



概率论笔记

作者：肖程哲

时间：July 10, 2022



苟日新，日日新，又日新

目录

第 1 章 概率基础	1
1.1 概率空间	1
1.2 古典概型与几何概型	3
1.2.1 古典概型	3
1.2.2 几何概型	3
1.3 条件概率	3
第 2 章 随机变量	5
2.1 随机变量的分布	5
2.2 多元随机变量	7
2.2.1 边际分布	8
2.2.2 独立	8
2.2.3 条件分布	9
2.3 随机变量的函数	10
2.3.1 分布函数法	10
2.3.2 Copula	11
2.3.3 概率密度函数法	12
2.3.4 矩母函数法	12
2.3.5 次序统计量	12
第 3 章 随机变量的数值特征	15
3.1 期望值	15
3.2 矩母函数与特征函数	15
3.3 熵与信息	15
第 4 章 常见分布	16
4.1 离散分布	16
4.2 连续分布	16
4.3 正态分布及其导出分布	16
第 5 章 概率极限	17
5.1 收敛	17
5.2 大数定理	17
5.3 中央极限定理	17
第 A 章 基本数学工具	18
A.1 排列与组合	18

第1章 概率基础

内容提要

- | | |
|------------------------------------|-----------------------------------|
| <input type="checkbox"/> 事件 | <input type="checkbox"/> 乘法法则 |
| <input type="checkbox"/> 古典概型与几何概率 | <input type="checkbox"/> 全概率公式 |
| <input type="checkbox"/> 条件概率与独立 | <input type="checkbox"/> Bayes 法则 |

1.1 概率空间

定义 1.1 (样本空间)

随机试验可能出现的结果称为样本点 (sample point), 用 ω 表示。样本的全体构成样本空间 (sample space), 用 Ω 表示。



定义 1.2 (事件的古典定义)

样本点 ω 的集合称为事件 (event)。



我们关心的随机现象被抽象为集合, 逻辑运算 (且, 或, 非, etc.) 对应成集合论运算 (交, 并, 补, etc.)。

性质 集合的运算性质:

交换律 $A \cup B = B \cup A, AB = BA$

结合律 $(A \cup B) \cup C = A \cup (B \cup C), (AB)C = A(BC)$

分配律

$$(A \cup B) \cup C = A \cup (B \cup C), \quad (1.1)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C). \quad (1.2)$$

对偶律 (De Morgan's laws)

$$\text{事件并的对立等于对立的交: } \overline{A \cup B} = \overline{A} \cap \overline{B}, \quad (1.3)$$

$$\text{事件交的对立等于对立的并: } \overline{A \cap B} = \overline{A} \cup \overline{B}. \quad (1.4)$$

为方便概率的定义, 并不把 Ω 的一切子集作为事件, 应避免不可测集的出现。

定义 1.3 (事件域)

事件构成的全体称为事件域 \mathcal{F} , 是 Ω 的子集族 (collection of subsets), 应满足 σ 代数 的要求:

- $\emptyset \in \mathcal{F}$, 无事发生;
- $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, 即 \mathcal{F} 对补集运算 (逻辑上的非) 封闭;
- $A_1, \dots, A_n, \dots \in \mathcal{F} \implies \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$, 即 \mathcal{F} 对可数交运算 (逻辑上的可数多个且) 封闭.



笔记 可数性是为了在数学上能够恰当地处理无穷的概念, 术语中的 σ 指的就是可数并。由对偶原理可得 σ 域同时对可数并运算封闭. 即 σ 域对逆, 并, 交, 差的可数次运算封闭。

事件域根据问题的不同要求适当选取, 在概率定义没有困难时, 应尽量选大, 通常以 Ω 的一切子集作为事件域. 当 Ω 给定后, 若某些子集必须作为事件处理, 能否找到包含他们的 σ 域?

命题 1.1

若给定 Ω 的一个非空集族 \mathcal{G} , 必存在 Ω 上唯一的 σ 域 $m(\mathcal{G})$, 满足下列性质:

- 包含 \mathcal{G}
- 若其他 σ 域包含 \mathcal{G} , 则必包含 $m(\mathcal{G})$

这个 $m(\mathcal{G})$ 称为包含 \mathcal{G} 的最小 σ 域, 或由 \mathcal{G} 扩张而成的 σ 域.



扩张, 或者称为延拓, 是数学中很重要的一个概念, 大抵是将某映射的定义域适当扩大, 不改变在初始定义域上的映射取值(注意值域可能是比较抽象的集合, 配备了某些操作之后被称为空间), 同时在扩大后的定义域上仍然保持某些优良的性质. 与此相对的概念是限制, 即关心局部上可能更加漂亮的性质, 把初始的定义域适当缩小.

证明 由于 Σ 的一切子集构成的集类包含 \mathcal{G} , 所以 m 存在. 再取 Σ 上满足此条件的 σ 域之交作为 $m(\mathcal{G})$ 即可.

特别地, 实数集 \mathbb{R} 的子集族 $\{(-\infty, x] : x \in \mathbb{R}\}$ 生成的 σ 代数 $\mathcal{B}_{\mathbb{R}}$ 称为 \mathbb{R} 上的 *Borel* 代数.

定义 1.4 (Borel 集)

设 \mathbb{R}^1 为全集, 形为 $[a, b)$ 构成的集类产生的 σ 域称为**一维 Borel** σ 域, 记为 \mathcal{B}_1 , 其中的元素称为**一维 Borel** 集



若 x, y 为任意实数, 由于:

$$\{x\} = \bigcap_{n=1}^{\infty} \left[x, x + \frac{1}{n} \right)$$

$$(x, y) = [x, y) - \{x\}$$

$$[x, y] = [x, y) + \{y\}$$

$$(x, y] = [x, y) + \{y\} - \{x\}$$

因此 \mathcal{B}_1 包含一切开区间, 闭区间, 单个实数, 可列个实数, 以及他们经可列次逆, 并, 交, 差运算的集合.

定义 1.5 (概率空间)

定义在事件域(非样本空间)上的集合函数 $P: \mathcal{F} \rightarrow \mathbb{R}$ 称为**概率**的条件是:

非负性 $P(A) \geq 0, \forall A \in \mathcal{F}$

规范性 $P(\Omega) = 1$; (如果没有这条就是一般的**有限测度**)

可列可加性 若 $A_1, \dots, A_n, \dots \in \mathcal{F}$ 两两不交, 即 $A_i \cap A_j = \emptyset, \forall i \neq j$, 则 $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

我们称 (Ω, \mathcal{F}, P) 为一个**概率空间** (probability space)



性质 概率的性质:

- $P(\Omega) = 1$;
- $P(A^c) = 1 - P(A)$;
- 若 $A \subseteq B$ 则 $P(A) \leq P(B)$;

推论 1.1 (加法公式)

基础形式:

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

一般形式:

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{i=1, \dots, n} P(A_i) - \sum_{\substack{i < j \\ i, j = 1, \dots, n}} P(A_i A_j) \\ &+ \sum_{\substack{i < j < k \\ i, j, k = 1, \dots, n}} P(A_i A_j A_k) - \dots + (-1)^{n-1} P(A_1 A_2 \dots A_n) \end{aligned}$$

特别地, 若事件出现个数相同时概率相等, 则可简化为:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = nP_1 - \binom{n}{2}P_2 + \binom{n}{3}P_3 - \dots + (-1)^{n-1}P_n$$



显然, 可列可加性可以推出有限可加性. 但是一般来讲, 由有限可加性并不能推出可列可加性. 设 $A_i \in \mathcal{F}, i = 1, 2, \dots$ 且两两互不相容, 若希望由有限可加性推出可列可加性, 则需要下式成立:

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

定义 1.6

对于 \mathcal{F} 上的集合函数 P , 若它对 \mathcal{F} 中任何一个单调不减的集序列 $\{S_n\}$ (即 $S_n \in \mathcal{F}, S_n \subseteq S_{n+1}$) 均满足:

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

则称它是下连续的.



故若令 $S_n = \sum_{i=1}^n A_i$, 则可列可加性条件等价于有限可加性加下连续.

1.2 古典概型与几何概型

1.2.1 古典概型

1.2.2 几何概型

古典概型的基本思想:

有限个样本点 所涉及的随机现象只有有限个样本点, 譬如为 n 个, 且这些事件是两两互不相容的;

等可能性 每个样本点发生的可能性相等

定义 1.7

若事件 A 含有 k 个样本点, 则事件 A 的概率为

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



笔记 事实上, 古典概型的大部分问题都能形象化地用摸球模型来描述以后我们经常研究摸球模型, 意义即在于此.

1.3 条件概率

定义 1.8 (条件概率)

令 $A, B \in \mathcal{F}$, 且 $P(B) > 0$ 称

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

为基于于 B 的条件概率 (probability conditional on B), 这仍然是一个概率测度.



定理 1.1 (乘法法则)

令 $A, B \in \mathcal{F}$, 且 $P(B) > 0$, 则

$$P(A \cap B) = P(A|B)P(B)$$

泛化后有:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\dots$$

**定理 1.2 (全概率公式)**

设 B_1, B_2, \dots, B_n 为样本空间 Ω 的一个分割, 且互不相容, 即 $\bigcup_{i=1}^n B_i = \Omega, B_i B_j = \emptyset \text{ for } i \neq j$ 如果 $P(B_i) > 0, i = 1, 2, \dots, n$, 则对任一事件 A 有

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i). \quad (1.5)$$



笔记 $P(A|B_i)$ 可视为事件 A 在 B_i 上的平均, $P(B_i)$ 则为其权重.

定理 1.3 (Bayes 定理)

设 B_1, B_2, \dots, B_n 为样本空间 Ω 的一个分割, 且互不相容, 即 $\bigcup_{i=1}^n B_i = \Omega, B_i B_j = \emptyset \text{ for } i \neq j$ 如果 $P(A) > 0, P(B_i) > 0, i = 1, 2, \dots, n$, 则

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^n P(B_j)P(A|B_j)} \quad (1.6)$$

**定义 1.9 (事件的独立性)**

如果 $A, B \in \mathcal{F}$ 满足

$$P(A \cap B) = P(A)P(B),$$

则称 A 与 B 独立 (independent), 记为 $A \perp\!\!\!\perp B$.

对于事件集 A_1, A_2, \dots, A_n , 若对于其中任意子集 $A_{i_1}, A_{i_2}, \dots, A_{i_n}$ 有:

$$P(A_{i_1} \cap \dots \cap A_{i_n}) = P(A_{i_1}) \cdots P(A_{i_n})$$

则称此事件集相互独立 (mutually independent)



笔记 当 $P(A) > 0$ 时, 我们有 $P(B|A) = P(B) \iff B \perp\!\!\!\perp A$, 由此可得到 B 独立于 A 的直观理解

性质 独立性是对称的, 即 $A \perp\!\!\!\perp B \iff B \perp\!\!\!\perp A$. 若两事件独立, 则其补集也独立.

$$\begin{array}{ccc} A & \longleftrightarrow & B \\ & \diagup \diagdown & \\ & & \\ A^c & \longleftrightarrow & B^c \end{array}$$

定义 1.10 (事件域的独立性)

若 $\mathcal{G} \subset \mathcal{F}$ 与 $\mathcal{H} \subset \mathcal{F}$ 满足

$$A \perp\!\!\!\perp B, \quad \forall A \in \mathcal{G}, B \in \mathcal{H},$$

则称 \mathcal{G} 与 \mathcal{H} 独立, 记为 $\mathcal{G} \perp\!\!\!\perp \mathcal{H}$



测度论告诉我们一个重要结果: 如果 \mathcal{G} 对交集运算封闭, 那么成立 $\mathcal{G} \perp\!\!\!\perp \mathcal{H} \implies \sigma(\mathcal{G}) \perp\!\!\!\perp \mathcal{H}$

第2章 随机变量

内容提要

- 离散与连续随机变量
- 一元与多元
- cdf, pmf, pdf
- 条件分布
- 独立随机变量
- 随机变量函数的分布
- 次序随机变量

在概率论中, 主要关心 X 取值于数值集合 \mathcal{X} 中某个子集 B 的可能性, 即希望得到 $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$. 概率论不关心具体的样本点 $\omega \in \Omega$, 将其记为 $\{X \in B\} = X^{-1}(B)$. 由于 \mathbb{P} 定义在 \mathcal{F} 上, 故需 $X^{-1}(B) \in \mathcal{F}$.

定义 2.1 (可测性)

设所有值得关心的 $B \subset \mathcal{X}$ 组成 $\mathcal{F}_{\mathcal{X}}$, 且 $\forall B \in \mathcal{F}_{\mathcal{X}}$ 都满足 $\{X \in B\} \in \mathcal{F}$, 则称 X 为 $\mathcal{F}/\mathcal{F}_{\mathcal{X}}$ 可测的. 当 $\mathcal{F}_{\mathcal{X}}$ 不引起混淆时, 简记为关于 \mathcal{F} 可测, 写作 $X \in \mathcal{F}$.



由于原像保持交、并、补等集合运算, 且 \mathcal{F} 是 σ 代数, 可将 $\mathcal{F}_{\mathcal{X}}$ 扩张为合适的最小的 σ 代数, 即 $\sigma(\mathcal{F}_{\mathcal{X}})$, 因此可测映射的定义不妨只考虑 $\mathcal{F}_{\mathcal{X}}$ 是 σ 代数的情况.

定义 2.2 (随机变量)

为了表示因随机性而变动的量, 称可测映射(measurable mapping)

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{X}, \mathcal{F}_{\mathcal{X}}), \quad \omega \in \Omega \mapsto X(\omega) \in \mathcal{X}$$

为随机元 (random element), 也称随机变量 (random variable). 其中 $\mathcal{F}_{\mathcal{X}}$



由于只考虑 $\mathcal{F}_{\mathcal{X}}$ 是 σ 代数的情况, 可将随机变量看作将原概率空间映射到新概率空间的方式. 新样本空间由 Borel 点集 构成, 对应的概率测度等于原像的.

注 使用随机变量 X 时, 有两个可能的含义:

- X 的 (随机) 取值
- X 的分布

定义 2.3 (离散与连续随机变量)

当 \mathcal{X} 是(至多可数的)离散点集, $\mathcal{F}_{\mathcal{X}}$ 由 \mathcal{X} 的所有子集组成, 则称其为离散随机变量 (discrete random variable).

当随机变量 $\mathcal{X} = \mathbb{R}^n$, 考虑 $\mathcal{F}_{\mathcal{X}}$ 为 $\{\prod_{i=1}^n (-\infty, x_i] : x_1, \dots, x_n \in \mathbb{R}\}$ 生成的 Borel 代数 (最小的 σ 代数), 则称其为连续随机变量 (continuous random variable).



2.1 随机变量的分布

定义 2.4

称随机元 X 诱导的概率测度

$$\mathbb{P}\{X \in \bullet\}, \bullet \in \mathcal{F}_{\mathcal{X}}$$

为 X 的概率分布 (distribution/law)



注 对于随机变量, 他的取值是随机的, 但他的分布是固定的

定义 2.5 (单变量分布函数)

一个函数 $F : \mathbb{R} \rightarrow [0, 1]$ 称为一个单变量分布函数, 当其满足以下性质时:

单调性 $F(x_1) \leq F(x_2), \forall x_1 < x_2$

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

有界性 $\lim_{n \rightarrow -\infty} F(x) = 0, \lim_{n \rightarrow \infty} F(x) = 1$



性质 $F(x)$ 最多只有可数个间断点

命题 2.1

对每个分布 $Q : \mathcal{B}_1 \rightarrow [0, 1]$ 都存在唯一一个分布函数 $F_Q : \mathbb{R} \rightarrow [0, 1]$ 使得 $F_Q(x) = Q[(-\infty, x)], \forall x \in \mathbb{R}$ 成立。

**命题 2.2**

对每个分布函数 $F : \mathbb{R} \rightarrow [0, 1]$ 都存在唯一一个分布 $Q_F : \mathcal{B}_1 \rightarrow [0, 1]$ 使得 $Q_F[(-\infty, x)] = F(x), \forall x \in \mathbb{R}$ 成立。

**定理 2.1**

分布函数可以唯一决定概率分布, 即:

$$Q_{F_Q} = Q, \quad F_{Q_F} = F$$

把随机变量 X 服从分布函数 $F(x)$ 简记作 $X \sim F(x)$



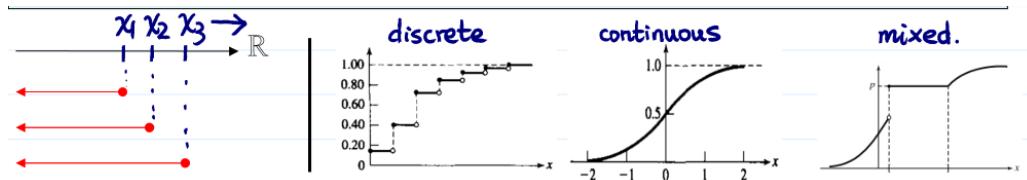
	离散	连续
一元随机变量	概率质量函数 (pmf)	概率密度函数 (pdf)
	累积分布函数 (cdf)	
	矩母函数/特征函数 (mgf/chf)	
多元随机变量	联合概率质量函数 (joint pmf)	联合概率密度函数 (joint pdf)
	联合累积分布函数 (joint cdf)	
	联合矩母函数/特征函数 (joint mgf/chf)	

定义 2.6 (累积分布函数)

此时 $X = (X_1, \dots, X_n)^\top$ 的分布由(累积)分布函数 (cumulative distribution function, c.d.f.)

$$F_X(x_1, \dots, x_n) = \mathbb{P}\{X_1 \leq x_1, \dots, X_n \leq x_n\}, \quad x_1, \dots, x_n \in \mathbb{R}.$$

唯一刻画. 把随机变量 X 服从分布函数 $F(x)$ 简记作 $X \sim F(x)$

**定义 2.7 (概率质量函数)**

当且仅当函数 $p(x)$ 满足下述条件时, 被称为概率质量函数 (probability mass function, p.m.f.):

- $p(x) \geq 0$
- $\sum_{x \in \mathcal{X}} p(x) = 1$



当 X 是离散型随机变量, 设 \mathcal{F}_X 由 \mathcal{X} 的所有子集组成, 此时 X 的分布由

$$p_X(x) = \mathbb{P}\{X = x\} = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}), \quad x \in \mathcal{X}$$

唯一刻画. 其与分布函数间的关系为:

- $F(x) = \sum_{t \leq x} P(X = t) = \sum_{t \leq x} p(t)$
- $p(x) = P(X = x) = F(x) - F(x-)$

定义 2.8 (概率密度函数)

当且仅当函数 $f(x)$ 满足下述条件时, 被称为概率密度函数 (probability density function, p.d.f.):

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x) dx = 1$



当 X 是连续型随机变量, 且 $F_X : \mathbb{R}^n \rightarrow [0, 1]$ 可微 (或者更一般地, 绝对连续), 此时 X 的分布由

$$f_X := \frac{\partial^n F_X}{\partial x_1 \cdots \partial x_n}$$

唯一刻画. 其与分布函数间的关系为:

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

注 即使对于 $f(x) > 0$ 的 x , $P(X = x) = x \int_x^x f(t) dt = 0$, 即连续型随机变量在实轴上任意一点的概率测度为零. 概率密度函数 $f(x)$ 代表的是在此位置上单位长度的概率, 可能是一个很大的值.

2.2 多元随机变量

定义 2.9 (随机向量)

若随机变量 $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$ 定义在同一概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上, 则称

$$X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$$

构成一个 n 维随机向量, 亦称 n 维随机变量.



命题 2.3

若 B_n 为 \mathbb{R}^n 上任一博雷尔点集, 有

$$\{X(\omega) \in B_n\} \in \mathcal{F}$$



定义 2.10

称 n 元函数

$$F(x_1, x_2, \dots, X_n) = \mathbb{P}\{X_1(\omega) < x_1, X_2(\omega) < x_2, \dots, X_n(\omega) < x_n\}$$

为随机向量 $X(\omega)$ 的联合分布函数 (joint cdf).



当 $n = 2$ 时, 有

$$\mathbb{P}((a_1, b_1) \leq X < (a_2, b_2)) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \quad (2.1)$$

性质 多元分布函数的一些性质:

1. 单调性: 关于每个变元是单调不减函数;

2.

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

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

3. 关于每个变元右连续.

4. 在二元场合, 还应该有: 对任意 $a_1 < b_1, a_2 < b_2$, 都有

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

为保证2.1式中的概率的非负性, 性质4是必须的, 而且由性质4可以推出单调性, 但存在着反例说明, 由单调性并不能保证性质4的成立(见习题12). 这是多元场合与一元场合的不同之处.

2.2.1 边际分布

定义 2.11

对于多维随机变量 X , 只考虑其中一个分量的分布时, 称其为 X 的 **边际分布或边缘分布**. 对于分量 X_i , 其**边缘分布函数** (marginal cdf) 为:

$$F_{X_i}(x_i) = \mathbb{P}\{X_i \leq x_i\} = F(\infty, \dots, x_i, \dots, \infty)$$

2.2.2 独立

定义 2.12 (独立随机变量)

若随机变量 $X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega))$ 联合分布函数可分解成各分量边缘分布函数的乘积, 即:

$$F(x_1, x_2, \dots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n), \quad \forall x_1, x_2, \dots, x_n \in \mathbb{R}$$

则称随机变量 X 各分量相互独立

注 对于一般的多元随机变量, 其各分量边缘分布不足以描述联合分布的情况. 但若其各分量独立则可以.

定理 2.2

对于连续情况:

$$\begin{aligned} F(x_1, x_2, \dots, x_n) &= F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n) \\ \Leftrightarrow f(x_1, x_2, \dots, x_n) &= f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n) \end{aligned}$$

对于离散情况:

$$\begin{aligned} F(x_1, x_2, \dots, x_n) &= F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n) \\ \Leftrightarrow p(x_1, x_2, \dots, x_n) &= p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n) \end{aligned}$$

定理 2.3

若随机变量 X, Y 独立, 则其变换 $Z = g(X), W = h(Y)$ 也独立.

泛化情况: 若随机向量 $\{X\}_n$ 各分类独立, 则其变换 $\{Y\}_n = g(\{X\}_n)$ 各分类也独立.

2.2.3 条件分布

定义 2.13

对一切使 $P(Y = y_j) = p_{\cdot j} = \sum_{i=1}^{+\infty} p_{ij} > 0$ 的 y_j , 称

$$p_{i|j} = P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}}, \quad i = 1, 2, \dots \quad (2.2)$$

为给定 $Y = y_j$ 条件下 X 的条件分布列. 若 $p_X(x) = 0$, 则定义其为 0.



设二维连续随机变量 (X, Y) 的联合密度函数为 $p(x, y)$, 边际密度函数为 $p_X(x), p_Y(y)$.

在离散随机变量场合, 其条件分布函数为 $P(X \leq x | Y = y)$. 但是, 因为连续随机变量取某个值的概率为零, 即 $P(Y = y) = 0$, 所以无法用条件概率直接计算 $P(X \leq x | Y = y)$, 一个很自然的想法是: 将 $P(X \leq x | Y = y)$ 看成是 $h \rightarrow 0$ 时 $P(X \leq x | y \leq Y \leq y + h)$ 的极限, 即

$$\begin{aligned} P(X \leq x | Y = y) &= \lim_{h \rightarrow 0} P(X \leq x | y \leq Y \leq y + h) \\ &= \lim_{h \rightarrow 0} \frac{P(X \leq x, y \leq Y \leq y + h)}{P(y \leq Y \leq y + h)} \\ &= \lim_{h \rightarrow 0} \frac{\int_{-\infty}^x \int_y^{y+h} p(u, v) dv du}{\int_y^{y+h} p_Y(v) dv} \\ &= \lim_{h \rightarrow 0} \frac{\int_{-\infty}^x \left\{ \frac{1}{h} \int_y^{y+h} p(u, v) dv \right\} du}{\frac{1}{h} \int_y^{y+h} p_Y(v) dv}. \end{aligned}$$

当 $p_Y(y), p(x, y)$ 在 y 处连续时, 由积分中值定理可得

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{1}{h} \int_y^{y+h} p_Y(v) dv &= p_Y(y), \\ \lim_{h \rightarrow 0} \frac{1}{h} \int_y^{y+h} p(u, v) dv &= p(u, y). \end{aligned}$$

所以

$$P(X \leq x | Y = y) = \int_{-\infty}^x \frac{p(u, y)}{p_Y(y)} du.$$

至此, 我们可以定义连续随机变量的条件分布如下.

定义 2.14

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

$$F(x|y) = \int_{-\infty}^x \frac{p(u, y)}{p_Y(y)} du, \quad (2.3)$$

$$p(x|y) = \frac{p(x, y)}{p_Y(y)}. \quad (2.4)$$

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

$$F(y|x) = \int_{-\infty}^y \frac{p(x, v)}{p_X(x)} dv \quad (2.5)$$

$$p(y|x) = \frac{p(x, y)}{p_X(x)}. \quad (2.6)$$



注 对于每一个固定的 x , $p_{Y|X}(y|x)$ 是一个关于 y 的概率质量函数; $f_{Y|X}(y|x)$ 是一个关于 y 的概率密度函数与概率三定理的对应:

乘法法则 $p_{XY}(x, y) = p_{Y|X}(y|x)p_X(x)$, $f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x)$

全概率公式 $p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x)$, $f_Y(y) = \int_{-\infty}^{+\infty} f_{Y|X}(y|x)f_X(x)dx$

$$\text{Bayes 原理 } p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_x p_{Y|X}(y|x)p_X(x)}, \quad f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x)f_X(x)dx}$$

2.3 随机变量的函数

在统计学中，常需要转化原始数据以获取其中信息，由此引出了研究随机变量的函数的需要。

定理 2.4 (事件法)

设 $Y = g(X)$ 是随机向量 $X = (X_1, X_2, \dots, X_n)$ 的函数，则 Y 的分布由 X 的分布通过下式决定：

$$\mathbb{P}\{Y \in B\} = \mathbb{P}\{X \in A\}, \quad A = \{\omega | g(\omega) \in B\}$$



此法是其他方法的基础，但使用不便，常用于离散随机变量。

例题 2.1 已知随机变量 X, Y 的联合概率质量函数为 $p(x, y)$ ，求 $Z = X + Y$ 的分布

$$p_Z = P(Z = z) = P(X + Y = z) = \sum_{x=-\infty}^{\infty} p(x, z - x)$$

若 X, Y 独立，则 $p_Z(z) = \sum_{x=-\infty}^{\infty} p_X(x)p_Y(z - x)dx$ ，为 p_X 与 p_Y 的卷积

2.3.1 分布函数法

通过下式获取随机变量的函数的分布：

$$F_Y(y) = \begin{cases} \int_{A_y} f_X(x)dx & , \quad A_y = \{x | g(x) \leq y\} \\ \sum_{x \in A_y} p_X(x) & \end{cases}$$

对每一个变换分别运用上式则可得到向量函数的分布。

例题 2.2 已知随机变量 X 的概率密度函数 $f_X(x)$ 与分布函数 $F_X(x)$ ，求 $Y = X^2$ 的分布

$$F_Y(y) = P(Y \leq y) = P(-\sqrt{y} < X \leq \sqrt{y}) + P(X = \sqrt{y} = F_X(\sqrt{y}) - F_X(-\sqrt{y})) + 0 \quad y \geq 0$$

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(\sqrt{y}) - \frac{d}{dy} F_X(\sqrt{y}) \\ &= f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} - f_X(-\sqrt{y}) \frac{-1}{2\sqrt{y}} \\ &= \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})) \end{aligned}$$

例题 2.3 已知随机变量 X, Y 的联合概率密度函数 $f(x, y)$ ，求 $Z = X + Y$ 的分布

$$F_Y(y) = P(Y \leq y) = P(-\sqrt{y} < X \leq \sqrt{y}) + P(X = \sqrt{y} = F_X(\sqrt{y}) - F_X(-\sqrt{y})) + 0 \quad y \geq 0$$

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(X + Y \leq z) \\ &= \iint_{x+y \leq z} f(x, y)dxdy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z-x} f(x, y)dxdy \\ &= \int_{-\infty}^z \int_{-\infty}^{\infty} f(x, v-x)dxdv, \quad y = v - x \\ f_Z(z) &= \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f(x, z-x)dx \end{aligned}$$

若 X, Y 独立，则 $f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$ ，为 f_X 与 f_Y 的卷积，与 2.1 类似

例题 2.4 已知随机变量 X, Y 的联合概率密度函数 $f(x, y)$, 求 $Z = \frac{Y}{X}$ 的分布

$$Q_z = \{(x, y) : y/x \leq z\} = \{(x, y) : x < 0, y \geq zx\} \cup \{(x, y) : x > 0, y \leq zx\}$$

$$\begin{aligned} F_Z(z) &= \iint_{Q_z} f(x, y) dx dy = \int_{-\infty}^0 \int_{xz}^{\infty} + \int_0^{\infty} \int_{-\infty}^{xz} f(x, y) dy dx \\ &= \int_{-\infty}^0 \int_z^{-\infty} + \int_0^{\infty} \int_{-\infty}^z xf(x, xv) dv dx \quad (\text{set } y = xv) \\ &= \int_{-\infty}^0 \int_{-\infty}^z (-x)f(x, xv) dv dx + \int_0^{\infty} \int_{-\infty}^z xf(x, xv) dv dx \\ &= \int_{-\infty}^z \int_{-\infty}^{\infty} |x|f(x, xv) dx dv \end{aligned}$$

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} |x|f(x, xz) dx$$

2.3.2 Copula

定义 2.15

设连续型实值随机变量 X 有分布函数 F , 易见 F 在 $\bar{\mathbb{R}} = [-\infty, +\infty]$ 上从 0 递增到 1. 定义相应的分位数函数 (quantile function) 为

$$F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad p \in [0, 1].$$



注 当 F 严格递增时, 这与一般的反函数定义相同.

定理 2.5

设连续型实值随机变量 X 有分布函数 F , 则 $F(X) \sim \text{Uniform}([0, 1])$



证明

$$\mathbb{P}\{F(X) \leq p\} = \mathbb{P}\{X \leq F^{-1}(p)\} = F(F^{-1}(p)) = p, \quad \forall p \in [0, 1].$$

定理 2.6

设连续型实值随机变量 X 有分布函数 F , 且设 $U \sim \text{Uniform}([0, 1])$, 则 $F^{-1}(U) \stackrel{d}{=} X$, 其中 $\stackrel{d}{=}$ 表示分布相同 (equal in distribution).



定理 2.7 (Sklar 定理)

考虑多个连续型实值随机变量 X_1, \dots, X_k , 记 X_i 的分布函数为 F_i . 我们称 $(F_1(X_1), \dots, F_k(X_k))$ 的分布函数 $C : [0, 1]^k \rightarrow [0, 1]$ 为相应的**Copula**, 适合

$$C(F_1(x_1), \dots, F_k(x_k)) = \mathbb{P}\{X_1 \leq x_1, \dots, X_k \leq x_k\}, \quad \forall x_1, \dots, x_k$$

这个结果称为**Sklar 定理**



这个结果, 在金融统计中有颇多应用. 稍作诠释的话, Copula 提取了变量间的相关性, 通过粘合边际能够恰好地表示总体. 人们可以构造各种各样的 Copula, 对真实世界进行建模.

2.3.3 概率密度函数法

定理 2.8

设连续随机变量 X 的概率密度函数为 $f_X(x)$. 令 $Y = g(X)$, 其中 g 为可微函数, 且严格单调, 则当 $y = g(x)$ 有定义时:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

否则为 0

若 g 为分段单调函数, 则分段计算上是结果, 再进行相加



定理 2.9

设连续随机向量 \mathbf{X} 的概率密度函数为 $f_{\mathbf{X}}(\mathbf{x})$. 令 $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, 其中 \mathbf{g} 为双射, 定义其逆函数为:

$$\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y}) = \mathbf{w}(\mathbf{y})$$

若 \mathbf{w} 存在连续偏导数, 则当 $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ 有定义时:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{y})) \left| \frac{\partial \mathbf{w}}{\partial \mathbf{y}} \right|$$

否则为 0



注 若 \mathbf{Y} 的维数 k 小于 \mathbf{X} 的维数 n , 可增补 $n - k$ 维的函数 $\mathbf{Z} = \mathbf{h}(\mathbf{X})$, 使得 (\mathbf{Y}, \mathbf{Z}) 满足条件, 再通过积分获取 \mathbf{Y} 的概率密度函数.

例题 2.5 已知随机变量 X_1, X_2 的联合概率密度函数 $f(x_1, x_2)$, 求 $Y_1 = \frac{X_2}{X_1}$ 的分布

令 $Y_2 = X_1$, 则:

$$\begin{aligned} x_1 &= y_2 \equiv w_1(y_1, y_2) \\ x_2 &= y_1 y_2 \equiv w_2(y_1, y_2) \\ \frac{\partial w_1}{\partial y_1} &= 0, \frac{\partial w_1}{\partial y_2} = 1, \frac{\partial w_2}{\partial y_1} = y_2, \frac{\partial w_2}{\partial y_2} = y_1 \\ J &= \begin{vmatrix} 0 & 1 \\ y_2 & y_1 \end{vmatrix} = -y_2 \end{aligned}$$

所以:

$$\begin{aligned} f_{Y_1 Y_2}(y_1, y_2) &= f_{X_1 X_2}(y_2, y_1 y_2) |y_2| \\ f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{Y_1 Y_2}(y_1, y_2) dy_2 = \int_{-\infty}^{\infty} f_{X_1 X_2}(y_2, y_1 y_2) |y_2| dy_2 \end{aligned}$$

2.3.4 矩母函数法

2.3.5 次序统计量

定义 2.16

设 X_1, X_2, \dots, X_n 为随机变量, 将其按大小排序后记为 $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, 则将 $x_{(i)}$ 称为该样本的第 i 个次序统计量. 其中

- 最小次序统计量定义为: $X_{(1)} = \min\{x_1, \dots, x_n\}$
- 最大次序统计量定义为: $X_{(n)} = \max\{x_1, \dots, x_n\}$
- 极差定义为: $R = X_{(n)} - X_{(1)}$
- 第 i 个间差定义为: $S_i = X_{(i)} - X_{(i-1)}$



注 虽然 X_1, X_2, \dots, X_n 独立, 但其次序统计量一般不独立

例题 2.6 若 X_1, X_2, \dots, X_n 独立同分布, 求 $X_{(1)}$ 与 $X_{(n)}$ 的分布

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leq x) = P(X_1 \leq x, \dots, X_n \leq x) \\ &= P(X_1 \leq x) \cdots P(X_n \leq x) \\ &= [F(x)]^n \\ f_{X_{(n)}}(x) &= \frac{d}{dx} F_{X_{(n)}}(x) \\ &= n f(x) [F(x)]^{n-1} \end{aligned}$$

$$\begin{aligned} 1 - F_{X_{(1)}}(x) &= P(X_{(1)} > x) = P(X_1 > x, \dots, X_n > x) \\ &= P(X_1 > x) \cdots P(X_n > x) \\ &= [1 - F(x)]^n \\ f_{X_{(1)}}(x) &= \frac{d}{dx} F_{X_{(1)}}(x) \\ &= n f(x) [1 - F(x)]^{n-1} \end{aligned}$$

定理 2.10

第 i 个次序统计量的概率密度函数为:

$$f_{X_{(k)}} = C(n; 1, k-1, n-k) f(x) [F(x)]^{k-1} [1 - F(x)]^{n-k}$$

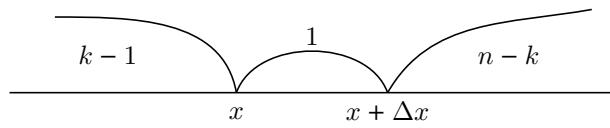


图 2.1: $X_{(k)}$ 取值的示意图

定理 2.11

次序统计量 $(x_{(i)}, x_{(j)})$ ($i < j$) 的联合分布密度函数为

$$\begin{aligned} p_{ij}(y, z) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} [F(y)]^{i-1} [F(z) - F(y)]^{j-i-1} \\ &\quad \cdot [1 - F(z)]^{n-j} p(y)p(z), \quad y \leq z, \end{aligned}$$



证明 对正 $\Delta y, \Delta z$ 以及 $y < z$, 事件 “ $x_{(i)} \in (y, y + \Delta y], x_{(j)} \in (z, z + \Delta z]$ ” 可以表示为“容量为 n 的样本 x_1, \dots, x_n 中有 $i-1$ 个观测值小于等于 y , 一个落入区间 $(y, y + \Delta y]$, $j-i-1$ 个落入区间 $(y + \Delta y, z]$, 一个落入区间 $(z, z + \Delta z]$, 而余下 $n-j$ 个大于 $z + \Delta z$ ”(见图 2.2).

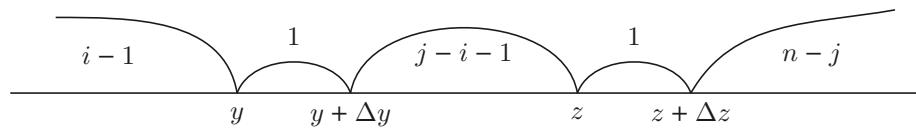


图 2.2: $x_{(i)}$ 与 $x_{(j)}$ 取值的示意图

例题 2.7 若 X_1, X_2, \dots, X_n 独立同分布, 求 $R = X_{(n)} - X_{(1)}$ 的分布

$$f_{X_{(1)}X_{(n)}}(s, t) = n(n-1)f(s)f(t)[F(t) - F(s)]^{n-2}\mathbb{I}(s \leq t)$$

$$f_R(r) = \mathbb{I}(r > 0) \int_{-\infty}^{\infty} f_{X_{(1)} X_{(n)}}(s, s+r) ds$$

第3章 随机变量的数值特征

3.1 期望值

定义 3.1

对于实值随机变量 $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ 和 (可测) 函数 $g : \mathbb{R} \rightarrow \mathbb{R}$, 称

$$\mathbb{E}[g(X)] = \int_{\Omega} g(X(\omega)) d\mathbb{P}(\omega) = \int_{\mathbb{R}} g(x) dF_X(x)$$

为 $g(X)$ 的 **均值** (mean) 或 **期望** (expectation). 期望算子 \mathbb{E} 是一个线性泛函, 仅适用于可积的随机变量.



一个重要结果是, 若 $g(X) \geq 0$, 则 $\mathbb{E}[g(X)] = 0 \implies g(X) \stackrel{\text{a.s.}}{=} 0$, 即 $\mathbb{P}\{g(X) = 0\} = 1$. 其证明可通过 **Markov 不等式**

$$\mathbb{P}\{g(X) \geq \varepsilon\} \leq \mathbb{E}[g(X)]/\varepsilon, \quad \forall \varepsilon > 0$$

完成, 其中需要用到概率的连续性, 即 $\lim_{n \rightarrow \infty} A_n = A \implies \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

刻画 X 的变动程度的量是其**方差** (variance)

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

考虑**均方误差** (mean squared error)

$$\text{MSE}(X; \theta) = \mathbb{E}[|X - \theta|^2], \quad \theta \in \mathbb{R},$$

通过**方差偏差分解** (variance-bias decomposition)

$$\text{MSE}(X; \theta) = \text{Var}(X) + |\mathbb{E}X - \theta|^2$$

可以说明 $\theta \mapsto \text{MSE}(X; \theta)$ 在 $\mathbb{E}X$ 处取到最小值 $\text{Var}(X)$. 投影 (projection) 和正交分解的思想在各种内积空间中应用广泛, 这里是 $\mathbb{E} = \text{proj}_{\mathbb{R}}$, 概率论中关于子事件域 \mathcal{G} (随机元 X , resp.) 的条件期望几何直观是 $\text{proj}_{\mathcal{G}}(\text{proj}_{\sigma(X)}, \text{resp.})$, 线性模型 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 中拟合值为 $\hat{\mathbf{y}} = \text{proj}_{\text{Col}(\mathbf{X})}\mathbf{y}$.

预处理随机变量有两个常用变换:

- **中心化** (centralization) $X \mapsto X - \mathbb{E}X$;
- **标准化** (standardization) $X \mapsto \frac{X - \mathbb{E}X}{\sqrt{\text{Var}(X)}}$.

3.2 矩母函数与特征函数

3.3 熵与信息

第4章 常见分布

4.1 离散分布

4.2 连续分布

4.3 正态分布及其导出分布

第 5 章 概率极限

5.1 收敛

5.2 大数定理

5.3 中央极限定理

附录 A 基本数学工具

A.1 排列与组合

全部组合分析公式的推导基于下列两条原理：

乘法原理 若进行 A_1 过程有 n_1 种方法，进行 A_2 过程有 n_2 种方法，则进行 A_1 过程后再接着进行 A_2 程共有 $n_1 \cdot n_2$ 种方法

加法原理 若进行 A_1 过程有 n_1 种方法，进行 A_2 过程有 n_2 种方法，假定 A_1 过程与 A_2 过程是并行的，则进行过程 A_1 或过程 A_2 的方法共有 $n_1 + n_2$ 种

排列与组合的定义及其计算公式如下。

1. **排列:** 从 n 个不同元素中任取 $r(r \leq n)$ 个元素排成一列 (考虑元素先后出现次序), 称此为一个排列, 此种排列的总数记为 P_n^r , 按乘法原理, 取出的第一个元素有 n 种取法, 取出的第二个元素有 $n - 1$ 种取法...取出的第 r 个元素有 $n - r + 1$ 种取法, 所以有

$$P_n^r = n \times (n - 1) \times \cdots \times (n - r + 1) = \frac{n!}{(n - r)!}. \quad (\text{A.1})$$

若 $r = n$, 则称为全排列, 记为 $n!$. 显然, 全排列 $P_n = n!$.

2. **重复排列:** 从 n 个不同元素中每次取出一个, 放回后再取下一个, 如此连续取 r 次所得的排列称为重复排列, 此种重复排列数共有 n^r 个. 注意这里的 r 允许大于 n .
3. **组合:** 从 n 个不同元素中任取 $r(r \leq n)$ 个元素并成一组 (不考虑元素间的先后次序), 称此为一个组合, 此种组合的总数记为 $\binom{n}{r}$ 或 C_n^r . 按乘法原理此种组合的总数为

$$\binom{n}{r} = \frac{P_n^r}{r!} = \frac{n(n - 1)\cdots(n - r + 1)}{r!} = \frac{n!}{r!(n - r)!}. \quad (\text{A.2})$$

在此规定 $0! = 1$ 与 $\binom{n}{0} = 1$.

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

上述四种排列组合及其总数计算公式, 在确定概率的古典方法中经常使用, 但在使用中要注意识别有序与无序、重复与不重复。