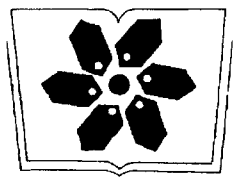
现代数学基础丛书 125

非参数蒙特卡罗检验及其应用

朱力行　许王莉 著

科学出版社
www.sciencep.com

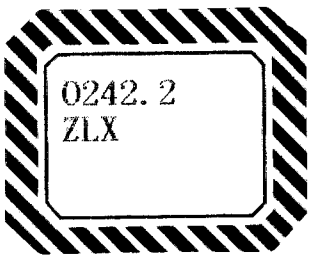


中国科学院科学出版基金资助出版

现代数学基础丛书 125

非参数蒙特卡罗检验及其应用

朱力行 许王莉 著



科学出版社
北京

内 容 简 介

本书提出一种新的产生参考数据的方法构造条件统计量，称之为非参数蒙特卡洛检验(NMCT). 全书共分 11 章：第 1 章介绍蒙特卡罗检验；第 2 章用 NMCT 方法检验 4 种类型的分布，并且说明此方法对这些类型的检验精确有效；第 3 章证明 NMCT 方法对 4 种情况是渐近有效的，而且 pn 相合；第 4~6 章研究了回归模型的模型检验问题，也说明了 Wild 自助法在某些情况下不相合；第 7~9 章研究了一些用自助逼近法可以实现的问题，但是 NMCT 方法也很容易实现，而且功效很好；第 10~11 章分别介绍协方差矩阵的同方差检验和参数型 coupula 函数的拟合检验.

本书特别适合重抽样逼近领域或者是将重抽样逼近技术应用到其他应用领域的研究人员，以及对拟合优度检验方向有兴趣的学者.

图书在版编目(CIP)数据

非参数蒙特卡罗检验及其应用/朱力行　许王莉著. —北京：科学出版社，2008

(现代数学基础丛书；125)

ISBN 978-7-03-022578-8

I. 非… II.①朱… ②许… III. ①蒙特卡罗法②非参数检验 IV. 0242.2　0212.1

中国版本图书馆 CIP 数据核字(2008)第 111833 号

责任编辑：陈玉琢　杨　然/责任校对：朱光光

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

铭浩彩色印装有限公司印刷

科学出版社编务公司排版制作

科学出版社发行　各地新华书店经销

*

2008 年 8 月第　一　版　　开本：B5 (720 × 1000)

2008 年 8 月第一次印刷　　印张：11 1/2

印数：1—3 000　　字数：210 000

定价：36.00 元

(如有印装质量问题，我社负责调换〈路通〉)

《现代数学基础丛书》序

对于数学研究与培养青年数学人才而言，书籍与期刊起着特殊重要的作用.许多成就卓越的数学家在青年时代都曾钻研或参考过一些优秀书籍，从中汲取营养，获得教益.

20 世纪 70 年代后期，我国的数学研究与数学书刊的出版由于文化大革命的浩劫已经破坏与中断了十余年，而在这期间国际上数学研究却在迅猛地发展着. 1978 年以后，我国青年学子重新获得了学习、钻研与深造的机会. 当时他们的参考书籍大多还是 50 年代甚至更早期的著述. 据此，科学出版社陆续推出了多套数学丛书，其中《纯粹数学与应用数学专著》丛书与《现代数学基础丛书》更为突出，前者出版约 40 卷，后者则逾 80 卷. 它们质量甚高，影响颇大，对我国数学研究、交流与人才培养发挥了显著效用.

《现代数学基础丛书》的宗旨是面向大学数学专业的高年级学生、研究生以及青年学者，针对一些重要的数学领域与研究方向，作较系统的介绍. 既注意该领域的基础知识，又反映其新发展，力求深入浅出，简明扼要，注重创新.

近年来，数学在各门科学、高新技术、经济、管理等方面取得了更加广泛与深入的应用，还形成了一些交叉学科. 我们希望这套丛书的内容由基础数学拓展到应用数学、计算数学以及数学交叉学科的各个领域.

这套丛书得到了许多数学家长期的大力支持，编辑人员也为其付出了艰辛的劳动. 它获得了广大读者的喜爱. 我们诚挚地希望大家更加关心与支持它的发展，使它越办越好，为我国数学研究与教育水平的进一步提高作出贡献.

杨　乐

2003 年 8 月

前　言

2005 年斯普林格出版社出版了朱力行的英文专著 *Nonparametric Monte Carlo Tests and Their Applications*, 这本书是基于朱力行和他的合作者的研究成果, 以及他在华东师范大学开设讨论班的相关资料写成的. 书中主要介绍了一种新的统计检验方法, 即非参数蒙特卡罗检验, 并将这种方法运用到各种半参数和非参数模型的检验问题.

在统计推断中, 用蒙特卡罗方法去逼近统计量的分布已成为非常重要的研究分支, 其主要思想是通过产生参考数据, 构造新的分布去逼近基于观测数据得到的统计量分布. 因而, 在这个研究领域, 如何产生参考数据至关重要. 在参数情况下, Barnard(1963) 首次提出蒙特卡罗检验 (MCT). MCT 有一些很好的性质, 非常类似于其后发展起来的参数自助法 (parametric bootstrap). 在这之后, 人们对 MCT 法有较多的研究, 如 MCT 最优性和计算功效的研究. 然而, 在半参数甚至非参数的情况下, 如何产生这样的参考数据是一个具有挑战性的难题.

非参数蒙特卡罗检验 (NMCT) 就是针对这个问题提出的. NMCT 的算法很容易实施, 并且在很多情况下, 检验精确有效. 此外, 逼近的精确性相对比较容易研究, 如第 3 章的相关内容就做了这方面的探讨.

我们一直在考虑写一个中文版, 以方便中文读者. 因而, 我们对这本专著的内容做了进一步的充实, 加进了第 11 章. 许王莉博士翻译和整理了全部的内容, 形成中文书稿的基本结构. 朱力行在此基础上做了进一步的整理. 我们也重写了中文稿的前言部分.

在此, 我们要感谢我们的主要合作者, 其中包括 Y. Fujikoshi, K. Naito, G. Neuhaus, K. W. Ng, W. Stute, K. C. Yuen; 第 8 章是许王莉和朱力行共同完成的, 并且是许王莉博士论文的一部分; 第 6 章是许王莉和朱若青一起完成的未发表的文章, 后者在文章中负责模拟部分并与华东师范大学的博士生武萍、於州和朱利平博士共同完成第 11 章. 2002~2003 年, 在华东师范大学开设讨论班期间, 复旦大学朱仲义博士和华东师范大学张志强博士对本书也给予了很好的建议.

英文专著 *Nonparametric Monte Carlo Tests and Their Applications* 得到香港大学和香港研究基金 (HKU7129/00P; HKU7181/02H HKU7060/04P) 的部分资助. 作为洪堡研究奖 (Humboldt Research Award) 的获得者, 朱力行在访问德国 Giessen 大学和 Hamburg 大学期间, 也受到德国 Alexander-von 洪堡基金的资助, 使他在教学之余完成斯普林格出版社的专著. 斯普林格出版社的编辑 John Kimmel 先生在撰写此书期间给予了极大帮助. 此后, 在我们撰写这本中文稿期间, 香港浸会大学和国家自然科学基金 (10701079) 提供了部分资助. 作为长江讲座教授, 朱力行

试读结束, 需要全本PDF请购买 www.ertongbook.com

也得到中国人民大学的支持. 特别是科学出版社的陈玉琢女士, 给予了有益的建议并专门为此书申请了出版基金. 作者在此一并表示深深的谢意.

朱力行
香港浸会大学
许王莉
中国人民大学
2007 年 3 月

目　录

第 1 章　蒙特卡罗检验

1.1　参数蒙特卡罗检验

对假设检验问题，在很多情况下，很难得到统计量在原假设下的精确分布或者极限分布，无法确定是否接受原假设的临界值点，此时可借助蒙特卡罗逼近的方法. 蒙特卡罗逼近是一种容易实施的方法，很多文献对它做了相关的研究. 文献 Bartlett(1963) 的讨论部分，首次描述了 MCT 的思想. Hope (1968) 证明在参数的情况下，如果没有讨厌参数，蒙特卡罗检验可能达到精确的显著性水平，即使与一致最优势 (UMP) 检验做比较，它的功效都很高. 在讨厌参数存在的情况下，MCT 也同样适用. 也就是，MCT 可应用在参数情况. 在空间模式研究中，Besag 和 Diggle (1977) 把 MCT 应用在随机变量分布中有讨厌参数的情况. 如果模拟可以基于原假设下最小充分统计量的观测值实现，Engen 和 Lillegård (1997) 用 MCT 逼近统计量的分布. 在具有讨厌参数的某些特定情况下，MCT 仍然可能达到精确的显著性水平. Zhu，Fang 和 Bhatti (1997) 构造投影追踪类型的 Crämer-von Mises 统计量检验参数族的分布. Hall 和 Titterington (1989) 说明在参数族的情况下，无论是否有讨厌参数，以及统计量渐近分布是否枢轴，由 MCT 逼近得到的误差要比由相应统计量的渐近分布带来的误差小；而且 MCT 可以区分以 $n^{-1/2}$ 的速度逼近原假设的备择假设. 这些结论进一步加强了 MCT 方法的理论依据.

举一个简单的例子解释如何用 MCT 方法. 考虑具有分布 $F(\cdot)$ 的独立同分布 (i.i.d.) 随机变量 $x_1,\cdots,x_n$，假设要检验 $F(\cdot)=G(\cdot,\theta)$ 是否成立，其中 θ 是未知参数，$G(\cdot)$ 为已知函数. 对这个检验问题的任何检验统计量，如 $T(x_1,\cdots,x_n)$，MCT 方法就是从分布 $G(\cdot,\hat{\theta})$ 中独立产生参考数据 $x_1',\cdots,x_n'$，计算相应统计量的值 $T(x_1',\cdots,x_n')$ 作为参考值，其中 $\hat{\theta}$ 是 θ 的估计. 如果 T 的值较大，拒绝原假设；对双边检验的情况不难做相应调整. 记 $T(x_1,\cdots,x_n)=T_0$，$T_1,\cdots,T_m$ 表示由蒙特卡罗得到的 m 个参考值. p 值的估计为

$$\hat{p}=k/(m+1),$$

其中，k 是 $T_0,T_1,\cdots,T_m$ 大于或者等于 T_0 的个数. 因此，给定水平 α，如果 $\hat{p}\leqslant\alpha$，拒绝原假设.

值得指出的是 20 世纪 80 年代发展的参数自助近似的具体步骤类似于上述的 MCT 步骤，具体可参考文献 Beran 和 Ducharme (1991).

1.2 非参数蒙特卡罗检验

1.2.1 方法论的动机

对于半参数或非参数的情况，在原假设下很难模拟参考数据计算统计量对应 MCT 的条件统计量. 主要困难在于即使在原假设下，模型不能用含有几个未知参数的具体模型刻画. 例如，检验现有的数据分布是否为椭球对称分布族 (简写为椭球分布). 如果对任何 $d \times d$ 正交矩阵 H ，存在形状矩阵 A 和位置向量 μ ，使 $HA(X-\mu)$ 和 $A(X-\mu)$ 的分布相同，称 d 维随机向量 X 服从椭球分布. 如果 X 二阶矩 Σ 有限， A 实际上就等于 $\Sigma^{-1/2}$ ，具体细节可参考文献 Fang， Kotz 和 Ng (1990). 从这个定义中，我们不难看出椭球分布不是参数族. *自助法* 是统计中非常重要的方法之一. Efron (1979) 首次提出这一方法，现在它已发展成解决上述问题的普遍适用的方法之一. Efron 自助法，也称为 *传统自助法*，它的基本思想是：从现有数据的经验分布中产生参考数据. 关于这种方法的研究很多，可参考 Davison 和 Hinkley (1997). Shao 和 Tu (1995) 对这个问题也做了全面的研究. 然而，用这个方法时必须注意几个问题. 第一，很难研究逼近的精确性或者渐近精确性，关于它的研究仍然停留在具体的某些问题中，并没有形成统一的方法，且相关的文献并不多，其中可参考文献 Singh (1981). 在一维变量的情况， Zhu 和 Fang (1994) 得到对应 Kolmogorov 统计量的自助统计量的精确分布，且证明它是 $\sqrt{n}$ 相合的. 就我们的知识而言，这是一篇唯一研究自助统计量准确分布的文章；第二，因为参考数据是从经验分布中产生，且经验分布收敛于数据的分布，自助逼近不能使统计量本身有效，可能渐近有效；第三，自助逼近有时不相合，对于这种不相合的修正也没有统一的方法. 从 n 个数据中产生 m 个数据是修正不相合的方法，但是在很多情况下，这种方法功效不好. 在回归分析中， Wu (1986) 提出减少方差估计偏差的新方法， Mammen (1992) 很好地发展了这种方法，并称之为 *Wild 自助法*，是一种重要的逼近方法. Wild 自助法已经成功地应用在许多不同的领域，特别是回归模型的检验，见 Härdle 和 Mammen (1993)， Stute, González Manteiga 和 Presedo Quindimil (1998). 在某些情况下，这种方法可以克服 Efron 的传统自助逼近法造成的不相合性，然而，并不是在所有情况下它都是相合的. 第 4 章对回归函数研究降维类型的检验中，给出一个例子说明 Wild 自助法的不相合性. 在第 6 章检验异方差性的问题中，也给出类似的例子；第四，在假设检验问题中，需慎重处理自助法产生的参考数据，否则可能降低检验的功效.

置换检验 是另一种产生参考数据的方法，见文献 Good (2000). 在有些情况下，它非常有效. 然而，如果只有一个数据，不能通过置换方法得到参考数据，在这种情况下这种方法的应用受到限制，且方法的实施也要花大量的计算时间.

自助法完全是非参数的统计方法论，它对模型结构以及数据分布的限制条件

试读结束，需要全本PDF请购买 www.ertongbook.com

很少. 因此, 如果模型并不是非参数的, 而是半参数结构, 如椭球对称分布, 我们可以用其他的蒙特卡罗逼近, 充分利用数据所提供的信息. 基于这些观测数据, 我们提出了非参数蒙特卡罗检验 (NMCT). 在第 2 章用 NMCT 方法检验四种类型的分布, 并且说明此方法对这些类型的检验精确有效. 如果第 2 章所研究的分布中含有讨厌参数, 在第 3 章证明 NMCT 方法对这种情况渐近有效的, 而且 $\sqrt{n}$ 相合, 然而根据自助逼近法不能得出这样的结论. 第 4~6 章研究了回归模型的模型检验问题, 第 4 章和第 6 章也说明了 Wild 自助法在某些情况下不相合. 第 7~9 章研究了一些用自助逼近法可以实现, NMCT 方法也很容易实现的问题, 而且功效很好. 在下面的两个小节, 分别给出了随机变量独立可分解时, 以及检验统计量可以渐近表示为线性统计量的函数时, NMCT 的具体实现过程.

1.2.2 基于可独立分解随机变量的 NMCT 方法

NMCT 最初的动机来自检验几类重要的多元分布, 现在已经发展成一般的方法论. 关于检验多元分布的具体细节见第 2 章.

我们经常用 4 种类型的多元分布: 椭球对称、反射对称、 Liouville-Dirichlet 和对称刻度混合分布. 关于这 4 种类型分布的定义, 可参见第 2 章和第 3 章. 这些分布族分别是正态、对称、 Beta 和平稳分布的推广, 见文献 Fang, Kotz 和 Ng (1990) 以及此文的参考文献.

关于椭球对称和反射对称分布的检验问题已经有一些研究. 例如, Aki(1993) , Baringhaus(1991), Beran(1979), Ghosh 和 Ruymgaart (1992), Heathcote, Rachev 和 Cheng(1995). 由于这些分布族, 如椭球对称分布, 不能用有限的参数完全刻画, 因此不能简单的用第 1.1 节中所提到的关于参数的 MCT 逼近统计量在原假设下的分布. 在假设检验问题中, 统计量在原假设下的极限分布通常很难确定临界值点, 可参见 Zhu, Fang, Bhatti 和 Bentler (1995). Diks 和 Tong (1999) 提出了 条件蒙特卡罗检验, 其中的思想是: 如果密度函数在等距紧集 G 下不变, G 轨道集是最小充分统计量, 在给定 G 轨道观测值的条件下模拟分析. 他们对不含讨厌参数的球和反射对称的多元分布做检验. Neuhaus 和 Zhu(1998), Zhu 和 Neuhaus (2003) 也对这两种类型的多元对称分布构造了条件检验过程.

接下来说明如何产生参考数据, 此方法依赖于分布的可独立分解性.

定义 1.2.1 *随机向量 X 称为独立可分解, 如果 $X = Y \cdot Z$ 依分布成立, 这里, Y 和 Z 独立, $Y \cdot Z$ 表示 Y 和 Z 点乘, 也就是: 如果 Y 和 Z 是 d 维向量, $Y \cdot Z = (Y^{(1)}Z^{(1)}, \cdots, Y^{(d)}Z^{(d)})$; 如果 Z 是一维的, $Y \cdot Z = (Y^{(1)}Z, \cdots, Y^{(d)}Z)$; 如果 Y 是一维的, $Y \cdot Z = (YZ^{(1)}, \cdots, YZ^{(d)})$.*

如果已知 Y 或 Z 的分布, 上面的可分解性是 MCT 步骤可实施的根据. 记 $x_1, \cdots, x_n$ 表示样本 n 的 i.i.d. 随机变量, 如果 x_i 在原假设下可独立分解为 $x_i = y_i \cdot z_i$, 则检验统计量 $T(x_1, \cdots, x_n)$ 等于 $T(y_1 \cdot z_1, \cdots, y_n \cdot z_n)$. NMCT 方法为: 给

定 $z_1,\cdots,z_n$，从 Y 的分布中独立产生一组参考数据 $y_1',\cdots,y_n'$，则可得相应统计量的值 $T(y_1'\cdot z_1,\cdots,y_n'\cdot z_n)$. 假设如果 T 值较大，原假设被拒绝；对双边检验问题不难做相应的调整. 记由原始数据得到的 T 为 T_0，通过蒙特卡罗产生 m 组参考数据，相应的得到 m 个值，分别记为 $T_1,\cdots,T_m$. 统计量 T 的 p 值的估计为

$$\hat{p}=k/(m+1),$$

其中，k 是 $T_0,T_1,\cdots,T_m$ 中大于等于 T_0 的个数. 给定水平 α，只要 $\hat{p}\leqslant\alpha$，拒绝原假设.

由于 $T(x_1,\cdots,x_n)$ 和 $T(y_1'\cdot z_1,\cdots,y_n'\cdot z_n)$ 同分布，而且给定 $z_1,\cdots,z_n$，它们有相同的条件分布，检验的可能精确有效. 下面的命题说明这个性质.

命题 1.2.1 *在原假设下，向量 X 可独立分解为 $Y\cdot Z$，那么，对任何 $0<\alpha<1$，*

$$\Pr(\hat{p}\leqslant\alpha)\leqslant\frac{[\alpha(m+1)]}{m+1},$$

其中，$[c]$ 表示 c 的整数部分.

证 在原假设下，给定 $z_1,\cdots,z_n$，$T_0,T_1,\cdots,T_m$ 条件独立同分布，如果 T_i 之间没有结，$\hat{p}$ 在 $\left\{\frac{1}{m+1},\cdots,\frac{m+1}{m+1}\right\}$ 均匀分布. 由于 $\hat{p}\leqslant\alpha$ 隐含 $k\leqslant[\alpha(m+1)]$，因此

$$\Pr(\hat{p}\leqslant\alpha\,|\,z_1,\cdots,z_n)=\frac{[\alpha(m+1)]}{m+1}.$$

如果 T_i 之间有结，且 $k\leqslant[\alpha(m+1)]$. 那么 T_0 至少比 T_i 中第 $m+1-[\alpha(m+1)]$ 个大的元素大. 因此，

$$\Pr(\hat{p}\leqslant\alpha\,|\,z_1,\cdots,z_n)\leqslant\frac{[\alpha(m+1)]}{m+1}.$$

对 z_i 求积分，证毕.

这个命题说明在变量可独立分解时，NMCT 方法可以精确有效. 相比较而言，自助法和置换检验并没有这个优点. 在第 2 章对上述所提到的四种类型的分布用蒙特卡罗逼近做检验.

1.2.3 基于随机加权的 NMCT 方法

如果假定的数据分布不具有独立分解的性质，本节建议用随机加权的方法产生参考数据. 这个方法可实施的根据是经验过程理论：随机加权经验过程的收敛性.

假定 $x_1,\cdots,x_n$ 表示 i.i.d. 的样本，如果检验统计量，如 $T_n=T(x_1,\cdots,x_n)$，可以重新写为 $\mathcal{T}\circ R_n$. 其中，$\mathcal{T}$ 是 R_n 的函数，R_n 是具有下述形式的过程：

$$R_n=\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}J(x_j,t),\,t\in S\right\}.$$

其中，子集 $S \subset R^d$；如果 S 为单点，R_n 表示随机变量；$E(J(X,t)) = 0$. $T_n(E_n) = \mathcal{T} \circ R_n(E_n)$ 表示对应于 T_n 的条件表达式，用这个条件表达式逼近 T_n 的分布，其中，

$$R_n(E_n) = \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} e_j J(x_j, t),\ t \in S \right\},$$

$E_n = \{e_1, \cdots, e_n\}$ 为独立于 x_j 的随机变量. 如果 e_j 以相同的概率取值 ± 1，这类随机加权的方法称为随机对称加权 (Pollard, 1984)；如果 e_j 是均值 0，方差 1 的正态分布，见文献 Dudley (1978)，Giné 和 Zinn (1984)，这种加权方法类似于 Wild 自助法 (Mammen, 1992). 假如 $x_1, \cdots, x_n$ 为可交换的随机变量，Van der Vaart 和 Wellner (2000) 称之为可交换自助法.

然而，大部分的检验统计量很难具有这样的表达式. 在很多情况下，$T_n(x_1, \cdots, x_n)$ 有下述渐近表达式：

$$T_n(x_1, \cdots, x_n, P_n) = \mathcal{T} \circ R_n + o_p(1), \tag{1.2.1}$$

其中，R_n 的表达式为 $n^{-1/2} \sum_{j=1}^{n} J(x_j, \psi, t)$，$E(J(X, \psi, t)) = 0$，$\psi$ 是感兴趣的未知参数，它可以是无限维的，如未知的光滑函数. 以下给出估计 p 值的一般步骤：

步骤 1 产生均值 0，方差 1 的独立随机变量 $e_j (j = 1, \cdots, n)$ 记 $E_n := (e_1, \cdots, e_n)$ 以及 R_n 的条件对应表达式

$$R_n(E_n, t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} e_j J(x_j, \hat{\psi}, t), \tag{1.2.2}$$

其中，$\hat{\psi}$ 是根据数据 $x_1, \cdots, x_n$ 得到 ψ 的相合估计. 对应的条件检验统计量为

$$T_n(E_n) = \mathcal{T} \circ R_n(E_n). \tag{1.2.3}$$

步骤 2 产生 m 组 E_n，记为 $E_n^{(i)} (i = 1, \cdots, m)$，相应得到 m 个 $T_n(E_n)$ 值，分别记为 $T_n(E_n^{(i)}), i = 1, \cdots, m$.

步骤 3 如果 T_n 的值较大，拒绝原假设. 对于双边检验问题，不难做出相应的调整. p 值的估计为 $\hat{p} = k/(m+1)$，其中，k 表示 $T_n(E_n^{(i)})$ 大于或者等于 T_n 的个数. 给定水平 α，如果 $\hat{p} \leqslant \alpha$，拒绝原假设.

命题 1.2.2 假定 e_i 是 i.i.d. 且具有紧支撑的变量，R_n 依分布收敛到连续的 Gaussian 过程 R，且存在 $a > 0$ 满足 $\hat{\psi} - \psi = O_p(n^{-a})$. 进一步假设对任何固定 $t \in S$，函数 J 关于 ψ 的两阶偏导数存在，且所有的偏导数关于 t 一致具有有限的一阶矩，则对于几乎所有序列 $(x_1, \cdots, x_n)$，$T_n(E_n)$ 和 T_n 的极限相同.

证 根据已知条件，可得

$$R_n(E_n,t) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} e_j J(z_j,\psi,t)+o_p(1). \tag{1.2.4}$$

也即是，$R_n(E_n,\cdot)$ 为经验过程. 根据文章 Van der Vaart 和 Wellner (2000) 中的定理 3.6.13，$R_n(E_n)$ 和 R_n 极限相同. 根据 T 的连续性，结论成立.

注释 1.2.1 下面举例说明检验统计量可以渐近地表示为式 (1.2.1) 关于线性统计量的函数. 考虑回归模型

$$Y=\Phi(X)+\varepsilon,$$

其中，$\Phi(\cdot)$ 是未知函数，Y 为 1 维响应随机变量，X 是独立 ε 的 p 维列随机向量. 假设检验的问题为

$$H_0: \quad \Phi(\cdot)\in\{\Phi_0(\cdot,\theta):\theta\in\Theta\},$$

其中，Φ_0 为给定的函数，Θ 为 q 维 Euclidean 空间 R^q 上的紧集. 因此，在原假设下，存在列向量 θ_0 满足 $\Phi(\cdot)=\Phi_0(\cdot,\theta_0)$. 通常使用的检验方法是基于残差构造统计量. 直观上说，如果残差比较大，则检验统计量的值可能较大，拒绝原假设. 假定 $(x_1,y_1),\cdots,(x_n,y_n)$ 为 i.i.d. 样本，根据注释 1.2.1 的想法，构造如下统计量：

得分类型的检验 $\hat{\varepsilon}_j=y_j-\hat{\Phi}_0(x_j,\hat{\theta}_0)(j=1,\cdots,n)$ 表示通过拟合回归函数 $\Phi_0(\cdot,\theta_0)$ 得到的残差，其中 $\hat{\theta}_0$ 为 θ_0 的相合估计. 得分类型的检验定义为

$$T_n=\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\hat{\varepsilon}_j w(x_j,\hat{\theta}_0)\right]^2$$

其中，$w(\cdot)$ 为待选择的权重函数. 记 $R_n=n^{-1/2}\sum\limits_{j=1}^{n}\hat{\varepsilon}_j w(x_j,\hat{\theta}_0)$. 如果 $\hat{\theta}_0$ 根据最小二乘法得到，不难证明在原假设下，假定某些正则条件成立，$\hat{\theta}_0-\theta_0$ 的渐近线性表示为 $\hat{\theta}_0-\theta_0=n^{-1}\sum\limits_{j=1}^{n}J_1(x_j,y_j,E(\Phi_0')^2,\theta_0)+o_p(1/\sqrt{n})$，

其中，

$$J_1(x_j,y_j,E(\Phi_0')^2,\theta_0)=:[E(\Phi_0'(X,\theta_0))(\Phi_0'(X,\theta_0))^\tau]^{-1}\Phi_0'(x_j,\theta_0)\varepsilon_j,$$

且 Φ_0' 为 Φ_0 关于 θ 的一阶导数. 显然有 $E(J_1(X,Y,E(\Phi_0')^2,\theta_0))=0$. 记

$$\begin{aligned} & J(x_j,y_j,E(\Phi_0')^2,E(\Phi_0'w),\theta_0)\\ = \;& \varepsilon_j w(x_j,\theta_0)-(J_1(x_j,y_j,E(\Phi_0')^2,\theta_0))^\tau E[\Phi_0'(X,\theta_0)w(X,\theta_0)]. \end{aligned}$$

简单推导可得 $R_n(t)$ 的渐近表达式为

$$R_n(t) = 1/\sqrt{n}\sum_{j=1}^{n} J(x_j, y_j, E(\Phi_0')^2, E(\Phi_0' w), \theta_0).$$

式 (1.2.1) 中的 $\mathcal{T}$ 在这里为平方函数.

Crämer-von Mises 类型和 Kolmogorov 类型的检验 记 $R_n(x) = n^{-1/2}\sum_{j=1}^{n} \hat{\varepsilon}_j w(x_j, \theta_0) I(x_j \leqslant x)$，其中，“$X \leqslant x$” 表示 X 的每个分量小于或等于对应于 x 的分量. 由上述关于得分类型的检验可知，R_n 的渐近表达式为 $n^{-1/2}\sum_{j=1}^{n} J(x_j, y_j, E(\Phi_0')^2, E(\Phi_0' w), \theta_0, w, x)$. Crämer–von Mises 类型的检验统计量为 $T_n = \int [R_n(X)]^2 \mathrm{d}\, F(X)$，Kolmogonov 类型的检验统计量为 $\sup_{t,x} |R_n(x)|$. $\mathcal{T}$ 在这里分别为积分和上确界函数.

需要指出的是，这里所提出的算法类似于 Wild 自助法 (如 Härdle 和 Mammen (1993)， Stute, González Manteiga 和 Presedo Quindimil (1998)). 算法的不同之处在于： Wild 自助法是产生样本 (X_i^*, Y_i^*)；而 NMCT 方法只用替换 R_n 中的 e_i. 可以证明，在对线性模型做拟合优度检验时，也就是说， $\Phi_0(x, \theta_0) = \theta_0^{\tau} x$，如果采用上述检验法， Wild 自助法 与 NMCT 等价. 但是对其他的模型做拟合优度检验，这个等价性未必成立. 第 4 章给出更详细的研究. 对于更一般的模型，如果用 Crämer–von Mises 检验做统计量， NMCT 和 Wild 自助法的等价性不存在，我们将在第 5 章做讨论.

第 2 章　多元分布的检验

本章研究多元分布的检验问题. 虽然在多元分析中, 对多元正态分布的检验仍然是研究的问题之一, 但是越来越多的工作开始致力于非参数情形的研究. 在多元分布中, 有一些重要的分布族. 本章考虑四类不同的分布族的检验问题, 本章的内容大部分来自文献 Zhu 和 Neuhaus (2000).

为了用第 1 章提到的 NMCT 方法检验多元分布族, 需要分析这些分布族是否具有独立可分解性 . X 表示 d 维随机变量, $X = Y \cdot Z$ 表示 X 和 $Y \cdot Z$ 的分布相同, 只研究 Y 和 Z 是否独立, 且 Y 的分布已知的情况.

2.1　四种类型的多元分布

情形 (a) 椭球对称分布

对于这类分布族, $X = U \cdot \|X\|$ 依分布成立. 其中, U 和 $\|X\|$ 独立, 且 U 是球 $S^d = \{a : \|a\| = 1, a \in R^d\}$ 上的均匀分布, $\|\cdot\|$ 表示 R^d 上的 Euclidean 范数. 不难看出 $Y = U$, $Z = \|X\|$. 多元 t 分布 (Fang, Kotz 和 Ng, 1990, 例 2.5) 和正态分布 $N(0, I_d)$ 属于这类分布族. 实际上我们取 $U = X/\|X\|$.

情形 (b) 反射对称分布族

对于这类分布族, 依分布有 $X = -X$. 由于依分布 $X = e \cdot X$ 成立, 其中 $e = \pm 1$ 的概率相同, X 独立可分解. 所以, $Y = e$, $Z = X$. $[-c, c]^d (c > 0)$ 上的均匀分布属于这类分布族.

情形 (c)Liouville-Dirichlet 分布族

对于这类分类族, 依分布有 $X = Y \cdot r$, 其中 Y 是独立于刻度变量 r 的 Dirichlet 分布 $D(\alpha)$, 参数 $\alpha = (\alpha_1, \cdots, \alpha_d)$ 已知; Y 的分量 $y^{(1)}, \cdots, y^{(d)}$ 满足 $B^d = \{(y^{(1)}, \cdots, y^{(d)}) \in R^d : y^{(i)} \geqslant 0, \sum_{i=1}^{d} y^{(i)} = 1\}$. 对于这类分布族, $Y = X/(\sum_{i=1}^{d} x^{(i)})$ 和 $Z = \sum_{i=1}^{d} x^{(i)}$, 其中 $X = (x^{(1)}, \cdots, x^{(d)})$. 这类分布族包括多元 Beta 和逆 Dirichlet 分布 (文献 Olkin, Rubin(1964); Guttman, Tiao(1965)).

情形 (d) 对称刻度混合分布

关于这类分布族, 对 $x \neq 0$, 存在刻度函数 $g(x)$ 满足 $g(x) = g(-x)$ 且 $g(x) \neq 0$. $x/g(x)$ 与 $g(x)$ 独立, 且 $x/g(x)$ 为空间 $C^d = \{y = (y^{(1)}, \cdots, y^{(d)}) \in R^d : g(y) = 1\}$ 上的均匀分布. 所以, 可以取 $Y = X/g(X)$ 以及 $Z = g(X)$. 这类分布族较大, 包含所有的球对称分布. 具有密度函数 $c\exp(-\sum_{i=1}^{d} |x^{(i)}|)$ 的 Laplace 分布也属于这类分布.

试读结束，需要全本PDF请购买 www.ertongbook.com