

概率论与数理统计
概率论与数理统计课程讲义
Probability and Mathematical Statistics

前言

课程与致谢

本讲义为笔者暑期学习《概率论与数理统计》期间，根据刘思齐老师的课堂讲授内容整理而成的学习笔记。衷心感谢刘老师一学期以来的精心备课，以及无私分享知识的精神。

刘老师的哔哩哔哩账号为：“我真的不懂分析”。

课程链接 <https://www.bilibili.com/video/BV19LAUz7E4T>

整理说明

本讲义在 Codex 的协助下完成了课堂内容转写、公式校订、图形重绘与 $\text{\LaTeX}$ 排版。在整理过程中，笔者参考《概率论与数理统计教程》的课程脉络，对各章内容进行了重新编排，并尽可能保留课堂中原有的推导过程与思考路径，使内容更加连贯、翔实，便于日后复习与查阅。

编写理念

本讲义希望在课堂笔记与系统教材之间取得适当的平衡：一方面保留学习过程中自然形成的直觉、疑问与思考痕迹，另一方面借助较为规范的数学语言，将零散的认识逐步整理为完整、清晰的知识结构。

正文中标有“◆”的内容为参考书中的补充条目，仅作简要介绍。

限于笔者的知识水平，讲义中难免存在疏漏与错误，敬请读者批评指正。

lim99913

于 2026 年 8 月 2 日夜

El tiempo se bifurca perpetuamente hacia innumerables futuros.

时间永不停息地分岔，通向无数个未来。

——《小径分岔的花园》

目录

第一章 随机事件与概率	8
1.1 随机事件及其运算	8
1.1.1 随机现象、样本空间与随机事件	8
1.1.2 事件间的关系与运算	8
1.1.3 事件域与 Borel 集	9
1.2 概率的定义及其确定方法	10
1.2.1 概率的公理化定义	10
1.2.2 等可能模型与频率思想	10
1.2.3 几何概率	11
1.3 概率的性质	12
1.3.1 单调性、加法公式与包含排斥	12
1.3.2 概率的连续性	13
1.4 条件概率	13
1.4.1 条件概率的定义	13
1.4.2 乘法公式与全概率公式	14
1.4.3 Bayes 公式	14
1.5 独立性	15
1.5.1 两个事件的独立性	15
1.5.2 多个事件的相互独立	16
第二章 随机变量及其分布	17
2.1 随机变量及其分布	17
2.1.1 随机变量的概念	17
2.1.2 分布函数的性质	17
2.1.3 离散型与连续型随机变量	18
2.2 随机变量的数学期望	19
2.2.1 期望的定义	19
2.2.2 期望的存在性与性质	20
2.3 随机变量的方差与标准差	21
2.4 常用离散分布	22
2.4.1 二项分布	22
2.4.2 Poisson 分布	22
2.4.3 超几何分布	23
2.4.4 几何分布与负二项分布	23
2.5 常用连续分布	24
2.5.1 正态分布	24
2.5.2 均匀分布	25
2.5.3 指数分布	25
2.5.4 Gamma 分布与 χ^2 分布	26
2.6 随机变量函数的分布	26
2.6.1 离散情形	26

2.6.2	连续情形：单调变换	27
2.7	分布的其他特征数	28
2.7.1	矩	28
2.7.2	分位数、中位数、变异系数、偏度与峰度	28
第三章	多维随机变量及其分布	29
3.1	多维随机变量与联合分布	29
3.1.1	随机向量与联合分布函数	29
3.1.2	常用多维分布	29
3.1.3	多元正态分布	30
3.2	边际分布与独立性	32
3.2.1	边际分布	32
3.2.2	独立性	34
3.3	多维随机变量函数的分布	34
3.3.1	离散映射、最大值与最小值	34
3.3.2	卷积公式	35
3.3.3	变量变换法	36
3.4	多维随机变量的特征数	39
3.4.1	函数的期望与运算性质	39
3.4.2	协方差、相关系数与协方差矩阵	39
3.5	条件分布与条件期望	40
3.5.1	条件分布	40
3.5.2	条件期望	42
第四章	大数定律与中心极限定理	44
4.1	◆ 随机变量序列的收敛性	44
4.2	◆ 特征函数	45
4.3	大数定律	45
4.3.1	Bernoulli 大数定律	45
4.3.2	Chebyshev、Markov 与 Khinchin 大数定律	46
4.3.3	强大数定律	46
4.4	中心极限定理	47
4.4.1	De Moivre–Laplace 定理	47
4.4.2	Lindeberg–Lévy 定理	47
4.4.3	独立但不同分布的情形	48
第五章	统计量及其分布	49
5.1	总体、样本与统计量	49
5.1.1	经验分布函数	49
5.2	样本均值、方差与矩	50
5.2.1	样本均值	50
5.2.2	样本方差	50
5.2.3	样本矩	51
5.3	次序统计量、分位数与箱线图	51

5.4	三大抽样分布	52
5.4.1	χ^2 分布	52
5.4.2	t 分布	52
5.4.3	F 分布	53
5.5	稳健统计量	54
5.6	充分统计量	54
第六章	参数估计	56
6.1	估计问题与点估计	56
6.2	估计量的评价	56
6.2.1	无偏性与渐近无偏性	56
6.2.2	有效性	57
6.2.3	均方误差	58
6.2.4	相合性	58
6.3	矩估计	59
6.3.1	基本方法	59
6.3.2	广义矩关系	59
6.4	最大似然估计	60
6.4.1	似然函数	60
6.4.2	得分、Fisher 信息与渐近正态性	61
6.5	◆最小方差无偏估计	61
6.6	Bayes 估计	62
6.6.1	先验、似然与后验	62
6.7	区间估计	63
6.7.1	置信区间与可信区间	63
6.7.2	枢轴量法	63
6.7.3	正态总体的区间估计	64
6.7.4	二项比例的大样本区间	65
6.7.5	两个正态总体的均值差	65
第七章	假设检验	66
7.1	统计假设的三种观点	66
7.1.1	参数模型	66
7.1.2	女士品茶	66
7.1.3	假设的形式	66
7.1.4	Fisher 情形: p 值	67
7.1.5	Neyman-Pearson 情形: 拒绝域	67
7.1.6	两类错误与功效函数	67
7.2	正态总体参数的假设检验	68
7.2.1	方差已知时均值的 Z 检验	68
7.2.2	方差未知时均值的 t 检验	69
7.2.3	假设检验与区间估计的关系	70
7.2.4	正态总体方差的检验	70
7.3	其他分布参数的假设检验	70

7.3.1	指数分布均值参数	70
7.3.2	Bernoulli 分布的成功概率	71
7.4	似然比与拟合优度	72
7.4.1	广义似然比检验	72
7.4.2	Pearson 卡方拟合优度检验	73
7.4.3	连续总体的拟合优度	75
7.4.4	◆参考书补充	75

lim99913

第一章 随机事件与概率

1.1 随机事件及其运算

1.1.1 随机现象、样本空间与随机事件

随机现象具有两项基本特征：其结果不能预先完全确定，但对其发生规律可以用概率加以描述。历史上，Hilbert 在 1900 年提出的第六问题涉及概率与物理的公理化；20 世纪 30 年代 Kolmogorov 建立了现代概率论的公理体系。

定义：样本空间

随机试验所有可能结果的集合称为样本空间，记为 Ω ；其中的元素称为样本点，常记为 ω 。

例：样本空间

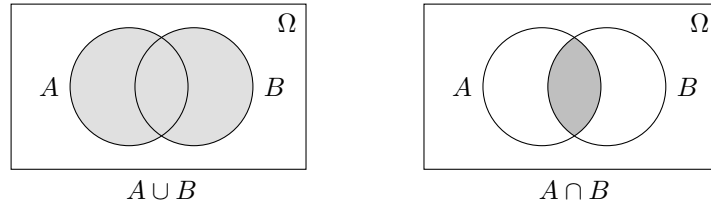
1. 抛一枚硬币： $\Omega = \{\text{正}, \text{反}\}$ ；
2. 掷一枚骰子： $\Omega = \{1, 2, 3, 4, 5, 6\}$ ；
3. 测量某物理量：可取 $\Omega = \mathbb{R}$ 。

样本空间的某些子集称为随机事件。事件发生，指实际出现的样本点落入该子集。若 Ω 有限或可数，常可把 Ω 的所有子集都作为事件；若 $\Omega = \mathbb{R}$ 或 $\mathbb{R}^n$ ，则不能不加限制地把所有子集都作为可测事件。

1.1.2 事件间的关系与运算

设 $A, B \subseteq \Omega$ 。

1. $A \subseteq B$ ：事件 A 发生必然导致 B 发生；
2. $A \subseteq B$ 且 $B \subseteq A$ 当且仅当 $A = B$ ；
3. $A \cap B = \emptyset$ ： A, B 互不相容；
4. $A \cup B$ ： A, B 至少一个发生；
5. $A \cap B$ ： A, B 同时发生；
6. $A^c = \Omega \setminus A$ ： A 的对立事件，且 $A \cap A^c = \emptyset$ ；
7. $\bigcup_{n \geq 1} A_n = \{\omega : \exists n, \omega \in A_n\}$ ；
8. $\bigcap_{n \geq 1} A_n = \{\omega : \forall n, \omega \in A_n\}$ ；
9. $A \setminus B = \{\omega \in A : \omega \notin B\} = A \cap B^c$ ；
10. $A \triangle B = (A \setminus B) \cup (B \setminus A)$ ，表示二者恰有一个发生。



图：事件的并与交

De Morgan 律为

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c,$$

以及可数形式

$$\left(\bigcup_{n=1}^{\infty} A_n \right)^c = \bigcap_{n=1}^{\infty} A_n^c, \quad \left(\bigcap_{n=1}^{\infty} A_n \right)^c = \bigcup_{n=1}^{\infty} A_n^c.$$

1.1.3 事件域与 Borel 集

定义： σ -代数

设 $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ 。若

$$\Omega \in \mathcal{F}, \quad A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}, \quad A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F},$$

则称 $\mathcal{F}$ 为 Ω 上的事件域或 σ -代数， $\mathcal{F}$ 中的元素称为事件。

由补集与可数并封闭可推出： $\mathcal{F}$ 对可数交、差与对称差也封闭。应注意 σ -代数要求的是可数封闭，不是对任意不可数运算封闭。

当 $\Omega = \mathbb{R}$ 时，通常取包含所有开区间（等价地，包含所有 $(-\infty, a)$ ）的最小 σ -代数，称为 Borel σ -代数 $\mathcal{B}(\mathbb{R})$ 。从半无限区间可以依次构造

$$\begin{aligned} [a, \infty) &= (-\infty, a)^c, \\ [a, b] &= [a, \infty) \cap (-\infty, b), \\ [a, b] &= \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b + \frac{1}{n} \right), \\ \{b\} &= [a, b] \setminus [a, b), \\ (a, b) &= [a, b] \setminus \{a, b\}. \end{aligned}$$

说明：为何不能取所有子集

在 $[0, 1]$ 上若把可数个互不相交的区间平移后要求同时满足平移不变性和可数可加性，会导向不可测集问题。Cantor 集虽长度为 0 且与 $\mathbb{R}$ 等势，仍是 Borel 集；真正的障碍来自选择公理下构造的不可测集。这也是实数样本空间上需指定 σ -代数的原因。

1.2 概率的定义及其确定方法

1.2.1 概率的公理化定义

定义：概率空间

在可测空间 $(\Omega, \mathcal{F})$ 上，若映射

$$\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R}, \quad A \mapsto \mathbb{P}(A)$$

满足

$$\mathbb{P}(A) \geq 0, \quad \mathbb{P}(\Omega) = 1,$$

以及对两两互不相容的 $A_1, A_2, \dots \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n),$$

则称 $(\Omega, \mathcal{F}, \mathbb{P})$ 为概率空间。

由公理立即得到

$$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(A) + \mathbb{P}(A^c) = 1, \quad 0 \leq \mathbb{P}(A) \leq 1.$$

例如取 $A_1 = \Omega, A_2 = A_3 = \dots = \emptyset$ 可得 $\mathbb{P}(\emptyset) = 0$ ；取 $A_1 = A, A_2 = A^c$ 可得补事件公式。

推导：由可数可加性推出基本性质

取 $A_1 = \Omega, A_n = \emptyset (n \geq 2)$ ，由两两不相容性，

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \sum_{n=2}^{\infty} \mathbb{P}(\emptyset),$$

故 $\mathbb{P}(\emptyset) = 0$ 。再将 $\Omega = A \dot{\cup} A^c$ 分解为不相交并，得到

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c).$$

若 $A \subseteq B$ ，则 $B = A \dot{\cup} (B \setminus A)$ ，所以

$$\mathbb{P}(B) - \mathbb{P}(A) = \mathbb{P}(B \setminus A) \geq 0.$$

说明：概率空间三要素

Ω 描述所有可能结果； $\mathcal{F}$ 指定哪些集合可作为事件； $\mathbb{P}$ 给这些事件赋予概率。离散有限情形常取 $\mathcal{F} = \mathcal{P}(\Omega)$ ；在 $\Omega = \mathbb{R}$ 时常取 $\mathcal{F} = \mathcal{B}(\mathbb{R})$ 。概率 $\mathbb{P}$ 还需根据具体模型确定。

1.2.2 等可能模型与频率思想

若 $\Omega = \{\omega_1, \dots, \omega_n\}$ 且各样本点等可能，则

$$\mathbb{P}(\{\omega_i\}) = \frac{1}{n}, \quad \mathbb{P}(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}.$$

频率方法以“频率趋于概率”为基础：在大量重复试验中，以事件 A 的发生频率 n_A/n 逼近 $\mathbb{P}(A)$ 。

例：有放回抽样

袋中有 N 个球，其中 n 个红球。每次有放回地抽取，共抽 k 次。令 A 为恰有 ℓ 次抽到红球，则

$$\text{card}(\Omega) = N^k, \quad \text{card}(A) = \binom{k}{\ell} n^{\ell} (N - n)^{k-\ell},$$

故令 $p = n/N$,

$$\mathbb{P}(A) = \binom{k}{\ell} p^{\ell} (1-p)^{k-\ell}.$$

这与 k 次 Bernoulli 试验中成功 ℓ 次的概率一致。

例：无放回抽样与超几何分布

仍有 N 个球，其中 n 个红球，不放回抽取 k 次。恰有 ℓ 个红球的概率为

$$\mathbb{P}(A) = \frac{\binom{n}{\ell} \binom{N-n}{k-\ell}}{\binom{N}{k}}.$$

若保留抽取次序，则分子分母同时乘以 $k!$ ，结果不变。对所有可行 ℓ 求和可得 Vandermonde 恒等式

$$\binom{N}{k} = \sum_{\ell} \binom{n}{\ell} \binom{N-n}{k-\ell}.$$

推导：无放回抽样的计数

若先忽略顺序，所有 k 元样本共有 $\binom{N}{k}$ 种。恰取到 ℓ 个红球时，须从 n 个红球中选 ℓ 个，再从 $N-n$ 个白球中选 $k-\ell$ 个，因此

$$\#A = \binom{n}{\ell} \binom{N-n}{k-\ell}, \quad \mathbb{P}(A) = \frac{\#A}{\#\Omega}.$$

若改用有序样本，则分子、分母同时乘以 $k!$ ，概率不变。对所有可行 ℓ 求和，因这些事件构成样本空间的分割，

$$\sum_{\ell} \binom{n}{\ell} \binom{N-n}{k-\ell} = \binom{N}{k}.$$

例：Bernoulli 试验

单次试验只有“成功/失败”，成功概率为 p ，失败概率为 $1-p$ 。独立重复 k 次，成功 ℓ 次的概率为

$$\binom{k}{\ell} p^{\ell} (1-p)^{k-\ell}.$$

1.2.3 几何概率

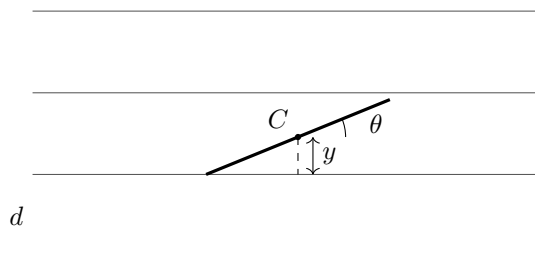
当样本点在区域 Ω 中均匀分布，可用长度、面积或体积 m 定义

$$\mathbb{P}(A) = \frac{m(A)}{m(\Omega)}.$$

例：Buffon 投针

平面上有间距 d 的平行线，随机投下一根长度 $\ell < d$ 的针。以针中心到最近平行线的距离 $y \in [0, d/2]$ 和针与平行线夹角 $\theta \in [0, \pi]$ 参数化。相交条件为 $y \leq (\ell/2) \sin \theta$ ，故

$$\mathbb{P}(\text{相交}) = \frac{\int_0^{\pi} (\ell/2) \sin \theta \, d\theta}{\pi(d/2)} = \frac{2\ell}{\pi d}.$$

图: Buffon 投针的参数 y 与 θ **例: 公平硬币的几何直觉**

若硬币近似圆柱, 直径 D 、厚度 T , 随机落地时“立边”与“正反面”的临界角满足 $\tan \theta = T/D$ 。例如取 $\theta = \pi/3$, 可得 $T = D/\sqrt{3}$ 作为一种几何平衡条件。

例: Bertrand 弦悖论

圆中“随机作一条弦”, 求弦长大于内接等边三角形边长的概率。不同的“随机”机制给出不同答案:

$$\frac{1}{2} \quad (\text{随机中点距圆心}), \quad \frac{1}{3} \quad (\text{随机端点/随机角}), \quad \frac{1}{4} \quad (\text{另一种中点取法}).$$

这说明几何概率必须明确样本空间上的测度, “等可能”不能脱离随机化机制。



图: Bertrand 弦悖论中的三种随机化机制

1.3 概率的性质

1.3.1 单调性、加法公式与包含排斥

若 $A \subseteq B$, 则

$$\mathbb{P}(A) \leq \mathbb{P}(B),$$

因为 $B = A \cup (B \setminus A)$ 且两部分不相容。一般地,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

对有限多个事件,

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i < j} \mathbb{P}(A_i \cap A_j) + \sum_{i < j < k} \mathbb{P}(A_i \cap A_j \cap A_k) - \dots.$$

例: 最大编号问题

有编号 $1, \dots, n$ 的球, 有放回抽取 m 次。令 A_k 为最大编号恰为 k , B_k 为最大编号不超过 k 。则

$$\mathbb{P}(B_k) = \left(\frac{k}{n}\right)^m, \quad \mathbb{P}(A_k) = \mathbb{P}(B_k) - \mathbb{P}(B_{k-1}) = \left(\frac{k}{n}\right)^m - \left(\frac{k-1}{n}\right)^m.$$

等价地也可把 A_k 分解为编号 k 出现 $\ell = 1, \dots, m$ 次并使用二项式定理。

例：错排与秘书问题

把 n 封信随机装入 n 个信封。令 A_i 为第 i 封信装入正确的信封，则“全部错排”的概率为

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^n A_i^c\right) &= 1 - \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \\ &= 1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + \frac{(-1)^n}{n!} \rightarrow e^{-1}.\end{aligned}$$

1.3.2 概率的连续性

结论：连续性

若 $A_1 \supseteq A_2 \supseteq \cdots$ ，则

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

若 $A_1 \subseteq A_2 \subseteq \cdots$ ，则

$$\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right).$$

这与分析中连续函数“极限可交换”的形式相似。

1.4 条件概率

1.4.1 条件概率的定义

定义：条件概率

若 $\mathbb{P}(B) > 0$ ，则在事件 B 已发生的条件下 A 的概率定义为

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

在等可能模型中，它等于 $m(A \cap B)/m(B)$ 。固定 B 后， $(\Omega, \mathcal{F}, \mathbb{P}(\cdot | B))$ （或把样本空间缩为 B ）仍构成概率空间。

例：两个孩子

已知一家有两个孩子。

1. 若已知“至少一个是男孩”，在 $\{bb, bg, gb, gg\}$ 等可能模型中，两个都是男孩的条件概率为 $1/3$ ；
2. 若已知“老大是男孩”，可能结果仅有 $\{bb, bg\}$ ，概率为 $1/2$ ；
3. 若加入“某个孩子是星期二出生的男孩”等筛选条件，条件事件本身发生机制必须说明，否则答案会依赖信息的获得方式。可以用 196 个等可能的“性别 $\times$ 生日”组合说明了这一点。

例：三门问题

奖品随机放在三扇门后，选手先选一扇门，主持人知道奖品位置并打开另一扇必为空门，再允许选手换门。主持人的开门规则携带信息；保持原选的胜率为 $1/3$ ，换门胜率为 $2/3$ 。不能把主持人打开后剩余两门简单看成各 $1/2$ 。

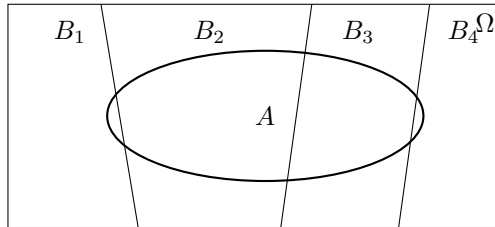
1.4.2 乘法公式与全概率公式

乘法公式：

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A | B) = \mathbb{P}(A)\mathbb{P}(B | A).$$

若 $B_1, \dots, B_n$ 两两不相容、 $\bigcup_{k=1}^n B_k = \Omega$ ，则称其为 Ω 的一个分割，且

$$A = \bigcup_{k=1}^n (A \cap B_k), \quad \mathbb{P}(A) = \sum_{k=1}^n \mathbb{P}(B_k)\mathbb{P}(A | B_k).$$



图：分割 $B_1, \dots, B_n$ 与事件 A 的交

推导：全概率公式

因为 $B_1, \dots, B_n$ 构成分割，

$$A = A \cap \Omega = A \cap \left(\bigcup_{k=1}^n B_k \right) = \bigcup_{k=1}^n (A \cap B_k),$$

且右侧各项互不相容。由有限可加性和乘法公式，

$$\mathbb{P}(A) = \sum_{k=1}^n \mathbb{P}(A \cap B_k) = \sum_{k=1}^n \mathbb{P}(B_k)\mathbb{P}(A | B_k).$$

例：敏感问题的随机化回答

为保护隐私，受访者先抛一枚成功概率为 s 的随机装置：成功则如实回答敏感问题，失败则对一个已知答案概率为 $1/2$ 的无关问题作答。若总体“是”的回答比例为 k/n ，敏感问题真实比例 x 满足

$$\frac{k}{n} = sx + (1-s)\frac{1}{2}, \quad x = \frac{\frac{k}{n} - \frac{1-s}{2}}{s}.$$

例：分层抽样

把全国人口按地域、年龄、性别、学历等分层为 $B_1, \dots, B_n$ 。总体某特征 A 的比例为

$$\mathbb{P}(A) = \sum_{k=1}^n \mathbb{P}(B_k)\mathbb{P}(A | B_k),$$

其中 $\mathbb{P}(B_k)$ 是该层在人口中的权重， $\mathbb{P}(A | B_k)$ 是层内比例。

1.4.3 Bayes 公式

结论：Bayes 公式

若 $B_1, \dots, B_n$ 构成分割，且 $\mathbb{P}(A) > 0$ ，则

$$\mathbb{P}(B_k | A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A | B_k)}{\sum_{\ell=1}^n \mathbb{P}(B_\ell)\mathbb{P}(A | B_\ell)}.$$

$\mathbb{P}(B_k)$ 称为先验概率， $\mathbb{P}(B_k | A)$ 称为后验概率。

推导：从乘法公式到 Bayes 公式

同一个联合事件可按两个方向分解：

$$\mathbb{P}(A \cap B_k) = \mathbb{P}(B_k)\mathbb{P}(A | B_k) = \mathbb{P}(A)\mathbb{P}(B_k | A).$$

于是

$$\mathbb{P}(B_k | A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A | B_k)}{\mathbb{P}(A)}.$$

再将分母 $\mathbb{P}(A)$ 用全概率公式展开，即得到 Bayes 公式。

例：医学检验与基础率

某病患病率为 0.0004。检测对患病者阳性的灵敏度为 0.99，对未患病者阳性的假阳性率为 0.001（实际数值应以具体检测模型为准）。若检测阳性，

$$\mathbb{P}(\text{患病} | \text{阳性}) = \frac{\mathbb{P}(\text{患病})\mathbb{P}(\text{阳性} | \text{患病})}{\mathbb{P}(\text{患病})\mathbb{P}(\text{阳性} | \text{患病}) + \mathbb{P}(\text{未患病})\mathbb{P}(\text{阳性} | \text{未患病})}.$$

极低的先验患病率会使一次阳性后的患病概率远低于直觉。连续两次独立检测阳性时，第一次后验可作为第二次的先验，后验概率会显著提高。

1.5 独立性

1.5.1 两个事件的独立性

定义：独立事件

若

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B),$$

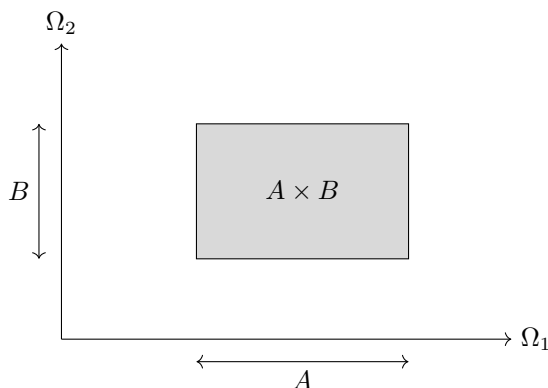
则称 A 与 B 相互独立。若 $\mathbb{P}(B) > 0$ ，这等价于 $\mathbb{P}(A | B) = \mathbb{P}(A)$ 。

可在乘积空间

$$(\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \times (\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$$

中理解独立性：对 $A \in \mathcal{F}_1, B \in \mathcal{F}_2$ ，规定

$$\mathbb{P}(A \times B) = \mathbb{P}_1(A)\mathbb{P}_2(B).$$



图：乘积空间中的矩形事件 $A \times B$

若 A, B 独立，则 A^c 与 B 、 A 与 B^c 、 A^c 与 B^c 均独立。另一方面，事件通常不与自身独立：

$$\mathbb{P}(A \cap A) = \mathbb{P}(A) \neq \mathbb{P}(A)^2,$$

除非 $\mathbb{P}(A) \in \{0, 1\}$ 。互不相容也不等于独立；若 $A \cap B = \emptyset$ 且 $\mathbb{P}(A), \mathbb{P}(B) > 0$ ，则二者不独立。

1.5.2 多个事件的相互独立

定义：相互独立

事件 $A_1, \dots, A_n$ 相互独立，是指对任意 $1 \leq i_1 < \dots < i_k \leq n$,

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \prod_{j=1}^k \mathbb{P}(A_{i_j}).$$

只验证两两独立并不足以推出相互独立。

例：两次抛硬币：两两独立但不相互独立

令 A 为第一次正面， B 为第二次正面， C 为两次结果相同。在 $\Omega = \{HH, HT, TH, TT\}$ 中，

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{4},$$

而各事件概率均为 $1/2$ ，所以任意两事件独立；但

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8},$$

故三者不相互独立。

例：整除事件

在 $\Omega = \{1, \dots, N\}$ 上均匀取数，令 $A = \{2 \mid \omega\}$ ， $B = \{3 \mid \omega\}$ 。当 $N = 6$ 时，

$$\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{3}, \quad \mathbb{P}(A \cap B) = \frac{1}{6},$$

二者独立；当 $N = 10$ 时，

$$\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{3}{10}, \quad \mathbb{P}(A \cap B) = \frac{1}{10},$$

二者不独立。独立性依赖整个概率模型，而不只是事件的文字描述。

例：几何分布的等待时间

每次抽卡成功率为 p ，独立重复。首次成功恰在第 k 次的概率

$$p_k = (1-p)^{k-1}p,$$

在前 k 次内成功的概率

$$q_k = \sum_{i=1}^k p_i = 1 - (1-p)^k.$$

例如，某些参数下可得到 $q_{99} \approx 0.950144$ 、 $q_{198} \approx 0.751142$ 、 $q_{398} \approx 0.900149$ ，用于提醒应按具体 p 与抽数代入，并区分“第 k 次首次成功”和“前 k 次至少一次成功”。

第二章 随机变量及其分布

2.1 随机变量及其分布

2.1.1 随机变量的概念

定义：随机变量

在概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上，函数 $X : \Omega \rightarrow \mathbb{R}$ 若满足对每个 $x \in \mathbb{R}$,

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F},$$

则称 X 为随机变量。该条件保证所有阈值事件都有定义良好的概率。

定义：分布函数

随机变量 X 的分布函数定义为

$$F_X(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

若 X 的分布函数为 F ，记作 $X \sim F$ 。

例：两次抛硬币

令 X 为正面出现次数。若 $\Omega = \{HH, HT, TH, TT\}$ 等可能，则

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0,$$

从而

$$F_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{4}, & 0 \leq x < 1, \\ \frac{3}{4}, & 1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

2.1.2 分布函数的性质

结论：分布函数的刻画

任一分布函数 F 满足：

1. $0 \leq F(x) \leq 1$;
2. F 单调不减;
3. F 右连续: $\lim_{t \downarrow x} F(t) = F(x)$;
4. $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow +\infty} F(x) = 1$.

反之，满足以上性质的函数必是某个随机变量的分布函数。

区间概率需注意端点：

$$\mathbb{P}(a < X \leq b) = F(b) - F(a), \quad \mathbb{P}(a \leq X \leq b) = F(b) - F(a-),$$

其中 $F(a-) = \lim_{x \uparrow a} F(x)$ ，且

$$\mathbb{P}(X = a) = F(a) - F(a-).$$

2.1.3 离散型与连续型随机变量

若 X 只取有限或可数个值 $x_1, x_2, \dots$ ，称其为离散型随机变量。记

$$p_k = \mathbb{P}(X = x_k), \quad p_k \geq 0, \quad \sum_k p_k = 1,$$

则

$$F_X(x) = \sum_{k: x_k \leq x} p_k.$$

离散型分布函数是阶梯函数，但其跳点可以无限多个、甚至稠密。

若存在非负可积函数 $p(x)$ 使

$$F_X(x) = \int_{-\infty}^x p(t) dt, \quad \int_{-\infty}^{\infty} p(x) dx = 1,$$

则称 X 为连续型随机变量， p 为概率密度函数 (p.d.f.)。此时 $F'_X(x) = p(x)$ 在几乎处处成立，并且单点概率 $\mathbb{P}(X = x) = 0$ 。但“概率为 0”不代表事件逻辑上不可能；“概率为 1”也不等于逻辑上必然。

例：区间均匀分布

若样本在 $[a, b]$ 上均匀，且 $X(\omega) = \omega$ ，则

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

端点取值不影响连续型分布。

例：Cauchy 分布

$$p(x) = \frac{1}{\pi(1+x^2)}, \quad F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x.$$

它是合法密度，但 $\int |x|p(x) dx = \infty$ ，故数学期望不存在。

例：Wigner 半圆律

$$p(x) = \begin{cases} \frac{1}{2\pi} \sqrt{4-x^2}, & |x| \leq 2, \\ 0, & |x| > 2. \end{cases}$$

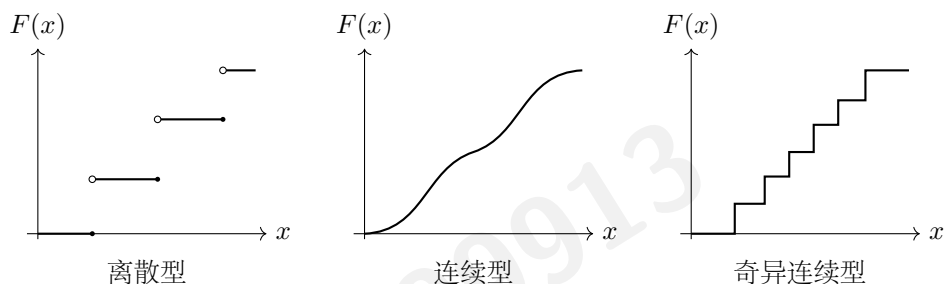
以随机实对称矩阵 $A = A^T$ （非对角元素取 ± 1 ）为例：适当归一化后的特征值经验分布在维数趋于无穷时趋近半圆律。

例：Cantor 随机变量

独立抛硬币无穷次，令 $Y_n \in \{0, 1\}$ ，并定义

$$X = \sum_{n=1}^{\infty} \frac{2Y_n}{3^n}.$$

X 的值域是 Cantor 集。其分布函数是 Cantor 魔鬼阶梯：连续但不绝对连续，且 $F'(x) = 0$ 几乎处处，因此不存在密度 p 使 $F(x) = \int_{-\infty}^x p(t) dt$ 。这说明随机变量不必是离散型或连续型。



图：三类典型分布函数的形状

2.2 随机变量的数学期望

2.2.1 期望的定义

定义：数学期望

若 X 为离散型，取值 x_k 、概率 p_k ，且 $\sum_k |x_k| p_k < \infty$ ，则

$$\mathbb{E}X = \sum_k x_k p_k.$$

若 X 为连续型，密度为 $p(x)$ ，且 $\int_{\mathbb{R}} |x| p(x) dx < \infty$ ，则

$$\mathbb{E}X = \int_{\mathbb{R}} x p(x) dx.$$

一般地可统一写为 Riemann–Stieltjes 积分

$$\mathbb{E}X = \int_{\mathbb{R}} x dF_X(x).$$

期望描述长期平均，但不是样本空间中“必定出现”的数，也不必是随机变量的某个可能取值。

例：分赌本问题

甲乙约定先赢三局者得 100 元，已知甲乙单局胜率之比为 2:1。比赛中断时，应按当前状态下双方最终获胜的条件概率分配赌本，而不是简单按已赢局数分配。可以把尚余的可能序列 $\omega_1, \omega_2, \omega_3$ 作为样本空间，按各序列概率计算最终收益 X ，再以

$$\mathbb{E}X = \sum_k \mathbb{P}(\omega_k)X(\omega_k)$$

求公平分配。这是 Pascal 与 Fermat 讨论的经典期望思想。

2.2.2 期望的存在性与性质

Cauchy 分布说明密度存在并不保证期望存在。另一个离散例子是：

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

虽然概率和为 1，但

$$\mathbb{E}X = \sum_{n=1}^{\infty} 2^n 2^{-n} = \infty.$$

结论：期望的线性

若相应期望存在，则

$$\mathbb{E}(c) = c, \quad \mathbb{E}(aX) = a\mathbb{E}X, \quad \mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y.$$

线性不要求 X, Y 独立。

例：最大编号的期望

从 $1, \dots, n$ 中不放回抽取 m 个数，令 X 为最大编号。则对 $k = m, \dots, n$,

$$\mathbb{P}(X = k) = \frac{\binom{k-1}{m-1}}{\binom{n}{m}},$$

于是

$$\begin{aligned} \mathbb{E}X &= \frac{1}{\binom{n}{m}} \sum_{k=m}^n k \binom{k-1}{m-1} \\ &= \frac{m}{\binom{n}{m}} \sum_{k=m}^n \binom{k}{m} = \frac{m(n+1)}{m+1}. \end{aligned}$$

因此由观测最大编号 k 可得到总体大小的矩估计形式 $\hat{n} = \frac{m+1}{m}k - 1$ 。

推导：最大编号分布的来源

要使最大编号等于 k ，抽出的 m 个数中必须包含 k ，其余 $m-1$ 个数只能从 $1, \dots, k-1$ 中选择。因此

$$\#\{X = k\} = \binom{k-1}{m-1}, \quad \mathbb{P}(X = k) = \frac{\binom{k-1}{m-1}}{\binom{n}{m}}.$$

再利用

$$k \binom{k-1}{m-1} = m \binom{k}{m}, \quad \sum_{k=m}^n \binom{k}{m} = \binom{n+1}{m+1},$$

便得到 $\mathbb{E}X = m(n+1)/(m+1)$ 。

例：St. Petersburg 悖论

连续抛公平硬币，首次正面出现在第 n 次时收益为 2^n ，故

$$\mathbb{P}(X = 2^n) = 2^{-n}, \quad \mathbb{E}X = \infty.$$

但现实参与者不会支付无限入场费。Bernoulli 以凹效用函数 $u(w)$ 解释风险厌恶，考察

$$\mathbb{E}[u(W - C + X)] \geq u(W)$$

来决定可接受费用 C ；例如可取 $u(w) = \log w$ 。该例说明财富的边际效用递减。

2.3 随机变量的方差与标准差

定义：方差与标准差

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2], \quad \sigma(X) = \sqrt{\text{Var}(X)}.$$

标准差与 X 具有相同量纲，方差则具有平方量纲。

结论：方差的性质

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2,$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X),$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\mathbb{E}(XY) - 2\mathbb{E}X \mathbb{E}Y.$$

若 $X = c$ 几乎处处，则 $\text{Var}(X) = 0$ ；反之，若 $\text{Var}(X) = 0$ ，则 $X = \mathbb{E}X$ 几乎处处。

最后一个反向结论可由 Chebyshev 不等式推出：令 $a = \mathbb{E}X$ ，对每个 $\varepsilon > 0$ ，

$$\mathbb{P}(|X - a| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2} = 0.$$

再用

$$\{X \neq a\} = \bigcup_{n=1}^{\infty} \{|X - a| \geq 1/n\}$$

得到 $\mathbb{P}(X = a) = 1$ 。

结论：Chebyshev 不等式

若 $\mathbb{E}X$ 和 $\text{Var}(X)$ 存在，则对 $\varepsilon > 0$ ，

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

证明只用到在集合 $|x - a| \geq \varepsilon$ 上 $(x - a)^2/\varepsilon^2 \geq 1$ ：

$$\begin{aligned} \mathbb{P}(|X - a| \geq \varepsilon) &= \int_{|x-a| \geq \varepsilon} p(x) \, dx \\ &\leq \int_{|x-a| \geq \varepsilon} \frac{(x-a)^2}{\varepsilon^2} p(x) \, dx \\ &\leq \frac{1}{\varepsilon^2} \int_{\mathbb{R}} (x-a)^2 p(x) \, dx. \end{aligned}$$

取 $a = \mathbb{E}X$ 即得结论。更一般地，

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

所以均方误差在 $a = \mathbb{E}X$ 时最小，最小值为 $\text{Var}(X)$ 。

2.4 常用离散分布

2.4.1 二项分布

在 n 重独立 Bernoulli 试验中, 每次成功概率为 p , 令 X 为成功次数, 则

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

记作 $X \sim \text{Bin}(n, p)$ 。二项式定理保证概率和为 1。

利用恒等式

$$k \binom{n}{k} = n \binom{n-1}{k-1}$$

可得

$$\mathbb{E}X = np.$$

再由 $k^2 = k(k-1) + k$,

$$\mathbb{E}X^2 = n(n-1)p^2 + np,$$

故

$$\text{Var}(X) = np(1-p).$$

推导：二项分布的前两阶矩

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

同理, 先算阶乘矩:

$$\mathbb{E}[X(X-1)] = n(n-1)p^2.$$

因为 $X^2 = X(X-1) + X$,

$$\mathbb{E}X^2 = n(n-1)p^2 + np,$$

从而

$$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = np(1-p).$$

2.4.2 Poisson 分布

若 n 很大、 p 很小, 而 $np = \lambda$ 保持适中, 则

$$\binom{n}{k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \left(\frac{\lambda}{n}\right)^k \longrightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

于是定义

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

记作 $X \sim \text{Pois}(\lambda)$ 。由指数级数, 概率和为 1, 并且

$$\mathbb{E}X = \lambda, \quad \mathbb{E}X^2 = \lambda^2 + \lambda, \quad \text{Var}(X) = \lambda.$$

推导：二项分布的 Poisson 极限

固定 k , 令 $p_n = \lambda/n$, 则

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.\end{aligned}$$

当 $n \rightarrow \infty$ 时, 四个因子依次趋于 $1, \lambda^k/k!, e^{-\lambda}, 1$, 所以

$$\mathbb{P}(X_n = k) \rightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

2.4.3 超几何分布

总体有 N 个个体, 其中 M 个为“成功”。不放回抽取 n 个, 令 X 为成功个数, 则

$$\mathbb{P}(X = k) = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}},$$

其中组合数越界时约定为 0。利用组合恒等式可得

$$\mathbb{E}X = n \frac{M}{N},$$

以及

$$\mathbb{E}X^2 = n \frac{M}{N} + \frac{M(M-1)n(n-1)}{N(N-1)}.$$

因此

$$\text{Var}(X) = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}.$$

最后的有限总体修正因子体现了无放回抽样中的负相关。

推导：超几何分布的期望

把 X 写成 n 个指示变量之和:

$$X = I_1 + \cdots + I_n, \quad I_j = \mathbf{1}\{\text{第 } j \text{ 次抽中成功个体}\}.$$

由对称性 $\mathbb{P}(I_j = 1) = M/N$, 故不必直接计算组合数求和,

$$\mathbb{E}X = \sum_{j=1}^n \mathbb{E}I_j = n \frac{M}{N}.$$

对 $i \neq j$,

$$\mathbb{E}(I_i I_j) = \mathbb{P}(I_i = I_j = 1) = \frac{M}{N} \frac{M-1}{N-1}.$$

把 $X^2 = \sum I_j + 2 \sum_{i < j} I_i I_j$ 代入即可得到二阶矩与方差。

2.4.4 几何分布与负二项分布

无限独立 Bernoulli 试验中, 首次成功出现的试验次数 X 满足

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

记作 $X \sim \text{Geom}(p)$ 。其无记忆性为

$$\mathbb{P}(X > m+n \mid X > m) = \mathbb{P}(X > n),$$

因为 $\mathbb{P}(X > n) = (1-p)^n$ 。

第 r 次成功出现的试验次数 X_r 满足

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

记作 $X_r \sim \text{NB}(r, p)$ (Pascal 分布)。由负二项级数可验证概率和为 1, 并可得

$$\mathbb{E}X_r = \frac{r}{p}, \quad \text{Var}(X_r) = \frac{r(1-p)}{p^2}.$$

2.5 常用连续分布

2.5.1 正态分布

若

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

则称 $X \sim N(\mu, \sigma^2)$ 。标准化变量

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$

标准正态分布函数记为

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

由对称性,

$$\Phi(0) = \frac{1}{2}, \quad \Phi(-x) = 1 - \Phi(x),$$

并且

$$\mathbb{P}(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

对 $X \sim N(\mu, \sigma^2)$,

$$\mathbb{E}X = \mu, \quad \mathbb{E}X^2 = \mu^2 + \sigma^2, \quad \text{Var}(X) = \sigma^2.$$

其中用到

$$\int_{\mathbb{R}} y^2 e^{-y^2/2} dy = \sqrt{2\pi},$$

可由分部积分得到。

经验上的 68-95-99.7 规则为

$$\mathbb{P}(|X - \mu| \leq \sigma) \approx 0.6826,$$

$$\mathbb{P}(|X - \mu| \leq 2\sigma) \approx 0.9545,$$

$$\mathbb{P}(|X - \mu| \leq 3\sigma) \approx 0.9973.$$

说明：de Moivre–Laplace 思想

对二项变量 $X_n \sim \text{Bin}(n, p)$,

$$\frac{X_n - np}{\sqrt{np(1-p)}}$$

在 n 大时近似标准正态。这给出了从二项分布过渡到正态分布的经典途径。

推导：正态分布的期望与方差

令 $Z = (X - \mu)/\sigma \sim N(0, 1)$ 。标准正态密度关于原点对称, 因此

$$\mathbb{E}Z = \int_{-\infty}^{\infty} z \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz = 0.$$

又由分部积分,

$$\begin{aligned}\mathbb{E}Z^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \left(-ze^{-z^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-z^2/2} dz \right) = 1.\end{aligned}$$

所以

$$\mathbb{E}X = \mathbb{E}(\mu + \sigma Z) = \mu, \quad \text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2.$$

2.5.2 均匀分布

若

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b], \end{cases}$$

则记作 $X \sim U(a, b)$ 。标准化

$$\frac{X-a}{b-a} \sim U(0, 1),$$

且

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

2.5.3 指数分布

若

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

则记作 $X \sim \text{Exp}(\lambda)$ 。其

$$F(x) = \begin{cases} 0, & x \leq 0, \\ 1 - e^{-\lambda x}, & x > 0, \end{cases} \quad \mathbb{P}(X > x) = e^{-\lambda x},$$

并且

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

指数分布具有无记忆性:

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t).$$

例: Poisson 过程的等待时间

若单位时间事件数服从 $\text{Pois}(\lambda)$, 则长度为 t 的区间内事件数服从 $\text{Pois}(\lambda t)$ 。首次事件等待时间 T 满足

$$\mathbb{P}(T > t) = \mathbb{P}(N(t) = 0) = e^{-\lambda t},$$

故 $T \sim \text{Exp}(\lambda)$ 。

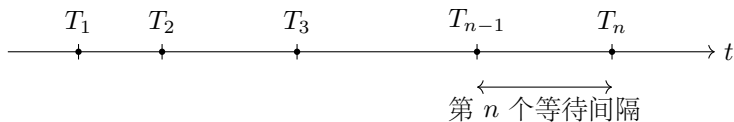


图: Poisson 过程中到达时刻与等待时间

2.5.4 Gamma 分布与 χ^2 分布

Gamma 函数定义为

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx, \quad \alpha > 0.$$

它满足

$$\Gamma(1) = 1, \quad \Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \quad \Gamma(n) = (n-1)!, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

若

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

则称 $X \sim \text{Gamma}(\alpha, \lambda)$ (此处 λ 为率参数)。当 $\alpha = 1$ 时即指数分布。

在 Poisson 过程中, 第 n 次事件发生时刻 T_n 的分布函数可写为

$$F_{T_n}(t) = \mathbb{P}(N(t) \geq n) = \sum_{k=n}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!}.$$

逐项求导并消去望远镜项, 得到

$$f_{T_n}(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t},$$

即 $T_n \sim \text{Gamma}(n, \lambda)$ 。

当 $\alpha = n/2, \lambda = 1/2$ 时,

$$p(x) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0,$$

这是自由度为 n 的 χ^2 分布, 是后续数理统计中的重要抽样分布。

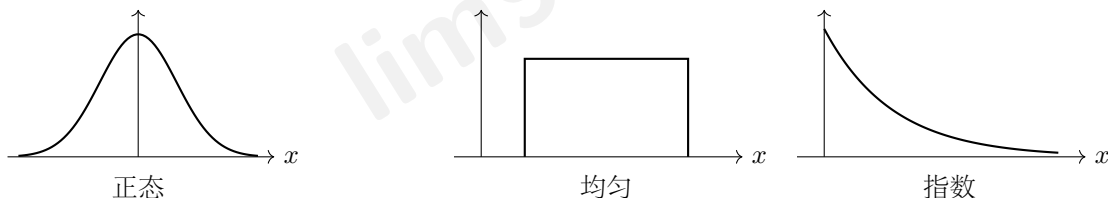


图: 正态、均匀与指数密度

2.6 随机变量函数的分布

2.6.1 离散情形

若 $Y = g(X)$, 而 X 的取值为 $x_1, x_2, \dots$, 则 Y 的可能取值为 $y_i = g(x_i)$ 。不同 x_i 可能映到同一个 y , 因此

$$\mathbb{P}(Y = y) = \sum_{i: g(x_i)=y} \mathbb{P}(X = x_i).$$

例如阈值函数

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

使 Y 成为 Bernoulli 变量, 且

$$\mathbb{P}(Y = 0) = \mathbb{P}(X < a), \quad \mathbb{P}(Y = 1) = \mathbb{P}(X \geq a).$$

2.6.2 连续情形：单调变换

结论：一维单调变换

设 X 有密度 p_X , $Y = g(X)$ 。若 g 严格单调且可微, 反函数为 $h = g^{-1}$, 则在 Y 的值域内

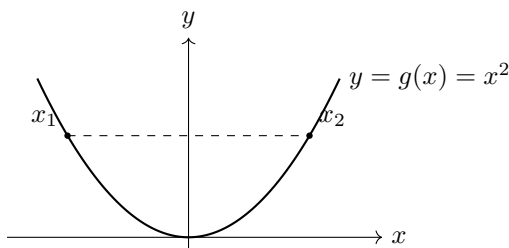
$$p_Y(y) = p_X(h(y)) |h'(y)|,$$

值域外密度为 0。

若 g 递增,

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq h(y)) = \int_{-\infty}^{h(y)} p_X(x) dx,$$

对 y 求导即得公式。若 g 递减, 绝对值处理方向反转。若 g 不单调, 应对所有反函数分支求和。



图：非单调变换中同一 y 对应多个反函数分支

推导：非单调变换的多分支公式

若方程 $g(x) = y$ 有有限个简单根 $x_1(y), \dots, x_m(y)$, 可把定义域切成若干单调区间。在每个区间上应用单调变换公式, 再把互不相容分支的概率相加:

$$p_Y(y) = \sum_{j=1}^m p_X(x_j(y)) \left| \frac{dx_j(y)}{dy} \right|.$$

例如 $Y = X^2$ 在 $y > 0$ 时有两个分支 $x_1 = \sqrt{y}$ 、 $x_2 = -\sqrt{y}$ 。

例：正态变量的线性变换

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

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

例：对数正态分布

若 $X \sim N(\mu, \sigma^2)$, 令 $Y = e^X$, 则 $Y > 0$ 且

$$p_Y(y) = \frac{1}{\sqrt{2\pi\sigma y}} \exp \left[-\frac{(\ln y - \mu)^2}{2\sigma^2} \right], \quad y > 0.$$

该分布记为对数正态分布。

结论：概率积分变换

若随机变量 X 的分布函数 F 连续, 则

$$U = F(X) \sim U(0, 1).$$

反过来, 若 $U \sim U(0, 1)$, 则以广义反函数 $X = F^{-1}(U)$ 可生成分布函数为 F 的随机变量。

例：标准正态平方

若 $X \sim N(0, 1)$, $Y = X^2$, 则当 $y > 0$,

$$F_Y(y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = 2\Phi(\sqrt{y}) - 1.$$

故

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

即 $Y \sim \text{Gamma}(1/2, 1/2) = \chi^2(1)$ 。

2.7 分布的其他特征数

2.7.1 矩

k 阶原点矩与中心矩分别为

$$\mu_k = \mathbb{E}(X^k), \quad \nu_k = \mathbb{E}[(X - \mathbb{E}X)^k].$$

特别地,

$$\mu_1 = \mathbb{E}X, \quad \nu_2 = \text{Var}(X) = \mu_2 - \mu_1^2.$$

由二项式展开,

$$\nu_k = \sum_{\ell=0}^k \binom{k}{\ell} \mu_\ell (-\mu_1)^{k-\ell}.$$

2.7.2 分位数、中位数、变异系数、偏度与峰度

对连续分布, p 分位数 x_p 可由

$$F_X(x_p) = p, \quad 0 < p < 1$$

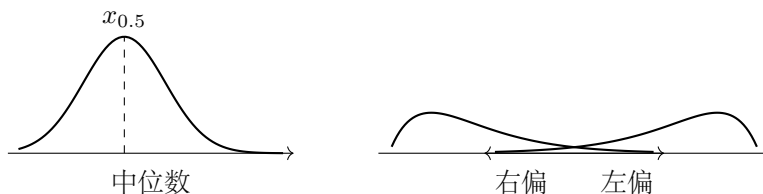
确定; 若有平台区间, 分位数可能不唯一, 可采用广义反函数约定。中位数是 $p = 1/2$ 分位数。

变异系数定义为

$$\text{CV}(X) = \frac{\sigma(X)}{\mathbb{E}X},$$

它消除量纲影响, 便于比较取值尺度不同的变量的相对离散程度 (通常要求 $\mathbb{E}X > 0$)。

偏度刻画分布的不对称性, 峰度刻画尾部与尖峭程度。可用左右长尾图示区分正偏与负偏, 并以不同尖峭程度的曲线说明峰度比较。



图：中位数与偏度的图形直观

第三章 多维随机变量及其分布

3.1 多维随机变量与联合分布

3.1.1 随机向量与联合分布函数

定义：随机向量

设 $(\Omega, \mathcal{F}, \mathbb{P})$ 为概率空间, $X_1, \dots, X_n$ 为其上的随机变量, 则

$$\mathbf{X} = (X_1, \dots, X_n)^T : \Omega \longrightarrow \mathbb{R}^n$$

称为 n 维随机变量或随机向量。

联合分布函数定义为

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

它对每个坐标单调不减、右连续, 并满足适当的边界条件。二维情形还要求任意矩形 $(a_1, b_1] \times (a_2, b_2]$ 的概率非负, 即

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

若 $\mathbf{X}$ 为离散型, 记

$$p_{i_1 \dots i_n} = \mathbb{P}(X_1 = x_{1i_1}, \dots, X_n = x_{ni_n}), \quad \sum_{i_1, \dots, i_n} p_{i_1 \dots i_n} = 1.$$

二维联合分布列常写成表格; 每个单元格是联合概率, 行和与列和分别给出边际概率。

若存在非负函数 $p_{\mathbf{X}}$ 使

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p_{\mathbf{X}}(t_1, \dots, t_n) dt_n \cdots dt_1,$$

且 $\int_{\mathbb{R}^n} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = 1$, 则称其为联合密度。

3.1.2 常用多维分布

例：独立指数变量

若 $X \sim \text{Exp}(\lambda_1)$ 、 $Y \sim \text{Exp}(\lambda_2)$ 且独立, 则

$$p_{X,Y}(x, y) = \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y} \mathbf{1}_{\{x>0, y>0\}}.$$

支持域是第一象限这一矩形型集合, 这与密度的乘积分解相一致。

例：多项分布

一次试验有 k 类结果, 其概率为 $p_1, \dots, p_k$, 且 $\sum_{j=1}^k p_j = 1$ 。独立重复 n 次, 令 X_j 为第 j 类出现次

数, 则

$$\mathbb{P}(X_1 = n_1, \dots, X_k = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k},$$

其中 $n_j \geq 0$ 且 $\sum_j n_j = n$. 归一化来自多项式定理:

$$\sum_{\sum n_j = n} \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} = (p_1 + \dots + p_k)^n = 1.$$

例: 多元超几何分布

总体分成 k 类, 第 j 类有 N_j 个, 共 $N = \sum_j N_j$ 个. 不放回抽取 n 个, 令 X_j 为抽到的第 j 类个数, 则

$$\mathbb{P}(X_1 = n_1, \dots, X_k = n_k) = \frac{\prod_{j=1}^k \binom{N_j}{n_j}}{\binom{N}{n}}, \quad \sum_j n_j = n.$$

例: 区域上的均匀分布

若平面区域 D 的面积为 S_D , 则在 D 上均匀分布的密度为

$$p(x, y) = \frac{1}{S_D} \mathbf{1}_{\{(x, y) \in D\}}.$$

因此子区域 $A \subseteq D$ 的概率就是面积比 S_A/S_D .

3.1.3 多元正态分布

定义: 多元正态分布

若 $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ 为正定矩阵, 且

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \sqrt{\det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

则记 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

推导: 从谱分解确定归一化常数

先不预设常数, 写成

$$p_{\mathbf{X}}(\mathbf{x}) = C \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right].$$

思路是先用正交变换消去交叉项, 再把 n 维积分拆成 n 个一维高斯积分. 因为 $\boldsymbol{\Sigma}$ 是实对称正定矩阵, 谱定理保证存在正交矩阵 P 及正对角矩阵

$$\Lambda = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$$

使

$$\boldsymbol{\Sigma} = P^\top \Lambda P, \quad \boldsymbol{\Sigma}^{-1} = P^\top \Lambda^{-1} P, \quad P^\top P = P P^\top = I.$$

令

$$\mathbf{y} = P(\mathbf{x} - \boldsymbol{\mu}).$$

由于 P 正交, $|\det P| = 1$, 所以平移和旋转都不改变体积元: $d\mathbf{y} = d\mathbf{x}$. 同时

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= (\mathbf{x} - \boldsymbol{\mu})^\top P^\top \Lambda^{-1} P (\mathbf{x} - \boldsymbol{\mu}) \\ &= \mathbf{y}^\top \Lambda^{-1} \mathbf{y} = \sum_{j=1}^n \frac{y_j^2}{\sigma_j^2}. \end{aligned}$$

于是归一化条件逐坐标分离：

$$\begin{aligned} 1 &= C \int_{\mathbb{R}^n} \exp \left[-\frac{1}{2} \sum_{j=1}^n \frac{y_j^2}{\sigma_j^2} \right] dy_1 \cdots dy_n \\ &= C \prod_{j=1}^n \left(\int_{\mathbb{R}} e^{-y_j^2/(2\sigma_j^2)} dy_j \right) \\ &= C \prod_{j=1}^n (\sqrt{2\pi} \sigma_j) = C(2\pi)^{n/2} \prod_{j=1}^n \sigma_j. \end{aligned}$$

而

$$\det \Sigma = \det(P^T) \det \Lambda \det P = \det \Lambda = \prod_{j=1}^n \sigma_j^2,$$

故 $\prod_j \sigma_j = \sqrt{\det \Sigma}$ ，最终

$$C = \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma}}.$$

推导：继续标准化：把椭球化为球

上一步只做了旋转。再令

$$\mathbf{z} = \Lambda^{-1/2} \mathbf{y} = \Lambda^{-1/2} P(\mathbf{x} - \boldsymbol{\mu}),$$

则

$$\mathbf{x} = \boldsymbol{\mu} + P^T \Lambda^{1/2} \mathbf{z}.$$

记 $A = P^T \Lambda^{1/2}$ ，就得到

$$\Sigma = AA^T, \quad \mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}, \quad \mathbf{Z} \sim N_n(\mathbf{0}, I).$$

在中间一步，

$$\mathbf{Y} = P(\mathbf{X} - \boldsymbol{\mu}) \sim N_n(\mathbf{0}, \Lambda).$$

因此旋转到特征向量坐标后，协方差矩阵成为对角阵；各坐标不相关。由于这里是联合正态，不相关进一步推出独立。再除以各自的标准差 σ_j ，才得到相互独立的标准正态坐标 $\mathbf{Z}$ 。这里

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= \mathbf{z}^T \mathbf{z}, \\ \left| \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \right| &= |\det A| = \sqrt{\det(AA^T)} = \sqrt{\det \Sigma}. \end{aligned}$$

所以一般多元正态可以看成：先取标准正态球形云团，再由 A 作伸缩与旋转，最后平移到 $\boldsymbol{\mu}$ 。反过来， $\mathbf{z} = A^{-1}(\mathbf{x} - \boldsymbol{\mu})$ 就是白化。协方差也直接验证为

$$\text{Cov}(\mathbf{X}) = \text{Cov}(A\mathbf{Z}) = A \text{Cov}(\mathbf{Z}) A^T = AA^T = \Sigma.$$

说明：二次型与等密度椭球

固定指数中的二次型

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c$$

得到等密度椭球。 Σ 的特征向量给出主轴方向，特征值 σ_j^2 的平方根 σ_j 给出各主轴相对长度；这正是上面“旋转后逐轴伸缩”的几何含义。

推导：二维参数形式从矩阵公式得到

设

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

先算行列式：

$$\det \Sigma = \sigma_X^2 \sigma_Y^2 - \rho^2 \sigma_X^2 \sigma_Y^2 = \sigma_X^2 \sigma_Y^2 (1 - \rho^2).$$

正定性要求 $\sigma_X, \sigma_Y > 0$ 且 $|\rho| < 1$ 。由二阶矩阵求逆公式，

$$\Sigma^{-1} = \frac{1}{\sigma_X^2 \sigma_Y^2 (1 - \rho^2)} \begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}.$$

令 $u = x - \mu_X$ 、 $v = y - \mu_Y$ ，则

$$\begin{pmatrix} u \\ v \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{1 - \rho^2} \left(\frac{u^2}{\sigma_X^2} - 2\rho \frac{uv}{\sigma_X\sigma_Y} + \frac{v^2}{\sigma_Y^2} \right).$$

再将 $\sqrt{\det \Sigma} = \sigma_X\sigma_Y\sqrt{1 - \rho^2}$ 代回一般密度，得到二维正态的常用参数形式：

$$p(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} \exp \left[-\frac{1}{2(1 - \rho^2)} \left\{ \frac{(x - \mu_X)^2}{\sigma_X^2} - 2\rho \frac{(x - \mu_X)(y - \mu_Y)}{\sigma_X\sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \right\} \right].$$

3.2 边际分布与独立性

3.2.1 边际分布

从联合分布函数取极限即可得到边际分布：

$$F_X(x) = \lim_{y \rightarrow +\infty} F_{X,Y}(x, y), \quad F_Y(y) = \lim_{x \rightarrow +\infty} F_{X,Y}(x, y).$$

离散情形对另一坐标求和；连续情形对另一坐标积分：

$$p_X(x) = \sum_j p_{X,Y}(x, y_j) \quad \text{或} \quad p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x, y) dy.$$

例：三角形区域的边际密度

设 (X, Y) 在三角形

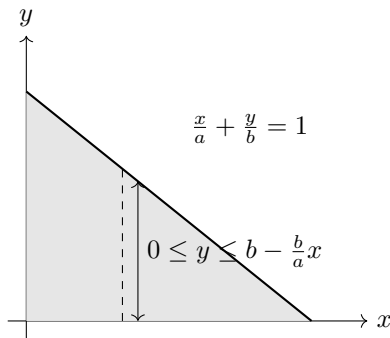
$$D = \{(x, y) : x \geq 0, y \geq 0, x/a + y/b \leq 1\}$$

上均匀。因 $S_D = ab/2$ ，联合密度为 $2/(ab)$ 。固定 $x \in [0, a]$ 时， $0 \leq y \leq b(1 - x/a)$ ，故

$$p_X(x) = \int_0^{b(1-x/a)} \frac{2}{ab} dy = \frac{2}{a} \left(1 - \frac{x}{a}\right), \quad 0 \leq x \leq a.$$

同理

$$p_Y(y) = \frac{2}{b} \left(1 - \frac{y}{b}\right), \quad 0 \leq y \leq b.$$



图：三角形支持域与固定 x 时的边际积分区间

结论：多元正态的边际分布

若

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right],$$

则 $\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$, $\mathbf{X}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_{22})$ 。

推导：沿最后一个坐标消元：边际仍为正态

先按“前 $n-1$ 个坐标与最后一个坐标”分块。记

$$\mathbf{x} = \begin{pmatrix} \tilde{\mathbf{x}} \\ x_n \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \tilde{\boldsymbol{\mu}} \\ \mu_n \end{pmatrix}, \quad \mathbf{y} = \mathbf{x} - \boldsymbol{\mu} = \begin{pmatrix} \tilde{\mathbf{y}} \\ y_n \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \tilde{\Sigma} & \mathbf{a} \\ \mathbf{a}^\top & \sigma_n^2 \end{pmatrix},$$

其中 $\tilde{\Sigma}$ 是正定的 $(n-1)$ 阶主子矩阵。为了把 y_n 从二次型中分离出来，定义 Schur 补

$$\tilde{\sigma}^2 = \sigma_n^2 - \mathbf{a}^\top \tilde{\Sigma}^{-1} \mathbf{a} > 0.$$

关键一步不是直接背分块逆公式，而是先把 Σ 作三因子分解：

$$\Sigma = \begin{pmatrix} I & 0 \\ \mathbf{a}^\top \tilde{\Sigma}^{-1} & 1 \end{pmatrix} \begin{pmatrix} \tilde{\Sigma} & 0 \\ 0 & \tilde{\sigma}^2 \end{pmatrix} \begin{pmatrix} I & \tilde{\Sigma}^{-1} \mathbf{a} \\ 0 & 1 \end{pmatrix}.$$

三者依次求逆并反序相乘，得到

$$\Sigma^{-1} = \begin{pmatrix} I & -\tilde{\Sigma}^{-1} \mathbf{a} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}^{-1} & 0 \\ 0 & 1/\tilde{\sigma}^2 \end{pmatrix} \begin{pmatrix} I & 0 \\ -\mathbf{a}^\top \tilde{\Sigma}^{-1} & 1 \end{pmatrix}.$$

先让最右侧矩阵作用在 $\mathbf{y}$ 上：

$$\begin{pmatrix} I & 0 \\ -\mathbf{a}^\top \tilde{\Sigma}^{-1} & 1 \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{y}} \\ y_n \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{y}} \\ y_n - \mathbf{a}^\top \tilde{\Sigma}^{-1} \tilde{\mathbf{y}} \end{pmatrix}.$$

于是自然出现新的坐标

$$z = y_n - \mathbf{a}^\top \tilde{\Sigma}^{-1} \tilde{\mathbf{y}}.$$

这只是把 y_n 平移一个依赖于 $\tilde{\mathbf{y}}$ 的量；固定 $\tilde{\mathbf{y}}$ 时 $dz = dy_n$ 。代回二次型便完成了“配方”：

$$\mathbf{y}^\top \Sigma^{-1} \mathbf{y} = \tilde{\mathbf{y}}^\top \tilde{\Sigma}^{-1} \tilde{\mathbf{y}} + \frac{z^2}{\tilde{\sigma}^2}.$$

行列式也从同一个三因子分解直接读出。左右两个三角矩阵行列式均为 1，所以

$$\det \Sigma = \det \tilde{\Sigma} \cdot \tilde{\sigma}^2.$$

现在对最后一个坐标积分：

$$\begin{aligned} p_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}}) &= \int_{\mathbb{R}} p_{\mathbf{X}}(\tilde{\mathbf{x}}, x_n) dx_n \\ &= \frac{e^{-\frac{1}{2}\tilde{\mathbf{y}}^T \tilde{\Sigma}^{-1} \tilde{\mathbf{y}}}}{(2\pi)^{n/2} \sqrt{\det \Sigma}} \int_{\mathbb{R}} e^{-z^2/(2\tilde{\sigma}^2)} dz \\ &= \frac{e^{-\frac{1}{2}\tilde{\mathbf{y}}^T \tilde{\Sigma}^{-1} \tilde{\mathbf{y}}}}{(2\pi)^{n/2} \sqrt{\det \tilde{\Sigma} \tilde{\sigma}}} \cdot \sqrt{2\pi} \tilde{\sigma} \\ &= \frac{1}{(2\pi)^{(n-1)/2} \sqrt{\det \tilde{\Sigma}}} \exp\left[-\frac{1}{2}\tilde{\mathbf{y}}^T \tilde{\Sigma}^{-1} \tilde{\mathbf{y}}\right]. \end{aligned}$$

因此删去最后一个分量后仍为 $N_{n-1}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$ 。重复这一消元过程，任意子向量的边际分布仍是相应均值子向量与协方差主子矩阵所确定的多元正态。

3.2.2 独立性

定义：随机变量的独立性

$X_1, \dots, X_n$ 相互独立，当且仅当对任意 $x_1, \dots, x_n$,

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \prod_{j=1}^n F_{X_j}(x_j).$$

离散型等价于联合分布列分解为边际分布列之积；连续型在几乎处处意义下等价于

$$p_{\mathbf{X}}(x_1, \dots, x_n) = \prod_{j=1}^n p_{X_j}(x_j).$$

说明：支持域也是检验线索

即使密度在支持域内能写成 $g(x)h(y)$ ，若支持域不是两个一维集合的笛卡尔积，通常仍不能推出独立。例如三角形均匀分布中，给定较大的 X 会压缩 Y 的可取范围。

对二维正态， $\rho = 0$ 时交叉项消失，密度分解，因此不相关等价于独立。这个结论是正态族的特殊性质，并不适用于一般分布。

3.3 多维随机变量函数的分布

3.3.1 离散映射、最大值与最小值

若 $\mathbf{X}$ 离散而 $Y = g(\mathbf{X})$ ，则所有映到同一 y 的点都要合并：

$$\mathbb{P}(Y = y) = \sum_{\mathbf{x}: g(\mathbf{x})=y} \mathbb{P}(\mathbf{X} = \mathbf{x}).$$

例：独立 Poisson 变量的可加性

若 $X \sim \text{Poisson}(\lambda_1)$ 、 $Y \sim \text{Poisson}(\lambda_2)$ 且独立，则对 $n \geq 0$,

$$\begin{aligned}\mathbb{P}(X + Y = n) &= \sum_{k=0}^n \mathbb{P}(X = k) \mathbb{P}(Y = n - k) \\ &= e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^n \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\ &= \frac{e^{-(\lambda_1 + \lambda_2)}}{n!} \sum_{k=0}^n \binom{n}{k} \lambda_1^k \lambda_2^{n-k} \\ &= \frac{(\lambda_1 + \lambda_2)^n}{n!} e^{-(\lambda_1 + \lambda_2)}.\end{aligned}$$

故 $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$ 。同理，独立二项变量在成功概率相同的条件下参数 n 可相加。

若 $X_1, \dots, X_n$ 独立，令 $M = \max_i X_i$ 、 $L = \min_i X_i$ ，则

$$F_M(y) = \mathbb{P}(X_1 \leq y, \dots, X_n \leq y) = \prod_{i=1}^n F_i(y),$$

$$F_L(y) = 1 - \mathbb{P}(X_1 > y, \dots, X_n > y) = 1 - \prod_{i=1}^n [1 - F_i(y)].$$

把样本由小到大排列所得 $X_{(1)} \leq \dots \leq X_{(n)}$ 称为顺序统计量，其中 $X_{(1)} = L$ 、 $X_{(n)} = M$ 。

3.3.2 卷积公式

令 $Z = X + Y$ 。连续情形先对半平面 $x + y \leq z$ 积分：

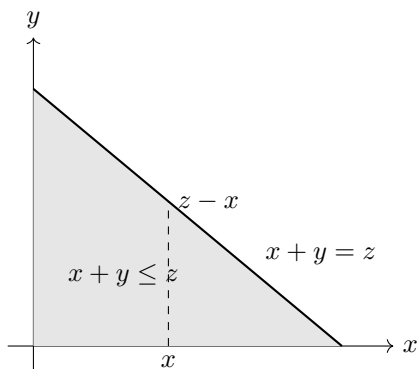
$$F_Z(z) = \iint_{x+y \leq z} p_{X,Y}(x,y) \, dx \, dy.$$

对 z 求导得到

$$p_Z(z) = \int_{\mathbb{R}} p_{X,Y}(x, z-x) \, dx.$$

若 X, Y 独立，

$$p_Z(z) = \int_{\mathbb{R}} p_X(x) p_Y(z-x) \, dx = (p_X * p_Y)(z).$$



图：卷积积分来自直线 $x + y = z$ 上的切片

推导：独立正态变量之和

设 $X \sim N(\mu_1, \sigma_1^2)$ 、 $Y \sim N(\mu_2, \sigma_2^2)$ 独立。卷积指数为

$$\frac{(x - \mu_1)^2}{\sigma_1^2} + \frac{(z - x - \mu_2)^2}{\sigma_2^2} = Ax^2 - 2Bx + C,$$

其中

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

配方 $Ax^2 - 2Bx + C = A(x - B/A)^2 + C - B^2/A$ 后，关于 x 的高斯积分给出 $\sqrt{2\pi/A}$ 。整理剩余指数与常数得

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

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

例：Gamma 与卡方分布的可加性

设 $X \sim \text{Ga}(\alpha_1, \lambda)$ 、 $Y \sim \text{Ga}(\alpha_2, \lambda)$ 独立，采用率参数化。对 $z > 0$,

$$\begin{aligned} p_{X+Y}(z) &= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z x^{\alpha_1-1} (z-x)^{\alpha_2-1} dx \\ &= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z^{\alpha_1+\alpha_2-1} \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1} dt \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1 + \alpha_2)} z^{\alpha_1+\alpha_2-1} e^{-\lambda z}. \end{aligned}$$

故形状参数相加。由于 $\chi_\nu^2 = \text{Ga}(\nu/2, 1/2)$ ，独立卡方变量的自由度也相加。特别地，独立标准正态 $Z_1, \dots, Z_n$ 满足 $\sum_i Z_i^2 \sim \chi_n^2$ 。

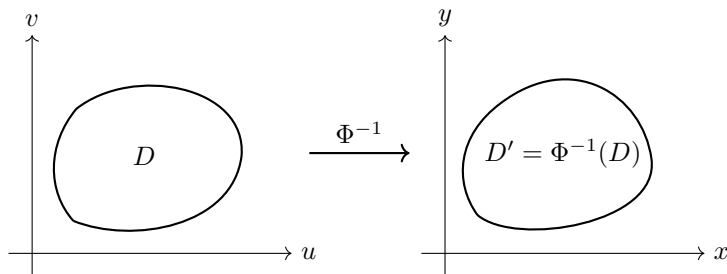
3.3.3 变量变换法

设

$$(U, V) = \Phi(X, Y) = (g_1(X, Y), g_2(X, Y)),$$

且在所考虑区域内 Φ 一一、连续可微，逆映射为 $(x, y) = (x(u, v), y(u, v))$ 。由面积元变换，

$$p_{U,V}(u, v) = p_{X,Y}(x(u, v), y(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|.$$



图：二维变量变换下区域与面积元的对应

推导：由体积分换元公式推出 Jacobian 因子

记逆映射为

$$\Psi = \Phi^{-1}, \quad (x, y) = \Psi(u, v) = (x(u, v), y(u, v)).$$

在 (u, v) 平面任取一个可测区域 D ，并令它在逆映射下对应的区域为

$$D' = \Psi(D) = \Phi^{-1}(D).$$

事件“ (U, V) 落入 D ”与事件“ (X, Y) 落入 D' ”完全相同，因此

$$\mathbb{P}((U, V) \in D) = \mathbb{P}((X, Y) \in D').$$

分别用两组随机变量的联合密度表示这个概率，得到

$$\int_D p_{U,V}(u, v) \, du \, dv = \int_{D'} p_{X,Y}(x, y) \, dx \, dy.$$

对右端使用二重积分的换元公式，以 $x = x(u, v)$ 、 $y = y(u, v)$ 代换：

$$\int_{D'} p_{X,Y}(x, y) \, dx \, dy = \int_D p_{X,Y}(x(u, v), y(u, v)) \left| \det \frac{\partial(x, y)}{\partial(u, v)} \right| \, du \, dv.$$

于是对任意区域 D 都有

$$\int_D p_{U,V}(u, v) \, du \, dv = \int_D p_{X,Y}(x(u, v), y(u, v)) \left| \det \frac{\partial(x, y)}{\partial(u, v)} \right| \, du \, dv.$$

由于该等式对任意 D 成立，两边的被积函数必几乎处处相等，故

$$p_{U,V}(u, v) = p_{X,Y}(x(u, v), y(u, v)) \left| \det \frac{\partial(x, y)}{\partial(u, v)} \right|.$$

同一证明在 n 维情形完全相同：若 $\mathbf{Y} = \Phi(\mathbf{X})$ 、 $\mathbf{X} = \Psi(\mathbf{Y})$ ，则

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\Psi(\mathbf{y})) |\det D\Psi(\mathbf{y})| \quad \text{a.e.}$$

例：和与差

令 $U = X + Y$ ， $V = X - Y$ ，则

$$x = \frac{u+v}{2}, \quad y = \frac{u-v}{2}, \quad \left| \det \frac{\partial(x, y)}{\partial(u, v)} \right| = \frac{1}{2}.$$

若 $X, Y \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ ，代入联合密度可得

$$(U, V) \sim N_2 \left(\begin{pmatrix} 2\mu \\ 0 \end{pmatrix}, \begin{pmatrix} 2\sigma^2 & 0 \\ 0 & 2\sigma^2 \end{pmatrix} \right).$$

二维正态中零协方差推出独立，故 U 与 V 独立。

例：乘积的分布

令 $U = XY$ ， $V = Y$ ，则 $x = u/v$ ， $y = v$ ，Jacobian 的绝对值为 $1/|v|$ 。若 X, Y 独立，

$$p_{U,V}(u, v) = p_X(u/v) p_Y(v) \frac{1}{|v|}, \quad p_U(u) = \int_{\mathbb{R}} p_X(u/v) p_Y(v) \frac{dv}{|v|}.$$

结论：正态向量的线性变换

若 $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, A 为可逆矩阵, 且 $\mathbf{Y} = A^{-1}\mathbf{X}$, 则变量变换给出

$$\mathbf{Y} \sim N_n(A^{-1}\boldsymbol{\mu}, A^{-1}\boldsymbol{\Sigma}A^{-\top}).$$

更一般地, 即使 B 非方阵, 仍有 $B\mathbf{X} + \mathbf{b} \sim N(B\boldsymbol{\mu} + \mathbf{b}, B\boldsymbol{\Sigma}B^\top)$ 。

推导：可逆线性变换下正态密度的整理

以 $\mathbf{X} = A\mathbf{Y}$ 为例。变量变换公式先给出

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(A\mathbf{y})|\det A|.$$

令

$$\tilde{\boldsymbol{\mu}} = A^{-1}\boldsymbol{\mu}, \quad \tilde{\boldsymbol{\Sigma}} = A^{-1}\boldsymbol{\Sigma}A^{-\top}.$$

则

$$A\mathbf{y} - \boldsymbol{\mu} = A(\mathbf{y} - \tilde{\boldsymbol{\mu}}), \quad \tilde{\boldsymbol{\Sigma}}^{-1} = A^\top \boldsymbol{\Sigma}^{-1} A,$$

所以指数中的二次型恰为

$$(A\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (A\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{y} - \tilde{\boldsymbol{\mu}})^\top \tilde{\boldsymbol{\Sigma}}^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}).$$

又因

$$\sqrt{\det \tilde{\boldsymbol{\Sigma}}} = \sqrt{\det(A^{-1})^2 \det \boldsymbol{\Sigma}} = \frac{\sqrt{\det \boldsymbol{\Sigma}}}{|\det A|},$$

Jacobian 因子也正好并入新的归一化常数, 故变换后仍为正态分布。

同一参数变化也可直接用期望与协方差核对:

$$\mathbb{E}\mathbf{Y} = A^{-1}\boldsymbol{\mu}, \quad \text{Cov}(\mathbf{Y}) = A^{-1} \text{Cov}(\mathbf{X}) A^{-\top} = A^{-1}\boldsymbol{\Sigma}A^{-\top}.$$

这与上面从密度中识别出的 $\tilde{\boldsymbol{\mu}}$ 、 $\tilde{\boldsymbol{\Sigma}}$ 完全一致。

推导：非方阵线性映射为何仍保持正态

把 $\mathbf{X}$ 写成标准正态的仿射像

$$\mathbf{X} = \boldsymbol{\mu} + C\mathbf{Z}, \quad \mathbf{Z} \sim N_r(\mathbf{0}, I), \quad CC^\top = \boldsymbol{\Sigma}.$$

那么对任意 $m \times n$ 矩阵 B ,

$$B\mathbf{X} + \mathbf{b} = (B\boldsymbol{\mu} + \mathbf{b}) + (BC)\mathbf{Z}.$$

右端仍是标准正态向量的线性像, 所以仍为多元正态; 其参数由矩阵运算直接给出:

$$\mathbb{E}(B\mathbf{X} + \mathbf{b}) = B\boldsymbol{\mu} + \mathbf{b},$$

$$\text{Cov}(B\mathbf{X} + \mathbf{b}) = BC \text{Cov}(\mathbf{Z}) C^\top B^\top = B\boldsymbol{\Sigma}B^\top.$$

这也说明线性像的协方差可以是半正定的: 当 B 降秩时, 分布集中在某个低维仿射子空间上。

3.4 多维随机变量的特征数

3.4.1 函数的期望与运算性质

若 $\mathbf{X}$ 连续、 $Y = g(\mathbf{X})$, 则

$$\mathbb{E}Y = \int_{\mathbb{R}^n} g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x};$$

离散情形改为对所有取值求和。这是“无须先求 Y 的分布”的期望公式。

例：指数样本最大值的期望

设 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, $M_n = \max_i X_i$ 。则对 $y \geq 0$,

$$F_{M_n}(y) = (1 - e^{-\lambda y})^n.$$

利用非负变量的尾积分公式,

$$E_n := \mathbb{E}M_n = \int_0^\infty [1 - F_{M_n}(y)] dy.$$

比较相邻两项:

$$\begin{aligned} E_{n+1} - E_n &= \int_0^\infty (1 - e^{-\lambda y})^n e^{-\lambda y} dy \\ &= \frac{1}{\lambda} \int_0^1 v^n dv = \frac{1}{\lambda(n+1)}. \end{aligned}$$

由 $E_1 = 1/\lambda$ 得

$$\mathbb{E}M_n = \frac{1}{\lambda} \left(1 + \frac{1}{2} + \dots + \frac{1}{n} \right) = \frac{H_n}{\lambda}.$$

期望的线性性不要求独立:

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y, \quad \mathbb{E}(aX) = a\mathbb{E}X.$$

若 X, Y 独立且期望存在, 则 $\mathbb{E}(XY) = \mathbb{E}X \mathbb{E}Y$ 。由此, 若 $X_1, \dots, X_n$ 独立同分布, 均值为 μ 、方差为 σ^2 , 则

$$\mathbb{E}\bar{X}_n = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

3.4.2 协方差、相关系数与协方差矩阵

定义：协方差与相关系数

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(XY) - \mathbb{E}X \mathbb{E}Y.$$

当两者方差均为正时,

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \in [-1, 1].$$

推导：为何相关系数的绝对值不超过 1

令 $X^\circ = X - \mathbb{E}X$ 、 $Y^\circ = Y - \mathbb{E}Y$ 。对任意实数 t ,

$$0 \leq \mathbb{E}(X^\circ - tY^\circ)^2 = \text{Var}(X) - 2t \text{Cov}(X, Y) + t^2 \text{Var}(Y).$$

右边关于 t 的二次式恒非负, 所以判别式不大于零:

$$4 \text{Cov}(X, Y)^2 - 4 \text{Var}(X) \text{Var}(Y) \leq 0.$$

因此

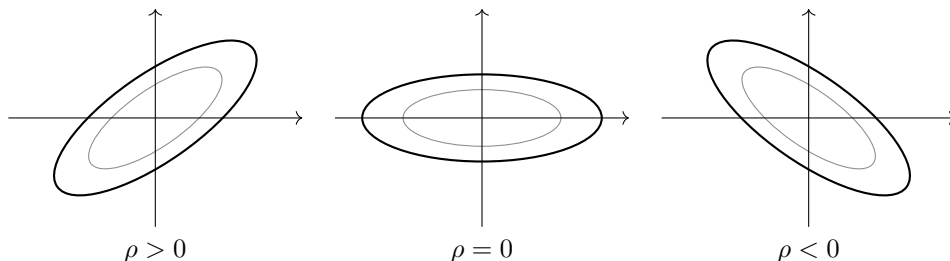
$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \text{Var}(Y)}, \quad |\rho_{X,Y}| \leq 1.$$

等号成立当且仅当存在 t 使 $X^\circ = tY^\circ$ 几乎处处成立，也就是两个变量之间存在几乎处处的线性关系。

协方差对每个变量线性，并且

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

独立蕴含不相关，反向一般不成立。等号 $|\rho| = 1$ 当且仅当 $Y = aX + b$ 几乎处处成立； $\rho = 0$ 仅表示线性关联为零。



图：相关系数的符号与二维等密度椭圆的方向

对随机向量 $\mathbf{X}$ ，记 $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$ ，

$$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] = (\text{Cov}(X_i, X_j))_{i,j}.$$

对任意 $\mathbf{a} \in \mathbb{R}^n$ ，

$$\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} = \text{Var}(\mathbf{a}^\top \mathbf{X}) \geq 0,$$

故协方差矩阵半正定；它正定当且仅当不存在非零 $\mathbf{a}$ 使 $\mathbf{a}^\top \mathbf{X}$ 几乎处处为常数。

例：相关系数的“贡献分解”

若 X, Y 不相关， $Z = aX + bY$ ，则

$$\sigma_Z^2 = a^2 \sigma_X^2 + b^2 \sigma_Y^2, \quad \rho_{X,Z} = \frac{a \sigma_X}{\sigma_Z}, \quad \rho_{Y,Z} = \frac{b \sigma_Y}{\sigma_Z},$$

因而 $\rho_{X,Z}^2 + \rho_{Y,Z}^2 = 1$ 。

例：方差最小的线性组合

有 n 个估计量 X_k ，取 $Y = \mathbf{a}^\top \mathbf{X}$ ，要求 $\mathbf{e}^\top \mathbf{a} = 1$ ，其中 $\mathbf{e} = (1, \dots, 1)^\top$ 。使 $\text{Var}(Y) = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$ 最小。拉格朗日函数

$$L(\mathbf{a}, \lambda) = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} - \lambda(\mathbf{e}^\top \mathbf{a} - 1)$$

给出 $2\boldsymbol{\Sigma} \mathbf{a} - \lambda \mathbf{e} = 0$ ，从而

$$\mathbf{a}_* = \frac{\boldsymbol{\Sigma}^{-1} \mathbf{e}}{\mathbf{e}^\top \boldsymbol{\Sigma}^{-1} \mathbf{e}}, \quad \min \text{Var}(Y) = \frac{1}{\mathbf{e}^\top \boldsymbol{\Sigma}^{-1} \mathbf{e}}.$$

若还要在固定方差与权重和为 1 下最大化期望 $\boldsymbol{\mu}^\top \mathbf{a}$ ，则引入第二个乘子，解形如 $\mathbf{a} = \boldsymbol{\Sigma}^{-1}(\alpha \boldsymbol{\mu} + \beta \mathbf{e})$ ；若另有限制 $a_i \geq 0$ ，则成为带不等式约束的二次规划问题。

3.5 条件分布与条件期望

3.5.1 条件分布

离散情形，若 $\mathbb{P}(Y = y_j) > 0$ ，

$$\mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)} = \frac{p_{ij}}{\sum_i p_{ij}}.$$

连续情形的单点事件 $\{Y = y\}$ 概率为零，不能直接套事件条件概率。由 $y < Y \leq y + h$ 的薄条取极限，得到

$$p_{X|Y}(x | y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}, \quad p_Y(y) > 0.$$

推导：从薄条条件概率得到条件密度

先固定 $h > 0$ ，对 $X \leq x$ 条件化：

$$\mathbb{P}(X \leq x | y < Y \leq y + h) = \frac{\int_{-\infty}^x \int_y^{y+h} p_{X,Y}(u, v) \, dv \, du}{\int_{\mathbb{R}} \int_y^{y+h} p_{X,Y}(u, v) \, dv \, du}.$$

分子、分母同时除以 h 。当 $h \downarrow 0$ ，在密度连续的点上，

$$\frac{1}{h} \int_y^{y+h} p_{X,Y}(u, v) \, dv \longrightarrow p_{X,Y}(u, y),$$

于是条件分布函数成为

$$F_{X|Y=y}(x) = \frac{\int_{-\infty}^x p_{X,Y}(u, y) \, du}{\int_{\mathbb{R}} p_{X,Y}(u, y) \, du} = \frac{\int_{-\infty}^x p_{X,Y}(u, y) \, du}{p_Y(y)}.$$

再对 x 求导，便得到 $p_{X|Y}(x | y) = p_{X,Y}(x, y)/p_Y(y)$ 。

相应地

$$p_{X,Y}(x, y) = p_{X|Y}(x | y)p_Y(y) = p_{Y|X}(y | x)p_X(x),$$

$$p_Y(y) = \int_{\mathbb{R}} p_{Y|X}(y | x)p_X(x) \, dx.$$

因此 Bayes 公式为

$$p_{X|Y}(x | y) = \frac{p_{Y|X}(y | x)p_X(x)}{\int_{\mathbb{R}} p_{Y|X}(y | u)p_X(u) \, du}.$$

例：Poisson 条件分布

若 $X \sim \text{Poisson}(\lambda_1)$ 、 $Y \sim \text{Poisson}(\lambda_2)$ 独立，则对 $k = 0, \dots, n$ ，

$$\begin{aligned} \mathbb{P}(X = k | X + Y = n) &= \frac{\mathbb{P}(X = k, Y = n - k)}{\mathbb{P}(X + Y = n)} \\ &= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-k}. \end{aligned}$$

故 $X | (X + Y = n) \sim \text{Bin}(n, \lambda_1/(\lambda_1 + \lambda_2))$ 。

例：区域均匀分布的切片

若 (X, Y) 在区域 D 上均匀，固定 y 后水平截线在 D 内形成区间 $[A_y, B_y]$ ，长度为 L_y ，则

$$p_{X|Y}(x | y) = \frac{1}{L_y} \mathbf{1}_{\{A_y \leq x \leq B_y\}}.$$

因此条件分布就是该水平截线上的一维均匀分布。类似地，条件 $aX + bY = c$ 可理解为沿直线与支持域交集进行归一化。

例：正态先验的连续更新

设未知量 $X \sim N(\mu_0, \sigma_0^2)$ ，观测 $Y = X + \varepsilon$ ，其中 $\varepsilon \sim N(0, \tau^2)$ 与 X 独立。忽略与 x 无关的因子，Bayes

公式的分子为

$$\begin{aligned} p_{Y|X}(y|x)p_X(x) &\propto \exp\left[-\frac{(y-x)^2}{2\tau^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right] \\ &= \exp\left[-\frac{1}{2}\left\{\left(\frac{1}{\tau^2} + \frac{1}{\sigma_0^2}\right)x^2 - 2\left(\frac{y}{\tau^2} + \frac{\mu_0}{\sigma_0^2}\right)x + C(y)\right\}\right]. \end{aligned}$$

令二次项系数为 $1/\sigma_1^2$ ，一次项系数为 $2\mu_1/\sigma_1^2$ ，再配成 $(x-\mu_1)^2/\sigma_1^2$ ，可得

$$X|Y=y \sim N(\mu_1, \sigma_1^2),$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2}, \quad \frac{\mu_1}{\sigma_1^2} = \frac{\mu_0}{\sigma_0^2} + \frac{y}{\tau^2}.$$

观测 $y_1, \dots, y_n$ 后递推得

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\tau^2}, \quad \frac{\mu_n}{\sigma_n^2} = \frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n y_i}{\tau^2}.$$

当 $n \rightarrow \infty$ ， $\mu_n - \bar{y} \rightarrow 0$ ，数据逐渐压过先验。

3.5.2 条件期望

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum x_i \mathbb{P}(X=x_i|Y=y), & \text{离散型,} \\ \int_{\mathbb{R}} x p_{X|Y}(x|y) dx, & \text{连续型.} \end{cases}$$

把右端看成 y 的函数 $g(y)$ ，则随机变量 $g(Y)$ 记为 $\mathbb{E}(X|Y)$ 。

例：二维正态的条件期望

若 (X, Y) 为二维正态，参数为 $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ ，将联合密度关于 x 配方得

$$\begin{aligned} &\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho \frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X \sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \\ &= \frac{1}{\sigma_X^2} \left[x - \mu_X - \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y) \right]^2 + (1-\rho^2) \frac{(y-\mu_Y)^2}{\sigma_Y^2}. \end{aligned}$$

给定 $Y=y$ 后，最后一项与 x 无关，会被条件密度的归一化常数吸收。而联合密度原指数外还有因子 $1/[2(1-\rho^2)]$ ，故

$$X|Y=y \sim N\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y} (y - \mu_Y), \sigma_X^2 (1-\rho^2)\right).$$

因此

$$\mathbb{E}(X|Y) = \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (Y - \mu_Y),$$

且

$$\text{Var}[\mathbb{E}(X|Y)] = \rho^2 \sigma_X^2.$$

结论：重期望公式

若 $\mathbb{E}|X| < \infty$ ，则

$$\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|Y)].$$

推导：重期望公式

连续情形由联合密度分解直接得到

$$\begin{aligned}\mathbb{E}X &= \iint_{\mathbb{R}^2} x p_{X,Y}(x, y) \, dx \, dy \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} x p_{X|Y}(x | y) \, dx \right] p_Y(y) \, dy \\ &= \int_{\mathbb{R}} \mathbb{E}(X | Y = y) p_Y(y) \, dy = \mathbb{E}[\mathbb{E}(X | Y)].\end{aligned}$$

例：三等奖项的反推

一次抽奖以概率 $p_1, p_2, 1 - p_1 - p_2$ 分别得到 N_1, N_2 和“再来一次”。设总收益为 X 。条件于第一次结果，

$$\mathbb{E}X = p_1 N_1 + p_2 N_2 + (1 - p_1 - p_2) \mathbb{E}X.$$

故

$$\mathbb{E}X = \frac{p_1 N_1 + p_2 N_2}{p_1 + p_2}.$$

例：随机项数之和

设 $X_1, X_2, \dots$ 独立同分布，且与取正整数值的 N 独立， $\mathbb{E}|X_1| < \infty, \mathbb{E}N < \infty$ 。令 $S_N = \sum_{i=1}^N X_i$ 。则

$$\mathbb{E}(S_N | N = m) = m \mathbb{E}X_1,$$

进而由重期望公式

$$\mathbb{E}S_N = \sum_{m \geq 1} m \mathbb{E}X_1 \mathbb{P}(N = m) = \mathbb{E}X_1 \mathbb{E}N.$$

第四章 大数定律与中心极限定理

4.1 ◆ 随机变量序列的收敛性

设 $X, X_1, X_2, \dots$ 定义在同一概率空间上。数列的“趋近”在随机情形下有多种含义；它们控制的对象不同，不能混为一谈。

定义：几种收敛方式

1. 若

$$\mathbb{P}\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\} = 1,$$

则称 X_n 几乎处处收敛于 X ，记 $X_n \xrightarrow{\text{a.s.}} X$ 。

2. 若对任意 $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0,$$

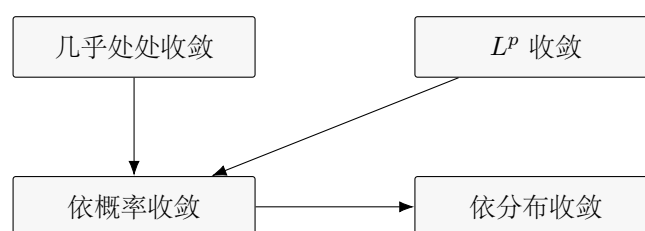
则称 X_n 依概率收敛于 X ，记 $X_n \xrightarrow{\mathbb{P}} X$ 。

3. 若对 X 的分布函数 F 的每个连续点 x ,

$$F_{X_n}(x) \rightarrow F_X(x),$$

则称 X_n 依分布收敛于 X ，记 $X_n \xrightarrow{d} X$ 。

4. 对 $p > 0$ ，若 $\mathbb{E}|X_n - X|^p \rightarrow 0$ ，则称 $X_n \xrightarrow{L^p} X$ 。



图：常用收敛方式的一般蕴含关系

结论：基本蕴含

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X, \quad X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{\mathbb{P}} X, \\ X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X.$$

逆命题一般不成立。但若 $X_n \xrightarrow{d} c$ ，极限为常数，则实际上 $X_n \xrightarrow{\mathbb{P}} c$ 。

说明：大数定律与中心极限定理的分工

大数定律研究样本均值是否靠近其期望，是“位置”或“一致性”问题；中心极限定理研究随机和减去均值并按标准差缩放后的分布形状，是“波动”问题。

4.2 ◆特征函数

定义：特征函数

随机变量 X 的特征函数为

$$\varphi_X(t) = \mathbb{E}(e^{itX}), \quad t \in \mathbb{R}.$$

因为 $|e^{itX}| = 1$ ，特征函数对任何随机变量都存在。

它满足

$$\varphi_X(0) = 1, \quad |\varphi_X(t)| \leq 1, \quad \varphi_X(-t) = \overline{\varphi_X(t)}.$$

若 $\mathbb{E}|X|^k < \infty$ ，则可在 0 附近求导，并有

$$\varphi_X^{(k)}(0) = i^k \mathbb{E}(X^k).$$

若 X, Y 独立，则

$$\varphi_{X+Y}(t) = \mathbb{E}(e^{itX} e^{itY}) = \varphi_X(t) \varphi_Y(t).$$

这把卷积变成乘法，是研究随机和的核心工具。

例：常用特征函数

分布	$\varphi_X(t)$
Bernoulli(p)	$1 - p + pe^{it}$
Poisson(λ)	$\exp\{\lambda(e^{it} - 1)\}$
$N(\mu, \sigma^2)$	$\exp\{i\mu t - \frac{1}{2}\sigma^2 t^2\}$
Exp(λ)	$\frac{\lambda}{\lambda - it}$

例如标准正态的特征函数可由对高斯密度配方，或解微分方程 $\varphi'(t) = -t\varphi(t)$ 、 $\varphi(0) = 1$ 得到 $\varphi(t) = e^{-t^2/2}$ 。

结论：唯一性与连续性定理

特征函数唯一决定分布。并且

$$X_n \xrightarrow{d} X \iff \varphi_{X_n}(t) \rightarrow \varphi_X(t) \quad (\forall t \in \mathbb{R}),$$

其中反向使用极限函数在 0 处连续且确为某个分布的特征函数。

4.3 大数定律

4.3.1 Bernoulli 大数定律

令 $X_k \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ ， $S_n = \sum_{k=1}^n X_k$ 。 S_n/n 是前 n 次试验中成功的频率，

$$\mathbb{E} \frac{S_n}{n} = p, \quad \text{Var} \left(\frac{S_n}{n} \right) = \frac{p(1-p)}{n}.$$

由 Chebyshev 不等式，对任意 $\varepsilon > 0$ ，

$$\mathbb{P} \left(\left| \frac{S_n}{n} - p \right| \geq \varepsilon \right) \leq \frac{p(1-p)}{n\varepsilon^2} \rightarrow 0.$$

因此

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} p.$$

这说明在大量重复试验中，频率依概率稳定于概率。

4.3.2 Chebyshev、Markov 与 Khinchin 大数定律

结论：Chebyshev 大数定律

若 $X_1, X_2, \dots$ 两两不相关，且存在常数 C 使 $\text{Var}(X_k) \leq C$ ，则

$$\frac{1}{n} \sum_{k=1}^n (X_k - \mathbb{E}X_k) \xrightarrow{\mathbb{P}} 0.$$

推导：Chebyshev 大数定律

由两两不相关，

$$\text{Var}\left(\frac{1}{n} \sum_{k=1}^n X_k\right) = \frac{1}{n^2} \sum_{k=1}^n \text{Var}(X_k) \leq \frac{C}{n}.$$

故对任意 $\varepsilon > 0$ ，

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n (X_k - \mathbb{E}X_k)\right| \geq \varepsilon\right) \leq \frac{C}{n\varepsilon^2} \rightarrow 0.$$

结论：◆ Markov 大数定律

若

$$\frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n X_k\right) \rightarrow 0,$$

则

$$\frac{1}{n} \sum_{k=1}^n (X_k - \mathbb{E}X_k) \xrightarrow{\mathbb{P}} 0.$$

Chebyshev 大数定律只是这一条件的一个容易验证的充分条件。

结论：Khinchin 弱大数定律

若 $X_1, X_2, \dots$ 独立同分布，且 $\mathbb{E}|X_1| < \infty$ ，记 $\mu = \mathbb{E}X_1$ ，则

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{\mathbb{P}} \mu.$$

该结论不要求方差存在。

4.3.3 强大数定律

结论：独立同分布情形的强大数定律

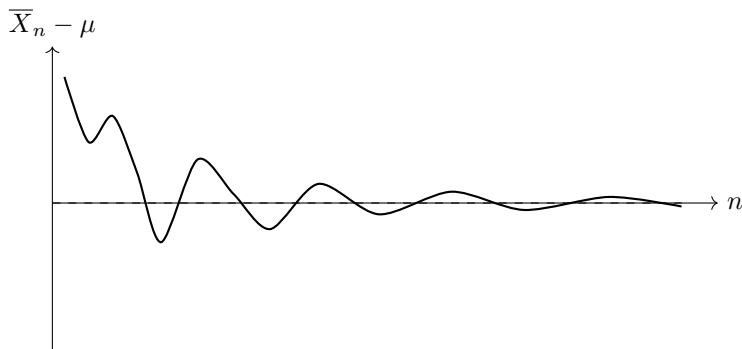
若 $X_1, X_2, \dots$ 独立同分布且 $\mathbb{E}|X_1| < \infty$ ，则

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu = \mathbb{E}X_1,$$

也就是

$$\mathbb{P}\left\{\omega : \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k(\omega) = \mu\right\} = 1.$$

强大数定律描述几乎每一条无限样本轨道；弱大数定律只要求偏离均值的概率趋于零。前者蕴含后者。图中一条典型轨道的波动逐渐缩小，但定理并不声称它单调收敛。



图：样本均值的典型轨道逐渐稳定在总体均值附近

4.4 中心极限定理

4.4.1 De Moivre–Laplace 定理

令 $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, $S_n = \sum_{k=1}^n X_k \sim \text{Bin}(n, p)$ 。标准化为

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

结论：De Moivre–Laplace 中心极限定理

对任意 $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

即 $Z_n \xrightarrow{d} N(0, 1)$ 。

◆ 在用正态分布近似二项概率时，离散端点与连续区间之间可作连续性修正：

$$\mathbb{P}(a \leq S_n \leq b) \approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

4.4.2 Lindeberg–Lévy 定理

结论：独立同分布中心极限定理

设 $X_1, X_2, \dots$ 独立同分布，

$$\mathbb{E}X_k = \mu, \quad \text{Var}(X_k) = \sigma^2 \in (0, \infty).$$

令 $S_n = \sum_{k=1}^n X_k$ ，则

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

等价地，对每个 $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x).$$

推导：◆ 特征函数证明的核心展开

令 $Y_k = (X_k - \mu)/\sigma$ ，则 $\mathbb{E}Y_k = 0$, $\mathbb{E}Y_k^2 = 1$ 。在 $t = 0$ 附近，

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2).$$

由独立性,

$$\varphi_{\frac{1}{\sqrt{n}} \sum_{k=1}^n Y_k}(t) = \left[\varphi_Y\left(\frac{t}{\sqrt{n}}\right) \right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right]^n \rightarrow e^{-t^2/2}.$$

$e^{-t^2/2}$ 是标准正态的特征函数, 故由连续性定理得到所需的依分布收敛。

例：均匀分布之和

若 $X_k \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$, 则 $\mu = 1/2$ 、 $\sigma^2 = 1/12$ 。取 $n = 12$,

$$\frac{\sum_{k=1}^{12} X_k - 6}{\sqrt{12 \cdot (1/12)}} = \sum_{k=1}^{12} X_k - 6$$

已可用 $N(0, 1)$ 作近似。这也是用若干均匀随机数近似生成正态随机数的经典想法。

4.4.3 独立但不同分布的情形

令 $X_1, X_2, \dots$ 独立, $\mathbb{E}X_k = \mu_k$ 、 $\text{Var}(X_k) = \sigma_k^2$, 并记

$$S_n = \sum_{k=1}^n X_k, \quad B_n^2 = \sum_{k=1}^n \sigma_k^2.$$

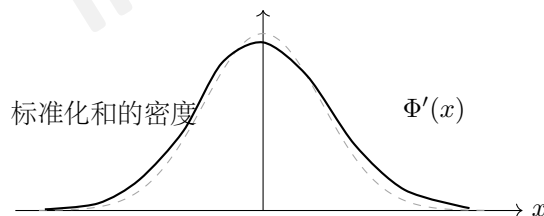
◆ 在适当的 Lindeberg 条件下,

$$\frac{S_n - \sum_{k=1}^n \mu_k}{B_n} \xrightarrow{d} N(0, 1).$$

Lindeberg 条件为对每个 $\varepsilon > 0$,

$$\frac{1}{B_n^2} \sum_{k=1}^n \mathbb{E}[(X_k - \mu_k)^2 \mathbf{1}_{\{|X_k - \mu_k| > \varepsilon B_n\}}] \rightarrow 0.$$

它排除了单个异常大的项支配总波动的情形。



图：中心极限定理：标准化和的分布趋近标准正态

说明：尺度的区别

大数定律说明 $\bar{X}_n - \mu \rightarrow 0$; 中心极限定理进一步指出其典型大小为 $n^{-1/2}$, 并给出

$$\sqrt{n}(\bar{X}_n - \mu) \text{ 的极限分布.}$$

因此前者回答“是否稳定”, 后者回答“以什么速度、按什么形状波动”。

第五章 统计量及其分布

5.1 总体、样本与统计量

数理统计从总体的概率模型出发，通过有限样本反推总体规律。其基本链条是

总体模型 $\rightarrow$ 抽取样本 $\rightarrow$ 构造统计量 $\rightarrow$ 统计推断.

总体可以理解为研究对象的全体，也可以抽象成具有分布函数 F 的随机变量 X 。从总体中独立重复抽取 n 次，得到

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F.$$

其中 $X_1, \dots, X_n$ 是样本随机变量；实验后得到的 $x_1, \dots, x_n$ 是一组样本观测值。若总体有密度或概率质量函数 $f(x; \theta)$ ，则联合分布为

$$f(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta).$$

定义：统计量

若

$$T = T(X_1, \dots, X_n)$$

不含任何未知参数，则称 T 为统计量。统计量本身仍是随机变量，因而有自己的分布。

若用有限维参数 $\theta \in \mathbb{R}^d$ 完全刻画分布族，称为参数模型；若模型同时含有限维参数和无限维干扰成分，称为半参数模型；非参数模型则不预设有有限维参数形式，例如只假定总体分布函数连续。

5.1.1 经验分布函数

给定样本，经验分布函数定义为

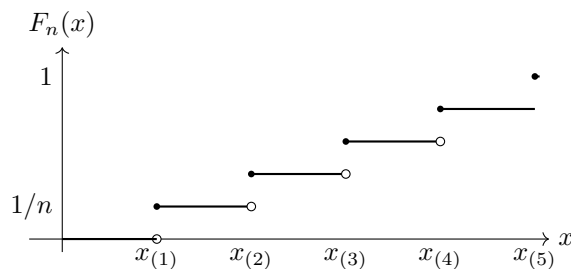
$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}.$$

对固定 x ，令 $Y_i = \mathbf{1}_{\{X_i \leq x\}}$ ，则 $Y_i \sim \text{Bernoulli}(F(x))$ ，所以

$$\mathbb{E}F_n(x) = F(x), \quad \text{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

由大数定律，

$$F_n(x) \xrightarrow{\mathbb{P}} F(x).$$



图：经验分布函数是一条右连续阶梯曲线

◆ 样本数据还可由频数表、直方图与茎叶图整理。它们只是对观测值的不同压缩，后续推断仍需回到相应统计量及其抽样分布。

5.2 样本均值、方差与矩

5.2.1 样本均值

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \mathbb{E}\bar{X} = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

若总体为正态分布，则

$$X_i \sim N(\mu, \sigma^2) \implies \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

可加性也给出许多直接例子：若 $X_i \sim \text{Poisson}(\lambda)$ ，则 $\sum X_i \sim \text{Poisson}(n\lambda)$ ；若 $X_i \sim \text{Bernoulli}(p)$ ，则 $\sum X_i \sim \text{Bin}(n, p)$ 。

5.2.2 样本方差

记

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

恒等式

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$$

给出

$$\mathbb{E} \sum_{i=1}^n (X_i - \bar{X})^2 = n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2.$$

因此

$$\mathbb{E}(S_n^2) = \frac{n-1}{n}\sigma^2, \quad \mathbb{E}(S^2) = \sigma^2.$$

推导：正态样本方差的线性代数结构

先令总体标准化为 $N(0, 1)$ 。写

$$\mathbf{X} = (X_1, \dots, X_n)^\top, \quad \mathbf{e} = (1, \dots, 1)^\top,$$

则

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \mathbf{X}^\top \mathbf{A} \mathbf{X}, \quad \mathbf{A} = \mathbf{I} - \frac{1}{n} \mathbf{e} \mathbf{e}^\top.$$

矩阵 $\mathbf{A}$ 是到 $\mathbf{e}^\perp$ 的正交投影，因为

$$\mathbf{A}^\top = \mathbf{A}, \quad \mathbf{A}^2 = \mathbf{I} - \frac{2}{n} \mathbf{e} \mathbf{e}^\top + \frac{1}{n^2} \mathbf{e} (\mathbf{e}^\top \mathbf{e}) \mathbf{e}^\top = \mathbf{A}.$$

若 $\mathbf{v} \perp \mathbf{e}$ ，则 $\mathbf{A}\mathbf{v} = \mathbf{v}$ ；而 $\mathbf{A}\mathbf{e} = \mathbf{0}$ 。所以 $\mathbf{A}$ 的特征值为

$$\underbrace{1, \dots, 1}_{n-1 \text{ 个}}, \quad 0,$$

亦可写出特征多项式

$$\det(\lambda \mathbf{I} - \mathbf{A}) = \lambda(\lambda - 1)^{n-1}.$$

由实对称矩阵的正交对角化, 存在正交矩阵 Q 使

$$Q^T A Q = \text{diag}(0, 1, \dots, 1).$$

取 Q 的第一列为 $e/\sqrt{n}$, 令 $Y = Q^T X$. 标准正态向量在正交变换下分布不变, 故 $Y_1, \dots, Y_n$ 仍相互独立且服从 $N(0, 1)$; 同时

$$Y_1 = \frac{X_1 + \dots + X_n}{\sqrt{n}} = \sqrt{n} \bar{X},$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = Y_2^2 + \dots + Y_n^2 \sim \chi^2(n-1).$$

第一坐标只描述均值方向, 其余坐标只描述残差方向; 两组坐标独立, 故 $\bar{X}$ 与 S^2 独立。

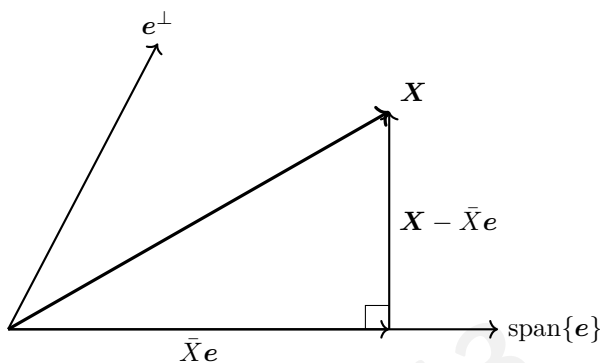


图: 样本向量分解为均值方向与残差子空间中的正交分量

对一般 $N(\mu, \sigma^2)$ 总体, 将 $Z_i = (X_i - \mu)/\sigma$ 代入即得

$$\bar{X} \perp S^2, \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

5.2.3 样本矩

样本 k 阶原点矩与中心矩分别为

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k.$$

特别地 $A_1 = \bar{X}$ 、 $B_1 = 0$ 、 $B_2 = S_n^2$ 。

5.3 次序统计量、分位数与箱线图

把样本从小到大排列:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

$X_{(k)}$ 称为第 k 个次序统计量, $X_{(1)}$ 、 $X_{(n)}$ 分别为样本最小值与最大值。

推导: 第 k 个次序统计量的分布

事件 $X_{(k)} \leq x$ 等价于至少有 k 个样本不超过 x 。因此

$$F_{X_{(k)}}(x) = \sum_{j=k}^n \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}.$$

若总体密度为 f ，求导并利用相邻项消去，得到

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} [1-F(x)]^{n-k} f(x).$$

直观上：选出一个观测落在 $[x, x+dx]$ ，另有 $k-1$ 个在左侧、 $n-k$ 个在右侧，再计数所有排列方式。

若总体为 $U(0,1)$ ，则

$$X_{(k)} \sim \text{Beta}(k, n-k+1).$$

进一步，对 $1 \leq i < j \leq n$ 及 $x < y$ ，

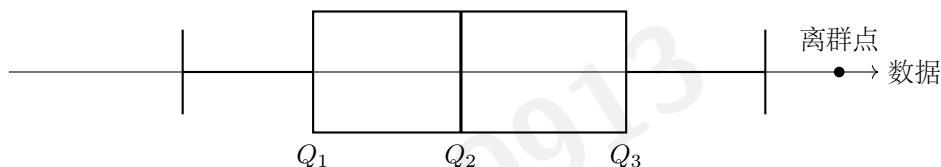
$$f_{X_{(i)}, X_{(j)}}(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(x)^{i-1} [F(y)-F(x)]^{j-i-1} [1-F(y)]^{n-j} f(x)f(y).$$

这是把样本分到 $(-\infty, x)$ 、 (x, y) 、 (y, ∞) 三段后计数得到的。

极差为

$$R = X_{(n)} - X_{(1)}.$$

样本中位数在 n 为奇数时取 $X_{((n+1)/2)}$ ，在 n 为偶数时通常取 $\frac{1}{2}(X_{(n/2)} + X_{(n/2+1)})$ 。样本 p 分位数则由靠近位置 np 的次序统计量给出；具体插值约定可因软件而异。



图：箱线图：四分位数、须与离群点

箱线图以 Q_1, Q_2, Q_3 为箱体的左端、中线、右端， $IQR = Q_3 - Q_1$ 。常用规则把 $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ 外的点单独标作离群点。把多组箱线图并排，可直接比较位置、离散程度、偏态与异常值。

5.4 三大抽样分布

5.4.1 χ^2 分布

若 $Z_1, \dots, Z_m \stackrel{\text{i.i.d.}}{\sim} N(0,1)$ ，则

$$V = Z_1^2 + \dots + Z_m^2 \sim \chi^2(m),$$

其密度为

$$f_V(v) = \frac{1}{2^{m/2}\Gamma(m/2)} v^{m/2-1} e^{-v/2}, \quad v > 0.$$

5.4.2 t 分布

若 $Z \sim N(0,1)$ 、 $V \sim \chi^2(n)$ 且二者独立，则

$$T = \frac{Z}{\sqrt{V/n}} \sim t(n).$$

推导: t 密度

令 $Y = T = Z/\sqrt{V/n}$ 、 $W = V$, 则 $Z = Y\sqrt{W/n}$, Jacobian 为 $\sqrt{W/n}$ 。于是

$$\begin{aligned} f_Y(y) &= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2 w/(2n)} \frac{w^{n/2-1} e^{-w/2}}{2^{n/2} \Gamma(n/2)} \sqrt{\frac{w}{n}} dw \\ &= \frac{1}{\sqrt{n\pi} 2^{(n+1)/2} \Gamma(n/2)} \int_0^\infty w^{(n+1)/2-1} \exp\left[-\frac{1}{2}\left(1 + \frac{y^2}{n}\right)w\right] dw. \end{aligned}$$

令 $z = \frac{1}{2}(1 + y^2/n)w$, 使用 Gamma 积分,

$$f_Y(y) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \Gamma(n/2)} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2}.$$

当 $n \rightarrow \infty$ 时, Gamma 函数比值与幂函数分别趋向相应正态常数和 $e^{-y^2/2}$, 故 $t(n)$ 渐近于 $N(0, 1)$ 。

由正态样本的独立性结论,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

5.4.3 F 分布

若 $V_1 \sim \chi^2(m)$ 、 $V_2 \sim \chi^2(n)$ 相互独立, 则

$$F = \frac{V_1/m}{V_2/n} \sim F(m, n).$$

推导: F 密度

取 $Y_1 = (V_1/m)/(V_2/n)$ 、 $Y_2 = V_2$, 则 $V_1 = (m/n)Y_1Y_2$, Jacobian 为 $(m/n)Y_2$ 。代入两个卡方密度并对 Y_2 积分:

$$\begin{aligned} f_{Y_1}(y) &= \frac{(m/n)^{m/2}}{2^{(m+n)/2} \Gamma(m/2) \Gamma(n/2)} y^{m/2-1} \\ &\cdot \int_0^\infty u^{(m+n)/2-1} \exp\left[-\frac{1}{2}\left(1 + \frac{m}{n}y\right)u\right] du \\ &= \frac{\Gamma((m+n)/2)}{\Gamma(m/2) \Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} y^{m/2-1} \left(1 + \frac{m}{n}y\right)^{-(m+n)/2}, \quad y > 0. \end{aligned}$$

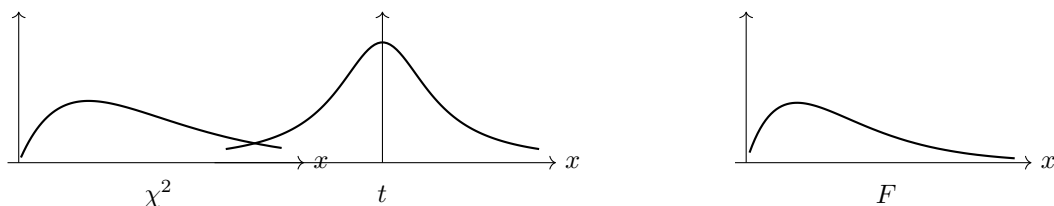


图: 三大抽样分布的典型密度形状

若有两组独立正态样本

$$X_1, \dots, X_m \sim N(\mu_1, \sigma_1^2), \quad Y_1, \dots, Y_n \sim N(\mu_2, \sigma_2^2),$$

则

$$\frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(m-1, n-1).$$

特别地, 当 $\sigma_1 = \sigma_2$ 时, $S_X^2/S_Y^2 \sim F(m-1, n-1)$ 。若还假设共同方差, 可取合并方差

$$S_p^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{m+n-2},$$

并有

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{1/m + 1/n}} \sim t(m + n - 2).$$

5.5 稳健统计量

Fisher 型离散程度用平方偏差

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

计算方便、适合实验设计，但少量远离中心的数据会因平方而产生很大影响。Eddington 型统计量改用绝对偏差

$$\frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|,$$

对离群点更稳健。若 $X_i \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ ，则

$$X_i - \bar{X} \sim N\left(0, \frac{n-1}{n} \sigma^2\right).$$

由 $Y \sim N(0, \tau^2)$ 时

$$\mathbb{E}|Y| = \frac{2}{\sqrt{2\pi}\tau} \int_0^\infty y e^{-y^2/(2\tau^2)} dy = \sqrt{\frac{2}{\pi}} \tau,$$

得到

$$\mathbb{E} \sum_{i=1}^n |X_i - \bar{X}| = \sqrt{\frac{2}{\pi}} \sqrt{n(n-1)} \sigma.$$

故可构造

$$\hat{\sigma}_{\text{Edd}} = \sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{n(n-1)}} \sum_{i=1}^n |X_i - \bar{X}|, \quad \mathbb{E} \hat{\sigma}_{\text{Edd}} = \sigma.$$

离群点可能来自测量误差，也可能来自模型中尚未认识的影响因素。稳健统计的目标，是使少量异常观测不至于支配整个估计结果。

5.6 充分统计量

定义：充分性

设总体分布为 $F(x; \theta)$ ，统计量 $T = T(X_1, \dots, X_n)$ 。若给定 $T = t$ 后样本的条件分布不依赖于 θ ，则称 T 是关于 θ 的充分统计量。

例：二项模型中的充分性

设 $X_i \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ ， $T = \sum X_i$ 。给定 $T = t$ ，任何恰含 t 个 1 的样本序列都有

$$\mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n | T = t) = \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}},$$

其中不含 θ ，故 T 充分。

例：正态总体

对 $N(\mu, \sigma^2)$ 总体， $T = (\bar{X}, S^2)$ 关于 (μ, σ^2) 充分。几何上，固定 $\bar{X} = \bar{x}$ 与 $S^2 = s^2$ 后，样本向量位于均值超平面内半径为 $\sqrt{(n-1)s^2}$ 的球面上。以正交坐标

$$y_1 = \sqrt{n} \bar{x}, \quad y_2^2 + \dots + y_n^2 = (n-1)s^2$$

表示, 再令残差坐标写成球坐标

$$(y_2, \dots, y_n) = r \mathbf{u}(\theta_1, \dots, \theta_{n-2}),$$

Jacobian 分解为 $r^{n-2}J(\boldsymbol{\theta})$ 。联合密度中的参数只出现在

$$\exp \left[-\frac{(\sqrt{n}\bar{x} - \sqrt{n}\mu)^2}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^2} \right],$$

而给定 $(\bar{X}, S^2)$ 后, 角度变量的条件密度只与 $J(\boldsymbol{\theta})$ 有关, 更明确地说, 它可写为

$$f(\boldsymbol{\theta} | \bar{X} = \bar{x}, S^2 = s^2) = \frac{J(\boldsymbol{\theta})}{\int J(\mathbf{u}) d\mathbf{u}},$$

不再含 μ, σ^2 。

结论：因子分解定理

设样本联合密度或概率质量函数为 $f(\mathbf{x}; \theta)$ 。统计量 T 关于 θ 充分, 当且仅当存在函数 g, h , 使

$$f(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta) h(\mathbf{x}),$$

其中 g 可含 θ , h 不含 θ 。

例：均匀总体

若 $X_i \stackrel{\text{i.i.d.}}{\sim} U(0, \theta)$, 则

$$f(\mathbf{x}; \theta) = \theta^{-n} \mathbf{1}_{\{X_{(n)} \leq \theta\}} \mathbf{1}_{\{X_{(1)} \geq 0\}}.$$

按因子分解定理, $X_{(n)}$ 是关于 θ 的充分统计量。

若 T 充分而 $S = \varphi(T)$ 保留了确定 T 所需的信息, 例如 $T = \psi(S)$, 则仍可由因子分解直接验证 S 充分。

◆ 充分统计量只回答“是否丢失关于参数的信息”, 并不自动保证维数最小。最小充分统计量、完备性以及 Rao-Blackwell 改进是进一步研究估计量结构的工具。

第六章 参数估计

6.1 估计问题与点估计

统计模型写作

$$\{F(x; \theta) : \theta \in \Theta\},$$

其中 Θ 为参数空间。估计的对象可以是参数本身、参数的函数 $\psi(\theta)$ ，也可以是与分布有关的其他数字特征。

定义：估计量与估计值

由样本构造的统计量

$$\hat{\psi} = g(X_1, \dots, X_n)$$

称为 $\psi(\theta)$ 的估计量；代入一组观测值后所得数值称为估计值。

点估计用一个数 $\hat{\theta}$ 估计 θ ；区间估计则构造随机区间 $[\hat{\theta}_L, \hat{\theta}_U]$ ，使它以指定概率覆盖未知参数。

6.2 估计量的评价

6.2.1 无偏性与渐近无偏性

定义：无偏估计

若对所有 $\theta \in \Theta$ 都有

$$\mathbb{E}_{\theta}(\hat{\psi}) = \psi(\theta),$$

则称 $\hat{\psi}$ 是 $\psi(\theta)$ 的无偏估计。若仅有

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\theta}(\hat{\psi}_n) = \psi(\theta),$$

则称其渐近无偏。

对 $N(\mu, \sigma^2)$ 总体，

$\bar{X}$ 是 μ 的无偏估计， $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的无偏估计。

相反， $S_n^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$ 有偏，因为 $\mathbb{E}S_n^2 = (n-1)\sigma^2/n$ 。

推导：正态总体中 σ 的无偏估计

自然的 $\sqrt{S^2}$ 并不无偏。由

$$Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

及卡方密度,

$$\begin{aligned}\mathbb{E}\sqrt{Y} &= \int_0^\infty y^{1/2} \frac{y^{(n-1)/2-1} e^{-y/2}}{2^{(n-1)/2} \Gamma((n-1)/2)} dy \\ &= \sqrt{2} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}.\end{aligned}$$

又 $\sqrt{S^2} = \sigma \sqrt{Y/(n-1)}$, 故

$$\mathbb{E}\sqrt{S^2} = \sqrt{\frac{2}{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \sigma.$$

于是

$$\hat{\sigma} = \sqrt{\frac{n-1}{2}} \frac{\Gamma((n-1)/2)}{\Gamma(n/2)} \sqrt{S^2}$$

是 σ 的无偏估计。

无偏估计未必总存在。若要估计 $N(\mu, \sigma^2)$ 中的 μ^2 ,

$$\mathbb{E}(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n},$$

所以在 σ^2 未知时可用

$$\widehat{\mu^2} = \bar{X}^2 - \frac{1}{n} S^2$$

作无偏估计。它可能取负值; 这不与无偏性矛盾, 因为无偏性只约束期望。把它截断为 $\max(\widehat{\mu^2}, 0)$ 后不再保持无偏。

例: 不存在无偏估计的情形

若 $X_i \sim \text{Bernoulli}(p)$, 任意统计量 $g(X_1, \dots, X_n)$ 的期望都是

$$\sum_{x_1=0}^1 \cdots \sum_{x_n=0}^1 g(x_1, \dots, x_n) p^{x_1+\cdots+x_n} (1-p)^{n-(x_1+\cdots+x_n)},$$

因此是 p 的次数不超过 n 的多项式。若 $\psi(p)$ 不是这种多项式, 例如 $\psi(p) = 1/p$, 就不存在基于固定样本量 n 的无偏估计。

6.2.2 有效性

定义: 有效性比较

设 $\hat{\psi}_1, \hat{\psi}_2$ 都是 $\psi(\theta)$ 的无偏估计。若对所有 θ ,

$$\text{Var}_\theta(\hat{\psi}_1) \leq \text{Var}_\theta(\hat{\psi}_2),$$

并至少在一个 θ_0 处严格小于, 则称 $\hat{\psi}_1$ 比 $\hat{\psi}_2$ 有效。

例如对 $N(\mu, \sigma^2)$ 总体,

$$\hat{\mu}_1 = \bar{X}, \quad \hat{\mu}_2 = X_1$$

均无偏, 但

$$\text{Var}(\hat{\mu}_1) = \frac{\sigma^2}{n} < \sigma^2 = \text{Var}(\hat{\mu}_2) \quad (n > 1).$$

例：\$U(0, \theta)\$ 中两个无偏估计的比较

记 \$X_{(n)} = \max_i X_i\$。因为

$$f_{X_{(n)}}(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 < y < \theta,$$

所以

$$\mathbb{E}X_{(n)} = \frac{n}{n+1}\theta, \quad \mathbb{E}X_{(n)}^2 = \frac{n}{n+2}\theta^2,$$

$$\text{Var}(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

两个无偏估计为

$$\hat{\theta}_1 = \frac{n+1}{n}X_{(n)}, \quad \hat{\theta}_2 = 2\bar{X}.$$

它们的方差分别为

$$\text{Var}(\hat{\theta}_1) = \frac{\theta^2}{n(n+2)}, \quad \text{Var}(\hat{\theta}_2) = 4\frac{\theta^2/12}{n} = \frac{\theta^2}{3n}.$$

当 \$n > 1\$ 时 \$\hat{\theta}_1\$ 更有效。

6.2.3 均方误差**定义：均方误差**

$$\text{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2.$$

加上并减去 \$\mathbb{E}\hat{\theta}\$，交叉项期望为零，得到偏差-方差分解

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\mathbb{E}\hat{\theta} - \theta)^2.$$

所以有偏估计也可能以较小方差换取更小 MSE。

例：有偏缩减可降低 MSE

仍在 \$U(0, \theta)\$ 中，考虑 \$\tilde{\theta} = dX_{(n)}\$。由上面的矩，

$$\frac{\text{MSE}(\tilde{\theta})}{\theta^2} = d^2 \frac{n}{(n+1)^2(n+2)} + \left(\frac{nd}{n+1} - 1 \right)^2.$$

对 \$d\$ 求导可得最优

$$d = \frac{n+2}{n+1}, \quad \text{MSE}(\tilde{\theta}) = \frac{\theta^2}{(n+1)^2}.$$

这小于无偏估计 \$\frac{n+1}{n}X_{(n)}\$ 的 \$\theta^2/[n(n+2)]\$。放弃一点无偏性，反而降低了总体误差。

6.2.4 相合性**定义：相合估计**

若对任意 \$\varepsilon > 0\$，

$$\mathbb{P}_\theta(|\hat{\theta}_n - \theta| \geq \varepsilon) \longrightarrow 0,$$

则称 \$\hat{\theta}_n\$ 是 \$\theta\$ 的相合估计。

结论：一个常用充分条件

若

$$\mathbb{E}_\theta \hat{\theta}_n \rightarrow \theta, \quad \text{Var}_\theta(\hat{\theta}_n) \rightarrow 0,$$

则由 Chebyshev 不等式, $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ 。

若 $\hat{\theta}_{1n}, \dots, \hat{\theta}_{kn}$ 分别相合估计 $\theta_1, \dots, \theta_k$, 且 g 连续, 则

$$g(\hat{\theta}_{1n}, \dots, \hat{\theta}_{kn}) \xrightarrow{\mathbb{P}} g(\theta_1, \dots, \theta_k).$$

例如 $\bar{X}$ 与 S^2 分别相合估计 μ 与 σ^2 ; $U(0, \theta)$ 中 $X_{(n)}$ 也相合, 因为其期望趋于 θ 、方差趋于零。若 $\bar{X}$ 相合估计 μ , 则 $\bar{X}^2$ 相合估计 μ^2 , 尽管 $\bar{X}^2$ 对有限 n 一般有偏。

6.3 矩估计

6.3.1 基本方法

总体前 k 阶矩

$$\mu_j(\theta_1, \dots, \theta_k) = \mathbb{E}_\theta(X^j)$$

是参数的函数。矩估计用样本矩

$$A_j = \frac{1}{n} \sum_{i=1}^n X_i^j$$

替换 μ_j , 解方程

$$A_j = \mu_j(\theta_1, \dots, \theta_k), \quad j = 1, \dots, k.$$

若 $\psi = g(\theta_1, \dots, \theta_k)$, 再用不变代入

$$\hat{\psi} = g(\hat{\theta}_1, \dots, \hat{\theta}_k).$$

例：指数分布

若 $X \sim \text{Exp}(\lambda)$, $\mathbb{E}X = 1/\lambda$, 则

$$\hat{\lambda}_1 = \frac{1}{\bar{X}}.$$

也可利用 $\text{Var}(X) = 1/\lambda^2$ 得到

$$\hat{\lambda}_2 = \frac{1}{\sqrt{S_n^2}}.$$

不同矩方程可能产生不同估计量, 需要再用方差、偏差或 MSE 比较。

例：均匀分布 $U(a, b)$

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

以 $\bar{X}$ 和 S_n^2 代替, 解得

$$\hat{a} = \bar{X} - \sqrt{3S_n^2}, \quad \hat{b} = \bar{X} + \sqrt{3S_n^2}.$$

6.3.2 广义矩关系

矩估计并不限于普通幂函数。若密度属于

$$f_\theta(x) = e^{-V(x)}$$

一类, 其中 V 为偶次多项式且最高次系数为正, 则由边界项消失,

$$0 = \int_{\mathbb{R}} \frac{d}{dx} (x^n e^{-V(x)}) dx = \int_{\mathbb{R}} [nx^{n-1} - x^n V'(x)] e^{-V(x)} dx.$$

这给出总体矩之间的递推关系。若 $V(x) = a_0 + a_1x + \dots + a_kx^k$, 可利用若干递推式和样本矩解出 $a_1, \dots, a_k$; 归一化条件再确定 a_0 。

6.4 最大似然估计

6.4.1 似然函数

对观测 $\mathbf{x} = (x_1, \dots, x_n)$, 似然函数把样本固定、把参数视为变量:

$$L(\theta; \mathbf{x}) = \begin{cases} \prod_{i=1}^n f(x_i; \theta), & \text{连续总体,} \\ \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i), & \text{离散总体.} \end{cases}$$

使 L 最大的参数值称为最大似然估计。通常改求 $\ell(\theta) = \log L(\theta)$ 的最大值; 对数是单调函数, 不改变极大点, 却把乘积化成求和。

例：最大似然的直观

有两个盒子：甲盒中红球比例为 0.99，乙盒中红球比例为 0.01。随机选定一个盒子后有放回地抽十次，结果十次全为红球。若模型只允许 $\theta \in \{\text{甲}, \text{乙}\}$ ，则

$$L(\text{甲}) = 0.99^{10}, \quad L(\text{乙}) = 0.01^{10}.$$

观测数据在甲模型下远更可能，因此最大似然原则选择甲盒。

例：Bernoulli 模型

若 $X_i \sim \text{Bernoulli}(p)$ ，则

$$L(p) = p^{\sum x_i} (1-p)^{n-\sum x_i} = p^{n\bar{x}} (1-p)^{n(1-\bar{x})}.$$

$$\ell'(p) = \frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0 \implies \hat{p} = \bar{x}.$$

而

$$\ell''(\hat{p}) = -\frac{n}{\bar{x}(1-\bar{x})} < 0,$$

所以确为极大点。

例：正态总体

令 $t = \sigma^2$ 。对 $N(\mu, t)$,

$$\ell(\mu, t) = -\frac{1}{2t} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \log t - \frac{n}{2} \log(2\pi).$$

得分方程为

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{t} \sum_{i=1}^n (x_i - \mu) = 0,$$

$$\frac{\partial \ell}{\partial t} = \frac{1}{2t^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2t} = 0.$$

故

$$\hat{\mu}_{\text{MLE}} = \bar{X}, \quad \widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S_n^2.$$

必要时再由 Hessian 或边界行为确认它是极大点。

例：均匀总体的边界极大

若 $X_i \sim U(0, \theta)$ ，则

$$L(\theta) = \theta^{-n} \mathbf{1}_{\{0 \leq X_{(1)} \leq X_{(n)} \leq \theta\}}.$$

在允许区域 $\theta \geq X_{(n)}$ 上， θ^{-n} 单调下降，所以

$$\hat{\theta}_{\text{MLE}} = X_{(n)}.$$

这里似然因指示函数而不光滑，不能机械地令导数为零。

结论：最大似然估计的不变性

若 $\hat{\theta}$ 是 θ 的 MLE，则在通常的良好定义条件下， $g(\hat{\theta})$ 是 $g(\theta)$ 的 MLE。例如正态总体中

$$\hat{\mu}_{\text{MLE}}^2 = \bar{X}^2, \quad \hat{\sigma}_{\text{MLE}} = \sqrt{S_n^2}.$$

6.4.2 得分、Fisher 信息与渐近正态性

单个观测的得分函数为

$$S_{\theta}(X) = \frac{\partial}{\partial \theta} \log f(X; \theta).$$

在可交换求导与积分且支持集不依赖参数等正则条件下，

$$\mathbb{E}_{\theta} S_{\theta}(X) = \int \frac{\partial \log f(x; \theta)}{\partial \theta} f(x; \theta) dx = \int \frac{\partial f(x; \theta)}{\partial \theta} dx = 0.$$

Fisher 信息为

$$I(\theta) = \mathbb{E}_{\theta} [S_{\theta}(X)^2] = -\mathbb{E}_{\theta} \left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right].$$

对向量参数 $\boldsymbol{\theta}$ ，信息矩阵写作

$$I(\boldsymbol{\theta}) = \mathbb{E} [\nabla_{\boldsymbol{\theta}} \log f(X; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log f(X; \boldsymbol{\theta})^{\top}].$$

结论：MLE 的渐近正态性

在相合性、可微性、可识别性等正则条件下，

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right),$$

即大样本下

$$\hat{\theta}_n \stackrel{\text{approx}}{\sim} N\left(\theta, \frac{1}{nI(\theta)}\right).$$

复杂模型中的似然方程未必有解析解，实际常借助数值优化求解。

◆ 当直接最大化含隐变量或缺失数据的似然较困难时，EM 算法交替进行：E 步计算给定当前参数的完全数据对数似然条件期望，M 步最大化该期望。每次迭代不会降低观测数据似然，但可能收敛到局部极大点。

6.5 ◆ 最小方差无偏估计

在所有无偏估计中方差处处最小者称为一致最小方差无偏估计 (UMVUE)。充分性说明如何压缩样本，完备性则可保证无偏函数的唯一性。若 T 充分，任一无偏估计 U 经

$$\tilde{U} = \mathbb{E}(U | T)$$

改进后仍无偏, 且 $\text{Var}(\tilde{U}) \leq \text{Var}(U)$; 这就是 Rao-Blackwell 思想。对满足正则条件的无偏估计, 还有 Cramér-Rao 下界

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{nI(\theta)}.$$

6.6 Bayes 估计

6.6.1 先验、似然与后验

频率派把 θ 看作固定但未知的数; Bayes 方法把参数空间视为概率空间, 令 θ 具有先验分布 $\pi(\theta)$ 。数据的条件分布仍为似然 $p(x | \theta)$, 联合密度为

$$h(x, \theta) = p(x | \theta)\pi(\theta).$$

边际密度

$$m(x) = \int_{\Theta} h(x, \theta) d\theta$$

是观测数据的无条件分布。Bayes 公式给出后验

$$\pi(\theta | x) = \frac{p(x | \theta)\pi(\theta)}{\int_{\Theta} p(x | u)\pi(u) du}.$$

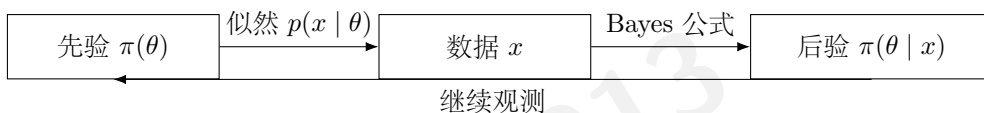


图: Bayes 更新: 先验经由数据变成后验

在平方损失下, Bayes 估计是后验均值

$$\hat{\theta}_B = \mathbb{E}(\theta | x).$$

例: Bernoulli 样本与均匀先验

设 $X_i | \theta \sim \text{Bernoulli}(\theta)$, $\theta \sim U(0, 1)$, 并记 $S = \sum_{i=1}^n X_i$ 。联合密度的参数部分为

$$h(\mathbf{x}, \theta) = \theta^S (1 - \theta)^{n-S}, \quad 0 < \theta < 1.$$

由 Beta 函数

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

得到

$$\begin{aligned} \theta | \mathbf{x} &\sim \text{Beta}(S+1, n-S+1), \\ \pi(\theta | \mathbf{x}) &= (n+1) \binom{n}{S} \theta^S (1-\theta)^{n-S}. \end{aligned}$$

因此平方损失下

$$\hat{\theta}_B = \mathbb{E}(\theta | \mathbf{x}) = \frac{S+1}{n+2}.$$

即使 $S = n$, 估计仍为 $(n+1)/(n+2)$, 不会武断地等于 1。

例: 正态-正态共轭

设

$$X_i | \theta \sim N(\theta, \tau^2), \quad \theta \sim N(\mu_0, \sigma_0^2),$$

其中 $\tau^2, \mu_0, \sigma_0^2$ 已知。把似然和先验中关于 θ 的二次项配方，得到后验仍为正态：

$$\theta | \mathbf{X} \sim N(\mu_n, \sigma_n^2),$$

$$\sigma_n^2 = \frac{1}{\sigma_0^{-2} + n\tau^{-2}}, \quad \mu_n = \frac{\mu_0/\sigma_0^2 + n\bar{X}/\tau^2}{1/\sigma_0^2 + n/\tau^2}.$$

后验均值是先验均值与样本均值按“精度”加权的結果。

6.7 区间估计

6.7.1 置信区间与可信区间

定义：频率派置信区间

若随机区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 满足对所有 $\theta \in \Theta$,

$$\mathbb{P}_\theta\{\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\} \geq 1 - \alpha,$$

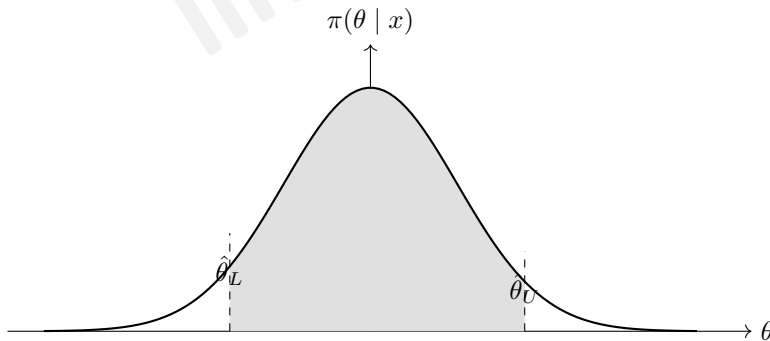
则称它为置信水平 $1 - \alpha$ 的双侧置信区间。

也可构造单侧置信下限或上限。若两端漏掉的概率均为 $\alpha/2$ ，称为等尾置信区间；若在覆盖率约束下区间长度最小，称为最短置信区间。

Bayes 可信区间在已经观测数据后满足

$$\mathbb{P}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U | x) = 1 - \alpha.$$

两者形式相近，概率解释却不同：频率派的参数固定、区间随机；Bayes 方法在后验分布下直接讨论参数的条件概率。



图：后验分布下概率为 $1 - \alpha$ 的可信区间

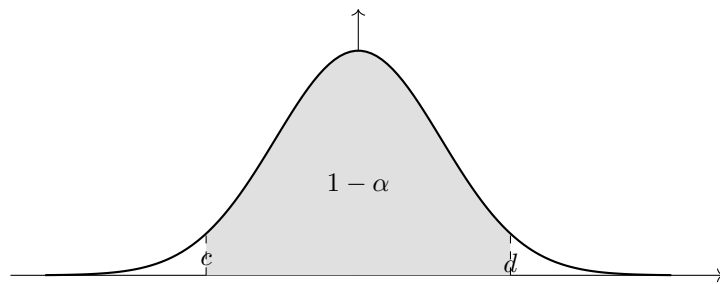
6.7.2 枢轴量法

定义：枢轴量

若 $G = G(\mathbf{X}, \theta)$ 的分布不依赖未知参数，则称 G 为枢轴量。

构造区间的步骤是：

1. 选择含 θ 的枢轴量 $G(\mathbf{X}, \theta)$;
2. 找常数 c, d ，使 $\mathbb{P}(c \leq G \leq d) = 1 - \alpha$;
3. 把不等式反解为 $\hat{\theta}_L(\mathbf{X}) \leq \theta \leq \hat{\theta}_U(\mathbf{X})$ 。



图：枢轴量的中央概率区间

例：\$U(0, \theta)\$ 的区间

最大次序统计量密度为

$$f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n}, \quad 0 < x < \theta.$$

令

$$Y = \frac{X_{(n)}}{\theta},$$

则 \$f_Y(y) = ny^{n-1}\$、\$F_Y(y) = y^n\$，分布与 \$\theta\$ 无关。若 \$0 < c < d < 1\$ 且

$$\mathbb{P}(c \leq Y \leq d) = d^n - c^n = 1 - \alpha,$$

则

$$\frac{X_{(n)}}{d} \leq \theta \leq \frac{X_{(n)}}{c}.$$

区间长度为

$$L = X_{(n)} \left(\frac{1}{c} - \frac{1}{d} \right).$$

在 \$d^n - c^n = 1 - \alpha\$ 下令 \$d = (1 - \alpha + c^n)^{1/n}\$，于是

$$d'(c) = \frac{c^{n-1}}{d^{n-1}}, \quad L'(c) = -\frac{1}{c^2} + \frac{c^{n-1}}{d^{n+1}} = \frac{c^{n+1} - d^{n+1}}{c^2 d^{n+1}} < 0.$$

故长度随 \$c\$ 增大而减小，最短区间在允许的最大 \$c\$，亦即边界 \$d = 1\$ 处取得：

$$\left[X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}} \right].$$

若要求等尾，则

$$c = \left(\frac{\alpha}{2} \right)^{1/n}, \quad d = \left(1 - \frac{\alpha}{2} \right)^{1/n}.$$

6.7.3 正态总体的区间估计

方差已知，估计均值

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

标准正态密度对称，最短的固定概率区间为中央区间，故

$$\Phi(d) - \Phi(c) = 1 - \alpha \implies d'(c) = \frac{\phi(c)}{\phi(d)} = e^{(d^2 - c^2)/2}.$$

若最小化标准化长度 \$d - c\$，驻点条件 \$d'(c) - 1 = 0\$ 给出 \$d^2 = c^2\$；因 \$c < d\$，必须有 \$c = -d\$。故

$$\mu \in \left[\bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

的置信水平为 \$1 - \alpha\$。

方差未知，估计均值

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

于是

$$\mu \in \left[\bar{X} - t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}} \right].$$

均值未知，估计方差

$$Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

等尾区间为

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \right].$$

由于卡方密度不对称，等尾区间一般不等于最短区间。若求最短区间，需要在

$$F_{\chi^2}(d) - F_{\chi^2}(c) = 1 - \alpha$$

下最小化 $1/c - 1/d$ 。

6.7.4 二项比例的大样本区间

若 $X_i \sim \text{Bernoulli}(p)$ ， $\bar{X}$ 近似满足

$$\frac{\bar{X} - p}{\sqrt{p(1-p)/n}} \stackrel{\text{approx}}{\sim} N(0, 1).$$

令 $\lambda = u_{1-\alpha/2}^2/n$ ，直接解不等式

$$\frac{n(\bar{X} - p)^2}{p(1-p)} \leq u_{1-\alpha/2}^2$$

得到 Wilson 区间

$$p \in \frac{\bar{X} + \lambda/2 \pm u_{1-\alpha/2} \sqrt{\bar{X}(1-\bar{X})/n + \lambda/(4n)}}{1 + \lambda}.$$

当 n 大时，它近似为熟悉的

$$\bar{X} \pm u_{1-\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

若希望近似区间长度不超过 L_0 ，利用 $p(1-p) \leq 1/4$ 可作保守设计：

$$L \lesssim \frac{u_{1-\alpha/2}}{\sqrt{n}} \leq L_0 \implies n \geq \left(\frac{u_{1-\alpha/2}}{L_0} \right)^2.$$

6.7.5 两个正态总体的均值差

设两组样本相互独立，

$$X_1, \dots, X_n \sim N(\mu_1, \sigma_1^2), \quad Y_1, \dots, Y_m \sim N(\mu_2, \sigma_2^2).$$

目标是估计 $\mu_1 - \mu_2$ 。根据 σ_1, σ_2 是否已知、未知时是否可认为相等，可分别使用正态枢轴量、合并方差 t 枢轴量，或自由度近似的 Welch 枢轴量。若已知 $\sigma_1 = c\sigma_2$ ，也可把共同尺度提取后重新构造枢轴量。

第七章 假设检验

7.1 统计假设的三种观点

7.1.1 参数模型

设参数空间为 $\Theta \subset \mathbb{R}^d$ 。对每个 $\theta \in \Theta$ ，给定一个分布 $F(x; \theta)$ 。在概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上取随机变量 $X_i: \Omega \rightarrow \mathbb{R}$ ，并有

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F(x; \theta).$$

假设检验是在这样的参数模型中，对总体参数提出假设，再由样本作出判断。

Fisher 观点 提出 $\theta \in \Theta_0$ ，再用数据判断这个命题是否受到支持。

Neyman–Pearson 观点 把检验看作决策问题：给出 H_0 （零假设）与 H_1 （对立假设），再把两者分别与决策行动联系起来。

Bayes 观点 在参数空间上引入先验，由数据得到后验密度 $\pi(\theta | x)$ ；若 $\mathbb{P}(\theta \in \Theta_0 | x) \geq 1 - \alpha$ ，则按这一标准接受相应假设。

7.1.2 女士品茶

例：女士品茶

十杯茶中五杯先加奶、五杯先加茶，而且这一配制结果已经确定。女士需要判断每杯的配制次序，结果两类各判断正确四杯，共判断正确八杯。

在 Fisher 的观点下，取

$$H_0: \text{女士不具备鉴别能力.}$$

若 H_0 成立，她的选择相当于从十杯中随机指出五杯。两类各判断正确四杯的概率为

$$\mathbb{P}_{H_0}(\text{两类各对四杯}) = \frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}} = \frac{25}{252} \approx 10\% > 5\%.$$

因此这组数据不足以支持她具有鉴别能力。若十杯全部判断正确，其概率为 $1/\binom{10}{5} = 1/252 < 5\%$ ，才会达到通常的显著性标准。

在 Neyman–Pearson 的观点下，写成 H_0 ：“不具备鉴别能力”与 H_1 ：“具备鉴别能力”，并分别联系行动： H_0 一侧不作进一步处理， H_1 一侧接受相应说明或继续实验。

7.1.3 假设的形式

原假设与对立假设分别写为

$$H_0: \theta \in \Theta_0, \quad H_1: \theta \in \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

这里并不强求 $\Theta_0 \cup \Theta_1 = \Theta$ 。常见形式包括

$$\theta = \theta_0, \quad \theta \leq \theta_0, \quad \theta > \theta_0.$$

若假设中的参数集合只有一个点, 称为**简单假设**; 否则称为 **复合假设**。

7.1.4 Fisher 情形: p 值

选择统计量

$$T = T(X_1, \dots, X_n),$$

使 T 越大表示样本越偏离 H_0 。对观测值 $x = (x_1, \dots, x_n)$, 定义

$$p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\{T(X) \geq T(x)\}.$$

它表示: 在 H_0 允许的参数中, 出现当前统计量或更极端统计量的最大概率。 $p(x)$ 越小, 说明观测 x 在 H_0 下越不典型; 通常当

$$p(x) < 0.05$$

时, 认为数据无法支持 H_0 。

7.1.5 Neyman–Pearson 情形: 拒绝域

记随机样本向量与它的观测值为

$$\mathbf{X} = (X_1, \dots, X_n), \quad \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n.$$

在样本空间中选取一个区域 W :

$$\mathbf{x} \in W \implies \text{拒绝 } H_0, \text{ 接受 } H_1, \quad \mathbf{x} \notin W \implies \text{不拒绝 } H_0.$$

W 称为**拒绝域**。若选取统计量 $T(\mathbf{X})$, 常把它写成

$$W = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) > c\}.$$

观测结果	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$\mathbf{x} \in W$	第一类错误	正确
$\mathbf{x} \notin W$	正确	第二类错误

7.1.6 两类错误与功效函数

第一类错误概率为

$$\alpha(\theta) = \mathbb{P}_\theta(\mathbf{X} \in W), \quad \theta \in \Theta_0.$$

第二类错误概率为

$$\beta(\theta) = \mathbb{P}_\theta(\mathbf{X} \notin W) = 1 - \mathbb{P}_\theta(\mathbf{X} \in W), \quad \theta \in \Theta_1.$$

定义功效函数

$$g(\theta) = \mathbb{P}_\theta(\mathbf{X} \in W), \quad \theta \in \Theta,$$

则

$$\alpha(\theta) = g(\theta) \quad (\theta \in \Theta_0), \quad \beta(\theta) = 1 - g(\theta) \quad (\theta \in \Theta_1).$$

实际构造拒绝域时, 先尽量使第一类错误小, 通常要求

$$\sup_{\theta \in \Theta_0} \alpha(\theta) \leq \alpha_0,$$

其中常取 $\alpha_0 = 0.05$; 在这一约束下, 再考察第二类错误。

推导：由功效函数确定临界值

设

$$X_i \sim N(\mu, \sigma_0^2), \quad H_0: \mu > \mu_0, \quad H_1: \mu \leq \mu_0,$$

并取拒绝域 $W = \{\bar{X} \leq c\}$ 。因为 $\bar{X} \sim N(\mu, \sigma_0^2/n)$ ，故

$$\begin{aligned} g(\mu) &= \mathbb{P}_\mu(\bar{X} \leq c) \\ &= \mathbb{P}_\mu\left(\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leq \frac{c - \mu}{\sigma_0/\sqrt{n}}\right) \\ &= \Phi\left(\frac{c - \mu}{\sigma_0/\sqrt{n}}\right). \end{aligned}$$

它随 μ 单调下降，因此 H_0 内最大的第一类错误出现在边界 $\mu \downarrow \mu_0$ 。令该极限等于 α_0 ，则

$$\Phi\left(\frac{c - \mu_0}{\sigma_0/\sqrt{n}}\right) = \alpha_0,$$

从而

$$c = \mu_0 + \frac{\sigma_0}{\sqrt{n}} u_{\alpha_0}.$$

这里 u_q 表示标准正态分布的下侧 q 分位数。

7.2 正态总体参数的假设检验

以下统一记 u_q 、 $t_q(\nu)$ 、 $\chi_q^2(\nu)$ 为相应分布的下侧 q 分位数。这里沿着前面的两条思路同时往下走：先由 Neyman-Pearson 的方法控制第一类错误并确定拒绝域，再计算 Fisher 意义下的 p 值，衡量观测结果对 H_0 的偏离程度。

7.2.1 方差已知时均值的 Z 检验

设

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma_0^2),$$

其中 σ_0^2 已知。边界零假设下的标准化统计量为

$$U = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1).$$

I. $H_0: \mu \leq \mu_0$ 对 $H_1: \mu > \mu_0$ 当 μ 增大时， $\bar{X}$ 整体向右移动，因此取 U 较大时拒绝 H_0 。先设拒绝域为

$$W_1 = \{U \geq c\}.$$

其功效函数为

$$g(\mu) = \mathbb{P}_\mu(U \geq c) = 1 - \Phi\left(c - \frac{\mu - \mu_0}{\sigma_0/\sqrt{n}}\right),$$

它随 μ 单调增加。因此在复合零假设 $\mu \leq \mu_0$ 中，第一类错误概率在边界 $\mu = \mu_0$ 处最大。令边界上的第一类错误等于 α ：

$$\mathbb{P}_{\mu_0}(U \geq c) = \alpha,$$

便有 $c = u_{1-\alpha}$ ，故

$$W_1 = \{U \geq u_{1-\alpha}\} = \left\{\bar{X} \geq \mu_0 + \frac{\sigma_0}{\sqrt{n}} u_{1-\alpha}\right\},$$

这就是 Neyman-Pearson 意义下的拒绝域。

再从 Fisher 的观点计算 p 值。记

$$u_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}.$$

在 H_0 中, 右尾概率仍在边界 $\mu = \mu_0$ 处最大, 所以

$$p_{\text{I}} = \sup_{\mu \leq \mu_0} \mathbb{P}_{\mu}(U \geq u_{\text{obs}}) = 1 - \Phi(u_{\text{obs}}).$$

II. $H_0: \mu \geq \mu_0$ 对 $H_1: \mu < \mu_0$ 这时较小的 U 才反对 H_0 。由正态分布的对称性, 或完全重复上面的功效函数计算, 可得

$$W_{\text{II}} = \{U \leq u_{\alpha}\} = \left\{ \bar{X} \leq \mu_0 + \frac{\sigma_0}{\sqrt{n}} u_{\alpha} \right\}.$$

Fisher 意义下的左尾概率为

$$p_{\text{II}} = \sup_{\mu \geq \mu_0} \mathbb{P}_{\mu}(U \leq u_{\text{obs}}) = \Phi(u_{\text{obs}}).$$

III. $H_0: \mu = \mu_0$ 对 $H_1: \mu \neq \mu_0$ 两侧的偏离都构成反对 H_0 的证据, 因而取统计量 $|U|$ 。在 H_0 下, $U \sim N(0, 1)$, 把显著性水平 α 平分到两侧, 得到

$$W_{\text{III}} = \{|U| \geq u_{1-\alpha/2}\}.$$

对应的 p 值为

$$p_{\text{III}} = \mathbb{P}_{\mu_0}(|U| \geq |u_{\text{obs}}|) = 2[1 - \Phi(|u_{\text{obs}}|)].$$

IV、V. 单点零假设的单侧备择 若把 I、II 中的复合零假设分别缩成单点, 得到

$$\text{IV: } H_0: \mu = \mu_0, \quad H_1: \mu > \mu_0,$$

$$\text{V: } H_0: \mu = \mu_0, \quad H_1: \mu < \mu_0.$$

由于上面功效函数关于 μ 的单调性, I 的最不利点本来就在 μ_0 , 所以 IV 与 I 使用同一个拒绝域; 同理, V 与 II 使用同一个拒绝域。这也解释了为什么单侧检验的零假设常写成含等号的复合形式, 而临界值仍只需在边界点计算。

7.2.2 方差未知时均值的 t 检验

当 σ^2 未知时, 以样本标准差

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

替代 σ , 并取

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \quad (H_0 \text{ 的边界上}).$$

记观测值为 t_{obs} 。与方差已知时完全相同, 只需把标准正态分位数换成 $t(n-1)$ 分位数。

I. $H_0: \mu \leq \mu_0$ 对 $H_1: \mu > \mu_0$

$$W_{\text{I}} = \{T \geq t_{1-\alpha}(n-1)\}, \quad p_{\text{I}} = 1 - F_{t_{n-1}}(t_{\text{obs}}).$$

II. $H_0: \mu \geq \mu_0$ 对 $H_1: \mu < \mu_0$

$$W_{\text{II}} = \{T \leq t_{\alpha}(n-1)\}, \quad p_{\text{II}} = F_{t_{n-1}}(t_{\text{obs}}).$$

III. $H_0: \mu = \mu_0$ 对 $H_1: \mu \neq \mu_0$

$$W_{\text{III}} = \{|T| \geq t_{1-\alpha/2}(n-1)\},$$

$$p_{\text{III}} = 2[1 - F_{t_{n-1}}(|t_{\text{obs}}|)],$$

其中 $F_{t_{n-1}}$ 是自由度 $n-1$ 的 t 分布函数。和前面一样，单点零假设配单侧备择时，拒绝域分别与 I、II 相同。

7.2.3 假设检验与区间估计的关系

双侧 t 检验的不拒绝域为

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| < t_{1-\alpha/2}(n-1).$$

把不等式改写为关于 μ_0 的条件，得到

$$\bar{X} - t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}} < \mu_0 < \bar{X} + t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}}.$$

这恰好表示 μ_0 落在置信水平 $1-\alpha$ 的双侧置信区间内。同理，右侧检验 $H_0: \mu \leq \mu_0$ 的不拒绝条件，对应于单侧置信下限

$$\mu_L = \bar{X} - t_{1-\alpha}(n-1) \frac{S}{\sqrt{n}}.$$

所以置信区间可以理解为：所有在相应显著性水平下不会被拒绝的参数值所成的集合。

7.2.4 正态总体方差的检验

设均值未知。检验 $\sigma^2 = \sigma_0^2$ 时使用枢轴量

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

记观测值为 q_{obs} 。

I. $H_0: \sigma^2 \leq \sigma_0^2$ 对 $H_1: \sigma^2 > \sigma_0^2$ 样本方差过大时拒绝 H_0 ，所以

$$W_{\text{I}} = \{Q \geq \chi_{1-\alpha}^2(n-1)\}, \quad p_{\text{I}} = 1 - F_{\chi_{n-1}^2}(q_{\text{obs}}).$$

II. $H_0: \sigma^2 \geq \sigma_0^2$ 对 $H_1: \sigma^2 < \sigma_0^2$ 样本方差过小时拒绝 H_0 ，所以

$$W_{\text{II}} = \{Q \leq \chi_{\alpha}^2(n-1)\}, \quad p_{\text{II}} = F_{\chi_{n-1}^2}(q_{\text{obs}}).$$

III. $H_0: \sigma^2 = \sigma_0^2$ 对 $H_1: \sigma^2 \neq \sigma_0^2$ 把显著性水平分配到卡方分布的两端，得到

$$W_{\text{III}} = \{Q \leq \chi_{\alpha/2}^2(n-1)\} \cup \{Q \geq \chi_{1-\alpha/2}^2(n-1)\},$$

$$p_{\text{III}} = 2 \min\{F_{\chi_{n-1}^2}(q_{\text{obs}}), 1 - F_{\chi_{n-1}^2}(q_{\text{obs}})\}.$$

7.3 其他分布参数的假设检验

7.3.1 指数分布均值参数

设

$$X_i \sim \text{Exp}(1/\theta), \quad \mathbb{E}X_i = \theta,$$

并记 $S = \sum_{i=1}^n X_i$ 。由 Gamma 分布的可加性，

$$S \sim \text{Ga}(n, 1/\theta), \quad f_S(s) = \frac{(1/\theta)^n}{\Gamma(n)} s^{n-1} e^{-s/\theta}, \quad s > 0.$$

推导：从 Gamma 分布变换到卡方分布

令

$$Y = \frac{2S}{\theta}, \quad S = \frac{\theta}{2}Y, \quad \left| \frac{ds}{dy} \right| = \frac{\theta}{2}.$$

于是

$$\begin{aligned} f_Y(y) &= f_S\left(\frac{\theta}{2}y\right) \frac{\theta}{2} \\ &= \frac{(1/\theta)^n}{\Gamma(n)} \left(\frac{\theta}{2}y\right)^{n-1} e^{-y/2} \frac{\theta}{2} \\ &= \frac{1}{2^n \Gamma(n)} y^{n-1} e^{-y/2}, \quad y > 0. \end{aligned}$$

这正是自由度 $2n$ 的卡方密度，故

$$\frac{2S}{\theta} \sim \chi^2(2n).$$

考虑

$$H_0: \theta \leq \theta_0, \quad H_1: \theta > \theta_0.$$

S 越大越支持 H_1 ，且 $\mathbb{P}_\theta(S \geq c)$ 随 θ 增加，故零假设内的最不利点是 $\theta = \theta_0$ 。令边界上的右尾概率为 α ，得到

$$\mathbb{P}_{\theta_0}\left(\frac{2S}{\theta_0} \geq \chi_{1-\alpha}^2(2n)\right) = \alpha.$$

因此

$$W = \left\{ \frac{2S}{\theta_0} \geq \chi_{1-\alpha}^2(2n) \right\}.$$

等价地，对给定样本，所有满足

$$\theta_0 \leq \frac{2S}{\chi_{1-\alpha}^2(2n)}$$

的零假设边界都会落入拒绝区域。这样写也说明了临界值随样本和分位数如何产生。

7.3.2 Bernoulli 分布的成功概率

设 $X_i \sim \text{Bernoulli}(p)$ ，则

$$S = \sum_{i=1}^n X_i \sim \text{Bin}(n, p).$$

对

$$H_0: p \leq p_0, \quad H_1: p > p_0,$$

自然取 $W = \{S \geq c\}$ 。第一类错误函数为

$$\alpha(p; c) = \mathbb{P}_p(S \geq c) = \sum_{k=c}^n \binom{n}{k} p^k (1-p)^{n-k}.$$

推导：二项分布上尾概率的单调性

使用恒等式 $k \binom{n}{k} = n \binom{n-1}{k-1}$ 和 $(n-k) \binom{n}{k} = n \binom{n-1}{k}$ ，逐项求导：

$$\begin{aligned} \frac{\partial \alpha}{\partial p} &= n \sum_{k=c}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &\quad - n \sum_{k=c}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1}. \end{aligned}$$

第一项令 $j = k - 1$ 后, 与第二项从 $j = c$ 到 $n - 1$ 的部分逐项抵消, 只留下 $j = c - 1$ 的首项, 因此

$$\frac{\partial \alpha}{\partial p} = n \binom{n-1}{c-1} p^{c-1} (1-p)^{n-c} > 0.$$

故上尾概率随 p 严格增加, 零假设内最大第一类错误在 $p = p_0$ 取得。

选择最小整数 c_0 , 使

$$\mathbb{P}_{p_0}(S \geq c_0) \leq \alpha_0, \quad \mathbb{P}_{p_0}(S \geq c_0 - 1) > \alpha_0.$$

由于离散性, 实际检验水平可能严格小于 α_0 。在 $p > p_0$ 上,

$$\beta(p) = \mathbb{P}_p(S < c_0) = 1 - \alpha(p; c_0)$$

随 p 单调下降。样本量较大时, 可由中心极限定理把标准化的 S 近似为标准正态变量, 从而转化为近似 Z 检验。

7.4 似然比与拟合优度

7.4.1 广义似然比检验

对 $H_0: \theta \in \Theta_0$ 与 $H_1: \theta \in \Theta \setminus \Theta_0$, 定义

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)} = \frac{L(\hat{\theta}_0; x)}{L(\hat{\theta}; x)}.$$

因为分子是在更小的参数空间上取上确界, $0 \leq \Lambda \leq 1$ 。 Λ 越小, 说明零假设的最佳解释仍远逊于完整模型的最佳解释, 故以 $\{\Lambda \leq c\}$ 构造拒绝域。

推导: 正态均值的似然比检验化为双侧 t 检验

设 $X_i \sim N(\mu, \sigma^2)$, 检验

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0,$$

其中 σ^2 未知。完整模型中的最大似然估计为

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2;$$

零假设下则只有方差需要极大化, 得到

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2.$$

在两个极大点处, 指数部分都等于 $e^{-n/2}$, 所以

$$\Lambda = \left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \mu_0)^2} \right]^{n/2}.$$

再利用平方和分解

$$\begin{aligned} \sum_{i=1}^n (X_i - \mu_0)^2 &= \sum_{i=1}^n [(X_i - \bar{X}) + (\bar{X} - \mu_0)]^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2, \end{aligned}$$

其中交叉项因 $\sum_i (X_i - \bar{X}) = 0$ 消失。于是

$$\begin{aligned}\Lambda &= \left[1 + \frac{(\bar{X} - \mu_0)^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right]^{-n/2} \\ &= \left(1 + \frac{T^2}{n-1} \right)^{-n/2}, \quad T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.\end{aligned}$$

Λ 随 $|T|$ 单调下降, 因此“ Λ 足够小”等价于“ $|T|$ 足够大”, 广义似然比检验与双侧 t 检验完全等价。

在适当的正则条件下, Wilks 定理给出

$$-2 \log \Lambda(X) \xrightarrow{d} \chi^2(d), \quad d = \dim \Theta - \dim \Theta_0.$$

因此即使 Λ 的有限样本精确分布很难求, 也可在大样本下使用卡方分布确定临界值。

7.4.2 Pearson 卡方拟合优度检验

设一次观测落入 k 个类别之一, 各类概率为 $p_1, \dots, p_k$ 。在 n 次独立观测中, 各类频数满足

$$(N_1, \dots, N_k) \sim \text{Multinomial}(n; p_1, \dots, p_k), \quad \sum_{i=1}^k N_i = n.$$

零假设给出理论概率 p_i , 相应的期望频数为 $E_i = np_i$ 。多项分布的矩为

$$\mathbb{E}N_i = np_i = E_i, \quad \text{Var}(N_i) = np_i(1 - p_i),$$

$$\text{Cov}(N_i, N_j) = -np_i p_j \quad (i \neq j).$$

令误差向量

$$\boldsymbol{\varepsilon} = (N_1 - E_1, \dots, N_k - E_k)^\top,$$

为什么不直接计算误差向量的模长? 一个最直接的想法是用

$$\|\boldsymbol{\varepsilon}\|^2 = \sum_{i=1}^k (N_i - E_i)^2$$

衡量观测频数与理论频数的距离。但各类误差的自然波动尺度并不相同: $\text{Var}(N_i) = np_i(1 - p_i)$ 随 p_i 改变。例如, 同样相差 2 次, 对期望频数 $E_i = 4$ 的类别可能很显著, 对 $E_i = 100$ 的类别却未必异常。另一方面, 由 $\sum_i N_i = \sum_i E_i = n$ 可知 $\sum_i \varepsilon_i = 0$, 各分量彼此相关; 普通欧氏模长却把它们当作同尺度、相互独立的坐标。因此需要先写出误差向量的协方差矩阵, 并按照它给出的尺度与相关性加权。若协方差矩阵可逆, 相应的距离应为

$$\boldsymbol{\varepsilon}^\top \Sigma^{-1} \boldsymbol{\varepsilon}.$$

并写

$$D = \text{diag}(p_1, \dots, p_k), \quad \boldsymbol{p} = (p_1, \dots, p_k)^\top.$$

则其协方差矩阵为

$$\Sigma = n(D - \boldsymbol{p}\boldsymbol{p}^\top).$$

由于 $D\mathbf{1} = \boldsymbol{p}$ 且 $\boldsymbol{p}^\top \mathbf{1} = 1$,

$$\Sigma \mathbf{1} = n(D\mathbf{1} - \boldsymbol{p}\boldsymbol{p}^\top \mathbf{1}) = 0.$$

这对应于约束 $\sum_i \varepsilon_i = 0$, 说明 k 个误差中只有 $k-1$ 个自由方向, 也说明 Σ 不能直接求逆。所以下面先用扰动法处理这个约束, 再让扰动趋于零。

推导：扰动法与 Pearson 统计量

为看清奇异协方差在约束子空间上的二次型，先作一个小扰动，令

$$\sum_{i=1}^k p_i = 1 - \delta, \quad \delta \neq 0.$$

此时 $D - \mathbf{p}\mathbf{p}^\top$ 可逆。由 Sherman-Morrison 公式，

$$(D - \mathbf{p}\mathbf{p}^\top)^{-1} = D^{-1} + \frac{D^{-1}\mathbf{p}\mathbf{p}^\top D^{-1}}{1 - \mathbf{p}^\top D^{-1}\mathbf{p}}.$$

注意

$$D^{-1}\mathbf{p} = \mathbf{1}, \quad \mathbf{p}^\top D^{-1}\mathbf{p} = \sum_{i=1}^k p_i = 1 - \delta,$$

所以

$$\Sigma^{-1} = \frac{1}{n} \left(D^{-1} + \frac{1}{\delta} \mathbf{1}\mathbf{1}^\top \right).$$

从而

$$\begin{aligned} \boldsymbol{\varepsilon}^\top \Sigma^{-1} \boldsymbol{\varepsilon} &= \frac{1}{n} \left(\boldsymbol{\varepsilon}^\top D^{-1} \boldsymbol{\varepsilon} + \frac{1}{\delta} (\mathbf{1}^\top \boldsymbol{\varepsilon})^2 \right) \\ &= \sum_{i=1}^k \frac{\varepsilon_i^2}{np_i} + \frac{1}{n\delta} \left(\sum_{i=1}^k \varepsilon_i \right)^2. \end{aligned}$$

扰动模型中 $\sum_i E_i = n(1 - \delta)$ 、 $\sum_i N_i = n$ ，故 $\sum_i \varepsilon_i = n\delta$ ，第二项等于 $n\delta$ 。令 $\delta \rightarrow 0$ ，便留下

$$\chi^2 = \sum_{i=1}^k \frac{\varepsilon_i^2}{np_i} = \sum_{i=1}^k \frac{(N_i - E_i)^2}{E_i}.$$

大样本下，多项分布的中心极限定理使误差在 $\mathbf{1}^\perp$ 这个 $k - 1$ 维子空间中渐近正态，因此

$$\chi^2 \xrightarrow{d} \chi^2(k - 1).$$

Pearson 检验的拒绝域为

$$W = \{\chi^2 \geq \chi_{1-\alpha}^2(k - 1)\}.$$

例：Mendel 豌豆实验

四类豌豆的观测频数为

$$(315, 108, 101, 32), \quad n = 556,$$

理论比例为 $9:3:3:1$ ，即

$$\mathbf{p} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right).$$

代入 Pearson 统计量得到

$$\chi^2 = \sum_{i=1}^4 \frac{(N_i - 556p_i)^2}{556p_i} \approx 0.47.$$

在 $\alpha = 0.05$ 下， $\chi_{0.95}^2(3) \approx 7.8$ ，且

$$p = \mathbb{P}\{\chi^2(3) \geq 0.47\} \approx 0.9254.$$

观测偏差很小，不能拒绝 $9:3:3:1$ 的理论比例。

7.4.3 连续总体的拟合优度

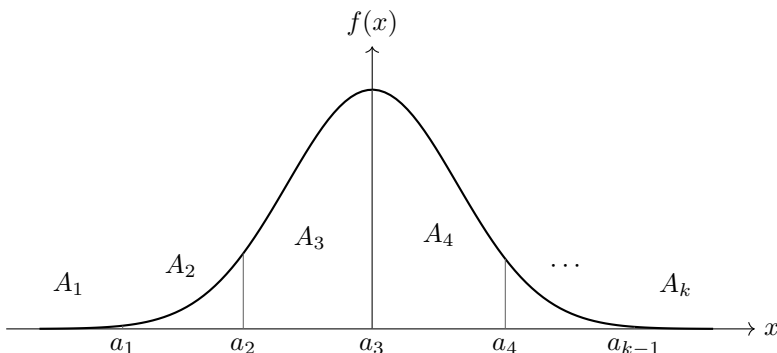
要检验连续总体的分布，可先把实轴分成互不相交的区域

$$A_i = (a_{i-1}, a_i], \quad i = 1, \dots, k,$$

其中 $a_0 = -\infty$ 、 $a_k = +\infty$ 。在候选分布函数 F 下，

$$p_i = \mathbb{P}(X \in A_i) = F(a_i) - F(a_{i-1}).$$

把样本按区间归类后，频数向量就是一个多项分布，可继续使用同一个 Pearson 统计量。



图：把连续总体按区间归类，再检验各区间的理论概率

因此卡方拟合优度检验本质上检验的是**整个分布**，而不只是某一个参数。分组以后，实际分类概率向量 $\mathbf{p}$ 属于有限维参数空间。若连续模型中的参数未知，需要先用样本估计参数；每估计一个独立参数，渐近自由度通常相应减少一维。

例：实验数据的末位数字

把末位数字 $0, 1, \dots, 9$ 看成十个类别。若记录机制没有末位偏好，零假设为

$$p_0 = p_1 = \dots = p_9 = \frac{1}{10}.$$

统计各数字的出现次数后，可用 Pearson 统计量检验它们是否与均匀分布相容。

模型必须说明是在观察数据之前指定，还是看过数据之后才提出。例如发现“第 5 类出现最多”后再检验第 5 类，会忽略事后挑选带来的额外机会；更合适的统计量可能是

$$Y = \max(N_1, \dots, N_k),$$

并在零假设下研究 $\mathbb{P}(Y \geq y_{\text{obs}})$ 。统计结论不仅取决于计算，也取决于假设是如何产生的。

7.4.4 ◆ 参考书补充

◆ 两个独立正态总体的均值差，可根据方差已知、未知但相等或未知且不等，分别使用 Z 检验、合并方差 t 检验或 Welch 型检验；成对观测则先取差值，转化为单总体均值检验。

◆ 二维列联表中的独立性检验同样使用 Pearson 统计量。理论频数由“行边际频率 $\times$ 列边际频率”估计，自由度为 $(r-1)(c-1)$ 。

◆ 正态性检验可先借助正态概率纸或分位数图观察样本是否近似落在一条直线上，再用专门的统计量作判断。图形方法便于发现偏态、厚尾与离群点，正式检验则负责量化这种偏离是否超过随机波动所能解释的范围。

◆ 当总体分布形式不便作参数化假定时，可采用游程检验、符号检验与秩和检验。这些非参数方法用次序、符号或秩构造统计量，通常以较少的分布假设换取更广的适用性。