为随机变量，通常已经给出它的概率分布的假设. 如经典计量经济模型中，人们通常假设随机扰动来自正态分布，有两个参数：期望均值(位置参数)与方差(离散程度参数)，这是决定正态分布形态的两个参数，一般这两个参数是未知的，需要进行统计估计. 写成统计形式即为 $x_1,\cdots,x_n \sim f(x;\theta)$，$f(x;\theta)$ 的函数形式是已知的，其中 θ 表示概率分布的未知参数向量，正态分布中 $\theta=(a,\sigma^2)^{\mathrm{T}}$，$a$ 为期望均值，σ^2 表方差，一个正态分布唯一地由它的期望均值与方差决定. 数理统计

中所提到的参数估计一般就是讨论如何估计函数 $f(x;\theta)$ 中的未知参数向量 θ. 如果预先并不假设随机变量所服从的分布类型，仅仅给出概率密度函数存在的条件，即 $x_1,\cdots,x_n\sim f(x)$，$f(x)$ 是一未知密度函数，则相应的计量经济模型为非参数的. 非参数密度估计要解决的是在一定的条件下，对未知密度函数 $f(x)$ 的估计. 由于预先并不假设 $f(x)$ 有具体的函数形式，故称为非参数密度估计. 这是首先要在第 2 章中介绍的，它的发展对后来的非参数回归估计起到了重要的推动作用.

其次，是非参数回归估计，这可以说是非线性回归模型讨论的一个延伸. 考察被解释随机变量 Y，解释变量向量 $\chi=(x_0,x_1,\cdots,x_k)$，其中的各分量可以是随机的，也可以是确定性的，则在 $E|Y|<\infty$ 条件下，Y 对 χ 的回归函数存在，且表示为 $m(x)=E(Y|\chi=x)$. 计量经济模型分析的基本问题之一是通过从 (χ,Y) 的可观测相互独立样本 (X_i,Y_i)，$i=1,2,\cdots,n$ 去估计回归函数 $m(x)$. 经典计量经济模型的分析中，处理的回归函数基本为给定特定函数形式的模型，如线性回归、非线性回归等. 此时，一般采用最小二乘法或它的扩展，如加权最小二乘法、稳健回归方法等. 在讨论数理统计学对计量经济学产生和发展所起的作用时，不能不首先提到最小二乘法. 最小二乘法自它最早被人们使用以来，已有 200 多年的历史，可广泛用于线性模型、可化为线性的非线性模型等，迄今仍是人们利用各种曲线和方程去拟合数据的最主要方法. 其原理是：对形同(1.1)并满足假设Ⅰ，Ⅱ和Ⅲ的计量经济模型，如改写(1.1)为 $\varepsilon=y-\chi\beta$，可以这样来解释误差 ε，即从观察向量 y 中扣除模型(1.1)的主要部分 $\chi\beta$ 后的剩余. 因而也称为残差，其平方和 $\varepsilon^{\mathrm{T}}\varepsilon=(y-\chi\beta)^{\mathrm{T}}(y-\chi\beta)$，最小二乘法原理就是要求 β 使残差平方和 $\varepsilon^{\mathrm{T}}\varepsilon$ 达到最小，而称那个达到最小值的 β(一般记为 $\hat{\beta}$)为最小二乘估计量. 无论对线性经典计量经济模型和 20 世纪 60 年代逐渐兴盛起来的非线性计量经济模型的分析，以及对传统最小二乘法和它的扩展来说，此时计量经济模型要解决的问题均属于参数回归的范畴. 前面已经指出，在给出的模型设定下，当满足一些假设条件时，都可以获得满意的参数估计值，从而获得较满意的计量模型分析结果.

但是，如果预先给出的模型设定不正确，如要考察的实际经济变量之间并不存在线性或多项式的相关关系，如果仍利用最小二乘法按线性回归或多项式回归模型分析，则结果是没有实际价值的，不能用于对经济现象的解释. 1964 年，美国的 G. Watson 和前苏联的 E. A. Nadaraya 分别在当时著名的刊物《Sankhya》和《Theory of Probability and Application》上各自独立发表了一种被称为核函数估计的统计方法，利用非参数密度估计的思想，直接对未知泛函数形式的回归函数 $m(x)=E(Y|\chi=x)$ 进行估计，并且在渐近形态下，给出了相关的统计性质. 后来被广泛使用的这种非参数回归的核估计方法，就以 Watson-Nadaraya 命名. 1977 年，美国人 Stone 在《统计年刊》上发表了他的著名论文，提出了对 $m(x)=E(Y|\chi=x)$ 进行估计的更为一般的权函数方法，并在理论上论证了这种方法的优良性(同样是

在大样本情况下). 无论是 Watson-Nadaraya 核估计还是更一般形式的权函数估计,因为不涉及回归函数 $m(x)=E(Y|\chi=x)$具体的数学形式,统称为非参数回归估计,并与非参数密度估计区别. 我们提到的非参数估计的内容,涵盖了非参数密度估计与非参数回归估计两部分. 在第 3 章将展开对非参数回归估计的讨论,除了详细介绍这里提到的方法外,也将介绍后来发展的更多的方法,如目前使用较广泛的局部多项式估计方法等.

可以看到,非参数估计方法,最大程度地放宽了经典计量经济模型的结构条件,特别是对微观计量经济模型的分析,能够在更宽松、更大自由度的条件下分析模型. 所以非参数估计方法成为继非线性系统分析方法后,在现代计量经济学中的又一研究热点,特别是近几年已经构成非参数计量经济学这一新的研究分支,并且在金融计量经济学研究中也占有重要地位. 为了能将非参数估计技术更方便地应用于计量经济模型的分析,本书在讨论这些估计方法的理论基础之上将着重介绍如何使用非参数估计技术及其在计量经济模型中的应用. 第 2 章主要介绍了非参数密度估计的理论、方法和基本性质,以及在线性模型未知误差密度的估计、金融股票市场收益率分布估计中的应用. 最后还介绍了 Copula 函数及它在金融资产市场相关结构研究中的应用,并提出非参数核密度估计——极大似然估计方法来估计和检验 Copula 函数中的未知参数,在放宽假设条件下选择合适的 Copula. 第 3 章介绍非参数回归估计的基本内容,包括介绍主要的非参数函数方法,如核权估计、最近邻估计、局部多项式估计、样条光滑和小波估计方法等. 除此之外,还介绍了如何利用交叉核实(CV)方法等方法来选择满意的光滑参数. 作为非参数回归估计在计量经济模型中的一个重要应用,还介绍了如何利用非参数回归技术来估计微观计量模型层面上居民的收入分布模型,给出一些实际应用结果.

金融时间序列的非参数估计研究已经成为金融计量经济学的重要部分,将在第 4 章给出相关的讨论,并介绍在研究金融资产收益率波动性非参数估计的结果,同时介绍非参数估计技术在金融计量研究中未来的发展. 本书侧重点不是研究非参数估计的理论,而是讨论其实际运用技术,所以在介绍应用结果时,也分别给出非参数估计方法实际运用中所涉及的模型解释、光滑参数选择等实际问题及部分使用的统计程序(如 S-Plus). 作为思考与练习,各章后面附有相应的习题. 读者可根据需要选用,在收集相关数据的基础上,运用附录中的程序进行实际计算和分析,以达到能掌握非参数估计技术的目的. 我们希望本书能引起广泛的讨论和评价,以期获得更全面、更合适的修改和建议.

参考文献

李子奈，叶阿忠. 2000. 高等计量经济学. 北京：清华大学出版社

Cameron A C, Pravin K T. 2005. Microeconometrics: Methods and Applications. New York: Cambridge University Press

Frish R. 1933. Econometrica. Editorial, 1:1～4

Green W. 2002. Econometric Analysis, 5th edition. Prentice Hall

Nadaraya E A. 1964. On estimating regression[J]. Theory Prob. Appl., 9:141～142

Stone C J. 1977. Consistent nonparametric regression. The Annals of Statistics, 5:595～620

Watson G S. 1964. Smooth regression analysis[J]. Sankhya Ser. A, 26:359～372

第 2 章　非参数密度估计及其应用

2.1　非参数密度估计简介

随机变量所服从的概率密度函数，或称为概率分布，是将数理统计学应用于解决实际问题最重要的基础之一，人们通过对随机变量的观测样本，在可以确定其泛函形式的概率密度函数的前提下，讨论其统计特征，如我们熟知的测量误差可视为来自正态密度分布，投掷一枚均匀硬币所得随机变量序列服从二项分布等. 人们已经习惯于在对某经济现象进行统计分析之前，假设该模型所含的随机变量来自一定的概率密度函数，所需要的仅仅是在观测样本基础上对概率密度函数中的未知参数进行估计. 如讨论一个人群集合中的收入分布，以变量 x 表示个人的月收入，人们可以较合理地认为：收入分布的密度曲线呈两头小、中间大的状态，也就是最高收入与最低收入人数所占比例较小，绝大多数人收入水平处于中等，可以粗略地假设 x 服从正态分布 $N(\mu,\sigma^2)$，其中期望 μ、方差 σ^2 未知. 可以通过对 x 的一组简单样本，如 $X_1,\cdots,X_n$ 来估计期望和方差 μ,σ^2 的大小，然后进行一系列的统计推断，如期望的估计值约等于 1500，认为该人群的月收入平均为 1500 元左右；而如果方差的估计值很大，则推断所搜集的样本值大小很分散等. 如果再进一步利用直方图等简单统计工具，或许还可以观察到密度曲线的粗略形态，如果右边尾部较长，可以推断高收入的人数较低，收入得多等. 人们可以由随机样本的密度函数形态粗略地推断总体分布的情况，帮助人们更好地理解某些经济现象. 只要确定总体分布的密度函数，就可以解决一系列的实际统计问题，如对一般计量经济模型的参数估计、假设检验、模型诊断等.

但由于研究对象的复杂性和多样性，以及随机抽样的时间性，要使某随机变量满足某给定的密度函数假设已经越来越多地受到质疑. 在分析很多常见经济现象中的随机变量时，如金融市场中的即时收益率(instantaneous return)和收益率的波动变化等，往往需要预先给出这些随机序列的密度函数. 在经典线性回归计量经济模型中，人们也通常给出扰动是来自正态分布的假设条件. 实际中，经济变量的概率密度往往是未知的，观测样本很难完全吻合预先给定的假设条件. 所以人们希望能够通过从随机变量抽取样本来研究随机变量的密度分布形态，最熟悉的即是直方图方法. 但直方图方法过于粗糙，估计精度较差，不能满足严格统计分析的需要. 随着数理统计学在 20 世纪中叶的迅速发展，1955 年美国的 M. Rosenblatt 首次提出对直方图方法的改进，正式提出了一维随机变量密度函数的非参数估计. 他

1956 年正式发表在美国《Ann. Math. Statistics》上的一篇文章，成为非参数密度函数估计早期文献中被引用最多的一篇. 非参数密度估计，是指随机变量 x 服从一未知的概率密度函数 $f(x)$，这里 $f(x)$ 只是一个记号，不具有任何具体的泛函形式，要利用 x 的一组简单抽样样本 $X_1,\cdots,X_n$ 对函数 $f(x)$ 进行统计估计. $f(x)$ 之所以称为“非参数”，是因为它不能通过设定有限个参数来确定. 自 1956 年 M. Rosenblatt 正式发表他的上述工作以来，已经有很多学者对密度函数的非参数估计做了大量的研究，其中 Parzen 的核估计法(1962)；Loftsgarden，Quesenberry 的非参数最近邻密度估计方法(Nearest Neighbor Estimate)(1965)；Schwartz 的正交多项式法(1967)；Kronmal，Tarter 的傅里叶级数法(1968)；Wahba 的样条光滑法(1975,1990)；Walter G，Blum J 的 δ 序列法(1979)；Walter，Ghorai(1992) 及 Donoho，Johnstone (1994，1995)的小波光滑等. 目前实际应用较多、计算软件较通用的是核密度估计，本章主要围绕核密度估计开展讨论.

本节主要按照历史发展的顺序，简要介绍三种非参数密度估计的方法. 这里只是假设随机变量 x 有一组简单随机样本 $X_1,\cdots,X_n$，x 的密度函数 $f(x)$ 未知. 如何估计未知函数 $f(x)$，通常可采取直方图、Rosenblatt 和 Parzen 三种估计方法.

1. Naive 估计(直方图方法)

直方图显然是人们最熟悉的，可以视为早期的一种简单、初等的非参数密度估计方法. 由概率论基础知识可知，随机变量 x 在直线区间 $[a,b]$ 上的概率为 $P(a\leqslant x\leqslant b)=\int_a^b f(x)\mathrm{d}x$，$f(x)$ 表示 x 的密度函数. 如记样本 $X_i,i=1,\cdots,n$，落入区间 $[a,b]$ 的点的总个数为 $\#\{i:1\leqslant i\leqslant n,a\leqslant X_i\leqslant b\}$，则由大数定律有

$$P(a\leqslant x\leqslant b)\approx \#\{i:1\leqslant i\leqslant n,a\leqslant X_i\leqslant b\}/n$$

因此

$$\frac{1}{b-a}\int_a^b f(x)\mathrm{d}x=\frac{P(a\leqslant x\leqslant b)}{b-a}\approx\frac{\#\{i:1\leqslant i\leqslant n,a\leqslant X_i\leqslant b\}}{n(b-a)} \tag{2.1}$$

当 $b-a$ 充分小时，$\frac{1}{b-a}\int_a^b f(x)\mathrm{d}x$ 可近似代表 $f(x)$ 在 $[a,b]$ 上的值，这就得到了 $f(x)$ 的一个估计. 这样的方法传统上称为直方图方法，是人们早期用来估计未知密度函数常用的一种方法. 迄今，在统计教科书和几乎所有的统计计算软件中，都仍包含直方图方法的介绍和计算作图. 基于上述原理，选择一个适当的正数 h，可以把全直线平分为一些长为 h 的小区间，任取一小区间，记为 I. 则与(2.1) 类似，考虑以频率

$$\frac{1}{nh}\#\{i:1\leqslant i\leqslant n,X_i\in I\} \tag{2.2}$$

来估计 $f(x)$.

例 2.1 直方图 2.1～图 2.3 描绘的是同一组样本数据的概率分布(或更准确地称为频率直方图). 图 2.1 的组距看起来恰到好处;图 2.2 的组距似乎过大,会突出平均化,掩盖数据自身变化的一些细节;而图 2.3 的组距过于窄小,显然图形受随机影响太大,会产生不规则的形状.

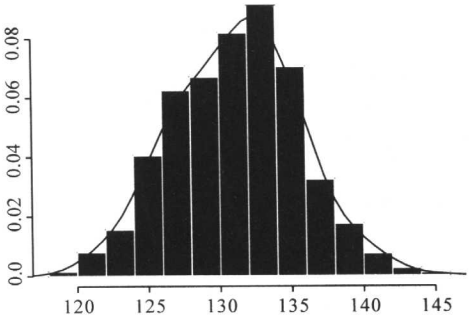

图 2.1 组距适中

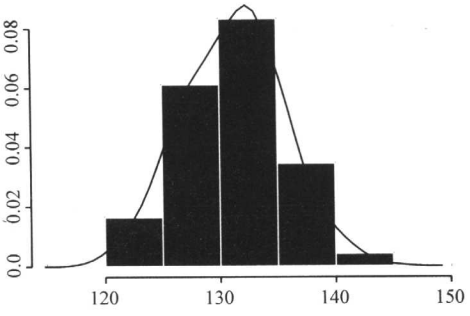

图 2.2 组距过大

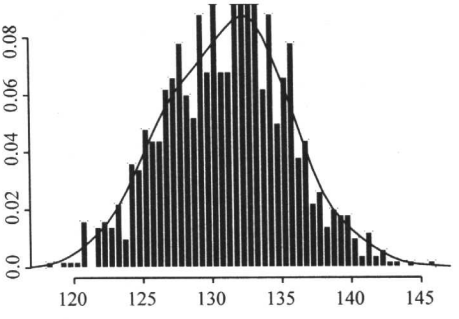

图 2.3 组距过小

为平衡这两种偏差,实际统计工作者一般借助经验,根据样本量的大小来选取合适的组距,即常数 h. 这种直方图的做法尽管粗糙,用于估计"真实"密度函数会有较大的偏差. 但由于简便易行,实际中仍在广泛使用. 如果进一步放宽 h 为固定的这一限制,让 h 随在给定区间上的样本个数的多寡而变,情况肯定会得到改善. 也就是说,可以根据样本个数在区间上不同区段的个数分布,选取不同的小区间长度 h. 直方图的不足是很明显的,由于估计为阶梯函数,图形不光滑,并且对每个小区间中心估计相对准确而区间边缘部分的估计较差.

2. Rosenblatt 估计

为了克服每个小区间边缘部分密度估计较差这一缺点,1955 年Rosenblatt对