应用数理统计方法

第三章 比较总体离散与分布特征的假设检验

3.1 一个或两个总体离散程度比较

3.2 多总体方差比较

应用举例

3.3 非正态分布的拟合度检验

3.4 正态检验

应用举例

假设检验问题

p.63

- **特征比较**

  两总体或多总体大小比较 2.2 – 2.5

  两总体或多总体离散程度比较 **3.1 – 3.2**

  两总体分布特征比较 3.4

  总体分布是否服从特定理论分布 3.3 – 3.4

- **影响因素**

  方差分析及补充分析 4.1 – 4.4

- **变量关系**

  相关分析 5.1 – 5.2

  回归分析 5.3 – 5.4

总体离散表征 复习

p.39-41

- **离散特征**

  个体的分散程度, 个体距分布中心的距离

- **正态分布总体的离散表征**

  离差 $\sum |x_i - \bar{x}|$

  平方和 $SS = \sum (x_i - \bar{x})^2$

  方差 $s^2 = SS/(n-1)$

  标准差 $s = (s^2)^{0.5}$

  变异系数 $V = 100\ s/\bar{x}$

总体离散特征比较方法

p.154

- **离散比较的假设检验**

  与大小比较类似 包括参数与非参数方法, 涉及一个, 两个或多个总体

  与大小比较不同 用途较少, 方法较少, 无对应关系

总体	参数方法		非参数方法举例	
一个	比较方差的卡方检验	**3.1.1**		
二个	比较方差的 F-检验	**3.1.2**	Siegel-Tukey 检验	**3.1.4**
	比较变异系数的 t-检验	**3.1.3**		
多个	比较方差的对数方差分析	**3.2.1**		
	比较方差的 Bartlett 检验	**3.2.2**		
	比较方差的 Fmax 检验	**3.2.3**		
	比较方差的 Cochran 检验	**3.2.4**		

离散比较检验方法选择

p.156

- **目的**

  确认关注的是离散特征

- **总体个数**

  一个, 两个或多个 如两个或多个湖泊湖水硬度的变异程度是否有差异

- **总体分布特征与参数选择**

  正态分布 – 参数方法 如方差比较, 变异系数比较, 对数方差分析

  非正态分布 – 非参数方法 如 **Siegel-Tukey** 检验

- **对立假设范畴**

  能否排除一半可能 单侧或双侧检验

假设检验 复习

p.64-65

- **基本思路**

  建立: 总体没有显著差别的原假设

  计算: 相伴概率 p, 即原假设成立条件下得到实际观察结果及更极端结果的概率

  – 更极端结果: 比实际观测结果更偏离原假设

  判断: 相伴概率太小则拒绝原假设, 否则接受原假设

3.1

3.1.1 比较一个总体方差和特定值的卡方检验

3.1.2 比较两个总体方差的 F-检验

3.1.3 比较两个总体变异系数的 t-检验

一个或两个总体离散程度比较的参数检验

p.157-164

- **检验方法 – 类比大小比较**

  总体方差与已知值比 H0 $\sigma^2=\sigma_0^2$ H1 $\sigma^2\neq\sigma_0^2$ X^2 检验 均值与已知指比较 t-检验

  两总体方差比较 H0 $\sigma_1^2=\sigma_2^2$ H1 $\sigma_1^2\neq\sigma_2^2$ F 检验 两均值比较 t-检验

  两总体变异系数比较 H0 $V_{p_1}=V_{p_2}$ H1 $V_{p_1}\neq V_{p_2}$ t 检验

- **离散程度比较的参数方法**

  略

总体	参数方法		非参数方法举例
一个	比较方差的卡方检验	3.1.1	
二个	比较方差的 F-检验	3.1.2	Siegel-Tukey 检验 **3.1.4**
	比较变异系数的 t-检验	3.1.3	
多个	比较方差的对数方差分析	**3.2.1**	
	比较方差的 Bartlett 检验	**3.2.2**	
	比较方差的 Fmax 检验	**3.2.3**	
	比较方差的 Cochran 检验	**3.2.4**	

3.1.4

3.1.4 比较两个总体离散程度的 Siegel-Tukey 检验

两总体离散程度比较的非参数方法

p.165-168

- **Siegel-Tukey 检验**

  数据 非参数方法, 基于秩数据

  假设 H0 两总体离散程度无显著差异

  H1 两总体离散程度有显著差异, 或一个比另一个大 (单侧)

  检验 计算检验统计量 **U**, 与 Mann-Whitney U 检验相似

总体	参数方法		非参数方法举例
一个	比较方差的卡方检验	**3.1.1**	
二个	比较方差的 F-检验	**3.1.2**	Siegel-Tukey 检验 3.1.4
	比较变异系数的 t-检验	**3.1.3**	
多个	比较方差的对数方差分析	**3.2.1**	
	比较方差的 Bartlett 检验	**3.2.2**	
	比较方差的 Fmax 检验	**3.2.3**	
	比较方差的 Cochran 检验	**3.2.4**	

Mann-Whitney U 检验 复习

p.102-104

- **方法**

  用途: 两个非对应非正态分布总体大小比较

  适用: 大样本量数据 样本量增加, 功效效率趋向 95%

- **思路**

  两样本一起求秩, 若两总体大小无显著差别, 两样本秩随机出现

  计算任一样本的秩和 两样本秩和信息相同

  若原假设成立, 秩和不应太大或太小 原假设成立条件下的小概率事件

  实际观测结果与更极端情况 区分单侧与双侧

极端, 小概率 ↑ 秩和＝1+2+ …+ 26＝ **351**

秩和＝1+2+3+5+8+…+53＝693

极端, 小概率 ↓ 秩和＝ 31+32+…+56＝**1130**

Siegel-Tukey 检验思路

p.165-166

- **与 Mann-Whitney U 检验比较**

  都是非参数方法 关注排序, 使用秩数据

  不同的求秩方法 相同的检验统计量与检验方法

  Mann-Whitney U 检验, 数据按大小次序排序求秩

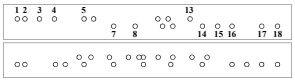
  Siegel-Tukey 检验, 数据按离散次序排序求秩

秩和 = 54

秩和 = 117

大小

离散



Siegel-Tukey 检验方法

p.165-166

- **与 Mann-Whitney U 检验比较**

  都是非参数方法　关注排序, 使用秩数据

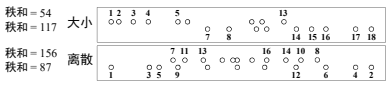
  不同的求秩方法　相同的检验统计量与检验方法

  Mann-Whitney U 检验, 数据按大小次序排序求秩

  Siegel-Tukey 检验, 数据按离散次序排序求秩

- **检验与判断**

  两方法相同　基于秩和计算检验统计量 U, 据此计算 *p*



应用数理统计方法

第三章比较总体离散与分布特征的假设检验

3.1 一个或两个总体离散程度比较

3.2 多总体方差比较

应用举例

3.3 非正态分布的拟合度检验

3.4 正态检验

应用举例

3.2

3.2.1 对数方差分析

3.2.2 Bartlett 检验

3.2.3 Fmax 检验

3.2.4 Cochran 检验

多总体方差检验方法

p.170-179

- **推荐方法**

  对数方差分析　相对严格, 需要样本量较大

- **其它方法**

  Bartlett 检验　对偏态也敏感, 缺乏专性

  Fmax 和 **Cochran** 检验　快速简便, 但不严格

总体	参数方法		非参数方法举例	
一个	比较方差的卡方检验	3.1.1		
二个	比较方差的 F-检验	3.1.2	Siegel-Tukey 检验	3.1.4
	比较变异系数的 t-检验	3.1.3		
多个	比较方差的对数方差分析	3.2.1		
	比较方差的 Bartlett 检验	3.2.2		
	比较方差的 Fmax 检验	3.2.3		
	比较方差的 Cochran 检验	3.2.4		

对数方差分析

p.171-173

- **思路**

  将每个样本拆分为几个样本量接近的子样本

  计算各子样本方差　获得每个样本方差的重复数据

  提供方差的随机波动信息　比较的基础

  类比大小比较　大小差异/随机波动

- **方法**

  H_0 样本来自方差相同的总体, H_1 样本来自方差不同的总体

  原始数据　x_{ij}, $i=1..a, j=1..n_i$

  拆分数据　x_{ijk}, $i=1..a, j=1..m_i, k=1..n_{ij}$ 子样本数约为样本量开方

  计算子样本方差及其对数　$\ln s^2_{ij}$, $i=1..a, j=1..m_i$

  针对方差比较的方差分析　参见 **4.2** 单因子方差分析

应用数理统计方法

第三章比较总体离散与分布特征的假设检验

3.1 一个或两个总体离散程度比较

3.2 多总体方差比较

应用举例

3.3 非正态分布的拟合度检验

3.4 正态检验

应用举例

应用实例 蝗虫的存活时间比较

Willis and Lewis, J. Econ. Entomol., 1957

- **问题与方法**

问题: 不喂食条件下雌雄蝗虫生存时间变异 (波动) 是否有差别

结果可用于采样设计

数据: 观察到的雌性与雌性蝗虫的生存时间分别为 **8.5±3.6** 天和 **4.8±0.9** 天

方法: 比较两个总体方差的 F-检验

假设: $H_0\ \sigma_1^2 = \sigma_2^2$, $H_1\ \sigma_1^2 \neq \sigma_2^2$

- **结果与讨论**

计算: $F = 3.6/0.9 = 4.0$, $p > 0.05$

结论: 没有显著差异

讨论: 拒绝原假设

表述为 "未发现有显著差异"

应用实例 气象参数插值效果评估

Liu and Zuo, Climate Change, 2012

- **问题与方法**

问题: 构建气象参数高分辨空间插值模型, 对模型进行评估

数据: 四地区三时段插值数据　　非正态分布

比较: 插值与观测, 时段之间　　非参数方法

方法: 大小比较用 Kruskall-Wallis 检验; 变异比较用 Siegel-Tukey检验

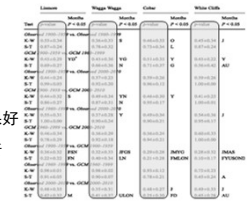
假设: 大小/变异均无显著差异

- **结果与讨论**

计算: p 值均远高于 0.05

结论: 大小/变异均 无显著差异, 插值结果好

讨论: 错误率 **β** 需要更深入的不确定性分析



应用实例 采样代表性评估

Chen et al., Sci. Total. Environ., 2019

- **问题与方法**

问题: 三种示踪剂在地下水中渗透速率的测定变异　氯化物, 硝酸盐, 阿特拉津

数据: 样地内与样地间重复观测　　前者视为重复数据

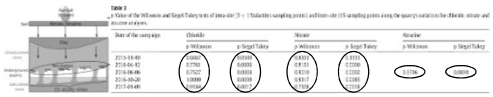
方法: **Wilcoxon** 比大小, **Siegel-Tukey** 比离散　　非正态分布数据

假设: 样地之间测定的大小/变异均无显著差异

- **结果讨论**

结果: 大小不显著, 离散显著

结论: 需增加样地重复数以确保数据可靠性　根据变异预估样本量 参见1.2



应用数理统计方法

第三章 比较总体离散与分布特征的假设检验

3.1 一个或两个总体离散程度比较

3.2 多总体方差比较

应用举例

3.3 非正态分布的拟合度检验

3.4 正态检验

应用举例

假设检验问题

p.63

- **特征比较**

两总体或多总体大小比较　　2.2 – 2.5

两总体或多总体离散程度比较　　3.1 – 3.2

两总体分布特征比较　　**3.4**

总体分布是否服从特定理论分布　3.3 – 3.4

- **影响因素**

方差分析及补充分析　　4.1 – 4.4

- **变量关系**

相关分析　5.1 – 5.2

回归分析　5.3 – 5.4

随机变量的分布特征 复习

p.20

- **总体分布特征**

个体在可能的取值范围内随机出现, 服从特定概率　　如黑白球, $PM_{2.5}$浓度

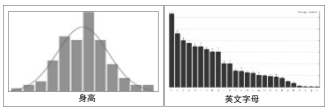
分布特征: 全部个体在可能取值范围内出现概率的分布

- **样本分布特征**

代表性样本表现出与总体类似的分布特征　不了解总体分布

通过对样本分布特征的描述判断总体分布特征

例 人群身高分布, 文献中英文字母分布, 鱼重量分布, 鱼种类分布



不同属性变量的分布 复习

p.31

- **变量属性 – 按观测水平分类**

  定量变量 连续量, 离散量, 顺序量　定性变量 类型量

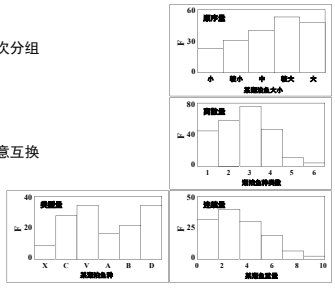
- **定量变量的频数分布图**

  包含次序信息 按取值大小顺次分组

  可用于某些分布检验

- **定性变量的频数分布图**

  无次序关系　不同类型可任意互换



累积分布 复习

p.32-35

- **累积频率/频数分布**

  频数/频率分布 特定取值范围内频数/频率的逐组累加

  概率密度函数　特定取值范围内密度函数的积分

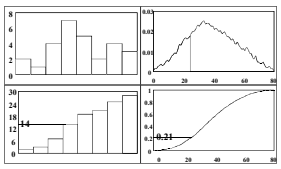
- **特点**

  与一般频率/频率分布信息量相同 表达方式不同

  表现特定范围内的累积效应

- **应用**

  如 Kolmogorov 检验 参见 3.4.1



分布检验问题

p.155

- **分布检验问题**

  两个经验分布比较

  一个经验分布与一个理论分布比较　如正态检验

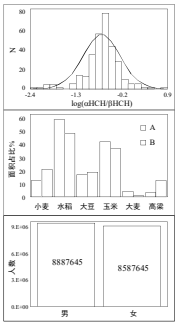
- **典型分布检验问题举例**

  观察数据是否服从正态分布？或对数正态分布？

  某百分数据作角变换后是否服从正态分布？

  两地的主要作物种类分布是否有差别？

  某地区婴儿男女比是否为 1？



分布检验方法

p.155

- **一般分布的拟合度检验**

  在取值范围内比较频数/频率　经验-经验, 经验-理论

  适用于除正态分布外的各类分布

- **正态分布检验**

  经验分布与理论正态分布比较　参数检验的前提

  用专门的拟合度检验等方法　一般拟合度检验不适用 参见3.3正态检验

分布类型	参数方法		其它方法	
正态分布	偏度-峰态检验	3.3.1	作图法, Rankits作图法	3.3.4
			Lillifors 检验	3.3.3
			Shapiro-Wilk 检验	3.3.2
			经验法则, David 快速检验	3.3.4
			D'Agostino 检验	3.3.4
一般分布			**Kolmogorov** 单侧检验	3.4.1
			卡方检验, G-检验	3.4.2-3
			二项检验	3.4.4

一般拟合度检验

p.196-198

- **基本思路**

  考察两个样本分布在不同取值范围内频率的吻合程度　定量或定性变量

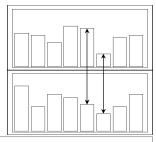
  不应用于正态分布检验　缺少构建对应理论正态分布的基本参数

- **推荐方法**

  卡方检验/G 检验　　类型数据, 等价, 仅算法不同

  Kolmogorov 拟合度检验 定量数据

  二项检验　　二分类类型数据



检验方法	检验对象	数据特点
Kolmogorov 检验	非正态分布	定量数据, 尤其是小样本量数据
卡方检验	**非正态分布**	类型数据, 所有频数 >1 且80%频数 > 4
G-检验	非正态分布	类型数据, 所有频数 > 1 且80%频数 > 4
二项检验	二项分布	二分类类型数据

3.4.2

3.4.2 拟合度卡方检验

3.4.1 Kolmogorov 拟合度检验

拟合度卡方检验思路

p.202

- **用途**

  比较两个类型变量总体的分布　经验与经验分布, 或经验与理论分布

  不应用于连续变量分类变换后数据　损失顺序信息

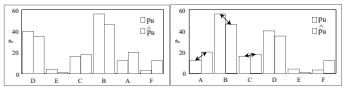
- **思路**

  若分布无显著差异的原假设成立, 频数大小差别不应太大　差别大, 概率小

  比较所有类别频数大小, 计算原假设成立条件下的相伴概率 与 α 比

  $X^2 = \Sigma[(p_i - \hat{p}_i)^2/\hat{p}_i]$, 相对差之和服从卡方分布　取平方强调远点

  相伴概率太低则拒绝原假设



拟合度卡方检验方法

p.202

- **数据**

  所有频数 > 0, 80%以上频数 ≥ 5　不符合要求时, 合并邻组使符合要求

- **假设**

  H_0: 两总体频数分布无显著差异　两经验分布, 经验分布与理论分布

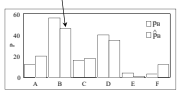
  H_1: 两总体频数分布有显著差异　无单双侧之分

- **计算与判断**

  计算相伴概率或检验统计量　$X^2 = \Sigma[(p_i - \hat{p}_i)^2/\hat{p}_i]$

  直接判断 $p < \alpha$

  间接判断 $X^2 > \chi^2_{\alpha[k-1]}$



3.4.1

3.4.2　拟合度卡方检验

3.4.1　Kolmogorov 拟合度检验

Kolmogorov 拟合度检验

p.198

- **用途**

  比较两个定量变量总体的分布　两个经验分布, 或经验分布与理论分布

- **与卡方检验比较**

  卡方检验　类型变量　不能利用顺序信息, 只应用于检验类型变量的分布

  Kolmogorov　定量变量　需要顺序信息

检验方法	检验对象	数据特点
Kolmogorov检验	非正态分布	定量数据, 尤其是小样本量数据
卡方检验	**非正态分布**	类型数据, 所有频数>1且80%频数>4
G-检验	非正态分布	类型数据, 所有频数>1且80%频数>4
二项检验	二项分布	二分类类型数据

Kolmogorov 拟合度检验思路

p.198

- **分布差异的不同表现**

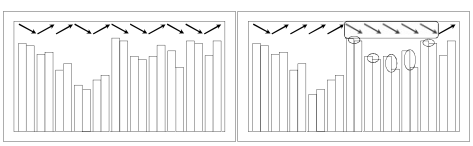
  对应频率差异大小　不显著时不应太大, 卡方检验与**Kolmogorov**的考察对象

  对应频率差异方向　不显著时应随机出现, **Kolmogorov**检验的考察指标

- **Kolmogorov 拟合度检验思路**

  卡方检验仅考察频率差异　差异平方和

  Kolmogorov检验兼顾差异大小和方向　最大累积差, 连续出现同方向差异叠加



利用累积分布

p.198

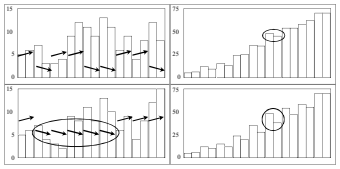
- **原始分布比较**

  拟合度卡方检验: 频数差加和, 与数据次序无关　适用于类型变量

- **累积分布比较**

  Kolmogorov 拟合度检验, 利用累积分布, 包含次序信息　适用于定量变量

  关注最极端情况, 即最大累积差　凸显连续出现同方向差别的叠加效应



分组统计 复习

p.i-v

- **分组统计 – 构建数据的频数/频率分布图**

  将取值范围分为若干组 组距相等或不等 参见3.4.1 Kolmogorov检验

  分别统计各组频数 **F** 样本数 或频率 **f** 频数/总样本量

  结果表达为频数/频率分布图 横坐标为分组取值范围, 纵坐标为频数或频率

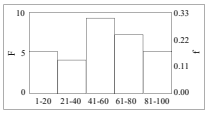
- **分组统计举例**

  数据 72, 6, 80, 7, 100, 77, 42, 77, 24, 10, 76, 57, 45, 81, 37, 28, 63, 48, 83, 98, 24, 41, 54, 59, 45, 41, 76, 18, 90, 12

  分组 6, 7, 10, 12, 18, 24, 24, 28, 37, 41, 41, 42, 45, 45, 48, 54, 57, 59, 63, 72, 76, 76, 77, 77, 80, 81, 83, 90, 98, 100

  统计 $F_i = 5, 4, 9, 7, 5$ $\Sigma F_i = n = 30$

  $f_i = 0.17, 0.13, 0.30, 0.23, 0.17$ $\Sigma f_i = 1$



每个个体自成一组的分组方案

p.199

- **等距分组统计**

  对连续数据分组, 构建频率/频数分布 一般分布与累积分布

  一般用等组距方式, 大多频数大于 1 组内合并会损失细节信息

- **不等组距分组统计**

  每个个体自成一组, 精细刻划累计分布 充分利用连续出现的同方向差异信息

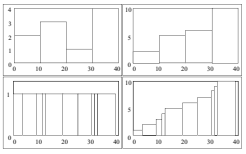
  例 3, 9, 11, 12, 19, 25, 31, 32, 33, 39 n = 10

  等距 1-10, 11-20, 21-30, 31-40

  频数 **2, 3, 1, 4** 累积频数 **2, 5, 6, 10**

  不等距 10 组, 组距不同, 频数为 **1**

  频数 **1, 1, … 1** 累积频数 **1, 2, … 10**



Kolmogorov拟合度检验方法

p.199

- **数据**

  频率数据 将频数数据转换为频率数据

  获得单一个体分组的累积频率分布 充分利用数据信息

- **假设**

  H_0 两总体分分布没有显著差异, H_1 两总体分布有显著差异

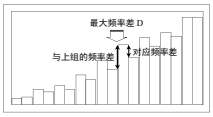
- **计算与检验**

  计算累积频率差, 确定最大累积频率差 对应组及与上一组

  计算相伴概率或检验统计量 **D**

  直接判断 $p < \alpha$

  间接判断 $D > D_{\alpha[n]}$



应用数理统计方法

第三章 比较总体离散与分布特征的假设检验

3.1 一个或两个总体离散程度比较

3.2 多总体方差比较

应用举例

3.3 非正态分布的拟合度检验

3.4 正态检验

应用举例

正态分布 复习

p.33

- **正态分布 – 高斯分布**

  数理统计中最重要的连续变量理论分布

  由均值 分布中心 和标准差 分布形状 两个参数定义

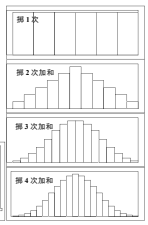
  标准正态分布 均值为 **0**, 标准差为 **1** 的正态分布 参见 1.3.3标准化

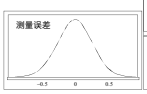
- **特点与重要性**

  集中性与对称性 有对称地向分布中心集中的趋势

  任意分布的均值趋向正态分布 如 多次掷骰子的均值

  参数方法的基础 均值和标准差分别反映大小和离散程度





正态检验

p.182

- **正态检验方法**

  特殊的分布检验 比较观测分布与理论正态分布是否有显著差异

  专门的检验方法 卡方检验不能利用顺序信息

  Kolmogorov检验则缺少构建待比较分布的总体参数 μ, σ

  适用方法 偏度峰态检验, Shapiro-Wilk检验, Lillifors检验

分布类型	参数方法		非参数方法	
正态分布	偏度-峰态检验	3.3.1	作图法, Rankits作图法	3.3.4
			Lillifors 检验	3.3.3
			Shapiro-Wilk 检验	3.3.2
			经验法则, David 快速检验	3.3.4
			D'Agostino 检验	3.3.4
非正态分布			**Kolmogorov** 单侧检验	**3.4.1**
			卡方检验, **G**-检验	**3.4.2-3**
			二项检验	**3.4.4**

正态检验的应用

p.183

- **目的**

  判断是否可以用参数方法进行后续分析 如用算术均值还是中位数表征大小

- **局限性**

  原假设: 与正态分布有显著差异

  拒绝原假设 与正态分布有显著差异, 应当用非参数方法 判断可靠

  接受原假设 不确定是否与正态分布有显著差异, 用参数方法 判断不可靠

  仍然会用参数方法

- **理解**

  经验 参照研究相关变量分布的文献

  理解 认识偏差的可能风险 如回归分析的方差同质性检验,参见5.3.1 最小二乘

正态检验方法

p.183

- **推荐方法**

  偏度峰态检验 判断对正态分布的偏离, 不是拟合度检验

  Shapiro-Wilk 检验 小样本非参数方法

  拟合度 **Lilliefors** 检验 **Kolmogorov**检验的改进

- **其他方法**

  简便, 但不严格

检验方法	方法类型	特点
偏度一峰度检验	参数方法	仅对偏度-峰态敏感
Shapiro-Wilk检验	非参数方法	小样本数据
Lilliefors 检验	非参数方法	**Kolmogorov** 方法的改进
作图法	非参数方法	直观、无概率意义
经验方法等	非参数方法	简便、无概率意义

3.3.1

3.3.1 正态分布的偏度-峰度检验

3.3.2 正态分布的 Shapiro-Wilk 检验

3.3.3 正态分布的 Lillifors 检验

偏度系数与峰态系数 复习

p.26-27

- **表征对正态分布偏离的参数**
- **偏度系数**

  描述两侧偏斜或拖尾程度 $g_1 = 1/ns^3 \Sigma (x_i - \bar{x})^3$, 正态 = 0, 左偏 < 0, 右偏 > 0

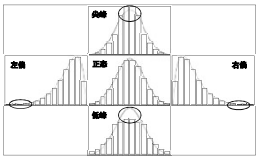
- **峰态系数**

  描述中部聚集趋势 $g_2 = 1/ns^4 \Sigma (x_i - \bar{x})^4 - 3$, 正态 = 0, 低峰 < 0, 尖峰 > 0

- **关于检验**

  "=" 表示无显著差异或"不显著"

  检验 参见 3.3 正态分布检验



偏度-峰态检验

p.184

- **思路**

  检验总体对正态分布的偏离 是否偏斜, 是否尖峰或低峰

  方法的局限性 如不能检验多峰分布 参见1.2.3

- **假设**

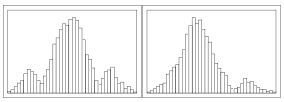
  H_0 $\gamma_i = 0$, H_1 $\gamma_i \neq 0$ ($\gamma_i < 0$, $\gamma_i > 0$) i = 1, 2, 有单双侧之分

- **计算与判断**

  计算相伴概率 p 或检验统计量 t_1, t_2

  直接判断 $p < \alpha$

  间接判断 $t_i > t_{\alpha[n-1]}$, $t_{2\alpha[n-1]}$



3.3.2

3.3.1 正态分布的偏度-峰度检验

3.3.2 正态分布的 Shapiro-Wilk 检验

3.3.3 正态分布的 Lillifors 检验

Shapiro-Wilk 检验

p.187-188

- **方法**

  适用于小样本量数据, 无单双侧之分

  需要专门用表　检验系数表, 检验临界值表 参见p.408

- **计算与检验**

  H_0 样本来自正态分布总体, H_1 非正态分布总体

  查取检验系数 $C_{i[n]}$, 计算相伴概率 p 或检验统计量 W

  直接判断 $p < \alpha$

  间接判断 $W < W_{\alpha[n]}$

3.3.3

3.3.1 正态分布的偏度-峰度检验

3.3.2 正态分布的 Shapiro-Wilk 检验

3.3.3 正态分布的 Lillifors 检验

Lilliefors 检验

p.189

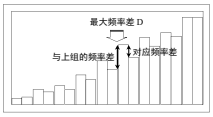
- **方法**

  源自 **Kolmogorov** 检验

  直接用 **Kolmogorov** 检验, 需要构建待比理论正态分布的总体参数 μ 和 σ

  如果用样本 $\bar{x}$ 和 s 替代总体 μ 和 σ, 会导入检验偏差

  Lillifors 检验修正了由此引入的偏差



其它正态检验方法

p.193

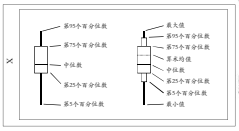
- **作图法**

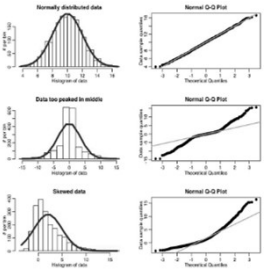
  用正态概率坐标 直观但非定量方法

- **经验法则**

  若以下关系成立, 则为正态分布

  $0.9 < M / \bar{x} < 1.1$, $\bar{x} > 3s$





综合检验方法

p.213

- **同时检验两个或多个特征**

  Kolmogorov 双侧检验　大小, 离散, 分布

  Wald-Wolfowitz 检验　大小, 偏斜, 集中趋势

  Fish 精确概率检验　小样本二分类数据

  两总体卡方检验　大小, 离散, 分布

  多总体卡方检验　大小, 离散, 分布

- **方法缺少专一性, 限制了用途**

  如果显著, 不知道那个特征显著

应用数理统计方法

第三章 比较总体离散与分布特征的假设检验

3.1 一个或两个总体离散程度比较

3.2 多总体方差比较

应用举例

3.3 非正态分布的拟合度检验

3.4 正态检验

应用举例

应用实例 文本相似性研究

Chen & Chen, Expert Syst. Appl. 2011

- **问题和方法**

问题: 将网页文本归入内容相似的类别 判据: 词频向量, 即词频分布

数据: 统计两个随机样本的词频分布

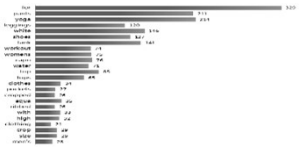
方法: 比较两个词频分布的卡方检验

假设: H_0 两总体词频分布无显著差异, H_1 两总体词频分布有显著差异

- **结果与讨论**

结果: $p < 0.05$ 拒绝原假设

结论: 两样本词频有显著差异



应用实例 多环芳烃被动采样器校验

Tao et al., ES&T 2007

- **问题和方法**

背景: 被动采样需用主动采样器校验　　成本低, 不用电源, 但无流量数据

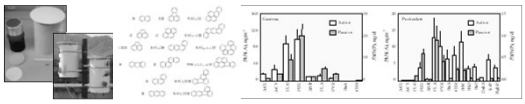
数据: 同步采样获得 16 种多环芳烃测定数据 气态与颗粒态

方法: 比较主动与被动样品分布谱的拟合度卡方检验

假设: H_0 谱分布无差别 H_1 谱分布有差别　　分别检验气态和颗粒态对象

- **结果与讨论**

结果: 气态 p>0.05 不显著; 颗粒态p<0.05显著 高环化合物被动采样效率极低



应用实例 环境辐射异常监测

Reinhart et al., Nucl. Inst. Method, 2015

- **问题和方法**

问题: 怀疑某地环境辐射异常

数据: 获得高分辨背景和样地伽马射线谱

方法: 比较两个连续变量分布射线谱的 **Kolmogorov**检验　判断样地是否异常

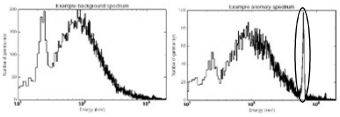
假设: H_0 分布谱无显著差别, H_1 分布谱有显著差别

- **结果与讨论**

结果: $p < 0.05$, 有显著差异

讨论: 累积差与高分辨

与卡方检验比较



应用实例 土壤微量元素含量统计分布表征

深圳土壤微量元素背景值报告

- **问题和方法**

问题: 表土微量元素背景值的统计分布特征

数据: 83 个样品, 11 种微量元素含量　铜,鎘,铅,锌,铬,钒,镍,汞,铁,锰,氟

方法: 用偏度-峰态检验从两方面判断数据是否服从正态分布

假设: H_0 $\gamma_i = 0$, H_1 $\gamma_i \neq 0$, i = 1, 2

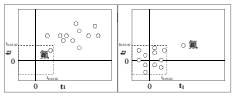
- **结果与讨论**

结果: 计算每个元素的 t_1, t_2 值, 与临界值比较　$t_1 - t_2$ 散点图, α=0.05 临界值边界

尖峰右拖尾, 对数正态分布　　$t > t_{\alpha[n-1]}$, 氟例外

对数变换, 成正态分布

讨论: 可直接用 p1–p2 散点图, 0.05 边界



应用实例 肠易激综合征分布特征

Turna et al., J. Psychiatric Res., 2019

- **问题和方法**

问题: 为研究强迫症与肠易激综合征的关系

需要首先判断变量的分布特征

数据: 募集强迫症实验组 n=21 与对照组 n=22　观测胃肠道症状严重程度**GSRS**

方法: 用**Shapiro-Wilk** 检验是否服从正态分布　小样本

假设: H_0 两组人群 **GSRS** 服从正态分布;

H_1 不服从正态分布

- **结果与讨论**

结果: 接受原假设, 可采用 t-检验进行后续大小比较

讨论: 关于接受原假设结论的局限性

应用实例 口罩对颗粒物的去除率的统计分布研究

Shen et al., EI, 2020

- **问题和方法**

问题: 研究六种口罩对空气中不同粒径颗粒物去除率的差异

数据: 模拟实验获得数据为百分数, 尝试角变换 $y' = \mathrm{Arcsin}\,(y^{0.5})$

若能获得正态分布数据, 后续分析可用参数方法

方法: 用**Lilliefors** 检验　分别检验原始数据和角变换后数据是否服从正态分布

假设: H_0 服从正态分布, H_1 不服从正态分布

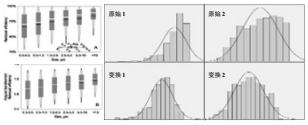
- **结果与讨论**

结果: 原始数据拒绝原假设

变换数据接收原假设

讨论: 高值缺失 泄露 ?

低成本传感器的局限性 ?



应用实例 北京大气颗粒物含量表征

Han et al., Environ. Pollut, 2015

- **问题和方法**

问题: 北京市 2014 年冬室外 $PM_{2.5}$ 浓度的频率分布特征

数据: 利用在线传感器, 观察到双峰分布 20 μg/m³ 左右和大于 100 μg/m³

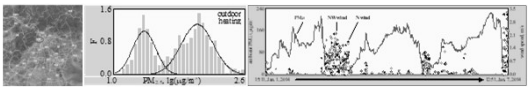
解释: 重污染与清洁天的快速转换　　地形与气象驱动, 风向与风速影响

方法: 对数变换后作偏度-峰态检验　　参照气象条件拆解两个总体

- **结果与讨论**

结果: 均为对数正态分布　　包括两个可拆分总体的双峰分布

讨论: 提供阐释动态变化成因的证据　　相对清洁与重污染情景



应用实例 肺癌易感性对风险评估的影响

Shen et al., Sci. Reports, 2014

- **问题和方法**

问题: 除环境与随机因素外, 肺癌发病率也受易感性影响

考虑与不考虑易感性的模拟结果差异　统计分布比较

数据: 模拟全球大气苯并芘浓度及肺癌风险　典型的对数正态分布

方法: 用**Kolmogorov**检验比较有无易感性的模拟结果　比较两个对数正态分布

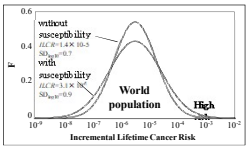
假设: H_0 两总体分布无显著差异, H_1 有显著差异 蓝与绿

- **结果与讨论**

结果: 易感性对人群风险有显著影响

讨论: 均值无差别, 但平均风险增加一倍

高风险人群贡献大幅上升



要点

- **离散检验**

参数方法　　卡方, F, t, 对数方差分析

非参数方法　　Siegel-Tukey

- **非正态分布的拟合度检验**

卡方-G　　考察频率相对差异

Kolmogorov　　考察差异大小和方向, 基于等组距分组的累积分布

- **正态检验**

偏度-峰态检验　优先推荐, 非单峰例外

Shapiro-Wilk　　小样本量

Lilliefors　　源自Kolmogorov, 也是拟合度检验

谢谢