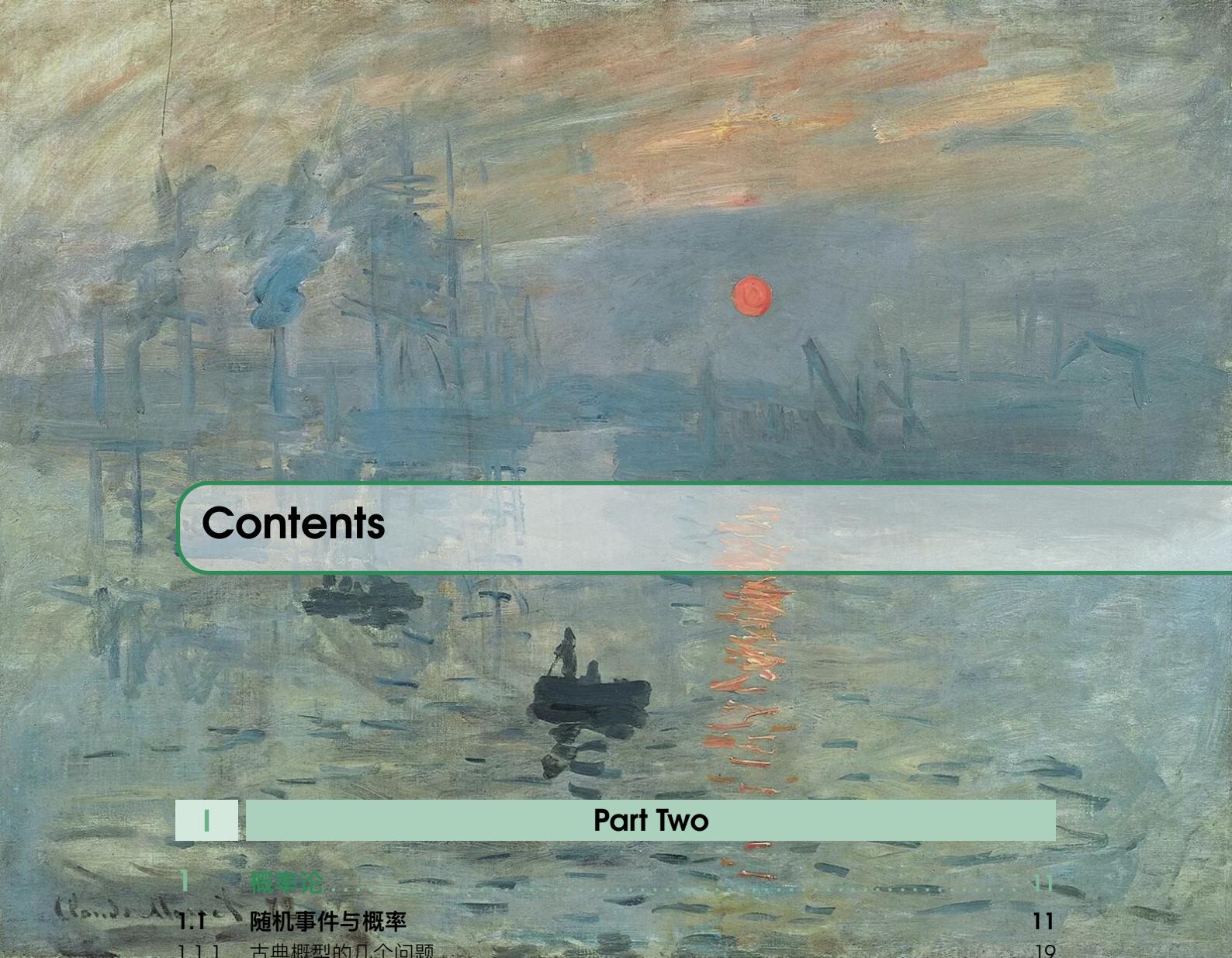


Statistics

Xin Wen



Contents

I

Part Two

1	概率论	11
1.1	随机事件与概率	11
1.1.1	古典概型的几个问题	19
1.2	随机变量	20
1.2.1	常用离散分布	24
1.2.2	常用连续分布	29
1.2.3	随机变量函数的分布	35
1.2.4	分布的其他特征数	38
1.3	多维随机变量及其分布	40
1.3.1	多维随机变量及其联合分布	40
1.3.2	边际分布与随机变量的独立性	42
1.3.3	多维随机变量函数的分布	45
1.3.4	多维随机变量的特征数	51
1.3.5	条件分布与条件期望	55
2	大数定律与中心极限定理	61
2.1	随机变量序列的两种收敛性	61
2.2	几种收敛之间的关系	66
2.3	特征函数	68
2.4	大数定律	73
2.4.1	强大数定理	76
2.5	中心极限定理	78

3	统计量及其分布	85
3.1	一些基本的定义	85
3.2	样本数据的整理与显示	88
3.3	统计量及其分布	89
3.4	抽样基本定理	94
3.5	充分统计量	102
3.6	Probability	104
3.7	Common Families of Distribution	105
3.7.1	Exponential Families	105
3.7.2	Inequalities	108
3.7.3	Identities	109
3.8	Properties of Random Sample	109
3.8.1	Convergence Concepts	109
4	参数估计	111
4.0.1	UE	111
4.0.2	判断相合性的一些定理	115
4.1	极大似然	117
4.2	统计量评估标准--最小方差无偏估计	123
4.2.1	完备统计量	130
4.2.2	信息量--最小方差的表达式	132
4.2.3	有效估计	137
4.3	贝叶斯估计	137
4.3.1	最小二乘估计	139
4.4	区间估计	141
5	假设检验	153
5.0.1	正态总体参数假设检验	163
5.1	其他分布的假设检验	171
5.1.1	指数分布参数的假设检验	171
5.2	似然比检验与分布拟合检验	173
5.2.1	p 值	178
5.3	优势检验	179
5.3.1	分布的 χ^2 拟合优度检验	183
5.4	正态性检验	186
5.5	非参数检验	189
6	方差分析	195
6.1	单因素方差分析	196
6.2	多重比较	202
6.3	方差齐性检验	205
6.4	一元线性回归	207
6.5	一元非线性回归	215

7	非参数检验	217
7.1	两匹配样本非参数检验	221
7.2	两独立样本的非参数检验	221

II

Part Three

8	贝叶斯统计	225
8.0.1	共轭先验分布	227
8.0.2	充分统计量	229
8.0.3	指数分布族	230
8.1	贝叶斯推断	231
8.1.1	可信区间	234
8.1.2	假设检验	235
8.1.3	简单原假设对复杂的备择假设	236
8.1.4	预测	237
8.1.5	似然原理	238
8.1.6	习题	238
8.2	先验分布的确定	241
8.2.1	利用先验信息确定先验分布	241
8.2.2	利用边缘分布 $m(x)$ 确定先验密度	241
8.2.3	Reference 先验的计算	251
8.3	决策	254
8.3.1	先验期望准则	255
8.3.2	后验风险准则	260
8.4	考试重点	261
8.5	期末复习	262
8.5.1	期中	262
8.5.2	第一章	262
8.5.3	第二章	266
8.6	第三章	268
8.7	第四章	272
8.8	第五章	272

III

Part Four

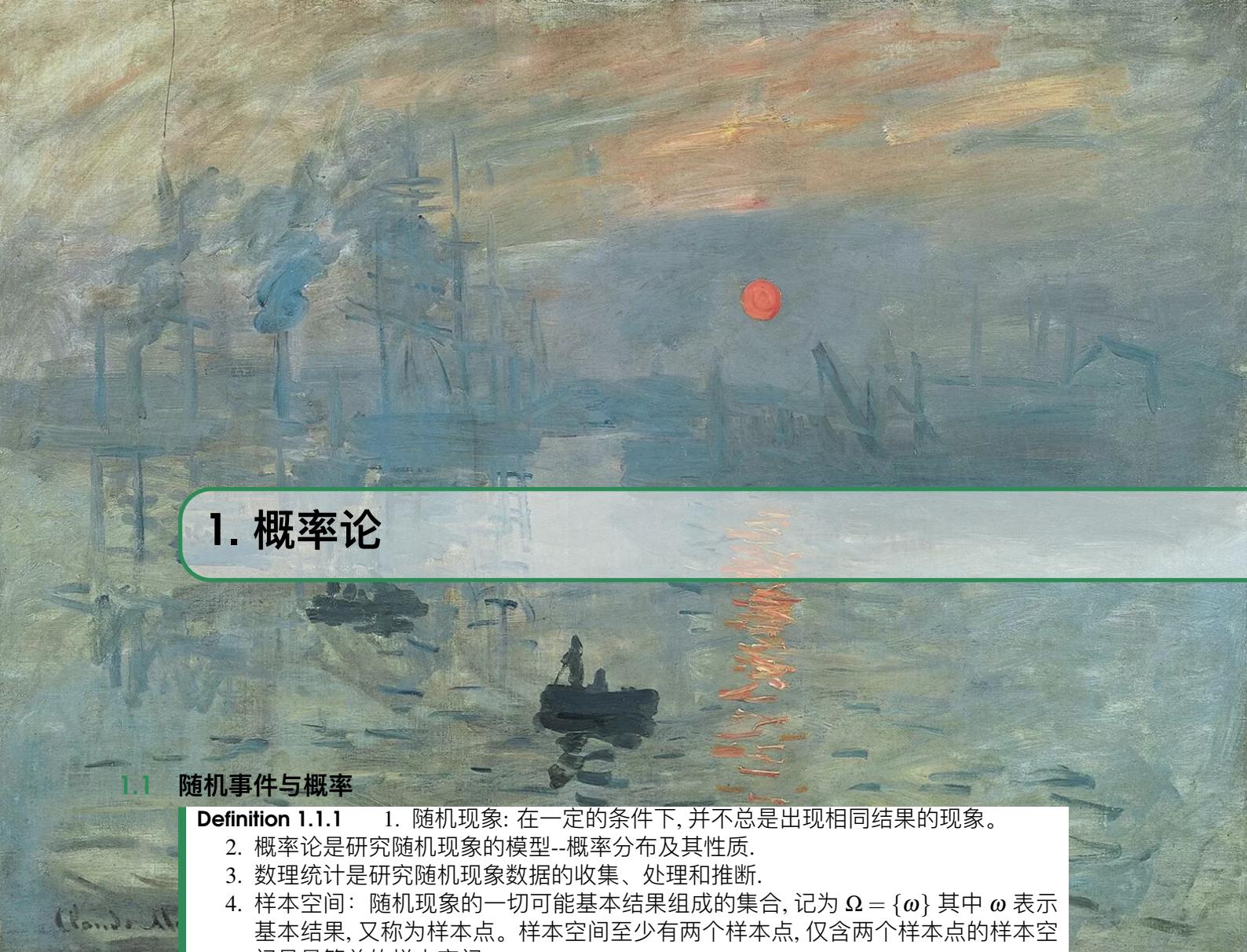
9	一元回归	277
9.1	参数估计	280
9.1.1	最小二乘估计 OLSE	280
9.2	极大似然估计	282
9.3	假设检验	285
9.3.1	t 检验	285
9.3.2	F 检验	285
9.3.3	相关系数检验	286
9.3.4	决定系数	286

9.4 残差检验	287
9.4.1 置信区间	288
9.4.2 控制问题	290
10 多元回归	291
10.0.1 参数估计	293
10.0.2 极大似然估计	295
10.1 显著性检验	297
10.2 违背基本假设的情况	302
10.2.1 异方差性	302
10.2.2 自相关	304
10.2.3 BOX-COX	307
10.2.4 异常点和强影响点	308
10.2.5 总结	309
11 自变量的选择和逐步回归和多重共线性	311
11.1 自变量的选择	311
11.1.1 最优子集	312
11.2 逐步回归	314
11.2.1 前进法	314
11.2.2 后退法	315
11.2.3 逐步回归	315
11.3 多重共线性	315
11.3.1 多重共线性的影响	315
11.3.2 多重共线性的诊断	316
11.3.3 消除多重共线性的方法	318
12 岭回归	319
12.0.1 岭回归估计的性质	319
12.0.2 岭参数的选择	321
13 主成分回归和偏最小二乘	323
13.1 主成分回归	323
13.2 偏最小二乘	324
14 非线性	329
14.1 非线性	329
14.2 Logistic 回归	330
15 多元正态及参数估计	333
15.0.1 多元正态的性质	333
15.1 假设检验	347
15.1.1 霍特林 (Hotelling) T^2 分布	352
15.1.2 威尔克斯 (Wilks) Λ 统计量及其分布	354
15.1.3 单总体均值向量的检验及置信域	356

16	线性模型	367
16.0.1	广义逆	368
16.0.2	偏序	381
16.0.3	矩阵微商	386
16.0.4	多元正态	394
17	线性模型参数估计	413
17.0.1	最小二乘估计	413
17.0.2	约束最小一乘估计	418
17.0.3	广义最小一乘估计	421
Bibliography		429
Articles		429
Books		429

Part Two

1	概率论	11
1.1	随机事件与概率	
1.2	随机变量	
1.3	多维随机变量及其分布	
2	大数定律与中心极限定理	61
2.1	随机变量序列的两种收敛性	
2.2	几种收敛之间的关系	
2.3	特征函数	
2.4	大数定律	
2.5	中心极限定理	
3	统计量及其分布	85
3.1	一些基本的定义	
3.2	样本数据的整理与显示	
3.3	统计量及其分布	
3.4	抽样基本定理	
3.5	充分统计量	
3.6	Probability	
3.7	Common Families of Distribution	
3.8	Properties of Random Sample	
4	参数估计	111
4.1	极大似然	
4.2	统计量评估标准--最小方差无偏估计	
4.3	贝叶斯估计	
4.4	区间估计	
5	假设检验	153
5.1	其他分布的假设检验	
5.2	似然比检验与分布拟合检验	
5.3	优势检验	
5.4	正态性检验	
5.5	非参数检验	
6	方差分析	195
6.1	单因素方差分析	
6.2	多重比较	
6.3	方差齐性检验	
6.4	一元线性回归	
6.5	一元非线性回归	
7	非参数检验	217
7.1	两匹配样本非参数检验	
7.2	两独立样本的非参数检验	



1. 概率论

1.1 随机事件与概率

- Definition 1.1.1**
1. 随机现象: 在一定的条件下, 并不总是出现相同结果的现象。
 2. 概率论是研究随机现象的模型--概率分布及其性质.
 3. 数理统计是研究随机现象数据的收集、处理和推断.
 4. 样本空间: 随机现象的一切可能基本结果组成的集合, 记为 $\Omega = \{\omega\}$ 其中 ω 表示基本结果, 又称为样本点。样本空间至少有两个样本点, 仅含两个样本点的样本空间是最简单的样本空间
 5. 随机事件: 随机现象的某些样本点组成的集合. 常用大写字母 A, B, C 等表示, Ω 表示必然事件, \emptyset 表示不可能事件.
 6. 随机变量: 用来表示随机现象结果的变量, 常用大写字母 X, Y, Z 等表示。
 7. 事件的表示有多种
 - (a) 用集合表示, 这是最基本形式;
 - (b) 用准确的语言表示
 - (c) 用等号或不等号把随机变量与某些实数联结起来表示.
 - 事件间的关系
 - (a) 包含关系: 如果属于 A 的样本点必属于 B , 即事件 A 发生必然导致事件 B 发生, 则称 A 被包含在 B 中, 记为 $A \subset B$
 - (b) 相等关系: 如果 $A \subset B$ 且 $B \subset A$, 则称 A 与 B 相等, 记为 $A = B$
 - (c) 互不相容: 如果 $A \cap B = \emptyset$, 即 A 与 B 不可能同时发生, 则称 A 与 B 互不相容.
 - R 对立事件一定是互不相容的事件, 即 $A \cap \bar{A} = \emptyset$. 但互不相容的事件不一定是对立事件.
 8. 事件运算
 - (a) 事件 A 与 B 的并: 事件 A 与 B 中至少有一个发生, 记为 $A \cup B$;
 - (b) 事件 A 与 B 的交: 事件 A 与 B 同时发生, 记为 $A \cap B$ 或 AB
 - (c) 事件 A 对 B 的差: 事件 A 发生而 B 不发生, 记为 $A - B$
 - (d) 对立事件: 事件 A 的对立事件, 即“ A 不发生”, 记为 \bar{A}

9. 事件的运算性质

- (a) 并与交满足结合律和交换律
 (b) 交对并满足分配律

$$A(B \cup C) = AB \cup AC$$

- (c) 并对交满足分配律

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

- (d) 棣莫弗公式 (对偶法则)

$$\begin{aligned} \overline{A \cup B} &= \bar{A} \cap \bar{B}, \quad \overline{A \cap B} = \bar{A} \cup \bar{B} \\ \overline{\bigcup_{i=1}^n A_i} &= \bigcap_{i=1}^n \overline{A_i}, \quad \overline{\bigcap_{i=1}^n A_i} = \bigcup_{i=1}^n \overline{A_i} \\ \overline{\bigcup_{i=1}^{\infty} A_i} &= \bigcap_{i=1}^{\infty} \overline{A_i}, \quad \overline{\bigcap_{i=1}^{\infty} A_i} = \bigcup_{i=1}^{\infty} \overline{A_i} \end{aligned}$$

10. 事件域: 含有必然事件 Ω , 并关于对立运算和可列并运算都封闭的事件类 \mathcal{F} 称为事件域, 又称为 σ 代数. 具体说, 事件域 \mathcal{F} 满足:

- (a) $\Omega \in \mathcal{F}$
 (b) 若 $A \in \mathcal{F}$, 则对立事件 $\bar{A} \in \mathcal{F}$
 (c) 若 $A_n \in \mathcal{F}, n = 1, 2, \dots$, 则可列并 $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

11. 两个常用的事件域

- (a) 离散样本空间 Ω (有限集或可列集) 内的一切子集组成的事事件域
 (b) 连续样本空间 Ω (如 \mathbb{R}, \mathbb{R}^2) 等内的一切博雷尔集, 如区间或矩形, 逐步扩展而成的事件域。

12. (分割) 把样本空间 Ω 划分为 n 个互不相容的事件 D_1, D_2, \dots, D_n 的行动称为 Ω 的一个分割, 记为 $\mathcal{D} = \{D_1, D_2, \dots, D_n\}$. 显然, 分割 \mathcal{D} 有性质:

$$\text{诸 } D_i \text{ 互不相容, , 且 } \bigcup_{i=1}^n D_i = \Omega$$

Theorem 1.1.1 — 科大概统第一章基础知识.	
	1. 主观概率优点: 有广泛的生活基础, 反映了主体的倾向性
2.	随机事件: 有一个明确的试验, 并且知道试验的所有可能的结果, 对试验有明确的描述 + 结果的不确定性
3.	随机事件的每一个结果叫做基本事件- ω
4.	互斥: 不能在同一次试验中都出现
5.	事件独立: B 发生与否与对 A 发生的概率不产生影响
6.	事件的划分: 事件之间两两互斥, 但是每次试验至少发生一个
7.	poisson 的推导: 把 0, 1 区间进行划分, 划分成 n 分, 每一份最多只可能发生一次, 并且每个区间发生与否之间是相互独立的
8.	指数分布描述寿命, 无记忆性的表述: 就是说, “元件在时刻 x 尚为正常工作的条件下, 其失效率总保持为某个常数 $\lambda > 0$, 与 x 无关; 期望是平均寿命, 所以参数越大, 寿命越短”

9. 考虑老化率不是常数，而是随着时间的增加，越来越大：指数分布是威布儿分布的特例
10. 随机变量独立：取值的概率不受其他随机变量的影响
11. 联合的概率密度函数可以拆解成 n 个函数的乘积，其中每个函数只与 x_i 有关，此时可以说这 n 个随机变量之间是独立的。这些函数与边缘概率密度函数之差常数 c 倍。(因为你在关于某一个随机变量积分的时候，其他的随机变量是常数)
12. $Y = X^2$ 时， y 的概率密度函数：

$$l(y) = \frac{1}{2}y^{-1/2}[f(\sqrt{y}) + f(-\sqrt{y})] \quad (y > 0)$$

13. 多元变换时要求一一对应，雅可比行列式是逆变换对于每个变量的导数矩阵的行列式
14. 正态分布的“再生性”：如果 Y 服从正态分布，而 Y 表成两个独立随机变量 X_1, X_2 之和，则 X_1, X_2 必都服从正态分布
15. Γ 函数换元 + 利用正态概率密度函数可以计算出 $\Gamma(1/2)$
16. 若 X_1, \dots, X_n 独立，且都服从指数分布，则

$$X = 2\lambda(X_1 + \dots + X_n) \sim \chi_{2n}^2$$

首先，由 X_i 的密度函数和 Gamma 分布的可加性，知 $2\lambda X_i$ 的密度区数为 $\frac{1}{2}e^{-x/2}(x > 0)$ ；当 $x \leq 0$ 时密度区数为 0。但在 Gamma 函数的定义式中令 $n = 2$ ，可知这正好是 χ_2^2 的密度函数因此 $2\lambda X_i \sim \chi_2^2$ 。再因 X_1, \dots, X_n 独立，利用 χ^2 的可加性可知

17. 数学期望要绝对收敛的原因：条件收敛的级数通过改变顺序可以收敛于任意一个预先给的数
18. 条件期望反映了随着随机变量 X 的取值 x 的变化对于 Y 的平均变化
19. 由于分布函数总是存在的，所以中位数也是总是存在的，而期望不是总都存在的
20. 泊松分布逼近二项分布。中心用于 p 固定，因而当 n 很大时 np 很大；而泊松逼近则用于 p 很小（可设想成 p 随 n 变化以趋向于 0）但 $np = \lambda$ 不太大时。共同之点是 n 必须相当大。
21. 数理统计学就是这样一门学科，它使用概率论和数学的方法，研究怎样收集（通过试验或观察）带有随机误差的数据，并在设定的模型（称为统计模型）之下，对这种数据进行分析（称为统计分析），以对所研究的问题做出推断（称为统计推断）。

Theorem 1.1.2 — 随机变量的独立和事件之间的关系. 如果连续变量 X_1, \dots, X_n 独立，则对任何 $a_i < b_i \quad (i = 1, \dots, n)$ ，由

$$A_1 = \{a_1 \leq X_1 \leq b_1\}, \dots, A_n = \{a_n \leq X_n \leq b_n\}$$

式定义的 n 个事件 A_1, \dots, A_n 也独立。反之，若对任何 $a_i < b_i \quad (i = 1, \dots, n)$ ，事件 A_1, \dots, A_n 独立，则变量 X_1, \dots, X_n 也独立。

Definition 1.1.2 — 概率的公理化定义. 定义在事件域 \mathcal{F} 上的一个实值函数 $P(A)$ 满足：

1. 非负性公理 若 $A \in \mathcal{F}$ ，则 $P(A) \geq 0$
2. 正则性公理 $P(\Omega) = 1$
3. 可列可加性公理 若 $A_1, A_2, \dots, A_n, \dots$ 互不相容，有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

则称 $P(A)$ 为事件 A 的概率, 称三元素 (Ω, \mathcal{F}, P) 为概率空间.

R 概率的公理化定义刻画了概率的本质, 概率是集合(事件)的函数, 若在事件域上给出一个函数, 当这个函数能满足上述三条公理, 就被称为概率; 当这个函数不能满足上述三条公理中任一条, 就被认为不是概率。

Definition 1.1.3 — 概率的定义. 1. (描述性) 把刻画事件 A 发生可能性大小的数量指标叫做事件 A 的概率
2. (统计定义) 在相同条件下重复进行 n 次试验, 当试验次数充分大时, 事件 A 发生的频率稳定的在常数 p 附近波动, 并随着试验次数的增大逐渐稳定于 p , 则称常数 p 为事件 A 的概率

Theorem 1.1.3 重复组合: 从 n 个不同元素中每次取出一个, 放回后再取下一个, 如此连续取 r 次所得的组合称为重复组合, 此种重复组合总数为 $\binom{n+r-1}{r}$. 注意: 这里的 r 也允许大于 n

重复组合数的得出可如下考虑: 将此 n 个元素画成 n 个盒子, 如果第 i 个元素取到过一次, 则在此盒子中用“0”作一记号. $|00||0|$ 第一个元素取到过 2 次, 第 2 个元素取到过 0 次, 第 3 个元素取到过 1 次. 因为共取 r 次, 所以共有 r 个“0”, $n+1$ 个“1”如此所有的 r 个“0”和 $n+1$ 个“1”中除了两端的那两个“1”不可以动外, 共有 $n+r-1$ 个“0”和“1”可随意放置, 不同的放置表示不同的取法. 因此重复组合数就等于在此 $n+r-1$ 个位置上任选 r 个放“0”, 或此 $n+r-1$ 个位置上任选 $n-1$ 个放“1”而 $\binom{n+r-1}{r}$ 和 $\binom{n+r-1}{n-1}$ 是相等的.

Theorem 1.1.4 — 确定概率的频率方法. 它的基本思想是:

1. 与考察事件 A 有关的随机现象可大量重复进行;
2. 在 n 次重复试验中, 记 $n(A)$ 为事件 A 出现的次数, 称

$$f_n(A) = \frac{n(A)}{n}$$

为事件 A 出现的频率

3. 频率的稳定值就是概率
4. 频率的稳定值就是概率

Theorem 1.1.5 — 确定概率的古典方法. 它的基本思想是:

1. 所涉及的随机现象只有有限个样本点, 脉如为 n 个
2. 每个样本点发生的可能性相等(称为等可能性)
3. 若事件 A 含有 k 个样本点, 则事件 A 的概率为

$$P(A) = \frac{\text{事件 } A \text{ 所含样本点的个数}}{\Omega \text{ 中所有样本点的个数}} = \frac{k}{n}$$

这样确定的概率常称为古典概率. 计算其分子与分母常用到排列与组合

Theorem 1.1.6 — 确定概率的几何方法. 它的基本思想是:

1. 如果一个随机现象的样本空间 Ω 充满某个区域, 其度量(长度, 面积或体积等) 大

小可用 S_Ω 表示

2. 任意一点落在度量相同的子区域内是等可能的:
3. 若事件 A 为 Ω 中某个子区域, 且其度量为 S_A , 则事件 A 的概率为

$$P(A) = \frac{S_A}{S_n}$$

这样确定的概率常称为几何概率, 计算其分子与分母要涉及长度、面积、体积等, 有时还请用重积分等工具.

■ **Example 1.1** (蒲丰投针问题) 平面上画有间隔为 $d(d > 0)$ 的等距平行线, 向平面任意投掷一枚长为 $l(l < d)$ 的针, 求针与任一平行线相交的概率。

Proof. 以 x 表示针的中点与最近一条平行线的距离, 又以 φ 表示针与此直线间的交角, 易知样本空间 Ω 满足

$$0 \leq x \leq d/2, \quad 0 \leq \varphi \leq \pi$$

由这两式可以确定 $x - \varphi$ 平面上的一个矩形 Ω , 这就是样本空间, 其面积为 $S_\Omega = d\pi/2$. 这时针与平行线相交 (记为事件 A) 的充要条件是

$$x \leq \frac{l}{2} \sin \varphi$$

由于针是向平面任意投掷的, 所以由等可能性知这是一个几何概率问题. 由此得

$$P(A) = \frac{S_A}{S_\Omega} = \frac{\int_0^\pi \frac{l}{2} \sin \varphi d\varphi}{\frac{d}{2}\pi} = \frac{2l}{d\pi}$$

如果 l, d 为已知, 则以 π 的值代入上式即可计算得 $P(A)$ 之值. 反之, 如果已知 $P(A)$ 的值, 则也可以利用上式去求 π , 而关于 $P(A)$ 的值, 可用从试验中获得的频率去近似它: 即投针 N 次, 其中针与平行线相交 n 次, 则频率 n/N 可作为 $P(A)$ 的估计值, 于是由

$$\frac{n}{N} \approx P(A) = \frac{2l}{d\pi}$$

可得

$$\pi \approx \frac{2lN}{dn}$$

■

Definition 1.1.4 — 主观方法. 确定概率的主观方法一个事件 A 的概率 $P(A)$ 是人们根据经验, 对该事件发生的可能性大小所作出的个人信念。概率是定义在事件域 \mathcal{F} 上的集合函数, 且满足三条公理. 前三种确定概率的方法自动满足三条公理, 而主观方法确定概率要加验证, 若不满足三条公理就不能称为概率.

Theorem 1.1.7 抽样模型 (包括返回抽样与不返回抽样) 与盒子模型可概括很多古典概率的计算问题, 应重点关注。

Proposition 1.1.8 — 概率的性质. 概率是定义在事件域 \mathcal{F} 上的非负、正则和可列可加的集合函数. 概率的运算性质受到集合 (事件) 关系及运算性质的制约, 或者说, 概率的运算性质是依据事件关系及运算性质而给出的。

1. $P(\emptyset) = 0$

2. 有限可加性: 若有限个事件 A_1, A_2, \dots, A_n 互不相容, 则有

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

3. 对立事件的概率对任一事件 A , 有

$$P(\bar{A}) = 1 - P(A)$$

4. 减法公式 (特定场合): 若 $A \supset B$, 则

$$P(A - B) = P(A) - P(B)$$

5. 单调性: 若 $A \supset B$, 则 $P(A) \geq P(B)$

6. 减法公式 (一般场合) 对任意两个事件 A, B , 有

$$P(A - B) = P(A) - P(AB)$$

7. 加法公式: 对任意两个事件 A, B , 有

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

对任意 n 个事件 A_1, A_2, \dots, A_n , 有

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i A_j A_k) \\ &\quad + \cdots + (-1)^{n-1} P(A_1 A_2 \cdots A_n) \end{aligned}$$

8. 半可加性对任意两个事件 A, B , 有

$$P(A \cup B) \leq P(A) + P(B)$$

9. 事件序列的极限

(a) 对 \mathcal{F} 中任一单调不减的事件序列 $F_1 \subset F_2 \subset \cdots \subset F_n \subset \cdots$, 称可列并 $\bigcup_{n=1}^{\infty} F_n$ 为 $\{F_n\}$ 的极限事件, 记为

$$\lim_{n \rightarrow \infty} F_n = \bigcup_{n=1}^{\infty} F_n$$

若 $\lim_{n \rightarrow \infty} P(F_n) = P(\lim_{n \rightarrow \infty} F_n)$, 则称概率 P 是下连续的.

(b) 对 \mathcal{F} 中任一单调不增的事件序列 $E_1 \supset E_2 \supset \cdots \supset E_n \supset \cdots$, 称可列交 $\bigcap_{n=1}^{\infty} E_n$ 为 $\{E_n\}$ 的极限事件, 记为

$$\lim_{n \rightarrow \infty} E_n = \bigcap_{n=1}^{\infty} E_n$$

若 $\lim_{n \rightarrow \infty} P(E_n) = P(\lim_{n \rightarrow \infty} E_n)$, 则称概率 P 是上连续的

10. 概率的连续性: 若 P 为事件域 \mathcal{F} 上的概率, 则 P 既是下连续的, 又是上连续的。
11. 若 P 是 \mathcal{F} 上满足 $P(\Omega) = 1$ 的非负集合函数, 则 P 具有可列可加性的充要条件是 P 具有有限可加性和下连续性。
12. (布尔不等式)

$$P(A \cup B) \leq P(A) + P(B)$$

13. (Bonferroni 不等式)

$$P(AB) \geq P(A) + P(B) - 1$$

Proof. 第十条的证明：证明先证 P 的下连续性。设 $\{F_n\}$ 是 \mathcal{F} 中一个单调不减的事件序列，即

$$\bigcup_{i=1}^{\infty} F_i = \lim_{n \rightarrow \infty} F_n$$

若定义 $F_0 = \emptyset$ ，则

$$\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} (F_i - F_{i-1})$$

由于 $F_{i-1} \subset F_i$ ，显然诸 $(F_i - F_{i-1})$ 两两不相容，再由可列可加性得

$$P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i - F_{i-1}) = \lim_{n \rightarrow \infty} \sum_{i=1}^n P(F_i - F_{i-1})$$

又由有限可加性得

$$\sum_{i=1}^n P(F_i - F_{i-1}) = P\left(\bigcup_{i=1}^n (F_i - F_{i-1})\right) = P(F_n)$$

所以

$$P\left(\lim_{n \rightarrow \infty} F_n\right) = \lim_{n \rightarrow \infty} P(F_n)$$

这就证得了 P 的下连续性。

再证 P 的上连续性。设 $\{E_n\}$ 是单调不增的事件序列，则 $\{\bar{E}_n\}$ 为单调不减的事件序列，由概率的下连续性得

$$\begin{aligned} 1 - \lim_{n \rightarrow \infty} P(E_n) &= \lim_{n \rightarrow \infty} [1 - P(E_n)] = \lim_{n \rightarrow \infty} P(\bar{E}_n) \\ &= P\left(\bigcup_{n=1}^{\infty} \bar{E}_n\right) = P\left(\overline{\left(\bigcap_{n=1}^{\infty} E_n\right)}\right) \\ &= 1 - P\left(\bigcap_{n=1}^{\infty} E_n\right) \end{aligned}$$

注意最后第二个等式用了德摩根公式。至此得

$$\lim_{n \rightarrow \infty} P(E_n) = P\left(\bigcap_{n=1}^{\infty} E_n\right)$$

这就证得了 P 的上连续性。 ■

Proof. 下证第十一条：下证充分性。设 $A_i \in \mathcal{F}, i = 1, 2, \dots$ 是两两不相容的事件序列，由有限可加性可知：对任意有限的 n 都有

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

这个等式的左边不超过 1, 因此正项级数 $\sum_{i=1}^{\infty} P(A_i)$ 收敛, 即

$$\lim_{n \rightarrow \infty} P\left(\bigcup_{i=1}^n A_i\right) = \lim_{n \rightarrow \infty} \sum_{i=1}^n P(A_i) = \sum_{i=1}^{\infty} P(A_i)$$

记

$$F_n = \bigcup_{i=1}^n A_i$$

则 $\{F_n\}$ 为单调不减的事件序列, 所以由下连续性得

$$\lim_{n \rightarrow \infty} P\left(\bigcup_{i=1}^n A_i\right) = \lim_{n \rightarrow \infty} P(F_n) = P\left(\bigcup_{n=1}^{\infty} F_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n\right)$$

■

Definition 1.1.5 — 条件概率. 1. 设事件 A 的概率为 $P(A)$, 若有新的信息 (概括为另一事件 B) 发生, 可能会对事件 A 发生的概率产生影响, 这就要研究条件概率 $P(A|B)$
2. 条件概率设 A, B 是两个事件, 若 $P(B) > 0$, 则称

$$P(A|B) = \frac{P(AB)}{P(B)}$$

为“在事件 B 发生下事件 A 发生的条件概率”, 简称条件概率. 它满足概率的三条公理。

3. 乘法公式

- (a) 若 $P(B) > 0$, 则 $P(AB) = P(B)P(A|B)$
- (b) 若 $P(A_1A_2 \cdots A_{n-1}) > 0$, 则

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_n|A_1A_2 \cdots A_{n-1})$$

4. 全概率公式 设 B_1, B_2, \dots, B_n 互不相容, 且 $\bigcup_{i=1}^n B_i = \Omega$, 如果 $P(B_i) > 0$ $i = 1, 2, \dots, n$, 则对任一事件 A 有

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i)$$

全概率公式提供了计算复杂事件概率的一条有效途径。

5. 贝叶斯公式 设 B_1, B_2, \dots, B_n 互不相容, 且 $\bigcup_{i=1}^n B_i = \Omega$, 如果 $P(A) > 0$ $P(B_i) > 0, i = 1, 2, \dots, n$, 则

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)}, \quad i = 1, 2, \dots, n$$

在贝叶斯公式中, 诸 $P(B_i)$ 称为 B_i 的先验 (试验以前) 概率, 而诸 $P(B_i|A)$ 称为 B_i 的后验 (试验以后) 概率, 它表示在“事件 A 发生”这个新信息后, 对 B_i 的概率作出的修正。

Definition 1.1.6 — 独立性. 1. 如果 $P(AB) = P(A)P(B)$, 则称事件 A 与事件 B 相互独立, 简称 A 与 B 独立. 否则称 A 与 B 不独立或相依。

- 2. 若事件 A 与 B 独立, 则 A 与 \bar{B} 独立, \bar{A} 与 B 独立, \bar{A} 与 \bar{B} 独立。
- 3. 多个事件的独立性设有 n 个事件 A_1, A_2, \dots, A_n , 如果对任意的 $1 \leq i < j < k < \dots \leq n$

n , 以下等式均成立

$$\left\{ \begin{array}{l} P(A_i A_j) = P(A_i) P(A_j) \\ P(A_i A_j A_k) = P(A_i) P(A_j) P(A_k) \\ \dots \\ P(A_1 A_2 \dots A_n) = P(A_1) P(A_2) \dots P(A_n) \end{array} \right.$$

则称此 n 个事件 A_1, A_2, \dots, A_n 相互独立.

4. 若 n 个事件相互独立, 则其任一部分与另一部分也相互独立. 特别把其中部分换为对立事件后, 所得诸事件亦相互独立。
5. 事件间的相互独立性是从实际中提炼出的一个重要概念, 一旦从实际 (相互有无影响) 认可诸事件相互独立性成立, 即可简化事件交的概率的计算。这是大多数情况下的思维过程, 不过还有不少情况下, 需要从定义判断事件间的独立性
6. 试验的独立性假如试验 E_1 的任一结果 (事件) 与试验 E_2 的任一结果 (事件) 都是相互独立的事件, 则称这两个试验相互独立。
7. n 重独立重复试验 (假如一个试验重复进行 n 次, 并各次试验间相互独立, 则称其为 n 重独立重复试验. 假如一个试验只可能有两个结果: A 与 \bar{A} , 则称其为伯努利试验. 假如一个伯努利试验重复进行 n 次, 并各次试验间相互独立, 则称其为 n 重伯努利试验。)

1.1.1 古典概率的几个问题

Theorem 1.1.9 抽样 (摸球): 从 n 张卡片中抽取 r 张. 抽样有两种方式: 放回, 不放回。对结果有两种记录: 讲序 (有序), 不讲序 (无序)

从 n 张卡片中抽取 r 张可能结果总数列于下表:

	放回	不放回
有序	n^r	A_n^r
无序	$\binom{n+r-1}{r}$	$\binom{n}{r}$

Theorem 1.1.10 分配 (占位): 把 r 只球投入 n 个房间. 每个房间容球的规则: 不限制, 有
限制 (只限一球)。球的个性: 可分辨 (编号), 不可分辨。 r 只球投入 n 个房间可能结果总
数列于下表:

把 r 只球投入 n 个房间还有另一种表述. 若以向量 (n_1, n_2, \dots, n_r) 记第 i 只球投
入的房间号码为 n_i $i = 1, 2, \dots, r$, 并称为房号向量. 再以向量 (r_1, r_2, \dots, r_n) 记第 j 号
房间落入的球数为 r_j , $j = 1, 2, \dots, n$, 并称为球数向量, 则可能的向量总数列于下表:

	不限制	有限制	不限制	有限制
可分辨	n^r	A_n^r	房号向量	n^r
不可分辨	$\binom{n+r-1}{r}$	$\binom{n}{r}$	球数向量	$\binom{n+r-1}{r}$

Theorem 1.1.11

$$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$$

$$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n-1}{r-1} + \binom{n-2}{r-1} + \dots + \binom{r-1}{r-1}$$

Theorem 1.1.12

$$(a_1 + a_2 + \cdots + a_n)^r = \sum_{r_1+r_2+\cdots+r_n=r} \frac{r!}{r_1!r_2!\cdots r_n!} a_1^{r_1} \cdots a_2^{r_2} \cdots a_n^{r_n}$$

共有 $\binom{n+r-1}{r}$ 个同类项别地, $(a+b)^r = \sum_{0 \leq k \leq r} \binom{r}{k} a^k b^{r-k}$ 共有 $r+1$ 个(同类)项

1.2 随机变量

Definition 1.2.1 随机变量: 定义在样本空间 Ω 上的实值函数 $X = X(\omega)$ 称为随机变量.

1. 仅取有限个或可列个值的随机变量称为离散随机变量;
2. 取值充满某个区间 (a, b) 的随机变量称为连续随机变量, 这里 a 可为 $-\infty, b$ 可为 $+\infty$

Definition 1.2.2 — 随机变量. 设 $\xi(\omega)$ 是定义于概率空间 (Ω, \mathcal{F}, P) 上的单值实函数, 如果对于直线上任一博雷尔点集 B , 有

$$\{\omega : \xi(\omega) \in B\} \in \mathcal{F}$$

则称 $\xi(\omega)$ 为随机变量, 而 $P\{\xi(\omega) \in B\}$ 称为随机变量 $\xi(\omega)$ 的概率分布.

R 随机变量是样本点的函数, 在测量之前, 我们只知道它可能会取什么值, 也就是样本空间, 不知道到底会取什么, 有随机性。

Definition 1.2.3 分布函数: 设 X 是一个随机变量, 对任意实数 x , 称 $F(x) = P(X \leq x)$ 为 X 的分布函数, 记为 $X \sim F(x)$. 任何一个随机变量都有分布函数。分布函数具有如下三条基本性质。

1. 单调性: $F(x)$ 是单调非减函数, 即对任意的 $x_1 < x_2$, 有 $F(x_1) \leq F(x_2)$
2. 有界性: 对任意的 x , 有 $0 \leq F(x) \leq 1$, 且

$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0, \quad F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = 1$$

3. 右连续性: $F(x)$ 是 x 的右连续函数, 即对任意的 x_0 , 有

$$\lim_{x \rightarrow x_0^+} F(x) = F(x_0)$$

即 $F(x_0 + 0) = F(x_0)$

反之, 可以证明: 具有上述三条性质的函数 $F(x)$ 一定是某一个随机变量的分布函数.

R 分布函数可以唯一决定概率分布

Definition 1.2.4 离散随机变量的概率分布列: 若离散随机变量 X 的可能取值是 $x_1, x_2, \dots, x_n, \dots$, 则称 X 取 x_i 的概率

$$p_i = p(x_i) = P(X = x_i), i = 1, 2, \dots, n$$

为 X 的概率分布列, 简称分布列. 分布列也可用列表方式来表示: 分布列 $p(x_i)$ 具有如

下两条基本性质:

1. 非负性: $p(x_i) \geq 0, i = 1, 2, \dots$
2. 正则性: $\sum_{i=1}^{+\infty} p(x_i) = 1$

Definition 1.2.5 离散随机变量 X 的分布函数为 $F(x) = \sum_{x_i \leq x} p(x_i)$, 它是有限级或可列无限级阶梯函数. 离散随机变量 X 取值于区间 $(a, b]$ 上的概率为 $P(a < X \leq b) = F(b) - F(a)$. 常数 c 可看作仅取一个值的随机变量 X , 即 $P(X = c) = 1$, 它的分布常称为单点分布或退化分布.

Definition 1.2.6 连续随机变量的概率密度函数: 记连续随机变量 X 的分布函数为 $F(x)$. 若存在一个非负可积函数 $p(x)$, 使得对任意实数 x , 有

$$F(x) = \int_{-\infty}^x p(t) dt$$

则称 $p(x)$ 为 X 的概率密度函数, 简称为密度函数. 密度函数 $p(x)$ 具有如下两条基本性质:

1. 非负性: $p(x) \geq 0$
2. 正则性: $\int_{-\infty}^{+\infty} p(x) dx = 1$

R 离散随机变量的分布函数 $F(x)$ 总是右连续的阶梯函数, 而连续随机变量的分布函数 $F(x)$ 一定是整个数轴上的连续函数

Question 1.1 概率密度函数都是连续的吗? 不一定. 例如均匀分布在定义域两个端点处不连续 ■

Question 1.2 分布函数和概率密度函数的联系? 由密度函数通过积分可确定分布函数, 虽然分布函数不能确定唯一的密度函数, 但在几乎处处相等意义上是唯一的, 且在密度函数连续点处有 $F(x) = f(x)$ ■

Definition 1.2.7 连续随机变量 X 的分布函数 $F(x)$ 是 $(-\infty, +\infty)$ 上的连续函数, 它可能在有限个点或可列个点上不可导, 除此以外, 有 $F'(x) = p(x)$ 连续随机变量 X 仅取一点值的概率恒为零, 从而有

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) \\ &= \int_a^b p(x) dx \end{aligned}$$

Proposition 1.2.1

1. 连续随机变量 X 的密度函数不唯一, 但它们必几乎处处相等, 即它们不相等处的点组成的集合的概率为零.
2. 分布在离散场合可以是分布列或分布函数, 这时称为离散分布; 在连续场合可以是密度函数或分布函数, 这时称为连续分布. 常用的是这两类分布, 但还存在既非离散又非连续的分布.
3. 分布函数需满足单调性、有界性、右连续

Theorem 1.2.2 设随机变量 X 的分布函数为 $F(x)$, 则可用 $F(x)$ 表示下列概率:

1. $P(X \leq a) = F(a)$

2. $P(X < a) = F(a - 0)$
3. $P(X > a) = 1 - P(X \leq a) = 1 - F(a)$
4. $P(X = a) = P(X \leq a) - P(X < a) = F(a) - F(a - 0)$
5. $P(X \geq a) = 1 - P(X < a) = 1 - F(a - 0)$
6. $P(|X| < a) = P(-a < X < a) = P(X < a) - P(X \leq -a) = F(a - 0) - F(-a)$

Definition 1.2.8 — 数学期望. 设随机变量 X 的分布用分布列 $p(x_i)$ 或用密度函数 $p(x)$ 表示, 若

$$\begin{cases} \sum_i |x_i| p(x_i) < +\infty, & \text{当 } X \text{ 为离散随机变量,} \\ \int_{-\infty}^{+\infty} |x| p(x) dx < +\infty, & \text{当 } X \text{ 为连续随机变量} \end{cases}$$

则称 $E(X) = \begin{cases} \sum_i x_i p(x_i), & \text{当 } X \text{ 为离散随机变量,} \\ \int_{-\infty}^{+\infty} x p(x) dx, & \text{当 } X \text{ 为连续随机变量} \end{cases}$ 为 X 的数学期望, 简称期望或均值, 且称 X 的数学期望存在. 否则称数学期望不存在. 数学期望是由分布决定的, 它是分布的位置特征. 如果两个随机变量同分布, 则其数学期望(存在的话)是相等的. 假如把概率看作质量、分布看作某物体的质母分布, 那么数学期望就是该物体的重心位置.

(R) 数学期望要求级数绝对收敛, 这也就是不管次序怎么样, 都是收敛的

Proposition 1.2.3 — 数学期望的性质. 以下所涉及的数学期望均假定其存在.

1. X 的某一函数 $g(X)$ 的数学期望为

$$E[g(X)] = \begin{cases} \sum_i g(x_i) p(x_i), & \text{在离散场合} \\ \int_{-\infty}^{+\infty} g(x) p(x) dx, & \text{在连续场合} \end{cases}$$

2. 若 c 是常数, 则 $E(c) = c$
3. 对任意常数 a , 有 $E(aX) = aE(X)$
4. 对任意的两个函数 $g_1(x)$ 和 $g_2(x)$, 有

$$E[g_1(X) \pm g_2(X)] = E[g_1(X)] \pm E[g_2(X)]$$

Definition 1.2.9 — 方差. (随机变量 X 对其期望 $E(X)$ 的偏差平方的数学期望(设其存在))

$$\text{Var}(X) = E[X - E(X)]^2$$

称为 X 的方差, 方差的正平方根 $\sigma(X) = \sigma_x = \sqrt{\text{Var}(X)}$ 称为 X 的标准差. 方差是由分布决定的, 它是分布的散布特征, 方差愈大, 分布愈分散; 方差愈小, 分布愈集中. 标准差与方差的功能相似, 只是量纲不同。

(R) 如果随机变量 X 的数学期望存在, 其方差不一定存在; 而 X 的方差存在时, 则 $E(X)$ 必定存在, 其原因在于 $|x| \leq x^2 + 1$ 总是成立的

Proposition 1.2.4 — 方差的性质. 以下所涉及的方差均假定其存在.

1. $\text{Var}(X) = E(X^2) - [E(X)]^2$
2. 若 c 是常数, 则 $\text{Var}(c) = 0$
3. 若 a, b 是常数, 则 $\text{Var}(aX + b) = a^2 \text{Var}(X)$
4. 若随机变量 X 的方差存在, 则 $\text{Var}(X) = 0$ 的充要条件是 X 几乎处处为某个常数 a , 即 $P(X = a) = 1$

5. 若 $c \neq E\xi$, 则 $D\xi < E(\xi - c)^2$

Proof. 最后一条的证明: 充分性是显然的, 下面证必要性. 设 $\text{Var}(X) = 0$, 这时 $E(X)$ 存在. 因为

$$\{|X - E(X)| > 0\} = \bigcup_{n=1}^{\infty} \left\{ |X - E(X)| \geq \frac{1}{n} \right\}$$

所以有

$$\begin{aligned} P(|X - E(X)| > 0) &= P\left(\bigcup_{n=1}^{\infty} \left\{ |X - E(X)| \geq \frac{1}{n} \right\}\right) \\ &\leq \sum_{n=1}^{\infty} P\left(|X - E(X)| \geq \frac{1}{n}\right) \\ &\leq \sum_{n=1}^{\infty} \frac{\text{Var}(X)}{(1/n)^2} = 0 \end{aligned}$$

其中最后一个不等式用到了切比雪夫不等式. 由此可知

$$P(|X - E(X)| > 0) = 0$$

因而有

$$P(|X - E(X)| = 0) = 1$$

即

$$P(X = E(X)) = 1$$

这就证明了结论, 且其中的常数 a 就是 $E(X)$

■

Theorem 1.2.5 — 切比雪夫不等式. 设 X 的数学期望和方差都存在, 则对任意常数 $\varepsilon > 0$, 有

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$$

或

$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{\text{Var}(X)}{\varepsilon^2}$$

切比雪夫不等式给出随机变量取值的大偏差(指事件 $\{|X - E(X)| \geq \varepsilon\}$)发生概率的上限, 该上限与分布的方差成正比

Proof. 若 $F(x)$ 是 ξ 的分布函数, 则显然有

$$\begin{aligned} D\xi &= \int_{-\infty}^{\infty} (x - E\xi)^2 dF(x) \\ &\geq \int_{|x-E\xi| \geq \varepsilon} (x - E\xi)^2 dF(x) \geq \int_{|x-E\xi| \geq \varepsilon} \varepsilon^2 dF(x) \\ &= \varepsilon^2 P\{|x - E\xi| \geq \varepsilon\} \end{aligned}$$

这就证得了不等式 (4.2.10)。有时把 (4.2.10) 改写成

$$P\{|\xi - E\xi| < \varepsilon\} \geq 1 - \frac{D\xi}{\varepsilon^2}$$

或

$$P\left\{\left|\frac{\xi - E\xi}{\sqrt{D\xi}}\right| \geq \delta\right\} \leq \frac{1}{\delta^2}$$

■

R 切比雪夫不等式表明：无论分布， ξ 落在 $(E\xi - \sigma\delta, E\xi + \sigma\delta)$ 中的概率均不小于 $1 - \frac{1}{\delta^2}$

Theorem 1.2.6 — 随机变量的标准化. 对任意随机变量 X , 如果 X 的数学期望和方差存在, 则称

$$X^* = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

为 X 的标准化随机变量, 此时有 $E(X^*) = 0, \text{Var}(X^*) = 1$

1.2.1 常用离散分布

Definition 1.2.10 — 退化分布. 若随机变量 α 只取常数值 c , 即

$$P\{\alpha = c\} = 1$$

这时分布函数为

$$I_c(x) = \begin{cases} 0, & x \leq c \\ 1, & x > c \end{cases}$$

这样的 α 并不随机, 但有时我们宁愿把它看作随机变量的退化情况更为方便, 因此称之为退化分布, 又称单点分布。

Definition 1.2.11 — 二项分布. 1. 若 X 的概率分布列为

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n$$

则称 X 服从二项分布, 记为 $X \sim b(n, p)$, 其中 $0 < p < 1$

2. 背景: n 重伯努利试验中成功的次数 X 服从二项分布 $b(n, p)$, 其中 p 为一次伯努利试验中成功发生的概率。
3. $n = 1$ 时的二项分布 $b(1, p)$ 称为二点分布, 或称 **0-1** 分布。因为当 $X \sim b(1, p)$ 时, X 可表示一次伯努利试验中成功的次数, 它只能取值 0 与 1
4. 二项分布 $b(n, p)$ 的数学期望和方差分别是 $E(X) = np, \text{Var}(X) = np(1-p)$
5. 若 $X \sim b(n, p)$, 则 $Y = n - X \sim b(n, 1-p)$, 其中 $Y = n - X$ 是 n 重伯努利试验中失败的次数。

Proof. 期望方差的推导：设随机变量 $X \sim b(n, p)$, 则

$$\begin{aligned} E(X) &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np[p + (1-p)]^{n-1} = np \end{aligned}$$

$$\begin{aligned} E(X^2) &= \sum_{k=0}^n k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n (k-1+1)k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=2}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} + np \\ &= n(n-1)p^2 \sum_{k=2}^n \binom{n-2}{k-2} p^{k-2} (1-p)^{(n-2)-(k-2)} + np \\ &= n(n-1)p^2 + np \end{aligned}$$

■

Definition 1.2.12 — 泊松分布. 若 X 的概率分布列为

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

则称 X 服从泊松分布, 记为 $X \sim P(\lambda)$, 其中参数 $\lambda > 0$

背景: 单位时间 (或单位面积, 单位产品等) 上某稀有事件 (这里稀有事件是指不经常发生的事件) 发生的次数常服从泊松分布 $P(\lambda)$, 其中 λ 为该稀有事件发生的强度。

Theorem 1.2.7 泊松分布 $P(\lambda)$ 的数学期望和方差分别是

$$E(X) = \lambda, \quad \text{Var}(X) = \lambda$$

Proof. 设随机变量 $X \sim P(\lambda)$, 则

$$E(X) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\ &= \sum_{k=1}^{\infty} [(k-1)+1] \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda^2 + \lambda \end{aligned}$$

■

Theorem 1.2.8 — 二项分布的泊松近似 (泊松定理) . 在 n 重伯努利试验中, 记事件 A 在一次试验中发生的概率为 p_n (与试验次数 n 有关), 如果当 $n \rightarrow +\infty$ 时, 有 $np_n \rightarrow \lambda$, 则

$$\lim_{n \rightarrow +\infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Proof. 证明 记 $np_n = \lambda_n$, 即 $p_n = \lambda_n/n$, 我们可得

$$\begin{aligned} \binom{n}{k} p_n^k (1 - p_n)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\ &= \frac{\lambda_n^k}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \end{aligned}$$

对固定的 k 有

$$\begin{aligned} \lim_{n \rightarrow \infty} \lambda_n &= \lambda \\ \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda_n}{n}\right)^{n-k} &= e^{-\lambda} \end{aligned}$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) = 1$$

从而

$$\lim_{n \rightarrow \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

■

Definition 1.2.13 — 超几何分布. 1. 若 X 的概率分布列为

$$P(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, \dots, r$$

则称 X 服从超几何分布, 记为 $X \sim h(n, N, M)$, 其中 $r = \min\{M, n\}$, 且 $M \leq N$ $n \leq N$. n, N, M 均为正整数.

2. 背景: 设有 N 个产品, 其中有 M 个不合格品. 若从中不放回地随机抽取 n 个, 则其中含有的不合格品的个数 X 服从超几何分布 $h(n, N, M)$
3. 超几何分布 $h(n, N, M)$ 的数学期望和方差分别是

$$E(X) = n \frac{M}{N}, \quad \text{Var}(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)}$$

4. 超几何分布的二项近似当 $n \ll N$ 时, 超几何分布 $h(n, N, M)$ 可用二项分布 $b(n, M/N)$ 近似, 即

$$\frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} \approx \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{其中 } p = \frac{M}{N}$$

5. 实际应用中, 在不返回抽样时, 常用超几何分布描述抽出样品中不合格品数的分布; 在返回抽样时, 常用二项分布 $b(n, p)$ 描述抽出样品中不合格品数的分布; 当批量 N 较大, 而抽出样品数 n 较小时, 不返回抽样可近似看作返回抽样。

Proof. 期望方差的证明: $X \sim h(n, N, M)$, 则 X 的数学期望为

$$E(X) = \sum_{k=0}^r k \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = n \frac{M}{N} \sum_{k=1}^r \frac{\binom{M-1}{k-1} \binom{N-M}{n-k}}{\binom{N-1}{n-1}} = n \frac{M}{N}$$

方差

$$\begin{aligned} E(X^2) &= \sum_{k=1}^r k^2 \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = \sum_{k=2}^r k(k-1) \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} + n \frac{M}{N} \\ &= \frac{M(M-1)}{\binom{N}{n}} \sum_{k=2}^r \binom{M-2}{k-2} \binom{N-M}{n-k} + n \frac{M}{N} \\ &= \frac{M(M-1)}{\binom{N}{n}} \binom{N-2}{n-2} + n \frac{M}{N} = \frac{M(M-1)n(n-1)}{N(N-1)} + n \frac{M}{N} \end{aligned}$$

■

Definition 1.2.14 — 几何分布. 1. 若 X 的概率分布列为

$$P(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots$$

则称 X 服从几何分布, 记为 $X \sim Ge(p)$, 其中 $0 < p < 1$

2. 背景: 在伯努利试验序列中, 成功事件 A 首次出现时的试验次数 X 服从几何分布 $Ge(p)$, 其中 p 为每次试验中事件 A 发生的概率。
3. 几何分布 $Ge(p)$ 的数学期望和方差分别显

$$E(X) = \frac{1}{p}, \quad \text{Var}(X) = \frac{1-p}{p^2}$$

4. 几何分布的无记忆性: 若 $X \sim Ge(p)$, 则对任意正整数 m 与 n 有

$$P(X > m+n | X > m) = P(X > n)$$

Proof. 数学期望为

$$\begin{aligned}
 E(X) &= \sum_{k=1}^{\infty} kpq^{k-1} = p \sum_{k=1}^{\infty} kq^{k-1} = p \sum_{k=1}^{\infty} \frac{dq^k}{dq} \\
 &= p \frac{d}{dq} \left(\sum_{k=0}^{\infty} q^k \right) = p \frac{d}{dq} \left(\frac{1}{1-q} \right) = \frac{p}{(1-q)^2} = \frac{1}{p} \\
 E(X^2) &= \sum_{k=1}^{\infty} k^2 pq^{k-1} = p \left[\sum_{k=1}^{\infty} k(k-1)q^{k-1} + \sum_{k=1}^{\infty} kq^{k-1} \right] \\
 &= pq \sum_{k=1}^{\infty} k(k-1)q^{k-2} + \frac{1}{p} = pq \sum_{k=1}^{\infty} \frac{d^2}{dq^2} q^k + \frac{1}{p} \\
 &= pq \frac{d^2}{dq^2} \left(\sum_{k=0}^{\infty} q^k \right) + \frac{1}{p} = pq \frac{d^2}{dq^2} \left(\frac{1}{1-q} \right) + \frac{1}{p} \\
 &= pq \frac{2}{(1-q)^3} + \frac{1}{p} = \frac{2q}{p^2} + \frac{1}{p}
 \end{aligned}$$

■

Proof. 无记忆性的证明：因为

$$P(X > n) = \sum_{k=n+1}^{\infty} (1-p)^{k-1} p = \frac{p(1-p)^n}{1-(1-p)} = (1-p)^n$$

所以对任意的正整数 m 与 n , 条件概率

$$\begin{aligned}
 P(X > m+n | X > m) &= \frac{P(X > m+n)}{P(X > m)} = \frac{(1-p)^{m+n}}{(1-p)^m} \\
 &= (1-p)^n = P(X > n)
 \end{aligned}$$

■

Theorem 1.2.9 若 η 是取正整数值的随机变量, 并且, 在已知 $\eta > k$ 的条件下, $\eta = k+1$ 的概率与 k 无关, 那么 η 服从几何分布.

Definition 1.2.15 — 帕斯卡分布. 在成功的概率为 p 的伯努利试验中, 若以 ζ 记第 r 次成功出现时的试验次数, 则 ζ 是随机变量, 取值 $r, r+1, \dots$ 其概率分布为帕斯卡分布:

$$P\{\zeta = k\} = \binom{k-1}{r-1} p^r q^{k-r}, \quad k = r, r+1, \dots$$

Definition 1.2.16 — 负二项分布. 对于任意实数 $r > 0$, 称

$$Nb(l; r, p) = \binom{-r}{l} p^r (-q)^l, \quad l = 0, 1, 2, \dots$$

为负二项分布。

Definition 1.2.17 — 负二项分布. 1. 若 X 的概率分布列为

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \dots$$

则称 X 服从负二项分布或巴斯卡分布, 记为 $X \sim Nb(r, p)$, 其中 r 为正整数 $0 < p < 1$

2. 背景: 在伯努利试验序列中, 成功事件 A 第 r 次出现时的试验次数 X 服从负二项分

- 布 $Nb(r, p)$, 其中 p 为每次试验中事件 A 发生的概率.
 3. $r=1$ 时的负二项分布为几何分布, 即 $Nb(1, p) = Ge(p)$
 4. 负二项分布 $Nb(r, p)$ 的数学期望和方差分别是

$$E(X) = r/p, \quad \text{Var}(X) = r(1-p)/p^2$$

5. 负二项分布的随机变量可以表示成 r 个独立同分布的几何分布随机变量之和, 即若 $X \sim Nb(r, p)$, 则 $X = X_1 + X_2 + \dots + X_r$, 其中 X_1, X_2, \dots, X_r 是相互独立、服从几何分布 $Ge(p)$ 的随机变量. 下图“1”表示 A , “0”表示 \bar{A} :

$$\underbrace{00 \dots 1}_{X_1} \underbrace{0.0 \dots 1}_{X_2} \dots \underbrace{000001}_{X_r} \underbrace{X}_X$$

这里随机变量间的相互独立是指一个变量的取值不影响其他变量的取值

1.2.2 常用连续分布

Theorem 1.2.10 连续型随机变量取个别值的概率为 0, 这与离散型随机变量截然不同. 此外, 上, 一个事件的概率等于 0, 这事件并不一定是不可能事件; 同样地, 一个事件的概率等于 1, 这事件也不一定是必然事件。

Definition 1.2.18 — 正态分布. 高斯用最大似然法导出正态分布是对概率论的重大贡献. 推导利用柯西方程。

1. 若 X 的密度函数和分布函数分别为

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty$$

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad -\infty < x < +\infty$$

则称 X 服从正态分布, 记作 $X \sim N(\mu, \sigma^2)$, 其中参数 $-\infty < \mu < +\infty, \sigma > 0$

2. 背景: 一个变量若是由大量微小的、独立的随机因素的叠加结果, 则此变量一定是正态变量(服从正态分布的变量). 测量误差就是由量具偏差、测量环境的影响、测量技术的影响、测量人员的心理影响等等随机因素叠加而成的, 所以测量误差常认为服从正态分布。
 3. 关于参数 μ : μ 是正态分布的数学期望, 即

$$E(X) = \mu$$

称 μ 为正态分布的位置参数. μ 是正态分布的对称中心, 在 μ 的左侧和 $p(x)$ 下的面积为 0.5; 在 μ 的右侧和 $p(x)$ 下的面积也为 0.5, 所以 μ 也是正态分布的中位数。若 $X \sim N(\mu, \sigma^2)$, 则 X 在离 μ 愈近取值的可能性愈大, 离 μ 愈远取值的可能性愈小。

4. 关于参数 σ σ^2 是正态分布的方差, 即

$$\text{Var}(X) = \sigma^2$$

σ 是正态分布的标准差, σ 愈小, 正态分布愈集中; σ 愈大, 正态分布愈分散. σ 又称为正态分布的尺度参数。若 $X \sim N(\mu, \sigma^2)$, 则其密度函数 $p(x)$ 在 $\mu \pm \sigma$ 处有两个拐点。

5. 称 $\mu = 0, \sigma = 1$ 时的正态分布 $N(0, 1)$ 为标准正态分布. 记 U 为标准正态变量, $\varphi(u)$ 和 $\Phi(u)$ 为标准正态分布的密度函数和分布函数. $\varphi(u)$ 和 $\Phi(u)$ 满足:
- $\varphi(-u) = \varphi(u)$
 - $\Phi(-u) = 1 - \Phi(u)$
- 对 $u > 0, \Phi(u)$ 的值有表可查.
6. 标准化变换: 若 $X \sim N(\mu, \sigma^2)$, 则 $U = (X - \mu)/\sigma \sim N(0, 1)$, 其中 $U = (X - \mu)/\sigma$ 称为 X 的标准化变换.
- 7.

$$E(U) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ue^{-\frac{u^2}{2}} du$$

被积函数是个奇函数

8.

$$\begin{aligned} \text{Var}(U) &= E(U^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}} du \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u d \left(-e^{-\frac{u^2}{2}} \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(-ue^{-\frac{u^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du \right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1 \end{aligned}$$

只算标准的, 其他的通过期望和方差的性质变换

9. 若 $X \sim N(\mu, \sigma^2)$, 则对任意实数 a 与 b , 有

$$\begin{aligned} P(X \leq b) &= \Phi\left(\frac{b-\mu}{\sigma}\right) \\ P(a < X) &= 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \\ P(a < X \leq b) &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

可见, 涉及正态变量的概率计算, 一般是化为标准正态变量的查表获得。

10. 正态分布的 3σ 原则: 设 $X \sim N(\mu, \sigma^2)$, 则

$$P(|X - \mu| < k\sigma) = \Phi(k) - \Phi(-k) = \begin{cases} 0.6826, & k = 1 \\ 0.9545, & k = 2 \\ 0.9973, & k = 3 \end{cases}$$

Definition 1.2.19 — 均匀分布. 1. 若 X 的密度函数和分布函数分别为

$$p(x) = \begin{cases} \frac{1}{b-a}, & a < x < b, \\ 0, & \text{其他} \end{cases} \quad F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases}$$

则称 X 服从区间 (a, b) 上的均匀分布, 记作 $X \sim U(a, b)$

2. 背景: 向区间 (a, b) 随机投点, 落点坐标 X 一定服从均匀分布 $U(a, b)$ 这里“随机投点”是指: 点落在任意相等长度的小区间上的可能性是相等的。

3. 均匀分布 $U(a, b)$ 的数学期望和方差分别是

$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

4. 称区间 $(0, 1)$ 上的均匀分布 $U(0, 1)$ 为标准均匀分布, 它是导出其他分布随机数的桥梁.

Proof.

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

$$E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$

■

Definition 1.2.20 — 指数分布. 1. 若 X 的密度函数和分布函数分别为

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0 \end{cases} \quad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

则称 X 服从指数分布, 记作 $X \sim Exp(\lambda)$, 其中参数 $\lambda > 0$

2. 背景: 若一个元器件 (或一台设备、或一个系统) 遇到外来冲击时即告失效, 则首次冲击来到的时间 X (寿命) 服从指数分布. 很多产品的寿命可认为服从或近似服从指数分布。
3. 指数分布 $Exp(\lambda)$ 的数学期望和方差分别为

$$E(X) = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

4. 指数分布的无记忆性: 若 $X \sim Exp(\lambda)$, 则对任意 $s > 0, t > 0$, 有

$$P(X > s+t | X > s) = P(X > t)$$

Proof. 设随机变量 $X \sim Exp(\lambda)$, 则

$$\begin{aligned} E(X) &= \int_0^\infty x \lambda e^{-\lambda x} dx = \int_0^\infty x d(-e^{-\lambda x}) \\ &= -xe^{-\lambda x} \Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^\infty = \frac{1}{\lambda} \end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \int_0^\infty x^2 d(-e^{-\lambda x}) \\ &= -x^2 e^{-\lambda x} \Big|_0^\infty + 2 \int_0^\infty x e^{-\lambda x} dx = \frac{2}{\lambda^2} \end{aligned}$$

■

Proof. 证明无记忆性 因为 $X \sim Exp(\lambda)$, 所以 $P(X > s) = e^{-\lambda s}, s > 0$. 又因为 $\{X > s+t\} \subseteq \{X > s\}$ 于是条件概率

$$P(X > s+t | X > s) = \frac{P(X > s+t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

■

Theorem 1.2.11 指数分布是唯一具有性质

$$P\{\xi \geq s+t | \xi \geq s\} = P\{\xi \geq t\}$$

的连续型分布.

■ **Example 1.2 — 泊松分布与指数分布的关系.** 如果某设备在任何长为 t 的时间 $[0, t]$ 内发生故障的次数 $N(t)$ 服从参数为 λt 的泊松分布, 则相继两次故障之间的时间间隔 T 服从参数为 λ 的指数分布。

Proof. 解 $N(t) \sim P(\lambda t)$, 即

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, \dots$$

注意到两次故障之间的时间间隔 T 是非负随机变量, 且事件 $\{T \geq t\}$ 说明此设备在 $[0, t]$ 内没有发生故障, 即 $\{T \geq t\} = \{N(t) = 0\}$, 由此我们得

当 $t < 0$ 时, 有 $F_T(t) = P(T \leq t) = 0$;

当 $t \geq 0$ 时, 有

$$F_T(t) = P(T \leq t) = 1 - P(T > t) = 1 - P(N(t) = 0) = 1 - e^{-\lambda t}$$

所以 $T \sim Exp(\lambda)$, 即相继两次故障之间的时间间隔 T 服从参数为 λ 的指数分布 ■

Theorem 1.2.12 — 埃尔朗分布. 若 $\xi(t)$ 是参数为 λt 的泊松过程, 以 W_r 记它的第 r 个跳跃发生的时刻. 事件 $\{W_r < t\}$ 发生表明第 r 个跳既出现在时刻 t 之前, 因此事件 $\{\xi(t) \geq r\}$ 发生, 即 $\{W_r < t\} \subset \{\xi(t) \geq r\}$. 反之, 若事件 $\{\xi(t) \geq r\}$ 发生, 即在时刻 t 时 $\xi(t)$ 之值不小于 r , 这时第 r 个跳跃已经出现过, 因此事件 $\{W_r < t\}$ 发生, 即有 $\{\xi(t) \geq r\} \subset \{W_r < t\}$. 综上所述可知

$$\{W_r < t\} = \{\xi(t) \geq r\}$$

对任意的正整数 r 及实数 $\lambda > 0$

$$p(x) = \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x}, \quad x \geq 0$$

是一个密度函数, 称为埃尔朗分布, 它是丹麦科学家埃尔朗 (Erlang) 在研究电话问题时引进的, 这些研究开包了排队论这一学科. 前面的推导说明, 泊松过程中第 r 个跳跃发生的时刻 W_r 埃尔朗分布。

Theorem 1.2.13 当 $r = 1$ 时, 埃尔朗分布化作指数分布. 另外, 若记

$$\begin{aligned} \tau_1 &= W_1 \\ \tau_r &= W_r - W_{r-1}, \quad r = 2, 3, \dots \end{aligned}$$

则 τ_r 表示泊松过程的第 r 个跳跃间隔, 用它们可以给出跳跃时刻 W_r 的如下表达式

$$W_r = \tau_1 + \tau_2 + \dots + \tau_r$$

可以证明, $\tau_1, \tau_2, \dots, \tau_r$ 服从参数为 λ 的指数分布, 且相互独立.

Definition 1.2.21 — 伽玛分布. 1. 伽玛函数 称 $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$ 为伽玛函数, 其中参数 $\alpha > 0$. 伽玛函数具有如下性质:

- (a) $\Gamma(1) = 1$
- (b) $\Gamma(1/2) = \sqrt{\pi}$
- (c) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$
- (d) $\Gamma(n+1) = n\Gamma(n) = n!$ (n 为自然数)

2. 伽玛分布: 若 X 的密度函数为

$$p(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

则称 X 服从伽玛分布, 记作 $X \sim Ga(\alpha, \lambda)$, 其中 $\alpha > 0$ 为形状参数, $\lambda > 0$ 为尺度参数

- 3. 背景: 若一个元器件 (或一台设备, 或一个系统) 能抵挡一些外来冲击, 但遇到第 k 次冲击时即告失效, 则第 k 次冲击来到的时间 X (寿命) 服从形状参数为 k 的伽玛分布 $Ga(k, \lambda)$
- 4. 伽玛分布 $Ga(\alpha, \lambda)$ 的数学期望和方差分别为

$$E(X) = \frac{\alpha}{\lambda}, \quad \text{Var}(X) = \frac{\alpha}{\lambda^2}$$

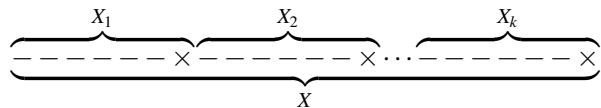
5. 伽玛分布的两个特例:

- (a) $\alpha = 1$ 时的伽玛分布就是指数分布, 即 $Ga(1, \lambda) = Exp(\lambda)$
- (b) 称 $\alpha = n/2, \lambda = 1/2$ 时的伽玛分布为自由度为 n 的 χ^2 (卡方) 分布, 记为 $\chi^2(n)$, 其密度函数为

$$p(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$\chi^2(n)$ 分布的期望和方差分别为 $E(X) = n, \text{Var}(X) = 2n$

- 6. 若形状参数为整数 k , 则伽玛变量可以表示成 k 个独立同分布的指数变量之和, 即若 $X \sim Ga(k, \lambda)$, 则 $X = X_1 + X_2 + \dots + X_k$, 其中 X_1, X_2, \dots, X_k 是相互独立且都服从指数分布 $Exp(\lambda)$ 的随机变量, 其中 \times 表示冲击来到的时间.



Proof. 期望方差的证明:

$$\begin{aligned} E(X) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x} dx = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \frac{1}{\lambda} = \frac{\alpha}{\lambda} \\ E(X^2) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1} e^{-\lambda x} dx = \frac{\Gamma(\alpha+2)}{\lambda^2 \Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\lambda^2} \end{aligned}$$

■

Definition 1.2.22 — 贝塔分布. 1. 称 $B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$ 为贝塔函数, 其中参数 $a > 0, b > 0$. 贝塔函数具有如下性质:

- (a) $B(a, b) = B(b, a)$

- (b) $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$
 2. 贝塔分布 若 X 的密度函数为

$$p(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, & 0 < x < 1 \\ 0, & \text{其他} \end{cases}$$

- 则称 X 服从贝塔分布, 记作 $X \sim Be(a, b)$, 其中 $a > 0, b > 0$ 都是形状参数
 3. 背景: 很多比率, 如产品的不合格品率, 机器的维修率、某商品的市场占有率为射击的命中率等都是在区间 $(0, 1)$ 上取值的随机变量, 贝塔分布 $Be(a, b)$ 可供描述这些随机变量之用, 而在应用中, 可调节 a 与 b 以适应实际中的要求
 4. 贝塔分布 $Be(a, b)$ 的数学期望和方差分别为

$$E(X) = \frac{a}{a+b}, \quad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

5. $a = b = 1$ 时的贝塔分布就是区间 $(0, 1)$ 上的均匀分布, 即 $Be(1, 1) = U(0, 1)$

Proof. 证明 Beta 分布和 Gamma 分布之间的关系

$$\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty x^{a-1} y^{b-1} e^{-(x+y)} dx dy$$

作变量变换 $x = uv, y = u(1-v)$, 其雅可比行列式 $J = -u$. 故

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^\infty \int_0^1 (uv)^{a-1} [u(1-v)]^{b-1} e^{-u} u du dv \\ &= \int_0^\infty u^{a+b-1} e^{-u} du \int_0^1 v^{a-1} (1-v)^{b-1} dv \\ &= \Gamma(a+b) B(a, b) \end{aligned}$$

■

Proof. 证明期望:

$$\begin{aligned} E(X) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^a (1-x)^{b-1} dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b} \end{aligned}$$

方差

$$\begin{aligned} E(X^2) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+1} (1-x)^{b-1} dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\ &= \frac{a(a+1)}{(a+b)(a+b+1)} \end{aligned}$$

■

Theorem 1.2.14 对数正态分布: $LN(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, x > 0 \quad (1.1)$$

期望

$$e^{\mu + \sigma^2/2}$$

方差

$$e^{2\mu + \sigma^2} (e^{\sigma^2} - 1)$$

Theorem 1.2.15 韦布尔分布

$$p(x) = F'(x), F(x) = 1 - \exp \left\{ - \left(\frac{x}{\eta} \right)^m \right\}, x > 0$$

期望

$$\eta \Gamma(1 + \frac{1}{m})$$

方差

$$\eta^2 \left[\Gamma \left(1 + \frac{2}{m} \right) - \Gamma^2 \left(1 + \frac{1}{m} \right) \right]$$

Theorem 1.2.16 — 伯努利过程与泊松过程. 若每隔 Δt 进行一次试验，则伯努利试验也可以看作一个随时间而变化的过程。在伯努利试验中，到时刻 $n\Delta t$ 为止，共进行 n 次试验，这时成功次数服从二项分布。而在泊松过程中，到时刻 t 的来到数则服从泊松分布。为等待第一次成功，伯努利试验中的等待时间服从几何分布而泊松过程中则服从指数分布。它们都有无记忆性。为等待第 r 次成功，伯努利试验中的等待时间服从帕斯卡分布；而泊松过程中则服从埃尔朗分布。

Theorem 1.2.17 关于单调函数的一般结果相同，不难推出分布函数具有如下性质：

1. 分布函数至多只有可列个不连续点
2. 对分布函数 $F(x)$ 有勒贝格分解

$$F(x) = c_1 F_1(x) + c_2 F_2(x) + c_3 F_3(x)$$

其中 $F_1(x)$ 是跳跃函数， $F_2(x)$ 是绝对连续函数， $F_3(x)$ 是所谓奇异函数，它们都是分布函数；而 $0 \leq c_i \leq 1, i = 1, 2, 3$ ，且 $c_1 + c_2 + c_3 = 1$

离散型是第一个为 1，连续性是第二个为 1

1.2.3 随机变量函数的分布

Definition 1.2.23 定义 3.3.1 设 $y = g(x)$ 是 \mathbf{R}^1 到 \mathbf{R}^1 上的一个映照，若对于一切 \mathbf{R}^1 中的博雷尔点集 B_1 均有

$$\{x : g(x) \in B_1\} \in \mathcal{B}_1$$

其中 \mathcal{B}_1 为 R^1 上博雷尔 σ 域，则称 $g(x)$ 是一元博雷尔（可测）函数。

Definition 1.2.24 设 $y = g(x_1, \dots, x_n)$ 是 \mathbf{R}^n 到 \mathbf{R}^1 上的一个映照若对一切 R^1 中的博雷尔点集 B_1 均有

$$\{(x_1, \dots, x_n) : g(x_1, \dots, x_n) \in B_1\} \in \mathcal{B}_n$$

其中 \mathcal{B}_n 为 \mathbf{R}^n 上博雷尔 σ 域, 则称 $g(x_1, \dots, x_n)$ 为 n 元博雷尔 (可测) 函数。

Definition 1.2.25 若 (ξ_1, \dots, ξ_n) 是 (Ω, \mathcal{F}, P) 上的随机向量, 而 $g(x_1, \dots, x_n)$ 是 n 元博雷尔函数, 则 $g(\xi_1, \dots, \xi_n)$ 是 (Ω, \mathcal{F}, P) 上的随机变量。

Theorem 1.2.18 设连续随机变量 X 的密度函数为 $p_X(x), Y = g(X)$

- 若 $y = g(x)$ 严格单调, 其反函数 $h(y)$ 有连续导函数, 则 $Y = g(X)$ 的密度函数为

$$p_Y(y) = \begin{cases} p_X[h(y)] |h'(y)|, & a < y < b \\ 0, & \text{其他} \end{cases}$$

其中 $a = \min\{g(-\infty), g(+\infty)\}, b = \max\{g(-\infty), g(+\infty)\}$

- 若 $y = g(x)$ 在不相重叠的区间 I_1, I_2, \dots 上逐段严格单调, 其反函数 $h_1(y), h_2(y), \dots$ 有连续导函数, 则 $Y = g(X)$ 的密度函数为

$$p_Y(y) = \sum_i p_X(h_i(y)) |h'_i(y)|$$

Theorem 1.2.19 正态变量的线性变换仍为正态变量: 若 X 服从正态分布 $N(\mu, \sigma^2)$, 则当 $a \neq 0$ 时, 有

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$$

■ **Example 1.3** 若 θ 服从 $[-\frac{\pi}{2}, \frac{\pi}{2}]$ 的均匀分布, $\psi = \tan \theta$, 试求 ψ 的密度函数 $q(y)$

Proof. 记 $y = \tan x$, 则 $x = \tan^{-1} y, \frac{dx}{dy} = \frac{1}{1+y^2}$, 因此由 (3.3.12) 知

$$q(y) = \frac{1}{\pi} \cdot \frac{1}{1+y^2}, -\infty < y < \infty$$

该分布称为柯西分布, 它没有期望 ■

■ **Example 1.4** 若 $\zeta \sim N(0, 1)$, 求 $\eta = \zeta^2$ 的密度函数. [解] 当 $y \leq 0$ 时, $G(y) = P\{\eta < y\} = 0$, 显然, 此时 $q(y) = 0$ 当 $y > 0$

$$\begin{aligned} G(y) &= P\{\eta < y\} = P\{\zeta^2 < y\} = P\{-\sqrt{y} < \zeta < \sqrt{y}\} \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 2 \int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \end{aligned}$$

因此 $\eta = \zeta^2$ 的密度函数为

$$q(y) = \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}}, \quad y > 0$$

■ **Example 1.5** 若 $\xi \sim N(\mu, \sigma^2)$, 求 $\eta = e^\xi$ 的密度函数. [解] 当 $y > 0$ 时

$$\begin{aligned} P\{\eta < y\} &= P\{e^\xi < y\} = P\{\xi < \ln y\} \\ &= \int_{-\infty}^{\ln y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \end{aligned}$$

所以, η 的密度函数为

$$q(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}}, \quad y > 0$$

的对数即 $\ln \eta = \xi$ 服从正态分布, 故称 η 所服从的分布为对数正态分布。

Definition 1.2.26 — 对数正态分布. 若 X 的密度函数为

$$p_x(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

则称 X 服从对数正态分布, 记为 $X \sim LN(\mu, \sigma^2)$, 其中 $-\infty < \mu < +\infty, \sigma > 0$.

Theorem 1.2.20 若 $X \sim LN(\mu, \sigma^2)$, 则

$$E(X) = e^{\mu + \sigma^2/2}, \text{Var}(X) = (e^{\sigma^2} - 1) \cdot e^{2\mu + \sigma^2}$$

Theorem 1.2.21 若 $X \sim LN(\mu, \sigma^2)$, 则

$$Y = \ln X \sim N(\mu, \sigma^2)$$

Theorem 1.2.22 若 $X \sim Ga(\alpha, \lambda)$, 则当 $k > 0$ 时, 有 $Y = kX \sim Ga(\alpha, \lambda/k)$

Theorem 1.2.23 若 X 的分布函数 $F_X(x)$ 为严格单调增的连续函数, 其反函数 $F_X^{-1}(y)$ 存在, 则 $Y = F_X(X)$ 服从 $(0, 1)$ 上的均匀分布 $U(0, 1)$

Proof. 下求 $Y = F_X(X)$ 的分布函数. 由于分布函数 $F_X(x)$ 仅在 $[0, 1]$ 区间上取值, 故当 $y < 0$ 时, 因为 $\{F_X(X) \leq y\}$ 是不可能事件, 所以

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = 0$$

当 $0 \leq y < 1$ 时, 有

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y \end{aligned}$$

当 $y \geq 1$ 时, 因为 $\{F_X(X) \leq y\}$ 是必然事件,) 所以

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = 1$$

综上所述, $Y = F_X(X)$ 的分布函数为

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ y, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

这正是 $(0, 1)$ 上均匀分布的分布函数, 所以 $Y \sim U(0, 1)$ ■

Theorem 1.2.24 上面的定理另外一种说法：若 $F(x)$ 是左连续的单调不成函数，且 $F(-\infty) = 0, F(+\infty) = 1$ ，则存在一个概率空间 (Ω, \mathcal{F}, P) 及其上的随机变量 $\xi(\omega)$ ，使 $\xi(\omega)$ 的分布函数正好是 $F(x)$

(R) 任一个连续随机变量 X 都可通过其分布函数 $F(x)$ 与均匀分布随机变量 U 发生关系。如 X 服从指数分布 $\text{Exp}(\lambda)$ ，其分布函数为 $F(x) = 1 - e^{-\lambda x}$ ，当 x 换为 X 后，有

$$U = 1 - e^{-\lambda X} \quad \text{或} \quad X = \frac{1}{\lambda} \ln \frac{1}{1-U}$$

后一式表明：由均匀分布 $U(0, 1)$ 的随机数（伪观察值） u_i 可得指数分布 $\text{Exp}(\lambda)$ 的随机数 $x_i = \frac{1}{\lambda} \ln \frac{1}{1-u_i}, i = 1, 2, \dots, n, \dots$ ，而均匀分布随机数在任何一个统计软件都可产生，从而指数分布（继而其他分布）随机数也可获得。而各种分布随机数的获得是进行随机模拟法（又称蒙特卡罗法）的基础。

1.2.4 分布的其他特征数

Definition 1.2.27 — k 阶矩。
 1. 称 $\mu_k = E(X^k)$ 为 X 的 k 阶原点矩。一阶原点矩就是数学期望
 2. 称 $v_k = E(X - E(X))^k$ 为 X 的 k 阶中心矩。二阶中心矩就是方差
 3.

$$v_k = E(X - E(X))^k = E(X - \mu_1)^k = \sum_{i=0}^k \binom{k}{i} \mu_i (-\mu_1)^{k-i}$$

4. 前 k 阶中心矩可用原点矩表示，如

$$\begin{aligned} v_1 &= 0 \\ v_2 &= \mu_2 - \mu_1^2 \\ v_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \\ v_4 &= \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4 \end{aligned}$$

■ **Example 1.6** 设随机变量 $X \sim N(0, \sigma^2)$ ，则

$$\begin{aligned} \mu_k &= E(X^k) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x^k \exp\left\{-\frac{x^2}{2\sigma^2}\right\} dx \\ &= \frac{\sigma^k}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^k \exp\left\{-\frac{u^2}{2}\right\} du \end{aligned}$$

在 k 为奇数时，上述被积函数是奇函数，故

$$\mu_k = 0, \quad k = 1, 3, 5, \dots$$

在 k 为偶数时，上述被积函数是偶函数，再利用变换 $z = u^2/2$ ，可得

$$\begin{aligned} \mu_k &= \sqrt{\frac{2}{\pi}} \sigma^k 2^{(k-1)/2} \int_0^{\infty} z^{(k-1)/2} e^{-z} dz = \sqrt{\frac{2}{\pi}} \sigma^k 2^{(k-1)/2} \Gamma\left(\frac{k+1}{2}\right) \\ &= \sigma^k (k-1)(k-3)\cdots 1, \quad k = 2, 4, 6, \dots \end{aligned}$$

故 $N(0, \sigma^2)$ 分布的前四阶原点矩为

$$\mu_1 = 0, \quad \mu_2 = \sigma^2, \quad \mu_3 = 0, \quad \mu_4 = 3\sigma^4$$

又因为 $E(X) = 0$ ，所以有原点矩等于中心矩，即 $\mu_k = v_k, k = 1, 2, \dots$

Definition 1.2.28 — 变异系数. 称比值 $C_v(X) = \frac{\sqrt{\text{Var}(X)}}{E(X)}$ 为 X 的变异系数. 变异系数是一个无量纲的量。

Definition 1.2.29 — 分位数. 设连续随机变量 X 的分布函数为 $F(x)$, 密度函数为 $p(x)$. 对任意 $p \in (0, 1)$

1. 称满足条件

$$F(x_p) = \int_{-\infty}^{x_p} p(x)dx = p$$

的 x_p 为此分布的 p 分位数, 又称下侧 p 分位数, 它把密度函数下的面积一分为二, 左侧面积恰好为 p

2. 称满足条件

$$1 - F(x'_p) = \int_{x'_p}^{+\infty} p(x)dx = p$$

的 x'_p 为此分布的上侧 p 分位数

3. 分位数与上侧分位数的转换公式: $x'_p = x_{1-p}; x_p = x'_{1-p}$
4. 称 $p = 0.5$ 时的 p 分位数 $x_{0.5}$ 为此分布的中位数, 即 $x_{0.5}$ 满足

$$F(x_{0.5}) = \int_{-\infty}^{x_{0.5}} p(x)dx = 0.5$$

5. 若随机变量 X 的密度函数 $p(x)$ 是偶函数, 则此分布的 p 分位数 x_p 满足: $x_p = -x_{1-p}$. 中位数为分布对称中心
6. 记标准正态分布的 p 分位数为 u_p . 因为标准正态密度函数是偶函数, 所以 $u_p = -u_{1-p}$. 竟如 $u_{0.25} = -u_{0.75} = -0.675$
7. 一般正态分布 $N(\mu, \sigma^2)$ 的 p 分位数 x_p 满足: $x_p = \mu + \sigma \times u_p$. 暨如 $N(10, 2^2)$ 的 0.25 分位数为 $x_{0.25} = 10 + 2u_{0.25} = 8.65$
8. 分布的矩有可能不存在, 但连续分布的分位数总存在. p 分位数 x_p 总是 p 的增函数.

Definition 1.2.30 — 偏度系数. 1. 称比值

$$\beta_s = \frac{E(X - E(X))^3}{[E(X - E(X))^2]^{3/2}}$$

为 X 的分布的偏度系数, 简称偏度

2. 偏度系数刻画的是分布的不对称程度, $|\beta_s|$ 愈大, 分布的对称性愈差
3. 任一对称分布的偏度 $\beta_s = 0$. 当 $\beta_s > 0$ 时, 分布为正偏 (又称右偏); 当 $\beta_s < 0$ 时, 分布为负偏 (又称左偏).

Definition 1.2.31 — 峰度系数. 设随机变量 X 的前四阶矩存在, 则如下比值减去 3

$$\beta_k = \frac{v_4}{v_2^2} - 3 = \frac{E(X - EX)^4}{[\text{Var}(X)]^2} - 3$$

称为 X (或分布) 的峰度系数, 简称峰度。

1. 峰度是描述分布尖峭程度和 (或) 尾部粗细的一个特征数
2. 峰度 β_k 是相对于正态分布而言的超出量, U 为标准正态变量, $E(U^4) = 3$

- 3. $\beta_k > 0$ 表示标准化后的分布比标准正态分布更尖峭和或尾部更粗
- 4. $\beta_k < 0$ 表示标准化后的分布比标准正态分布更平坦和或尾部更细

1.3 多维随机变量及其分布

1.3.1 多维随机变量及其联合分布

Definition 1.3.1 — n 维随机变量. 如果 $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$ 是定义在同一个样本空间 $\Omega = \{\omega\}$ 上的 n 个随机变量, 则称 $X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$ 为 n 维随机变量, 或 n 元随机变量, 或随机向量。

Definition 1.3.2 — 联合分布函数. 对任意的 n 个实数 x_1, x_2, \dots, x_n, n 个事件 $X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n$ 同时发生的概率

$$F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$$

称为 n 维随机变量 (X_1, X_2, \dots, X_n) 的联合分布函数.

Proposition 1.3.1 二维随机变量 (X, Y) 的联合分布函数 $F(x, y) = P(X \leq x, Y \leq y)$ 具有如下四条基本性质:

1. 单调性 $F(x, y)$ 分别对 x 或 y 是单调不减的.
2. 有界性 对任意的 x 和 y , 有 $0 \leq F(x, y) \leq 1$, 且

$$F(-\infty, y) = F(x, -\infty) = 0, \quad F(+\infty, +\infty) = 1$$

3. 右连续性 对每个变量都是右连续的, 即

$$F(x+0, y) = F(x, y), \quad F(x, y+0) = F(x, y)$$

4. 非负性 对任意的 $a < b, c < d$ 有

$$P(a < X \leq b, c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c) \geq 0$$

可以证明: 具有上述四条性质的二元函数 $F(x, y)$ 一定是某个二维随机变量的分布函数. 注意: 事件 " $X \leq x, Y \leq y$ " 常可用平面上的无穷直角区域表示.

Theorem 1.3.2 多元分布函数

1. 单调性: 关于每个变元是单调不减函数
- 2.

$$F(x_1, x_2, \dots, -\infty, \dots, x_n) = 0$$

$$F(+\infty, +\infty, \dots, +\infty) = 1$$

3. 关于每个变元左连续。
4. 在二元场合, 还应该有: 对任意 $a_1 < b_1, a_2 < b_2$, 都有

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0$$

满足 (ii), (iii), (iv) 这三条性质的二元函数是某二维随机变量的分布函数

Definition 1.3.3 — 联合分布列. 如果 (X, Y) 只取有限个或可列个数对 (x_i, y_j) , 则称 (X, Y) 为二维离散随机变量, 称 $p_{ij} = P(X = x_i, Y = y_j), i, j = 1, 2, \dots$ 为 (X, Y) 的联合分布列

Proposition 1.3.3 联合分布列的基本性质:

1. 非负性: $p_{ij} \geq 0$
2. 正则性: $\sum_{i=1}^{+\infty} \sum_{j=1}^{+\infty} p_{ij} = 1$

Definition 1.3.4 — 列联表. 列联表中, 中间部分是 (ξ, η) 的联合概率分布, 而边沿部分是 ξ 及 η 的概率分布, 它们由联合分布经同一行或同一列的相加而得出来. 在列联表中, ξ 与 η 的概率分布处于表的边沿部位, 因此称为边际分布。

Definition 1.3.5 — 联合密度函数. 如果存在二元非负函数 $p(x, y)$, 使得二维随机变量 (X, Y) 的分布函数 $F(x, y)$ 可表示为

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y p(u, v) dv du$$

则称 (X, Y) 为二维连续随机变量, 称 $p(x, y)$ 为 (X, Y) 的联合密度函数. 联合密度函数的基本性质:

1. 非负性: $p(x, y) \geq 0$
 2. 正则性: $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) dx dy = 1$
- 在 $F(x, y)$ 偏导数存在的点上有

$$p(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

若 G 为平面上的一个区域, 则有

$$P((X, Y) \in G) = \int_G p(x, y) dx dy$$



两个随机变量的边缘分布都是一样的, 联合分布一样吗? --否

■ **Example 1.7** n 维分布而言, 存在着 $n - 1$ 维, $n - 2$ 维, \dots , 2 维, 1 维的边际分布。在二维场合, 两个随机变量可以都是离散型的, 也可以都是连续型的, 上举例子都是如此; 但是也可以一个是离散型的, 另一个却是连续型的. 进一步, 也容易举出既非离散型又非连续型的例子, 这时整个概率测度集中在一个不可列的一维点集上, 因此也不存在密度函数. 例如, 若 $\xi \sim U[0, 1]$, 令 $\eta = \xi^2$, 则 (ξ, η) 既非离散型又没有联合密度函数。

Definition 1.3.6 — 多项分布. 在 n 次独立重复试验中, 如果每次试验有 r 个可能结果: A_1, A_2, \dots, A_r , 且每次试验中 A_i 发生的概率为 $p_i = P(A_i), i = 1, 2, \dots, r$. $p_1 + p_2 + \dots + p_r = 1$. 记 X_i 为 n 次独立重复试验中 A_i 出现的次数, $i = 1, 2, \dots, r$. 则 (X_1, X_2, \dots, X_r) 服从多项分布, 又称 r 项分布, 记为 $M(n, p_1, p_2, \dots, p_r)$, 其联合分布列为

$$P(X_1 = n_1, X_2 = n_2, \dots, X_r = n_r) = \frac{n!}{n_1! n_2! \dots n_r!} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$$

其中 $n = n_1 + n_2 + \dots + n_r$ 当 $r = 2$ 时, 即为二项分布.

Definition 1.3.7 — 多维超几何分布. 有 N 个对象, 共分 r 类, 其中第 i 类对象有 N_i 个, $N = N_1 + N_2 + \dots + N_r$, 从中随机取出 n 个, 若记 X_i 为取出的 n 个对象中第 i 类对象的个

数, $i = 1, 2, \dots, r$, 则 (X_1, X_2, \dots, X_r) 服从 r 维超几何分布, 其联合分布列为

$$P(X_1 = n_1, X_2 = n_2, \dots, X_r = n_r) = \frac{\binom{N_1}{n_1} \binom{N_2}{n_2} \cdots \binom{N_r}{n_r}}{\binom{N}{n}}$$

其中 $n_1 + n_2 + \cdots + n_r = n$

Definition 1.3.8 — 多维均匀分布. 设 D 为 R^n 中的一个有界区域, 其度量 (平面上为面积, 空间为体积等) 为 S_D , 如果多维随机变量 (X_1, X_2, \dots, X_n) 的联合密度函数为

$$p(x_1, x_2, \dots, x_n) = \begin{cases} \frac{1}{S_D}, & (x_1, x_2, \dots, x_n) \in D \\ 0, & \text{其他} \end{cases}$$

则称 (X_1, X_2, \dots, X_n) 服从 D 上的多维均匀分布, 记为 $(X_1, X_2, \dots, X_n) \sim U(D)$

Definition 1.3.9 — 二元正态分布. 如果二维随机变量 (X, Y) 的联合密度函数为

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}, \quad -\infty < x, y < +\infty$$

则称 (X, Y) 服从二元正态分布, 记为 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. 其中

$$E(X) = \mu_1, \quad E(Y) = \mu_2; \quad \text{Var}(X) = \sigma_1^2, \quad \text{Var}(Y) = \sigma_2^2; \quad -1 \leq \rho \leq 1$$

1.3.2 边际分布与随机变量的独立性

Definition 1.3.10 — 边际分布函数. 若二维随机变量 (X, Y) 的联合分布函数为 $F(x, y)$ 则称

$$F_x(x) = F(x, +\infty) = \lim_{y \rightarrow +\infty} F(x, y), \quad -\infty < x < +\infty$$

为 X 的边际分布. 称

$$F_y(y) = F(+\infty, y) = \lim_{x \rightarrow +\infty} F(x, y), \quad -\infty < y < +\infty$$

为 Y 的边际分布.

Definition 1.3.11 — 边际分布列. 若二维离散随机变量 (X, Y) 的联合分布列为 $\{p_{ij}\}$, 则称

$$p_{i \cdot} = \sum_{j=1}^{+\infty} p_{ij}, \quad i = 1, 2, \dots$$

为 X 的边际分布列. 称

$$p_{\cdot j} = \sum_{i=1}^{+\infty} p_{ij}, \quad j = 1, 2, \dots$$

为 Y 的边际分布列。

Definition 1.3.12 — 边际密度函数. 若二维连续随机变量 (X, Y) 的联合密度函数为 $p(x, y)$ 则称

$$p_x(x) = \int_{-\infty}^{+\infty} p(x, y) dy, \quad -\infty < x < +\infty$$

为 X 的边际密度函数. 称

$$p_Y(y) = \int_{-\infty}^{+\infty} p(x, y) dx, \quad -\infty < y < +\infty$$

为 Y 的边际密度函数。

■ **Example 1.8** 多项分布的一维边际分布仍为二项分布下面先证三项分布的边际分布为二项分布. 设二维随机变量 (X, Y) 服从三项分布 $M(n, p_1, p_2, p_3)$, 其联合分布列为

$$P(X = i, Y = j) = \frac{n!}{i! j! (n-i-j)!} p_1^i p_2^j (1-p_1-p_2)^{n-i-j}$$

$i, j = 1, 2, \dots, n, i+j \leq n$ 对上式分别乘以和除以 $(1-p_1)^{n-i}/(n-i)!$, 再对 j 从 0 到 $n-i$ 求和, 并记 $p'_2 = p_2/(1-p_1)$, 则可得

$$\begin{aligned} \sum_{j=0}^{n-i} P(X = i, Y = j) &= \frac{n!}{i!(n-i)!} p_1^i (1-p_1)^{n-i} \cdot \\ &\quad \sum_{j=0}^{n-i} \binom{n-i}{j} \left(\frac{p_2}{1-p_1}\right)^j \left(1 - \frac{p_2}{1-p_1}\right)^{n-i-j} \\ &= \frac{n!}{i!(n-i)!} p_1^i (1-p_1)^{n-i} [p'_2 + (1-p'_2)]^{n-i} \\ &= \frac{n!}{i!(n-i)!} p_1^i (1-p_1)^{n-i} \end{aligned}$$

所以 $X \sim b(n, p_1)$. 同理可证 $Y \sim b(n, p_2)$. 用类似的方法可以证明: 若多维随机变量 $(X_1, X_2, \dots, X_r) \sim M(n, p_1, p_2, \dots, p_r)$, 则 $X_i \sim b(n, p_i), i = 1, 2, \dots, r$

Theorem 1.3.4 二维正态分布的边际分布为一维正态分布

Proof. 设二维随机变量 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. 先把二维正态密度函数 $p(x, y)$ 的指数部分

$$-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1 \sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right]$$

改写成

$$-\frac{1}{2} \left[\rho \frac{x-\mu_1}{\sigma_1 \sqrt{1-\rho^2}} - \frac{y-\mu_2}{\sigma_2 \sqrt{1-\rho^2}} \right]^2 - \frac{(x-\mu_1)^2}{2\sigma_1^2}$$

再对积分

$$\int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} \left[\rho \frac{x-\mu_1}{\sigma_1 \sqrt{1-\rho^2}} - \frac{y-\mu_2}{\sigma_2 \sqrt{1-\rho^2}} \right]^2 \right\} dy$$

作变换 (注意把 x 看作常量)

$$t = \rho \frac{x - \mu_1}{\sigma_1 \sqrt{1 - \rho^2}} - \frac{y - \mu_2}{\sigma_2 \sqrt{1 - \rho^2}}$$

则

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{\infty} p(x,y) dy \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \sigma_2 \sqrt{1-\rho^2} \int_{-\infty}^{\infty} \exp\left\{-\frac{t^2}{2}\right\} dt \end{aligned}$$

注意到上式中的积分恰好等于 $\sqrt{2}$

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}$$

这正是一维正态分布 $N(\mu_1, \sigma_1^2)$ 的密度函数, 即 $X \sim N(\mu_1, \sigma_1^2)$. 同理可证 $Y \sim N(\mu_2, \sigma_2^2)$. 由此可见

1. 二维正态分布的边际分布中不含参数 ρ
2. 这说明二维正态分布 $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, 0.1)$ 与 $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, 0.2)$ 的边际分布是相同的。
3. 具有相同边际分布的多维联合分布可以是不同的。

■

(R)

1. 由高维联合分布可以获得低维的边际分布, 反之不一定
2. 不同的联合分布可以有相同边际分布
3. 多项分布的边际分布仍为低维的多项分布或二项分布
4. 二维正态分布的边际分布为一维正态分布

Definition 1.3.13 设 n 维随机变量 (X_1, X_2, \dots, X_n) 的联合分布函数为 $F(x_1, x_2, \dots, x_n)$, 且 $F_{x_i}(x_i)$ 为 X_i 的边际分布函数. 如果对任意 n 个实数 x_1, x_2, \dots, x_n , 有

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F_{x_i}(x_i)$$

则称 X_1, X_2, \dots, X_n 相互独立. 否则称 X_1, X_2, \dots, X_n 不相互独立, 或相关.

Definition 1.3.14 设 n 维离散随机变量 (X_1, X_2, \dots, X_n) 的联合分布列为 $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$, 且 $P(X_i = x_i)$ 为 X_i 的边际分布列, $i = 1, 2, \dots, n$. 如果对其任意 n 个取值 x_1, x_2, \dots, x_n , 有

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$$

则称 X_1, X_2, \dots, X_n 相互独立的. 否则称 X_1, X_2, \dots, X_n 不相互独立, 或相关.

Definition 1.3.15 设 n 维连续随机变量 (X_1, X_2, \dots, X_n) 的联合密度函数为 $p(x_1, x_2, \dots, x_n)$,

且 $p_{x_i}(x_i)$ 为 X_i 的边际密度函数. 如果对任意 n 个实数 x_1, x_2, \dots, x_n , 有

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p_{x_i}(x_i)$$

则称 X_1, X_2, \dots, X_n 相互独立的. 否则称 X_1, X_2, \dots, X_n 不相互独立, 或相关.

Theorem 1.3.5

1. 多维随机变量间相互独立, 必导致其部分随机变量与另一部分随机变量相互独立
2. 多维随机变量相互独立, 其联合分布可由其边际分布唯一确定
3. 多维随机变量间的独立性可从定义出发加以判别, 也可从实际背景加以判别
4. 多维随机变量间的独立性假设, 可给理论研究和实际运用带来很多方便之处

1.3.3 多维随机变量函数的分布

Definition 1.3.16 — 最大值, 最小值分布. 设 (X_1, X_2, \dots, X_n) 是相互独立、同分布的 n 维连续随机变量, 其共同的密度函数和分布函数分别为 $p(x)$ 和 $F(x)$, 记

$$Y = \min\{X_1, X_2, \dots, X_n\}, \quad Z = \max\{X_1, X_2, \dots, X_n\}$$

$$\begin{aligned} F_Y(y) &= 1 - [1 - F(y)]^n; \quad p_Y(y) = n[1 - F(y)]^{n-1} p(y) \\ F_Z(z) &= [F(z)]^n; \quad p_Z(z) = n[F(z)]^{n-1} p(z) \end{aligned}$$

Theorem 1.3.6 讨论 (ξ_1^*, ξ_n^*) 最大最小值的联合分布.

$$\text{记 } G(x, y) = P\{\xi_1^* < x, \xi_n^* < y\}$$

若 $x \geq y$, 则

$$\begin{aligned} G(x, y) &= P\{\xi_1^* < x, \xi_n^* < y\} \\ &= P\{\xi_n^* < y\} = [F(y)]^n \end{aligned}$$

若 $x < y$, 则

$$\begin{aligned} G(x, y) &= P\{\xi_1^* < x, \xi_n^* < y\} \\ &= P\{\xi_n^* < y\} - P\{\xi_1^* \geq x, \xi_n^* < y\} \\ &= [F(y)]^n - [F(y) - F(x)]^n \end{aligned}$$

其联合密度函数为

$$q(x, y) = \begin{cases} 0, & x \geq y \\ n(n-1)[F(y) - F(x)]^{n-2} p(x)p(y), & x < y \end{cases}$$

Theorem 1.3.7 最后, 我们来求极差 $R = \xi_n^* - \xi_1^*$ 的分布密度函数 $f_R(r)$, 显然对 $r \leq$

$0, f_R(r) = 0$, 若 $r > 0$, 则

$$\begin{aligned} P\{R < r\} &= \iint_{y-x < r} q(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x+r} q(x, y) dy \right] dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^r q(x, x+z) dz \right] dx \\ &= \int_{-\infty}^r \left[\int_{-\infty}^{\infty} q(x, x+z) dx \right] dz \end{aligned}$$

因此

$$\begin{aligned} f_R(r) &= \int_{-\infty}^{\infty} q(x, x+r) dx \\ &= n(n-1) \int_{-\infty}^{\infty} [F(x+r) - F(x)]^{n-2} p(x) p(x+r) dx \end{aligned}$$

Theorem 1.3.8 — 变量变换法. 若变换 $\begin{cases} u = g_1(x, y), \\ v = g_2(x, y) \end{cases}$, 存在唯一的反函数 $\begin{cases} x = x(u, v), \\ y = y(u, v) \end{cases}$ 变换的雅可比行列式

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \left(\frac{\partial(u, v)}{\partial(x, y)} \right)^{-1} = \left(\begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} \right)^{-1} \neq 0$$

则二维连续随机变量 (X, Y) 的函数 $\begin{cases} U = g_1(X, Y) \\ V = g_2(X, Y) \end{cases}$ 的联合密度函数为

$$p_{v,v}(u, v) = p_{X,Y}(x(u, v), y(u, v)) |J|$$

Theorem 1.3.9 — 卷积公式. 卷积公式用于求随机变量和 $Z = X + Y$ 的分布.

1. 离散场合的卷积公式 $Z = X + Y$ 的分布列为

$$\begin{aligned} P(Z = z_k) &= \sum_{i=1}^{+\infty} P(X = x_i, Y = z_k - x_i) \\ \text{或} &= \sum_{j=1}^{+\infty} P(X = z_k - y_j, Y = y_j) \end{aligned}$$

当 X 与 Y 独立时,

$$\begin{aligned} P(Z = z_k) &= \sum_{i=1}^{+\infty} P(X = x_i) P(Y = z_k - x_i) \\ \text{或} &= \sum_{j=1}^{+\infty} P(X = z_k - y_j) P(Y = y_j) \end{aligned}$$

其中诸 x_i 为 X 的取值, 诸 y_j 为 Y 的取值.

2. 连续场合的卷积公式 $Z = X + Y$ 的密度函数为

$$p_Z(z) = \int_{-\infty}^{+\infty} p_{X,Y}(x, z-x) dx$$

或 $= \int_{-\infty}^{+\infty} p_{X,Y}(z-y, y) dy$

当 X 与 Y 独立时

$$p_Z(z) = \int_{-\infty}^{+\infty} p_X(x) p_Y(z-x) dx$$

或 $= \int_{-\infty}^{+\infty} p_X(z-y) p_Y(y) dy$

Theorem 1.3.10 — 积的公式. $U = X \cdot Y$ 的密度函数为

$$p_U(u) = \int_{-\infty}^{+\infty} p_{X,Y}(u/v, v) \frac{1}{|v|} dv$$

或 $= \int_{-\infty}^{+\infty} p_{X,Y}(v, u/v) \frac{1}{|v|} dv$

当 X 与 Y 相互独立时,

$$p_U(u) = \int_{-\infty}^{+\infty} p_X(u/v) p_Y(v) \frac{1}{|v|} dv$$

或 $= \int_{-\infty}^{+\infty} p_X(v) p_Y(u/v) \frac{1}{|v|} dv$

Theorem 1.3.11 — 商的公式. $U = X/Y$ 的密度函数为

$$p_U(u) = \int_{-\infty}^{+\infty} p_{X,Y}(uv, v) |v| dv$$

或 $= \int_{-\infty}^{+\infty} p_{X,Y}(v, v/u) \frac{|v|}{u^2} dv$

当 X 与 Y 独立时

$$p_U(u) = \int_{-\infty}^{+\infty} p_X(uv) p_Y(v) |v| dv$$

或 $= \int_{-\infty}^{+\infty} p_X(v) p_Y(v/u) \frac{|v|}{u^2} dv$

Theorem 1.3.12 — 分布的可加性. 若同一类分布的独立随机变量和的分布仍属于此类分布, 则称此类分布具有可加性. 以下一些常用分布具有可加性:

1. 二项分布: 若 $X \sim b(n, p), Y \sim b(m, p)$, 且 X 与 Y 独立, 则 $Z = X + Y \sim b(n+m, p)$. 注意这里两个二项分布中的参数 p 要相同.
2. 泊松分布: 若 $X \sim P(\lambda_1), Y \sim P(\lambda_2)$, 且 X 与 Y 独立, 则 $Z = X + Y \sim P(\lambda_1 + \lambda_2)$

3. 正态分布: 若 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 且 X 与 Y 独立, 则 $Z =$

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

4. 伽玛分布: 若 $X \sim Ga(\alpha_1, \lambda), Y \sim Ga(\alpha_2, \lambda)$, 且 X 与 Y 独立, 则 $Z = X + Y \sim Ga(\alpha_1 + \alpha_2, \lambda)$. 注: 这里两个伽玛分布中的尺度参数 λ 要相同.

5. χ^2 分布: 若 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$, 且 X 与 Y 独立, 则 $Z = X + Y \sim \chi^2(n_1 + n_2)$

Theorem 1.3.13 — (泊松分布的可加性). 设随机变量 $X \sim P(\lambda_1), Y \sim P(\lambda_2)$, 且 X 与 Y 独立, 证明 $Z = X + Y \sim P(\lambda_1 + \lambda_2)$

Proof. 首先指出, $Z = X + Y$ 可取 $0, 1, 2, \dots$ 所有非负整数. 而事件 $\{Z = k\}$ 是如下诸互不相容事件

$$\{X = i, Y = k - i\}, \quad i = 0, 1, \dots, k$$

的并, 再考虑到独立性, 则对任意非负整数 k , 有

$$P(Z = k) = \sum_{i=0}^k P(X = i)P(Y = k - i)$$

这个概率等式被称为离散场合下的卷积公式. 利用此公式可得

$$\begin{aligned} P(Z = k) &= \sum_{i=0}^k \left(\frac{\lambda_1^i}{i!} e^{-\lambda_1} \right) \left(\frac{\lambda_2^{k-i}}{(k-i)!} e^{-\lambda_2} \right) \\ &= \frac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1 + \lambda_2)} \sum_{i=0}^k \frac{k!}{i!(k-i)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^i \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{k-i} \\ &= \frac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1 + \lambda_2)} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^k \\ &= \frac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1 + \lambda_2)}, \quad k = 0, 1, \dots \end{aligned}$$

这表明 $Z = X + Y \sim P(\lambda_1 + \lambda_2)$, 结论得证. 注意 $X - Y$ 不服从泊松分布. ■

Theorem 1.3.14 — (二项分布的可加性). 设随机变量 $X \sim b(n, p), Y \sim b(m, p)$, 且 X 与 Y 独立, 证明 $Z = X + Y \sim b(n+m, p)$

Proof. 首先指出, $Z = X + Y$ 可取 $0, 1, 2, \dots, n+m$ 等 $n+m+1$ 个不同的值, 利用离散场合的卷积公式, 可把事件 $\{Z = k\}$ 的概率表示为

$$P(Z = k) = \sum_{i=0}^k P(X = i)P(Y = k - i)$$

在二项分布场合, 上式中有些事件是不可能事件:

1. 当 $i > n$ 时, $\{X = i\}$ 是不可能事件, 所以只需考虑 $i \leq n$
2. 当 $k - i > m$ 时, $\{Y = k - i\}$ 是不可能事件, 所以只需考虑 $i \geq k - m$

因此记

$$a = \max\{0, k-m\}, \quad b = \min\{n, k\}$$

则

$$\begin{aligned} P(Z=k) &= \sum_{i=a}^b P(X=i)P(Y=k-i) \\ &= \sum_{i=a}^b \binom{n}{i} p^i (1-p)^{n-i} \cdot \binom{m}{k-i} p^{k-i} (1-p)^{m-(k-i)} \\ &= p^k (1-p)^{n+m-k} \sum_{i=a}^b \binom{n}{i} \binom{m}{k-i} \end{aligned}$$

利用超几何分布可证明上式组合乘积的和满足：

$$\sum_{i=a}^b \frac{\binom{n}{i} \binom{m}{k-i}}{\binom{n+m}{k}} = 1 \quad \text{或} \quad \sum_{i=a}^b \binom{n}{i} \binom{m}{k-i} = \binom{n+m}{k}$$

代回原式，可得

$$P(Z=k) = \binom{n+m}{k} p^k (1-p)^{n+m-k}, \quad k = 0, 1, \dots, n+m$$

这表明 $Z = X + Y \sim b(n+m, p)$, 即在参数 p 相同情况下, 二项分布的卷积仍是二项分布, 即 $b(n, p) * b(m, p) = b(n+m, p)$. ■

Theorem 1.3.15 显然, $\xi_i \sim B(n, p_i), i = 1, 2, \dots, r$. 因此

$$E\xi_i = np_i, \quad D\xi_i = np_i(1-p_i)$$

为求协方差或相关系数, 可用下面技巧: 注意到

$$\xi_i + \xi_j \sim B(n, p_i + p_j)$$

因此

$$E(\xi_i + \xi_j) = n(p_i + p_j), \quad D(\xi_i + \xi_j) = n(p_i + p_j)(1 - p_i - p_j)$$

Theorem 1.3.16 — (正态分布的可加性) . 设随机变量 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ 且 X 与 Y 独立, 证明 $Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

Proof. 首先指出 $Z = X + Y$ 仍在 $(-\infty, \infty)$ 上取值, 利用卷积公式可得

$$p_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} \left[\frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\} dy$$

对上式被积函数中的指数部分按 y 的幂次展开, 再合并同类项, 不难得出

$$\frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} = A \left(y - \frac{B}{A} \right)^2 + \frac{(z-\mu_1-\mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

其中

$$A = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \quad B = \frac{z - \mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}$$

代入原式，可得

$$p_z(z) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\frac{(z - \mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right\} \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{A}{2}\left(y - \frac{B}{A}\right)^2\right\} dy$$

利用正态密度函数的正则性，上式中的积分应为 $\sqrt{2\pi}/\sqrt{A}$ ，于是

$$p_z(z) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left\{-\frac{1}{2}\frac{(z - \mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right\}$$

这正是均值为 $\mu_1 + \mu_2$ ，方差为 $\sigma_1^2 + \sigma_2^2$ 的正态密度函数。上述结论表明：两个独立的正态变量之和仍为正态变量，其分布中的两个参数分别对应相加，即

$$N(\mu_1, \sigma_1^2) * N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

显然，这个结论可以推广到有限个独立正态变量之和的场合。 ■

Theorem 1.3.17 — (伽玛分布的可加性) . 设随机变量 $X \sim Ga(\alpha_1, \lambda), Y \sim Ga(\alpha_2, \lambda)$ 且 X 与 Y 独立，证明 $Z = X + Y \sim Ga(\alpha_1 + \alpha_2, \lambda)$

Proof. 首先指出 $Z = X + Y$ 仍在 $(0, \infty)$ 上取值，所以当 $z \leq 0$ 时， $p_z(z) = 0$ 。而当 $z > 0$ 时，可用卷积公式，此时被积函数 $p_X(z-y)p_Y(y)$ 的非零区域为 $0 < y < z$ ，故

$$\begin{aligned} p_z(z) &= \frac{\lambda^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z (z-y)^{\alpha_1-1} e^{-\lambda(z-y)} \cdot y^{\alpha_2-1} e^{-\lambda y} dy \\ &= \frac{\lambda^{\alpha_1 + \alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z (z-y)^{\alpha_1-1} y^{\alpha_2-1} dy \\ &= \frac{\lambda^{\alpha_1 + \alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1 + \alpha_2 - 1} \int_0^1 (1-t)^{\alpha_1-1} t^{\alpha_2-1} dt \end{aligned}$$

最后的积分是贝塔函数，它等于 $\Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1 + \alpha_2)$ 。代入上式得

$$p_z(z) = \frac{\lambda^{\alpha_1 + \alpha_2}}{\Gamma(\alpha_1 + \alpha_2)} z^{\alpha_1 + \alpha_2 - 1} e^{-\lambda z}$$

这正是形状参数为 $\alpha_1 + \alpha_2$ ，尺度参数仍为 λ 的伽玛分布。这个结论表明：两个尺度参数相同的独立的伽玛变量之和仍为伽玛变量，其尺度参数不变，而形状参数相加，即

$$Ga(\alpha_1, \lambda) * Ga(\alpha_2, \lambda) = Ga(\alpha_1 + \alpha_2, \lambda)$$

显然这个结论可推广到有限个尺度参数相同的独立伽玛变量之和上。 ■

Corollary 1.3.18 1. m 个独立同分布的指数变量之和为伽玛变量，即

$$\underbrace{\text{Exp}(\lambda) * \text{Exp}(\lambda) * \cdots * \text{Exp}(\lambda)}_{m \uparrow} = Ga(m, \lambda)$$

2. m 个独立的 χ^2 变量之和为 χ^2 变量 (χ^2 分布的可加性), 即

$$\chi^2(n_1) * \chi^2(n_2) * \cdots * \chi^2(n_m) = \chi^2(n_1 + n_2 + \cdots + n_m)$$

Theorem 1.3.19 1. 设 X_1, X_2, \dots, X_n 独立同分布, 都服从二点分布 $b(1, p)$, 则 $X_1 + X_2 + \cdots + X_n$ 服从二项分布 $b(n, p)$
 2. 设 X_1, X_2, \dots, X_n 独立同分布, 都服从几何分布 $Ge(p)$, 则 $X_1 + X_2 + \cdots + X_n$ 服从负二项分布 $Nb(n, p)$
 3. 设 X_1, X_2, \dots, X_n 独立同分布, 都服从指数分布 $Exp(\lambda)$, 则 $X_1 + X_2 + \cdots + X_n$ 服从伽玛分布 $Ga(n, \lambda)$

Theorem 1.3.20 随机变量 $\xi_1, \xi_2, \dots, \xi_n$ 相互独立的充要条件是对一切一维博雷尔点集 A_1, A_2, \dots, A_n 成立

$$\begin{aligned} & P\{\xi_1 \in A_1, \xi_2 \in A_2, \dots, \xi_n \in A_n\} \\ &= P\{\xi_1 \in A_1\} P\{\xi_2 \in A_2\} \cdots P\{\xi_n \in A_n\} \end{aligned}$$

Theorem 1.3.21 若 ξ_1, \dots, ξ_n 是相互独立的随机变量, 则 $f_1(\xi_1), \dots, f_n(\xi_n)$ 也是相互独立的, 这里 $f_i(i = 1, \dots, n)$ 是任意的一元博雷尔函数。

Proof. 对任意的一维博雷尔点集 A_1, \dots, A_n 有

$$\begin{aligned} & P\{f_1(\xi_1) \in A_1, \dots, f_n(\xi_n) \in A_n\} \\ &= P\{\xi_1 \in f_1^{-1}(A_1), \dots, \xi_n \in f_n^{-1}(A_n)\} \\ &= P\{\xi_1 \in f_1^{-1}(A_1)\} \cdots P\{\xi_n \in f_n^{-1}(A_n)\} \\ &= P\{f_1(\xi_1) \in A_1\} \cdots P\{f_n(\xi_n) \in A_n\} \end{aligned}$$

■

1.3.4 多维随机变量的特征数

Theorem 1.3.22 — 多维随机变量函数的数学期望. 若二维随机变量 (X, Y) 的分布用联合分布列 $P(X = x_i, Y = y_j)$ 或用联合密度函数 $p(x, y)$ 表示, 则 $Z = g(X, Y)$ 的数学期望 (假设存在) 为

$$E[g(X, Y)] = \begin{cases} \sum_i \sum_j g(x_i, y_j) P(X = x_i, Y = y_j), & \text{在离散场合} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) p(x, y) dx dy, & \text{在连续场合.} \end{cases}$$

对 n 维随机变量结论是类似的.

Proposition 1.3.23 数学期望与方差的运算性质: 以下均假定有关的期望和方差存在。

1. $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$
2. 若 X_1, X_2, \dots, X_n 相互独立, 则

$$E(X_1 X_2 \cdots X_n) = E(X_1) E(X_2) \cdots E(X_n)$$

$$\text{Var}(X_1 \pm X_2 \pm \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)$$

Definition 1.3.17 — 协方差. 若 $E[(X - E(X))(Y - E(Y))]$ 存在, 则称 $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$ 为 X 与 Y 的协方差。

1. 当 $\text{Cov}(X, Y) > 0$ 时, 称 X 与 Y 正相关, 即 X 与 Y 同时增加或同时减少
2. 当 $\text{Cov}(X, Y) < 0$ 时, 称 X 与 Y 负相关, 即 X 增加 Y 减少, 或 X 减少 Y 增加.
3. 当 $\text{Cov}(X, Y) = 0$ 时, 称 X 与 Y 不相关.

Theorem 1.3.24 若 ξ 与 η 都是二值随机变量, 则不相关性与独立性是等价的。

$$|P(AB) - P(A)P(B)| \leq \frac{1}{4}$$

Proposition 1.3.25 协方差的性质

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
2. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
3. 若 X 与 Y 相互独立, 则 $\text{Cov}(X, Y) = 0$, 反之不然
4. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
5. 对任意的常数 a , 有 $\text{Cov}(X, a) = 0$
6. 对任意的常数 a, b , 有 $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$
7. 对任意二维随机变量 (X, Y) , 有

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X, Y)$$

对任意 n 个随机变量 X_1, X_2, \dots, X_n , 有

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=1}^{i-1} \text{Cov}(X_i, X_j)$$

Theorem 1.3.26 — 许瓦兹不等式. 对任意二维随机变量 (X, Y) , 若 X 与 Y 的方差都存在, 则有

$$[\text{Cov}(X, Y)]^2 \leq \text{Var}(X)\text{Var}(Y)$$

Proof. 证明不妨设 $\sigma_x^2 > 0$, 因为当 $\sigma_x^2 = 0$ 时, 则 X 几乎处处为常数, 因而其与 Y 的协方差亦为零, 从而两端皆为零, 结论成立. 若 $\sigma_x^2 > 0$ 成立, 考虑 t 的如下二次函数:

$$g(t) = E[t(X - E(X)) + (Y - E(Y))]^2 = t^2\sigma_X^2 + 2t \cdot \text{Cov}(X, Y) + \sigma_Y^2$$

由于上述的二次三项式非负, 平方项系数 σ_x^2 为正, 所以其判别式小于或等于零, 即

$$[2\text{Cov}(X, Y)]^2 - 4\sigma_X^2\sigma_Y^2 \leq 0$$

移项后即得施瓦茨不等式. ■

Definition 1.3.18 — 相关系数. 设 (X, Y) 是一个二维随机变量, 且 $\text{Var}(X) > 0, \text{Var}(Y) > 0$ 则称

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}$$

为 X 与 Y 的(线性)相关系数。

Proposition 1.3.27 相关系数的性质

1. $-1 \leq \text{Corr}(X, Y) \leq 1$

2. $\text{Corr}(X, Y)$ 与 $\text{Cov}(X, Y)$ 同号, 或同为零.
3. $\text{Corr}(X, Y) = \text{Cov}(X^*, Y^*)$, 其中 X^*, Y^* 分别为 X, Y 的标准化随机变量。
4. $\text{Corr}(X, Y) = \pm 1$ 的充要条件是 X 与 Y 间几乎处处有线性关系, 即存在

$a(a \neq 0)$ 与 b , 使得 $P(Y = aX + b) = 1$. 其中当 $\text{Corr}(X, Y) = 1$ 时, 有 $a > 0$; 当

$\text{Corr}(X, Y) = -1$ 时, 有 $a < 0$

5. 在二维正态分布 $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ 场合, 不相关与独立是等价的.

Proof. 第四条的证明: 证明充分性. 若 $Y = aX + b$ ($X = cY + d$ 也一样), 则将

$$\text{Var}(Y) = a^2 \text{Var}(X), \quad \text{Cov}(X, Y) = a \text{Cov}(X, X) = a \text{Var}(X)$$

代入相关系数的定义中得

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{a \text{Var}(X)}{|a| \text{Var}(X)} = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}$$

必要性: 因为

$$\text{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) = 2[1 \pm \text{Corr}(X, Y)]$$

所以当 $\text{Corr}(X, Y) = 1$ 时, 有

$$\text{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0$$

由此得

$$P\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = c\right) = 1$$

或

$$P\left(Y = \frac{\sigma_y}{\sigma_x}X - c\sigma_y\right) = 1$$

这就证明了: 当 $\text{Corr}(X, Y) = 1$ 时, Y 与 X 几乎处处为线性正相关. 当 $\text{Corr}(X, Y) = -1$ 时, 得

$$\text{Var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_Y}\right) = 0$$

由此得

$$P\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_Y} = c\right) = 1$$

或

$$P\left(Y = -\frac{\sigma_Y}{\sigma_X}X + c\sigma_Y\right) = 1$$

这也证明了: 当 $\text{Corr}(X, Y) = -1$ 时, Y 与 X 几乎处处为线性负相关. ■

(R)

1. 相关系数 $\text{Corr}(X, Y)$ 刻画了 X 与 Y 之间的线性关系强弱, 因此也常称其为“线性相关系数”
2. 若 $\text{Corr}(X, Y) = 0$, 则称 X 与 Y 不相关. 不相关是指 X 与 Y 之间没有线性关系, 但 X 与 Y 之间可能有其他的函数关系.
3. 若 $\text{Corr}(X, Y) = 1$, 则称 X 与 Y 完全正相关; 若 $\text{Corr}(X, Y) = -1$, 则称 X 与 Y 完全负相关.

Definition 1.3.19 — 随机向量的数学期望与协方差阵. 记 n 维随机向量为 $X = (X_1, X_2, \dots, X_n)'$, 以下假设所涉及的数学期望、方差、协方差均存在.

1. 随机向量 X 的数学期望向量为

$$E(X) = (E(X_1), E(X_2), \dots, E(X_n))'$$

2. 随机向量 X 的协方差阵为

$$E[(X - E(X))(X - E(X))'] = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Var}(X_n) \end{pmatrix}$$

也记以上的协方差阵为 $\text{Cov}(X)$, 或记成 Σ .

3. 随机向量 X 的协方差阵 $\text{Cov}(X) = (\text{Cov}(X_i, X_j))_{n \times n}$ 是一个对称的非负定矩阵.

Proof. 第三条的证明: 证因为 $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, 所以对称性是显然的. 下证非负定性. 因为对任意的 n 维实向量 $c = (c_1, c_2, \dots, c_n)'$, 有

$$\begin{aligned} c' \text{Cov}(X) c &= (c_1, c_2, \dots, c_n) \begin{pmatrix} \text{Var}(X_1) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & & \vdots \\ \text{Cov}(X_n, X_1) & \cdots & \text{Var}(X_n) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} \\ &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \sum_{j=1}^n E\{[c_i(X_i - E(X_i))] [c_j(X_j - E(X_j))]\} \\ &= E\left\{\sum_{i=1}^n \sum_{j=1}^n [c_i(X_i - E(X_i))] [c_j(X_j - E(X_j))]\right\} \\ &= E\left\{\left[\sum_{i=1}^n c_i(X_i - E(X_i))\right] \left[\sum_{j=1}^n c_j(X_j - E(X_j))\right]\right\} \\ &= E\left[\sum_{i=1}^n c_i(X_i - E(X_i))\right]^2 \geqslant 0 \end{aligned}$$

所以矩阵 $\text{Cov}(X)$ 是非负定的, 定理得证. ■

Definition 1.3.20 — n 元正态分布. 设 n 维随机向量 $X = (X_1, X_2, \dots, X_n)'$ 的协方阵为 $\Sigma =$

$\text{Cov}(X)$, 数学期望向量为 $a = (a_1, a_2, \dots, a_n)'$. 又记 $x = (x_1, x_2, \dots, x_n)'$, 则由密度函数

$$p(x_1, x_2, \dots, x_n) = p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x-a)' \Sigma^{-1} (x-a) \right\}$$

定义的分布称为 n 元正态分布, 记为 $X \sim N(a, \Sigma)$

Theorem 1.3.28 — n 元正态分布. 设 n 维随机变量 $X = (X_1, X_2, \dots, X_n)'$ 的协方差矩阵为 $B = \text{Cov}(X)$, 数学期望向量为 $a = (a_1, a_2, \dots, a_n)'$. 又记 $x = (x_1, x_2, \dots, x_n)'$, 则由密度函数

$$p(x_1, x_2, \dots, x_n) = p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |B|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x-a)' B^{-1} (x-a) \right\}$$

定义的分布称为 n 元正态分布, 记为 $X \sim N(a, B)$. 其中 $|B|$ 表示 B 的行列式, B^{-1} 表示 B 的逆阵, $(x-a)'$ 表示向量 $(x-a)$ 的转置. 若记 $B^{-1} = (r_{ij})$, 则 (3.4.13) 式可写成

$$p(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |B|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \sum_{i,j=1}^n r_{ij} (x_i - a_i)(x_j - a_j) \right\}$$

在 $n = 2$ 的场合, 若取数学期望向量和协方差矩阵分别为

$$a = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad B = \begin{pmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 \end{pmatrix}$$

1.3.5 条件分布与条件期望

离散随机变量的条件分布

Definition 1.3.21 — 条件分布列. 对一切使 $P(Y = y_j) = p_{\cdot j} = \sum_{i=1}^{+\infty} p_{ij} > 0$ 的 y_j , 称

$$p_{i|j} = P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}}, \quad i = 1, 2, \dots$$

为给定 $Y = y_j$ 条件下 X 的条件分布列. 同理, 对一切使 $P(X = x_i) = p_{i\cdot} = \sum_{j=1}^{+\infty} p_{ij} > 0$ 的 x_i , 称

$$p_{j|i} = P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i\cdot}}, \quad j = 1, 2, \dots$$

为给定 $X = x_i$ 条件下 Y 的条件分布列.

Definition 1.3.22 — 条件分布函数. 给定 $Y = y_j$ 条件下 X 的条件分布函数为

$$F(x|y_j) = \sum_{x_i \leq x} P(X = x_i | Y = y_j) = \sum_{x_i \leq x} p_{i|j}$$

给定 $X = x_i$ 条件下 Y 的条件分布函数为

$$F(y|x_i) = \sum_{y_j \leq y} P(Y = y_j | X = x_i) = \sum_{y_j \leq y} p_{j|i}$$

Example 1.9 设随机变量 X 与 Y 相互独立, 且 $X \sim P(\lambda_1), Y \sim P(\lambda_2)$. 在已知 $X + Y = n$ 的条件下, 求 X 的条件分布.

Proof. 解因为独立泊松变量的和仍为泊松变量, 即 $X + Y \sim P(\lambda_1 + \lambda_2)$, 所以

$$\begin{aligned} P(X = k | X + Y = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} \\ &= \frac{\frac{\lambda_1^k}{k!} e^{-\lambda_1} \cdot \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2}}{\frac{(\lambda_1 + \lambda_2)^n}{n!} e^{-(\lambda_1 + \lambda_2)}} \\ &= \frac{n!}{k!(n-k)!} \frac{\lambda_1^k \lambda_2^{n-k}}{(\lambda_1 + \lambda_2)^n} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k}, \quad k = 0, 1, \dots, n \end{aligned}$$

在 $X + Y = n$ 的条件下, X 服从二项分布 $b(n, p)$, 其中 $p = \lambda_1 / (\lambda_1 + \lambda_2)$

Definition 1.3.23 — 连续随机变量的条件分布. 对一切使 $p_Y(y) > 0$ 的 y , 给定 $Y = y$ 条件下 X 的条件密度函数和条件分布函数分别为

$$p(x|y) = \frac{p(x,y)}{p_Y(y)}, \quad F(x|y) = \int_{-\infty}^x p(u|y) du = \int_{-\infty}^x \frac{p(u,y)}{p_Y(y)} du$$

类似对一切使 $p_X(x) > 0$ 的 x , 给定 $X = x$ 条件下 Y 的条件密度函数和条件分布函数分别为

$$p(y|x) = \frac{p(x,y)}{p_X(x)}, \quad F(y|x) = \int_{-\infty}^y p(v|x) dv = \int_{-\infty}^y \frac{p(x,v)}{p_X(x)} dv$$

Theorem 1.3.29 连续场合的全概率公式和贝叶斯公式

1. 全概率公式的密度函数形式:

$$p_Y(y) = \int_{-\infty}^{+\infty} p_X(x)p(y|x) dx, \quad p_X(x) = \int_{-\infty}^{+\infty} p_Y(y)p(x|y) dy$$

2. 贝叶斯公式的密度函数形式:

$$p(x|y) = \frac{p_X(x)p(y|x)}{\int_{-\infty}^{+\infty} p_X(x)p(y|x) dx}, \quad p(y|x) = \frac{p_Y(y)p(x|y)}{\int_{-\infty}^{+\infty} p_Y(y)p(x|y) dy}$$

Definition 1.3.24 — 条件数学期望. 条件分布的数学期望 (若存在) 称为条件期望, 其定义如下:

$$E(X|Y = y) = \begin{cases} \sum_i x_i P(X = x_i | Y = y), & (X, Y) \text{ 为二维离散随机变量,} \\ \int_{-\infty}^{+\infty} xp(x|y) dx, & (X, Y) \text{ 为二维连续随机变量} \end{cases}$$

$$E(Y|X = x) = \begin{cases} \sum_j y_j P(Y = y_j | X = x), & (X, Y) \text{ 为二维离散随机变量,} \\ \int_{-\infty}^{+\infty} yp(y|x) dy, & (X, Y) \text{ 为二维连续随机变量.} \end{cases}$$

Theorem 1.3.30 1. 条件期望具有数学期望的一切性质。

2. 条件期望 $E(X|Y = y)$ 是 y 的函数, 记为 $g(y)$, 它是另一个随机变量 $g(Y) = E(X|Y)$

的取值. $E(X|Y)$ 与 $E(X|Y=y)$ 虽然都称为条件期望, 但含义不同. 前者是特定的随机变量, 后者是其取值。

Definition 1.3.25 在 $\xi=x$ 的条件下, η 的条件数学期望定义为

$$E\{\eta|\xi=x\} = \int_{-\infty}^{\infty} yp(y|x)dy$$

■ **Example 1.10** 正态分布条件分布:

$$E\{\eta|\xi=x\} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)$$

这时条件数学期望 $E\{\eta|\xi=x\}$ 是 x 的线性函数. 在正态分布中, 最佳预报是线性预报。

这个事实可以解释为: 残差中已不再包含对预测 η 有用的知识。因此观察值 η 被分解为两个不相关的随机变量之和:

$$\eta = \hat{\eta} + (\eta - \hat{\eta})$$

以上是二阶矩理论, 或称均值-方差理论。预测值与残差不相关。

Theorem 1.3.31 — 重期望公式. 设 (X, Y) 是二维随机变量, 若 $E(X)$ 存在, 则

$$E(X) = E[E(X|Y)]$$

注意: 该公式中里面的期望是用条件分布 $p(x|y)$ 计算的, 外面的期望是用 y 的分布 $p(y)$ 计算的

Definition 1.3.26 — 二维正态分布. $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ 的两个条件分布仍是正态分布, 即 $X|Y=y \sim N(g_1(y), \sigma_1^2(1-\rho^2))$

$$\text{其中 } g_1(y) = E(X|Y=y) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2)$$

$$Y|X=x \sim N(g_2(x), \sigma_2^2(1-\rho^2))$$

$$\text{其中 } g_2(x) = E(Y|X=x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)$$

可见, 在二维正态分布中, 一个变量的条件期望是另一个变量取值的线性函数, 常称为一元线性回归方程

Theorem 1.3.32 — 随机个随机变量和的数学期望. 设 X_1, X_2, \dots 为一列独立同分布的随机变量, 随机变量 N 只取正整数值, 且 N 与 $\{X_n\}$ 独立, 证明

$$E\left(\sum_{i=1}^N X_i\right) = E(X_1)E(N)$$

Proof. 由重期望定理知

$$\begin{aligned}
 E\left(\sum_{i=1}^N X_i\right) &= E\left[E\left(\sum_{i=1}^N X_i|N\right)\right] \\
 &= \sum_{n=1}^{\infty} E\left(\sum_{i=1}^N X_i|N=n\right) P(N=n) \\
 &= \sum_{n=1}^{\infty} E\left(\sum_{i=1}^n X_i\right) P(N=n) \\
 &= \sum_{n=1}^{\infty} nE(X_1) P(N=n) \\
 &= E(X_1) \sum_{n=1}^{\infty} nP(N=n) = E(X_1) E(N)
 \end{aligned}$$

■

Definition 1.3.27 假定我们研究的随机试验 α 只有有限个不相容的结果 A_1, A_2, \dots, A_n , 它们相应的概率为 $p(A_1), p(A_2), \dots, p(A_n)$, 满足 $\sum_{i=1}^n p(A_i) = 1$,

$$H(\alpha) = -\sum_{i=1}^n p(A_i) \log p(A_i)$$

为试验 α 的熵 (entropy)

Theorem 1.3.33 函数 $\varphi(x) = -x \log x$ 是凸函数, 考虑凸函数的性质: (延森 (Jensen) 不等式) 设 $\varphi(x)$ 是 $[a, b]$ 上的上凸函数, 而 x_1, x_2, \dots, x_n 是 $[a, b]$ 中的任意点, $\lambda_1, \lambda_2, \dots, \lambda_n$ 是和为 1 的正数, 则

$$\sum_{i=1}^n \lambda_i \varphi(x_i) \leq \varphi\left(\sum_{i=1}^n \lambda_i x_i\right)$$

等号成立当且仅当诸 x_i 相等.

Theorem 1.3.34 在有 n 个可能结果的试验中, 等概试验具有最大熵其值为 $\log n$

Theorem 1.3.35 以 $\alpha\beta$ 记这两个试验联合起来所构成的新试验, 于是试验 $\alpha\beta$ 的可能结果为 $A_k B_l, k = 1, 2, \dots, m, l = 1, 2, \dots, n$, 相应的概率为 $p(A_k B_l)$. 按定义

$$H(\alpha\beta) = -\sum_{k,l} p(A_k B_l) \log p(A_k B_l)$$

若试验 α 与试验 β 独立, 则

$$H(\alpha\beta) = H(\alpha) + H(\beta)$$

Definition 1.3.28 — 条件熵. 设 α, β 是前述两个试验, 以 $p(B_l|A_k)$ 记试验 α 出现结果 A_k

的条件下试验 β 出现结果 B_l 的概率, 则

$$H_{A_k}(\beta) = - \sum_{l=1}^n p(B_l|A_k) \log p(B_l|A_k)$$

是在试验 α 出现 A_k 的条件下, 试验 β 的熵. 我们称平均值

$$H_\alpha(\beta) = \sum_{k=1}^m p(A_k) H_{A_k}(\beta)$$

为在试验 α 实现的条件下试验 β 的条件熵.

- Proposition 1.3.36**
1. 熵加法法则: $H(\alpha\beta) = H(\alpha) + H_\alpha(\beta)$
 2. $H_\alpha(\beta) \leq H(\beta)$
 3. $H(\alpha\beta) = H(\alpha) + H_\alpha(\beta) \leq H(\alpha) + H(\beta)$

Definition 1.3.29 记

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta)$$

并称之为含在试验 α 中的有关试验 β 的信息量.

Definition 1.3.30 — 分布函数的熵.

$$H(\alpha) = - \int_{-\infty}^{\infty} p(x) \log p(x) dx$$

$$H(\alpha\beta) = - \iint f(x, y) \log f(x, y) dx dy$$

$$H_\alpha(\beta) = - \iint f(x, y) \log \frac{f(x, y)}{p(x)} dx dy$$

$$H_\beta(\alpha) = - \iint f(x, y) \log \frac{f(x, y)}{q(y)} dx dy$$

Definition 1.3.31

$$P(s) = \sum_{k=0}^{\infty} p_k s^k$$

为 ξ 的母函数 (generating function) .



2. 大数定律与中心极限定理

2.1 随机变量序列的两种收敛性

Definition 2.1.1 — 依概率收敛. 设 $\{X_n\}$ 为一随机变量序列, X 为一随机变量. 如果对任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow +\infty} P\{|X_n - X| < \varepsilon\} = 1$$

or

$$P(|X_n - X| \geq \varepsilon) \rightarrow 0 (n \rightarrow \infty)$$

则称 $\{X_n\}$ 依概率收敛于 X , 记作 $X_n \xrightarrow{P} X$.

Theorem 2.1.1 — 依概率收敛与服从大数定律间的关系. 设 $\{X_n\}$ 为一随机变量序列, 记 $Y_n = \frac{1}{n} \sum_{i=1}^n X_i, E(Y_n) = \frac{1}{n} \sum_{i=1}^n E(X_i)$. 则 $\{X_n\}$ 服从大数定律等价于 $Y_n - E(Y_n) \xrightarrow{P} 0$

Theorem 2.1.2 — 依概率收敛的四则运算. 如果 $X_n \xrightarrow{P} a, Y_n \xrightarrow{P} b$, 则有

1. $X_n \pm Y_n \xrightarrow{P} a \pm b$
2. $X_n \times Y_n \xrightarrow{P} a \times b$
3. $X_n \div Y_n \xrightarrow{P} a \div b (b \neq 0)$

Proof.

$$\begin{aligned}
 P\left(\left|\frac{1}{Y_n} - \frac{1}{b}\right| \geq \varepsilon\right) &= P\left(\left|\frac{Y_n - b}{Y_0 b}\right| \geq \varepsilon\right) \\
 &= P\left(\left|\frac{Y_n - b}{b^2 + b(Y_n - b)}\right| \geq \varepsilon, |Y_n - b| < \varepsilon\right) \\
 &\quad + P\left(\left|\frac{Y_n - b}{b^2 + b(Y_n - b)}\right| \geq \varepsilon, |Y_n - b| \geq \varepsilon\right) \\
 &\leq P\left(\frac{|Y_n - b|}{b^2 - \varepsilon|b|} \geq \varepsilon\right) + P(|Y_n - b| \geq \varepsilon) \\
 &= P(|Y_n - b| \geq (b^2 - \varepsilon|b|)\varepsilon) + P(|Y_n - b| \geq \varepsilon) \rightarrow 0
 \end{aligned}$$

■

Theorem 2.1.3 — 按分布收敛、弱收敛—分布函数的点点收敛. 设 $\{F_n(x)\}$ 是随机变量序列 $\{X_n\}$ 的分布函数列, $F(x)$ 为 X 的分布函数. 若对 $F(x)$ 的任一连续点 x , 都有 $\lim_{n \rightarrow +\infty} F_n(x) = F(x)$, 则称 $\{F_n(x)\}$ 弱收敛于 $F(x)$, 记作 $F_n(x) \xrightarrow{W} F(x)$, 也称 $\{X_n\}$ 按分布收敛于 X , 记作 $X_n \xrightarrow{L} X$

Theorem 2.1.4 设 $\{F_n(x)\}$ 是实变量 x 的非降函数列, D 是 \mathbf{R}^1 上的稠密集. 若对于 D 中的所有点, 序列 $\{F_n(x)\}$ 收敛于 $F(x)$, 则对 $F(x)$ 的一切连续点 x 有

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

Theorem 2.1.5 — 海莱第一定理. 任一一致有界的非降函数列 $\{F_n(x)\}$ 中必有一子序列 $\{F_{n_k}(x)\}$ 弱收敛于某一有界的非降函数 $F(x)$

Theorem 2.1.6 — 海莱第二定理. 设 $f(x)$ 是 $[a, b]$ 上的连续函数, 又 $\{F_n(x)\}$ 是在 $[a, b]$ 上弱收敛于函数 $F(x)$ 的一致有界非降函数序列, 且 a 和 b 是 $F(x)$ 的连续点, 则

$$\lim_{n \rightarrow \infty} \int_a^b f(x) dF_n(x) = \int_a^b f(x) dF(x)$$

Theorem 2.1.7 — 拓广的海莱第二定理. 设 $f(x)$ 在 $(-\infty, \infty)$ 上有界连续, 又 $\{F_n(x)\}$ 是 $(-\infty, \infty)$ 上弱收敛于函数 $F(x)$ 的一致有界非降函数序列, 且

$$\lim_{n \rightarrow \infty} F_n(-\infty) = F(-\infty), \quad \lim_{n \rightarrow \infty} F_n(\infty) = F(\infty)$$

则

$$\lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} f(x) dF_n(x) = \int_{-\infty}^{\infty} f(x) dF(x)$$

Theorem 2.1.8 — 依概率收敛与按分布收敛间的关系. 1. $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X$

2. $X_n \xrightarrow{P} c \Leftrightarrow X_n \xrightarrow{L} c$ (其中 c 为常数)

Proof. 证明设随机变量 X, X_1, X_2, \dots 的分布函数分别为 $F(x), F_1(x), F_2(x), \dots$ 为证 $X_n \xrightarrow{L} X$, 相当于证 $F_n(x) \xrightarrow{W} F(x)$, 所以只需证: 对所有的 x , 有

$$F(x-0) \leq \underline{\lim}_{n \rightarrow \infty} F_n(x) \leq \overline{\lim}_{n \rightarrow \infty} F_n(x) \leq F(x+0)$$

因为若上式成立, 则当 x 是 $F(x)$ 的连续点时, 有 $F(x-0) = F(x+0)$, 由此即可得 $F_d(x) \xrightarrow{W} F(x)$. 先令 $x' < x$, 则

$$\begin{aligned} \{X \leq x'\} &= \{X_n \leq x, X \leq x'\} \cup \{X_n > x, X \leq x'\} \\ &\subset \{X_n \leq x\} \cup \{|X_n - X| \geq x - x'\} \end{aligned}$$

从而有

$$F(x') \leq F_n(x) + P(|X_n - X| \geq x - x')$$

由 $X_n \xrightarrow{P} X$, 得 $P(|X_n - X| \geq x - x') \rightarrow 0 (n \rightarrow \infty)$. 所以有

$$F(x') \leq \liminf_{n \rightarrow \infty} F_n(x)$$

再令 $x' \rightarrow x$, 即得

$$F(x-0) \leq \underline{\lim}_{n \rightarrow \infty} F_n(x)$$

同理可证, 当 $x'' > x$ 时, 有

$$\overline{\lim}_{n \rightarrow \infty} F_n(x) \leq F(x'')$$

令 $x'' \rightarrow x$, 即得

$$\overline{\lim}_{n \rightarrow \infty} F_n(x) \leq F(x+0)$$

■

Proof. 下证充分性. 记 X_n 的分布函数为 $F_n(x), n = 1, 2, \dots$ 因为常数 c 的分布函数(退化分布)为

$$F(x) = \begin{cases} 0, & x < c \\ 1, & x \geq c \end{cases}$$

所以对任意的 $\varepsilon > 0$, 有

$$\begin{aligned} P(|X_n - c| \geq \varepsilon) &= P(X_n \geq c + \varepsilon) + P(X_n \leq c - \varepsilon) \\ &\leq P(X_n > c + \varepsilon/2) + P(X_n \leq c - \varepsilon) \\ &= 1 - F_n(c + \varepsilon/2) + F_n(c - \varepsilon) \end{aligned}$$

由于 $x = c + \varepsilon/2$ 和 $x = c - \varepsilon$ 均为 $F(x)$ 的连续点, 且 $F_n(x) \xrightarrow{W} F(x)$, 所以当 $n \rightarrow \infty$ 时, 有

$$F_n(c + \varepsilon/2) \rightarrow F(c + \varepsilon/2) = 1, \quad F_n(c - \varepsilon) \rightarrow F(c - \varepsilon) = 0$$

由此得

$$P(|X_n - c| \geq \varepsilon) \rightarrow 0$$

即 $X_n \xrightarrow{P} c$. 定理证毕. ■

■ **Example 2.1** 设随机变量 X 的分布列为

$$P(X = -1) = \frac{1}{2}, \quad P(X = 1) = \frac{1}{2}$$

令 $X_n = -X$, 则 X_n 与 X 同分布, 即 X_n 与 X 有相同的分布函数, 故 $X_n \xrightarrow{L} X$ 但对任意的 $0 < \varepsilon < 2$, 有

$$P(|X_n - X| \geq \varepsilon) = P(2|X| \geq \varepsilon) = 1 \rightarrow 0$$

即 X_n 不是依概率收敛于 X

Definition 2.1.2 — r 阶收敛. (r 阶收敛) 设对随机变量 ξ_n 及 ξ 有 $E|\xi_n|^r < \infty$ $E|\xi|^r < \infty$, 其中 $r > 0$ 为常数, 如果

$$\lim_{n \rightarrow \infty} E|\xi_n - \xi|^r = 0$$

则称 $\{\xi_n\}$ r 阶收敛 (convergence in r -order mean) 于 ξ , 并记为

$$\xi_n \xrightarrow{r} \xi$$

Theorem 2.1.9 — r 阶收敛与依概率收敛的关系.

$$\xi_n \xrightarrow{r} \xi \Rightarrow \xi_n \xrightarrow{P} \xi$$

Proof. 先证对于任意 $\varepsilon > 0$, 成立

$$P\{|\xi_n - \xi| \geq \varepsilon\} \leq \frac{E|\xi_n - \xi|^r}{\varepsilon^r}$$

事实上, 若以 $F(x)$ 记 $\xi_n - \xi$ 的分布函数, 则仿切比雪夫不等式的证明可得

$$\begin{aligned} P\{|\xi_n - \xi| \geq \varepsilon\} &= \int_{|x| \geq \varepsilon} dF(x) \\ &\leq \int_{|x| \geq \varepsilon} \frac{|x|^r}{\varepsilon^r} dF(x) \leq \frac{1}{\varepsilon^r} \int_{-\infty}^{\infty} |x|^r dF(x) \\ &= \frac{E|\xi_n - \xi|^r}{\varepsilon^r} \end{aligned}$$

不等式是切比雪夫不等式的推广, 通常称作马尔可夫不等式



$r = 2$ 为均方收敛

■ **Example 2.2** 举例子说明依概率收敛推不出 r 阶收敛: 取 $\Omega = (0, 1]$, \mathcal{F} 为 $(0, 1]$ 中博雷尔点集全体所构成的 σ 域, P 为勒贝格测度. 定义 $\xi(\omega) \equiv 0$ 及

$$\xi_n(\omega) = \begin{cases} n^{1/r}, & 0 < \omega \leq \frac{1}{n} \\ 0, & \frac{1}{n} < \omega \leq 1 \end{cases}$$

显然对一切 $\omega \in \Omega$, $\xi_n(\omega) \rightarrow \xi(\omega)$, 又对于任意的 $\varepsilon > 0$

$$P\{|\xi_n(\omega) - \xi(\omega)| \geq \varepsilon\} \leq \frac{1}{n}$$

因此 $\xi_n \xrightarrow{P} \xi$, 但是

$$E|\xi_n - \xi|^r = \left(n^{1/r}\right)^r \cdot \frac{1}{n} = 1$$

Definition 2.1.3 — 以概率 1 收敛. 如果

$$P \left\{ \lim_{n \rightarrow \infty} \xi_n(\omega) = \xi(\omega) \right\} = 1$$

则称 $\{\xi_n(\omega)\}$ 以概率 1 收敛 (convergence in probability 1) 于 $\xi(\omega)$ 又称 $\{\xi_n(\omega)\}$ 几乎处处收敛于 $\xi(\omega)$, 记为 $\xi_n(\omega) \xrightarrow{\text{a.s.}} \xi(\omega)$

Theorem 2.1.10 — 博雷尔-康特立引理. 1. 若随机事件序列 $\{A_n\}$ 满足

$$\sum_{n=1}^{\infty} P(A_n) < \infty$$

则

$$P \left\{ \overline{\lim}_{n \rightarrow \infty} A_n \right\} = 0, \quad P \left\{ \underline{\lim}_{n \rightarrow \infty} \bar{A}_n \right\} = 1$$

2. 若 $\{A_n\}$ 是相互独立的随机事件序列, 则

$$\sum_{n=1}^{\infty} P(A_n) = \infty$$

成立的充要条件为

$$P \left\{ \overline{\lim}_{n \rightarrow \infty} A_n \right\} = 1 \quad \text{或} \quad P \left\{ \underline{\lim}_{n \rightarrow \infty} \bar{A}_n \right\} = 0$$

Theorem 2.1.11 若 $\xi_n(\omega) (n = 1, 2, \dots), \xi(\omega)$ 是随机变量, 则

$$\begin{aligned} & \{\omega : \lim_{n \rightarrow \infty} \xi_n(\omega) = \xi(\omega)\} \\ &= \{\omega : \bigcap_{m=1}^{\infty} \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} (|\xi_n(\omega) - \xi(\omega)| < \frac{1}{m})\} \end{aligned}$$

下面两个式子都表示了依概率 1 收敛

$$\begin{aligned} & P \left\{ \bigcap_{m=1}^{\infty} \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left(|\xi_n(\omega) - \xi(\omega)| < \frac{1}{m} \right) \right\} = 1 \\ & P \left\{ \bigcup_{m=1}^{\infty} \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \left(|\xi_n(\omega) - \xi(\omega)| \geq \frac{1}{m} \right) \right\} = 0 \end{aligned}$$

Theorem 2.1.12

$$\xi_n(\omega) \xrightarrow{\text{a.s.}} \xi(\omega) \Rightarrow \xi_n(\omega) \xrightarrow{P} \xi(\omega)$$

■ **Example 2.3** 举例说明依概率收敛推不出依概率 1 收敛: 取 $\Omega = (0, 1]$, \mathcal{F} 为 $(0, 1]$ 中博雷尔点集全体所构成的 σ 域, P 为勒贝格测度, 令

$$\eta_{ki}(\omega) = \begin{cases} 1, & \omega \in \left(\frac{i-1}{k}, \frac{i}{k} \right] \quad i = 1, 2, \dots, k \\ 0, & \omega \in \left(\frac{i-1}{k}, \frac{i}{k} \right] \quad k = 1, 2, \dots \end{cases}$$

定义

$$\begin{aligned}\xi_1(\omega) &= \eta_{11}(\omega), \xi_2(\omega) = \eta_{21}(\omega), \xi_3(\omega) = \eta_{22}(\omega) \\ \xi_4(\omega) &= \eta_{31}(\omega), \xi_5(\omega) = \eta_{32}(\omega), \dots\end{aligned}$$

一般 $\xi_n(\omega) = \eta_{ki}(\omega)$, 其中 $n = i + \frac{k(k-1)}{2}$, 这样定义的 $\{\xi_n(\omega)\}$ 是一列随机变量. 但对于任何一个 $\omega \in (0, 1]$, $\xi_n(\omega)$ 必有无限个 k, i 使其取值 0, 也有无限个 k, i 使其取值 1, 因此 $\{\xi_n(\omega)\}$ 不是以概率 1 收敛于 0. 但是另一方面, 对任意的 $\varepsilon > 0$

$$P\{|\eta_{ki}(\omega)| \geq \varepsilon\} \leq \frac{1}{k}$$

当 $n \rightarrow \infty$ 时, 由 $n = \frac{k(k-1)}{2} + i \leq \frac{k(k-1)}{2} + k$, 知道 $k \rightarrow \infty$, 因此

$$\lim_{n \rightarrow \infty} P\{|\xi_n(\omega)| \geq \varepsilon\} = \lim_{n \rightarrow \infty} P\{|\eta_{ki}(\omega)| \geq \varepsilon\} = 0$$

所以 $\{\xi_n(\omega)\}$ 依概率收敛于 0.



不难验证, $\{\xi_n(\omega)\}$ 是 r 阶收敛于 0 的, 上面的例子提供了 r 阶收敛推不出以概率 1 收敛的例子.

Theorem 2.1.13 强大数定理和弱大数定理之间的差别是: 一个以概率 1 收敛, 一个以概率收敛

2.2 几种收敛之间的关系

almost uniformly (a.u.), written as $f_n \rightarrow_{\text{a. u.}} f$, if $\forall \delta > 0, \exists A \in \mathfrak{N}$, s.t. $\mu(A) < \delta$ and $f_n \rightarrow f$ uniformly on $\mathcal{X} \setminus A$ which is equivalent to

$$\forall \varepsilon > 0, \quad \lim_{m \rightarrow \infty} \mu \left(\bigcup_{n=m}^{\infty} \{|f_n - f| > \varepsilon\} \right) = 0$$

Theorem 2.2.1 1. 以分布收敛: $F_n(x) \rightarrow F(x)$, 与随机变量无关, x 和 $-x$ 的例子
2. 以概率收敛: $P(|X_n - X| \leq \varepsilon) \rightarrow 1$, 当 $n \rightarrow \infty$, 在乎的随机变量的值; in measure, written as $f_n \rightarrow_n f$, if

$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow \infty} \mu \{ |f_n - f| > \varepsilon \} = 0$$

3. 均方收敛: $E(X_n - X)^p \rightarrow 0$, 当 $n \rightarrow \infty$, $p \geq 1$ 在乎的也是随机变量的值
4. 几乎处处: $P(X_n \rightarrow X) = 1$, 当 $n \rightarrow \infty$ 在乎的也是随机变量的值; almost everywhere (a.e.), written as $f_n \rightarrow_{\text{a.e.}} f$, if

$$\mu \left\{ \lim_{n \rightarrow \infty} f_n \neq f \right\} = 0$$

which is equivalent to

$$\forall \varepsilon > 0, \quad \mu \left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{ |f_n - f| > \varepsilon \} \right) = 0$$

值相同了, 分布一定相同: $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$

2. 利用切比雪夫不等式: $P(|X - \mu| \geq \varepsilon) \leq \frac{E(X-\mu)^2}{\varepsilon^2}$. 因此 $X_n \xrightarrow{L^2} X \Rightarrow X_n \xrightarrow{P} X$
3. a.s. 和 L(k) 是最强的两种收敛, 依概率收敛次之, 依分布收敛再次。
4. 如果 almost sure convergence, 则一定有 convergence in probability, 如果 convergence in L(k), 则一定有 convergence in probability, 如果 convergence in probability, 则一定有 convergence in distribution。
5. almost sure convergence 和 convergence in L(k) 之间没有必然关系
6. 对于 convergence in L(k) 来说, 更大的 k 可以推出更小的 L(k) 收敛, 但反之不可
7. convergence in probability 也有等价的判定方法, 若任取 X_n 的一个子列, 都有这一子列的子列 almost sure 收敛 X, 则 X_n 依概率收敛于 X

Theorem 2.2.2 1. 一致收敛: $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ 关于 $x \in E$ 一致地成立.

2. 近一致收敛: 任给 $\delta > 0$, 存在 E 的可测子集 E_δ , 使在 E_δ 上 $\{f_n\}$ 一致收敛于 f , 而 $m(E \setminus E_\delta) < \delta$
3. 几乎处处收敛 $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ 对于几乎所有的 $x \in E$ 成立.
4. 测度收敛: 任给 $\varepsilon > 0$, $\lim_{n \rightarrow \infty} m\{x : |f_n(x) - f(x)| \geq \varepsilon\} = 0$
5. 平均收敛:

$$\lim_{n \rightarrow \infty} \int_E |f_n(x) - f(x)|^p dx = 0$$

6. 弱收敛: 当 $1 < p < +\infty$ 时, 对每个 $g \in L^q(E)$, $1/p + 1/q = 1$, 有

$$\lim_{n \rightarrow \infty} \int_E f_n(x) g(x) dx = \int_E f(x) g(x) dx$$

$p = 1$ 时, 对每个 $g \in L^\infty(E)$, 有

$$\lim_{n \rightarrow \infty} \int_E f_n(x) g(x) dx = \int_E f(x) g(x) dx$$

在测度有限的前提下

1. 1 → 2, 3, 5
2. 2 ⇔ 3
3. 2, 3, 5 → 4
4. 5 → 6

测度无限时

1. 1 → 2, 3
2. 2 → 3, 4
3. 5 → 4, 6

Theorem 2.2.3 — 马尔可夫 (Markov) 不等式. 对于任意非负随机变量 X

$$P(X > t) \leq \frac{E(X)}{t}, \forall t > 0$$

Proof.

$$E(X) = \int_0^\infty xf(x)dx = \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \geq \int_t^\infty xf(x)dx \geq t \int_t^\infty f(x)dx = tP(X > t)$$



Theorem 2.2.4 — 切比雪夫 (Chebyshev) 不等式.

$$P(|x - \mu| \geq t) \leq \frac{\sigma^2}{t^2}, P(|Z| \geq k) \leq \frac{1}{k^2}$$

其中, $Z = \frac{x-\mu}{\sigma}$

Proof.

$$P(|X - \mu| \geq t) = P(|X - \mu|^2 \geq t^2) \leq \frac{E(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}$$

令 $t = k\sigma$ 则第二个不等式得证。 ■

2.3 特征函数

Definition 2.3.1 — 特征函数的定义. 设 X 是一个随机变量, 称 $\varphi(t) = E(e^{itx})$ 为 X 的特征函数, 其表达式如下

$$\varphi(t) = E(e^{itx}) = \begin{cases} \sum_i e^{itx_i} P(X = x_i), & \text{在离散场合} \\ \int_{-\infty}^{+\infty} e^{itx} p_x(x) dx, & \text{在连续场合} \end{cases} \quad -\infty < t < +\infty$$

由于 $|e^{itx}| = \sqrt{\cos^2 tx + \sin^2 tx} = 1$, 所以随机变量 X 的特征函数 $\varphi(t)$ 总是存在的。利用欧拉公式 $e^{i\theta} = \cos \theta + i \sin \theta$ 可把不少问题中的复变函数问题转化为实变函数问题进行处理. 上述由密度函数 $p_x(x)$ 求其特征函数的公式常称傅里叶变换.

Proposition 2.3.1 — 特征函数的性质. 1. $|\varphi(t)| \leq \varphi(0) = 1$

- 2. $\varphi(-t) = \overline{\varphi(t)}$, 其中 $\overline{\varphi(t)}$ 表示 $\varphi(t)$ 的共轭
- 3. 若 $Y = aX + b$, 其中 a, b 是常数, 则

$$\varphi_Y(t) = e^{ibt} \varphi_X(at)$$

- 4. 若 X 与 Y 是相互独立的随机变量, 则

$$\varphi_{x+y}(t) = \varphi_X(t) \cdot \varphi_Y(t)$$

- 5. 若 $E(X^l)$ 存在, 则 $\varphi_X(t)$ 可 l 次求导, 且对 $1 \leq k \leq l$, 有

$$\varphi^{(k)}(0) = i^k E(X^k)$$

故

$$E(X) = \frac{\varphi'(0)}{i}, \quad \text{Var}(X) = -\varphi''(0) + (\varphi'(0))^2$$

Proof. 因为 $E(X^l)$ 存在, 也就是

$$\int_{-\infty}^{\infty} |x|^l p(x) dx < \infty$$

于是含参变量 t 的广义积分 $\int_{-\infty}^{\infty} e^{itx} p(x) dx$ 可以对 t 求导 l 次, 于是对 $0 \leq k \leq l$ 有

$$\varphi^{(k)}(t) = \int_{-\infty}^{\infty} i^k x^k e^{ix} p(x) dx = i^k E(X^k e^{ix})$$

令 $t = 0$ 即得

$$\varphi^{(k)}(0) = i^k E(X^k)$$



6. 一致连续性 特征函数 $\varphi(t)$ 在 $(-\infty, +\infty)$ 上一致连续
 7. 非负定性 特征函数 $\varphi(t)$ 是非负定的, 即对任意正整数 n , 及 n 个实数 t_1, t_2, \dots, t_n 和 n 个复数 z_1, z_2, \dots, z_n , 有

$$\sum_{k=1}^n \sum_{j=1}^n \varphi(t_k - t_j) z_k \bar{z}_j \geq 0$$

8. 逆转公式: 设 $F(x)$ 和 $\varphi(t)$ 分别为 X 的分布函数和特征函数, 则对 $F(x)$ 的任意两个连续点 $x_1 < x_2$, 有

$$F(x_2) - F(x_1) = \lim_{T \rightarrow +\infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-itx_1} - e^{-itx_2}}{it} \varphi(t) dt$$

9. 唯一性定理: 随机变量的分布函数由其特征函数唯一决定
 10. 若连续随机变量 X 的密度函数为 $p(x)$, 特征函数为 $\varphi(t)$. 如果 $\int_{-\infty}^{+\infty} |\varphi(t)| dt < +\infty$

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi(t) dt$$

这个公式又称傅里叶逆变换。

(R)

$$\begin{aligned}\varphi(t) &= \int_{-\infty}^{\infty} e^{itx} p(x) dx \\ p(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt\end{aligned}$$

Theorem 2.3.2 若随机变量 ξ 有 n 阶矩存在, 则它的特征函数可作如下展开:

$$f(t) = 1 + (\text{i}t) E\xi + \frac{(\text{i}t)^2}{2!} E\xi^2 + \dots + \frac{(\text{i}t)^n}{n!} E\xi^n + o(t^n)$$

Theorem 2.3.3 分布函数序列 $\{F_n(x)\}$ 弱收敛于分布函数 $F(x)$ 的充要条件是 $\{F_n(x)\}$ 的特征函数序列 $\{\varphi_n(t)\}$ 收敛于 $F(x)$ 的特征函数 $\varphi(t)$

Theorem 2.3.4 — 正极限定理--莱维·克拉默定理. 设分布函数列 $\{F_n(x)\}$ 弱收敛于某一分布函数 $F(x)$, 则相应的特征函数列 $\{f_n(t)\}$ 收敛于特征函数 $f(t)$, 且在 t 的任一有限区间内收敛是一致的.

Theorem 2.3.5 — 逆极限定理. 设特征函数列 $\{f_n(t)\}$ 收敛于某一函数 $f(t)$, 且 $f(t)$ 在 $t = 0$ 连续, 则相应的分布函数列 $\{F_n(x)\}$ 弱收敛于某一分布函数 $F(x)$, 而且 $f(t)$ 是 $F(x)$ 的特征函数.

Theorem 2.3.6 — 常见的分布的特征函数. 1. 单点分布 $P(X = a) = 1$, 其特征函数为

$$\varphi(t) = e^{\text{i}ta}$$

2. 0-1 分布 $P(X=x) = p^x(1-p)^{1-x}$, $x=0,1$, 其特征函数为

$$\varphi(t) = p e^{it} + q$$

其中 $q = 1 - p$

3. 泊松分布 $P(\lambda)$ $P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $k=0,1,\dots$, 其特征函数为

$$\varphi(t) = \sum_{k=0}^{\infty} e^{ikt} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

4. 均匀分布 $U(a,b)$ 因为密度函数为

$$p(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{其他} \end{cases}$$

所以其特征函数为

$$\varphi(t) = \int_a^b \frac{e^{itx}}{b-a} dx = \frac{e^{ibt} - e^{iat}}{it(b-a)}$$

5. 标准正态分布 $N(0,1)$ 因为密度函数为

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$$

所以其特征函数为

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(itx)^n}{n!} e^{-\frac{x^2}{2}} dx \quad (2.1)$$

$$= \sum_{n=0}^{\infty} \frac{(it)^n}{n!} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n e^{-\frac{x^2}{2}} dx \right] \quad (2.2)$$

上式中方括号内正是标准正态分布的 n 阶矩 $E(X^n)$. 当 n 为奇数时 $E(X^n) = 0$ 当 n 为偶数时, 如 $n = 2m$ 时

$$E(X^n) = E(X^{2m}) = (2m-1)!! = \frac{(2m)!}{2^m \cdot m!}$$

代回原式, 可得标准正态分布的特征函数

$$\varphi(t) = \sum_{m=0}^{\infty} \frac{(it)^{2m}}{(2m)!} \cdot \frac{(2m)!}{2^m \cdot m!} = \sum_{m=0}^{\infty} \left(-\frac{t^2}{2} \right)^m \frac{1}{m!} = e^{-\frac{t^2}{2}}$$

6. 指数分布 $\text{Exp}(\lambda)$ 因为密度函数为

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

所以其特征函数为

$$\begin{aligned}\varphi(t) &= \int_0^\infty e^{itx} \lambda e^{-\lambda x} dx \\ &= \lambda \left\{ \int_0^\infty \cos(tx) e^{-\lambda x} dx + i \int_0^\infty \sin(tx) e^{-\lambda x} dx \right\} \\ &= \lambda \left\{ \frac{\lambda}{\lambda^2 + t^2} + i \frac{t}{\lambda^2 + t^2} \right\} = \left(1 - \frac{it}{\lambda} \right)^{-1}\end{aligned}$$

7. $b(n, p)$: $p_k = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n$

$$\varphi_Y(t) = (pe^{it} + q)^n$$

8. 正态分布 $N(\mu, \sigma^2)$ 设随机变量 $Y \sim N(\mu, \sigma^2)$, 则 $X = (Y - \mu)/\sigma \sim N(0, 1)$.

$$\varphi_X(t) = e^{-\frac{t^2}{2}}$$

所以由 $Y = \sigma X + \mu$ 得

$$\varphi_Y(t) = \varphi_{\sigma X + \mu}(t) = e^{i\mu t} \varphi_x(\sigma t) = \exp \left\{ i\mu t - \frac{\sigma^2 t^2}{2} \right\}$$

9. 伽玛分布 $Ga(n, \lambda)$ 设随机变量 $Y \sim Ga(n, \lambda)$, 则 $Y = X_1 + X_2 + \dots + X_n$, 其中 X_i 独立同分布, 且 $X_i \sim \text{Exp}(\lambda)$.

$$\varphi_{x_i}(t) = \left(1 - \frac{it}{\lambda} \right)^{-1}$$

所以由独立随机变量和的特征函数为特征函数的积, 得

$$\varphi_Y(t) = (\varphi_{X_1}(t))^n = \left(1 - \frac{it}{\lambda} \right)^{-n}$$

进一步, 当 α 为任一正实数时, 我们可得 $Ga(\alpha, \lambda)$ 分布的特征函数为

$$\varphi(t) = \left(1 - \frac{it}{\lambda} \right)^{-\alpha}$$

10. $\chi^2(n)$ 分布 因为 $\chi^2(n) = Ga(n/2, 1/2)$, 所以 $\chi^2(n)$ 分布的特征函数为

$$\varphi(t) = (1 - 2it)^{-n/2}$$

(R)

$$\begin{aligned}N(\mu, \sigma^2) : \varphi(t) &= \exp \left\{ i\mu t - \frac{\sigma^2 t^2}{2} \right\} \\ N(0, 1) : \varphi(t) &= e^{-t^2/2}\end{aligned}$$

Theorem 2.3.7 若 X_j 相互独立, 且 $X_j \sim N(\mu_j, \sigma_j^2), j = 1, 2, \dots, n$, 则

$$\sum_{j=1}^n X_j \sim N\left(\sum_{j=1}^n \mu_j, \sum_{j=1}^n \sigma_j^2\right)$$

■ **Example 2.4** 若 X_λ 服从参数为 λ 的泊松分布, 证明:

$$\lim_{\lambda \rightarrow \infty} P\left(\frac{X_\lambda - \lambda}{\sqrt{\lambda}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Proof. 证明 已知 X_λ 的特征函数为 $\varphi_\lambda(t) = \exp\{\lambda(e^{it} - 1)\}$, 故 $Y_\lambda = \frac{X_\lambda - \lambda}{\sqrt{\lambda}}$ 的特征函数为

$$g_\lambda(t) = \varphi_\lambda\left(\frac{t}{\sqrt{\lambda}}\right) \exp\{-i\sqrt{\lambda}t\} = \exp\left\{\lambda\left(e^{i\frac{t}{\sqrt{\lambda}}} - 1\right) - i\sqrt{\lambda}t\right\}$$

对任意的 t , 有

$$\exp\left\{i\frac{t}{\sqrt{\lambda}}\right\} = 1 + \frac{it}{\sqrt{\lambda}} - \frac{t^2}{2! \lambda} + o\left(\frac{1}{\lambda}\right)$$

于是

$$\lambda\left(e^{i\frac{t}{\sqrt{\lambda}}} - 1\right) - i\sqrt{\lambda}t = -\frac{t^2}{2} + \lambda \cdot o\left(\frac{1}{\lambda}\right) \rightarrow -\frac{t^2}{2}, \quad \lambda \rightarrow \infty$$

从而有

$$\lim_{\lambda \rightarrow \infty} g_\lambda(t) = e^{-t^2/2}$$

而 $e^{-t^2/2}$ 正是标准正态分布 $N(0, 1)$ 的特征函数. ■

Theorem 2.3.8 n 个由一切阶矩 $\{\mu_n\}_{n \geq 1}$ 可唯一决定概率分布的充分条件, 其中 $\mu_n = E(X^n), n = 1, 2, \dots$ 唯一决定的意思是两个相等则相等

1. 若

$$\overline{\lim}_{n \rightarrow \infty} \frac{m_n^{1/n}}{n} < \infty, m_n = E|X|^n$$

则矩 $\{\mu_n\}_{n \geq 1}$ 唯一决定概率分布

2. 若

$$\overline{\lim}_{n \rightarrow \infty} \frac{(\mu_{2n})^{1/2n}}{2n} < \infty$$

则矩 $\{\mu_n\}_{n \geq 1}$ 唯一决定概率分布.

3. 若

$$\sum_{n=0}^{\infty} \frac{1}{(\mu_{2n})^{1/2n}} = \infty$$

则矩 $\{\mu_n\}_{n \geq 1}$ 唯一决定概率分布

■ **Example 2.5** 对正态分布 $N(0, \sigma^2)$ 其各阶矩为

$$\mu_{2n-1} = 0, \quad \mu_{2n} = \frac{(2n)!}{2^n n!} \sigma^{2n}, \quad n = 1, 2, \dots$$

利用斯特林公式

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \exp\left\{-\frac{\theta_n}{12n}\right\}, \quad 0 < \theta_n < 1$$

可验证上述第二个充分条件成立. 故上述各阶矩是只有正态分布 $N(0, \sigma^2)$ 才有的矩。

Theorem 2.3.9 — 波赫纳尔-辛钦. 函数 $f(t)$ 是特征函数的充要条件是: $f(t)$ 非负定, 连续, 且 $f(0) = 1$

Definition 2.3.2 如果对任意的正整数 n 及复数 $\lambda_1, \lambda_2, \dots, \lambda_n$ 均有

$$\sum_{k=1}^n \sum_{j=1}^n C_{k-j} \lambda_k \bar{\lambda}_j \geq 0$$

则称复数列 $C_n (n = 0, \pm 1, \pm 2, \dots)$ 是非负定的.

Theorem 2.3.10 — 赫格洛茨. 数列 $C_n (n = 0, \pm 1, \pm 2, \dots)$ 可以表为

$$C_n = \int_{-\pi}^{\pi} e^{inx} dG(x)$$

的充要条件是它是非负定的, 其中 $G(x)$ 是 $[-\pi, \pi]$ 上有界、非降、左连续函数。

2.4 大数定律

大数定理描述的不是简单的数列极限的问题, 而是概率中的事件发生的可能性: 当 n 很大时, 事件 $\{s_n/n = 1\}$ 和 $\{s_n/n = 0\}$ 的概率都很微小

Definition 2.4.1 随机变量序列 $\{X_n\}$ 服从大数定律设 $\{X_n\}$ 为随机变量序列, 若对任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i) \right| < \varepsilon \right\} = 1$$

则称 $\{X_n\}$ 服从大数定律.

Theorem 2.4.1 — 伯努利大数定律. 设 μ_n 为 n 重伯努利试验中事件 A 发生的次数, p 为每次试验中 A 出现的概率, 则对任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{\mu_n}{n} - p \right| < \varepsilon \right\} = 1$$

这是第一个大数定律, 它表明: 事件发生的频率是依概率收敛于该事件的概率, 这就是“频率稳定于概率”的含义, 也是“用频率去估计概率”的依据。

Theorem 2.4.2 — (Chebychev's Inequality). Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$

$$P(g(X) \geq r) \leq \frac{Eg(X)}{r}$$

Proof.

$$\begin{aligned} Eg(X) &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \\ &\geq \int_{\{x:g(x) \geq r\}} g(x)f_X(x)dx \quad (g \text{ is nonnegative}) \\ &\geq r \int_{\{x:g(x) \geq r\}} f_X(x)dx \\ &= rP(g(X) \geq r) \end{aligned}$$

Rearranging now produces the desired inequality. ■

Proof. 因为 $s_n \sim b(n, p)$, 由切比雪夫不等式得

$$1 \geq P\left(\left|\frac{s_n}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{\text{Var}\left(\frac{s_n}{n}\right)}{\varepsilon^2} = 1 - \frac{p(1-p)}{n\varepsilon^2}$$

当 $n \rightarrow \infty$ 时, 上式右端趋于 1, 因此

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{s_n}{n} - p\right| < \varepsilon\right) = 1$$
■

Example 2.6 — 蒙特卡洛随机投点计算定积分面积. 对于一般区间 $[a, b]$ 上的定积分

$$J' = \int_a^b g(x)dx$$

作线性变换 $y = (x-a)/(b-a)$, 即可化成 $[0, 1]$ 区间上的积分. 进一步若 $c \leq g(x) \leq d$, 可令

$$f(y) = \frac{1}{d-c}[g(a + (b-a)y) - c]$$

则 $0 \leq f(y) \leq 1$. 此时有

$$J' = \int_a^b g(x)dx = S_0 \int_0^1 f(y)dy + c(b-a)$$

其中 $S_0 = (b-a)(d-c)$. 这说明以上用蒙特卡罗方法计算定积分方法带有普遍意义。

Definition 2.4.2 设有一随机变量序列 $\{X_n\}$, 假如它满足: 对任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E(X_i)\right| < \varepsilon\right) = 1$$

则称该随机变量序列 $\{X_n\}$ 服从大数定律.

Theorem 2.4.3 — 切比雪夫大数定律. 设 $\{X_n\}$ 为一列两两不相关的随机变量序列, 若每个 X_i 的方差存在, 且有共同的上界, 即 $\text{Var}(X_i) \leq c, i = 1, 2, \dots$, 则 $\{X_n\}$ 服从大数定律.

Proof.

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \leq \frac{c}{n}$$

■

(R) 没有要求同分布, 伯努利大数定律是切比雪夫大数定律的特例.

Theorem 2.4.4 — 泊松大数定律. 如果在一个独立试验序列中, 事件 A 在第 k 次试验中出现的概率等于 p_k , 以 μ_n 记在前 n 次试验中事件 A 出现的次数, 则对任意 $\epsilon > 0$, 都有

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{\mu_n}{n} - \frac{p_1 + p_2 + \dots + p_n}{n}\right| < \epsilon\right\} = 1$$

Theorem 2.4.5 — 马尔可夫大数定律. 对随机变量序列 $\{X_n\}$, 若有

$$\frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \rightarrow 0 \quad (n \rightarrow \infty)$$

则 $\{X_n\}$ 服从大数定律. 上式被称为马尔可夫条件.

(R) 马尔可夫大数定律的重要性在于: 对 $\{X_n\}$ 已经没有任何同分布, 独立性, 不相关的假定. 切比雪夫大数定律显然可由马尔可夫大数定律推出.

Theorem 2.4.6 — 辛钦大数定律. 设 $\{X_n\}$ 为一独立同分布的随机变量序列, 若 X_i 的数学期望存在, 则 $\{X_n\}$ 服从大数定律.

(R) 方差存在, 期望一定存在

Proof. 设 $\{X_n\}$ 独立同分布, 且 $E(X_i) = a, i = 1, 2, \dots$. 现在要证明

$$\frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{P} a, \quad n \rightarrow \infty$$

为此记

$$Y_n = \frac{1}{n} \sum_{k=1}^n X_k$$

由定理2.1.8 知, 只需证 $Y_n \xrightarrow{L} a$. 又由定理2.3.3 知, 只需证 $\varphi_{Y_n}(t) \rightarrow e^{iat}$ 因为 $\{X_n\}$ 同分布, 所以它们有相同的特征函数, 记这个特征函数为 $\varphi(t)$. 因为 $\varphi'(0)/i = E(X_i) = a$, 从而 $\varphi(t)$

在 0 点展开式为

$$\varphi(t) = \varphi(0) + \varphi'(0)t + o(t) = 1 + iat + o(t)$$

再由 $\{X_n\}$ 的独立性知 Y_n 的特征函数为

$$\varphi_{Y_n}(t) = \left[\varphi\left(\frac{t}{n}\right) \right]^n = \left[1 + ia\frac{t}{n} + o\left(\frac{1}{n}\right) \right]^n$$

对任意的 t 有

$$\lim_{n \rightarrow \infty} \left[\varphi\left(\frac{t}{n}\right) \right]^n = \lim_{n \rightarrow \infty} \left[1 + ia\frac{t}{n} + o\left(\frac{1}{n}\right) \right]^n = e^{iat}$$

而 e^{iat} 正是退化分布的特征函数，由此证得了 $Y_n \xrightarrow{P} a$. 至此定理得证. ■

(R)

辛钦大数定律提供了求随机变量数学期望 $E(X)$ 的近似值的方法. 设想对随机变量 X 独立重复地观察 n 次, 第 k 次观察值为 X_k , 则 X_1, X_2, \dots, X_n 应该是相互独立的, 且它们的分布应该与 X 的分布相同. 所以, 在 $E(X)$ 存在的条件下, 按照辛钦大数定律, 当 n 足够大时, 可以把平均观察值

$$\frac{1}{n} \sum_{i=1}^n X_i$$

作为 $E(X)$ 的近似值. 这样做法的一个优点是我们可以不必去管 X 的分布究竟是怎样的, 我们的目的只是寻求数学期望的近似值. 用观察值的平均去近似随机变量的均值.

Corollary 2.4.7 由辛钦大数定律我们很容易地得出: 如果 $\{X_n\}$ 为一独立同分布的随机变量序列, 且 $E\{|X_i|^k\}$ 存在, 其中 k 为正整数, 则 $\{X_n^k\}$ 服从大数定律, 所以可以将 $\frac{1}{n} \sum_{i=1}^n X_i^k$ 作为 $E(X_i^k)$ 的近似值

■ **Example 2.7 — (用蒙特卡罗方法计算定积分 (平均值法))**. 计算定积分

$$J = \int_0^1 f(x) dx$$

设随机变量 X 服从 $(0, 1)$ 上的均匀分布, 则 $Y = f(X)$ 的数学期望为

$$E(f(X)) = \int_0^1 f(x) dx = J$$

所以估计 J 的值就是估计 $f(X)$ 的数学期望的值. 由辛钦大数定律, 可以用 $f(X)$ 的观察值的平均去估计 $f(X)$ 的数学期望的值. 具体做法如下: 先用计算机产生 n 个 $(0, 1)$ 上均匀分布的随机数 $x_i, i = 1, 2, \dots, n$, 然后对每个 x_i 计算 $f(x_i)$ 最后得 J 的估计值为

$$J \approx \frac{1}{n} \sum_{i=1}^n f(x_i)$$

2.4.1 强大数定理

Theorem 2.4.8 — 博雷尔. 设 μ_n 是事件 A 在 n 次独立试验中的出现次数, 在每次试验中事件 A 出现的概率均为 p , 那么当 $n \rightarrow \infty$ 时,

$$P\left\{\frac{\mu_n}{n} \rightarrow p\right\} = 1$$

Theorem 2.4.9 — 强大数定理说明频率稳定于概率. 强大数定理的一个结论, 即当试验次数无限增加时, 频率将趋于概率, 博雷尔强大数定律正给出了这个结果. 从伯努利大数定律并不能引申出这个结论, 它只断言一个不等式 $|\frac{\mu_n}{n} - p| < \varepsilon$ 成立的概率可以大于 $1 - \eta$, 不论 η 是什么正数; 但是事件

$$\left|\frac{\mu_{n+1}}{n+1} - p\right| \geq \varepsilon, \left|\frac{\mu_{n+2}}{n+2} - p\right| \geq \varepsilon, \dots, \left|\frac{\mu_{2n}}{2n} - p\right| \geq \varepsilon, \dots$$

中至少有一个发生仍是可能的, 因为它是可列个事件之并, 而我们只知道每个事件的概率很小. 但博雷尔强大数定律则断言 $|\frac{\mu_n}{n} - p|$ 以概率 1 变得很小, 而且保持很小. 虽然从逻辑上讲, 在投硬币时每次都出现正面是可能的, 这时 $\frac{\mu_n}{n} = 1$, 因而 $\frac{\mu_n}{n} \rightarrow p$ 并不成立, 但是强大数定律断言了这种事件发生的概率为 0.

Definition 2.4.3 设 $\{\xi_i\}$ 是独立随机变量序列, 若

$$P\left\{\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\xi_i - E\xi_i) = 0\right\} = 1$$

则称它满足强大数定律.

Theorem 2.4.10 — 噶依克-瑞尼不等式. 若 $\{\xi_i\}$ 是独立随机变量序列, $D\xi_i = \sigma_i^2 < \infty$, ($i = 1, 2, \dots$), 而 $\{C_n\}$ 是一列正的非增常数序列, 则对任意正整数 m, n ($m < n$) $\varepsilon > 0$, 均有

$$\begin{aligned} P\left\{\max_{m \leq j \leq n} C_j \left| \sum_{i=1}^j (\xi_i - E\xi_i) \right| \geq \varepsilon\right\} \\ \leq \frac{1}{\varepsilon^2} \left(C_m^2 \sum_{j=1}^m \sigma_j^2 + \sum_{j=m+1}^n C_j^2 \sigma_j^2 \right) \end{aligned}$$

Theorem 2.4.11 — 科尔莫戈罗夫不等式. 设 $\xi_1, \xi_2, \dots, \xi_n$ 是独立随机变量, 方差有限, 则对任意 $\varepsilon > 0$, 成立

$$P\left\{\max_{1 \leq j \leq n} \left| \sum_{i=1}^j (\xi_i - E\xi_i) \right| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2} \sum_{j=1}^n D\xi_j$$

Theorem 2.4.12 — 科尔莫戈罗夫强大数定律. 设 $\{\xi_i\}, i = 1, 2, \dots$ 是独立随机变量序列, 且 $\sum_{n=1}^{\infty} \frac{D\xi_n}{n^2} < \infty$, 则成立

$$P\left\{\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\xi_i - E\xi_i) = 0\right\} = 1$$

Theorem 2.4.13 — 独立同分布场合的强大数定律--科尔莫戈罗夫. 设 ξ_1, ξ_2, \dots 是相互独立相同分布的随机变量序列, 则

$$\frac{1}{n}(\xi_1 + \xi_2 + \dots + \xi_n) \xrightarrow{a.s.} a$$

成立的充要条件是 $E\xi_i$ 存在且等于 a

2.5 中心极限定理

Definition 2.5.1 中心极限定理研究独立随机变量和的极限分布在什么条件下为正态分布的命题. 大数定理是给出频率会趋于概率, 中心极限定理是给出事件发生次数的渐进分布

Theorem 2.5.1 — 林德伯格—莱维中心极限定理. 设 $\{X_n\}$ 是独立同分布的随机变量序列, 且 $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2 > 0$. 记

$$Y_n^* = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

则对任意实数 y , 有

$$\lim_{n \rightarrow \infty} P(Y_n^* \leq y) = \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-t^2/2} dt$$

Proof. 设 $X_n - \mu$ 的特征函数为 $\varphi(t)$, 则 Y_n^* 的特征函数为

$$\varphi_{Y_n}(t) = \left[\varphi \left(\frac{t}{\sigma\sqrt{n}} \right) \right]^n$$

又因为 $E(X_n - \mu) = 0$, $\text{Var}(X_n - \mu) = \sigma^2$, 所以有

$$\varphi'(0) = 0, \quad \varphi''(0) = -\sigma^2$$

于是特征函数 $\varphi(t)$ 有展开式

$$\begin{aligned} \varphi(t) &= \varphi(0) + \varphi'(0)t + \varphi''(0)\frac{t^2}{2} + o(t^2) \\ &= 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2) \end{aligned}$$

从而有

$$\lim_{n \rightarrow \infty} \varphi_{Y_n^*}(t) = \lim_{n \rightarrow \infty} \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right]^n = e^{-t^2/2}$$

而 $e^{-t^2/2}$ 正是 $N(0, 1)$ 分布的特征函数, 定理得证. ■



标准化后的随机变量趋于标准正态



林德贝格-莱维定理有广泛应用: 在实际工作中, 只要 n 足够大, 便可以把独立同分布的随机变量之和当作是正态变量.

■ **Example 2.8** 设随机变量 X 服从 $(0, 1)$ 上的均匀分布, 则其数学期望与方差分别为 $1/2$ 和 $1/12$. 由此得 12 个相互独立的 $(0, 1)$ 上均匀分布随机变量和的数学期望与方差分别为 6 和 1. 因此我们可以如下产生正态分布 $N(\mu, \sigma^2)$ 的随机数

1. 从计算机中产生 12 个 $(0, 1)$ 上均匀分布的随机数, 记为 x_1, x_2, \dots, x_{12}
2. 计算 $y = x_1 + x_2 + \dots + x_{12} - 6$, 则由林德伯格 – 莱维中心极限定理知, 可将 y 近似看成来自标准正态分布 $N(0, 1)$ 的一个随机数.
3. 计算 $z = \mu + \sigma y$, 则可将 z 看成来自正态分布 $N(\mu, \sigma^2)$ 的一个随机数
4. 重复 (1) – (3) n 次, 就可得到 $N(\mu, \sigma^2)$ 分布的 n 个随机数. 从这个例子可以看出, 由 12 个均匀分布的随机数得到 1 个正态分布的随机数是利用了林德伯格-莱维中心极限定理。

中心极限定理的两个应用: 正态随机数和误差估计

■ **Example 2.9** (数值计算中的误差分析) 在数值计算中, 任何实数 x 都只能用定位数的小数 x 来近似. 现在如果要求 n 个实数 $x_i (i = 1, 2, \dots, n)$ 的和 S , 在数值计算中, 只能用 x_i 的近似数 x'_i 来得到 S 的近似数 S' , 记个别误差为 $\varepsilon_i = x_i - x'_i$, 则总误差为

$$S - S' = \sum_{i=1}^n x_i - \sum_{i=1}^n x'_i = \sum_{i=1}^n \varepsilon_i$$

若在数值计算中, 取 k 位小数, 则可认为 ε_i 服从区间 $(-0.5 \times 10^{-k}, 0.5 \times 10^{-k})$ 上的均匀分布, 且相互独立. 下面我们来估计总误差. 一种粗略的估计方法是: 由于 $|\varepsilon_i| \leq 0.5 \times 10^{-k}$, 所以

$$\left| \sum_{i=1}^n \varepsilon_i \right| \leq n \times 0.5 \times 10^{-k}$$

现在用中心极限定理来估计: 因为 $\{\varepsilon_i\}$ 居立同分布, 且

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \frac{10^{-2k}}{12}$$

因此对总误差有

$$E\left(\sum_{i=1}^n \varepsilon_i\right) = 0, \quad \text{Var}\left(\sum_{i=1}^n \varepsilon_i\right) = \frac{n10^{-2k}}{12}$$

由林德伯格-莱维中心极限定理知, 对任意的 $z > 0$, 有

$$\begin{aligned} P\left(\left|\sum_{i=1}^n \varepsilon_i\right| \leq z\right) &\approx \Phi\left(\frac{z\sqrt{12}}{\sqrt{n10^{-2k}}}\right) - \Phi\left(-\frac{z\sqrt{12}}{\sqrt{n10^{-2k}}}\right) \\ &= 2\Phi\left(\frac{z\sqrt{12}}{\sqrt{n10^{-2k}}}\right) - 1 \end{aligned}$$

要从上式中求出总误差的上限 z , 可令上式右边的概率为 0.99, 由此得

$$\Phi\left(\frac{z\sqrt{12}}{\sqrt{n10^{-2k}}}\right) = 0.995$$

再查标准正态分布函数的 0.995 分位数得

$$\frac{z\sqrt{12}}{\sqrt{n}10^{-2k}} = 2.575$$

由此解得

$$\begin{aligned} z &= \frac{2.575\sqrt{n \times 10^{-2k}}}{\sqrt{12}} = 0.7433 \times \sqrt{n \times 10^{-2k}} \\ &= 0.7433 \times \sqrt{n} \times 10^{-k} \end{aligned}$$

也就是我们有 99% 的把握程度, 可以说

$$\left| \sum_{i=1}^n \varepsilon_i \right| \leq 0.7433 \times \sqrt{n} \times 10^{-k}$$

Theorem 2.5.2 — 棣莫弗-拉普拉斯. 若 μ_n 是 n 次伯努利试验中事件 A 出现的次数, $0 < p < 1$, 则对任意有限区间 $[a, b]$: (i) 当 $a \leq x_k \equiv \frac{k-np}{\sqrt{npq}} \leq b$ 及 $n \rightarrow \infty$ 时, 一致地有

$$P\{\mu_n = k\} \div \left(\frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x_k^2} \right) \rightarrow 1$$

(ii) 当 $n \rightarrow \infty$ 时, 一致地有

$$P\left\{ a \leq \frac{\mu_n - np}{\sqrt{npq}} < b \right\} \rightarrow \int_a^b \varphi(x) dx$$

$$\text{其中 } \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} (-\infty < x < \infty)$$

Theorem 2.5.3 — 二项分布的正态近似--棣莫弗-拉普拉斯中心极限定理. 设 n 重伯努利试验中, 事件 A 在每次试验中出现的概率为 $p (0 < p < 1)$, 记 μ_n 为 n 次试验中事件 A 出现的次数, 且记

$$Y_n^* = \frac{\mu_n - np}{\sqrt{np(1-p)}}$$

则对任意实数 y , 有

$$\lim_{n \rightarrow +\infty} P(Y_n^* \leq y) = \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-t^2/2} dt$$

R 与 (泊松定理)“二项分布的泊松近似”相比, 一般在 p 较小时, 用泊松分布近似较好; 而在 $np > 5$ 和 $n(1-p) > 5$ 时, 用正态分布近似较好

Theorem 2.5.4 — 近似中的修正. 因为二项分布是离散分布, 而正态分布是连续分布, 所以用正态分布作为二项分布的近似计算中, 作些修正可以提高精度. 若 $k_1 < k_2$ 均为整数,

一般先作如下修正后再用正态近似

$$\begin{aligned} P(k_1 \leq \mu_n \leq k_2) &= P(k_1 - 0.5 < \mu_n < k_2 + 0.5) \\ &= \Phi\left(\frac{k_2 + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Theorem 2.5.5 — 三类近似计算问题. 若记 $\beta = \Phi(y)$, 则由中心极限定理给出的近似式 $P(Y_n^* \leq y) \approx \Phi(y) = \beta$ 可用来解三类计算问题:

1. 已知 n, y 求 β (求概率)
2. 已知 n, β 求 y (求分位数)
3. 已知 y, β 求 n (求样本量)

Theorem 2.5.6 — 多元中心极限定理. 若 p 维随机向量 $\xi_1, \xi_2, \dots, \xi_n, \dots$ 相互独立, 具有相同的分布, 其数学期望为 μ , 协方差阵为 Σ , 则

$$\eta_n = \{(\xi_1 - \mu) + (\xi_2 - \mu) + \dots + (\xi_n - \mu)\} / \sqrt{n}$$

的极限分布为 $N(\mathbf{0}, \Sigma)$

■ **Example 2.10** 中多项分布的随机向量, 可以看作 n 个相互独立相同分布随机向量之和, 由定理多元正态中心极限定理可知多项分布渐近于正态分布, 真正维数为 $r-1$ 。

独立不同分布下的中心极限定理

为了使极限分布是正态分布, 还要求各个加项“均匀地小”--林条件

Theorem 2.5.7 — 独立不同分布下的中心极限定理. 设 $\{X_n\}$ 为独立随机变量序列, 且 $E(X_i) = \mu_i, \text{Var}(X_i) = \sigma_i^2, i = 1, 2, \dots$,

$$\text{记 } Y_n = \sum_{i=1}^n X_i, \quad B_n = \sqrt{\text{Var}(Y_n)} = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}$$

1. 林德伯格条件: 若诸 X_i 为连续随机变量, 其密度函数为 $p_i(x)$, 对任意的 $\tau > 0$, 称

$$\lim_{n \rightarrow \infty} \frac{1}{\tau^2 B_n^2} \sum_{i=1}^n \int_{|x-\mu_i| > \tau B_n} (x - \mu_i)^2 p_i(x) dx = 0$$

为林德伯格条件。

2. 林德伯格中心极限定理: 设独立随机变量序列 $\{X_n\}$ 满足林德伯格条件, 则对任意的 x , 有

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{B_n} \sum_{i=1}^n (X_i - \mu_i) \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

3. 假如独立随机变量序列 $\{X_n\}$, 具有同分布和方差有限的条件, 则必定满足林德伯格条件, 即林德伯格—莱维中心极限定理是林德伯格中心极限定理的特例.
4. 李雅普诺夫中心极限定理设 $\{X_n\}$ 为独立随机变量序列, 若存在 $\delta > 0$ 满足

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^{2+\delta}} \sum_{i=1}^n E(|X_i - \mu_i|^{2+\delta}) = 0$$

则对任意的 x , 有

$$\lim_{n \rightarrow \infty} P \left\{ \frac{1}{B_n} \sum_{i=1}^n (X_i - \mu_i) \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Proof. 设 $\{X_n\}$ 是一个相互独立的随机变量序列, 它们具有有限的数学期望和方差

$$E(X_i) = \mu_i, \quad \text{Var}(X_i) = \sigma_i^2, \quad i = 1, 2, \dots$$

要讨论随机变量的和 $Y_n = \sum_{i=1}^n X_i$, 我们先将其标准化, 即将它减去均值、除以标准差, 由于

$$\begin{aligned} E(Y_n) &= \mu_1 + \mu_2 + \dots + \mu_n \\ \sigma(Y_n) &= \sqrt{\text{Var}(Y_n)} = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2} \end{aligned}$$

且记 $\sigma(Y_n) = B_n$, 则 Y_n 的标准化为

$$Y_n^* = \frac{Y_n - (\mu_1 + \mu_2 + \dots + \mu_n)}{B_n} = \sum_{i=1}^n \frac{X_i - \mu_i}{B_n}$$

如果要求 Y_n^* 中各项 $\frac{|X_i - \mu_i|}{B_n}$ “均匀地小”, 即对任意的 $\tau > 0$, 要求事件

$$A_{ni} = \left\{ \frac{|X_i - \mu_i|}{B_n} > \tau \right\} = \{|X_i - \mu_i| > \tau B_n\}$$

发生的可能性小, 或直接要求其概率趋于 0. 为达到这个目的, 我们要求

$$\lim_{n \rightarrow \infty} P \left(\max_{1 \leq i \leq n} |X_i - \mu_i| > \tau B_n \right) = 0$$

因为

$$\begin{aligned} P \left(\max_{1 \leq i \leq n} |X_i - \mu_i| > \tau B_n \right) &= P \left(\bigcup_{i=1}^n (|X_i - \mu_i| > \tau B_n) \right) \\ &\leq \sum_{i=1}^n P(|X_i - \mu_i| > \tau B_n) \end{aligned}$$

若设诸 X_i 为连续随机变量, 其密度函数为 $p_i(x)$, 则

$$\begin{aligned} \text{上式右边} &= \sum_{i=1}^n \int_{|x - \mu_i| > \tau B_n} p_i(x) dx \\ &\leq \frac{1}{\tau^2 B_n^2} \sum_{i=1}^n \int_{|x - \mu_i| > \tau B_n} (x - \mu_i)^2 p_i(x) dx \end{aligned}$$

因此, 只要对任意的 $\tau > 0$, 有

$$\lim_{n \rightarrow \infty} \frac{1}{\tau^2 B_n^2} \sum_{i=1}^n \int_{|x - \mu_i| > \tau B_n} (x - \mu_i)^2 p_i(x) dx = 0$$

就可保证 Y_n^* 中各加项“均匀地小”, 上述称为林德伯格条件. ■

Proof. 设 $\{X_n\}$ 是独立同分布的随机变量序列, 为确定起见, 设诸 X_n 显连续随机变量, 其共同的密度函数为 $p(x), \mu_i = \mu, \sigma_i = \sigma$. 这时 $B_n = \sigma\sqrt{n}$, 由此得

$$\begin{aligned} & \frac{1}{B_n^2} \sum_{i=1}^n \int_{|x-\mu|>rB_n} (x-\mu_i)^2 p(x) dx \\ &= \frac{n}{n\sigma^2} \int_{|x-\mu|>\tau\sigma/\sqrt{n}} (x-\mu)^2 p(x) dx \end{aligned}$$

因为方差存在, 即

$$\text{Var}(X_i) = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx < \infty$$

所以其尾部积分一定有

$$\lim_{n \rightarrow \infty} \int_{|x-\mu|>+\infty\sqrt{n}} (x-\mu)^2 p(x) dx = 0$$

故林德伯格条件满足。 ■

■ **Example 2.11** 例 4. 4.9 一份考卷由 99 个题目组成, 并按由易到难顺序排列. 某学生答对第 1 题的概率为 0.99, 答对第 2 题的概率为 0.98. 一般地, 他答对第 i 题的概率为 $1 - i/100, i = 1, 2, \dots$, 假如该学生同答各题目是相互独立的, 并且要正确回答其中 60 个以上 (包括 60 个) 题目才算通过考试. 试计算该学生通过考试的可能性多大? 解设

$$X_i = \begin{cases} 1, & \text{若学生答对第 } i \text{ 题,} \\ 0, & \text{若学生答错第 } i \text{ 题.} \end{cases}$$

于是诸 X_i 相互独立, 且服从不同的二点分布

$$P(X_i = 1) = p_i = 1 - \frac{i}{100}, \quad P(X_i = 0) = 1 - p_i = \frac{i}{100} \quad i = 1, 2, \dots, 99$$

而我们要求的是

$$P\left(\sum_{i=1}^{99} X_i \geq 60\right)$$

为使用中心极限定理, 我们可以设想从 X_{100} 开始的随机变量都与 X_{99} , 同分布且相互独立. 下面我们用 $\delta = 1$ 来验证随机变量序列 $\{X_n\}$ 满足李雅普诺夫条件, 因为

$$\begin{aligned} B_n &= \sqrt{\sum_{i=1}^n \text{Var}(X_i)} = \sqrt{\sum_{i=1}^n p_i(1-p_i)} \rightarrow \infty (n \rightarrow \infty) \\ E(|X_i - p_i|^3) &= (1-p_i)^3 p_i + p_i^3 (1-p_i) \leq p_i(1-p_i) \end{aligned}$$

于是

$$\frac{1}{B_n^3} \sum_{i=1}^n E(|X_i - p_i|^3) \leq \frac{1}{[\sum_{i=1}^n p_i(1-p_i)]^{1/2}} \rightarrow 0 (n \rightarrow \infty)$$

即 $\{X_n\}$ 满足李雅普诺夫条件, 所以可以使用中心极限定理.

Theorem 2.5.8 — 费勒条件. 林德贝格条件不是中心极限定理成立的必要条件. 不过, 费勒进一步指出, 假如下面条件得到满足:

$$\lim_{n \rightarrow \infty} \max_{k \leq n} \frac{b_k}{B_n} = 0$$

则林德贝格条件也是中心极限定理成立的必要条件。

Theorem 2.5.9 — 费勒条件等价条件.

$$\begin{aligned}\lim_{n \rightarrow \infty} B_n &= \infty \\ \lim_{n \rightarrow \infty} \frac{b_n}{B_n} &= 0\end{aligned}$$

Theorem 2.5.10 对 $n = 1, 2, \dots$ 及任意的 t

$$\left| e^{it} - 1 - \frac{it}{1!} - \dots - \frac{(it)^{n-1}}{(n-1)!} \right| \leq |t|^n$$

Theorem 2.5.11 对于任何满足 $|a_k| \leq 1$ 及 $|b_k| \leq 1$ ($k = 1, 2, \dots, n$) 的复数, 有

$$|a_1 a_2 \cdots a_n - b_1 b_2 \cdots b_n| \leq \sum_{k=1}^n |a_k - b_k|$$

Proof.

$$a_1 a_2 - b_1 b_2 = (a_1 - b_1) a_2 + (a_2 - b_2) b_1$$

因此

$$|a_1 a_2 - b_1 b_2| \leq |a_1 - b_1| + |a_2 - b_2|$$

用归纳法即得 ■

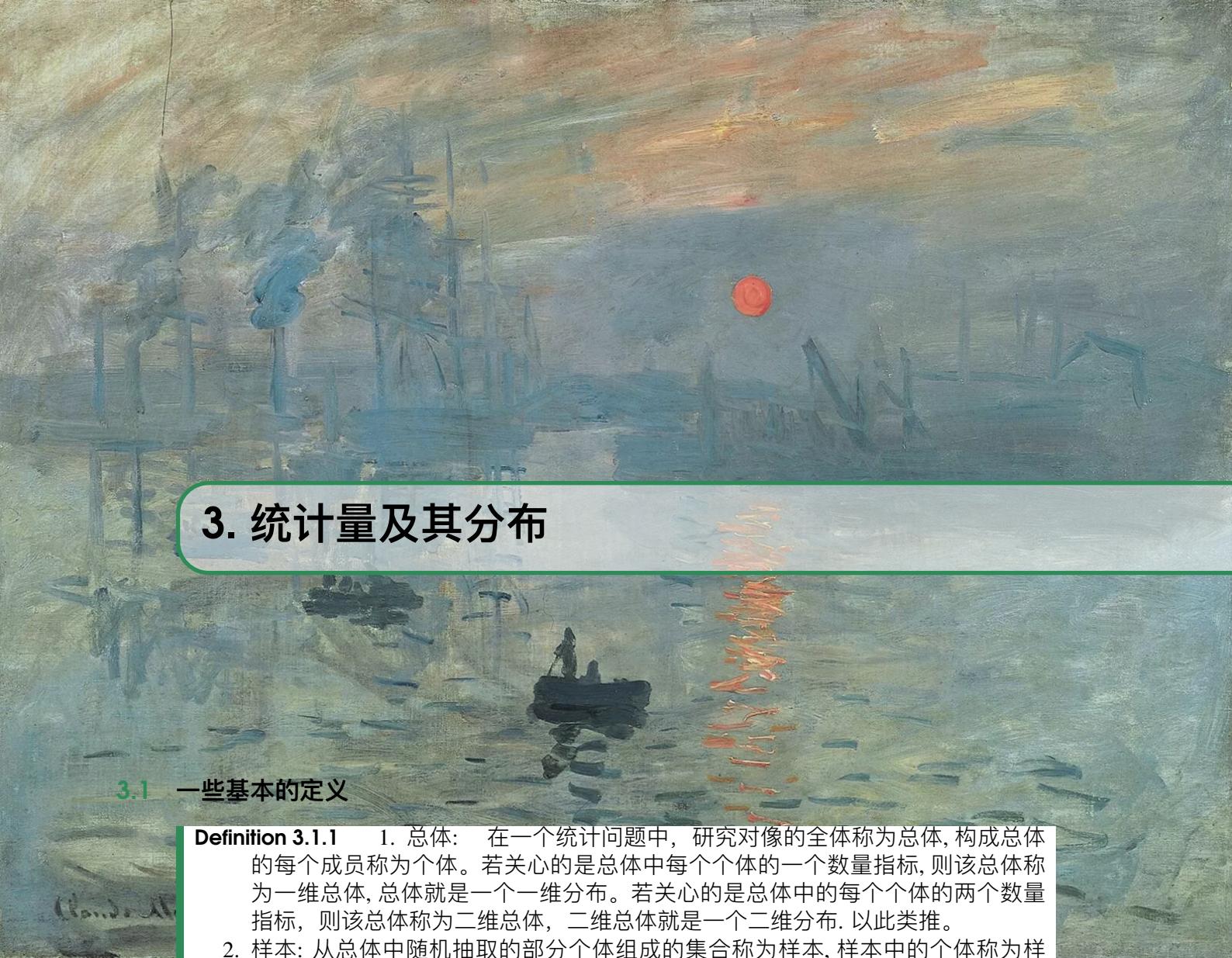
Theorem 2.5.12 若 $\varphi(t)$ 是特征函数, 则 $e^{\varphi(t)-1}$ 也是特征函数, 特别地

$$\left| e^{\varphi(t)-1} \right| \leq 1$$

Theorem 2.5.13 若 ξ_1, ξ_2, \dots 是独立随机变量序列, 存在常数 K_n , 使 $\max_{1 \leq j \leq n} |\xi_j| \leq K_n$ ($n = 1, 2, \dots$), 且 $\lim_{n \rightarrow \infty} \frac{K_n}{B_n} = 0$, 则

$$P \left\{ \sum_{k=1}^n \frac{\xi_k - a_k}{B_n} < x \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

Theorem 2.5.14 伯努利大数定律是用矩法证明的; 棣莫弗-拉普拉斯定理则通过利用斯特灵公式进行渐近估计而得到。



3. 统计量及其分布

3.1 一些基本的定义

- Definition 3.1.1** 1. 总体: 在一个统计问题中, 研究对象的全体称为总体, 构成总体的每个成员称为个体。若关心的是总体中每个个体的一个数量指标, 则该总体称为一维总体, 总体就是一个一维分布。若关心的是总体中的每个个体的两个数量指标, 则该总体称为二维总体, 二维总体就是一个二维分布. 以此类推。
2. 样本: 从总体中随机抽取的部分个体组成的集合称为样本, 样本中的个体称为样品, 样品个数称为样本容量或样本量。样本常用 n 个指标值 x_1, x_2, \dots, x_n 表示. 它可看作 n 维随机变量, 又可看作其观察值。
3. 既然样本是随机变量, 就有概率分布, 于是, 这个概率分布就称为样本分布。
4. 样本具有所谓的二重性: 一方面, 由于样本是从总体中随机抽取的, 抽取前无法预知它们的数值. 因此, 样本是随机变量, 用大写字母 X, X_2, \dots, X_n 表示; 另一方面, 样本在抽取以后经观测就有确定的观测值, 因此, 样本又是一组数值. 此时用小写字母 x_1, x_2, \dots, x_n 表示是恰当的.
5. 分组样本: 只知样本观测值所在区间, 而不知具体值的样本称为分组样本。缺点: 与完全样本相比损失部分信息。优点: 在样本量较大时, 用分组样本既简明扼要, 又能帮助人们更好地认识总体
6. “简单随机抽样”有如下两个要求:
- 样本具有随机性, 即要求总体中每一个个体都有同等机会被选入样本, 这便意味着每一样品 x_i 与总体 X 有相同的分布。
 - 样本要有独立性, 即要求样本中每一样品的取值不影响其他样品的取值, 这意味着 x_1, x_2, \dots, x_n 相互独立. 用简单随机抽样方法得到的样本称为简单随机样本, 也简称样本。
7. 简单随机样本: 若样本 x_1, x_2, \dots, x_n 是 n 个相互独立的具有同一分布 (总体分布) 的随机变量, 则称该样本为简单随机样本, 仍简称样本。
8. 若总体的分布函数为 $F(x)$, 则其样本的(联合)分布函数为 $\prod_{i=1}^n F(x_i)$
9. 若总体的密度函数为 $p(x)$, 则其样本的(联合)密度函数为 $\prod_{i=1}^n p(x_i)$
10. 若总体的分布列为 $\{p(x_i)\}$, 则其样本的(联合)分布列为 $\prod_{i=1}^n p(x_i)$

Theorem 3.1.1 — 科大概统--统计部分. 1. 非参数总体：总体分布不能通过若干个未知参数表达出来
2. 经验分布函数：

$$F_n(x) = \{X_1, \dots, X_n \text{ 中不大于 } x \text{ 的个数}\} / n$$

3. 点估计分为：矩估计、极大似然估计、bayes 估计。
4. 点估计的定义：每当有了样本 X_1, \dots, X_n , 就代入函数 $\hat{\theta}_1(X_1, \dots, X_n)$ 中算出一个值，用来作为 θ_1 的估计值。为着这样的特定目的而构造的统计量 $\hat{\theta}_1$ 叫做 (θ_1) 的估计量。由于未知参数 θ_1 是数轴上的一个点，用 $\hat{\theta}_1$ 去估计 θ_1 , 等于用一个点去估计另一个点，所以这样的估计叫做点估计
5. 当 X_1, \dots, X_n 固定而把 L 看做 $\theta_1, \dots, \theta_k$ 的函数时，它称为“似然函数”。取定 X 之后，看作是各种 θ 的似然度，极大似然的思想就是找那个似然度最大的估计
6. 评价点估计的好坏：
 - (a) 估计量的某种指标，如无偏性
 - (b) 具体的数量指标，如均方误差
7. 无偏性表示没有系统误差，但是随机误差是一直都存在的；从大数定理角度来理解无偏性： $\sum_{i=1}^N \hat{g}(X_1^{(i)}, \dots, X_n^{(i)}) / N$, 依概率收敛到被估计的值 $g(\theta_1, \dots, \theta_k)$
8. 标准差不是无偏估计： $\sigma^2 = E(S^2) = \text{Var}(S) + (ES)^2$, 由于方差总非负: $\text{Var}(S) \geq 0$, 有 $\sigma \geq E(S)$. 因而 $E(S) \leq \sigma$
9. 信息量越大，下界越小，表示 $g(\theta)$ 的无偏估计更有可能达到较小的方差—即有可能被估计得更准确一些. $g(\theta)$ 是通过样本去估计的， $g(\theta)$ 能估得更准，表示样本所含的信息量愈大.
10. 指数分布的样本均值是 umvue
11. 相合估计：如果当样本大小无限增加时，估计量依概率收敛于被估计的值，则称该估计量是相合估计。
12. 大样本性质：渐进正态和相合性
13. 区间估计：用一个区间去估计参数
14. 区间估计：平均 100 次中有 95 次的确包含所要估计的值. 一旦算出具体区间，就不能再说它有 95% 的机会包含要估计的值了.
15. 枢轴量：
 - (a) σ^2 : $(n-1)S^2/\sigma^2$
 - (b) 指数分布中的 λ : 由于 $2n\lambda\bar{X} \sim \chi^2_{2n}$, $2n\lambda\bar{X}$ 可作为枢轴变量
16. 如果给定一个区间估计，如果想控制区间的长度，如果涉及到样本均值或者样本方差，可以使用两阶段方法：先抽出样本 X_1, \dots, X_n , 算出样本标准差 S . 根据 S 的大小决定追加抽样的数目， S 愈大，追加抽样次数愈多.
17. 大样本法：二项分布： $(Y_n - np) / \sqrt{np(1-p)} \sim N(0, 1)$
18. 大样本法：poisson 分布： $(Y_n - n\lambda) / \sqrt{n\lambda}$ 近似地有分布 $N(0, 1)$
19. 针对总体均值的估计，其中均值方差均未知。根据中心极限定理 $\sqrt{n}(\bar{X} - \theta) / \sigma \sim N(0, 1)$. 但此处 σ 未知，仍不能以 $\sqrt{n}(\bar{X} - \theta) / \sigma$ 作为枢轴变量. 因为 n 相当大，样本均方差 S 是 σ 的一个相合估计，故可近似地用 S 代 σ , 得

$$\sqrt{n}(\bar{X} - \theta) / S \sim N(0, 1)$$

Theorem 3.1.2 — 陈家鼎. 1. 数理统计学研究怎样有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断或预测，直至为采取一定的决策和行动提供依据和建议。

2. 样本方差 S^2 是 $\text{Var}_\theta(X)$ 的强相合估计.
3. 从因子分解定理可以知道: 如果极大似然估计存在, 则一定是充分统计量的函数
4. 对于指数分布组:

$$f(x, \theta) = S(\theta)h(x) \exp \left\{ \sum_{k=1}^m C_k(\theta) T_k(x) \right\}$$

其中 $\theta = (\theta_1, \dots, \theta_m) \in \Theta, S(\theta) > 0, h(x) \geq 0, \Theta$ 是 m 维欧氏空间中的开集。此时样本 (X_1, \dots, X_n) 的联合密度函数(或概率函数)为

$$\prod_1^n f(x_i, \theta) = [S(\theta)]^n \prod_{i=1}^n h(x_i) \exp \left\{ \sum_{k=1}^m C_k(\theta) \sum_{i=1}^n T_k(x_i) \right\}$$

可见, $\varphi(X_1, \dots, X_n) = (\sum_1^n T_1(X_i), \dots, \sum_1^n T_m(X_i))$ 是 θ 的充分统计量。

5. 完全性: 称统计量 $\varphi(X_1, \dots, X_n)$ 是完全的, 若对任何(Borel 可测) 函数 $u()$, 只要 $E_\theta u[\varphi(X_1, \dots, X_n)] = 0$ (对一切 θ) 就可推出 $P_\theta(u[\varphi(X_1, \dots, X_n)] = 0) = 1$ (对一切 θ)。这里 P_θ 是与参数 θ 相应的概率, E_θ 是与 P_θ 相应的数学期望。
6. 若参数 θ 的集合 Θ 有内点, 则指类型分布族中的充分统计量

$$\varphi = \left(\sum_1^n T_1(x_i), \dots, \sum_1^n T_m(x_i) \right)$$

是完全的

7. Blackwell - Lehmann - Sheffe 定理: 若 $\varphi(X_1, \dots, X_n)$ 是完全的充分统计量, $\psi[\varphi(X_1, \dots, X_n)]$ 是 $g(\theta)$ 的无偏估计, 则

$$\psi[\varphi(X_1, \dots, X_n)]$$

就是 $g(\theta)$ 的最小方差无偏估计.

8. C-R 下界推导: 利用无偏估计, 写出其无偏的定义式, 接着对其求导 + 求导和积分可以交换, 最后利用 Schwarz 不等式
9. 对于正态分布来说, 对于均值和方差如果一个知道, 另外一个的矩估计就是 UMVUE
10. 两点分布的样本均值是 p 的 umvue
11. 强相合估计: 称估计量 $\varphi_n(X_1, \dots, X_n)$ ($n \geq 1$) 是 $g(\theta)$ 的强相合估计, 若

$$P \left(\lim_n \varphi_n = g(\theta) \right) = 1$$

强相合估计一定是相合估计

12. 根据大数定律, 可知矩估计是强相合估计
13. MLE 在满足一定条件下也是强相合的: 设 X 有密度函数 $f(x, \theta)$, 其中 $\theta \in (a, b)$ ($-\infty \leq a < b \leq \infty$) 且满足下列条件:
 - (a) 对一切 $\theta_1 \neq \theta_2$

$$\mu \{x : f(x, \theta_1) > 0 \text{ 且 } f(x, \theta_1) \neq f(x, \theta_2)\} > 0$$

这里 μ 表示 Lebesgue 测度。

- (b) 似然函数 $L(x_1, \dots, x_n; \theta) = \prod_1^n f(x_i, \theta)$ 是 $\dot{\theta}$ 的单峰函数, 即存在 $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$ 使得 $L(x_1, \dots, x_n; \theta)$ 在 $(a, \hat{\theta}_n]$ 上严格增, 在 $[\hat{\theta}_n, b)$ 上严格减.

则 θ 的最大似然估计 $\hat{\theta}_n$ (基于样本 X_1, \dots, X_n) 存在且唯一, 而且 $\hat{\theta}_n$ 是 θ 的强相合估计.

14. 常用不等式: $\ln x < x - 1 (x > 0 \text{ 且 } x \neq 1)$

15. 点估计可能会有差距, 但是差距有多少不知道; 区间估计是保障点在这个区间里。

Definition 3.1.2 当总体分布为 F , 而 X_1, \dots, X_n 为独立同分布的样本时, 我们常称 X_1, \dots, X_n 是从总体 F 中抽出的**简单随机样本或独立同分布的样本** (independent identical distribution, 简记为 iid)

Definition 3.1.3 — 简单随机抽样. 如果试验是非破坏性的且总体是有限的, 抽取应该是有放回的; 如果试验是破坏性的总体应该是无限的或是很大的。上述思想的抽样方法称为**简单随机抽样**。

Definition 3.1.4 — 统计量. 设 $((\xi_1, \dots, \xi_n))$ 为总体 ξ 的样本, $T(x_1, \dots, x_n)$ 为样本空间 X 上的实值波雷尔可测函数. 如果 $T(\xi_1, \dots, \xi_n)$ 中不包含任何未知参数, 则称 $T(\xi_1, \dots, \xi_n)$ 为一个**统计量**.

3.2 样本数据的整理与显示

Definition 3.2.1 — 经验分布函数. 若将样本观察值 x_1, x_2, \dots, x_n 由小到大进行排列, 得有序样本 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, 用有序样本定义如下函数

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)} \\ k/n, & \text{当 } x_{(k)} \leq x < x_{(k+1)}, k = 1, 2, \dots, n-1 \\ 1, & \text{当 } x \geq x_{(n)} \end{cases}$$

则称 $F_n(x)$ 为该样本的经验分布函数, 满足分布函数的三条要求。

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mu(x - \xi_k), x \in R$$

$\sum_{k=1}^n \mu(x - \xi_k)$ 表示小于 x 的那些样品 ξ_k 的个数.

Proposition 3.2.1 1. $F_n(x) \xrightarrow{P} F(x), \forall x \in R$

2. $E[F_n(x) - F(x)]^2 \rightarrow 0, \forall x \in R$

3. $P\{\lim_n F_n(x) = F(x)\} = 1, \forall x \in R$

4. $\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} N(0, 1), \forall x \in R$

Theorem 3.2.2 — 格里文科定理 (Glivenko-Cantelli). 设 x_1, x_2, \dots, x_n 是取自总体分布函数为 $F(x)$ 的样本, $F_n(x)$ 是该样本的经验分布函数, 则当 $n \rightarrow +\infty$ 时, 有

$$P\left(\sup_{-\infty < x < +\infty} |F_n(x) - F(x)| \rightarrow 0\right) = 1$$

此定理表明: 当 n 相当大时, 经验分布函数 $F_n(x)$ 是总体分布函数 $F(x)$ 的一个良好的近似。

Theorem 3.2.3 — (Dvoretzky, Kiefer, and Wolfowitz). 对于任意给定的自然数 n , 设 X_1, \dots, X_n 为取自总体分布 $F(x)$ 的 iid 样本, $F_n(x)$ 为其经验分布函数, 记

$$D_n = \sup_x |F_n(x) - F(x)|$$

对于 iid 样本 X_1, \dots, X_n , 存在一个不依赖分布函数的正常数 C

$$P\{D_n > d\} \leq Ce^{-2nd^2}, \forall d > 0$$

上述不等式称为 D_n 的指类型不等式, 它是上述三位作者于 1956 年证明的 (Ann. Math. Statist. 27, 642-669). 另外, Kolmogorov 于 1933 年给出了 D_n 在一维时的极限分布如下:

Theorem 3.2.4 (Kolmogorov) 当 F 为一维且连续时, 我们有

$$\lim_{n \rightarrow \infty} P\left\{ n^{1/2} D_n \leq d \right\} = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 d^2}, \forall d > 0$$

关于 D_n 的应用, 请见 §6.4 的 Kolmogorov 检验.

Theorem 3.2.5 — 频数频率分布表. 由样本数据 x_1, x_2, \dots, x_n 制作频数频率分布表的操作步骤如下:

1. 确定组数 k
2. 确定每组组距, 通常取每组组距相等为 d : 组距 $d = (\text{样本最大观测值} - \text{样本最小观测值}) / \text{组数}$.
3. 确定每组组限: 各组区间端点为 $a_0, a_0 + d = a_1, a_0 + 2d = a_2, \dots, a_0 + kd = a_k$, 形成如下的分组区间

$$(a_0, a_1], (a_1, a_2], \dots, (a_{k-1}, a_k]$$

4. 统计样本数据落入每个区间的频数, 并计算频率。综合上述, 列入表中, 即得该样本的频数频率分布表。

该表依赖于分组, 不同的分组方式有不同的频数频率分布表。

Theorem 3.2.6 — 样本数据的图形表示. 1. 直方图

- (a) 利用频数频率分布表上的区间 (横坐标) 和频数 (纵坐标) 可作出频数直方图
- (b) 若把纵坐标改为频率就得频率直方图
- (c) 若把纵坐标改为频率 / 组距, 就得到单位频率直方图. 这时长条矩形的面积之和为 1

此三种直方图的差别仅在纵坐标的设置上, 直方图本身并无变化。

2. 茎叶图: 把样本中的每个数据分为茎与叶, 把茎放于一侧, 叶放于另一侧, 就得到一张该样本的茎叶图。比较两个样本时, 可画出背靠背的茎叶图。其叶图保留数据中全部信息。当样本量较大, 数据很分散, 横跨二, 三个数量级时, 茎叶图并不适用。

3.3 统计量及其分布

Definition 3.3.1 统计量: 不含未知参数的样本函数称为统计量. 统计量的分布称为抽样分布

Definition 3.3.2 由于统计量是作为随机变量的样本的函数, 故它也有概率分布, 于是, 我们称统计量的概率分布为该统计量的抽样分布 (Sampling Distribution)

Definition 3.3.3 — 样本均值. 样本 x_1, \dots, x_n 的算术平均值称为样本均值, 记为 \bar{x} 分组样本均值:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i f_i$$

其中 n 为样本量, k 为组数, x_i 与 f_i 为第 i 组的组中值与频数, 分组样本均值是完全样本均值的一种较好的近似. 样本均值是样本的位置特征, 样本中大多数值位于 \bar{x} 左右.

Proposition 3.3.1 样本均值的性质:

1. 样本数据 x_i 对样本均值 \bar{x} 的偏差之和为零

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

2. 样本数据 x_i 与样本均值 \bar{x} 的偏差平方和最小, 即对任意的实数 c 有

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - c)^2$$

3. 若总体分布为 $N(\mu, \sigma^2)$, 则 \bar{x} 的精确分布为 $N(\mu, \sigma^2/n)$
4. 若总体分布未知, 但其期望 μ 与方差 σ^2 存在, 则当 n 较大时, \bar{x} 的渐近分布为 $N(\mu, \sigma^2/n)$, 这里渐近分布是指 n 较大时的近似分布.

Proof. 3.

$$\sum_{i=1}^n x_i \sim N(n\mu, n\sigma^2)$$

4.

由中心极限定理, $\sqrt{n}(\bar{x} - \mu)/\sigma \xrightarrow{L} N(0, 1)$, 这表明 n 较大时 \bar{x} 的渐近分布为 $N(\mu, \sigma^2/n)$.

■

(R) 这里是 MLE 的渐进正态性吗

Definition 3.3.4 — 样本方差与样本标准差. 样本方差有两个, 样本方差 s_*^2 . 与样本无偏方差 s^2

$$s_*^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

实际中常用的是无偏样本方差 s , 这是因为: 当 σ^2 为总体方差时, 总有

$$E(s_*^2) = \frac{n-1}{n} \sigma^2, \quad E(s^2) = \sigma^2$$

这表明: s_*^2 有系统偏小的误差, 而 s^2 无此系统偏差. 今后称 s^2 为样本方差. $s = \sqrt{s^2}$ 为样本标准差.

样本方差是样本的散布特征, s^2 愈大样本愈分散, s^2 愈小分布愈集中, 样本标准差 s 与样本均值 \bar{x} 有相同单位, 使用更频繁, 但 s 的计算必须通过 s^2 才能获得。 s^2 的计算有如下三个公式可供选用:

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum x_i^2 \frac{(\sum x_i)^2}{n} \right] = \frac{1}{n-1} (\sum x_i^2 - n\bar{x}^2)$$

在分组样本场合, 样本方差的近似计算公式为

$$s^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^k f_i x_i^2 - n\bar{x}^2 \right)$$

其中 k 为组数, x_i, f_i 分别为第 i 个区间的组中值与频数, \bar{x} 为分组样本的均值.

Theorem 3.3.2 样本方差的自由度: 如果把 \bar{X} 代入 $\sum_{i=1}^n (X_i - \bar{X})^2$, 则可知它是一个如下形式的二次型: $\sum_{i,j=1}^n a_{ij} X_i X_j$ ($a_{ij} = a_{ji}$), 且不难验证矩阵 $A = (a_{ij})$ 的秩为 $n-1$

Definition 3.3.5 设 X_1, \dots, X_n 为样本, 则称 S_n/\bar{X} 为样本变异系数 (Coefficient of Variability).

Definition 3.3.6 — 样本矩及其区数. 1. 样本的 k 阶原点矩

$$a_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

样本均值 \bar{x} 为样本的一阶原点矩

2. 样本的 k 阶中心矩

$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

样本方差 s_*^2 和 s^2 都为样本的二阶中心矩

3. 样本变异系数

$$C_r = s/\bar{x}$$

4. 样本的偏度

$$\hat{\beta}_s = b_3/b_2^{3/2}$$

样本偏度 $\hat{\beta}_s$ 反映了总体分布密度曲线与对称性的偏离方向和程度。如果数据完全对称, 则不难看出 $b_3 = 0$. 对不对称的数据则 $b_3 \neq 0$. 这里用 b_3 除以 $b_2^{3/2}$ 是为了消除量纲的影响。如果 $\hat{\beta}_s$ 明显小于 0 表示分布的左尾长, 即样本中有几个特小的数。

5. 样本的峰度

$$\hat{\beta}_k = \frac{b_4}{b_2^2} - 3$$

样本峰度 $\hat{\beta}_k$ 是反映总体分布密度曲线在其峰值附近的陡峭程度和尾部粗细的统计量。当 $\hat{\beta}_k$ 明显大于 0 时, 分布密度曲线在其峰值附近比正态分布来得陡, 尾部更细, 称为尖顶型; 当 $\hat{\beta}_k$ 明显小于 0 时, 分布密度曲线在其峰值附近比正态分布来得平坦, 尾部更粗, 称为平顶型

Definition 3.3.7 — 次序统计量及其分布. 设 x_1, \dots, x_n 是取自某总体的一个样本, $x_{(i)}$ 称为该样本的第 i 个次序统计量, 假如 $x_{(i)}$ 的每次取值是将每次所得的样本观测值由小到大排序后得到的第 i 个观测值。

1. $x_{(1)} = \min\{x_1, \dots, x_n\}$ 称为该样本的最小次序统计量
2. $x_{(n)} = \max\{x_1, \dots, x_n\}$ 称为该样本的最大次序统计量
3. $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ 称为该样本的次序统计量.
4. $R = x_{(n)} - x_{(1)}$ 称为样本极差.

Theorem 3.3.3 设总体 X 的密度函数为 $p(x)$, 分布函数为 $F(x)$, x_1, \dots, x_n 为样本, 则有

1. 样本第 k 个次序统计量 $x_{(k)}$ 的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!} (F(x))^{k-1} (1-F(x))^{n-k} p(x)$$

2. 样本第 i 个与第 j 个次序统计量的联合密度函数为

$$p_{ij}(y, z) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(y)]^{i-1} [F(z) - F(y)]^{j-i-1} \\ \cdot [1 - F(z)]^{n-j} p(y) p(z), \quad y \leq z, 1 \leq i < j \leq n$$

Proof. 对任意的实数 x , 考虑次序统计量 $x_{(k)}$ 取值落在小区间 $(x, x + \Delta x]$ 内这一事件, 它等价于“样本容量为 n 的样本中有 1 个观测值落在 $(x, x + \Delta x]$ 之间而有 $k-1$ 个观测值小于等于 x , 有 $n-k$ 个观测值大于 $x + \Delta x$ 。”

样本的每一个分量小于等于 x 的概率为 $F(x)$, 落入区间 $(x, x + \Delta x]$ 的概率为 $F(x + \Delta x) - F(x)$, 大于 $x + \Delta x$ 的概率为 $1 - F(x + \Delta x)$, 而将 n 个分量分成这样的三组, 总的分法有 $\frac{n!}{(k-1)!!(n-k)!}$ 种。于是, 若以 $F_k(x)$ 记 $x_{(k)}$ 的分布函数, 则由多项分布可得

$$F_k(x + \Delta x) - F_k(x) \approx \frac{n!}{(k-1)!(n-k)!} (F(x))^{k-1} (F(x + \Delta x) - F(x)) (1 - F(x + \Delta x))^{n-k}$$

两边除以 Δx , 并令 $\Delta x \rightarrow 0$, 即有

$$p_k(x) = \lim_{\Delta x \rightarrow 0} \frac{F_k(x + \Delta x) - F_k(x)}{\Delta x} \\ = \frac{n!}{(k-1)!(n-k)!} (F(x))^{k-1} p(x) (1 - F(x))^{n-k}$$

其中 $p_k(x)$ 的非零区间与总体的非零区间相同。特别, 令 $k=1$ 和 $k=n$ 即得到最小次序统计量 $x_{(1)}$ 和最大次序统计量 $x_{(n)}$ 的密度函数分别为

$$p_1(x) = n \cdot (1 - F(x))^{n-1} p(x) \quad p_n(x) = n \cdot (F(x))^{n-1} p(x)$$

Proof. 对增量 $\Delta y, \Delta z$ 以及 $y < z$, 事件

$$\{x_{(i)} \in (y, y + \Delta y], x_{(j)} \in (z, z + \Delta z]\}$$

可以表述为“容量为 n 的样本 x_1, \dots, x_n 中有 $i-1$ 个观测值小于等于 y , 一个落入区间 $(y, y + \Delta y]$, $j-i-1$ 个落入区间 $(y + \Delta y, z]$, 一个落入区间 $(z, z + \Delta z]$, 而余下 $n-j$ 个大于 $z + \Delta z$ ”于是由多项分布可得

$$\begin{aligned} & P(x_{(i)} \in (y, y + \Delta y), x_{(j)} \in (z, z + \Delta z)) \\ & \approx \frac{n!}{(i-1)!1!(j-i-1)!1!(n-j)!} [F(y)]^{i-1} p(y) \Delta y [F(z) - F(y + \Delta y)]^{j-i-1} p(z) \Delta z [1 - F(z + \Delta z)]^{n-j} \end{aligned}$$

考虑到 $F(x)$ 的连续性, , 当 $\Delta y \rightarrow 0, \Delta z \rightarrow 0$ 时有 $F(y + \Delta y) \rightarrow F(y), F(z + \Delta z) \rightarrow F(z)$ 于是

$$\begin{aligned} p_{ij}(y, z) &= \lim_{\Delta y \rightarrow 0, \Delta z \rightarrow 0} \frac{P(x_{(i)} \in (y, y + \Delta y), x_{(j)} \in (z, z + \Delta z))}{\Delta y \cdot \Delta z} \\ &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(y)]^{i-1} [F(z) - F(y)]^{j-i-1} [1 - F(z)]^{n-j} p(y) p(z) \end{aligned}$$

定理得证。 ■

■ **Example 3.1** 设总体分布为 $U(0, 1)$, x_1, x_2, \dots, x_n 为样本, 则其第 k 个次序统计量的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k}, \quad 0 < x < 1$$

这是贝塔分布 $Be(k, n-k+1)$, 从而有 $E(x_{(k)}) = \frac{k}{n+1}$

Theorem 3.3.4 — (次序统计量). 设 $X_1, \dots, X_n \sim \text{iid } F(x)$, 求次序统计量 $X_{(k)}$ 的抽样分布解
实际上, 我们有

$$\begin{aligned} F_k(x) &= P\{X_{(k)} < x\} = P\{\text{在 } X_1, \dots, X_n \text{ 中至少有 } k \text{ 个小于 } x\} \\ &= \sum_{m=k}^n P\{\text{在 } X_1, \dots, X_n \text{ 中恰有 } m \text{ 个小于 } x\} \\ &= \sum_{m=k}^n \binom{n}{m} F^m(x) (1-F(x))^{n-m} \end{aligned}$$

即

$$F_k(x) = \sum_{m=k}^n \binom{n}{m} F^m(x) (1-F(x))^{n-m}$$

如总体分布 $F(x)$ 有密度函数 $f(x)$, 则 $X_{(k)}$ 地有密度, 且少

$$f_k(x) = k \binom{n}{k} F^{k-1}(x) (1-F(x))^{n-k} f(x)$$

Definition 3.3.8 — 样本中位数与样本分位数. x_1, \dots, x_n 是取自某总体的样本, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 为该样本的次序统计量, 则样本中位数 $m_{0.5}$ 定义为

$$m_{0.5} = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ 为奇数,} \\ \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \text{ 为偶数.} \end{cases}$$

而样本的 p 分位数 m_p 定义为

$$m_p = \begin{cases} x_{([np+1])}, & np \text{ 不是整数} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}), & np \text{ 是整数} \end{cases}$$

其中 $[x]$ 表示小于或等于 x 的最大整数. 中位数对样本的极端值有抗干扰性, 或称有稳健性.

Definition 3.3.9 — 样本分位数的渐近分布. 设总体的密度函数为 $p(x)$, x_p 为总体的 p 分位数. 若 $p(x)$ 在 x_p 处连续且 $p(x_p) > 0$, 则当 n 充分大时, 有

$$m_p \doteq N\left(x_p, \frac{p(1-p)}{n \cdot p^2(x_p)}\right) \quad m_{0.5} \doteq N\left(x_{0.5}, \frac{1}{4n \cdot p^2(x_{0.5})}\right)$$

Definition 3.3.10 — 五数概括与箱线图. 五数概括是指用样本的五个次序统计量

$$x_{\min} = x_{(1)}, Q_1 = m_{0.25}, Q_2 = m_{0.5}, Q_3 = m_{0.75}, x_{\max} = x_{(n)}$$

大致描述一个样本的轮廓, 其图形表示称为箱线图. 当样本量较大时, 箱线图可用来对总体分布形状进行大致的判断

Definition 3.3.11 — (样本偏度 (Skewness) 与峰度 (Kurtosis)). 称

$$\sqrt{n} \sum_{i=1}^n (X_i - \bar{X})^3 / \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]^{3/2}$$

与

$$n \sum_{i=1}^n (X_i - \bar{X})^4 / \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]^2 - 3$$

为样本偏度与峰度.

3.4 抽样基本定理

统计量的构造	抽样分布密度函数	期望	方差
$\chi^2 = x_1^2 + x_2^2 + \cdots + x_n^2$	$P(y) = \frac{1}{\Gamma(\frac{n}{2}) 2^{n/2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} (y > 0)$	n	$2n$
$F = \frac{(y_1^2 + \cdots + y_m^2)/m}{(x_1^2 + \cdots + x_n^2)/n}$	$p(y) = \frac{\Gamma(\frac{m+n}{2})(\frac{m}{n})^{m/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} y^{\frac{m-1}{2}} \cdot (1 + \frac{m}{n}y)^{-\frac{m+n}{2}}$	$\frac{n}{n-2} (n > 2)$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} (n > 4)$
$t = \frac{y}{\sqrt{(x_1^2 + \cdots + x_n^2)/n}}$	$p(y) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}}$	$0 (n > 1)$	$\frac{n}{n-2} (n > 2)$

Definition 3.4.1 — χ^2 分布. 设 X_1, X_2, \dots, X_n 独立同分布于标准正态分布 $N(0, 1)$, 则

$$\chi^2 = X_1^2 + \cdots + X_n^2$$

的分布称为自由度为 n 的 χ^2 分布, 记为 $\chi^2 \sim \chi^2(n)$

若随机变量 $X \sim N(0, 1)$, 则 $X^2 \sim Ga(1/2, 1/2)$, 根据伽玛分布的可加性立有

$$\chi^2 \sim Ga(n/2, 1/2) = \chi^2(n)$$

由此可见, $\chi^2(n)$ 分布是伽玛分布的特例, 故 $\chi^2(n)$ 分布的密度函数为

$$p(y) = \frac{(1/2)^{\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad y > 0$$

该密度函数的图像是一个只取非负值的偏态分布, $E(\chi^2) = n$, $\text{Var}(\chi^2) = 2n$

Proposition 3.4.1 1. 随着 n 的增大, 它的对称性越来越好, 峰度越来越小。

2. 随着 n 的增大, 其图形越来越像正态分布的概率密度函数。

3. 随着 n 的增大, 它的图形越来越向右移动, 且尾部越来越大。

4. χ^2 的特征函数为 $\psi(t) = (1 - 2it)^{-\frac{n}{2}}$

■ **Example 3.2** 设 x_1, x_2, \dots, x_n 是来自正态分布 $N(\mu, \sigma^2)$ 的一个样本, 其中 μ 是已知常数, 求统计量

$$T = \sum_{i=1}^n (x_i - \mu)^2$$

的分布。

Proof. 令 $y_i = (x_i - \mu)/\sigma, i = 1, \dots, n$, 则 y_1, \dots, y_n 是独立同分布随机变量, 其共同分布为 $N(0, 1)$

$$\frac{T}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n y_i^2 \sim \chi^2(n)$$

而 T 的密度函数为

$$p(t) = \frac{1}{(2\sigma^2)^{n/2} \Gamma(n/2)} e^{-\frac{1}{2\sigma^2} t^{\frac{n}{2}-1}}, \quad t > 0$$

这就是伽玛分布 $Ga\left(\frac{n}{2}, \frac{1}{2\sigma^2}\right)$

■

Theorem 3.4.2 — 抽样基本定理. 设 x_1, \dots, x_n 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 其样本均值和样本方差分别为

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{和} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

则有

1. \bar{x} 与 s^2 相互独立
2. $\bar{x} \sim N(\mu, \sigma^2/n)$
3. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

Proof. x_1, \dots, x_n 的联合密度函数为

$$p(x_1, \dots, x_n) = (2\pi\sigma^2)^{-n/2} e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n x_i^2 - 2n\bar{x}\mu + n\mu^2}{2\sigma^2} \right\}$$

记 $X = (x_1, \dots, x_n)'$, 取一个 n 维正交矩阵 A , 其第一行的每一个元素均为 $1/\sqrt{n}$, 如

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2 \cdot 1}} & -\frac{1}{\sqrt{2 \cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3 \cdot 2}} & \frac{1}{\sqrt{3 \cdot 2}} & -\frac{2}{\sqrt{3 \cdot 2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}$$

令 $Y = (y_1, \dots, y_n)' = AX$, 则该变换的雅可比 (Jaccobi) 行列式为 1, 且注意到

$$\bar{x} = \frac{1}{\sqrt{n}} y_1$$

$$\sum_{i=1}^n y_i^2 = Y'Y = X'A'AX = \sum_{i=1}^n x_i^2$$

于是 y_1, \dots, y_n 的联合密度函数为

$$p(y_1, \dots, y_n) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n y_i^2 - 2\sqrt{n}y_1\mu + n\mu^2}{2\sigma^2} \right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=2}^n y_i^2 + (y_1 - \sqrt{n}\mu)^2}{2\sigma^2} \right\}$$

由此, $Y = (y_1, \dots, y_n)'$ 的各个分量相互独立, 且都服从正态分布, 其方差均为 σ^2 而均值并不完全相同, y_2, \dots, y_n 的均值为 0, y_1 的均值为 $\sqrt{n}\mu$, 这就证明了结论 (2). 由于

$$(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - (\sqrt{n}\bar{x})^2$$

$$= \sum_{i=1}^n y_i^2 - y_1^2 = \sum_{i=2}^n y_i^2$$

这证明了结论 (1), 由于 y_2, \dots, y_n 独立同分布于 $N(0, \sigma^2)$, 于是

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=2}^n \left(\frac{y_i}{\sigma} \right)^2 \sim \chi^2(n-1)$$

■

Theorem 3.4.3 如果一组随机样本的均值与样本方差独立, 则总体分布必为正态 (见 Stuart and Ord (1987))。事实上, 一般情况下样本均值与方差的协方差为 $\text{Cov}(\bar{X}, S_n^2) = \frac{v_3}{n}$, 其中

$$v_k = E(X - E(X))^k$$

Theorem 3.4.4 — (Cochran). 设 X_1, \dots, X_n 相互独立, 且 $X_i \sim N(\mu_i, \sigma^2), i = 1, \dots, n$, 令 $\mathbf{X} = (X_1, \dots, X_n)'$. 又设 A_1, \dots, A_m 是 m 个 n 阶非负定阵, 且 $A_1 + \dots + A_m = I_n$ (n 阶单位阵), $\sum_{i=1}^m rk(A_i) = n$. 记

$$\xi_i = \mathbf{X}' A_i \mathbf{X}, i = 1, \dots, m, \mu = (\mu_1, \dots, \mu_n)'$$

则

1. ξ_1, \dots, ξ_m 相互独立

2. 如果 $\mu' A_1 \mu = 0$, 则 $\xi_1 / \sigma^2 \sim \chi^2(rk(A_1))$

Proof. 记 $n_i = rk(A_i), i = 1, \dots, m$. 因为 A_i 是 n 阶非负定阵, 则存在一个 $n \times n_i$ 阵 B_i , 使得 $A_i = B_i B_i'$. 记 $B = \begin{pmatrix} B_1 : B_2 : \dots : B_m \end{pmatrix}$, 则由已知条件可知 B 是一个 n 阶正交阵. 作如下正交变换:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = B' \mathbf{X} = \begin{pmatrix} B'_1 \mathbf{X} \\ \vdots \\ B'_m \mathbf{X} \end{pmatrix}$$

即

$$B'_i \mathbf{X} = \begin{pmatrix} Y_{n_1+\dots+n_{i-1}+1} \\ Y_{n_1+\dots+n_{i-1}+2} \\ \vdots \\ Y_{n_1+\dots+n_{i-1}+n_i} \end{pmatrix}, i = 1, \dots, m$$

其中 $n_0 = 0$ 由于 B 是一正交阵, 故由习题 1.4 知道: Y_1, \dots, Y_n 独立, 且 $Y_i \sim N(\beta_i, \sigma^2), i = 1, \dots, n$, 其

$$\text{中 } \boldsymbol{\beta} = (\beta_1, \dots, \beta_n)' = B' \boldsymbol{\mu}$$

又因为

$$\xi_i = \mathbf{X}' A_i \mathbf{X} = \mathbf{X}' B_i B_i' \mathbf{X} = (B'_i \mathbf{X})' B'_i \mathbf{X} = \sum_{j=n_1+\dots+n_{i-1}+1}^{n_1+\dots+n_{i-1}+n_i} Y_j^2, i = 1, \dots, m$$

故 ξ_1, \dots, ξ_m 相互独立, 第一点得证. 下证第二点. 由已知条件知, $\mu' A_1 \mu = 0$, 即 $\mu' B_1 B_1' \mu = 0$, 于是, $B'_1 \boldsymbol{\mu} = \mathbf{0}$, 即 $\beta_i = 0, i = 1, \dots, n_1$

又由于 $\xi_1 / \sigma^2 = \sum_{i=1}^{n_1} Y_i^2 / \sigma^2$, 而 $Y_1, \dots, Y_{n_1} \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, 故 $\xi_1 / \sigma^2 \sim \chi^2(n_1)$

■

Definition 3.4.2 — F 分布. 设随机变量 $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n), X_1$ 与 X_2 独立, 则称

$$F = \frac{X_1/m}{X_2/n}$$

的分布是自由度为 m 与 n 的 F 分布, 记为 $F \sim F(m, n)$, 其中 m 称为分子自由度, n 称为分母自由度. 下面分两步来导出 F 分布的密度函数。

1. 我们导出 $Z = \frac{X_1}{X_2}$ 的密度函数, 若记 $p_1(x)$ 和 $p_2(x)$ 分别为 $\chi^2(m)$ 和 $\chi^2(n)$ 的密度函数, 根据独立随机变量商的分布的密度函数的公式的密度函数为

$$\begin{aligned} p_z(z) &= \int_0^\infty x_2 p_1(zx_2) p_2(x_2) dx_2 \\ &= \frac{z^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2}) 2^{\frac{m+n}{2}}} \int_0^\infty x_2^{\frac{m+n}{2}-1} e^{-\frac{x_2}{2}(1+z)} dx_2 \end{aligned}$$

运用变换 $u = \frac{x_2}{2}(1+z)$, 可得

$$p_z(z) = \frac{z^{\frac{m}{2}-1} (1+z)^{-\frac{m+n}{2}}}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} \int_0^\infty u^{\frac{m+n}{2}-1} e^{-u} du$$

最后的定积分为伽玛函数 $\Gamma(\frac{m+n}{2})$, 从而

$$p_z(z) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} z^{\frac{m}{2}-1} (1+z)^{-\frac{m+n}{2}}, \quad z > 0$$

2. 我们导出 $F = \frac{n}{m}Z$ 的密度函数, 对 $y > 0$, 有

$$\begin{aligned} p_F(y) &= p_Z\left(\frac{m}{n}y\right) \cdot \frac{m}{n} \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} \left(\frac{m}{n}y\right)^{\frac{m}{2}-1} \left(1 + \frac{m}{n}y\right)^{-\frac{m+n}{2}} \cdot \frac{m}{n} \\ &= \frac{\Gamma(\frac{m+n}{2}) (\frac{m}{n})^{\frac{m}{2}}}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} y^{\frac{m}{2}-1} \left(1 + \frac{m}{n}y\right)^{-\frac{m+n}{2}} \end{aligned}$$

这就是自由度为 m 与 n 的 F 分布的密度函数. 该密度函数的图像是一个只取非负值的偏态分布.

由 F 分布的构造知, 若 $F \sim F(m, n)$, 则有 $1/F \sim F(n, m)$, 故对给定 $\alpha (0 < \alpha < 1)$

$$\alpha = P\left(\frac{1}{F} \leqslant F_\alpha(n, m)\right) = P\left(F \geqslant \frac{1}{F_\alpha(n, m)}\right)$$

从而

$$P\left(F \leqslant \frac{1}{F_\alpha(n, m)}\right) = 1 - \alpha$$

这说明

$$F_\alpha(n, m) = \frac{1}{F_{1-\alpha}(m, n)}$$

Theorem 3.4.5 从上述定义可以看出, 当 $n \rightarrow \infty$ 时, 统计量 mF 依概率收敛于一个自由度为 m 的 χ^2 随机变量. 同理, 当 $m \rightarrow \infty$ 时, 依概率收敛于自由度为 n 的 χ^2 随机变量.

Theorem 3.4.6 — F 分布的另外一种推导. 定理 1.3.8 由上面定义的 F 分布的 PDF 为

$$f(x; m, n) = \begin{cases} 0, & x < 0 \\ \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right) \left(\frac{mx}{n}\right)^{m/2-1} \left(1 + \frac{mx}{n}\right)^{-(m+n)/2}, & x > 0 \end{cases}$$

证明: 由于 $\xi \sim \chi^2(m), \eta \sim \chi^2(n)$, 且二者独立, 故它们的联合 PDF 为

$$f(x, y) = \frac{1}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} e^{-(x+y)/2} x^{m/2-1} y^{n/2-1}, x > 0, y > 0$$

作如下变换:

$$\begin{cases} u = x + y \\ v = \frac{x}{y} \frac{n}{m} \end{cases}$$

则 (U, V) 的联合 PDF 为

$$\begin{aligned} g(u, v) &= \frac{1}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} e^{-u/2} u^{\frac{m+n}{2}-2} \\ &\quad \left(\frac{m}{n}\right)^{m/2-1} \frac{v^{\frac{m}{2}-1}}{(1+mv/n)^{\frac{m+n}{2}-2}} \frac{m}{n} \frac{u}{(1+mv/n)^2} \\ &= \frac{1}{2^{\frac{m+n}{2}} \Gamma((m+n)/2)} e^{-u/2} u^{\frac{m+n}{2}-2} \\ &\quad \frac{\Gamma((m+n)/2)(m/n)^{m/2}}{\Gamma(m/2)\Gamma(n/2)} \frac{v^{m/2-1}}{(1+mv/n)^{\frac{m+n}{2}}} \end{aligned}$$

则知 U, V 相互独立, 且 $U \sim \chi^2(m+n)$, 而第二部分即为 V 的 PDF, 即为所求.

Corollary 3.4.7 设 $\xi \sim \chi^2(m), \eta \sim \chi^2(n)$, 且 ξ 与 η 相互独立, 则 $Y = \xi + \eta$ 与 $Z = \xi/\eta$ 相互独立。

Theorem 3.4.8 设 X_1, \dots, X_n iid $\sim N(0, \sigma^2)$, $Q_i, i = 1, \dots, k$ 是秩为 n_i 的关于 X_1, \dots, X_n 的非负定二次型, 且 $\sum_{i=1}^k Q_i = \sum_{i=1}^n X_i^2, n_1 + \dots + n_k = n$

$$F_{ij} = \frac{Q_i/n_i}{Q_j/n_j} \sim F(n_i, n_j)$$

Definition 3.4.3 — t 分布. 设随机变量 X_1 与 X_2 独立且 $X_1 \sim N(0, 1), X_2 \sim \chi^2(n)$, 则称 $t = \frac{X_1}{\sqrt{X_2/n}}$ 的分布为自由度为 n 的 t 分布, 记为 $t \sim t(n)$.

下面导出 t 分布的密度函数. 由标准正态密度函数的对称性知, X_1 与 $-X_1$ 有相同分布, 从而 t 与 $-t$ 有相同分布. 这意味着: 对任意实数 y 有

$$P(0 < t < y) = P(0 < -t < y) = P(-y < t < 0)$$

于是

$$P(0 < t < y) = \frac{1}{2} P(t^2 < y^2)$$

由 F 变量构造可知, $t^2 = \frac{X_1^2}{X_2/n} \sim F(1, n)$, 将上式两边关于 y 求导可得 t 分布的密度函数为

$$\begin{aligned} p_t(y) &= y p_F(y^2) = \frac{\Gamma(\frac{1+n}{2}) (\frac{1}{n})^{\frac{1}{2}}}{\Gamma(\frac{1}{2}) \Gamma(\frac{n}{2})} (y^2)^{\frac{1}{2}-1} \left(1 + \frac{1}{n} y^2\right)^{-\frac{1+n}{2}} y \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < y < \infty \end{aligned}$$

这就是自由度为 n 的 t 分布的密度函数 t 分布的密度函数的图像是一个关于纵轴对称的分布, 与标准正态分布的密度函数形状类似, 只是峰比标准正态分布低一些, 尾部的概率比标准正态分布的大一些

Theorem 3.4.9 由于对于固定的 x , 我们有

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} = e^{-x^2/2}$$

故当 n 很大时, t 分布的 PDF 接近于标准正态分布的 PDF. (其常数可用 Stirling 公式:

$$n! = \sqrt{2\pi n} n^n e^{-n}$$
 求得.

Theorem 3.4.10 设 $\xi \sim N(\mu, \sigma^2)$, $\eta/\sigma^2 \sim \chi^2(n)$, 且 ξ, η 相互独立, 则易知 $T = \frac{\xi - \mu}{\sqrt{\eta/n}} \sim t(n)$

Proposition 3.4.11 — t 分布的相关性质. 1. 自由度为 1 的 t 分布就是标准柯西分布, 它的均值不存在.
 2. $n > 1$ 时, t 分布的数学期望存在且为 0
 3. $n > 2$ 时, t 分布的方差存在, 且为 $n/(n-2)$
 4. 当自由度较大 (如 $n \geq 30$) 时, t 分布可以用 $N(0, 1)$ 分布近似.

Theorem 3.4.12 — 重要结论 1. 设 x_1, \dots, x_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, 则有

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \sim t(n-1)$$

其中 \bar{x} 为样本均值, s 为样本标准差.

Proof. 由于

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

原式等于

$$\frac{\sqrt{n}(\bar{x} - \mu)}{s} = \frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}}$$

由于分子是标准正态变量, 分母的根号里是自由度为 $n-1$ 的 χ^2 变量除以它的自由度, 且分子与分母相互独立, 由 t 分布定义可知 $t \sim t(n-1)$, 推论证完. ■

Theorem 3.4.13 — 重要结论 2. 设 x_1, \dots, x_m 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, \dots, y_n 是来自 $N(\mu_2, \sigma_2^2)$ 的样本, 且此两样本相互独立, 则有

$$F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1)$$

其中 s_x^2, s_y^2 分别是两个样本方差. 若 $\sigma_1^2 = \sigma_2^2$, 则

$$F = s_x^2 / s_y^2 \sim F(m-1, n-1)$$

Proof. 由两样本独立可知, s_x^2 与 s_y^2 相互独立, 由定理 3.4.2 知

$$\frac{(m-1)s_x^2}{\sigma_1^2} \sim \chi^2(m-1), \quad \frac{(n-1)s_y^2}{\sigma_2^2} \sim \chi^2(n-1)$$

由 F 分布定义可知 $F \sim F(m-1, n-1)$ ■

Proposition 3.4.14 1. $t(n)$ 分布的密度函数呈“中间高, 两边低, 左右对称”, 与标准正态曲线类似, 但峰比 $N(0, 1)$ 低, 两侧尾部概率比 $N(0, 1)$ 大. 当自由度 $n \rightarrow \infty$ 时, $t(n)$ 分布趋向 $N(0, 1)$ 分布. 当 $n > 30$ 时, 两者相差已不大, 可用 $N(0, 1)$ 分位数代替 $t(n)$ 分位数.

2. 关于 $t(n)$ 分布分位数有 $t_\alpha(n) + t_{1-\alpha}(n) = 0$ (互为相反数) 关于 $F(m, n)$ 分布分位数有 $F_\alpha(m, n) \cdot F_{1-\alpha}(n, m) = 1$ (互为倒数)
3. $t_{(n)}^2 = F(1, n)$
4. 由于 t 分布的密度函数关于 0 对称, 故其分位数间有如下关系

$$t_\alpha(n) = -t_{1-\alpha}(n)$$

Theorem 3.4.15 设 x_1, \dots, x_m 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, \dots, y_n 是来自 $N(\mu_2, \sigma_2^2)$ 的样本, 且此两样本相互独立, 记

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

其中

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

设 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 并记

$$s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = \frac{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}{m+n-2}$$

则

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

Proof. 由 $\bar{x} \sim N(\mu_1, \sigma^2/m)$, $\bar{y} \sim N(\mu_2, \sigma^2/n)$, \bar{x} 与 \bar{y} 独立, 故有

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2\right)$$

所以

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0, 1)$$

由定理3.4.2知, $\frac{(m-1)s_x^2}{\sigma^2} \sim \chi^2(m-1)$, $\frac{(n-1)s_y^2}{\sigma^2} \sim \chi^2(n-1)$, 且它们相互独立, 则由可加性知

$$\frac{(m+n-2)s_w^2}{\sigma^2} = \frac{(m-1)s_x^2 + (n-1)s_y^2}{\sigma^2} \sim \chi^2(m+n-2)$$

由于 $\bar{x} - \bar{y}$ 与 s_w^2 相互独立, 根据 t 分布的定义即可得到. ■

3.5 充分统计量

Definition 3.5.1 — 充分统计量. 设 x_1, \dots, x_n 是来自总体分布函数为 $F(x; \theta)$ 的一个样本, 统计量

$$T = T(x_1, \dots, x_n)$$

称为 θ 的充分统计量 (也称为该分布的充分统计量), 如果在给定 T 的取值后, x_1, \dots, x_n 的条件分布与 θ 无关. 其中条件分布可以是条件分布列 (离散场合) 或条件密度函数 (连续场合)。

Definition 3.5.2 (充分统计量 (Sufficient Statistics)) 对于某分布族 $\mathcal{F} = \{F_\theta(x) : \theta \in \Theta\}$ $\forall F \in \mathcal{F}$, 设 X_1, \dots, X_n 是来自 F 的样本, $T = T(X_1, \dots, X_n)$ 是一统计量. 如在给定 $T = t$ 下样本 (X_1, \dots, X_n) 的条件概率分布与总体分布 F 或参数 θ 无关, 则称统计量 T 是此分布族 \mathcal{F} 的充分统计量, 也称统计量 T 是参数 θ 的充分统计量。

■ **Example 3.3** 设总体为二点分布 $b(1, \theta)$, X_1, \dots, X_n 为样本, 令 $T = X_1 + \dots + X_n$, 则在给定 T 的取值后, 对任意的一组 x_1, \dots, x_n ($\sum_{i=1}^n x_i = t$), 有

$$\begin{aligned} & P(X_1 = x_1, \dots, X_n = x_n | T = t) \\ &= \frac{P(X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i)}{P(\sum_{i=1}^n X_i = t)} \\ &= \frac{\prod_{i=1}^{n-1} P(X_i = x_i) \cdot P(X_n = t - \sum_{i=1}^{n-1} x_i)}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \\ &= \frac{\prod_{i=1}^{n-1} \theta^{x_i} (1-\theta)^{1-x_i} \cdot \theta^{t-\sum_{i=1}^{n-1} x_i} (1-\theta)^{1-t+\sum_{i=1}^{n-1} x_i}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} \end{aligned}$$

■ **Example 3.4** 设 x_1, \dots, x_n 是来自 $N(\mu, 1)$ 的样本, $T = \bar{x}$, 则 $T \sim N(\mu, 1/n)$, 作变换

$$x_1 = x_1, \quad \dots, \quad x_{n-1} = x_{n-1}, \quad t = \bar{x}$$

其雅可比行列式为 n , 故 x_1, \dots, x_{n-1}, t 的联合密度函数为

$$\begin{aligned} p(x_1, \dots, x_{n-1}, t; \mu) &= n(2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \left[\sum_{i=1}^{n-1} (x_i - \mu)^2 + \left(nt - \sum_{i=1}^{n-1} x_i - \mu \right)^2 \right] \right\} \\ &= n(2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \left[\sum_{i=1}^{n-1} x_i^2 + n\mu^2 + (nt)^2 + \left(\sum_{i=1}^{n-1} x_i \right)^2 - 2nt\mu - 2nt \sum_{i=1}^{n-1} x_i \right] \right\} \\ &= n(2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \left[n(t - \mu)^2 + \sum_{i=1}^{n-1} x_i^2 + \left(\sum_{i=1}^{n-1} x_i - nt \right)^2 - nt^2 \right] \right\} \end{aligned}$$

从而条件密度函数 $p_\mu(x_1, \dots, x_{n-1} | T = t)$ 为

$$\begin{aligned} p_\mu(x_1, \dots, x_{n-1} | T = t) &= \frac{p_\mu(x_1, \dots, x_{n-1}, t)}{p_\mu(t)} \\ &= \frac{n(2\pi)^{-1/2} \exp\left\{-\frac{1}{2}\left[n(t-\mu)^2 + \sum_{i=1}^{n-1} x_i^2 + (\sum_{i=1}^{n-1} x_i - nt)^2 - n^2\right]\right\}}{(2\pi/n)^{-1/2} \exp\left\{-\frac{n}{2}(t-\mu)^2\right\}} \\ &= \sqrt{n}(2\pi)^{-(n-1)/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n-1} x_i^2 + \left(\sum_{i=1}^{n-1} x_i - nt\right)^2 - nt^2\right]\right\} \end{aligned}$$

该分布与 μ 无关, 这说明 \bar{x} 是 μ 的充分统计量.

Theorem 3.5.1 — 因子分解定理. 设总体的概率函数为 $f(x; \theta)$, x_1, \dots, x_n 为其样本, 则 $T = T(x_1, \dots, x_n)$ 为充分统计量的充分必要条件是: 存在如下两个函数 $g(t, \theta)$, 它是通过统计量 T 的取值 t 而依赖于样本的函数 $h(x_1, \dots, x_n)$, 它是样本的函数, 与 θ 无关. 使得

$$f(x_1, \dots, x_n; \theta) = g(T(x_1, \dots, x_n), \theta) h(x_1, \dots, x_n)$$

■ **Example 3.5 例 5.5.5** 设 x_1, \dots, x_n 是取自总体 $N(\mu, \sigma^2)$ 的样本, $\theta = (\mu, \sigma^2)$ 是未知的, 则联合密度函数为

$$\begin{aligned} p(x_1, \dots, x_n; \theta) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i\right)\right\} \end{aligned}$$

取

$$t_1 = \sum_{i=1}^n x_i, t_2 = \sum_{i=1}^n x_i^2$$

并令

$$g(t_1, t_2, \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2} (t_2 - 2\mu t_1)\right\}, h(X) = 1$$

则由因子分解定理, $T = (t_1, t_2) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ 是充分统计量. 进一步, 我们指出这个统计量与 (\bar{x}, s^2) 是一一对应的, 所以, 正态总体下常用的 (\bar{x}, s^2) 是 $\theta = (\mu, \sigma^2)$ 的充分统计量. 事实上, 本例中不难看出有如下分解:

$$p(x_1, \dots, x_n; \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n(\bar{x}-\mu)^2 + ns^2}{2\sigma^2}\right\}$$

Theorem 3.5.2 充分统计量的一一对应变换仍是充分统计量

Theorem 3.5.3 如果统计量 T 是参数 θ 的充分统计量, 且 $S(t)$ 是单值可逆的, 则 $S(T)$ 也是 θ 的充分统计量。

Theorem 3.5.4 几个充分统计量

1. 设 X_1, \dots, X_n 是来自几何分布的 iid 样本, 即

$$P\{X = x\} = \theta(1 - \theta)^x, x = 0, 1, \dots$$

其中 $0 < \theta < 1$, 则 $T = \sum_{i=1}^n X_i$ 是充分统计量.

2. 设 X_1, \dots, X_n 是来自 Poisson 分布 $P(\lambda)$ 的 iid 样本, 则 $T = \sum_{i=1}^n X_i$ 是 λ 的充分统计量。
3. 设 X_1, \dots, X_n 为来自正态总体 $N(\mu, 1)$ 的 iid 样本, 则 $T = \sum_{i=1}^n X_i$ 是 μ 的充分统计量
4. 因子分解定理可知 $(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2)$ 是 (μ, σ^2) 的充分统计量.
5. Γ 分布: 特别地, 当 α 已知时, 可验证 $T_2 = \sum_{i=1}^n X_i$ 是 λ 的充分统计量; 当 λ 已知时, $T_1 = \prod_{i=1}^n X_i$ 是 α 的充分统计量.

Theorem 3.5.5 考虑 Bernoulli 分布 $b(1, p)$ 的充分统计量.

Proof. 此时随机样本 X_1, \dots, X_n 的联合 CDF 为

$$P\{X_1 = x_1, \dots, X_n = x_n\} = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}$$

于是, 由因子分解定理可知, 如下统计量

$$\begin{aligned} T_1 &= (X_1, \dots, X_n) \\ T_2 &= (X_1 + X_2, X_3, \dots, X_n) \\ &\dots \\ T_n &= X_1 + \dots + X_n \end{aligned}$$

均是充分统计量。 ■

Theorem 3.5.6 — (次序统计量的充分性). 对于分布族 \mathcal{F} , 设 $F \in \mathcal{F}, X_1, \dots, X_n$ 为来自 F 的样本, 只要 X_1, \dots, X_n 是 iid 的, 则不论 \mathcal{F} 如何, 其次序统计量 $X_{(1)}, \dots, X_{(n)}$ 都是充分的.

Definition 3.5.3 — (极小充分统计量). 设 S 是分布族 F 的充分统计量, 如对 F 的任一充分统计量 T , 均存在函数 $f(\cdot)$, 使得 $S = f(T)$, 则称 S 是此分布族 \mathcal{F} 的极小充分统计量. 我们常用的充分统计量基本都是极小的。

3.6 Probability

Definition 3.6.1 Given a sample space S and an associated σ -algebra \mathcal{B} , a probability function is a function P with domain \mathcal{B} that satisfies

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{3.1}$$

3.7 Common Families of Distribution

3.7.1 Exponential Families

Definition 3.7.1 — 指数型分布族. 设 $\mathcal{F} = \{f(x; \theta) : \theta \in \Theta\}$ 是某参数分布族, 如果 $f(x; \theta)$ 可以表示成

$$f(x; \theta) = c(\theta) \exp \left\{ \sum_{i=1}^k c_i(\theta) T_i(x) \right\} h(x)$$

则称此分布族为指数型分布族, 其中 k 为正整数, $c(\theta) > 0, h(x) > 0$ 为了方便, 人们常把上面的指数型密度函数写成如下的典则形式 (canonical form):

$$f(x; \eta) = \exp \left\{ \sum_{i=1}^k \eta_i T_i(x) - a(\eta) \right\} h(x), \eta \in \Xi = \{\eta(\theta) : \theta \in \Theta\}$$

其中 η 称为自然参数, Ξ 称为自然参数空间如果自然参数空间 Ξ 包含一个开集, 则称此指数型分布族为满秩的。

Definition 3.7.2 A family of pdfs is called an exponential family if it can be expressed as

$$f(x|\theta) = h(x)c(\theta) \exp \left(\sum_{i=1}^k w_i(\theta) t_i(x) \right) \quad (3.2)$$

Here $h(x) \geq 0$ and $t_1(x), \dots, t_k(x)$ are real-valued functions of the observation x (they cannot depend on θ), and $c(\theta) \geq 0$ and $w_1(\theta), \dots, w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (they cannot depend on x).

■ **Example 3.6** Many common families introduced in the previous section are exponential families. These include the continuous families-normal, gamma, and beta, and the discrete families-binomial, Poisson, and negative binomial.

■ **Example 3.7** (Binomial exponential family) Let n be a positive integer and consider the binomial (n, p) family with $0 < p < 1$. The pmf is

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} (1-p)^n \left(\frac{p}{1-p} \right)^x \\ &= \binom{n}{x} (1-p)^n \exp \left(\log \left(\frac{p}{1-p} \right) x \right) \end{aligned} \quad (3.3)$$

Define

$$h(x) = \begin{cases} \binom{n}{x} & x = 0, \dots, n \\ 0 & \text{otherwise} \end{cases} \quad c(p) = (1-p)^n, 0 < p < 1$$

$$w_1(p) = \log \left(\frac{p}{1-p} \right), \quad 0 < p < 1, \quad \text{and} \quad t_1(x) = x$$

Then we have

$$f(x|p) = h(x)c(p) \exp [w_1(p)t_1(x)]$$

which is of the form with $k = 1$. In particular, note that $h(x) > 0$ only if $x = 0, \dots, n$ and $c(p)$ is defined only if $0 < p < 1$. This is important, as equation must match for all values of x and is an exponential family only if $0 < p < 1$ (so the functions of the parameter are only defined here). Also, the parameter values $p = 0$

Theorem 3.7.1 If X is a random variable with pdf or pmf of the form 16.39 then

$$\mathrm{E} \left(\sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right) = -\frac{\partial}{\partial \theta_j} \log c(\theta) \quad (3.4)$$

$$\mathrm{Var} \left(\sum_{i=1}^k \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X) \right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\theta) - \mathrm{E} \left(\sum_{i=1}^k \frac{\partial^2 w_i(\theta)}{\partial \theta_j^2} t_i(X) \right) \quad (3.5)$$

■ **Example 3.8** From the 3.7, we have

$$\begin{aligned} \frac{d}{dp} w_1(p) &= \frac{d}{dp} \log \frac{p}{1-p} = \frac{1}{p(1-p)} \\ \frac{d}{dp} \log c(p) &= \frac{d}{dp} n \log(1-p) = \frac{-n}{1-p} \end{aligned} \quad (3.6)$$

then from the previous theorem, we have

$$\mathrm{E} \left(\frac{1}{p(1-p)} X \right) = \frac{n}{1-p} \Rightarrow E(X) = np.$$

■ **Example 3.9** (Normal exponential family) Let $f(x|\mu, \sigma^2)$ be the $n(\mu, \sigma^2)$ family of pdfs, where $\theta = (\mu, \sigma)$, $-\infty < \mu < \infty, \sigma > 0$. Then

$$\begin{aligned} f(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{\mu^2}{2\sigma^2} \right) \exp \left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} \right) \end{aligned}$$

Define

$$\begin{aligned} h(x) &= 1 \text{ for all } x \\ c(\theta) = c(\mu, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{\mu^2}{2\sigma^2} \right), \quad -\infty < \mu < \infty, \sigma > 0 \\ w_1(\mu, \sigma) &= \frac{1}{\sigma^2}, \quad \sigma > 0; \quad w_2(\mu, \sigma) = \frac{\mu}{\sigma^2}, \sigma > 0 \\ t_1(x) &= -x^2/2; \quad \text{and} \quad t_2(x) = x \end{aligned}$$

Then

$$f(x|\mu, \sigma^2) = h(x)c(\mu, \sigma) \exp [w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x)]$$

which is the form 16.39 with $k = 2$. Note again that the parameter functions are defined only over the range of the parameter

Theorem 3.7.2 The set of x for which $f(x|\theta) > 0$ cannot depend on θ in an exponential family. So we introduce indicator function of a set A to deal with this problem

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

■ **Example 3.10**

$$f(x|\theta) = \theta^{-1} \exp(1 - (x/\theta)), 0 < \theta < x < \infty$$

, which wrote with indicator functions as

$$f(x|\theta) = \theta^{-1} \exp\left(1 - \left(\frac{x}{\theta}\right)\right) I_{[\theta, \infty)}(x)$$

the last term has relationship with θ and x , so this pdf is not exponential family.

Definition 3.7.3 An exponential family is sometimes reparameterized as

$$f(x|\eta) = h(x)c^*(\eta) \exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)$$

Here the $h(x)$ and $t_i(x)$ functions are the same as in the original parameterization 16.39. The set

$$\mathcal{H} = \left\{ \eta = (\eta_1, \dots, \eta_k) : \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x)\right) dx < \infty \right\}$$

is called the natural parameter space for the family. For the values of $\eta \in \mathcal{H}$, we must have $c^*(\eta) = [\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x)\right) dx]^{-1}$ to ensure that the pdf integrates to 1. since the original $f(x|\theta)$ in 16.39 is a pdf or pmf, the set $\{\eta = (w_1(\theta), \dots, w_k(\theta)) : \theta \in \Theta\}$ must be a subset of the natural parameter space. But there may be other values of $\eta \in \mathcal{H}$ also. \mathcal{H} is convex.

Definition 3.7.4 A curved exponential family is a family of densities of the form 16.39 for which the dimension of the vector θ is equal to $d < k$. If $d = k$, the family is a **full exponential family**.

■ **Example 3.11** (A curved exponential family) The normal family of Example3.9 is a full exponential family. However, if we assume that $\sigma^2 = \mu^2$, the family becomes curved. We then have

$$\begin{aligned} f(x|\mu) &= \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{(x-\mu)^2}{2\mu^2}\right) \\ &= \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2}\right) \exp\left(-\frac{x^2}{2\mu^2} + \frac{x}{\mu}\right) \end{aligned}$$

For the normal family the full exponential family would have parameter space $(\mu, \sigma^2) = \mathfrak{R} \times (0, \infty)$, while the parameter space of the curved family $(\mu, \sigma^2) = (\mu, \mu^2)$ is a parabola.

■ **Example 3.12** (Noraml approximations)If x_1, \dots, x_n is a sample from a Poisson(λ), then $\bar{X} = \sum_i X_i/n$ is approximately

$$\bar{X} \sim \backslash(\lambda, \lambda/n)$$

a curved exponential family.

Definition 3.7.5 $c_1(p) = \ln(p/(1-p))$ 被称为 logit 函数. 如果有 n 个独立的二项随机变量: $X_i \sim B(n_i, p_i)$ 和一个协变量 Z , 且满足

$$\eta_i = \alpha + \beta Z_i$$

则称上述模型为 Logit 模型, 其中 $\eta_i = \ln(p_i/(1-p_i))$

Definition 3.7.6 一个 PDF $f(x; \theta)$ 的支撑集 (support set) 被定义为

$$S = \{x : f(x; \theta) > 0\}$$

在指数型分布族的定义 1.3.9 中, 我们必须要求其支撑集与参数无关。

3.7.2 Inequalities

Theorem 3.7.3 (Chebychev's Ineq) Let X be a random variable, and let $g(x)$ be a nonnegative function, the for any $r > 0$, we have

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r} \quad (3.7)$$

Proof.

$$\begin{aligned} E[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &\geq \int_{\{x: g(x) \geq r\}} g(x) f_X(x) dx \\ &\geq r \int_{\{x: g(x) \geq r\}} f_X(x) dx \\ &= r P(g(X) \geq r) \end{aligned} \quad (3.8)$$

■ **Example 3.13** When let $g(x) = (x - \mu)^2 / \sigma^2$, and $r = t^2$, then

$$P\left(\frac{(X - \mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{1}{t^2} E\left(\frac{(X - \mu)^2}{\sigma^2}\right) = \frac{1}{t^2} \quad (3.9)$$

$$\begin{aligned} P(|X - \mu| \geq t\sigma) &\leq \frac{1}{t^2} \\ P(|X - \mu| < t\sigma) &\geq 1 - \frac{1}{t^2} \end{aligned} \quad (3.10)$$

From previous formula, we can get *threeσrule* no matter X 's distribution. ■

■ **Example 3.14** (A normal prob ineq) If Z is standard normal, then

$$P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}, \quad \text{for all } t > 0 \quad (3.11)$$

Proof.

$$\begin{aligned} P(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-x^2/2} dx \\ &\leq \frac{1}{\sqrt{2\pi}} \int_t^{\infty} \frac{x}{t} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} \end{aligned} \quad (3.12)$$

since $x/t > 1$ for $x > t$, and $P(|Z| \geq t) = 2P(Z \geq t)$ ■

■

3.7.3 Identities

Theorem 3.7.4 Let $X_{\alpha,\beta}$ denote a gamma(α, β) random variable, where $\alpha > 1$ then for any constant a, b , we have

$$P(a < X_{\alpha,\beta} < b) = \beta(f(a|\alpha, \beta) - f(b|\alpha, \beta)) + P(a < X_{\alpha-1,\beta} < b) \quad (3.13)$$

Proof. By definition and integration by parts $u = x^{\alpha-1}$ and $dv = e^{-x/\beta} dx$ and property $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ ■

3.8 Properties of Random Sample

3.8.1 Convergence Concepts

The main situation to be considered in this section is when the sample size to approach infinity and how the behavior of certain sample quantities as this happens.

Definition 3.8.1 — (Convergence in Probability). A sequence of random variables X_1, X_2, \dots converges in probability to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq \varepsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n \rightarrow \infty} P(|X_n - X| < \varepsilon) = 1 \quad (3.14)$$



i.i.d. is not required.

Definition 3.8.2 — (Almost surely). A sequence of random variables X_1, X_2, \dots converges almost surely to a random variable X if, for every $\varepsilon > 0$

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1 \quad (3.15)$$

A random variable is a real-valued function defined on a sample space S . If a sample space S has elements denoted by s , then $X_n(s)$ and $X(s)$ are all function defined on S . Previous definition states that X_n converges to X almost surely if the function $X_n(s)$ converge to $X(s)$ for all $s \in S$ except perhaps for $s \in N$, where $N \subset S$ and $P(N) = 0$. This converge is similar to pointwise convergence of a sequence of functions, except that the convergence need not occur on a set with probability 0.

■ **Example 3.15** Let the sample space S be the closed interval $[0, 1]$ with the uniform probability distribution. Define random variables $X_n(s) = s + s^n, X(s) = s, s \in [0, 1]. X_n(1) = 2, X(1) = 1, X_n$ converges to X almost surely.

$$\begin{aligned} X_1(s) &= s + I_{[0,1]}(s), & X_2(s) &= s + I_{[0,\frac{1}{2}]}(s), & X_3(s) &= s + I_{[\frac{1}{2},1]}(s) \\ X_4(s) &= s + I_{[0,\frac{1}{3}]}(s), & X_5(s) &= s + I_{[\frac{1}{3},\frac{2}{3}]}(s), & X_6(s) &= s + I_{[\frac{2}{3},1]}(s) \end{aligned} \quad (3.16)$$

$n \rightarrow \infty, P(|X_n - X| \geq \varepsilon) \rightarrow 0$ so X_n converges to X in probability, however X_n does not converge to X almost surely. $s = \frac{3}{8}, X_1(s) = 1\frac{3}{8}, X_2(s) = 1\frac{3}{8}, X_3(s) = \frac{3}{8}, X_4(s) = \frac{3}{8}, X_5(s) = 1\frac{3}{8}, X_6(s) = \frac{3}{8}$ No pointwise convergence occurs for this sequence.

R (里斯) If a sequence converges in prob, it is possible to find a subsequence that converges almost surely.

Theorem 3.8.1 — (Weak Law of Large Numbers). Let X_1, X_2, \dots be i.i.d random variables with $E(X) = \mu$, and $\text{Var}(X) = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$. Then for every $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1 \quad (3.17)$$

that is \bar{X} converges in probability to μ .

Proof. By Chebychev's inequality 3.7.3,

$$P(|\bar{X}_n - \mu| \geq \varepsilon) = P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{E(\bar{X}_n - \mu)^2}{\varepsilon^2} = \frac{\text{Var}\bar{X}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \quad (3.18)$$

■

R (Strong Law of Large Numbers)

$$P\left(\lim_{n \rightarrow \infty} |\bar{X}_n - \mu| < \varepsilon\right) = 1 \quad (3.19)$$

4. 参数估计

Definition 4.0.1 统计中的参数常指以下几种情况

1. 分布中所含的未知参数 θ 及其某个函数 $g(\theta)$
2. 分布的各种特征数, 如期望、方差、中位数等。参数 θ 可能取值的范围 Θ 称为参数空间。

R

参数估计的两种形式: 点估计与区间估计。

Definition 4.0.2 — 点估计. 参数的点估计是指: 对未知参数 θ 选用一个统计量 $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ 的取值作为 θ 的估计值, $\hat{\theta}$ 就是 θ 的点估计(量), 简称估计. 好的点估计来自好的统计思想。

- Definition 4.0.3**
1. 无偏估计 (Unbiased Estimation, 简记为 UE)
 2. 一致最小方差无偏估计 (Uniformly Minimum Variance Unbiased Estimation, 简记为 UMVUE)
 3. 极大似然估计 (Maximum Likelihood Estimation, 简记为 MLE) 等

4.0.1 UE

Definition 4.0.4 — 无偏性与可估参数 (UE). 如果 $T(X)$ 是未知参数 θ 的函数 $g(\theta)$ 的一个估计量, 且满足

$$E_{\theta} T(\mathbf{X}) = g(\theta), \forall \theta \in \Theta$$

则称 T 是 $g(\theta)$ 的 UE, 也称 $g(\theta)$ 是可估的 (Estimable), 其中 E_{θ} 表示期望是在分布 f_{θ} 下进行的。

R

对于正态总体, 我们不难验证, 样本均值 \bar{X} 及样本方差 S_n^2 分别是总体均值与方差的 UE

Definition 4.0.5 (渐近 UE) 如果 $T(X)$ 是 $g(\theta)$ 的一个有偏估计量, 且满足

$$\lim_{n \rightarrow \infty} E_\theta T(X_1, \dots, X_n) = g(\theta), \forall \theta \in \Theta$$

则称 T 是 $g(\theta)$ 的渐近 UE.

Theorem 4.0.1 关于无偏估计, 我们有如下几个注解:

1. 如果 $T(X)$ 是 $g(\theta)$ 的有偏估计, 则称 $E_\theta T(X) - g(\theta)$ 为其偏差 (bias)
2. 无偏估计是从多次重复的角度而引出的一个概念, 从期望的定义不难看出, 尽管一次估计, $T(x)$ 的值不一定恰好等于参数真值 $g(\theta)$, 但当大量重复使用时, 其多次估计的平均值即等于参数值 $g(\theta)$
3. 一个参数的无偏估计可能不是唯一的, 也可能不存在, 也可能不合理。

Theorem 4.0.2 对任一总体而言, 样本均值是总体均值的无偏估计。当总体 k 阶矩存在时, 样本 k 阶原点矩 a_k 是总体 k 阶原点矩 μ_k 的无偏估计. 但对 k 阶中心矩则不一样, 譬如, 样本方差 s_*^2 就不是总体方差 σ^2 的无偏估计。

Theorem 4.0.3 — 无偏性不具有不变性. 若 $\hat{\theta}$ 是 θ 的无偏估计, 一般而言, 其函数 $g(\hat{\theta})$ 不是 $g(\theta)$ 的无偏估计, 除非 $g(\theta)$ 是 θ 的线性函数。

■ **Example 4.1** 设总体为 $N(\mu, \sigma^2)$, x_1, \dots, x_n 是样本, 我们已经指出 s^2 是 σ^2 的无偏估计. 下面来考察 s 是否是 σ 的无偏估计. 由定理3.4.2, $Y = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$, 其密度函数为

$$p(y) = \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} y^{-\frac{n-1}{2}} e^{-\frac{y}{2}}, \quad y > 0$$

从而

$$\begin{aligned} E(Y^{1/2}) &= \int_0^\infty y^{1/2} p(y) dy \\ &= 2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right) \int_0^\infty y^{\frac{n-1}{2}-1} e^{-\frac{y}{2}} dy \\ &= \frac{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} = \sqrt{2} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \end{aligned}$$

由此, 我们有

$$Es = \frac{\sigma}{\sqrt{n-1}} E(Y^{1/2}) = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \cdot \sigma \equiv \frac{\sigma}{c_n}$$

Theorem 4.0.4 — (刀切法,Jackknife)--缩小误差. 设 $T(x)$ 是基于样本 $x = (x_1, x_2, \dots, x_n)$ 的关于参数 $g(\theta)$ 的估计量, 且满足 $E_\theta T(x) = g(\theta) + O\left(\frac{1}{n}\right)$. 如以 $x_{(-i)}$ 表示从样本中删去 x_i 后的向量, 则 $T(x)$ 的刀切统计量定义为

$$T_J(x) = nT(x) - \frac{n-1}{n} \sum_{i=1}^n T(x_{(-i)})$$

可以证明: 由上式定义的刀切统计量具有如下性质:

$$E_{\theta} T_J(x) = g(\theta) + O\left(\frac{1}{n^2}\right)$$

并且其方差不会增大.

Question 4.1 减少偏差的方法有哪些: 刀切法和自助法 ■

■ **Example 4.2** 设总体为 $b(1, \theta), x_1, \dots, x_n$ 是其样本, 又设 $g(\theta) = \theta^2$, 则 $T(x) = \bar{x}^2$ 是 $g(\theta)$ 的一个估计,

$$ET(x) = \theta^2 + \frac{\theta(1-\theta)}{n} = g(\theta) + O\left(\frac{1}{n}\right)$$

下面应用刀切法, 注意到

$$T(x_{(-i)}) = \left(\frac{\sum_{j=1}^n x_j - x_i}{n-1} \right)^2 = \frac{n^2 \bar{x}^2 + x_i^2 - 2nx_i \bar{x}}{(n-1)^2}$$

于是

$$\begin{aligned} T_J(x) &= n\bar{x}^2 - \frac{n-1}{n} \sum_{i=1}^n \frac{n^2 \bar{x}^2 + x_i^2 - 2nx_i \bar{x}}{(n-1)^2} \\ &= \frac{n\bar{x}^2}{n-1} - \frac{\sum_{i=1}^n x_i^2}{n(n-1)} \end{aligned}$$

可以验证

$$ET_J(x) = g(\theta)$$

■ **Example 4.3 — 不可估.** 设总体为二点分布 $b(1, p), 0 < p < 1, x_1, \dots, x_n$ 是样本, 说明参数 $\theta = 1/p$ 是不可估的.

Proof. 首先, $T = x_1 + \dots + x_n$ 是充分统计量, $T \sim b(n, p)$. 若有一个 $\hat{\theta} = \hat{\theta}(t)$ 是 θ 的无偏估计, 则有

$$E_{\theta}(\hat{\theta}) = \sum_{i=1}^n \binom{n}{i} \hat{\theta}(i) p^i (1-p)^{n-i} = \frac{1}{p}$$

或

$$\sum_{i=1}^n \binom{n}{i} \hat{\theta}(i) p^{i+1} (1-p)^{n-i} - 1 = 0, \quad 0 < p < 1$$

这是 p 的 $n+1$ 次方程, 最多有 $n+1$ 个实根, 要使它对 $(0, 1)$ 中所有的 p 都成立是不可能的, 故参数 $\theta = 1/p$ 是不可估的. ■

Definition 4.0.6 — 有效性. 设 $\hat{\theta}_1, \hat{\theta}_2$ 是 θ 的两个无偏估计, 如果对任意的 $\theta \in \Theta$ 有

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$$

且至少有一个 $\theta \in \Theta$ 使得上述不等号严格成立, 则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

Definition 4.0.7 — 矩法估计. 利用如下两句“替换原理”而获得的估计量称为矩估计: 用样本矩替换总体矩, 这里的矩可以是原点矩, 也可以是中心矩。用样本矩的函数去替换相应的总体矩的函数。

$$\mu_k = EX^k, \quad v_k = E(X - \mu_1)^k$$

另外, 由大数定律和中心极限定理可知, 样本矩是总体矩的一个很好的估计

1. 在总体分布未知场合, 可用矩法对一些参数作出估计, 如:
 - (a) 用样本均值 \bar{x} 估计总体均值 $E(X)$
 - (b) 用样本方差 s^2 估计总体方差 $Var(X)$
 - (c) 用事件 A 出现的频率估计事件 A 发生的概率 $p(A)$
 - (d) 用样本分位数估计总体分位数。
2. 在总体分布列或分布密度函数形式已知场合, 在有关各阶矩存在的条件下, 用“总体矩等于样本矩”列出矩方程(组), 解之即得分布中未知参数的矩估计. 其中尽量选用低阶矩。

■ **Example 4.4** 考虑总体均值与方差的矩估计(没有要求分布): 设 X_1, \dots, X_n 是一组简单随机样本, 且总体二阶矩存在, 记 $\mu = EX_1, \sigma^2 = Var X_1$, 则由矩估计法知, 其估计方程为

$$\begin{cases} \hat{\mu} = a_1 = \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\mu}_2 = \hat{\mu}^2 + \hat{\sigma}^2 = a_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$

由此可求得总体均值与方差的矩估计为

$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S_n^2 \end{cases}$$

由此例可以看出, 总体均值的矩估计是样本均值, 而总体方差的矩估计却不是样本方差,

Definition 4.0.8 — 相合性. 根据格里纹科定理, 随着样本量的不断增大, 经验分布函数逼近真实分布函数, 因此完全可以要求估计量随着样本量的不断增大而逼近参数真值, 这就是相合性。

设 $\theta \in \Theta$ 为未知参数, $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$ 是 θ 的一个估计量, n 是样本容量, 若对任何一个 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0, \quad \forall \theta \in \Theta$$

则称 $\hat{\theta}_n$ 为参数 θ 的相合估计. 相合性本质上就是按概率收敛, 它是估计量的一个基本要求, 即当样本量不断增大时, 相合估计按概率收敛于未知参数

Definition 4.0.9 设统计量 T_n 是总体参数 $g(\theta)$ 的估计量.

1. 如果当 $n \rightarrow \infty$ 时, T_n 依概率收敛于 $g(\theta)$, 即 $\forall \theta \in \Theta$ 及 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\{|T_n - g(\theta)| \geq \varepsilon\} = 0$$

则称 T_n 是 $g(\theta)$ 的(弱)相合估计.

2. 如果当 $n \rightarrow \infty$ 时, T_n 以概率 1 收敛于 $g(\theta)$, 即 $\forall \theta \in \Theta$, 有

$$P\left\{\lim_{n \rightarrow \infty} T_n = g(\theta)\right\} = 1$$

则称 T_n 是 $g(\theta)$ 的强相合估计.

3. 如果当 $n \rightarrow \infty$ 时, T_n 依 r 阶矩收敛于 $g(\theta)$, 即 $\forall \theta \in \Theta$, 有

$$\lim_{n \rightarrow \infty} E_\theta |T_n - g(\theta)|^r = 0$$

则称 T_n 是 $g(\theta)$ 的 r 阶矩相合估计. 当 $r = 2$ 时, 称为均方相合估计. 由概率论知识知道, 强相合 \Rightarrow 弱相合, r 阶矩相合 \Rightarrow 弱相合, 反之不成立, 且强相合与 r 阶相合之间没有包含关系.

■ **Example 4.5** 设 x_1, x_2, \dots 是来自正态总体 $N(\mu, \sigma^2)$ 的样本序列, 则由辛钦大数定律及依概率收敛的性质知:

1. \bar{x} 是 μ 的相合估计.
2. s_*^2 是 σ^2 的相合估计.
3. s^2 也是 σ^2 的相合估计.

可见参数的相合估计不止一个

Theorem 4.0.5 设 X_1, \dots, X_n 是来自分布族 $\{f(x, \theta) : \theta \in \Theta\}$ 的 iid 样本, 且 $E|X_1|^p < \infty$ (p 为正整数), 则样本的 k ($1 \leq k \leq p$) 阶原点矩是总体 k 阶原点矩的相合估计, 即

$$a_k = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} \mu_k = EX^k$$

Theorem 4.0.6 如果 T_n 是 $g(\theta)$ 的相合估计, c_n, d_n 是两个常数列, 且 $\lim_n c_n = 0, \lim_n d_n = 1$ 则 $d_n T_n + c_n$ 也是 $g(\theta)$ 的相合估计.

Theorem 4.0.7 如果 T_n 是 $g(\theta)$ 的渐近无偏估计, 且满足

$$\lim_{n \rightarrow \infty} \text{Var}_\theta T_n = 0, \forall \theta \in \Theta$$

则 T_n 既是 $g(\theta)$ 的相合估计, 也是均方相合估计.

Theorem 4.0.8 如果 T_n 是 θ 的相合估计, $g(x)$ 在点 $x = \theta$ 处连续, 则 $g(T_n)$ 也是 $g(\theta)$ 的相合估计.

Theorem 4.0.9 — 讨论 MLE 的相合性. 为此假设 X_1, \dots, X_n 为来自 $f(x; \theta)$ 的 IID 样本, 为简单起见, 我们只考虑单参数的情况, 设参数空间 Θ 是一个开区间, 对数似然函数为 $l(\theta; x) = \sum_{i=1}^n \ln f(x_i; \theta)$ 如 $\ln f(x; \theta)$ 在 Θ 上可微, 并设 $f(x; \theta)$ 是可识别的, 即 $\forall \theta \neq \theta', \{x; f(x; \theta) \neq f(x; \theta')\}$ 不是零测集. 则似然方程在 $n \rightarrow \infty$ 时以概率 1 有解, 且此解关于 θ 是相合的.

(R) 定理给出了似然方程有解且其解是相合的条件. 这个解有可能是局部极大值.

4.0.2 判断相合性的一些定理

Theorem 4.0.10 设 $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$ 是 θ 的一个估计量, 若

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta, \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$$

则 $\hat{\theta}_n$ 是 θ 的相合估计.

Proof. 对任意的 $\varepsilon > 0$, 由切比雪夫不等式有

$$P\left(\left|\hat{\theta}_n - E\hat{\theta}_n\right| \geq \frac{\varepsilon}{2}\right) \leq \frac{4}{\varepsilon^2} \text{Var}(\hat{\theta}_n)$$

另一方面, 由 $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$ 可知, 当 n 充分大时有

$$\left|E\hat{\theta}_n - \theta\right| < \frac{\varepsilon}{2}$$

注意到此时如果

$$\left|\hat{\theta}_n - E\hat{\theta}_n\right| < \frac{\varepsilon}{2}$$

就有

$$\left|\hat{\theta}_n - \theta\right| \leq \left|\hat{\theta}_n - E\hat{\theta}_n\right| + \left|E\hat{\theta}_n - \theta\right| < \varepsilon$$

故

$$\left\{\left|\hat{\theta}_n - E\hat{\theta}_n\right| < \frac{\varepsilon}{2}\right\} \subset \left\{\left|\hat{\theta}_n - \theta\right| < \varepsilon\right\}$$

等价地

$$\left\{\left|\hat{\theta}_n - E\hat{\theta}_n\right| \geq \frac{\varepsilon}{2}\right\} \supset \left\{\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right\}$$

由此即有

$$P\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq P\left(\left|\hat{\theta}_n - E\hat{\theta}_n\right| \geq \frac{\varepsilon}{2}\right) \leq \frac{4}{\varepsilon^2} \text{Var}(\hat{\theta}_n) \rightarrow 0(n \rightarrow \infty)$$

定理得证. ■

Theorem 4.0.11 若 $\hat{\theta}_{n1}, \dots, \hat{\theta}_{nk}$ 分别是 $\theta_1, \dots, \theta_k$ 的相合估计, $\eta = g(\theta_1, \dots, \theta_k)$ 是 $\theta_1, \dots, \theta_k$ 的连续函数, 则

$$\hat{\eta}_n = g(\hat{\theta}_{n1}, \dots, \hat{\theta}_{nk})$$

是 η 的相合估计.

Proof. 由函数 g 的连续性, 对任意给定的 $\varepsilon > 0$, 存在一个 $\delta > 0$, 当 $|\hat{\theta}_j - \theta_j| < \delta, j = 1, \dots, k$, 有

$$|g(\hat{\theta}_{n1}, \dots, \hat{\theta}_{nk}) - g(\theta_1, \dots, \theta_k)| < \varepsilon$$

又由 $\hat{\theta}_{n1}, \dots, \hat{\theta}_{nk}$ 的相合性, 对给定的 $\delta > 0$, 对任意给定的 $v > 0$, 存在正整数 N , 使得 $n \geq N$ 时

$$P\left(\left|\hat{\theta}_{nj} - \theta_j\right| \geq \delta\right) < v/k, \quad j = 1, \dots, k$$

从而有

$$P\left(\bigcap_{j=1}^k \left\{\left|\hat{\theta}_{nj} - \theta_j\right| < \delta\right\}\right) = 1 - P\left(\bigcup_{j=1}^k \left\{\left|\hat{\theta}_{nj} - \theta_j\right| \geq \delta\right\}\right)$$

$$\begin{aligned} &\geq 1 - \sum_{j=1}^k P(|\hat{\theta}_{nj} - \theta_j| \geq \delta) \\ &> 1 - k \cdot v/k = 1 - v \end{aligned}$$

根据 (6.2.3), $\bigcap_{j=1}^k \{|\hat{\theta}_{mj} - \theta_j| < \delta\} \subset \{|\hat{\eta}_n - \eta| < \varepsilon\}$, 故有

$$P(|\hat{\eta}_n - \eta| < \varepsilon) > 1 - v$$

由 v 的任意性, 定理得证. ■

Proposition 4.0.12 矩法估计一般都是相合估计.

1. 样本均值是总体均值的相合估计.
2. 样本标准差是总体标准差的相合估计.
3. 样本变异系数 s/\bar{x} 是总体变异系数的相合估计

(3) 大数定律。

Definition 4.0.10 — 相合渐近正态估计. (CAN) 设 T_n 是参数 $g(\theta)$ 的相合估计量, 如存在与样本容量 n 有关的定义于参数空间 Θ 上的函数 $\mu(\theta), \sigma_n(\theta)$, 且 $\sigma_n(\theta) > 0$, 使得当 $n \rightarrow \infty$ 时有

$$\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \xrightarrow{\mathcal{L}} N(0, 1)$$

则称 T_n 为 $g(\theta)$ 的 CAN 估计, 也称 T_n 渐近正态 $N(\mu_n, \sigma_n^2)$, 记为 $T_n \sim AN(\mu_n, \sigma_n^2)$ 关于统计量 T_n 的渐近正态性, 由概率论知识知道, 其渐近的均值与方差显然不是唯一的。

Theorem 4.0.13 If T_n is $AN(\mu_n, \sigma_n^2)$, then also T_n is $AN(\mu'_n, \sigma'^2_n)$ if and only if

$$\frac{\sigma'_n}{\sigma_n} \rightarrow 1, \quad \frac{\mu'_n - \mu_n}{\sigma_n} \rightarrow 0$$

If T_n is $AN(\mu_n, \sigma_n^2)$, then also $a_n T_n + b_n$ is $AN(\mu_n, \sigma_n^2)$ if and only if

$$a_n \rightarrow 1, \quad \frac{\mu_n(a_n - 1) + b_n}{\sigma_n} \rightarrow 0$$

利用概率论知识, 可以证明样本均值、样本方差、样本标准差都是渐近正态的。

Theorem 4.0.14 — 样本分位数的相合渐近正态性. 设 ξ_p 表示总体的 p 分位数, $f(x)$ 表示总体的 PDF. 如果 $f(\xi_p) > 0$ 且 $f(x)$ 在 ξ_p 点连续, 则当 $n \rightarrow \infty$ 时, 有

$$\sqrt{n}(m_{n,p} - \xi_p) \xrightarrow{\mathcal{L}} N(0, p(1-p)/f^2(\xi_p))$$

4.1 极大似然

R 矩估计不要求分布; 似然估计一定要知道总体分布

Definition 4.1.1 (似然函数 (Likelihood Function)) 对于分布族 $\{f(x, \theta), \theta \in \Theta\}$, 如以 $f(\mathbf{x}, \theta)$ 记其 n 个样本的联合概率分布, 则对于给定的样本观测值 $x = (x_1, \dots, x_n)$, 我们称 $f(\mathbf{x}, \theta)$ 为参数的似然函数, 简称为似然函数, 并记之为

$$L(\theta, \mathbf{x}) = f(\mathbf{x}, \theta), \forall \theta \in \Theta$$

Table 4.1: 常见的似然估计
极大似然估计

分布	矩估计	
0-1 分布	$\hat{p} = \bar{\xi}$	$\hat{p} = \bar{\xi}$
$b(n, p)$	$\hat{p} = \frac{n}{\sum \xi_i}$	$\hat{p} = \frac{n}{\sum \xi_i}$
$P(\lambda)$	$\hat{\lambda} = \bar{\xi}$	$\hat{\lambda} = \bar{\xi}$
$N(\mu, \sigma^2)$	$\hat{\mu} = \bar{\xi}$	$\hat{\mu} = \bar{\xi}$
	$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$	$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$
$E(\lambda)$	$\hat{\lambda} = \frac{1}{\bar{\xi}}$	$\hat{\lambda} = \frac{1}{\bar{\xi}}$
$U(0, \theta)$	$\hat{\theta} = 2\bar{\xi}$	$\hat{\theta} = \xi_{(n)}$

称 $\ln L(\theta, \mathbf{x})$ 为对数似然函数, 记为 $l(\theta, \mathbf{x})$ 或 $l(\theta)$

Theorem 4.1.1 从上一定义可以看出, 似然函数与样本联合概率分布相同, 但二者的含义却不同: 后者是固定参数值为 θ 下关于样本 \mathbf{x} 的函数, 它的取值空间为样本空间 \mathcal{X} ; 而似然函数则是固定样本值 \mathbf{x} 下关于参数 θ 的函数, 其在参数空间 Θ 上取值. 为考察二者的区别, 我们不妨把参数 θ 样本分别看作“原因”和“结果”. 当给定参数后, 样本联合分布将告诉我们哪个样本将以多大的概率被观测到; 反过来, 当有了样本后, 似然函数将告诉我们如何最有可能地取参数的估计。

Theorem 4.1.2 — 最大似然估计. 最大似然估计, 利用“最大似然原理”获得的估计, 它只能在总体概率函数形式已知的情况下使用, 具体步骤如下: 设总体的概率函数为 $p(x; \theta), \theta \in \Theta, x_1, \dots, x_n$ 是来自该总体的样本. 。

1. 写出似然函数

$$L(\theta) = L(\theta; x_1, \dots, x_n) = p(x_1; \theta) \cdot p(x_2; \theta) \cdots \cdots p(x_n; \theta)$$

2. 使似然函数 $L(\theta)$ 达到最大的统计量 $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ 称为 θ 的最大似然估计, 简称 MLE, 即 $\hat{\theta}$ 满足

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

注意: 使得对数似然函数 $\ln L(\theta)$ 最大的 $\hat{\theta}$ 也使似然函数 $L(\theta)$ 最大, 寻找最大值可以从定义出发, 也可以对 $l(\theta) = \ln L(\theta)$ 使用微分法, 后者更为常用.

(R)

1. MLE 可能不唯一 (前面已经讲过)
2. MLE 依赖于总体的分布函数, 如不知样本的分布, 则无法求得其感兴趣参数的 MLE.

■ **Example 4.6** 例 6.3.4 对正态总体 $N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$ 是二维参数, 设有样本 x_1, \dots, x_n ,

则似然函数及其对数分别为

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ \ln L(\mu, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi) \end{aligned}$$

将 $\ln L(\mu, \sigma^2)$ 分别关于两个分量求偏导并令其为 0 即得到似然方程组

$$\begin{aligned} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} &= \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0 \end{aligned}$$

解此方程组，可得 μ 的最大似然估计为

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

σ^2 的最大似然估计

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s_*^2$$

利用二阶导函数矩阵的非正定性可以说明上述估计使得似然函数取极大值。

R 正态的两个估计相同；均匀的两个不同

Theorem 4.1.3 均匀分布总体 $U(0, \theta)$ ，参数 θ 的矩估计和 MLE 是不一样的。为了对二者作一比较，我们记

$$\hat{\theta}_M = 2\bar{X}, \quad \hat{\theta}_L = X_{(n)}$$

R 我们不能利用微分法求其 MLE，其原因是此时似然函数的支撑集依赖于未知参数 θ

Theorem 4.1.4 — MLE 一是充分统计量的函数. 事实上，如果 T 是充分统计量，则由因子分解定理可知， $f(\mathbf{x}; \theta) = g(\theta, T(\mathbf{x}))h(\mathbf{x})$ ，于是， $l(\theta; \mathbf{x}) = \ln g(\theta, T(\mathbf{x})) + \ln h(\mathbf{x})$ ，由此可见，为使对数似然达到最大，只需使 $g(\theta, T(\mathbf{x}))$ 达到最大，所以，MLE 肯定是充分统计量 $T(x)$ 的函数。

Theorem 4.1.5 — 最大似然估计的不变性. 若 $\hat{\theta}$ 是 θ 的最大似然估计，则对任一函数 $g(\theta), g(\hat{\theta})$ 是其最大似然估计。如 $g(\theta)$ 是 1-1 映射，且 $\hat{\theta}$ 是 θ 的 MLE，则可以证明 $g(\hat{\theta})$ 也是 $g(\theta)$ 的 MLE。

Theorem 4.1.6 对于指数族分布, 似然方程如有解, 则必唯一. 对于单参数指数族 $f(x; \theta) = \exp\{\theta T(x) + a(\theta)\}h(x)$, 其似然方程为

$$\frac{\partial l(\theta; x)}{\partial \theta} = \sum T(x_i) + na'(\theta) = 0$$

由于 $ET(X) = -a'(\theta)$, $\text{Var } T(X) = -a''(\theta) > 0$ (利用 $\int f(x; \theta)dx = 1$, 之后两边关于 θ 求导即可) 故如上式有解, 必唯一。如果上式无解, 则其 MLE 在 Θ 的边界达到。

(R) 一般的 MLE 不是唯一的

- **Example 4.7**
 1. 标准差 σ 的 MLE 是 $\hat{\sigma} = s$
 2. 概率 $P(X < 3) = \Phi\left(\frac{3-\mu}{\sigma}\right)$ 的 MLE 是 $\Phi\left(\frac{3-\bar{x}}{s_*}\right)$
 3. 总体 0.90 分位数 $x_{0.90} = \mu + \sigma \cdot u_{0.90}$ 的 MLE 是 $\bar{x} + s_* u_{0.90}$, 其中 $u_{0.90}$ 为标准正态分布的 0.90 分位数。

Theorem 4.1.7 — EM 算法. 当分布中有讨厌参数或数据为截尾或缺失时, 其 MLE 的求取是比较困难的。EM 算法, 其含义是把求 MLE 的过程分两步走,

1. 第一步求期望 (E 步), 以便把讨厌的部分去掉: 在已有观测数据 y 及第 i 步估计值 $\theta = \theta^{(i)}$ 的条件下, 求基于完全数据的对数似然函数的期望 (即把其中与 z 有关的部分积分掉):

$$Q(\theta|y, \theta^{(i)}) = E_z l(\theta; y, z) \quad (4.1)$$

右边的期望是关于 Z 在 $\theta = \theta^{(i)}$ 的条件下求取的, 而其余的参数不变, 故左边与 $\theta^{(i)}$ 有关。

2. 第二步求极大值 (M 步). 重复使用这两步直至收敛可得 MLE 的近似解: 求 $Q(\theta|y, \theta^{(i)})$ 关于 θ 的最大值 $\theta^{(i+1)}$, 即找 $\theta^{(i+1)}$ 使得

$$Q(\theta^{(i+1)}|y, \theta^{(i)}) = \max_{\theta} Q(\theta|y, \theta^{(i)}) \quad (4.2)$$

这样就完成了由 $\theta^{(i)}$ 到 $\theta^{(i+1)}$ 的一次迭代. 重复 4.1 和 4.2 式, 直至收敛即可得到 θ 的 MLE.

- **Example 4.8** 设一次试验可能有四个结果, 其发生的概率分别为 $\frac{1}{2} - \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1+\theta}{4}, \frac{\theta}{4}$, 其中 $\theta \in (0, 1)$, 现进行了 197 次试验, 四种结果的发生次数分别为 75, 18, 70, 34. 试求 θ 的 MLE. 解以 y_1, y_2, y_3, y_4 表示四种结果发生的次数, 此时总体分布为多项分布故其似然函数

$$\begin{aligned} L(\theta; y) &\propto \left(\frac{1}{2} - \frac{\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1+\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ &\propto (2-\theta)^{y_1} (1-\theta)^{y_2} (1+\theta)^{y_3} \theta^{y_4} \end{aligned}$$

要由此式求解 θ 的 MLE 是比较麻烦的

我们可以通过引入 2 个变量 z_1, z_2 后, 使得求解变得比较容易. 现假设第一种结果可以分成两部分, 其发生概率分别为 $\frac{1-\theta}{4}$ 和 $\frac{1}{4}$, 令 z_1 和 $y_1 - z_1$ 分别表示落入这两部分的次数; 再假设第三种结果分成两部分, 其发生概率分别为 $\frac{\theta}{4}$ 和 $\frac{1}{4}$ 令 z_2 和 $y_3 - z_2$ 分别表示落入这两部分的次数. 显然, z_1, z_2 是我们人为引入的, 它是不可观测的 (在文献中称之为 latent variable,

即潜变量). 也称数据 (y, z) 为完全数据 (complete data), 而观测到的数据 y 称为不完全数据. 此时, 完全数据的似然函数

$$\begin{aligned} L(\theta; y, z) &\propto \left(\frac{1}{4}\right)^{y_1-z_1} \left(\frac{1-\theta}{4}\right)^{z_1+y_2} \left(\frac{1}{4}\right)^{y_3-z_2} \left(\frac{\theta}{4}\right)^{z_2+y_4} \\ &\propto \theta^{z_2+y_4} (1-\theta)^{z_1+y_2} \end{aligned}$$

其对数似然为

$$l(\theta; y, z) = (z_2 + y_4) \ln \theta + (z_1 + y_2) \ln(1 - \theta)$$

当 y 及 θ 已知时,

$$z_1 \sim b\left(y_1, \frac{1-\theta}{2-\theta}\right), z_2 \sim b\left(y_3, \frac{\theta}{1+\theta}\right)$$

以 z_1 为例说明它的分布. A_1 表示第一种结果出现, B_1, B_2 分别表示我们所定义的两个事件, $A_1 = B_1 \cup B_2$. 由定义知它们是独立的, 且

$$P(A_1) = \frac{1}{2} - \frac{\theta}{4}, P(B_1) = \frac{1-\theta}{4}, P(B_2) = \frac{1}{4}, \text{故 } P(B_1|A_1) = \frac{P(B_1)}{P(A_1)} = \frac{1-\theta}{2-\theta}$$

从而在 $Y_1 = y_1$ 的条件下

$$z_1 \sim b\left(y_1, \frac{1-\theta}{2-\theta}\right)$$

故

$$\begin{aligned} Q(\theta|y, \theta^{(i)}) &= \left(E(z_2|y, \theta^{(i)}) + y_4\right) \ln \theta + \left(E(z_1|y, \theta^{(i)}) + y_2\right) \ln(1 - \theta) \\ &= \left(\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4\right) \ln \theta + \left(\frac{1-\theta^{(i)}}{2-\theta^{(i)}} y_1 + y_2\right) \ln(1 - \theta) \end{aligned}$$

其 M 步即为上式两边关于 θ 求导, 并令其等于 0, 即

$$\frac{\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4}{\theta^{(i+1)}} - \frac{\frac{1-\theta^{(i)}}{2-\theta^{(i)}} y_1 + y_2}{1-\theta^{(i+1)}} = 0$$

解之, 得如下迭代公式:

$$\theta^{(i+1)} = \frac{\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4}{\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4 + \frac{1-\theta^{(i)}}{2-\theta^{(i)}} y_1 + y_2}$$

R 在很一般的条件下, EM 算法是收敛的

Definition 4.1.2 参数 θ 的相合估计 $\hat{\theta}_n$ 称为渐近正态的, 若存在趋于 0 的非负常数序列 $\sigma_n(\theta)$, 使得

$$\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)}$$

依分布收敛于标准正态分布. 这时也称 $\hat{\theta}_n$ 服从渐近正态分布

$$N(\theta, \sigma_n^2(\theta))$$

记为

$$\hat{\theta}_n \sim AN(\theta, \sigma_n^2(\theta)) . \sigma_n^2(\theta)$$

称为 $\hat{\theta}_n$ 的渐近方差.

上述定义中趋于 0 的数列 $\sigma_n(\theta)$ 表示着估计量 $\hat{\theta}_n(x)$ 依概率收敛于 θ 的速度.

1. 若 $\sigma_n(\theta)$ 趋于 0 的速度过快, 则其比值会趋于 ∞ ;

2. 若 $\sigma_n(\theta)$ 趋于 0 的速度过慢, 则其比值会趋于 0;

3. 只有当 $\sigma_n(\theta)$ 趋于 0 的速度不快不慢时, 其比值才可能按分布收敛于 $N(0, 1)$.

所以 $\sigma_n(\theta)$ 趋于 0 的速度就是 $\hat{\theta}_n$ 依概率收敛于 θ 的速度. 故把 $\sigma_n^2(\theta)$ 称为渐近方差是适当的.

■ **Example 4.9** 设总体为泊松分布 $P(\lambda)$, λ 的估计: 样本均值, 即

$$\hat{\lambda}_n = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

由中心极限定理

$$(\hat{\lambda}_n - \lambda) / \sqrt{\lambda/n}$$

依分布收敛于 $N(0, 1)$, 因此, $\hat{\lambda}_n$ 是渐近正态的, 且

$$\hat{\lambda}_n \sim AN(\lambda, \lambda/n)$$

这里常数序列 $\sigma_n(\lambda) = \sqrt{\lambda/n} \rightarrow 0$. 它表示 $\hat{\lambda}_n$ 依概率收敛于 λ 的速度为 $1/\sqrt{n}$, 大多数渐近正态估计都是以 $1/\sqrt{n}$ 速度依概率收敛于被估参数。

Theorem 4.1.8 — MLE 的渐近正态性. 设总体 X 有密度函数 $p(x; \theta)$, $\theta \in \Theta$, Θ 为非退化区间, 假定

1. 对任意的 x , 偏导数 $\frac{\partial \ln p}{\partial \theta}$, $\frac{\partial^2 \ln p}{\partial \theta^2}$ 和 $\frac{\partial^3 \ln p}{\partial \theta^3}$ 对所有 $\theta \in \Theta$ 都存在
2. $\forall \theta \in \Theta$, 有

$$\left| \frac{\partial p}{\partial \theta} \right| < F_1(x), \quad \left| \frac{\partial^2 p}{\partial \theta^2} \right| < F_2(x), \quad \left| \frac{\partial^3 \ln p}{\partial \theta^3} \right| < F_3(x)$$

其中函数 $F_1(x), F_2(x), F_3(x)$ 满足

$$\int_{-\infty}^{\infty} F_1(x) dx < \infty, \quad \int_{-\infty}^{\infty} F_2(x) dx < \infty \quad \sup_{\theta \in \Theta} \int_{-\infty}^{\infty} F_3(x) p(x; \theta) dx < \infty$$

3.

$$\forall \theta \in \Theta, 0 < I(\theta) \equiv \int_{-\infty}^{\infty} \left(\frac{\partial \ln p}{\partial \theta} \right)^2 p(x; \theta) dx < \infty$$

若 x_1, \dots, x_n 是来自该总体的样本, 则存在未知参数 θ 的最大似然估计

$$\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$$

且 $\hat{\theta}_n$ 具有相合性和渐近正态性, 即

$$\hat{\theta}_n \sim AN\left(\theta, \frac{1}{nI(\theta)}\right)$$

其中 n 为样本容量

$$I(\theta) = \int_{-\infty}^{\infty} \left(\frac{\partial \ln p}{\partial \theta} \right)^2 p(x; \theta) dx$$

为费希尔信息量.

■ **Example 4.10** 设 x_1, \dots, x_n 是来自 $N(\mu, \sigma^2)$ 的样本, 可以验证该总体分布在 σ^2 已知或 μ 已知时均满足定理4.1.8的三个条件.

1. 在 σ^2 已知时, μ 的 MLE 为 $\hat{\mu} = \bar{x}$, 由定理4.1.8 知, $\hat{\mu}$ 服从渐近正态分布下面求 $I(\mu)$

$$\ln p(x) = -\ln \sqrt{2\pi} - \frac{1}{2} \ln \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2} \frac{\partial \ln p}{\partial \mu} = \frac{x - \mu}{\sigma^2} I(\mu) = E \left(\frac{x - \mu}{\sigma^2} \right)^2 = \frac{1}{\sigma^2}$$

从而有 $\hat{\mu} \sim AN(\mu, \sigma^2/n)$, 这与 $\hat{\mu}$ 的精确分布相同.

2. 在 μ 已知时, σ^2 的 MLE 为 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$, 下求 $I(\sigma^2)$

$$\begin{aligned} \frac{\partial \ln p}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} (x - \mu)^2 = \frac{(x - \mu)^2 - \sigma^2}{2\sigma^4} \\ I(\sigma^2) &= \frac{E[(X - \mu)^2 - \sigma^2]^2}{4\sigma^8} \\ &= \frac{\text{Var}((X - \mu)^2)}{4\sigma^8} = \frac{1}{2\sigma^4} \end{aligned}$$

从而有 $\hat{\sigma}^2 \sim AN(\sigma^2, 2\sigma^4/n)$

4.2 统计量评估标准--最小方差无偏估计

相合性和渐近正态性是在大样本场合下评价估计好坏的两个重要标准, 在样本量不是很大时, 人们更加倾向于使用一些基于小样本的评价标准, 对无偏估计使用方差, 对有偏估计使用均方误差。

Definition 4.2.1 — 均方误差. 设 $\hat{\theta}$ 是 θ 的一个估计 (无偏的或有偏的), 则称

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + (E\hat{\theta} - \theta)^2$$

为 $\hat{\theta}$ 的均方误差. 均方误差较小意味着: $\hat{\theta}$ 不仅方差较小, 而且偏差 $(E\hat{\theta} - \theta)$ 也小, 所以均方误差是评价点估计的最一般标准。

(R)

1. 使均方误差一致最小的估计量一般是不存在的, 但两个估计好坏可用均方误差评估
2. 在无偏估计类中使均方误差最小就是使方差最小。

Definition 4.2.2 设 X_1, \dots, X_n 是来自分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$ 中某一分布的样本, $g(\theta)$ 是一参数函数, 以 (g) 表示用来估计 $g(\theta)$ 的某些估计量的集合, 如果存在一个 $T^* \in \mathcal{E}(g)$, 使得对 任 $T \in \mathcal{E}(g)$ 均有

$$E_\theta(T^* - g(\theta))^2 \leq E_\theta(T - g(\theta))^2, \forall \theta \in \Theta$$

则称 T^* 为 $g(\theta)$ 的在 $\mathcal{E}(g)$ 中的一致最小均方误差估计, 也称满足 (2.4.3) 式的统计量 T^* 在均方意义下优于 T . 当 T 是 $g(\theta)$ 的 UE 时, 其 MSE 就是它的方差。

Definition 4.2.3 — 一致最小均方误差估计. 设有样本 x_1, \dots, x_n , 对待估参数 θ , 设有一个估计类, 称 $\hat{\theta}(x_1, \dots, x_n)$ 是该估计类中 θ 的一致最小均方误差估计, 如果对该估计类中另外任意一个 θ 的估计 $\tilde{\theta}$, 在参数空间 Θ 上都有

$$\text{MSE}_\theta(\hat{\theta}) \leq \text{MSE}_\theta(\tilde{\theta})$$

一致最小均方误差估计通常是在一个确定的估计类中进行的, 若不对估计加以限制 (即考虑所有的估计) 可能不存在。

■ **Example 4.11** 例 2.4.5 考虑正态总体 $N(\mu, \sigma^2)$ 中关于 σ^2 的形如 cS_n^2 的一致最小均方误差估计. 解在前面我们讲过样本方差 S_n^2 是 σ^2 的一个 UE, 且 $\text{Var} S_n^2 = \frac{2\sigma^4}{n-1} = \text{MSE}(S_n^2)$. 下面我们在形如 cS_n^2 的估计类中找一个 MSE 最小的估计量。事实上, 由于

$$\begin{aligned} \text{MSE}(cS_n^2) &= E(cS_n^2 - \sigma^2)^2 \\ &= E[c(S_n^2 - \sigma^2) - \sigma^2(1-c)]^2 = \sigma^4 \left[\frac{2c^2}{n-1} + (1-c)^2 \right] \end{aligned}$$

如令 $f(c) = \frac{2c^2}{n-1} + (1-c)^2$, 则知它在 $c = \frac{n-1}{n+1}$ 处达到最小 $\frac{2}{n+1}$. 于是, 如取

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

则知 $\hat{\sigma}^2$ 不是 σ^2 的 UE, 但

$$\frac{2\sigma^4}{n+1} = \text{MSE}(\hat{\sigma}^2) < \text{MSE}(S_n^2) = \frac{2\sigma^4}{n-1}$$

此例说明, 无偏准则与均方误差准则是从两个不同角度考察一个估计量好坏的。但当二者发生矛盾时, 更应重视均方误差准。我们应注意, 这里所说的一致并不是指在估计类 (g) 中一致, 而是指式关于 $\theta \in \Theta$ 一致. 因此, 有时一致最小均方误差估计并不存在. 此时, 为了找到一个一致最小均方误差估计, 通常都是通过来缩小估计类来求得。

一致最小均方误差估计对于无偏估计来说就是一致最小方差无偏估计

Definition 4.2.4 — 一致最小方差无偏估计. 设 $\hat{\theta}$ 是 θ 的一个无偏估计, 如果对另外任意一个 θ 的无偏估计 $\bar{\theta}$, 在参数空间 $\Theta = \{\theta\}$ 上都有

$$\text{Var}_\theta(\hat{\theta}) \leq \text{Var}_\theta(\bar{\theta})$$

则称 $\hat{\theta}$ 是 θ 的一致最小方差无偏估计, 简记为 UMVUE

Theorem 4.2.1 — 一致最小方差无偏估计判断准则. 设 $X = (x_1, \dots, x_n)$ 是来自某总体的一个样本, $\hat{\theta} = \hat{\theta}(X)$ 是 θ 的一个无偏估计, $\text{Var}(\hat{\theta}) < \infty$. 则 $\hat{\theta}$ 是 θ 的 UMVUE 的充要条件是, 对任意一个满足 $E(\varphi(X)) = 0$ 和 $\text{Var}(\varphi(X)) < \infty$ 的 $\varphi(X)$, 都有

$$\text{Cov}_\theta(\hat{\theta}, \varphi) = 0, \quad \forall \theta \in \Theta$$

Proof. 先证充分性. 对 θ 的任意一个无偏估计 $\tilde{\theta}$, 令 $\varphi = \tilde{\theta} - \hat{\theta}$, 则

$$E(\varphi) = E(\tilde{\theta}) - E(\hat{\theta}) = 0$$

于是

$$\begin{aligned}\text{Var}(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 \\ &= E[(\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta)]^2 \\ &= E(\varphi^2) + \text{Var}(\hat{\theta}) + 2\text{Cov}(\varphi, \hat{\theta}) \\ &\geq \text{Var}(\hat{\theta})\end{aligned}$$

这表明 $\hat{\theta}$ 在 θ 无偏估计类中方差一致最小.

采用反证法证必要性. 设 $\hat{\theta}$ 是 θ 的 UMVUE, $\varphi(x)$ 满足 $E_\theta(\varphi(x)) = 0, \text{Var}_\theta(\varphi(x)) < \infty$, 尚若在参数空间 Θ 中有一个 θ_0 使得 $\text{Cov}_{\theta_0}(\hat{\theta}, \varphi(x)) \stackrel{\Delta}{=} a \neq 0$, 取 $b = -\frac{a}{\text{Var}_{\theta_0}(\varphi(x))} \neq 0$ 则

$$b^2 \text{Var}_{\theta_0}(\varphi(x)) + 2ab = b(-a + 2a) = -\frac{a^2}{\text{Var}_{\theta_0}(\varphi(x))} < 0$$

令 $\tilde{\theta} = \hat{\theta} + b\varphi(x)$, 则

$$E_\theta(\tilde{\theta}) = E_\theta(\hat{\theta}) + bE_\theta(\varphi(x)) = \theta$$

这说明 $\tilde{\theta}$ 也是 θ 的无偏估计但其方差

$$\begin{aligned}\text{Var}_{\theta_0}(\tilde{\theta}) &= E_{\theta_0}(\hat{\theta} + b\varphi(x) - \theta)^2 \\ &= E_{\theta_0}(\hat{\theta} - \theta)^2 + b^2 E_{\theta_0}(\varphi(x))^2 + 2bE_{\theta_0}((\hat{\theta} - \theta)\varphi(x)) \\ &= \text{Var}_{\theta_0}(\hat{\theta}) + b^2 \text{Var}_{\theta_0}(\varphi(x)) + 2ab \\ &< \text{Var}_{\theta_0}(\hat{\theta})\end{aligned}$$

这与 $\hat{\theta}$ 是 θ 的 UMVUE 矛盾, 这就证明了对参数空间 Θ 中任意的 θ 都有 $\text{Cov}_\theta(\hat{\theta}, \varphi(x)) = 0$. ■

■ **Example 4.12** 设 x_1, \dots, x_n 是来自指数分布 $\text{Exp}(1/\theta)$ 的样本, 则根据因子分解定理可知, $T = x_1 + \dots + x_n$ 是 θ 的充分统计量, 由于 $ET = n\theta$, 所以 $\bar{x} = T/n$ 是 θ 的无偏估计. 设 $\varphi = \varphi(x_1, \dots, x_n)$ 是 0 的任一无偏估计, 则

$$E\varphi(T) = \int_0^\infty \cdots \int_0^\infty \varphi(x_1, \dots, x_n) \cdot \prod_{i=1}^n \left\{ \frac{1}{\theta} \cdot e^{-x_i/\theta} \right\} dx_1 \cdots dx_n = 0$$

即

$$\int_0^\infty \cdots \int_0^\infty \varphi(x_1, \dots, x_n) \cdot e^{-(x_1+\dots+x_n)/\theta} dx_1 \cdots dx_n = 0$$

两端对 θ 求导, 得

$$\int_0^\infty \cdots \int_0^\infty \frac{n\bar{x}}{\theta^2} \varphi(x_1, \dots, x_n) \cdot e^{-(x_1+\dots+x_n)/\theta} dx_1 \cdots dx_n = 0$$

这说明 $E(\bar{x} \cdot \varphi) = 0$, 从而

$$\text{Cov}(\bar{x}, \varphi) = E(\bar{x} \cdot \varphi) - E(\bar{x}) \cdot E(\varphi) = 0$$

由定理4.2.1, x 是 θ 的 UMVUE

Theorem 4.2.2 — 利用充分统计量对无偏估计取期望可以一直构造方差小的统计量. 设总体概率函数是 $p(x; \theta)$, x_1, \dots, x_n 是其样本, $T = T(x_1, \dots, x_n)$ 是 θ 的充分统计量, 则对 θ 的任一无偏估计 $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$, 令 $\tilde{\theta} = E(\hat{\theta}|T)$ 则 $\tilde{\theta}$ 也是 θ 的无偏估计, 且

$$\text{Var}(\tilde{\theta}) \leq \text{Var}(\hat{\theta})$$

Proof. 由于 $T = T(x_1, \dots, x_n)$ 是充分统计量, 故而 $\tilde{\theta} = E(\hat{\theta}|T)$ 与 θ 无关, 因此它也是 θ 的一个估计(统计量), 根据重期望公式, 有

$$E(\tilde{\theta}) = E[E(\hat{\theta}|T)] = E(\hat{\theta}) = \theta$$

故 $\tilde{\theta}$ 是 θ 的无偏估计. 再考察其方差

$$\begin{aligned}\text{Var}(\tilde{\theta}) &= E[(\hat{\theta} - \tilde{\theta})^2] \\ &= E(\hat{\theta} - \tilde{\theta})^2 + E(\tilde{\theta} - \theta)^2 - 2E[(\hat{\theta} - \tilde{\theta})(\tilde{\theta} - \theta)]\end{aligned}$$

由于

$$\begin{aligned}E[(\hat{\theta} - \tilde{\theta})(\tilde{\theta} - \theta)] &= E\{E[(\hat{\theta} - \tilde{\theta})(\tilde{\theta} - \theta)|T]\} \\ &= E\{(\tilde{\theta} - \theta) \cdot E[(\hat{\theta} - \tilde{\theta})|T]\} = 0\end{aligned}$$

由此即有

$$\text{Var}(\tilde{\theta}) = E(\hat{\theta} - \tilde{\theta})^2 + \text{Var}(\tilde{\theta})$$

由于上式右端第一项非负, 这就证明了第二个结论。 ■

(R) 如果无偏估计不是充分统计量的函数, 则将之对充分统计量求条件期望可以得到一个新的无偏估计, 该估计的方差比原来的估计的方差要小, 从而降低了无偏估计的方差. 换言之, 考虑 θ 的估计问题只需要在基于充分统计量的函数中进行即可。

$$\hat{\theta} = E(\hat{\theta}_1|T=t)$$

θ_1 无偏估计不好, 只需要对充分统计量 T 求条件期望.

Theorem 4.2.3 设 $T(X)$ 是参数分布族 $\mathcal{F} = \{f(x; \theta) : \theta \in \Theta\}$ 的一个充分统计量, $S(\mathbf{X})$ 是参数 $g(\theta)$ 的一个UE, 则

$$h(T) = E[S(\mathbf{X})|T(\mathbf{X})]$$

是 $g(\theta)$ 的一个UE, 且

$$\text{Var}_{\theta}(h(T)) \leq \text{Var}_{\theta}(S(\mathbf{X})), \forall \theta \in \Theta$$

其中等号成立的充要条件是 $P\{S(\mathbf{X}) = h(T)\} = 1$

Proof. 因为 T 是充分统计量, 故由充分统计量的定义知, 在给定 T 下, 样本 $\mathbf{X} = (X_1, \dots, X_n)$ 的条件分布与参数 θ 无关. 因此, 由定义的 $h(T)$ 是一统计量. 另外, 由于

$$E_{\theta}(h(T)) = E_{\theta}[E(S(\mathbf{X})|T(\mathbf{X}))] = E_{\theta}(S(\mathbf{X})) = g(\theta), \forall \theta \in \Theta$$

故 $h(T)$ 也是 $g(\theta)$ 的一个 UE. 又因为

$$\begin{aligned}\text{Var}_{\theta}(S(\mathbf{X})) &= E_{\theta}[S(\mathbf{X}) - g(\theta)]^2 \\ &= E_{\theta}[S(\mathbf{X}) - h(T) + h(T) - g(\theta)]^2 \\ &= E_{\theta}[S(\mathbf{X}) - h(T)]^2 + \text{Var}_{\theta}(h(T)) + 2E_{\theta}[(S(\mathbf{X}) - h(T))(h(T) - g(\theta))]\end{aligned}$$

又由于

$$\begin{aligned}E_{\theta}[(S(\mathbf{X}) - h(T))(h(T) - g(\theta))] &= E_{\theta}\{E_{\theta}[(S(\mathbf{X}) - h(T))(h(T) - g(\theta))|T]\} \\ &= E_{\theta}\{(h(T) - g(\theta)) \cdot E_{\theta}(S(\mathbf{X}) - h(T))|T\} \\ &= E_{\theta}\{(h(T) - g(\theta)) \cdot E_{\theta}[(S(\mathbf{X}) - h(T))|T]\} \\ &= E_{\theta}\{(h(T) - g(\theta)) \cdot [(E_{\theta}S(\mathbf{X})|T) - h(T)]\} = 0\end{aligned}$$

故

$$\text{Var}_{\theta}(S(\mathbf{X})) = E_{\theta}(S(\mathbf{X}) - h(T))^2 + \text{Var}_{\theta}(h(T)) \geq \text{Var}_{\theta}(h(T)), \forall \theta \in \Theta$$

且等号成立当且仅当 $E_{\theta}(S(\mathbf{X}) - h(T))^2 = 0$, 即 $P\{S(\mathbf{X}) = h(T)\} = 1$

从定理可以看出, 当我们找到充分统计量后, 任一个 UE 都可以得到改进, 并且, 为找到 UMVUE, 我们仅需在基于充分统计量的无偏估计类中去找即可. ■

Definition 4.2.5 为此, 我们引入如下两个无偏估计类:

$$\begin{aligned}U &= \{T : E_{\theta}T = g(\theta), E_{\theta}T^2 < \infty, \forall \theta \in \Theta\} \\ U_0 &= \{T : E_{\theta}T = 0, E_{\theta}T^2 < \infty, \forall \theta \in \Theta\}\end{aligned}$$

它们分别表示 $g(\theta)$ 与 0 的具有二阶矩的无偏估计类.

Theorem 4.2.4 对于参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$, 设 $g(\theta)$ 可估. 则估计量 $T_0 \in U$ 是 $g(\theta)$ 的一个 UMVUE 的充要条件是

$$E_{\theta}(v T_0) = 0, \forall \theta \in \Theta, \forall v \in U_0$$

Proof. 先证必要性: 反证 设 $T_0 \in U$ 是 $g(\theta)$ 的一个 UMVUE, 但条件不成立, 即存在 $v_0 \in U_0$ 和 $\theta_0 \in \Theta$, 使得

$$E_{\theta_0}(v_0 T_0) \neq 0$$

因为 $v_0 \in U_0$, 故对于任意的 λ , 有 $T_{\lambda} = T_0 - \lambda v_0 \in U$, 而

$$E_{\theta_0}T_{\lambda}^2 = E_{\theta_0}(T_0 - \lambda v_0)^2 = E_{\theta_0}T_0^2 + \lambda^2 E_{\theta_0}v_0^2 - 2\lambda E_{\theta_0}(T_0 v_0)$$

由于 $E_{\theta_0}(v_0 T_0) \neq 0$, 则一定能找到 $\lambda_0 (= E_{\theta_0}(v_0 T_0) / E_{\theta_0}v_0^2)$, 使得

$$E_{\theta_0}T_{\lambda_0}^2 < E_{\theta_0}T_0^2$$

即 $\text{Var}_{\theta_0}T_{\lambda_0} < \text{Var}_{\theta_0}T_0$, 与 T_0 是一个 UMVUE 矛盾. 下证充分性. 设 T_0 满足条件式, 则对于任一 $T \in U$, 由于 $T - T_0 \in U_0$, 则

$$E_{\theta}[(T - T_0)T_0] = 0, \forall \theta \in \Theta$$

即

$$E_\theta T_0^2 = E_\theta(T_0 T), \forall \theta \in \Theta$$

于是由 Cauchy-Schwartz 不等式知,

$$E_\theta T_0^2 \leq (E_\theta T_0^2)^{1/2} (E_\theta T^2)^{1/2}, \forall \theta \in \Theta$$

即

$$E_\theta T_0^2 \leq E_\theta T^2, \forall \theta \in \Theta$$

又由于 T_0, T 均是 $g(\theta)$ 的 UE, 且 T 是任意的, 故由上式可知, T_0 是 $g(\theta)$ 的 UMVUE. ■

Theorem 4.2.5 如果 $g(\theta)$ 的无偏估计类 U 非空, 则其 UMVUE 最多有一个.

Proof. 如果 $T \neq T_0$ 是 $g(\theta)$ 的两个 UMVUE, 则必有

$$E_\theta T = E_\theta T_0 = g(\theta), \text{Var}_\theta T_0 = \text{Var}_\theta T, \forall \theta \in \Theta$$

于是, $T - T_0 \in U_0$. 因为 T 是一个 UMVUE, 故由上述定理知

$$E_\theta [(T - T_0) T] = 0, \forall \theta \in \Theta$$

又由于 T_0 也是一个 UMVUE, 故由定理知

$$E_\theta [(T - T_0) T_0] = 0, \forall \theta \in \Theta$$

综合以上两式, 我们有

$$E_\theta (T - T_0)^2 = E_\theta [(T - T_0)(T - T_0)] = 0, \forall \theta \in \Theta$$

故 $P_\theta \{T = T_0\} = 1, \forall \theta \in \Theta$ 从证明知道, UMVUE 在以概率 1 相等的意义下是唯一的. ■

Corollary 4.2.6 — UMVUE 的线性函数还是 UMVUE. 如果 T_1 和 T_2 分别是 $g_1(\theta)$ 和 $g_2(\theta)$ 的 UMVUE, 则对于任给的常数 $a, b, aT_1 + bT_2$ 是 $ag_1(\theta) + bg_2(\theta)$ 的 UMVUE.

■ **Example 4.13** 设 X_1, \dots, X_n 为来自正态总体 $N(\mu, \sigma^2)$ 的 iid 样本, 试求 μ, σ^2 的 UMVUE.

Proof. 此时样本的联合 PDF 为

$$f(\mathbf{x}; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

则任给 $v \in U_0$, 有

$$\int \cdots \int v(\mathbf{x}) \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} d\mathbf{x} = 0, \forall \mu, \sigma^2 \quad (4.3)$$

上式两边关于 μ 求导, 综合 4.3 式后, 得到

$$\int \cdots \int v(\mathbf{x}) \left(\sum_{i=1}^n x_i \right) \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} d\mathbf{x} = 0, \forall \mu, \sigma^2 \quad (4.4)$$

如取 $T(\mathbf{X}) = \sum_{i=1}^n X_i/n = \bar{X}$, 则上式即为

$$\int \cdots \int v(\mathbf{x}) T(\mathbf{x}) f(\mathbf{x}; \mu, \sigma^2) d\mathbf{x} = 0, \forall \mu, \sigma^2$$

又由于 $T(X)$ 是 μ 的 UE, 故由定理16.0.26知, $T(\mathbf{X}) = \bar{X}$ 是 μ 的 UMVUE. 在4.4式两边再关于 μ 求导, 综合4.3式后, 有

$$\int \cdots \int v(\mathbf{x}) T^2(\mathbf{x}) f(\mathbf{x}; \mu, \sigma^2) d\mathbf{x} = 0, \forall \mu, \sigma^2$$

另外, 在4.3式两边关于 σ^2 求导, 有

$$\int \cdots \int v(\mathbf{x}) \left[\sum_{i=1}^n (x_i - \mu)^2 \right] f(\mathbf{x}; \mu, \sigma^2) d\mathbf{x} = 0, \forall \mu, \sigma^2$$

综合三式, 我们得到

$$\int \cdots \int v(\mathbf{x}) \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] f(\mathbf{x}; \mu, \sigma^2) d\mathbf{x} = 0, \forall \mu, \sigma^2$$

又由于 $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的 UE, 则由 (*₄) 式和定理16.0.26知, 样本方差 S_n^2 是总体方差 σ^2 的 UMVUE. ■



[UMVUE 一定是无偏的] 从上例可以看出, 由于修正的样本方差 S_n^{*2} 不是 σ^2 的 UE, 故它肯定不是 σ^2 的 UMVUE.

■ **Example 4.14** 设 X_1, \dots, X_n 为来自指数分布族 $E(\lambda)$ 的 iid 样本, 试求总体均值 $1/\lambda$ 的 UMVUE.

Proof. 可知, 对于某分布族, 当其充分统计量存在时, 我们只需要在基于充分统计量的无偏估计类中求 UMVUE 即可。对于本例, 我们可由因子分解定理容易验证 $T = \sum_{i=1}^n X_i$ 是其充分统计量。于是, 我们将只在充分统计量的无偏估计类中求取 $1/\lambda$ 的 UMVUE. $T = \sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$, 因此, T 的 PDF 为

$$f_T(x; \lambda) = \frac{1}{(n-1)!} \lambda^n x^{n-1} e^{-\lambda x}, x > 0$$

且易证 T/n 是 $1/\lambda$ 的 UE. $\forall v(T) \in U_0$, 即 $v(T)$ 是 T 的函数且为 0 的无偏估计, 即

$$0 = E v(T) = \frac{1}{(n-1)!} \int_0^\infty v(x) \lambda^n x^{n-1} e^{-\lambda x} dx, \forall \lambda > 0$$

也即

$$0 = \int_0^\infty v(x) x^{n-1} e^{-\lambda x} dx$$

上式两边关于 λ 求导, 有

$$\int_0^\infty x v(x) x^{n-1} e^{-\lambda x} dx = 0, \forall \lambda > 0$$

于是, 由定理16.0.26知, $T = \bar{X}$ 是 $1/\lambda$ 的 UMVUE. ■

■ **Example 4.15** 设 X_1, \dots, X_n 为来自 $U(0, \theta)$ 的 iid 样本, 试求 θ 的 UMVUE.

Proof. $T = X_{(n)}$ 是 θ 的充分统计量, 且 $X_{(n)}$ 的 PDF 为

$$f_T(t; \theta) = nt^{n-1}/\theta^n, 0 < t < \theta$$

由此容易验证 $\frac{n+1}{n}T$ 是 θ 的一个 UE, 下面验证它也是 θ 的 UMVUE. 为此, 任给 0 的一个 UE, $v(T)$ 即

$$\int_0^\theta v(t)nt^{n-1}/\theta^n dt = 0, \forall \theta > 0$$

两边关于 θ 求导, 得

$$v'(\theta) = 0, \forall \theta > 0$$

即此时 0 的无偏估计肯定为 0, 于是由定理 16.0.26 可知, $\frac{n+1}{n}X_{(n)}$ 是 θ 的 UMVUE. 同样, 因为 $EX_1 = \frac{\theta}{2}$, 于是可知总体均值 $\theta/2$ 的 UMVUE 为 $\frac{n+1}{2n}X_{(n)}$ ■

Corollary 4.2.7 样本均值不一定是总体均值的 UMVUE.

Question 4.2 零的 UE 就是零, 什么时候这个结论均成立? ■

Theorem 4.2.8 — 充分性原则. 1. 任一参数 θ 的 UMVUE 不一定存在, 若存在, 则它一定是充分统计量的函数 (MLE 和 UMVUE 都是充分统计量的函数)
 2. 若 θ 的某个无偏估计 $\hat{\theta}$ 不是充分统计量 $T = T(x_1, \dots, x_n)$ 的函数, 则通过条件期望可以获得一个新的无偏估计 $\tilde{\theta} = E(\hat{\theta}|T)$, 且方差比原估计的方差要小
 3. 考虑 θ 的估计时, 只需要在其充分统计量的函数中寻找即可, 该说法对所有统计推断都是正确的. 这便是充分性原则。

4.2.1 完备统计量

Question 4.3 1. 对于一个统计量 $S(X)$, 其定义域为 X , 而其期望 $E_\theta S(X)$ 的定义域是 Θ , 这就是说, 这里的期望 E_θ 相当于一个从 X 到 Θ 的变换. 对于一个变换而言, 一个自然问题是: 这个变换是 1-1 的吗?
 2. 我们在前面也说过, 如果充分统计量存在, 则我们可以仅在充分统计量的无偏估计类中去寻找参数的 UMVUE. 如果我们知道基于充分统计量的无偏估计类仅有一个元素, 则它肯定就是 UMVUE. 那何时这个估计类仅有一个元素呢? ■

Definition 4.2.6 — 完备统计量 (Complete Statistics). 引进完备统计量 (Complete Statistics) 的概念. 它是 Lehman 和 Scheffe 于 1950 年提出的. 有的书上也称之为完全统计量. 对于参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$, 设 $T(\mathbf{X})$ 为一统计量如对任何满足条件

$$E_\theta g(T(\mathbf{X})) = 0, \forall \theta \in \Theta$$

的统计量 $g(T)$, 都有

$$P_\theta\{g(T) = 0\} = 1, \forall \theta \in \Theta$$

则称统计量 $T(X)$ 是完备统计量.

Question 4.4 — (完备分布族)---针对分布族. 对于参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$, 如果对于任一函数 $\psi(x)$ 由

$$E_\theta \psi(X) = 0, \forall \theta \in \Theta$$

总可推出 $P_\theta\{\psi(X) = 0\} = 1$, 则称此分布族 \mathcal{F} 是完备的. ■

Theorem 4.2.9 1. 如果 T 是完备的, $S = \psi(T)$, 且 ψ 可测, 则 S 也是完备的;
2. 充分完备统计量是极小充分统计量, 但反之不成立。

■ **Example 4.16** 考虑二项分布族 $\{B(n, p) : 0 < p < 1\}$ 的完备性.

Proof. 设函数 $\psi(x)$ 满足

$$E_p \psi(X) = \sum_{x=0}^n \psi(x) \binom{n}{x} p^x (1-p)^{n-x} = 0, \forall 0 < p < 1$$

令 $\theta = p/(1-p)$, 则 $\theta > 0$, 且上式可以写成

$$\sum_{x=0}^n \psi(x) \binom{n}{x} \theta^x = 0, \forall \theta > 0$$

由于上式左边是一个 n 次多项式, 故有 $\psi(x) = 0, x = 0, 1, \dots, n$, 故二项分布是完备的。 ■

■ **Example 4.17** 正态分布族 $N(0, \sigma^2)$ 是不完备的.

Proof. 我们注意到, 正态分布的 PDF 是偶函数, 故对于任何一个奇函数, 比如 $\psi(x) = x$, 我们均有

$$E_\sigma \psi(X) = 0, \forall \sigma^2 > 0$$

但是 $P_\sigma\{X = 0\} = 0$. 这说明此正态分布族是不完备的. ■

Theorem 4.2.10 设 X_1, \dots, X_n 是来自参数分布族 $\{f(x, \theta) : \theta \in \Theta\}$ 的 iid 样本, T 是 θ 的充分完备统计量。如果 θ 可估, 且 $S(\mathbf{X})$ 是 θ 的一个 UE, 则 $S_0(T) = E[S(\mathbf{X})|T]$ 是 θ 的唯一的 UMVUE.

Proof. 设 S_1, S_2 是 θ 的任意两个 UE, 则由定理 16.0.26 可知, $E(S_1|T), E(S_2|T)$ 均是 θ 的 UE, 即

$$E_\theta [E(S_1|T)] = E_\theta [E(S_2|T)] = \theta, \forall \theta \in \Theta \quad (4.5)$$

且

$$\begin{aligned} \text{Var}_\theta (E(S_1|T)) &\leq \text{Var}_\theta (S_1), \forall \theta \in \Theta \\ \text{Var}_\theta (E(S_2|T)) &\leq \text{Var}_\theta (S_2), \forall \theta \in \Theta \end{aligned}$$

另由 5.3 式可知, $E(S_1|T) - E(S_2|T)$ 是 0 的无偏估计, 即

$$E_\theta [E(S_1|T) - E(S_2|T)] = 0, \forall \theta \in \Theta$$

又由于 T 是完备统计量, 故由定义知

$$P_{\theta}\{E(S_1|T) = E(S_2|T)\} = 1, \forall \theta \in \Theta$$

这说明在概率 1 的意义下, $E(S_1|T)$ 与 $E(S_2|T)$ 相等, 又由于 S_1 与 S_2 的任意性, 故 $S_0 = E(S|T)$ 是唯一的 UMVUE. ■

Theorem 4.2.11 设 $\{f(x, \eta) : \eta \in \Xi\}$ 是 k 个参数的标准指类型分布族, 即

$$f(x, \eta) = c(\eta) \exp \left\{ \sum_{j=1}^k \eta_j T_j(x) \right\} h(x)$$

其中 $\eta = (\eta_1, \dots, \eta_k)' \in \Xi$, 这里

$$\Xi = \left\{ (\eta_1, \dots, \eta_k) : 0 < \int_{\mathcal{X}} \exp \left\{ \sum_{j=1}^k \eta_j T_j(x) \right\} h(x) dx < \infty \right\}$$

为自然参数空间。则

1. 统计量 $T(\mathbf{X}) = (T_1, \dots, T_k)$ 是充分的;
2. 如果 Ξ 包含一个开集 (此时, 称此分布族为满秩的), 则 $T(\mathbf{X}) = (T_1, \dots, T_k)$ 是完备的.

■ **Example 4.18** 考虑正态总体 $N(\mu, \sigma^2)$ 中参数的 UMVUE.

Proof. 容易验证正态分布族是指类型分布, 且 $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ 是 (μ, σ^2) 的充分完备统计量. 于是, 由于样本均值 \bar{X} , 样本方差 S_n^2 分别是 μ, σ^2 的 UE, 故它们是 UMVUE. ■

■ **Example 4.19** 求 Gamma 分布族 $\Gamma(\alpha, \lambda)$ 的充分完备统计量.

Proof. 设 X_1, \dots, X_n 为来自 $\Gamma(\alpha, \lambda)$ 的 iid 样本, 由其联合 PDF 为

$$f(\mathbf{x}; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} = \frac{\lambda^\alpha}{\Gamma(\alpha)} \exp \left\{ (\alpha-1) \sum_{i=1}^n \ln x_i - \lambda \sum_{i=1}^n x_i \right\}$$

于是, 由定理可知, 统计量 $(\sum_{i=1}^n X_i, \sum_{i=1}^n \ln X_i)$ 是 (α, λ) 的充分完备统计量. ■

Theorem 4.2.12 — (次序统计量的完备性). 设 X_1, \dots, X_n 是来自分布族 F 的分布函数为 F 的 iid 样本, 如果

1. F 是凸的;
2. $\forall a < b$, 记 $S = [a, b]$, 由 $F(b) - F(a) > 0$ 可导出 $P\{X_1 < x | X_1 \in S\} \in \mathcal{F}$

则该样本的次序统计量 $X_{(1)}, \dots, X_{(n)}$ 关于分布族 \mathcal{F} 是完备的.

4.2.2 信息量--最小方差的表达式

Definition 4.2.7 — 费希尔信息量 $I(\theta)$. (正则分布) 如果单参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$ 具有如下五个条件:

1. 参数空间 Θ 是直线上的开区间 (有限、无限或半无限);
2. 导数 $\frac{\partial f(x, \theta)}{\partial \theta}$ 存在 ($\forall \theta \in \Theta$)
3. 支撑集与 θ 无关;

4. 其 PDF $f(x, \theta)$ 的积分与微分运算可以互换, 即

$$\frac{d}{d\theta} \int_{-\infty}^{+\infty} f(x, \theta) dx = \int_{-\infty}^{+\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx$$

5.

$$I(\theta) = E_{\theta} \left(\frac{\partial}{\partial \theta} \ln f(X, \theta) \right)^2 \text{ 存在, 且 } I(\theta) > 0$$

则称此分布族为 C-R 正则分布族, 其中条件 (1)-(5) 也称为正则条件, $I(\theta)$ 称为该分布族的 Fisher 信息量 (Information).

则称该期望

$$I(\theta) = E \left[\frac{\partial}{\partial \theta} \ln p(x; \theta) \right]^2$$

为总体分布的费希尔 (Fisher) 信息量。

如果二阶导数对一切 $\theta \in \Theta$ 都存在, 则 $I(\theta)$ 还可用下式计算

$$I(\theta) = -E \left[\frac{\partial^2}{\partial \theta^2} \ln p(x; \theta) \right]$$

R 很多的统计结果都与费希尔信息量有关. 如最大似然估计的渐近方差, 无偏估计的方差的下界等都与费希尔信息量 $I(\theta)$ 有关. $I(\theta)$ 越大”可被解释为总体分布中包含未知参数 θ 的信息越多.

Theorem 4.2.13 — 常用分布的费希尔信息量. 1. 二点分布 $b(1, p)$ 的费希尔信息量 $I(p) = [p(1-p)]^{-1}$

2. 泊松分布 $p(\lambda)$ 的费希尔信息量 $I(\lambda) = \lambda^{-1}$
3. 指数分布 $Exp(\lambda)$ 的费希尔信息量 $I(\lambda) = \lambda^2$
4. 正态分布 $N(\mu, 1)$ 的费希尔信息量 $I(\mu) = 1$
5. 正态分布 $N(0, \sigma^2)$ 的费希尔信息量 $I(\sigma^2) = \frac{1}{2\sigma^4}$
6. 正态分布 $N(\mu, \sigma^2)$ 的费希尔信息量 (信息矩阵)

$$I(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$$

Theorem 4.2.14 可以验证, 我们常用的单参数指数型分布族是正则的。当然, 常用的分布中也有不是正则的, 如均匀分布族 $\{U(0, \theta) : \theta > 0\}$

Theorem 4.2.15 考虑 iid 样本的联合 PDF, 则可以证明

$$E_{\theta} \left(\frac{\partial}{\partial \theta} \ln f(\mathbf{X}, \theta) \right)^2 = nI(\theta)$$

值得注意的是, Fisher 信息量 $I((\theta))$ 依赖于我们所选择的参数. 事实上, 设 $\theta = h(\xi)$, 其

中 h 可微, 则 X 所包含的关于 ξ 的信息量就是

$$I^*(\xi) = I(h(\xi)) [h'(\xi)]^2$$

为理解 Fisher 信息量的意义, 我们先考虑一个估计量的方差下界。设 S 是 $g(\theta)$ 的一个估计, $\psi(X, \theta)$ 是任一具有有限二阶矩的函数, 则容易证明如下的协方差不等式:—施瓦茨不等式

$$\text{Var}(S) \geq \frac{[\text{Cov}(S, \psi)]^2}{\text{Var}(\psi)}$$

上一不等式给出了估计量 S 的方差下界, 但对于实际应用没有任何意义, 由于此不等式右边包含 S . 但若 $\text{Cov}(S, \psi)$ 只与 $E_\theta S = g(\theta)$ 有关, 则上式的确提供了无偏估计 S 的方差的一个下界。于是, 我们有如下结论。

Theorem 4.2.16 (Blyth 定理) $\text{Cov}(S, \psi)$ 只通过 $g(\theta) = E_\theta(S)$ 依赖于 S 的充要条件是: $\forall \theta$, 有

$$\text{Cov}(U, \psi) = 0, \forall U \in U_0$$

其中 U_0 为零的无偏估计类。

Proof. $\text{Cov}(S, \psi)$ 只通过 $g(\theta) = E_\theta(S)$ 依赖于 S 等价于对任何两个满足条件 $E_\theta S_1 =$

$E_\theta S_2$ 的估计 S_1, S_2 , 必有 $\text{Cov}(S_1, \psi) = \text{Cov}(S_2, \psi)$

而对于任两个无偏估计 S_1, S_2

$$0 = \text{Cov}(S_1, \psi) - \text{Cov}(S_2, \psi) = \text{Cov}(S_1 - S_2, \psi) \iff \text{Cov}(U, \psi) = 0, \forall U \in U_0$$

■

Theorem 4.2.17 — C-R 不等式. 设 $T = T(x_1, \dots, x_n)$ 是未知参数 $g(\theta)$ 的一个无偏估计, 若 $g'(\theta) = \frac{\partial g(\theta)}{\partial \theta}$ 存在, 且对 Θ 中一切 θ , 对

$$g(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \dots, x_n) \prod_{i=1}^n p(x_i; \theta) dx_1 \cdots dx_n$$

的微商可在积分号下进行, 即

$$\begin{aligned} g'(\theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \dots, x_n) \frac{\partial}{\partial \theta} \left(\prod_{i=1}^n p(x_i; \theta) \right) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \dots, x_n) \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n p(x_i; \theta) \right] \prod_{i=1}^n p(x_i; \theta) dx_1 \cdots dx_n \end{aligned}$$

则在费希尔信息量 $I(\theta)$ 也存在的条件下4.2.7有

$$\text{Var}(T) \geq [g'(\theta)]^2 / (nI(\theta)) \tag{4.6}$$

上式称为克拉美-罗 ($C - R$) 不等式, $[g'(\theta)]^2 / (nI(\theta))$ 称为 $g(\theta)$ 的无偏估计的方差的 C - R 下界, 简称 $g(\theta)$ 的 C - R 下界. 特别, 对 θ 的无偏估计 $\hat{\theta}$ 有 $\text{Var}(\hat{\theta}) \geq (nI(\theta))^{-1}$

Proof. 证明以连续总体为例加以证明. 由 $\int_{-\infty}^{\infty} p(x_i; \theta) dx_i = 1, i = 1, \dots, n$, 两边对 θ 求导, 由于积分与微分可交换次序, 于是有

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} p(x_i; \theta) dx_i = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta} \ln p(x_i; \theta) \right] p(x_i; \theta) dx_i \\ &= E \left[\frac{\partial}{\partial \theta} \ln p(x_i; \theta) \right] \end{aligned}$$

而

$$\begin{aligned} E(Z^2) &= \text{Var}(Z) = \sum_{i=1}^n \text{Var} \left(\frac{\partial}{\partial \theta} \ln p(x_i; \theta) \right) \\ &= \sum_{i=1}^n E \left[\frac{\partial}{\partial \theta} \ln p(x_i; \theta) \right]^2 = nI(\theta) \end{aligned}$$

又由 $g'(\theta) = E(T \cdot Z) = E((T - g(\theta)) \cdot Z)$, 据施瓦茨不等式, 有

$$[g'(\theta)]^2 \leq E[(T - g(\theta))^2] \cdot E(Z^2) = \text{Var}(T) \text{Var}(Z)$$

由此, 4.6 得证. 关于离散总体可类似证明. 如果4.6中等号成立, 则称 $T = T(x_1, \dots, x_n)$ 是 $g(\theta)$ 的有效估计, 有效估计一定是 UMVUE. ■

(R) $g(\theta)$ 的 $C - R$ 下界并不是对任意参数函数 $g(\theta)$ 的无偏估计的方差都可达到. 但能达到 $C - R$ 下界的 $g(\theta)$ 的估计 $T = T(x_1, \dots, x_n)$ 一定是 $g(\theta)$ 的 UMVUE. 方差达到 $C - R$ 下界的无偏估计称为有效估计.

(R) 且由 Cauchy-Schwartz 不等式知, 上述等号成立的条件为: 存在 $c(\theta) \neq 0$, 使得

$$\frac{\partial \ln f(\mathbf{X}, \theta)}{\partial \theta} = c(\theta)(T(\mathbf{X}) - g(\theta))$$

以概率 1 成立。

(R) 这为寻找一致最小方差无偏估计给出了一种方法. 如果有一个 $g(\theta)$ 的无偏估计 $\hat{g}(X_1, X_2, \dots, X_n)$, 它的方差达到了 $C - R$ 下界. 则该估计就是 $g(\theta)$ 的一致最小方差无偏估计.

(R) 如果不是 iid, $nI(\theta)$ 改为 $E_{\theta} \left(\frac{\partial \ln f(\mathbf{X}, \theta)}{\partial \theta} \right)^2$

Theorem 4.2.18 从定理的证明不难看出, 对于正则分布族, 我们有如下结论:

$$E_{\theta} \left[\frac{\partial}{\partial \theta} \ln f(X, \theta) \right] = 0, \quad I(\theta) = \text{Var}_{\theta} \left[\frac{\partial}{\partial \theta} \ln f(X, \theta) \right] = -E_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \ln f(X, \theta) \right]$$

■ **Example 4.20** 设 X_1, \dots, X_n 为来自 Bernoulli 分布 $b(1, p)$ 的 iid 样本, $p \in (0, 1)$. 试求 p 的 UE 的方差的 C-R 下界。

Proof. 容易验证此分布族是正则的, 且其 Fisher 信息量为

$$I(p) = E_p \left(\frac{\partial \ln f(X, p)}{\partial p} \right)^2 = \sum_{k=0}^1 \left(\frac{\partial \ln f(k, p)}{\partial p} \right)^2 f(k, p)$$

其中 $f(k, p) = P\{X = k\} = p^k(1-p)^{1-k}$. 于是

$$I(p) = \left(\frac{\partial \ln(1-p)}{\partial p} \right)^2 (1-p) + \left(\frac{\partial \ln p}{\partial p} \right)^2 p = \frac{1}{p(1-p)}$$

这样由信息不等式知, p 的 UE 的方差的 C-R 下界为 $\frac{1}{nI(p)} = \frac{p(1-p)}{n}$ 另外, 由于 $\sum_{i=1}^n X_i \sim B(n, p)$, 故 $\text{Var } \bar{X} = \frac{p(1-p)}{n}$ 达到了 C-R 下界, 故可知 \bar{X} 是 p 的 UMVUE. 这与我们前面讲结论一致。 ■

■ **Example 4.21** 设 X_1, \dots, X_n 是来自 $N(\mu, \sigma^2)$ 的 iid 样本, 试考虑 μ 和 σ^2 的 UE 的方差的下界。

Proof. 容易验证此时的正态分布族是正则的, 且因为

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

则

$$\begin{aligned} \frac{\partial \ln f(x, \mu, \sigma^2)}{\partial \mu} &= \frac{x-\mu}{\sigma^2}, & \frac{\partial^2 \ln f(x, \mu, \sigma^2)}{\partial \mu^2} &= -\frac{1}{\sigma^2} \\ \frac{\partial \ln f(x, \mu, \sigma^2)}{\partial \sigma^2} &= \frac{(x-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}, & \frac{\partial^2 \ln f(x, \mu, \sigma^2)}{\partial (\sigma^2)^2} &= -\frac{(x-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4} \\ \frac{\partial^2 \ln f(x, \mu, \sigma^2)}{\partial \mu \sigma^2} &= -\frac{x-\mu}{\sigma^4} \end{aligned}$$

于是, 此时的 Fisher 信息阵为

$$I(\mu, \sigma^2) = -E_{\mu, \sigma^2} \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{X-\mu}{\sigma^4} \\ -\frac{X-\mu}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

由于可知

$$I(\mu) = 1/\sigma^2, I(\sigma^2) = 1/2\sigma^4$$

则 μ 和 σ^2 的 UE 的方差的 C-R 下界分别为 $\sigma^2/n, 2\sigma^4/n$ 因为 $\text{Var } \bar{X} = \sigma^2/n$ 达到了 C-R 下界, 故 \bar{X} 是 μ 的 UMVUE. ■

另外, 由例 2.4.7 知, 样本方差 S_n^2 是 σ^2 的 UMVUE, 但是 $\text{Var}(S_n^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$, 故知 σ^2 的 UMVUE 并没有达到 C-R 下界。上一个例子说明我们不能用一个 UE 的方差是否达到 C-R 下界来判断它是否是 UMVUE, 并且也说明, 定理 2.6.2 与定理 2.6.3 给出的 C-R 下界并不是最大的一个, 还有必要进行改进。另外, 我们也有必要对达到与没有达到 C-R 下界的 UE 进行一定的区别。

■ **Example 4.22** 设总体为正态分布 $N(0, \sigma^2)$, 它满足定义的所有条件, 下面计算它的费希尔信息量. 由于 $p(x; \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}$, 注意到 $x^2/\sigma^2 \sim \chi^2(1)$, 故

$$\begin{aligned} I(\sigma^2) &= E \left[\frac{\partial}{\partial \sigma^2} \ln p(x; \sigma^2) \right]^2 \\ &= E \left(\frac{x^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)^2 \\ &= \frac{1}{4\sigma^4} \text{Var} \left(\frac{x^2}{\sigma^2} \right) = \frac{1}{2\sigma^4} \end{aligned}$$

对于参数 σ^2 , 其 C-R 下界为 $2\sigma^4/n$. 对于它的一个估计量 $T(X) = \sum_{i=1}^n X_i^2/n$, 因为 $E T(X) = \sigma^2$, $\text{Var} T(X) = 2\sigma^4/n$, 故知它是一个 UE 且达到了 C-R 下界, 于是它是 σ^2 的 UMVUE.

若 x_1, \dots, x_n 是样本, 则 σ^2 的无偏估计的 C-R 下界为 $\frac{2\sigma^4}{n}$, 而 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$ 是 σ^2 的无偏估计, 其方差达到了 C-R 下界, 故 $\hat{\sigma}^2$ 是 σ^2 的 UMVUE. 另一方面, 令 $\sigma = g(\sigma^2) = \sqrt{\sigma^2}$, 则 σ 的 C-R 下界为

$$\frac{[g'(\sigma^2)]^2}{nI(\sigma^2)} = \frac{[1/(2\sigma)]^2}{n/(2\sigma^4)} = \frac{\sigma^2}{2n}$$

σ 的无偏估计为

$$\hat{\sigma} = \sqrt{\frac{n}{2} \cdot \frac{\Gamma(n/2)}{\Gamma((n+1)/2)} \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}}$$

可以证明, 这是 σ 的 UMVUE, 且其方差大于 C-R 下界. 这表明所有 σ 的无偏估计的方差都大于其 C-R 下界。

4.2.3 有效估计

Definition 4.2.8 — (有效估计及效率). 设 $T(X)$ 是 $g(\theta)$ 的一个 UE, 则比值

$$e_n = \frac{(g'(\theta))^2 / nI(\theta)}{\text{Var}_\theta T(\mathbf{X})}$$

为 $T(X)$ 的效率. 如果 $e_n = 1$, 则称 $T(X)$ 为 $g(\theta)$ 的有效估计. 如 $\lim_{n \rightarrow \infty} e_n = 1$, 则称 $T(\mathbf{X})$ 为 $g(\theta)$ 的渐近有效估计.

(R) 有的 UMVUE 能达到 C-R 下界, 但有的则不可以,

4.3 贝叶斯估计

Definition 4.3.1

$$P(H_k|A) = \frac{P(H_k)P(A|H_k)}{\sum_i P(H_i)P(A|H_i)}, k = 1, 2, \dots$$

其中 $\cup H_i = \Omega$ 是必然事件的一个划分. 当时我们称 $P(H_k)$ 为先验 (prior), 而 $P(H_k|A)$ 称为后验 (posterior).

- Definition 4.3.2 — 贝叶斯统计推断使用的三种信息.**
1. 总体信息, 总体分布或总体所属分布族提供的信息
 2. 样本信息, 从总体中抽取样本所提供的信息
 3. 先验信息, 在试验前人们对要做的问题在经验上和资料上所占有的信息

Theorem 4.3.1 — 贝叶斯统计的基本观点. 任一未知量 θ 都可看作一个随机变量, 用一个概率分布来描述未知参数是最好的办法, 这个分布称为先验分布。

Definition 4.3.3 — 贝叶斯公式的密度函数形式. 总体依赖于参数 θ 的概率函数在贝叶斯统计中记为 $p(x|\theta)$, 它表示在随机变量 θ 取某个给定值时总体的条件概率函数:

1. 根据参数 θ 的先验信息设法确定先验分布 $\pi(\theta)$
2. 从贝叶斯观点看, 样本 x_1, \dots, x_n 的产生分两步进行. 首先从先验分布 $\pi(\theta)$ 产生一个样本 θ_0 , 然后从 $p(x_1, \dots, x_n|\theta_0)$ 中产生一组样本, 这时样本的联合条件概率函数为

$$p(x_1, \dots, x_n|\theta_0) = \prod_{i=1}^n p(x_i|\theta_0)$$

这个分布综合了总体信息和样本信息。 θ_0 是未知的, 它是按先验分布 $\pi(\theta)$ 产生的。为把先验信息综合进去, 不能只考虑 θ_0 , 对 θ 的其他值发生的可能性也要加以考虑, 故要用 $\pi(\theta)$ 进行综合. 这样一来, 样本 x_1, \dots, x_n 和参数 θ 的联合分布为

$$h(x_1, \dots, x_n, \theta) = p(x_1, \dots, x_n|\theta) \cdot \pi(\theta)$$

这个联合分布把总体信息, 样本信息和先验信息三种可用信息都综合进去了

Definition 4.3.4 — 后验分布. 分析的目的是要对未知参数 θ 作统计推断。

在没有样本信息时, 人们只能依据先验分布对 θ 作出推断。在有了样本观察值 x_1, \dots, x_n 之后, 则应依据 $h(x_1, \dots, x_n, \theta)$ 对 θ 作出推断。由于 $h(x_1, \dots, x_n, \theta)$ 可分解为

$$h(x_1, \dots, x_n, \theta) = \pi(\theta|x_1, \dots, x_n) m(x_1, \dots, x_n)$$

其中

$$m(x_1, \dots, x_n) = \int_{\theta} h(x_1, \dots, x_n, \theta) d\theta = \int_{\theta} p(x_1, \dots, x_n|\theta) \pi(\theta) d\theta$$

是 x_1, \dots, x_n 的边际概率函数, 它与 θ 无关, 不含 θ 的任何信息. 因此能用来对 θ 作出推断的仅是条件分布 $\pi(\theta|x_1, \dots, x_n)$, 它的计算公式是

$$\pi(\theta|x_1, \dots, x_n) = \frac{h(x_1, \dots, x_n, \theta)}{m(x_1, \dots, x_n)} = \frac{p(x_1, \dots, x_n|\theta) \pi(\theta)}{\int_{\theta} p(x_1, \dots, x_n|\theta) \pi(\theta) d\theta}$$

这个条件分布称为 θ 的后验分布, 它集中了总体, 样本和先验中有关 θ 的一切信息。后验分布 $\pi(\theta|x_1, \dots, x_n)$ 的计算公式就是用密度函数表示的贝叶斯公式. 它是用总体和样本对先验分布 $\pi(\theta)$ 作调整的结果, 贝叶斯统计的一切推断都基于后验分布进行。

Theorem 4.3.2 — 后验分布. 贝叶斯估计基于后验分布

$$\pi(\theta|x_1, \dots, x_n)$$

对 θ 所作的贝叶斯估计有多种，常用有如下三种：

1. 使用后验分布的密度函数最大值作为 θ 的点估计，称为最大后验估计
2. 使用后验分布的中位数作为 θ 的点估计，称为后验中位数估计
3. 使用后验分布的均值作为 θ 的点估计，称为后验期望估计。这是使用最为频繁的贝叶斯估计。

Definition 4.3.5 — 共轭先验分布. 设 θ 是总体参数， $\pi(\theta)$ 是其先验分布，若对任意的样本观测值得到的后验分布 $\pi(\theta|X)$ 与 $\pi(\theta)$ 属于同一个分布族，则称该分布族是 θ 的共轭先验分布（族）

1. 二项分布 $b(n, \theta)$ 中的成功概率 θ 的共轭先验分布是贝塔分布 $Be(a, b)$
2. 泊松分布 $P(\theta)$ 中的均值 θ 的共轭先验分布且伽马分布 $Ga(\alpha, \lambda)$
3. 在方差已知时，正态均值 θ 的共轭先验分布是正态分布 $N(\mu, \tau^2)$
4. 在均值已知时，正态方差 σ^2 的共轭先验分布是倒伽马分布 $IGa(\alpha, \lambda)$
5. (若 $X \sim Ga(\alpha, \lambda)$ ，则 X^{-1} 的分布称为倒伽马分布 $IGa(\alpha, \lambda)$)

Definition 4.3.6 — 超参数. 超参数：先验分布中的未知参数称为超参数。应尽力对各种先验信息进行加工获得超参数的估计。

4.3.1 最小二乘估计

介绍最小二乘估计 (Least Square Estimation, 简记为 LSE)、最优线性无偏估计 (Best Linear Unbiased Estimation, 简记为 BLUE) 和加权最小二乘估计 (Weighted LSE.)

Definition 4.3.7 — Gauss-Markov 模型.

$$\begin{cases} \mathbf{y} = X\beta + \varepsilon \\ E\varepsilon = \mathbf{0}, \quad \text{Var } \varepsilon = \sigma^2 \mathbf{I}_n \end{cases}$$

其中 y 为 n 维观测向量， X 为 $n \times p$ 阶设计矩阵， β 为 p 维未知参数， σ^2 未知。

Definition 4.3.8 — (LSE). Gauss-Markov 模型中，如果

$$(\mathbf{y} - X\hat{\beta})'(\mathbf{y} - X\hat{\beta}) = \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)$$

则称令为 β 的 LSE.

Theorem 4.3.3 记 $Q(\beta) = (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta)$ ，则知，求 LSE 等价于求 $Q(\beta)$ 的最小值。解下面的方程：

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X'\mathbf{y} + 2X'X\beta = 0$$

此方程称为 **正规方程** (normal equation)，也即

$$X'X\beta = X'\mathbf{y}$$

如果设计矩阵 X 是列满秩的，则上述的正规方程的解唯一，且为

$$\hat{\beta}_{LS} = (X'X)^{-1}X'\mathbf{y}$$

这即是 β 的 LSE.

可以看出，LSE 是观测向量 y 的线性函数，这一点在以后的证明过程中非常有用。

R 当 X 不是列满秩时, 上式中的逆不存在, 但可以用广义逆代替。

Theorem 4.3.4 对于前面的 Gauss-Markov 模型, 如果设计矩阵 X 是列满秩的, 且以 $\hat{\beta}$ 记 β 的 LSE, 则

1. $\hat{\beta}$ 是 β 的 UE; (如 X 不是列满秩的, 此结论仍然正确)
2. $\text{Var } \hat{\beta} = \sigma^2 (X'X)^{-1}$

Theorem 4.3.5 — (BLUE). 对于前面的 Gauss-Markov 模型, 如果 β 的一个线性 UE $\tilde{\beta} = Cy$ 满足

$$\text{Var } \tilde{\beta} \leq \text{Var}(Ay)$$

则称 $\tilde{\beta}$ 是 β 的 BLUE, 其中 C, A 是 $p \times n$ 阶矩阵, 且满足 $ECy = EAy = \beta$

Theorem 4.3.6 对于前面的 Gauss-Markov 模型, 如果 X 是列满秩的, 则 LSE 是的 BLUE。

Theorem 4.3.7 对于的 Gauss-Markov 模型, 如果 X 是列满秩的, 则 $\forall \mathbf{a} \in \mathbf{R}^p, \mathbf{a}'\hat{\beta}$ 是 $\mathbf{a}'\beta$ 的唯一的 BLUE.

Theorem 4.3.8 我们注意到在 Gauss-Markov 模型中, 还有一个参数 σ^2 仍需估计. 实际上, 如记

$$SSE = (\mathbf{y} - X\hat{\beta})'(\mathbf{y} - X\hat{\beta})$$

它被称为残差平方和, 并且 $SSE = Q(\hat{\beta})$. 如果 X 是列满秩的, 则可以证明

$$\hat{\sigma}^2 = SSE/(n-p)$$

是 σ^2 的 UE. 另外, 由于它是基于 β 的 LSE 得到的, 故称之为 σ^2 的 LSE. 如果已知它是正态的, 则可以证明上述 LSE 是 UMVUE.

Theorem 4.3.9 对于的 Gauss-Markov 模型, 如果 X 是列满秩的, 且 $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_p)$, 则 LSE 式中的 $\hat{\beta}$ 及 $\hat{\sigma}^2$ 分别是 β 和 σ^2 的 UMVUE. 另外, $\forall \mathbf{a}' \in \mathbf{R}^p, \mathbf{a}'\hat{\beta}$ 也是 $\mathbf{a}'\beta$ 的 UMVUE.

Definition 4.3.9 — 加权最小二乘估计--处理不是等方差的. 在上一小节的讨论中, 我们始终假设各观测间的方差是相等的, 但在许多实际问题, 每次观测误差的精度可能是不一样的, 于是本节将讨论这一问题的 LSE。

此时, 我们有如下的广义 Gauss-Markov 模型:

$$\begin{cases} \mathbf{y} = X\beta + \varepsilon \\ E\varepsilon = \mathbf{0}, \quad \text{Var } \varepsilon = \sigma^2 \Sigma \end{cases}$$

其中 $\Sigma > 0$ 为已知正定阵。

对于此模型, 我们仍然可以利用之前的 LSE 公式来作为 β 的估计, 并且也是无偏的, 但它不是 BLUE, 这是由于在证明 BLUE 过程中要用到观测的方差。

对于此广义的 Gauss-Markov 模型, 由于 $\Sigma > 0$, 故存在 n 阶对称阵 B , 使 $\Sigma = B^2$. 作

变换 $z = B^{-1}\mathbf{y}$, $\tilde{X} = B^{-1}X$, 则有如下的 Gauss-Markov 模型:

$$\begin{cases} \mathbf{z} = \tilde{X}\beta + \varepsilon \\ E\varepsilon = \mathbf{0}, \quad \text{Var } \varepsilon = \sigma^2 \mathbf{I}_n \end{cases}$$

这样, 由此模型得到 β 的 LSE 如下:

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'z = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\mathbf{y}$$

我们称之为 β 的加权的或广义的最小二乘估计, 并且由定理4.3.7知, 它也是 BLUE.

4.4 区间估计

Theorem 4.4.1 — 陈家鼎.

1. 极轴量的精确分布与未知参数无关
2. 一个总体的样本均值 \bar{X} 与样本离差 $\sum_1^n (X_i - \bar{X})^2$ 相互独立, 则这个总体必服从正态分布。
- 3.

Theorem 4.4.2 — 正交分布保持正态分布. 设 X_1, X_2, \dots, X_n 相互独立, 且

$$X_i \sim N(\mu_i, \sigma^2) \quad (i = 1, 2, \dots, n)$$

$A = (a_{ij})$ 是 n 阶正交矩阵

$$Y_i = \sum_{k=1}^n a_{ik} X_k \quad (i = 1, 2, \dots, n)$$

则 Y_1, \dots, Y_n 相互独立, 且

$$Y_i \sim N\left(\sum_{k=1}^n a_{ik} \mu_k, \sigma^2\right) \quad (i = 1, 2, \dots, n)$$

Theorem 4.4.3 — t 分布对于正态分布的近似. t 分布的密度函数是 x 的偶函数。当自由度 $n \geq 25$ 时 t 分布的密度曲线与标准正态分布的密度曲线极为接近。可以证明

$$\lim_n p_n(x) = \varphi(x) \quad (\text{一切 } x)$$

这里 $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$ 是标准正态分布密度。证明是不难的, 请读者自己完成 (提示: 用斯特林公式)。

Theorem 4.4.4 — 经验分布函数. (Glivenko - Cantelli) 设:

$$D_n = \sup_x |F_n(x) - F(x)|$$

则

$$P\left(\lim_n D_n = 0\right) = 1$$

Definition 4.4.1 — (区间估计 (interval estimation)). 设 X_1, \dots, X_n 为来自参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$ 的样本, θ 为一维未知参数. 如果 $\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})$ 为两个统计量, 且 $\hat{\theta}_L(\mathbf{X}) \leq \hat{\theta}_U(\mathbf{X})$, 则称随机区间 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 为 θ 的一个区间估计. 既然是估计, 就应有一个好坏的衡量标准. 当参数真值为 θ 时, 我们自然希望随机区间 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 包含 θ 的概率 $P_\theta \{ \hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X}) \}$ 越大越好. 这个概率就称为置信度或置信水平 (confidence level). 由于这个置信度依赖于参数真值, 故我们自然希望对于参数空间 Θ 中的每一个, 其置信水平都很大.

Definition 4.4.2 — (置信系数). 设随机区间 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 为的一个区间估计, 则称

$$\inf_{\theta \in \Theta} P_\theta \{ \hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X}) \}$$

为该区间估计的置信系数.

如果置信度不依赖于未知参数, 则置信系数就是置信度.

Definition 4.4.3 — 置信区间. 设 θ 是总体的一个参数, 其参数空间为 Θ, x_1, \dots, x_n 是来自该总体的样本, 对给定的一个 $\alpha (0 < \alpha < 1)$, 若有两个统计量 $\hat{\theta}_L = \hat{\theta}_L(x_1, \dots, x_n)$ 和 $\hat{\theta}_U = \hat{\theta}_U(x_1, \dots, x_n)$, 使得对任意的 $\theta \in \Theta$, 有

$$P_\theta (\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \geq 1 - \alpha$$

则称随机区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 是 θ 的置信水平为 $1 - \alpha$ 的置信区间, 或简称 $[\hat{\theta}_L, \hat{\theta}_U]$ 是 θ 的 $1 - \alpha$ 置信区间, $\hat{\theta}_L$ 和 $\hat{\theta}_U$ 分别称为 θ 的双侧置信下限和置信上限. 这里置信水平 $1 - \alpha$ 的含义是指在大量使用该置信区间时, 至少有 $100(1 - \alpha)\%$ 的区间含有 θ ; 随机区间 $(\hat{\theta}_1, \hat{\theta}_2)$ 将以概率 $1 - \alpha$ 覆盖参数 θ

Definition 4.4.4 — 同等置信区间. 在上述记号下, 若对给定的 $\alpha (0 < \alpha < 1)$, 对任意的 $\theta \in \Theta$, 有

$$P_\theta (\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

则称 $[\hat{\theta}_L, \hat{\theta}_U]$ 为 θ 的 $1 - \alpha$ 同等置信区间. 同等置信区间是把给定的置信水平 $1 - \alpha$ 用足了. 常在总体为连续分布场合下可以实现.

Definition 4.4.5 — 置信限. 在上述记号下, 若对给定的 $\alpha (0 < \alpha < 1)$ 和任意的 $\theta \in \Theta$, 有

$$P_\theta (\hat{\theta}_L \leq \theta) \geq 1 - \alpha, \quad \forall \theta \in \Theta$$

则称 $\hat{\theta}_L$ 是 θ 的置信水平为 $1 - \alpha$ 的 (单侧) 置信下限. 假如等号对一切 $\theta \in \Theta$ 成立, 则称 $\hat{\theta}_L$ 是 θ 的 $1 - \alpha$ 同等置信下限. 若对给定的 $\alpha (0 < \alpha < 1)$ 和任意的 $\theta \in \Theta$, 有

$$P_\theta (\hat{\theta}_U \geq \theta) \geq 1 - \alpha$$

则称 $\hat{\theta}_U$ 是 θ 的置信水平为 $1 - \alpha$ 的 (单侧) 置信上限. 若等号对一切 $\theta \in \Theta$ 成立则称 $\hat{\theta}_U$ 是 $1 - \alpha$ 同等置信上限.

Definition 4.4.6 定义 3.1.5 (置信域 (Confident Region)) 设 X_1, \dots, X_n 是来自参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta \subset R^k\}$ 的样本, $\theta = (\theta_1, \dots, \theta_k)'$. 如果统计量 $S(\mathbf{X})$ 满足

1. 对任一样本观测值 $x, S(x)$ 是 Θ 的一个子集,
2. 对给定的 $\alpha \in (0, 1), P_\theta \{ \theta \in S(\mathbf{X}) \} \geq 1 - \alpha, \forall \theta \in \Theta$

则称 $S(\mathbf{X})$ 是 θ 的置信水平为 $1 - \alpha$ 的置信域, 而概率 $P_\theta\{\theta \in S(\mathbf{X})\}$ 在 Θ 上的下确界就称为置信系数.

Definition 4.4.7 — 权轴量法. 寻找同等置信区间常采用权轴量法, 其步骤如下:

1. 设法构造一个样本和 θ 的函数 $G = G(x_1, \dots, x_n, \theta)$, 使得 G 的分布不依赖于未知参数. 此种 G 被称为权轴量
2. 适当地选择两个常数 c, d , 使对给定的 $\alpha(0 < \alpha < 1)$, 有 $P(c \leq G \leq d) = 1 - \alpha$
3. 若能将 $c \leq G \leq d$ 进行不等式等价变形化为 $\hat{\theta}_L \leq \theta \leq \hat{\theta}_U$, 则有

$$P_\theta(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha,$$

最后的 $[\hat{\theta}_L, \hat{\theta}_U]$ 就是 θ 的 $1 - \alpha$ 同等置信区间.

关于置信区间的构造有两点说明

1. 满足置信水平要求的 c 与 d 通常不唯一. 若有可能, 应选平均长度 $E(\hat{\theta}_U - \hat{\theta}_L)$ 达到最短的 c 与 d , 这在 G 的分布为对称分布场合通常容易实现.
2. 实际中, 选平均长度 $E(\hat{\theta}_U - \hat{\theta}_L)$ 尽可能短的 c 与 d 往往很难实现, 因此常这样选择 c 与 d , 使得两个尾部概率各为 $\frac{\alpha}{2}$, 即

$$P(G < c) = P(G > d) = \frac{\alpha}{2}$$

这样的置信区间称为等尾置信区间. 这是在 G 的分布为偏态分布场合常采用的方法.

Definition 4.4.8 1. 找一个与待估参数 $g(\theta)$ 无关的统计量 T . 一般是它的一个很好的点估计
2. 设法找出 T 和 $g(\theta)$ 的某函数 $S(T, g(\theta))$, 使得 $S(T, g(\theta))$ 的分布 $F(x)$ 与 θ 无关
3. 适当地选取两个常数 c, d , 使对给定的 $\alpha \in (0, 1)$, 有

$$P_\theta\{c \leq S(T, g(\theta)) \leq d\} = 1 - \alpha,$$

即 $F(d) - F(c) = 1 - \alpha$ (一般取 $d = F_{\alpha/2}, c = F_{1-\alpha/2}$)

4. 如果能把不等式 $c \leq S(T, g(\theta)) \leq d$ 等价地改写成 $\hat{\theta}_L(\mathbf{X}) \leq g(\theta) \leq \hat{\theta}_U(\mathbf{X})$, 其中 $\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})$ 只与 c, d 和 T 有关, 而与 θ 无关. 则 $[\hat{\theta}_L(\mathbf{X}), \hat{\theta}_U(\mathbf{X})]$ 就是 $g(\theta)$ 的置信水平为 $1 - \alpha$ 的置信区间.

Theorem 4.4.5 置信区间的解释: 我们知道 [5.419, 5.613] 是一个具体的区间, 而 μ 是一个虽未知但其值却是确定的数. 于是, 区间 [5.419, 5.613] 或者包含 μ , 或者不包含 μ , 二者必居其一. 那置信水平 0.95 如何解释呢? 事实上, 置信水平的确切含义是: 如果我们把上述试验重复 100 次, 则将得到 100 个这样的置信水平为 0.95 的置信区间. 于是, 我们有理由相信在这 100 个区间中, 至少有 95 次确实包含所要估计的参数的真值. 而一旦得到具体的区间, 我们就不能再说它有 $(1 - \alpha) \times 100\%$ 的可能性包含要估计的参数真值了.

■ **Example 4.23** 设 x_1, \dots, x_n 是来自均匀总体 $U(0, \theta)$ 的一个样本, 试对设定的 $\alpha(0 < \alpha < 1)$ 给出 θ 的 $1 - \alpha$ 同等置信区间. 解我们采用权轴量法分三步进行.

1. 我们已知 θ 的最大似然估计为样本的最大次序统计量 $x_{(n)}$, 而 $x_{(n)} / \theta$ 的密度函数为

$$p(y; \theta) = ny^{n-1}, \quad 0 < y < 1$$

它与参数 θ 无关, 故可取 $x_{(n)} / \theta$ 作为权轴量 G

2. 由于 $x_{(n)} / \theta$ 的分布函数为 $F(y) = y^n, 0 < y < 1$, 故 $P(c \leq x_{(n)} / \theta \leq d) = d^n - c^n$ 因此我

们可以适当的选择 c 和 d 满足

$$d^n - c^n = 1 - \alpha$$

3. 利用不等式变形可容易地给出 θ 的 $1 - \alpha$ 同等置信区间为 $[x_{(n)}/d, x_{(n)}/c]$, 该区间的平均长度为 $(\frac{1}{c} - \frac{1}{d})Ex_{(n)}$. 不难看出, 在 $0 \leq c < d \leq 1$ 及 $d^n - c^n = 1 - \alpha$ 的条件下, 当 $d = 1, c = \sqrt[n]{\alpha}$ 时, $\frac{1}{c} - \frac{1}{d}$ 取最小值, 这说明 $[x_{(n)}, x_{(n)}/\sqrt[n]{\alpha}]$ 是 θ 的此类区间估计中置信水平为 $1 - \alpha$ 最短置信区间.

Theorem 4.4.6 — 常用的置信区间. 1. 设 x_1, \dots, x_n 是来自 $N(\mu, \sigma^2)$ 的样本, \bar{x} 为样本均值, s 为样本标准差, u_p 为标准正态分布的 p 分位数, $t_p(k)$ 为自由度是 k 的 t 分布 $t(k)$ 的 p 分位数, $\chi_p^2(k)$ 为自由度是 k 的 χ^2 分布 $\chi^2(k)$ 的 p 分位数, 取置信水平 $1 - \alpha$, 则

(a) σ 已知时 μ 的置信区间为

$$[\bar{x} - u_{1-\alpha/2}\sigma/\sqrt{n}, \bar{x} + u_{1-\alpha/2}\sigma/\sqrt{n}]$$

(b) σ 未知时 μ 的置信区间为

$$[\bar{x} - t_{1-\alpha/2}s/\sqrt{n}, \bar{x} + t_{1-\alpha/2}s/\sqrt{n}]$$

(c) $\sigma^2(\mu$ 未知) 的置信区间为

$$\left[\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \right]$$

(d) $\sigma(\mu$ 已知) 的置信区间为

$$\left[\frac{s\sqrt{n-1}}{\sqrt{\chi_{1-\alpha/2}^2(n-1)}}, \frac{s\sqrt{n-1}}{\sqrt{\chi_{\alpha/2}^2(n-1)}} \right]$$

Theorem 4.4.7 — σ 已知时 μ 的置信区间. 在这种情况下, 由于 μ 的点估计为 \bar{x} , 其分布为 $N(\mu, \sigma^2/n)$, 因此枢轴量可选为 $G = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, c 和 d 应满足 $P(c \leq G \leq d) = \Phi(d) - \Phi(c) = 1 - \alpha$, 经过不等式变形可得

$$P_\mu(\bar{x} - d\sigma/\sqrt{n} \leq \mu \leq \bar{x} - c\sigma/\sqrt{n}) = 1 - \alpha$$

该区间长度为 $(d - c)\sigma/\sqrt{n}$. 由于标准正态分布为单峰对称的, 在 $\Phi(d) - \Phi(c) = 1 - \alpha$ 的条件下, 当 $d = -c = u_{1-\alpha/2}$ 时, $d - c$ 达到最小, 由此给出了 μ 的 $1 - \alpha$ 同等置信区间为

$$[\bar{x} - u_{1-\alpha/2}\sigma/\sqrt{n}, \bar{x} + u_{1-\alpha/2}\sigma/\sqrt{n}]$$

这是一个以 \bar{x} 为中心, 半径为 $u_{1-\alpha/2}\sigma/\sqrt{n}$ 的对称区间, 常将之表示为 $\bar{x} \pm u_{1-\alpha/2}\sigma/\sqrt{n}$

Theorem 4.4.8 — σ 未知时 μ 的置信区间. 这时可用 t 统计量, 因为 $t = \frac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t(n-1)$, 因此 t 可以用来作为枢轴量. 可得到 μ 的 $1 - \alpha$ 置信区间为

$$\bar{x} \pm t_{1-\alpha/2}(n-1)s/\sqrt{n}$$

此处 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ 是 σ^2 的无偏估计.

Theorem 4.4.9 — μ 已知的情况下 σ^2 的估计. σ^2 可用样本方差 s^2 估计. $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$, 由于 χ^2 分布是偏态分布, 寻找平均长度最短区间很难实现, 一般都改为寻找等尾置信区间: 把 α 平分为两部分, 在 χ^2 分布两侧各截面积为 $\alpha/2$ 的部分, 即采用 χ^2 的两个分位数 $\chi_{a/2}^2(n-1)$ 和 $\chi_{1-\alpha/2}^2(n-1)$ 它们满足

$$P\left(\chi_{a/2}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{1-\alpha/2}^2\right) = 1 - \alpha$$

由此给出 σ^2 的 $1 - \alpha$ 置信区间为

$$\left[(n-1)s^2 / \chi_{1-\alpha/2}^2(n-1), \quad (n-1)s^2 / \chi_{a/2}^2(n-1) \right] \quad (4.7)$$

将 4.7 的两端开方即得到标准差 σ 的 $1 - \alpha$ 置信区间。

Theorem 4.4.10 设 x_1, \dots, x_m 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, \bar{x} 为其样本均值, s_x 为其样本标准差, y_1, \dots, y_n 是来自 $N(\mu_2, \sigma_2^2)$ 的样本, \bar{y} 为其样本均值, s_y 为其样本标准差 $u_p, t_p(k)$ 含义同上, $F_p(k_1, k_2)$ 为自由度是 (k_1, k_2) 的 F 分布 $F(k_1, k_2)$ 的 p 分位数, 取置信水平 $1 - \alpha$, 则

1. σ_1^2 与 σ_2^2 均已知时, $\mu_1 - \mu_2$ 的置信区间为

$$\left[\bar{x} - \bar{y} - u_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}, \quad \bar{x} - \bar{y} + u_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} \right]$$

2. $\sigma_1^2 = \sigma_2^2$ 未知时, $\mu_1 - \mu_2$ 的置信区间为

$$\left[\bar{x} - \bar{y} - \sqrt{\frac{m+n}{mn}} s_w t_{1-\alpha/2}(m+n-2), \quad \bar{x} - \bar{y} + \sqrt{\frac{m+n}{mn}} s_w t_{1-\alpha/2}(m+n-2) \right]$$

其中

$$s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$$

3. $\frac{\sigma_1^2}{\sigma_2^2} = \theta$ 已知时, $\mu_1 - \mu_2$ 的置信区间为

$$\left[\bar{x} - \bar{y} - \sqrt{\frac{m\theta+n}{mn}} s_t t_{1-\alpha/2}(m+n-2), \quad \bar{x} - \bar{y} + \sqrt{\frac{m\theta+n}{mn}} s_t t_{1-\alpha/2}(m+n-2) \right]$$

其中

$$s_t^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2 / \theta}{m+n-2}$$

4. m 与 n 都很大时, $\mu_1 - \mu_2$ 的近似置信区间为

$$\left[\bar{x} - \bar{y} - u_{1-\alpha/2} \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}, \quad \bar{x} - \bar{y} + u_{1-\alpha/2} \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}} \right]$$

5. 一般场合下 $\mu_1 - \mu_2$ 的近似置信区间为

$$[\bar{x} - \bar{y} - s_0 t_{1-\alpha/2}(l), \quad \bar{x} - \bar{y} + s_0 t_{1-\alpha/2}(l)]$$

其中 $s_0^2 = s_x^2/m + s_y^2/n$, $l = \frac{s_0^4}{\frac{s_x^4}{m^2(m-1)} + \frac{s_y^4}{n^2(n-1)}}$

6. 方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间为

$$\left[\frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{1-\alpha/2}(m-1, n-1)}, \quad \frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{\alpha/2}(m-1, n-1)} \right]$$

Theorem 4.4.11 — σ^2 和 σ_2^2 已知时. 此时有 $\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n})$, 取枢轴量为

$$u = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

沿用前面多次用过的方法可以得到 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为

$$\bar{x} - \bar{y} \pm u_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Theorem 4.4.12 — $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知时. 此时有 $\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, (\frac{1}{m} + \frac{1}{n}) \sigma^2)$

$$\frac{(m-1)s_x^2 + (n-1)s_y^2}{\sigma^2} \sim \chi^2(m+n-2)$$

由于 $\bar{x}, \bar{y}, s_x^2, s_y^2$ 相互独立, 故可构造如下服从 t 分布 $t(m+n-2)$ 的枢轴量

$$t = \sqrt{\frac{mn(m+n-2)}{m+n}} \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2}} \sim t(m+n-2)$$

记 $s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$, 则 $\mu_1 - \mu_2$ 的置信区间为

$$\bar{x} - \bar{y} \pm \sqrt{\frac{m+n}{mn}} s_w t_{1-\alpha/2}(m+n-2)$$

Theorem 4.4.13 — $\sigma_2^2/\sigma_1^2 = c$ 已知时. 由于

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right) = N\left(\mu_1 - \mu_2, \sigma_1^2 \left(\frac{1}{m} + \frac{c}{n}\right)\right)$$

有

$$\frac{(m-1)s_x^2 + (n-1)s_y^2/c}{\sigma_1^2} = \frac{(m-1)s_x^2}{\sigma_1^2} + \frac{(n-1)s_y^2}{\sigma_2^2} \sim \chi^2(m+n-2)$$

由于 $\bar{x}, \bar{y}, s_x^2, s_y^2$ 相互独立, 仍可构造如下服从 t 分布 $t(m+n-2)$ 的枢轴量

$$t = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2/c}} \sqrt{\frac{mn(m+n-2)}{mc+n}} \sim t(m+n-2)$$

记 $s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2/c}{m+n-2}$, 则 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为

$$\bar{x} - \bar{y} \pm \sqrt{\frac{mc+n}{mn}} s_w t_{1-\alpha/2}(m+n-2)$$

Theorem 4.4.14 — 当 $m=n$ 时. 因为 $Z_i = Y_i - X_i \sim N(\delta, \sigma_1^2 + \sigma_2^2), i=1, \dots, n$, 且相互独立, 于是知 $\bar{Z} \sim N(\delta, (\sigma_1^2 + \sigma_2^2)/n)$

$$\begin{aligned} \sum_{i=1}^n (Z_i - \bar{Z})^2 / (\sigma_1^2 + \sigma_2^2) &\sim \chi^2(n-1). \text{ 故有} \\ \frac{\sqrt{n(n-1)(\bar{Z}-\delta)}}{\sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}} &\sim t(n-1) \end{aligned}$$

由此可得 δ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\left[\bar{Z} \pm t_{\alpha/2}(n-1) \frac{\sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}}{\sqrt{n(n-1)}} \right]$$

Theorem 4.4.15 — 当 m 和 n 都很大时的近似置信区间. 若对 σ_1^2, σ_2^2 没有什么信息, 当 m, n 都很大时, 由中心极限定理知

$$\frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}} \stackrel{d}{\sim} N(0, 1)$$

由此可给出 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 近似置信区间为

$$\bar{x} - \bar{y} \pm u_{1-\alpha/2} \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

Theorem 4.4.16 — 以上情况都不满足时的一般情况. 令 $s_0^2 = s_x^2/m + s_y^2/n$ 取近似枢轴量

$$T = [\bar{x} - \bar{y} - (\mu_1 - \mu_2)] / s_0$$

此时 T 既不服从 $N(0, 1)$ 也不服从 t 分布. 但近似服从自由度为 l 的 t 分布, 其中 l 由公式

$$l = \frac{s_0^4}{\frac{s_x^4}{m^2(m-1)} + \frac{s_y^4}{n^2(n-1)}}$$

决定, 一般不为整数, 可以取与 l 最接近的整数代替之. 于是, 近似地有 $T \sim t(l)$, 从而可得 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 近似置信区间为

$$\bar{x} - \bar{y} \pm s_0 t_{1-\alpha/2}(l)$$

Theorem 4.4.17 — 大样本置信区间. 在样本量充分大时, 可用渐近分布来构造近似的置信区间, 一个典型的例子是关于比例 p 的置信区间. 设 x_1, \dots, x_n 是来自二点分布 $b(1, p)$ 的样本, 现要求 p 的 $1 - \alpha$ 置信区间. 由中心极限定理知, 样本均值 \bar{x} 的渐近分布为 $N(p, \frac{p(1-p)}{n})$, 因此有

$$u = \frac{\bar{x} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

这个 u 可作为近似枢轴量, 对给定 α , 利用标准正态分布的 $1 - \alpha/2$ 分位数 $u_{1-\alpha/2}$ 可得

$$P\left(\left|\frac{\bar{x} - p}{\sqrt{p(1-p)/n}}\right| \leq u_{1-\alpha/2}\right) \approx 1 - \alpha$$

括号里的事件等价于

$$(\bar{x} - p)^2 \leq u_{1-\alpha/2}^2 p(1-p)/n$$

记 $\lambda = u_{1-\alpha/2}^2$, 上述不等式可化为

$$\left(1 + \frac{\lambda}{n}\right)p^2 - \left(2\bar{x} + \frac{\lambda}{n}\right)p + \bar{x}^2 \leq 0$$

左侧 p 的二次三项式的判别式

$$\left(2\bar{x} + \frac{\lambda}{n}\right)^2 - 4\left(1 + \frac{\lambda}{n}\right)\bar{x}^2 = \frac{4\bar{x}(1-\bar{x})}{n}\lambda + \frac{\lambda^2}{n^2} > 0$$

故此二次三项式的图形是开口向上并与 x 轴有两个交点的曲线. 记此两个交点的横坐标为 p_L 和 p_U , 则有

$$P(p_L \leq p \leq p_U) = 1 - \alpha$$

这里 p_L 和 p_U 是该二次三项式的两个根, 它们可表示为

$$\frac{1}{1 + \frac{\lambda}{n}} \left(\bar{x} + \frac{\lambda}{2n} \pm \sqrt{\frac{\bar{x}(1 - \bar{x})}{n} \lambda + \frac{\lambda^2}{4n^2}} \right)$$

由于 n 比较大, 在实用中通常略去 λ/n 项, 于是可将置信区间近似为

$$\left[\bar{x} - u_{1-\alpha/2} \sqrt{\frac{\bar{x}(1 - \bar{x})}{n}}, \quad \bar{x} + u_{1-\alpha/2} \sqrt{\frac{\bar{x}(1 - \bar{x})}{n}} \right]$$

Theorem 4.4.18 — 样本量的确定. 控制比率 p 的 $1 - \alpha$ 置信区间长度不超过 $2d_0$ 的最小样本量为 $n \geq (u_{1-\alpha/2}/2d_0)^2$

Theorem 4.4.19 — σ_1^2/σ_2^2 的置信区间. 由于 $(m-1)s_x^2/\sigma_1^2 \sim \chi^2(m-1)$, $(n-1)s_y^2/\sigma_2^2 \sim \chi^2(n-1)$, 且 s_x^2 与 s_y^2 相互独立, 故可仿照 F 变量构造如下枢轴量:

$$F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1)$$

对给定的置信水平 $1 - \alpha$, 由

$$P \left(F_{\alpha/2}(m-1, n-1) \leq \frac{s_x^2}{s_y^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\alpha/2}(m-1, n-1) \right) = 1 - \alpha$$

经不等式变形即给出 σ_1^2/σ_2^2 的如下的 $1 - \alpha$ 置信区间

$$\left[\frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{1-\alpha/2}(m-1, n-1)}, \quad \frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{\alpha/2}(m-1, n-1)} \right]$$

从上倒下四种情况的分布

参数情况	μ^2 的置信区间	σ^2 的置信水平	分布
σ^2 已知	$\left[\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$		$N(0, 1)$
σ^2 未知	$\left[\bar{X} - t_{\alpha/2}(n-1) \frac{S_n}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \frac{S_n}{\sqrt{n}} \right]$		$t(n-1)$
μ 已知		$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\alpha/2}^2(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\alpha/2}^2(n)} \right]$	$\chi^2(n)$
μ 未知		$\left[\frac{(n-1)S_n^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S_n^2}{\chi_{1-\alpha/2}^2(n-1)} \right]$	$\chi^2(n-1)$

σ_1^2, σ_2^2	$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$N(0, 1)$	$\hat{X} - \bar{Y} \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
均已知			

讨厌参数	待估参数	置信区间
$\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知	$\mu_2 - \mu_1$	$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(m+n-2) \sqrt{\frac{m+n}{mn(m+n-2)} (Q_1^2 + Q_2^2)} \right]$
$\sigma_2^2 / \sigma_1^2 = \theta$ 已知	$\mu_2 - \mu_1$	$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(m+n-2) \sqrt{\frac{m\theta+n}{mn(m+n-2)} (Q_1^2 + Q_2^2/\theta)} \right]$
$m = n$ 时	$\mu_2 - \mu_1$	$\left[\bar{Z} \pm t_{\alpha/2}(n-1) \frac{\sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}}{\sqrt{n(n-1)}} \right]$
m, n 充分大 一般情况	$\mu_2 - \mu_1$	$\left[\bar{Y} - \bar{X} \pm u_{\alpha/2} S_{mn}^* \right]$
	$\mu_2 - \mu_1$	$\left[\bar{Y} - \bar{X} \pm t_{\alpha/2}(r) S_{mn}^* \right]$
μ_1, μ_2 未知	σ_1^2 / σ_2^2	$\left[\frac{S_{1m}^2 / S_{2n}^2}{F_{\alpha/2}(m-1, n-1)}, \frac{S_{1m}^2 / S_{2n}^2}{F_{1-\alpha/2}(m-1, n-1)} \right]$

其中 $Q_1^2 = \sum_{i=1}^m (X_i - \bar{X})^2, Q_2^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2, Z_i = Y_i - X_i, S_{mn}^* = S_{1m}^2/m + S_{2n}^2/n$

补充：对于 $\mu_2 - \mu_1$ 的估计，档

补充：方差比在均值已知的情况：枢轴量

$$\frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1 \sigma_1^2}{\sum_{i=1}^{n_2} (Y_i - \mu_2)^2 / n_2 \sigma_2^2}$$

服从的分布

$$F(n_1, n_2)$$

置信区间的上下限

$$\frac{1}{f_{\alpha/2}(n_1, n_2)} \cdot \frac{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{n_1 \sum_{i=1}^{n_2} (Y_i - \mu_2)^2}$$

$$\frac{1}{f_{1-\alpha/2}(n_1, n_2)} \cdot \frac{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{n_1 \sum_{i=1}^{n_2} (Y_i - \mu_2)^2}$$

Theorem 4.4.20 二项分布中的 p 的置信区间：当 n 很大的时候，有

$$\frac{\bar{X} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

故

$$\left(\bar{X} - z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \quad \bar{X} + z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right)$$

Theorem 4.4.21 poisson 也可以用正态近似：

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \sim N(0, 1)$$

$$\left(\bar{X} - z_{a/2} \sqrt{\frac{\bar{X}}{n}}, \bar{X} + z_{a/2} \sqrt{\frac{\bar{X}}{n}} \right)$$

R

1. 估计精度: 置信区间的区间长度 $\hat{\theta}_2 - \hat{\theta}_1$ 是区间估计的估计精度, 区间长度越短, 精度越高。
2. 置信度: 置信度的大小反映了这个区间估计的可靠程度, 即随机区间包含待估参数的概率的大小。
3. 唯一一种同时增大置信度和精度的方法: 增加样本量

5. 假设检验

Definition 5.0.1 α 为显著性水平 (significant level), 它越小, 获得显著性结果越难, 即越难拒绝 H_0 , Fisher 称这样的检验为显著性检验。需要在实验前设置好

Definition 5.0.2 — 功效函数, 势函数. 设总体分布包含若干个未知参数 $\theta_1, \dots, \theta_k$. H_0 是关于这些参数的一个原假设, 设有了样本 X_1, \dots, X_n , 而 Φ 是基于这些样本而对 H_0 所做的一个检验. 则称检验 Φ 的功效函数为

$$\beta_\phi(\theta_1, \dots, \theta_k) = P_{\theta_1, \dots, \theta_k} (\text{在检验 } \Phi \text{ 之下}, H_0 \text{ 被否定})$$

它是未知参数 $\theta_1, \dots, \theta_k$ 的函数。

$$\begin{aligned} L_W(\theta) &= P(\text{接受 } H_0 | \theta) \\ &= P((X_1, \dots, X_n) \in W | \theta) \\ \rho_W(\theta) &= P(\text{拒绝 } H_0 | \theta) \\ &= P((X_1, \dots, X_n) \notin W | \theta) \end{aligned}$$

$L_W(\theta)$ 叫做检验法 (否定域) 的操作特性函数 (简称 OC 函数) $\rho_W(\theta)$ 叫做 W 的功效函数. 显然 $L_w(\theta) = 1 - \rho_W(\theta)$ 当 $\theta \in \Theta_0$ 时, $\rho_W(\theta)$ 表示犯第一类错误的概率. 当 $\theta \in \Theta_0$ 时 $1 - \rho_w(\theta)$ 表示犯第二类错误的概率.

Theorem 5.0.1 — 功效函数与犯错之间的联系. 若以 $\theta_1, \dots, \theta_k$ 记总体分布的参数, $\beta_\Phi(\theta_1, \dots, \theta_k)$ 记检验 Φ 的功效函数, 则犯第一、二类错误的概率 $\alpha_{1\Phi}(\theta_1, \dots, \theta_k)$ 和 $\alpha_{2\Phi}(\theta_1, \dots, \theta_k)$ 分

别为

$$\alpha_{1\phi}(\theta_1, \dots, \theta_k) = \begin{cases} \beta_\Phi(\theta_1, \dots, \theta_k), & \text{当 } (\theta_1, \dots, \theta_k) \in H_0 \text{ 时} \\ 0, & \text{当 } (\theta_1, \dots, \theta_k) \in H_1 \text{ 时} \end{cases}$$

$$\alpha_{2\phi}(\theta_1, \dots, \theta_k) = \begin{cases} 0, & \text{当 } (\theta_1, \dots, \theta_k) \in H_0 \text{ 时} \\ 1 - \beta_\Phi(\theta_1, \dots, \theta_k), & \text{当 } (\theta_1, \dots, \theta_k) \in H_1 \text{ 时} \end{cases}$$

这里, H_1 是对立假设。选择一个检验 Φ 时, 要使其功效函数 β_Φ 在 H_0 上尽量小而在 H_1 上尽量大. 先保证第一类错误的概率不超过某指定值 α (α 通常较小, 最常用的是 $\alpha = 0.05$ 和 0.01 , 有时也用到 $0.001, 0.10$, 以至 0.20 等值), 再在这个限制下, 使第二类错误的概率尽可能小。

Definition 5.0.3 称 W 是检验水平为 α 的无偏否定域, 若对一切 $\theta \in \Theta_1$, 有

$$\rho_W(\theta) \geq \alpha$$

无偏性是很自然的要求: “假设”在它真实时遭拒绝的概率不大于它虚假时遭拒绝的概率。

Definition 5.0.4 称 W 是水平为 α 的一致最大功效无偏否定域, 若 W 是水平为 α 的无偏否定域, 而且对任何水平为 α 的无偏否定域 \tilde{W} 恒有

$$\rho_W(\theta) \geq \rho_{\tilde{W}}(\theta) \quad (\text{一切 } \theta \in \Theta_1)$$

以后将会看到, 大量常用的否定域是一致最大功效无偏的 (UMPU).

考虑简单对简单的假设

$$H_0 : \theta = \theta_1 \leftrightarrow H_a : \theta = \theta_2$$

设 X 的样本是 X_1, \dots, X_n , 应怎样确定最好的否定域呢? 记

$$\underline{x} = (x_1, \dots, x_n), \quad d\underline{x} = dx_1 dx_2 \cdots dx_n, \quad L(\underline{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

Theorem 5.0.2 — (Neyman - Pearson 引理). 给定 $\alpha \in (0, 1)$, 设

$$W_0 = \{\underline{x} : L(\underline{x}, \theta_2) > \lambda_0 L(\underline{x}, \theta_1)\}$$

(这里 $\lambda_0 \geq 0$) 适合

$$\int_{W_0} \cdots \int L(\underline{x}, \theta_1) d\underline{x} = \alpha$$

则对任何否定域 $W \subset R^n$, 只要 $\rho_W(\theta_1) \leq \alpha$, 就一定有

$$\rho_{W_0}(\theta_2) \geq \rho_W(\theta_2)$$

换句话说, W_0 是所有检验水平不超过 α 的否定域中犯第二类错误的概率最小的一个。

Proof. 设 W 是任何满足 $\rho_W(\theta_1) \leq \alpha$ 的否定域, 则

$$\begin{aligned}
& \rho_{W_0}(\theta_2) - \rho_W(\theta_2) \\
&= P((X_1, \dots, X_n) \in W_0 \mid \theta_2) - P((X_1, \dots, X_n) \in W \mid \theta_2) \\
&= \int_{W_0} \cdots \int L(\underline{x}, \theta_2) d\underline{x} - \int_W \cdots \int L(\underline{x}, \theta_2) d\underline{x} \\
&= \int_{W_0 - W} \cdots \int L(\underline{x}, \theta_2) d\underline{x} - \int_{W - W_0} \cdots \int L(\underline{x}, \theta_2) d\underline{x} \\
&\geq \lambda_0 \left[\int_{W_0 - W} \cdots \int L(\underline{x}, \theta_1) d\underline{x} - \int_{W - W_0} \cdots \int L(\underline{x}, \theta_1) d\underline{x} \right] \\
&= \lambda_0 \left[\int_{W_0} \cdots \int L(\underline{x}, \theta_1) d\underline{x} - \int_W \cdots \int L(\underline{x}, \theta_1) d\underline{x} \right] \\
&= \lambda_0 [\alpha - \rho_W(\theta_1)] \geq 0
\end{aligned}$$

■

Definition 5.0.5 N - P 引理告诉我们, 似然比检验法具有最优性。N - P 引理断言否定域具有最优性 (犯第二类错误的概率最小)。这个否定域可表示为:

$$W_0 = \{x : \lambda > \lambda_0\}$$

其中 $\lambda = \lambda(x) = L(\underline{x}, \theta_2) / L(\underline{x}, \theta_1)$ 叫做似然比. 这个否定域确定的检验法叫做似然比检验法。

■ **Example 5.1 — N-P 定理的应用.** 设 $X \sim N(\mu, 1), \mu \in \Theta = \{0, 2\}$, 检验问题是

$$H_0 : \mu = 0 \leftrightarrow H_a : \mu = 2$$

给定 $\alpha = 0.05$, 设样本是 X_1, \dots, X_n , 我们来求出最大功效的否定域?

Proof. X 的密度函数是

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

根据 N - P 引理只须找 $\{(x_1, \dots, x_n) : \lambda > \lambda_0\}$ 型的否定域, 其中

$$\lambda = \prod_1^n f(x_i, 2) / \prod_1^n f(x_i, 0)$$

由知 $\lambda = e^{2n\bar{x}-2n}$, 为了 $\lambda > \lambda_0$ 必须且只须 $\bar{X} > \frac{\ln \lambda_0}{2n} + 1$, 记 $C = \frac{\ln \lambda_0}{2n} + 1$, 故应选 C 满足 $P(\bar{X} > C \mid 0) = 0.05$, 这里 $\bar{X} = \frac{1}{n} \sum_1^n X_i \sim N(0, \frac{1}{n})$ (当 $\mu = 0$ 时), 于是

$$\sqrt{n}\bar{X} \sim N(0, 1)$$

所以 C 应满足:

$$\int_{\sqrt{n}C}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = 0.05$$

查表知 $\sqrt{n}C \doteq 1.65$, $C = \frac{1.65}{\sqrt{n}}$, 最大功效的否定域是 $W_0 = \{(x_1, \dots, x_n) : \bar{x} > \frac{1.65}{\sqrt{n}}\}$, 其中 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. 这个否定域是由统计量 \bar{X} 来确定的。 ■

Definition 5.0.6 — 单参数指数分布组的定义.

$$f(x, \theta) = S(\theta)h(x)e^{Q(\theta)V(x)}$$

Theorem 5.0.3 — 单参数指数分布组的 UMP. 设 X 的分布密度有单参数指数分布组型。给定检验问题

$$H_0 : \theta \leq \theta_1 \leftrightarrow H_a : \theta > \theta_1$$

对 $\alpha \in (0, 1)$, 若存在 C 满足

$$P\left(\sum_1^n V(X_i) > C \mid \theta_1\right) = \alpha$$

则

$$W_0 = \left\{ (x_1, \dots, x_n) : \sum_1^n V(x_i) > C \right\}$$

是检验水平为 α 的一致最大功效的否定域。

■ **Example 5.2 — 单参数指数分布组的定义.** 设 $X \sim N(\mu, \sigma_0^2)$ (σ_0 已知) 检验问题是

$$H_0 : \mu \leq \mu_0 \leftrightarrow H_a : \mu > \mu_0$$

其中 μ_0 是已知的. 设样本值是 x_1, \dots, x_n , 如何检验 H_0 ? 此时

$$\begin{aligned} f(x, \theta) &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(x-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{\mu^2}{2\sigma_0^2}\right\} \cdot \exp\left\{-\frac{x^2}{2\sigma_0^2}\right\} \cdot \exp\left\{\frac{\mu x}{\sigma_0^2}\right\} \end{aligned}$$

对比 (3.1) 知 $V(x_i) = x_i$, 依定理 3.1 知一致最大功效的否定域是

$$W_0 = \left\{ (x_1, \dots, x_n) : \sum_{i=1}^n x_i > C \right\}$$

其中 C 满足

$$P\left(\sum_{i=1}^n X_i > C \mid \mu_0\right) = \alpha$$

但 $\mu = \mu_0$ 时

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma_0^2}{n}\right)$$

故

$$\begin{aligned} P\left(\sum_{i=1}^n X_i > C \mid \mu_0\right) &= P\left(\bar{X} > \frac{C}{n} \mid \mu_0\right) \\ &= 1 - \Phi\left(\frac{Cn^{-\frac{1}{2}} - \sqrt{n}\mu_0}{\sigma_0}\right) \end{aligned}$$

若 $\alpha = 0.05$, 查表知 $\Phi(1.65) \doteq 0.95$. 于是

$$Cn^{-\frac{1}{2}} - \sqrt{n}\mu_0 = 1.65\sigma_0$$

故

$$C = n\mu_0 + 1.65\sigma_0\sqrt{n}$$

这时否定域 (3.6) 可以写成:

$$W_0 = \left\{ (x_1, \dots, x_n) : \bar{x} > \bar{\mu}_0 + \frac{1.65}{\sqrt{n}}\sigma_0 \right\}$$

Theorem 5.0.4 设 X 有分布密度

$$f(x, \theta) = S(\theta)h(x)e^{\alpha(\theta)V(x)}$$

这里 $S(\theta) > 0, h(x) \geq 0, Q(\theta)$ 是 θ 的严格增函数. 给定检验问题:

$$H_0 : \theta \in (\theta_1, \theta_2) \leftrightarrow H_a : \theta \in (\theta_1, \theta_2)$$

设 $X = (X_1, \dots, X_n)$ 是 X 的样本, 且

$$W_0 = \left\{ (x_1, \dots, x_n) : C_1 < \sum_{i=1}^n V(x_i) < C_2 \right\}$$

若 $P(X \in W_0 | \theta_i) = \alpha (i = 1, 2)$, 则 W_0 是水平为 α 的一致最大功效的否定域 (UMP 检验)。

Theorem 5.0.5 设 X 有分布密度

$$f(x, \theta) = S(\theta)h(x)e^{Q(\theta)V(x)}$$

$\theta \in (a, b) (-\infty \leq a < b \leq \infty)$, 其中 $S(\theta) > 0, h(x) \geq 0, Q(\theta)$ 是 θ 的严格增连续函数。给定检验问题:

$$H_0 : \theta \in [\theta_1, \theta_2] \leftrightarrow H_a : \theta \in [\theta_1, \theta_2] \\ (a < \theta_1 < \theta_2 < b)$$

设 $X = (X_1, \dots, X_n)$ 是 X 的样本, 且

$$W_0 = \left\{ (x_1, \dots, x_n) : \sum_{i=1}^n V(x_i) \text{ 小于 } C_1 \text{ 或大于 } C_2 \right\}$$

若 $C_1 < C_2$ 使得

$$P(X \in W_0 | \theta_i) = \alpha \quad (i = 1, 2)$$

则 W_0 是水平为 α 的一致最大功效的无偏否定域 (UMPU 检验)。

Definition 5.0.7 设 Φ 是原假设 H_0 的一个检验, $\beta_\phi(\theta_1, \dots, \theta_k)$ 为其功效函数, α 为常数 ($0 \leq \alpha \leq 1$) . 如果

$$\beta_\phi(\theta_1, \dots, \theta_k) \leq \alpha \quad (\text{对任何 } (\theta_1, \dots, \theta_k) \in H_0)$$

则称 Φ 为 H_0 的一个水平为 α 的检验, 或者说检验 Φ 的水平为 α , 检验 Φ 有水平 α

Definition 5.0.8 — 一致最优检验. 设 Φ 为一个水平 α 的检验, 若对任何其他一个水平 α 的检验 g , 必有

$$\beta_\phi(\theta_1, \dots, \theta_k) \geq \beta_g(\theta_1, \dots, \theta_k) \quad (\text{对任何 } (\theta_1, \dots, \theta_k) \in H_1)$$

这里 H_1 为对立假设, 则称 Φ 是假设检验问题 $H_0 : H_1$ 的一个水平 α 的一致最优检验。简单地说, 水平 α 的一致最优检验, 就是在一切水平 α 的检验中, 其功效函数在对立假设 H_1 上处处达到最大者. 或者说, 是在一切其第一类错误概率不超过 α 的检验中, 第二类错误概率处处达到最小者. 一致表示的是对于每个 θ 来说都大

Definition 5.0.9 — 显著性. 从统计学的观点看, 达到显著性无非是指: 在给定水平上, 差异 $\bar{Y} - \bar{X}$ 已不能仅由随机性来解释, 而也有 $\theta_2 > \theta_1$ 的原因. 所以, 你可以简单地把“显著性检验”理解为“希望原假设被否定的那种检验”. 显著性检验的特点不在于检验自身, 而在于其在使用中的含义如何。

Theorem 5.0.6 — 截尾寿命检验. 1. 定数截尾法: 取 n 个元件做试验, 定下一个自然数 $r < n$, 试验进行到有 r 个元件失效时止. 把到此时为止, 全部 n 个元件的工作时间加起来记为 T , 即为

$$T = Y_1 + \dots + Y_r + (n - r)Y_r$$

Y_1 是最先失效的那个元件的失效时刻 (从时刻 0 开始算起), Y_2 为第二个失效的元件的失效时刻, 以此类推, 第 r 个失效元件在时刻 Y , 试验也就到此为止, 余下尚有 $n - r$ 个未失效元件, 它们已工作的总时间为 $(n - r)Y_r$. 这样得到 T 的表达式。不难理解: T 愈大, 就愈使我们相信元件的平均寿命大. 因此, 比如说, 一个合理检验为: ψ : 当 $T \geq C$ 时接受原假设 H'_0 , 不然就否定 H'_0 可以证明: 当参数值为 λ 时, $2\lambda T \sim \chi^2_{2r}$. 由此出发, 仿照前面的推理, 就不难在给定检验水平 α 之下定出式中的 C 为

$$C = \chi^2_{2r}(1 - \alpha) / (2\lambda_0)$$

2. 定时截尾法指定一个时刻 T_0 , 拿 n 个元件做试验, 直到时刻 T_0 为止. 把到这时为止全部 n 个元件的工作总时间加起来记为 T^* , 算法是: 若某个元件在 T_0 之前的某个时刻 t 已失效, 则该元件的工作时间为 t ; 若到 T_0 时刻仍未失效, 则该元件的工作时间为 T_0 . ** 显然, 平均寿命愈大, 则 T^* 愈倾向于取较大的值, 于是得出检验 ψ' : 当 $T^* \geq C$ 时接受 H'_0 , 不然就否定 H'_0 可以证明: 近似地有 $2\lambda T^* \sim \chi^2_{2u+1}$. 这里, u 是到时刻 T_0 停试时已失效的元件个数. 由此出发, 仿照前面的推理, 即可定出在给定水平 α 时 C 的近似值 (因 $2\lambda T^* \sim \chi^2_{2u+1}$ 只是近似成立), 为 $C = \chi^2_{2u+1}(1 - \alpha) / (2\lambda_0)$

Theorem 5.0.7 — 科大概统--假设检验. 1. 统计方法的大小样本之分, 不在于样本大小 n 多大 (这无清楚界线), 而全看其是否使用有关变量的极限分布。

2. 拟合优度检验/皮尔逊卡方检验:

$$\begin{aligned} Z &= \sum (\text{理论值} - \text{经验值})^2 / \text{理论值} \\ &= \sum_{i=1}^k (np_i - v_i)^2 / (np_i) \end{aligned}$$

H_0 成立, 则在样本大小 $n \rightarrow \infty$ 时, Z 的分布趋向于自由度为 $k-1$ 的 χ^2 分布, 即 χ_{k-1}^2

3. 列联表是一种按照两属性做双向分类的表, 原假设是检验两属性独立

4. 如果总体分布为离散形式, 但是可以取的点是无数个时, 对区间进行划分

Theorem 5.0.8 — 简单假设下奈曼--皮尔逊引理的解释. 原假设 H_0 和对立假设 H_1 中都只包含一个分布.

H_0 : 总体有密度 $f_0(x)$

H_1 : 总体有密度 $f_1(x)$

设 X_1, \dots, X_n 为样本, 则 (X_1, \dots, X_n) 的密度, 在 H_0 和 H_1 之下, 分别为 $g_0(y) = f_0(x_1) \cdots f_0(x_n)$ 和 $g_1(y) = f_1(x_1) \cdots f_1(x_n)$. 这里已简记 $y = (x_1, \dots, x_n)$. 求这个问题的水平 α 的检验, 转化为下述数学问题: 找 y 空间中的一个区域 Q , 作为检验的否定域 (当 (X_1, \dots, X_n) 落在 Q 内时否定 H_0 , 不然就接受 H_0). 为使 Q 达到最优, 就必须在条件

$$\int_Q g_0(y) dy \leq \alpha$$

之下使 $\int_Q g_1(y) dy$ 达到最大. 很容易看出: 为达到这一点, Q 必须这样取: 把值 $g_1(y)/g_0(y)$ 大的那些 y 收进来.

Theorem 5.0.9 — 奈—皮基本引理. 水平 α 的一致最优检验 φ 的否定域 Q 应如下取: 找常数 C , 使

$$Q = \{y \mid g_1(y)/g_0(y) > C\} \quad (5.1)$$

而满足

$$\int_Q g_0(y) dy = \alpha \quad (5.2)$$

Proof. 5.2 式保证了检验 φ 的水平为 α , 现设 φ' 为另一水平 α' 的检验, 其否定域为 Q' . 记 Q 与 Q' 的公共部分为 R . Q_1 记 Q 中去掉 R 的剩余部分, Q'_1 记 Q' 中去掉 R 的剩余部分, 则易见

$$\int_Q g_1(y) dy - \int_{Q'} g_1(y) dy = \int_{Q_1} g_1(y) dy - \int_{Q'_1} g_1(y) dy \quad (5.3)$$

由于 φ' 有水平 α' , 有

$$\int_{Q'} g_0(y) dy \leq \alpha'$$

再由 (5.2) 式, 知

$$\int_{Q_1} g_0(y) dy \geq \int_{Q'_1} g_0(y) dy \quad (5.4)$$

因为 Q'_1 在 Q 之外, 按 5.1 式, 当 y 属于 Q'_1 时, 有 $g_1(y) \leq Cg_0(y)$. 而当 y 属于 Q_1 时有 $g_1(y) > Cg_0(y)$. 故

$$\begin{aligned} \int_{Q_1} g_1(y) dy &\geq C \int_{Q_1} g_0(y) dy \\ \int_{Q'_1} g_1(y) dy &\leq C \int_{Q'_1} g_0(y) dy \end{aligned}$$

由此及 5.3 式, 5.14 式, 即知

$$\int_Q g_1(y) dy \geq \int_{Q'} g_1(y) dy$$

即检验 φ 的功效总不小于 φ' 的功效, 由于 φ' 是任取的水平 α 的检验, 故证明了 φ 是水平 α 的一致最优检验。 ■

Theorem 5.0.10 在 H_0 中取定一值 θ_0 , 对 H_1 中的值 θ_1 建立假设检验问题

$$H'_0 : \theta_0; \quad H'_1 : \theta_1$$

按奈一皮基本引理, 求出其水平 α 的一致最优检验 φ , 如果 φ 符合以下两个条件, 则它必须是原问题 $H_0 : H_1$ 的一个水平 α 的一致最优检验

1. 检验 φ 也是 $H_0 : H_1$ 的水平 α 的检验
2. 检验 φ 不依赖于 θ_1 值.

Proof. 设 φ' 为 $H_0 : H_1$ 的任一水平 α 的检验, 则它必是 (5) 式的一个水平 α 的检验. 这很显然: 以 $\beta_{\varphi'}(\theta)$ 记 φ' 的功效函数. φ' 为 $H_0 : H_1$ 的水平 α 检验, 意味着 $\beta_{\varphi'}(\theta)$ 在 H_0 上处处不超过 α , 因而特别在 θ_0 点不超过 α . 这样, φ 和 φ' 都是 (5) 式的水平 α 的检验, 而 φ 是 (5) 式的水平 α 的一致最优检验, 故 $\beta_\varphi(\theta_1) \geq \beta_{\varphi'}(\theta_1)$. 因为这个事实对 H_1 中任一个 θ_1 都成立, 即知 φ 为 $H_0 : H_1$ 的水平 α 的一致最优检验。 ■

Definition 5.0.10 [非中心 t 分布] 设 X 与 Y 独立, $X \sim N(0, 1)$, $Y \sim \chi_n^2$. 又设 δ 为常数, 则随机变量 $Z = (X + \delta)/\sqrt{\frac{1}{n}Y}$ 的分布称为自由度 n 、非中心参数 δ 的非中心 t 分布, 记为 $Z \sim t_{n,\delta}$. $t_{n,\delta}$ 的分布函数将记为 $F_{n,\delta}(x)$.

Proposition 5.0.11 若 $\delta_2 > \delta_1$, 则 $F_{n,\delta_2}(x) \leq F_{n,\delta_1}(x)$. 事实上, 记

$$Z_i = (X + \delta_i)/\sqrt{\frac{1}{n}Y} \quad (i = 1, 2)$$

X, Y 如上文所述, 则有 $Z_1 < Z_2$, 故对任何 x 有 $P(Z_1 \leq x) \geq P(Z_2 \leq x)$, 即

$$F_{n,\delta_1}(x) \geq F_{n,\delta_2}(x)$$

Definition 5.0.11 设有样本 X , 取值于样本空间 \mathcal{X} , 且知道样本来自某一个参数分布族 $\{F(x, \theta) : \theta \in \Theta\}$ 其中 Θ 为参数空间。设 $\Theta_0 \subset \Theta$, 且 $\Theta_0 \neq \emptyset$, 则命题 $H_0 : \theta \in \Theta_0$ 称为一个假设或零假设 (null-hypothesis). 如记 $\Theta_1 = \Theta - \Theta_0$, 则命题 $H_1 : \theta \in \Theta_1$ 称为 H_0 的对立假设或备选假设 (alternative hypothesis). 于是, 我们感兴趣的假设就是

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1$$

我们称此假设的检验问题为假设检验问题。

Definition 5.0.12 1. 假设: 参数空间 $\Theta = \{\theta\}$ 的非空子集或有关参数 θ 的命题, 称为统计假设, 简称假设

2. 原假设, 根据需要而设立的假设, 常记为 $H_0 : \theta \in \Theta_0$

3. 备择假设, 在原假设被拒绝后而采用 (接受) 的假设, 常记为 $H_1 : \theta \in \Theta_1$

要求 $\Theta_0 \cap \Theta_1 = \emptyset$, 即原假设 H_0 与备择假设 H_1 不含公共参数. 如果假设 H_0 (或 H_1) 只含一个点, 则称该假设是简单假设, 否则称为复杂假设

检验: 对原假设 $H_0 : \theta \in \Theta_0$ 作出是否拒绝 H_0 的判断的法则称为检验法则, 简称检验. 检验有两个结果:

1. “原假设不正确”, 称为拒绝原假设, 或称检验显著;

2. “原假设正确”, 称为接受 (保留) 原假设, 或称检验不显著

统计假设检验的着力点不在于说明原假设正确, 而在于说明原假设不正确, 因为用一个样本去说明原假设正确是根据不足的, 而用一个样本推翻原假设是合理的, 因此统计学在讨论假设检验时着力点在于建立检验的拒绝域.

Theorem 5.0.12 — 检验问题. 由原假设 H_0 和备择假设 H_1 组成一个需要作判断的问题称为检验问题

1. 参数假设检验问题, 两个假设都是有关参数的命题组成的检验问题;

2. 非参数假设检验问题, 两个假设都是有关分布的命题组成的检验问题

常用的参数假设检验问题有如下三个, 其中 θ_0 是已知常数

1. $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$

2. $H_0 : \theta \geq \theta_0$ vs $H_1 : \theta < \theta_0$

3. $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

其中 (1) 与 (2) 又称单侧检验问题, 因为一个假设位于另一个假设的一侧, (3) 称为双侧检验问题, 因为备择假设位于原假设的两侧

Definition 5.0.13 — 两类错误及其发生概率. 1. 原假设 H_0 正确, 但被拒绝, 这种判断错误称为犯第一类错误, 其发生概率称为犯第一类错误的概率, 或称拒真概率, 常记为 α

2. 原假设 H_0 不真, 但被接受, 这种判断错误称为犯第二类错误, 其发生概率称为犯第二类错误的概率, 或称受伪概率, 常记为 β

决策 H_0 为真 H_1 为真

接受 H_0 正确 Type II

拒绝 H_0 Type I 正确

为了方便记忆, 我们分别称第一、二类错误为拒真

与纳伪。

第一类错误概率: $\alpha = P_{\theta} \{ \mathbf{X} \in W \}, \theta \in \Theta_0$, 也记为 $P \{ \mathbf{X} \in W | H_0 \}$ 第二类错误概率: $\beta = P_{\theta} \{ \mathbf{X} \in \bar{W} \}, \theta \in \Theta_1$, 也记为 $P \{ \mathbf{X} \in \bar{W} | H_1 \}$

$$\alpha = P \{ \text{拒绝 } H_0 | H_0 \text{ 为真} \}$$

$$\beta = P\{\text{接受} H_0 | H_1 \text{ 为真}\}$$

Theorem 5.0.13 — 假设检验的基本步骤.

1. 建立假设. 根据要求建立原假设 H_0 和备择假设 H_1
2. 选择检验统计量, 给出拒绝域 W 的形式。用于对原假设 H_0 作出判断的统计量称为检验统计量。使原假设被拒绝的样本观察值所在区域称为拒绝域, 常用 W 表示。一个拒绝域 W 唯一确定一个检验法则, 反之, 一个检验法则唯一确定一个拒绝域 W .
3. 选择显著性水平 $\alpha(0 < \alpha < 1)$
4. 给出拒绝域: 由概率等式 $P(W) = \alpha$ 确定具体的拒绝域.
5. 作出判断
 - (a) 当样本 $(x_1, \dots, x_n) \in W$, 则拒绝 H_0 , 即接受 H_1
 - (b) 当样本 $(x_1, \dots, x_n) \in \bar{W}$, 则接受 H_0

Definition 5.0.14 — 势函数或功效函数 (power function). 设检验问题 $H_0 : \theta \in \Theta_0 \text{ vs } H_1 : \theta \in \Theta_1$ 的拒绝域为 W , 则样本观测值 x_1, \dots, x_n 落在拒绝域 W 内的概率称为该检验的势函数, 记为

$$g(\theta) = P_\theta((x_1, \dots, x_n) \in W), \quad \theta \in \Theta_0 \cup \Theta_1$$

由势函数 $g(\theta)$ 容易得到犯两类错误的概率

$$g(\theta) = \begin{cases} \alpha(\theta), & \theta \in \Theta_0 \\ 1 - \beta(\theta), & \theta \in \Theta_1 \end{cases}$$

Definition 5.0.15 — (势函数). 对于假设的一个检验方法 ψ , 其拒绝域记为 W , 则我们称

$$\beta_\psi(\theta) = P_\theta\{\mathbf{X} \in W\}, \forall \theta \in \Theta$$

为此检验的势函数. 从上一定义可以看出, 当 $\theta \in \Theta_0$ 时, 此检验犯第一类错误的概率等于其势函数 $\beta_\psi(\theta)$; 而当 $\theta \in \Theta_1$ 时, 检验犯第二类错误的概率等于 $1 - \beta_\psi(\theta)$

(R)

1. 对于固定的样本容量, 我们找不到一个检验方法, 使得其第一二类错误概率均达到最小;
2. 第二类错误概率不易求出, 由于它依赖于备选假设中的参数。

Definition 5.0.16 — (显著性水平). 对于检验 ψ 和事先给定的 $\alpha \in (0, 1)$, 如果它满足

$$P_\theta\{\mathbf{X} \in W\} \leq \alpha, \forall \theta \in \Theta_0$$

则称 α 是检验 ψ 的水平或显著性水平, 也称 ψ 为显著性水平 α 的检验.

Definition 5.0.17 — 显著性检验. 水平为 α 的检验 对检验问题 $H_0 : \theta \in \Theta_0 \text{ vs } H_1 : \theta \in \Theta_1$ 中, 如果一个检验犯第一类错误的概率 $\alpha(\theta)$ 不超过事先给定的显著性水平 α , 即

$$\alpha(\theta) \leq \alpha, \theta \in \Theta_0$$

则称该检验为水平为 α 的检验, 或称显著性检验。在实际使用中 α 不宜选得过小, α 过小会导致 β 过大, 应在适当控制 α 中制约 β . 最常用的选择是 $\alpha = 0.05$, 有时也选用 $\alpha = 0.10$ 或 $\alpha = 0.01$

Definition 5.0.18 P 值的定义：在一个假设检验问题中，利用观测值能够做出拒绝原假设的最小显著性水平称为 P -值。对于 P -值的理解，一般情况下有几种认识：

1. 拒绝原假设的最小显著性水平。
2. 观察到的（实例样本的）显著性水平。
3. 表示对原假设的支持程度，是用于确定是否应该拒绝原假设的另一种方法。
4. 一种概率，一种在原假设为真的前提下出现观察样本以及更极端情况的概率。
5. 我们在拒绝原假设的犯的第一类错误，而所规定的显著性水平（具有主观性）是事先给定的犯第一类错误的最大错误。

Theorem 5.0.14 — 对 p 值的理解. α 是犯第一类错误的概率，也就当 H_0 为真时，要拒绝原假设的概率，简称为弃真。犯第一类错误的概率也就是小概率事件发生的概率。 α 的意思是当样本表明我要拒绝原假设的时候，我是拒绝还是不拒绝。这要根据小事件发生的可能性，如果小事件发生的概率是 0.05，但是我这一次抽样得到的样本却不支持原假设，那么我就认为我的原假设是错的，我要拒绝原假设。 p -value 是利用样本计算得到拒绝原假设的最小显著性水平。也就是我犯第一类错误的可能性。 p -value 的意思是我小概率事件发生的概率是 0.02，这么小的概率但是这次却发生了，我几乎不会犯错，所以我很有信心地拒绝原假设。

Theorem 5.0.15 检验的 p 值在一个假设检验问题中，利用样本观察值能够做出拒绝原假设的最小显著性水平称为该检验的 p 值，引入检验的 p 值的好处是：

1. 它比较客观，避免了事先确定显著性水平；
2. 由检验的 p 值与人们心目中的显著性水平 α 进行比较
 - (a) 如果 $\alpha \geq p$, 则在显著性水平 α 下拒绝 H_0
 - (b) 如果 $\alpha < p$, 则在显著性水平 α 下应保留 H_0
3. 检验的 p 值的计算是复杂的，会涉及各种抽样分布。如今统计软件都有计算 p 值的功能，因此这对使用者反而是方便，它不需要备用各种抽样分布的分位数表，而只需要观察计算机的输出的 p 值多少就可以做出判断。

5.0.1 正态总体参数假设检验

单总体

设 x_1, \dots, x_n 是来自 $N(\mu, \sigma^2)$ 的样本，考虑如下三种关于 μ 的检验问题

$$\begin{aligned} \text{I } H_0 : \mu \leq \mu_0 &\quad \text{vs} \quad H_1 : \mu > \mu_0 \\ \text{I } H_0 : \mu \geq \mu_0 &\quad \text{vs} \quad H_1 : \mu < \mu_0 \\ \text{III } H_0 : \mu = \mu_0 &\quad \text{vs} \quad H_1 : \mu \neq \mu_0 \end{aligned}$$

其中 μ_0 是已知常数。由于正态总体含两个参数，总体方差 σ^2 已知与否对检验有

Theorem 5.0.16 — σ 已知时的 μ 检验. 对于

$$\text{I } H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

所示的单侧检验问题 I，由于 μ 的点估计是 \bar{x} ，且 $\bar{x} \sim N(\mu, \sigma^2/n)$ ，故选用检验统计量

$$u = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

是恰当的。直觉告诉我们：当样本均值 \bar{x} 不超过设定均值 μ_0 时，应倾向于接受原假设；当样本均值 \bar{x} 超过 μ_0 时，应倾向于拒绝原假设。可是，在有随机性存在的场合，如果 \bar{x} 比

μ_0 太一点就拒绝原假设似乎不当, 只有当 \bar{x} 比 μ_0 大到一定程度时拒绝原假设才是恰当的, 这就存在一个临界值 c , 拒绝域为

$$W_1 = \{(x_1, \dots, x_n) : u \geq c\}$$

常简记为 $\{u \geq c\}$, 若要求检验的显著性水平为 α , 则 c 满足

$$P_{\mu_0}(u \geq c) = \alpha$$

由于在 $\mu = \mu_0$ 时 $u \sim N(0, 1)$, 故 $c = u_{1-\alpha}$, 最后的拒绝域为

$$W_1 = \{u \geq u_{1-\alpha}\}$$

该检验用的检验统计量是 u 统计量, 故一般称为 u 检验. 该检验的势函数是 μ 的函数, 它可用正态分布写出, 具体如下: 对 $\mu \in (-\infty, \infty)$

$$\begin{aligned} g(\mu) &= P_{\mu}(X \in W_1) = P_{\mu}(u \geq u_{1-\alpha}) \\ &= P_{\mu}\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq u_{1-\alpha}\right) \\ &= P_{\mu}\left(\frac{\bar{x} - \mu + \mu - \mu_0}{\sigma/\sqrt{n}} \geq u_{1-\alpha}\right) \\ &= P_{\mu}\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + u_{1-\alpha}\right) \\ &= 1 - \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} + u_{1-\alpha}\right) \end{aligned}$$

由此可见, 势函数是 μ 的增函数. 由增函数性质知, 只要 $g(\mu_0) = \alpha$ 就可保证在 $\mu \leq \mu_0$ 时有 $g(\mu) \leq \alpha$. 所以上述求出的检验是显著性水平为 α 的检验.

用 p 值进行检验的方法

对给定的样本观测值, 可以计算出相应的检验统计量 u 的值, 记为 $u_0 = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$, 这里的 \bar{x} 是样本观测值. 因 u 是服从标准正态分布的随机变量, 令

$$p_1 = P(u \geq u_0) = 1 - \Phi(u_0) \quad (5.5)$$

此即说明 $u_0 = u_{1-p}$, 于是由正态分布函数的反函数的单调性有如下结论:

1. 当 $p > \alpha$ 时, $u_{1-\alpha} > u_0$, 于是观测值不在拒绝域里, 应接受原假设.
2. 当 $p \leq \alpha$ 时, $u_{1-\alpha} \leq u_0$, 于是观测值落在拒绝域里, 应拒绝原假设, 5.5计算出的值就是该检验的 p 值.

Theorem 5.0.17 — σ 已知时的 μ 检验. 对双侧检验问题 III, 也可类似进行讨论, 只不过检验的 p 值稍有不同. 仍选用 u 作为检验统计量, 考虑到备择假设 H_1 分散在两侧, 故其拒绝域亦应在两侧, 即拒绝域应有如下形式

$$W_{III} = \{|u| \geq c\}$$

对给定的显著性水平 $\alpha(0 < \alpha < 1)$, 由 $P_{\mu_0}(|u| \geq c) = \alpha$ 可定出 $c = u_{1-\alpha/2}$, 最后的拒绝域为

$$W_{III} = \{|u| \geq u_{1-\alpha/2}|$$

下面介绍双侧检验的 p 值的计算. 在检验统计量分布对称场合, 双侧检验的 p 值的计算与单侧检验是类似的, 不对称场合我们在后面介绍。仿上, 令

$$p_{III} = P(|u| \geq |u_0|) = 2(1 - \Phi(|u_0|)) \quad (5.6)$$

此即说明 $|u_0| = u_{1-\alpha/2}$, 这里要用到 u_0 的绝对值是因为对双侧假设检验, 观察值可能为正, 协可能为负, 二者机会相同, 于是有类似的结论:

1. 当 $p > \alpha$ 时, $u_{1-\alpha/2} > |u_0|$, 于是观测值不在拒绝域里, 应接受原假设.
2. 当 $p \leq \alpha$ 时, $u_{1-\alpha/2} \leq |u_0|$, 于是观测值落在拒绝域里, 应拒绝原假设。由此可以看出, 5.6计算出的值就是该检验的 p 值。

Theorem 5.0.18 — σ 未知时的 t 检验. 对检验问题 I, 由于 σ 未知, 将未知的 σ 替换成样本标准差 s , 这就形成 t 检验统计量

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

由抽样分布基本定理知, 在 $\mu = \mu_0$ 时 $t \sim t(n-1)$, 从而检验问题 I 的拒绝域为

$$W_I = \{t \geq t_{1-\alpha}(n-1)\}$$

检验的 p 值是类似的, 对给定的样本观测值, 可以计算出相应的检验统计量的值, 记为 $t_0 = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$, 这里的 \bar{x}, s 是样本观测值, 记 t 是服从自由度是 $n-1$ 的 t 分布的随机变量, 则

$$p_I = P(t \geq t_0)$$

对另两组检验问题讨论是完全类似于上一小节的, 罗列结果如下: 检验问题 II 的拒绝域为

$$W_{II} = \{t \leq t_\alpha(n-1)\}$$

p 值为

$$p_{II} = P(t \leq t_0)$$

检验问题 III 的拒绝域为

$$W_{III} = \{|t| \geq t_{1-\alpha/2}(n-1)\}$$

p 值为

$$p_{III} = P(|t| \geq |t_0|)$$

同样可证明这三个检验都是显著性水平为 α 的检验。

Table 5.1: 单样本正态总体均值的显著性检验

方差	假设	检验统计量	拒绝域	名字
$\sigma^2 = \sigma_0^2$	$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0$	$U = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0}$	$\{ U > u_{\alpha/2}\}$	双边 u 检验
	$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu < \mu_0$		$\{U < -u_{\alpha}\}$	单边 u 检验
	$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0$		$\{U > u_{\alpha}\}$	单边 u 检验
	$H_0 : \mu \leq \mu_0 \longleftrightarrow H_1 : \mu > \mu_0$		$\{U > u_{\alpha}\}$	单边 u 检验
	$H_0 : \mu \geq \mu_0 \longleftrightarrow H_1 : \mu < \mu_0$		$\{U < -u_{\alpha}\}$	单边 u 检验
σ^2 未知	$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0$	$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S_n}$	$\{ T > t_{\alpha/2}(n-1)\}$	双边 t 检验
	$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu < \mu_0$		$\{T < -t_{\alpha}(n-1)\}$	单边检验
	$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0$		$\{T > t_{\alpha}(n-1)\}$	单边 t 检验
	$H_0 : \mu \leq \mu_0 \longleftrightarrow H_1 : \mu > \mu_0$		$\{T > t_{\alpha}(n-1)\}$	单边 t 检验
	$H_0 : \mu \geq \mu_0 \longleftrightarrow H_1 : \mu < \mu_0$		$\{T < -t_{\alpha}(n-1)\}$	单边 t 检验

检验法	H_0	H_1	检验统计量	拒绝域	p 值
u 检验 (σ 已知)	$\mu \leq \mu_0$	$\mu > \mu_0$	$u = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$\{u \geq u_{1-\alpha}\}$	$1 - \Phi(u_0)$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$\{u \leq u_{\alpha}\}$	$\Phi(u_0)$
	$\mu = \mu_0$	$\mu \neq \mu_0$		$\{ u \geq u_{1-\alpha/2}\}$	$2(1 - \Phi(u_0))$
t 检验 (σ 未知)	$\mu \leq \mu_0$	$\mu > \mu_0$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$\{t \geq t_{1-\alpha}(n-1)\}$	$P(t \geq t_0)$
	$\mu \geq \mu_0$	$\mu < \mu_0$		$\{t \leq t_{\alpha}(n-1)\}$	$P(t \leq t_0)$
	$\mu = \mu_0$	$\mu \neq \mu_0$		$\{ t \geq t_{1-\alpha/2}(n-1)\}$	$P(t \geq t_0)$

$u_0 = \sqrt{n}(\bar{x} - \mu_0)/\sigma, t_0 = \sqrt{n}(\bar{x} - \mu_0)/s, t$ 是服从 $t(n-1)$ 的随机变量.

Theorem 5.0.19 — 假设检验和置信区间之间的关系. 首先考虑双侧检验问题 III, 显著性水平为 α 的检验的接受域为

$$\overline{W}_{\text{III}} = \left\{ |\bar{x} - \mu_0| \leq \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right\}$$

它可以改写为

$$\overline{W}_{\text{m}} = \left\{ \bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \leq \mu_0 \leq \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right\}$$

这里 μ_0 并无限制, 若让 μ_0 在 $(-\infty, \infty)$ 内取值, 就可得到 μ 的 $1 - \alpha$ 置信区间 $\bar{x} \pm \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$. 反之, 若有一个如上的 $1 - \alpha$ 置信区间, 也可获得关于 $H_0 : \mu = \mu_0$ 的显著性水平为 α 的显著性检验. 所以, 正态均值 μ 的 $1 - \alpha$ 置信区间”与“关于 $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ 的双侧检验问题的显著性水平为 α 的检验”是一一对应的。

两个正态总体均值差的检验

设 x_1, \dots, x_m 是来自正态总体 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, \dots, y_n 是来自另一个正态总体 $N(\mu_2, \sigma_2^2)$ 的样本, 两个样本相互独立. 考虑如下三类检验问题

- I $H_0 : \mu_1 - \mu_2 \leq 0$ vs $H_1 : \mu_1 - \mu_2 > 0$
- II $H_0 : \mu_1 - \mu_2 \geq 0$ vs $H_1 : \mu_1 - \mu_2 < 0$
- III $H_0 : \mu_1 - \mu_2 = 0$ vs $H_1 : \mu_1 - \mu_2 \neq 0$

这里对常用的两种情形进行讨论。

Theorem 5.0.20 — σ_1, σ_2 已知时的两样本 u 检验. 此时 $\mu_1 - \mu_2$ 的点估计 $\bar{x} - \bar{y}$ 的分布完全已知

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

由此可采用 u 检验方法, 检验统计量为

$$u = (\bar{x} - \bar{y}) / \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}}$$

在 $\mu_1 = \mu_2$ 时, $u \sim N(0, 1)$. 检验的拒绝域取决于备择假设的具体内容. 检验问题 I , 检验的拒绝域与 p 值分别为

$$W_I = \{u \geq u_{1-\alpha}\}, \quad p_I = 1 - \Phi(u_0)$$

其中 $u_0 = (\bar{x} - \bar{y}) / \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}}$ 是由样本计算得到的检验统计量的值. 对检验问题 II , 检验的拒绝域与 p 值分别为

$$W_{II} = \{u \leq u_\alpha\}, \quad p_{II} = \Phi(u_0)$$

对检验问题 III, 检验的拒绝域与 p 值分别为

$$W_{III} = \{|u| \geq u_{1-\alpha/2}\}, \quad p_{III} = 2(1 - \Phi(|u_0|))$$

Theorem 5.0.21 — $\sigma_1 = \sigma_2 = \sigma$ 但未知时的两样本 t 检验. 在 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 但未知时, 首先

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right) \sigma^2\right)$$

其次, 由于

$$\frac{1}{\sigma^2} \sum_{i=1}^m (x_i - \bar{x})^2 \sim \chi^2(m-1), \quad \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 \sim \chi^2(n-1)$$

故

$$\frac{1}{\sigma^2} \left(\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2 \right) \sim \chi^2(m+n-2)$$

记

$$s_w^2 = \frac{1}{m+n-2} \left[\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 \right]$$

于是有

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_w \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

这就给出了检验统计量为

$$t = \frac{(\bar{x} - \bar{y})}{s_w \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

对检验问题 I, 检验的拒绝域与 p 值分别为

$$W_I = \{t \geq t_{1-\alpha}(m+n-2)\}, \quad p_I = P(t \geq t_0)$$

其中

$$t_0 = \frac{(\bar{x} - \bar{y})}{s_w \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

是由样本计算得到的检验统计量的值, t 是服从自由度是 $n+m-2$ 的 t 分布的统计量.

对检验问题 II, 检验的拒绝域与 p 值分别为

$$W_{II} = \{t \leq t_\alpha(m+n-2)\}, \quad p_{II} = P(t \leq t_0)$$

对检验问题 III, 检验的拒绝域与 p 值分别为

$$W_{III} = \{|t| \geq t_{1-\alpha/2}(m+n-2)\}, \quad p_{III} = P(|t| \geq |t_0|)$$

Theorem 5.0.22 — 成对数据检验. 在对两个总体均值进行比较时, 若数据是成对出现的, 则应采用成对数据检验. 成对数据检验就是把两个总体均值的比较通过成对数据的差转变为对单个总体均值的检验, 方法完全等同. 成对数据的获得事先要作周全的安排(即试验设计). 在获得成对数据时不要发生“错位”, 从而失去“成对数据”的信息。

在正态性假定下, $d = x - y \sim N(\mu, \sigma_d^2)$, 其中 $\mu = \mu_1 - \mu_2$, $\sigma_d^2 = \sigma_1^2 + \sigma_2^2$. 原先要比较 μ_1 与 μ_2 的大小, 如今则转化为考察 μ 是否为零, 即考察如下检验问题:

$$H_0: \mu = 0 \quad \text{vs} \quad H_1: \mu \neq 0$$

即把双样本的检验问题转化为单样本 t 检验问题. 这时检验的 t 统计量为

$$t_2 = \bar{d} / (s_d / \sqrt{n})$$

其中

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i, \quad s_d = \left(\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 \right)^{1/2}$$

在给定显著性水平 α 下, 该检验问题的拒绝域是

$$W_1 = \{|t_2| \geq t_{1-\alpha/2}(n-1)\}$$

这就是成对数据的 t 检验.

正态总体方差的检验

Theorem 5.0.23 — 单个正态总体方差的 χ^2 检验. 设 x_1, \dots, x_n 是来自 $N(\mu, \sigma^2)$ 的样本, 对方差亦可考虑如下三个检验问题

$$\begin{array}{lll} \text{I} & H_0: \sigma^2 \leq \sigma_0^2 & \text{vs} \quad H_1: \sigma^2 > \sigma_0^2 \\ \text{II} & H_0: \sigma^2 \geq \sigma_0^2 & \text{vs} \quad H_1: \sigma^2 < \sigma_0^2 \\ \text{III} & H_0: \sigma^2 = \sigma_0^2 & \text{vs} \quad H_1: \sigma^2 \neq \sigma_0^2 \end{array}$$

其中 σ_0^2 是已知常数. 此处通常假定 μ 未知, 它们采用的检验统计量是相同的, 均为

$$\chi^2 = (n-1)s^2/\sigma_0^2$$

在 $\sigma^2 = \sigma_0^2$ 时, $\chi^2 \sim \chi^2(n-1)$, 于是, 若取显著性水平为 α , 则对应三个检验问题的显著性水平为 α 的检验的拒绝域依次为

$$\begin{aligned} W_{\text{I}} &= \left\{ \chi^2 \geq \chi_{1-\alpha}^2(n-1) \right\} \\ W_{\text{II}} &= \left\{ \chi^2 \leq \chi_{\alpha}^2(n-1) \right\} \\ W_{\text{III}} &= \left\{ \chi^2 \leq \chi_{\alpha/2}^2(n-1) \text{ 或 } \chi^2 \geq \chi_{1-\alpha/2}^2(n-1) \right\} \end{aligned}$$

我们亦可给出检验的 p 值. 对单侧检验, 想法是类似的, 记 $\chi_0^2 = (n-1)s^2/\sigma_0^2$ 是由样本计算得到的检验统计量的值, χ^2 表示自由度为 $n-1$ 的 χ^2 分布的统计量, 则检验问题 I, II 的 p 值分别为 $p_{\text{I}} = P(\chi^2 \geq \chi_0^2)$, $p_{\text{II}} = P(\chi^2 \leq \chi_0^2)$. 双侧检验的拒绝域在两侧, 用 χ_0^2 可算得两个尾部概率 $P(\chi^2 \leq \chi_0^2)$ 和 $P(\chi^2 \geq \chi_0^2)$, 其和为 1, 其中必有一个 ≤ 0.5 . 检验的注意力总放在拒绝域上, 故应从中选一个与 $\alpha/2$ 比较, 从而检验问题 III 的 p 值为

$$p_{\text{III}} = 2 \min \{P(\chi^2 \geq \chi_0^2), P(\chi^2 \leq \chi_0^2)\}$$

对这样定义的 p 值, 有:

1. 当 $p_{\text{III}} \leq \alpha$ 时, 应拒绝原假设.
2. 当 $p_{\text{III}} > \alpha$ 时, 应接受原假设.

由于方差与标准差是一一对应关系, 不等式 $\sigma^2 \leq \sigma_0^2$ 等价于 $\sigma \leq \sigma_0$, 故上述讨论一样适用于对标准差的检验问题

Theorem 5.0.24 — 两个正态总体方差比的 F 检验. 设 x_1, \dots, x_m 是来自 $N(\mu_1, \sigma_1^2)$ 的样本, y_1, \dots, y_n 是来自 $N(\mu_2, \sigma_2^2)$ 的样本. 考虑如下三个假设检验问题

$$\begin{array}{lll} \text{I} & H_0: \sigma_1^2 \leq \sigma_2^2 & \text{vs} \quad H_1: \sigma_1^2 > \sigma_2^2 \\ \text{II} & H_0: \sigma_1^2 \geq \sigma_2^2 & \text{vs} \quad H_1: \sigma_1^2 < \sigma_2^2 \\ \text{III} & H_0: \sigma_1^2 = \sigma_2^2 & \text{vs} \quad H_1: \sigma_1^2 \neq \sigma_2^2 \end{array}$$

此处 μ_1, μ_2 均未知, 记 s_x^2, s_y^2 分别是由 x_1, \dots, x_m 算得的 σ_1^2 的无偏估计和由 y_1, \dots, y_n 算得的 σ_2^2 的无偏估计 (两个都是样本方差), 则可建立如下的检验统计量

$$F = \frac{s_x^2}{s_y^2}$$

检验法	H_0	H_1	检验统计量	拒绝域	p 值
χ^2 检验	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$		$x^2 \geq x_{1-\alpha}^2(n-1)$	$P(\chi^2 \geq x_0^2)$
	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 \leq \chi_\alpha^2(n-1)$	$P(\chi^2 \leq \chi_0^2)$
	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$		$\chi^2 \leq \chi_{\alpha/2}^2(n-1)$ 或 $\chi^2 \geq \chi_{1-\alpha/2}^2(n-1)$	$2 \min \{P(\chi^2 \leq \chi_0^2), P(\chi^2 \geq \chi_0^2)\}$
F 检验	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$		$F \geq F_{1-\alpha}(m-1, n-1)$	$P(F \geq F_0)$
	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F = \frac{s_x^2}{s_y^2}$	$F \leq F_\alpha(m-1, n-1)$	$P(F \leq F_0)$
	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$		$F \leq F_{\alpha/2}(m-1, n-1)$ 或 $F \geq F_{1-\alpha/2}(m-1, n-1)$	$2 \min \{P(F \leq F_0), P(F \geq F_0)\}$

均值	假设	检验统计量	拒绝域
$\mu = \mu_0$	$\sigma^2 = \sigma_0^2 \longleftrightarrow \sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma_0^2}$	$\{\chi^2 < \chi_{1-\alpha/2}^2(n)\} \cup \{\chi^2 > \chi_{\alpha/2}^2(n)\}$
	$\sigma^2 = \sigma_0^2 \longleftrightarrow \sigma^2 < \sigma_0^2$		$\{\chi^2 < \chi_{1-\alpha}^2(n)\}$
	$\sigma^2 = \sigma_0^2 \longleftrightarrow \sigma^2 > \sigma_0^2$		$\{\chi^2 > \chi_\alpha^2(n)\}$
	$\sigma^2 \leq \sigma_0^2 \longleftrightarrow \sigma^2 > \sigma_0^2$		$\{\chi^2 > \chi_\alpha^2(n)\}$
	$\sigma^2 \geq \sigma_0^2 \longleftrightarrow \sigma^2 < \sigma_0^2$		$\{\chi^2 < \chi_{1-\alpha}^2(n)\}$
μ 未知	$\sigma^2 = \sigma_0^2 \longleftrightarrow \sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma_0^2}$	$\{\chi^2 < \chi_{1-\alpha/2}^2(n-1)\} \cup \{\chi^2 > \chi_{\alpha/2}^2(n-1)\}$
	$\sigma^2 = \sigma_0^2 \longleftrightarrow \sigma^2 < \sigma_0^2$		$\{\chi^2 < \chi_{1-\alpha}^2(n-1)\}$
	$\sigma^2 = \sigma_0^2 \longleftrightarrow \sigma^2 > \sigma_0^2$		$\{\chi^2 > \chi_\alpha^2(n-1)\}$
	$\sigma^2 \leq \sigma_0^2 \longleftrightarrow \sigma^2 > \sigma_0^2$		$\{\chi^2 > \chi_\alpha^2(n-1)\}$
	$\sigma^2 \geq \sigma_0^2 \longleftrightarrow \sigma^2 < \sigma_0^2$		$\{\chi^2 < \chi_{1-\alpha}^2(n-1)\}$

当 $\sigma_1^2 = \sigma_2^2$ 时, $F \sim F(m-1, n-1)$, 由此给出三个检验问题对应的拒绝域依次为

$$\begin{aligned} W_I &= \{F \geq F_{1-\alpha}(m-1, n-1)\} \\ W_{II} &= \{F \leq F_\alpha(m-1, n-1)\} \\ W_{III} &= \{F \leq F_{\alpha/2}(m-1, n-1) \text{ 或 } F \geq F_{1-\alpha/2}(m-1, n-1)\} \end{aligned}$$

此时检验的 p 值的讨论与前述 χ^2 是相似的, 记 $F_0 = s_x^2/s_y^2$ 是由样本计算得到的检验统计量的值, F 表示服从 $F(m-1, n-1)$ 的统计量, 则检验问题 I, II, III 的 p 值分别为

$$\begin{aligned} p_I &= P(F \geq F_0) \\ p_{II} &= P(F \leq F_0) \\ p_{III} &= 2 \min \{P(F \geq F_0), P(F \leq F_0)\} \end{aligned}$$

Table 5.2: 两样本正态总体方差的显著性检验

讨厌参数	假设	检验统计量	拒绝域
μ_1, μ_2 已知	$\sigma_1^2 = \sigma_2^2 \longleftrightarrow \sigma_1^2 \neq \sigma_2^2$	$F = \frac{\sum_{i=1}^m (X_i - \mu_1)^2 / m}{\sum_{i=1}^n (Y_i - \mu_2)^2 / n}$	$\{F \leq F_{1-\alpha/2}(m, n)\} \cup \{F \geq F_{\alpha/2}(m, n)\}$
	$\sigma_1^2 \leq \sigma_2^2 \longleftrightarrow \sigma_1^2 > \sigma_2^2$		$\{F \geq F_\alpha(m, n)\}$
	$\sigma_1^2 \geq \sigma_2^2 \longleftrightarrow \sigma_1^2 < \sigma_2^2$		$\{F \leq F_{1-\alpha}(m, n)\}$
μ_1, μ_2 未知	$\sigma_1^2 = \sigma_2^2 \longleftrightarrow \sigma_1^2 \neq \sigma_2^2$	$F = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 / (m-1)}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)}$	$\{F \leq F_{1-\alpha/2}(m-1, n-1)\} \cup \{F \geq F_{\alpha/2}(m-1, n-1)\}$
	$\sigma_1^2 \leq \sigma_2^2 \longleftrightarrow \sigma_1^2 > \sigma_2^2$		$\{F \geq F_\alpha(m-1, n-1)\}$
	$\sigma_1^2 \geq \sigma_2^2 \longleftrightarrow \sigma_1^2 < \sigma_2^2$		$\{F \leq F_{1-\alpha}(m-1, n-1)\}$

两样本正态总体均值的显著性检验

方差	假设	检验统计量	拒绝域
σ_1^2, σ_2^2 已知	$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \mu_1 \neq \mu_2$	$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}}$	$\{ U \geq u_{\alpha/2}\}$
	$H_0: \mu_1 \leq \mu_2 \longleftrightarrow H_1: \mu_1 > \mu_2$		$\{U \geq u_\alpha\}$
	$H_0: \mu_1 \geq \mu_2 \longleftrightarrow H_1: \mu_1 < \mu_2$		$\{U \leq -u_\alpha\}$
$\sigma_1^2 = \sigma_2^2$ 未知	$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \mu_1 \neq \mu_2$	$T = \sqrt{\frac{mn}{m+n}} \frac{\bar{X} - \bar{Y}}{S_{mn}^*$	$\{ T \geq t_{\alpha/2}(m+n-2)\}$
	$H_0: \mu_1 \leq \mu_2 \longleftrightarrow H_1: \mu_1 > \mu_2$		$\{T \geq t_\alpha(m+n-2)\}$
	$H_0: \mu_1 \geq \mu_2 \longleftrightarrow H_1: \mu_1 < \mu_2$		$\{T \leq -t_\alpha(m+n-2)\}$
σ_1^2, σ_2^2 未知时 m, n 充分大时	$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \mu_1 \neq \mu_2$	$U = \frac{\bar{X} - \bar{Y}}{\sqrt{S_{1m}^2/m + S_{2n}^2/n}}$	$\{ U \geq u_{\alpha/2}\}$
	$H_0: \mu_1 \leq \mu_2 \longleftrightarrow H_1: \mu_1 > \mu_2$		$\{U \geq u_\alpha\}$
	$H_0: \mu_1 \geq \mu_2 \longleftrightarrow H_1: \mu_1 < \mu_2$		$\{U \leq -u_\alpha\}$
σ_1^2, σ_2^2 未知 m, n 不都充分大时	$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \mu_1 \neq \mu_2$	$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_{1m}^2/m + S_{2n}^2/n}}$	$\{ T \geq t_{\alpha/2}(r)\}$
	$H_0: \mu_1 \leq \mu_2 \longleftrightarrow H_1: \mu_1 > \mu_2$		$\{T \geq t_\alpha(r)\}$
	$H_0: \mu_1 \geq \mu_2 \longleftrightarrow H_1: \mu_1 < \mu_2$		$\{T \leq -t_\alpha(r)\}$

$$S_{mn}^{*2} = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n-2}$$

5.1 其他分布的假设检验

5.1.1 指数分布参数的假设检验

Theorem 5.1.1 设 x_1, \dots, x_n 是来自指数分布 $\text{Exp}(1/\theta)$ 的样本, θ 为其均值, 现考虑关于 θ 的如下检验问题

$$\text{I} \quad H_0: \theta \leq \theta_0 \quad \text{vs} \quad H_1: \theta > \theta_0$$

为寻找检验统计量, 我们考察参数 θ 的充分统计量 \bar{x} . 在 $\theta = \theta_0$ 时, $n\bar{x} = \sum_{i=1}^n x_i \sim \text{Ga}(n, 1/\theta_0)$, 由伽玛分布性质可知

$$\chi^2 = \frac{2n\bar{x}}{\theta_0} \chi^2(2n)$$

于是可用 χ^2 作为检验统计量并利用 $\chi^2(2n)$ 的分位数建立检验的拒绝域, 对检验问题 I, 拒绝域形式为 $W_1 = \{\chi^2 \geq c\}$, 对给定的显著性水平 α , 可由 $P(W_1) = \alpha$ 获得拒绝域如下

$$W_1 = \{\chi^2 \geq \chi^2_{1-\alpha}(2n)\}$$

p 值的讨论, 记 $\chi_0^2 = \frac{2n\bar{x}}{\theta_0}$ 为由样本算得的检验统计量值, 则检验的 p 值为 $p_I = P(\chi^2 \geq \chi_0^2)$, 其中 χ^2 表示服从 $\chi^2(2n)$ 分布的随机变量. 关于 θ 的另两种检验问题处理方法类似. 对检验问题

$$\text{II } H_0 : \theta \geq \theta_0 \quad \text{vs} \quad H_1 : \theta < \theta_0 \quad \text{和 III } H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

检验统计量不变, 拒绝域以及检验的 p 值分别为

$$\begin{aligned} W_{\text{II}} &= \left\{ \chi^2 \leq \chi_{\alpha}^2(2n) \right\}, \quad p_{\text{II}} = P(\chi^2 \leq \chi_0^2) \\ W_{\text{III}} &= \left\{ \chi^2 \leq \chi_{\alpha/2}^2(2n) \text{ 或 } \chi^2 \geq \chi_{1-\alpha/2}^2(2n) \right\}, \quad p_{\text{III}} = 2 \min \{ P(\chi^2 \geq \chi_0^2), P(\chi^2 \leq \chi_0^2) \} \end{aligned}$$

Theorem 5.1.2 — 比率 p 检验. 比率 p 可看作某事件发生的概率, 即可看作二点分布 $b(1, p)$ 中的参数. 作 n 次独立试验, 以 x 记该事件发生的次数, 则 $x \sim b(n, p)$. 我们可以根据 x 检验关于 p 的一些假设. 先考虑如下单边假设检验问题

$$I \quad H_0 : p \leq p_0 \quad \text{vs} \quad H_1 : p > p_0$$

直观上看, 一个显然的检验方法是取如下的拒绝域 $W = \{x \geq c\}$, 由于 x 只取整数值, 故 c 可限制在非负整数中. 然而, 一般情况下对给定的 α , 不一定能正好取到一个 c 使

$$P(x \geq c; p_0) = \sum_{i=c}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} = \alpha$$

能恰巧使得成立的 c 值是罕见的. 在这种情况下, 较常见的是找一个 c_0 , 使得

$$\sum_{i=c_0}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} > \alpha > \sum_{i=c_0+1}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} \quad (5.7)$$

于是, 可取 $c = c_0 + 1$, 此时相当于把显著性水平由 α 降低到 $\sum_{i=c_0+1}^n \binom{n}{i} p_0^i (1-p_0)^{n-i}$ 因为它可保证 5.7 的左侧不大于 α , 从而是显著性水平为 α 的检验. 事实上, 在离散场合使用 p 值作检验较为简便, 这时可以不用找 c_0 , 而只需根据观测值 $x = x_0$ 计算检验的 p 值:

$$p = P_{p_0}(x \geq x_0)$$

并将之与事先给定的显著性水平比较大小即可, 其中 x 为服从 $b(n, p_0)$ 的随机变量. 譬如, $n = 40, p_0 = 0.1, x_0 = 8$, 则

$$p = 1 - 0.9^{40} - \binom{40}{1} 0.1 \times 0.9^{39} - \cdots - \binom{40}{7} 0.1^7 \times 0.9^{33} = 0.0419$$

于是, 若取 $\alpha = 0.05$, 由于 $p < \alpha$, 则应拒绝原假设.

Theorem 5.1.3 — 大样本时的 p 检验. 如果样本量较大, 人们还经常采用如下的大样本检验. 其一般思路如下: 设 x_1, \dots, x_n 是来自某总体的样本, 又设该总体均值为 θ , 方差为 θ 的函数, 记为 $\sigma^2(\theta)$, 譬如, 对二点分布 $b(1, \theta)$, 其方差 $\theta(1-\theta)$ 是均值 θ 的函数, 则对下

列三类假设检验问题

$$\begin{array}{lll} \text{I} & H_0 : \theta \leq \theta_0 & \text{vs} \quad H_1 : \theta > \theta_0 \\ \text{II} & H_0 : \theta \geq \theta_0 & \text{vs} \quad H_1 : \theta < \theta_0 \\ \text{III} & H_0 : \theta = \theta_0 & \text{vs} \quad H_1 : \theta \neq \theta_0 \end{array}$$

在样本容量 n 充分大时, 利用中心极限定理, $\bar{x} \sim N(\theta, \sigma^2(\theta)/n)$, 故在 $\theta = \theta_0$ 时可采用如下检验统计量

$$u = \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\sigma^2(\hat{\theta})}} \sim N(0, 1)$$

其中 $\hat{\theta}$ 为 θ 的 MLE, 并由此可近似地确定拒绝域. 对应上述三类检验问题的拒绝域依次为

$$\begin{aligned} W_{\text{I}} &= \{u \geq u_{1-\alpha}\} \\ W_{\text{II}} &= \{u \leq u_\alpha\} \\ W_{\text{III}} &= \{|u| \geq u_{1-\alpha/2}\} \end{aligned}$$

5.2 似然比检验与分布拟合检验

Theorem 5.2.1 — (似然比统计量).

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$$

设 X_1, \dots, X_n 为来自分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$ 的 iid 样本, 对于感兴趣的假设 (4.5.1), 令

$$\lambda(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} f(\mathbf{X}, \theta)}{\sup_{\theta \in \Theta} f(\mathbf{X}, \theta)}$$

则我们称统计量 $\lambda(X)$ 为假设的似然比 (Likelihood Ratio), 有时也称之为广义似然比。从 $\lambda(X)$ 的定义不难看出, 如果 $\lambda(X)$ 的值很小, 则说明 $\theta \in \Theta_0$ 的可能性要比 $\theta \in \Theta$ 的可能性小, 于是, 我们有理由认为 H_0 不成立。这样, 我们有如下的似然比检验。

Definition 5.2.1 — (似然比检验). 似然比统计量 $\lambda(X)$ 作为假设的检验统计量, 且取其拒绝域为 $\{\lambda(\mathbf{x}) \leq c\}$ 时, 其中临界值 c 满足

$$P_\theta\{\lambda(\mathbf{X}) \leq c\} \leq \alpha, \forall \theta \in \Theta_0$$

则称此检验为水平 α 的似然比检验 (Likelihood Ratio Test, 简记为 LRT).

Theorem 5.2.2 — Wilks 定理. 假设在 H_0 下我们对 k 维参数向量 $\theta = (\theta_1, \dots, \theta_k)^T$ 有 r 个约束 (则 θ 中只有 $k-r$ 个分量可以是自由的). 在适当条件下, $-2 \ln \lambda(\mathbf{X})$ 的极限零分布是 χ_r^2 分布, 即收敛到自由度是备选假设和原假设下自由参数个数之差的卡方分布。该定理是由 Wilks 在 1938 年的数

■ **Example 5.3** 设 X_1, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的 iid 样本, μ, σ^2 均未知. 试求假设

$$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0$$

的水平为 α 的似然比检验.

Proof.

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{\sup_{\theta \in \Theta_0} f(\mathbf{x}, \theta)}{\sup_{\theta \in \Theta} f(\mathbf{x}, \theta)} = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right)^{n/2} \\ &= \left[\frac{1}{\frac{\sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]^{n/2} = \left[\frac{1}{1 + n \frac{(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]^{n/2} \\ &= \left[1 + \frac{T^2}{n-1} \right]^{-n/2}\end{aligned}$$

另外，从上式可知，此时的似然比统计量与传统的 t 统计量的平方成反比，于是，两个检验统计量的拒绝域有如下关系：

$$\{\lambda(\mathbf{x}) \leq c\} \iff \{|T(\mathbf{x})| \geq d\}$$

又因为当 H_0 成立时， $T \sim t(n-1)$ ，故我们取 $d = t_{\alpha/2}(n-1)$ 即可控制其第一类错误概率不超过 α 。由此可见，此时的似然比检验与我们前面讲过的双边 t 检验完全等价。 ■

Definition 5.2.2 — 似然比检验--茆版. 设 x_1, \dots, x_n 为来自密度函数为 $p(x; \theta), \theta \in \Theta$ 的样本，则对检验问题

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1 = \Theta - \Theta_0$$

可用似然比统计量

$$\Lambda = \frac{\sup_{\theta \in \Theta} p(x_1, \dots, x_n; \theta)}{\sup_{\theta \in \Theta_0} p(x_1, \dots, x_n; \theta)}$$

检验统计量，该检验称为似然比 (Likelihood Ratio) 检验，有时也称之为广义似然比检验。

检验统计量也可以写为

$$\Lambda = \frac{p(x_1, \dots, x_n; \hat{\theta})}{p(x_1, \dots, x_n; \hat{\theta}_0)}$$

其中 $\hat{\theta}$ 表示在全参数空间 Θ 上 θ 的最大似然估计， $\hat{\theta}_0$ 表示在原假设成立时子参数空间 Θ_0 上 θ 的最大似然估计。拒绝域为 $W = \{\Lambda \geq c\}$ ，其中临界值 c 由

$$P_{\theta}(\Lambda \geq c) \leq \alpha (\forall \theta \in \Theta_0)$$

确定。称此检验为显著性水平 α 的似然比检验 (likelihood ratio test)，简记为 LRT。

似然比检验方法是产生检验统计量的另一种思路。该似然比检验统计量没有统一的精确分布形式，但其对数似然比的 2 倍， $2 \ln \Lambda$ ，渐近服从 $\chi^2(k)$ 分布，其中 k 为 Λ 中独立参数个数。

■ **Example 5.4** 例 7.4.1 设 x_1, \dots, x_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本， μ, σ^2 均未知。试求检验问题

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

的显著性水平为 α 的似然比检验。解 记 $\theta = (\mu, \sigma^2)$ ，样本联合密度函数为

$$p(x_1, \dots, x_n; \theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

两个参数空间分别为

$$\Theta_0 = \{(\mu_0, \sigma^2) | \sigma^2 > 0\}, \quad \Theta = \{(\mu, \sigma^2) | \mu \in \mathbf{R}, \sigma^2 > 0\}$$

利用微分法, 我们容易求得在 Θ 上 $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ 分别为 μ 与 σ^2 的 MLE, 在 Θ_0 上 $\frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2$ 是 σ^2 的 MLE, 代回各自似然函数后, 可得

$$\begin{aligned} \sup_{\theta \in \Theta_0} p(x_1, \dots, x_n; \theta) &= \left[2\pi \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2 \right]^{-n/2} e^{-n/2} \\ \sup_{\theta \in \Theta} p(x_1, \dots, x_n; \theta) &= \left[2\pi \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{-n/2} e^{-n/2} \end{aligned}$$

于是, 其似然比统计量为

$$\begin{aligned} \Lambda(x_1, \dots, x_n) &= \frac{\sup_{\theta \in \Theta} p(x_1, \dots, x_n; \theta)}{\sup_{\theta \in \Theta_0} p(x_1, \dots, x_n; \theta)} = \left(\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{n/2} \\ &= \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{n/2} = \left(1 + \frac{t^2}{n-1} \right)^{n/2} \end{aligned}$$

其中 $t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$ 是 t 检验统计量. 从上式可知, 此时的似然比统计量 Λ 是传统的 t 统计量平方的严增函数, 于是, 两个检验统计量的拒绝域有如下等价关系:

$$\{\Lambda(x_1, \dots, x_n) \geq c\} \Leftrightarrow \{|t| \geq d\}$$

且由 t 的分位数可定出 Λ 的分位数. 又因为当 H_0 成立时, $t \sim t(n-1)$, 若我们取 $d = t_{1-\alpha/2}(n-1)$ 则用 $c = \left[1 + \frac{\alpha^2}{n-1} \right]^{n/2}$ 就可控制用 Λ 犯第一类错误的概率不超过 α . 由此可见, 此时的似然比检验与我们前面讲过的双侧 t 检验完全等价.

Theorem 5.2.3 — 总体可以分成 k 类: A_1, \dots, A_k 时的分布拟合优度检验. 原假设

$$H_0 : P(A_i) = p_i, \quad i = 1, 2, \dots, k, \tag{5.8}$$

其中诸 $p_i \geq 0$, 且 $\sum_{i=1}^k p_i = 1$

数据: 对总体作 n 次观察, k 个类各出现的频数分别为 n_1, \dots, n_k , 且 $\sum_{i=1}^k n_i = n$. 分两种情况给出检验统计量及其拒绝域:

1. 诸 p_i 均已知, 由于当 H_0 成立时, 在 n 个样本中属于 A_i 类的理论个数”或“期望个数”为 np_i , 而我们实际观测到的值为 n_i , 故当 H_0 成立时, n_i 与 np_i 应相差不大. 于是检验统计量

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \tag{5.9}$$

来衡量“理论个数”与实际个数间的差异. 分子是实际观测数与期望观测数的偏差的平方, 而 $\frac{(n_i - np_{i0})^2}{np_{i0}}$ 可以看成是 $(n_i - np_{i0})^2$ 的规范化, 所以提供了实际观测数与期望观测数接近程度的一个度量, 当 H_0 为真时, 它的值应该比较小, 所以, 拒绝域为

$$W = \{\chi^2 \geq \chi^2_{1-\alpha}(k-1)\}$$

2. 诸 p_i 不完全已知, 在同样的条件下, 可以先用最大似然估计方法估计出这 r 个未

知参数, 然后再算出 p_i 的估计值 $\hat{p}_i (i = 1, 2, \dots, k)$ 这时, 统计量检验统计量

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \quad (5.10)$$

拒绝域为

$$\{\chi^2 \geq \chi^2_{1-\alpha}(k-r-1)\}$$

其中 r 为 p_1, p_2, \dots, p_k 中独立参数个数 (也就是 p_i 依赖于这 r 个参数), \hat{p}_i 为 p_i 的最大似然估计.

这个检验被皮尔逊 (K. Pearson) 称为 χ^2 拟合优度检验, $p = P(\chi^2 \geq \chi^2_0)$ 被称为拟合优度, p 值愈大拟合优度愈好, p 值愈小拟合优度愈差, 从而拒绝原假设 H_0 . 这就是检验的 p 值最原始的想法. 两种情况都要求是大样本, 需要样本个数 n 大于等于 5. 拟合优度对于检验分布时, 区间的划分有要求

Theorem 5.2.4 在前述各项假定下, 在 H_0 成立时, 对 5.9 的检验统计量有

$$\chi^2 \xrightarrow{L} \chi^2(r-1)$$

Proof. 对最简单的 $r=2$ 给出证明. 当 $r=2$ 时, $n_1 \sim b(n, p_{10})$, 且, $p_{10} + p_{20} = 1, n_1 + n_2 = n, n_1 - np_{10} = (n - n_2) - n(1 - p_{20}) = np_{20} - n_2$, 故

$$\chi^2 = \frac{(n_1 - np_{10})^2}{np_{10}} + \frac{(n_2 - np_{20})^2}{np_{20}} = \frac{(n_1 - np_{10})^2}{np_{10}p_{20}}$$

而由中心极限定理可知,

$$\frac{n_1 - np_{10}}{\sqrt{np_{10}p_{20}}} \xrightarrow{L} N(0, 1)$$

故 $\chi^2 \xrightarrow{L} \chi^2(1)$. 一般场合的证明此处从略. ■

Theorem 5.2.5 — 从似然比检验得到上述皮尔逊 χ^2 检验统计量. 样本联合分布为

$$P_\theta(X_1 = x_1, \dots, X_n = x_n) = p_1^{n_1} \cdots p_r^{n_r} = \prod_{i=1}^r p_i^{n_i}$$

由此可求得

$$\sup_{\theta \in \Theta} P_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^r \left(\frac{n_i}{n}\right)^{n_i} \quad \sup_{\theta \in \Theta_0} P_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^r p_{i0}^{n_i}$$

于是, 其似然比统计量为

$$\Lambda(x_1, \dots, x_n) = \prod_{i=1}^r \left(\frac{n_i}{np_{i0}}\right)^{n_i}$$

另外, 由于

$$\begin{aligned}
 & \ln \Lambda(x_1, \dots, x_n) \\
 &= \sum_{i=1}^r n_i \ln \frac{n_i}{np_{i0}} \\
 &= \sum_{i=1}^r [np_{i0} + (n_i - np_{i0})] \ln \left(1 + \frac{n_i - np_{i0}}{np_{i0}} \right) \\
 &= \sum_{i=1}^r [np_{i0} + (n_i - np_{i0})] \left\{ \frac{n_i - np_{i0}}{np_{i0}} - \frac{1}{2} \left(\frac{n_i - np_{i0}}{np_{i0}} \right)^2 + o_p(n^{-2}) \right\} \\
 &\approx \frac{1}{2} \sum_{i=1}^r \frac{(n_i - np_{i0})^2}{np_{i0}} + o_p(n^{-1})
 \end{aligned}$$

所以, $2 \ln \Lambda(x_1, \dots, x_n) \approx \sum_{i=1}^r \frac{(n_i - np_{i0})^2}{np_{i0}}$, 由单调性, 此处似然比检验与皮尔逊引进的 χ^2 拟合优度检验等价.

Theorem 5.2.6 — Pearson χ^2 拟合优度检验.. 检验统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i}$$

拒绝域

$$W = \{\chi^2 \geq \chi_{\alpha}^2(r-1)\}$$

可以检验理论频数和实际频数是否符合; 也可以检验不含有未知参数的分布: 对于下面一般的分布假设

$$H_0: F(x) \equiv F_0(x)$$

我们仍可利用上述的 χ^2 拟合优度检验, 其中 $F_0(x)$ 为一个完全已知的分布函数 (形式及参数均已知) 此时, 我们可以把 $(-\infty, +\infty)$ (或样本空间) 分成 r 个互不相交的区间:

$$(-\infty, \infty) = \bigcup_{i=1}^r I_i = (-\infty, a_1) \cup [a_1, a_2] \cup \dots \cup [a_{r-1}, \infty)$$

且以 n_i 记落在第 i 个区间 I_i 内的样本个数, 再记

$$\begin{aligned}
 p_1 &= F(a_1), & p_{10} &= F_0(a_1) \\
 p_2 &= F(a_2) - F(a_1), & p_{20} &= F_0(a_2) - F_0(a_1) \\
 &\dots & &\dots \\
 p_r &= 1 - F(a_{r-1}), & p_{r0} &= 1 - F_0(a_{r-1})
 \end{aligned}$$

则我们可以利用统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_{i0})^2}{np_{i0}}$$

R 在一般情形下, 分点的选取应保证落在每个区间的样本点个数不小于 5, 且总的样本量不应小于 30。

通过上面分析，我们可以看到，当 F_0 中含有未知参数时，上述拟合优度检验无法实施。

Theorem 5.2.7 处理待检验分布含有未知参数的情况：

$$H_0 : F(x) \equiv F_0(x; \theta_1, \dots, \theta_k)$$

其中 $F_0(x; \theta_1, \dots, \theta_k)$ 是依赖于 k 个未知参数的形式已知的分布。用样本估计未知参数后，可以得到 \hat{p}_{i0} ，利用下面的统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}}$$

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \xrightarrow{\mathcal{L}} \chi^2(r - 1 - k)$$

5.2.1 p 值

单边情形

$$W = \{\mathbf{X} : T(\mathbf{X}) > c\} \text{ 或 } W = \{\mathbf{X} : T(\mathbf{X}) < c\}$$

定义 4.6.1 对于拒绝域形如 (4.6.2) 的单边检验，当给定样本观测值 \mathbf{x}^0 后，称

$$p(\mathbf{x}^0) = \sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) \geq T(\mathbf{x}^0)\} \text{ 或 } \sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) \leq T(\mathbf{x}^0)\}$$

为此检验的 p 值。

Theorem 5.2.8 对于给定的 $\alpha \in (0, 1)$ ，如存在常数 c 满足 $\sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) > c\} = \alpha$ ，则样本 x^0 落入拒绝域 $W = \{\mathbf{X} : T(\mathbf{X}) > c\}$ 的充要条件是其 p 值 $p(\mathbf{x}^0)$ 小于 α 。

Proof. 对于给定的样本值 x^0 ，如果 $p(\mathbf{x}^0) < \alpha$ ，即 $\sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) \geq T(\mathbf{x}^0)\} < \alpha$ ，而由已知条件知常数 c 满足 $\sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) > c\} = \alpha$ ，则必有 $T(\mathbf{x}^0) > c$ ，这就是说样本 x^0 落入了拒绝域 W 反之，如果样本 x^0 落入拒绝域 W ，即 $T(\mathbf{x}^0) > c$ ，则存在一个正数 $\varepsilon > 0$ ，使得 $T(\mathbf{x}^0) - \varepsilon > c$ 于是

$$\begin{aligned} p(T(\mathbf{x}^0)) &= \sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) \geq T(\mathbf{x}^0)\} \\ &\leq \sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) > T(\mathbf{x}^0) - \varepsilon\} \\ &< \sup_{\theta \in \Theta_0} P_\theta \{T(\mathbf{X}) > c\} = \alpha \end{aligned}$$

■

(R) 样本值 x^0 落入水平为 α 的拒绝域 $W = \{\mathbf{X} : T(\mathbf{X}) > c\}$ 的充要条件是此样本的 p 值小于 α 。换句话说，当且仅当样本值的 p 值小于 α 时拒绝 H_0 ，也就是说，p 值是可以拒绝原假设的水平的最小值。我们注意到，引入 p 值的最大优点在于：在做检验时，我们不需要事先给定此检验的显著性水平 α ，而通过计算当前样本的 p 值知道，对一切大于此 p 值的 α ，则错误拒绝 H_0 的概率不超过 α

5.3 优势检验

Definition 5.3.1 — (检验). 设 $\phi(\mathbf{X})$ 是定义在 \mathcal{X} 上的可测函数, 满足 $0 \leq \phi(\mathbf{x}) \leq 1$, 则称 $\phi(\mathbf{X})$ 为检验函数, 简称检验。如果 $\phi(\mathbf{x})$ 仅取 0,1 两值时, 则称为非随机化检验, 否则, 就称为随机化检验, 其势函数为

$$\beta_\phi(\theta) = E_\theta \phi(\mathbf{X})$$

从上述定义可以看出, 上一章我们称拒绝域的示性函数为一个检验, 而本章则不然

Theorem 5.3.1 对于一个随机化检验 $\phi(x)$, 当其取值为 1 时, 我们拒绝原假设; 当其取值为 0 时, 我们不能拒绝原假设; 当其取值为 $\delta \in (0, 1)$ 时, 我们如下处理: 取一个来自于 $b(1, \delta)$ 的随机数, 如此随机数为 1, 则拒绝原假设, 否则不能拒绝原假设。

另外, 当检验统计量为连续随机变量时, 我们采用的是非随机化检验。当检验统计量为离散随机变量时, 就有可能采用随机化检验。这一点, 我们将在后面接触到。

Theorem 5.3.2 设 $\phi_1(\mathbf{X})$ 和 $\phi_2(\mathbf{X})$ 是检验问题 $H_0 \longleftrightarrow H_1$ 的检验函数, 如果它们的势函数相同, 即

$$E_\theta \phi_1(\mathbf{X}) = E_\theta \phi_2(\mathbf{X}), \forall \theta \in \Theta$$

则称检验函数 $\phi_1(\mathbf{X})$ 和 $\phi_2(\mathbf{X})$ 等价.

Theorem 5.3.3 设 X_1, \dots, X_n 是来自分布族 $\{f(x, \theta) : \theta \in \Theta\}$ 的样本. $T(\mathbf{X})$ 是参数 θ 的充分统计量。则对手任意一个检验函数 $\phi(\mathbf{X})$, 存在另一个只依赖于 $T(\mathbf{X})$ 的检验函数与它等价.

Proof. 定义一个新统计量 $\psi(t) = E[\phi(\mathbf{X})|T(\mathbf{X}) = t]$. 由于 $0 \leq \phi(\mathbf{x}) \leq 1$, 故由条件期望性质知, $\psi(t) \in [0, 1]$, 故它也是一个检验函数. 另外, 又由全期望公式有

$$E_\theta \psi(T(\mathbf{X})) = E_\theta [E(\phi(\mathbf{X})|T(\mathbf{X}))] = E_\theta \phi(\mathbf{X}), \forall \theta \in \Theta$$

故知 $\psi(T(\mathbf{X}))$ 与 $\phi(\mathbf{x})$ 是等价的.

这个定理告诉我们, 当 θ 的充分统计量存在时, 关于此参数的检验问题, 我们仅需在由充分统计量构成的检验函数中去寻找就可以了。这就是假设检验中的“充分性原则”。 ■

Definition 5.3.2 — (MP 检验). 对于参数分布族 $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$, 设我们感兴趣的假设为

$$H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta = \theta_1 (\theta_1 \neq \theta_0)$$

并设有两个水平为 α 的检验函数 $\phi_1(\mathbf{X}), \phi_2(\mathbf{X})$, 即满足

$$E_{\theta_0} \phi_i(\mathbf{X}) \leq \alpha, \quad i = 1, 2$$

如果

$$\beta_{\phi_1}(\theta_1) \geq \beta_{\phi_2}(\theta_1)$$

则称检验 ϕ_1 比 ϕ_2 有效. 如果检验 ϕ_1 对于任一个水平小于等于 α 的检验 ϕ_2 , 上式均成立, 则称 ϕ_1 是假设的水平 α 的**最优势检验** (Most Powerful Test, 简记为 MPT). 从上述定义可以看出, 一个 MP 检验, 不仅控制其第一类错误概率, 而且也控制其第二类错误概率, 且比一般水平 α 的显著性检验有效。

Theorem 5.3.4 — (Neyman-Pearson 基本引理). 对于参数分布族 $\{f(x; \theta) : \theta \in \Theta = \{\theta_0, \theta_1\}\}$ 则关于检验问题

$$H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta = \theta_1 (\theta_1 \neq \theta_0)$$

我们有如下结论:

1. 对给定的 $\alpha \in (0, 1)$, 存在一个检验函数 $\phi(x)$ 及常数 $k \geq 0$, 使得

$$\phi(x) = \begin{cases} 1, & f(x; \theta_1) > kf(x; \theta_0) \\ 0, & f(x; \theta_1) \leq kf(x; \theta_0) \end{cases} \quad (5.11)$$

且

$$E_{\theta_0} \phi(X) = \alpha \quad (5.12)$$

2. 由 5.11 和 5.12 式确定的检验函数 $\phi(x)$ 是检验问题 5.12 的水平为 α 的 MPT.
3. 如果 $\phi'(x)$ 是此检验问题的水平 α 的 MPT, 则一定存在常数 $k \geq 0$, 使得 $\phi'(x)$ 满足 5.11 式. 又如 $\phi'(x)$ 满足 $E_{\theta_1} \phi'(X) < 1$, 则它也满足 5.12 式.

(R)

1. 从 N-P 引理的证明可以看出, MPT 是似然比统计量 $\lambda(X) = f(X; \theta_1)/f(X; \theta_0)$ 的函数, 这种检验我们也称之为似然比检验;
2. 如果似然比 $\lambda(X)$ 的分布是连续的, 则假设的 MPT 是非随机化的检验; 如果 $\lambda(X)$ 的分布是离散的, 则假设的 MPT 有可能是随机化的;
3. 对于非参数假设, N-P 引理仍是正确的。

Corollary 5.3.5 如 $\phi(X)$ 为假设的水平 α 的 MPT, 则必有 $\beta_\phi(\theta_1) \geq \alpha$. 如 $0 < \alpha < 1$

且 $f(x; \theta_1) \neq f(x; \theta_0)$, 则 $\beta_\phi(\theta_1) > \alpha$

■ **Example 5.5** 例 5.2.1 设 X_1, \dots, X_n 是来自 $N(\mu, 1)$ 的 IID 样本, 试考虑假设

$$H_0 : \mu = 0 \longleftrightarrow H_1 : \mu = \mu_1 (> 0)$$

的水平 α 的 MPT. 解此时样本的联合 PDF 为

$$f(x; \mu) = (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

由此可求得似然比统计量为

$$\lambda(x) = \exp \{n\mu_1 \bar{x} - n\mu_1^2/2\}$$

由此可见, $\lambda(x)$ 与 \bar{x} 成正比, 于是, 由 N-P 引理知, 水平 α 的 MPT 为

$$\phi(X) = \begin{cases} 1, & \bar{X} > k \\ 0, & \bar{X} \leq k \end{cases}$$

并满足 $E_{\mu=0}\phi(X) = \alpha$. 又由于在 H_0 成立时, $\bar{X} \sim N(0, 1/n)$, 故上面 MPT 中的 k 满足

$$\alpha = P_{H_0}\{\bar{X} > k\} = P_{H_0}\{\sqrt{n}\bar{X} > \sqrt{n}k\} = 1 - \Phi(\sqrt{n}k)$$

即 $k = u_\alpha/\sqrt{n}$ 综上而述, 知此假设的水平 α 的 MPT 为

$$\phi(X) = \begin{cases} 1, & \bar{X} > u_\alpha/\sqrt{n} \\ 0, & \bar{X} < u_\alpha/\sqrt{n} \end{cases}$$

■ **Example 5.6** 设 X_1, \dots, X_n 为来自 Poisson 分布 $P(\lambda)$ 的 IID 样本, 求假设

$$H_0 : \lambda = 1 \longleftrightarrow H_1 : \lambda = \lambda_1 (> 1)$$

水平为 α 的 MP 检验。

Proof. 此时的似然比统计量为

$$\lambda(x) = \prod_{i=1}^n f(x_i; \lambda_1) / \prod_{i=1}^n f(x_i; \lambda_0) = \lambda_1^{\sum_{i=1}^n x_i} \exp\{-n(\lambda_1 - 1)\}$$

由于它关于 $T(x) = \sum_{i=1}^n x_i$ 单调上升, 故由 N-P 引理知, 其水平为 α 的 MP 检验为

$$\phi(x) = \begin{cases} 1, & T(x) > k \\ \delta, & T(x) = k \\ 0, & T(x) < k \end{cases}$$

其中 k, δ 满足 $E_{H_0}\phi(X) = \alpha$. 由于当 H_0 成立时, $T(X) \sim P(n)$, 故满足

$$\alpha = P_{H_0}\{T(X) > k\} + \delta P_{H_0}\{T(X) = k\} = \sum_{i=k+1}^{\infty} \frac{e^{-n} n^i}{i!} + \delta \frac{n^k}{k!} e^{-n}$$

如果存在整数 k_0 , 使得 $\sum_{i=k_0+1}^{\infty} \frac{e^{-n} n^i}{i!} = \alpha$, 则此时的 MP 检验为

$$\phi(x) = \begin{cases} 1, & T(x) > k_0 \\ 0, & T(x) \leq k_0 \end{cases}$$

如果存在整数 k_0 , 使得

$$c_1 = \sum_{i=k_0+1}^{\infty} \frac{e^{-n} n^i}{i!} < \alpha < \sum_{i=k_0}^{\infty} \frac{e^{-n} n^i}{i!} = c_2$$

■

Definition 5.3.3 — (一致最优势检验)--复杂检验. 对于参数分布族 $\{f(x, \theta) : \theta \in \Theta\}$, 设感兴趣的假设为

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1$$

如果对于两个水平为 α 的检验 ϕ_1, ϕ_2 , 即

$$E_\theta \phi_i \leq \alpha, \forall \theta \in \Theta_0, i = 1, 2$$

且满足

$$\beta_{\phi_1}(\theta) = E_\theta \phi_1 \geq E_\theta \phi_2 = \beta_{\phi_2}(\theta), \forall \theta \in \Theta_1 \quad (5.13)$$

则称检验 ϕ_1 一致优于检验 ϕ_2 . 如果存在一个检验 ϕ_1 , 使对任何水平为 α 的检验 ϕ_2 , 均有5.13成立, 则称检验 ϕ_1 是假设的水平 α 的一致最优势 (Uniformly Most Powerful, 简记为 UMP) 检验。

Theorem 5.3.6 设有如下三个检验问题

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1 \quad (5.14)$$

$$H_0 : \theta \in \Theta_{01} \longleftrightarrow H_1 : \theta \in \Theta_1 \quad (5.15)$$

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta = \theta_1 \quad (5.16)$$

其中 $\Theta_{01} \subset \Theta_0, \theta_1 \in \Theta_1$, 并设 $\phi(x)$ 是检验问题 5.14 的水平 α 的检验, 则

1. 如果 $\phi(x)$ 是检验问题 5.15 的水平 α 的 UMPT, 则 $\phi(x)$ 也是 5.14 的水平 α 的 UMPT.
2. $\phi(x)$ 是检验问题 5.14 的水平 α 的 UMPT 的充要条件是, 对任意的 $\theta_1 \in \Theta_1, \phi(x)$ 是检验问题 5.16 的水平 α 的 MPT.

Theorem 5.3.7 设 $\phi(x)$ 是 5.14 的水平 α 的检验, 如果对某个 $\theta_0 \in \Theta_0$ 及任意一个 $\theta_1 \in \Theta_1, \phi(x)$ 都是假设

$$H_0 : \theta = \theta_0 \longleftrightarrow H_1 : \theta = \theta_1$$

的水平 α 的 MPT, 则 $\phi(x)$ 也是 5.14 的水平 α 的 UMPT.

■ **Example 5.7** 之前例子中给出的 MP 检验也是相应假设的 UMP, 即, 对于正态总体 $N(\mu, 1)$, 假设

$$H_0 : \mu = 0 \longleftrightarrow H_1 : \mu > 0$$

的水平为 α 的 UMPT 为

$$\phi(X) = \begin{cases} 1, & \bar{X} > u_\alpha / \sqrt{n} \\ 0, & \bar{X} < u_\alpha / \sqrt{n} \end{cases}$$

而假设

$$H_0 : \mu = 0 \longleftrightarrow H_1 : \mu < 0$$

的水平为 α 的 UMPT 为

$$\phi(X) = \begin{cases} 1, & \bar{X} < -u_\alpha / \sqrt{n} \\ 0, & \bar{X} > -u_\alpha / \sqrt{n} \end{cases}$$

Theorem 5.3.8 总而言之, 我们注意到:

1. 如果简单假设对简单假设的 MPT 不依赖于备选假设中的参数值, 则可适当扩大备选假设成复杂假设;
2. 如果简单假设对简单假设的 MPT 的势函数是单调的, 则也可以适当扩大原假设成复杂假设。
3. 但是, 我们要注意到, 并不是所有的复杂假设对复杂假设的 UMPT 都是存在

的，它的存在性不仅依赖于总体分布，而且还依赖于假设的复杂情况。本节将只考虑针对单调似然比分布族的两种单边假设和针对指数型分布族的三种双边假设的 UMP 检验问题。

Definition 5.3.4 如果一个检验的势函数在 Θ 之子集 Θ' 上保持不变，则称之为关于集合 Θ' 的相似检验 (similar test)

$$\begin{aligned} H_0 : \theta = \theta_0 &\longleftrightarrow H_1 : \theta \neq \theta_0 \\ H_0 : \theta_1 \leq \theta \leq \theta_2 &\longleftrightarrow H_1 : \theta < \theta_1 \text{ 或 } \theta > \theta_2 \end{aligned}$$

假设我们感兴趣的假设为

$$H_0 : \theta \in \Theta_0 \longleftrightarrow H_1 : \theta \in \Theta_1$$

定义 5.4.1 (无偏检验) 设 $\phi(x)$ 是上述假设的一个检验，如果其势函数 $E_\theta \phi(X)$ 满足条件

$$E_\theta \phi(X) \leq \alpha, \forall \theta \in \Theta_0, \quad E_\theta \phi(X) \geq \alpha, \forall \theta \in \Theta_1$$

则称之为水平 α 的无偏检验。

Definition 5.3.5 设 $\phi(x)$ 是上述假设的一个检验，如果其势函数 $E_\theta \phi(X)$ 满足

$$E_\theta \phi(X) = \alpha, \forall \theta \in \{\Theta_0 \text{ 与 } \Theta_1 \text{ 的公共边界}\}$$

则称之为边界相似检验。

Definition 5.3.6 对于上述假设，如果存在一个水平为 α 的无偏检验 $\phi(x)$ ，使得对于任何水平为 α 的无偏检验 $\phi'(x)$ ，均满足

$$E_\theta \phi(X) \geq E_\theta \phi'(X), \forall \theta \in \Theta_1$$

则称检验 $\phi(x)$ 是水平 α 的一致最优势无偏检验，记为 UMPUT.

Theorem 5.3.9 由上述定义，我们知道它们之间的关系为：

1. 如果无偏检验 $\phi(x)$ 的势函数 $E_\theta \phi(X)$ 是连续的，则它一定是相似的；
2. 水平为 α 的 UMP 一定是水平为 α 的无偏检验；
3. 如果上述检验问题的所有检验函数的势函数都是连续的，且一个水平为 α 的检验 $\phi(x)$ 是最优势的相似检验，则它必是水平 α 的 UMPUT；
4. MLR 分布族的单边假设的 UMP 是一致最优势的边界相似检验。

由此可以看出水平为 α 的检验类 Φ_α ，水平为 α 的无偏检验类 Φ_α^u ，水平为 α 的相似检验类 Φ_α^s 之间的关系为： $\Phi_\alpha^u \subset \Phi_\alpha$ ；相似检验不一定是水平 α 的，更不一定是无偏检验，但当势函数连续时， $\Phi_\alpha^u \subset \Phi_\alpha^s$

5.3.1 分布的 χ^2 拟合优度检验

Theorem 5.3.10 设 x_1, x_2, \dots, x_n 是来自总体 $F(x)$ 的样本，有时，需要检验的原假设是

$$H_0 : F(x) = F_0(x)$$

其中 $F_0(x)$ 称为理论分布，它可以是一个完全已知的分布，也可以是一个仅依赖于有限个实参数且分布形式已知的分布函数。这个分布检验问题就是检验观测数据是否与理论分布相符合。在样本容量较大时，这类问题可以用 χ^2 拟合优度检验来解决。这类问题可以

分以下两种情况来讨论。

Theorem 5.3.11 — 总体 X 为离散分布. 设总体 X 为取有限或可列个值 a_1, a_2, \dots 的离散随机变量, 我们把某些 a_i 合并为一类, 使得 a_1, a_2, \dots 被分为有限个类 A_1, A_2, \dots, A_r , 并满足样本观测值 x_1, x_2, \dots, x_n 落入每一个 A_i 内的个数 n_i 不小于 5. 记 $P(X \in A_i) = p_i (i = 1, 2, \dots, r)$ 那么,

假设 H_0 : 总体分布 $F(x) = F_0(x)$

就转化为如下

假设 H_0 : A_i 所占的比例为 $p_i (i = 1, 2, \dots, r)$

这样, 离散分布的拟合检验与前述分类数据的检验问题就完全一样了。

Theorem 5.3.12 — 总体 X 为连续分布. 设总体 X 为连续随机变量, 分布函数为 $F_0(x)$, : 选 $r - 1$ 个实数 $a_1 < a_2 < \dots < a_{r-1}$, 将实数族分为 r 个区间

$(-\infty, a_1], (a_1, a_2], \dots, (a_{r-1}, \infty)$

当观测值落入第 i 个区间时, 就把它看作属于第 i 类, 因此, 这 r 个区间就相当于 r 个类. 在 H_0 为真时, 记

$$p_i = P(a_{i-1} \leq X < a_i) = F_0(a_i) - F_0(a_{i-1}), i = 1, 2, \dots, r$$

其中 $a_0 = -\infty, a_r = \infty$, 以 n_i 表示样本的观测值 x_1, \dots, x_n 落入区间 $(a_{i-1}, a_i]$ 内的个数 ($i = 1, 2, \dots, r$), 接下来的做法就与总体只取有限个值的情况一样了。

Definition 5.3.7 — 列联表. (具有多个分类指标) 分析按两个或多个特征分类的频数数据, 这种数据通常称为**交叉分类数据**, 它们一般都以表格的形式给出, 称为**列联表**. 如, 如下二维列联表, 又称 2x2 表或四格表5.3。 $r \times c$ 的二维列联表: 总体按两个属性 A 与 B 分类, A 有 r 个类: A_1, \dots, A_r , B 有 c 个类: B_1, \dots, B_c , 共有 rc 个类, 若进行 n 次试验, 其中所属 A_i 又属 B_j 的结果有 n_{ij} 个, 按矩阵排列, 就得 $r \times c$ 二维列联表。

$A \setminus B$	1	\cdots	j	\cdots	c	和
1	n_{11}	\cdots	n_{1j}	\cdots	n_{1c}	$n_{1 \cdot}$
\vdots	\vdots		\vdots		\vdots	\vdots
i	n_{i1}	\cdots	n_{ij}	\cdots	n_{ic}	$n_{i \cdot}$
\vdots	\vdots		\vdots		\vdots	\vdots
r	n_{r1}	\cdots	n_{rj}	\cdots	n_{rc}	$n_{r \cdot}$
列和	$n_{\cdot 1}$	\cdots	$n_{\cdot j}$	\cdots	$n_{\cdot c}$	n

Table 5.3: 二位联列表

性别	视觉	
	正常	色盲
男	535	65
女	382	18

Definition 5.3.8 — 列联表的独立性检验. 列联表分析的基本问题是, 考察各属性之间有无关联, 即判别两属性是否独立. 在 $r \times c$ 表中, 若以 $p_{i\cdot}$, $p_{\cdot j}$ 和 p_{ij} 分别表示总体中的个体仅属 A_i , 仅属于 B_j 和同时属于 A_i 与 B_j 的概率, 可得一个二维离散分布表, 则“A、B 两属性独立”的假设可以表述为原假设

$$H_0 : P(A_i B_j) = p_{ij} = p_{i\cdot} p_{\cdot j} = P(A_i) P(B_j), i = 1, \dots, r, j = 1, \dots, c,$$

其意为: 属性 A 与 B 相互独立. 这种情况是 p_{ij} 不完全已知时的分布拟合检验. 这里诸 p_{ij} 共有 rc 个参数, 在原假设 H_0 成立时, 这 rc 个参数 p_{ij} 由 $r+c$ 个参数 $p_{1\cdot}, \dots, p_{r\cdot}$ 和 $p_{\cdot 1}, \dots, p_{\cdot c}$ 决定. 在这后 $r+c$ 个参数中存在两个约束条件: $\sum_{i=1}^r p_{i\cdot} = 1$, $\sum_{j=1}^c p_{\cdot j} = 1$, 所以, 此时 p_{ij} 实际上由 $r+c-2$ 个独立参数所确定. 在诸 p_{ij} 未知(常见)场合, 检验统计量是为

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - n \hat{p}_{ij})^2}{n \hat{p}_{ij}}$$

在原假设 H_0 成立时上式近似服从自由度为 $rc - (r+c-2) - 1 = (r-1)(c-1)$ 的 χ^2 分布. 其中诸 \hat{p}_{ij} 是在 H_0 成立下得到的 p_{ij} 的最大似然估计, 其表达式为

$$\hat{p}_{ij} = \hat{p}_{i\cdot} \hat{p}_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}$$

对给定显著性水平 α ($0 < \alpha < 1$) 在 n 较大场合该检验的拒绝域为

$$W = \{\chi^2 \geq x_{1-\alpha}^2((r-1)(c-1))\}$$

- Theorem 5.3.13**
1. 当分类指标为连续变量时, 我们仍然可以先用前面的离散化方法对变量离散后, 再用 χ^2 拟合优度检验对二者的独立性进行检验。
 2. 当带有未知参数时, Pearson 认为与参数已知的极限零分布相同, 而 Fisher 则从 $r = s = 2$ 的特殊列联表情况指出了此错误。
 3. 我们还可以把上述针对二维列联表的 χ^2 拟合优度检验推广到高维列联情况, 并且在对列联表数据进行分析时, 也要注意其维数的重要性。

Theorem 5.3.14 我们可以将上述结果推广到多维的情况. 下面考虑三维列联表的独立性检验问题. 假设一个观测包含三个指标: A, B, C , 且分别有 r, s, t 个水平, 且以 n_{ijk} 表示在 n 个样本中属于 $A_i \cap B_j \cap C_k$ 的样本个数. 此时感兴趣的假设可能有三类: 三个指标相互独立; 一个指标和另两个指标独立; 一个指标给定的条件下, 另两个指标独立。前两个类似于二维列联表的独立性检验问题, 故从略. 下面考虑第三类检验问题. 假设此时我们感兴趣的假设为 H_0 : 给定指标 C 下, 指标 A 与 B 独立. 基于二维统计量的导出方法, 我们可以得到此时的似然比统计量为:

$$\Lambda = \frac{\prod_{i=1}^r \prod_{j=1}^s \prod_{k=1}^t \left(\frac{n_{ijk} n_{\cdot jk}}{n n_{\cdot \cdot k}}\right)^{n_{ijk}}}{\prod_{i=1}^r \prod_{j=1}^s \prod_{k=1}^t \left(\frac{n_{ijk}}{n}\right)^{n_{ijk}}}$$

且当 H_0 成立及 $n \rightarrow \infty$ 时, 我们有

$$-2 \ln \Lambda \xrightarrow{\mathcal{L}} \chi^2(t(r-1)(s-1))$$

Theorem 5.3.15 — Kolmogorov 检验. 由于样本经验分布函数 $F_n(x)$ 是 $F(x)$ 的一个很好的估计, 故我们知道, 当 (6.2.5) 式的 H_0 成立时, $F_n(x)$ 与 $F_0(x)$ 应相差不大, 于是, 我们可以用统计量

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

来衡量 $F(x)$ 与 $F_0(x)$ 间的差别, 且拒绝域为 $\{D_n \geq c\}$. 此检验是由 Kolmogorov 于 1933 年提出的, 并给出了 D_n 的极限零分布(见定理 1.2.3), 于是, 我们称之为 Kolmogorov 检验. 对于 D_n 的精确分布, 请见下面的定理. 定理 6.4.1 如果 $F_0(x)$ 连续, 则当 H_0 成立时, 有

$$P\left\{D_n < \lambda + \frac{1}{2n}\right\} = \begin{cases} 0, & \lambda < 0 \\ \int_{\frac{1}{2n}-\lambda}^{\frac{1}{2n}+\lambda} \int_{\frac{2}{2n}-\lambda}^{\frac{3}{2n}+\lambda} \cdots \int_{\frac{2n-1}{2n}-\lambda}^{\frac{2n-1}{2n}+\lambda} f(\mathbf{x}) d\mathbf{x}, & 0 \leq \lambda < \frac{2n-1}{2n} \\ 1, & \lambda \geq \frac{2n-1}{2n} \end{cases}$$

$$\text{其中 } f(\mathbf{x}) = \begin{cases} n!, & 0 < x_1 < \cdots < x_n < 1 \\ 0, & \text{否则.} \end{cases}$$

Theorem 5.3.16 1. 当 F_0 为完全已知的连续分布时, Kolmogorov 检验优于 χ^2 拟合优度检验. 但如果 F_0 是离散的或含有未知参数, 则 Kolmogorov 检验无法应用.

2. 当 F_0 为含有未知参数的连续分布时, 如用样本去估计未知参数, 则定理 1.2.3 及定理 6.4.1 的结论均不成立, 故此时我们无法利用 Kolmogorov 检验, 但 χ^2 拟合优度检验是可以

3.

注 6.4.3 对手两样本分布假设检验问题, 即设 X_1, \dots, X_m 和 Y_1, \dots, Y_n 为分别来自总体 F 和 G 的 iid 样本, 且全样本独立, 则我们可以用统计量

$$D_{m,n} = \sup_x |F_n(x) - G_m(x)|$$

进行分布检验, 且拒绝域为 $\{D_{m,n} \geq c\}$, 其中 F_n, G_m 分别表示总体 X, Y 的经验分布函数. 手 Smirnov 给出了此统计量的极限分布, 故我们称之为 Smirnov 检验. 由手它是 Kolmogorov 检验的自然推广, 故有的书也统称之为 Kolmogorov-Smirnov 检验. 注 6.4.4 Kolmogorov 检验很难处理高维数据的分布检验, 而 χ^2 拟合优度检验则与一维的类似。

5.4 正态性检验

Theorem 5.4.1 当样本量 $n \leq 50$ 时, 用 W 检验; 当样本量 $n > 50$ 时, 用 D 检验. 接下去讨论的检验是

$$H_0 : F(x) \in \left\{ \Phi\left(\frac{x-\mu}{\sigma}\right) : \mu \in R, \sigma > 0 \right\}$$

Definition 5.4.1 — 正态性概率纸 (图) 检验. 正态概率图是一种特殊的坐标图, 其横坐标是等间隔的, 纵坐标是按标准正态分布函数值给出的, 关于纵坐标的划分则是不等间距的。具体步骤如下:

- 首先将样本观察值按从小到大的次序排列: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

2. 对每一个 i , 计算修正频率

$$\hat{F}_i = (i - 0.375)/(n + 0.25), i = 1, 2, \dots, n$$

将 \hat{F}_i 看作概率 $F(x_{(i)})$ 的估计

3. 将点 $(x_{(i)}, \hat{F}_i), i = 1, 2, \dots, n$ 逐一点在正态概率纸上

判断: 若诸点在一条直线附近, 则认为该样本来自正态总体若诸点明显不在一条直线附近, 则认为该样本不是来自正态分布总体。如果从正态概率纸上确认总体是非正态分布时, 可从如下变换

$$y = \ln x, \quad y = 1/x, \quad y = \sqrt{x}$$

中选一个数据作变换, 然后用 $(y_{(i)}, \hat{F}_i)$ 再描点, 判断变换后的数据是否来自正态分布

1. 若 $\ln x \sim N(\mu, \sigma^2)$, 则 $x \sim LN(\mu, \sigma^2)$,
2. 若 $\frac{1}{x} \sim N(\mu, \sigma^2)$, 则 $x \sim IN(\mu, \sigma^2)$ (倒正态分布),
3. 若 $\sqrt{x} \sim N(\mu, \sigma^2)$, 则 x 服从非中心 χ^2 分布 (更一般的 χ^2 分布)

“修正频率”: 对应第 i 个观测值 $x_{(i)}$ 的累积分布函数值 $F(x_{(i)}) = P(X \leq x_{(i)})$ 是一个概率, 可用频率作出估计, 即

$$\hat{F}(x_{(i)}) = \frac{\text{样本中小于等于 } x_{(i)} \text{ 的个数}}{\text{样本量}} = \frac{i}{n}$$

当 $i = n$ 时该频率为 1, 这意味着 x 的取值最大为 $x_{(n)}$, 不可能再超过 $x_{(n)}$, 这往往与实际不符, 对此需要修正. 常见的有如下两个修正频率

$$\hat{F}(x_{(i)}) = \frac{i}{n+1}, \quad \hat{F}(x_{(i)}) = \frac{i-3/8}{n+1/4}$$

这里是用的是第二个

Definition 5.4.2 — 夏皮洛-威尔克 (Shapiro - wilk) 检验. 适用于 $8 \leq n \leq 50$ 的情况。具体步骤如下

1. 首先将观察值按从小到大的次序排列: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
2. 从附表中查得对应 n 的系数 a_1, \dots, a_n , 其中 $a_{n+1-i} = -a_i, i = 1, 2, \dots, [n/2]$ 计算检验统计量

$$W = \frac{\left[\sum_{i=1}^{[n/2]} a_i (x_{(n+1-i)} - x_{(i)}) \right]^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2}$$

3. 拒绝域 $W = \{W \leq W_a\}$, 其中 W_a 可查表

Theorem 5.4.2 — 夏皮洛-威尔克 (Shapiro - wilk) 检验的具体推导. W 检验统计量是 n 个数对 $(x_{(i)}, a_i)$ 的相关系数的平方, 因此, 对 $x_{(i)}$ 或对 a_i 作线性变换, 该统计量值不变.

设 x_1, x_2, \dots, x_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 为布 $N(0, 1)$ 的次序样本, 显然, 有

$$x_{(i)} = \sigma u_{(i)} + \mu, \quad i = 1, 2, \dots, n$$

由于 $N(0,1)$ 不含任何未知参数, 故其次序统计量 $u_{(i)}$ 的前二阶矩可算出, 并记为

$$\begin{aligned} E(u_{(i)}) &= m_i, \quad i = 1, 2, \dots, n \\ \text{Cov}(u_{(i)}, u_{(j)}) &= v_{ij}, \quad i, j = 1, 2, \dots, n \\ m &= \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{pmatrix} = (m_1, m_2, \dots, m_n)' \\ V &= (v_{ij})_{n \times n} \end{aligned}$$

用 m_i 代替 $u_{(i)}$ 时会引起误差, 若记误差为 ε_i , 则可转化为

$$x_{(i)} = \sigma m_i + \mu + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (5.17)$$

其中 $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ 是均值为零, 协方差矩阵为 $\sigma^2 V$ 的 n 维随机向量. 作一个直角坐标系, 横轴表示 $x_{(i)}$, 纵轴表示 m_i , 根据 5.17 式, 在这个坐标系中, n 个点 $(x_{(1)}, m_1), (x_{(2)}, m_2), \dots, (x_{(n)}, m_n)$ 应该大致成一条直线, 微小的误差是由 ε_i 引起的.

怎样定量地衡量这些点接近直线的程度呢? 这正是 W 检验的出发点. 夏皮诺和威尔克首先研究了 $x = (x_{(1)}, x_{(2)}, \dots, x_{(n)})'$ 与 $m = (m_1, m_2, \dots, m_n)'$ 的相关系数

$$r^2 = \frac{\left[\sum_{i=1}^n (x_{(i)} - \bar{x})(m_i - \bar{m}) \right]^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2 \sum_{i=1}^n (m_i - \bar{m})^2} \quad (5.18)$$

显然, r^2 越是接近 1, $x = (x_{(1)}, x_{(2)}, \dots, x_{(n)})'$ 与 $m = (m_1, m_2, \dots, m_n)'$ 的线性关系越是明显. 所以, 若用 r^2 检验假设 H_0 : 总体分布为 $N(\mu, \sigma^2)$, 那么该检验的拒绝域具有形式 $\{r^2 \leq c\}$, 其中 c 为某个常数.

3. 由于 $m = (m_1, m_2, \dots, m_n)'$ 完全确定, 由标准正态分布的对称性可以证明 $m_i = -m_{n+1-i}$, 即

$$m_1 = -m_n, \quad m_2 = -m_{n-1}, \quad \dots, \quad m_{[n/2]} = -m_{n+1-[n/2]}$$

且当 n 为奇数时, $m_{(n+1)/2} = 0$. 由这一性质得

$$\sum_{i=1}^n m_i = 0, \quad \bar{m} = \frac{1}{n} \sum_{i=1}^n m_i = 0$$

4. 根据 3 中结论, 5.18 的分子部分可简化, 注意到

$$\sum_{i=1}^n (x_{(i)} - \bar{x})(m_i - \bar{m}) = \sum_{i=1}^n m_i x_{(i)} = \sum_{i=1}^{[n/2]} m_i (x_{(i)} - x_{(n-i+1)})$$

若记 $b_i = m_i / \sum m_j^2$, 则可验证: $\hat{\sigma}_1 = \sum_{i=1}^n b_i x_{(i)}$ 是正态标准差 σ 的线性无偏估计 (LUE), 若把 5.18 变为

$$r^2 = \frac{\left(\sum_{i=1}^{[n/2]} b_i (x_{(i)} - x_{(n-i+1)}) \right)^2}{\sum_{i=1}^n (x_{(i)} - \bar{x})^2} = \frac{\sum_{i=1}^n m_i^2}{n-1} \cdot \frac{\hat{\sigma}_1^2}{s^2} \quad (5.19)$$

若忽视与样本无关的常数, 则在正态性假设 H_0 为真时, 上式中分子与分母分别都是正态方差 σ^2 的无偏估计, 相差不会很大. 可当 H_0 为假时, 其差别就大了. 因为分子是在 H_0 为真下构造, 而分母 s^2 是通用的.

5. 为了在正态性假设 H_0 为假时能扩大分子与分母的差异, 更好区别非正态分布的能力, 夏皮诺与威尔克把5.19 的分子换用 σ 的最小方差线性无偏估计 (BLUE), 然后再正则化, 使最后得到的如下检验统计量 W 位于 0 与 1 之间。

$$W = \frac{\left(\sum_{i=1}^{[n/2]} a_i (x_{(i)} - x_{(n-i+1)})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

其中诸 a_i 如下得到:

$$C = (c_1, c_2, \dots, c_n)' = \frac{m'V^{-1}}{m'V^{-1}m} \quad a = (a_1, a_2, \dots, a_n)' = \frac{C'}{\sqrt{C'C}} = \frac{m'V^{-1}}{\sqrt{m'V^{-1}V^{-1}m}}$$

就是最常用的 W 检验的计算公式, 其中诸 a_i 可查附表。

Definition 5.4.3 — EP 检验即爱浚斯-普利 (Epps-Pulley) 检验. 爱海斯-普利检验对多种备择假设有效率, 其出发点是利用样本的特征函数与正态分布的特征函数的差的模的平方产生的一个加权积分得到的。EP 检验统计量定义为

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n} \sum_{i=2}^n \sum_{j=1}^{i-1} \exp \left\{ \frac{-(x_j - x_i)^2}{2s_*^2} \right\} - \sqrt{2} \sum_{i=1}^n \exp \left\{ \frac{-(x_i - \bar{x})^2}{4s_*^2} \right\}$$

其中 \bar{x}, s_*^2 就是前述的样本均值和 (除以 n 的) 样本方差。

其拒绝域为

$$\{T_{EP} \geq T_{1-\alpha, EP}(n)\}$$

当 $n > 200$ 时, 统计量 T_{EP} 的分位数可以用 $n = 200$ 时的分位数代替. 对小于 200 而不在表内的 n 可采用线性插值的方法得到近似的分位数。注意: 样本观察值的次序是随机的, 但一经选定后在计算 T_{EP} 中必须保持不变.

Theorem 5.4.3 — D 检验.

$$T = \sum_{i=1}^n \left(i - \frac{n+1}{2} \right) X_{(i)}$$

作为 σ 的一个线性无偏估计. 再考虑到 n 的阶, 我们用统计量

$$D_n = \frac{\sum_{i=1}^n \left(i - \frac{n+1}{2} \right) X_{(i)}}{n^{3/2} \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

作为正态性检验统计量, 且当它远离 1 时, 我们有理由拒绝正态性假设。

5.5 非参数检验

Definition 5.5.1 — 游程检验. 在怀疑数据可能不符合随机选取的原则时可对数据进行游程检验。以 0 为界的一连串的或以 1 为界的一连串的 0 称为一个游程。

设 x_1, \dots, x_n 为依时间顺序连续得到的一组样本观测值序列. 要考察如下原假设是否成立.

H_0 : 样本观察值序列符合随机抽取的原则

为此记样本中位数为 m , 把序列中小于 m 的那些 x_i 换成 0; 大于或等于 m 的那些 x_i 换成 1. 设序列中 0 和 1 的个数分别为 n_1 和 n_2 , 设 R 表示样本序列的总游程数, 过大或者过小都认为不是随机的, 则检验的拒绝域为

$$\{R \leq c_1\} \cup \{R \geq c_2\}$$

其中临界值 c_1 和 c_2 根据 H_0 为真时 R 的分布确定, 可查表。当 n_1, n_2 都不太大时, 游程检验用 p 值进行更加方便, 记 R 的观测值为 R_0 , 则

$$p = 2 \min \{P(R \leq R_0), P(R \geq R_0)\}$$

其中概率可用下述两式进行计算

$$P(R = 2k) = \frac{2 \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_1}}, \quad k = 1, 2, \dots, \left[\frac{n}{2} \right]$$

$$P(R = 2k+1) = \frac{\binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k} + \binom{n_1 - 1}{k} \binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_1}}, \quad k = 1, 2, \dots, \left[\frac{n-1}{2} \right]$$

对比较大的 n_1, n_2 , 可以使用渐进分布. 当样本随机地取自同一总体, n_1, n_2 都趋于无穷且 n_1/n_2 趋于常数 c 时, 有

$$\frac{R - \frac{2n_1}{1+c}}{\sqrt{\frac{4cn_1}{(1+c)^2}}} \xrightarrow{L} N(0, 1)$$

当 n_1, n_2 都比较大时, 上式中的 c 可用 n_1/n_2 代替, 从而对给定的显著性水平 α 两个临界值可近似取为

$$c_1 = \left[\frac{2n_1 n_2}{n_1 + n_2} \left(1 + \frac{u_{\alpha/2}}{\sqrt{n_1 + n_2}} \right) \right], c_2 = \left[\frac{2n_1 n_2}{n_1 + n_2} \left(1 + \frac{u_{1-\alpha/2}}{\sqrt{n_1 + n_2}} \right) \right] + 1$$

研究表明, 当 n_1, n_2 都大于 20 时, 上式近似效果是足够好的.

Proof. 上述两个公式的推导: 由于 0 的游程和 1 的游程是交替出现的, 所以, 当 $R = 2k$ 时, 必有 0 的游程和 1 的游程个数都是 k ; 当 $R = 2k+1$ 时, 要么 0 的游程个数是 k 且 1 的游程个数都是 $k+1$, 要么 1 的游程个数是 k 且 0 的游程个数是 $k+1$ 出现“ k 个 0 的游程”意味着 n_1 个 0 被分成 k 组, 这相当于将 n_1 个不可分辨的球放入 k 个盒中且没有一个盒子是空的, 依重复组合公式, 共有

$$\binom{(n_1 - k) + (k - 1)}{n_1 - k} = \binom{n_1 - 1}{k - 1}$$

种可能. 所以, 在 $R = 2k$ 时, k 个 0 的游程共有 $\binom{n_1 - 1}{k - 1}$ 种不同方式的安排, 类似地, k 个 1 的游程也共有 $\binom{n_2 - 1}{k - 1}$ 种不同方式的安排. 把 0 的游程和 1 的游程合在一起有两种可能, 一种是 0 的游程在前面而 1 的游程在后面, 另一种则反过来, 1 的游程在前面而 0 的游程在后面. 由此可知, 出现 $R = 2k$ 的情况一共有

$$2 \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 1}$$

种可能, 而总的等可能结果一共有 $\binom{n_1 + n_2}{n_1}$ 个, 从而有

$$P(R = 2k) = \frac{2 \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_1}}, \quad k = 1, 2, \dots, \left[\frac{n}{2} \right]$$

类似地, 在 $R = 2k + 1$ 时, 要么有 k 个 0 的游程和 $k + 1$ 个 1 的游程, 要么有 $k + 1$ 个 0 的游程和 k 个 1 的游程. 对前一种情况, 一定是以 1 的游程开始也以 1 的游程收尾, 共有 $\binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k}$ 种可能, 对后一种情况, 共有 $\binom{n_1 - 1}{k} \binom{n_2 - 1}{k - 1}$ 种可能, 于是有

$$P(R = 2k + 1) = \frac{\binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k} + \binom{n_1 - 1}{k} \binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_1}}, \quad k = 1, 2, \dots, \left[\frac{n - 1}{2} \right]$$

■

Theorem 5.5.1 — 游程检验还可以用于检验两个总体是否有相同分布. 设 x_1, \dots, x_n 是来自总体 X 的样本, y_1, \dots, y_m 是来自总体 Y 的样本, 将两组样本合并在一起, 按由小到大的顺序排列如下:

$$z_1 \leq z_2 \leq \dots \leq z_{n+m}$$

引入 w_i 如下: 若 z_i 是来自总体 X 的观察, 则 $w_i = 0$, 否则 (即 z_i 是来自总体 Y 的观察) $w_i = 1$. 这样, 我们就得到一个由 0 与 1 两个元素组成的序列

$$w_1, w_2, \dots, w_{n+m} \tag{5.20}$$

1. 在 X 与 Y 有相同的分布时, $x_1, \dots, x_n, y_1, \dots, y_m$ 可以看作是从同一个总体中抽取的样本, 因而序列 5.20 的总游程数 R 将是较大的.
2. 在 X 与 Y 的分布不相同时, 序列 5.20 的总游程数 R 将是较小的. 例如, 若总体 X 与 Y 分得很开, 以至于它们的样本观测值彼此不重叠时, R 的值接近于 2, 这时就有把握认为“两个分布不相同”.

在一般场合, 对于原假设 H_0 : 两个总体分布相同, 混合样本的游程数越小就趋向于拒绝原假设, 故检验的拒绝域应具有形式 $\{R \leq c\}$, 其中 c 为临界值, 它可以由 $P(R = 2k)$ 和 $P(R = 2k + 1)$ 两式算出, 当然也可以据此算出检验的 p 值。

Definition 5.5.2 — 符号检验. 符号检验主要用来对总体 p 分位数 x_p 进行检验. 设 x_1, x_2, \dots, x_n 是来自总体 X 的样本, 欲检验

$$H_0 : x_p \leq x_0, \quad \text{vs} \quad H_1 : x_p > x_0$$

其中 x_0 是某个给定常数, 记

$$y_i = \begin{cases} 1, & x_i > x_0 \\ 0, & x_i \leq x_0 \end{cases}$$

记 $\theta = P(y_i = 1) = P(x_i - x_0 > 0)$, 则 y_1, y_2, \dots, y_n 可以看作是来自二点分布 $b(1, \theta)$ 的样本, 从而 $S^+ = \sum_{i=1}^n y_i \sim b(n, \theta)$, 在 H_0 为真时, 有

$$\begin{aligned} \theta &= P(y_i = 1) = P(x_i - x_0 > 0) \\ &\leq P(x_i - x_p > 0) = 1 - P(x_i - x_p \leq 0) \\ &= 1 - F(x_p) = 1 - p \end{aligned}$$

反之, 若 $\theta \leq 1 - p$, 则 $\theta = P(x_i - x_0 > 0) \leq P(x_i - x_p > 0)$, 从而 $x_p \leq x_0$, 亦即 H_0 为真. 检验问题等价于检验问题

$$H_0 : \theta \leq 1 - p \quad \text{vs} \quad H_1 : \theta > 1 - p$$

检验问题是二项分布参数的检验问题, 符号检验统计量为 $S^+ = \sum_{i=1}^n y_i$, 拒绝域为

$$W = \{S^+ \geq c\}, c = \inf_k \left\{ k \mid \sum_{i=k}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \right\}$$

上述统计量 $S^+ = \sum_{i=1}^n y_i$ 亦称作符号统计量. 利用符号统计量所做的检验称作符号检验。

对符号检验, 使用检验的 p 值较为简便. 记 S_0^+ 为符号统计量的观测值, 即 S^+ 为 x_1, x_2, \dots, x_n 中取正数的个数。在原假设成立时, S^+ 的取值不应过大也不应过小. 在 H_0 为真时, S^+ 服从二项分布 $b(n, 0.5)$, 从而, 可确定常数则检验的 p 值为

$$p = P(S^+ \geq S_0^+) = \sum_{i=S_0^+}^n b(i; n, 1 - p)$$

其中 $b(i; n, 1 - p) = \binom{n}{i} (1 - p)^i p^{n-i}$ 表示二项分布的概率函数.

H_0	H_1	拒绝域形式	检验的 p 值
$x_p \geq x_0$	$x_p < x_0$	$W_I = \{S^+ \leq c\}$	$\sum_{i=0}^{S_0^+} b(i; n, 1 - p)$
$x_p = x_0$	$x_p \neq x_0$	$W_{III} = \{S^+ \leq c_1 \text{ 或 } S^+ \geq c_2\}$	$2 \min \left\{ \sum_{i=0}^{S_0^+} b(i; n, 1 - p), \sum_{i=S_0^+}^n b(i; n, 1 - p) \right\}$

Definition 5.5.3 符号检验的一个不足: 它只利用了观测值与中心位置之差的正负号, 而没有考虑到这些差的绝对值大小. 事实上, 这些差的绝对值大小度量了观测值既离中心的远近, 如果把两者结合起来, 自然可以期望检验效果更好, 秩和检验主要用于对对称分布的分布中心进行检验。

设 x_1, \dots, x_n 是来自连续分布 $F(x)$ 的简单随机样本, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 是其观测值的有序样本, 则观测值 x_i 在有序样本中的序号 r 称为 x_i 的秩, 记为 $R_i = r$. 注: 由于 $F(x)$ 是连续函数, 因此, R_i 不能唯一确定的概率为 0

设 x_1, \dots, x_n 是来自连续总体的样本, R_i 是 x_i 的秩, 则 $R = (R_1, \dots, R_n)$ 称为 (x_1, \dots, x_n) 的**秩统计量**. 由 R 导出的统计量, 也称为秩统计量. 基于秩统计量的检验方法称为秩检验.

Theorem 5.5.2 — 符号秩和检验. 设连续总体关于某个参数 θ 对称, 其分布函数记为 $F(x - \theta)$, 要检验的假设为:

$$H_0: \theta = 0 \quad \text{vs} \quad H_1: \theta \neq 0$$

设 x_1, \dots, x_n 是样本, 记 R_i 为 $|x_i|$ 在 $(|x_1|, \dots, |x_n|)$ 中的秩, 符号秩和统计量

$$W^+ = \sum_{i=1}^n R_i I(x_i > 0)$$

拒绝域为

$$\left\{ W^+ \leq W_{\alpha/2}^+(n) \right\} \cup \left\{ W^+ \geq W_{1-\alpha/2}^+(n) \right\}$$

附表给出了 $n \leq 50$ 时满足条件 $P(W^+ \leq W_\alpha^+(n)) \leq \alpha$ 的 $W_\alpha^+(n)$, 至于满足条件

$$P(W^+ \geq W_{1-\alpha}^+(n)) \leq \alpha \text{ 的 } W_{1-\alpha}^+(n)$$

它等于

$$\frac{1}{2}n(n+1) - W_\alpha^+(n)$$

当 $n > 50$ 时可采用正态近似

$$\frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{L} N(0, 1)$$

计算检验临界值。

$$E(W^+) = \frac{n(n+1)}{4}, \quad \text{Var}(W^+) = \frac{n(n+1)(2n+1)}{24}$$

R 称相同的几个变量值为一个“结”, 结外的秩是唯一的, 结内的秩通常取这些相继秩数的算术平均值作为各个变量的秩. 这样规定后, 11, 12, 12, 15 四个双测值的秩就唯一确定了, 它们依次是 1, 2.5, 2, 5, 4.

Definition 5.5.4 — 秩和检验. 秩和检验常用来对两个总体的位置进行比较: 设有两个总体, 其分布类型假定是一致的, 但分布的中心位置可能不同.

设有两个总体, 分布函数分别为 $F(x - \theta_1)$ 和 $F(x - \theta_2)$, x_1, \dots, x_m 和 y_1, \dots, y_n 分别为其样本, , 对应的样本 $x_1, \dots, x_m, y_1, \dots, y_n$ 进行排序, 对应的秩为

$$R = (Q_1, Q_2, \dots, Q_m, R_1, R_2, \dots, R_n)$$

检验统计量定义为

$$W = \sum_{i=1}^n R_i$$

此即 y_1, y_2, \dots, y_n 在混合样本中的秩的和, 通常称为 **Wilcoxon 秩和统计量**.

在比较 θ_1, θ_2 的大小时, 假设与拒绝域分别为

1.

$$H_0 : \theta_1 \leq \theta_2 \quad \text{vs} \quad H_1 : \theta_1 > \theta_2, W_1 = \{W \leq W_{\alpha}(m, n)\}$$

2.

$$H_0 : \theta_1 \geq \theta_2 \quad \text{vs} \quad H_1 : \theta_1 < \theta_2, W_n = \{W \geq W_{1-\alpha}(m, n)\}$$

3.

$$H_0 : \theta_1 = \theta_2 \quad \text{vs} \quad H_1 : \theta_1 \neq \theta_2, W_M = \{W \leq W_{\alpha/2}(m, n) \text{ 或 } W \geq W_{1-\alpha/2}(m, n)\}$$

在样本容量较大时, 可以用大样本近似公式

$$W^* = \frac{W - \frac{n(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0, 1)$$

作上述检验. 在 m 与 n 都大于等于 20 时, 此种近似效果很好.

1. 若数据中有结, 则取平均秩。
2. 若记 W_1 和 W_2 分别为两组样本 (x_1, \dots, x_m) 和 (y_1, \dots, y_n) 在混合样本中的秩和, 则

$$W_1 + W_2 = 1 + 2 + \dots + (m+n) = \frac{(m+n)(m+n+1)}{2}$$

是一个常数. 用 W_1 与用 W_2 作为检验的统计量是等价的. 为编表方便, 不失一般性, 假定 $m \geq n$. 此假定的含义是指使用观测值个数少的那组观测值对应的秩和作为检验统计量

3. 可以证明 W 的分布关于 $\frac{1}{2}n(m+n+1)$ 是对称的, 因此附表给出了满足条件 $P\{W \leq c\} \leq \alpha$ 的临界值 c , 如果需要满足条件 $P\{W \geq d\} \leq \alpha$ 的临界值 d , 它等于 $n(m+n+1) - c$
- 4.

$$E(W) = \frac{n(m+n+1)}{2}, \quad \text{Var}(W) = \frac{mn(m+n+1)}{12}$$

Definition 5.5.5 — 广义似然比检验.

6. 方差分析

Theorem 6.0.1 — 科大概统--回归分析。 1. 回归分析着重在寻求变量之间近似的函数关系，相关分析致力于寻求一些数量性的指标，以刻画有关变量之间关系深浅的程度。方差分析着重考虑一个或一些变量对一特定变量的影响有无及大小，由于其方法是基于样本方差的分解。

2. 中心化：

$$Y_i = \beta_0 + \beta_1 (X_i - \bar{X}) + e_i \quad (i = 1, \dots, n)$$

其关系是

$$\beta_1 = b_1, \quad \beta_0 = b_0 + b_1 \bar{X}$$

之后的结果都是基于中心化后的方程
 $\hat{\beta}_0 = \bar{Y}$

$$\begin{aligned} 3. \quad \hat{\beta}_1 &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n (X_i - \bar{X}) Y_i / \sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned}$$

4. 两个估计都是无偏的，并且时不相关的，这对于没有中心化的方程是不成立的
5. 残差的方差的无偏估计： $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \delta_i^2$
6. 当 e_i 服从正态分布 $N(0, \sigma^2)$ 时，有

$$\sum_{i=1}^n \delta_i^2 / \sigma^2 \sim \chi_{n-2}^2$$

自由度少两个是因为有两个 β 需要估计

7. 残差可以考察 e_1, \dots, e_n 独立同分布 $E(e_i) = 0, \text{Var}(e_i) = \sigma^2$ ($i = 1, \dots, n$) 假设是否正确
8. 多元线性回归的参数估计: $\hat{\beta} = (XX')^{-1}XY_{(n)}$

6.1 单因素方差分析

Definition 6.1.1 — 单因子方差分析的统计模型. 注意什么是因子, 什么是水平。考察了一个因子, 称其为单因子试验. 通常, 在单因子试验中, 记因子为 A , 设其有 r 个水平, 记为 A_1, A_2, \dots, A_r , 在每一水平下考察的指标可以看成一个总体, 现有 r 个水平, 故有 r 个总体, 假定:

1. 每一总体均为正态总体, 记为 $N(\mu_i, \sigma_i^2), i = 1, \dots, r$
2. 各总体的方差相同, 记为 $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_r^2 = \sigma^2$
3. 从每一总体中抽取的样本是相互独立的, 则所有的试验结果 y_{ij} 都相互独立。

上述假设可以通过正态性检验和方差齐性检验得到。

我们要做的工作是比较各水平下的均值是否相同, 即要对如下的一个假设进行检验,

$$H_0: \mu_1 = \mu_2 = \dots = \mu_r \quad (6.1)$$

其备择假设为

$$H_1: \mu_1, \mu_2, \dots, \mu_r \text{ 不全相等}$$

如果 H_0 成立, 因子 A 的 r 个水平均值相同, 称因子 A 的 r 个水平间没有显著差异, 简称 b; 反之, 当 H_0 不成立时, 因子 A 的 r 个水平均值不全相同, 这时称因子 A 的不同水平间有显著差异, 简称**因子 A 显著**。

Definition 6.1.2 方差分析是检验多个正态总体的均值比较的检验, 假设方差相同。检验方差的是方差齐性检验。

Theorem 6.1.1 为对假设6.1进行检验, 需要从每一水平下的总体抽取样本, 设从第 i 个水平下的总体获得 m 个试验结果 (设各水平下试验的重复数相同, 记 y_{ij} 表示第 i 个总体的第 j 次重复试验结果, 共得如下 $r \times m$ 个试验结果

$$y_{ij}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, m$$

其中 r 为水平数, m 为重复数, i 为水平编号, j 为重复序号. 在水平 A_i 下的试验结果 y_{ij} 与该水平下的指标均值 μ_i 一般总是有差距的, 记 $\varepsilon_{ij} = y_{ij} - \mu_i$, ε_{ij} 称为随机误差. 于是有试验结果 y_{ij} 的**数据结构式**

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

引入总均值与水平效应: 诸 μ_i 的平均 (所有试验结果的均值的平均)

$$\mu = \frac{1}{r} (\mu_1 + \dots + \mu_r) = \frac{1}{r} \sum_{i=1}^r \mu_i$$

为总均值, 也称一般平均. 称第 i 水平下的均值 μ_i 与总均值 μ 的差

$$a_i = \mu_i - \mu, \quad i = 1, 2, \dots, r$$

为因子 A 的第 i 水平的**主效应**, 简称为 A_i 的水平效应. 容易看出

$$\sum_{i=1}^r a_i = 0$$

$$\mu_i = \mu + a_i$$

这表明第 i 个总体均值是由总均值与该水平效应叠加而成的, 写出**单因子方差分析的统计模型**

$$\begin{cases} y_{ij} = \mu + a_i + \varepsilon_{ij}, & i = 1, 2, \dots, r, j = 1, 2, \dots, m, \\ \sum_{i=1}^r a_i = 0 \\ \varepsilon_{ij} \text{ 相互独立, 且都服从 } N(0, \sigma^2) \end{cases}$$

所有 ε_{ij} 可作为来自 $N(0, \sigma^2)$ 的一个样本, 在上述数据结构式下, $y_{ij} \sim N(\mu + a_i, \sigma^2)$. 要检验的假设检验可改写为

$$H_0 : a_1 = a_2 = \dots = a_r = 0 \quad \text{vs} \quad H_1 : a_1, \dots, a_r, \text{ 不全为 } 0$$

Definition 6.1.3 — 方差分析表.

$$T_i = \sum_{j=1}^m y_{ij}, \quad \bar{y}_{i \cdot} = \frac{T_i}{m}, \quad i = 1, 2, \dots, r$$

$$T = \sum_{i=1}^r T_i, \quad \bar{y} = \frac{T}{r \cdot m} = \frac{T}{n}$$

$n = r \cdot m =$ 总试验次数.

Definition 6.1.4 — 组内偏差与组间偏差. 数据 y_{ij} 与总平均 \bar{y} 间的偏差可用 $y_{ij} - \bar{y}$ 表示, 它可分解为两个偏差之和

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_{i \cdot}) + (\bar{y}_{i \cdot} - \bar{y})$$

记

$$\bar{\varepsilon}_{i \cdot} = \frac{1}{m} \sum_{j=1}^m \varepsilon_{ij}, \quad \bar{\varepsilon} = \frac{1}{r} \sum_{i=1}^r \bar{\varepsilon}_{i \cdot} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^m \varepsilon_{ij}$$

由于

$$y_{ij} - \bar{y}_{i \cdot} = (\mu_i + \varepsilon_{ij}) - (\mu_i + \bar{\varepsilon}_{i \cdot}) = \varepsilon_{ij} - \bar{\varepsilon}_{i \cdot}$$

所以 $y_{ij} - \bar{y}_{i \cdot}$ 仅反映组内数据与组内平均的随机误差, 称为**组内偏差**. 而

$$\bar{y}_{i \cdot} - \bar{y} = (\mu_i + \bar{\varepsilon}_{i \cdot}) - (\mu + \bar{\varepsilon}) = a_i + \bar{\varepsilon}_{i \cdot} - \bar{\varepsilon}$$

$\bar{y}_{i \cdot} - \bar{y}$ 除了反映随机误差外, 还反映了第 i 个水平效应, 称为**组间偏差**.

Definition 6.1.5 — 偏差平方和及其自由度. k 个数据的偏差平方和, 有时简称平方和

$$Q = (y_1 - \bar{y})^2 + \cdots + (y_k - \bar{y})^2 = \sum_{i=1}^k (y_i - \bar{y})^2$$

. 偏差平方和常用来度量若干个数据分散的程度, 它是用来度量若干个数据间差异 (即波动) 的大小的一个重要的统计量。在构成偏差平方和 Q 的 k 个偏差 $y_1 - \bar{y}, \dots, y_k - \bar{y}$ 间有一个恒等式

$$\sum_{i=1}^k (y_i - \bar{y}) = 0$$

这说明在 Q 中独立的偏差只有 $k - 1$ 个. 在统计学中把平方和中独立偏差个数称为该平方和的自由度, 常记为 f , 如 Q 的自由度为 $f_Q = k - 1$. 自由度是偏差平方和的一个重要参数。

Theorem 6.1.2 — 平方和分解式.

$$S_r = S_A + S_e, \quad f_r = f_A + f_e$$

若记

$$\bar{y}_{ij} = \frac{1}{m} \sum_{j=1}^m y_{ij}, \quad \bar{y} = \frac{1}{rm} \sum_{i=1}^r \sum_{j=1}^m y_{ij} = \frac{1}{r} \sum_{i=1}^r \bar{y}_{i\cdot}$$

上述诸平方和分别为

$$S_r = \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y})^2$$

称为总平方和, 其自由度 $f_r = n - 1$

$$S_A = m \sum_{i=1}^r (\bar{y}_{i\cdot} - \bar{y})^2$$

由于组间差异数除了随机误差外, 还反映了效应间的差异, 故由效应不同引起的数据差异可用组间偏差平方和表示, 也称为因子 A 的偏差平方和, 其自由度 $f_A = r - 1$

$$S_e = \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y}_{i\cdot})^2$$

称为组内平方和或误差平方和, 其自由度 $f_e = n - r$

(R) 数据 y_{ij} 的平移 $y' = y_{ij} - a$ 不会改变其平方和的值. 用此性质可简化计算.

Proof.

$$\sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}) = \sum_{i=1}^r \left[(\bar{y}_{i\cdot} - \bar{y}) \sum_{j=1}^m (y_{ij} - \bar{y}_{i\cdot}) \right] = 0$$

故有

$$\begin{aligned} S_T &= \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y})^2 = \sum_{i=1}^r \sum_{j=1}^m [(y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y})]^2 \\ &= S_e + S_A + 2 \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}) = S_e + S_A \end{aligned}$$

自由度之间也有这种关系 ■

Definition 6.1.6 — 均方. 引入了均方, 它定义为

$$MS = \frac{Q}{f_Q}$$

其意为平均每个自由度上有多少平方和。均方可以消除不同的自由度之间的影响。

Theorem 6.1.3 运用抽样分布基本定理和单因素方差分析的统计模型可得到

1.

$$S_e / \sigma^2 \sim \chi^2(n - r)$$

从而 $E(S_e) = (n - r)\sigma^2$

2.

$$E(S_A) = (r - 1)\sigma^2 + m \sum_{i=1}^r a_i^2$$

进一步, 若 H_0 成立, 则有

$$S_A / \sigma^2 \sim \chi^2(r - 1)$$

3. S_A 与 S_e 独立.

Proof.

$$S_e = \sum_{i=1}^r \sum_{j=1}^m (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2$$

诸 $\varepsilon_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, m$ 独立同分布于 $N(0, \sigma^2)$, 由定理 3.4.2 知

$$\frac{1}{\sigma^2} \sum_{j=1}^m (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2, i = 1, \dots, r$$

相互独立, 其共同分布为 $\chi^2(m - 1)$, 由 χ^2 分布的可加性, 有

$$\frac{S_e}{\sigma^2} \sim \chi^2(n - r)$$

这给出 $E(S_e / \sigma^2) = n - r = f_e$. 类似地,

$$S_A = m \sum_{i=1}^r (a_i + \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon})^2$$

Table 6.1: 单因子方差分析

来源	平方和	自由度	均方	F 比	p 值
因子	S_A	$f_A = r - 1$	$MS_A = S_A/f_A$	$F = MS_A/MS_e$	p
误差	S_e	$f_e = n - r$	$MS_e = S_e/f_e$		
总和	S_r	$f_r = n - 1$			

由定理 3.4.2 知, 对每个 i , 平方和

$$\sum_{j=1}^m (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2$$

与均值 $\bar{\varepsilon}_{i\cdot}$ 独立, 从而 $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \dots, \bar{\varepsilon}_{r\cdot}$ 与 S_e 独立, 而 S_A 只是 $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \dots, \bar{\varepsilon}_{r\cdot}$ 的函数, 由此 (3) 得证.

S_A 的期望是

$$E(S_A) = m \sum_{i=1}^r a_i^2 + E \left[m \sum_{i=1}^r (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon})^2 \right]$$

由于诸误差均值 $\bar{\varepsilon}_{1\cdot}, \bar{\varepsilon}_{2\cdot}, \dots, \bar{\varepsilon}_{r\cdot}$ 独立同分布于 $N(0, \sigma^2/m)$, 故

$$\frac{1}{\sigma^2} \sum_{i=1}^r m(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon})^2 \sim \chi^2(r-1)$$

于是, $E \left[\sum_{i=1}^r m(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon})^2 \right] = (r-1)\sigma^2$, 在 H_0 成立下, $S_A/\sigma^2 \sim \chi^2(r-1)$

■

Theorem 6.1.4 — 假设检验. 在 H_0 成立下

$$F = MS_A/MS_e \sim F(f_A, f_e)$$

对给定的显著性水平 α ($0 < \alpha < 1$) 其拒绝域为

$$W = \{F \geq F_{1-\alpha}(f_A, f_e)\}$$

其中 $F_{1-\alpha}(f_A, f_e)$

1. 若 $F \geq F_{1-\alpha}(f_A, f_e)$, 则认为因子 A 显著, 即诸正态均值间有显著差异
2. 若 $F < F_{1-\alpha}(f_A, f_e)$, 则说明因子 A 不显著, 即接受原假设 H_0

检验的 p 值: 若以 X 记服从 $F(f_A, f_e)$ 的随机变量, F 为统计量, $F = MS_A/MS_e$, 的观测值为 F_0 , 则 $p = P(X \geq F_0)$

给出常用计算公式

$$S_T = \sum_{i=1}^r \sum_{j=1}^m y_{ij}^2 - \frac{T^2}{n}$$

$$S_A = \frac{1}{m} \sum_{i=1}^r T_i^2 - \frac{T^2}{n}$$

$$S_e = S_T - S_A$$

- Theorem 6.1.5 — 点估计.**
1. 总均值 μ 的估计 $\hat{\mu} = \bar{y}$
 2. 水平均值 μ_i 的估计 $\hat{\mu}_i = \bar{y}_i, i = 1, 2, \dots, r$
 3. 主效应 a_i 的估计 $\hat{a}_i = \bar{y}_i - \bar{y}, i = 1, 2, \dots, r$
 4. 误差方差 σ^2 的估计 $\hat{\sigma}^2 = MS_e = S_e/f_e$
 5. μ_i 的 $1 - \alpha$ 置信区间为 $\bar{y}_i \pm \hat{\sigma} \cdot t_{1-\alpha/2}(f_e) / \sqrt{m}$

Proof. 推导：诸 y_{ij} 相互独立，且 $y_{ij} \sim N(\mu + a_i, \sigma^2)$ ，因此，可使用最大似然方法求出总均值 μ ，各水平效应 a_i 和误差方差 σ^2 的估计。首先，写出似然函数

$$L(\mu, a_1, \dots, a_r, \sigma^2) = \prod_{i=1}^r \prod_{j=1}^m \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_{ij} - \mu - a_i)^2}{2\sigma^2} \right\} \right\}$$

其对数似然函数为

$$l(\mu, a_1, \dots, a_r, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{l=1}^r \sum_{j=1}^m (y_{lj} - \mu - a_l)^2$$

求偏导，得似然方程为求偏导，得似然方程为

$$\begin{cases} \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \mu - a_i) = 0 \\ \frac{\partial l}{\partial a_i} = \frac{1}{\sigma^2} \sum_{j=1}^m (y_{ij} - \mu - a_i) = 0, \quad i = 1, \dots, r \\ \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \mu - a_i)^2 = 0 \end{cases}$$

考虑到约束条件总效应为 0，可求出前述各参数的最大似然估计为

$$\begin{aligned} \hat{\mu} &= \bar{y} \\ \hat{a}_i &= \bar{y}_i - \bar{y}, \quad i = 1, \dots, r \\ \hat{\sigma}_M^2 &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2 = \frac{S_e}{n} \end{aligned}$$

由最大似然估计的不变性，各水平均值 μ_i 的最大似然估计为

$$\hat{\mu}_i = \bar{y}_i.$$

由于 $\hat{\sigma}_n^2$ 不是 σ^2 的无偏估计，实用中通常采用如下误差方差的无偏估计

$$\hat{\sigma}^2 = MS_e$$

讨论各水平均值 μ_i 的置信区间。知 $\bar{y}_i \sim N(\mu_i, \sigma^2/m)$, $S_e/\sigma^2 \sim \chi^2(f_e)$ ，且两者独立，故

$$\frac{\sqrt{m}(\bar{y}_i - \mu_i)}{\sqrt{S_e/f_e}} \sim t(f_e)$$

由此给出 A_i 的水平均值 μ_i 的 $1 - \alpha$ 的置信区间为

$$\bar{y}_i \pm \hat{\sigma} \cdot t_{1+\alpha/2}(f_e) / \sqrt{m}$$

$\hat{\sigma}$ 由上面得到



Theorem 6.1.6 单因子试验的统计分析可得如下三个结果：

1. 因子 A 是否显著
2. 试验误差方差 σ^2 的估计
3. 诸水平均值 μ_i 的点估计与区间估计 (此项在因子 A 不显著时无需进行)

Theorem 6.1.7 重复数不相等的情形下的方差分析

1. 设从第 i 个水平下的总体获得 m_i 个试验结果, 记为 $y_{i1}, y_{i2}, \dots, y_{im_i}$, $i = 1, 2, \dots, r$, 故总试验次数为 $n = m_1 + m_2 + \dots + m_r$, 从而, 其统计模型为

$$\begin{cases} y_{ij} = \mu_i + \varepsilon_{ij}, & i = 1, 2, \dots, r, \quad j = 1, 2, \dots, m_i \\ \text{各 } \varepsilon_{ij} \text{ 相互独立, 且都服从 } N(0, \sigma^2) \end{cases}$$

2. 数据结构式及其参数估计基本同前, 但要注意以下两点:
 - (a) 总均值

$$\mu = \frac{1}{n} \sum_{i=1}^r m_i \mu_i$$

- (b) 主效应的约束条件为

$$\sum_{i=1}^r m_i a_i = 0$$

- (c) 基本假定、平方和分解、方差分析及判断准则都和前面一样, 只是因子 A 的平方和 S_A 的计算公式略有不同: 记 $n = \sum_{i=1}^r m_i$, 则

$$S_A = \frac{T_1^2}{m_1} + \frac{T_2^2}{m_2} + \dots + \frac{T_r^2}{m_r} - \frac{T^2}{n}$$

6.2 多重比较

Theorem 6.2.1 — 水平均值差的置信区间. 在方差分析中, 如果经过 F 检验拒绝原假设, 表明因子 A 是显着的, 即水平对应的水平均值不全相等, 此时, 我们还需要进一步确认哪些水平均值间是确有差异的, 哪些水平均值间无显著差异。指定的一对水平 A_i 与 A_j :

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \sim N \left(\mu_i - \mu_j, \left(\frac{1}{m_i} + \frac{1}{m_j} \right) \sigma^2 \right)$$

有 $S_e / \sigma^2 \sim \chi^2(f_e)$, 且两者独立, 故

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\sqrt{\left(\frac{1}{m_i} + \frac{1}{m_j} \right) \frac{S_e}{f_e}}} \sim t(f_e)$$

由此给出 $\mu_i - \mu_j$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm \sqrt{\left(\frac{1}{m_i} + \frac{1}{m_j} \right) \hat{\sigma}^2} \cdot t_{1-\frac{\alpha}{2}}(f_e)$$

其中 $\hat{\sigma}^2 = S_e/f_e$ 是 σ^2 的无偏估计.

Definition 6.2.1 对每一组 (i, j) 式给出的区间的置信水平都是 $1 - \alpha$, 但对多个这样的区间, 要求其同时成立, 其同时发生的概率

$$P\left(\bigcap_{i=1}^k E_i\right) = 1 - P\left(\bigcup_{i=1}^k \bar{E}_i\right) \geqslant 1 - \sum_{i=1}^k P(\bar{E}_i) = 1 - k\alpha$$

这说明它们同时发生的概率可能比 $1 - \alpha$ 小很多. 为了使它们同时发生的概率不低于 $1 - \alpha$, 一个办法是把每个事件发生的概率提高到 $1 - \alpha/k$. 比如, 如果我们同时考虑所有的 $k = r(r-1)/2$ 组水平均值差 $\mu_i - \mu_j$ 的置信区间, 则将 $t_{1-\alpha/2}(f_e)$ 替换为 $t_{1-\alpha/(2k)}(f_e)$ 即可. 这将导致每个置信区间过长, 联合置信区间的精度很差, 引入多重比较。

在 $r(r > 2)$ 个水平均值中同时比较任意两个水平均值间有无明显差异的问题称为**多重比较**, 多重比较即要以显著性水平 α 同时检验如下 $r(r-1)/2$ 个假设:

$$H_0^{ij}: \mu_i = \mu_j, \quad 1 \leq i < j \leq r$$

直观地看, 当 H_0^{ij} 成立时, $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|$ 不应过大, 故在同时考察 $\binom{r}{2}$ 个假设 H_0^{ij} 时, 诸 H_0^{ij} 中至少有一个不成立就构成多重比较的拒绝域. 故

$$W = \bigcup_{1 \leq i < j \leq r} \{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| \geq c_{ij}\}$$

诸临界值 c_{ij} 应在原假设成立时由 $P(W) = \alpha$ 确定. 下面分重复数相等和不等分别介绍临界值的确定。

Theorem 6.2.2 — 重复数相等场合的 T 法. 在重复数相等时, 由对称性有 c_{ij} 相等, 记为 c . 记 $\hat{\sigma}^2 = S_e/f_e$, 则

$$t_i = \frac{\bar{y}_{i\cdot} - \mu_i}{\hat{\sigma}/\sqrt{m}} \sim t(f_e)$$

于是当原假设式成立时, $\mu_1 = \dots = \mu_r = \mu$, 故有

$$\begin{aligned} P(W) &= P\left(\bigcup_{1 \leq i < j \leq r} \{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| \geq c\}\right) \\ &= 1 - P\left(\bigcap_{1 \leq i < j \leq r} \{|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| < c\}\right) \\ &= 1 - P\left(\max_{1 \leq i < j \leq r} |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| < c\right) \\ &= P\left(\max_{1 \leq i < j \leq r} |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| \geq c\right) \\ &= P\left(\max_{1 \leq i < j \leq r} \left| \frac{(\bar{y}_{i\cdot} - \mu) - (\bar{y}_{j\cdot} - \mu)}{\hat{\sigma}/\sqrt{m}} \right| \geq \frac{c}{\hat{\sigma}/\sqrt{m}}\right) \\ &= P\left(\max_i \frac{(\bar{y}_{i\cdot} - \mu)}{\hat{\sigma}/\sqrt{m}} - \min_j \frac{(\bar{y}_{j\cdot} - \mu)}{\hat{\sigma}/\sqrt{m}} \geq \frac{c}{\hat{\sigma}/\sqrt{m}}\right) \end{aligned}$$

这里 $q(r, f_e) = \max_i \frac{(\bar{y}_i - \mu)}{\hat{\sigma}/\sqrt{m}} - \min_j \frac{(\bar{y}_j - \mu)}{\hat{\sigma}/\sqrt{m}}$ 称为 t 化极差统计量, $q(r, f_e)$ 的分布不易导出, 但知它的分布只与自由度 f_e 和水平数 r 有关, 而与参数 μ, σ^2 无关, 也与 m 无关, 该分布可由随机模拟方法得到, 方法如下(不妨设 $\mu = 0, \sigma^2 = 1, m = 1$): 对给定的 r 和 f_e

1. 从标准正态分布 $N(0, 1)$ 产生 r 个随机数 x_1, \dots, x_r , 将该 r 个随机数按从小到大排序得到 $x_{(1)}$ 和 $x_{(r)}$
2. 从自由度为 f_e 的 χ^2 分布 $\chi^2(f_e)$ 产生一个随机数 y
3. 计算 $q = (x_{(r)} - x_{(1)}) / \sqrt{y}$
4. 重复(1)到(3) N 即得 $q(r, f_e)$ 的 N 个观测值. 由此可获得 $q(r, f_e)$ 的各种分位数. 于是

$$P(W) = P(q(r, f_e) \geq \sqrt{mc}/\hat{\sigma}) = \alpha$$

可以得出

$$c = q_{1-\alpha}(r, f_e) \hat{\sigma} / \sqrt{m}$$

其中 $q_{1-\alpha}(r, f_e)$ 表示 $q(r, f_e)$ 的 $1 - \alpha$ 分位数. 至此, 可将重复数相同时多重比较的步渠总结如下: 对给定的显著性水平 α , 查多重比较的分位数 $q_{1-\alpha}(r, f)$ 表, 计算

$$c = q_{1-\alpha}(r, f_e) \hat{\sigma} / \sqrt{m}$$

比较诸 $|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|$ 与 c 的大小, 若与 c 的大小, 若

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| \geq c$$

则认为水平 A_i 与水平 A_j 间有显著差异, 反之, 则认为水平 A_i 与水平 A_j 间无明显差别. 这一方法最早由图基(Turkey)提出, 因此称为 T 法.

Theorem 6.2.3 — 重复数不等场合的 S 法. 在重复数不等时, 沿用上面的记号, 我们有

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}} \hat{\sigma}} \sim t(f_e)$$

在假设原假设成立时, $\mu_1 = \dots = \mu_r = \mu$, 于是有

$$t_{ij} = \frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot})}{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}} \hat{\sigma}} \sim t(f_e) \quad \text{或} \quad F_{ij} = \frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot})^2}{\left(\frac{1}{m_i} + \frac{1}{m_j}\right) \hat{\sigma}^2} \sim F(1, f_e)$$

从而可以要求 $c_{ij} = c \sqrt{\frac{1}{m_i} + \frac{1}{m_j}}$, 类似于重复数相等时的推导, 有

$$\begin{aligned} P(W) &= P \left[\bigcup_{1 \leq i < j \leq r} \left(|\bar{y}_{i \cdot} - \bar{y}_{j \cdot}| \geq c \sqrt{\frac{1}{m_i} + \frac{1}{m_j}} \right) \right] \\ &= P \left(\max_{1 \leq i < j \leq r} \frac{|\bar{y}_{i \cdot} - \bar{y}_{j \cdot}|}{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}} \hat{\sigma}} \geq \frac{c}{\hat{\sigma}} \right) \\ &= P \left(\max_{1 \leq i < j \leq r} \frac{(\bar{y}_{i \cdot} - \bar{y}_{j \cdot})^2}{\left(\frac{1}{m_i} + \frac{1}{m_j}\right) \hat{\sigma}^2} \geq \frac{c^2}{\hat{\sigma}^2} \right) \\ &= P \left(\max_{1 \leq i < j \leq r} F_{ij} \geq (c/\hat{\sigma})^2 \right) \end{aligned}$$

可以证明

$$\frac{\max_{1 \leq i < j \leq r} F_{ij}}{r-1} \sim F(r-1, f_e)$$

从而由 $P(W) = \alpha$ 可推出 $(\frac{c}{\hat{\sigma}})^2 = (r-1)F_{1-\alpha}(r-1, f_e)$, 亦即

$$c_{ij} = \sqrt{(r-1)F_{1-\alpha}(r-1, f_e) \left(\frac{1}{m_i} + \frac{1}{m_j}\right) \hat{\sigma}^2}$$

6.3 方差齐性检验

Definition 6.3.1 — 方差齐性检验. 方差齐性即诸方差相等, 是方差分析的基本假定之一, 方差齐性检验就是检验这个假定是否成立. 方差相等成为**方差齐性**. 该检验问题的一对假设为

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 \quad \text{vs} \quad H_1: \text{诸} \sigma_i^2 \text{ 不全相等.}$$

Theorem 6.3.1 — Hartley 检验. (在重复数相等场合使用) 在重复次数均为 m 时, 采用 Hartley 检验, 检验统计量是

$$H = \frac{\max \{s_1^2, s_2^2, \dots, s_r^2\}}{\min \{s_1^2, s_2^2, \dots, s_r^2\}}$$

其中 s_i^2 是第 i 个水平 A_i 下重复试验数据的样本方差. 直观上看, 当 H_0 成立, 即诸方差相等时, H 的值应接近于 1, 当 H 的值较大时, 诸方差间的差异就大, H 愈大, 诸方差间的差异就愈大, 这时应拒绝 H_0 拒绝域为

$$W = \{H > H_{1-\alpha}(r, f)\}$$

其中 α 为显著性水平, $f = m-1$, $H_{1-\alpha}(r, f)$ 是统计量 H 的分布的 $1-\alpha$ 分位数

Theorem 6.3.2 — Bartlett 检验. (可在重复数不等场合使用, 但样本量不得低于 5) 在重复

次数不等且每个水平下试验次数均不低于 5 时, 可采用 Bartlett 检验, 检验统计量为

$$B = \frac{1}{C} \left[f_e \ln MS_e - \sum_{i=1}^r f_i \ln s_i^2 \right]$$

其中诸 s_i^2 同上, $f_i = m_i - 1$ 为 s_i^2 的自由度, $MS_e = \frac{1}{f_e} \sum_{i=1}^r f_i s_i^2$, $f_e = \sum_{i=1}^r f_i$ 为误差方差的自由度. $C = 1 + \frac{1}{3(r-1)} \left(\sum_{i=1}^r \frac{1}{f_i} - \frac{1}{f_e} \right)$, 拒绝域为

$$W = \{B > \chi_{1-\alpha}^2(r-1)\}$$

其中 $\chi_{1-\alpha}^2(r-1)$ 是自由度为 $r-1$ 的 χ^2 分布的 $1-\alpha$ 分位数.

Proof. 在单因子方差分析中有 r 个样本, 设第 i 个样本方差为

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i\cdot})^2 = \frac{Q_i}{f_i}, \quad i = 1, 2, \dots, r$$

其中 m_i 为第 i 个样本的容量 (即试验重复次数), $Q_i = \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$ 与 $f_i = m_i - 1$ 为该样本的偏差平方和及自由度. 由于误差均方

$$MS_e = \frac{1}{f_e} \sum_{i=1}^r Q_i = \sum_{i=1}^r \frac{f_i}{f_e} s_i^2$$

它是 r 个样本方差 $s_1^2, s_2^2, \dots, s_r^2$ 的 (加权) 算术平均数. 而相应的 r 个样本方差的几何平均数记为 GMS_e , 它是

$$GMS_e = \left[(s_1^2)^{f_1} (s_2^2)^{f_2} \cdots (s_r^2)^{f_r} \right]^{1/f_e}$$

其中 $f_e = f_1 + f_2 + \cdots + f_r = \sum_{i=1}^r (m_i - 1) = n - r$ 由于几何平均数总不会超过算术平均数, 故有

$$GMS_e \leq MS_e$$

其中等号成立当且仅当诸 s_i^2 彼此相等, 若诸 s_i^2 间的差异愈大, 则此两个平均值相差也愈大. 由此可见, 当诸总体方差相等时, 其样本方差间不应相差较大, 从而比值 MS_e/GMS_e 接近于 1. 反之, 在比值 MS_e/GMS_e 较大时, 就意味着诸样本方差差异较大, 从而反映诸总体方差差异也较大. 这个结论对此比值的对数也成立. 从而检验表示的一对假设的拒绝域应是

$$W = \{\ln(MS_e/GMS_e) > d\}$$

Bartlett 证明了: 在大样本场合, $\ln(MS_e/GMS_e)$ 的某个函数近似服从自由度为 $r-1$ 的 χ^2 分布. 具体是

$$B = \frac{f_e}{C} (\ln MS_e - \ln GMS_e) \sim \chi^2(r-1)$$

其中

$$C = 1 + \frac{1}{3(r-1)} \left[\sum_{i=1}^r \frac{1}{f_i} - \frac{1}{f_e} \right]$$

且 C 通常会大于 1 根据上述结论, 可取

$$B = \frac{1}{C} \left[f_e \ln MS_e - \sum_{i=1}^r f_i \ln s_i^2 \right]$$

作为检验统计量, 对给定的显著性水平 α , 检验的拒绝域为

$$W = \{B \geq \chi^2_{1-\alpha}(r-1)\}$$

考虑到这里 χ^2 分布是近似分布, 在诸样本量 m_i 均不小于 5 时使用上述检验是适当的. ■

Theorem 6.3.3 — 修正的 Bartlett 检验. (在样本量相等或不等, 样本量较小或较大均可使用) 一般场合, 可采用修正的 Bartlett 检验, 检验统计量

$$B' = \frac{f_2 BC}{f_1(A - BC)}$$

其中 B 与 C 同前 $f_1 = r - 1, f_2 = \frac{r+1}{(C-1)^2}, A = \frac{f_2}{2-C+2/f_2}$, 在原假设成立下, Box 还证明了统计量 B' 的近似分布是 F 分布 $F(f_1, f_2)$, 对给定的显著性水平 α , 拒绝域为

$$W = \{B' > F_{1-\alpha}(f_1, f_2)\}$$

其中 $F_{1-\alpha}(f_1, f_2)$ 是相应 F 分布的 $1 - \alpha$ 分位数, f_2 的值可能不是整数, 这时可通过对 F 分布的分位数表施行线性内插法获得。

6.4 一元线性回归

Definition 6.4.1 — 一元线性回归. 考察两个变量 x 与 y 之间是否存在线性相关关系, 其中 x 是一般 (可控) 变量, y 是随机变量, 其线性相关关系可表示如下 (可用散点图显示):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

其中 β_0 为截距, β_1 为斜率, ε 为随机误差, 常假设 $\varepsilon \sim N(0, \sigma^2)$. 这里 $\beta_0, \beta_1, \sigma^2$ 是三个待估参数. 上式表明, y 与 x 之间有线性关系, 但受到随机误差的干扰。

对 x 与 y 通过试验或观察可得 n 对数据 (注: 数据是成对的, 不允许错位). 在 y 与 x 之间存在线性关系的假设下, 有如下统计模型:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n \\ \text{各 } \varepsilon_i \text{ 独立同分布, 其分布为 } N(0, \sigma^2) \end{cases} \quad (6.2)$$

利用成对数据可获得 β_0 与 β_1 的估计, 设估计分别为 $\hat{\beta}_0$ 与 $\hat{\beta}_1$, 则称

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

为 **回归方程**, 其图形称为回归直线. 给定 $x = x_0$ 后, 称 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 为回归值 (在不同场合也称其为拟合值),

Theorem 6.4.1 — 参数估计. 用最小二乘法可得 β_0 与 β_1 的无偏估计

$$\begin{cases} \hat{\beta}_1 = l_{xy}/l_{xx} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

其中 $\bar{x} = \frac{1}{n} \sum x_i, \bar{y} = \frac{1}{n} \sum y_i$ (此处 Σ 表示 $\sum_{i=1}^n$, 下同)

$$\begin{aligned} l_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x} \cdot \bar{y} = \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \\ l_{xx} &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \\ l_{yy} &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = \sum y_i^2 - \frac{1}{n} (\sum y_i)^2 \end{aligned}$$

Theorem 6.4.2 在模型6.2下, 有

1. $\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}}\right) \sigma^2\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right)$
2. $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}} \sigma^2$
3. 对给定的 $x_0, \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right) \sigma^2\right)$

Proof. 利用 $\sum (x_i - \bar{x}) = 0$, 可把 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 改写为

$$\hat{\beta}_1 = \frac{l_{xy}}{l_{xx}} = \sum \frac{x_i - \bar{x}}{l_{xx}} y_i \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{l_{xx}} \right] y_i$$

它们是独立正态变量 y_1, y_2, \dots, y_n 的线性组合, 故都服从正态分布, 下面分别求其期望与方差.

$$\begin{aligned} E(\hat{\beta}_1) &= \sum \frac{x_i - \bar{x}}{l_{xx}} E(y_i) = \sum \frac{x_i - \bar{x}}{l_{xx}} (\beta_0 + \beta_1 x_i) = \beta_1 \\ \text{Var}(\hat{\beta}_1) &= \sum \left(\frac{x_i - \bar{x}}{l_{xx}} \right)^2 \text{Var}(y_i) = \sum \frac{(x_i - \bar{x})^2}{l_{xx}^2} \sigma^2 = \frac{\sigma^2}{l_{xx}} \\ E(\hat{\beta}_0) &= E(\bar{y}) - E(\hat{\beta}_1) \bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \\ \text{Var}(\hat{\beta}_0) &= \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{l_{xx}} \right]^2 \text{Var}(y_i) = \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \right) \sigma^2 \end{aligned}$$

这就证明了 (1). 进一步, 考虑到诸 y_i 之间的独立性, 可得

$$\begin{aligned} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= \text{Cov}\left(\sum \left[\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{l_{xx}} \right] y_i, \sum \frac{x_i - \bar{x}}{l_{xx}} y_i\right) \\ &= \sum \left[\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{l_{xx}} \right] \frac{x_i - \bar{x}}{l_{xx}} \sigma^2 = -\frac{\bar{x}}{l_{xx}} \sigma^2 \end{aligned}$$

这就证明了 (2). 为证明 (3), 注意到 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 也是 y_1, y_2, \dots, y_n 的线性组合它也服从正态分布, 只需求出其期望与方差即可。

$$\begin{aligned} E(\hat{y}_0) &= E(\hat{\beta}_0) + E(\hat{\beta}_1) x_0 = \beta_0 + \beta_1 x_0 = E(y_0) \\ \text{Var}(\hat{y}_0) &= \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{\beta}_1) x_0^2 + 2 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \left[\left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \right) + \frac{x_0^2}{l_{xx}} - 2 \frac{x_0 \bar{x}}{l_{xx}} \right] \sigma^2 = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}} \right] \sigma^2 \end{aligned}$$

■

(R)

1. $\hat{\beta}_0, \hat{\beta}_1$ 分别是 β_0, β_1 的无偏估计
2. \hat{y}_0 是 $E(y_0) = \beta_0 + \beta_1 x_0$ 的无偏估计.
3. 除 $\bar{x} = 0$ 外, $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 是相关的.

4. 要提高 $\hat{\beta}_0, \hat{\beta}_1$ 的估计精度 (即降低它们的方差) 就要求 n 大, l_{xx} 大 (即要求 x_1, x_2, \dots, x_n 较分散)

Theorem 6.4.3 — 回归方程的显著性检验. 回归方程的显著性检验就是要对如下一对假设作出判断:

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_1: \beta_1 \neq 0$$

对此可采用如下两种等价的检验方法.

1. F 检验如下的平方和分解式是非常重要的, 它在许多统计领域得到应用:

$$S_T = S_R + S_e, f_T = f_R + f_e$$

其中

$$S_T = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n} (\sum y_i)^2 = l_{yy} \text{ 是总平方和, 其自由度 } f_r = n - 1$$

$$S_R = \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \hat{\beta}_1 l_{xy} = \hat{\beta}_1^2 l_{xx} \text{ 是回归平方和, 其自由度 } f_R = 1$$

$$S_e = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \text{ 是残差平方和, 其自由度 } f_e = n - 2.$$

而 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 是在 $x = x_i$ 的回归值 (拟合值), 它与实测值 y_i 通常是不相等的. 在原假设 H_0 成立的条件下, 检验统计量

$$F = \frac{S_R}{S_e/(n-2)} \sim F(1, n-2)$$

拒绝域为

$$W = \{F \geq F_{1-\alpha}(1, n-2)\}$$

2. t 检验: 由于

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{l_{xx}}\right), \frac{S_e}{\sigma^2} \sim \chi^2(n-2)$$

且与 $\hat{\beta}_1$ 相互独立, 因此在 H_0 为真时, 有

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}} \sim t(n-2)$$

其中 $\hat{\sigma} = \sqrt{S_e/(n-2)}$, 由于 $\sigma_{\hat{\beta}_1} = \sigma/\sqrt{l_{xx}}$, 因此称 $\hat{\sigma}_{\hat{\beta}_1} = \hat{\sigma}/\sqrt{l_{xx}}$ 为 $\hat{\beta}_1$ 的标准误, 即 $\hat{\beta}_1$ 的标准差的估计. 对给定的显著性水平 α , 拒绝域为

$$W = \{|t| > t_{1-\alpha/2}(n-2)\}$$

注意到 $t^2 = F$, 因此, t 检验与 F 检验是等同的.

Proof.

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

这组方程称为**正规方程组**

$\hat{\beta}_0, \hat{\beta}_1$ 满足正规方程组因此有

$$\begin{aligned} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \Rightarrow \sum (y_i - \hat{y}_i) = 0 \\ \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i &= 0 \Rightarrow \sum (y_i - \hat{y}_i) x_i = 0 \end{aligned}$$

利用 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$, 可得

$$\begin{aligned} \sum (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}) &= \sum (y_i - \hat{y}_i) [\hat{\beta}_1 (x_i - \bar{x})] \\ &= \hat{\beta}_1 [\sum (y_i - \hat{y}_i) x_i - \sum (y_i - \hat{y}_i) \bar{x}] = 0 \end{aligned}$$

从而

$$S_T = \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

即

$$S_T = S_R + S_e$$

■

Theorem 6.4.4 设 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, 其中 $\varepsilon_i, \dots, \varepsilon_n$ 相互独立, 且

$$E \varepsilon_i = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2, \quad i = 1, \dots, n$$

沿用上面的记号, 有

$$\begin{aligned} E(S_R) &= \sigma^2 + \beta_1^2 l_{xx} \\ E(S_e) &= (n-2)\sigma^2 \end{aligned}$$

上式说明 $\hat{\sigma}^2 = S_e/(n-2)$ 是 σ^2 的无偏估计.

Proof. 首先我们可以写出 S_R 的简化公式:

$$S_R = \sum (\hat{y}_i - \bar{y})^2 = \sum [\bar{y} + \hat{\beta}_1 (x_i - \bar{x}) - \bar{y}]^2 = \hat{\beta}_1^2 l_{xx}$$

从而

$$\begin{aligned} E(S_R) &= E(\hat{\beta}_1^2) l_{xx} = \left[\text{Var}(\hat{\beta}_1) + (E\hat{\beta}_1)^2 \right] \cdot l_{xx} \\ &= \left(\frac{\sigma^2}{l_{xx}} + \beta_1^2 \right) l_{xx} = \sigma^2 + \beta_1^2 l_{xx} \end{aligned}$$

另外

$$\begin{aligned} S_e &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum (\beta_0 + \beta_1 x_i + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum \left[(\hat{\beta}_0 - \beta_0)^2 + x_i^2 (\hat{\beta}_1 - \beta_1)^2 + \varepsilon_i^2 + 2(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)x_i \right. \\ &\quad \left. - 2(\hat{\beta}_0 - \beta_0)\varepsilon_i - 2(\hat{\beta}_1 - \beta_1)x_i\varepsilon_i \right] \end{aligned}$$

故

$$\begin{aligned} E(S_e) &= n \text{Var}(\hat{\beta}_0) + \sum x_i^2 \text{Var}(\hat{\beta}_1) + n \text{Var}(\varepsilon) + 2n\bar{x}\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &\quad - 2 \sum E(\hat{\beta}_0 \varepsilon_i) - 2 \sum x_i E(\hat{\beta}_1 \varepsilon_i) \end{aligned}$$

将 $\hat{\beta}_0, \hat{\beta}_1$ 写成 y_1, y_2, \dots, y_n 的线性组合, 利用 y_j 与 ε_i ($i \neq j$) 的独立性, 有

$$\begin{aligned} E(\hat{\beta}_0 \varepsilon_i) &= E\left[\varepsilon_i \sum_j \left(\frac{1}{n} - \frac{(x_j - \bar{x})}{l_{xx}}\right) y_j\right] = \left(\frac{1}{n} - \frac{(x_i - \bar{x})}{l_{xx}}\right) \sigma^2 \\ E(\hat{\beta}_1 \varepsilon_i) &= E\left[\varepsilon_i \sum_j \frac{x_j - \bar{x}}{l_{xx}} y_j\right] = \frac{x_i - \bar{x}}{l_{xx}} \sigma^2 \end{aligned}$$

由此即有

$$\sum E(\hat{\beta}_0 \varepsilon_i) = \sigma^2, \quad \sum x_i E(\hat{\beta}_1 \varepsilon_i) = \sigma^2$$

从而

$$\begin{aligned} E(S_e) &= n \left[\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \right] \sigma^2 + \sum \frac{x_i^2}{l_{xx}} \sigma^2 + n \sigma^2 - \frac{2n\bar{x}^2}{l_{xx}} \sigma^2 - 2\sigma^2 - 2\sigma^2 \\ &= (1+n-4)\sigma^2 + \frac{1}{l_{xx}} \sum (x_i - \bar{x})^2 \sigma^2 = (n-2)\sigma^2 \end{aligned}$$

■

Theorem 6.4.5 设 y_1, \dots, y_n 相互独立, 且 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \dots, n$, 则在上述记号下, 有

1. $S_e/\sigma^2 \sim \chi^2(n-2)$
2. 若 H_0 成立, 则有 $S_R/\sigma^2 \sim \chi^2(1)$
3. S_R 与 S_e, \bar{y} 独立 (或 $\hat{\beta}_1$ 与 S_e, \bar{y} 独立)

Proof. 取 $n \times n$ 正交矩阵 A , 具有如下形式:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n-2,1} & a_{n-2,2} & \cdots & a_{n-2,n} \\ (x_1 - \bar{x})/\sqrt{l_{xx}} & (x_2 - \bar{x})/\sqrt{l_{xx}} & \cdots & (x_n - \bar{x})/\sqrt{l_{xx}} \\ 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{pmatrix}$$

由正交性，可得如下一些约束条件

$$\begin{aligned}\sum_j a_{ij} = 0, \quad \sum_j a_{ij}x_j = 0, \quad \sum_j a_{ij}^2 = 1, \quad i = 1, 2, \dots, n-2 \\ \sum_k a_{ik}a_{jk} = 0, \quad 1 \leq i < j \leq n-2\end{aligned}$$

这里共有 $n(n-2)$ 个未知参数，约束条件有 $3(n-2) + \binom{n-2}{2} = (n-2)(n+3)/2$ 个，只要 $n \geq 3$ ，未知参数个数就不少于约束条件数，因此必定有解。令

$$\mathbf{Z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = AY = A \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j}y_j \\ \vdots \\ \sum_j a_{n-2,j}y_j \\ \sum_j \frac{x_j - \bar{x}}{\sqrt{l_{xx}}}y_j \\ \sum_j \frac{1}{\sqrt{n}}y_j \end{pmatrix}$$

其中

$$\begin{aligned}z_{n-1} &= \frac{\sum (x_i - \bar{x})y_i}{\sqrt{l_{xx}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{l_{xx}}} = \frac{l_{xy}}{\sqrt{l_{xx}}} = \sqrt{l_{xx}}\hat{\beta}_1 \\ z_n &= \frac{1}{\sqrt{n}} \sum y_i = \sqrt{n}\bar{y}\end{aligned}$$

则 \mathbf{Z} 仍然服从 n 维正态分布，且其期望与协差阵分别为

$$E\mathbf{Z} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \beta_1\sqrt{l_{xx}} \\ \sqrt{n}(\beta_0 + \beta_1\bar{x}) \end{pmatrix}, \quad \text{Var}(Z) = A \text{Var}(Y)A' = \sigma^2 I_n$$

这表明 z_1, z_2, \dots, z_n 相互独立 z_1, z_2, \dots, z_{n-2} 的共同分布为

$$N(0, \sigma^2), z_{n-1} \sim N(\beta_1\sqrt{l_{xx}}, \sigma^2), z_n \sim N(\sqrt{n}(\beta_0 + \beta_1\bar{x}), \sigma^2)$$

由于

$$\sum z_i^2 = \sum y_i^2 = S_T + n\bar{y}^2 = S_R + S_e + n\bar{y}^2$$

而 $z_n = \sqrt{n}\bar{y}, z_{n-1} = \sqrt{l_{xx}}\hat{\beta}_1 = \sqrt{S_R}$ ，于是有 $z_1^2 + z_2^2 + \dots + z_{n-2}^2 = S_e$ ，所以 S_e, S_R, \bar{y} 三者相互独立，并有

$$S_e/\sigma^2 = \sum_{i=1}^{n-2} (z_i/\sigma^2) \sim \chi^2(n-2)$$

在 $\beta_1 = 0$ 时，

$$S_R/\sigma^2 = (z_{n-1}/\sigma)^2 \sim \chi^2(1)$$



Definition 6.4.2 — 相关系数及其检验. 1. 相关系数: 对容量为 n 的二维样本 $(x_i, y_i), i = 1, 2, \dots, n$ 的线性相关程度可用如下(样本)相关系数度量

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$$

利用施瓦茨不等式可以证明: 样本相关系数也满足 $|r| \leq 1$, 其中等号成立条件是存在两个实数 a 与 b , 使得对 $i = 1, \dots, n$ 几乎处处有 $y_i = a + bx_i$

- (a) $r = \pm 1, n$ 个点完全在一条直线上, 此时两者之间可能是确定性关系
- (b) $r > 0$, 当 x 增加时, y 有线性增加趋势, 此时称正相关
- (c) $r < 0$, 当 x 增加时, y 反而有线性减少趋势, 此时称负相关
- (d) $r = 0, n$ 个点可能毫无规律, 也可能呈某种曲线趋势, 此时称不(线性)相关.

2. 相关系数的检验记 ρ 为二维总体的相关系数, 于是可建立如下假设:

$$H_0: \rho = 0 \quad \text{vs} \quad H_1: \rho \neq 0$$

对此, 采用检验统计量 $r = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$, 拒绝域为

$$W = \{|r| > r_{1-\alpha}(n-2)\}$$

其中 $r_{1-\alpha}(n-1)$ 是 $|r|$ 分布的 $1-\alpha$ 分位数

3. 检验统计量 r 与 F 统计量之间关系

$$r^2 = \frac{l_{xy}^2}{l_{xx}l_{yy}} = \frac{S_R}{S_T} = \frac{S_R}{S_R + S_e} = \frac{S_R/S_e}{S_R/S_e + 1}$$

而

$$F = \frac{MS_R}{MS_e} = \frac{S_R}{S_e/(n-2)} = \frac{(n-2)S_R}{S_e}$$

两者综合, 可得

$$r^2 = \frac{F}{F + (n-2)}$$

这表明 $|r|$ 是 F 的严格增函数, 所以相关系数检验与前面的 F 检验也是等价的. 这表明, $|r|$ 是 F 的严格单调增函数, 故可以从 F 分布的 $1-\alpha$ 分位数 $F_{1-\alpha}(1, n-2)$ 得到 $|r|$ 的 $1-\alpha$ 分位数为

$$c = r_{1-\alpha}(n-2) = \sqrt{\frac{F_{1-\alpha}(1, n-2)}{F_{1-\alpha}(1, n-2) + n-2}}$$

Theorem 6.4.6 — 估计与预测—回归方程的应用. 1. 当 $x = x_0$ 时, $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 是

$$E(y_0) = \beta_0 + \beta_1 x_0$$

的点估计

2. 当 $x = x_0$ 时, $E(y_0) = \beta_0 + \beta_1 x_0$ 的置信水平由 $1-\alpha$ 的置信区间是 $[\hat{y}_0 - \delta_0, \hat{y}_0 + \delta_0]$,

其中

$$\delta_0 = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}, \quad \hat{\sigma} = \sqrt{MS_e}$$

3. 当 $x = x_0$ 时, $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ 的 $1 - \alpha$ 预测区间是 $[\hat{y}_0 - \delta, \hat{y}_0 + \delta]$, 其中

$$\delta = \delta(x_0) = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}$$

注: $E(y_0)$ 是未知参数, 而 y_0 是随机变量. 对 $E(y_0)$ 谈论的是置信区间, 对 y_0 谈论的是预测区间, 两者是不同的, 显然, 预测区间要比置信区间宽很多. 要提高预测区间(置信区间也一样)的精度, 即要使 δ (或 δ_0) 较小, 这要求

1. 增大样本量 n
2. 增大 l_{xx} , 即要求 x_1, x_2, \dots, x_n 较为分散
3. 使 x_0 靠近 \bar{x}

(R) 当 n 较大时(如 $n > 30$), t 分布可以用正态分布近似, 进一步, 若 x_0 与 \bar{x} 相差不大时, δ 可以近似取为

$$\delta \approx \hat{\sigma}u_{1-\alpha/2}$$

其中 $u_{1-\alpha/2}$ 是标准正态分布的 $1 - \alpha/2$ 分位数.

Proof. 为得到 $E(y_0)$ 的区间估计, 我们需要知道 y_0 的分布.

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\beta_0 + \beta_1 x_0, \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right] \sigma^2\right)$$

并且有, $S_e/\sigma^2 \sim \chi^2(n-2)$, 且与 $\hat{y}_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$ 相互独立, 记

$$\hat{\sigma}^2 = \frac{S_e}{n-2}$$

则

$$\frac{(\hat{y}_0 - E y_0) / \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}} \sigma}{\sqrt{\frac{S_e}{\sigma^2} / (n-2)}} = \frac{\hat{y}_0 - E y_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2)$$

于是 $E(y_0)$ 的 $1 - \alpha$ 的置信区间是

$$[\hat{y}_0 - \delta_0, \hat{y}_0 + \delta_0]$$

其中

$$\delta_0 = t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}$$

Proof. 事实上, $y_0 = E(y_0) + \varepsilon$, 由于通常假定 $\varepsilon \sim N(0, \sigma^2)$, 因此, y_0 的最可能取值仍然为 \hat{y}_0 , 于是, 我们可以使用以 \hat{y}_0 为中心的一个区间

$$(\hat{y}_0 - \delta, \hat{y}_0 + \delta) \quad (6.3)$$

作为 y_0 的取值范围, 为确定 δ 的值, 我们需要如下的结果: 由于 y_0 与 \hat{y}_0 独立, 故

$$y_0 - \hat{y}_0 \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right] \sigma^2\right)$$

因此有

$$\frac{y_0 - \hat{y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2)$$

从6.3表示的预测区间中 δ 的表达式为

$$\delta = \delta(x_0) = t_{1-\alpha/2}(n-2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}$$

■

6.5 一元非线性回归

Definition 6.5.1 — 非线性函数形式. 根据二维样本的散点图确定可能的非线性函数形式, 部分常见的非线性函数及其图形如下表. 注意: 若有两个或两个以上非线性函数可用, 可对它们分别拟合非线性回归并根据后面提及的一些准则进行选择。

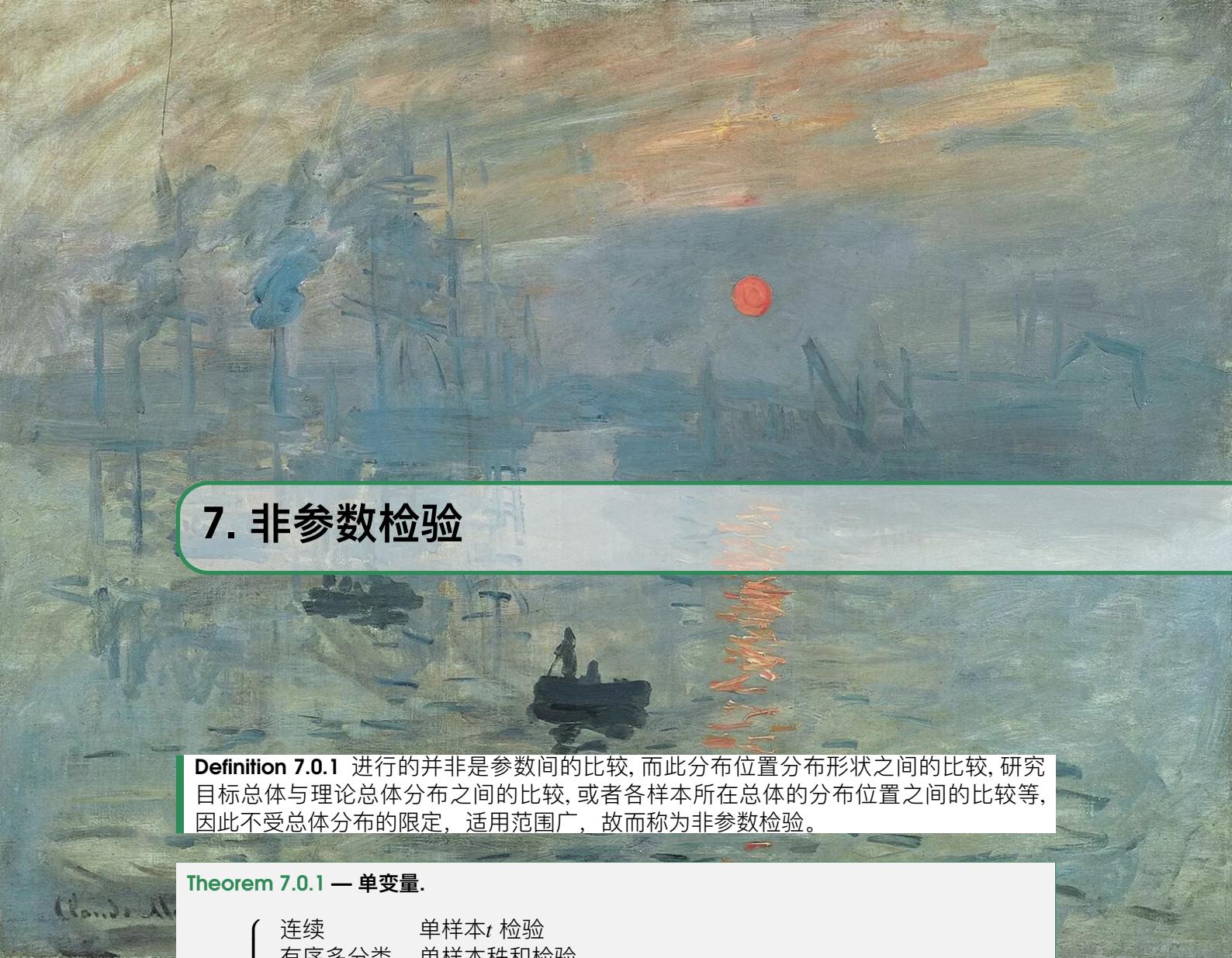
Theorem 6.5.1 — 参数估计. 通过适当变换, 把非线性函数转化为线性函数形式, 然后对未知参数寻求最小二乘估计. 譬如由 $\frac{1}{y} = a + \frac{b}{x}$ 可转化为 $v = a + bu$ 只要令 $\frac{1}{y} = v, \frac{1}{x} = u$ 即可.

Theorem 6.5.2 — 评价标准. 常用的曲线回归方程的好坏评价标准有两个:

1. 决定系数 $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$, 愈大愈好

2. 剩余标准差 $s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$ 愈小愈好

这两个评价标准是一致的, 只是从两个侧面作出评价。



7. 非参数检验

Definition 7.0.1 进行的并非是参数间的比较, 而此分布位置分布形状之间的比较, 研究目标总体与理论总体分布之间的比较, 或者各样本所在总体的分布位置之间的比较等, 因此不受总体分布的限定, 适用范围广, 故而称为非参数检验。

Theorem 7.0.1 — 单变量.

连续	单样本t检验
有序多分类	单样本秩和检验
无序多分类	单样本卡方检验
二分类	二项分布确切概率法

Theorem 7.0.2 — 应变量: 连续变量. 1. 单个自变量:

连续	相关分析, 回归分析
有序多分类	单因素方差分析, 解释结果时利用有序信息
无序多分类	单因素方差分析
二分类	两样本t检验

2. 多个自变量:

连续变量为主	线性回归模型
分类变量为主	方差分析模型, 和回归模型实际上等价

在应用 t 检验进行两样本均数的比较时, 要求数据满足以下 3 个条件。

1. 独立性, 各观察值之间是相互独立的, 不能相互影响。
2. 正态性, 各个样本均来自于正态分布的总体。
3. 方差齐性, 各个样本所在总体的方差相等。

方差齐性检验可通过 Levene's 检验来进行, 其假设为

1. $H_0: \sigma_1^2 = \sigma_2^2$, 两总体方差相同。
 2. $H_1: \sigma_1^2 \neq \sigma_2^2$, 两总体方差不同。
- Levene's 检验的实质是将两组数据的方差进行比较, 其统计量的计算公式为

$$F = s_1^2/s_2^2, v_1 = n_1 - 1, v_2 = n_2 - 1$$

其中分子为较大的方差, 如果两组方差的比值较大, 其所对应的 P 值小于设定的检验水准, 则按照小概率反证法原理拒绝 H_0 , 认为两组所在总体的方差不齐。

Theorem 7.0.3 — 配对 t 检验--本质上是单样本 t 检验. 相应的假设为

1. $H_0: \mu_d = 0$, 两种处理没有差别。
2. $H_1: \mu_d \neq 0$, 两种处理存在差别。

其统计量的计算公式为

$$t = \frac{\bar{d} - 0}{s_{\bar{d}}} = \frac{\bar{d}}{s/\sqrt{n}}, df = n - 1 (n \text{ 为对子数})$$

Theorem 7.0.4 — 应变量: 有序分类变量. 1. 单个自变量:

连续	有序分类的 Logistic 回归
有序多分类	秩相关分析、CMH 卡方
无序多分类	多样本秩和检验(H 检验)
二分类	两样本秩和检验 (W 检验)

2. 多个自变量:

连续变量为主	有序分类的判别分析, 有序分类的 Logistic 回归
分类变量为主	有序分类的 Logistic 回归

Theorem 7.0.5 — 应变量: 无序分类变量. 1. 单个自变量:

连续	无序分类的 Logistic 回归
有序多分类	可将自/应变量交换后进行分析
无序多分类	卡方检验, 深入分析时可用对数线性模型
二分类	卡方检验

2. 多个自变量:

连续变量为主	判别分析、无序分类的 Logistic 回归
分类变量为主	无序分类的 Logistic 回归

Theorem 7.0.6 — 应变量: 二分类变量. 1. 单个自变量:

连续	两分类 Logistic 回归
有序多分类	可将自/应变量交换后进行分析
无序多分类	卡方检验, 两分类 Logistic 回归
二分类	四格表卡方检验, 确切概率法

2. 多个自变量:

$$\left\{ \begin{array}{l} \text{连续变量为主 判别分析, 两分类 Logistic 回归} \\ \text{分类变量为主 两分类 Logistic 回归} \end{array} \right.$$

- Theorem 7.0.7 — 多元分析方法.**
1. 考察的特征需要由多个应变量来表示, 同时研究多个自变量对它们的影响: 多元方差分析模型、多元回归模型。
 2. 希望将变量/记录分成若干个类别, 但类别数不清楚, 或各类别的特征不明: 聚类分析。
 3. 已知分类情况, 研究目的是希望建立判别方程, 对之后新进入的案例进行所属类别的预测: 判别分析。
 4. 需要探索多个连续变量间的内在联系或数据的内在结构: 因子分析。
 5. 需要探索多个分类变量间的内在联系或数据的内在结构: 对应分析。
 6. 考察多个概念间的相似程度, 并寻找受访者用于评价相似性的标准: 多维尺度分析。
 7. 生存时间和生存结局都是需要关心的因素, 同时数据中存在大量的失访: 生存分析。
 8. 得到的是时间序列数据, 需要根据历史资料对之后的情形加以预测: 时间序列模型。

Definition 7.0.2 — Bootstrap. 1. 判断原参数估计值是否准确;

2. 计算出更准确的可信区间, 判断得出的统计学结论是否正确。

Bootstrap 方法的基本思想为: 在原始数据的范围内做有放回的抽样, 样本含量仍为 n , 原始数据中每个观察单位每次被抽到的概率相等, 为 $1/n$, 所得样本称为 Bootstrap 样本。于是可得到任何一个参数 θ 的一个估计值 $\hat{\theta}^{(b)}$, 重复抽取这样的样本若干次, 记为 B 。例如 $B = 1000$, 就得到该参数的 1 000 个估计值, 则参数 θ 的标准误的 Bootstrap 估计为

$$s_{\hat{\theta}} = \left\{ \sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2 / (B-1) \right\}^{1/2}$$

其中, $\hat{\theta}^*(.) = \sum_{b=1}^B \hat{\theta}^*(b)/B$, 根据其性质可以估计得到 θ 的一些性质, 如 $\hat{\theta}^{(b)}$ 的分布是否为正态, $\theta^{(b)}$ 的均数及标准差(误), θ 的可信区间等

(R)

1. 中位数不受极端值的影响, 截尾均值
2. 变异系数消除了量纲的影响

- Theorem 7.0.8 — 检验.**
1. 正态检验: Kolmogorov - Smirnov (K-S) 单样本检验, 分布拟合优度检验, 其方法是将一个变量的累积分布函数与特定分布进行比较。
 2. 游程检验 (Runs Test) 是对二分变量的随机检验, 它可用于判断观察值的顺序是否为随机

Theorem 7.0.9 — 方差分析. 方差分析的条件: 独立性、正态性和方差齐性。

方差分析是基于变异分解的思想进行的, 在单因素方差分析中, 整个样本的变异可以看成由如下两个部分构成:

$$\text{总变异} = \text{随机变异} + \text{处理因素导致的变异}$$

其中随机变异是永远存在的, 确定处理因素导致的变异是否存在就是所要达到的研究目标, 即只要能证明它不等于 0, 就等同于证明了处理因素的确存在影响。

$$\text{总变异} = \text{随机变异} + \text{处理因素导致的变异}$$

$$\text{总变异} = \text{组内变异} + \text{组间变异}$$

组内变异由随机变异引起, 组间变异由两种变异导致。检验统计量是: 组间比上组内。

在 H_0 成立时, 处理所造成的各组间均数的差异应为 0(理论上应为 0, 但由于抽样误差不可能恰好为 0) , 即

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

于是, 组间变异将主要由随机误差构成, 即组间变异的值应当接近组内变异。于是检验统计量 F 值应当不会太大, 且接近于 1 ; 否则, F 值将会偏离 1, 并且各组间的不一致程度越强, F 值越大。

方差分析的原假设和备择假设分别为

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$H_1 : k$ 个总体均数不同或者不全相同沿用上面的变量标记方式, 有检验统计量为

$$F_{k-1, N-k} = \frac{MS_B}{MS_W} = \frac{SS_B/(k-1)}{SS_W/(N-k)}$$

方差分析表

变异来源	离差平方和	自由度	均方	F	P
组间变异	SS _B	$k - 1$	MS _B	MS _B /MS _W	$P = \{F_{t-1, k(m-1)} \geq F\}$ 如
组内变异	SS _w	$k(n_i - 1)$	MS _w		
总变异	SS _T	$N - 1$	MS _r		

果假设检验拒绝了 H_0 , 可以得出多个样本不是来自同一总体的结论。但是到底这些样本来自于几个不同的总体, 这次假设检验还不能回答这个问题, 而需要进一步进行单因素不同水平间的多重比较 (Multiple - Comparison)

Definition 7.0.3 — 评估标准. 1. CER: 比较误差, 即每进行一次比较犯一类错误的概率。

2. EERC: 在完全无效假设下的试验误差率, 即在 H_0 成立时做完全部比较所犯一类错误的概率。
3. MEER: 最大试验误差率, 即在任何完全或部分无效假设下做完全部比较所犯一类错误的最大概率值。

如前所述, 当无效假设实际上成立, 各组均数无差别时, k 组完全两两比较的次数 $c = k(k - 1)/2$, 做完所有这些比较犯第一类错误的概率为 $1 - (1 - \alpha_{ij})^c$, 此即 EERC, 所做的方差分析的实质也就是控制 EERC 为所设定的水准。因此, 进行一类错误控制时最直接的想法就是将总的

Definition 7.0.4 多重比较分为两种类型: 计划好的和非计划的。所谓计划好的多重比较 (Planned Comparisons), 即在收集数据之前便决定了要通过多重比较来考察多个组与某个特定组间的差别或者某几个特定组间彼此的差别; 而非计划的多重比较 (Unplanned Comparisons, PostHoc Comparisons) 只有在方差分析得到有统计学意义的 F 值后才有必

要进行, 是一种探索性的分析。

1. S-N-K 法: 经常在有关统计学教材中出现, 全称为 Student-Newman-Keuls 法。它实质上是根据预先指定的准则将各组均数分为多个子集, 利用 studentized range 分布来进行假设检验, 并根据所要检验的均数的个数调整总的一类错误概率不超过 α
2. Tukey 法: 即 Tukey's Honestly Significant Difference 法, 应用这种方法要求各组样本含量相同。它也是利用 studentized range 分布来进行各组均数间的比较的, 与 SNK 法不同的是, 它用于控制所有比较中最大的一类错误的概率, 即 MEER 不超过 α_0

Definition 7.0.5 秩 (Rank) 及秩统计量: 对于样本 X_1, \dots, X_n , 按由小到大的顺序排成一列, 若 X_i 在这列中占据第 R_i 位, 称 X_i 的秩为 R_i , $R_i = \sum_{j=1}^n I(X_j \leq X_i)$, 即小于或等于 X_i 的样本点个数, 称 $R = (R_1, \dots, R_n)$ 是原样本的秩统计量。

Definition 7.0.6 结 (Ties) 和结统计量: 在许多情况下, 数据中会有相同的值出现, 此时如果排秩就会出现同秩的现象, 这种情况称为数据中的结。

7.1 两匹配样本非参数检验

原假设和备则假设: $H_0: \text{差值的总体中位数 } M_d = 0; H_1: \text{差值的总体中位数 } M_d \neq 0$

Theorem 7.1.1 — 符号检验. 符号检验原理是如果两个配对样本实际上无区别, 则将数据样本相减所得到的差值应该一半正, 一半负。将差值为正的个数记为 S^+ , 差值为负的个数记为 S^- , 按照中位数的意义, 若 $H_0: M = M_0$ 成立, 那么 S^*, S^- 应大体相等, S^*, S^- 都服从二项分布 $B(n, 0.5)$ 。当 S^*, S^- 过大或过小, 或者 $\min(S^*, S^-)$ 过小时, 就有理由拒绝 H_0

Wilcoxon 符号秩检验考虑样本差数的符号, 同时又考虑到差数的顺序。不同的符号代表了在中心位置的哪一边, 而差的绝对值代表了距离中心的远近, 两者结合会更有效(注意该秩和检验利用的是样本差数的顺序, 而不是样本差数数值本身, 在这方面又比参数检验利用样本数值本身的信息逊色)。 $|d_i|$ 表示数据对的差值。对 $|d_i|$ 由低到高进行排秩, 相同的差异将被赋予平均秩, 若 X, Y 具有相同的分布, 那么 $P(d_i > 0) = P(d_i < 0)$ 。把 $|d_i|$ 看成单样本, 令 W^+ 表示 $|d_i > 0|$ 的秩和, W^- 表示 $|d_i < 0|$ 的秩和。检验统计量取 $W = \min(W^+, W^-)$, 在文献中也记为统计量 T ; 当 H_0 (差值的总体中位数 $M_d = 0$), 应该离对称轴 $n(n+1)/4$ 不远

7.2 两独立样本的非参数检验

Theorem 7.2.1 — Mann-Whitney U 检验. 检验两独立样本是否来自同一个样本, 没有三条条件, 基本思想是把两组样本混合后, 分别求两组样本的秩

Theorem 7.2.2 — 多样本的独立性检验. Kruskal-Wallis H 检验--克罗斯考尔和瓦里斯

Theorem 7.2.3 — 多个相关样本的非参数检验. 1. Friedman 检验也称为弗里德曼双向评价方差分析, 该方法的基本思想是: 由于区组间的差异是各式各样的, 只有同区组的处理值的比较才有意义, 一个观察值的秩是在某一区组中的秩, 而不是对所有数据而言的。因此应当独立地在每一个区组内分别对数据进行排秩, 这样就可以消除区组间的差异以检验各种处理之间是否存在差异。该检验的假设如下:

$H_0: M_1 = \dots = M_k$ (所有的位置参数都相等)

$H_{1,:}$: 至少有一个 M_i 与其他不同 (不是所有位置参数都相等)

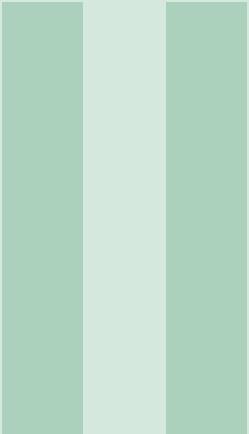
2. 多人评价标准是否一致: 原假设为 H_0 : 这些评估 (对于不同个体) 是不相关的或者是随机的; 而备择假设为 H_1 : 评估是正相关的或者是一致的。F 只能知道一致吗
3. K 可以知道一致性程度: Kendall 协和系数: W 愈接近 1, b 个变量间的正相关性愈好, 即表现的一致性愈强; 反之, W 愈接近 0, 变量间的正相关性愈差, 一致性愈弱。因此与 Friedman 检验相比, Kendall 协和系数不仅可以检验 k 个相关样本是否来自同一总体, 还能检验 b 个变量间的相关性。它表亦的是 k 个指标间相互关联的程度 (一致性程度), 取值在 0 ~ 1 之间。
4. 解决二元变量会出现很多结的情况: Cochran 检验

Theorem 7.2.4 — χ^2 检验的作用.

1. 检验某个连续变量的分布是否与某种理论分布相一致。如是否符合正态分布、是否服从均匀分布, 是否服从 Poisson 分布等。
2. 检验某个分类变量各类的出现概率是否等于指定概率。
3. 检验某两个分类变量是否相互独立。
4. 检验控制某种或某几种分类因素的作用以后, 另两个分类变量是否相互独立。
5. 检验某两种方法的结果是否一致。

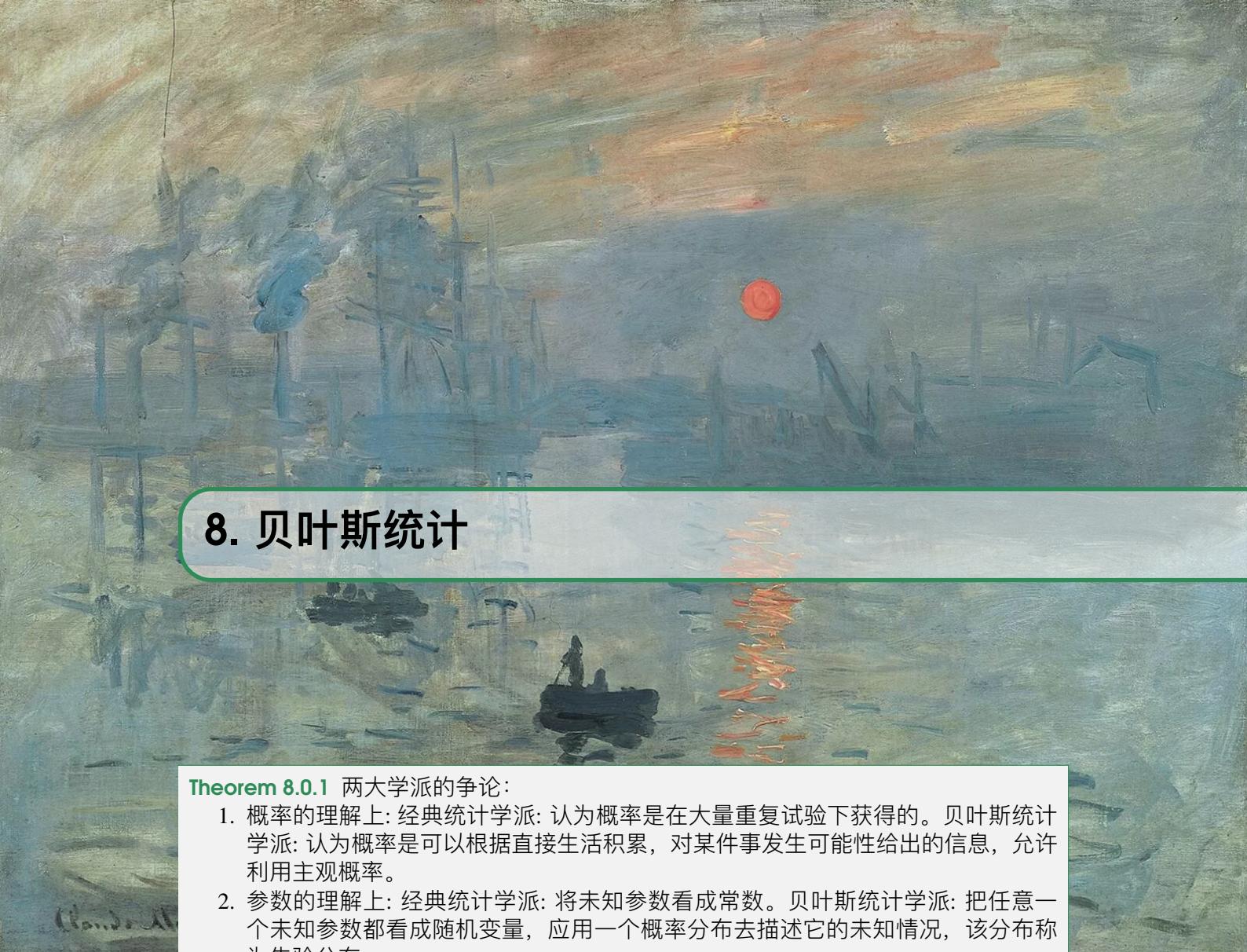
Theorem 7.2.5 — 相关系数检验.

1. 皮尔逊检验: 要求 XY 都服从正态分布, 不能有极值
2. Spearman 等级相关系数: 利用两变量的秩次大小进行线性相关分析的,



Part Three

8	贝叶斯统计	225
8.1	贝叶斯推断	
8.2	先验分布的确定	
8.3	决策	
8.4	考试重点	
8.5	期末复习	
8.6	第三章	
8.7	第四章	
8.8	第五章	



8. 贝叶斯统计

Theorem 8.0.1 两大学派的争论:

- 概率的理解上: 经典统计学派: 认为概率是在大量重复试验下获得的。贝叶斯统计学派: 认为概率是可以根据直接生活积累, 对某件事发生可能性给出的信息, 允许利用主观概率。
- 参数的理解上: 经典统计学派: 将未知参数看成常数。贝叶斯统计学派: 把任意一个未知参数都看成随机变量, 应用一个概率分布去描述它的未知情况, 该分布称为先验分布。
- 先验信息的利用与非: 经典统计学派: 不利用; 贝叶斯学派: 利用
- 样本信息的利用: 经典统计学派: 把样本看做是来自总体的信息, 研究的是总体, 不局限数据本身。贝叶斯统计学: 是重视样本观察值, 而对尚未发生的样本观察值不予考虑。
- 对可信区间和置信区间的认识上不同: 经典统计学派: 把真值看做常量, 置信水平为 $1 - \alpha$, m 次使用这个区间时, 大概有 $m(1 - \alpha)$ 个可以覆盖住 θ ; 贝叶斯统计学派: 将参数的真值看成是变量, 可信水平表示 θ 落入在可信区间内的概率。例如: $p\{x_1 \leq \theta \leq x_2\} = 0.9$ 表示 θ 落入 $[x_1, x_2]$ 内的概率为 0.9。

共同点: 1) 都承认样本有概率分布; 2) 概率的计算遵循共同的准则。

Definition 8.0.1

- 总体信息: 总体分布或所属分布族提供给我们的信息。
- 样本信息: 从总体抽取的样本提供给我们的信息。
- 先验信息: 在抽样之前有关统计推断的一些信息。

Theorem 8.0.2 贝叶斯的三个假设

- 设总体指标 X 有依赖于参数 θ 的密度函数, 在经典统计中常记为 $p(x; \theta)$ 或 $p_\theta(x)$, 它表示在参数空间 $\Theta = \{\theta\}$ 中不同的 θ 对应不同的分布。可在贝叶斯统计中记为 $p(x|\theta)$, 它表示在随机变量 θ 给定某个值时, 总体指标 X 的条件分布。
- 从总体 $p(x|\theta)$ 中随机抽取样本 X_1, X_2, \dots, X_n , 该样本中含有 θ 的有关信息是**样本信息**。

Definition 8.0.2 — 先验分布. 1. 将总体中的未知参数 $\theta \in \Theta$ 看成一取值于 Θ 的随机变量, 它有一概率分布, 记为 $\pi(\theta)$, 称为参数 θ 的**先验分布**.
2. 样本 X_1, X_2, \dots, X_n 和参数 θ 的联合密度函数:

$$h(x_1, x_2, \dots, x_n, \theta) = p(x_1, x_2, \dots, x_n | \theta) \pi(\theta)$$

3. 样本 X_1, X_2, \dots, X_n 的边际分布或无条件分布.

$$m(x_1, x_2, \dots, x_n) = \int_{\Theta} p(x_1, x_2, \dots, x_n | \theta) \pi(\theta) d\theta$$

4. $\pi(\theta | x_1, \dots, x_n)$ 称为 θ 的后验密度函数, 或后验分布. 这就是贝叶斯公式的密度函数形式

$$\pi(\theta | x) = \frac{h(x, \theta)}{m(x)} = \frac{p(x | \theta) \pi(\theta)}{\int_{\Theta} p(x | \theta) \pi(\theta) d\theta} \quad (8.1)$$

■ **Example 8.1** 为了提高某产品的质量, 公司经理考虑增加投资来改进生产设备, 预计需投资 90 万元, 但从投资效果看, 下属部门有二种意见:

θ_1 : 改进生产设备后, 高质量产品可占 90%

θ_2 : 改进生产设备后, 高质量产品可占 70%

经理当然希望 θ_1 发生, 公司效益可得很大提高, 投资改进设备也是合算的. 但根据下属二个部门过去建议被采纳的情况, 经理认为, θ_1 的可信程度只有 40%, θ_2 的可信程度是 60%. 即

$$\pi(\theta_1) = 0.4, \quad \pi(\theta_2) = 0.6$$

这二个都是经理的主观概率. 经理不想仅用过去的经验来决策此事, 想慎重一些通过小规模试验后观其结果再定. 为此做了一项试验, 试验结果(记为 A)如下:

A: 试制五个产品, 全是高质量的产品

经理对这次试验结果很高兴, 希望用此试验结果来修改他原先对 θ_1 和 θ_2 的看法, 即要求后验概率 $\pi(\theta_1 | A)$ 与 $\pi(\theta_2 | A)$. 这可用贝叶斯公式的离散形式来完成. 如今已有先验概率 $\pi(\theta_1)$ 与 $\pi(\theta_2)$. 还需要二个条件概率 $P(A | \theta_1)$ 与 $P(A | \theta_2)$. 这可用二项分布算得,

$$P(A | \theta_1) = 0.9^5 = 0.590, \quad P(A | \theta_2) = 0.7^5 = 0.168$$

由全概率公式可算得 $P(A) = P(A | \theta_1) \pi(\theta_1) + P(A | \theta_2) \pi(\theta_2) = 0.337$. 最后可算得,

$$\pi(\theta_1 | A) = P(A | \theta_1) \pi(\theta_1) / P(A) = 0.236 / 0.337 = 0.700$$

$$\pi(\theta_2 | A) = P(A | \theta_2) \pi(\theta_2) / P(A) = 0.101 / 0.337 = 0.300$$

这表明, 经理根据试验 A 的信息调整自己的看法, 把对 θ_1 与 θ_2 的可信程度由 0.4 和 0.6 调整到 0.7 和 0.3. 后者是综合了经理的主观概率和试验结果而获得的, 要比主观概率更有吸引力, 更贴近当今的实际, 这就是贝叶斯公式的应用. 经过实验 A 后, 经理对增加投资改进质量的兴趣增大.

但因投资额大, 还想再做一次小规模试验, 观其结果再作决策. 为此又做了一批试验, 试验结果(记为 B)如下:

B: 试制 10 个产品, 有 9 个是高质量产品

经理对此试验结果更为高兴。希望用此试验结果对 θ_1 与 θ_2 再作一次调整。为此把上次后验概率看作这次的先验概率, 即

$$\pi(\theta_1) = 0.7, \pi(\theta_2) = 0.3$$

用二项分布还可算得

$$\begin{aligned} P(B|\theta_1) &= 10(0.9)^9(0.1) = 0.387 \\ P(B|\theta_2) &= 10(0.7)^9(0.3) = 0.121 \end{aligned}$$

由此可算得 $P(B) = 0.307$ 和后验概率 $\pi(\theta_1|B) = 0.883, \pi(\theta_2|B) = 0.117$ 经理看到, 经过二次试验, θ_1 (高质量产品可占 90%) 的概率已上升到 0.883, 到可以下决心的时候了, 他能以 88.3% 的把握保证此项投资能取得较大经济效益。



注意从题目中提取先验信息

8.0.1 共轭先验分布

1. 写出样本的似然函数, 注意个数, 注意定义域
2. 确定先验分布
3. 计算后验分布

Theorem 8.0.3 — 正态均值 (方差已知) 的共轭先验分布是正态分布. 设 x_1, \dots, x_n 是来自正态分布 $N(\theta, \sigma^2)$ 的一组样本观察值。其中 σ^2 已知。此样本的似然函数为:

$$P(x|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right\}$$

$$-\infty < x_1, \dots, x_n < +\infty$$

现取另一个正态分布 $N(\mu, \tau^2)$ 作为正态均值 θ 的先验分布, 即

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau} \exp \left\{ -\frac{(\theta - \mu)^2}{2\tau^2} \right\}, -\infty < \theta < +\infty$$

其中 μ 与 τ^2 为已知. 结论是:

$$\sigma_0^2 = \frac{\sigma^2}{n}, A = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2}, B = \frac{\bar{x}}{\sigma_0^2} + \frac{\mu}{\tau^2}, C = \frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu^2}{\tau^2}$$

后验分布也就是正态分布, 两个参数的值为

$$\mu_1 = \frac{B}{A} = \frac{\bar{x}\sigma_0^{-2} + \mu\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}}, \quad \frac{1}{\tau_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2} \quad (8.2)$$

■ **Example 8.2** 例 1.3.2 二项分布中的成功概率 θ 的共轭先验分布是贝塔分布。设总体 $X \sim b(n, \theta)$, 其密度函数中与 θ 有关部分 (核) 为 $\theta^x(1-\theta)^{n-x}$ 。又设 θ 的先验分布为贝塔分布 $Be(\alpha, \beta)$, 其核为 $\theta^{\alpha-1}(1-\theta)^{\beta-1}$, 其中 α, β 已知, 从而可写出 θ 的后验分布

$$\pi(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}, 0 < \theta < 1$$

立即可以看出, 这是贝塔分布 $\text{Be}(\alpha+x, \beta+n-x)$ 的核, 故此后验密度为

$$\pi(\theta|x) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)} \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1}, 0 < \theta < 1$$

从上述二个例子可以看出: 只要先验的核均总体分布的核类似, 则此先验一定是共轭先验。

■ **Example 8.3** 例 1.3.5 设 x_1, \dots, x_n 是来自正态分布 $N(\theta, \sigma^2)$ 的一个样本观测值, 其中 θ 已知, 现要寻求方差 σ^2 的共轭先验分布, 由于该样本的似然函数为

$$\begin{aligned} p(x|\sigma^2) &= \left\{ \frac{1}{\sqrt{2\pi}\sigma} \right\}^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right\} \\ &\propto \left(\frac{1}{\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 \right\} \end{aligned}$$

设 X 服从伽玛分布 $\text{Ga}(\alpha, \lambda)$, 其中 $\alpha > 0$ 为形状参数, $\lambda > 0$ 为尺度参数, 其密度函数为

$$p(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0$$

通过概率运算可以求得 $Y=X^{-1}$ 的密度函数

$$p(y|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y} \right)^{\alpha+1} e^{-\lambda/y}, y > 0$$

这个分布称为倒伽玛分布, 记为 $\text{IGa}(\alpha, \lambda)$, 其均值为 $E(y) = \lambda/(\alpha-1)$ 。假如取此倒伽玛分布为 σ^2 的先验分布, 其中参数 α 与 λ 已知, 则其密度函数为

$$\pi(\sigma^2) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2} \right)^{\alpha+1} e^{-\lambda/\sigma^2}, \sigma^2 > 0$$

于是 σ^2 的后验分布为

$$\begin{aligned} \pi(\sigma^2|x) &\propto p(x|\sigma^2) \pi(\sigma^2) \\ &\propto \left(\frac{1}{\sigma^2} \right)^{a+\frac{n}{2}+1} \exp \left\{ -\frac{1}{\sigma^2} \left[\lambda + \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \right] \right\} \end{aligned}$$

容易看出, 这仍是倒伽玛分布 $\text{IGa}\left(\alpha + \frac{n}{2}, \lambda + \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right)$, 这表明, 倒伽玛分布 $\text{IGa}(\alpha, \lambda)$ 是正态方差 σ^2 的共轭先验分布。

总体分布	参数	共轭先验分布
二项分布	成功概率	贝塔分布 $\text{Be}(\alpha, \beta)$
泊松分布	均值	伽玛分布 $\text{Ga}(\alpha, \lambda)$
指数分布	均值的倒数	伽玛分布 $\text{Ga}(\alpha, \lambda)$
正态分布 (方差已知)	均值	正态分布 $N(\mu, \tau^2)$
正态分布 (均值已知)	方差	倒伽玛分布 $\text{IGa}(\alpha, \lambda)$

Definition 8.0.3 超参数定义: 先验分布中所含的未知参数称为超参数. 方法

1. 利用先验矩
2. 利用先验分位数:
3. 利用先验矩和先验分位数

8.0.2 充分统计量

Definition 8.0.4 经典统计中充分统计量是这样定义的: 设 $x = (x_1, \dots, x_n)$ 是来自分布函数 $F(x|\theta)$ 的一个样本, $T = T(x)$ 是统计量, 假如在给定 $T(x) = t$ 的条件下, x 的条件分布与 θ 无关的话, 则称该统计量为 θ 的充分统计量

Theorem 8.0.4 — 因子分解. 一个统计量 $T(x)$ 对参数 θ 是充分的充要条件是存在一个 t 与 θ 的函数 $g(t, \theta)$ 和一个样本 x 的函数 $h(x)$, 使得对任一样本 x 和任意 θ , 样本的密度 $p(x|\theta)$ 可表示为它们的乘积, 即

$$p(x|\theta) = g(T(x), \theta)h(x)$$

■ **Example 8.4** 例 1.6. 2 设 $x = (x_1, \dots, x_n)$ 是来自正态分布 $N(\theta, 1)$ 的一个样本, 大家知道样本均值 \bar{x} 是 θ 的充分统计量, 若 θ 的先验分布取为正态分布 $N(0, \tau^2)$, 其中 τ^2 已知, 那么 θ 的后验分布可用充分统计量 \bar{x} 的分布算得, 即

$$\begin{aligned} \pi(\theta|\bar{x}) &\propto \exp\left\{-\frac{\pi}{2}(\bar{x}-\theta)^2 - \frac{\theta^2}{2\tau^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}[\theta^2(n+\tau^{-2}) - 2n\theta\bar{x}]\right\} \\ &\propto \exp\left\{-\frac{n+\tau^{-2}}{2}\left(\theta - \frac{n\bar{x}}{n+\tau^{-2}}\right)^2\right\} \\ &= N\left(\frac{n\bar{x}}{n+\tau^{-2}}, \frac{1}{n+\tau^{-2}}\right) \end{aligned}$$

Theorem 8.0.5 设 $x = (x_1, \dots, x_n)$ 是来自密度函数 $p(x|\theta)$ 的一个样本, $T = T(x)$ 是统计量, 它的密度函数为 $p(t|\theta)$, 又设 $\mathcal{H} = \{\pi(\theta)\}$ 是 θ 的某个先验分布族, 则 $T(x)$ 为 θ 的充分统计量的充要条件是对任一先验分布 $\pi(\theta) \in \mathcal{H}$, 有

$$\pi(\theta|T(x)) = \pi(\theta|x)$$

即用样本分布 $p(x|\theta)$ 算得的后验分布与统计量 $T(x)$ 算得的后验分布是相同的。

■ **Example 8.5** 设总体 X 服从两点分布 $B(1, p)$, 即

$$P\{X=x\} = p^x(1-p)^{1-x}, x=0, 1$$

其中 $0 < p < 1$, $(X_1, \dots, X_n)^T$ 是来自总体 X 的一个样本, 试证 $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ 是参数 p 的充分统计量。证明 由于 $X_i \sim B(1, p)$, 易知 $n\bar{X} = \sum_{i=1}^n X_i \sim B(n, p)$, 即有 $P\{n\bar{X} = k\} = C_n^k p^k (1-p)^{n-k}$, $k = 0, 1, \dots, n$ 设 $(X_1, \dots, X_n)^T$ 为样本值, 其中 $x_i = 0$ 或 1 。当已知 $\sum_{i=1}^n x_i = k$

即 $\bar{X} = \frac{k}{n}$ 时, 样本 $(X_1, \dots, X_n)^T$ 的条件概率

$$\begin{aligned} & P\left\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid \bar{X} = \frac{k}{n}\right\} \\ &= \frac{P\left\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, \bar{X} = \frac{k}{n}\right\}}{P\left\{\bar{X} = \frac{k}{n}\right\}} \\ &= \begin{cases} \frac{P\left\{X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}, X_n = k - \sum_{i=1}^{n-1} x_i\right\}}{P\{n\bar{X} = k\}} & \sum_{i=1}^n x_i = k \\ 0 & \sum_{i=1}^n x_i \neq k \end{cases} \\ &= \begin{cases} \frac{p^{\sum x_i} (1-p)^{n-\sum x_i}}{C_n^k p^k (1-p)^{n-k}}, & \sum_{i=1}^n x_i = k \\ 0, & \sum_{i=1}^n x_i \neq k \end{cases} \\ &= \begin{cases} \frac{1}{C_n^k}, & \sum_{i=1}^n x_i = k \\ 0, & \sum_{i=1}^n x_i \neq k \end{cases} \end{aligned}$$

与 p 无关, 故为充分统计量

8.0.3 指数分布族

Definition 8.0.5 — 单参数指教分布族的定义. 设总体 X 或 $X|\theta$ 的分布密度 $p(\mathbf{x}|\theta)$ 为:

$$p(x|\theta) = g(x)h(\theta)\exp\{t(x)\varphi(\theta)\}$$

其中 g, h, t, φ 为一般的函数记号, 则称 $p(\mathbf{x}|\theta)$ 属于指数分布族。

■ **Example 8.6** 正态分布 $N(\mu, \sigma^2)$ 当 σ^2 已知时,

$$\begin{aligned} p(x|\mu) &= (2\pi\sigma^2)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \left[(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}\right] \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{\frac{x\mu}{\sigma^2}\right\} \end{aligned}$$

它属于指数分布族. 显然, $N(\mu, \sigma^2)$ 当 μ 已知, σ^2 未知时, 是单参数 σ^2 的指教分布族.

Definition 8.0.6 — 指数分布族中的参数的共轭先验密度. 若 $X|\theta$ 的分布属于指数分布族 $p(x|\theta) = g(x)h(\theta)\exp\{t(x)\varphi(\theta)\}$ 故似然函数 $l((\theta|\vec{x})$ (其中 $\vec{x} = (x_1, x_2, \dots, x_n)$) 为 i.i.d. 的样本) 为

$$l(\theta|\vec{x}) \propto [h(\theta)]^n \exp\left\{\sum t(x_i)\varphi(\theta)\right\}$$

共轭先验密度 $\pi(\theta)$ 为: $\pi(\theta) \propto [h(\theta)]^\gamma \exp\{\tau\varphi(\theta)\}$ 即指教分布族的共轭族 Π 为:

$$\Pi = [h(\theta)]^\gamma \exp\{\tau\varphi(\theta)\}$$

Definition 8.0.7 — 两参数指热分布族及其共轭分布组. 设 $X \sim p(x|\theta, \varphi)$, θ, φ 皆未知, 若 $p(x|\theta, \varphi) = g(x)h(\theta, \varphi)\exp\{t(x)\psi(\theta, \varphi) + u(x)\chi(\theta, \varphi)\}$ 称 X 的分布属于两参数指教族。若 $\vec{x} = (x_1, x_2, \dots, x_n)$ 为 i.i.d. 的样本, X 的分布为两参数的指教分布族, 似然函数 $l(\theta, \varphi|\vec{x})$ 可写为: $l(\theta, \varphi|\vec{x}) \propto [h(\theta, \varphi)]^n \exp\{\sum [t(x_i)\psi(\theta, \varphi) + u(x_i)\chi(\theta, \varphi)]\}$ 共轭先验族 Π 为以下形式:

$$\pi(\theta, \varphi) \propto [h(\theta, \varphi)]^\gamma \exp\{\alpha\psi(\theta, \varphi) + \beta\chi(\theta, \varphi)\}$$

Theorem 8.0.6 两参数 θ, φ 的后验及先验的一种求法

设 $\pi(\theta, \varphi|x)$ 为已知 i.i.d 的样本 x 的关于 θ, φ 的后验, X 的模型为 $p(x|\theta, \varphi)$, θ, φ 皆未知, 则

$$\pi(\theta, \varphi|\vec{x}) = \pi(\varphi|\vec{x})\pi(\theta|\varphi, \vec{x})$$

相应的, 也有类似的先验公式, 设 $\pi(\theta, \varphi)$ 为 θ, φ 的先验密度, 则:

$$\pi(\theta, \varphi) = \pi(\varphi)\pi(\theta|\varphi)$$

8.1 贝叶斯推断

Theorem 8.1.1 贝叶斯的特点: 统计推断的操作都是从后验分布中提取信息。

Definition 8.1.1 — 条件观点. 后验分布 $\pi(\theta|x)$ 是在样本 x 给定下 θ 的条件分布, 基于后验分布的统计推断就意味着只考虑已出现的数据(样本观察值), 而认为未出现的数据与推断无关, 这重要的观点被称为“条件观点”, 基于这种观点提出的统计推断方法被称为条件方法, 它与大家熟悉的频率方法之间还是有很大的差别。

Definition 8.1.2 — 参数估计. 参数估计: 寻取后验分布的某个特征量.

1. 使后验密度 $\pi(\theta|x)$ 达到最大的值 $\hat{\theta}_{MD}$ 称为最大后验估计;
2. 后验分布的中位数 $\hat{\theta}_{Me}$ 称为 θ 的后验中位数估计
3. 后验分布的期望值 $\hat{\theta}_E$ 称为 θ 的后验期望估计
4. 正态分布三种估计重合

这三个估计也都称为 θ 的贝叶斯估计, 记为 $\hat{\theta}_B$, 在不引起混乱时也记为 $\hat{\theta}$

Theorem 8.1.2

■ **Example 8.7** 作为一个数值例子, 我们考虑对一个儿童做智力测验, 设测验结果 $X \sim N(\theta|100, 225)$, 其中 θ 在心理学中定义为该儿童的智商, 根据过去多次测验, 可设 $\theta \sim N(100, 225)$, 应用上述方法, 在 $n = 1$ 时, 可得在给定 $X=x$ 条件下, 该儿童智商 θ 的后验分布是正态分布 $N(\mu_1, \sigma_1^2)$, 其中

$$\begin{aligned}\mu_1 &= \frac{100 + 100 + 225x}{100 + 225} = \frac{400 + 9x}{13} \\ \sigma_1^2 &= \frac{100 \times 225}{100 + 225} = \frac{900}{13} = 69.23 = (8.32)^2\end{aligned}$$

假如该儿童这次测验得分为 115 分, 则他的智商的贝叶斯估计为

$$\hat{\theta}_B = \frac{400 + 9 \times 115}{13} = 110.38$$

■ **Example 8.8** 为估计不合格品率 θ , 今从一批产品中随机抽取 n 件, 其中不合格品数 X 服从二项分布 $b(n, \theta)$, 国内外诸多文献上都取贝塔分布 $Be(\alpha, \beta)$ 作为 θ 的先验分布, 它的众数为 $(\alpha - 1)/(\alpha + \beta - 2)$, 它的期望为 $\alpha/(\alpha + \beta)$, 这里假设 α 与 β 已知, 由共轭先验分布可知, 这时 θ 的后验分布仍为贝塔分布 $Be(\alpha + x, \beta + n - x)$, 这时 θ 的最大后验估计 $\hat{\theta}_{MD}$ 和后验期望估计 $\hat{\theta}_E$ 分别为

$$\hat{\theta}_{MD} = \frac{\alpha + x - 1}{\alpha + \beta + n - 2} \quad \hat{\theta}_E = \frac{\alpha + x}{\alpha + \beta + n}$$

Theorem 8.1.3 在二项分布场合, θ 的最大后验估计就是经典统计中的极大似然估计. θ 的后验期望值估计 $\hat{\theta}_E$ 要比最大后验估量 $\hat{\theta}_{MD}$ 更合适一些.

Theorem 8.1.4 求最大后验估计常用方法: 对参数 θ 求导, 判断单调性, 求极值。

Theorem 8.1.5 为估计不合格品率 θ , 今从一批产品中随机抽取 n 件, 其中不合格品数 X 服从二项分布 $b(n, \theta)$, 国内外诸多文献上都取贝塔分布 $Be(\alpha, \beta)$ 作为 θ 的先验分布, 它的众数为 $(\alpha - 1)/(\alpha + \beta - 2)$, 它的期望为 $\alpha/(\alpha + \beta)$, 这里假设 α 与 β 已知, 由共轭先验分布可知(见例 1.3.4), 这时 θ 的后验分布仍为贝塔分布 $Be(\alpha + x, \beta + n - x)$, 这时 θ 的最大后验估计 $\hat{\theta}_{MD}$ 和后验期望估计 $\hat{\theta}_E$ 分别为

$$\begin{aligned}\hat{\theta}_{MD} &= \frac{\alpha + x - 1}{\alpha + \beta + n - 2} \\ \hat{\theta}_E &= \frac{\alpha + x}{\alpha + \beta + n}\end{aligned}$$

■ **Example 8.9** 例 2.2.3 设 x 是来自如下指数分布的一个观察值。

$$p(x|\theta) = e^{-(x-\theta)}, x \geq \theta$$

又取柯西分布作为 θ 的先验分布, 即

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}, -\infty < \theta < \infty$$

这时可得后验密度

$$\pi(\theta|x) = \frac{e^{-(x-\theta)}}{m(x)(1+\theta^2)\pi}, \theta \leq x$$

为了寻找 θ 的最大后验估计 $\hat{\theta}_{MD}$, 我们对后验密度使用微分法, 可得

$$\begin{aligned}\frac{d}{d\theta} \pi(\theta|x) &= \frac{e^{-x}}{m(x)\pi} \left[\frac{e^\theta}{1+\theta^2} - \frac{2\theta e^\theta}{(1+\theta^2)^2} \right] \\ &= \frac{e^{-x} e^\theta (\theta - 1)^2}{m(x)(1+\theta^2)^2 \pi} \geq 0\end{aligned}$$

由于 $\pi(\theta|x)$ 的非减性, 考虑到 θ 的取值不能超过 x , 故 θ 的最大后验估计应为 $\hat{\theta}_{MD} = x$

Definition 8.1.3 设参数 θ 的后验分布 $\pi(\theta|x)$, 贝叶斯估计为 $\hat{\theta}$, 则 $(\theta - \hat{\theta})^2$ 的后验期望

$$MSE(\hat{\theta}|x) = E^{\theta|x}(\theta - \hat{\theta})^2$$

称为 $\hat{\theta}$ 的后验均方差, 而其平方根 $[MSE(\hat{\theta}|x)]^{1/2}$ 称为 $\hat{\theta}$ 的后验标准误, 其中符号 $E^{\theta|x}$ 表示用条件分布 $\pi(\theta|x)$ 求期望, 当 $\hat{\theta}$ 为 θ 的后验期望 $\hat{\theta}_E = E(\theta|x)$ 时, 则

$$MSE(\hat{\theta}_E|x) = E^{\theta|x}(\theta - \hat{\theta}_E)^2 = \text{Var}(\theta|x)$$

称为后验方差, 其平方根 $[\text{Var}(\theta|x)]^{1/2}$ 称为后验标准差。后验均方差与后验方差有如下关

系：

$$\begin{aligned}\text{MSE}(\hat{\theta}|x) &= E^{\theta|x}(\theta - \hat{\theta})^2 \\ &= E^{\theta|x}[(\theta - \hat{\theta}_E) + (\hat{\theta}_E - \hat{\theta})]^2 \\ &= \text{Var}(\theta|x) + (\hat{\theta}_E - \hat{\theta})^2\end{aligned}$$

■ **Example 8.10** 设一批产品的不合格品率为 θ , 检查是一个接一个地进行, 直到发现第一个不合格品停止检查, 若设 X 为发现第一个不合格品时已检查的产品数, 则 X 服从几何分布, 其分布列为

$$P(X = x|\theta) = \theta(1 - \theta)^{x-1}, x = 1, 2, \dots$$

假如其中参数 θ 只能为 $1/4, 2/4$ 和 $3/4$ 三个值, 并以相同概率取这三个值, 如今只获得一个样本观察值 $x = 3$, 要求 θ 的最大后验估计 $\hat{\theta}_{MD}$, 并计算它的误差。在这个问题中, θ 的先验分布为

$$P\left(\theta = \frac{i}{4}\right) = \frac{1}{3}, i = 1, 2, 3$$

在 θ 给定下, $X = 3$ 的条件概率为

$$P(X = 3|\theta) = \theta(1 - \theta)^2$$

于是联合概率为

$$P\left(X = 3, \theta = \frac{i}{4}\right) = \frac{1}{3} \cdot \frac{i}{4} \cdot \left(1 - \frac{i}{4}\right)^2$$

$X = 3$ 的无条件概率为

$$P(X = 3) = \frac{1}{3} \left[\frac{1}{4} \left(\frac{3}{4}\right)^2 + \frac{2}{4} \left(\frac{2}{4}\right)^2 + \frac{3}{4} \left(\frac{1}{4}\right)^2 \right] = \frac{5}{48}$$

于是在 $X = 3$ 条件下, θ 的后验分布列为

$$P(\theta = i/4|X = 3) = \frac{P(X = 3, \theta = i/4)}{P(X = 3)} = \frac{4i}{5} \left(1 - \frac{i}{4}\right)^2, i = 1, 2, 3$$

或
$$\frac{\theta}{P(\theta = i/4|X = 3)} \mid \begin{array}{ccc} 1/4 & 2/4 & 3/4 \\ 9/20 & 8/20 & 3/20 \end{array}$$
 可以看出, θ 的最大后验估计 $\hat{\theta}_{MD} = 1/4$ 为了计算此贝叶斯估计的误差, 我们先计算上述后验分布的均值与方差, 容易算得

$$E(\theta|X = 3) = 17/4 \quad E(\theta^2|X = 3) = 17/80$$

于是后验方差 $\text{Var}(\theta|X = 3) = 17/80 - (17/40)^2 = 51/1600$, 利用前述公式, 最大后验估计 $\hat{\theta}_{MD}$ 的后验均方差为

$$\begin{aligned}\text{MSE}(\hat{\theta}|X = 3) &= \text{Var}(\theta|X = 3) + (\hat{\theta}_{MD} - \bar{\theta})^2 \\ &= \frac{51}{1600} + \left(\frac{1}{4} - \frac{17}{40}\right)^2 = \frac{1}{16}\end{aligned}$$

而其后验标准误为 $[\text{MSE}(\hat{\theta}|X = 3)]^{1/2} = 1/4$

■ **Example 8.11** 在选用共轭先验分布下, 不合格品率 θ 的后验分布为贝塔分布, 它的后验方差为

$$\text{Var}(\theta|x) = \frac{(\alpha+x)(b+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

其中 n 为样本量, x 为样本中不合格品数, α 与 β 为先验分布中的二个超参数。若取 $\alpha = \beta = 1$, 则其后验方差为

$$\text{Var}(\theta|x) = \frac{(x+1)(n-x+1)}{(n+2)^2(n+3)}$$

这时 θ 的后验期望估计 $\hat{\theta}_E$ 和最大后验估计 $\hat{\theta}_{MD}$ 分别为

$$\hat{\theta}_E = \frac{x+1}{n+2}, \quad \hat{\theta}_{MD} = \frac{x}{n}$$

显然, $\hat{\theta}_E$ 的后验均方差就是上述 $\text{Var}(\theta|x)$, $\hat{\theta}_{MD}$ 的后验均方差为

$$\text{MSE}(\hat{\theta}_{MD}|x) = \frac{(x+1)(n-x+1)}{(n+2)(n+3)} + \left(\frac{x+1}{n+2} - \frac{x}{n} \right)^2$$

8.1.1 可信区间

当参数 θ 的后验分布 $\pi(\theta|x)$ 获得以后, 立即可计算 θ 落在某区间 $[a,b]$ 内的后验概率, 譬如为 $1-\alpha$, 即

$$P(a \leq \theta \leq b|x) = 1-\alpha$$

反之, 若给定概率 $1-\alpha$, 要找一个区间 $[a,b]$, 使上式成立, 这样求得的区间就是 θ 的贝叶斯区间估计, 又称为可信区间, 这是在 θ 为连续随机变量场合, 若 θ 为离散随机变量, 对给定的概率 $1-\alpha$, 满足上式的区间 $[a,b]$ 不一定存在, 这时只有略微放大上式左端概率, 才能找到 a 与 b , 使得

$$P(a \leq \theta \leq b|x) > 1-\alpha$$

这样的区间也是 θ 的贝叶斯可信区间, 它的一般定义如下。

Definition 8.1.4 — 贝叶斯可信区间. 设参数 θ 的后验分布为 $\pi(\theta|x)$, 对给定的样本 x 和概率 $1-\alpha$ ($0 < \alpha < 1$), 若存在这样的二个统计量 $\hat{\theta}_L = \hat{\theta}_L(x)$ 与 $\hat{\theta}_U = \hat{\theta}_U(x)$, 使得

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U|x) \geq 1-\alpha$$

则称区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 为参数 θ 的可信水平为 $1-\alpha$ 贝叶斯可信区间, 或简称为 θ 的 $1-\alpha$ 可信区间。而满足

$$P(\theta \geq \hat{\theta}_L|x) \geq 1-\alpha$$

的 $\hat{\theta}_L$ 称为 θ 的 $1-\alpha$ (单侧) 可信下限。满足

$$P(\theta \leq \hat{\theta}_U|x) \geq 1-\alpha$$

的 $\hat{\theta}_U$ 称为 θ 的 $1-\alpha$ (单侧) 可信上限。

Theorem 8.1.6 正态的计算公式:

$$P(\mu_1 - \sigma_1 \alpha_{-\alpha/2} \leq \theta \leq \mu_1 + \sigma_1 \alpha_{1-\alpha/2}) = 1 - \alpha$$

其中 $\alpha_{1-\alpha/2}$ 是标准正态分布的 $1 - \alpha/2$ 分位数。

最理想的可信区间应是区间长度最短, 这只要把具有最大后验密度的点都包含在区间内, 而在区间外的点上的后验密度函数值不超过区间内的后验密度函数值, 这样的区间称为最大后验密度 (Highest Posterior Density, 简称 HPD) 可信区间, 它的一般定义如下:

Definition 8.1.5 设参数 θ 的后验密度为 $\pi(\theta|x)$, 对给定的概率 $1 - \alpha (0 < \alpha < 1)$ 若在直线上存在这样一个子集 C , 满足下列二个条件: $1^{\circ} P(C|x) = 1 - \alpha$

8.1.2 假设检验

在贝叶斯统计中处理假设检验问题是直截了当的, 在获得后验分布 $\pi(\theta|x)$ 后, 即可计算二个假设 H_0 与 H_1 的后验概率

$$\alpha_i = P(\Theta_i|x) d\theta, i = 0, 1$$

然后比较 α_0 与 α_1 的大小, 当后验概率比 (或称后验机会比) $\alpha_0/\alpha_1 > 1$ 时接受 H_0 ; 当 $\alpha_0/\alpha_1 < 1$ 时接受 H_1 ; 当 $\alpha_0/\alpha_1 \approx 1$ 时, 不宜做判断, 尚而进一步抽样或进一步搜集先验信息。

Definition 8.1.6 设两个假设 Θ_0 与 Θ_1 的先验概率分别为 π_0 与 π_1 , 后验概率分别为 α_0 与 α_1 , 则称

$$B^\pi(x) = \frac{\text{后验机会比}}{\text{先验机会比}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

为贝叶斯因子。贝叶斯因子 $B^\pi(x)$ 是数据 x 支持 Θ_0 的程度。

Theorem 8.1.7 — 简单假设 $\Theta_0 = \{\theta_0\}$ 对简单假设 $\Theta_1 = \{\theta_1\}$. 在这种场合, 这两种简单假设的后验概率分别为

$$\alpha_0 = \frac{\pi_0 p(x|\theta_0)}{\pi_0 p(x|\theta_0) + \pi_1 p(x|\theta_1)}, \alpha_1 = \frac{\pi_1 p(x|\theta_1)}{\pi_0 p(x|\theta_0) + \pi_1 p(x|\theta_1)}$$

其中 $p(x|\theta)$ 为样本的分布, 这时后验机会比为

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 p(x|\theta_0)}{\pi_1 p(x|\theta_1)}$$

欲要拒绝原假设 $\Theta_0 = \{\theta_0\}$, 则必须有 $\alpha_0/\alpha_1 < 1$, 或

$$\frac{p(x|\theta_1)}{p(x|\theta_0)} > \frac{\pi_0}{\pi_1}$$

Theorem 8.1.8 — 复杂假设 Θ_0 对复杂假设 Θ_1 . 我们把先验分布 $\pi(\theta)$ 限制在 $\Theta_0 \cup \Theta_1$ 上, 并

令

$$\begin{aligned} g_0(\theta) &\propto \pi(\theta)I_{\Theta_0}(\theta) \\ g_1(\theta) &\propto \pi(\theta)I_{\Theta_1}(\theta) \end{aligned}$$

于是先验分布可改写为

$$\begin{aligned} \pi(\theta) &= \pi_0g_0(\theta) + \pi_1g_1(\theta), \theta \in \Theta_0 \cup \Theta_1 \\ &= \begin{cases} \pi_0g_0(\theta), \theta \in \Theta_0 \\ \pi_1g_1(\theta), \theta \in \Theta_1 \end{cases} \end{aligned}$$

其中 π_0 与 π_1 分别是 Θ_0 与 Θ_1 上的先验概率, g_0 与 g_1 分别是 Θ_0 与 Θ_1 上的概率密度函数。在这些记号下, 后验概率比为

$$\frac{\alpha_0}{\alpha_1} = \frac{\int_{\Theta_0} p(x|\theta)\pi_0g_0(\theta)d\theta}{\int_{\Theta_1} p(x|\theta)\pi_1g_1(\theta)d\theta}$$

于是贝叶斯因子可表示为

$$B^\pi(x) = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} = \frac{\int_{\Theta_0} p(x|\theta)g_0(\theta)d\theta}{\int_{\Theta_1} p(x|\theta)g_1(\theta)d\theta} = \frac{m_0(x)}{m_1(x)}$$

8.1.3 简单原假设对复杂的备择假设

Theorem 8.1.9 合理的原假设与备择假设应是

$$H_0 : \theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon], H_1 : \theta \notin [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$$

有效的方法是对 $\theta = \theta_0$ 给一个正概率 π_0 , 而对 $\theta \neq \theta_0$ 给一个加权密度 $\pi g_1(\theta)$, 即 θ 的先验密度为

$$\pi(\theta) = \pi_0 I_{\theta_0}(\theta) + \pi_1 g_1(\theta)$$

其中 $I_{(\theta_0)}$ 为 $\theta = \theta_0$ 的示性函数, $\pi_1 = 1 - \pi_0$, $g_1(\theta)$ 为 $\theta \neq \theta_0$ 上的一个正常密度函数这里可把 π_0 看作近似的实际假设 $H_0 : \theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ 上的先验概率, 如此的先验分布是由离散和连续两部分组合而成。设样本分布为 $p(x|\theta)$, 利用上述先验分布容易获得样本 x 的边缘分布

$$\begin{aligned} m(x) &= \int_{\theta} p(x|\theta)\pi(\theta)d\theta \\ &= \pi_0 p(x|\theta_0) + \pi_1 m_1(x) \end{aligned}$$

其中 (第一个等号可作为符号理解)

$$m_1(x) = \int_{\theta \neq \theta_0} p(x|\theta)g_1(\theta)d\theta$$

(2. 4. 6) 从而简单原假设与复杂备择假设 (记为 $\Theta_1 = \{\theta \neq \theta_0\}$) 的后验概率分别为

$$\begin{aligned} \pi(\Theta_0|x) &= \pi_0 p(x|\theta_0)/m(x) \\ \pi(\Theta_1|x) &= \pi_1 m_1(x)/m(x) \end{aligned}$$

后验机会比为

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0}{\pi_1} \frac{p(x|\theta_0)}{m_1(x)}$$

从而贝叶斯因子为

$$B^\pi(x) = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{p(x|\theta_0)}{m_1(x)}$$

这一简单表达式要比后验概率计算容易很多, 故实际中常常是先计算 $B^\pi(x)$, 然后再计算 $\pi(\Theta_0|x)$, 因为由贝叶斯因子的定义和 $\alpha_0 + \alpha_1 = 1$ 可推得

$$\pi(\Theta_0|x) = \left[1 + \frac{1 - \pi_0}{\pi_0} \frac{1}{B^\pi(x)} \right]^{-1}$$

8.1.4 预测

Theorem 8.1.10 — 无观察数据. 设随机变量 $X \sim p(x|\theta)$, 在无 X 的观察数据时, 利用先验分布 $\pi(\theta)$ 容易获得未知的、但可观察的数据 x 的分布

$$m(x) = \int_{\theta} p(x|\theta) \pi(\theta) d\theta$$

这个分布常被称为 X 的边缘分布, 但它还有一个更富于内含的名称是“先验预测分布”, 这里的先验是指对过去的数据没有要求, 预测是指它是可观察量的分布, 由此先验预测分布就可从中提取有用信息作出未来观察值的预测值或未来观察值的预测区间, 譬如用 $m(x)$ 的期望值. 中位数或众数作为预测值, 或确定 90% 的预测区间 $[a,b]$, 使得

$$P^x(a \leq X \leq b) = 0.90$$

其中 P^x 指用分布 $m(x)$ 计算概率。

Theorem 8.1.11 — 有观察数据. 另一种情况是: 在有 X 的观察数据 $x = (x_1, \dots, x_n)$ 时, 利用后验分在 $\pi(\theta|x)$ 容易获得未知观察值的分布, 如要预测同一总体 $p(x|\theta)$ 的未来观察值, 则在

$$m(x|x) = \int_{\theta} p(x|\theta) \pi(\theta|x) d\theta$$

如要预测另一总体 $g(z|\theta)$ 的未来观察值, 则有

$$m(z|x) = \int_{\theta} g(z|\theta) \pi(\theta|x) d\theta$$

这里 $m(x|x)$ 或 $m(z|x)$ 都称为“后验预测分布”, 有此后验预测分布后, 类似地从中提取有用信息作出未来观察值的预测值或预测区间, 譬如用 $m(z|x)$ 的期望值, 中位数或众数作为 z 的预测值, 或确定 90% 的预测区间 $[a,b]$, 使得

$$P^{z|x}(a \leq Z \leq b|x) = 0.90$$

其中 $P^{z|x}$ 是指用预测分布 $m(z|x)$ 计算概率。

Theorem 8.1.12 这个问题的一般提法是: 在 n 次相互独立的贝努里试验成功了 x 次, 现要对未来的 k 次相互独立的贝努里试验中成功次数 z 作出预测, 这里的贝努里试验中的成功射击的命中等。若设成功概率为 θ , 则样本 x 的似然函数为.

$$L(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

若取 θ 的共轭先验分布 $Be(\alpha, \beta)$, 则其后验密度为

$$\pi(\theta|x) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}$$

新的样本 z 的似然函数为

$$L(z|\theta) = \binom{k}{z} \theta^z (1-\theta)^{k-z}$$

于是在给定 x 时, z 的后验预测分布为

$$\begin{aligned} m(z|x) &= \int_0^1 \binom{k}{z} \theta^z (1-\theta)^{k-z} \pi(\theta|x) d\theta \\ &= \binom{k}{z} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \int_0^1 \theta^{z+x+\alpha-1} (1-\theta)^{k-z+n-x+\beta-1} d\theta \\ &= \binom{k}{z} \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \frac{\Gamma(z+x+\alpha)\Gamma(k-z+n-x+\beta)}{\Gamma(n+k+\alpha+\beta)} \end{aligned}$$

8.1.5 似然原理

Theorem 8.1.13 — 似然原理. 1. 有了观察值 x 之后, 在关于 θ 的推断和决策时, 所有与试验有关的 θ 信息均被包含在似然函数 $L(\theta)$ 之中。
2. 如果有两个似然函数是成比例的, 并且比例常数与 θ 无关, 则他们关于 θ 含有相同的信息.

8.1.6 习题

Exercise 8.1 — 点估计. 设不合格品率的先验分在为贝塔分布 $Be(5, 10)$, 在下血抽样信息下逐次求 θ 的最大后验估计与后验期望估计。抽样信息: 随机抽检 20 个产品, 发现 17 个合格品. ■

Proof. 最大后验估计: 首先求的后验分布, 再根据单调性等信息, 求得最大值。首先知道总体信息是贝塔分布, 以及样本信息是二项分布, 样本量是三个次品。根据后验概率分布的计算公式得

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta)$$

$$\begin{aligned} P(x|\theta) &= P\{X=1, X=1, X=1, X=0, \dots\} \\ &= \theta^3 (1-\theta)^{17} \end{aligned}$$

$$\begin{aligned} \pi(\theta|x) &\propto p(x|\theta)\cdot\pi(\theta) \\ &= \frac{\Gamma(15)}{\Gamma(5)\Gamma(10)} \cdot \theta^4 \cdot (1-\theta)^9 \theta^3 \cdot (1-\theta)^{17} \\ &= A \cdot \theta^7 (1-\theta)^{26} \end{aligned}$$

$$\begin{aligned}
\frac{d\pi(\theta|x)}{d\theta} &= 7\theta^6(1-\theta)^{26} - 26\theta^7(1-\theta)^{25} \\
&= \theta^6(1-\theta)^{25}[7-7\theta-26\theta] \\
&= \theta^6(1-\theta)^{25}(7-33\theta) = 0 \\
\Rightarrow \theta_{max} &= \frac{7}{33}
\end{aligned}$$

首先计算出后验概率的常数 A

$$\int_{-\infty}^{+\infty} A\theta^7(1-\theta)^{26}d\theta = 1$$

根据 *Beta* 分布的定义, 可知

$$A = \frac{\Gamma(35)}{\Gamma(8)\Gamma(27)}$$

, 故后验期望估计为

$$\hat{\theta}_E = \frac{\alpha}{\alpha + \beta} = \frac{8}{35}$$

■

Exercise 8.2 设 x_1, x_2, \dots, x_n 是来自如下指数分布的一个观察值.

$$p(x|\theta) = e^{-(x-\theta)}, x \geq \theta$$

又取柯西分布作为 θ 的先验分布, 即

$$\theta \sim \pi(\theta) = \frac{1}{\pi(1+\theta^2)}, -\infty < \theta < +\infty$$

求 θ 的最大后验估计。

■

Proof.

$$p(x|\theta) = \prod_{i=1}^n e^{-(x_i-\theta)} \forall x_i \geq 0 \quad (8.3)$$

$$\pi(\theta|x) = e^{n\theta - \sum x_i} \cdot \frac{1}{\pi(1+\theta^2)} \quad (8.4)$$

注意上式中的后验概率密度函数的取值范围是

$$\theta \leq \min\{x_1, \dots, x_n\}$$

对上式子对 θ 求导,

$$\begin{aligned}
\frac{d\pi(\theta|x)}{d\theta} &= \frac{1}{\pi(1+\theta^2)} e^{n\theta - \sum x_i} \cdot n - e^{n\theta - \sum x_i} \left(\frac{1}{\pi(1+\theta^2)} \right)^2 2\theta \\
&= \left[\frac{1}{\pi(1+\theta^2)} \right]^2 e^{n\theta - \sum x_i} \cdot [n(\pi + \pi\theta^2) - 2\theta] \\
&\geq 0
\end{aligned}$$

故 θ 最大后验估计为 $\min\{x_1, \dots, x_n\}$

■

Exercise 8.3 对正态分布 $N(\theta, 1)$ 作观察, 获得三个观察值: 2,3,4, 若 θ 的先验分在为 $N(3, 1)$, 求 θ 的 0.95 可信区间。 ■

Proof. 套用正态分布的公式

$$\begin{aligned}\mu &= 3 \quad \tau^2 = 1 \quad \sigma^2 = 1 \quad \sigma_0^2 = \frac{1}{3} \\ \hat{\theta}_B &= \frac{1}{3+1} \cdot 3 + \frac{3}{3+1} \cdot 3 \quad \sigma_1^2 = \frac{1}{4}\end{aligned}$$

可信区间为

$$\begin{aligned}[3 - 0.5 \times 1.96, 3 + 0.5 \times 1.96] \\ = [2.02, 3.98]\end{aligned}$$



Exercise 8.4 对正态分在 $N(\theta, 1)$ 作观察, 获得三个观察值: 2,3,4, 若 θ 的先验分在为 $N(2, 1)$, 针对检验 $H_0: \theta \leq 3, H_1: \theta > 3$, 试做出判断, 并求贝叶斯因子。 ■

Proof. 与上一题步骤一样, 得到后验分布为

$$N\left(\frac{11}{4}, \frac{1}{4}\right)$$

$$P(x \leq 3) = \Phi\left(\frac{3 - \frac{11}{4}}{a\sqrt{5}}\right) = 0.6915$$

$$P(x \geq 3) = 0.3085$$

$$\frac{\alpha_0}{\alpha_1} > 1$$

接受 H_0 , 拒绝 H_1

$$\pi_0 = 0.8413 \quad \pi_1 = 0.1587$$

$$\Rightarrow B^\pi(x) = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = 0.0582$$



Exercise 8.5 某同学抛拼硬币 15 次, 正面出现的次数是 15 次, 正面出现的几率 θ 的共轭先验分布为 $Be(3, 2)$, 请问第 16 次拼硬币, 硬币出现正面还是反面, 请做出你推断. ■

Proof. 总体信息: 15 次伯努利实验, 知道样本信息, 知道先验信息。首先求的后验分布, 套用之前的公式得到后验为 $Be(18, 2)$,

$$m(z|x) = \binom{1}{z} \frac{\Gamma(20)}{\Gamma(18)\Gamma(2)} \frac{\Gamma(z+18)\Gamma(3-z)}{\Gamma(21)}$$

$$m(1|x) = A \cdot \Gamma(19)\Gamma(2) = A \cdot 18!$$

$$m(0|x) = A \cdot \Gamma(18)\Gamma(3) = A \cdot 17! \cdot 2$$

第十六次为正面



8.2 先验分布的确定

Definition 8.2.1 — 主观概率. 贝叶斯学派认为: 一个事件的概率是人们根据经验对该事件发生可能性所给出个人信念。这样给出的概率称为主观概率。

Theorem 8.2.1 确定的主观概率都必须满足概率的三条公理, 即

1. 非负性公理: 对任一事件 $A, 0 \leq P(A) \leq 1$
2. 正则性公理: 必然事件的概率为 1
3. 可列可加性公理: 对可列个互不相容的事件 A_1, A_2, \dots , 有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

当发现所确定的主观概率与这三条公理及其推出的性质有不和谐时, 必须立即修正, 直到和谐为止。这时给出的主观概率才能称得上概率。

8.2.1 利用先验信息确定先验分布

接下去介绍连续型确定先验分布的方法:

直方图法

有数据使用直方图:

1. 将参数空间分为一些小区间
2. 在每个小区间上决定主观概率
3. 绘制频率直方图
4. 在直方图上作一条光滑的曲线, 此曲线就是 $\pi(\theta)$ 。在做光滑曲线时, 尽量在每个小区间上使用得曲线下的面积与直方图的面积相等。

选定先验密度函数形式再估计其超参数

Theorem 8.2.2 这个方法的要点如下 (1) 根据先验信息选定 θ 的先验密度函数 $\pi(\theta)$ 的形式, 如选其共轭先验分布。(2) 当先验分布中含有未知参数(称为超参数)时, 譬如 $\pi(\theta) = \pi(\theta; \alpha, \beta)$, 给出超参数 α, β 的估计值使 $\pi(\theta; \alpha, \hat{\beta})$ 最接近先验信息



如果有两个甚至多个先验分布都满足给定的先验信息, 则要看情况选择: ((1) 假如这两个先验分在差异不大, 对后验分布影响也不大, 则可任选一个; ((2) 如果我们面临着两个差异极大的先验分布可供选择时, 一定要根据实际情况慎重选择。

定分度法与变分度法

Definition 8.2.2 定分变法和变分度法都是通过专家咨询得到各种主观概率的方法, 然后加工整理成累积分布的概率曲线的方法。专家的信誉要高, 经验要多。一般决策者更加愿意使用变分度法

(1) 定分度法: 把参数可能取值的区间分为长度相等的小区间, 每次在每个小区间上请专家给出主观概率。(2) 变分度法: 该法是把参数可能取值的区间逐次分为机会相等的两个小区间, 这里的分点由专家确定。

相对似然法

Definition 8.2.3 对 Θ 中的各种点的直观“似然”进行比较, 再按确定了的值画图, 即可得到先验密度草图。使用范围: 此法大多用于 $\Theta \in (-\infty, +\infty)$ 的有限子区间的情形。

8.2.2 利用边缘分布 $m(x)$ 确定先验密度

Definition 8.2.4 — 边缘分布 $m(x)$. 设总体 X 的密度函数为 $p(x|\theta)$ 它含有未知参数 θ 若 θ 的先验分布选用形式已知的密度函数 $\pi(\theta)$, 则可算得 X 的边缘分布 (即无条件分布):

$$m(x) = \begin{cases} \int_{\Theta} p(x|\theta)\pi(\theta)d\theta, & \text{当 } \theta \text{ 为连续时} \\ \sum_{\theta \in \Theta} p(x|\theta)\pi(\theta), & \text{当 } \theta \text{ 为离散时} \end{cases}$$

当先验分布含有未知参数, 如 $\pi(\theta) = \pi(\theta|\lambda)$, 则边缘分布 $m(x)$ 依赖于 λ , 可记为 $m(x|\lambda)$

■ **Example 8.12** 设给定 θ 时总体 X 服从 $X \sim N(\theta, \sigma^2)$, σ^2 已知. θ 的先验分布选取为正态分布, $\theta \sim N(\mu_\pi, \sigma_\pi^2)$ 求 X 的边缘分布 $m(x)$

Proof.

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\}$$

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\sigma_\pi} \exp\left\{-\frac{1}{2\sigma_\pi^2}(\theta-\mu_\pi)^2\right\}$$

于是边缘分布

$$m(x) = \frac{1}{2\pi\sigma\sigma_\pi} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_\pi)^2}{\sigma_\pi^2}\right]\right\} d\theta$$

若令

$$A = \frac{1}{\sigma^2} + \frac{1}{\sigma_\pi^2}, B = \frac{x}{\sigma^2} + \frac{\mu_\pi}{\sigma_\pi^2}, C = \frac{x^2}{\sigma^2} + \frac{\mu_\pi^2}{\sigma_\pi^2}$$

则可算得

$$\begin{aligned} m(x) &= \frac{1}{2\pi\sigma\sigma_\pi} \exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \times \sqrt{\frac{2\pi}{A}} \\ &= \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_x^2)}} \exp\left\{-\frac{AC - B^2}{2A}\right\} \end{aligned}$$

由于

$$AC - B^2 = (x - \mu_x)^2 / \sigma^2 \sigma_\pi^2$$

$$\begin{aligned} m(x) &= \int_{-\infty}^{+\infty} f(x|\theta)\pi(\theta)d\theta \\ &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma\sigma_\pi} \exp\left\{-\frac{1}{2\sigma^2}(x-\theta)^2\right\} \exp\left\{-\frac{1}{2\sigma_\pi^2}(\theta-\mu_\pi)^2\right\} d\theta \\ &= \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_\pi^2)}} \exp\left\{-\frac{(x-\mu_\pi)^2}{2(\sigma^2 + \sigma_\pi^2)}\right\} \end{aligned}$$

边缘分布 $m(x)$ 是正态分布 $N(\mu_\pi, \sigma_\pi^2 + \sigma^2)$

Definition 8.2.5 — 混合分布. 设随机变量 X 以概率 π 在总体 F_1 中取值, 以概率 $1 - \pi$ 在总体 F_2 中取值. 若 $F(x|\theta_1)$ 和 $F(x|\theta_2)$ 分别是这两个总体的分布函数, 则 X 的分布函数为:

$$F(x) = \pi F(x|\theta_1) + (1 - \pi)F(x|\theta_2)$$

或用密度函数(或率密度)表示:

$$p(x) = \pi p(x|\theta_1) + (1 - \pi)p(x|\theta_2)$$

称分布 $F(x)$ 为 $F(x|\theta_1)$ 和 $F(x|\theta_2)$ 的混合分布.

Definition 8.2.6 — 混合样本的概念.

$$F(x) = (\pi)F(x|\theta_1) + (1 - \pi)F(x|\theta_2)$$

从混合分布中抽出的样本称为混合样本。这里的 π 和 $1 - \pi$ 可以看作一个新随机变量的分布,

$$p(\theta = \theta_1) = \pi = \pi(\theta_1), p(\theta = \theta_2) = 1 - \pi = \pi(\theta_2)$$

Theorem 8.2.3 1. 从混合分布 $F(x)$ 中抽取一个样品 x_1 , 相当于二次抽样:

- (a) 第一次: 从 $\pi(\theta)$ 中抽取一个样品 θ
- (b) 第二次: 若 $\theta = \theta_1$, 则从 $F(x|\theta_1)$ 中抽一个样品, 这个样品就是 x_1 若 $\theta = \theta_2$, 则从 $F(x|\theta_2)$ 中抽一个样品, 这个样品就是 x_1

2. 若从混合分布抽取一个容量为 n 的样本, 则约有 $n\pi(\theta_1)$ 个来自 $F(x|\theta_1)$, 约有 $n\pi(\theta_2)$ 个来自 $F(x|\theta_2)$

Definition 8.2.7 — 先验选择的 ML-II 方法. 在边缘分布 $m(x)$ 的表示式中, 若 $p(x|\theta)$ 已知, 则 $m(x)$ 的大小反映 $\pi(\theta)$ 的合理程度, 这里把 $m(x)$ 记为 $m^\pi(x)$ 。当观察值 x 对二个不同的先验分布 π_1 和 π_2 , 有

$$m^{\pi_1}(x) > m^{\pi_2}(x)$$

时, 人们可认为, 数据 x 对 π_1 比对 π_2 提供更多支持。于是把 m^π 看作 π 的似然函数是合理的。设 $\Gamma = \{\pi(\theta|\lambda), \lambda \in \Lambda\}$ 为所考虑的先验类, , 且 x_1, x_2, x_n 来自以 Γ 中的函数为先验的混合样本, 若存在 $\hat{\pi} \in \Gamma(\hat{\lambda} \in \Lambda)$ 满足 (对观测数据 x_1, x_2, \dots, x_n)

$$m(\vec{x}|\hat{\lambda}) = \sup_{\lambda \in \Lambda} \prod_{i=1}^n m(x_i|\lambda)$$

则 $\hat{\pi}$ 称为 II 型极大似然先验, 或简称为 ML-II 先验 MLS 说明: 这里将 $m(x)$ 看成似然函数! λ 是未知参数。

■ **Example 8.13** 设 $X \sim N(\theta, \sigma^2)$, 其中 σ^2 已知, 又设 $\theta \sim N(\mu_\pi, \sigma_\pi^2)$, 已算得

$$m(x|\mu_\pi, \sigma_\pi^2) = N(\mu_\pi, \sigma_\pi^2 + \sigma^2)$$

若有来自 $m(x|\mu_x, \sigma_\pi^2)$ 的混合样本 x_1, x_2, \dots, x_n , 则超参数 μ_x, σ_π^2 的似然函数为

$$\begin{aligned} m(x|\mu_\pi, \sigma_\pi^2) &= [2\pi(\sigma_\pi^2 + \sigma^2)]^{-\frac{n}{2}} \exp \left\{ -\frac{\sum(x_i - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma^2)} \right\} \\ &= 2\pi(\sigma_\pi^2 + \sigma^2)^{-\frac{n}{2}} \exp \left\{ \frac{-ns_n^2}{2(\sigma_\pi^2 + \sigma^2)} \right\} \exp \left\{ \frac{-n(\bar{x} - \mu_\pi)^2}{(\sigma_\pi^2 + \sigma^2)} \right\} \end{aligned}$$

其中

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

从上式容易看出, 当不考虑 σ_π^2 时, μ_π 在 \bar{x} 处达到最大, 所以 $\hat{\mu}_\pi = \bar{x}$ 应是 μ_π 的 ML-II 选择将 μ_π 以 \bar{x} 代入上式, 只剩 σ_π^2 的函数

$$\phi(\sigma_\pi^2) = [2\pi(\sigma_\pi^2 + \sigma^2)]^{-\frac{n}{2}} \exp \left\{ \frac{-ns_n^2}{2(\sigma_\pi^2 + \sigma^2)} \right\}$$

对 $\ln \phi(\sigma_\pi^2)$ 微分并令为零, 可得似然方程

$$\frac{d \ln \phi(\sigma_\pi^2)}{d \sigma_\pi^2} = \frac{-n/2}{\sigma_\pi^2 + \sigma^2} + \frac{-ns_n^2}{2(\sigma_\pi^2 + \sigma^2)^2} = 0$$

解之, 可得 $\sigma_\pi^2 = S_n^2 - \sigma^2$ 。若 $S_n^2 < \sigma^2$, 导致 σ_π^2 为负值, 这不合情理, 故令 $\sigma_\pi^2 = 0$ 。于是 σ_π^2 的 ML-II 估计为

$$\hat{\sigma}_\pi^2 = \begin{cases} 0, & \text{当 } s_n^2 < \sigma^2 \\ s_n^2 - \sigma^2, & \text{当 } s_n^2 \geq \sigma^2 \end{cases}$$

从而所求的 ML-II 先验为 $\pi = N(\hat{\mu}_\pi, \hat{\sigma}_\pi^2)$

Definition 8.2.8 — 选择先验分布的矩方法. 当先验密度函数 $\pi(\theta|\lambda)$ 的形式已知, 可利用先验矩与边缘分布矩之间的关系寻求超参数的估计。这种方法称为先验选择的矩方法。该方法的具体步骤是:

1. 计算总体分布 $p(x|\theta)$ 的期望 $\mu(\theta)$ 和方差

$$\sigma^2(\theta) mu(\theta) = E^{x|\theta}(X)$$

$$\sigma^2(\theta) = E^{x|\theta}[X - \mu(\theta)]^2$$

$E^{x|\theta}$ 表示用 θ 给定下的条件分布 $p(x|\theta)$ 求期望。

2. 计算边缘密度 $m(x|\lambda)$ 的期望 $\mu_m(\lambda)$ 和方差 $\sigma_m^2(\lambda)$, 其中:

$$\begin{aligned} \mu_m(\lambda) &= E^{x|\lambda}(X) = \int_x xm(x|\lambda) dx \\ &= \int_x x \int_\Theta p(x|\theta) \pi(\theta|\lambda) d\theta dx \\ &= \int_\Theta \pi(\theta|\lambda) d\theta \int_x x p(x|\theta) dx \\ &= \int_\Theta \mu(\theta) \pi(\theta|\lambda) d\theta \\ &= E^{\theta|\lambda}[\mu(\theta)] \end{aligned}$$

$$m(x) = \begin{cases} \int_{\Theta} p(x|\theta)\pi(\theta)d\theta, & \text{当 } \theta \text{ 为连续时} \\ \sum_{\theta \in \Theta} p(x|\theta)\pi(\theta), & \text{当 } \theta \text{ 为离散时} \end{cases}$$

$$\begin{aligned}\sigma_m^2(\lambda) &= E^{x|\lambda} [X - \mu_m(\lambda)]^2 \\ &= \int_x (x - \mu_m(\lambda))^2 \int_{\Theta} p(x|\theta)\pi(\theta|\lambda)d\theta dx \\ &= \int_{\Theta} \pi(\theta|\lambda)d\theta \int_x (x - \mu_m(\lambda))^2 p(x|\theta)dx \\ &= \int_{\Theta} E^{x|\theta} (x - \mu_m(\lambda))^2 \pi(\theta|\lambda)d\theta\end{aligned}$$

其中

$$\begin{aligned}E^{x|\theta} [X - \mu_m(\lambda)]^2 &= E^{x|\theta} [X - \mu(\theta) + \mu(\theta) - \mu_m(\lambda)]^2 \\ &= E^{x|\theta} [X - \mu(\theta)]^2 + E^{x|\theta} [\mu(\theta) - \mu_m(\lambda)]^2 \\ &= \sigma^2(\theta) + [\mu(\theta) - \mu_m(\lambda)]^2\end{aligned}$$

代入：

$$\begin{aligned}\sigma_m^2(\lambda) &= \int_{\Theta} E^{x|\theta} (x - \mu_m(\lambda))^2 \pi(\theta|\lambda)d\theta \\ &= \int_{\Theta} [\sigma^2(\theta) + (\mu(\theta) - \mu_m(\lambda))^2] \pi(\theta|\lambda)d\theta\end{aligned}$$

得

$$\sigma_m^2(\lambda) = E^{\theta|\lambda} [\sigma^2(\theta)] + E^{\theta|\lambda} [\mu(\theta) - \mu_m(\lambda)]^2$$

3. 混合样本矩带入总体矩：例如：当先验分布中仅含二个超参数时，即 $\lambda = (\lambda_1, \lambda_2)$ 可用混合样本 $x = (x_1, x_2, \dots, x_n)$ 再用样本矩代替边缘分布的矩，列出如下方程

$$\hat{\mu}_m = E^{\theta|\lambda} [\mu(\theta)] \quad \hat{\sigma}_m^2(\lambda) = E^{\theta|\lambda} [\sigma^2(\theta)] + E^{\theta|\lambda} [\mu(\theta) - \mu_m(\lambda)]^2$$

得超参数 $\lambda = (\lambda_1, \lambda_2)$ 的估计 $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$ 从而得先验分布 $\pi(\theta|\hat{\lambda})$

■ **Example 8.14** 设总体 $X \sim Exp(\theta)$ ，其密度函数为

$$p(x|\theta) = \theta e^{-\theta x}, x > 0$$

参数 θ 的先验分布取伽玛分布 $Ga(\alpha, \lambda)$ 其密度函数

$$\pi(\theta|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\lambda\theta}, \theta > 0$$

现有混合样本的均值 $\hat{\mu}_m$ 和方差 $\hat{\sigma}_m^2$ ，要寻求超参数 α, λ 的矩估计。

Proof. 1. 计算指数分布 $Exp(\theta)$ 的期望和方差

$$\mu(\theta) = \theta^{-1} \quad \sigma^2(\theta) = \theta^{-2}$$

2. 计算三个先验矩。

$$\begin{aligned}
 E^{\theta|\lambda}[\mu(\theta)] &= E^{\theta|\lambda}[\theta^{-1}] \\
 &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{\alpha-2} e^{-\lambda\theta} d\theta \\
 &= \frac{\lambda}{\alpha-1} \\
 E^{\theta|\lambda}[\sigma^2(\theta)] &= E^{\theta|\lambda}[\theta^{-2}] \\
 &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^{\alpha-3} e^{-\lambda\theta} d\theta \\
 &= \frac{\lambda^2}{(\alpha-1)(\alpha-2)} \\
 E^{\theta|\lambda}[(\mu(\theta) - \mu_m(\lambda))^2] &= E^{\theta|\lambda} \left[\theta^{-1} - \frac{\lambda}{\alpha-1} \right]^2 \\
 &= E^{\theta|\lambda}[\theta^{-2}] - \frac{2\lambda}{\alpha-1} E^{\theta|\lambda}[\theta^{-1}] + \frac{\lambda^2}{(\alpha-1)^2} \\
 &= \frac{\lambda^2}{(\alpha-1)(\alpha-2)} - \frac{2\lambda}{\alpha-1} \frac{\lambda}{\alpha-1} + \frac{\lambda^2}{(\alpha-1)^2} \\
 &= \frac{\lambda^2}{(\alpha-1)(\alpha-2)} - \frac{\lambda^2}{(\alpha-1)^2}
 \end{aligned}$$

把上述三式代入，即得边缘分布的期望与方差

$$\begin{aligned}
 \mu_m(\lambda) &= \frac{\lambda}{\alpha-1} \\
 \sigma_m^2(\lambda) &= \frac{2\lambda^2}{(\alpha-1)(\alpha-2)} - \frac{\lambda^2}{(\alpha-1)^2} \\
 &= \left(\frac{\lambda}{\alpha-1} \right)^2 \frac{\alpha}{\alpha-2}
 \end{aligned}$$

3. 用样本矩代替边缘分布的矩，列出方程

$$\begin{aligned}
 \hat{\mu}_m &= \frac{\lambda}{\alpha-1} \\
 \hat{\sigma}_m^2 &= \left(\frac{\lambda}{\alpha-1} \right)^2 \frac{\alpha}{\alpha-2}
 \end{aligned}$$

把第一方程代入第二方程中去，可得

$$\hat{\sigma}_m^2 = \hat{\mu}_m^2 \frac{\alpha}{\alpha-2}$$

解之可得

$$\hat{\alpha} = \frac{2\hat{\sigma}_m^2}{\hat{\sigma}_m^2 - \hat{\mu}_m^2}$$

再代回第一方程，即得

$$\hat{\lambda} = (\alpha-1)\hat{\mu}_m = \frac{\hat{\sigma}_m^2 + \hat{\mu}_m^2}{\hat{\sigma}_m^2 - \hat{\mu}_m^2}$$

这就是超参数和矩估计。 ■

■ **Example 8.15** 设总体 $X \sim N(\theta, 1)$, 其中参数 θ 的先验分布取共轭先验 $N(\mu_\pi, \sigma_\pi^2)$. 试估计两个参数的值.

Proof. 这时总体均值 $\mu(\theta) = \theta$, 而总体方差 $\sigma^2(\theta) = 1$ 与参数无关。边缘分布 $m(x|\lambda)$ 的均值与方差, 其中 $\lambda = (\mu_x, \sigma_x^2)$

$$\begin{aligned}\mu_m(\lambda) &= E^{\theta|\lambda}[\mu(\theta)] \\ &= E^{\theta|\lambda}(\theta) = \mu_\pi \\ \sigma_m^2 &= E^{\theta|\lambda}[\mu^2(\theta)] + E^{\theta|\lambda}[\mu(\theta) - \mu_m(\lambda)]^2 \\ &= E^{\theta|\lambda}[1] + E^{\theta|\lambda}[\theta - \mu_\pi]^2 \\ &= 1 + \sigma_\pi^2\end{aligned}$$

由过去的经验, 决策者认为边缘分布的均值为 10, 方差为 3。即 $\hat{\mu}_m = 10, \hat{\sigma}_m^2 = 3$ 。于是所列方程为

$$\begin{aligned}\mu_n &= 10 \\ 1 + \sigma_\pi^2 &= 3\end{aligned}$$

解之, 可得 $\hat{\mu}_\pi = 10, \hat{\sigma}_\pi^2 = 2$ 。即 θ 的先验分布为 $N(10, 2)$ ■

Definition 8.2.9 — 贝叶斯假设. 所谓参数 θ 的无信息先验分布是指除参数 0 的取值范围 Θ 和 θ 在总体分布中的地位之外, 再也不包含 θ 的任何信息的先验分布。把“不包含 θ 的任何信息”这句话理解为对 θ 的任何可能值, 都没有偏爱, 都是同等的。因此把 θ 的取值范围上的“均匀”分布看作 θ 的先验分布, 即

$$\pi(\theta) = \begin{cases} c, \theta \in \Theta \\ 0, \theta \notin \Theta \end{cases}$$

其中 Θ 是 θ 的取值范围, c 是一个容易确定的常数。这一方法通常被称为贝叶斯假设, 又称拉普拉斯 (Laplace) 先验。

使用贝叶斯假设的麻烦:

1. 当 θ 为无限区间时, 如为 $(0, \infty)$ 或 $(-\infty, +\infty)$ 时, 在 Θ 上无法定义一个正常的均匀分布。
2. 贝叶斯假设不满足变换下的不变性。

Definition 8.2.10 — 广义先验分布. 设总体 $X \sim p(x|\theta), \theta \in \Theta$. 若 θ 的先验分布 $\pi(\theta)$ 满足下列条件

1. $\pi(\theta) \geq 0$, 且 $\int_{\Theta} \pi(\theta) d\theta = \infty$
2. 由此决定的后验密度 $\pi(\theta|x)$ 是正常的密度函数, 则称 $\pi(\theta)$ 为 θ 的广义先验密度

Definition 8.2.11 — 尺度参数的无信息先验. 设总体 X 的密度函数具有形式

$$\frac{1}{\sigma} p\left(\frac{x}{\sigma}\right)$$

, 其中 σ 称为尺度参数, 参数空间为 $R^+ = (0, \infty)$ 这类密度的全体称为尺度参数族。正态分布 $N(0, \sigma^2)$ 和形状参数已知的伽玛分布都是这个分布族的成员。

Theorem 8.2.4 — σ 的无信息先验. 设让 X 改变比例尺, 即得 $Y = cX(c > 0)$. 类似地定义

$\eta = c\sigma$, 即让参数 σ 同步变化, 不难算出 Y 的密度函数为

$$\frac{1}{\eta} p\left(\frac{y}{\eta}\right)$$

仍是尺度参数族。

1. 若 X 的样本空间为 R^1 , 则 Y 的样本空间也为 R^1 ;
2. 若 X 的样本空间为 R^+ , 则 Y 的样本空间也为 R^+

此外 σ 的参数空间与 η 的参数空间都为 R^+ , 可见 (X, σ) 问题与 (y, η) 问题的统计结构完全相同, 故 σ 的无信息先验 $\pi(\sigma)$ 与 η 的无信息先验 $\pi^*(\eta)$ 应相同, 即

$$\pi(\tau) = \pi^*(\tau)$$

另一方面, 由变换 $\eta = c\sigma$ 可以得 η 的无信息先验

$$\pi^*(\eta) = \left| \frac{d\sigma}{d\eta} \right| \pi(\sigma) = \frac{1}{c} \pi\left(\frac{\eta}{c}\right)$$

比较可得

$$\pi(\eta) = \frac{1}{c} \pi\left(\frac{\eta}{c}\right)$$

取 $\eta = c$, 则有

$$\pi(c) = \frac{1}{c} \pi(1)$$

为方便计算, 令 $\pi(1) = 1$, 可得 σ 的无信息先验为

$$\pi(\sigma) = \sigma^{-1}, \sigma > 0$$

这仍是一个不正常先验, 在很多场合它可成为广义先验。

Definition 8.2.12 — Jeffreys 先验. 在较为一般场合, Jeffreys 用 Fisher 信息量 (阵) 给出未知参数 θ 的无信息先验。在贝叶斯统计中它被用来表示无信息先验分布。

Definition 8.2.13 信息量设总体的密度函数 (或分布列) 为 $p(x|\theta)$, 其中 $\theta = (\theta_1, \dots, \theta_p)' \in \Theta \subset R^p$, 如果

1. 参数空间 Θ 是 R^p 上的开矩形
2. 分布的支撑 $A = \{x : p(x|\theta) > 0\}$ 与 θ 无关
3. 对数似然 $l = \ln p(x|\theta)$ 对 θ_i 的偏导数 $\frac{\partial l}{\partial \theta_i}, i = 1, \dots, p, \theta \in \Theta$ 都存在, 常称随机向量

$$S_\theta(x) = \left(\frac{\partial l}{\partial \theta_1}, \dots, \frac{\partial l}{\partial \theta_p} \right)'$$

为记分向量或记分函数

4. 对 $p(x|\theta)$ 的积分与微分运算可以交换
5. 对一切 $1 \leq i, j \leq p$, 有

$$I_{ij}(\theta) = E_\theta \left\{ \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right\} < \infty, \theta \in \Theta$$

则称该分布族 $\{p(x|\theta), \theta \in \Theta\}$ 为 Cramer – Rao 正则分布族, 称 $\mathbf{C} - \mathbf{R}$ 正则族。在 C-R 正则族前提下, 记分向量 $S_\theta(x)$ 的方差协方差阵

$$I(\theta) = \text{Var}_\theta [S_\theta(x)] = E [S_\theta(x) S'_\theta(x)] = (I_{ij}(\theta))_{p \times p}$$

称为该分布族中参数 $\theta = (\theta_1, \dots, \theta_p)'$ 的 Fisher 信息阵, 称 θ 的信息阵。说明

1. 定义中的 C-R 正则族是 Fisher 信息阵存在的条件, 不是所有的分布都是 C-R 正则族, 如均匀分布。
2. Fisher 信息阵常被解释为是分布族中所含参数 θ 的信息量。
3. $p = 1$ 和 $p = 2$ 是两种常用的情况, 它们的 Fisher 信息阵分别为

$$I(\theta) = E \left(\frac{\partial l}{\partial \theta} \right)^2$$

$$I(\theta) = \begin{pmatrix} E \left(\frac{\partial l}{\partial \theta_1} \right)^2 & E \left(\frac{\partial l}{\partial \theta_1} \frac{\partial l}{\partial \theta_2} \right) \\ E \left(\frac{\partial l}{\partial \theta_1} \frac{\partial l}{\partial \theta_2} \right) & E \left(\frac{\partial l}{\partial \theta_2} \right)^2 \end{pmatrix}$$

4. 若总体的二阶导数均存在, Fisher 信息阵中的元素有简便的计算公式

$$I_{ij}(\theta) = E \left\{ \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right\} = -E \left\{ \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right\}$$

5. 若 x_1, x_2, \dots, x_n 来自 $C-R$ 正则族中分布的一个样本, 则该样本的 Fisher 信息阵 $I_n(\theta)$ 是原信息阵 $I(\theta)$ 的 n 倍。
6. $\pi(\theta) \propto [\det I(\theta)]^{\frac{1}{2}}$ 为无信息先验。

Proof. 第四点的推导: 事实上, 利用积分与微分次序可交换可知 $E \left(\frac{\partial l}{\partial \theta_i} \right) = 0, i = 1, \dots, p$ 因为

$$E \left(\frac{\partial l}{\partial \theta_i} \right) = \int \frac{\partial l}{\partial \theta_i} p(x|\theta) dx = \frac{\partial}{\partial \theta_i} \int p(x|\theta) dx = 0$$

对上式中 θ_j 再进行一次微分后仍为 0, 即

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_j} \left[E \frac{\partial l}{\partial \theta_i} \right] = \frac{\partial}{\partial \theta_j} \int \frac{\partial l}{\partial \theta_i} p(x|\theta) dx \\ &= \int \frac{\partial}{\partial \theta_j} \left(\frac{\partial l}{\partial \theta_i} p(x|\theta) \right) dx \\ &= \int \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} p(x|\theta) dx + \int \frac{\partial l}{\partial \theta_i} \frac{\partial p(x|\theta)}{\partial \theta_j} dx \end{aligned}$$

上式第二个积分中

$$\frac{\partial p(x|\theta)}{\partial \theta_j} = \frac{\partial \ln p(x|\theta)}{\partial \theta_j} p(x|\theta) = \frac{\partial l}{\partial \theta_j} p(x|\theta)$$

代回原式即得

$$E \left(\frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right) + E \left(\frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right) = 0$$

移项即得。特别, 当 $i = j$ 时有 $E \left(\frac{\partial l}{\partial \theta_i} \right)^2 = -E \left(\frac{\partial^2 l}{\partial \theta_i^2} \right)$

■

Proof. 第五点的证明：这是因为样本的联合分布为 $\phi(x|\theta) = \prod_{\theta=1}^n p(x_i|\theta)$, 从而其记分向量为

$$= \frac{\partial}{\partial \theta} \ln p(x|\theta) = \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n p(x_i|\theta) = \sum_{i=1}^n S_\theta(x_i)$$

再由样本各分量的独立同分布性质可知

$$I_n(\theta) = \text{Var}(S_\theta(x)) = \sum_{i=1}^n \text{Var}(S_\theta(x_i)) = nI(\theta)$$

可见，由总体分布的 Fisher 信息阵很容易获得样本分布的 Fisher 信息阵。 ■

Theorem 8.2.5 寻求分布的一般步骤:

1. 写出样本的对数似然函数:

$$l(\theta|\vec{x}) = \ln \left[\prod_{i=1}^n p(x_i|\theta) \right] = \sum_{i=1}^n \ln p(x_i|\theta)$$

2. 求样本的信息阵:

$$I_{ij}(\theta) = E \left\{ \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right\} = -E \left\{ \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right\}, i, j = 1, \dots, p$$

特别地，单参数的情形: $I(\theta) = E^{x|\theta} \left(-\frac{\partial^2 l}{\partial \theta^2} \right)$

3. θ 的无信息先验密度为:

$$\pi(\theta) \propto [\det I(\theta)]^{\frac{1}{2}}$$

信息阵 I (的行列式)

■ **Example 8.16** 设 θ 为成功概率则在 n 次独立试验中成功次数服从二项分布, 即: $P(X=x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$, $x=0, 1, 2, \dots, n$ 试求参数 θ 的 Jeffreys 先验.

Proof. 对数似然函数为: $l = x \ln \theta + (n-x) \ln(1-\theta) + \ln C_n^x$

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}$$

$$I(\theta) = E \left(-\frac{\partial^2 l}{\partial \theta^2} \right) = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$$

Jeffreys 先验

$$\pi(\theta) \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}}, \theta \in (0, 1)$$

■

四、Reference 先验基本准则: 给定观测数据, 参数的先验分布和后验分布之间 K-L 距离最大。注意: (1) 模型中没有讨厌参数时, Reference 先验就是 Jeffreys 先验, 特别对于单参数的模型 (2) 模型中存在讨厌参数时, Reference 先验与 Jeffreys 先验会不同。

1. Reference 先验们定义定义 4 设样本 \vec{x} 的分布为 $p(\vec{x}|\theta)$, θ 为参数向量, θ 的先验分布 $\pi(\theta)$ 属于先验分布族 $P = \{\pi(\theta) > 0 : \int_0 \pi(\theta)d\theta < \infty\}$ θ 的后验分布为 $\pi(\theta|\vec{x})$ 先验分布 $\pi(\theta)$ 到后验分布 $\pi(\theta|\vec{x})$ 的 K-L 距离

$$KL(\pi(\theta), \pi(\theta|\vec{x})) = \int_{\Theta} \pi(\theta|\vec{x}) \ln \left(\frac{\pi(\theta|\vec{x})}{\pi(\theta)} \right) d\theta$$

$$E[KL(\pi(\theta), \pi(\theta|\vec{x}))]$$

$$\triangleq I_{\pi(\theta)}(\theta, \vec{x}) = \int_x p(\vec{x})dx \int_{\Theta} \pi(\theta|\vec{x}) \ln \left(\frac{\pi(\theta|\vec{x})}{\pi(\theta)} \right) d\theta$$

达到最大即 $I_{\pi^*(\theta)}(\theta, \vec{x}) = \max_{\pi(\theta)} \{I_{\pi(\theta)}(\theta, \vec{x})\}$ 其中 $p(\vec{x}) = \int_{\Theta} \pi(\theta)p(\vec{x}|\theta)d\theta$, 记 $\pi^*(\theta) = \arg \max_{\pi(\theta)} \{I_{\pi(\theta)}(\theta, \vec{x})\}$ $\pi^*(\theta)$ 称为参数 θ 的 reference 作验. $I_{\pi(\theta)}(\theta, \vec{x})$ 可视为样本可以提供的关于参数 (向量) 的信息 (Bernardo, 1979). reference 先验的合理性是: $I_{\pi(\theta)}$ 越大, 先验提供的信息越少.

8.2.3 Reference 先验的计算

Definition 8.2.14 — Reference 先验的计算. 从信息量准则出发, 其基本准则为: 给定观测数据, 使得参数的先验分布和后验分布之间 Kullback-Liebler(K-L) 距离最大。当模型中没有讨厌参数时, reference 先验就是 Jeffreys 先验, 特别是对于单参数模型。当模型中存在讨厌参数时, 则 reference 先验和 Jeffreys 先验会不同。

Definition 8.2.15 设样本 x 的分布为 $p(x|\theta)$, 其中 θ 为参数 (向量), θ 的先验分布为 $\pi(\theta)$, 后验分布为 $\pi(\theta|x)$ 。称 $\pi^*(\theta)$ 为参数 θ 的 reference 先验, 如果它在先验分布类 $\mathcal{P} = \{\pi(\theta) > 0 : \int_{\Theta} \pi(\theta|x)d\theta < \infty\}$ 中, 先验分布 $\pi(\theta)$ 到后验分布 $\pi(\theta|x)$ 的 $K-L$ 距离

$$KL(\pi(\theta), \pi(\theta|x)) = \int_{\Theta} \pi(\theta|x) \log \left(\frac{\pi(\theta|x)}{\pi(\theta)} \right) d\theta$$

关于样本的平均

$$I_{\pi(\theta)}(\theta, x) = \int_X \left[\int_{\Theta} \pi(\theta|x) \log \left(\frac{\pi(\theta|x)}{\pi(\theta)} \right) d\theta \right] p(x)dx$$

达到最大, 即

$$\pi^*(\theta) = \arg \max_{\pi(\theta)} I_{\pi(\theta)}(\theta, x)$$

其中 $p(x) = \int_{\Theta} \pi(\theta)p(x|\theta)d\theta$

Theorem 8.2.6 — R 的计算--渐进法. 设 X 表示一个试验的观测结果, 向量 $X^{(k)} = (X_1, \dots, X_k)$ 的分量独立同分布于总体 X , 令

$$I_{\pi(\theta)}(\theta, x^{(k)}) = \int_P p(x^{(k)}) dx^{(k)} \int_{\Theta} \pi(\theta|x^{(k)}) \ln \frac{\pi(\theta|x^{(k)})}{\pi(\theta)} d\theta$$

最大化 $I_{\pi(\theta)}(\theta, x^{(k)})$, 获得 $\pi_k(\theta) = \arg \max_{\pi(\theta)} I_{\pi(\theta)}(\theta, x^{(k)})$

$$\begin{aligned} I_{\pi(\theta)}(\theta, x^{(k)}) &= \int_P p(x^{(k)}) dx^{(k)} \int_{\Theta} \pi(\theta | x^{(k)}) \ln \frac{\pi(\theta | x^{(k)})}{\pi(\theta)} d\theta \\ &= \int_{\Theta} \pi(\theta) d\theta \int_P p(x^{(k)} | \theta) \ln \frac{\pi(\theta | x^{(k)})}{\pi(\theta)} dx^{(k)} \\ &= \int_{\Theta} \pi(\theta) d\theta \int_P p(x^{(k)} | \theta) [\ln \pi(\theta | x^{(k)}) - \ln \pi(\theta)] dx^{(k)} \\ &= \int_{\Theta} \pi(\theta) d\theta \int_P p(x^{(k)} | \theta) [\ln \pi(\theta | x^{(k)}) - \ln \pi(\theta)] dx^{(k)} \\ &= \int_{\Theta} \pi(\theta) d\theta \left[\theta \int_P p(x^{(k)} | \theta) \ln \pi(\theta | x^{(k)}) dx^{(k)} - \ln \pi(\theta) \right] \\ &= \int_{\Theta} \pi(\theta) \ln \frac{f_k(\theta)}{\pi(\theta)} d\theta \end{aligned}$$

其中

$$f_k(\theta) = \exp \left\{ \int_P p(x^{(k)} | \theta) \ln \pi(\theta | x^{(k)}) dx^{(k)} \right\}$$

利用变分法求解, 得 $\pi_k(\theta) \propto f_k(\theta)$ Berger 等 (2009) 在适当条件下证明: θ 的 reference 先验为 $\pi^*(\theta) = \lim_{k \rightarrow \infty} \frac{f_k(\theta)}{f_k(\theta_0)}$, 其中 θ_0 是参数空间的一个内点。

Theorem 8.2.7 — 当存在多余参数时 Reference 先验的计算. 设似然函数为 $p(x|\theta, \lambda)$, 此时 θ 是感兴趣的参数, λ 是多余参数。

1. 固定 θ , 用标准的 Reference 先验方法获得 $\pi(\lambda|\theta)$ 若 λ 是一维的, 将 θ 看成常数, 计算 Jeffreys 先验, 获得 $\pi(\lambda|\theta)$
2. 如果 $\pi(\lambda|\theta)$ 是正常的先验, 则 λ 亦积分得到:

$$p(x|\theta) = \int p(\vec{x}|\theta, \lambda) \pi(\lambda|\theta) d\lambda$$

3. 基于 $p(x|\theta)$, 利用 Reference 先验方法获得 $\pi(\theta)$ 若 θ 是一维的, 利用 $p(x|\theta)$ 去计算 Jeffreys 先验, 获得 $\pi(\theta)$
4. θ 和 λ 联合先验为 $\pi(\theta, \lambda) = \pi(\lambda|\theta)\pi(\theta)$

对多于两个参数的情形, 将感兴趣参数按降序排列, 重复使用上述方法。



reference 先验的合理性在于: $I_{\pi(\theta)}(\theta, \mathbf{x})$ 越大, 先验提供的信息越少。

Definition 8.2.16 — 连续型随机变量的最大熵先验. 设随机变量是连续型的, 称

$$E_n(\pi) = -E^\pi \left[\ln \frac{\pi(\theta)}{\pi_0(\theta)} \right] = - \int_0^\infty \pi(\theta) \ln \frac{\pi(\theta)}{\pi_0(\theta)} d\theta$$

为随机变量 θ 的熵, 其中 $\pi_0(\theta)$ 为问题的自然的“不变”的无信息先验。

Definition 8.2.17 — 最大熵先验. “无信息”意味着“不确定性最大”，故无信息先验分布应是最大熵所对应的分布，所以最大熵先验的思想概括为：在满足给定约束条件的先验分布中寻找熵最大的先验分布

Definition 8.2.18 — 连续型随机变量的最大熵先验. 设随机变量 θ 是连续型的，称

$$E_n(\pi) = -E^\pi \left[\ln \frac{\pi(\theta)}{\pi_0(\theta)} \right] = -\int_{\Theta} \pi(\theta) \ln \frac{\pi(\theta)}{\pi_0(\theta)} d\theta$$

为随机变量 θ 的熵，其中 $\pi_0(\theta)$ 为问题的自然的“不变”的无信息先验

Theorem 8.2.8 设 θ 为连续型随机变量， θ 的先验分布 $\pi(\theta)$ 满足下列条件：

$$E^\pi(g_k(\theta)) = \int_{\Theta} g_k(\theta) \pi(\theta) d\theta = \mu_k, k = 1, 2, \dots, m$$

其中 g_k, μ_k 表亦已知的函数和已知的常数，则满足上式的且使 $E_n(\pi)$ 最大化的解为

$$\tilde{\pi}(\theta) = \frac{\pi_0(\theta) \exp \{ \sum_{k=1}^m \lambda_k g_k(\theta) \}}{\int_{\Theta} \pi_0(\theta) \exp \{ \sum_{k=1}^m \lambda_k g_k(\theta) \} d\theta}$$

其中 $\lambda_1, \dots, \lambda_m$ 使得当 $\pi = \tilde{\pi}$ 时，即]

$$\int_{\Theta} g_k(\theta) \tilde{\pi}(\theta) d\theta = \mu_k, \quad k = 1, 2, \dots, m \text{ 都成立.}$$

Definition 8.2.19 — 离散型随机变量的最大熵先验. 设随机变量 θ 是离散型的，它的取值为 $\theta_1, \theta_2, \dots$ 令 $\pi(\theta)$ 为 θ 的概率分布， $\pi(\theta_i) = p_i (i = 1, 2, \dots)$ ，则称

$$E_n(\pi) = -\sum_i \pi(\theta_i) \ln \pi(\theta_i) = -\sum_i p_i \ln p_i$$

规定 $0 \cdot \ln 0 = 0$.

Theorem 8.2.9 设 θ 为离散型随机变量，取值为 $\theta_1, \theta_2, \dots$ (至多可列个值)， θ 的先验分布满足下列条件：

$$E^\pi(g_k(\theta)) = \sum_i g_k(\theta_i) \pi(\theta_i) = \mu_k, \quad k = 1, 2, \dots, m$$

其中 g_k, μ_k 表示已知的函数和已知的常数，则满足上式的且使 $E_n(\pi)$ 最大化的解为

$$\tilde{\pi}(\theta_i) = \frac{\exp \{ \sum_{k=1}^m \lambda_k g_k(\theta_i) \}}{\sum_i \exp \{ \sum_{k=1}^m \lambda_k g_k(\theta_i) \}} i = 1, 2, \dots$$

其中 $\lambda_1, \dots, \lambda_m$ 使得当 $\pi = \tilde{\pi}$ 时， $\sum_i g_k(\theta_i) \tilde{\pi}(\theta_i) = \mu_k, \quad k = 1, 2, \dots, m$ 都成立.

■ **Example 8.17** 例 4 设 θ 的参数空间为 $\Theta = \{0, 1, 2, \dots\}$ ，且 θ 的先验分布满足条件： $E^\pi(\theta) = 5$. 求的最大熵先验.

Proof. 由定理 1 约束条件，可得 $m = 1, g_1(\theta) = \theta, \quad \mu_1 = 5 \tilde{\pi}(\theta_i) = \frac{e^{\lambda_1 \theta_i}}{\sum_i e^{\lambda_1 \theta_i}} = (1 - e^{\lambda_1}) e^{\lambda_1 \theta_i}, \quad \theta_i = i, i = 0, 1, 2, \dots$ 令 $p = 1 - e^{\lambda_1}, E^\pi(\theta) = (1 - p)/p = 5$, 求 e^{λ_1} 得 $= 5/6$



8.3 决策

这种人与自然界(或社会)的博将问题称为决策问题。

- Definition 8.3.1 — 三大基本要素.**
1. 状态集 $\Theta = \{\theta\}$, 其中每个元素 θ 表示自然界(或社会)可能出现的一种状态, 所有可能状态的全体组成状态集。在实际中常会遇到这样的决策问题, 其自然界(或社会)所处的状态可用一个实数表示, 这样的状态又称为状态参数, 简称为参数。其状态集 Θ 常由一些实数组成, 这样的状态集又称为参数空间。
 2. 行动集 $\mathcal{A} = \{a\}$, 其中每个元素 a 表示人对自然界(或社会)可能采取的一个行动, 所有此种行动的全部就是行动集。
 3. 收益函数 $Q(\theta, a)$ 。其中 θ 可以是状态集 Θ 中任一个状态, a 可以是行动集 A 中任一个行动, 函数值 $Q(\theta_i, a_j) = Q_{ij}$ 表示当自然界(或社会)处于状态 θ_i , 而人们选取行动 a_j 时所得到(经济上)的收益大小。收益函数的值可正可负, 其正值表示盈利, 负值表示亏损, 收益函数的单位常用货币单位, 但有时也用其它容易比较好坏的单位。收益函数是今后用来评价人们选取的行动是好是坏的基础。当状态集和行动集都仅含有有限个元素时, 如 $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}, \alpha = \{a_1, a_2, \dots, a_m\}$, 收益函数也只取 nm 个值, 这 nm 个值可以有规律地排成一个矩阵。收益矩阵

$$Q = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ Q_{11} & Q_{12} & \cdots & Q_{1m} \\ Q_{21} & Q_{22} & \cdots & Q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{n1} & Q_{n2} & \cdots & Q_{nm} \end{pmatrix} \begin{matrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{matrix}$$

Definition 8.3.2 在给定的决策问题中, A 中的行动 a_1 称为是容许的, 假如在 A 中不存在满足如下两个条件的行动 a_2

1. 对所有的 $\theta \in \Theta$, 有 $Q(\theta, a_2) \geq Q(\theta, a_1)$
2. 至少有一个 θ , 可使上述不等式严格成立。假如这样的 a_2 存在的话, 则称 a_1 是非容许的, 假如二个行动 a_1 和 a_2 的收益函数在 Θ 上处处相等, 则称行动 a_1 与 a_2 是等价的。

Theorem 8.3.1 — 悲观准则. 悲观准则, 又称差中求好准则。悲观准则由下列二步组成:

1. 对每一个行动选出最小的收益
2. 在所有选出的最小收益中选取最大值。此最大值对应的行动就是悲观准则下的最优行动。

悲观准则是一种保守的决策准则。它是在最不利的状态发生情况下, 尽量争取较多的收益。它也是一种稳妥的策略。

Theorem 8.3.2 — 乐观准则. 乐观准则, 又称好中求好准则。乐观准则, 由下列两步组成。第一步,

1. 对每一个行动选出量大的收益值
2. 在所有选出的最大收益值中选取相对最大值, 此最大值所对应的行动就是在乐观准则下寻得的最优行动。

上述两步说明, 乐观准则就是设想最有利的状态发生情况下, 尽量争取最多的收益

Theorem 8.3.3 — 折中准则, 又称赫维斯 (Hurwicz) 准则. 折中准则

1. 在 0 与 1 之间选一个数 a , 称为乐观系数, 用它来表示决策者对面临的决策问题所

持的乐观程度，愈接近于 1，决策者愈乐观；愈接近于 0，决策者愈悲观。

2. 对每一个行动 a 计算

$$H(\alpha) = \alpha \max_{\theta \in \Theta} Q(\theta, a) + (1 - \alpha) \min_{\theta \in \Theta} Q(\theta, a)$$

这里 $\max_{\theta \in \Theta} Q(\theta, a)$ 表示行动 a 的最大收益值， $\min_{\theta \in \Theta} Q(\theta, a)$ 表示行动 a 的最小收益值。 $\theta \in \Theta$

3. 取行动 a_0 ，使得 $H(a_0)$ 达到最大，即

$$H(a_0) = \max_{a \in \mathcal{A}} H(a)$$

此种 a_0 就是折中准则下的最优行动。

8.3.1 先验期望准则

Definition 8.3.3 — 最优行动. 对给定的决策问题，若在状态集 Θ 上有一个正常的先验分布 $\pi(\theta)$ ，则收益函数 $Q(\theta, a)$ 对 $\pi(\theta)$ 的期望与方差

$$\begin{aligned}\bar{Q}(a) &= E^\theta Q(\theta, a) \\ \text{Var}[Q(\theta, a)] &= E^\theta Q(\theta, a)^2 - [E^\theta Q(\theta, a)]^2\end{aligned}$$

分别称为先验期望收益和收益的先验方差。使先验平均收益达到最大的行动 a'

$$\bar{Q}(a') = \max_{a \in A} \bar{Q}(a)$$

称为先验期望准则下的**最优行动**。若此种最优行动不止一个，其中先验方差达到最小的行动称为二阶矩准则下的最优行动。

(R)

1. 定义中的先验分布只能用正常先验分布，而不能采用广义先验分布
2. 如果在比较先验期望收益的大小时，有两个或两个以上的行动使先验期望收益同时达到最大，这时才需要比较先验方差的大小做出决策
3. 使用合理的先验信息，按照先验期望准则和二阶矩准则进行决策，所得结果更加可信

Theorem 8.3.4 在先验分布不变的情况下，收益函数 $Q(\theta, a)$ 的线性变换 $kQ(\theta, a) + c (k > 0)$ 不会改变先验期望准则下的最优行动。

Proof. 记 $Q_1(\theta, a) = kQ(\theta, a) + c$ ，其中 $k > 0$ ，于是 Q_1 的先验期望收益为

$$\begin{aligned}\bar{Q}_1(a) &= E^\theta Q_1(\theta, a) \\ &= kE^\theta Q(\theta, a) + c \\ &= k\bar{Q}(a) + c\end{aligned}$$

由于 $k > 0$ ， $\bar{Q}_1(a)$ 与 $\bar{Q}(a)$ 同时达到最大值。故不会改变决策结果。 ■

Theorem 8.3.5 设 Θ_1 为状态集 Θ 的一个非空子集，假如在 Θ_1 上的收益函数 $Q(\theta, a)$ 都加上一个常数 c ，而在 Θ 上的先验分布不变，则在先验期望准则下的最优行动不变。

Proof. 根据假设, 新的收益函数为

$$Q_2(\theta, a) = \begin{cases} Q(\theta, a) + c, & \text{当 } \theta \in \Theta_1 \text{ 时} \\ Q(\theta, a), & \text{当 } \theta \in \Theta - \Theta_1 \end{cases}$$

在状态集 Θ 是连续的情况下, Q_2 的先验期望为

$$\begin{aligned} \bar{Q}_2(a) &= E^\theta Q_2(\theta, a) = \int_{\Theta} Q_2(\theta, a) \pi(\theta) d\theta \\ &= \int_{\Theta_1} [Q(\theta, a) + c] \pi(\theta) d\theta + \int_{\Theta - \Theta_1} Q(\theta, a) \pi(\theta) d\theta \\ &= \int_{\Theta} Q(\theta, a) \pi(\theta) d\theta + c \int_{\Theta_1} \pi(\theta) d\theta \\ &= \bar{Q}(a) + cP(\theta \in \Theta_1) \end{aligned}$$

可以看出, 上式第二项 $cP(\theta \in \Theta_1)$ 是与行动 a 无关的量, 故 $\bar{Q}_2(a)$ 与 $\bar{Q}(a)$ 在同一个行动上达到最大值。即在先验期望准则下的最优行动是不变。类似地, 在状态集 Θ 为离散场合亦可证得上述结果。先验期望准则的上述二个性质可以使得寻求最优行动时得以简化。甚至在归纳收益函数时不在乎原点与单位的选取。 ■

Definition 8.3.4 我们所讲的损失是指: “该赚而没有赚到的钱”。在一个决策问题中假设状态集为 $\Theta = \{\theta\}$, 行动集为 $A = \{a\}$, 而定义在 $\theta \times A$ 上的二元函数 $L(\theta, a)$ 称为损失函数, 假如它能表示在自然界(或社会)处于状态 θ 而人们采取行动 a 时对人们引起的(经济的)损失。

构成决策问题的三要素: $\Theta = \{\theta\}, A = \{a\}, L(\theta, a)$

给定收益函数 $Q(\theta, a)$

$$L(\theta, a) = \max_{a \in A} Q(\theta, a) - Q(\theta, a)$$

或给定支付函数 $W(\theta, a)$

$$L(\theta, a) = W(\theta, a) - \min_{a \in A} W(\theta, a)$$

Theorem 8.3.6 — 悲观. 第一步, 对每个行动 a , 选出最大损失值, 记为

$$\max_{\theta \in \Theta} L(\theta, a), a \in A$$

第二步, 在所有选出的最大损失中再选出最小者 a , 则 a 满足

$$\min_{a \in A} \max_{\theta \in \Theta} L(\theta, a) = \max_{\theta \in \Theta} L(\theta, a')$$

则称 a' 为悲观准则下的最优行动. 这是一种保守策略. 不求零损失, 但愿少损失.

Definition 8.3.5 对给定的决策问题在状态集 Θ 上有一个正常的先验分布 $\pi(\theta)$, 则损失函数 $L(\theta, a)$ 对 $\pi(\theta)$ 的

$$\bar{L}(a) = E^\theta [L(\theta, a)]$$

$$\text{Var}[L(\theta, a)] = E^\theta [L(\theta, a)]^2 - \left\{ E^\theta L(\theta, a) \right\}^2$$

分别称为先验期望损失和损失的先验方差使先验期望损失达到最小的行动 a

$$\bar{L}(a') = \min_{a \in A} \bar{L}(a)$$

称为先验期望准则下的最优行动。若此种最优行动不止一个，其中先验方差达到最小的行动称为二阶矩准则下的最优行动。

1. 定义中的先验分布只能用正常先验分布，而不能采用广义先验分布。
2. 损失的先验方差有着特别的意义：
 - (a) 可以作为挑选最优行动的标准（在平均先验损失相等或者相差不大时）。
 - (b) 衡量风险的大小。
3. 使用合理的先验信息，按照先验期望准则和二阶矩准则进行决策，所得结果更加可信。

■ **Example 8.18** 按季节出售的某种应时商品，每售出 1kg 获利润 6 元，如果到季末有剩余则净亏损 2 元/kg。在市场需求量一时无法弄清的情况下，该商店的经理应作如何决策？这是经营决策问题，市场需求量 θ 是状态参数。据历史资料 $\Theta = [8, 16]$ 。经理可采取的行动 a 也应在这个区间 $A = \Theta = [8, 16]$

$$\begin{aligned}\bar{L}(a) &= E^\theta [L(\theta, a)] = \int_8^a 6(\theta - a)\pi(\theta)d\theta + \int_a^{16} 2(a - \theta)\pi(\theta)d\theta \\ &= \left(\frac{3}{2}\theta^2 - \frac{3}{4}a\theta \right) \Big|_8^a + \left(\frac{1}{4}a\theta - \frac{1}{2}\theta^2 \right) \Big|_a^{16} \\ &= \frac{1}{2}a^2 - 14a + 104 \quad \frac{d}{da} L(a) = a - 14 \quad \frac{d^2L(a)}{da^2} = 1 > 0\end{aligned}$$

$$\begin{aligned}Q(\theta, a) &= \begin{cases} 6a & a \leq \theta < 16 \\ 6\theta - 2(a - \theta) & 8 < \theta < a \\ a = 14 & a = 14 \Rightarrow Q(14) = 66 \end{cases} \quad \text{当购进数量 } a \text{ 等于市场需求量时，收益达到最大为 } 6\theta \\ L(\theta, a) &= \begin{cases} 6(\theta - a), a \leq \theta & L(a) = \frac{1}{2}a^2 - 14a + 104 \\ 2(a - \theta), a > \theta & a = 14 \Rightarrow L(14) = 6 \end{cases}\end{aligned}$$

Theorem 8.3.7 不论什么场合，今后总要求损失函数是非负的，即

$$L(\theta, a) \geq 0$$

当 θ 与 a 都为实数， A 与 θ 之间的偏离， θ 的影响。

$$L(\theta, a) = \lambda(\theta) |a - \theta|^k$$

Theorem 8.3.8 1. 平方损失函数

$$L(\theta, a) = (a - \theta)^2 \quad L(\theta, a) = \lambda(\theta) (a - \theta)^2$$

这是在统计决策中用得最多的损失函数。

2. 线性损失函数

$$L(\theta, a) = \begin{cases} k_0(\theta - a), a \leq \theta & L(\theta, a) = \begin{cases} k_0(\theta)(\theta - a), a \leq \theta \\ k_1(\theta)(a - \theta), a \geq \theta \end{cases} \end{cases}$$

3. 0-1 损失函数

$$L(\theta, a) = \begin{cases} 0, & |\theta - a| \leq \varepsilon \\ k(\theta), & |\theta - a| > \varepsilon \end{cases}$$

4. 多元二次损失函数

$$L(\theta, a) = \sum_{i=1}^p \omega_i (a_i - \theta_i)^2$$

Definition 8.3.6 — 二行动线性决策问题的损失函数. 若某一决策问题只有两个行动 a_1, a_2 , 而在每个行动下的收益函数都是状态 θ (连续或离散) 的线性函数, 则

$$Q(\theta, a) = \begin{cases} b_1 + m_1 \theta, & a = a_1 \\ b_2 + m_2 \theta, & a = a_2 \end{cases} \quad (m_1 > m_2)$$

则称此决策问题为二行动线性决策问题。

该决策问题对应的损失函数

$$\begin{aligned} L(\theta, a_1) &= \begin{cases} (b_2 - b_1) + (m_2 - m_1) \theta, & \theta \leq \theta_0 \\ 0, & \theta > \theta_0 \end{cases} \\ L(\theta, a_2) &= \begin{cases} 0, & \theta \leq \theta_0 \\ (b_1 - b_2) + (m_1 - m_2) \theta, & \theta > \theta_0 \end{cases} \end{aligned}$$

Definition 8.3.7 — 效用及效用函数. 设决策问题的各可行方案有多种可能的结果值 m (货币量), 依据决策者的主观愿望和价值倾向, 每个结果值对决策者均有不同的价值和作用。反映结果值 m 对决策者的价值和作用称为效用。

Theorem 8.3.9 效用函数: m 表示结果的价值, U 表示其效用, U 是价值 m 的函数, 即

$$U = U(m)$$

, 其曲线称为效用曲线。效用函数的作用: 它衡量消费者从消费既定的商品组合中所获得满足的程度。

Definition 8.3.8 — 直线型效用函数. 持有直线型效用的决策者, 是对风险无所谓的人, 这种人对风险大小在态度上现无所区别。每一点的收益与受损对他的效用的增减都是相当的, 这是一种富豪所拥有的典型态度, 某些大公司常抱此类态度。对这类人就没有必要去寻求效用曲线了, 他可直接用收益期望值来选择最优行动。

$$U(am_1 + (1-a)m_2) = aU(m_1) + (1-a)U(m_2)$$

直线型效用函数表明: 收益与效用成线性关系。 a_1 : 以概率 α 可获 m_1 元, 以概率 $1-\alpha$ 可获 m_2 元 a_2 : 肯定获得 $\mu = \alpha m_1 + (1-\alpha)m_2$ 元。

Definition 8.3.9 — 上凸型效用函数. 持有上凸型效用的决策者总是选择 a_2 这类人对风险非常厌恶, 对亏损非常敏感。对无风险行动感兴趣。所以这类曲线又称为保守型效用曲线。人们作过调查, 大多数普通人在大部分时间里都是持此类效用曲线。如退休基金公司的经理也都持此类效曲线, 因为对他们来说, 风险就意味着退休人员领不到退休金或

少领退休金, 这将会影响社会安定。

$$U(am_1 + (1-a)m_2) > aU(m_1) + (1-a)U(m_2)$$

a_1 : 以概率 α 可获 m_1 元, 以概率 $1-\alpha$ 可获 m_2 元 | a_2 : 肯定获得 $\mu = \alpha m_1 + (1-\alpha)m_2$ 元。

Definition 8.3.10 — 下凸型效用函数. 持有下见型效用的决策者总是选择 a_1 , 这类人对高收益特别敏感, 而对亏损比较迟针。所以这类曲线又称为冒险型效用曲线。持这类效用曲线的人常常是顾后果的, 或宁愿投机结果的期望值是负的, 也不愿保持见状, 为获高额收益是不顾获得这项收益的概率非常小的这一事实。一些投机商开发商常倾向于此类效用曲线。并曲线。

$$U(am_1 + (1-a)m_2) < aU(m_1) + (1-a)U(m_2)$$

a_1 : 以概率 α 可获 m_1 元, 以概率 $1-\alpha$ 可获 m_2 元 | a_2 : 肯定获得 $\mu = \alpha m_1 + (1-\alpha)m_2$ 元。

Definition 8.3.11 — 混合型效用函数. 有人作过调查, 相当一部分人持此效用曲线。事实上, 一般情况下, 一个人也并非一点也不敢冒险, 有些决策者对在他能承受范围内的风险还是敢冒险的, 超过了他的承受能力, 变为保守型了: 另外一些决策者, 在他渴望得到一笔钱时, 也会孤注一郑, 敢于采冒险行动, 但一旦得到这笔钱后, 就唯恐失去它, 从而转向保守型了。 $m < m_0$ 效用曲线是下品的, 决策者是敢于冒险的。 $m > m_0$ 效用曲线是上兄的, 决策者是敢于保守的。点 m_0 处是决策者冒险型与保守型的分界点, 这类曲线称为有抛点型效用曲线, m_0 为抛点。

■ **Example 8.19** 4、用效用函数作决策用效用函数做决策时通常是比较先验期望效用进行选择。例 2 $U(m) = 0.62 \ln(0.004m + 1)$ a_1

$$Q = \begin{bmatrix} 10 & 5 \\ -1 & 5 \end{bmatrix} \quad \begin{array}{ll} \theta_1 & \pi(\theta_1) = 0.5 \\ \theta_2 & \pi(\theta_2) = 0.5 \end{array}$$

■ **Example 8.20** 5、从效用函数到损失

$$U(m) = U(Q(\theta, a))$$

从收益到损失一样, 亦常把效用换算为损失, 换算公式也类似于从收益换算到损失的公式。

$$L(\theta, a) = \sup_{\substack{a \in A \\ a_1 \quad a_2}} U(\theta, a) - U(\theta, a)$$

$$U = \begin{pmatrix} 0.0243 & 0.0123 \\ -0.0025 & 0.0123 \end{pmatrix} \Rightarrow L = \begin{pmatrix} a_1 & a_2 \\ 0 & 0.012 \\ 0.0148 & 0 \end{pmatrix}$$

$$\pi(\theta_1) = \pi(\theta_2) = 0.5$$

$$\bar{L}(a_1) = 0.0074 \bar{L}(a_2) = 0.006$$

a_2 为先验期望准则下最优行动。这与用效用函数做决策的结果相同。

■ **Definition 8.3.12** 先验信息: 从状态集 $\Theta = \{\theta\}$ 中概括出来的, 即为 $\pi(\theta)$

Definition 8.3.13 试验信息或抽样信息: 把自然界或社会的状态放在有关的环境中去观察、实验、抽样, 因此, 从获得的样本中去了解当今状态 θ 的最新信息.

(R) 关键: 确定一个可观察的随机变量 X , 其概率分布中恰好把它当作未知参数。

Theorem 8.3.10 当一个贝叶斯决策问题给定后. 约定已知下列条件:

1. 有一个可观察的随机变量 X , 其密度函数 $p(x|\theta)$ 依赖于未知参数 θ , 且 $\theta \in \Theta$. Θ 是状态集.
2. 在参数空间 Θ 上有一个先验分布 $\pi(\theta)$
3. 有一个行动集 $A = \{a\}$. 在对 θ 做点估计时, 一般 $A = \Theta$ 在对 θ 做区间估计时, 行动 a 就是一个区间, Θ 的一切可能的区间构成行动集 A ; 在对 θ 作假设检验时, A 只有两个行动: 接受和拒绝。
4. 在 $\Theta \times A$ 上定义了一个损失函数 $L(\theta, a)$, 则说一个贝叶斯决策问题给定了。

Definition 8.3.14 对上述两种信息的使用情况, 形成不同的决策问题:

1. 仅使用先验信息的决策问题称为无数据的决策问题。
2. 仅使用抽样信息的决策问题称为统计决策问题。
3. 先验信息和抽样信息都用的决策问题称为贝叶斯决策问题。

8.3.2 后验风险准则

Theorem 8.3.11 — 后验风险. 样本包含了 θ 的信息, 在此利用样本来进行决策。(1) 对可观测的随机变量 X , 获得样本 $x = (x_1, x_2, \dots, x_n)$, 得到样本的似然函数 $p(\cdot | \theta)$, 利用 θ 的先验分布 $\pi(\theta)$, 可以得到 θ 的后验密度函数 $\pi(\theta|x)$ (2) 把损失函数 $L(\theta, a)$ 对后验分布 $\pi(\theta|x)$ 求期望称为后验风险, 记为 $R(a|x)$

$$R(a|x) = E^{\theta_x}[L(\theta, a)] = \begin{cases} \sum_i L(\theta_i, a) \pi(\theta_i|x) & \theta \text{ 为离散型} \\ \int_{\Theta} L(\theta, a) \pi(\theta|x) d\theta & \theta \text{ 为连续型} \end{cases}$$

使后验风险达到最小的行动 a'

$$R(a'|x) = \min_{a \in A} R(a|x)$$

称为后验风险准则下的最优行动.

Definition 8.3.15 二、决策函数定义 1 在贝叶斯决策问题中, 从样本空间 $x = \{x = (x_1, x_2, \dots, x_n)\}$ 到行动集 A 上的一个映射 $\delta(x)$ 称为该决策问题的一个决策函数。所有从 χ 到 A 上的决策函数组成的类称为决策函数类. 用 D 表示, 即 $D = \{\delta(x)\}$ 注:

1. 当行动集 A 是某个实数集时, 上述决策函数就是统计量。决策函数允许其值不是实数而是某个行动。
2. 在无数据的决策问题中我们面临的是行动集 A , 并在行动集 A 中选取行动 a , 使其先验期望损失(或称先验风险)最小
3. 在贝叶斯决策问题中, 我们面临的是决策函数类 D , 要在 D 中选取决策函数 $\delta(x)$, 使其后验风险最小.

Definition 8.3.16 — 后验风险准则. 在给定的贝叶斯决策问题中, $D = \{\delta(x)\}$ 是其决策函数类, 则称 $R(\delta|x) = E^{\theta_x}[L(\theta, \delta(x))], x \in \chi, \theta \in \Theta$ 为决策函数 $\delta = \delta(x)$ 的后验风险. 若在

决策函数类 D 中存在这样的决策函数 $\delta' = \delta'(x)$ 它在 D 中具有最小的后验风险, 即

$$R(\delta'|x) = \min_{\delta \in D} R(\delta|x)$$

则称 $\delta'(x)$ 为后验风险准则下的最优决策函数, 或称贝叶斯决策函数或贝叶斯解。当状态集 Θ 和行动集 A 相同且为实数集时, $\delta'(x)$ 又称为 θ 的贝叶斯解或贝叶斯估计。常记为: $\hat{\theta} = \hat{\theta}(x), \hat{\theta}_B = \hat{\theta}_B(x)$ 注意: 此处先验分布可以使用广义先验分布。

Theorem 8.3.12 所谓“给定贝叶斯决策问题”主要是指给定如下三个前提:

1. 样本 x 的联合密度函数 $p(x|\theta)$
2. 参数空间 Θ 上的先验分布 $\pi(\theta)$
3. 定义在 $\Theta \times A$ 上的损失函数 $L(\theta, a)$

这三个前提中任一个改了, 贝叶斯决策问题就改变了, 从而贝叶斯解或贝叶斯估计也会随着改变。

■ **Example 8.21** 设 $x = \vec{x} = (x_1, x_2, \dots, x_n)$, 是来自正态分布 $N(\theta, 1)$ 的一个样本, 参数 θ 的先验分布为共轭先验分布 $N(0, \tau^2)$, 其中 τ^2 已知, 损失函数为

$$L(\theta, \delta(x)) = \begin{cases} 0, & |\delta - \theta| \leq \varepsilon \\ 1, & |\delta - \theta| > \varepsilon \end{cases}$$

求参数 θ 的贝叶斯估计。

$$\text{Proof. } \pi(\theta|\vec{x}) = N\left(\frac{\sum_{i=1}^n x_i}{n+\tau^{-2}}, \frac{1}{n+\tau^{-2}}\right) \quad \delta = \frac{\sum_{i=1}^n x_i}{n+\tau^{-2}}$$

$$\begin{aligned} R(\delta|\vec{x}) &= \int_{-\infty}^{\infty} L(\theta, \delta) \pi(\theta|\vec{x}) d\theta = p^{\theta|\vec{x}}(|\delta - \theta| > \varepsilon) \\ &= 1 - p^{\theta|\vec{x}}(|\delta - \theta| \leq \varepsilon) \end{aligned}$$

■

8.4 考试重点

1. 第一章: 计算后验概率分布, 离散 + 连续, 利用 bayes 公式; 利用似然函数得到共轭先验分布; 后验概率分布只需要利用核
2. 充分统计量来求后验概率分布 + 利用贝叶斯公式
3. 经典和贝叶斯的差别
4. 贝叶斯推断: 计算未知参数的点估计, 区间估计, 假设检验
5. 先验信息: 无信息先验: 均匀分布, bayes 假设; Jeff 先验计算
6. 描述决策问题, 状态; 在各种准则下可以选择最优行为
7. 描述; 贝叶斯; EVPI, 等计算, 纯收益
8. 概率确定; 信息利用; 充分统计量性质; 共轭的特征
9. 5 大题
10. 证明充分统计量
11. 证明是共轭先验
12. 最大后验估计和后验期望估计
13. 可信区间
14. 后验概率 + 贝叶斯因子 + 判断
15. 第一章课后题
16. 第二章课后题
17. 32 页例题

18. 5.3 的证明, $E\theta$ 的几次方用后验证证明
19. 给损失函数来做决策
20. 课后题 5.18 改数

8.5 期末复习

8.5.1 期中

1. 贝叶斯推断使用的信息: 样本信息; 总体信息; 专家经验; 历史数据
2. 贝叶斯和传统频率学派的不同: 概率的定义; 參數的看法
3. 共轭先验依赖参数
4. 条件方法: 只考虑已出现的观察值, 对没有出现的不考虑
5. 贝叶斯推断中使用的是条件方法
6. 无信息满足不变形吗? 错
7. 两大统计学都可以使用直方图
8. 边缘概率密度函数可以用来预测和确定先验分布
9. 无信息通过参数的信息来确定的: J , R , 最大熵
- 10.

8.5.2 第一章

Theorem 8.5.1 — 贝叶斯和经典的区别. 两大学派的争论:

1. 概率的理解上: 经典统计学派: 认为概率是在大量重复试验下获得的。贝叶斯统计学派: 认为概率是可以根据直接生活积累, 对某件事发生可能性给出的信息, 允许利用主观概率。
2. 参數的理解上: 经典统计学派: 将未知參數看成常数。贝叶斯统计学派: 把任意一个未知參數都看成随机变量, 应用一个概率分布去描述它的未知情况, 该分布称为先验分布。
3. 先验信息的利用与非: 经典统计学派: 不利用; 贝叶斯学派: 利用
4. 样本信息的利用: 经典统计学派: 把样本看做是来自总体的信息, 研究的是总体, 不局限数据本身。贝叶斯统计学: 是重视样本观察值, 而对尚未发生的样本观察值不予考虑。
5. 对可信区间和置信区间的认识上不同: 经典统计学派: 把真值看做常量, 置信水平为 $1 - \alpha$, m 次使用这个区间时, 大概有 $m(1 - \alpha)$ 个可以覆盖住 θ ; 贝叶斯统计学派: 将參數的真值看成是变量, 可信水平表示 θ 落入在可信区间内的概率。例如: $p\{x_1 \leq \theta \leq x_2\} = 0.9$ 表示 θ 落入 $[x_1, x_2]$ 内的概率为 0.9。
6. 贝叶斯假设检验的优点: 不需要检验统计量和抽样分布; 不需要显著性水平; 可以推广到多个, 只需要选择后验概率最大的

共同点: 1) 都承认样本有概率分布; 2) 概率的计算遵循共同的准则。



三种信息: 总体信息, 样本信息, 先验信息 (经验, 历史材料)

Theorem 8.5.2 贝叶斯思想

1. 设总体指标 X 有依赖于參數 θ 的密度函数, 在经典统计中常记为 $p(x; \theta)$ 或 $p_\theta(x)$, 它表示在参数空间 $\Theta = \{\theta\}$ 中不同的 θ 对应不同的分布。可在贝叶斯统计中记为 $p(x|\theta)$, 它表示在随机变量 θ 给定某个值时, 总体指标 X 的条件分布。
2. 根据參數 θ 的先验信息确定先验分布 $\pi(\theta)$
3. 从总体 $p(x|\theta)$ 中随机抽取样本 X_1, X_2, \dots, X_n , 该样本中含有 θ 的有关信息是样本信息。

4. 区分似然函数和联合分布函数的意义
 5. 样本和参数的联合分布

$$h(x, \theta) = p(x | \theta)\pi(\theta)$$

6. 给定样本的参数的后验分布

$$h(x, \theta) = \pi(\theta | x)m(x)$$

其中 $m(x)$ 是 x 的边缘密度函数。

$$m(x) = \int_{\Theta} h(x, \theta) d\theta = \int_{\Theta} p(x | \theta)\pi(\theta) d\theta$$

$$\pi(\theta | x) = \frac{h(x, \theta)}{m(x)} = \frac{p(x | \theta)\pi(\theta)}{\int_{\Theta} p(x | \theta)\pi(\theta) d\theta}$$

离散形式：

$$\pi(\theta_i | x) = \frac{p(x | \theta_i)\pi(\theta_i)}{\sum_j p(x | \theta_j)\pi(\theta_j)}, \quad i = 1, 2, \dots$$

贝叶斯公式的事件形式：

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{\sum_{j=1}^n P(A | B_j)P(B_j)}, \quad i = 1, 2, \dots, n$$

Theorem 8.5.3

$$\begin{aligned}\Gamma(s) &= \int_0^{+\infty} x^{s-1} e^{-x} dx \quad \text{其中 } s > 0 \\ \Gamma(n+1) &= n\Gamma(n) \quad \Gamma(n+1) = n! \quad \Gamma(1) = 1 \\ B(\alpha, \beta) &= \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx \quad \text{其中 } \alpha > 0, \beta > 0 \\ B(\alpha, \beta) &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\end{aligned}$$

若随机变量 X 的密度函数为：

$$p(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

$\alpha > 0$ 形状参数, $\lambda > 0$ 尺度参数. 则称随机变量 X 服从**伽玛分布**, 记为 $X \sim Ga(\alpha, \lambda)$

若随机变量 X 的密度函数为：

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (x \in [0, 1])$$

$\alpha > 0, \beta > 0$. 则称随机变量 X 服从**Beta 分布**, 记为

$$X \sim \beta_1(\alpha, \beta) \text{ 或者 } X \sim Be(\alpha, \beta)$$

Proposition 8.5.4 1. 当 $\alpha = \beta = 1$ 时, $\beta_1(1, 1)$ 型分布即为区间 $[0, 1]$ 上的均匀分布, 利用这个性质可以计算分布函数乘积的分布

2. 当 $\alpha = \beta = 1/2, \beta_1(1/2, 1/2)$ 型分布称为反正弦分布, 密度函数为:

$$p(x) = \frac{1}{\sqrt{\pi x(1-x)}}, 0 < x < 1$$

3. 设 $X_i \in U(0, 1)$, 则 $X_{(k)}$ 的密度函数为:

$$p(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k}, 0 < x < 1$$

即: $X_{(k)} \sim \beta_1(k, n-k+1)$

4. 设 $X \sim Ga(\alpha, \lambda)$, 则 $E(X) = \frac{\alpha}{\lambda}$, $D(X) = \frac{\alpha}{\lambda^2}$

5. 设 $X \sim Be(\alpha, \beta)$, 则 $E(X) = \frac{\alpha}{\alpha+\beta}$

$$D(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$$

Theorem 8.5.5 — 共轭先验.

1. 先计算似然函数, 再确定先验分布, 最后计算后验分布

2. 正态均值(方差已知)的共轭先验分布是正态分布. 总体分布 $N(\theta, \sigma^2)$, θ 的先验分布 $\pi(\theta)$: $\theta \sim N(\mu, \tau^2)$, 其中 μ, τ^2 已知, 后验分布: $\theta | \vec{x} \sim N(\mu_1, \tau_1^2)$

$$\sigma_0^2 = \frac{\sigma^2}{n}, A = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2}, B = \frac{\bar{x}}{\sigma_0^2} + \frac{\mu}{\tau^2}$$

后验分布也就是正态分布, 两个参数的值为

$$\mu_1 = \frac{B}{A} = \frac{\bar{x}\sigma_0^{-2} + \mu\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}}, \quad \frac{1}{\tau_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2} \quad (8.5)$$

3. 二项分布中成功概率 θ 的共轭先验分布是贝塔分布.
4. 倒伽玛分布 $IGa(\alpha, \lambda)$ 是正态方差 σ^2 的共轭先验分布.

Theorem 8.5.6 — 超参数的确认方法.

1. 矩估计: $\bar{\theta} = \frac{1}{k} \sum_{i=1}^k \theta_i$, $s_\theta^2 = \frac{1}{k-1} \sum_{i=1}^k (\theta_i - \bar{\theta})^2$
2. 分数估计: $\begin{cases} \int_0^{\theta_L} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = 0.25 \\ \int_{\theta_U}^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = 0.25 \end{cases}$

Theorem 8.5.7 — 充分统计量.

1. 用统计量 T 代替原始样本并且不损失任何有关 θ 的信息.
2. 设 $X = (X_1, X_2, \dots, X_n)$ 是来自分布函数 $F(x|\theta)$ 的样本, $T = T(\vec{X})$ 是统计量, 若在给定 $T = t$ 的条件下 \vec{X} 的条件分布与 θ 无关的话, 则称该统计量 T 是的充分统计量.
3. 充分统计量的一个重要特性: 当得到充分统计量 T 的某个取值 t 之后, 而失去原样本的观察值也没有关系. 因为我们可以根据上述的条件分布来构造某个随机试验, 从中可以获得来自总体的一个新样本, 这个样本虽然不能恢复原样本的原状, 但是它与原样本所含有的有关参数 θ 的信息是一样的.
4. 因子分解定理: 一个统计量 $T(x)$ 对参数 θ 是充分的充要条件是存在一个 t 与 θ 的函数 $g(t, \theta)$ 和一个样本 x 的函数 $h(x)$, 使得对任一样本 x 和任意 θ , 样本的密度 $p(x|\theta)$ 可表示为它们的乘积, 即

$$p(x | \theta) = g(T(x), \theta)h(x)$$

也就是一个与参数无关, 只与样本 x 有关; 一个与参数有关, 但与样本的关系仅仅通过统计量 T 来表示

5. 充分统计量不唯一
6. 设 $x = (x_1, \dots, x_n)$ 是来自密度函数 $p(x | \theta)$ 的一个样本, $T = T(x)$ 是统计量, 它的密度函数为 $p(t | \theta)$, 又设 $\mathcal{H} = \{\pi(\theta)\}$ 是 θ 的某个先验分布族, 则 $T(x)$ 为 θ 的充分统计量的充要条件是对任一先验分布 $\pi(\theta) \in \mathcal{H}$, 有

$$\pi(\theta | T(x)) = \pi(\theta | x)$$

即用样本分布 $p(x | \theta)$ 算得的后验分布与统计量 $T(x)$ 算得的后验分布是相同的。

■ **Example 8.22** 设 $x = (x_1, \dots, x_n)$ 是来自正态分布 $N(\theta, 1)$ 的一个样本, 样本均值 \bar{x} 是 θ 的充分统计量, 若 θ 的先验分布取为正态分布 $N(0, \tau^2)$, 其中 τ^2 已知, 那么 θ 的后验分布可用充分统计量 \bar{x} 的分布算得, 即

$$\pi(\theta | \bar{x}) \propto \exp \left\{ -\frac{\pi}{2} (\bar{x} - \theta)^2 - \frac{\theta^2}{2\tau^2} \right\} \quad (8.6)$$

$$\propto \exp \left\{ -\frac{1}{2} [\theta^2 (n + \tau^{-2}) - 2n\theta\bar{x}] \right\} \quad (8.7)$$

$$\propto \exp \left\{ -\frac{n + \tau^{-2}}{2} \left(\theta - \frac{n\bar{x}}{n + \tau^{-2}} \right)^2 \right\} \quad (8.8)$$

$$= N \left(\frac{n\bar{x}}{n + \tau^{-2}}, \frac{1}{n + \tau^{-2}} \right) \quad (8.9)$$

■ **Example 8.23** 求证 $B(1, \theta)$ 的充分统计量是 $T = \sum x_i$

Theorem 8.5.8 — 指数分布族. 1. 单参数: $p(x | \theta) = g(x)h(\theta) \exp\{t(x)\varphi(\theta)\}$

2. 似然函数

$$l(\theta | \vec{x}) \propto [h(\theta)]^n \exp \left\{ \sum t(x_i) \varphi(\theta) \right\}$$

则共轭先验密度 $\pi(\theta)$ 为: $\pi(\theta) \propto [h(\theta)]^\gamma \exp\{\tau\varphi(\theta)\}$ 即指数分布族的共循族 Π 为:

$$\Pi = [h(\theta)]^\gamma \exp\{\tau\varphi(\theta)\}$$

3. 两参数的形式: $p(x | \theta, \varphi) = g(x)h(\theta, \varphi) \exp\{t(x)\psi(\theta, \varphi) + u(x)\chi(\theta, \varphi)\}$
4. $\pi(\theta, \varphi) \propto [h(\theta, \varphi)]^\gamma \exp\{\alpha\psi(\theta, \varphi) + \beta\chi(\theta, \varphi)\}$

■ **Example 8.24** 正态分布 $N(\mu, \sigma^2)$ 当 σ^2 已知,

$$\begin{aligned} p(x | \mu) &= (2\pi\sigma^2)^{\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \\ &= \left[(2\pi\sigma^2)^{-\frac{1}{2}} \exp \left\{ -\frac{x^2}{2\sigma^2} \right\} \right] \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} \exp \left\{ \frac{x\mu}{\sigma^2} \right\} \end{aligned}$$

为指数分布族 $N(\mu, \sigma^2)$ 当 μ 已知, σ^2 未知, 是单参数 σ^2 的指教分布族。

总体分布	参数	共轭先验分布
二项分布	成功概率	贝塔分布 $\text{Be}(\alpha, \beta)$
泊松分布	均值	伽玛分布 $\text{Ga}(\alpha, \lambda)$
指数分布	均值的倒数	伽玛分布 $\text{Ga}(\alpha, \lambda)$
正态分布 (方差已知)	均值	正态分布 $N(\mu, \tau^2)$
正态分布 (均值已知)	方差	倒伽玛分布 $\text{IGa}(\alpha, \lambda)$



均匀分布的后验概率是 beta 分布；二项分布的共轭先验是 beta 分布

8.5.3 第二章

Definition 8.5.1 参数估计：寻取后验分布的某个特征量。

1. 使后验密度 $\pi(\theta|x)$ 达到最大的值 $\hat{\theta}_{MD}$ 称为最大后验估计；
2. 后验分布的中位数 $\hat{\theta}_{Me}$ 称为 θ 的后验中位数估计
3. 后验分布的期望值 $\hat{\theta}_E$ 称为 θ 的后验期望估计（最好）
4. 正态分布三种估计重合

这三个估计也都称为 θ 的贝叶斯估计，记为 $\hat{\theta}_B$ ，在不引起混乱时也记为 $\hat{\theta}$

■ **Example 8.25** 设 x 是来自如下指数分布的一个观察值。

$$p(x|\theta) = e^{-(x-\theta)}, x \geq \theta$$

又取柯西分布作为 θ 的先验分布，即

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}, -\infty < \theta < \infty$$

这时可得后验密度

$$\pi(\theta|x) = \frac{e^{-(x-\theta)}}{m(x)(1+\theta^2)\pi}, \theta \leq x$$

为了寻找 θ 的最大后验估计 $\hat{\theta}_{MD}$ ，我们对后验密度使用微分法，可得

$$\begin{aligned} \frac{d}{d\theta}\pi(\theta|x) &= \frac{e^{-x}}{m(x)\pi} \left[\frac{e^\theta}{1+\theta^2} - \frac{2\theta e^\theta}{(1+\theta^2)^2} \right] \\ &= \frac{e^{-x}e^\theta(\theta-1)^2}{m(x)(1+\theta^2)^2\pi} \geq 0 \end{aligned}$$

由于 $\pi(\theta|x)$ 的非减性，考虑到 θ 的取值不能超过 x ，故 θ 的最大后验估计应为

$$\hat{\theta}_{MD} = x$$

■ **Example 8.26** 为估计不合格品率 θ ，今从一批产品中随机抽取 n 件，其中不合格品数 X 服从二项分布 $b(n, \theta)$ ，国内外诸多文献上都取贝塔分布 $\text{Be}(\alpha, \beta)$ 作为 θ 的先验分布，它的众数为 $(\alpha-1)/(\alpha+\beta-2)$ ，它的期望为 $\alpha/(\alpha+\beta)$ ，这里假设 α 与 β 已知，由共轭先验分布可知，这时 θ 的后验分布仍为贝塔分布 $\text{Be}(\alpha+x, \beta+n-x)$ ，这时 θ 的最大后验估计 $\hat{\theta}_{MD}$ 和后验期望估计 $\hat{\theta}_E$ 分别为

$$\hat{\theta}_{MD} = \frac{\alpha+x-1}{\alpha+\beta+n-2} \quad \hat{\theta}_E = \frac{\alpha+x}{\alpha+\beta+n}$$

Definition 8.5.2 设参数 θ 的后验分布 $\pi(\theta|x)$, 贝叶斯估计为 $\hat{\theta}$, 则 $(\theta - \hat{\theta})^2$ 的后验期望

$$\text{MSE}(\hat{\theta}|x) = E^{\theta|x}(\theta - \hat{\theta})^2$$

称为 $\hat{\theta}$ 的后验均方差, 而其平方根 $[\text{MSE}(\hat{\theta}|x)]^{\frac{1}{2}}$ 称为 $\hat{\theta}$ 的后验标准误, 其中符号 $E^{\theta|x}$ 表示用后验分布 $\pi(\theta|x)$ 求期望, 当 $\hat{\theta}$ 为 θ 的后验期望 $\hat{\theta}_E = E(\theta|x)$ 时, 则

$$\text{MSE}(\hat{\theta}_E|x) = E^{\theta|x}(\theta - \hat{\theta}_E)^2 = \text{Var}(\theta|x)$$

称为后验方差, 其平方根 $[\text{Var}(\theta|x)]^{\frac{1}{2}}$ 称为后验标准差。后验均方差与后验方差有如下关系:

$$\begin{aligned}\text{MSE}(\hat{\theta}|x) &= E^{\theta|x}(\theta - \hat{\theta})^2 \\ &= E^{\theta|x}[(\theta - \hat{\theta}_E) + (\hat{\theta}_E - \hat{\theta})]^2 \\ &= \text{Var}(\theta|x) + (\hat{\theta}_E - \hat{\theta})^2\end{aligned}$$

Theorem 8.5.9 — 点估计--求后验期望估计的方法.

1. 利用条件公式计算联合分布
2. 再次利用条件公式计算后验分布
3. 最大后验估计就是后验概率最大的那个; 后验期望估计就对后验函数取期望; 后验均方差使用公式: $\text{MSE}(\hat{\theta}|x) = \text{Var}(\theta|x) + (\hat{\theta}_E - \hat{\theta})^2$

 HPD

Theorem 8.5.10 — 假设检验. 在获得后验分布 $\pi(\theta|x)$ 后, 即可计算二个假设 H_0 与 H_1 的后验概率

$$\alpha_i = P(\Theta_i|x) d\theta, i = 0, 1$$

然后比较 α_0 与 α_1 的大小,

1. 当后验概率比(或称后验机会比) $\alpha_0/\alpha_1 > 1$ 时接受 H_0 ;
2. 当 $\alpha_0/\alpha_1 < 1$ 时接受 H_1 ;
3. 当 $\alpha_0/\alpha_1 \approx 1$ 时, 不宜做判断, 尚而进一步抽样或进一步搜集先验信息。

Definition 8.5.3 设两个假设 Θ_0 与 Θ_1 的先验概率分别为 π_0 与 π_1 , 后验概率分别为 α_0 与 α_1 , 则称

$$B^\pi(x) = \frac{\text{后验机会比}}{\text{先验机会比}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

为贝叶斯因子。贝叶斯因子 $B^\pi(x)$ 是数据 x 支持原假设 Θ_0 的程度。



1. 不需要检验统计量和抽样分布
2. 不需要显著性水平
3. 可以推广到多个, 只需要选择后验概率最大的

Theorem 8.5.11 1. 简单对简单的拒绝域: $\alpha_0/\alpha_1 < 1$, 或

$$\frac{p(x|\theta_1)}{p(x|\theta_0)} > \frac{\pi_0}{\pi_1}$$

2. 复杂对复杂: 后验概率比为

$$\frac{\alpha_0}{\alpha_1} = \frac{\int_{\Theta_0} p(x|\theta) \pi_0 g_0(\theta) d\theta}{\int_{\Theta_1} p(x|\theta) \pi_1 g_1(\theta) d\theta}$$

于是贝叶斯因子可表示为

$$B^\pi(x) = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{\int_{\Theta_0} p(x|\theta) g_0(\theta) d\theta}{\int_{\Theta_1} p(x|\theta) g_1(\theta) d\theta} = \frac{m_0(x)}{m_1(x)}$$

3. 简单对复杂:

$$\pi(\theta) = \pi_0 I_{\theta_0}(\theta) + \pi_1 g_1(\theta)$$

后验机会比为

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0}{\pi_1} \frac{p(x|\theta_0)}{m_1(x)}$$

从而贝叶斯因子为

$$B^\pi(x) = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = \frac{p(x|\theta_0)}{m_1(x)}$$

$$\alpha_0 = \pi(\Theta_0 | x) = \left[1 + \frac{1 - \pi_0}{\pi_0} \frac{1}{B^\pi(x)} \right]^{-1}$$



贝叶斯因子对样本信息变化是敏感的, 先验信息对信息变化是迟钝的

Theorem 8.5.12 — 预测. 获得预测分布

1. 无数据: 使用 X 的边缘分布的中位数, 期望等
2. 如要预测同一总体 $p(x|\theta)$ 的未来观察值, 则在

$$m(x|x) = \int_{\theta} p(x|\theta) \pi(\theta|x) d\theta$$

如要预测另一总体 $g(z|\theta)$ 的未来观察值, 则有

$$m(z|x) = \int_{\theta} g(z|\theta) \pi(\theta|x) d\theta$$

8.6 第三章

Theorem 8.6.1 — 确定主观概率的方法. 1. 离散: 利用对立事件的比较确定主观概率

2. 利用专家意见确定主观概率
3. 向多位专家咨询确定主观概率
4. 充分利用历史资料, 考虑现有信息加以修正
5. 连续: 直方图法
6. 选定先验密度附数形式再估计其超参数
7. 定分度法与变分度法
8. 相对似然法 (确定最大最小 + 插点)
9. 边缘分布 $m(x)$
10. 混合分布:

$$p(x) = \pi p(x|\theta_1) + (1 - \pi)p(x|\theta_2)$$

11. 设 $\Gamma = \{\pi(\theta|\lambda), \lambda \in \Lambda\}$ 为所考虑的先验类, , 且 x_1, x_2, x_n 来自以 Γ 中的函数为先验的混合样本, 若存在 $\hat{\pi} \in \Gamma (\hat{\lambda} \in \Lambda)$ 满足 (对观测数据 x_1, x_2, \dots, x_n)

$$m(\vec{x}|\hat{\lambda}) = \sup_{\lambda \in \Lambda} \prod_{i=1}^n m(x_i|\lambda)$$

则 $\hat{\pi}$ 称为 II 型极大似然先验

12. 选择先验分布的矩方法: 当先验密度函数 $\pi(\theta|\lambda)$ 的形式已知, 可利用先验矩与边缘分布矩之间的关系寻求超参数的估计。具体步骤是:
- (a) 计算总体分布 $p(x|\theta)$ 的期望 $\mu(\theta) = E^{x|\theta}(X)$ 和方差 $\sigma^2(\theta) = E^{x|\theta}[X - \mu(\theta)]^2$
 - (b) 利用先验分布, 计算边缘密度 $m(x|\lambda)$ 的期望

$$\mu_m(\lambda) = E^{\theta|\lambda}[\mu(\theta)]$$

和方差

$$\sigma_m^2(\lambda) = E^{\theta|\lambda}[\sigma^2(\theta)] + E^{\theta|\lambda}[\mu(\theta) - \mu_m(\lambda)]^2$$

- (c) 混合样本矩带入总体矩: 例如: 当先验分布中仅含二个超参数时, 即 $\lambda = (\lambda_1, \lambda_2)$ 可用混合样本 $x = (x_1, x_2, \dots, x_n)$ 再用样本矩代替边缘分布的矩, 列出如下方程

$$\hat{\mu}_m = E^{\theta|\lambda}[\mu(\theta)] \quad \hat{\sigma}_m^2(\lambda) = E^{\theta|\lambda}[\sigma^2(\theta)] + E^{\theta|\lambda}[\mu(\theta) - \mu_m(\lambda)]^2$$

得超参数 $\lambda = (\lambda_1, \lambda_2)$ 的估计 $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$ 从而得先验分布 $\pi(\theta|\hat{\lambda})$

Theorem 8.6.2 — 贝叶斯假设. 把“不包含 θ 的任何信息”这句话理解为对 θ 的任何可能值, 都没有偏爱, 都是同等的。因此把 θ 的取值范围上的“均匀”分布看作 θ 的先验分布, 即

$$\pi(\theta) = \begin{cases} c, & \theta \in \Theta \\ 0, & \theta \notin \Theta \end{cases}$$

其中 Θ 是 θ 的取值范围, c 是一个容易确定的常数。这一行法通常被称为贝叶斯假设, 又称拉普拉斯 (Laplace) 先验。

使用贝叶斯假设的麻烦:

1. 当 θ 为无限区间时, 如为 $(0, \infty)$ 或 $(\infty, +\infty)$ 时, 在 Θ 上无法定义一个正常的均匀分布。
2. 贝叶斯假设不满足变换下的不变性。

Definition 8.6.1 — 广义先验分布. 设总体 $X \sim p(x|\theta), \theta \in \Theta$. 若 θ 的先验分布 $\pi(\theta)$ 满足下列条件

1. $\pi(\theta) \geq 0$, 且 $\int_{\Theta} \pi(\theta) d\theta = \infty$
2. 由此决定的后验密度 $\pi(\theta|x)$ 是正常的密度函数, 则称 $\pi(\theta)$ 为 θ 的广义先验密度

Theorem 8.6.3

1. $p(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right), \mu \in (-\infty, \infty), \sigma \in (0, \infty)$ 位置-尺度参数组: 正态, 指数, 均匀
2. $\sigma = 1$ 称作位置参数族, 无信息先验 $\pi(\theta) \propto 1, \theta \in \Theta$
3. $\mu = 0$ 时位置-尺度参数族称为尺度参数族。无信息先验: $\pi(\sigma) \propto \frac{1}{\sigma}, \sigma \in (0, \infty)$

Definition 8.6.2 — 正则组, 信息量. 设总体的密度函数(或分布列)为 $p(x|\theta)$, 其中 $\theta = (\theta_1, \dots, \theta_p)' \in \Theta \subset R^p$, 如果

1. 参数空间 Θ 是 R^p 上的开矩形
2. 分布的支撑 $A = \{x : p(x|\theta) > 0\}$ 与 θ 无关
3. 对数似然 $l = \ln p(x|\theta)$ 对 θ_i 的偏导数 $\frac{\partial l}{\partial \theta_i}, i = 1, \dots, p, \theta \in \Theta$ 都存在, 常称随机向量

$$S_{\theta}(x) = \left(\frac{\partial l}{\partial \theta_1}, \dots, \frac{\partial l}{\partial \theta_p} \right)'$$

为记分向量或记分函数

4. 对 $p(x|\theta)$ 的积分与微分运算可以交换
5. 对一切 $1 \leq i, j \leq p$, 有

$$I_{ij}(\theta) = E_{\theta} \left\{ \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right\} < \infty, \theta \in \Theta$$

则称该分布族 $\{p(x|\theta), \theta \in \Theta\}$ 为 Cramer – Rao 正则分布族, 称 **C – R 正则族**。在 C-R 正则族前提下, 记分向量 $S_{\theta}(x)$ 的方差协方差阵

$$I(\theta) = \text{Var}_{\theta}[S_{\theta}(x)] = E[S_{\theta}(x)S'_{\theta}(x)] = (I_{ij}(\theta))_{p \times p}$$

称为该分布族中参数 $\theta = (\theta_1, \dots, \theta_p)'$ 的 Fisher 信息阵, 称 θ 的信息阵。

Theorem 8.6.4

1. 定义中的 C-R 正则族是 Fisher 信息阵存在的条件, 不是所有的分布都是 C-R 正则族, 如均匀分布。
2. Fisher 信息阵常被解释为是分布族中所含参数 θ 的信息量。
3. $p = 1$ 和 $p = 2$ 是两种常用的情况, 它们的 Fisher 信息阵分别为

$$I(\theta) = E \left(\frac{\partial l}{\partial \theta} \right)^2$$

$$I(\theta) = \begin{pmatrix} E \left(\frac{\partial l}{\partial \theta_1} \right)^2 & E \left(\frac{\partial l}{\partial \theta_1} \frac{\partial l}{\partial \theta_2} \right) \\ E \left(\frac{\partial l}{\partial \theta_1} \frac{\partial l}{\partial \theta_2} \right) & E \left(\frac{\partial l}{\partial \theta_2} \right)^2 \end{pmatrix}$$

4. 若总体的二阶导数均存在,Fisher 信息阵中的元素有简便的计算公式

$$I_{ij}(\theta) = E \left\{ \frac{\partial l}{\partial \theta_i} \frac{\partial l}{\partial \theta_j} \right\} = -E \left\{ \frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right\}$$

5. 若 x_1, x_2, \dots, x_n 来自 C^-R 正则族中分布的一个样本, 则该样本的 Fisher 信息阵 $I_n(\theta)$ 是原信息阵 $I(\theta)$ 的 n 倍。
 6. Jeff 先验: $\pi(\theta) \propto [\det I(\theta)]^{\frac{1}{2}}$ 为 θ 无信息先验.

Theorem 8.6.5 — 步骤.

1. 求样本的对数似然函数
2. 求样本的信息阵
3. 行列式开平方

Definition 8.6.3 — Reference 先验的计算. 基本准则为: 给定观测数据, 使得参数的先验分布和后验分布之间 Kullback-Liebler(K-L) 距离最大。

1. 当模型中没有讨厌参数时, reference 先验就是 Jeffreys 先验, 特别是对于单参数模型。
2. 当模型中存在讨厌参数时, 则 reference 先验和 Jeffreys 先验不同。

Definition 8.6.4 设样本 x 的分布为 $p(x|\theta)$, 其中 θ 为参数(向量), θ 的先验分布为 $\pi(\theta)$, 后验分布为 $\pi(\theta|x)$. 称 $\pi^*(\theta)$ 为参数 θ 的 reference 先验, 如果它在先验分布类 $\mathcal{P} = \{\pi(\theta) > 0 : \int_{\Theta} \pi(\theta|x)d\theta < \infty\}$ 中, 先验分布 $\pi(\theta)$ 到后验分布 $\pi(\theta|x)$ 的 $K-L$ 距离

$$KL(\pi(\theta), \pi(\theta|x)) = \int_{\Theta} \pi(\theta|x) \log \left(\frac{\pi(\theta|x)}{\pi(\theta)} \right) d\theta$$

关于样本的平均

$$I_{\pi(\theta)}(\theta, x) = \int_{\mathcal{X}} \left[\int_{\Theta} \pi(\theta|x) \log \left(\frac{\pi(\theta|x)}{\pi(\theta)} \right) d\theta \right] p(x) dx$$

达到最大, 即

$$\pi^*(\theta) = \arg \max_{\pi(\theta)} I_{\pi(\theta)}(\theta, x)$$

其中 $p(x) = \int_{\Theta} \pi(\theta)p(x|\theta)d\theta$

Definition 8.6.5 — 最大熵先验. "无信息" 意味着"不确定性最大", 故无信息先验分布应是最大熵所对应的分布, 所以最大熵先验的思想概括为: 在满足给定约束条件的先验分布中寻找熵最大的先验分布

Theorem 8.6.6 — 离散形式. 设 θ 为离散型随机变量, 取值为 $\theta_1, \theta_2, \dots$ (至多可列个值), θ 的先验分布满足下列条件:

$$E^{\pi}(g_k(\theta)) = \sum_i g_k(\theta_i) \pi(\theta_i) = \mu_k, \quad k = 1, 2, \dots, m$$

其中 g_k, μ_k 表示已知的函数和已知的常数, 则满足上式的且使 $E_n(\pi)$ 最大化的解为

$$\tilde{\pi}(\theta_i) = \frac{\exp\{\sum_{k=1}^m \lambda_k g_k(\theta_i)\}}{\sum_i \exp\{\sum_{k=1}^m \lambda_k g_k(\theta_i)\}} i = 1, 2, \dots$$

其中 $\lambda_1, \dots, \lambda_m$ 使得当 $\pi = \tilde{\pi}$ 时, $\sum_i g_k(\theta_i) \tilde{\pi}(\theta_i) = \mu_k, k = 1, 2, \dots, m$ 都成立.

■ **Example 8.27** 设 θ 的参数空间为 $\Theta = \{0, 1, 2, \dots\}$, 且 θ 的先验分布满足条件: $E^\pi(\theta) = 5$. 求的最大熵先验.

Proof. 由约束条件, 可得 $m = 1, g_1(\theta) = \theta, \mu_1 = 5$ $\tilde{\pi}(\theta_i) = \frac{e^{\lambda_1 \theta_i}}{\sum_i e^{\lambda_1 \theta_i}} = (1 - e^{\lambda_1}) e^{\lambda_1 \theta_i}, \theta_i = i, i = 0, 1, 2, \dots$ 令 $p = 1 - e^{\lambda_1}, E^\pi(\theta) = (1 - p)/p = 5$, 求 e^{λ_1} 得 $= 5/6$

■

8.7 第四章

Theorem 8.7.1 1. 决策三要素: 状态集, 行为集, 收益函数

2. 推断: 理清因果关系; 决策: 做决定
3. 损失函数被称为第四章信息
4. 乐观: 好中求好; 悲观: 坏中求好; 折中: 乐观系数

$$H(a) = \alpha \max_{\theta \in \Theta} Q(\theta, a) + (1 - \alpha) \min_{\theta \in \Theta} Q(\theta, a)$$

这里的最大和最小是在一列里面找

5. 先验期望准则: 收益函数 Q 对先验分布的期望和方差成为先验期望收益和先验方差; 在多个达到最大值才比方差; 不可使用广义先验
6. 离散型的比较先验均值和方差; 连续型的求的先验期望, 再对其求导得到最大值; 先验期望具有不变性
7. 损失和亏损的意思不同
8. 给定收益函数 $Q(\theta, a)$

$$L(\theta, a) = \max_{a \in A} Q(\theta, a) - Q(\theta, a)$$

或给定支付函数 $W(\theta, a)$

$$L(\theta, a) = W(\theta, a) - \min_{a \in A} W(\theta, a)$$

9. 损失函数: 悲观: 最大的损失里面找最小的
10. 用损失函数作决策更好, 因为信息使用的更加多
11. 同样的可以定义先验期望损失和先验方差
12. 两行动线性决策的损失函数: 找到交点
13. 常用的效用函数: 直线: 有钱的; 上凸: 保守的; 下凸: 冒险的; 混合的: 先冒险再保守, 先下凸再上凸
14. 效益矩阵带入效用函数可以得到效用矩阵; 但是损失矩阵和效用之间的关系

$$L(\theta, a) = \sup_{a \in A} U(\theta, a) - U(\theta, a)$$

8.8 第五章

Theorem 8.8.1 1. 仅使用先验信息的决策问题称为无数据的决策问题；仅使用抽样信息的决策问题称为统计决策问题；先验信息和抽样信息都用的决策问题称为贝叶斯决策问题。

2. 后验风险：样本的似然函数 + 先验分布 == 后验分布；后验风险就是损失函数对后验分布做期望；使后验风险达到最小的行动 a'

$$R(a'|x) = \min_{a \in A} R(a|x)$$

称为后验风险准则下的最优行动。

- 3. 做题的时候注意离散情况和连续情况
- 4. 决策函数是样本空间到行为集上的映射
- 5. 在给定的贝叶斯决策问题中, $D = \{\delta(x)\}$ 是其决策函数类, 则称

$$R(\delta|x) = E^{\theta|x}[L(\theta, \delta(x))], x \in \chi, \theta \in \Theta$$

为决策函数 $\delta = \delta(x)$ 的后验风险。若在决策函数类 D 中存在这样的决策函数 $\delta' = \delta'(x)$ 它在 D 中具有最小的后验风险, 即

$$R(\delta'|x) = \min_{\delta \in D} R(\delta|x)$$

则称 $\delta'(x)$ 为后验风险准则下的最优决策函数, 或称贝叶斯决策函数或贝叶斯解。

- 6. 注意: 此处先验分布可以使用广义先验分布。
- 7. 所谓“给定贝叶斯决策问题”主要是指给定如下三个前提:
 - (a) 样本 x 的联合密度函数 $p(x|\theta)$
 - (b) 参数空间 Θ 上的先验分布 $\pi(\theta)$
 - (c) 定义在 $\Theta \times A$ 上的损失函数 $L(\theta, a)$
 这三个前提中任一个改了, 贝叶斯决策问题就改变了, 从而贝叶斯解或贝叶斯估计也会随着改变。
- 8. 给损失函数, 求贝叶斯估计: 首先计算出后验分布; 计算 $R(\delta | \vec{x})$, 寻找参数使得其达到最小值
- 9. 假设检验问题: $H_0: x$ 来自 $p_0(x), H_1: x$ 来自 $p_1(x)$, 欠!

Theorem 8.8.2 — 常见损失函数的贝叶斯估计. 1. 在平方损失函数 $L(\theta, \delta(x)) = (\delta - \theta)^2$ 下, θ 的贝叶斯估计为后验均值, 即 $\delta_B(x) = E(\theta | \vec{x})$

2. 在加权平方损失函数 $L(\theta, \delta(x)) = \lambda(\theta)(\delta - \theta)^2$ 下, 的贝叶斯估计为

$$\delta_B(x) = \frac{E[\lambda(\theta)\theta | \vec{x}]}{E[\lambda(\theta) | \vec{x}]}$$

3. 在参数向量 $\theta' = (\theta_1, \theta_2, \dots, \theta_k)$ 的场合下, 对多元二次损失函数 $L(\theta, \delta) = (\delta - \theta)^T Q(\delta - \theta), Q$ 为正定矩阵则 θ 的贝叶斯估计为后验均值向量:

$$\delta_B(x) = E(\theta | \vec{x}) = \begin{pmatrix} E(\theta_1 | \vec{x}) \\ \vdots \\ E(\theta_k | \vec{x}) \end{pmatrix}$$

4. 如果求平凡损失下 θ^{-1} 的贝叶斯估计, 可以使用随机变量函数的后验期望, 或者求出逆的分布

5. 在绝对值损失函数 $L(\theta, \delta) = |\delta - \theta|$ 下, θ 的贝叶斯估计为后验分布的中位数: 证明思路: 分类讨论去掉绝对值, 在进行放缩
6. 在线性损失函数 $L(\theta, \delta) = \begin{cases} k_0(\theta - \delta), & \delta \leq \theta \\ k_1(\delta - \theta), & \delta > \theta \end{cases}$ 下, θ 的贝叶斯估计为后验分布的 $k_0 / (k_0 + k_1)$ 下分位数。

R 常见的后验分布的推导和伽马分布, beta 分布的期望和方差

Theorem 8.8.3 有限个行为的假设检验: 根据题设计算出每个假设的损失函数; 计算后验分布; 计算 $R(a_1 | x = 115)$, 比较小大

- Theorem 8.8.4**
1. 完全信息期望, 每种 θ 下的最大值乘上对应的 θ 的先验概率; 完全信息下的收益: $E^\theta[\max Q(\theta, a_i)]$ 先验期望收益: $\max_{a_i} E^\theta[Q(\theta, a_i)]$ 收益函数对先验分布求期望
 2. **完全信息先验期望**: $EVPI = E^\theta[\max Q(\theta_i, a_j)] - \max_{a_j} E^\theta[Q(\theta_i, a_j)]$
 3. 找到最优行为后, 可对损失函数求先验期望就是完全信息先验期望。 a' 是先验期望准则下的最优行动, 则在 a' 下的损失函数 $L(\theta, a')$ 的先验期望 $E^\theta[L(\theta, a')]$ 称为完全信息先验期望值, 记为先验 $EVPI = E^\theta[L(\theta, a')]$
 4. 设 $\pi(\theta | x)$ 为样本 $x = (x_1, x_2, \dots, x_n)$ 给定下 θ 的后验分布, $\delta(x)$ 为据此后验分布所确定的贝叶斯决策函数, 用 $\pi(\theta | x)$ 计算在 $\delta'(x)$ 下损失函数 $L(\theta, \delta'(x))$ 的后验期望可称为完全信息后验期望值, 记为

$$\text{后验}EVPI = E^{\theta|x}[L(\theta, \delta'(x))]$$

用样本的边缘分布 $m(x)$ 对 $E^{\theta|x}[L(\theta, \delta'(x))]$ 再求一次期望, 并称为后验 EVPI 期望值, 即后验 EVPI 期望值 = $E^x\{E^{\theta|x}[L(\theta, \delta')]\}$

5. 在一个贝叶斯决策问题中 a' 是先验期望准则下的最优行动, $\delta'(x)$ 时后验准则下的最优决策函数, 则先验 EVPI 与后验 EVPI 期望值的差称为**抽样信息期望值**, 记为

$$EVSI = E^\theta[L(\theta, a')] - E^x\{E^{\theta|x}[L(\theta, \delta'(x))]\}$$

Definition 8.8.1

$$C(n) = C_f + C_v \cdot n \quad (n \geq 1), \text{ 当 } n = 0 \text{ 时, } C(n) \text{ 规定为 } 0$$

抽样净益, 记为 ENGS, 即

$$ENGS(n) = EVSI(n) - C(n)$$

n 满足 $ENGS(n) > 0$! 使得抽样净益达到最大的样本量 n^* 称为最佳样本量, 即 n^* 满足

$$ENGS(n^*) = \max_{n>0} ENGS(n)$$

若最佳样本量不止一个, 就选其中最小的一个作为最佳样本量。

$$\text{给出上界: } n^* \leq \frac{\text{先验}EVPI - C_f}{C_v}$$

Part Four

9	一元回归	277
9.1	参数估计	
9.2	极大似然估计	
9.3	假设检验	
9.4	残差检验	
10	多元回归	291
10.1	显著性检验	
10.2	违背基本假设的情况	
11	自变量的选择和逐步回归和多重共线性	311
11.1	自变量的选择	
11.2	逐步回归	
11.3	多重共线性	
12	岭回归	319
13	主成分回归和偏最小二乘	323
13.1	主成分回归	
13.2	偏最小二乘	
14	非线性	329
14.1	非线性	
14.2	Logistic 回归	
15	多元正态及参数估计	333
15.1	假设检验	
16	线性模型	367
17	线性模型参数估计	413
	Bibliography	429
	Articles	
	Books	

9. 一元回归

Definition 9.0.1 — 回归分析和相关分析的异同. 回归分析和相关分析都是研究变量间关系的统计学课题。在应用中，两种分析方法经常相互结合和渗透，但它们研究的重点和应用面不同。它们的差别主要有以下几点：

1. 在回归分析中，变量 y 称为因变量，处在被解释的特殊地位。在相关分析中，变量 y 与变量 x 处于平等的地位，即研究变量 y 与变量 x 的密切程度与研究变量 x 与变量 y 的密切程度是一回事。
2. 相关分析中所涉及的变量 y 与 x 全是随机变量。而回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量。通常的回归模型中，我们总是假定 x 是非随机的确定变量。
3. 相关分析的研究主要是为刻画两类变量间线性相关的密切程度。而回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制。

Theorem 9.0.1 — 主要内容.

1. 线性回归
 - 一元线性回归
 - 多元线性回归
 - 多个因变量与多个自变量的回归
2. 回归诊断
 - 讨论如何从数据推断回归模型基本假设的合理性
 - 当基本假设不成立时如何对数据进行修正
 - 判定回归方程拟合的效果
 - 选择回归函数的形式
3. 回归变量的选择
 - 自变量选择的准则
 - 逐步回归分析方法
4. 参数估计方法的改进
 - 岭回归
 - 主成分回归
 - 偏最小二乘法
5. 非线性回归
 - 一元非线性回归
 - 分段回归
 - 多元非线性回归

6. 含有定性变量的回归 $\begin{cases} \text{自变量含定性变量的情况} \\ \text{因变量是定性变量的情况} \end{cases}$

Definition 9.0.2 1. 理论回归方程：假设将所有的观测值都带入方程时，计算两个参数

$$E(y|x) = \alpha + \beta x$$

2. 经验回归方程：

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$\hat{\alpha}, \hat{\beta}$ 经验回归常数和经验回归系数

Theorem 9.0.2 随机误差项主要包括下列因素的影响：

1. 由于人们认识的局限或时间、费用、数据质量等的制约未引入回归模型但又对回归被解释变量 y 有影响的因素
2. 样本数据的采集过程中变量观测值的观测误差
3. 理论模型设定的误差
4. 其他随机因素

Definition 9.0.3 — 回归模型的一般形式. 如果变量 x_1, x_2, \dots, x_p 与随机变量 y 之间存在着相关关系，通常就意味着每当 x_1, x_2, \dots, x_p 取定值后， y 便有相应的概率分布与之对应。随机变量 y 与相关变量 x_1, x_2, \dots, x_p 之间的概率模型为

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon$$

分别为 x, y 解释变量和被解释变量，模型中含有确定的部分即回归函数，也含有不确定的部分 ε .

当概率模型式中回归函数为线性函数时，即有

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

其中， $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 为未知参数，常称它们为**回归系数**。线代回归模型的“线性”是针对未知参数 $\beta_i (i = 0, 1, 2, \dots, p)$ 而言的。回归解释变量的线性是非本质的，因为解释变量是非线性时，常可以通过变量的替换把它转化成线性的。

如果 $(x_{i1}, x_{i2}, \dots, x_{ip}; y_i) (i = 1, 2, \dots, n)$ 是上式中变量 $(x_1, x_2, \dots, x_p; y)$ 的一组观测值，则线性回归模型可表示为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Theorem 9.0.3 为了估计模型参数的需要，古典线性回归模型通常应满足以下几个基本假设：

1. 解释变量 x_1, x_2, \dots, x_p 是非随机变量，观测值 $x_{i1}, x_{i2}, \dots, x_{ip}$ 是常数
2. 等方差及不相关的假定条件为

$$\begin{cases} E(\varepsilon_i) = 0, & i = 1, 2, \dots, n \\ \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} & i, j = 1, 2, \dots, n \end{cases}$$

这个条件称为高斯-马尔柯夫 (Gauss-Markov) 条件，简称 G-M 条件。在此条件下，便可以得到关于回归系数的最小二乘估计及误差项方差 σ^2 估计的一些重要性质，如回归系数的最小二乘估计是回归系数的最小方差线性无偏估计等。

3. 正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), & i = 1, 2, \dots, n \\ \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ 相互独立} \end{cases}$$

在此条件下便可得到关于回归系数的最小二乘估计及 σ^2 估计的进一步的结果，如它们分别是回归系数及 σ^2 的最小方差无偏估计等，并且可以进行回归的显著性检验及区间估计。

4. 通常为了便于数学上的处理，还要求 $n > p$ ，即样本量的个数要多于解释变量的个数。

Theorem 9.0.4 对线性回归模型通常要研究的问题有：

1. 如何根据样本 $((x_{i1}, x_{i2}, \dots, x_{ip}; y_i) | i = 1, 2, \dots, n)$ 求出 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 及方差 σ^2 的估计。
2. 对回归方程及回归系数的种种假设进行检验。
3. 如何根据回归方程进行预测和控制，以及如何进行实际问题的结构分析。

Definition 9.0.4 ——一元线性回归. 一元线性理论回归模型：

$$y = \beta_0 + \beta_1 x + \varepsilon$$

假设

$$\begin{cases} E(\varepsilon) = 0 \\ \text{var}(\varepsilon) = \sigma^2 \end{cases}$$

两边同时取期望

$$E(y|x) = \beta_0 + \beta_1 x$$

得到回归方程。以下把条件期望 $E(y|x)$ 简记为 $E(y)$ 。

一般情况下，对我们所研究的某个实际问题，如果获得的 n 组样本观测值 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 符合模型式，则

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

有

$$\begin{cases} E(\varepsilon_i) = 0 \\ \text{var}(\varepsilon_i) = \sigma^2 \end{cases} \quad i = 1, 2, \dots, n$$

上公式成为一元线性回归样本模型。

R 区分线性模型和线性方程。

对上式两边分别求数学期望和方差，得

$$E(y_i) = \beta_0 + \beta_1 x_i, \quad \text{var}(y_i) = \sigma^2, \quad i = 1, 2, \dots, n$$

表明随机变量 y_1, y_2, \dots, y_n 的期望不等，方差相等，因而 y_1, y_2, \dots, y_n 是独立的随机变量，但并不是同分布的。而 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 是独立同分布的随机变量。利用数据对两个 β 进行估计，得到

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

为 y 关于 x 的一元线性经验回归方程。下一步是假设检验，需要分布：

$$\varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, n$$

在 ε_i 服从正态分布的假定下，进一步有随机变量 y_i 也服从正态分布

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, 2, \dots, n$$

写成矩阵形式： $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

于是理论回归模型式表示为

$$\begin{cases} y = x\beta + \varepsilon \\ E(\varepsilon) = 0 \\ \text{var}(\varepsilon) = \sigma^2 I_n \end{cases}$$

其中， I_n 为 n 阶单位矩阵。

9.1 参数估计

9.1.1 最小二乘估计 OLSE

Definition 9.1.1 为了由样本数据得到回归参数 β_0 和 β_1 的理想估计值，我们将使用普通最小二乘估计（ordinary least square estimation, OLSE）。对每一个样本观测值 (x_i, y_i) ，最小二乘法考虑观测值 y_i 与其回归值 $E(y_i) = \beta_0 + \beta_1 x_i$ 的离差越小越好，综合考虑 n 个离差值，定义离差平方和为

$$\begin{aligned} Q(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - E(y_i))^2 \\ &= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

所谓最小二乘法，就是寻找参数 β_0, β_1 的估计值 $\hat{\beta}_0, \hat{\beta}_1$ ，使离差平方和达到极小，即寻找 $\hat{\beta}_0, \hat{\beta}_1$ ，满足

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

$\hat{\beta}_0, \hat{\beta}_1$ 就称为回归参数 β_0, β_1 的最小二乘估计。称

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

为 $y_i (i = 1, 2, \dots, n)$ 的回归拟合值，简称回归值或拟合值。称

$$e_i = y_i - \hat{y}_i$$

为 $y_i (i = 1, 2, \dots, n)$ 的残差。求解参数的最小二乘估计，由于离差平方和是二次函数，故对其求极值

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} \Big|_{\beta_0=\hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} \Big|_{\beta_1=\hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

经整理后，得正规方程组

$$\begin{cases} n\hat{\beta}_0 + (\sum_{i=1}^n x_i) \hat{\beta}_1 = \sum_{i=1}^n y_i \\ (\sum_{i=1}^n x_i) \hat{\beta}_0 + (\sum_{i=1}^n x_i^2) \hat{\beta}_1 = \sum_{i=1}^n x_i y_i \end{cases}$$

求解以上正规方程组得 β_0, β_1 的最小二乘估计为

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

其中

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

记

$$\begin{aligned} L_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2 \\ L_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \end{aligned}$$

则可简写为

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = L_{xy}/L_{xx} \end{cases}$$

易知， $\hat{\beta}_1$ 可以等价地表示为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

或

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n (\bar{x})^2}$$

由 $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ 可知

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

可见回归直线 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 是通过样本重心点 (\bar{x}, \bar{y}) 的直线。

Proposition 9.1.1 — 残差的性质.

$$\sum_{i=1}^n e_i = 0$$

$$\sum_{i=1}^n x_i e_i = 0$$

即残差的平均值为 0, 残差以自变量 x 的加权平均值为 0。

9.2 极大似然估计

Definition 9.2.1 最大似然估计 (maximum likelihood estimation, MLE) 也可以作为回归参数的估计方法。最大似然估计是利用总体的分布密度或概率分布的表达式及其样本所提供的信息求未知参数估计量的一种方法。

当总体 X 为连续型分布时, 设其分布密度族为 $\{f(x; \theta), \theta \in \Theta\}$, 假设总体 X 的一个独立同分布的样本为 x_1, x_2, \dots, x_n 。其似然函数为

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta)$$

最大似然估计应在一切 θ 中选取使随机样本 (X_1, X_2, \dots, X_n) 落在点 (x_1, x_2, \dots, x_n) 附近的概率最大的 $\hat{\theta}$ 为未知参数 θ 真值的估计值, 即选取 $\hat{\theta}$ 满足

$$L(\hat{\theta}; x_1, x_2, \dots, x_n) = \max_{\theta} L(\theta; x_1, x_2, \dots, x_n)$$

假设得到观测数据, 由于 MLE 需要概率密度函数, 故需要分布。假设 $\varepsilon_i \sim N(0, \sigma^2)$ 时, 知 y_i 服从如下正态分布

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

y_i 的分布密度为

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}, \quad i = 1, 2, \dots, n$$

于是 y_1, y_2, \dots, y_n 的似然函数为

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^n f_i(y_i) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\} \end{aligned}$$

由于 L 的极大化与 $\ln(L)$ 的极大化是等价的，所以取对数似然函数为

$$\ln(L) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

求极大值，等价于对 $\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$ 求极小值，到此又与最小二乘原理完全相同。因而 $\hat{\beta}_0, \hat{\beta}_1$ 的最大似然估计就是最小二乘估计。另外，由最大似然估计还可以得到 σ^2 的估计值为

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2\end{aligned}$$

这个估计量是 σ^2 的有偏估计。在实际应用中，常用无偏估计量

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2\end{aligned}$$

作为 σ^2 的估计量。

R 在此需要注意的是，以上最大似然估计是在 $\varepsilon_i \sim N(0, \sigma^2)$ 的正态分布假设下求得的，而最小二乘估计则对分布假设没有要求。两种估计的结果是一样的，但是一个需要残差分布的假设，一个不需要。

另外， y_1, y_2, \dots, y_n 是独立的正态分布样本，但并不是同分布的。期望值 $E(y_i) = \beta_0 + \beta_1 x_i$ 不相等，但这并不妨碍最大似然方法的应用。

Proposition 9.2.1 — 最小二乘估计的性质. 1. 线性：就是估计量 $\hat{\beta}_0, \hat{\beta}_1$ 为随机变量 y_i 的线性函数。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} y_i$$

其中， $\frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}$ 是 y_i 的常数，所以 $\hat{\beta}_1$ 是 y_i 的线性组合。同理可以证明 $\hat{\beta}_0$ 是 y 的线性组合。因为 y_i 为随机变量，所以 $\hat{\beta}_1$ 作为 y_i 的线性组合， $\hat{\beta}_1$ 亦为随机变量，因此，各有其概率分布、均值、方差、标准差及两者的协方差。

2. $\hat{\beta}_0, \hat{\beta}_1$ 的无偏性：由于 x_i 是非随机变量， $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $E(\varepsilon_i) = 0$ ，因而有

$$E(y_i) = \beta_0 + \beta_1 x_i$$

故

$$\begin{aligned}E(\hat{\beta}_1) &= \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} E(y_i) \\ &= \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1\end{aligned}$$

证得 $\hat{\beta}_1$ 是 β_1 的无偏估计, 其中用到 $\sum(x_i - \bar{x}) = 0$, $\sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})^2$ 。同理可证 $\hat{\beta}_0$ 是 β_0 的无偏估计。无偏估计的意义是, 如果屡次变更数据、反复求 β_0, β_1 的估计值, 这两个估计量没有高估或低估的系统趋向, 它们的平均值将趋于 β_0, β_1

$$\begin{aligned} E(\hat{y}) &= E(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \beta_0 + \beta_1 x_i \\ &= E(y) \end{aligned}$$

这表明回归值 \hat{y} 是 $E(y)$ 的无偏估计, 也说明 \hat{y} 与真实值 y 的平均值是相同的

3. 方差: 由 y_1, y_2, \dots, y_n 相互独立, $\text{var}(y_i) = \sigma^2$, 得

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \sum_{i=1}^n \left[\frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} \right]^2 \text{var}(y_i) \\ &= \frac{\sigma^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \end{aligned}$$

回归系数 $\hat{\beta}_1$ 不仅与随机误差的方差 σ^2 有关, 而且与自变量 x 的取值离散程度有关。如果 x 的取值比较分散, 即 x 的波动较大, 则 $\hat{\beta}_1$ 的波动就小, β_1 的估计值 $\hat{\beta}_1$ 就比较稳定。

$$\begin{aligned} \text{var}(\hat{\beta}_0) &= \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \sigma^2 \\ \hat{\beta}_0 &\sim N\left(\beta_0, \left(\frac{1}{n} + \frac{(\bar{x})^2}{L_{xx}}\right) \sigma^2\right) & \hat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right) \end{aligned}$$

另外, 还可得到 $\hat{\beta}_0, \hat{\beta}_1$ 的协方差

$$\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{L_{xx}} \sigma^2$$

上式说明, 在 $\bar{x} = 0$ 时, $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不相关, 在正态假定下独立; 在 $\bar{x} \neq 0$ 时不独立。它揭示了回归系数之间的关系状况。在前边我们曾给出回归模型随机误差项 ε_i 等方差及不相关的假定条件, 这个条件称为高斯-马尔柯夫条件, 即

$$\begin{cases} E(\varepsilon_i) = 0, & i = 1, 2, \dots, n \\ \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} & i, j = 1, 2, \dots, n \end{cases}$$

在此条件下可以证明, $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 分别是 β_0 与 β_1 的最佳线性无偏估计 (best linear unbiased estimator, BLUE), 也称为最小方差线性无偏估计。BLUE 即指在 β_0 和 β_1 的一切线性无偏估计中, 它们的方差最小。

进一步可知, 对固定的 x_0 来讲

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

也是 y_1, y_2, \dots, y_n 的线性组合, 且

$$\hat{y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_n}\right) \sigma^2\right)$$

由此可见, \hat{y}_0 是 $E(y_0)$ 的无偏估计, 且 \hat{y}_0 的方差随给定的 x_0 值与 \bar{x} 的距离 $|x_0 - \bar{x}|$ 的增大而增大。即当给定的 x_0 与 x 的样本平均值 \bar{x} 相差较大时, \hat{y}_0 的估计值波动就增大。这说明在实际应用回归方程进行控制和预测时, 给定的 x_0 值不能偏离样本均值太大, 否则, 用回归方程无论是作因素分析还是作预测, 效果都不会理想。

9.3 假设检验

9.3.1 t 检验

检验的原假设是

$$H_0: \beta_1 = 0$$

对立假设是

$$H_1: \beta_1 \neq 0$$

回归系数的显著性检验就是要检验自变量 x 对因变量 y 的影响程度是否显著。如果原假设 H_0 成立，则因变量 y 与自变量 x 之间并没有真正的线性关系，也就是说，自变量 x 的变化对因变量 y 并没有影响。由于， $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$ 因而当原假设 $H_0: \beta_1 = 0$ 成立时有

$$\hat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{L_{xx}}\right)$$

此时 $\hat{\beta}_1$ 在零附近波动，构造 t 统计量

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/L_{xx}}} = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

其中

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

是 σ^2 的无偏估计，称 $\hat{\sigma}$ 为回归标准差。统计量就是回归系数的最小二乘估计值除以其标准差的样本估计值。当原假设 $H_0: \beta_1 = 0$ 成立时， t 统计量服从自由度为 $n-2$ 的 t 分布。给定显著性水平 α ，双侧检验的临界值为 $t_{\alpha/2}$ 。当 $|t| \geq t_{\alpha/2}$ 时拒绝原假设 $H_0: \beta_1 = 0$ ，认为 β_1 显著不为零，因变量 y 对自变量 x 的一元线性回归成立，当 $|t| < t_{\alpha/2}$ 时接受原假设 $H_0: \beta_1 = 0$ ，认为 β_1 为零，因变量 y 对自变量 x 的一元线性回归不成立。

Definition 9.3.1 P-value 是显著性概率值 (significance probability value)

用 P 值代替 t 值作判定有几方面的优越性：

1. 用 P 值作检验不需要查表，只需直接用 P 值与显著性水平 α 相比，当 P 值 $\leq \alpha$ 时即拒绝原假设 H_0 ，当 P 值 $> \alpha$ 时即接受原假设 H_0 ，而用 t 值作检验需要查 t 分布表求临界值。
2. 用 P 值作检验具有可比性，而用 t 值作检验与自由度有关，可比性差。
3. 用 P 值作检验可以准确地知道检验的显著性，实际上 P 值就是犯弃真错误的真实概率，也就是检验的真实显著性。

9.3.2 F 检验

对线性回归方程显著性的另外一种检验是 F 检验，F 检验是根据平方和分解式，直接从回归效果检验回归方程的显著性。平方和分解式是

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

其中，

1. $\sum_{i=1}^n (y_i - \bar{y})^2$ 称为总离差平方和，简记为 SST 或 $S_{\text{总}}$ 或 L_{yy} ，SST 表示 Sum of Squares for Total。

方差来源	自由度	平方和	均方	F 值	P 值
回归	1	SSR	SSR/1	SSR/1	$P(F > F \text{ 值}) = P \text{ 值}$
残差	$n - 2$	SSE	$\text{SSE}/(n - 2)$		
总和	$n - 1$	SST			

2. $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 称为回归平方和, 简记为 SSR 或 $S_{\text{回}}$, R 表示 Regression
 3. $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 称为残差平方和, 简记为 SSE 或 $S_{\text{残}}$, E 表示 Error
 因而平方和分解式可以简写为

$$\text{SST} = \text{SSR} + \text{SSE}$$

检验统计量

$$F = \frac{\text{SSR}/1}{\text{SSE}/(n-2)}$$

在假设下, 当原假设 $H_0: \beta_1 = 0$ 成立时, F 服从自由度为 $(1, n-2)$ 的 F 分布。当 F 值大于临界值 $F_a(1, n-2)$ 时, 拒绝 H_0 , 说明回归方程显著, x 与 y 有显著的线性关系。一元线性回归方差分析

9.3.3 相关系数检验

由于一元线性回归方程讨论的是变量 x 与变量 y 之间的线性关系, 所以可以用变量 x 与 y 之间的相关系数来检验回归方程的显著性。设 $(x_i, y_i) (i = 1, 2, \dots, n)$ 是 (x, y) 的 n 组样本观测值, 我们称

$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \end{aligned}$$

为 x 与 y 的简单相关系数, 简称相关系数。其中, L_{xy}, L_{xx}, L_{yy} 与前面的定义相同。相关系数 r 表示 x 和 y 的线性关系的密切程度。 ± 1 表示正相关, 完全的线性。0 表示一点线性都没有

利用回归系数 $\hat{\beta}_1$ 的表达式可得

$$r = \frac{L_{sy}}{\sqrt{L_{xx}L_{xy}}} = \hat{\beta}_1 \sqrt{\frac{L_{xx}}{L_{xy}}}$$

由上式可以得到一个很有用的结论, 即一元线性回归的回归系数 $\hat{\beta}_1$ 的符号与相关系数 r 的符号相同。这里需要指出的是, 相关系数有个明显的缺点, 就是它接近 1 的程度与数据组数 n 有关, 这样容易给人一种假象。因为, 当 n 较小时, 相关系数的绝对值容易接近 1; 当 n 较大时, 相关系数的绝对值容易偏小。特别是当 $n = 2$ 时, 相关系数的绝对值总为 1。因此在样本量 n 较小时, 我们仅凭相关系数较大就说变量 x 与 y 之间有密切的线性关系, 就显得过于草率。

一元的情况, 三种假设的结果是一样的。

9.3.4 决定系数

由回归平方和与残差平方和的意义我们知道, 如果在总离差平方和中回归平方和所占的比重越大, 则线性回归效果就越好, 这说明回归直线与样本观测值的拟合优度越好; 如果残

差平方和所占的比重大，则回归直线与样本观测值拟合得就不理想。这里把回归平方和与总离差平方和之比定义为决定系数 (coefficient of determination)，也称为判定系数，确定系数，记为 r^2 ，即

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

由关系式

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

可以证明决定系数 r^2 正好是相关系数 r 的平方。即

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{L_{xy}^2}{L_{xx} L_{yy}} = (r)^2$$

决定系数 r^2 是一个回归直线与样本观测值拟合优度的相对指标，反映了因变量的变异中能用自变量解释的比例。其数值在 0 - 1 之间，可以用百分数表示。如果决定系数 r^2 接近 1，说明因变量不确定性的绝大部分能由回归方程解释，回归方程拟合优度就好；反之，如 r^2 不大，说明回归方程的效果不好，应进行修改，可以考虑增加新的自变量或者使用曲线回归。



1. 当样本量较小时，与前面在讲述相关系数时所强调的一样，此时即使得到一个大的决定系数，这个决定系数也很可能是虚假现象。为此可以结合样本量和自变量个数对决定系数做调整，计算调整的决定系数。
2. 即使样本量并不小，决定系数很大，例如 0.9，也不能肯定自变量与因变量之间的关系就是线性的，这是因为有可能曲线回归的效果更好。尤其是当自变量的取值范围很窄时，线性回归的效果通常较好，这样的线性回归方程是不能用于外推预测的。可以用模型失拟检验 (lack of fit test) 来判定因变量与自变量之间的关系是线性关系还是曲线关系，如果是曲线关系到底是哪一种曲线关系。这种检验需要对自变量有重复观测数据，而经济数据建模通常不能得到重复观测，这时可以用下面一节介绍的残差分析方法来判定回归方程的正确性。
3. 当你算出一个很小的决定系数 r^2 ，例如 $r^2 = 0.1$ 时，与相关系数的显著性检验相似，这时如果样本量 n 不大，就会得到线性回归不显著的检验结论，而在样本量 n 很大时，会得出线性回归显著的结论。不论检验结果是否显著，这时都应该尝试改进回归的效果，例如增加自变量，改用曲线回归等。

9.4 残差检验

之前的检验只能来说明线性模型是有效的，是可以用的，但是拟合的好还是不好要进行残差检验

残差 $e_i = y_i - \hat{y}_i$ ， n 对数据产生 n 个残差值。残差是实际观测值 y 与通过回归方程给出的回归值之差，残差 e_i 可以看做误差项 ε_i 的估计值。残差 $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ ，误差项 $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ ，比较两个表达式可以正确区分残差 e_i 与误差项 ε_i 的异同。

Proposition 9.4.1 — 残差的性质. 1. 残差是个估计，所以可以取数学期望 $E(e_i) = 0$

Proof.

$$\begin{aligned} E(e_i) &= E(y_i) - E(\hat{y}_i) \\ &= (\beta_0 + \beta_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0 \end{aligned}$$



2.

$$\begin{aligned}\text{var}(e_i) &= \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{L_{xx}}\right] \sigma^2 \\ &= (1 - h_{ii}) \sigma^2\end{aligned}$$

其中, $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{L_{xx}}$, 称为杠杆值。 $0 < h_{ii} < 1$, 当 x_i 靠近 \bar{x} 时, h_{ii} 的值接近 0, 相应的残差方差就大。当 x_i 远离 \bar{x} 时, h_{ii} 的值接近 1, 相应的残差方差就小。也就是说, 靠近 \bar{x} 附近的点相应的残差方差较大, 远离 \bar{x} 附近的点相应的残差方差较小。实际上, 远离 \bar{x} 的点数目必然较少, 回归线容易“照顾”到这样的少数点。使得回归线接近这些点, 因而远离 \bar{x} 附近的 x_i 相应的残差方差较小。

3. 残差满足约束条件: $\sum_{i=1}^n e_i = 0, \sum_{i=1}^n x_i e_i = 0$, 此这表明残差 e_1, e_2, \dots, e_n 是相关的, 不是独立的。

在残差分析中, 一般认为超过 $\pm 2\hat{\sigma}$ 或 $\pm 3\hat{\sigma}$ 的残差为异常值,

1. 标准化残差

$$\text{ZRE}_i = \frac{e_i}{\hat{\sigma}}$$

2. 学生化残差

$$\text{SRE}_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_i}}$$

标准化残差使残差具有可比性, $|\text{ZRE}_i| > 3$ 的相应观测值即判定为异常值, 这简化了判定工作, 但是没有解决方差不等的问题。学生化残差则进一步解决了方差不等的问题, 因而在寻找异常值时, 用学生化残差优于用普通残差, 认为 $|\text{SRE}_i| > 3$ 的相应观测值为异常值。

R 如何判断观察值是否存在异常值, 方差相等用 ZRE, 方差不等用 SRE

9.4.1 置信区间

在实际应用中, 我们主要关心回归系数 $\hat{\beta}_1$ 的精度, 因而这里只推导 $\hat{\beta}_1$ 的置信区间。由于 $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$ 可得

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / L_{xx}}} = \frac{(\hat{\beta}_1 - \beta_1) \sqrt{L_{xx}}}{\hat{\sigma}}$$

服从自由度为 $n - 2$ 的 t 分布。因而

$$P\left(\left|\frac{(\hat{\beta}_1 - \beta_1) \sqrt{L_{xx}}}{\hat{\sigma}}\right| < t_{a/2}(n - 2)\right) = 1 - \alpha$$

上式等价于

$$P\left(\hat{\beta}_1 - t_{a/2} \frac{\hat{\sigma}}{\sqrt{L_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{a/2} \frac{\hat{\sigma}}{\sqrt{L_{xx}}}\right) = 1 - \alpha$$

即得 β_1 的置信度为 $1 - \alpha$ 的置信区间为

$$\left(\hat{\beta}_1 - t_{a/2} \frac{\hat{\sigma}}{\sqrt{L_{xx}}}, \hat{\beta}_1 + t_{a/2} \frac{\hat{\sigma}}{\sqrt{L_{xx}}}\right)$$

对因变量的区间预测又分为两种情况：一种是因变量新值的区间预测；另一种是因变量新值的平均值的区间预测。1. 因变量新值的区间预测为了给出新值 y_0 的置信区间，需要首先求出其估计值 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 的分布。由于 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 都是 y_1, y_2, \dots, y_n 的线性组合，因而 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 也是 y_1, y_2, \dots, y_n 的线性组合，在正态假定下 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ 服从正态分布，其期望值为 $E(\hat{y}_0) = \beta_0 + \beta_1 x_0$ ，以下计算其方差，首先

$$\begin{aligned}\hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0 \\ &= \sum_{i=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}} \right] y_i\end{aligned}$$

因而有

$$\begin{aligned}\text{var}(\hat{y}_0) &= \sum_{i=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{L_{xx}} \right]^2 \text{var}(y_i) \\ &= \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right] \sigma^2\end{aligned}$$

从而得

$$\hat{y}_0 \sim N \left\{ \beta_0 + \beta_1 x_0, \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right] \sigma^2 \right\}$$

记

$$h_{00} = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}$$

为新值 x_0 的杠杆值，则上式简写为

$$\hat{y}_0 \sim N(\beta_0 + \beta_1 x_0, h_{00} \sigma^2)$$

\hat{y}_0 是先前独立观测到的随机变量 y_1, y_2, \dots, y_n 的线性组合，现在小麦产量的新值 y_0 与先前的观测值是独立的，所以 y_0 与 \hat{y}_0 是独立的。因而

$$\begin{aligned}\text{var}(y_0 - \hat{y}_0) &= \text{var}(y_0) + \text{var}(\hat{y}_0) \\ &= \sigma^2 + h_{00} \sigma^2\end{aligned}$$

再由回归值 \hat{y} 是 $E(y)$ 的无偏估计式知 $E(y_0 - \hat{y}_0) = 0$ ，于是有

$$y_0 - \hat{y}_0 \sim N(0, (1 + h_{00}) \sigma^2)$$

进而可知统计量

$$t = \frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}} \hat{\sigma} \sim t(n-2)$$

可得

$$P \left(\left| \frac{y_0 - \hat{y}_0}{\sqrt{1 + h_{00}}} \hat{\sigma} \right| \leqslant t_{a/2}(n-2) \right) = 1 - \alpha$$

由此可以求得 y_0 的置信概率为 $1 - \alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \sqrt{1+h_{00}} \hat{\sigma}$$

当样本量 n 较大, $|x_0 - \bar{x}|$ 较小时, h_{00} 接近零, y_0 的置信度为 95% 的置信区间近似为

$$\hat{y}_0 \pm 2\hat{\sigma}$$

提高精度: 增加样本量, 并且数据分散。 x_0 与样本均值越靠近越好

因变量新值的平均值的区间估计。对于前面提出的小麦产量问题, 如果该地区的一大片麦地每亩施肥量同为 x_0 , 那么这一大片地小麦的平均亩产如何估计呢? 这个问题就是要估计平均值 $E(y_0)$, $E(y_0)$ 的点估计仍为 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, 但是其区间估计却与因变量单个新值 y_0 的置信区间 (2.80) 式有所不同. 由于 $E(y_0) = \beta_0 + \beta_1 x_0$ 是常数, 由 (2.73) 式知

$$\hat{y}_0 - E(y_0) \sim N \left(0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}} \right) \sigma^2 \right)$$

进而可得置信水平为 $1 - \alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \sqrt{h_{00}} \hat{\sigma}$$

9.4.2 控制问题

要求 y 在一定范围内, 如何控制 x

$1 - \alpha$ 的概率保证把目标值 y 控制在 $T_1 < y < T_2$ 中, 即

$$P(T_1 < y < T_2) = 1 - \alpha$$

其中, α 是事先给定的小的正数, $0 < \alpha < 1$ 我们通常用近似的预测区间来确定 x . 如果 $\alpha = 0.05$, 根据近似置信区间

$$\hat{y}_0 \pm 2\hat{\sigma}$$

可由下述不等式组

$$\begin{cases} \hat{y}(x) - 2\hat{\sigma} > T_1 \\ \hat{y}(x) + 2\hat{\sigma} < T_2 \end{cases}$$

求出 x 的取值区间, 将 $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ 代入求得

当 $\hat{\beta}_1 > 0$ 时, 得

$$\frac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1}$$

当 $\hat{\beta}_1 < 0$ 时, 得

$$\frac{T_2 - 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1} < x < \frac{T_1 + 2\hat{\sigma} - \hat{\beta}_0}{\hat{\beta}_1}$$

10. 多元回归

Definition 10.0.1 设随机变量 y 与一般变量 x_1, x_2, \dots, x_p 的线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

其中, $\beta_0, \beta_1, \dots, \beta_p$ 是 $p+1$ 个未知参数, β_0 称为回归常数, β_1, \dots, β_p 称为回归系数。 y 称为被解释变量(因变量), x_1, x_2, \dots, x_p 是 p 个可以精确测量并控制的一般变量, 称为解释变量(自变量)。

随机误差项假定

$$\begin{cases} E(\varepsilon) = 0 \\ \text{var}(\varepsilon) = \sigma^2 \end{cases}$$

称

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

为理论回归方程。对一个实际问题, 如果我们获得 n 组观测数据 $(x_{i1}, x_{i2}, \dots, x_{ip}; y_i) (i = 1, 2, n)$, 则线性回归模型 (3.1) 式可表示为

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2 \\ \dots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

写成矩阵形式为

$$y = X\beta + \varepsilon$$

其中

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

X 是一个 $n \times (p+1)$ 阶矩阵, 称为回归设计矩阵或资料矩阵。

Theorem 10.0.1 — 基本假定.

1. 解释变量 x_1, x_2, \dots, x_p 是确定性变量, 不是随机变量, 且要求 $\text{rank}(X) = p+1 < n$
2. 这里的 $\text{rank}(X) = p+1 < n$, 表明设计矩阵 X 中的自变量列之间不相关, 样本量的个数应大于解释变量的个数, X 是一满秩矩阵。
3. 随机误差项具有零均值和等方差, 即

$$\begin{cases} E(\varepsilon_i) = 0, & i = 1, 2, \dots, n \\ \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} & i, j = 1, 2, \dots, n \end{cases}$$

这个假定常称为高斯-马尔柯夫条件。 $E(\varepsilon_i) = 0$, 即假设观测值没有系统误差随机误差项 ε_i 的平均值为零。随机误差项 ε_i 的协方差为零, 表明随机误差项在不同的样本点之间是不相关的 (在正态假定下即为独立的), 不存在序列相关, 并且有相同的精度。

4. 正态分布的假定条件为

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), & i = 1, 2, \dots, n \\ \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ 相互独立} \end{cases}$$

对于多元线性同归的矩阵模型, 这个条件便可表示为

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

由上述假定和多元正态分布的性质可知, 随机向量 y 服从 n 维正态分布, 回归模型式的期望向量

$$\begin{aligned} E(y) &= X\beta \\ \text{var}(y) &= \sigma^2 I_n \end{aligned}$$

因此

$$y \sim N(X\beta, \sigma^2 I_n)$$

对一般情况下含有 p 个自变量的多元线性回归而言, 每个回归系数 β_i 表示在回归方程中其他自变量保持不变的情况下, 自变量 x_i 每增加一个单位时因变量 y 的平均增加程度。因此也把多元线性回归的回归系数称为偏回归系数 (partial regression coefficient), 本书则仍简称为回归系数。

10.0.1 参数估计

最小二乘，依旧是考了离差平方和公式

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p) &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2 \\ &= \min_{\beta_0, \beta_1, \beta_2, \dots, \beta_p} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ij})^2 \end{aligned}$$

求偏导得到

$$\left\{ \begin{array}{l} \frac{\partial Q}{\partial \beta_0} \Big|_{\beta_0=\hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip}) = 0 \\ \frac{\partial Q}{\partial \beta_1} \Big|_{\beta_1=\hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip}) x_{i1} = 0 \\ \frac{\partial Q}{\partial \beta_2} \Big|_{\beta_2=\hat{\beta}_2} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip}) x_{i2} = 0 \\ \dots \\ \frac{\partial Q}{\partial \beta_p} \Big|_{\beta_p=\hat{\beta}_p} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip}) x_{ip} = 0 \end{array} \right.$$

以上方程组经整理后，得出用矩阵形式表示的正规方程组

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$$

$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$ 当 $(\mathbf{X}\mathbf{X}^{-1})$ 存在时，即得回归参数的最小二乘估计为 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 称

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

为经验回归方程。

向量 $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)'$ 为因变量向量 $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ 的回归值。由 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 可得

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

矩阵 $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的作用是把因变量向量 \mathbf{y} 变为拟合值向量 $\hat{\mathbf{y}}$ ，从形式上看是给 \mathbf{y} 戴上了一顶帽子，因而形象地称矩阵 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 为帽子矩阵，记为 H ，于是 $\hat{\mathbf{y}} = H\mathbf{y}$ 显然帽子矩阵 $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 是 n 阶对称矩阵，同时还是幂等矩阵，即 $H = H^2$ 。帽子矩阵 H 也是一个投影阵，从代数学的观点看， $\hat{\mathbf{y}}$ 是 \mathbf{y} 在自变量 \mathbf{X} 生成的空间上的投影，这个投影过程就是把 \mathbf{y} 左乘矩阵 H ，因此称 H 为投影阵。帽子矩阵 $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的主对角线元素记为 h_i ，可以证明，帽子矩阵 H 的迹为 $\text{tr}(H) = \sum_{i=1}^n h_{ii} = p+1$ 证明只需根据迹的性质 $\text{tr}(AB) = \text{tr}(BA)$ ，因而

$$\begin{aligned} \text{tr}(H) &= \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) \\ &= \text{tr}(\mathbf{I}_{p+1}) = p+1 \end{aligned}$$

称

$$e_i = y_i - \hat{y}_i$$

为 $y_i (i=1, 2, \dots, n)$ 的残差。称 $e = (e_1, e_2, \dots, e_n)' = \mathbf{y} - \hat{\mathbf{y}}$ 为回归残差向量。将 $\hat{\mathbf{y}} = H\mathbf{y}$ 代入得， $e = \mathbf{y} - H\mathbf{y} = (\mathbf{I} - H)\mathbf{y}$

记 $\text{cov}(e, e) = (\text{cov}(e_i, e_j))_{n \times n}$ 为残差向量 e 的协方差阵，或称为方差阵，记为 $D(e)$ 。因而

$$\begin{aligned} D(e) &= \text{cov}(e, e) \\ &= \text{cov}((I - H)y, (I - H)y) \\ &= (I - H)\text{cov}(y, y)(I - H)' \\ &= \sigma^2(I - H)I_n(I - H)' \\ &= \sigma^2(I - H) \end{aligned}$$

于是有

$$D(e_i) = (1 - h_{ii})\sigma^2, \quad i = 1, 2, \dots, n$$

根据 (3.17) 式可知，残差满足关系式

$$\left\{ \begin{array}{l} \sum e_i = 0 \\ \sum e_i x_{i1} = 0 \\ \dots \\ \sum e_i x_{ip} = 0 \end{array} \right.$$

即残差的平均值为 0，残差对每个自变量的加权平均为 0。可以用矩阵表示为 $X'e = 0$

误差项方差 σ^2 的无偏估计为

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-p-1} \text{SSE} = \frac{1}{n-p-1} (e'e) \\ &= \frac{1}{n-p-1} \sum_{i=1}^n e_i^2 \end{aligned}$$

证明注意 $E(\sum_{i=1}^n e_i^2) = \sum_{i=1}^n D(e_i)$ ，前边在由正规方程组求 $\hat{\beta}$ 时，要求 $(XX)^{-1}$ 必须存在，即 $X'X$ 是一非奇异矩阵

$$|X'X| \neq 0$$

由线性代数可知。 $X'X$ 为 $p+1$ 阶满秩矩阵

$$\text{rank}(X'X) = p+1$$

必须有

$$\text{rank}(X) \geq p+1$$

而 X 为 $n \times (p+1)$ 阶矩阵，于是应有

$$n \geq p+1$$

这是一个重要的结论，我们在多元线性回归模型的基本假定中用过它，这里就更清楚这个假定的重要意义。结论说明，要想用普通最小二乘法估计多元线性回归模型的未知参数，样本量必须不少于模型中参数的个数。在后面关于回归方程的假设检验中也少不了这一假设，否则检验无任何意义。

10.0.2 极大似然估计

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I_n) \end{aligned}$$

即 ε 服从多变量正态分布，那么 y 的概率分布为

$$y \sim N(X\beta, \sigma^2 I_n)$$

这时，似然函数为

$$L = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)\right)$$

其中的未知参数且 β 和 σ^2 ，最大似然估计就是选取使似然函数 L 达到最大的 $\hat{\beta}$ 和 $\hat{\sigma}^2$ 。要使 L 达到最大，对两边同时取自然对数，得

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)$$

仅在最后一项中含有 β ，显然使上式达到最大，等价于使

$$(y - X\beta)'(y - X\beta)$$

达到最小，这又完全与普通最小二乘估计一样。**故在正态假定下，回归参数 β 的最大似然估计与普通最小二乘估计完全相同**，即

$$\hat{\beta} = (X'X)^{-1} X'y$$

误差项方差 σ^2 的最大似然估计为

$$\sigma^2 = \frac{1}{n} \text{SSE} = \frac{1}{n} (e'e)$$

这是 σ^2 的有偏估计，但它满足一致性。在大样本的情况下，这是 σ^2 的渐近无偏估计。

Proposition 10.0.2 — 参数估计量的性质. 1. X 是固定的设计矩阵，因此， $\hat{\beta}$ 是 y 的一个线性变换。

2. $\hat{\beta}$ 是 β 的无偏估计

Proof.

$$\begin{aligned} E(\hat{\beta}) &= E\left((X'X)^{-1} X'y\right) \\ &= (X'X)^{-1} X'E(y) \\ &= (X'X)^{-1} X'E(X\beta + \varepsilon) \\ &= (X'X)^{-1} X'X\beta \\ &= \beta \end{aligned}$$

■

3. $D(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

Proof.

$$\begin{aligned}
 D(\hat{\beta}) &= \text{cov}(\hat{\beta}, \hat{\beta}) \\
 &= \text{cov}\left((X'X)^{-1}X'y, (X'X)^{-1}X'y\right) \\
 &= (X'X)^{-1}X'\text{cov}(y, y)(X'X)^{-1}' \\
 &= (X'X)^{-1}X'\sigma^2 X(X'X)^{-1} \\
 &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
 &= \sigma^2(X'X)^{-1}
 \end{aligned}$$

■

当 $p = 1$ 时即一元线性回归的情况，此时

$$\begin{aligned}
 X'X &= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \\
 (X'X)^{-1} &= \frac{1}{|X'X|} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \\
 &= \frac{1}{nL_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \\
 &= \begin{bmatrix} \frac{1}{nL_{xx}} \sum_{i=1}^n x_i^2 & -\frac{\bar{x}}{L_{xx}} \\ -\frac{\bar{x}}{L_{xx}} & \frac{1}{L_{xx}} \end{bmatrix}
 \end{aligned}$$

再由

$$D(\hat{\beta}) = \begin{pmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{var}(\hat{\beta}_1) \end{pmatrix}$$

4. 为了分析 $\hat{\beta}$ 各分量之间的相关程度，更方便的工具是采用 $\hat{\beta}$ 的相关阵。以一元线性回归为例， $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ 的相关阵为
5.

$$\begin{aligned}
 R(\hat{\beta}) &= \begin{bmatrix} 1 & \frac{\text{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_0)} \sqrt{\text{var}(\hat{\beta}_1)}} \\ \frac{\text{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_0)} \sqrt{\text{var}(\hat{\beta}_1)}} & 1 \end{bmatrix} \\
 &= \begin{bmatrix} 1 & -\frac{\bar{x}}{\sqrt{\frac{1}{n} \sum x_i^2}} \\ -\frac{\bar{x}}{\sqrt{\frac{1}{n} \sum x_i^2}} & 1 \end{bmatrix}
 \end{aligned}$$

6. 高斯-马尔柯夫 (G-M) 定理在实际应用中，我们关心的一个主要问题是预测。预测函数

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \cdots + \hat{\beta}_p x_{p0}$$

是 $\hat{\beta}$ 的线性函数，因而我们希望 $\hat{\beta}$ 的线性函数的波动越小越好。设 c 为任一 $p+1$ 维常数向量，我们希望回归系数向量 β 的估计值 $\hat{\beta}$ 具有如下性质：

- (a) $c' \hat{\beta}$ 是 $c' \beta$ 的无偏估计
- (b) $c' \beta$ 的方差要小。下面的一个重要性质告诉我们普通最小二乘估计 $\hat{\beta}$ 正好满足上述条件。

Theorem 10.0.3 — 高斯-马尔柯夫定理. 在假定 $E(y) = X\beta$, $D(y) = \sigma^2 I_n$ 时, β 的任一线性函数 $c'\beta$ 的最小方差线性无偏估计 (BLUE) 为 $c'\hat{\beta}$, 其中, c 是任一 $p+1$ 维常数向量, $\hat{\beta}$ 是 β 的最小二乘估计。

Proposition 10.0.4 定理说明了用普通最小二乘估计得到的 $\hat{\beta}$ 是理想的估计量。关于这条性质, 请读者注意以下四点:

1. 取常数向量 c 的第 $j (j = 0, 1, \dots, p)$ 个分量为 1, 其余分量为 0, 这时 G-M 定理表明最小二乘估计 $\hat{\beta}_j$ 是 β_j 的最小方差线性无偏估计。
2. 可能存在 y_1, y_2, \dots, y_n 的非线性函数, 作为 $c'\beta$ 的无偏估计, 比最小二乘估计 $c'\hat{\beta}$ 的方差更小。
3. 可能存在 $c'\hat{\beta}$ 的有偏估计, 在某种意义 (例如均方误差最小) 上比最小二乘估计 $c'\hat{\beta}$ 更好。
4. 在正态假定下, $c'\hat{\beta}$ 是 $c'\beta$ 的最小方差无偏估计。也就是说, 既不可能存在 y_1, y_2, \dots, y_n 的非线性函数, 也不可能存在 y_1, y_2, \dots, y_n 的其他线性函数, 作为 $c'\beta$ 的无偏估计, 比最小二乘估计 $c'\hat{\beta}$ 的方差更小。
5. $\text{cov}(\hat{\beta}, e) = \mathbf{0}$ 此性质说明与 $\hat{\beta}$ 与 e 不相关, 在正态假定下, $\hat{\beta}$ 与 e 不相关等价于 $\hat{\beta}$ 与 e 独立, 从而 $\hat{\beta}$ 与 $\text{SSE} = e'e$ 独立。
6. 当 $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ 时, 则
 - (a) $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$
 - (b) $\text{SSE}/\sigma^2 \sim \chi^2(n-p-1)$

10.1 显著性检验

F 检验对多元线性回归方程的显著性检验就是要看自变量 x_1, x_2, \dots, x_p 从整体上对随机变量 y 是否有明显的影响。为此提出原假设

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

为了建立对 H_0 进行检验的 F 统计量, 仍然利用总离差平方和的分解式, 即

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

简写为

$$\text{SST} = \text{SSR} + \text{SSE}$$

构造 F 检验统计量如下

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)}$$

在正态假设下, 当原假设 $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ 成立时, F 服从自由度为 $p, n-p-1$ 的 F 分布。于是, 可以利用 F 统计量对回归方程的总体显著性进行检验。对于给定的数据, $i = 1, 2, \dots, n$, 计算出 SSR 和 SSE , 进而得到 F 的值, 其计算过程列在方差分析表中, 再由给定的显著性水平 α 查 F 分布表, 得临界值 $F_a(p, n-p-1)$

显然, 如果某个自变量 x_j 对 y 的作用不显著, 那么在回归模型中, 它的系数 β_j 就取值为零。因此, 检验变量 x_j 是否显著, 等价于检验假设

$$H_{0j}: \beta_j = 0, \quad j = 1, 2, \dots, p$$

如果接受原假设 H_{0j} , 则 x_j 不显著: 如果拒绝原假设 H_{0j} , 则 x_j 是显著的。由于

$$\hat{\beta} \sim N\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

记

$$(X'X)^{-1} = (c_{ij}), \quad i, j = 0, 1, 2, \dots, p$$

于是有

$$\begin{aligned} E(\hat{\beta}_j) &= \beta_j, \text{var}(\hat{\beta}_j) = c_{jj}\sigma^2 \\ \hat{\beta}_j &\sim N(\beta_j, c_{jj}\sigma^2), \quad j = 0, 1, 2, \dots, p \end{aligned}$$

据此可以构造 t 统计量

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

其中

$$\hat{\sigma} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^n e_i^2} = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

是回归标准差。当原假设 $H_{0j}: \beta_j = 0$ 成立时, t_j 统计量服从自由度为 $n-p-1$ 的 t 分布。给定显著性水平 α , 查出双侧检验的临界值 $t_{a/2}$ 。当 $|t_j| \geq t_{a/2}$ 时, 拒绝原假设 $H_{0j}: \beta_j = 0$, 认为 β_j 显著不为零, 自变量 x_j 对因变量 y 的线性效果显著; 当 $|t_j| < t_{a/2}$ 时, 接受原假设 $H_{0j}: \beta_j = 0$, 认为 β_j 为零, 自变量 x_j 对因变量 y 的线性效果不显著。在一元线性回归中, 回归系数显著性的 t 检验与回归方程显著性的 F 检验是等价的, 而在多元线性回归中, 这两种检验是不等价的。F 检验显著, 说明 y 对自变量 x_1, x_2, \dots, x_p 整体的线性回归效果是显著的, 但不等于 y 对每个自变量 x_i 的效果都显著。反之, 某个或某几个 x_i 的系数不显著, 回归方程显著性的 F 检验仍有可能是显著的。可以从另外一个角度考虑自变量 x_j 的显著性。 y 对自变量 x_1, x_2, \dots, x_p 线性回归的残差平方和为 SSE, 回归平方和为 SSR, 在剔除掉 x_j 后, 用 y 对其余的 $p-1$ 个自变量作回归, 记所得的残差平方和为 $SSE_{(j)}$, 回归平方和为 $SSR_{(j)}$, 则自变量 x_j 对回归的贡献为 $\Delta SSR_{(j)} = SSR - SSR_{(j)}$, 称为 x_j 的偏回归平方和。由此构造偏 F 统计量

$$F_j = \frac{\Delta SSR_{(j)}/1}{SSE/(n-p-1)}$$

当原假设 $H_{0j}: \beta_j = 0$ 成立时, 偏 F 统计量 F_j 服从自由度为 $(1, n-p-1)$ 的 F 分布, 此 F 检验与 t 检验是一致的, 可以证明 $F_j = t_j^2$ 。当从回归方程中剔除变元时, 回归平方和减少, 残差平方和增加。根据平方和分解式可知, $\Delta SSR_{(j)} = \Delta SSE_{(j)} = SSE_{(j)} - SSE$ 。反之, 往回归方程中引入变元时, 回归平方和增加, 残差平方和减少, 两者的增减量同样相等。

三、回归系数的置信区间当我们有了参数向量 β 的估计量 $\hat{\beta}$ 时, $\hat{\beta}$ 与 β 的接近程度如何? 这就需构造 β_j 的一个区间, 以 $\hat{\beta}_j$ 为中心的区间, 该区间以一定的概率包含 β_j 由 (3.39) 式可知

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p-1)$$

一元线性回归系数区间估计的推导过程。可得 β_j 的置信度为 $1-\alpha$ 的置信区间为

$$(\hat{\beta}_j - t_{a/2}\sqrt{c_{jj}}\hat{\sigma}, \hat{\beta}_j + t_{a/2}\sqrt{c_{jj}}\hat{\sigma})$$

拟合优度用于检验回归方程对样本观测值的拟合程度。在一元线性回归中，定义了样本决定系数 $r^2 = SSR/SST$ ，在多元线性回归中，定义样本决定系数为

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

样本决定系数 R^2 的取值在 [0, 1] 区间内， R^2 越接近 1，表明同归拟合的效果越好； R^2 越接近 0，表明回归拟合的效果越差。与 F 检验相比， R^2 可以更清楚直观地反映回归拟合的效果，但不能作为严格的显著性检验。称

$$R = \sqrt{R^2} = \sqrt{\frac{SSR}{SST}}$$

为 y 关于 x_1, x_2, \dots, x_p 的样本复相关系数。在两个变量的简单相关系数中，相关系数有正负之分，而复相关系数表示的是因变量 y 与全体自变量之间的线性关系，它的符号不能由某一个自变量的回归系数的符号来确定，因而都取正号。用复相关系数 R 来表示回归方程对原有数据拟合程度的好坏，它衡量作为一个整体的 x_1, x_2, \dots, x_p 与 y 的线性关系。

n 比较大时， R^2 等于 0.7 左右比较好

产生舍入误差有两个主要原因：一是回归分析计算中数据量级有很大差异，比如数据 892 976 与 0.582 这样的大小相差悬殊的数据出现在同一个计算中；二是设计矩阵 X 的列向量近似线性相关时， $X'X$ 为病态矩阵，其逆矩阵 $(X'X)^{-1}$ 就会产生较大的误差。

多元线性回归模型的一般形式（3.1）式为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

其经验回归方程（3.19）式为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

此经验回归方程经过样本中心 $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p, \bar{y})$ ，将坐标原点移至样本中心。即作坐标变换

$$\begin{aligned} x'_{ij} &= x_{ij} - \bar{x}_j, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, p \\ y'_i &= y_i - \bar{y}, \quad i = 1, 2, \dots, n \end{aligned}$$

上述经验方程式转变为

$$\hat{y}' = \hat{\beta}_1 x'_1 + \hat{\beta}_2 x'_2 + \dots + \hat{\beta}_p x'_p$$

即中心化经验回归方程。中心化经验回归方程的常数项为 0，而回归系数的最小二乘估计值 $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 保持不变，这一点是容易理解的。这是因为坐标系的平移变换只改变直线的截距，不改变直线的斜率。

求解线性回归方程时，通常先对数据中心化，求出中心化经验方程式，再由

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 - \dots - \hat{\beta}_p \bar{x}_p$$

求出常数项估计值 $\hat{\beta}_0$

为了消除量纲不同和数量级的差异所带来的影响，就需要将样本数据作标准化处理，然后用最小二乘法估计未知参数、求得标准化回归系数。样本数据的标准化公式为

$$\begin{aligned} x_{ij}^* &= \frac{x_{ij} - \bar{x}_j}{\sqrt{L_{jj}}}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, p \\ y_i^* &= \frac{y_i - \bar{y}}{\sqrt{L_{yy}}}, \quad i = 1, 2, \dots, n \end{aligned}$$

其中

$$L_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

是自变量 $x_j (j = 1, 2, \dots, p)$ 的离差平方和。用最小二乘法求出标准化的样本数据 $(x_{i1}^*, x_{i2}^*, \dots, x_{ip}^*; y_i^*)$ 的经验回归方程，记为

$$\hat{y}^* = \hat{\beta}_1^* x_1^* + \hat{\beta}_2^* x_2^* + \dots + \hat{\beta}_p^* x_p^*$$

其中， $\hat{\beta}_1^*, \hat{\beta}_2^*, \dots, \hat{\beta}_p^*$ 为 y 对自变量 x_1, x_2, \dots, x_p 的标准化回归系数。标准化包括了中心化，因而标准化的回归常数项为 0。容易验证，标准化回归系数与普通最小二乘回归系数之间存在关系式

$$\hat{\beta}_j^* = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}} \hat{\beta}_j, \quad j = 1, 2, \dots, p$$

普通最小二乘估计 $\hat{\beta}_j$ 表示在其他变量不变的情况下，自变量 x_j 的每单位的绝对变化引起的因变量均值的绝对变化量。标准化回归系数 $\hat{\beta}^*$ 表示自变量 x_j 的 1% 相对变化（相对于 $\sqrt{L_{jj}}$ ）引起的因变量均值的相对变化百分数（相对于 $\sqrt{L_{yy}}$ ）。

当自变量所使用的单位不同时。用普通最小二乘估计建立的回归方程，其回归系数不具有可比性，得不到合理的解释。

复相关系数 R 反映了 y 与一组自变量的相关性，是整体和共性指标：简单相关系数反映的是两个变量间的相关性，是局部和个性指标。由样本观测值 $x_{i1}, x_{i2}, \dots, x_{ip} (i = 1, 2, \dots, n)$ ，分别计算 x_i 与 x_j 之间的简单相关系数 r_{ij} ，得自变量样本相关阵

$$r = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

注意相关阵是对称矩阵。记

$$X^* = (x_{ij}^*)_{n \times p}$$

表示中心标准化的设计阵，则相关阵可表示为

$$r = (X^*)' X$$

进一步求出 y 与每个自变量 x_i 的相关系数 r_{yi} ，得增广的样本相关阵为

$$\tilde{r} = \begin{bmatrix} 1 & r_{y1} & r_{y2} & \cdots & r_{yp} \\ r_{1y} & 1 & r_{12} & \cdots & r_{1p} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{py} & r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Definition 10.1.1 在多元线性回归分析中，当其他变量被固定后，给定的任两个变量之间的相关系数叫偏相关系数。偏相关系数可以度量 $p+1$ 个变量 y, x_1, x_2, \dots, x_p 之中任意两个变量的线性相关程度，而这种相关程度是在固定其余 $p-1$ 个变量的影响下的线性相关。

1. 两个自变量的偏决定系数二元线性回归模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

记 $\text{SSE}(x_2)$ 是模型中只含有自变量 x_2 时 y 的残差平方和, $\text{SSE}(x_1, x_2)$ 是模型中同时含有自变量 x_1 和 x_2 时 y 的残差平方和。因此, 模型中已含有 x_2 时, 再加入 x_1 使 y 的剩余变差的相对减少量为

$$r_{y_1:2}^2 = \frac{\text{SSE}(x_2) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_2)}$$

此即模型中已含有 x_2 时, y 与 x_1 的偏决定系数。同样地, 模型中已含有 x_1 时, y 与 x_2 的偏决定系数为

$$r_{y_2:1}^2 = \frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_1)}$$

2. 一般情况当模型中已含有 x_2, \dots, x_p 时, y 与 x_1 的偏决定系数为

$$r_{y_1:2,\dots,p} = \frac{\text{SSE}(x_2, \dots, x_p) - \text{SSE}(x_1, x_2, \dots, x_p)}{\text{SSE}(x_2, \dots, x_p)}$$

偏决定系数与回归系数显著性检验的偏 F 值是等价的。

偏决定系数的平方根称为偏相关系数, 其符号与相应的回归系数的符号相同。

偏相关系数与回归系数显著性检验的 t 值是等价的。

偏相关系数反映的是变量间的相关性, 因而并不需要有处于特殊地位的变量 y , 我们可以对任意 p 个变量 x_1, x_2, \dots, x_p 定义它们之间的偏相关系数。记

$$r_{ij} = \frac{L_{ij}}{\sqrt{L_{ii} \cdot L_{jj}}}$$

表示两个变量 x_i, x_j 之间的简单相关系数, $r = (r_{ij})_{p \times p}$ 为 x_1, x_2, \dots, x_p 的相关阵, 则在固定 x_3, \dots, x_p 保持不变时, x_1 与 x_2 之间的偏相关系数为

$$r_{12;3,\dots,p} = \frac{-\Delta_{12}}{\sqrt{\Delta_{11} \cdot \Delta_{22}}}$$

其余变量间偏相关系数的定义依此类推, 这个定义与用 (3.58) 式的平方根的定义是等价的。其中符号 Δ_{ij} 表示相关阵 $(r_{ij})_{p \times p}$ 第 i 行第 j 列元素的代数余子式, 注意相关阵 $(r_{ij})_{p \times p}$ 是对称矩阵。容易验证以下关系

$$r_{12:3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

在多元回归中, 应注意简单相关系数只是两变量局部的相关性质, 而并非整体的性质。所以在多元线性回归分析中我们并不看重简单相关系数, 而认为偏相关系数才是真正反映因变量 y 与自变量 x_i 以及自变量 x_i 与 x_j 的相关性的数值。根据偏相关系数, 可以判断哪些自变量对因变量的影响较大, 而选择必须考虑的自变量, 对于那些对因变量影响较小的自变量, 则可以舍去不顾。在剔除某个自变量时, 可以结合偏相关系数考虑。

在最小二乘前要检验数据是否满足假设的条件

10.2 违背基本假设的情况

针对之前的基本假设，很多情况是不满足的，分别是一种是计量经济建模中常说的异方差性。即

$$\text{var}(\varepsilon_i) \neq \text{var}(\varepsilon_j), \quad \text{当 } i \neq j \text{ 时}$$

另一种是自相关性，即

$$\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0, \quad \text{当 } i \neq j \text{ 时}$$

10.2.1 异方差性

导致的问题

1. 最小二乘估计量不再具有最小方差的优良性
2. 当存在异方差时，普通最小二乘估计存在以下问题：
 - (a) 参数估计值虽是无偏的，但不是最小方差线性无偏估计。
 - (b) 参数的显著性检验失效。
 - (c) 回归方程的应用效果极不理想。

如何检验

1. 残差图是否有规律，如果有则异方差性
2. 等级相关系数法又称斯皮尔曼（Spearman）检验。
 - (a) 第一步，作 y 关于 x 的普通最小二乘回归，求出 ε_i 的估计值，即 e_i 的值。
 - (b) 第二步，取 e_i 的绝对值，即 $|e_i|$ ，把 x_i 和 $|e_i|$ 按递增或递减的次序排列后分成等级，按下式计算出等级相关系数

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n d_i^2$$

其中， n 为样本量： d_i 为对应于 x_i 和 $|e_i|$ 的等级的差数。

- (c) 第三步，做等级相关系数的显著性检验。在 $n > 8$ 的情况下，用下式对样本等级相关系数 r_s 进行 t 检验。检验统计量为

$$t = \frac{\sqrt{n-2}r_s}{\sqrt{1-r_s^2}}$$

如果 $t \leq t_{a/2}(n-2)$ ，可以认为异方差性问题不存在；如果 $t > t_{a/2}(n-2)$ ，说明 x_i 与 $|e_i|$ 之间存在系统关系，异方差性问题存在。



计算残差绝对值 $|e_i|$ 与自变量 x_i 的相关性时采用 Spearman 等级相关系数，而不采用 Pearson 简单相关系数，这是由于等级相关系数可以反映非线性相关的情况，而简单相关系数不能如实反映非线性相关的情况。

Definition 10.2.1 加权最小二乘法 (weighted least square, WLS)

$$\begin{aligned} Q_w(\beta_0, \beta_1) &= \sum_{i=1}^n w_i (y_i - E(y_i))^2 \\ &= \sum_{i=1}^n w_i (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

其中， w_i 为给定的第 i 个观测值的权数。加权最小二乘估计就是寻找参数 β_0, β_1 的估计值 $\hat{\beta}_{0w}, \hat{\beta}_{1w}$ ，使离差平方和 Q_w 达到极小。如果所有的权数相等，即 w_i 都等于某个常数，

问题就成为普通最小二乘法。可以证明加权最小二乘估计为

$$\hat{\beta}_{0w} = \bar{y}_w - \hat{\beta}_{1w}\bar{x}_w \hat{\beta}_{1w} = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w) (y_i - \bar{y}_w)}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}$$

其中, $\bar{x}_w = \frac{1}{\sum w_i} \sum w_i x_i$ 为自变量的加权平均

$$\bar{y}_w = \frac{1}{\sum w_i} \sum w_i y_i \text{ 为因变量的加权平均。}$$

在使用加权最小二乘法时, 为了消除异方差性的影响, 使各项地位相同, 观测值的权数应该是观测值误差项方差的倒数, 即

$$w_i = \frac{1}{\sigma_i^2}$$

其中, σ_i^2 为第 i 个观测值误差项的方差。所以误差项方差较大的观测值接受较小的权数; 误差项方差较小的观测值接受较大的权数。在实际问题的研究中, 误差项的方差 σ_i^2 通常是未知的, 但是, 当误差项方差随自变量水平以系统的形式变化时, 我们可以利用这种关系。如, 已知误差项方差 σ_i^2 与 x_i^2 成比例, 那么 $\sigma_i^2 = kx_i^2$, 其中 k 为比例系数。

权数 w_i 为

$$w_i = \frac{1}{kx_i^2}$$

因为比例系数 k 在参数估计中可以消去, 所以可以直接使用权数

$$w_i = \frac{1}{x_i^2}$$

误差项方差与 x 的幂函数 x^m 成比例, 其中, m 为待定的未知参数。此时权函数为

$$w_i = \frac{1}{x_i^m}$$

Definition 10.2.2 — 多元线性回归模型. 对于一般的多元线性回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

当误差项 ε_i 存在异方差时, 加权离差平方和为

$$Q_w = \sum_{i=1}^n w_i (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2$$

其中, w_i 为给定的第 i 个观测值的权数。加权最小二乘估计就是寻找参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$

的估计值 $\hat{\beta}_{0w}, \hat{\beta}_{1w}, \hat{\beta}_{2w}, \dots, \hat{\beta}_{pw}$, 使 Q_w 达到极小。记

$$W = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix}$$

可以证明, 加权最小二乘估计的矩阵可表达为

$$\hat{\beta}_w = (X'WX)^{-1}X'Wy$$

Theorem 10.2.1 — 权函数的确定方法. 多元线性回归有多个自变量, 通常取权函数 W 为某个自变量 $x_j (j = 1, 2, \dots, p)$ 的幂函数, 即 $W = x_j^m$, 在 x_1, x_2, \dots, x_p 这 p 个自变量中, 应该取哪一个自变量呢? 只需计算每个自变量 x_j 与普通残差的等级相关系数, 选取等级相关系数最大的自变量构造权函数。

10.2.2 自相关

Definition 10.2.3 — 自相关. $cov(\varepsilon_i, \varepsilon_j) \neq 0$, 则称随机误差项之间存在自相关现象。这里的自相关现象不是指两个或两个以上的变量之间的相关关系, 而是指一个变量前后期数值之间的相关关系。

Theorem 10.2.2 — 序列相关性带来的问题.

1. 参数的估计值不再具有最小方差线性无偏性。
2. 均方误差 (MSE) 可能严重低估误差项的方差。
3. 容易导致对 t 值评价过高, 常用的 F 检验和 t 检验失效。如果忽视这一点可能导致得出回归参数统计检验为显著, 但实际上并不显著的严重错误结论。
4. 当存在序列相关时, $\hat{\beta}$ 仍然是 β 的无偏估计量, 但在任一特定的样本中, $\hat{\beta}$ 可能严重歪曲 β 的真实情况, 即最小二乘估计量对抽样波动变得非常敏感。
5. 如果不加处理地运用普通最小二乘法估计模型参数。用此模型进行预测和结构分析将会带来较大的方差甚至错误的解释。

Theorem 10.2.3 — 诊断方法.

1. 图示检验法: 绘制 e_t, e_{t-1} 的散点图。按照时间顺序绘制回归残差项 e_t 的图形, 如果随时间 t 变化有规律, 或者符号交替变化, 有蛛网现象。

2. 自相关系数法: 误差序列 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 的自相关系数定义为

$$\rho = \frac{\sum_{t=2}^n \varepsilon_t \varepsilon_{t-1}}{\sqrt{\sum_{t=2}^n \varepsilon_t^2} \sqrt{\sum_{t=2}^n \varepsilon_{t-1}^2}}$$

自相关系数 ρ 的取值范围是 $[-1, 1]$, 当 ρ 接近 1 时, 表明误差序列存在正相关, 当 ρ 接近 -1 时, 表明误差序列存在负相关。在实际应用中, 误差序列 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 的真实值是未知的, 需要用其估计值 e_t 代替, 得自相关系数的估计值为

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sqrt{\sum_{t=2}^n e_t^2} \sqrt{\sum_{t=2}^n e_{t-1}^2}}$$

$\hat{\rho}$ 作为自相关系数 ρ 的估计值与样本量有关，需要做统计显著性检验才能确定自相关性的存在，通常采用下面介绍的 DW 检验代替对 $\hat{\rho}$ 的检验。

3. DW 检验是杜宾 (J. Durbin) 和沃特森 (G. S. Watson) 提出的适用于小样本的一种检验方法。DW 检验只能用于检验随机扰动项具有一阶自回归形式的序列相关问题。随机扰动项的一阶自回归形式为

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

为了检验序列的相关性，构造的假设是

$$H_0 : \rho = 0$$

为了检验上述假设。构造 DW 统计量首先要求算出回归估计式的残差 e_t ，定义 DW 统计量为

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

其中， $e_t = y_i - \hat{y}_i (t = 1, 2, \dots, n)$

下面我们推导出 DW 值的取值范围。

$$DW = \frac{\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 - 2 \sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_t^2}$$

如果认为 $\sum_{t=2}^n e_t^2$ 与 $\sum_{t=2}^n e_{t-1}^2$ 近似相等，则由 (4.14) 式得

$$DW \approx 2 \left[1 - \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_t^2} \right]$$

同样，在认为 $\sum_{t=2}^n e_t^2$ 与 $\sum_{t=2}^n e_{t-1}^2$ 近似相等时，则由 (4.11) 式得

$$\hat{\rho} \approx \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=2}^n e_t^2} \quad (10.1)$$

因此，(4.15) 式可以写为

$$DW \approx 2(1 - \hat{\rho})$$

因而 DW 值与 $\hat{\rho}$ 的对应关系如表 4.4 所示。

$\hat{\rho}$	DW	误差项的白相关性
-1	4	完全负自相关
(-1,0)	(2,4)	负白相关
0	2	无自相关
(0,1)	(0,2)	正自相关
1	0	完全正自相关

$0 \leq DW \leq d_L$	误差项 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 间存在正自相关
$d_L < DW \leq d_U$	不能判定是否有自相关
$d_U < DW < 4 - d_U$	误差项 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 间无自相关
$4 - d_U \leq DW < 4 - d_L$	不能判定是否有自相关
$4 - d_L \leq DW \leq 4$	误差项 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 间存在负自相关



DW 检验缺点和局限性：

- (a) DW 检验有两个不能确定的区域，一旦 DW 值落在这两个区域，就无法判断，这时，只有增大样本量或选取其他方法。
- (b) DW 统计量的上、下界表要求 $n > 15$ ，这是因为样本如果再小，利用残差就很难对自相关的存在性作出比较正确的诊断。
- (c) DW 检验不适合随机项具有高阶序列相关的检验。

Theorem 10.2.4 — 自相关处理方法. 1. 迭代法：以一元线性回归模型为例，设一元线性同归模型的误差项存在一阶自相关

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad \begin{cases} E(u_t) = 0, & t = 1, 2, \dots, n \\ \text{cov}(u_t, u_s) = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s \end{cases} & t, s = 1, 2, \dots, n \end{cases}$$

第二个公式表明误差项 ε_t 存在一阶自相关，第三个公式表明 u_t 满足关于随机扰动项的基本假设

根据回归模型 (4.18) 式，有

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}$$

将 (4.21) 式两端乘以 ρ 。用 (4.18) 式成去乘以 ρ 的 (4.21) 式，则有

$$(y_t - \rho y_{t-1}) = (\beta_0 - \rho \beta_0) + \beta_1 (x_t - \rho x_{t-1}) + (\varepsilon_t - \rho \varepsilon_{t-1})$$

令

$$\begin{aligned} y'_t &= y_t - \rho y_{t-1} \\ x'_t &= x_t - \rho x_{t-1} \\ \beta'_0 &= \beta_0(1 - \rho), \beta'_1 = \beta_1 \end{aligned}$$

于是

$$y'_t = \beta'_0 + \beta'_1 x'_t + u_t$$

模型有独立随机误差项，它满足线性回归模型的基本假设，用普通最小二乘法估计的参数估计量具有通常的优良性。其中 $\hat{\rho} \approx 1 - \frac{1}{2}DW$ 。如果做一次没有消除，可以再次做检验再消去。

2. 差分法：差分法就是用增量数据代替原来的样本数据，将原来的回归模型变为差分形式的模型。一阶差分法通常适用于原模型存在较高程度的一阶自相关的情况。在迭代法 (4.22) 式中，当 $\rho = 1$ 时，得

$$(y_1 - y_{t-1}) = \beta_1 (x_t - x_{t-1}) + (\varepsilon_t - \varepsilon_{t-1})$$

以 $\Delta y_t = y_t - y_{t-1}$, $\Delta x_t = x_t - x_{t-1}$ 代之, 得

$$\Delta y_t = \beta_1 \Delta x_t + u_t$$

上式不存在序列的自相关, 它是以差分数据 Δy_t , 和 Δx_t 为样本的回归方程。对不带有常数项的回归方程仍用最小二乘法, 但它与前面的带有常数项的情形稍有不同, 它是回归直线过原点的回归方程。得

$$\hat{\beta}_1 = \frac{\sum_{t=2}^n \Delta y_t \Delta x_t}{\sum_{t=2}^n \Delta x_t^2}$$

一阶差分法的应用条件是自相关系数 $\rho = 1$, 在实际应用中, ρ 接近 1 时就采用差分法而不用迭代法。这有两个原因:

- (a) 迭代法需要用样本估计自相关系数 ρ 对 ρ 的估计误差会影响迭代法的使用效率;
- (b) 差分法比迭代法简单, 人们在建立时序数据的回归模型时, 更习惯于用差分法。

在自相关回归中, 回归预测值 \hat{y}_t 不是使用估计值 $\hat{\beta}_0 + \hat{\beta}_1 x_t$ 计算, 而是用

$$\hat{y}_t = \hat{\beta}'_0 + \hat{\rho} y_{t-1} + \hat{\beta}'_1 (x_t - \hat{\rho} x_{t-1})$$

计算出 \hat{y}_t 后, 再用 $y_t - \hat{y}_t$ 计算 e'_t , 这里 e'_t 是随机误差项 u_t 的估计值。另外一种计算 \hat{y}_t 的方法是对 $\hat{\beta}_0 + \hat{\beta}_1 x_t$ 做修正。在误差项没有自相关时, 我们实际上就是直接用估计值 $\hat{\beta}_0 + \hat{\beta}_1 x_t$ 作为回归预测值 \hat{y}_t 。现在误差项存在自相关 $\varepsilon_t = \rho e_{t-1} + u_t$, 需要从残差 e_t 中提取出有用的信息对估计值 $\hat{\beta}_0 + \hat{\beta}_1 x_t$ 做修正, 其中 $e_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_t)$ 是误差项 ε_t 的估计值。注意其中的系数估计值 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是按照关系式 $\hat{\beta}_0 = \hat{\beta}'_0 / (1 - \hat{\rho})$ 和 $\hat{\beta}_1 = \hat{\beta}'_1$ 根据迭代法的参数估计值推算的, 并不是普通最小二乘的估计值, 残差 e_t 也不是普通最小二乘的残差。计算过程如下:

$$\begin{aligned} t = 1 \text{ 时, 取 } \hat{y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 x_1, e_1 = y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1) \\ t \geq 2 \text{ 时, 取 } \hat{y}_t &= \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\rho} e_{t-1}, e_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_t) \end{aligned}$$

10.2.3 BOX-COX

Definition 10.2.4 — BOX-COX. BOX-COX 变换是对因变量 y 所作的如下变换:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

其中, λ 是待定参数。此变换要求 y 的各分量都大于 0, 否则可用下面推广的 BOX-COX 变换:

$$y^{(\lambda)} = \begin{cases} \frac{(y+a)^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y+a), & \lambda = 0 \end{cases}$$

即先对 y 做平移, 使得 $y+a$ 的各个分量都大于 0 后再做 BOX-COX 变换。对于不同的 λ , 所作的变换也不同, 所以这是一个变换族。它包含了一些常用变换, 如对数变换 $\lambda = 0$, 平方根变换 $\lambda = 1/2$ 和倒数变换 $\lambda = -1$ 。通过此变换, 我们寻找合适的 λ , 使

得变换后

$$y^{(\lambda)} = \begin{pmatrix} y_1^{(\lambda)} \\ y_2^{(\lambda)} \\ \vdots \\ y_n^{(\lambda)} \end{pmatrix} \sim N_n(\mathbf{X}\beta, \sigma^2 I)$$

从而符合线性回归模型的各项假设：误差各分量等方差、不相关等。事实上。BOX-COX 变换不仅可以处理异方差问题，还能处理自相关、误差非正态、回归函数非线性等情况。经过计算可得 λ 的最大似然估计

$$L_{\max}(\lambda) = (2\pi e \hat{\sigma}_\lambda^2)^{-\frac{n}{2}} |J|$$

其中， $\hat{\sigma}_\lambda^2 = \frac{1}{n} \text{SSE}(\lambda, y^{(\lambda)})$, $|J| = \prod_{i=1}^n \left| \frac{dy_i^{(\lambda)}}{dy_i} \right| = \prod_{i=1}^n y_i^{\lambda-1}$ 令 $y^{(\lambda)} = \frac{y^{(\lambda)}}{|J|}$, 对 $L_{\max}(\lambda)$ 取对数并略去与 λ 无关的常数项，可得：

$$\ln L_{\max}(\lambda) = -\frac{n}{2} \ln \text{SSE}(\lambda, z^{(\lambda)})$$

为找出 λ ，使得 $L_{\max}(\lambda)$ 达到最大，只需使 $\text{SSE}(\lambda, z^{(\lambda)})$ 达到最小即可。它的解析解比较难找，通常是给出一系列 λ 的值，计算对应的 $\text{SSE}(\lambda, z^{(\lambda)})$ ，取使得 $\text{SSE}(\lambda, z^{(\lambda)})$ 达到最小的 λ 即可。

10.2.4 异常点和强影响点

一元时画图就可以看出来，多元的时候根据情况采取不同的方法异常值分为两种情况：一种是关于因变量 y 异常；另一种是关于自变量 x 异常。以下分别讨论这两种情况。一、关于因变量 y 的异常值在残差分析中，认为超过 $\pm 3\hat{\sigma}$ 的残差为异常值。由于普通残差 e_1, e_2, \dots, e_n 的方差 $D(e_i) = (1 - h_{ii})\sigma^2$ 不等，用 e_i 作判断会带来一定的麻烦。类似于一元线性回归，在多元线性回归中，同样可以引入标准化残差 ZRE 和学生化残差 SRE 概念，以改进普通残差的性质。分别为：标准化残差

$$\text{ZRE}_i = \frac{e_i}{\hat{\sigma}}$$

学生化残差

$$\text{SRE}_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

其中， h_{ii} 为帽子矩阵 $H = X(X'X)^{-1}X'$ 的主对角线元素。标准化残差使残差具有可比性， $|ZRE_i| > 3$ 的相应观测值即判定为异常值，这简化了判定工作，但是没有解决方差不等的问题。学生化残差则进一步解决了方差不等的问题，比标准化残差又有所改进。但是当观测数据中存在关于 y 的异常观测值时，普通残差、标准化残差、学生化残差这三种残差都不再适用。这是由于异常值把回归线拉向自身，使异常值本身的残差减少，而其余观察值的残差增大，这时回归标准差 $\hat{\sigma}$ 也会增大，因而用 3σ 准则不能正确分辨出异常值。解决这个问题的方法是改用删除残差。删除残差的构造思想是：在计算第 i 个观测值的残差时，用删除掉这第 i 个观测值的其余 $n-1$ 个观测值拟合回归方程，计算出第 i 个观测值的删除拟合值 $\hat{y}_{(i)}$ 这个删除拟合值与第 i 个值无关，不受第 i 个值是否为异常值的影响，由此定义第 i 个观测值的删除残差为

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

删除残差 $e_{(i)}$ 较普通残差更能如实反映第 i 个观测值的异常性。可以证明

$$e_{(i)} = \frac{e_i}{1 - h_i}$$

进一步，我们可以给出第 i 个观测值的删除学生化残差，记为 $SRE_{(i)}$ 。删除学生化残差 $SRE_{(i)}$ 的公式推导比较复杂，本书在此不加证明直接给出其表达式

$$SRE_{(i)} = SRE_i \left(\frac{n-p-2}{n-p-1-SRE_i^2} \right)^{\frac{1}{2}}$$

二、关于自变量 x 的异常值对回归的影响 $D(e_i) = (1 - h_{ii})\sigma^2$ ，其中， h_{ii} 为帽子矩阵中主对角线的第 i 个元素，它是调节 e_i 方差大小的杠杆，因而称 h_{ii} 为第 i 个观测值的杠杆值。类似于一元线性回归，多元线性回归的杠杆值 h_{ii} 也是表示自变量的第 i 次观测值与自变量平均值之间距离的远近。根据 (3.24) 式，较大的杠杆值的残差偏小，这是因为杠杆值大的观测点远离样本中心，能够把回归方程拉向自身，因而把杠杆值大的样本点称为强影响点。强影响点并不一定是 y 值的异常值点，因而强影响点并不总会对回归方程造成不良影响。但是强影响点对回归效果通常有较强的影响，我们对强影响点应该有足够的重视，这是由于以下两个原因：(1) 在实际问题中，因变量与自变量的线性关系只是在一定的范围内成立，强影响点远离样本中心，因变量与自变量之间可能不再是线性函数关系，因而在选择回归函数的形式时，要侧重于强影响点；(2) 即使线性回归形式成立，但是强影响点远离样本中心，能的把回归方程拉向自身，使回归方程产生偏移。

由于强影响点并不总是 y 的异常值点，因而不能单纯根据杠杆值 h_{ii} 的大小判断强影响点是否异常。为此，我们引入库克距离，用来判断强影响点是否为 y 的异常值点。库克距离的计算公式为

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

库克距离反映了杠杆值 h_i 与残差 e_i 大小的一个综合效应。根据 (3.22) 式， $\text{tr}(H) = \sum_{i=1}^n h_{ii} = p+1$ ，则杠杆值 h_{ii} 的平均值为

$$\bar{h} = \frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{p+1}{n}$$

这样，一个杠杆值 h_{ii} 如果大于 2 倍或 3 倍的 \bar{h} 就认为是大的。一个粗略的标准是：当 $D_i < 0.5$ 时，认为不是异常值点；当 $D_i > 1$ 时，认为是异常值点。

10.2.5 总结

R 需要注意的是，加权最小二乘估计并不能消除异方差，只是能够消除或减弱异方差的不良影响。当存在异方差时，普通最小二乘估计不再具有最小方差线性无偏估计等良好的性质，而加权最小二乘估计可以改进估计的性质。加权最小二乘估计给误差项方差小的项加一个大的权数，给误差项方差大的项加一个小的权数，因此加强了小方差项的地位，使离差平方和中各项的作用相同。如果把误差项加权，那么加权的误差项 $\sqrt{w_i}e_i$ 是等方差的。从残差图来看，普通最小二乘估计只能照顾到残差大的项，而小残差项往往有整体的正偏或负偏。加权最小二乘估计的残差图，对大残差和小残差拟合得都好，大残差和小残差都没有整体的正偏或负偏。

常见的变量变换有如下几种：

1. 如果 σ_i^2 与 $E(y_i)$ 存在一定的比例关系，使用 $y' = \sqrt{y}$
2. 如果 σ_i 与 $E(y_i)$ 存在一定的比例关系，使用 $y' = \ln(y)$
3. 如果 $\sqrt{\sigma_i}$ 与 $E(y_i)$ 存在一定的比例关系，使用 $y' = \frac{1}{y}$

方差稳定变换在改变误差项方差的同时，也会改变误差项的分布和回归函数的形式。

1. 因而当误差项服从正态分布, 因变量与自变量之间遵从线性回归函数关系, 只是误差项存在异方差时, 应该采用加权最小二乘估计, 以消除异方差的影响。
2. 当误差项不仅存在异方差, 而且误差项不服从正态分布, 因变量与自变量之间也不尊从线性回归函数关系时, 应该采用方差稳定变换。

用迭代法处理序列相关并不总是有效。主要原因是当误差项正自相关时。10.1 式往往低估自相关参数 ρ 。如果这种偏差严重, 就会显著地降低迭代法的效率。对于误差项一阶自相关回归模型 $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ 式, 用迭代法得到的 (4.28) 式回归方程适于作短期预测。如果要作长期预测, 可直接使用回归方程 $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$ 这里 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 不是普通最小二乘估计值, 而是根据?? 式, 用公式 $\hat{\beta}_0 = \hat{\beta}'_0 / (1 - \hat{\rho})$ 和 $\hat{\beta}_1 = \hat{\beta}'_1$ 针换得到的。一阶差分法是自相关参数 $\rho = 1$ 时的迭代法, 一阶差分模型的一个重要特征是它没有截距项, 得到的差分回归线通过原点。一阶差分法是对原始数据的一种修正, 有时一阶差分法可能会过度修正, 使得差分数据中出现负自相关的误差项。因此, 从一定意义上说, 差分法要慎用。只有当 $\rho = 1$ 或者接近 1 时, 差分法的效果才会好。



DW 检验也有一些局限性。尤其是 DW 检验有两个不能确定结果的区域, 对于这种状况, 一般需要增大样本量。但在实际问题的研究中, 样本量的获取往往受到一定限制。为了克服 DW 检验的这一局限, J. Durbin 和 G. S. Watson 给出了一个近似的检验, 在使用下界 d_L 和上界 d_U 的 DW 检验得不到确定结果时可以使用。另外在 DW 表中, 变量个数较多, 样本量 n 较小时会出现 $d_U > 2$ 的情形, 这正是这种方法的一个不太合理的地方。在多元线性回归中, 一定要注意 n 与 p 的匹配问题。回归检验法也很受人们的推崇。回归检验法需要首先应用普通最小二乘法估计模型并求出 ε 的估计值 e , 然后以 e_t 为被解释变量, 以各种可能的相关量, 诸如 e_{t-1}, e_{t-2} 等作为解释变量分别进行线性拟合

$$\begin{aligned} e_t &= \beta e_{t-1} + u_t \\ e_t &= \beta_1 e_{t-1} + \beta_2 e_{t-2} + u_t \end{aligned}$$



如果一个异常值数据是准确的, 但是找不到对它的合理解释, 与剔除这个观测值相比, 一个更稳健的方法是抑制它的影响。最小绝对离差和法是一种稳健估计方法, 它具有对异常值和不合适的模型不敏感的性质。最小绝对离差和法是寻找参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$, 使绝对离差和达到极小, 即寻找 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$, 满足

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p) &= \sum_{i=1}^n |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip}| \\ &= \min_{\beta_0, \beta_1, \beta_2, \dots, \beta_p} \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip}| \end{aligned}$$

依照求出的 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 就称为回归参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 的最小绝对离差和估计

11. 自变量的选择和逐步回归和多重共线性

11.1 自变量的选择

设我们研究的某一实际问题涉及对因变量有影响的因素共有 m 个，由因变量 y 和 m 个自变量 x_1, x_2, \dots, x_m 构成的回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \varepsilon \quad (11.1)$$

因为模型 11.1 式是因变量 y 与所有自变量 x_1, x_2, \dots, x_m 的回归模型，故称 11.1 式为全回归模型。如果从所有可供选择的 m 个变量中挑选出 p 个，记为 x_1, x_2, \dots, x_p ，由所选的 p 个自变量组成的同归模型为

$$y = \beta_{0p} + \beta_{1p} x_1 + \beta_{2p} x_2 + \dots + \beta_{pp} x_p + \varepsilon_p \quad (11.2)$$

相对全模型而言，我们称模型 11.2 式为选模型。选模型 11.2 式的 p 个自变量。

为了方便，把模型 11.1 式的参数向量 β 和 σ^2 的估计记为

$$\hat{\beta}_m = (X'_m X_m)^{-1} X'_m y$$
$$\hat{\sigma}_m^2 = \frac{1}{n-m-1} SSE_m$$

把模型 11.2 式的参数向量 β 和 σ^2 的估计记为

$$\hat{\beta}_p = (X'_p X_p)^{-1} X'_p y$$

$$\hat{\sigma}_p^2 = \frac{1}{n-p-1} SSE_p$$

Theorem 11.1.1 — 应该使用全模型而使用选模型的后果.

- 在 x_j 与 x_{p+1}, \dots, x_m 的相关系数不全为 0 时，选模型回归系数的最小二乘估计是全模型相应参数的有偏估计，即 $E(\hat{\beta}_{jp}) = \beta_{ip} \neq \beta_j (j = 1, 2, \dots, p)$
- 选模型的预测是有偏的。给定新自变量值， $x_{om} = (x_{01}, x_{02}, \dots, x_{om})'$ ，因变量新值为 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_m x_{0m} + \varepsilon_0$ ，用选模型的预测值 $\hat{y}_{0p} = \hat{\beta}_{0p} + \hat{\beta}_{1p} x_{01} + \hat{\beta}_{2p} x_{02} + \dots + \hat{\beta}_{pp} x_{0p}$ ，作为 y_0 的预测值是有偏的，即 $E(\hat{y}_{0p} - y_0) \neq 0$

3. 选模型的参数估计有较小的方差。选模型的最小二乘参数估计为 $\hat{\beta}_p =$
4. $(\hat{\beta}_{0p}, \hat{\beta}_{1p}, \dots, \hat{\beta}_{pp})'$, 全模型的最小二乘参数估计为 $\hat{\beta}_m = (\hat{\beta}_{0m}, \hat{\beta}_{1m}, \dots, \hat{\beta}_{mm})'$, 这一性质说明 $D(\hat{\beta}_{jp}) \leq D(\hat{\beta}_{jm}) (j = 0, 1, \dots, p)$
5. 选模型的预测残差有较小的方差。选模型的预测残差为 $e_{0p} = \hat{y}_{0p} - y_0$ 全模型的预测残差为 $e_{0m} = \hat{y}_{0m} - y_0$, 其中 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_m x_{0m} + \epsilon$, 则有 $D(e_{0p}) \leq D(e_{0m})$
6. 记 $\beta_{m-p} = (\beta_{p+1}, \dots, \beta_m)'$, 用全模型对 β_{m-p} 的最小二乘估计为 $\hat{\beta}_{m-p} = (\hat{\beta}_{p+1}, \dots, \hat{\beta}_m)'$, 则在 $D(\hat{\beta}_{m-p}) \geq \beta_{m-p} \beta'_{m-p}$ 的条件下, $E(e_{0p})^2 = D(e_{0p}) + (E(e_{0p}))^2 \leq D(e_{0m})$, 即选模型预测的均方误差比全模型预测的方差更小。

(R)

1. 性质 1 和性质 2 表明, 当全模型正确时, 而我们舍去了 $m-p$ 个自变量, 用剩下的 p 个自变量去建立选模型, 参数估计值是全模型相应参数的有偏估计, 用其作预测, 预测值也是有偏的。这是误用选模型产生的弊端。
2. 性质 3 和性质 4 表明, 用选模型去作预测, 残差的方差比用全模型去作预测的方差小, 尽管用选模型所作的预测是有偏的, 但得到的预测残差的方差下降了。这说明尽管全模型正确, 误用选模型是有弊也有利的。
3. 性质 5 说明即使全模型正确, 但如果其中有一些自变量对因变量影响很小或回归系数方差过大, 我们丢掉这些变量之后, 用选模型去预测可以提高预测的精度。由此可见, 如果模型中包含了一些不必要的自变量, 模型的预测精度就会下降。

11.1.1 最优子集

在第 3 章, 我们从数据与模型拟合优劣的角度出发, 认为残差平方和 SSE 最小的回归方程就是最好的, 还用复相关系数 R 来衡量回归拟合的好坏。然而下面的讨论将会看到上述两种方法都有明显的不足。我们把选模型的残差平方和记为 SSE_{p+1} 时, 相应的残差平方和记为 SSE_{p+1} 。根据最小二乘估计的原理, 增加自变量时残差平方和将减少, 减少自变量时残差平方和将增加。因此有

$$SSE_{p+1} \leq SSE_p$$

又记它们的复决定系数分别为: $R_{p+1}^2 = 1 - SSE_{p+1}/SST$, $R_p^2 = 1 - SSE_p/SST$ 。由于 SST 是因变量的离差平方和, 与自变量无关, 因而

$$R_{p+1}^2 \geq R_p^2$$

即当自变量子集在扩大时, 残差平方和随之减少, 而复决定系数 R^2 随之增大。因此, 如果按残差平方和越小越好的原则来选择自变量子集, 或者为提高复决定系数, 不论什么变量只要多取就行, 则毫无疑问选的变量越多越好。这样由于变量的多重共线性, 给变量的回归系数估计值带来不稳定性, 加上变量的测量误差积累和参数数目增加, 将使估计值的误差增大。如此构造的回归模型稳定性差, 使得为增大复相关系数 R 而付出了模型参数估计稳定性差的代价。因此残差平方和、复相关系数或样本决定系数都不能作为选择变量的准则。下面从不同的角度给出几个常用的准则。

准则 1 自由度调整复决定系数达到最大。当给模型增加自变量时, 复决定系数也随之逐步增大, 然而复决定系数增大的代价是残差自由度的减少, 因为残差自由度等于样本个数与自变量个数之差。自由度小意味着估计和预测的可靠性低。这表明一个回归方程涉及的自变量很多时, 回归模型的拟合从表面上看是良好的, 而区间预测和区间估计的幅度则变大, 以至失去实际意义。这里回归模型的拟合良好掺杂了一些虚假成分。为了克服样本决定系数的这一缺点, 我们设法对 R^2 给予适当的修正, 使得只有加入有意义的变量时, 经

过修正的样本决定系数才会增加，这就是所谓的自由度调整复决定系数。设 R_a^2 为调整的复决定系数， n 为样本量， p 为自变量的个数，则

$$R_a^2 = 1 - \frac{n-1}{n-p-1} (1-R^2)$$

显然有 $R_a^2 \leq R^2$ ， R_a^2 随着自变量的增加并不一定增大。尽管 $1-R^2$ 随着变量的增加而减少，但由于其前面的系数 $(n-1)/(n-p-1)$ 起折扣作用，才使 R_a^2 随着自变量的增加并不一定增大。当所增加的自变量对回归的贡献很小时， R_a^2 反而可能减少。

在一个实际问题的回归建模中，自由度调整复决定系数 R_a^2 越大，所对应的回归方程越好。从拟合优度的角度追求最优，则所有回归子集中 R_a^2 最大者对应的回归方程就是最优方程。从另外一个角度考虑回归的拟合效果，回归误差项方差 σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{1}{n-p-1} SSE$$

此无偏估计式中也加入了惩罚因子 $n-p-1$ ， $\hat{\sigma}^2$ 实际上就是用自由度 $n-p-1$ 作平均的平均残差平方和。当自变量个数从 0 开始增加时，SSE 逐渐减小，作为除数的惩罚因子 $n-p-1$ 也随之减少。一般来说，当自变量个数从 0 开始增加时， $\hat{\sigma}^2$ 先是开始下降，而后稳定下来，当自变量个数增加到一定数量后， $\hat{\sigma}^2$ 又开始增加。这是因为刚开始时，随着自变量个数的增加，SSE 能如快速减小，虽然作为除数的惩罚因子 $n-p-1$ 也随之减小，但由于 SSE 成小的速度更快，因而 $\hat{\sigma}^2$ 是趋于减小的。当自变量数目增加到一定程度，重要的自变量基本都选上了，这时再增加自变量，SSE 减少的幅度不大，以至于抵消不了除数 $n-p-1$ 的减小，最终又导致了 $\hat{\sigma}^2$ 的增加。

这两个准则是等价的，

$$R_a^2 = 1 - \frac{n-1}{SST} \hat{\sigma}^2$$

由于 SST 是与回归无关的固定值。

准则 2 赤池信息量 AIC 达到最小。赤池信息量准则 (Akaike information criteri on, AIC)。AIC 准则既可用来作回归方程自变量的选择，又可用于时间序列分析中自回归模型的定阶。

对一般情况，设模型的似然函数为 $L(\theta, x)$ ， θ 的维数为 p ， x 为随机样本 (在回归分析中随机样本为 $y = (y_1, y_2, \dots, y_n)'$)，则 AIC 定义为

$$AIC = -2 \ln L(\hat{\theta}_L, x) + 2p$$

其中， $\hat{\theta}_L$ 为 θ 的最大似然估计； p 为未知参数的个数。式中右边第一项是似然函数的对数乘以 -2，第二项惩罚因子是未知参数个数的 2 倍。我们知道，似然函数越大的估计量越好，而 AIC 是似然函数的对数乘-2 再加上惩罚因子 $2p$ ，因而使 AIC 达到最小的模型是最优模型。下面我们讨论把 AIC 用于回归模型的选择。假定回归模型的随机误差项 ε 服从正态分布，即

$$\varepsilon \sim N(0, \sigma^2)$$

在这个正态假定下，回归参数的最大似然估计

$$\ln L_{\max} = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2} SSE$$

将 $\hat{\sigma}_L^2 = \frac{1}{n} SSE$ 代入得

$$\ln L_{\max} = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln\left(\frac{SSE}{n}\right) - \frac{n}{2}$$

这里似然函数中的未知参数个数为 $p+2$, 略去与 p 无关的常数, 得回归模型的 AIC 公式为

$$\text{AIC} = n \ln(\text{SSE}) + 2p$$

在回归分析的建模过程中, 对每一个回归子集计算 AIC, 其中 AIC 最小者所对应的模型是最优回归模型。准则 3 C_p 统计量达到最小。马洛斯 (Mallows) 根据性质 5, 即使全模型正确, 但仍有可能选模型有更小的预测误差。 C_p 正是根据这一原理提出来的。考虑在 n 个样本点上, 用选模型 (5.2) 式作回归预测, 预测值与期望值的相对偏差平方和为

$$\begin{aligned} J_p &= \frac{1}{\sigma^2} \sum_{i=1}^n (\hat{y}_{ip} - E(y_i))^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n \left(\hat{\beta}_{0p} + \hat{\beta}_{1p}x_{i1} + \cdots + \hat{\beta}_{pp}x_{ip} - (\beta_0 + \beta_1x_{i1} + \cdots + \beta_mx_{im}) \right)^2 \end{aligned}$$

可以证明, J_p 的期望值是

$$E(J_p) = \frac{E(\text{SSE}_p)}{\sigma^2} - n + 2(p+1)$$

略去无关的常数 2, 据此构造出 C_p 统计量为

$$\begin{aligned} C_p &= \frac{\text{SSE}_p}{\hat{\sigma}^2} - n + 2p \\ &= (n-m-1) \frac{\text{SSE}_p}{\text{SSE}_m} - n + 2p \end{aligned}$$

其中, $\hat{\sigma}^2 = \frac{1}{n-m-1} \text{SSE}_m$ 为全模型中 σ^2 的无偏估计。这样我们得到一个选择变量的 C_p 准则: 选择使 C_p 最小的自变量子集, 这个自变量子集对应的回归方程就是最优回归方程。

11.2 逐步回归

这一部分介绍的三种选择自变量个数的方法都是采取 $F_j = \frac{\Delta \text{SSR}_{(j)}/1}{\text{SSE}/(n-p-1)}$

11.2.1 前进法

前进法的思想是变量由少到多, 每次增加一个, 直至没有可引入的变量为止。具体做法是首先将全部 m 个自变量分别对因变量 y 建立 m 个一元线性回归方程并分别计算这 m 个一元回归方程的 m 个回归系数的 F 检验值, 记为 $\{F_1^1, F_2^1, \dots, F_m^1\}$, 选其最大者记为

$$F_j^1 = \max \{F_1^1, F_2^1, \dots, F_m^1\}$$

给定显著性水平 α , 若 $F_j^1 \geq F_a(1, n-2)$, 则首先将 x_j 引入回归方程, 为了方便设 x_j 就是 x_1 。接下来因变量 y 分别与 $(x_1, x_2), (x_1, x_3), \dots, (x_1, x_m)$ 建立 $m-1$ 个二元线性回归方程, 对这 $m-1$ 个回归方程中 x_2, x_3, \dots, x_m 的回归系数进行 F 检验, 计算 F 值, 记为 $\{F_2^2, F_3^2, \dots, F_m^2\}$, 选其最大者记为

$$F_j^2 = \max \{F_2^2, F_3^2, \dots, F_m^2\}$$

若 $F_j^2 \geq F_a(1, n-3)$, 则接着将 x_j 引入回归方程。依上述方法接着做下去, 直至所有未被引入方程的自变量的 F 值均小于 $F_a(1, n-p-1)$ 时为止。这时, 得到的回归方程就是最终确定的方程。每步检验中的界值 $F_a(1, n-p-1)$ 与自变量数目 p 有关, 在用软件计算时, 我们实际使用的是显著性 P 值 (或记为 Sig) 作检验。

11.2.2 后退法

后退法与前进法相反，首先用全部 m 个变量建立一个回归方程，然后在这 m 个变量中选择一个最不重要的变量，将它从方程中剔除。在第 3 章的回归系数的显著性检验中，用的就是这种思想，将回归系数检验的 F 值最小者对应的自变量剔除。设对 m 个回归系数进行 F 检验，记求得的 F 值为 $\{F_1^m, F_2^m, \dots, F_m^m\}$ ，选其最小者记为

$$F_j^m = \min \{F_1^m, F_2^m, \dots, F_m^m\}$$

给定显著性水平 α ，若 $F_j^m \leq F_{\alpha}(1, n - m - 1)$ ，则首先将 x_j 从回归方程中易除，为了方便，设 x_j 就是 x_m ，接着对剩下的 $m - 1$ 个自变量重新建立回归方程，进行回归系数的显著性检验，像上面那样计算出 F_j^{m-1} ，如果又有 $F_j^{m-1} \leq F_{\alpha}(1, n - (m - 1) - 1)$ ，则易除 x_j ，重新建立 y 关于 $m - 2$ 个自变量的回归方程，依此类推，直至回归方程中所剩余的 p 个自变量的 F 检验值均大于临界值 $F_{\alpha}(1, n - p - 1)$ ，没有可剔除的自变量为止。这时，得到的回归方程就是最终确定的方程。

11.2.3 逐步回归

逐步回归的基本思想是有进有出。具体做法是将变量一个一个引入，每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。引入一个变量或从回归方程中剔除一个变量，为逐步回归的一步，每一步都要进行 F 检验，以确保每次引入新的变量之前回归方程中只包含显著的变量。这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著自变量从回归方程中易除为止。这样就避免了前进法和后退法各自的缺陷，保证了最后所得的回归子集是最优回归子集。在逐步回归法中需要注意的一个问题是引入自变量和剔除自变量的显著性水平 α 值是不同的，要求引入自变量的显著性水平 $\alpha_{\text{进}}$ 于剔除自变量的显著性水平 $\alpha_{\text{出}}$ 否则可能产生“死循环”。也就是当 $\alpha_{\text{进}} \geq \alpha_{\text{出}}$ 时，如果某个自变量的显著性 P 值在 $\alpha_{\text{进}}$ 与 $\alpha_{\text{出}}$ 之间，那么这个自变量将被引入、易除，再引入、再易除，循环往复，以至无穷。

11.3 多重共线性

Definition 11.3.1 — 多重共线性. 多元线性回归模型有一个基本假设，就是要求设计矩阵 X 的秩 $\text{rank}(X) = p + 1$ ，即要求 X 中的列向量之间线性无关。如果存在不全为零的 $p + 1$ 个数 $c_0, c_1, c_2, \dots, c_p$ ，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \dots + c_p x_{ip} = 0, \quad i = 1, 2, \dots, n \quad (11.3)$$

则自变量 x_1, x_2, \dots, x_p 之间存在完全多重共线性。在实际问题中，完全的多重共线性并不多见，常见的是 11.3 式近似成立的情况，即存在不全为零的 $p+1$ 个数 $c_0, c_1, c_2, \dots, c_p$ ，使得

$$c_0 + c_1 x_n + c_2 x_{i2} + \dots + c_p x_{ip} \approx 0, \quad i = 1, 2, \dots, n \quad (11.4)$$

当自变量 x_1, x_2, \dots, x_p 存在 11.4 式的关系时，称自变量 x_1, x_2, \dots, x_p 之间存在多重共线性 (multi-collinearity)，也称为复共线性。在实际经济问题的多

11.3.1 多重共线性的影响

设回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

存在完全的多重共线性，即设计矩阵 X 的列向量存在不全为零的一组数 $c_0, c_1, c_2, \dots, c_p$ ，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \dots + c_p x_{ip} = 0, \quad i = 1, 2, \dots, n$$

设计矩阵 X 的秩 $\text{rank}(X) < p + 1$, 此时 $|X'X| = 0$, 正规方程组 $X'X\hat{\beta} = X'y$ 的解不唯一, $(XX)^{-1}$ 不存在, 回归参数的最小二乘估计表达式 $\hat{\beta} = (X'X)^{-1}X'y$ 不成立。在实际问题的研究中, 经常见到的是近似共线性的情形, 即存在不全为零的一组数 $c_0, c_1, c_2, \dots, c_p$, 使得

$$c_0 + c_1x_{i1} + c_2x_{i2} + \dots + c_px_{ip} \approx 0, \quad i = 1, 2, \dots, n$$

此时设计矩阵 X 的秩 $\text{rank}(X) = p + 1$ 虽然成立, 但是此时 $|X'X| \approx 0$, $(X'X)^{-1}$ 的对角线元素很大, $\hat{\beta}$ 的方差阵 $D(\hat{\beta}) = \sigma^2(X'X)^{-1}$ 的对角线元素很大, 而 $D(\hat{\beta})$ 的对角线元素即 $\text{var}(\hat{\beta}_0), \text{var}(\hat{\beta}_1), \dots, \text{var}(\hat{\beta}_p)$, 因而 $\beta_0, \beta_1, \dots, \beta_p$ 的估计精度很低。这样, 虽然用普通最小二乘估计能得到 β 的无偏估计, 但估计量 $\hat{\beta}$ 的方差很大, 不能正确判断解释变量对被解释变量的影响程度, 甚至导致估计量的经济意义无法解释。当完全相关时, $r = 1$, 方差将无穷大。

11.3.2 多重共线性的诊断

二、方差扩大因子法对自变量作中心标准化, 则 $X^*X^* = (r_{ij})$ 为自变量的相关阵。记

$$C = (c_{ij}) = (X^*X^*)^{-1}$$

称其主对角线元素 $VIF_j = c_{jj}$ 为自变量 x_j 的方差扩大因子 (variance inflation factor, VIF)。根据 $D(\hat{\beta}) = \sigma^2(X'X)^{-1}$ 式可知

$$\text{var}(\hat{\beta}_j) = c_{jj}\sigma^2/L_{jj}, \quad j = 1, 2, \dots, p \quad (11.5)$$

其中, L_{jj} 为 x_j 的离差平方和, 由 11.5 式可知, 用 c_{jj} 作为衡量自变量 x_j 的方差扩大程度的因子是恰如其分的。记 R_j^2 为自变量 x_j 对其余 $p - 1$ 个自变量的复决定系数, 可以证明

$$c_{jj} = \frac{1}{1 - R_j^2}$$

可以知道 $VIF_j \geq 1$

R_j^2 度量了自变量 x_j 与其余 $p - 1$ 个自变量的线性相关程度, 这种相关程度越强, 说明自变量之间的多重共线性越严重, R_j^2 就越接近 1, VIF 就越大。反之, x_j 与其余 $p - 1$ 个自变量的线性相关程度越弱, 自变量间的多重共线性就越弱, R_j^2 就越接近零, VIF 就越接近 1。由此可见, VIF 的大小反映了自变量之间是否存在多重共线性, 因此可由它来度量多重共线性的严重程度。经验表明, 当 $VIF_j \geq 10$ 时, 就说明自变量 x_j 与其余自变量之间有严重的多重共线性, 且这种多重共线性可能会过度地影响最小二乘估计值。也可以用 p 个自变量所对应的方差扩大因子的平均数来度量多重共线性。当

$$\overline{VIF} = \frac{1}{p} \sum_{j=1}^p VIF_j$$

远远大于 1 时, 就表示存在严重的多重共线性问题。对于只含两个解释变量 x_1 和 x_2 的回归方程, 判断它们是否存在多重共线性实际上就是计算 x_1 和 x_2 的样本决定系数 R_{12}^2 , 如果 R_{12}^2 很大, 则认为 x_1 与 x_2 可能存在严重的多重共线性。因为 R^2 和样本量 n 有关, 当样本量较小时, R^2 容易接近 1, 就像我们曾说的, $n = 2$ 时, 两点总能连成一条直线, $R^2 = 1$ 。所以我们认为当样本量还不算小, 而 R^2 接近 1 时, 可以肯定存在严重的多重共线性。当某自变量 x_j 对其余 $p - 1$ 个自变量的复决定系数 R_j^2 超过一定界限时, SPSS 软件将拒绝这个自变量 x_j 进入回归模型。称 $Tol_j = 1 - R_j^2$ 为自变量 x_j 的容忍度 (tolerance), SPSS 软件的默认容忍度为 0.0001。也就是说, 当 $R_j^2 > 0.9999$ 时, 自变量 x_j 将被自动拒绝在回归方程之外, 除非我们修改容忍度的默认值。

特征根分析

根据矩阵行列式的性质，矩阵的行列式等于其特征根的连乘积。因而，当行列式 $|X'X| \approx 0$ 时，矩阵 $X'X$ 至少有一个特征根近似为零。反之可以证明，当矩阵 $X'X$ 至少有一个特征根近似为零时， X 的列向量间必存在多重共线性，证明如下：记 $X = (X_0, X_1, \dots, X_p)$ ，其中 $X_i (i = 0, 1, \dots, p)$ 为 X 的列向量， $X_0 = (1, 1, \dots, 1)$ 是元素全为 1 的 n 维列向量。 λ 是矩阵 $X'X$ 的一个近似为零的特征根， $\lambda \approx 0$ ， $c = (c_0, c_1, \dots, c_p)'$ 是对应于特征根 λ 的单位特征向量，则

$$X'Xc = \lambda c \approx 0$$

上式两边左乘 c ，得

$$c'X'Xc \approx 0$$

从而有 $Xc \approx 0$ 即

$$c_0X_0 + c_1X_1 + \dots + c_pX_p \approx 0$$

写成分量形式即

$$c_0 + c_1x_{i1} + c_2x_{i2} + \dots + c_px_{ip} \approx 0, \quad i = 1, 2, \dots, n$$

这正是定义的多重共线性关系。

如果矩阵 $X'X$ 有多个特征根近似为零，在上面的证明中，取每个特征根的特征向量为标准化正交向量，即可证明： $X'X$ 有多少个特征根接近零，设计矩阵 X 就有多少个多重共线性关系，并且这些多重共线性关系的系数向量就等于接近零的那些特征根对应的特征向量。

特征根近似为零的标准如何确定呢？可以用下面介绍的条件数确定。记 $X'X$ 的最大特征根为 λ_m ，我们称

$$k_i = \sqrt{\frac{\lambda_m}{\lambda_i}}, \quad i = 0, 1, 2, \dots, p$$

为特征根 λ_i 的条件数 (condition index)。在其他一些书中，条件数定义为 $k_i = \lambda_m/\lambda_i$ ，没有开平方根，SPSS 软件是采用的 (6.10) 式开平方根的，这一点请读者注意。条件数度量了矩阵 $X'X$ 的特征根的散布程度，可以用它来判断多重共线性是否存在以及多重共线性的严重程度。通常认为 $0 < k < 10$ 时，设计矩阵 X 没有多重共线性； $10 < k < 100$ 时，存在较强的多重共线性； $k \geq 100$ 时，存在严重的多重共线性。方差比例 (Variance Proportions)：如果某几个自变量的方差比例值在某一行同时较大（接近 1），则这几个自变量间就存在多重共线性。

多重共线性的其他情况

1. 当增加或易除一个自变量，或者改变一个观测值时，回归系数的估计值发生较大变化，就认为回归方程存在严重的多重共线性。
2. 从定性分析角度来看，当一些重要的自变量在回归方程中没有通过显著性检验时，可初步判断存在严重的多重共线性。
3. 当有些自变量的回归系数所带正负号与定性分析结果违背时，认为存在多重共线性。
4. 自变量的相关矩阵中，当自变量间的相关系数较大时，认为可能存在多重共线性。
5. 当一些重要的自变量的回归系数的标准误差较大时，认为可能存在多重共线性。

11.3.3 消除多重共线性的方法

1. 解释变量的 VIF 如果大于 10, 可以考虑剔除
2. 增加样本量, L_{11} 会增加, 从而

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{(1 - r_{12}^2)L_{11}}$$

减少。在实践中, 当我们所选的变量个数接近样本量 n 时, 自变量间就容易产生共线性。所以在运用回归分析研究经济问题时, 要尽可能使样本量 n 远大于自变量个数 p_0

12. 岭回归

多元线性回归模型的矩阵形式为

$$y = X\beta + \varepsilon$$

参数 β 的普通最小二乘估计为

$$\hat{\beta} = (X'X)^{-1}X'y$$

当自变量 x_j 与其余自变量间存在多重共线性时，

$$\text{var}(\hat{\beta}_j) = c_{jj}\sigma^2/L_{jj}$$

很大， $\hat{\beta}_j$ 就很不稳定，在具体取值上与真值有较大的偏差。

Definition 12.0.1 — 岭回归. 岭回归 (ridge regression, RR)：当自变量间存在多重共线性， $|X'X| \approx 0$ 时，我们设想给 $X'X$ 加上一个正常数量阵 $k(k > 0)$ ，那么 $X'X + kI$ 接近奇异的程度就会比 $X'X$ 接近奇异的程度小得多。考虑到变量的量纲问题，先对数据作标准化，为了计算方便，标准化后的设计阵仍然用 X 表示，定义为

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'y \quad (12.1)$$

我们称12.1式为 β 的岭回归估计，其中， k 称为岭参数。由于假设 X 已经标准化，所以 $X'X$ 就是自变量样本相关阵。12.1式中 y 可以标准化也可以不标准化，如果 y 也经过标准化，那么12.1式计算的实际是标准化岭回归估计。 $\hat{\beta}(k)$ 作为 β 的估计应比最小二乘估计 $\hat{\beta}$ 稳定，当 $k = 0$ 时的岭回归估计 $\hat{\beta}(0)$ 就是普通最小二乘估计。因为岭参数 k 不是唯一确定的，所以得到的岭回归估计 $\hat{\beta}(k)$ 实际是回归参数 β 的一个估计族。

12.0.1 岭回归估计的性质

在本节关于岭回归估计的性质的讨论中，假定12.1式中因变量观测向量 y 未经标准化。

Proposition 12.0.1 $\hat{\beta}(k)$ 是回归参数 β 的有偏估计

Proof.

$$\begin{aligned} E[\hat{\beta}(k)] &= E\left((X'X + kI)^{-1}X'y\right) \\ &= (X'X + kI)^{-1}X'E(y) \\ &= (X'X + kI)^{-1}X'X\beta \end{aligned}$$

显然只有当 $k = 0$ 时, $E[\hat{\beta}(0)] = \beta$; 当 $k \neq 0$ 时, $\hat{\beta}(k)$ 是 β 的有偏估计。要特别强调的是 $\hat{\beta}(k)$ 不再是 β 的无偏估计, 有偏性是岭回归估计的一个重要特性。 ■

Proposition 12.0.2 在认为岭参数 k 是与 y 无关的常数时, $\hat{\beta}(k) = (X'X + kI)^{-1}X'y$ 是最小二乘估计 $\hat{\beta}$ 的一个线性变换, 也是 y 的线性函数。

$$\begin{aligned} \hat{\beta}(k) &= (X'X + kI)^{-1}X'y \\ &= (X'X + kI)^{-1}X'X(X'X)^{-1}X'y \\ &= (X'X + kI)^{-1}X'X\hat{\beta} \end{aligned}$$

所以, 岭估计 $\hat{\beta}(k)$ 是最小二乘估计 $\hat{\beta}$ 的一个线性变换, 根据定义式 $\hat{\beta}(k) = (X'X + kI)^{-1}X'y$ 知 $\hat{\beta}(k)$ 也是 y 的线性函数。这里需要注意的是, 在实际应用中, 由于岭参数 k 总是要通过数据来确定, 因而 k 也依赖于 y , 因此从本质上说, $\hat{\beta}(k)$ 并非 $\hat{\beta}$ 的线性变换, 也不是 y 的线性函数。

性质 3 对任意 $k > 0$, $\|\hat{\beta}\| \neq 0$, 总有

$$\|\hat{\beta}(k)\| < \|\hat{\beta}\|$$

这里 $\|\cdot\|$ 是向量的模, 等于向量各分量的平方和。这个性质表明 $\hat{\beta}(k)$ 可看成由 $\hat{\beta}$ 进行某种向原点的压缩。从 $\hat{\beta}(k)$ 的表达式可以看到, 当 $k \rightarrow \infty$ 时, $\hat{\beta}(k) \rightarrow 0$ 即 $\hat{\beta}(k)$ 化为零向量。

性质 4 以 MSE 表示估计向量的均方误差, 则存在 $k > 0$, 使得 $\text{MSE}[\hat{\beta}(k)] < \text{MSE}(\hat{\beta})$ 即

$$\sum_{j=1}^p E\left[\hat{\beta}_j(k) - \beta_j\right]^2 < \sum_{j=1}^p D(\hat{\beta}_j)$$

当岭参数 k 在 $(0, \infty)$ 内变化时, $\hat{\beta}_j(k)$ 是 k 的函数, 在平面坐标系上把函数 $\hat{\beta}_j(k)$ 描画出来, 画出的曲线称为岭迹。在岭回归中, 岭迹分析可用来了解各自变量的作用及自变量间的相互关系。

1. 在图 12.1 (a) 中, $\hat{\beta}_j(0) = \hat{\beta}_j > 0$, 且比较大。从古典回归分析的观点看, 应将 x_j 看作对 y 有重要影响的因素。但 $\hat{\beta}_j(k)$ 的图形显示出相当的不稳定当 k 从零开始略增加时, $\hat{\beta}_j(k)$ 显著地下降, 而且迅速趋于零, 因而失去预测能力。从岭回归的观点看, x_j 对 y 不起重要作用, 甚至可以剔除这个变量。
2. 与图 7.2 (a) 相反的情况见图 7.2 (b), $\hat{\beta}_j = \hat{\beta}_j(0) > 0$, 但很接近 0. 从古典回归分析的观点看, x_j 对 y 的作用不大。但随着 k 略增加, $\hat{\beta}_j(k)$ 骤然变为负值, 从岭回归的观点看, x_j 对 y 有显著影响。
3. 在图 7.2 (c) 中, $\hat{\beta}_j = \hat{\beta}_j(0) > 0$, 说明 x_j 比较显著, 但当 k 增加时, $\hat{\beta}_j(k)$ 迅速下降, 且稳定为负值。从古典回归分析的观点看, x_j 是对 y 有正影响的显著因素。从岭回归的观点看, x_j 被看做对 y 有负影响的因素。
4. (4) 在图 7.2 (d) 中, $\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k)$ 都很不稳定, 但其和却大体上稳定。这种情况往往发生在自变量 x_1 和 x_2 的相关性很大的场合, 即在 x_1 和 x_2 之间存在多重共线性。因此, 从变量选择的观点看, 两者只要保存一个就够了。这可用来解释某些回归系数估计的符号不合理的情形, 从实际观点看, β_1 和 β_2 不应有相反的符号。岭回归分析

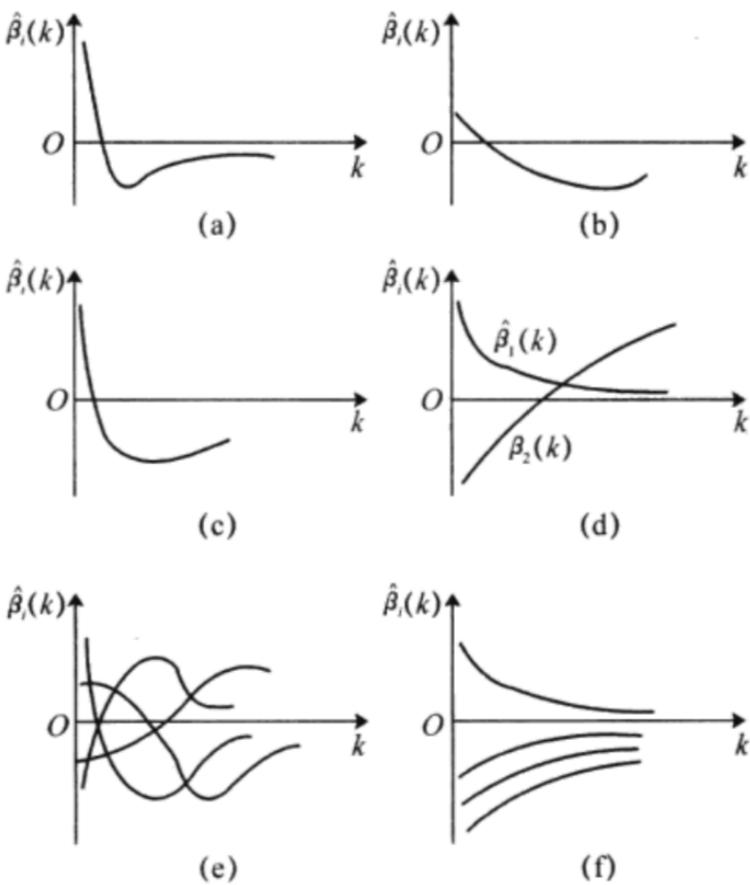


Figure 12.1: RidgeRegress

的结果对这一点提供了一种解释。(5)从全局看，岭迹分析可用来估计在某一具体实例中最小二乘估计是否适用。把所有回归系数的岭迹都描在一张图上，如果这些岭迹线的不稳定性很大，整个系统呈现比较“乱”的局面，往往就使人怀疑最小二乘估计是否很好地反映了真实情况，如图 7.2 (e) 所示。如果情况如图 7.2 (f) 那样，则我们必须对最小二乘估计可以有更大的信心。当情况介于 (e) 和 (f) 之间时，我们必须适当地选择 k 值。

12.0.2 岭参数的选择

我们的目的是要选择使 $MSE(\hat{\beta}(k))$ 达到最小的 k ，最优 k 值依赖于未知参数 β 和 σ^2 ，因而在实际应用中必须通过样本来确定。

岭迹法

改善，岭参数 k 值的选择就显得尤为重要。选择 k 值的一般原则是：(1) 各回归系数的岭估计基本稳定。(2) 用最小二乘估计时符号不合理的回归系数，其岭估计的符号变得合理。(3) 回归系数没有不合乎经济意义的绝对值。(4) 残差平方和增加不太多。

二、方差扩大因子法 方差扩大因子 c_j 可以度量多重共线性的严重程度，一般当 $c_{jj} > 10$

时，模型就有严重的多重共线性。计算岭估计 $\hat{\beta}(k)$ 的协方差阵，得

$$\begin{aligned} D(\hat{\beta}(k)) &= \text{cov}(\hat{\beta}(k), \hat{\beta}(k)) \\ &= \text{cov}\left((X'X + kI)^{-1}X'y, (X'X + kI)^{-1}X'y\right) \\ &= (X'X + kI)^{-1}X'\text{cov}(y, y)X(X'X + kI)^{-1} \\ &= \sigma^2 (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \\ &= \sigma^2 c(k) \end{aligned}$$

其中，矩阵 $c(k) = (X'X + kI)^{-1}X'X(X'X + kI)^{-1}$ ，其对角元素 $c_{jj}(k)$ 为岭估计的方差扩大因子。不难看出， $c_{jj}(k)$ 随着 k 的增大而减少。应用方差扩大因子选择 k 的经验做法是：选择 k 使所有方差扩大因子 $c_{jj}(k) \leq 10$ 。当 $c_{ij}(k) \leq 10$ 时，所对应的 k 值的岭估计 $\hat{\beta}(k)$ 就会相对稳定。

岭回归的一个重要应用是选择变量，选择变量通常的原则是：(1) 在岭回片的计算中，假定设计矩阵 X 已经中心化和标准化，这样可以直接比较标准化岭回归系数的大小。我们可以剔除掉标准化岭回归系数比较稳定且绝对值很小的自变量。(2) 当 k 值较小时，标准化岭回归系数的绝对值并不很小，但是不稳定，随着 k 的增加迅速趋于零。像这样岭回归系数不稳定、振动趋于零的自变量，我们也可以予以剔除。(3) 剔除标准化岭回归系数很不稳定的自变量。如果有若干个岭回归系数不稳定，究竟剔除几个，剔除哪几个，并无一般原则可循，需根据易除某个变量后重新进行岭回归分析的效果来确定。

13. 主成分回归和偏最小二乘

13.1 主成分回归

Definition 13.1.1 主成分分析 (principal components analysis, PCA) 也称主分量分析, 是用一种降维的思想, 在损失很少信息的前提下把多个指标利用正交旋转变换转化为几个综合指标的多元统计分析方法。通常把转化生成的综合指标称为主成分, 其中每个主成分都是原始变量的线性组合, 且各个主成分之间互不相关。设对某一事物的研究涉及 p 个指标, 分别用 X_1, X_2, \dots, X_p 表示, 这 p 个指标构成的 p 维随机向量为 $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ 。设随机向量 \mathbf{X} 的均值为 $\boldsymbol{\mu}$, 协方差矩阵为 Σ 对 \mathbf{X} 进行线性变换, 可以形成新的综合变量, 用 \mathbf{Y} 表示, 新的综合变量可以由原来的变量线性表示, 即满足下式:

$$\begin{cases} Y_1 = \mu_{11}X_1 + \mu_{12}X_2 + \dots + \mu_{1p}X_p \\ Y_2 = \mu_{21}X_1 + \mu_{22}X_2 + \dots + \mu_{2p}X_p \\ \dots \\ Y_p = \mu_{p1}X_1 + \mu_{p2}X_2 + \dots + \mu_{pp}X_p \end{cases}$$

由于可以任意地对原始变量进行上述线性变换, 得到的综合变量 \mathbf{Y} 的统计特性也不尽相同, 因此为了取得较好的效果, 我们总是希望 $Y_i = \boldsymbol{\mu}'_i \mathbf{X}$ 的方差尽可能大且各 Y_i 之间互相独立, 由于

$$\text{var}(Y_i) = \text{var}(\boldsymbol{\mu}'_i \mathbf{X}) = \boldsymbol{\mu}'_i \Sigma \boldsymbol{\mu}_i$$

而对于任意常数 c , 有

$$\text{var}(c\boldsymbol{\mu}'_i \mathbf{X}) = c\boldsymbol{\mu}'_i \Sigma \boldsymbol{\mu}_i c = c^2 \boldsymbol{\mu}'_i \Sigma \boldsymbol{\mu}_i$$

因此, 对 $\boldsymbol{\mu}_i$ 不加限制时, 可使 $\text{var}(Y_i)$ 任意增大, 问题将变得没有意义。我们将线性变换约束在下面的原则之下: (1) $\boldsymbol{\mu}'_i \boldsymbol{\mu}_i = 1$, 即 $\mu_{i1}^2 + \mu_{i2}^2 + \dots + \mu_{ip}^2 = 1 (i = 1, 2, \dots, p)$ (2) Y_i 与 Y_j 不相关 ($i \neq j; i, j = 1, 2, \dots, p$) (3) Y_1 是 X_1, X_2, \dots, X_p 的所有满足原则 1 的线性组合中方差最大者: Y_2 是与 Y_1 不相关的 X_1, X_2, \dots, X_p 的所有线性组合中方差最大者

;……; Y_p 是与 Y_1, Y_2, \dots, Y_{p-1} 都不相关的 X_1, X_2, \dots, X_p 的所有线性组合中方差最大者。基于以上三条原则决定的综合变量 Y_1, Y_2, \dots, Y_p , 分别称为原始变量的第一、第二……第 p 个主成分。其中, 各综合变量在总方差中占的比重依次递减。在实际研究工作中, 通常只挑前几个方差最大的主成分, 从而达到简化系统结构、抓住问题实质的目的。

引论: 设矩阵 $A' = A$, 将 A 的特征值 $\lambda_1, \lambda_2, \dots, \lambda_p$ 依大小顺序排列, 不妨设 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$, $\gamma_1, \gamma_2, \dots, \gamma_p$ 为矩阵 A 各特征值对应的标准正交向量, 则对任意向量 x , 有

$$\max_{x \neq 0} \frac{x'Ax}{x'x} = \lambda_1, \dots, \min_{x \neq 0} \frac{x'Ax}{x'x} = \lambda_p$$

结论: 设随机向量 $X = (X_1, X_2, \dots, X_p)'$ 的协方差矩阵为 ■, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 为 Σ 的特征值, $\gamma_1, \gamma_2, \dots, \gamma_p$ 为矩阵 Σ 各特征值对应的标准正交向量, 则第 i 个主成分为

$$Y_i = \gamma_1 i X_1 + \gamma_2 i X_2 + \dots + \gamma_p i X_p, i = 1, 2, \dots, p$$

此时,

$$\begin{aligned}\text{var}(Y_i) &= \gamma_i^2 \Sigma \gamma_i = \lambda_i \\ \text{cov}(Y_i, Y_j) &= \gamma_i^2 \Sigma \gamma_j = 0, i \neq j\end{aligned}$$

由以上结论, 我们把 X_1, X_2, \dots, X_p 的协方差阵 Σ 的非零特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ 对应的标准化特征向量 $\gamma_1, \gamma_2, \dots, \gamma_p$ 分别作为系数向量, $Y_1 = \gamma_1^T X, Y_2 = \gamma_2^T X, \dots, Y_p = \gamma_p^T X$ 分别称为随机向量 X 的第一主成分、第二主成分……第 p 主成分。

Proposition 13.1.1 — 主成分的性质. 1. Y 的协方差阵为对角矩阵 Λ 。其中对角线上的值为 $\lambda_1, \lambda_2, \dots, \lambda_p$

2. 记 $\Sigma = (\sigma_{ij})_{p \times p}$, 有 $\sum_{i=1}^p \lambda_i = \sum_{i=1}^p \sigma_{ii}$ 称 $\alpha_k = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}$ ($k = 1, 2, \dots, p$) 为第 k 个主成分 Y_k 的方差贡献率称 $\frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i}$ 为主成分 Y_1, Y_2, \dots, Y_m 的累积贡献率。 $\sum_{i=1}^p \lambda_i$
3. $\rho(Y_k, X_i) = \mu_{ki} \sqrt{\lambda_k} / \sqrt{\sigma_{ii}}$ ($k, i = 1, 2, \dots, p$) 其中, 第 k 个主成分 Y_k 与原始变量 X_i 的相关系数 $\rho(Y_k, X_i)$ 称为因子负荷量。因子负荷量是主成分解释中非常重要的解释依据, 因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因。
4. $\sum_{i=1}^p \rho^2(Y_k, X_i) \sigma_{ii} = \lambda_k$
5. $\sum_{i=1}^p \rho^2(Y_k, X_i) = \frac{1}{\sigma_{ii}} \sum_{k=1}^p \lambda_k \mu_{kk}^2 = 1$

X_i 与前 m 个主成分 Y_1, Y_2, \dots, Y_m 的全相关系数平方和称为 Y_1, Y_2, \dots, Y_m 对原始变量 X_i 的方差贡献率 v_i , 即 $v_i = \frac{1}{\sigma_{ii}} \sum_{k=1}^m \lambda_k \mu_{kk}^2$ ($i = 1, 2, \dots, p$)。这一定义说明前 m 个主成分提取了原始变量 X_i 中 v_i 的信息, 由此可以判断提取的主成分说明原始变量的能力。

13.2 偏最小二乘

n 为数据个数, k 为变量数。当 $k > n$ 时, $\mathbf{X}'\mathbf{X}$ 是一个奇异矩阵, 无法求逆。主成分回归 (PCR) 就不求 $\mathbf{X}'\mathbf{X}$ 的逆, 而直接求 $\mathbf{X}'\mathbf{X}$ 的特征根。把它的非零的特征根记为 λ_i , 如果有 r 个, r 就是 $\mathbf{X}'\mathbf{X}$ 的秩, 将它们按大小顺序排出, 得 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$, 相应的特征向量分别记为 $\alpha_1, \alpha_2, \dots, \alpha_r$, 它们均为 $k \times 1$ 向量, 令 α_i 的分量为 α_{ij} , 即 $\alpha'_i = (\alpha_{i1}, \dots, \alpha_{ik})$, 又令

$$z_i = \alpha'_i X = \alpha_{i1} x_1 + \alpha_{i2} x_2 + \dots + \alpha_{ik} x_k, \quad i = 1, 2, \dots, r$$

则 z_1, z_2, \dots, z_r 都是 x_1, x_2, \dots, x_k 的线性函数, $r < k$, 且 $r < n$, 因此将 y 对 z_1, z_2, \dots, z_r 或 z_1, z_2, \dots, z_r 的一部分做同归就可以了, 这就是 PCR 的主要想法。

PCR 虽然解决了 $k > n$ 这一矛盾, 但它选 z_i 的方法与因变量 y 无关, 只在自变量 x_1, x_2, \dots, x_k 中去寻找有代表性的 z_1, z_2, \dots, z_r 。偏最小二乘 (partial least squares, PLS) 在

这一点上就与 PCR 不同，它寻找 x_1, x_2, \dots, x_k 的线性函数时，考虑与 y 的相关性，选择与 y 相关性较强又能方便算得的 x_1, x_2, \dots, x_k 的线性函数。它的算法是最小二乘，但是它只选 x_1, x_2, \dots, x_k 中与 y 有相关性的变量，不考虑全部 x_1, x_2, \dots, x_k 的线性函数，只考虑偏向与 y 有关的一部分所以称为偏最小二乘。

(y, x) 共观测了 n 组数据 $(y_1, x_1), \dots, (y_n, x_n)$ ，于是 x, y 的线性回归方程为

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, & \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}\end{aligned}$$

当 x_i, y_i 这些数据的均值为 0 时， $\hat{\beta}_0 = 0, \hat{\beta}_1$ 就有简单的形式，即有

$$\begin{cases} y = \hat{\beta} x \\ \hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{x' y}{x' x} \end{cases} \quad (13.1)$$

其中， $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ 为观测值向量。PLS 就是反复利用 13.1 式 0
首先将数据

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

中心化，中心化之后得到的 \tilde{y}_i, \tilde{x}_i 相应的各自的均值都是 0. 即 y 和 $X = (x_i)$ 满足

$$\sum_{i=1}^n y_i = 0, \quad \sum_{i=1}^n x_{ti} = 0, \quad i = 1, 2, \dots, k$$

将 y 对每个自变量 x_i 单独作回归，用 13.1 式就得

$$\hat{y}(x_i) = \frac{x'_i y}{x'_i x_i} x_i, \quad x_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{ni} \end{pmatrix}, \quad i = 1, 2, \dots, k$$

我们用 x_i 表示资料向量， x_i 表示自变量（不是数据）上式告诉我们与 y 有关的 x_i 的线性组合，应该是右端的量，将右端的量加权后，用 ω_i 记相应的权，就得到

$$\sum_{i=1}^k \omega_i \frac{x'_i y}{x'_i x_i} x_i$$

权 ω_i 可以有很多种选择，比较简单的是 $\omega_i = x'_i x_i$ ，代入上式就得 $\sum_{i=1}^k (x'_i y) x_i$ ，可见这个 x_i 的线性组合是应入选的变量。令

$$t_1 = \sum_{i=1}^k (x'_i y) x_i \quad (13.2)$$

它相应的 n 个资料是

$$t_1 = \sum_{i=1}^k (x'_i y) x_i$$

容易看出, 13.2 式的 t_1 中, 它的系数与 y 有关, 而不像 PCR 与 y 无关。将 t_1 作为自变量, 对 y 求回归, 用13.1 式就有

$$\hat{y}(t_1) = \frac{t_1'y}{t_1't_1} t_1$$

利用上式预测 y , 得预测值向量 $\hat{y}(t_1)$, 即有

$$\hat{y}(t_1) = \frac{t_1'y}{t_1't_1} t_1$$

于是得残差 $y^{(1)} = y - \hat{y}(t_1)$ 。考虑到残差 $y^{(1)}$ 中不再含 t_1 的信息, 因此各个自变量 x_i 的作用对 y 而言, 含 t_1 的部分已不具新的信息, 都应删去。也就是将每个自变量 x_i 对 t_1 求回归, 得回归方程 (还是用13.1 式) 和预测值

$$\hat{x}_i(t_1) = \frac{t_1'x_i}{t_1't_1} t_1, \quad i = 1, 2, \dots, k$$

x_i 相应的残差, $x_i^{(1)} = x_i - \hat{x}_i(t_1)$ ($i = 1, 2, \dots, k$), 于是将 $y^{(1)}, x_1^{(1)}, \dots, x_k^{(1)}$ 作为新的原始资料, 重复上述步骤, 逐步求得 t_1, t_2, \dots, t_r, r 是 $\mathbf{X}'\mathbf{X}$ 的秩。最后利用 y 对 t_2, t_3, \dots, t_r 用普通最小二乘做回归, 经过变量间的转换, 最终可得到 y 对 x_1, x_2, \dots, x_k 的回归方程, 这种求得回归方程的方法就称为 PLS 法, 即偏最小二乘法。

Wold 算法 (1) $y \rightarrow y_0, X \rightarrow X_0, 0 \rightarrow \hat{y}_0, 0 \rightarrow \hat{X}_0$ (2) 对 $a = 1$ 到 r 做: (3) $t_a = X_{a-1}X_{a-1}'y_{a-1}$ (4) $\hat{y}_a = \frac{t_a t_a'}{t_a t_a} y_{a-1} + \hat{y}_{a-1}$ (5) $y_a = y_{a-1} - \frac{t_a t_a'}{t_a t_a} y_{a-1}$ (6) $\hat{\mathbf{X}}_a = \frac{t_a t_a'}{t_a t_a} X_{a-1}$ (7) $X_a = X_{a-1} - \hat{\mathbf{X}}_a$ (8) $X_a'X_a$ 中主对角元素近似 0, 就退出

一个更为简单的算法。这个证明利用了回归方程是观测向量 y 在自变量资料向量所张成的子空间中的投影, 所以逐次求出 t_1, t_2, \dots, t_r 的投投影矩阵是

$$P_{t_i} = t_i(t_i't_i)^{-1}t_i' = \frac{t_i t_i'}{t_i't_i}$$

$$P_{t_i}y = \frac{t_i'y}{t_i't_i} t_i$$

我们用 $(t_1), (t_1, t_2), \dots, (t_1, t_2, \dots, t_r)$ 分别表示由 t_1 张成的子空间, t_1, t_2 张成的子空间, 等等, \hat{y}_a 就是在 (t_1, t_2, \dots, t_a) 上的 y 的投影。如引人记号

$$S = X'X, \quad s = X'y, \quad s_1 = s, \quad s_r = S^{k-1}s, \quad k = 1, 2, \dots, r$$

Helland 证明了

$$(t_1, t_2, \dots, t_a) = (X_{s_1}, X_{s_2}, \dots, X_{s_a})$$

对 $a = 1, 2, \dots, r$ 都成立。于是 PLS 算法可改为: Helland 算法 (1) $S = X'X, s = X'y$ (2) 对 $a = 1$ 到 r 做: (3) $S_a = S^{a-1}s$ (4) y 对 $X_{s_1}, X_{s_2}, \dots, X_{s_a}$ 做普通最小二乘回归得 \hat{y}_a (5) 选择合适的 \hat{y}_a 上述算法中都存在一个问题, 就是这个算法何时结束, 什么是合适的 a , 是否一定要算到某个 X_a 中的一列是 0 为止?

运用 PLS 的情况下, 大部分都使用交叉验证 (cross-validation) 法: 现在从资料 Xy 中删去第 l 组资料, 即删去 (y, x_n, \dots, x_k) , 删去后的 X, y 用 $X(-l), y(-l)$ 表示; 用 $X(-l), y(-l)$ 作为原始资料, 用 PLS 方法算出预测方程中 \hat{y}_a 的表达式, 然后用 $\hat{y}_a(-l)$ 表示这个预测方程的预测值, 将 x_{l1}, \dots, x_{lk} 代入 $\hat{y}_a(-l)$, 得它的预测值为 $\hat{y}_{al}(-l)$, 残差 $\hat{y}_l - \hat{y}_{al}(-l)$ 就反映了第 a 步预测方程的好坏在第 1 组资料上的体现, 于是

$$\sum_{l=1}^n (\hat{y}_l - \hat{y}_{al}(-l))^2$$

	普通最小二乘法、岭回归、变量选择	主成分回归、偏最小二乘法
各种回归方法的假设条件	自变量是独立的 自变量的值必须是精确的 残差必须是随机的	自变量可以是相关的 自变量的值可以有误差 残差可以有一定的结构

就在整体上反映了第 a 步预测方程的好坏，把这个值记为损失 $L(a)$ ，自然应该选 a 使 $L(a)$ 达到最小，即应该选 a_* 使

$$L(a_*) = \min_{1 \leq a \leq r} L(a)$$

所以 $L(a)$ 的计算没有必要添加新的程序、实际上重复使用就行了，当 n 不大时，更为方便。正因为使用了这个交叉验证方法，选出的预测方程往往效果比较好。

14. 非线性

14.1 非线性

Theorem 14.1.1

英文名称	中文名称	方程形式
Linear	线性函数	$y = b_0 + b_1 t$
Logarithm	对数函数	$y = b_0 + b_1 \ln t$
Inverse	逆函数	$y = b_0 + b_1 / t$
Quadratic	二次曲线	$y = b_0 + b_1 t + b_2 t^2$
Cubic	三次曲线	$y = b_0 + b_1 t + b_2 t^2 + b_3 t^3$
Power	幂函数	$y = b_0 t^{b_1}$
Compound	复合函数	$y = b_0 b_1^t$
S	S形函数	$y = \exp(b_0 + b_1 / t)$
Logistic	逻辑函数	$y = \frac{1}{\frac{1}{u} + b_0 b_1^t}$ u 是预先给定的常数
Growth	增长曲线	$y = \exp(b_0 + b_1 t)$
Exponent	指数函数	$y = b_0 \exp(b_1 t)$

Definition 14.1.1 双曲函数

$$y = \frac{x}{ax + b}$$

或等价地表示为

$$\frac{1}{y} = a + b \frac{1}{x}$$

14.2 Logistic 回归

处理 0-1 型

$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Definition 14.2.1 Logistic 回归方程为

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \quad i = 1, 2, \dots, c$$

其中，c 为分组数据的组数，将以上回归方程作线性化变换，令

$$p'_i = \ln\left(\frac{p_i}{1-p_i}\right)$$

变换称为逻辑 (logit) 变换，变换后的线性回归模型为

$$p'_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

是一个普通的一元线性回归模型。拟合了因变量为定性变量的回归模型，但是异方差性没有结果，应该使用加权最小二乘

Theorem 14.2.1 — 未分组数据的 Logistic 回归模型. 设 y 是 0-1 型变量， x_1, x_2, \dots, x_p 是与 y 相关的确定性变量， n 组观测数据为 $(x_{i1}, x_{i2}, \dots, x_{ip}; y_i)$ ($i = 1, 2, \dots, n$)，其中， y_1, y_2, \dots, y_n 是取值 0 或 1 的随机变量， y_i 与 $x_{i1}, x_{i2}, \dots, x_{ip}$ 的关系如下

$$E(y_i) = \pi_i = f(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})$$

其中，函数 $f(x)$ 是值域在 [0, 1] 区间内的单调增函数。对于 Logistic 回归

$$f(x) = \frac{e^x}{1+e^x}$$

于是 y_i 是均值为 $\pi_i = f(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})$ 的 0-1 型分布，概率函数为

$$\begin{aligned} P(y_i = 1) &= \pi_i \\ P(y_i = 0) &= 1 - \pi_i \end{aligned}$$

可以把 y_i 的概率函数合写为

$$P(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad y_i = 0, 1; i = 1, 2, \dots, n$$

于是， y_1, y_2, \dots, y_n 的似然函数为

$$L = \prod_{i=1}^n P(y_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

对似然函数取自然对数，得

$$\begin{aligned}\ln L &= \sum_{i=1}^n [y_i \ln \pi_i + (1 - y_i) \ln (1 - \pi_i)] \\ &= \sum_{i=1}^n \left[y_i \ln \frac{\pi_i}{(1 - \pi_i)} + \ln (1 - \pi_i) \right]\end{aligned}$$

对于 Logistic 回归，将

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$$

代入得

$$\ln L = \sum_{i=1}^n [y_i (\beta_0 + \beta_1 x_n + \cdots + \beta_p x_{ip})] \quad (14.1)$$

$$- \ln (1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})) \quad (14.2)$$

最大似然估计就是选取 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 使 14.1 式达到极大。

Definition 14.2.2 Probit 回归称为单位概率回归，与 Logistic 同归相似，也是拟合 0—1 型因变量回归的方法，其回归函数是

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

用样本比例 p_i 代替概率 π_i ，表示为样本回归模型

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

15. 多元正态及参数估计

Definition 15.0.1 把 p 个随机变量放在一起得 $X = (X_1, X_2, \dots, X_p)'$ 为一个 p 维随机向量, 如果同时对 p 个变量作一次观测, 得观测值: $(x_{11}, x_{12}, \dots, x_{1p}) = X'_{(1)}$, 它是一个样品. 观测 n 次得 n 个样品: $X'_{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip}) (i = 1, 2, \dots, n)$, 而 n 个样品就构成一个样本。常把 n 个样品排成一个 $n \times p$ 矩阵, 称为样本数据阵(或样本资料阵), 记并

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} == \begin{bmatrix} X'_{(1)} \\ X'_{(2)} \\ \vdots \\ X'_{(n)} \end{bmatrix} == (X_1, X_2, \dots, X_p)$$

矩阵 X 的第 i 行: $X'_{(i)} = (x_{i1}, x_{i2}, \dots, x_{ip}) (i = 1, 2, \dots, n)$ 表示对第 i 个样品的观测值, 在具体观测之前, 它是一个 p 维的随机向量. 矩阵 X 的第 j 列

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \quad (j = 1, 2, \dots, p)$$

表示对第 j 个变量的 n 次观测, 在具体观测之前, 它是一个 n 维随机向量; 而样本数据阵 X 是一个随机阵。

15.0.1 多元正态的性质

Theorem 15.0.1 $\theta_1, \theta_2, \dots, \theta_n$ iid, $\theta_i \sim N(0, I_p)$, T 为正交矩阵, $\eta = T'\theta$, 则 $\eta_1, \eta_2, \dots, \eta_n$ 也 iid, $\eta_i \sim N(0, I_p)$, $\eta \sim N_n(0, I_n)$

Corollary 15.0.2 设 $\eta \sim N_n(\theta, \sigma^2 I_n)$, T 为 n 阶正交矩阵, 则 $\xi \equiv$

$$T \left(\frac{\eta - \theta}{\sigma} \right) \sim N_n(0, I_n)$$

Corollary 15.0.3 设 $\eta \sim N_n(\theta, \Sigma)$, 则存在正交矩阵 T , 使由变换 $\zeta = T'(\eta - \theta)$ 确定的随机向量 $\zeta \sim N_n(0, \Lambda)$, 其中 Λ 为对角矩

Theorem 15.0.4 设 $\eta \sim N_n(\theta, \Sigma)$, 则 η 的特征函数为

$$\varphi_\eta(t) = \exp \left\{ jt' \theta - \frac{1}{2} t' \Sigma t \right\}$$

其中 $j = \sqrt{-1}, t' = (t_1, \dots, t_n)$

Theorem 15.0.5 $\eta \sim N_n(\theta, \Sigma), A_{m \times n}, \text{rank}(A) = m, a$ 为 m 维常数列向量, $\xi = A\eta + a$, 则 m 为随机向量 $\xi \sim N_m(A\theta + a, A\Sigma A')$

Corollary 15.0.6 正态随机向量的任一子向量仍是正态随机向量。

Theorem 15.0.7 $\eta \sim N_n(\theta, \Sigma), \eta' = (\eta_1, \eta_2, \dots, \eta_n)$, 则 $\eta_1, \eta_2, \dots, \eta_n$ 相互独立的充要条件是它们两两不相关。

Theorem 15.0.8 if $\eta \sim \mathcal{N}(0, 1), A^2 = A$, then $\eta' A \eta \sim \chi^2(\text{tr}(A))$

Theorem 15.0.9 If

$$\xi \sim \mathcal{N}_n(\theta, \sigma^2 I_n), A = A', A^2 = A$$

then

$$\frac{(\xi - \theta)' A (\xi - \theta)}{\sigma^2} \sim \chi^2(\text{tr}(A))$$

Theorem 15.0.10 If

$$\eta' = (\eta_1, \eta_2, \dots, \eta_n), E(\eta) = \theta, \theta' = (\theta_1, \theta_2, \dots, \theta_n), \text{Var}(\eta) = \sigma^2 I_n, A = A'$$

then

1. $E(\eta' A \eta) = \sigma^2 \text{tr}(A) + \theta' A \theta$
2. if $\eta \sim \mathcal{N}_n(0, \sigma^2 I_n)$, then $\text{Var}(\eta' A \eta) = 2\sigma^4 \text{tr}(A)$

Theorem 15.0.11 $A_{n \times n} = A', B_{m \times n}, BA = 0, \eta \sim \mathcal{N}_n(\theta, \sigma^2 I_n)$, then $B\eta$ 和 $\eta' A \eta$ 是独立的

Theorem 15.0.12 $A = A', B = B', BA = 0, \eta \sim \mathcal{N}_n(\theta, \sigma^2 I_n)$, then $\eta' A \eta$ 和 $\eta' B \eta$ 是独立的

Theorem 15.0.13 $Q_i \sim \chi^2(r_i), i = 1, 2, r_1 > r_2$, and $Q_1 - Q_2$ and Q_2 are indep, then

$$Q_1 - Q_2 \sim \chi^2(r_1 - r_2)$$

Theorem 15.0.14 $\eta \sim \mathcal{N}_n(\theta, \Sigma)$, then $(\eta - \theta)'\Sigma^{-1}(\eta - \theta) \sim \chi^2(n)$

Theorem 15.0.15 若 $\xi = (\xi_1, \dots, \xi_n)^T$ 服从 n 元正态分布 $N(\mu, \Sigma)$, 而 C 为任意 $m \times n$ 阵, 则 $\eta = C\xi$ 服从 m 元正态分布 $N(C\mu, C\Sigma C^T)$

Theorem 15.0.16 若 ξ 服从 n 元正态分布 $N(\mu, \Sigma)$, 则存在一个正交变换 U , 使得 $\eta = U\xi$ 是一个具有独立正态分布分量的随机向量, 它的数学期望为 $U\mu$, 而它的方差分量是 Σ 的特征值.

Theorem 15.0.17 在正交变换下, 多维正态变量保持其独立、同方差性不变。

Theorem 15.0.18 若 $\xi \sim N(\mu, \Sigma)$, 其中 Σ 为 n 阶正定阵, 则

$$(\xi - \mu)^T \Sigma^{-1} (\xi - \mu) \sim \chi_n^2$$

Definition 15.0.2 — 边缘分布. 设 $X^{(1)}$ 为 r 维随机向量, $X^{(2)}$ 为 $p - r$ 维随机向量. 若 p 维随机向量 $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}$, 则 $X^{(1)}$ 的边缘分布为

$$f_1(x^{(1)}) = f_1(x_1, \dots, x_r) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_p) dx_{r+1} \cdots dx_p$$

$X^{(2)}$ 的边缘分布为

$$f_2(x^{(2)}) = f_2(x_{r+1}, \dots, x_p) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_p) dx_1 \cdots dx_r$$

■ **Example 15.1** 例 2. 1.1 设二维随机向量 $X = (X_1, X_2)'$ 的联合密度函数为

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} \left[1 + x_1 x_2 e^{-\frac{1}{2}(x_1^2 + x_2^2)} \right]$$

试求 X_1 和 X_2 关于随机向量 X 的边缘密度。

Proof. 首先可验证 $f(x_1, x_2)$ 满足联合密度函数的两条性质. 再利用边缘密度的计算公式, 有

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)} \left[1 + x_1 x_2 e^{-\frac{1}{2}(x_1^2 + x_2^2)} \right] dx_2 \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}x_1^2} \left[\int_{-\infty}^{\infty} e^{-\frac{1}{2}x_2^2} dx_2 + x_1 e^{-\frac{1}{2}x_1^2} \int_{-\infty}^{\infty} x_2 e^{-\frac{1}{2}x_2^2} dx_2 \right] \\ &= \frac{1}{2\pi} e^{-\frac{1}{2}x_1^2} [\sqrt{2\pi} + 0] = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \end{aligned}$$

类似可得出 $X_2 \sim N(0, 1)$



Definition 15.0.3 3. 随机向量 X 和 Y 的协方差阵若 X_i 和 Y_j 的协方差 $\text{Cov}(X_i, Y_j)$ 存在 ($i = 1, \dots, p; j = 1, \dots, q$), 则称

$$\begin{aligned}\text{COV}(X, Y) &= E[(X - E(X))(Y - E(Y))'] \\ &= \begin{bmatrix} \text{Cov}(X_1, Y_1) & \text{Cov}(X_1, Y_2) & \cdots & \text{Cov}(X_1, Y_q) \\ \text{Cov}(X_2, Y_1) & \text{Cov}(X_2, Y_2) & \cdots & \text{Cov}(X_2, Y_q) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X_p, Y_1) & \text{Cov}(X_p, Y_2) & \cdots & \text{Cov}(X_p, Y_q) \end{bmatrix}\end{aligned}$$

为随机向量 X 和 Y 的协方差阵. 若

$$\text{COV}(X, Y) = O \quad (\text{其中 } O \text{ 表示零矩阵})$$

则称 X 与 Y 不相关.

Definition 15.0.4 — 方差. 若 X_i 和 Y_i 的协方差 $\text{Cov}(X_i, Y_i)$ 存在 ($i, j = 1, 2, \dots, p$), 称 $R = (r_{ij})_{p \times p}$ 为 X 的相关阵, 其中

$$r_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)}\sqrt{\text{Var}(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \quad (i, j = 1, 2, \dots, p)$$

这里

$$\text{Var}(X_i) = \text{Cov}(X_i, X_i) \stackrel{\text{def}}{=} \sigma_{ii}$$

为随机变量 X_i 的方差, 而 $\sqrt{\sigma_{ii}}$ 为 X_i 的标准差 ($i = 1, 2, \dots, p$) 若记 $V^{1/2} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$ 为标准差矩阵, 则

$$\Sigma = V^{1/2} R V^{1/2} \quad \text{或} \quad R = (V^{1/2})^{-1} \Sigma (V^{1/2})^{-1}$$

Proposition 15.0.19 设 X, Y 是随机向量, A, B 是常数矩阵, 则

$$\begin{aligned}\text{E}(AX) &= A\text{E}(X) \\ \text{E}(AXB) &= A\text{E}(X)B \\ \text{D}(AX) &= A\text{D}(X)A' \\ \text{COV}(AX, BY) &= A\text{COV}(X, Y)B'\end{aligned}$$

Proposition 15.0.20 若 X, Y 相互独立, 则 $\text{COV}(X, Y) = O_{p \times q}$; 反之不一定成立.

Proposition 15.0.21 随机向量 $X = (X_1, X_2, \dots, X_p)'$ 的协方差阵 $\text{D}(X) = \Sigma$ 是对称非负定矩阵.

Proof. 因为 $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, 所以 $\Sigma = \Sigma'$. 对任给

$\alpha = (\alpha_1, \dots, \alpha_p)'$, 有

$$\begin{aligned}\alpha' \Sigma \alpha &= (\alpha_1, \dots, \alpha_p) E[(X - E(X))(X - E(X))'] \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} \\ &= E[\alpha'(X - E(X)) \cdot (X - E(X))' \alpha] \\ &= E[(\alpha'(X - E(X)))^2] \geq 0\end{aligned}$$

所以 $\Sigma \geq 0$, 即 Σ 为非负定矩阵. ■

Proposition 15.0.22 $\Sigma = L^2$, 其中 L 为非负定矩阵.

Proof. 由于 $\Sigma \geq 0$ (非负定), 利用线性代数中实对称阵的对角化定理, 存在正交矩阵 Γ , 使得

$$\begin{aligned}\Sigma &= \Gamma \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} \Gamma' \quad (\text{其中 } \lambda_i \geq 0) \\ &= \Gamma \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_p} \end{bmatrix} \Gamma' \cdot \Gamma \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_p} \end{bmatrix} \Gamma' \\ &= L^2\end{aligned}$$

其中 $L = \Gamma \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p}) \Gamma'$, 且 $L = L'$, 所以 $L \geq 0$. ■

Corollary 15.0.23 当矩阵 $\Sigma > 0$ (正定) 时, 矩阵 L 也称为 Σ 的平方根矩阵, 记为 $\Sigma^{1/2}$. 若令 $A = \Gamma \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$, 则协方差阵 Σ 还有如下分解 $\Sigma = AA'$ (A 为非退化方阵).

Definition 15.0.5 设 $U = (U_1, \dots, U_q)'$ 为随机向量, U_1, \dots, U_q 相互独立且同 $N(0, 1)$ 分布; 设 μ 为 p 维常数向量, A 为 $p \times q$ 常数矩阵则称 $X = AU + \mu$ 的分布为 p 元正态分布, 或称 X 为 p 维正态随机向量, 记为 $X \sim N_p(\mu, AA')$ 简单地说, 由 q 个相互独立的标准正态随机变量的一些线性组合所构成的随机向量的分布, 称其为多元正态分布。

Definition 15.0.6 在一元统计中, 若 $X \sim N(\mu, \sigma^2)$, 则 X 的特征函数为

$$\varphi(t) = E(e^{itX}) = \exp \left[it\mu - \frac{1}{2}t^2\sigma^2 \right]$$

将其推广到多维正态随机向量的情况有如下性质。

Proposition 15.0.24 设 $U = (U_1, \dots, U_q)'$ 为随机向量, U_1, \dots, U_q 相互独立且同 $N(0, 1)$ 分布; 令 $X = AU + \mu$, 则 X 的特征函数为

$$\Phi_X(t) = \exp \left[it'\mu - \frac{1}{2}t'AA't \right]$$

Proof.

$$\begin{aligned}\Phi_X(t) &= E(e^{it'X}) = E(e^{i^2t'(\mu+AU)}) \\ &= \exp(it'\mu) \cdot E(e^{it'AU}) \quad (\text{令 } s' = t'A = (s_1, \dots, s_q)) \\ &= \exp(it'\mu) \cdot E(e^{i(s_1U_1+\dots+s_qU_q)}) \\ &= \exp(it'\mu) \cdot \prod_{j=1}^q E(e^{is_jU_j}) \quad (\text{因 } U_1, \dots, U_q \text{ 独立}) \\ &= \exp(it'\mu) \cdot \prod_{j=1}^q \exp \left(-\frac{1}{2}s_j^2 \right) \quad (U_j \sim N(0, 1)) \\ &= \exp \left(it'\mu - \frac{1}{2}s's \right) = \exp \left(it'\mu - \frac{1}{2}t'AA't \right)\end{aligned}$$

■

Definition 15.0.7 若 p 维随机向量 X 的特征函数为

$$\Phi_X(t) = \exp \left[i t' \mu - \frac{1}{2} t' \Sigma t \right] \quad (\Sigma \geq 0)$$

则称 X 服从 p 元正态分布, 记为 $X \sim N_p(\mu, \Sigma)$

Proposition 15.0.25 设 $X \sim N_p(\mu, \Sigma)$, B 为 $s \times p$ 常数矩阵, d 为 s 维常向量, 令 $Z = BX + d$, 则

$$Z \sim N_s(B\mu + d, B\Sigma B')$$

Proposition 15.0.26 推论 设 $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-r}^r \sim N_p(\mu, \Sigma)$, 将 μ, Σ 剖分关

$$\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}_{p-r}^r, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}_{p-r}^r$$

则 $X^{(1)} \sim N_r(\mu^{(1)}, \Sigma_{11})$, $X^{(2)} \sim N_{p-r}(\mu^{(2)}, \Sigma_{22})$ 证明取 $B_1 = (I_r \ O)$ (其中 I_r 为 r 阶单位矩阵, O 为 $r \times (p-r)$ 零矩阵), r 维向量 $d_1 = 0$, 由性质 2 即得

$$X^{(1)} = B_1 X + d_1 \sim N_r(\mu^{(1)}, \Sigma_{11})$$

(R) 此推论指出, 多元正态分布的边缘分布仍为正态分布. 但反之, 若随机向量的任何边缘分布均为正态分布, 也不一定能导出该随机向量服从多元正态分布

Proposition 15.0.27 设 $X = (X_1, \dots, X_p)'$ 为 p 维随机向量, 则 X 服从 p 元正态分布 \Leftrightarrow 对任一 p 维实向量 a , $\xi = a'X$ 是一维正态随机变量.

Proof. \Rightarrow (必要性) : 若 $X \sim N_p(\mu, \Sigma)$, 对任一实向量 $a = (a_1, \dots, a_p)'$, 取 $B = a', d = 0$, 由性质 2 即得

$$\xi = a'X = \sum_{i=1}^p a_i X_i \sim N(a'\mu, a'\Sigma a)$$

(充分性) : 因对任给实向量 $t \in R^p$, $\xi = t'X \sim$ 一元正态分布, 可知 ξ 的各阶矩存在, 故 $E(X_i), \text{Cov}(X_i, X_j)$ ($i, j = 1, \dots, p$) 存在. 记 $E(X) = \mu, D(X) = \Sigma$ 对任意给定的 $t \in R^p$, $\xi = t'X \sim N(t'\mu, t'\Sigma t)$, 且 ξ 的特征函数为

$$\Phi_\xi(\theta) = E(e^{i\theta\xi}) = \exp \left[i\theta(t'\mu) - \frac{1}{2}\theta^2(t'\Sigma t) \right]$$

取 $\theta = 1$

$$\Phi_\xi(1) = E(e^{i\xi}) = E(e^{it'X}) = \Phi_X(t) = \exp \left[it'\mu - \frac{1}{2}t'\Sigma t \right]$$

$X \sim N_p(\mu, \Sigma)$

Definition 15.0.8 若 p 维随机向量 X 的任意线性组合均服从一元正态分布, 则称 X 为 p 维正态随机向量. 一元正态随机变量的密度函数是

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

这个式子又可改写为

$$f(x) = \frac{1}{(2\pi)^{1/2} |\sigma^2|^{1/2}} \exp \left[-\frac{1}{2}(x-\mu)' (\sigma^2)^{-1} (x-\mu) \right]$$

Definition 15.0.9 设 $X \sim N_p(\mu, \Sigma)$, 且 $\Sigma > 0$ (正定), 则 X 的联合密度函数为

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu) \right]$$

Example 15.2 二元正态密度函数的几何图形. 我们把具有等密度的点的轨迹称为等高线(面). 显然当 $p = 2$ 时

$$\begin{aligned} f(x_1, x_2) &= C \\ &\Leftrightarrow \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \cdot \sigma_2} \\ &\quad + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 = a^2 (a \geq 0) \end{aligned}$$

它是一族中心在 $(\mu_1, \mu_2)'$ 的椭圆. 一般的 p 元正态密度函数的等高面为

$$(x-\mu)' \Sigma^{-1} (x-\mu) = a^2 \quad (a \geq 0)$$

相关系数为正则 1, 3 象限

Theorem 15.0.28 定理 2.3.1 设 p 维随机向量 $X \sim N_p(\mu, \Sigma)$

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \sim N_p \left(\begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

则

$X^{(1)}$ 与 $X^{(2)}$ 相互独立 $\iff \Sigma_{12} = 0$

(即 $X^{(1)}$ 与 $X^{(2)}$ 互不相关).

Corollary 15.0.29 设 $X = (X_1, \dots, X_p)' \sim N_p(\mu, \Sigma)$, 若 Σ 为对角矩阵, 则 X_1, \dots, X_p 相互独立.

Theorem 15.0.30

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11,2}^{-1} & -\Sigma_{11,2}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11,2}^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11,2}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \end{bmatrix}$$

其中 $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$.

Theorem 15.0.31 设 $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-r}^r \sim N_p(\mu, \Sigma) (\Sigma > 0)$, 则当 $X^{(2)}$ 给定时, $X^{(1)}$ 的条件分

布为

$$\left(X^{(1)} | X^{(2)} \right) \sim N_r(\mu_{1 \cdot 2}, \Sigma_{11 \cdot 2})$$

其中

$$\begin{aligned}\mu_{1 \cdot 2} &= \mu^{(1)} + \Sigma_{12} \Sigma_{22}^{-1} (x^{(2)} - \mu^{(2)}) \\ \Sigma_{11 \cdot 2} &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\end{aligned}$$

1. $X^{(2)}$ 与 $X^{(1)} - \Sigma_{12} \Sigma_{22}^{-1} X^{(2)}$ 相互独立
2. $X^{(1)}$ 与 $X^{(2)} - \Sigma_{21} \Sigma_{11}^{-1} X^{(1)}$ 相互独立
3. $(X^{(2)} | X^{(1)}) \sim N_{p-r}(\mu_{2 \cdot 1}, \Sigma_{22 \cdot 1})$, 其中

$$\begin{aligned}\mu_{2 \cdot 1} &= \mu^{(2)} + \sum_{21} \Sigma_{11}^{-1} (x^{(1)} - \mu^{(1)}) \\ \Sigma_{22 \cdot 1} &= \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}\end{aligned}$$

Definition 15.0.10 设

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-r}^r \sim N_p \left(\begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

又已知 $X^{(2)}$ 给定时 $X^{(1)}$ 的条件分布为

$$\left(X^{(1)} | X^{(2)} \right) \sim N(\mu_{1 \cdot 2}, \Sigma_{11 \cdot 2})$$

则称

$$\mu_{1 \cdot 2} = \mu^{(1)} + \Sigma_{12} \Sigma_{22}^{-1} (x^{(2)} - \mu^{(2)})$$

为**条件期望**, 记为 $E(X^{(1)} | X^{(2)})$; 并称 $\mu_{1 \cdot 2}$ 为 $X^{(1)}$ 对 $X^{(2)}$ 的回归, 称

$$\Sigma_{12} \Sigma_{22}^{-1} \stackrel{\text{def}}{=} B$$

为**回归系数**. 记

$$\Sigma_{11 \cdot 2} = (\sigma_{ij \cdot r+1, \dots, p})_{r \times r} \quad (i, j = 1, \dots, r)$$

称

$$r_{ij \cdot r+1, \dots, p} = \frac{\sigma_{ij \cdot r+1, \dots, p}}{\sqrt{\sigma_{ii \cdot r+1, \dots, p}} \cdot \sqrt{\sigma_{jj \cdot r+1, \dots, p}}}$$

为当 $X^{(2)} = (X_{r+1}, \dots, X_p)'$ 给定时, X_i 与 X_j ($i, j = 1, 2, \dots, r$) 的**偏相关系数**。

Definition 15.0.11 设 $Z = \begin{bmatrix} X \\ Y \end{bmatrix}_1^p \sim N_{p+1} \left(\begin{bmatrix} \mu_X \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{Xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$, 则称

$$R = \left(\frac{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}}{\sigma_{yy}} \right)^{1/2}$$

为 Y 与 $X = (X_1, X_2, \dots, X_p)'$ 的全相关系数

Definition 15.0.12 — 最佳预况. 我们考虑 $r = 1$, 记 $X^{(1)} = Y, g(x^{(2)}) = \mathbf{E}(Y|X^{(2)})$, 则对任意函数 $\varphi()$, :

$$\mathbf{E} \left[(Y - g(x^{(2)}))^2 \right] \leq \mathbf{E} \left[(Y - \varphi(x^{(2)}))^2 \right]$$

即在均方差最小的准则下, 条件期望 $g(x^{(2)})$ 是对 Y 的最佳预测函数。

Definition 15.0.13 — 拉直运算. 所谓拉直运算, 就是将矩阵拉成一个长向量, 通过它来建立矩阵和向量之间的联系. 设随机矩阵 X 是一个 $n \times p$ 矩阵, 用 X 的列向量 X_1, X_2, \dots, X_p 组成一个 np 维向量, 记为

$$\text{Vec}(X) = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} = (x_{11}, x_{21}, \dots, x_{n1}, \dots, x_{1p}, x_{2p}, \dots, x_{np})'$$

符号“Vec”称为拉直运算. 如果将矩阵 X 的行向量(样品)拉直为一个 np 维向量, 用拉直运算的符号可记为

$$\text{Vec}(X') = \begin{bmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{bmatrix} = (x_{11}, x_{12}, \dots, x_{1p}, \dots, x_{n1}, x_{n2}, \dots, x_{np})'$$

对称矩阵的拉直运算. 设 S 是 p 阶对称随机阵, 在 S 矩阵中只包含 $p(p+1)/2$ 个不同的随机变量, 故将其拉直为 p^2 维向量是不合适的, 应拉成 $p(p+1)/2$ 维向量. 设 $S = (S_{ij})_{p \times p}$ 为 p 阶对称矩阵, 令

$$\text{Svec}(S) = (S_{11}, \dots, S_{p1}, S_{22}, \dots, S_{p2}, \dots, S_{pp})'$$

为 $p(p+1)/2$ 维向量. 符号“Svec”称为对称矩阵的拉直运算.

Definition 15.0.14 2. 克罗内克积 设 $A = (a_{ij})$ 和 B 分别为 $n \times p$ 和 $m \times q$ 的矩阵, A 和 B 的克罗内克积 $A \otimes B$ 定义为

$$A \otimes B = (a_{ij}B) = \begin{bmatrix} a_{11}B & \cdots & a_{1p}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{np}B \end{bmatrix}$$

它是 $mn \times pq$ 矩阵. 在多元统计分析中**克罗内克积又称矩阵的直积**,

Definition 15.0.15 设 $X_{(i)} = (x_{i1}, \dots, x_{ip})'$ ($i = 1, \dots, n$) 为来自 p 元正态总体 $N_p(\mu, \Sigma)$ 的随

机样本 (独立同分布), 记随机阵 $X = (x_{ij})_{n \times p}$, 利用拉直运算及矩阵的直积的定义和性质, 可知

$$\text{Vec}(X') \sim N_{np}(\mathbf{1}_n \otimes \mu, I_n \otimes \Sigma)$$

由矩阵的直积的定义, np 维随机向量 $\text{Vec}(X)$ 的均值向量和协方差阵分别为

$$\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \mathbf{1}_n \otimes \mu, \quad \begin{bmatrix} \Sigma & \cdots & O \\ \vdots & & \vdots \\ 0 & \cdots & \Sigma \end{bmatrix} = I_n \otimes \Sigma$$

当随机阵 X 按行拉直后, 如果有

$$\text{Vec}(X') \sim N_{np}(\mathbf{1}_n \otimes \mu, I_n \otimes \Sigma)$$

则称 X 服从矩阵正态分布, 记作

$$X \sim N_{n \times p}(M, I_n \otimes \Sigma)$$

其中

$$\text{Vec}(M') = \mathbf{1}_n \otimes \mu = (\mu_1, \dots, \mu_p, \dots, \mu_1, \dots, \mu_p)'$$

即

$$X \sim N_{n \times p}(M, I_n \otimes \Sigma) \iff \text{Vec}(X') \sim N_{np}(\text{Vec}(M'), I_n \otimes \Sigma)$$

其中

$$M = \begin{bmatrix} \mu_1 & \cdots & \mu_p \\ \vdots & & \vdots \\ \mu_1 & \cdots & \mu_p \end{bmatrix} = \mathbf{1}_n \mu' = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} (\mu_1, \dots, \mu_p)$$

Proposition 15.0.32 随机阵正态分布有如下有用的性质: 设 $X \sim N_{n \times p}(M, I_n \otimes \Sigma)$, A 为 $k \times n$ 常数矩阵, B 为 $q \times p$ 常数矩阵, D 为 $k \times q$ 常数矩阵, 令 $Z = AXB' + D$, 则

$$Z \sim N_{k \times q}(AMB' + D, (AA') \otimes (B\Sigma B'))$$

Definition 15.0.16 — 样本均值向量 \bar{X} .

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_{(i)} = (\bar{x}_1, \dots, \bar{x}_p)' = \frac{1}{n} X' \mathbf{1}_n$$

其中

$$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai} \quad (i = 1, 2, \dots, p)$$

Definition 15.0.17 — 样本离差阵 (又称交叉乘积阵).

$$\begin{aligned} A &= \sum_{a=1}^n (X_{(a)} - \bar{X}) (X_{(a)} - \bar{X})' = X'X - n\bar{X}\bar{X}' \\ &= X' \left[I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n' \right] X \stackrel{\text{def}}{=} (a_{ij})_{p \times p} \end{aligned}$$

其中

$$a_{ij} = \sum_{a=1}^n (x_{ai} - \bar{x}_i) (x_{aj} - \bar{x}_j) \quad (i, j = 1, 2, \dots, p)$$

Definition 15.0.18 — 样本协方差阵 S :

$$S = \frac{1}{n-1} A = (s_{ij})_{p \times p} \quad \left(\text{或 } S^* = \frac{1}{n} A \right)$$

其中

$$s_{ii} = \sum_{a=1}^n (x_{ai} - \bar{x}_i)^2 \quad (i = 1, 2, \dots, p)$$

称为变量 X_i 的样本方差; 样本方差的平方根 $\sqrt{s_{ii}}$ 称为变量 X_i 的样本标准差.

Definition 15.0.19 — 样本相关阵 R .

$$R = (r_{ij})_{p \times p}$$

其中

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}} \sqrt{s_{jj}}} \quad \text{或} \quad \frac{a_{ij}}{\sqrt{a_{ii}} \sqrt{a_{jj}}} \quad (i, j = 1, 2, \dots, p)$$

Theorem 15.0.33 — 似然函数 $L(\mu, \Sigma)$. 把随机数据阵 X 按行拉直后形成的 np 维长向量 $\text{Vec}(X')$ 的联合密度函数看成未知参数 μ, Σ 的函数, 并称为样本 $X_{(i)} (i = 1, \dots, n)$ 的似然

函数, 记为 $L(\mu, \Sigma)$

$$\begin{aligned}
L(\mu, \Sigma) &= \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (x_{(i)} - \mu)' \Sigma^{-1} (x_{(i)} - \mu) \right] \\
&= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_{(i)} - \mu)' \Sigma^{-1} (x_{(i)} - \mu) \right] \\
&= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp \left[-\frac{1}{2} \sum_{i=1}^n \text{tr} \left((x_{(i)} - \mu)' \Sigma^{-1} (x_{(i)} - \mu) \right) \right] \\
&= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp \left[-\frac{1}{2} \sum_{i=1}^n \text{tr} \left(\Sigma^{-1} (x_{(i)} - \mu) (x_{(i)} - \mu)' \right) \right] \\
&= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \exp \left[\text{tr} \left(-\frac{1}{2} \Sigma^{-1} \sum_{i=1}^n (x_{(i)} - \mu) (x_{(i)} - \mu)' \right) \right] \\
&\stackrel{\text{def}}{=} \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} \text{etr} \left(-\frac{1}{2} \Sigma^{-1} \sum_{i=1}^n (x_{(i)} - \mu) (x_{(i)} - \mu)' \right)
\end{aligned}$$

其中

$$\begin{aligned}
&\sum_{i=1}^n (x_{(i)} - \mu) (x_{(i)} - \mu)' \\
&= \sum_{i=1}^n (x_{(i)} - \bar{X} + \bar{X} - \mu) (x_{(i)} - \bar{X} + \bar{X} - \mu)' \\
&= \sum_{i=1}^n (x_{(i)} - \bar{X}) (x_{(i)} - \bar{X})' + n(\bar{X} - \mu)(\bar{X} - \mu)' \\
&= A + n(\bar{X} - \mu)(\bar{X} - \mu)'
\end{aligned}$$

由于 $\ln x$ 是 x 的单调函数, $L(\mu, \Sigma)$ 与 $\ln L(\mu, \Sigma)$ 有相同的最大值点. 以下只须讨论 $\ln L(\mu, \Sigma)$ 的最大值问题.

Theorem 15.0.34 设 B 为 p 阶正定矩阵, 则

$$\text{tr} B - \ln |B| \geq p$$

且等号成立的充分必要条件是 $B = I_p$

Proof. 因为 $B > 0$, 所以 B 的全部特征值 $\lambda_1, \dots, \lambda_p > 0$, 且 $|B| = \lambda_1 \cdots \lambda_p$. 利用不等式 $\ln(1+x) \leq x$ (当 $x+1 > 0$), 可得

$$\begin{aligned}
\ln |B| &= \sum_{i=1}^p \ln \lambda_i = \sum_{i=1}^p \ln (1 + \lambda_i - 1) \\
&\leq \sum_{i=1}^p (\lambda_i - 1) = \text{tr}(B) - p
\end{aligned}$$

所以

$$\text{tr} B - \ln |B| \geq p$$

因不等式 $\ln(1+x) \leq x$ 中的等号仅当 $x=0$ 时成立, 故引理给出的不等式仅当 $\lambda_i - 1 = 0 (i=1, \dots, p)$ 时成立, 即 $B = I_p$ 反之, 当 $B = I_p$ 时, $\ln|I_p| = 0, \text{tr}B = p$, 故引理给出的不等式中的等号成立。 ■

Theorem 15.0.35 设 $X_{(i)} (i=1, \dots, n)$ 是多元正态总体 $N_p(\mu, \Sigma)$ 的随机样本, $n > p$, 则 μ, Σ 的最大似然估计为 $\hat{\mu} = \bar{X}$, $\hat{\Sigma} = \frac{1}{n}A$

Theorem 15.0.36 设 $X_{(t)} = (x_{t1}, \dots, x_{tp})' (t=1, \dots, n)$ 独立同 $N_p(\mu, \Sigma)$ 分布, 且 $\Sigma > 0$, 记

$$\bar{X} = \frac{1}{n} \sum_{t=1}^n X_{(t)}, \quad A = \sum_{t=1}^n (X_{(t)} - \bar{X})(X_{(t)} - \bar{X})'$$

设 \bar{X} 和 A 分别为 p 元正态总体 $N_p(\mu, \Sigma)$ 的样本均值向量和样本离差阵, 则

1. $\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$
2. $A = \sum_{t=1}^{n-1} Z_t Z_t'$, 其中 Z_1, \dots, Z_{n-1} 独立同 $N_p(0, \Sigma)$ 分布
3. \bar{X} 和 A 相互独立
4. $P\{A > 0\} = 1 \iff n > p$

Proof. 设 Γ 是 n 阶正交矩阵, 具有以下形式

$$\Gamma = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & & \vdots \\ r_{(n-1)1} & \cdots & r_{(n-1)n} \\ 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{bmatrix} = (r_{ij})_{n \times n}$$

令

$$Z = \begin{bmatrix} Z'_1 \\ \vdots \\ Z'_n \end{bmatrix} = \Gamma \begin{bmatrix} X'_{(1)} \\ \vdots \\ X'_{(n)} \end{bmatrix} = \Gamma X$$

即

$$Z_t = (X_{(1)}, \dots, X_{(n)}) \begin{bmatrix} r_{t1} \\ \vdots \\ r_{tn} \end{bmatrix} \quad (t=1, \dots, n)$$

为 p 维随机向量. 因 Z_t 是 p 维正态随机向量 $X_{(1)}, \dots, X_{(n)}$ 的线性组合, 故 Z_t 也是 p 维正态随机向量, 且

$$\begin{aligned} E(Z_t) &= \sum_{i=1}^n r_{ti} E(X_{(i)}) = \begin{cases} 0, & \text{当 } t \neq n \text{ 时,} \\ \sqrt{n}\mu, & \text{当 } t = n \text{ 时} \end{cases} \\ \text{Cov}(Z_a, Z_\beta) &= E[(Z_a - E(Z_a))(Z_\beta - E(Z_\beta))'] \\ &= \sum_{i=1}^n r_{ai} r_{\beta i} \Sigma = \begin{cases} O, & \text{当 } \alpha \neq \beta \text{ 时,} \\ \Sigma, & \text{当 } \alpha = \beta \text{ 时.} \end{cases} \end{aligned}$$

(1) 因为 $Z_n = \frac{1}{\sqrt{n}} \sum_{a=1}^n X_{(a)} = \sqrt{n}X \sim N_p(\sqrt{n}\mu, \Sigma)$, 故有

$$\bar{X} = \frac{1}{\sqrt{n}} Z_n \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$$

(2) 因为

$$\begin{aligned} \sum_{a=1}^n Z_a Z'_a &= (Z_1, \dots, Z_n) \begin{bmatrix} Z'_1 \\ \vdots \\ Z'_n \end{bmatrix} = Z' Z \\ &= X' \Gamma' \cdot \Gamma X = X' X = \sum_{a=1}^n X_{(a)} X'_{(a)} \end{aligned}$$

且

$$\begin{aligned} \sum_{a=1}^{n-1} Z_a Z'_a &= \sum_{a=1}^n X_{(a)} X'_{(a)} - Z_n Z'_n = \sum_{a=1}^n X_{(a)} X'_{(a)} - n \bar{X} \bar{X}' \\ &= \sum_{a=1}^n (X_{(a)} - \bar{X}) (X_{(a)} - \bar{X})' = A \end{aligned}$$

(3) 因 $A = \sum_{a=1}^{n-1} Z_a Z'_a$ 是 Z_1, \dots, Z_{n-1} 的函数, \bar{X} 是 Z_n 的函数, 而 Z_1, \dots, Z_{n-1} 与 Z_n 相互独立, 故 A 与 \bar{X} 也相互独立. (4) 记 $B = (Z_1, \dots, Z_{n-1})$, 则 $A = BB'$, 以下来证明: $P\{A > 0\} = 1$ 的充要条件是 $n > p$ 因为 $A = BB'$, B 是 $p \times (n-1)$ 矩阵. 显然 $\text{rank}(A) = \text{rank}(B)$ 当 A 为正定矩阵时 A 的秩是 p , 故 B 的秩也是 p . 从而 $p < n$ 反之, 设 $n > p$, 我们来证明 $P\{A > 0\} = 1$, 为此只须证 $P\{B$ 的前 p 列线性相关 } = 0. 容易看出

$$\begin{aligned} P\{B \text{ 的前 } p \text{ 列线性相关}\} &= P\{Z_1, \dots, Z_p \text{ 线性相关}\} \\ &\leq \sum_{i=1}^p P\{Z_i \text{ 可表成 } Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_p \text{ 的线性组合}\} \\ &= p \cdot P\{Z_1 \text{ 可表成 } Z_2, \dots, Z_p \text{ 的线性组合}\} \\ &= p \cdot E[P\{Z_1 \text{ 可表成 } z_2, \dots, z_p \text{ 的线性组合} | Z_2 = z_2, \dots, Z_p = z_p\}] \\ &= p \cdot E[P\{Z_1 \text{ 落入由 } z_2, \dots, z_p \text{ 张成的子空间} | Z_2 = z_2, \dots, Z_p = z_p\}] \\ &= p \cdot E[P\{\text{存在 } p \text{ 维常向量 } \alpha \neq 0, \text{ 使 } \alpha' Z_1 = |Z_2 = z_2, \dots, Z_p = z_p\}] \\ &= p \cdot E(0) = 0 \end{aligned}$$

在证明过程中用到以下事实: 由于 $Z_1 \sim N_p(0, \Sigma)$, 而 $\Sigma > 0$ (正定), 对常向量 $\alpha \neq 0$, $\alpha' Z_1 \sim N(0, \alpha' \Sigma \alpha)$, 且 $\alpha' \Sigma \alpha > 0$, 即 $P\{\alpha' Z_1 = 0\} = 0$. 或者说 Z_1 取值落入任何维数小于 p 的子空间的概率是 0. ■

Proposition 15.0.37 — 无偏性.

$$E(\bar{X}) = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n E(x_{i1}) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n E(x_{ip}) \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n \mu_1 \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \mu_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \mu$$

故 \bar{X} 是 μ 的无偏估计.

$$\begin{aligned} E(A) &= E\left(\sum_{a=1}^{n-1} Z_a Z'_a\right) = \sum_{a=1}^{n-1} (E(Z_a Z'_a)) \\ &= \left(\sum_{a=1}^{n-1} D(Z_a)\right) = (n-1)\Sigma \end{aligned}$$

因而 Σ 的最大似然估计 $\hat{\Sigma} = \frac{1}{n} A$ 不是无偏估计. 为了得到无偏估计量, 常作如下修正: 令 $S = \frac{1}{n-1} A$, 则 S 是 Σ 的无偏估计. 常称 $\bar{X} = (\bar{x}_1, \dots, \bar{x}_p)'$ 为样本均值; $S = \frac{1}{n-1} A$ 为样本协方差阵.

Proposition 15.0.38 — 有效性. 可以证明 \bar{X}, S 是 μ, Σ 的“最小方差”无偏估计量, 即 \bar{X}, S 是 μ Σ 的有效估计量 (见参考文献 [2])。

Proposition 15.0.39 — 相合性 (一致性). 可以证明当 $n \rightarrow \infty$ 时 $\bar{X}, \hat{\Sigma}$ 是 μ, Σ 的强相合估计. 实际上, 因 $E(\bar{X}) = \mu$, 由强大数定律知

$$P\left\{\lim_n \bar{X} = \mu\right\} = 1$$

另一方面, 因 $\hat{\Sigma} = \frac{1}{n} \sum_{\alpha=1}^{n-1} Z_\alpha Z'_\alpha$, 而 Z_1, \dots, Z_{n-1} 相互独立同分布共同分布是 $N_p(0, \Sigma)$, 而 $E(Z_\alpha Z'_\alpha) = \Sigma (\alpha = 1, \dots, n-1)$. 再利用强大数定律知

$$P\left\{\lim_n \hat{\Sigma} = \Sigma\right\} = 1$$

Proposition 15.0.40 还可以证明 $\bar{X}, \hat{\Sigma}$ 是 μ, Σ 的充分统计量; \bar{X} 是 μ 的极小极大估计量 (最大风险达最小); 且估计量具有渐近正态性.

Definition 15.0.20 设参数向量 θ 的变化范围是 $\Theta \in \mathbb{R}^k$. $L(\theta)$ 是似然函数. 设 $w = g(\theta)$ 是 Θ 到 Θ^* 上的博雷尔 (Borel) 可测映射, 这里 Θ^* 是 \mathbb{R}^k 的子集. 对任何 $w \in \Theta^*$, 令

$$M(w) = \text{Sup}_{\{\theta: g(\theta)=w\}} L(\theta)$$

Definition 15.0.21 称 $M(w)$ 为函数 $g(\theta)$ 诱导出的似然函数.

Definition 15.0.22 若 \hat{w} 满足 $M(\hat{w}) = \text{Sup}_w M(w)$, 则称 \hat{w} 是 $g(\theta)$ 的最大似然估计.

Theorem 15.0.41 若 $\hat{\theta}$ 是 θ 的最大似然估计, 则 $\hat{w} = g(\hat{\theta})$ 是 $g(\theta)$ 的最大似然估计.

Corollary 15.0.42 既然多元正态分布 $N_p(\mu, \Sigma)$ 的参数 μ 和 Σ 有最大似然估计量 $\hat{\mu} = \bar{X}, \hat{\Sigma} = \frac{1}{n} A$, 函数 $g(\mu, \Sigma)$ 的最大似然估计为

$$g\left(\bar{X}, \frac{1}{n} A\right)$$

■ **Example 15.3** 设 p 维正态随机向量 $X = (X_1, \dots, X_p)', X_i, X_j$ 的相关系数为

$$\rho_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i) \cdot \text{Var}(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \cdot \sigma_{jj}}}$$

其中 σ_{ij} 是协方差阵 Σ 的第 i 行第 j 列的元素. ρ_{ij} 的最大似然估计量 r_{ij}

$$r_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii} \cdot \hat{\sigma}_{jj}}} = \frac{a_{ij}}{\sqrt{a_{ii} \cdot a_{jj}}}$$

15.1 假设检验

Definition 15.1.1 分量独立的 n 维随机向量 X 的二次型: 设 $X_i \sim N_1(\mu_i, \sigma^2)$ ($i = 1, \dots, n$), 且相互独立, 记

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

则 $X \sim N_n(\mu, \sigma^2 I_n)$, 其中 $\mu = (\mu_1, \dots, \mu_n)'$

Proposition 15.1.1 当 $\mu_i = 0(i = 1, \dots, n), \sigma^2 = 1$ 时, 则

$$\xi = X'X = \sum_{i=1}^n X_i^2 \sim \chi^2(n)$$

Proposition 15.1.2 当 $\mu_i = 0(i = 1, 2, \dots, n), \sigma^2 \neq 1$ 时, 则有

$$\frac{1}{\sigma^2} X'X \sim \chi^2(n) \quad (\text{或记为 } X'X \sim \sigma^2 \chi^2(n))$$

Definition 15.1.2 当 $\mu_i \neq 0(i = 1, 2, \dots, n), X'X$ 的分布常称为非中心 χ^2 分布.

Definition 15.1.3 设 n 维随机向量 $X \sim N_n(\mu, I_n)(\mu \neq 0)$, 则称随机变量 $\xi = X'X$ 为服从 n 个自由度, 非中心参数 $\delta = \mu'\mu = \sum_{i=1}^n \mu_i^2$ 的 χ^2 分布, 记为 $X'X \sim \chi^2(n, \delta)$ 或 $X'X \sim \chi_n^2(\delta)$ 当 $X \sim N_n(\mu, \sigma^2 I_n), \mu \neq 0$, 且 $\sigma^2 \neq 1$ 时,

$$Y_i = \frac{1}{\sigma} X_i$$

显然

$$Y_i \sim N\left(\frac{\mu_i}{\sigma}, 1\right) \quad (i = 1, \dots, n)$$

则

$$Y'Y = \frac{1}{\sigma^2} X'X \sim \chi_n^2(\delta)$$

其中 $\delta = \frac{1}{\sigma^2} \mu'\mu$

Theorem 15.1.3 设 $X \sim N_n(0_n, \sigma^2 I_n)$, A 为对称矩阵, 且 $\text{rank}(A) = r$, 则二次型 $X'AX/\sigma^2 \sim \chi^2(r) \iff A^2 = A$ (A 为对称幂等矩阵)

Theorem 15.1.4 设 $X \sim N_n(\mu, \sigma^2 I_n), A = A'$, 则

$$\frac{1}{\sigma^2} X'AX \sim \chi^2(r, \delta)$$

其中 $\delta = \frac{1}{\sigma^2} \mu'A\mu \iff A = A^2$ (A 为对称幂等矩阵) 且 $\text{rank}(A) = r(r \leq n)$

Theorem 15.1.5 二次型与线性函数的独立性: 设 $X \sim N_n(\mu, \sigma^2 I_n)$, A 为 n 阶对称矩阵, B 为 $m \times n$ 矩阵, 令 $\xi = X'AX, Z = BX, Z$ 为 m 维随机向量, 若 $BA = O$, 则 BX 和 $X'AX$ 相互独立.

Theorem 15.1.6 若 BX 和 $X'AX$ 相互独立, 则 $BA = O$

Theorem 15.1.7 两个二次型相互独立的条件: 设 $X \sim N_n(\mu, \sigma^2 I_n)$, A, B 为 n 阶对称矩阵, 则

$$AB = O \iff X'AX \text{ 与 } X'BX \text{ 相互独立.}$$

Theorem 15.1.8 — p 维随机向量的二次型. 设 $X \sim N_p(\mu, \Sigma), \Sigma > 0$, 则 $X'\Sigma^{-1}X \sim \chi^2(p, \delta)$, 其中

$$\delta = \mu'\Sigma^{-1}\mu$$

Theorem 15.1.9 设 $X \sim N_p(\mu, \Sigma), \Sigma > 0, A$ 为对称矩阵, $\text{rank}(A) = r$. 则

$$(X - \mu)'A(X - \mu) \sim \chi^2(r) \iff \Sigma A \Sigma A \Sigma = \Sigma A \Sigma$$

Theorem 15.1.10 设 $X \sim N_p(\mu, \Sigma), \Sigma > 0, A$ 和 B 为 p 阶对称矩阵, 则

$$\begin{aligned} (X - \mu)'A(X - \mu) \text{ 与 } (X - \mu)'B(X - \mu) \text{ 独立} \\ \iff \Sigma A \Sigma B \Sigma = O_{p \times p} \end{aligned}$$

Definition 15.1.4 — 非中心 t 分布. 设 $X \sim N(\delta, 1)$ 与 $Y \sim \chi^2(n)$ 相互独立, 令

$$T = \frac{X\sqrt{n}}{\sqrt{Y}}$$

则称 T 的分布为具有 n 个自由度、非中心参数为 δ 的非中心 t 分布记为 $T \sim t(n, \delta)$

Definition 15.1.5 — 非中心 F 分布. 设 $X \sim \chi^2(m, \delta)$ 与 $Y \sim \chi^2(n)$ 独立, 令

$$F = \frac{X/m}{Y/n}$$

则称 F 的分布为具有自由度为 m, n 和非中心参数为 δ 的 F 分布, 记

$$\text{为 } F \sim F(m, n, \delta)$$

4. 非中心 χ^2 分布, 非中心 t 分布和非中心 F 分布的成用一元统计中, 关于在一个正态总体 $N(\mu, \sigma^2)$ 的均值检验中, 检验 $H_0: \mu = \mu_0$ 时, 检验统计量为

$$T = \frac{\bar{X} - \mu_0}{\sqrt{s^2/n}} \stackrel{H_0}{\sim} t(n-1)$$

否定域为 $\{|T| > \lambda\}$, 其中 λ 满足: $P\{|T| > \lambda\} = \alpha$ (显著性水平)

1. 当否定 H_0 时, 可能犯第一类错误, 且第一类错误的概率 = $P\{\text{“以真当假”}\} = P\{|T| > \lambda | \mu = \mu_0\} = \alpha$
2. 当 H_0 相容时, 可能犯第二类错误, 且第二类错误的概率 = $P\{\text{“以假当真”}\} = P\{|T| \leq \lambda | \mu \neq \mu_0\}$, 设 $\mu = \mu_1 \neq \mu_0$

$$P\left\{\left|\frac{\bar{X} - \mu_1 + (\mu_1 - \mu_0)}{\sqrt{s^2/n}}\right| \leq \lambda | \mu = \mu_1\right\} = \beta$$

此时检验统计量 $T \sim t(n-1, \delta)$ ($\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma$) 利用非中心 t 分布可以计算第二类错误 β 的值, 从而得到检验法的功效函数为 $1 - \beta$ 类似地, 非中心 χ^2 分布和非中心 F 分布在一元统计的相应检验中, 将应用非中心分布来计算第二类错误。

二、威沙特 (Wishart) 分布威沙特分布是一元统计中 χ^2 分布的推广. 多元正态总体 $N_p(\mu, \Sigma)$ 中, 常用样本均值向量 \bar{X} 作为 μ 的估计, 样本协方差阵

$$S = \frac{1}{n-1} A$$

作为 Σ 的估计.

$$\bar{X} \sim N_p\left(\mu, \frac{\Sigma}{n}\right)$$

一元统计中, 用样本方差

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{(i)} - \bar{X})^2$$

作为 σ^2 的估计, 而且知道

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_{(i)} - \bar{X})^2 \sim \chi^2(n-1)$$

推广到 p 元正态总体, 样本协方差阵 $S = \frac{1}{n-1} A$ 及随机阵 A (离差阵

设 $X_{(a)}$ ($a = 1, \dots, n$) 为来自总体 $N_p(0, \Sigma)$ 的随机样本, 记 $X = (X_{(1)}, \dots, X_{(n)})'$ 为 $n \times p$ 样本数据阵. 考虑随机阵

$$W = \sum_{i=1}^n X_{(i)} X'_{(i)} = (X_{(1)}, \dots, X_{(n)}) \begin{bmatrix} X'_{(1)} \\ \vdots \\ X'_{(n)} \end{bmatrix} = X' X$$

的分布. 当 $p = 1$ 时 (总体 $X \sim N_1(0, \sigma^2)$)

$$W = \sum_{i=1}^n X_{(i)}^2 = (X_{(1)}, \dots, X_{(n)}) \begin{bmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{bmatrix} = X' X \sim \sigma^2 \chi^2(n)$$

在一元正态总体情况下,

$$\xi = \frac{1}{\sigma^2} \sum_{i=1}^n X_{(i)}^2 \sim \chi^2(n)$$

Definition 15.1.6 — 威沙特分布的定义. 设 $X_{(a)} \sim N_p(0, \Sigma)$ ($a = 1, \dots, n$) 相互独立, 记 $X = (X_{(1)}, \dots, X_{(n)})'$ 为 $n \times p$ 矩阵, 则称随机阵

$$W = \sum_{a=1}^n X_{(a)} X'_{(a)} = X' X$$

的分布为威沙特分布, 记为 $W \sim W_p(n, \Sigma)$ 显然, $p = 1$ 时, $X_{(a)} \sim N(0, \sigma^2)$, 此时

$$W = \sum_{a=1}^n X_{(a)}^2 \sim \sigma^2 \chi^2(n)$$

即 $W_1(n, \sigma^2)$ 就是 $\sigma^2 \chi^2(n)$. 当 $p = 1, \sigma^2 = 1$ 时, $W_1(n, 1)$ 就是 $\chi^2(n)$
一般地, 设 $X_{(a)} \sim N_p(\mu, \Sigma) (\alpha = 1, \dots, n)$ 相互独立, 记

$$M = \begin{bmatrix} \mu_1 & \cdots & \mu_p \\ \vdots & & \vdots \\ \mu_1 & \cdots & \mu_p \end{bmatrix} = \mathbf{1}_n \mu'$$

则称 $W = X'X$ 服从非中心参数为 Δ 的非中心威沙特分布, 记为 $W \sim W_p(n, \Sigma, \Delta)$, 其中

$$\Delta = M'M = (\mathbf{1}_n \mu')' (\mathbf{1}_n \mu') = \mu \mathbf{1}'_n \mathbf{1}_n \mu' = n \mu \mu'$$

当 $X_{(a)} \sim N_p(\mu_a, \Sigma) (\alpha = 1, \dots, n)$ 相互独立时, 非中心参数

$$\Delta = \sum_{\alpha=1}^n \mu_\alpha \mu'_\alpha \quad \text{或} \quad \Delta = M'M$$

这里

$$M = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1p} \\ \vdots & & \vdots \\ \mu_{n1} & \cdots & \mu_{np} \end{bmatrix} = \begin{bmatrix} \mu'_1 \\ \vdots \\ \mu'_n \end{bmatrix}$$

其中 p 为随机阵 W 的阶数, n 为自由度, 一元统计中的 σ^2 对应 p 元统计中的协方差阵 Σ .

威沙特分布的性质

Proposition 15.1.11 设 $X_{(a)} \sim N_p(\mu, \Sigma) (\alpha = 1, \dots, n)$ 相互独立, 则样本离差阵 A 服从威沙特分布, 即

$$A = \sum_{a=1}^n (X_{(a)} - \bar{X}) (X_{(a)} - \bar{X})' \sim W_p(n-1, \Sigma)$$

Proof.

$$A = \sum_{a=1}^{n-1} Z_a Z_a'$$

而 $Z_a \sim N_p(0, \Sigma) (\alpha = 1, \dots, n-1)$ 相互独立,

$$A \sim W_p(n-1, \Sigma)$$

■

Proof. 关于自由度 n 具有可加性: 设 $W_i \sim W_p(n_i, \Sigma) (i = 1, \dots, k)$ 相互独立, 则

$$\sum_{i=1}^k W_i \sim W_p(n, \Sigma), \quad \text{其中 } n = n_1 + \dots + n_k$$

■

Proposition 15.1.12 设 p 阶随机阵 $W \sim W_p(n, \Sigma)$, C 是 $m \times p$ 常数矩阵, 则 m 阶随机阵 CWC' 也服从威沙特分布, 即

$$CWC' \sim W_m(n, C\Sigma C')$$

Proof. 因 $W \stackrel{d}{=} \sum_{\alpha=1}^n Z_\alpha Z'_\alpha \sim W_p(n, \Sigma)$, 其中 $Z_\alpha \sim N_p(0, \Sigma) (\alpha = 1, \dots, n)$ 相互独立. 令 $Y_a = CZ_a$, 则 $Y_a \sim N_m(0, C\Sigma C')$. 故

$$\sum_{a=1}^n Y_a Y'_a = \sum_{a=1}^n CZ_a \cdot Z'_a C' \stackrel{d}{=} CWC' \sim W_m(n, C\Sigma C')$$

特别地:

1. $aW \sim W_p(n, a\Sigma) (a > 0, \text{ 为常数})$ 在性质 3 中只须取 $C = \sqrt{a}I_p$, 即得此结论.
2. 设 $l' = (l_1, \dots, l_p)$, 则 $l'Wl = \xi \sim W_1(n, l'\Sigma l)$, 即

$$\xi \sim \sigma^2 \chi^2(n) \quad (\text{其中 } \sigma^2 = l'\Sigma l)$$

在性质 3 中只须取 $C = l'$, 即得此结论. ■

Proposition 15.1.13 $X_{(\alpha)} \sim N_p(0, \Sigma) (\alpha = 1, \dots, n)$ 相互独立, 其中

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}_{p-r}^r$$

又已知随机阵

$$W = \sum_{a=1}^n X_{(a)} X'_{(\alpha)} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}_{p-r}^r \sim W_p(n, \Sigma)$$

则

1. $W_{11} \sim W_r(n, \Sigma_{11}), W_{22} \sim W_{p-r}(n, \Sigma_{22})$
2. 当 $\Sigma_{12} = O$ 时, W_{11} 与 W_{22} 相互独立

Proposition 15.1.14 设 $W \sim W_p(n, \Sigma)$, 记 $W_{22.1} = W_{22} - W_{21}W_{11}^{-1}W_{12}$, 则

$$W_{22.1} \sim W_{p-r}(n-r, \Sigma_{22.1})$$

其中 $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, 且 $W_{22.1}$ 与 W_{11} 相互独立.

Proposition 15.1.15 设随机阵 $W \sim W_p(n, \Sigma)$, 则 $E(W) = n\Sigma$

Proposition 15.1.16 设 $X \sim N_{n \times p}(M, I_n \otimes \Sigma), A$ 为 n 阶对称矩阵, 则

$$X'AX \sim W_p(r, \Sigma, \Delta)$$

其中 $\Delta = M'AM \iff A^2 = A$, 且 $\text{rank}(A) = r$ 这是一元统计中 n 维观测向量 X 的二次型分布 在 p 维情况下的推广。

Proposition 15.1.17 设 $X \sim N_{n \times p}(M, I_n \otimes \Sigma), A$ 和 B 均为 n 阶对称幂等矩阵, 则 $X'AX$ 与 $X'BX$ 相互独立 $\iff AB = O$ 这是一元统计中 $(p=1)n$ 维观测向量 X 的两个二次型相互独立的条件在 p 维情况下的推广.

15.1.1 霍特林 (Hotelling) T^2 分布

Definition 15.1.7 — 霍特林 T^2 分布的定义. 一元统计中, 若 $X \sim N(0, 1), \xi \sim \chi^2(n), X$ 与 ξ 相互独立, 则随机变量

$$t = \frac{X}{\sqrt{\xi/n}} \sim t(n)$$

下面把 $t^2 = nX^2/\xi = nX'\xi^{-1}X$ 的分布推广到 p 元总体. 设总体 $X \sim N_p(0, \Sigma)$, 随机阵

$W \sim W_p(n, \Sigma)$, 我们来讨论 $T^2 = nX'W^{-1}X$ 的分布。

Definition 15.1.8 $X \sim N_p(0, \Sigma)$, 随机阵 $W \sim W_p(n, \Sigma)$ ($\Sigma > 0, n \geq p$), 且 X 与 W 相互独立, 则称统计量 $T^2 = nX'W^{-1}X$ 为霍特林 T^2 统计量, 其分布称为服从 n 个自由度的 T^2 分布, 记为

$$T^2 \sim T^2(p, n)$$

更一般地, 若 $X \sim N_p(\mu, \Sigma)$ ($\mu \neq 0$), 则称 T^2 的分布为非中心霍特林 T^2 分布, 记为 $T^2 \sim T^2(p, n, \mu)$

霍特林 T^2 分布的性质

Proposition 15.1.18 设 $X_{(a)}$ ($a = 1, \dots, n$) 是来自 p 元总体 $N_p(\mu, \Sigma)$ 的随机样本, \bar{X} 和 A 分别是正态总体 $N_p(\mu, \Sigma)$ 的样本均值向量和样本离差阵, 则统计量

$$\begin{aligned} T^2 &= (n-1)[\sqrt{n}(\bar{X} - \mu)]' A^{-1} [\sqrt{n}(\bar{X} - \mu)] \\ &= n(n-1)(\bar{X} - \mu)' A^{-1} (\bar{X} - \mu) \\ &\sim T^2(p, n-1) \end{aligned}$$

Proof. 因 $\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$, 则 $\sqrt{n}(\bar{X} - \mu) \sim N_p(0, \Sigma)$ 而 $A \sim W_p(n-1, \Sigma)$, 且 A 与 \bar{X} 相互独立. 由定义知

$$T^2 \sim T^2(p, n-1)$$

■

Proposition 15.1.19 T^2 与 F 分布的关系: 设 $T^2 \sim T^2(p, n)$, 则

$$\frac{n-p+1}{np} T^2 \sim F(p, n-p+1)$$

在一元统计中, 若 $t = \frac{X}{\sqrt{\xi/n}} \sim t(n)$, 则 $t^2 = \frac{X^2/1}{\xi/n} \sim F(1, n)$ 当 $p=1$ 时, 一元总体 $X \sim N(0, \sigma^2)$, $X_{(a)}$ ($a = 1, \dots, n$) 为来自总体 X 的随机样本, 则

$$W = \sum_{a=1}^n X_{(a)} X'_{(a)} = \sum_{a=1}^n X_{(a)}^2 \sim W_1(n, \sigma^2) \quad (\text{即 } \sigma^2 \chi^2(n))$$

所以

$$\frac{n}{n} T^2 = nX'W^{-1}X = \frac{nX^2}{W} = \frac{(X/\sigma)^2}{(W/\sigma^2 n)} \sim F(1, n)$$

一般地,

$$\begin{aligned} &\frac{n-p+1}{p} \cdot \frac{T^2}{n} \stackrel{d}{=} \frac{n-p+1}{p} X'W^{-1}X \\ &= \frac{n-p+1}{p} X'\Sigma^{-1}X / \frac{X'W^{-1}X}{X'W^{-1}X} \stackrel{d}{=} \frac{n-p+1}{p} \cdot \frac{\xi}{\eta} \\ &= \frac{\xi/p}{\eta/(n-p+1)} \sim F(p, n-p+1) \end{aligned}$$

其中 $\xi = X'\Sigma^{-1}X \sim \chi^2(p, \delta)$ ($\delta = 0$). 还可证明

$$\eta = \frac{X'W^{-1}X}{X'W^{-1}X} \sim \chi^2(n-p+1)$$

且 ξ 与 η 独立

Proposition 15.1.20 设 $X_{(\alpha)} (\alpha = 1, 2, \dots, n)$ 为来自 p 元总体 $N_p(\mu, \Sigma)$ 的随机样本. \bar{X}, A 分别为样本均值向量和样本离差阵. 记

$$T^2 = n(n-1)\bar{X}A^{-1}\bar{X}$$

则

$$\frac{n-p}{p} \frac{T^2}{n-1} \sim F(p, n-p, \delta)$$

其中 $\delta = n\mu'\Sigma^{-1}\mu$.

Proposition 15.1.21 一元统计中 ($p=1$ 时), t 统计量与参数 σ^2 无关.

Proposition 15.1.22 — T^2 统计量的分布只与 p, n 有关, 而与 Σ 无关. 设 $U \sim N_p(0, I_p), W_0 \sim W_p(n, I_p)$, U 和 W_0 相互独立, 则

$$nU'W_0^{-1}U = nX'W^{-1}X \sim T^2(p, n)$$

事实上, 因 $X \sim N_p(0, \Sigma) (\Sigma > 0), W \sim W_p(n, \Sigma)$, 则 $\Sigma^{-1/2}X \sim N_p(0, I_p)$, 且 $\Sigma^{-1/2}W\Sigma^{-1/2} \sim W_p(n, I_p)$, 因此

$$U \stackrel{d}{=} \Sigma^{-1/2}X, \quad W_0 \stackrel{d}{=} \Sigma^{-1/2}W\Sigma^{-1/2}$$

所以 $nU'W_0^{-1}U \stackrel{d}{=} nX'W^{-1}X \sim T^2(p, n)$

Proposition 15.1.23 — T^2 统计量对非退化变换保持不变.. 设 $X_{(\alpha)} (\alpha = 1, \dots, n)$ 是来自 p 元总体 $N_p(\mu, \Sigma)$ 的随机样本, \bar{X}_x 和 A_x 分别表示正态总体 X 的样本均值向量和样本离差阵, 则由性质 1 有

$$T_x^2 = n(n-1)(\bar{X}_x - \mu)'A_x^{-1}(\bar{X}_x - \mu) \sim T^2(p, n-1)$$

令 $Y_{(\alpha)} = CX_{(\alpha)} + d (\alpha = 1, \dots, n)$, 其中 C 为 $p \times p$ 非退化常数矩阵, d 为 p 维常向量, 则可以证明

$$T_y^2 = T_x^2$$

15.1.2 威尔克斯 (Wilks) Λ 统计量及其分布

Definition 15.1.9 — 威尔克斯 Λ 分布的定义. 一元统计中, 设 $\xi \sim \chi^2(m), \eta \sim \chi^2(n)$, 且相互独立, 则

$$F = \frac{\xi/m}{\eta/n} \sim F(m, n)$$

在两个总体 $(N(\mu_1, \sigma_x^2) \text{ 和 } N(\mu_2, \sigma_y^2))$ 方差齐性检验中 ($H_0: \sigma_x^2 = \sigma_y^2$), 设 $X_{(i)} (i = 1, \dots, m)$ 为来自 $N(\mu_1, \sigma_x^2)$ 的随机样本, $Y_{(j)} (j = 1, \dots, n)$ 为来自 $N(\mu_2, \sigma_y^2)$ 的随机样本, 取 σ_x^2 和 σ_y^2 的估计量 (样本方差) 分别为

$$s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (X_{(i)} - \bar{X})^2 \quad \text{和} \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_{(i)} - \bar{Y})^2$$

则检验统计量

$$F = \frac{s_x^2}{s_y^2} \stackrel{H_0}{\sim} F(m-1, n-1)$$

在 p 元总体 $N_p(\mu, \Sigma)$ 中, 协方差阵 Σ 的估计量为

$$\hat{\Sigma} = \frac{1}{n-1} A \quad \left(\text{或 } \frac{1}{n} A \right)$$

在检验 $H_0 : \Sigma_1 = \Sigma_2$ 时, 如何用一个数值来描述对矩阵的离散程度的估计呢? 一般可用矩阵的行列式. 迹或特征值等数量指标来描述总体的分散程度。

设 $X \sim N_p(\mu, \Sigma)$, 则称协方差阵的行列式 $|\Sigma|$ 为 X 的广义方差. 若 $X_{(a)} (\alpha = 1, \dots, n)$ 为 p 元总体 X 的随机样本, A 为样本离差阵, 则称 $|\frac{1}{n} A|$ 或 $|\frac{1}{n-1} A|$ 为样本广义方差. 有了广义方差的概念后, 在多元统计的协方差阵齐性检验中, 类似一元统计, 可考虑两个广义方差之比构成的统计量--威尔克斯统计量的分布。

Definition 15.1.10 设 $A_1 \sim W_p(n_1, \Sigma), A_2 \sim W_p(n_2, \Sigma) (\Sigma > 0, n_1 \geq p)$, 且 A_1 与 A_2 独立, 则称广义方差之比

$$\Lambda = \frac{|A_1|}{|A_1 + A_2|}$$

为威尔克斯统计量或 Λ 统计量, 其分布称为威尔克斯分布, 记为

$$\Lambda \sim \Lambda(p, n_1, n_2)$$

当 $p=1$ 时, Λ 统计量的分布正是一元统计中的参数为 $n_1/2, n_2/2$ 的 β 分布 (记为. $\beta(n_1/2, n_2/2)$)

Λ 统计量与 T^2 或 F 统计量的关系

Theorem 15.1.24 当 $n_2 = 1$ 时, 设 $n_1 = n > p$, 则

$$\Lambda(p, n, 1) \stackrel{d}{=} \frac{1}{1 + \frac{1}{n} T^2(p, n)}$$

或

$$\begin{aligned} T^2(p, n) &= n \cdot \frac{1 - \Lambda(p, n, 1)}{\Lambda(p, n, 1)} \\ \frac{n-p+1}{np} T^2 &= \frac{n-p+1}{p} \frac{1 - \Lambda}{\Lambda} = F(p, n-p+1) \end{aligned}$$

Theorem 15.1.25 当 $n_2 = 2$ 时, 设 $n_1 = n > p$, 则

$$\frac{n-p+1}{p} \frac{1 - \sqrt{\Lambda(p, n, 2)}}{\sqrt{\Lambda(p, n, 2)}} \stackrel{d}{=} F(2p, 2(n-p+1))$$

Theorem 15.1.26 当 $p=1$ 时, 则

$$\frac{n_1}{n_2} \frac{1 - \Lambda(1, n_1, n_2)}{\Lambda(1, n_1, n_2)} \stackrel{d}{=} F(n_2, n_1)$$

利用 $\Lambda(1, n_1, n_2)$ 就是 $\beta(n_1/2, n_2/2)$, 以及 β 分布与 F 分布的关系即得此结论。

Theorem 15.1.27 当 $p = 2$ 时, 则

$$\frac{n_1 - 1}{n_2} \cdot \frac{1 - \sqrt{\Lambda(2, n_1, n_2)}}{\sqrt{\Lambda(2, n_1, n_2)}} = F(2n_2, 2(n_1 - 1))$$

Theorem 15.1.28 当 $n_2 > 2, p > 2$ 时, 可用 χ^2 统计量或 F 统计量近似. 博克斯 (Box)(1949) 给出以下结论: 设 $\Lambda \sim \Lambda(p, n_1, n_2)$, 则当 $n \rightarrow \infty$ 时

$$-r \ln \Lambda \sim \chi^2(pn_2)$$

$r = n_1 - \frac{1}{2}(p - n_2 + 1)$ 当 n 不太大时也有一些近似分布。

Theorem 15.1.29 若 $\Lambda \sim \Lambda(p, n_1, n_2)$, 则存在 $B_k \sim \beta\left(\frac{n_1-p+k}{2}, \frac{n_2}{2}\right)$ ($k = 1, \dots, p$) 相互独立, 使得

$$\Lambda \stackrel{d}{=} B_1 B_2 \cdots B_p$$

Theorem 15.1.30 若 $n_2 < p$, 则

$$\Lambda(p, n_1, n_2) \stackrel{d}{=} \Lambda(n_2, p, n_1 + n_2 - p)$$

结论 2 是一元统计中 $F(n, m) = \frac{1}{F(m, n)}$ 的推广.

15.1.3 单总体均值向量的检验及置信域

Theorem 15.1.31 — 均值向量的检验. 设总体 $X \sim N_p(\mu, \Sigma)$, 随机样本 $X_{(a)} (\alpha = 1, \dots, n)$. 检验

$$H_0: \mu = \mu_0 \quad (\mu_0 \text{ 为已知向量}), \quad H_1: \mu \neq \mu_0$$

1. 当 $\Sigma = \Sigma_0$ 已知时均值向量的检验因

$$\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma_0\right), \quad \sqrt{n}(\bar{X} - \mu) \sim N_p(0, \Sigma_0)$$

利用二次型分布的结论, 知

$$(\bar{X} - \mu)' \left(\frac{1}{n}\Sigma_0\right)^{-1} (\bar{X} - \mu) \sim \chi^2(p)$$

取检验统计量为

$$T_0^2 = n(\bar{X} - \mu_0)' \Sigma_0^{-1} (\bar{X} - \mu_0) \stackrel{H_0 \text{ 下}}{\sim} \chi^2(p)$$

按照传统的检验方法, 对给定的显著性水平 α , 查 χ^2 分布临界值表得 λ_a , 使 $P\{T_0^2 > \lambda_a\} = \alpha$, 则否定域为 $\{T_0^2 > \lambda_a\}$ 由样本值 $x_{(a)} (\alpha = 1, \dots, n)$, 计算 \bar{X} 及 T_0^2 值, 若 $T_0^2 > \lambda_a$, 则否定 H_0 , 否则 H_0 相容.

假设在 H_0 成立情况下, 随机变量 $T_0^2 \sim \chi^2(p)$, 由样本值计算得到 T_0^2 的值为 d , 同时可以计算以下概率值:

$$p = P\{T_0^2 \geq d\}$$

常称此概率值为显著性概率值, 或简称为 p 值。对给定的显著性水平 α , 当 $p < \alpha$ 时, 则在显著性水平 α 下否定假设 H_0 ; 在这种情况下, 可能犯“以真当假”的第一类错误, 且 α 就是犯第一类错误的概率。当 $p \geq \alpha$ 时, 则在显著性水平 α 下 H_0 相容; 在这种情况下, 可能犯“以假当真”的第二类错误, 且犯第二类错误的概率 β 为

$$\beta = P\{T_0^2 \leq \lambda_a \mid \text{当 } \mu = \mu_1 \neq \mu_0\}$$

其中检验统计量 $T_0^2 \sim \chi^2(p, \delta)$, 非中心参数

$$\delta = n(\mu_1 - \mu_0)' \Sigma_0^{-1} (\mu_1 - \mu_0)$$

p 值的直观含义可这样看, 检验统计量 T_0^2 的大小反映 \bar{X} 与 μ_0 的偏差大小, 当 H_0 成立时 T_0^2 值应较小。现由观测数据计算 T_0^2 值并 d ; 当 H_0 成立时统计量 $T_0^2 \sim \chi^2(p)$, 由 χ^2 分布可计算该统计量 $\geq d$ 的概率值(即 p 值)。比如 $p = 0.02 < \alpha = 0.05$, 这时出现一个比小概率标准 ($\alpha = 0.05$) 还要小的事件 $\{T_0^2 \geq d\}$ 。也就是说, 在 $\mu = \mu_0$ 假设下, 观察数据中极少情况会出现 T_0^2 的值大于等于 d 值, 故在 0.05 显著性水平下有足够的证据否定原假设, 即认为 μ 与 μ_0 有显著地差异。又比如当 $p = 0.22 \geq \alpha = 0.05$ 时, 表示在 $\mu = \mu_0$ 的假设下, 观测数据中经常会出现 T_0^2 的值大于等于 d 值的情况, 故在 0.05 显著性水平下没有足够的证据否定原假设, 即认为 μ 与 μ_0 没有显著地差异。

Theorem 15.1.32 — 当 Σ 未知时均值向量的检验. 当 $p = 1$ 时(一元统计), 取检验统计量为

$$t = \frac{(\bar{X} - \mu_0) \sqrt{n}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_{(i)} - \bar{X})^2}} \sim t(n-1)$$

或等价地取检验统计量

$$t^2 = n(\bar{X} - \mu_0)' \left(\frac{1}{n-1} \sum_{i=1}^n (X_{(i)} - \bar{X})^2 \right)^{-1} (\bar{X} - \mu_0)$$

推广到多元, 考虑统计量

$$T^2 = n(\bar{X} - \mu_0)' \left(\frac{1}{n-1} A \right)^{-1} (\bar{X} - \mu_0)$$

因而

$$\bar{X} \stackrel{H_0 \text{T}}{\sim} N_p \left(\mu_0, \frac{1}{n} \Sigma \right), \quad \sqrt{n}(\bar{X} - \mu_0) \stackrel{H_0 \text{T}}{\sim} N_p(0, \Sigma)$$

样本离差阵并

$$A = \sum_{a=1}^n (X_{(a)} - \bar{X})(X_{(a)} - \bar{X})' \sim W_p(n-1, \Sigma)$$

由定义可知

$$\begin{aligned} T^2 &= (n-1) \cdot [\sqrt{n}(\bar{X} - \mu_0)]' A^{-1} [\sqrt{n}(\bar{X} - \mu_0)] \\ &= (n-1)n(\bar{X} - \mu_0)' A^{-1} (\bar{X} - \mu_0) \sim T^2(p, n-1) \end{aligned}$$

再利用 T^2 与 F 分布的关系, 检验统计量取并

$$\begin{aligned} F &= \frac{(n-1)-p+1}{(n-1)p} T^2 \stackrel{H_0 \text{ 下}}{\sim} F(p, (n-1)-p+1) \\ &\stackrel{H_0 \text{ F}}{\sim} F(p, n-p) \end{aligned}$$

Theorem 15.1.33 — 最大似然比原理. 设 p 元总体的密度函数为 $f(x, \theta)$, 其中 θ 是未知参数, 且 $\theta \in \Theta$ (参数空间), 又设 Θ_0 是 Θ 的子集, 我们希望对下列假设:

$$H_0: \theta \in \Theta_0, \quad H_1: \theta \notin \Theta_0$$

作出判断, 这就是假设检验问题. 称 H_0 为原假设 (或零假设), H_1 为对立假设 (或备择假设). 从总体 X 抽取容量为 n 的样本 $X_{(t)} (t = 1, \dots, n)$. 把样本的联合密度函数

$$L(x_{(1)}, \dots, x_{(n)}; \theta) = \prod_{t=1}^n f(x_{(t)}; \theta)$$

记为 $L(X; \theta)$, 并称它为样本的似然函数. 引入统计量

$$\lambda = \max_{\theta \in \Theta_0} L(X; \theta) / \max_{\theta \in \Theta} L(X; \theta)$$

它是样本 $X_{(t)} (t = 1, \dots, n)$ 的函数, 常称 λ 为似然比统计量. 由于 $\Theta_0 \subset \Theta$, 从而 $0 \leq \lambda \leq 1$ 由最大似然比原理知, 如果 λ 取值太小, 说明 H_0 为真时观测到此样本 $X_{(t)} (t = 1, \dots, n)$ 的概率比 H_0 为不真时观测到此样本 $X_{(t)} (t = 1, \dots, n)$ 的概率要小得多. 故有理由认为假设 H_0 不成立, 所以从似然比出发, 以上检验问题的否定域为

$$\{\lambda(X_{(1)}, \dots, X_{(n)}) < \lambda_a\}$$

按传统的检验方法, λ_a 是由显著性水平 α 确定的临界值, 它满足当 H_0 成立时使得:

$$P\{\lambda(X_{(1)}, \dots, X_{(n)}) < \lambda_a\} = \alpha$$

为了得到 λ_a , 必须研究似然比统计量 λ 的抽样分布. 在一些特殊的情况下, 可以得到 λ 的精确分布; 但在很多情况下是得不到 λ 的精确分布的. 当样本量很大且满足一定正则条件时, $-2 \ln \lambda$ 的抽样分布与 χ^2 分布十分接近.

Theorem 15.1.34 当样本容量 n 很大时,

$$-2 \ln \lambda = -2 \ln \left[\left(\max_{\theta \in \Theta_0} L(X; \theta) / \max_{\theta \in \Theta} L(X; \theta) \right) \right]$$

近似服从自由度为 f 的 χ^2 分布, 其中 $f = \Theta$ 的维数 $-\Theta_0$ 的维数.

Theorem 15.1.35 当 Σ 未知时检验均值向量 $\mu = \mu_0$ 的似然比统计量, 并讨论它的分布.

设样本的似然函数为 $L(\mu, \Sigma)$. 检验均值向量 $\mu = \mu_0$ 的似然比统计量

$$\lambda = \max_{\mu=\mu_0, \Sigma>0} L(\mu_0, \Sigma) / \max_{\mu, \Sigma>0} L(\mu, \Sigma)$$

上面比式的分母当 $\mu = \bar{X}, \Sigma = \frac{1}{n}A$ 时达最大值, 且最大值并

$$\max_{\mu, \Sigma>0} L(\mu, \Sigma) = (2\pi)^{-np/2} \left| \frac{1}{n}A \right|^{-n/2} e^{-np/2}$$

上面比式的分子当

$$\Sigma = \frac{1}{n} \sum_{i=1}^n (X_{(i)} - \mu_0) (X_{(i)} - \mu_0)' = \frac{1}{n} A_0$$

时达最大值, 经过一些推导, 得到

$$\frac{|A|}{|A_0|} = \frac{1}{1 + n(\bar{X} - \mu_0)' A^{-1} (\bar{X} - \mu_0)} = \frac{1}{1 + \frac{1}{n-1} T^2}$$

其中

$$T^2 = (n-1)n(\bar{X} - \mu_0)' A^{-1} (\bar{X} - \mu_0) \stackrel{H_0}{\sim} T^2(p, n-1)$$

否定域:

$$\{\lambda < \lambda_\alpha\} \Leftrightarrow \{T^2 > T_a^2\} \Leftrightarrow \{F > F_a\}$$

$$\text{其中 } F = \frac{n-p}{p} \frac{T^2}{n-1} \stackrel{H_0}{\sim} F(p, n-p)$$

Theorem 15.1.36 三、置信域与联立置信区间在一元统计中, 讨论均值的假设检验问题本质上也等价于求均值的置信区间. 下面就单个多维正态总体均值向量的置信域的概念作为一元统计中置信区间的推广给出简单介绍. 1. 置信域假设 $X_{(t)} (t = 1, 2, \dots, n)$ 来自 p 元正态总体 $N_p(\mu, \Sigma)$ (Σ 未知), 由前面的讨论可知

$$T^2 = n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \sim T^2(p, n-1)$$

或者

$$F = \frac{n-p}{(n-1)p} T^2 \sim F(p, n-p)$$

任给置信度 $1 - \alpha$, 查 F 分布临界值表得 F_α 满足

$$P\{F \leq F_\alpha\} = 1 - \alpha$$

则均值向量 μ 的置信度为 $1 - \alpha$ 的置信域为

$$T^2 = n(X - \mu)' S^{-1} (\bar{X} - \mu) \leq \frac{(n-1)p}{n-p} F_\alpha$$

该置信域是一个中心在 \bar{X} 的椭球。当检验假设 $H_0 : \mu = \mu_0$ 时, 若 μ_0 落入上述置信域内, 即

$$T^2 = n(\bar{X} - \mu_0)' S^{-1} (\bar{X} - \mu_0) \leq \frac{(n-1)p}{n-p} F_a$$

则在显著性水平 α 下, H_0 相容; 若 μ_0 没有落入上述置信域内, 则否定 H_0 . 可见在多元统计中, 讨论均值向量的假设检验问题本质上也等价于求均值向量的置信域。

Theorem 15.1.37 对任意的 a , 考虑 $a'\mu$ 的置信区间便能够得到所要的联立置信区间. 置信区间:

$$a'\bar{X} - t_{a/2} \frac{\sqrt{a'Sa}}{\sqrt{n}} \leq a'\mu \leq a'\bar{X} + t_{a/2} \frac{\sqrt{a'Sa}}{\sqrt{n}} \quad (15.1)$$

其中 $t_{a/2}$ 满足: $P\{|t| \leq t_{a/2}\} = 1 - \alpha$ (这里 $t \sim t(n-1)$) 显然通过选择不同的系数向量 a , 便可得到 μ 的若干个线性组合的置信度为 $1 - \alpha$ 的置信区间; 但请注意, 这时总的置信度不再是 $1 - \alpha$, 而比 $1 - \alpha$ 低. 下面给出构造所有 $a'\mu$ 的联立置信区间估计的 Scheffe 方法。对给定的样本 $X_{(t)} (t = 1, 2, \dots, n)$ 和系数向量 a , 若全体 $a'\mu$ 值的置信区间是由 16.25 式给出的, 则不等式

$$t^2 = \frac{n[a'(\bar{X} - \mu)]^2}{a'Sa} \leq t_{a/2}^2$$

成立. 若让 a 变化, 求所有 $a'\mu$ 的联立置信区间, 那么应将 16.25 式的右边换上更大的常数才较为合理. 为此来求最大值

$$\begin{aligned} \max_{a \neq 0} t^2 &= \max_{a \neq 0} \frac{n[a'(\bar{X} - \mu)]^2}{a'Sa} \\ \max_{a \neq 0} \frac{n[a'(\bar{X} - \mu)]^2}{a'Sa} &= n(\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) = T^2 \end{aligned}$$

且最大值在 a 与 $S^{-1}(\bar{X} - \mu)$ 成比例时达到.

Theorem 15.1.38 假设 $X_{(t)} (t = 1, 2, \dots, n)$ 为来自 p 元正态总体 $N_p(\mu, \Sigma) (\Sigma > 0$ 未知) 的随机样本, 则对所有的 a , 区间

$$[a'\bar{X} - d, a'\bar{X} + d] \quad (\text{其中 } d = \sqrt{\frac{(n-1)p}{n(n-p)} F_\alpha a'Sa})$$

包含 $a'\mu$ 的概率为 $1 - \alpha$ (其中 F_α 满足

$$P\{F \leq F_\alpha\} = 1 - \alpha$$

式).

Theorem 15.1.39 由于置信概率由 T^2 分布确定, 称给出的联立置信区间为 T^2 区间. 在 T^2 区间中, 若取 $a = e_i = (0, \dots, 1, \dots, 0)'$, 我们便同时得到 $\mu_i (i = 1, \dots, p)$ 的置信度均为

$1 - \alpha$ 的 T^2 区间

$$\bar{x}_i - c\sqrt{\frac{s_{ii}}{n}} \leq \mu_i \leq \bar{x}_i + c\sqrt{\frac{s_{ii}}{n}} \quad (\text{其中 } c = \sqrt{\frac{(n-1)p}{(n-p)} F_\alpha})$$

其中 s_{ii} 为样本协方差阵 S 的第 i 个对角元素. 请注意: 如果在 16.25 式中取 $a = e_i (i = 1, \dots, p)$, 即每次考虑一个分量的置信区间, 则得到单个 $\mu_i (i = 1, 2, \dots, p)$ 的置信度为 $1 - \alpha$ 的置信区间若把这 p 个区间合在一起构成 $\mu_i (i = 1, 2, \dots, p)$ 的联立置信区间, 其置信度比 $1 - \alpha$ 低. 请读者仔细比较在统计意义上的差别.

多个正态总体均值向量的检验---多元方差分析

Theorem 15.1.40 — 元方差分析. $A = \sum_{i=1}^k A_i$ 称为组内离差阵

$$B = \sum_{i=1}^k n_i (\bar{X}^{(i)} - \bar{X}) (\bar{X}^{(i)} - \bar{X})'$$

称为组间离差阵. 根据直观想法及用似然比原理得到检验 H_0 的统计量为

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|A|}{|T|}$$

易见:

1. 因 $A_i \sim W_p(n_i - 1, \Sigma)$ 且相互独立 ($i = 1, \dots, k$), 由可加性可得

$$A = \sum_{i=1}^k A_i \sim W_p(n - k, \Sigma) \quad (n = n_1 + \dots + n_k)$$

2. 在 H_0 下, $T \sim W_p(n - 1, \Sigma)$
3. 还可以证明在 H_0 下, $B \sim W_p(k - 1, \Sigma)$, 且 B 与 A 相互独立.

根据 Λ 分布的定义, 可知

$$\Lambda = \frac{|A|}{|A+B|} \stackrel{H_0}{\sim} \Lambda(p, n - k, k - 1)$$

给定显著性水平 α , 查威尔克斯分布临界值表, 可得 λ_α , 使

$$P\{\Lambda < \lambda_\alpha\} = \alpha$$

故否定域 $W = \{\Lambda < \lambda_\alpha\}$. 当手头没有威尔克斯临界值表时, 可用 χ^2 分布或 F 分布来近似, 即由 Λ 的函数的近似分布进行检验

Theorem 15.1.41 — 独立性检验. 设总体 $X \sim N_p(\mu, \Sigma)$, 将 X 剖分为 k 个子向量, 而 μ 和 Σ 也相应剖分为

$$X = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(k)} \end{bmatrix}_{p_k}^{p_1}, \mu = \begin{bmatrix} \mu^{(1)} \\ \vdots \\ \mu^{(k)} \end{bmatrix}_{p_k}^{p_1}, \Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} \\ \vdots & & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} \end{bmatrix}_{p_k}^{p_1}$$

其中 $p = p_1 + \dots + p_k$, 且知 p_t 维子向量 $X^{(t)} \sim N_{p_t}(\mu^{(t)}, \Sigma_{tt}) (t = 1, \dots, k)$. 若 k 个随机子

向量相互独立，则可把 p 维（高维）随机向量的问题化为 k 个低维随机向量的问题来处理，这在处理多元统计分析的许多问题中将带来极大的方便。

在第二章中，我们已介绍过若 $X^{(1)}, \dots, X^{(k)}$ 相互独立 $\iff \Sigma_{ij} = O$ (对一切 $i \neq j$)。因此检验 $X^{(1)}, \dots, X^{(k)}$ 是否相互独立的问题等价于检验对任意两个子向量，协方差阵 Σ_{ij} 是否等于 O (对一切 $i \neq j$) 在正态总体下，独立性检验可化为检验：

$$H_0: \Sigma_{ij} = O \text{ (一切 } i \neq j), \quad H_1: \Sigma_{ij} \neq 0, \text{ 至少有一对 } i \neq j$$

设 $X_{(a)} (\alpha = 1, \dots, n, n > p)$ 为来自总体 X 的随机样本。将 $X_{(a)}$ 样本均值向量 \bar{X} 和样本离差阵

$$A = \sum_{j=1}^n (X_{(j)} - \bar{X}) (X_{(j)} - \bar{X})'$$

得到似然比统计量

$$\lambda = \prod_{i=1}^k \left| \frac{A_{ii}}{n} \right|^{-n/2} / \left| \frac{1}{n} A \right|^{-n/2} = \left[\frac{|A|}{\prod_{i=1}^k |A_{ii}|} \right]^{n/2} \stackrel{\text{def}}{=} V^{n/2}$$

$$\ln \lambda = \frac{n}{2} \ln V$$

博克斯 (Box) 证明了，在 H_0 成立下当 $n \rightarrow \infty$ 时

$$-b \ln V \sim \chi^2(f)$$

其中

$$b = n - \frac{3}{2} - \frac{p^3 - \sum_a^k p_a^3}{3(p^2 - \sum_a^k p_a^2)} \\ f = \frac{1}{2} [p(p+1) - \sum_{a=1}^k p_a (p_a + 1)]$$

Theorem 15.1.42 设 $X_{(a)} = (X_{a1}, \dots, X_{ap})'$ ($\alpha = 1, \dots, n$) 是来自 p 元总体 X 的随机样本，试问总体 X 是否服从 $N_p(\mu, \Sigma)$ 分布？

若总体 $X = (X_1, \dots, X_p)' \sim N_p(\mu, \Sigma)$ ，利用多元正态分布的一些性质可知以下结论（记 $\mu = (\mu_1, \dots, \mu_p)', \Sigma = (\sigma_{ij})_{p \times p}$ ）

1. 每个分量 $X_i \sim N(\mu_i, \sigma_{ii})$ ($i = 1, \dots, p$)
2. 任意两个分量 (X_i, X_j) 为二元正态分布
3. 设 $l = (l_1, \dots, l_p)'$ 为任给的 p 维常向量，令 $\xi = l'X$ ，则

$$\xi \sim N_1(l'\mu, l'\Sigma l)$$

4. 令 $\eta = (X - \mu)'\Sigma^{-1}(X - \mu)$ ，则 $\eta \sim \chi^2(p)$
5. 正态随机向量 X 的概率密度等高线为椭球。

一维边缘分布的正态性检验

Theorem 15.1.43 设 p 维随机向量 $X = (X_1, \dots, X_p)'$ ，检验分量 $X_i \sim N(\mu_i, \sigma_i^2)$ ($i = 1, \dots, p$)。若要把 p 元正态性检验化为 p 个一元数据的正态性检验，常用的检验方法有以下

1. χ^2 检验法：这是适用于连续型或离散型随机变量分布的拟合优度检验方法，也称

为皮尔逊 (Pearson) χ^2 检验法.

2. 科尔莫戈罗夫 (Kolmogorov) 检验法: 这是适用于连续型分布的拟合优度检验方法, 当然也适用于正态性检验。以下几种方法是仅适用于正态分布的检验法。
3. 偏峰检验法。
4. W(Wilks) 检验和 D 检验。
5. Q-Q(Quantile-Quantile) 图检验法。
6. P-P(Probability-Probability) 图检验法: 这是与 Q-Q 图检验法类似的图示检验法. Q-Q 图检验法绘制散点 $(q_i, x_{(i)}^*)$ 的散布图, 其中 $q_i = \Phi^{-1}(p_i)$ 为正态总体的 p_i 分位数

$$p_i = \frac{i - 0.5}{n} \quad (i = 1, \dots, n)$$

$x_{(i)}^*$ 为样本的 p_i 分位数. 如果总体 X 为一元正态总体, 这些散点应散布在一条直线上。另外, 我们还可以绘制另一对数据点: $(p_i, F(x_{(i)}^*))$ ($i = 1, \dots, n$) 的散布图, 记 $F_n(\cdot)$ 为经验分布函数. 因为

$$p_i = F_n(x_{(i)}^*) \approx F(x_{(i)}^*) = \Phi\left(\frac{x_{(i)}^* - \mu}{\sigma}\right)$$

故这些散点也应散布在直线上. 这里 $F(x_{(i)}^*)$ 是正态总体 $X(X \sim N(\mu, \sigma^2))$ 的分布函数在点 $x_{(i)}^*$ 上的值, 即随机事件 $\{X \leq x_{(i)}^*\}$ 的概率 (Probability) 值, 而 p_i 是由经验分布函数 $F_n(x)$ 得到的样本分布函数在点 $x_{(i)}^*$ 上的值, 故称此散布图为 P-P 图, 利用此图得到的检验法称为 P-P 图检验法。

7. “ 3σ ”原则检验法: 如果总体 $X \sim N(\mu, \sigma^2)$, 根据“ 3σ ”原则可知

$$p_k = P\{\mu - k\sigma < X < \mu + k\sigma\} = \begin{cases} 0.683, & \text{当 } k = 1 \text{ 时,} \\ 0.954, & \text{当 } k = 2 \text{ 时,} \\ 0.997, & \text{当 } k = 3 \text{ 时.} \end{cases}$$

$X_{(1)}, X_{(2)}, \dots, X_{(n)}$ 是来自总体 X 的样本 (假设样本容量 n 很大) 又 X 为正态分布, 则样品点落入区域 $(\mu - k\sigma, \mu + k\sigma)$ 的比例 \hat{p}_k 与以上列出的概率 p_k 应是相差不多的. 利用大数定律可知, \hat{p}_k 近似为正态分布.

二元数据的正态性检验

Theorem 15.1.44 — 等概椭圆检验法. 设 $X = (X_1, X_2, \dots, X_p)'$ 为 p 维随机向量, X 的任意两个分量的 n 次观测数据记为 $X_{(i)} = (X_{i1}, X_{i2})'$ ($i = 1, \dots, n$). 下面介绍检验二元观测数据是否来自二元正态分布的方法--等概椭圆检验法: 若二维随机向量 $X = (X_1, X_2)'$ $\sim N_2(\mu, \Sigma)$, 则 X 的概率密度函数等高线

$$f(x_1, x_2) = a \Leftrightarrow (X - \mu)' \Sigma^{-1} (X - \mu) = b^2$$

上式右边是中心在 (μ_1, μ_2) 由 $(X - \mu)' \Sigma^{-1} (X - \mu) = b^2$ 决定的椭圆. 由本章 §3.1 中所介绍的知识可知

$$D^2 = (X - \mu)' \Sigma^{-1} (X - \mu) \sim \chi^2(2)$$

对给定 $p_0 \in (0, 1)$, 则存在 d_0 , 使

$$P\{D^2 \leq d_0\} = p_0$$

■ **Example 15.4** 比如取 $p_0 = 1/2$, 由统计软件计算或者查有关临界值表, 可得 $d_0 = 1.386$; 当 $p_0 = 0.25$ 时 $d_0 = 0.575$; $p_0 = 0.75$ 时 $d_0 = 2.773, \dots$. 而

$$P\{(X - \mu)' \Sigma^{-1} (X - \mu) \leq 1.386\} = 0.5$$

表示样品点 $X_{(i)}$ (落入由 1.386 指定的椭圆内的概率为 1/2. 利用这一结论, 可对二元观测数据的正态性进行检验

Theorem 15.1.45 — p 元数据的正态性检验. 设 $X_{(\alpha)} = (X_{\alpha 1}, \dots, X_{\alpha p})'$ ($\alpha = 1, \dots, n$) 为来自 p 元总体 X 的随机样本. 检验

$$H_0: X \sim N_p(\mu, \Sigma), \quad H_1: X \text{ 不服从 } N_p(\mu, \Sigma)$$

1. χ^2 统计量的 Q-Q 图检验法 (或 P-P 图检验法) 这是由正态分布的性质的结论 4 构造的检验法. 在 H_0 下, 将样品 X 到总体中心 μ 的马氏距离 $D^2(X, \mu)$ 记为 D^2 , 则有

$$D^2 = (X - \mu)' \Sigma^{-1} (X - \mu) \sim \chi^2(p)$$

以下构造的检验方法就是检验统计量 D^2 是否有 $D^2 \sim \chi^2(p)$ 成立. 直观的想法是: 由样品 $X_{(\alpha)}$ 计算 $D_{\alpha}^2 (\alpha = 1, \dots, n)$, 对 D_{α}^2 排序:

$$D_{(1)}^2 \leq D_{(2)}^2 \leq \dots \leq D_{(n)}^2$$

统计量 D^2 的经验分布函数取为

$$F_n(D_{(t)}^2) = \frac{t - 0.5}{n} \stackrel{\text{def}}{=} p_t \approx H(D_{(t)}^2 | p)$$

其中 $H(D_{(t)}^2 | p)$ 表示 $\chi^2(p)$ 的分布函数在 $D_{(t)}^2$ 的值. 设 χ^2 分布的 p_t 分位数为 χ_t^2 , 显然 χ_t^2 满足: $H(\chi_t^2 | p) = p_t$, 即 χ^2 分布的 p_t 分位数 $\chi_t^2 = H^{-1}(p_t | p)$. 又由经验分布得到样本的 p_t 分位数 $D_{(t)}^2 = F_n^{-1}(p_t)$. 若 $H(x | p) \approx F_n(x)$, 应有

$$D_{(t)}^2 \approx \chi_t^2$$

绘制点 $(D_{(t)}^2, \chi_t^2)$ 的散布图, 当 X 为正态总体时, 这些点应散布在一条直线上. 这种检验法其实就是 χ^2 分布的 Q-Q 图检验法. 类似地, 也可以绘制点 $(p_t, H(D_{(t)}^2 | p))$ 的散布图, 当 X 为正态总体时, 这些点也应散布在一条直线上. 这种检验法其实就是 χ^2 分布 (有时表示为“卡方分布”) 的 P-P 图检验法. 具体检验步骤如下:

1. 由 n 个 p 样品点 $X_{(\alpha)}$ ($\alpha = 1, \dots, n$) 计算样本均值 \bar{X} 和样本协方差阵 S :

$$S = \frac{1}{n-1} \sum_{\alpha=1}^n (X_{(\alpha)} - \bar{X})(X_{(\alpha)} - \bar{X})'$$

2. 计算样品点 $X_{(t)}$ 到 \bar{X} 的马氏距离

$$D_t^2 = (X_{(t)} - \bar{X})' S^{-1} (X_{(t)} - \bar{X}) \quad (t = 1, \dots, n)$$

3. 对马氏距离 D_t^2 按从小到大的次序排序：

$$D_{(1)}^2 \leq D_{(2)}^2 \leq \cdots \leq D_{(n)}^2$$

4. 计算 $p_t = \frac{t-0.5}{n}, n = 1, 2, \dots, n$, 及 χ_t^2 , 其满足

$$H(\chi_t^2 | p) = p_t \quad \left(\text{或计算 } H(D_{(t)}^2 | p) \text{ 的值} \right)$$

5. 以马氏距离为横坐标, χ^2 分位数为纵坐标作平面坐标系, 用 n 个点 $(D_{(t)}^2, \chi_t^2)$ 绘制散布图, 即得到 χ^2 分布的 Q-Q 图; 或者用另 n 个点 $(p_t, H(D_{(t)}^2 | p))$ 绘制散布图, 即得 χ^2 分布的 P-P 图. 个点
6. 考察这 n 个点是否散布在一条通过原点, 斜率为 1 的直线上, 若是, 接受数据来自 p 元正态总体的假设; 否则拒绝正态性假设。

Theorem 15.1.46 — 主成分检验法. 设 $X_{(i)} = (X_{i1}, X_{i2}, \dots, X_{ip})'$ ($i = 1, \dots, n$) 为来自 p 元总体 $X = (X_1, \dots, X_p)'$ 的观测数据(样本), 检验 $H_0: X \sim N_p(\mu, \Sigma)$, $H_1: X$ 不服从 $N_p(\mu, \Sigma)$ 设样本协方差阵 S 的特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, 相应的特征向量为 l_1, l_2, \dots, l_p , 记 $l_t = (l_{1t}, l_{2t}, \dots, l_{pt})'$. 令

$$Z_t = l_{1t}X_1 + l_{2t}X_2 + \cdots + l_{pt}X_p \quad (t = 1, 2, \dots, p)$$

即新变量 Z_1, \dots, Z_p 是 X_1, \dots, X_p 的线性组合. 则可以证明: Z_1, \dots, Z_p 是相互独立的; p 元观测数据提供的信息大部分可由前几个新变量所提供. 这时 p 元数据的正态性检验可化为几个相互独立的新变量的一元数据的正态性检验. 这些新变量在第七章中被称为主成分。故检验法称为主成分检验法. 如果正态性假设不能成立, 一般应考虑对数据进行变换, 使非正态更接近正态, 然后对变换后的数据进行统计分析.

16. 线性模型

矩阵论的相关性质

- Proposition 16.0.1**
1. 若将 a_1, a_2, \dots, a_k 排成 $n \times k$ 矩阵 $A = (a_1, a_2, \dots, a_k)$, 则 S_0 可表为 $S_0 = \{x = At, t \in R_k\}$, 它是 A 的列向量张成的子空间, 记为 $S_0 = \mathcal{M}(A)$.
 2. R_n 的任一子空间都是某一矩阵的列向量张成的子空间.
 3. $\dim \mathcal{M}(A) = \text{rk}(A)$
 4. $\mathcal{M}(A) \subset \mathcal{M}(A : B)$, 特别若 $b_j, j = 1, 2, \dots, l$ 可表为 a_1, a_2, \dots, a_k 的线性组合, 则 $\mathcal{M}(A) = \mathcal{M}(A : B)$
 5. 设 A 为 $n \times k$ 矩阵, 记 A^\perp 为满足条件 $A'A^\perp = 0$ (inner product) 且具有最大秩的矩阵, 则

$$\mathcal{M}\left(A^\perp\right) = \mathcal{M}(A)^\perp$$

6. 对任意矩阵 A , 恒有 $\mathcal{M}(A) = \mathcal{M}(AA')$.

Proof. 显然 $\mathcal{M}(AA') \subset \mathcal{M}(A)$, 故只需证明 $\mathcal{M}(A) \subset \mathcal{M}(AA')$. 从另一个角度来证明, 即补空间。事实上, 对任给 $x \perp \mathcal{M}(AA')$, 有 $x'AA' = 0$. 右乘 x , 得 $x'AA'x = \|A'x\|^2 = 0$, 故 $A'x = 0$. 于是 $x \perp \mathcal{M}(A)$. 明所欲证. ■

7. 设 $A_{n \times m}, H_{k \times m}$, 则
 - (a) $S = \{Ax : Hx = 0\}$ 是 $\mathcal{M}(A)$ 的子空间
 - (b) $\dim(S) = \text{rk} \begin{pmatrix} A \\ H \end{pmatrix} - \text{rk}(H)$

Proof. 第一结论的证明是简单的, 由于不加条件时已经是 $\mathcal{M}(A)$ 的子空间了, 现在(1)相当于缩小了原来的空间, 当然还是子空间。现证(2). 不妨设 $\text{rk}(H) = k (k < m)$,

则存在 $m \times m$ 可逆阵 Q , 使得 $H_{k \times m} Q_{m \times m} = \begin{pmatrix} I_k & : & 0 \end{pmatrix}$. 于是

$$\begin{aligned}\dim(S) &= \dim \left\{ \begin{pmatrix} A \\ H \end{pmatrix} x : Hx = 0 \right\} = \dim \left\{ \begin{pmatrix} A_{n \times m} \\ H_{k \times m} \end{pmatrix} Qx : HQx = 0 \right\} \\ &= \dim \left\{ \begin{pmatrix} U_1^{n \times k} & U_2^{n \times m-k} \\ I_k & 0 \end{pmatrix} x : (I_k : 0)x = 0 \right\} = \dim \{U_2 x_{(2)} : x_{(2)} \text{ 任意}\} \\ &= \operatorname{rk}(U_2) = \stackrel{(i)}{\operatorname{rk}} \begin{pmatrix} U_1 & U_2 \\ I_k & 0 \end{pmatrix} - \operatorname{rk}(I_k) = \operatorname{rk} \begin{pmatrix} A \\ H \end{pmatrix} - \operatorname{rk}(H)\end{aligned}\tag{16.1}$$

(i) 是因为可以进行初等行变换, 计算得第一项的秩就是 $\operatorname{rk}(U_2) + \operatorname{rk}(I_k)$. 其中 $(U_1 : U_2) = AQ$, $x = \begin{pmatrix} x_{(1)} \\ x_{(2)} \end{pmatrix}$, $x_{(1)}$ 为 $k \times 1$ 向量, $x_{(2)}$ 为 $(m-k) \times 1$ 向量. 定理证毕 ■

R 倒数第二个等号有问题

8. 设 $\mathcal{M}(A) \cap \mathcal{M}(B) = \{0\}$, 则 $\mathcal{M}(A'B^\perp) = \mathcal{M}(A')$

Proof. 因为

$$\mathcal{M}(A'B^\perp) = \left\{ A'x, x = B^\perp t, t \text{ 任意} \right\} = \{A'x, B'x = 0\}$$

依上一条性质及假设条件, 有

$$\dim \mathcal{M}(A'B^\perp) = \operatorname{rk} \begin{pmatrix} A' \\ B' \end{pmatrix} - \operatorname{rk}(B') = \operatorname{rk}(A : B) - \operatorname{rk}(B) = \operatorname{rk}(A) = \dim(\mathcal{M}(A'))$$

(倒数第二个等号用了条件) 但

$$\mathcal{M}(A'B^\perp) \subset \mathcal{M}(A')$$

于是

$$\mathcal{M}(A'B^\perp) = \mathcal{M}(A')$$

■

16.0.1 广义逆

Definition 16.0.1 — A^- 广义逆的定义. 对矩阵 $A_{m \times n}$, 一切满足方程组

$$AXA = A$$

的矩阵 X , 称为矩阵 A 的广义逆, 记为 A^-

Theorem 16.0.2 设 A 为 $m \times n$ 矩阵, $\operatorname{rk}(A) = r$. 若

$$A = P \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q$$

■

这里 P 和 Q 分别为 $m \times m, n \times n$ 的可逆阵, 则

$$A^- = Q^{-1} \begin{pmatrix} I_r & B \\ C & D \end{pmatrix} P^{-1}$$

这里 B, C 和 D 为适当阶数的任意矩阵.(由于 B, C, D 可以是任意的, 所以广义逆不唯一)

Proof. 设 X 为 A 的广义逆, 则有

$$\begin{aligned} AXA = A &\iff P \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} QXP \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q = P \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q \\ &\iff \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} QXP \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

若记

$$QXP = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

则上式

$$\iff \begin{pmatrix} B_{11} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \iff B_{11} = I_r$$

于是, $AXA = A \iff X = Q^{-1} \begin{pmatrix} I_r & B_{12} \\ B_{21} & B_{22} \end{pmatrix} P^{-1}$, 其中 B_{12}, B_{21} 和 B_{22} 任意. 证毕. ■

- Corollary 16.0.3**
1. 对任意矩阵 A , A^- 总是存在的(任何一个矩阵都有秩, 那么通过初等变换就可以换成于单位矩阵相似);
 2. A^- 唯一 $\iff A$ 为可逆方阵. 此时 $A^- = A^{-1}$
 3. $\text{rk}(A^-) \geq \text{rk}(A) = \text{rk}(A^-A) = \text{rk}(AA^-)$
 4. 若 $\mathcal{M}(B) \subset \mathcal{M}(A), \mathcal{M}(C) \subset \mathcal{M}(A')$, 则 $C'A^-B$ 与 A^- 的选择无关.

Proof. 前三条结论不难从定理 16.0.2 及广义逆的定义得到。第四条只要注意到, 假设条件 $\mathcal{M}(B) \subset \mathcal{M}(A), \mathcal{M}(C) \subset \mathcal{M}(A')$ 蕴涵着, 存在矩阵 T_1, T_2 使得 $B = AT_1, C = A'T_2$, 就可证明所要结论. 证毕. ■

- Corollary 16.0.4** 对任一矩阵 A ,

1. $A(A'A)^-A'$ 与广义逆 $(A'A)^-$ 的选择无关;
2. $A(A'A)^-A'A = A, A'A(A'A)^-A' = A'$

Proof. 1. 由之前的性质知 $\mathcal{M}(A') = \mathcal{M}(A'A)$, 故存在矩阵 B , 使得 $A' = A'AB$. 于是, $A(A'A)^-A' = B'A'A(A'A)^-A'AB = B'A'AB$, 与 $(A'A)^-$ 无关.
2. 记 $F = A(A'A)^-A'A - A$, 利用广义逆的定义, 可以验证: $F'F = 0$. 于是 $F = 0$. 第一式得证. 同法. ■

Theorem 16.0.5 设 $Ax = b$ 为一相容方程组, 则

1. 对任一广义逆 A^- , $x = A^-b$ 必为解--广义逆一定是方程组的解;
2. 齐次方程组 $Ax = 0$ 的通解为 $x = (I - A^-A)z$, 这里 z 为任意的向量, A^- 为任意固定的一个广义逆;

3. $Ax = b$ 的通解为

$$x = A^{-}b + (I - A^{-}A)z \quad (16.2)$$

其中 A^{-} 为任一固定的广义逆, z 为任意向量.

Proof. 1. 由相容性假设知, 存在 x_0 , 使 $Ax_0 = b$. 故对任一 $A^{-}, A(A^{-}b) = AA^{-}Ax_0 = Ax_0 = b$. 即 $A^{-}b$ 为解.

2. 设 x_0 为 $Ax = 0$ 的任一解, 即 $Ax_0 = 0$, 那么

$$x_0 = (I - A^{-}A)x_0 + A^{-}Ax_0 = (I - A^{-}A)x_0$$

即任一解都取 $(I - A^{-}A)z$ 的形式. 反过来, 对任一的 z , 因 $A(I - A^{-}A)z = (A - AA^{-}A)z = 0$, 故 $(I - A^{-}A)z$ 必为解.

3. 任取定一个广义逆 A^{-} , 由(1)知 $x_1 = A^{-}b$ 为方程组 $Ax = b$ 的一个特解. 由(2)知 $x_2 = (I - A^{-}A)z$ 为齐次方程组 $Ax = 0$ 的通解. 依非齐次线性方程组的解结构定理知, $x_1 + x_2$ 为 $Ax = b$ 的通解. 证毕. ■

Theorem 16.0.6 设 $Ax = b$ 为相容线性方程组, 且 $b \neq 0$, 那么, 当 A^{-} 取遍 A 的所有的广义逆时, $x = A^{-}b$ 构成了该方程组的全部解.

Proof. 证明由两部分组成. 其一, 要证对每一个 $A^{-}, x = A^{-}b$ 为 $Ax = b$ 的解, 这已在前一定理中证明过了. 其二, 要证对 $Ax = b$ 的任一解 x_0 , 必有在一个 A^{-} , 使 $x_0 = A^{-}b$. 由方程组解的结构(16.2)知, 存在 A 的一个广义逆 G 及 z_0 , 使得

$$x_0 = Gb + (I - GA)z_0$$

因 $b \neq 0$, 故总存在矩阵 U , 使得 $z_0 = Ub$. 例如, 可取 $U = z_0(b'b)^{-1}b'$. 于是

$$x_0 = Gb + (I - GA)Ub = (G + (I - GA)U)b \triangleq Hb$$

其中 $H = G + (I - GA)U$. 易验证 H 为一个 A^{-} . 定理得证. ■



定理 16.0.5 的(3)和定理 16.0.6 给出了相容线性方程组解集的两种表示. 在(16.2)中, A^{-} 是固定的, $(I - A^{-}A)z$ 为任意项. 而在定理 16.0.6 中, A^{-} 是变的, 是任意的. 这两种表示各有其方便之处, 在以后的讨论中我们要经常用到它们.

Theorem 16.0.7 设

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

可逆. 若 $|A_{11}| \neq 0$, 则

$$A^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22,1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22,1}^{-1} \\ -A_{22,1}^{-1}A_{21}A_{11}^{-1} & A_{22,1}^{-1} \end{pmatrix}$$

若 $|A_{22}| \neq 0$, 则

$$A^{-1} = \begin{pmatrix} A_{11,2}^{-1} & -A_{11,2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}A_{11,2}^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}A_{11,2}^{-1}A_{12}A_{22}^{-1} \end{pmatrix},$$

其中 $A_{22.1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$, $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$

Proof. 证明 若 $|A_{11}| \neq 0$, 则有

$$\begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22.1} \end{pmatrix} \quad (16.3)$$

此式证明了 $A_{22.1}$ 的可逆性. 两边求逆矩阵, 容易得到

$$\begin{aligned} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22.1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \\ &= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22.1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22.1}^{-1} \\ -A_{22.1}^{-1}A_{21}A_{11}^{-1} & A_{22.1}^{-1} \end{pmatrix} \end{aligned}$$

■

矩阵 A 不可逆时考虑广义逆

Theorem 16.0.8 — 分块矩阵的广义逆. 1. 若 A_{11}^{-1} 存在, 则

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^- = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}A_{22.1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22.1}^{-1} \\ -A_{22.1}^{-1}A_{21}A_{11}^{-1} & A_{22.1}^{-1} \end{pmatrix}.$$

2. 若 A_{22}^{-1} 存在, 则

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^- = \begin{pmatrix} A_{11.2}^- & -A_{11.2}^-A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}A_{11.2}^- & A_{22}^{-1} + A_{22}^{-1}A_{21}A_{11.2}^-A_{12}A_{22}^{-1} \end{pmatrix}.$$

3. 若

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \geq 0$$

则

$$A^- = \begin{pmatrix} A_{11}^- + A_{11}^-A_{12}A_{22.1}^{-1}A_{21}A_{11}^- & -A_{11}^-A_{12}A_{22.1}^{-1} \\ -A_{22.1}^{-1}A_{21}A_{11}^- & A_{22.1}^- \end{pmatrix} \quad (16.4)$$

或

$$A^- = \begin{pmatrix} A_{11.2}^- & -A_{11.2}^-A_{12}A_{22}^- \\ -A_{22}^-A_{21}A_{11.2}^- & A_{22}^- + A_{22}^-A_{21}A_{11.2}^-A_{12}A_{22}^- \end{pmatrix} \quad (16.5)$$

其中 $A_{22.1} = A_{22} - A_{21}A_{11}^-A_{12}$, $A_{11.2} = A_{11} - A_{12}A_{22}^-A_{21}$

Proof. 我们只证明 (1) 和 (3),(2) 的证明与 (1) 类似. 先证 (1). 当 A_{11}^{-1} 存在时, (16.3) 式仍成立. 于是根据事实: $B = PCQ$, P, Q 可逆, 则 $B^- = Q^{-1}C^-P^{-1}$ (证明留作习题), 有

$$\begin{aligned} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^- &= \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22.1} \end{pmatrix}^- \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \\ &= \begin{pmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22.1}^- \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{pmatrix} \end{aligned}$$

这里, 我们利用了事实:

$$\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22,1}^{-1} \end{pmatrix}$$

是准对角阵

$$\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22,1} \end{pmatrix}$$

的广义逆。把上面三个矩阵乘开来, 即得所证. 再证 (3). 因 $A \geq 0$, 故存在矩阵 $B = (B_1 : B_2)$, 使得

$$A = B'B = \begin{pmatrix} B'_1 B_1 & B'_1 B_2 \\ B'_2 B_1 & B'_2 B_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

由推论 16.0.4 的 (2), 有

$$\begin{aligned} A_{21} A_{11}^{-1} A_{11} &= B'_2 B_1 (B'_1 B_1)^{-1} B'_1 B_1 = B'_2 B_1 = A_{21} \\ A_{11} A_{11}^{-1} A_{12} &= B'_1 B_1 (B'_1 B_1)^{-1} B'_1 B_2 = B'_1 B_2 = A_{12} \end{aligned}$$

于是, 和 (16.3) 相类似, 有

$$\begin{pmatrix} I & 0 \\ -A_{21} A_{11}^{-1} & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & -A_{11}^{-1} A_{12} \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22,1} \end{pmatrix} \quad (16.6)$$

依此事实及用与前面完全相同的方法, 可得

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I & -A_{11}^{-1} A_{12} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22,1}^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -A_{21} A_{11}^{-1} & I \end{pmatrix}$$

将此三矩阵相乘, 即得所证. 用类似方法可证第二种表达式. 定理证毕. ■

从定理证明过程可以看出, 我们所求到的广义逆只是 A^- 的一部分. 因此, 定理中的 A^- 的四个表达式, 应理解为右端是 A 的广义逆. 这一点并不影响我们后面的应用. 因为在线性模型估计理论中, 我们所关心的量都与 A^- 的选择无关.

定理的条件 A_{11}^{-1} 或 A_{22}^{-1} 存在或 $A \geq 0$ 还可以进一步削弱. 因为, 由 $\mathcal{M}(A_{12}) \subset \mathcal{M}(A_{11})$ 和 $\mathcal{M}(A'_{21}) \subset \mathcal{M}(A'_{11})$ 可推出 $A_{11} A_{11}^- A_{12} = A_{12}$ 和 $A_{21} A_{11}^- A_{12} = A_{21}$, 于是, (16.6) 成立. 因此, (16.4) 和 (16.5) 也成立. 故得

Corollary 16.0.9 对矩阵

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

若 $\mathcal{M}(A_{12}) \subset \mathcal{M}(A_{11}), \mathcal{M}(A'_{21}) \subset \mathcal{M}(A'_{11})$, 则 (16.4) 和 (16.5) 成立.

上述 A^- 有无数个, 接下去引入 M-P 广义逆

Definition 16.0.2 设 A 为任一矩阵, 若 X 满足下述四个条件:

$$AXA = A, XAX = X, (AX)' = AX, (XA)' = XA \quad (16.7)$$

则称矩阵 X 为 A 的 Moore-Penrose 广义逆, 记为 A^+ . 有时称 (16.7) 为 Penrose 方程.

Lemma 16.1 — 奇异值分解. 设矩阵 $A_{m \times n}$ 的秩为 r , 记为 $\text{rk}(A) = r$, 则存在两个正交方阵 $P_{m \times m}, Q_{n \times n}$, 使

$$A = P \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} Q' \quad (16.8)$$

其中 $\Lambda_r = \text{diag}(\lambda_1, \dots, \lambda_r), \lambda_i > 0, i = 1, 2, \dots, r$. $\lambda_1^2, \dots, \lambda_r^2$ 为 $A'A$ 的非零特征根. ■

Proof. 因为 $A'A$ 为对称阵, 故存在正交方阵 $Q_{n \times n}$, 致

$$Q'A'AQ = \begin{pmatrix} \Lambda_r^2 & 0 \\ 0 & 0 \end{pmatrix}$$

记 $B = AQ$, 上式即为

$$B'B = \begin{pmatrix} \Lambda_r^2 & 0 \\ 0 & 0 \end{pmatrix}$$

这说明 B 的列向量互相正交, 且前 r 个列向量长度分别为 $\lambda_1, \dots, \lambda_r$, 后 $n - r$ 个列向量为零向量. 于是, 存在一正交方阵 $P_{m \times m}$, 使得

$$B = P \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix}$$

再由 $B = AQ$. 立得 (16.8). 证毕. 通常称 $\lambda_1, \dots, \lambda_r$ 为 A 的奇异值. 利用这个引理, 可以构造地给出 A^+ ■

Theorem 16.0.10 1. 设 A 有分解式 (16.8), 则

$$A^+ = Q \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} P' \quad (16.9)$$

2. 对任何矩阵 A, A^+ 惟一".

Proof. 1. 很容易直接验证, (16.9) 的右端满足 (16.7).

2. 设 X 和 Y 都是 A^+ , 由 (16.7) 的四个条件知

$$\begin{aligned} X &= XAX = X(AX)' = XX'A' = XX'(AYA)' = X(AX)'(AY)' = (XAX)AY \\ &= XAY = (XA)'YAY = A'X'A'Y'Y = A'Y'Y = (YA)'Y = YAY = Y \end{aligned}$$

这就证明了惟一性. ■

(R) A^+ 也是一定存的

因为 A^+ 是一个特殊的 A^- , 因此, 它除了具有 A^- 的全部性质外, 还有下列性质.

Corollary 16.0.11 1. $(A^+)^+ = A$
 2. $(A^+)' = (A')^+$
 3. $I \geq A^+A$
 4. $\text{rk}(A^+) = \text{rk}(A)$
 5. $A^+ = (A'A)^+A' = A'(AA')^+$
 6. $(A'A)^+ = A^+(A')^+$

7. 设 a 为一非零向量, 则 $a^+ = a'/\|a\|^2$

8. 若 A 为对称方阵, 它可表为

$$A = P \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} P'$$

这里 P 为正交阵, $\Lambda_r = \text{diag}(\lambda_1, \dots, \lambda_r)$, $r = \text{rk}(A)$, 则

$$A^+ = P \begin{pmatrix} \Lambda_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} P'$$

从定理 16.0.5 知, 对相容线性方程组 $Ax = b$, $x_0 = A^+b$ 必为解.

Theorem 16.0.12 在相容线性方程组 $Ax = b$ 的解集中, $x_0 = A^+b$ 为长度最小者.

Proof. 由 (16.2), $Ax = b$ 的通解可表为

$$x = A^+b + (I - A^+A)z$$

于是

$$\begin{aligned} \|x\|^2 &= (A^+b + (I - A^+A)z)'(A^+b + (I - A^+A)z) \\ &= \|x_0\|^2 + z'(I - A^+A)^2z + 2b'(A^+)'(I - A^+A)z \\ &= \|x_0\|^2 + z'(I - A^+A)^2z \geq \|x_0\|^2 \end{aligned} \quad (16.10)$$

根据上面的性质 5: $A^+ = (A'A)^+A'$, 有 $A'^+ = (AA')^+A$, 所以 $(A^+)'(I - A^+A) = (A^+)' - (A^+)'A^+A = 0$ 和 $z'(I - A^+A)^2z \geq 0$ 对任意的 z 成立. 在 (16.10) 中, 等号成立 $\iff (I - A^+A)z = 0 \iff x = A^+b$. 证毕. ■

上面我们所讨论的广义逆 A^- 和 A^+ , 是满足 (16.7) 第一条和全部四条的两个极端情况. 自然我们还可以定义满足四个条件中任一个、任两个或任三个的广义逆.

幂等矩阵

Definition 16.0.3 若方阵 $A_{n \times n}$ 满足 $A^2 = A$, 则称 A 为幂等阵 (idempotent matrix).

Theorem 16.0.13 幂等阵的特征根只能为 0 或 1.

Theorem 16.0.14 对任意的矩阵 A ,

1. $A^-A, AA^-, I - A^-A$, 和 $I - AA^-$ 都是幂等阵. 特别, $A^+A, AA^+, I - A^+A$ 和 $I - AA^+$ 都是幂等阵;
2. 若 A 为对称幂等阵, 则 $A^+ = A$

Proof. 从定义容易验证 (1), 利用定理 16.0.13 和推论 16.0.11 之 (8), 立得 (2). ■

Theorem 16.0.15 1. 若 $A_{n \times n}$ 幂等, 则 $\text{tr}(A) = \text{rk}(A)$

2. $A_{n \times n}$ 幂等 $\iff \text{rk}(A) + \text{rk}(I - A) = n$

Proof. 1. 设 $\text{rk}(A) = r$, 则存在可逆方阵 P, Q , 使

$$A = P \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q$$

将 P, Q 分块: $P = (P_1 : P_2)$, 其中 P_1 为 $n \times r$ 的矩阵, $Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$, 其中 Q_1 为 $r \times n$ 的矩阵, 于是 $A = P_1 Q_1$. 另一方面, 由 $A^2 = A$, 得到

$$\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q P \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

故 $Q_1 P_1 = I_r$. 所以 $\text{tr}(A) = \text{tr}(P_1 Q_1) = \text{tr}(Q_1 P_1) = \text{tr}(I_r) = r = \text{rk}(A) \cdot (1)$ 得证.

2. 必要性是显然的. 事实上, 由 A 的幂等性知, $I - A$ 也幂等. 利用刚证过的性质, 有

$$n = \text{tr}(I_n) = \text{tr}(I_n - A + A) = \text{tr}(I_n - A) + \text{tr}(A) = \text{rk}(I_n - A) + \text{rk}(A)$$

反过来, 设 $\text{rk}(A) = r$, 则 $Ax = 0$ 有 $n - r$ 个线性无关的解, 它们是对应于特征根零的 $n - r$ 个线性无关的特征向量. 由 $\text{rk}(I - A) = n - r$ 知, $Ax = x$ 有 r 个线性无关的解, 它们是对应于特征根 1 的 r 个线性无关的特征向量. 因为这 n 个特征向量线性无关, 于是 A 相似于

$$\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

即存在可逆阵 P , 使

$$A = P \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} P^{-1}$$

故 $A^2 = A$. 证毕. ■

Theorem 16.0.16 设 $P_{n \times n}$ 为对称幂等阵, $\text{rk}(P) = r$, 则存在秩为 r 的 $A_{n \times r}$, 使 $P = A(A'A)^{-1}A'$

Proof. 因 P 为对称幂等阵, 故存在正交阵 $R = (R_1 : R_2)$, 使得

$$P = R \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} R' = (R_1 \quad R_2) \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} R'_1 \\ R'_2 \end{pmatrix} = R_1 R'_1 = R_1 (R'_1 R_1)^{-1} R'_1$$

这里用到了 $R'_1 R_1 = I_r$. 再令 $A = R_1$, 定理得证. ■

现在我们讨论正交投影和正交投影阵.

Definition 16.0.4 — 正交投影阵. 设 $x \in R_n, S$ 为 R_n 的一个线性子空间. 对 x 作分解

$$x = y + z, \quad y \in S, \quad z \in S^\perp \tag{16.11}$$

则称 y 为 x 在 S 上的正交投影. 若 P 为 n 阶方阵, 使得对一切 $x \in R_n$, (16.11) 定义的 y 满足 $y = Px$, 则称 P 为向 S 的正交投影阵.

我们知道, 对 R_n 的任一子空间 S , 都可以找到矩阵 $A_{n \times m}$, 使得 $S = \mathcal{M}(A)$ 所以, 下面的定理给出了正交投影阵的表示.

Theorem 16.0.17 设 A 为 $n \times m$ 矩阵, P_A 为向 $\mathcal{M}(A)$ 的正交投影阵, 则 $P_A = A(A'A)^{-1}A'$

Proof. 记 B 为一矩阵, 使得 $\mathcal{M}(B) = \mathcal{M}(A)^\perp$, 则对任一 $x \in R_n$, 有分解 $x = A\alpha + B\beta$, 这里 α, β 为适当维数的列向量. 依定义, $P_A x = P_A A\alpha + P_A B\beta = A\alpha$ 对一切 α, β 都成立. 故正交投影阵 P_A 满足矩阵方程组

$$\begin{cases} P_A A = A \\ P_A B = 0 \end{cases}$$

由第二方程推得, $\mathcal{M}(P_A) \subset \mathcal{M}(B)^\perp = \mathcal{M}(A)$. 于是, 存在矩阵 $U, P'_A = AU$. 代入第一方程, 得 $U'A'A = A$. 此方程组是相容的, 由定理 16.0.6, $U = (A'A)^{-1}A'$. 于是

$$P_A = U'A' = A \left((A'A)^{-1} \right)' A' = A (A'A)^{-1} A'$$

这里应用了推论 16.0.4 之 (1) 及 $((A'A)^{-1})'$ 仍为一个 $(A'A)^{-1}$. 定理证毕. 因为 $P_A = A (A'A)^{-1} A'$ 与广义逆选择无关, 所以正交投影阵是惟一的. ■

Theorem 16.0.18 P 为正交投影阵 $\iff P$ 为对称幂等阵.

Proof. 设 P 为向 $\mathcal{M}(A)$ 的正交投影阵, 由上一定理, $P = A (A'A)^{-1} A' = A (A'A)^+ A'$, 对称性得证. 利用推论 16.0.4(2), 有

$$P^2 = A (A'A)^{-1} A' A (A'A)^{-1} A' = A (A'A)^{-1} A' = P$$

必要性得证. 充分性即定理 16.0.16. 证毕. ■

Theorem 16.0.19 n 阶方阵 P 为正交投影阵 \iff 对任给 $x \in R_n$

$$\|x - Px\| = \inf \|x - u\|, \quad u \in \mathcal{M}(P) \quad (16.12)$$

Proof. 先证必要性. 任取 $u \in \mathcal{M}(P), v \in \mathcal{M}(P)^\perp$, 记 $y = u + v$, 则 $u = Py$.

$$\begin{aligned} \|x - u\|^2 &= \|x - Py\|^2 = \|x - Px + Px - Py\|^2 \\ &= \|(x - Px) + P(x - y)\|^2 \\ &= \|x - Px\|^2 + \|P(x - y)\|^2 + 2x'(I - P)P(x - y) \\ &= \|x - Px\|^2 + \|P(x - y)\|^2 \\ &\geq \|x - Px\|^2 \end{aligned} \quad (16.13)$$

等号成立 $\iff Px = Py$, 即 $u = Px$. 必要性得证.

充分性. 若 (16.12) 成立, 我们首先证明

$$x'(I - P)'P(x - y) = 0, \quad \text{对一切 } x, y \text{ 成立.}$$

用反证法. 假设存在 x_0 和 y_0 , 使得

$$x_0'(I - P)'P(x_0 - y_0) = c \neq 0$$

可以假定 $c < 0$. 因为若 $c > 0$, 则取满足 $x_0 - y_1 = -(x_0 - y_0)$ 的 y_1 代替 y_0 , 便化为 $c < 0$ 的情形. 取 y 满足 $x_0 - y = \varepsilon(x_0 - y_0)$, 并记 $u = Py$, 则

$$\begin{aligned} \|x_0 - u\|^2 &= \|x_0 - Py\|^2 \\ &= \|x_0 - Px_0\|^2 + \|P(x_0 - y)\|^2 + 2x_0'(I - P)P(x_0 - y) \\ &= \|x_0 - Px_0\|^2 + \varepsilon^2 \|P(x_0 - y_0)\|^2 + 2\varepsilon x_0'(I - P)P(x_0 - y_0) \\ &= \|x_0 - Px_0\|^2 + \varepsilon^2 \|P(x_0 - y_0)\|^2 + 2\varepsilon c \end{aligned}$$

因 $c < 0$, 故取 $\varepsilon > 0$ 充分小, 可使上式后两项小于零. 于是

$$\|x_0 - u\|^2 < \|x_0 - Px_0\|^2$$

这与 (16.12) 矛盾, 这就证明了 (16.13). 因 ?? 对一切 x 和 y 成立, 故 $\mathcal{M}(P)$ 与 $\mathcal{M}(I-P)$ 正交. 据此易推知, $\text{rk}(P) + \text{rk}(I-P) = n$. 所以, 对任意 $x \in R_n$, 有分解式

$$x = Px + (I-P)x, \quad Px \in \mathcal{M}(P), \quad (I-P)x \in \mathcal{M}(P)^\perp$$

依定义, P 为向 $\mathcal{M}(P)$ 的正交投影阵. 定理证毕. ■

这个定理刻画了正交投影阵的距离最短性, 即在线性子空间 $\mathcal{M}(P)$ 的所有向量中, 只有 x 的正交投影阵 Px 到 x 的距离 $\|x - Px\|$ 最短. 这个结果在最小二乘估计理论中有重要应用. 在一定的条件下, 正交投影阵的和, 差, 积仍为正交投影阵, 这些结果概括在如下三个定理中.

Theorem 16.0.20 设 P_1 和 P_2 为两个正交投影阵, 则

1. $P = P_1 + P_2$ 为正交投影 $\iff P_1P_2 = P_2P_1 = 0$
2. 当 $P_1P_2 = P_2P_1 = 0$ 时, $P = P_1 + P_2$ 为向 $\mathcal{M}(P_1) \oplus \mathcal{M}(P_2)$ 上的正交投影.

Proof. 1. 充分性易证, 下证必要性. 假设 P 是一个正交投影阵, 根据定理 16.0.18 知 $P^2 = P$. 于是

$$P_1P_2 + P_2P_1 = 0 \tag{16.14}$$

用 P_1 分别左乘和右乘 (16.14) 得到

$$\begin{aligned} P_1P_2 + P_1P_2P_1 &= 0 \\ P_2P_1 + P_1P_2P_1 &= 0 \end{aligned} \tag{16.15}$$

把上两式相加, 并利用 (16.14), 得到

$$P_1P_2P_1 = 0 \tag{16.16}$$

再由 (16.15) 和 (16.16), 便得到 $P_1P_2 = P_2P_1 = 0$

2. 我们只需要证明

$$\mathcal{M}(P) = \mathcal{M}(P_1) \oplus \mathcal{M}(P_2)$$

对任一 $y \in \mathcal{M}(P)$, 存在 $x \in R^n$, 使得 $y = Px$, 于是

$$y = Px = P_1x + P_2x = y_1 + y_2$$

这里 $y_i = P_ix \in \mathcal{M}(P_i)$, $i = 1, 2$, 且从 $P_1P_2 = 0$ 可推知 $y_1 \perp y_2$. 定理证毕. ■

Theorem 16.0.21 设 P_1 和 P_2 为两个正交投影阵, 则

1. $P = P_1P_2$ 也为正交投影阵 $\iff P_1P_2 = P_2P_1 = P_2$
2. 当 $P_1P_2 = P_2P_1$ 时, $P = P_1P_2$ 为向 $\mathcal{M}(P_1) \cap \mathcal{M}(P_2)$ 上的正交投影阵.

Theorem 16.0.22 设 P_1 和 P_2 为两个正交投影阵, 则

1. $P = P_1 - P_2$ 为正交投影阵 $\iff P_1P_2 = P_2P_1 = P_2$
2. 当 $P = P_1 - P_2$ 为正交投影阵时, P 为向 $\mathcal{M}(P_1) \cap \mathcal{M}(P_2)^\perp$ 上的正交投影.

特征值不等式

设 A 为 $n \times n$ 实对称阵, 我们用 $\lambda_1(A), \dots, \lambda_n(A)$ 表示 A 的特征值. 在不致引起混淆时, 也简记为 $\lambda_1, \dots, \lambda_n$. 记 $\varphi_1, \dots, \varphi_n$ 为对应的标准正交化特征向量. 我们总假定 $\lambda_1(A) \geq \dots \geq \lambda_n(A)$, 并称 $\lambda_i(A)$ 为 A 的第 i 个顺序特征值.

Theorem 16.0.23 — Rayleigh-Ritz. 设 A 为 $n \times n$ 对称阵, 则

1. $\sup_{x \neq 0} \frac{x'Ax}{x'x} = \varphi_1' A \varphi_1 = \lambda_1$
2. $\inf_{x \neq 0} \frac{x'Ax}{x'x} = \varphi_n' A \varphi_n = \lambda_n$

Proof. 记 $\Phi = (\varphi_1, \dots, \varphi_n)$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$. 对任意 $x \in R_n$, 存在向量 t 使 $x = \Phi t$. 故

$$\frac{x'Ax}{x'x} = \frac{t'\Lambda t}{t't} = \sum_{i=1}^n \lambda_i \omega_i \leq \lambda_1 \sum_{i=1}^n \omega_i = \lambda_1$$

这里 $\omega_i = t_i^2 / \sum t_j^2 \geq 0$, $\sum_{i=1}^n \omega_i = 1$, 并且等号成立 $\iff \omega_1 = 1, \omega_i = 0, i > 1 \iff x = a\varphi_1$, 其中 a 为数. (1) 得证. ■

Corollary 16.0.24 对任一 n 阶对称方阵 $A = (a_{ij})$, 总有 $\lambda_n \leq a_{ii} \leq \lambda_1, i = 1, \dots, n$

Corollary 16.0.25 设 A 为 $n \times n$ 对称阵, 则

1. $\sup_{\substack{\varphi_i' x = 0 \\ i=1, \dots, k}} \frac{x'Ax}{x'x} = \varphi_{k+1}' A \varphi_{k+1} = \lambda_{k+1}$
2. $\inf_{\substack{\varphi_i' x = 0 \\ i=1, \dots, k}} \frac{x'Ax}{x'x} = \varphi_n' A \varphi_n = \lambda_n$
3. $\sup_{\substack{\varphi_i' x = 0 \\ i=k+1, \dots, n}} \frac{x'Ax}{x'x} = \varphi_1' A \varphi_1 = \lambda_1$
4. $\inf_{\substack{\varphi_i' x = 0 \\ i=k+1, \dots, n}} \frac{x'Ax}{x'x} = \varphi_k' A \varphi_k = \lambda_k$

Theorem 16.0.26 设 A 为 $n \times n$ 对称阵, B 为 $n \times k$ 对称阵, 则

1. $\inf_B \sup_{B'x=0} \frac{x'Ax}{x'x} = \sup_{\Phi_k' x = 0} \frac{x'Ax}{x'x} = \varphi_{k+1}' A \varphi_{k+1} = \lambda_{k+1}$
2. $\sup_B \inf_{B'x=0} \frac{x'Ax}{x'x} = \inf_{\Phi_{(k)}' x = 0} \frac{x'Ax}{x'x} = \varphi_{n-k}' A \varphi_{n-k} = \lambda_{n-k}$

其中 $\Phi_k, \Phi_{(k)}$ 分别表示 $\Phi = (\varphi_1, \dots, \varphi_n)$ 的前 k 列和后 k 列

Proof. 1. 记 $x = \Phi y$, 则

$$\begin{aligned} \sup_{B'x=0} \frac{x'Ax}{x'x} &= \sup_{H'y=0} \frac{y'\Lambda y}{y'y} \geq \sup_{H'(y'_1 \ 0)'=0} \frac{y'_1 \Lambda_1 y_1}{y'_1 y_1} \\ &\geq \inf_{H'(y'_1 \ 0)'=0} \frac{y'_1 \Lambda_1 y_1}{y'_1 y_1} \geq \inf_{y_1 \neq 0} \frac{y'_1 \Lambda_1 y_1}{y'_1 y_1} = \lambda_{k+1} \end{aligned}$$

其中 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, $H = \Phi' B$, $\Lambda_1 = \text{diag}(\lambda_1, \dots, \lambda_{k+1})$, $y' = (y'_1 y'_2)$, $y_1 : (k+1) \times 1$, 于是

$$\inf_B \sup_{B'x=0} \frac{x'Ax}{x'x} \geq \lambda_{k+1}$$

再由推论 2.4.2 之 (1) 知,

$$\sup_{\Phi'x=0} \frac{x'Ax}{x'x} = \varphi_{k+1}' A \varphi_{k+1} = \lambda_{k+1}$$

明所欲证.

2. (2) 用与 (1) 同样的记号

$$\begin{aligned} \inf_{B'x=0} \frac{x'Ax}{x'x} &= \inf_{H'y=0} \frac{y'\Lambda y}{y'y} \leq \inf_{H'(0 \ y'_2)'=0} \frac{y'_2 \Lambda_2 y_2}{y'_2 y_2} \\ &\leq \sup_{H'(0 \ y'_2)'} = 0 \quad \frac{y'_2 \Lambda_2 y_2}{y'_2 y_2} \leq \sup_{y_2} \frac{y'_2 \Lambda_2 y_2}{y'_2 y_2} = \lambda_{n-k} \end{aligned}$$

其中 $\Lambda_2 = \text{diag}(\lambda_{n-k}, \dots, \lambda_n)$, $y' = (y'_1, y'_2), y_2 : (n-k) \times 1$. 那么

$$\sup_B \inf_{B'x=0} \frac{x'Ax}{x'x} \leq \lambda_{n-k}$$

由推论 16.0.25 之 (4) 知

$$\inf_{\Phi'_{(k)}x=0} \frac{x'Ax}{x'x} = \varphi'_{n-k} A \varphi_{n-k} = \lambda_{n-k}$$

这就完成了定理的证明. ■

Theorem 16.0.27 — Sturm 分离定理. 设 A 为 $n \times n$ 对称阵, 记

$$A_r = \begin{pmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rr} \end{pmatrix}, \quad r = 1, \dots, n$$

为 A 的顺序主子式, 则

$$\lambda_{i+1}(A_{r+1}) \leq \lambda_i(A_r) \leq \lambda_i(A_{r+1}), \quad i = 1, 2, \dots, r$$

Proof. 先证第一不等式, 记 g_i 为 A_r 对应于特征根 $\lambda_i(A_r)$ 的标准正交化特征向量, $i = 1, \dots, r$, 依推论 16.0.25 (1), 得

$$\lambda_i(A_r) = \sup_{\substack{g'_j x=0 \\ j=1, \dots, i-1}} \frac{x' A_r x}{x' x} = \sup_{\substack{y_{r+1}=0 \\ (g'_j, 0)y=0 \\ j=1, \dots, i-1}} \frac{y' A_{r+1} y}{y' y} \geq \inf_B \sup_{B' y=0} \frac{y' A_{r+1} y}{y' y} = \lambda_{i+1}(A_{r+1})$$

其中 $y : (r+1) \times 1, B : (r+1) \times i$, 这里应用了定理 16.0.26.

再证第二个不等式. 记 $\psi_i, i = 1, \dots, r+1$ 为 A_{r+1} 对应特征根 $\lambda_i(A_{r+1}), i = 1, \dots, r+1$, 的标准正交化特征向量, 类似地, 有

$$\begin{aligned} \lambda_i(A_{r+1}) &= \sup_{\substack{\psi'_j y=0 \\ j=1, \dots, i-1}} \frac{y' A_{r+1} y}{y' y} \geq \sup_{\substack{y_{r+1}=0 \\ \psi'_j y=0 \\ j=1, \dots, i-1}} \frac{y' A_{r+1} y}{y' y} \\ &= \sup_{\substack{\tilde{\psi}'_j x=0 \\ j=1, \dots, i-1}} \frac{x' A_r x}{x' x} \geq \inf_B \sup_{B' x=0} \frac{x' A_r x}{x' x} = \lambda_i(A_r) \end{aligned}$$

其中 $\psi_j = (\tilde{\psi}'_{j \times 1} *)'$, $B : r \times (i-1)$. 定理得证. ■

Theorem 16.0.28 — Weyl 定理. 设 A 和 B 皆为 $n \times n$ 的对称阵, 则

$$\lambda_i(A) + \lambda_n(B) \leq \lambda_i(A+B) \leq \lambda_i(A) + \lambda_1(B), \quad i = 1, \dots, n$$

Proof. 设 $x'x = 1$, 显然有

$$x'Ax + \min(x'Bx) \leq x(A+B)x \leq x'Ax + \max(x'Bx)$$

根据定理 16.0.23 有

$$\lambda_i(A) + \lambda_n(B) \leq \lambda_i(A+B) \leq \lambda_i(A) + \lambda_1(B)$$

证毕。 ■

Weyl 定理给出了 $A+B$ 特征根的上、下界.

Theorem 16.0.29 — Poincare 分离定理. 设 $A_{n \times n}$ 为对称阵, P 为 $n \times k$ 的列正交阵, 即 $P'P = I_k$, 则

$$\lambda_{n-k+i}(A) \leq \lambda_i(P'AP) \leq \lambda_i(A), \quad i = 1, \dots, k$$

Proof. 将 P 扩充为正交方阵 $\tilde{P} = (P|Q)$, 记

$$H = \tilde{P}'A\tilde{P} = \begin{pmatrix} P'AP & P'AQ \\ Q'AP & Q'AQ \end{pmatrix}$$

H_k 为 H 的 k 阶顺序主子阵. 注意到 $H_k = P'AP, H_n = \tilde{P}'A\tilde{P}$, 利用 sturm 定理有

$$\begin{aligned} \lambda_i(A) &= \lambda_i(\tilde{P}'A\tilde{P}) \geq \lambda_i(P'AP) = \lambda_i(H_k) \geq \lambda_{i+1}(H_{k+1}) \geq \dots \\ &\geq \lambda_{i+(n-k)}(H_n) = \lambda_{i+n-k}(\tilde{P}'A\tilde{P}) = \lambda_{n-k+i}(A) \end{aligned}$$

即 $\lambda_i(A) \geq \lambda_i(P'AP)$ 和 $\lambda_i(P'AP) \geq \lambda_{n-k+i}(A)$. 证毕。 ■

Theorem 16.0.30 — Kantorovich 不等式. 设 $A_{n \times n}$ 为正定阵, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ 为 A 的特征根, 则

$$1 \leq \frac{x'Ax \cdot x'A^{-1}x}{(x'x)^2} \leq \frac{1}{4} \frac{(\lambda_1 + \lambda_n)^2}{\lambda_1 \lambda_n}$$

Proof. 左边的不等式容易从 Cauchy-Schwarz 不等式 (见本章末习题) 得到. 现证右边不等式, 首先

$$\frac{x'Ax \cdot x'A^{-1}x}{(x'x)^2} \leq \frac{1}{4} \frac{(\lambda_1 + \lambda_n)^2}{\lambda_1 \lambda_n} \iff x'Ax \cdot x'A^{-1}x \leq \frac{\lambda_1 + \lambda_n}{2} \cdot \frac{\lambda_1^{-1} + \lambda_n^{-1}}{2}$$

其中 $x'x = 1$. 设 Q 为正交方阵, 致 $A = Q\Lambda Q'$, $\Lambda = \text{diag}(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n)$. 记 $u = Q'x$, 上式 \iff

$$u'\Lambda u \cdot u'\Lambda^{-1}u \leq \frac{\lambda_1 + \lambda_n}{2} \cdot \frac{\lambda_1^{-1} + \lambda_n^{-1}}{2} \iff u' \left(\frac{2}{\lambda_1 + \lambda_n} \Lambda \right) u \cdot u' \left(\frac{2}{\lambda_1^{-1} + \lambda_n^{-1}} \Lambda^{-1} \right) u \leq 1$$

其中 $u'u = 1$. 利用几何平均小于算术平均, 则上式的一个充分条件为, 对一切 u

$$u' \left(\frac{\Lambda}{\lambda_1 + \lambda_n} + \frac{\Lambda^{-1}}{\lambda_1^{-1} + \lambda_n^{-1}} \right) u \leq u'u$$

而此式又

$$\begin{aligned} &\iff \frac{\lambda_i}{\lambda_1 + \lambda_n} + \frac{\lambda_i^{-1}}{\lambda_1^{-1} + \lambda_n^{-1}} \leq 1, \quad i = 1, \dots, n \\ &\iff (\lambda_i - \lambda_1)(\lambda_i - \lambda_n) \leq 0, \quad i = 1, \dots, n \end{aligned}$$

明所欲证. ■

Theorem 16.0.31 — (Wielandt). 设 A 为 $n \times n$ 正定对称阵, $\lambda_1 \geq \dots \geq \lambda_n > 0$ 为 A 的特征值, 则对任意一对正交向量 x 和 y , 有

$$|x'Ay|^2 \leq \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^2 x'Ax \cdot y'Ay \quad (16.17)$$

且存在正交向量 x 和 y , 使 (16.17) 的等号成立.

Proof. 显然我们只需对 $\|x\| = 1, \|y\| = 1$ 的正交向量证明 (16.17). 设 x 和 y 为任一对标准正交向量, 定义

$$B = (x, y)' A (x, y)$$

这里 B 是一个 2×2 正定对称阵, 记其特征值为 $\mu_1 \geq \mu_2 > 0$. 根据 Poincare 定理, 我们有

$$\lambda_1 \geq \mu_1 \geq \mu_2 \geq \lambda_n$$

另一方面

$$\begin{aligned} 1 - \frac{|x'Ay|^2}{x'Ax \cdot y'Ay} &= 4 \frac{x'Ax \cdot y'Ay - |x'Ay|^2}{(x'Ax + y'Ay)^2 - (x'Ax - y'Ay)^2} = \frac{4 \det B}{\text{tr}(B)^2 - (x'Ax - y'Ay)^2} \\ &= \frac{4\mu_1\mu_2}{(\mu_1 + \mu_2)^2 - (\mu_1 - \mu_2)^2} \geq \frac{4\mu_1\mu_2}{(\mu_1 + \mu_2)^2} \end{aligned} \quad (16.18)$$

这里等号成立当且仅当 $x'Ax = y'Ay$, 且 x, y 为一对标准正交向量. (16.18) 可以改写为

$$\frac{|x'Ay|^2}{x'Ax \cdot y'Ay} \leq 1 - \frac{4\mu_1\mu_2}{(\mu_1 + \mu_2)^2} = \left(\frac{\mu_1 - \mu_2}{\mu_1 + \mu_2} \right)^2 = \left(\frac{\mu_1/\mu_2 - 1}{\mu_1/\mu_2 + 1} \right)^2$$

因头右端是 μ_1/μ_2 的单调函数, 结合 (16.18), 得

$$\frac{|x'Ay|^2}{x'Ax \cdot y'Ay} \leq \left(\frac{\lambda_1/\lambda_n - 1}{\lambda_1/\lambda_n + 1} \right)^2 = \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^2$$

(16.17) 得证. 若记 φ_1 和 φ_n 分别为对应于 λ_1 和 λ_n 的 A 的标准正交化特征向量, 则容易验证, 当 $x = (\varphi_1 + \varphi_n)/\sqrt{2}, y = (\varphi_1 - \varphi_n)/\sqrt{2}$, 等号成立. 定理证毕. ■

16.0.2 偏序

Definition 16.0.5 — Lowner 偏序. 设 A, B 为两个 n 阶对称阵, 若 $B - A \geq 0$, 即 $B - A$ 为半正定阵, 则称 A 低于 B , 记为 $B \geq A$ 或 $A \leq B$. 类似, $A > B$ 表明 $A - B$ 为正定阵容易验证, 对称阵的这种关系满足下列性质:

1. 自反性: $A \geq A$
2. 传递性: 若 $A \geq B, B \geq C$, 则 $A \geq C$
3. 若 $A \geq B, B \geq A$, 则 $A = B$ 这种关系被称为 Lowner 偏序.

因为并非任意两个对称阵都有这种关系, 所以称其为偏序.

定理 2.5.1(单调性) 设 A, B 为两个 n 阶对称阵. (1) 若 $A \geq B$, 则 $\lambda_i(A) \geq \lambda_i(B), i = 1, \dots, n$ (2) 若 $A > B$, 则 $\lambda_i(A) > \lambda_i(B), i = 1, \dots, n$ 此结果可由 Weyl 定理直接得到. 但注意定理 2.5.1 的逆定理未必成立. 例如, 设 $A = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, B = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. 由此立即可得如下推论

Corollary 16.0.32 设 $A \geq B \geq 0$, 则

1. $\text{tr}(A) \geq \text{tr}(B)$
2. $|A| \geq |B|$
3. $\text{rk}(A) \geq \text{rk}(B)$

Theorem 16.0.33 设 A 和 B 为两个 n 阶对称阵, P 为 $n \times k$ 矩阵.

1. 若 $A \geq B$, 则 $P'AP \geq P'BP$
2. 若 $\text{rk}(P) = k, A > B$, 则 $P'AP > P'BP$.

Proof. 1. 由 $A \geq B$ 的定义知, 对任意 $x \in R_n$, 有 $x'(A - B)x \geq 0$, 于是, 对任意 $x \in R_n$

$$x(P'AP - P'BP)x = (Px)'(A - B)(Px) \geq 0$$

此即 $P'AP - P'BP \geq 0$

2. 设 $A > B, \text{rk}(P) = k$, 则对任意 $x \neq 0$, 我们有 $Px \neq 0$, 因此对任意 $x \in R_k(x \neq 0)$, 有

$$x(P'AP - P'BP)x = (Px)'(A - B)(Px) > 0$$

故 $P'AP - P'BP > 0$

■

Theorem 16.0.34 设 $A \geq B \geq 0$, 则

$$\mathcal{M}(B) \subset \mathcal{M}(A)$$

Proof. 首先从定义知: $A \geq B \iff$ 对任意 $x, x'Ax \geq x'Bx$. 若 $x \in \mathcal{M}(A)^\perp$, 则 $x'Ax = 0$, 进而有 $x'Bx = 0$, 也就是 $x \in \mathcal{M}(B)^\perp$, 这就证明了 $\mathcal{M}(A)^\perp \subset \mathcal{M}(B)^\perp$, 因此 $\mathcal{M}(B) \subset \mathcal{M}(A)$. 证毕. ■

Definition 16.0.6 — 半正定方阵的平方根阵. 若 $A \geq 0$, 其所有特征根 $\lambda_i \geq 0$, 则算术平方根 $\lambda_i^{1/2}$ 都是实数. Φ 为 λ_i 对应的 n 个标准正交化特征向量为列组成的矩阵, 记

$$\Lambda^{1/2} = (\lambda_1^{1/2}, \dots, \lambda_n^{1/2})$$

定义

$$A^{1/2} = \Phi \Lambda^{1/2} \Phi'$$

称 $A^{1/2}$ 为 A 的平方根阵。因此

$$\left(A^{1/2}\right)^2 = \Phi\Lambda^{1/2}\Phi'\Phi\Lambda^{1/2}\Phi' = \Phi\Lambda\Phi' = A$$

显然, $A^{1/2} \geq 0$ 如果 $A > 0$, 则不难证明 $A^{1/2} > 0$. 因此, 我们可以求 $A^{1/2}$ 的逆矩阵, 记之为 $A^{-1/2}$, 即 $A^{-1/2} = (A^{1/2})^{-1}$. 利用 Φ 为正交阵, 可以推出

$$A^{-1/2} = \Phi\Lambda^{-1/2}\Phi'$$

其中

$$\Lambda^{-1/2} = \text{diag}\left(\lambda_1^{-1/2}, \dots, \lambda_n^{-1/2}\right)$$

Theorem 16.0.35 设 $A \geq 0, B \geq 0$, 则下面的命题等价.

1. $A \geq B$
2. $\mathcal{M}(B) \subseteq \mathcal{M}(A)$, 对任意的 $x \in \mathcal{M}(A), x'(A - B)x \geq 0$
3. $\mathcal{M}(B) \subseteq \mathcal{M}(A), \lambda_1(BA^-) \leq 1$, 这里 $\lambda_1(BA^-)$ 与 A^- 的选择无关.

Proof. 1. 由定理 16.0.34, (1) \implies (2), 下面证 (2) \implies (1). 设 $x \in R_n$, 且

$$x = y + z, \quad y \in \mathcal{M}(A), z \in \mathcal{M}(A)^\perp$$

则 $Az = 0$, 故 $Bz = 0$. 由于 $y \in \mathcal{M}(A)$, 我们有

$$x'(A - B)x = y'(A - B)y \geq 0$$

即 $A \geq B$, 因此 (2) \iff (1).

2. 下面我们证 (1) \iff (3) 根据定理 16.0.33, 16.0.34, 我们不难证明

$$A \geq B \iff (A^+)^{1/2}(A - B)(A^+)^{1/2} \geq 0, \quad \mathcal{M}(A) \subseteq \mathcal{M}(B)$$

令

$$\begin{aligned} M_1 &= (A^+)^{1/2}A(A^+)^{1/2} \\ M_2 &= (A^+)^{1/2}B(A^+)^{1/2} \end{aligned}$$

注意到 $M_1 = M_1^2 = M_1'$, $\mathcal{M}(M_1) = \mathcal{M}(A)$, 因此 M_1 为向 $\mathcal{M}(A)$ 上的正交投影阵. 由于 $(A^+)^{1/2}AA^+ = (A^+)^{1/2}$, 因此 M_1 与 M_2 可交换, 即 $M_1M_2 = M_2M_1 = M_2$. 于是 M_1 和 M_2 有相同的正交特征向量 $\varphi_1, \dots, \varphi_n$. 不失一般性, 设 $\varphi_1, \dots, \varphi_r$ 为 $\mathcal{M}(A)$ 的一组标准正交基, 且 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$ 是 M_2 对应的特征根, 注意到 M_2 与 BA^- 有相同的特征值, 且由于 $\mathcal{M}(B) \subseteq \mathcal{M}(A), BA^-$ 的特征值与 A^- 的选择无关, 因此 $\lambda_1, \dots, \lambda_r$ 也为 BA^- 的特征值. 记 $\phi_1 = (\varphi_1, \dots, \varphi_r), \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ 故

$$M_1 - M_2 = \phi_1(I_r - \Lambda)\phi_1' \geq 0 \iff \lambda_1 \leq 1$$

因此证明了 (1) \iff (3)



Corollary 16.0.36 设 $A \geq 0, B \geq 0$, 则

1. 若 $\text{rk}(A) = \text{rk}(B)$, 则 $A \geq B$ 当且仅当 $B^+ \geq A^+$
2. 若 $B > 0$, 则 $A \geq B$ 当且仅当 $B^{-1} \geq A^{-1}$; $A > B$ 当且仅当 $B^{-1} > A^{-1}$

Proof. 从定理 16.0.35 推导

$$\begin{aligned} A \geq B &\iff \mathcal{M}(B) \subseteq \mathcal{M}(A), \lambda_1(BA^+) \leq 1 \\ B^+ \geq A^+ &\iff \mathcal{M}(A^+) \subseteq \mathcal{M}(B^+), \lambda_1(BA^+) \leq 1 \end{aligned}$$

从 $\text{rk}(A) = \text{rk}(B)$, 得 $\mathcal{M}(B) = \mathcal{M}(A), \mathcal{M}(A^+) = \mathcal{M}(B^+)$. 由于 $\mathcal{M}(A) = \mathcal{M}(A^+), \mathcal{M}(B) = \mathcal{M}(B^+)$, 故 (1) 得证.

(2) 可由 (1) 直接得到. 证毕. ■

下面我们考虑 $A > B$ 与 $A^2 \geq B^2$ 的关系.

Lemma 16.2 设 A 为 $n \times n$ 实方阵, $\lambda_1(A), \sigma_1(A)$ 分别为它的最大特征根和最大奇异值, 则 $|\lambda_1(A)| \leq \sigma_1(A)$ ■

Proof. 设 x 为 A 的对立于 $\lambda_1(A)$ 的单位特征向量, 则

$$(\lambda_1(A))^2 = x'A'Ax \leq \lambda_1(A'A) = \sigma_1^2(A)$$

故引理得证. ■

Theorem 16.0.37 设 A, B 为两个半正定阵, 则

1. $A^2 \geq B^2 \implies A \geq B$
2. 若 $AB = BA$, 则 $A \geq B \implies A^k \geq B^k \geq 0, k$ 为任意正整数.

Proof. 1. 应用定理 16.0.35 知

$$A^2 \geq B^2 \iff \mathcal{M}(B^2) \leq \mathcal{M}(A^2), \quad \lambda_1(B^2(A^2)^+) \leq 1$$

由于 $\mathcal{M}(B^2) = \mathcal{M}(B), \mathcal{M}(A^2) = \mathcal{M}(A), \lambda_1(B^2(A^2)^+) = \lambda_1(B(A^2)^+B) = (\sigma_1(BA^+))^2$ 注意到 $A \geq 0, B \geq 0$, 故 $\sigma_1(BA^+) > 0$. 因此 $\mathcal{M}(B) \subseteq \mathcal{M}(A), \sigma_1(BA^+) \leq 1$, 依引理 16.2, 有

$$\lambda_1(BA^+) \leq \sigma_1(BA^+)$$

由定理 16.0.35, $A \geq B$.

2. 因为 $AB = BA$, 故存在正交阵 Q , 使得 A, B 同时对角化, 即

$$\begin{aligned} A &= Q\Lambda Q', \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n), \lambda_i \geq 0, i = 1, \dots, n \\ B &= Q\Delta Q', \quad \Delta = \text{diag}(\sigma_1, \dots, \sigma_n), \sigma_1 \geq 0, i = 1, \dots, n \end{aligned}$$

据此容易证明 $A^+B = BA^+ = Q\Lambda^+\Delta Q' \geq 0, (A^+)^+B^+ = B^+(A^+)^+ = Q[(\Lambda^+)^+\Delta^+]Q' \geq 0$, 故

$$\begin{aligned} \lambda_1 &= ((A^+)^+B^k) = [\lambda_1(A^+B)]^k \leq 1 \\ \mathcal{M}(A^k) &= \mathcal{M}(A), \mathcal{M}(B^k) \mathcal{M}(A) \end{aligned}$$

因此有 $A^k \geq B^k$. 证毕. ■

定理 16.0.37 中, 条件 $AB = BA$ 是必要条件. 总的来说, $A \geq B$ 并不一定有 $A^2 \geq B^2$ 成立.

矩阵的两种特殊运算: Kronecker 乘积与向量化运算

Definition 16.0.7 设 $A = (a_{ij})$ 和 $B = (b_{ij})$ 分别为 $m \times n, p \times q$ 的矩阵, 定义矩阵 $C = (a_{ij}B)$. 这是一个 $mp \times nq$ 的矩阵, 称为 A 和 B 的 Kronecker 乘积, 记为 $C = A \otimes B$, 即

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

Proposition 16.0.38 — 这种乘积具有下列性质. 1. $0 \otimes A = A \otimes 0 = 0$

2. $(A_1 + A_2) \otimes B = (A_1 \otimes B) + (A_2 \otimes B), A \otimes (B_1 + B_2) = (A \otimes B_1) + (A \otimes B_2)$
3. $(\alpha A) \otimes (\beta B) = \alpha \beta (A \otimes B)$
4. $(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1 A_2) \otimes (B_1 B_2)$
5. $(A \otimes B)' = A' \otimes B'$
6. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, 和以前一样, 应理解为: $A^{-1} \otimes B^{-1}$ 为 $A \otimes B$ 的广义逆, 但不必是全部广义逆. 特别 $(A \otimes B)^+ = A^+ \otimes B^+$. 当 A, B 都可逆时, 有 $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

Theorem 16.0.39 设 A, B 分别为 $n \times n, m \times m$ 的方阵, $\lambda_1, \dots, \lambda_n$ 和 μ_1, \dots, μ_m 分别为 A, B 的特征值, 则

1. $\lambda_i \mu_j, i = 1, \dots, n, j = 1, \dots, m$ 为 $A \otimes B$ 的特征值且 $|A \otimes B| = |A|^m |B|^n$;
2. $\text{tr}(A \otimes B) = \text{tr}(A) \text{tr}(B)$
3. $\text{rk}(A \otimes B) = \text{rk}(A) \text{rk}(B)$
4. 若 $A \geq 0, B \geq 0$, 则 $A \otimes B \geq 0$

Proof. 1. 记 A, B 的 Jordan 标准形分别为

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ 0 & \lambda_2 & * & \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}, \quad \Delta = \begin{pmatrix} \mu_1 & & & \\ 0 & \mu_2 & * & \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \mu_m \end{pmatrix}$$

依 Jordan 分解, 存在可逆阵 P 和 Q , 使得 $A = P\Lambda P^{-1}, B = Q\Delta Q^{-1}$, 利用 Kronecker 乘积的性质, 得

$$A \otimes B = (P\Lambda P^{-1}) \otimes (Q\Delta Q^{-1}) = (P \otimes Q)(\Lambda \otimes \Delta)(P \otimes Q)^{-1}$$

即 $A \otimes B$ 相似于上三角阵 $\Lambda \otimes \Delta$, 后者的对角元为 $\lambda_i \mu_j, i = 1, \dots, n, j = 1, \dots, m$ 所以, 这些 λ_i, μ_j 为 $A \otimes B$ 的全部特征根, 又

$$|A \otimes B| = |\Lambda \otimes \Delta| = \prod_{i=1}^n \prod_{j=1}^m \lambda_i \mu_j = \left(\prod_{i=1}^n \lambda_i \right)^m \left(\prod_{j=1}^m \mu_j \right)^n = |A|^m |B|^n$$

证毕。

2. 由 (1) 立得 (2) 和 (4), (3) 可由秩的定义直接导出。 ■

Definition 16.0.8 — 向量化. 设 $A_{m \times n} = (a_1, a_2, \dots, a_n)$, 定义 $mn \times 1$ 的向量

$$\text{Vec}(A) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

这是把矩阵 A 按列向量依次排成的向量, 往往称这个程序为矩阵的向量化.

Proposition 16.0.40 — 向量化运算具有下列性质.

1. $\text{Vec}(A + B) = \text{Vec}(A) + \text{Vec}(B)$
2. $\text{Vec}(\alpha A) = \alpha \text{Vec}(A)$, 这里 α 为数
3. $\text{tr}(AB) = (\text{Vec}(A'))' \text{Vec}(B)$
4. $\text{tr}(A) = \text{tr}(AI) = \text{tr}(IA) = (\text{Vec}(I_n))' \text{Vec}(A)$
5. 设 a 和 b 分别为 $n \times 1, m \times 1$ 向量, 则 $\text{Vec}(ab') = b \otimes a$
6. $\text{Vec}(ABC) = (C' \otimes A) \text{Vec}(B)$
7. 设 $X_{m \times n} = (x_1, \dots, x_n)$ 为随机矩阵,

$$\text{Cov}(x_i, x_j) = E(x_i - Ex_i)(x_j - Ex_j)' = v_{ij}\Sigma$$

记 $V = (v_{ij})_{n \times n}$, 则

$$\text{Cov}(\text{Vec}(X)) = V \otimes \Sigma$$

$$\text{Cov}(\text{Vec}(X')) = \Sigma \otimes V$$

$$\text{Cov}(\text{Vec}(TX)) = V \otimes (T\Sigma T')$$

这里 T 为非随机矩阵.

Proof. 证明 6: 设 $C_{m \times n} = (c_{ij}) = (c_1, \dots, c_n), B = (b_1, \dots, b_m)$, 依定义

$$\begin{aligned} (C' \otimes A) \text{Vec}(B) &= \begin{pmatrix} c_{11}A & c_{21}A & \cdots & c_{m1}A \\ c_{12}A & c_{22}A & \cdots & c_{m2}A \\ \vdots & \vdots & & \vdots \\ c_{1n}A & c_{2n}A & \cdots & c_{mn}A \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \\ &= \begin{pmatrix} A\Sigma c_{j1}b_j \\ A\Sigma c_{j2}b_j \\ \vdots \\ A\Sigma c_{jn}b_j \end{pmatrix} = \begin{pmatrix} ABc_1 \\ ABc_2 \\ \vdots \\ ABc_n \end{pmatrix} = \text{Vec}(ABC) \end{aligned}$$

■

16.0.3 矩阵微商

Definition 16.0.9 假设 X 为 $n \times m$ 矩阵, $y = f(X)$ 为 X 的一个实值函数, 矩阵

$$\frac{\partial y}{\partial X} \triangleq \left(\begin{array}{cccc} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1m}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2m}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y}{\partial x_{n1}} & \frac{\partial y}{\partial x_{n2}} & \cdots & \frac{\partial y}{\partial x_{nm}} \end{array} \right)_{n \times m}$$

称为 y 对 X 的微商.

没有特殊声明, 以下都假定矩阵 X 中的 mn 个变量 $x_{ij}, i = 1, 2, \dots, n, j = 1, \dots, m$ 都

是独立自变量.

为此, 我们来回顾,

1. 一元微积分中的导数 (标量对标量的导数) 与微分有联系: $df = f'(x)dx$
2. 多元微积分中的梯度 (标量对向量的导数) 也与微分有联系: $df = \sum_{i=1}^n \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f}{\partial x}^T dx$,
这里第一个等号是全微分公式, 第二个等号表达了梯度与微分的联系: 全微分 df 是梯度向量 $\frac{\partial f}{\partial x}$ ($n \times 1$) 与微分向量 dx ($n \times 1$) 的内积,

受此启发, 我们将矩阵导数与微分建立联系: $df = \sum_{i=1}^m \sum_{j=1}^n \frac{\partial f}{\partial X_{ij}} dX_{ij} = \text{tr}\left(\frac{\partial f}{\partial X} dX\right)$. 其中 tr 代表迹 (trace) 是方阵对角线元素之和, 满足性质: 对尺寸相同的矩阵 A, B , $\text{tr}(A^T B) = \sum_{i,j} A_{ij} B_{ij}$, 即 $\text{tr}(A^T B)$ 是矩阵 A, B 的内积。与梯度相似, 这里第一个等号是全微分公式, 第二个等号表达了矩阵导数与微分的联系: 全微分 df 是导数 $\frac{\partial f}{\partial X}$ ($m \times n$) 与微分矩阵 dX ($m \times n$) 的内积。

Proposition 16.0.41 — 矩阵求导的性质.

1. 加减法: $d(X \pm Y) = dX \pm dY$; 矩阵乘法: $d(XY) = (dX)Y + XdY$: 转置: $d(X^T) = (dX)^T$; 迹: $d\text{tr}(X) = \text{tr}(dX)$
2. 逆: $dX^{-1} = -X^{-1}dXX^{-1}$ 。此式可在 $XX^{-1} = I$ 两侧求微分来证明。
3. 行列式: $d|X| = \text{tr}(X^\# dX)$, 其中 $X^\#$ 表示 X 的伴随矩阵, 在 X 可逆时又可以写作 $d|X| = |X|\text{tr}(X^{-1}dX)$ 。此式可用 Laplace 展开来证明, 详见张贤达《矩阵分析与应用》第 279 页。
4. 逐元素乘法: $d(X \odot Y) = dX \odot Y + X \odot dY$, \odot 表示尺寸相同的矩阵 X, Y 逐元素相乘。
5. 逐元素函数: $d\sigma(X) = \sigma'(X) \odot dX$, $\sigma(X) = [\sigma(X_{ij})]$ 是逐元素标量函数运算, $\sigma'(X) = [\sigma'(X_{ij})]$ 是逐元素求导数。例如 $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$, $d\sin(X) = \begin{bmatrix} \cos X_{11} dX_{11} & \cos X_{12} dX_{12} \\ \cos X_{21} dX_{21} & \cos X_{22} dX_{22} \end{bmatrix} = \cos(X) \odot dX$

由于在求导时, 矩阵的范数时 trace, 故介绍迹的相关性质

Proposition 16.0.42

1. 标量套上迹: $a = \text{tr}(a)$
2. 转置: $\text{tr}(A^T) = \text{tr}(A)$ 。
3. 线性: $\text{tr}(A \pm B) = \text{tr}(A) \pm \text{tr}(B)$ 。
4. 矩阵乘法交换: $\text{tr}(AB) = \text{tr}(BA)$, 其中 A 与 B^T 尺寸相同。两侧都等于 $\sum_{i,j} A_{ij} B_{ji}$
5. 矩阵乘法/逐元素乘法交换: $\text{tr}(A^T(B \odot C)) = \text{tr}((A \odot B)^T C)$, 其中 A, B, C 尺寸相同。两侧都等于 $\sum_{i,j} A_{ij} B_{ij} C_{ij}$

■ **Example 16.1** 对于不是 X 来求导的情况, 例如最常见的情形是 $Y = AXB$, 此时 $df = \text{tr}\left(\frac{\partial f}{\partial Y} dY\right) = \text{tr}\left(\frac{\partial f}{\partial Y} AdXB\right) = \text{tr}\left(B \frac{\partial f}{\partial Y} AdX\right) = \text{tr}\left(\left(A^T \frac{\partial f}{\partial Y} B^T\right)^T dX\right)$, 可得到 $\frac{\partial f}{\partial X} = A^T \frac{\partial f}{\partial Y} B^T$ 。注意这里 $dY = (dA)XB + AdXB + AXdB = AdXB$, 由于 A, B 是常量 $dA = 0, dB = 0$, 以及我们使用矩阵乘法交换的迹技巧交换了 $\frac{\partial f}{\partial Y} AdX$ 与 B 。

■ **Example 16.2** $f = a^T X b$, 求 $\frac{\partial f}{\partial X}$ 其中 a 是 $m \times 1$ 列向量, X 是 $m \times n$ 矩阵, b 是 $n \times 1$ 列向量, f 是标量。

解: 先使用矩阵乘法法则求微分, $df = da^T X b + a^T dX b + a^T X db = a^T dX b$, 注意这里的 a, b 是常量, $da = \mathbf{0}, db = \mathbf{0}$ 。由于 df 是标量, 它的迹等于自身, $df = \text{tr}(df)$, 套上迹并做矩阵乘法交换: $df = \text{tr}(a^T dX b) = \text{tr}(ba^T dX) = \text{tr}((ab^T)^T dX)$, 注意这里我们根据 $\text{tr}(AB) = \text{tr}(BA)$ 交换了 $a^T dX$ 与 b 。对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X} dX\right)$, 得到 $\frac{\partial f}{\partial X} = ab^T$

注意: 这里不能用 $\frac{\partial f}{\partial X} = a^T \frac{\partial X}{\partial X} b = ?$, 导数与矩阵乘法的交换是不合法则的运算 (而微分是合法的)。有些资料在计算矩阵导数时, 会略过求微分这一步, 这是逻辑上解释不通的。

■ **Example 16.3** $f = a^T \exp(Xb)$, 求 $\frac{\partial f}{\partial X}$, 其中 a 是 $m \times 1$ 列向量, X 是 $m \times n$ 矩阵, b 是

$n \times 1$ 列向量, \exp 表示逐元素求指数, f 是标量。

解: 先使用矩阵乘法、逐元素函数法则求微分: $df = a^T(\exp(Xb) \odot (dXb))$, 再套上述并做交换: $df = \text{tr}(a^T(\exp(Xb) \odot (dXb))) = \text{tr}((a \odot \exp(Xb))^T dXb) = \text{tr}(b(a \odot \exp(Xb))^T dX) = \text{tr}(((a \odot \exp(Xb))b^T)^T dX)$, 注意这里我们先根据 $\text{tr}(A^T(B \odot C)) = \text{tr}((A \odot B)^T C)$ 交换了 $a, \exp(Xb)$ 与 dXb , 再根据 $\text{tr}(AB) = \text{tr}(BA)$ 交换了 $(a \odot \exp(Xb))^T dX$ 与 b 。对照导数与微分的联系 $df = \text{tr}\left(\frac{\partial f}{\partial X} dX\right)$, 得到 $\frac{\partial f}{\partial X} = (a \odot \exp(Xb))b^T$

■ **Example 16.4** $f = \text{tr}(Y^T MY)$, $Y = \sigma(WX)$, 求 $\frac{\partial f}{\partial X}$, 其中 W 是 $l \times m$ 矩阵, X 是 $m \times n$ 矩阵, Y 是 $l \times n$ 矩阵, M 是 $l \times l$ 对称矩阵, σ 是逐元素函数, f 是标量。

解: 先求 $\frac{\partial f}{\partial Y}$, 求微分, 使用矩阵乘法、转置法则: $df = \text{tr}((dY)^T MY) + \text{tr}(Y^T M dY) = \text{tr}(Y^T M^T dY) + \text{tr}(Y^T M dY) = \text{tr}(Y^T(M + M^T)dY)$, 对照导数与微分的联系, 得到 $\frac{\partial f}{\partial Y} = (M + M^T)Y = 2MY$, 注意这里 M 是对称矩阵。为求 $\frac{\partial f}{\partial X}$ 写出 $df = \text{tr}\left(\frac{\partial f}{\partial Y} dY\right)$, 再将 dY 用 dX 表示出来代入, 并使用矩阵乘法/逐元素乘法交换 $df = \text{tr}\left(\frac{\partial f}{\partial Y} (\sigma'(WX) \odot (WdX))\right) = \text{tr}\left(\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right)^T WdX\right)$, 对照导数与微分的联系, 得到 $\frac{\partial f}{\partial X} = W^T\left(\frac{\partial f}{\partial Y} \odot \sigma'(WX)\right) = W^T((2M\sigma(WX)) \odot \sigma'(WX))$.

■ **Example 16.5 — 线性回归.** $l = \|Xw - y\|^2$, 求 w 的最小二乘估计, 即求 $\frac{\partial l}{\partial w}$ 的零点。其中 y 是 $m \times 1$ 列向量, X 是 $m \times n$ 矩阵, w 是 $n \times 1$ 列向量, l 是标量。

解: 这是标量对向量的导数, 不过可以把向量看做矩阵的特例。先将向量模平方改写成向量与自身的内积: $l = (Xw - y)^T(Xw - y)$, 求微分, 使用矩阵乘法、转置等法则: $dl = (Xdw)^T(Xw - y) + (Xw - y)^T(Xdw) = 2(Xw - y)^T X dw$, 注意这里 $X dw$ 和 $Xw - y$ 是向量, 两个向量的内积满足 $u^T v = v^T u$. 对照导数与微分的联系 $dl = \frac{\partial l}{\partial w}^T dw$, 得到 $\frac{\partial l}{\partial w} = 2X^T(Xw - y) \cdot \frac{\partial l}{\partial w} = \mathbf{0}$ 即 $X^T X w = X^T y$, 得到 w 的最小二乘估计为 $w = (X^T X)^{-1} X^T y$

■ **Example 16.6 — 例子.** 1. 设 a, x 均为 $n \times 1$ 向量, $y = a'x$, 则 $\frac{\partial y}{\partial x} = a$

2. 设 $A_{n \times n}$ 对称, $x_{n \times 1}, y = x'Ax$, 则 $\frac{\partial y}{\partial x} = 2Ax$

3. 记矩阵 $X_{m \times m}$ 的元素 x_{ij} 的代数余子式为 X_{ij} , 则

$$\frac{\partial |X|}{\partial X} = (X_{ij})_{m \times m} = |X| (X^{-1})'$$

结果容易从 $|X| = \sum_{j=1}^m x_{ij} X_{ij}$ 和 X_{ij} 中不包含 x_{ij} 导出。

Theorem 16.0.43 设 Y 和 X 分别为 $m \times n, p \times q$ 矩阵, Y 的每个元素 y_{ij} 是 X 元素的函数, 又 $u = u(Y)$, 则

$$\frac{\partial u}{\partial X} = \sum_{ij} \left(\frac{\partial u}{\partial Y} \right)_{ij} \frac{\partial (Y)_{ij}}{\partial X}$$

其中 $\left(\frac{\partial u}{\partial Y} \right)_{ij}$ 表示矩阵 $\frac{\partial u}{\partial Y}$ 的 (i, j) 元, $(Y)_{ij}$ 表示矩阵 Y 的 (i, j) 元 y_{ij}

结论容易从复合函数的求导法则

$$\frac{\partial u}{\partial x_{kl}} = \sum_{ij} \frac{\partial u}{\partial y_{ij}} \frac{\partial y_{ij}}{\partial x_{kl}} = \sum_{ij} \left(\frac{\partial u}{\partial Y} \right)_{ij} \cdot \frac{\partial (Y)_{ij}}{\partial x_{kl}}$$

得到。

■ **Example 16.7** 1.

$$\frac{\partial|Y|}{\partial X} = \sum_{ij} \left(\frac{\partial|Y|}{\partial Y} \right)_{ij} \frac{\partial(Y)_{ij}}{\partial X} = \sum_{ij} (Y_{kl})_{ij} \frac{\partial(Y)_{ij}}{\partial X} = \sum_{ij} |Y| (Y^{-1})'_{ij}$$

$\frac{\partial(Y)_{ij}}{\partial X}$, 其中 Y_{kl} 表示矩阵 Y 的元素 y_{kl} 的代数余子式, (Y_{kl}) 表示由这些代数余子式组成的矩阵.

2.

$$\frac{\partial \ln|Y|}{\partial X} = \frac{1}{|Y|} \frac{\partial|Y|}{\partial X} = \sum_{ij} (Y^{-1})'_{ij} \frac{\partial(Y)_{ij}}{\partial X}$$

我们用 $E_{ij}(m \times n)$ 表示 (i, j) 元为 1, 其余元素全为零的矩阵. 在不致引起混淆的情况下, 常常把阶数 $m \times n$ 略去. 利用这个记号, 则有

$$\frac{\partial y}{\partial X_{m \times n}} = \left(\frac{\partial y}{\partial x_{ij}} \right) = \sum_{ij} E_{ij}(m \times n) \frac{\partial y}{\partial x_{ij}}$$

$$3. \frac{\partial|AXB|}{\partial X} = |AXB| A' ((AXB)^{-1})' B'$$

Proof. 记 $Y = AXB$, 利用之前的例子, 有

$$\frac{\partial|AXB|}{\partial X} = \sum_{ij} |Y| (Y^{-1})'_{ij} \frac{\partial(Y)_{ij}}{\partial X} = |AXB| \sum_{ij} ((AXB)^{-1})'_{ij} \frac{\partial(AXB)_{ij}}{\partial X}$$

因为

$$\frac{\partial(AXB)_{ij}}{\partial X} = \left(\frac{\partial(AXB)_{ij}}{\partial x_{kl}} \right) = (a_{ik} b_{lj}) = A'E_{ij}B'$$

于是

$$\begin{aligned} \frac{\partial|AXB|}{\partial X} &= |AXB| \sum_{ij} ((AXB)^{-1})'_{ij} \cdot A'E_{ij}B' \\ &= |AXB| A' \left[\sum_{ij} ((AXB)^{-1})'_{ij} E_{ij} \right] B' = |AXB| A' ((AXB)^{-1})' B' \end{aligned}$$

最后一式利用了

$$\sum_{ij} (A)_{ij} E_{ij} = \sum_{ij} a_{ij} E_{ij} = A$$

■

$$4. \frac{\partial \ln|AXB|}{\partial X} = A' ((AXB)^{-1})' B'$$

Theorem 16.0.44 — 转换定理. 设 X 和 Y 分别为 $n \times m, p \times q$ 矩阵, A, B, C, D 分别为 $p \times m, n \times q, p \times n, m \times q$ 矩阵 (可以是 X 的函数), 则下列两条是等价的

1. $\frac{\partial Y}{\partial x_{ij}} = AE_{ij}(m \times n)B + CE'_{ij}(m \times n)D, \quad i = 1, \dots, m, j = 1, \dots, n$
2. $\frac{\partial(Y)_{ij}}{\partial X} = A'E_{ij}(p \times q)B' + DE'_{ij}(p \times q)C, \quad i = 1, \dots, p, j = 1, \dots, q$

这里

$$\frac{\partial Z}{\partial t} = \begin{pmatrix} \frac{\partial z_{11}}{\partial t} & \frac{\partial z_{12}}{\partial t} & \cdots & \frac{\partial z_{1n}}{\partial t} \\ \frac{\partial z_{21}}{\partial t} & \frac{\partial z_{22}}{\partial t} & \cdots & \frac{\partial z_{2n}}{\partial t} \\ \vdots & \vdots & & \vdots \\ \frac{\partial z_{m1}}{\partial t} & \frac{\partial z_{m2}}{\partial t} & \cdots & \frac{\partial z_{mn}}{\partial t} \end{pmatrix}_{m \times n} \quad (16.19)$$

Proof. $Z_{m \times n} = (z_{ij}(t))$, 它是矩阵 $Z = (z_{ij}(t))$ 对自变量 t 的微商. 证明 记 $e'_i = (0, \dots, 0, 1, 0, \dots, 0)$, 即 e_i 是第 i 个元素为 1, 其余元素全为零的向量. 则 $E_{ij} = e_i e'_j$. 首先注意到

$$\begin{aligned} e'_k (AE_{ij}B + CE'_{ij}D) e_l &= e'_k A e_i e'_j B e_l + e'_k C e_j e'_l D e_l \\ &= e'_i A e_k e'_l B e_j + e'_l D e_l e'_k C e_j \\ &= e'_i (A e_k e'_l B + D e_l e'_k C) e_j \\ &= e'_i (A' E_{kl} B' + D E_{lk} C) e_j \end{aligned}$$

若 (1) 成立, 则

$$\left(\frac{\partial Y}{\partial x_{ij}} \right)_{kl} = e'_i (AE_{ij}B + CE_{ij}D) e_l = e'_i (A' E_{kl} B' + D E_{lk} C) e_j$$

但是, 由 (16.19), 有

$$\left(\frac{\partial Y}{\partial x_{ij}} \right)_{kl} = \left(\frac{\partial y_{kl}}{\partial x_{ij}} \right) = \left(\frac{\partial y_{kl}}{\partial X} \right)_{ij}$$

于是

$$\left(\frac{\partial y_{kl}}{\partial X} \right)_{ij} = e'_i (A' E_{kl} B' + D E_{lk} C) e_j, \quad \text{对一切 } i, j,$$

此即 (2). 同法可从 (2) \Rightarrow (1). ■

Corollary 16.0.45 设 X, Y 分别为 $m \times n, p \times q$ 矩阵, A_k, B_k, C_k, D_k 分别为 $p \times m, n \times q, p \times n, m \times q$ 矩阵 (可以是 X 的函数), 则下列两条是等价的.

1. $\frac{\partial Y}{\partial x_{ij}} = \sum_k A_k E_{ij} (m \times n) B_k + \sum_l C_l E'_{ij} (m \times n) D_l, \quad i = 1, \dots, m, j = 1, \dots, n$
2. $\frac{\partial (Y)_{ij}}{\partial X} = \sum_k A'_k E_{ij} (p \times q) B'_k + \sum_l D_l E'_{ij} (p \times q) C_l, \quad i = 1, \dots, p, j = 1, \dots, q$

转换定理是求矩阵微商的一个重要工具, 从定理 16.0.43 我们看到, 为求 $\frac{\partial u}{\partial X}$ 需要求 $\frac{\partial (Y)_{ij}}{\partial X}$, 但在很多情况下, 这是困难的. 转换定理给出了利用 $\frac{\partial Y}{\partial x_{ij}}$ 求 $\frac{\partial (Y)_{ij}}{\partial X}$ 的途径, 前者往往是比较容易的.

■ **Example 16.8** $\frac{\partial \ln|X'AX|}{\partial X} = 2AX(X'AX)^{-1}$, 其中 A 对称.

Proof. 依定理 16.0.43, 及之前的例子, 得

$$\begin{aligned} \frac{\partial \ln|X'AX|}{\partial X} &= \sum_{i,j} \frac{1}{|X'AX|} |X'AX| \left((X'AX)^{-1} \right)'_{ij} \cdot \frac{\partial (X'AX)_{ij}}{\partial X} \\ &= \sum_{i,j} \left((X'AX)^{-1} \right)'_{ij} \frac{\partial (X'AX)_{ij}}{\partial X} \end{aligned} \quad (16.20)$$

因为

$$\frac{\partial(X'AX)}{\partial x_{ij}} = \frac{\partial X'}{\partial x_{ij}}(AX) + X' \cdot \frac{\partial AX}{\partial x_{ij}} = E'_{ij}AX + X'AE_{ij} \quad (16.21)$$

由转换定理, 应得

$$\frac{\partial(X'AX)_{ij}}{\partial X} = AXE'_{ij} + AXE_{ij}$$

代入 (16.20), 得到

$$\begin{aligned} \frac{\partial \ln|X'AX|}{\partial X} &= \sum_{i,j} \left((X'AX)^{-1} \right)'_{ij} (AXE'_{ij} + AXE_{ij}) \\ &= AX \left[\sum_{i,j} \left((X'AX)^{-1} \right)'_{ij} E'_{ij} + \sum_{i,j} \left((X'AX)^{-1} \right)'_{ij} E_{ij} \right] \\ &= 2AX (X'AX)^{-1} \end{aligned}$$

■

■ Example 16.9

$$\frac{\partial \text{tr}(XAX')}{\partial X} = X(A+A')$$

Proof.

$$\text{左边} = \frac{\partial \text{tr}(XAX')}{\partial X} = \sum_i \frac{\partial(XAX')_{ii}}{\partial X} \quad (16.22)$$

与 (16.21) 同样的方法可推得

$$\frac{\partial(XAX')}{\partial x_{ij}} = E_{ij}AX' + XAE'_{ij}$$

由转化定理, 有

$$\frac{\partial(XAX')_{ij}}{\partial X} = E_{ij}XA' + E'_{ij}XA$$

代入 (16.22) 得

$$\frac{\partial \text{tr}(XAX')}{\partial X} = \sum_i (E_{ii}XA' + E'_{ii}XA) = X(A+A')$$

证毕。 ■

用完全同样的方法可以证明以下结果.

■ Example 16.10 1.

$$\frac{\partial \text{tr}(AXB)}{\partial X} = A'B', \quad \text{特别 } \frac{\partial \text{tr}(AX)}{\partial X} = A'$$

2.

$$\frac{\partial \text{tr}(X'AXB)}{\partial X} = AXB + A'XB'$$

上面的讨论都是假定 X 的分量是独立自变量、然而，有时会碰到 X 的分量不独立的情况。其中较重要的是， X 为对称阵，这时 $x_{ij} = x_{ji}$ 。对这种情况，矩阵微商公式略显复杂。

以下记 $\text{diag}(A) = \text{diag}(a_{11}, \dots, a_{nn})$

■ **Example 16.11** 设 X 为 $n \times n$ 对称阵，则

$$\frac{\partial |X|}{\partial X} = |X| (2X^{-1} - \text{diag}(X^{-1}))$$

Proof. 为求 $\frac{\partial |X|}{\partial x_{11}}, \frac{\partial |X|}{\partial x_{1j}}$ ，将 $|X|$ 按第一行展开，得

$$|X| = \sum_{j=1}^n x_{1j} X_{1j}$$

于是

$$\begin{aligned} \frac{\partial |X|}{\partial x_{11}} &= X_{11} \\ \frac{\partial |X|}{\partial x_{12}} &= X_{12} + x_{12} \frac{\partial X_{12}}{\partial x_{12}} + \frac{\partial}{\partial x_{12}} [x_{13} X_{13} + \dots + x_{1n} X_{1n}] \end{aligned} \quad (16.23)$$

若用 $X_{ij,kl}$ 表示 x_{ij} 的余子式中 (k,l) 元的代数余子式，将 X_{1j} 按第一行展开，得

$$\begin{aligned} X_{1j} &= x_{21} X_{1j,21} + x_{22} X_{1j,22} + \dots + x_{2j-1} X_{1j,2j-1} \\ &\quad + x_{2j+1} X_{1j,2j+1} + \dots + x_{2n} X_{1j,2n}, \quad j = 2, \dots, n \end{aligned}$$

因为 $X_{1j,2k}, j = 2, \dots, n, k = 1, \dots, n$ 都与 x_{21} 无关，所以

$$\sum_{j=2}^n x_{1j} X_{1j} = x_{21} \sum_{j=2}^n x_{1j} X_{1j,21} + (\text{与 } x_{21} \text{ 无关的项})$$

代入 (16.23)，我们得到

$$\frac{\partial |X|}{\partial x_{12}} = X_{12} + \sum_{j=2}^n x_{1j} X_{1j,21} = 2X_{12}$$

同理

$$\frac{\partial |X|}{\partial x_{ij}} = 2X_{ij}, \quad \frac{\partial |X|}{\partial x_{ii}} = X_{ii}$$

结论得证。 ■

■ **Example 16.12** 设 X 为对称阵，则

$$\frac{\partial \ln |X|}{\partial X} = 2X^{-1} - \text{diag}(X^{-1})$$

■ **Example 16.13** 设 X 为对称阵，则

$$\frac{\partial \text{tr}(AX)}{\partial X} = A + A' - \text{diag}(A)$$

Proof. 因对任一矩阵 A , 总有 $A = \sum_{i,j} a_{ij} E_{ij}$, 所以

$$\frac{\partial \text{tr}(AX)}{\partial X} = \sum_{i,j} E_{ij} \frac{\partial \text{tr}(AX)}{\partial x_{ij}} = \sum_{i,j} E_{ij} \text{tr}\left(\frac{\partial AX}{\partial x_{ij}}\right)$$

从 $x = X'$, 有

$$\text{tr}\left(\frac{\partial AX}{\partial x_{ij}}\right) = \begin{cases} a_{ii}, & i = j \\ a_{ij} + a_{ji}, & i \neq j \end{cases}$$

代入上式, 得

$$\frac{\partial \text{tr}(AX)}{\partial X} = \sum_i a_{ii} E_{ii} + \sum_{i \neq j} (a_{ij} + a_{ji}) E_{ij} = A + A' - \text{diag}(A)$$

■

$y = f(X)$	$\frac{\partial y}{\partial X}$
$a'x$	a
$x'Ax$	$2Ax$
$ X $	$\begin{cases} X (X^{-1})', & X \text{ 对称} \\ X (2X^{-1} - \text{diag}(X^{-1})) & \end{cases}$
$ AXB $	$ AXB A'((AXB)^{-1})'B'$
$\ln AXB $	$A'((AXB)^{-1})'B'$
$\ln X'AX $	$2AX(X'AX)^{-1}$ (A 对称)
$\text{tr}(XAX')$	$X(A+A')$
$\text{tr}(AXB)$	$A'B'$
$\text{tr}(X'AXB)$	$AXB + A'XB'$
$\ln X $	$2X^{-1} - \text{diag}(X^{-1})$ (X 对称)
$\text{tr}(AX)$	$A + A' - \text{diag}(A)$ (X 对称)

■ Example 16.14

$$\frac{\partial}{\partial t} \ln|A(t)| = \text{tr}\left(A^{-1}(t) \frac{\partial A(t)}{\partial t}\right)$$

其中 $A(t)$ 为矩阵, t 为标量.

$$\begin{aligned} \frac{\partial}{\partial t} \ln|A(t)| &= |A(t)|^{-1} \frac{\partial|A(t)|}{\partial t} = \frac{1}{|A(t)|} \sum_i \sum_{j \leq j} \frac{\partial|A(t)|}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial t} \\ &= \frac{1}{|A(t)|} \sum_i \sum_{j \leq j} (2 - \sigma_{ij}) |A_{ij}| \frac{\partial a_{ij}}{\partial t} = \frac{1}{|A(t)|} \sum_i \sum_j |A_{ij}| \frac{\partial a_{ij}}{\partial t} \\ &= \sum_i \sum_j \frac{|A_{ij}|}{|A|} \frac{\partial a_{ij}}{\partial t} = \sum_i \sum_j a^{ij} \frac{\partial a_{ij}}{\partial t} \\ &= \text{tr}(A^{-1})' \frac{\partial A}{\partial t} = \text{tr}\left(A^{-1} \frac{\partial A}{\partial t}\right) \end{aligned}$$

其中 $A^{-1} = (a^{ij})$

■ Example 16.15

$$\frac{\partial A^{-1}(t)}{\partial t} = -A^{-1}(t) \frac{\partial A^{-1}(t)}{\partial t} A^{-1}(t)$$

■ Example 16.16

$$\frac{\partial A^{-1}(t)}{\partial t} = -A^{-1}(t) \frac{\partial A(t)}{\partial t} A^{-1}(t)$$

Proof. 由于 $A(t)A^{-1}(t) = I$, 故有

$$\frac{\partial A(t)}{\partial t} A^{-1}(t) + A(t) \frac{\partial A^{-1}(t)}{\partial t} = 0$$

因此

$$\frac{\partial A^{-1}(t)}{\partial t} = -A^{-1}(t) \frac{\partial A(t)}{\partial t} A^{-1}(t)$$

■

最后我们简要提一下矩阵对矩阵的微商. 设 Y 和 X 分别为 $m \times n, p \times q$ 矩阵, 且 Y 的元素 y_{ij} 为 X 的函数. 记

$$\frac{\partial Y}{\partial X} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x_{11}} & \frac{\partial y_{11}}{\partial x_{12}} & \cdots & \frac{\partial y_{11}}{\partial x_{pq}} \\ \frac{\partial y_{12}}{\partial x_{11}} & \frac{\partial y_{12}}{\partial x_{12}} & \cdots & \frac{\partial y_{12}}{\partial x_{pq}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_{mn}}{\partial x_{11}} & \frac{\partial y_{mn}}{\partial x_{12}} & \cdots & \frac{\partial y_{mn}}{\partial x_{pq}} \end{pmatrix}$$

称为 Y 对 X 的微商. 容易看到

$$\frac{\partial Y}{\partial X} = \left(\text{Vec} \left(\frac{\partial Y}{\partial x_{11}} \right)', \text{Vec} \left(\frac{\partial Y}{\partial x_{12}} \right)', \dots, \text{Vec} \left(\frac{\partial Y}{\partial x_{pq}} \right)' \right)$$

它把求 $\frac{\partial Y}{\partial X}$ 转化为求 $\frac{\partial Y}{\partial x_{ij}}$, 在一些情况下, 这会带来不少方便.

■ Example 16.17 设 $Y = AXB$, 则 $\frac{\partial Y}{\partial X} = A \otimes B'$

Proof. 因为 $\frac{\partial Y}{\partial x_{ij}} = AE_{ij}B$, 于是

$$\text{Vec} \left(\frac{\partial Y}{\partial x_{ij}} \right)' = \text{Vec} (B'E_{ji}A') = (A \otimes B') \text{Vec} (E_{ji})$$

所以

$$\frac{\partial Y}{\partial X} = ((A \otimes B') \text{Vec} (E_{11}), \dots, (A \otimes B') \text{Vec} (E_{pq})) = A \otimes B'$$

证毕. ■

若 Y, X, A, B 分别为 $n \times m, n \times m, n \times n, m \times m$ 矩阵, 则变换 $Y = AXB$ 的 Jacobi 行列式为

$$\left| \frac{\partial Y}{\partial X} \right| = |A \otimes B'| = |A|^m |B|^n$$

16.0.4 多元正态

Theorem 16.0.46 设 A 为 $m \times n$ 非随机矩阵, X 和 b 分别为 $n \times 1$ 和 $m \times 1$ 随机向量, 记 $Y = AX + b$, 则

$$E(Y) = AE(X) + E(b)$$

Definition 16.0.10 n 维随机向量 X 的协方差阵定义为

$$\text{Cov}(X) = E[(X - EX)(X - EX)']$$

这是一个 $n \times n$ 对称阵, 它的 (i, j) 元为 $\text{Cov}(X_i, X_j) = E[(X_i - EX_i)(X_j - EX_j)]$

Proposition 16.0.47 $\text{tr Cov}(X) = \sum_{i=1}^n \text{Var}(X_i)$, 这里 $\text{tr } A$ 表示方阵 A 的迹, 即对角元素之和.

Theorem 16.0.48 设 X 为 $n \times 1$ 随机向量, 则它的协方差阵必为半正定的对称阵

Proof. 对称性是显然的. 下面证明它是半正定的. 事实上, 对任意 $n \times 1$ 非随机向量 c , 考虑随机变量 $Y = c'X$ 的方差. 根据定义, 我们有

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(c'X) = E[(c'X - E(c'X))^2] \\ &= E[(c'X - E(c'X))(c'X - E(c'X))] \\ &= c'E[(X - EX)(X - EX)']c \\ &= c'\text{Cov}(X)c \end{aligned}$$

因为左端总是非负的, 于是对一切 c , 右端也是非负的. 根据定义, 这说明矩阵 $\text{Cov}(X)$ 是半正定的. 定理证毕. ■

Theorem 16.0.49 设 A 为 $m \times n$ 阵, X 为 $n \times 1$ 随机向量, $Y = AX$, 则 $\text{Cov}(Y) = A\text{Cov}(X)A'$

Proof.

$$\begin{aligned} \text{Cov}(Y) &= E[(Y - EY)(Y - EY)'] \\ &= E[(AX - E(AX))(AX - E(AX))'] \\ &= AE[(X - EX)(X - EX)']A' \\ &= A\text{Cov}(X)A' \end{aligned}$$

设 X 和 Y 分别为 $n \times 1, m \times 1$ 随机向量, 它们的协方差阵定义为

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)']$$

Theorem 16.0.50 设 X 和 Y 分别为 $n \times 1, m \times 1$ 随机向量, A 和 B 分别为 $p \times n, q \times m$ 非随机矩阵, 则

$$\text{Cov}(AX, BY) = A\text{Cov}(X, Y)B'$$

Proof.

$$\begin{aligned}\text{Cov}(AX, BY) &= E[(AX - E(AX))(BY - E(BY))'] \\ &= AE[(X - EX)(Y - EY)']B' \\ &= A\text{Cov}(X, Y)B'\end{aligned}$$

■

Definition 16.0.11 — 二次型. 假设 $X = (X_1, \dots, X_n)'$ 为 $n \times 1$ 随机向量, A 为 $n \times n$ 对称阵, 则随机变量

$$X'AX = \sum_{i=1}^n \sum_{j=1}^n a_{ij}X_iX_j$$

称为 X 的二次型.

Theorem 16.0.51 设 $E(X) = \mu, \text{Cov}(X) = \Sigma$, 则

$$E(X'AX) = \mu'A\mu + \text{tr}(A\Sigma) \quad (16.24)$$

Proof. 因为

$$\begin{aligned}X'AX &= (X - \mu + \mu)'A(X - \mu + \mu) \\ &= (X - \mu)'A(X - \mu) + \mu'A(X - \mu) + (X - \mu)'A\mu + \mu'A\mu\end{aligned} \quad (16.25)$$

利用定理 16.0.46, 有

$$E[\mu'A(X - \mu)] = E(\mu'AX) - \mu'A\mu = \mu'AE(X) - \mu'A\mu = 0$$

于是 (16.25) 式中第二、三两项的均值都等于零. 为了证明 (16.24), 只需证明

$$E[(X - \mu)'A(X - \mu)] = \text{tr}(A\Sigma)$$

注意到

$$E[(X - \mu)'A(X - \mu)] = E[\text{tr}(X - \mu)'A(X - \mu)]$$

利用矩阵迹的性质: $\text{tr}(AB) = \text{tr}(BA)$, 并交换求均值和求迹的次序, 上式变为

$$\begin{aligned}E[(X - \mu)'A(X - \mu)] &= E[\text{tr}(X - \mu)'A(X - \mu)] \\ &= E\text{tr}[A(X - \mu)(X - \mu)'] \\ &= \text{tr}AE[(X - \mu)(X - \mu)'] \\ &= \text{tr}(A\Sigma)\end{aligned}$$

定理证毕. ■



(在定理证明中, 我们应用了一个很重要的技巧. 这就是, 首先注意到二次型 $(X - \mu)'A(X - \mu)$ 的迹就是它本身, 然后利用迹的可交换性 $\text{tr}(AB) = \text{tr}(BA)$, 交换 $A(X - \mu)$ 与 $(X - \mu)'$ 的位置, 最后再交换求 $E(\cdot)$ 和 $\text{tr}(\cdot)$ 的次序. 这样一来, 把求 $E[(X - \mu)'A(X - \mu)]$ 的问题归结为求协方差阵 $E[(X - \mu)(X - \mu)'] = \Sigma$. 这个技巧在后面的讨论中会多次用到.)

Corollary 16.0.52 在定理 16.0.51 的假设条件下,

1. 若 $\mu = 0$, 则 $E(X'AX) = \text{tr}(A\Sigma)$
2. 若 $\Sigma = \sigma^2 I$, 则 $E(X'AX) = \mu'A\mu + \sigma^2 \text{tr}(A)$
3. 若 $\mu = 0, \Sigma = I$, 则 $E(X'AX) = \text{tr}(A)$

■ **Example 16.18** 假设一维总体的均值为 μ , 方差为 σ^2 . X_1, \dots, X_n 为从此总体中抽取的随机样本, 试求样本方差

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

的均值, 这里 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

Proof. 记 $Q = (n-1)S^2, X = (X_1, \dots, X_n)'$. 我们首先把 Q 表示为 X 的一个二次型. 用 $\mathbf{1}_n$ (在不会引起误解时也常用 $\mathbf{1}$) 表示所有元素为 1 的 n 维向量, 则 $E(X) = \mu \mathbf{1}_n, \text{Cov}(X) = \sigma^2 I_n$. 另外

$$\begin{aligned} \bar{X} &= \frac{1}{n} \mathbf{1}' X \\ X - \bar{X} \mathbf{1} &= X - \frac{1}{n} \mathbf{1} \mathbf{1}' X = (I_n - \frac{1}{n} \mathbf{1} \mathbf{1}') X = CX \end{aligned}$$

这里 $C = I_n - \frac{1}{n} \mathbf{1} \mathbf{1}'$, 这是一个对称幂等阵, 即 $C^2 = C, C' = C$. 于是

$$Q = \sum_{i=1}^n (X_i - \bar{X})^2 = (X - \bar{X} \mathbf{1})'(X - \bar{X} \mathbf{1}) = (CX)'CX = X'CX \quad (16.26)$$

应用定理 16.0.51, 得

$$E(Q) = (E(X))'C(E(X)) + \sigma^2 \text{tr}(C) = \mu^2 \mathbf{1}' C \mathbf{1} + \sigma^2 \text{tr}(C)$$

容易验证

$$C \mathbf{1} = 0, \text{tr}(C) = n-1$$

故有

$$E(Q) = \sigma^2(n-1)$$

因而

$$E(S^2) = \sigma^2$$

这就得到了所要得的结论. 这个例子证明了初等数理统计中的一个重要事实: 不管总体的具体分布形式如何, 样本方差总是总体方差的一个无偏估计. ■

现在我们先导出二次型 $X'AX$ 的方差公式.

Theorem 16.0.53 设随机变量 $X_i, i = 1, \dots, n$ 相互独立, $E(X_i) = \mu_i, \text{Var}(X_i) = \sigma_i^2, m_r = E(X_i - \mu_i)^r, r = 3, 4$. $A = (a_{ij})_{n \times n}$ 为对称阵. 记 $X' = (X_1, \dots, X_n), \mu' = (\mu_1, \dots, \mu_n)$, 则

$$\text{Var}(X'AX) = (m_4 - 3\sigma^4)a'a + 2\sigma^4 \text{tr}(A^2) + 4\sigma^2 \mu'A^2\mu + 4m_3 \mu'Aa$$

其中 $a' = (a_{11}, \dots, a_{nn})$, 即 A 的对角元组成的列向量.

Proof. 首先注意到

$$\text{Var}(X'AX) = E(X'AX)^2 - [E(X'AX)]^2 \quad (16.27)$$

由定理 16.0.51, 及 $E(X) = \mu$, $\text{Cov}(X) = \sigma^2 I$, 我们有

$$E(X'AX) = \mu'A\mu + \sigma^2 \text{tr}(A) \quad (16.28)$$

所以我们的主要问题是计算 (16.27) 中的第一项. 将 $X'AX$ 改写为

$$X'AX = (X - \mu)'A(X - \mu) + 2\mu'A(X - \mu) + \mu'A\mu$$

将其平方, 得到

$$\begin{aligned} (X'AX)^2 &= [(X - \mu)'A(X - \mu)]^2 + 4[\mu'A(X - \mu)]^2 \\ &\quad + (\mu'A\mu)^2 + 2\mu'A\mu[(X - \mu)'A(X - \mu) + 2\mu'A(X - \mu)] \\ &\quad + 4\mu'A(X - \mu)(X - \mu)'A(X - \mu) \end{aligned}$$

令 $Z = X - \mu$, 则 $E(Z) = 0$. 再次利用定理 16.0.51, 推得

$$E(X'AX)^2 = E(Z'AZ)^2 + 4E(\mu'AZ)^2 + (\mu'A\mu)^2$$

$$+ 2\mu'A\mu(\sigma^2 \text{tr}(A)) + 4E[\mu'AZZ'AZ]$$

下面逐个计算上式所含的每个均值. 由

$$(Z'AZ)^2 = \sum_i \sum_j \sum_k \sum_l a_{ij}a_{kl}Z_iZ_jZ_kZ_l$$

及 Z_i 的独立性导出的事实:

$$E(Z_iZ_jZ_kZ_l) = \begin{cases} m_4, & \text{若 } i = j = k = l \\ \sigma^4, & \text{若 } i = j, k = l; i = k, j = l; i = l, j = k \\ 0, & \text{其它,} \end{cases}$$

便有

$$\begin{aligned} E(Z'AZ)^2 &= m_4 \left(\sum_{i=1}^n a_{ii}^2 \right) + \sigma^4 \left(\sum_{i \neq k} a_{ii}a_{kk} + \sum_{i \neq j} a_{ij}^2 + \sum_{i \neq j} a_{ij}a_{ji} \right) \\ &= (m_4 - 3\sigma^4)a'a + \sigma^4[(\text{tr}(A))^2 + 2\text{tr}(A^2)] \end{aligned} \quad (16.29)$$

而

$$\begin{aligned} E(\mu'AZ)^2 &= E(\mu'AZ \cdot \mu'AZ) = E(Z'A\mu\mu'AZ) \\ &= \text{tr}(A\mu\mu'A) \cdot \sigma^2 = \sigma^2\mu'A^2\mu \end{aligned} \quad (16.30)$$

最后, 若记 $b = A\mu$, 则

$$E(\mu'AZ \cdot Z'AZ) = \sum_i \sum_j \sum_k b_i a_{jk} E(Z_iZ_jZ_k)$$

因为

$$E(Z_iZ_jZ_k) = \begin{cases} m_3, & \text{若 } i = j = k \\ 0, & \text{其它} \end{cases}$$

所以

$$E(\mu'AZ \cdot Z'AZ) = m_3 \sum_i b_i a_{ii} = m_3 b'a = m_3 \mu'Aa \quad (16.31)$$

将 (16.29) (16.31) 代入 (16.28), 再将 (16.27) 和 (16.28) 代入 (16.26), 便得到了要证的结果. 定理证毕. ■

多元正态分布

若随机变量 X 具有密度函数

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

则称 X 为具有均值 μ , 方差 σ^2 的正态随机变量, 记为 $N(\mu, \sigma^2)$. 推广到多元情形, 我们可以做如下定义:

Definition 16.0.12 设 n 维随机向量 $X = (X_1, \dots, X_n)$ 具有密度函数

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)} \quad (16.32)$$

其中 $x = (x_1, \dots, x_n)'$, $-\infty < x_i < +\infty, i = 1, \dots, n$, $\mu = (\mu_1, \dots, \mu_n)'$, Σ 是正定矩阵, 则称 X 为 n 维正态随机向量, 记为 $N_n(\mu, \Sigma)$. 在不致引起混淆的情况下, 也简记为 $N(\mu, \Sigma)$, 这里 μ 和 Σ 分别为分布参数.

我们首先证明, 其中的参数 μ 为 X 的均值向量, Σ 为 X 的协方差阵. 在 (16.32) 中, 用到了 Σ^{-1} , 因此我们假定 Σ 是正定阵, 记为 $\Sigma > 0$. 用 $\Sigma^{\frac{1}{2}}$ 记 Σ 的平方根阵, 记 $\Sigma^{-\frac{1}{2}}$ 为 $\Sigma^{\frac{1}{2}}$ 的逆矩阵, 即 $\Sigma^{-\frac{1}{2}} = (\Sigma^{\frac{1}{2}})^{-1}$. 定义

$$Y = \Sigma^{-\frac{1}{2}}(X - \mu) \quad (16.33)$$

故 $X = \Sigma^{\frac{1}{2}}Y + \mu$, 于是 Y 的密度函数为 $g(y) = f(\Sigma^{\frac{1}{2}}y + \mu) |J|$, 这里 J 为变换的 Jacobi 行列式,

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} = \left| \Sigma^{\frac{1}{2}} \right| = |\Sigma|^{\frac{1}{2}}$$

从 (16.32) 得到 Y 的密度函数

$$g(y) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}y'y} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{y_i^2}{2}} = \prod_{i=1}^n f(y_i)$$

这里

$$f(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y_i^2}{2}}$$

是标准正态分布的密度函数。这表明, Y 的 n 个分量的联合密度等于每个分量的密度函数的乘积. 于是, Y 的 n 个分量相互独立, 且 $Y_i \sim N(0, 1), i = 1, \dots, n$. 因而有 $E(Y) = 0, \text{Cov}(Y) = I$. 利用关系 $X = \Sigma^{\frac{1}{2}}Y + \mu$ 及定理 ?? 和定理 ??, 得 $E(X) = \mu, \text{Cov}(X) = \Sigma$. 这就完成了所要的证明. 从定义可以看出, 多元正态分布完全由它的均值向量 μ 和协方差阵 Σ 所确定. 特别, 若 $\mu = 0, \Sigma = I$, 此时称 X 服从标准正态分布 $N(0, I)$, 它的概率密度函数有如下形式

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}$$

容易证明, 它的 n 个分量 x_1, \dots, x_n 皆服从 $N(0, 1)$ 且相互独立. 上述定义是用概率密度函数定义分布的, 它需要假设协方差阵 $\Sigma > 0$. 下面我们引进多元正态分布的另一种定义.

Definition 16.0.13 设 X 为 n 维随机向量. 若存在 $n \times r$ 的列满秩矩阵 A , 使得 $X = AU + \mu$, 这里 $U = (u_1, \dots, u_r)', u_i \sim N(0, 1)$ 且相互独立, μ 为 $n \times 1$ 非随机向量, 则称 X 服从均值为 μ 、协方差阵为 $\Sigma = AA'$ 的多元正态向量, 记为 $X \sim N_n(\mu, \Sigma)$, 在不致引起混淆时简记为 $X \sim N(\mu, \Sigma)$

这个定义是由我国统计学先驱许宝绿先生提出的, 他把多元正态向量定义为若干个相互独立的一元标准正态分布随机变量的线性变换. 在这个定义中, Σ 可以是半正定的, 即 $|\Sigma| = 0$, 这时的分布称为奇异正态分布. 如果限制 $\Sigma > 0$, 则这个定义与定义一是等价的. 事实上, 从 (16.33) 及其后的证明我们可以把 X 表示为 $X = \Sigma^{\frac{1}{2}}Y + \mu$, 这里 $Y_i \sim N(0, 1), i = 1, \dots, n$ 独立. 据此式, 两种定义的等价性是显然的. 定义二不仅仅是把多元正态的定义推广到奇异正态的情形, 而且根据这种定义, 容易推导多元正态分布的一些性质.

Theorem 16.0.54 设 $X \sim N_n(\mu, \Sigma), \Sigma \geq 0, B$ 为 $m \times n$ 任意实矩阵, 则 $Y = BX \sim N(B\mu, B\Sigma B')$

Proof. 设 $\text{rk}(\Sigma) = r$, 根据定义二, 存在 $n \times r$ 矩阵 $A, \text{rk}(A) = r, X$ 可表示为

$$X = AU + \mu, \quad AA' = \Sigma, \quad U \sim N(0, I_r)$$

于是

$$Y = BAU + B\mu$$

再用定义二, 定理得证. 这个定理表明, 多元正态向量的任意线性变换仍为正态向量. ■

Corollary 16.0.55 设 $X \sim N_n(\mu, \Sigma), \Sigma > 0$, 则

$$Y = \Sigma^{-\frac{1}{2}}X \sim N\left(\Sigma^{-\frac{1}{2}}\mu, I_n\right)$$

注意, 这里 X 的诸分量可以是彼此相关且方差互不相等, 但经过变换过的 Y 的诸分量相互独立, 且方差皆为 1. 这个推论表明, 我们可以用一个线性变换把诸分量相关且方差不等的多元正态向量变换为多元标准正态向量.

Corollary 16.0.56 设 $X \sim N_n(\mu, \sigma^2 I), Q$ 为 $n \times n$ 正交阵, 则 $QX \sim N_n(Q\mu, \sigma^2 I)$. 本推论表明, 诸分量相互独立且具有等方差的正态向量, 经过正交变换后, 变为诸分量仍然相互独立且具有等方差的正态向量.

Theorem 16.0.57 — 推导多元正态的概率密度函数. 现在我们来求 $X \sim N(\mu, \Sigma), \Sigma \geq 0$ 的概率密度函数. 设 $\text{rk}(\Sigma) = r < n, Q = (Q_1 : Q_2)$ 为 Σ 的标准正交化特征向量组成的正交阵, Q_1 为 $n \times r$ 矩阵, 其 r 个列对应于非零特征根 $\lambda_1, \dots, \lambda_r, Q_2$ 为 $n \times (n-r)$, 其 $n-r$ 个列皆对应于特征根零. 记 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$, 则

$$\begin{aligned} Q'\Sigma Q &= \begin{pmatrix} Q'_1 \\ Q'_2 \end{pmatrix} \Sigma \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \\ &= \begin{pmatrix} Q'_1 \Sigma Q_1 & Q'_1 \Sigma Q_2 \\ Q'_2 \Sigma Q_1 & Q'_2 \Sigma Q_2 \end{pmatrix} = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

考虑线性变换

$$\begin{aligned} Y_{(1)} &= Q'_1 X \\ Y_{(2)} &= Q'_2 X \end{aligned}$$

依定理 16.0.54, 有

$$\begin{aligned} Y_{(1)} &= Q'_1 X \sim N_r(Q'_1 \mu, \Lambda) \\ Y_{(2)} &= Q'_2 X \sim N_{n-r}(Q'_2 \mu, 0) \end{aligned} \quad (16.34)$$

由 (16.34) 推得, $Q'_2 X = Q'_2 \mu$, 以概率为 1 成立. 这等价于 $Q'_2(X - \mu) = 0$, 以概率为 1 成立. 即

$$X - \mu \in \mathcal{M}(Q_1), \quad \text{以概率为 1 成立.} \quad (16.35)$$

因 $\Sigma = Q_1 \Lambda Q'_1$, 所以 $\mathcal{M}(\Sigma) = \mathcal{M}(Q_1)$. 我们推得 (16.34) 等价于

$$X - \mu \in \mathcal{M}(\Sigma), \quad \text{以概率为 1 成立} \quad (16.36)$$

. 另一方面, 从 (16.34) 得, $Y_{(1)}$ 的概率密度函数为

$$g(y_{(1)}) = (2\pi)^{-\frac{r}{2}} |\Lambda|^{-1/2} \exp \left\{ -\frac{1}{2} (y_{(1)} - Q'_1 \mu)' \Lambda^{-1} (y_{(1)} - Q'_1 \mu) \right\} \quad (16.37)$$

作变换 $x = Qy$. 由 Q 的正交性, 该变换的 Jacobi 行列式 $|Q| = \pm 1$. 又 $y_{(1)} = Q'_1 x$ 从 (16.37) 得到的密度函数

$$\begin{aligned} f(x) &= (2\pi)^{-\frac{r}{2}} \left(\prod_{i=1}^r \lambda_i \right)^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' Q_1 \Lambda^{-1} Q'_1 (x - \mu) \right\} \\ &= (2\pi)^{-\frac{r}{2}} \left(\prod_{i=1}^r \lambda_i \right)^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^+ (x - \mu) \right\} \end{aligned} \quad (16.38)$$

由 (16.36) 知, $(x - \mu)' \Sigma^- (x - \mu)$ 与广义逆 Σ^- 选择无关, 于是

$$(x - \mu)' \Sigma^+ (x - \mu) = (x - \mu)' \Sigma^- (x - \mu)$$

综合 (16.35) 和 (16.37), 我们得到如下结论: 若 $X \sim N_n(\mu, \Sigma)$, $\text{rk}(\Sigma) = r$, 则 $x - \mu$ 以概率为 1 落在子空间 $M(\Sigma)$ 内, 且在此子空间内有密度函数 (关于该子空间的 Lebesgue 测度)

$$(2\pi)^{-\frac{r}{2}} \left(\prod_{i=1}^r \lambda_1, \dots, \lambda_r \right)^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' \Sigma^- (x - \mu) \right\}$$

这个结果是由 Khatri(见文献 [71]) 得到的. 把上面的结果归纳起来, 即为

Theorem 16.0.58 设 $X \sim N_n(\mu, \Sigma)$, 则

1. 当 $\Sigma > 0$ 时, X 具有密度 (16.32)
2. 当 $\text{rk}(\Sigma) = r < n$ 时, $X - \mu$ 以概率为 1 落在子空间 $\mathcal{M}(\Sigma)$ 内, 且在此子空间内具有密度 (16.38).

应用定义二, 我们也很容易获得多元正态分布的特征函数. 我们知道 $N(0, 1)$ 的特征函

数为

$$\varphi(t) = e^{-\frac{t^2}{2}}$$

于是 $U \sim N_r(0, I_r)$ 的特征函数为

$$\varphi_u(t) = e^{-\frac{t^2}{2}}, \quad t \in R_r$$

记 $i = \sqrt{-1}$, 那么 $X = AU + \mu$ 的特征函数

$$\begin{aligned} \varphi_x(t) &= Ee^{it'X} = E\left(e^{it'(AU+\mu)}\right) \\ &= e^{it'\mu} E\left(e^{it'AU}\right) = e^{it'\mu} \varphi_u(A't) \\ &= e^{it'\mu} e^{-\frac{t'AA't}{2}} \\ &= e^{it'\mu - \frac{t'AA't}{2}}, \quad t \in R_n \end{aligned}$$

因为由概率论中的惟一性定理, 我们知道, 随机变量的分布是由它的特征函数惟一确定的, 于是我们证明了如下定理:

Theorem 16.0.59 $X \sim N_n(\mu, \Sigma)$ 当且仅当它的特征函数为

$$\varphi_x(t) = e^{it'\mu - \frac{(t'\Sigma)t}{2}}, \quad t \in R_n$$

Theorem 16.0.60 具有均值向量 μ , 协方差阵为 Σ 的随机向量 X 服从多元正态分布当且仅当对任意实向量 $c, c'X \sim N(c'\mu, c'\Sigma c)$, 这里 $\Sigma \geq 0$

Proof. 必要性由定义二直接推出, 也可以从定理 16.0.54 导出. 现在证明充分性. 若对任意 $c, c'X \sim N(c'\mu, c'\Sigma c)$, 则对一切 $t \in R$, 有

$$\varphi_{c'X}(t) = e^{itc'\mu - \frac{(c'\Sigma_c)t^2}{2}}$$

特别令 $t = 1$

$$\varphi_{c'X}(1) = e^{ic'\mu - \frac{(c'\Sigma_c)}{2}} = \varphi_x(c)$$

于是随机向量 X 的特征函数

$$\varphi_x(c) = e^{ic'\mu - \frac{(c'\Sigma_c)}{2}}$$

由定理 16.0.59, 这正是 $N(\mu, \Sigma)$ 的特征函数. 依惟一性定理, 知 $X \sim N(\mu, \Sigma)$. 证毕. ■



若 $X \sim N(\mu, \Sigma)$, 当 $\Sigma > 0$ 时, 对任意 $c \in R_n$, 若 $c \neq 0, c'\Sigma c > 0$, 则 $c'X$ 是非退化的一元正态变置. 若 $\Sigma \geq 0, rk(\Sigma) = r < p$, 即便 $c \neq 0$, 可能有 $c'\Sigma c = 0$. 这时 $P(c'X = c'\mu) = 1$, $c'X$ 是退化的一元正态随机变量. 事实上, 对任意 $c \in \mathcal{M}(\Sigma)^\perp$ 都有 $P(c'X = c'\mu) = 1$

■ **Example 16.19** 设 X_1, \dots, X_n 为从正态总体 $N(\mu, \sigma^2)$ 抽取的简单随机样本, 则样本均值 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. 事实上, 若记 $X = (X_1, \dots, X_n)', c = (\frac{1}{n}, \dots, \frac{1}{n})'$, 则 $\bar{X} = c'X$. 依 3.3.4 知 \bar{X} 服从正态分布. 其余结论的证明是容易的, 留给读者作练习. 在定理中取 $c' = (0, \dots, 0, 1, 0, \dots, 0)$, 则 $c'X = X_i, c'\mu = \mu_i, c'\Sigma c = \sigma_{ii}$.

于是我们有如下推论:

Corollary 16.0.61 设 $X \sim N_n(\mu, \Sigma)$, $\mu = (\mu_1, \dots, \mu_n)$, $\Sigma = (\sigma_{ij})$, 则 $X_i \sim N(\mu_i, \sigma_{ii})$ $i = 1, \dots, n$ 这个推论表明, 若 $X = (X_1, \dots, X_n)'$ 为 n 维正态向量, 则它的任一分量也是正态向量(包括退化情形)。但反过来的结论未必成立, 即 X_1, \dots, X_n 均为正态变量, $X = (X_1, \dots, X_n)'$ 未必为正态向量·我们可以举出很多这样的例子, 下面就是其中的一个。

■ **Example 16.20** 设 (X, Y) 的联合密度函数为

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \left[1 - \frac{xy}{(x^2+1)(y^2+1)} \right], \quad -\infty < x, y < +\infty$$

显然这不是二元正态分布的密度函数, 而 X 和 Y 的边缘分布为 $N(0, 1)$ 事实上

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} - \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{xy}{(x^2+1)(y^2+1)} e^{-\frac{1}{2}(x^2+y^2)} dy \end{aligned}$$

上式第二项被积函数对固定的 x 是 y 的奇函数, 因此第二项积分等于零. 于是

$$\begin{aligned} f_1(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} dy \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \end{aligned}$$

这里利用了 $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 1$ 这就证明了 $X \sim N(0, 1)$. 在 $f(x, y)$ 表达式中, x, y 的地位完全对称, 故 $Y \sim N(0, 1)$ 也成立.

这个例子容易推广到多元情形、设 X_1, \dots, X_n 的联合密度为

$$f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} \left[1 - \frac{\prod_{i=1}^n x_i}{\prod_{i=1}^n (x_i^2 + 1)} \right]$$

显然, X_1, \dots, X_n 联合分布不是 n 元正态, 但用前面同样的方法, 可以证明 $X_i \sim$

$$N(0, 1), \quad i = 1, \dots, n$$

现在我们来讨论多元正态的进一步性质, 先讨论边缘分布. 在以下讨论中, 无特殊声明, 总假设 $\Sigma \geq 0$, 即 Σ 不必是正定阵. 将 X, μ, Σ 做如下分块

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

这里 X_1, μ_1 皆为 $m \times 1$ 向量, Σ_{11} 为 $m \times m$ 矩阵.

Theorem 16.0.62 设 $X \sim N_n(\mu, \Sigma)$, 则 $X_1 \sim N_m(\mu_1, \Sigma_{11})$, $X_2 \sim N_{n-m}(\mu_2, \Sigma_{22})$

Proof. X_1 的特征函数为

$$\varphi_{x_1}(t) = \varphi_x(t_1, \dots, t_m, 0, \dots, 0) = e^{it'\mu_1 - \frac{t'\Sigma_1 t}{2}}$$

依定理 3.3.3 知 $X_1 \sim N_m(\mu_1, \Sigma_{11})$, 同理可证 $X_2 \sim N_{n-m}(\mu_2, \Sigma_{22})$, 定理证毕. 注意: 这个定理也可以用定理 3.3.1 来证明. ■

Theorem 16.0.63 设 $X \sim N_n(\mu, \Sigma)$, 则 X_1 和 X_2 独立当且仅当 $\Sigma_{12} = 0$

证明

Proof. 设 $t \in R_n$, $t = (t'_1, t'_2)', t_1 \in R_m, t_2 \in R_{n-m}$. $\varphi_x(t)$, $\varphi_{x_1}(t_1)$, $\varphi_{x_2}(t_2)$ 分别表示 X , X_1 和 X_2 的特征函数. 于是

$$\begin{aligned}\Sigma_{12} = 0 &\iff t' \Sigma t = t'_1 \Sigma_{11} t_1 + t'_2 \Sigma_{22} t_2 \\ &\iff \varphi_x(t) = \varphi_{x_1}(t_1) \varphi_{x_2}(t_2)\end{aligned}$$

利用如下事实: 随机向量独立当且仅当它们的联合特征函数等于它们的边缘特征函数的乘积. 这就证明了我们的结论, 定理证毕. 这个定理刻画了多元正态分布的一个重要性质, 相互独立与不相关是等价的.

如果限于非奇异正态分布, 当 $\Sigma_{12} = 0$ 时, 则 (16.32) 可分解为

$$f(x) = f_1(x_1) f_2(x_2)$$

其中

$$\begin{aligned}f_1(x_1) &= \frac{1}{(2\pi)^{m/2} |\Sigma_{11}|} e^{-\frac{1}{2}(x_1 - \mu_1)' \Sigma_{11}^{-1} (x_1 - \mu_1)} \\ f_2(x_2) &= \frac{1}{(2\pi)^{(n-m)/2} |\Sigma_{22}|} e^{-\frac{1}{2}(x_2 - \mu_2)' \Sigma_{22}^{-1} (x_2 - \mu_2)}\end{aligned}$$

这里 $f_1(x_1)$ 和 $f_2(x_2)$ 分别是 $X_1 \sim N_m(\mu_1, \Sigma_{11})$ 和 $X_2 \sim N_{n-m}(\mu_2, \Sigma_{22})$ 的密度函数. 因为从 $\Sigma > 0$ 可推出 $\Sigma_{ii} > 0$, 因此非奇异正态分布的边缘分布也是非奇异的. ■

例 3.3.3

■ **Example 16.21** 二元正态分布从初等概率统计教科书我们已经知道, 二元正态分布密度为

$$\begin{aligned}f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\cdot \exp\left\{\frac{1}{2(1-\rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \right\}\end{aligned}$$

若写成 (3.3.1) 的形式, 则其中的 μ 和 Σ 分别为

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

它们分别是二元正态向量的均值向量和协方差阵, ρ 表示相关系数. 因为 $|\Sigma| = (1 - \rho^2) \sigma_1^2 \sigma_2^2$, 所以为了保证 Σ 可逆, 我们要求 $|\rho| < 1$ 当 $\rho = 0$ 时, $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$, 依定理 3.3 .6 知, 此时 X_1 与 X_2 相互独立, 且的联合密度可分解为

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}}$$

可见 $X_i \sim N(\mu_i, \sigma_i^2)$, 且相互独立.

下面我们讨论多元正态的条件分布.

Theorem 16.0.64 设 $X \sim N_n(\mu, \Sigma)$, 对 X, μ, Σ 做如 (3.3.10) 的分块, 则给定 $X_1 = x_1$ 时, X_2 的条件分布为

$$X_2 | X_1 = x_1 \sim N_{n-m}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22.1})$$

这里 $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$

证明

Proof.

$$C = \begin{pmatrix} I_m & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix}$$

做变换 $Y = CX$, 则 $Y \sim N_n(C\mu, C\Sigma C')$. 利用 (2.2.11) 和 (2.2.12) 得

$$\Sigma_{21} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11} = 0, \quad \Sigma_{12} - \Sigma_{11}(\Sigma_{11}^{-1})'\Sigma_{12} = 0$$

于是

$$\begin{aligned} C\Sigma C' &= \begin{pmatrix} I_m & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{n-m} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_m & -(\Sigma_{11}^{-1})'\Sigma_{12} \\ 0 & I_{n-m} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22.1} \end{pmatrix} \end{aligned}$$

即

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1 \end{pmatrix} \sim N_n \left(\begin{pmatrix} \mu_1 \\ \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22.1} \end{pmatrix} \right)$$

于是

$$\begin{aligned} X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1 &\sim N_{n-m}(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22.1}) \\ X_1 &\sim N_m(\mu_1, \Sigma_{11}) \end{aligned}$$

且二者相互独立. 故给定 $X_1 = x_1$

$$X_2 \sim N_m(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22.1})$$

证毕。 ■

从这个定理我们可以获得如下重要事实:

$$E(X_2 | X_1 = x_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) = (\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1) + \Sigma_{21}\Sigma_{11}^{-1}x_1$$

即给定 $X_1 = x_1, X_2$ 的条件均值是 x_1 的线性函数. 由定理 3.3.2 的证明, 我们知道, $X_1 - \mu_1 \in \mathcal{M}(\Sigma_{11})$ (以概率为 1), 而 $\mathcal{M}(\Sigma_{21}) \subset \mathcal{M}(\Sigma_{11})$, 所以, $\Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ 与广义逆 Σ_{11}^{-1} 的选择无关. 从定理的证明, 我们还可以有如下推论:

Corollary 16.0.65 1. $X_1 - \Sigma_{12}\Sigma_{22}^-X_2 \sim N_m(\mu_1 - \Sigma_{12}\Sigma_{22}^-\mu_2, \Sigma_{11.2})$ 且与 $X_2 \sim N_{n-m}(\mu_2, \Sigma_{22})$ 相互独立, 其中 $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21}$
 2. $X_2 - \Sigma_{21}\Sigma_{11}^-X_1 \sim N_{n-m}(\mu_2 - \Sigma_{21}\Sigma_{11}^-\mu_1, \Sigma_{22.1})$ 且与 $X_1 \sim N_m(\mu_1, \Sigma_{11})$ 相互独立.

正态变量的二次型

设 $X \sim N_n(\mu, \Sigma), A_{n \times n}$ 为实对称阵. 本节的目的是研究 $X'AX$ 的性质, 特别是在什么条件下, 二次型 $X'AX$ 服从 χ^2 分布, 并讨论 χ^2 分布的一些重要性质. 以下我们总假设 $\Sigma > 0$.

Theorem 16.0.66 1. 设 $X \sim N_n(\mu, \Sigma), A_{n \times n}$ 对称, 则 $\text{Var}(X'AX) = 2\text{tr}(A\Sigma)^2 + 4\mu'A\Sigma A\mu$
 2. 设 $X \sim N_n(\mu, \sigma^2 I), A_{n \times n}$ 对称, 则 $\text{Var}(X'AX) = 2\sigma^4 \text{tr}(A^2) + 4\sigma^2 \mu'A^2\mu$

Proof. 1. 记 $Y = \Sigma^{-\frac{1}{2}}X$, 则 $Y \sim N_n(\Sigma^{-\frac{1}{2}}\mu, I)$, 所以 Y 的分量相互独立, $\text{Var}(X'AX) = \text{Var}(Y'\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}Y)$, 注意到对正态分布

$$\begin{aligned} m_3 &= E(Y_i - EY_i)^3 = 0 \\ m_4 &= E(Y_i - EY_i)^4 = 3 \end{aligned}$$

应用定理 16.0.53, 便得到第一条结论.

2. 这是(1)的特殊情况. 定理证毕. ■

定义 3.4.1 设 $X \sim N_n(\mu, I_n)$. 随机变母 $Y = X'X$ 的分布称为自由度为 n 非中心参数为 $\lambda = \mu'\mu$ 的 χ^2 分布, 记为 $Y \sim \chi^2_{n,\lambda}$. 当 $\lambda = 0$ 时, 称 Y 的分布为中心 χ^2 分布, 记为 $Y \sim \chi^2_n$

Theorem 16.0.67 χ^2 分布具有下述性质:

1. (可加性) 设 $Y_i \sim \chi^2_{n_i, \lambda_i}, i = 1, \dots, k$, 且相互独立, 则

$$Y_1 + \dots + Y_k \sim \chi^2_{n, \lambda}$$

这里 $n = \sum n_i, \lambda = \sum \lambda_i$

$$2. E(\chi^2_{n, \lambda}) = n + \lambda, \quad \text{Var}(\chi^2_{n, \lambda}) = 2n + 4\lambda$$

Proof. 1. 根据定义易得. 下证(2).

2. 设 $Y \sim \chi^2_{n, \lambda}$, 则依定义, Y 可表示为

$$Y = X_1^2 + \dots + X_{n-1}^2 + X_n^2$$

其中 $X_i \sim N(0, 1), i = 1, \dots, n-1$, $X_n \sim N(\sqrt{\lambda}, 1)$, 且相互独立, 于是

$$\begin{aligned} E(Y) &= \sum_{i=1}^n E(X_i^2) \\ \text{Var}(Y) &= \sum_{i=1}^n \text{Var}(X_i^2) \end{aligned} \tag{16.39}$$

因为

$$E(X_i^2) = \text{Var}(X_i) + E(X_i)^2 = \begin{cases} 1, & i = 1, \dots, n-1 \\ 1+\lambda, & i = n \end{cases}$$

代入 (16.39), 第一条结论得证. 直接计算可得

$$\begin{aligned} EX_i^4 &= 3, \quad i = 1, \dots, n-1 \\ EX_n^4 &= \lambda^2 + 6\lambda + 3 \end{aligned}$$

于是

$$\text{Var}(X_i^2) = EX_i^4 - (EX_i^2)^2 = 3 - 1 = 2, \quad i = 1, \dots, n-1$$

$$\text{Var}(X_n^2) = EX_n^4 - (EX_n^2)^2 = 2 + 4\lambda$$

代入 (16.39) 便证明了第二条结论. 设 $X \sim N_n(0, \Sigma), \Sigma > 0$, 依定义容易证明二次型 $X'\Sigma^{-1}X \sim \chi_n^2$. 事实上, 记 $Y = \Sigma^{-\frac{1}{2}}X$, 则 $Y^* \sim N_n(0, I)$. 于是

$$X'\Sigma^{-1}X = \left(\Sigma^{-\frac{1}{2}}X\right)' \left(\Sigma^{-\frac{1}{2}}X\right) = Y'Y \sim \chi_n^2$$

■

对于正态向量的一般二次型, 我们有下面的定理.

Theorem 16.0.68 设 $X \sim N_n(\mu, I_n)$, A 对称, 则 $X'AX \sim \chi_{r, \mu'A\mu}^2 \iff A$ 幂等,

$$\text{rk}(A) = r$$

Proof. 先证充分性. 设 A 幂等、对称, 且 $\text{rk}(A) = r$, 依定理 16.0.13, A 的特征根只能为 0 或 1, 于是存在正交方阵 Q , 使得

$$A = Q' \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q$$

令 $Y = QX$, 则 $Y \sim N_n(Q\mu, I_n)$, 对 Y 和 Q 做分块

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix}, Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$$

其中 $Y_{(1)} : r \times 1, Q_1 : r \times n$. 于是

$$X'AX = Y' \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Y = Y_{(1)}' Y_{(1)} \sim \chi_{r, \lambda}^2$$

其中 $\lambda = (Q_1\mu)' Q_1\mu = \mu' Q_1' Q_1\mu = \mu' A\mu$ 再证必要性. 设 $\text{rk}(A) = t$. 因 A 对称, 故存在正交方阵 Q , 使得

$$A = Q' \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} Q$$

其中 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_t)$. 我们只需证明 $\lambda_i = 1, i = 1, \dots, t, t = r$. 令 $Y = QX$, 则 $Y \sim N_n(Q\mu, I_n)$.

□

$$c = Q\mu = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

则

$$X'AX = Y' \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} Y = \sum_{j=1}^t \lambda_j Y_j^2$$

这里 $Y' = (Y_1 \cdots Y_n)$, $Y_j \sim N(c_j, 1)$ 且相互独立, $j = 1, \dots, t$. 依特征函数的定义, 不难算出 $\lambda_j Y_j^2$ 的特征函数为

$$g_j(z) = (1 - 2i\lambda_j z)^{-\frac{1}{2}} \exp \left\{ \frac{i\lambda_j z}{1 - 2i\lambda_j z} c_j^2 \right\}$$

利用独立随机变量之和的特征函数等于它们的特征函数之积, 由 (3.4.3) 得 $X'AX$ 的特征函数

$$\prod_{j=1}^t (1 - 2i\lambda_j z)^{-\frac{1}{2}} \exp \left\{ \frac{i\lambda_j z}{1 - 2i\lambda_j z} c_j^2 \right\}$$

我们再来计算 $\chi_{r,\lambda}^2$ 的特征函数. 设 $u \sim \chi_{r,\lambda}^2$, $\lambda = \mu' A \mu$, 记 $u = u_1^2 + \dots + u_r^2$, 其中 $u_1 \sim N(\lambda^{1/2}, 1)$, $u_j \sim N(0, 1)$, $j \geq 2$. 和刚才同样的道理, 得 u 的特征函数,

$$(1 - 2iz)^{-\frac{r}{2}} \exp \left\{ \frac{i\lambda z}{1 - 2iz} \right\}$$

依假设, $X'AX \sim \chi_{r,\lambda}^2$. 于是 (3.4.4) 和 (3.4.5) 应该相等. 比较两者的奇点及其个数知, $\lambda_j = 1$, $j = 1, \dots, t$, 且 $t = r$. 必要性得证. 定理证毕. ■

Corollary 16.0.69 设 $A_{n \times n}$ 对称, $X \sim N_n(\mu, I)$, 那么 $X'AX \sim \chi_k^2$, 即中心 χ^2 分布 $\iff A$ 幂等, $\text{rk}(A) = k$, $A\mu = 0$

Corollary 16.0.70 设 $A_{n \times n}$ 对称, $X \sim N_n(0, I)$, 那么 $X'AX \sim \chi_k^2 \iff A$ 幂等 $\text{rk}(A) = k$

Corollary 16.0.71 设 $A_{n \times n}$ 对称, $X \sim N_n(\mu, \Sigma)$, $\Sigma > 0$, 那么 $X'AX \sim \chi_{k;\lambda}^2$, $\lambda =$

$$\mu' A \mu \iff A \Sigma A = A$$

定理 3.4.3 及其推论把判定正态变量二次型服从 χ^2 分布的问题化为研究相应的二次型矩阵的问题, 而后者往往很容易处理. 因此, 这些结果是判定 χ^2 分布的很有效的工具.

■ **Example 16.22** 设 $X \sim N_n(C\beta, \sigma^2 I)$, $\text{rk}(C) = r$. 利用推论 16.0.69 容易证明, $X'[I - C(C'C)^- C']X / \sigma^2 \sim \chi_{n-r}^2$. 事实上, 该二次型的矩阵 $A = I - C(C'C)^- C'$ 是幂等阵, 依定理 16.0.15, 有

$$\text{rk}(A) = \text{tr}(A) = \text{tr}(I - C(C'C)^- C') = n - \text{tr}(C(C'C)^- C') = n - \text{rk}(C(C'C)^- C')$$

再利用推论 ??(3) 得 $\text{rk}(A) = n - \text{rk}(C'C) = n - \text{rk}(C) = n - r$. 又因 $AC = 0$, 根据推论 16.0.69,

$$X'[I - C(C'C)^- C']X / \sigma^2 \sim \chi_{n-r}^2$$

Theorem 16.0.72 设 $X \sim N_n(\mu, I)$, $X'AX = X'A_1X + X'A_2X \sim \chi^2_{r;\lambda}$, $X'A_1X \sim \chi^2_{s;\lambda_1}$, $A_2 \geq 0$, 其中 $\lambda = \mu'A\mu$, $\lambda_1 = \mu'A_1\mu$. 则

1. $X'A_2X \sim \chi^2_{r-s;\lambda_2}$, $\lambda_2 = \mu'A_2\mu$
2. $X'A_1X$ 和 $X'A_2X$ 相互独立,
3. $A_1A_2 = 0$

Proof. 因 $X'AX \sim \chi^2_{r;\lambda}$, 由定理 3.4 .3 知, A 幂等, $\text{rk}(A) = r$, 于是, 存在正交方阵 P , 使得

$$P'AP = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

因 $A \geq A_1, A \geq A_2$, 于是

$$\begin{aligned} P'A_1P &= \begin{pmatrix} B_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B_1 : r \times r \\ P'A_2P &= \begin{pmatrix} B_2 & 0 \\ 0 & 0 \end{pmatrix}, \quad B_2 : r \times r \end{aligned}$$

由假设 $X'A_1X \sim \chi^2_{s;\lambda_1}$, 推得 $A_1^2 = A_1$, 于是 $B_1^2 = B_1$. 故存在正交阵 $Q_{r \times r}$, 使

$$Q'B_1Q = \begin{pmatrix} I_s & 0 \\ 0 & 0 \end{pmatrix}$$

记

$$S' = \begin{pmatrix} Q' & 0 \\ 0 & I_{n-r} \end{pmatrix} P'$$

则 S 为正交阵, 且使

$$S'AS = S'A_1S + S'A_2S$$

形为

$$\begin{pmatrix} I_s & 0 & 0 \\ 0 & I_{r-s} & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} I_s & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_{r-s} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

做变换 $Y = SX$, 依定理 16.32, 有 $Y \sim N_n(S\mu, I)$. 于是

$$\begin{aligned} X'AX &= Y'S'ASY = \sum_{i=1}^r Y_i^2 \\ X'A_1X &= Y'S'A_1SY = \sum_{i=1}^s Y_i^2 \\ X'A_2X &= Y'S'A_2SY = \sum_{i=s+1}^r Y_i^2 \end{aligned}$$

因为 Y_1, \dots, Y_n 相互独立, 所以 $X'A_1X$ 与 $X'A_2X$ 相互独立. 再依定义, $X'A_2X \sim \chi^2_{r-s;\lambda_2}$, X

$$A_1A_2 = S \begin{pmatrix} I_s & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} S' S \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_{r-s} & 0 \\ 0 & 0 & 0 \end{pmatrix} S' = 0$$

(3) 得证. 定理证毕. ■

Corollary 16.0.73 设 $X \sim N_n(\mu, I)$, A_1, A_2 对称, $X'A_1X$ 和 $X'A_2X$ 都服从 χ^2 分布, 则它们相互独立 $\iff A_1A_2 = 0$

Proof. 充分性. 令 $A = A_1 + A_2$. 由 $A_1A_2 = 0$, 可推出 $A_2A_1 = (A_1A_2)' = 0$ 因此, 由 A_1, A_2 的蔡等性得

$$A^2 = (A_1 + A_2)^2 = A_1^2 + A_2^2 + A_1A_2 + A_2A_1 = A_1 + A_2 = A$$

必要性. 若 $X'A_1X$ 与 $X'A_2X$ 相互独立, 则 $X'AX$ 也服从 χ^2 分布, 再由定理 16.0.72(3), 结论得证. ■

上面两个定理很容易推广到 $\text{Cov}(X) = \Sigma > 0$ 的情形.

Corollary 16.0.74 设 $X \sim N_n(\mu, \Sigma)$, $\Sigma > 0$, $X'AX = X'A_1X + X'A_2X \sim \chi_{r, \lambda_1}^2$, $X'A_1X \sim \chi_{s, \lambda_2}^2$, $A_2 \geq 0$, 则

1. $X'A_2X \sim \chi_{r-s, \lambda_3}^2$
2. $X'A_1X$ 与 $X'A_2X$ 相互独立,
3. $A_1\Sigma A_2 = 0$

其中 $\lambda_i, i = 1, 2, 3$ 为非中心参数, 不再精确写出.

Corollary 16.0.75 设 $X \sim N_n(\mu, \Sigma)$, $\Sigma > 0$, A_1, A_2 对称, $X'A_1X$ 与 $X'A_2X$ 都服从 χ^2 分布. 则它们相互独立 $\iff A_1\Sigma A_2 = 0$. 在这个推论中, 我们要求 $X'A_1X$ 与 $X'A_2X$ 都服从 χ^2 分布. 事实上, 从下一节我们可以看出, 这个条件是可以放弃的.

正态变量的二次型与线性型的独立性

设 $X \sim N_n(\mu, \Sigma)$, A, B 皆为 n 阶对称阵, C 为 $m \times n$ 矩阵. 本节将建立二次型 $X'AX, X'BX$ 和线性型 CX 相互独立的条件. 这些结果在线性模型的参数估计和假设检验中将有重要应用.

Theorem 16.0.76 设 $X \sim N_n(\mu, I)$, A 为 $n \times n$ 对称阵, C 为 $m \times n$ 矩阵. 若 $CA = 0$ 则 CX 和 $X'AX$ 相互独立.

Proof. 由 A 的对称性, 知存在标准正交阵 P , 使得

$$P'AP = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}$$

这里 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$, $\lambda_i \neq 0$, $\text{rk}(A) = r$. 由 $CA = 0$ 可推得 $CPP'AP = 0$. 这等价于

$$CP \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} = 0 \tag{16.40}$$

若记

$$D = CP = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$$

由 (16.40) 推得 $D_{11} = 0, D_{21} = 0$. 于是 D 就变为

$$D = \begin{pmatrix} 0 & D_{12} \\ 0 & D_{22} \end{pmatrix} \triangleq (0 : D_1), \quad D_1 : m \times (n - r)$$

将 P 做对应分块: $P = (P_1 : P_2)$, P_1 为 $n \times r$. 那么

$$C = DP' = (0 : D_1) \begin{pmatrix} P'_1 \\ P'_2 \end{pmatrix} = D_1 P'_2 \quad (16.41)$$

$$A = P \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} P' = P_1 \Lambda P'_1 \quad (16.42)$$

记 $Y = P'X$, 依定理 16.0.54, 我们知道

$$Y = \begin{pmatrix} Y_{(1)} \\ Y_{(2)} \end{pmatrix} = \begin{pmatrix} P'_1 X \\ P'_2 X \end{pmatrix} \sim N_n(P\mu, I)$$

显然, $Y_{(1)}$ 和 $Y_{(2)}$ 相互独立. 但由 (16.41) 和 (16.42), 有

$$\begin{aligned} CX &= D_1 P'_2 X = D_1 Y_{(2)} \\ X'AX &= X' P_1 \Lambda P'_1 X = Y'_{(1)} \Lambda Y_{(1)} \end{aligned}$$

因 CX 只依赖于 $Y_{(2)}$, 而 $X'AX$ 只依赖于 $Y_{(1)}$, 所以 CX 与 $X'AX$ 独立, 定理得证. ■

■ **Example 16.23** 设 X_1, \dots, X_n 为取自 $N(0, \sigma^2)$ 的随机样本, 则样本均值 \bar{X} 与样本方差 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 相互独立. 事实上, 若记 $X = (X_1, \dots, X_n)$, $1 = (1, \dots, 1)'$, 即 1 为所有分量全为 1 的 $n \times 1$ 向量, $X \sim N_n(0, \sigma^2 I)$, 则

$$\bar{X} = \frac{1}{n} \mathbf{1}' X, \quad (n-1)S^2 = X' CX$$

这里

$$C = I_n - \frac{1}{n} \mathbf{1} \mathbf{1}'$$

容易验证 $\mathbf{1}' C = 0$, 由定理 16.0.76 知 \bar{X} 与 S^2 独立.

Corollary 16.0.77 设 $X \sim N_n(\mu, \Sigma)$, $\Sigma > 0$, $A_{n \times n}$ 为对称阵. 若 $C\Sigma A = 0$, 则 CX 与 $X'AX$ 相互独立.

Theorem 16.0.78 设 $X \sim N_n(\mu, I)$, A, B 皆 $n \times n$ 对称, 若 $AB = 0$, 则 $X'AX$ 与 $X'BX$ 相互独立.

Proof. 由 $AB = 0$ 及 A, B 的对称性, 立得 $BA = 0$, 于是 $AB = BA$, 故存在正交阵 P , 可使 A, B 同时对角化, 即

$$\begin{aligned} P'AP &= \Lambda_1 = \text{diag} \left(\lambda_1^{(1)}, \dots, \lambda_n^{(1)} \right) \\ P'BP &= \Lambda_2 = \text{diag} \left(\lambda_1^{(2)}, \dots, \lambda_n^{(2)} \right) \end{aligned}$$

由 $AB = 0 \Rightarrow \Lambda_1 \Lambda_2 = 0$, 即

$$\lambda_i^{(1)} \text{ 和 } \lambda_i^{(2)} \text{ 至少有一个为 } 0, \quad i = 1, \dots, n \quad (16.43)$$

令

$$Y = P'X = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$$

则 $Y \sim N_n(P'\mu, I)$, 于是 Y 的诸分量 Y_1, \dots, Y_p 相互独立, 但

$$\begin{aligned} X'AX &= X'P\Lambda_1P'X = Y'\Lambda_1Y \\ X'BX &= X'P\Lambda_2P'X = Y'\Lambda_2Y \end{aligned}$$

根据 (16.43), $X'AX$ 与 $X'BX$ 所依赖的 Y 分量不同, 故 $X'AX$ 与 $X'BX$ 相互独立. 定理得证. 这个定理的逆也是对的, 即设 $X \sim N_n(\mu, I)$, A, B 皆 $n \times n$ 对称, 若 $X'AX$ 与 $X'BX$ 相互独立, 则 $AB = 0$. 这个事实的证明此处就略去了, 详见文境 [42] 和 [88], 后者还把定理推广到奇异正态分布的情形。 ■

Corollary 16.0.79 设 $X \sim N_n(\mu, \Sigma)$, $\Sigma > 0$, A, B 皆 $n \times n$ 对称. 若 $A\Sigma B = 0$, 则 $X'AX$ 与 $X'BX$ 相互独立.

17. 线性模型参数估计

17.0.1 最小二乘估计

我们讨论线性模型

$$y = X\beta + e, \quad E(e) = 0, \quad \text{Cov}(e) = \sigma^2 I \quad (17.1)$$

的参数 β 和 σ^2 的估计问题, 这里 y 为 $n \times 1$ 观测向量, X 为 $n \times p$ 的设计矩阵. β 为 $p \times 1$ 未知参数向量, e 为随机误差, σ^2 为误差方差, $\sigma^2 > 0$. 如果 $\text{rk}(X) = r \leq p$ 称 (17.1) 为降秩线性模型, 否则, 称为满秩线性模型. 我们先讨论 β 的估计问题. 获得参数向量的估计的基本方法是最小二乘法, 其思想是, β 的真值应该使误差向量 $e = y - X\beta$ 达到最小, 也就是它的长度平方

$$Q(\beta) = \|e\|^2 = \|y - X\beta\|^2 = (y - X\beta)'(y - X\beta)$$

达到最小。因此, 我们应该通过求 $Q(\beta)$ 的最小值来求 β 的估计. 注意到

$$Q(\beta) = y'y - 2y'X\beta + \beta'X'X\beta$$

利用矩阵微商公式 (见第二章)

$$\frac{\partial y'X\beta}{\partial \beta} = X'y, \quad \frac{\partial \beta'X'X\beta}{\partial \beta} = 2X'X\beta$$

于是

$$\frac{\partial Q(\beta)}{\partial \beta} = -X'y + 2X'X\beta$$

令其等于 0, 得到

$$X'X\beta = X'y \quad (17.2)$$

称之为正则方程.

因为向量 $X'y \in \mathcal{M}(X') = \mathcal{M}(X'X)$, 于是正则方程 (17.2) 是相容的. 根据定理 16.0.6, 正则方程 (17.2) 的解为

$$\hat{\beta} = (X'X)^{-}X'y \quad (17.3)$$

这里 $(X'X)^{-}$ 是 $X'X$ 的任意一个广义逆. 根据函数极值理论, 我们知道 $\hat{\beta}$ 只是函数 $Q(\beta)$ 的驻点. 我们还需证明它确实使 $Q(\beta)$ 达到最小. 事实上, 对任意一个 β

$$\begin{aligned} Q(\beta) &= \|y - X\beta\|^2 = \|y - X\hat{\beta} + X(\hat{\beta} - \beta)\|^2 \\ &= \|y - X\hat{\beta}\|^2 + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) + 2(\hat{\beta} - \beta)'X'(y - X\hat{\beta}) \end{aligned}$$

因为 $\hat{\beta}$ 满足正则方程 (17.2), 于是上式第三项为 0, 而第二项总是非负的, 于是

$$Q(\beta) \geq \|y - X\hat{\beta}\|^2 = Q(\hat{\beta}) \quad (17.4)$$

此式表明, $\hat{\beta}$ 确使 $Q(\beta)$ 达到最小.

现在我们再进一步证明, 使 $Q(\beta)$ 达到最小的必是 $\hat{\beta}$. 事实上, (4.1.4) 等号成立, 当且仅当

$$(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) = 0$$

等价地

$$X(\hat{\beta} - \beta) = 0$$

不难证明, 上式又等价于

$$X'X\beta = X'X\hat{\beta} = X'y$$

这就证明了, 使 $Q(\beta)$ 达到最小值的点必为正则方程的解 $\hat{\beta} = (X'X)^{-}X'y$. 若 $\text{rk}(X) = p$, 则 $X'X$ 可逆, 这时, $\hat{\beta} = (X'X)^{-1}X'y$, 且有 $E(\hat{\beta}) = \beta$, 即 $\hat{\beta}$ 是 β 的无偏估计. 这时, 我们称 $\hat{\beta} = (X'X)^{-1}X'y$ 为 β 的最小二乘估计 (least squares estimate, 简记为 LS 估计).

若 $\text{rk}(X) < p$, 则 $E(\hat{\beta}) \neq \beta$, 即 $\hat{\beta}$ 不是 β 的无偏估计. 更进一步, 此时根本不存在 β 的线性无偏估计. 事实上, 若存在 $p \times n$ 矩阵 A , 使得 Ay 为 β 的线性无偏估计, 即要求 $E(Ay) = AX\beta = \beta$, 对一切 β 成立. 必存在 $AX = I_p$. 但因 $\text{rk}(AX) \leq \text{rk}(X) < p = \text{rk}(I_p)$, 这就与 $AX = I_p$ 相矛盾. 因此, 这样的矩阵 A 根本不存在. 这表明当 $\text{rk}(X) < p$ 时, β 没有线性无偏估计, 此时我们称 β 是不可估的. 但是, 退一步, 我们可以考虑 β 的线性组合 $c'\beta$, 这就导致了可估的定义.

Definition 17.0.1 若存在 $n \times 1$ 向量 a , 使得 $E(a'y) = c'\beta$ 对一切 β 成立, 则称 $c'\beta$ 是可估函数 (estimable function).

定理 $c'\beta$ 是可估函数 $\Leftrightarrow c \in \mathcal{M}(X')$ 证明 $c'\beta$ 是可估函数 \Leftrightarrow 存在 $a_{n \times 1}$, 使得 $E(a'y) = c'\beta$, 对一切 β 成立 $\Leftrightarrow a'X\beta = c'\beta$, 对一切 β 成立 $\Leftrightarrow c = X'a$. 证毕. 这个定理告诉我们, 使 $c'\beta$ 可估的全体 $p \times 1$ 向量 c 构成子空间 $\mathcal{M}(X')$. 于是, 若 c_1, c_2 为 $p \times 1$ 向量, 使 $c_1'\beta$ 和 $c_2'\beta$ 均可估, 那么, 对任意两个数 α_1, α_2 , 线性组合 $\alpha_1 \cdot c_1'\beta + \alpha_2 \cdot c_2'\beta$ 都是可估的. 若 c_1 和 c_2 为线性无关, 则称可估函数 $c_1'\beta$ 和 $c_2'\beta$ 是线性无关的. 显然, 对于一个线性模型, 线性无关的可估函数组最多含有 $\text{rk}(X) = r$ 个可估函数. 另外, 对于任一可估函数, $c'\hat{\beta}$ 与 $(X'X)^{-}$ 的选择无关, 是唯一的. 事实上, 由 $c'\beta$ 的可估性, 知存在向量 $a_{n \times 1}$, 使得 $c = X'a$, 于是

$$c'\hat{\beta} = c'(X'X)^{-}X'y = a'X(X'X)^{-}X'y = a'X(X'X)^{+}X'y$$

这里利用了 $X(X'X)^{-}X'$ 与广义逆 $(X'X)^{-}$ 选择无关, 故 $c'\hat{\beta}$ 也与 $(X'X)^{-}$ 的选择无关. 此时还有 $E(c'\hat{\beta}) = a'X(X'X)^{-}X'X\beta = a'X\beta = c'\beta$, 即 $c'\hat{\beta}$ 为 $c'\beta$ 的无偏估计. 于是, 我们给出如下定义. 定义 4.1.2 对可估函数 $c'\beta$, 称 $c'\hat{\beta}$ 为 $c'\beta$ 的 LS 估计. 对于线性模型 (??), 记 $X = (x_1, \dots, x_n)'$, 则这个模型的分量形式为

$$\begin{aligned} y_i &= x'_i\beta + e_i, \quad i = 1, \dots, n \\ E(e_i) &= 0, \quad \text{Cov}(e_i, e_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases} \end{aligned} \quad (17.5)$$

再记 $\mu_i = x'_i\beta$, $\mu = (\mu_1, \dots, \mu_n)' = X\beta = E(y)$, 即 μ 为观测向量 y 的均值向量. 它是 n 个可估函数, 但其中只有 $r = \text{rk}(X)$ 个是线性无关的. μ 的 LS 估计为

$$\hat{\mu} = X\hat{\beta} = X(X'X)^{-}X'y = P_Xy \quad (17.6)$$

这里 $P_X = X(X'X)^{-}X'$ 是向 $\mathcal{M}(X)$ 上的正交投影阵. 可见均值向量 μ 的 LS 估计就是 y 向 $\mathcal{M}(X')$ 上的正交投影.

对任一可估函数 $c'\beta$, 虽然它的 LS 估计 $c'\hat{\beta}$ 是唯一的, 但是它可能有很多个线性无偏估计. 事实上, 若记 $\mathcal{M}(X)^\perp$ 为 $\mathcal{M}(X)$ 的正交补空间. 设 $a'y$ 为 $c'\beta$ 的一个无偏估计, 那么对任意 $b \in \mathcal{M}(X)^\perp$, $(a+b)'y$ 也是 $c'\beta$ 的一个无偏估计. 此因 $E(a+b)'y = E(a'y) + E(b'y) = c'\beta + b'X'\beta = c'\beta$. 这样一来, 对任意线性函数 $c'\beta$ 它的线性无偏估计的个数有三种情况:

1. 一个也没有, 这时它是不可估的;
2. 只有一个, 这出现在 $\text{rk}(X) = n$ 的情形, 因为此时 $\mathcal{M}(X)^\perp = 0$;
3. 有无穷多个.

当 $c'\beta$ 可估时, 在其线性无偏估计当中, 方差最小者称为最佳线性无偏估计 (best linear unbiased estimate) 以下简记为 BLU 估计. 下面的定理表明, LS 估计就是 BLU 估计. 定理 4.1.2(Gauss-Markov 定理) 对任意的可估函数 $c'\beta$, LS 估计 $c'\hat{\beta}$ 为其唯一的 BLU 估计. 证明

Proof. 前面已证 $c'\hat{\beta}$ 为 $c'\beta$ 的无偏估计, 而线性性是显然的. 现证 $c'\hat{\beta}$ 的方差最小. 首先

$$\text{Var}(c'\hat{\beta}) = \text{Var}\left(c'(X'X)^{-}X'y\right) = \sigma^2 c'(X'X)^{-}X'X(X'X)^{-}c$$

由 $c'\beta$ 的可估性, 知存在向量 $\alpha_{n \times 1}$, 使得 $c = X'\alpha$, 于是, 利用 $X'X(X'X)^{-}X' = X'$ 得到

$$\text{Var}(c'\hat{\beta}) = \sigma^2 c'(X'X)^{-}X'X(X'X)^{-}X'\alpha = \sigma^2 c'(X'X)^{-}c$$

另一方面, 设 $a'y$ 为 $c'\beta$ 的任一无偏估计, 于是 a 满足: $X'a = c$. 这样

$$\begin{aligned} \text{Var}(a'y) - \text{Var}(c'\hat{\beta}) &= \sigma^2 [a'a - c'(X'X)^{-}c] \\ &= \sigma^2 (a' - c'(X'X)^{-}X') (a - X(X'X)^{-}c) \\ &= \sigma^2 \|a - X(X'X)^{-}c\|^2 \geq 0 \end{aligned}$$

并且等号成立 $\iff a' = c'(X'X)^{-}X' \iff a'y = c'\hat{\beta}$. 定理证毕. ■

这个重要的定理奠定了 LS 估计在线性模型参数估计理论中的地位. 由于它所刻画的 LS 估计在线性无偏估计类中的最优性, 使得人们长期以来把 LS 估计当作线性模型 (17.1) 的唯一最好的估计. 但是, 到了 20 世纪 60 年代, 许多研究表明, 在一些情况下 LS 估计的性质并不很好. 如果采用另外一个度量估计优劣的标准, LS 估计并不一定是最优的, 这些将留在第六章详细讨论.

Corollary 17.0.1 设 $\psi = c'_i \beta, i = 1, \dots, k$ 都是可估函数, $\alpha_i, i = 1, \dots, k$ 是实数, 则 $\psi = \sum_{i=1}^k \alpha_i \psi_i$ 也是可估的, 且 $\hat{\psi} = \sum_{i=1}^k \alpha_i \hat{\psi}_i = \sum_{i=1}^k \alpha_i c'_i \hat{\beta}$ 是 ψ 的 BLU 估计.

Corollary 17.0.2 设 $c' \beta$ 和 $d' \beta$ 是两个可估函数, 则

$$\begin{aligned}\text{Var}(c'_i \hat{\beta}) &= \sigma^2 c' (X' X)^{-} c \\ \text{Cov}(c'_i \hat{\beta}, d'_j \hat{\beta}) &= \sigma^2 c' (X' X)^{-} d\end{aligned}\tag{17.7}$$

并且上述两式与所含广义逆的选择无关.

这两个推论的证明也不困难, 留给读者完成. 现在我们讨论误差方差 σ^2 的估计. 记

$$\hat{e} = y - X \hat{\beta} = (I - P_X) y\tag{17.8}$$

称 \hat{e} 为残差向量. 它作为误差向量的一个“估计”, 对研究关于误差假设的合理性起着重要作用. 容易证明, 残差向量 \hat{e} 满足 $E(\hat{e}) = 0, \text{Cov}(\hat{e}) = \sigma^2 (I - P_X)$. 基于 \hat{e} , 我们可以构造 σ^2 的如下估计

$$\hat{\sigma}^2 = \frac{\hat{e}' \hat{e}}{n-r} = \frac{\|y - X \hat{\beta}\|^2}{n-r}\tag{17.9}$$

这里 $r = \text{rk}(X)$

Theorem 17.0.3 $\hat{\sigma}^2$ 是 σ^2 的无偏估计.

Proof. 因 $I - P_X$ 为幂等阵, 于是

$$\hat{e}' \hat{e} = y' (I - P_X) y$$

利用定理 3.2.1

$$E(\hat{e}' \hat{e}) = (X \beta)' (I - P_X) X \beta + \text{tr}(I - P_X) \text{Cov}(y) = \sigma^2 \text{tr}(I - P_X)$$

这里利用了 $(I - P_X) X = 0$. 利用迹和幂等阵的性质

$$E(\hat{e}' \hat{e}) = \sigma^2 [n - \text{tr}(P_X)] = \sigma^2 [n - \text{rk}(X)]$$

明所欲证. ■

为方便计, 通常也称 $\hat{\sigma}^2$ 为 σ^2 的 LS 估计.

对于线性模型 (17.1), 若我们进一步假设误差向量 e 服从多元正态分布, 则称相应的模型为正态线性模型, 记为

$$y = X \beta + e, \quad e \sim N(0, \sigma^2 I)\tag{17.10}$$

下面我们研究在这个模型下, LS 估计的性质.

Theorem 17.0.4 对正态线性模型 (17.10), 设 $c' \beta$ 为任一可估函数, 则

1. LS 估计 $c' \hat{\beta}$ 是 $c' \beta$ 的极大似然估计 (maximum likelihood estimate, 简记为 ML 估计), 且 $c' \hat{\beta} \sim N(c' \beta, \sigma^2 c' (X' X)^{-} c)$
2. $\frac{n-r}{n} \hat{\sigma}^2$ 为 σ^2 的 ML 估计, $A \frac{(n-r) \hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$
3. $c' \hat{\beta}$ 与 $\hat{\sigma}^2$ 相互独立, 这里 $\hat{\beta} = (X' X)^{-} X' y, r = \text{rk}(X)$

Proof. 记 $\mu = X\beta$, 考虑 μ 和 σ^2 的似然函数

$$L(\mu, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \|y - \mu\|^2}$$

取对数, 略去常数项, 得

$$\log L(\mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|y - \mu\|^2$$

对均值向量 μ 的 LS 估计 $\hat{\mu} = X\hat{\beta}$, 我们有

$$\|y - \hat{\mu}\|^2 = \|y - X\hat{\beta}\|^2 = \min_{\mu=X\beta} \|y - \mu\|^2$$

于是, 对每一个固定的 σ^2

$$\log L(\hat{\mu}, \sigma^2) \geq \log L(\mu, \sigma^2)$$

而

$$\log L(\hat{\mu}, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|y - \hat{\mu}\|^2$$

在 $\tilde{\sigma}^2 = \frac{1}{n} \|y - \hat{\mu}\|^2$ 达到最大. 于是 $\hat{\mu} = X\hat{\beta}$ 和 $\tilde{\sigma}^2$ 分别为 μ 和 σ^2 的 ML 估计. 对任一可估函数 $c'\beta$, 存在 $\alpha \in R^n$, 使得 $c = X'\alpha$. 于是, $c'\beta = \alpha'X\beta = \alpha'\mu$. 由 ML 估计的不变性, $c'\beta$ 的 ML 估计为 $\alpha'\hat{\mu}$, 注意到 $c'\hat{\beta} = \alpha'X\hat{\beta} = \alpha'\hat{\mu}$. 这就证明 LS 估计 $c'\hat{\beta}$ 为 ML 估计. 又因 $c'\hat{\beta} = c'(X'X)^{-}X'y$ 为 y 的线性函数, 而 $y \sim N_n(X\beta, \sigma^2 I)$, 依定理 16.0.60 知, $c'\hat{\beta} \sim N(c'(X'X)^{-}X'X\beta, \sigma^2 c'(X'X)^{-}c)$, 但由 $c'\beta$ 的可估性, 容易推出 $c'(X'X)^{-}X'X = c'$, 于是(1)得证.

(2) 的第一条结论已证. 因为 $P_X X = X$, 所以

$$\begin{aligned} \frac{(n-r)\hat{\sigma}^2}{\sigma^2} &= \frac{\hat{e}'\hat{e}}{\sigma^2} = \frac{y'(I-P_X)y}{\sigma^2} \\ &= \frac{e'(I-P_X)e}{\sigma^2} = z'(I-P_X)z \end{aligned}$$

其中 $z = e/\sigma \sim N_n(0, I)$. 由 $I - P_X$ 的斯等性及 $\text{rk}(I - P_X) = \text{tr}(I - P_X) = n - \text{tr}(P_X) = n - \text{rk}(X) = n - r$, 利用定理 16.0.68, 即得 $(n-r)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-r}^2$. 为证 $c'\hat{\beta}$ 与 $\hat{\sigma}^2$ 的独立性, 只要注意到 $c'\hat{\beta}$ 与 $\hat{\sigma}^2$ 分别为正态向量 y 的线性型和二次型, 根据定理 ?? 和 $c'(X'X)^{-}X'(I - P_X) = 0$, 结论可直接推得. 定理证毕. ■

从这个定理我们看出, 对于可估函数 $c'\beta$, 它的 LS 估计和 ML 估计是相同的. 但是, 对于误差方差 σ^2 , 两者就不同了. 它们只差一个因子, 很明显 ML 估计 $\tilde{\sigma}^2$ 是有偏的, $E(\tilde{\sigma}^2) = \frac{n-r}{n} \sigma^2 < \sigma^2$, 即在平均意义上讲, ML 估计 $\tilde{\sigma}^2$ 偏小. 在前面的 Guass-Markov 定理中, 我们证明了可估函数 $c'\beta$ 的 LS 估计 $c'\hat{\beta}$ 在线性无偏类中是方差最小的. 然而对于正态线性模型, 我们有下面更强的结果.

Theorem 17.0.5 对于正态线性模型 (17.10),

1. $T_1 = y'y$ 和 $T_2 = X'y$ 为完全充分统计量,
2. 对任一可估函数 $c'\beta, c'\hat{\beta}$ 为其惟一的最小方差无偏估计 (minimum variance unbiased estimate, 简记为 MVU 估计).

Proof. 观测向量 y 的概率密度函数为

$$\begin{aligned} f(y) &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta) \right\} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} y'y + \frac{1}{\sigma^2} y'X\beta - \frac{1}{2\sigma^2} \beta'X'X\beta \right\} \end{aligned}$$

记 $\theta_1 = -\frac{1}{2\sigma^2}$, $\theta_2 = \frac{\beta}{\sigma^2}$, 它们是所谓的自然参数, 则上式可改写为

$$f(y) = \frac{1}{(-\pi)^{\frac{n}{2}}} \theta_1^n e^{\frac{1}{4\theta_1} \theta_2' X' X \theta_2} \exp \{ \theta_1 T_1 + \theta_2 T_2 \}$$

这样, 我们把 $f(y)$ 表成了指数族的自然形式. 其参数空间

$$\Theta = \left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}; \quad \theta_1 < 0, \quad \theta_2 \in R^p \right\}$$

依文献 [25], p.59, 定理 ?? 知, $T_1 = y'y$ 和 $T_2 = X'y$ 为完全充分统计量.

对任一可估函数 $c'\beta$, 其 LS 估计 $c'\hat{\beta} = c'(X'X)^{-1}T_2$, 误差方差 σ^2 的 LS 估计 $\hat{\sigma}^2 = (T_1 - T_2(X'X)^{-1}T_2)/(n-r)$, 它们都是完全充分统计量的函数. 同时我们知道, 它们都是无偏估计, 依 Lehmann-Scheffe 定理 (参见文献 [25], p.58) 立即推出, $c'\hat{\beta}$ 和 $\hat{\sigma}^2$ 分别是 $c'\beta$ 和 σ^2 的惟一 MVU 估计. 定理证毕. ■

对任一可估函数 $c'\beta$, 这个定理和 Guass-Markov 定理都建立了它的 LS 估计 $c'\hat{\beta}$ 的方差最小性, 两者的区别在于, 本定理在误差服从正态分布的条件下, 证明了 LS 估计 $c'\hat{\beta}$ 在所有的 (线性的和非线性) 无偏估计类中方差最小. 而 Guass-Markov 定理只证明了 $c'\hat{\beta}$ 在线性无偏类中方差最小性.

例 4.1.1 设 μ 为一物体的重量, 现对该物体测量 n 次, 其测量值记为 y_1, \dots, y_n 通常我们用 $\bar{y} = \sum y_i/n$ 来估计 μ , 现在我们来研究估计 \bar{y} 的优良性. 如果测量过程没有系统误差, 则 y_i 可表示为

$$y_i = \mu + e_i, \quad i = 1, \dots, n$$

将其写成线性模型的矩阵形式

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

假设 $e = (e_1, \dots, e_n)'$ 满足 Guass-Markov 假设. 容易计算出 μ 的 LS 估计 $\hat{\mu} = (X'X)^{-1}X'y = \sum_{i=1}^n y_i/n = \bar{y}$, 即观测值的算数平均值为物体重量 μ 的 LS 估计. 且从 Guass-Markov 定理我们知道, 在 $y = (y_1, \dots, y_n)'$ 所有线性函数组成的无偏估计类中, 列具有最小方差. 如果我们进一步假设误差服从多元正态分布, 那么在所有无偏估计类中, \bar{y} 仍然具有最小方差. 这些结果充分显示了 \bar{y} 作为 μ 的估计的优良性质.

17.0.2 约束最小一乘估计

对线性模型 (17.1), 在上节, 我们导出了可估函数 $c'\beta$ 和 σ^2 的没有任何附带约束条件的最小二乘估计, 并讨论了它们的基本性质, 但是在检验问题的讨论中或其它一些场合, 我们需要求带一定约束条件的最小二乘估计. 假设

$$H\beta = d \tag{17.11}$$

是一个相容线性方程组，其中 H 为 $k \times p$ 的已知矩阵，且秩为 k , $\mathcal{M}(H') \subset \mathcal{M}(X')$, 于是 $H\beta$ 是 k 个线性无关的可估函数, d 为 $k \times 1$ 已知向量. 本节用 Lagrange 乘子法求模型 (17.1) 满足线性约束 (17.11) 的最小二乘估计. 记

$$H = \begin{pmatrix} h'_1 \\ \vdots \\ h'_k \end{pmatrix}, \quad d = \begin{pmatrix} d_1 \\ \vdots \\ d_k \end{pmatrix} \quad (17.12)$$

则线性约束 (17.11) 可以改写为

$$h'_i \beta = d_i, \quad i = 1, \dots, k \quad (17.13)$$

我们的问题是在 (17.13) 的 k 个条件下求 β 使 $Q(\beta) = \|y - X\beta\|^2$ 达到最小值. 为了应用 Lagrange 乘子法, 构造辅助函数

$$\begin{aligned} F(\beta, \lambda) &= \|y - X\beta\|^2 + 2 \sum_{i=1}^k \lambda_i (h'_i \beta - d_i) \\ &= \|y - X\beta\|^2 + 2\lambda'(H\beta - d) \\ &= (y - X\beta)'(y - X\beta) + 2\lambda'(H\beta - d) \end{aligned}$$

其中 $\lambda = (\lambda_1, \dots, \lambda_k)'$ 为 Lagrange 乘子, 对函数 $F(\beta, \lambda)$ 求对 β 的偏导数, 整理并令它们等于零, 得到

$$X'X\beta = X'y - H'\lambda \quad (17.14)$$

然后求解 (17.14) 和 (17.11) 组成的联立方程组, 记它们的解为 $\hat{\beta}_H$ 和 $\hat{\lambda}_H$. 因为 $\mathcal{M}(H') \subset \mathcal{M}(X')$, 所以 (17.14) 关于 β 是相容的, 其解

$$\hat{\beta}_H = (X'X)^{-1}X'y - (X'X)^{-1}H'\hat{\lambda}_H = \hat{\beta} - (X'X)^{-1}H'\hat{\lambda}_H \quad (17.15)$$

代入 (17.11) 得

$$d = H\hat{\beta}_H = H\hat{\beta} - H(X'X)^{-1}H'\hat{\lambda}_H$$

等价地

$$H(X'X)^{-1}H'\hat{\lambda}_H = (H\hat{\beta} - d) \quad (17.16)$$

这是一个关于 $\hat{\lambda}_H$ 的线性方程组. 因为 H 的秩为 k , 且 $\mathcal{M}(H') \subset \mathcal{M}(X')$, 于是 $H(X'X)^{-1}H'$ 跟所包含广义逆的选择无关. 故可知它是 $k \times k$ 的可逆矩阵, 因而 (17.16) 有唯一解

$$\hat{\lambda}_H = \left(H(X'X)^{-1}H' \right)^{-1} (H\hat{\beta} - d)$$

将 $\hat{\lambda}_H$ 代入 (17.15) 得到

$$\hat{\beta}_H = \hat{\beta} - (X'X)^{-1}H' \left(H(X'X)^{-1}H' \right)^{-1} (H\hat{\beta} - d) \quad (17.17)$$

现在我们证明 $\hat{\beta}_H$ 确实是线性约束 $H\beta = d$ 下 β 的最小二乘解. 为此我们只需证明如下两点:

1. $H\hat{\beta}_H = d$

2. 对一切满足 $H\beta_H = d$ 的 β , 都有

$$\|y - X\beta\|^2 \geq \|y - X\widehat{\beta}_H\|^2$$

根据 (17.17) 结论 (a) 是很容易验证的. 为了证明 (b), 我们将平方和 $\|y - X\beta\|^2$ 作分解

$$\begin{aligned} \|y - X\beta\|^2 &= \|y - X\widehat{\beta}\|^2 + (\widehat{\beta} - \beta)'X'X(\widehat{\beta} - \beta) \\ &= \|y - X\widehat{\beta}\|^2 + (\widehat{\beta} - \widehat{\beta}_H + \widehat{\beta}_H - \beta)'X'X(\widehat{\beta} - \widehat{\beta}_H + \widehat{\beta}_H - \beta) \\ &= \|y - X\widehat{\beta}\|^2 + (\widehat{\beta} - \widehat{\beta}_H)'X'X(\widehat{\beta} - \widehat{\beta}_H) + (\widehat{\beta}_H - \beta)'X'X(\widehat{\beta}_H - \beta) \\ &= \|y - X\widehat{\beta}\|^2 + \|X(\widehat{\beta} - \widehat{\beta}_H)\|^2 + \|X(\widehat{\beta}_H - \beta)\|^2 \end{aligned} \quad (17.18)$$

这里我们利用了 (17.15) 及 $\mathcal{M}(H') \subset \mathcal{M}(X')$ 导出的下述关系

$$(\widehat{\beta} - \widehat{\beta}_H)'X'X(\widehat{\beta}_H - \beta) = \widehat{\lambda}'H(\widehat{\beta}_H - \beta) = \widehat{\lambda}'(H\widehat{\beta}_H - H\beta) = \widehat{\lambda}'(d - d) = 0$$

这个等式对一切满足 $H\beta = d$ 的 β 都成立. (17.18) 式表明, 对一切满足 $H\beta = d$ 的 β , 总有

$$\|y - X\beta\|^2 \geq \|y - X\widehat{\beta}\|^2 + \|X(\widehat{\beta} - \widehat{\beta}_H)\|^2 \quad (17.19)$$

且等号成立当且仅当 (17.18) 式的第三项等于零, 也就是 $X\beta = X\widehat{\beta}_H$. 于是在 (17.19) 中用 $X\widehat{\beta}_H$ 替代 $X\beta$, 等式成立, 即

$$\|y - X\widehat{\beta}_H\|^2 = \|y - X\widehat{\beta}\|^2 + \|X(\widehat{\beta} - \widehat{\beta}_H)\|^2$$

综合 (17.19) 和 (??), 便证明了结论 (b).

Theorem 17.0.6 对于线性模型 (4.1.1), 设 H 为 $k \times p$ 矩阵, $\text{rk}(H) = k$, $\mathcal{M}(H') \subset \mathcal{M}(X')$, 且 $H\beta = d$ 相容, 则

1. $\widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta} - d)$ 为 β 在线性约束条件 $H\beta = d$ 下的约束 LS 解, $H\widehat{\beta}_H$ 为 $H\beta$ 的约束 LS 估计, 这里 $\widehat{\beta} = (X'X)^{-1}X'y$
2. 若 $\text{rk}(X) = p$, 则 $\widehat{\beta}_H = \widehat{\beta} - (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(H\widehat{\beta} - d)$ 为 β 的约束 LS 估计, 这里 $\widehat{\beta} = (X'X)^{-1}X'y$

看书上一个例子

Theorem 17.0.7 在定理 17.0.6 假设下, 在参数区域 $H\beta = d$ 上, $\widehat{\sigma}_H^2$ 是 σ^2 的无偏估计.

Proof. 由 (??), 得

$$E\|y - X\beta\|^2 = E\|y - X\widehat{\beta}\|^2 + E\|X(\widehat{\beta} - \widehat{\beta}_H)\|^2$$

由上节知 $E\|y - X\beta\|^2 = (n - r)\sigma^2$. 对上式第二项应用定理 3.2.1, 得

$$\begin{aligned}
 & E\left\|X\left(\hat{\beta} - \hat{\beta}_H\right)\right\|^2 \\
 &= E(H\hat{\beta} - d)' \left(H(X'X)^{-1}H'\right)^{-1} (H\hat{\beta} - d) \\
 &= (H\beta - d)' \left(H(X'X)^{-1}H'\right)^{-1} (H\beta - d) + \text{tr}\left[\left(H(X'X)^{-1}H'\right)^{-1} \text{Cov}(H\hat{\beta})\right] \\
 &= \delta + \text{tr}(\sigma^2 I_k) \\
 &= \delta + k\sigma^2
 \end{aligned} \tag{17.20}$$

这里 $\delta = (H\beta - d)' \left(H(X'X)^{-1}H'\right)^{-1} (H\beta - d)$. 于是我们证明了

$$E\left\|X\left(\hat{\beta} - \hat{\beta}_H\right)\right\|^2 = (n - r - k)\sigma^2 + \delta$$

显然, 在参数区域 $H\beta = d$ 上, $\delta = 0$. 定理证毕. ■

17.0.3 广义最小一乘估计

到目前为止, 我们的讨论都假定误差协方差阵为 $\sigma^2 I$ 的情形。但是, 哀观上存在着许多线性模型, 其误差协方差阵具有形式 $\sigma^2 \Sigma$, 并且 Σ 往往包含未知参数. 暂时我们先假设 Σ 是已知正定方阵, σ^2 为未知参数. 于是本节讨论线性模型:

$$y = X\beta + e, \quad E(e) = 0, \quad \text{Cov}(e) = \sigma^2 \Sigma \tag{17.21}$$

的参数 β, σ^2 的估计问题, 其中 $\Sigma > 0$. 因为假设了 $\Sigma > 0$, 故存在惟一的正定对称阵 $\Sigma^{\frac{1}{2}}$. 用 $\Sigma^{-\frac{1}{2}}$ 左乘 (4.3.1), 并记 $\tilde{y} = \Sigma^{-\frac{1}{2}}y, \tilde{X} = \Sigma^{-\frac{1}{2}}X, u = \Sigma^{-\frac{1}{2}}e$, 则得到

$$\tilde{y} = \tilde{X}\beta + u, \quad E(u) = 0, \quad \text{Cov}(u) = \sigma^2 I \tag{17.22}$$

这就化为以前讨论过的情形了。对模型 (17.22) 用最小二乘法求 β 的 LS 解, 即解 $Q(\beta) = \|\tilde{y} - \tilde{X}\beta\|^2$ 的最小值问题. 等价地, 解

$$\min Q(\beta) = \min(y - X\beta)' \Sigma^{-1} (y - X\beta)$$

正则方程组为

$$X'\Sigma^{-1}X\beta = X'\Sigma^{-1}y$$

于是, β 的 LS 解为

$$\beta^* = (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}y$$

称为广义最小二乘解。特别, 当 $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2), \sigma_i^2, i = 1, \dots, n$ 已知时, 称 β^* 为加权最小二乘解. 因为 (17.21) 和以前讨论的模型只是误差协方差阵不同, 而线性函数 $c'\beta$ 的可估性又与协方差阵无关, 于是, 对模型 (17.21), $c'\beta$ 可估的充要条件仍为 $c \in \mathcal{M}(X')$. 我们称 $c'\beta^*$ 为可估函数 $c'\beta$ 的广义最小二乘估计 (generalized least squares estimate), 简记为 GLS 估计. 对应地, 当 Σ 为对角阵时, 称 $c'\beta^*$ 为可估函数 $c'\beta$ 的加权最小二乘估计 (weighted least squares estimate), 简记为 WLS 估计. 当 $\text{rk}(X_{n \times p}) = p$ 时, β 可估, 称 β^* 为 β 的 GLS 估计. 因为导出(?)的方法是由 Aitken(1934) 首先提出的, 所以文献中也称 $c'\beta^*$ 和 β^* 为 Aitken 估计. 对应于 Gauss-Markov 定理, 我们有

Theorem 17.0.8 对任一可估函数 $c'\beta, c'\beta^*$ 为 $c'\beta$ 的 BLU 估计, 其方差为 $\sigma^2 c'(X'\Sigma^{-1}X)^{-1}c$

Proof. 因为 $c \in \mathcal{M}(X') = \mathcal{M}(X'\Sigma^{-1}X)$, 故存在向量 α 使得 $c = X'\Sigma^{-1}X\alpha$ 于是

$$\begin{aligned}\text{Var}(c'\beta^*) &= \sigma^2 c(X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}c \\ &= \sigma^2 c(X'\Sigma^{-1}X)^{-1}c\end{aligned}$$

设 $a'y$ 为 $c'\beta$ 的任一无偏估计, 则 $c = X'a$, 故

$$\begin{aligned}\text{Var}(a'y) - \text{Var}(c'\beta^*) &= \sigma^2 \left(a'\Sigma a - c(X'\Sigma^{-1}X)^{-1}c \right) \\ &= \sigma^2 \left(a'\Sigma a - a'X'(X'\Sigma^{-1}X)^{-1}X'a \right) \\ &= \sigma^2 \left(b'b - b'Q(Q'Q)^{-1}Q'b \right) \\ &= \sigma^2 b'(I - P_Q)b \geq 0\end{aligned}$$

其中 $b = \Sigma^{1/2}a, Q = \Sigma^{-1/2}X, P_Q = Q(Q'Q)^{-1}Q'$. 这就证明了 $c'\beta^*$ 的方差最小性上式等号成立 $\iff (I - P_Q)b = 0 \iff b = P_Qb \iff a = \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}c \iff a'y = c'\beta^*$. 惟一性得证 $c'\beta^*$ 的无偏性是显然的. 定理证毕. ■

根据 GLS 解 β^* , 我们可以给出 σ^2 的无偏估计, 记

$$e^* = y - X\beta^* = y - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y = \Sigma^{\frac{1}{2}}(I - P_{\Sigma^{-\frac{1}{2}}X})\Sigma^{-\frac{1}{2}}y$$

称为残差向量. 容易证明

$$\begin{aligned}E(e^*) &= 0 \\ \text{Cov}(e^*) &= \sigma^2 \Sigma^{\frac{1}{2}}(I - P_{\Sigma^{-\frac{1}{2}}X})\Sigma^{-\frac{1}{2}}\end{aligned}$$

记 $r = \text{rk}(X)$, 定义

$$\sigma^{2*} = (y - X\beta^*)' \Sigma^{-1} (y - X\beta^*) / (n - r) = \frac{e^{*'} \Sigma^{-1} e^*}{n - r}$$

类似于定理 4.1.3, 定理 4.1.4 和定理 4.1.5, 可以证明

Theorem 17.0.9 σ^{2*} 为 σ^2 的无偏估计.

Theorem 17.0.10 设 $e \sim N(0, \sigma^2 \Sigma), \Sigma(> 0)$ 已知, 则

1. 对任一可估函数 $c'\beta, c'\beta^*$ 为 $c'\beta$ 的 ML 估计, 且 $c'\beta^* \sim N(c'\beta, \sigma^2 c(X'\Sigma^{-1}X)^{-1}c)$
2. $\frac{n-r}{n}\sigma^{2*}$ 为 σ^2 的 ML 估计, $H(n-r)\sigma^{2*}/\sigma^2 \sim \chi^2_{n-r}$
3. $c'\beta^*$ 与 σ^{2*} 相互独立
4. 当 $\text{rk}(X_{n \times p}) = p$ 时, β^* 为 β 的 ML 估计, $\beta^* \sim N(\beta, \sigma^2 (X'\Sigma^{-1}X)^{-1})$, 且与 σ^{2*} 相互独立;
5. 若 $c'\beta$ 可估, 则 $c'\beta^*$ 为其惟一 MVU 估计
6. σ^{2*} 为 σ^2 的惟一 MVU 估计.

如果我们忽略 $\text{Cov}(e) = \sigma^2 \Sigma \neq \sigma^2 I$. 而按以前的 $\text{Cov}(e) = \sigma^2 I$ 情形来处理, 这就导致了 LS 解 $(X'X)^{-1}X'y$, 这样一来, 对任一可估函数 $c'\beta$, 我们就有了两个估计: LS 估计 $c'\hat{\beta}$ 和 GLS

估计 $c'\beta^*$, 两者都是无偏估计, 而后者是 BLU 估计. 一般来说, $c'\hat{\beta} \neq c'\beta^*$, 即 LS 估计和 BLU 估计不一定相等, 这是和 $\text{Cov}(e) = \sigma^2 I$ 情形所不同的. 特别, 当 $\text{rk}(X_{n \times p}) = p$ 时, β 的 LS 估计 $\hat{\beta} = (X'X)^{-1}X'y$, 而 GLS 估计 $\beta^* = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$, 它们都是 β 的无偏估计, 但协方差阵分别为

$$\begin{aligned}\text{Cov}(\beta^*) &= \sigma^2 (X'\Sigma^{-1}X)^{-1} \\ \text{Cov}(\hat{\beta}) &= \sigma^2 (X'X)^{-1} X' \Sigma X (X'X)^{-1}\end{aligned}$$

根据定理 17.0.8, 立即可推得 $\text{Cov}(\hat{\beta}) \geq \text{Cov}(\beta^*)$, 即

$$(X'X)^{-1} X' \Sigma X (X'X)^{-1} \geq (X'\Sigma^{-1}X)^{-1}$$

这里 $A \geq B$ 意为 $A - B \geq 0$, 此式表明 β^* 优于 $\hat{\beta}$.

书上有个例题 \$4.4\$ 最小二乘统一理论对于线性模型

$$y = X\beta + e, \quad E(e) = 0, \text{Cov}(e) = \sigma^2 \Sigma \quad (17.23)$$

如果 $|\Sigma| = 0$, 则称该模型为奇异线性模型, 对于这样的模型, 因为 Σ^{-1} 不存在, 所以我们不能通过最小化 (4.3.3) 所定义的 $Q(\beta)$ 来求得 β 的最小二乘估计. 20 世纪 60 年代以来, 许多统计学家研究了这种模型的参数估计, 提出了几种估计方法. 在这些估计方法中, 著名统计学家 Rao 应用推广的最小二乘法所导出的估计以其形式简单便于理论研究而得到普遍采用. 本节的目的是讨论这个方法. 对于奇异线性模型, 因为 Σ^{-1} 不存在, 于是 (4.3.3) 的 $Q(\beta)$ 无定义. 如果用任一广义逆 Σ^- 代替 Σ^{-1} , 把 $Q(\beta)$ 定义为 $Q(\beta) = (y - X\beta)' \Sigma^- (y - X\beta)$, 因为这样的 $Q(\beta)$ 与所含的广义逆 Σ^- 有关, 取不同的广义逆得到不同的 $Q(\beta)$, 因而 (4.3.3) 失去意义, 于是对于奇异线性模型, 一个核心的问题是寻找一个新矩阵 T , 它能够充当 (4.3.3) 中 Σ^{-1} 所担负的作用. Rao^[86] 成功地解决了这个问题. 他定义

$$T = \Sigma + XUX', \quad \text{其中 } U \geq 0, \quad \text{rk}(T) = \text{rk}(\Sigma : X) \quad (17.24)$$

然后定义

$$Q(\beta) = (y - X\beta)' T^- (y - X\beta) \quad (17.25)$$

用最小化 $Q(\beta)$ 求出最小值点

$$\beta^* = (X'T^-X)^{-1}X'T^-y \quad (17.26)$$

后面我们将证明, 对任一可估函数 $c'\beta, c'\beta^*$ 为其 BLU 估计. 这个结论既适用于设计阵 X 列满秩或列降秩的情形, 又适用于 Σ 奇异或非奇异的情形. 正是由于这个原因, 通常把这个结果称为最小二乘统一理论, 参见文献 [86]. 在 T 的定义中, 包含一个可以选择的半正定阵 U . 事实上满足条件的方阵 U 是很多的. 例如, 一个简单的选择是 $U = I_p$, 这是因为等式

$$\text{rk}(\Sigma + XX') = \text{rk}(\Sigma : X)$$

对一切 Σ 和 X 都成立. 另外, 当 $\Sigma > 0$ 时, 可取 $U = 0$, 此时 $T = \Sigma$, (4.4.4) 就变成了 (4.3.5). 为了证明 $c'\beta^*$ 为 $c'\beta$ 的 BLU 估计, 先证明几个预备事实. 引理 4.4.1 对于线性模型 (4.4.1), 不管 $\Sigma > 0$ 或 $\Sigma \geq 0, y \in \mathcal{M}(\Sigma : X)$ 总是成立. 证明 因为 $\Sigma \geq 0$, 将 Σ 分解为 $\Sigma = LL'$, 这里 L 为 $n \times t$ 矩阵, $t = \text{rk}(\Sigma) = \text{rk}(L)$. 记 $e = L\varepsilon, E(\varepsilon) = 0, \text{Cov}(\varepsilon) = \sigma^2 I_n$, 则 y 可表为如下新线性模型的形式

$$y = X\beta + L\varepsilon, \quad E(\varepsilon) = 0, \quad \text{Cov}(\varepsilon) = \sigma^2 I_n$$

于是 $y \in \mathcal{M}(X : L)$. 再利用 $\mathcal{M}(L) = \mathcal{M}(LL') = \mathcal{M}(\Sigma)$, 结论得证.

Lemma 17.1 对 (17.24) 所定义的 T , 总有

1. $\mathcal{M}(T) = \mathcal{M}(\Sigma : X)$
2. $X'T^{-}X, X'T^{-}y$ 和 $(y - X\beta)T^{-}(y - X\beta)$ 都与广义逆 T^{-} 的选择无关.

■

Proof. 证明 (1) 是 (17.24) 的直接推论. 因为 $y \in \mathcal{M}(T), \mathcal{M}(X) \subset \mathcal{M}(T), y - X\beta \in (2)$, 引理证毕. 这个推论表明, (17.25) 所定义的 $Q(\beta)$ 与所含的广义逆 T^{-} 的选择无关, 同时也可以证明, 对任一可估函数 $c'\beta, c'\beta^* = c'(X'T^{-}X)^{-}X'T^{-}y$ 也与所含的广义逆的选择无关. ■

引理 4.4.3 对于线性模型 (4.4.1), 可估函数 $c'\beta$ 的一个无偏估计 $a'y$ 为 BLU 估计, 当且仅当它满足

$$\text{Cov}(a'y, b'y) = 0$$

这里 $b'y$ 为零的任一无偏估计, 即 $E(b'y) = 0$ 证明 设 $l'y$ 为 $c'\beta$ 的任一无偏估计, 则 $l \rightarrow$ 定可表示为 $l = a + b$, 对某个满足 $X'b = 0$ 的 b . 于是

$$\text{Var}(l'y) = \text{Var}(a'y) + \text{Var}(b'y) + \text{Cov}(a'y, b'y) \quad (17.27)$$

由 (4.4.5), 充分性部分得证. 下面用反证法来证明必要性. 设 $a'y$ 为 $c'\beta$ 的 BLU 估计. 若存在一个 b_0 , 满足 $X'b_0 = 0$, 但有 $\text{Cov}(a'y, b'_0y) = d \neq 0$, 不妨设 $d < 0$. 若不然, 只需取 $-b_0$ 代替 b_0 , 就可化为 $d < 0$ 的情形. 用 $b = \alpha b_0$ 代替 (4.4.5) 中的 b , 则 (4.4.5) 为 α 的二次三项式, 且一次项为负数, 故必存在 α_0 使此二次三项式的后面两项之和取负值. 取 $l_0 = a + \alpha_0 b_0$, 必有

$$\text{Var}(l'_0y) < \text{Var}(a'y),$$

这与 $a'y$ 为 BLU 估计相矛盾. 引理得证. 现在证明如下重要定理. 定理 4.4.1 对于线性模型 (4.4.1) 和任一可估函数 $c'\beta$ 有 (1) $c'\beta^* = c'(X'T^{-}X)^{-}X'T^{-}y$ 为 $c'\beta$ 的 BLU 估计 (2) $\text{Var}(c'\beta^*) = \sigma^2 c' [(X'T^{-}X)^{-} - U] c$ 证明 (1) 由 $c'\beta$ 的可估性, 知存在 $n \times 1$ 的向量 t , 使得 $c' = t'X$. 利用 $X(X'T^{-}X)^{-}X'T^{-}X = X$, 于是

$$E(c'\beta^*) = t'X(X'T^{-}X)^{-}X'T^{-}X\beta = t'X\beta = c'\beta$$

无偏性得证. 以下我们应用引理 4.4.3 来证明 $c'\beta^*$ 在线性无偏估计类中是方差最小的. 对任一满足 $X'b = 0$ 的向量 b , 总有

$$\begin{aligned} \text{Cov}(a'y, b'y) &= \sigma^2 c' (X'T^{-}X)^{-} X'T^{-}\Sigma b \\ &= \sigma^2 c' (X'T^{-}X)^{-} X'T^{-}Tb \\ &= \sigma^2 c' (X'T^{-}X)^{-} X'b = 0 \end{aligned}$$

这里我们利用了 $X'T^{-}T = X'$ 和 $X'b = 0$. 根据引理 4.4.3, $c'\beta^*$ 为 $c'\beta$ 的 BLU 估计. (2) 首先注意到, 在表达式

$$\text{Var}(c'\beta^*) = \sigma^2 c' (X'T^{-}X)^{-} X'T^{-}\Sigma T^{-'}X \left((X'T^{-}X)^{-} \right)' c$$

中, $\left((X'T^{-}X)^{-} \right)'$ 和 $T^{-'}$ 可分别用 $(X'T^{-}X)^{-}$ 和 T^{-} 所替代, 于是

$$\text{Var}(c'\beta^*) = \sigma^2 c' (X'T^{-}X)^{-} X'T^{-}\Sigma T^{-}X (X'T^{-}X)^{-} c$$

再用 $T - XUX'$ 代替其中的 Σ , 得到

$$\begin{aligned}\text{Var}(c'\beta^*) &= \sigma^2 \left[c' (X'T^-X)^- X'T^-TT^-X (X'T^-X)^- c \right. \\ &\quad \left. - \sigma^2 c' (X'T^-X)^- X'T^-XUX'T^-X (X'T^-X)^- c \right]\end{aligned}$$

再利用 $c' = t'X$ 和 $X'T^-T = X'$, 上式右端第一项变为

$$\begin{aligned}c' (X'T^-X)^- X'T^-X (X'T^-X)^- c \\ &= t'X (X'T^-X)^- X'T^-X (X'T^-X)^- X't \\ &= t'X (X'T^-X)^+ X'T^-X (X'T^-X)^+ X't \\ &= t'X (X'T^-X)^+ X't \\ &= c' (X'T^-X)^+ c\end{aligned}$$

而对右端第二项, 利用 $c' = t'X$ 和 $X(X'T^-X)^+ X'T^-X = X'$, 得

$$\begin{aligned}c' (X'T^-X)^- X'T^-XUX'T^-X (X'T^-X)^- c \\ &= t'X (X'T^-X)^- X'T^-XUX'T^-X (X'T^-X)^- X't \\ &= t'XUX't = c'Uc\end{aligned}$$

定理得证. 若 $\text{rk}(X) = p$, 则 β 的 BLU 估计为 $\beta^* = (X'T^-X)^{-1}X'T^-y$. 若 X 为列降秩, 这时需要全体可估函数的估计. 在这种情况下, 可改为讨论均值向量 $\mu = X\beta$, 这是因为任一可估函数都可表为 μ 的线性组合, 容易证明它的 BLU 估计为

$$\mu^* = X\beta^* = X(X'T^-X)^- X'T^-y$$

且

$$\text{Cov}(\mu^*) = \sigma^2 X \left[(X'T^-X)^- - U \right] X'$$

下面的推论是一个具有广泛应用的重要特殊情形。推论 4.4.1

Corollary 17.0.11 对于线性模型 (4.4.1), 若 $\mathcal{M}(X) \subset \mathcal{M}(\Sigma)$, 则对任一可估函数 $c'\beta$, 它的 BLU 估计为

$$\begin{aligned}c'\beta^* &= c'(X'\Sigma^-X)^- X'\Sigma^-y \\ \text{Var}(c'\beta^*) &= \sigma^2 c'(X'\Sigma^-X)^- c,\end{aligned}\tag{17.28}$$

并且所有表达式与所包含的广义逆选择无关, 特别, 当 $\text{rk}(X) = p$ 时, $\beta^* = (X'\Sigma^-X)^{-1}X'\Sigma^-y$ 为 β 的 BLU 估计, 它的协方差阵为 $\text{Cov}(\beta^*) = \sigma^2 (X'\Sigma^-X)^{-1}$

证明 因为在条件 $\mathcal{M}(X) \subset \mathcal{M}(\Sigma)$ 下, 在 (4.4.2) 是中的 U 可取为零矩阵, 这时 $T = \Sigma$. 证毕. 我们知道, 当 $\Sigma > 0$ 时, 对任一可估函数 $c'\beta$, 它的 BLU 估计为

$$\begin{aligned}c'\beta^* &= c'(X'\Sigma^{-1}X)^- X'\Sigma^{-1}y \\ \text{Var}(c'\beta^*) &= \sigma^2 c'(X'\Sigma^{-1}X)^- c\end{aligned}$$

与 (4.4.6) 相比较, 我们发现, 当 $|\Sigma| = 0$ 时, 只要 $\mathcal{M}(X) \subset \mathcal{M}(\Sigma)$, Σ^- 就能够担负起 $\Sigma > 0$ 时 Σ^{-1} 所起的作用. 注 1 条件 $\mathcal{M}(X) \subset \mathcal{M}(\Sigma)$ 是任一可估函数 $c'\beta$ 的 BLU 估计为 (4.4.6) 的

充分条件, 但它并不必要. 例如, 在线性模型(4.4.1)中, 若 $X = \begin{pmatrix} \mathbf{1}_n \\ X_1 \end{pmatrix}$, 这里 $\mathbf{1}_n = (1, \dots, 1)'$, 即 n 个元素皆为 1 的 n 维向量, X_1 为任意的 $n \times (p-1)$ 矩阵, $\Sigma = I_n - \mathbf{1}_n \mathbf{1}_n' / n$, 即 Σ 为中心化矩阵, 这是一个零等阵, 单位阵 I_n 和 Σ 本身都是 Σ 的广义逆. 由定理 4.5.1 可以证明, 在这个模型里, 任一可估函数的 LS 估计都是它的 BLU 估计, 这相当于在(4.4.6)中取 Σ^- 为 I_n , 但是条件 $\mathcal{M}(X) \subset \mathcal{M}(\Sigma)$ 并不成立. 定理 4.4.2

$$\sigma^{2*} = (y - X\beta^*)' T^- (y - X\beta^*) / q$$

为 σ^2 的无偏估计, 其中 $q = \text{rk}(T) - \text{rk}(X)$ 证明因为

$$\begin{aligned} & E(y - X\beta^*)' T^- (y - X\beta^*) \\ &= \text{tr}[T^- E(y - X\beta^*)(y - X\beta^*)'] \end{aligned}$$

直接计算 $E(y - X\beta^*)(y - X\beta^*)'$ 并将所得表达式中的 Σ 用 $T - XUX'$ 代替, 再利用关系式

$$X(X'T^-X)^{-1}X'T^-X' = X, \quad X'T^-T = X'$$

得到

$$\begin{aligned} & E(y - X\beta^*)' T^- (y - X\beta^*) \\ &= \sigma^2 \text{tr}[T^- T - T^- X(X'T^-X)^{-1}X'] \end{aligned}$$

注意到 T^-T 和 $(X'T^-X)^{-1}X'T^-X$ 都是幂等阵, 利用幂等阵的性质: 若 A 为幂等阵, 则 $\text{rk}(A) = \text{tr}(A)$, 以及对任意矩阵 B , 有 $\text{rk}(B^-B) = \text{rk}(B)$, 于是有

$$\begin{aligned} & E(y - X\beta^*)' T^- (y - X\beta^*) \\ &= \sigma^2 [\text{rk}(T^-T) - \text{rk}((X'T^-X)^{-1}X'T^-X)] \\ &= \sigma^2 [\text{rk}(T) - \text{rk}(X')] \\ &= \sigma^2 [\text{rk}(T) - \text{rk}(X)] \\ &= \sigma^2 q \end{aligned}$$

定理证毕. 注 2 对任一可估函数 $c'\beta$, 它的 BLU 估计 $c'\beta^*$ 及其方差以及估计 σ^{2*} 都与所含的广义逆无关, 因此都可以用对应的 Moore-Penrose 广义逆代替, 即

$$\begin{aligned} c'\beta^* &= c'(X'T^+X)^+ X'T^+y \\ \text{Var}(c'\beta^*) &= T^+ c' [(X'T^+X)^+ - U] c \\ \sigma^{2*} &= (y - X\beta^*)' T^+ (y - X\beta^*) / q \end{aligned}$$

另外, 这些表达式还都与 T 的选择无关, 只要它满足(4.4.2). 为简单计常取 $U = I$, 这时 $T = \Sigma + XX'$. 特别当 $\mathcal{M}(X) \subset \mathcal{M}(\Sigma)$ 时, 取 $U = 0$, 即 $T = \Sigma$, 于是

$$\begin{aligned} c'\beta^* &= c'(X'\Sigma^+X)^+ X'\Sigma^+y \\ \text{Var}(c'\beta^*) &= \sigma^2 c'(X'\Sigma^+X)^+ c \end{aligned}$$

下面我们讨论 β 的几种估计, 以后我们总假定 $\text{rk}(X) = k$ 引理 4.4.4 (1) P, Q, P_1 和 J_{NT} 都是对称零等阵, 其秩分别为 $N, N(T-1), N-1$ 和 1 (2) P_1, Q 和 J_{NT} 两两正交, 即 $P_1Q = 0, P_1J_{NT} = 0, QJ_{NT} = 0$

(3) $PQ = 0, PJ_{NT} = J_{NT}, PP_1 = P_1P = P_1$ 这些事实的证明并不困难, 但它们对后面结论的证明是很关键的. 假定 σ_μ^2 和 σ_e^2 已知, 则 β 的 BLU 估计可表示为

$$\beta^*(\sigma^2) = \left(\frac{X'P_1X}{\sigma_\mu^2} + \frac{X'QX}{\sigma_e^2} \right)^{-1} \left(\frac{X'P_1y}{\sigma_\mu^2} + \frac{X'Qy}{\sigma_e^2} \right) \quad (17.29)$$

$\sigma^2 = (\sigma_1^2, \sigma_e^2)'$, 它的协方差阵为

$$\text{Cov}(\beta^*(\sigma^2)) = \left(\frac{X'P_1X}{\sigma_1^2} + \frac{X'QX}{\sigma_e^2} \right)^{-1} \quad (17.30)$$

但是, 在实际应用中, 因为 σ_μ^2 和 σ_e^2 都是未知的, 因此 $\beta^*(\sigma^2)$ 并不能付诸应用. 这时我们有两种处理方法: 一种是先设法获得 σ_μ^2 和 σ_e^2 的某种估计, 然后代入 (4.4.9). 通常把所得的估计称为两步估计。关于这种估计, 我们将在后面讨论. 另一种方法是寻求不包含 σ_μ^2 和 σ_e^2 的估计, 例如, LS 估计

$$\hat{\beta} = (X'P_1X + X'QX)^{-1}(X'P_1y + X'Qy) \quad (17.31)$$

和 Within 估计

$$\hat{\beta}_W = (X'QX)^{-1}X'Qy \quad (17.32)$$

以及 Between 估计

$$\hat{\beta}_B = (X'P_1X)^{-1}X'P_1y \quad (17.33)$$

比较 (4.4.11) 和 (4.4.9) 知, LS 估计可以看做是在 (4.4.9) 中令 $\sigma_1^2 = \sigma_e^2$, 即 $\sigma_\mu^2 = 0$ 时产生的. 而 Within 估计和 Between 估计的获得稍微复杂一点, 需要对两个变换模型应用最小二乘统一理论才能获得。

对模型 (4.4.8) 分别左乘 P_1 和 Q , 得到

$$P_1y = P_1X\beta + u_1 \quad (17.34)$$

$$Qy = QX\beta + u_2 \quad (17.35)$$

这里 $u_1 = P_1u, u_2 = Qu$. u_1 和 u_2 的均值皆为零, 它们的协方差阵分别为

$$V_1 = \text{Cov}(u_1) = \sigma_1^2 P_1 \quad (17.36)$$

$$V_2 = \text{Cov}(u_2) = \sigma_e^2 Q \quad (17.37)$$

因为 P_1 和 Q 都是幂等阵, 所以这两个模型都是奇异线性模型. 因为 $\mathcal{M}(P_1X) \subset \mathcal{M}(P_1), \mathcal{M}(QX) \subset \mathcal{M}(Q)$, 故由推论 4.4.1 容易证明 $\hat{\beta}_W$ 和 $\hat{\beta}_B$ 分别是从模型 (4.4.14) 和 (4.4.15) 求到的 β 的 BLU 估计. 这里我们总是假定 $(X'P_1X)^{-1}$ 和 $(X'QX)^{-1}$ 是存在的, 这在经济数据分析中总是成立的. 容易验证, $\hat{\beta}_W$ 和 $\hat{\beta}_B$ 的协方差阵分别为

$$\text{Cov}(\hat{\beta}_W) = \sigma_e^2 (X'QX)^{-1}$$

和

$$\text{Cov}(\hat{\beta}_B) = \sigma_1^2 (X'P_1X)^{-1}$$

84.5 LS 估计的稳健性

虽然稳健性 (robustness) 这种统计思想在统计文献中由来已久, 并且从 20 世纪 20 年代就开始受到统计学家的重视, 但“稳健性”一词只是到了 1953 年才由 G.E.P.Box 第一次明确提出来. 直观地讲, 稳健性是指统计推断关于统计模型即假设条件具有相对稳定性. 这就是说, 当模型假设发生某种微小变化时, 相应地统计推断与只有微小改变. 这时, 我们就说

统计推断关于这种微小变化具有稳健性。例如，本章开头几节的讨论中，关于线性模型有一个重要的假设是 $\text{Cov}(e) = \sigma^2 I$ 。在此条件下，证明了可估函数 $c'\beta$ 的 LS 估计 $c'\hat{\beta}$ 是 BLU 估计。但是在应用上我们不可能要求一个实际问题完全满足这一假设。事实上，我们也根本无法知道，它确实满足这条假设。只能通过分析或检验，判断假设 $\text{Cov}(e) = \sigma^2 I$ 是否大致上可以接受。因此，我们总是希望当实际的 $\text{Cov}(e)$ 与 $\sigma^2 I$ 相差不太远时，LS 估计 $c'\hat{\beta}$ 仍然保持原来的最优性或即便不是最优的，但不要变得很坏，大体上还“过得去”。若是这样的话，我们就说 LS 估计关于协方差阵是稳健的。相反，如果出现失之毫厘，谬之千里的情况，这个估计就不具有稳健性，应用起来就得特别谨慎。稳健性总是相对于模型的某种变化而言的。例如，上面举的例子是 LS 估计关于协方差阵变化的稳健性。我们自然也可以讨论它关于设计阵的稳健性，或者它的某一条性质关于误差分布的稳健性等等。

应该说，稳健性是每一种统计推断都应当具有的性质。因此，统计文献中有了稳健设计，稳健检验等概念。足见稳健性的研究已经渗透到统计学的很多分支。前面已经说过，在某种意义上讲，稳健性就是稳定性。在数学的其它分支，我们也可以找到与之相当的概念。例如，常微分方程中十分重要的稳定性理论，就是专门研究方程的解关于初始条件的稳定性。又如在非线性规划中，也有类似的解的稳定性概念。一节我们主要讨论 LS 估计关于协方差阵的稳健性。考虑线性模型

$$y = X\beta + e, \quad E(e) = 0, \quad \text{Cov} = \sigma^2 \Sigma$$

这里 $\Sigma \geq 0$ 已知、对任一可估函数 $c'\beta$ ，它的 LS 估计为

$$c'\hat{\beta} = c'(X'X)^{-1}X'y$$

我们知道，当 $\text{Cov}(e) = \sigma^2 I$ 时，它是 BLU 估计。现在尽管协方差阵 $\text{Cov}(e) = \sigma^2 \Sigma \neq \sigma^2 I$ ，我们希望 $c'\hat{\beta}$ 关于误差协方差阵的这种变化具有稳健性，即 $c'\hat{\beta}$ 仍然是 BLU 估计：

$$c'\hat{\beta} = c'\beta^* \tag{17.38}$$

这里 β^* 由上节最小二乘统一理论给出（见定理 4.4.4）。下面两个定理回答了这个问题。记 Z 为 $n \times (n-r)$ 且秩为 $n-r$ 的矩阵，满足 $X'Z=0$ ，这里 $r=\text{rk}(X)$ 。不失一般性，以下讨论中假设 $\sigma^2 = 1$ 。定理 4.5.1 对于线性模型 (4.5.1) 和任一可估函数 $c'\beta$ ，(17.38) 成立当且仅当下列条件之一成立。(1)

$$X'\Sigma Z = 0 \tag{17.39}$$

(2)

$$\Sigma = X\Lambda_1 X' + Z\Lambda_2 Z' \tag{17.40}$$

(3)

$$\Sigma = X D_1 X' + Z D_2 Z' + I \tag{17.41}$$

其中 $\Lambda_1, \Lambda_2, D_1$ 和 D_2 为任意对称阵，但使 $\Sigma \geq 0$

Bibliography

[Articles](#)

[Books](#)

