

# 概率论与数理统计

(人工智能或计算机专业用书)

DRAFT

DO NOT DISTRIBUTE



# 目 录

第 1 章 随机事件与概率	1
1.1 随机事件及其运算	1
1.2 频率与概率公理化	7
1.3 古典概型与几何概型	12
1.4 组合计数*	19
习题	26
第 2 章 条件概率与独立性	29
2.1 条件概率	29
2.2 全概率公式和贝叶斯公式	33
2.3 事件独立性	38
2.4 案例分析	45
习题	50
第 3 章 离散型随机变量	51
3.1 离散型随机变量及分布列	52
3.2 离散型随机变量的期望	53
3.3 离散型随机变量的方差	57
3.4 常用离散型随机变量	59
3.5 案例分析	67
习题	71
第 4 章 连续型随机变量	73
4.1 分布函数	73
4.2 概率密度函数	76
4.3 连续型随机变量的期望和方差	80
4.4 常用连续型随机变量	82
4.5 连续随机变量函数的分布	91
4.6 常用分布的随机数*	95
习题	98

第 5 章 多维随机向量 .....	101
5.1 二维联合分布函数 .....	101
5.2 二维离散型随机向量 .....	104
5.3 二维连续型随机向量 .....	106
5.4 随机变量的独立性 .....	110
5.5 条件分布 .....	115
5.6 多维随机变量函数的分布 .....	119
5.7 多维正太分布 .....	128
习题 .....	133
第 6 章 多维随机向量的数字特征 .....	137
6.1 多维随机向量函数的期望 .....	137
6.2 协方差 .....	140
6.3 相关系数 .....	145
6.4 条件期望 .....	147
6.5 随机向量的数学期望与协方差阵 .....	150
6.6 应用案例 .....	151
习题 .....	153
第 7 章 集中不等式 (Concentration) .....	155
7.1 基础不等式 .....	156
7.2 Chernoff 不等式 .....	159
7.3 Bennet 和 Bernstein 不等式 .....	167
7.4 应用: 随机投影 (Random Projection) .....	169
习题 .....	172
第 8 章 大数定律及中心极限定理 .....	175
8.1 大数定律 .....	175
8.2 中心极限定理 .....	177
习题 .....	181
第 9 章 统计的基本概念 .....	183
9.1 总体 (population) 与样本 (sample) .....	183
9.2 常用统计量 .....	184
9.3 Beta 分布、 $\Gamma$ 分布、Dirichlet 分布 .....	187
9.4 正态总体抽样分布定理 .....	192
习题 .....	197

第 10 章 参数估计 .....	199
10.1 点估计 .....	199
10.2 估计量的评价标准 .....	204
10.3 区间估计 .....	209
第 11 章 假设检验(Hypothesis Testing) .....	217
11.1 正态总体期望的假设检验 .....	219
11.2 正态分布的方差假设检验. ....	222
11.3 非参假设检验 .....	222
习题 .....	225



# 第 1 章 随机事件与概率

对自然界和人类社会存在的各种现象进行观察, 会发现有一类现象在一定条件下是必然发生的, 常被称为 **必然现象**, 又称 **确定性现象**. 例如, 太阳从东边升起; 成熟的苹果会从树上掉下来; 在标准大气压下, 水在  $0^{\circ}\text{C}$  以下会结冰, 加热到  $100^{\circ}\text{C}$  以上会沸腾; 平面三角形两边之和大于第三边; 等等. 这些现象发生的条件与结果之间具有确定性关系, 可用确切的数学函数进行描述.

然而在自然界和人类社会也往往存在着另一类不确定的现象. 例如, 我们今晚能否观察到流星; 随意投掷一枚硬币, 可能正面朝上, 也可能反面朝上; 当你穿过马路时, 遇见的信号灯可能是绿色, 也可能是红色; 两位相恋的人最终能否走在一起; 等等. 这些现象在一定条件下可能出现这种结果, 也可能出现那种结果, 出现的结果并不唯一, 而事先不能确定哪种结果会出现, 常被称为 **随机现象**, 这类现象发生的条件与结果之间具有不确定性关系, 无法通过确切的数学函数进行刻画.

随机现象尽管在一次观察中无法确定哪种结果发生, 表现出不确定性或偶然性. 然而经过人们长期的研究后发现: 在大量重复的实验中, 随机现象的结果却表现出固有的规律性, 即 **统计规律性**. 例如多次重复投掷一枚硬币, 得到的正面朝上的次数和反面朝上的次数几乎总是差不多. 因此随机现象通常表现出二重属性:

- **偶然性**: 对随机现象进行一次观察, 其结果表现出不确定性;
- **必然性**: 对随机现象进行大量重复观察, 其结果呈现出固有的统计规律性.

概率论与数理统计是研究和揭示随机现象统计规律性的一门学科, 应用几乎遍及所有的科学技术领域、行业生产、国民经济生活等, 如法国著名数学家拉普拉斯 (P. S. Laplace, 1794-1827) 所言: “对生活的大部分, 最重要的问题实际上都是概率问题”. 图灵奖得主 Y. LeCun 在其自传中指出: “历史上大多数重要成果的出现都是偶然事件... 所有的努力都是为了提高概率”. 而对现实生活中的每个人而言: 所有的努力都是为了提高成功的概率.

## 1.1 随机事件及其运算

为研究和揭示随机现象的规律, 通常需要在相同的条件下重复进行一系列实验和观察, 常被称为 **随机试验**, 或简称为 **试验**. 一般用  $E$  或  $E_1, E_2, E_3, \dots$  表示, 本书所提及的试验均是随机试验. 下面给出一些随机试验的例子:

$E_1$ : 随意抛一枚硬币, 观察正面朝上还是反面朝上.

$E_2$ : 随意抛一枚骰子, 观察出现的点数.

$E_3$ : 统计某地区一年内出生的婴儿数量.

$E_4$ : 随机选取一盏电灯, 测试其寿命.

这些试验具有一些共有的特点: 试验在相同的条件下可重复进行, 具有多种结果, 我们已知每次试验所有可能的结果, 但在每次试验之前不确定出现哪种结果. 例如抛硬币有正面/反面朝上两种结果, 在相同的条件下可以重复进行, 且每次试验前不确定正面/反面朝上. 概括而言, 随机试验具有以下三个特点:

- **可重复**: 在相同的条件下试验可重复进行;
- **多结果**: 试验结果不唯一, 所有可能发生的结果事先明确已知;
- **不确定**: 试验前无法预测或确定哪一种结果会发生.

### 1.1.1 样本空间与随机事件

随机试验尽管在每次试验前不能确定发生的结果, 但其所有可能发生的结果却是事先已知的. 将随机试验  $E$  所有可能的结果构成的集合称为试验  $E$  的 **样本空间**, 记为  $\Omega$ . 样本空间  $\Omega$  中的每个元素, 即试验  $E$  的每种结果, 称为 **样本点**, 记为  $\omega$ .

例如前一页所述的四种试验, 其样本空间分别为:

试验  $E_1$  的样本空间为  $\Omega_1 = \{\text{正面}, \text{反面}\}$ , 样本点分别为  $\omega_1 = \text{正面}$ ,  $\omega_2 = \text{反面}$ .

试验  $E_2$  的样本空间为  $\Omega_2 = \{1, 2, 3, 4, 5, 6\}$ , 样本点分别为  $\omega_1 = 1, \omega_2 = 2, \dots, \omega_6 = 6$ .

试验  $E_3$  的样本空间为  $\Omega_3 = \{0, 1, 2, \dots\}$ , 样本点为任意非负整数.

试验  $E_4$  的样本空间为  $\Omega_4 = \{t: t \geq 0\}$ , 样本点为任意非负数.

包含有限个样本点的样本空间称为 **有限样本空间**, 如样本空间  $\Omega_1$  和  $\Omega_2$ . 包含无限但可列多个样本点的样本空间称为 **可列样本空间**, 如样本空间  $\Omega_3$ . 有限样本空间和无限可列样本空间统称为 **离散样本空间**. 包含无限不可列个样本点的样本空间称为 **不可列样本空间**, 如样本空间  $\Omega_4$ .

在随机试验中, 通常关心具有某些特性的样本点构成的集合, 称之为 **随机事件**, 简称为 **事件**, 一般用大写字母  $A, B, C, \dots$  表示. 随机事件的本质是集合, 由单个或某些样本点所构成的集合, 是样本空间  $\Omega$  的子集. 如果随机试验的结果是事件  $A$  中包含的元素, 则称 **事件  $A$  发生**.

只包含一样本点的事件称为 **基本事件**, 包含两个或两个以上样本点的事件称为 **复合事件**. 样本空间  $\Omega$  包含所有样本点, 是其自身的子集, 每次试验必然发生, 因而称  $\Omega$  为 **必然事件**. 另一方面, 如果某事件在每次试验中都不发生, 则该事件不可能包含任何样本点, 我们用空集符号  $\emptyset$  表示, 且称  $\emptyset$  为 **不可能事件**.

**例 1.1** 随机试验  $E$ : 抛一枚骰子观察其出现的点数, 其样本空间  $\Omega = \{1, 2, \dots, 6\}$ , 则:

事件  $A$  表示抛骰子的点数为 2, 则  $A = \{2\}$  为基本事件;

事件  $B$  表示抛骰子的点数为偶数, 则  $B = \{2, 4, 6\}$ ;

事件  $C$  表示抛骰子的点数大于 7, 则  $C = \emptyset$  为不可能事件;

事件  $D$  表示抛骰子的点数小于 7, 则  $D = \Omega$  为必然事件.



## 1.1.2 随机事件的关系与运算

随机事件的本质是样本空间的子集, 因此随机事件的关系与运算可类比于集合论的关系与运算. 下面默认随机试验的样本空间为  $\Omega$ , 用  $A, B, A_i$  ( $i = 1, 2, \dots$ ) 表示样本空间  $\Omega$  中的随机事件.

- 1) **包含事件** 若事件  $A$  发生必将导致事件  $B$  发生, 则称 **B 包含 A**, 记为  $A \subset B$  或  $B \supset A$ .

若  $A \subset B$  且  $B \subset A$ , 则称事件  $A$  与  $B$  **相等**, 记为  $A = B$ .

- 2) **事件的并/和** 若事件  $A$  和  $B$  中至少有一个发生所构成的事件称为 **事件 A 与 B 的并 (或和) 事件**, 记为  $A \cup B$ , 即

$$A \cup B = \{\omega: \omega \in A \text{ 或 } \omega \in B\}.$$

类似地, 事件  $A_1, A_2, \dots, A_n$  中至少有一个发生所构成的事件称为事件  $A_1, A_2, \dots, A_n$  的并事件, 记为

$$\bigcup_{i=1}^n A_i = A_1 \cup A_2 \cup \dots \cup A_n = \{\omega: \exists i \in [n] \text{ s.t. } \omega \in A_i\},$$

称  $\bigcup_{i=1}^{\infty} A_i$  为可列个事件  $A_1, A_2, \dots$  的并事件.

- 3) **事件的交/积** 若事件  $A$  和  $B$  同时发生所构成的事件称为 **事件 A 与 B 的交 (或积) 事件**, 记为  $A \cap B$  或  $AB$ , 即

$$AB = A \cap B = \{\omega: \omega \in A \text{ 且 } \omega \in B\}.$$

类似地, 事件  $A_1, A_2, \dots, A_n$  同时发生所构成的事件称为事件  $A_1, A_2, \dots, A_n$  的交事件, 记为

$$A_1 A_2 \dots A_n = \bigcap_{i=1}^n A_i = A_1 \cap A_2 \cap \dots \cap A_n = \{\omega: \forall i \in [n] \text{ s.t. } \omega \in A_i\},$$

称  $\bigcap_{i=1}^{\infty} A_i$  为可列个事件  $A_1, A_2, \dots$  的交事件.

- 4) **事件的差** 若事件  $A$  发生但事件  $B$  不发生所构成的事件称为 **事件 A 与 B 的差**, 记  $A - B$ ,

$$A - B = A - AB = A\bar{B} = \{\omega: \omega \in A \text{ 且 } \omega \notin B\}.$$

- 5) **对立/逆事件** 对事件  $A$  而言, 所有不属于事件  $A$  的基本事件所构成的事件称为 **事件 A 的对立事件 或 逆事件**, 记为  $\bar{A}$ , 即  $\bar{A} = \Omega - A$ .

根据定义可知  $\bar{\bar{A}} = A$ ,  $\bar{A} \cap A = \emptyset$  和  $\Omega = A \cup \bar{A}$ .

- 6) **互不相容/互斥事件** 若事件  $A$  和事件  $B$  不能同时发生, 即  $A \cap B = \emptyset$ , 则称事件  $A$  和  $B$  是 **互不相容的 或 互斥的**.

若事件  $A_1, A_2, \dots, A_n$  中任意两事件不可能同时发生, 即对任意  $i \neq j$  有  $A_i \cap A_j = \emptyset$  成立, 则称  $n$  个事件  $A_1, A_2, \dots, A_n$  是 **互不相容的 或 互斥的**, 类似地定义可列个互不相容的事件. 对立的事件是互不相容的, 但互不相容的事件并不一定是对立事件.

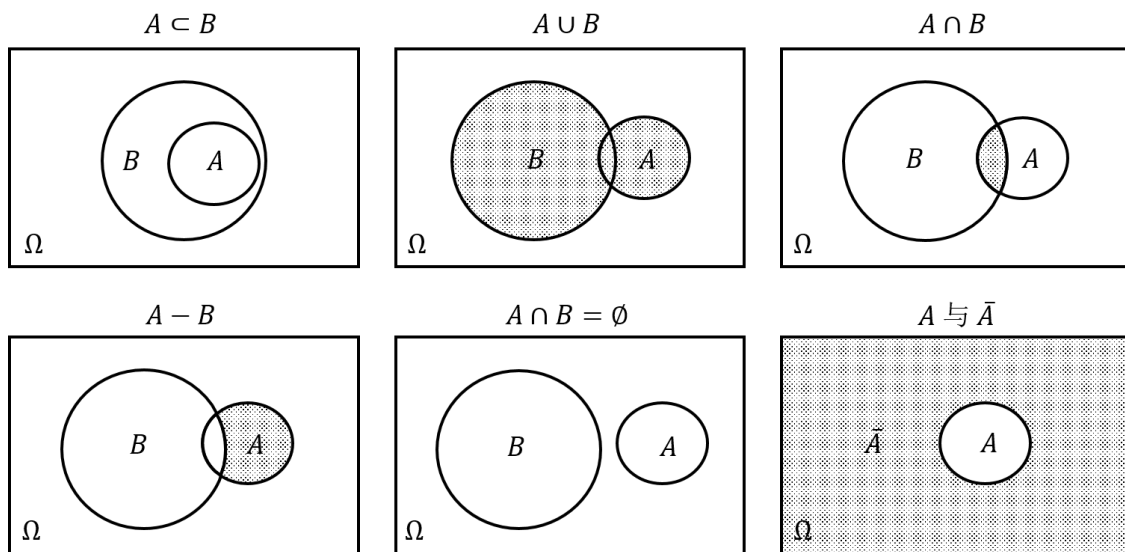


图 1.1 事件关系或运算通过韦恩图表示,  $A \cup B$ ,  $A \cap B$ ,  $A - B$ ,  $\bar{A}$  分别为图中阴影部分所示

如图 1.1 所示, 借助集合论的韦恩图 (Venn Diagram) 可直观地表示事件之间的关系或运算. 例如, 在  $A \subset B$  的图示中, 矩形表示样本空间  $\Omega$ , 椭圆  $A$  和  $B$  分别表示事件  $A$  和  $B$ , 椭圆  $B$  包含椭圆  $A$  则表示事件  $A \subset B$ ; 在  $A \cup B$  的图示中阴影部分表示并事件  $A \cup B$ .

根据定义可知道事件还满足下面的规律, 相关证明读者可参考集合的运算规律.

- 交换律:  $A \cup B = B \cup A$ ,  $A \cap B = B \cap A$ ;
- 结合律:  $(A \cup B) \cup C = A \cup (B \cup C)$ ,  $(A \cap B) \cap C = A \cap (B \cap C)$ ;
- 分配律:  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ ,  $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ ;
- 德摩根 (De Morgen) 律:  $\overline{A \cup B} = \bar{A} \cap \bar{B}$ ,  $\overline{A \cap B} = \bar{A} \cup \bar{B}$ .

上面的四条规律对有限个或可列个事件均成立, 例如对德摩根律有

$$\overline{\bigcup_{i=1}^n A_i} = \bigcap_{i=1}^n \bar{A}_i, \quad \overline{\bigcap_{i=1}^n A_i} = \bigcup_{i=1}^n \bar{A}_i, \quad \overline{\bigcup_{i=1}^{+\infty} A_i} = \bigcap_{i=1}^{+\infty} \bar{A}_i, \quad \overline{\bigcap_{i=1}^{+\infty} A_i} = \bigcup_{i=1}^{+\infty} \bar{A}_i.$$

此外, 若事件  $A \subset B$ , 有  $AB = A$  和  $A \cup B = B$  成立.

例 1.2 设  $A, B, C$  为三个随机事件, 则有

- 事件  $A$  与  $B$  同时发生, 而事件  $C$  不发生的事件可表示为  $AB\bar{C}$  或  $AB - C$ ;
- 这三个事件中至少有一个发生的事件可表示为  $A \cup B \cup C$ ;
- 这三个事件中恰好有一个发生的事件可表示为  $(A\bar{B}\bar{C}) \cup (\bar{A}B\bar{C}) \cup (\bar{A}\bar{B}C)$ ;

- 这三个事件中至多有一个发生的事件可表示为  $(A\bar{B}\bar{C}) \cup (\bar{A}B\bar{C}) \cup (\bar{A}\bar{B}C) \cup (\bar{A}\bar{B}\bar{C})$  或  $\overline{AB \cup AC \cup BC}$ ;
- 这三个事件中至少有两个发生的事件可表示为  $AB \cup AC \cup BC$ ;
- 这三个事件中至多有两个发生的事件可表示为  $\bar{A} \cup \bar{B} \cup \bar{C}$ ;
- 这三个事件中恰好有两个发生的事件可表示为  $AB\bar{C} \cup AC\bar{B} \cup BC\bar{A}$ .

**例 1.3** 设  $A, B, C$  为三个随机事件, 证明

$$(\bar{A} \cup B)(A \cup B)(\bar{A} \cup \bar{B})(A \cup \bar{B}) = \emptyset \quad \text{和} \quad (A - B) \cup (B - C) = (A \cup B) - BC.$$

**证明** 根据事件的分配律有  $(\bar{A} \cup B)(A \cup B) = (A \cap \bar{A}) \cup B = B$  以及  $(\bar{A} \cup \bar{B})(A \cup \bar{B}) = \bar{B}$ , 由此可得  $(\bar{A} \cup B)(A \cup B)(\bar{A} \cup \bar{B})(A \cup \bar{B}) = B \cap \bar{B} = \emptyset$ .

根据事件的差  $A - B = A\bar{B}$  可得  $(A - B) \cup (B - C) = (A\bar{B}) \cup (B\bar{C})$ . 根据事件的分配律和德摩根律有

$$\begin{aligned} (A \cup B) - BC &= (A \cup B)\overline{BC} = (A \cup B) \cap (\bar{B} \cup \bar{C}) \\ &= (A\bar{B}) \cup (A\bar{C}) \cup (B\bar{B}) \cup (B\bar{C}) = (A\bar{B}) \cup (A\bar{C}) \cup (B\bar{C}). \end{aligned}$$

由此可知  $((A - B) \cup (B - C)) \subset ((A \cup B) - BC)$ . 只需进一步证明  $A\bar{C} \subset (A\bar{B}) \cup (B\bar{C})$ , 对任意  $x \in A\bar{C}$ , 有  $x \in A$  且  $x \in \bar{C}$ , 再根据  $x \in B$  或  $x \in \bar{B}$  有  $x \in A\bar{B}$  或  $x \in B\bar{C}$  成立.

事件的关系与运算可类比于集合的关系与运算, 表 1.1 简要地给出了概率论和集合论之间对应关系, 即概率统计中事件的关系与运算可通过集合的方式进行描述.

**表 1.1** 概率论与集合论之间相关概念的对应关系

符号	概率论	集合论
$\Omega$	必然事件, 样本空间	全集
$\emptyset$	不可能事件	空集
$\omega$	基本事件	元素
$A$	随机事件	子集
$\bar{A}$	事件 $A$ 的对立事件	集合 $A$ 的补集
$\omega \in A$	事件 $A$ 发生	元素 $\omega$ 属于集合 $A$
$A \subset B$	事件 $A$ 发生导致 $B$ 发生	集合 $B$ 包含集合 $A$
$A = B$	事件 $A$ 与 $B$ 相等	集合 $A$ 与 $B$ 相等
$A \cup B$	事件 $A$ 与 $B$ 的并	集合 $A$ 与 $B$ 的并集
$A \cap B$	事件 $A$ 与 $B$ 的交	集合 $A$ 与 $B$ 的交集
$A - B$	事件 $A$ 与 $B$ 的差	集合 $A$ 与 $B$ 的差集
$AB = \emptyset$	事件 $A$ 与 $B$ 互不相容	集合 $A$ 与 $B$ 无相同元素

### 1.1.3 可测空间\*

设  $\Omega$  是一个样本空间, 用  $2^\Omega$  表示样本空间  $\Omega$  所有子集所构成的集合, 称为  $\Omega$  的 **幂集**, 即样本空间  $\Omega$  上所有事件所构成的集合. 对可列的样本空间, 将幂集  $2^\Omega$  中的元素都看作事件没什么不妥; 但对无限不可列样本空间, 一般情形下不将样本空间  $\Omega$  的一切子集都作为事件, 这将对概率的计算带来不可克服的困难. 例如, 在几何概型 (见 1.3.2 节) 中将不可测集作为事件则难以计算概率. 为了更好地刻画随机事件, 本节引入可测空间.

**定义 1.1** 设  $\Omega$  是一个样本空间且  $\Sigma \subseteq 2^\Omega$ , 若  $\Sigma$  满足以下三个条件:

- 必然事件  $\Omega \in \Sigma$ ;
- 若任意  $A \in \Sigma$ , 则有补集  $\bar{A} \in \Sigma$ ;
- 若任意  $A_i \in \Sigma$  ( $i = 1, 2, \dots$ ), 则有  $\bigcup_{i=1}^{+\infty} A_i \in \Sigma$ ,

则称  $\Sigma$  是样本空间  $\Omega$  的  $\sigma$  代数 (又称  $\sigma$  域), 称  $\Sigma$  中的元素 (一个子集) 是 **可测集**, 以及  $(\Omega, \Sigma)$  是一个 **可测空间**.

$\sigma$  代数  $\Sigma$  本质上是一个集合, 其每一个元素也是集合, 即  $\Sigma$  是  $\Omega$  一些子集所构成的集合. 若  $\Sigma$  是一个  $\sigma$  代数, 则  $\Sigma$  中每个元素都是可测集. 根据可测空间定义可知

$$\emptyset \in \Sigma, \quad \bigcup_{i=1}^n A_i \in \Sigma, \quad \bigcap_{i=1}^n A_i \in \Sigma, \quad \bigcap_{i=1}^{+\infty} A_i \in \Sigma.$$

给定样本空间  $\Omega$ , 最小的  $\sigma$  代数为  $\Sigma = \{\emptyset, \Omega\}$ , 最大的  $\sigma$  代数为  $\Sigma = 2^\Omega$ .

所关注的非空事件集合  $\mathcal{F} \subset 2^\Omega$  有时不一定满足  $\sigma$  代数, 此时可以构造包含  $\mathcal{F}$  的最小  $\sigma$  代数:

**定义 1.2** 给定样本空间  $\Omega$  和非空事件集合  $\mathcal{F} \subset 2^\Omega$ , 记  $\sigma(\mathcal{F})$  为包含  $\mathcal{F}$  的最小  $\sigma$  代数. 即若  $\Sigma$  是一个  $\sigma$  代数且  $\mathcal{F} \subset \Sigma$ , 则有  $\sigma(\mathcal{F}) \subset \Sigma$ .

例如当  $\mathcal{F} = \{A\}$  时, 即集合  $\mathcal{F}$  仅包含单一事件  $A$ , 则最小  $\sigma$  代数为  $\sigma(\mathcal{F}) = \{\emptyset, A, \bar{A}, \Omega\}$ . 对一般的事件集合  $\mathcal{F}$ , 有最小  $\sigma$  代数  $\sigma(\mathcal{F}) = \bigcap_{\mathcal{F} \subset \Sigma} \Sigma$ .

对有限或可列的样本空间  $\Omega$ , 一般考虑  $\sigma$  代数  $\Sigma = 2^\Omega$ ; 而当样本空间  $\Omega$  为实数集  $\mathbb{R}$  时, 一般考虑博雷尔  $\sigma$  代数, 即由有限或可列个开区间 (或闭区间) 构成的  $\sigma$  代数, 记为  $\mathfrak{R}_1$ , 即

$$\begin{aligned} \mathfrak{R}_1 &= \sigma(\{(a, b): a < b \in \mathbb{R}\}) = \sigma(\{[a, b]: a < b \in \mathbb{R}\}) = \sigma(\{(-\infty, b), b \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty, b], b \in \mathbb{R}\}) = \sigma(\{(a, +\infty), a \in \mathbb{R}\}) = \sigma(\{[a, +\infty), a \in \mathbb{R}\}). \end{aligned}$$

上式成立的原因有  $[a, b] = \bigcap_{n=1}^{\infty} (a + 1/n, b - 1/n)$ ,  $(a, b) = [a, b] \setminus a \setminus b$ , 以及可列次的逆、并、交等运算. 类似可定义  $n$  维博雷尔  $\sigma$  代数  $\mathfrak{R}_n$ .

## 1.2 频率与概率公理化

随机事件在一次试验中可能发生、也可能不发生, 我们通常关心随机事件发生的可能性究竟有多大, 最好能用介于 0 和 1 之间的一个数来进行刻画. 为此首先引入频率, 用以描述随机事件发生的频繁程度, 在此基础上引入事件的概率.

### 1.2.1 频率

**定义 1.3** 随机事件  $A$  在相同条件下重复进行的  $n$  次试验中出现了  $n_A$  次, 则称  $f_n(A) = n_A/n$  为事件  $A$  在  $n$  次试验中发生的 **频率**, 并称  $n_A$  为事件  $A$  发生的 **频数**.

事件的频率在一定程度上反应了事件发生的可能性, 若事件发生的频率越大, 则事件  $A$  发生越频繁, 因而事件在一次试验中发生的可能性越大. 根据上面的定义可知频率具有如下性质:

1° 对任意事件  $A$  有  $f_n(A) \in [0, 1]$ ;

2° 对必然事件  $\Omega$  有  $f_n(\Omega) = 1$ ;

3° 对互不相容的事件  $A_1, A_2, \dots, A_k$  有  $f_n(A_1 \cup A_2 \cup \dots \cup A_k) = f_n(A_1) + f_n(A_2) + \dots + f_n(A_k)$ .

性质 1° 和 2° 根据定义显然成立. 对互不相容的事件  $A_1, A_2, \dots, A_k$ , 并事件  $A_1 \cup A_2 \cup \dots \cup A_k$  发生的频数等于每个事件  $A_i$  发生的频数之和, 由此可知性质 3° 成立.

频率在实际中通常表现出一定的随机性, 例如, 在相同条件下进行两轮  $n$  次试验, 每轮试验中事件  $A$  发生的频率往往不同. 其次, 随着试验次数  $n$  的增加, 事件  $A$  发生的频率  $f_n(A)$  会发生一定的变化, 表现出一定的随机性.

尽管频率表现出一定的随机性, 但经过大量重复的试验, 事件的频率通常在一个确定的常数  $p$  附近摆动, 而且随着试验次数的增大, 摆幅越来越小, 频率也越来越稳定于常数  $p$ , 将这种规律称为 **频率的稳定性**. 例如历史上有多人做过重复投掷硬币的试验, 下表列出了其中一些试验统计结果:

**表 1.2** 历史上多人重复投掷硬币的试验结果

实验者	投掷总数	正面朝上的频数	正面朝上的频率
德摩根	2048	1061	0.5181
蒲丰	4040	2048	0.5069
K. 皮尔逊	12000	6019	0.5016
K. 皮尔逊	24000	12012	0.5005

我们也可以利用计算机产生随机数对投掷硬币的试验进行仿真, 图 1.2 给出了相应的试验结果. 这些研究结果均表明: 尽管对不同的投掷总数, 正面朝上的频率并不相同, 但随着投掷次数的增加, 正面朝上的频率越来越接近常数  $1/2$ , 即频率逐渐稳定于  $1/2$ . 这种频率的稳定性即通常所说的统计规律性, 是随机事件本身所固有的客观属性, 可用于度量事件发生的可能性大小.

**定义 1.4** 随机事件  $A$  在大量重复试验中发生的频率总是稳定地在一个常数  $p$  附近摆动, 且随着试验次数的增加而摆幅逐渐越小, 则称常数  $p$  为事件  $A$  发生的 **概率**, 记为  $P(A) = p$ .



图 1.2 任意投掷硬币, 正面朝上频率的趋势

该定义又称 **概率的统计定义**, 其概率称为 **统计概率**, 提供了计算随机事件概率的一种方法, 即当试验次数足够多时, 可用频率来给出事件概率的近似值.

另一方面, 概率的统计定义存在着数学上的不严谨性, 在实际中也不太可能每一个事件做大量重复试验来计算频率, 以此近似概率. 受到频率的稳定性及其性质的启发, 下面我们给出严谨的概率公理化定义.

### 1.2.2 概率公理化

20 世纪 30 年代, 前苏联数学家柯尔莫哥洛夫 (A. Kolmogorov) 提出了概率论的公理化体系, 通过基本的性质给出了概率的严格定义, 建立可媲美于欧氏几何公理化的理论体系.

**定义 1.5 (概率公理化)** 在可测空间  $(\Omega, \Sigma)$  上, 若函数  $P: \Sigma \rightarrow R$  满足以下条件:

- 1° **非负性**: 对任意  $A \in \Sigma$  有  $P(A) \geq 0$ ;
- 2° **规范性**: 对样本空间  $\Omega$  有  $P(\Omega) = 1$ ;
- 3° **可列可加性**: 若  $A_1, A_2, \dots, A_n, \dots \in \Sigma$  是可列个互不相容的事件, 即  $A_i A_j = \emptyset$  ( $i \neq j$ ), 有
 
$$P(A_1 \cup A_2 \cup \dots \cup A_n \cup \dots) = P(A_1) + P(A_2) + \dots + P(A_n) + \dots,$$

则称  $P(A)$  为随机事件  $A$  的 **概率**, 称  $(\Omega, \Sigma, P)$  为 **概率测度空间** 或 **概率空间** (probability space).

概率  $P(A)$  是定义在可测空间  $(\Omega, \Sigma)$  上的实值函数, 满足非负性、规范性和可列可加性三条公理, 该定义简明扼要地刻画了概率的本质, 为现代概率论奠定了基础, 公理化体系是概率论发展历史上的一个里程碑, 从此概率论被公认为数学的一个分支.

根据概率公理化的定义, 可以推导出很多概率的性质.

**性质 1.1** 对不可能事件  $\emptyset$  有  $P(\emptyset) = 0$ .

**证明** 令  $A_i = \emptyset$  ( $i = 1, 2, \dots$ ), 则有  $\emptyset = \bigcup_{i=1}^{\infty} A_i$  且  $A_i \cap A_j = \emptyset$  ( $i \neq j$ ). 根据公理 3° 有

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

再根据公理 1° 可知  $P(\emptyset) = 0$ .

不可能事件  $\emptyset$  的概率为 0, 但概率为 0 的事件并不一定是不可能事件; 同理, 必然事件  $\Omega$  的概率为 1, 但概率为 1 的事件并不一定是必然事件. 反例参考后面所学的几何概型或连续随机变量.

**性质 1.2 (有限可加性)** 若  $A_1, A_2, \dots, A_n$  是两两不相容的事件, 则

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

**证明** 令  $A_i = \emptyset$  ( $i > n$ ), 则有  $\bigcup_{i=1}^n A_i = \bigcup_{i=1}^{\infty} A_i$ , 且  $A_1, A_2, \dots, A_n, A_{n+1}, \dots$  是两两互不相容事件. 根据公理 3° 可知

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

性质得证.

**性质 1.3** 对任意事件  $A$ , 有  $P(\bar{A}) = 1 - P(A)$ .

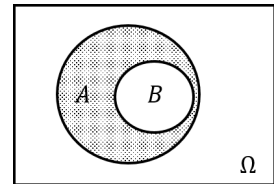
**证明** 由于  $\Omega = \bar{A} \cup A$ , 以及事件  $A$  与  $\bar{A}$  互不相容, 根据有限可加性有  $1 = P(\Omega) = P(A) + P(\bar{A})$ .

**性质 1.4** 若事件  $B \subset A$ , 则有  $P(A - B) = P(A) - P(B)$  和  $P(B) \leq P(A)$ .

**证明** 若  $B \subset A$ , 如右图所示有  $A = B \cup (A - B)$ , 根据定义可知  $B$  与  $A - B$  互不相容. 由有限可加性有

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

再根据公理 1° 有  $P(A - B) = P(A) - P(B) \geq 0$ , 从而得到  $P(A) \geq P(B)$ .

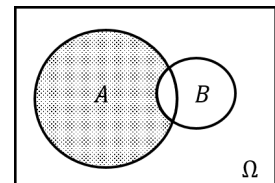


**性质 1.5** 对任意事件  $A$  和  $B$ , 有

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

**证明** 根据  $A = (A - B) \cup (AB)$ , 以及  $A - B$  与  $AB$  互斥, 有

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



再根据  $A \cup B = (A - B) \cup B$ , 以及  $A - B$  与  $B$  互斥, 有

$$P(A - B) = P(A \cup B) - P(B).$$

**性质 1.6 (容斥原理)** 对任意随机事件  $A$  和  $B$  有

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

**证明** 因  $A \cup B = (A - B) \cup (AB) \cup (B - A)$ , 以及  $A - B$ ,  $B - A$ ,  $AB$  两两互不相容, 由有限可加性可知

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

再将  $P(A - B) = P(A) - P(AB)$  和  $P(B - A) = P(B) - P(AB)$  代入上式即可完成证明.

类似地, 对三个随机事件  $A, B, C$  有

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

对  $n$  个随机事件  $A_1, A_2, \dots, A_n$  有

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i A_j) + \sum_{i < j < k} P(A_i A_j A_k) + \cdots + (-1)^{n-1} P(A_1 \cdots A_n).$$

对  $n$  个随机事件  $A_1, A_2, \dots, A_n$  的容斥原理可进一步简写为

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{r=1}^n (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(A_{i_1} \cdots A_{i_r}).$$

**性质 1.7 (Union Bound 或 布尔不等式)** 对事件  $A_1, A_2, \dots, A_n$  有

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n).$$

**证明** 我们利用数学归纳法进行证明. 当  $n = 2$  时, 由容斥原理有

$$P(A \cup B) = P(A) + P(B) - P(AB) \leq P(A) + P(B). \quad (1.1)$$

假设当  $n = k$  时性质成立, 对  $n = k + 1$  有

$$\begin{aligned} P(A_1 \cup \cdots \cup A_{k+1}) &= P((A_1 \cup \cdots \cup A_k) \cup A_{k+1}) \\ &\leq P(A_1 \cup \cdots \cup A_k) + P(A_{k+1}) \leq P(A_1) + \cdots + P(A_k) + P(A_{k+1}), \end{aligned}$$



这里第一个不等式成立是根据式 (1.1), 而第二个不等式成立是根据归纳假设. 完成证明.

根据数学归纳法可类似得到下列不等式:

**推论 1.1 (Bonferroni 不等式)** 对事件  $A_1, A_2, \dots, A_n$  有

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

**例 1.4** 设  $P(A) = p$ ,  $P(B) = q$ ,  $P(AB) = r$ , 用  $p, q, r$  分别表示事件的概率: 1)  $P(\bar{A} \cup \bar{B})$ , 2)  $P(\bar{A}B)$ ; 3)  $P(\bar{A} \cup B)$ ; 4)  $P(\bar{A} \cap \bar{B})$ .

**解** 对问题 1), 根据事件的德摩根律有

$$P(\bar{A} \cup \bar{B}) = P(\overline{AB}) = 1 - r.$$

对问题 2), 根据差事件的定义

$$P(\bar{A}B) = P(B - A) = P(B) - P(AB) = q - r.$$

对问题 3), 根据容斥原理有

$$P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A}B) = 1 - p + q - (q - r) = 1 - p + r.$$

对问题 4), 根据德摩根律与容斥原理有

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 + r - p - q.$$

**例 1.5** 设三个随机事件  $A, B, C$  满足  $P(A) = P(B) = P(C) = 1/4$ ,  $P(AB) = 0$ ,  $P(AC) = P(BC) = 1/16$ , 求事件  $A, B, C$  中至少有一个事件发生的概率.

**解** 首先根据三个事件的容斥原理有

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

$$= 3/4 - 1/8 + P(ABC).$$

根据  $P(AB) = 0$  和  $ABC \subset AB$  可知

$$0 \leq P(ABC) \leq P(AB) = 0$$

由此可知事件  $A, B, C$  中至少有一个事件发生的概率为  $5/8$ .

### 1.3 古典概型与几何概型

本节介绍两种历史较为久远的经典概率模型: 古典概型与几何概型.

#### 1.3.1 古典概型

首先研究一类简单的随机现象, 它是概率论早期最重要的研究对象, 其发展在概率论中具有重要的意义, 并在产品质量抽样检测等问题中具有广泛的应用.

**定义 1.6 (古典概型)** 如果试验  $E$  满足:

- 试验的结果只有有限种可能, 即样本空间  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ , 其中  $\omega_i$  为基本事件,
- 每种结果发生的可能性相同, 即  $P(\{\omega_i\}) = P(\{\omega_j\})$  ( $i \neq j$ ),

则称该类试验称为 **古典概型**, 又称 **等可能概型**.

根据上述定义以及  $P(\Omega) = 1$  可知: 每个基本事件发生的概率为  $P(\{\omega_i\}) = 1/n$ , 若事件  $A$  包含  $k$  个基本事件  $\{\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_k}\}$ , 则事件  $A$  发生的概率为

$$P(A) = k/n = |A|/|\Omega|,$$

这里  $|A|$  表示事件  $A$  包含的事件的个数. 很显然古典概型的概率满足概率公理化体系的三条公理.

计算古典概率的本质是计数 (Counting), 计数是组合学研究的重要内容, 我们将在 1.4 节详细的介绍各种计数方法, 这里仅仅介绍一些基本原理和排列组合:

- **加法原理:** 若一项工作可以用两种不同的过程  $\mathcal{A}_1$  和  $\mathcal{A}_2$  完成, 且过程  $\mathcal{A}_1$  和  $\mathcal{A}_2$  分别有  $n_1$  和  $n_2$  种方法, 则完成该工作有  $n_1 + n_2$  种方法.
- **乘法原理:** 若一项工作需要依次通过  $\mathcal{A}_1$  和  $\mathcal{A}_2$  两过程, 且过程  $\mathcal{A}_1$  和  $\mathcal{A}_2$  分别有  $n_1$  和  $n_2$  种方法, 则完成该工作有  $n_1 \times n_2$  种方法.

上述两条原理可进一步推广到多个过程的情况.

**排列:** 从  $n$  个不同的元素中无放回地取出  $r$  个元素进行排列, 此时既要考虑取出的元素, 也要顾及其排列顺序, 则有  $(n)_r = n(n-1) \cdots (n-r+1)$  种不同的排列. 若  $r = n$  时称全排列, 有  $n!$  种.

**组合:** 从  $n$  个不同的元素中无放回地取出  $r$  个元素, 取出的元素之间无顺序关系, 共有  $\binom{n}{r}$  种不同的取法, 其中

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{(n)_r}{r!}, \quad \text{且记} \quad \binom{n}{0} = 1.$$

这里  $\binom{n}{r}$  称为 **组合数** 或 **二项系数**, 它是二项展开式  $(a+b)^n = \sum_{r=0}^n \binom{n}{r} a^r b^{n-r}$  中项  $a^r b^{n-r}$  的系数.

很多经典的数学问题都可归纳为古典概型, 下面介绍一些典型的例子:

**例 1.6** 将  $n$  个不同的球随机放入  $N$  ( $N \geq n$ ) 个不同的盒子中, 事件  $A$  表示恰有  $n$  个盒子且每盒一球; 事件  $B$  表示指定的  $n$  个盒子中各有一球; 事件  $C$  表示指定一盒子恰有  $m$  个球. 求事件  $A, B, C$  发生的概率. (盒子的容量不限, 放入同一个盒子内的球无顺序排列区别)

**解** 将  $n$  个不同的球随机放入  $N$  个不同的盒子中, 共有  $N^n$  种不同的放法. 而对事件  $A$ , 有  $(N)_n = N!/n!$  种不同的放法, 因此

$$P(A) = \frac{(N)_n}{N^n} = \frac{N!}{N^n n!}.$$

对事件  $B$ , 有  $n!$  种不同的放法, 因此

$$P(B) = \frac{n!}{N^n}.$$

对事件  $C$ , 可分为两步: 第一步在指定的盒子内放入  $m$  个球, 有  $\binom{n}{m}$  种不同的放法; 第二步将剩下的  $n-m$  个球放入  $N-1$  个盒子, 有  $(N-1)^{n-m}$  种不同的放法. 因此

$$P(C) = \frac{\binom{n}{m}(N-1)^{n-m}}{N^n}.$$

生日问题是概率历史上有名的数学问题, 研究某次集会的  $k$  个人中至少有两人生日相同的概率, 或者有  $k$  人的班级中至少两人生日相同的概率.

**例 1.7 (生日问题)** 有  $k$  个人 ( $k < 365$ ), 每个人的生日等可能地出现于 365 天中的任意一天, 求至少两人生日相同的概率.

**解** 用  $A$  表示至少有两人生日相同的事件, 其对立事件  $\bar{A}$  表示任意两人生日均不相同的事件.  $k$  个人的生日共有  $365^k$  种可能, 而  $k$  个人的生日两两互不相同的有  $(365)_k$  种可能. 因此

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{(365)_k}{365^k}.$$

易知当  $k = 30$  时,  $P(A) = 70.6\%$ ; 当  $k = 40$  时,  $P(A) = 89.1\%$ ; 当  $k = 50$  时,  $P(A) = 97\%$ ; 当  $k = 60$  时,  $P(A) = 99.4\%$ ; 当  $k = 100$  时,  $P(A) = 99.99\%$ .

下面介绍古典概型计算中一类典型问题, 在产品质量检测等方面广泛应用.

**例 1.8** 设一批  $N$  件产品中  $M$  件次品, 现从  $N$  件产品中不放回地任选  $n$  件, 求其中恰有  $k$  件次品的概率.

**解** 用  $A$  表示恰有  $k$  件次品的事件. 从  $N$  件产品中任选  $n$  件, 有  $\binom{N}{n}$  种不同的选法; 在所选取的  $n$  件产品中, 有  $k$  件次品以及  $n-k$  件正品, 即从  $M$  件次品中选出  $k$  件次品, 从  $N-M$  件正品中选出  $n-k$  件正品, 因此有  $\binom{M}{k}\binom{N-M}{n-k}$  种不同的取法. 由此可得

$$P(A) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad (1.2)$$

该概率  $P(A)$  被称为 **超几何概率**.

在例 1.8 中若为有放回地任选  $n$  件, 则每次抽到一件非次品的概率为  $(N-M)/N$ , 抽到一件次品的概率为  $M/N$ , 因此  $n$  件中恰有  $k$  件次品的概率为

$$\binom{n}{k} \left(\frac{M}{N}\right)^k \left(\frac{N-M}{N}\right)^{n-k}.$$

抽签是人们引入随机性的一个简单例子, 广泛应用于各种体育赛事或日常生活中. 关于抽签的公平性, 即抽签结果虽然不同但出现这种结果的可能性相同, 需要通过计算概率来进行验证:

**例 1.9 (抽签问题)** 袋中有  $a$  个不同的白球,  $b$  个不同的红球, 假设有  $k$  个人依次随机无放回地从袋中取一个球, 问第  $i$  个人 ( $i \leq k$ ) 取出红球的概率是多少?

**解** 用  $A$  表示第  $i$  个人取到红球的事件. 若  $k$  个人依次随机无放回地从袋中取一个球, 则有  $(a+b)_k$  种不同的取法. 若事件  $A$  发生, 第  $i$  个人取到红球, 它可能是  $b$  个红球中的任意一个, 有  $b$  种取法; 其它剩余的  $k-1$  个球可以从  $a+b-1$  个球中取出, 有  $(a+b-1)_{k-1}$  种不同的取法. 因此事件  $A$  的概率为

$$P(A) = \frac{b(a+b-1)_{k-1}}{(a+b)_k} = \frac{b}{a+b}.$$

由此可知第  $i$  个人取到红球的概率为  $b/(a+b)$ , 与  $i$  的大小无关, 即抽签先后顺序对抽签的结果没有影响, 由此证明了抽签的公平性.

**例 1.10 (匹配问题)** 有  $n$  对夫妻参加一次聚会, 现将所有参会人员任意分成  $n$  组, 每组一男一女, 问至少有一对夫妻被分到同一组的概率是多少?

**解** 用  $A$  表示至少有一对夫妻被分到同一组的事件, 以及  $A_i$  表示第  $i$  对夫妻 ( $i \in [n]$ ) 被分到同一组的事件, 于是有  $A = A_1 \cup A_2 \cup \cdots \cup A_n$ . 根据容斥原理有

$$P(A) = P\left(\bigcup_{i=1}^n A_i\right) = \sum_{r=1}^n (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(A_{i_1} \cdots A_{i_r}).$$

对任意  $r \in [n]$ , 考虑事件  $A_{i_1} \cdots A_{i_r}$  概率, 若参会人员任意分成  $n$  组且每组一男一女, 共有  $n!$  种不同的分法, 若将第  $i_1, i_2, \dots, i_r$  对夫妻分别分组, 则有  $(n-r)!$  种不同的分法. 根据等可能性原则有

$$P(A_{i_1} \cdots A_{i_r}) = \frac{(n-r)!}{n!}.$$

而和式  $\sum_{i_1 < \dots < i_r} P(A_{i_1} \cdots A_{i_r})$  中共有  $\binom{n}{r}$  项, 由此可得

$$\sum_{i_1 < \dots < i_r} P(A_{i_1} \cdots A_{i_r}) = \binom{n}{r} \frac{(n-r)!}{n!} = \frac{1}{r!},$$

于是事件  $A$  发生的概率

$$P(A) = 1 - \frac{1}{2!} + \frac{1}{3!} + \cdots + (-1)^{n+1} \frac{1}{n!}.$$

当  $n$  较大时, 利用泰勒展式  $e^x = 1 + x + x^2/2! + \cdots + x^n/n! + \cdots$  以及令  $x = -1$  有

$$e^{-1} = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} + \cdots \approx \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!},$$

由此近似有  $P(A) = 1 - 1/e = 0.632$ .

在概率计算的过程中, 有时可适当利用概率的性质来简化计算, 例如,

**例 1.11** 从  $\{1, 2, \dots, 9\}$  数中有放回取  $n$  个, 试求取出  $n$  个数的乘积被 10 整除的概率.

**解** 令  $A = \{\text{取出 } n \text{ 个整数的乘积能被 } 10 \text{ 整除}\}$ ,  $B = \{\text{取出的 } n \text{ 个数中有偶数}\}$ ,  $C = \{\text{取出的 } n \text{ 个数中至少有一个 } 5\}$ , 于是有  $A = BC$ . 直接计算事件  $B$  发生的概率较难, 我们因此考虑  $B$  的对立事件的概率

$$P(\bar{B}) = P(\{\text{取出的 } n \text{ 个数中无偶数}\}) = P(\{\text{取出的 } n \text{ 个数只包括 } 1, 3, 5, 7, 9\}) = 5^n/9^n.$$

同理可得

$$P(\bar{C}) = 8^n/9^n \quad \text{和} \quad P(\bar{B}\bar{C}) = 4^n/9^n.$$

根据概率的性质有

$$P(A) = 1 - P(\overline{BC}) = 1 - P(\bar{B} \cup \bar{C}) = 1 - P(\bar{B}) - P(\bar{C}) + P(\bar{B}\bar{C}) = 1 - \frac{5^n}{9^n} - \frac{8^n}{9^n} + \frac{4^n}{9^n}.$$

### 1.3.2 几何概型

古典概型考虑有限的样本空间, 即有限个等可能的基本事件, 然而在很多实际应用中受到了限制. 本节研究可能有无限多种结果的随机现象, 具有如下两个特点:

- **样本空间无限可测** 样本空间包含无限不可列个样本点, 但可以用几何图形 (如一维线段、二位平面区域、或三维空间区域等) 来表示, 其相应的几何测度 (如长度、面积、体积等) 是一个非零有限的实数,
- **基本事件等可能性** 每个基本事件发生的可能性大小相等, 从而使得每个事件发生的概率与该事件的几何测度相关, 与具体位置无关,

称为 **几何概型**. 其形式化定义如下:

**定义 1.7** 在一个测度有限的区域  $\Omega$  内等可能性投点, 落入  $\Omega$  内的任意子区域  $A$  的可能性与  $A$  的测度成正比, 与  $A$  的位置与形状无关, 这样的概率模型称之为**几何概型**. 事件  $A$  发生的概率为

$$P(A) = \frac{A \text{ 的测度}}{\Omega \text{ 的测度}} = \frac{\mu(A)}{\mu(\Omega)}.$$

根据上述定义可验证几何概型的概率满足三条公理. 下面给出几何概型的案例.

**例 1.12** 将一根长度为  $l$  的木棍随意折成三段, 这三段能构成平面三角形的概率是多少?

**解** 在此例中将一根木棍折成三段有无穷种可能, 根据其随意性任何一种折法的可能性大小相等, 且木棍的长度可度量, 由此采用几何概型. 用  $x, y$  分别表示第一段、第二段木棍的长度, 第三段的长度为  $l - x - y$ , 由此可得样本空间

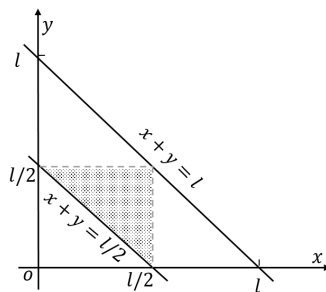
$$\Omega = \{(x, y): x > 0, y > 0, l - x - y > 0\}.$$

用  $A$  表示折成的三段能构成平面三角形的事件, 而构成平面三角形的条件是任意两边之和大于第三边, 由此可得

$$\begin{aligned} A &= \{(x, y): x + y > l - x - y, l - y > x, l - x > y\} \\ &= \{(x, y): x + y > l/2, y < l/2, x < l/2\}. \end{aligned}$$

如右图所示, 计算事件  $A$  发生的概率为

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{(l/2)^2/2}{l^2/2} = \frac{1}{4}.$$



**例 1.13** 假设一乘客到达汽车站的时间是任意的, 客车间隔一段时间发班, 请规划最长的间隔发车时间, 才能确保乘客候车等待时间不超过 20 分钟的概率大于 80%.

**解** 设客车的间隔时间为  $l$  ( $l > 20$ ), 选择特定的连续的  $l$  分钟为样本空间, 则乘客到达时间的样本空间为  $\Omega = \{x: 0 < x \leq l\}$ . 用  $B$  表示乘客的等待时间超过 20 分钟的事件, 而事件  $B$  发生则可知乘客到达车站的时间在 0 与  $l - 20$  之间, 即

$$B = \{x: 0 < x < l - 20\}.$$

可知事件  $B$  发生的概率小于或等于 20%, 即

$$P(B) = \frac{l - 20}{l} \leq 0.2,$$

求解可得  $l \leq 25$ .

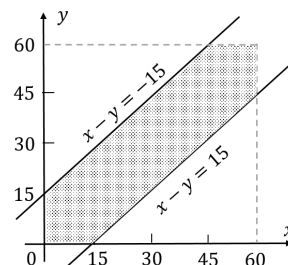
**例 1.14 (会面问题)** 两银行经理约定中午 12:00 – 13:00 到某地会面, 两人到达时间随机, 先到者等另一人 15 分钟后离开, 求两人见面的概率.

**解** 用  $x, y$  分别表示两人的到达时间 (分钟), 则样本空间  $\Omega = \{(x, y) | 0 \leq x, y \leq 60\}$ . 用  $A$  表示两人见面的事件, 则

$$A = \{(x, y) | |x - y| \leq 15\} = \{(x, y) | x - y \leq 15 \text{ 且 } x - y \geq -15\}.$$

根据右图计算事件  $A$  发生的概率

$$P(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{60^2 - 45^2}{60^2} = \frac{7}{16}.$$



进一步思考: 若两银行经理非常聪明且都非常希望能促成此次见面, 但没有通讯方式进行联系, 能否找出一些策略来解决会面问题?

很多几何概型的概率可通过计算机模拟仿真来近似计算, 即 **统计模拟法** 或 **蒙特卡洛 (Monte Carlo) 法**. 先构造相应的概率模型, 再进行计算机模拟试验, 用统计的方法计算其估计值, 作为所求问题的近似值. 例如, 可利用蒙特卡洛法来近似计算例 1.14 的概率, 伪代码如下:

```

输入参数: 试验总次数  $N$ .                %% 取较大正整数  $N$ , 更能精确计算两人见面的概率
初始化: 事件  $A$  最初发生的次数  $n_A \leftarrow 0$ .    %% 此处事件  $A$  表示两人见面的事件
For  $i = 1 : N$ 
     $x \leftarrow \text{Random}(0, 60), y \leftarrow \text{Random}(0, 60)$ .    %% 在区间  $(0, 60)$  以随机任意选取两个数
    If  $|x - y| \leq 15$  then
         $n_A \leftarrow n_A + 1$ .                %% 若两人见面则频数+1
    End
End
输出概率:  $n_A/N$ .

```

接下来介绍几何概型的一个经典问题, 由法国科学家蒲丰于 1777 年提出.

**例 1.15 (投针问题)** 平面上有两条平行线, 相距为  $a$ , 向此平面任投一长度为  $l$  ( $l < a$ ) 的针, 求此针与任一平行线相交的概率.

**解** 用  $x$  表示针的中点到最近的一条平行线的距离, 用  $\theta$  表示针与平行线的夹角. 针与平行线的位置关系图 1.3 所示.

容易知道  $x \in [0, a/2]$  和  $\theta \in [0, \pi]$ , 以  $\Omega$  表示边长分别为  $a/2$  和  $\pi$  的长方形, 用  $A$  表示针与平行线相交的事件, 若事件  $A$  发生则必有  $x \leq l \sin(\theta)/2$  成立, 由此得到事件  $A$  的发生如图 1.3 中阴影部分所示. 求解可得

$$P(A) = \frac{A \text{ 的面积}}{\Omega \text{ 的面积}} = \frac{\int_0^\pi l \sin(\theta)/2 d\theta}{a\pi/2} = \frac{2l}{a\pi}. \quad (1.3)$$



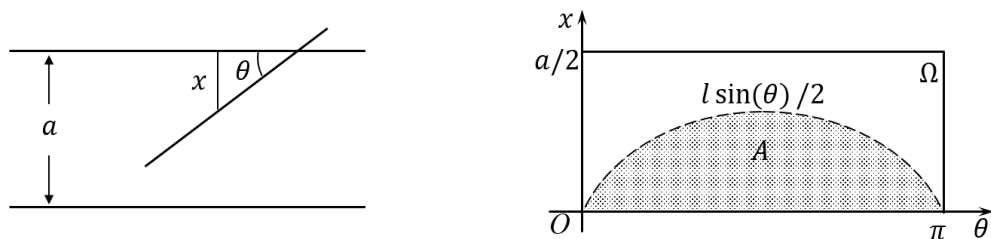


图 1.3 蒲丰投针问题

在上例中, 事件  $A$  发生的概率包含圆周率  $\pi$ , 由此蒲丰设想出计算  $\pi$  的概率近似方法, 通过频率来近似计算事件  $A$  发生的概率, 再根据 (1.3) 计算圆周率  $\pi$ , 即

$$\frac{n_A}{n} \approx P(A) = \frac{2l}{a\pi} \quad \Rightarrow \quad \pi \approx \frac{2ln}{an_A}.$$

这里  $n$  表示试验的总次数, 而  $n_A$  表示事件  $A$  发生的次数. 历史上有多人根据蒲丰的设想还真做了试验来近似计算圆周率  $\pi$ , 有兴趣的读者可查找相关文献.

在二十世纪之前, 很多人都相信只要找到合适的等可能性描述, 概率是可以被唯一定义的. 然而贝特朗 (Bertrand) 对这种观点提出了质疑, 他通过下面的一个具体例子说明: 几何型等的等可能性概率存在多种看似合理但相互矛盾的结果.

**例 1.16 (贝特朗奇论)** 在半径为 1 的圆内随机地取一条弦, 求其弦长超过该圆内接等边三角形边长  $\sqrt{3}$  的概率.

**解** 对于“等可能性”或“随机性”含义的不同解释, 这个问题存在着多种不同答案的解决方法, 下面给出三种不同的方法:

- (1) 任何与圆相交的弦一般有两个交点, 不妨在圆上先固定其中一点, 以此点为顶点作一个等边三角形, 只有落入此三角形内的弦才满足弦长超过  $\sqrt{3}$ . 这种弦的另一端跑过的弧长为整个圆周的  $1/3$ , 故所求概率等于  $1/3$  (如图 1.4-a).
- (2) 弦长只跟到圆心的距离有关, 与方向无关, 因此可以假定它垂直于某一条直径. 当且仅当它与圆心的距离小于  $1/2$  时, 其弦长才会大于  $1/3$ , 因此此时所求的概率为  $1/2$  (如图 1.4-b).
- (3) 弦可被弦的中心点唯一确定, 当且仅当弦的中心点位于半径为  $1/2$  的同心圆内时, 弦长才会大于  $1/3$ . 半径为  $1/2$  的小圆面积为  $1/4$ , 大圆面积为 1, 故所求概率等于  $1/4$  (如图 1.4-c).

同一问题却有三种不同的答案, 其根本原因在于取弦时采用不同的等可能性假定. 第一种方法假定端点在圆周上均匀分布, 第二种方法假定弦的中点在直径上均匀分布, 第三种方法假定弦的中心点在小圆内均匀分布. 这三种方法采用三种不同的随机试验, 对于各自的随机试验而言, 它们都是正确的, 因此在概率的计算中一定要明确具体的试验.



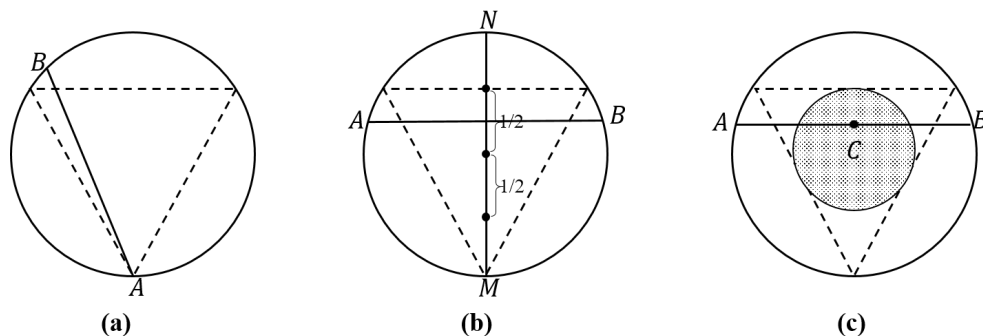


图 1.4 贝特朗奇论

贝特朗奇论在现代概率的发展中起到过重要作用, 上述例子由贝特朗于 1899 年提出, 以此反驳了“等可能性可完全定义概率”的观点. 从此概率论开始向公理化方面发展, 从应满足的基本性质来定义概率, 而不是某些具体事件的概率. 正因为如此, 希尔伯特于 1900 年在巴黎举行的第二届数学家大会上提出了著名的 23 个数学问题, 其中第六个问题就是概率公理化.

与古典概型一样, 几何概型的研究有助于发现概率的一些基本性质, 有助于对某些概率问题的直观理解和具体计算.

## 1.4 组合计数\*

组合计数研究满足一定条件的计数对象的数目, 概率论中的很多问题都可以通过计算一个事件发生的数目来解决, 如古典概型. 此外, 组合计数在人工智能、计算机等领域具有广泛的应用, 本节将介绍经典的组合计数: 十二重计数 (The twelvefold way).

十二重计数由著名的组合学家 G.-C. Rota (1932-1999) 提出, 最初的问题表述为研究一定条件下两个集合之间映射的数目. 为了问题的可理解性, 我们采用《计算机程序设计艺术》中的表述: 将  $n$  个不同或相同的球, 放入  $m$  个不同或相同的箱子, 在无任何限制、或每个箱子至多或至少放一个球的条件下, 研究在这十二种情形下分别有多少种不同的方法数. 我们首先给出十二重计数的结果, 如表 1.3 所示, 相关知识和符号说明将在后续小节逐一介绍.

表 1.3 十二重计数.

$n$ 个球	$m$ 个箱子	无任何限制	每个箱子至多放一球	每个箱子至少放一球
不同	不同	$m^n$	$(m)_n$	$m!S(n, m)$
相同	不同	$\binom{n+m-1}{n}$	$\binom{m}{n}$	$\binom{n-1}{m-1}$
不同	相同	$\sum_{k=1}^m S(n, k)$	$\begin{cases} 1 & n \leq m \\ 0 & n > m \end{cases}$	$S(n, m)$
相同	相同	$\sum_{k=1}^m p(n, k)$	$\begin{cases} 1 & n \leq m \\ 0 & n > m \end{cases}$	$p(n, m)$

### 1.4.1 排列、环排列、组合与多重组合

前面介绍了排列, 即从  $n$  个不同的元素中取出  $r$  个元素进行排列, 需考虑取出的元素以及其排列顺序, 有  $(n)_r = n(n-1)\cdots(n-r+1)$  种不同的排列方法. 若  $r = n$  时称全排列, 有  $n!$  种方法.

若从  $n$  个不同的元素中取出  $r$  个元素排成一个圆环, 称为 **环排列**.

如右图所示, 从顺时针看 a-b-c-a, b-c-a-b 和 c-a-b-c 是同一个环排列, 但 a-c-b-a 则不是. 因此从  $n$  个不同的元素中取出  $r$  个元素进行环排列, 每一个环排列对应于  $r$  种不同的直线排列, 而且不同的环排列对应的直线排列互不相同. 因此有

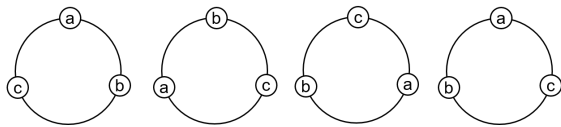


图 1.5 环排列.

**定义 1.8** 若从  $n$  个不同的元素中取出  $r$  个元素排成一个圆环, 有  $(n)_r/r$  种不同的排法, 称  $(n)_r/r$  为 **环排列数**. 特别地,  $n$  个不同元素的环排列数为  $(n-1)!$ .

**例 1.17** 将  $n$  对夫妇任意安排在一张圆桌, 求任何一对夫妻都被安排坐在一起的概率.

**解** 用  $\Omega$  表示将  $n$  对夫妇任意安排在一张圆桌时所有可能的环排列, 以及用  $A$  表示任何一对夫妻都被安排坐在一起的事件. 则根据环排列可知

$$|\Omega| = (2n-1)! \quad \text{和} \quad |A| = 2^n(n-1)!.$$

由此可得任何一对夫妻都被安排坐在一起的概率  $2^n(n-1)!/(2n-1)!$ .

前面介绍了组合数, 即从  $n$  不同的元素中选取  $r$  个元素, 取出的元素之间无顺序关系, 有  $\binom{n}{r}$  种不同的方法, 称为组合数. 现将组合数的概念进行推广到多重组合数.

**定义 1.9** 将  $n$  个不同的元素分成  $k$  组, 每组分别有  $r_1, r_2, \dots, r_k$  个元素, 组内元素无顺序关系, 即满足  $n = r_1 + \dots + r_k$  且  $r_1, r_2, \dots, r_k$  为正整数, 则有

$$\binom{n}{r_1, r_2, \dots, r_k} = \binom{n}{r_1} \binom{n-r_1}{r_2} \cdots \binom{n-r_1-r_2-\cdots-r_{k-1}}{r_k} = \frac{n!}{r_1! r_2! \cdots r_k!}$$

种不同的方法, 称  $\binom{n}{r_1, r_2, \dots, r_k}$  为 **多重组合数**.

对于一个  $n$  次多项式有:

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{n=r_1+r_2+\cdots+r_k} \binom{n}{r_1, r_2, \dots, r_k} x_1^{r_1} x_2^{r_2} \cdots x_k^{r_k},$$

因此多重组合数又被称为多项式系数.

根据定义可知组合数本质上属于多重组合数, 即  $\binom{n}{r} = \binom{n}{r, n-r}$ .

以前研究集合的元素都是互不相同的, 我们引入多重集的概念.

**定义 1.10** 若集合中的元素是可以重复的, 且重复的元素是完全相同、不可分辨的, 则称该集合为 **多重集**. 例如多重集  $A = \{1, 1, 1, 2, 2, 2, 3, 3, 4\}$ .

假设多重集  $A$  有  $k$  类不同的元素, 每类元素的个数分别为  $r_1, r_2, \dots, r_k$ , 即  $n = r_1 + r_2 + \dots + r_k$ . 若将此多重集  $A$  中的所有元素排列成一排, 则相当于从  $n$  个位置中选取  $r_1$  个位置放第一类元素, 再从剩下的从  $n - r_1$  个位置中选取  $r_2$  个位置放第二类元素,  $\dots$ , 从最后  $r_k$  个位置放第  $k$  类元素. 因此该多重集  $A$  有

$$\binom{n}{r_1, r_2, \dots, r_k}$$

种不同的排列方法, 即多重组合数.

根据排列组合数有

$n$ 个球	$m$ 个箱子	无任何限制	每个箱子至多放一球
不同	不同	$m^n$	$(m)_n$
相同	不同		$\binom{m}{n}$

#### 1.4.2 整数的有序分解

本节研究将  $n$  个完全相同、不可分辨的球放入  $m$  个不同的箱子, 有多少种不同的方法数. 鉴于球完全相同且不可分辨, 可以对问题进行转化: 假设第一个箱子有  $x_1$  个球, 第二个箱子有  $x_2$  个球,  $\dots$ , 第  $m$  个箱子有  $x_m$  个球, 这里  $x_1, x_2, \dots, x_m$  是非负的整数, 并满足

$$x_1 + x_2 + \dots + x_m = n.$$

因此将  $n$  个相同的球放入  $m$  个不同的箱子等价于上述方程的非负整数解, 有如下定理:

**定理 1.1** 方程  $x_1 + x_2 + \dots + x_m = n$  有  $\binom{n+m-1}{m-1} = \binom{n+m-1}{n}$  种不同的非负整数解.

**证明** 这里将通过构造一一对应关系给出组合证明. 将  $n$  个相同的球对应于  $n$  个圈 ‘ $\circ$ ’, 将  $m$  个箱子与  $m$  条竖线 ‘ $|$ ’ 进行关联. 现将  $n$  个圆圈和  $m - 1$  条竖线排列成一行, 最后在排列末尾再加入一条竖线, 如下所示:

$$\underbrace{\circ \circ \circ \circ}_{x_1} | | \dots | \underbrace{\circ \circ \circ \circ}_{x_i} | \dots | \underbrace{\circ \circ}_{x_m} |.$$

从左向右看, 用  $x_1$  表示第一条竖线之前圆圈的个数, 用  $x_i$  表示第  $i$  条竖线与第  $i - 1$  条竖线之间圆圈的个数 ( $2 \leq i \leq m$ ). 由此可知方程  $x_1 + x_2 + \dots + x_m = n$  的非负整数解与上述的排列之间存在一一对应关系, 而这种排列有

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

种不同的方法, 即为所求方程  $x_1 + x_2 + \dots + x_m = n$  非负整数解的个数.

例如, 方程  $x_1 + x_2 + x_3 = 10$  有  $\binom{12}{2} = 66$  种不同的非负整数解, 因此将 10 个相同的球放入 3 个不同的箱子有 66 种不同的放法.

定理 1.1 给出了方程  $x_1 + x_2 + \cdots + x_m = n$  非负整数解的个数, 根据该定理可以进一步研究该方程的正整数解的个数, 以及不等式  $x_1 + x_2 + \cdots + x_m < n$  非负整数解或正整数解的个数. 例如,

**推论 1.2** 方程  $x_1 + x_2 + \cdots + x_m = n$  ( $m \leq n$ ) 有  $\binom{n-1}{m-1} = \binom{n-1}{n-m}$  种不同的正整数解.

**解** 引入新变量  $x'_1 = x_1 - 1, x'_2 = x_2 - 1, \cdots, x'_m = x_m - 1$ , 则方程  $x_1 + x_2 + \cdots + x_m = n$  的正整数解等价于方程

$$x'_1 + x'_2 + \cdots + x'_m = n - m$$

的非负整数解. 根据定理 1.1 可知上述方程有

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

种不同的正整数解.

**例 1.18** 求多项式  $(x_1 + x_2 + \cdots + x_m)^n$  的展开式中有多少种不同的展开项.

**解** 根据多项式的展开式有

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{r_1, r_2, \dots, r_m \text{ 非负整数且和为 } n} \binom{n}{r_1, r_2, \dots, r_m} x_1^{r_1} x_2^{r_2} \cdots x_m^{r_m},$$

不同的展开项意味着各个变量不同的多项式次数, 此时与方程  $x_1 + x_2 + \cdots + x_m = n$  的非负整数解建立一一对应关系, 因此多项式  $(x_1 + x_2 + \cdots + x_m)^n$  有  $\binom{n+m-1}{m-1}$  种不同的展开项.

根据整数的有序分解有:

$n$ 个球	$m$ 个箱子	无任何限制	每个箱子至少放一球
相同	不同	$\binom{n+m-1}{m-1}$	$\binom{n-1}{m-1}$

#### 1.4.3 第二类 Stirling 数 (The Stirling number of the second kind)

本节研究将  $n$  个不同的球放入  $m$  个相同的箱子, 有多少种不同的放法, 这里箱子完全相同不可分辨, 可以通过箱子里放置的不同的球加以区分. 该问题在组合学中有另一种表述: 将  $n$  个不同的元素分成  $m$  个非空子集 (block) 的划分数, 即第二类 Stirling 数:

**定义 1.11** 将  $n$  个不同的元素分成  $m$  个非空子集的划分数, 称为 **第二类 Stirling 数**, 记为  $S(n, m)$ .

例如考虑三个不同的元素  $\{1, 2, 3\}$ , 分成  $m = 1, 2, 3$  个非空的子集, 不同的划分情况如下:

- 若分成  $m = 1$  个非空的子集, 则有  $\{1, 2, 3\}$ , 因此  $S(3, 1) = 1$ ;
- 若分成  $m = 2$  个非空的子集, 则有  $\{\{1\}, \{2, 3\}\}, \{\{2\}, \{1, 3\}\}, \{\{3\}, \{1, 2\}\}$ , 因此  $S(3, 2) = 3$ ;

- 若分成  $m = 3$  个非空的子集, 则有  $\{\{1\}, \{2\}, \{3\}\}$ , 因此  $S(3, 3) = 1$ .

根据第二类 Stirling 数的定义可知, 当  $n \geq 1$  时有

$$S(n, n) = 1, \quad S(n, 1) = 1, \quad S(n, 0) = 0.$$

当  $m > n \geq 1$  时有  $S(n, m) = 0$ . 按惯例设  $S(0, 0) = 1$ . 对第二类 Stirling 数有如下递推关系:

**定理 1.2** 对  $n \geq m \geq 1$  有

$$S(n, m) = mS(n-1, m) + S(n-1, m-1).$$

**证明** 根据定义可知将  $\{1, 2, \dots, n\}$  划分成  $m$  个非空的子集, 有  $S(n, m)$  种不同的划分数. 将这些不同的划分可分成两种情况考虑:

- 若元素  $n$  被划分为单独的子集  $\{n\}$ , 则其它剩余的元素被划分成  $m-1$  个非空的子集, 此时有  $S(n-1, m-1)$  种不同的划分数;
- 若元素  $n$  未被划分为单独的子集, 其它剩余元素被划分成  $m$  个非空的子集, 有  $S(n-1, m)$  种不同的划分数; 再将元素  $n$  放入已经划分好的  $m$  个子集之一, 共  $mS(n-1, m)$  种划分数.

由此完成证明.

根据上面的递推关系, 利用归纳法证明可得

**推论 1.3** 第二类 Stirling 数满足

$$S(n, m) = \frac{1}{m!} \sum_{i=0}^m (-1)^i \binom{m}{i} (m-i)^n \quad \text{和} \quad \sum_{m=1}^n S(n, m) (x)_m = x^n,$$

这里  $(x)_m = x(x-1)\cdots(x-m+1)$ .

根据第二类 Stirling 数有

$n$ 个球	$m$ 个箱子	无任何限制	每个箱子至多放一球	每个箱子至少放一球
不同	不同			$m!S(n, m)$
不同	相同	$\sum_{k=1}^m S(n, k)$	$\begin{cases} 1 & n \leq m \\ 0 & n > m \end{cases}$	$S(n, m)$

#### 1.4.4 正整数的无序分拆 (Partition)

本节研究将  $n$  个相同的球放入  $m$  个相同的箱子, 球与箱子都是完全相同、不可分辨的, 只能通过箱子内不同的球的个数加以区别. 该问题在组合学中有另一种表述: 将正整数  $n$  划分成  $m$  个无序的正整数之和, 即 **正整数的无序分拆**.

**定义 1.12** 将正整数  $n$  划分成  $m$  个无序的正整数之和, 有多少种不同的划分数记为  $p(n, m)$ .

例如考虑正整数 7 的无序划分, 相关分拆和划分数  $p(n, m)$  如下表:

$m = 1$	7	$p(7, 1) = 1$
$m = 2$	6 + 1, 5 + 2, 4 + 3	$p(7, 2) = 3$
$m = 3$	5 + 1 + 1, 4 + 2 + 1, 3 + 3 + 1, 3 + 2 + 2	$p(7, 3) = 4$
$m = 4$	4 + 1 + 1 + 1, 3 + 2 + 1 + 1, 2 + 2 + 2 + 1	$p(7, 4) = 3$
$m = 5$	3 + 1 + 1 + 1 + 1, 2 + 2 + 1 + 1 + 1	$p(7, 5) = 2$
$m = 6$	2 + 1 + 1 + 1 + 1 + 1	$p(7, 6) = 1$
$m = 7$	1 + 1 + 1 + 1 + 1 + 1 + 1	$p(7, 7) = 1$

通过上面的观察发现, 将正整数  $n$  划分成  $m$  个无序的正整数, 等价于下面方程的解

$$x_1 + x_2 + \cdots + x_m = n \quad \text{s.t.} \quad x_1 \geq x_2 \geq \cdots \geq x_m \geq 1.$$

根据定义可知, 当  $n \geq 1$  时有

$$p(n, n) = 1, \quad p(n, 1) = 1, \quad p(n, 0) = 0,$$

当  $m > n \geq 1$  时有  $p(n, m) = 0$ , 按惯例设  $p(0, 0) = 1$ . 对  $p(n, m)$  有如下递推关系:

**定理 1.3** 对  $n \geq m \geq 1$  有

$$p(n, m) = p(n - 1, m - 1) + p(n - m, m) \quad \text{和} \quad p(n, m) = \sum_{i=1}^m p(n - m, i).$$

**证明** 将正整数  $n$  划分成  $m$  个无序的正整数之和, 有  $p(n, m)$  种不同的划分方法. 针对任意一种划分  $x_1 + x_2 + \cdots + x_m = n$  ( $x_1 \geq x_2 \geq \cdots \geq x_m \geq 1$ ), 可以考虑两种情况:

- 若最小部分  $x_m = 1$ , 则  $x_1 + x_2 + \cdots + x_{m-1} = n - 1$  是整数  $n - 1$  的  $m - 1$  部分的无序划分, 有  $p(n - 1, m - 1)$  种不同的划分数;
- 若最小部分  $x_m > 1$ , 则  $x_1 - 1 + x_2 - 1 + \cdots + x_m - 1 = n - m$  是整数  $n - m$  的  $m$  部分的无序划分, 有  $p(n - m, m)$  种不同的划分数.

由此证明  $p(n, m) = p(n - 1, m - 1) + p(n - m, m)$ .

对第二个等式的证明, 考虑任何一种划分  $x_1 + x_2 + \cdots + x_m = n$  ( $x_1 \geq x_2 \geq \cdots \geq x_m \geq 1$ ), 设  $y_j = x_j - 1$ , 则有

$$y_1 + y_2 + \cdots + y_m = n - m \quad \text{s.t.} \quad y_1 \geq y_2 \geq \cdots \geq y_m \geq 0.$$

考虑  $y_1, y_2, \dots, y_m$  非零元的个数, 假设恰好有  $i$  个非零元, 则有  $p(n-m, i)$  种不同的解, 由此证明

$$p(n, m) = \sum_{i=1}^m p(n-m, i).$$

下面给出了  $p(n, m)$  的有效估计, 相关证明超出了本书的范围.

**定理 1.4** 对整数  $n \geq m \geq 1$  有

$$\frac{1}{m!} \binom{n-1}{m-1} \leq p(n, m) \leq \frac{1}{m!} \binom{n-1+m(m-1)/2}{m-1}.$$

给定整数  $m \geq 1$ , 当  $n$  非常大或  $n \rightarrow \infty$  有

$$p(n, m) \approx \frac{n^{m-1}}{m!(m-1)!}.$$

根据正整数的无序分拆有

$n$ 个球	$m$ 个箱子	无任何限制	每个箱子至多放一球	每个箱子至少放一球
相同	相同	$\sum_{k=1}^m p(n, k)$	$\begin{cases} 1 & n \leq m \\ 0 & n > m \end{cases}$	$p(n, m)$

## 习题

1.1 简述: 频率与概率的关系, 随机现象中的二重性, 对立与互不相容事件的关系.

1.2 i) 对任意事件  $A$  和  $B$ , 简化  $(A - AB) \cup B$  和  $\overline{(\bar{A} \cup B)}$ ;

ii) 若事件  $A, B, C$  两两互不相容, 简化  $(A \cup B) - C$ .

1.3 班级有  $n$  个同学参加考试, 用  $A_i$  表示第  $i$  个同学通过考试的事件, 用他们表示以下事件:

i) 只有第一位同学未通过考试;      ii) 至少有一位同学未通过考试;

iii) 恰好有一位同学未通过考试;      iv) 至少有两位同学未通过考试;

v) 至多有两位同学未通过考试;      vi) 所有同学通过了考试.

1.4 证明  $n$  个事件的德摩根律, 即对任意  $n$  个事件  $A_1, A_2, \dots, A_n$  有

$$\overline{\bigcup_{i=1}^n A_i} = \bigcap_{i=1}^n \bar{A}_i \quad \text{和} \quad \overline{\bigcap_{i=1}^n A_i} = \bigcup_{i=1}^n \bar{A}_i.$$

1.5 已知事件  $A, B, C$  满足  $P(A) = 1/3$ ,  $P(B) = 1/5$ ,  $P(C) = 1/6$ ,  $P(AB) = 1/20$ ,  $P(AC) = 1/20$ ,  $P(BC) = 1/60$  和  $P(ABC) = 1/100$ , 求事件  $\bar{A}B$ ,  $\bar{A} \cup \bar{B}$ ,  $A \cup B \cup C$ ,  $\bar{A}\bar{B}\bar{C}$ ,  $\bar{A}\bar{B}C$  和  $(\bar{A}\bar{B}) \cup C$  的概率.

1.6 若事件  $A, B$  的概率分别为  $P(A) = 0.6$  和  $P(B) = 0.9$ , 求  $P(AB)$  的最大值和最小值, 并说明在怎样的情形下取得.

1.7 若事件  $A$  和  $B$  满足  $P(AB) = P(\bar{A}\bar{B})$  且概率  $P(B) = 1/4$ , 求概率  $P(A)$ .

1.8 若事件  $A$  和  $B$  满足  $P(A) = 0.1$  和  $P(\bar{A}\bar{B}) = 0.7$ , 求概率  $P(B - A)$ .

1.9 证明: 对任意  $n$  个事件  $A_1, A_2, \dots, A_n$  有

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i A_j) + \sum_{i < j < k} P(A_i A_j A_k) + \dots + (-1)^{n-1} P(A_1, A_2, \dots, A_n).$$

1.10 已知 16 件产品中有 4 件是次品, 不放回地任取两次, 每次任取一件产品, 求事件的概率: i) 两件均是次品; ii) 一件正品和一件次品; iii) 第二次取出正品.

1.11 将  $n$  个男生和两个女生任意排成一列, 两女生间恰有  $k$  个男生 ( $2 < k < n$ ) 的概率是多少.

1.12 将  $n$  个男生和  $m$  个女生任意排成一列 ( $m < n$ ), 问任意两女生不相邻的概率是多少; 若排列成一圆环, 问任意两女生不相邻的概率又是多少.

1.13 从 1 到 1000 中随机取一个整数, 求取到的整数既不能被 6 整除又不能被 9 整除的概率.



- 1.14 两个不同的箱子中分别装有  $n$  个苹果, 若随机选一个箱子并拿走其中一个苹果, 求一个箱子里没有苹果时另一箱子还剩下  $k$  个苹果的概率 ( $k \in [n]$ ).
- 1.15 有  $m$  个相同或不同的白球和  $n$  个相同或不同的红球, 随机取出依次排成一列, 求第  $k$  次取出红球的概率 (分四种情况讨论).
- 1.16 将 3 个不同的球放入 4 个不同的杯子, 求杯子中球的最大个数分别为 1, 2, 3 的概率.
- 1.17 袋中有  $a$  个不同的白球,  $b$  个不同的红球, 假设有  $k$  个人依次任意无放回地从袋中取一个球, 问第  $i$  个人 ( $i \leq k$ ) 取出红球的概率是多少; 若为任意无放回地取球, 第  $i$  个人 ( $i \leq k$ ) 取出红球的概率又是多少.
- 1.18 一张圆桌有  $2n$  个位置, 将  $n$  对夫妻任意安排入座圆桌, 求任意一对夫妻不相邻的概率.
- 1.19 在区间  $[0, 1]$  内随机取两数, 求两数之积小于  $1/4$  的概率.
- 1.20 利用计算机编程计算: 在  $[0, 1]$  区间内任意取 4 个数  $a, b, c, d$ , 求事件

$$A = \{a^2 + \sin(b) + a \cdot e^c \leq d\}$$

发生的概率 (要求写出伪代码以及概率保留小数点后 5 位).

- 1.21 已知多重集  $A = \{a, a, a, b, b, b, b, c, c\}$ , 求  $A$  有多少种不同的排列.
- 1.22 对正整数  $m, n$  以及  $r < n$ , 证明:

$$\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}, \quad \binom{m+n}{r} = \sum_{i=0}^r \binom{m}{i} \binom{n}{r-i}, \quad \binom{2n}{n} = \sum_{i=0}^n \binom{n}{i}^2.$$

- 1.23 从  $m$  个不同的元素中无放回/有放回地取出  $r$  个元素进行排列, 分别有多少种不同的排法; 若从  $m$  个不同的元素中无放回/有放回地取出  $r$  个元素, 分别有多少种不同的取法.
- 1.24 求方程  $x_1 + x_2 + \dots + x_k \leq n$  的正整数解、非负整数解的个数 ( $n$  为正整数).
- 1.25 求方程  $x_1 + x_2 + \dots + x_k < n$  的正整数解、非负整数解的个数 ( $n$  为正整数).
- 1.26 利用第二类 Stirling 数的递推关系证明:

$$S(n, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n.$$



## 第2章 条件概率与独立性

前面关于事件概率的讨论都是在整个样本空间上进行, 不需考虑其它条件或限制因素. 然而在很多实际问题中, 我们往往关心随机事件在一定附加信息 (条件) 下发生的概率, 即条件概率. 它是概率论中一个非常重要且实用的概念. 可以帮助我们更好地分析和理解复杂的随机事件, 同时也有助于简化复杂事件概率的计算.

### 2.1 条件概率

#### 2.1.1 条件概率

首先来看一个例子, 随意投掷一枚骰子观察点数, 其样本空间  $\Omega = \{1, 2, \dots, 6\}$ . 用  $A$  表示观察到奇数点的事件, 则事件  $A = \{1, 3, 5\}$ , 根据古典概型有  $P(A) = 1/2$ . 用  $B$  表示观察到 3 点的事件, 根据古典概型有  $P(B) = 1/6$ .

现在考虑在事件  $A$  发生的情况下事件  $B$  发生的概率, 记为  $P(B|A)$ . 由于  $A = \{1, 3, 5\}$  且每种结果等可能发生, 由此可得条件概率

$$P(B|A) = 1/3 > P(B).$$

用事件  $C$  表示观察到 2 点, 根据古典概型有  $P(C) = 1/6$ , 但在事件  $A$  发生的情况下事件  $C$  不可能发生, 因此条件概率

$$P(C|A) = 0 < P(C).$$

由此可知一个随机事件发生的概率可能随着条件的改变而改变, 同时通过观察可以发现

$$P(B|A) = 1/3 = \frac{P(AB)}{P(A)} \quad \text{和} \quad P(C|A) = 0 = \frac{P(AC)}{P(A)}.$$

针对一般情形, 我们将上述关系作为条件概率的定义.

**定义 2.1** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 随机事件  $A \in \Sigma$  且  $P(A) > 0$ . 对任意事件  $B \in \Sigma$ , 称

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

为事件  $A$  发生的条件下事件  $B$  发生的概率, 简称 **条件概率** (conditional probability).

对任意事件  $A \in \Sigma$  有  $P(A) = P(A\Omega)/P(\Omega) = P(A|\Omega)$  成立, 因而任何随机事件的概率可以看作必然事件下的条件概率. 根据条件概率的定义有  $P(AB) = P(A)P(B|A)$ . 在本书后续章节中, 若出现条件概率  $P(B|A)$ , 一般都默认  $P(A) > 0$ .

根据条件概率的定义, 不难验证条件概率  $P(\cdot|A)$  具有以下一些基本性质:

1° **非负性**: 对任意事件  $B$  有  $P(B|A) \geq 0$ ;

2° **规范性**: 对样本空间  $\Omega$  有  $P(\Omega|A) = 1$ ;

3° **可列可加性**: 若  $B_1, B_2, \dots, B_n, \dots$  是可列无穷个互不相容的事件, 即  $B_i B_j = \emptyset$  ( $i \neq j$ ), 有

$$P(B_1 \cup B_2 \cup \dots \cup B_n \cup \dots | A) = P(B_1|A) + P(B_2|A) + \dots + P(B_n|A) + \dots.$$

从概率和条件概率的定义可知  $P(B|A) = P(AB)/P(A) \geq 0$  以及  $P(\Omega|A) = P(A\Omega)/P(A) = 1$ , 由此验证公理 1° 和公理 2°. 若可列个事件  $B_1, B_2, \dots, B_n, \dots$  是两两互不相容的, 则可列个事件  $AB_1, AB_2, \dots, AB_n, \dots$  也是两两互不相容的, 根据分配律有

$$P\left(\bigcup_{i=1}^{\infty} B_i \middle| A\right) = \frac{P(A(\bigcup_{i=1}^{\infty} B_i))}{P(A)} = \frac{P(\bigcup_{i=1}^{\infty} AB_i)}{P(A)} = \sum_{i=1}^{\infty} \frac{P(AB_i)}{P(A)} = \sum_{i=1}^{\infty} P(B_i|A)$$

由此可知公理 3° 成立. 由于条件概率满足概率的三条公理, 因此条件概率  $P(\cdot|A)$  仍然是一种概率.

**性质 2.1 (容斥原理)** 对随机事件  $A, B_1$  和  $B_2$  且满足  $P(A) > 0$ , 有

$$P(B_1 \cup B_2 | A) = P(B_1|A) + P(B_2|A) - P(B_1 B_2 | A).$$

**证明** 由条件概率的定义有

$$P(B_1 \cup B_2 | A) = P((B_1 \cup B_2) \cap A) / P(A).$$

再根据随机事件的分配律和容斥原理有

$$P((B_1 \cup B_2) \cap A) = P(AB_1 \cup AB_2) = P(AB_1) + P(AB_2) - P(AB_1 B_2),$$

上式两边同时除以  $P(A)$  即可完成证明.

**性质 2.2** 对随机事件  $A$  和  $B$  且满足  $P(A) > 0$ , 有  $P(\bar{B}|A) = 1 - P(B|A)$ .

**证明** 根据容斥原理有

$$1 = P(\Omega|A) = P(B \cup \bar{B}|A) = P(B|A) + P(\bar{B}|A) - P(B\bar{B}|A)$$

再根据事件  $B$  和  $\bar{B}$  互不相容有  $P(B\bar{B}|A) = 0$ , 从而完成证明.

事件  $A$  发生的条件下事件  $B$  发生的概率, 可以将  $A$  看作新的样本空间、而忽略以前的样本空间  $\Omega$ , 由此可以发现: **条件概率的本质是缩小了有效的样本空间**. 这也为计算条件概率提供了一种方法: 空间缩减法, 在新的样本空间  $A$  下考虑事件  $B$  发生的概率.

**例 2.1** 盒子中有 4 个不同的产品, 其中 3 个一等品, 1 个二等品. 从盒子中不放回随机取两次产品. 用  $A$  表示第一次拿到一等品的事件,  $B$  表示第二次取到一等品的事件, 求条件概率  $P(B|A)$ .

**解** 将盒子中 3 个一等产品分别编号为 1, 2, 3, 二等品编号 4. 用  $i$  和  $j$  分别表示第一、二次抽取的产品的编号, 由此可得

$$\begin{aligned}\Omega &= \{(i, j): i \neq j, i, j \in [4]\}, & A &= \{(i, j): i \neq j, i \neq 4\}, \\ B &= \{(i, j): i \neq j, j \neq 4\}, & AB &= \{(i, j): i \neq j, i, j \in [3]\}.\end{aligned}$$

计算可得  $|\Omega| = 12$ ,  $|A| = 9$ ,  $|B| = 9$  以及  $AB = 6$ . 根据古典概型有

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{4}, \quad P(B|A) = \frac{P(AB)}{P(A)} = \frac{1/2}{3/4} = \frac{2}{3}.$$

也可以采用 **样本空间缩减法** 来求解此问题: 当事件  $A$  发生后, 剩下 2 只一等品, 1 只二等品, 因此直接得到  $P(B|A) = 2/3$ .

**例 2.2** 箱子中有  $a$  个红球和  $b$  个白球, 依次任意无放回地取出  $n$  个球 ( $n \leq a + b$ ), 其中包括  $k$  个白球 ( $k \leq b$ ), 在此情形下求第一次取出白球的概率.

**解** 用  $A$  表示第一次取出白球的事件, 用  $B$  表示依次取出  $n$  个球中包括  $k$  个白球的事件. 根据题意和超几何分布有

$$P(A) = \frac{b}{a+b}, \quad P(B) = \frac{\binom{a}{n-k} \binom{b}{k}}{\binom{a+b}{n}}, \quad P(B|A) = \frac{\binom{a}{n-k} \binom{b-1}{k-1}}{\binom{a+b-1}{n-1}}.$$

所求条件概率

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

也可以采用 **样本空间缩减法** 来求解此问题: 在事件  $B$  发生的情况下, 即选中的  $n$  个球中有  $k$  个白球, 由于任何一球被第一次选中的可能性一样, 因此事件  $A$  发生的概率为  $k/n$ .

### 2.1.2 乘法公式

随机事件  $A$  和  $B$  满足  $P(A) > 0, P(B) > 0$ , 根据条件概率的定义可知

$$P(AB) = P(A)P(B|A) = P(B)P(A|B).$$

将上式进一步推广, 根据条件概率的定义有下面的乘法公式:

**定理 2.1** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 若随机事件  $A_1, A_2, \dots, A_n \in \Sigma$ , 且满足条件  $P(A_1 A_2 \cdots A_{n-1}) > 0$ , 则有

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

**例 2.3** 假设一批灯泡有 100 只, 其中有次品 10 只, 其余为正品. 不放回抽取地每次抽取一只, 求第三次才是正品的概率.

**解** 用  $A_i$  表示第  $i$  次抽到正品的事件 ( $i \in [3]$ ), 事件  $B$  表示第 3 次才抽到的正品, 则有  $B = \bar{A}_1 \bar{A}_2 A_3$ . 根据乘法公式有

$$P(B) = P(\bar{A}_1 \bar{A}_2 A_3) = P(\bar{A}_1)P(\bar{A}_2|\bar{A}_1)P(A_3|\bar{A}_1 \bar{A}_2) = \frac{10}{100} \times \frac{9}{99} \times \frac{90}{98} = \frac{9}{1078}.$$

**例 2.4** 设  $n$  把钥匙中只有一把能打开门. 不放回随机取出一把开门, 求第  $k$  次打开门的概率.

**解** 用  $A_i$  表示第  $i$  次没有打开门的事件, 则第  $k$  次打开门的事件可表示为  $A_1 A_2 \cdots A_{k-1} \bar{A}_k$ , 根据乘法公式有

$$\begin{aligned} P(A_1 A_2 \cdots A_{k-1} \bar{A}_k) &= P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \cdots P(A_{k-1}|A_1 \cdots A_{k-2})P(\bar{A}_k|A_1 A_2 \cdots A_{k-1}) \\ &= \frac{n-1}{n} \times \frac{n-2}{n-1} \times \cdots \times \frac{n-(k-1)}{n-(k-2)} \times \frac{1}{n-(k-1)} = \frac{1}{n} \end{aligned}$$

也可以根据 **抽签原理** 来求解该问题: 第  $k$  次打开门的概率与  $k$  无关, 每次打开门的概率相同, 共  $n$  把钥匙, 因此第  $k$  次打开门的概率为  $1/n$ .

**例 2.5** 假设有  $n$  对夫妻参加活动, 被随机分成  $n$  组, 每组一男一女, 求  $n$  对夫妻恰好两两被分到一组的概率.

**解** 用  $A_i$  表示第  $i$  对夫妻被分到同一组的事件, 则  $n$  对夫妻恰好两两被分到一组的事件可表示为  $A_1 A_2 \cdots A_n$ . 根据乘法公式有

$$\begin{aligned} P(A_1 A_2 \cdots A_n) &= P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \cdots P(A_n|A_1 A_2 \cdots A_{n-1}) \\ &= \frac{1}{n} \times \frac{1}{n-1} \times \cdots \times \frac{1}{1} = \frac{1}{n!}. \end{aligned}$$

**例 2.6** 第一个箱子里有  $n$  个不同的白球, 第二个箱子里有  $m$  个不同的红球, 从第一个箱子任意取走一球, 再从第二个箱子里任意取走一球放入第一个箱子, 依次进行, 直至第一、第二个箱子都为空, 求第一个箱子最后一次取走的球是白球的概率.

**解** 假设第一个箱子里的白球分别标号为  $1, 2, \cdots, n$ , 用  $A_i$  表示第一个箱子最后取走的是第  $i$  号白球的事件. 由此可知事件  $A_1, A_2, \cdots, A_n$  是两两互不相容的, 且第一个箱子最后一次取走的球是白球的事件可表示为  $A_1 \cup A_2 \cup \cdots \cup A_n$ , 根据事件的对称性可得其概率为

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^n P(A_i) = nP(A_1).$$

若事件  $A_1$  发生, 则从第一个箱子中取走的  $m+n-1$  个球均不是第 1 号白球, 用事件  $B_j$  表示第  $j$  次从第一个箱子里取走的球不是第 1 号白球, 即  $A_1 = B_1 B_2 \cdots B_{m+n-1}$ . 根据乘法公式有

$$\begin{aligned} P(A_1) &= P(B_1)P(B_2|B_1) \cdots P(B_m|B_1 B_2 \cdots B_{m-1}) \times P(B_{m+1} B_{m+2} \cdots B_{m+n-1} | B_1 B_2 \cdots B_m) \\ &= \left(1 - \frac{1}{n}\right)^m P(B_{m+1} B_{m+2} \cdots B_{m+n-1} | B_1 B_2 \cdots B_m) = \left(1 - \frac{1}{n}\right)^m \times \frac{1}{n}. \end{aligned}$$

由此可知第一个箱子最后一次取走的球是白球的概率为  $(1 - 1/n)^m$ .

**例 2.7** 假设箱子里有  $m$  个红球和  $n$  个白球, 现随机取出一球后放回, 并加入  $c$  个与取出球同色的球, 求前两次取出红球、后两次取出白球的概率.

**解** 用  $A_i$  表示第  $i$  次抽到红球的事件 ( $i \in [2]$ ), 事件  $B_i$  表示第  $i$  次抽到白球的事件 ( $i = 3, 4$ ), 我们有

$$\begin{aligned} P(A_1) &= \frac{m}{m+n}, & P(A_2|A_1) &= \frac{m+c}{m+n+c}, \\ P(B_1|A_1 A_2) &= \frac{n}{m+n+2c}, & P(B_2|A_1 A_2 B_1) &= \frac{n+c}{m+n+3c}. \end{aligned}$$

根据乘法公式有

$$P(A_1 A_2 B_1 B_2) = \frac{mn(m+c)(n+c)}{(m+n)(m+n+c)(m+n+2c)(m+n+3c)}$$

上述例子可用来作为疾病传染的粗略解释, 每取出一球代表疾病的一次传染, 每次传染将增加再传染的可能性.

## 2.2 全概率公式和贝叶斯公式

本节介绍概率计算中两个重要的公式: 全概率公式和贝叶斯公式.

### 2.2.1 全概率公式

全概率公式是概率论中最基本的公式之一, 将一个复杂事件的概率计算分解为若干简单事件的概率计算. 具体而言, 将一个复杂事件分解为若干不相容的简单事件之和, 通过分别计算简单事件的概率, 利用概率的可加性得到复杂事件的概率. 首先定义样本空间的一个分割.

**定义 2.2** 设  $A_1, A_2, \cdots, A_n$  是样本空间  $\Omega$  的一组事件, 若满足:

- i) 任意两个事件是互不相容性的, 即  $A_i \cap A_j = \emptyset$  ( $i \neq j$ );
- ii) 完备性  $\Omega = A_1 \cup A_2 \cup \cdots \cup A_n$ ,

则称事件  $A_1, A_2, \cdots, A_n$  为样本空间  $\Omega$  的一个 **分割**, 亦称 **完备事件组**.

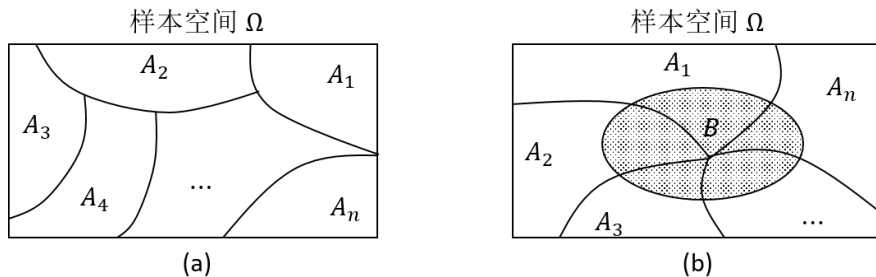


图 2.1 样本空间的分割与事件

如图 2.1(a) 所示, 若  $A_1, A_2, \dots, A_n$  为样本空间  $\Omega$  的一个分割, 则每次试验时在  $A_1, A_2, \dots, A_n$  中有且仅有一个事件发生. 对任何事件  $A \subseteq \Omega$ , 事件  $A$  与对立事件  $\bar{A}$  构成样本空间  $\Omega$  的一个分割.

基于样本空间的分割, 下面给出全概率公式:

**定理 2.2** 设事件  $A_1, A_2, \dots, A_n$  是样本空间  $\Omega$  的一个分割, 对任意事件  $B$  有

$$P(B) = \sum_{i=1}^n P(BA_i) = \sum_{i=1}^n P(A_i)P(B|A_i),$$

该公式被称为 **全概率公式** (Law of total probability).

**证明** 该定理的证明本质上是对加法和乘法事件的综合运用. 首先根据分配律有

$$B = B \cap \Omega = B \cap \left( \bigcup_{i=1}^n A_i \right) = \bigcup_{i=1}^n BA_i$$

由  $A_i \cap A_j = \emptyset$  可得  $BA_i \cap BA_j = \emptyset$ , 由概率的有限可列可加性有

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

由于任意事件和其对立事件构成一个分割, 对任意概率非零的事件  $A$  和  $B$  有

$$P(B) = P(AB) + P(\bar{A}B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A}).$$

如图 2.1(b), 还可以从另一个角度来理解全概率公式: 将事件  $B$  看作一个结果, 将事件  $A_1, A_2, \dots, A_n$  看作产生该结果的若干原因, 针对不同原因事件  $B$  发生的概率 (即条件概率  $P(B|A_i)$ ) 各不相同, 而到底是哪一种原因具有随机性. 具体而言, 每一种原因发生的概率  $P(A_i)$  是已知的, 以及每一种原因对结果  $B$  的影响  $P(B|A_i)$  已知, 则可以计算结果  $P(B)$ .

下面来看一些例子:

**例 2.8** 小明参加一次人工智能竞赛, 目前的排名不理想, 分析其原因: 方法不够新颖的概率为 50%, 通过设计新方法后取得理想排名的概率为 50%; 程度代码有误的概率为 30%, 通过纠正代码后



取得理想排名的概率为 60%; 数据不充分的概率为 20%, 通过采集更多数据后取得理想排名的概率为 80%. 求小明最后取得理想排名的概率.

**解** 用  $B$  表示小明最后取得理想排名的事件, 用  $A_1, A_2, A_3$  分别表示方法不够新颖、程度代码有误、数据不充分这三个事件, 根据题意有

$$P(A_1) = 50\%, P(A_2) = 30\%, P(A_3) = 20\%, P(B|A_1) = 50\%, P(B|A_2) = 60\%, P(B|A_3) = 80\%.$$

小明最后取得理想排名的概率

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = 59\%.$$

**例 2.9** 随意抛  $n$  次硬币, 证明正面朝上的次数是偶数 (或奇数) 的概率为  $1/2$ .

**证明** 用事件  $A$  表示前  $n-1$  次抛硬币正面朝上的次数为偶数, 其对立事件  $\bar{A}$  表示前  $n-1$  次抛硬币朝上的次数为奇数, 事件  $B$  表示前  $n$  次硬币朝上的次数为偶数. 于是有

$$P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) = \frac{P(A)}{2} + \frac{P(\bar{A})}{2} = \frac{1}{2}.$$

方法二: 直接计算概率. 若正面朝上的次数是偶数, 则随意抛  $n$  次硬币中正面朝上的次数为偶数分别有  $\{0, 2, 4, \dots, 2k\}$  ( $2k \leq n$ ), 根据概率公式直接计算有

$$\sum_{0 \leq k \leq n/2} \binom{n}{2k} \left(\frac{1}{2}\right)^{2k} \left(\frac{1}{2}\right)^{n-2k} = \frac{1}{2^n} \sum_{0 \leq k \leq n/2} \binom{n}{2k} = \frac{1}{2},$$

这里使用公式  $\sum_{0 \leq k \leq n/2} \binom{n}{2k} = 2^{n-1}$ .

**例 2.10** 假设有  $n$  个箱子, 每个箱子里有  $a$  只白球和  $b$  只红球, 现从第一个箱子取出一个球放入第二个箱子, 第二个箱子取出一个球放入第三个箱子, 依次类推, 求从最后一个箱子取出一球是红球的概率.

**解** 用  $A_i$  表示从第  $i$  个箱子取出红球的事件 ( $i \in [n]$ ), 则  $\bar{A}_i$  表示从第  $i$  个箱子取出白球的事件. 则有

$$P(A_1) = b/(a+b) \quad \text{和} \quad P(\bar{A}_1) = a/(a+b).$$

根据全概率公式有

$$P(A_2) = P(A_1)P(A_2|A_1) + P(\bar{A}_1)P(A_2|\bar{A}_1) = \frac{b}{a+b} \times \frac{b+1}{a+b+1} + \frac{a}{a+b} \times \frac{b}{a+b+1} = \frac{b}{a+b}.$$

由此可知  $P(\bar{A}_2) = a/(a+b)$ . 依次类推重复上述过程  $n-1$  次, 最后一个箱子中取出一球是红球的概率为  $b/(a+b)$ .

### 2.2.2 贝叶斯公式

贝叶斯公式也是概率论中最基本的公式之一, 在结果发生的情况下探讨是由何种原因导致结果. 具体而言, 假设有  $A_1, A_2, \dots, A_n$  种原因导致事件  $B$  发生, 贝叶斯公式研究在事件  $B$  发生情况下由原因  $A_i$  导致的概率, 即条件概率  $P(A_i|B)$ .

**定理 2.3** 设  $A_1, A_2, \dots, A_n$  是样本空间  $\Omega$  的一个分割, 用  $B$  表示任一事件且满足  $P(B) > 0$ . 对任意  $1 \leq i \leq n$  有

$$P(A_i|B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)},$$

该公式被称为 **贝叶斯公式** (Bayes' formula).

贝叶斯公式由条件概率和全概率公式直接推导可得. 由于任何事件和其对立事件都是样本空间的一个分割, 对任意概率非零的事件  $A$  和  $B$  有

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

全概率公式和贝叶斯公式应用的条件是相同的, 但解决的问题不同: 将事件  $A_1, A_2, \dots, A_n$  看作事件  $B$  发生的“原因”, 而事件  $B$  是伴随着原因  $A_1, A_2, \dots, A_n$  而发生的“结果”. 若知道各种原因  $P(A_i)$ , 以及在该原因下事件  $B$  发生的概率  $P(B|A_i)$ , 此时利用全概率公式计算结果事件  $B$  发生的概率; 若结果事件  $B$  已经发生, 此时利用贝叶斯公式探讨是由某原因  $A_i$  导致该结果的概率  $P(A_i|B)$ .

贝叶斯公式被应用于生活中的很多决策问题, 与决策理论密切相关, 下面来看一个简单的例子:

**例 2.11** 小明参加一次人工智能竞赛, 目前的排名不理想, 分析其原因: 方法不够新颖的概率为 50%, 通过设计新方法后取得理想排名的概率为 50%; 程度代码有误的概率为 30%, 通过纠正代码后取得理想排名的概率为 60%; 数据不充分的概率为 20%, 通过采集更多数据后取得理想排名的概率为 80%. 因为时间有限, 小明只能选择三种方案 (设计新方法、纠正代码、采集更多数据) 中一种, 想要取得理想的排名, 小明应该选择哪一种方案.

**解** 用  $B$  表示小明最后取得理想排名的事件, 用  $A_1, A_2, A_3$  分别表示方法不够新颖、程度代码有误、数据不充分这三个事件, 根据题意有

$$P(A_1) = 50\%, P(A_2) = 30\%, P(A_3) = 20\%, P(B|A_1) = 50\%, P(B|A_2) = 60\%, P(B|A_3) = 80\%.$$

小明最后取得理想排名的概率

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = 59\%.$$

根据贝叶斯公式有  $P(A_1|B) = P(A_1)P(B|A_1)/P(B) = 25/59$ , 同理可得  $P(A_2|B) = 18/59$  和  $P(A_3|B) = 16/59$ . 因此小明应该选择设计新方法来获得理想排名的概率更高.

**例 2.12 (三囚徒问题)** 三犯人  $a, b, c$  均被判为死刑, 法官随机赦免其中一人, 看守知道谁被赦免但不会说. 犯人  $a$  问看守:  $b$  和  $c$  谁会被执行死刑? 看守的策略: i) 若赦免  $b$ , 则说  $c$ ; ii) 若赦免  $c$ , 则说  $b$ ; iii) 若赦免  $a$ , 则以  $1/2$  的概率说  $b$  或  $c$ . 看守回答  $a$ : 犯人  $b$  会被执行死刑. 犯人  $a$  兴奋不已, 因为自己生存的概率为  $1/2$ . 犯人  $a$  将此事告诉犯人  $c$ ,  $c$  同样高兴, 因为他觉得自己的生存几率为  $2/3$ . 那么谁才是正确的呢?

**解** 用事件  $A, B, C$  分别表示犯人  $a, b, c$  被赦免, 由题意可知

$$P(A) = P(B) = P(C) = 1/3.$$

用事件  $D$  表示看守人说犯人  $b$  被执行死刑, 则有

$$P(D|A) = 1/2 \quad P(D|B) = 0 \quad P(D|C) = 1.$$

由全概率公式有

$$P(D) = P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) = 1/2.$$

由贝叶斯公式有

$$P(A|D) = P(A)P(D|A)/P(D) = 1/3 \quad \text{和} \quad P(C|D) = P(C)P(D|C)/P(D) = 2/3,$$

所以犯人  $a$  的推断不正确, 犯人  $c$  的推断正确.

贝叶斯公式提出了重要的推理逻辑, 在概率统计以及日常生活中存在多方面的应用. 假定  $A_1, A_2, \dots, A_n$  是导致结果事件  $B$  的“原因”, 概率  $P(A_i)$  被称为 **先验概率** (prior probability), 反映了各种“原因”的可能性大小, 一般都是根据先前的经验总结而成. 若现在试验产生了事件  $B$ , 这个信息有助于探讨事件发生的“原因”, 条件概率  $\Pr(A_i|B)$  被称 **后验概率** (posterior probability), 反映了试验之后对各种“原因”发生可能性的新知识.

例如, 医生为诊断病人患了疾病  $A_1, A_2, \dots$  中哪一种疾病, 可以对病人进行检查, 确定某个指标  $B$  (如血糖、血脂、血钙等), 从而帮助诊断, 此时可以采用贝叶斯公式来计算相关概率. 根据以往的数据资料确定先验概率  $P(A_i)$ , 即人们患各种疾病的可能性; 在通过医学知识确定概率  $P(B|A_i)$ , 最后通过贝叶斯公式计算后验概率  $P(A_i|B)$ . 在实际应用中, 可能检验多个指标  $B$ , 综合所有的后验概率进行诊断. 在自动诊断和辅助诊断的专家系统中, 这种方法非常实用.

贝叶斯公式使用中最存在争议之处在于先验的选取, 在很多实际应用中往往都根据以往的数据而得出的, 符合概率的频率解释, 但需要以往大量的历史数据, 在实际应用中通常难以满足. 其次, 在很多应用中先验概率可能由某一种主观的方式给出, 例如对未来宏观经济形势 (或对某人诚信度) 的判断, 这种将概率解释为信任程度的做法明显带有主观性, 通常被称为 **主观概率**.

伊索寓言“孩子与狼”讲一个小孩每天到山上放羊, 山里有狼出没, 第一天他在山上喊“狼来了! 狼来了!”, 山下的村民们闻声便去打狼, 到了山上发现没有狼; 第二天仍是如此; 第三天狼真来了, 可无论小孩怎么喊叫, 也没有人来救他, 因为前二次他说了谎话, 人们不再相信他了. 我们可以将这个寓言抽象为一个主观概率的例子, 并利用贝叶斯公式来分析这个寓言中村民们的心理活动.

**例 2.13** 假设村民们对这个小孩的印象一般, 认为小孩说谎话和说真话的概率相同, 均为  $1/2$ . 假设说谎话的小孩喊狼来了时狼真来的概率为  $1/3$ , 而说真话的小孩喊狼来了时狼真来的概率为  $3/4$ . 若第一天、第二天上山均没有发现狼, 请分析村民们的心理活动.

**解** 用  $B_1$  和  $B_2$  分别表示第一天和第二天狼来了的事件, 用  $A_1$  表示小孩第一天说谎话的事件, 用  $A_2$  表示在第一天狼没有的情况下小孩第二天说谎话的事件, 根据题意可知

$$P(A_1) = P(\overline{A_1}) = 1/2, P(B_1|A_1) = 1/3, P(B_1|\overline{A_1}) = 3/4, P(B_2|A_2) = 1/3, P(B_2|\overline{A_2}) = 3/4.$$

第一天村民上山打狼但没有发现狼, 根据贝叶斯公式可知村民们对说谎话小孩的认识发生了改变, 体现在

$$P(A_2) = P(A_1|\overline{B_1}) = \frac{P(\overline{B_1}|A_1)P(A_1)}{P(\overline{B_1}|A_1)P(A_1) + P(\overline{B_1}|\overline{A_1})P(\overline{A_1})} = \frac{8}{11} \approx 0.7273, \quad P(\overline{A_2}) = \frac{3}{11}.$$

此时, 村民对这个小孩说谎话的概率从 50% 调整到 72.72%.

第二天村民上山打狼还是没有发现狼, 根据贝叶斯公式可知村民们对说谎话小孩的认识又发生了改变, 体现在

$$P(A_2|\overline{B_2}) = \frac{P(\overline{B_2}|A_2)P(A_2)}{P(\overline{B_2}|A_2)P(A_2) + P(\overline{B_2}|\overline{A_2})P(\overline{A_2})} = \frac{64}{73} \approx 0.8767.$$

此时, 村民对这个小孩说谎话的概率从 72.72% 调整到 87.67%.

这表明村民们经过两次上当, 对这个小孩说谎话的概率从 50% 上升到 87.67%, 给村民留下这种印象, 他们听到第三次呼叫时不会再上山打狼.

## 2.3 事件独立性

前面的例子表明, 在事件  $A$  发生的条件下事件  $B$  发生的条件概率  $P(B|A)$ , 通常不等于事件  $B$  发生的概率  $P(B)$  (无任何附加条件), 即  $P(B|A) \neq P(B)$ , 也就是说“事件  $A$  发生通常会改变事件  $B$  发生的可能性”. 然而在有些特殊情形下, 事件  $A$  的发生对事件  $B$  的发生可能没有任何影响, 这就是本节所研究的事件独立性.

### 2.3.1 两事件的独立性

**定义 2.3** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 若事件  $A, B \in \Sigma$  且满足  $P(AB) = P(A)P(B)$ , 则称事件  $A$  与  $B$  是相互独立的, 简称独立.

根据定义可知任何事件与不可能事件 (或必然事件) 是相互独立的. 设两事件  $A$  和  $B$  是相互独立的, 且满足  $P(A)P(B) > 0$ , 则有

$$P(AB) = P(A)P(B) \Leftrightarrow P(B|A) = P(B) \Leftrightarrow P(A|B) = P(A).$$

**性质 2.3** 若事件  $A$  与  $B$  相互独立, 则  $A$  与  $\bar{B}$ ,  $\bar{A}$  与  $B$ ,  $\bar{A}$  与  $\bar{B}$  都互相独立.

**证明** 根据事件差公式  $P(A - B) = P(A) - P(AB)$  有

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

同理可证  $P(\bar{A}B) = P(\bar{A})P(B)$ . 利用容斥原理有

$$\begin{aligned} P(\bar{A}\bar{B}) &= 1 - P(A \cup B) = 1 - P(A) - P(B) + P(AB) \\ &= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(\bar{A})P(\bar{B}), \end{aligned}$$

从而完成证明.

如何判断事件的独立性? 根据定义直接计算进行判断:

**例 2.14** 从一副扑克 (不含大王、小王) 中随机抽取一张扑克, 用事件  $A$  表示抽到 10, 事件  $B$  表示抽到黑色的扑克. 事件  $A$  与  $B$  是否独立?

**解** 根据问题可知一副扑克 (不含大王、小王) 52 张, 黑色扑克 26 张, 4 张 10, 根据古典概型有

$$P(A) = 4/52 = 1/13, \quad P(B) = 1/2.$$

由此可得  $P(AB) = 2/52 = 1/26 = P(A)P(B)$ , 根据定义可知事件  $A$  和  $B$  是相互独立的.

也可以根据实际问题判断事件的独立性, 例如

- 两人独立射击打靶、且互不影响, 因此两人中靶的事件相互独立;
- 从  $n$  件产品中随机抽取两件, 事件  $A_i$  表示第  $i$  件是合格品. 若有放回抽取则事件  $A_1$  与  $A_2$  相互独立; 若不放回则不独立;
- 机器学习的经典假设是训练数据独立同分布采样.

独立与互斥之间的关系: 若事件  $A$  和  $B$  是独立的, 有  $P(AB) = P(A)P(B)$ , 独立性与概率相关, 反映事件的概率属性; 若事件  $A$  和  $B$  是互斥的, 有  $AB = \emptyset$ , 互斥性与事件的运算关系相关, 与概率无关, 因此独立性与互不相容性反映事件不同的性质.

类似于条件概率, 可以定义概率论中的条件独立性, 即在一定条件下两事件是相互独立的.

**定义 2.4** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 事件  $C \in \Sigma$  有  $P(C) > 0$  成立, 若事件  $A, B \in \Sigma$  满足

$$P(AB|C) = P(A|C)P(B|C) \quad \text{或} \quad P(A|BC) = P(A|C),$$

则称事件  $A$  和  $B$  在  $C$  发生的情况下是 **条件独立的** (conditional independent).

下面给出一个关于条件独立性的例子:

**例 2.15** 假设一个箱子中有  $k+1$  枚不均匀的硬币, 投掷第  $i$  枚硬币时正面向上的概率为  $i/k$  ( $i = 0, 1, 2, \dots, k$ ). 现从箱子中任意取出一枚硬币、并任意重复投掷多次, 若前  $n$  次正面向上, 求第  $n+1$  次正面向上的概率.

**解** 用  $A$  表示第  $n+1$  次投掷正面向上的事件, 用  $B$  表示前  $n$  次投掷都正面向上的事件, 用  $C_i$  表示从箱子中取出第  $i$  枚硬币的事件 ( $i = 0, 1, 2, \dots, k$ ). 根据条件概率的定义可知

$$P(A|B) = P(AB)/P(B).$$

根据全概率公式和条件独立性有

$$P(AB) = \sum_{i=0}^k P(C_i)P(AB|C_i) = \sum_{i=0}^k P(C_i)P(A|C_i)P(B|C_i) = \frac{1}{k+1} \sum_{i=0}^k \frac{i^{n+1}}{k^{n+1}},$$

以及

$$P(B) = \sum_{i=0}^k P(C_i)P(B|C_i) = \frac{1}{k+1} \sum_{i=0}^k \frac{i^n}{k^n},$$

由此可得

$$P(A|B) = \frac{\sum_{i=0}^k (i/k)^{n+1}}{\sum_{i=0}^k (i/k)^n}.$$

当  $k$  非常大或  $k \rightarrow +\infty$  时可利用积分近似

$$\frac{1}{k} \sum_{i=1}^k (i/k)^n \approx \int_0^1 x^n dx = \frac{1}{n+1} \quad \text{和} \quad \frac{1}{k} \sum_{i=1}^k (i/k)^{n+1} \approx \int_0^1 x^{n+1} dx = \frac{1}{n+2},$$

此时有  $P(A|B) \approx (n+1)/(n+2)$ .

### 2.3.2 多个事件的独立性

**定义 2.5** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 若事件  $A, B, C \in \Sigma$  且满足

- 事件两两独立, 即  $P(AB) = P(A)P(B)$ ,  $P(AC) = P(A)P(C)$  和  $P(BC) = P(B)P(C)$ ,
- $P(ABC) = P(A)P(B)P(C)$ ,

则称事件  $A, B, C$  是 **相互独立的**.

根据定义可知若事件  $A, B, C$  是相互独立的, 则事件  $A, B, C$  是两两相互独立的; 但反之不一定成立, 还需满足  $P(ABC) = P(A)P(B)P(C)$ . 下面给出一个简单的例子说明: 三事件的两两独立并不能得出三事件相互独立.

**[Bernstein 反例]** 一个均匀的正四面体, 第一面是红色, 第二面是白色, 第三面是黑色, 第四面同时有红、白、黑三种颜色. 随意投掷一次该四面体, 用  $A, B, C$  分别表示红色、白色、黑色朝下的事件, 因为有一面同时包含三种颜色, 有

$$P(A) = P(B) = P(C) = 1/2 \quad \text{和} \quad P(AB) = P(BC) = P(AC) = 1/4,$$

由此可得事件  $A, B, C$  两两独立. 但由于

$$P(ABC) = 1/4 \neq 1/8 = P(A)P(B)P(C),$$

由此可知  $A, B, C$  不是相互独立的.

**定义 2.6** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 若事件  $A_1, A_2, \dots, A_n \in \Sigma$  中任意  $k$  个事件是相互独立的 ( $k \geq 2$ ), 即满足

- 对任意  $1 \leq i_1 < i_2 \leq n$  有  $P(A_{i_1}A_{i_2}) = P(A_{i_1})P(A_{i_2})$  成立;
- 对任意  $1 \leq i_1 < i_2 < i_3 \leq n$  有  $P(A_{i_1}A_{i_2}A_{i_3}) = P(A_{i_1})P(A_{i_2})P(A_{i_3})$  成立;
- $\dots \dots$
- $P(A_1A_2 \cdots A_n) = P(A_1)P(A_2) \cdots P(A_n)$ ,

则称事件  $A_1, A_2, \dots, A_n$  是 **相互独立的**.

根据定义可知,  $n$  个事件的相互独立性应满足  $\binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n - n - 1$  个等式, 同样  $n$  个事件的相互独立性与两两独立性是不同的概念. 类似地可以定义多个事件的条件独立性. 下面看一个关于独立性的例子.

**例 2.16** 三人独立破译一份密码, 每人单独能破译的概率分别为  $1/5, 1/3, 1/4$ , 问三人中至少有一人能破译密码的概率.

**解** 用事件  $A_i$  表示第  $i$  个人破译密码 ( $i \in [3]$ ), 根据题意有

$$P(A_1) = 1/5, \quad P(A_2) = 1/3, \quad P(A_3) = 1/4.$$

根据容斥原理和独立性, 三人中至少有一人能破译密码的概率为

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3) = 3/5.$$



也可以根据对偶性和独立性来求解该问题, 三人中至少有一人能破译密码的概率为

$$P(A_1 \cup A_2 \cup A_3) = 1 - P(\bar{A}_1 \bar{A}_2 \bar{A}_3) = 1 - P(\bar{A}_1)P(\bar{A}_2)P(\bar{A}_3) = 1 - \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{3}{4} = 3/5.$$

从上例可知: 尽管每个人能破译密码的概率都不大于  $1/3$ , 但三人独立进行破译, 则至少有一人破译密码的概率则为  $3/5$ , 由此提高了破译密码的概率. 我们可以将类似问题推广到更一般的情况.

若事件  $A_1, A_2, \dots, A_n$  相互独立, 发生的概率分别为  $p_1, p_2, \dots, p_n$ , 则事件  $A_1, A_2, \dots, A_n$  中至少有一事件发生的概率为

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - P(\bar{A}_1 \bar{A}_2 \dots \bar{A}_n) = 1 - (1 - p_1)(1 - p_2) \dots (1 - p_n);$$

此外, 事件  $A_1, A_2, \dots, A_n$  中至少有一事件不发生的概率为

$$P(\bar{A}_1 \cup \bar{A}_2 \cup \dots \cup \bar{A}_n) = 1 - P(A_1 A_2 \dots A_n) = 1 - p_1 p_2 \dots p_n.$$

由此可知: 尽管每个事件发生的概率  $p_i$  都非常小, 但若  $n$  非常大, 则  $n$  个相互独立的事件中“至少有一事件发生”或“至少有一事件不发生”的概率可能很大.

**定义 2.7 (小概率原理)** 若事件  $A$  在一次试验中发生的概率非常小, 但经过多次独立地重复试验, 事件  $A$  的发生是必然的, 称之为 **小概率原理**.

小概率原理可通过严格的数学证明得到: 若事件  $A_1, A_2, \dots, A_n, \dots$  独立且每事件发生的概率  $P(A_i) = p > 0$  非常小, 则有

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - P(\bar{A}_1 \bar{A}_2 \dots \bar{A}_n) = 1 - (1 - p)^n \rightarrow 1 \quad \text{当} \quad n \rightarrow \infty,$$

即独立重复多次的小概率事件亦可成立必然事件.

**例 2.17** 冷战时期美国的导弹精度 90%, 苏联的导弹精度 70%, 但苏联的导弹数量特别多, 导弹的数量能否弥补精度的不足?

**解** 假设每次独立发射  $n$  枚导弹, 用事件  $A_i$  表示第  $i$  枚导弹命中目标, 则  $n$  枚导弹击中目标的概率为

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = 1 - (1 - 0.7)^n \geq 0.9 \quad \Rightarrow \quad n \geq 2,$$

因此每次独立发射 2 枚导弹, 击中目标的概率高于 90%.

**例 2.18** 假设市场上有  $m$  种不同类型的邮票, 一位集邮爱好者收集第  $i$  种邮票的概率为  $p_i$ , 且  $p_1 + p_2 + \dots + p_m = 1$ . 假设每次集邮都是独立同分布的, 若现已收集到  $n$  张邮票, 用  $A_i$  表示至少收集到第  $i$  种类型邮票的事件, 求概率  $P(A_i)$ ,  $P(A_i \cup A_j)$  以及  $P(A_i | A_j)$  ( $i \neq j$ ).



解 根据题意有

$$P(A_i) = 1 - P(\overline{A_i}) = 1 - P(\text{收集的 } n \text{ 张邮票中没有第 } i \text{ 种类型邮票}) = 1 - (1 - p_i)^n.$$

同理可得

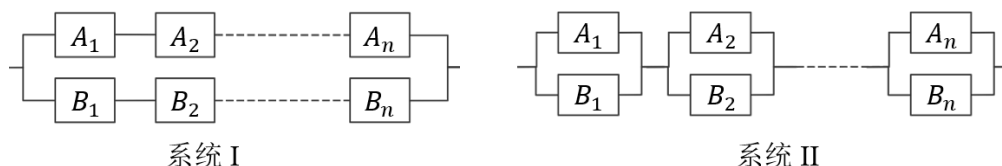
$$P(A_i \cup A_j) = 1 - P(\overline{A_i} \cap \overline{A_j}) = 1 - (1 - p_i - p_j)^n.$$

利用容斥原理和条件概率的定义有

$$\begin{aligned} P(A_i \cup A_j) &= \frac{P(A_i A_j)}{P(A_j)} = \frac{P(A_i) + P(A_j) - P(A_i \cup A_j)}{P(A_j)} \\ &= \frac{1 - (1 - p_i)^n - (1 - p_j)^n + (1 - p_i - p_j)^n}{1 - (1 - p_j)^n}. \end{aligned}$$

为保证系统的可靠性, 近代电子系统通常由多个独立的元件构成, 一个元件能正常工作的概率称为这个元件的可靠性. 由元件组成的系统能正常工作的概率称为系统的可靠性.

**例 2.19** 设构成系统的每个元件的可靠性均为  $p$  ( $0 < p < 1$ ), 且各元件是否正常工作是相互独立的. 设有  $2n$  个元件按下图所示, 两种不同连接方式构成两个不同的系统, 比较这两种系统的可靠性大小.



**解** 用事件  $A_i$  和  $B_i$  表示图中所对应的元件正常工作 ( $i = 1, 2, \dots, n$ ). 可以发现系统 I 有两条通路, 它能正常工作当且仅当两条通路至少有一条能正常工作, 而每一条通路能正常工作当且仅当它的每个元件能正常工作, 因此有系统 I 的可靠性为

$$\begin{aligned} &P((A_1 A_2 \cdots A_n) \cup (B_1 B_2 \cdots B_n)) \\ &= P(A_1 A_2 \cdots A_n) + P(B_1 B_2 \cdots B_n) - P(A_1 A_2 \cdots A_n B_1 B_2 \cdots B_n) = 2p^n - p^{2n} = p^n(2 - p^n). \end{aligned}$$

系统 II 由  $n$  对并联元件  $\{A_i, B_i\}$  组成, 它能正常工作当且仅当每对并联元件组能够正常工作, 因此系统 II 的可靠性为

$$P\left(\bigcap_{i=1}^n (A_i \cup B_i)\right) = \prod_{i=1}^n P(A_i \cup B_i) = (2p - p^2)^n = p^n(2 - p)^n.$$

利用数学归纳法可证明当  $n \geq 2$  时有  $(2 - p)^n > 2 - p^n$  成立, 由此可知系统 II 的可靠性更好.

## 2.3.3 Borel-Cantelli 引理\*

Borel-Cantelli 引理常常被用来计算事件的概率为 0 或 1, 首先介绍一个有用的引理:

**引理 2.1** 若数列  $\{p_i\}_{i=1}^n$  满足  $p_i \in [0, 1]$  和  $\sum_{i=1}^{\infty} p_i = +\infty$ , 则有  $\prod_{i=1}^{\infty} (1 - p_i) = 0$ .

**证明** 对任意  $x \in [0, 1]$ , 有  $\ln(1 - x) \leq -x$ , 于是得到

$$\ln \prod_{i=1}^{\infty} (1 - p_i) \leq \ln \prod_{i=1}^n (1 - p_i) = \sum_{i=1}^n \ln(1 - p_i) \leq \sum_{i=1}^n -p_i.$$

分别对上式两边取极限  $n \rightarrow +\infty$  有  $\ln \prod_{i=1}^{\infty} (1 - p_i) = -\infty$ , 由此完成证明.

根据上述引理, 我们可以证明如下定理

**定理 2.4 (Borel-Cantelli 引理)** 设  $(\Omega, \Sigma, P)$  是一个概率空间, 以及事件系列  $A_i \in \Sigma$ , 令事件  $A = \bigcap_{n=1}^{+\infty} \bigcup_{i=n}^{+\infty} A_i$ , 则有

- 若  $\sum_{i=1}^{\infty} P(A_i) < +\infty$  则有  $P(A) = 0$ ;
- 若  $\sum_{i=1}^{\infty} P(A_i) = +\infty$  且事件  $\{A_i\}$  相互独立, 则有  $P(A) = 1$ .

该定理考虑事件序列  $\{A_i\}_{i=1}^{+\infty}$  中属于无限多  $A_i$  的基本事件的概率和. 在定理第二不妨中事件  $A_i$  之间相互独立

**证明** 根据无穷级数  $\sum_{i=1}^{\infty} P(A_i) < +\infty$  收敛的性质可知  $\lim_{n \rightarrow \infty} \sum_{i=n}^{\infty} P(A_i) = 0$ . 根据题意可知  $A \subseteq \bigcup_{i=n}^{+\infty} A_i$ , 利用 Union bounds 有

$$P(A) \leq P\left(\bigcup_{i=n}^{+\infty} A_i\right) \leq \sum_{i=n}^{+\infty} P(A_i),$$

上式两边同时对  $n \rightarrow +\infty$  取极限证明  $P(A) = 0$ .

针对第二个问题, 不妨设  $B_n = \bigcup_{i=n}^{+\infty} A_i$ , 由此可知  $A = \bigcap_{n=1}^{+\infty} B_n$ . 给定任意正整数  $m > n \geq 1$ , 利用德摩根律和独立性假设有

$$P(\overline{B_n}) = P\left(\bigcap_{i=n}^{\infty} \overline{A_i}\right) = \lim_{m \rightarrow +\infty} P\left(\bigcap_{i=n}^m \overline{A_i}\right) = \lim_{m \rightarrow +\infty} \prod_{i=n}^m P(\overline{A_i}) = \prod_{i=n}^{+\infty} (1 - P(A_i))$$

根据引理 2.1 可得  $P(\overline{B_n}) = 0$ , 结合德摩根律进一步有

$$P(\bar{A}) = P\left(\bigcup_{n=1}^{+\infty} \overline{B_n}\right) \leq \sum_{n=1}^{+\infty} P(\overline{B_n}) = 0,$$

由此完成证明.

## 2.4 案例分析

下面将利用本节知识来解决一些实际的问题, 值得注意的是贝叶斯公式在人工智能的决策任务中有诸多的应用, 例如朴素贝叶斯分类器等, 由于涉及到多维随机变量相关, 我们将在后面的章节中介绍贝叶斯公式的应用.

### 2.4.1 多项式相等

有两个较为复杂的多项式, 例如

$$\begin{aligned} F(x) &= (x+2)^7(x+3)^5 + (x+1)^{100} + (x+2)(x+3) + x^{20}, \\ G(x) &= (x+3)^{100} - (x+1)^{25}(x+2)^{30} + x^{20} + (x-2)(x-3) \cdots (x-100). \end{aligned}$$

是否存在一种方法验证  $F(x) \equiv G(x)$ .

最容易想到的方法是将多项式全部展开, 合并同类项, 比较多项式每项的系数, 若相应的系数完全相同则有  $F(x) \equiv G(x)$ . 但这种方法通常需要较高的计算时间开销, 当多项式较复杂时更加困难, 是否存在一种简单快捷的验证方法.

我们介绍一种利用随机性来求解该问题的简单方法: 不妨假设  $F(x)$  和  $G(x)$  的最高次 (或多项式的度) 不超过  $d$ , 考虑从集合  $[100d] = \{1, 2, \dots, 100d\}$  中等可能随意选取一个数  $r$ , 然后计算  $F(r)$  和  $G(r)$ , 若  $F(r) \neq G(r)$  则返回  $F(x) \not\equiv G(x)$ ; 否则返回  $F(x) \equiv G(x)$ . 下面分析该方法的正确性:

- 若多项式  $F(x) \equiv G(x)$ , 则该方法得到“正确”结果, 因为对任意  $r \in [100d]$  都有  $F(r) = G(r)$ .
- 若多项式  $F(x) \not\equiv G(x)$  且  $F(r) \neq G(r)$ , 则该方法也得到“正确”结果, 因为找到了一个  $r \in [100d]$  使得  $F(r) \neq G(r)$  成立.
- 若多项式  $F(x) \not\equiv G(x)$  但  $F(r) = G(r)$ , 则该方法得到“错误”结果. 当  $F(x) \not\equiv G(x)$  时, 依然存在  $r \in [100d]$  使得  $F(r) = G(r)$  成立, 此时  $r$  是多项式  $F(x) - G(x) = 0$  的一个实数根. 根据代数知识不超过  $d$  次多项式  $F(x) - G(x) = 0$  至多有  $d$  个实数根, 而  $r$  从  $[100d]$  中等可能随机选取, 因此有

$$P[F(r) = G(r)] \leq d/100d = 1/100.$$

利用独立性可以进一步提高方法返回“正确”的概率: 从集合  $[100d]$  中独立地随意选取  $k$  ( $< d$ ) 个数  $r_1, r_2, \dots, r_k$ . 若存在  $r_i$  使得  $F(r_i) \neq G(r_i)$  成立, 则返回  $F(x) \not\equiv G(x)$ , 否则返回  $F(x) \equiv G(x)$ .

这里仅分析该方法返回“错误”结果发生的概率, 当  $F(x) \not\equiv G(x)$  时出现  $F(r_1) = G(r_1), F(r_2) = G(r_2), \dots, F(r_k) = G(r_k)$  的概率, 根据事件的独立性与前面的分析有

$$P\left(\bigcap_{i=1}^k \{F(r_i) = G(r_i)\}\right) = \prod_{i=1}^k P(F(r_i) = G(r_i)) \leq 1/100^k,$$

因此显著提高了方法返回“正确”结果的概率.

### 2.4.2 大矩阵乘法

本节考虑利用概率随机性来快速验证矩阵乘法的问题. 假设给定三个矩阵  $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \{0, 1\}^{n \times n}$ , 其中  $n$  非常大, 例如  $n \geq 10000000$ , 我们研究的问题: 验证下面的矩阵乘法是否成立

$$\mathbf{AB} \stackrel{?}{=} \mathbf{C}.$$

若直接采用矩阵乘法计算  $\mathbf{AB}$ , 然后再与矩阵  $\mathbf{C}$  进行比较, 则计算复杂开销为  $O(n^3)$ . 也可以采用更为精妙的算法, 比如采用分治策略, 目前最好的确定性算法的计算复杂开销为  $O(n^{2.37})$ , 我们采用概率的随机方法进一步降低计算开销.

类似于验证多项式  $F(x) \equiv G(x)$  的方法, 我们随机选取一个向量  $\bar{\mathbf{r}} = (r_1, r_2, \dots, r_n)^\top$ , 其中元素  $r_1, r_2, \dots, r_n$  都是从  $\{0, 1\}$  中独立等可能随机选取所得. 下面验证

$$\mathbf{A}\bar{\mathbf{B}}\bar{\mathbf{r}} = \mathbf{A}(\mathbf{B}\bar{\mathbf{r}}) \stackrel{?}{=} \mathbf{C}\bar{\mathbf{r}}.$$

计算  $\mathbf{A}(\mathbf{B}\bar{\mathbf{r}})$  和  $\mathbf{C}\bar{\mathbf{r}}$ , 以及比较两个向量是否相等的计算复杂开销为  $O(n^2)$ . 若  $\mathbf{A}(\mathbf{B}\bar{\mathbf{r}}) \neq \mathbf{C}\bar{\mathbf{r}}$  则可以直接得到结果  $\mathbf{AB} \neq \mathbf{C}$ ; 而  $\mathbf{A}(\mathbf{B}\bar{\mathbf{r}}) = \mathbf{C}\bar{\mathbf{r}}$  则不能直接得到结果  $\mathbf{AB} = \mathbf{C}$ , 此时可以将上述过程独立地进行  $k$  次, 以此用较大的概率保证有  $\mathbf{AB} = \mathbf{C}$  成立. 该过程被称为 Freivalds 算法, 如下所示:

```

输入: 矩阵  $\mathbf{A}, \mathbf{B}, \mathbf{C}$ 
输出: 是/否          %% 验证  $\mathbf{AB} \stackrel{?}{=} \mathbf{C}$ 
-----
For  $i = 1 : k$ 
    随机选择向量  $\bar{\mathbf{r}}_i = (r_{i1}, r_{i2}, \dots, r_{in})$ , 其每个元素是从  $\{0, 1\}$  独立等可能随机采样所得
    计算向量  $\bar{\mathbf{p}}_i = \mathbf{A}(\mathbf{B}\bar{\mathbf{r}}_i) - \mathbf{C}\bar{\mathbf{r}}_i$ 
    If  $\{\bar{\mathbf{p}}_i \text{ 不是零向量}\}$  then
        返回“否”
    EndIf
EndFor
返回“是”.

```

关于有效性, 若算法返回“否”, 则必有  $\mathbf{AB} \neq \mathbf{C}$ , 因为找到了一个  $\bar{\mathbf{r}}$  使得  $\mathbf{A}(\mathbf{B})\bar{\mathbf{r}} \neq \mathbf{C}\bar{\mathbf{r}}$  成立; 若算法返回“是”, 则不一定有  $\mathbf{AB} = \mathbf{C}$  成立, 但我们可以给出以较大的概率保证有  $\mathbf{AB} = \mathbf{C}$  成立.

**定理 2.5** 设随机向量  $\bar{\mathbf{r}}_1, \bar{\mathbf{r}}_2, \dots, \bar{\mathbf{r}}_k \in \{0, 1\}^n$  中每个元素都是从  $\{0, 1\}$  独立等可能随机选取, 若  $\mathbf{AB} \neq \mathbf{C}$ , 则有

$$P \left[ \bigcap_{i=1}^k \{ \mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_i = \mathbf{C}\bar{\mathbf{r}}_i \} \right] \leq \frac{1}{2^k}.$$

根据该定理可以选择  $k = \log_2 n$ , 则 Freivalds 算法计算复杂度为  $O(n^2 \log n)$ , 若算法返回“否”, 则有  $\mathbf{AB} \neq \mathbf{C}$ ; 若返回“是”, 则有  $P(\mathbf{AB} = \mathbf{C})$  成立的概率超过  $1 - 1/n$ .

**证明** 首先根据随机向量  $\bar{\mathbf{r}}_1, \bar{\mathbf{r}}_2, \dots, \bar{\mathbf{r}}_k$  的独立同分布性有

$$P\left[\bigcap_{i=1}^k \{\mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_i = \mathbf{C}\bar{\mathbf{r}}_i\}\right] = \prod_{i=1}^k P[\{\mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_i = \mathbf{C}\bar{\mathbf{r}}_i\}] = (P[\{\mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_1 = \mathbf{C}\bar{\mathbf{r}}_1\}])^k. \quad (2.1)$$

若  $\mathbf{AB} \neq \mathbf{C}$ , 则必有  $\mathbf{D} = (d_{ij})_{n \times n} = \mathbf{AB} - \mathbf{C} \neq (0)_{n \times n}$ , 此时不妨假设  $d_{11} \neq 0$ . 随机向量  $\bar{\mathbf{r}}_1 = (r_{11}, r_{12}, \dots, r_{1n})^\top$  中每个元素都是从  $\{0, 1\}$  独立等可能随机选取, 由于结果返回“是”可知  $\mathbf{D}\bar{\mathbf{r}}_1 = 0$ , 由此可得

$$d_{11}r_{11} + d_{12}r_{12} + \dots + d_{1n}r_{1n} = 0 \implies r_{11} = -\frac{d_{12}r_{12} + \dots + d_{1n}r_{1n}}{d_{11}}.$$

因此无论  $r_{12}, \dots, r_{1n}$  取何值, 等式  $d_{11}r_{11} + d_{12}r_{12} + \dots + d_{1n}r_{1n} = 0$  是否成立可根据  $r_{11}$  的值决定. 再根据  $P(r_{11} = 0) = P(r_{11} = 1) = 1/2$  得到等式  $d_{11}r_{11} + d_{12}r_{12} + \dots + d_{1n}r_{1n} = 0$  成立的概率不超过  $1/2$ , 因此有

$$P[\{\mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_1 = \mathbf{C}\bar{\mathbf{r}}_1\}] \leq 1/2.$$

结合 (2.1) 完成证明.

证明的思想又被称为 **延迟决策原理** (Principle of deferred decision), 当有多个随机变量解决一个问题时, 可以先着重考虑其中一个或一些变量, 而让其它剩余的变量保持随机性, 即延迟甚至不需考虑剩余变量对决策的影响. 在上面的证明过程中, 针对多个随机变量  $r_{11}, r_{12}, \dots, r_{1n}$ , 我们着重考虑随机变量  $r_{11}$ , 通过  $r_{11}$  概率的取值直接解决问题, 而没有考虑其它变量的可能性.

### 2.4.3 隐私问题的调查\*

现实生活中的每个人都有一些隐私或秘密, 相关信息不希望被外人知晓, 然而对于一些具有社会普遍性的隐私问题, 我们需要对此进行一定的了解和调查, 例如在校大学生有抑郁倾向的同学占有多少比例, 家庭不和谐的同学占有多少比例, 等等. 这些信息属于个人隐私不便直接调查, 需要设计一种好的方案, 使被调查者愿意作出真实回答、又能较好地保护个人隐私.

经过多年研究与实践, 心理学家和统计学家设计了一种巧妙的方案, 核心是如下两个问题:

**[问题 A:]** 你的生日是否在 7 月 1 日之前?

**[问题 B:]** 你是否有抑郁的倾向?

再准备一个箱子, 里面装有  $m$  个白球和  $n$  个红球. 被调查者随机抽取一球, 若抽到白球回答问题 A, 否则回答问题 B. 在问卷的答案上只有两选项: “是”或“否”, 无论哪个问题都只需选择“是”或“否”, 最后将答卷放入一个投票箱内密封.

上述的抽球与回答过程都在一间无人的房间内进行, 任何外人都不知道被调查者抽到什么颜色的球, 也不知道被调查者的答案, 以此保护个人隐私. 如果向被调查者解释清楚了该调查方案并严格执行, 那么被调查者很容易确信他/她参加这次调查不会泄露个人隐私, 从而愿意配合调查.

当有  $N > 500$  位学生参加调查后, 就可以打开投票箱进行统计. 设有  $N_y$  张答卷选择“是”, 根据频率与概率的关系有

$$P(\text{一个学生回答“是”}) \approx N_y/N.$$

设一个学生有抑郁倾向的概率为  $p$ , 即

$$P(\text{一个学生回答“是”}|\text{红球}) = p.$$

不妨假设每个学生的生日是等可能事件, 因此一个学生在 7 月 1 日之前出生的概率为  $1/2$ , 即

$$P(\text{一个学生回答“是”}|\text{白球}) = 1/2.$$

根据全概率公式有

$$P(\text{一个学生回答“是”}) = P(\text{一个学生回答“是”}|\text{红球})P(\text{红球}) + P(\text{一个学生回答“是”}|\text{白球})P(\text{白球}).$$

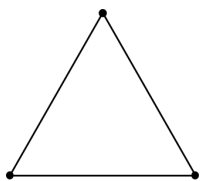
由此可得

$$\frac{N_y}{N} \approx \frac{m}{m+n} \times \frac{1}{2} + \frac{n}{m+n} \times p,$$

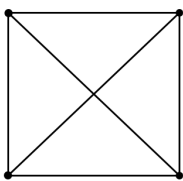
进一步估计出具有抑郁倾向的学生比例为  $p \approx (m+n)N_y/nN - m/2n$ .

#### 2.4.4 完全图着色\*

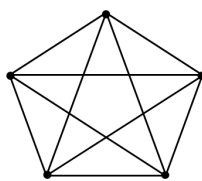
设平面上有  $n$  个顶点, 其中任意三个顶点不在同一条直线上, 用  $n(n-1)/2$  条边将这些顶点连接起来的图称为  $n$  个顶点的 **完全图**, 例如三个、四个、五个顶点的完全图如下所示:



三个顶点的完全图



四个顶点的完全图



五三个顶点的完全图

将图中的每条边都分别染成红色或蓝色, 给定两正整数  $n \geq 10$  和  $k > n/2$ , 是否存在一种染色方法, 使得图上任意  $k$  个顶点相对应的  $k(k-1)/2$  条边不是同一颜色?

我们利用概率的方法来求解该问题: 假设每条边等可能独立地被染成红色或蓝色, 即每条边为红色或为蓝色的概率均为  $1/2$ . 从  $n$  个不同顶点中选出  $k$  个顶点有  $\binom{n}{k}$  种不同的选法, 分别对应于  $\binom{n}{k}$  个包含有  $k$  个顶点的子集, 这里将  $k$  个顶点的子集分别标号为  $1, 2, \dots, \binom{n}{k}$ .

用  $E_i$  表示第  $i$  个子集中  $k(k-1)/2$  条边染成相同颜色的事件, 根据题意可得

$$P(E_i) = 2(1/2)^{k(k-1)/2} \quad i = 1, 2, \dots, \binom{n}{k}.$$

若存在  $k$  个顶点, 其相应的  $k(k-1)/2$  条边是同一颜色的事件可表示为  $\bigcup_{i=1}^{\binom{n}{k}} E_i$ . 根据布尔不等式有

$$P\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq \sum_{i=1}^{\binom{n}{k}} P(E_i) = \binom{n}{k} (1/2)^{k(k-1)/2-1}.$$

当  $n \geq 10$  和  $k > 2/n$  时有  $P\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq 1$ , 因此事件“完全图中任意  $k$  个顶点, 其相应的  $k(k-1)/2$  条边不是同一颜色”的概率大于零. 这意味着至少存在一种染色方法, 使得对任意  $k$  顶点集合所对应的  $k(k-1)/2$  边染色不全相同.

这种将概率用于求解纯粹确定性问题的方法称为 **概率化方法** (probabilistic method), 在计算机或人工智能中证明存在性时经常用到.

上述分析说明了完全图染色满足要求的存在性, 但并没有告诉我们如何涂颜色: 一种方法是随机涂色, 然后检查所涂的颜色是否满足所要求的性质; 若不成再重复进行直到成功为止.

## 习题

- 2.1 阐述独立与互不相容的关系.
- 2.2 若事件  $A, B, C$  独立, 证明:  $A$  与事件  $B \cup C$  独立.
- 2.3 设事件  $A$  和  $B$  满足  $P(A)P(B) > 0$ , 证明: 若两事件独立则不互斥; 若两事件互斥则不独立.
- 2.4 小明同学参加时长为 1 小时的竞赛, 假设他在  $a \in [0, 1]$  小时内完成竞赛的概率为  $a/2$ . 已知小明在半小时后仍未完成, 求他最后要用完一小时的条件概率.
- 2.5 随机掷两次骰子, 已知第一次掷 6 点, 求两次点数之和不小于 10 的概率.
- 2.6 (三门问题) 在一电视节目中, 参赛者看到三扇关闭的门, 已知一门后面是汽车, 其它两门后面是山羊, 选中什么则获得什么, 主持人知道三门后有什么. 当参赛者选定一扇门但未开启, 此时节目主持人则开启剩下有山羊的一扇门. 问题: 若参赛者允许重新选择, 是否换一扇门?
- 2.7 已知事件  $A$  为病人被诊断为肝癌, 事件  $C$  为病人患有肝癌,  $P(A|C) = 0.95$ ,  $P(\bar{A}|\bar{C}) = 0.9$ ,  $P(C) = 0.0004$ . 求  $P(C|A)$ .
- 2.8 书上的习题: 书 26 页到 28 页: 22, 27, 28, 30, 31, 32, 33, 37, 39.

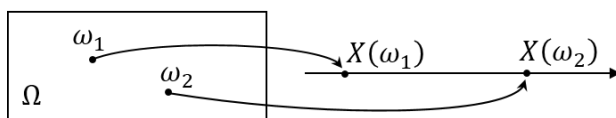


### 第3章 离散型随机变量

在很多随机现象中, 随机试验的结果可能与某些数值直接相关, 例如, 抛一枚骰子的点数分别为  $1, 2, \dots, 6$ ; 国家一年内出生的婴儿数; 一批出厂的产品中包含的废品数; 等等. 有些看起来与数值无关的随机现象, 也可以通过数值来描述, 例如在抛硬币试验中, 每次试验结果为正面或反面朝上, 与数值没无关, 我们可以用 1 表示‘正面朝上’, 用 0 表示‘反面朝上’, 通过数值进行描述. 针对更一般的随机事件  $A$ , 也可与数值产生联系, 如

$$X = \begin{cases} 1 & \text{如果事件 } A \text{ 发生,} \\ 0 & \text{如果事件 } A \text{ 不发生.} \end{cases}$$

针对随机现象中的每种结果  $\omega$  (即每个基本事件), 都能与实数值  $X(\omega)$  建立某种数值对应关系, 并且随着基本事件  $\omega$  的不同而  $X(\omega)$  的取值也不同, 称这样的函数  $X = X(\omega)$  为随机变量, 如下图所示:



**定义 3.1** 设  $\Omega$  是一个样本空间, 如果对每个基本事件  $\omega \in \Omega$ , 都对应于一个实数  $X(\omega)$ , 称这样的单射实值函数  $X(\omega) : \Omega \rightarrow \mathbb{R}$  为 **随机变量** (random variable), 一般简写为  $X$ .

随机变量  $X$  的取值随试验结果的不同而不同, 具有一定的随机性; 由于各试验结果的出现具有一定的概率,  $X$  的取值具有统计规律性, 因此随机变量与普通函数存在着本质的不同. 通过随机变量来描述随机现象或随机事件, 使得我们可以利用各种数学分析工具, 通过对随机变量的研究来分析随机现象. 可以用  $\{X \leq -\infty\}$  表示不可能事件, 以及  $\{X \leq +\infty\}$  表示必然事件. 一般用大写字母  $X, Y, Z$  表示随机变量. 下面给出一些随机变量的例子:

- 抛一枚骰子, 用随机变量  $X$  表示出现的点数, 则随机变量  $X \in \{1, 2, 3, 4, 5, 6\}$ . 出现的点数不超过 4 的事件可表示为  $\{X \leq 4\}$ ; 出现偶数点的事件可表示为  $\{X = 2, 4, 6\}$ .
- 用随机变量  $X$  表示一盏电灯的寿命, 其取值为  $[0, +\infty)$ , 电灯寿命不超过 500 小时的事件可表示为  $\{X \leq 500\}$ .

根据取值的类型, 可将随机变量分为离散型随机变量和非离散型随机变量. 若随机变量  $X$  的取值是有限的、或无限可列的, 则称  $X$  为 **离散型随机变量**; 若随机变量  $X$  的取值是无限不可列的, 则称  $X$  为 **非离散型随机变量**. 本章主要研究离散型随机变量.

### 3.1 离散型随机变量及分布列

离散型随机变量的取值是有限或无限可列的, 要完全刻画它的概率属性, 需要首先了解它所有可能的取值, 以及这些取值发生的概率.

**定义 3.2** 设随机变量  $X$  所有可能的取值为  $x_1, x_2, \dots, x_k, \dots$ , 事件  $\{X = x_k\}$  的概率为

$$p_k = P(X = x_k), \quad k = 1, 2, \dots,$$

称之为随机变量  $X$  的 **概率分布列** 或 **概率分布**, 简称 **分布列**.

概率分布列能一目了然的看出随机变量的取值以及相应的概率, 也可以通过下面的表格给出:

$X$	$x_1$	$x_2$	$\cdots$	$x_n$	$\cdots$
$P$	$p_1$	$p_2$	$\cdots$	$p_n$	$\cdots$

根据概率的非负性和完备性可知分布列应具有如下性质:

**性质 3.1** 随机变量  $X$  的分布列  $p_k = P(X = x_k)$  满足  $p_k \geq 0$  和  $\sum_k p_k = 1$ .

反之, 任何满足上面两条性质的数列  $\{p_k\}$ , 都可以作为一个随机变量的分布列.

**例 3.1** 设随机变量  $X$  的分布列  $P(X = k) = c/4^k$  ( $k = 0, 1, 2, \dots$ ), 求概率  $P(X = 1)$ .

**解** 根据概率的完备性有

$$1 = \sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{c}{4^k} = \frac{4}{3}c,$$

求解得到  $c = 3/4$ , 进一步有  $P(X = 1) = 3/16$ .

**例 3.2** 给定常数  $\lambda > 0$ , 随机变量  $X$  的分布列  $p(X = i) = c\lambda^i/i!$  ( $i \geq 0$ ), 求  $P(X > 2)$ .

**解** 根据概率的完备性有

$$1 = \sum_{i=0}^{\infty} P(X = i) = c \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = c \cdot e^{\lambda}$$

从而得到  $c = e^{-\lambda}$ , 进一步得到

$$P(X > 2) = 1 - P(X \leq 2) = 1 - p_0 - p_1 - p_2 = 1 - e^{-\lambda}(1 + \lambda + \lambda^2/2).$$

### 3.2 离散型随机变量的期望

针对一个具体的问题, 完全求解出概率分布列可能不是一件容易的事; 很多时候也不需要知道精确的概率分布列, 而是要掌握它的整体特征. 例如, 在统计某个地区的工资水平时, 我们可能更关心该地区工资的平均水平、贫富差距等特征, 而不是每个人的具体工资. 这些刻画随机变量某些方面特征的数值称为 **随机变量的数字特征**.

数字特征在概率统计中起着重要的作用, 它从宏观的角度刻画了随机变量某些基本特性, 有助于对随机变量的总体理解. 针对一些常用的随机变量, 我们可能只需要知道他们的一些数字特征, 就可以完全确定其概率分布. 常用的数字特征包括随机变量的期望、方差、相关系数和矩等, 本节介绍离散型随机变量的期望和性质, 其它数字特征在后续章节中介绍.

**定义 3.3** 设离散型随机变量  $X$  的分布列为  $p_k = P(X = x_k) (k \geq 1)$ . 若级数

$$\sum_{k=1}^{\infty} p_k x_k$$

绝对收敛, 则称该级数和为随机变量  $X$  的 **期望** (expectation), 又称为 **均值** (mean), 记为  $E(X)$ , 即

$$E(X) = \sum_{k=1}^{\infty} p_k x_k .$$

期望  $E(X)$  反映随机变量  $X$  的平均值, 由随机变量的分布列决定, 是常量而不是变量, 其本质是随机变量的取值  $x_i$  根据概率  $p_i$  加权所得. 级数的绝对收敛确保了期望  $E(X)$  的唯一性, 即级数和不会随级数各项次序的改变而改变. 除非特别说明, 我们通常都直接利用定义计算期望, 不需考虑其绝对收敛性.

**例 3.3** 随意掷一枚骰子,  $X$  表示观察到的点数, 求  $E[X]$ .

**解** 随机变量  $X$  的取值为  $1, 2, \dots, 6$ , 且每点等可能发生, 其分布列为  $P(X = k) = 1/6, k \in [6]$ . 因此随机变量  $X$  的期望为  $E(X) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 7/2$ .

我们来看一个期望不存在的例子:

**例 3.4** 设随机变量  $X$  的分布列为  $P(X = (-2)^k/k) = 1/2^k, k = 1, 2, \dots$ , 求期望  $E(X)$ .

**解** 尽管根据定义有

$$E(X) = \sum_{k=1}^{+\infty} P\left(X = \frac{(-2)^k}{k}\right) \frac{(-2)^k}{k} = \sum_{k=1}^{+\infty} \frac{(-1)^k}{k} = -\ln 2.$$

但是

$$\sum_{k=1}^{+\infty} P\left(X = \frac{(-2)^k}{k}\right) \left|\frac{(-2)^k}{k}\right| = \sum_{k=1}^{+\infty} \frac{1}{k} \rightarrow +\infty.$$

该级数并非绝对收敛, 其级数和可能随着求和顺序的改变而改变, 级数和并非唯一的数值, 故该随机变量的期望  $E(X)$  不存在.

**例 3.5** 若  $n$  把钥匙中只有一把能开门, 现随机选取一把钥匙开门, 若打不开门则去掉该钥匙, 再随机选取剩下的钥匙进行尝试, 求打开门需要尝试的平均次数.

**解** 设随机变量  $X$  表示尝试开门的次数, 其分布列为

$$P(X = k) = \frac{\binom{n-1}{k-1}}{\binom{n}{k-1}} \cdot \frac{1}{n-k+1} = \frac{1}{n} \quad k \in [n].$$

因此打开门需要尝试的平均次数

$$E(X) = \sum_{k=1}^n \frac{k}{n} = \frac{(1+n)n}{2n} = \frac{n+1}{2}.$$

下面介绍期望的一些性质, 除非特别说明, 这些性质不仅对离散型随机变量成立, 对其它任何类型的随机变量都成立. 为了证明的可读性, 在证明过程中仅考虑离散型随机变量.

**性质 3.2** 设  $c \in \mathbb{R}$  是常数, 若随机变量  $X \equiv c$ , 则  $E(X) = c$ .

**性质 3.3** 若随机变量  $X$  的取值非负, 即  $X \geq 0$ , 则  $E(X) \geq 0$ .

**性质 3.4** 对随机变量  $X$  和常数  $a, b \in \mathbb{R}$ , 有  $E(aX + b) = aE(X) + b$ .

**证明** 设随机变量  $X$  的分布列为  $p_k = P(X = x_k)$ , 则随机变量  $Y = aX + b$  的分布列为  $p_k = P(Y = ax_k + b)$ , 进而有

$$E[aX + b] = \sum_{k \geq 1} (ax_k + b)p_k = a \sum_{k \geq 1} x_k p_k + b \sum_{k \geq 1} p_k = aE[X] + b.$$

**性质 3.5** 若离散型随机变量  $X$  所有可能的取值为非负整数  $\{0, 1, 2, \dots\}$ , 则

$$E(X) = \sum_{i=1}^{+\infty} P(X \geq i).$$

**证明** 根据期望的定义有

$$E[X] = \sum_{j=1}^{+\infty} jP(X = j) = \sum_{j=1}^{+\infty} \sum_{i=1}^j P(X = j) = \sum_{i=1}^{+\infty} \sum_{j=i}^{+\infty} P(X = j) = \sum_{i=1}^{+\infty} P(X \geq i),$$

由此完成证明.

针对随机变量的函数的期望, 有如下定理:

**定理 3.1** 设离散型随机变量  $X$  的分布列为  $p_k = P(X = x_k)$  ( $k \geq 1$ ). 对任意的实值函数  $g: \mathbb{R} \rightarrow \mathbb{R}$ , 若级数  $\sum_{k \geq 1} g(x_k)p_k$  绝对收敛, 则有

$$E[g(X)] = \sum_{k=1}^{\infty} g(x_k)p_k.$$

**证明** 证明的核心思想是利用无穷级数的绝对收敛确保任意重排后的级数和等于原级数和. 根据题意有  $X$  的分布列为  $p_k = P(X = x_k)$  和随机变量函数  $Y = g(X)$  有

$X$	$x_1$	$x_2$	$\cdots$	$x_n$	$\cdots$
$P$	$p_1$	$p_2$	$\cdots$	$p_n$	$\cdots$
$Y$	$y_1$	$y_2$	$\cdots$	$y_n$	$\cdots$

其中  $y_i = g(x_i)$ . 上面的表格给出了随机变量  $X$  的分布列, 但并没给出随机变量  $Y$  的分布列, 因为可能存在  $y_i = g(x_i) = y_j = g(x_j)$ . 为了得到随机变量  $Y$  的分布列, 需要将  $x_1, x_2, \dots, x_n, \dots$  进行重新排列分组为

$$\underbrace{x_{1,1}, x_{1,2}, \dots, x_{1,k_1}}_{y'_1 = g(x_{1,j}) \ (j \in [k_1])}, \underbrace{x_{2,1}, x_{2,2}, \dots, x_{2,k_2}}_{y'_2 = g(x_{2,j}) \ (j \in [k_2])}, \dots, \underbrace{x_{n,1}, x_{n,2}, \dots, x_{n,k_n}}_{y'_n = g(x_{n,j}) \ (j \in [k_n])}, \dots$$

满足当  $i \neq j$  时有  $y'_i \neq y'_j$  成立. 由此可得随机变量  $Y$  的分布列为

$$P(Y = y'_i) = \sum_{j=1}^{k_i} p_{i,j} = \sum_{k \geq 1, y'_i = g(x_k)} p_k,$$

进一步得到随机变量  $Y$  的期望为

$$E[Y] = \sum_{i=1}^{\infty} y'_i P[Y = y'_i] = \sum_{i=1}^{\infty} y'_i \sum_{j=1}^{k_i} p_{i,j} = \sum_{i=1}^{\infty} \sum_{j=1}^{k_i} g(x_{i,j}) p_{i,j} = \sum_{k=1}^{\infty} g(x_k) p_k,$$

最后一个等式成立是因为绝对收敛的无穷级数在重排前与重排后其级数和不变.

基于上述定理, 我们可以直接计算随机变量  $Y = g(X)$  的期望, 而不需要知道  $Y$  的分布列, 即通过  $X$  的分布列计算期望  $E[Y]$ . 此外基于该定理有

**推论 3.1** 设  $X$  是离散型随机变量, 以及  $g_i: \mathbb{R} \rightarrow \mathbb{R}$  是实值函数 ( $i \in [n]$ ). 若期望  $E(g_i(X))$  存在, 则对任意常数  $c_1, c_2, \dots, c_n$  有  $E(\sum_{i=1}^n c_i g_i(X)) = \sum_{i=1}^n c_i E(g_i(X))$  成立.

基于此推论很容易得到

$$E(X^4 + \sin(X) + 4) = E(X^4) + E(\sin(X)) + 4.$$

最后探讨当函数  $g(x)$  满足什么样的性质时, 期望  $E(g(X))$  和  $g(E(X))$  之间都存在一定的比较关系. 相关的知识在实际应用和科研中具有重要意义, 因为即使不知道随机变量的具体概率分布, 仍可以对期望进行一定的估计或推理. 看一个例子: 设离散型随机变量  $X$  的分布列为  $P(X=1)=P(X=2)=P(X=0)=1/3$ , 很容易发现

$$(E(X))^2 \leq E(X^2) \quad \text{和} \quad \sqrt{E(X)} \geq E(\sqrt{X}).$$

针对更一般的情况, 考虑两类函数:

**定义 3.4** 设函数  $g: [a, b] \rightarrow \mathbb{R}$ , 对任意  $x_1, x_2 \in [a, b]$  和  $\lambda \in [0, 1]$ ,

- 若  $g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2)$ , 则称函数  $g(x)$  是定义在  $[a, b]$  上的 **凸函数**;
- 若  $g(\lambda x_1 + (1-\lambda)x_2) \geq \lambda g(x_1) + (1-\lambda)g(x_2)$ , 则称函数  $g(x)$  是定义在  $[a, b]$  上的 **凹函数**.

凸函数和凹函数具有很多良好的数学性质, 例如凸函数的一阶导数单调性、二阶导数小于或等于零, 大家可以参考一些数学分析或优化书籍. 下面介绍著名的 **琴生不等式** (Jensen's inequality).

**定理 3.2** 设随机变量  $X \in [a, b]$  和实值函数  $g: [a, b] \rightarrow \mathbb{R}$ ,

- 若  $g(x)$  在  $[a, b]$  上是凸函数, 则有  $g(E(X)) \leq E(g(X))$ ;
- 若  $g(x)$  在  $[a, b]$  上是凹函数, 则有  $g(E(X)) \geq E(g(X))$ .

定理 3.2 中的不等式对所有的随机变量都成立. 即使在不知道随机变量  $X$  的概率分布情况下, 根据该定理可知

$$(E(X))^2 \leq E(X^2), \quad \sqrt{E(X)} \geq E(\sqrt{X}) \quad \text{和} \quad e^{E(X)} \leq E(e^X).$$

**证明** 这里仅给出离散型随机变量具有有限个取值和凸函数的证明. 设随机变量  $X$  的取值为  $x_1, x_2, \dots, x_n$ , 以及它的分布列为  $p_k = P(X = x_k) > 0$ , 易知  $\sum_k p_k = 1$ . 我们需要证明的不等式为

$$g(p_1 x_1 + p_2 x_2 + \dots + p_n x_n) \leq p_1 g(x_1) + p_2 g(x_2) + \dots + p_n g(x_n). \quad (3.1)$$

针对上式采用归纳法证明, 当  $n=2$  时利用凸函数的定义结论显然成立. 不妨假设当  $n=m-1$  时 (3.1) 成立, 下面证明当  $n=m$  时 (3.1) 亦成立. 首先有

$$\begin{aligned} g(p_1 x_1 + p_2 x_2 + \dots + p_m x_m) &= g\left(p_1 x_1 + (1-p_1) \left[ \frac{p_2}{1-p_1} x_2 + \dots + \frac{p_m}{1-p_1} x_m \right]\right) \\ &\leq p_1 g(x_1) + (1-p_1) g\left(\frac{p_2}{1-p_1} x_2 + \dots + \frac{p_m}{1-p_1} x_m\right), \end{aligned}$$

这里将凸函数的定义应用到两个点  $x_1$  和  $x'_1 = (x_2 p_2 + \dots + x_m p_m)/(1-p_1)$ . 容易发现  $p_i/(1-p_1) \geq 0$

且  $\sum_{i=2}^m p_i/(1-p_1) = 1$ , 根据归纳假设有

$$g\left(\frac{p_2}{1-p_1}x_2 + \cdots + \frac{p_m}{1-p_1}x_m\right) \leq \frac{p_2}{1-p_1}g(x_2) + \cdots + \frac{p_m}{1-p_1}g(x_m),$$

由此可完成证明.

### 3.3 离散型随机变量的方差

数学期望反映了随机变量的平均值, 在很多实际应用中我们不仅仅要知道随机变量的平均值, 还需要进一步了解随机变量的取值与期望之间的偏离程度. 例如, 考虑三个随机变量  $X, Y$  和  $Z$ , 它们的分布列分别为

$$P(X=0)=1; \quad P(Y=1)=P(Y=-1)=1/2; \quad P(Z=2)=1/5, P(Z=-1/2)=4/5.$$

容易得到  $E(X)=E(Y)=E(Z)=0$ , 即三个随机变量的期望相同. 然而很显然这三个随机变量存在着明显的差异, 如何刻画它们的不同之处, 可以考虑三个随机变量的取值与期望的偏离程度, 即本节所研究随机变量的方差.

**定义 3.5** 设离散随机变量  $X$  的分布列为  $p_k = P(X=x_k) > 0$ , 若期望  $E(X)$  和  $E(X-E(X))^2$  存在, 则称  $E(X-E(X))^2$  为随机变量  $X$  的 **方差** (variance), 记为  $\text{Var}(X)$ , 即

$$\text{Var}(X) = E(X-E(X))^2 = \sum_k p_k (x_k - E(X))^2 = \sum_k p_k \left( x_k - \sum_k x_k p_k \right)^2. \quad (3.2)$$

称  $\sqrt{\text{Var}(X)}$  为 **标准差** (standard deviation), 记为  $\sigma(X)$ .

结合方差的定义和期望的性质有

$$\begin{aligned} \text{Var}(X) &= E(X-E(X))^2 = E(X^2 - 2XE(X) + E^2(X)) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 = E(X^2) - (E(X))^2, \end{aligned}$$

由此给出方差的另一种定义

$$\text{Var}(X) = E(X^2) - (E(X))^2. \quad (3.3)$$

尽管方差的两种定义等价, 然而在实际应用中却存在着不同的用处, (3.3) 给出了方差的物理含义, 而 (3.2) 更有利于方差的计算, 例如,

**例 3.6** 设随机变量  $X$  的分布列为  $P(X=x_k) = 1/n$  ( $k \in [n]$ ), 这里  $x_1, x_2, \dots, x_n$  的数值各不相同, 需遍历数据几次才能计算出方差  $\text{Var}(X)$ .

**解** 若采用  $\text{Var}(X) = E(X-E(X))^2$ , 则需要遍历数据  $x_1, x_2, \dots, x_n$  两次, 第一次计算期望  $E(X)$ , 第二次计算方差  $\text{Var}(X)$ .

若采用  $\text{Var}(X) = E(X^2) - (E(X))^2$ , 则只需要遍历数据  $x_1, x_2, \dots, x_n$  一次, 在遍历数据的过程中计算  $E(X^2)$  和  $(E(X))^2$ , 从而计算方差. 在此过程中也不需要全部数据  $x_1, x_2, \dots, x_n$  存在内存中, 可以一个个轮流存取数据.

下面给出方差的性质:

**性质 3.6** 设  $c \in \mathbb{R}$  是常数, 若随机变量  $X \equiv c$ , 则  $\text{Var}(X) = 0$ .

**性质 3.7** 对随机变量  $X$  和常数  $a, b \in \mathbb{R}$ , 有

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

**证明** 根据期望的性质有  $E(aX + b) = aE(X) + b$ , 代入可得

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

值得注意的是, 方差通常不具有线性性, 即  $\text{Var}(f(X) + g(X)) \neq \text{Var}(f(X)) + \text{Var}(g(X))$ .

**性质 3.8** 对随机变量  $X$  和常数  $a \in \mathbb{R}$ , 有

$$\text{Var}(X) = E(X - E(X))^2 \leq E(X - a)^2.$$

**证明** 根据期望的性质有

$$\begin{aligned} E(X - c)^2 &= E(X - E(X) + E(X) - c)^2 \\ &= E(X - E(X))^2 + E[(X - E(X))(E(X) - c)] + (E(X) - c)^2 \\ &= E(X - E(X))^2 + (E(X) - c)^2 \\ &\geq E(X - E(X))^2, \end{aligned}$$

从而完成证明.

**定理 3.3 (Bhatia-Davis不等式)** 对随机变量  $X \in [a, b]$ , 有

$$\text{Var}[X] \leq (b - E(X))(E(X) - a) \leq (b - a)^2/4.$$

**证明** 对任意随机变量  $X \in [a, b]$ , 有

$$(b - X)(X - a) \geq 0,$$

两边同时对随机变量取期望, 整理可得

$$E(X^2) \leq (a + b)E(X) - ab.$$



根据方差的定义有

$$\text{Var}(X) = E(X^2) - (E(X))^2 = -(E(X))^2 + (a+b)E(X) - ab = (b - E(X))(E(X) - a).$$

利用二次函数  $f(t) = (b-t)(t-a) = -t^2 + (a+b)t - ab$  的最大值可得

$$(b - E(X))(E(X) - a) \leq (b - a)^2/4.$$

### 3.4 常用离散型随机变量

本节介绍几种常用的离散型随机变量, 并研究它们的性质.

#### 3.4.1 0-1分布

0-1分布是概率统计中最经典、最简单的分布, 是很多概率模型的基础.

**定义 3.6** 设随机变量  $X$  的分布列  $P(X=1)=p$ ,  $P(X=0)=1-p$ , 等价于

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

则称随机变量  $X$  服从 **参数为  $p$  的 0-1 分布**, 又称 **两点分布**, 或 **伯努利分布** (Bernoulli distribution), 记  $X \sim \text{Ber}(p)$ .

0-1 分布也可以通过表格表示为

$X$	0	1
$P$	$1-p$	$p$

根据定义容易得到

**引理 3.1** 若随机变量  $X \sim \text{Ber}(p)$ , 则有  $E(X) = p$  和  $\text{Var}(X) = p(1-p)$ .

由此可知 0-1 分布也可由它的数学期望唯一确定.

若一次试验只考虑事件  $A$  发生或不发生两种情况, 称这样的试验为 **伯努利试验**, 可以通过 0-1 分布来描述伯努利试验:

$$X = \begin{cases} 1 & \text{若事件 } A \text{ 发生,} \\ 0 & \text{否则.} \end{cases}$$

此时容易得到  $E[X] = P(A)$ , 即随机变量  $X$  的期望等于事件  $A$  发生的概率.

#### 3.4.2 二项分布

伯努利试验只考虑事件  $A$  发生或不发生两种结果, 不妨设事件  $A$  发生的概率  $P(A) = p \in (0, 1)$ . 将一个伯努利试验独立重复地进行  $n$  次, 称这一系列重复的独立试验为  **$n$  重伯努利试验**. 它是一种非常重要的概率模型, 衍生出很多的概率分布.

在  $n$  重伯努利试验中, 我们关心随机事件  $A$  发生了多少次, 用随机变量  $X$  表示, 其所有可能的取值为  $0, 1, 2, \dots, n$ . 随机事件  $\{X = k\}$  表示在  $n$  重伯努利试验中事件  $A$  发生了  $k$  次, 到底是哪  $k$  次发生的, 共有  $\binom{n}{k}$  种不同的情况. 针对一种具体的情况, 不妨设前  $k$  次事件  $A$  发生, 后  $n - k$  次事件  $A$  不发生, 此种情况发生的概率为

$$\underbrace{p \times p \times \cdots \times p}_k \times \underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{n-k} = p^k (1-p)^{n-k}.$$

由此可知在  $n$  重伯努利试验中事件  $A$  发生了  $k$  次的概率为  $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ .

**定义 3.7** 若随机变量  $X$  的分布列为

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, 2, \dots, n, \quad (3.4)$$

则称随机变量  $X$  服从 **参数为  $n$  和  $p$  的二项分布** (binomial distribution), 记  $X \sim B(n, p)$ .

容易发现 (3.4) 中  $P(X = k)$  是二项式  $(1-p+xp)^n$  展开式中  $x^k$  项的系数, 所以该分布被称为二项分布. 进一步可检验

$$\sum_{k=0}^n P(X = k) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + 1-p)^n = 1.$$

若  $n = 1$ , 则二项分布退化为 0-1 分布, 即  $B(1, p) = \text{Ber}(p)$ . 关于二项分布的数字特征有

**引理 3.2** 若随机变量  $X \sim B(n, p)$ , 则有  $E(X) = np$  和  $\text{Var}(X) = np(1-p)$ .

若知道二项分布的期望和方差, 可反解出参数  $n$  和  $p$ , 因此二项分布可由它的期望和方差唯一确定.

**证明** 根据定义有

$$E(X) = \sum_{k=0}^n P(X = k)k = \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k} = (1-p)^n \sum_{k=1}^n \binom{n}{k} k \left(\frac{p}{1-p}\right)^k.$$

对二项展开式  $(1+x)^n = \sum_{k=0}^n \binom{n}{k} x^k$  两边同时求导后乘  $x$  可得

$$nx(1+x)^{n-1} = \sum_{k=1}^n \binom{n}{k} k x^k,$$

将  $x = p/(1-p)$  代入上式可得

$$E(X) = (1-p)^n \sum_{k=0}^n \binom{n}{k} k \left(\frac{p}{1-p}\right)^k = (1-p)^n \frac{np}{1-p} \frac{1}{(1-p)^{n-1}} = np.$$

对于方差, 首先计算

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

对二项展开式  $(1+x)^n = \sum_{k=0}^n \binom{n}{k} x^k$  两边同时求导两次后乘  $x^2$  可得

$$n(n-1)x^2(1+x)^{n-2} = \sum_{k=2}^n \binom{n}{k} k(k-1)x^k,$$

将  $x = p/(1-p)$  带入上式有

$$E(X^2) = n(n-1)p^2 + np = n^2p^2 + np(1-p),$$

从而得到  $\text{Var}(X) = E[X^2] - (E[X])^2 = np(1-p)$ .

下面给出几个二项分布的概率分布示意图:

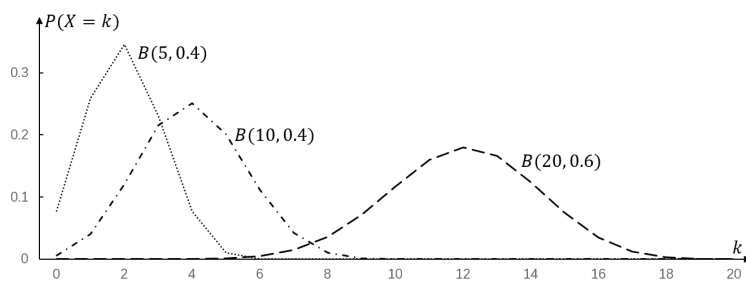


图 3.1 二项分布的概率分布示意图

根据上述分布图可知: 若随机变量  $X \sim B(n, p)$ , 则有  $P(X = k)$  从一开始单调递增, 然后一直单调递减, 一般在期望  $np$  附近的整数点取得最大值. 也可严格证明: 当  $k \in [0, np + p]$  时  $P(X = k)$  单调递增, 当  $k \in [np + p, n]$  时  $P(X = k)$  单调递减.

**例 3.7** 假设有两个箱子, 每个箱子里放了  $n$  个球, 现在任意选取一个箱子拿走其中一球 (不放回), 重复这一过程, 求一个箱子中的球拿光而另一个箱子还剩下  $r$  个球的概率.

**解** 两个箱子分别被表示为第一个箱子和第二个箱子, 考虑的伯努利试验: 在箱子选取过程中是否选取第一个箱子? 用事件  $A$  表示选取第一个箱子, 根据题意有  $P(A) = 1/2$ . 因此可以共发现进行了  $2n - r$  重伯努利试验, 用  $X$  表示事件  $A$  发生的次数, 于是有

$$X \sim B(2n - r, 1/2).$$

最后所求概率为

$$\begin{aligned} & P(X = n) + P(X = n - r) \\ &= \binom{2n-r}{n} (1/2)^n (1/2)^{n-r} + \binom{2n-r}{n-r} (1/2)^{n-r} (1/2)^n = \binom{2n-r}{n} / 2^{2n-r-1}, \end{aligned}$$

由此完成证明.

**例 3.8** 一个系统由  $n$  个独立的元件组成, 每个元件能正常工作的概率为  $p$ , 若该系统中至少有一半的元件能正常工作则整个系统有效, 在什么情况下 5 个元件的系统比 3 个元件的系统更有效?

**解** 用  $X$  表示  $n$  个元件能正常工作的元件数, 则有  $X \sim B(n, p)$ . 由此可知包含有 5 个元件的系统有效的概率为

$$\binom{5}{3} p^3 (1-p)^2 + \binom{5}{4} p^4 (1-p) + \binom{5}{5} p^5 = p^3 (6p^2 - 15p + 10),$$

而包含有 3 个元件的系统有效的概率为

$$\binom{3}{2} p^2 (1-p) + \binom{3}{3} p^3 = p^2 (3 - 2p).$$

当  $p^3 (6p^2 - 15p + 10) > p^2 (3 - 2p)$  时, 即当  $3(p-1)^2 (2p-1) > 0$  时 5 个元件的系统比 3 个元件的系统更有效, 此时  $p > 1/2$ .

### 3.4.3 泊松分布

泊松分布是概率论中另一种重要的分布, 用于描述大量试验中稀有事件出现次数的概率模型. 例如, 一个月内网站的访问量, 一个小时内公共汽车站来到的乘客数, 书中一页出现错误的语法数, 一天中银行办理业务的顾客数, 一年内中国发生的地震次数等.

**定义 3.8** 给定常数  $\lambda > 0$ , 若随机变量  $X$  的分布列为

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

则称随机变量  $X$  服从 **参数为  $\lambda$  的泊松分布** (Poisson distribution), 记为  $X \sim P(\lambda)$ .

容易验证  $P(X = k) = \lambda^k e^{-\lambda} / k! \geq 0$ , 并根据指数的泰勒展式  $e^x = \sum_{k=0}^{\infty} x^k / k!$  有

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

关于泊松分布的数字特征有:

**引理 3.3** 若随机变量  $X \sim P(\lambda)$ , 则有  $E(X) = \lambda$  和  $\text{Var}(X) = \lambda$ .

因此泊松分布可由期望或方差唯一确定.

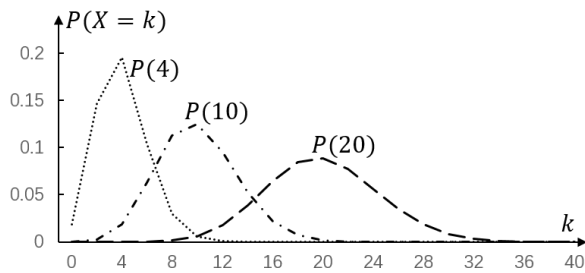
**证明** 根据指数的泰勒展开式有  $e^x = \sum_{k=0}^{\infty} x^k/k!$  有

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

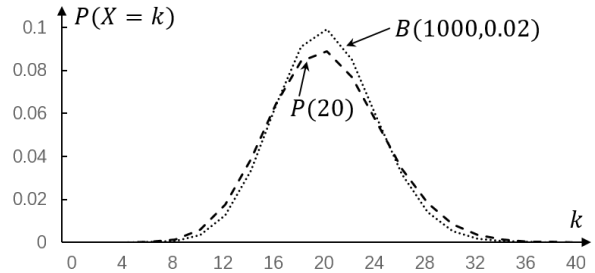
对于随机变量的方差, 首先计算

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

从而得到  $\text{Var}(X) = E[X^2] - (E[X])^2 = \lambda$ .



a) 泊松分布示意图



b) 二项分布  $B(1000, 0.02)$  与泊松分布  $P(20)$  的逼近性

图 3.2 泊松分布示意图、以及泊松分布与二项分布的逼近图

从图 3.2(a) 中可以观察发现: 若随机变量  $X \sim P(\lambda)$ , 则有  $P(X=k)$  从一开始单调递增, 然而一致单调递减, 在期望  $\lambda$  附近取得最大值. 其次, 泊松分布与二项分布的分布图之间有一定的相似性, 如图 3.2(b) 所示, 下面的定理给出了二者之间的近似关系:

**定理 3.4 (泊松定理)** 设  $\lambda > 0$  任意给定的常数,  $n$  是一个正整数, 若  $np_n = \lambda$ , 则对任意给定的非负整数  $k$ , 有

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

**证明** 由  $p_n = \lambda/n$ , 有

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

当  $n \rightarrow \infty$  时有  $(1 - \frac{\lambda}{n})^{\frac{n-k}{\lambda} \lambda} \rightarrow e^{-\lambda}$  以及  $\frac{n-k}{n} \lambda \rightarrow \lambda$ , 从而完成证明.

泊松分布的应用: 若随机变量  $X \sim B(n, p)$ , 当  $n$  比较大而  $p$  比较小时, 令  $\lambda = np$ , 有

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}.$$

即利用泊松分布近似计算二项分布. 针对彩票中奖、火山爆发、洪水泛滥、意外事故等小概率事件, 当试验的次数较多时, 可以将  $n$  重伯努利试验中小概率事件发生的次数近似服从泊松分布.

**例 3.9** 设有 80 台同类型设备独立工作, 每台发生故障的概率为 0.01, 一台设备发生故障时只能由一人处理, 考虑两种方案: I) 由四人维护, 每人单独负责 20 台; II) 由三人共同维护 80 台. 哪种方案更为合理?

**解** 首先讨论方案 I), 用事件  $A_i$  表示第  $i$  人负责的设备发生故障不能及时维修, 用  $X_i$  为第  $i$  人负责的 20 台设备同一时刻发生故障的台数, 则有  $X \sim B(20, 0.01)$ , 根据泊松定理有近似有  $X \sim P(0.2)$ , 进一步有

$$P(A_i) = P(X_i \geq 2) = 1 - P(X = 0) - P(X = 1) \approx 1 - \sum_{k=0}^2 \frac{(0.2)^k}{k!} e^{-0.2} \approx 0.0175.$$

因四人独立维修, 有设备发生故障时而不能及时的概率

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) \geq P(A_1) \approx 0.0175.$$

对方案 II): 设随机变量  $Y$  为 80 台设备中同一时刻发生故障的台数, 则  $Y \sim B(80, 0.01)$ , 根据泊松定理有近似有  $Y \sim P(0.8)$ , 则有设备发生故障不能及时维修的概率为

$$P(Y \geq 4) = 1 - \sum_{k=0}^3 P(Y = k) \approx 1 - \sum_{k=0}^3 \frac{(0.8)^k}{k!} e^{-0.8} \approx 0.0091.$$

由此比较可知方案 II) 更优.

**例 3.10** 一个公共汽车站有很多路公交车, 若一个时间段内到站的乘客数  $X \sim P(\lambda)$  ( $\lambda > 0$ ), 所有到站的乘客是相互独立的、且选择 D1 路公交车的概率为  $p$  ( $p > 0$ ), 求乘坐 D1 路公交车的乘客数  $Y$  的分布.

**解** 设一个时间段内到站的乘客数为  $k$ , 该事件发生的概率

$$P(X = k) = \lambda^k e^{-\lambda} / k!.$$

根据题意可知到达公交站的  $k$  个人中乘坐 D1 的人数服从参数为  $k$  和  $p$  的二项分布  $B(k, p)$ , 即

$$P(Y = i | X = k) = \binom{k}{i} p^i (1-p)^{k-i}.$$

根据全概率公式和指数函数  $e^x$  的泰勒展开式有

$$\begin{aligned}
 P(Y = i) &= \sum_{k=i}^{+\infty} P(X = k)P(Y = i|X = k) = p^i e^{-\lambda} \sum_{k=i}^{+\infty} \binom{k}{i} \frac{\lambda^k}{k!} (1-p)^{k-i} \\
 &= \frac{(p\lambda)^i e^{-\lambda}}{i!} \sum_{k=i}^{+\infty} \frac{((1-p)\lambda)^{k-i}}{(k-i)!} = \frac{(p\lambda)^i e^{-\lambda}}{i!} \sum_{k=0}^{+\infty} \frac{((1-p)\lambda)^k}{(k)!} \\
 &= \frac{(p\lambda)^i e^{-\lambda}}{i!} e^{(1-p)\lambda} = \frac{(p\lambda)^i e^{-p\lambda}}{i!},
 \end{aligned}$$

由此可知乘坐 D1 路公交车的乘客数  $Y \sim P(p\lambda)$ .

#### 3.4.4 几何分布

在多重 Bernoulli 试验中, 设事件  $A$  发生的概率为  $p$ . 用随机变量  $X$  表示事件  $A$  首次发生需要的试验次数, 事件  $\{X = k\}$  发生当且仅当事件  $A$  在前  $k-1$  次不发生而第  $k$  次发生, 根据多重 Bernoulli 试验的独立性可知概率  $P(X = k) = (1-p)^{k-1}p$ .

**定义 3.9** 设  $p \in (0, 1)$  是一个常数, 若随机变量  $X$  的分布列为

$$P(X = k) = (1-p)^{k-1}p \quad (k \geq 1), \quad (3.5)$$

称  $X$  服从 **参数为  $p$  的几何分布** (geometric distribution), 记  $X \sim G(p)$ .

容易得到  $P(X = k) \geq 0$  以及

$$\sum_{k=1}^{\infty} P(X = k) = p \sum_{k=1}^{\infty} (1-p)^{k-1} = p \times \frac{1}{1-(1-p)} = 1,$$

从而验证了 (3.5) 构成概率分布列. 几何分布有一个重要的性质: **无记忆性** (memoryless property).

**定理 3.5** 设随机变量  $X \sim G(p)$ , 对任意正整数  $m, n$ , 有

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

**证明** 根据几何分布的定义, 对任何正整数  $k$  有

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

根据条件概率的定义有

$$P(X > m+n | X > m) = \frac{P(X > m+n)}{P(X > m)} = \frac{(1-p)^{m+n}}{(1-p)^m} = (1-p)^n = P(X > n),$$

这里利用事件  $\{X > m + n\} \cap \{X > m\} = \{X > m + n\}$ , 从而完成证明.

几何分布无记忆性的直观解释: 假设以前经历了  $m$  次失败, 从当前起至成功的次数与  $m$  无关. 例如, 一人赌博时前面总输, 觉得下一次应该会赢了, 然而无记忆性告诉大家: 下一次是否会赢与前面输了多少次没有任何关系.

关于几何分布的数字特征, 我们有

**引理 3.4** 若随机变量  $X \sim G(p)$  ( $0 < p < 1$ ), 则有

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

**证明** 根据几何分布的定义有

$$P(X \geq i) = \sum_{k=i}^{+\infty} P(X = k) = p \sum_{k=i}^{+\infty} (1-p)^{k-1} = (1-p)^{i-1}.$$

对于非负整数的随机变量  $X$  有

$$E(X) = \sum_{i=1}^{+\infty} P(X \geq i) = \sum_{i=1}^{+\infty} (1-p)^{i-1} = 1/p.$$

对于随机变量  $X$  的方差, 首先计算

$$E(X^2) = \sum_{k=1}^{\infty} k^2 p (1-p)^{k-1} = p \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-1} + 1/p.$$

对级数展开式  $(1-x)^{-1} = \sum_{k=0}^{\infty} x^k$  两边先求二阶导后乘  $x$  有

$$\sum_{k=2}^{\infty} k(k-1)x^{k-1} = \frac{2x}{(1-x)^3}.$$

令  $x = 1-p$  代入可得  $E(X^2) = (2-p)/p^2$ . 最后有  $\text{Var}(X) = E(X^2) - (EX)^2 = (1-p)/p^2$ .

**例 3.11** 在古代非常重视生男孩但生存资源有限, 于是规定: 每个家庭可生一个男孩, 如果没有男孩则可以继续生育直至有一个男孩; 若已有一个男孩, 则不再生育. 不妨假设每个家庭生男孩的概率为  $p = 1/2$ , 问题: 1) 一个家庭恰好有  $n$  个小孩的概率; 2) 一个家庭至少有  $n$  个小孩的概率; 3) 男女比例是否会失衡?

**解** 用随机变量  $X$  表示一个家庭的小孩个数, 其取值为  $\{1, 2, \dots\}$ , 根据题意可知  $X$  服从参数为  $p = 1/2$  的几何分布, 因此一个家庭恰好有  $n$  个小孩的概率为

$$P(X = n) = p(1-p)^{n-1} = 1/2^n.$$



一个家庭至少有  $n$  个小孩的概率为

$$P(X \geq n) = \sum_{k=n}^{+\infty} P(X = k) = 1/2^{n-1}.$$

至于男女比例是否会失衡, 考虑一个家庭平均的孩子个数为  $E[X] = 1/p = 2$ , 由此可知在平均的情形下, 一个家庭的小孩男女比例 1:1, 因此不会造成男女失衡.

几何分布考虑在多重试验中事件  $A$  首次发生时所进行的试验次数, 可以进一步考虑事件  $A$  第  $r$  次发生时所进行的试验次数. 设随机事件  $A$  发生的概率为  $p \in (0, 1)$ , 用  $X$  表示事件  $A$  第  $r$  次成功时发生的试验次数, 则  $X$  取值  $r, r+1, r+2, \dots$ , 其分布列为

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

称随机变量  $X$  服从 **参数为  $r$  和  $p$  的负二项分布** 或 **帕斯卡分布**. 可以验证上述概率构成一个分布列, 以及随机变量  $X$  的期望  $E(X) = r/p$  和方差  $\text{Var}(X) = r(1-p)/p^2$ . 相关证明将作为练习题.

### 3.5 案例分析

#### 3.5.1 德国坦克问题

在二战期间, 同盟国一直在努力确定德国坦克的生产数量, 有助于对德国战力的评估. 这个问题可描述为: 德国生产了  $n$  辆坦克, 编号分别为  $1, 2, \dots, n$ . 盟军在战斗中任意击毁了  $k$  辆坦克, 被击毁的坦克编号分别为  $x_1, x_2, \dots, x_k$ , 能否通过被击毁的坦克编号来估计  $n$  的大小, 即估计德国生产了多少辆坦克.

在没有其它信息的情况下, 不妨假设被随机击毁的坦克是等可能事件, 即第  $i$  辆坦克被击毁的概率为  $1/n$ . 可以将问题看作从集合  $\{1, 2, \dots, n\}$  中不放回随机抽取  $k$  个数, 用  $X$  表示抽到的  $k$  个数中的最大数. 则  $X$  的取值为  $\{k, k+1, \dots, n\}$  以及概率

$$P(X = i) = \binom{i-1}{k-1} / \binom{n}{k} \quad (i = k, k+1, \dots, n).$$

于是得到

$$E(X) = \binom{n}{k}^{-1} \sum_{i=k}^n \binom{i-1}{k-1} i.$$

针对上面的求和表达式, 可以考虑从  $n+1$  个元素中选取  $k+1$  个元素, 共有  $\binom{n+1}{k+1}$  种不同的方法. 将这些不同的方法分情况讨论, 按照选取的  $k+1$  个元素中最大元素  $i = k+1, k+2, \dots, n+1$  进行分类; 若最大元素为  $i$ , 则有  $\binom{i-1}{k}$  种不同的方法. 于是有

$$\binom{n+1}{k+1} = \sum_{i=k+1}^{n+1} \binom{i-1}{k} = \sum_{i=k}^n \binom{i}{k} = \sum_{i=k}^n \frac{i}{k} \binom{i-1}{k-1},$$

代入期望  $E(X)$  可得

$$E(X) = k \binom{n}{k}^{-1} \sum_{i=k}^n \binom{i-1}{k-1} \frac{i}{k} = k \binom{n+1}{k+1} / \binom{n}{k} = \frac{k(n+1)}{k+1}.$$

由于仅做了一次观察, 将观察中  $k$  个数的最大值近似期望  $E[X]$ , 即  $E(X) \approx \max(x_1, x_2, \dots, x_n)$ , 由此估计

$$n \approx \max(x_1, x_2, \dots, x_n) \left(1 + \frac{1}{k}\right) - 1,$$

从而完成  $n$  的估计.

例如, 如果观察到被击毁坦克编号分别为 17, 68, 94, 127, 135, 212, 根据上面的推到可估计出

$$n \approx 212 \times (1 + 1/6) - 1 = 246.$$

针对德国坦克数量的实际估计情况见下表, 可以发现利用上述所提的统计估计方法接近德国的实际产量, 比英国的情报估计准确得多.

时间	统计估计	英国情报估计	德国实际产量
1940-06	169	1000	122
1941-06	244	1550	271
1942-08	327	1550	342

### 3.5.2 集卡活动

很多小朋友喜欢各种集卡活动, 如奥特曼卡和叶罗丽卡等. 事实上很多成年人也对集卡游戏并不陌生, 例如 80 年代的葫芦娃洋画、或 90 年代的小虎队旋风卡等. 问题可以描述为: 市场上有  $n$  种不同类型的卡片, 假设一个小朋友每次都能以等可能概率、独立地收集一张卡片, 问一个小朋友在平均情况下至少要收集多少张卡才能收集齐  $n$  种不同类型的卡片.

这里先补充一个需要用到的引理, 后面将给出详细的证明:

**引理 3.5** 对任意的随机变量  $X_1, X_2, \dots, X_n$  有

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n).$$

用  $X$  表示收集齐  $n$  种不同类型的卡片所需要的收集次数, 用  $X_k$  表示收集齐第  $k-1$  种和第  $k$  种不同类型卡片之间所需要的收集次数 ( $k \in [n]$ ), 于是有  $X = X_1 + X_2 + \dots + X_n$ . 我们的问题是计算期望  $E(X)$ .

很容易发现随机变量  $X_k$  服从参数为  $p_k$  的几何分布. 当已经收集到  $k-1$  种不同类型的卡片时, 再获得一张新卡的概率

$$p_k = 1 - (k-1)/n.$$

根据几何分布的性质有  $E[X_k] = 1/p_k = n/(n-k+1)$ . 利用引理 3.5 有

$$E(X) = E\left(\sum_{k=1}^n X_k\right) = \sum_{k=1}^n E(X_k) = \sum_{k=1}^n \frac{n}{n-k+1} = n \sum_{k=1}^n \frac{1}{k} = nH(n),$$

这里  $H(n)$  表示参数为  $n$  的调和数, 即  $H(n) = \sum_{k=1}^n 1/k$ . 关于调和数有

**引理 3.6** 调和数  $H(n) \in [\ln(n+1), 1 + \ln(n)]$ .

**证明** 因为函数  $1/x$  在  $x \in (0, +\infty)$  单调递减, 有

$$\ln(n+1) = \int_{x=1}^{n+1} \frac{1}{x} dx \leq \sum_{k=1}^n \frac{1}{k} = 1 + \sum_{k=2}^n \frac{1}{k} \leq 1 + \int_{x=1}^n \frac{1}{x} dx = 1 + \ln(n).$$

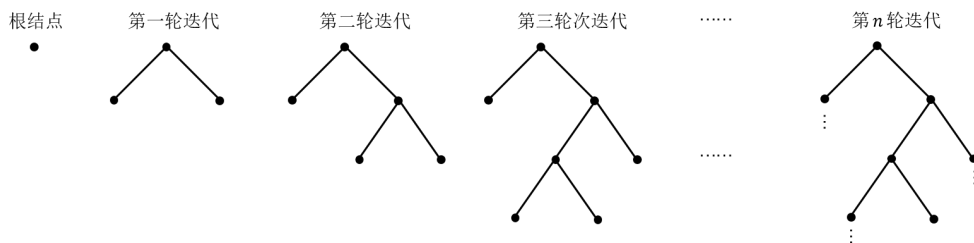
最后得到  $n \ln(n+1) \leq E(X) \leq n + n \ln n$ .

### 3.5.3 随机二叉树叶子结点的高度

在机器学习中, 随机树和随机森林是一类经典的分类或回归算法, 随机树叶子结点的高度估计对学习算法性能的分析具有重要作用. 本节考虑完全随机的二叉树中一个叶子结点的平均高度. 随机二叉树的构造过程非常简单: 首先给定二叉树的根结点, 然后在每一轮的迭代过程中执行以下两步操作:

- 在当前所有的叶子结点中随机选择一个叶子结点作为划分结点;
- 被选中的叶子结点变成一个内部结点, 生长出左、右两个叶子结点.

重复上述过程  $n$  步, 最后得到具有  $n$  个叶子结点的随机二叉树. 在这一构造过程中, 最关键的一步是随机选择的叶子结点作为划分结点. 随机二叉树构造的示意图如下所示:



**图 3.3** 随机二叉树构造的示意图

一个叶子结点的高度是从根节点到该叶子结点的路径中边的条数. 求解的问题: 在最后生成的随机二叉树中, 求任意一个叶结点的平均高度.

用随机变量  $X$  表示任意给定的一个叶结点的高度, 并用随机变量  $X_i$  表示在第  $i$  轮迭代过程中该叶子的祖先结点是否恰好被选中作为划分结点, 而在第  $i$  轮迭代过程中恰好有  $i$  个叶结点, 则有

$$X_i = \text{Ber}(1/i) \quad \text{且} \quad X = X_1 + X_2 + \cdots + X_n.$$

根据期望的性质和引理 3.6 有

$$E[X] = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n 1/i = H(n) \in [\ln(n+1), 1 + \ln(n)] .$$

由此可知一个叶子结点的平均高度为  $\Theta(\ln n)$ .

## 习题

- 3.1** 从  $\{1, 2, \dots, 10\}$  中有放回地任取 5 个数, 令  $X$  表示五个数中的最大值, 求  $X$  的分布列, 并求在无放回地情况下的分布列.
- 3.2** 将一枚骰子任意投掷三次, 用  $X$  表示三次中得到最小点的点数, 求  $X$  的分布列及期望.
- 3.3** 有 4 个盒子编号分别为 1, 2, 3, 4. 将 3 个不同的球随机放入 4 个盒子中, 同一盒子内的球无顺序关系, 用  $X$  表示有球盒子的最小编号, 求  $E(X)$ .
- 3.4** 设离散型随机变量  $X \in [a, b]$  的取值有有限中可能, 实值函数  $g: [a, b] \rightarrow \mathbb{R}$  是凹函数, 证明  $g(E(X)) \geq E(g(X))$ .
- 3.5** 若随机变量  $X \sim B(n, p)$ , 证明当  $k \in [0, np + p]$  时  $P(X = k)$  单调递增, 当  $k \in [np + p, n]$  时  $P(X = k)$  单调递减.
- 3.6** 设随机变量  $X$  服从参数为  $\lambda$  的泊松分布, 且  $P(X = 1) = P(X = 2)$ , 求  $P(X \geq 4)$ .
- 3.7** 设随机变量  $X$  的取值为  $r, r + 1, \dots$  以及事件  $\{X = k\}$  的概率为

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad p \in (0, 1), \quad k = r, r+1, r+2, \dots,$$

检验上面的概率构成一个分布列.

- 3.8** 设随机变量  $X$  的分布列为

$$P(X = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad p \in (0, 1), \quad k = r, r+1, r+2, \dots,$$

证明: 随机变量  $X$  的期望  $E(X) = r/p$  和方差  $\text{Var}(X) = r(1-p)/p^2$

- 3.9** 现需要 100 个符合规格的元件, 从市场上购买该元件的废品率为 0.01, 现准备在市场上买  $100 + x$  个元件, 要使得其中至少有 100 个符合规格元件的概率大于 0.95, 求  $x$  的最小值?
- 3.10** 设随机变量的分布列为  $P(X = (-5)^k/k) = 4/5^k$  ( $k = 1, 2, \dots$ ), 证明  $X$  的期望不存在.
- 3.11** 一个箱子中有一个白球和一个红球, 若从箱子中随机摸到一个白球则再放入一个白球, 若摸到一个红球则结束这个游戏. 证明: 游戏结束时的摸球次数的期望不存在.



## 第 4 章 连续型随机变量

### 4.1 分布函数

离散型随机变量利用概率分布列将随机变量的取值和对应的概率全部罗列出来. 然而一些随机现象的试验结果可能不止可列个取值, 此时不能一一列举出来, 例如候车的等待时间、一个地区的降雨量、一盏电灯的寿命等. 特别地, 对于连续性随机变量, 它在任意一个特定值的概率为 0 (将在 4.2 节介绍), 此时用分布列来描述这一类型的随机变量就根本行不通.

对于一些非离散型随机变量, 我们可能更关心在某个区间内的概率, 而不是它在某个特定点值的概率. 例如, 对于一盏电灯而言, 我们关心其寿命大于 1000 个小时的概率, 而不是恰好 1005 个小时的概率. 针对这些随机现象, 我们关注于随机变量  $X$  在一个区间  $[x_1, x_2]$  上的概率  $P(x_1 \leq X \leq x_2)$ . 为此引入分布函数的概念:

**定义 4.1** 给定随机变量  $X$ , 对任意实数  $x \in (-\infty, +\infty)$ , 函数

$$F(x) = P(X \leq x)$$

称为随机变量  $X$  的 **分布函数** (cumulative distribution function).

分布函数  $F(x)$  是定义在  $(-\infty, +\infty)$  的普通函数, 将普通函数与随机事件的概率关联起来, 有利于利用数学分析的知识来研究随机变量. 分布函数不限制随机变量的类型, 无论是离散型随机变量还是非离散型随机变量, 都有各自的分布函数.

分布函数的本质是概率, 考虑随机事件  $\{X \in (-\infty, x]\}$  的概率. 对任意实数  $x_1 < x_2$  有

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1).$$

若已知随机变量  $X$  的分布函数  $F(x)$ , 则可以知道  $X$  落入任意区间  $(x_1, x_2]$  上的概率, 因此分布函数完整地刻画了随机变量的统计规律性. 分布函数具有良好的分析性质:

**定理 4.1** 分布函数  $F(x)$  具有如下性质:

- 单调性: 若  $x_1 < x_2$ , 则  $F(x_1) \leq F(x_2)$ ;
- 规范性:  $F(x) \in [0, 1]$ , 且  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ ,  $F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = 1$ ;
- 右连续性:  $F(x+0) = \lim_{\Delta x \rightarrow 0^+} F(x + \Delta x) = F(x)$ .

**证明** 根据概率的非负性, 对任意  $x_1 < x_2$  有

$$F(x_2) - F(x_1) = P(x_1 < X \leq x_2) \geq 0.$$

根据规范性有

$$\begin{aligned} 1 = P(-\infty < X < +\infty) &= \sum_{n=-\infty}^{+\infty} P(n < X \leq n+1) = \sum_{n=-\infty}^{+\infty} F(n+1) - F(n) \\ &= \lim_{n \rightarrow -\infty} F(n) - \lim_{m \rightarrow +\infty} F(m). \end{aligned}$$

根据  $F(x)$  的单调性有  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = \lim_{n \rightarrow -\infty} F(n)$  和  $F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = \lim_{n \rightarrow +\infty} F(n)$ , 以及结合  $F(-\infty), F(+\infty) \in [0, 1]$  和  $F(+\infty) - F(-\infty) = 1$  可得

$$F(-\infty) = 0 \quad \text{和} \quad F(+\infty) = 1.$$

针对右连续性, 设  $\{x_n\}_{n=1}^{\infty}$  是一个单调下降的数列且  $x_n \rightarrow x$ , 则有

$$F(x_1) - F(x) = P(x < X \leq x_1) = \sum_{n=1}^{+\infty} F(x_n) - F(x_{n+1}) = F(x_1) - \lim_{n \rightarrow +\infty} F(x_n).$$

于是得到  $\lim_{n \rightarrow +\infty} F(x_n) = F(x)$ , 再结合函数  $F(x)$  的单调性有

$$F(x+0) = \lim_{n \rightarrow +\infty} F(x_n) = F(x),$$

由此完成证明.

通过上面的证明发现, 分布函数的三条基本性质, 分别对应于概率的三条公理. 因此, 任何分布函数都满足三条基本性质, 而满足上面三条基本性质的函数必是某随机变量的分布函数.

有了分布函数, 就很容易计算随机变量  $X$  在很多区间上的概率, 例如

$$\begin{aligned} P(X > a) &= 1 - F(a) \\ P(X < a) &= F(a-0) = \lim_{x \rightarrow a^-} F(x) \\ P(X = a) &= F(a) - F(a-0) \\ P(X \geq a) &= 1 - F(a-0) \\ P(a \leq X \leq b) &= F(b) - F(a-0). \end{aligned}$$

针对离散型的随机变量  $X$ , 设其分布列为  $p_k = P(X = x_k)$  ( $k = 1, 2, \dots$ ), 根据概率的可列可加性可得  $X$  的分布函数为

$$F(x) = P(X \leq x) = \sum_{k: x_k \leq x} p_k. \quad (4.1)$$

**例 4.1** 随机变量  $X$  的分布列为  $P(X = -1) = P(X = 3) = 1/4$  和  $P(X = 2) = 1/2$ , 求  $X$  的分布函数.



解 当  $x < -1$  时, 根据 (4.1) 有

$$F(x) = P(X \leq x) = P(\emptyset) = 0;$$

当  $-1 \leq x < 2$  时, 根据 (4.1) 有

$$F(x) = P(X \leq x) = P(X = -1) = \frac{1}{4};$$

当  $2 \leq x < 3$  时, 根据 (4.1) 有

$$F(x) = P(X \leq x) = P(X = -1) + P(X = 2) = \frac{3}{4};$$

当  $x \geq 3$  时有  $F(x) = 1$ . 如图 4.1(a) 所示, 分布函数  $F(x)$  是一条阶梯形的曲线, 在  $x = -1, 2, 3$  处有跳跃点.

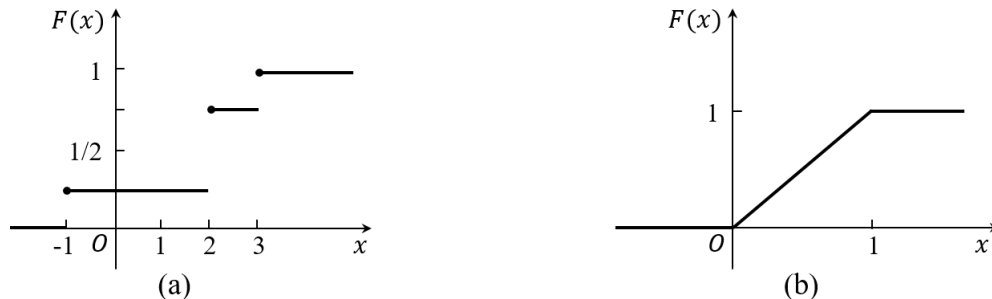


图 4.1 图 (a) 和 (b) 分别给出了例 4.1 和 4.2 的分布函数

**例 4.2** 在  $[0, 1]$  区间随机抛一个点, 用  $X$  表示落点的坐标, 假设  $X$  落入  $[0, 1]$  区间内任一子区间的概率与区间长度成正比, 求  $X$  的分布函数.

解 设随机变量  $X$  的分布函数为  $F(x)$ , 其中  $x \in [0, 1]$ , 当  $x < 0$  时有  $F(x) = 0$ ; 当  $x > 1$  时有  $F(x) = 1$ . 当  $x \in [0, 1]$  时有

$$F(x) = P(X \leq x) = kx.$$

根据  $F(1) = 1$  求解可得  $k = 1$ . 从而得到  $X$  的分布函数为

$$F(x) = \begin{cases} 0 & x < 0, \\ x & 0 \leq x \leq 1, \\ 1 & x > 1. \end{cases}$$

如图 4.1(b) 所示, 分布函数  $F(x)$  是一条连续的折线.

**例 4.3** 随机变量  $X$  的分布函数  $F(x) = A + B \arctan x$ ,  $x \in (-\infty, +\infty)$ , 求  $P(X \leq 1)$ .

解 由分布函数的性质有

$$0 = F(-\infty) = \lim_{x \rightarrow -\infty} A + B \arctan x = A - \pi B/2,$$

$$1 = F(+\infty) = \lim_{x \rightarrow +\infty} A + B \arctan x = A + \pi B/2,$$

求解可得  $A = 1/2$  和  $B = 1/\pi$ , 从而得到  $P(X \leq 1) = 3/4$ .

## 4.2 概率密度函数

离散型随机变量的取值是有限个或可列个离散的点, 本节研究连续型随机变量, 即随机变量的取值充满整个区间  $[a, b]$  或  $(a, +\infty)$ , 例如火车的到站时间、或一盏灯泡的寿命等. 离散型和连续型随机变量是实际应用中常遇到的两种随机变量.

**定义 4.2** 设随机变量  $X$  的分布函数为  $F(x)$ , 如果存在可积函数  $f(x)$ , 使得对任意实数  $x$  有

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

则称  $X$  为 **连续型随机变量**, 函数  $f(x)$  为随机变量  $X$  的 **概率密度函数** (probability density function), 简称 **密度函数**.

下面给出概率密度函数的一系列性质:

**引理 4.1** 概率密度函数  $f(x)$  满足非负性  $f(x) \geq 0$  和规范性  $\int_{-\infty}^{+\infty} f(t) dt = 1$ .

任意概率密度函数必然满足非负性和规范性; 而对满足非负性和规范性的任意函数  $f(x)$ , 其必为某个随机变量的密度函数, 并有分布函数为  $F(x) = \int_{-\infty}^x f(t) dt$ , 密度函数完整地刻画了随机变量的统计规律. 分布函数和密度函数都能刻画连续随机变量的统计规律, 但密度函数在图形上对各种分布特征的显示要优越得多, 比分布函数更常用.

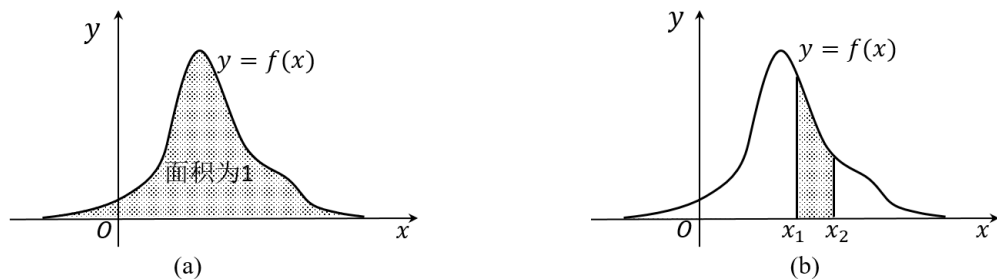


图 4.2 概率密度函数的几何解释

根据规范性可知曲线  $y = f(x)$  与  $x$  轴所围成的面积为 1 (如图 4.2(a) 所示). 对任意  $x_1 < x_2$ , 有

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t) dt.$$

由此给出概率密度的几何解释: 随机变量  $X$  落入区间  $(x_1, x_2]$  的概率等于由  $x$  轴,  $x = x_1$ ,  $x = x_2$  和  $y = f(x)$  所围成的曲边梯形的面积, 如图 4.2(b) 所示.

**引理 4.2** 对连续随机变量  $X$ , 分布函数  $F(x)$  在整个实数域上连续; 若密度函数  $f(x)$  在  $x$  点连续, 则分布函数  $F(x)$  在  $x$  点可导, 且有  $F'(x) = f(x)$ .

**证明** 该引理根据函数的积分性质直接可得: 若函数  $f(x)$  在实数域上可积, 则积分函数

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

在实数域上连续; 若函数  $f(x)$  在实数域上连续, 则  $F(x) = \int_{-\infty}^x f(t)dt$  在实数域上可导, 且有  $F'(x) = f(x)$  成立.

**引理 4.3** 对任意常数  $c$  和连续型随机变量  $X$ , 有  $P(X = c) = 0$ .

**证明** 对任意  $\Delta x > 0$  有事件  $\{X = x\} \subset \{X \in (x - \Delta x, x]\}$ , 根据积分中值定理有

$$P(X = x) \leq \lim_{\Delta x \rightarrow 0} P(x - \Delta x \leq X \leq x) = \lim_{\Delta x \rightarrow 0} \int_{x-\Delta x}^x f(t)dt \leq \lim_{\Delta x \rightarrow 0} f(\xi)\Delta x = 0,$$

其中  $\xi = \arg \max_{x \in (x-\Delta x, x]} f(x)$ , 根据概率的非负性完成证明.

根据上面的引理, 一个事件的概率为 0, 不能推出该事件是不可能事件; 一个事件的概率为 1, 也不能推出该事件是必然事件. 此外, 连续随机变量的概率无需强调端点, 因为

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b).$$

因为  $f(x) \neq 0 = P(X = x)$ , 由此说明概率密度函数不是概率.

若  $f(x)$  在点  $x$  连续, 由连续性定义有

$$\lim_{\Delta x \rightarrow 0} \frac{P(x - \Delta x \leq X \leq x + \Delta x)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\int_{x-\Delta x}^{x+\Delta x} f(t)dt}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{2\Delta x \cdot f(\xi)}{\Delta x} = 2f(x),$$

其中  $\xi \in (x - \Delta x, x + \Delta x)$ . 由此可得

$$P(x - \Delta x \leq X \leq x + \Delta x) \approx 2f(x)\Delta x,$$

若概率密度  $f(x)$  越大, 则  $X$  在  $x$  附近取值的概率越大.

例 4.4 设随机变量  $X$  的密度函数

$$f(x) = \begin{cases} x & 0 < x \leq 1 \\ a - x & 1 < x < 2 \\ 0 & \text{其它,} \end{cases}$$

求其分布函数  $F(x)$ .

解 根据概率密度的规范性有

$$1 = \int_{-\infty}^{+\infty} f(t)dt = \int_0^1 tdt + \int_1^2 (a-t)dt = a - 1,$$

从而求解出  $a = 2$ , 于是得到具体的密度函数  $f(x)$ . 当  $x \leq 0$  时有  $F(x) = 0$ ; 当  $0 < x \leq 1$  时, 有

$$F(x) = \int_0^x f(t)dt = x^2/2;$$

当  $1 < x \leq 2$  时, 有

$$F(x) = \int_0^1 f(t)dt + \int_1^x f(t)dt = 1/2 + \int_1^x (2-t)dt = -x^2/2 + 2x - 1;$$

当  $x \geq 2$  时有  $F(x) = 1$ . 综合可得

$$F(x) = \begin{cases} 0 & x \leq 0, \\ x^2/2 & 0 < x \leq 1, \\ -x^2/2 + 2x - 1 & 1 < x \leq 2, \\ 1 & x \geq 2. \end{cases}$$

随机变量  $X$  的密度函数和分布函数如图 4.3 所示.

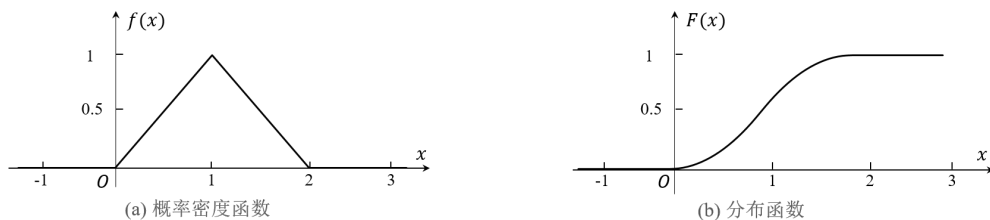


图 4.3 例 4.4 中随机变量  $X$  的密度函数和分布函数图

**例 4.5** 对连续随机变量  $X$ , 当  $x \in (0, 3)$  时密度函数  $f(x) = cx^2$ , 在其它点的密度函数  $f(x) = 0$ . 设随机变量

$$Y = \begin{cases} 2, & X \leq 1 \\ X, & X \in (1, 2) \\ 1, & X \geq 2 \end{cases}$$

求随机变量  $Y$  的分布函数, 以及计算概率  $P(Y \geq X)$ .

**解** 根据概率密度函数的规范性有  $1 = \int_{-\infty}^{+\infty} f(t)dt = 9c$ , 由此可得  $c = 1/9$ .

用  $F_Y(y)$  表示随机变量  $Y$  的分布函数. 当  $y < 1$  时, 有  $F_Y(y) = P(Y \leq y) = 0$ ; 当  $y \geq 2$  时, 有  $F_Y(y) = P(Y \leq y) = 1$ ; 当  $1 \leq y < 2$  时有

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(Y = 1) + P(1 < Y \leq y) \\ &= P(X \geq 2) + P(1 < X \leq y) = \int_2^3 t^2/9dt + \int_1^y t^2/9dt = (18 + y^3)/27. \end{aligned}$$

由此可得随机变量  $Y$  的分布函数为

$$F_Y(y) = \begin{cases} 0, & y < 1, \\ (18 + y^3)/27, & y \in [1, 2), \\ 1, & y \geq 2. \end{cases}$$

可以观察发现随机变量  $Y$  不是连续型随机变量, 也不是离散型随机变量. 最后计算概率

$$P(X \leq Y) = P(X < 2) = \int_0^2 t^2/9dt = 8/27.$$

**例 4.6** 已知一个靶半径为 2 米的圆盘, 击中靶上任一同心圆盘上的点的概率与该圆盘的面积成正比. 假设射击都能击中靶, 用  $X$  表示击中点与圆心的距离, 求  $X$  的概率密度函数.

**解** 根据题意分析随机变量  $X$  的分布函数  $F(x)$ . 当  $x < 0$  时有  $F(x) = 0$ ; 当  $0 \leq x \leq 2$  时有

$$F(x) = P(X \leq x) = P(0 \leq X \leq x) = kx^2.$$

根据分布函数的性质有  $F(2) = 1 = 4k$ , 求解可得  $k = 1/4$ , 进一步得到  $X$  的概率密度

$$f(x) = \begin{cases} x/2 & 0 \leq x \leq 2 \\ 0 & \text{其它.} \end{cases}$$

### 4.3 连续型随机变量的期望和方差

本节研究连续型随机变量的期望和方差, 有利于了解这些变量的整体数字特征.

**定义 4.3** 设连续随机变量  $X$  的概率密度函数为  $f(x)$ , 若积分  $\int_{-\infty}^{+\infty} |x|f(x)dx$  收敛, 称  $\int_{-\infty}^{+\infty} xf(x)dx$  为随机变量  $X$  的 **期望**, 记为  $E(X)$ , 即

$$E(X) = \int_{-\infty}^{+\infty} tf(t)dt.$$

类似于离散型随机变量, 连续型随机变量的期望具有以下一些性质:

**引理 4.4 (线性关系)** 对任意任意常数  $a, b$  和连续随机变量  $X$ , 有  $E(aX + b) = aE(X) + b$ .

**引理 4.5 (Jensen 不等式)** 对连续随机变量  $X$  和函数  $g(x)$ ,

- 若  $g(x)$  是凸函数, 则有  $g(E(X)) \leq E[g(X)]$ ;
- 若  $g(x)$  是凹函数, 则有  $g(E(X)) \geq E[g(X)]$ .

对于非负的连续型随机变量, 也可以利用  $P(X > t)$  来直接计算期望:

**引理 4.6** 若连续型随机变量  $X \geq 0$ , 则有

$$E[X] = \int_0^{\infty} P(X > t)dt.$$

该定理对随机变量函数  $Y = g(X) \geq 0$  也成立, 即  $E[g(X)] = \int_0^{\infty} P(g(X) > t)dt$ .

**证明** 设随机变量  $X$  的概率密度为  $f(x)$ , 首先观察得到

$$X = \int_0^X 1dt = \int_0^{+\infty} \mathbb{I}[t < X]dt = \int_0^{+\infty} \mathbb{I}[X > t]dt,$$

这里  $\mathbb{I}[\cdot]$  表示指示函数, 如果论断为真, 其值为 1, 否则为 0. 两边同时取期望有

$$\begin{aligned} E[X] &= E\left[\int_0^{+\infty} \mathbb{I}[X > t]dt\right] \\ &= \int_0^{+\infty} \left[\int_0^{+\infty} \mathbb{I}[x > t]f(x)dt\right]dx \quad (\text{积分换序}) \\ &= \int_0^{+\infty} \left[\int_0^{+\infty} \mathbb{I}[x > t]f(x)dx\right]dt \\ &= \int_0^{+\infty} \left[\int_0^t \mathbb{I}[x > t]f(x)dx + \int_t^{+\infty} \mathbb{I}[x > t]f(x)dx\right]dt \\ &= \int_0^{+\infty} \left[\int_t^{+\infty} f(x)dx\right]dt = \int_0^{+\infty} P(X > t)dt. \end{aligned}$$

对于连续随机变量函数的期望有

**定理 4.2** 设随机变量  $X$  的密度函数为  $f(x)$ , 且  $\int_{-\infty}^{+\infty} g(t)f(t)dt$  绝对可积, 则

$$E(g(X)) = \int_{-\infty}^{+\infty} g(t)f(t)dt.$$

该定理表明, 若已知随机变量  $X$  的密度函数为  $f(x)$ , 以及随机变量函数  $Y = g(X)$ , 可以直接利用随机变量  $X$  的密度函数来计算  $Y$  的期望, 而不需要知道随机变量  $Y$  的密度函数.

**证明** 该定理对一般的可积函数  $g(x)$  均成立, 但证明过程却非常复杂, 这里仅给出非负随机变量函数  $g(x) \geq 0$  的证明. 根据引理 4.6 有

$$\begin{aligned} E[g(X)] &= \int_0^{+\infty} P(g(X) \geq t)dt = \int_0^{+\infty} \int_{x: g(x) \geq t} f(x)dxdt \\ &= \int_{x: g(x) \geq 0} \int_0^{g(x)} f(x)dt dx = \int_{x: g(x) \geq 0} g(x)f(x)dx = \int_{-\infty}^{+\infty} g(x)f(x)dx, \end{aligned}$$

由此完成证明.

下面介绍物理学中用到的柯西分布 (Cauchy distribution), 它的期望不存在.

**例 4.7** 设连续随机变量  $X$  的密度函数为  $f(x) = 1/\pi(1+x^2)$  ( $x \in \mathbb{R}$ ), 求期望  $E(X)$ .

因为积分

$$\int_{-\infty}^{+\infty} \frac{|x|}{\pi(1+x^2)} dx = 2 \int_0^{+\infty} \frac{x}{\pi(1+x^2)} dx = \frac{1}{\pi} [\ln(1+x^2)]_0^{+\infty} = +\infty,$$

由此可知期望  $E(X)$  不存在.

**例 4.8** 古人运送粮草, 如果早到每天需要的存储费用  $c$  元, 如果晚到每天需要的延期费用为  $C$  元. 粮草在运送过程中存在天气、路况等不确定因素, 因此运送需要的天数是随机的, 其概率密度函数为  $f(x)$ , 问什么时候出发才能使费用的期望值最小?

**解** 用随机变量  $X$  表示实际的运送天数, 分布函数为  $F(x)$ . 不妨假设提前了  $t$  天出发, 那么所需费用为

$$\ell_t(X) = \begin{cases} c(t-X) & X \leq t, \\ C(X-t) & X > t. \end{cases}$$

因此可得

$$\begin{aligned} E[\ell_t(X)] &= \int_0^{+\infty} \ell_t(x)f(x)dx = \int_0^t c(t-x)f(x)dx + \int_t^{+\infty} C(x-t)f(x)dx \\ &= ctF(t) - c \int_0^t xf(x)dx + C \int_t^{+\infty} xf(x)dx - Ct(1-F(t)). \end{aligned}$$

对上式中的  $t$  求导、并令导数为零可得

$$\frac{d}{dt} E[\ell_t(X)] = cF(t) - C(1 - F(t)) = (c + C)F(t) - C.$$

求解可得期望最小的天数  $t^*$  满足

$$F(t^*) = C/(C + c).$$

**定义 4.4** 设连续随机变量  $X$  的概率密度为  $f(x)$ , 若  $\int_{-\infty}^{+\infty} (t - E(X))^2 f(t) dt$  收敛, 称为随机变量  $X$  的方差, 记为  $\text{Var}(X)$ , 即

$$\text{Var}(X) = E(X - E(X))^2 = \int_{-\infty}^{+\infty} (t - E(X))^2 f(t) dt.$$

其等价性定义为

$$\text{Var}(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2 = \int_{-\infty}^{+\infty} t^2 f(t) dt - \left( \int_{-\infty}^{+\infty} t f(t) dt \right)^2.$$

类似于离散型随机变量, 连续型随机变量的方差具有如下性质:

- 对连续型随机变量  $X$  和常数  $a, b$ , 有  $\text{Var}(aX + b) = a^2 \text{Var}(X)$ ;
- 对连续型随机变量  $X$  和常数  $a$ , 有  $\text{Var}(X) = E(X - E(X))^2 \leq E(X - a)^2$ ;
- 对连续型随机变量  $X \in [a, b]$ , 有  $\text{Var}(X) = (b - E(X))(E(X) - a) \leq (b - a)^2/4$ .

#### 4.4 常用连续型随机变量

下面介绍几种常用的连续型随机变量.

##### 4.4.1 均匀分布(uniform distribution)

**定义 4.5** 若随机变量  $X$  的概率密度函数为

$$f(x) = \begin{cases} 1/(b - a) & x \in [a, b] \\ 0 & \text{其它,} \end{cases}$$

称  $X$  在区间  $[a, b]$  上服从 **均匀分布**, 记  $X \sim U(a, b)$ .

根据上面的定义很容易发现服从均匀分布的随机变量落入区间任何一点的概率相同. 对任意实数  $x \in \mathbb{R}$  有  $f(x) \geq 0$  且

$$\int_{-\infty}^{+\infty} f(t) dt = \int_{-\infty}^a f(t) dt + \int_a^b f(t) dt + \int_b^{+\infty} f(t) dt = \int_a^b \frac{1}{b - a} dt = 1.$$



若随机变量  $X \sim U(a, b)$ , 则  $X$  落入内任一子区间  $[x, x + \Delta]$  的概率

$$P(x \leq X \leq x + \Delta) = \int_x^{x+\Delta} \frac{1}{b-a} dt = \frac{\Delta}{b-a} .$$

该概率与子区间的具体位置  $x$  无关, 而与子区间长度  $\Delta$  成正比, 由此给出了均匀分布的几何解释: 若随机变量  $X \sim U(a, b)$ , 则  $X$  落入  $[a, b]$  内任一子区间的概率与该区间的长度成正比, 与位置无关.

根据分布函数的定义可知  $X \sim U(a, b)$  的分布函数为

$$F(x) = \begin{cases} 0 & x \leq a \\ (x-a)/(b-a) & a < x < b \\ 1 & x \geq b \end{cases}$$

随机变量  $X \sim U(a, b)$  的密度函数和分布函数的示意图如下:

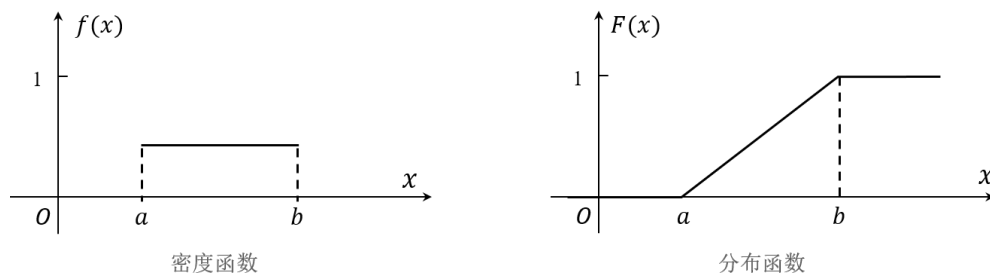


图 4.4 随机变量  $X \sim U(a, b)$  的密度函数和分布函数

**定理 4.3** 若随机变量  $X \sim U(a, b)$ , 则有

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

**证明** 根据期望的定义有

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} t f(t) dt = \frac{1}{b-a} \int_a^b t dt = \frac{a+b}{2} , \\ E(X^2) &= \int_{-\infty}^{+\infty} t^2 f(t) dt = \frac{1}{b-a} \int_a^b t^2 dt = \frac{a^2 + ab + b^2}{3} , \end{aligned}$$

从而得到方差

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} .$$

**例 4.9** 已知随机变量  $X \sim U(a, b)$ , 对  $a < c < d < b$ , 求  $P(X \leq c | X \leq d)$ .

**解** 根据条件概率的定义有

$$P(X \leq c | X \leq d) = \frac{P(\{X \leq d\} \cap \{X \leq c\})}{P(X \leq d)} = \frac{P(X \leq c)}{P(X \leq d)} = \frac{c - a}{d - a},$$

即在  $X \leq d$  的条件下, 随机变量  $X$  服从  $U(a, d)$ .

**例 4.10** 设随机变量  $\xi \sim U(-3, 6)$ , 试求方程  $4x^2 + 4\xi x + (\xi + 2) = 0$  有实根的概率.

**解** 易知随机变量  $\xi$  的概率密度函数

$$f(t) = \begin{cases} 1/9 & x \in [-3, 6] \\ 0 & \text{其它.} \end{cases}$$

设事件  $A$  表示方程有实根, 于是有

$$\begin{aligned} P(A) &= P((4\xi)^2 - 4 \times 4 \times (\xi + 2) \geq 0) = P((\xi + 1)(\xi - 2) \geq 0) \\ &= P(\{\xi \geq -1\} \cap \{\xi \geq 2\} \geq 0) + P(\{\xi \leq -1\} \cap \{\xi \leq 2\} \geq 0) \\ &= P(\xi \leq -1) + P(\xi \geq 2) = \int_{-3}^{-1} \frac{1}{9} dt + \int_2^6 \frac{1}{9} dt = \frac{2}{3}. \end{aligned}$$

#### 4.4.2 指数分布

指数分布常用于电话的通话时间和银行的服务等待时间, 也可以用于描述动物和电子元件的寿命, 在可靠性理论和排队论中具有广泛的应用.

**定义 4.6** 给定常数  $\lambda > 0$ , 若随机变量  $X$  的密度函数

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{其它,} \end{cases}$$

称  $X$  服从 **参数为  $\lambda$  的指数分布**, 记  $X \sim e(\lambda)$ .

对任意实数  $x$  有密度函数  $f(x) \geq 0$ , 以及

$$\int_{-\infty}^{+\infty} f(t) dt = \int_0^{+\infty} \lambda e^{-\lambda t} dt = [-e^{-\lambda t}]_0^{+\infty} = 1.$$

根据指数分布的密度函数很容易得到分布函数, 即当  $x \leq 0$  时, 分布函数  $F(x) = 0$ ; 当  $x > 0$  时, 分布函数

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

指数分布的密度函数和分布函数如图 4.5 所示.

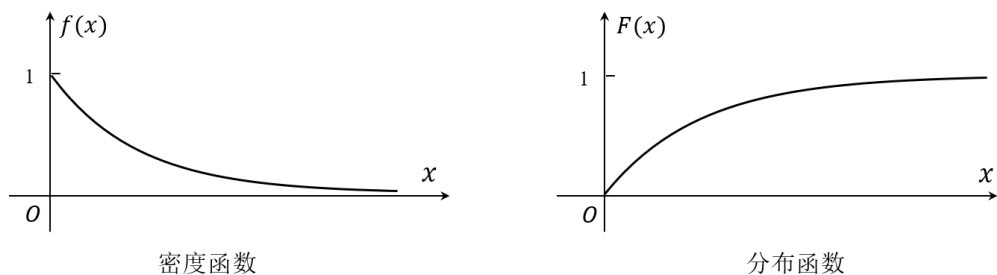


图 4.5 指数分布的密度函数和分布函数

**引理 4.7** 若随机变量  $X \sim e(\lambda)$ , 则  $E(X) = 1/\lambda$  和  $\text{Var}(X) = 1/\lambda^2$ .

**证明** 根据连续函数的定义有

$$E(X) = \int_0^{+\infty} t\lambda e^{-\lambda t} dt = \left[ -te^{-\lambda t} \right]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda t} dt = -\frac{1}{\lambda} \left[ e^{-\lambda t} \right]_0^{+\infty} = \frac{1}{\lambda},$$

对非负的随机变量  $X \geq 0$ , 可以利用引理 4.6 和分布函数  $F(x) = P(X \leq x) = 1 - e^{-\lambda x}$  有

$$E[X] = \int_0^{+\infty} P(X > t) dt = \int_0^{+\infty} 1 - F(t) dt = \int_0^{+\infty} e^{-\lambda t} dt = \frac{1}{\lambda}.$$

对于方差, 首先计算

$$E(X^2) = \lambda \int_0^{+\infty} t^2 e^{-\lambda t} dt = \left[ -t^2 e^{-\lambda t} \right]_0^{+\infty} + \int_0^{+\infty} 2te^{-\lambda t} dt = \frac{2}{\lambda} E(X) = \frac{2}{\lambda^2},$$

于是得到  $\text{Var}(X) = E(X^2) - [E(X)]^2 = 1/\lambda^2$ .

下面研究指数分布的一个重要性质: 指数分布的无记忆性.

**定理 4.4** 若随机变量  $X \sim e(\lambda)$ , 则对任意  $s > 0, t > 0$ , 有

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

**证明** 对任意  $x > 0$ , 根据分布函数有  $P(X > x) = 1 - F(x) = e^{-\lambda x}$ , 从而得到

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

定理得证.

指数分布是唯一具有无记忆性的连续型随机变量, 例如一盏灯泡的寿命  $X$  服从指数分布, 若已经使用了  $s$  个小时, 则再使用  $t$  个小时的概率与已使用过  $s$  个小时无关, 将这个经历给“忘记”了.

**引理 4.8** 若随机变量  $X_1, X_2, \dots, X_n$  是相互独立的、且分别服从参数为  $\lambda_1, \lambda_2, \dots, \lambda_n$  的指数分布, 则有

$$X = \min\{X_1, X_2, \dots, X_n\} \sim e(\lambda_1 + \lambda_2 + \dots + \lambda_n).$$

**解** 这里随机变量的相互独立性可以理解为随机变量取不同值的随机事件相互独立. 计算随机变量  $X$  的分布函数

$$\begin{aligned} F_X(x) &= P(X \leq x) = 1 - P(\min(X_1, X_2, \dots, X_n) > x) \\ &= 1 - \prod_{i=1}^n P(X_i > x) = 1 - \prod_{i=1}^n \exp(-\lambda_i x) = 1 - \exp\left(-x \sum_{i=1}^n \lambda_i\right), \end{aligned}$$

由此完成证明.

#### 4.4.3 正态分布

正态分布是概率统计中最重要的一种分布, 最早由法国数学家棣莫弗 (De Moivre, 1667-1754) 在 1730s 提出, 用于近似抛硬币试验中随机事件的概率, 即中心极限定理的雏形. 德国数学家高斯 (Gauss, 1777-1855) 在 1800s 首次将正态分布应用于预测天文学中星体的位置, 由此才展示出正态分布的应用价值, 后来发现很多随机现象可以通过正态分布来描述, 正态分布因此被称为高斯分布.

正态分布在概率统计中的重要性主要体现在以下几方面:

- 现实生活中很多随机现象需要用正态分布进行描述, 如人的身高或体重, 某地区的降雨量等;
- 很多分布可以通过正态分布来进行近似计算, 如后面所学的中心极限定理;
- 数理统计中常用的统计分布是由正态分布导出的, 如后面所学的  $\chi^2$  分布、 $t$  分布和  $F$  分布.

**定义 4.7** 给定任何实数  $\mu$  和  $\sigma > 0$ , 若随机变量  $X$  的概率密度为

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

称随机变量  $X$  服从 **参数为  $(\mu, \sigma^2)$  的正态分布** (normal distribution), 又称为 **高斯分布** (Gaussian distribution), 记为  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

特别地, 当  $\mu = 0$  和  $\sigma = 1$  时的正态分布  $\mathcal{N}(0, 1)$  被称为 **标准正态分布**, 此时密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad x \in (-\infty, +\infty).$$

对任意  $x \in (-\infty, +\infty)$  有  $f(x) \geq 0$ , 利用极坐标变换 ( $x = r \cos \theta, y = r \sin \theta$ ) 有

$$\left( \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right)^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy$$

$$= \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\frac{r^2}{2}} r dr = \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\frac{r^2}{2}} d\frac{r^2}{2} = 2\pi ,$$

由此验证了  $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 1$ , 利用简单的变量替换可验证一般正太分布的密度函数.

关于标准正太分布和一般的正太分布, 有如下关系:

**定理 4.5** 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则有  $Y = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ ; 若随机变量  $X \sim \mathcal{N}(0, 1)$ , 则有  $Y = \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$ .

**证明** 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则  $Y = (X - \mu)/\sigma$  的分布函数

$$F_Y(y) = P[Y \leq y] = P[X - \mu \leq y\sigma] = P[X \leq y\sigma + \mu] = \int_{-\infty}^{\mu+y\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt .$$

令  $x = (t - \mu)/\sigma$ , 代入上面的分布函数有

$$F_Y(y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx ,$$

由此可证  $Y \sim \mathcal{N}(0, 1)$ .

另一方面, 若随机变量  $X \sim \mathcal{N}(0, 1)$ , 则有  $Y = \sigma X + \mu$  的分布函数

$$F_Y(y) = P(Y \leq y) = P(\sigma X + \mu \leq y) = P(X \leq (y - \mu)/\sigma) = \int_{-\infty}^{(y-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt .$$

令  $t = (x - \mu)/\sigma$ , 代入上面的分布函数有

$$F_Y(y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx ,$$

由此可证  $Y \sim \mathcal{N}(\mu, \sigma^2)$ .

关于正太分布的数字特征有

**定理 4.6** 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则有  $E(X) = \mu$  和  $\text{Var}(X) = \sigma^2$ ; 特别地, 若随机变量  $X \sim \mathcal{N}(0, 1)$ , 则有  $E(X) = 0$  和  $\text{Var}(X) = 1$ .

正太分布的两个参数分别表示正太分布的期望和方差.

**证明** 这里仅仅证明标准正太分布的期望为 0 和方差为 1, 结合定理 4.5 直接可得  $X \sim \mathcal{N}(\mu, \sigma^2)$  的期望和方差. 若随机变量  $X \sim \mathcal{N}(0, 1)$ , 根据奇函数在对称的区间上积分为 0 有

$$E(X) = \int_{-\infty}^{+\infty} \frac{t}{\sqrt{2\pi}} e^{-t^2/2} dt = 0 .$$

根据方差的定义和分部积分有

$$\text{Var}(X) = \int_{-\infty}^{+\infty} \frac{t^2}{\sqrt{2\pi}} e^{-t^2/2} dt = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t de^{-t^2/2} = \left[ \frac{te^{-t^2/2}}{-\sqrt{2\pi}} \right]_{t=-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-t^2/2} dt = 1.$$

由此完成证明.

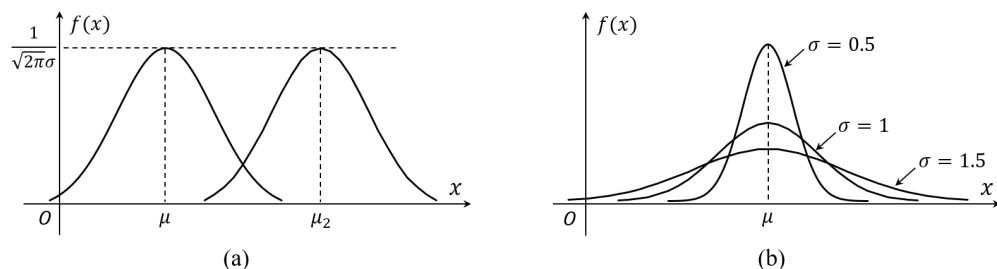


图 4.6 正太分布的密度函数

正太分布的密度函数如图 4.7(a) 所示, 具有以下一些特点:

- 1) 曲线  $f(x)$  关于  $x = \mu$  对称, 先单调递增, 之后单调递减, 在  $x = \mu$  处取最大值  $1/\sqrt{2\pi}\sigma$ . 说明随机变量  $X$  的取值主要集中在  $x = \mu = E(X)$  附近, 离  $x = \mu$  越远的区间概率越小.
- 2) 根据  $\lim_{x \rightarrow \pm\infty} f(x) = 0$  可知曲线  $f(x)$  以  $x$  轴为渐近线; 根据  $f''(x) = 0$  可知曲线  $f(x)$  的拐点为  $x = \mu \pm \sigma$ .
- 3) 固定标准差  $\sigma$  而改变期望  $\mu$  的值, 曲线  $f(x)$  形状不变, 仅沿  $x$  轴左右平行移动, 如图 4.7(a).
- 4) 固定期望  $\mu$  而改变标准差  $\sigma$  的值, 曲线  $f(x)$  的对称点不变, 但最大值  $1/\sqrt{2\pi}\sigma$  和拐点  $x = \mu \pm \sigma$  发生了改变. 如图 4.7(b) 所示: 当  $\sigma$  越小, 曲线顶峰越高, 曲线越陡峭, 分布越集中, 方差越小; 反之  $\sigma$  越大, 曲线顶峰越低, 曲线越平坦, 分布越分散, 方差越大.

关于正太分布的概率估计, 有下面的不等式:

**定理 4.7** 若  $X \sim \mathcal{N}(0, 1)$ , 对任意  $\epsilon > 0$  有

$$P(X \geq \epsilon) \leq \frac{1}{2} e^{-\epsilon^2/2}$$

$$P(|X| \geq \epsilon) \leq \min \left\{ 1, \sqrt{\frac{2}{\pi}} \frac{1}{\epsilon} e^{-\epsilon^2/2} \right\}.$$

在上面的定理中, 第一个不等式具有广泛的应用, 在  $\epsilon \in (0, 1)$  时对真实的概率有更好的估计; 第二个不等式被称为 Mill 不等式, 在  $\epsilon \in (1, +\infty)$  时对真实的概率有更好的估计.

**证明** 针对第一个不等式, 我们有

$$\begin{aligned} P(X \geq \epsilon) &= \int_{\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \\ &\leq e^{-\epsilon^2/2} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{2} e^{-\epsilon^2/2}. \end{aligned}$$

对于 Mill 不等式, 根据  $\mathcal{N}(0, 1)$  的概率密度  $f(x) = e^{-x^2/2}/\sqrt{2\pi}$  有  $f'(x) = -xf(x)$ , 进一步可得

$$\begin{aligned} P(|X| \geq \epsilon) &= 2 \int_{\epsilon}^{+\infty} f(t) dt = 2 \int_{\epsilon}^{+\infty} \frac{tf(t)}{t} dt \\ &\leq 2 \int_{\epsilon}^{+\infty} \frac{tf(t)}{\epsilon} dt = -2 \int_{\epsilon}^{+\infty} \frac{f'(t)}{\epsilon} dt = -\frac{2}{\epsilon} [f(t)]_{\epsilon}^{+\infty} = \frac{2}{\sqrt{2\pi}\epsilon} e^{-\epsilon^2/2}. \end{aligned}$$

由此完成证明.

若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则有分布函数

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

该分布函数没有显示的表达式, 只能求数值解, 函数如图 4.7(a) 所示. 为便于研究正态分布的分布函数, 利用定理 4.5 可将其它正态分布都转化为标准正态分布  $\mathcal{N}(0, 1)$ , 设其分布函数为

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

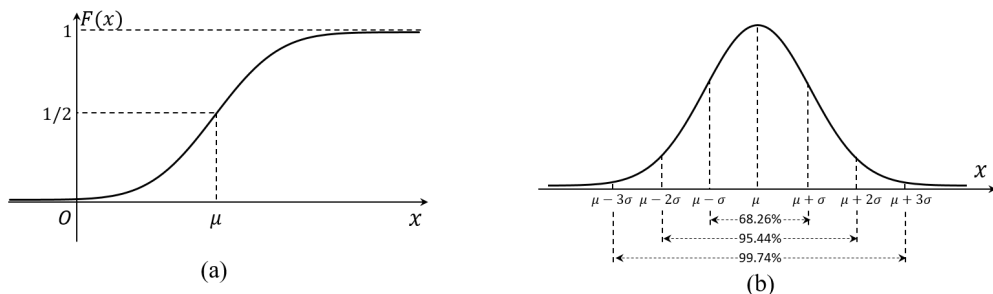
表 4.1 给出了标准正态分布  $\Phi(x)$  的函数表, 在计算具体的概率时可供查询. 下面给出关于分布函数  $\Phi(x)$  的一些性质:

- 1) 根据对称性有  $\Phi(x) + \Phi(-x) = 1$ .
- 2) 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则对任意实数  $a < b$  有

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

- 3) 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则对任意实数  $k > 0$  有

$$P(|x - \mu| < k\sigma) = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1.$$

图 4.7 正太分布函数和  $3\sigma$  原则

特别的, 当  $k = 1, 2, 3$  时通过表 4.1 有

$$P(|x - \mu| < \sigma) = 0.6826, \quad P(|x - \mu| < 2\sigma) = 0.9544, \quad P(|x - \mu| < 3\sigma) = 0.9974.$$

如图 4.7(b) 所示, 尽管随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$  的取值范围为整个实数域  $\mathbb{R}$ , 但其取值落在  $[\mu - 3\sigma, \mu + 3\sigma]$  之外的概率不超过千分之三, 也就是  $X \sim \mathcal{N}(\mu, \sigma^2)$  的取值几乎总在  $[\mu - 3\sigma, \mu + 3\sigma]$  之内, 这就是人们所说的“ $3\sigma$  原则”, 在实际的统计推断, 特别是产品质量检测中具有重要的应用.

4) 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 且已知  $P(X < c) = p$ , 则有

$$p = P(X < c) = P\left(\frac{X - \mu}{\sigma} < \frac{c - \mu}{\sigma}\right) = \Phi\left(\frac{c - \mu}{\sigma}\right),$$

由此可反解出  $c = \mu + \sigma\Phi^{-1}(p)$ . 这里  $\Phi^{-1}(x)$  表示标准正太分布函数  $\Phi(x)$  的反函数, 可根据表 4.1 由里向外查得, 例如  $\Phi^{-1}(0.5871) = 0.22$ .

**例 4.11** 已知某公司员工每个月的工资服从正太分布  $\mathcal{N}(6000, \sigma^2)$ , 问题:

- i) 若已知标准差  $\sigma = 500$ , 求工资在 5000 与 7000 之间的员工在公司中占比多少?
- ii) 当标准差  $\sigma$  为何值时, 工资在 5000 与 7000 之间的员工在公司中占比为 0.803?

**解** 用随机变量  $X$  表示公司员工每个月的工资, 则  $X \sim \mathcal{N}(6000, \sigma^2)$ . 针对问题 i), 当  $\sigma = 500$  时通过查询表 4.1 有

$$P(5000 \leq X \leq 7000) = P\left(-2 \leq \frac{X - 6000}{500} \leq 2\right) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0.9544.$$

针对问题 ii) 有

$$P(5000 \leq X \leq 7000) = P\left(-\frac{1000}{\sigma} \leq \frac{x - 6000}{\sigma} \leq \frac{1000}{\sigma}\right) = 2\Phi\left(\frac{1000}{\sigma}\right) - 1 = 0.803,$$

于是得到  $\Phi(1000/\sigma) = 0.9015$ , 通过由内自外查表 4.1 有  $\sigma \approx 775.2$ .



表 4.1 标准正态分布表  $\Phi(x) = \int_{-\infty}^x e^{-t^2/2}/\sqrt{2\pi}dt$ .

$x$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

## 4.5 连续随机变量函数的分布

当知道一个随机变量的概率分布后, 经常会考虑它的一些函数的分布, 例如已知一个圆的直径  $X$  服从均匀分布  $U(a, b)$ , 则可以考虑圆的面积  $Y = \pi(X/2)^2$  的分布. 一般地, 若已知随机变量  $X$  的分布函数, 以及  $g(x)$  是定义在随机变量  $X$  所有可能取值的函数, 则称  $Y = g(X)$  为随机变量  $X$  的函数, 很显然  $Y$  也是随机变量. 研究的问题可以归纳为: 若已知随机变量  $X$  的概率分布和函数  $g(x)$ , 如何求解随机变量  $Y = g(X)$  的概率分布.

设离散型随机变量  $X$  的分布列为  $p_k = P(X = x_k) (k = 1, 2, \dots)$ , 求解随机变量  $Y = g(X)$  的分布列较为简单, 将相等的项  $g(x_i) = g(x_j)$  合并, 相应的概率相加即可. 例如

**例 4.12** 若随机变量  $X$  的概率分布列为  $P(X = k) = 1/2^k (k = 1, 2, \dots)$ , 求随机变量  $Y = \cos(\pi X/2)$  的分布列.

**解** 当  $n = 1, 2, \dots$  时有

$$\cos(k\pi/2) = \begin{cases} 1 & k = 4n \\ 0 & k = 2n - 1 \\ -1 & k = 4n - 2, \end{cases}$$

所以  $\cos(\pi X/2)$  的取值为  $\{-1, 0, +1\}$ , 其概率分别为

$$\begin{aligned} P(Y = 1) &= \sum_{n=1}^{\infty} P(X = 4n) = \sum_{n=1}^{\infty} \frac{1}{2^{4n}} = \frac{1}{15}, \\ P(Y = 0) &= \sum_{n=1}^{\infty} P(X = 2n - 1) = \sum_{n=1}^{\infty} \frac{1}{2^{2n-1}} = \frac{2}{3}, \\ P(Y = -1) &= \sum_{n=1}^{\infty} P(X = 4n - 2) = \sum_{n=1}^{\infty} \frac{1}{2^{4n-2}} = \frac{4}{15}, \end{aligned}$$

由此给出随机变量  $Y$  的分布列.

针对连续型随机变量, 一般采用概率密度函数来刻画其概率分布. 若已知函数  $g(x)$  和随机变量  $X$  的概率密度函数为  $f_X(x)$ , 求解随机变量  $Y = g(X)$  的概率密度  $f_Y(y)$ , 通常分为以下两步:

- 1) 求解随机变量  $Y$  的分布函数  $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \int_{g(x) \leq y} f_X(x) dx$ ;
- 2) 求解随机变量  $Y$  的密度函数  $f_Y(y) = F'_Y(y)$ .

求解此类问题常用到的数学工具是积分求导公式: 设函数  $F(y) = \int_{\psi(y)}^{\varphi(y)} f(x) dx$ , 则有

$$F'(y) = f(\varphi(y))\varphi'(y) - f(\psi(y))\psi'(y). \quad (4.2)$$

下面来看两个例子.

**例 4.13** 设随机变量  $X$  的概率密度为  $f_X(x)$ , 求  $Y = X^2$  的概率密度.

**解** 根据分布函数的定义可知

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y).$$

当  $y \leq 0$  时有  $F_Y(y) = 0$ ; 当  $y > 0$  时有

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq x \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x) dx ,$$

根据 (4.2) 得到密度函数

$$f_Y(y) = F'_Y(y) = f_X(\sqrt{y})/2\sqrt{y} + f_X(-\sqrt{y})/2\sqrt{y} .$$

综上所述有

$$f_Y(y) = \begin{cases} (f_X(-\sqrt{y}) + f_X(\sqrt{y})) / 2\sqrt{y} & y > 0 , \\ 0 & y \leq 0 . \end{cases}$$

**例 4.14** 设随机变量  $X$  的密度函数为

$$f(x) = \begin{cases} x/8 & 0 < x < 4 \\ 0 & \text{其它,} \end{cases}$$

求  $Y = 2X + 8$  的密度函数.

**解** 首先求解分布函数

$$F_Y(y) = P(Y \leq y) = P(2X + 8 \leq y) = P(X \leq (y - 8)/2) = F_X((y - 8)/2) ,$$

求导可得密度函数

$$f_Y(y) = f_X((y - 8)/2)/2 = \begin{cases} (y - 8)/32 & (y - 8)/2 \in [0, 4] \\ 0 & \text{其它} \end{cases} = \begin{cases} (y - 8)/32 & y \in [8, 16] \\ 0 & \text{其它} . \end{cases}$$

针对一般情况有如下定理:

**定理 4.8** 设随机变量  $X$  的概率密度是定义在实数域上的函数  $f_X(x)$ , 函数  $y = g(x)$  处处可导且严格单调 (即  $g'(x) > 0$  或  $g'(x) < 0$ ), 令其反函数  $x = g^{-1}(y) = h(y)$ , 则  $Y = g(X)$  的概率密度为

$$f_Y(y) = \begin{cases} f_X(h(y))|h'(y)| & y \in (\alpha, \beta) \\ 0 & \text{其它,} \end{cases}$$

其中  $\alpha = \min\{g(-\infty), g(+\infty)\}$  和  $\beta = \max\{g(-\infty), g(+\infty)\}$ .

可将上述定理推广至区间函数  $x \in [a, b]$ , 上述定理依旧成立, 此时有  $\alpha = \min\{g(a), g(b)\}$  和  $\beta = \max\{g(a), g(b)\}$ .

**证明** 证明思路与前面的例题类似, 这里不妨假设  $g'(x) > 0$ , 同理可以考虑  $g'(x) < 0$  的情况. 根据  $g'(x) > 0$  可知其反函数  $x = h(y)$  也严格单调, 且  $g(x) \in [\alpha, \beta]$ . 因此, 当  $y \leq \alpha$  时, 有  $F_Y(y) = 0$ ; 当  $y \geq \beta$  时有  $F_Y(y) = 1$ ; 当  $\alpha < y < \beta$  时,

$$F_Y(y) = P(g(X) < y) = P(X \leq h(y)) = F(h(y)).$$

于是有随机变量  $Y$  的概率密度

$$f_Y(y) = F'(h(y)) \cdot h'(y) = f_X(h(y)) \cdot h'(y).$$

根据  $x = h(y)$  严格单调可知  $h'(y) > 0$ .

**定理 4.9** 设  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 则  $Y = aX + b$  ( $a > 0$ ) 服从正太分布  $\mathcal{N}(a\mu + b, a^2\sigma^2)$ .

**证明** 设函数  $g(x) = ax + b$ , 可得  $\alpha = -\infty$ ,  $\beta = +\infty$ , 以及  $y = g(x)$  的反函数为

$$x = h(y) = (y - b)/a,$$

且有  $h'(y) = 1/a$ . 根据定理 4.8 可知

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{y-b}{a} - \mu\right)^2 / 2\sigma^2} = \frac{1}{\sqrt{2\pi}a\sigma} e^{-(y-b-a\mu)^2 / 2a^2\sigma^2},$$

由此证明了  $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ .

**例 4.15** 设随机变量  $X$  的分布函数  $F_X(x)$  是严格单调的连续函数, 则  $Y = F_X(X) \sim U(0, 1)$ .

**证明** 令  $Y = F_X(X)$  的分布函数为  $G(y)$ , 则

$$G(y) = P(Y \leq y) = P(F_X(X) \leq y).$$

因为分布函数  $F_X(x) \in [0, 1]$ , 当  $y < 0$  时有  $G(y) = 0$ ; 当  $y \geq 1$  时有  $G(y) = 1$ ; 当  $y \in [0, 1]$  时, 由于  $F_X(X)$  严格单调, 所以  $F_X^{-1}(y)$  存在且严格单调, 于是有

$$G(y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

综上所述有密度函数

$$f_Y(y) = \begin{cases} 1 & y \in [0, 1] \\ 0 & \text{其它} \end{cases}.$$

## 4.6 常用分布的随机数\*

随机数在计算机仿真学和密码学等领域具有广泛的应用, 通过产生大量的随机数据来实现对真实世界的模拟. 可以利用物理随机过程来产生真实的随机数, 例如反复抛掷硬币、骰子、抽签、摇号等, 这些方法可以得到质量很高的随机数, 但其数量和类型通常较少、难以满足实际的需求.

现在的主流方法是使用计算机产生伪随机数, 通过计算机的确定算法来生成“类似”真随机数的伪随机数. 由于算法给出的结果总是确定的, 所以伪随机数并不是真正的随机数, 但是好的伪随机数序列与真实随机数序列表现几乎相同, 很难进行区分.

本节首先介绍如何生成在  $(0, 1)$  上均匀分布的随机数, 然后根据此随机数构造其它常用分布的随机数. 这里仅给出具体的构造方法, 关于其中的原理, 有兴趣的读者可以参考相关书籍.

### 4.6.1 区间 $(0, 1)$ 上均匀分布的随机数

目前有很多方法生成  $(0, 1)$  上均匀分布的随机数, 这里介绍最常用的线性同余法. 通过下面的迭代方式产生一系列随机数  $x_1, x_2, \dots, x_k$ ,

$$x_i = ax_{i-1} + c \quad \text{mod } m,$$

其中  $1 \leq i \leq k \leq m$ , 常数  $x_0$  为初始给定的种子. 为了获得较好的随机性, 通常  $m$  的取值应足够大, 如  $m = 2^k$  ( $k = 31, 63$ ), 常数  $c$  与  $m$  互质, 常数  $a - 1$  被  $m$  的因子整除, 例如一种可行的选择是

$$x_i = 31415926x_{i-1} + 453806245 \quad \text{mod } 2^{31}.$$

最后得到在区间  $(0, 1)$  上均匀分布的随机数  $x_1/m, x_2/m, \dots, x_k/m$ . 很多时候会根据实际情况选择不同的初始种子  $x_0$ . 有了在  $(0, 1)$  上均匀分布的随机数, 则可以构造一些服从常用分布的随机数.

### 4.6.2 常用离散型分布的随机数

**定理 4.10** 若随机变量  $X \sim U(0, 1)$ , 以及  $F(y)$  是某个离散型随机变量的分布函数, 其可能的取值为  $\{y_1, y_2, \dots\}$ , 不妨假设  $y_1 < y_2 < \dots$ . 设随机变量

$$Y = \begin{cases} y_1 & X \leq F(y_1) \\ y_i & X \in (F(y_{i-1}), F(y_i)] \quad (i \geq 2), \end{cases}$$

则  $Y$  的分布函数  $F_Y(y) = F(y)$  (如图 4.8 所示).

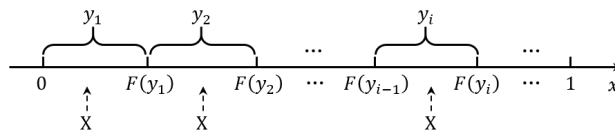


图 4.8 离散分布的随机数生成

**证明** 当  $y < y_1$  时有

$$F_Y(y) = P(Y \leq y) \leq P(Y < y_1) = 0 = F(y) .$$

当  $y_1 \leq y < y_2$  时根据  $X \sim U(0, 1)$  有

$$F_Y(y) = P(Y \leq y) = P(Y = y_1) = P(X \leq F(y_1)) = F(y_1) = F(y) .$$

当  $y \geq y_2$  时再次根据  $X \sim U(0, 1)$  有

$$F_Y(y) = P(Y \leq y) = \sum_{y_i: y_i \leq y} P(Y = y_i) = \sum_{y_i: y_i \leq y} P(X \in (F(y_{i-1}), F(y_i)]) = \max_{y_i: y_i \leq y} F(y_i) = F(y) .$$

定理 4.10 说明可以利用均匀分布的随机数来构造其它离散型随机变量的随机数, 以下以伯努利分布为例, 可类似生成其它常用分布的随机数.

**例 4.16** 已知在区间  $(0, 1)$  上均匀分布的随机数  $x_1, x_2, \dots, x_k$  ( $k \geq 1$ ), 如何生成参数为  $p$  的伯努利分布的随机数.

**解** 设随机变量  $Y$  服从参数为  $p$  的伯努利分布, 则  $Y$  的分布函数  $F(0) = 1 - p$  和  $F(1) = 1$ . 根据定理 4.10, 当  $1 \leq i \leq k$  时构造随机数

$$y_i = \begin{cases} 0 & x_i \leq F(0) = 1 - p \\ 1 & x_i \in (F(0), F(1)] = (1 - p, 1] , \end{cases}$$

其服从参数  $p$  的伯努利分布.

#### 4.6.3 常用连续型分布的随机数

**定理 4.11** 若随机变量  $X \sim U(0, 1)$ , 且  $F(y)$  是某一个连续的分布函数, 很显然反函数  $F^{-1}(y)$  存在, 则随机变量

$$Y = F^{-1}(X)$$

的分布函数为  $F_Y(y) = F(y)$ .

**证明** 根据分布函数的定义和  $F(x)$  的单调性有

$$F_Y(y) = P(Y \leq y) = P(F^{-1}(X) \leq y) = P(X \leq F(y)) .$$

再根据分布函数  $F(y)$  的非负性和随机变量  $X \sim U(0, 1)$  有

$$F_Y(y) = F(y) ,$$

由此完成证明.

**例 4.17** 若已知在区间  $(0, 1)$  上均匀分布的随机数  $x_1, x_2, \dots, x_k$  ( $k \geq 1$ ), 如何生成参数为  $\lambda = 4$  的指数分布的随机数.

**解** 若随机变量  $Y$  服从参数为  $\lambda = 4$  的指数分布, 则有连续的分函数  $F_Y(y) = 1 - e^{-4\lambda y}$ , 以及反函数  $F_Y^{-1}(y) = -\ln(1 - y)/4$ . 根据定理 4.11 可构造随机数

$$y_i = -\ln(1 - x_i)/4 \quad (i \in [k])$$

服从参数为  $\lambda = 4$  的指数分布.

由于正太分布的分函数不存在显示表达式, 所以它的反函数也不存在显示表达式, 因此不能利用定理 4.11 来直接生成服从正太分布的随机数. 这里简要介绍一种生成标准正太分布的随机数的方法、以及一些相关的结论, 详细的证明可参考相关书籍.

设  $X$  和  $Y$  是相互独立的标准正太分布随机变量, 则它们的极坐标

$$\begin{cases} R = \sqrt{X^2 + Y^2} \\ \theta = \arctan(Y/X) \end{cases}$$

相互独立, 且  $R^2$  服从参数为  $1/2$  的指数分布, 而  $\theta$  服从在  $(0, 2\pi)$  上的均匀分布. 设相互独立的随机变量  $X_1$  和  $X_2$  在区间  $(0, 1)$  上服从均匀分布, 则有

$$R = (-2 \ln X_1)^{1/2} \quad \text{和} \quad \theta = 2\pi X_2 .$$

由此可得相互独立的标准正太分布随机变量

$$X = R \cos(\theta) = (-2 \ln X_1)^{1/2} \cos(2\pi X_2) \quad \text{和} \quad Y = R \sin(\theta) = (-2 \ln X_1)^{1/2} \sin(2\pi X_2) ,$$

这种产生标准正太分布随机变量的方法被称为 **Box-Muller 方法**.

## 习题

4.1 若随机变量  $X$  的取值区间为  $[1, c]$ , 且落入  $[1, c]$  任意小区间的概率与小区间的长度成正比, 求  $X$  的分布函数.

4.2 用随机变量  $X$  表示某银行从下午开始营业起到第一个顾客到达的等待时间 (分), 设  $X$  的分布函数为

$$F(x) = \begin{cases} 1 - c \exp(-x/8) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

求这些事件的概率: 1)  $P(X \leq 4)$ ; 2)  $P(X \geq 8)$ ; 3)  $P(4 \leq X \leq 8)$ ; 4)  $P(X \leq 4 \text{ 或 } X \geq 8)$ ; 5)  $P(X = 6)$ .

4.3 若随机变量  $X$  的分布函数

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

求随机变量  $X$  的密度函数.

4.4 若随机变量  $X$  的密度函数

$$f(x) = \begin{cases} x & x \in [0, 1) \\ c - x & x \in [1, 2) \\ 0 & \text{其它,} \end{cases}$$

求随机变量  $X$  的分布函数, 并画出分布函数和密度函数.

4.5 已知长方形的宽服从均匀分布  $U(0, 2)$  (单位: 米), 以及长方形的面积为 10 (单位: 平方米), 求长方形的周长的期望与方差.

4.6 设随机变量  $X$  的概率密度为

$$f(x) = \begin{cases} Ae^{-x} & x > 0 \\ 0 & x \leq 0. \end{cases}$$

求  $Y = e^{-2X}$  的期望.

4.7 已知随机变量  $X \sim U(0, 1)$ , 对任意  $\lambda > 0$  求  $E[\lambda^{\max(X, 1-X)}]$ .



4.8 电池的故障是电动汽车的核心问题, 设相继两次事故之间的时间  $T$  服从参数为  $1/40$  的指数分布, 求概率  $P(X > 45)$ , 以及求最小的  $\tau$  使得  $P(X > t) \geq 60\%$ .

4.9 设乘客在一公交车站等待公交车的时间服从参数为  $1/6$  的指数分布, 某乘客若等待时间超过 10 分钟则换乘出租车离开. 该乘客一个月内有 10 天乘公交站 (每天是否乘出租车相互独立), 用  $Y$  表示该乘客因未等到公交车而换乘出租车的次数, 求  $Y$  的分布函数.

4.10 设随机变量  $X$  服从瑞利分布, 其概率密度为

$$f(x) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} & x > 0 \\ 0 & x \leq 0, \end{cases}$$

求期望  $E(X)$  和方差  $\text{Var}(X)$ .

4.11 设随机变量  $X \sim \mathcal{N}(4, 49)$ , 问题: i) 求概率  $P(3 < X \leq 7)$ ,  $P(|X| \leq 4)$  和  $P > 6$ ; ii) 求常数  $c$  使得  $P(X > c) = P(X \leq c)$ ; iii) 求常数  $a$  至多有多大时满足  $P(X > d) \geq 0.9$ .

4.12 设一批产品的寿命  $X \sim \mathcal{N}(180, \sigma^2)$  ( $\sigma > 0$ ), 要保证有  $P(140 < X \leq 220) \geq 0.9$  成立, 则  $\sigma$  最大值是多少.

4.13 证明

$$\int_{-\infty}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \sqrt{2\pi}\sigma.$$

4.14 若  $X \sim N(0, 1)$ , 对任意实数  $\epsilon > 0$ , 求证

$$P(X \geq \epsilon) \geq \frac{1}{3} e^{-\frac{(\epsilon+1)^2}{2}}$$

4.15 已知随机变量  $X$  的概率密度函数  $f_X(x)$ , 求随机变量  $Y = |X|$  的概率密度  $f_Y(y)$ .

4.16 若随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$ , 求随机变量  $Y = e^X$  的密度函数.

4.17 若随机变量  $X \sim \mathcal{N}(0, 1)$ , 求随机变量  $Y = X^2$  和  $Z = 2X^2 + 1$  的密度函数.

4.18 若随机变量  $X \sim e(\lambda)$ , 求随机变量  $Y = aX$  ( $a > 0$ ) 的概率分布.

4.19 若随机变量  $X \sim U(0, 1)$ , 求随机变量  $Y = -\ln X$  和  $Z = e^X$  的概率分布.

4.20 设随机变量  $X$  的密度函数为

$$f(x) = \begin{cases} 2x/\pi & x \in (0, \pi) \\ 0 & \text{其它} . \end{cases}$$

求  $Y = \sin X$  的密度函数.

- 4.21 大作业:** 编程生成 100000 在  $(0, 1)$  区间上均匀分布的随机数, 并以此生成伯努利分布、二项分布、泊松分布、几何分布、指数分布、正太分布的随机数, 需要将各种分布的参数作为输入变量, 最后查资料验证方法的正确性.

## 第5章 多维随机向量

前面讨论了一维随机变量的概率分布, 然而很多在实际问题中, 随机现象可能需要两种或两种以上的随机因素来描述, 仅仅用一个随机变量是不够的, 需要多个随机变量. 例如, 为了考察某地区儿童的身体素质时, 可以同时考虑他们的身高、体重、肺活量、视力等, 此时至少需要四个随机变量来描述. 这些随机变量之间可能存在某些关联, 因此分别对每个随机变量单独进行研究是不够的, 需要将其看作一个整体来研究, 即多维随机向量.

**定义 5.1** 设  $X_1 = X_1(\omega), X_2 = X_2(\omega), \dots, X_n = X_n(\omega)$  是定义在同一样本空间  $\Omega$  上的  $n$  个随机变量, 由它们构成的向量  $(X_1, X_2, \dots, X_n)$  称为  $n$  维随机向量, 或称  $n$  维随机变量.

一维随机变量可以看作多维随机变量的一种特殊情况, 本章主要讨论二维随机向量及其分布, 同理可讨论二维以上的随机向量.

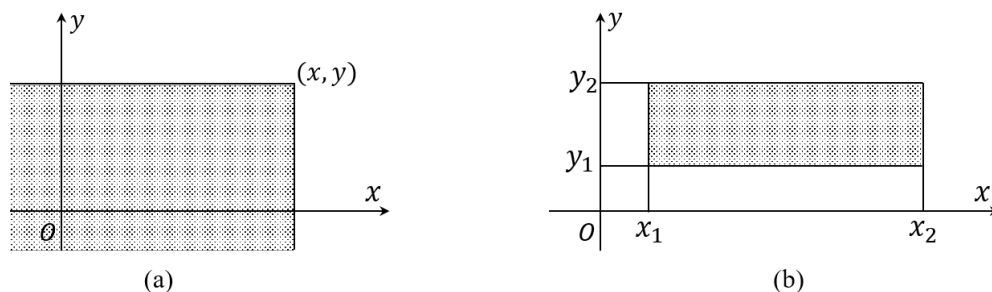
### 5.1 二维联合分布函数

类似于一维随机变量, 我们用分布函数来研究二维随机向量的概率特性.

**定义 5.2** 设  $(X, Y)$  为二维随机向量, 对任意实数  $x$  和  $y$ ,

$$F(x, y) = P(X \leq x, Y \leq y)$$

称为二维随机向量  $(X, Y)$  的 **分布函数**, 或称随机变量  $X$  和  $Y$  的 **联合分布函数** (joint cumulative probability distribution function).



**图 5.1** 随机向量  $(X, Y)$  的分布函数  $F(x, y)$  和概率  $P(x_1 < X \leq x_2, y_1 < Y \leq y_2)$

若将  $(X, Y)$  看作平面上随机点的坐标, 则分布函数  $F(x, y)$  的值表示随机向量  $(X, Y)$  落入以  $(x, y)$  为顶点的左下方无穷区域的概率, 如图 5.1(a) 所示. 再根据图 5.1(b) 可知, 随机向量  $(X, Y)$  落

入矩形区域  $\{(x, y): x_1 < x \leq x_2, y_1 < y \leq y_2\}$  的概率

$$P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) .$$

二维随机向量  $(X, Y)$  的分布函数  $F(x, y)$  具有以下性质:

- 1) 分布函数  $F(x, y)$  对每个变量都是单调不减的, 即对任意固定的实数  $y$ , 当  $x_1 > x_2$  时有  $F(x_1, y) \geq F(x_2, y)$ ; 对任意固定的实数  $x$ , 当  $y_1 > y_2$  时有  $F(x, y_1) \geq F(x, y_2)$ .
- 2) 对任意实数  $x$  和  $y$ , 分布函数  $F(x, y) \in [0, 1]$ , 而且

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

- 3) 分布函数  $F(x, y)$  关于每个变量右连续, 即

$$F(x, y) = F(x + 0, y) \quad \text{和} \quad F(x, y) = F(x, y + 0) .$$

- 4) 对任意实数  $x_1 < x_2$  和  $y_1 < y_2$  有

$$F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \geq 0 .$$

任何的分布函数  $F(x, y)$  都满足上述四条性质, 前三条性质与一维随机变量类似, 第四条性质根据图 5.1(b) 直接可证. 反之, 任何满足上面四条性质的二元函数  $F(x, y)$  都可看成某二维随机向量的分布函数.

值得说明的是, 当二元函数  $F(x, y)$  仅仅满足前面的三条性质时, 并不一定能成为某二维随机向量的分布函数, 例如

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

很容易验证  $F(x, y)$  仅仅满足前面的三条性质, 但因为

$$F(1, 1) - F(1, -1) - F(-1, 1) + F(-1, -1) = -1 ,$$

如图 5.2(a) 所示, 不满足第四条性质因此不构成一个分布函数.

根据随机向量  $(X, Y)$  的联合分布函数  $F(x, y)$ , 还可以研究每个随机变量的统计特征, 即将  $X$  和  $Y$  看做单独的随机变量, 通过联合分布函数  $F(x, y)$  来研究随机变量  $X$  和  $Y$  的分布函数  $F_X(x)$  和  $F_Y(y)$ , 即边缘分布函数.

**定义 5.3** 设二维随机向量  $(X, Y)$  的联合分布函数为  $F(x, y)$ , 称

$$F_X(x) = P(X \leq x) = P(X \leq x, y < +\infty) = F(x, +\infty) = \lim_{y \rightarrow +\infty} F(x, y) ,$$

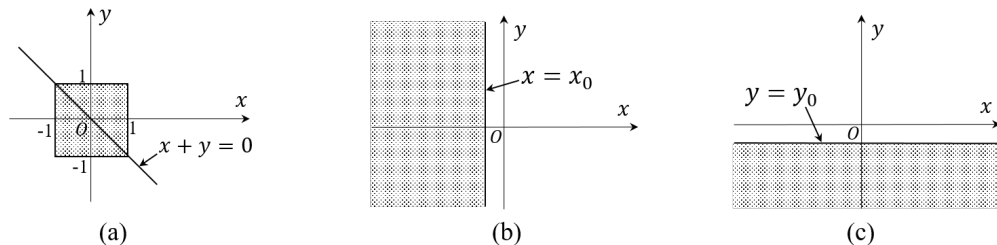


图 5.2 分布函数第四条性质的反例和边缘分布

为  $(X, Y)$  关于随机变量  $X$  的 **边缘分布函数** (marginal distribution function). 类似地定义  $(X, Y)$  关于随机变量  $Y$  的边缘分布函数

$$F_Y(y) = P(Y \leq y) = P(Y \leq y, x < +\infty) = F(+\infty, y) = \lim_{x \rightarrow +\infty} F(x, y).$$

边缘分布函数  $F_X(x_0)$  和  $F_Y(y_0)$  的值分别表示随机向量  $(X, Y)$  落入图 5.2(b) 和 5.2(c) 中阴影部分的概率. 下面来看一个例子.

**例 5.1** 设二维随机向量  $(X, Y)$  的分布函数为

$$F(x, y) = A(B + \arctan \frac{x}{2})(C + \arctan \frac{y}{3}) \quad (x, y \in \mathbb{R}).$$

求随机变量  $X$  与  $Y$  的边缘分布函数, 以及概率  $P(Y > 3)$ .

**解** 对任意  $x \in (-\infty, +\infty)$  和  $y \in (-\infty, +\infty)$ , 根据分布函数的性质有

$$\begin{aligned} 1 &= F(+\infty, +\infty) = A(B + \frac{\pi}{2})(C + \frac{\pi}{2}), \\ 0 &= F(x, -\infty) = A(B + \arctan \frac{x}{2})(C - \frac{\pi}{2}), \\ 0 &= F(-\infty, y) = A(B - \frac{\pi}{2})(C + \arctan \frac{y}{3}). \end{aligned}$$

求解上述方程可得

$$C = \frac{\pi}{2}, \quad B = \frac{\pi}{2}, \quad A = \frac{1}{\pi^2}.$$

从而得到  $F(x, y) = (\pi/2 + \arctan x/2)(\pi/2 + \arctan y/3)/\pi^2$ , 进一步得到

$$F_X(x) = \lim_{y \rightarrow \infty} \frac{1}{\pi^2} (\frac{\pi}{2} + \arctan \frac{x}{2})(\frac{\pi}{2} + \arctan \frac{y}{3}) = \frac{1}{\pi} (\frac{\pi}{2} + \arctan \frac{x}{2}),$$

同理可得

$$F_Y(y) = \frac{1}{\pi} (\frac{\pi}{2} + \arctan \frac{y}{3}).$$

最后得到

$$P(Y > 3) = 1 - P(Y \leq 3) = 1 - F_Y(3) = 1 - \left( \frac{1}{2} + \frac{1}{\pi} \arctan 1 \right) = \frac{1}{4}.$$

## 5.2 二维离散型随机向量

**定义 5.4** 若二维随机向量  $(X, Y)$  的取值是有限个或无限可列的, 则称  $(X, Y)$  为 **二维离散型随机向量**. 设离散型随机向量  $(X, Y)$  所有可能的取值为  $(x_i, y_j)$  ( $i, j = 1, 2, \dots$ ), 则称

$$p_{ij} = P(X = x_i, Y = y_j)$$

为二维随机向量  $(X, Y)$  的 **联合分布列**, 简称 **分布列**.

二维随机向量分布列具有下列性质:

$$p_{ij} \geq 0 \quad \text{和} \quad \sum_{i,j} p_{ij} = 1.$$

通过随机向量  $(X, Y)$  的联合分布列  $p_{ij}$ , 还可以研究每个随机变量的统计特征, 例如随机变量  $X$  的 **边缘分布列** 为

$$P(X = x_i) = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = \sum_{j=1}^{\infty} p_{ij} = p_{i\cdot},$$

以及随机变量  $Y$  的 **边缘分布列** 为

$$P(Y = y_j) = \sum_{i=1}^{\infty} P(X = x_i, Y = y_j) = \sum_{i=1}^{\infty} p_{ij} = p_{\cdot j}.$$

二维随机向量的联合分布列和边缘分布列可通过表 5.1 来进行表示.

**表 5.1** 二维随机向量的概率分布表

$X \backslash Y$	$y_1$	$y_2$	$\cdots$	$y_j$	$\cdots$	$p_{i\cdot} = \sum_j p_{ij}$
$x_1$	$p_{11}$	$p_{12}$	$\cdots$	$p_{1j}$	$\cdots$	$p_{1\cdot}$
$x_2$	$p_{21}$	$p_{22}$	$\cdots$	$p_{2j}$	$\cdots$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
$x_i$	$p_{i1}$	$p_{i2}$	$\cdots$	$p_{ij}$	$\cdots$	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$
$p_{\cdot j} = \sum_i p_{ij}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	1

根据二维随机变量  $(X, Y)$  的联合分布列  $p_{ij}$ , 可以得到它们的联合分布函数

$$F(x, y) = \sum_{x_i \leq x, y_j \leq y} p_{ij},$$

和边缘分布函数

$$F_X(x) = \sum_{x_i \leq x} p_{i\cdot} = \sum_{x_i \leq x} \sum_{j=1}^{+\infty} p_{ij} \quad \text{和} \quad F_Y(y) = \sum_{y_j \leq y} p_{\cdot j} = \sum_{y_j \leq y} \sum_{i=1}^{+\infty} p_{ij}.$$

**例 5.2** 假设某地区有 15% 的家庭没小孩, 20% 的家庭有一个小孩, 35% 的家庭有两个小孩, 30% 的家庭有三个小孩, 且假设每个小孩为男孩或女孩是相互独立且等可能的. 随机选择一个家庭, 用随机变量  $X, Y$  分别表示该家庭中男孩和女孩的个数, 求  $P(X \geq 1)$ ,  $P(Y \leq 2)$  和  $P(X \leq Y)$ .

**解** 根据题意有  $X, Y$  的所有可能取值为  $\{0, 1, 2, 3\}$ , 进一步有联合分布列

$$\begin{aligned} P(X = i, Y = j) &= P(\text{选择的家庭有 } i + j \text{ 个小孩, 其中 } i \text{ 个男孩和 } j \text{ 个女孩}) \\ &= P(\text{选择的家庭有 } i + j \text{ 个小孩})P(i \text{ 个男孩和 } j \text{ 个女孩} | \text{选择的家庭有 } i + j \text{ 个小孩}) \\ &= \binom{i+j}{i} \frac{1}{2^{i+j}} P(\text{选择的家庭有 } i + j \text{ 个小孩}), \end{aligned}$$

由此可得联合分布列和边缘分布列为

$X \backslash Y$	0	1	2	3	$p_{i\cdot}$
0	0.1500	0.1000	0.0875	0.0375	0.3750
1	0.1000	0.175	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2000
3	0.0375	0	0	0	0.0375
$p_{\cdot j}$	0.3750	0.3875	0.2000	0.0375	1

最后得到

$$P(X \geq 1) = 0.625, \quad P(Y \leq 2) = 0.9625, \quad P(X \leq Y) = 0.6625.$$

最后介绍一种常用的多维离散分布: 多项分布, 它本质上是二项分布的推广, 可用于机器学习中的多分类问题. 假设试验  $E$  有  $n$  种可能的结果  $A_1, A_2, \dots, A_n$ , 每种结果发生的概率  $p_i = P(A_i)$ , 则有  $p_1 + p_2 + \dots + p_n = 1$ .

将试验  $E$  独立重复地进行  $m$  次, 用  $X_1, X_2, \dots, X_n$  分别表示事件  $A_1, A_2, \dots, A_n$  发生的次数, 则每个随机变量  $X_i$  的取值为  $\{0, 1, 2, \dots, m\}$  且满足  $X_1 + X_2 + \dots + X_n = m$ , 则随机向量  $(X_1, X_2, \dots, X_n)$  服从多项分布, 其严格的定义如下:

**定义 5.5** 若  $n$  维随机向量  $(X_1, X_2, \dots, X_n)$  的分布列为

$$P(X_1 = k_1, X_2 = k_2, \dots, X_n = k_n) = \binom{m}{k_1, k_2, \dots, k_n} p_1^{k_1} p_2^{k_2} \dots p_n^{k_n},$$

其中  $k_1, k_2, \dots, k_n$  是非负的整数且满足  $k_1 + k_2 + \dots + k_n = m$ , 则称随机向量  $(X_1, X_2, \dots, X_n)$  服从参数为  $m, p_1, p_2, \dots, p_n$  的 **多项分布** (multinomial distribution), 记为  $(X_1, X_2, \dots, X_n) \sim M(m, p_1, p_2, \dots, p_n)$ .

很容易验证  $P(X_1 = k_1, X_2 = k_2, \dots, X_n = k_n) \geq 0$  以及

$$\begin{aligned} & \sum_{k_i \geq 0, k_1 + k_2 + \dots + k_n = m} P(X_1 = k_1, X_2 = k_2, \dots, X_n = k_n) \\ &= \sum_{k_i \geq 0, k_1 + k_2 + \dots + k_n = m} \binom{m}{k_1, k_2, \dots, k_n} p_1^{k_1} p_2^{k_2} \dots p_n^{k_n} = (p_1 + p_2 + \dots + p_n)^m = 1. \end{aligned}$$

当  $n = 2$  时多项分布简化为二项分布.

**引理 5.1** 若多维随机向量  $(X_1, X_2, \dots, X_n) \sim M(m, p_1, p_2, \dots, p_n)$ , 则每个随机变量  $X_i$  的边缘分布是二项分布  $B(m, p_i)$ .

根据  $X_i$  的实际含义, 考虑事件  $A_i$  发生或不发生的伯努利试验, 则有  $X_i \sim (m, p_i)$ . 另一种方法是通过多项分布的定义直接计算, 我们将其作为一个作业题.

### 5.3 二维连续型随机向量

**定义 5.6** 设二维随机向量  $(X, Y)$  的分布函数为  $F(x, y)$ , 如果存在二元非负可积函数  $f(x, y)$ , 使得对任意实数  $x$  和  $y$  有

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

则称  $(X, Y)$  为 **二维连续型随机向量**, 称  $f(x, y)$  为二维随机向量  $(X, Y)$  的 **密度函数**, 或称随机变量  $X$  和  $Y$  的 **联合密度函数**.

联合密度函数  $f(x, y)$  满足如下性质:

1) 非负性: 对任意实数  $x$  和  $y$  有  $f(x, y) \geq 0$ .

2) 规范性:  $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$ .

任何满足上面两条性质的二元函数  $f(x, y)$  可以成为某随机向量  $(X, Y)$  的联合密度函数.

3) 若  $G$  为平面上的一个区域, 则点  $(X, Y)$  落入  $G$  的概率为

$$P((X, Y) \in G) = \iint_{(x, y) \in G} f(x, y) dx dy,$$

在几何上可以看作是以  $G$  为底面,  $z = f(x, y)$  为顶面的柱体体积, 如图 5.3(a) 所示.



4) 若密度函数  $f(x, y)$  在  $(x, y)$  连续, 则联合分布函数  $F(x, y)$  和密度函数  $f(x, y)$  满足

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

根据此性质、并利用多元泰勒展开式有

$$\begin{aligned} & \lim_{\substack{\Delta x \rightarrow 0^+ \\ \Delta y \rightarrow 0^+}} \frac{P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)}{\Delta x \Delta y} \\ &= \lim_{\substack{\Delta x \rightarrow 0^+ \\ \Delta y \rightarrow 0^+}} \frac{F(x + \Delta x, y + \Delta y) - F(x + \Delta x, y) - F(x, y + \Delta y) + F(x, y)}{\Delta x \Delta y} \\ &= \frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y) , \end{aligned}$$

由此可知

$$P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y) \approx f(x, y) \Delta x \Delta y ,$$

概率  $f(x, y)$  的值反映了二维随机向量  $(X, Y)$  落入  $(x, y)$  邻域内概率的大小.

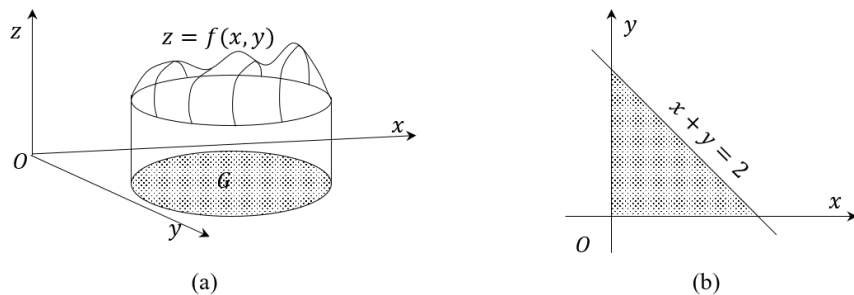


图 5.3 二维密度函数的几何意义和例 5.3 的积分区域

根据  $(X, Y)$  的联合密度函数  $f(x, y)$ , 还可以研究每个随机变量  $X$  和  $Y$  的密度函数  $f_X(x)$  和  $f_Y(y)$ . 首先考虑随机变量  $X$  的边缘分布

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(X \leq x, Y < \infty) = F(x, +\infty) \\ &= \int_{-\infty}^x \int_{-\infty}^{+\infty} f(t, y) dt dy = \int_{-\infty}^x \left( \int_{-\infty}^{+\infty} f(t, y) dy \right) dt , \end{aligned}$$

对上式两边求导得到  $X$  的边缘概率密度

$$f_X(x) = F'_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy .$$

同理分析随机变量  $Y$  的边缘分布, 于是得到边缘概率密度的严格定义.

**定义 5.7** 设二维随机向量  $(X, Y)$  的联合密度函数为  $f(x, y)$ , 则随机变量  $X$  和  $Y$  的 **边缘密度函数** 分别为

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad \text{和} \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx .$$

**例 5.3** 设二维随机变量  $(X, Y)$  的概率密度为

$$f(x, y) = \begin{cases} ce^{-(3x+4y)} & x > 0, y > 0 \\ 0 & \text{其它} . \end{cases}$$

求: 1) 常数  $c$ ; 2) 联合分布函数  $F(x, y)$ ; 3)  $X$  和  $Y$  的边缘概率密度; 4) 概率  $P(X + Y \leq 2)$ .

**解** 根据密度函数的性质可知

$$1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ce^{-(3x+4y)} dx dy = \frac{c}{12},$$

求解出  $c = 12$ . 当  $x > 0$  和  $y > 0$  时有

$$F(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = \int_0^x \int_0^y 12e^{-(3x+4y)} dx dy = (1 - e^{-3x})(1 - e^{-4y}) ,$$

进一步根据边缘概率密度的定义有

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_0^{+\infty} 12e^{-(3x+4y)} dy = 3e^{-3x} ,$$

同理可得  $f_Y(y) = 4e^{-4y}$ . 最后计算概率  $P(X + Y \leq 2)$ , 其积分区域如图 5.3(b) 所示, 有

$$P(X + Y \leq 2) = 12 \int_0^2 dx \int_0^{2-x} e^{-(3x+4y)} dy = 3 \int_0^2 e^{-3x} (1 - e^{-8+4x}) dx = 1 - 4e^{-6} + 3e^{-8} .$$

### 5.3.1 常用二维连续分布

下面介绍两种常用的二维连续分布: 均匀分布和正太分布.

**定义 5.8** 设  $G$  为平面上一个有界的区域, 其面积为  $A_G$ , 若二维随机向量  $(X, Y)$  的联合密度函数为

$$f(x, y) = \begin{cases} 1/A_G & (x, y) \in G \\ 0 & (x, y) \notin G, \end{cases}$$

则称  $(X, Y)$  服从区域  $G$  上的 **二维均匀分布**.

二维均匀分布在区域  $G$  上每一点等可能发生, 本质上就是 (平面) 几何概型的随机向量描述. 这里以圆的均匀分布为例, 可类似考虑三角形、椭圆等平面上一个有界区域的均匀分布.

**例 5.4** 在一个以坐标原点为中心、半径为  $R$  的圆内等可能随机投点. 用随机向量  $(X, Y)$  分别表示落点的横坐标和纵坐标, 求: 随机向量  $(X, Y)$  的联合密度函数, 边缘密度函数, 以及  $(X, Y)$  落入  $X^2 + Y^2 \leq r^2$  ( $0 < r \leq R$ ) 的概率.

**解** 很容易得到圆的面积为  $\pi R^2$ , 由此可知随机向量  $(X, Y)$  的联合密度函数

$$f(x, y) = \begin{cases} 1/\pi R^2 & x^2 + y^2 \leq R^2 \\ 0 & x^2 + y^2 > R^2 \end{cases}.$$

对于随机变量  $X$  的边缘密度函数, 当  $x^2 \leq R^2$  时有

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_{x^2 + y^2 \leq R^2} \frac{1}{\pi R^2} dy = \int_{-\sqrt{R^2 - x^2}}^{+\sqrt{R^2 - x^2}} \frac{1}{\pi R^2} dy = \frac{2}{\pi R^2} \sqrt{R^2 - x^2},$$

同理可得随机变量  $Y$  的边缘密度函数. 最后所求概率

$$P(X^2 + Y^2 \leq r^2) = \iint_{x^2 + y^2 \leq r^2} \frac{1}{\pi R^2} dx dy = \frac{r^2}{R^2}.$$

二维连续分布中最重要的是二维正太分布, 其定义如下:

**定义 5.9** 对任意实数  $x, y$ , 若随机向量  $(X, Y)$  的密度函数为

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right),$$

其中常数  $\mu_x, \mu_y \in (-\infty, +\infty)$ ,  $\sigma_x, \sigma_y \in (0, +\infty)$  以及  $\rho \in (-1, 1)$ , 则称  $(X, Y)$  服从 **二维正太分布**, 记  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ .

下面研究二维正态分布的性质:

**定理 5.1** 设二维随机向量  $(X, Y)$  服从正态分布  $\mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则有随机变量  $X$  和  $Y$  的边缘分布分别为  $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$  和  $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ .

**证明** 这里将证明随机变量  $X$  的边缘密度函数, 可同理证明  $Y$  的边缘密度函数. 首先将二维随机向量  $(X, Y)$  的联合密度函数  $f(x, y)$  分解为

$$f(x, y) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{(y-\mu_y-\rho\sigma_y(x-\mu_x)/\sigma_x)^2}{2\sigma_y^2(1-\rho^2)}\right). \quad (5.1)$$

因此联合密度函数等于两个一维正太分布  $\mathcal{N}(\mu_x, \sigma_x)$  和  $\mathcal{N}(\mu_y + \rho\sigma_y(x - \mu_x)/\sigma_x, \sigma_y^2(1 - \rho^2))$  的密度函数的乘积. 给定  $x, \mu_x \in (-\infty, +\infty), \sigma_x > 0, \rho \in (-1, 1)$  则有

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{(y - \mu_y - \rho\sigma_y(x - \mu_x)/\sigma_x)^2}{2\sigma_y^2(1-\rho^2)}\right) dy = 1,$$

于是得到

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2}\right).$$

由此完成证明.

定理 5.1 说明正太分布的边缘分布还是正太分布, 并给出了二维正太分布前四个参数的意义, 即随机变量  $X$  和  $Y$  的期望和方差, 第五个参数反应了两个随机变量的密切程度, 我们将在后面介绍.

二维联合分布可以唯一确定它们的边缘分布, 但反之不成立, 即使知道两个随机变量的边缘分布, 也不足以决定联合分布. 例如, 两个边缘分布为  $\mathcal{N}(\mu_x, \sigma_x)$  和  $\mathcal{N}(\mu_y, \sigma_y)$ , 因为不能确定  $\rho$  的值而不能确定它们的联合分布. 基于 (5.1), 我们还可以验证二维正太分布的规范性

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1,$$

以及二维正太分布的密度函数本质是两个 (一维) 正太分布的密度函数的乘积.

## 5.4 随机变量的独立性

前面第二章介绍了随机事件的独立性, 即独立的随机事件  $A$  和  $B$  满足  $P(AB) = P(A)P(B)$ . 本节介绍概率统计中另一个重要的概念: 随机变量的独立性. 考虑两个随机变量, 若一个随机变量的取值对另一个随机变量没有什么影响, 则称两个随机变量相互独立. 下面给出严格的数学定义:

**定义 5.10** 设二维随机向量  $(X, Y)$  的联合分布函数为  $F(x, y)$ , 以及  $X$  和  $Y$  的边缘分布函数分别为  $F_X(x)$  和  $F_Y(y)$ , 若对任意的实数  $x$  和  $y$  有

$$F(x, y) = F_X(x)F_Y(y),$$

则称随机变量  $X$  与  $Y$  相互独立.

根据上面的定义可知, 随机变量  $X$  与  $Y$  相互独立等价于随机事件  $\{X \leq x\}$  和  $\{Y \leq y\}$  对任意实数  $x$  和  $y$  都相互独立; 容易发现常数  $c$  与任意随机变量相互独立.

对于离散型随机向量, 可以考虑通过分布列来刻画它的统计规律, 关于独立性有

**定理 5.2** 设二维离散型随机向量  $(X, Y)$  的分布列为  $p_{ij} = P(X = x_i, Y = y_j)$  ( $i, j = 1, 2, \dots$ ), 以及  $X$  和  $Y$  的边缘分布列为  $p_{i\cdot} = P(X = x_i)$  和  $p_{\cdot j} = P(Y = y_j)$ , 则随机变量  $X$  和  $Y$  相互独立的充要条件是  $p_{ij} = p_{i\cdot}p_{\cdot j}$ .

**证明** 首先证明必要性, 根据定义 5.10 分布函数的独立性有

$$\begin{aligned}
 p_{i,j} &= F(x_i, y_j) - F(x_{i-1}, y_j) - F(x_i, y_{j-1}) + F(x_{i-1}, y_{j-1}) \\
 &= F_X(x_i)F_Y(y_j) - F_X(x_{i-1})F_Y(y_j) - F_X(x_i)F_Y(y_{j-1}) + F_X(x_{i-1})F_Y(y_{j-1}) \\
 &= (F_X(x_i) - F_X(x_{i-1}))F_Y(y_j) - (F_X(x_i) - F_X(x_{i-1}))F_Y(y_{j-1}) \\
 &= p_{i\cdot}F_Y(y_j) - p_{i\cdot}F_Y(y_{j-1}) = p_{i\cdot}p_{\cdot j} .
 \end{aligned}$$

其次证明充分性, 根据  $p_{ij} = p_{i\cdot}p_{\cdot j}$  ( $i, j = 1, 2, \dots$ ) 有

$$F(x_m, y_n) = \sum_{i \leq m} \sum_{j \leq n} p_{ij} = \sum_{i \leq m} \sum_{j \leq n} p_{i\cdot}p_{\cdot j} = \sum_{i \leq m} p_{i\cdot} \times \sum_{j \leq n} p_{\cdot j} = F_X(x_m)F_Y(y_n) .$$

由此完成证明.

**例 5.5** 设离散型随机变量  $X$  和  $Y$  相互独立且它们的取值均为  $\{1, 2, 3\}$ , 已知  $P(Y = 1) = 1/3$ ,  $P(X = 1, Y = 1) = P(X = 2, Y = 1) = 1/8$  和  $P(X = 1, Y = 3) = 1/16$ , 求  $X$  和  $Y$  的联合分布列和边缘分布列.

**解** 根据边缘分布列的定义有

$$P(X = 3, Y = 1) = P(Y = 1) - P(X = 1, Y = 1) - P(X = 2, Y = 1) = 1/12 ,$$

再根据定理 5.2 有  $P(X = 1) = P(X = 2) = 3/8$  和  $P(X = 3) = 1/4$ , 同理计算其它概率, 最后得到的分布列为

$X \backslash Y$	1	2	3	$p_{i\cdot}$
1	1/8	3/16	1/16	3/8
2	1/8	3/16	1/16	3/8
3	1/12	1/8	1/24	1/4
$p_{\cdot j}$	1/3	1/2	1/6	

对于连续型随机向量, 一般可以通过密度函数来进行刻画, 关于独立性有

**定理 5.3** 设二维随机向量  $(X, Y)$  的联合密度函数为  $f(x, y)$ , 及  $X$  和  $Y$  的边缘密度函数分别为  $f_X(x)$  和  $f_Y(y)$ , 则随机变量  $X$  和  $Y$  相互独立的充要条件是  $f(x, y) = f_X(x)f_Y(y)$ .

**证明** 首先证明必要性: 若二维连续随机变量满足  $F(x, y) = F_X(x)F_Y(y)$ , 则有

$$\int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv = \int_{-\infty}^x f_X(u) du \int_{-\infty}^y f_Y(v) dv ,$$

对上式两边同时求偏导有

$$f(x, y) = f_X(x)f_Y(y) .$$

其次证明充分性: 若  $f(x, y) = f_X(x)f_Y(y)$ , 则有

$$\begin{aligned} F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv = \int_{-\infty}^x \int_{-\infty}^y f_X(u) f_Y(v) du dv \\ &= \int_{-\infty}^x f_X(u) du \int_{-\infty}^y f_Y(v) dv = F_X(x) F_Y(y) , \end{aligned}$$

由此完成证明.

下面介绍关于随机变量独立性的一些性质:

**性质 5.1** 若随机变量  $X$  和  $Y$  相互独立, 则对任意给定的集合  $A, B \subseteq \mathbb{R}$ , 事件  $\{X \in A\}$  和事件  $\{Y \in B\}$  相互独立.

**证明** 该引理对离散型和连续型随机变量均成立, 这里我们详细证明连续随机变量情形. 根据独立性有  $f(x, y) = f_X(x)f_Y(y)$ , 由此可得

$$\begin{aligned} P(X \in A, Y \in B) &= \iint_{x \in A, y \in B} f(x, y) dx dy \\ &= \iint_{x \in A, y \in B} f_X(x) f_Y(y) dx dy = \int_{x \in A} f_X(x) dx \int_{y \in B} f_Y(y) dy = P(X \in A) P(Y \in B) , \end{aligned}$$

引理得证.

**性质 5.2** 设随机变量  $X$  和  $Y$  相互独立, 以及  $f(x)$  和  $g(y)$  是连续或分段连续的函数, 则有  $f(X)$  与  $g(Y)$  相互独立.

该定理对离散型和连续型随机变量均成立, 这里没给出它的证明是因此其超出了本书的范围. 根据此引理, 若随机变量  $X$  与  $Y$  相互独立, 则  $X^2$  与  $Y^3$  相互独立, 以及  $\sin X$  与  $\cos Y$  也相互独立.

下面给出一种判断两个随机变量独立性的简单方法:

**性质 5.3** 如果存在两个函数  $h(x)$  和  $g(y)$ , 使得  $X$  和  $Y$  的联合密度函数  $f(x, y)$  对任意实数  $x$  和  $y$  均有

$$f(x, y) = h(x)g(y) ,$$

则随机变量  $X$  和  $Y$  相互独立.

**证明** 不妨假设

$$C_1 = \int_{-\infty}^{+\infty} h(x) dx \quad \text{和} \quad C_2 = \int_{-\infty}^{+\infty} g(y) dy .$$

根据密度函数的规范性有

$$1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) = \int_{-\infty}^{+\infty} h(x) dx \int_{-\infty}^{+\infty} g(y) dy = C_1 C_2 .$$

随机变量  $X$  和  $Y$  的边缘分布分别为

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = h(x) C_2 \quad \text{和} \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx = C_1 g(y) .$$

于是得到

$$f_X(x) f_Y(y) = C_1 C_2 h(x) g(y) = h(x) g(y) = f(x, y) ,$$

由此完成证明.

**定理 5.4** 设二维随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则  $X$  与  $Y$  独立的充要条件为  $\rho = 0$ .

**证明** 首先证明必要性. 若随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则  $X$  和  $Y$  的边缘分布分别为  $\mathcal{N}(\mu_x, \sigma_x^2)$  和  $\mathcal{N}(\mu_y, \sigma_y^2)$ . 当  $\rho = 0$  时, 根据二维正太分布的定义有

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right) = f_X(x) f_Y(y) .$$

其次证明充分性. 若  $X$  与  $Y$  相互独立, 则对任意实数  $x$  和  $y$  均有  $f(x, y) = f_X(x) f_Y(y)$  成立, 即

$$\begin{aligned} & \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right) \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} \exp\left(-\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho}{\sigma_x\sigma_y} (x-\mu_x)(y-\mu_y) \right]\right) , \end{aligned}$$

取  $x = \mu_x$  和  $y = \mu_y$  求解出  $\rho = 0$ , 由此完成证明。

**例 5.6** 设二维随机向量的密度函数

$$f(x, y) = \begin{cases} cxe^{-y} & 0 < x < y < +\infty \\ 0 & \text{其它,} \end{cases}$$

问随机变量  $X$  与  $Y$  是否相互独立?

**解** 如图 5.4(a) 所示, 利用密度函数的规范性和分部积分有

$$1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = c \int_0^{+\infty} dy \int_0^y xe^{-y} dx = c .$$

当  $x > 0$  时随机变量  $X$  的边缘概率密度为

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_x^{+\infty} x e^{-y} dy = x e^{-x}.$$

同理当  $y > 0$  时随机变量  $Y$  的边缘概率密度为

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx = \int_0^y x e^{-y} dx = \frac{1}{2} y^2 e^{-y}.$$

因为  $f(x, y) \neq f_X(x)f_Y(y)$  可得随机变量  $X$  与  $Y$  不独立.

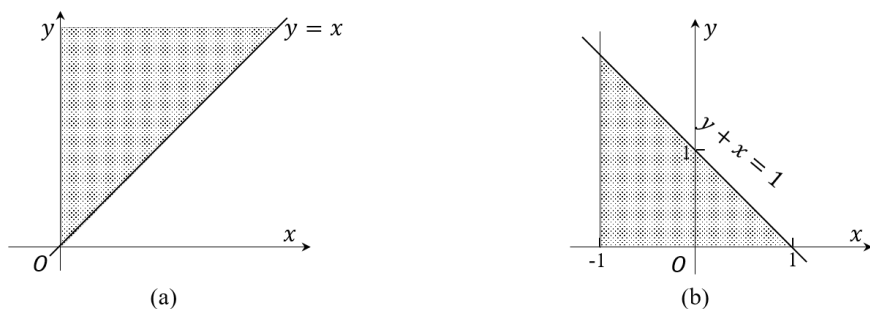


图 5.4 例 5.6 和 5.7 的积分区域

**例 5.7** 设随机变量  $X$  与  $Y$  相互独立, 且  $X$  服从  $[-1, 1]$  均匀分布,  $Y$  服从参数为  $\lambda = 2$  的指数分布, 求  $P(X + Y \leq 1)$ .

**解** 首先有随机变量  $X$  与  $Y$  的边缘概率密度分别为

$$f_X(x) = \begin{cases} \frac{1}{2} & x \in [-1, 1] \\ 0 & \text{其它} \end{cases} \quad \text{和} \quad f_Y(y) = \begin{cases} 2e^{-2y} & y \geq 0 \\ 0 & \text{其它} \end{cases}$$

根据独立性可得随机变量  $X$  与  $Y$  的联合概率密度

$$f(x, y) = \begin{cases} e^{-2y} & -1 \leq x \leq 1, y \geq 0 \\ 0 & \text{其它} \end{cases}.$$

所求积分区域如图 5.4(b) 所示, 最后得到

$$P(X + Y \leq 1) = \int_{-1}^1 dx \int_0^{1-x} e^{-2y} dy = \frac{3}{4} + \frac{1}{4} e^{-4}.$$



## 5.5 条件分布

前面第二章讨论了随机事件的条件概率, 即在事件  $B$  发生的条件下事件  $A$  发生的条件概率  $P(A|B) = P(AB)/P(B)$ . 可同理考虑随机变量的条件分布, 在给定随机变量  $Y$  具体取值的条件下, 考虑随机变量  $X$  的概率分布. 下面分离散和连续两种情形进行讨论.

### 5.5.1 离散型随机变量的条件概率

**定义 5.11** 设离散型随机变量  $(X, Y)$  的分布列为  $p_{ij} = P(X = x_i, Y = y_j)$  ( $i, j = 1, 2, \dots$ ), 对于给定的边缘概率  $p_{\cdot j} = P(Y = y_j) = \sum_{i=1}^{+\infty} p_{ij} > 0$ ,

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

称为在  $Y = y_j$  条件下随机变量  $X$  的 **条件分布列** (conditional probability distribution). 可以类似定义在  $X = x_i$  条件下随机变量  $Y$  的条件分布列.

条件分布本质上也是一种概率分布, 具有分布的性质. 例如,

- **非负性:** 对任意整数  $i \geq 1$  有  $P(X = x_i | Y = y_j) \geq 0$ ;

- **规范性:**

$$\sum_{i=1}^{\infty} P(X = x_i | Y = y_j) = \sum_{i=1}^{+\infty} \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \sum_{i=1}^{+\infty} \frac{p_{ij}}{p_{\cdot j}} = 1.$$

- 若离散随机变量  $X$  和  $Y$  相互独立, 则有

$$P(X = x_i | Y = y_j) = P(X = x_i) \quad \text{和} \quad P(Y = y_j | X = x_i) = P(Y = y_j).$$

在后面的章节中, 若出现条件概率  $P(X = x_i | Y = y_j)$ , 一般默认概率  $P(Y = y_j) > 0$ . 条件分布列也可以通过下面的表格给出:

$X$	$x_1$	$x_2$	$\cdots$	$x_n$	$\cdots$
$P(X = x_i   Y = y_j)$	$p_{1j}/p_{\cdot j}$	$p_{2j}/p_{\cdot j}$	$\cdots$	$p_{nj}/p_{\cdot j}$	$\cdots$

**例 5.8** 一个选手击中目标的概率为  $p$ , 射中两次目标为止, 用  $X$  表示首次击中目标所进行的射击次数, 用  $Y$  表示第二次射中目标所进行的射击次数, 求  $X$  和  $Y$  的联合分布和条件分布.

**解** 随机变量  $X = m$  表示首次击中目标射击了  $m$  次,  $Y = n$  表示第二次次击中目标射击了  $n$  次, 则  $X$  和  $Y$  的联合分布列为

$$P\{X = m, Y = n\} = f(x, y) = \begin{cases} p^2(1-p)^{n-2} & 1 \leq m < n < \infty \\ 0 & \text{其它} \end{cases}$$

于是得到随机变量  $X$  的边缘分布列

$$P\{X = m\} = \sum_{n=m+1}^{\infty} P\{X = m, Y = n\} = \sum_{n=m+1}^{\infty} p^2(1-p)^{n-2} = p(1-p)^{m-1},$$

以及随机变量  $Y$  的边缘分布列

$$P\{Y = n\} = \sum_{m=1}^{n-1} P\{X = m, Y = n\} = \sum_{m=1}^{n-1} p^2(1-p)^{n-2} = (n-1)p^2(1-p)^{n-2} \quad (n \geq 2).$$

当  $n \geq 2$  时, 随机变量  $X$  在  $Y = n$  条件下的分布列

$$P\{X = m|Y = n\} = \frac{P\{X = m, Y = n\}}{P\{Y = n\}} = \frac{p^2(1-p)^{n-2}}{(n-1)p^2(1-p)^{n-2}} = \frac{1}{n-1} \quad (1 \leq m \leq n-1).$$

当  $m \geq 1$  时, 随机变量  $Y$  在  $X = m$  条件下的分布列

$$P\{Y = n|X = m\} = \frac{P\{X = m, Y = n\}}{P\{X = m\}} = \frac{p^2(1-p)^{n-2}}{p(1-p)^{m-1}} = p(1-p)^{n-m-1} \quad (n > m).$$

### 5.5.2 连续型随机向量的条件分布

连续型随机向量  $(X, Y)$  对任意实数  $x, y$  都有  $P(X = x) = 0$  和  $P(Y = y) = 0$ , 因此不能用条件概率的公式直接推导连续型随机向量的条件分布, 但可以通过下面的极限方式来考虑.

当  $P(y \leq Y \leq y + \epsilon) > 0$  时, 利用积分中值定理来求解条件分布函数

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{\epsilon \rightarrow 0^+} P\{X \leq x | y \leq Y \leq y + \epsilon\} = \lim_{\epsilon \rightarrow 0^+} \frac{P\{X \leq x, y \leq Y \leq y + \epsilon\}}{P\{y \leq Y \leq y + \epsilon\}} \\ &= \lim_{\epsilon \rightarrow 0^+} \frac{\int_{-\infty}^x \int_y^{y+\epsilon} f(u, v) du dv}{\int_y^{y+\epsilon} f_Y(u) dv} = \lim_{\epsilon \rightarrow 0^+} \frac{\epsilon \int_{-\infty}^x f(u, y + \theta_1 \epsilon) du}{\epsilon f_Y(y + \theta_2 \epsilon)} \quad \theta_1, \theta_2 \in (0, 1) \\ &= \frac{\int_{-\infty}^x f(u, y) du}{f_Y(y)} = \int_{-\infty}^x \frac{f(u, y)}{f_Y(y)} du, \end{aligned}$$

进一步得到条件密度函数  $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$ . 在后面的章节中, 若出现条件密度函数  $f_{X|Y}(x|y)$  (或  $f_{Y|X}(y|x)$ ), 一般都默认  $f_Y(y) > 0$  (或  $f_X(x) > 0$ ). 下面给详细的定义:

**定义 5.12** 设连续型随机变量  $(X, Y)$  的联合概率密度为  $f(x, y)$ , 以及  $X$  和  $Y$  的边缘概率密度分别为  $f_X(x)$  和  $f_Y(y)$ . 对任意给定的  $f_Y(y) > 0$ , 称  $f(x, y)/f_Y(y)$  为在  $Y = y$  条件下随机变量  $X$  的 **条件密度函数** (conditional probability density function), 记为

$$f_{X|Y}(x|y) = f(x, y)/f_Y(y),$$

以及在  $Y = y$  条件下  $X$  的 **条件分布函数** (conditional cumulative distribution function) 为

$$F_{X|Y}(x|y) = P\{X \leq x | Y = y\} = \int_{-\infty}^x f_{X|Y}(u|y) du .$$

对任意给定的  $f_X(x) > 0$ , 可类似定义在  $X = x$  条件下  $Y$  的条件密度函数和条件分布函数分别为

$$f_{Y|X}(y|x) = f(x, y)/f_X(x) \quad \text{和} \quad F_{Y|X}(y|x) = \int_{-\infty}^y f_{Y|X}(v|x) dv .$$

条件密度函数本质上是密度函数, 具有以下性质:

- **非负性:** 对任意实数  $x, y$  有  $f_{Y|X}(y|x) \geq 0$ .
- **规范性:** 对任意实数  $y$ :  $f_Y(y) > 0$  有

$$\int_{-\infty}^{+\infty} f_{X|Y}(x|y) dx = \int_{-\infty}^{+\infty} \frac{f(x, y)}{f_Y(y)} dx = \frac{1}{f_Y(y)} \int_{-\infty}^{+\infty} f(x, y) dx = 1 .$$

- **乘法公式:**  $f(x, y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y)$ .
- 若随机变量  $X$  和  $Y$  相互独立, 则有

$$f_{Y|X}(y|x) = f_Y(y) \quad \text{和} \quad f_{X|Y}(x|y) = f_X(x) .$$

根据条件概率的乘法公式有

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{+\infty} f(x, y) dx} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{+\infty} f(y|x)f_X(x) dx} ,$$

可以将其看作 **密度函数的贝叶斯公式**. 目前可能有三种途径来构造二维随机向量的联合分布函数:

- 根据实际问题或实际数据归纳为  $f(x, y)$ ;
- 根据随机变量的独立性有  $f(x, y) = f_X(x)f_Y(y)$ ;
- 根据乘法公式  $f(x, y) = f_X(x)f_{Y|X}(y|x)$ .

关于二维正太分布有

**定理 5.5** 若随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则在  $Y = y$  的条件下随机变量  $X$  服从正太分布  $\mathcal{N}(\mu_x - \rho\sigma_x(y - \mu_y)/\sigma_y, (1 - \rho^2)\sigma_x^2)$ , 以及在  $X = x$  的条件下随机变量  $Y$  服从正太分布  $\mathcal{N}(\mu_y - \rho\sigma_y(x - \mu_x)/\sigma_x, (1 - \rho^2)\sigma_y^2)$ .

**证明** 若随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则随机变量  $X$  的边缘分布为  $\mathcal{N}(\mu_x, \sigma_x^2)$ , 即

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2}\right) ,$$

此外可以将二维随机向量  $(X, Y)$  的联合密度函数  $f(x, y)$  分解为

$$f(x, y) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{(x - \mu_x)^2}{2\sigma_x^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{(y - \mu_y - \rho\sigma_y(x - \mu_x)/\sigma_x)^2}{2\sigma_y^2(1-\rho^2)}\right).$$

根据乘法公式  $f(x, y) = f_X(x)f_{Y|X}(y|x)$  可得

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{(y - \mu_y - \rho\sigma_y(x - \mu_x)/\sigma_x)^2}{2\sigma_y^2(1-\rho^2)}\right),$$

即正太分布  $\mathcal{N}(\mu_y + \rho\sigma_y(x - \mu_x)/\sigma_x, (1 - \rho^2)\sigma_y^2)$ . 同理可证在  $Y = y$  的条件下随机变量  $X$  的条件分布为  $\mathcal{N}(\mu_x + \rho\sigma_x(y - \mu_y)/\sigma_y, (1 - \rho^2)\sigma_x^2)$ .

**例 5.9** 设二维随机向量  $(X, Y)$  的密度函数

$$f(x, y) = \begin{cases} e^{-x/y}e^{-y}/y & x > 0, y > 0 \\ 0 & \text{其它,} \end{cases}$$

求  $P(X > 1|Y = y)$ .

**解** 积分区域如图 5.5(a) 所示, 求解随机变量  $Y$  的边缘分布为

$$f_Y(y) = \int_0^{+\infty} e^{-x/y}e^{-y}/y dx = e^{-y} \left[-e^{-x/y}\right]_0^{+\infty} = e^{-y} \quad (y > 0).$$

进而得到在  $Y = y$  条件下  $X$  的条件概率密度为

$$f_{X|Y}(x|y) = e^{-x/y}/y.$$

最后求解得到

$$P(X > 1|Y = y) = \int_1^{\infty} e^{-x/y}/y dx = -\left[e^{-x/y}\right]_1^{\infty} = e^{-1/y}.$$

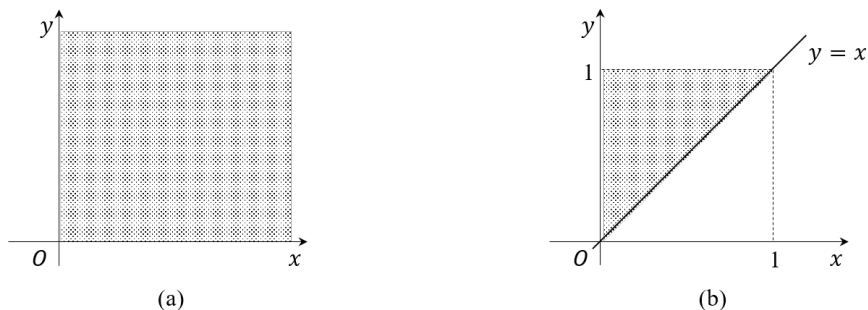


图 5.5 例 5.9 和 5.10 的积分区域

**例 5.10** 设随机变量  $X \sim U(0, 1)$ , 在观察到  $X = x$  的条件下随机变量  $Y \sim U(x, 1)$ , 求随机变量  $Y$  的概率密度.

**解** 随机变量  $X \sim U(0, 1)$ , 在随机变量  $X = x$  的条件下  $Y \sim U(x, 1)$ , 于是当  $x > 0$  时有

$$f_{Y|X}(y|x) = 1/(1-x) .$$

根据条件概率乘积公式有

$$f(x, y) = f_X(x)f_{Y|X}(y|x) = \begin{cases} 1/(1-x) & 0 < x < y < 1, \\ 0 & \text{其它.} \end{cases}$$

积分区域如图 5.5(b) 所示, 当  $y > 0$  时随机变量  $Y$  的边缘分布

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)dx = \int_0^y \frac{1}{1-x}dx = -\ln(1-y) .$$

## 5.6 多维随机变量函数的分布

已知二维随机向量  $(X, Y)$  的概率分布, 如何求解随机变量函数  $Z = g(X, Y)$  的概率分布. 下面分离散型和连续型随机变量两种情况进行讨论.

### 5.6.1 二维离散型随机向量函数

已知二维离散型随机向量  $(X, Y)$  的联合分布列, 求函数  $Z = g(X, Y)$  的分布列相对简单. 首先针对  $X, Y$  的各种取值, 计算随机变量  $Z$  的值, 然后对相同的  $Z$  值合并, 对应的概率相加. 下面研究两个相互独立的离散型随机变量之和, 即离散型随机变量的卷积公式:

**定理 5.6** 设离散型随机变量  $X$  与  $Y$  相互独立、且它们的分布列分别为  $a_i = P(X = i)$  和  $b_j = P(Y = j)$  ( $i, j = 0, 1, \dots$ ), 则随机变量  $Z = X + Y$  的分布列为

$$P(Z = k) = \sum_{i=0}^k a_i b_{k-i} .$$

**证明** 对任意非负整数  $i$  和  $j$ , 根据独立性可知

$$P(X = i, Y = j) = P(X = i)P(Y = j) = a_i b_j .$$

因此随机变量  $Z$  的分布列为

$$\begin{aligned} P(Z = k) &= P(X + Y = k) \\ &= \sum_{i=0}^k P(X = i, Y = k - i) = \sum_{i=0}^k P(X = i)P(Y = k - i) = \sum_{i=0}^k a_i b_{k-i} , \end{aligned}$$

定理得证.

基于定理 5.6, 可以得到一系列推论:

**推论 5.1** 若随机变量  $X \sim B(n_1, p)$  和  $Y \sim B(n_2, p)$  相互独立, 则随机变量

$$Z = X + Y \sim B(n_1 + n_2, p).$$

**证明** 根据二项分布的定义, 当  $i = 0, 1, \dots, n_1$  和  $j = 0, 1, \dots, n_2$  有

$$P(X = i) = \binom{n_1}{i} p^i (1-p)^{n_1-i} \quad \text{和} \quad P(Y = j) = \binom{n_2}{j} p^j (1-p)^{n_2-j}.$$

对  $k = 0, 1, \dots, n_1 + n_2$ , 根据定理 5.6 有

$$\begin{aligned} P[Z = k] &= \sum_{i=0}^k P[X = i] P[Y = k - i] = \sum_{i=0}^k \binom{n_1}{i} p^i (1-p)^{n_1-i} \binom{n_2}{k-i} p^{k-i} (1-p)^{n_2-(k-i)} \\ &= p^k (1-p)^{n_1+n_2-k} \sum_{i=0}^k \binom{n_1}{i} \binom{n_2}{k-i} = \binom{n_1+n_2}{k} p^k (1-p)^{n_1+n_2-k}. \end{aligned}$$

利用归纳法和推论 5.1, 若相互独立的随机变量  $X_i \sim \text{Ber}(p) = B(1, p)$  ( $i \in [n]$ ), 则随机变量

$$X = X_1 + X_2 + \dots + X_n \sim B(n, p).$$

即随机变量  $X \sim B(n, p)$  可以看作  $n$  个相互独立的服从参数为  $p$  的伯努利分布随机变量之和.

**推论 5.2** 若随机变量  $X \sim P(\lambda_1)$  和  $Y \sim P(\lambda_2)$  相互独立, 则随机变量

$$Z = X + Y \sim P(\lambda_1 + \lambda_2).$$

**证明** 根据泊松分布的定义, 对任意非负整数  $i$  和  $j$  有

$$P(X = i) = \lambda_1^i e^{-\lambda_1} / i! \quad \text{和} \quad P(Y = j) = \lambda_2^j e^{-\lambda_2} / j!.$$

对任意非负整数  $k$ , 根据定理 5.6 有

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

## 5.6.2 二维连续型随机向量函数

已知二维连续型随机向量  $(X, Y)$  的联合概率密度为  $f(x, y)$ , 求随机变量  $Z = g(X, Y)$  的概率密度, 一般先求解分布函数

$$F_Z(z) = P(Z \leq z) = P(g(x, y) \leq z) = \iint_{g(x, y) \leq z} f(x, y) dx dy ,$$

再对分布函数  $F_Z(z)$  求导得到密度函数

$$f_Z(z) = F'_Z(z) .$$

**例 5.11** 设服从标准正态分布的两个随机变量  $X$  和  $Y$  相互独立, 求随机变量  $Z_1 = \sqrt{X^2 + Y^2}$  和  $Z_2 = X^2 + Y^2$  的密度函数.

**解** 根据独立性有随机变量  $X$  和  $Y$  的联合密度函数

$$f(x, y) = f_X(x)f_Y(y) = e^{-(x^2+y^2)/2}/2\pi \quad (x, y \in \mathbb{R}) .$$

当  $z_1 \leq 0$  时, 根据  $Z_1 = \sqrt{X^2 + Y^2}$  很显然有分布函数  $F_{Z_1}(z_1) = 0$ ; 当  $z_1 > 0$  时有

$$F_{Z_1}(z_1) = P(Z_1 \leq z_1) = P\left(\sqrt{X^2 + Y^2} \leq z_1\right) = \iint_{X^2+Y^2 \leq z_1^2} e^{-(x^2+y^2)/2}/2\pi dx dy ,$$

利用极坐标积分变换  $x = r \cos \theta$  和  $y = r \sin \theta$  有

$$F_{Z_1}(z_1) = \int_0^{2\pi} \int_0^{z_1} \frac{r}{2\pi} e^{-r^2/2} d\theta dr = \int_0^{z_1} r e^{-r^2/2} dr = 1 - e^{-z_1^2/2} .$$

由此得到随机变量  $Z_1$  的密度函数为

$$f_{Z_1}(z_1) = \begin{cases} z_1 e^{-z_1^2/2} & z_1 > 0 \\ 0 & z_1 \leq 0 . \end{cases}$$

上述分布称为 **瑞利分布** (Rayleigh distribution), 该分布常用于通信等领域. 同理可证随机变量  $Z_2 \sim e(1/2)$ , 即

$$f_{Z_2}(z_2) = \begin{cases} z_2 e^{-z_2/2}/2 & z_2 > 0 \\ 0 & z_2 \leq 0 . \end{cases}$$

5.6.2.1 和的分布  $Z = X + Y$ 

**引理 5.2** 设二维随机向量  $(X, Y)$  的联合密度函数为  $f(x, y)$ , 则有  $Z = X + Y$  的密度函数

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z-x)dx \quad \text{或} \quad f_Z(z) = \int_{-\infty}^{+\infty} f(z-y, y)dy .$$

**解** 首先求解  $Z = X + Y$  的分布函数

$$F_Z(z) = P(Z \leq z) = P(X + Y \leq z) = \iint_{x+y \leq z} f(x, y)dx dy = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{z-x} f(x, y)dy ,$$

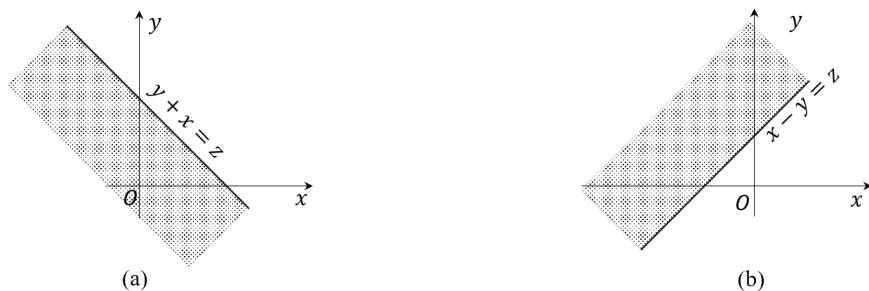
这里考虑的积分区域为  $\{(x, y): x + y \leq z\}$ , 如图 5.6(a) 所示. 利用变量替换  $u = y + x$  并积分换序有

$$F_Z(z) = \int_{-\infty}^z \left( \int_{-\infty}^{+\infty} f(x, u-x)dx \right) du ,$$

两边同时对  $z$  求导数可得

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z-x)dx .$$

同理可证  $f_Z(z) = \int_{-\infty}^{+\infty} f(z-y, y)dy$ .



**图 5.6** 函数  $Z = X + Y$  和  $Z = X - Y$  的积分区域

类似考虑随机变量  $Z = X - Y$ , 其积分区域  $\{(x, y): x - y \leq z\}$  如图 5.6(b) 所示, 得到  $Z = X - Y$  的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, x-z)dx \quad \text{或} \quad f_Z(z) = \int_{-\infty}^{+\infty} f(z+y, y)dy .$$

若随机变量  $X$  和  $Y$  相互独立, 则有  $f(x, y) = f_X(x)f_Y(y)$ . 结合引理 5.2 给出下面著名的定理:

**定理 5.7 (卷积公式)** 若连续型随机变量  $X$  与  $Y$  相互独立, 且它们的密度函数分别为  $f_X(x)$  和  $f_Y(y)$ , 则随机变量  $Z = X + Y$  的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dy .$$



**推论 5.3** 若随机变量  $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$  和  $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$  相互独立, 则随机变量

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2).$$

根据上面的推论很容易得到  $X - Y \sim \mathcal{N}(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$ . 该结论可以推广到  $n$  个随机变量, 设随机变量  $X_1, X_2, \dots, X_n$  相互独立、且  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , 则随机变量

$$Z = X_1 + X_2 + \dots + X_n \sim \mathcal{N}(\mu_1 + \mu_2 + \dots + \mu_n, \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2).$$

**证明** 若随机变量  $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$  和  $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ , 则根据正太分布的性质有

$$X' = X - \mu_x \sim \mathcal{N}(0, \sigma_x^2) \quad \text{和} \quad Y' = Y - \mu_y \sim \mathcal{N}(0, \sigma_y^2).$$

因此只需证明  $Z = X' + Y' \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$ . 根据卷积公式有

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{(z-x)^2}{2\sigma_2^2}\right) dx \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \exp\left(-\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1^2\sigma_2^2} \left(x - \frac{\sigma_1^2 z}{\sigma_1^2 + \sigma_2^2}\right)^2 - \frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right) dx \\ &= \frac{\exp\left(-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right)}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} \times \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{2\pi}\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \exp\left(-\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1^2\sigma_2^2} \left(x - \frac{\sigma_1^2 z}{\sigma_1^2 + \sigma_2^2}\right)^2\right) dx \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right), \end{aligned}$$

最后一个等式成立是因为正太分布的规范性.

**例 5.12** 设随机变量  $X \sim U(0, 1)$  和  $Y \sim U(0, 1)$  相互独立, 求  $Z = X + Y$  的概率密度.

**解** 根据卷积公式有

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx.$$

根据区间  $(0, 1)$  上的均匀分布, 当  $x \in [0, 1]$  时有  $f_X(x) = 1$ ; 当  $z-x \in [0, 1]$  时有  $f_Y(z-x) = 1$ . 由此可得非零的积分区域为  $\{x \in [0, 1], z-x \in [0, 1]\}$ , 如图 5.7(a) 所示. 当  $z \leq 0$  或  $z \geq 2$  时有  $f_Z(z) = 0$ ; 当  $z \in (0, 1)$  时有

$$f_Z(z) = \int_0^z 1 dz = z;$$

当  $z \in [1, 2)$  时有

$$f_Z(z) = \int_{z-1}^1 dx = 2 - z.$$

综上所述, 随机变量  $Z = X + Y$  的概率密度

$$f_Z(z) = \begin{cases} z & z \in [0, 1] \\ 2 - z & z \in [1, 2] \\ 0 & \text{其它} . \end{cases}$$

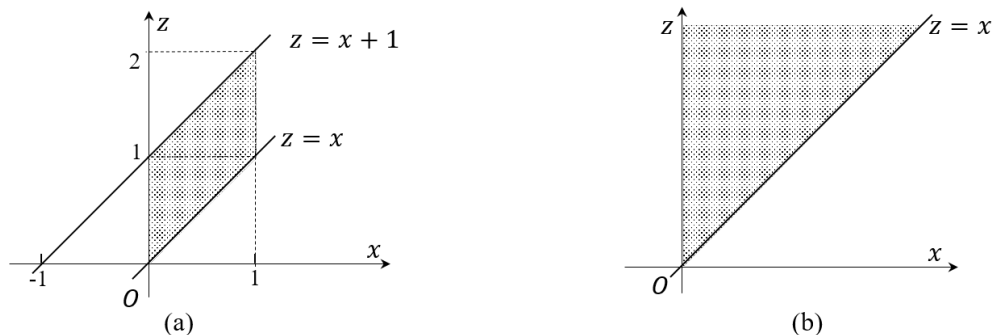


图 5.7 例 5.12 和 5.13 中积分区域示意图

**例 5.13** 设随机变量  $X \sim e(\lambda)$  和  $Y \sim e(\lambda)$  相互独立, 求  $Z = X + Y$  的概率密度.

**解** 由卷积公式可得

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx .$$

根据指数分布的定义, 当  $x \geq 0$  时有  $f_X(x) = \lambda \exp(-\lambda x)$ ; 当  $z-x \geq 0$  时有  $f_Y(z-x) = \lambda \exp(-\lambda(z-x))$ , 因此积分区域  $\{x \in [0, +\infty), z-x \in [0, +\infty)\}$  如图 5.7(b) 所示. 当  $z \geq 0$  时有

$$f_Z(z) = \lambda^2 \int_0^z \exp(-\lambda x) \exp(-\lambda(z-x)) dx = \lambda^2 z \exp(-\lambda z) .$$

### 5.6.3 随机变量的乘/除法分布

**定理 5.8** 设二维随机变量  $(X, Y)$  的概率密度为  $f(x, y)$ , 则随机变量  $Z = XY$  的概率密度为

$$f_{XY}(z) = \int_{-\infty}^{+\infty} \frac{1}{|x|} f(x, \frac{z}{x}) dx ;$$

随机变量  $Z = Y/X$  的概率密度为

$$f_{Y/X}(z) = \int_{-\infty}^{+\infty} |x| f(x, xz) dx .$$

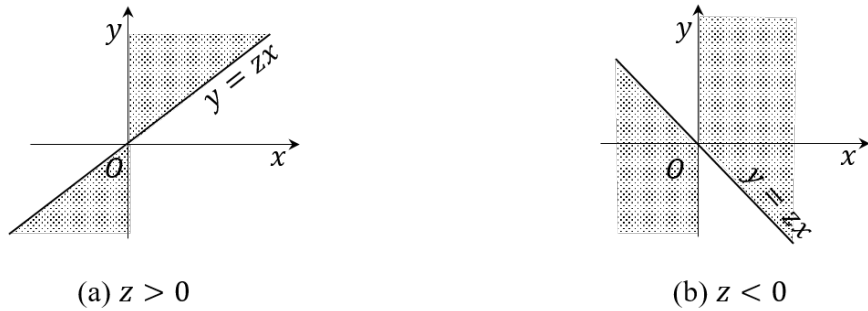
**证明** 这里给出随机变量  $Z = Y/X$  的概率密度详细证明, 同理给出  $Z = XY$  的概率密度. 首先考虑分布函数

$$\begin{aligned}
 F_{Y/X}(z) &= P(Y/X \leq z) = \iint_{y/x \leq z} f(x, y) dx dy \\
 &= \iint_{x < 0, y \geq zx} f(x, y) dx dy + \iint_{x > 0, y \leq zx} f(x, y) dx dy \\
 &= \int_{-\infty}^0 dx \int_{zx}^{+\infty} f(x, y) dy + \int_0^{+\infty} dx \int_{-\infty}^{xz} f(x, y) dy,
 \end{aligned}$$

如图 5.8 所示, 这里考虑积分区域为  $\{(x, y): x > 0, y < xz\} \cup \{(x, y): x < 0, y > xz\}$ . 变量替换  $t = y/x$  有

$$\begin{aligned}
 F_{Y/X}(z) &= \int_{-\infty}^0 dx \int_z^{-\infty} x f(x, tx) dt + \int_0^{+\infty} dx \int_{-\infty}^z x f(x, tx) dt \\
 &= \int_{-\infty}^0 \int_{-\infty}^z (-x) f(x, tx) dt dx + \int_0^{+\infty} \int_{-\infty}^z x f(x, tx) dt dx \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^z |x| f(x, tx) dt dx = \int_{-\infty}^z dt \int_{-\infty}^{+\infty} |x| f(x, tx) dx,
 \end{aligned}$$

对分布函数求导即可完成证明.



**图 5.8** 随机变量  $Z = Y/X$  的积分区域

**推论 5.4** 若标准正太分布的随机变量  $X$  和  $Y$  相互独立, 则随机变量  $Z = Y/X$  服从柯西分布.

**证明** 根据独立性和定理 5.8, 对任意实数  $z$  有

$$\begin{aligned}
 f_Z(z) &= \int_{-\infty}^{+\infty} |x| f(x, xz) dx = \int_{-\infty}^{+\infty} |x| f_X(x) f_Y(xz) dx \\
 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} |x| e^{-x^2(1+z^2)/2} dx = \frac{1}{\pi} \int_0^{+\infty} x e^{-x^2(1+z^2)/2} dx
 \end{aligned}$$

$$= \frac{1}{\pi} \left[ -\frac{e^{-x^2(1+z^2)/2}}{1+z^2} \right]_0^{+\infty} = \frac{1}{\pi(1+z^2)},$$

推论得证.

#### 5.6.4 最大值和最小值的分布

**定理 5.9** 设随机变量  $X_1, \dots, X_n$  相互独立、且其分布函数分别为  $F_{X_1}(x_1), \dots, F_{X_n}(x_n)$ , 则随机变量  $Y = \max(X_1, X_2, \dots, X_n)$  的分布函数为

$$F_Y(y) = F_{X_1}(y)F_{X_2}(y) \cdots F_{X_n}(y),$$

以及随机变量  $Z = \min(X_1, X_2, \dots, X_n)$  的分布函数为

$$F_Z(z) = 1 - (1 - F_{X_1}(z))(1 - F_{X_2}(z)) \cdots (1 - F_{X_n}(z)).$$

**证明** 根据独立性, 随机变量  $Y = \max(X_1, X_2, \dots, X_n)$  的分布函数为

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(\max(X_1, X_2, \dots, X_n) \leq y) = P(X_1 \leq y, X_2 \leq y, \dots, X_n \leq y) \\ &= P(X_1 \leq y)P(X_2 \leq y) \cdots P(X_n \leq y) = F_{X_1}(y)F_{X_2}(y) \cdots F_{X_n}(y). \end{aligned}$$

随机变量  $Z = \min(X_1, X_2, \dots, X_n)$  的分布函数为

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(\min(X_1, X_2, \dots, X_n) \leq z) = 1 - P(\min(X_1, X_2, \dots, X_n) > z) \\ &= 1 - P(X_1 > z, X_2 > z, \dots, X_n > z) = 1 - P(X_1 > z)P(X_2 > z) \cdots P(X_n > z) \\ &= 1 - (1 - F_{X_1}(z))(1 - F_{X_2}(z)) \cdots (1 - F_{X_n}(z)), \end{aligned}$$

定理得证.

根据定理 5.9 有

**推论 5.5** 设  $X_1, X_2, \dots, X_n$  是独立同分布的随机变量, 其分布函数和密度函数分别为  $F(x)$  和  $f(x)$ , 则随机变量  $Y = \max(X_1, X_2, \dots, X_n)$  的分布函数和密度函数分别为

$$F_Y(y) = (F(y))^n \quad \text{和} \quad f_Y(y) = n(F(y))^{n-1}f(y),$$

以及随机变量  $Z = \min(X_1, X_2, \dots, X_n)$  的分布函数和密度函数分别为

$$F_Z(z) = 1 - (1 - F(z))^n \quad \text{和} \quad f_Z(z) = n(1 - F(z))^{n-1}f(z).$$

**例 5.14** 假设随机变量  $X$  与  $Y$  相互独立, 且有  $X \sim e(\alpha)$  和  $Y \sim e(\beta)$ , 求随机变量  $Z_1 = \max(X, Y)$  和  $Z_2 = \min(X, Y)$  的概率密度.

**解** 根据指数随机变量的定义可知随机变量  $X$  和  $Y$  的概率密度为

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x} & x > 0 \\ 0 & x \leq 0 \end{cases} \quad \text{和} \quad f_Y(y) = \begin{cases} \beta e^{-\beta y} & y > 0 \\ 0 & y \leq 0 \end{cases}.$$

于是得到随机变量  $Z_1$  的分布函数为

$$F_{Z_1}(z_1) = F_X(z_1)F_Y(z_1) = \int_{-\infty}^{z_1} f_X(t)dt \int_{-\infty}^{z_1} f_Y(t)dt.$$

当  $z_1 \leq 0$  时由  $F_{Z_1}(z_1) = 0$ ; 当  $z_1 > 0$  时

$$F_{Z_1}(z_1) = \int_0^{z_1} f_X(t)dt \int_0^{z_1} f_Y(t)dt = \int_0^{z_1} \alpha e^{-\alpha t} dt \int_0^{z_1} \beta e^{-\beta t} dt = (1 - e^{-\alpha z_1})(1 - e^{-\beta z_1}).$$

两边对  $z_1$  求导可得其概率密度为

$$f_{Z_1}(z_1) = \begin{cases} \alpha e^{-\alpha z_1} + \beta e^{-\beta z_1} - (\alpha + \beta)e^{-(\alpha+\beta)z_1} & z_1 > 0 \\ 0 & z_1 \leq 0 \end{cases}.$$

同理可得随机变量  $Z_2$  的分布函数和概率密度分别为

$$F_{Z_2}(z_2) = \begin{cases} 1 - e^{-(\alpha+\beta)z_2} & z_2 > 0 \\ 0 & z_2 \leq 0 \end{cases} \quad f_{Z_2}(z_2) = \begin{cases} (\alpha + \beta)e^{-(\alpha+\beta)z_2} & z_2 > 0 \\ 0 & z_2 \leq 0 \end{cases}.$$

### 5.6.5 随机变量的联合分布函数

已知随机向量  $(X, Y)$  的联合概率密度为  $f(x, y)$ , 随机变量  $U$  和  $V$  是  $X$  和  $Y$  的函数, 如何求解  $(U, V)$  的联合分布. 具体而言, 设

$$\begin{cases} U = u(X, Y) \\ V = v(X, Y), \end{cases}$$

已知随机向量  $(X, Y)$  的联合概率密度为  $f(x, y)$ , 如何求解二维随机向量  $(U, V)$  的联合分布. 这里二元函数  $u(\cdot, \cdot)$  和  $v(\cdot, \cdot)$  具有连续的偏导, 并满足

$$\begin{cases} u = u(x, y) \\ v = v(x, y) \end{cases} \quad \text{存在唯一的反函数} \quad \begin{cases} x = x(u, v) \\ y = y(u, v). \end{cases}$$

**定理 5.10** 设随机变量  $U = u(X, Y)$  和  $V = v(X, Y)$  有连续偏导, 且存在反函数  $X = x(U, V)$  和  $Y = y(U, V)$ . 若  $(X, Y)$  的联合概率密度为  $f(x, y)$ , 则  $(U, V)$  的联合密度为

$$f_{UV}(u, v) = f_{XY}(x(u, v), y(u, v))|J|,$$

其中  $J$  为变换的雅可比行列式不为零, 即

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

上述结论可推广到一般的  $n$  维随机变量.

**推论 5.6** 设  $X$  和  $Y$  是相互独立的标准正太分布随机变量, 则有随机变量  $R = X^2 + Y^2$  与  $\theta = \arctan(Y/X)$  相互独立, 且有  $R \sim e(1/2)$  以及  $\theta \sim U(0, 2\pi)$ .

**证明** 令  $R = u(x, y) = x^2 + y^2$  和  $\Theta = v(x, y) = \arctan(y/x)$ . 于是得到雅可比行列式

$$J = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix}^{-1} = \begin{vmatrix} 2x & 2y \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{vmatrix}^{-1} = \frac{1}{2}.$$

由此可得  $R$  与  $\Theta$  的联合分布为

$$f_{R \times \Theta}(r, \theta) = f_{X \times Y}(\sqrt{r} \cos \theta, \sqrt{r} \sin \theta)|J| = \frac{1}{4\pi} \exp(-r/2) = \frac{1}{2} \exp(-r/2) \times \frac{1}{2\pi}.$$

由此可以发现  $R \sim e(1/2)$  和  $\Theta \sim U(0, 2\pi)$ , 且  $R$  和  $\Theta$  相互独立. 推论得证.

## 5.7 多维正太分布

本节将二维随机向量及其分布推广到多维随机向量, 二维与多维随机变量没有本质性的区别, 只是相关的概念和结论的扩展.

**定义 5.13** 设  $(X_1, X_2, \dots, X_n)$  为  $n$  维随机向量, 对任意实数  $x_1, x_2, \dots, x_n$ , 称函数

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

为  $n$  维随机向量  $(X_1, X_2, \dots, X_n)$  的分布函数, 或随机变量  $X_1, X_2, \dots, X_n$  的联合分布函数. 若存在可积函数  $f(x_1, x_2, \dots, x_n)$ , 使得对任意实数  $x_1, x_2, \dots, x_n$  有

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(u_1, u_2, \dots, u_n) du_1 du_2 \cdots du_n,$$

则称  $(X_1, X_2, \dots, X_n)$  为连续型随机向量, 以及  $f(x_1, x_2, \dots, x_n)$  为  $n$  维联合密度函数.

类似于二维密度函数,  $n$  维联合密度函数具有以下性质:

- 非负性: 对任意实数  $x_1, x_2, \dots, x_n$  有  $f(x_1, x_2, \dots, x_n) \geq 0$ .
- 规范性:  $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n = 1$ .
- 设  $G$  是  $n$  维空间的一片区域, 则有

$$P((X_1, X_2, \dots, X_n) \in G) = \int \dots \int_G f(u_1, u_2, \dots, u_n) du_1 du_2 \dots du_n .$$

- 若  $f(x_1, x_2, \dots, x_n)$  在点  $(x_1, x_2, \dots, x_n)$  处连续, 则有

$$\frac{\partial F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n} = f(x_1, x_2, \dots, x_n) .$$

随机向量  $(X_1, X_2, \dots, X_n)$  中任意  $k$  个向量所构成的随机向量 ( $k \leq n$ ), 它的分布函数和密度函数被称为  $k$  维边缘分布函数和  $k$  维边缘密度函数. 例如随机向量  $(X_1, X_2, \dots, X_n)$  前  $k$  维随机向量的边缘分布函数和边缘密度函数分布为

$$\begin{aligned} F_{X_1, X_2, \dots, X_k}(x_1, x_2, \dots, x_k) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k) = \lim_{\substack{x_{k+1} \rightarrow +\infty \\ \dots \\ x_n \rightarrow +\infty}} F(x_1, x_2, \dots, x_n) \\ f_{X_1, X_2, \dots, X_k}(x_1, x_2, \dots, x_k) &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(u_1, \dots, u_k, u_{k+1}, \dots, u_n) du_{k+1} \dots du_n . \end{aligned}$$

还可以定义  $n$  个随机变量的独立性和两个随机向量的独立性.

**定义 5.14** 若随机向量  $(X_1, X_2, \dots, X_n)$  的联合分布函数  $F(x_1, x_2, \dots, x_n)$  满足

$$F(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \dots F_{X_n}(x_n) ,$$

则称  $X_1, X_2, \dots, X_n$  相互独立. 若随机向量  $X = (X_1, X_2, \dots, X_m)$  和  $Y = (Y_1, Y_2, \dots, Y_n)$  的联合分布函数  $F(x_1, \dots, x_m, y_1, \dots, y_n)$  满足

$$F(x_1, \dots, x_m, y_1, \dots, y_n) = F_X(x_1, \dots, x_m) F_Y(y_1, \dots, y_n) ,$$

则称 随机向量  $X$  和  $Y$  相互独立.

上面的独立性也可以通过联合密度函数和边缘密度函数来定义. 多维随机向量中最重要的常用分布是多维正太分布.

**定义 5.15** 给定一个向量  $\boldsymbol{\mu} \in \mathbb{R}^n$  和正定矩阵  $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ , 对任意实数向量  $\boldsymbol{x} = (x_1, \dots, x_n)^T$ , 若随机向量  $X = (X_1, X_2, \dots, X_n)$  的密度函数为

$$f(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/2\right),$$

则称随机向量  $X$  服从参数为  $\boldsymbol{\mu}$  和  $\boldsymbol{\Sigma}$  的多维正态分布 (multivariate normal distribution), 记

$$X = (X_1, X_2, \dots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

在上面的定义中,  $|\boldsymbol{\Sigma}|$  表示矩阵  $\boldsymbol{\Sigma}$  的行列式, 因为其正定性可以确保  $|\boldsymbol{\Sigma}|^{-1/2}$  有意义. 特别地, 当  $n = 2$  时, 设

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{和} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix},$$

则定义 5.9 和定义 5.15 中关于二维正太分布的密度函数尽管表达形式不同, 但两者完全相等, 相关证明将作为一个练习题.

当  $\boldsymbol{\mu} = \mathbf{0}_n$  (全为零的  $n$  维向量), 以及  $\boldsymbol{\Sigma} = I_n$  ( $n \times n$  单位阵) 时, 正太分布  $\mathcal{N}(\mathbf{0}_n, I_n)$  被称为  $n$  维标准正太分布, 此时它的密度函数为

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{x_1^2 + x_2^2 + \dots + x_n^2}{2}\right).$$

不难发现,  $n$  维标准正太分布可以看作是相互独立的  $n$  个标准正太分布随机变量的联合分布, 也容易验证  $n$  标准正太分布的密度函数满足

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(\boldsymbol{x}) dx_1 \dots dx_n = \prod_{i=1}^n \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right) dx_i = 1.$$

对于正定矩阵  $\boldsymbol{\Sigma}$  通过特征值分解有

$$\boldsymbol{\Sigma} = \boldsymbol{U}^T \boldsymbol{\Lambda} \boldsymbol{U},$$

这里  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  是由特征值构成的对角阵,  $\boldsymbol{U}$  是特征向量所构成的正交矩阵. 基于特征值分解可以将任意  $n$  维正态分布转化为  $n$  维标准正态分布.

**定理 5.11** 设  $n$  维随机向量  $X = (X_1, X_2, \dots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , 以及正定矩阵  $\boldsymbol{\Sigma}$  的特征值分解为  $\boldsymbol{\Sigma} = \boldsymbol{U}^T \boldsymbol{\Lambda} \boldsymbol{U}$ , 则随机向量

$$Y = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{U} (X - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}_n, I_n).$$



**证明** 根据  $Y = \mathbf{\Lambda}^{-1/2}\mathbf{U}(X - \boldsymbol{\mu})$  可得  $X = \mathbf{U}^T \mathbf{\Lambda}^{1/2} Y + \boldsymbol{\mu}$ , 已知  $X$  的概率密度函数为

$$f_X(\mathbf{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})/2\right).$$

根据  $n$  维随机变量函数 (定理 5.10 的多维情况) 的概率密度公式有

$$f_Y(\mathbf{y}) = f_X\left(\mathbf{U}^T \mathbf{\Lambda}^{1/2} \mathbf{y} + \boldsymbol{\mu}\right) \left|\mathbf{U}^T \mathbf{\Lambda}^{1/2}\right|,$$

其中  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ . 根据特征值分解  $\boldsymbol{\Sigma} = \mathbf{U}^T \mathbf{\Lambda} \mathbf{U}$  有

$$\left|\mathbf{U}^T \mathbf{\Lambda}^{1/2}\right| = |\boldsymbol{\Sigma}|^{1/2},$$

以及将  $\mathbf{x} = \mathbf{U}^T \mathbf{\Lambda}^{1/2} \mathbf{y} + \boldsymbol{\mu}$  代入有

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

由此可得随机向量  $Y = \mathbf{\Lambda}^{-1/2}\mathbf{U}(X - \boldsymbol{\mu})$  的密度函数为

$$f_Y(\mathbf{y}) = (2\pi)^{-n/2} \exp\left(-\mathbf{y}^T \mathbf{y}/2\right),$$

定理得证.

多维正太分布有下面的性质, 其证明将作为一个练习题.

**定理 5.12** 设随机向量  $X = (X_1, X_2, \dots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , 则有

$$Y = \mathbf{A}X + \mathbf{b} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

其中  $|\mathbf{A}| \neq 0$ ,  $\mathbf{A} \in \mathbb{R}^{n \times n}$  和  $\mathbf{b} \in \mathbb{R}^{n \times 1}$ .

对于多维正太分布还有下面一些重要的性质:

**定理 5.13** 设随机向量  $X = (X_1, X_2, \dots, X_n)^T$  和  $Y = (Y_1, Y_2, \dots, Y_m)^T$ , 以及

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}\right),$$

其中  $\boldsymbol{\mu}_x = (\mu_{x_1}, \mu_{x_2}, \dots, \mu_{x_n})^T$ ,  $\boldsymbol{\mu}_y = (\mu_{y_1}, \mu_{y_2}, \dots, \mu_{y_m})^T$ ,  $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^T \in \mathbb{R}^{m \times n}$ ,  $\boldsymbol{\Sigma}_{xx} \in \mathbb{R}^{n \times n}$  和  $\boldsymbol{\Sigma}_{yy} \in \mathbb{R}^{m \times m}$ , 则有

- 随机向量  $X$  和  $Y$  的边缘分布分别为  $X \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$  和  $Y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy})$ ;
- 随机向量  $X$  与  $Y$  相互独立的充要条件是  $\boldsymbol{\Sigma}_{xy} = (\mathbf{0})_{m \times n}$  (元素全为零的  $m \times n$  矩阵);

- 在  $X = \mathbf{x}$  的条件下随机向量  $Y \sim \mathcal{N}(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy})$ ;
- 在  $Y = \mathbf{y}$  的条件下随机向量  $X \sim \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx})$ .

我们这里仅仅给出结论, 不给出具体的证明. 有兴趣的读者可以查询资料或自己动手, 证明的核心是矩阵的分块, 可以借鉴二维正太分布的证明.

## 习题

5.1 设二维随机变量  $(X, Y)$  的分布函数为  $F(x, y)$ , 求概率  $P(X > x, Y > y)$ .

5.2 设随机变量  $(X, Y)$  的分布函数

$$F(x, y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-x-y} & x \geq 0, y \geq 0 \\ 0 & \text{其它,} \end{cases}$$

求  $X$  和  $Y$  的边缘分布函数和边缘密度函数.

5.3 设连续非负的随机变量  $X$  和  $Y$  相互独立,  $X$  的边缘分布函数为  $F_X(x)$ ,  $Y$  的边缘密度函数为  $f_Y(y)$ , 证明

$$P(X < Y) = \int_0^{+\infty} F_X(x) f_Y(x) dx.$$

5.4 设随机变量  $(X, Y)$  的密度函数

$$f(x, y) = \begin{cases} ce^{-y} & 0 \leq x < y < +\infty \\ 0 & \text{其它,} \end{cases}$$

求  $X$  和  $Y$  的边缘密度函数.

5.5 设相互独立的随机变量  $X \sim e(\lambda_1)$  和  $Y \sim e(\lambda_2)$ , 其中  $\lambda_1 > 0, \lambda_2 > 0$ . 求  $P(X < Y)$ .

5.6 给定  $\alpha > 0$ , 设随机变量  $(X, Y)$  的分布函数

$$F(x, y) = \begin{cases} (1 - e^{-\alpha x})y & x \geq 0, 0 \leq y \leq 1 \\ 1 - e^{-\alpha x} & x \geq 0, y > 1 \\ 0 & \text{其它,} \end{cases}$$

求  $X$  和  $Y$  的独立性.

5.7 设二维随机变量  $(X, Y)$  的概率密度

$$f(x, y) = \begin{cases} x^2 + axy & 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0 & \text{其它,} \end{cases}$$

求  $P(X + Y \geq 1)$ .

5.8 设二维随机变量  $(X, Y)$  的概率密度

$$f(x, y) = \begin{cases} cxy & 0 \leq x \leq y \leq 1 \\ 0 & \text{其它,} \end{cases}$$

求  $P(X \leq 1/2)$ .

5.9 若多维随机向量  $(X_1, X_2, \dots, X_r) \sim M(n, p_1, p_2, \dots, p_r)$ , 则每个随机变量  $X_i$  ( $i \in [r]$ ) 的边缘分布为二项分布  $B(n, p_i)$ . (利用联合分布函数的定义证明)

5.10 设随机变量  $X, Y, Z$  服从  $(0, 1)$  上的均匀分布且相互独立, 求概率  $P(X \geq YZ)$ .

5.11 设随机向量  $(X, Y)$  的密度函数为

$$f(x, y) = \begin{cases} 1 & x \in (0, 1), |y| < x \\ 0 & \text{其它,} \end{cases}$$

求条件概率密度  $f_{Y|X}(y|x)$  和  $f_{X|Y}(x|y)$ .

5.12 设随机向量  $(X, Y)$  的密度函数为

$$f(x, y) = \begin{cases} e^{-y} & y > x > 0 \\ 0 & \text{其它} \end{cases}$$

求条件概率密度  $f_{X|Y}(x|y)$ .

5.13 设随机向量  $(X, Y)$  的密度函数为

$$f(x, y) = \begin{cases} x + y & x, y \in (0, 1) \\ 0 & \text{其它,} \end{cases}$$

求  $Z_1 = X + Y$  和  $Z_2 = XY$  的密度函数.

5.14 设随机向量  $(X, Y)$  的密度函数为

$$f(x, y) = \begin{cases} A(x + y)e^{-x-y} & x > 0, y > 0 \\ 0 & \text{其它,} \end{cases}$$

求随机变量  $X$  与  $Y$  的独立性, 以及  $Z = X + Y$  的密度函数.

**5.15** 设随机向量  $(X, Y)$  的密度函数为

$$f(x, y) = \begin{cases} Ae^{-x-y} & 0 < x < 1, y > 0 \\ 0 & \text{其它} \end{cases}$$

求 1) 常数  $A$ ; 2)  $X$  与  $Y$  的边缘密度函数, 3)  $Z = \max(X, Y)$  的密度函数.

**5.16** 设随机变量  $X \sim U(0, 1)$  和  $Y \sim e(1)$  相互独立, 求  $Z = X + Y$  的概率密度.

**5.17** 若随机变量  $X \sim e(\lambda_1)$  和  $Y \sim e(\lambda_2)$  相互独立, 求  $Z = X + Y$  的分布函数和概率密度.

**5.18** 若随机变量  $X \sim e(1)$  和  $Y \sim e(1)$  相互独立, 求  $Z = Y/X$  的概率密度.

**5.19** 若随机变量  $X \sim G(p_1)$  和  $Y \sim G(p_2)$  相互独立, 求  $Z = X + Y$  的分布列.

**5.20** 若相互独立的随机变量  $X$  和  $Y$  分别服从参数为  $\lambda_1$  和  $\lambda_2$  的泊松分布, 求在  $X + Y = n$  的条件下  $X$  的条件分布.

**5.21** 证明: 定义 5.9 和定义 5.15 中关于二维正太分布的密度函数完全相等.

**5.22** 证明定理 5.12.

**5.23** 证明定理 5.13.



## 第6章 多维随机向量的数字特征

### 6.1 多维随机向量函数的期望

前面介绍了一维随机变量及其函数的期望, 下面研究二维随机向量函数的期望, 同理可推广到维度更多的随机变量.

**定理 6.1** 设二维离散型随机向量  $(X, Y)$  的分布列为  $p_{ij} = P(X = x_i, Y = y_j)$ , 则随机向量函数  $g(X, Y)$  的期望为

$$E[g(X, Y)] = \sum_{i,j} g(x_i, y_j) p_{ij};$$

设二维连续型随机向量  $(X, Y)$  的密度函数为  $f(x, y)$ , 则随机向量函数  $g(X, Y)$  的期望为

$$E[g(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy.$$

定理的证明超出了本书的范畴而被略去. 当  $g(X, Y) = X$  时, 对二维连续型随机变量有

$$E[X] = \int_{-\infty}^{+\infty} x dx \int_{-\infty}^{+\infty} f(x, y) dy = \int_{-\infty}^{+\infty} x f_X(x) dx,$$

随机向量  $(X, Y)$  中  $X$  的期望就是其边缘分布的期望, 同理可得  $Y$  的期望, 对离散情况结论也成立.

当  $g(X, Y) = (X - E(X))^2$  时有

$$E[g(X, Y)] = \int_{-\infty}^{+\infty} (x - E(X))^2 dx \int_{-\infty}^{+\infty} f(x, y) dy = \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x) dx = \text{Var}(X),$$

$X$  的方差就是边缘分布的方差, 同理可得  $Y$  的方差. 根据期望的定义有

**性质 6.1** 若随机变量  $X, Y$  满足  $X \geq Y$ , 则有  $E[X] \geq E[Y]$ .

**性质 6.2** 对任意随机变量  $X, Y$  有  $E[X + Y] = E[X] + E[Y]$ .

**证明** 这里仅仅给出连续情况的证明, 同理可以考虑类型随机变量.

$$\begin{aligned} E(X + Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y) f(x, y) dx dy = \int_{-\infty}^{+\infty} x dx \int_{-\infty}^{+\infty} f(x, y) dy \\ &\quad + \int_{-\infty}^{+\infty} y dy \int_{-\infty}^{+\infty} f(x, y) dx = \int_{-\infty}^{+\infty} x f_X(x) dx + \int_{-\infty}^{+\infty} y f_Y(y) dy = E(X) + E(Y), \end{aligned}$$

由此完成证明.

性质 6.2 可推广到  $n$  个随机变量, 即  $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$ .

性质 6.3 对相互独立的随机变量  $X$  和  $Y$ , 有

$$E[XY] = E[X]E[Y];$$

对任意随机变量  $X$  和  $Y$ , 有 Cauchy-Schwartz 不等式

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}.$$

**证明** 这里仅仅给出连续情况的证明, 同理可以考虑类型随机变量. 根据随机变量  $X$  与  $Y$  的独立性有  $f(x, y) = f_X(x)f_Y(y)$ , 由此可得

$$\begin{aligned} E[XY] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x, y)dxdy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf_X(x)yf_Y(y)dxdy \\ &= \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy = E(X)E(Y). \end{aligned}$$

对任意随机变量  $X$  和  $Y$ , 以及对任意  $t \in \mathbb{R}$  有  $E[(X + tY)^2] \geq 0$ , 即任意  $t \in \mathbb{R}$ ,

$$t^2E[Y^2] + E[X^2] + 2tE[XY] \geq 0.$$

因此有  $\Delta = 4[E(XY)]^2 - 4E[X^2]E[Y^2] \leq 0$ , 即  $|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$ .

根据性质 6.3, 若随机变量  $X$  和  $Y$  相互独立, 则有

$$E[h(X)g(Y)] = E[h(X)]E[g(Y)];$$

若随机变量  $X_1, X_2, \cdots, X_n$  相互独立, 则有

$$E[X_1X_2 \cdots X_n] = E[X_1]E[X_2] \cdots E[X_n].$$

性质 6.4 设随机变量  $X$  与  $Y$  相互独立, 则有

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y).$$

**证明** 根据方差的定义有

$$\begin{aligned} \text{Var}(X \pm Y) &= E[(X - EX \pm (Y - EY))^2] \\ &= E(X - EX)^2 + E(Y - EY)^2 \pm 2E[(X - EX)(Y - EY)] \\ &= \text{Var}(X) + \text{Var}(Y) \pm 2E[(X - EX)(Y - EY)]. \end{aligned}$$



根据性质 6.3, 我们有  $E[(X - EX)(Y - EY)]E(X - EX)E(Y - EY) = 0$ , 由此完成证明.

**例 6.1** 设随机变量  $X \sim \mathcal{N}(0, 1)$  和  $Y \sim \mathcal{N}(0, 1)$  相互独立, 求  $E[\max(X, Y)]$ .

**解** 根据独立性定义可得随机变量  $X$  和  $Y$  的联合概率密度为

$$f(x, y) = f_X(x)f_Y(y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}.$$

考虑区域  $D_1 = \{(x, y): x \geq y\}$  和  $D_2 = \{(x, y): x < y\}$ , 如图 6.1 所示. 于是得到

$$\begin{aligned} E[\max(X, Y)] &= \int \int_{D_1} xf(x, y)dx dy + \int \int_{D_2} yf(x, y)dx dy \\ &= \int_{-\infty}^{+\infty} dy \int_y^{+\infty} xf(x, y)dx + \int_{-\infty}^{+\infty} dx \int_x^{+\infty} yf(x, y)dy \\ &= 2 \int_{-\infty}^{+\infty} dy \int_y^{+\infty} xf(x, y)dx = \frac{1}{\pi} \int_{-\infty}^{+\infty} dy \int_y^{+\infty} xe^{-\frac{x^2+y^2}{2}} dx \\ &= \frac{1}{\pi} \int_{-\infty}^{+\infty} e^{-y^2} dy = \frac{1}{\sqrt{\pi}} \end{aligned}$$

最后一个等式成立是因为  $\int_{-\infty}^{+\infty} e^{-y^2} dy = \sqrt{\pi}$ .

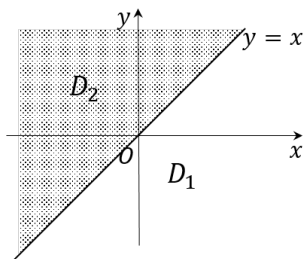


图 6.1 例 6.1 的积分区域  $D_1$  和  $D_2$ .

**例 6.2** 某水果超市在每星期一进货一定数量的新鲜水果, 假设一周内出售水果的件数  $X \sim U(10, 20)$ . 若这一周内出售一件水果获利 10 元, 若不能出售则因为水果过期而每件亏损 4 元, 求水果超市的最优进货策略.

**解** 不妨假设水果超市每周进货  $n$  件 ( $10 \leq n \leq 20$ ), 则它的周利润为

$$Y = \begin{cases} 10n & X \geq n \\ 10X - 4(n - X) & X < n \end{cases}.$$

由于周利润  $Y$  是关于  $X$  的随机变量, 只能考虑在期望下的最优策略

$$E[Y] = \sum_{i=10}^{n-1} (10i - 4(n - i))P(X = i) + \sum_{i=n}^{20} 10nP(X = i)$$

$$= \sum_{i=10}^{n-1} \frac{14i - 4n}{10} + \sum_{i=n}^{20} n = (-7n^2 + 243n + 630)/10 .$$

上式对  $n$  求一阶导数并令其等于零, 求解可得  $n = 17.36$ , 最后取  $n = 17$ .

## 6.2 协方差

随机变量的期望或方差仅涉及变量自身的统计信息, 没有刻画变量之间的统计信息. 本节引入一个新的统计特征: 协方差, 用于描述随机变量  $X$  和  $Y$  相互关系的数字特征.

**定义 6.1** 设二维随机向量  $(X, Y)$  的期望  $E[(X - E(X))(Y - E(Y))]$  存在, 则称其为  $X$  和  $Y$  的协方差, 记为

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

协方差是两个随机变量与它们各自期望的偏差之积的期望, 由于偏差可正可负, 因此协方差可正可负. 根据协方差的定义容易发现

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

下面进一步研究协方差的性质:

**性质 6.5** 对任意随机变量  $X, Y$  和常数  $c$  有

$$\text{Cov}(X, c) = 0 \quad \text{和} \quad \text{Cov}(X, Y) = \text{Cov}(Y, X) .$$

**性质 6.6** 对任意随机变量  $X, Y$  和常数  $a, b$  有

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y) \quad \text{和} \quad \text{Cov}(X + a, Y + b) = \text{Cov}(X, Y) .$$

**证明** 根据协方差的定义有

$$\text{Cov}(aX, bY) = E[(aX - E(aX))(bY - E(bY))] = abE[(X - E(X))(Y - E(Y))] ,$$

以及

$$\begin{aligned} \text{Cov}(X + a, Y + b) &= E[(X + a - E(X + a))(Y + b - E(Y + b))] \\ &= E[(X - E(X))(Y - E(Y))] . \end{aligned}$$

**性质 6.7** 对任意随机变量  $X_1, X_2, Y$  有

$$\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) .$$

**证明** 根据协方差的定义有

$$\begin{aligned} \text{Cov}(X_1 + X_2, Y) &= E[(X_1 + X_2 - E(X_1) - E(X_2))(Y - E(Y))] \\ &= E[(X_1 - E(X_1))(Y - E(Y))] + E[(X_2 - E(X_2))(Y - E(Y))] \\ &= \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) . \end{aligned}$$

可将性质 6.7 推广到多个随机变量: 对随机变量  $X_1, X_2, \dots, X_n$  和  $Y_1, Y_2, \dots, Y_m$  有

$$\text{Cov}\left(\sum_i^n X_i, \sum_j^m Y_j\right) = \sum_i^n \sum_j^m \text{Cov}(X_i, Y_j) ,$$

以及进一步有

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

**性质 6.8** 若随机变量  $X$  与  $Y$  相互独立, 则有  $\text{Cov}(X, Y) = 0$ , 但反之不成立.

**证明** 若  $X$  与  $Y$  相互独立, 则有  $E(XY) = E(X)E(Y)$ , 于是得到

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0 .$$

但反方向并不成立, 即若随机变量  $X$  与  $Y$  满足  $\text{Cov}(X, Y) = 0$ , 并不能得到  $X$  与  $Y$  相互独立. 下面来看一个反例, 设随机变量  $X$  的分布列为

$$P(X = -1) = P(X = 0) = P(X = 1) = 1/3 ,$$

很容易得到  $E(X) = 0$ . 引入一个新的随机变量  $Y$

$$Y = \begin{cases} 0 & X \neq 0 \\ 1 & X = 0 \end{cases} ,$$

可以发现  $E(XY) = 0$ , 最后得到

$$\text{Cov}(X, Y) = E[XY] - E(X)E(Y) = 0 .$$

但随机变量  $X$  与  $Y$  显然不是相互独立的, 因为

$$P(X = 0, Y = 0) = 0 \quad \text{而} \quad P(X = 0)P(Y = 0) = 2/9 .$$

由此完成证明.

**性质 6.9** 对任意随机变量  $X$  与  $Y$  有

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

等式成立的充要条件是  $Y = aX + b$  几乎处处成立, 即  $X$  与  $Y$  几乎处处存在线性关系.

**证明** 根据 Cauchy-Schwartz 不等式有

$$\begin{aligned} |\text{Cov}(X, Y)| &= |E[(X - E(X))(Y - E(Y))]| \\ &\leq \sqrt{E[(X - E(X))^2]E[(Y - E(Y))^2]} = \sqrt{\text{Var}(X)\text{Var}(Y)} . \end{aligned}$$

下面证明等号成立的充要条件. 若  $Y = aX + b$  几乎处处成立, 则有

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

由此直接验证  $(\text{Cov}(X, Y))^2 = \text{Var}(X)\text{Var}(Y)$ . 另一方面, 可以考虑函数

$$\begin{aligned} f(t) &= E[t(X - EX) - (Y - EY)]^2 \\ &= t^2 E[X - E(X)]^2 - 2tE[(X - E(X))(Y - E(Y))] + E[Y - E(Y)]^2 , \end{aligned}$$

根据方程的性质和条件  $(\text{Cov}(X, Y))^2 = \text{Var}(X)\text{Var}(Y)$  有

$$\Delta = 4(E[(X - EX)(Y - EY)])^2 - 4E(X - EX)^2 E(Y - EY)^2 = 0 .$$

由此存在一个根  $t_0$  使得  $f(t_0) = 0$  成立, 即

$$f(t_0) = E[(t_0(X - EX) - (Y - EY))^2] = 0 ,$$

由此可得  $Y = t_0(X - E(X)) + E(Y) = aX + b$  几乎处处成立.

**定理 6.2** 若随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则有  $\text{Cov}(X, Y) = \rho\sigma_x\sigma_y$ .

**证明** 根据协方差的定义有

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y)f(x, y)dx dy ,$$

其中

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right).$$

考虑变量变换

$$\begin{cases} u = (x - \mu_x)/\sigma_x \\ v = (y - \mu_y)/\sigma_y \end{cases} \quad \text{于是有} \quad \begin{cases} x = u\sigma_x + \mu_x \\ y = v\sigma_y + \mu_y \end{cases}.$$

容易得到其雅可比行列式为  $J = \sigma_x\sigma_y$ , 于是有

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{\sigma_x\sigma_y}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} uv \exp\left(-\frac{u^2 + v^2 - 2\rho uv}{2(1-\rho^2)}\right) dudv \\ &= \frac{\sigma_x\sigma_y}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} v \exp(-v^2/2) dv \int_{-\infty}^{+\infty} u \exp\left(-\frac{(u-\rho v)^2}{2(1-\rho^2)}\right) du \\ &= \frac{\sigma_x\sigma_y}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \rho v^2 \exp(-v^2/2) dv = \rho\sigma_x\sigma_y. \end{aligned}$$

其中第三个等式成立是因为利用正太分布  $\mathcal{N}(\rho v, 1-\rho^2)$  的期望, 而最后一个不等式成立是因为利用标准正太分布的方差.

结合定理 5.4 和定理 6.2 得到

**推论 6.1** 若随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则  $X$  和  $Y$  相互独立的充要条件是协方差  $\text{Cov}(X, Y) = 0$ .

**例 6.3** 设随机变量  $X_1, X_2, \dots, X_n$  相互独立且服从正太分布, 方差为  $\sigma^2$ . 记  $\bar{X} = \sum_{i=1}^n X_i/n$ , 讨论  $\bar{X}$  和  $\bar{X} - X_i$  的独立性.

**解** 根据正太分布的性质可以知道  $\bar{X}$  和  $\bar{X} - X_i$  都服从正太分布, 而正太分布的独立性可通过协方差来研究. 根据协方差的性质有

$$\text{Cov}(\bar{X}, \bar{X} - X_i) = \text{Cov}(\bar{X}, \bar{X}) - \text{Cov}(\bar{X}, X_i) = \text{Var}(\bar{X}) - \text{Cov}\left(\sum_{j=1}^n \frac{X_j}{n}, X_i\right).$$

根据  $X_1, X_2, \dots, X_n$  的相互独立性有

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{\sigma^2}{n} \quad \text{和} \quad \text{Cov}\left(\sum_{j=1}^n \frac{X_j}{n}, X_i\right) = \frac{1}{n} \text{Cov}(X_i, X_i) = \frac{\sigma^2}{n},$$

于是得到  $\text{Cov}(\bar{X}, \bar{X} - X_i) = 0$ . 根据推论 6.1 得到  $\bar{X}$  和  $\bar{X} - X_i$  是相互独立的.

例 6.4 随机变量  $(X, Y)$  联合概率密度为

$$f(x, y) = \begin{cases} (x+y)/8 & 0 \leq x \leq 2, 0 \leq y \leq 2 \\ 0 & \text{其它} \end{cases},$$

求  $\text{Cov}(X, Y)$  和  $\text{Var}(X+Y)$ .

解 根据协方差的定义有  $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$ , 需计算

$$E[X] = E[Y] = \int_0^2 \int_0^2 x(x+y)/8 dx dy = 7/6 \quad \text{和} \quad E[XY] = \int_0^2 \int_0^2 xy(x+y)/8 dx dy = 4/3,$$

由此可得  $\text{Cov}(X, Y) = -1/36$ . 进一步计算

$$E[X^2] = E[Y^2] = \int_0^2 \int_0^2 x^2(x+y)/8 dx dy = 5/3,$$

由此可得  $\text{Var}(X) = \text{Var}(Y) = 5/3 - (7/6)^2 = 11/36$ . 最后得到

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) = 11/18 - 1/18 = 5/9.$$

例 6.5 (匹配问题) 有  $n$  对夫妻参加一次聚会, 将所有参会人员任意分成  $n$  组, 每组一男一女, 用  $X$  表示夫妻两人被分到一组的对数, 求  $X$  的期望和方差.

解 用  $X_i$  表示第  $i$  对夫妻是否被分到一组, 即

$$X_i = \begin{cases} 1 & \text{第 } i \text{ 对夫妻被分到一组} \\ 0 & \text{否则} \end{cases},$$

则有  $X = X_1 + X_2 + \cdots + X_n$ . 随机变量  $X_i$  的分布列为

$$P(X_i = 1) = (n-1)!/n! = 1/n \quad \text{和} \quad P(X_i = 0) = 1 - 1/n.$$

于是得到期望

$$E(X) = E(X_1 + X_2 + \cdots, X_n) = E(X_1) + E(X_2) + \cdots + E(X_1) = 1.$$

对任意  $i \neq j$  有

$$P(X_i = 1, X_j = 1) = (n-2)!/n! = 1/n(n-1),$$

由此得到

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E(X_i)E(X_j) = 1/n^2(n-1),$$

最后根据协方差的性质有

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i \neq j} \text{Cov}(X_i, X_j) = 1 .$$

### 6.3 相关系数

两个随机变量之间的关系可分为独立与非独立, 在非独立中又可以分为线性关系和非线性关系. 非线性关系较为复杂, 目前尚无好的办法来处理. 但线性相关程度可以通过线性相关系数来刻画, 下面给出具体的定义:

**定义 6.2** 设  $(X, Y)$  为二维随机向量, 如果它们的方差  $\text{Var}(X)$  和  $\text{Var}(Y)$  存在且都不为零, 则称  $\text{Cov}(X, Y)/\sqrt{\text{Var}(X)\text{Var}(Y)}$  为  $X$  与  $Y$  的 **线性相关系数**, 简称 **相关系数** (correlation coefficient), 记为

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} .$$

若  $\rho_{XY} > 0$ , 称  $X$  与  $Y$  **正相关**; 若  $\rho_{XY} < 0$ , 称  $X$  与  $Y$  **负相关**; 若  $\rho_{XY} = 0$ , 称  $X$  与  $Y$  **不相关**.

相关系数是根据协方差和方差所定义, 与协方差同号, 可以看作是对协方差的一种规范化. 相关系数的很多性质可以通过协方差获得:

- 根据协方差性质 6.9 可知

$$|\rho_{XY}| \leq 1 ,$$

$|\rho_{XY}| = 1$  的充要条件为  $X$  与  $Y$  几乎处处有线性关系  $Y = aX + b$ .

- 根据协方差性质 6.8 可知, 若  $X$  与  $Y$  相互独立, 则  $X$  与  $Y$  不相关 ( $\rho_{XY} = 0$ ), 但反之不成立;
- 随机变量  $X$  与  $Y$  不相关, 仅仅表示  $X$  与  $Y$  之间不存在线性关系, 可能存在其他关系. 例如, 设随机变量  $X \sim U(-1/2, 1/2)$  和  $Y = \cos(X)$ , 则有

$$\text{Cov}(X, Y) = E(X \cos(X)) - E(X)E(\cos(X)) = E[X \cos(X)] = \int_{-1/2}^{1/2} x \cos(x) dx = 0 .$$

根据定理 6.2 和推论 6.1 有

**定理 6.3** 若随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则  $X$  与  $Y$  的相关系数  $\rho_{XY} = \rho$ , 以及  $X$  与  $Y$  独立的充要条件是  $X$  与  $Y$  不相关.

独立与不相关的等价性仅限于正太分布随机变量, 对其它类型并不一定成立, 详情见性质 6.8 证明中的反例.

根据相关系数、协方差和方差的定义很容易得到如下几个不相关的等价条件.

**定理 6.4** 设随机变量  $X$  和  $Y$  的方差存在且都不为零, 以下几个条件相互等价:

- $X$  和  $Y$  不相关, 即  $\rho_{XY} = 0$ ;
- 协方差  $\text{Cov}(X, Y) = 0$ ;
- $E(XY) = E(X)E(Y)$ ;
- $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ .

**例 6.6** 设随机变量  $X \sim \mathcal{N}(\mu, \sigma^2)$  和  $Y \sim \mathcal{N}(\mu, \sigma^2)$  相互独立. 求  $Z_1 = \alpha X + \beta Y$  和  $Z_2 = \alpha X - \beta Y$  的相关系数 ( $\alpha, \beta \neq 0$ ).

**解** 根据正态分布的定义有

$$\begin{aligned}\text{Cov}(Z_1, Z_2) &= \text{Cov}(\alpha X + \beta Y, \alpha X - \beta Y) = (\alpha^2 - \beta^2)\sigma^2 \\ \text{Var}(Z_1) &= \text{Cov}(\alpha X + \beta Y, \alpha X + \beta Y) = (\alpha^2 + \beta^2)\sigma^2 \\ \text{Var}(Z_2) &= \text{Cov}(\alpha X - \beta Y, \alpha X - \beta Y) = (\alpha^2 + \beta^2)\sigma^2,\end{aligned}$$

由此可知  $\rho_{XY} = (\alpha^2 - \beta^2)/(\alpha^2 + \beta^2)$ .

**例 6.7** 设离散型随机向量  $(X_1, X_2, \dots, X_n)$  服从多项分布  $M(m, p_1, p_2, \dots, p_n)$ , 对任意  $i \neq j$ , 求  $X_i$  和  $X_j$  的相关系数.

根据多项分布的定义有  $X_1 + X_2 + \dots + X_n = m$ , 当  $X_i$  越大时  $X_j$  应该越小, 因此直观而言这两随机变量之间应该是负相关的.

**解** 根据多项分布的性质有  $X_i$  和  $X_j$  的边缘分布分别

$$X_i \sim B(m, p_i) \quad \text{和} \quad X_j \sim B(m, p_j).$$

由此可得  $\text{Var}(X_i) = mp_i(1 - p_i)$  和  $\text{Var}(X_j) = mp_j(1 - p_j)$ . 直接求解  $\text{Cov}(X_i, X_j)$  存在一定的难度. 对每个  $k \in [m]$ , 引入随机变量

$$Y_i^k = \begin{cases} 1 & \text{若第 } k \text{ 次实验的结果为 } i \\ 0 & \text{其它} \end{cases} \quad \text{和} \quad Y_j^k = \begin{cases} 1 & \text{若第 } k \text{ 次实验的结果为 } j \\ 0 & \text{其它} \end{cases}.$$

由此可得

$$X_i = Y_i^1 + Y_i^2 + \dots + Y_i^m \quad \text{和} \quad X_j = Y_j^1 + Y_j^2 + \dots + Y_j^m.$$

根据第  $k$  次试验和第  $l$  次试验相互独立 ( $k \neq l$ ), 以及  $Y_i^k Y_j^k = 0$  有

$$\text{Cov}(Y_i^k, Y_j^l) = 0, \quad \text{以及} \quad \text{Cov}(Y_i^k, Y_j^k) = E[Y_i^k Y_j^k] - E(Y_i^k)E(Y_j^k) = -p_i p_j,$$



根据协方差的性质有

$$\text{Cov}(X_i, X_j) = \sum_{k=1}^m \text{Cov}(Y_i^k, Y_j^k) + \sum_{k \neq l} \text{Cov}(Y_i^k, Y_j^l) = -mp_i p_j .$$

最后得到  $X_i$  和  $X_j$  的相关系数为

$$\rho = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i)\text{Var}(X_j)}} = \frac{-mp_i p_j}{\sqrt{mp_i(1-p_i)}\sqrt{mp_j(1-p_j)}} = -\frac{\sqrt{p_i p_j}}{\sqrt{(1-p_i)(1-p_j)}} .$$

## 6.4 条件期望

前一章介绍了条件分布, 基于条件分布可以考虑条件期望, 分离散和连续性随机变量两种情况.

**定义 6.3** 设  $(X, Y)$  为连续型随机变量, 在  $Y = y$  条件下  $X$  的条件密度函数为  $f_{X|Y}(x|y)$ , 称

$$E(X|y) = E(X|Y = y) = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx$$

为在  $Y = y$  条件下  $X$  的 **条件期望**. 设  $(X, Y)$  为离散型随机变量, 在  $Y = y$  条件下  $X$  的条件分布列为  $P(X = x_i|Y = j)$ , 称

$$E(X|y) = E(X|Y = y) = \sum_i x_i P(X = x_i|Y = j)$$

为在  $Y = y$  条件下  $X$  的 **条件期望**.

条件期望  $E[X|y]$  一般都与  $y$  相关, 是  $y$  的函数, 而 (无条件) 期望  $E(X)$  是一个具体的常数. 在上面的定义中, 我们都默认期望存在. 条件期望是条件分布的期望, 具有期望的一切性质:

- 对任意常数  $a, b$  有  $E(aX_1 + bX_2|Y) = aE(X_1|Y) + bE(X_2|Y)$ ;
- 对离散型随机变量  $(X, Y)$  和函数  $g(X)$  有

$$E(g(X)|Y) = \sum_i g(x_i) P(X = x_i|Y = y) ;$$

对连续型随机变量  $(X, Y)$  和函数  $g(X)$  有

$$E(g(X)|Y) = \int_{-\infty}^{+\infty} g(x) f(x|Y = y) dx ;$$

- 设随机向量  $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ , 则在  $Y = y$  的条件下随机变量  $X$  服从正太分布  $\mathcal{N}(\mu_x - \rho\sigma_x(y - \mu_y)/\sigma_y, (1 - \rho^2)\sigma_x^2)$ , 由此可得  $E(X|y) = \mu_x - \rho\sigma_x(y - \mu_y)/\sigma_y$ .

下面给出了计算期望的另一种方法.

**定理 6.5** 对二维随机变量  $(X, Y)$  有

$$E(X) = E_Y(E(X|Y)) = \begin{cases} \sum_{y_j} E(X|y_j)P(Y = y_j) & \text{离散型随机变量,} \\ \int_{-\infty}^{\infty} E(X|y)f_Y(y)dy & \text{连续型随机变量.} \end{cases}$$

**证明** 对连续型随机变量  $(X, Y)$ , 不妨假设其联合密度函数为  $f(x, y)$ , 根据条件概率有

$$E[X] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x, y)dydx = \int_{-\infty}^{+\infty} f_Y(y) \int_{-\infty}^{+\infty} xf_{X|Y}(x|y)dx dy = \int_{-\infty}^{+\infty} E(X|y)f_Y(y)dy.$$

对离散型随机变量  $(X, Y)$ , 根据条件概率和全概率公式有

$$\begin{aligned} E[X] &= \sum_i x_i P_X(X = x_i) = \sum_i \sum_j x_i P(X = x_i, Y = y_j) \\ &= \sum_i \sum_j x_i P(X = x_i|Y = y_j)P(Y = y_j) \\ &= \sum_j P(Y = y_j) \sum_i x_i P(X = x_i|Y = y_j) \\ &= \sum_j P(Y = y_j)E[X|Y = y_j] = E_Y[E[X|Y]]. \end{aligned}$$

下面介绍与全概率公式相对于的一个公式: **全期望公式** (law of total expectation), 在期望的计算起到重要作用.

**定理 6.6** 设  $A_1, A_2, \dots, A_n$  是样本空间  $\Omega$  的一个分割, 即  $A_i A_j = \emptyset$  和  $\Omega = \cup_{i=1}^n A_i$ . 对任意随机变量  $X$  有

$$E[X] = E[X|A_1]P(A_1) + E[X|A_2]P(A_2) + \dots + E[X|A_n]P(A_n),$$

特别地, 随机事件  $A$  与其对立事件  $\bar{A}$  构成样本空间  $\Omega$  的一个划分, 对任意随机变量  $X$  有

$$E[X] = E[X|A]P(A) + E[X|\bar{A}]P(\bar{A}).$$

**证明** 对于随机变量  $X$  和  $A_1, A_2, \dots, A_n$ , 首先引入新的离散随机变量  $Y = 1, 2, \dots, n$  满足

随机事件  $Y = i$  发生的充要条件是  $X \in A_i$ .

根据定理 6.5 可知

$$E(X) = E_Y(E(X|Y)) = \sum_{i=1}^n E(X|Y = i)P(Y = i) = \sum_{i=1}^n E(X|A_i)P(A_i).$$

**例 6.8** 设  $(X, Y)$  的联合概率密度为

$$f(x, y) = \begin{cases} \exp(-y) & 0 < x < y < +\infty \\ 0 & \text{其它} \end{cases},$$

求条件期望  $E(X|y)$ .

**解** 首先计算  $Y$  的边缘密度函数, 当  $y > 0$  时

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)dx = \int_0^y \exp(-y)dx = y \exp(-y),$$

由此得到在  $Y = y$  的条件下  $X$  的条件分布

$$f_{X|Y}(x|y) = f(x, y)/f_Y(y) = 1/y \quad (0 < x < y < +\infty).$$

最后得到条件期望

$$E(X|y) = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx = \int_0^y x/y dx = y/2.$$

**例 6.9** 一矿工被困在有三个门的矿井里, 第一个门通一坑道, 沿此坑道走 3 小时可使他到达安全地点; 第二个门可使他走 5 小时后义回到原地; 第三个门可使他走 7 小时后也回到原地. 如设此矿工在任何时刻都等可能地选定其中一个门, 试问他到达安全地点平均要用多长时间?

**解** 用  $X$  为该矿工到达安全地点所需时间, 用  $Y$  为他所选的门, 根据全期望公式有

$$E(X) = E(X|Y=1)P(Y=1) + E(X|Y=2)P(Y=2) + E(X|Y=3)P(Y=3),$$

其中  $P(Y=1) = P(Y=2) = P(Y=3) = 1/3$ ,  $E(X|Y=1) = 3$ . 用  $E(X|Y=2)$  表示矿工从第二个门出去要到达安全地点所需平均时间. 而他沿此坑道走 5 小时又转回原地, 而一旦返回原地, 问题就与当初他还没有进第二个门之前一样, 因此他要到达安全地点平均还需再用  $E(X)$  小时. 同理可以考虑  $E(X|Y=2)$ , 故有

$$E(X|Y=2) = 5 + E(X) \quad \text{和} \quad E(X|Y=3) = 7 + E(X).$$

于是得到

$$E(X) = (3 + 5 + E(X) + 7 + E(X))/3.$$

求解出  $E(X) = 15$  (小时), 该矿工到达安全地点平均需要 15 小时.

### 6.5 随机向量的数学期望与协方差阵

**定义 6.4** 设随机向量  $X = (X_1, X_2, \dots, X_n)^\top$ , 称

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

为随机向量  $X$  的期望, 以及称

$$\text{Cov}(X) = \Sigma = \begin{pmatrix} \text{Cov}(X_1, X_1) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & & \vdots \\ \text{Cov}(X_n, X_1) & \cdots & \text{Cov}(X_n, X_n) \end{pmatrix}$$

为随机变量  $X$  的协方差矩阵.

下面介绍协方差矩阵的一些性质:

**定理 6.7** 随机向量  $X = (X_1, X_2, \dots, X_n)$  的协方差矩阵是对称半正定的矩阵.

**证明** 对任意  $i \neq j$ , 根据协方差的性质

$$\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i),$$

可知协方差矩阵是对称的. 对于半正定性的证明, 首先引入新的函数

$$f(t_1, t_2, \dots, t_n) = (t_1, t_2, \dots, t_n) \begin{pmatrix} \text{Cov}(X_1, X_1) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & & \vdots \\ \text{Cov}(X_n, X_1) & \cdots & \text{Cov}(X_n, X_n) \end{pmatrix} (t_1, t_2, \dots, t_n)^\top,$$

由此得到

$$f(t_1, t_2, \dots, t_n) = E \left( \left( \sum_{i=1}^n t_i (X_i - E(X_i)) \right)^2 \right) \geq 0,$$

由此完成证明.

**定理 6.8** 设多维正态分布  $X = (X_1, X_2, \dots, X_n)^\top \sim N(\mu, \Sigma)$ , 则有

$$\mu = (E[X_1], E[X_2], \dots, E[X_n])^\top \quad \text{和} \quad \Sigma = [\text{Cov}(X_i, X_j)]_{n \times n}.$$

## 6.6 应用案例

有时我们能观察到随机变量  $X$  的值, 需要对随机变量  $Y$  的值进行预测, 即选择一个函数  $g(x)$ , 使得  $g(X)$  接近预测值  $Y$ . 选择函数  $g(x)$  的一个准则是最小化  $E[(Y - g(x))^2]$ . 关于最优的函数  $g(X)$ , 有如下结论:

**定理 6.9** 对任意函数  $g(x)$  和随机变量  $X$  和  $Y$ , 有

$$E([Y - g(X)]^2) \geq E([Y - E(Y|X)]^2) .$$

**证明** 根据定理 6.5 只需证明对任意给定  $X$  有

$$E([Y - g(X)]^2|X) \geq E([Y - E(Y|X)]^2|X) , \quad (6.1)$$

对上式两边分别对  $X$  求期望可完成证明. 下面考虑如何上面的条件期望不等式, 首先有

$$\begin{aligned} E([Y - g(X)]^2|X) &= E([Y - E(Y|X) + E(Y|X) - g(X)]^2|X) \\ &= E([Y - E(Y|X)]^2|X) + E([E(Y|X) - g(X)]^2|X) + 2E([Y - E(Y|X)][E(Y|X) - g(X)]|X) , \end{aligned}$$

当给定  $X$  后,  $[E(Y|X) - g(X)]^2$  和  $E(Y|X)$  都是常数, 因此有

$$E([Y - E(Y|X)][E(Y|X) - g(X)]|X) = [E(Y|X) - g(X)]E([Y - E(Y|X)]|X) = 0 ,$$

结合上面两式完成 (6.1) 的证明.

很多情况下很难知道随机变量  $X$  和  $Y$  的联合分布, 有些情况下即使知道联合分布计算  $E(Y|X)$  也非常复杂. 若已知随机变量  $X$  和  $Y$  的一些统计量, 依然可以很好地估计出  $X$  和  $Y$  的最优线性预测, 例如,

**例 6.10** 设随机变量  $X$  和  $Y$  的期望、方差、相关系数分别为  $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho$ , 其中  $\sigma_x > 0, \sigma_y > 0, \rho \in [-1, +1]$ , 求解最优的线性预测  $Y = aX + b$  使得  $E((Y - aX - b)^2)$  最小化.

**解** 首先设函数

$$\begin{aligned} F(a, b) &= E((Y - aX - b)^2) \\ &= E(Y^2) - 2aE(Y) - 2bE(XY) + a^2 + 2abE(X) + b^2E(X^2) . \end{aligned}$$

求函数  $F(a, b)$  的最小值, 可以考虑令  $a$  和  $b$  的偏导等于零, 即

$$\begin{cases} \partial F(a, b)/\partial a = 2a + 2bE(X) - 2E(Y) = 0 \\ \partial F(a, b)/\partial b = 2bE(X^2) + 2aE(X) - 2E(XY) = 0 . \end{cases}$$

求解上面的方程组可得

$$\begin{cases} a = E(Y) + bE(X) = \mu_y - \rho\sigma_y\mu_x/\sigma_x \\ b = \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\rho\sigma_x\sigma_y}{\sigma_x^2} = \rho\sigma_y/\sigma_x . \end{cases}$$

由此给出  $Y$  的最优线性预测为

$$Y = \rho\sigma_y(X - \mu_x)/\sigma_x + \mu_y ,$$

在最优线性预测下预测的均分误差

$$\begin{aligned} & E((Y - \rho\sigma_y(X - \mu_x)/\sigma_x - \mu_y)^2) \\ &= E((Y - \mu_y)^2) + \rho^2\sigma_y^2 E((X - \mu_x)^2) / \sigma_x^2 - 2\rho\sigma_y E((X - \mu_x)(Y - \mu_y)) / \sigma_x \\ &= \sigma_y^2 + \rho^2\sigma_y^2 - 2\rho^2\sigma_y^2 = \sigma_y^2(1 - \rho^2) . \end{aligned}$$

由此可以看出, 当  $\rho^2 \rightarrow 1$  时最优线性预测的均方误差接近零.

## 习题

- 6.1 随机变量  $X$  与  $Y$  独立, 且  $\text{Var}(X) = 6$  和  $\text{Var}(Y) = 3$ , 求  $\text{Var}(2X \pm Y)$ .
- 6.2 设随机变量  $X \sim P(2)$  和  $Y \sim \mathcal{N}(-2, 4)$  相互独立, 求  $E[(X - Y)^2]$ .
- 6.3 在长度为  $l$  的线段上任取两点  $X, Y$ , 求期望  $E[\min(X, Y)]$  和  $E[|X - Y|]$ .
- 6.4 随机变量  $X \sim \mathcal{N}(-1, 2)$  和  $Y \sim \mathcal{N}(1, 8)$ , 且相关系数  $\rho_{XY} = -1/2$ . 求  $\text{Var}(X + Y)$ .
- 6.5 对任意实数  $x, y \in (0, 1)$  且满足  $x + y \leq 1$ , 证明  $xy/(1 - x)(1 - y) \in (0, 1)$ .
- 6.6 证明: 定义 5.9 和定义 5.15 中关于二维正太分布的密度函数完全相等.





## 第 7 章 集中不等式 (Concentration)

给定一个训练数据集

$$S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\},$$

其中  $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$  表示第  $i$  个训练样本的特征 (feature),  $y_i \in \mathcal{Y} = \{0, 1\}$  表示第  $i$  个训练样本的标记 (二分类). 假设  $\mathcal{D}$  是空间  $\mathcal{X} \times \mathcal{Y}$  的一个未知不可见的联合分布. 机器学习的经典假设是训练数据集  $S_n$  中每个数据  $(\mathbf{x}_i, y_i)$  是根据分布  $\mathcal{D}$  独立同分布采样所得.

给定一个函数或分类器  $f: \mathcal{X} \rightarrow \{0, 1\}$ , 定义函数  $f$  在训练数据集  $S_n$  上的分类错误率为

$$\hat{R}(f, S_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(f(\mathbf{x}_i) \neq y_i),$$

这里  $\mathbb{I}(\cdot)$  表示指示函数, 当论断为真时其返回值为 1, 否则为 0. 在实际应用中我们更关心函数  $f$  对未见数据的分类性能, 即函数  $f$  在分布  $\mathcal{D}$  上的分类错误率, 称之为 ‘泛化错误率’

$$R(f, \mathcal{D}) = E_{(\mathbf{x}, y) \sim \mathcal{D}}(\mathbb{I}(f(\mathbf{x}) \neq y)) = \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[f(\mathbf{x}) \neq y].$$

由于分布  $\mathcal{D}$  不可知, 不能直接计算  $R(f, \mathcal{D})$ , 但我们已知训练数据集  $S_n$  和训练错误率  $\hat{R}(f, S_n)$ , 如何基于训练错误率  $\hat{R}(f, S_n)$  来有效估计  $R(f, \mathcal{D})$ ? 我们可以将问题归纳为

$$\Pr_{S_n \sim \mathcal{D}^n} \left[ |\hat{R}(f, S_n) - R(f)| \geq t \right] \text{ 是否足够小?}$$

即能否以很大的概率保证

$$|\hat{R}(f, S_n) - R(f)| < t.$$

从而在理论上保证  $\hat{R}(f, S_n)$  是  $R(f)$  的一个有效估计. 上述性质在机器学习被称为 ‘泛化性’, 是机器学习模型理论研究的根本性质, 研究模型能否从可见的训练数据推导出对未见数据的处理能力.

首先来看一种简单的例子:

**例 7.1** 假设训练数据集  $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$  根据分布  $\mathcal{D}$  独立采样所得, 分类器  $f$  在训练集  $S_n$  的错误率为零 (全部预测正确), 求分类器  $f$  在分布  $\mathcal{D}$  上的错误率介于 0 和  $\epsilon$  之间的概率 ( $\epsilon > 0$ ).

**解** 设随机变量

$$X_i = \mathbb{I}[f(\mathbf{x}_i) \neq y_i] \quad (i \in [n]),$$

根据数据集的独立同分布假设可知  $X_1, X_2, \dots, X_n$  是独立同分布的随机变量. 令  $p = E[X_i]$ , 则有  $X_i \sim \text{Ber}(p)$ . 分类器  $f$  在训练集  $S_n$  的错误率为零, 且在分布  $\mathcal{D}$  上的错误率大于  $\epsilon$  的概率为

$$\begin{aligned} \Pr \left[ \sum_{i=1}^n X_i = 0, p > \epsilon \right] &\leq \Pr \left[ \sum_{i=1}^n X_i = 0 | p > \epsilon \right] \\ &= \Pr [X_1 = 0, X_2 = 0, \dots, X_n = 0 | p > \epsilon] \quad (\text{根据独立性假设}) \\ &= \prod_{i=1}^n \Pr [X_i = 0 | p > \epsilon] \leq (1 - \epsilon)^n \leq \exp(-n\epsilon). \end{aligned}$$

因此当分类器  $f$  在训练集  $S_n$  的错误率为零且  $p \in (0, \epsilon)$  的概率至少以  $1 - \exp(-n\epsilon)$  成立.

对上例的求解进一步进行归纳, 设随机变量

$$X_i = \mathbb{I}(f(\mathbf{x}_i) \neq y_i),$$

则机器学习问题可通过概率统计抽象描述为: 假设有  $n$  个独立同分布的随机变量  $X_1, X_2, \dots, X_n$ , 如何从  $n$  个独立同分布的随机变量中以很大概率地获得期望  $E[X]$  的一个估计, 即

$$\Pr \left[ \left| \frac{1}{m} \sum_{i=1}^m X_i - E(X_i) \right| > \epsilon \right] \quad \text{非常小.}$$

后续研究将不再给出机器学习的实际应用, 仅仅讨论概率论中的随机变量, 但大家要了解随机变量背后的实际应用.

## 7.1 基础不等式

首先给出一些基础的概率或期望不等式. 首先研究著名的 Markov 不等式:

**定理 7.1** 对任意随机变量  $X \geq 0$  和  $\epsilon > 0$ , 有

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}.$$

**证明** 利用全期望公式考虑随机事件  $X \geq \epsilon$  有

$$E[X] = E[X | X \geq \epsilon]P(X \geq \epsilon) + E[X | X \leq \epsilon]P(X \leq \epsilon) \geq P(X \geq \epsilon)\epsilon,$$

从而完成证明.

利用 Markov 不等式可得到一系列有用的不等式:

**推论 7.1** 对任意随机变量  $X$  和  $\epsilon \geq 0$ , 以及单调递增的非负函数  $g(x)$ , 有

$$P(X \geq \epsilon) \leq \frac{E[g(X)]}{g(\epsilon)}.$$

利用 Markov 不等式可以推导 Chebyshev 不等式:

**定理 7.2 (Chebyshev 不等式)** 设随机变量  $X$  的均值为  $\mu$ , 则有

$$P(|X - \mu| > \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}.$$

**证明** 根据 Markov 不等式有

$$P(|X - \mu| > \epsilon) = P((X - \mu)^2 \geq \epsilon^2) \leq \frac{E(X - \mu)^2}{\epsilon^2} = \frac{\text{Var}(X)}{\epsilon^2}.$$

**例 7.2** 设随机变量  $X$  和  $Y$  的期望分别为  $-1$  和  $1$ , 方差分别为  $2$  和  $8$ , 以及  $X$  和  $Y$  的相关系数为  $-1/2$ , 利用 Chebyshev 不等式估计概率  $P(|X + Y| \geq 6)$  的上界.

**解** 根据随机变量  $X$  和  $Y$  的相关系数为  $-1$  可知

$$\text{Cov}(X, Y) = \rho_{XY} \sqrt{\text{Var}(X)\text{Var}(Y)} = -2.$$

由  $E[X + Y] = 0$ , 利用 Chebyshev 不等式有

$$\begin{aligned} P(|X + Y| \geq 6) &= P(|X + Y - E[X + Y]| \geq 6) \\ &\leq \text{Var}(X + Y)/36 = (\text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y))/36 = 1/6. \end{aligned}$$

比 Chebyshev 不等式更紧地 Cantelli 不等式, 又被成为单边 Chebyshev 不等式.

**引理 7.1** 随机变量  $X$  的均值  $\mu > 0$ , 方差  $\sigma^2$ , 则对任意  $\epsilon > 0$  有

$$P(X - \mu \geq \epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2} \quad \text{和} \quad P(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

**证明** 设随机变量  $Y = X - \mu$ , 有  $E(Y) = 0$  以及  $\text{Var}(Y) = \sigma^2$ . 对任意  $t > 0$  有

$$\begin{aligned} P(X - \mu \geq \epsilon) &= P(Y \geq \epsilon) = P(Y + t \geq \epsilon + t) \leq P((Y + t)^2 \geq (\epsilon + t)^2) \\ &\leq \frac{E((Y + t)^2)}{(\epsilon + t)^2} = \frac{\sigma^2 + t^2}{(\epsilon + t)^2}. \end{aligned}$$

对  $(\sigma^2 + t^2)/(\epsilon + t)^2$  求关于  $t$  的最小值, 求解可得  $t = \sigma^2/\epsilon$ , 由此得到

$$P(X - \mu \geq \epsilon) \leq \min_{t>0} \frac{\sigma^2 + t^2}{(\epsilon + t)^2} = \frac{\sigma^2}{\epsilon^2 + \sigma^2}.$$

另一方面, 对任意  $t > 0$  有

$$\begin{aligned} P(X - \mu \leq -\epsilon) &= P(Y \leq -\epsilon) = P(Y - t \leq -\epsilon - t) \leq P((Y + t)^2 \geq (\epsilon + t)^2) \\ &\leq \frac{E((Y + t)^2)}{(\epsilon + t)^2} = \frac{\sigma^2 + t^2}{(\epsilon + t)^2}, \end{aligned}$$

同理完成证明.

下面介绍 Chebyshev 不等式的推论.

**推论 7.2** 设独立同分布的随机变量  $X_1, X_2, \dots, X_n$  满足  $E(X_i) = \mu$  和  $\text{Var}(X_i) \leq \sigma^2$ , 对任意实数  $\epsilon > 0$  有

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

**证明** 根据 Chebyshev 不等式有

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right).$$

而独立同分布的假设有

$$\text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^n X_i \right) = \frac{1}{n} \text{Var}(X_i) \leq \frac{\sigma^2}{n}.$$

由此得到

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right) \leq \frac{\sigma^2}{n\epsilon^2},$$

从而完成证明.

**例 7.3** 设分类器  $f$  在训练集  $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$  的错误率为  $\hat{p} > 0$ , 求分类器  $f$  在分布  $\mathcal{D}$  上的错误率在  $(9\hat{p}/10, 11\hat{p}/10)$  之间的概率.

**解** 设  $X_i = \mathbb{I}[f(\mathbf{x}_i) \neq y_i]$  ( $i \in [n]$ ), 则这些随机变量是独立同分布的. 训练错误率

$$\hat{p} = \sum_{i=1}^n X_i / n.$$

设分类器  $f$  在分布  $\mathcal{D}$  上的错误率为  $p$ , 则  $X_i \sim \text{Ber}(p)$  以及

$$p = E[X_i] = E \left[ \frac{1}{n} \sum_{i=1}^n X_i \right],$$

根据独立性假设和 Chebyshev 不等式有

$$\Pr[|p - \hat{p}| > \epsilon] \leq \frac{1}{\epsilon^2} \text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2} .$$

取  $\epsilon = \hat{p}/10$  有

$$\Pr[|p - \hat{p}| > \hat{p}/10] \leq \frac{25}{n\hat{p}^2} .$$

**引理 7.2 (Young 不等式)** 给定正常数  $a, b$ , 对任意满足  $1/p + 1/q = 1$  的正实数  $p, q$  有

$$ab \leq a^p/p + b^q/q .$$

**证明** 根据凸函数性质有

$$\begin{aligned} ab &= \exp(\ln(ab)) = \exp(\ln a + \ln b) \\ &= \exp \left( \frac{1}{p} \ln a^p + \frac{1}{q} \ln b^q \right) \leq \frac{1}{p} \exp(\ln a^p) + \frac{1}{q} \exp(\ln b^q) = \frac{1}{p} a^p + \frac{1}{q} b^q . \end{aligned}$$

引理得证.

根据 Young 不等式可证明著名的 Hölder 不等式.

**引理 7.3 (Hölder 不等式)** 设  $X$  和  $Y$  是随机变量, 若正数  $p, q$  满足  $1/p + 1/q = 1$ , 则有

$$E(|XY|) \leq [E(|X|^p)]^{1/p} [E(|Y|^q)]^{1/q} .$$

特别地, 当  $p = q = 2$  时 Hölder 不等式变成为 Cauchy-Schwartz 不等式.

**证明** 设  $c = [E(|X|^p)]^{1/p}$  和  $d = [E(|Y|^q)]^{1/q}$ , 根据 Young 不等式有

$$\frac{|XY|}{cd} = \frac{|X|}{c} \frac{|Y|}{d} \leq \frac{1}{p} \frac{|X|^p}{c^p} + \frac{1}{q} \frac{|Y|^q}{d^q} .$$

对上式两边同时取期望有

$$\frac{E(|XY|)}{cd} \leq \frac{1}{p} \frac{E(|X|^p)}{c^p} + \frac{1}{q} \frac{E(|Y|^q)}{d^q} = \frac{1}{p} + \frac{1}{q} = 1 ,$$

从而完成证明.

## 7.2 Chernoff 不等式

首先给出随机变量的矩生成函数 (Moment Generating Function) 的定义.

**定义 7.1** 定义随机变量  $X$  的矩生成函数为

$$M_X(t) = E[e^{tX}].$$

下面给出关于矩生成函数的一些性质:

**定理 7.3** 设随机变量  $X$  的矩生成函数为  $M_X(t)$ , 对任意  $n \geq 1$  有

$$E[X^n] = M_X^{(n)}(0),$$

这里  $M_X^{(n)}(t)$  表示矩生成函数在  $t = 0$  的  $n$  阶导数, 而  $E[X^n]$  被称为随机变量  $X$  的  $n$  阶矩 (moment).

**证明** 由 Taylor 公式有

$$e^{tX} = \sum_{i=0}^{\infty} \frac{(tX)^i}{i!}.$$

两边同时取期望有

$$E[e^{tX}] = \sum_{i=0}^{\infty} \frac{t^i}{i!} E[X^i].$$

对上式两边分别对  $t$  求  $n$  阶导数并取  $t = 0$  有  $M_X^{(n)}(0) = E[X^n]$ .

**定理 7.4** 对随机变量  $X$  和  $Y$ , 如果存在常数  $\delta > 0$ , 使得当  $t \in (-\delta, \delta)$  时有  $M_X(t) = M_Y(t)$  成立, 那么  $X$  与  $Y$  有相同的分布.

上述定理表明随机变量的矩生成函数可唯一确定随机变量的分布, 其证明超出了本书的范围. 若随机变量  $X$  与  $Y$  独立, 则有

$$M_{X+Y}(t) = E[e^{(X+Y)t}] = E[e^{tX} e^{tY}] = E[e^{tX}] \cdot E[e^{tY}] = M_X(t) M_Y(t).$$

于是得到

**推论 7.3** 对任意独立的随机变量  $X$  和  $Y$  有  $M_{X+Y}(t) = M_X(t) M_Y(t)$ .

下面将利用矩生成函数来证明一系列不等式. 给定任意随机变量  $X$  和任意  $t > 0$  和  $\epsilon > 0$ , 利用 Markov 不等式有

$$\Pr[X \geq E[X] + \epsilon] = \Pr[e^{tX} \geq e^{tE[X] + t\epsilon}] \leq e^{-t\epsilon - tE[X]} E[e^{tX}].$$

特别地, 有

$$\Pr[X \geq \epsilon] \leq \min_{t>0} \left\{ e^{-t\epsilon - tE[X]} E[e^{tX}] \right\}.$$

类似地, 对任意  $\epsilon > 0$  和  $t < 0$  有

$$\Pr[X \leq E[X] - \epsilon] = \Pr[e^{tX} \geq e^{tE[X] - t\epsilon}] \leq e^{t\epsilon - tE[X]} E[e^{tX}].$$

同理有

$$\Pr[X \leq \epsilon] \leq \min_{t < 0} \left\{ e^{t\epsilon - tE[X]} E[e^{tX}] \right\}.$$

上述方法称为‘**Chernoff 方法**’, 是证明集中不等式一种最根本最重要的方法. 下面将针对特定的分布或特定的条件, 先求解矩生成函数  $E[e^{tX}]$ , 然后求解最小值  $t$  的取值.

### 7.2.1 二值随机变量的 Chernoff 不等式

**定理 7.5** 设随机变量  $X_1, X_2, \dots, X_n$  相互独立且满足  $X_i \sim \text{Ber}(p_i)$ , 令  $\mu = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$ . 对任意  $\epsilon > 0$  有

$$\Pr \left[ \sum_{i=1}^n X_i \geq (1 + \epsilon)\mu \right] \leq \left( \frac{e^\epsilon}{(1 + \epsilon)^{(1 + \epsilon)}} \right)^\mu;$$

对任意  $\epsilon \in (0, 1)$  有

$$\Pr \left[ \sum_{i=1}^n X_i \geq (1 + \epsilon)\mu \right] \leq e^{-\mu\epsilon^2/3}.$$

上述第一个不等式给出了最紧的不等式上界, 第二个不等式是第一个不等式的适当放松.

**证明** 令  $\bar{X} = \sum_{i=1}^n X_i$ . 对任意  $t > 0$ , 根据 Chernoff 方法有

$$\Pr[\bar{X} \geq (1 + \epsilon)\mu] = \Pr[e^{t\bar{X}} \geq e^{t(1 + \epsilon)\mu}] \leq e^{-t(1 + \epsilon)\mu} E[e^{t\bar{X}}].$$

利用随机变量的独立性以及  $1 + x \leq e^x$  有

$$\begin{aligned} E[e^{t\bar{X}}] &= E[e^{\sum_{i=1}^n tX_i}] = \prod_{i=1}^n E[e^{tX_i}] \\ &= \prod_{i=1}^n [(1 - p_i) + p_i e^t] = \prod_{i=1}^n [1 + p_i(e^t - 1)] \\ &\leq \exp \left( \sum_{i=1}^n p_i(e^t - 1) \right) = \exp(\mu(e^t - 1)). \end{aligned}$$

由此可得

$$\Pr[\bar{X} \geq (1 + \epsilon)\mu] \leq \exp(-t(1 + \epsilon)\mu + \mu(e^t - 1)).$$

对上式求最小值得  $t_{\min} = \ln(1 + \epsilon)$ , 代入可得

$$\Pr[\bar{X} \geq (1 + \epsilon)\mu] \leq \left( \frac{e^\epsilon}{(1 + \epsilon)^{(1 + \epsilon)}} \right)^\mu.$$

对第二个不等式, 只需证明当  $\epsilon \in (0, 1)$  有

$$f(\epsilon) = \ln \left( \frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}} \right) + \frac{\epsilon^2}{3} = \epsilon - (1+\epsilon) \ln(1+\epsilon) + \frac{\epsilon^2}{3} \leq 0.$$

易知  $f(0) = 0$  和  $f(1) < 0$ . 当  $\epsilon \in (0, 1)$ ,

$$f'(\epsilon) = -\ln(1+\epsilon) + 2\epsilon/3, \quad f''(\epsilon) = -\frac{1}{1+\epsilon} + \frac{2}{3}.$$

于是得到  $f'(0) = 0$ ,  $f'(1) = -0.0265 < 0$  和  $f'(1/2) = -0.0721 < 0$ , 由连续函数性质有  $f'(\epsilon) \leq 0$ , 即函数  $f(\epsilon)$  在  $[0, 1]$  上单调递减. 当  $\epsilon \geq 0$  时有  $f(\epsilon) \leq f(0) = 0$ , 所以  $\exp(f(\epsilon)) \leq 1$ .

下面的定理给出了  $\Pr[\sum_{i=1}^n X_i \leq (1-\epsilon)\mu]$  的估计, 证明作为练习题留给大家完成.

**定理 7.6** 设随机变量  $X_1, X_2, \dots, X_n$  相互独立且满足  $X_i \sim \text{Ber}(p_i)$ , 令  $\mu = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$ . 对任意  $\epsilon \in (0, 1)$  有

$$\Pr \left[ \sum_{i=1}^n X_i \leq (1-\epsilon)\mu \right] \leq \left( \frac{e^{-\epsilon}}{(1-\epsilon)^{(1-\epsilon)}} \right)^\mu \leq \exp(-\mu\epsilon^2/2).$$

**定义 7.2** 若随机变量  $X \in \{+1, -1\}$  满足

$$\Pr(X = +1) = \Pr(X = -1) = 1/2,$$

则称  $X$  为 Rademacher 随机变量.

我们有如下定理:

**定理 7.7** 对  $n$  个独立的 Rademacher 随机变量  $X_1, X_2, \dots, X_n$ , 有

$$\Pr \left( \frac{1}{n} \sum_{i=1}^n X_i \geq \epsilon \right) \leq \exp(-n\epsilon^2/2) \quad \text{和} \quad \Pr \left( \frac{1}{n} \sum_{i=1}^n X_i \leq -\epsilon \right) \leq \exp(-n\epsilon^2/2).$$

**证明** 根据 Taylor 展开式有

$$\frac{1}{2} \exp(t) + \frac{1}{2} \exp(-t) = \sum_{i \geq 0} \frac{t^{2i}}{(2i)!} \leq \sum_{i \geq 0} \frac{(t^2/2)^i}{i!} = \exp(t^2/2).$$

若随机变量  $X \in \{+1, -1\}$  且满足  $\Pr(X = 1) = \Pr(X = -1) = 1/2$ , 则有

$$E[e^{tX}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} \leq \exp(t^2/2).$$




对任意  $t > 0$ , 根据 Chernoff 方法有

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) &\leq \exp(-nt\epsilon) E\left[\exp\left(\sum_{i=1}^n tX_i\right)\right] \\ &= \exp(-nt\epsilon) \prod_{i=1}^n E[\exp(tX_i)] \leq \exp(-nt\epsilon + nt^2/2).\end{aligned}$$

通过对上式右边求最小值解得  $t = \epsilon$ , 带入上式得到

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) \leq \exp(-n\epsilon^2/2).$$

同理证明另一个不等式.

 **推论 7.4** 对独立同分布的随机变量  $X_1, X_2, \dots, X_n$  满足  $P(X_1 = 0) = P(X_1 = 1) = 1/2$ , 有

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{2} \geq \epsilon\right) \leq \exp(-2n\epsilon^2) \quad \text{和} \quad \Pr\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{2} \leq -\epsilon\right) \leq \exp(-2n\epsilon^2).$$

### 7.2.2 有界随机变量的 Chernoff 不等式

本节研究有界的随机变量  $X_i \in [a, b]$  的 Chernoff 不等式. 首先介绍著名的 Chernoff 引理.

**引理 7.4** 设随机变量  $X \in [0, 1]$  的期望  $\mu = E[X]$ . 对任意  $t > 0$  有

$$E[e^{tX}] \leq \exp(t\mu + t^2/8).$$

**证明** 由凸函数的性质可知

$$e^{tX} = e^{tX+(1-X)0} \leq Xe^t + (1-X)e^0,$$

两边再同时取期望有

$$E(e^{tX}) \leq 1 - \mu + \mu e^t = \exp(\ln(1 - \mu + \mu e^t)).$$

令  $f(t) = \ln(1 - \mu + \mu e^t)$ , 我们有  $f(0) = 0$  以及

$$f'(t) = \frac{\mu e^t}{1 - \mu + \mu e^t} \Rightarrow f'(0) = \mu.$$

进一步有

$$f''(t) = \frac{\mu e^t}{1 - \mu + \mu e^t} - \frac{\mu^2 e^{2t}}{(1 - \mu + \mu e^t)^2} \leq 1/4.$$

根据泰勒中值定理有

$$f(t) = f(0) + tf'(0) + f''(\xi)t^2/2 \leq t\mu + t^2/8.$$

引理得证.

由上面的 Chernoff 引理进一步推导出

**推论 7.5** 设随机变量  $X \in [a, b]$  的期望  $\mu = E[x]$ . 对任意  $t > 0$  有

$$E(e^{tX}) \leq \exp(\mu t + t^2(b-a)^2/8).$$

根据上述推论, 我们得到有界随机变量的 Chernoff 不等式:

**定理 7.8** 假设  $X_1, \dots, X_n$  是  $n$  独立的随机变量、且满足  $X_i \in [a, b]$ . 对任意  $\epsilon > 0$  有

$$\begin{aligned} \Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E[X_i] \geq \epsilon \right] &\leq \exp(-2n\epsilon^2/(b-a)^2), \\ \Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E[X_i] \leq -\epsilon \right] &\leq \exp(-2n\epsilon^2/(b-a)^2). \end{aligned}$$

**证明** 这里给出第一个不等式的证明, 第二个不等式证明作为习题. 对任意  $t > 0$ , 根据 Chernoff 方法有

$$\begin{aligned} &\Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E[X_i] \geq \epsilon \right] \\ &= \Pr \left[ \sum_{i=1}^n t(X_i - E[X_i]) \geq nt\epsilon \right] \\ &\leq \exp(-nt\epsilon) E \left[ \exp \left( \sum_{i=1}^n t(X_i - E[X_i]) \right) \right] \\ &= \exp(-nt\epsilon) \prod_{i=1}^n E[\exp(t(X_i - E[X_i]))]. \end{aligned}$$

根据 Chernoff 引理, 对任意  $X_i \in [a, b]$  有

$$E[\exp(t(X_i - E[X_i]))] \leq \exp((b-a)^2 t^2/8).$$

由此得到

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E[X_i] \geq \epsilon \right] \leq \exp(-nt\epsilon + nt^2(b-a)^2/8).$$

对上式右边取最小值求解  $t = 4\epsilon/(b-a)^2$ , 然后带入上式可得:

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n E[X_i] \geq \epsilon \right] \leq \exp(-2n\epsilon^2/(b-a)^2).$$

从而完成证明.

### 7.2.3 Gaussian 和 Sub-Gaussian 随机变量不等式

首先考虑独立同分布的 Gaussian 随机变量:

**定理 7.9** 设随机变量  $X_1, \dots, X_n$  相互独立、且服从  $X_i \sim \mathcal{N}(\mu, \sigma)$ , 对任意  $\epsilon > 0$  有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] = \Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \leq -\epsilon \right] \leq \frac{1}{2} \exp(-n\epsilon^2/2\sigma^2).$$

**证明** 对随机变量  $X_i \sim \mathcal{N}(\mu, \sigma)$ , 根据正太分布的性质有

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \sim \mathcal{N}(0, \sigma^2/n) \Rightarrow \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu) \sim \mathcal{N}(0, 1).$$

若  $X' \sim \mathcal{N}(0, 1)$ , 对任意  $\epsilon > 0$ , 根据以前的定理有

$$P(X' \geq \epsilon) \leq \frac{1}{2} e^{-\epsilon^2/2}.$$

因此得到

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] = \Pr \left[ \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu) \geq \epsilon\sqrt{n}/\sigma \right] \leq \frac{1}{2} \exp(-n\epsilon^2/2\sigma^2),$$

定理得证.

下面定义 Sub-Gaussian 随机变量, 将有界随机变量和 Gaussian 随机变量统一起来:

**定义 7.3** 对任意  $t \in (-\infty, +\infty)$ , 若随机变量  $X$  满足

$$E[e^{(X-E[X])t}] \leq \exp(bt^2/2),$$

则称随机变量  $X$  是服从参数为  $b$  的亚高斯 (Sub-Gaussian) 随机变量.

亚高斯随机变量表示随机变量的尾部分布不会比一个高斯分布更严重.

**例 7.4** 对任意有界的随机变量  $X \in [a, b]$ , 根据 Chernoff 引理有

$$E[e^{(X-\mu)t}] \leq \exp(t^2(b-a)^2/8),$$

即有界的随机变量是参数为  $(b-a)^2/4$  的亚高斯随机变量.

**例 7.5** 如果随机变量  $X$  服从高斯分布  $\mathcal{N}(\mu, \sigma^2)$ , 则有

$$E[e^{(X-\mu)t}] = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{xt} e^{-x^2/2\sigma^2} dx = e^{\sigma^2 t^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(t\sigma-x/\sigma)^2/2} d(x/\sigma) = e^{\sigma^2 t^2/2}.$$

Gaussian 随机变量是参数为  $\sigma^2$  的亚高斯随机变量.

由前面的例子可知高斯随机变量和有界的随机变量都是亚高斯随机变量. 根据 Chernoff 方法有

**定理 7.10** 设  $X_1, \dots, X_n$  是  $n$  个独立的且参数为  $b$  的亚高斯随机变量, 对任意  $\epsilon > 0$  有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp(-n\epsilon^2/2b) \quad \text{和} \quad \Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \leq -\epsilon \right] \leq \exp(-n\epsilon^2/2b).$$

**证明** 对任意  $t > 0$ , 根据 Chernoff 方法有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq e^{-tn\epsilon} \prod_{i=1}^n E[e^{(X_i - \mu)t}] \leq e^{-tn\epsilon + nbt^2/2},$$

通过求解上式最小值可得  $t_{\min} = \epsilon/b$ , 代入完成证明.

对亚高斯型随机变量, 还可以给出最大值期望的估计:

**定理 7.11** 设  $X_1, \dots, X_n$  是  $n$  个相互独立的、参数为  $b$  的亚高斯随机变量, 且满足  $E[X_i] = 0$ , 我们有

$$E \left[ \max_{i \in [n]} X_i \right] \leq \sqrt{2b \ln n}.$$

**证明** 对任意  $t > 0$ , 根据 Jensen 不等式有

$$\begin{aligned} \exp \left( t E \left[ \max_{i \in [n]} X_i \right] \right) &\leq E \left[ \exp \left( t \max_{i \in [n]} X_i \right) \right] \\ &= E \left[ \max_{i \in [n]} \exp(tX_i) \right] \leq \sum_{i=1}^n E[\exp(tX_i)] \leq n \exp(t^2 b/2). \end{aligned}$$

对上式两边同时取对数整理可得

$$E \left[ \max_{i \in [n]} X_i \right] \leq \frac{\ln n}{t} + \frac{bt}{2}.$$

通过求解上式最小值可得  $t_{\min} = \sqrt{2 \ln n/b}$ , 代入完成证明.

前面所讲的概率不等式, 可以用另外一种表达形式给出, 这里以定理 7.10 为例: 假设  $X_1, \dots, X_n$  是独立的、且参数为  $b$  的亚高斯随机变量, 对任意  $\epsilon > 0$  有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp(-n\epsilon^2/2b).$$

令  $\delta = \exp(-n\epsilon^2/2b)$ , 求解出

$$\epsilon = \sqrt{2b \ln(1/\delta)/n},$$

代入整理可得: 至少以  $1 - \delta$  的概率有下面的不等式成立

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \leq \sqrt{\frac{2b}{n} \ln \frac{1}{\delta}}.$$

前面讲的所有不等式都可以采用  $1 - \delta$  的形式描述.

### 7.3 Bennet 和 Bernstein 不等式

通过考虑随机变量的方差, 可能推导出更紧地集中不等式, 下面介绍两个基于方差的不等式.

**定理 7.12 (Bennet不等式)** 设  $X_1, \dots, X_n$  是独立同分布的随机变量且满足  $X_i - E[X_i] \leq 1$ , 其均值为  $\mu$  和方差为  $\sigma^2$ , 我们有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp \left( -\frac{n\epsilon^2}{2\sigma^2 + 2\epsilon/3} \right).$$

**证明** 对任意  $t > 0$ , 根据 Chernoff 方法有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq e^{-nt\epsilon} E \left[ \exp \left( \sum_{i=1}^n t(X_i - \mu) \right) \right] = e^{-nt\epsilon} \left( E[e^{t(X_1 - \mu)}] \right)^n.$$

设  $Y = X_1 - \mu$ , 利用公式  $\ln z \leq z - 1$  得到

$$\begin{aligned} \ln E[e^{t(X_1 - \mu)}] &= \ln E[e^{tY}] \leq E[e^{tY}] - 1 = t^2 E \left[ \frac{e^{tY} - tY - 1}{t^2 Y^2} Y^2 \right] \\ &\leq t^2 E \left[ \frac{e^t - t - 1}{t^2} Y^2 \right] = (e^t - t - 1) \sigma^2 \end{aligned}$$

这里利用  $tY \leq t$  以及  $(e^z - z - 1)/z^2$  是一个非单调递减的函数. 进一步有

$$e^t - t - 1 \leq \frac{t^2}{2} \sum_{k=0}^{\infty} (t/3)^k = \frac{t^2}{2(1-t/3)}.$$

因此可得

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp \left( -nt\epsilon + \frac{nt^2\sigma^2}{2(1-t/3)} \right).$$

猜出  $t = \epsilon/(\sigma^2 + \epsilon/3)$ , 带入完成证明.

对于 Bennet 不等式, 令

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp(-n\epsilon^2/(2\sigma^2 + 2\epsilon/3)) = \delta,$$

可以给出不等式的另外一种表述: 至少以  $1 - \delta$  的概率有以下不等式成立

$$\frac{1}{n} \sum_{i=1}^n X_i \leq \mu + \frac{2 \ln 1/\delta}{3n} + \sqrt{\frac{2\sigma^2}{n} \ln \frac{1}{\delta}}.$$

当方法  $\sigma^2$  非常小, 或趋于 0 时, 得到更紧的收敛率  $\bar{X}_n - \mu \leq O(1/n)$ .

下面考虑另一种基于方差的不等式, 与 Bennet 不等式不同之处在于约束随机变量的矩:

**定理 7.13 (Bernstein不等式)** 设  $X_1, \dots, X_n$  是独立同分布的随机变量, 其均值为  $\mu$  和方差为  $\sigma^2$ , 若存在常数  $b > 0$ , 使得对任意正整数  $m \geq 2$  有  $E[X_i^m] \leq m!b^{m-2}\sigma^2/2$ , 那么我们有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp \left( -\frac{n\epsilon^2}{2\sigma^2 + 2b\epsilon} \right).$$

**证明** 对任意  $t > 0$ , 根据 Chernoff 方法有

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq e^{-nt\epsilon} E \left[ \exp \left( \sum_{i=1}^n (X_i - \mu) \right) \right] = e^{-nt\epsilon - n\mu t} (E[e^{tX_1}])^n$$

利用公式  $\ln z \leq z - 1$  有

$$\ln E[e^{tX_1}] \leq E[e^{tX}] - 1 = \sum_{m=1}^{\infty} E[X^m] \frac{t^m}{m!} \leq t\mu + \frac{t^2\sigma^2}{2} \sum_{m=2}^{\infty} (bt)^{m-2} = t\mu + \frac{t^2\sigma^2}{2(1-bt)}.$$

由此可得

$$\Pr \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \geq \epsilon \right] \leq \exp \left( -nt\epsilon + \frac{t^2n\sigma^2}{2(1-bt)} \right)$$

取  $t = \epsilon/(\sigma^2 + b\epsilon)$  完成证明.

**例 7.6** 给出 Bernstein 不等式的  $1 - \delta$  表述.

## 7.4 应用: 随机投影 (Random Projection)

设高维空间  $\mathbb{R}^d$  有  $n$  个点  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  ( $d$  非常大, 如 100 万或 1 亿). 处理这样一个高维的问题很难, 实际中的一种解决方案是能否找到一个保距变换:  $f: \mathbb{R}^d \rightarrow \mathbb{R}^k$  ( $k \ll d$ ), 使得以较大概率有

$$(1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 \leq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2.$$

随机投影广泛应用于高维的机器学习问题, 例如最近邻、 $k$ -近邻、降维、聚类等问题.

随机投影可以简单的表示为

$$f(\mathbf{x}) = \mathbf{x}P/c,$$

其中  $P$  是一个  $d \times k$  的随机矩阵, 其每个元素之间相互独立,  $c$  为一常数 (根据随机矩阵  $P$  确定). 下面介绍三种常见的随机矩阵:

- $P = (p_{ij})_{d \times k} \in \mathbb{R}^{d \times k}$ ,  $p_{ij} \sim \mathcal{N}(0, 1)$ , 此时  $c = \sqrt{k}$ ;
- $P = (p_{ij})_{d \times k} \in \{-1, 1\}^{d \times k}$ ,  $p_{ij}$  为 Rademacher 随机变量, 即  $\Pr(p_{ij} = 1) = \Pr(p_{ij} = -1) = 1/2$ , 此时  $c = \sqrt{k}$ ;
- $P = (p_{ij})_{d \times k} \in \{-1, 0, 1\}^{d \times k}$ , 满足  $\Pr(p_{ij} = 1) = \Pr(p_{ij} = -1) = 1/6$  和  $\Pr(p_{ij} = 0) = 2/3$ , 此时  $c = \sqrt{k/3}$ . 【主要用于 sparse 投影, 减少计算量】

下面我们重点理论分析 Gaussian 随机变量, 其它随机变量可参考相关资料, 对 Gaussian 随机变量, 这里介绍著名的 Johnson - Lindenstrauss 引理, 简称 JL 引理.

**引理 7.5** 设  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  为  $\mathbb{R}^d$  空间的  $n$  个点, 随机矩阵  $P = (p_{ij})_{d \times k} \in \mathbb{R}^{d \times k}$ ,  $p_{ij} \sim \mathcal{N}(0, 1)$  且每个元素相互独立, 令

$$\mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{x}_i P / \sqrt{k}, \quad i \in [n]$$

将  $d$  维空间中  $n$  个点  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  通过随机矩阵  $P$  投影到  $k$  维空间. 对任意  $\epsilon \in (0, 1/2)$ , 当  $k \geq 8 \log 2n / (\epsilon^2 - \epsilon^3)$  时至少以  $1/2$  的概率有

$$(1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \leq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \quad (i, j \in [n]).$$

**证明** 下面分三步证明 J-L 引理.

**第一步:** 对任意非零  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ , 首先证明

$$E \left[ \left\| \mathbf{x}P / \sqrt{k} \right\|_2^2 \right] = \|\mathbf{x}\|_2^2,$$

即在期望的情况下, 随机投影变换前后的点到原点的距离相同. 根据  $P = (p_{ij})_{d \times k}$  ( $p_{ij} \sim \mathcal{N}(0, 1)$ ) 有

$$\begin{aligned} E \left[ \left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \right] &= E \left[ \sum_{j=1}^k \left( \sum_{i=1}^d \frac{x_i p_{ij}}{\sqrt{k}} \right)^2 \right] = \sum_{j=1}^k \frac{1}{k} E \left[ \left( \sum_{i=1}^d x_i p_{ij} \right)^2 \right] \\ &= \sum_{j=1}^k \frac{1}{k} \sum_{i=1}^d x_i^2 = \frac{1}{k} \sum_{j=1}^k \|\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2. \end{aligned}$$

**第二步:** 对任意非零  $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ , 证明

$$\Pr \left[ \left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq \exp(-(\epsilon^2 - \epsilon^3)k/4).$$

将矩阵  $P$  表示为  $P = (P_1, P_2, \dots, P_k)$ , 其中  $P_i$  ( $i \in [d]$ ) 是一个  $d \times 1$  的列向量, 令  $v_j = \mathbf{x}P_j / \|\mathbf{x}\|_2$ , 即

$$(v_1, v_2, \dots, v_k) = \left( \frac{\mathbf{x}}{\|\mathbf{x}\|_2} P_1, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} P_2, \dots, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} P_k \right).$$

根据 Gaussian 分布的性质有  $v_j \sim \mathcal{N}(0, 1)$ , 且  $v_1, v_2, \dots, v_k$  是  $k$  个独立的随机变量. 对任意  $t \in (0, 1/2)$ , 根据 Chernoff 方法有

$$\begin{aligned} \Pr \left[ \left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] &= \Pr \left[ \left\| \frac{\mathbf{x}P}{\|\mathbf{x}\|_2} \right\|_2^2 \geq (1 + \epsilon)k \right] \\ &= \Pr \left[ \sum_{j=1}^k v_j^2 \geq (1 + \epsilon)k \right] \leq e^{-(1+\epsilon)kt} \left( E[e^{t \sum_{j=1}^k v_j^2}] \right)^k = e^{-(1+\epsilon)kt} \left( E[e^{tv_1^2}] \right)^k. \end{aligned}$$

对标准 Gaussian 分布有

$$E[e^{tv_1^2}] = \int_{-\infty}^{+\infty} \frac{e^{tu^2}}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \int_{-\infty}^{+\infty} \frac{e^{-\frac{u^2}{2}(1-2t)}}{\sqrt{2\pi}} du = \frac{1}{\sqrt{1-2t}},$$

代入可得

$$\Pr \left[ \left\| \mathbf{x}P/\sqrt{k} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq \left( \frac{e^{-2(1+\epsilon)t}}{1-2t} \right)^{k/2}.$$

上式右边对  $t$  求最小解得  $t_{\min} = \frac{\epsilon}{2(1+\epsilon)}$ , 代入可得

$$\Pr \left[ \left\| \mathbf{x}P/\sqrt{k} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq ((1 + \epsilon)e^{-\epsilon})^{k/2}.$$



设  $f(\epsilon) = \ln(1 + \epsilon)$ , 根据  $\epsilon \in (0, 1/2)$  有

$$f'(\epsilon) = \frac{1}{1 + \epsilon}, f''(\epsilon) = -\frac{1}{(1 + \epsilon)^2}, f'''(\epsilon) = \frac{2}{(1 + \epsilon)^3} \leq 2.$$

根据泰勒中值定理有

$$f(\epsilon) = f(0) + f'(0)\epsilon + \frac{f''(0)\epsilon^2}{2!} + \frac{f'''(\xi)\epsilon^3}{3!} \leq \epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^2}{3} \leq \epsilon - \frac{\epsilon^2 - \epsilon^3}{2}.$$

于是得到

$$\Pr \left[ \left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq e^{-k(\epsilon^2 - \epsilon^3)/4}.$$

同理可证

$$\Pr \left[ \left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \leq (1 - \epsilon) \|\mathbf{x}\|_2^2 \right] \leq e^{-k(\epsilon^2 - \epsilon^3)/4}.$$

**第三步:** 对任意给定  $i \neq j$ , 根据第二步的结论可知

$$\Pr[\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2] \leq e^{-k(\epsilon^2 - \epsilon^3)/4},$$

$$\Pr[\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \leq (1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2] \leq e^{-k(\epsilon^2 - \epsilon^3)/4}.$$

由于  $i, j \in [n]$ , 因此共有  $n(n - 1)$  对  $(i, j)$ , 根据 Union 不等式有

$$\Pr \left[ \exists i \neq j: \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \quad \text{或} \quad \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \leq (1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \right] \leq 2n^2 e^{-k(\epsilon^2 - \epsilon^3)/4},$$

设  $2n^2 e^{-k(\epsilon^2 - \epsilon^3)/4} \leq 1/2$ , 求解  $k \geq 8 \log 2n / (\epsilon^2 - \epsilon^3)$ . 引理得证.

## 习题

7.1 设随机变量  $X$  的期望  $E[X] = \mu > 0$ , 方差为  $\sigma^2$ , 证明对任意  $\epsilon > 0$  有

$$P(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

7.2 设随机变量  $X$  和  $Y$  满足  $E(X) = -2$ ,  $E(Y) = 2$ ,  $\text{Var}(X) = 1$ ,  $\text{Var}(Y) = 4$ ,  $\rho_{XY} = -1/2$ . 利用 Chebyshev 不等式估计  $\Pr(|X + Y| \geq 6)$  的上界.

7.3 独立同分布随机变量  $X_1, X_2, \dots, X_n$  满足  $E[X_i] = \mu$  和  $\text{Var}(X_i) \leq v$ . 证明对任意  $\epsilon > 0$  有

$$\left[ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right] \leq \frac{v}{n\epsilon^2}.$$

7.4 阐述什么是 chernoff 方法。

7.5 随机变量  $X_1, X_2, \dots, X_n$  相互独立且满足  $X_i \sim \text{Ber}(p_i)$  ( $p_i > 0$ ). 利用 chernoff 方法给出下列概率的上界

$$P \left[ \frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \geq \epsilon \right] \quad \text{和} \quad P \left[ \frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \leq -\epsilon \right].$$

7.6 若独立同分布随机变量  $X_1, X_2, \dots, X_n$  满足  $X_i \in \{a, b\}$  ( $b > a$ ) 且  $P(X_i = a) = P(X_i = b) = 1/2$ . 求下列概率的上界

$$P \left[ \frac{1}{n} \sum_{i=1}^n \left( X_i - \frac{a+b}{2} \right) \geq \epsilon \right] \quad \text{和} \quad \Pr \left[ \frac{1}{n} \sum_{i=1}^n \left( X_i - \frac{a+b}{2} \right) \leq -\epsilon \right].$$

7.7 随机变量  $X_1, X_2, \dots, X_n$  相互独立且满足  $X_i \sim \text{Ber}(p_i)$  ( $p_i > 0$ ). 证明对任意  $0 < \epsilon < 1$  有不等式

$$P \left[ \sum_{i=1}^n X_i \geq (1 + \epsilon) \sum_{i=1}^n p_i \right] \leq e^{-\mu\epsilon^2/3}.$$

7.8 随机变量  $X \in [a, b]$  且期望  $\mu = \mathbb{E}[x]$ , 证明对任意  $t > 0$  有

$$\mathbb{E} [e^{tx}] \leq \exp (\mu t + t^2(b-a)^2/8)$$

7.9 利用 chernoff 方法证明: 设  $X_1, X_2, \dots, X_k$  是  $k$  个独立的随机变量, 且  $X_i \sim N(0, 1)$ , 则有

$$\Pr \left( \sum_{i=1}^k X_i^2 \geq (1 + \epsilon)k \right) \leq \exp (-k(\epsilon^2 - \epsilon^3)/4)$$

7.10 证明 Bennet 不等式.

**7.11** 证明 Bernstein 不等式.

**7.12** 已知 Bernstein 不等式

$$P \left[ \frac{1}{n} \sum_{i=1}^n (x_i - \mu) \geq \epsilon \right] \leq \exp \left( \frac{-n\epsilon^2}{2\sigma^2 + 2b\epsilon} \right),$$

给出其等价  $1 - \delta$  描述。

**7.13** 已知独立同分布随机变量  $X_1, X_2, \dots, X_n$  满足  $X_i \sim N(\mu, \sigma^2)$ , 给出  $E[\max_{i \in [n]} \{X_i\}]$  的上界, 并给出严格证明。

**7.14** 假设训练数据集  $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$  根据分布  $\mathcal{D}$  独立采样所得, 分类器  $f$  在训练集  $S_n$  的错误率为零 (全部预测正确), 求分类器  $f$  在分布  $\mathcal{D}$  上的错误率介于 0 和  $\epsilon$  之间的概率 ( $\epsilon > 0$ ).

**7.15** 假设训练数据集  $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$  根据分布  $\mathcal{D}$  独立采样所得, 分类器  $f$  在训练集  $S_n$  的错误率为  $\hat{p}$ , 求分类器  $f$  在分布  $\mathcal{D}$  上的错误率介于  $\hat{p} - \epsilon$  和  $\hat{p} + \epsilon$  之间的概率 ( $\epsilon > 0$ ).



## 第8章 大数定律及中心极限定理

### 8.1 大数定律

给定随机变量  $X_1, X_2, \dots, X_n$ , 这些随机变量的均值 (算术平均值) 为

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

当  $n$  非常大时, 大数定律考虑随机变量的均值是否具有稳定性.

**定义 8.1 (依概率收敛)** 设  $X_1, X_2, \dots, X_n, \dots$  是一随机变量序列,  $a$  是一常数, 如果对任意  $\epsilon > 0$  有

$$\lim_{n \rightarrow \infty} \Pr\{|X_n - a| < \epsilon\} = 1 \quad \text{或} \quad \lim_{n \rightarrow \infty} \Pr\{|X_n - a| > \epsilon\} = 0,$$

则称随机变量序列  $X_1, X_2, \dots, X_n, \dots$  依概率收敛于  $a$ , 记  $X_n \xrightarrow{P} a$ .

问题: 与数列极限的区别? 下面我们给出依概率的性质:

- 1) 若  $X_n \xrightarrow{P} a$  且函数  $g: \mathbb{R} \rightarrow \mathbb{R}$  在  $X = a$  点连续, 则  $g(X_n) \xrightarrow{P} g(a)$ .
- 2) 若  $X_n \xrightarrow{P} a, Y_n \xrightarrow{P} b$ , 函数  $g: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  在点  $(X, Y) = (a, b)$  处连续, 则  $g(X_n, Y_n) \xrightarrow{P} g(a, b)$ .

例如: 如果  $X_n \xrightarrow{P} a$  和  $Y_n \xrightarrow{P} b$ , 那么  $X_n + Y_n \xrightarrow{P} a + b$  和  $X_n Y_n \xrightarrow{P} ab$ .

**定理 8.1 (大数定律)** 若随机变量序列  $X_1, X_2, \dots, X_n, \dots$  满足

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \frac{1}{n} \sum_{i=1}^n E[X_i],$$

则称  $\{X_n\}$  服从大数定律.

大数定理刻画了随机变量的均值 (算术平均值) 依概率收敛于期望的均值 (算术平均值). 下面介绍几种大数定律:

**定理 8.2 (马尔可夫 Markov 大数定律)** 如果随机变量序列  $X_1, X_2, \dots, X_n, \dots$  满足

$$\frac{1}{n^2} \text{Var} \left( \sum_{i=1}^n X_i \right) \rightarrow 0 \quad n \rightarrow \infty,$$

则  $\{X_n\}$  服从大数定理.

马尔可夫大数定律不要求随机变量序列  $X_1, X_2, \dots, X_n, \dots$  相互独立或同分布, 其证明直接通过 Chebyshev 不等式有

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \right| \geq \epsilon \right] \leq \frac{1}{n^2 \epsilon^2} \text{Var} \left( \sum_{i=1}^n X_i \right) \rightarrow 0 \quad n \rightarrow \infty.$$

**定理 8.3 (切比雪夫 Chebyshev 大数定律)** 设随机变量序列  $X_1, X_2, \dots, X_n, \dots$  相互独立, 且存在常数  $c > 0$  使得  $\text{Var}(X_n) \leq c$ , 则  $\{X_n\}$  服从大数定律.

此处独立的随机变量可以修改为‘不相关随机变量’. 证明直接通过切比雪夫不等式

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \right| \geq \epsilon \right] \leq \frac{1}{\epsilon^2 n^2} \text{Var} \left( \sum_{i=1}^n X_i \right) \leq \frac{c}{n \epsilon^2} \rightarrow 0 \quad n \rightarrow \infty.$$

**定理 8.4 (辛钦 Khintchine 大数定律)** 设  $X_1, X_2, \dots, X_n, \dots$  为独立同分布随机变量序列, 且每个随机变量的期望  $E[X_i] = \mu$  存在, 则  $\{X_n\}$  服从大数定律.

辛钦大数定律不要求方差一定存在, 其证明超出了本书范围.

**定理 8.5 (Bernoulli 大数定律)** 设随机变量序列  $X_n \sim B(n, p)$  ( $p > 0$ ), 对任意  $\epsilon > 0$  有

$$\lim_{n \rightarrow \infty} \Pr \left[ \left| \frac{X_n}{n} - p \right| \geq \epsilon \right] = 0,$$

即  $X_n/n \xrightarrow{P} p$ .

定理的证明依据二项分布的性质: 独立同分布随机变量  $Y_1, Y_2, \dots, Y_n$  满足  $Y_i \sim \text{Ber}(p)$ , 则

$$X_n = \sum_{i=1}^n Y_i \sim B(n, p).$$

于是得到

$$\lim_{n \rightarrow \infty} \Pr \left[ \left| \frac{X_n}{n} - p \right| \geq \epsilon \right] = \lim_{n \rightarrow \infty} \Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n Y_i - E[Y_i] \right| \geq \epsilon \right] \leq \frac{1}{\epsilon^2 n^2} \text{Var} \left( \sum_{i=1}^n Y_i \right) = \frac{p(1-p)}{\epsilon^2 n} \rightarrow 0.$$

如何判断随机变量序列  $X_1, X_2, \dots, X_n, \dots$  满足大数定律:

- 若随机变量独立同分布, 则利用辛钦大数定律查看期望是否存在;
- 对非独立同分布随机变量, 则利用 Markov 大数定律判断方差是否趋于零.

**例 8.1** 独立的随机变量序列  $X_1, X_2, \dots, X_n, \dots$  满足  $\Pr\{X_n = n^{1/4}\} = \Pr\{X_n = -n^{1/4}\} = 1/2$ . 证明  $\{X_n\}$  服从大数定律.

**证明** 根据题意可得  $E[X_i] = 0$ , 以及  $\text{Var}(X_i) = E[X_i^2] = i^{1/2}$ , 根据 Chebysheve 不等式和独立性有

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n X_i \right| \geq \epsilon \right] \leq \frac{1}{n^2 \epsilon^2} \text{Var} \left( \sum_{i=1}^n X_i \right) = \frac{1}{n^2 \epsilon^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{\epsilon^2} \frac{1}{n^2} \sum_{i=1}^n i^{1/2} \leq \frac{1}{\epsilon^2 \sqrt{n}}$$

再根据

$$\sum_{i=1}^n i^{1/2} \leq \sum_{i=1}^n \int_i^{i+1} i^{1/2} dx \leq \sum_{i=1}^n \int_i^{i+1} x^{1/2} dx = \int_1^{n+1} x^{1/2} dx = 2((n+1)^{3/2} - 1)/3$$

由此可得当  $n \rightarrow +\infty$  时有

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n X_i \right| \geq \epsilon \right] \leq \frac{2((n+1)^{3/2} - 1)/3}{\epsilon^2 n^2} \rightarrow 0$$

大数定律小结:

- Markov 大数定律: 若随机变量序列  $\{X_i\}$  满足  $\text{Var}(\sum_{i=1}^n X_i)/n^2 \rightarrow 0$ , 则满足大数定律;
- Chebyshev 大数定律: 若独立随机变量序列  $\{X_i\}$  满足  $\text{Var}(X_i) \leq c$ , 则满足大数定律;
- Khintchine 大数定律: 若独立同分布随机变量序列  $\{X_i\}$  期望存在, 则满足大数定律;
- Bernoulli 大数定律: 对二项分布  $X_n \sim B(n, p)$ , 有  $X_n/n \xrightarrow{P} p$ .

## 8.2 中心极限定理

对独立的随机变量序列  $X_1, X_2, \dots, X_n, \dots$ , 我们考虑标准化后随机变量

$$Y_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n E(X_i)}{\sqrt{\text{Var}(\sum_{i=1}^n X_i)}}$$

的极限分布是否为服从正态分布. 首先介绍依分布收敛.

**定义 8.2** 设随机变量  $Y$  的分布函数为  $F_Y(y) = \Pr(Y \leq y)$ , 以及随机变量序列  $Y_1, Y_2, \dots, Y_n, \dots$  的分布函数分别为  $F_{Y_n}(y) = \Pr(Y_n \leq y)$ , 如果

$$\lim_{n \rightarrow \infty} \Pr[Y_n \leq y] = \Pr[Y \leq y], \quad \text{即} \quad \lim_{n \rightarrow \infty} F_{Y_n}(y) = F_Y(y),$$

则称随机变量序列  $Y_1, Y_2, \dots, Y_n, \dots$  依分布收敛于  $Y$ , 记  $Y_n \xrightarrow{d} Y$ .

下面介绍独立同分布中心极限定理, 又被称为林德贝格-勒维 (Lindeberg-Lévy) 中心极限定理”:

**定理 8.6** 设独立同分布的随机变量  $X_1, X_2, \dots, X_n, \dots$  的期望  $E(X_1) = \mu$  和方差  $\text{Var}(X_1) = \sigma^2$ , 则

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

前面介绍标准正态分布的分布函数为  $\Phi(x)$ , 则上述中心极限定理等价于

$$\lim_{n \rightarrow \infty} \Pr[Y_n \leq y] = \Phi(y).$$

随机变量  $Y_n$  是随机变量  $X_1, X_2, \dots, X_n$  的标准化, 其极限服从标准正态分布. 当  $n$  足够大时近似有  $Y_n \sim \mathcal{N}(0, 1)$ , 中心极限定理的变形公式为

$$\sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2), \quad \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

大数定律给出了当  $n \rightarrow \infty$  时随机变量平均值  $\frac{1}{n} \sum_{i=1}^n X_i$  的趋势, 而中心极限定理给出了  $\frac{1}{n} \sum_{i=1}^n X_i$  的具体分布.

**例 8.2** 设一电压接收器同时接收到 20 个独立同分布的信号电压  $V_k$  ( $k \in [20]$ ), 且  $V_k \sim U(0, 10)$ , 求电压和大于 105 的概率.

**解** 根据题意可知独立同分布的随机变量  $V_1, V_2, \dots, V_{20}$  服从均匀分布  $U(0, 10)$ , 于是有  $E(V_k) = 5$  和  $\text{Var}(V_k) = 100/12 = 25/3$ . 设  $V = \sum_{k=1}^{20} V_k$ , 则有

$$E(V) = 100 \quad \text{Var}(V) = 500/3.$$

根据中心极限定理近似有

$$\frac{V - E(V)}{\sqrt{\text{Var}(V)}} = \frac{V - 100}{\sqrt{500/3}} \sim \mathcal{N}(0, 1).$$

根据标准正态分布的分布函数  $\Phi(x)$  有

$$\Pr(V \geq 105) = \Pr\left(\frac{V - 100}{\sqrt{500/3}} \geq \frac{105 - 100}{\sqrt{500/3}}\right) = \Pr\left(\frac{V - 100}{\sqrt{500/3}} \geq 0.387\right) = 1 - \Phi(0.387).$$

查表完成证明.

**例 8.3** 某产品装箱, 每箱重量是随机的, 假设其期望是 50 公斤, 标准差为 5 公斤. 若最大载重量为 5 吨, 问每车最多可装多少箱能以 0.997 以上的概率保证不超载?

**解** 假设最多可装  $n$  箱不超重, 用  $X_i$  表示第  $i$  箱重量 ( $i \in [n]$ ), 有  $E(X_i) = 50$  和  $\text{Var}(X_i) = 25$ . 设总重量  $X = \sum_{i=1}^n X_i$ , 则有  $E(X) = 50n$  和  $\text{Var}(X) = 25n$ . 由中心极限定理近似有

$$(X - 50n)/\sqrt{25n} \sim \mathcal{N}(0, 1).$$



根据标准正态分布的分布函数  $\Phi(x)$  有

$$\Pr(X \leq 5000) = \Pr\left(\frac{X - 50n}{\sqrt{25n}} \leq \frac{5000 - 50n}{\sqrt{25n}}\right) = \Phi\left(\frac{5000 - 50n}{\sqrt{25n}}\right) > 0.977 = \Phi(2).$$

根据分布函数的单调性有

$$\frac{1000 - 10n}{\sqrt{n}} > 2 \implies 1000n^2 - 2000n + 1000^2 > 4n.$$

求解可得  $n > 102.02$  或  $n < 98.02$ , 根据由题意可知  $n = 98$ .

下面介绍另一个中心极限定理: 棣莫弗-拉普拉斯 (De Moivre-Laplace) 中心极限定理:

**推论 8.1** 设随机变量  $X_n \sim B(n, p)$ , 则

$$Y_n = \frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

由此中心极限定理可知: 当  $n$  非常大时随机变量  $X_n \sim B(n, p)$  满足  $X_n \overset{\text{近似}}{\sim} \mathcal{N}(np, np(1-p))$ , 从而有如下近似估计:

$$\Pr[X_n \leq y] = \Pr\left[\frac{X_n - np}{\sqrt{np(1-p)}} \leq \frac{y - np}{\sqrt{np(1-p)}}\right] \approx \Phi\left(\frac{y - np}{\sqrt{np(1-p)}}\right).$$

针对上式, 可以考虑三种问题: i) 已知  $n$  和  $\Pr[X_n \leq y]$ , 求  $y$ ; ii) 已知  $n$  和  $y$ , 求  $\Pr[X_n \leq y]$ ; iii) 已知  $y$  和  $\Pr[X_n \leq y]$ , 求  $n$ . 下面看三个例子:

**例 8.4** 车间有 200 台独立工作的车床, 每台工作的概率为 0.6, 工作时每台耗电 1 千瓦, 至少供电多少千瓦才能以 99.9% 的概率保证正常生产.

**解** 设工作的车床数为  $X$ , 则  $X \sim B(200, 0.6)$ . 设至少供电  $y$  千瓦. 根据棣莫弗-拉普拉斯中心定理近似有  $X \sim \mathcal{N}(120, 48)$ , 进一步有

$$\Pr(X \leq y) \geq 0.999 \implies \Pr\left(\frac{X - 120}{\sqrt{48}} \leq \frac{y - 120}{\sqrt{48}}\right) \approx \Phi\left(\frac{y - 120}{\sqrt{48}}\right) \geq 0.999 = \Phi(3.1).$$

所以有  $\frac{y-120}{\sqrt{48}} \geq 3.1$ , 求解可得  $y \geq 141$ .

**例 8.5** 系统由 100 个相互独立的部件组成, 每部件损坏率为 0.1, 至少 85 个部件正常工作系统才能运行, 求系统运行的概率.

**解** 设  $X$  是损坏的部件数, 则  $X \sim B(100, 0.1)$ , 有  $E(X) = 10$  和  $\text{Var}(X) = 9$ . 根据棣莫弗-拉普拉斯中心定理近似有  $X \sim \mathcal{N}(10, 9)$ , 求系统运行的概率为

$$\Pr(X \leq 15) = \Pr\left(\frac{X - 10}{\sqrt{9}} \leq \frac{15 - 10}{\sqrt{9}}\right) \approx \Phi(5/3).$$

**例 8.6** 一次电视节目调查中调查  $n$  人, 其中  $k$  人观看了电视节目, 因此收看比例  $k/n$  作为电视节目收视率  $p$  的估计, 要以 90% 的概率有  $|k/n - p| \leq 0.05$  成立, 需要调查多少对象?

**解** 用  $X_n$  表示  $n$  个调查对象中收看节目的人数, 则有  $X_n \sim B(n, p)$ . 根据棣莫弗-拉普拉斯中心定理近似有  $(X_n - np)/\sqrt{np(1-p)} \sim \mathcal{N}(0, 1)$ , 进一步有

$$\begin{aligned} \Pr\left[\left|\frac{X_n}{n} - p\right| \leq 0.05\right] &= \Pr\left[\frac{|X_n - np|}{n} \leq 0.05\right] = \Pr\left[\frac{|X_n - np|}{\sqrt{np(1-p)}} \leq \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right] \\ &= \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) \end{aligned}$$

对于标准正太分布函数有  $\Phi(-\alpha) = 1 - \Phi(\alpha)$  以及  $p(1-p) \leq 1/4$ , 于是有

$$\Pr\left[\left|\frac{X_n}{n} - p\right| \leq 0.05\right] = 2\Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 > 2\Phi(\sqrt{n}/10) - 1 > 0.9.$$

所以  $\Phi(\sqrt{n}/10) \geq 0.95$ , 查表解得  $n \geq 271$ .

对独立不同分布的随机变量序列, 有李雅普诺夫 (Lyapunov) 中心极限定理:

**定理 8.7** 设独立随机变量  $X_1, X_2, \dots, X_n, \dots$  的期望  $E[X_n] = \mu_n$  和方差  $\text{Var}(X_n) = \sigma_n^2 > 0$ . 记  $B_n^2 = \sum_{k=1}^n \sigma_k^2$ , 若存在  $\delta > 0$ , 当  $n \rightarrow \infty$  时有

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E[|X_k - \mu_k|^{2+\delta}] \rightarrow 0$$

成立, 则有

$$Y_n = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n E(X_k)}{\sqrt{\text{Var}(\sum_{k=1}^n X_k)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

中心极限定理小结:

- 独立同分布中心极限定理: 若  $E[X_k] = \mu$  和  $\text{Var}(X_k) = \sigma^2$ , 则  $\sum_{k=1}^n X_k \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2)$ ;
- 棣莫弗-拉普拉斯中心极限定理: 若  $X_k \sim B(k, p)$ , 则  $X_k \xrightarrow{d} \mathcal{N}(np, np(1-p))$ ;
- 独立不同分布中心极限定理: 李雅普诺夫定理.

## 习题

- 8.1 设随机变量  $X$  的期望  $E[X] = \mu > 0$ , 方差为  $\sigma^2$ , 证明对任意  $\epsilon > 0$  有

$$P(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

- 8.2 设随机变量  $X$  和  $Y$  满足  $E(X) = -2$ ,  $E(Y) = 2$ ,  $\text{Var}(X) = 1$ ,  $\text{Var}(Y) = 4$ ,  $\rho_{XY} = -1/2$ . 利用Chebyshev不等式估计  $\Pr(|X + Y| \geq 6)$  的上界.

- 8.3 独立同分布随机变量  $X_1, X_2, \dots, X_n$  满足  $E[X_i] = \mu$  和  $\text{Var}(X_i) \leq v$ . 证明对任意  $\epsilon > 0$  有

$$\left[ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right] \leq \frac{v}{n\epsilon^2}.$$

- 8.4 阐述什么是chernoff方法。

- 8.5 随机变量  $X_1, X_2, \dots, X_n$  相互独立且满足  $X_i \sim \text{Ber}(p_i)$  ( $p_i > 0$ ). 利用chernoff方法给出下列概率的上界

$$P \left[ \frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \geq \epsilon \right] \quad \text{和} \quad P \left[ \frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \leq -\epsilon \right].$$

- 8.6 若独立同分布随机变量  $X_1, X_2, \dots, X_n$  满足  $X_i \in \{a, b\}$  ( $b > a$ ) 且  $P(X_i = a) = P(X_i = b) = 1/2$ . 求下列概率的上界

$$P \left[ \frac{1}{n} \sum_{i=1}^n \left( X_i - \frac{a+b}{2} \right) \geq \epsilon \right] \quad \text{和} \quad \Pr \left[ \frac{1}{n} \sum_{i=1}^n \left( X_i - \frac{a+b}{2} \right) \leq -\epsilon \right].$$

- 8.7 随机变量  $X_1, X_2, \dots, X_n$  相互独立且满足  $X_i \sim \text{Ber}(p_i)$  ( $p_i > 0$ ). 证明对任意  $0 < \epsilon < 1$  有不等式

$$P \left[ \sum_{i=1}^n X_i \geq (1 + \epsilon) \sum_{i=1}^n p_i \right] \leq e^{-\mu\epsilon^2/3}.$$

- 8.8 随机变量  $X \in [a, b]$  且期望  $\mu = \mathbb{E}[x]$ , 证明对任意  $t > 0$  有

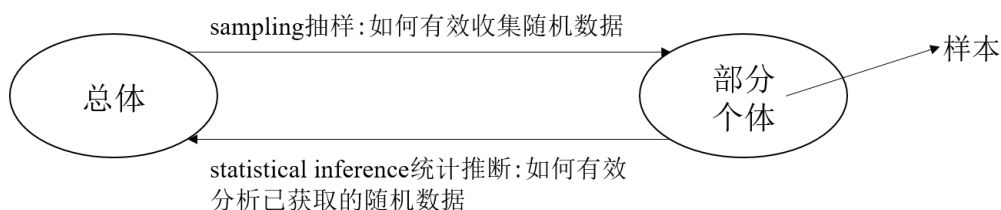
$$\mathbb{E}[e^{tx}] \leq \exp(\mu t + t^2(b-a)^2/8)$$



## 第9章 统计的基本概念

到 19 世纪末 20 世纪初, 随着近代数学和概率论的发展, 诞生了统计学.

统计学: 以概率论为基础, 研究如何有效收集研究对象的随机数据, 以及如何运用所获得的数据揭示统计规律的一门学科. 统计学的研究内容具体包括: 抽样、参数估计、假设检验等.



### 9.1 总体 (population) 与样本 (sample)

‘总体’是研究问题所涉及的对象全体; 总体中每个元素称为‘个体’. 总体分为有限或无限总体. 例如: 全国人民的收入是总体, 一个人的收入是个体.

在研究总体时, 通常关心总体的某项或某些数量指标, 总体中的每个个体是随机试验的一个观察值, 即随机变量  $X$  的值. 对总体的研究可转化为对随机变量  $X$  的分布或数字特征的研究, 后面总体与随机变量  $X$  的分布不再区分, 简称总体  $X$ .

总体: 研究对象的全体  $\Rightarrow$  数据  $\Rightarrow$  随机变量 (分布未知).

样本: 从总体中随机抽取一些个体, 一般表示为  $X_1, X_2, \dots, X_n$ , 称  $X_1, X_2, \dots, X_n$  为取自总体  $X$  的随机样本, 其样本容量为  $n$ .

抽样: 抽取样本的过程.

样本值: 观察样本得到的数值, 例如:  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  为样本观察值或样本值.

样本的二重性: i) 就一次具体观察而言, 样本值是确定的数; ii) 不同的抽样下, 样本值会发生变化, 可看作随机变量.

**定义 9.1 (简单随机样本)** 称样本  $X_1, X_2, \dots, X_n$  是总体  $X$  的简单随机样本, 简称样本, 是指样本满足: 1) 代表性, 即  $X_i$  与  $X$  同分布; 2) 独立性, 即  $X_1, X_2, \dots, X_n$  之间相互独立.

本书后面所考虑的样本均为简单随机样本.

设总体  $X$  的联合分布函数为  $F(x)$ , 则  $X_1, X_2, \dots, X_n$  的联合分布函数为

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

若总体  $X$  的概率密度为  $f(x)$ , 则样本  $X_1, X_2, \dots, X_n$  的联合概率密度为

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

若总体  $X$  的分布列  $\Pr(X = x_i)$ , 则样本  $X_1, X_2, \dots, X_n$  的联合分布列为

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

## 9.2 常用统计量

为研究样本的特性, 我们引入统计量:

**定义 9.2** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本,  $g(X_1, X_2, \dots, X_n)$  是关于  $X_1, X_2, \dots, X_n$  的一个连续、且不含任意参数的函数, 称  $g(X_1, X_2, \dots, X_n)$  是一个 **统计量**.

由于  $X_1, X_2, \dots, X_n$  是随机变量, 因此统计量  $g(X_1, X_2, \dots, X_n)$  是一个随机变量. 而  $g(x_1, x_2, \dots, x_n)$  为  $g(X_1, X_2, \dots, X_n)$  的一次观察值. 下面研究一些常用统计量.

假设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 定义 **样本均值** 为

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

根据样本的独立同分布性质有

**引理 9.1** 设总体  $X$  的期望为  $E[X] = \mu$ , 方差  $\text{Var}(X) = \sigma^2$ , 则有

$$E[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \sigma^2/n, \quad \bar{X} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

假设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 定义 **样本方差** 为

$$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2.$$

**引理 9.2** 设总体  $X$  的期望为  $E[X] = \mu$ , 方差  $\text{Var}(X) = \sigma^2$ , 则有

$$E[S_0^2] = \frac{n-1}{n} \sigma^2.$$

**证明** 根据  $E[X_i^2] = \sigma^2 + \mu^2$  有

$$E(\bar{X}^2) = E \left[ \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^2 \right] = \frac{1}{n^2} E \left[ \left( \sum_{i=1}^n X_i \right)^2 \right] = \frac{1}{n^2} E \left[ \sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j \right] = \frac{\sigma^2}{n} + \mu^2,$$

于是有

$$E(S_0^2) = E(X_i^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

由此可知样本方差  $S_0^2$  与总体方差  $\sigma^2$  之间存在偏差.

进一步定义 **样本标准差** 为:

$$S_0 = \sqrt{S_0^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

定义 **修正后的样本方差** 为:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{即} \quad S^2 = \frac{n}{n-1} S_0^2,$$

**引理 9.3** 设总体  $X$  的期望为  $E[X] = \mu$ , 方差  $\text{Var}(X) = \sigma^2$ , 则有

$$E[S^2] = \sigma^2.$$

**证明** 根据期望的性质有

$$E[S^2] = E\left[\frac{n}{n-1} S_0^2\right] = \frac{n}{n-1} E[S_0^2] = \sigma^2.$$

假设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 定义 **样本  $k$  阶原点矩** 为:

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots.$$

定义 **样本  $k$  阶中心矩** 为:

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

**例 9.1** 设总体  $X \sim \mathcal{N}(20, 3)$ , 从总体中抽取两独立样本, 容量分别为 10 和 15. 求这两个样本均值之差的绝对值大于 0.3 的概率.

**解** 设  $X_1, X_2, \dots, X_{10}$  和  $X'_1, X'_2, \dots, X'_{15}$  分别为来自总体  $X \sim \mathcal{N}(20, 3)$  的两个独立样本. 根据正态分布的性质有

$$\bar{X}_1 = \frac{1}{10} \sum_{i=1}^{10} X_i \sim \mathcal{N}(20, 3/10), \quad \bar{X}_2 = \frac{1}{15} \sum_{i=1}^{15} X'_i \sim \mathcal{N}(20, 1/5).$$

进一步根据正态分布的性质有  $\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}(0, 1/2)$ , 于是可得

$$\Pr(|\bar{X}_1 - \bar{X}_2| > 0.3) = 2 - 2\Phi(0.3/\sqrt{1/2}).$$

假设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 定义 **最小次序统计量** 和 **最大次序统计量** 分别为:

$$X_{(1)} = \min\{X_1, X_2, \dots, X_n\} \quad \text{和} \quad X_{(n)} = \max\{X_1, X_2, \dots, X_n\},$$

以及定义 **样本极差** 为

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

设总体  $X$  的分布函数为  $F(x)$ , 则有

$$F_{X_{(1)}}(x) = \Pr(X_{(1)} \leq x) = 1 - \Pr(X_{(1)} > x) = 1 - (1 - F(x))^n, \quad F_{X_{(n)}}(x) = F^n(x).$$

**定理 9.1** 设总体  $X$  的密度函数为  $f(x)$ , 分布函数为  $F(x)$ ,  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本, 则第  $k$  次序统计量  $X_{(k)}$  的分布函数和密度函数分别为

$$\begin{aligned} F_k(x) &= \sum_{r=k}^n \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r} \\ f_k(x) &= \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x). \end{aligned}$$

**证明** 根据题意有第  $k$  次序统计量  $X_{(k)}$  的分布函数为

$$\begin{aligned} F_k(x) &= \Pr[X_{(k)} \leq x] = \Pr[X_1, X_2, \dots, X_n \text{ 中至少有 } k \text{ 个随机变量 } \leq x] \\ &= \sum_{r=k}^n \Pr[X_1, X_2, \dots, X_n \text{ 中恰有 } r \text{ 个随机变量 } \leq x, n-r \text{ 个随机变量 } > x] \\ &= \sum_{r=k}^n \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r}. \end{aligned}$$

利用恒等式

$$\sum_{r=k}^n \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{(k-1)!(n-k)!} \int_0^p t^{k-1} (1-t)^{n-k} dt \quad (r \in [n], p \in [0, 1])$$

由此可知

$$F_k(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} dt,$$

根据积分函数求导完成证明.



### 9.3 Beta 分布、 $\Gamma$ 分布、Dirichlet 分布

首先介绍两积分函数.

**定义 9.3 (Beta-函数)** 对任意给定  $\alpha_1 > 0$  和  $\alpha_2 > 0$ , 定义 Beta 函数为

$$\text{Beta}(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1-1} (1-x)^{\alpha_2-1} dx,$$

有些书简记为  $B(\alpha_1, \alpha_2)$ , 被称为第一类欧拉积分函数.

根据数学分析可知  $\text{Beta}(\alpha_1, \alpha_2)$  在定义域  $(0, +\infty) \times (0, +\infty)$  连续. 利用变量替换  $t = 1 - x$ , 根据定义有

$$\begin{aligned} \text{Beta}(\alpha_1, \alpha_2) &= \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1} dt = \int_1^0 (1-x)^{\alpha_1-1} x^{\alpha_2-1} d(1-x) \\ &= \int_0^1 x^{\alpha_2-1} (1-x)^{\alpha_1-1} dx = \text{Beta}(\alpha_2, \alpha_1), \end{aligned}$$

由此可知 Beta 函数的对称性:  $\text{Beta}(\alpha_1, \alpha_2) = \text{Beta}(\alpha_2, \alpha_1)$ .

**定义 9.4 ( $\Gamma$ -函数)** 对任意给定  $\alpha > 0$ , 定义  $\Gamma$ -函数为

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx,$$

又被称为第二类欧拉积分函数.

**性质 9.1** 对  $\Gamma$ -函数, 有  $\Gamma(1) = 1$  和  $\Gamma(1/2) = \sqrt{\pi}$ , 以及对  $\alpha > 1$  有  $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ .

**证明** 根据定义有

$$\Gamma(1) = \int_0^{+\infty} e^{-x} dx = 1.$$

利用变量替换  $x = t^{1/2}$  有

$$\Gamma(1/2) = \int_0^{+\infty} t^{-\frac{1}{2}} e^{-t} dt = \int_0^{+\infty} x^{-1} e^{-x^2} dx^2 = 2 \int_0^{+\infty} e^{-x^2} dx = \int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}.$$

进一步有

$$\Gamma(\alpha) = - \int_0^{\infty} x^{\alpha-1} de^{-x} = -[x^{\alpha-1} e^{-x}]_0^{+\infty} + (\alpha-1) \int_0^{+\infty} x^{\alpha-2} e^{-x} dx = (\alpha-1)\Gamma(\alpha-1)$$

对任意正整数  $n$ , 根据上面的性质有

$$\Gamma(n) = (n-1)!$$

关于 Beta 函数和  $\Gamma$ -函数, 有如下关系:

**定理 9.2** 对任意给定  $\alpha_1 > 0$  和  $\alpha_2 > 0$ , 有

$$\text{Beta}(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}.$$

**证明** 根据  $\Gamma$ -函数的定义有

$$\Gamma(\alpha_1)\Gamma(\alpha_2) = \int_0^{+\infty} t^{\alpha_1-1} e^{-t} dt \int_0^{+\infty} s^{\alpha_2-1} e^{-s} ds = \int_0^{+\infty} \int_0^{+\infty} e^{-(t+s)} t^{\alpha_1-1} s^{\alpha_2-1} dt ds.$$

引入变量替换  $x = t + s$  和  $y = t/(t + s)$ , 反解可得  $t = xy$  和  $s = x - xy$ , 计算雅可比行列式有

$$\begin{vmatrix} \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \\ \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \end{vmatrix} = \begin{vmatrix} y & x \\ 1-y & -x \end{vmatrix} = -x.$$

同时有  $x \in (0, +\infty)$  和  $y \in (0, 1)$  成立, 由此可得

$$\begin{aligned} \Gamma(\alpha_1)\Gamma(\alpha_2) &= \int_0^1 \int_0^{+\infty} e^{-x} x^{\alpha_1-1} y^{\alpha_1-1} x^{\alpha_2-1} (1-y)^{\alpha_2-1} |x| dx dy \\ &= \int_0^1 \int_0^{+\infty} e^{-x} x^{\alpha_1+\alpha_2-1} y^{\alpha_1-1} (1-y)^{\alpha_2-1} dx dy \\ &= \int_0^{+\infty} e^{-x} x^{\alpha_1+\alpha_2-1} dx \int_0^1 y^{\alpha_1-1} (1-y)^{\alpha_2-1} dy \\ &= \Gamma(\alpha_1 + \alpha_2) \text{Beta}(\alpha_1, \alpha_2) \end{aligned}$$

定理得证.

根据上述定理可知

**推论 9.1** 对任意  $\alpha_1 > 1$  和  $\alpha_2 > 0$ , 有

$$\text{Beta}(\alpha_1, \alpha_2) = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 1} \text{Beta}(\alpha_1 - 1, \alpha_2).$$

**证明** 根据前面的定理有

$$\text{Beta}(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} = \frac{(\alpha_1 - 1)\Gamma(\alpha_1 - 1)\Gamma(\alpha_2)}{(\alpha_1 + \alpha_2 - 1)\Gamma(\alpha_1 + \alpha_2 - 1)} = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 1} \text{Beta}(\alpha_1 - 1, \alpha_2).$$

**定义 9.5** 对任意  $\alpha_1, \alpha_2, \dots, \alpha_k > 0$ , 定义多维 Beta 函数为

$$\text{Beta}(\alpha_1, \alpha_2, \dots, \alpha_k) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}.$$

下面介绍三种分布:

**定义 9.6 (Beta 分布)** 给定  $\alpha_1 > 0$  和  $\alpha_2 > 0$ , 若随机变量  $X$  的概率密度为

$$f(x) = \begin{cases} \frac{x^{\alpha_1-1}(1-x)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)} & x \in (0, 1) \\ 0 & \text{其它.} \end{cases}$$

称  $X$  服从参数为  $\alpha_1$  和  $\alpha_2$  的 Beta 分布, 记  $X \sim B(\alpha_1, \alpha_2)$ .

**定理 9.3** 若随机变量  $X \sim B(\alpha_1, \alpha_2)$ , 则有

$$E[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2} \quad \text{和} \quad \text{Var}(X) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)}.$$

**证明** 根据期望的定义有

$$\begin{aligned} E[X] &= \frac{1}{B(\alpha_1, \alpha_2)} \int_0^1 x \cdot x^{\alpha_1-1} (1-x)^{\alpha_2-1} dx = \frac{B(\alpha_1+1, \alpha_2)}{B(\alpha_1, \alpha_2)} = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \\ E[X^2] &= \frac{1}{B(\alpha_1, \alpha_2)} \int_0^1 x^{\alpha_1+1} (1-x)^{\alpha_2-1} dx = \frac{B(\alpha_1+2, \alpha_2)}{B(\alpha_1, \alpha_2)} = \frac{\alpha_1+1}{\alpha_1 + \alpha_2 + 1} \frac{\alpha_1}{\alpha_1 + \alpha_2}, \end{aligned}$$

由此可得

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{\alpha_1(1+\alpha_1)}{(\alpha_1 + \alpha_2)(\alpha_1 + \alpha_2 + 1)} - \left(\frac{\alpha_1}{\alpha_1 + \alpha_2}\right)^2 = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2 (\alpha_1 + \alpha_2 + 1)}.$$

**例 9.2** 设独立同分布随机变量  $X_1, X_2, \dots, X_n$  服从均匀分布  $\mathcal{U}(0, 1)$ , 记  $X_{(k)}$  为其顺序统计量, 则

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

**证明** 若随机变量  $X_i \sim U(0, 1)$  ( $i \in [n]$ ), 则当  $x \in (0, 1)$  时其分布函数  $F(x) = x$ . 由此可得到第  $k$  个统计量  $X_{(k)}$  的概率密度函数

$$\begin{aligned} f(x) &= \frac{n!}{(k-1)!(n-k)!} (F(x))^{k-1} (1-F(x))^{n-k} = \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k} \\ &= \frac{1}{B(k, n-k+1)} x^{k-1} (1-x)^{n-k}. \end{aligned}$$

下面定义  $\Gamma$  分布:

**定义 9.7** 如果随机变量  $X$  的概率密度

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

其中  $\alpha > 0$  和  $\lambda > 0$ , 则称随机变量  $X$  服从参数为  $\alpha$  和  $\lambda$  的  $\Gamma$  分布, 记为  $X \sim \Gamma(\alpha, \lambda)$ .

**定理 9.4** 若随机变量  $X \sim \Gamma(\alpha, \lambda)$ , 则有  $E(X) = \alpha/\lambda$  和  $\text{Var}(X) = \alpha/\lambda^2$ .

**证明** 根据期望的定义有

$$E[X] = \int_0^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\lambda x} dx = \frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)} \int_0^{\infty} \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\lambda x} dx = \alpha/\lambda.$$

以及

$$E[X^2] = \int_0^{\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\lambda x} dx = \frac{\Gamma(\alpha+2)}{\lambda^2 \Gamma(\alpha)} \int_0^{\infty} \frac{\lambda^{\alpha+2}}{\Gamma(\alpha+2)} x^{\alpha+1} e^{-\lambda x} dx = \alpha(\alpha+1)/\lambda^2,$$

由此可得

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \alpha(\alpha+1)/\lambda^2 - \alpha^2/\lambda^2 = \alpha/\lambda^2.$$

我们有  $\Gamma$  分布的可加性:

**定理 9.5** 若随机变量  $X \sim \Gamma(\alpha_1, \lambda)$  和  $Y \sim \Gamma(\alpha_2, \lambda)$ , 且  $X$  与  $Y$  相互独立, 则  $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$ .

**证明** 设随机变量  $Z = X + Y$ , 根据独立同分布随机变量和函数的分布有随机变量  $Z$  的概率密度为

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_0^z \frac{\lambda^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1-1} e^{-\lambda x} \frac{\lambda^{\alpha_2}}{\Gamma(\alpha_2)} (z-x)^{\alpha_2-1} e^{-\lambda(z-x)} dx \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\lambda z} \int_0^z x^{\alpha_1-1} (z-x)^{\alpha_2-1} dx \end{aligned}$$

令变量替换  $x = zt$  有

$$\int_0^z x^{\alpha_1-1} (z-x)^{\alpha_2-1} dx = z^{\alpha_1+\alpha_2-1} \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1} dt = z^{\alpha_1+\alpha_2-1} \mathcal{B}(\alpha_1, \alpha_2)$$

在利用 Beta 函数的性质

$$\mathcal{B}(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1} dt = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

代入完成证明.

特别地, 若随机变量  $X \sim \Gamma(1/2, 1/2)$ , 则其密度函数为

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{1}{2}x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

**例 9.3** 若随机变量  $X \sim \mathcal{N}(0, 1)$ , 则有  $X^2 \sim \Gamma(1/2, 1/2)$ .

**解** 首先求解随机变量函数  $Y = X^2$  的分布函数. 当  $y \leq 0$  时有  $F_Y(y) = 0$ ; 当  $y > 0$  时有

$$F_Y(y) = \Pr(X^2 \leq y) = \Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,$$

由此得到概率密度为  $f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}$ . 从而得到  $X^2 \sim \Gamma(1/2, 1/2)$ .

下面介绍 Dirichlet 分布:

**定义 9.8** 给定  $\alpha_1, \alpha_2, \dots, \alpha_k \in (0, +\infty)$ , 若多元随机向量  $X = (X_1, X_2, \dots, X_k)$  的密度函数为

$$f(x_1, x_2, \dots, x_k) = \begin{cases} \frac{x_1^{\alpha_1-1} x_2^{\alpha_2-1} \dots x_k^{\alpha_k-1}}{\text{Beta}(\alpha_1, \alpha_2, \dots, \alpha_k)} & \sum_{i=1}^k x_i = 1, x_i > 0 (i \in [k]), \\ 0 & \text{其它} \end{cases}$$

则称  $X$  服从参数为  $\alpha_1, \alpha_2, \dots, \alpha_k$  的 Dirichlet 分布, 记  $X \sim \text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_k)$ .

Dirichlet 分布是 Beta 分布的一种推广, 当  $k = 2$  时 Dirichlet 分布退化为 Beta 分布.

**定理 9.6** 若随机向量  $X = (X_1, X_2, \dots, X_k) \sim \text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_k)$ , 设  $\tilde{\alpha} = \alpha_1 + \alpha_2 + \dots + \alpha_k$  和  $\tilde{\alpha}_i = \alpha_i / \tilde{\alpha}$ , 则

$$E[X_i] = \tilde{\alpha}_i \quad \text{和} \quad \text{Cov}(X_i, X_j) = \begin{cases} \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\tilde{\alpha}+1} & i = j, \\ -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha}+1} & i \neq j. \end{cases}$$

**证明** 根据期望的定义有

$$\begin{aligned} E[X_i] &= \frac{\int \int_{\sum_i x_i=1, x_i \geq 0} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \dots x_k^{\alpha_k-1} \cdot x_i dx_1 \dots dx_k}{\text{Beta}(\alpha_1, \alpha_2, \dots, \alpha_k)} \\ &= \frac{\text{Beta}(\alpha_1, \dots, \alpha_i + 1, \dots, \alpha_k)}{\text{Beta}(\alpha_1, \dots, \alpha_i, \dots, \alpha_k)} = \frac{\alpha_i}{\alpha_1 + \alpha_2 + \dots + \alpha_k} = \tilde{\alpha}_i. \end{aligned}$$

若  $i = j$ , 则有

$$\text{Cov}(X_i, X_i) = E[X_i^2] - (E[X_i])^2 = \frac{\text{Beta}(\alpha_1, \dots, \alpha_i + 2, \dots, \alpha_k)}{\text{Beta}(\alpha_1, \dots, \alpha_i, \dots, \alpha_k)} - (\tilde{\alpha}_i)^2 = \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\tilde{\alpha}+1}.$$

若  $i \neq j$ , 则有

$$\begin{aligned} \text{Cov}(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] = \frac{\text{Beta}(\alpha_1, \dots, \alpha_i + 1, \dots, \alpha_j + 1, \dots, \alpha_k)}{\text{Beta}(\alpha_1, \dots, \alpha_i, \dots, \alpha_j, \dots, \alpha_k)} - \tilde{\alpha}_i \tilde{\alpha}_j \\ &= \frac{\alpha_i \alpha_j}{\tilde{\alpha}(\tilde{\alpha}+1)} - \tilde{\alpha}_i \tilde{\alpha}_j = -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha}+1}. \end{aligned}$$

## 9.4 正态总体抽样分布定理

### 9.4.1 $\chi^2$ 分布

**定义 9.9** 若  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \mathcal{N}(0, 1)$  的一个样本, 称  $Y = X_1^2 + X_2^2 + \dots + X_n^2$  为服从自由度为  $n$  的  $\chi^2$  分布, 记  $Y \sim \chi^2(n)$ .

根据  $X_1^2 \sim \Gamma(1/2, 1/2)$  和  $\Gamma$  函数的可加性可得  $Y \sim \Gamma(n/2, 1/2)$ . 于是有随机变量  $Y$  的概率密度为

$$f_Y(y) = \begin{cases} \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} & y > 0 \\ 0 & y \leq 0 \end{cases}$$

下面研究  $\chi^2$  分布的性质:

**定理 9.7** 若随机变量  $X \sim \chi^2(n)$ , 则  $E(X) = n$  和  $\text{Var}(X) = 2n$ ; 若随机变量  $X \sim \chi^2(m)$  和  $Y \sim \chi^2(n)$  相互独立, 则  $X + Y \sim \chi^2(m + n)$ ;

**证明** 若随机变量  $X \sim \chi^2(n)$ , 则有  $X = X_1^2 + X_2^2 + \dots + X_n^2$ , 其中  $X_1, X_2, \dots, X_n$  是总体为  $X' \sim \mathcal{N}(0, 1)$  的一个样本. 我们有

$$\begin{aligned} E[X] &= E[X_1^2 + X_2^2 + \dots + X_n^2] = nE[X_1^2] = n, \\ \text{Var}(X) &= n\text{Var}(X_1^2) = n[E(X_1^4) - (E(X_1^2))^2] = n(E(X_1^4) - 1). \end{aligned}$$

计算

$$E(X_1^4) = \int_{-\infty}^{+\infty} \frac{x^4}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = - \int_{-\infty}^{+\infty} \frac{x^3}{\sqrt{2\pi}} de^{-\frac{x^2}{2}} = 3 \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 3$$

可得  $\text{Var}(X) = 2n$ .

若随机变量  $X \sim \mathcal{N}(0, 1)$ , 则

$$E(X^k) = \begin{cases} (k-1)!! & k \text{ 为偶数} \\ 0 & k \text{ 为奇数} \end{cases}$$

其中  $(2k)!! = 2k \cdot (2k-2) \cdot \dots \cdot 2$  和  $(2k+1)!! = (2k+1) \cdot (2k-1) \cdot \dots \cdot 1$ .

**例 9.4** 设  $X_1, X_2, X_3, X_4$  是来自于总体  $\mathcal{N}(0, 4)$  的样本, 以及  $Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$ . 求  $a, b$  取何值时,  $Y$  服从  $\chi^2$  分布, 并求其自由度.

**解** 根据正态分布的性质有  $X_1 - 2X_2 \sim \mathcal{N}(0, 20)$  和  $3X_3 - 4X_4 \sim \mathcal{N}(0, 100)$ , 因此

$$\frac{X_1 - 2X_2}{2\sqrt{5}} \sim \mathcal{N}(0, 1), \quad \frac{3X_3 - 4X_4}{10} \sim \mathcal{N}(0, 1),$$

所以当  $a = 1/20, b = 1/100$  时有  $Y \sim \chi^2(2)$  成立.

分布可加性:

- 如果  $X \sim \mathcal{N}(\mu_1, a_1^2)$  和  $Y \sim \mathcal{N}(\mu_2, a_2^2)$ , 且  $X$  与  $Y$  独立, 那么  $X \pm Y \sim \mathcal{N}(\mu_1 \pm \mu_2, a_1^2 + a_2^2)$ ;
- 如果  $X \sim B(n_1, p)$  和  $Y \sim B(n_2, p)$ , 且  $X$  与  $Y$  独立, 那么  $X + Y \sim B(n_1 + n_2, p)$ ;
- 如果  $X \sim P(\lambda_1)$  和  $Y \sim P(\lambda_2)$ , 且  $X$  与  $Y$  独立, 那么  $X + Y \sim P(\lambda_1 + \lambda_2)$ ;
- 如果  $X \sim \Gamma(\alpha_1, \lambda)$  和  $Y \sim \Gamma(\alpha_2, \lambda)$ , 且  $X$  与  $Y$  独立, 那么  $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$ .
- 如果  $X \sim \chi(m)$  和  $Y \sim \chi(n)$ , 且  $X$  与  $Y$  独立, 那么  $X + Y \sim \chi(m + n)$ .

#### 9.4.2 $t$ 分布 (student distribution)

**定义 9.10** 随机变量  $X \sim \mathcal{N}(0, 1)$  和  $Y \sim \chi^2(n)$  相互独立, 则随机变量

$$T = \frac{X}{\sqrt{Y/n}}$$

服从自由度为  $n$  的  $t$ -分布, 记  $T \sim t(n)$ .

随机变量  $T \sim t(n)$  的概率密度为

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad x \in (-\infty, +\infty).$$

由此可知  $t$ -分布的密度函数  $f(x)$  是偶函数. 当  $n > 1$  为偶数时有

$$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} = \frac{(n-1)(n-3)\cdots 5 \cdot 3}{2\sqrt{n}(n-2)(n-4)\cdots 4 \cdot 2};$$

当  $n > 1$  为奇数时有

$$\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} = \frac{(n-1)(n-3)\cdots 4 \cdot 2}{\pi\sqrt{n}(n-2)(n-4)\cdots 5 \cdot 3}.$$

当  $n \rightarrow \infty$  时, 随机变量  $T \sim t(n)$  的概率密度

$$f(x) \rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

因此当  $n$  足够大时,  $f(x)$  可被近似为  $\mathcal{N}(0, 1)$  的密度函数.

#### 9.4.3 $F$ 分布

**定义 9.11** 设随机变量  $X \sim \chi^2(m)$  和  $Y \sim \chi^2(n)$  相互独立, 称随机变量

$$F = \frac{X/m}{Y/n}$$

服从自由度为  $(m, n)$  的  $F$ -分布, 记  $F \sim F(m, n)$ .

随机变量  $F \sim F(m, n)$  的概率密度为

$$f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})(\frac{m}{n})^{\frac{m}{2}} x^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(1+\frac{mx}{n})^{\frac{m+n}{2}}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

若随机变量  $F \sim F(m, n)$ , 则  $\frac{1}{F} \sim F(n, m)$ .

课题练习:

- 独立同分布随机变量  $X_1, X_2, \dots, X_n$  满足  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ , 求  $\sum_{i=1}^n (X_i - \mu_i)^2 / \sigma_i^2$  的分布.
- 设  $X_1, X_2, \dots, X_9$  和  $Y_1, Y_2, \dots, Y_9$  是分别来自总体  $\mathcal{N}(0, 9)$  的两个独立样本, 求  $(X_1 + X_2 + \dots + X_9) / \sqrt{Y_1^2 + Y_2^2 + \dots + Y_9^2}$  的分布.
- 设  $X_1, X_2, \dots, X_{2n}$  来自总体  $\mathcal{N}(0, \sigma_2)$  的样本, 求  $(X_1^2 + X_3^2 + \dots + X_{2n-1}^2) / (X_2^2 + X_4^2 + \dots + X_{2n}^2)$  的分布.

#### 9.4.4 正态分布的抽样分布定理

**定理 9.8** 设  $X_1, X_2, \dots, X_n$  是来自总体  $\mathcal{N}(\mu, \sigma^2)$  的样本, 则有

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}), \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

**定理 9.9** 设  $X_1, X_2, \dots, X_n$  是来自总体  $\mathcal{N}(\mu, \sigma^2)$  的样本, 其样本均值和修正样本方差分别为

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

则有  $\bar{X}$  和  $S^2$  相互独立, 且

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

此定理证明参考书的附件.

**定理 9.10** 设  $X_1, X_2, \dots, X_n$  是来自总体  $\mathcal{N}(\mu, \sigma^2)$  的样本, 其样本均值和修正样本方差分别为

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

则有

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



**证明** 根据前面两个定理可知  $(\bar{X} - \mu)/\sigma\sqrt{n} \sim \mathcal{N}(0, 1)$  和  $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ , 于是有

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \bigg/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$$

**定理 9.11** 设  $X_1, X_2, \dots, X_m$  和  $Y_1, Y_2, \dots, Y_n$  分别来自总体  $\mathcal{N}(\mu_X, \sigma^2)$  和  $\mathcal{N}(\mu_Y, \sigma^2)$  的两个独立样本, 令其样本均值分别  $\bar{X}$  和  $\bar{Y}$ , 修正样本方差分别为  $S_X^2$  和  $S_Y^2$ , 则

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

**证明** 根据正太分布的性质有  $\bar{X} \sim \mathcal{N}(\mu_X, \sigma^2/m)$  和  $\bar{Y} \sim \mathcal{N}(\mu_Y, \sigma^2/n)$ , 以及

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma^2\right),$$

进一步有

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathcal{N}(0, 1).$$

根据定理 9.9 有  $\frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2(m-1)$  和  $\frac{(n-1)S_Y^2}{\sigma^2} \sim \chi^2(n-1)$ , 由此得到

$$\frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi^2(m+n-2).$$

从而完成证明.

**定理 9.12** 设  $X_1, X_2, \dots, X_m$  和  $Y_1, Y_2, \dots, Y_n$  分别来自总体  $\mathcal{N}(\mu_X, \sigma_X^2)$  和  $\mathcal{N}(\mu_Y, \sigma_Y^2)$  的两个独立样本, 令其修正样本方差分别为  $S_X^2$  和  $S_Y^2$ , 则有

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

**证明** 根据定理 9.9 有  $\frac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2(m-1)$  和  $\frac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n-1)$ , 由此得到

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

课堂习题:

- 若随机变量  $X \sim t(n)$ , 求  $Y = X^2$  的分布.

- 设  $X_1, X_2, \dots, X_5$  是来自总体  $\mathcal{N}(0, 1)$  的样本, 令  $Y = c_1(X_1 + X_3)^2 + c_2(X_2 + X_4 + X_5)^2$ . 求常数  $c_1, c_2$  使  $Y$  服从  $\chi^2$  分布.
- 设  $X_1, X_2$  是来自总体  $\mathcal{N}(0, \sigma^2)$  的样本, 求  $\frac{(X_1 + X_2)^2}{(X_1 - X_2)^2}$  的分布.

#### 9.4.5 分位数(点)

**定义 9.12** 对给定  $\alpha \in (0, 1)$  和随机变量  $X$ , 称满足  $\Pr(X > \lambda_\alpha) = \alpha$  的实数  $\lambda_\alpha$  为上侧  $\alpha$  分位数(点).

对正态分布  $X \sim \mathcal{N}(0, 1)$ , 给定  $\alpha \in (0, 1)$ , 满足  $\Pr(X > \mu_\alpha) = \int_{\mu_\alpha}^{\infty} f(x)dx = \alpha$  的点  $\mu_\alpha$  称为正态分布上侧  $\alpha$  分位点, 由对称性可知  $\mu_{1-\alpha} = -\mu_\alpha$ .

对  $\chi^2(n)$  分布  $X \sim \chi^2(n)$ , 给定  $\alpha \in (0, 1)$ , 满足  $\Pr(X \geq \chi_\alpha^2(n)) = \alpha$  的点  $\chi_\alpha^2(n)$  称为  $\chi^2(n)$  分布上侧  $\alpha$  分位点. 当  $n \rightarrow \infty$  时有  $\chi_\alpha^2(n) \approx \frac{1}{2}(\mu_\alpha + \sqrt{2n-1})^2$ , 其中  $\mu_\alpha$  表示正态分布上侧  $\alpha$  分位点.

对  $t$ -分布  $X \sim t(n)$ , 给定  $\alpha \in (0, 1)$ , 满足  $\Pr(X > t_\alpha(n)) = \alpha$  的点  $t_\alpha(n)$  称为  $t(n)$ -分布上侧  $\alpha$  分位点. 由对称性可知  $t_{(1-\alpha)}(n) = -t_\alpha(n)$ .

对  $F$ -分布  $X \sim F(m, n)$ , 给定  $\alpha \in (0, 1)$ , 满足  $\Pr[X > F_\alpha(m, n)] = \alpha$  的点  $F_\alpha(m, n)$  称为  $F(m, n)$  分布上侧  $\alpha$  分位点.

对于  $F$ -分布, 有如下性质:

**引理 9.4** 对  $F$  分布的分位点有

$$F_{(1-\alpha)}(m, n) = \frac{1}{F_\alpha(n, m)}.$$

**证明** 设  $X \sim F(m, n)$ , 根据定义有

$$1 - \alpha = \Pr(X > F_{1-\alpha}(m, n)) = \Pr\left(\frac{1}{X} < \frac{1}{F_{1-\alpha}(m, n)}\right) = 1 - \Pr\left(\frac{1}{X} \geq \frac{1}{F_{1-\alpha}(m, n)}\right).$$

再根据  $1/X \sim F(n, m)$ , 结合上式有

$$\alpha = \Pr\left(\frac{1}{X} \geq \frac{1}{F_{1-\alpha}(m, n)}\right) = \Pr\left(\frac{1}{X} > \frac{1}{F_{1-\alpha}(m, n)}\right)$$

于是有  $F_\alpha(n, m) = 1/F_{1-\alpha}(m, n)$ .

课堂习题:

- 设  $X_1, X_2, \dots, X_{10}$  是总体  $\mathcal{N}(\mu, 1/4)$  的样本, i) 若  $\mu = 0$ , 求  $\Pr(\sum_{i=1}^{10} X_i^2 \geq 4)$ ; ii) 若  $\mu$  未知, 求  $\Pr(\sum_{i=1}^{10} (X_i - \bar{X})^2 \geq 2.85)$ .
- 设  $X_1, X_2, \dots, X_{25}$  是总体  $\mathcal{N}(12, \sigma^2)$  的样本, i) 若  $\sigma = 2$ , 求  $\Pr(\sum_{i=1}^{25} X_i/25 \geq 12.5)$ ; ii) 若  $\sigma$  未知但知道修正样本方差为  $S^2 = 5.57$ , 求  $\Pr(\sum_{i=1}^{25} X_i/25 \geq 12.5)$ .

## 习题

9.1 设随机变量  $X$  的期望  $E[X] = \mu > 0$ , 方差为  $\sigma^2$ , 证明对任意  $\epsilon > 0$  有

$$P(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$



## 第 10 章 参数估计

设总体  $X$  的分布函数为  $F(X, \theta)$ , 其中  $\theta$  为未知参数(也可向量为向量). 现从总体中抽取一样本  $X_1, X_2, \dots, X_n$ , 如何依据样本估计参数  $\theta$ , 或  $\theta$  的函数  $g(\theta)$ , 此类问题称为参数估计问题. 内容包括: 点估计, 估计量标准, 区间估计.

### 10.1 点估计

#### 10.1.1 矩估计法

总体  $X$  的  $k$  阶矩:  $a_k = E[X^k]$

样本  $k$  阶矩:  $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$

用相应的样本矩去估计总体矩, 求解参数  $\theta$  的方法称为 **矩估计法**. 矩估计法的理论基础是大数定理:  $X_1, X_2, \dots, X_n$  为 i.i.d. 的随机变量, 若  $E(X) = \mu$ , 则当  $n \rightarrow \infty$  时有

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mu.$$

推论: 若  $E[X^k] = a_k$  存在, 则当  $n \rightarrow \infty$  时有

$$\frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} a_k = E[X^k].$$

还可利用中心矩进行估计:

总体  $X$  的  $k$  阶中心矩:  $b_k = E[(X - E(X))^k]$

样本  $k$  阶中心矩:  $B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$

矩估计方法: 总体  $X$  的分布函数  $F$  包含  $m$  个未知参数  $\theta_1, \theta_2, \dots, \theta_m$ ,

- 1) 求总体  $X$  的  $k$  阶矩:  $a_k = a_k(\theta_1, \theta_2, \dots, \theta_m) = E[X^k]$ ,  $k \in [m]$  ( $a_k$  一般为  $\theta_1, \theta_2, \dots, \theta_m$  的函数).
- 2) 计算样本的  $k$  阶矩:  $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ .
- 3) 令样本矩等于总体矩  $A_k = a_k = a_k(\theta_1, \theta_2, \dots, \theta_m)$  ( $k = 1, 2, \dots, m$ ), 得到  $m$  个关于  $\theta_1, \theta_2, \dots, \theta_m$  的方程组.
- 4) 求解方程组得到估计量  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m$ .

**例 10.1** 设总体  $X$  的概率密度函数

$$f(x) = \begin{cases} (\alpha + 1)x^\alpha & x \in (0, 1) \\ 0 & \text{其它,} \end{cases}$$

设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 求参数  $\alpha$  的矩估计.

**解** 首先计算总体  $X$  的期望

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^1 x(\alpha+1)x^{\alpha+1}dx = \frac{\alpha+1}{\alpha+2}.$$

样本  $X$  的均值  $\bar{X} = \sum_{i=1}^n X_i/n$ . 样本矩等于总体矩有

$$E(X) = \frac{\alpha+1}{\alpha+2} = \bar{X},$$

求解可得  $\alpha = (2\bar{X} - 1)/(1 - \bar{X})$ .

**例 10.2** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 以及总体  $X$  的密度函数为

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x-\mu}{\theta}} & x \geq \mu \\ 0 & \text{其它,} \end{cases}$$

其中  $\theta > 0$ , 求  $\mu$  和  $\theta$  的矩估计.

**解** 设随机变量  $Y = X - \mu$ , 则  $Y$  服从参数为  $1/\theta$  的指数分布, 有

$$E(Y) = \theta \quad \text{和} \quad \text{Var}(Y) = \theta^2.$$

由此可得  $E(X) = \mu + \theta$  和  $\text{Var}(X) = \theta^2$ . 计算对应的样本矩

$$A_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad B_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

求解方程组

$$\mu + \theta = A_1 \quad \text{和} \quad \theta^2 = B_2,$$

解得  $\mu = \bar{X} - \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/n}$  和  $\theta = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/n}$ .

课堂习题:

- 求正态总体  $\mathcal{N}(\mu, \sigma^2)$  的  $\mu, \sigma^2$  的矩估计法.
- 求总体  $X \sim \mathcal{U}(a, b)$  中  $a, b$  的矩估计法.

## 10.1.2 最大似然估计法

设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的一个样本. 若总体  $X$  为离散型随机变量, 其分布列为  $\Pr(X = x) = \Pr(X = x; \theta)$ , 则样本  $X_1, X_2, \dots, X_n$  的分布列为

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n \Pr(x_i; \theta).$$

这里  $L(\theta)$  表示样本  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  发生的概率.

若总体  $X$  为连续型随机变量, 其概率密度为  $f(x; \theta)$ , 则  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  的联合概率密度为

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta).$$

根据概率密度定义可知  $L(\theta)$  越大, 样本  $(X_1, X_2, \dots, X_n)$  落入  $(x_1, x_2, \dots, x_n)$  的邻域内概率越大.

综合上述离散和连续两种随机变量, 统称  $L(\theta)$  为样本  $X_1, X_2, \dots, X_n$  的似然函数, 可以发现  $L(\theta)$  是  $\theta$  的函数, 若

$$\hat{\theta} = \arg \max_{\theta} L(x_1, x_2, \dots, x_n; \theta),$$

则称  $\hat{\theta}$  为  $\theta$  的最大似然估计量. 直觉而言: 最大似然估计量  $\hat{\theta}$  是使观测值  $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$  出现的概率最大.

求解最大似然估计量的步骤如下:

- i) 计算对数似然函数  $\log(L(x_1, x_2, \dots, x_n; \theta))$ ;
- ii) 求对数似然函数中参数  $\theta$  的一阶偏导, 令其等于零;
- iii) 求解方程组得到最大似然估计量  $\hat{\theta}$ .

**例 10.3** 设  $X_1, X_2, \dots, X_n$  是取自总体  $X \sim B(1, p)$  的样本, 求参数  $p$  的最大似然估计.

**解** 首先计算似然函数

$$L(p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i},$$

从而得到对数似然函数

$$\ln L(p) = \sum_{i=1}^n X_i \ln p + \left( n - \sum_{i=1}^n X_i \right) \ln(1-p),$$

求一阶偏导并令其为零可得

$$\frac{\partial \ln L(p)}{\partial p} = \frac{1}{p} \sum_{i=1}^n X_i - \frac{1}{1-p} \left( n - \sum_{i=1}^n X_i \right) = 0.$$

由此求解  $p = \sum_{i=1}^n X_i/n = \bar{X}$ . [验证矩估计法]

下面讨论 **最大似然估计不可变性**

**性质 10.1** 设  $\mu(\theta)$  为  $\theta$  的函数, 且存在反函数  $\theta = \theta(\mu)$ . 若  $\hat{\theta}$  是  $\theta$  的最大似然估计, 则  $\hat{\mu} = \mu(\hat{\theta})$  是  $\mu$  的最大似然估计.

**例 10.4** 设  $X_1, X_2, \dots, X_n$  为总体  $X \sim \mathcal{N}(\mu, \sigma^2)$  的样本, 求  $\mu$  和  $\sigma > 0$  的最大似然估计.

**解** 根据高斯分布知  $X$  的概率密度为  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ . 样本  $X_1, X_2, \dots, X_n$  的似然函数为

$$L(\mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}\right).$$

其对数似然函数为  $\ln L(\mu, \sigma) = -n \ln(2\pi)^{1/2} - n \ln \sigma - \sum_{i=1}^n (X_i - \mu)^2 / 2\sigma^2$ . 对参数  $\mu$  求导计算可得

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^n (X_i - \mu) = 0 \implies \mu = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

对  $\sigma$  求导计算可得

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (X_i - \mu)^2 = 0 \implies \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

根据最大似然估计的不变性可知方差  $\sigma^2$  的最大似然估计为  $\sigma^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n$ . 下面进行验证最大似然估计的不变性: 设  $X_1, X_2, \dots, X_n$  为总体  $X \sim \mathcal{N}(\mu, \nu)$  的样本, 求  $\mu$  和  $\nu$  的最大似然估计. 根据题意可知样本  $X_1, X_2, \dots, X_n$  的对数似然函数为

$$\ln L(\mu, \nu) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \nu - \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\nu}.$$

对参数  $\mu$  求偏导计算其最大似然估计  $\mu = \sum_{i=1}^n X_i / n = \bar{X}$ , 对  $\nu$  求偏导计算可得

$$\frac{\partial \ln L(\mu, \nu)}{\partial \nu} = -\frac{n}{2\nu} + \frac{1}{2\nu^2} \sum_{i=1}^n (X_i - \mu)^2 = 0 \implies \nu = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

从而完成验证.

**例 10.5** 设总体  $X$  的密度函数为

$$f(x) = \begin{cases} (\alpha + 1)x^\alpha & x \in (0, 1) \\ 0 & \text{其它} \end{cases}$$

设  $X_1, X_2, \dots, X_n$  是总体  $X$  的样本, 求  $\alpha$  的最大似然估计.



解 首先得到似然函数为

$$L(\alpha) = (\alpha + 1)^n \prod_{i=1}^n X_i^\alpha = (\alpha + 1)^n (X_1 X_2 \cdots X_n)^\alpha,$$

以及其对数似然函数  $\ln L(\alpha) = n \ln(\alpha + 1) + \alpha \ln(X_1 X_2 \cdots X_n)$ . 求导并令偏导为零有

$$\frac{\partial \ln L(\alpha)}{\partial \alpha} = \frac{n}{\alpha + 1} + \ln(X_1 X_2 \cdots X_n) = 0,$$

求解得

$$\alpha = \frac{-n}{\sum_{i=1}^n \ln(X_i)} - 1 = \frac{-1}{\frac{1}{n} \sum_{i=1}^n \ln(X_i)} - 1.$$

对上例, 矩估计值为  $\alpha = (2\bar{X} - 1)/(1 - \bar{X})$ , 因此矩估计值与最大似然估计值可能不同.

**例 10.6** 设  $X_1, X_2, \cdots, X_n$  是总体  $X \sim \mathcal{U}(a, b)$  的样本, 求  $a$  和  $b$  的最大似然估计.

解 当  $x \in [a, b]$  时, 总体  $X$  的概率密度为  $f(x) = 1/(b - a)$ , 其它情况为零, 因此似然函数为

$$L(a, b) = \begin{cases} \frac{1}{(b-a)^n} & a \leq X_1, X_2, \cdots, X_n \leq b \\ 0 & \text{其它} \end{cases}$$

直接求偏导无法解出  $a$  和  $b$ , 此时可以从最大似然的定义出发, 应使得  $b$  尽可能小且  $a$  尽可能大, 但需满足  $a \leq X_1, X_2, \cdots, X_n \leq b$ , 因此最大似然估计量为:

$$b = \max\{X_1, X_2, \cdots, X_n\} \quad \text{和} \quad a = \min\{X_1, X_2, \cdots, X_n\}.$$

**例 10.7** 设  $X_1, X_2, \cdots, X_n$  是总体  $X$  的样本, 以及总体  $X$  的概率密度为

$$f(x) = \begin{cases} \theta e^{-(x-\mu)\theta} & x \geq \mu \\ 0 & \text{其它,} \end{cases}$$

求  $\mu$  和  $\theta$  的最大似然估计.

解 首先计算似然函数为

$$L(\theta, \mu) = \begin{cases} \theta^n e^{-\theta \sum_{i=1}^n (X_i - \mu)} & X_i \geq \mu \\ 0 & \text{其它} \end{cases}$$

进一步得到对数似然函数为

$$\ln L(\theta, \mu) = n \ln \theta - \theta \sum_{i=1}^n (X_i - \mu).$$

求偏导、并令偏导等于零有

$$\frac{\partial \ln L(\theta, \mu)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^n (X_i - \mu) = 0 \Rightarrow \theta = \frac{1}{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)},$$

另一方面有

$$\frac{\partial \ln L(\theta, \mu)}{\partial \mu} = n\theta = 0 \Rightarrow \theta = 0,$$

此时无法求解  $\theta$  和  $\mu$  的最大似然估计. 回到似然函数的定义

$$L(\theta, \mu) = \begin{cases} \theta^n e^{-\theta \sum_{i=1}^n (X_i - \mu)} & X_1, X_2, \dots, X_n \geq \mu \\ 0 & \text{其它} \end{cases}$$

可以发现  $\mu$  越大似然函数  $L(\theta, \mu)$  越大, 但须满足  $X_i \geq \mu$  ( $i \in [n]$ ). 由此可得最大似然估计

$$\hat{\mu} = \min\{X_1, X_2, \dots, X_n\},$$

进一步求解可得

$$\hat{\theta} = \frac{1}{\frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})}.$$

## 10.2 估计量的评价标准

前一节已经讲过不同的点估计方法, 不同的估计方法可能得到不同的估计值, 自然涉及到一个问题: 采用哪一种估计量更好, 或更好的标准是什么呢? 估计量的常用标准: 无偏性, 有效性, 一致性.

### 10.2.1 无偏性

**定义 10.1** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 令  $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$  是  $\theta$  的一个估计量, 若

$$E_{X_1, X_2, \dots, X_n} [\hat{\theta}] = E_{X_1, X_2, \dots, X_n} [\hat{\theta}(X_1, X_2, \dots, X_n)] = \theta$$

则称  $\hat{\theta}$  为  $\theta$  的无偏估计.

无偏估计不要求估计值  $\hat{\theta}$  在任意情况下都等于  $\theta$ , 但在期望的情形下有  $E(\hat{\theta}) = \theta$  成立. 其意义在于无系统性偏差, 无偏性是一种对估计量常见而且重要的标准.

首先看看如下例子:

**例 10.8 (样本  $k$  阶原点矩为总体  $k$  阶原点矩的无偏估计)** 设  $X_1, X_2, \dots, X_n$  是总体  $X$  的样本, 若  $E[X^k]$  存在, 则  $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$  是总体  $a_k = E[X^k]$  的无偏估计.

**例 10.9** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 其期望为  $\mu$ , 方差为  $\sigma^2$ , 则: 1)  $S_0^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n$  是  $\sigma^2$  的有偏估计; 2)  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$  是  $\sigma^2$  的无偏估计.

注意  $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$  是  $\theta$  的无偏估计, 但并不一定有  $g(\hat{\theta})$  是  $g(\theta)$  的无偏估计, 这是因为  $E[\hat{\theta}] = \theta$  并不能推导出  $E[g(\hat{\theta})] = g(\theta)$ . 例如

$$E[\bar{X}] = E[X] = \mu \quad \text{但} \quad E[(\bar{X})^2] \neq \mu^2.$$

**例 10.10** 设  $X_1, X_2, \dots, X_n$  是总体  $X$  的样本, 以及总体  $X$  的概率密度为

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x \geq 0 \\ 0 & x < 0, \end{cases}$$

证明:  $\bar{X} = \sum_{i=1}^n X_i/n$  和  $n \min\{X_1, X_2, \dots, X_n\}$  均是  $\theta$  的无偏估计.

**证明** 根据期望和指数分布的性质有

$$E[\bar{X}] = E[X] = \theta,$$

由此可知  $\bar{X}$  是  $E[X]$  的无偏估计. 设随机变量  $Z = \min\{X_1, X_2, \dots, X_n\}$ , 则有

$$\begin{aligned} F_Z(z) &= \Pr[Z \leq z] = 1 - \Pr[Z > z] \\ &= 1 - \Pr[X_1 > z] \Pr[X_2 > z] \cdots \Pr[X_n > z] \\ &= 1 - \prod_{i=1}^n (1 - \Pr[X_i \leq z]) = \begin{cases} 0 & z < 0 \\ 1 - e^{-nz/\theta} & z \geq 0. \end{cases} \end{aligned}$$

于是当  $z \geq 0$  时有

$$\Pr[Z > z] = 1 - F_Z(z) = e^{-nz/\theta}.$$

根据期望的性质有

$$E[Z] = \int_0^{+\infty} \Pr[Z > z] dz = \int_0^{+\infty} e^{-nz/\theta} dz = \frac{\theta}{n}.$$

于是有  $\theta = E[nZ]$  成立.

### 10.2.1.1 有效性

参数可能存在多个无偏估计, 若  $\hat{\theta}_1$  和  $\hat{\theta}_2$  都是  $\theta$  的无偏估计, 则可以比较方差

$$\text{Var}(\hat{\theta}_1) = E[(\hat{\theta}_1 - \theta)^2] \quad \text{和} \quad \text{Var}(\hat{\theta}_2) = E[(\hat{\theta}_2 - \theta)^2].$$

一般而言: 方差越小, 无偏估计越好.

**定义 10.2** 设  $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$  和  $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$  是  $\theta$  的两个无偏估计, 若

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

则称  $\theta_1$  比  $\theta_2$  有效.

**例 10.11** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 且  $X$  的概率密度为

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x \geq 0 \\ 0 & x < 0 \end{cases},$$

令  $Z = \min\{X_1, X_2, \dots, X_n\}$ , 证明: 当  $n > 1$  时  $\bar{X} = \sum_{i=1}^n X_i/n$  比  $nZ$  有效.

**证明** 根据独立性有

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\theta^2}{n}.$$

根据例 10.10 可知随机变量  $Z$  的概率密度为

$$f(z) = \begin{cases} 0 & z < 0 \\ \frac{n}{\theta} e^{-\frac{nz}{\theta}} & z \geq 0 \end{cases}$$

从而得到

$$\text{Var}(nZ) = n^2 \text{Var}(Z) = n^2 \frac{\theta^2}{n^2} = \theta^2,$$

因此当  $n \geq 1$  时有  $\text{Var}(\bar{X}) \leq \text{Var}(nZ)$  成立, 故  $\bar{X}$  比  $nZ$  有效.

**例 10.12** 设  $X_1, X_2, \dots, X_n$  是总体  $X$  的样本, 且  $E(X) = \mu$  和  $\text{Var}(X) = \sigma^2$ . 设常数  $c_1, c_2, \dots, c_n \geq 0$  满足  $\sum_{i=1}^n c_i = 1, c_i \neq 1/n$ , 求证:  $\bar{X}$  比  $\sum_{i=1}^n c_i X_i$  有效.

**证明** 根据样本的独立同分布条件有

$$E[\bar{X}] = \mu \quad \text{和} \quad \text{Var}(\bar{X}) = \sigma^2/n.$$

根据期望的性质有  $E[\sum_{i=1}^n c_i X_i] = \mu$ , 进一步有

$$\text{Var}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \text{Var}(X_i) = \sigma^2 \sum_{i=1}^n c_i^2 \geq \frac{\sigma^2}{n}$$

这里利用不等式  $\sum_{i=1}^n c_i^2/n \geq (\sum_{i=1}^n c_i/n)^2 = 1/n^2$ , 所以有  $\text{Var}(\sum_{i=1}^n c_i X_i) \geq \text{Var}(\bar{X})$ .

下面定义有效统计量:

**定理 10.1 (Rao-Crammer 不等式)** 设随机变量  $X$  的概率密度为  $f(x; \theta)$  或分布函数为  $F(x; \theta)$ , 令

$$\text{Var}_0(\theta) = \frac{1}{nE\left[\left(\frac{\partial \ln f(X; \theta)}{\partial \theta}\right)^2\right]} \quad \text{或} \quad \text{Var}_0(\theta) = \frac{1}{nE\left[\left(\frac{\partial \ln F(X; \theta)}{\partial \theta}\right)^2\right]},$$

对任意的无偏估计量  $\hat{\theta}$  有

$$\text{Var}(\hat{\theta}) \geq \text{Var}_0(\theta),$$

称  $\text{Var}_0(\theta)$  为估计量  $\hat{\theta}$  方差的下界. 当  $\text{Var}(\hat{\theta}) = \text{Var}_0(\theta)$  时称  $\hat{\theta}$  为达到方差下界的无偏估计量, 此时  $\hat{\theta}$  为最有效估计量, 简称有效估计量.

**例 10.13** 设  $X_1, X_2, \dots, X_n$  为总体  $X$  的样本, 令总体  $X$  的密度函数为

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & x \leq 0, \end{cases}$$

证明:  $\theta$  的最大似然估计为有效估计量.

**解** 首先计算对数似然函数

$$\ln L(\theta) = -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^n X_i \quad \Rightarrow \quad \hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i,$$

进一步得到统计量的方差

$$\text{Var}(\hat{\theta}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{\theta^2}{n}.$$

同时考察

$$\ln f(X; \theta) = -\ln \theta - \frac{X}{\theta}, \quad \frac{\partial \ln f(X; \theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{X}{\theta^2}$$

所以

$$E\left[\frac{\partial \ln f(X; \theta)}{\partial \theta}\right]^2 = E\left[\left(-\frac{1}{\theta} + \frac{X}{\theta^2}\right)^2\right] = \frac{1}{\theta^4} E[(X - E[X])^2] = \frac{1}{\theta^2},$$

从而得到  $\text{Var}_0(X) = \theta^2/n = \text{Var}(\hat{\theta})$ , 因此  $\theta$  的最大似然估计是有效估计量.

### 10.2.1.2 一致性

**定义 10.3** 设  $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$  是  $\theta$  的一个估计量, 若当  $n \rightarrow \infty$  时有  $\hat{\theta}_n \xrightarrow{P} \theta$  成立, 即对任意  $\epsilon > 0$  有

$$\lim_{n \rightarrow \infty} \Pr[|\hat{\theta}_n - \theta| > \epsilon] = 0,$$

则称  $\hat{\theta}_n$  为  $\theta$  的一致估计量.

估计量的一致性刻画了在足够多样本情形下估计量  $\hat{\theta}$  能有效逼近真实值  $\theta$ , 一致性是对估计的基本要求, 不满足一致性的估计量一般不予考虑. 下面给出满足一致性的充分条件:

**定理 10.2** 设  $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$  是  $\theta$  的一个估计量, 若满足以下两个条件:

$$\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta \quad \text{和} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0,$$

则  $\hat{\theta}_n$  为  $\theta$  的一致估计量.

**证明** 根据  $\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta$  知道对任意  $\epsilon > 0$ , 存在一个  $N_0$ , 当  $n \geq N_0$  有  $|E[\hat{\theta}_n] - \theta| \leq \epsilon/2$ , 于是有

$$\lim_{n \rightarrow \infty} \Pr \left[ |E[\hat{\theta}_n] - \theta| > \epsilon/2 \right] = 0.$$

根据 Chebyshev 不等式有

$$\lim_{n \rightarrow 0} \Pr \left[ \left| \hat{\theta}_n - E[\hat{\theta}_n] \right| > \epsilon/2 \right] \leq \lim_{n \rightarrow 0} \frac{4}{\epsilon} \text{Var}(\hat{\theta}_n) = 0$$

再根据

$$\Pr \left[ |\hat{\theta}_n - \theta| > \epsilon \right] \leq \Pr \left[ \left| \hat{\theta}_n - E[\hat{\theta}_n] \right| > \epsilon/2 \right] + \Pr \left[ |E[\hat{\theta}_n] - \theta| > \epsilon/2 \right]$$

完成证明.

**定理 10.3** 设  $\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \dots, \hat{\theta}_{n_k}$  分别为  $\theta_1, \theta_2, \dots, \theta_k$  满足一致性的估计量, 对连续函数  $g: \mathbb{R}^n \rightarrow \mathbb{R}$ , 有函数  $\hat{\eta}_n = g(\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \dots, \hat{\theta}_{n_k})$  是  $\eta = g(\theta_1, \theta_2, \dots, \theta_k)$  满足一致性的估计量.

根据大数定理可知样本的  $k$  阶矩是总体  $k$  阶矩的一致估计量. 矩估计法得到的估计量一般是一致估计量. 最大似然估计量在一定条件下是一致性估计量.

**例 10.14** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 以及总体  $X$  的密度函数为

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & x < 0 \end{cases},$$

则样本均值  $X_n = \sum_{i=1}^n X_i/n$  为  $\theta$  的无偏、有效、一致估计量.

由前面的例子可知估计的无偏性和有效性, 一致性可根据  $E[X_n] = \theta$  以及

$$\lim_{n \rightarrow \infty} \text{Var}(\bar{X}) = \lim_{n \rightarrow \infty} \frac{\theta^2}{n} = 0.$$

**例 10.15** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim U(0, \theta)$  的样本, 证明:  $\theta$  的最大似然估计量是一致估计量.

**证明** 根据前面的例题可知  $\theta$  的最大似然估计为  $\hat{\theta}_n = \max(X_1, X_2, \dots, X_n)$ . 设随机变量  $Z = \max(X_1, X_2, \dots, X_n)$ , 则由  $Z$  的分布函数

$$F_Z(z) = \Pr[Z \leq z] = \Pr[\max(X_1, X_2, \dots, X_n) \leq z] = \prod_{i=1}^n \Pr[X_i \leq z] = \begin{cases} 1 & z > \theta \\ (\frac{z}{\theta})^n & z \in [0, \theta] \\ 0 & z < 0. \end{cases}$$

由此得到当  $z \in [0, \theta]$  时随机变量  $Z$  的密度函数  $f_Z(z) = nz^{n-1}/\theta^n$ , 进一步有

$$E[\hat{\theta}_n] = E[Z] = \int_0^\theta \frac{nz^n}{\theta^n} dz = \frac{n}{n+1}\theta,$$

因此  $\hat{\theta}$  是  $\theta$  的有偏估计. 另一方面有

$$E[Z^2] = \int_0^\theta \frac{nz^{n+1}}{\theta^n} dz = \frac{n}{n+2}\theta^2,$$

从而得到

$$\text{Var}(\hat{\theta}_n) = \text{Var}(Z) = E[Z^2] - (E[Z])^2 = \frac{n}{n+2}\theta^2 - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n}{(n+1)^2(n+2)}\theta^2,$$

于是有

$$\lim_{n \rightarrow \infty} E[\hat{\theta}_n] = \theta \quad \text{和} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0,$$

由此可得  $\hat{\theta}$  是  $\theta$  的有偏、但一致估计量.

### 10.3 区间估计

区间估计问题: 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本,  $\theta$  为总体  $X$  的分布函数  $F(x, \theta)$  的未知参数, 根据样本估计  $\theta$  的范围  $(\hat{\theta}_1, \hat{\theta}_2)$ , 其中  $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$  和  $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$ , 使得以较大的概率保证有  $\theta \in (\hat{\theta}_1, \hat{\theta}_2)$  成立. 具体而言, 对任意给定  $\alpha \in (0, 1)$ , 有

$$\Pr[\hat{\theta}_1(X_1, X_2, \dots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \dots, X_n)] \geq 1 - \alpha.$$

**定义 10.4 (置信区间与置信度)** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 总体  $X$  的分布函数含未知参数  $\theta$ , 找出统计量  $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$  和  $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \dots, X_n)$  ( $\hat{\theta}_1 < \hat{\theta}_2$ ), 使得

$$\Pr[\hat{\theta}_1 < \theta < \hat{\theta}_2] \geq 1 - \alpha$$

成立, 则称  $1 - \alpha$  为置信度,  $[\hat{\theta}_1, \hat{\theta}_2]$  为  $\theta$  的置信度为  $1 - \alpha$  的置信区间.

注意: 置信区间  $[\hat{\theta}_1, \hat{\theta}_2]$  是随机区间,  $1 - \alpha$  为该区间包含  $\theta$  的概率/可靠程度. 若  $\alpha = 0.05$ , 则置信度为 95%. 通常采用 95% 的置信度, 有时也可 99% 或 90% 等. 说明:

- i)  $\hat{\theta}_2 - \hat{\theta}_1$  反映了估计精度, 长度越小精度越大.
- ii)  $\alpha$  反映了估计的可靠度,  $\alpha$  越小可靠度越高.
- iii) 给定  $\alpha$ , 区间  $[\hat{\theta}_1, \hat{\theta}_2]$  的选取并不唯一确定, 通常选长度最小的一个区间.

置信区间的求解方法: **枢轴变量法**.

- 1) 先找一样本函数  $W(X_1, X_2, \dots, X_n; \theta)$  包含待估参数  $\theta$ , 但不含其它参数, 函数  $W$  的分布已知, 称  $W$  为枢轴变量.
- 2) 给定置信度  $1 - \alpha$ , 根据  $W$  的分布找出临界值  $a$  和  $b$ , 使得  $\Pr[a < W < b] = 1 - \alpha$  成立.
- 3) 根据  $a < W < b$  解出  $\hat{\theta}_1 < \theta < \hat{\theta}_2$ , 则  $(\hat{\theta}_1, \hat{\theta}_2)$  为  $\theta$  的置信度为  $1 - \alpha$  的置信区间.

### 10.3.1 正态总体, 方差已知, 求期望的区间估计

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \mathcal{N}(\mu, \sigma^2)$  的样本, 若方差  $\sigma^2$  已知. 给定  $\alpha \in (0, 1)$ , 确定置信度为  $1 - \alpha$  下  $\mu$  的置信区间  $[\hat{\theta}_1, \hat{\theta}_2]$ . 令样本均值为  $\bar{X} = \sum_{i=1}^n X_i/n$ , 根据正态分布的性质找出枢轴变量:

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

给定置信度  $1 - \alpha$ , 找出临界值  $a$  和  $b$  使得

$$\Pr[a < W < b] = 1 - \alpha.$$

根据正态分布的性质、对称性和上分位点可知

$$\Pr[W \geq \mu_{\alpha/2}] = 1 - \alpha/2 \quad \text{和} \quad \Pr[W \leq -\mu_{\alpha/2}] = 1 - \alpha/2.$$

求解可得  $a = -\mu_{\alpha/2}$  和  $b = \mu_{\alpha/2}$ . 于是有

$$\Pr[-\mu_{\alpha/2} < W < \mu_{\alpha/2}] = 1 - \alpha.$$

根据  $W = (\bar{X} - \mu)/(\sigma/\sqrt{n})$  可得

$$\Pr\left[\bar{X} - \frac{\sigma}{\sqrt{n}}\mu_{\alpha/2} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}}\mu_{\alpha/2}\right] = 1 - \alpha.$$

**例 10.16** 某地区儿童身高服从正态分布, 现随机抽查 9 人, 高度分别为 115, 120, 131, 115, 109, 115, 115, 105, 110, 已知  $\sigma^2 = 7$  和置信度为 95%, 求期望  $\mu$  的置信区间 ( $\mu_{0.025} = 1.96$ ).

### 10.3.2 正态总体, 方差未知, 求期望的区间估计

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \mathcal{N}(\mu, \sigma^2)$  的样本, 若方差  $\sigma^2$  未知, 考虑期望  $\mu$  的置信度为  $1 - \alpha$  的置信区间. 设  $\bar{X} = \sum_{i=1}^n X_i/n$  和  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ , 根据正态总体抽样定理可知:

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

由此设枢轴变量

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



给定置信度  $1 - \alpha$ , 设临界值  $a$  和  $b$  满足

$$\Pr[a \leq W \leq b] = 1 - \alpha \Rightarrow b = t_{\alpha/2}(n-1), a = -t_{\alpha/2}(n-1).$$

整理可得

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

### 10.3.3 正态总体, 求方差 $\sigma^2$ 的置信区间

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \mathcal{N}(\mu, \sigma^2)$  的样本, 考虑方差  $\sigma^2$  的置信度为  $1 - \alpha$  的置信区间. 设修正样本方差  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$ , 根据正态总体抽样定理有

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

由此设枢轴变量  $W = (n-1)S^2/\sigma^2$ , 设临界值  $a$  和  $b$  满足

$$\Pr[a \leq W \leq b] = 1 - \alpha.$$

根据  $\chi^2$  分布的不对称性, 采用概率对称的区间

$$\Pr[W \leq a] = \Pr[b \leq W] = \alpha/2 \Rightarrow b = \chi_{\alpha/2}^2(n-1), a = \chi_{1-\alpha/2}^2(n-1).$$

根据枢轴变量  $W = (n-1)S^2/\sigma^2$  可得

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

### 10.3.4 双正态总体情形

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$  的样本, 设  $Y_1, Y_2, \dots, Y_m$  是总体  $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$  的样本, 令

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i, \quad S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_2^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2.$$

考虑  $\mu_1 - \mu_2$  和  $\sigma_1^2/\sigma_2^2$  的置信度为  $1 - \alpha$  的区间估计.

1) 已知方差  $\sigma_1^2$  和  $\sigma_2^2$ , 求  $\mu_1 - \mu_2$  的置信区间. 根据正态分布的性质有

$$\bar{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n}\right), \quad \bar{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{m}\right) \quad \bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right),$$

进一步有

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1).$$

于是设枢轴变量

$$W = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1),$$

求解置信区间

$$\Pr \left[ \bar{X} - \bar{Y} - \mu_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + \mu_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right] = 1 - \alpha.$$

2) 若  $\sigma_1^2$  和  $\sigma_2^2$  未知, 但已知  $\sigma_1^2 = \sigma_2^2$ , 设

$$S_W = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2},$$

则考虑枢轴变量

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

于是有

$$\Pr \left[ -t_{\alpha/2}(n+m-2) < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_W \sqrt{\frac{1}{n} + \frac{1}{m}}} < t_{\alpha/2}(n+m-2) \right] = 1 - \alpha.$$

3) 求方差比  $\sigma_1^2/\sigma_2^2$  的置信度为  $1 - \alpha$  的置信区间. 设枢轴变量

$$W = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1),$$

根据  $F$  分布的不对称性, 采用概率对称的区间

$$\Pr[W \leq a] = \Pr[W \geq b] = \alpha/2 \quad \Rightarrow \quad b = F_{\frac{\alpha}{2}}(n-1, m-1), \quad a = F_{1-\alpha/2}(n-1, m-1).$$

由此可得置信区间

$$\Pr \left[ \frac{S_1^2}{S_2^2} \frac{1}{F_{\alpha/2}(n-1, m-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\alpha/2}(n-1, m-1)} \right] = 1 - \alpha.$$

### 10.3.5 单侧置信区间

对某些实际问题, 我们往往只关心置信区间的上限或下限, 例如, 次品率只关心上限, 产品的寿命只关心下限, 由此引入单侧置信区间及其估计.

**定义 10.5 (单侧置信区间)** 给定  $\alpha \in (0, 1)$ , 若样本  $X_1, \dots, X_n$  的统计量  $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \dots, X_n)$  满足

$$\Pr[\theta > \hat{\theta}_1] \geq 1 - \alpha,$$

则称  $(\hat{\theta}_1, +\infty)$  为  $\theta$  的置信度为  $1 - \alpha$  的单侧置信区间,  $\hat{\theta}_1$  称为单侧置信下限.

同理定义单侧置信上限. 对正态总体, 可以将相关置信区间的估计都扩展到单侧置信估计, 枢轴变量的定理类似, 我们将不再重复讨论, 下面仅举两个实例:

**例 10.17** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \mathcal{N}(\mu, \sigma^2)$  的样本, 若方差  $\sigma^2$  已知, 求  $\mu$  的置信度为  $1 - \alpha$  的单侧置信下限和上限.

**解** 设样本均值  $\bar{X} = \sum_{i=1}^n X_i/n$ , 根据正态分布的性质考虑枢轴变量

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

于是有

$$\Pr\left[\frac{\bar{X} - \mu}{S/\sqrt{n}} < \mu_\alpha\right] = 1 - \alpha, \quad \Pr\left[\frac{\bar{X} - \mu}{S/\sqrt{n}} > -\mu_\alpha\right] = 1 - \alpha,$$

整理计算完成估计.

**例 10.18** 从一批出厂的灯泡中随机抽取 10 盏灯泡, 测试其寿命分别为: 1000, 1500, 1250, 1050, 950, 1000, 1150, 1050, 950, 1000, (单位: 小时). 假设这批灯泡的寿命服从正态分布, 求这批灯泡平均寿命的置信度为 95% 的单侧置信下限.

**解** 首先计算样本均值和样本修正方差分别为

$$\bar{X} = \sum_{i=1}^{10} X_i/10 = 1090 \quad \text{和} \quad S^2 = \sum_{i=1}^{10} (X_i - \bar{X})^2/9 = 8800/3.$$

根据正态分布的性质考虑枢轴变量

$$W = \frac{\bar{X} - \mu}{S/3} \sim t(9),$$

于是有

$$\Pr\left[\frac{\bar{X} - \mu}{S/3} < t_{0.05}(9)\right] = 0.95,$$

查表  $t_{0.05}(9) = 1.833$  可得

$$\mu > \bar{X} - t_{0.05}(9)S/3 = 1090 - \sqrt{8800/3} \times 1.833/3 > 1056.$$

### 10.3.6 非正态分布的区间估计

设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本, 若总体  $X$  的分布未知或非正态分布, 我们可以给出总体期望  $\mu = E[X]$  的区间估计, 方法分为两种: 利用 Concentration 不等式和中心极限定理.

- (1) 首先考虑 Concentration 不等式, 若总体  $X \in [a, b]$ , 设  $\bar{X} = \sum_{i=1}^n X_i/n$ , 根据 Concentration 不等式有

$$\Pr[|\mu - \bar{X}| \geq \epsilon] \leq 2 \exp(-2n\epsilon^2/(b-a)^2).$$

令  $\alpha = 2 \exp(-2n\epsilon^2/(b-a)^2)$  求解  $\epsilon = \sqrt{(b-a)^2 \ln(2/\alpha)/n}$ , 于是有

$$\Pr\left[\bar{X} - \sqrt{(b-a)^2 \ln(2/\alpha)/n} < \mu < \bar{X} + \sqrt{(b-a)^2 \ln(2/\alpha)/n}\right] > 1 - \alpha.$$

可基于其它 Concentration 不等式给出类似的置信区间估计, 以及其它 sub-Gaussian 型随机变量的期望的置信区间估计.

- (2) 利用中心极限定理, 求枢轴变量的近似分布, 再给出置信区间估计. 设总体  $X$  的期望  $E(X) = \mu$ , 方差  $\text{Var}(X) = \sigma^2$ , 利用中心极限定理设枢轴变量

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

枢轴变量  $W$  的分布近似于标准正态分布  $\mathcal{N}(0, 1)$ . 当方差  $\sigma^2$  已知时有

$$\Pr\left[-\mu_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \mu_{\alpha/2}\right] \approx 1 - \alpha.$$

当方差  $\sigma^2$  未知时, 用修正样本方差  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$  代替方差  $\sigma^2$ , 于是有

$$\Pr\left[-\mu_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < \mu_{\alpha/2}\right] \approx 1 - \alpha.$$

**例 10.19** 设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim \text{Ber}(p)$  的样本, 求  $p$  的置信度为  $1 - \alpha$  的区间估计.

**解** 根据 Bernoulli 分布的性质有  $X_i \in \{0, 1\}$  以及  $p = E[X]$ , 根据 Chernoff 不等式有

$$\Pr[|\bar{X} - p| > \epsilon p] \leq 2 \exp(-n\epsilon^2/3),$$

设  $\alpha = 2 \exp(-n\epsilon^2/3)$ , 于是有

$$\Pr\left[\bar{X} - \sqrt{3p \ln(2/\alpha)/n} < p < \bar{X} + \sqrt{3p \ln(2/\alpha)/n}\right] \geq 1 - \alpha,$$

最后求解  $p$  的置信区间.

方法二: 根据 Bernoulli 分布的性质有  $E[X] = p$  和  $\text{Var}(X) = p(1-p)$ , 设枢轴变量

$$W = \frac{n\bar{X} - np}{\sqrt{np(1-p)}}$$

根据中心极限定理可知  $W$  近似于标准正态分布  $\mathcal{N}(0, 1)$ . 于是有

$$\Pr \left[ -\mu_{\alpha/2} < \frac{n\bar{X} - np}{\sqrt{np(1-p)}} < \mu_{\alpha/2} \right] \approx 1 - \alpha.$$

最后求解  $p$  的近似置信区间.



## 第 11 章 假设检验(Hypothesis Testing)

根据样本信息来检验关于总体的某个假设是否正确, 此类问题称为 **假设检验问题**, 可分为两类:

- 参数检验问题: 总体分布已知, 检验某未知参数的假设;
- 非参数检验问题: 总体分布未知时的假设检验问题.

假设检验的方法: 先假设所做的假设  $H_0$  成立, 然后从总体中取样, 根据样本的取值来判断是否有‘不合理’的现象出现, 最后做出接受或者拒绝所做假设的决定. ‘不合理’的现象指小概率事件在一次事件中几乎不会发生.

**例 11.1** 某产品出厂检验规定次品率  $p \leq 0.04$  才能出厂, 现从 10000 件产品中任抽取 12 件, 发现 3 件是次品, 问该批产品是否该出厂; 若抽样结果有 1 件次品, 问该批产品是否该出厂?

**解** 首先做出假设  $H_0: p \leq 0.04$ . 若假设  $H_0$  成立, 设随机变量  $X \sim B(12, p)$ ,

$$\Pr[X = 3] = \binom{12}{3} p^3 (1-p)^9 \leq 0.0097.$$

由此可知这是一个小概率事件, 一次试验不应该发生, 但却发生了, 故不合理, 原假设  $H_0: p \leq 0.04$  不成立, 即  $p > 0.04$ , 该批产品不能出厂.

若  $X = 1$  则

$$\Pr[X = 1] = p(1-p)^{11} \binom{12}{1} \geq 0.306.$$

这不是小概率事件, 没理由拒绝原假设  $H_0$ , 产品可以出厂.

注: 当  $X = 1$  情况下, 若直接利用参数估计

$$p = 1/12 = 0.083 > 0.04.$$

若仅仅采用参数估计而不用假设检验, 则不能出厂, 因此参数估计与假设检验是两回事.

在假设检验中, 需要对‘不合理’的小事件给出一个定性描述, 通常给出一上界  $\alpha$ , 当一事件发生的概率小于  $\alpha$  时则成为小概率事件. 通常取  $\alpha = 0.05, 0.1, 0.01$ , 其具体取值根据实际问题而定. 在假定  $H_0$  成立下, 根据样本提供的信息判断出不合理现象 (概率小于  $\alpha$  的事件发生), 则认为假设  $H_0$  不显著,  $\alpha$  被称为显著水平.

注意: 不否定假设  $H_0$  并不是肯定假设  $H_0$  一定成立, 而只能说差异不够显著, 没达到否定的程度, 所以假设检验被称为“显著性检验”.

前面的例子初步介绍了假设检验的基本思想和方法, 下面再进一步说明假设检验的一般步骤:

**例 11.2** 假设某产品的重量服从  $\mathcal{N}(500, 16)$ , 随机取出 5 件产品, 测得重量为 509, 507, 498, 502, 508, 问产品的期望是否正常? (显著性水平  $\alpha = 0.05$ )

**解** 下面给出假设检验的一般步骤:

- 第一步: 提出原假设  $H_0: \mu = 500$  和备择假设  $H_1: \mu \neq 500$ ;
- 第二步: 设计检验统计量, 在原假设  $H_0$  成立下的条件下求出其分布. 令样本均值  $\bar{X} = \sum_{i=1}^5 X_i/5 = 504.8$ , 设检验统计量为

$$Z = \frac{\bar{X} - 500}{\sqrt{16/5}} \sim \mathcal{N}(0, 1).$$

检验统计量能衡量差异大小且分布已知.

- 第三步: 给定显著性水平  $\alpha = 0.05$ , 查表得到临界值  $\mu_{0.025} = 1.96$ , 使得

$$\Pr[|Z| > 1.96] = 0.05$$

成为一个小事件, 从而得到否定域  $\{Z: |Z| > 1.96\}$ .

- 第四步: 将样本值代入计算统计量  $Z$  的实测值

$$|Z| = \frac{|\bar{X} - 500|}{\sqrt{16/5}} = \frac{4.8}{4/\sqrt{5}} = 1.2 \times \sqrt{5} = 2.68 > 1.96.$$

根据实测值  $Z$  落入否定域  $\{Z: |Z| > 1.96\}$ , 从而拒绝原假设  $H_0$ .

由此归纳出假设检验的一般步骤:

- 1) 根据实际问题提出原假设  $H_0$  和备择假设  $H_1$ ;
- 2) 确定检验统计量 (分布已知);
- 3) 确定显著性水平  $\alpha$ , 并给出拒绝域;
- 4) 由样本计算统计量的实测值, 判断是否接受原假设  $H_0$ .

假设检验可分为如下三类:

- 原假设  $H_0: \mu = \mu_0$  和备选假设  $H_1: \mu \neq \mu_0$ , 称为 **双边假设检验**;
- 原假设  $H_0: \mu \leq \mu_0$  和备选假设  $H_1: \mu > \mu_0$ , 称为 **右边检验**;
- 原假设  $H_0: \mu \geq \mu_0$  和备选假设  $H_1: \mu < \mu_0$ , 称为 **左边检验**.

右边检验和左边检验又被通称为双边检验.



下面研究假设检验是否会犯错, 假设检验的核心是先假设原判断假设  $H_0$  成立, 然后根据样本的取值来判断是否有‘不合理’的现象出现, 即“小概率”原理, 然而小概率事件在一次试验中不发生并不意味着小概率事件不发生. 可能发生如下两种错误:

- 第 I 类错误: “弃真”, 即当  $H_0$  为真时, 我们仍可能拒绝  $H_0$ .
- 第 II 类错误: “存伪”, 即当  $H_0$  不成立时, 我们仍可能接受  $H_0$ .

两类错误如下表格所示

假设检验的决定	真实情况: $H_0$ 为真	真实情况: $H_0$ 为假
拒绝 $H_0$	第 I 类错误	正确
接受 $H_0$	正确	第 II 类错误

设犯第 I 类错误的概率为  $\alpha$ , 即显著性水平, 第 II 类错误的概率用  $\beta$  表示, 即

$$\alpha = \Pr[\text{拒绝 } H_0 | H_0 \text{ 为真}] \quad \beta = \Pr[\text{接受 } H_0 | H_0 \text{ 为假}].$$

这两类错误互相关联, 当样本容量固定时, 一类错误概率的减少导致另一类错误概率的增加. Neyman-Pearson 原则: 在控制第 I 类错误的前提下, 尽可能减小第 II 类错误的概率.

## 11.1 正态总体期望的假设检验

### 11.1.1 方差已知的单个正态总体的期望检验 (Z 检验)

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim N(\mu, \sigma^2)$  的样本, 若方差  $\sigma^2$  已知, 检验原假设  $H_0: \mu = \mu_0$  和备择假设  $H_1: \mu \neq \mu_0$ . 设样本均值为  $\bar{X} = \sum_{i=1}^n X_i/n$ , 根据正态分布的性质选择检验统计量

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

给定显著性水平  $\alpha$ , 得到拒绝域为  $|Z| \geq \mu_{\alpha/2}$ , 这种检验方法称为 **Z 检验法**.

关于 Z 检验法的双边和单边检验有

- 原假设  $H_0: \mu = \mu_0$  和备择假设  $H_1: \mu \neq \mu_0$ , 拒绝域为  $\{Z: |Z| \geq \mu_{\alpha/2}\}$ ;
- 原假设  $H_0: \mu \geq \mu_0$  和备择假设  $H_1: \mu < \mu_0$ , 拒绝域为  $\{Z: Z \leq -\mu_{\alpha}\}$ ;
- 原假设  $H_0: \mu \leq \mu_0$  和备择假设  $H_1: \mu > \mu_0$ , 拒绝域为  $\{Z: Z \geq \mu_{\alpha}\}$ .

**例 11.3** 已知某产品的重量  $X \sim \mathcal{N}(4.55, 0.108^2)$ , 现随机抽取 5 个产品, 其质量分别为 4.28, 4.40, 4.42, 4.35, 4.27. 问产品的期望在  $\alpha = 0.05$  下有无显著性变化. ( $\mu_{0.025} = 1.96$ )

**解** 首先提出原假设  $H_0: \mu = 4.55$  和备择假设  $H_1: \mu \neq 4.55$ . 若  $H_0$  成立, 选择检验量

$$Z = \frac{\bar{X} - 4.55}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

求得拒绝域为  $|Z| \geq \mu_{\alpha/2} = 1.96$ . 计算样本均值可知  $\bar{X} = 4.364$ , 于是有

$$\frac{\bar{X} - 4.55}{0.108/\sqrt{5}} = 3.851 > 1.96,$$

由此可拒绝  $H_0$ , 说明有显著变化.

**例 11.4** 某灯泡平均寿命要求不低于 1000 小时被称为‘合格’, 已知灯泡的寿命  $X \sim \mathcal{N}(\mu, 100^2)$ , 现在随机抽取 25 件, 其样本均值为  $\bar{X} = 960$ . 在显著性水平  $\alpha = 0.05$  的情况下, 检验这批灯泡是否合格. ( $\mu_{0.05} = 1.645$ )

**解** 首先提出原假设  $H_0: \mu \geq 1000$  和备择假设  $H_1: \mu < 1000$ . 若  $H_0$  成立, 选择假设统计量

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

由此得到假设拒绝域为:  $Z < -\mu_{\alpha} = -1.645$ . 根据样本均值  $\bar{X} = 960$  可知观察值

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = -2.0 < -1.645$$

由此可拒绝  $H_0$ , 认为这篇灯泡不合格.

### 11.1.2 方差未知的单个正态总体的期望检验 (t 检验)

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim N(\mu, \sigma^2)$  的样本, 若方差  $\sigma^2$  未知, 检验原假设  $H_0: \mu = \mu_0$  和备择假设  $H_1: \mu \neq \mu_0$ . 设样本均值为  $\bar{X} = \sum_{i=1}^n X_i/n$  和样本修正方差  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ , 根据正态分布的性质选择检验统计量

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1).$$

给定显著性水平  $\alpha$ , 得到拒绝域为  $|t| \geq t_{\alpha/2}(n-1)$ , 这种检验方法称为 **t 检验法**.

关于 t 检验法的双边和单边检验有

- i) 原假设  $H_0: \mu = \mu_0$  和备择假设  $H_1: \mu \neq \mu_0$ , 拒绝域为  $\{t: |t| \geq t_{\alpha/2}(n-1)\}$ ;
- ii) 原假设  $H_0: \mu \geq \mu_0$  和备择假设  $H_1: \mu < \mu_0$ , 拒绝域为  $\{t: t \leq -t_{\alpha}(n-1)\}$ ;
- iii) 原假设  $H_0: \mu \leq \mu_0$  和备择假设  $H_1: \mu > \mu_0$ , 拒绝域为  $\{t: t \geq t_{\alpha}(n-1)\}$ .

### 11.1.3 方差已知的两个正态总体的期望差检验

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim N(\mu_1, \sigma_1^2)$  的样本, 以及  $Y_1, Y_2, \dots, Y_m$  是来自总体  $Y \sim N(\mu_2, \sigma_2^2)$  的样本, 若方差  $\sigma_1^2$  和  $\sigma_2^2$  已知, 检验原假设  $H_0: \mu_1 - \mu_2 = \delta$  和备择假设  $H_1: \mu_1 - \mu_2 \neq \delta$

(注:  $\delta$  为常数). 设样本均值  $\bar{X} = \sum_{i=1}^n X_i/n$  和  $\bar{Y} = \sum_{i=1}^m Y_i/m$ , 根据正态分布的性质有

$$U = \frac{\bar{X} - \bar{Y} - \delta}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0, 1).$$

给定显著性水平  $\alpha$ , 其双边和单边检验有

- i) 原假设  $H_0: \mu_1 - \mu_2 = \delta$  和备择假设  $H_1: \mu_1 - \mu_2 \neq \delta$ , 拒绝域为  $\{U: |U| \geq \mu_{\alpha/2}\}$ ;
- ii) 原假设  $H_0: \mu_1 - \mu_2 \geq \delta$  和备择假设  $H_1: \mu_1 - \mu_2 < \delta$ , 拒绝域为  $\{U: U \leq -\mu_{\alpha}\}$ ;
- iii) 原假设  $H_0: \mu_1 - \mu_2 \leq \delta$  和备择假设  $H_1: \mu_1 - \mu_2 > \delta$ , 拒绝域为  $\{U: U \geq \mu_{\alpha}\}$ .

#### 11.1.4 方差未知但相等的两个正态总体的期望差检验

略, 以后补上

#### 11.1.5 基于成对 (pairwise) 数据的检验

在很多实际应用中, 为了比较两种方法或两种产品的差异, 往往会得到一批成对的观察值, 然后基于观察的数据分析判断方法会产品是否具有显著的区别, 这种方法称为 **成对 (pairwise) 比较法**.

**例 11.5** 假设有两种学习方法  $A$  和  $B$ , 在 9 个数据集上取得的效果如下表

数据集	1	2	3	4	5	6	7	8	9
方法 $A$	0.6	0.9	0.8	0.7	0.6	0.9	0.8	0.9	0.7
方法 $B$	0.7	0.95	0.7	0.6	0.7	0.9	0.9	0.8	0.6

问这两种方法在  $\alpha = 0.05$  下是否有显著性区别?

上述问题可进一步形式化为: 假设观察到  $n$  对互相独立的随机变量  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ , 其中  $X_1, X_2, \dots, X_n$  和  $Y_1, Y_2, \dots, Y_n$  分别是总体  $X$  和  $Y$  的两个样本, 检验这两种方法是否性能相同, 即检验总体  $X$  和  $Y$  的期望是否相等. 因为对相同的数据集  $i$  而言,  $X_i$  和  $Y_i$  不能被认为相互独立. 由此假设

$$Z = X - Y \sim \mathcal{N}(\mu, \sigma^2),$$

并提出原假设  $H_0: \mu = 0$  和备择假设  $H_1: \mu \neq 0$ , 方差  $\sigma^2$  未知, 因此考虑统计  $t$  检验量. 设  $Z_i = X_i - Y_i$  ( $i \in [n]$ ), 可得样本均值和方差分别为

$$\bar{Z} = \sum_{i=1}^n \frac{Z_i}{n} \quad \text{和} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2.$$

由此得到统计检验量

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

在显著性水平  $\alpha$  下得到拒绝域为:  $|t| > t_{\alpha/2}(n-1)$ . 下面给出例 11.5 详细求解.

解 设随机变量  $Z_i = X_i - Y_i$  ( $i \in [10]$ ), 可得样本均值  $\bar{Z} = 0.0056$  和方差  $S^2 = 0.009$ , 由此可得观察值

$$|t| = \frac{|\bar{Z}|}{S/\sqrt{n}} = \frac{0.0056}{0.9} \approx 0.062 < t_{0.025}(8) = 2.3060,$$

由此说明这两种方法没有显著性区别.

## 11.2 正态分布的方差假设检验.

### 11.2.1 单个正态总体的方差检验 ( $\chi^2$ 检验)

设  $X_1, X_2, \dots, X_n$  是来自总体  $X \sim N(\mu, \sigma^2)$  的样本, 检验原假设  $H_0: \sigma^2 = \sigma_0^2$  和备择假设  $H_1: \sigma^2 \neq \sigma_0^2$ . 设样本修正方差  $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$ , 根据正态总体抽样定理选择检验统计量

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

给定显著性水平  $\alpha$  求解拒绝域, 这种检验方法称为  $\chi^2$  检验法.

关于  $\chi^2$  检验法的双边和单边检验有

- i) 原假设  $H_0: \sigma^2 = \sigma_0^2$  和备择假设  $H_1: \sigma^2 \neq \sigma_0^2$ , 拒绝域为:  $\{\chi^2 \geq \chi_{\frac{\alpha}{2}}^2(n-1)\} \cup \{\chi^2 \leq \chi_{1-\frac{\alpha}{2}}^2(n-1)\}$ .
- ii) 原假设  $H_0: \sigma^2 \geq \sigma_0^2$  和备择假设  $H_1: \sigma^2 < \sigma_0^2$ . 拒绝域为:  $\{\chi^2 \leq \chi_{1-\alpha}^2(n-1)\}$ .
- iii) 原假设  $H_0: \sigma^2 \leq \sigma_0^2$  和备择假设  $H_1: \sigma^2 > \sigma_0^2$ . 拒绝域为:  $\{\chi^2 \geq \chi_{\alpha}^2(n-1)\}$ .

### 11.2.2 两个正态总体的方差比检验 ( $F$ 检验)

略

## 11.3 非参数假设检验

前面的内容讨论整体分布类型已知 (正态总体) 的参数假设检验问题. 本节讨论总体分布的假设检验问题, 因为所研究的检验是如何利用子样去拟合总体分布, 所以又被称分布的拟合优度检验.

### 11.3.1 $\chi^2$ 检验法

设总体  $X$  的分布函数  $F(x)$  具体形式未知. 根据样本  $X_1, \dots, X_n$  来检验关于总体的假设:

$$H_0: F(x) = F_0(x)$$

其中  $F_0(x)$  为某确定的分布函数.

若总体  $X$  为离散随机变量:  $H_0: \Pr[X = x_i] = p_i$  ( $i = 1, 2, \dots$ )

若总体  $X$  为连续随机变量:  $H_0: X$  的密度函数  $p(x) = p_0(x)$

若  $p_i$  或  $p_0(x)$  包含未知参数, 此时首先用极大似然估计/矩估计估计未知参数.

下面介绍  $\chi^2$  检验法: 将随机试验结果的全体  $\Omega$  分成  $k$  个互不相容的事件  $A_1, A_2, \dots, A_k$ , 且  $\cup_{i=1}^k A_i = \Omega$ . 根据假设  $H_0: F(x) = F_0(x)$  计算概率  $p_i = \Pr(A_i)$ . 对样本  $X_1, \dots, X_n$ , 事件  $A_i$  出现的频率为  $n_i/n$ . 当假设  $H_0$  为真时, 频率  $n_i/n$  与概率  $p_i$  差异不应太大. 基于这种思想, Pearson 构造了检验统计量:

$$W = \sum_{i=1}^K \frac{(n_i - np_i)^2}{np_i}$$

称为 Pearson  $\chi^2$  统计量.

**定理 11.1** 若分布函数  $F_0(x)$  不包含未知参数, 当  $H_0$  为真时 (无论  $H_0$  中的分布属于什么分布), 统计量

$$W = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(k-1)$$

证明超出了本书的范围. 给定显著性水平  $\alpha$ , 若  $W > \chi_\alpha^2(k-1)$  则拒绝  $H_0$ .

**例 11.6** 实验 E 有四种不同的结果  $\{A, B, C, D\}$ . 现进行如下实验: 独立重复实验直到结果 A 发生为止. 记录下抛掷的次数, 如此试验 200 次, 结果如下表. 试问该试验是否为均匀分布?

重复次数	1	2	3	4	$\geq 5$
频数	56	48	32	28	36

**解** 首先提出原假设  $H_0$ : 均匀分布. 用随机变量  $X$  表示试验结果 A 发生时重复的试验次数, 有

$$p_1 = P(X=1) = \frac{1}{4} \quad p_2 = P(X=2) = \frac{3}{4} \times \frac{1}{4} \quad p_3 = P(X=3) = \left(\frac{3}{4}\right)^2 \cdot \frac{1}{4}$$

$$p_4 = P(X=4) = \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4} \quad p_5 = P(X=5) = 1 - \frac{1}{4} - \frac{3}{16} - \left(\frac{3}{4}\right)^3 \cdot \frac{1}{4}$$

计算检验统计量

$$W = \sum_{i=1}^5 \frac{(n_i - np_i)^2}{np_i} = 18.21$$

根据统计量实值  $W > \chi_{0.05}^2(4) = 9.488$ , 因此不服从均匀分布.

上例指定了分布的具体分布形式. 在许多实际问题中, 假设  $H_0$  只确定了总体分布的类型, 分布中还包含未知参数, 如

$$H_0: F(x) = F_0(x; \theta_1, \theta_2, \dots, \theta_r)$$

其中  $F_0$  已知,  $\theta_1, \theta_2, \dots, \theta_r$  未知. 从样本  $X_1, X_2, \dots, X_n$  中得到估计值  $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r$ , 代入得

$$H_0: F(x) = F_0(x; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r)$$

将子样分成  $k$  组:  $a_0 < a_1 < \dots < a_k$  且  $A_1 \in [a_0, a_1], A_2 \in [a_1, a_2] \dots A_k = [a_{k-1}, a_k]$ . 总体  $X$  落入  $A_i$  的概率为

$$\hat{p}_i = p(x \in A_i | \hat{\theta}_1 \dots \hat{\theta}_r)$$

检验估计量  $W$  为

$$W = \sum_{i=1}^k \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

**定理 11.2** 当  $n \rightarrow +\infty$  时, 有  $W \xrightarrow{d} \chi^2(k-r-1)$  成立.

### 11.3.1.1 独立性检验

设  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$  是总体  $(X, Y)$  的样本, 通过样本考虑二元总体  $(X, Y)$  中随机变量  $X$  与  $Y$  的独立性. 将随机变量  $X$  和  $Y$  的取值分成  $r$  个和  $s$  个互不相交的区间  $A_1, A_2, \dots, A_r$  和  $B_1, B_2, \dots, B_s$ . 用  $n_{ij}$  表示落入区域  $A_i \times B_j$  的频数. 设  $n_{i\cdot} = \sum_{j=1}^s n_{ij}$  和  $n_{\cdot j} = \sum_{i=1}^r n_{ij}$  为边缘之和, 则  $n = \sum_{i,j} n_{ij}$ . 建立如下二元联立表:

	$B_1$	$B_2$	$\cdots$	$B_s$	$n_{i\cdot}$
$A_1$	$n_{11}$	$n_{12}$	$\cdots$	$n_{1s}$	$n_{1\cdot}$
$A_2$	$n_{21}$	$n_{22}$	$\cdots$	$n_{2s}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$A_r$	$n_{r1}$	$n_{r2}$	$\cdots$	$n_{rs}$	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	$n$

首先提出假设  $H_0: X$  与  $Y$  相互独立. 记

$$p_{ij} = \Pr(X \in A_i, Y \in B_j) \quad p_{i\cdot} = P(X \in A_i) = \sum_{j=1}^s p_{ij} \quad p_{\cdot j} = P(Y \in B_j) = \sum_{i=1}^r p_{ij}$$

若假设  $H_0$  成立, 则  $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ . 利用矩估计/最大似然估计得

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

设计假设检验统计量

$$W = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}} = n \sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - n \sim \chi^2((r-1)(s-1))$$

在显著性水平为  $\alpha$  时有  $W \sim \chi^2((r-1)(s-1))$  成立, 由此得到拒绝域为:  $W > \chi_{\alpha}^2((r-1)(s-1))$ , 即在此范围内不接受随机变量  $X$  与  $Y$  独立.

## 习题

11.1 设随机变量  $X$  的期望  $E[X] = \mu > 0$ , 方差为  $\sigma^2$ , 证明对任意  $\epsilon > 0$  有

$$P(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

