随机问题

Enlightments

作者：syqwq
时间：2025.11.18
邮箱：3422403944@qq.com

Contents

1. 测度论与概率论 3
1.1. 三大重要收敛定理 3
1.2. 为什么要有概率密度函数? 7
1.3. 2025-10-11 概率论作业 11
1.4. 条件期望 13

Chapter 1
测度论与概率论

1.1 三大重要收敛定理

定理 1.1.1 单调收敛定理(Monotone Convergence Theorem - MCT)

若 $(\Omega, \mathcal{F}, \mu)$ 是一个测度空间，$\{f_i\}_{i=1}^{\infty}$ 是一族 $\Omega \to \mathbb{R}$ 的可测函数，满足 $f_i \uparrow f$ a.e. 且 $\int f_1 \,\mathrm{d}\mu > -\infty$. 则积分和极限可以交换，且积分也单调收敛，即

$$\int f_i \,\mathrm{d}\mu \uparrow \int f \,\mathrm{d}\mu$$

注解 1.1.1 热气球

- 想象 f_i 是一个热气球在时刻 i 的体积，$\int f_i \,\mathrm{d}\mu$ 是它的总浮力.
- 条件 $f_i \uparrow f$ 意味着这个热气球只增不减，不断地膨胀，最终趋近于一个最终形态 f.
- MCT 定理说：既然热气球一直在稳定地、单调地变大，那么它浮力的极限，理所当然就是它最终形态的浮力. 这个过程中没有“泄气”或者其他诡异的事情发生.

Proof. 我们将证明分为两个主要部分: 首先处理更简单也更基础的非负函数情况, 然后利用它来证明一般情况.

先假设 $f_1 \geq 0$ 相当于证明等式:

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu$$

考虑一边，即证明

$$\lim \int f_n \,\mathrm{d}\mu \leq \int f \,\mathrm{d}\mu$$

这是显然的. 因为 $f_n \uparrow f$ 意味着 $\forall n, f_n(\omega) \leq f(\omega)$，从而 $\int f_n \,\mathrm{d}\mu \leq \int f \,\mathrm{d}\mu$. 因为 $\int f_n \,\mathrm{d}\mu$ 单调有上界，因此 $\lim \int f_n \,\mathrm{d}\mu$ 存在，且 $\lim \int f_n \,\mathrm{d}\mu \leq \int f \,\mathrm{d}\mu$.

接下来考虑另一侧

$$\lim \int f_n \,\mathrm{d}\mu \geq \int f \,\mathrm{d}\mu$$

回顾 $\int f \,\mathrm{d}\mu$ 的定义是

$$\int f\,\mathrm{d}\mu := \sup\left\{\int g\,\mathrm{d}\mu \;\middle|\; 0 \le g \le f, g\text{为简单函数}\right\}$$

如果我们可以证明对于任意的简单函数 g，都有 $c := \lim \int f_n\,\mathrm{d}\mu \ge \int g\,\mathrm{d}\mu$，那么一定可以得到 c 比他们的上确界大，从而得证.

为此，任取简单函数 g s.t. $0 \le g \le f$，再取一个略小于 1 的数 $a := 1 - \varepsilon$. 构造集合

$$E_n := \{\omega \mid f_n(\omega) \ge a g(\omega)\}$$

这个集合 E_n 是指那些 f_n 的值已经“长得足够高”，至少达到了 g 的 a 倍的点 ω 的集合. 由于 f_n 单调递增，因此这个集合也单调，即 $E_1 \subseteq E_2 \subseteq \cdots$. 因为 $f_n(\omega) \uparrow f(\omega)$ 且 $f(\omega) \ge g(\omega)$，这个集合会不断扩大，直至整个 g 的支撑集，也就是

$$\bigcup_{n=1}^{\infty} E_n = \{\omega \mid g(\omega) > 0\}$$

从而，得到

$$\int f_n\,\mathrm{d}\mu \ge \int_{E_n} f_n\,\mathrm{d}\mu \ge a\int_{E_n} g\,\mathrm{d}\mu$$

不等式两侧取极限得到

$$c \ge a \cdot \lim_{n\to\infty}\int_{E_n} g\,\mathrm{d}\mu = a\cdot\int g\,\mathrm{d}\mu$$

因为 g 是任取的，且 a 可以任意接近 1，因此就有

$$c \ge \sup\int g\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$$

从而 f_1 非负部分证完了.

接下来，证明另一部分.

我们构造函数列 $f_n{}' = f_n - f_1$，则 $f_n{}' \ge 0$ 且 $f_n{}' \uparrow f' := f - f_1$，因此现在的 $f_n{}'$ 满足上面情况的所有条件，从而我们得到

$$\int f_n{}'\,\mathrm{d}\mu \uparrow \int f'\,\mathrm{d}\mu \Leftrightarrow \int (f_n - f_1)\,\mathrm{d}\mu \uparrow \int (f - f_1)\,\mathrm{d}\mu$$

由于 $\int f_1\,\mathrm{d}\mu > -\infty$ 利用积分的线性性，得到

$$\int f_n\,\mathrm{d}\mu \uparrow \int f\,\mathrm{d}\mu$$

■

定理 1.1.2 Fatou 引理(Fatou's Lemma)

若 $(\Omega, \mathcal{F}, \mu)$ 是一个测度空间，$\{f_i\}_{i=1}^{\infty}$ 是一族 $\Omega \to \mathbb{R}$ 的**非负**可测函数. 则

$$\int \liminf_{i\to\infty} f_i\,\mathrm{d}\mu \le \liminf_{i\to\infty}\int f_i\,\mathrm{d}\mu$$

注解 1.1.2 一个可能漏水的水桶

- 想象 $\int f_i \, \mathrm{d}\mu$ 是一个桶在时刻 i 的总水量. f_i 是水的分布.
- f_i 不要求单调，你可以把水在桶里晃来晃去，甚至有些水会溅出来.
- $\liminf f_i$ 是在晃动过程中，每个位置“最终”稳定留下的水位.
- **Fatou 引理**说: 最终桶里剩下的水量 (左边), 小于等于你每次测量水量的极限 (右边) .
- 为什么是不等式? 因为在晃动的过程中 (f_i 不单调), 可能有一部分“质量” (积分值) 泄露掉了或“蒸发”了 (数学上叫“跑到无穷远处”) . 所以你最终看到的实体 ($\liminf f_i$) 的积分，可能会比积分的极限要小.

Proof. 我们手里的工具是强大的**单调收敛定理(MCT)**，但它的使用条件很苛刻，要求函数序列是单调递增的. 而 Fatou 引理的 $\{f_i\}$ 序列不保证单调性. 所以，证明的核心思想是：从不单调的 $\{f_i\}$ 出发，构造出一个相关的、新的、单调递增的序列 $\{g_j\}$ ，然后对 $\{g_j\}$ 应用 MCT，最后再把结果和 $\{f_i\}$ 联系起来.

我们先回顾 $\liminf$ 的定义:

$$\liminf_i f_i := \sup_j \inf_{i \geq j} f_i$$

受到启发，我们定义新的函数列 g_j

$$g_j := \inf\{f_i \mid i \geq j\}$$

也就是 g_j 是原序列 f_i 从第 j 项开始的 “尾巴” 的下确界. 由于所选取的集合越来越小，因此下确界是递增的，从而有 $g_1 \leq g_2 \leq \cdots$，这就意味着我们构造出了一个单调递增的序列.

而根据 $\liminf$ 的定义，这个单调递增序列 $\{g_j\}$ 的极限（也就是它的上确界 $\sup g_j$ ）正好就是 $\liminf f_i$. 所以我们有: $g_j \uparrow \liminf f_i$. 因为 f_i 都非负，所以 g_j 也非负. 应用 MCT，我们得到

$$\lim_{j \to \infty} \int g_j \, \mathrm{d}\mu = \int \liminf_{i \to \infty} f_i \, \mathrm{d}\mu$$

而这就是我们要证明的不等式的左边，接下来我们考虑右边.

由于 g_j 的定义 $g_j = \inf\{f_i \mid i \geq j\}$，于是 $\forall i \geq j, f_i \geq g_j$，利用积分的单调性有

$$\forall i \geq j, \int f_i \, \mathrm{d}\mu \geq \int g_j \, \mathrm{d}\mu$$

既然 $\int g_j \, \mathrm{d}\mu$ 比这里的每一项都要小，那么一定小于等于他们的下确界，从而有关键不等式

$$\int g_j \, \mathrm{d}\mu \leq \inf_{i \geq j} \int f_i \, \mathrm{d}\mu$$

我们对上面这个不等式两边同时取 $j \to \infty$ 的下极限 $\liminf$

$$\lim_{j \to \infty} \int g_j \, \mathrm{d}\mu = \liminf_{j \to \infty} \int g_j \, \mathrm{d}\mu \leq \liminf_{j \to \infty} \left(\inf_{i \geq j} \int f_i \, \mathrm{d}\mu \right)$$

由于 g_j 是单调收敛的序列，因此下极限等于极限等于 $\int \liminf_{i\to\infty} f_i \,\mathrm{d}\mu$. 在这个不等式的右边，被取极限的部分正好符合 $\liminf$ 的定义，因此就等于 $\liminf \int f_j \,\mathrm{d}\mu$，把两侧的分析结果带入，就得到了我们要证明的 Fatou 引理

$$\int \liminf_{i\to\infty} f_i \,\mathrm{d}\mu \le \liminf_{i\to\infty} \int f_i \,\mathrm{d}\mu$$

■

定理 1.1.3 控制收敛定理 (Dominated Convergence Theorem - DCT)

若 $(\Omega, \mathcal{F}, \mu)$ 是一个测度空间，$\{f_i\}_{i=1}^{\infty}$ 是一族 $\Omega \to \mathbb{R}$ 的可测函数，g 是一个可积函数. 如果 $|f_i| \le g$ a.e. 且 $\forall \omega \in \Omega, f_i(\omega) \to f(\omega)$ (逐点收敛)，则 f 也是可积的，且积分和极限可交换，即

$$\int f_i \,\mathrm{d}\mu \to \int f \,\mathrm{d}\mu \Leftrightarrow \int f \,\mathrm{d}\mu = \int \lim_{n\to\infty} f_n \,\mathrm{d}\mu = \lim_{n\to\infty} \int f_n \,\mathrm{d}\mu$$

注解 1.1.3 有盖子的桶

- 这就像 **Fatou 引理**里那个可能漏水的桶，但现在我们给它加了一个坚固的盖子 (g).
- $f_i \to f$ 意味着桶里的水最终会形成一个稳定的形态 f .
- $|f_i| \le g$ 且 $\int g \,\mathrm{d}\mu < \infty$ 这个“盖子”的作用是阻止任何水的泄露和蒸发. 因为所有的水都被限制在一个总容量有限的空间里.
- **DCT 定理**说：既然一滴水都没有损失，那么水量的极限，自然就等于极限形态下水的总量.

Proof. 证明的核心思想是构造两条“单调的夹板”来挤压 原始的序列 $\{f_i\}$ 本身不一定是单调的，所以我们不能直接用 MCT. 因此，我们从 $\{f_i\}$ 出发，构造出两个新的、单调的函数序列，一个从下方逼近 f ，一个从上方逼近 f ，像两块夹板一样把 f_i 夹在中间. 然后我们分别对这两条“单调的夹板”应用 MCT.

我们定义两个辅助序列:

(1) 地板序列 $f_i^{\wedge}$

$$f_i^{\wedge} := \inf\{f_j \mid j \ge i\}$$

这个序列是单调递增的，并且收敛到 $\liminf f_i$. 因为 $f_i \to f$，所以 $f_i^{\wedge} \uparrow f$.

(2) 天花板序列 $f_i^{\vee}$

$$f_i^{\vee} := \sup\{f_j \mid j \ge i\}$$

这个序列是单调递减的，并且收敛到 $\limsup f_i$. 因为 $f_i \to f$，所以 $f_i^{\vee} \downarrow f$.

通过定义，我们得到

$$f_i^{\wedge} \le f_i \le f_i^{\vee} \tag{1.1.1}$$

由于 g 的存在保证了有限且可积，因此对于地板序列应用 MCT，我们得到

$$\lim_{i\to\infty}\int f_i^{\wedge}\,\mathrm{d}\mu=\int f\,\mathrm{d}\mu$$

类似的，应用 MCT（递减版），我们得到

$$\lim_{i\to\infty}\int f_i^{\vee}\,\mathrm{d}\mu=\int f\,\mathrm{d}\mu$$

对 式 (1.1.1) 取积分，得到

$$\int f_i^{\wedge}\,\mathrm{d}\mu\le\int f_i\,\mathrm{d}\mu\le\int f_i^{\vee}\,\mathrm{d}\mu$$

再通过取极限，应用夹逼定理得到

$$\lim_{i\to\infty}\int f_i\,\mathrm{d}\mu=\int f\,\mathrm{d}\mu$$

■

1.2 为什么要有概率密度函数?

为了接下来的讨论，我们首先要在一些概念的定义上达成共识.

定义 1.2.1 可测函数

对于两个可测空间 $(\Omega,\mathcal{F}),(S,\mathcal{S})$，如果存在一个函数 $X:\Omega\to S$ 满足

$$X^{-1}(B):=\{\omega\mid X(\omega)\in B\}\in\mathcal{F},\forall B\in\mathcal{S}$$

那么，称 X 是一个从 $(\Omega,\mathcal{F})$ 到 $(S,\mathcal{S})$ 的**可测函数**.

注解 1.2.1

这说的就是对于任何一个事件 $\omega\in\mathcal{F}$，我们固然可以得知他的结果 $X(\omega)$，而可测函数的定义是在说，我任意给定一个 $B\in\mathcal{S}$，他一定有 $\mathcal{F}$ 中的可测事件与他对应.

定义 1.2.2 随机向量、随机变量

在 定义 1.2.1 中，如果 $(S,\mathcal{S})=\left(\mathbb{R}^d,\mathcal{R}^d\right)$，当 $d>1$ 时，我们称 X 为**随机向量**；当 $d=1$ 时，我们称 X 为**随机变量**.

我们为了研究一个事件发生的概率，例如抛出一枚硬币是否是正面朝上，一个很直观的想法就是把所有可能的情况列出来，而随机变量的作用就是给某些情况的集合一个“标签”，就如同拎葡萄，抓着柄就抓了一串，并且随机变量的定义也保证了这样拎出来的集合一定也是可测的. 这样就可以让我们在讨论某些具有共性的集合的时候更加得心应手.

在抛硬币中，我们假设 X 是一个随机变量，它的取值为 0（反面）或 1（正面）. 那么，我们可以很清晰地通过**分布列（PMF）**来研究整个样本空间中每个事件发生的概率：

	$X=0$	$X=1$

$$P \Big| \quad 1/2 \qquad 1/2$$

但是，当问题来到连续型随机变量时，这就并不奏效了.

考虑这样一个场景，我们有一个理想的随机数生成器，会随机地、均匀地从 $[0,1]$ 中生成一个数字 X. 现在，让我们尝试用分布列的逻辑来描述它. 这意味着我们要给 $[0,1]$ 区间内的每一个点都赋予一个概率值 $P(X=x)$. 那么此时，灾难性的问题出现了:

- 区间内有多少点? 无穷个.
- 我们应该给每个 $P(X=x)$ 赋予多大的值?
  - 假设我们给每个点的概率是一个很小的 $\varepsilon > 0$，那么总概率

$$\sum_{x\in[0,1]} P(X=x) = \sum_{x\in[0,1]} \varepsilon = \infty \times \varepsilon = \infty$$

    一发不可收拾了 😭
  - 既然不能是正数，那么只能是 0 了. 但是

$$\sum_{x\in[0,1]} P(X=x) = \sum_{x\in[0,1]} 0 = 0$$

    这与 $P(\Omega)=1$ 的要求不符，也不对啊 😭

分布列这个工具，它的“给每个点分配一块概率，然后加总”的核心思想，在面对连续变量的“不可数无穷”特性时，彻底失效了. 无论怎么给单点分配概率，总概率要么是无穷，要么是 0，永远凑不成那个必需的“1”.

既然纠结于单点的概率是死路一条，于是我们换一个角度: 能不能不讨论单点的概率，而是讨论一个区间的概率? 这个思想的转变，就是引入 PDF 的直觉来源.

(1) **放弃“点的质量”，拥抱“点的密度”**

想象一根一米长的金属棒，总质量是 1 千克. 你问: “在 0.5 米那个点上，质量是多少? ” 答案是 0. 一个没有长度的点，自然没有质量. 但是，你可以问: “在 0.5 米那个点附近，质量的密度是多少? ” 这就有意义了，比如是 1.2 公斤/米.

密度这个概念，描述的不是一个点的质量，而是这个点周围质量的集中程度.

(2) **PDF 就是概率上的密度**

概率密度函数 (PDF) $f(x)$ 就扮演了这个“密度”的角色. $f(x)$ 本身的值不是概率，就像 “kg/m^3” 不是 “kg” 一样. **$f(x)$ 的值越大，表示随机变量落在 x 附近的可能性就越“密集”或越“集中”.**

(3) **如何从“密度”得到“概率”?**

对于那根金属棒，如何从密度得到一段区间的质量? 答案是积分! 把从 0.4 米到 0.6 米的密度函数积分，就得到了这段的质量.

$$m = \int_L \lambda \, \mathrm{d}\ell$$

同理，对于连续型随机变量，要计算它落在区间 $[a,b]$ 内的概率，我们就需要对概率密度函数在这个区间上进行积分:

$$P(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x$$

这个模型完美解决了以前的所有问题: 总概率 $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$, 而单点的概率 $P(X = a) = \int_a^a f(x)\,\mathrm{d}x = 0$. 所有矛盾都消失了.

因此, 我们就引出了如下的定义.

定义 1.2.3 连续型随机变量

在概率空间 $(\Omega, \mathcal{F}, P)$ 上, 如果随机变量 $X : \Omega \to \mathbb{R}$ 满足:

存在一个可测函数 $f : \mathbb{R} \to [0, \infty)$ 使得

$$P(X \le x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, -\infty < x < \infty$$

那么称 X 为**连续型随机变量**, f 为其**概率密度函数（PDF）**.

但是, 这里有一个小问题: 我们之前的讨论都是对于区间 $[a, b]$ 的概率, 而定义中却是 $(-\infty, x]$ 的概率, 这是否合理呢? 答案是肯定的. 事实上, 由所有形如 $(-\infty, x]$ 的半无限区间构成的集合, 是整个实数轴上 Borel 代数 $\mathcal{B}(\mathbb{R})$ 的一个**生成元**, 也就是通过可数次交并补操作得到整个 $\mathcal{B}(\mathbb{R})$. 而对于具有这样性质的集合, 我们有一个很神奇的定理

定理 1.2.1 Carathéodory 扩张定理

设 $\mathcal{A}$ 是集合 X 上的代数, μ_0 是集合 $\mathcal{A}$ 上的一个预测度. 令 μ^* 是由 μ_0 导出的外测度. 那么所有 μ^*-可测集构成的集合 $\mathcal{M}$ 满足:

(1) $\mathcal{M}$ 是一个 σ-代数

(2) $\mathcal{A} \subseteq \mathcal{M}$

(3) 这个测度 μ 是预测度 μ_0 的一个扩张, 即对于任何在原始代数中的集合 $A \in \mathcal{A}$, 都有 $\mu(A) = \mu_0(A)$.

如果在上述条件下, 预测度 μ_0 是 σ-有限的（即存在一系列 $A_i \in \mathcal{A}$ s.t. $\bigcup_{i=1}^{\infty} A_i = X$ 且 $\mu_0(A_i) < \infty$）, 那么这个扩张是唯一的.

既然所有我们关心的区间（以及更复杂的集合）都可以由 $(-\infty, x]$ 这种“基本砖块”通过集合运算搭建而成, 那么根据 定理 1.2.1 中测度的唯一性, 我们可以得出一个结论:

只要我们在“生成元”这个集合上定义好了概率, 那么所有由它生成的更复杂的集合（所有 Borel 集）上的概率也就被唯一地确定了.

感谢大家的补充解释, 以下是我基于个人理解的一些整理. 👀 💦 💦 💦

同样从金属棒入手, 想象一根一米长的密度不均匀的金属棒. 我们现在有两种测度:

(1) 质量测度 μ: 这是一个函数, 你给它棒上的任意一段区间 A , 它会返回这段区间的质量 $\mu(A)$

(2) 长度测度 λ: 这是一个函数, 你给它任意一段区间 A , 它会返回这段区间的长度 $\lambda(A)$. 这本质上就是勒贝格测度.

这根棒的总质量 $\mu([0,1])$ 是 1 千克. 但由于密度不均, 我们不能简单地说“质量等于长度”. 那么, 我们如何描述在**每一个点 x 上, 质量相对于长度的“集中程度”**呢?

对于这种问题, 我们考虑“求导”的思想. 在微积分中, $\mathrm{d}y/\mathrm{d}x$ 描述了 y 相对于 x 的变化率. 类似的, 我们想定义一个 “测度对于测度的导数” $\mathrm{d}\mu/\mathrm{d}\lambda$. 这个导数是一个函数, 我们称为密度函数 $\rho(x)$, 他在某个点 x 的值, 代表了该点单位长度上所拥有的质量.

那么怎么求这个值? 我们可以想象在点 x 附近取一个极小的区间 ΔA , 然后计算这个区间的质量与长度之比: $\mu(\Delta A)/\lambda(\Delta A)$. 当这个小区间缩向点 x 时, 这个比值的极限就是点 x 的密度 $\rho(x)$. 一旦我们求出了这个密度函数 $\rho(x)$ (也就是那个“导数”), 我们就可以通过积分来反求任何一段区间的总质量:

$$\mu(A) = \int_A \rho(x)\,\mathrm{d}\lambda(x)$$

这个过程, 就是从导数 (密度) 通过积分还原原函数 (质量) 的过程.

那么, 接下来问题回到概率论中. 现在我们也有两种测度:
(1) 概率测度 P_X: 这是由随机变量 X 导出的测度. 你给它实数轴上任意一个事件 (Borel 集) A , 它返回 X 落入 A 的概率 $P_{X(A)}$. 这对应了上面例子中的“质量测度 μ”.
(2) 勒贝格测度 λ: 这是实数轴上标准的“长度”测度.

在进行求导之前, 必须满足一个关键条件, 就是测度的绝对连续性 (Absolute Continuity).

定义 1.2.4 测度的绝对连续性

如果 μ,ν 是同一个测度空间上的两个测度, 那么我们称 μ 相对于 ν 是**绝对连续的**, 如果对于任何集合 A , 只要 $\nu(A)=0$, 那么 $\mu(A)=0$, 记为 $\mu \ll \nu$. 形式化地说就是

$$\mu \ll \nu \Leftrightarrow (\nu(A)=0 \Rightarrow \mu(A)=0)$$

在这里, 我们说概率测度 P_X 相对于勒贝格测度 λ 是绝对连续的, 如果对于任何集合 A , 只要它的“长度”为零 $\lambda(A)=0$, 那么它的“概率”也必须为零 $P_X(A)=0$, 显然单点集就是这样的零测集.

注解 1.2.2 绝对连续性

直观含义就是“没有长度的地方, 就没有概率”. 这意味着概率不会“凭空”集中在一些没有长度的“点”或“尘埃”上. 这完美地排除了离散概率 (集中在单点上), 并从根本上解释了为什么连续随机变量在单点的概率为 0.

定理 1.2.2 拉东-尼科迪姆定理 (The Radon-Nikodym Theorem)

如果测度空间 (X,Σ) 上的一个 σ-有限测度 ν 关于另一个 σ-有限测度 μ 绝对连续, 那么存在一个在 X 上的非负可测函数 $f: X \to [0,\infty)$, 也就是 μ 关于 ν 的密度, 或者 Radon-Nikodym 导数 $f = \dfrac{\mathrm{d}\mu}{\mathrm{d}\nu}$. 对所有的可测集合 A , 都有:

$$\mu(A) = \int_A f\,\mathrm{d}\nu$$

那么在概率论的语境下，如果概率测度 P_X 相对于勒贝格测度 λ 是绝对连续的，那么根据拉东-尼科迪姆定理，必然存在一个唯一的（几乎处处唯一）非负可积函数 f_X，使得对于任何事件 A，都有：

$$P_X(A) = \int_A f_X(x)\,\mathrm{d}\lambda(x)$$

这个必然存在的函数 f_X 就是 P_X 相对于 λ 的 Radon-Nikodym 导数，我们给这个函数起了一个专门的名字：**概率密度函数 (PDF)**.

1.3 2025-10-11 概率论作业

问题 1.3.1

设 X 为仅取非负整数的离散随机变量，若其数学期望存在，证明：

(1) $E(X) = \sum_{k=1}^{\infty} P(X \geq k)$

(2) $\sum_{k=0}^{\infty} kP(X > k) = \frac{1}{2}(E(X^2) - E(X))$

Proof.

(1) 不难发现，这是由于无穷级数收敛，从而求和可以交换次序.

(2) 同理，由于无穷级数求和收敛，根据等差数列求和公式易证.

■

问题 1.3.2

甲、乙、丙三人进行比赛，规定每局两个人比赛，胜者与第三人比赛，依次循环，直至有一人连胜两次为止，此人即为冠军.而每次比赛双方取胜的概率都是 1/2，现假定甲、乙两人先比，试求各人得冠军的概率.

Solution. 假设 (X, Y) 表示在上一场比赛中，X 战胜了 Y. 记甲,乙,丙分别为 A, B, C. 那么比赛的可能状态有 $(A,B),(B,A),(A,C),(C,A),(B,C),(C,B)$. 记 $P_A(X,Y)$ 为在状态 (X,Y) 下，A 最终获得冠军的概率. 那么，由于初始条件，我们得到

$$P(A) = \frac{1}{2}(P_A(A,B) + P_A(B,A))$$

接下来，考虑各个状态下的转移概率：

(1) (A,B) 下，如果 A 获胜了，那么直接赢得比赛；如果 A 战败了，那么进入 (C,A) 状态. 因此

$$P_A(A,B) = \frac{1}{2}(1 + P_A(C,A))$$

(2) (B,A) 下，如果 B 获胜了，B 直接赢了，此时 A 获胜的概率为 0；如果 B 战败了，那么进入 (C,B) 状态. 因此

$$P_A(B,A) = \frac{1}{2}(0 + P_A(C,B))$$

(3) (A,C) 下，如果 A 赢了，那么 A 直接获胜；如果 A 输了，那么进入 (B,A) 状态. 因此

$$P_A(A,C)=\frac{1}{2}(1+P_A(B,A))$$

(4) (C,A) 下，如果 C 赢了，那么 C 直接获胜，此时 A 获胜的概率为 0；如果 C 输了，那么进入 (B,C) 状态. 因此

$$P_A(C,A)=\frac{1}{2}(0+P_A(B,C))$$

(5) (B,C) 下，如果 B 赢了，那么 B 直接获胜了，此时 A 获胜的概率为 0；如果 B 输了，那么进入 (A,B) 状态. 因此

$$P_A(B,C)=\frac{1}{2}(0+P_A(A,B))$$

(6) (C,B) 下，如果 C 赢了，那么 C 直接获胜了，此时 A 获胜的概率为 0；如果 C 输了，那么进入 (A,C) 状态. 因此

$$P_A(C,B)=\frac{1}{2}(0+P_A(A,C))$$

联立以上的 6 个方程，得到

$$P_A(A,B)=\frac{4}{7}$$

$$P_A(B,A)=\frac{1}{7}$$

因此得到

$$P(A)=\frac{1}{2}(P_A(A,B)+P_A(B,A))=\frac{5}{14}$$

⬢

1.4 条件期望

今天上课的时候，概率论老师说了这样一句话：

条件分布，条件期望很深刻，因为条件是人为创造的，就和集合中的辅助线一样.

这不禁引发我的思考……

1.4.1 为什么说它是“辅助线”?

在平面几何中, 你面对一个复杂的图形, 直接求证很难. 但是, 如果你人为地画一条线 (辅助线), 图形突然就被分割成了两个大家熟悉的三角形，原本复杂的关系瞬间变得清晰明了.

在概率论中，全期望公式 (Law of Total Expectation) 就是这条辅助线

$$E[X] = E[E[X|Y]]$$

- **原本的问题**: 直接求 X 的期望 $E[X]$ 可能非常困难, 因为 X 的分布太复杂, 混杂了各种因素.
- **引入辅助线 (Y)**：我们人为引入一个随机变量 Y. 这个 Y 在原题中可能根本没出现，是我们为了解题“创造”出来的.
- **化整为零**：一旦我们固定了 Y 的值（比如 $Y = y$），在这个条件下，X 的行为可能变得异常简单（变成了简单的三角形）.
- **最后整合**：算出简单的 $E[X|Y]$ 后，再对 Y 求期望，就把问题解决了.

1.4.2 例子们

问题 1.4.1

假设你开了一家商店. 每天进店的顾客人数 N 是随机的，均值为 $E[N] = 100$ 人. 每位顾客消费的金额 X_i 也是随机的，均值为 $E[X] = 50$ 元. 顾客人数和每人的消费金额是独立的. 问：你这一天总营业额 S 的期望是多少?

Solution. 直接做（很难）：总金额 $S = X_1 + X_2 + \cdots + X_N$. 请注意，这里的项数 N 本身就是随机的！你不能直接用线性性质拆开，因为你不知道有多少项. 如果 N 和 X 的分布很复杂，求 S 的分布简直是噩梦.

加“辅助线” (Conditioning)：我们引入条件：**假设我们知道今天来了 n 个人（即 $N = n$）**.

(1) **在条件 $N = n$ 下**：总金额 S 就变成了固定的 n 个随机变量之和：$S = X_1 + \cdots + X_n$. 这时候求期望就太简单了：

$$E[S|N = n] = E[X_1 + \cdots + X_n] = n \cdot E[X]$$

这意味着，随机变量 $E[S|N] = N \cdot E[X]$.

(2) **使用全期望公式**:

$$E[S] = E[E[S|N]] = E[N \cdot E[X]]$$

因为 $E[X]$ 是常数，提出来：

$$E[S] = E[X] \cdot E[N] = 50 \cdot 100 = 5000$$

妙处：我们通过引入条件 N，把一个变动项数的求和问题，转化为了简单的乘法问题. ●

问题 1.4.2

一只老鼠被困在矿井里，面前有三个门：
- 门 1：通向一条隧道，走 2 小时后回到原点（没出去）.
- 门 2：通向一条隧道，走 3 小时后回到原点（没出去）.
- 门 3：通向一条隧道，走 5 小时后直接逃出矿井.

老鼠每次随机选一个门（概率各 1/3），如果回到原点，它会像失忆一样重新随机选. 问：老鼠逃出去的平均时间 $E[T]$ 是多少?

Solution. 过程懒得写了，留作习题（也比较经典哈哈哈）. 通过 Conditioning，我们利用了过程的自相似性 (Self-similarity)，把一个无穷级数求和问题变成了一个代数方程. ●

1.4.3 深刻的理解：条件期望是一个“投影”

如果在更高等的数学（希尔伯特空间理论）中看，条件期望 $E[X|Y]$ 实际上是随机变量 X 在由 Y 生成的信息空间上的**正交投影 (Orthogonal Projection)**.

- X 是一个包含所有细节的、起伏不定的函数.
- $E[X|Y]$ 是一个**平滑**后的版本，它去掉了 X 中那些与 Y 无关的“噪音”，只保留了 X 随 Y 变化而变化的“趋势”.

这就像辅助线：辅助线往往是图形的对称轴或高线，它抓住了图形最本质的骨架. 条件期望也是抓住了随机变量 X 在条件 Y 下的“骨架”.

具体而言，想象一个巨大的无限维空间，里面的每一个点不是坐标，而是一个随机变量 X.

- **向量**：随机变量 X.
- 长度（范数）：$\|X\| = \sqrt{E[X^2]}$.
- 角度（内积）：$\langle Z, X\rangle = E[XZ]$. 如果 $E[XZ] = 0$，我们就说 X 和 Z 是**正交（垂直）**的.
- 距离：$\|X - Z\| = \sqrt{E[(X-Z)^2]}$. 这是“均方误差”.

这是一个 L^2 空间.

现在，假设我们观测到了随机变量 Y.

- Y 带来的信息生成了一个**子空间 (Subspace)**，记为 $\mathcal{M}_Y$.
- 在这个子空间里的所有向量，都是 Y 的函数，即形如 $g(Y)$ 的随机变量.
- **直观理解**：$\mathcal{M}_Y$ 就像是三维空间里的一个“平面”. 我们只能在这个平面上活动，因为我们只知道 Y.

现在，有一个随机变量 X（目标），它漂浮在这个平面 $\mathcal{M}_Y$ 的外面（因为 X 包含了一些 Y 无法解释的随机性）.

任务：我们要在这个平面 $\mathcal{M}_Y$ 上找到一个点 Z，使得 Z 离 X 最近. 也就是要最小化“距离”：

$$\min_{Z\in\mathcal{M}_Y} \|X - Z\|^2 = \min_{g} E[(X - g(Y))^2]$$

几何直觉：在几何学中，从平面外一点 X 到平面 $\mathcal{M}_Y$ 的最短距离，就是从 X 向平面做**垂线**. 垂足就是**正交投影**.

根据希尔伯特空间的投影定理，这个最佳逼近点 Z 必须满足：**误差向量 $(X-Z)$ 必须垂直于平面上的任意向量 W**. （使用变分法，考虑 $J(\varepsilon)=\|(X-Z)-\varepsilon W\|^2$，然后求导可得等价于内积为 0） 用内积公式写出来就是：

$$\forall W\in\mathcal{M}_Y, E[(X-Z)W]=0\Leftrightarrow E[XW]=E[ZW]$$

惊人的事实：数学家发现，满足这个几何性质的 Z，正是我们定义的条件期望 $E[X|Y]$！（证明在后面呜呜呜）

所以：**$E[X|Y]$ 就是 X 在 Y 所生成的子空间上的正交投影**.

- 它去掉了 X 中与 Y “垂直”（无关）的噪音部分.
- 它保留了 X 中“平行”（相关）于 Y 的部分.
- 这就是为什么在最小均方误差（MSE）意义下，$E[X|Y]$ 是 X 的最佳估计.

考虑随机变量独立性的一般定义

定义 1.4.1 独立随机变量

随机变量 X 和 Y 独立，当且仅当

$$P(AB)=P(A)\cdot P(B)$$

这意味着， $P(B)=P(B|A)$，这相当于是 $E[X|Y]=E[X]$. 因为如果说条件期望是“投影”，那么独立性就意味着**“投影之后只剩下一个常数中心”**.

当我们做投影时：

- **一般情况（相关）**：比如 B 和 A 有夹角（相关）. 把 B 投影到 A 的轴上，你会得到一个随 A 变化的影子（长短不一）. 这就是 $P(B|A)$ 随着 A 变化.
- **独立情况**：B 的“变化部分”完全垂直于 A 的轴.
  - 你把 B 强行投影到 A 的空间上.
  - 因为 B 的波动方向和 A 完全无关（垂直），所以 B 的波动在 A 上的投影是 0.
  - 投影剩下的结果是什么？只剩下 B 的**平均值（重心）**.

如果 A,B 独立：

(1) 向量 $\mathbb{1}_B$ 的“变化方向”与向量 $\mathbb{1}_A$ 的空间完全垂直.

(2) 如果你站在 A 的角度去看 B，你完全看不到 B 的任何“起伏”或“趋势”.

(3) 你只能看到 B 的平均大小，也就是 $P(B)$.

1.4.4 一些应用

把条件期望看作正交投影，简直是概率论里的“降维打击”. 一旦你戴上这副几何眼镜，很多原本需要繁琐积分证明的性质，瞬间就变成了直观的几何公理.

这里有三个最“妙”的结论，它们从几何角度看是显然的，但在统计和机器学习中威力无穷.

定理 1.4.1 重期望公式

假设 $\sigma(Z) \subset \sigma(Y)$，则

$$E[E[X|Y]|Z] = E[X|Z]$$

几何视角（秒杀）：想象三个空间：

(1) **全空间**：包含 X 的大宇宙.

(2) **中空间** ($\mathcal{M}_Y$)：由 Y 生成的子空间（比如这是一个平面）.

(3) **小空间** ($\mathcal{M}_Z$)：由 Z 生成的子空间. 因为 Z 的信息比 Y 少，所以 $\mathcal{M}_Z$ 是 $\mathcal{M}_Y$ 里面的一条直线.

公式的含义：

- $E[X|Y]$：先把 X 投影到**平面** Y 上，得到影子 X_Y.
- $E[\cdot\,|Z]$：再把刚才那个影子 X_Y，投影到**直线** Z 上.

(如果 Z 是常数轴 $\mathcal{M}_0$，就得到 $E[X] = E[E[X|Y]]$)

结论：先把一个点垂直投影到平面上，再把平面上的影子垂直投影到平面内的直线上，**这等价于直接把那个点垂直投影到直线上！**

影子的影子，就是原始物体的影子. 这就是为什么内层的 Y 直接“消失”了.

定理 1.4.2 全方差公式 (Law of Total Variance)

$$\mathrm{Var}(X) = E[\mathrm{Var}(X|Y)] + \mathrm{Var}(E[X|Y])$$

总方差 = 组内方差的均值 + 组间方差. 这是统计学中方差分析 (ANOVA) 的基石.

几何视角（勾股定理）：我们回顾投影的几何结构.

- 向量 X 是原始变量.
- 向量 $Z = E[X|Y]$ 是投影（最佳预测）.
- 向量 $\varepsilon = X - E[X|Y]$ 是误差（残差）.

根据正交投影的性质，**预测值 Z 和误差 ε 是垂直（正交）的！**

在欧几里得几何中，直角三角形的斜边平方等于两直角边平方和：

$$\|X\|^2 = \|Z + \varepsilon\|^2 = \|Z\|^2 + \|\varepsilon\|^2$$

(这里我们要考虑中心化后的向量，即减去均值 $E[X]$)

对应到概率论中，长度的平方就是方差：

- 斜边平方 $\to \mathrm{Var}(X)$ (**总波动**)
- 直角边 1 平方 $\to \mathrm{Var}(E[X|Y])$ (**被 Y 解释的波动/信号的能量**)
- 直角边 2 平方 $\to E[\mathrm{Var}(X|Y)]$ (**没被 Y 解释的波动/噪音的能量**)

妙处：你不需要死记硬背复杂的方差公式，只需要画一个直角三角形：**总风险 = 预测模型能解释的风险 + 预测模型剩下的纯噪音**.

定理 1.4.3 已知因子提出

如果 $f(Y)$ 是已知的，它可以从期望中提出：

$$E[f(Y)\cdot X|Y]=f(Y)\cdot E[X|Y]$$

几何视角（线性性质）：在这个几何空间里，随机变量 X 是向量. 而 $f(Y)$ 是什么？因为 $f(Y)$ 是 $\sigma(Y)$-可测的，它位于投影的目标子空间 $\mathcal{M}_Y$ 里. 在这个子空间的视角下，$f(Y)$ 不像是一个向量，而更像是一个**“标量” (Scalar)** 或者说系数.

这个性质相当于说：投影是一个线性算子. $\text{proj}(c\cdot v)=c\cdot\text{proj}(v)$

虽然 c（即 $f(Y)$）本身是随机的，但相对于子空间 $\mathcal{M}_Y$ 而言，它的值是确定的（Fixing the condition）. 所以它就像常数一样，可以直接提出来.

接下来补一些证明，我也看不懂.

假设我们有概率空间 $(\Omega,\mathcal{F},P)$，

- X 是该空间上的一个随机变量，且满足 $E[|X|]<\infty$（即 $X\in L^1(\Omega,\mathcal{F},P)$）.
- $\mathcal{G}$ 是 $\mathcal{F}$ 的一个子 σ-代数（$\mathcal{G}\subseteq\mathcal{F}$）. 这意味着 $\mathcal{G}$ 包含的信息比 $\mathcal{F}$ 少（颗粒度更粗）.

我们需要在较小的 σ-代数 $\mathcal{G}$ 上寻找一个随机变量来近似 X. 为此，我们在 $(\Omega,\mathcal{G})$ 上定义一个新的集函数 ν：

$$\nu(A)=\int_A X\,\mathrm{d}P,\quad\forall A\in\mathcal{G}$$

注意：
(1) 如果 $X\geq 0$，则 ν 是 $(\Omega,\mathcal{G})$ 上的一个**测度**.
(2) 如果 X 取一般值，则 ν 是 $(\Omega,\mathcal{G})$ 上的一个**符号测度（Signed Measure）**.

观察 ν 和 P（限制在 $\mathcal{G}$ 上，记为 $P|_{\mathcal{G}}$）的关系：对于任意 $A\in\mathcal{G}$，如果 $P(A)=0$，那么根据积分的性质，必有 $\nu(A)=\int_A X\,\mathrm{d}P=0$.

这说明 ν 对于 $P|_{\mathcal{G}}$ 是 绝对连续 的（记作 $\nu\ll P|_{\mathcal{G}}$）.

根据 Radon-Nikodym 定理，存在一个唯一的（在 P-几乎处处意义下）$\mathcal{G}$-可测函数 Z，使得对于任意 $A\in\mathcal{G}$：

$$\nu(A)=\int_A Z\,\mathrm{d}P|_{\mathcal{G}}=\int_A Z\,\mathrm{d}P$$

这个 Radon-Nikodym 导数 $Z=\dfrac{\mathrm{d}\nu}{\mathrm{d}P|_{\mathcal{G}}}$ 就被定义为 X 关于 $\mathcal{G}$ 的条件期望. 记作 $E[X|\mathcal{G}]$.

定义 1.4.2 条件期望

$Y=E[X|\mathcal{G}]$ 是满足以下两个条件的唯一随机变量（几乎处处）
(1) **可测性**: Y 是 $\mathcal{G}$-可测的.
(2) **局部平均性**: 对于任意 $A\in\mathcal{G}$，有 $\int_A Y\,\mathrm{d}P=\int_A X\,\mathrm{d}P$.

为了讨论“正交投影”，我们需要内积结构，因此我们将讨论范围限制在 L^2 空间（平方可积随机变量空间）.

几何背景

- 令 $H = L^2(\Omega, \mathcal{F}, P)$ 为希尔伯特空间，其内积定义为 $\langle X, Y\rangle = E[XY]$.
- 令 $M = L^2(\Omega, \mathcal{G}, P)$. 由于 $\mathcal{G} \subseteq \mathcal{F}$，$M$ 是 H 的一个 **闭子空间**.

目标： 证明对于任意 $X \in H$，$E[X|\mathcal{G}]$ 是 X 在子空间 M 上的正交投影. 即证明 $X - E[X|\mathcal{G}]$ 垂直于子空间 M 中的任意向量.

Proof. 设 $Y = E[X|\mathcal{G}]$，我们要证明两点：

(1) $Y \in \mathcal{G}$ （即 Y 在子空间内）.

(2) $(X - Y) \perp M$ （即误差向量垂直于子空间）.

根据定义， Y 是 $\mathcal{G}$-可测的，且由 Jensen 不等式

$$E[Y^2] = E[E[X|\mathcal{G}]^2] \leq E[E[X^2\ |\mathcal{G}]] = E[X^2] < \infty$$

因此 $Y \in M$. 我们需要证明对于任意的 $Z \in M$（即任意 $\mathcal{G}$-可测的平方可积随机变量），都有：

$$\langle X - Y, Z\rangle = 0$$

展开有

$$\langle X - Y, Z\rangle = E[(X - E[X|\mathcal{G}]) \cdot Z] = \int_\Omega (X - E[X|\mathcal{G}]) \cdot Z \, \mathrm{d}P$$

根据条件期望的定义， $\forall A \in \mathcal{G}, \int_A (X - E[X|\mathcal{G}])\, \mathrm{d}P = 0$，这意味着 $(X - E[X|\mathcal{G}])$ 与 任何 $\mathbb{1}_A (A \in \mathcal{G})$ 正交. 由于 $\mathcal{G}$-可测简单函数是示性函数的线性组合，且简单函数在 $L^2(\mathcal{G})$ 中稠密，因此

$$E[(X - E[X|\mathcal{G}])Z] = \langle X - Y, Z\rangle = 0, \quad \forall Z \in M$$

■

由于 $E[X|\mathcal{G} \in L^2(\mathcal{G})$ 且误差向量 $X - E[X|\mathcal{G}]$ 垂直于整个子空间 $L^2(\mathcal{G})$，根据希尔伯特空间的投影定理（Projection Theorem），$E[X|\mathcal{G}]$ 正是 X 在 $L^2(\mathcal{G})$ 上的正交投影.