

# 概率论与数理统计

授课教师：唐宏岩

# 前言

本讲义基于清华大学数学系唐宏岩老师于 2023 - 2024 学年秋季学期开设的《概率论与数理统计》课程，用于辅助同学们课后复习，助教尽量做到每周课后两天内更新。

由于时间与能力所限，本讲义可能不会出现大段的文字论述（但会包含重要的定义、定理与公式等）。但是，对许多基本概念的深入理解是非常有必要的，同学们可以在浏览时检查自己是否能够回忆起课上的内容，对掌握不够扎实的地方，鼓励大家查阅参考书或在课程群提问以解决问题。

由于此为教学团队第一年尝试整理讲义，诸如格式编排、内容完整性方面可能存在许多不足，欢迎大家联系我提出宝贵的意见与建议。

曹子尧

2023 年 9 月

# 目录

|              |    |
|--------------|----|
| 前言           | i  |
| 第一部分 初等概率论   | 1  |
| 第一章 事件的概率    | 2  |
| 1.1 概率的发展史   | 2  |
| 1.2 随机试验与事件  | 2  |
| 1.3 事件的运算    | 3  |
| 1.4 概率的几种解释  | 3  |
| 1.5 概率的公理化定义 | 3  |
| 1.6 条件概率     | 4  |
| 1.7 事件的独立性   | 5  |
| 1.8 Bayes 公式 | 6  |
| 第二章 随机变量     | 7  |
| 2.1 一维随机变量   | 7  |
| 2.2 离散随机变量   | 9  |
| 2.3 常见离散分布   | 10 |
| 2.4 连续随机变量   | 11 |
| 2.5 常见连续分布   | 11 |
| 2.6 随机变量的函数  | 13 |
| 第三章 联合分布     | 15 |
| 3.1 随机向量     | 15 |
| 3.2 离散分布     | 15 |
| 3.3 连续分布     | 16 |

|             |                  |           |
|-------------|------------------|-----------|
| 3.4         | 边际分布             | 16        |
| 3.5         | 条件分布             | 17        |
| 3.6         | 独立性              | 17        |
| 3.7         | 随机向量的函数          | 18        |
| <b>第四章</b>  | <b>随机变量的数字特征</b> | <b>21</b> |
| 4.1         | 期望               | 21        |
| 4.2         | 分位数              | 21        |
| 4.3         | 方差               | 22        |
| 4.4         | 协方差与相关系数         | 22        |
| 4.5         | 矩                | 23        |
| 4.6         | 矩母函数             | 24        |
| 4.7         | 条件期望             | 26        |
| <b>第五章</b>  | <b>不等式与极限定理</b>  | <b>28</b> |
| 5.1         | 概率不等式            | 28        |
| 5.2         | 大数定律             | 29        |
| 5.3         | 中心极限定理           | 30        |
| <b>第二部分</b> | <b>统计推断</b>      | <b>32</b> |
| <b>第六章</b>  | <b>参数估计</b>      | <b>35</b> |
| 6.1         | 矩估计              | 35        |
| 6.2         | 极大似然估计           | 35        |
| 6.3         | 优良性准则            | 36        |
| 6.4         | 置信区间             | 38        |
| 6.5         | Bayes 估计         | 41        |

# 第一部分

## 初等概率论

# 第一章 事件的概率

## 1.1 概率的发展史

赌博中的 de Méré's Problem: 连续掷一个均匀六面骰 4 次, 获得至少一次 “6” 的概率为  $1 - (\frac{5}{6})^4 \approx 0.5177$ ; 而连续掷两个均匀六面骰 24 次, 获得至少一次 “对 6” 的概率为  $1 - (35/36)^{24} \approx 0.4914$ 。

Pascal 和 Fermat 的通信中使用初等数学的方法, 首创了概率论相当多的数学理论, 虽然当时没有总结成通用的定理。

Laplace 创立了采用分析方法的分析概率论。

Kolmogorov 利用测度论方法发展了现代概率理论。

## 1.2 随机试验与事件

**定义 1.1.** 概率论中的随机试验指的是符合下面两个特点的试验:

1. 不能预先确知结果
2. 可以预测所有可能的结果

**定义 1.2.** 样本空间是指一个试验的所有可能结果的集合, 常用  $\Omega$  表示。

**定义 1.3.** 事件是样本空间的一个良定义子集。

一次随机试验中, 一个事件可能发生或不发生。

下面是一些常见的事件:

1. 全事件  $\Omega$  (必然事件)
2. 空事件  $\emptyset$  (不可能事件)
3. 基本事件  $\{a\}$ , 其中  $a \in \Omega$ , 即仅包含单一试验结果的事件

## 1.3 事件的运算

由于事件是集合，因此事件之间可以进行集合之间的运算，如：

1. 余  $A^c = \Omega \setminus A$
2. 和  $A + B = A \cup B = (A^c \cap B^c)^c$
3. 差  $A - B = A \setminus B$
4. 积  $AB = A \cap B = (A^c \cup B^c)^c$

集合的 De Morgan's laws 也适用于事件： $(\bigcup_n A_n)^c = \bigcap_n A_n^c$ 。

事件的运算像集合的运算一样，可以用 Venn 图来表示。

## 1.4 概率的几种解释

对于概率这一数学概念，人们形成了几种从不同角度出发的解释：

1. 古典解释：基于等可能性的解释
2. 频率解释：基于大量重复试验的解释（频率学派采用的解释）
3. 主观解释：概率是一种对确信程度的度量（Bayes 学派采用的解释）

## 1.5 概率的公理化定义

我们用  $2^\Omega$  表示  $\Omega$  的幂集，即  $\Omega$  的所有子集组成的集合。

**定义 1.4.** 事件集类  $\mathcal{F} \subset 2^\Omega$  必须满足所谓  $\sigma$ -代数的性质：

1.  $\Omega \in \mathcal{F}$
2.  $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ （对补运算的封闭性）
3.  $A_i \in \mathcal{F}, \forall i \in \mathbb{N}^* \Rightarrow \bigcup_{i=1}^\infty A_i \in \mathcal{F}$ （对可列并的封闭性）

**例 1.1.**  $\Omega = \{a, b, c, d\}$ ，以下是一些合法的事件集类：

1.  $\mathcal{F}_1 = 2^\Omega$
2.  $\mathcal{F}_2 = \{\Omega, \emptyset\}$
3.  $\mathcal{F}_3 = \{\Omega, \emptyset, \{a, b\}, \{c, d\}\}$

**定义 1.5.** (Kolmogorov) 概率函数  $P: \mathcal{F} \rightarrow \mathbb{R}$  是满足以下三条公理的映射：

1.  $P(A) \geq 0, \forall A \in \mathcal{F}$
2.  $P(\Omega) = 1$
3.  $A_i \in \mathcal{F}, \forall i \in \mathbb{N}^*, A_i A_j = \emptyset, \forall i \neq j \Rightarrow P(\sum_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$ （加法公理/可列可加性）

我们称  $(\Omega, \mathcal{F}, P)$  是一个概率空间。

**命题 1.1.** 关于概率空间, 有如下性质:

1.  $P(A) \leq 1, \forall A \in \mathcal{F}$
2.  $P(\emptyset) = 0$
3.  $P(A) + P(A^c) = 1$
4.  $A_i \in \mathcal{F}, \forall i \in \{1, 2, \dots, n\}, A_i A_j = \emptyset, \forall i \neq j \Rightarrow P(\sum_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$  (有限可加性)
5.  $A \subset B \Rightarrow P(A) \leq P(B)$  (我们称事件  $A$  蕴含事件  $B$ )
6.  $P(A_1 + \dots + A_n) = \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) + \dots + (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} A_{i_2} \dots A_{i_r}) + \dots + (-1)^{n+1} P(A_1 \dots A_n)$  (容斥公式)

特别地,  $P(A + B) = P(A) + P(B) - P(AB)$ 。

**例 1.2.** (配对问题)

有  $n$  个人, 每人有一顶帽子。现将所有帽子放到一起, 再随机分配给每人一顶, 考虑无人拿到自己的帽子的概率。

为此, 设事件  $A_i$  为“第  $i$  个人拿到自己的帽子”, 则  $P(A_i) = 1/n$ 。

利用容斥公式, 至少一人拿到自己帽子的概率为  $P(A_1 + \dots + A_n) = \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) + \dots + (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} A_{i_2} \dots A_{i_r}) + \dots + (-1)^{n+1} P(A_1 \dots A_n)$ ,

其中  $\sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} A_{i_2} \dots A_{i_r}) = \frac{(n-r)!}{n!} \binom{n}{r} = \frac{1}{r!}$ , 即  $P(A_1 + \dots + A_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots + (-1)^{r+1} \frac{1}{r!} + \dots + (-1)^{n+1} \frac{1}{n!}$ 。

所求概率  $P_n = 1 - P(A_1 + \dots + A_n) = 1 - (1 - \frac{1}{2!} + \dots + (-1)^{n+1} \frac{1}{n!}) \rightarrow e^{-1} (n \rightarrow +\infty)$ 。

思考: 恰有  $k$  个人拿到自己的帽子的概率?

## 1.6 条件概率

**定义 1.6.** 若  $P(B) > 0$ , 定义条件概率  $P(A|B) = \frac{P(AB)}{P(B)}$ 。

通常, 我们计算条件概率的方法有两种:

1. 在缩小 (受限) 的样本空间 (要求事件  $B$  发生) 上, 考虑事件  $A$  发生的概率
2. 根据定义计算

一种常用的形式是  $P(AB) = P(A|B)P(B) = P(B|A)P(A)$ , 这可以视作是求解两个事件的积的概率的方法 (乘法法则)。

**例 1.3.** 掷一个均匀六面骰,  $\Omega = \{1, 2, 3, 4, 5, 6\}, A = \{2, 3, 4, 5\}, B = \{1, 3, 5\}$ , 则  $P(A) = 4/6, P(B) = 3/6, P(AB) = 2/6, P(A|B) = \frac{P(AB)}{P(B)} = 2/3$ 。



**例 1.4.** 袋子中有 8 个红球和 4 个白球，无放回地取出两个球，利用组合数可知，两个都是红球的概率为  $\frac{\binom{8}{2}}{\binom{12}{2}}$ 。

用条件概率可以简化计算： $P(R_1 R_2) = P(R_1)P(R_2|R_1) = \frac{8}{12} \times \frac{7}{11}$ 。

更一般地，我们有  $P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2) \cdots P(A_n|A_1 A_2 \cdots A_{n-1})$ ，常用于序贯发生的一系列事件的积的概率求解。

**例 1.5.** 回忆上一节的“配对问题”。我们有  $P(A_{i_1} A_{i_2} \cdots A_{i_r}) = P(A_{i_1})P(A_{i_2}|A_{i_1}) \cdots P(A_{i_r}|A_{i_1} \cdots A_{i_{r-1}}) = \frac{1}{n} \times \frac{1}{n-1} \times \cdots \times \frac{1}{n-(r-1)} = \frac{(n-r)!}{n!}$ 。

**命题 1.2.** 对于给定的事件  $B$ ， $P(\cdot|B) : \mathcal{F} \rightarrow \mathbb{R}$  是概率函数，即  $(\Omega, \mathcal{F}, P(\cdot|B))$  仍是概率空间。

对于上述命题的证明，只需验证  $P(\cdot|B)$  满足概率的三条公理即可。

这提示我们，条件概率也是一种概率，如果我们将  $P(A)$  称为观察到事件  $B$  之前  $A$  的“先验概率”，则  $P(A|B)$  就是相应的“后验概率”。

一个常见的迷思是：观测到事件  $A$  已经发生后，是否可以说事件  $A$  发生的概率  $P(A) = 1$ ？学过条件概率之后，我们知道答案是否定的，实际上是后验概率  $P(A|A) = 1$ 。

## 1.7 事件的独立性

**定义 1.7.** 若  $P(AB) = P(A)P(B)$ ，则称事件  $A, B$  相互独立。

如果  $P(B) > 0$ ，我们注意到  $A, B$  独立等价于  $P(A|B) = P(A)$ 。

**命题 1.3.** 若  $A, B$  独立，则  $A^c, B$  独立。

**定义 1.8.** 若  $P(ABC) = P(A)P(B)P(C)$ ，且  $A, B, C$  两两独立，则称事件  $A, B, C$  独立。

注意，仅有  $A, B, C$  两两独立，不能推出三者独立。

**定义 1.9.** 若对于事件列  $\{A_i\}_{i=1}^\infty$ ，任意取有限个事件  $A_{i_1}, A_{i_2}, \dots, A_{i_r}$ ，都有  $P(A_{i_1} A_{i_2} \cdots A_{i_r}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_r})$ ，则称  $\{A_i\}_{i=1}^\infty$  相互独立。

**例 1.6.** 每周开奖的彩票，各次中奖率均为  $10^{-5}$  且独立，问连续十年（520 周）不中奖的概率？令事件  $A_i$  为第  $i$  周不中奖，则  $P(A_i) = 1 - 10^{-5}$ ，故  $P(A_1 \cdots A_{520}) = (1 - 10^{-5})^{520} \approx 0.9948$ 。

**定义 1.10.** 若事件  $A, B, E$  满足  $P(AB|E) = P(A|E)P(B|E)$ ，则我们称  $A, B$  关于  $E$  条件独立。

注意，条件独立性和独立性之间没有蕴含关系。

## 1.8 Bayes 公式

### 定理 1.1. (全概率公式)

设  $\{B_i\}$  是  $\Omega$  的一个分割, 即

1.  $\sum_i B_i = \Omega$
2.  $B_i B_j = \emptyset, \forall i \neq j$
3.  $P(B_i) > 0, \forall i$

则  $P(A) = P(\sum_i (AB_i)) = \sum_i P(AB_i) = \sum_i P(A|B_i)P(B_i)$ 。

注:  $\{B_i\}$  可以是有限集合, 或可数无穷集合。

**例 1.7.** 对于调查问卷中的敏感问题 (如 “你是否有过某病史”), 被调查者可能会有所顾虑而做出虚假的回答。为保护被调查者的隐私, 同时取得其信任, 考虑引入一个 “保护性问题”, 即不具有敏感性的问题 (如 “你是否会游泳”), 并让被调查者以抛硬币的方式, 随机抽取一个问题回答。这样, 抽到敏感问题的、确有过该病史的被调查者在回答 “是” 时也无须有病史暴露之虞。

设人群中, 敏感问题答案为 “是” 的比例为  $p$  (未知), 保护性问题答案为 “是” 的比例为  $q$  (假设已知), 则若收集到  $n$  个被调查者的结果, 其中  $k$  个为 “是”, 我们便有  $\frac{1}{2}p + \frac{1}{2}q \approx \frac{k}{n}$ , 可以据此得到  $p$  的估计。

### 定理 1.2. (Bayes 公式 / Bayes 准则)

设  $\{B_i\}$  是  $\Omega$  的一个分割, 则  $P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)}$ 。

### 例 1.8. (假阳性悖论)

对于一种流行病,  $A$  表示一个人检查呈阳性,  $B$  表示此人确实患病。设  $P(B) = 10^{-4}$ ,  $P(A|B) = 0.99$ ,  $P(A|B^c) = 10^{-3}$ , 则一个检查呈阳性的人真的患病的概率仅为  $P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \approx 9\%$ 。

如果再次检测仍呈阳性, 且两次检测效率不变, 结果彼此独立, 则此人真的患病的概率为

$$P(B|A_1 A_2) = \frac{P(A_1 A_2|B)P(B)}{P(A_1 A_2|B)P(B) + P(A_1 A_2|B^c)P(B^c)} = \frac{P(A_1|B)P(A_2|B)P(B)}{P(A_1|B)P(A_2|B)P(B) + P(A_1|B^c)P(A_2|B^c)P(B^c)} \approx 99\%。$$

## 第二章 随机变量

### 2.1 一维随机变量

**定义 2.1.** 随机变量是样本空间上的实值函数。

注意，上述定义是不严格的。

更严谨的定义：若对于可测空间  $(\Omega, \mathcal{F})$  和函数  $X : \Omega \rightarrow \mathbb{R}$ ，有  $\forall x \in \mathbb{R}, \{\omega | X(\omega) \leq x\} \in \mathcal{F}$ ，则称  $X$  是  $(\Omega, \mathcal{F})$  上的随机变量。其中“可测空间”是指  $\mathcal{F}$  是样本空间  $\Omega$  上的  $\sigma$ -代数。此处不要求“概率空间”，即随机变量的定义并不依赖概率测度  $P$  的存在。

**例 2.1.** 下表展示了两个随机变量。其中“像集”即  $\{X(\omega) | \omega \in \Omega\}$ 。

| 试验                | 样本空间 $\Omega$                  | 随机变量 $X$              | 像集                    |
|-------------------|--------------------------------|-----------------------|-----------------------|
| 随机调查 50 人对某议题支持与否 | $\Omega_1 = \{0, 1\}^{50}$     | $X_1 = \text{“1”的个数}$ | $\{0, 1, \dots, 50\}$ |
| 随机抽取一名北京成年市民      | $\Omega_2 = \text{所有北京成年市民之集}$ | $X_2 = \text{其年收入}$   | $\mathbb{R}$          |

注意，我们经常用“ $X_1 = 20$ ”、“ $X_2 > 100000$ ”等简化的记号来表示事件。例如，前者实际上指的是  $\{\omega \in \Omega_1 | X_1(\omega) = 20\}$ 。

诸如此类的试验结果集合需是事件，这体现出前述的随机变量严谨定义的意义。事实上，如果满足该严谨定义，则对于任意可测集  $I \subset \mathbb{R}$ ，都有  $\{\omega \in \Omega | X(\omega) \in I\} \in \mathcal{F}$ 。

随机变量是试验结果的数值摘要，起到一种概括的作用。随机变量的“随机”要素来自于样本点  $\omega \in \Omega$  的随机选择。在实际应用中，随机变量常常比样本空间具有更直观的意义。

随机变量可以分为：

1. 离散型：至多可数多个取值
2. 连续型：区间型取值（非严格定义）
3. 其他

“其他”中的一个非常特殊的子类是所谓的混合型随机变量。

**定义 2.2.** 对于随机变量  $X$  和  $\mathbb{R}$  的可测子集  $I$  (例如  $I = (a, b]$ ), 令  $X^{-1}(I) = \{\omega \in \Omega | X(\omega) \in I\} \subset \Omega$  为  $I$  的原像集, 我们定义记号  $P(X \in I)$  表示 “ $X$  的取值在  $I$  中的概率”, 其值为  $P(X^{-1}(I))$ 。

例如,  $P(a < X \leq b) = P(\{\omega | X(\omega) \in (a, b]\})$ 。

**定义 2.3.**  $F_X(x) = P(X \leq x), \forall x \in \mathbb{R}$  称为随机变量  $X$  的累积分布函数 (Cumulative Distribution Function, CDF)。下标  $X$  在无歧义时可省略。

我们有  $P(a < X \leq b) = F(b) - F(a)$ 。

**例 2.2.** 令  $X$  表示掷两个均匀六面骰所得的点数和, 则  $X$  的分布表 (详见 2.2 节) 为

|     |      |      |      |      |      |      |      |      |      |      |      |
|-----|------|------|------|------|------|------|------|------|------|------|------|
| $X$ | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | 11   | 12   |
| $P$ | 1/36 | 2/36 | 3/36 | 4/36 | 5/36 | 6/36 | 5/36 | 4/36 | 3/36 | 2/36 | 1/36 |

相应的 CDF 见图 2.1。



图 2.1:  $X$  的 CDF 图象

注: 由于软件限制, 各个阶跃点的绘制方式不太规范, 实际上从其左侧逼近应该为一个空圈, 例如  $F(3) = 3/36$  而不是  $1/36$ 。另外,  $\forall x < 2, F(x) = 0; \forall x \geq 12, F(x) = 1$ 。

**命题 2.1.** CDF 的性质:

1.  $F$  单调递增 (未必严格单调递增)
2.  $\lim_{x \rightarrow +\infty} F(x) = 1, \lim_{x \rightarrow -\infty} F(x) = 0$
3.  $F$  右连续

可以证明, 上述三条性质是任意函数  $F: \mathbb{R} \rightarrow \mathbb{R}$  成为 CDF 的充要条件。

思考: 如果我们将 CDF 的定义改为  $P(X < x)$ , 上述性质会如何变化?

**命题 2.2.** 若  $X, Y$  为随机变量, 则  $aX + bY, XY, X/Y$  (需  $Y \neq 0$ ) 都是随机变量。一般地, 若  $g$  为可测函数, 则  $g(X, Y)$  是随机变量。

**定义 2.4.** 设  $X_1, X_2$  的 CDF 分别为  $F_1, F_2$ , 我们称  $X_1$  与  $X_2$  同分布, 若  $\forall x \in \mathbb{R}, F_1(x) = F_2(x)$ 。

**命题 2.3.** 随机变量  $X_1$  与  $X_2$  同分布的一个充要条件是  $\forall$  可测集  $I \subset \mathbb{R}, P(X_1 \in I) = P(X_2 \in I)$ 。

注意, 同分布不等价于“同变量”, 即两个同分布的变量的取值不一定恒等。

**例 2.3.** 掷一次硬币,  $X$  表示正面向上次数,  $Y$  表示反面向上次数, 显然  $X$  与  $Y$  同分布, 但取值不等。

## 2.2 离散随机变量

**定义 2.5.** 离散随机变量  $X$  的概率质量函数 (Probability Mass Function, PMF)  $f$  是指该随机变量取各个可能值的概率, 即  $f(x) = P(X = x), \forall x \in \mathbb{R}$ 。可以用分布表的形式展示各个可能取值与概率的对应关系。

**命题 2.4.** 如果离散随机变量  $X$  的所有可能取值为  $\{x_i\}$ , 则  $X$  的 PMF 具有如下性质:

1.  $f(x_i) = p_i \geq 0, \forall i$
2.  $\sum_i p_i = 1$
3.  $F(x) = \sum_{x_i \leq x} f(x_i)$

**定义 2.6.** 离散随机变量  $X$  的期望定义为  $E(X) = \sum_i x_i p_i$ 。

我们称  $X$  的期望存在, 当且仅当  $\sum_i |x_i| p_i < +\infty$ 。

当期望存在时, 其方差定义为  $\text{Var}(X) = \sum_i (x_i - E(X))^2 p_i = E((X - E(X))^2) = E(X^2) - E^2(X)$ 。

当方差有限时, 称其算术平方根为  $X$  的标准差, 记作  $\text{SD}(X)$ 。

注意, 通常我们所说的一个随机变量的均值指的就是期望。

标准化指的是对  $X$  作线性变换  $\frac{X - \mu}{\sigma}$ , 其中  $\mu$  和  $\sigma$  分别为  $X$  的期望和标准差, 得到均值为 0, 标准差为 1 的随机变量。

对于可测函数  $g$ ,  $g(X)$  也是随机变量, 其期望  $E(g(X)) = \sum_i g(x_i) p_i$ 。

期望反映了随机变量的集中趋势, 而方差反映了其分散程度。

## 2.3 常见离散分布

**定义 2.7.** 称一个随机变量  $X$  服从 *Bernoulli* 分布, 若  $\exists p \in (0, 1)$ ,  $X$  的取值集合为  $\{0, 1\}$ , 且  $P(X = 1) = p, P(X = 0) = 1 - p$ . 记作  $X \sim B(p)$ .

$B(p)$  中的  $p$  称为该 *Bernoulli* 分布的参数。后续介绍的其他分布同理。

我们常将两种取值分别称为“成功”和“失败”。

计算可得, 若  $X \sim B(p)$ , 则  $E(X) = p, \text{Var}(X) = p(1 - p)$ 。

**定义 2.8.** 称一个随机变量  $X$  服从二项分布, 若  $\exists N \in \mathbb{N}^*, p \in (0, 1)$ ,  $X$  的取值集合为  $\{0, 1, \dots, N\}$ , 且  $P(X = k) = \binom{N}{k} p^k (1 - p)^{N-k} (k \in \{0, 1, \dots, N\})$ . 记作  $X \sim B(N, p)$ 。

我们常将  $k$  理解为“ $N$  次独立 *Bernoulli* 试验中的成功次数”。

计算可得, 若  $X \sim B(N, p)$ , 则  $E(X) = Np, \text{Var}(X) = Np(1 - p)$ 。

**定义 2.9.** 称一个随机变量  $X$  服从 *Poisson* 分布, 若  $\exists \lambda > 0$ ,  $X$  的取值集合为  $\mathbb{N}$ , 且  $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} (k \in \mathbb{N})$ . 记作  $X \sim P(\lambda)$ 。

计算可得, 若  $X \sim P(\lambda)$ , 则  $E(X) = \lambda, \text{Var}(X) = \lambda$ 。

对 *Poisson* 分布的一种常见理解是“一段时间内某个小概率事件发生的次数”所服从的分布。例如, 观察时间  $(0, 1]$  内某路口的交通事故数  $X$ , 将  $(0, 1]$  区间等分成  $n$  个小区间, 即  $l_i = (\frac{i-1}{n}, \frac{i}{n}] (i = 1, 2, \dots, n)$ 。考虑到  $n$  很大时, 每个区间的长度很小, 我们作如下假设:

1. 每段区间内, 至多发生一次事故
2.  $l_i$  上发生一次事故的概率与区间长度  $(1/n)$  成正比, 为  $p = \lambda/n$
3. 各区间内是否发生事故彼此独立

则  $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \rightarrow \frac{\lambda^k e^{-\lambda}}{k!} (n \rightarrow +\infty)$ , 即  $X \sim P(\lambda)$ 。

**例 2.4.** 设某医院平均每天出生婴儿数为  $\lambda$ , 则接下来  $t$  天内出生婴儿数服从参数为  $t\lambda$  的 *Poisson* 分布。

对于一般的二项分布  $X \sim B(N, p)$ , 若  $p$  很小,  $N$  很大, 而  $\lambda = Np$  不太大, 则近似有  $X \sim P(\lambda)$ , 且近似误差不超过  $\min\{p, Np^2\}$ 。

进一步, 若  $N$  次 *Bernoulli* 试验并非严格独立, 但满足弱相依条件, 则 *Poisson* 分布仍为一种较好的近似。

**例 2.5.** (配对问题)

$A_i$  表示第  $i$  个人拿到自己的帽子, 则  $P(A_i) = 1/n, P(A_i | A_j) = \frac{1}{n-1} (j \neq i)$ , 当  $n$  很大时,  $1/n$

和  $\frac{1}{n-1}$  很接近, 可以认为满足弱相依条件。

记  $X$  为拿到自己帽子的人数, 则  $X$  近似服从参数为  $\lambda = np = n \cdot \frac{1}{n} = 1$  的 Poisson 分布, 即  $P(X = k) \approx \frac{e^{-1}}{k!}$ 。

我们用常规做法检查这种近似是否合理。首先考虑指定的某  $k$  人, 记事件  $E$  表示这  $k$  人拿到自己的帽子, 事件  $F$  表示其余  $(n - k)$  人未拿到自己的帽子, 则  $P(EF) = P(E)P(F|E) = \frac{(n-k)!}{n!} \cdot P_{n-k}$ , 其中  $P_{n-k}$  为  $(n - k)$  人随机拿帽子时无人拿对的概率。那么我们有  $P(X = k) = \binom{n}{k} P(EF) = \frac{1}{k!} P_{n-k} \rightarrow \frac{e^{-1}}{k!} (n \rightarrow +\infty)$ 。这说明前述的近似是较好的。

## 2.4 连续随机变量

**定义 2.10.** 对随机变量  $X$ , 若存在  $f: \mathbb{R} \rightarrow [0, +\infty)$ , 使得  $\forall$  可测集  $I \subset \mathbb{R}$ , 都有  $P(X \in I) = \int_I f(x)dx$ , 则称  $X$  为连续型随机变量,  $f$  称为其概率密度函数 (Probability Density Function, PDF)。

**命题 2.5.** 连续随机变量  $X$  的 PDF 具有如下性质:

1.  $\int_{-\infty}^{+\infty} f(x)dx \equiv 1$
2.  $P(a < X \leq b) = \int_a^b f(x)dx = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$
3.  $P(X = a) \equiv 0, \forall a \in \mathbb{R}$
4. 若  $f$  在  $x_0$  处连续, 则  $P(x_0 - \delta < X < x_0 + \delta) = \int_{x_0 - \delta}^{x_0 + \delta} f(t)dt \approx f(x_0) \cdot 2\delta$
5.  $F(x) = \int_{-\infty}^x f(t)dt$  连续, 且若  $f$  在  $x$  处连续, 有  $F'(x) = f(x)$
6. PDF 若存在, 则不唯一 (可以修改其在任意零测集上的值, 得到不同的 PDF)

**定义 2.11.** 连续随机变量  $X$  的期望定义为  $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$ 。

我们称  $X$  的期望存在, 当且仅当  $\int_{-\infty}^{+\infty} |x|f(x)dx < +\infty$ 。

当期望存在时, 其方差定义为  $\text{Var}(X) = \int_{-\infty}^{+\infty} (x - E(x))^2 f(x)dx = E((X - E(X))^2) = E(X^2) - E^2(X)$ 。

当方差有限时, 称其算术平方根为  $X$  的标准差, 记作  $\text{SD}(X)$ 。

对于可测函数  $g$ ,  $g(X)$  也是随机变量, 其期望  $E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx$ 。

## 2.5 常见连续分布

**定义 2.12.** 称一个连续型随机变量  $X$  服从均匀分布, 若其 PDF 为  $f(x) = \frac{1}{b-a}(x \in (a, b))$ ,  $f$  在其余各处取 0。记作  $X \sim U(a, b)$ 。

我们常将  $X \sim U(0, 1)$  称为随机数。

计算可得, 若  $X \sim U(a, b)$ , 则  $E(X) = \frac{a+b}{2}$ ,  $\text{Var}(X) = \frac{(b-a)^2}{12}$ 。

**定义 2.13.** 称一个连续型随机变量  $X$  服从正态分布, 若其 PDF 为  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$  ( $\sigma > 0$ )。记作  $X \sim N(\mu, \sigma^2)$ 。

计算可得, 若  $X \sim N(\mu, \sigma^2)$ , 则  $E(X) = \mu$ ,  $\text{Var}(X) = \sigma^2$ 。

著名的“经验法则”见图 2.2。



图 2.2: 经验法则

$X \sim N(\mu, \sigma^2)$  的充要条件是  $Y = \frac{X-\mu}{\sigma} \sim N(0, 1)$ 。我们将  $N(0, 1)$  称为标准正态分布。

**定义 2.14.** 称一个连续型随机变量  $X$  服从指数分布, 若其 PDF 为  $f(x) = \lambda e^{-\lambda x}$  ( $\lambda > 0, x > 0$ ),  $f$  在其余各处取 0。记作  $X \sim \text{Exp}(\lambda)$ 。

指数分布常用于刻画等待时间、寿命等。

计算可得, 若  $X \sim \text{Exp}(\lambda)$ , 则  $E(X) = 1/\lambda$ ,  $\text{Var}(X) = 1/\lambda^2$ 。

指数分布有另一种符号约定, 以  $\beta = 1/\lambda$  为参数, 一些数学软件可能采用此种约定。

指数分布的 CDF 为  $F(x) = 1 - e^{-\lambda x}$  ( $x > 0$ ), 所谓的“尾概率”为  $P(X > x) = 1 - F(x) = e^{-\lambda x}$  ( $x > 0$ )。



**例 2.6.** 设某医院平均每天出生婴儿数为  $\lambda$ ，现在观察到一名婴儿出生，则接下来  $t$  天内有婴儿出生的概率为  $P(X \leq t)$ ，其中  $X$  表示到下一个婴儿出生所需等待的时间。

记  $N(t)$  为  $t$  天内出生婴儿数，我们已经知道  $N(t) \sim P(t\lambda)$ ，则  $P(X > t) = P(N(t) = 0) = e^{-\lambda t}$ ，故  $P(X \leq t) = 1 - e^{-\lambda t}$ 。我们发现  $X$  服从参数为  $\lambda$  的指数分布。

我们从另一个角度理解指数分布。

首先引入失效率或危险率的概念。设  $X$  为连续型随机变量（表示某种零件的寿命），其 CDF 为  $F(x)$ ，且  $F(0) = 0$ 。考虑条件概率  $P(x < X < x + dx | X > x) = \frac{P(x < X < x + dx)}{P(X > x)} = \frac{F(x+dx) - F(x)}{1 - F(x)} \approx \frac{F'(x)}{1 - F(x)} dx$ ，即“年龄”为  $x$  的零件不能继续工作的条件概率密度为  $\frac{F'(x)}{1 - F(x)}$ ，我们称其为瞬时失效率  $\lambda(x)$ ，则  $F(x) = 1 - e^{-\int_0^x \lambda(t) dt}$ 。

在“无老化”假设下，即  $\lambda(t) \equiv \lambda$  不随时间变化，则  $F(x) = 1 - e^{-\lambda x} (x > 0)$ ， $X$  服从指数分布。

指数分布有所谓“无记忆性”： $P(X > t+s | X > s) = \frac{P(X > t+s)}{P(X > s)} = e^{-\lambda t} = P(X > t) (t, s > 0)$ 。

“无老化”假设并不总是成立。为此，我们可以进行一定程度的改进，例如令  $\lambda(x) = \alpha \frac{x^{\alpha-1}}{\beta^\alpha} (x > 0, \alpha, \beta > 0 \text{ 为常数})$ ，则  $F(x) = 1 - e^{-(\frac{x}{\beta})^\alpha} (x > 0)$ ，称之为 Weibull 分布。当  $\alpha = 1$  时，Weibull 分布退化为参数为  $1/\beta$  的指数分布。

总览至此我们介绍过的各个分布的参数，可以将其大致分为以下几类：

1. 位置参数：决定了分布平移到的位置，通常在 PMF/PDF 中体现为  $f(x) = g(x - \cdot)$  的形式，如正态分布的参数  $\mu$
2. 尺度参数：决定了分布伸缩的程度，通常在 PMF/PDF 中体现为  $f(x) = g(\frac{x}{\cdot})$  的形式，如正态分布的参数  $\sigma$ 、Weibull 分布的参数  $\beta$
3. 形状参数：决定了分布的形状，如 Weibull 分布的参数  $\alpha$

## 2.6 随机变量的函数

对于随机变量  $X$  和可测函数  $g$ ， $Y = g(X)$  也是随机变量。特别地，若  $X$  为离散型随机变量，则  $Y$  也离散。但若  $X$  为连续型随机变量， $Y$  未必连续。

**例 2.7.**  $X \sim \text{Exp}(\lambda)$ ， $Y = \begin{cases} 0, & X \leq t_0, \\ 1, & X > t_0, \end{cases}$  其中  $t_0 > 0$  为常数，则  $Y \sim B(e^{-\lambda t_0})$ 。

**例 2.8.** 设  $X$  为连续型随机变量，PDF 为  $f(x)$ ，考虑  $Y = X^2$ 。

从 CDF 入手， $\forall y > 0, P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f(x) dx$ ，我们有  $Y$  的 PDF 为  $l(y) = \frac{d}{dy} P(Y \leq y) = \frac{1}{2\sqrt{y}} (f(\sqrt{y}) + f(-\sqrt{y})) (y > 0)$ 。

特别地，若  $X \sim N(0, 1)$ ，称  $Y$  服从自由度为 1 的  $\chi^2$ -分布，读作“卡方分布”。

若  $Y = g(X)$  为随机变量，我们可以计算  $Y$  的分布如下：

- $P(Y = y) = P(g(X) = y) = P(X \in g^{-1}(y))$
- $P(Y \leq y) = P(g(X) \leq y) = P(X \in g^{-1}((-\infty, y]))$

## 第三章 联合分布

### 3.1 随机向量

**定义 3.1.** 称  $(X_1, X_2, \dots, X_n) : \Omega \rightarrow \mathbb{R}^n$  为  $(n \text{ 维})$  随机向量, 若  $\{X_i\}_{i=1}^n$  均为随机变量。

**定义 3.2.**  $n$  维随机向量的 (联合) (累积) 分布函数 (CDF) 定义为  $F(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n), \forall (x_1, \dots, x_n) \in \mathbb{R}^n$ 。

对于  $n = 2$  (二元分布) 的情形, 我们常用  $(X, Y)$  来表示随机向量, 对应的 CDF 为  $F(x, y)$ 。

### 3.2 离散分布

**定义 3.3.** 称  $n$  维随机向量  $(X_1, \dots, X_n)$  是离散的, 当且仅当  $\{X_i\}_{i=1}^n$  均为离散随机变量。

离散随机向量  $(X_1, \dots, X_n)$  的 (联合) 概率质量函数 (PMF) 定义为  $f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n), \forall (x_1, \dots, x_n) \in \mathbb{R}^n$ 。

**命题 3.1.** 离散随机向量  $(X_1, \dots, X_n)$  的 PMF 具有如下性质:

1.  $f(x_1, \dots, x_n) \geq 0, \forall (x_1, \dots, x_n) \in \mathbb{R}^n$
2.  $\sum_{x_i \in \{X_i(\omega) | \omega \in \Omega\}, \forall i \in \{1, \dots, n\}} f(x_1, \dots, x_n) \equiv 1$

注意第 2 条性质中求和的项数为至多可数, 原因是有限个至多可数集的笛卡尔积仍是至多可数集。

**例 3.1.** 设  $\{B_i\}_{i=1}^n$  为  $\Omega$  的一个分割 (分割的定义见 1.8 节),  $P(B_i) = p_i \geq 0, \forall i \in \{1, \dots, n\}$ ,  $\sum_{i=1}^n p_i = 1$ 。

进行  $N$  次独立试验, 设  $\forall i \in \{1, \dots, n\}$ , 有  $X_i$  个试验结果落在  $B_i$  中, 则若  $k_1 + \dots + k_n = N$ , 其中  $k_i$  均为非负整数, 我们有  $P(X_1 = k_1, \dots, X_n = k_n) = \binom{N}{k_1, \dots, k_n} p_1^{k_1} \dots p_n^{k_n}$ 。其中  $\binom{N}{k_1, \dots, k_n} = \frac{N!}{k_1! \dots k_n!}$  为多项式  $(a_1 + \dots + a_n)^N$  中  $a_1^{k_1} \dots a_n^{k_n}$  项的系数。

我们称  $(X_1, \dots, X_n)$  服从多项分布。

### 3.3 连续分布

**定义 3.4.** 对  $n$  维随机向量  $(X_1, \dots, X_n)$ , 若存在  $f: \mathbb{R}^n \rightarrow [0, +\infty)$ , 使得  $\forall$  可测集  $Q \subset \mathbb{R}^n$ , 都有  $P((X_1, \dots, X_n) \in Q) = \int_Q f(x_1, \dots, x_n) dx_1 \cdots dx_n$ , 则称  $(X_1, \dots, X_n)$  为连续型随机向量,  $f$  称为其 (联合) 概率密度函数 (PDF)。

**命题 3.2.** 连续随机向量  $(X_1, \dots, X_n)$  的 PDF 具有如下性质:

1.  $\int_{\mathbb{R}^n} f(x_1, \dots, x_n) dx_1 \cdots dx_n \equiv 1$
2. 以  $n=2$  为例,  $F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(t, s) ds dt$ ,  $f(a, b) = \frac{\partial^2 F}{\partial x \partial y}(a, b)$ , a.e.

其中 a.e. 表示 “almost everywhere”。

**例 3.2.** 矩形域上的均匀分布的 PDF:  $f(x, y) = \begin{cases} \frac{1}{(b-a)(d-c)}, & (x, y) \in (a, b) \times (c, d), \\ 0, & \text{其他.} \end{cases}$

**例 3.3.** 二元正态分布  $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$  的 PDF:

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_2}{\sigma_2})^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2})}, \forall (x, y) \in \mathbb{R}^2, \sigma_1, \sigma_2 > 0, |\rho| < 1.$$

令  $\mathbf{x} = \begin{bmatrix} \frac{x-\mu_1}{\sigma_1} \\ \frac{y-\mu_2}{\sigma_2} \end{bmatrix}$ ,  $W = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$ ,  $W = A^T A$  为正定矩阵  $W$  的 Cholesky 分解, 则  $-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_2}{\sigma_2})^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2}) = -\frac{1}{2}\mathbf{x}^T W \mathbf{x} = -\frac{1}{2}\mathbf{x}^T A^T A \mathbf{x} = -\frac{1}{2}(A\mathbf{x})^T (A\mathbf{x})$ .

上述 Cholesky 分解的结果为  $A = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} 1 & -\rho \\ 0 & \pm\sqrt{1-\rho^2} \end{bmatrix}$  或  $A = \frac{1}{\sqrt{1-\rho^2}} \begin{bmatrix} -1 & \rho \\ 0 & \pm\sqrt{1-\rho^2} \end{bmatrix}$ 。

### 3.4 边际分布

对  $n$  维随机向量  $(X_1, \dots, X_n)$ , 称  $F_i(x) = P(X_i \leq x) = P(X_i \leq x, -\infty < X_j < +\infty, \forall j \neq i)$  为  $X_i$  的边际分布。

例如, 若  $n=2$ , 随机向量  $(X, Y)$  有 CDF  $F(x, y)$ , 则  $X$  的边际分布为  $F_X(x) = P(X \leq x) = P(X \leq x, Y \in \mathbb{R}) = \lim_{y \rightarrow +\infty} P(X \leq x, -\infty < Y \leq y) = \lim_{y \rightarrow +\infty} F(x, y)$ 。

若  $n=3$ , 随机向量  $(X, Y, Z)$  有 CDF  $F(x, y, z)$ , 则  $X$  的边际分布为  $F_X(x) = \lim_{y, z \rightarrow +\infty} F(x, y, z)$ ,  $(X, Y)$  的边际分布为  $F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P(X \leq x, Y \leq y, -\infty < Z < +\infty) = \lim_{z \rightarrow +\infty} F(x, y, z)$ 。

**例 3.4.** 设二维随机向量  $(X, Y)$  的 CDF 为  $F(x, y)$ , 则  $\forall a, b \in \mathbb{R}, P(X > a, Y > b) = 1 - F_X(a) - F_Y(b) + F(a, b)$ 。

对于离散型随机向量, 以  $n = 2$  为例, 定义边际 PMF 为  $P(X = x) = \sum_y P(X = x, Y = y)$ 。

对于连续型随机向量, 以  $n = 2$  为例, 设联合 PDF 为  $f(x, y)$ , 则  $F_X(x) = P(X \leq x, Y \in \mathbb{R}) = \int_{-\infty}^x \int_{-\infty}^{+\infty} f(t, s) ds dt$ , 则  $X$  的边际 PDF 为  $f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy$ 。

**例 3.5.**  $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , 则  $f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$ , 即  $X \sim N(\mu_1, \sigma_1^2)$ 。同理  $Y \sim N(\mu_2, \sigma_2^2)$ 。

### 3.5 条件分布

以  $n = 2$  为例说明条件分布的概念, 考虑随机向量  $(X, Y)$ 。

对于离散型随机向量, 设联合 PMF 为  $P(X = a_i, Y = b_j) = p_{ij} \geq 0, \sum_{i,j} p_{ij} \equiv 1$ , 则在  $Y = b_j$  条件下的  $X$  的条件 PMF 为  $P(X = a_i | Y = b_j) = \frac{P(X=a_i, Y=b_j)}{P(Y=b_j)} = \frac{p_{ij}}{\sum_k p_{kj}}$ 。条件 PMF 满足  $\sum_i P(X = a_i | Y = b_j) \equiv 1, \forall j$ 。

对于连续型随机向量, 设联合 PDF 为  $f(x, y)$ , 首先考虑条件概率  $P(X \leq x | y \leq Y \leq y + dy) = \frac{P(X \leq x, y \leq Y \leq y + dy)}{P(y \leq Y \leq y + dy)} = \frac{\int_{-\infty}^x \int_y^{y+dy} f(t, s) ds dt}{\int_y^{y+dy} f_Y(s) ds}$ , 对  $x$  求导得  $X$  在  $y \leq Y \leq y + dy$  条件下的条件 PDF 为  $\frac{\int_y^{y+dy} f(x, s) ds}{\int_y^{y+dy} f_Y(s) ds} \rightarrow \frac{f(x, y)}{f_Y(y)} (dy \rightarrow 0)$ 。

**定义 3.5.** 对于连续型随机向量  $(X, Y)$ , 设联合 PDF 为  $f(x, y)$ , 若  $f_Y(y) > 0$ , 则称  $X$  在  $Y = y$  条件下的条件 PDF 为  $f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$ 。

可以验证  $f_{X|Y}(x|y)$  满足 PDF 的各性质。

相应的条件 CDF 为  $F_{X|Y}(a|y) = P(X \leq a | Y = y) = \int_{-\infty}^a f_{X|Y}(x|y) dx$ 。

我们熟知的各个定理均有适用于连续型随机向量的版本:

1.  $f(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$  (乘法法则)
2.  $f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_{-\infty}^{+\infty} f_{X|Y}(x|y)f_Y(y) dy$  (全概率公式)
3.  $f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{+\infty} f_{X|Y}(x|y)f_Y(y) dy}$  (Bayes 公式)

**例 3.6.**  $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , 则  $f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}\sigma_2} \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{(y-(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)))^2}{2(1-\rho^2)\sigma_2^2}}$ , 即  $Y|X = x \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), (1 - \rho^2)\sigma_2^2)$ 。

### 3.6 独立性

**定义 3.6.** 设二维随机向量  $(X, Y)$  的 CDF 为  $F(x, y)$ , 若  $F(x, y) = F_X(x)F_Y(y), \forall x, y \in \mathbb{R}$ , 则称  $X, Y$  相互独立。

可以证明, 对于二维离散型 (或连续型) 随机向量  $(X, Y)$ ,  $X, Y$  相互独立的充要条件是  $f(x, y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$ , 其中  $f(x, y)$  为联合 PMF (或 PDF)。

**定义 3.7.** 设  $n$  维随机向量  $(X_1, \dots, X_n)$  的 CDF 为  $F(x_1, \dots, x_n)$ , 若  $F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n), \forall x_1, \dots, x_n \in \mathbb{R}$ , 则称  $X_1, \dots, X_n$  相互独立。

可以证明, 对于  $n$  维离散型 (或连续型) 随机向量  $(X_1, \dots, X_n)$ ,  $X_1, \dots, X_n$  相互独立的充要条件是  $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n), \forall x_1, \dots, x_n \in \mathbb{R}$ , 其中  $f(x_1, \dots, x_n)$  为联合 PMF (或 PDF)。

**定理 3.1.** 1. 若  $X_1, \dots, X_n$  相互独立, 则  $\forall m \in \{1, \dots, n-1\}$ , 可测函数  $g_1, g_2$ , 有  $Y_1 = g_1(X_1, \dots, X_m)$  与  $Y_2 = g_2(X_{m+1}, \dots, X_n)$  相互独立。  
2. 若  $n$  维连续型随机向量  $(X_1, \dots, X_n)$  的联合 PDF  $f(x_1, \dots, x_n) = g_1(x_1) \cdots g_n(x_n), \forall x_1, \dots, x_n \in \mathbb{R}, g_i : \mathbb{R} \rightarrow [0, +\infty), \forall i \in \{1, \dots, n\}$ , 则  $X_1, \dots, X_n$  相互独立且  $\forall i \in \{1, \dots, n\}$ ,  $X_i$  的边缘 PDF  $f_i$  与  $g_i$  相差常数因子。

**例 3.7.** 设  $(X, Y)$  服从如图 3.1 的三角形域上的均匀分布, 即  $f(x, y) = \begin{cases} c, & (x, y) \in \text{三角形域}, \\ 0, & \text{其他} \end{cases}$ , 则  $X, Y$  不独立。



图 3.1: 三角形域上的均匀分布

## 3.7 随机向量的函数

本节中, 我们考虑给定随机向量  $(X_1, \dots, X_n)$  和可测函数  $g$ , 如何求  $Y = g(X_1, \dots, X_n)$  的分布。

首先介绍 “直接法”。

**例 3.8.**  $X_i \sim B(n_i, p) (i = 1, 2)$  独立,  $Y = X_1 + X_2$ , 则  $\forall k \in \{0, 1, \dots, n_1 + n_2\}$ ,

$$\begin{aligned}
 P(Y = k) &= P(X_1 + X_2 = k) \\
 &= \sum_{k_1=0}^k P(X_1 = k_1, X_2 = k - k_1) \\
 &= \sum_{k_1=0}^k P(X_1 = k_1)P(X_2 = k - k_1) \\
 &= \sum_{k_1=0}^k \binom{n_1}{k_1} p^{k_1} (1-p)^{n_1-k_1} \binom{n_2}{k-k_1} p^{k-k_1} (1-p)^{n_2-(k-k_1)} \\
 &= \left( \sum_{k_1=0}^k \binom{n_1}{k_1} \binom{n_2}{k-k_1} \right) p^k (1-p)^{n_1+n_2-k} \\
 &= \binom{n_1+n_2}{k} p^k (1-p)^{n_1+n_2-k}
 \end{aligned}$$

因此  $Y \sim B(n_1 + n_2, p)$ 。

**例 3.9.** 随机向量  $(X_1, X_2)$  有联合 PDF  $f(x_1, x_2)$ , 且  $X_1 > 0$ , 考虑  $Y = X_2/X_1$ , 有  $\forall y \in \mathbb{R}$ ,  $P(Y \leq y) = P(\frac{X_2}{X_1} \leq y) = P(X_2 \leq X_1 y) = \int_D f(x_1, x_2) dx_1 dx_2 = \int_0^{+\infty} \int_{-\infty}^{yx_1} f(x_1, x_2) dx_2 dx_1$ , 作  $x_2 = x_1 t$  换元得  $P(Y \leq y) = \int_0^{+\infty} \int_{-\infty}^y f(x_1, x_1 t) x_1 dt dx_1$ , 故  $Y$  的 PDF 为  $l(y) = \int_0^{+\infty} x_1 f(x_1, yx_1) dx_1$ 。



图 3.2: 区域  $D$  的范围, 其中边界线的斜率为  $y$

接下来介绍“密度函数变换法”。

设随机向量  $(X_1, X_2)$  有联合 PDF  $f(x_1, x_2)$ , 且有可逆可微的映射关系  $\begin{cases} Y_1 = g_1(X_1, X_2) \\ Y_2 = g_2(X_1, X_2) \end{cases}$ ,

据此解出逆映射  $\begin{cases} X_1 = h_1(Y_1, Y_2) \\ X_2 = h_2(Y_1, Y_2) \end{cases}$ , 则对于任意可测集  $A$ , 若  $(h_1, h_2)$  将  $A$  映射到集合  $B$ , 则

由可逆性可知  $B$  在  $(g_1, g_2)$  的映射下的值域为  $A$ 。因此我们有  $P((Y_1, Y_2) \in A) = P((X_1, X_2) \in B)$ 。

$B) = \int_B f(x_1, x_2) dx_1 dx_2 = \int_A f(h_1(y_1, y_2), h_2(y_1, y_2)) |J| dy_1 dy_2$ , 其中  $J$  为 Jacobi 行列式  $\det \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{bmatrix}$ , 因此  $(Y_1, Y_2)$  的联合 PDF 为  $l(y_1, y_2) = f(h_1(y_1, y_2), h_2(y_1, y_2)) |J|$ 。

**例 3.10.** 随机向量  $(X_1, X_2)$  有联合 PDF  $f(x_1, x_2)$ , 为求  $Y = X_1 + X_2$  的 PDF, 引入  $Z = X_1$ , 则  $\begin{cases} X_1 = Z \\ X_2 = Y - Z \end{cases}$ , Jacobi 行列式为  $\det \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} = -1$ , 故  $(Y, Z)$  的联合 PDF 为  $f(z, y - z) | -1 | = f(z, y - z)$ ,  $Y$  的边际 PDF 为  $l_Y(y) = \int_{-\infty}^{+\infty} f(z, y - z) dz$ 。

上例中, 若  $X_1, X_2$  相互独立, 则  $f(x_1, x_2) = f_1(x_1)f_2(x_2) \Rightarrow l_Y(y) = \int_{-\infty}^{+\infty} f_1(z)f_2(y - z)dz$ , 这称之为  $f_1$  和  $f_2$  的卷积, 记作  $f_1 * f_2$ 。

特别地, 若  $(X_1, X_2) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , 则  $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$ 。

利用上述随机向量的函数的 PDF 求解方法, 可以得到所谓卡方分布 ( $\chi^2$ -分布)、 $t$ -分布和  $F$ -分布的 PDF。这些分布的表达式较为复杂, 在此不一一罗列。它们在统计推断部分会大量使用, 感兴趣的同学可以查阅资料, 简单了解一下它们与标准正态分布的联系。



## 第四章 随机变量的数字特征

### 4.1 期望

离散型和连续型随机变量的期望分别参见定义 2.6 和定义 2.11。

对于随机向量，期望自然推广定义为  $E((X_1, \dots, X_n)) = (E(X_1), \dots, E(X_n))$ 。

**命题 4.1.** 期望有如下性质：

1. 离散型和连续型随机向量的函数的期望  $E(g(X_1, \dots, X_n))$  分别等于

$$\sum_{x_i \in \{X_i(\omega) | \omega \in \Omega\}, \forall i \in \{1, \dots, n\}} g(x_1, \dots, x_n) f(x_1, \dots, x_n) \text{ 和 } \int_{\mathbb{R}^n} g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

其中  $g$  为可测函数， $f$  分别为联合 PMF 与联合 PDF

2.  $E(aX + bY) = aE(X) + bE(Y), \forall$  常数  $a, b \in \mathbb{R}$
3. 若  $X_1, \dots, X_n$  相互独立，则  $E(X_1 \cdots X_n) = E(X_1) \cdots E(X_n)$

### 4.2 分位数

**定义 4.1.** 设  $X$  为连续型随机变量，若  $P(X \leq m) = F(m) = 1/2$ ，则称  $m$  为  $X$  的中位数。

和均值一样，中位数也是随机变量集中趋势的一种刻画。中位数不一定唯一。

若  $m$  是连续型随机变量  $X$  的中位数，则  $P(X < m) = P(X > m) = 1/2$ 。

以下给出更一般的中位数定义。

**定义 4.2.** 对随机变量  $X$ ，若  $P(X < m) \leq 1/2$ ，且  $P(X > m) \leq 1 - 1/2 = 1/2$ ，则称  $m$  为  $X$  的中位数。

**例 4.1.** 设离散型随机变量  $X$  的分布表为

|     |     |     |      |      |
|-----|-----|-----|------|------|
| $X$ | 1   | 2   | 3    | 4    |
| $P$ | 1/3 | 1/2 | 1/12 | 1/12 |

则其中位数为 2。

**定义 4.3.** 对随机变量  $X$ ,  $\forall \alpha \in (0, 1)$ , 若  $P(X < a) \leq \alpha$  且  $P(X > a) \leq 1 - \alpha$ , 则称  $a$  为  $X$  的 (下侧)  $\alpha$ -分位数。

上述定义的  $\alpha$ -分位数是不唯一的。为了唯一性, 我们考虑定义  $F^{-1}(\alpha) = \inf\{x | F(x) \geq \alpha\}$ 。

我们给出众数 (mode) 的方便定义:  $f(x)$  的最大值点, 其中  $f(x)$  为 PMF 或 PDF。由于 PDF 可在任意零测集上修改取值, 故这一定义并非严谨的。

## 4.3 方差

离散型和连续型随机变量的方差分别参见定义 2.6 和定义 2.11。

方差的意义: 若  $X$  为收益率, 则  $SD(X)$  称为波动率, 刻画了风险的大小。我们定义变异系数  $CV = \frac{SD(X)}{\mu}$ , 其中  $\mu = E(X) \neq 0$ 。

**命题 4.2.** 方差有如下性质:

1.  $\text{Var}(C) \equiv 0, C$  为常数
2.  $\text{Var}(CX) = C^2 \text{Var}(X)$
3.  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2E((X - E(X))(Y - E(Y)))$ , 且若  $X, Y$  独立, 则  $E((X - E(X))(Y - E(Y))) = 0$

## 4.4 协方差与相关系数

对随机变量  $X, Y$ , 设  $E(X) = \mu_1, E(Y) = \mu_2, \text{Var}(X) = \sigma_1^2, \text{Var}(Y) = \sigma_2^2$ 。

**定义 4.4.** 称  $X$  与  $Y$  的协方差  $\text{Cov}(X, Y) = E((X - \mu_1)(Y - \mu_2))$ 。

**命题 4.3.** 协方差有如下性质:

1.  $\text{Cov}(X, X) = \text{Var}(X)$
2.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3.  $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$
4.  $\text{Cov}(aX_1 + bX_2 + c, Y) = a\text{Cov}(X_1, Y) + b\text{Cov}(X_2, Y), \forall$  常数  $a, b, c \in \mathbb{R}$

**定义 4.5.** 称  $X$  与  $Y$  的 (线性) 相关系数  $\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1 \sigma_2} = E\left(\frac{X - \mu_1}{\sigma_1} \frac{Y - \mu_2}{\sigma_2}\right)$ 。

若  $\text{Corr}(X, Y) = 0$ , 称  $X, Y$  不相关。

**定理 4.1.** 相关系数有如下性质:

1. 若  $X, Y$  相互独立, 则  $X, Y$  不相关 (反之未必成立)

2.  $|\text{Corr}(X, Y)| \leq 1$ , 且等号成立当且仅当  $\exists a, b, P(Y = aX + b) = 1$ , 即  $Y = aX + b, \text{a.s.}$

其中 a.s. 表示 “almost surely”。

为证明上述定理的 (2), 首先我们利用 Cauchy-Schwartz 不等式证明引理: 对随机变量  $U, V$ , 有  $E^2(UV) \leq E(U^2)E(V^2)$ , 且等号成立当且仅当  $\exists t_0 \in \mathbb{R}, P(V = t_0 U) = 1$ 。接下来令  $U = \frac{X - \mu_1}{\sigma_1}, V = \frac{Y - \mu_2}{\sigma_2}$ , 即得。

当  $\text{Corr}(X, Y) = \pm 1$ , 可以证明  $a = \pm \sigma_2 / \sigma_1$ 。

**例 4.2.**  $X \sim N(0, 1), Y = X^2$ , 则  $X$  与  $Y$  不相关, 但不独立。

**例 4.3.**  $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , 则

$$\begin{aligned} \text{Corr}(X, Y) &= E\left(\frac{X - \mu_1}{\sigma_1} \frac{Y - \mu_2}{\sigma_2}\right) \\ &= \int_{\mathbb{R}^2} \frac{x - \mu_1}{\sigma_1} \frac{y - \mu_2}{\sigma_2} \frac{1}{2\pi\sigma_1\sigma_2} \frac{1}{\sqrt{1 - \rho^2}} e^{-\frac{1}{2(1 - \rho^2)}\left(\left(\frac{x - \mu_1}{\sigma_1}\right)^2 + \left(\frac{y - \mu_2}{\sigma_2}\right)^2 - 2\rho \frac{x - \mu_1}{\sigma_1} \frac{y - \mu_2}{\sigma_2}\right)} dx dy \end{aligned}$$

进行换元  $(u, v)^T = A\left(\frac{x - \mu_1}{\sigma_1}, \frac{y - \mu_2}{\sigma_2}\right)^T$ , 其中  $A$  的定义参见例 3.3, 则指数上的项化为  $-\frac{1}{2}(u^2 + v^2)$ , 这一步实质上是进行了二次型的标准化。后续过程留作习题, 最终计算结果为  $\text{Corr}(X, Y) = \rho$ 。

## 4.5 矩

**定义 4.6.** 对  $k = 1, 2, \dots$ , 称  $E((X - c)^k)$  为  $X$  关于  $c$  点的  $k$  阶矩。特别地,  $c = 0$  的情况下称为  $k$  阶原点矩,  $c = E(X)$  的情况下称为  $k$  阶中心矩。

根据定义可知,  $E(X)$  为 1 阶原点矩, 而 1 阶中心矩恒等于 0;  $\text{Var}(X) = E(X^2) - E^2(X)$  为 2 阶中心矩。

若  $E(X) = \mu, \text{SD}(X) = \sigma$ , 我们称  $E\left(\left(\frac{X - \mu}{\sigma}\right)^k\right) = \frac{E((X - \mu)^k)}{\sigma^k}$  为  $k$  阶标准矩。

1 阶标准矩恒等于 0, 2 阶标准矩恒等于 1, 3 阶标准矩称为  $X$  的偏度系数, 记作  $\text{Skew}(X)$ 。

**例 4.4.**  $X \sim N(0, 1)$ , 则  $\text{Skew}(X) = \int_{-\infty}^{+\infty} x^3 f(x) dx = 0$ , 其中  $f$  为  $X$  的 PDF。

我们称偏度系数  $< 0$  的分布为 “负偏” 或 “左偏”, 如图 4.1。

5 阶以上的奇数阶标准矩计算更复杂, 受噪声影响更大。

4 阶标准矩称为  $X$  的峰度系数, 记作  $\text{Kurt}(X)$ 。由于正态分布的峰度系数恒等于 3, 因此常定义超额峰度系数为  $\text{Kurt}(X) - 3$ 。

我们经常将  $\mu \pm \sigma$  以内的范围称为 “峰”, 范围在 “峰” 以外但在  $\mu \pm 2\sigma$  以内的范围称为 “肩”, 范围在 “肩” 以外的部分称为 “尾”。

通常, 峰度系数  $> 3$  表现为相对于正态分布 “尖峰厚尾”, 如图 4.2。



图 4.1: 负偏分布

图 4.2: “Leptokurtic” 一词的含义即峰度系数  $> 3$ 

## 4.6 矩母函数

**定义 4.7.** 记  $M_X(t) = E(e^{tX})$ , 若  $M_X(t)$  在  $t = 0$  的某邻域内存在, 则称其为  $X$  的矩母函数 (Moment Generating Function, MGF), 否则称  $X$  的矩母函数不存在。

**例 4.5.** 若  $X \sim \text{Exp}(\lambda)$ , 则  $M_X(t) = E(e^{tX}) = \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t}, t < \lambda$ 。

**例 4.6.** 若  $X \sim N(0, 1)$ , 则  $M_X(t) = E(e^{tX}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx} e^{-\frac{x^2}{2}} dx = e^{\frac{t^2}{2}}, t \in \mathbb{R}$ 。

**命题 4.4.** 矩母函数有如下性质:

1.  $M_X(0) \equiv 1$
2.  $Y = aX + b$ , 则  $M_Y(t) = E(e^{tY}) = E(e^{t(aX+b)}) = e^{tb} M_X(at)$

**例 4.7.** 若  $Y \sim N(\mu, \sigma^2)$ , 令  $Y = \sigma X + \mu$ , 则  $X \sim N(0, 1)$ , 故  $M_Y(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{(\sigma t)^2}{2}} = e^{\frac{\sigma^2 t^2}{2} + \mu t}, t \in \mathbb{R}$ 。

矩母函数可以用于确定矩。

**定理 4.2.** 随机变量  $X$  的  $n$  阶 (原点) 矩与其矩母函数有如下关系:  $E(X^n) = M_X^{(n)}(0)$ 。

**证明.** 由 Taylor 展开有  $M_X(t) = \sum_{n=0}^{+\infty} M_X^{(n)}(0) \frac{t^n}{n!}$ , 又  $M_X(t) = E(e^{tX}) = E(\sum_{n=0}^{+\infty} X^n \frac{t^n}{n!}) = \sum_{n=0}^{+\infty} E(X^n) \frac{t^n}{n!}$ , 得到结论。  $\square$

**例 4.8.** 若  $X \sim N(0, 1)$ , 则  $M_X(t) = e^{\frac{t^2}{2}} = \sum_{n=0}^{+\infty} \frac{(\frac{t^2}{2})^n}{n!} = \sum_{n=0}^{+\infty} \frac{(2n)!}{2^n n!} \frac{t^{2n}}{(2n)!}$ , 因此我们得出  $E(X^{2n}) = \frac{(2n)!}{2^n n!}$ ,  $E(X^{2n+1}) \equiv 0$  ( $n = 0, 1, \dots$ )。

由此可以计算  $\text{Var}(X) = E(X^2) = 1$ ,  $\text{Kurt}(X) = E(X^4) = \frac{4!}{2^2 \cdot 2!} = 3$ 。

矩母函数还可以用于确定分布。

**定理 4.3.** 若存在  $a > 0$ , 使得  $M_X(t) = M_Y(t), \forall t \in (-a, a)$ , 则  $X, Y$  同分布。

**例 4.9.** 若随机变量  $X$  的矩母函数  $M_X(t) = \frac{1}{2}e^{-t} + \frac{1}{4} + \frac{1}{8}e^{4t} + \frac{1}{8}e^{5t}$ , 则  $X$  为离散型随机变量, 分布表为

|     |     |     |     |     |
|-----|-----|-----|-----|-----|
| $X$ | -1  | 0   | 4   | 5   |
| $P$ | 1/2 | 1/4 | 1/8 | 1/8 |

一般地, 若离散型随机变量  $X$  有 PMF  $P(X = k) = p_k$  ( $\sum_k p_k \equiv 1$ ), 则其 MGF 为  $M_X(t) = E(e^{tX}) = \sum_k e^{tk} p_k$ 。

注意, 各阶矩均相同的随机变量未必同分布。

**例 4.10.** 设连续型随机变量  $X_1$  和  $X_2$  的 PDF 分别为  $f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-\frac{(\log x)^2}{2}}, x > 0$  和  $f_2(x) = f_1(x)(1 + \sin(2\pi \log x)), x > 0$  ( $X_1$  服从对数正态分布), 则  $E(X_2^n) = E(X_1^n) + \int_0^{+\infty} x^n f_1(x) \sin(2\pi \log x) dx$ , 其中后一项通过换元  $y = \log x - n$  可以证明为 0, 即  $X_1$  和  $X_2$  同矩但不同分布。

下面我们运用矩母函数, 研究独立随机变量和的分布。

**定理 4.4.** 若随机变量  $X, Y$  独立,  $Z = X + Y$ , 则  $M_Z(t) = M_X(t)M_Y(t)$ 。

证明.  $M_Z(t) = E(e^{tZ}) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = M_X(t)M_Y(t)$ , 其中最后一个等号利用了独立性。□

推而广之, 若  $\{X_i\}_{i=1}^n$  相互独立,  $Z = X_1 + \dots + X_n$ , 则  $M_Z(t) = \prod_{i=1}^n M_{X_i}(t)$ 。

**例 4.11.** 若  $\{X_i\}_{i=1}^n$  相互独立且服从正态分布, 则  $X_1 + \dots + X_n$  也服从正态分布。

以  $n = 2$  为例说明。设  $X_i \sim N(\mu_i, \sigma_i^2)$  ( $i = 1, 2$ ), 则  $M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = e^{\frac{\sigma_1^2 t^2}{2} + \mu_1 t} e^{\frac{\sigma_2^2 t^2}{2} + \mu_2 t} = e^{\frac{1}{2}(\sigma_1^2 + \sigma_2^2) + (\mu_1 + \mu_2)t}$ , 对应  $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$  的 MGF, 再由 MGF 确定分布可得结论。

定义随机向量  $(X_1, \dots, X_n)$  的 MGF 为  $M_{X_1, \dots, X_n}(t_1, \dots, t_n) = E(e^{t_1 X_1 + \dots + t_n X_n})$ 。

以下简介类似 MGF 的其他函数:

1. 概率母函数 (Probability Generating Function, PGF), 仅针对非负整数取值的离散型随机变量  $X$ , 设其 PMF 为  $P(X = k) = p_k$ , 则其 PGF 定义为  $E(t^X) = \sum_{k=0}^{+\infty} p_k t^k, t \in [-1, 1]$ , 或对于  $t \in (0, 1]$ , 等于  $E(e^{X \log t}) = M_X(\log t)$ 。
2. 特征函数, 定义为  $E(e^{itX})$ , 其中  $i^2 = -1$ 。

## 4.7 条件期望

我们定义条件期望  $E(Y|X \in A) = \begin{cases} \sum_i y_i P(Y = y_i | X \in A) \\ \int_{-\infty}^{+\infty} y f_{Y|X}(y|X \in A) dy \end{cases}$ , 两种定义分别针对  $Y$  为离散型和连续型随机变量。

进而, 我们定义  $E(Y|x) = E(Y|X = x) = \begin{cases} \sum_i y_i P(Y = y_i | X = x) \\ \int_{-\infty}^{+\infty} y f_{Y|X}(y|x) dy \end{cases}$ , 注意到这是一个  $x$

的函数, 记作  $h(x)$ 。将其作用在  $X$  上, 得到  $h(X) = E(Y|X)$ , 这是一个  $X$  的函数 (称为  $Y$  对  $X$  的回归函数), 因此是一个新的随机变量。

**例 4.12.**  $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ , 则  $E(Y|x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)$ 。

**例 4.13.** 甲、乙两种同类产品, 平均使用寿命分别为 10 年和 15 年, 市场占有率分别为 60% 和 40%, 随机买一个, 则期望寿命是  $10 \times 60\% + 15 \times 40\% = 12$  年, 我们发现这个计算过程可以表示为  $E(Y) = E(Y|X = 1)P(X = 1) + E(Y|X = 2)P(X = 2) = h(1)P(X = 1) + h(2)P(X = 2) = E(h(X)) = E(E(Y|X))$ , 其中  $X = 1$  表示抽到甲产品,  $X = 0$  表示抽到乙产品,  $Y$  表示抽到的产品的寿命。

一般地, 我们有以下定理:

**定理 4.5.** (全期望公式)

对于随机向量  $(X, Y)$ , 有  $E(Y) = E(E(Y|X))$ 。

证明. 以连续型为例。设  $(X, Y)$  的联合 PDF 为  $f(x, y)$ , 有  $E(Y|x) = \int_{-\infty}^{+\infty} y f_{Y|X}(y|x) dy = \int_{-\infty}^{+\infty} y \frac{f(x, y)}{f_X(x)} dy$ , 故  $E(Y) = \int_{-\infty}^{+\infty} y f_Y(y) dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f(x, y) dx dy = \int_{-\infty}^{+\infty} E(Y|x) f_X(x) dx = E(E(Y|X))$ 。□

一般地, 对于可测函数  $g$ , 我们有  $E(g(X, Y)) = E(E(g(X, Y)|X))$ 。

**定理 4.6.** 对于随机向量  $(X, Y)$  和任意可测函数  $g: \mathbb{R} \rightarrow \mathbb{R}$ , 都有  $E((Y - g(X))^2) \geq E((Y - E(Y|X))^2)$ , 即条件期望是均方误差意义下的最优预测。

证明. 类比期望的性质  $E((Y - c)^2) \geq E((Y - E(Y))^2), \forall c \in \mathbb{R}$ , 我们有  $E((Y - g(X))^2|X) \geq E((Y - E(Y|X))^2|X), \forall g: \mathbb{R} \rightarrow \mathbb{R}$  可测, 两边对  $X$  求期望即得。  $\square$

我们经常用到最优线性预测, 即  $\min_{a,b} E((Y - (aX + b))^2)$ , 这种“均方意义上的最优”称之为最小二乘 (least square)。

# 第五章 不等式与极限定理

## 5.1 概率不等式

**定理 5.1.** (Markov 不等式)

若随机变量  $Y \geq 0$ , 则  $\forall a > 0$ , 有  $P(Y \geq a) \leq \frac{E(Y)}{a}$ 。

证明. 取示性变量  $I = \begin{cases} 1, & Y \geq a, \\ 0, & Y < a, \end{cases}$  则  $I \leq Y/a$ , 故  $P(Y \geq a) = E(I) \leq E(Y/a) = E(Y)/a$ 。

□

**定理 5.2.** (Chebyshev 不等式)

若随机变量  $Y$  的方差  $\text{Var}(Y)$  存在, 则  $\forall a > 0$  有  $P(|Y - E(Y)| \geq a) \leq \frac{\text{Var}(Y)}{a^2}$ 。

证明.  $P(|Y - E(Y)| \geq a) = P((Y - E(Y))^2 \geq a^2) \leq \frac{E((Y - E(Y))^2)}{a^2} = \frac{\text{Var}(Y)}{a^2}$ 。

□

这告诉我们, 如果  $\text{Var}(Y) = 0$ , 则  $P(Y = E(Y)) = 1$  (即 a.s.)。

**定理 5.3.** (Chernoff 不等式)

对于任意随机变量  $Y$ ,  $\forall a > 0, t > 0$ , 有  $P(Y \geq a) \leq \frac{E(e^{tY})}{e^{ta}}$ 。

证明.  $\forall t > 0, P(Y \geq a) = P(e^{tY} \geq e^{ta}) \leq \frac{E(e^{tY})}{e^{ta}}$ 。

□

**例 5.1.** 若  $X \sim N(0, 1)$ , 则

1. 根据 Markov 不等式,  $P(|X| \geq 3) \leq \frac{E(|X|)}{3} = \frac{1}{3}\sqrt{\frac{2}{\pi}} \approx 0.27$ ;
2. 根据 Chebyshev 不等式,  $P(|X| \geq 3) \leq \frac{\text{Var}(X)}{3^2} = \frac{1}{9} \approx 0.11$ ;
3. 根据 Chernoff 不等式,  $\forall t > 0, P(|X| \geq 3) = 2P(X \geq 3) \leq 2\frac{E(e^{tX})}{e^{3t}} = 2e^{\frac{t^2}{2}-3t}$ , 取最小值点  $t = 3$ , 得  $P(|X| \geq 3) \leq 2e^{-\frac{9}{2}} \approx 0.022$ ;
4. 根据经验法则,  $P(|X| \geq 3) \approx 0.003$ 。



## 5.2 大数定律

设随机变量  $X_1, \dots, X_n$  独立同分布, 均值  $E(X_i) = \mu$ , 方差  $\text{Var}(X_i) = \sigma^2 > 0$ , 则样本均值  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ , 其均值  $E(\bar{X}) = \mu$ , 方差  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \rightarrow 0 (n \rightarrow +\infty)$ 。

**定理 5.4.** (Khinchin 弱大数定律)

设随机变量  $X_1, \dots, X_n$  独立同分布, 均值  $E(X_i) = \mu$ , 方差  $\text{Var}(X_i) = \sigma^2 > 0$ , 则  $\forall \epsilon > 0$ , 有  $\lim_{n \rightarrow +\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$ , 或等价地,  $\lim_{n \rightarrow +\infty} P(|\bar{X} - \mu| < \epsilon) = 1$ 。

证明. 由 Chebyshev 不等式,  $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n} \frac{1}{\epsilon^2} \rightarrow 0 (n \rightarrow +\infty)$ .  $\square$

$\forall \epsilon > 0, \forall \alpha > 0$ , 如果我们将  $\epsilon$  和  $(1 - \alpha)$  分别称为精度和置信度, 则根据 Khinchin 弱大数定律,  $\exists N \in \mathbb{N}^+$ , 当  $n \geq N$  时,  $P(|\bar{X} - \mu| < \epsilon) \geq 1 - \alpha$ , 即  $\bar{X}$  至少以概率  $(1 - \alpha)$  落在区间  $(\mu - \epsilon, \mu + \epsilon)$  内。

换句话说, 当样本量足够大时, 有很大的概率  $\bar{X} \approx \mu$ , 其中  $\mu$  为未知的总体均值。

我们将  $X_i \sim B(p)$  这一特例称之为 Bernoulli 大数定律。

通过更进一步的讨论可以证明, 上述定理中关于方差的条件可以去掉, 结论仍正确。

此外, 我们还有对 Khinchin 弱大数定律的若干推广, 如

1. 要求  $X_i$  两两不相关,  $\text{Var}(X_i)$  一致有界, 我们就得到了 Chebyshev 大数定律;
2. 要求  $\text{Var}(\bar{X}) \rightarrow 0 (n \rightarrow +\infty)$ , 我们就得到了 Markov 大数定律。

**定义 5.1.** 我们称  $Y_n$  依概率收敛于  $Y$ , 记作  $Y_n \xrightarrow{P} Y$ , 如果  $\forall \epsilon > 0$ , 有  $\lim_{n \rightarrow +\infty} P(|Y_n - Y| \geq \epsilon) = 0$ 。

用上述定义, 弱大数定律可以表述为  $\bar{X} \xrightarrow{P} \mu$ 。

**定理 5.5.** (Kolmogorov 强大数定律)

设随机变量  $X_1, \dots, X_n$  独立同分布, 均值  $E(X_i) = \mu$ , 则  $P(\lim_{n \rightarrow +\infty} \bar{X} = \mu) = 1$ 。

考虑  $X_i \sim B(p)$  的特殊情形, 则  $\bar{X}$  称之为频率, 根据 Kolmogorov 强大数定律,  $P(\lim_{n \rightarrow +\infty} \bar{X} = p) = 1$ , 这说明概率的频率解释是合理的。

**定义 5.2.** 我们称  $Y_n$  以概率 1 收敛于  $Y$ , 又称几乎必然收敛于  $Y$ , 记作  $Y_n \xrightarrow{\text{a.s.}} Y$ , 如果  $P(\lim_{n \rightarrow +\infty} Y_n = Y) = 1$ 。

用上述定义, 强大数定律可以表述为  $\bar{X} \xrightarrow{\text{a.s.}} \mu$ 。

**例 5.2.** (Monte Carlo 积分)

设我们要计算  $g(x) > 0$  在区间  $[a, b]$  上的定积分, 首先取一个适当的  $c > \sup\{g(x)|x \in [a, b]\}$ ,

设  $(X_i, Y_i)$  独立且服从区域  $[a, b] \times [0, c]$  上的均匀分布, 记  $I_i = \begin{cases} 1, & Y_i \leq g(X_i), \\ 0, & Y_i > g(X_i), \end{cases}$  则  $I_i \sim B(p)$ ,

其中  $p = \frac{\int_a^b g(x)dx}{c(b-a)}$ , 于是  $\bar{I} = \frac{1}{n} \sum_{i=1}^n I_i \approx p$ , 从而  $\int_a^b g(x)dx \approx c(b-a)\bar{I}$ .

**例 5.3.** 我们通过一个例子来考察一下上面介绍的两种收敛性的区别。

设概率空间  $(\Omega, \mathcal{F}, P)$ , 其中  $\Omega = [0, 1]$ ,  $\omega$  在  $\Omega$  上均匀分布. 定义随机变量序列  $\forall \omega \in \Omega, Y_1(\omega) = \omega + I_{[0,1]}(\omega), Y_2(\omega) = \omega + I_{[0,1/2]}(\omega), Y_3(\omega) = \omega + I_{[1/2,1]}(\omega), Y_4(\omega) = \omega + I_{[0,1/3]}(\omega), Y_5(\omega) = \omega + I_{[1/3,2/3]}(\omega), Y_6(\omega) = \omega + I_{[2/3,1]}(\omega), \dots$ , 则  $Y_n(\omega)$  依概率收敛于  $Y(\omega) = \omega$ , 但不以概率 1 收敛于  $Y(\omega)$ , 因为  $\forall \omega_0 \in \Omega, Y_n(\omega_0)$  无极限。

## 5.3 中心极限定理

**定理 5.6.** 设随机变量  $X_1, \dots, X_n$  独立同分布, 均值  $E(X_i) = \mu$ , 方差  $\text{Var}(X_i) = \sigma^2 > 0$ , 则  $\forall x \in \mathbb{R}, \lim_{n \rightarrow +\infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x)$ , 其中  $\Phi(x)$  为标准正态分布的 CDF。或等价地,  $\lim_{n \rightarrow +\infty} P\left(\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$ 。

证明. 只对  $X_i$  的 MGF 存在的情形给出证明。

不失一般性, 假设  $\mu = 0, \sigma^2 = 1$ , 令  $M(t) = E(e^{tX_i})$ , 则  $M(0) = 1, M'(0) = 0, M''(0) = 1$ , 于是  $E(e^{t\frac{X_1 + \dots + X_n}{\sqrt{n}}}) = M^n\left(\frac{t}{\sqrt{n}}\right)$ , 而根据 Taylor 展开,  $M\left(\frac{t}{\sqrt{n}}\right) = 1 + 0 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{t^2}{n}\right)$ , 故  $E(e^{t\frac{X_1 + \dots + X_n}{\sqrt{n}}}) = (1 + \frac{t^2}{2n} + o(\frac{t^2}{n}))^n \rightarrow e^{t^2/2} (n \rightarrow +\infty)$ , 此为  $N(0, 1)$  的 MGF, 这说明  $\frac{X_1 + \dots + X_n}{\sqrt{n}}$  的分布趋近于  $N(0, 1)$ 。  $\square$

上述定理通常称为 Lindeberg-Lévy CLT, 可推广至不同分布的情形。

如果将定理中的  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  理解为标准化的过程, 则不难得出  $\bar{X}$  近似服从  $N(\mu, \frac{\sigma^2}{n})$ ,  $X_1 + \dots + X_n$  近似服从  $N(n\mu, n\sigma^2)$ 。

**例 5.4.** (De Moivre-Laplace CLT)

设  $X_i \sim B(p)$ , 则  $\sum_{i=1}^n X_i \sim B(n, p)$ , 当  $n$  充分大时, 可以近似地认为  $\sum_{i=1}^n X_i \sim N(np, np(1-p))$ , 于是我们可近似计算  $P(t_1 \leq \sum_{i=1}^n X_i \leq t_2) = P\left(\frac{t_1 - np}{\sqrt{np(1-p)}} \leq \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \leq \frac{t_2 - np}{\sqrt{np(1-p)}}\right) \approx \Phi(y_2) - \Phi(y_1)$ , 其中  $y_1 = \frac{t_1 - np - \frac{1}{2}}{\sqrt{np(1-p)}}$ ,  $y_2 = \frac{t_2 - np + \frac{1}{2}}{\sqrt{np(1-p)}}$ , 其中  $\frac{1}{2}$  是连续性修正项。

**定义 5.3.** (依分布收敛)

我们称  $Y_n$  依分布收敛于  $Y$ , 记作  $Y_n \xrightarrow{d} Y$ , 如果  $\lim_{n \rightarrow +\infty} F_{Y_n}(x) = F_Y(x), \forall x \in \mathbb{R}$ 。

用上述定义, CLT 可以表述为  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$ , 其中  $Z \sim N(0,1)$ , 或简记为  $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightarrow N(0,1)$ 。

**例 5.5.** (选举问题)

设  $p$  为选民真实支持度 (未知), 随机抽样调查  $n$  人 (假设  $n$  远远小于总人数  $N$ , 可以近似有放回抽样), 样本支持比例  $P_n = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ , 其中  $X_i \sim B(p)$  且独立, 表示第  $i$  个人是否支持。

设置精度  $\epsilon = 0.03$ , 置信度  $1-\alpha = 95\%$ , 则至少需要  $n$  为多少, 才能保证  $P(|P_n - p| < \epsilon) \geq 1-\alpha$ ? 根据 CLT, 我们有  $P(|P_n - p| \geq \epsilon) \approx 2 \left( 1 - \Phi\left(\frac{\epsilon}{\sqrt{p(1-p)/n}}\right) \right) \leq \alpha$ , 于是  $n \geq \frac{z_{\alpha/2}^2 p(1-p)}{\epsilon^2}$ , 其中  $z_{\alpha/2}$  为标准正态分布的上  $\alpha/2$  分位数, 代入最大值点  $p = \frac{1}{2}$ , 我们得到  $n \geq \frac{z_{\alpha/2}^2}{4} \epsilon^2$ , 代入  $\epsilon = 0.03, \alpha = 0.05$ , 得到  $n \geq 1068$ 。这一结果与  $N$  无关!

## 第二部分

## 统计推断

# 统计引言

统计学是一门从数据中获得信息的学问。根据 Claude Shannon 的信息论，所谓的信息就是不确定性的分解。

数理统计通常包括数据收集、数据分析和统计推断三部分。

**例.** 检测某厂的一大批电子元件产品的寿命，我们关注的问题是“判断产品是否合格”。这个问题的“总体”就是所需检测的这批元件的寿命，更具体地说，是元件寿命这一随机变量  $X$  的分布。

统计学上所谓总体，就是指一个概率分布。而统计分析问题就是研究对象全体所服从的分布的某个数字特征，来了解总体变量  $X$  的分布。

总体可以分为有限总体、无限总体等，其中有限总体在个体数量很多时可以近似看作无限总体。

所谓的“虚拟总体”是一种无限总体，并无实际存在的个体集合，而是一个假想的、潜在的无限个体集合，如测量讲桌的长度所得到的测量值，可以视为来自一个虚拟总体。

我们将一族概率分布称为一个统计模型。

**例.** 正态分布族  $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$  就是一个统计模型。

模型可以分为参数模型和非参数模型，正态分布族就是一个参数模型。非参数模型是指不能用少数几个参数决定的模型，例如对某总体  $X$ ，我们限定  $X$  连续， $E(X)$  存在或属于某个取值范围等条件，但不用具体的若干参数去精确描述  $X$  的分布，这就是一个非参数模型。

样本是指从总体中抽取的一组观测值  $X_1, \dots, X_n$ ，其中每个  $X_i$  来自总体  $X$ ，而  $n$  称为样本容量。

抽样方式分为试验与观测，后者又可以分为完全观测和不完全观测。

若  $X_1, \dots, X_n$  独立同分布，且  $X_i \sim X$ ，则称  $X_1, \dots, X_n$  为来自总体  $X$  的一个随机样本。对于有限总体，这需要有放回地抽样。

简单随机抽样是指当总体个数  $N$  有限，从中无放回地抽取  $n$  个个体，每个个体被抽取的概率相同。这种情况下，任意容量为  $n$  的样本都有相同的出现概率，为  $\frac{1}{\binom{N}{n}}$ 。

---

抽样方式的选择有很多需要注意的地方，否则可能属于不当抽样。

**定义.** 统计量定义为样本的函数，即  $T(X_1, \dots, X_n)$ 。

统计量是完全由样本决定的量，因此也是随机变量。统计量可以看作一种对数据进行简化的方式。

**例.** 设  $X_1, \dots, X_n$  独立同分布，均值  $E(X_i) = \mu$ ，则以下是一些常用的统计量：

1. 样本均值  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ;
2. 样本方差  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ ;
3. 当  $\mu$  已知时， $\bar{X} - \mu$  是统计量；当  $\mu$  未知时， $\bar{X} - \mu$  不是统计量。

总体决定样本，故我们可以通过样本来推断总体的性质，这就是统计推断。统计推断又可以分为经典方法（频率学派的）以及 Bayes 方法。

**例.** 设总体满足  $Y = aX + \epsilon$ ，其中  $X$  为自变量， $Y$  为因变量， $\epsilon$  为误差。这是一个参数模型。假设我们抽取的样本为  $(X_1, Y_1), \dots, (X_n, Y_n)$ ，则：

- 若  $a$  未知，可通过观测各  $(X_i, Y_i)$  来估计  $a$ ，这属于模型推断、参数估计的范畴；
- 若  $a$  已知，可通过观测  $Y_i$  来估计  $X_i$ ，这属于变量推断、模型应用的范畴。

**例.** 假设元件寿命  $X \sim \text{Exp}(\lambda)$ ，如何通过样本估计  $\lambda$  的值？这是一个参数估计问题。

假设元件的合格标准是  $E(X) \geq L$ ，但  $E(X)$  未知。考虑制定一种可操作的检验标准，当  $\bar{X} \geq l$  时，我们就认为元件合格。这种标准如何制定？这是一个假设检验问题。

## 第六章 参数估计

### 6.1 矩估计

设  $X_1, \dots, X_n$  为独立同分布的样本, 我们定义样本矩如下:

1.  $k$  阶样本原点矩  $\mu_k = \frac{1}{n} \sum_{i=1}^n X_i^k$
2.  $k$  阶样本中心矩  $m_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$

根据大数定律,  $\mu_k \rightarrow E(X^k)$ 。

矩估计就是用样本矩去估计参数。

**例 6.1.** 设  $X_1, \dots, X_n$  独立同分布,  $X_i \sim N(\mu, \sigma^2)$ , 则  $\mu = E(X) \approx \mu_1 = \bar{X}$ ,  $\sigma^2 = \text{Var}(X) \approx m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 。

**例 6.2.** 设  $X_1, \dots, X_n$  独立同分布,  $X_i \sim \text{Exp}(\lambda)$ , 则  $\lambda = E(X)^{-1} \approx \mu_1^{-1} = \frac{1}{\bar{X}}$ , 或  $\lambda = \text{Var}(X)^{-1/2} \approx m_2^{-1/2}$ 。

我们发现上例中  $\lambda$  可以有两种不同的矩估计, 一个基本原则是尽量用低阶矩。

### 6.2 极大似然估计

设  $(X_1, \dots, X_n)$  的联合分布 (PMF 或 PDF) 为  $f(x_1, \dots, x_n; \theta)$ , 其中  $\theta$  为未知参数。

对于观测  $(X_1, \dots, X_n)$ , 我们定义似然函数 (likelihood function) 为  $L(\theta) = f(X_1, \dots, X_n; \theta)$ 。

对于离散情形,  $L(\theta)$  就是当参数为  $\theta$  时出现观测  $(X_1, \dots, X_n)$  的概率。

随机变量  $X_1, \dots, X_n$  的一个实现是指一次观测到的具体数据, 记为  $x_1, \dots, x_n$ 。

若  $X_1, \dots, X_n$  独立同分布, 来自总体  $f_1(x; \theta)$  (PMF 或 PDF), 则  $f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_1(x_i; \theta)$ , 似然函数  $L(\theta) = \prod_{i=1}^n f_1(X_i; \theta)$ 。

**例 6.3.** 设  $X_1, \dots, X_n$  独立同分布,  $X_i \sim N(\mu, \sigma^2)$ ,  $\mu$  和  $\sigma^2$  未知, 则  $f_1(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ , 似然函数  $L(\theta) = L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$ 。

**定义 6.1.**  $\theta^* = \underset{\theta}{\operatorname{argmax}} L(\theta)$  称为  $\theta$  的极大似然估计 (MLE)。

注意  $\theta^* = \theta^*(X_1, \dots, X_n)$  是一个随机变量, 因为它是  $X_1, \dots, X_n$  的函数。

**例 6.4.** 上例中, 解方程  $\frac{\partial \log L}{\partial \mu} = 0$  和  $\frac{\partial \log L}{\partial (\sigma^2)} = 0$  (称它们为似然方程), 得  $\mu^* = \bar{X}$  和  $(\sigma^2)^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 。

此处 MLE 的结果与矩估计一致, 这是偶然现象, 对于一般分布不总成立。

**命题 6.1.** MLE 有重要的所谓不变性: 设  $\theta^*$  是  $\theta$  的 MLE,  $g(\theta)$  是  $\theta$  的可测函数, 则  $g(\theta^*)$  是  $g(\theta)$  的 MLE。例如, 如果上例中选择  $\theta = (\mu, \sigma)$ , 则  $\sigma^* = \sqrt{(\sigma^2)^*}$  是  $\sigma$  的 MLE。

**例 6.5.** 设  $X_1, \dots, X_n$  独立同分布,  $X_i \sim U(0, \theta), \theta > 0$  未知,  $L(\theta) = \begin{cases} \frac{1}{\theta^n}, & X_i \in (0, \theta), \forall i, \\ 0, & \text{其他,} \end{cases}$  则  $\theta^* = \max\{X_1, \dots, X_n\}$ 。

**例 6.6.** 设  $X_1, \dots, X_n$  独立同分布,  $X_i$  的 PDF 为  $f_1(x; \theta) = \frac{1}{\pi(1+(x-\theta)^2)} (x \in \mathbb{R})$ ,  $\theta$  未知, 即  $X_1, \dots, X_n$  服从 Cauchy 分布。

- 由于 Cauchy 分布的任意阶矩都不存在, 故不能用矩估计。
- 若采用 MLE 方法, 似然方程为  $\sum_{i=1}^n \frac{X_i - \theta}{1 + (X_i - \theta)^2} = 0$ , 当  $n$  较大时, 此方程有很多的根且无显式解, 故 MLE 方法也不理想。
- 一种可能的对  $\theta$  的估计: 由于  $\theta$  为中位数, 因此用样本中位数作为  $\theta$  的估计。

这个例子告诉我们, 统计方法不是唯一的, 也没有绝对的优劣。

需要指出, MLE 不一定是唯一的。

MLE 的另一局限性是它需要分布的具体函数形式, 而矩估计不需要。

此外, 如果似然函数在最大值点附近变化过于平缓, 则可能不利于通过迭代等方法有效计算。

## 6.3 优良性准则

无论是矩估计还是极大似然估计, 都是用样本的函数来估计总体的参数, 对每个参数给出一个估计值, 这样的估计称为点估计。

用于估计参数的函数  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  称为估计量, 其分布 (依赖于  $\theta$ ) 称为抽样分布, 其标准差  $\sqrt{\operatorname{Var}(\hat{\theta})}$  称为标准误 (差) (Standard error), 记为  $\operatorname{Se} = \operatorname{Se}(\hat{\theta})$ 。

在选择估计量时, 有若干准则。首先介绍所谓无偏性。

我们称  $E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$  为  $\hat{\theta}$  的偏差 (bias)。



**定义 6.2.** 设  $\hat{\theta}$  是  $\theta$  的估计量, 若  $\forall \theta, E(\hat{\theta} - \theta) = 0$ , 则称  $\hat{\theta}$  为  $\theta$  的一个无偏估计 (量)。

由上述定义可知, 无偏性指的是无系统偏差。

一般地, 若  $\hat{g}(X_1, \dots, X_n)$  是对  $\theta$  的函数  $g(\theta)$  的估计, 且满足  $\forall \theta, E(\hat{g}(X_1, \dots, X_n)) = g(\theta)$ , 则称  $\hat{g}(X_1, \dots, X_n)$  是  $g(\theta)$  的一个无偏估计。

对于无偏估计  $\hat{g}(X_1, \dots, X_n)$ , 若进行  $N$  组抽样, 第  $m$  组样本记作  $X_1^{(m)}, \dots, X_n^{(m)}$ , 则由大数定律,  $\frac{1}{N} \sum_{m=1}^N \hat{g}(X_1^{(m)}, \dots, X_n^{(m)})$  会收敛到  $E(\hat{g}(\theta)) = g(\theta)$ 。

在实际应用中, 无偏的重要性视情况而定。

**例 6.7.** 若随机变量  $X$  的均值  $\mu$  和方差  $\sigma^2$  均未知, 则由  $E(\bar{X}) = \mu$  知  $\bar{X}$  是  $\mu$  的无偏估计。而二阶矩  $m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X} - \mu)^2$ , 有  $E(m_2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$ , 故  $m_2$  不是  $\sigma^2$  的无偏估计 (系统偏小)。

样本方差  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  中的  $(n-1)$  是所谓的无偏修正, 满足  $E(S^2) = \sigma^2$ , 故  $S^2$  是  $\sigma^2$  的无偏估计。

**例 6.8.** 若随机变量  $X \sim U(0, \theta)$ , 则矩估计  $\hat{\theta} = 2\bar{X}$  为  $\theta$  的无偏估计, 而 MLE  $\theta^* = \max\{X_1, \dots, X_n\}$ , 有  $E(\theta^*) = \frac{n}{n+1}\theta$ , 故  $\theta^*$  不是  $\theta$  的无偏估计。

这个例子说明, MLE 不一定是无偏的。

下面介绍均方误差准则。

我们定义均方误差 (MSE) 为  $E((\hat{\theta} - \theta)^2) = \text{Var}(\hat{\theta}) + E^2(\hat{\theta} - \theta)$ , 其中等号右边的两项分别反映了精确度 (precision) 和准确度 (accuracy)。

**定义 6.3.** 若  $\hat{\theta}_1, \hat{\theta}_2$  为  $\theta$  的无偏估计, 且  $\forall \theta, \text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$ , 且存在一个  $\theta$  的值使得不等号严格成立, 则称  $\hat{\theta}_1$  在均方误差意义下优于  $\hat{\theta}_2$ 。

**例 6.9.** 若随机变量  $X$  的均值  $\mu$  未知, 方差为  $\sigma^2$ , 则  $\bar{X}, \frac{1}{2}(X_1 + X_2), X_1$  都是  $\mu$  的无偏估计, 它们各自的方差为  $\frac{\sigma^2}{n}, \frac{\sigma^2}{2}, \sigma^2$ , 故若  $n > 2$ , 则  $\bar{X}$  在均方误差意义下优于  $\frac{1}{2}(X_1 + X_2)$ , 而  $\frac{1}{2}(X_1 + X_2)$  在均方误差意义下优于  $X_1$ 。

**定义 6.4.** 若  $\hat{\theta}_0$  是  $\theta$  的无偏估计, 且  $\forall \hat{\theta}$  为  $\theta$  的无偏估计, 都有  $\forall \theta, \text{Var}(\hat{\theta}_0) \leq \text{Var}(\hat{\theta})$ , 则称  $\hat{\theta}_0$  是  $\theta$  的最小方差无偏估计 (MVUE)。

**例 6.10.** 若  $X \sim N(\mu, \sigma^2)$ , 则  $E(m_2) = \frac{n-1}{n} \sigma^2, E(S^2) = \sigma^2$ , 但  $E((m_2 - \sigma^2)^2) < E((S^2 - \sigma^2)^2)$ , 故  $m_2$  在均方误差意义下优于  $S^2$ 。尽管  $m_2$  是有偏的, 但它有更小的方差, 总的来说其 MSE 更小。

接下来介绍一些大样本性质。所谓大样本性质，是指样本容量  $n$  趋于无穷时  $\hat{\theta}$  的性质。

首先是渐进无偏性。若  $\lim_{n \rightarrow +\infty} E(\hat{\theta} - \theta) = 0$ ，则称  $\hat{\theta}$  具有渐进无偏性。

然后是相合性。若  $\forall \epsilon > 0, \lim_{n \rightarrow +\infty} P(|\hat{\theta} - \theta| \geq \epsilon) = 0$ ，则称  $\hat{\theta}$  是  $\theta$  的相合估计。

$\hat{\theta}$  是  $\theta$  的相合估计，当且仅当  $\hat{\theta} \xrightarrow{P} \theta$ 。例如，根据弱大数定律， $\bar{X}$  是  $\mu$  的相合估计。

相合性是良好点估计的自然要求。

**例 6.11.** 若随机变量  $X$  的均值为  $\mu$ ，方差为  $\sigma^2$ ，考虑  $m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\bar{X} - \mu)^2$ ，由大数定律， $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \xrightarrow{P} E((X_i - \mu)^2) = \sigma^2$ ，而  $(\bar{X} - \mu)^2 \xrightarrow{P} 0$ ，故  $m_2 \xrightarrow{P} \sigma^2$ ，即  $m_2$  是  $\sigma^2$  的相合估计。同时， $S^2 = \frac{n}{n-1} m_2 \xrightarrow{P} \sigma^2$ ，故  $S^2$  也是  $\sigma^2$  的相合估计。

最后是渐进正态性。若  $\frac{\hat{\theta} - \theta}{\text{Se}(\hat{\theta})} \xrightarrow{d} Z \sim N(0, 1)$ ，则称  $\hat{\theta}$  是  $\theta$  的渐进正态估计。

例如，根据 CLT， $\bar{X}$  是  $\mu$  的渐进正态估计，且  $\text{Se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ 。

若  $\hat{\theta}$  是  $\theta$  的渐进正态估计，则当  $n$  充分大时，近似有  $\hat{\theta} \sim N(\theta, \text{Se}^2(\hat{\theta}))$ 。

## 6.4 置信区间

**定义 6.5.**  $\forall \alpha \in (0, 1)$ ， $\hat{\theta}_i = \hat{\theta}_i(X_1, \dots, X_n) (i = 1, 2)$  为统计量，若  $P(\hat{\theta}_1 < \theta < \hat{\theta}_2) \geq 1 - \alpha$ ，则称  $(\hat{\theta}_1, \hat{\theta}_2)$  为  $\theta$  的一个  $(1 - \alpha)$ -置信的（双侧）区间估计。

$(1 - \alpha)$  称为置信水平，置信系数或置信度是指置信水平中的最大者，这三个术语都是针对方法而言的。 $\alpha$  通常取 0.05, 0.01, 0.1 等。

通常用  $E(\hat{\theta}_2 - \hat{\theta}_1)$  来刻画区间估计的精度。我们遵循可靠度优先原则，即先保证置信水平，然后再提升精度。

**例 6.12.** 设  $X_1, \dots, X_n$  独立同分布， $X_i \sim N(\mu, \sigma^2)$ ， $\mu$  未知， $\sigma^2$  已知，则由  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ ，有  $\bar{X} - \mu \sim N(0, \frac{\sigma^2}{n})$ 。为给出  $\mu$  的区间估计，我们的目标是寻找  $c_1, c_2$  使得  $P(\bar{X} - c_1 < \mu < \bar{X} + c_2) \geq 1 - \alpha$ ，这等价于  $P(-c_2 < \bar{X} - \mu < c_1) \geq 1 - \alpha$ 。设  $\alpha_1 = P(\bar{X} - \mu \leq -c_2)$ ， $\alpha_2 = P(\bar{X} - \mu \geq c_1)$ ，一个自然的选择是令  $\alpha_1 = \alpha_2 = \alpha/2$ （事实上这也是能够使精度最高的选择）。记  $z_{\frac{\alpha}{2}}$  为  $N(0, 1)$  的上  $\frac{\alpha}{2}$ -分位数，即  $\Phi(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$ ，则  $P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$ ，从而  $P(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$ ，故  $(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$  是  $\mu$  的一个  $(1 - \alpha)$ -置信的区间估计。

若  $\alpha = 0.05$ ，则  $z_{\frac{\alpha}{2}} \approx 1.96 \approx 2$ 。

上述区间估计的一种理解是：若用  $\bar{X}$  来估计  $\mu$ ，则绝对误差  $|\bar{X} - \mu|$  在  $(1 - \alpha)$ -置信下不超过  $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ 。

区间的半长度为  $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ , 如果给定精度, 例如取  $\epsilon > 0$ , 要求  $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \epsilon$ , 则  $n \geq (\frac{z_{\frac{\alpha}{2}} \sigma}{\epsilon})^2$ , 即样本容量至少为  $(\frac{z_{\frac{\alpha}{2}} \sigma}{\epsilon})^2$  时有  $(1 - \alpha)$ -置信使绝对误差不超过  $\epsilon$ 。这一推理可以理解为  $(\alpha, \epsilon, n)$  三个变量之间存在的关系。

**例 6.13.** 设  $X_1, \dots, X_n$  独立同分布,  $X_i \sim N(\mu, \sigma^2)$ ,  $\mu, \sigma^2$  未知, 首先估计  $\sigma^2$ 。注意到,  $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \sim \chi^2(n-1)$ , 同样令  $\alpha_1 = \alpha_2 = \alpha/2$ , 有  $(\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)}, \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)})$  是  $\sigma^2$  的一个  $(1 - \alpha)$ -置信的区间估计, 其中  $\chi_{\frac{\alpha}{2}}^2(n-1)$  和  $\chi_{1-\frac{\alpha}{2}}^2(n-1)$  分别为  $\chi^2(n-1)$  的上  $\frac{\alpha}{2}$ -分位数和下  $\frac{\alpha}{2}$ -分位数。

接下来估计  $\mu$ , 可以证明,  $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$  且与  $\frac{(n-1)S^2}{\sigma^2}$  独立, 从而  $\frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2} \frac{\sigma^2}{n-1}}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$ , 故  $(\bar{X} - t_{\frac{\alpha}{2}}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}}(n-1) \frac{S}{\sqrt{n}})$  是  $\mu$  的一个  $(1 - \alpha)$ -置信的区间估计, 其中  $t_{\frac{\alpha}{2}}(n-1)$  为  $t(n-1)$  的上  $\frac{\alpha}{2}$ -分位数。

**例 6.14.** 若  $X \sim N(\mu_1, \sigma^2), Y \sim N(\mu_2, \sigma^2)$ , 且  $X, Y$  独立, 下面估计均值差  $\mu_1 - \mu_2$ 。设随机样本为  $X_1, \dots, X_n$  和  $Y_1, \dots, Y_m$ , 则  $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$ , 有  $\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1)$ 。同时, 由  $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1)$  和  $\frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{\sigma^2} = \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1)$ , 且  $\frac{(n-1)S_1^2}{\sigma^2}$  与  $\frac{(m-1)S_2^2}{\sigma^2}$  独立, 有  $\frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2)$ , 故  $\frac{\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \frac{\sigma^2}{n+m-2}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2)$ , 其中  $S^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ , 于是  $(\bar{X} - \bar{Y} - t_{\frac{\alpha}{2}}(n+m-2) S \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + t_{\frac{\alpha}{2}}(n+m-2) S \sqrt{\frac{1}{n} + \frac{1}{m}})$  是  $\mu_1 - \mu_2$  的一个  $(1 - \alpha)$ -置信的区间估计。

类似点估计, 区间估计也有对应的大样本方法, 即所谓渐近置信区间。

**例 6.15.** (选举问题)

设  $p$  为未知的真实支持率, 样本容量  $n = 1200$ , 其中有 684 人支持, 即观测比例为  $\frac{684}{1200} \approx 0.57$ , 下面给出  $p$  的一个  $1 - \alpha = 95\%$  置信的区间估计。

记  $X_i$  为第  $i$  个人的态度, 1 表示支持, 0 表示不支持,  $X_i \sim B(p) (i = 1, 2, \dots, n)$  且独立, 记观测比例  $P_n = P_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ , 有  $E(P_n) = p, \text{Var}(P_n) = \frac{p(1-p)}{n}$ , 由 CLT, 近似有  $\frac{P_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$ 。但是, 由于  $p$  未知, 则分母上的标准误未知, 故我们无法直接利用这一分布给出置信区间。记  $\sigma^2 = p(1-p)$ , 下面采用几种不同方法给出其估计  $\hat{\sigma}^2$ 。

1. 用  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  估计  $\sigma^2$ , 于是近似有  $\frac{P_n - p}{\sqrt{\frac{S^2}{n}}} \sim N(0, 1)$ , 对应的置信区间为

$$(P_n - z_{\frac{\alpha}{2}} \sqrt{\frac{S^2}{n}}, P_n + z_{\frac{\alpha}{2}} \sqrt{\frac{S^2}{n}}) \approx (0.542, 0.598).$$

2. 用  $m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = P_n(1 - P_n)$  估计  $\sigma^2$ , 于是近似有  $\frac{P_n - p}{\sqrt{\frac{P_n(1-P_n)}{n}}} \sim N(0, 1)$ , 对

$$\text{应的置信区间为 } (P_n - z_{\frac{\alpha}{2}} \sqrt{\frac{P_n(1-P_n)}{n}}, P_n + z_{\frac{\alpha}{2}} \sqrt{\frac{P_n(1-P_n)}{n}}) \approx (0.542, 0.598).$$

3. 用  $p(1-p)$  的最大值  $\frac{1}{4}$  来估计  $\sigma^2$ , 于是近似有  $\frac{P_n - p}{\frac{1}{2}\sqrt{\frac{1}{n}}} \sim N(0, 1)$ , 对应的置信区间为  $(P_n - z_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}, P_n + z_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}) \approx (0.542, 0.598)$ 。

注意我们这里采用了近似分布, 因此只能说置信水平近似是  $(1-\alpha)$ , 且近似的程度取决于总体分布和样本容量  $n$  的大小。

下面介绍利用 MLE 构建置信区间的方法。

设总体分布的 PDF 或 PMF 为  $f(x; \theta)$ , 有随机样本  $X_1, \dots, X_n$ , 则似然函数  $L(\theta) = \prod_{i=1}^n f(X_i; \theta)$ , 对数似然函数  $\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(X_i; \theta)$ 。

**定理 6.1.** 若  $f$  满足一定的光滑性条件,  $\theta^*$  为  $\theta$  的 MLE, 则存在  $\sigma_n > 0$ , 使得  $\frac{\theta^* - \theta}{\sigma_n} \rightarrow N(0, 1)$ 。

根据 Taylor 展开, 对于  $\theta^*$  附近的  $\theta$ , 有  $0 = \ell'(\theta^*) = \ell'(\theta) + \ell''(\theta)(\theta^* - \theta) + o(\theta^* - \theta)$ , 从而  $\theta^* - \theta \approx -\frac{\ell'(\theta)}{\ell''(\theta)}$ , 即  $\sqrt{n}(\theta^* - \theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)}$ 。

由  $\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial \log f(X_i; \theta)}{\partial \theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}$ , 其中  $f_\theta$  表示  $f$  对  $\theta$  的偏导数, 记  $Y_i = \frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}$ , 则  $Y_1, \dots, Y_n$  独立同分布, 且  $E(Y_i) = E\left(\frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}\right) = \int_{-\infty}^{+\infty} \frac{f_\theta(x; \theta)}{f(x; \theta)} f(x; \theta) dx = \int_{-\infty}^{+\infty} f_\theta(x; \theta) dx = \frac{\partial}{\partial \theta} \int_{-\infty}^{+\infty} f(x; \theta) dx = \frac{\partial}{\partial \theta} 1 = 0$ ,  $\text{Var}(Y_i) = E(Y_i^2) = E\left(\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta}\right)^2\right)$  记作  $I(\theta)$ 。根据 CLT, 我们有  $\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \rightarrow N(0, I(\theta))$ 。

一般地, 我们称  $I_n(\theta) = E((\ell'(\theta))^2) = E\left(\left(\sum_{i=1}^n \frac{\partial \log f(X_i; \theta)}{\partial \theta}\right)^2\right)$  为 Fisher 信息量, 展开得  $I_n(\theta) = \sum_{i=1}^n E\left(\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta}\right)^2\right) + \sum_{i \neq j} E\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta} \frac{\partial \log f(X_j; \theta)}{\partial \theta}\right)$ , 由于  $X_1, \dots, X_n$  独立同分布, 有  $E\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta} \frac{\partial \log f(X_j; \theta)}{\partial \theta}\right) = E\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta}\right) E\left(\frac{\partial \log f(X_j; \theta)}{\partial \theta}\right) = 0$ , 从而  $I_n(\theta) = nI(\theta)$ 。

注意到  $\frac{\partial^2 \log f(X_i; \theta)}{\partial \theta^2} = \frac{\partial}{\partial \theta} \left(\frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}\right) = \frac{f_{\theta\theta}(X_i; \theta)f(X_i; \theta) - f_\theta(X_i; \theta)f_\theta(X_i; \theta)}{f^2(X_i; \theta)} = \frac{f_{\theta\theta}(X_i; \theta)}{f(X_i; \theta)} - \left(\frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}\right)^2$ , 故  $E\left(\frac{\partial^2 \log f(X_i; \theta)}{\partial \theta^2}\right) = E\left(\frac{f_{\theta\theta}(X_i; \theta)}{f(X_i; \theta)} - \left(\frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}\right)^2\right)$ , 其中  $E\left(\frac{f_{\theta\theta}(X_i; \theta)}{f(X_i; \theta)}\right) = \int_{-\infty}^{+\infty} \frac{f_{\theta\theta}(x; \theta)}{f(x; \theta)} f(x; \theta) dx = \int_{-\infty}^{+\infty} f_{\theta\theta}(x; \theta) dx = \frac{\partial}{\partial \theta} \int_{-\infty}^{+\infty} f_\theta(x; \theta) dx = \frac{\partial}{\partial \theta} 0 = 0$ , 即  $E\left(\frac{\partial^2 \log f(X_i; \theta)}{\partial \theta^2}\right) = -E\left(\left(\frac{f_\theta(X_i; \theta)}{f(X_i; \theta)}\right)^2\right) = -I(\theta)$ 。则根据弱大数定律有  $-\frac{1}{n}\ell''(\theta) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(X_i; \theta)}{\partial \theta^2} \xrightarrow{P} I(\theta)$ 。

至此, 有结论  $\sqrt{n}(\theta^* - \theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \rightarrow N(0, \frac{1}{I(\theta)})$ , 即  $\frac{\theta^* - \theta}{\sqrt{\frac{1}{nI(\theta)}}} \rightarrow N(0, 1)$ , 即定理 6.1 中的  $\sigma_n = \sqrt{\frac{1}{nI(\theta)}}$ 。  $\theta$  是未知的, 但构造置信区间时  $I(\theta)$  可以用  $I(\theta^*)$  估计, 即  $\frac{\theta^* - \theta}{\sqrt{\frac{1}{nI(\theta^*)}}} \rightarrow N(0, 1)$ 。

对选举问题,  $f(x; p) = p^x(1-p)^{1-x}$ ,  $I(p) = E\left(\left(\frac{\partial \log f(X_i; p)}{\partial p}\right)^2\right) = E\left(\left(\frac{X_i - p}{p(1-p)}\right)^2\right) = \frac{1}{p(1-p)}$ , 于是  $\frac{P_n - p}{\sqrt{\frac{1}{nI(P_n)}}} = \frac{P_n - p}{\sqrt{\frac{P_n(1-P_n)}{n}}} \rightarrow N(0, 1)$ , 据此构造的置信区间与前面的第二种方法相同。

最后介绍一个近似估计两正态总体的均值差的例子。

**例 6.16.** 设总体为  $X \sim N(\mu_1, \sigma_1^2)$ ,  $Y \sim N(\mu_2, \sigma_2^2)$ ,  $X, Y$  独立,  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$  均未知, 随机样本  $X_1, \dots, X_n; Y_1, \dots, Y_m$ , 则  $\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0, 1)$ 。由于  $\sigma_1^2, \sigma_2^2$  未知, 我们用  $S_1^2, S_2^2$  分别估计

之, 于是近似有  $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}} \sim N(0, 1)$ , 对应  $\mu_1-\mu_2$  的置信区间为  $(\bar{X}-\bar{Y}-z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}, \bar{X}-\bar{Y}+z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}})$ 。

## 6.5 Bayes 估计

Bayes 学派看待世界的视角与频率学派不同。简单来说, 在 Bayes 方法中, 对未知参数  $\theta$  的认知可以由概率分布来刻画, 设对应的随机变量为  $\Theta$ , 则  $\theta$  为  $\Theta$  的实现值。在搜集数据前对  $\Theta$  的分布的认知  $f_{\Theta}(\theta)$  称为先验分布。将试验观测抽象为随机变量  $X$ , 当参数为  $\theta$  时, 观测数据的分布为  $f_{X|\Theta}(x|\theta)$ , 称为样本分布。当观测到数据  $x$  后, 可以利用 Bayes 公式来更新对  $\Theta$  的认知, 得到后验分布  $f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)}$ 。这样, 我们就可以利用后验分布来对  $\Theta$  进行推断。

**例 6.17.** 某枚硬币正面向上的概率为未知参数  $\theta$ , 设先验分布为  $f_{\Theta}(\theta) = 1$  ( $\theta \in (0, 1)$ ) (无信息先验, 体现了所谓的同等无知原则, 是 Bayes 统计常用假设)。现抛硬币  $n$  次, 观测到正面向上的次数为  $x$ 。

记  $X$  为  $n$  次中正面向上的次数, 则给定  $\theta$  时,  $X \sim B(n, \theta)$ , 即样本分布  $f_{X|\Theta}(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$  ( $x = 1, \dots, n$ )。于是  $X$  与  $\Theta$  的联合分布为  $f(x, \theta) = f_{X|\Theta}(x|\theta)f_{\Theta}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ ,  $X$  的边缘 PMF 为  $f_X(x) = \int_0^1 f(x, \theta)d\theta = \binom{n}{x} \int_0^1 \theta^x(1-\theta)^{n-x}d\theta = \binom{n}{x} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{1}{n+1}$ , 则后验分布为  $f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{f_X(x)} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}$ , 即  $\Theta|X = x \sim \beta(x+1, n-x+1)$ 。

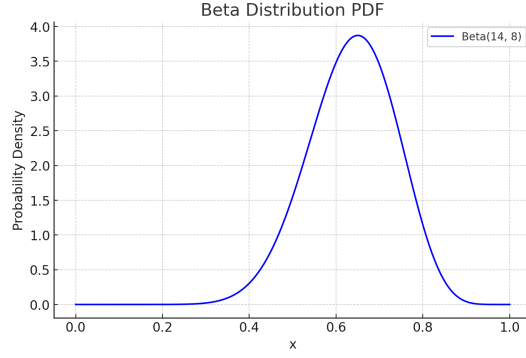
其中,  $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$  为 Gamma 函数, 满足  $\Gamma(x+1) = x\Gamma(x)$ ,  $\Gamma(1) = 1$ , 对于正整数  $n$ ,  $\Gamma(n+1) = n!$ 。而  $\beta(a, b)$  表示 Beta 分布, 其 PDF 为  $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$  ( $x \in (0, 1)$ )。若  $X \sim \beta(a, b)$ , 则  $E(X) = \frac{a}{a+b}$ ,  $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$ 。均匀分布  $U(0, 1)$  即  $\beta(1, 1)$  分布。

上例中, 若  $n = 20, x = 13$ , 则后验分布为  $\beta(14, 8)$ , 其 PDF 图象如 6.1。计算可知,  $P(\Theta > \frac{1}{2}) \approx 0.91$ , 而  $\Theta < \frac{1}{4}$  的可能性很小。

已知了后验分布后, 如何给出参数  $\theta$  的合理估计呢? 常用方法如:

1. 后验众数  $\hat{\theta}_1$ , 即  $\beta(x+1, n-x+1)$  的 PDF 最大值点  $\frac{x}{n}$  (恰与 MLE 一致)
2. 后验均值  $\hat{\theta}_2 = E(\Theta|X = x) = \frac{x+1}{n+2}$
3. 后验中位数  $\hat{\theta}_3$

上例中还可以进一步证明, 若选取先验为  $\beta(a, b)$ , 则后验分布为  $\beta(x+a, n-x+b)$ , 此时后验均值为  $\frac{x+a}{n+a+b} = \frac{a+b}{n+a+b} \frac{a}{a+b} + \frac{n}{n+a+b} \frac{x}{n}$ , 即后验均值是先验均值  $\frac{a}{a+b}$  与样本均值  $\frac{x}{n}$  的加权平均, 权重分别为  $\frac{a+b}{n+a+b}$  和  $\frac{n}{n+a+b}$ 。

图 6.1:  $\beta(14, 8)$  的 PDF 图象

Bayes 方法根据后验分布给出区间估计, 称之为 可信区间。具体来说, 就是要找到  $a, b \in \mathbb{R}$ , 使  $P(a < \Theta < b | X = x) \geq 1 - \alpha$ 。具体的选取方式如:

1. 最大后验区间 (通常用于单峰情形), 可以直观理解为用一条平行于横轴的线自上而下扫描, 直到截取后验 PDF 的面积为  $(1 - \alpha)$
2. 等尾区间, 即令  $P(\Theta < a | X = x) = P(\Theta > b | X = x) = \frac{\alpha}{2}$

**例 6.18.** 设总体分布为  $X \sim N(\mu, \sigma^2)$ , 其中  $\sigma^2$  已知, 有随机样本  $X_1, \dots, X_n$ , 取  $\mu$  的先验分布  $f(\mu) \propto 1$  (无信息先验, 这不是一个合理的分布, 理解为一种广义 PDF), 则样本分布为  $f(x_1, \dots, x_n | \mu) \propto \prod_{i=1}^n e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$ , 后验分布为  $f(\mu | x_1, \dots, x_n) \propto f(x_1, \dots, x_n | \mu) f(\mu) \propto \prod_{i=1}^n e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \propto e^{-\frac{n\mu^2 - 2\mu \sum_{i=1}^n x_i}{2\sigma^2}} \propto e^{-\frac{n(\mu - \bar{x})^2}{2\sigma^2}}$ , 即  $\mu$  的后验分布为  $N(\bar{x}, \frac{\sigma^2}{n})$ 。