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

对自然和社会所发生的各种现象或问题进行观察, 会发现在一定条件下很多结果是必然发生的, 称之为 **必然现象**, 又称 **确定性现象**. 例如, 成熟的苹果从树上掉落下来; 在标准大气压下, 水在 0°C 以下会结冰, 加热到 100°C 以上会沸腾; 平面上三角形两边之和大于第三边; 等等. 必然现象发生的条件与结果之间具有确定性关系.

然而在实际生活中往往也会面对大量带有不确定性的现象. 例如, 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; 当你乘坐公交车时, 需在站台等待公交车多长的时间; 今晚的夜空能否观察到流星; 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, 而事先不确定哪种结果会出现, 称之为 **随机现象**, 随机现象发生的条件与结果之间具有不确定性关系.

随机现象发生的条件和结果之间具有不确定性联系, 无法通过确切的数学函数来刻画. 尽管在一次观察中随机现象无法确定哪种结果发生, 具有一定的偶然性; 然而在大量重复实验和观察下, 随机现象的结果却具有一定的规律性. 例如, 多次重复随机投掷一枚硬币得到的正面/反面朝上数几乎相同; 公交车的等待时间按照一定的规律; 等等. 因而随机现象具有二重属性:

- **偶然性**: 对随机现象进行一次观察, 其结果具有不确定性;
- **必然性**: 对随机现象进行大量重复观察, 其结果呈现一定的统计规律性.

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, 其应用几乎遍及所有科学技术领域、行业生产、国民经济与生活等. 正如法国著名数学家拉普拉斯 (Laplace, 1794-1827) 所言：“对生活的大部分, 最重要的问题实际上只是概率问题”, 图灵奖得主 Y. LeCun 在其自传中指出：“历史上多数研究成果的出现是偶然事件... 所有努力都是为了提升概率”. 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, 称之为 **随机试验**, 简称为 **试验**. 一般用 E 或 E_1, E_2, E_3, ... 表示, 本书所提及的试验均是随机试验.

下面给出一些例子:

E_1: 随意抛一枚硬币, 观察正面/反面朝上的情况.

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同的特点：每次试验的所有可能结果已知, 如抛硬币有正面/反面朝上两种结果, 实验可以在相同的条件下重复地进行, 在实验之前不确定出现那种结果. 概况起来, 随机试验具有以下三个特点:

- **可重复**: 可在相同的条件下随机试验可重复进行;
- **多结果**: 试验的结果不唯一, 所有可能发生的结果事先明确可知;
- **不确定**: 试验前无法预测/确定哪一种结果会发生.

1.1.1 随机事件

尽管随机试验在试验前不能确定试验的结果, 但其所有可能发生的结果事先是可知的. 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 Ω. 样本空间 Ω 的每个元素, 即试验 E 的每一种结果, 称为 **样本点**, 记为 ω.

例如在前面所述的试验中,

试验 E_1 的样本空间为 $\Omega_1 = \{$正面, 反面$\}$, 样本点分别为 ω_1=正面, ω_2=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 $\omega_1 = 1$, $\omega_2 = 2$, …, $\omega_6 = 6$.

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

试验 E_4 的样本空间为 $\Omega_4 = \{t\colon t \geqslant 0\}$, 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 Ω_1 和 Ω_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 Ω_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 Ω_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 $A, B, C, \ldots$ 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 Ω 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**. 样本空间 Ω 包含所有样本点, 是其自身的子集, 每次试验必然发生, 因而称事件 Ω 为 **必然事件**. 空集 $\emptyset$ 不包含任意样本点, 也是样本空间的子集, 在每次试验中均不发生, 称空集 $\emptyset$ 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

$$\Omega = \{1, 2, \ldots, 6\}.$$

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其运算可类比于集合论的关系和运算. 下面默认随机试验的样本空间为 Ω, 而 A, B, A_i $(i = 1, 2, \ldots)$ 表示样本空间 Ω 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
$$A \cup B = \{\omega \colon \omega \in A \text{ 或 } \omega \in B\}.$$
类似地, 事件 $A_1, A_2, \ldots, A_n$ 中至少有一个发生所构成的事件称为事件 $A_1, A_2, \ldots, A_n$ 的并事件, 记为
$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n = \{\omega \colon \exists\, i \in [n] \text{ s.t. } \omega \in A_i\}.$$
称 $\bigcup_{i=1}^{\infty} A_i$ 为可列个事件 $A_1, A_2, \cdots$ 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
$$A \cap B = \{\omega \colon \omega \in A \text{ 且 } \omega \in B\}.$$
类似地, 事件 $A_1, A_2, \ldots, A_n$ 同时发生所构成的事件称为事件 $A_1, A_2, \ldots, A_n$ 的交事件, 记为
$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n = \{\omega \colon \forall i \in [n] \text{ s.t. } \omega \in A_i\}.$$
称 $\bigcap_{i=1}^{\infty} A_i$ 为可列个事件 $A_1, A_2, \cdots$ 的交事件.

4) **事件的差** 若事件 A 发生而同时事件 B 不发生所构成的事件称为 **事件 A 与 B 的差**, 记为 $A - B$, 即
$$A - B = A - AB = A\bar{B} = \{\omega \colon \omega \in A \text{ 且 } \omega \notin B\}.$$

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** (或 **逆**) **事件**, 记为 $\bar{A}$, 即 $\bar{A} = \Omega - A$. 容易得到 $\bar{A} \cap A = \emptyset$ 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, 则称事件 A 和 B 是 **互不相容** (或 **互斥**) **的**, 即
$$A \cap B = \emptyset.$$
注意: 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. 例如基本事件是互不相容的, 但超过 2 个样本点的样本空间中基本事件不是对立事件.

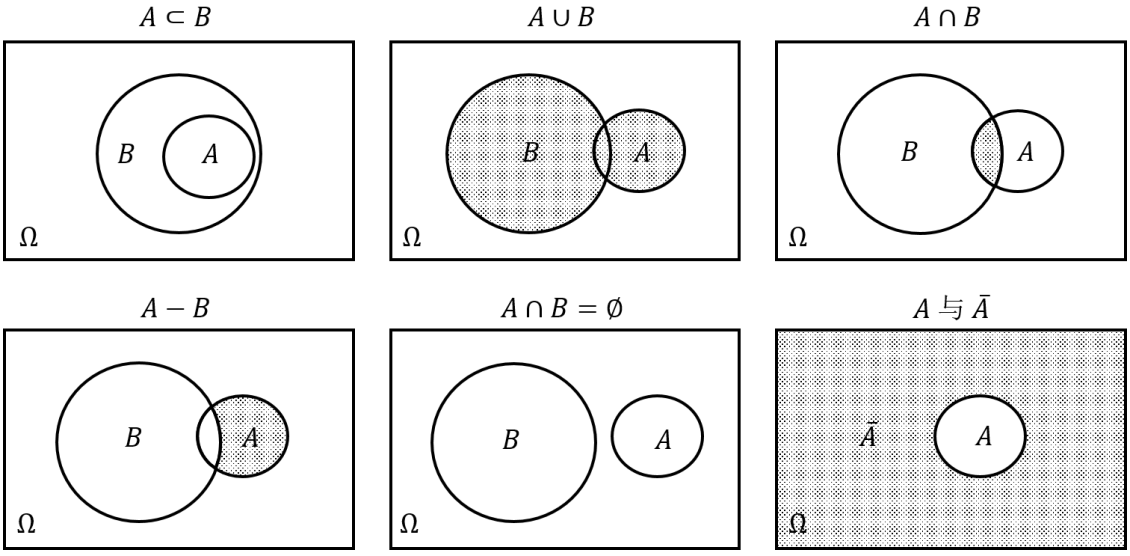


图 1.1 事件关系或运算通过 Venn 图表示, $A\cup B$, $A\cap B$, $A-B$, $\bar{A}$ 分别为阴影部分

借助集合论的 Venn 图, 事件之间的关系或运算可用图 1.1 表示. 例如, 在 $A\subset B$ 的图示中, 矩形表示样本空间 Ω, 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 $A\cup B$.

根据前面的定义, 可以发现事件还满足下面的规律, 相关证明读者可参考集合的运算.

- 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: $(A\cup B)\cap C=(A\cap C)\cup(B\cap C)$, $(A\cap B)\cup C=(A\cup C)\cap(B\cup C)$;
- 对偶律: $\overline{A\cup B}=\bar{A}\cap\bar{B}$, $\overline{A\cap B}=\bar{A}\cup\bar{B}$, 对偶律又称德 · 摩根 (De Morgen) 律.

若事件 $A\subset B$, 有 $AB=A$ 和 $A\cup B=B$. 上述四条规律对有限个或可列个事件均成立, 例如, 对偶律满足

$$\overline{\bigcup_{i=1}^{n}A_i}=\bigcap_{i=1}^{n}\overline{A_i},\qquad \overline{\bigcap_{i=1}^{n}A_i}=\bigcup_{i=1}^{n}\overline{A_i}.$$

例 1.2 设 A,B,C 为任意三个随机事件, 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 为任意三个随机事件, 证明

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

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 $((A - B) \cup (B - C)) \subset ((A \cup B) - BC)$, 另一方面只需证明 $(A\bar{C}) \subset ((A\bar{B}) \cup (B\bar{C}))$, 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.

事件间的关系类比于运算和集合间的关系与运算, 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 简要地给出了概率论和集合论相关概率的对应关系.

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
Ω	必然事件, 样本空间	全集
$\emptyset$	不可能事件	空集
ω	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 ω 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

1.2 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, 我们首先引入频率, 用以描述随机事件发生的频繁程度, 然后介绍刻画随机事件发生可能性大小的数, 即事件的概率.

1.2.1 频率

定义 1.1 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**.

直观而言, 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, 反之亦然. 根据上述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 Ω 有 $f_n(\Omega) = 1$;

3° 对 k 个互不相容的事件 $A_1, A_2, \ldots, A_k$ 有

$$f_n(A_1 \cup A_2 \cup \cdots \cup A_k) = f_n(A_1) + f_n(A_2) + \cdots + f_n(A_k).$$

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

频率在实际中往往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.

尽管频率会随着试验次数 n 的变化表现出一定的随机性, 但在大量重复的试验中, 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, 历史上多人进行重复投掷硬币的试验, 下面给出了一些人的试验统计结果:

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

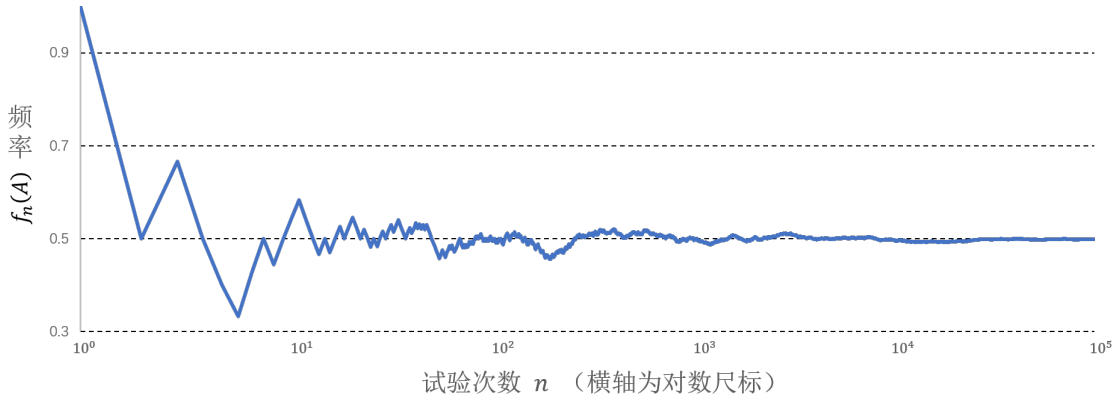


图 1.2 任意投掷硬币, 正面朝上频率的趋势

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 给出了实验结果. 这些实验结果均表明, 尽管对不同的投掷总数, 正面朝上的频率不并相同, 但随着投掷次数的增加, 正面朝上的频率越来越接近一个常数 (0.5), 即频率逐渐稳定于 0.5. 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 1.2 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 $P(A) = p$.

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发生的概率的一种方法, 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. 然而概率的统计定义存在数学上的不严谨性, 而在实际中几乎不可能每一个事件做大量重复的试验来计算频率, 进而近似概率, 以此刻画事情发生的可能性. 受到频率的稳定性及其性质的启发, 下一节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, 期望建立媲美于欧氏几何公理化的理论体系.

定义 1.3 (概率公理化定义) 随机试验 E 所对应的样本空间 Ω 中每一个随机事件 A, 均赋予一实数 $P(A)$, 且满足以下条件:

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

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

3° **可列可加性**: 若 $A_1, A_2, \ldots, A_n, \ldots$ 是可列无穷个互不相容的事件, 即 $A_iA_j = \emptyset$ $(i \neq j)$, 有

$$P(A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_n) + \cdots ;$$

则称 $P(A)$ 为随机事件 A 的 **概率**.

由定义可知概率 $P(A)$ 是在样本空间 Ω 中所有随机事件做构成的集合上定义的实值函数. 概率的三条公理 (非负性、规范性和可列可加性) 简明扼要地刻画了概率的定义, 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.

在后续章节中将证明当 $n \to \infty$ 时频率 $f_n(A)$ 在一定意义下接近概率 $P(A)$, 因此用概率 $P(A)$ 刻画事件 A 发生的可能性大小是合理的. 我们将根据概率的三条公理化定义, 推导概率的重要性质.

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

证明 令 $A_i = \emptyset\ (i = 1, 2, \ldots)$, 则有 $\emptyset = \cup_{i=1}^{\infty} A_i$ 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 $P(\emptyset) = 0$.

注: 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 Ω 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连续函数或几何概型在一个样本点处的概率.

性质 1.2 (有限可加性) 若 $A_1, A_2, \cdots, A_n$ 是两两不相容事件, 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 $A_1, A_2, \cdots, A_n, A_{n+1}, \ldots$ 是两两互不相容事件. 根据公理 3° 可知

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

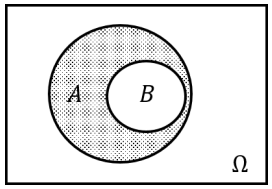
性质得证.

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 $P(B) \leqslant P(A)$ 和 $P(A - B) = P(A) - P(B)$.

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



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

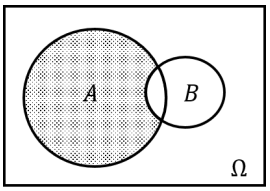
再根据公理 1° 有 $P(A - B) = P(A) - P(B) \geqslant 0$, 从而得到 $P(A) \geqslant P(B)$.

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 Inclusion-Exclusion Principle) 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 $A_1,A_2,\ldots,A_n$ 有

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

对 n 个随机事件 $A_1,A_2,\ldots,A_n$ 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) 对事件 $A_1,A_2,\ldots A_n$ 有

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

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

$$P(A\cup B)=P(A)+P(B)-P(AB)\leqslant P(A)+P(B). \tag{1.1}$$

假设当 $n=k$ 时性质成立, 对 $n=k+1$ 有

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

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推得到 Bonferroni 不等式: 对事件 $A_1,A_2,\ldots A_n$ 有

$$\begin{aligned}P\left(\bigcup_{i=1}^n A_i\right)&\leqslant\sum_{i=1}^n P(A_i);\\P\left(\bigcup_{i=1}^n A_i\right)&\geqslant\sum_{i=1}^n P(A_i)-\sum_{i<j}P(A_iA_j);\\P\left(\bigcup_{i=1}^n A_i\right)&\leqslant\sum_{i=1}^n P(A_i)-\sum_{i<j}P(A_iA_j)+\sum_{i<j<k}P(A_iA_jA_k);\end{aligned}$$

可依次类推相关不等式.

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), 根据事件的对偶律有

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

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), 根据对偶律与容斥原理有

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

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

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) \\ &= 3/4 - 1/8 + P(ABC). \end{aligned}$$

根据 $P(AB) = 0$ 和 $ABC \subset AB$ 可知

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

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 1.4 (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 $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, 其中 w_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 $\{\omega_{i_1}, \omega_{i_1}, \ldots, \omega_{i_k}\}$, 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.

在使用古典概型计算概率时需注意每个基本事件发生的可行性大小是否相等, 例如,

例 1.6 在相同条件下连续两次抛一枚均匀硬币, 此试验观察的结果: A) 两正面, B) 两反面, C)一正一反. 根据古典概型可知

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

然而这种结论不正确, 因为这三个事件发生的可能性不同. 正确的理解是事件 $C=\{C_1,C_2\}$, 其中 C_1 表示先正后反的事件, C_2 表示先反后正的事件, 从而有

$$P(A)=P(B)=P(C_1)=P(C_2)=1/4.$$

古典概率计算的本质是计数 (Counting), 下面介绍一些基本的计数原理, 更为详细的计数方法将在下一节介绍. 首先介绍计数的两条基本原理:

- **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.

下面介绍无放回的排列组合, 更为复杂的排列组合将在下一节介绍:

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 $(a+b)^n=\sum_{r=0}^{n}\binom{n}{r}a^rb^{n-r}$ 中项 a^rb^{n-r} 的系数.

例 1.7 将 n 个不同的球随机放入 N $(N\geqslant n)$ 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 只不同的球随机放入 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

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

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

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

很多实际问题与上述例子具有相同的数学模型, 如经典的生日问题:

例 1.8 (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, $P(A) = 99.99\%$.

例 1.9 设一批 N 件产品中有 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

$$P(A) = \binom{M}{k}\binom{N-M}{n-k} \Big/ \binom{N}{n}. \tag{1.2}$$

上例是古典概型中一个典型问题, 其概率 (1.2) 称为 **超几何概率**, 在产品质量检测等方面广泛应用. 在例 1.9 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

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

下面分析抽签的先后顺序是否会对抽签的概率产生影响.

例 1.10 (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 ($i \leqslant k$) 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

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

由此例可知第 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. 在上例中, 袋中有 a 个不同的白球和 b 个不同的红球, 或若

k 个人依次随机有放回地从袋中取一个球, 则第 i 个人 $(i \leqslant k)$ 取出红球的概率又是多少? 这里留给读者进一步思考.

在计算概率的过程中, 有时可适当利用概率的性质, 例如,

例 1.11 从 $\{1,2,\ldots,9\}$ 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 $A=\{$取出 n 个整数的乘积能被 10 整除$\}$, $B=\{$取出的 n 个数中有偶数$\}$, $C=\{$取出的 n 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

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

同理可得

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

根据概率的性质有

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

例 1.12 (Matching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 $i_1,i_2,\cdots,i_r$ 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 $\sum_{i_1<\cdots<i_r}P(A_{i_1}\cdots A_{i_r})$ 中共有 $\binom{n}{r}$ 项, 由此可得

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

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 $P(A) = 1 - 1/e = 0.632$.

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, 在很多实际应用中受到了限制. 本节介绍另一种特殊的随机现象, 具有如下特征:

- **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 1.5 在一个测度有限的区域 Ω 内等可能性投点, 落入 Ω 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 1.13 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

$$\Omega = \{x \colon 0 < x \leqslant l\}.$$

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

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

可知事件 B 发生的概率小于或等于 20%, 即

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

求解可得 $l \leqslant 25$.

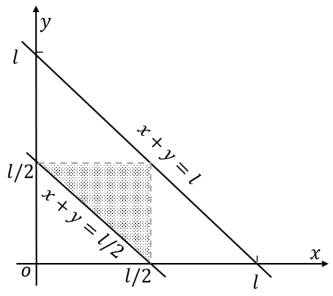
例 1.14 将一根长度为 l 的木棍随意折成三段, 这三段能构成平面三角形的概率是多少?

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

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

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

$$\begin{aligned} A &= \{(x,y)\colon x+y>l-x-y, 1-y>y, l-x>x\} \\ &= \{(x,y)\colon x+y>l/2, y<l/2, x<l/2\}. \end{aligned}$$



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 1.15 (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

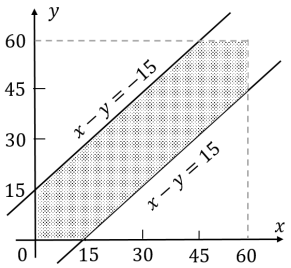
$$\Omega = \{(x,y)\big|0 \leqslant x, y \leqslant 60\}.$$

用 A 表示两人见面的事件, 则

$$A = \{(x,y)\big||x-y| \leqslant 15\} = \{(x,y)\big|x-y \leqslant 15 \text{ 且 } x-y \geqslant -15\}.$$

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 1.15 的概率, 伪代码如下:

$n_A \leftarrow 0$
For $i = 1 : N$
    $x \leftarrow \text{Random}(0, 60)$
    $y \leftarrow \text{Random}(0, 60)$
    If $|x - y| \leqslant 15$ then
        $n_A \leftarrow n_A + 1$
    Endif
Endfor
Return n_A/N

在上述程序中取 N 是一个很大的数, 通过计算机程序近似计算可得两人的见面概率.

1.4 组合计数*

在古典概型中, 概率的计算往往都与组合计数密切相关, 同时组合计数在人工智能、计算机等领域具有广泛的应用, 因此本节将简要地介绍经典的组合计数: 十二重计数 (The twelvefold way). 该计数由著名组合学大师 G.-C. Rota (1932-1999) 首先提出, 最初的问题表述为在满足一定条件下的两个集合之间函数映射的个数.

为可读性起见, 这里采用更为简洁的表述: 将 n 只不同 (或相同) 的球放入 m 个不同 (或相同) 的箱子, 考虑在无任何限制、或每个箱子至多放一球、或每个箱子至少放一球这三种条件下有多少种不同的放法. 下表首先给出相应的计数结果, 后面将一一说明.

表 1.3 将 n 只球放入 m 个箱子, 在三种条件下各有多少种不同的放法.

n 只球	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	$\sum_{k=1}^{m} S(n,k)$	$\begin{cases} 1 & n \leqslant m \\ 0 & n > m \end{cases}$	$S(n,m)$
相同	相同	$\sum_{k=1}^{m} p(n,k)$	$\begin{cases} 1 & n \leqslant m \\ 0 & n > m \end{cases}$	$p(n,m)$

1.4.1 排列、环排列、组合与多重组合

在古典概型中, 我们简要地介绍了排列, 即从 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.

若从 n 个不同的元素中无放回地取出 r 个元素排成一个圆环, 称为 **环排列**.

画图

画图

画图

按照顺时针看: 环排列 a-b-c-a, b-c-a-b 和 c-a-b-c 是同一个环排列, 而 a-c-b-a 则为不同的环排列. 因此对从 n 个不同的元素中无放回地取出 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

定义 1.6 若从 n 个不同的元素中无放回地取出 r 个元素排成一个圆环, 有 $(n)_r/r$ 种不同的排法, 称为 **环排列数**, 特别地, n 个不同元素的环排列数为 $(n-1)!$.

在前一节简要地介绍了组合数, 即从 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. 这里给出一些关于组合数的恒等式, 其证明将作为练习题.

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

下面将组合的概念进行推广到多重组合.

定义 1.7 将 n 个不同的元素分成 k 组, 组内元素无顺序关系, 每组分别有 $r_1, r_2, \cdots, r_k$ 个元素, 即 $n=r_1+\cdots+r_k$ 且 $r_1, r_2, \cdots, r_k$ 为正整数, 则有

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

种不同的分组方法, 称 $\binom{n}{r_1,r_2,\ldots,r_k}$ 为 **多重组合数**.

根据上述定义可知组合数本质上也属于多重组合数, 即 $\binom{n}{r}=\binom{n}{r,n-r}$. 可以得到多重组合数与多项式系数的关系:

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

以前研究集合的元素都是互不相同的, 我们这里引入多重集的概念.

定义 1.8 若集合中的元素是可以重复的, 且重复的元素之间不可分辨, 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 $r_1, r_2, \cdots, r_k$, 即 $n=r_1+r_2+\cdots+r_k$. 若将此多重集 A 中的所有元素排列成一排, 则相当于从 n 个位置中选取出 r_1 个位置放第一类元素, 再从剩下的从 $n-r_1$ 个位置中选取出 r_2 个位置放第二类元素, $\cdots$, 从最后 r_k 个位置放第 k 类元素. 因此该多重集 A 有

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

种不同的排列方法, 即多重组合数.

根据排列组合数, 有如下结论:

n 只球	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 只完全相同、不可分辨的球放入 m 个不同的箱子, 此时有多少种不同的放法. 鉴于球完全相同且不可分辨, 一种放的结果可描述为第一个箱子有 x_1 个球, 第二个箱子有 x_2 个球, $\cdots$, 第 m 个箱子有 x_m 个球, 这里 $x_1, x_2, \cdots, x_m$ 是非负的整数, 并满足

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

因此将 n 只相同的球放入 m 个不同的箱子等价于上述方程的非负整数解. 关于此问题有如下定理:

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

证明 将通过构造一一对应关系给出组合证明. 将 n 只相同的球分别对应于 n 个完全相同且不可分辨的圆圈 '$\circ$', 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, 如下例所示:

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

从左向右看, 第一条竖线之前圆圈的个数用 x_1 表示, 第 i 条竖线与第 $i-1$ 条竖线之间圆圈的个数用 x_i 表示 $(2 \leqslant i \leqslant m)$. 由此可知方程 $x_1 + x_2 + \cdots + x_m = n$ 与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, 即为方程 $x_1 + x_2 + \cdots + x_m = n$ 非负整数解的个数.

由此定理可知方程 $x_1 + x_2 + x_3 = 10$ 有 66 种不同的非负整数解, 将 10 个相同的球放入 3 个不同的箱子有 66 种不同的放法.

对该问题可进一步推广, 例如将 n 只相同的球放入 m 个不同的箱子, 每个箱子至少放入一球, 则有多少种不同放法, 该问题等价于方程 $x_1 + x_2 + \cdots + x_k = n$ 有多少种不同的正整数解, 我们有如下推论:

推论 1.1 方程 $x_1 + x_2 + \cdots + x_m = n$ $(m \leqslant n)$ 有 $\binom{n-1}{m-1} = \binom{n-1}{n-m}$ 个不同的正整数解.

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.

同理还可以研究一些不等式的非负整数解、正整数解的个数, 相关求解将作为练习题.

例 1.16 求不等式 $x_1+x_2+\cdots+x_m\leqslant n$ 有多少种不同的非负整数解、正整数解.

例 1.17 在多项式 $(x_1+x_2+\cdots+x_m)^n$ 的展开式中, 一共有多少种不同的展开项?

解 根据多项式的展开式有

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

因此不同的展开项即为不同的项及其次数, 与 $x_1+x_2+\cdots+x_m=n$ 的非负整数解一一对应, 由此可知多项式 $(x_1+x_2+\cdots+x_m)^n$ 有 $\binom{n+m-1}{m-1}$ 种不同的展开项.

根据整数的有序分解有如下结论:

n 只球	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 只不同的球放入 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 个非空的子集的划分数, 即第二类 Stirling 数:

定义 1.9 将 n 个不同的元素分成 m 个非空的子集, 不同的划分数称为 **第二类 Stirling 数**, 记为 $S(n,m)$.

我们首先以集合 $\{1,2,3\}$ 为例, 讨论不同的划分数:

- 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 $S(3,3)=1$.

遵从惯例设 $S(0,0)=1$. 当 $n\geqslant 1$ 时有 $S(n,n)=S(n,1)=1$ 和 $S(n,0)=0$. 当 $m>n\geqslant 1$ 时有 $S(n,m)=0$. 针对更一般情况, 第二类 Stirling 数有如下的递推关系:

定理 1.2 对 $n \geqslant 1, m \geqslant 1$ 有

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

证明 根据定义可知将集合 $\{1,2,\ldots,n\}$ 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以分成两种情况考虑:

- 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 $S(n-1,m)$ 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, 并利用归纳法证明可得

推论 1.2 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 只球	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	$\sum_{k=1}^{m} S(n,m)$	$\begin{cases}1 & n \leqslant m\\ 0 & n > m\end{cases}$	$S(n,m)$

1.4.4 正整数的无序分拆 (Partition)

本节研究将 n 只相同的球放入 m 个相同的箱子, 即球与箱子均是完全相同不可分辨, 只能通过箱子内不同的球的个数进行区别. 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 1.10 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 $p(n,m)$.

这里以正整数 7 为例考虑的无序划分数 $p(n,m)$ 的情况, 如下表:

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

通过观察发现: 正整数 n 划分成 m 个无序的正整数, 等价于下面方程的正整数解

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

遵从惯例设 $p(0,0)=1$. 当 $n\geqslant 1$ 时有 $p(n,n)=p(n,1)=1$ 和 $p(n,0)=0$. 当 $m>n\geqslant 1$ 时有 $p(n,m)=0$. 对更一般情况, 有如下递推关系:

性质 1.8 对 $n\geqslant 1$ 和 $m\geqslant 1$ 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. 对其中任意一种划分 $x_1+x_2+\cdots+x_m=n$ $(x_1\geqslant x_2\geqslant\cdots\geqslant x_m\geqslant 1)$, 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 $p(n,m)=p(n-1,m-1)+p(n-m,m)$.

对任何一种划分 $x_1+x_2+\cdots+x_m=n$ $(x_1\geqslant x_2\geqslant\cdots\geqslant x_m\geqslant 1)$, 设 $y_j=x_j-1$, 则有

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

考虑 $y_1,y_2,\cdots,y_m$ 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

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

性质 1.9 对整数 $n \geqslant 1$ 和 $m \geqslant 1$ 有

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

给定整数 $m \geqslant 1$, 当 n 非常大或 $n \to \infty$ 有

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

根据正整数的无序分拆, 有

n 只球	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	$\sum_{k=1}^{m} p(n,k)$	$\begin{cases} 1 & n \leqslant m \\ 0 & n > m \end{cases}$	$p(n,m)$

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 个事件的对偶律, 即对任意 n 个事件 A_1,A_2,...,A_n 有

$$\overline{\bigcup_{i=1}^{n} A_i}=\bigcap_{i=1}^{n}\overline{A_i} \qquad \text{和} \qquad \overline{\bigcap_{i=1}^{n} A_i}=\bigcup_{i=1}^{n}\overline{A_i}.$$

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 $\bar{A}B$, $\bar{A}\cup\bar{B}$, $A\cup B\cup C$, $\bar{A}\bar{B}\bar{C}$, $\bar{A}\bar{B}C$ 和 $(\bar{A}\bar{B})\cup C$.

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, 并说明在怎么样的情形下取得.

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,...,A_n 有

$$P(\bigcup_{i=1}^{n} A_i)=\sum_{i=1}^{n}P(A_i)-\sum_{i<j}P(A_iA_j)+\sum_{i<j<k}P(A_iA_jA_k)+...+(-1)^{n-1}P(A_1,A_2,...,A_n).$$

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 有 m 只相同或不同的白球和 n 只相同或不同的红球, 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).

1.14 将 3 只不同的球放入 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.

1.15 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 ($i \leqslant k$) 取出红球的概率是多少; 若为任意无放回地取球, 第 i 个人 ($i \leqslant k$) 取出红球的概率又是多少.

1.16 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.

1.17 在区间 $[0, 1]$ 内随机取两数, 求两数之积小于 $1/4$ 的概率.

1.18 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

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

发生的概率 (要求写出伪代码以及概率保留小数点后 5 位).

1.19 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.

1.20 对正整数 m, n 以及 $r < n$, 证明:

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

1.21 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.

1.22 求方程 $x_1 + x_2 + ... + x_k \leqslant n$ 的正整数解、非负整数解的个数 (n 为正整数).

1.23 求方程 $x_1 + x_2 + ... + x_k < n$ 的正整数解、非负整数解的个数 (n 为正整数).

1.24 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

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, 它是概率论中非常重要的概念之一. 条件概率可以帮助我们更好地分析和理解复杂的随机事件, 同时也有助于复杂事件概率的计算.

2.1 条件概率

前一章所讨论的概率均是在整个样本空间上进行, 无任何其它的条件或限制因素. 然而在很多情况下我们往往需要考虑在一定条件下某一随机事件发生的概率. 首先来看一个直观的例子.

例 2.1　随意掷一枚骰子观察点数, 其样本空间 $\Omega = \{1, 2, \cdots, 6\}$, 用事件 B 表示观察到 3 点, 根据古典概型可知 $P(B) = 1/6$. 用事件 A 表示观察到奇数点, 则有事件 $A = \{1, 3, 5\}$, 根据古典概型可知 $P(A) = 1/2$.

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

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

用事件 C 表示观察到 2 点, 同理可知 $P(C) = 1/6$, 以及在事件 A 发生的情况下事件 C 发生的概率

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

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. 此外还可以通过观察发现:

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

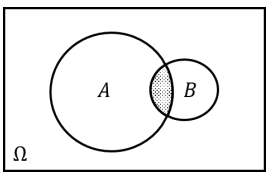
针对一般情形, 下面给出条件概率的形式化定义:

定义 2.1　设 A 和 B 为样本空间下 Ω 的随机事件, 且 $P(A) > 0$, 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**.

考虑事件 A 发生的条件下考虑事件 B 发生的概率, 可将 A 看作新的样本空间, 即缩减的样本空间, 由此可知条件概率的本质是缩小了有效的样本空间, 由此给出了计算条件概率的一种方法, 空间缩减法, 在



后面的例子中会具体介绍. 此外, 根据上述定义可知任何随机事件的概率可以看作必然事件下的条件概率, 即 $P(A)=P(A)/P(\Omega)$.

对任何给定的事件 A 满足 $P(A)>0$, 其条件概率 $P(\cdot|A)$ 满足概率定义的三条公理:

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

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

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

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

根据概率和条件的定义可知 $P(B|A)=P(AB)/P(B)\geqslant 0$ 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 $B_1,B_2,\cdots,B_n,\cdots$ 是两两互不相容的, 则可列个事件 $AB_1,AB_2,\cdots,AB_n,\cdots$ 也是两两互不相容的, 根据分配律有

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

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. 下面继续研究条件概率的其它性质:

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

$$P(B_1\cup B_2|A)=P(B_1|A)+P(B_2|A)-P(B_1B_2|A).$$

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

$$P((B_1\cup B_2)\cap A)=P(AB_1\cup AB_2)=P(AB_1)+P(AB_2)-P(AB_1B_2),$$

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

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.

例 2.2 盒子中有 4 只不同的产品, 其中 3 只一等品, 1 只二等品. 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 只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

$$\Omega = \{(i,j)\colon i \neq j, i,j \in [4]\}, \qquad A = \{(i,j)\colon i \neq j, i \neq 4\},$$
$$B = \{(i,j)\colon i \neq j, j \neq 4\}, \qquad AB = \{(i,j), i \neq j, i,j \in [3]\}.$$

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 2.3 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.

解 用 i 和 j 分别表示第一、二次掷骰子的点数, 由此可得样本空间 $\Omega = \{(i,j)\colon i,j \in [6]\}$, 用 $A = \{(6,j)\colon j \in [6]\}$ 表示第一次掷 6 点的事件, 用 B 表示两次点数之和小于 10 的事件, 则有

$$A = \{(6,j)\colon j \in [6]\} \quad 和 \quad B = \{(i,j)\colon i + j \geqslant 10\}$$

于是有事件 $AB = \{(6,j)\colon j \in \{4,5,6\}\}$, 从而得到 $P(B|A) = P(AB)/P(A) = 1/2$.

也可以采用 **样本空间缩减法** 来求解此问题: 在第一次掷 6 点的条件下, 第二次掷骰子的样本空间 $\Omega' = \{1,2,\cdots,6\}$, 事件 $B = \{\text{两次点数之和} \geqslant 10\} = \{4,5,6\}$, 于是有 $P(B|A) = 1/2$.

2.1.1 乘法公式

对随机事件 A 和 B 且满足 $P(A) > 0$, 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:

性质 2.3 对随机事件 $A_1, A_2, \ldots, A_n$ 且满足 $P(A_1A_2\cdots A_{n-1}) > 0$, 有

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

例 2.4 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 $B = \bar{A}_1\bar{A}_2A_3$. 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 2.5 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 $A_1A_2\cdots A_{k-1}\bar{A}_k$, 根据乘法公式有

$$\begin{aligned} & P(A_1A_2\cdots A_{k-1}\bar{A}_k) \\ = \ & P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_{k-1}|A_1\cdots A_{k-2})P(\bar{A}_k|A_1A_2\cdots A_k) \\ = \ & \frac{n-1}{n} \times \frac{n-2}{n-1} \times \cdots \times \frac{n-(k-1)}{n-(k-2)} \times \frac{1}{n-(k-1)} = \frac{1}{n} \end{aligned}$$

我们也可以根据 **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 2.6 (匹配问题) 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 $A_1A_2\cdots A_n$. 根据乘法公式有

$$\begin{aligned} & P(A_1A_2\cdots A_n) \\ = \ & P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1}) \\ = \ & \frac{1}{n} \times \frac{1}{n-1} \times \cdots \times \frac{1}{1} = \frac{1}{n!}. \end{aligned}$$

例 2.7 第一个箱子里有 n 个不同的白球, 第二个箱子里有有 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 $A_1 = B_1B_2\cdots B_{m+n-1}$. 根据乘法公式有

$$\begin{aligned} P(A_1) &= P(B_1)P(B_2|B_1)\cdots P(B_m|B_1B_2\cdots B_{m-1}) \times P(B_{m+1}B_{m+2}\cdots B_{m+n-1}|B_1B_2\cdots B_m) \\ &= \left(1-\frac{1}{n}\right)^m P(B_{m+1}B_{m+2}\cdots B_{m+n-1}|B_1B_2\cdots B_m) = \left(1-\frac{1}{n}\right)^m \times \frac{1}{n}. \end{aligned}$$

由此可知第一个箱子最后一次取走的球是白球的概率为 $(1-1/n)^m$.

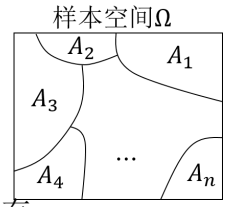
2.1.2 全概率公式

利用条件概率可以将一个复杂事件的概率计算问题进行简化, 这就是本节所讲的 **全概率公式**, 是概率论中最基本的公式之一. 其本质是对加法和乘法事件的综合运用: 对任意互不相容的事件 A, B 有 $P(A\cup B) = P(A)+P(B)$; 对任意事件 A, B 满足 $P(A)>0$ 有 $P(AB) = P(A)P(B|A)$.

首先定义样本空间的一个划分.

定义 2.2 若随机事件 $A_1, A_2, \cdots, A_n$ 满足: i) 任意两两事件是互不相容性的 (或互斥的), 即 $A_i \cap A_j = \emptyset$ $(i \neq j)$; ii) 完备性 $\Omega = \bigcup_{i=1}^n A_i$, 则称事件 $A_1, A_2, \cdots, A_n$ 为空间 Ω 的一个 **划分**.

特别地, 当 $n=2$ 时有 $A_1 = \bar{A}_2$, 即 A_1 与 A_2 互为对立事件. 若 $A_1, A_2, \cdots, A_n$ 为样本空间的一个划分, 则每次试验时事件 $A_1, A_2, \cdots, A_n$ 有且仅有一个事件发生.



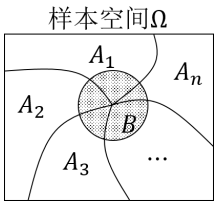
基于样本空间的划分, 下面介绍全概率公式:

定理 2.1 若事件 $A_1, A_2, \cdots, A_n$ 为样本空间 Ω 的一个划分, 对任意事件 B 有

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

称之为 **全概率公式** (Law of total probability).

可以将事件 B 看作某一过程的结果, 将 $A_1, A_2, \cdots, A_n$ 看作产生该结果的若干原因. 若 i) 每一种原因已知, 即 $P(A)$ 已知; ii) 每一种原因对结果 B 的影响已知, 即 $P(B|A_k)$ 已知, 则 $P(B)$ 可计算.



证明 根据分配律有

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

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

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

例 2.8 同一种型号产品由三家工厂生产, 其生产的市场份额分别为 30%, 50%, 20%, 三家工厂的次品率分别为 2%, 1%, 1%. 求这批产品中任取一件是次品的概率.

解 用事件 B 表示任取一件是次品, 事件 A_i 表示取自第 i 家工厂的产品 ($i \in [3]$). 于是有

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = 1.3\%.$$

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

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

还可以采用 **直接计算概率** 求解该问题. 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 $\{0, 2, 4, \ldots, 2k\}$ ($2k \leqslant n$), 根据概率公式直接计算有

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

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

还可以采用 **推迟决定原则** (Principle of deferred decision) 来求解该问题. 无论前 $n-1$ 次中正面朝上的次数为奇数或偶数, 前 n 次正面朝上次数的奇偶性取决于最后一次, 机会各半.

例 2.10 假设有 n 个箱子, 每个箱子里有 30 只白球和 20 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 $\overline{A_i}$ 表示从第 i 个箱子取出白球的事件. 则有

$$P(A_1) = 2/5 \qquad 和 \qquad P(\overline{A_1}) = 3/5$$

根据全概率公式有

$$P(A_2) = P(A_1)P(A_2|A_1) + P(\overline{A_1})P(A_2|\overline{A_1}) = \frac{2}{5} \times \frac{21}{51} + \frac{3}{5} \times \frac{20}{51} = \frac{2}{5}.$$

由此可知 $P(\overline{A_2}) = 3/5$. 依次类推重复上述过程 $n-1$ 次, 最后一个箱子取出一球是红球的概率为 2/5.

2.1.3 贝叶斯公式

基于全概率公式, 我们可以介绍概率论中另一个重要的公式: **贝叶斯公式** (Bayes' law). 其研究在一种结果已发生的情况下是何种原因导致该结果, 确切的说: 观察到事件 B 已经发生的条件下,

寻找导致 B 发生原因的概率. 贝叶斯公式给出了相应的答案.

定理 2.2 设 $A_1, A_2, \cdots, A_n$ 为样本空间 Ω 的一个划分, 且事件 B 满足 $P(B) > 0$. 对任意 $1 \leqslant i \leqslant n$ 有

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

贝叶斯公式的一种直觉解释: 将事件 B 看作结果, 将 $A_1, A_2, \cdots, A_n$ 看作产生结果的若干种原因, 如果 i) 每一种原因发生的概率 $P(A_i)$ 已知; ii) 每一种原因 A_i 对结果 B 的影响已知, 即概率 $P(B|A_i)$ 已知, 则可求事件 B 由第 i 种原因引起的概率 $P(A_i|B)$.

贝叶斯公式中每项都有特定的名称: $\Pr(A_i)$ 被称为事件 A_i 的 **先验 (prior) 概率**, 之所有称为 '先验' 是因为不考虑事件 B 的任何因素; $P(B) = \sum_{j=1}^{n} P(A_j)P(B|A_j)$ 被称为 **证据 (evidence) 概率**; $\Pr(A_i|B)$ 被称为事件 A_i 在事件 B (证据) 发生的情况下的 **后验 (posterior) 概率**; $P(B|A_i)$ 被称为 **似然度 (likelihood)**. 因此贝叶斯公式可以进一步写为

$$\text{后验概率} = \frac{\text{先验概率} \times \text{似然度}}{\text{证据概率}} = \text{常量} \times \text{似然度},$$

由此可知后验概率与似然度成正比.

对贝叶斯公式, 当 $n = 2$ 时有

推论 2.1 对事件 A 和 B 且满足 $P(B) > 0$, 有

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

例 2.11 设一个班级中智商高、中、低的同学各占三分之一, 若智商高、中、低的同学分别考得好成绩的概率是 90%, 70%, 50%, 求任意选一个同学考得好成绩的概率, 以及任意选择一个考得好的同学是低智商的概率.

解 任意选择一个同学, 用 A_1, A_2, A_3 分别表示该同学具有高、中、低智商的事件, 用 B 表示该同学考得好成绩的事件. 根据题意可知

$$P(A_1) = P(A_2) = P(A_3) = 1/3, \quad P(B|A_1) = 0.9, \quad P(B|A_2) = 0.7, \quad P(B|A_3) = 0.5.$$

根据全概率公式可知

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = 0.7.$$

根据贝叶斯公式可知

$$P(A_3|B) = \frac{P(A_3)P(B|A_3)}{P(B)} = \frac{0.5}{0.7} \times \frac{1}{3} = \frac{5}{27}.$$

例 2.12 已知事件A为病人被诊断为肝癌, 事件C为病人患有肝癌, $P(A|C)=0.95$, $P(\overline{A}|\overline{C})=0.9$, $P(C)=0.0004$. 求$P(C|A)$.

上面的例子仅作为课堂练习题, 这里不再讲解.

例 2.13 (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 $2/3$. 那么谁错了?

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

$$P(A|D)=\frac{P(A)P(D|A)}{P(D)}=\frac{1}{3} \quad P(C|D)=\frac{P(C)P(D|C)}{P(D)}=\frac{2}{3}$$

所以犯人 a 的推断不正确, 犯人 c 的推断正确.

与三囚徒类似的问题是如下三门问题, 这里仅给出问题的描述, 求解方案与上面类似.

例 2.14 (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?

2.2 独立性

在一般情况下, 由条件概率定义知

$$P(B|A)=P(AB)/P(A)\neq P(B),$$

即事件 A 发生对事件 B 的发生有影响. 然而在很多情况下, 事件 A 的发生对事件 B 的发生可能没有任何影响, 这是本节研究的事件独立性.

2.2.1 两事件的独立性

定义 2.3 若事件 A, B 满足 $P(AB) = P(A)P(B)$, 则称 **事件 A 与 B 相互独立**.

根据上面的定义可知, 对事件 A 和 B 满足 $P(A)P(B) > 0$, 有

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

根据定义还可以发现任何事件与不可能事件 (或必然事件) 相互独立.

性质 2.4 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 2.15 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, 因此

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

另一方面有

$$P(AB) = 2/52 = 1/26 = P(A)P(B),$$

由此事件 A 和 B 相互独立.

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;

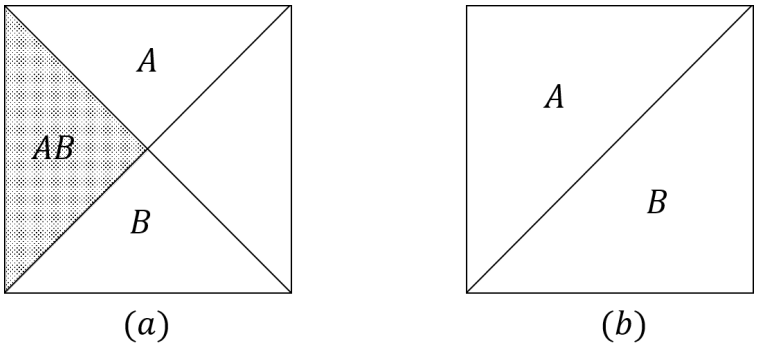


图 2.1 假设落入正方形每个点的可能性完全相同, 即几何概型

- 机器学习的经典假设是训练数据独立同分布采样.

现在我们讨论独立性与互不相容性 (互斥性) 之间的关系: 事件 A 和 B 独立, 根据定义可知 $P(AB) = P(A)P(B)$, 独立性与概率相关, 反映事件的概率属性;

若事件 A 和 B 互不相容, 根据定义有 $AB = \emptyset$, 互斥性与事件的运算关系相关, 与概率无关. 因此独立与互斥反映事件不同的性质, 无必然联系.

如图 2.1(a) 所示: 事件 A 和 B 独立并不意味着事件 A 和 B 互斥, 如图 2.1(b) 所示: 事件 A 和 B 互斥并不意味着事件 A 和 B 独立. 我们进一步有

性质 2.5 事件 A 和 B 满足 $P(A)P(B) > 0$, 若事件 A 和 B 独立则 A 和 B 不互斥; 若事件 A 和 B 互斥则 A 和 B 不独立.

证明 若事件 A 和 B 独立且 $P(A)P(B) > 0$, 有

$$P(AB) = P(A)P(B) > 0$$

事件 A 和 B 不互斥; 另一方面, 若事件 A 和 B 互斥且 $P(A)P(B) > 0$, 有

$$P(AB) = 0 \neq P(A)P(B)$$

事件 A 和 B 不独立.

若事件 A 和 B 互斥且 $P(A)P(B) > 0$, 下面哪些说法正确?

a) $P(B|A) > 0$, b) $P(A|B) = 0$, c) A, B不独立, d) $P(A|B) = P(A)$.

若事件 A 和 B 独立且 $P(A)P(B) > 0$, 下面哪些说法正确?

a) $P(B|A) > 0$, b) $P(A|B) = P(A)$, c) $P(A|B) = 0$, d) $P(AB) = P(A)P(B)$.

2.2.2 多个事件的独立性

定义 2.4 若事件 A, B, C 满足 $P(AB) = P(A)P(B)$, $P(AC) = P(A)P(C)$, $P(BC) = P(B)P(C)$ 且 $P(ABC) = P(A)P(B)P(C)$, 则称 **事件 $\mathbf{A}, \mathbf{B}, \mathbf{C}$ 相互独立**.

根据定义可知: 事件 A, B, C 相互独立和事件 A, B, C 的两两独立不同, 由事件 A, B, C 相互独立可知事件 A, B, C 两两独立; 反之不一定成立, 还需满足 $P(ABC) = P(A)P(B)P(C)$.

下面定义 n 个事件的独立性:

定义 2.5 若事件 $A_1, A_2, \cdots, A_n$ 中任意 k 个事件独立, 即对任意 $k \in [n]$ 有

$$P(A_{i_1} \cdots A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

其中 $1 \leqslant i_1 \leqslant i_2 \leqslant \cdots i_k \leqslant n$, 则称 **事件 $\mathbf{A_1}, \mathbf{A_2}, \cdots, \mathbf{A_n}$ 相互独立**.

同样需要注意 n 个事件的相互独立性与两两独立性的区别. 下面来看一个独立性的例子.

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

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

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

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3) \\ &= \frac{1}{5} + \frac{1}{4} + \frac{1}{3} - \frac{1}{20} - \frac{1}{15} - \frac{1}{12} + \frac{1}{60} = 0.6 \end{aligned}$$

我们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

$$P(A_1 \cup A_2 \cup A_3) = 1 - P(\bar{A}_1\bar{A}_2 \cdots \bar{A}_n) = 1 - P(\bar{A}_1)P(\bar{A}_2)P(\bar{A}_3) = 1 - \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{3}{4} = 0.6.$$

从上例可知: 尽管每个人能破译密码的概率都小于 $1/2$, 但三人独立进行破译, 则至少有一人破译密码的概率则为 $2/3$, 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 n 个事件 $A_1, A_2, \cdots, A_n$ 相互独立, 以及其发生的概率分别为 $p_1, p_2, \cdots, p_n$, 则事件 $A_1, A_2, \cdots, A_n$ 中至少有一事件发生的概率为

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

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

$$P(\bar{A}_1 \cup \bar{A}_2 \cup \cdots \cup \bar{A}_n) = 1 - P(A_1A_2 \cdots A_n) = 1 - p_1p_2 \cdots p_n.$$

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 2.6 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 $A_1, A_2, \cdots, A_n, \cdots$ 独立且每事件发生的概率 $P(A_i) = p > 0$ 非常小, 则有

$$P(A_1A_2\cdots A_n) = 1 - P(\bar{A}_1\bar{A}_2\cdots\bar{A}_n) = 1 - (1-p)^n \to 1 \quad \text{当} \quad n \to \infty,$$

即独立重复多次的小概率事件亦可成立必然事件.

还可以进一步研究: 若独立事件 $A_1, A_2, \cdots, A_n$ 发生的概率 $P(A_i) = p\ (i \in [n])$, 则 n 个事件中恰有 k 个事件发生的概率为 $\binom{n}{k}p^k(1-p)^{n-k}$.

例 2.17 冷战时期美国的导弹精度 99%, 苏联的导弹精度 60%, 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?

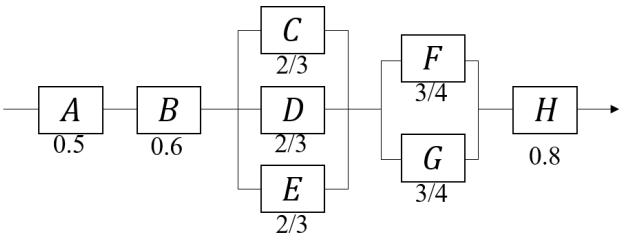
解 假设每次独立发射 n 枚导弹, 用事件 A_i 表示第 i 枚导弹命中目标, 则 n 枚导弹击中目标的概率为

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = 1 - (1-0.6)^n \geqslant 0.99 \quad \Rightarrow \quad n \geqslant 5,$$

因此每次独立发射 5 枚导弹, 击中目标的概率高于 99%.

在上例中, 若美国的导弹精度为 90%, 苏联的导弹精度为 70%, 则苏联每次只需独立发射两枚导弹即可达到 91%.

例 2.18 一串电路如下图所示: A, B, C, D, E, F, G 是电路元件, 电路元件各自下方的数字表示正常工作的概率. 若各电路元件之间相互独立. 求电路正常工作的概率.



解 用事件 W 表示电路正常工作, 则有 $W = A \cap B \cap (C \cup D \cup E) \cap (F \cup G) \cap H$. 根据独立性假设有

$$P(W) = P(A)P(B)P(C \cup D \cup E)P(F \cup G)P(H).$$

根据 $P(C\cup D\cup E)=1-P(\bar{C})P(\bar{D})P(\bar{E})=1-(2/3)^3=19/27$ 和 $P(F\cup G)=1-P(\bar{E})P(\bar{G})=7/16$, 可得 $P(W)=133/1800$.

2.3 案例分析

本节研究的问题: 给定矩阵 $A,B,C\in\{0,1\}^{n\times n}$ (n 非常大, 如 $n\geqslant 10000000$), 验证 $AB=C$ 是否成立? 若直接执行矩阵乘法运算、并验证等式是否成立, 计算复杂度为 $O(n^3)$; 若采用分治法, 计算复杂度为 $O(n^{\log_2 7})$, 目前最好的计算复杂度为 $O(n^{2.37})$. 为进一步降低计算复杂度, 可利用独立性验证 $AB=C$ 是否成立?

独立随机产生一个向量 $r\in\{0,1\}^n$, 判断

$$A(Br)=Cr?$$

计算 $A(Br)$ 和 Cr 的复杂度均为 $O(n^2)$. 若 $A(Br)\neq Cr$ 则直接可得 $AB\neq C$; 若 $A(Br)=Cr$ 并不能得出 $AB=C$. 将上述过程独立进行 K 次, 可以证明以较大的概率有 $AB=C$ 成立, 该过程被称为 Freivalds 算法.

- -

Freivalds 算法

- -

Input: A,B,C

Output: Yes/No

- -

For $i=1:K$

  Select a random vector $r=(r_1,r_2,\cdots,r_n)$ with $P(r_j=0)=P(r_j=1)=1/2$ $(j\in[n])$

  Compute $p=A\times(Br)-Cr$

  If $p\neq 0$ then

    Return 'No'.

  EndIf

EndFor

Return 'Yes'.

- -

首先发现该算法的计算复杂度为 $O(Kn^2)$, 若 K 比较小则显著降低了计算复杂度. 进一步研究算法的有效性, 若返回 'No', 则必然有 $AB\neq C$; 若返回 'Yes', 然而并不一定有 $AB=C$ 成立, 下面研究当算法返回 'Yes' 时 $AB=C$ 成立的概率.

设 $D=AB-C\neq 0$, 则 D 中必存在一些元素不为 0, 不妨令 $d_{11}\neq 0$. 对任意一轮循环, 不妨设随机向量 $r=\{r_1,r_2,\cdots,r_n\}$, 根据返回 'Yes' 可知 $Dr=0$, 进一步可得向量 Dr 的第一个元素等于

0, 即

$$\sum_{j=1}^{n} d_{1j} r_j = 0 \quad \Longrightarrow \quad r_1 = -\frac{1}{d_{11}} \sum_{j=2}^{n} d_{1j} r_j$$

无论 $r_2, \ldots, r_n$ 取何值, 等式 $\sum_{j=1}^{n} d_{1j} r_j = 0$ 是否成立由 r_1 的值决定. 根据 $P(r_1 = 0) = P(r_1 = 1) = 1/2$ 可知 $\sum_{j=1}^{n} d_{1j} r_j = 0$ 成立的概率不超过 $1/2$. 因此在 K 轮独立的循环中, 等式 $\sum_{j=1}^{n} d_{1j} r_j = 0$ 成立的概率不超过 $1/2^K$.

取 $K = \log_2 n$, 则算法 Freivalds 计算复杂度为 $O(n^2 \log n)$, 若算法返回 'No', 则 $AB \neq C$; 若返回 'Yes', 则有

$$P(AB = C) > 1 - 1/n,$$

即至少以 $1 - 1/n$ 的概率有 $AB = C$ 成立.

习题

2.1 阐述独立与互不相容的关系.

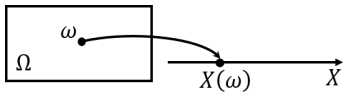
2.2 若事件 A, B, C 独立, 证明: A 与事件 $B \cup C$ 独立.

2.3 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

有些随机试验的结果是数值的, 例如, 抛一枚骰子的点数分别为 $1,2,\cdots,6$; 国家一年出生的婴儿数分别为 $0,1,2,\cdots,n,\cdots$. 有些试验结果可能与数值, 但可以用数值来表示, 例如, 抛一枚硬币, 用 0 表示 '正面朝上', 用 1 表示 '正面朝下'. 流星坠落地球的落脚点用坐标纬度表示. 当试验结果用数值表示时, 可以引入一个变量来表示随机事件, 由此产生随机变量的概念.

将样本空间 Ω 中每个样本点 ω 与一个实数 $X(\omega)$ 相对应, $X(\omega)$ 是 ω 的实值函数, 称实值函数 $X(\omega):\Omega\to\mathbb{R}$ 为随机变量 (random variable), 简写为 r.v., 一般用大写字母 X,Y,Z 表示. $X(\omega)$ 随样本点 ω 的不同而取不同的值, 例如:



- 抛一枚骰子, 用随机变量 X 表示出现的点数, 则随机变量 $X\in[6]$. 出现的点数不超过 4 的事件可表示为 $\{X\leqslant 4\}$; 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 $\{X\leqslant 500\}$.

通过随机变量来形式化描述随机现象或随机事件, 从而利用数学工具来研究概率, 例如 $\{X\leqslant -\infty\}$ 表示不可能事件, 以及 $\{X\leqslant +\infty\}$ 表示必然事件.

根据随机变量的取值, 可分为离散型随机变量和连续型随机变量. 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 X 的取值是有限或无限可列的, 不妨假设其取值为 $x_1,x_2,\cdots,x_n,\cdots$, 事件 $\{X=x_k\}$ 的概率记为

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

称之为随机变量 X 的 **分布列**.

分布列包含了随机变量的取值和概率, 从而完整地刻画了离散随机变量的概率属性, 也可以用表格表示分布列, 如下所示:

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有

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

下面来看看一些离散随机变量的例子:

例 3.1 设随机变量 X 的分布列 $P(X = k) = c/4^k$ $(k = 0, 1, 2, \cdots)$, 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 $p_i = c\lambda^i/i!$ $(i \geqslant 0)$, 求 $P(X > 2)$.

解 根据概率的完备性有

$$1 = \sum_{i=0}^{\infty} p_i = c \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = c \cdot e^{\lambda}$$

从而得到 $c = e^{-\lambda}$, 进一步得到

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

例 3.3 从 $\{1, 2, \ldots, 10\}$ 中不放回随机任意取 5 个数, 令随机变量 X 表示所取 5 个数中的最大值, 求 X 的分布列.

解 由题意可知 X 的取值为 $5, 6, 7, 8, 9, 10$, 且

$$P(X = k) = \binom{k-1}{4} \Big/ \binom{10}{5} \quad (k = 5, 6, \ldots, 10).$$

由此可得 X 的分布列表格为

X	5	6	7	8	9	10
P	1/252	5/252	15/252	35/252	70/252	126/252

3.2 离散型随机变量的期望和方差

随机变量的取值具有一定的随机性, 我们希望研究随机变量的一些不变量, 用以刻画随机变量的特征, 最常见的特征是期望与方差.

3.2.1 期望

定义 3.1 设离散型随机变量 X 的分布列为 $P(X = x_k) = p_k$ $(k \geqslant 1)$, 若级数 $\sum_{k=1}^{\infty} p_k x_k$ 绝对收敛, 称级数和为随机变量 X 的 **期望** (expectation), 又被称为 **均值** (mean) 或 **加权平均** (weighted

average), 记为 $E(X)$, 即

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

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. 级数的绝对收敛保证了级数和不随级数各项次序的改变而改变, 因此期望 $E(X)$ 反映了 X 可能值的平均值, 不会随次序改变而改变. 根据随机变量随机变量 X 的分布列可直接计算其期望.

例 3.4 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 $1,2,\cdots,6$, 且每点等可能发生, 其分布列为 $P(X=i)=1/6\ (i\in[6])$. 因此随机变量 X 的期望为

$$E(X)=(1+2+3+4+5+6)/6=3.5.$$

例 3.5 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号码, 求 $E(X)$.

解 先给出 X 的分布列

$$P(X=1)=\frac{\binom{3}{1}3^2+\binom{3}{2}3+1}{4^3}=\frac{37}{64},\quad P(X=2)=\frac{\binom{3}{1}2^2+\binom{3}{2}2+1}{4^3}=\frac{19}{64},$$
$$P(X=3)=\frac{\binom{3}{1}+\binom{3}{2}+1}{4^3}=\frac{7}{64},\qquad P(X=4)=\frac{1}{64}.$$

进一步可得

$$E(X)=\frac{37}{64}+2\cdot\frac{19}{64}+3\cdot\frac{7}{64}+4\cdot\frac{1}{64}=\frac{25}{16}.$$

例 3.6 有 n 把钥匙只有一把能打开门, 随机选取一把试开门, 若打不开则除去, 求打开门需要尝试次数的期望.

解 设随机变量 X 表示尝试开门的次数, 其分布列为

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

进一步可得打开门次数的平均数

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

根据期望的定义有如下性质:

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

性质 3.3 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 $P(X = x_k) = p_k$, 则随机变量 $Y = aX + b$ 的分布列为 $P(Y = ax_k + b) = p_k$, 进而有

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

对随机变量函数的期望, 有如下定理:

定理 3.1 设离散型随机变量 X 的分布列为 $P(X = x_k) = p_k$ $(k \geqslant 1)$, 若 $g : \mathbb{R} \to \mathbb{R}$ 是连续的函数, 且级数 $\sum_{k\geqslant 1} g(x_k)p_k$ 绝对收敛, 则有

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

根据该定理可知当计算随机变量 $Y = g(X)$ 的期望时, 不需计算 Y 的分布列, 只需利用 X 的分布列即可计算期望 $E[Y]$.

证明 证明的思想是利用绝对收敛保证无穷级数任意重排后的级数仍收敛于原无穷级数的和. 根据题意有 X 的分布列为 $P(X = x_k) = p_k$ 以及随机变量函数 $Y = g(X)$ 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	$\cdots$

注意 y_i 可能等于 y_j $(i \neq j)$, 因此 $P(Y = y_j) = p_j$ 不是随机变量 Y 的分布列. 为构造 Y 的分布列, 我们将 $x_1, x_2, \ldots, x_n, \ldots$ 进行重新分组,

$$\underbrace{x_{1,1}, x_{1,2}, \cdots, x_{1,k_1}}_{y_1'=g(x_{1,j})\ (j\in[k_1])},\quad \underbrace{x_{2,1}, x_{2,2}, \cdots, x_{2,k_2}}_{y_2'=g(x_{2,j})\ (j\in[k_2])},\quad \cdots,\quad \underbrace{x_{n,1}, x_{n,2}, \cdots, x_{n,k_n}}_{y_n'=g(x_{n,j})\ (j\in[k_n])},\quad \cdots$$

其中 $y_i' \neq y_j'$ $(i \neq j)$. 由此可得随机变量 Y 的分布列为

$$P[Y = y_i'] = \sum_{j=1}^{k_i} p_{i,j} = \sum_{k\geqslant 1, y_i'=g(x_k)} p_k,$$

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级数重排后其和不变.

推论 3.1 对离散型随机变量 X 和连续函数 $g_i:\mathbb{R}\to\mathbb{R}$ $(i\in[n])$, 若每个函数的期望 $E(g_i(X))$ 存在, 则对任意常数 $c_1,c_2,\ldots,c_n$ 有

$$E\left(c_1g_1(X)+c_2g_2(X)+\cdots+c_ng_n(X)\right)=\sum_{i=1}^{n}c_iE(g_n(X)).$$

证明 根据定理 3.1有

$$\begin{aligned}&E\left(c_1g_1(X)+c_2g_2(X)+\cdots+c_ng_n(X)\right)\\ =&\sum_{k=1}^{\infty}p_k(c_1g_1(x_k)+c_2g_2(x_k)+\cdots+c_ng_n(x_k))\\ =&\sum_{i=1}^{n}c_i\sum_{k=1}^{\infty}p_kg_i(x_k)=\sum_{i=1}^{n}c_iE(g_i(X)).\end{aligned}$$

由此容易得到 $E(X^2+X+\sin X+4)=E(X^2)+E(X)+E(\sin X)+4$.

在实际应用中往往不知道随机变量的分布, 但需要对期望进行一定的估计, 为此需要引入一些不等式. 给定随机变量函数 $Y=g(X)$, 下面探讨 $E(g(X))$ 和 $g(E(X))$ 之间的大小关系.

定义 3.2 若函数 $g\colon[a,b]\to\mathbb{R}$ 对任意 $x_1,x_2\in[a,b]$ 和 $\lambda\in[0,1]$, 有 $g(\lambda x_1+(1-\lambda)x_2)\leqslant\lambda g(x_1)+(1-\lambda)g(x_2)$ 成立, 称函数 $g(x)$ 是定义在 $[a,b]$ 上的 **凸函数**;

若对任意 $x_1,x_2\in[a,b]$ 和 $\lambda\in[0,1]$, 有 $g(\lambda x_1+(1-\lambda)x_2)\geqslant\lambda g(x_1)+(1-\lambda)g(x_2)$ 成立, 则称函数 $g(x)$ 是定义在 $[a,b]$ 上的 **凹函数**.

下面介绍著名的 **琴生不等式** (Jensen's inequality), 常用于各种推导估计.

定理 3.2 对离散型随机变量 $X\in[a,b]$ 和连续凸函数 $g\colon[a,b]\to\mathbb{R}$, 有

$$g(E(X))\leqslant E(g(X));$$

对离散型随机变量 $X\in[a,b]$ 和连续凹函数 $g\colon[a,b]\to\mathbb{R}$, 有

$$g(E(X))\geqslant E(g(X)).$$

证明 为了证明的简洁起见, 这里考虑有限的样本空间和凸函数情况, 可类似考虑其它情况. 设随机变量 X 的取值为 $x_1,x_2,\cdots,x_n$, 其分布列为 $P(X=x_k)=p_k\geqslant0$, 根据概率性质有 $\sum_k p_k=1$. 我们需要证明 $g(E(X))\leqslant E[g(X)]$, 即不等式

$$g(p_1x_1+p_2x_2+\cdots+p_nx_n)\leqslant p_1g(x_1)+p_2g(x_2)+\cdots+p_ng(x_n).\tag{3.1}$$

这里对 n 采用归纳法证明, 当 $n=2$ 时由凸函数的定义直接可证. 不妨假设 $n=m-1$ 时成立 $(m\geqslant 3)$, 下面证明当 $n=m$ 亦成立. 首先有

$$\begin{aligned} g(p_1x_1+p_2x_2+\cdots+p_mx_m) &= g\left(p_1x_1+(1-p_1)\left[\frac{p_2}{1-p_1}x_2+\cdots+\frac{p_m}{1-p_1}x_m\right]\right) \\ &\leqslant p_1g(x_1)+(1-p_1)g\left(\frac{p_2}{1-p_1}x_2+\cdots+\frac{p_m}{1-p_1}x_m\right) \end{aligned}$$

这里利用 $g(p_1x_1+(1-p_1)x_1')\leqslant p_1g(x_1)+(1-p_1)g(x_1')$, 其中 $x_1'=(x_2p_2+\cdots+x_mp_m)/(1-p_1)$. 容易发现 $p_i/(1-p_1)\geqslant 0$ 且 $\sum_{i=2}^m p_i/(1-p_1)=1$, 根据归纳假设有

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

代入即可完成证明.

对任意离散型随机变量 X, 根据 Jensen 不等式有

$$(E(X))^2\leqslant E(X^2)\qquad 和\qquad e^{E(X)}\leqslant E(e^X).$$

3.2.2 方差

数学期望反映了 X 取值的平均值, 这里考虑三个随机变量 X, Y 和 Z, 其分布列分别为

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

尽管随机变量的均值一样, 即 $E(X)=E(Y)=E(Z)=0$, 但是与期望的偏离程度确又有很大的不同, 我们本节研究随机变量 X 与期望 $E(X)$ 的偏离程度, 即方差.

定义 3.3 设离散随机变量 X 的分布列为 $p_k=P(X=x_k)$ $(k\geqslant 0)$, 若期望 $E(X)=\sum_k x_kp_k$ 存在, 以及 $E(X-E(X))^2=\sum_k p_k(x_k-E(X))^2$ 存在, 称 $E(X-E(X))^2$ 为随机变量 X 的方差 (variance), 记为 $\mathrm{Var}(X)$ 或 $D(X)$, 即

$$\mathrm{Var}(X)=\mathrm{D}(X)=E(X-E(X))^2=\sum_k p_k(x_k-E(X))^2=\sum_k p_k\left(x_k-\sum_k x_kp_k\right)^2,$$

称 $\sqrt{\mathrm{Var}(X)}$ 为标准差 (standard deviation), 记为 $\sigma(X)$.

根据定义可知随机变量取值的顺序改变不会改变方差的大小, 根据期望的性质有

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

由此给出方差的另一种等价性定义

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2.$$

尽管方差的两种定义完全等价, 但在实际应用中可能带来不同的计算量.

例 3.7 设随机变量 X 的分布列为 $P(X = x_i) = 1/n$ $(i \in [n])$, 试问: 计算随机变量 X 的方差需要遍历 $x_1, x_2, \cdots, x_n$ 几遍.

解 若利用定义 $\mathrm{Var}(X) = E(X - E(X))^2$, 则需要遍历数据 $x_1, x_2, \cdots, x_n$ 两遍, 第一遍计算期望 $E(X)$, 第二遍计算方差 $\mathrm{Var}(X)$. 若利用定义 $\mathrm{Var}(X) = E(X^2) - (E(X))^2$, 则只需要遍历 $x_1, x_2, \cdots, x_n$ 一遍, 可在线 (online) 计算方差, 不需存储数据.

下面给出方差的性质:

性质 3.4 若随机变量 $X \equiv c$, 则 $\mathrm{Var}(X) = 0$.

性质 3.5 对随机变量 X 和常数 $a, b \in \mathbb{R}$, 有

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

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

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

一般情况下方差不具有线性性, 即 $\mathrm{Var}[f(X) + g(X)] \neq \mathrm{Var}[f(X)] + \mathrm{Var}[g(X)]$.

性质 3.6 对随机变量 X 和常数 $a \in \mathbb{R}$, 有

$$\mathrm{Var}(X) = E(X - E(X))^2 \leqslant E(X - a)^2.$$

证明 我们有

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

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

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

证明 对任意随机变量 $X \in [a,b]$, 有 $(b-X)(X-a) \geqslant 0$, 两边同时对随机变量取期望, 整理可得

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

根据方差的定义有

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

进一步对二次函数 $f(t) = -t^2 + (a+b)t - ab$ 求最大值, 可得 $(b - E(X))(E(X) - a) \leqslant (b-a)^2/4$.

3.3 常用的离散型随机变量

本节介绍几种常用的离散型随机变量, 并研究其数字特征.

3.3.1 离散均匀分布

定义 3.4 设随机变量 X 的取值为 $x_1, x_2, \cdots, x_n$, 且 $P(X = x_i) = 1/n$, 称 X 服从离散型均匀分布.

由定义可知

$$\begin{aligned} E(X) &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\ \mathrm{Var}(X) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2. \end{aligned}$$

下面来看一个离散型均匀分布的例子.

例 3.8 (德国坦克问题) 假设德国生产了 N 辆坦克, 编号分别为 $1, 2, \cdots, N$, 盟军战斗中随机击毁了 k 辆, 被随机击毁坦克编号分别为 $x_1, x_2, \ldots, x_k$, 如何估计 N 的大小.

解 对题意进行分析, 坦克被随机击毁可看作坦克被击毁的服从离散型均匀分布, 即 $P[X = i] = 1/N$ ($i \in [N]$), 因此有 $E(X) = (1+N)/2$, 可用被击毁坦克编号的平均值去近似, 即

$$E(X) = \frac{1+N}{2} \approx \frac{1}{k}\sum_{i=1}^{k} x_i \qquad \Longrightarrow \qquad N \approx \frac{2}{k}\sum_{i=1}^{k} x_i - 1.$$

若 k 较大时上述表达式对 N 有一个较好的估计; 但 k 较小时, 往往更关注于被毁坦克的最大编号 m.

方法二 问题转化为从 $\{1,2,\cdots,N\}$ 中以不放回随机抽取 k 个数, 观察到 k 个数中最大数为 m, 如何利用 m 和 k 估计 N. 假设随机变量 X 表示抽取 k 个数中的最大数, 其分布列为

$$P(X=i)=\binom{i-1}{k-1}\Big/\binom{N}{k}\quad(k\leqslant i\leqslant N).$$

由此得到

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

为计算期望 $E(X)$, 考虑从 $N+1$ 个元素中随机取 $k+1$ 个元素, 可等价于按所抽取 $k+1$ 个元素中最大元进行分类, 即

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

代入 $E(X)$ 可得

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

由于仅做了一次观察, 可以将一次观察的最大值 m 看作为 $E[X]$ 的近似, 即

$$m\approx E(X)=\frac{k(N+1)}{k+1}\quad\Longrightarrow\quad N\approx m\left(1+\frac{1}{k}\right)-1,$$

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 $N=212\times(1+1/6)-1\approx 246$. 针对二战德国坦克数量的实际估计情况可参见下表, 统计估计比情报估计准确得多, 接近德国的实际产量.

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

3.3.2 0-1分布

定义 3.5 随机变量 X 的取值为 $\{0,1\}$, 其分布列 $P(X=1)=p$, $P(X=0)=1-p$, 称 X 服从参数为 p 的 $0-1$ 分布, 又称两点分布, 或 Bernoulli 分布, 记 $X\sim\mathrm{Ber}(p)$.

由定义可知

$$E(X)=p,\qquad \mathrm{Var}(X)=p-p^2=p(1-p).$$

$0-1$分布是很多概率模型的基础.

3.3.3 二项分布

Bernoulli 试验只有两个结果: A 和 $\bar{A}$, 设 $P(A)=p$ ($p\in[0,1]$), 因此 $P(\bar{A})=1-p$. 将 Bernoulli 试验独立重复地进行 n 次, 称 n 重 Bernoulli 试验, 是一种非常重要的概率模型, 具有广泛的应用.

定义 3.6 用随机变量 X 表示 n 重 Bernoulli 试验试验中事件 A 发生的次数, 则 X 的取值为 $0,1,\cdots,n$, 其分布列为

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}\qquad k\in\{0,1,2,\ldots,n\}$$

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

我们称之为二项分布是因为与二项展开式 $(a+b)^n=\sum_{k=1}^n\binom{n}{k}a^kb^{n-k}$ 相似, 容易验证

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

对二项分布, 我们有

性质 3.7 对随机变量 $X\sim B(n,p)$ 有

$$E(X)=np\qquad 和\qquad \mathrm{Var}(X)=np(1-p).$$

证明 由定义可知

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

为计算 $E(X)$, 对二项展开式 $(1+x)^n=\sum_{k=0}^n\binom{n}{k}x^k$ 两边求导数有

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

将 $x=p/(1-p)$ 带入可得

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

对于方差, 首先计算

$$E(X^2)\quad=\quad\sum_{k=0}^{n}k^2\binom{n}{k}p^k(1-p)^{n-k}$$

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

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

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

将 $x = p/(1-p)$ 带入有

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

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

例 3.9 有 5 个选择题, 每个选择题有 4 种答案, 只有一种正确, 求一学生随机猜对 4 个选择题的概率?

解 将每一个选择题看作一次 Bernoulli 试验, 事件 A 表示猜正确, 则有 $P(A) = 1/4$. 整个问题等价于 5 重 Bernoulli 试验, 用 X 表示学生猜对题的个数, 则 $X \sim B(5, 1/4)$, 从而得到

$$
P(X=4) = \binom{5}{4}\cdot\frac{1}{4^4}\cdot\frac{3}{4} = \frac{15}{4^5}.
$$

3.3.4 几何分布

定义 3.7 在多重 Bernoulli 试验, 设事件 A 发生的概率为 p. 用随机变量 X 表示事件 A 首次发生时的试验次数, 则 X 的取值为 $1, 2, \cdots$, 其分布列为

$$
P(X=k) = (1-p)^{k-1}p \qquad (k \geqslant 1)
$$

称 X 服从参数为 p 的几何分布, 记为 $X \sim G(p)$.

首先可知 $P(X=k) = (1-p)^{k-1}p \geqslant 0$ 以及

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

从而验证了几何分布构成一个分布列. 对几何分布, 我们有

性质 3.8 若随机变量 $X \sim G(p)$ $(0 < p < 1)$, 则有

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

证明 根据期望的定义有

$$E(X) = \sum_{k=1}^{\infty} kP(X=k) = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1}.$$

对级数展开式 $(1-x)^{-1} = \sum_{k=0}^{\infty} x^k$ 两边求导有

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

令 $x = 1-p$ 可证 $E(X) = 1/p$. 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 $(1-x)^{-1} = \sum_{k=0}^{\infty} x^k$ 两边求二阶导有

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

令 $x = 1-p$ 可得 $E(X^2) = (2-p)/p^2$, 于是有 $\mathrm{Var}(X) = E(X^2) - (EX)^2 = (1-p)/p^2$.

下面给出几何分布的一个重要性质: 无记忆性 (memoryless property).

定理 3.4 设随机变量 $X \sim G(p)$, 对任意正整数 m, n, 有

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

几何分布无记忆性的直观解释: 假设现在已经历 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, 觉得下一次应该赢了, 然而无记忆性给出下一次是否赢与前面输了多少次无关.

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 $\{X>m+n\}\cap\{X>m\}=\{X>m+n\}$.

例 3.10 古人非常重视生男孩且资源有限, 规定每个家庭可生一个男孩, 如果没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. 多年后男女比例是否会失衡?

解 对一个家庭而言, 用随机变量 X 表示该家庭的小孩个数, 则 $X=1,2,\cdots$, 以及

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

这里 $p=1/2$ 表示生男孩的概率. 根据几何分布的期望有一个家庭小孩数的期望为 $E[X]=1/p=2$, 由此可得一个家庭的小孩男女比例 $1:1$.

3.3.5 Pascal/负二项分布

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, 可以这个问题进一步推广到事件 A 第 r 次发生时所进行的试验次数. 具体而言, 在多次 Bernoulli 试验中, 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 $r,r+1,r+2,\cdots$, 其分布列为

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

称随机变量 X 服从参数为 r 和 p 的 **负二项分布**.

易知 $P(X=k)\geqslant 0$, 下面证明

$$\sum_{k=r}^{\infty}P(X=k)=p^r\sum_{k=r}^{\infty}\binom{k-1}{r-1}(1-p)^{k-r}=1.$$

设 $q=1-p$, 根据泰勒展式有

$$p^{-r}=(1-q)^{-r}=\sum_{t=0}^{\infty}\binom{t+r-1}{r-1}q^t=\sum_{k=r}^{\infty}\binom{k-1}{r-1}(1-p)^{k-r}\quad(\text{令 }k=t+r).$$

定理 3.5 设随机变量 X 服从参数为 $p\in(0,1)$ 和 $r>0$ 的负二项分布, 则有

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

证明 对于期望 $E(X)$ 有

$$\begin{aligned} E(X) &= \sum_{k=r}^{\infty} k \cdot P(X=k) = \sum_{k=r}^{\infty} k\binom{k-1}{r-1} p^r (1-p)^{k-r} \\ &= \frac{r}{p}\sum_{k=r}^{\infty}\binom{k}{r} p^{r+1}(1-p)^{k-r} = \frac{r}{p}\sum_{k=r}^{\infty}\binom{k+1-1}{r+1-1} p^{r+1}(1-p)^{k-r} = \frac{r}{p}, \end{aligned}$$

这里利用

$$\sum_{k=r}^{\infty}\binom{k+1-1}{r+1-1} p^{r+1}(1-p)^{k-r} = 1.$$

对 $E(X^2)$ 的计算, 类似有

$$\begin{aligned} E(X^2) &= \sum_{k=r}^{\infty} k^2 P(X=k) = \sum_{k=r}^{\infty} k^2\binom{k-1}{r-1} p^r (1-p)^{k-r} \\ &= \frac{r}{p}\sum_{k=r}^{\infty}(k+1-1)\binom{k}{r} p^r (1-p)^{k-r} \\ &= \frac{r(r+1)}{p^2}\sum_{k=r}^{\infty}\binom{k+1}{r+1} p^{r+1}(1-p)^{k-r} - \frac{r}{p}\sum_{k=r}^{\infty}\binom{k}{r} p^r (1-p)^{k-r} \\ &= \frac{r(r+1)}{p^2} - \frac{r}{p}, \end{aligned}$$

由此可得

$$Var(X) = E(X^2) - (E(X))^2 = \frac{r(r+1)}{p^2} - \frac{r}{p} - \frac{r^2}{p^2} = \frac{r(1-p)}{p^2}.$$

3.3.6 泊松分布

定义 3.8 若随机变量 X 的分布列为

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

其中 $\lambda > 0$ 是一个给定的常数, 称随机变量 X 服从参数为 λ 的泊松分布, 记为 $X \sim P(\lambda)$.

容易验证 $P(X=k) = \lambda^k e^{-\lambda}/k! \geqslant 0$, 并根据泰勒展式 $e^{\lambda} = \sum_{k=0}^{\infty} \lambda^k/k!$ 有

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

泊松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: 在一段时间内电话收到的呼叫次数, 放射物在一段时间内放射的粒子数, 一段时间内通过某路口的出租车数, 一本书中一页出现的语法错误数, 一天内到一所银行办理业务的顾客数等.

性质 3.9 对任意给定的 $\lambda > 0$, 若 $X \sim P(\lambda)$, 则

$$E(X) = \lambda \qquad 和 \qquad \mathrm{Var}(X) = \lambda.$$

证明 根据期望的定义和幂级数 $e^{\lambda} = \sum_{k=0}^{\infty} \lambda^k/k!$ 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

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

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

例 3.11 设随机变量 X 服从参数为 λ 的泊松分布, 且 $P(X=1) = P(X=2)$, 求 $P(X \geqslant 4)$.

解 根据泊松分布的定义可知 $P(X=k) = \lambda^k e^{-\lambda}/k!$ 和 $P(X=1) = P(X=2)$ 可得

$$\lambda e^{-\lambda} = \lambda^2 e^{-\lambda}/2 \quad \Longrightarrow \quad \lambda = 2,$$

进一步得到

$$P(X \geqslant 4) = 1 - P(X=0) - P(X=1) - P(X=2) - P(X=3) = 1 - 5e^{-2} - 4e^{-2}/3.$$

下面研究二项分布和泊松分布的关系, 即泊松定理:

定理 3.6 对任意给定的常数$\lambda > 0$, n 为任意正整数, 设 $np_n = \lambda$, 则对任意给定的非负整数 k, 有

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

证明 由 $p_n = \lambda/n$, 有

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

$$= \frac{\lambda^k}{k!}(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})\left(1-\frac{\lambda}{n}\right)^{\frac{n}{\lambda}\frac{n-k}{n}\lambda}$$

当 $n\to\infty$ 时有 $(1-\frac{\lambda}{n})^{\frac{n}{\lambda}}\to e^{-1}$ 以及 $\frac{n-k}{n}\lambda\to\lambda$, 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.

例 3.12 射击训练每次命中目标的概率为 0.002, 现射击 1000 次, 求命中目标在 10 次与 50 次之间的概率. (用泊松近似计算)

解 将 1000 次射击可看作 1000 重 Bernoulli 试验, 设随机变量 X 表示 1000 射击训练中射中目标的次数, 则 $X\sim B(1000,0.002)$, 利用泊松分布近似, 则可以看作 $X\sim P(2)$, 于是有

$$P(500\leqslant X\leqslant 600)=\sum_{k=10}^{50}\binom{1000}{k}(0.002)^k0.998^{1000-k}\approx\sum_{k=10}^{50}\frac{2^k}{k!}e^{-2}.$$

例 3.13 有 80 台同类型设备独立工作, 发生故障的概率是 0.01, 一台设备发生故障时只能由一人处理, 考虑方案: I) 由四人维护, 每人单独负责 20 台; II) 由三人共同维护 80 台. 哪种方案更为可取?

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

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

因四人独立维修, 有设备发生故障时而不能及时的概率

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

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 $X\sim P(0.8)$, 则有设备发生故障不能及时维修的概率为

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

由此比较可知方案 II) 更优.

3.4 案例分析: 随机二叉树叶结点的平均高度

初始为一个根结点, 再每一次迭代过程中, 随机选择一个叶子结点, 将该叶子结点分裂为左、右叶子结点, 由此重复进行 k 次, 求此随机树一个叶结点的平均高度.

习题

3.1 设随机变量 $X \sim B(n,p)$,证明: $E(X) = np$ 和 $\mathrm{Var}(X) = np(1-p)$.

3.2 设随机变量 $X \sim G(p)$, 证明: $E(X) = 1/p$ 和 $\mathrm{Var}(X) = (1-p)/p^2$.

3.3 设随机变量 X 服从参数为 r 和 p 的负二项分布,证明: $E(X) = r/p$ 和 $\mathrm{Var}(X) = r(1-p)/p^2$.

3.4 设随机变量 $X \sim P(\lambda)$,证明: $E(X) = \lambda$ 和 $\mathrm{Var}(X) = \lambda$.

3.5 初始为一个根结点, 再每一次迭代过程中, 随机选择一个叶子结点, 将该叶子结点分裂为左、右叶子结点, 由此重复进行 k 次, 求此随机树一个叶结点的平均高度.

3.6 从 $\{1,2,...,10\}$ 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.

3.7 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?

3.8 书 55 页 2, 3 题.

3.9 书 113 页 2, 3 题.

3.10 书 114 页 4, 6 题.

第 4 章　连续型随机变量

4.1 概念与性质

4.1.1 分布函数

定义 4.1　给定任意随机变量 X 和 实数 $x \in (-\infty, +\infty)$, 函数

$$F(x) = P(X \leqslant x)$$

称为随机变量 X 的 **分布函数**, 分布函数的本质是概率.

对任意实数 $x_1 < x_2$, 有

$$P(x_1 < X \leqslant x_2) = P(X \leqslant x_2) - P(X \leqslant x_1) = F(x_2) - F(x_1).$$

分布函数 $F(x)$ 具有如下性质:

- 单调性: 若 $x_1 < x_2$, 则 $F(x_1) \leqslant F(x_2)$;
- 规范性: $F(x) \in [0, 1]$, 且 $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$, $F(+\infty) = \lim_{x \to +\infty} F(x) = 1$;
- 右连续性: $F(x + 0) = \lim_{\Delta x \to 0^+} F(x + \Delta x) = F(x)$.

任一分布函数必满足上述三性质, 而满足上述三性质的函数必是某随机变量的分布函数, 因此分布函数可由上述三性质完全刻画.

可利用分布函数 $F(x)$ 表示随机事件的概率, 例如

$$\begin{aligned}
P(X > a) &= 1 - F(a) \\
P(X < a) &= F(a - 0) = \lim_{x \to a^-} F(x) \\
P(X = a) &= F(a) - F(a - 0) \\
P(X \geqslant a) &= 1 - F(a - 0) \\
P(a \leqslant X \leqslant b) &= F(b) - F(a - 0).
\end{aligned}$$

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, 有

$$F(x)=P(X\leqslant x)=P(\emptyset)=0;$$

当 $-1\leqslant x<2$ 时, 有

$$F(x)=P(X\leqslant x)=P(X=-1)=\frac{1}{4};$$

当 $2\leqslant x<3$ 时, 有

$$F(x)=P(X\leqslant x)=P(X=-1)+P(X=2)=\frac{3}{4};$$

当 $x\geqslant 3$ 时有 $F(x)=1$.

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

$$F(x)=P(X\leqslant x)=kx.$$

根据 $F(1)=1$ 求解可得 $k=1$. 从而得到 X 的分布函数为

$$F(x)=\begin{cases}0 & x<0,\\ x & 0\leqslant x\leqslant 1,\\ 1 & x>1.\end{cases}$$

例 4.3 随机变量 X 的分布函数 $F(x)=A+B\arctan x$, $x\in(-\infty,+\infty)$, 求 $P(X\leqslant 1)$.

解 由分布函数的性质有

$$0=F(-\infty)=\lim_{x\to-\infty}A+B\arctan x=A-\pi B/2,$$
$$1=F(+\infty)=\lim_{x\to+\infty}A+B\arctan x=A+\pi B/2,$$

求解可得 $A=1/2$ 和 $B=1/\pi$, 从而得到 $P(X\leqslant 1)=3/4$.

4.1.2 概率密度函数

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

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

成立, 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**, 简称 **概率密度**.

根据分布函数的性质可得到概率密度函数的性质:

性质 4.1 概率密度函数 $f(x)$ 满足非负性 $f(x) \geqslant 0$ 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, 而对任意满足非负性和规范性的函数函数 $f(x)$, 可引入新的随机变量 X, 其分布函数为 $G(x) = \int_{-\infty}^{x} f(t)dt$. 由此说明概率密度函数可由函数的非负性和规范性完全刻画.

对任意 $x_1 \leqslant x_2$, 有

$$P(x_1 < X \leqslant x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t)dt.$$

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 x 轴, $x = x_1$, $x = x_2$ 和 $y = f(x)$ 所围成的曲边梯形的面积.

定理 4.1 对连续随机变量 X, 其分布函数 $F(x)$ 在整个实数域上连续; 若 $f(x)$ 在 x 点连续, 则 $F(x)$ 在 x 点可导, 且 $F'(x) = f(x)$.

证明 根据函数的积分性质: 若 $f(x)$ 在 $[a, b]$ 上可积, 则 $\phi(x) = \int_a^x f(t)dt$ 在 $[a, b]$ 上连续. 若 $f(x)$ 在 $[a, b]$ 上连续, 则 $\phi(x) = \int_a^x f(t)dt$ 在 $[a, b]$ 上可导, 且 $\phi'(x) = f(x)$.

性质 4.2 对连续型随机变量 X 和常数 $x \in (-\infty, +\infty)$, 有 $P(X = x) = 0$.

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

$$P(X = x) \leqslant \lim_{\Delta x \to 0} P(x - \Delta x \leqslant X \leqslant x) = \lim_{\Delta x \to 0} \int_{x-\Delta x}^{x} f(t)dt \leqslant \lim_{\Delta x \to 0} f(\xi)\Delta x = 0,$$

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.

由此可知, 连续随机变量无需考虑端点, 即

$$P(a \leqslant X \leqslant b) = P(a < X < b) = P(a \leqslant X < b) = P(a < X \leqslant b),$$

同时说明了概率密度函数不是概率, 即 $P(X = x) = 0 \neq f(x)$.

若 $f(x)$ 在点 x 连续, 由连续性定义有

$$\lim_{\Delta x \to 0} \frac{P(x - \Delta x \leqslant X \leqslant x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\int_{x-\Delta x}^{x+\Delta x} f(t)dt}{\Delta x} = \lim_{\Delta x \to 0} \frac{2\Delta x \cdot f(\xi)}{\Delta x} = 2f(x),$$

其中 $\xi \in (x - \Delta x, x + \Delta x)$. 由此可得

$$P(x - \Delta x \leqslant X \leqslant x + \Delta x) \approx 2f(x)\Delta x,$$

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 设连续随机变量 X 的密度函数

$$f(x)=\begin{cases}c(4x-2x^2) & 0<x<2\\ 0 & \text{其它},\end{cases}$$

求概率 $P(X>1)$.

解 根据概率密度函数的规范性有

$$1=\int_{-\infty}^{+\infty}f(t)dt=\int_0^2 c(4t-2t^2)dt=\frac{8}{3}c,$$

得到 $c=3/8$, 所以

$$P(X>1)=\int_1^{\infty}f(t)dt=\int_1^2 f(t)dt=\int_1^2\frac{3}{8}(4t-2t^2)dt=\frac{1}{2}.$$

例 4.5 设随机变量 X 的密度函数

$$f(x)=\begin{cases}x & 0<x\leqslant 1\\ a-x & 1<x<2\\ 0 & \text{其它},\end{cases}$$

求其分布函数 $F(x)$.

解 根据概率密度的规范性, 有

$$1=\int_{-\infty}^{+\infty}f(t)dt=\int_0^1 tdt+\int_1^2(a-t)dt=\frac{1}{2}+a-2+\frac{1}{2}=a-1,$$

求解可得 $a=2$, 于是有

$$f(x)=\begin{cases}x & 0\leqslant x\leqslant 1\\ 2-x & 1<x<2\\ 0 & \text{其它}.\end{cases}$$

当 $x\leqslant 0$ 时有 $F(x)=0$; 当 $0<x\leqslant 1$ 时, 有

$$F(x)=\int_0^x f(t)dt=x^2/2;$$

当 $1<x\leqslant 2$ 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 $x \geqslant 2$ 时有 $F(x) = 1$. 综合可得

$$F(x) = \begin{cases} 0 & x \leqslant 0, \\ x^2/2 & 0 < x \leqslant 1, \\ -x^2/2 + 2x - 1 & 1 < x \leqslant 2, \\ 1 & x \geqslant 2. \end{cases}$$

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 $0 \leqslant x \leqslant 2$ 时有

$$F(x) = P(X \leqslant x) = P(0 \leqslant X \leqslant x) = kx^2.$$

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

$$f(x) = \begin{cases} x/2 & 0 \leqslant x \leqslant 2 \\ 0 & \text{其它}. \end{cases}$$

4.1.3 连续随机变量的期望和方差

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 $\int_{-\infty}^{+\infty} xf(x)dx$ 绝对收敛, 称 $\int_{-\infty}^{+\infty} xf(x)dx$ 为随机变量 X 的 **期望**, 记为 $E(X)$, 即

$$E(X) = \int_{-\infty}^{+\infty} tf(t)dt.$$

与离散性随机变量类似, 连续随机变量的期望具有如下一些列性质:

性质 4.3 (线性关系) 对任意任意常数 a, b 和随机变量 X, 有

$$E(aX + b) = aE(X) + b;$$

对常数 $c_1, \ldots, c_n$ 和连续函数 $g_1(x), \ldots, g_n(x)$, 有

$$E\left(\sum_{i=1}^{n} c_i g_i(X)\right) = \sum_{i=1}^{n} c_i E(g_i(X)).$$

性质 4.4 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

$$E(g(X)) = \int_{-\infty}^{+\infty} g(t)f(t)dt.$$

性质 4.5 (Jensen 不等式) 对连续随机变量 X 和凸函数 $f(x)$ 有

$$f(E(X)) \leqslant E[f(X)];$$

对连续随机变量 X 和凹函数 $f(x)$ 有

$$f(E(X)) \geqslant E[f(X)].$$

例 4.7 设随机变量 X 的密度函数为 $f(x)=\begin{cases} cx & x\in[0,1] \\ 0 & \text{其它} \end{cases}$, 求 $E(X^m)$ (m 为正整数).

解 根据概率密度函数的规范性 $\int_{-\infty}^{+\infty} f(t)dt = 1$, 求解可得 $c=2$, 进一步有

$$E(X^m) = \int_0^1 t^m \cdot 2t dt = 2\int_0^1 t^{m+1} dt = \frac{2}{m+2}.$$

在很多实际问题中可能不知道随机变量的分布, 因此不能直接计算期望, 但可以估计概率 $P(X>t)$ 来计算期望:

定理 4.2 对非负随机变量 X, 有

$$E[X] = \int_0^\infty P(X>t)dt.$$

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

$$\begin{aligned} E[X] &= E\left[\int_0^{+\infty} \mathbb{I}[X>t]dt\right] \\ &= \int_{-\infty}^{+\infty}\int_0^{+\infty} \mathbb{I}[x>t]f(x)dtdx \qquad \text{(积分换序)} \\ &= \int_0^{+\infty}\left[\int_{-\infty}^{+\infty} \mathbb{I}[x>t]f(x)dx\right]dt \\ &= \int_0^{+\infty}\left[\int_{-\infty}^{t} \mathbb{I}[x>t]f(x)dx + \int_t^{+\infty} \mathbb{I}[x>t]f(x)dx\right]dt \\ &= \int_0^{+\infty}\left[\int_t^{+\infty} f(x)dx\right]dt = \int_0^{+\infty} P(X>t)dt. \end{aligned}$$

例 4.8 利用此定理计算例 4.7 中随机变量 X 的期望 $E(X)=2/3$.

根据上述定理有如下推理:

推论 4.1 对随机变量 X 和连续函数 $g(x)$, 有

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx = \int_{0}^{+\infty} P(g(X) > t)dt.$$

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 $\mathrm{Var}(X)$, 即

$$\mathrm{Var}(X) = E(X - E(X))^2 = \int_{-\infty}^{+\infty}(t - E(X))^2 f(t)dt.$$

其等价性定义为

$$\mathrm{Var}(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2 = \int_{-\infty}^{+\infty} t^2 f(t)dt - \left(\int_{-\infty}^{+\infty} tf(t)dt\right)^2.$$

性质 4.6 对任意常数 a, b 和随机变量 X, 有 $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

4.2 常用连续型随机变量

本章介绍三种常用连续型随机变量.

4.2.1 均匀分布(uniform distribution)

给定区间 $[a, b]$, 考虑一个随机变量 X, 其落入区间 $[a, b]$ 内任何一个点的概率相等.

定义 4.5 若随机变量 X 的概率密度为

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

称 X 服从区间 $[a, b]$ 上的均匀分布, 记 $X \sim U(a, b)$.

对任意 $x \in (-\infty, +\infty)$ 有$f(x) \geqslant 0$, 且

$$\int_{-\infty}^{+\infty} f(t)dt = \int_{-\infty}^{a} f(t)dt + \int_{a}^{b} f(t)dt + \int_{b}^{+\infty} f(t)dt = \int_{a}^{b} \frac{1}{b-a}dt = 1.$$

均匀分布的几何解释: 若 $X \sim U(a, b)$, 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, 与该区间的位置无关.

根据分布函数的定义可知 $X \sim U(a,b)$ 的分布函数为

$$F(x) = \begin{cases} 0 & x \leqslant a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geqslant b \end{cases}$$

定理 4.3 若 $X \sim U(a,b)$, 则

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

证明 根据期望和方差的定义有

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} tf(t)dt = \frac{1}{b-a}\int_a^b tdt = \frac{a+b}{2} \\ E(X^2) &= \int_{-\infty}^{+\infty} t^2 f(t)dt = \frac{1}{b-a}\int_a^b t^2 dt = \frac{a^2+ab+b^2}{3}, \end{aligned}$$

从而得到方差

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

例 4.9 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 ξ 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

$$\begin{aligned} P(A) &= P((4\xi)^2 - 4\times 4\times(\xi+2) \geqslant 0) \\ &= P((\xi+1)(\xi-2) \geqslant 0) = P(\xi \leqslant -1) + P(\xi \geqslant 2) \\ &= \int_{-3}^{-1} \frac{1}{9}dt + \int_2^6 \frac{1}{9}dt = \frac{2}{3}. \end{aligned}$$

例 4.10 已知随机变量 $X \sim U(0,1)$, 对任意 $\lambda > 0$ 求 $E[\lambda^{\max(X,1-X)}]$.

4.2.2 指数分布

定义 4.6 给定常数 $\lambda > 0$, 若随机变量 X 的密度函数

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & \text{其它}, \end{cases}$$

称 X 服从参数为 λ 的指数分布, 记 $X \sim e(\lambda)$.

指数分布一般用于时间等待等实际问题. 对任意 $x \in (-\infty, +\infty)$ 有 $f(x) \geqslant 0$, 进一步有

$$\int_{-\infty}^{+\infty} f(t)dt = \int_0^{+\infty} \lambda e^{-\lambda t}dt = \left[-e^{-\lambda t}\right]_0^{+\infty} = 1.$$

对于指数函数的分布函数: 当 $x \leqslant 0$ 时有 $F(x) = 0$; 当 $x > 0$ 时,

$$F(x) = \int_0^x \lambda e^{-\lambda t}dt = 1 - e^{-\lambda x}.$$

定理 4.4 若随机变量 $X \sim e(\lambda)$, 则

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

证明 根据连续函数的定义有

$$\begin{aligned} E(X) &= \int_0^{+\infty} t\lambda e^{-\lambda t}dt = \left[-te^{-\lambda t}\right]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda t}dt = -\frac{1}{\lambda}\left[e^{-\lambda t}\right]_0^{+\infty} = \frac{1}{\lambda}, \\ E(X^2) &= \lambda\int_0^{+\infty} t^2 e^{-\lambda t}dt = \left[-t^2 e^{-\lambda t}\right]_0^{+\infty} + \int_0^{+\infty} 2te^{-\lambda t}dt = \frac{2}{\lambda}E(X) = \frac{2}{\lambda^2}, \end{aligned}$$

于是得到 $\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = 1/\lambda^2$.

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 4.5 给定常数 $\lambda > 0$, 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 根据指数分布函数的性质: 对任意 $x > 0$, 有 $P(X > x) = 1 - F(x) = e^{-\lambda x}$, 从而直接验证 $P(X > s + t|X > t) = P(X > s)$.

指数分布是唯一具有无记忆性的连续型随机变量.

例 4.11 打一次公用电话所用时间 $X \sim e(1/10)$, 如果某人刚好在你前面使用公用电话, 求你需等待 $10 \sim 20$ 分钟的概率.

解 根据指数分布函数有

$$P(10 \leqslant X \leqslant 20) = F(20) - F(10) = e^{-1} - e^{-2} \approx 0.2325.$$

4.2.3 正态分布

定义 4.7 给定 $u \in (-\infty, +\infty)$ 和 $\sigma > 0$, 如果随机变量 X 的概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad x \in (-\infty, +\infty),$$

称 X 服从参数为 (μ, σ^2) 的正态分布 (Normal distribution), 又被称为高斯分布 (Gaussian distribution), 记 $X \sim \mathcal{N}(\mu, \sigma^2)$.

特别地, 若 $\mu = 0$ 和 $\sigma = 1$, 称 $\mathcal{N}(0, 1)$ 为标准正态分布, 其密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad x \in (-\infty, +\infty).$$

对任意 $x \in (-\infty, +\infty)$ 有 $f(x) \geqslant 0$, 进一步有

$$\begin{aligned}\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx\right)^2 &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dxdy \\ &= \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\frac{r^2}{2}} r dr = \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\frac{r^2}{2}} d\frac{r^2}{2} = 2\pi,\end{aligned}$$

这里使用极坐标变换, 由此可证 $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = 1$.

下面考虑正太分布概率密度 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 的图形:

1) 关于直线 $x = \mu$ 对称, 即 $f(\mu - x) = f(\mu + x)$.

2) 当 $x = \mu$ 时取最大值 $f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$.

3) 概率密度函数的二阶导数

$$f''(x) = \frac{1}{\sqrt{2\pi}\sigma^5} e^{-\frac{(x-\mu)^2}{2\sigma^2}} ((x-\mu)^2 - \sigma^2),$$

可得其拐点为 $x = \mu \pm \sigma$. 根据

$$\lim_{x\to\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = 0,$$

可得渐近线为 $y = 0$.

4) 当 σ 固定时, 改变 μ 的值, $f(x)$ 沿 x 轴左右平行移动, 不改变其形状.

5) 当 μ 固定时, 改变 σ 的值, 根据 $f(x)$ 的最大值 $f(\mu)=\frac{1}{\sqrt{2\pi}\sigma}$可知: 当 σ 越小, 图形越陡, $X\sim\mathcal{N}(\mu,\sigma)$ 落入 μ 附近的概率越大; 反之 σ 越大, 图形越平坦, X 落入 μ 附近的概率越小.

定理 4.6 若 $X\sim\mathcal{N}(\mu,\sigma^2)$, 则

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

若 $X\sim\mathcal{N}(0,1)$, 则 $Y=\sigma X+\mu\sim\mathcal{N}(\mu,\sigma^2)$.

证明 若 $X\sim\mathcal{N}(\mu,\sigma^2)$, 随机变量 Y 的分布函数

$$F_Y(y)=P[Y\leqslant y]=P[X-\mu\leqslant y\sigma]=P[X\leqslant y\sigma+\mu]=\int_{-\infty}^{\mu+y\sigma}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

令 $x=(t-\mu)/\sigma$, 代入得到分布函数

$$F_Y(y)=\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx,$$

由此可得 $Y\sim\mathcal{N}(0,1)$. 若 $X\sim\mathcal{N}(0,1)$, 则

$$F_Y(y)=P(Y\leqslant y)=P(\sigma X+\mu\leqslant y)=P(X\leqslant(y-\mu)/\sigma)=\int_{-\infty}^{(y-\mu)/\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt.$$

令 $t=(x-\mu)/\sigma$, 代入得到

$$F_Y(y)=\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx.$$

定理 4.7 若 $X\sim\mathcal{N}(\mu,\sigma^2)$, 则

$$E(X)=\mu\qquad\text{和}\qquad\mathrm{Var}(X)=\sigma^2.$$

特别地, 若 $X\sim\mathcal{N}(0,1)$, 则 $E(X)=0$ 和 $\mathrm{Var}(X)=1$.

证明 若 $X\sim\mathcal{N}(0,1)$, 根据期望的定义有

$$E(X)=\int_{-\infty}^{+\infty}\frac{t}{\sqrt{2\pi}}e^{-t^2/2}dt=0$$

因为奇函数在对称的区间上积分为 0. 进一步有

$$\mathrm{Var}(X)=\int_{-\infty}^{+\infty}\frac{t^2}{\sqrt{2\pi}}e^{-t^2/2}dt=-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}tde^{-t^2/2}=\left[te^{-t^2/2}\right]_{t=-\infty}^{+\infty}+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-t^2/2}dt=1.$$

如果 $Y\sim\mathcal{N}(\mu,\sigma^2)$, 则 $(Y-\mu)/\sigma\sim\mathcal{N}(0,1)$, 于是有

$$0=E((Y-\mu)/\sigma)=(E(Y)-\mu)/\sigma\quad\Rightarrow\quad E(Y)=\mu,$$

$$1 = \mathrm{Var}((Y-\mu)/\sigma) = \mathrm{Var}(Y)/\sigma^2 \quad \Rightarrow \quad \mathrm{Var}(Y) = \sigma^2.$$

下面给出正太分布的估计:

定理 4.8 若 $X \sim \mathcal{N}(0,1)$, 对任意 $\epsilon > 0$ 有

$$P(X \geqslant \epsilon) \leqslant \frac{1}{2} e^{-\epsilon^2/2};$$

[Mill 不等式] 若$X \sim \mathcal{N}(0,1)$, 对任意$\epsilon > 0$ 有

$$P(|X| \geqslant \epsilon) \leqslant \min\left\{1, \sqrt{\frac{2}{\pi}} \frac{1}{\epsilon} e^{-\epsilon^2/2}\right\}.$$

证明 对第一个不等式, 我们有

$$\begin{aligned} P(X \geqslant \epsilon) &= \int_\epsilon^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \\ &\leqslant e^{-\epsilon^2/2} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{2} e^{-\epsilon^2/2}. \end{aligned}$$

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

$$\begin{aligned} P(|X| \geqslant \epsilon) &= 2\int_\epsilon^{+\infty} f(t) dt = 2\int_\epsilon^{+\infty} \frac{t f(t)}{t} dt \\ &\leqslant 2\int_\epsilon^{+\infty} \frac{t f(t)}{\epsilon} dt = -2\int_\epsilon^{+\infty} \frac{f'(t)}{\epsilon} dt = -\frac{2}{\epsilon} \left[f(t)\right]_\epsilon^{+\infty} = \frac{2}{\sqrt{2\pi}\epsilon} e^{-\epsilon^2/2}. \end{aligned}$$

若随机变量 $X \sim \mathcal{N}(\mu, \sigma)$, 根据 Mill 不等式有

$$P\left(\left|\frac{X-\mu}{\sigma}\right| \geqslant 3\right) \leqslant \min\left\{1, \sqrt{\frac{2}{\pi}} \frac{1}{3e^{4.5}}\right\} \leqslant 0.003,$$

由此可得若 $X \sim \mathcal{N}(\mu, \sigma)$, 则有

$$P(u - 3\sigma \leqslant X \leqslant u + 3\sigma) \geqslant 99.7\%,$$

因此在工程应用中, 一般通常认为

$$P(|X - \mu| \leqslant 3\sigma) \approx 1.$$

正太分布 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的分布函数为

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

此分布函数无闭式解, 将 $\mathcal{N}(\mu, \sigma)$ 转为为标准正太分布 $\mathcal{N}(0, 1)$, 设 $\mathcal{N}(0, 1)$ 的分布函数为

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

通过查表或计算机进行计算.

例 4.12 设随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 求概率 $P(a \leqslant X \leqslant b)$.

解 将 $\mathcal{N}(\mu, \sigma)$ 转化为标准正太分布 $\mathcal{N}(0, 1)$, 有

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

4.3 连续随机变量函数的分布

前面研究了常用的连续型随机变量, 本节进一步研究连续随机变量的函数. 该问题可描述为: 对给定的连续函数 $g(x) : \mathbb{R} \to \mathbb{R}$, 已知连续随机变量 X 的概率密度为 $f_X(x)$, 求解新的随机变量 $Y = g(X)$ 的概率密度 $f_Y(y)$? 该问题的求解一般可分为如下两步:

1) 求解 $Y = g(X)$ 的分布函数 $F_Y(y) = P(Y \leqslant y) = P(g(X) \leqslant y) = \int_{g(x) \leqslant y} f_X(x)dx$;

2) 利用分布函数和概率密度之间的关系求解密度函数 $f_Y(y) = F_Y'(y)$.

求解该问题常用到的数学工具为积分求导公式: 设函数 $F(y) = \int_{\psi(y)}^{\varphi(y)} f(x)dx$, 则有

$$F'(y) = f(\varphi(y))\varphi'(y) - f(\psi(y))\psi'(y).$$

例 4.13 设连续型随机变量 X 的密度函数

$$f(x) = \begin{cases} x/8 & 0 < x < 4 \\ 0 & \text{其它}, \end{cases}$$

求 $Y = 2X + 8$ 的密度函数.

解 求解分布函数

$$F_Y(y) = P(Y \leqslant y) = P(2X + 8 \leqslant y) = P(X \leqslant \frac{y-8}{2}) = F_X(\frac{y-8}{2}),$$

可得密度函数

$$f_Y(y)=f_X(\frac{y-8}{2})\cdot\frac{1}{2}=\begin{cases}\frac{y-8}{32} & \frac{y-8}{2}\in[0,4]\\ 0 & \text{其它}\end{cases}=\begin{cases}\frac{y-8}{32} & y\in[8,16]\\ 0 & \text{其它.}\end{cases}$$

例 4.14 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 首先求解分布函数 $F_Y(y)=P(Y\leqslant y)=P(X^2\leqslant y)$. 当 $y\leqslant 0$ 时有 $F_Y(y)=0$; 当 $y>0$ 时有

$$F_Y(y)=P(X^2\leqslant y)=P(-\sqrt{y}\leqslant x\leqslant\sqrt{y})=\int_{-\sqrt{y}}^{\sqrt{y}}f_X(x)dx,$$

进一步得到密度函数

$$f_Y(y)=F_Y'(y)=f_X(\sqrt{y})\cdot\frac{1}{2\sqrt{y}}+f_X(-\sqrt{y})\cdot\frac{1}{2\sqrt{y}}.$$

最后得到

$$f_Y(y)=\begin{cases}\frac{1}{2\sqrt{y}}(f_X(-\sqrt{y})+f_X(\sqrt{y})) & y>0\\ 0 & y\leqslant 0\end{cases}$$

例 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 $f_Y(y)$.

若连续函数 $g(x)$ 满足一定的条件, 可以直接写出概率密度函数:

定理 4.9 设随机变量 X 的概率密度为 $f_X(x)$, 其中 $x\in(-\infty,+\infty)$. 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 证明类似于前面的思路, 这里不妨假设 $g'(x)>0$, 可同理考虑 $g'(x)<0$). 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 $y\leqslant\alpha$ 时, 有 $F_Y(y)=0$; 当 $y\geqslant\beta$ 时有 $F_Y(y)=1$; 当 $\alpha<y<\beta$ 时,

$$F_Y(y)=P(g(X)<y)=P(X\leqslant h(y))=F(h(y)).$$

于是可得随机变量 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 4.10 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 4.9 可知

$$f_Y(y) = \frac{1}{a} f_X(\frac{y-b}{a}) = \frac{1}{a}\frac{1}{\sqrt{2\pi}\sigma} e^{-(\frac{y-b}{a}-\mu)^2)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi}a\sigma} e^{-(y-b-a\mu)^2/2a^2\sigma^2},$$

由此证明 $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

例 4.16 连续随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 证明 $Y = e^X$ 的概率密度函数为

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}y\sigma} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}} & y > 0 \\ 0 & y \leqslant 0 \end{cases}$$

定理 4.11 设随机变量 X 的分布函数是严格单调的连续函数, 则 $Y = F(X) \sim U(0, 1)$.

证明 令 $Y = F(X)$ 的分布函数为 $G(y)$, 则

$$G(y) = P(Y \leqslant y) = P(F(X) \leqslant y).$$

由于分布函数 $F(x) \in [0, 1]$, 所以当 $y < 0$ 时有 $G(y) = 0$; 当 $y \geqslant 1$ 时有 $G(y) = 1$; 当 $y \in [0, 1]$ 时, 由于 $F(X)$ 严格单调, 所以 $F^{-1}(y)$ 存在且严格单调, 于是有 $G(y) = P(F(X) \leqslant y) = P(X \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y$. 于是得到分布函数

$$G(y) = \begin{cases} 0 & y < 0 \\ y & 0 \leqslant y \leqslant 1 \\ 1 & y \geqslant 1. \end{cases},$$

以及密度函数

$$f_Y(y) = \begin{cases} 1 & y \in [0, 1] \\ 0 & \text{其它}. \end{cases}$$

习题

4.1 书57 58页: 18, 19, 20, 21, 23, 24, 25 题, 书 115 页第 18 题.

4.2 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.

4.3 设随机变量 X 的概率密度为

$$f(x)=\begin{cases}Ae^{-x} & x>0\\ 0 & x\leqslant 0.\end{cases}$$

求 $Y=e^{-2X}$ 的期望.

4.4 证明

$$\int_{-\infty}^{+\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt=\sqrt{2\pi}\sigma.$$

4.5 设随机变量 X 的分布函数是严格单调的连续函数 $F(x)$，证明 $Y=F(X)\sim U(0,1)$

4.6 若 $X\sim N(0,1)$,对任意实数 $\epsilon>0$, 求证

$$P(X\geqslant\epsilon)\geqslant\frac{1}{3}e^{-\frac{(\epsilon+1)^2}{2}}$$

4.7 查资料学习Γ分布并计算期望与方差. Γ分布的概率密度函数为

$$f(x)=\begin{cases}\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & x>0\\ 0 & x\leqslant 0.\end{cases}$$

4.8 书58-59页: 26, 32, 34, 35, 36, 37题.

第 5 章　多维随机变量及其分布

前面所研究的随机现象均由单一因素决定, 即一维随机变量. 然而实际问题中很多随机现象往往由两个或多个随机因素造成的, 需用多个随机变量描述. 例如: 导弹攻击点的坐标 (经度、纬度); 学生的高考成绩 (语文、数学、英语等).

定义 5.1　设 $X = X(\omega)$ 和 $Y = Y(\omega)$ 为定义在样本空间 Ω 上的随机变量, 由它们构成的向量 (X, Y) 称为二维随机变量.

二维随机向量又称为二维随机变量, 需要将 (X, Y) 看作一个整体, 不能分开看待, 在几何上 (X, Y) 可看作平面上的随机点.

5.1 二维随机变量的分布函数

首先研究二维随机变量的分布函数:

定义 5.2　设 (X, Y) 为二维随机变量, 对任意实数 $x \in (-\infty, +\infty)$ 和 $y \in (-\infty, +\infty)$,

$$F(x, y) = P(X \leqslant x, Y \leqslant y)$$

称为二维随机变量 (X, Y) 的分布函数, 或称为随机变量 X 和 Y 的联合分布函数.

二维随机变量分布函数 $F(x, y)$ 几何意义: 随机点 (X, Y) 落入以 (x, y) 为右上定点无穷矩形的概率. 对二维随机变量分布函数, 有如下性质:

- 分布函数 $F(x, y)$ 对每个变量单调不减: 固定 y, 当 $x_1 > x_2$ 时有 $F(x_1, y) \geqslant F(x_2, y)$; 同理固定 x, 当 $y_1 > y_2$ 时有 $F(x, y_1) \geqslant F(x, y_2)$;
- 对任意 $x \in (-\infty, +\infty)$ 和 $y \in (-\infty, +\infty)$, 分布函数 $F(x, y) \in [0, 1]$, 且

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

- 分布函数 $F(x, y)$ 关于每个变量右连续.

根据分布函数可推导概率:

$$P(x_1 < X \leqslant x_2, y_1 < Y \leqslant y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1).$$

设随机变量 (X, Y) 的联合分布函数为 $F(x, y)$, 将随机变量 X 和 Y 单独看依然是随机变量. 可以根据随机变量的联合分布函数 $F(x, y)$ 研究随机变量 X 和 Y 的分布函数 $F_X(x)$ 和 $F_Y(y)$.

定义 5.3 设二维随机变量 (X,Y) 的联合分布函数为 $F(x,y)$, 称

$$F_X(x)=P(X\leqslant x)=P(X\leqslant x,y<+\infty)=F(x,+\infty)=\lim_{y\to+\infty}F(x,y),$$

为随机变量 X 的边缘分布函数. 同理定义随机变量 Y 的边缘分布函数为:

$$F_Y(y)=P(Y\leqslant y)=P(Y\leqslant y,x<+\infty)=F(+\infty,y)=\lim_{x\to+\infty}F(x,y).$$

例 5.1 设二维随机变量 (X,Y) 的联合分布函数为

$$F(x,y)=A(B+\arctan\frac{x}{2})(C+\arctan\frac{y}{3})(x,y\in\mathbb{R}).$$

求随机变量 X 与 Y 的边缘分布函数和概率 $P(Y>3)$.

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

$$F_X(x)=\lim_{y\to\infty}\frac{1}{\pi^2}(\frac{\pi}{2}+\arctan\frac{x}{2})(\frac{\pi}{2}+\arctan\frac{y}{3})=\frac{1}{\pi}(\frac{\pi}{2}+\arctan\frac{x}{2})$$

同理可得

$$F_Y(y)=\frac{1}{\pi}(\frac{\pi}{2}+\arctan\frac{y}{3})$$

最后得到

$$P(Y>3)=1-P(Y\leqslant 3)=1-F_Y(3)=1-\left(\frac{1}{2}+\frac{1}{\pi}\arctan 1\right)=\frac{1}{4}.$$

前面讲过独立的随机事件 A 和 B 有 $P(AB)=P(A)P(B)$, 下面介绍随机变量的独立性:

定义 5.4 设 X,Y 为二维随机变量, 对任意 $x\in(-\infty,+\infty)$ 和 $y\in(-\infty,+\infty)$, 若事件 $X\leqslant x$ 和 $Y\leqslant y$ 相互独立, 即

$$P(X\leqslant x,Y\leqslant y)=P(X\leqslant x)\cdot P(Y\leqslant y)\quad\Leftrightarrow\quad F(x,y)=F_X(x)F_Y(y),$$

则称随机变量 X 与 Y 相互独立.

设随机变量 X 与 Y 相互独立, 则 $f(X)$ 与 $g(Y)$ 也相互独立, 其中 $f(x)$ 和 $g(y)$ 是连续或分段连续函数. 例如: 若随机变量 X 与 Y 相互独立, 则 X^2 与 Y^3 相互独立, $\sin X$ 与 $\cos Y$ 相互独立.

5.2 二维离散型随机变量

定义 5.5 若二维随机变量 (X,Y) 的取值是有限个或无限可列的, 称 (X,Y) 为二维离散型随机变量. 设离散型随机变量 (X,Y) 的取值分别为 $(x_i,y_j), i=1,2,\cdots, j=1,2,\cdots$, 则称

$$p_{ij}=P(X=x_i,Y=y_j)$$

为 (X,Y) 的联合分布列. 二维随机变量的联合分布列可表示为

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$

根据分布列的性质可知 $p_{ij}\geqslant 0$ 和 $\sum_{i,j}p_{ij}=1$. 根据二维随机变量 (X,Y) 的联合分布列 p_{ij}, 可得到随机变量 X 的边缘分布列

$$P(X=x_i)=\sum_{j=1}^{\infty}P(X=x_i,Y=y_j)=\sum_{j=1}^{\infty}p_{ij}=p_{i\cdot}\ .$$

同理可得随机变量 Y 的边缘分布列

$$P(Y=y_j)=\sum_{i=1}^{\infty}P(X=x_i,Y=y_j)=\sum_{i=1}^{\infty}p_{ij}=p_{\cdot j}\ .$$

将二维随机变量的联合分布列和边缘分布表示在同一个表格中有

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	$p_{i\cdot}$
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}	$\cdots$	$p_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$
$p_{\cdot j}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	1

例 5.2 有三个数 $1,2,3$, 随机变量 X 表示从这三个数中随机地抽取一个数, 随机变量 Y 表示从 1 到 X 中随机抽取一个数. 求 (X,Y) 的联合分布列和边缘分布列.

解 由题意可知随机变量 X 和 Y 的取值为 $1,2,3$: 当 $X=1$ 时有 $Y=1$; 当 $X=2$ 时有 Y 等可能取 $1,2$; 当 $X=3$ 时 Y 等可能取 $1,2,3$. 从而得到

X \ Y	1	2	3	$p_{i\cdot}$
1	1/3	0	0	1/3
2	1/6	1/6	0	1/3
3	1/9	1/9	1/9	1/3
$p_{\cdot j}$	11/18	5/18	1/9	1

定义 5.6 对离散型随机变量 (X,Y), 若对所有 (x_i,y_j) 有

$$P(X=x_i,Y=y_j)=P(X=x_i)P(Y=y_j),\quad \text{即}\quad p_{ij}=p_{i\cdot}p_{\cdot j}$$

称离散随机变量 X 与 Y 相互独立.

定理 5.1 对二维离散型随机变量 (X,Y), 定义 5.4 与上述定义等价, 即对所有 (x_i,y_j) 有

$$F(x_i,y_j)=F_X(x_i)F_Y(y_j)\quad\Longleftrightarrow\quad p_{i,j}=p_{i\cdot}p_{\cdot j}.$$

证明 首先证明必要性, 根据定义 5.4 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, 根据定义 5.6 有

$$F(x_i,y_j)=\sum_{l\leqslant i}\sum_{k\leqslant j}p_{lk}=\sum_{l\leqslant i}\sum_{k\leqslant j}p_{l\cdot}p_{\cdot k}=\sum_{l\leqslant i}p_{l\cdot}\sum_{k\leqslant j}p_{\cdot k}=F_X(x_i)F_Y(y_j).$$

定理 5.2 设离散随机变量 X 和 Y 独立, 对任意集合 $A,B\in\mathbb{R}$, 有事件 $X\in A$ 和 $Y\in B$ 独立.

证明 对离散型随机变量, 不放假设 $A=\{x_1,x_2,\ldots,x_k\}$ 和 $B=\{y_1,y_2,\ldots,y_l\}$, 则有

$$P(X\in A,Y\in B)=\sum_{i=1}^{k}\sum_{j=1}^{l}p_{ij}=\sum_{i=1}^{k}\sum_{j=1}^{l}p_{i\cdot}p_{\cdot j}=\sum_{i=1}^{k}p_{i\cdot}\sum_{j=1}^{l}p_{\cdot j}=P(X\in A)P(Y\in B).$$

例 5.3 设离散型 X,Y 独立, 求解 (X,Y) 的联合分布律为

X \ Y	y_1	y_2	y_3	$p_{i\cdot}$
x_1		$1/8$		
x_2	$1/8$			
$p_{\cdot j}$	$1/6$			

求解可得

X \ Y	y_1	y_2	y_3	$p_{i\cdot}$
x_1	$1/24$	$1/8$	$1/12$	$1/4$
x_2	$1/8$	$3/8$	$1/4$	$3/4$
$p_{\cdot j}$	$1/6$	$1/2$	$1/3$	1

例 5.4 将两个球 A,B 放入编号为 $1,2,3$ 的三个盒子中, 用随机变量 X 放入 1 号盒的球数, 用随机变量 Y表示 放入 2 号盒的球数, 判断 X 和 Y 是否独立.

解 由题意可知

X \ Y	0	1	2	$p_{i\cdot}$
0	$1/9$	$2/9$	$1/9$	$4/9$
1	$2/9$	$2/9$	0	$4/9$
2	$1/9$	0	0	$1/9$
$p_{\cdot j}$	$4/9$	$4/9$	$1/9$	

由此可知 $P(X=2,Y=2)\neq P(X=2)P(Y=2)$, 所以 X 和 Y 不独立.

5.3 二维连续型随机变量

定义 5.7 设二维随机变量的分布函数为 $F(x,y)$, 如果存在二元非负可积函数 $f(x,y)$ 使得对任意实数对 (x,y) 有

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)dudv,$$

则称 (X,Y) 为二维连续型随机变量, 称 $f(x,y)$ 称为二维随机变量 (X,Y) 的概率密度, 或称为随机变量 X 和 Y 的联合概率密度.

根据概率密度的定义可知概率密度函数 $f(x,y)$ 满足如下性质:

1) $f(x,y)\geqslant 0$.

2) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x,y)dxdy=1$.

3) 若 $f(x,y)$ 在 (x,y) 连续, 则 $f(x,y)=\frac{\partial^2 F(x,y)}{\partial x\partial y}$.

4) 若 G 为平面上的一个区域, 则点 (X,Y) 落入 G 的概率为

$$P((X,Y)\in G)=\int\int_{(x,y)\in G}f(x,y)dxdy.$$

例 5.5 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 $P(0<X<1,0<Y<2)$.

解 根据概率密度的性质可知

$$1=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(x,y)dxdy=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}ce^{-(3x+4y)}dxdy=\frac{c}{12},$$

由此可得 $c=12$. 进一步可得

$$P(0<X<1,0<Y<2)=12\int_0^1\int_0^2 e^{-(3x+4y)}dxdy=(1-e^{-3})(1-e^{-8}).$$

例 5.6 设二维随机变量 (X,Y) 的概率密度

$$f(x,y)=\begin{cases} x^2+axy & 0\leqslant x\leqslant 1,0\leqslant y\leqslant 2 \\ 0 & \text{其它,} \end{cases}$$

求 $P(X+Y\geqslant 1)$.

给定二维随机变量的联合概率密度 $f(x,y)$, 下面定义随机变量 X 和 Y 的边缘概率密度:

定义 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 X 和 Y 的边缘概率密度为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy \qquad f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx.$$

上述的边缘概率密度可完全根据边缘分布函数 $F_X(x)$ 的定义导出, 首先可知随机变量 X 的边缘分布函数为

$$\begin{aligned} F_X(x) &= P(X\leqslant x)=P(X\leqslant x,Y<\infty)=F(x,+\infty) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{+\infty}f(t,y)dtdy=\int_{-\infty}^{x}\left(\int_{-\infty}^{+\infty}f(t,y)dy\right)dt, \end{aligned}$$

由此可得随机变量 X 的边缘概率密度为

$$f_X(x)=F_X'(x)=\int_{-\infty}^{+\infty}f(x,y)dy.$$

例 5.7 设二维随机变量 (X,Y) 的概率密度

$$f(x,y)=\begin{cases}cxy & 0\leqslant x\leqslant y\leqslant 1\\ 0 & \text{其它},\end{cases}$$

求 $P(X\leqslant 1/2)$.

解 根据概率密度的性质有

$$1=\int_0^1\int_x^1 cxydydx=c\int_0^1 x(1-x^2)/2dx=c/8,$$

由此可解 $c=8$. 当 $0\leqslant x\leqslant 1$ 时随机变量 X 的边缘概率密度为

$$f_X(x)=\int_{-\infty}^{\infty}f(x,y)dy=\int_x^1 8xydy=4x(1-x^2),$$

进一步有

$$P(X\leqslant\frac{1}{2})=\int_0^{\frac{1}{2}}4x(1-x^2)dx=\frac{7}{16}.$$

下面定义二维连续随机变量的独立性:

定义 5.9 对任意 $x\in(-\infty,+\infty)$ 和 $y\in(-\infty,+\infty)$, 若二维连续随机变量 (X,Y) 的概率密度与边缘概率密度满足

$$f(x,y)=f_X(x)f_Y(y),$$

则称随机变量 X 和 Y 相互独立.

对连续随机变量, 上述独立性定义与基于分布函数的独立性 (定义 5.4) 等价, 即有如下定理:

定理 5.3 设二维随机变量 (X,Y) 的概率密度为 $f(x,y)$, 则有

$$F(x,y)=F_X(x)F_Y(y)\quad\Longleftrightarrow\quad f(x,y)=f_X(x)f_Y(y).$$

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

$$\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)dudv=\int_{-\infty}^{x}f_X(u)du\int_{-\infty}^{y}f_Y(v)dv$$

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

$$F(x,y)\quad=\quad\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)dudv=\int_{-\infty}^{x}\int_{-\infty}^{y}f_X(u)f_Y(v)dudv$$

$$= \int_{-\infty}^{x} f_X(u)du \int_{-\infty}^{y} f_Y(v)dv = F_X(x)F_Y(y).$$

例 5.8 设二维随机变量的密度函数

$$f(x,y) = \begin{cases} cxe^{-y} & 0 < x < y < +\infty \\ 0 & \text{其它}, \end{cases}$$

问 X 与 Y 是否独立.

解 根据概率密度的性质有

$$1 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)dxdy = c \int_{0}^{+\infty} dy \int_{0}^{y} xe^{-y}dx = c.$$

当 $x > 0$ 时随机变量 X 的边缘概率密度为

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_{x}^{+\infty} xe^{-y}dy = xe^{-x}.$$

同理当 $y > 0$ 时随机变量 Y 的边缘概率密度为

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx = \int_{0}^{y} xe^{-y}dx = \frac{1}{2}y^2e^{-y}.$$

由此可得随机变量 X 与 Y 不独立.

例 5.9 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 $P(X + Y \leqslant 1)$.

解 根据均匀分布和指数分布的定义有随机变量 X 与 Y 的边缘概率密度为

$$f_X(x) = \begin{cases} \frac{1}{2} & x \in [-1,1] \\ 0 & \text{其它} \end{cases} \quad \text{和} \quad f_Y(y) = \begin{cases} 2e^{-2y} & y \geqslant 0 \\ 0 & \text{其它} \end{cases}$$

根据随机变量的独立性可得随机变量 X 与 Y 的联合概率密度

$$f(x,y) = \begin{cases} e^{-2y} & -1 \leqslant x \leqslant 1, y \geqslant 0 \\ 0 & \text{其它} \end{cases}.$$

由此可得

$$P(X + Y \leqslant 1) = \int_{-1}^{1} dx \int_{0}^{1-x} e^{-2y}dy = \frac{3}{4} + \frac{1}{4}e^{-4}.$$

对常见的二维随机变量, 我们这里仅仅考虑二维正态 (Gaussian) 分布. 其定义如下

定义 5.10 设 $|\rho|<1$, 令

$$\mu=(\mu_x,\mu_y)^\top=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}\qquad 和\qquad \Sigma=\begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix}.$$

若随机变量 X 和 Y 的联合概率密度函数为

$$\begin{aligned}f(x,y)&=(2\pi)^{-2/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\xi-\mu)^\top\Sigma^{-1}(\xi-\mu)\right)\qquad \xi=(x,y)^\top\\ &=\frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho}{\sigma_x\sigma_y}(x-\mu_x)(y-\mu_y)\right]\right)\end{aligned}$$

这里利用 $|\Sigma|=(1-\rho^2)\sigma_x^2\sigma_y^2$, 以及

$$\Sigma^{-1}=\frac{1}{(1-\rho^2)\sigma_x^2\sigma_y^2}\begin{pmatrix}\sigma_y^2 & -\rho\sigma_x\sigma_y\\ -\rho\sigma_x\sigma_y & \sigma_x^2\end{pmatrix}=\frac{1}{1-\rho^2}\begin{pmatrix}1/\sigma_x^2 & -\rho/\sigma_x\sigma_y\\ -\rho/\sigma_x\sigma_y & 1/\sigma_y^2\end{pmatrix},$$

则称随机变量 X 和 Y 服从参数为 μ 和 Σ 的正太分布, 记为 $(X,Y)\sim\mathcal{N}(\mu,\Sigma)$.

若 $\mu=(0,0)^\top$ 和 Σ 为二维单位阵, 则称为二维标准正态分布. 下面研究二维正态分布的性质:

定理 5.4 设二维随机变量 (X,Y) 服从正态分布 $\mathcal{N}(\mu,\Sigma)$, 其中

$$\mu=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}\qquad 和\qquad \Sigma=\begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix},$$

则有边缘分布为 $X\sim\mathcal{N}(\mu_x,\sigma_x^2)$ 和 $Y\sim\mathcal{N}(\mu_y,\sigma_y^2)$.

证明 根据边缘概率密度的定义和正态分布性质有

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}e^{-\frac{1}{2(1-\rho^2)}[(\frac{y-\mu_y}{\sigma_y})^2-2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y}+(\frac{x-\mu_x}{\sigma_x})^2]}dy$$

令 $t=\frac{y-\mu_y}{\sigma_y}$, 有 $dy=\sigma_y dt$, 进一步得到

$$f_X(x)=\frac{e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2}}{2\pi\sqrt{1-\rho^2}\sigma_x}\int_{-\infty}^{+\infty}e^{-\frac{(t-\rho(x-\mu_x)/\sigma_x)^2}{2(1-\rho^2)}}dt=\frac{e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2}}{\sqrt{2\pi}\sigma_x},$$

这里利用了正态分布 $N(\rho(x-\mu_x)/\sigma_x,1-\rho^2)$ 的密度函数满足

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

上述定理说明正太分布的边缘分布还是正太分布.

定理 5.5 若二维随机变量 $(X,Y)\sim\mathcal{N}(\mu,\Sigma)$, 则 X 与 Y 独立的充要条件为 $\Sigma=\begin{pmatrix}\sigma_x^2 & 0\\ 0 & \sigma_y^2\end{pmatrix}$.

证明 若随机变量 $(X,Y)\sim\mathcal{N}(\mu,\Sigma)$, 根据定理 5.4 可知

$$f_X(x)f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_x}e^{-(x-\mu_x)^2/2\sigma_x^2}\times\frac{1}{\sqrt{2\pi}\sigma_y}e^{-(y-\mu_y)^2/2\sigma_y^2}.$$

必要性证明: 当 $\Sigma=\begin{pmatrix}\sigma_x^2 & 0\\ 0 & \sigma_y^2\end{pmatrix}$ 时有

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}[(\frac{x-\mu_x}{\sigma_x})^2+(\frac{y-\mu_y}{\sigma_y})^2]}=f_X(x)f_Y(y).$$

充分性证明: 若 X 与 Y 独立, 则对任意 $x\in(-\infty,+\infty)$ 和 $y\in(-\infty,+\infty)$ 有 $f(x,y)=f_X(x)f_Y(y)$ 成立, 即

$$\begin{aligned}&\frac{1}{\sqrt{2\pi}\sigma_x}e^{-(x-\mu_x)^2/2\sigma_x^2}\times\frac{1}{\sqrt{2\pi}\sigma_y}e^{-(y-\mu_y)^2/2\sigma_y^2}\\ &\quad=\frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho}{\sigma_x\sigma_y}(x-\mu_x)(y-\mu_y)\right]\right)\end{aligned}$$

令 $x=\mu_x$ 和 $y=\mu_y$, 代入上式求解可得 $\rho=0$.

下面进一步研究多维正态 (Gaussian) 分布, 其定义如下:

定义 5.11 设向量 $\mu\in\mathbb{R}^n$ 和正定矩阵 $\Sigma\in\mathbb{R}^{n\times n}$, 若随机向量 $(X_1,X_2,\cdots,X_n)$ 的概率密度函数为

$$f(x_1,\ldots,x_n)=(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\xi-\mu)^\top\Sigma^{-1}(\xi-\mu)\right)$$

其中 $\xi=(x_1,\ldots,x_n)^\top$, 则称随机向量 $(X_1,X_2,\cdots,X_n)$ 服从参数为 μ 和 Σ 的多维正态分布, 记

$$(X_1,X_2,\cdots,X_n)\sim\mathcal{N}(\mu,\Sigma).$$

对多维正态分布, 有如下定理:

定理 5.6 设随机向量 $(X,Y)=(X_1,X_2,\cdots,X_n,Y_1,Y_2,\cdots,Y_m)\sim\mathcal{N}(\mu,\Sigma)$, 其中

$$\mu=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\quad \mu_x=(\mu_{x_1},\mu_{x_2},\cdots,\mu_{x_n})^\top,\quad \mu_y=(\mu_{y_1},\mu_{y_2},\cdots,\mu_{y_m})^\top,\quad \Sigma=\begin{pmatrix}\sum_{xx} & \sum_{yx}\\ \sum_{xy} & \sum_{yy}\end{pmatrix},$$

则有

- 随机向量 X 和 Y 分布服从 $X \sim \mathcal{N}(\mu_x, \Sigma_{xx})$和$Y \sim \mathcal{N}(\mu_y, \Sigma_{yy})$;
- 随机向量 X 与 Y 相互独立的充要条件是 $\Sigma = \begin{pmatrix} \sum_{xx} & 0 \\ 0 & \sum_{yy} \end{pmatrix}$.

这里我们回顾正定矩阵的特征值分解, 对正定矩阵 Σ, 其特征值分解为

$$\Sigma = U^\top \Lambda U$$

其中 $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$ 为特征值构成的对角阵, U 为特征向量构成的正交矩阵. 我们有如下多维正态分布的标准正态化:

定理 5.7 若随机向量 $X = (X_1, X_2, \cdots, X_n) \sim \mathcal{N}(\mu, \Sigma)$, 且正定矩阵 Σ 的特征值分解为 $\Sigma = U^\top \Lambda U$, 则随机向量

$$Y = \Lambda^{-1/2} U(X - \mu) \sim \mathcal{N}(\mathbf{0}_n, I_n),$$

其中 $\mathbf{0}_n$ 为全为零的 n 维向量, I_n 表示 $n \times n$ 的单位阵.

证明 根据 $Y = \Lambda^{-1/2} U(X - \mu)$ 可得 $X = U^\top \Lambda^{1/2} Y + \mu$, 已知 X 的概率密度函数为

$$p_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \mu)^\top \Sigma^{-1} (\boldsymbol{x} - \mu) \right)$$

根据函数的概率密度公式有

$$p_Y(\boldsymbol{y}) = p_X(U^\top \Lambda^{1/2} \boldsymbol{y} + \mu) |U^\top \Lambda^{1/2}| = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \boldsymbol{y}^\top \boldsymbol{y} \right).$$

由此可得完成证明.

下面研究多维标准正态分布的一些特征:

定理 5.8 若随机向量 $X = (X_1, X_2, \cdots, X_n) \sim \mathcal{N}(\mathbf{0}_n, I_n)$, 则其概率密度函数为 $p_X(\boldsymbol{x})$, 则有

$$\int p_X(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

证明 根据概率密度的定义有

$$\int \int p_X(\boldsymbol{x}) d\boldsymbol{x} = \int \int \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \boldsymbol{x}^\top \boldsymbol{x} \right) d\boldsymbol{x} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x_i^2/2} dx_i = 1.$$

定理 5.9 若随机向量 $X = (X_1, X_2, \cdots, X_n) \sim \mathcal{N}(\mu, \Sigma)$, 则

$$Y = AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$$

其中 $A \in \mathbb{R}^{m\times n}$ 和 $b \in \mathbb{R}^{m\times 1}$.

证明作为练习题, 仅仅要求证明 $m = n$ 且 $|A| \neq 0$ 的情况.

5.4 多维随机变量函数的分布

本节我们研究已知 (X, Y) 的分布, 求 $Z = g(X, Y)$ 的分布. 若 $X_1, X_2, \cdots, X_n$ 是 n 维离散型随机变量, 那么 $Z = g(X_1, X_2, \cdots, X_n)$ 为一维随机变量, 其分布列可以通过如下两步求得:

i) 对 $X_1, X_2, \cdots, X_n$ 的各种取值, 计算随机变量 Z 的取值;

ii) 对相同的 Z 值, 合并其概率.

例 5.10 设 (X, Y) 的联合分布列为

X \ Y	0	1	2
0	1/4	1/6	1/8
1	1/4	1/8	1/12

求 $Z_1 = X + Y$ 和 $Z_2 = XY$ 的分布列.

解 通过简单计算、合并可得 Z_1 和 Z_2 的分布列分别为:

Z_1	0	1	2	3
P	1/4	5/12	1/4	1/12

Z_2	0	1	2
P	19/24	1/8	1/12

.

对于连续随机变量 (X, Y), 其联合概率密度为 $f(x, y)$, 如何求解随机变量 $Z = g(X, Y)$ 的概率密度. 针对此类问题, 主要求解思路为分布函数法, 即:

i) 求 $Z = g(X, Y)$ 的分布函数

$$F_Z(z) = P(Z \leqslant z) = P(g(x, y) \leqslant z) = \int\int_{g(x,y)\leqslant z} f(x, y)dxdy.$$

ii) 求 Z 的密度函数

$$f_Z(z) = F_Z'(z).$$

5.4.1 极大极小分布

设随机变量 X 和 Y 相互独立, 其分布函数分别为 $F_X(x)$ 和 $F_Y(y)$, 求随机变量

$$Z_1 = \max(X, Y) \quad \text{和} \quad Z_2 = \min(X, Y)$$

的分布函数和概率密度函数. 首先求 Z_1 的分布函数为

$$\begin{aligned} F_{Z_1}(z_1) &= P(Z_1 \leqslant z_1) \\ &= P(\max(X,Y) \leqslant z_1) = P(X \leqslant z_1, Y \leqslant z_1) \\ &= P(X \leqslant z_1)P(Y \leqslant z_1) = F_X(z_1)F_Y(z_1). \end{aligned}$$

进一步求解 Z_2 的分布函数为

$$\begin{aligned} F_{Z_2}(z_2) &= P(Z_2 \leqslant z_2) \\ &= P(\min(X,Y) \leqslant z_2) = 1 - P(\min(X,Y) > z_2) \\ &= 1 - P(X \geqslant z_2)P(Y > z_2) = 1 - (1 - F_X(z_2))(1 - F_Y(z_2)). \end{aligned}$$

上述结论可进一步推广到 n 个独立的随机变量有

引理 5.1 设 $X_1, X_2, \cdots, X_n$ 为 n 个相互独立的随机变量, 其分布函数分别为 $F_{X_i}(x_i)$, 则随机变量 $Y = \max(X_1, X_2, \cdots, X_n)$ 的分布函数为

$$F_Y(y) = F_{X_1}(y)F_{X_2}(y)\cdots F_{X_n}(y),$$

随机变量 $Z = \min(X_1, X_2, \cdots, X_n)$ 的分布函数为

$$F_Z(z) = 1 - (1 - F_{X_1}(z))(1 - F_{X_2}(z))\cdots(1 - F_{X_n}(z)).$$

特别地, 当 $X_1, X_2, \cdots, X_n$ 独立同分布时, 则有

$$F_Y(y) = (F_{X_1}(y))^n \quad \text{和} \quad F_Z(z) = 1 - (1 - F_{X_1}(z))^n.$$

根据分布函数可进一步求得概率密度.

例 5.11 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

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x} & x \geqslant 0 \\ 0 & x \leqslant 0 \end{cases} \quad \text{和} \quad f_Y(y) = \begin{cases} \beta e^{-\beta y} & y \geqslant 0 \\ 0 & y \leqslant 0. \end{cases}$$

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 $z_1 \leqslant 0$ 时由 $F_{Z_1}(z_1) = 0$; 当 $z_1 > 0$ 时

$$F_{Z_1}(z_1) = \int_0^{z_1} f_X(t)dt \int_0^{z_1} f_Y(t)dt = \int_0^{z_1} \alpha e^{-\alpha t}dt \int_0^{z_1} \beta e^{-\beta y}dt = (1 - e^{-\alpha z_1})(1 - e^{-\beta z_1}).$$

两边对 z_1 求导可得其概率密度为

$$f_{Z_1}(z_1) = \begin{cases} \alpha e^{-\alpha z_1} + \beta e^{-\beta z_1} - (\alpha + \beta)e^{-(\alpha+\beta)z_1} & z_1 \geqslant 0 \\ 0 & z_1 < 0. \end{cases}$$

同理可得随机变量 Z_2 的分布函数和概率密度分别为

$$F_{Z_2}(z_2) = \begin{cases} 1 - e^{-(\alpha+\beta)z_2} & z_2 \geqslant 0 \\ 0 & z_2 < 0 \end{cases} \qquad f_{Z_2}(z_2) = \begin{cases} (\alpha + \beta)e^{-(\alpha+\beta)z_2} & z_2 \geqslant 0 \\ 0 & z_2 < 0. \end{cases}$$

5.4.2 和的分布 $Z = X + Y$

引理 5.2 设随机变量 (X, Y) 的联合密度为 $f(x, y)$, 则随机变量 $Z = X + Y$ 的概率密度为

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - x)dx \quad \text{或} \quad f_Z(z) = \int_{-\infty}^{+\infty} f(z - y, y)dy.$$

解 首先求解分布函数

$$\begin{aligned} F_Z(z) &= P(Z \leqslant z) = P(X + Y \leqslant z) \\ &= \int\int_{x+y\leqslant z} f(x, y)dxdy = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{z-x} f(x, y)dy \\ &= \int_{-\infty}^{+\infty} dx \int_{-\infty}^{z} f(x, u - x)du \qquad (\text{变量替换} u = y + x) \\ &= \int_{-\infty}^{z} (\int_{-\infty}^{+\infty} f(x, u - x)dx)du \end{aligned}$$

两边同时对 z 求导数可得

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - x)dx.$$

下面给出著名的卷积公式:

定理 5.10 若连续随机变量 X 与 Y 相互独立, 其概率密度函数分别为 $f_X(x)$ 和 $f_Y(y)$, 则随机变量 $Z = X + Y$ 的密度函数为

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z - x)dx = \int_{-\infty}^{+\infty} f_X(z - y)f_Y(y)dy.$$

若离散随机变量 X 与 Y 独立, 其分布列为 $a_i = P(X = i)$ 和 $b_j = P(Y = j)$ $(i, j = 0, 1, \cdots)$, 则随机变量 $Z = X + Y$ 的分布列为

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

对于常见的分布, 我们有如下系列定理:

定理 5.11 若随机变量 $X \sim B(n_1, p)$ 和 $Y \sim B(n_2, p)$ 独立, 则

$$Z = X + Y \sim B(n_1 + n_2, p).$$

根据此定理可推出: 若 $X_i \sim Ber(p) = B(1, p)$, 那么 $\sum_{i=1}^{n} X_i \sim B(n, p)$.

证明 由卷积公式可得

$$\begin{aligned} P[Z = k] &= \sum_{i=0}^{k} P[X = i]P[Y = k - i] \\ &= \sum_{i=0}^{k} \binom{n_1}{i} p^i (1-p)^{n_1 - i} \binom{n_2}{k-i} p^{k-i} (1-p)^{n_2 - (k-i)} \\ &= p^k (1-p)^{n_1 + n_2 - k} \sum_{i=0}^{k} \binom{n_1}{i} \binom{n_2}{k-i} \\ &= \binom{n_1 + n_2}{k} p^k (1-p)^{n_1 + n_2 - k}. \end{aligned}$$

定理 5.12 若随机变量 $X \sim P(\lambda_1)$ 和 $Y \sim P(\lambda_2)$ 相互独立, 则

$$Z = X + Y \sim P(\lambda_1 + \lambda_2).$$

证明 由泊松分布的定义可知: 当 $i \geqslant 0$ 和 $j \geqslant 0$ 时有

$$P(X = i) = \frac{\lambda_1^i}{i!} e^{-\lambda_1} \quad \text{和} \quad P(Y = j) = \frac{\lambda_2^j}{j!} e^{-\lambda_2}.$$

根据卷积公式有

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

定理 5.13 若随机变量 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 和 $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ 相互独立, 则

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

证明 首先证明随机变量 $Z = X' + Y' = X - \mu_1 + Y - \mu_2 \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$, 其中 $X' \sim \mathcal{N}(0, \sigma_1^2)$ 和 $Y' \sim \mathcal{N}(0, \sigma_2^2)$. 根据卷积公式有

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\sigma_1^2} - \frac{(z-x)^2}{2\sigma_2^2}} dx \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(x - \frac{\sigma_1^2 z}{\sigma_1^2+\sigma_2^2}\right)^2 - \frac{z^2}{2(\sigma_1^2+\sigma_2^2)}} dx \\ &= \frac{e^{-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}}}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sqrt{2\pi}\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(x - \frac{\sigma_1^2 z}{\sigma_1^2+\sigma_2^2}\right)^2} dx \\ &= \frac{e^{-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}}}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}. \end{aligned}$$

由此可得 $Z \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$, 进一步证明 $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

例 5.12 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.

例 5.13 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 由卷积公式可得

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx.$$

由 $X \sim U(0,1)$ 和 $Y \sim U(0,1)$ 可知: 当 $x \in [0,1]$ 时有 $f_X(x) = 1$; 当 $z - x \in [0,1]$ 时有 $f_Y(z-x) = 1$, 即积分区域为 $\{x \in [0,1], z - x \in [0,1]\}$. 由此可得

- 当 $z \leqslant 0$ 或 $z \geqslant 2$ 时, 有 $f_Z(z) = 0$;
- 当 $z \in (0,1)$ 时, 有 $f_Z(z) = \int_0^z 1 dz = z$;
- 当 $x \in [1,2)$ 时, 有 $f_Z(z) = \int_{z-1}^1 dx = 2 - z$.

例 5.14 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.

解 由卷积公式有

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx.$$

由于 $X \sim U(0,1)$ 和 $Y \sim e(1)$ 可知 $f_X(x) f_Y(z-x) \neq 0$ 的区域为 $\{x \in [0,1], z \geqslant x\}$. 因此有

- 当 $z \leqslant 0$ 时有 $f_Z(z) = 0$;
- 当 $0 \leqslant z \leqslant 1$ 时有 $f_Z(z) = \int_0^z e^{-(z-x)}dx = e^{-z}(e^z - 1) = 1 - e^z$;
- 当 $z \geqslant 1$时有 $f_Z(z) = \int_0^1 e^{-(z-x)}dx = e^{-z}(e^1 - 1) = (e-1)e^{-z}$.

5.4.3 随机变量的乘/除法分布

定理 5.14 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

$$\begin{aligned} F_{Y/X}(z) &= P(Y/X \leqslant z) = \int\int_{y/x \leqslant z} f(x, y)dxdy \\ &= \int\int_{x<0, y \geqslant zx} f(x, y)dxdy + \int\int_{x>0, y \leqslant zx} f(x, y)dxdy \\ &= \int_{-\infty}^{0} dx \int_{zx}^{+\infty} f(x, y)dy + \int_{0}^{+\infty} dx \int_{-\infty}^{xz} f(x, y)dy. \end{aligned}$$

变量替换 $t = y/x$ 有

$$\begin{aligned} F_{Y/X}(z) &= \int_{-\infty}^{0} dx \int_{z}^{-\infty} xf(x, tx)dt + \int_{0}^{+\infty} dx \int_{-\infty}^{z} xf(x, tx)dt \\ &= \int_{-\infty}^{0} \int_{-\infty}^{z} (-x)f(x, tx)dtdx + \int_{0}^{+\infty} \int_{-\infty}^{z} xf(x, tx)dtdx \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z} |x| f(x, tx)dtdx = \int_{-\infty}^{z} dt \int_{-\infty}^{+\infty} |x| f(x, tx)dx \end{aligned}$$

求导可得概率密度函数.

5.4.4 随机变量的联合分布函数

已知随机变量 (X, Y) 的联合概率密度为 $f(x, y)$, 设 (X, Y) 的函数

$$U = u(X, Y) \quad V = v(X, Y)$$

如何求 (U, V) 的联合分布, 有如下结论:

定理 5.15 若 $U=u(X,Y)$ 和 $V=v(X,Y)$ 有连续偏导, 且存在反函数

$$x=x(u,v)\quad y=y(u,v),$$

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, 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.

5.5 多维随机变量的数学特征

5.5.1 多维随机变量的期望

定理 5.16 设二维离散型随机变量 (X,Y) 的分布列为 $p_{ij}=P(X=x_i,Y=y_j)$, 则随机变量 $Z=g(X,Y)$ 的期望为

$$E[Z]=E[g(X,Y)]=\sum_{i,j}g(x_i,y_j)p_{ij};$$

设二维连续随机变量 (X,Y) 的概率密度为 $f(x,y)$, 则随机变量 $Z=g(X,Y)$ 的期望为

$$E[Z]=E[g(X,Y)]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,y)f(x,y)dxdy.$$

例 5.15 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

$$f(x,y)=f_X(x)f_Y(y)=\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}.$$

于是得到

$$\begin{aligned}E[\max(X,Y)] &= \int\int_{D_1}xf(x,y)dxdy+\int\int_{D_2}yf(x,y)dxdy\\ &= \int_{-\infty}^{+\infty}dy\int_{y}^{+\infty}xf(x,y)dx+\int_{-\infty}^{+\infty}dx\int_{x}^{+\infty}yf(x,y)dy\\ &= 2\int_{-\infty}^{+\infty}dy\int_{y}^{+\infty}xf(x,y)dx=\frac{1}{\pi}\int_{-\infty}^{+\infty}dy\int_{y}^{+\infty}xe^{-\frac{x^2+y^2}{2}}dx\\ &= \frac{1}{\pi}\int_{-\infty}^{+\infty}e^{-y^2}dy=\frac{1}{\sqrt{\pi}}\end{aligned}$$

最后一个等式成立是因为$\int_{-\infty}^{+\infty} e^{-y^2}dy = \sqrt{\pi}$.

例 5.16 在长度为 1 米的线段上任取两点 X, Y, 求 $E[\min(X,Y)], E[|X-Y|]$.

定理 5.17 对任意随机变量 X, Y 和常数 a, b, 有

$$E[aX+bY] = aE[X] + bE[Y];$$

对独立随机变量 X 和 Y, 以及任意函数 h, g, 有

$$E[XY] = E[X]E[Y] \quad 和 \quad E[h(X)g(Y)] = E[h(X)]E[g(Y)];$$

对任意随机变量 X 和 Y, 有 Cauchy-Schwartz 不等式

$$E[XY] \leqslant \sqrt{E[X^2]E[Y^2]}.$$

证明 设随机变量 X, Y 的联合概率密度为 $f(x,y)$, 则

$$\begin{aligned} E[aX+bY] &= \int\int (ax+by)f(x,y)dxdy \\ &= a\int\int xf(x,y)dxdy + b\int\int yf(x,y)dxdy = aE(X) + bE(Y). \end{aligned}$$

若随机变量 X 与 Y 独立, 则有

$$\begin{aligned} E[XY] &= \int\int xyf(x,y)dxdy = \int\int xf_X(x)yf_Y(y)dxdy \\ &= \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy = E(X)E(Y). \end{aligned}$$

对任意随机变量 X 与Y, 以及对任意 $t \in \mathbb{R}$ 有 $E[(X+tY)^2] \geqslant 0$ 成立, 即任意 $t \in \mathbb{R}$,

$$t^2E[Y^2] + E[X^2] + 2tE[XY] \geqslant 0.$$

因此有 $\Delta = 4[E(XY)]^2 - 4E[X^2]E[Y^2] \leqslant 0$, 即 $E(XY) \leqslant \sqrt{E(X^2)E(Y^2)}$.

5.5.2 协方差

定理 5.18 对任意随机变量 X 与 Y 有

$$Var(X+Y) = Var(X) + Var(Y) + 2E[(X-E(X))(Y-E(Y))].$$

特别地, 当 X 与 Y 独立时有

$$Var(X+Y) = Var(X) + Var(Y).$$

证明 令 $Z = X + Y$, 有

$$\begin{aligned} Var(Z) &= E[(Z - EZ)^2] = E[(X - EX + Y - EY)^2] \\ &= E(X - EX)^2 + E(Y - EY)^2 + 2E[(X - EX)(Y - EY)] \\ &= Var(X) + Var(Y) + 2E[(X - EX)(Y - EY)]. \end{aligned}$$

若 X 与 Y 独立, 则 $2E[(X - EX)(Y - EY)] = 0$, 所以 $Var(X + Y) = Var(X) + Var(Y)$.

定义 5.12 定义随机变量 X 和 Y 的协方差为

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

根据协方差定义和定理 5.18 有

$$Cov(X, X) = Var(X) \quad 和 \quad Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y).$$

下面研究协方差的性质.

性质 5.1 对任意随机变量 X, Y 和常数 c, 有

$$Cov(X, Y) = Cov(Y, X) \quad 和 \quad Cov(X, c) = 0.$$

性质 5.2 对任意常数 a 和 b, 随机变量 X 和 Y, 有

$$Cov(aX, bY) = abCov(X, Y) \quad 和 \quad Cov(X + a, Y + b) = Cov(X, Y).$$

证明 根据协方差的定义有

$$\begin{aligned} Cov(aX, bY) &= E[(aX - E(aX))(bY - E(bY))] = abE[(X - E(X))(Y - E(Y))]; \\ Cov(X + a, Y + b) &= E[(X + a - E(X + a))(Y + b - E(Y + b))] = E[(X - E(X))(Y - E(Y))]. \end{aligned}$$

性质 5.3 对任意随机变量 X_1, X_2, Y, 有

$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y).$$

证明 我们有

$$\begin{aligned} &Cov(X_1 + X_2, Y) = E[(X_1 + X_2 - E(X_1) - E(X_2))(Y - E(Y))] \\ &= E[(X_1 - E(X_1))(Y - E(Y))] + E[(X_2 - E(X_2))(Y - EY)] = Cov(X_1, Y) + Cov(X_2, Y). \end{aligned}$$

由此性质可进一步得到: 对随机变量$X_1, X_2, \ldots, X_n$和$Y_1, Y_2, \ldots, Y_m$, 有

$$Cov\left(\sum_i^n X_i, \sum_j^m Y_j\right) = \sum_i^n \sum_j^m Cov(X_i, Y_j),$$

以及进一步有

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

性质 5.4 若随机变量 X 与 Y 独立, 则有 $Cov(X, Y) = 0$; 但反之不成立.

证明 若 X 与 Y 独立, 则

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E[X - E(X)]E[Y - E(Y)] = 0.$$

反之不成立, 例如随机变量 X 的分布列为

X	-1	0	1
P_i	$1/3$	$1/3$	$1/3$

当 $X \neq 0$ 时随机变量 $Y = 0$, 否则$Y = 1$, 根据联合分布列可知则 X与Y 不独立, 但此时有

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E[XY] - E(X)E(Y) = 0.$$

性质 5.5 对任意随机变量 X 与 Y 有

$$(Cov(X, Y))^2 \leqslant Var(X)Var(Y)$$

等号成立的充要条件是 $Y = aX + b$ (即 X 与 Y 之间存在线性关系).

证明 由 Cauchy-Schwartz 不等式有

$$\begin{aligned} Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &\leqslant \sqrt{E[(X - E(X))^2]E[(Y - E(Y))^2]} = \sqrt{Var(X)Var(Y)}. \end{aligned}$$

下面证明等号成立的充要条件. 若 $Y = aX + b$, 则

$$Cov(X, Y) = Cov(X, aX + b) = aVar(X), \quad Var(Y) = a^2 Var(X),$$

所以

$$Cov^2(X,Y) = a^2Var^2(X) = Var(X)a^2Var(X) = Var(X)Var(Y).$$

另一方面, 若 $(Cov(X,Y))^2 = Var(X)Var(Y)$ 则有

$$(E[(X-EX)(Y-EY)])^2 = E(X-EX)^2E(Y-EY)^2,$$

设

$$\begin{aligned} f(t) &= E[t(X-EX)-(Y-EY)]^2 \\ &= t^2E[X-E(X)]^2 - 2tE[(X-E(X))(Y-E(Y))] + E[Y-E(Y)]^2 \end{aligned}$$

根据一元二次方程的性质 $\Delta = 4(E[(X-EX)(Y-EY)])^2 - 4E(X-EX)^2E(Y-EY)^2 = 0$ 可得方程 $f(t)=0$恰有一重根 t_0. 由此得到

$$f(t_0) = 0 \equiv E[(t_0(X-EX)-(Y-EY))^2]$$

根据 $(t_0(X-EX)-(Y-EY))^2 \geqslant 0$ 可得 $Y = t_0(X-E(X)) + E(Y) = aX+b$.

例 5.17 随机变量 X 与 Y 独立, 且 $Var(X)=6$ 和 $Var(Y)=3$, 求 $Var(2X \pm Y)$.

例 5.18 随机变量 $X \sim P(2)$ 和 $Y \sim \mathcal{N}(-2,4)$, 且 X 与 Y 独立, 则$E[(X-Y)^2]$.

根据性质 5.5可知

$$\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \leqslant 1.$$

等号成立的充要条件是 X 与 Y 存在线性相关. 上式一定程度上反应了随机变量 X 和 Y 的线性相关程度, 由此引入一个新概念: 相关系数.

定义 5.13 设 X 和 Y 为二维随机变量, 如果 $Var(X), Var(Y)$ 存在且不为 0, 则称

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

为 X 与 Y 的相关系数, 简记 ρ.

关于相关系数, 我们需要注意:

- 使用相关系数而不是 $Cov(X,Y)$, 主要是规范 $|\rho_{XY}| \leqslant 1$, 而 $Cov(X,Y)$ 受数值大小影响;
- 相关系数 $|\rho_{XY}| \leqslant 1$: 若 $\rho > 0$, X 与 Y 正相关; 若 $\rho < 0$, X 与 Y 负相关; $|\rho_{XY}| = 1$ 的充要条件为 X 与 Y 有线性关系 $Y = aX+b$. 本质上 ρ_{XY} 刻画了 X,Y 的线性相关程度, 又称为“线性相关系数”;

- 相关系数 $\rho = 0$ 称 X 与 Y 不相关(线性不相关). 独立 $\Rightarrow$ 不相关, 不相关 $\not\Leftarrow$ 独立;
- 随机变量 X 与 Y 不相关, 仅表示 X 与 Y 之间无线性关系, 还可能存在其他关系. 例如: $X \sim U[-\frac{1}{2}, \frac{1}{2}]$, $Y = cos(X)$. 易有$E(X) = 0$,

$$\begin{aligned} Cov(X,Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[X \cdot \cos(X) - XE(\cos(X))] = E[X \cdot \cos(X)] = \int_{-\frac{1}{2}}^{\frac{1}{2}} x \cdot \cos(x)dx = 0. \end{aligned}$$

定理 5.19 对方差不为零的随机变量 X 和 Y, 下述条件相互等价:

- $\rho_{XY} = 0$
- $Cov(X,Y) = 0$
- $E(XY) = E(X)E(Y)$
- $Var(X \pm Y) = Var(X) + Var(Y)$

定理 5.20 对二维正态分布

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right),$$

有 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, 以及 $Cov(X,Y) = \rho\sigma_1\sigma_2$, 即参数 ρ 为 X 与 Y 的相关系数; X 与 Y 独立 $\Longleftrightarrow$ X 与 X 不相关 (此结论仅限于正太分布).

例 5.19 随机变量 (X,Y) 联合概率密度为

$$f(x,y) = \begin{cases} (x+y)/8 & 0 \leqslant x \leqslant 2, 0 \leqslant y \leqslant 2 \\ 0 & \text{其它} \end{cases}$$

求 $Cov(X,Y), Var(X+Y)$.

解 根据协方差的定义有 $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$, 需要计算

$$\begin{aligned} E[X] &= \int_0^2 \int_0^2 x(x+y)/8dxdy = 7/6, \\ E[Y] &= \int_0^2 \int_0^2 y(x+y)/8dxdy = 7/6, \\ E[XY] &= \int_0^2 \int_0^2 xy(x+y)/8dxdy = 4/3, \end{aligned}$$

由此可得 $Cov(X,Y)=4/3-(7/6)^2=-1/36$. 进一步计算

$$\begin{aligned} E[X^2] &= \int_0^2\int_0^2 x^2(x+y)/8dxdy=5/3, \\ E[Y^2] &= \int_0^2\int_0^2 y^2(x+y)/8dxdy=5/3, \end{aligned}$$

由此可得 $Var(X)=Var(Y)=5/3-(7/6)^2=11/36$, 由此可得

$$\begin{aligned} Var(X+Y) &= Var(X)+Var(Y)+2Cov(X,Y)=11/18-1/18=5/9, \\ \rho_{XY} &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}=-1/11. \end{aligned}$$

例 5.20 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

$$\begin{aligned} Cov(Z_1,Z_2) &= Cov(\alpha X+\beta Y,\alpha X-\beta Y)=(\alpha^2-\beta^2)\sigma^2 \\ Var(Z_1) &= Cov(\alpha X+\beta Y,\alpha X+\beta Y)=(\alpha^2+\beta^2)\sigma^2 \\ Var(Z_2) &= Cov(\alpha X-\beta Y,\alpha X-\beta Y)=(\alpha^2+\beta^2)\sigma^2 \end{aligned}$$

由此可知 $\rho_{XY}=(\alpha^2-\beta^2)/(\alpha^2+\beta^2)$.

例 5.21 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 $Var(X+Y)$.

5.5.3 随机向量的数学期望与协方差阵

定义 5.14 设 $X=(X_1,X_2,\cdots,X_n)^\top$, 则随机向量的期望

$$E(X)=(E(X_1),E(X_2),\cdots,E(X_n))^\top,$$

称随机变量 X 的协方差矩阵为

$$Cov(X)=\Sigma=\begin{pmatrix} Cov(X_1,X_1) & \cdots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & \cdots & Cov(X_2,X_n) \\ \vdots & & \vdots \\ Cov(X_n,X_1) & \cdots & Cov(X_n,X_n) \end{pmatrix}.$$

定理 5.21 随机变量 X 的协方差矩阵是对称半正定的矩阵.

证明 证明根据函数的

$$\begin{aligned} f(t_1,t_2,\ldots,t_n) &= (t_1,t_2,\ldots,t_n)\begin{pmatrix} Cov(X_1,X_1) & \cdots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & \cdots & Cov(X_2,X_n) \\ \vdots & & \vdots \\ Cov(X_n,X_1) & \cdots & Cov(X_n,X_n) \end{pmatrix}(t_1,t_2,\ldots,t_n)^\top \\ &= E[(t_1(X_1-E[X_1])+t_2(X_2-E[X_2])+\cdots+t_n(X_n-E[X_n]))^2] \geqslant 0. \end{aligned}$$

定理 5.22 设多维正态分布 $X=(X_1,X_2,\ldots,X_n)^\top \sim N(\mu,\Sigma)$, 则有

$$\mu=(E[X_1],E[X_2],\ldots,E[X_n])^\top \quad 和 \quad \Sigma=[Cov(X_i,X_j)]_{n\times n}.$$

对多维随机变量 $X=(X_1,X_2,\cdots,X_n)\ \sim\mathcal{N}(\mu,\Sigma)$, 有

- 每个变量X_i的边缘分布是正态分布;
- $X_1,X_2,\cdots,X_n$ 相互独立 $\Longleftrightarrow$ $X_1,X_2,\cdots,X_n$ 相互不相关;
- $X\sim\mathcal{N}(\mu,\Sigma)\Longleftrightarrow\sum_{i=1}^n a_iX_i$ 是正态分布(对任意非全为0常数$a_1,a_2,\ldots,a_n$).

5.6 条件分布与条件期望

前面学过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

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

相关概念可推广到随机变量: 给定随机变量 Y 取值条件下求随机变量 X 的概率分布, 即条件分布.

首先考虑离散型随机变量:

定义 5.15 设二维离散型随机变量 (X,Y) 的分布列为 $\{p_{ij}\}$, 若 Y 的边缘分布 $P(Y=y_j)=p_{\cdot j}>0$, 称

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

在 $Y=y_j$ 条件下随机变量 X 的条件分布列.

类似可定义在 $X=x_i$ 条件下随机变量 Y 的条件分布列. 条件分布是一种概率分布, 具有分布的性质. 例如, 非负性 $P(X=x_i|Y=y_j)\geqslant 0$, 规范性 $\sum_{i=1}^\infty P(X=x_i|Y=y_j)=1$ 等性质.

例 5.22 一个选手击中目标的概率为 p, 射击进行到击中两次目标为止, 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:

$$P\{X=m,Y=n\}=f(x,y)=\begin{cases}p^2(1-p)^{n-2} & 1\leqslant m<n<\infty\\ 0 & \text{其它}\end{cases}$$

由此可得 X 的边缘分布列为

$$P\{X=m\}=\sum_{n=m+1}^{\infty}P\{X=m,Y=n\}=\sum_{n=m+1}^{\infty}p^2(1-p)^{n-2}=p^2\frac{(1-p)^{m-1}}{1-(1-p)}=p(1-p)^{m-1}.$$

同理得到随机变量 Y 的边缘分布列为

$$P\{Y=n\}=\sum_{m=1}^{n-1}P\{X=m,Y=n\}=\sum_{m=1}^{n-1}p^2(1-p)^{n-2}=(n-1)p^2(1-p)^{n-2}(n=2,3,\cdots).$$

因此, 当 $n=2,3,\cdots$ 时, 随机变量 X 在 $Y=n$ 条件下的分布列为:

$$P\{X=m|Y=n\}=\frac{P\{X=m,Y=n\}}{P\{Y=n\}}=\frac{p^2(1-p)^{n-2}}{(n-1)p^2(1-p)^{n-2}}=\frac{1}{n-1}\quad m=1,2,\cdots,n-1.$$

当 $m=1,2,3,\cdots$ 时, 随机变量 Y 在 $X=m$ 条件下的分布列为:

$$P\{Y=n|X=m\}=\frac{P\{X=m,Y=n\}}{P\{X=m\}}=\frac{p^2(1-p)^{n-2}}{p(1-p)^{m-1}}=p(1-p)^{n-m-1}\quad n=m+1,m+2,\cdots$$

对于连续型随机变量 (X,Y), 对任意 $x,y\in(-\infty,+\infty)$, 有 $P(X=x)=0$ 和 $P(Y=y)=0$ 成立, 因此不能利用离散随机变量的条件概率推导连续随机变量的条件分布. 下面给出条件概率的定义:

定义 5.16 设连续随机变量 (X,Y) 的联合概率密度为 $f(x,y)$, 以及 Y 的边缘概率密度为 $f_Y(y)$, 对任意 $f_Y(y)>0$, 称

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$$

在 $Y=y$ 条件下随机变量 X 的条件概率密度; 称

$$F_{X|Y}(x|y)=P\{X\leqslant x|Y=y\}=\int_{-\infty}^{x}f_{X|Y}(u|y)du$$

为 $Y=y$ 条件下 X 的条件分布函数.

类似可定义在 $X=x$ 条件下随机变量 Y 的条件概率密度和分布函数分别为

$$f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)} \qquad F_{Y|X}(y|x)=\int_{-\infty}^{y}\frac{f(x,v)}{f_X(x)}dv.$$

下面给出条件概率的一种解释, 这里以 $f_{X|Y}=\frac{f(x,y)}{f_Y(y)}$ 为例, 首先分布函数有

$$F_{X|Y}(x|y)=\lim_{\epsilon\to 0^+}P\{X\leqslant x|y\leqslant Y\leqslant y+\epsilon\}=\lim_{\epsilon\to 0^+}\frac{P\{X\leqslant x,y\leqslant Y\leqslant y+\epsilon\}}{P\{y\leqslant Y\leqslant y+\epsilon\}}.$$

根据积分中值定理有

$$\frac{P\{X\leqslant x,y\leqslant Y\leqslant y+\epsilon\}}{P\{y\leqslant Y\leqslant y+\epsilon\}}=\frac{\int_{-\infty}^{x}\int_{y}^{y+\epsilon}f(u,v)dudv}{\int_{y}^{y+\epsilon}f_Y(u)dv}=\frac{\epsilon\int_{-\infty}^{x}f(u,y+\theta_1\epsilon)du}{\epsilon f_Y(y+\theta_2\epsilon)}$$

其中$\theta_1,\theta_2\in(0,1)$. 当$\epsilon\to 0^+$时, 有

$$F_{X|Y}(x|y)=\lim_{\epsilon\to 0^+}P\{X\leqslant x|y\leqslant Y\leqslant y+\epsilon\}=\frac{\int_{-\infty}^{x}f(u,y)du}{f_Y(y)}=\int_{-\infty}^{x}\frac{f(u,y)}{f_Y(y)}du,$$

由此可得条件概率密度 $f_{X|Y}(x|y)=f(x,y)/f_Y(y)$. 下面给出条件概率的性质:

引理 5.3 (乘法公式) 对于随机变量 X 和 Y, 有

$$\begin{aligned}f(x,y) &= f_X(x)f_{Y|X}(y|x), \quad (f_X(x)>0),\\ f(x,y) &= f_Y(y)f_{X|Y}(x|y), \quad (f_Y(y)>0).\end{aligned}$$

若随机变量 X 和 Y 相互独立, 则有联合概率密度 $f(x,y)=f_X(x)f_Y(y)$, 由此可得

引理 5.4 如果随机变量 X 和 Y 相互独立, 则有

$$f_{Y|X}(y|x)=f_Y(y) \quad \text{和} \quad f_{X|Y}(x|y)=f_X(x).$$

由此可根据条件概率来可判别随机变量 (X,Y) 的独立性. 下面看几个条件概率的例子:

例 5.23 设二维随机变量 (X,Y) 的概率密度

$$f(x,y)=\begin{cases}\frac{e^{-x/y}e^{-y}}{y} & 0<x<+\infty,0<y<+\infty\\ 0 & \text{其它,}\end{cases}$$

求 $P(X>1|Y=y)$.

解 首先求解随机变量 Y 的边缘分布为

$$f_Y(y)=\int_0^{+\infty}\frac{e^{-x/y}e^{-y}}{y}dx=e^{-y}[-e^{-\frac{x}{y}}]_0^{+\infty}=e^{-y}\qquad(y>0).$$

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

$$P(X>1|Y=y)=\int_1^{\infty}\frac{e^{-x/y}}{y}dx=-e^{-x/y}|_1^{\infty}=e^{-\frac{1}{y}}.$$

例 5.24 已知随机变量 $X\sim U(0,1)$, 当观察到 $X=x$ 的条件下, 随机变量 $Y\sim U(x,1)$. 求 Y 的概率密度.

解 根据题意可知 $X\sim U(0,1)$, 在随机变量 $X=x$ 的条件下 $Y\sim U(x,1)$, 即 $f_{Y|X}(y|x)=1/(1-x)$. 根据条件概率乘积公式有

$$f(x,y)=f_X(x)f_{Y|X}(y|x)=\begin{cases}\frac{1}{1-x} & 0<x<y<1,\\ 0 & \text{其它}.\end{cases}$$

根据联合分布求解随机变量 Y 的边缘分布

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx=\begin{cases}\int_0^y\frac{1}{1-x}dx=-\ln(1-y) & y>0,\\ 0 & \text{其它}.\end{cases}$$

例 5.25 设随机变量 (X,Y) 的概率密度为

$$f(x,y)=\begin{cases}e^{-y} & y>x>0\\ 0 & \text{其它}\end{cases}$$

求条件概率密度 $f_{X|Y}(x|y)$.

定理 5.23 多维正太分布的条件分布是正太分布.

证明 为简单起见仅给出二维正太分布的详细证明. 设随机变量 $(X,Y)\sim\mathcal{N}(\mu,\Sigma)$, 其中

$$\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\quad\text{和}\quad\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}.$$

下面证明在 $Y=y$ 的条件下随机变量 X 服从正态分布. 首先给出二维正太分布的联合分布

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2}\frac{1}{(1-\rho)^2}[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}]}.$$

以及随机变量 Y 的边缘分布 $Y\sim\mathcal{N}(\mu_2,\sigma_2^2)$. 于是得到条件概率

$$\begin{aligned}f_{X|Y}(x|y) &= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\rho^2\frac{(y-\mu_2)^2}{\sigma_2^2}]}\\ &= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}[\frac{x-\mu_1}{\sigma_1}+\frac{\rho(y-\mu_2)}{\sigma_2}]^2}\\ &= \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}[x-\mu_1+\sigma_1^2\rho^2(y-\mu_2)/\sigma_2^2]^2}\end{aligned}$$

由此可知在 $Y=y$ 的条件下 $X\sim\mathcal{N}(\mu_1-\sigma_1^2\rho(y-\mu_2)/\sigma_2^2,\sigma_1^2(1-\rho^2))$.

5.6.1 条件期望

定义 5.17 对二维离散随机变量 (X,Y), 在 $Y=y$ 条件下随机变量 X 的期望为

$$E[X|Y=y]=\sum_{i=1}^{\infty}x_iP(X=x_i|Y=y);$$

对二维连续随机变量 (X,Y), 在 $Y=y$ 条件下随机变量 X 的期望为

$$E[X|Y=y]=\int_{-\infty}^{+\infty}xf(x|y)dx.$$

注意 $E[X|Y=y]$ 是 y 的函数, 对条件期望有如下重要性质:

定理 5.24 对离散随机变量 $X_1,X_2,\cdots,X_n$ 及常数 $c_1,c_2,\cdots,c_n$ 有

$$E\left[\sum_{i=1}^{n}c_iX_i|Y=y\right]=\sum_{i=1}^{n}c_iE[X_i|Y=y].$$

定理 5.25 (全期望公式, law of total expectation) 对随机变量 X 和事件 A 有

$$E[X]=E[X|A]P(A)+E[X|\bar{A}](1-P(A))$$

其中事件 $\bar{A}$ 为事件 A 的补.

全期望公式对应于全概率公式的期望版本, 在很多应用中有重要的性质。

证明 此定理对离散和连续随机变量都成立, 为证明简单起见, 这里给出离散情况下的详细证明. 根据概率的性质有

$$\begin{aligned} E[X] &= \sum_i x_i P(X=x_i) = \sum_i x_i [P(X=x_i, A) + P(X=x_i, \bar{A})] \\ &= \sum_i x_i P(X=x_i|A)P(A) + \sum_i x_i P(X=x_i|\bar{A})P(\bar{A}) \\ &= P(A)\sum_i x_i P(X=x_i|A) + P(\bar{A})\sum_i x_i P(X=x_i|\bar{A}) \\ &= P(A)E[X|A] + P(\bar{A})E[X|\bar{A}]. \end{aligned}$$

该定理有一个关于随机变量的定理:

定理 5.26 对二维随机变量 (X,Y) 有

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

特别地, 对二维离散随机变量有

$$E[X] = E[E(X|Y)] = \sum_{Y=y_j} P[Y=y_j]E[X|Y=y_j].$$

证明 利用全概率公式有

$$\begin{aligned} E[X] &= \sum_i x_i P(X=x_i) = \sum_i \sum_j x_i P(X=x_i, Y=y_j) \\ &= \sum_i \sum_j x_i P(X=x_i, Y=y_j) \\ &= \sum_j P(Y=y_j) \sum_i x_i \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} \\ &= \sum_j P(Y=y_j) \sum_i x_i P(X=x_i|Y=y_j) \\ &= \sum_j P(Y=y_j) E[X|Y=y_j] = E_Y[E[X|Y]]. \end{aligned}$$

待加入连续随机变量的证明.

习题

5.1 已知二维随机变量 (X,Y) 的分布函数为 $F(x,y)$, 求概率 $P(X>x,Y>y)$.

第 6 章　集中不等式 (Concentration)

给定一个训练数据集

$$S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\},$$

其中 $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ 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 $(\boldsymbol{x}_i, y_i)$ 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 $f\colon \mathcal{X} \to \{0, 1\}$, 定义函数 f 在训练数据集 S_n 上的分类错误率为

$$\hat{R}(f, S_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i),$$

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'

$$R(f, \mathcal{D}) = E_{(\boldsymbol{x}, y) \sim \mathcal{D}}(\mathbb{I}(f(\boldsymbol{x}) \neq y)) = \Pr_{(\boldsymbol{x}, y) \sim \mathcal{D}}[f(\boldsymbol{x}) \neq y].$$

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

$$\Pr_{S_n \sim \mathcal{D}^n} \left[|\hat{R}(f, D_n) - R(f)| \geqslant t \right] \text{ 是否足够小?}$$

即能否以很大的概率保证

$$|\hat{R}(f, D_n) - R(f)| < t.$$

从而在理论上保证 $\hat{R}(f, D_n)$ 是 $R(f)$ 的一个有效估计. 上述性质在机器学习中被称为 ' 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 6.1　假设训练数据集 $S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ 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 ϵ 之间的概率 ($\epsilon > 0$).

解　设随机变量

$$X_i = \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i] \qquad (i \in [n]),$$

根据数据集的独立同分布假设可知 $X_1, X_2, \ldots, X_n$ 是独立同分布的随机变量. 令 $p = E[X_i]$, 则有 $X_i \sim \mathrm{Ber}(p)$. 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 ϵ 的概率为

$$\begin{aligned}
&\Pr\left[\sum_{i=1}^{n} X_i = 0, p > \epsilon\right] \leqslant \Pr\left[\sum_{i=1}^{n} X_i = 0 | p > \epsilon\right] \\
&= \Pr\left[X_1 = 0, X_2 = 0, \ldots, X_n = 0 | p > \epsilon\right] \qquad \text{(根据独立性假设)} \\
&= \prod_{i=1}^{n} \Pr\left[X_i = 0 | p > \epsilon\right] \leqslant (1-\epsilon)^n \leqslant \exp(-n\epsilon).
\end{aligned}$$

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

$$X_i = \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i),$$

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 $X_1, X_2, \ldots, X_n$, 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.

6.1 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 6.1 对任意随机变量 $X \geqslant 0$ 和 $\epsilon > 0$, 有

$$P(X \geqslant \epsilon) \leqslant \frac{E(X)}{\epsilon}.$$

证明 利用全期望公式考虑随机事件 $X \geqslant \epsilon$ 有

$$E[X] = E[X|X \geqslant \epsilon]P(X \geqslant \epsilon) + E[X|X \leqslant \epsilon]P(X \leqslant \epsilon) \geqslant P(X \geqslant \epsilon)\epsilon$$

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 6.1 对任意随机变量 X 和 $\epsilon \geqslant 0$, 以及单调递增的非负函数 $g(x)$, 有

$$P(X \geqslant \epsilon) \leqslant \frac{E[g(X)]}{g(\epsilon)}.$$

利用 Markov 不等式可以推导 Chebyshev 不等式:

定理 6.2 (Chebyshev 不等式) 设随机变量 X 的均值为 μ, 则有

$$P(|X-\mu| > \epsilon) \leqslant \frac{Var(X)}{\epsilon^2}.$$

证明 根据 Markov 不等式有

$$P(|X-\mu| > \epsilon) = P((X-\mu)^2 \geqslant \epsilon^2) \leqslant \frac{E(X-\mu)^2}{\epsilon^2} = \frac{Var(X)}{\epsilon^2}.$$

例 6.2 设随机变量 $X \sim N(-1,2)$ 和 $Y \sim N(1,8)$, 且 X 和 Y 的相关系数为 -1, 利用 Chebyshev 不等式求 $P(|X+Y| \geqslant 6) \leqslant$?

解 根据随机变量 X 和 Y 的相关系数为 -1 可知

$$\mathrm{Cov}(X,Y) = -\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} = -4.$$

由 $E[X+Y]=0$, 利用 Chebyshev 不等式有

$$\begin{aligned} P(|X+Y| \geqslant 6) &= P(|X+Y-E[X+Y]| \geqslant 6) \\ &\leqslant \mathrm{Var}(X+Y)/36 = (\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y))/36 = 1/18. \end{aligned}$$

下面补充一个Chebysheve不等式的应用例子:

例 6.3 设随机变量X和Y满足$E(X)=-2$, $E(Y)=2$, $\mathrm{Var}(X)=1$, $Var(Y)=4$, $\rho_{XY}=-1/2$. 利用Chebyshev不等式估计$\Pr(|X+Y| \geqslant 6)$的上界.

解 根据期望的线性关系有$E[X+Y]=0$, 根据相关系数的定义有

$$\rho_{XY} = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{Var(X)}\sqrt{Var(Y)}} = -\frac{1}{2}.$$

由此可得$Var(X+Y) = Var(X)+Var(Y)+2E_{XY}[X-E(X)][Y-E(Y)] = 3$. 根据Chebysheve不等式有$\Pr\{|X+Y| \geqslant 6\} \leqslant \mathrm{Var}(X+Y)/36 = 1/12$.

比 Chebyshev 不等式更紧地 Cantelli 不等式, 又被成为单边 Chebyshev 不等式.

引理 6.1 随机变量 X 的均值 $\mu > 0$, 方差 σ^2, 则对任意 $\epsilon > 0$ 有

$$P(X - \mu \geqslant \epsilon) \leqslant \frac{\sigma^2}{\sigma^2 + \epsilon^2} \quad \text{和} \quad P(X - \mu \leqslant -\epsilon) \leqslant \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

证明 设随机变量 $Y = X - \mu$, 有 $E(Y) = 0$ 以及 $Var(Y) = \sigma^2$. 对任意 $t > 0$ 有

$$\begin{aligned} P(X - \mu \geqslant \epsilon) &= P(Y \geqslant \epsilon) = P(Y + t \geqslant \epsilon + t) \leqslant P((Y + t)^2 \geqslant (\epsilon + t)^2) \\ &\leqslant \frac{E((Y + t)^2)}{(\epsilon + t)^2} = \frac{\sigma^2 + t^2}{(\epsilon + t)^2} \end{aligned}$$

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

$$P(X - \mu \geqslant \epsilon) \leqslant \min_{t>0} \frac{\sigma^2 + t^2}{(\epsilon + t)^2} = \frac{\sigma^2}{\epsilon^2 + \sigma^2}.$$

另一方面, 对任意 $t > 0$ 有

$$\begin{aligned} P(X - \mu \leqslant -\epsilon) &= P(Y \leqslant -\epsilon) = P(Y - t \leqslant -\epsilon - t) \leqslant P((Y + t)^2 \geqslant (\epsilon + t)^2) \\ &\leqslant \frac{E((Y + t)^2)}{(\epsilon + t)^2} = \frac{\sigma^2 + t^2}{(\epsilon + t)^2} \end{aligned}$$

同理完成证明.

下面介绍 Chebyshev 不等式的推论.

推论 6.2 设独立同分布的随机变量 $X_1, X_2, \ldots, X_n$ 满足 $E(X_i) = \mu$ 和 $Var(X_i) \leqslant \sigma^2$, 对任意 $\epsilon > 0$ 有

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \epsilon\right) \leqslant \frac{\sigma^2}{n\epsilon^2}$$

证明 根据 Chebyshev 不等式有

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \epsilon\right) \leqslant \frac{1}{\epsilon^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right).$$

而独立同分布的假设有

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}Var(X_i) \leqslant \frac{\sigma^2}{n}.$$

由此得到

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \epsilon\right) \leqslant \frac{\sigma^2}{n\epsilon^2},$$

从而完成证明.

例 6.4 设分类器 f 在训练集 $S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ 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 $X_i = \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i]$ $(i \in [n])$, 则这些随机变量是独立同分布的. 训练错误率

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

设分类器 f 在分布 $\mathcal{D}$ 上的错误率为 p, 则 $X_i \sim \mathrm{Ber}(p)$ 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

$$\Pr[|p - \hat{p}| > \epsilon] \leqslant \frac{1}{\epsilon^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{\epsilon^2 n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{p(1-p)}{n\epsilon^2} \leqslant \frac{1}{4n\epsilon^2}$$

取 $\epsilon = \hat{p}/10$ 有

$$\Pr\left[|p - \hat{p}| > \hat{p}/10\right] \leqslant \frac{25}{n\hat{p}^2}.$$

引理 6.2 (Young 不等式) 给定常数 $a > 0, b > 0$, 对满足 $1/p + 1/q = 1$ 的实数 $p > 0, q > 0$ 有

$$ab \leqslant \frac{1}{p}a^p + \frac{1}{q}b^q.$$

证明 根据凸函数性质有

$$\begin{aligned} ab &= \exp(\ln(ab)) = \exp(\ln a + \ln b) = \exp\left(\frac{1}{p}\ln a^p + \frac{1}{q}\ln b^q\right) \\ &\leqslant \frac{1}{p}\exp\left(\ln a^p\right) + \frac{1}{q}\exp\left(\ln b^q\right) = \frac{1}{p}a^p + \frac{1}{q}b^q. \end{aligned}$$

引理得证.

根据 Young 不等式可证明著名的 Hölder 不等式.

引理 6.3 (Hölder 不等式) 对任意随机变量 X 和 Y 以及实数 $p > 0$ 和 $q > 0$ 满足$1/p+1/q = 1$, 有

$$E(|XY|) \leqslant (E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}.$$

特别地, 当 $p = q = 2$ 时 Hölder 不等式变成为 Cauchy-Schwartz 不等式.

证明 设 $c = (E(|X|^p))^{\frac{1}{p}}$ 和 $d = (E(|Y|^q))^{\frac{1}{q}}$, 根据 Young 不等式有

$$\frac{|XY|}{cd} = \frac{|X|}{c}\frac{|Y|}{d} \leqslant \frac{1}{p}\frac{|X|^p}{c^p} + \frac{1}{q}\frac{|Y|^q}{d^q}.$$

对上式两边同时取期望有

$$\frac{E(|XY|)}{cd} \leqslant \frac{1}{p}\frac{E(|X|^p)}{c^p} + \frac{1}{q}\frac{E(|Y|^q)}{d^q} = \frac{1}{p} + \frac{1}{q} = 1,$$

从而完成证明.

6.2 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 6.1 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 6.3 设随机变量 X 的矩生成函数为 $M_X(t)$, 对任意 $n \geqslant 1$ 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 Tayler 公式有

$$e^{tX} = \sum_{i=1}^{\infty} \frac{(tX)^i}{i!}.$$

两边同时取期望有

$$E[e^{tX}] = \sum_{i=1}^{\infty} \frac{t^i}{i!} E[X^i].$$

对上式两边分别对 t 求 n 阶导数并取 $t = 0$ 有 $M_X^{(n)}(t) = E[X^n]$.

定理 6.4 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 6.3 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

$$\Pr[X \geqslant E[X]+\epsilon] = \Pr[e^{tX} \geqslant e^{tE[X]+t\epsilon}] \leqslant e^{-t\epsilon-tE[X]}E[e^{tX}].$$

特别地, 有

$$\Pr[X \geqslant \epsilon] \leqslant \min_{t>0}\left\{e^{-t\epsilon-tE[X]}E\left[e^{tX}\right]\right\}.$$

类似地, 对任意 $\epsilon>0$ 和 $t<0$ 有

$$\Pr[X \leqslant E[X]-\epsilon] = \Pr[e^{tX} \geqslant e^{tE[X]-t\epsilon}] \leqslant e^{t\epsilon-tE[X]}E[e^{tX}].$$

同理有

$$\Pr[X \leqslant \epsilon] \leqslant \min_{t<0}\left\{e^{t\epsilon-tE[X]}E\left[e^{tX}\right]\right\}.$$

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.

6.2.1 二值随机变量的 Chernoff 不等式

定理 6.5 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立且满足 $X_i \sim Ber(p_i)$, 令 $\mu = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$. 对任意 $\epsilon>0$ 有

$$\Pr\left[\sum_{i=1}^n X_i \geqslant (1+\epsilon)\mu\right] \leqslant \left(\frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}}\right)^\mu;$$

对任意 $\epsilon \in (0,1)$ 有

$$\Pr\left[\sum_{i=1}^n X_i \geqslant (1+\epsilon)\mu\right] \leqslant e^{-\mu\epsilon^2/3}.$$

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

$$\Pr[\bar{X} \geqslant (1+\epsilon)\mu] = \Pr[e^{t\bar{X}} \geqslant e^{t(1+\epsilon)\mu}] \leqslant e^{-t(1+\epsilon)\mu}E[e^{t\bar{X}}].$$

利用随机变量的独立性以及 $1+x \leqslant e^x$ 有

$$\begin{aligned} E[e^{t\bar{X}}] &= E[e^{\sum_{i=1}^n tX_i}] = \prod_{i=1}^n E[e^{tX_i}] \\ &= \prod_{i=1}^n [(1-p_i)+p_ie^t] = \prod_{i=1}^n [1+p_i(e^t-1)] \end{aligned}$$

$$\leqslant \quad \exp\left(\sum_{i=1}^{n} p_i(e^t-1)\right) = \exp(\mu(e^t-1)).$$

由此可得

$$\Pr[\bar{X} \geqslant (1+\epsilon)\mu] \leqslant \exp\left(-t(1+\epsilon)\mu + \mu(e^t-1)\right).$$

对上式求最小值解得 $t_{\min} = \ln(1+\epsilon)$, 代入可得

$$\Pr[\bar{X} \geqslant (1+\epsilon)\mu] \leqslant \left(\frac{e^{\epsilon}}{(1+\epsilon)^{(1+\epsilon)}}\right)^{\mu}.$$

对第二个不等式, 只需证明当 $\epsilon \in (0,1)$ 有

$$f(\epsilon) = \ln\left(\frac{e^{\epsilon}}{(1+\epsilon)^{(1+\epsilon)}}\right) + \frac{\epsilon^2}{3} = \epsilon - (1+\epsilon)\ln(1+\epsilon) + \frac{\epsilon^2}{3} \leqslant 0.$$

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 $f^{'}(\epsilon) \leqslant 0$, 即函数 $f(\epsilon)$ 在 $[0,1]$ 上单调递减. 当 $\epsilon \geqslant 0$ 时有 $f(\epsilon) \leqslant f(0) = 0$, 所以 $\exp(f(\epsilon)) \leqslant 1$.

下面的定理给出了 $\Pr[\sum_{i=1}^{n} X_i \leqslant (1-\epsilon)\mu]$ 的估计, 证明作为练习题留给大家完成.

定理 6.6 设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立且满足 $X_i \sim Ber(p_i)$, 令 $\mu = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p_i$. 对任意 $\epsilon \in (0,1)$ 有

$$\Pr\left[\sum_{i=1}^{n} X_i \leqslant (1-\epsilon)\mu\right] \leqslant \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{(1-\epsilon)}}\right)^{\mu} \leqslant \exp(-\mu\epsilon^2/2).$$

定义 6.2 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 6.7 对 n 个独立的 Rademacher 随机变量 $X_1, X_2, \cdots, X_n$, 有

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \epsilon\right) \leqslant \exp(-n\epsilon^2/2) \quad \text{和} \quad \Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leqslant -\epsilon\right) \leqslant \exp(-n\epsilon^2/2).$$

证明 根据 Taylor 展开式有

$$\frac{1}{2}\exp(t)+\frac{1}{2}\exp(-t)=\sum_{i\geqslant 0}\frac{t^{2i}}{(2i)!}\leqslant\sum_{i\geqslant 0}\frac{(t^2/2)^i}{i!}=\exp(t^2/2).$$

若随机变量 $X\in\{+1,-1\}$ 且满足 $\Pr(X=1)=\Pr(X=-1)=1/2$, 则有

$$E[e^{tX}]=\frac{1}{2}e^t+\frac{1}{2}e^{-t}\leqslant\exp(t^2/2).$$

对任意 $t>0$, 根据 Chernoff 方法有

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{i=1}^{n}X_i\geqslant\epsilon\right) &\leqslant \exp(-nt\epsilon)E\left[\exp\left(\sum_{i=1}^{n}tX_i\right)\right]\\ &= \exp(-nt\epsilon)\prod_{i=1}^{n}E\left[\exp(tX_i)\right]\leqslant\exp(-nt\epsilon+nt^2/2).\end{aligned}$$

通过对上式右边求最小值解得 $t=\epsilon$, 带入上式得到

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}X_i\geqslant\epsilon\right)\leqslant\exp(-n\epsilon^2/2).$$

同理证明另一个不等式.

推论 6.4 对独立同分布的随机变量 $X_1,X_2,\cdots,X_n$ 满足 $P(X_1=0)=P(X_1=1)=1/2$, 有

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{2}\geqslant\epsilon\right)\leqslant\exp(-2n\epsilon^2)\quad\text{和}\quad\Pr\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{2}\leqslant-\epsilon\right)\leqslant\exp(-2n\epsilon^2).$$

6.2.2 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 6.4 设随机变量 $X\in[0,1]$ 的期望 $\mu=E[X]$. 对任意 $t>0$ 有

$$E[e^{tX}]\leqslant\exp(t\mu+t^2/8).$$

证明 由凸函数的性质可知

$$e^{tX}=e^{tX+(1-X)0}\leqslant Xe^t+(1-X)e^0,$$

两边再同时取期望有

$$E(e^{tX})\leqslant 1-\mu+\mu e^t=\exp(\ln(1-\mu+\mu e^t)).$$

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

$$f''(t) = \frac{\mu e^t}{1 - \mu + \mu e^t} - \frac{\mu^2 e^{2t}}{(1 - \mu + \mu e^t)^2} \leqslant 1/4.$$

根据泰勒中值定理有

$$f(t) = f(0) + tf'(0) + f''(\xi)t^2/2 \leqslant t\mu + t^2/8.$$

引理得证.

由上面的 Chernoff 引理进一步推导出

推论 6.5 设随机变量 $X \in [a, b]$ 的期望 $\mu = E[x]$. 对任意 $t > 0$ 有

$$E(e^{tX}) \leqslant \exp(\mu t + t^2(b - a)^2/8).$$

根据上述推论, 我们得到有界随机变量的 Chernoff 不等式:

定理 6.8 假设 $X_1, \ldots, X_n$ 是 n 独立的随机变量、且满足 $X_i \in [a, b]$. 对任意 $\epsilon > 0$ 有

$$\begin{aligned}
\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \geqslant \epsilon\right] &\leqslant \exp(-2n\epsilon^2/(b-a)^2), \\
\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \leqslant -\epsilon\right] &\leqslant \exp(-2n\epsilon^2/(b-a)^2).
\end{aligned}$$

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

$$\begin{aligned}
&\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \geqslant \epsilon\right] \\
&= \Pr\left[\sum_{i=1}^{n} t(X_i - E[X_i]) \geqslant nt\epsilon\right] \\
&\leqslant \exp(-nt\epsilon) E\left[\exp\left(\sum_{i=1}^{n} t(X_i - E[X_i])\right)\right] \\
&= \exp(-nt\epsilon) \prod_{i=1}^{n} E\left[\exp(t(X_i - E[X_i]))\right].
\end{aligned}$$

根据 Chernoff 引理, 对任意 $X_i \in [a,b]$ 有

$$E\left[\exp(t(X_i - E[X_i]))\right] \leqslant \exp((b-a)^2 t^2/8).$$

由此得到

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \geqslant \epsilon\right] \leqslant \exp(-nt\epsilon + nt^2(b-a)^2/8).$$

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:

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \geqslant \epsilon\right] \leqslant \exp(-2n\epsilon^2/(b-a)^2).$$

从而完成证明.

6.2.3 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 6.9 设随机变量 $X_1, \ldots, X_n$ 相互独立、且服从 $X_i \sim \mathcal{N}(\mu, \sigma)$, 对任意 $\epsilon > 0$ 有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geqslant \epsilon\right] = \Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \leqslant -\epsilon\right] \leqslant \frac{1}{2}\exp(-n\epsilon^2/2\sigma^2).$$

证明 对随机变量 $X_i \sim \mathcal{N}(\mu, \sigma)$, 根据正太分布的性质有

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \sim \mathcal{N}(0, \sigma^2/n) \quad \Rightarrow \quad \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i - \mu) \sim \mathcal{N}(0,1).$$

若 $X' \sim \mathcal{N}(0,1)$, 对任意 $\epsilon > 0$, 根据以前的定理有

$$P(X' \geqslant \epsilon) \leqslant \frac{1}{2}e^{-\epsilon^2/2}.$$

因此得到

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geqslant \epsilon\right] = \Pr\left[\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i - \mu) \geqslant \epsilon\sqrt{n}/\sigma\right] \leqslant \frac{1}{2}\exp(-n\epsilon^2/2\sigma^2),$$

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 6.3 对任意 $t \in (-\infty, +\infty)$, 若随机变量 X 满足

$$E[e^{(X-E[X])t}] \leqslant \exp(bt^2/2),$$

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 6.5 对任意有界的随机变量 $X \in [a,b]$, 根据 Chernoff 引理有

$$E[e^{(X-\mu)t}] \leqslant \exp(t^2(b-a)^2/8),$$

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 6.6 如果随机变量 X 服从高斯分布 $\mathcal{N}(\mu,\sigma^2)$, 则有

$$E[e^{(X-\mu)t}] = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{xt} e^{-x^2/2\sigma^2} dx = e^{\sigma^2t^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(t\sigma - x/\sigma)^2/2} d(x/\sigma) = e^{\sigma^2t^2/2}.$$

Gaussian 随机变量是参数为 σ^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 6.10 设 $X_1,\ldots,X_n$ 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

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) \geqslant \epsilon\right] \leqslant \exp(-n\epsilon^2/2b) \quad \text{和} \quad \Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) \leqslant -\epsilon\right] \leqslant \exp(-n\epsilon^2/2b).$$

证明 对任意 $t > 0$, 根据 Chernoff 方法有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) \geqslant \epsilon\right] \leqslant e^{-tn\epsilon} \prod_{i=1}^{n} E[e^{(X_i-\mu)t}] \leqslant e^{-tn\epsilon + nbt^2/2},$$

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 6.11 设 $X_1,\ldots,X_n$ 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

$$E\left[\max_{i\in[n]} X_i\right] \leqslant \sqrt{2b\ln n}.$$

证明 对任意 $t > 0$, 根据 Jensen 不等式有

$$\exp\left(tE\left[\max_{i\in[n]} X_i\right]\right) \leqslant E\left[\exp\left(t\max_{i\in[n]} X_i\right)\right]$$

$$= \quad E\left[\max_{i\in[n]}\exp\left(tX_i\right)\right] \leqslant \sum_{i=1}^{n} E[\exp(tX_i)] \leqslant n\exp(t^2b/2).$$

对上式两边同时取对数整理可得

$$E\left[\max_{i\in[n]} X_i\right] \leqslant \frac{\ln n}{t} + \frac{bt}{2}.$$

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 6.10 为例: 假设 $X_1, \ldots, X_n$ 是独立的、且参数为 b 的亚高斯随机变量, 对任意 $\epsilon > 0$ 有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geqslant \epsilon\right] \leqslant \exp(-n\epsilon^2/2b).$$

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

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \leqslant \sqrt{\frac{2b}{n}\ln\frac{1}{\delta}}.$$

前面讲的所有不等式都可以采用 $1-\delta$ 的形式描述.

6.3 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 6.12 (Bennet不等式) 设 $X_1, \ldots, X_n$ 是独立同分布的随机变量且满足 $X_i - E[X_i] \leqslant 1$, 其均值为 μ 和方差为 σ^2, 我们有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geqslant \epsilon\right] \leqslant \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2\epsilon/3}\right).$$

证明 对任意 $t > 0$, 根据 Chernoff 方法有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geqslant \epsilon\right] \leqslant e^{-nt\epsilon}E\left[\exp\left(\sum_{i=1}^{n} t(X_i - \mu)\right)\right] = e^{-nt\epsilon}\left(E[e^{t(X_1-\mu)}]\right)^n.$$

设 $Y = X_1 - \mu$, 利用公式 $\ln z \leqslant z - 1$ 得到

$$\ln E[e^{t(X_1-\mu)}] \quad = \quad \ln E[e^{tY}] \leqslant E[e^{tY}] - 1 = t^2E\left[\frac{e^{tY} - tY - 1}{t^2Y^2}Y^2\right]$$

$$\leqslant \quad t^2 E\left[\frac{e^t - t - 1}{t^2} Y^2\right] = (e^t - t - 1)\sigma^2$$

这里利用 $tY \leqslant t$ 以及 $(e^z - z - 1)/z^2$ 是一个非单调递减的函数. 进一步有

$$e^t - t - 1 \leqslant \frac{t^2}{2} \sum_{k=0}^{\infty} (t/3)^k = \frac{t^2}{2(1 - t/3)}.$$

因此可得

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n - \mu) \geqslant \epsilon\right] \leqslant \exp\left(-nt\epsilon + \frac{nt^2\sigma^2}{2(1 - t/3)}\right).$$

猜出 $t = \epsilon/(\sigma^2 + \epsilon/3)$, 带入完成证明.

对于 Bennet 不等式, 令

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n - \mu) \geqslant \epsilon\right] \leqslant \exp(-n\epsilon^2/(2\sigma^2 + 2\epsilon/3)) = \delta,$$

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

$$\frac{1}{n}\sum_{i=1}^{n} X_n \leqslant \mu + \frac{2\ln 1/\delta}{3n} + \sqrt{\frac{2\sigma^2}{n}\ln\frac{1}{\delta}}.$$

当方法 σ^2 非常小, 或趋于 0 时, 得到更紧的收敛率 $\bar{X}_n - \mu \leqslant O(1/n)$.

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 6.13 (Bernstein不等式) 设 $X_1, \ldots, X_n$ 是独立同分布的随机变量, 其均值为 μ 和方差为 σ^2, 若存在常数 $b > 0$, 使得对任意正整数 $m \geqslant 2$ 有 $E[X_i^m] \leqslant m! b^{m-2}\sigma^2/2$, 那么我们有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n - \mu) \geqslant \epsilon\right] \leqslant \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + 2b\epsilon}\right).$$

证明 对任意 $t > 0$, 根据 Chernoff 方法有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n - \mu) \geqslant \epsilon\right] \leqslant e^{-nt\epsilon} E\left[\exp\left(\sum_{i=1}^{n}(X_i - \mu)\right)\right] = e^{-nt\epsilon - n\mu t}\left(E[e^{tX_1}]\right)^n$$

利用公式 $\ln z \leqslant z - 1$ 有

$$\ln E[e^{tX_1}] \leqslant E[e^{tX}] - 1 = \sum_{m=1}^{\infty} E[X^m]\frac{t^m}{m!} \leqslant t\mu + \frac{t^2\sigma^2}{2}\sum_{m=2}^{\infty}(bt)^{m-2} = t\mu + \frac{t^2\sigma^2}{2(1 - bt)}.$$

由此可得

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\geqslant\epsilon\right]\leqslant=\exp\left(-nt\epsilon+\frac{t^2n\sigma^2}{2(1-bt)}\right)$$

取 $t=\epsilon/(\sigma^2+b\epsilon)$ 完成证明.

例 6.7 给出 Bernstein 不等式的 $1-\delta$ 表述.

6.4 应用: 随机投影 (Random Projection)

设高维空间 $\mathbb{R}^d$ 有 n 个点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ (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: $f:\mathbb{R}^d\to\mathbb{R}^k$ ($k\ll d$), 使得以较大概率有

$$(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\leqslant\|f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|_2^2\leqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2.$$

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

$$f(\boldsymbol{x})=\boldsymbol{x}P/c,$$

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 6.5 设 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 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

$$\boldsymbol{y}_i=f(\boldsymbol{x}_i)=\boldsymbol{x}_iP/\sqrt{k},\qquad i\in[n]$$

将 d 维空间中 n 个点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 通过随机矩阵 P 投影到 k 维空间. 对任意 $\epsilon\in(0,1/2)$, 当 $k\geqslant 8\log 2n/(\epsilon^2-\epsilon^3)$ 时至少以 $1/2$ 的概率有

$$(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\leqslant\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\qquad(i,j\in[n]).$$

证明 下面分三步证明 J-L 引理.

第一步: 对任意非零 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)\in\mathbb{R}^d$, 首先证明

$$E\left[\left\|\boldsymbol{x}P/\sqrt{k}\right\|_2^2\right]=\|\boldsymbol{x}\|_2^2,$$

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

$$\begin{aligned}E\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\right] &= E\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{d}\frac{x_ip_{ij}}{\sqrt{k}}\right)^2\right]=\sum_{j=1}^{k}\frac{1}{k}E\left[\left(\sum_{i=1}^{d}x_ip_{ij}\right)^2\right]\\ &= \sum_{j=1}^{k}\frac{1}{k}\sum_{i=1}^{d}x_i^2=\frac{1}{k}\sum_{j=1}^{k}\|\boldsymbol{x}\|_2^2=\|\boldsymbol{x}\|_2^2.\end{aligned}$$

第二步: 对任意非零 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)\in\mathbb{R}^d$, 证明

$$\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant\exp(-(\epsilon^2-\epsilon^3)k/4).$$

将矩阵 P 表示为 $P=(P_1,P_2,\ldots,P_k)$, 其中 P_i ($i\in[d]$) 是一个 $d\times1$ 的列向量, 令 $v_j=\boldsymbol{x}P_j/\|\boldsymbol{x}\|_2$, 即

$$(v_1,v_2,\ldots,v_k)=\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}P_1,\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}P_2,\ldots,\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}P_k\right).$$

根据 Gaussian 分布的性质有 $v_j\sim\mathcal{N}(0,1)$, 且 $v_1,v_2,\ldots,v_k$ 是 k 个独立的随机变量. 对任意 $t\in(0,1/2)$, 根据 Chernoff 方法有

$$\begin{aligned}&\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]=\Pr\left[\left\|\frac{\boldsymbol{x}P}{\|\boldsymbol{x}\|_2}\right\|_2^2\geqslant(1+\epsilon)k\right]\\ =\ &\Pr\left[\sum_{j=1}^{k}v_j^2\geqslant(1+\epsilon)k\right]\leqslant e^{-(1+\epsilon)kt}\left(E[e^{t\sum_{j=1}^{k}v_j^2}]\right)^k=e^{-(1+\epsilon)kt}\left(E[e^{tv_1^2}]\right)^k.\end{aligned}$$

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

$$\Pr\left[\left\|\boldsymbol{x}P/\sqrt{k}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant\left(\frac{e^{-2(1+\epsilon)t}}{1-2t}\right)^{k/2}.$$

上式右边对 t 求最小解得 $t_{\min}=\frac{\epsilon}{2(1+\epsilon)}$, 代入可得

$$\Pr\left[\left\|\boldsymbol{x}P/\sqrt{k}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant\left((1+\epsilon)e^{-\epsilon}\right)^{k/2}.$$

设 $f(\epsilon)=\ln(1+\epsilon)$, 根据 $\epsilon\in(0,1/2)$ 有

$$f^{'}(\epsilon)=\frac{1}{1+\epsilon},f^{''}(\epsilon)=-\frac{1}{(1+\epsilon)^2},f^{'''}(\epsilon)=\frac{2}{(1+\epsilon)^3}\leqslant 2.$$

根据泰勒中值定理有

$$f(\epsilon)=f(0)+f^{'}(0)\epsilon+\frac{f^{''}(0)\epsilon^2}{2!}+\frac{f^{'''}(\xi)\epsilon^3}{3!}\leqslant\epsilon-\frac{\epsilon^2}{2}+\frac{\epsilon^2}{3}\leqslant\epsilon-\frac{\epsilon^2-\epsilon^3}{2}.$$

于是得到

$$\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4}.$$

同理可证

$$\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1-\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4}.$$

第三步: 对任意给定 $i\neq j$, 根据第二步的结论可知

$$\Pr[\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4},$$
$$\Pr[\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4}.$$

由于 $i,j\in[n]$, 因此共有 $n(n-1)$ 对 (i,j), 根据 Union 不等式有

$$\Pr\left[\exists i\neq j\colon\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\quad\text{或}\quad\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\right]\leqslant 2n^2e^{-k(\epsilon^2-\epsilon^3)/4},$$

设 $2n^2e^{-k(\epsilon^2-\epsilon^3)/4}\leqslant 1/2$, 求解 $k\geqslant 8\log 2n/(\epsilon^2-\epsilon^3)$. 引理得证.

习题

6.1 设随机变量 X 的期望 $E[X]=\mu>0$, 方差为 σ^2, 证明对任意 $\epsilon>0$ 有

$$P(X-\mu\leqslant-\epsilon)\leqslant\frac{\sigma^2}{\sigma^2+\epsilon^2}.$$

6.2 设随机变量X和Y满足$E(X)=-2$, $E(Y)=2$, $\mathrm{Var}(X)=1$, $Var(Y)=4$, $\rho_{XY}=-1/2$. 利用Chebyshev不等式估计$\Pr(|X+Y|\geqslant 6)$的上界.

6.3 独立同分布随机变量 $X_1,X_2,\cdots,X_n$ 满足 $E[X_i]=\mu$ 和 $\mathrm{Var}(X_i)\leqslant v$. 证明对任意 $\epsilon>0$ 有

$$\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\geqslant\epsilon\right]\leqslant\frac{v}{n\epsilon^2}\ .$$

6.4 阐述什么是chernoff方法。

6.5 随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且满足 $X_i\sim\mathrm{Ber}(p_i)$ $(p_i>0)$. 利用chernoff方法给出下列概率的上界

$$P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-E[X_i])\geqslant\epsilon\right]\quad\text{和}\quad P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-E[X_i])\leqslant-\epsilon\right]\ .$$

6.6 若独立同分布随机变量 $X_1,X_2,\cdots,X_n$ 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

$$P\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{a+b}{2}\right)\geqslant\epsilon\right]\quad\text{和}\quad\Pr\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{a+b}{2}\right)\leqslant-\epsilon\right]\ .$$

6.7 随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且满足 $X_i\sim\mathrm{Ber}(p_i)$ $(p_i>0)$. 证明对任意 $0<\epsilon<1$ 有不等式

$$P\left[\sum_{i=1}^{n}X_i\geqslant(1+\epsilon)\sum_{i=1}^{n}p_i\right]\leqslant e^{-\mu\epsilon^2/3}\ .$$

6.8 随机变量 $X\in[a,b]$ 且期望 $\mu=\mathbb{E}[x]$, 证明对任意 $t>0$ 有

$$\mathbb{E}\left[e^{tx}\right]\leqslant\exp\left(\mu t+t^2(b-a)^2/8\right)$$

6.9 利用chernoff方法证明: 设$X_1,X_2,\cdots,X_k$ 是k 个独立的随机变量, 且$X_i\sim N(0,1)$, 则有

$$\Pr\left(\sum_{i=1}^{k}X_i^2\geqslant(1+\epsilon)k\right)\leqslant\exp\left(-k\left(\epsilon^2-\epsilon^3\right)/4\right)$$

6.10 证明 Bennet 不等式.

6.11 证明 Berstein 不等式.

6.12 已知 Berstein 不等式

$$P\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)\geqslant\epsilon\right]\leqslant\exp\left(\frac{-n\epsilon^2}{2\sigma^2+2b\epsilon}\right),$$

给出其等价 $1-\delta$ 描述。

6.13 已知独立同分布随机变量 $X_1, X_2, \cdots, X_n$ 满足 $X_i \sim N\left(\mu, \sigma^2\right)$, 给出 $E\left[\max_{i\in[n]}\{X_i\}\right]$ 的上界, 并给出严格证明。

第 7 章　大数定律及中心极限定理

7.1 大数定律

给定随机变量 $X_1, X_2, \cdots, X_n$, 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 7.1 (依概率收敛) 设 $X_1, X_2, \cdots, X_n, \cdots$ 是一随机变量序列, a 是一常数, 如果对任意 $\epsilon > 0$ 有

$$\lim_{n\to\infty} \Pr\{|X_n - a| < \epsilon\} = 1 \quad 或 \quad \lim_{n\to\infty} \Pr\{|X_n - a| > \epsilon\} = 0,$$

则称随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 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 $g\colon \mathbb{R} \to \mathbb{R}$ 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 $g\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ 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 $X_nY_n \xrightarrow{P} ab$.

定理 7.1 (大数定律) 若随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 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 7.2 (马尔可夫 Markov 大数定律) 如果随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足

$$\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \to 0 \qquad n \to \infty,$$

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 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立或同分布, 其证明直接通过 Chebyshev 不等式有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geqslant \epsilon\right] \leqslant \frac{1}{n^2\epsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \to 0 \qquad n \to \infty.$$

定理 7.3 (切比雪夫 Chebyshev 大数定律) 设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立, 且存在常数 $c > 0$ 使得 $\mathrm{Var}(X_n) \leqslant c$, 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

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geqslant \epsilon\right] \leqslant \frac{1}{\epsilon^2 n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \leqslant \frac{c}{n\epsilon^2} \to 0 \qquad n \to \infty.$$

定理 7.4 (辛钦 Khintchine 大数定律) 设 $X_1, X_2, \cdots, X_n, \cdots$ 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 7.5 (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

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - p\right| \geqslant \epsilon\right] = 0,$$

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 $Y_1, Y_2, \ldots, Y_n$ 满足 $Y_i \sim \mathrm{Ber}(p)$, 则

$$X_n = \sum_{i=1}^{n} Y_i \sim \mathrm{B}(n, p).$$

于是得到

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - p\right| \geqslant \epsilon\right] = \lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - E[Y_i]\right| \geqslant \epsilon\right] \leqslant \frac{1}{\epsilon^2 n^2}\mathrm{Var}(\sum_{i=1}^{n} Y_i) = \frac{p(1-p)}{\epsilon^2 n} \to 0.$$

如何判断随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 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 7.1 独立的随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 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 $\mathrm{Var}(X_i)=E[X_i^2]=i^{1/2}$, 根据 Chebysheve 不等式和独立性有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right|\geqslant\epsilon\right]\leqslant\frac{1}{n^2\epsilon^2}\mathrm{Var}(\sum_{i=1}^{n}X_i)=\frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}Var(X_i)=\frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}i^{1/2}\leqslant\frac{1}{\epsilon^2\sqrt{n}}$$

再根据

$$\sum_{i=1}^{n}i^{1/2}\leqslant\sum_{i=1}^{n}\int_i^{i+1}i^{1/2}dx\leqslant\sum_{i=1}^{n}\int_i^{i+1}x^{1/2}dx=\int_1^{n+1}x^{1/2}dx=2((n+1)^{3/2}-1)/3$$

由此可得当 $n\to+\infty$ 时有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right|\geqslant\epsilon\right]\leqslant\frac{2((n+1)^{3/2}-1)/3}{\epsilon^2n^2}\to0$$

大数定律小结:

- Markov 大数定律: 若随机变量序列 $\{X_i\}$ 满足 $\mathrm{Var}(\sum_{i=1}^{n}X_n)/n^2\to0$, 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 $\mathrm{Var}(X_i)\leqslant c$, 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 $X_n\sim\mathrm{B}(n,p)$, 有 $X_n/n\xrightarrow{P}p$.

7.2 中心极限定理

对独立的随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$, 我们考虑标准化后随机变量

$$Y_n=\frac{\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}E(X_i)}{\sqrt{\mathrm{Var}(\sum_{i=1}^{n}X_i)}}$$

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 7.2 设随机变量 Y 的分布函数为 $F_Y(y)=\Pr(Y\leqslant y)$, 以及随机变量序列 $Y_1,Y_2,\cdots,Y_n,\cdots$ 的分布函数分别为 $F_{Y_n}(y)=\Pr(Y_n\leqslant y)$, 如果

$$\lim_{n\to\infty}\Pr[Y_n\leqslant y]=\Pr[Y\leqslant y],\quad\text{即}\quad\lim_{n\to\infty}F_{Y_n}(y)=F_Y(y),$$

则称随机变量序列 $Y_1,Y_2,\cdots,Y_n,\cdots$ 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 7.6 设独立同分布的随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 的期望 $E(X_1) = \mu$ 和方差 $Var(X_1) = \sigma^2$, 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

$$\lim_{n\to\infty} \Pr[Y_n \leqslant y] = \Phi(y).$$

随机变量 Y_n 是随机变量 $X_1, X_2, \ldots, X_n$ 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

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2), \qquad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

大数定律给出了当 $n \to \infty$ 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 7.2 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 $V_1, V_2, \ldots, V_{20}$ 服从均匀分布 $U(0,10)$, 于是有 $E(V_k) = 5$ 和 $\mathrm{Var}(V_k) = 100/12 = 25/3$. 设 $V = \sum_{k=1}^{20} V_k$, 则有

$$E(V) = 100 \qquad \mathrm{Var}(V) = 500/3.$$

根据中心极限定理近似有

$$\frac{V - E(V)}{\sqrt{\mathrm{Var}(V)}} = \frac{V - 100}{\sqrt{500/3}} \sim \mathcal{N}(0,1).$$

根据标准正态分布的分布函数 $\Phi(x)$ 有

$$\Pr(V \geqslant 105) = \Pr\left(\frac{V-100}{\sqrt{500/3}} \geqslant \frac{105-100}{\sqrt{500/3}}\right) = \Pr\left(\frac{V-100}{\sqrt{500/3}} \geqslant 0.387\right) = 1 - \Phi(0.387).$$

查表完成证明.

例 7.3 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 $\mathrm{Var}(X_i) = 25$. 设总重量 $X = \sum_{i=1}^{n} X_i$, 则有 $E(X) = 50n$ 和 $\mathrm{Var}(X) = 25n$. 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

$$\Pr(X \leqslant 5000) = \Pr\left(\frac{X-50n}{\sqrt{25n}} \leqslant \frac{5000-50n}{\sqrt{25n}}\right) = \Phi\left(\frac{5000-50n}{\sqrt{25n}}\right) > 0.977 = \Phi(2).$$

根据分布函数的单调性有

$$\frac{1000-10n}{\sqrt{n}} > 2 \Longrightarrow 1000n^2 - 2000n + 1000^2 > 4n.$$

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 7.1 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 $X_n \overset{近似}{\sim} \mathcal{N}(np, np(1-p))$, 从而有如下近似估计:

$$\Pr[X_n \leqslant y] = \Pr\left[\frac{X_n - np}{\sqrt{np(1-p)}} \leqslant \frac{y-np}{\sqrt{np(1-p)}}\right] \approx \Phi\left(\frac{y-np}{\sqrt{np(1-p)}}\right).$$

针对上式, 可以考虑三种问题: i) 已知 n 和 $\Pr[X_n \leqslant y]$, 求 y; ii) 已知 n 和 y, 求 $\Pr[X_n \leqslant y]$; iii)已知 y 和 $\Pr[X_n \leqslant y]$, 求 n. 下面看三个例子:

例 7.4 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

$$\Pr(X \leqslant y) \geqslant 0.999 \quad \Rightarrow \quad \Pr\left(\frac{X-120}{\sqrt{48}} \leqslant \frac{y-120}{\sqrt{48}}\right) \approx \Phi\left(\frac{y-120}{\sqrt{48}}\right) \geqslant 0.999 = \Phi(3.1).$$

所以有 $\frac{y-120}{\sqrt{48}} \geqslant 3.1$, 求解可得 $y \geqslant 141$.

例 7.5 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 $\mathrm{Var}(X) = 9$. 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

$$\Pr(X \leqslant 15) = \Pr\left(\frac{X-10}{\sqrt{9}} \leqslant \frac{15-10}{\sqrt{9}}\right) \approx \Phi(5/3).$$

例 7.6 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 $|k/n - p| \leqslant 0.05$ 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

$$\begin{aligned}\Pr\left[\left|\frac{X_n}{n} - p\right| \leqslant 0.05\right] &= \Pr\left[\frac{|X_n - np|}{n} \leqslant 0.05\right] = \Pr\left[\frac{|X_n - np|}{\sqrt{np(1-p)}} \leqslant \frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right] \\ &= \Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right)\end{aligned}$$

对于标准正太分布函数有 $\Phi(-\alpha) = 1 - \Phi(\alpha)$ 以及 $p(1-p) \leqslant 1/4$, 于是有

$$\Pr\left[\left|\frac{X_n}{n} - p\right| \leqslant 0.05\right] = 2\Phi\left(\frac{0.05\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 > 2\Phi\left(\sqrt{n}/10\right) - 1 > 0.9.$$

所以 $\Phi\left(\sqrt{n}/10\right) \geqslant 0.95$, 查表解得 $n \geqslant 271$.

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 7.7 设独立随机变量 $X_1, X_2, \ldots, X_n, \ldots$ 的期望 $E[X_n] = \mu_n$ 和方差 $\mathrm{Var}(X_n) = \sigma_n^2 > 0$. 记 $B_n^2 = \sum_{k=1}^n \sigma_k^2$, 若存在 $\delta > 0$, 当 $n \to \infty$ 时有

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E[|X_k - \mu_k|^{2+\delta}] \to 0$$

成立, 则有

$$Y_n = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n E(X_k)}{\sqrt{Var(\sum_{k=1}^n X_k)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

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 $\mathrm{Var}(X_k) = \sigma^2$, 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

7.1 设随机变量 X 的期望 $E[X]=\mu>0$, 方差为 σ^2, 证明对任意 $\epsilon>0$ 有

$$P(X-\mu\leqslant-\epsilon)\leqslant\frac{\sigma^2}{\sigma^2+\epsilon^2}.$$

7.2 设随机变量X和Y满足$E(X)=-2$, $E(Y)=2$, $\mathrm{Var}(X)=1$, $Var(Y)=4$, $\rho_{XY}=-1/2$. 利用Chebyshev不等式估计$\Pr(|X+Y|\geqslant 6)$的上界.

7.3 独立同分布随机变量 $X_1,X_2,\cdots,X_n$ 满足 $E[X_i]=\mu$ 和 $\mathrm{Var}(X_i)\leqslant v$. 证明对任意 $\epsilon>0$ 有

$$\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|\geqslant\epsilon\right]\leqslant\frac{v}{n\epsilon^2}\ .$$

7.4 阐述什么是chernoff方法。

7.5 随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且满足 $X_i\sim\mathrm{Ber}(p_i)$ $(p_i>0)$. 利用chernoff方法给出下列概率的上界

$$P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-E[X_i])\geqslant\epsilon\right]\quad\text{和}\quad P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-E[X_i])\leqslant-\epsilon\right]\ .$$

7.6 若独立同分布随机变量 $X_1,X_2,\cdots,X_n$ 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

$$P\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{a+b}{2}\right)\geqslant\epsilon\right]\quad\text{和}\quad\Pr\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{a+b}{2}\right)\leqslant-\epsilon\right]\ .$$

7.7 随机变量 $X_1,X_2,\cdots,X_n$ 相互独立且满足 $X_i\sim\mathrm{Ber}(p_i)$ $(p_i>0)$. 证明对任意 $0<\epsilon<1$ 有不等式

$$P\left[\sum_{i=1}^{n}X_i\geqslant(1+\epsilon)\sum_{i=1}^{n}p_i\right]\leqslant e^{-\mu\epsilon^2/3}\ .$$

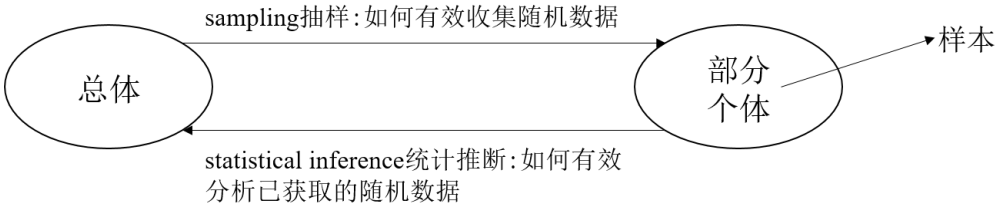
7.8 随机变量 $X\in[a,b]$ 且期望 $\mu=\mathbb{E}[x]$, 证明对任意 $t>0$ 有

$$\mathbb{E}\left[e^{tx}\right]\leqslant\exp\left(\mu t+t^2(b-a)^2/8\right)$$

第 8 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.



8.1 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 $X_1, X_2, \cdots, X_n$, 称 $X_1, X_2, \cdots, X_n$ 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: $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ 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 8.1 (简单随机样本) 称样本 $X_1, X_2, \cdots, X_n$ 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 $X_1, X_2, \cdots, X_n$ 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 $X_1, X_2, \cdots, X_n$ 的联合分布函数为

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

若总体 X 的概率密度为 $f(x)$, 则样本 $X_1, X_2, \cdots, X_n$ 的联合概率密度为

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

若总体 X 的分布列 $\Pr(X = x_i)$, 则样本 $X_1, X_2, \cdots, X_n$ 的联合分布列为

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

8.2 常用统计量

为研究样本的特性, 我们引入统计量:

定义 8.2 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, $g(X_1, X_2, \cdots, X_n)$ 是关于 $X_1, X_2, \cdots, X_n$ 的一个连续、且不含任意参数的函数, 称 $g(X_1, X_2, \cdots, X_n)$ 是一个 **统计量**.

由于 $X_1, X_2, \cdots, X_n$ 是随机变量, 因此统计量 $g(X_1, X_2, \cdots, X_n)$ 是一个随机变量. 而 $g(x_1, x_2, \cdots, x_n)$ 为 $g(X_1, X_2, \cdots, X_n)$ 的一次观察值. 下面研究一些常用统计量.

假设 $X_1, X_2, \cdots, X_n$ 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 8.1 设总体 X 的期望为 $E[X] = \mu$, 方差 $\mathrm{Var}(X) = \sigma^2$, 则有

$$E[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \sigma^2/n, \qquad \bar{X} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n).$$

假设 $X_1, X_2, \cdots, X_n$ 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 8.2 设总体 X 的期望为 $E[X] = \mu$, 方差 $\mathrm{Var}(X) = \sigma^2$, 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 σ^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:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{即} \qquad S^2 = \frac{n}{n-1}S_0^2,$$

引理 8.3 设总体 X 的期望为 $E[X] = \mu$, 方差 $\mathrm{Var}(X) = \sigma^2$, 则有

$$E[S^2] = \sigma^2.$$

证明 根据期望的性质有

$$E[S^2] = E\left[\frac{n}{n-1}S_0^2\right] = \frac{n}{n-1}E\left[S_0^2\right] = \sigma^2.$$

假设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 定义 **样本 k 阶原点矩** 为:

$$A_k = \frac{1}{n}\sum_{i=1}^{n}X_i^k, \qquad k = 1, 2, \cdots.$$

定义 **样本 k 阶中心矩** 为:

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, \qquad k = 1, 2, \cdots.$$

例 8.1 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 $X_1, X_2, \ldots, X_{10}$ 和 $X_1', X_2', \ldots, X_{15}'$ 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

$$\bar{X}_1 = \frac{1}{10}\sum_{i=1}^{10}X_i \sim \mathcal{N}(20, 3/10), \qquad \bar{X}_2 = \frac{1}{15}\sum_{i=1}^{15}X_i' \sim \mathcal{N}(20, 1/5).$$

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 $X_1, X_2, \cdots, X_n$ 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:

$$X_{(1)} = \min\{X_1, X_2, \cdots, X_n\} \qquad \text{和} \qquad X_{(n)} = \max\{X_1, X_2, \cdots, X_n\},$$

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

$$F_{X_{(1)}}(x) = \Pr(X_{(1)} \leqslant x) = 1 - \Pr(X_{(1)} > x) = 1 - (1 - F(x))^n, \qquad F_{X_{(n)}}(x) = F^n(x).$$

定理 8.1 设总体 X 的密度函数为 $f(x)$, 分布函数为 $F(x)$, $X_1, X_2, \cdots, X_n$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

$$\begin{aligned} F_k(x) &= \Pr[X_{(k)} \leqslant x] = \Pr[X_1, X_2, \cdots, X_n \text{ 中至少有 } k \text{ 个随机变量 } \leqslant x] \\ &= \sum_{r=k}^{n} \Pr[X_1, X_2, \cdots, X_n \text{ 中恰有 } r \text{ 个随机变量 } \leqslant x,\ n-r \text{ 个随机变量 } > x] \\ &= \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r}. \end{aligned}$$

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.

8.3 Beta 分布、Γ 分布、Dirichlet 分布

首先介绍两积分函数.

定义 8.3 (Beta-函数) 对任意给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 定义 Beta 函数为

$$\mathrm{Beta}(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1} dx,$$

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 $\mathrm{Beta}(\alpha_1, \alpha_2)$ 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

$$\begin{aligned} \mathrm{Beta}(\alpha_1, \alpha_2) &= \int_0^1 t^{\alpha_1 - 1}(1-t)^{\alpha_2 - 1} dt = \int_1^0 (1-x)^{\alpha_1 - 1} x^{\alpha_2 - 1} d(1-x) \\ &= \int_0^1 x^{\alpha_2 - 1}(1-x)^{\alpha_1 - 1} dx = \mathrm{Beta}(\alpha_2, \alpha_1), \end{aligned}$$

由此可知 Beta 函数的对称性: $\mathrm{Beta}(\alpha_1, \alpha_2) = \mathrm{Beta}(\alpha_2, \alpha_1)$.

定义 8.4 (Γ-函数) 对任意给定 $\alpha > 0$, 定义 Γ-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 8.1 对 Γ-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 Γ-函数, 有如下关系:

定理 8.2 对任意给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 有

$$\mathrm{Beta}(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}.$$

证明 根据 Γ-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

$$\begin{aligned} \Gamma(\alpha_1)\Gamma(\alpha_2) &= \int_0^1 \int_0^{+\infty} e^{-x} x^{\alpha_1 - 1} y^{\alpha_1 - 1} x^{\alpha_2 - 1} (1 - y)^{\alpha_2 - 1} |x| dx dy \\ &= \int_0^1 \int_0^{+\infty} e^{-x} x^{\alpha_1 + \alpha_2 - 1} y^{\alpha_1 - 1} (1 - y)^{\alpha_2 - 1} dx dy \\ &= \int_0^{+\infty} e^{-x} x^{\alpha_1 + \alpha_2 - 1} dx \int_0^1 y^{\alpha_1 - 1} (1 - y)^{\alpha_2 - 1} dy \\ &= \Gamma(\alpha_1 + \alpha_2)\mathrm{Beta}(\alpha_1, \alpha_2) \end{aligned}$$

定理得证.

根据上述定理可知

推论 8.1 对任意 $\alpha_1 > 1$ 和 $\alpha_2 > 0$, 有

$$Beta(\alpha_1, \alpha_2) = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 1} Beta(\alpha_1 - 1, \alpha_2).$$

证明 根据前面的定理有

$$\mathrm{Beta}(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} = \frac{(\alpha_1 - 1)\Gamma(\alpha_1 - 1)\Gamma(\alpha_2)}{(\alpha_1 + \alpha_2 - 1)\Gamma(\alpha_1 + \alpha_2 - 1)} = \frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 1}\mathrm{Beta}(\alpha_1 - 1, \alpha_2).$$

定义 8.5 对任意 $\alpha_1, \alpha_2, \cdots, \alpha_k > 0$, 定义多维 Beta 函数为

$$\mathrm{Beta}(\alpha_1, \alpha_2, \cdots, \alpha_k) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)}{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}.$$

下面介绍三种分布:

定义 8.6 (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 α_1 和 α_2 的 Beta 分布，记 $X \sim B(\alpha_1, \alpha_2)$.

定理 8.3 若随机变量 $X \sim B(\alpha_1, \alpha_2)$, 则有

$$E[X] = \frac{\alpha_1}{\alpha_1+\alpha_2} \quad \text{和} \quad \mathrm{Var}(X) = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}.$$

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

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{\alpha_1(1+\alpha_1)}{(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1)} - (\frac{\alpha_1}{\alpha_1+\alpha_2})^2 = \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}.$$

例 8.2 设独立同分布随机变量 $X_1, X_2, \cdots, X_n$ 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 Γ 分布:

定义 8.7 如果随机变量 X 的概率密度

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

其中 $\alpha > 0$ 和 $\lambda > 0$, 则称随机变量 X 服从参数为 α 和 λ 的 Γ 分布, 记为$X \sim \Gamma(\alpha, \lambda)$.

定理 8.4 若随机变量 $X \sim \Gamma(\alpha, \lambda)$, 则有 $E(X) = \alpha/\lambda$ 和 $\mathrm{Var}(X) = \alpha/\lambda^2$.

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

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \alpha(\alpha+1)/\lambda^2 - \alpha^2/\lambda^2 = \alpha/\lambda^2.$$

我们有 Γ 分布的可加性:

定理 8.5 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

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-\frac{1}{2}x} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

例 8.3 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 $y \leqslant 0$ 时有 $F_Y(y) = 0$; 当 $y > 0$ 时有

$$F_Y(y) = \Pr(X^2 \leqslant y) = \Pr(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,$$

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 8.8 给定 $\alpha_1, \alpha_2, \cdots, \alpha_k \in (0, +\infty)$, 若多元随机向量 $X = (X_1, X_2, \cdots, X_k)$ 的密度函数为

$$f(x_1, x_2, \cdots, x_k) = \begin{cases} \frac{x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_k^{\alpha_k - 1}}{\mathrm{Beta}(\alpha_1, \alpha_2, \cdots, \alpha_k)} & \sum_{i=1}^k x_i = 1,\ x_i > 0\ (i \in [k]), \\ 0 & \text{其它} \end{cases}$$

则称 X 服从参数为 $\alpha_1, \alpha_2, \cdots, \alpha_k$ 的 Dirichlet 分布, 记 $X \sim \mathrm{Dir}(\alpha_1, \alpha_2, \cdots, \alpha_k)$.

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 8.6 若随机向量 $X = (X_1, X_2, \cdots, X_k) \sim \mathrm{Dir}(\alpha_1, \alpha_2, \cdots, \alpha_k)$, 设 $\tilde{\alpha} = \alpha_1 + \alpha_2 + \cdots + \alpha_k$ 和 $\tilde{\alpha}_i = \alpha_i / \tilde{\alpha}$, 则

$$E[X_i] = \tilde{\alpha}_i \qquad \text{和} \qquad Cov(X_i, X_j) = \begin{cases} \frac{\tilde{\alpha}_i(1 - \tilde{\alpha}_i)}{\tilde{\alpha} + 1} & i = j, \\ -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha} + 1} & i \neq j. \end{cases}$$

证明 根据期望的定义有

$$\begin{aligned} E[X_i] &= \frac{\int\int_{\sum_i x_i = 1, x_i \geqslant 0} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_k^{\alpha_k - 1} \cdot x_i dx_1 \cdots dx_k}{\mathrm{Beta}(\alpha_1, \alpha_2, \cdots, \alpha_k)} \\ &= \frac{\mathrm{Beta}(\alpha_1, \cdots, \alpha_i + 1, \cdots, \alpha_k)}{\mathrm{Beta}(\alpha_1, \cdots, \alpha_i, \cdots, \alpha_k)} = \frac{\alpha_i}{\alpha_1 + \alpha_2 + \cdots + \alpha_k} = \tilde{\alpha}_i. \end{aligned}$$

若 $i = j$, 则有

$$\mathrm{Cov}(X_i, X_i) = E[X_i^2] - (E[X_i])^2 = \frac{Beta(\alpha_1, \cdots, \alpha_i + 2, \cdots, \alpha_k)}{Beta(\alpha_1, \cdots, \alpha_i, \cdots, \alpha_k)} - (\tilde{\alpha}_i)^2 = \frac{\tilde{\alpha}_i(1 - \tilde{\alpha}_i)}{\tilde{\alpha} + 1}.$$

若 $i \neq j$, 则有

$$\begin{aligned} Cov(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] = \frac{\mathrm{Beta}(\alpha_1, \cdots, \alpha_i + 1, \cdots, \alpha_j + 1, \cdots, \alpha_k)}{\mathrm{Beta}(\alpha_1, \cdots, \alpha_i, \cdots, \alpha_j, \cdots, \alpha_k)} - \tilde{\alpha}_i \tilde{\alpha}_j \\ &= \frac{\alpha_i \alpha_j}{\tilde{\alpha}(\tilde{\alpha} + 1)} - \tilde{\alpha}_i \tilde{\alpha}_j = -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha} + 1}. \end{aligned}$$

8.4 正态总体抽样分布定理

8.4.1 χ^2 分布

定义 8.9 若 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(0,1)$ 的一个样本, 称 $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ 为服从自由度为 n 的 χ^2 分布, 记 $Y \sim \chi^2(n)$.

根据 $X_1^2 \sim \Gamma(1/2, 1/2)$ 和 Γ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

$$f_Y(y) = \begin{cases} \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} & y > 0 \\ 0 & y \leqslant 0 \end{cases}$$

下面研究 χ^2 分布的性质:

定理 8.7 若随机变量 $X \sim \chi^2(n)$, 则 $E(X) = n$ 和 $\mathrm{Var}(X) = 2n$; 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 $X = X_1^2 + X_2^2 + \cdots + X_n^2$, 其中 $X_1, X_2, \cdots, X_n$ 是总体为 $X' \sim \mathcal{N}(0,1)$ 的一个样本. 我们有

$$\begin{aligned} E[X] &= E[X_1^2 + X_2^2 + \cdots + X_n^2] = nE[X_1^2] = n, \\ \mathrm{Var}(X) &= n\mathrm{Var}(X_1^2) = n[E(X_1^4) - (E(X_1^2))^2] = n(E(X_1^4) - 1). \end{aligned}$$

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 $\mathrm{Var}(X) = 2n$.

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 $(2k)!! = 2k \cdot (2k-2) \cdot \cdots \cdot 2$ 和 $(2k+1)!! = (2k+1) \cdot (2k-1) \cdot \cdots \cdot 1$.

例 8.4 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 χ^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

$$\frac{X_1 - 2X_2}{2\sqrt{5}} \sim \mathcal{N}(0,1), \qquad \frac{3X_3 - 4X_4}{10} \sim \mathcal{N}(0,1),$$

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 $X + Y \sim P(\lambda_1 + \lambda_1)$;
- 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 $X + Y \sim \chi(m + n)$.

8.4.2 t 分布 (student distribution)

定义 8.10 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

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad x \in (-\infty, +\infty).$$

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 $n \to \infty$ 时, 随机变量 $T \sim t(n)$ 的概率密度

$$f(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

因此当 n 足够大时, $f(x)$ 可被近似为 $\mathcal{N}(0, 1)$ 的密度函数.

8.4.3 F 分布

定义 8.11 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

$$f(x)=\begin{cases}\frac{\Gamma(\frac{m+n}{2})(\frac{m}{n})^{\frac{m}{2}}x^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})(1+\frac{mx}{n})^{\frac{m+n}{2}}} & x>0\\ 0 & x\leqslant 0\end{cases}$$

若随机变量 $F\sim F(m,n)$, 则 $\frac{1}{F}=F(n,m)$.

课题练习:

- 独立同分布随机变量 $X_1,X_2,\cdots,X_n$ 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 $X_1,X_2,\cdots,X_9$ 和 $Y_1,Y_2,\cdots,Y_9$ 是分别来自总体 $\mathcal{N}(0,9)$ 的两个独立样本, 求 $(X_1+X_2+\cdots+X_9)/\sqrt{Y_1^2+Y_1^2+\cdots+Y_9^2}$ 的分布.
- 设 $X_1,X_2,\ldots,X_{2n}$ 来自总体 $\mathcal{N}(0,\sigma_2)$ 的样本, 求 $(X_1^2+X_3^2+\cdots+X_{2n-1}^2)/(X_2^2+X_4^2+\cdots+X_{2n}^2)$ 的分布.

8.4.4 正态分布的抽样分布定理

定理 8.8 设 $X_1,X_2,\cdots,X_n$ 是来自总体 $\mathcal{N}(\mu,\sigma^2)$ 的样本, 则有

$$\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i\sim\mathcal{N}(\mu,\frac{\sigma^2}{n}),\qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim\mathcal{N}(0,1).$$

定理 8.9 设 $X_1,X_2,\cdots,X_n$ 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

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

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 8.10 设 $X_1,X_2,\cdots,X_n$ 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

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

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 8.11 设 $X_1,X_2,\cdots,X_m$ 和 $Y_1,Y_2,\cdots,Y_n$ 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 8.9 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 8.12 设 $X_1,X_2,\cdots,X_m$ 和 $Y_1,Y_2,\cdots,Y_n$ 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 8.9 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 $X_1, X_2, \cdots, X_5$ 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 χ^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.

8.4.5 分位数(点)

定义 8.12 对给定 $\alpha \in (0,1)$ 和随机变量 X, 称满足 $\Pr(X > \lambda_\alpha) = \alpha$ 的实数 λ_α 为上侧 α 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 μ_α 称为正态分布上侧 α 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 $\Pr(X \geqslant \chi^2_\alpha(n)) = \alpha$ 的点 $\chi^2_\alpha(n)$ 称为 $\chi^2(n)$ 分布上侧 α 分位点. 当 $n \to \infty$ 时有 $\chi^2_\alpha(n) \approx \frac{1}{2}(\mu_\alpha + \sqrt{2n-1})^2$, 其中 μ_α 表示正态分布上侧 α 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 α 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 α 分位点.

对于 F-分布, 有如下性质:

引理 8.4 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

$$1 - \alpha = \Pr(X > F_{1-\alpha}(m,n)) = \Pr\left(\frac{1}{X} < \frac{1}{F_{1-\alpha}(m,n)}\right) = 1 - \Pr\left(\frac{1}{X} \geqslant \frac{1}{F_{1-\alpha}(m,n)}\right).$$

再根据 $1/X \sim F(n,m)$, 结合上式有

$$\alpha = \Pr\left(\frac{1}{X} \geqslant \frac{1}{F_{1-\alpha}(m,n)}\right) = \Pr\left(\frac{1}{X} > \frac{1}{F_{1-\alpha}(m,n)}\right)$$

于是有 $F_\alpha(n,m) = 1/F_{1-\alpha}(m,n)$.

课堂习题:

- 设 $X_1, X_2, \cdots, X_{10}$ 是总体 $\mathcal{N}(\mu, 1/4)$ 的样本, i) 若 $\mu = 0$, 求 $\Pr(\sum_{i=1}^{10} X_i^2 \geqslant 4)$; ii) 若 μ 未知, 求 $\Pr(\sum_{i=1}^{10}(X_i - \bar{X})^2 \geqslant 2.85)$.
- 设 $X_1, X_2, \cdots, X_{25}$ 是总体 $\mathcal{N}(12, \sigma^2)$ 的样本, i) 若 $\sigma = 2$, 求 $\Pr(\sum_{i=1}^{25} X_i/25 \geqslant 12.5)$; ii) 若 σ 未知但知道修正样本方差为 $S^2 = 5.57$, 求 $\Pr(\sum_{i=1}^{25} X_i/25 \geqslant 12.5)$.

习题

8.1 设随机变量 X 的期望 $E[X] = \mu > 0$, 方差为 σ^2, 证明对任意 $\epsilon > 0$ 有

$$P(X - \mu \leqslant -\epsilon) \leqslant \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$

第 9 章　参数估计

设总体 X 的分布函数为 $F(X,\theta)$, 其中 θ 为未知参数(也可为向量). 现从总体中抽取一样本 $X_1, X_2, \cdots, X_n$, 如何依据样本估计参数 θ, 或 θ 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.

9.1 点估计

9.1.1 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 θ 的方法称为 **矩估计法**. 矩估计法的理论基础是大数定理: $X_1, X_2, \cdots, X_n$ 为 i.i.d. 的随机变量, 若 $E(X) = \mu$, 则当 $n \to \infty$ 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 $n \to \infty$ 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 $\theta_1, \theta_2, \cdots, \theta_m$,

1) 求总体 X 的 k 阶矩: $a_k = a_k(\theta_1, \theta_2, \cdots, \theta_m) = E[X^k],\ k \in [m]$ (a_k 一般为 $\theta_1, \theta_2, \cdots, \theta_m$ 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 $A_k = a_k = a_k(\theta_1, \theta_2, \cdots, \theta_m)$ $(k = 1, 2, \cdots, m)$, 得到 m 个关于 $\theta_1, \theta_2, \cdots, \theta_m$ 的方程组.

4) 求解方程组得到估计量 $\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_m$.

例 9.1　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 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, 求参数 α 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 9.2 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, 以及总体 X 的密度函数为

$$f(x) = \begin{cases} \frac{1}{\theta}e^{-\frac{x-\mu}{\theta}} & x \geqslant \mu \\ 0 & \text{其它}, \end{cases}$$

其中 $\theta > 0$, 求 μ 和 θ 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

$$E(Y) = \theta \quad \text{和} \quad \mathrm{Var}(Y) = \theta^2.$$

由此可得 $E(X) = \mu + \theta$ 和 $\mathrm{Var}(X) = \theta^2$. 计算对应的样本矩

$$A_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad B_2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

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 μ, σ^2 的矩估计法.
- 求总体 $X \sim \mathcal{U}(a, b)$ 中 a, b 的矩估计法.

9.1.2 最大似然估计法

设 $X_1, X_2, \cdots, X_n$ 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 $X_1, X_2, \cdots, X_n$ 的分布列为

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} \Pr(x_i; \theta).$$

这里 $L(\theta)$ 表示样本 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ 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 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ 的联合概率密度为

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

根据概率密度定义可知 $L(\theta)$ 越大, 样本 $(X_1, X_2, \cdots, X_n)$ 落入 $(x_1, x_2, \cdots, x_n)$ 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 $X_1, X_2, \cdots, X_n$ 的似然函数, 可以发现 $L(\theta)$ 是 θ 的函数, 若

$$\hat{\theta} = \arg\max_{\theta} L(x_1, x_2, \cdots, x_m; \theta),$$

则称 $\hat{\theta}$ 为 θ 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 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$ 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 $\log(L(x_1, x_2, \cdots, x_m; \theta)$;

ii) 求对数似然函数中参数 θ 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 9.3 设 $X_1, X_2, \cdots, X_n$ 是取自总体 $X \sim \mathrm{B}(1, p)$ 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 9.1 设 $\mu(\theta)$ 为 θ 的函数, 且存在反函数 $\theta=\theta(\mu)$. 若 $\hat{\theta}$ 是 θ 的最大似然估计, 则 $\hat{\mu}=\mu(\hat{\theta})$ 是 μ 的最大似然估计.

例 9.4 设 $X_1,X_2,\cdots,X_n$ 为总体 $X\sim\mathcal{N}(\mu,\sigma^2)$ 的样本, 求 μ 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 $X_1,X_2,\cdots,X_n$ 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 μ 求导计算可得

$$\frac{\partial\ln L(\mu,\sigma)}{\partial\mu}=\sum_{i=1}^{n}(X_i-\mu)=0\quad\Longrightarrow\quad\mu=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X},$$

对 σ 求导计算可得

$$\frac{\partial\ln L(\mu,\sigma)}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2=0\quad\Longrightarrow\quad\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

根据最大似然估计的不变性可知方差 σ^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 $X_1,X_2,\cdots,X_n$ 为总体 $X\sim\mathcal{N}(\mu,\nu)$ 的样本, 求 μ 和 ν 的最大似然估计. 根据题意可知样本 $X_1,X_2,\cdots,X_n$ 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 μ 求偏导计算其最大似然估计 $\mu=\sum_{i=1}^{n}X_i/n=\bar{X}$, 对 ν 求偏导计算可得

$$\frac{\partial\ln L(\mu,\nu)}{\partial\nu}=-\frac{n}{2\nu}+\frac{1}{2\nu^2}\sum_{i=1}^{n}(X_i-\mu)^2=0\quad\Longrightarrow\quad\nu=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

从而完成验证.

例 9.5 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 $X_1,X_2,\cdots,X_n$ 是总体 X 的样本, 求 α 的最大似然估计.

解 首先得到似然函数为

$$L(\alpha)=(\alpha+1)^n\prod_{i=1}^{n}X_i^{\alpha}=(\alpha+1)^n(X_1X_2\cdots X_n)^{\alpha},$$

以及其对数似然函数 $\ln L(\alpha)=n\ln(\alpha+1)+\alpha\ln(X_1X_2\cdots X_n)$. 求导并令偏导为零有

$$\frac{\partial\ln L(\alpha)}{\partial\alpha}=\frac{n}{\alpha+1}+\ln(X_1X_2\cdots X_n)=0,$$

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 9.6 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

$$L(a,b)=\begin{cases}\frac{1}{(b-a)^n} & a\leqslant X_1,X_2,\cdots,X_n\leqslant b\\ 0 & \text{其它}\end{cases}$$

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 $a\leqslant X_1,X_2,\cdots,X_n\leqslant b$, 因此最大似然估计量为:

$$b=\max\{X_1,X_2,\cdots,X_n\}\qquad\text{和}\qquad a=\min\{X_1,X_2,\cdots,X_n\}.$$

例 9.7 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

$$f(x)=\begin{cases}\theta e^{-(x-\mu)\theta} & x\geqslant\mu\\ 0 & \text{其它},\end{cases}$$

求 μ 和 θ 的最大似然估计.

解 首先计算似然函数为

$$L(\theta,\mu)=\begin{cases}\theta^n e^{-\theta\sum_{i=1}^{n}(X_i-\mu)} & X_i\geqslant\mu\\ 0 & \text{其它}\end{cases}$$

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

$$\frac{\partial \ln L(\theta,\mu)}{\partial \theta}=\frac{n}{\theta}-\sum_{i=1}^{n}(X_i-\mu)=0 \quad \Rightarrow \quad \theta=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)},$$

另一方面有

$$\frac{\partial \ln L(\theta,\mu)}{\partial \mu}=n\theta=0 \quad \Rightarrow \quad \theta=0,$$

此时无法求解 θ 和 μ 的最大似然估计. 回到似然函数的定义

$$L(\theta,\mu)=\begin{cases}\theta^n e^{-\theta\sum_{i=1}^{n}(X_i-\mu)} & X_1,X_2,\cdots,X_n\geqslant\mu\\ 0 & \text{其它}\end{cases}$$

可以发现 μ 越大似然函数 $L(\theta,\mu)$ 越大, 但须满足 $X_i\geqslant\mu$ $(i\in[n])$. 由此可得最大似然估计

$$\hat{\mu}=\min\{X_1,X_2,\cdots,X_n\},$$

进一步求解可得

$$\hat{\theta}=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})}.$$

9.2 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.

9.2.1 无偏性

定义 9.1 设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本, 令 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是 θ 的一个估计量, 若

$$E_{X_1,X_2,\ldots,X_n}\left[\hat{\theta}\right]=E_{X_1,X_2,\ldots,X_n}\left[\hat{\theta}(X_1,X_2,\cdots,X_n)\right]=\theta$$

则称 $\hat{\theta}$ 为 θ 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 θ, 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 9.8 (样本 k 阶原点矩为总体 k 阶原点矩的无偏估计) 设 $X_1,X_2,\cdots,X_n$ 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 9.9 设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本, 其期望为 μ, 方差为 σ^2, 则: 1) $S_0^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/n$ 是 σ^2 的有偏估计; 2) $S^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1)$ 是 σ^2 的无偏估计.

注意 $\hat{\theta}=\hat{\theta}(X_1,X_2,\ldots,X_n)$ 是 θ 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

$$E[\bar{X}]=E[X]=\mu \qquad \text{但} \qquad E[(\bar{X})^2]\neq\mu^2.$$

例 9.10 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

$$f(x;\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}} & x\geqslant 0\\ 0 & x<0,\end{cases}$$

证明: $\bar{X}=\sum_{i=1}^{n}X_i/n$ 和 $n\min\{X_1,X_2,\cdots,X_n\}$ 均是 θ 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 $Z=\min\{X_1,X_2,\cdots,X_n\}$, 则有

$$\begin{aligned}F_Z(z) &= \Pr[Z\leqslant z]=1-\Pr[Z>z]\\ &= 1-\Pr[X_1>z]\Pr[X_2>z]\cdots\Pr[X_n>z]\\ &= 1-\prod_{i=1}^{n}(1-\Pr[X_i\leqslant z])=\begin{cases}0 & z<0\\ 1-e^{-nz/\theta} & z\geqslant 0.\end{cases}\end{aligned}$$

于是当 $z\geqslant 0$ 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.

9.2.1.1 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 θ 的无偏估计, 则可以比较方差

$$\mathrm{Var}(\hat{\theta}_1)=E[(\hat{\theta}_1-\theta)^2] \qquad \text{和} \qquad \mathrm{Var}(\hat{\theta}_2)=E[(\hat{\theta}_2-\theta)^2].$$

一般而言: 方差越小, 无偏估计越好.

定义 9.2 设 $\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 和 $\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n)$ 是 θ 的两个无偏估计, 若

$$\mathrm{Var}(\hat{\theta}_1)\leqslant\mathrm{Var}(\hat{\theta}_2),$$

则称 θ_1 比 θ_2 有效.

例 9.11 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, 且 X 的概率密度为

$$f(x;\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x}{\theta}} & x\geqslant 0\\ 0 & x<0\end{cases},$$

令 $Z=\min\{X_1, X_2, \cdots, X_n\}$, 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

$$\mathrm{Var}(\bar{X})=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)=\frac{\theta^2}{n}.$$

根据例 9.10 可知随机变量 Z 的概率密度为

$$f(z)=\begin{cases}0 & z<0\\ \frac{n}{\theta}e^{-\frac{nz}{\theta}} & z\geqslant 0\end{cases}$$

从而得到

$$\mathrm{Var}(nZ)=n^2\mathrm{Var}(Z)=n^2\frac{\theta^2}{n^2}=\theta^2,$$

因此当 $n\geqslant 1$ 时有 $\mathrm{Var}(\bar{X})\leqslant\mathrm{Var}(nZ)$ 成立, 故 $\bar{X}$ 比 nZ 有效.

例 9.12 设 $X_1, X_2, \cdots, X_n$ 是总体 X 的样本, 且 $E(X)=\mu$ 和 $\mathrm{Var}(X)=\sigma^2$. 设常数 $c_1, c_2, \cdots, c_n\geqslant 0$ 满足 $\sum_{i=1}^{n}c_i=1, c_i\neq 1/n$, 求证: $\bar{X}$ 比 $\sum_{i=1}^{n}c_iX_i$ 有效.

证明 根据样本的独立同分布条件有

$$E\left[\bar{X}\right]=\mu \quad 和 \quad \mathrm{Var}\left(\bar{X}\right)=\sigma^2/n.$$

根据期望的性质有 $E[\sum_{i=1}^{n}c_iX_i]=\mu$, 进一步有

$$\mathrm{Var}\left(\sum_{i=1}^{n}c_iX_i\right)=\sum_{i=1}^{n}c_i^2\mathrm{Var}(X_i)=\sigma^2\sum_{i=1}^{n}c_i^2\geqslant\frac{\sigma^2}{n}$$

这里利用不等式 $\sum_{i=1}^{n}c_i^2/n\geqslant(\sum_{i=1}^{n}c_i/n)^2=1/n^2$, 所以有 $\mathrm{Var}(\sum_{i=1}^{n}c_iX_i)\geqslant\mathrm{Var}(\bar{X})$.

下面定义有效统计量:

定理 9.1 (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

$$\mathrm{Var}_0(\theta)=\frac{1}{nE\left[\left(\frac{\partial\ln f(X;\theta)}{\partial\theta}\right)^2\right]} \quad 或 \quad \mathrm{Var}_0(\theta)=\frac{1}{nE\left[\left(\frac{\partial\ln F(X;\theta)}{\partial\theta}\right)^2\right]},$$

对任意的无偏估计量 $\hat{\theta}$ 有

$$\mathrm{Var}(\hat{\theta}) \geqslant \mathrm{Var}_0(\theta),$$

称 $\mathrm{Var}_0(\theta)$ 为估计量 $\hat{\theta}$ 方差的下界. 当 $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}_0(\theta)$ 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 9.13 设 $X_1, X_2, \cdots, X_n$ 为总体 X 的样本, 令总体 X 的密度函数为

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & x \leqslant 0, \end{cases}$$

证明: θ 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

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{\theta^2}{n}.$$

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 $\mathrm{Var}_0(X) = \theta^2/n = \mathrm{Var}(\hat{\theta})$, 因此 θ 的最大似然估计是有效估计量.

9.2.1.2 一致性

定义 9.3 设 $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ 是 θ 的一个估计量, 若当 $n \to \infty$ 时有 $\hat{\theta}_n \xrightarrow{P} \theta$ 成立, 即对任意 $\epsilon > 0$ 有

$$\lim_{n \to 0} \Pr[|\hat{\theta}_n - \theta| > \epsilon] = 0,$$

则称 $\hat{\theta}_n$ 为 θ 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 θ, 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 9.2 设 $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ 是 θ 的一个估计量, 若满足以下两个条件:

$$\lim_{n \to \infty} E\left[\hat{\theta}_n \right] = \theta \quad \text{和} \quad \lim_{n \to \infty} \mathrm{Var}\left(\hat{\theta}_n \right) = 0,$$

则 $\hat{\theta}_n$ 为 θ 的一致估计量.

证明 根据 $\lim_{n\to\infty} E[\hat{\theta}_n] = \theta$ 知道对任意 $\epsilon > 0$, 存在一个N_0, 当 $n \geqslant N_0$ 有 $|E[\hat{\theta}_n] - \theta| \leqslant \theta/2$, 于是有

$$\lim_{n\to\infty} \Pr\left[|E[\hat{\theta}_n] - \theta| > \epsilon/2\right] = 0.$$

根据 Chebyshev 不等式有

$$\lim_{n\to 0} \Pr\left[\left|\hat{\theta}_n - E[\hat{\theta}_n]\right| > \epsilon/2\right] \leqslant \lim_{n\to 0} \frac{4}{\epsilon}\mathrm{Var}(\hat{\theta}_n) = 0$$

再根据

$$\Pr\left[|\hat{\theta}_n - \theta| > \epsilon\right] \leqslant \Pr\left[\left|\hat{\theta}_n - E[\hat{\theta}_n]\right| > \epsilon/2\right] + \Pr\left[\left|E[\hat{\theta}_n] - \theta\right| > \epsilon/2\right]$$

完成证明.

定理 9.3 设 $\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \cdots, \hat{\theta}_{n_k}$ 分别为 $\theta_1, \theta_2, \cdots, \theta_k$ 满足一致性的估计量, 对连续函数 $g\colon \mathbb{R}^n \to \mathbb{R}$, 有函数 $\hat{\eta}_n = g(\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \cdots, \hat{\theta}_{n_k})$ 是 $\eta = g(\theta_1, \theta_2, \cdots, \theta_k)$ 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 9.14 设$X_1, X_2, \cdots, X_n$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 θ 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

$$\lim_{n\to\infty} Var(\bar{X}) = \lim_{n\to\infty} \frac{\theta^2}{n} = 0.$$

例 9.15 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim U(0,\theta)$ 的样本, 证明: θ 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 θ 的最大似然估计为 $\hat{\theta}_n = \max(X_1, X_2, \cdots, X_n)$. 设随机变量 $Z = \max(X_1, X_2, \cdots, X_n)$, 则由 Z 的分布函数

$$F_Z(z) = \Pr[Z \leqslant z] = \Pr[\max(X_1, X_2, \cdots, X_n) \leqslant z] = \prod_{i=1}^n \Pr[X_i \leqslant z] = \begin{cases} 1 & z > \theta \\ (\frac{z}{\theta})^n & z \in [0,\theta] \\ 0 & z < 0. \end{cases}$$

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

$$E\left[\hat{\theta}_n\right] = E[Z] = \int_0^\theta \frac{nz^n}{\theta^n} dz = \frac{n}{n+1}\theta,$$

因此 $\hat{\theta}$ 是 θ 的有偏估计. 另一方面有

$$E\left[Z^2\right] = \int_0^\theta \frac{nz^{n+1}}{\theta^n} dz = \frac{n}{n+2}\theta^2,$$

从而得到

$$\mathrm{Var}\left(\hat{\theta}_n\right) = \mathrm{Var}(Z) = E[Z^2] - (E[Z])^2 = \frac{n}{n+2}\theta^2 - (\frac{n\theta}{n+1})^2 = \frac{n}{(n+1)^2(n+2)}\theta^2,$$

于是有

$$\lim_{n\to\infty} E[\hat{\theta}_n] = \theta \quad \text{和} \quad \lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0,$$

由此可得 $\hat{\theta}$ 是 θ 的有偏、但一致估计量.

9.3 区间估计

区间估计问题: 设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本, θ 为总体 X 的分布函数 $F(x,\theta)$ 的未知参数, 根据样本估计 θ 的范围 $(\hat{\theta}_1, \hat{\theta}_2)$, 其中 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \cdots, X_n)$, 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

$$\Pr\left[\hat{\theta}_1(X_1, X_2, \cdots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \cdots, X_n)\right] \geqslant 1-\alpha.$$

定义 9.4 (置信区间与置信度) 设 $X_1, X_2 \cdots, X_n$ 是来自总体 X 的样本, 总体 X 的分布函数含未知参数 θ, 找出统计量 $\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和 $\hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \cdots, X_n)$ $(\hat{\theta}_1 < \hat{\theta}_2)$, 使得

$$\Pr\left[\hat{\theta}_1 < \theta < \hat{\theta}_2\right] \geqslant 1-\alpha$$

成立, 则称 $1-\alpha$ 为置信度, $[\hat{\theta}_1, \hat{\theta}_2]$ 为 θ 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 θ 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) α 反映了估计的可靠度, α 越小可靠度越高.

iii) 给定 α, 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 $W(X_1, X_2, \cdots, X_n; \theta)$ 包含待估参数 θ, 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 θ 的置信度为 $1-\alpha$ 的置信区间.

9.3.1 正态总体, 方差已知, 求期望的区间估计

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 若方差 σ^2 已知. 给定 $\alpha \in (0,1)$, 确定置信度为 $1-\alpha$ 下 μ 的置信区间 $[\hat{\theta}_1, \hat{\theta}_2]$. 令样本均值为 $\bar{X} = \sum_{i=1}^{n} X_i/n$, 根据正太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

$$\Pr[W \geqslant \mu_{\alpha/2}] = 1-\alpha/2 \quad \text{和} \quad \Pr[W \leqslant -\mu_{\alpha/2}] = 1-\alpha/2.$$

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 9.16 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 μ 的置信区间 $(\mu_{0.025} = 1.96)$.

9.3.2 正态总体, 方差未知, 求期望的区间估计

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 若方差 σ^2 未知, 考虑期望 μ 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

$$\Pr[a \leqslant W \leqslant b] = 1-\alpha \quad \Rightarrow \quad b = t_{\alpha/2}(n-1),\ a = -t_{\alpha/2}(n-1).$$

整理可得

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

9.3.3 正态总体, 求方差 σ^2 的置信区间

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 考虑方差 σ^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

$$\Pr[a \leqslant W \leqslant b] = 1-\alpha.$$

根据 χ^2 分布的不对称性, 采用概率对称的区间

$$\Pr[W \leqslant a] = \Pr[\geqslant b] = \alpha/2 \quad \Rightarrow \quad b = \chi^2_{\alpha/2}(n-1), \quad a = \chi^2_{1-\alpha/2}(n-1).$$

根据枢轴变量 $W = (n-1)S^2/\sigma^2$ 可得

$$\Pr\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right] = 1-\alpha.$$

9.3.4 双正态总体情形

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 的样本, 设 $Y_1, Y_2, \cdots, Y_m$ 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 σ_1^2/σ_2^2 的置信度为 $1-\alpha$ 的区间估计.

1) 已知方差 σ_1^2 和 σ_2^2, 求 $\mu_1 - \mu_2$ 的置信区间. 根据正态分布的性质有

$$\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n}), \quad \bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{m}) \quad \bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}),$$

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 σ_1^2 和 σ_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 σ_1^2/σ_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

$$\Pr[W\leqslant a]=\Pr[W\geqslant b]=\alpha/2 \quad \Rightarrow \quad b=F_{\frac{\alpha}{2}}(n-1,m-1), \quad a=F_{1-\alpha/2}(n-1,m-1).$$

由此可得置信区间

$$\Pr\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n-1,m-1)}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n-1,m-1)}\right]=1-\alpha.$$

9.3.5 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 9.5 (单侧置信区间) 给定 $\alpha \in (0,1)$, 若样本 $X_1,\cdots,X_n$ 的统计量 $\hat{\theta}_1 = \hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 满足

$$\Pr\left[\theta > \hat{\theta}_1\right] \geqslant 1-\alpha,$$

则称 $(\hat{\theta}_1,+\infty)$ 为 θ 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 9.17 设 $X_1,X_2,\cdots,X_n$ 是来自总体 $X\sim\mathcal{N}(\mu,\sigma^2)$ 的样本, 若方差 σ^2 已知, 求 μ 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 9.18 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

$$\bar{X}=\sum_{i=1}^{10}X_i/10=1090\quad 和\quad S^2=\sum_{i=1}^{10}(X_i-\bar{X})^2/9=8800/3.$$

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

$$\mu>\bar{X}-t_{0.05}(9)S/3=1090-\sqrt{8800/3}\times 1.833/3>1056.$$

9.3.6 非正态分布的区间估计

设 $X_1,X_2,\cdots,X_n$ 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

$$\Pr\left[|\mu - \bar{X}| \geqslant \epsilon\right] \leqslant 2\exp(-2n\epsilon^2/(b-a)^2).$$

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 $\mathrm{Var}(X) = \sigma^2$, 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 σ^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 σ^2 未时, 用修正样本方差 $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ 代替方差 σ^2, 于是有

$$\Pr\left[-\mu_{\alpha/2} < \frac{\bar{X} - \mu}{S^2/\sqrt{n}} < \mu_{\alpha/2}\right] \approx 1-\alpha.$$

例 9.19 设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim \mathrm{Ber}(p)$ 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

$$\Pr[|\bar{X} - p| > \epsilon p] \leqslant 2\exp(-np\epsilon^2/3),$$

设 $\alpha = 2\exp(-np\epsilon^2/3)$, 于是有

$$\Pr\left[\bar{X} - \sqrt{3p\ln(2/\alpha)/n} < p < \bar{X} + \sqrt{3p\ln(2/\alpha)/n}\right] \geqslant 1-\alpha,$$

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 $\mathrm{Var}(X) = p(1-p)$, 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 10 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 10.1　某产品出厂检验规定次品率 $p \leqslant 0.04$ 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 $H_0 : p \leqslant 0.04$. 若假设 H_0 成立, 设随机变量 $X \sim \mathrm{B}(12, p)$,

$$\Pr[X = 3] = \binom{12}{3} p^3 (1-p)^9 \leqslant 0.0097.$$

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 $H_0 : p \leqslant 0.04$ 不成立, 即 $p > 0.04$, 该批产品不能出厂.

若 $X = 1$ 则

$$\Pr[X = 1] = p(1-p)^{11} \binom{12}{1} \geqslant 0.306.$$

这不是小概率事件, 没理由拒绝原假设H, 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 α, 当一事件发生的概率小于 α 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的现象 (概率小于 α 的事件发生), 则认为假设 H_0 不显著, α 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 10.2 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

$$\Pr\left[|Z|>1.96\right]=0.05$$

  成为一个小事件, 从而得到否定域 $\{Z\colon |Z|>1.96\}$.

- 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 $\{Z\colon |Z|>1.96\}$, 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 α, 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 $H_0:\mu\leqslant\mu_0$ 和备选假设 $H_1:\mu>\mu_0$, 称为 **右边检验**;
- 原假设 $H_0:\mu\geqslant\mu_0$ 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 α, 即显著性水平, 第 II 类错误的概率用 β 表示, 即

$$\alpha = \Pr\left[\text{拒绝 } H_0 | H_0 \text{ 为真}\right] \qquad \beta = \Pr\left[\text{接受 } H_0 | H_0 \text{ 为假}\right].$$

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. Neymam-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.

10.1 正态总体期望的假设检验

10.1.1 方差已知的单个正态总体的期望检验 (Z 检验)

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim N(\mu, \sigma^2)$ 的样本, 若方差 σ^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 α, 得到拒绝域为 $|Z| \geqslant \mu_{\alpha/2}$, 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

i) 原假设 $H_0: \mu = \mu_0$ 和备择假设 $H_1: \mu \neq \mu_0$, 拒绝域为 $\{Z : |Z| \geqslant \mu_{\alpha/2}\}$;

ii) 原假设 $H_0: \mu \geqslant \mu_0$ 和备择假设 $H_1: \mu < \mu_0$, 拒绝域为 $\{Z : Z \leqslant -\mu_\alpha\}$;

iii) 原假设 $H_0: \mu \leqslant \mu_0$ 和备择假设 $H_1: \mu > \mu_0$, 拒绝域为 $\{Z : Z \geqslant \mu_\alpha\}$.

例 10.3 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 $|Z| \geqslant \mu_{\alpha/2} = 1.96$. 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 10.4 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 $H_0: \mu \geqslant 1000$ 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.

10.1.2 方差未知的单个正态总体的期望检验 (t 检验)

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim N(\mu, \sigma^2)$ 的样本, 若方差 σ^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 α, 得到拒绝域为 $|t| \geqslant t_{\alpha/2}(n-1)$, 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 $\{t : |t| \geqslant t_{\alpha/2}(n-1)\}$;

ii) 原假设 $H_0: \mu \geqslant \mu_0$ 和备择假设 $H_1: \mu < \mu_0$, 拒绝域为 $\{t : t \leqslant -t_\alpha(n-1)\}$;

iii) 原假设 $H_0: \mu \leqslant \mu_0$ 和备择假设 $H_1: \mu > \mu_0$, 拒绝域为 $\{t : t \geqslant t_\alpha(n-1)\}$.

10.1.3 方差已知的两个正态总体的期望差检验

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim N(\mu_1, \sigma_1^2)$ 的样本, 以及 $Y_1, Y_2, \cdots, Y_m$ 是来自总体 $Y \sim N(\mu_2, \sigma_2^2)$ 的样本, 若方差 σ_1^2 和 σ_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: δ 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 α, 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 $\{U:|U|\geqslant\mu_{\alpha/2}\}$;

ii) 原假设 $H_0:\ \mu_1-\mu_2\geqslant\delta$ 和备择假设 $H_1:\ \mu_1-\mu_2<\delta$, 拒绝域为 $\{U:U\leqslant-\mu_\alpha\}$;

iii) 原假设 $H_0:\ \mu_1-\mu_2\leqslant\delta$ 和备择假设 $H_1:\ \mu_1-\mu_2>\delta$, 拒绝域为 $\{U:U\geqslant\mu_\alpha\}$.

10.1.4 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

10.1.5 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 10.5 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 $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$, 其中 $X_1,X_2,\ldots,X_n$ 和 $Y_1,Y_2,\ldots,Y_n$ 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 $H_1:\ \mu\neq0$, 方差 σ^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

$$\bar{Z}=\sum_{i=1}^{n}\frac{Z_i}{n}\qquad\text{和}\qquad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})^2.$$

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 α 下得到拒绝域为: $|t|>t_{\alpha/2}(n-1)$. 下面给出例 10.5 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.

10.2 正态分布的方差假设检验.

10.2.1 单个正态总体的方差检验 (χ^2 检验)

设 $X_1, X_2, \cdots, X_n$ 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

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

给定显著性水平 α 求解拒绝域, 这种检验方法称为 χ^2 **检验法**.

关于 χ^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: $\{\chi^2 \geqslant \chi^2_{\frac{\alpha}{2}}(n-1)\} \cup \{\chi^2 \leqslant \chi^2_{1-\frac{\alpha}{2}}(n-1)\}$.

ii) 原假设 $H_0:\ \sigma^2 \geqslant \sigma_0^2$ 和备择假设 $H_1:\ \sigma^2 < \sigma_0^2$. 拒绝域为: $\{\chi^2 \leqslant \chi^2_{1-\alpha}(n-1)\}$.

iii) 原假设 $H_0:\ \sigma^2 \leqslant \sigma_0^2$ 和备择假设 $H_1:\ \sigma^2 > \sigma_0^2$. 拒绝域为: $\{\chi^2 \geqslant \chi^2_{\alpha}(n-1)\}$.

10.2.2 两个正态总体的方差比检验 (F 检验)

略

10.3 非参假设检验

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.

10.3.1 χ^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 $X_1, \ldots, X_n$ 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: $H_0 : \Pr[X = x_i] = p_i\ \ (i = 1, 2, ...)$

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 χ^2 检验法: 将随机试验结果的全体 Ω 分成 k 个互不相容的事件 $A_1, A_2, \ldots, A_k$, 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 $X_1, \ldots, X_n$, 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 χ^2 统计量.

定理 10.1 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 α, 若 $W > \chi^2_\alpha(k-1)$ 则拒绝 H_0.

例 10.6 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	$\geqslant 5$
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

$$p_1 = P(X=1) = \frac{1}{4} \quad p_2 = P(X=2) = \frac{3}{4} \times \frac{1}{4} \quad p_3 = P(X=3) = (\frac{3}{4})^2 \cdot \frac{1}{4}$$

$$p_4 = P(X=4) = (\frac{3}{4})^3 \cdot \frac{1}{4} \quad p_5 = P(X=5) = 1 - \frac{1}{4} - \frac{3}{16} - (\frac{3}{4})^3 \cdot \frac{1}{4}$$

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 $W > \mathcal{X}^2_{0.05}(4) = 9.488$, 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

$$H_0 : F(x) = F_0(x; \theta_1, \theta_2, \ldots, \theta_r)$$

其中 F_0 已知, $\theta_1, \theta_2, \ldots, \theta_r$ 未知. 从样本 $X_1, X_2, \ldots, X_n$ 中得到估计值 $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r$, 代入得

$$H_0 : F(x) = F_0(x; \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r)$$

将子样分成 k 组: $a_0 < a_1 < \ldots < a_k$ 且 $A_1 \in [a_0, a_1], A_2 \in [a_1, a_2] \ldots A_k = [a_{k-1}, a_k]$. 总体 X 落入 A_i 的概率为

$$\hat{p_i} = p(x \in A_i | \hat{\theta_1} \ldots \hat{\theta_r})$$

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 10.2 当 $n\to+\infty$ 时, 有 $W\xrightarrow{d}\chi^2(k-r-1)$ 成立.

10.3.1.1 独立性检验

设 $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$ 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 $A_1,A_2,\cdots,A_r$ 和 $B_1,B_2,\cdots,B_s$. 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

$$p_{ij}=\Pr(X\in A_i,Y\in B_j)\qquad p_{i\cdot}=P(X\in A_i)=\sum_{j=1}^{s}p_{ij}\qquad p_{\cdot j}=P(Y\in B_j)=\sum_{i=1}^{r}p_{ij}$$

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

$$\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n},\qquad \hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n}.$$

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 α 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

10.1 设随机变量 X 的期望 $E[X] = \mu > 0$, 方差为 σ^2, 证明对任意 $\epsilon > 0$ 有

$$P(X - \mu \leqslant -\epsilon) \leqslant \frac{\sigma^2}{\sigma^2 + \epsilon^2}.$$