多元分布拟合优度检验基础

● 苏岩 著

科学出版社

多元分布拟合优度检验基础

苏 岩 著

科学出版社
北京

内 容 简 介

本书系统阐述多元分布拟合优度检验基础知识与理论进展. 全书共 11 章: 第 1 章为拟合优度检验概述; 第 2 章介绍矩阵代数, 用于构造球面均匀分布的特征检验; 第 3 章介绍概率极限理论; 第 4 章讨论垂直密度表示; 第 5 章介绍球面调和函数, 用于构造球面均匀分布的光滑检验; 第 6 章介绍球面与球体概率分布, 用于多元分布检验功效模拟; 第 7 章讨论概率密度核估计, 用于中心相似分布的拟合优度检验; 第 8 章介绍一元分布的拟合优度检验; 第 9 章介绍球面均匀分布的拟合优度检验; 第 10 章介绍实心区域均匀分布的拟合优度检验; 第 11 章介绍椭球对称分布的拟合优度检验.

本书可作为高等院校数学、统计学和计量经济学专业的高年级大学生、研究生教材, 亦可作为数学、经济、金融、生物医学、工业工程和模式识别等领域的教师、统计应用工作者的参考书.

图书在版编目(CIP)数据

多元分布拟合优度检验基础/苏岩著. —北京: 科学出版社, 2016.3
ISBN 978-7-03-047562-6

I.①多… II.①苏… III.① 多元分布–拟合优度检验 IV. ①O212.1

中国版本图书馆 CIP 数据核字(2016) 第 046722 号

责任编辑: 陈玉琢 / 责任校对: 钟 洋
责任印制: 张 倩 / 封面设计: 陈 静

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

新科印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2016 年 3 月第 一 版 开本: 720×1000 1/16
2016 年 3 月第一次印刷 印张: 18 1/4
字数: 380 000
定价: 108.00 元
(如有印装质量问题, 我社负责调换)

前　言

多元正态分布是经典多元分析的基本假设, 椭球对称分布是广义多元分析的基本假设. 对于多元正态分布, 椭球对称分布的拟合优度检验可转化为对球面均匀分布的拟合优度检验. 因此, 球面均匀分布的拟合优度检验显得尤为重要. 椭球对称分布的拟合优度检验, 为广义多元分析与多元时间序列模型在实际中的应用打下基础. 全书共 11 章.

第 1 章概述拟合优度检验的理论与方法, 介绍基于球的多元分布拟合优度检验研究主线. 第 2 章包括广义逆矩阵、矩阵分解及矩阵微商等内容, 用于球面均匀分布的特征检验. 第 3 章讨论随机向量的收敛模式与概率极限理论. 第 4 章介绍垂直密度表示及随机数生成. 第 5 章介绍球面微积分、球调和函数概念与性质, 用以构造球面均匀分布光滑检验统计量. 第 6 章介绍球面旋转对称分布, 基于正交球调和函数的球面 Beran 分布族、中心相似分布等的概念与性质. 本章内容用于多元分布拟合优度检验功效的随机模拟. 第 7 章讨论一元概率密度核估计、多元概率密度核估计、基于球极投影变换密度核估计和球面密度核估计. 本章内容可用于边界未知中心相似分布的拟合优度检验. 第 8 章介绍一元分布光滑检验, 基于经验分布函数的 EDF 型检验, 基于间隔的 (0,1) 均匀分布 Gini 检验及功效模拟. 本章内容将用于多元分布的拟合优度检验. 第 9 章介绍基于广义逆的球面均匀分布特征检验与基于球调和函数的球面均匀分布光滑检验. 功效模拟显示这两种球面均匀性检验具有一定的互补性. 球心在原点的单位球面具有完美的几何对称性. 球面均匀分布的一阶矩、二阶矩分别对应着它的基本特征: 质心与惯量矩这两个球面均匀性物理特性. 基于 1 阶和 2 阶球调和函数的球面均匀性光滑检验对应着其质心和惯量矩联合检验. 第 10 章介绍边界已知实心区域均匀分布的拟合优度检验及其特殊情形: 单位球均匀分布与单纯形均匀分布的拟合优度检验. 第 11 章介绍椭球对称分布与多元正态分布的特征检验、光滑检验、多元线性模型误差分布的拟合优度检验.

杨振海教授对作者关于多元分布拟合优度检验研究给予了热情鼓励和指导, 作者的研究合作者对本书给出了非常有益的建议, 在此作者向他们表示深深的谢意.

由于作者水平所限, 书中不妥之处在所难免, 恳请广大读者批评指正.

苏　岩

2015 年 12 月

目　录

第1章 绪 论

总体分布是构成统计模型的基本要素, 统计推断离不开对总体分布的假设. 因此, 概率分布的构造和拟合优度检验在统计理论和应用中有着特殊地位, 它们的理论和方法始终受到人们的重视. 在统计理论及应用研究中, 正态分布居主导地位, 中心极限定理保证在一定条件下, 随机变量和具有渐近正态性. 多元数据的处理, 主要是基于多元正态分布进行统计分析的, 时间序列分析中误差分布一般假设为正态分布, 信息论中高斯分布是其基本假设.

随着研究的深入及精确推断的要求, 人们发现在一些应用问题中, 概率分布为正态分布的假设是不成立的. 例如, 收益序列通常呈现的 "尖峰厚尾态" 不宜假设为正态分布, 备选的概率分布有 t 分布及 GED 密度函数等. 制定中国男子服装 12 项指标构成的随机向量不满足多元正态分布, 只是 12 项指标的若干子集构成的随机向量服从多元正态分布. 这些随机变量或向量按正态分布假设进行统计分析会出现明显的误差. 总体分布的正态性拟合优度检验, 总体分布的非正态性条件下, 备选分布的构造与拟合优度检验在应用中具有重要意义.

1.1 一元概率分布

概率分布是用来描述随机现象的基本工具, 任何统计方法都离不开概率分布的概念和各种具体分布的性质, Johnson 等在 20 世纪 70 年代编著了 *Distribution in Statistics* 一书, 共 4 册. 方开泰等 (1987) 编著了《统计分布》一书, 详细介绍了一维随机变量的对数正态分布、χ^2 分布、Gamma 分布、Beta 分布、柯西分布、Logistic 分布、极值分布、Laplace 分布. 对数正态分布出现在许多领域之中, 如针刺麻醉的镇痛效果, 流行病蔓延时间的长短, 某些电器寿命等. Logistic 分布最初用于描述生长曲线, 现也广泛应用于经济社会统计中. χ^2 分布是由正态分布派生出来的分布, 在统计检验中有广泛的应用, 例如, 时间序列模型白噪声的检验可由渐近 χ^2 分布来进行. Gamma 分布和 Beta 分布应用于可靠性统计及先验分布. 极值分布用于描述洪水等灾害性自然现象, Laplace 分布可用于稳健统计分析.

1.2 多元概率分布

为了解决多元数据的总体概率分布问题, 统计学者提出了各种概率分布的构造

方法. Fang 等 (1990a) 著述的 *Symmetric Multivariate and Related Distributions*, 研究了球对称分布、椭球对称分布、多元 L_1 模对称分布、多元 Liouville 分布、α 对称分布等的构造, 详细讨论了各种分布的性质, 例如, Kotz 型分布、Pearson II 型分布、多元柯西分布和多元 t 分布等的特征函数, 矩、边际分布和条件分布等. 椭球对称分布是多元正态分布的自然推广, 这些分布族成为多元正态分布的备选分布, 相对一般意义下的多元分析即为广义多元分析, 可以通过密度定义、特征函数、随机表示、正交变换、垂直密度表示等方法构造多元分布.

1. 密度定义

设 X 服从 $N(\mu,\sigma^2)$, 其概率密度为

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right\},$$

由此猜想多元正态分布概率密度为

$$c|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

$\boldsymbol{x},\boldsymbol{\mu}\in\mathbf{R}^d$, Σ 为 $d\times d$ 矩阵, $\Sigma>0$, c 为归一常数.

多元正态分布概率密度表达式中, $(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ 等于常数时, 其密度值相等. 一个自然的推广是, 构造多元概率密度为 $|\Sigma|^{-1/2}g((\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$, 由此得到椭球分布概率密度.

2. 特征函数

设 X 服从 $N(\mu,\sigma^2)$, 则 X 的特征函数为

$$\exp\left\{\mathrm{i}t\mu-\frac{1}{2}\sigma^2t^2\right\}=\exp\left\{\mathrm{i}t\mu-\frac{1}{2}t(\sigma^2)t\right\}.$$

因此多元正态分布的特征函数应为

$$\exp\left\{\mathrm{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right\},$$

其中 $\boldsymbol{t},\boldsymbol{\mu}\in\mathbf{R}^d$ 且 $\Sigma>0$.

3. 随机表示

设 $X\sim N(\mu,\sigma^2)$, 则有随机表示

$$X\overset{d}{=}\mu+\sigma Y,$$

其中 $Y\sim N(0,1)$. 设 $Y_1,\cdots,Y_m$ i.i.d. $N(0,1)$, $\boldsymbol{Y}=(Y_1,\cdots,Y_m)^{\mathrm{T}}$ 且

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + A\boldsymbol{Y},$$

其中 $\boldsymbol{X}, \boldsymbol{\mu} \in \mathbf{R}^d$, A 为 $d \times m$ 矩阵, 则 $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$, $\Sigma = AA^{\mathrm{T}}$.

由随机向量的随机表示, 可以得到一些主要的多元概率分布.

(1) d 维随机向量 $\boldsymbol{Y}$ 称为服从球对称分布, 当且仅当

$$\boldsymbol{Y} \overset{d}{=} R\boldsymbol{U}_d,$$

其中 R 是非负随机变量, $\boldsymbol{U}_d$ 是 d 维单位球面上的均匀分布随机向量, R 与 $\boldsymbol{U}_d$ 独立. 称 $\boldsymbol{U}_d$ 为球对称分布的 L_2 模球面均匀分布基.

(2) 设 $\boldsymbol{Y}$ 服从 m 维球对称分布. d 维随机向量 $\boldsymbol{X}$ 称为服从参数为 $\boldsymbol{\mu}, \Sigma$ 的椭球对称分布, 当且仅当

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + A\boldsymbol{Y},$$

其中 A 为 $d \times m$ 阶矩阵, $AA^{\mathrm{T}} = \Sigma$ 且 $\mathrm{rank}(\Sigma) = m$.

(3) $\mathbf{R}_+^d$ 上的 d 维随机向量 $\boldsymbol{X}$ 称为服从多元 L_1 模对称分布, 当且仅当

$$\boldsymbol{X} \overset{d}{=} R\boldsymbol{U}_d,$$

其中 $\mathbf{R}_+^d = \{(x_1, \cdots, x_d)^{\mathrm{T}} : x_i \geqslant 0,\ i = 1, \cdots, d\}$,

$$\boldsymbol{B}_d = \left\{ \boldsymbol{y} = (y_1, \cdots, y_d)^{\mathrm{T}} : \sum_{i=1}^{d} y_i = 1, \boldsymbol{y} \in \mathbf{R}_+^d \right\}.$$

$\boldsymbol{U}_d$ 服从 $\boldsymbol{B}_d$ 上的均匀分布, $R \geqslant 0$ 与 $\boldsymbol{U}_d$ 独立. 称 $\boldsymbol{U}_d$ 为多元 L_1 模对称分布的 L_1 模球面均匀分布基.

(4) $\mathbf{R}_+^d$ 上的 d 维随机向量 $\boldsymbol{X}$ 称为服从多元 Liouville 分布, 当且仅当

$$\boldsymbol{X} \overset{d}{=} R\boldsymbol{Y},$$

R 与 $\boldsymbol{Y}$ 独立, $\boldsymbol{Y}$ 服从 $\boldsymbol{B}_d$ 上的 Dirichlet 分布, 即 $\boldsymbol{Y} \sim D_d(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)^{\mathrm{T}}$. 称 $\boldsymbol{Y}$ 为多元 Liouville 分布的 Dirichlet 分布基.

当 $R = 1$ 时, 多元 Liouville 分布成为 Dirichlet 分布. 当 $\boldsymbol{Y}$ 服从 $\boldsymbol{B}_d$ 上的均匀分布即参数分量均为 1 的 Dirichlet 分布 $D_d(1, \cdots, 1)$ 时, 多元 L_1 模对称分布成为多元 Liouville 分布的特例.

4. 正交变换

设 $O(d)$ 表示 $d \times d$ 正交矩阵群. d 维随机向量 $\boldsymbol{X}$ 称为服从球对称分布, 当且仅当对任意 $\Gamma \in O(d)$ 有 $\boldsymbol{X} \overset{d}{=} \Gamma\boldsymbol{X}$.

5. 垂直密度表示

Troutt(1991) 首先提出了垂直密度表示 (vertical density repyesentation, VDR), 给出了 I-型垂直密度表示 (Type I VDR). 设随机向量 $\boldsymbol{X}_d$ 具有概率密度 $f(\boldsymbol{x}_d)$, 记 $V = f(\boldsymbol{X}_d)$. $D_{(d,[f])}(x) = \{\boldsymbol{x}_d \in \mathbf{R}^d,\ f(\boldsymbol{x}_d) \geqslant x\}, L_d(A)$ 为 A 的 Lebesgue 测度, $A \in \mathbf{R}^d$. 若 $L_d(D_{(d,[f])}(v))$ 可微, 则 V 的概率密度为

$$g(v) = -v\frac{\mathrm{d}}{\mathrm{d}v}L_d(D_{(d,[f])}(v)), \tag{1.2.1}$$

且概率密度 $f(\boldsymbol{x}_d)$ 可表示为

$$f(\boldsymbol{x}_d) = \int_0^{f_0} h_v(\boldsymbol{x}_d|v)g(v)\mathrm{d}v, \tag{1.2.2}$$

$h_v(\boldsymbol{x}_d|v)$ 是给定 $V = v$ 条件下 $\boldsymbol{X}_d$ 的条件概率密度, $f_0 = \sup\{f(\boldsymbol{x}_d) :\ \boldsymbol{x}_d \in \mathbf{R}^d\}$. 称 $g(v)$ 为垂直密度, 称式 (1.2.1) 为 Type I VDR. 对有限值 f_0, 定义 $W = V/f_0$, W 的概率密度记为 $p_W(w)$. 称 $p_W(w)$ 为规范化垂直密度.

Troutt 没有给出 $h_v(\cdot|v)$ 的表示, Fang 等 (2001) 提出了 II-型垂直密度表示 (Type II VDR), 给出了 $h_v(\cdot|v)$ 的表达式.

Yang 等 (2003) 基于垂直密度表示, 提出了中心相似分布: $\boldsymbol{X}_d = R\boldsymbol{U}_d$. 此时假定 $\boldsymbol{U}_d$ 在基本集 D_0 服从均匀分布, R 是非负随机变量且与 $\boldsymbol{U}_d$ 独立. 做为例子, 得到零均值的多元正态分布是中心相似分布的特例. Yang 等给出了中心相似分布另一种表达等价形式:

$$\boldsymbol{X} = R\boldsymbol{U}_d = R^*\boldsymbol{W}_d, \quad R^* = R\|\boldsymbol{U}_d\|, \tag{1.2.3}$$

得到 $\boldsymbol{W}_d = \boldsymbol{U}_d/\|\boldsymbol{U}_d\|$ 的概率密度公式, 此时单位球面上的随机向量 $\boldsymbol{W}_d$ 不再服从球面上的均匀分布. 中心相似分布也可由球对称分布直接推广得出. 替换球对称分布定义中的 L_2 模球面均匀分布基 $\boldsymbol{U}_d$ 为 L_2 模球面非均匀分布基 $\boldsymbol{W}_d$, 则得到中心相似分布. 中心相似分布的构造与表示体现了 d 维欧氏空间与球面之间的几何关系. 一般形式下的中心相似分布为椭球对称分布的推广形式.

1.3　拟合优度检验概述

拟合优度检验在统计理论研究和实际应用中有着重要作用, 做具体数据统计分析面临的重要任务是确定数据服从何种概率分布. 设 $X_1, \cdots, X_n$ 是来自总体分布为 F 的样本, 拟合优度检验就是如何检验假设

$$H_0 :\ F \in \boldsymbol{P}_0; \quad H_1 :\ F \overline{\in} \boldsymbol{P}_0, \tag{1.3.1}$$

对立假设也可取为 $H_1:\ F\in\boldsymbol{P}_1,\ \boldsymbol{P}_0\cap\boldsymbol{P}_1=\varnothing$. $\boldsymbol{P}_0$, $\boldsymbol{P}_1$ 是具有特定性质的概率分布组成的分布族. 典型的做法是构造检验统计量, 求出检验统计量的精确分布或渐近分布. 对给定显著水平 α, 作出接受或拒绝原假设 H_0 的结论.

1.3.1 一元概率分布的拟合优度检验

杨振海等 (2011) 著述的《拟合优度检验》, 结合实际统计分析经验, 系统研究了一元概率分布拟合优度检验的各种方法: Q-Q 图形方法、χ^2 型检验、光滑检验、基于经验分布的 EDF 型检验、拟合优度检验中的变换方法和常见分布的拟合优度检验等, 讨论了局部对立假设序列下检验统计量的渐近分布, 针对具体分布, 分析拟合优度检验的模拟功效.

1. Pearson χ^2 检验

Pearson(1900) 提出了著名的 Pearson χ^2 检验统计量. 设 $X_1,\cdots,X_n$ 为来自 X 的样本, 总体分布为 F, 要检验假设

$$H_0:\ F=F_0.$$

设 X 的样本空间可分为 m 个两两互斥的集 $B_1,\cdots,B_m$ 的并, 且 $P(X\in B_i)=p_i$, $n_i=\#\{X_j\in B_i,\ j=1,\cdots,n\}$, $i=1,\cdots,m, n=\sum_{i=1}^m n_i$.

当 F_0 完全已知时, Pearson χ^2 检验统计量

$$\chi_P^2=\sum_{i=1}^m(n_i-np_i)^2\Big/np_i$$

的渐近分布为 χ^2_{m-1}.

当 F_0 不完全已知, 含有 $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_q)^{\mathrm{T}}$, q 个未知参数时, 由似然方程组

$$\sum_{i=1}^m\frac{n_i}{p_i(\boldsymbol{\theta})}\frac{\partial p_i(\boldsymbol{\theta})}{\partial\theta_j}=0,\quad j=1,\cdots,q \tag{1.3.2}$$

得到 p_i 基于分组样本的 MLE 估计 $\widehat{p}_i$, $i=1,\cdots,m$, 且有

$$\chi^2_{\mathrm{PF}}=\sum_{i=1}^m(n_i-n\widehat{p}_i)^2\Big/n\widehat{p}_i$$

的渐近分布为 χ^2_{m-q-1}.

设 $f(\cdot,\boldsymbol{\theta})$ 是 $F(\cdot,\boldsymbol{\theta})$ 的密度函数, $\widehat{\boldsymbol{\theta}}$ 是 $\boldsymbol{\theta}$ 的 MLE 估计, 即 $\boldsymbol{\theta}$ 满足方程

$$\sum_{i=1}^n\frac{\partial\log f(X_i,\boldsymbol{\theta})}{\partial\theta_j}=0,\quad j=1,\cdots,q. \tag{1.3.3}$$

记 $p_i(\boldsymbol{\theta})$ 的估计为 $p_i(\widehat{\boldsymbol{\theta}})$, 称

$$\chi^2_{\mathrm{PX}} = \sum_{i=1}^{m} \left(n_i - np_i(\widehat{\boldsymbol{\theta}})\right)^2 \Big/ np_i(\widehat{\boldsymbol{\theta}})$$

为 Chernoff-Lehmann 检验统计量. χ^2_{PX} 的渐近分布为 $\chi^2_{m-q-1} + \sum_{i=1}^{q} \lambda_i(\boldsymbol{\theta})\chi^2_{1i}$, 其中 χ^2_{m-q-1}, $\chi^2_{11}, \cdots, \chi^2_{1q}$ 相互独立, $\chi^2_{1i} \sim \chi^2_1$, $\lambda_i(\boldsymbol{\theta})$ 是依赖于 $\boldsymbol{\theta}$ 的参数, $i = 1, \cdots, q$.

2. EDF 型检验

一类著名拟合优度检验是 EDF 型检验, 考虑假设 $H_0: F = F_0$; $H_1: F \neq F_0$. 设 $X_1, \cdots, X_n$ 是来自连续分布函数 F 的样本, F_n 记为样本 $X_1, \cdots, X_n$ 的经验分布, 即

$$F_n(x) = \frac{\sharp\{X_i : X_i \leqslant x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x]}(X_i), \quad x \in \mathbf{R},$$

其中 $\sharp\{\cdot\}$ 表示集合 $\{\cdot\}$ 中元素的个数, I_A 表示示性函数, 则 F_n 是 F 的强相合估计. 当原假设成立时, F_n 与 F 应充分接近. 以 $\rho(F_n, F_0)$ 描述两者的距离, 当 $\rho(F_n, F_0)$ 过大时, 拒绝原假设. 若将 $\rho(F_n, F_0)$ 取作一致距离时, 就是 Kolmogorov-Smirnov 统计量, 定义为

$$D_n(F) = \sup_{x \in \mathbf{R}} |F_n(x) - F_0(x)|, \tag{1.3.4}$$

且有

$$\lim_{n\to\infty} P(\sqrt{n} D_n(F) \leqslant t) = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \mathrm{e}^{-2j^2t^2}, \quad t > 0. \tag{1.3.5}$$

当以 L_2 距离建立检验统计量时, 就形成均方型统计量, 通常称为 Cramér-von Mises 型统计量, 定义为

$$R_n^2 = n\int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 g(x) \mathrm{d}F_0(x).$$

令 $t = F_0(x)$, $G_n(\cdot)$ 为 $U(0,1)$ 的经验分布函数, 则

$$R_n^2 = n\int_0^1 (G_n(t) - t)^2 \phi(t) \mathrm{d}t,$$

$\phi(t)$ 是权函数, 其取不同函数时可得不同的统计量. 取 $\phi(t) = 1$ 时, 得到 Cramér-von Mises 统计量 W_n^2, 即

$$W_n^2 = n\int_0^1 (G_n(t) - t)^2 \mathrm{d}t.$$

取 $\phi(t) = (t(1-t))^{-1}$ 时, 得到 Anderson-Darling 统计量 A_n, 即

$$A_n^2 = n\int_0^1 \frac{(G_n(t) - t)^2}{t(1-t)} \mathrm{d}t.$$

Watson 统计量 U_n^2 是中心化的 W_n^2 统计量, 定义为

$$U_n^2 = n\int_0^1 [G_n(t) - t - E(G_n(t) - t)]^2 \mathrm{d}t.$$

当 H_0 成立时

$$W_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \frac{Z_j^2}{j^2\pi^2},$$

其中 $Z_1, Z_2, \cdots$ 为 i.i.d. N(0, 1) 随机变量序列.

基于 Cramér-von Mises 型统计量的检验是无向检验 (omnibus test), 只提供否定 H_0 的信息, 不能提供在何种特性上偏离原假设的信息. 仅给出 W_n^2 的值, 很难说出数据所服从分布的特征. 在应用中, 当数据的分布与原假设有显著差异时, 不仅要指出存在差异, 而且应指出是什么性质的偏离引起的显著差异. Durbin 等 (1972) 对 W_n^2 的分解及分量特性进行研究, 将 W_n^2 分解成若干分量, 可做特定的某分量对某种偏离敏感性功效分析. 设 $X_1, \cdots, X_n$ 是来自 $F(\cdot, \boldsymbol{\theta})$ 的样本, $\boldsymbol{\theta}$ 是 d 维参数. 考虑局部对立假设序列

$$H_0: \ \boldsymbol{\theta} = \boldsymbol{\theta}_0; \quad H_1: \ \boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\nu} n^{-1/2},$$

$\boldsymbol{\nu}$ 是 d 维常量参数. 功效模拟显示, 对正态性检验, A_n^2 优于 W_n^2, U_n^2.

3. 光滑检验

对于简单假设 $H_0: F = F_0; H_1: F \neq F_0$ 的拟合优度检验，可通过变换 $U_i = F_0(X_i), i = 1, \cdots, n$, 将检验样本 $X_1, \cdots, X_n$ 的总体是否服从 F_0 ，变换为检验 U_i, $i = 1, \cdots, n$ 的总体是否服从 (0, 1) 上的均匀分布. 只要 F 是连续的, 通过积分变换, 都转化为同一问题：(0, 1) 上均匀分布的检验.

Neyman(1937) 提出了光滑检验, 作变换 $U_i = F_0(X_i)$, $i = 1, \cdots, n$, 由 U_i, $i = 1, \cdots, n$ 检验假设

$$H_0: f(u) = 1; \quad u \in (0, 1), \tag{1.3.6}$$

$f(u)$ 是均匀分布的概率密度. Neyman 将对立假设取为包含参数的正交多项式的指数函数, 即将原假设分布嵌于一个参数分布族中, 将分布的拟合优度检验转化为参数的假设检验. k 阶对立概率密度定义为

$$g_k(u) = C(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} \theta_i h_i(u)\right\}, \quad 0 < u < 1, \tag{1.3.7}$$

其中 $\{h_i(u)\}$ 是关于均匀分布的正交多项式集, $C(\boldsymbol{\theta})$ 为规范化常数, $\boldsymbol{\theta}$=$(\theta_1, \cdots, \theta_k)^{\mathrm{T}}$. 这样，假设检验 (1.3.6) 转换为

$$H_0: \ \boldsymbol{\theta} = \mathbf{0}; \quad H_1: \ \boldsymbol{\theta} \neq \mathbf{0}. \tag{1.3.8}$$

光滑检验统计量定义为

$$\Psi_k^2 = \sum_{i=1}^{k} Z_i^2, \quad Z_i = \sum_{j=1}^{n} h_i(U_j) \Big/ \sqrt{n}. \tag{1.3.9}$$

检验统计量 Ψ_k^2 的渐近分布是自由度为 k 的 χ^2 分布 χ_k^2. 当 Ψ_k^2 取值偏大时, 拒绝原假设 H_0(1.3.6). $\{h_i(u)\}$ 是 (0,1) 上的正交多项式, 即

$$\int_0^1 h_s(u)h_t(u)\mathrm{d}u = \delta_{st}, \tag{1.3.10}$$

$$\delta_{st} = 0,\ s \neq t;\ \delta_{st} = 1,\ s = t;\ s,t = 0,1,2,\cdots.$$

任意分布完全已知的一元概率分布的拟合优度检验, 可以转化为 $(0,1)$ 上均匀分布的拟合优度检验. 故 $(0,1)$ 上的光滑检验具有重要意义. 相应地, 可构造含未知参数一元概率分布的光滑检验. 功效模拟表明一元分布光滑检验优于其他常见的 Omnibus 检验. Kallenberg 等 (1995) 转述了 Ranyer 等 (1990) 的总结“don't use those other methods-use a smooth test”.

4. 基于间隔的均匀分布 $U(0,1)$ 的拟合优度检验

上述拟合优度检验问题的三类常用检验是 Pearson χ^2 检验, 基于经验分布函数的 EDF 检验和光滑检验. Rayner 等 (1989) 的分析表明, Kolmogorov-Smirnov 及 Cramer-von Mises 检验能够较好地探测分布函数间的差异, 而基于间隔的拟合优度检验尤其在探测密度差异方面有较高的检验功效. 针对均匀分布的分布特性, Greenwood 提出了基于间隔的 $(0,1)$ 上均匀分布的拟合优度检验.

5. 条件积分变换

设 $X \sim F(x)$, 分布函数 $F(x)$ 连续, 则 $U = F(X) \sim U(0,1)$. 这样, 通过积分变换, 可将对 $F(x)$ 的拟合优度检验转换为对 $U(0,1)$ 的拟合优度检验. 上述结论可推广至多元分布的情形.

1.3.2　多元概率分布的拟合优度检验

Muirhead(1982) 介绍了球对称分布和椭球对称分布的概念与性质, 探讨了椭球对称分布条件下的多元分析. Fang 等 (1990b) 详细介绍了椭球对称分布条件下的广义多元分析. Anderson(2003) 在*An Introduction to Multivariate Statistical Analysis* (3rd ed.) 一书中的多章内容中增加讨论了椭球对称分布假设下的统计推断. 椭球对称分布包含多元正态分布、多元 t 分布、多元柯西分布和多元 Laplace 分布. 多元分布的拟合优度检验是多元分析的应用基础.

原则上, 一元概率分布的拟合优度检验方法可推广到多元概率分布的检验. Justel 等 (1997) 基于条件积分变换, 推广一元 Kolmogorov-Smirnov 检验到多元 Kolmogorov-Smirnov 检验, Huffer 等 (2007) 基于 Pearson χ^2 检验, 做椭球对称分布的拟合优度检验. 实际上, 多元概率分布的拟合优度检验更为复杂, 检验方法的成熟性尚处于发展和检验阶段. 新型多元概率分布的构造被更多地讨论, 以适应多元数据的多样性. Jones(2002) 讨论了偏球对称分布 (skewing spherically symmetric distribution) 的构造, Fang 等 (2002) 讨论了偏椭球对称分布 (meta-elliptical distribution) 的构造. 已有文献对实际多元数据的分布检验, 仅限于多元正态性检验. 利用生成不同分布的随机向量, 做椭球对称分布拟合优度检验的功效模拟.

多元概率分布一般含有未知参数, 当不易通过数据变换消去未知参数时, 需由样本估计未知参数. 一方面, 检验统计量极限分布的利用, 要求有较大的样本容量. 另一方面, 拟合优度检验的相合性是对偏大的样本容量, 倾向拒绝原假设. 这是一个矛盾, 解决的方法是模拟检验统计量趋向极限分布的收敛速度, 以确定适当的样本容量. 当检验统计量收敛速度较慢时, 利用检验统计量的模拟分位点进行检验.

1. 球面均匀性的特征

记 $\boldsymbol{X}=(X_1,\cdots,X_d)^{\mathrm{T}}$, $\|\boldsymbol{X}\|_\alpha=\left(\sum_{i=1}^d|X_i|^\alpha\right)^{1/\alpha}$. Szablowski(1998) 研究了 d 维单位球球面上 L_α 模均匀分布的特性, 即 $d-1$ 维类似柯西分布及球面上混合均匀分布性质, 给出了球上 L_α 模均匀分布的定义, 证明了若 $\boldsymbol{X}$ 是单位球上对称分布, 同时 $\boldsymbol{X}$ 分量的商服从 $d-1$ 维 α-柯西分布, 则 $\boldsymbol{X}$ 服从单位球上的 L_α 模均匀分布, 且认为 α-柯西分布较球面上 L_α 模均匀分布更为简单实用. 方开泰等 (1990) 分析了三维球面上均匀分布、单峰分布、双峰分布、环形带分布及对称环型带的特征.

(1) 均匀分布: 样本点在球面上分布均匀, 没有趋势向.

(2) 单峰分布: 样本点聚集在某一方向附近, 该方向为峰向. 若样本点还关于峰向旋转对称, 则为单极分布.

(3) 双峰分布: 样本点聚集在两个方向附近. 若还关于两个相反的峰向转动旋转对称, 则称为双极分布.

(4) 环形带分布: 若样本点聚集在球面的某一大圆附近, 利用转动惯量及特征根研究球面上分布特性.

2. 多元正态分布的拟合优度检验

在多元分析理论中, 多元正态分布是其基本假设. 多元正态分布的拟合优度检验始终是理论和应用研究关注的热点. 椭球对称分布 (包括多元正态分布) 的总体偏度为 0, d 多维多元正态分布的峰度为 $d(d+2)$. Mardia(1970) 提出了多元偏度