概率论与数理统计-提纲

Westfox

December 2024

前言：如何认识“概率论”和“数理统计”

我们常常说，抛一枚硬币，“花面朝上”发生的概率是 $\frac{1}{2}$。但是要进一步地问，这个所谓 $\frac{1}{2}$ 的概率，具体是什么呢?

我们会说，这是某一个事件发生的可能性。但是什么叫做 $\frac{1}{2}$ 的可能性呢？你可能会争辩说，扔多枚硬币，会有 $\frac{1}{2}$ 的硬币会“花面朝上”，这就是所谓 $\frac{1}{2}$ 的可能性。这种说法被称为“概率的频率定义”，即认为对于同一分布独立试验，当试验次数趋向于无穷的时候，频率会趋向于一个值，这个值就是概率。

可是频率定义的说法似乎并不让人满意。我们一定要进行很多次甚至无限次试验，这个“概率”的概念才有意义吗？一个带有“无穷”的定义，显然有一些牵强。而且中学时，教科书也强调过频率和概率并不是同一个概念。那么我们该如何理解，“概率”真正的含义呢?

我们可以进一步探寻概率这一数值的含义。考虑另一个情景，一个黑箱里有三个球：两个白球，一个红球；我们现在从这个黑箱里任意摸出一个球来，观察它的颜色。此时，我们会说摸出白球的概率为 $\frac{2}{3}$，摸出红球的概率为 $\frac{1}{3}$

这时候我们所说的概率是什么？摸出什么仍然是不确定的，而黑箱里有多少个白球仿佛也不一定确保实际中摸 9 次一定能出来 6 个白球；我们知道的只是：箱子里，白球的数量是红球的两倍。

在箱子里，有一种不可见的对称性：如果能够摸出白球，那么我们也能以同样的方式摸出红球：球与球之间的关系是对称的，它们在“被摸出”这件事上是对等的。但现在，我们知道有 2 个白球在箱子里，由于这三个球之间都是对等的，我们知道，在能够获取的“球”之中，白球所占有的“份量”

是红球的两倍。

在这个例子中，概率成为了一种“份量”的描述：它描述的是在一次观测中（摸一个球），所有可能的状态之中（摸出的球），某一种特征（颜色）在所有状态之下的“份量”。

这里的“特征”描述的是我们在观测之中所关心的量，并且我们按这个量将样本进行了分组。概率是对于这样的组别而言的，而组别的分类方式不是唯一的，比如我们也可以根据球的大小对于黑箱里的球进行分类，从而得到一个新的“份量系统”。这对应着 σ-代数的概念。

所以，概率描述的是所有状态组成的空间之下，某一种特征所占有的“份量”，或者说“权重”。这种权重被用数值表示。不同特征之间的权重取决于这一数值的比值。

那么，我们也就理解了，为何我们要把概率定义成一种“归一的”量：我们把总体的“份量”之和看作 1，这样每一个特征的数值便能够直接地表示其份量对于总体的份量之比，或者说，其自身的权重对于总体的权重之比。

为什么一定要做这样的比较呢？因为概率衡量的并不是一种绝对的数量，而是某种匀质之下的“对比”。在上一个例子中，我们认为“如果能够摸出白球，那么我们也能以同样的方式摸出红球”。这里的“2：1”并不是数量上的 2：1，而是一种对称性的整体之间的 2：1。我们可以有 200 个白球，100 个红球，当把红球当作一个整体时，白球相当于有 2 个整体，他们之间被“选中”的这一关系是对称的，而相较于红球而言，白球有两份这种“对称的整体”，因而被视为有更高的“权重”。在这里，红球可以有 1 个，10 个，100 个，但是他们都是被当作整体看待的，其权重的考量取决于其与白球的这一状态之间的比较，而不在于其实际上的数量是多少。

在汉语里，“权”本身就是“比较”的意思，“权，然后知轻重”。

现在我们理解了，概率是在所有的状态之下，用来比较不同特征之间“被观测到”的可能性区别的一种权重系统。它代表了，黑箱之下有怎样的“球”。但是箱子里如何并不代表“抽取球”（也就是所谓“观测”）的情况如何，而另一方面，箱子里球的这种分量好像确实影响着客观的规律。对于一个放有 100 个白球，1 个红球的箱子，我们当然知道是有可能抽出红球的（不然为什么我们那么热衷于抽奖和乐透呢:D）；但是经验告诉我们，在大多数的抽取中，我们只能抽到白球，而并非红球。所以，“箱子里”和“抽出来”之间看上去是存在一种联系的。问题是，如何理解这种联系呢？

那前面的 2 白 1 红来举例，一种理解是，由于确实存在着对称性上的差别，“理想情况下”，抽取的情况应该与这种对称性一致。这里的“理想”，也可以理解成一种“相信”或者“认为”：我们“认为”在一次抽取中，出现白球的“合理性”是出现红球的“两倍”，尽管这两种状态本身都是“合理的”。也就是说，我们“双倍地相信”，抽取出的应该是白球。

也就是说，我们预期之中，观测到的结果应该和对应的样本空间中的状态一致；这种“预期”在一次抽样中被认定为：我们更“信服”权重高的样本在观测中出现，这种“信服”的程度和其权重成正比。这种对概率的理解被称为“主观概率 (Subjective Probability)”。这种认识更契合数理统计的“最大似然估计”，“贝叶斯方法”。（也就是说，这些方法的设计更注重体现这一精神）

另一种对于这种联系的阐释是，当观测次数足够多时，最终观测到的结果中的比例会与样本空间中的这种权重保持一致。这种认识被称为概率的“频率定义”。在这种认识之下，我们更关注在概率加权之下的“整体性”表现，如数学期望、方差等。这种认知更契合数理统计中的“假设检验”、“无偏估计方法”。

但是在这种联系中，由于“随机性”，我们事实上是可以建立很多这样的权重系统（概率系统）的；这些权重系统都是合理的，因为即使在某一个权重系统中，一个特定的特征所被赋予的权重低，也不代表其不可能发生，它仍然能够描述实际观测到的事实；因此，我们是无法得知，某一个现实系统中“真正的状态之间的权重”是如何的。这有一种“不可知论”的意味，但这确实是概率论和数理统计所面临难题。更确切地，它将概率论和数理统计区分开来：我们的观测，与我们所创设的概率系统之间，有一种独立性；我们必须要做一些假设和取舍，才能将他们建立起联系。这些假设包括“最大似然”等。

在这一语境之下，概率更像是一种“赋值”，“赋”一字展示了一种主观性，是我们让它拥有这样的结构的，它本身的结构可能并不是这样。数理统计的任务，就是建立一套合乎直觉的标准，确保我们所建立起的概率系统，能够“更为合理”。

本笔记会按照如上的思想，介绍概率论和数理统计之中的相关内容。

同时推荐阅读：Linde, W. (2016). Probability theory: A first course in probability theory and statistics. De Gruyter. https://doi.org/10.1515/9783110466195

接下来，我们具体再区分一下“概率论”和“数理统计”。

概率论在本质上可以被视为一种“赋值”系统，更确切地说，是一种“权重”系统。这里的“权”具有比较的含义，因此，概率所代表的数值并非绝对值，而是相对意义下的数值，用来表示部分与整体之间的关系。这种关系通过**归一化性质**（即所有可能事件的概率总和为 1）得以体现。

概率系统的构建与性质

概率系统的核心在于为事件赋值，这种赋值在数学理论上具有一定的“任意性”。所谓“任意性”，是指我们可以按照以下规则自由构造概率系统：

对任意事件 $A, B \in \mathcal{F}$，只要满足：

- **非负性**：$P(A) \geq 0$；
- **归一性**：$P(\Omega) = 1$，即全集的概率为 1；
- **可加性**：对于两两互不相交的事件族 $\{A_i\}$，有

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

那么这个赋值就被认为是一个合法的概率系统。

这种赋值行为并不直接决定它是否具有现实意义。例如，我们可以为两个事件 A 和 B 分别赋予概率 0.4 和 0.6，只要满足上述规则，这个系统在理论上就是正确的。

概率的意义

尽管概率的赋值本质上是一种数学操作，但在实际应用中，赋值通常具有特定的意义。概率最常见的含义是描述事件发生的可能性，即所谓的“发生性”。在这种语境下：

- 较大的概率值表示事件发生的可能性更大；
- 较小的概率值表示事件发生的可能性更小。

因此，概率大小反映了事件在整体样本空间中的“重要性”或“权重”。

概率论的研究内容

概率论是一门研究概率系统的普遍性质、特殊概率系统的性质，以及基于已知概率系统推导其他信息的理论。例如：

- **普遍性质**：随机变量、概率密度等基本概念；
- **特殊概率系统**：如正态分布、泊松分布等特定概率分布；
- **信息推导**：如随机变量统计特征、条件概率、贝叶斯定理等，用于研究事件之间的依赖关系。

概率论的目标是建立一个严谨的数学框架，用来描述随机现象的规律性。尽管概率系统的构建具有一定的“任意性”，但在现实问题中，合理的概率系统通常是基于实验数据或理论推导得到的。

从概率论到数理统计

由于概率系统的构建是一种人为的赋值行为，它不一定能够完全符合现实中的随机现象。事实上，真实世界中的随机现象复杂多变，我们所创建的概率系统几乎不可能完全表述实际的“真实系统”。因此，我们需要一种方法论，通过对实际观测数据的分析和处理，去“找到”或“逼近”那个真实的概率系统。

数理统计就是这样一门理论，它的主要任务包括：

- **估计真实系统**：通过采样数据，估计随机变量的分布和参数，例如利用样本均值估计总体均值。
- **验证假设**：通过假设检验，判断某种假设是否与数据相符，例如检验两组数据是否来自相同分布。
- **预测与推断**：根据数据推断未来的可能结果，或描述数据之间的关系。

概率论与数理统计的关系

概率论与数理统计是密切相关但又各自独立的学科：

- 概率论提供了描述随机现象的数学框架，用于定义事件和概率的关系；
- 数理统计基于概率论的理论，从实际数据中推断随机现象的规律。

可以将两者的关系理解为：概率论是“自上而下”的理论构建，它从抽象的数学规则出发定义随机现象；而数理统计是“自下而上”的实践探索，它从具体的事实和数据中寻找随机现象的数学描述。

总结

概率论的本质是赋予事件一种比较意义上的权重，其核心在于“部分与全体之比”。**数理统计**则是利用观测数据，从事实中逼近真实的概率系统的方法论。两者共同构成了研究随机现象的数学基础，一个侧重于理论，一个侧重于实践，它们相辅相成，密不可分。

概率论部分

(一) 测度论

测度论的知识，以一种规范的公理框架，回答了一个“概念性”的问题：概率究竟是什么？

1 测度空间

- 可能的状态（全集）被划分并区分成满足特定性质的集合系统，形成 σ-代数。
- σ-代数中的元素之间具有闭合性，特别是在并、交和补运算下的封闭性，使得它们之间更倾向于呈现一种可加性的关系。
- 测度为 σ-代数中的这些被划分归类的群组赋值，并保持 σ-代数的结构性质，例如可数可加性。

测度空间是一种数学框架，用于为一个集合的子集系统性地赋予“大小”或“值”。

需要注意的是，这里的测度还没有被赋予“可能性”的含义。从更为广义的测度空间的定义以及各种各样的测度空间中，我们可以更好理解“测度”的“赋值”含义，进而理解概率的实质其实是一种“赋值”。

它由以下三个基本要素组成：

1.1 全集 (Ω)

Ω 表示所有可能状态的集合，称为全集或状态空间。它包含了我们讨论的所有可能情形。

1.2 σ-代数（$\mathcal{F}$）

为了对全集 Ω 进行区分和分类，我们引入一个 σ-代数 $\mathcal{F}$。σ-代数是 Ω 的一个子集的集合，它满足以下性质：

- **包含全集**：$\Omega \in \mathcal{F}$，即全集本身是 σ-代数的一个元素；
- **对补集封闭**：如果 $A \in \mathcal{F}$，则 $\Omega \setminus A \in \mathcal{F}$，即对于任意一个元素，其补集也是 σ-代数中的元素；
- **对可数并封闭**：如果 $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F}$，则 $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$，即任意可数多个元素的并集仍属于 σ-代数。

σ-代数可以看作是对全集 Ω 的一种划分，这种划分具有逻辑一致性并能进行加和操作。

1.3 测度（μ）

测度 μ 是一个函数：

$$\mu : \Omega \mapsto \mathbb{R}$$

它为 σ-代数中的每个元素赋予一个非负值，且满足以下性质：

- **非负性**：$\mu(A) \geq 0, \forall A \in \mathcal{F}$；
- **空集的测度为零**：$\mu(\emptyset) = 0$；
- **可数可加性**：如果 $\{A_i\}_{i=1}^{\infty}$ 是 $\mathcal{F}$ 中两两互不相交的集合族，则

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

(需要注意的是，我们这里谈论的是标准测度，满足”非负性“；作为其推广，还有一种测度被称作**“符号测度”（带号测度）**，其不满足非负性)

因此，一个测度空间可以形式化地表示为三元组 $(\Omega, \mathcal{F}, \mu)$，其中：

- Ω 是全集，
- $\mathcal{F}$ 是定义在 Ω 上的 σ-代数，
- μ 是定义在 $\mathcal{F}$ 上的测度。

直观理解

可以将测度空间的构成类比为一个分类系统：

- 全集 Ω 是所有可能的状态；
- σ-代数 $\mathcal{F}$ 是对这些状态的有序划分，确保每个划分都能组合出其他划分；
- 测度 μ 是对这些划分赋予“大小”，并且保证这种大小是可加的。

示例说明

为了更直观地理解测度空间的构成和应用，以下通过两个例子来说明这一框架在实际中的作用。

例 1：几何中的面积测度

考虑几何中的面积计算：

- 全集 Ω 表示平面中的某个区域，例如矩形区域 $[0,1]\times[0,1]$；
- σ-代数 $\mathcal{F}$ 表示该区域内所有的可测子集，例如矩形区域中的任意几何形状；
- 测度 μ 表示这些子集的面积，例如：

$$\mu(A) = \text{子集 } A \text{ 的面积}$$

在这种情况下，测度空间 $(\Omega, \mathcal{F}, \mu)$ 可以用于描述几何区域的大小关系。例如，对于 $A = [0, 0.5]\times[0,1]$ 和 $B = [0.5,1]\times[0,1]$，有：

$$\mu(A) = \mu(B) = 0.5 \quad \text{且} \quad \mu(A\cup B) = \mu(A) + \mu(B) = 1$$

例 2：概率论中的事件发生可能性

考虑概率论中的一个抛硬币实验：

- 全集 Ω 表示所有可能的结果，例如 $\Omega = \{\text{正面}, \text{反面}\}$；
- σ-代数 $\mathcal{F}$ 表示事件的集合，例如 $\mathcal{F} = \{\emptyset, \{\text{正面}\}, \{\text{反面}\}, \{\text{正面}, \text{反面}\}\}$；

- 测度 μ 表示每个事件的概率，例如：

$$\mu(\{\text{正面}\}) = 0.5, \quad \mu(\{\text{反面}\}) = 0.5, \quad \mu(\Omega) = 1$$

在这种情况下，测度空间 $(\Omega, \mathcal{F}, \mu)$ 用于描述每个事件发生的可能性。例如，对于事件 {正面} 和 {反面}：

$$\mu(\{\text{正面}\} \cup \{\text{反面}\}) = \mu(\{\text{正面}\}) + \mu(\{\text{反面}\}) = 1$$

这表明，测度空间不仅可以用于几何上的面积计算，也可以应用于描述概率中的事件发生关系。

这样，一个测度空间 $(\Omega, \mathcal{F}, \mu)$ 就成为一个完整的框架，用于研究“分类”和“大小”的关系，例如几何中的面积、概率中的事件发生可能性等。

测度空间广泛应用于积分理论、概率论和函数分析等数学领域，是研究集合及其属性的重要工具。

最后强调测度是**对于 σ 代数上的元素的赋值**！这一点在区分概率和随机变量的时候至关重要（随机变量的定义域是样本空间 Ω）

2 概率测度

> - “归一化”是概率测度根本的特性。
> - “概率”一词本身意味着其所赋予的量值被用来衡量事件发生的可能性，如果这些值并不代表可能性，一般这个测度不被称为“概率”。
> - 测度为 σ-代数中的这些被划分归类的群组赋值，并保持 σ-代数的结构性质，例如可数可加性。

比如，在全文开始的例子里（黑箱摸球），箱子里的每一个球是一个“状态”，它们的集合组成了**“样本空间”**；所考察的颜色是一种“划分标准”，对应着**“σ-代数”**，从而将样本点聚类划分；概率是为其发生性赋予的“大小”，是一种**“测度”**。

在前面已经全面介绍了测度空间的概念及其基本构成 $(\Omega, \mathcal{F}, \mu)$。在此基础上，“概率测度”是对测度的一种特殊约束与解释，将测度理论的“赋

值”思想应用到事件发生可能性上。它与一般测度的本质区别在于以下两点：

2.1 归一化与概率解释

与一般测度相比，概率测度 P 对全集 Ω 的测度值被严格固定为 1，即：

$$P(\Omega) = 1$$

这一归一化条件确保了 P 的取值范围在 $[0,1]$ 之内，使得每个事件都能被解释为具有“发生可能性”的量化指标。这种将测度结果映射为介于 0 与 1 之间的数值，为概率论中“事件发生概率”的概念奠定了严格的数学基础。

2.2 从“大小”到“可能性”的转变

一般测度 μ 可以理解为对集合大小、长度、面积或体积的衡量，数值可为零、有限或无限。而概率测度 P 的数值不仅有限（始终不超过 1），更赋予了实际的含义——它表示事件发生的可能性。通过这一解释上的转化，概率测度在描述随机现象时，不仅维持了测度的数学结构特性（如可数可加性），还为不确定性的量化提供了自然且直观的框架。

需要指出的是，“归一化”是概率测度根本的特性，它同时预示着有界性，使得其可以用来代表“可能性”，但这种代表对于这样定义的测度而言**并不是必须的**。在我们的理解中，“概率”一词本身意味着**其所赋予的量值被用来衡量事件发生的可能性**，如果这些值并不代表可能性，一般这个测度不被称为“概率”。

若一个测度空间所选用的测度为概率测度，则这个空间本身也叫**测度空间**，其中的全集也被称作**样本空间**，其 σ-代数也被称作**事件空间**

总而言之，概率测度是在测度理论基础上附加了一项关键条件（归一性）与明确的解释（概率意义）后所得到的特殊测度。这一特殊性使得概率测度能够直接应用于随机现象的分析和预测，成为概率论研究和应用的核心工具。

（二）概率空间

先前我们已经提到，概率所表示的赋值系统是一种特殊的测度；在此章中，我们进一步分析概率空间具体可以被定义成怎样的结构，以及不同结构所具有的性质。

由于样本空间由三部分组成，我们也可以从这三部分出发，思考如何区分不同种类的概率空间以及相应的性质。这些区分/性质包括：

- 样本空间的种类
- σ-代数的种类
- 概率测度的种类
  - 离散和连续
  - 一维和高维

1 样本空间、σ-代数与概率测度之间的关系

- 样本空间确定了 σ-代数的基本形式，但是 σ-代数本身的结构可以有不同的变化
- 样本空间决定了 σ-代数能够达到的最高的复杂程度，但是 σ-代数并不一定要和样本空间本身一样复杂
- σ-代数具有怎样的结构，我们基本可以认为概率测度具有怎样的结构
- σ-代数是从“集合”的角度实现这个结构，而概率测度是从“数值”的角度实现了这个结构

在概率论中，$(\Omega, \mathcal{F}, P)$ 这三者的相互关系是核心：Ω 表示所有“可能性”的集合（即样本空间），$\mathcal{F}$ 是对这些可能性的一种满足 σ-代数结构的划分或分类，而 P 则为 $\mathcal{F}$ 中的每个事件（子集）赋予符合概率公理的数值。

样本空间与 σ-代数的关系

由于 σ-代数本身是对于样本空间的“分割”，而且这种“分割”方式是相对多样的，所以我们不难推测两者之间的关系：样本空间确定了 σ-代数的基本形式，但是 σ-代数本身的结构可以有不同的变化；样本空间决定了 σ-代数能够达到的最高的复杂程度，但是 σ-代数并不一定要和样本空间本身一样复杂

比如，对于不同的样本空间，其 σ-代数可以为：

- **一维样本空间**：一维样本空间是指样本点仅具有一个特征的情况。例如，当样本空间为 $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ 时，其中每个 ω_i 代表一个可能的实验结果。此时，样本空间的每个元素对应于同一个性状上不同的表型。

- **多维样本空间**：当样本点包含多个特征，例如 $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$，则可定义多维的 Borel σ-代数或其他更复杂的 σ-代数。如果某些维度可以独立拆解，可以通过笛卡尔积的方式生成乘积 σ-代数（Product σ-algebra）。例如，假设 $\Omega = \mathbb{R}^n$ 为 n 维空间，则其 σ-代数通常由笛卡尔积生成，表示多个独立或相关变量的组合。

- **离散样本空间**：离散样本空间指的是包含有限或可数无限个样本点的样本空间。例如，抛硬币实验的样本空间为 $\Omega = \{\text{Heads}, \text{Tails}\}$，或者抛掷骰子的样本空间 $\Omega = \{1, 2, 3, 4, 5, 6\}$。对于此类样本空间，其可能出现的状态是容易确定的，其能够产生的最复杂的 σ-代数也是有限或可数无限的。

- **连续样本空间**：连续样本空间指的是包含不可数无穷个样本点的样本空间，通常与测量相关，如时间、温度等。例如，$\Omega = \mathbb{R}$ 或 $\Omega = [0, 1]$ 表示所有实数或在区间 $[0, 1]$ 内的所有实数。最常见的 σ-代数是由拓扑生成的 Borel σ-代数（Borel σ-algebra）。但是，连续样本空间也可以有简单的 σ-代数。例如，考虑样本空间 $\Omega = \mathbb{R}$，其可以包含简单的 σ-代数，如：$\{\emptyset, (-\infty, 0], (0, \infty), \mathbb{R}\}$。这种 σ 代数仅包含一些基础的区间，并且在某些情况下足够用来描述相关的概率模型。

σ-代数决定概率测度的形式

由于概率测度是定义在 σ-代数代数上的，所以 σ-代数具有怎样的结构，我们基本可以认为概率测度具有怎样的结构；其区别在于，σ-代数是从“集合”的角度实现这个结构，而概率测度是从“数值”的角度实现了这个结构

- 当 σ-代数比较简单（例如有限可数），对应的概率测度通常是**离散型**或**可数可加型**；
- 当 σ-代数由笛卡尔积构成，如 $\mathcal{F}_X \otimes \mathcal{F}_Y$，可以定义**联合概率测度**（Joint Probability Measure），在研究多维随机变量时非常重要；
- 当 σ-代数非常庞大，如 Borel σ-代数，通常可定义**连续型**或**绝对连续型**概率测度，常与 Lebesgue 测度紧密联系。

2 样本空间的类型

在概率论中，样本空间是描述所有可能实验结果的基础结构。根据不同的特性和结构，样本空间可以分为多种类型。本文将介绍几种主要的样本空间类型，包括离散样本空间、连续样本空间、一维与高维样本空间、乘积样本空间，以及具备附加结构的样本空间。

2.1 离散样本空间与连续样本空间

2.1.1 离散样本空间

- 样本空间由有限或可数无限个互不相同的基本事件组成。
- 每个基本事件都可以被明确地列举出来，例如掷骰子的结果 $\{1, 2, 3, 4, 5, 6\}$。

离散样本空间是指样本空间中的基本事件是有限个或可数无限个的。这类样本空间适用于那些结果可以被逐一列举或编号的实验。

2.1.2 连续样本空间

- 样本空间包含不可数多个基本事件，通常对应于实数区间内的连续取值。
- 适用于描述测量类实验，如温度、时间、长度等。

连续样本空间是指样本空间中的基本事件构成一个不可数的集合，通常是实数上的区间。这类样本空间适用于那些结果取值在一个连续范围内的实验。

2.2 一维与高维样本空间

2.2.1 一维样本空间

- 样本空间的每个样本点只包含一个属性或特征。
- 可以产生多个相互正交的 σ-代数，反映同一属性的不同划分。

一维样本空间指的是样本空间中的每个样本点只有一个独立的属性或特征。例如，单次掷骰子的结果只涉及骰子的一个面值。

2.2.2 高维样本空间

- 样本空间中的每个样本点包含多个相互独立或正交的属性。
- 维度的数量取决于样本空间的数学结构，如多个变量或特征的组合。

高维样本空间指的是样本空间中的每个样本点包含多个属性或特征。这些属性通常是相互独立或正交的，使得样本空间的结构更加复杂和多样化。例如，在多变量统计分析中，每个样本点可能包含多个测量值。

2.3 乘积样本空间

2.3.1 定义与构造

- 乘积样本空间由多个单一样本空间通过笛卡尔积构成。
- 每个样本点对应于各个单一样本空间中结果的组合。

乘积样本空间（Cartesian Sample Space）是通过多个单一样本空间的笛卡尔积构造而成的。这种结构特别适用于同时考虑多个随机变量的情况，每个随机变量对应一个独立的样本空间。

2.3.2 直观理解与应用

在多变量实验中，每个随机变量可以被视为高维空间中的一个维度，整个样本空间则是各个随机变量样本空间的笛卡尔积。例如，抛掷两枚骰子的样本空间就是两个离散样本空间 $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$ 的乘积。

2.4 具有附加结构的样本空间

有时，样本空间不仅仅是一个集合，还可以赋予额外的数学结构，以反映元素之间的关系。这些附加结构包括度量空间和拓扑空间等。

2.4.1 度量空间

- 在度量空间中，定义了元素之间的距离。
- 适用于表示空间位置或连续测量的样本空间，如物理空间中的位置点。

度量空间是一种赋予样本空间距离概念的结构，使得可以测量样本点之间的“距离”。例如，二维平面上的样本空间可以使用欧几里得距离来定义样本点之间的距离。

2.4.2 拓扑空间

- 拓扑空间赋予样本空间一种“开集”与“闭集”的概念。
- 在高级概率模型中，拓扑结构有助于处理连续性和极限等性质。

拓扑空间为样本空间定义了“开集”与“闭集”的概念，允许研究样本空间的连续性和边界性质。这在处理复杂的概率模型和函数空间时尤为重要。

直观理解与示例

为了更好地理解不同类型的样本空间，以下通过具体例子进行说明。

例 1：抛硬币实验

- **样本空间**：$\Omega = \{\text{正面}, \text{反面}\}$，这是一个有限的离散样本空间。
- **维度**：一维样本空间，每个样本点只有一个属性——硬币的结果。
- **乘积样本空间**：若抛两次硬币，样本空间为

$$\begin{aligned}\Omega \times \Omega = \{&(\text{正面}, \text{正面}),\\ &(\text{正面}, \text{反面}),\\ &(\text{反面}, \text{正面}),\\ &(\text{反面}, \text{反面})\}\end{aligned}$$

形成一个二维离散样本空间。

例 2：温度测量

- **样本空间**：$\Omega = \mathbb{R}$，即所有实数，构成一个连续样本空间。
- **维度**：一维样本空间，每个样本点代表一个具体的温度值。
- **附加结构**：可以赋予度量空间结构，使用标准的绝对值距离来衡量温度值之间的距离。

例 3：股票价格变化

- **样本空间**：$\Omega = \mathbb{R}^n$，表示在 n 个时间点上的股票价格，形成一个高维连续样本空间。
- **乘积结构**：每个时间点的价格可以视为一个独立的随机变量，其样本空间的乘积构成整个样本空间。
- **拓扑结构**：可以赋予样本空间拓扑结构，以研究价格变化的连续性和极限行为。

通过上述例子，可以看出不同类型的样本空间在实际应用中的多样性和灵活性。理解样本空间的结构是深入学习概率论和统计学的基础。

3 σ-代数的类型

σ-代数（σ-Algebra）是测度理论和概率论中的基本概念，用于系统地描述样本空间中的事件结构。不同类型的 σ-代数具有不同的性质和应用场景。本文将介绍几种常见的 σ-代数类型，包括幂集、平凡 σ-代数、Borel σ-代数、生成 σ-代数以及乘积 σ-代数，并提供相应的形式定义和直观解释。

3.1 幂集 σ-代数

3.1.1 定义与性质

- 幂集 σ-代数是样本空间所有子集组成的 σ-代数。
- 对于任意集合 Ω，幂集 2^Ω 满足 σ-代数的所有公理。

幂集 σ-代数，记作 2^Ω，是包含样本空间 Ω 所有可能子集的 σ-代数。由于它包含了所有子集，因此在 σ-代数中具有最大的复杂性和灵活性。

3.1.2 直观理解与应用

幂集 σ-代数适用于那些需要考虑样本空间中所有可能事件的情形。然而，在实际应用中，由于幂集通常包含不可数多个子集，其在处理无限样本空间时可能导致复杂性过高，因此在许多情况下会选择更为简洁的 σ-代数结构。

3.2 平凡 σ-代数

3.2.1 定义与性质

- 平凡 σ-代数仅包含全集和空集两个元素。
- 形式化定义为 $\mathcal{F} = \{\emptyset, \Omega\}$。

平凡 σ-代数是最简单的 σ-代数，仅包含样本空间 Ω 和空集 $\emptyset$。这种 σ-代数适用于无法区分任何非平凡事件的情形。

3.2.2 直观理解与应用

在平凡 σ-代数下，唯一可以讨论的事件是整个样本空间和不发生任何事件。这在某些极端或理论性的分析中可能有用，但在实际概率论中应用较少。

3.3 Borel σ-代数

3.3.1 定义与性质

- Borel σ-代数是实数集上的 σ-代数，生成自所有开集。
- 记作 $\mathcal{B}$，它包含所有可以通过开集的可数并、可数交及补集运算得到的集合。

Borel σ-代数是实数集 $\mathbb{R}$ 上的重要 σ-代数，由所有开集通过 σ-代数运算生成。它是处理实值随机变量及其分布的基础。

3.3.2 直观理解与应用

Borel σ-代数包含了大多数在分析和概率论中遇到的常见集合，如开区间、闭区间、半开区间等。由于其结构的丰富性，Borel σ-代数广泛应用于测度理论、实分析以及概率分布的定义中。

3.4 生成 σ-代数

3.4.1 形式定义

设 $\mathcal{A}$ 是样本空间 Ω 的一个子集系统，$\sigma(\mathcal{A})$ 表示由 $\mathcal{A}$ 生成的最小 σ-代数，即满足：

$$\mathcal{A} \subseteq \sigma(\mathcal{A}),$$

且 $\sigma(\mathcal{A})$ 包含所有满足 σ-代数性质的集合。

3.4.2 性质与构造

生成 σ-代数 $\sigma(\mathcal{A})$ 具有以下性质：

- **包含性**：任何包含 $\mathcal{A}$ 的 σ-代数都包含 $\sigma(\mathcal{A})$。
- **封闭性**：对于 $\mathcal{A}$ 中的任意集合，通过可数并、可数交和补集运算得到的所有集合均属于 $\sigma(\mathcal{A})$。

3.4.3 直观理解与应用

生成 σ-代数将一组初始事件通过 σ-代数的运算封闭性扩展为一个完整的 σ-代数。这一过程类似于通过最小生成系统构建一个包含所有必要事件的框架，确保所有基于初始事件的复杂事件都被包含在内。

3.4.4 示例

假设 $\Omega = \{a, b, c\}$，令 $\mathcal{A} = \{\{a\}, \{b, c\}\}$。则由 $\mathcal{A}$ 生成的 σ-代数 $\sigma(\mathcal{A})$ 为：

$$\sigma(\mathcal{A}) = \{\emptyset, \{a\}, \{b, c\}, \{a, b, c\}\}$$

3.5 乘积 σ-代数

3.5.1 定义与构造

设 Ω_1 和 Ω_2 分别是两个样本空间，其对应的 σ-代数为 $\mathcal{F}_1$ 和 $\mathcal{F}_2$。则乘积 σ-代数 $\mathcal{F}_1 \otimes \mathcal{F}_2$ 定义为在 $\Omega_1 \times \Omega_2$ 上的最小 σ-代数，使得对于所有 $A \in \mathcal{F}_1$ 和 $B \in \mathcal{F}_2$，集合 $A \times B$ 属于 $\mathcal{F}_1 \otimes \mathcal{F}_2$。

3.5.2 性质与应用

乘积 σ-代数允许在多维样本空间中处理多个独立随机变量的联合分布。它确保了各个维度上的事件能够通过简单的笛卡尔积运算组合起来，同时保持 σ-代数的封闭性。

3.5.3 直观理解与应用

在处理多个随机变量时，乘积 σ-代数提供了一种系统化的方法来定义联合事件。例如，考虑两个独立的抛硬币实验，样本空间为 $\{\text{正面}, \text{反面}\} \times \{\text{正面}, \text{反面}\}$，其乘积 σ-代数包含所有可能的组合事件，如 $\{\text{正面}\} \times \{\text{反面}\}$。

3.6 联合 σ-代数

3.6.1 定义与性质

设 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 是同一样本空间 Ω 上的两个 σ-代数。联合 σ-代数 $\mathcal{F}_1 \vee \mathcal{F}_2$ 定义为包含 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 的最小 σ-代数。形式上：

$$\mathcal{F}_1 \vee \mathcal{F}_2 = \sigma(\mathcal{F}_1 \cup \mathcal{F}_2)$$

3.6.2 性质与应用

联合 σ-代数结合了两个 σ-代数中的所有事件及其通过 σ-代数运算生成的事件。这在需要同时考虑多个事件系统的情况下尤为重要，如多重实验或多重测量情景。

3.6.3 直观理解与应用

联合 σ-代数允许在一个样本空间中同时处理多个独立的事件系统。例如，若 $\mathcal{F}_1$ 描述了某一属性的事件，而 $\mathcal{F}_2$ 描述了另一属性的事件，联合 σ-代数则包含了所有可能的组合事件及其逻辑运算结果。

3.7 子 σ-代数

3.7.1 定义与性质

- 子 σ-代数是某一 σ-代数的子集，同时自身也是一个 σ-代数。
- 形式上，若 $\mathcal{G} \subseteq \mathcal{F}$ 且 $\mathcal{G}$ 满足 σ-代数的所有公理，则 $\mathcal{G}$ 是 $\mathcal{F}$ 的子 σ-代数。

子 σ-代数是在一个更大的 σ-代数中自身构成 σ-代数的子集。它通常用于限制考虑的事件范围或聚焦于特定的事件子系统。

3.7.2 直观理解与应用

在实际应用中，子 σ-代数常用于定义条件概率或处理部分可观测的信息。例如，在一个复杂的概率模型中，子 σ-代数可以表示对某些随机变量信息的“过滤”，从而进行条件分析。

3.7.3 示例

设 $\Omega = \{a, b, c, d\}$，并设 $\mathcal{F} = \{\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$。则 $\mathcal{G} = \{\emptyset, \{a, b\}, \{a, b, c, d\}\}$ 不是 σ-代数，因为它不对补集封闭；然而:

$\mathcal{G}' = \{\emptyset, \{a, b, c, d\}\}$

是 $\mathcal{F}$ 的子 σ-代数。

3.8 完备 σ-代数

3.8.1 定义与性质

- 完备 σ-代数是指任何属于 σ-代数的零测度集的子集也属于该 σ-代数。
- 如果一个 σ-代数满足：对于所有的 $A \in \mathcal{F}$ 和 $\mu(A) = 0$，任意子集 $B \subseteq A$ 也属于 $\mathcal{F}$，则称 $\mathcal{F}$ 是完备的。

完备 σ-代数是在一个 σ-代数的基础上进一步完善，使得所有零测度集的子集也被包含在内。换句话说，除了 σ-代数本身的所有事件外，任何可以从这些事件中通过包含零测度集的方式衍生出的事件也属于该 σ-代数。

3.8.2 直观理解与应用

完备 σ-代数确保了在处理概率测度时，不会遗漏任何“微小”事件，尤其是那些在概率上几乎不可能发生的事件。这在实际应用中尤为重要，例如在实分析和概率论中，处理几乎处处成立的性质时，完备 σ-代数提供了必要的框架。

3.8.3 示例

考虑实数集上的 Borel σ-代数 $\mathcal{B}$，如果我们将所有 Lebesgue 零测度的集合也加入到 $\mathcal{B}$ 中，则得到的 σ-代数是完备的。这意味着不仅所有 Borel 集合属于该 σ-代数，而且所有 Lebesgue 零测度集合及其子集也属于该 σ-代数。

3.9 子 σ-代数之间的独立性

3.9.1 定义与性质

- 子 σ-代数之间的独立性指的是两个 σ-代数中的事件在概率上是独立的。
- 形式上，若 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 是样本空间 Ω 上的两个 σ-代数，则 $\mathcal{F}_1$ 与 $\mathcal{F}_2$ 独立当且仅当对于所有 $A \in \mathcal{F}_1$ 和 $B \in \mathcal{F}_2$，有：

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

子 σ-代数之间的独立性表示两个 σ-代数中的事件在概率上相互独立，意味着一个 σ-代数中的事件发生与否不影响另一个 σ-代数中事件发生的概率。这种独立性可以看作是两个 σ-代数之间存在“正交性”。

3.9.2 直观理解与应用

独立性意味着对应特征的出现之间没有任何关联或影响。例如，在多维随机变量中，不同维度上的随机变量可以通过独立的 σ-代数来描述其独立性，从而简化联合概率的计算和分析。

3.9.3 示例

设样本空间 $\Omega = \{a, b, c, d\}$，概率测度定义为 $P(a) = P(b) = P(c) = P(d) = 0.25$。定义两个子 σ-代数：

$$\mathcal{F}_1 = \{\emptyset, \{a, b\}, \{c, d\}, \Omega\}$$

$$\mathcal{F}_2 = \{\emptyset, \{a, c\}, \{b, d\}, \Omega\}$$

则 $\mathcal{F}_1$ 与 $\mathcal{F}_2$ 是独立的 σ-代数，因为对于任意 $A \in \mathcal{F}_1$ 和 $B \in \mathcal{F}_2$，有：

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

例如：

$$P(\{a, b\} \cap \{a, c\}) = P(\{a\}) = 0.25 = 0.5 \times 0.5 = P(\{a, b\}) \cdot P(\{a, c\})$$

3.10 总结

不同类型的 σ-代数通过其结构和生成方式在概率论和测度理论中扮演着不同的角色。幂集 σ-代数具有最大的灵活性但复杂度高，平凡 σ-代数极为简单而适用范围有限，Borel σ-代数在实分析中广泛应用，生成 σ-代数提供了从基础事件构建复杂事件系统的方法，乘积 σ-代数适用于多维随机变量的联合分析，而子 σ-代数则用于限定和条件化事件系统，完备 σ-代数确保所有零测度集的子集也被包含，而子 σ-代数之间的独立性则用于描述不同事件系统之间的相互独立性。理解这些不同类型的 σ-代数有助于在实际问题中选择合适的数学工具进行建模和分析。

4 概率测度及其类型

- 根据所关注子集的“可数性”与“不可数性”以及所依赖的基准测度，概率测度一般可分为离散型与连续型，或者介于两者之间的混合型。
- 连续型概率测度的构造离不开对 Lebesgue 测度的依托，后者提供了“长度/面积/体积”这套**基准刻度**，使得我们能在不可数的实数轴或高维空间中，对区间或其他可测集合的概率进行**平滑**、**连续**的刻画。这与离散型测度中“单点也可有正测度”的做法形成了鲜明对比。
- **联合概率测度**强调仍然**依托于同一个**样本空间，通过生成或合并子 σ-代数进行刻画；
- **乘积概率测度**是在**新的**笛卡尔积空间上定义，常用于多个原本独立的系统相互组合的场景。
- **边缘概率测度**（Marginal Measure）本质上是将不感兴趣的维度进行**求和**或**积分**“合并”掉

4.1 离散概率测度与连续概率测度

根据所关注的可测集合在某种意义上（常见于基于基准测度的区分）是否为“可数”或“不可数”，以及事件测度是否依赖于“点集”或“区间”，我们可以将概率测度粗略地划分为离散型与连续型（当然也有一些混合型情形，这里暂不展开）。

离散概率测度

离散性与可数子集

当样本空间中真正“有可能发生”或“我们关注”的事件只在一个**有限或可数无穷**的子集中具有正测度，而对其他点都赋值 0 时，便形成了典型

的离散概率测度。具体而言，如果存在可数集合 $\{\omega_i\}_{i=1}^{\infty}$，满足

$$\Omega = \bigcup_{i=1}^{\infty} \{\omega_i\} \quad \text{且} \quad P(\{\omega_i\}) = p_i,$$

并且

$$\sum_{i=1}^{\infty} p_i = 1,$$

那么 P 称为离散型概率测度。

- 在离散型测度中，**单点**通常具有正测度（即 $P(\{\omega_i\}) > 0$），这种测度往往借由 **概率质量函数 (PMF, Probability Mass Function)** 来描述。
- 对于任何事件 $A \subset \Omega$（可测），因为它可以表示为若干单点的并集，那么

$$P(A) = \sum_{\omega_i \in A} P(\{\omega_i\}).$$

例：伯努利测度与泊松测度

- **伯努利测度**：$\Omega = \{\text{正面}, \text{反面}\}$，$P(\{\text{正面}\}) = p$，$P(\{\text{反面}\}) = 1 - p$，这是有限可数离散测度的简单示例。
- **泊松测度**：$\Omega = \{0, 1, 2, \dots\}$，$P(\{k\}) = e^{-\lambda} \frac{\lambda^k}{k!}$。这是可数无穷集合上的典型离散测度。

连续概率测度

Lebesgue 测度简介

在构建**连续型概率测度**的理论时，我们往往需要借助一个“基准测度”，使得概率测度能够**相对于**该基准进行定义。对实数轴 $\mathbb{R}$ 或更高维空间 $\mathbb{R}^n$ 来说，最常用的基准测度便是 **Lebesgue 测度**（记作 λ 或 m）。简单而言：

- 当 $n = 1$ 时，Lebesgue 测度可以直观地理解为**区间的长度**；在更高维 $\mathbb{R}^n$ 时，则相应地对应**体积**、**面积**等概念的推广。
- Lebesgue 测度能够对“绝大多数”日常关心的子集进行测量，并具备**平移不变性**、**可数可加性**等核心性质。

- 在 Lebesgue 测度中，任何单点 $\{x\}$ 或者有限/可数多个点的集合都具有**零测度**；任意区间/矩形的测度等于它们的“长度/面积/体积”。

得益于这些性质，Lebesgue 测度成为刻画“连续结构”事件集时最自然、最通用的参考；很多连续型概率测度都以它为基准来进行定义（也就是所谓的**绝对连续**——见下文）。

基于 Lebesgue 测度的绝对连续性

当我们在更广泛的情境下考虑**不可数无穷**个可能状态（比如 $\mathbb{R}$ 上的区间）并且对点集合赋予的测度通常是 0 时，我们可以获得所谓的**连续型概率测度**。具体来说：

$$\forall\, x \in \mathbb{R}, \quad P(\{x\}) = 0.$$

满足上述性质且与 Lebesgue 测度 λ 存在一种“**绝对连续**”关系（即若 $\lambda(A) = 0$，则 $P(A) = 0$）的测度，通常称为 **连续型概率测度**。直观地说，如果一个事件集 A 在 Lebesgue 意义上没有“面积”（即 $\lambda(A) = 0$），那么它在这个概率测度下也应当没有“可能性”（即 $P(A) = 0$）。

在这种测度下，我们往往能用**概率密度函数 (PDF, Probability Density Function)** 来辅助描述。对于任意区间 $[a, b] \subset \mathbb{R}$，有

$$P([a, b]) = \int_a^b f(x)\, dx,$$

其中 $f(x)$ 为非负可积函数并满足

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

这里的 $f(x)$ 称为该概率测度相对于 Lebesgue 测度的 **Radon-Nikodým 导数**或**概率密度函数**，它能在一定程度上反映出该测度是如何在数轴上分布的。

为什么叫“连续”

“连续”一词的本质涵义在于：这种测度对点或者任何“零测度”集合都赋 0，且可以用一个可积函数 f 的**积分**来表示。与离散情形相比，连续型测度没有“点质量”或“原子”，因此对区间的测度随着区间端点的连续变

化而进行**平滑**地改变，没有任何“跳点”式的离散集中质量。

例：指数测度与正态测度

- **指数测度**：设样本空间为 $\Omega = \mathbb{R}_{\geq 0}$，若概率密度函数为

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

则对任意可测集合 $A \subset \mathbb{R}_{\geq 0}$ 有

$$P(A) = \int_A \lambda e^{-\lambda x}\, dx.$$

该分布常用于刻画具有**无记忆**性质的事件间隔时间。

- **正态测度**：设样本空间为 $\Omega = \mathbb{R}$，若概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\tfrac{(x-\mu)^2}{2\sigma^2}\Big),$$

则对区间 $[a, b]$，

$$P([a,b]) \;= \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\tfrac{(x-\mu)^2}{2\sigma^2}\Big)\, dx.$$

正态分布是应用极为广泛、理论地位极高的一类**连续型**分布，与许多极限定理、噪声分布、误差分析等密切相关。

小结： 连续型概率测度的构造离不开对 Lebesgue 测度的依托，后者提供了“长度/面积/体积”这套**基准刻度**，使得我们能在不可数的实数轴或高维空间中，对区间或其他可测集合的概率进行**平滑**、**连续**的刻画。这与离散型测度中“单点也可有正测度”的做法形成了鲜明对比。

4.2 高维与混合情形

高维样本空间

在实际问题中，样本空间往往是**多维或高维**结构。当我们说“高维”时，并非仅仅指坐标轴的维度（如三维坐标系），还可能指**属性维度**或**特征维度**。例如：

- 一个球体不仅只有“位置”可以变化，还可能包含“大小、颜色、表面光滑程度、材质”等多个维度，每一个维度都可以看作一个坐标轴，相互组合后构成一个高维空间；
- 在多元统计中，常常要分析 $(X_1, X_2, \ldots, X_n)$ 这样的**多维随机变量**或**向量随机变量**。每个分量 X_i 代表一个不同的度量或属性（如温度、压力、流量等），这些分量共同构成一个样本的“多维描述”。

在这种情境下，若我们将所有可能的“多重属性组合”视作样本空间 Ω 的元素，那么 Ω 可能就在 $\mathbb{R}^n$（甚至更一般的度量空间）上。为此，需要在 $\mathbb{R}^n$ 上引入相应的 **σ-代数**和**测度结构**，才能对高维事件（可测集合）进行系统刻画。

为什么样本空间可以是“高维的”？ 究其根本，是因为现实中可观测、可度量的特征往往不止一个维度：在低维情形下（例如“抛硬币”只有正反面），我们只需关心单一的结果；但在更复杂的系统中，每个**独立或可区分**的属性（物理量、化学量、生物学量等）都可能扩充样本空间的维度。高维样本空间为我们提供了更灵活、更全面的刻画方式。

联合概率测度与乘积概率测度

当我们研究**多个随机变量**时，可以从两个不同的角度来谈“联合”：

1. **联合概率测度（Joint Probability Measure）**：

   这是在**同一个**样本空间 $(\Omega, \mathcal{F}, P)$ 上，先拥有多个子 σ-代数（例如 $\sigma(X)$ 和 $\sigma(Y)$），再把这些子 σ-代数合并生成

   $$\sigma(X, Y) \;=\; \sigma(X) \;\vee\; \sigma(Y).$$

   在此框架下，“联合测度”只是**同一个测度** P 在 $\sigma(X, Y)$ 上的限制与刻画，对应的是

   $$P\big(\{\omega \in \Omega \mid X(\omega) \in A,\; Y(\omega) \in B\}\big),$$

   其中 A, B 分别是 X, Y 所在的值域中的可测集合。换言之，这是一种**更一般**的描述方式，不需要显式去构造高维（或笛卡尔积）空间。

2. **乘积概率测度（Product Probability Measure）**：

   若我们有两个(或多个)**分别定义好的**概率空间 $(\Omega_1, \mathcal{F}_1, P_1)$ 和 $(\Omega_2, \mathcal{F}_2, P_2)$，我们也可以在**笛卡尔积**空间

   $$\Omega \;=\; \Omega_1 \times \Omega_2, \quad \mathcal{F} \;=\; \mathcal{F}_1 \otimes \mathcal{F}_2$$

   上定义一个新的测度 P，满足对可测矩形 $A_1 \times A_2$ ($A_1 \in \mathcal{F}_1$, $A_2 \in \mathcal{F}_2$) 有

   $$P(A_1 \times A_2) \;=\; P_1(A_1)\, P_2(A_2).$$

   称之为**乘积测度**。当 P_1 与 P_2 对应的随机现象可视为“独立”时，这个定义尤为常见。

   进一步地，若我们有 n 个空间 $(\Omega_i, \mathcal{F}_i, P_i)$，就可依次构造

   $$P \;=\; P_1 \otimes P_2 \otimes \cdots \otimes P_n$$

   在

   $$\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n, \quad \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \cdots \otimes \mathcal{F}_n$$

   上的乘积测度，这在**高维空间**中研究多重独立随机机制时十分便利。

两者的差异与联系：

- **定义空间的差异**：
  - **联合概率测度**强调仍然**依托于同一个**样本空间，通过生成或合并子 σ-代数进行刻画；
  - **乘积概率测度**是在**新的**笛卡尔积空间上定义，常用于多个原本独立的系统相互组合的场景。
- **高维空间的便利性**：
  - 在很多应用中，如果随机变量之间“正好独立”，就可以借用“乘积空间”构造**乘积测度**，并且具有分离性（$P(A_1 \times A_2) = P_1(A_1) \cdot P_2(A_2)$）的优雅性质；
  - 若随机变量并非独立，依然可以在相同的笛卡尔积空间上定义一个更**一般的联合测度**（不一定分解为乘积），或者直接在原空间通

过 $\sigma(X) \vee \sigma(Y)$ 等方式表征联合分布。

- **抽象测度论的视角**：
  - 无论是**同一个空间**上的联合分布，还是**笛卡尔积空间**上的乘积分布，都可以用“推送测度（Pushforward Measure）”来相互转化：若在 Ω 上定义了 (X, Y)，则可在 $\Omega_1 \times \Omega_2$ 上自然地产生一个联合分布 $(X, Y)_* P$；
  - 反之，若在乘积空间已有测度 $P_1 \otimes P_2$，则可以反推到原空间（若存在相应的映射）来理解它对应的联合事件测度。

边缘概率测度

当我们有了高维或联合分布后，往往也会关心“某几个分量”的概率测度，而并不在意其它分量具体取值如何。例如，在 (X_1, X_2, X_3) 的三维随机向量中，我们可能只关注 X_1 发生在某个区间的概率，而不在意 X_2、X_3 的取值。此时就需要使用**边缘概率测度**（Marginal Measure）。它本质上是将不感兴趣的维度进行**求和**或**积分**“合并”掉。例如：

$$P_{X_1}(A) \;=\; \int_{\Omega_2} \int_{\Omega_3} P\Big(A \times d\omega_2 \times d\omega_3\Big),$$

其中 A 是 X_1 所在空间的一部分，而 Ω_2、Ω_3 分别代表 X_2、X_3 的整个取值域，并通过对另外两个维度的所有可能取值进行“累加”或“积分”来得到 X_1 的**边缘分布**。

边缘测度的要点

- 无论离散（求和）还是连续（积分），边缘化操作都是把不关心的维度“合并”在一起；
- 边缘分布并不保留对其它维度事件的区分能力，因此信息量会有所丢失，但在许多情形下却是我们真正需要分析的目标，比如只关心某个单一特征的分布情况。

高维概率测度的常见示例

- n **维正态分布**：在 $\mathbb{R}^n$ 上定义，具有密度函数

$$f(\mathbf{x}) \;=\; \frac{1}{\sqrt{(2\pi)^n\,\det(\Sigma)}}\exp\Big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\,\Sigma^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\Big),$$

  这里 $\mathbf{x}\in\mathbb{R}^n$, $\boldsymbol{\mu}\in\mathbb{R}^n$ 为均值向量，Σ 为对称正定协方差矩阵。

- **Dirichlet 分布**：在符合概率单纯形（如 $\{(x_1,x_2,\ldots,x_n)\mid x_i\geq 0,\sum_i x_i=1\}$）上定义的一类重要分布，广泛用于贝叶斯统计建模。

混合型测度

在一些场景中，我们会遇到既含离散部分又含连续部分的**混合分布**。例如，“观测结果可能完全集中在某几个值上，也可能在某些区间里连续分布”。对于混合分布，我们可以把测度拆解为离散成分和连续成分之和：

$$P(A) \;=\; \sum_i p_i\,\mathbf{1}_{\{\omega_i\in A\}} \;+\; \int_A f(x)\,dx,$$

其中 $\{\omega_i\}$ 的测度由离散部分给出，区间部分的测度由连续部分给出。更复杂的情况下，还可拓展到**高维空间**或**多变量情形**，比如在 $\mathbb{R}^n$ 中部分维度离散、部分维度连续的状况，以及带有独特几何结构的混合型分布。

综上所述，高维样本空间和混合型测度在理论与应用中都极为常见：当多个随机机制**相互独立**时，在**笛卡尔积**空间上使用**乘积测度**可以大幅简化分析；而**联合概率测度**则是更通用的观点，不一定需要高维空间来显性构造，也可以直接在同一个样本空间的子 σ-代数上进行描述。无论在高维还是低维情形，核心仍是对可测集合的**概率测度**赋值，这也与离散型、连续型、以及混合型的基本概念保持一致。

4.3 经典概率分布及其测度特征

在概率论中，我们经常会根据实际现象的离散性或连续性，选取不同的分布来建构概率测度。以下简要介绍常见的离散分布与连续分布，并为每一种分布给出相应的测度定义、特殊性质与常用场景。

需要特别注意的是：实际上这里的“分布”指的是随机变量语境下的测度，由于其重要性我们在这里提前提到，对于“分布”相关的概念请查看本

笔记的（三）随机变量的第二章节!

4.3.1 伯努利分布（Bernoulli Distribution）

定义及测度形式 伯努利分布通常被定义在样本空间 $\Omega = \{0, 1\}$ 上，其测度可表示为:

$$P(\{1\}) = p, \quad P(\{0\}) = 1 - p, \quad 0 \leq p \leq 1.$$

这是一个最简单的离散分布例子，每个样本点（0 或 1）具有正测度。

特殊性质

- **两点空间**: 只有“成功/失败”两种可能性，结构极度简单;
- **刻画基本试验结果**: 常用于描述“一次性事件是否发生”的测度。

应用场景

- **二分类问题**: 如生产线上检测产品合格（1）或不合格（0）;
- **原子事件**: 在更复杂分布（如二项分布）中常作为“单次试验测度”构成基元。

4.3.2 二项分布（Binomial Distribution）

定义及测度形式 设 $\Omega = \{0, 1, \ldots, n\}$，在 Ω 上的二项分布测度可写为:

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

其中 p 为单次伯努利测度中“成功”的概率。

特殊性质

- **有限可数支持**: 只在 $\{0, \ldots, n\}$ 上取正值;
- **与伯努利的关系**: 由 n 次独立伯努利测度构成的“总成功次数”的测度。

应用场景

- **固定次数试验的统计**：如“独立抽样 n 次，关注其中成功的总次数”；
- **质量检测中的成功率估计**：多次测试合格产品数量的分布等。

4.3.3 超几何分布（Hypergeometric Distribution）

定义及测度形式 超几何分布往往刻画在“从有限人群中不放回抽取”这一背景下的测度。设 $\Omega = \{0, 1, \ldots, k_{\max}\}$，其中 $k_{\max} \leq n$，若在总体 N 个元素中有 K 个“标记”（或成功），抽取 n 个元素后，成功元素数目为 k 的测度为：

$$P(\{k\}) \;=\; \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, \ldots, \min(K, n).$$

特殊性质

- **无放回**：与二项分布不同，抽取后不回放，总体内部结构会变化；
- **有限资源**：对有限量样本的真实情况描述更贴近实际（如抽卡、实物检测等）。

应用场景

- **质量检验**：从有限批次产品中抽样检测，不放回时合格品数量的分布；
- **生物统计**：从有限群体（带有标记个体）中不放回抽取后所得到的标记数。

4.3.4 泊松分布（Poisson Distribution）

定义及测度形式 泊松分布定义在 $\Omega = \{0, 1, 2, \ldots\}$（可数无穷集），其测度为：

$$P(\{k\}) \;=\; e^{-\lambda}\,\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots, \quad \lambda > 0.$$

特殊性质

- **单参数分布**：只需 λ 即可刻画整个分布结构；

- **稀疏事件**：常用于描述单位区间或单位时间内“随机到达”或“随机发生”次数的测度；
- **极限逼近**：当二项分布的 n 很大、p 很小且 $np = \lambda$ 时，泊松分布可视为二项分布的极限。

应用场景

- **随机到达过程**：电话呼入、网络流量、放射性粒子衰变等；
- **排队论**：描述顾客、车辆等在一段时间内到达队列的次数。

4.3.5 几何分布（Geometric Distribution）[1]

定义及测度形式 几何分布也定义于可数无穷集（如 $\{1, 2, 3, \dots\}$），其测度最常见的形式是：

$$P(\{k\}) \;=\; (1-p)^{k-1}\, p, \quad k = 1, 2, 3, \dots, \quad 0 < p \leq 1.$$

这可以解释为“第一次成功出现在第 k 次试验”的概率。

特殊性质

- **记忆无损性**：几何分布与指数分布一样，具有“无记忆”属性；
- **单峰分布**：取值在正整数上且分布呈单调递减或单调递增的形式（取决于定义方式）。

应用场景

- **重复试验**：描述在伯努利试验中“直到第一次成功所需的次数”；
- **故障测试**：机器或系统在重复尝试后首次成功/首次故障发生的测度。

4.3.6 Gamma 分布及相关分布

定义及测度形式

[1] 有时也被称为“集合分布”，但主流称谓为“几何分布”。

- **Gamma 分布**：定义在 $\mathbb{R}_{>0}$ 上的连续测度，一般形式为

$$P(A) \;=\; \int_A f(x)\,dx, \quad f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

  其中 $\alpha > 0,\ \beta > 0$。

- χ^2 **分布**：是 Gamma 分布在 $\alpha = \frac{\nu}{2},\ \beta = \frac{1}{2}$ 时的特例，常写作 χ^2_ν，其中 ν 为自由度。

- **Beta 分布**：定义域通常取在 $(0,1)$，密度函数形式为

$$f(x) = \frac{x^{\alpha-1}\,(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)},$$

  其中 $\mathrm{B}(\alpha,\beta)$ 为 Beta 函数。Beta 分布不直接是 Gamma 分布，但常与 Gamma 分布联系紧密（比如 Beta 分布可以被视为两个 Gamma 分布比值的归一化）。

特殊性质

- **Gamma 分布**：若将泊松过程的到达事件视为一种度量，Gamma 分布可用来描述“第 α 个事件到达时间”。

- χ^2 **分布**：在统计推断中经常出现，如方差分析、卡方检验等。

- **Beta 分布**：在先验-后验分析里扮演重要角色，可以自然地表达“区间 $(0,1)$ 上的比例/概率”的不确定性。

应用场景

- **寿命与可靠性**：Gamma 分布常用于描述设备寿命或故障间隔；

- **假设检验与推断**：χ^2 分布、Beta 分布均在统计检验与参数估计中极为常见。

4.3.7 正态分布（Normal Distribution）

定义及测度形式 正态分布又称高斯分布，其测度在 $\mathbb{R}$ 上以密度函数给出：

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\, \exp\!\Big(-\frac{(x-\mu)^2}{2\,\sigma^2}\Big), \quad x \in \mathbb{R}.$$

对可测集合 A 的测度即为 $P(A) = \int_A f(x)\,dx$。

特殊性质

- **对称性**：关于 $x = \mu$ 对称，且最高峰在 μ；
- **无限可分与中心极限定理**：大量独立同分布随机现象的和往往逼近正态分布；
- **光滑分布**：在统计、信号处理等领域最常见的“误差或噪声”分布模型。

应用场景

- **误差分析**：描述测量误差、观测噪声等；
- **大数定律与中心极限定理**：大量独立效应综合下出现的极限分布。

4.3.8 指数分布（Exponential Distribution）

定义及测度形式 指数分布常定义在 $\mathbb{R}_{\geq 0}$ 上，其密度函数为

$$f(x) \;=\; \lambda e^{-\lambda x}, \quad x \geq 0.$$

对可测子集 $A \subset [0, \infty)$ 的测度 $P(A) = \int_A \lambda e^{-\lambda x}\,dx$。

特殊性质

- **无记忆性**：在连续分布中具备无记忆性的典型代表；
- **单参数分布**：只需一个正参数 λ 表征整体形状和尺度。

应用场景

- **到达过程**：泊松过程的两个相邻事件的发生间隔；
- **可靠性分析**：描述机件的无故障工作时间、电子元件的寿命等。

4.3.9 柯西分布（Cauchy Distribution）

定义及测度形式 柯西分布是定义在 $\mathbb{R}$ 上的连续测度，常见形态为

$$f(x) \;=\; \frac{1}{\pi}\,\frac{\gamma}{(x-x_0)^2+\gamma^2},$$

其中 $x_0 \in \mathbb{R}$ 为位置参数，$\gamma > 0$ 为尺度参数。

特殊性质

- **厚尾分布**：尾部衰减慢于正态分布，导致期望与方差都不定义（或无穷）；
- **自相似性**：分布形状在缩放和平移下保持相同形式。

应用场景

- **信号与谐振**：描述物理信号中的谐振峰形（洛伦兹线型）；
- **Robust 统计**：因其重尾特征，在抵抗极端值方面具有研究价值。

4.3.10 均匀分布（Uniform Distribution）

定义及测度形式 在区间 $[a,b]$ 上的均匀分布，测度可由常数密度给出：

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

故若 $A \subset [a,b]$，则 $P(A) = \int_A \frac{1}{b-a}\,dx$。

特殊性质

- **等可能性**：任何相同长度的子区间测度相等；
- **简单分布**：常作为“先验不确定性”或“缺省假设”时的基本模型。

应用场景

- **随机采样**：在区间内生成“无偏”的随机位置；
- **模拟**：对未知分布进行蒙特卡罗模拟时常用作初步试验。

4.3.11 t 分布（Student's t Distribution）

定义及测度形式 t 分布定义在 $\mathbb{R}$ 上，参数为自由度 $\nu > 0$，其密度函数为

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

特殊性质

- **对称分布**：关于 $x = 0$ 对称，形状类似正态但尾部更厚；
- **自由度控制**：ν 值越大，分布越接近正态分布，当 $\nu \to \infty$ 时极限逼近正态。

应用场景

- **小样本推断**：在样本量不大时（自由度小），t 分布常用于参数估计与区间估计；
- **假设检验**：配合样本方差估计，构建 t 检验。

4.3.12 F 分布（F Distribution）

定义及测度形式 F 分布定义在 $\mathbb{R}_{\geq 0}$ 上，常写作 $F(d_1, d_2)$，密度函数为

$$f(x) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}},$$

其中 d_1, d_2 称为自由度参数。

特殊性质

- **非对称**：定义域在非负实数上，具单峰结构且右偏；
- **比率分布**：可视为两个独立的卡方（或等效 Gamma）分布归一化后的比率；
- **与方差分析相关**：常用于比较两组方差或方差分析模型中的均方比。

应用场景

- **方差检验**：判断两总体方差是否相等；
- **多因素方差分析**：在回归模型或多元统计分析中用于显著性检验。

回顾整个概率空间的建构思路：先划定一个**样本空间** Ω，再以 **σ-代数** $\mathcal{F}$ 对其进行可测集合的系统划分，最后以**概率测度** P 来对这些集合赋以数值。无论是离散型、连续型还是混合型分布，其最根本的一致性在于满足测度的可数可加性、非负性和总和为 1 的性质，而它们之间最大的区别在于“赋值方式”以及对“点集”或“区间集”的度量差异。

在进一步的研究与应用中，**随机变量**、**期望**、**方差**等概念以及各种收敛定理，都是基于此处对概率空间和概率测度的严谨定义。理解这些概念的本质，需要时刻紧扣“**测度是定义在事件（可测集合）上的赋值函数**”这一根基，并留意离散或连续结构下的区别与联系。

5 条件概率测度

> **概要：** 条件概率测度能够刻画在给定部分信息（或观测）后的事件发生可能性，并在测度论框架下表现为对概率空间的**重新分配**或“坍缩”。当我们观测到某一信息（对应于一个子 σ-代数或事件）时，对与该信息不兼容的区域通常赋予**零**测度，从而 **focus on** 兼容信息的部分。

5.1 引言与背景

最直观的情形是：“如果已经知道某件事发生了，那么另一件事发生的概率是多少？”——这是经典的条件概率问题。但从测度论角度看，条件概率有不同层次的概念，分别应对：

1. 针对单一已知事件的简单情形（单事件条件概率）；
2. 面对“部分信息”或“子 σ-代数”时，需要的关于 σ-代数的条件概率；
3. 以及作为一系列“随点变化的测度核”而存在的正则条件概率（*Regular Conditional Probability, RCP*）。

后者可以理解为真正意义上的“测度重新分配”：一旦在样本空间 Ω 的某点（信息）被观测到，不与该信息兼容的其他部分自然**测度变为 0**；仅在与该信息兼容的“子空间”上保留概率质量。

5.2 单事件条件概率：最基础的比值

定义

给定事件对 $A, B \in \mathcal{F}$，并且假设 $P(B) > 0$，则

$$P(A \mid B) \;=\; \frac{P(A \cap B)}{P(B)}.$$

这是经典教科书中最常见的表达式，代表了“**当** B **已经发生时**，A 发生的概率是多少”。等价地，可视为在 B 外的区域概率被**压缩**为 0，所有“概率流” **focus on** B 上。

性质与解释

- **仅针对单个已知事件**：B 被固定且有正概率；
- **结果是一个常数**：$P(A \mid B) \in [0, 1]$。

5.3 关于 σ-代数的条件概率：在信息分割下的随机变量

为何要引入子 σ-代数？

在许多场景中，已知信息并不是“B 发生”这样单一事件，而是一种**更广泛的部分信息**。形式上，可以由一个**子 σ-代数** $\mathcal{G} \subseteq \mathcal{F}$ 表示。要在这种“分割”或“观测”下，描述任意事件 A 发生的概率时，我们就需要**关于 σ-代数的条件概率**。

定义

令 $\mathcal{G} \subseteq \mathcal{F}$ 为子 σ-代数。对于任何事件 $A \in \mathcal{F}$，若随机变量

$$P(A \mid \mathcal{G})(\omega)$$

满足：

- 它是 $\mathcal{G}$-**可测**的（即对每个 ω，该值只依赖于 ω 在 $\mathcal{G}$ 的分块）；
- 对所有 $E \in \mathcal{G}$，有

$$\int_E P(A \mid \mathcal{G})(\omega)\, dP(\omega) \;=\; P(A \cap E),$$

则称此 $P(A \mid \mathcal{G})(\omega)$ 为**关于 $\mathcal{G}$ 的条件概率**。

“测度的缩放”直观　不同的 $\omega \in E$ 可能让 $P(A \mid \mathcal{G})$ 取同样或不同的值，具体依赖 ω 在 $\mathcal{G}$ 里的分块信息。当一个样本点 ω 出现时，我们可以认为这个子 σ 代数里的事件 $\mathcal{G}$ 发生了，从而另我们的概率测度 focus on 这个事件

5.4　正则条件概率（RCP）：基于观测的全测度重新分配

动机

对固定 A 而言，$P(A \mid \mathcal{G})(\omega)$ 是一个 $\mathcal{G}$-可测函数；但若我们想对所有 $A \in \mathcal{F}$ 在给定信息 $\mathcal{G}$ 下如何被**全面**重估，就需要引入随 ω 变化的“**测度核**”：

$$\nu(\omega, A),$$

使得对每个 ω，$\nu(\omega, \cdot)$ 都是一个**新的概率测度**。

定义（非正式）

一个**正则条件概率**是映射

$$\nu : \Omega \times \mathcal{F} \;\longrightarrow\; [0,1],$$

满足：

1. 对固定的 ω，$\nu(\omega, \cdot)$ 是一个**概率测度**；
2. 对固定的 A，映射 $\omega \mapsto \nu(\omega, A)$ 在 $\mathcal{G}$ 下可测；
3. 一致性：$\displaystyle\int_E \nu(\omega, A)\, dP(\omega) \;=\; P(A \cap E), \quad \forall A \in \mathcal{F},\ E \in \mathcal{G}.$

“坍缩”/“focus on”理解 本质上，我们可以看作定义了一系列的概率测度，这些测度是我们考虑概率测度收缩在子 σ 代数的不同事件上而产生的，所以我们对于一个观测到的样本点 ω，可以找到对应的那个事件，进而从这一族测度中找到那个对应的概率测度，然后将概率测度 focus on 到这个测度上

- **一旦观测到** ω，概率对与“ω 携带的 $\mathcal{G}$ 信息”不兼容的区域**设为 0**；只在兼容部分继续分配；

- $\nu(\omega, \cdot)$ 就是“在 ω 所携带信息下，对所有事件的**新分布**”；

- 这也正是贝叶斯统计中“后验分布”的数学抽象：依据观测修正原分布，从而只在满足观测的部分赋予正概率。

5.5 在条件概率测度中讨论独立性

子 σ-代数之间的独立性

回顾之前给出的定义：若 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 是样本空间 Ω 上的两个子 σ-代数，则二者独立指**对任意** $A \in \mathcal{F}_1$ 和 $B \in \mathcal{F}_2$，都有

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

此时可将独立性理解为：“$\mathcal{F}_1$ 所携带的信息”和“$\mathcal{F}_2$ 所携带的信息”在概率上**彼此互不影响**。”

在条件测度下的解释

若我们在 $\mathcal{F}_2$ 的信息上做条件，则对于 $\mathcal{F}_1$ 中的事件 A，有

$$P\big(A \,\big|\, \mathcal{F}_2\big)(\omega) \;=\; P(A),$$

几乎处处成立（对于绝大多数 ω），就意味着**在得知 $\mathcal{F}_2$ 的信息**后，事件 $A \in \mathcal{F}_1$ 的概率值**不发生改变**。也可以说，“对 $\mathcal{F}_2$ 的条件化**不**影响 $\mathcal{F}_1$ 里的事件”。这正是**独立**的核心思想在条件概率测度层面的体现：一旦子 σ-代数独立，则其包含的信息对另一个子 σ-代数的事件概率无影响。

RCP 视角

在正则条件概率 $\nu(\omega, \cdot)$ 里，若 $\mathcal{F}_1$ 与 $\mathcal{F}_2$ 独立，并且我们根据 $\mathcal{F}_2$ 进行“坍缩”，则

$$\nu(\omega, A) \;=\; P(A), \quad \text{对几乎所有} \omega \in \Omega,\ \forall A \in \mathcal{F}_1.$$

表示在任何给定的观测 ω（只要它对应于 $\mathcal{F}_2$ 的信息），$\nu(\omega, \cdot)$ 对 $\mathcal{F}_1$ 的事件给出的概率和原测度 $P(\cdot)$ 相同。这完美刻画了“彼此独立导致观测不改变分布”的事实。

5.6 比较与小结

单事件 $P(A \mid B)$

- **对象：** 固定事件 B（$P(B) > 0$）。
- **结果形式：** 一个数值 $\in [0, 1]$。
- **直观：** Ω 中与 B 不兼容的区域测度设为 0，概率 **focus on** B 并进行归一化。

关于 σ-代数的 $P(A \mid \mathcal{G})$

- **对象：** 一个子 σ-代数 $\mathcal{G}$。
- **结果形式：** 一个 $\mathcal{G}$-可测随机变量；
- **直观：** 不再是简单地“事件 B 发生”，而是“只知道 ω 的某些属性”，在对应子空间对 A 的概率进行再评估。

正则条件概率 $\nu(\omega, \cdot)$

- **对象：** 一族随点 ω 变化的概率测度；
- **结果形式：** 对每个 ω 都给出一个完整的分布 $\nu(\omega, A)$；
- **直观：** “在 ω 对应的观测信息下，对所有事件的发生概率如何**全面**重估”，对不兼容信息的部分直接赋 0。

“focus on” 式理解：
一旦观测（或已知信息）选定了子空间，原概率测度会把该子空间外的所有事件的概率设为 0，并对子空间内的事件做**归一化**分配。这个过程可类比物理上的 “坍缩” 或 “投影”：我们只保留了与观测相容的轨道，而其余轨道的概率则被压缩到 0。

总体而言：条件概率测度从最简单的**单事件**情形（通过简单比值计算），一直延伸到对**子 σ-代数**的条件化，再到随观测点变化的 **RCP**；每一个层次都体现了 “得到新信息后，对不符该信息的区域测度**设为 0**，对剩余区域**重新归一化**” 的核心思路。而**独立性**在此框架下表明：若两子 σ-代数独立，则对其中一个做条件化，不会改变另一个子 σ-代数里事件的测度。

（三）随机变量

概率空间的概念已经被完备地定义了起来，可是还有一个问题：对于现实中遇到的一些问题，其样本空间往往比较复杂、抽象。如从黑箱里摸球，只考虑颜色，其样本空间是摸到红球，摸到白球。这样一来，其描述就会变得很困难。有没有好的方法能够用来描述这些抽象的样本空间呢?

随机变量正是这样一种工具，其通过把抽象的样本空间映射到一个简单的数值空间，并在此新空间中保留远来概率空间的 σ-代数和概率测度的性质，以此更方便地讨论原来的空间。

在此，我们会讨论有关随机变量的以下概念：

1. 随机变量的基本思想：样本空间的变换
2. 与概率空间中相关概念的关联
    - 分布，PMF（概率质量函数），PDF（概率密度函数），分布函数
    - 随机向量，联合分布，边缘分布与随机性
3. 随机变量的数字特征
4. 分析随机变量的方法
    - 特征函数与矩母函数
    - 次序统计量
    - 随机变量的函数与代数运算
    - 随机变量分布的模拟方法

1 随机变量的基本概念

- 随机变量是一种**可测映射**，将原有的“样本空间”中的事件与“数值（或更一般空间）”上的可测集合关联起来。
- 引入随机变量的核心原因，是为了便于对“原本可能复杂且抽象的样本空间”进行数值化分析，从而将概率问题转化为数值空间的问题。
- 随机变量会**诱导**出一个新的概率测度（推前测度），使我们可以只在数值空间上操作，简化研究和计算。

1.1 随机变量的定义

定义： 设 $(\Omega, \mathcal{F}, P)$ 为一个概率空间。若有一个映射

$$X : \Omega \longrightarrow S,$$

其中 $(S, \mathcal{S})$ 是另一个可测空间（即 S 是一个集合且 $\mathcal{S}$ 是其上的 σ-代数），并且对**所有** $B \in \mathcal{S}$ 都满足

$$X^{-1}(B) \in \mathcal{F},$$

则称 X 为一个**随机变量（Random Variable）**。换言之，X 保证了：从数值空间 $(S, \mathcal{S})$ 中**任意**一个可测集合往回看，其**原像** $X^{-1}(B)$ 始终是原概率空间 $(\Omega, \mathcal{F}, P)$ 中的可测事件。

需要强调的是：

1. X 作为一个函数，必须对 Ω 中**每一个**样本点都有定义；不存在“只对部分点定义，其他点未定义”的情况，否则就不是一个完整的函数，也就不能算随机变量。
2. 该映射只要求“对任意可测集 $B \in \mathcal{S}$，其原像可测”。它并不要求 X 是“满射（surjective）”或“双射（bijection）”。换言之，随机变量无须把 Ω 映到 S 的全部元素，也无需一个值只对应一个样本点。

3. 虽然不要求满射或双射,但每个点 $\omega \in \Omega$ 仍会对应到某一个 $X(\omega) \in S$。对于每个 $x \in S$，集合

$$X^{-1}(\{x\}) \;=\; \{\omega \in \Omega : X(\omega) = x\}$$

被称为 "x 的原像" 或者 "x 的纤维（fiber）"。将所有不同 x 的原像收集起来，它们会**两两不相交**并且并集是 Ω，从而在样本空间中形成一个**分割（partition）**。这些原像及其并集自然生成一个**子-σ-代数**，记作 $\sigma(X)$，它包含所有可以 "通过 X 的取值" 来刻画的事件。形式上:

$$\sigma(X) \;=\; \left\{X^{-1}(B) \mid B \in \mathcal{S}\right\} \;\subseteq\; \mathcal{F}.$$

该子-σ-代数包含所有由 X 的数值诱导出来的可测事件。

4. 对于随机变量中，那些原像不为原样本空间中的样本点的值，我们认为其原像为空集，对应的概率测度（分布）为 0

进一步说明:

- 对于有限或可数的 Ω，若 $\mathcal{F}$ 是所有子集所组成的离散 σ-代数，则任何函数 $X : \Omega \to S$ 都是可测的。
- 在更一般的情形下，X 的可测性需要与 $\mathcal{F}$ 和 $\mathcal{S}$ 相匹配，即对 $\mathcal{S}$ 中**任意**集合 B，$X^{-1}(B)$ 必须落在 $\mathcal{F}$ 里。
- "随机变量" 是为我们在数值空间上做分析（例如计算期望、方差等）提供了基础；然而，它并不必是 "单射" 或 "满射"，只要它符合上面的可测性即可。

实值随机变量

在最常用的情况下，我们取

$$S = \mathbb{R}, \quad \mathcal{S} = \mathcal{B}(\mathbb{R}),$$

其中 $\mathcal{B}(\mathbb{R})$ 表示实数上的 Borel σ-代数。此时 $X(\omega)$ 常被简化记为 "ω 的数值"，并称 X 为**实值随机变量**。

1.2 可测性与 σ-代数的拉回

可测性的含义

对任意可测集 $B \in \mathcal{S}$，其原像

$$X^{-1}(B) \;=\; \{\omega \in \Omega : X(\omega) \in B\}$$

必须属于 $\mathcal{F}$。换言之，“X 的取值落在 B 内”这个事件，在原空间中是可测事件。这正是随机变量最基本的要求，即**可测性 (Measurability)**。

> **直观理解：**你可以把“S 上的可测集合 B”想象成“数值空间中的某些区间、子集等”，一旦指定了“$X \in B$”这件事，我们需要在样本空间 Ω 里找到一个对应的“$\omega \in \Omega$ 满足 $X(\omega) \in B$”的集合，并要求它**绝对**能被测度 P 正确处理。

1.3 推前测度（Push-forward measure）

定义：给定随机变量 X 以及原概率测度 P，在 $(S, \mathcal{S})$ 上定义新测度 P_X：

$$P_X(B) \;=\; P\big(X^{-1}(B)\big), \quad \forall\, B \in \mathcal{S}.$$

这一定义将原空间中发生概率 $P(A)$ 的“事件” A 转换为数值空间中集合 B 的“发生概率” $P_X(B)$。因此，P_X 称为**推前测度**，也被称作随机变量 X 的**分布**。

含义与优势

- **简化分析：**只需在较直观的数值空间 $(S, \mathcal{S})$ 上操作，而不必处理原本复杂的 Ω 和 $\mathcal{F}$。
- **信息不丢失：**通过“原像” $X^{-1}(B)$ 与 B 的映射关系，保留了关于事件可测性以及概率的全部信息。

1.4 $(S, \mathcal{S}, P_X)$：新的概率空间

分布空间

由上述推前测度 P_X 可得：

$$(S, \mathcal{S}, P_X)$$

本身就构成一个概率空间。我们常称它为“X 的**分布空间**”或“X 的**像空间**”。在此空间中，元素是“随机变量取到的值”，而不是“原始 ω”。

- 当 X 是离散型实值随机变量时，P_X 在某些点集（有限或可数集）上赋予正概率。
- 当 X 是连续型实值随机变量时，P_X 相对于 Lebesgue 测度可具有一个概率密度函数 $f_X(x)$。

意义

- **摆脱原本样本空间的束缚：**在很多应用中，Ω 可能是某些极其复杂乃至抽象的集合，但只要借助随机变量 X 将它“投影”或“映射”到一个更易处理的数值空间，就能便捷地进行概率分析、统计推断或数值计算。
- **保留关键概率结构：**由于 $P_X(B) = P(X \in B)$，“概率”本身没有丢失或伪造，依旧遵守测度性质（非负性、可数可加性等），从而确保一切基于分布的理论（如期望、方差、特征函数等）都能在 $(S, \mathcal{S}, P_X)$ 上成立。

一个简单例子

掷骰子场景：

- 原空间 $\Omega = \{1, 2, 3, 4, 5, 6\}$，$\mathcal{F} = 2^{\Omega}$，$P$ 为等概率分配（即 $P(\{\omega\}) = \frac{1}{6}$）。
- 定义 $X(\omega) = \omega^2$。这时，$S = \{1, 4, 9, 16, 25, 36\}$，且 $\mathcal{S}$ 可取所有子集。

- 对于子集 $B = \{1, 4\} \subseteq S, P_X(B) = P(X \in \{1, 4\}) = P(\omega \in \{1, 2\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$。

由此可见，通过映射 X，原有的样本空间变换到了数的集合上，并在数值域里形成了新测度 P_X。

小结

通过引入随机变量，我们建立了一个从**原始概率空间** $(\Omega, \mathcal{F}, P)$ 到**数值空间** $(S, \mathcal{S}, P_X)$ 的映射关系。这个映射既保留了 σ-代数的可测性结构，又将概率测度“推前”到新的空间里。结果是，我们往往能够在**更直观、易操作**的数值体系上分析随机现象，极大简化了对事件概率、分布特征以及后续统计量的研究与计算。

2 随机变量的分布与分布函数

概要：

- 随机变量所诱导的“推前测度” P_X 是对数值域（如 $\mathbb{R}$）中的可测集合赋值。
- 但在实际使用中，直接去处理“任意可测集合”可能过于复杂；相反，我们经常定义**概率质量函数**（PMF）、**概率密度函数**（PDF）或**累积分布函数**（CDF）来更简便地刻画分布的整体信息。
- 其中，CDF 是核心工具：它既能统一离散与连续情形，又能唯一确定推前测度 P_X。

需要特别注意的是，分布 (distribution) 这个概念本身与“推前测度”是等同的！分布就是分布空间的测度本身。

而 PMF、PDF 和 CDF 是仅仅针对于随机变量而言才成立的概念，因为他们是“函数”！其定义域是数集（而不是更为抽象的集合）；其与分布的关系是在随机变量（且为值域为实数集）的语境之下，可以用这些函数来充分地描述特定种类的分布。

“分布函数” (Distribution function) 这个概念等同于累积分布函数 (Cumulative distribution function)。

2.1 离散情形：概率质量函数 (PMF)

定义与性质

当随机变量 X 的取值集合至多可数（有限或可数无穷）时，分布空间 $(S, \mathcal{S}, P_X)$ 上的正概率仅落在若干点集上。此时可定义**概率质量函数**（PMF）：

$$p_X(x) \;=\; P_X(\{x\}) \;=\; P\big(X = x\big).$$

此外，由于**全部概率**之和为 1，对于离散情形有

$$\sum_{x \in S} p_X(x) \;=\; 1.$$

直观理解

- **记录每个数值点的概率：** 对任意 $x \in S$，$p_X(x)$ 告诉我们 “X 恰好等于 x” 的可能性大小。
- **简洁性：** 一旦列出（或写出一个公式给出）全部 $p_X(x)$，就能完整表达推前测度 P_X 在每个单点集 $\{x\}$ 的取值，从而在离散情形下**唯**一确定分布。

注意

即使我们有了 PMF，仍可以通过 $p_X(x)$ 的累加来计算形如 $P_X(\{x : x \in A\})$ 或 $P_X((a, b])$ 等事件的概率，避免直接在原 σ-代数中操作。

2.2 连续情形：概率密度函数 (PDF)

绝对连续与 PDF

若推前测度 P_X 相对于 Lebesgue 测度绝对连续，则存在非负可积函数 $f_X(x)$，使得

$$P_X\big((a, b]\big) \;=\; \int_a^b f_X(x)\, dx, \quad \forall\, a < b.$$

我们称 $f_X(x)$ 为**概率密度函数**（PDF）。它满足

$$\int_{-\infty}^{+\infty} f_X(x)\, dx \;=\; 1,$$

并对所有 x 有 $f_X(x) \geq 0$。

可视化理解

- **连续分布**意味着随机变量不在任何单点处累积正概率，且“概率集中”可用曲线（密度函数）来刻画。
- 当我们说“X 属于某区间 (a, b) 的概率”时，只需积分 $f_X(x)$ 即可。

与推前测度的关系

虽然 P_X 是定义在所有 Borel 集上，但在**绝对连续**的场景下，任何 Borel 集的概率都可以用 f_X 对该集合的积分来表示（配合 Lebesgue 测度）。这样，就不需要在原像中逐个分析 Borel 集。

2.3 累积分布函数 (CDF)

定义

累积分布函数（CDF）是一个从 $\mathbb{R}$ 映射到 $[0, 1]$ 的函数 F_X，定义为：

$$F_X(x) \;=\; P\big(X \leq x\big) \;=\; P_X\big((-\infty, x]\big).$$

无论 X 是离散型、连续型还是混合型，都能用 F_X 一致地定义其分布。

基本性质

1. **单调不减：** 若 $a \leq b$，则 $F_X(a) \leq F_X(b)$。
2. **右连续：** $\lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = F_X(x)$。
3. **边界：** $F_X(-\infty) = 0$ 与 $F_X(+\infty) = 1$。

此外，在任何实值随机变量场合，F_X 都能够**唯**一地决定推前测度 P_X：一旦给出 F_X，就能复原 $P_X((-\infty, x]) = F_X(x)$ 以及其他 Borel 集的概率值（借助分段构造、右连续性等）。

为什么使用 CDF 而不是直接用 P_X？

- **P_X 的定义域是 "Borel 集"**：要想"直接使用" P_X，意味着要面对整套 Borel σ-代数的所有集合，操作起来并不直观。
- **CDF 提供一个简单实函数描述：**只需告诉人们每个点 x 处的 $F_X(x) = P_X((-\infty, x])$，就能得到分布的整体信息。这种从 $\mathbb{R}$ 到 $[0,1]$ 的函数在可视化、计算和理论研究中都非常便利。
- **统一描述离散、连续和混合分布：**无论密度函数是否存在，或 PMF 是否有限多个点，CDF 都可以平滑地囊括不同类型的分布，且理论性质非常完备（单调、右连续等）。

一个例子

分段定义 CDF：

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.2, & 0 \le x < 1, \\ 0.6, & 1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

这是一个**混合型**或**分段常数型** CDF 示例：表明在区间 $[0,1)$ 段随机变量概率增加 0.2，在区间 $[1,2)$ 再增加 0.4（总和到 0.6），最后在 $[2,\infty)$ 拉升至 1。从中我们可读取各区间所对应的概率，且满足单调、右连续等性质。

特别提示：分布函数与推前测度的区别

- **推前测度** P_X：是定义在 $(S, \mathcal{S})$ 上的测度，严格说它给每一个 Borel 集合 B 一个概率值 $P_X(B)$。
- **分布函数** F_X：是定义在 $\mathbb{R}$ 上的一个函数（由 $x \mapsto P_X((-\infty, x])$ 给出），并不直接把"所有 $B \in \mathcal{B}(\mathbb{R})$" 映射到 $[0,1]$，而仅向每个**实数点** x 输出 P_X 在 $(-\infty, x]$ 上的测度值。

- **原因**：我们通常只需要掌握“到某个数值 x 的累计概率”，就能高效地描述/重建分布。相对来说，去操控“任意 Borel 集”太笼统，不如用 CDF 搭配一些基础理论就能获得所有所需概率信息。

小结

- **PMF、PDF 与 CDF 三兄弟**：根据随机变量的类型（离散 / 连续 / 混合），我们可选择合适方式来刻画 P_X。
- **CDF 的核心地位**：它可以同时处理不同类型分布，并且由 F_X 可以**唯**一地推回完整的推前测度 P_X。
- **简化与可操作性**：相比直接使用 P_X（在 Borel σ-代数上），这些函数式描述在理论与实践中都极具简洁与可操作性。例如，计算概率、查表、定量比较不同分布等，CDF 都是最常见工具。

3 随机向量

概要： 当我们要同时刻画多个随机变量时，往往会将它们组合成一个**随机向量**（random vector）。这意味着对样本空间中的每个点，都可以映射出一个多维数值，从而构成向量形式的取值。随机向量的引入让我们能够分析不同分量之间的联合分布、边缘分布，以及它们是否具有相互独立的性质等。

3.1 多元随机变量（随机向量）

定义

设 $(\Omega, \mathcal{F}, P)$ 是一个概率空间。若有 n 个实值随机变量 $\{X_1, X_2, \ldots, X_n\}$ 定义在同一 Ω 上，则我们可以构造一个映射

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) : \Omega \longrightarrow \mathbb{R}^n,$$

并称 $\mathbf{X}$ 为**随机向量**。它的**分量**分别是 $X_1, X_2, \ldots, X_n$。

特别要注意的是，这一个随机变量是将同一个样本空间映射到了一个高维的向量上，而不是对于每个随机变量独立地选取不同的样本点；所以其原像是一个积事件而非同样是一个向量

可测性

若对任意 Borel 集 $B \subseteq \mathbb{R}^n$，都有

$$\mathbf{X}^{-1}(B) \;=\; \{\omega \in \Omega : (X_1(\omega), \ldots, X_n(\omega)) \in B\} \;\in\; \mathcal{F},$$

则称 $\mathbf{X}$ 可测。由于 X_i 各自都是**实值随机变量**（可测映射），它们的笛卡尔组合 $\mathbf{X}$ 在自然定义的可测结构（$\mathbb{R}^n$ 上的 Borel σ-代数）下仍保持可测性。

> **直观理解：** 将 Ω 上所有可能状态 ω 一次性“映射”到一个 n 维向量 $(x_1, x_2, \ldots, x_n)$ 中，这就使得我们能够在 $\mathbb{R}^n$ 的坐标系里分析“同时观察到 X_1 在某一区间、X_2 在另一区间、…”的事件。

3.2 联合分布与边缘分布

3.2.1 联合测度

随机向量 $\mathbf{X}$ 诱导了一个**联合测度** $P_{\mathbf{X}}$（也可记为 $P_{X_1,\ldots,X_n}$）在 $\mathbb{R}^n$ 的 Borel σ-代数上，定义为

$$P_{\mathbf{X}}(B) \;=\; P\big(\mathbf{X}^{-1}(B)\big), \quad \forall\, B \in \mathcal{B}(\mathbb{R}^n).$$

这可以看作是每个分量 X_i 联合在一起后的“整体分布”，也被称为**联合分布 (Joint Distribution)**。与一维情形类似，我们可以通过“$\mathbb{R}^n$ 上的可测集合及其原像”来定义、表征并计算事件在联合维度上的概率。

联合分布的可视化理解

- 当 $n = 2$ 时，我们有 $\mathbf{X} = (X, Y)$。它在 $\mathbb{R}^2$ 上产生一个概率测度。对于任意“矩形”或更复杂的 Borel 集，都可以找到相应的概率值，如 $P\big(X \in [a, b], Y \in [c, d]\big)$。

- 若满足“绝对连续”条件，则可以定义**联合 PDF（概率密度函数）** $f_{X,Y}(x,y)$，并通过 $P((X,Y)\in A)=\iint_A f_{X,Y}(x,y)\,dx\,dy$ 来计算概率。
- 在有限可数取值的情况下，则可定义**联合 PMF（概率质量函数）** $p_{X,Y}(x,y)=P(X=x,Y=y)$。

3.2.2 边缘分布

若我们只关心随机向量的某一个分量，譬如在 $\mathbf{X}=(X_1,X_2)$ 中只关心 X_1，则可以将与其它坐标（例如 X_2）相关的所有情形**合并**起来：

$$P_{X_1}(B_1) \;=\; \sum_{x_2\,\in\,\mathcal{X}_2} P_{X_1,X_2}(B_1\times\{x_2\}) \quad \text{(离散情形)},$$

或

$$P_{X_1}(B_1) \;=\; \int_{x_2\,\in\,\mathcal{X}_2} P_{X_1,X_2}\big(B_1\times dx_2\big) \quad \text{(连续情形)}.$$

这就称为**边缘分布**（Marginal Distribution）。对高维情形亦然，可以通过在除目标分量之外的坐标上**积分**或**求和**，来获得任意一个子向量的分布。

> **投影视角：**将高维联合分布做“投影”到其中某个（或几个）坐标轴上，而忽略（积分/求和）其他维度，如把 $\mathbb{R}^n$ 的分布投影到第 i 维坐标上。

3.3 随机向量与联合 σ-代数

联合 σ-代数的概念

从测度论角度来看，若我们将 σ-代数 $\mathcal{F}_i=\sigma(X_i)$ 表示“随机变量 X_i 所产生的 σ-代数”，那么**随机向量 $(X_1,\dots,X_n)$ 的 σ-代数**就是由这 n 个 σ-代数**联合生成**：

$$\sigma(X_1,\dots,X_n) \;=\; \sigma\big(\mathcal{F}_1,\mathcal{F}_2,\dots,\mathcal{F}_n\big).$$

直观地说，“$\mathbf{X}$ 取值在某可测集 $B\subseteq\mathbb{R}^n$”可拆分/组合成“各分量 X_i 落在哪些子集”这类事件的并、交、补关系。于是，对于随机向量的所有**联合事件**，其所在的 σ-代数就是由各分量生成的最细 σ-代数。

为何要引入联合 σ-代数？

- **完整刻画多维信息：**要同时描述 $(X_1, \ldots, X_n)$ 的全部可能事件，必须考虑各分量**同时**满足某些条件的情形，而这自然是各单变量 σ-代数的最小闭包。
- **测度分配的一致性：**当定义联合测度 $P_{X_1,\ldots,X_n}$，我们要保证任何与 $(X_1, \ldots, X_n)$ 相关的“组合事件”都能在此 σ-代数中找到对应原像并赋予概率。

3.4 随机向量的独立性

概念

当随机向量 $\mathbf{X} = (X_1, \ldots, X_n)$ 的各分量本身要保持“无关性”时，通常说“$X_1, \ldots, X_n$ 相互独立”。这一独立性可等价表述为：由 X_i 生成的 σ-代数 $\mathcal{F}_i$ 两两（乃至更高阶）独立。具体而言，

$$P\big(X_1 \in B_1, \ldots, X_n \in B_n\big) \;=\; P(X_1 \in B_1) \times \ldots \times P(X_n \in B_n),$$

对所有 $B_1, \ldots, B_n$ 皆成立时，称之为**相互独立**。

简化作用

如果 $(X_1, \ldots, X_n)$ 相互独立，则很多运算与概率计算可因式分解。例如，联合 PDF（若存在）可以写成

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \;=\; \prod_{i=1}^{n} f_{X_i}(x_i).$$

这极大地简化了对多元分布的研究与计算，也是实际建模中常见的“独立性假设”的根源。

小结

要点回顾：

- **随机向量**是把多个随机变量组合在一起的可测映射；它为我们提供在 $\mathbb{R}^n$ 上的联合分布。
- **联合分布**描述了各维度之间的联合概率结构，通过“投影（积分/求和）”可以得到**边缘分布**。
- **联合 σ-代数**体现了多维事件间的综合可测性，并在定义**联合测度**时起到不可或缺的作用。
- 若各维度**独立**，则在联合测度上可以进行因式分解，从而极大地简化了高维运算与分析。

简而言之，随机向量是将多个随机变量**集成**到一个**高维映射**中进行统一研究的核心工具，它所带来的联合分布、边缘分布以及独立性概念，为我们深入理解多元随机现象提供了坚实的理论框架。

4 随机变量的若干数量特征

在研究随机变量时，我们往往关心其所诱导的分布**在数值上的刻画**。这些数值刻画可以帮助我们理解分布的中心位置、离散程度、变量间关系以及更高阶的规律。本节将介绍数学期望、方差、协方差及相关概念的**形式化定义**及其意义。

4.1 数学期望（Mean）

定义及性质

定义： 设 X 是一个实值随机变量，P_X 为 X 在 $\mathbb{R}$ 上诱导的分布测度。若下式中的积分绝对可积，则

$$E[X] \;=\; \int_{\Omega} X(\omega)\,\mathrm{d}P(\omega) \;=\; \int_{-\infty}^{+\infty} x\,\mathrm{d}P_X(x).$$

在这里，$\int_{\Omega} X(\omega)\,\mathrm{d}P(\omega)$ 代表了**将随机变量 $X(\omega)$ 作为可测函数**在原空间上做的积分；而 $\int_{-\infty}^{+\infty} x\,\mathrm{d}P_X(x)$ 则利用了“推前测度” P_X 的形式在数值域 $\mathbb{R}$

上求积分。两者等价，体现了**随机变量可测性与测度转换**的完整性。

> **直观理解：** 数学期望常被视为“中心位置”或“平均水平”。当 X 具备可积性时，$E[X]$ 反映出在大量重复实验（或采样）下，“长期平均值”会稳定趋近于 $E[X]$。

4.2 方差（Variance）

定义与解释

对于一个可积随机变量 X，若 $E[(X - E[X])^2]$ 存在（有限），则定义

$$\mathrm{Var}(X) \;=\; E\big[(X - E[X])^2\big].$$

它表示随机变量围绕其均值 $E[X]$ 波动程度的平均度量。在实际中，方差越大表示 X 越“不稳定”或“分散度”越高；方差越小表示 X 越“集中”。

4.3 协方差（Covariance）及相关系数

协方差

当我们同时关心两个随机变量 X 与 Y 之间如何共同波动时，可以考虑“协方差”：

$$\mathrm{Cov}(X,Y) \;=\; E\big[(X - E[X])\,(Y - E[Y])\big].$$

若 $\mathrm{Cov}(X,Y) > 0$，大致意味着 X、Y 往往**同步增减**；若 < 0，则表明二者往往**一涨一跌**；$= 0$ 通常意味着**无线性相关**（但不一定表示绝对不相关）。

相关系数

$$\rho_{X,Y} \;=\; \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

称为**相关系数**（Correlation Coefficient）。它是一个归一化量，介于 $[-1,1]$ 之间，数值越靠近 ± 1，表示线性相关性越强。

4.4 更高阶矩与协方差矩阵

高阶矩

对于单个随机变量 X，可以定义 k 阶矩：

$$E[X^k], \quad k = 1, 2, 3, \ldots$$

它们可以进一步刻画分布在更高维度上的形状特征（例如偏度、峰度等）。

协方差矩阵

在多维随机向量 $\mathbf{X} = (X_1, \ldots, X_n)$ 中，若我们关心各分量间两两的协方差，可以构造一个 $n \times n$ 的矩阵：

$$\mathrm{Cov}(\mathbf{X}) \;=\; \big[\mathrm{Cov}(X_i, X_j)\big]_{1 \le i,j \le n},$$

称为**协方差矩阵**。该矩阵在多元统计、机器学习中扮演关键角色，如刻画数据云分布的“形状”和“扩散范围”等。

4.5 典型分布的数量特征简介

在应用中，常会遇到多种分布类型：如伯努利分布、二项分布、超几何分布、泊松分布（离散情形）；或正态分布、指数分布、柯西分布、均匀分布、χ^2 分布、t 分布、F 分布（连续情形）等。这些分布常有以下典型数量特征：

分布的均值与方差总结

以下汇总了常见分布的**均值 (Mean)** 与**方差 (Variance)**，并在注释中标明所需参数及适用条件。需要注意的是，有些分布的均值或方差可能并不存在（例如柯西分布的情形）。

- **伯努利分布**：$X \sim \mathrm{Bernoulli}(p)$，$X \in \{0, 1\}$，

$$E[X] \;=\; p, \quad \mathrm{Var}(X) \;=\; p(1-p).$$

- **二项分布**：$X \sim \mathrm{Binomial}(n, p)$，

$$E[X] \;=\; np, \quad \mathrm{Var}(X) \;=\; np(1-p).$$

- **超几何分布**：$X \sim \mathrm{Hypergeometric}(N, K, n)$
  - N：总体大小
  - K：总体中“成功”元素的个数
  - n：抽取次数（不放回）

  则
$$E[X] \;=\; n \cdot \frac{K}{N}, \quad \mathrm{Var}(X) \;=\; n \cdot \frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}.$$

- **泊松分布**：$X \sim \mathrm{Poisson}(\lambda)$，

$$E[X] \;=\; \lambda, \quad \mathrm{Var}(X) \;=\; \lambda.$$

- **离散均匀分布**（“集合分布”的一种情形）：若 X 在 $\{1, 2, \ldots, m\}$ 上取值且各点概率均等，

$$P(X = k) = \frac{1}{m}, \;\; k = 1, 2, \ldots, m.$$

  则
$$E[X] \;=\; \frac{m+1}{2}, \quad \mathrm{Var}(X) \;=\; \frac{m^2-1}{12}.$$

- **Gamma 分布**：$X \sim \mathrm{Gamma}(\alpha, \beta)$
  - 这里一般取 $\alpha > 0$ 为形状 (shape)，$\beta > 0$ 为率 (rate) 参数
  - (有些文献也用 $\theta = 1/\beta$ 作为尺度 (scale))

  则
$$E[X] \;=\; \frac{\alpha}{\beta}, \quad \mathrm{Var}(X) \;=\; \frac{\alpha}{\beta^2}.$$

- **卡方分布**：$\chi^2(k)$
  - 这是 Gamma 分布的特例：$\chi^2(k) \equiv \mathrm{Gamma}\left(\frac{k}{2}, \frac{1}{2}\right)$

$$E[X] \;=\; k, \quad \mathrm{Var}(X) \;=\; 2\,k.$$

- **Beta 分布**：$X \sim \mathrm{Beta}(\alpha, \beta)$，

$$E[X] = \frac{\alpha}{\alpha+\beta}, \quad \mathrm{Var}(X) = \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}.$$

- **正态分布**：$X \sim \mathcal{N}(\mu, \sigma^2)$，

$$E[X] = \mu, \quad \mathrm{Var}(X) = \sigma^2.$$

- **指数分布**：$X \sim \mathrm{Exp}(\lambda)$，

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

- **柯西分布**：$X \sim \mathrm{Cauchy}(x_0, \gamma)$

  - x_0 为位置参数，$\gamma > 0$ 为尺度参数

  该分布**均值与方差都不存在**（即都不定义，或可认为不收敛）。

- **连续均匀分布**：$X \sim \mathrm{Uniform}(a, b)$

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

- **t 分布**：$X \sim t(\nu)$

  - ν 是自由度 (degrees of freedom)
  - 当 $\nu > 1$ 时，$E[X] = 0$
  - 当 $\nu > 2$ 时，$\mathrm{Var}(X) = \frac{\nu}{\nu-2}$
  - 否则均值或方差不存在/不定义

- **F 分布**：$X \sim F(d_1, d_2)$

  - d_1, d_2 为自由度参数
  - 若 $d_2 > 2$，则 $E[X] = \frac{d_2}{d_2-2}$
  - 若 $d_2 > 4$，则 $\mathrm{Var}(X) = \frac{2\,d_2^2\,(d_1+d_2-2)}{d_1\,(d_2-2)^2\,(d_2-4)}$

说明：

- 柯西分布 (Cauchy) 的均值和方差均不存在（分布太“重尾”）。
- t 分布和 F 分布在自由度较小时，均值或方差同样不存在。
- 以上仅列出常见分布的主要形式，更多分布之均值方差可参阅统计分布手册或专业文献。

小结

回顾：

- **数学期望**与**方差**是最核心的两个数量特征，用于刻画随机变量的“中心”和“离散度”。
- **协方差**与**相关系数**拓展到两个随机变量之间的线性依赖程度，多维场景中构成**协方差矩阵**。
- 较高阶的**矩**与其衍生量（如偏度、峰度）能更深入地描述分布特性；多元情形下的协方差矩阵在统计与机器学习里扮演重要角色。
- 针对典型分布，我们往往能直接写出这些数量特征的**封闭形式**或**已知公式**，在应用时可快速使用。

5 特征函数与矩母函数

在研究随机变量的分布特性时，特征函数（Characteristic Function）和矩母函数（Moment Generating Function, MGF）是两种功能强大的工具。它们能够在分布分析、卷积运算以及极限定理等方面提供简洁、有效的手段。下文先分别介绍特征函数与矩母函数的定义、性质与常见应用，再简要对比两者的联系与区别。

5.1 特征函数（Characteristic Function）

5.1.1 定义

对于一个实值随机变量 X，其**特征函数** $\varphi_X(t)$ 定义为：

$$\varphi_X(t) \;=\; E\big[e^{\,itX}\big] \;=\; \int_{-\infty}^{+\infty} e^{\,itx}\,\mathrm{d}P_X(x),$$

其中 i 表示虚数单位，$t \in \mathbb{R}$。在此定义下：

- 特征函数将分布的概率信息“映射”到复平面中，可视为分布的傅里叶变换形式。
- 无论 X 是否具有有限矩（即可积高阶矩是否存在），特征函数都必定存在且为**酉可测（bounded, continuous）**函数。

5.1.2 核心性质

- **唯一性**：特征函数在很大程度上能够**唯一确定**随机变量的分布。若有两个随机变量的特征函数一致，则它们的分布相同。
- **卷积化简（独立性）**：若 X、Y 独立，则其和 $X+Y$ 的特征函数满足

  $$\varphi_{X+Y}(t) \;=\; \varphi_X(t)\,\varphi_Y(t),$$

  将原本在分布层面为“卷积”的运算化简成“乘积”运算。
- **连续性与可微性**：$\varphi_X(t)$ 在 $t=0$ 处可导，且 $\varphi_X(0)=1$。在一定条件下可进一步展开为幂级数，从而获得随机变量的矩信息（若矩存在）。

5.1.3 常见分布的特征函数示例

- **正态分布** $X \sim \mathcal{N}(\mu,\sigma^2)$：

  $$\varphi_X(t) \;=\; \exp\!\big(i\,\mu\,t - \tfrac{1}{2}\sigma^2 t^2\big).$$

- **泊松分布** $X \sim \mathrm{Poisson}(\lambda)$：

  $$\varphi_X(t) \;=\; \exp\!\big(\,\lambda\big(e^{\,it}-1\big)\big).$$

- **指数分布** $X \sim \mathrm{Exp}(\lambda)$：

$$\varphi_X(t) \;=\; \frac{\lambda}{\lambda - it}, \quad (\text{Impose } \mathrm{Re}(\lambda - it) > 0 \text{ 以保证存在性}).$$

5.2 特征函数的独立性、连续性定理及应用

特征函数在研究随机变量之和、极限定理等问题时具有不可或缺的地位。本节在介绍其**乘积性质**的同时，还将补充**连续性定理**等关键内容，并基于此给出中心极限定理（CLT）的一个典型证明思路。

基本性质与连续性 在介绍独立性与应用之前，先回顾特征函数的一些**通用性质**（以下“X”均指实值随机变量）：

1. $\varphi_X(0) = 1$. 由定义可得，$\varphi_X(0) = E[e^{i\cdot 0\cdot X}] = E[1] = 1$.
2. $\varphi_X(-t) = \overline{\varphi_X(t)}$. 特征函数对于**实值**随机变量是“共轭对称”的；特别是 $|\varphi_X(t)| = |\varphi_X(-t)|$.
3. $\big|\varphi_X(t)\big| \le 1$. 这是因为 $\big|e^{i\,t\,X}\big| = 1$ 对任意实 X 成立，因此 $|\varphi_X(t)| = |E[e^{i\,t\,X}]| \le E[|e^{i\,t\,X}|] = 1$.
4. φ_X 在 $(-\infty, \infty)$ 上是一致连续的：对任何 $t, h \in \mathbb{R}$，

$$\big|\varphi_X(t+h) - \varphi_X(t)\big| \;\le\; E\big|e^{i(t+h)X} - e^{itX}\big| \;\le\; E\big(\dots\big) \;<\; \dots$$

   具体估计可得出该函数是一致连续的。

5. 若 $aX + b$ 表示随机变量的仿射变换，则

$$\varphi_{aX+b}(t) \;=\; E\big[e^{i\,t\,(aX+b)}\big] \;=\; e^{i\,t\,b}\,\varphi_X(a\,t).$$

   这对于研究位置/尺度变换时十分便利。

连续性定理（Lévy Continuity Theorem） 该定理是用特征函数分析分布收敛的**基石**，其核心内容是：

$$\text{若}\varphi_{X_n}(t) \xrightarrow{n\to\infty} \varphi(t) \quad (\text{逐点收敛于某函数}\varphi),$$

且 φ 在 0 处**连续**并且 $\varphi(0)=1$，那么 φ 一定是**某个概率分布**的特征函数；并且若这一极限分布的特征函数是 φ，则说明随机变量序列 X_n **按分布收敛**到与该特征函数对应的分布。

> **要点：** 在分布收敛（或称弱收敛）的讨论中，检验点态收敛 + 在 0 处连续 + 正确的归一化便足以判断极限函数 φ 是合法的特征函数。这也是中心极限定理等极限证明中所使用的关键一环。

乘积性质：独立性与卷积简化 特征函数最著名的性质之一：若随机变量 X、Y **独立**，则它们和的特征函数是各自特征函数的**乘积**：

$$\varphi_{X+Y}(t) \;=\; \varphi_X(t)\,\varphi_Y(t).$$

原因为：

$$\varphi_{X+Y}(t) \;=\; E\big[e^{\,i\,t\,(X+Y)}\big] \;=\; E\big[e^{\,i\,t\,X}\,e^{\,i\,t\,Y}\big] \;=\; E\big[e^{\,i\,t\,X}\big]\,E\big[e^{\,i\,t\,Y}\big] \;=\; \varphi_X(t)\,\varphi_Y(t),$$

其中独立性保证“期望可拆分”。这意味着**卷积**（独立变量之和的分布）的研究可转化为**乘积**的分析，大大简化了分布运算与极限推导。

应用：中心极限定理（CLT）证明思路（特征函数法） 中心极限定理表明：若 $\{X_k\}$ 是一组独立同分布（i.i.d.）随机变量，且 $E[X_k]=\mu$、$\mathrm{Var}(X_k)=\sigma^2>0$，令

$$S_n \;=\; X_1+X_2+\cdots+X_n,$$

则在 n 足够大后，适当标准化的随机变量

$$Z_n \;=\; \frac{S_n-n\,\mu}{\sigma\,\sqrt{n}}$$

的分布趋近于标准正态 $\mathcal{N}(0,1)$。

特征函数法的详细步骤如下：

1. **中心化与标准化**令 $Y_i=\frac{X_i-\mu}{\sigma}$。则 $E[Y_i]=0$，$\mathrm{Var}(Y_i)=1$，且 $Z_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Y_i$。

2. **特征函数展开**由独立性得

$$\varphi_{Z_n}(t) \;=\; E\Big[e^{\,i\,t\,(\frac{1}{\sqrt{n}}\sum Y_i)}\Big] \;=\; \prod_{i=1}^{n}\varphi_{Y_i}\Big(\tfrac{t}{\sqrt{n}}\Big).$$

因为 Y_i 同分布，令 $\varphi_{Y_i}=\varphi_{Y_1}$，则

$$\varphi_{Z_n}(t) \;=\; \Big[\varphi_{Y_1}\big(\tfrac{t}{\sqrt{n}}\big)\Big]^n.$$

3. **泰勒（或幂级数）展开**当 $t/\sqrt{n}$ 足够小时，可对 $\varphi_{Y_1}\big(t/\sqrt{n}\big)$ 做展开。注意 Y_1 有 $E[Y_1]=0,\mathrm{Var}(Y_1)=1$，故附近展开给出：

$$\varphi_{Y_1}\Big(\tfrac{t}{\sqrt{n}}\Big) \;=\; 1 \;-\; \tfrac{t^2}{2\,n}\,E\big[Y_1^2\big] \;+\; o\Big(\tfrac{t^2}{n}\Big).$$

进一步合并常量，约得：

$$\varphi_{Y_1}\Big(\tfrac{t}{\sqrt{n}}\Big) \;=\; 1 \;-\; \tfrac{t^2}{2\,n} \;+\; o\Big(\tfrac{1}{n}\Big).$$

将其**乘方**至 n 次，可用 $(1+x/n)^n\approx e^x$ 的极限技巧，得到

$$\varphi_{Z_n}(t) \;=\; \Big[1-\tfrac{t^2}{2\,n}+o(\tfrac{1}{n})\Big]^n \;\longrightarrow\; e^{-\frac{t^2}{2}},\quad (n\to\infty).$$

4. **用连续性定理收尾**我们看到 $\varphi_{Z_n}(t)\to e^{-\frac{t^2}{2}}$ 并且 $e^{-\frac{t^2}{2}}$ 是 $\mathcal{N}(0,1)$ 的特征函数。

$$\varphi_{\mathcal{N}(0,1)}(t)=e^{-\frac{t^2}{2}}.$$

由**特征函数的连续性定理**，可知若 φ_{Z_n} 逐点收敛到某个连续且在 0 处等于 1 的函数，那么该函数便是某概率分布的特征函数；又因分布由特征函数唯一决定，故 Z_n 在分布上收敛到 $\mathcal{N}(0,1)$。

以上就是中心极限定理的完整特征函数证明框架。它在分析极限时**无需直接处理卷积**，大幅减少了技术复杂度。

更多应用与总结

- **多元正态分布：**一旦多元随机向量的任何线性组合均呈正态，其特征函数可写成 “$\exp\big(i\,\mathbf{t}^\top\boldsymbol{\mu}-\frac{1}{2}\,\mathbf{t}^\top\Sigma\,\mathbf{t}\big)$” 的形式，反向也可从该形式识别

出“它是高斯分布”。

- **随机过程：**在更一般的随机过程（如鞅、马氏过程）或稳定分布研究中，特征函数与独立性、渐近收敛等也紧密相连。

  小结：

  - 乘积性质使得独立随机变量之和的分析变得易于处理，卷积问题转化为简单的乘积形式；
  - 连续性定理为**极限过程**搭建了严谨的收敛判据：只要点态收敛并保证在 0 处的连通性，就能说明分布收敛；
  - 中心极限定理等重要结果正是由此衍生，体现了特征函数法在极限理论中的强大威力；
  - 再结合多维场景，可轻松识别并表征高斯或其他可判定的分布家族。

5.3 矩母函数（MGF, Moment Generating Function）

5.3.1 定义

对于随机变量 X，若下式在某邻域内收敛，则可定义**矩母函数**（MGF）$M_X(t)$：

$$M_X(t) \;=\; E\big[e^{\,tX}\big] \;=\; \int_{-\infty}^{+\infty} e^{\,tx}\,\mathrm{d}P_X(x),$$

其中 $t \in \mathbb{R}$。MGF 存在与否与随机变量是否具有“指数型可积”关系紧密；若 X 的尾部衰减速度不足，则 $M_X(t)$ 可能在 $t>0$ 或 $t<0$ 出现发散情形。

5.3.2 性质与应用

- **存在性局限**：MGF 并不一定总是存在（如柯西分布就没有有限的 MGF），所以不能像特征函数那样“随时可用”。
- **推导各阶矩**：若 $M_X(t)$ 在 $t=0$ 附近可展开为幂级数，则

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)\bigg|_{t=0} \;=\; E\big[X^k\big].$$

- **简化卷积**：若 X、Y 独立，则 $M_{X+Y}(t) = M_X(t)\,M_Y(t)$，与特征函数类似，可以将分布的卷积运算转换为函数的乘积运算。

5.3.3 常见分布的 MGF 示例

- **正态分布** $X \sim \mathcal{N}(\mu, \sigma^2)$：

$$M_X(t) \;=\; \exp\!\Big(\mu t + \tfrac{1}{2}\sigma^2 t^2\Big).$$

- **指数分布** $X \sim \mathrm{Exp}(\lambda)$：

$$M_X(t) \;=\; \frac{\lambda}{\lambda - t}, \quad (\text{需满足 } t < \lambda).$$

- **二项分布** $X \sim \mathrm{Bin}(n, p)$：

$$M_X(t) \;=\; \Big(1 - p + p\,e^{\,t}\Big)^{n}.$$

特征函数与 MGF 的对比

- **存在性：**
  - 特征函数 $\varphi_X(t)$ 对任意实值随机变量总是存在并且处处连续；
  - MGF $M_X(t)$ 需要“可积到 e^{tX}”的尾部条件，不一定存在。
- **表达形式：**
  - 特征函数是 $\exp\big(i\,t\,X\big)$ 的期望，往往带有复数因子 i；
  - MGF 则是 $\exp\big(t\,X\big)$ 的期望，在真实数轴上分析，但有收敛范围的限制。
- **应用场景：**
  - 特征函数：在极限定理、独立和卷积分析（如中心极限定理、鞅收敛理论等）中更常见，且不受存在性限制；
  - MGF：只要存在，获取各阶矩、推导多项式展开时常更为方便，亦能用于卷积简化。

小结

要点总结：

- **特征函数** $\varphi_X(t) = E[e^{i\,t\,X}]$ 在复平面上表征分布，**总是存在**、**可唯一重构**分布，并在处理独立性的卷积问题时提供了**乘积简化**。
- **矩母函数** $M_X(t) = E[e^{\,t\,X}]$ 若存在，可通过其幂级数展开**提取各阶矩**，在可用情形下也常用于简化运算与建模。
- 二者皆可视为“对分布的一种变换”并具有**唯一确定分布**的作用，但应用范围和存在性条件上略有差异。实际分析中往往因分布性质而择优使用特征函数或 MGF。

6 次序统计量（Order Statistics）

当我们观测到一个样本序列 $\{X_1, \ldots, X_n\}$ 时，有时更感兴趣的是这些观测值**按数值大小排列**后的信息，而不仅仅是它们各自的原顺序。此时，就会用到“次序统计量（Order Statistics）”的概念。

6.1 基本定义

给定样本： $\big(X_1, X_2, \ldots, X_n\big)$。将其从小到大**排序**后，得到新的序列

$$X_{(1)} \;\le\; X_{(2)} \;\le\; \ldots \;\le\; X_{(n)},$$

其中 $X_{(k)}$ 称为 **第 k 个次序统计量**。例如：

- $X_{(1)}$ 称为**样本最小值**；
- $X_{(n)}$ 称为**样本最大值**；
- 若 $k = \frac{n+1}{2}$（适当取整），则 $X_{(k)}$ 会与样本中位数紧密关联。

含序 vs. 不含序： 在研究普通随机变量序列时，我们并不关心各变量间的数值大小关系，仅关注其**联合分布**。而次序统计量会**明**

确地将样本点进行数值上的排名，从而构造出与“分位数”、“极值”等问题密切相关的新变量族。

6.2 用途与示例

- **极值理论：**关心 $X_{(n)}$（最大值）或 $X_{(1)}$（最小值）随样本规模增大的分布行为，在保险风险、工程可靠性等领域常见。
- **分位数估计：**统计中常用 $X_{(k)}$ 来近似某个分位点（如中位数、四分位数）。

一个简单例子：样本最小值分布

设 $\{X_i\}_{i=1}^n$ 为**独立同分布**（i.i.d.）随机变量，且设它们共有一个连续分布函数 $F_X(x)$。记 $M_n = X_{(1)} = \min(X_1, \ldots, X_n)$。我们想求 $P(M_n > x)$：

$$\begin{aligned} P\big(M_n > x\big) &= P\big(X_1 > x, X_2 > x, \ldots, X_n > x\big) \\ &= \prod_{i=1}^{n} P\big(X_i > x\big) \quad \text{(因独立性)} \\ &= \big[1 - F_X(x)\big]^n. \end{aligned}$$

所以

$$F_{M_n}(x) \;=\; P\big(M_n \le x\big) \;=\; 1 - \big[1 - F_X(x)\big]^n.$$

从而可知最小值的分布函数：

$$F_{M_n}(x) = 1 - [\,1 - F_X(x)\,]^n.$$

类似地，若令 $X_{(n)} = \max(X_1, \ldots, X_n)$，则可得

$$F_{X_{(n)}}(x) = \big[F_X(x)\big]^n.$$

如果想要看更多例子，可以参看【Probability Theory A First Course in Probability Theory and Statistics (Werner Linde)】一书的 P141

小结

要点回顾：

- **次序统计量**是将样本**按值排序**后得到的一组随机变量，常用于极值研究和分位数分析。
- 与“自然顺序”无关的随机变量序列比较，次序统计量通过数值排序增加了新的结构性信息。
- **极值**（最小值、最大值）是最简单的次序统计量，可显式得到其分布公式——通常蕴含独立性假设。
- 对更一般的第 k 个次序统计量 $X_{(k)}$，也有类似的可解析分布形式，应用在各种统计与概率推断场景中十分广泛。

7 随机变量的函数与代数运算

在概率论中，常见的情形是我们对已有的随机变量进行函数变换或代数组合，以得到新的随机变量，进而研究它们的分布和特征。下文将结合单一函数映射、多随机变量组合等多角度，阐述如何在分布层面完成相应计算，并辅以常见场景下的公式示例。

7.1 单一函数映射

可测映射：为什么可测？

若 X 是一个实值随机变量，且 $g:\mathbb{R}\to\mathbb{R}$ 是**可测函数**（如 Borel 可测、分段连续等），那么

$$Y \;=\; g(X)$$

本质上仍是一个随机变量。原因在于：**原像**的可测性得到保留——对任意 Borel 集 $B\subseteq\mathbb{R}$，

$$Y^{-1}(B) \;=\; \{\omega\in\Omega : g(X(\omega))\in B\} \;=\; \{\omega\in\Omega : X(\omega)\in g^{-1}(B)\} \;\in\; \mathcal{F}.$$

由此可知，Y 在概率空间 $(\Omega,\mathcal{F},P)$ 上确为可测映射。

新的分布：推前测度

既然 $Y = g(X)$ 是新的随机变量，那么它在 $\mathbb{R}$ 上诱导出一个分布 P_Y。通过“**推前测度**”思想，可用 X 的原分布 P_X 来确定：

$$P_Y(B) \;=\; P\big(Y \in B\big) \;=\; P\big(g(X) \in B\big) \;=\; P\big(X \in g^{-1}(B)\big) \;=\; P_X\big(g^{-1}(B)\big).$$

具体到连续或离散情形时，还可以选用 PDF/PMF 或 CDF 的方法来进行**更直观**的计算。

> **示例：线性变换**设 $Y = aX + b$（$a \neq 0$，$a, b \in \mathbb{R}$）。当 X 具有 PDF $p_X(x)$ 时，Y 的 PDF 可写为
>
> $$p_Y(t) \;=\; \frac{1}{|a|}\, p_X\Big(\frac{t-b}{a}\Big), \quad t \in \mathbb{R}.$$
>
> 若使用 CDF，则有
>
> $$F_Y(t) = \; F_X\Big(\frac{t-b}{a}\Big), \quad (\text{若} a > 0),$$
>
> $$\text{或者} \;\; F_Y(t) = \; 1 - F_X\Big(\frac{t-b}{a}\Big), \quad (\text{若} a < 0).$$

7.2 多随机变量情形

从单变量到向量：映射在高维上的推广

若我们有一个随机向量 $\mathbf{X} = (X_1, \ldots, X_n)$，以及一个可测函数 $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$，则定义

$$\mathbf{Z} \;=\; \mathbf{g}(\mathbf{X}) \;=\; \big(g_1(X_1, \ldots, X_n),\, g_2(X_1, \ldots, X_n),\, \ldots,\, g_m(X_1, \ldots, X_n)\big)$$

依旧是一个随机向量。它的分布可在 $\mathbb{R}^m$ 上通过**联合测度** $P_{\mathbf{X}}$ 进行推前计算：

$$P_{\mathbf{Z}}(B) \;=\; P\big(\mathbf{Z} \in B\big) \;=\; P\big(\mathbf{X} \in \mathbf{g}^{-1}(B)\big), \quad B \subseteq \mathcal{B}(\mathbb{R}^m).$$

> **示例：线性变换**对于向量 $\mathbf{X} \in \mathbb{R}^n$，若定义 $\mathbf{Z} = A\mathbf{X} + \mathbf{b}$（其中 A 为 $m \times n$ 矩阵，$\mathbf{b} \in \mathbb{R}^m$ 为常向量），这是最常用的仿射变换。易

知

$$P\big(\mathbf{Z}\in B\big) \;=\; P\big(\mathbf{X}\in A^{-1}(B-\mathbf{b})\big),$$

进而可通过 $\mathbf{X}$ 的联合分布来分析 $\mathbf{Z}$ 在 $\mathbb{R}^m$ 上的分布。

7.3 随机变量函数的基本计算公式

7.3.1 线性变换：$Y = aX + b$

如上所示，若 X 具有 CDF $F_X(x)$，则

$$F_Y(t) \;=\; \begin{cases} F_X\big(\frac{t-b}{a}\big), & a>0,\\ 1-F_X\big(\frac{t-b}{a}\big), & a<0. \end{cases}$$

对 PDF 而言，若 X 具备 p_X，则

$$p_Y(t) \;=\; \frac{1}{|a|}\,p_X\Big(\frac{t-b}{a}\Big).$$

7.3.2 两个随机变量相加/相减

令 $Z = X \pm Y$。若 X, Y 独立且有 PDF p_X, p_Y，则

$$p_Z(z) \;=\; \int_{-\infty}^{+\infty} p_X(x)\,p_Y\big(z-x\big)\,\mathrm{d}x,$$

称为**卷积**。若 X, Y 的分布为离散型，可对应地使用**离散卷积**公式

$$p_Z(z) \;=\; \sum_{x\in\mathcal{X}} p_X(x)\,p_Y\big(z-x\big).$$

在更深入分析时，可考虑**特征函数**或 **MGF** 的**乘积性质**以简化多次叠加情形。

证明为：

为说明上述公式的来由，以下以 $Z = X + Y$ 为例进行推导。假设 X, Y 独立且均为连续型随机变量，其 PDF 分别为 f_X 与 f_Y。先写出 Z 的分布函数：

$$F_Z(z) \;=\; P(Z\le z) \;=\; P(X+Y\le z).$$

由于 X 与 Y 独立，联合 PDF 可写为 $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$。于是，

$$F_Z(z) \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x)\, f_Y(y)\, \mathrm{d}y\, \mathrm{d}x.$$

将 $f_X(x)$ 提到外层积分：

$$F_Z(z) \;=\; \int_{-\infty}^{\infty} f_X(x) \left[\int_{-\infty}^{z-x} f_Y(y)\, \mathrm{d}y \right] \mathrm{d}x \;=\; \int_{-\infty}^{\infty} f_X(x)\, F_Y(z-x)\, \mathrm{d}x.$$

为了得到 PDF f_Z，对 $F_Z(z)$ 关于 z 做导数：

$$f_Z(z) \;=\; \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) \;=\; \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{\infty} f_X(x)\, F_Y\big(z-x\big)\, \mathrm{d}x.$$

在满足适当条件（如 Fubini 定理可交换微分与积分）的前提下，

$$f_Z(z) \;=\; \int_{-\infty}^{\infty} f_X(x)\, \frac{\mathrm{d}}{\mathrm{d}z} F_Y\big(z-x\big)\, \mathrm{d}x \;=\; \int_{-\infty}^{\infty} f_X(x)\, f_Y\big(z-x\big)\, \mathrm{d}x.$$

这便是卷积形式

$$f_Z(z) \;=\; (f_X * f_Y)(z),$$

即 $Z = X + Y$ 的 PDF 等于 X, Y 各自 PDF 的卷积。

7.3.3 随机变量相乘 / 做商

若我们关心 $Z = X \cdot Y$ 或 $Z = \frac{X}{Y}$，理论上也可通过积分或求和拿到其分布，但往往较加法卷积更复杂。例如，在连续情形且 $Y > 0$ 的假设下，有经典的公式（以下为参考形式）：

$$Z = X \cdot Y: \quad r_Z(x) \;=\; \int_0^{\infty} p_X\Big(\tfrac{x}{y}\Big)\, \frac{p_Y(y)}{y}\, \mathrm{d}y,$$

$$Z = \tfrac{X}{Y}: \quad r_Z(x) \;=\; \int_0^{\infty} y\, p_X(x\, y)\, p_Y(y)\, \mathrm{d}y.$$

这需要再区分 $y > 0$ 或 $y < 0$ 等符号区间，视具体情况而定。

总结与展望

要点回顾：

- **函数映射**：一个可测函数 $g(\cdot)$ 作用于随机变量 X 后可生成新随机变量 $g(X)$。其分布可通过“原像”或“积分/求和”直接计算。
- **线性变换、加减法**：是最常见的基本操作；独立情形下可用**卷积**（或特征函数乘积）来处理和的分布，简化复杂度。
- **乘商及更广泛函数**：概念上同理，但计算公式更繁琐。必要时可借助分区间积分或符号分析。
- **多维映射**：当随机向量 $\mathbf{X}$ 映射到 $\mathbf{Z} = \mathbf{g}(\mathbf{X})$ 时，也可用联合分布的“原像”法推得新向量分布。线性变换等在高维统计与机器学习中尤其常见。

随着函数形式和随机变量维度的提升，相应的分布获取方法虽在原理上不变，但在具体计算和推导上会变得更加多样化与灵活。在后续的分析中，**特征函数法**、**MGF**、以及**马尔可夫核**等更高级工具也都能扮演重要角色，以支持更复杂的随机函数运算与概率推断。

8 用简单方法模拟随机变量的分布

在实际中，如果我们想要在计算机中（或者通过某些可重复的实验手段）“模拟”某个概率分布所对应的随机变量，就需要先寻找一类**易于实现的随机源**。最常见的“易于实现”方式是：假设我们能够（在算法或实验上）**获得一个** $[0,1]$ **上的均匀随机数**，然后通过**恰当的函数变换**将它“转化”成想要的分布。下文先介绍如何模拟 $[0,1]$ 上的均匀分布，再说明如何利用它去模拟离散和连续分布（如泊松、指数、正态等）。

8.1 模拟 $[0,1]$ 上均匀分布

一个最简单且常被引用的“理想模型”是**无限次抛硬币**：

$$\omega = (X_1, X_2, X_3, \dots), \quad \text{其中} X_n = \begin{cases} 0, & \text{概率}1/2, \\ 1, & \text{概率}1/2. \end{cases}$$

相互独立的 $\{X_n\}$ 就像一个二进制序列，我们可以将其拼接成

$$U = 0.X_1X_2X_3\dots$$

在二进制小数下 interpret 成一个 $[0,1]$ 区间内的数。从测度论角度看，这个 U 事实上服从 $[0,1]$ 上的均匀分布 $\mathrm{Unif}(0,1)$——虽然在实际机器中无法真的进行无限次抛硬币，但这一思路为我们构建了一种理想的“均匀随机源”。

8.2 模拟离散分布：区间切割法

算法思路

若想模拟一个离散随机变量 X，其可能的取值集合为 $\{x_k\}_{k=0}^{\infty}$（或从 1 开始），概率分配为

$$P\{X = x_k\} = p_k, \quad \text{其中} \sum_k p_k = 1.$$

可以**在** $[0,1]$ **上切割**出一系列不相交区间来对应各 x_k。设定义：

$$I_0 = [0, p_0), \quad I_1 = \Big[p_0,\ p_0 + p_1\Big), \quad I_2 = \Big[p_0 + p_1,\ p_0 + p_1 + p_2\Big), \dots$$

其中

$$\sum_{i=0}^{k-1} p_i \le \sum_{i=0}^{k} p_i.$$

这样，区间 I_k 的长度就是 p_k。再定义一个**映射函数** $f : [0,1] \to \{x_k\}$，令

$$f(u) \ = \ x_k \quad \text{若} u \in I_k.$$

步骤：

1. 先从均匀随机源中生成一个 $U \sim \mathrm{Unif}(0,1)$；

2. 检查 U 落在哪个 I_k 中；若 $U \in I_k$，则输出 x_k。

这样就得到了一个随机变量 $f(U)$，其分布满足

$$P\{f(U) = x_k\} \;=\; P\{U \in I_k\} \;=\; |I_k| \;=\; p_k.$$

故 $f(U)$ 模拟的正是该离散分布。

泊松分布示例

泊松分布 $X \sim \mathrm{Pois}(\lambda)$ 满足

$$P\{X = k\} \;=\; \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots.$$

对应地，可定义

$$I_0 = [\,0,\; e^{-\lambda}\,), \quad I_1 = \Big[e^{-\lambda},\; e^{-\lambda}{+}\lambda e^{-\lambda}\Big), \quad I_2 = \Big[\sum_{j=0}^{1} \frac{\lambda^j e^{-\lambda}}{j!},\; \sum_{j=0}^{2} \frac{\lambda^j e^{-\lambda}}{j!}\Big), \ldots$$

再令 $f(u) = k$ 若 $u \in I_k$。从均匀分布采样一个 U，即可将其转换为泊松分布的取值。

8.3 模拟连续分布：反函数法

CDF 与伪反函数

对一个连续随机变量 X，其累积分布函数（CDF）$F(x) = P(X \le x)$ 是单调不减且右连续；记为

$$F(x) \;=\; \int_{-\infty}^{x} p(t)\,\mathrm{d}t,$$

若存在可微部分则 $p(t)$ 则为 PDF。为了对所有情形进行覆盖（包括平坦段），我们引入**广义逆函数**或**左连续反函数**：

$$F^{-}(s) \;=\; \inf\big\{\, t \in \mathbb{R} : F(t) \ge s \big\}, \quad s \in [0,1].$$

（也常见记法 F^{-1}、$\inf\{t : F(t) \ge s\}$ 等。）

算法思路

1. 依旧先从均匀随机源中获取一个 $U \sim \mathrm{Unif}(0,1)$；

2. 定义 $Y = F^{-}(U)$，其中 F^{-} 表示上文定义的伪反函数。

那么可证明 Y 与 X 同分布，换言之

$$P(Y \le t) \;=\; P\big(F^{-}(U) \le t\big) \;=\; P\big(U \le F(t)\big) \;=\; F(t).$$

为便于展示，该推导可分行写作：

$$\begin{aligned} P\{F^{-}(U) \le t\} &= P\Big(U \le \sup\{u : F^{-}(u) \le t\}\Big) \\ &= P\big(U \le F(t)\big) \\ &= F(t). \end{aligned}$$

这便是所谓**反函数法**或 **Inverse Transform Sampling**。它在模拟指数分布、均匀分布甚至一些不太规则的连续分布时都极其常用。

简单示例：指数分布

若 $X \sim \mathrm{Exp}(\lambda)$，则

$$F(x) = 1 - e^{-\lambda x}, \quad x \ge 0.$$

其**伪反函数**为

$$F^{-}(s) \;=\; -\frac{1}{\lambda}\ln(1-s), \quad 0 \le s < 1.$$

若我们生成 $U \sim \mathrm{Unif}(0,1)$，则

$$X = \; F^{-}(U) = \; -\frac{1}{\lambda}\ln\big(1-U\big)$$

便服从 $\mathrm{Exp}(\lambda)$。实际上由于 U 与 $1-U$ 分布相同，亦常写作 $-\frac{1}{\lambda}\ln U$。

小结

总结要点：

- **基础随机源：**通过（理想化）无限次掷硬币或算法伪随机数，可获得 $U \sim \text{Unif}(0,1)$。
- **离散分布：**用区间切割法，将 $[0,1]$ 分段，并在段上分别赋予相应概率，利用 $f(U)$ 达到模拟目的。
- **连续分布：**借助 CDF 的“广义逆函数” F^-，使用 $X = F^-(U)$ 生成目标分布；即**反函数法**。
- **逻辑基础：**以上方法都是基于“将简单分布（均匀）**推前测度**到更复杂分布”的思路，从而将模拟问题转化为“$[0,1]$ 区间上的测度变换”。

通过这样的思路，我们可以在计算机或实验上，先产生“近似的”$\text{Unif}(0,1)$ 随机数，再通过区间分割或伪反函数等手段模拟出几乎任何想要的分布（只要其 CDF 信息已知或易于处理）。在实践中，再搭配高效算法或分片逼近技巧，就可以高效且灵活地完成各种随机模拟需求。

数理统计部分

在概率论的部分中，我们建设了一套概率测度的系统，用来数值地度量事件发生的可能性。现在，我们将其应用到现实中的随机系统中，帮我们预测其性质。首先，如果要取得这种性质的预测，我们需要知道具体应该采用哪种概率测度，否则我们就不能运用随机变量和分布的性质来预测其行为。随之而来的一个问题是，有太多太多的概率测度都是合理的了，因为只要概率不为 0，就算出现一些与分布性质完全错误的样本，我们也不能完全否定这一概率测度的正确性。比如，一个街头魔术师掷了 100 次骰子，每次都掷到了 6。在这种情况下，我们能够明显地察觉出，骰子点数的分布应该不是均匀分布，因为如果是均匀分布的话出现这一事件的概率实在太小了，基本不可能发生；但是我们却也不能否定“均匀分布”完全是错误的，因为均匀分布确实给出了这种可能。

如何寻找到一个最能描述现实中随机事件的发生状态的概率系统？为了解决这一问题而产生的方法论便是所谓的“数理统计”。数理统计的思想是，通过观测样本，并制定一定的标准（如最大似然、统计无偏），来选取一套最为合理的概率系统，从而按照一个我们认为可信的方式来预测随机事件的性质。

这个被构建出来的系统被称为**概率统计模型**。其正规的定义是：

随机变量的极限定理告诉我们，观察的样本越多，样本总体上的行为就越稳定，从而我们能够更容易地选出一个好的概率系统

无偏估计和最大似然估计是两种最主要的“选取概率模型的方法”，最大似然估计认为“对于产生的样本，在所有的概率统计模型之中，那个此样本发生的概率最大的概率统计模型是最优的”，而无偏估计认为“若某一种估计方式在概率模型的期望之下能够给出参数的准确值，则对于取得的样本采用此估计方式所得到的概率统计模型是最优的”

假设检验是通过合适的构造，以样本为依据，以收缩参数空间，使得概率模型的可选择范围尽可能小

置信空间是以“真值所确定的概率测度之下，样本发生的概率不低于

100(1-α)” % 来确定真值存在的范围

一般而言，这四种统计方法是一起出现的，共同给出一个合理的概率统计模型。

非常推荐参看 Probability Theory A First Course in Probability Theory and Statistics (Werner Linde) 的数理统计部分。

(一) 随机变量序列及其极限定理

在现实中，当我们考虑某些随机试验时，往往会构造出一个「无限大」的样本空间。以掷骰子为例：如果我们从理论上允许无限次投掷，那么各种可能的投掷序列（如第 1 次投到 3、第 2 次投到 6，……）所形成的样本空间确实是无限的。然而在实际场景中，我们未必真的进行无限次投掷；也许只投 3 次就结束了。在「无限次投掷」的大样本空间里，「只投 3 次」的情形可以视为该大空间中的一个「边缘」部分。随着投掷次数的增加，我们对事件的刻画会变得越来越细，逐渐接近大空间中整体分布的特征。

为了解释这种「在有限次/无限次重复试验下，如何刻画样本空间分布演变」的问题，通常会引入**随机变量序列**的概念。可以把序列中的每一个随机变量视作在「进行到第 n 次试验」时所形成的一种分布表征，而当 n 越来越大时，这种表征就变得越来越复杂，进而能代表在更多次迭代后形成的整体分布规律。最终，「随机变量序列的极限」可以与「无限次迭代所产生的样本空间的分布」建立对应关系。

在现实生活中，我们也常常关心这种情况；一方面是我们可以用序列的方法产生更为复杂的样本空间，另一方面这种“无穷次重复产生的样本空间”在分布上有一些非常好的性质（比如说趋向于 0 的方差），同时其本身保有某些单次重复的分布的特征（如均值）

也就是说：

1. 构造复杂样本空间：通过逐次迭代或重复，可以产生更加复杂、多样的试验结果，丰富了我们对随机现象的建模能力。

2. 分布的稳定性：随着重复次数的增加，一些随机变量的方差会缩小，或者说我们对它们平均行为的把握度会提高。

3. 与单次分布特征的联系：尽管重复让分布更稳定，但它仍然保留了一定的「单次分布」的核心属性（如均值等），这为我们研究单次分布提

供了统计上的方法论依据。

所以，对于这种复杂的样本空间，我们可以通过随机变量序列的方式，研究其分布性质的演变

同时，我们可以利用“无穷次重复产生的样本空间”的分布特性，通过构造重复实验的方法，来研究单次分布的特征

在这一背景下，人们会使用**大数定律**（Law of Large Numbers, LLN）和**中心极限定律**（Central Limit Theorem, CLT）来探讨「无限次重复实验」的一些重要性质。

大数定律揭示了在什么条件下，随着重复次数的增多，观测量会收敛到某个稳定的特征值（通常是期望），即所谓的「频率稳定性」;

中心极限定律则进一步说明了在收敛的过程中，这些观测量如何近似服从某种分布（常见的是正态分布）。

两大定律也指出了一个事实：虽然单次试验结果是随机的，但若让同一分布下的试验重复足够多次，整体表现会趋向于「较小的随机性」。换言之，「通过大量重复可以消解随机性」——这正是频率统计学派的重要基石，在实验方法和科学研究中都发挥了深远的影响（典型例子是为了减小测量误差，经常多次测量并取平均值）。

本节将依次简要介绍这三个要点：**随机变量序列**、**大数定律 (LLN)** 与**中心极限定律 (CLT)**，从而说明在「无限次重复产生的样本空间」里，分布如何演变以及在何种意义上表现出稳定性和规律性。

1 随机变量序列及其收敛

背景与直观理解

在现实当中，我们构造的样本空间往往会有**无穷**的可能状态（例如可以无限次投掷骰子）。这意味着在理论层面，“投 3 次”这种有限试验只是整个无穷次投掷所构成样本空间中的一个“边缘事件”。当我们不断累积试验次数，就能从一个较粗糙的事件（比如“只投 3 次”）走向更“精细”的事件描述（比如“投了 10 次或更多”），这就逐渐逼近或揭示了样本空间的整体结构。

随机变量序列正是用来刻画这种“随试验次数迭代、愈加细致的分布行为”的重要工具。它不仅能表示“第 n 次试验”这一有限时刻的分布情况，也能在极限意义上表示当试验次数趋于无穷时的统计性质。

1.1 有限随机变量序列

(有限随机变量序列) 在同一个概率空间 $(\Omega, \mathcal{F}, P)$ 中，挑出有限个随机变量 $(X_1, \ldots, X_n)$；这就构成一个有限随机变量序列。比如进行 3 次骰子试验时，可令 X_1, X_2, X_3 分别表示每次的骰子点数。

对于有限序列，我们常将其视为一个 $\mathbb{R}^n$ 上的**随机向量**，具有联合分布 $P_{(X_1,\ldots,X_n)}$。当 n 增加时，所能刻画的“组合事件”更加复杂。

1.2 无穷随机变量序列

(无穷随机变量序列) 若我们在 $(\Omega, \mathcal{F}, P)$ 中拥有可数多（无限个）随机变量 $X_1, X_2, \ldots$，则称之为一个无穷随机变量序列。这恰好对应了“可观念上的无限重复试验”。

Lim inf 与 Lim sup 的补充说明

在测度理论或概率论的语境下，对于随机变量序列 $\{X_n\}$，我们通常定义：

$$\liminf_{n\to\infty} X_n \;=\; \sup_{m\geq 1} \inf_{n\geq m} X_n, \quad \limsup_{n\to\infty} X_n \;=\; \inf_{m\geq 1} \sup_{n\geq m} X_n.$$

如果几乎处处（a.s.）有 $\liminf X_n = \limsup X_n$，则我们称 $\{X_n\}$ **几乎处处收敛**，并将它们的共同极限记为 $\lim_{n\to\infty} X_n$.

集合视角的解释： 当我们转而讨论事件序列 $\{A_n\}$（或一般的集合序列）时，也有类似的下极限与上极限定义：

$$\liminf_{n\to\infty} A_n \;=\; \bigcup_{m=1}^{\infty} \bigcap_{n\geq m} A_n, \quad \limsup_{n\to\infty} A_n \;=\; \bigcap_{m=1}^{\infty} \bigcup_{n\geq m} A_n.$$

$\liminf_{n\to\infty} A_n$ 包含的是“从某个 N 开始就一直在后续所有 A_n 里”的元素或事件（最终不再离开）。

$\limsup_{n\to\infty} A_n$ 包含的是“可以在无限多个 A_n 中出现”的元素或事件（可能断断续续，但总能反复出现）。

在概率论中，“$\liminf A_n$ 发生”意味着事件 A_n 最终会一直发生（从某个时刻起不再失效）；而“$\limsup A_n$ 发生”意味着事件 A_n 会发生无穷多次。

二者的联系： 对于随机变量序列而言，$\liminf X_n$ 与 $\limsup X_n$ 分别可视作“从某个索引后，X_n 总是大于/小于某值”的极端行为。

如果它们相等（a.s.），就表示序列不再在两端摆动，而几乎处处稳定收敛到某个值。

虽然这里我们针对随机变量序列给出定义和直观解释，背后的思路与事件（集合）序列的极限概念在本质上是一致的。

1.3 无穷随机变量序列的四种收敛方式

要研究 $\{X_n\}$ 在 $n \to \infty$ 时的行为，可从不同角度定义收敛：

1) 几乎处处收敛 (Almost Sure Convergence)

$$X_n \xrightarrow{\text{a.s.}} X \iff P\Big(\{\omega : X_n(\omega) \to X(\omega)\}\Big) = 1.$$

即除了一个概率为 0 的例外集外，ω 都满足 $X_n(\omega) \to X(\omega)$。这是最强的一种点态收敛。

2) 依概率收敛 (Convergence in Probability)

$$X_n \xrightarrow{P} X \iff \forall \varepsilon > 0, \quad P\big(|X_n - X| > \varepsilon\big) \to 0.$$

含义：X_n 偏离 X 超过任何 ε 的概率趋于 0。比 a.s. 收敛稍弱，但常用于统计推断等应用。

3) 依分布收敛 (Convergence in Law / Distribution)

$$X_n \xrightarrow{d} X \iff F_{X_n}(t) \to F_X(t) \quad (\text{在 } t \text{ 的连续点处}).$$

F_{X_n} 是 X_n 的分布函数。它仅要求**分布函数**逐点收敛，是最弱的一种收敛形式。

4) L^2 (或 L^α) 收敛

$$X_n \xrightarrow{L^2} X \iff E\big[(X_n - X)^2\big] \to 0.$$

可以更一般地定义 L^α 收敛：$E\big[|X_n - X|^\alpha\big] \to 0$。此种收敛侧重随机变量间的均方误差或更高阶误差。

各收敛方式的层级关系 一般地，若 $\{X_n\}$ 在 L^2（或更一般的 L^α）意义下收敛于 X，则可推出 $X_n \xrightarrow{P} X$（依概率收敛），而依概率收敛又能推出依分布收敛。另一方面，若 X_n 几乎处处收敛于 X，也会推出依概率收敛（从而推出依分布收敛）。因此从强到弱可大致排为：

$$L^\alpha \text{ 收敛} \implies \text{依概率收敛} \implies \text{依分布收敛}.$$

$$\text{a.s. 收敛} \implies \text{依概率收敛} \implies \text{依分布收敛}.$$

需要注意的是，L^α 收敛与 a.s. 收敛并非总能互相推出，还需考虑各自的适用条件与随机变量本身的有界性或矩条件等。

2 大数定律

大数定律（LLN）展示了：当我们把同一分布的随机变量**独立且同分布**地重复采样并取平均后，这个平均值会“逐渐稳定”在某个数（往往是分布的期望）周围。它可以解释“通过大量重复实验可以减小随机误差”这一事实，也是**频率统计学派**立论的基础。

2.1 弱大数定律

Chebyshev 弱大数定律

内容 设 $\{X_k\}$ 为 i.i.d. 随机变量，$E[X_k] = \mu$ 且 $\mathrm{Var}(X_k) = \sigma^2 < \infty$，定义

$$\overline{X}_n \;=\; \frac{1}{n}\sum_{k=1}^{n} X_k.$$

则

$$\overline{X}_n \xrightarrow{P} \mu,$$

即在概率意义下收敛到 μ。

证明（Chebyshev 不等式） Chebyshev 不等式由 Markov 不等式衍生而来：

$$P\big(|\overline{X}_n - \mu| > \varepsilon\big) \;\le\; \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2} \;\to 0.$$

因此 $\overline{X}_n \to \mu$ 依概率收敛。

Khinchin 弱大数定律 (General WLLN)

内容 仍是 i.i.d. 但只假设 $E[X_1] = \mu$（不必要求方差有限），则依然有

$$\overline{X}_n \xrightarrow{P} \mu.$$

思路 先将 X_k 截断（例如限制到 $|X_k| \le M$），使其方差可控，再用 Chebyshev 或其他工具处理 “尾部” 的影响——可以做得任意小。此法常称 “截尾技巧”。

Kolmogorov 弱大数定律

在更一般的情形下（可允许部分非 i.i.d. 结构，只要满足一定独立或相依条件），也有各种 WLLN 的推广版本，如 Kolmogorov、Marcinkiewicz-Zygmund 等定理。

L^2 弱大数定律

若在 i.i.d. 情形下，且 $E[X_1^2] < \infty$，则可进一步得到

$$\overline{X}_n \xrightarrow{L^2} \mu,$$

即在均方意义下收敛。具体证明是基于 $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$ 的事实。

2.2 强大数定律

内容 强大数定律给出“几乎处处收敛”的保证。例如 Kolmogorov SLLN：对于 i.i.d. X_k，若 $E[|X_1|] < \infty$，则

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu.$$

证明思路 往往要用到 **Kolmogorov 不等式**和 **Borel-Cantelli 引理**等高级技术，控制“大偏差”出现的概率之和可收敛，从而保证其发生仅是有限次，不影响几乎处处结论。核心证明采用特征函数和小量分析的方法。

3 中心极限定律 (CLT)

大数定律告诉我们：$\overline{X}_n$ 会收敛到期望 μ，意味着“偏离”在大样本下会缩小。然而，这个“偏离”本身如何分布？CLT 告诉我们“偏离的归一化”——最典型的就是 $(S_n - n\mu)/(\sigma\sqrt{n})$——将趋于正态分布。

内容

设 $\{X_k\}$ i.i.d.，$E[X_k] = \mu$，$\mathrm{Var}(X_k) = \sigma^2 < \infty$。令

$$S_n = \sum_{k=1}^{n} (X_k - \mu), \quad Z_n = \frac{S_n}{\sigma\sqrt{n}}.$$

则

$$Z_n \xrightarrow{d} \mathcal{N}(0,1),$$

即分布收敛于标准正态。

证明：特征函数法

- 设 $\varphi_{X_1}(t)$ 为 $X_1 - \mu$ 的特征函数；展开可得 $\varphi_{X_1}(t) \approx 1 - \frac{t^2\sigma^2}{2} + o(t^2)$。
- 独立性使 S_n 的特征函数为 $\left[\varphi_{X_1}(t)\right]^n$。
- 令 $Z_n = \frac{S_n}{\sigma\sqrt{n}}$，则 $\varphi_{Z_n}(t) = \varphi_{S_n}\left(\frac{t}{\sqrt{n}\,\sigma}\right) = \left[\varphi_{X_1}\left(\frac{t}{\sqrt{n}\,\sigma}\right)\right]^n \to e^{-\frac{t^2}{2}}$.

- $e^{-t^2/2}$ 即是 $\mathcal{N}(0,1)$ 的特征函数，由此可得 $Z_n \xrightarrow{d} \mathcal{N}(0,1)$。

总结与思考

要点回顾：

- **序列与无限样本空间**：通过随机变量序列 $\{X_n\}$，我们可刻画“无限次实验”背后的概率结构，并讨论其在极限下的行为。
- **收敛形式**：几乎处处、依概率、依分布、L^α 四类收敛各有不同强度与应用场景。
- **大数定律**：说明通过重复可让平均值稳定地逼近分布的均值，从而“消解”单次试验的随机性；弱与强之分对应不同的收敛定义。
- **中心极限定律**：刻画了偏离平均的归一化形态会趋向正态分布，成为误差分析与统计推断等领域的基石。
- **实验与方法论影响**：这一结论告诉我们：在同一分布下，**无穷次重复**会显现“显著稳定性”，为近似确定性的结论提供统计保证，这正是科学实验与频率统计学中“多次测量”的根本依据。

通过大数定律与中心极限定律，我们能更加深刻地认识“随机性如何在大量重复下产生规律性”这一现代统计学精髓，也因此在数理统计、数据分析乃至科学实验中广泛应用。

（二）概率模型的建立与选择

1 假设检验

在统计学研究中，我们经常需要基于样本信息来对总体的某个参数或分布形式进行判断。为此，我们提出若干关于总体的假设，然后利用相应的

检验方法来判定这些假设是否有足够的依据被拒绝或暂时无法拒绝。这样的流程就是**假设检验（Hypothesis Testing）**。

1.1 基本概念：零假设与备选假设

在假设检验的框架中，我们通常从两个对立的假设出发：

- **零假设（Null Hypothesis, H_0）**：这是我们最初的或传统上被接受的关于总体参数或分布的假设，常常表示“不存在差异”或“不发生某种效应”等中性假设；
- **备选假设（Alternative Hypothesis, H_1 或 H_a）**：这是与零假设对立的观点，意味着“存在差异”或“发生某种效应”等。

我们通过假设检验的过程来判断是否**在数据的支持下“拒绝”零假设**。一旦拒绝零假设，也就等价于接受了备选假设。值得注意的是：**未拒绝零假设并不意味着它就一定成立**，而是现有数据的证据还不足以推翻它。

1.2 假设检验的核心思想

设想存在一个包含所有可能样本的样本空间 X。对零假设 H_0 和备选假设 H_1，我们通过一个**检验规则**或**检验统计量**来决定是否拒绝零假设。这种检验规则可形式化为：给定样本 $x \in X$，若 x 落在某个“拒绝域”中，则拒绝 H_0；否则保留（或“暂不拒绝”）H_0。

在数学形式上，一个**假设检验**可以写作 $T = (X_0, X_1)$，其中 X_0 为**接受域**（或“不拒绝域”），X_1 为**拒绝域**。这两个集合将整个样本空间 X 划分为两个部分，且 $X_0 \cup X_1 = X$、$X_0 \cap X_1 = \emptyset$。

1.3 拒绝域与接受域

- **拒绝域**（Reject Region）：若观测到的样本点 $x \in X_1$，则我们有理由认为 H_0 不成立，从而拒绝 H_0。
- **接受域**（Accept Region）：若观测到的样本点 $x \in X_0$，则我们暂时无法拒绝 H_0。需要强调的是，这并不代表“H_0 就是真的”，而只意味着“我们没有足够的样本证据推翻 H_0”。

1.4 显著性检验

在选定拒绝域时，我们希望保证一定的**显著性水平**（significance level），以控制我们做出错误决策的概率。为此，需要先了解两种常见的错误形式，以及功效函数的概念。

1.4.1 第一类错误与第二类错误

- **第一类错误（Type I error）**：当 H_0 实际上为真，但依据观测到的样本却将其拒绝。
- **第二类错误（Type II error）**：当 H_0 实际上是假的，然而我们并没有拒绝它，也就是把一个错误的零假设当作“还不能被推翻”。

一般地，用 α 表示第一类错误的概率上限，并将其称为检验的显著性水平 (significance level)。第二类错误的概率可记作 β，对应该备选假设成立时却不拒绝零假设的情况。

1.4.2 功效函数（Power Function, 检验功效）

检验功效描述了在不同的真实参数条件下，检验拒绝 H_0（即识别出 H_0 为假）或“揭示假设真相”的概率。形式上，对于一个检验 $T = (X_0, X_1)$ 和总体参数 θ，我们定义：

$$\Phi_T(\theta) \;=\; P_\theta\big(X \in X_1\big),$$

即当真正的参数是 θ 时，样本落入拒绝域的概率。若 θ 属于备选假设所指定的参数空间，我们希望 $\Phi_T(\theta)$ 尽量大（检验有较高“查出错误 H_0”的能力）。

1.4.3 α-显著性检验

假设给定一个显著性水平 $\alpha \in (0, 1)$，我们要求检验的第一类错误概率不超过 α。也就是对所有满足 H_0 的参数取值 θ_0，应有

$$P_{\theta_0}(X \in X_1) \;\leq\; \alpha.$$

满足上述条件的检验称为 α-**显著性检验**或 α-**检验**。这意味着，在零假设真实成立时，我们将它错误地拒绝的概率被严格控制在 α 之内。

需要特别注意：当我们设计检验时往往倾向于**严控第一类错误**的概率（即 α）在一个较小水平上，例如 0.05 或 0.01 等。但这也常常导致第二类错误的概率（β）可能较大。因此，“接受 H_0”并不意味着它就肯定正确，更多情况下只表示“我们没有足够证据让我们在 α 的显著性水平上拒绝它”。

1.4.4 一致最有力 α 检验

在给定的 α 显著性水平下，不同的检验（不同的 (X_0, X_1) 划分方式）会有不同的功效函数 $\Phi_T(\theta)$。若某一检验 T_1 的功效函数在备选假设覆盖的整个参数空间上均**不小于**另一检验 T_2，我们就说 T_1 比 T_2 更有力。若存在一个检验 T^*，对所有 α-检验来说都没有比它更高的功效函数，则称 T^* 为**（一致）最有力 α 检验**。在简单假设检验中，著名的 **Neyman-Pearson 引理**给出了寻找最有力检验的一般方法。

1.5 常见的拒绝域构造

我们通常会根据检验统计量的分布特征来构造拒绝域，并对应地给出检验的临界值。以下是几种常见的情形。

1.5.1 One-Sided Test 与 Two-Sided Test

根据备选假设 H_1 对参数的可能偏离方向不同，我们会有：

- **单边检验（One-Sided Test）**：例如 $H_0 : \mu = \mu_0$ 对 $H_1 : \mu > \mu_0$，或 $H_1 : \mu < \mu_0$。此时拒绝域多半选择在分布的一侧，如“观察到的统计量显著大于某临界值”。
- **双边检验（Two-Sided Test）**：例如 $H_0 : \mu = \mu_0$ 对 $H_1 : \mu \neq \mu_0$。此时拒绝域往往在分布的两端，如“观察到的统计量在过大或过小时拒绝”。

1.5.2 对二项分布的检验

若观测数据服从**二项分布** $\mathrm{Bin}(n, p)$，假设 $H_0 : p = p_0$ 与 $H_1 : p > p_0$（单边），我们可以考虑统计量 $X \sim \mathrm{Bin}(n, p)$ 自身。设定一个临界值 k，若

$X \geq k$, 则拒绝 H_0。此时 k 由控制 $P_{p_0}(X \geq k) \leq \alpha$ 而确定。双边检验则考虑同时控制 X 的偏大与偏小情况，建立相应的拒绝域 $\{X \leq k_1\} \cup \{X \geq k_2\}$。

1.5.3 对正态分布的检验与 Fisher's Theorem

若观测的数据 $X_1, X_2, \ldots, X_n$ 来自正态分布 $N(\mu, \sigma^2)$，我们关心 μ 或 σ^2 的假设检验时，会用到一些经典的**正态分布诱导的统计量**（如样本均值 $\bar{X}$、样本方差 S^2 等）。之所以能采用这些统计量并得到 t 分布、χ^2 分布、F 分布等，背后关键的结果正是 **Fisher 定理（Fisher's Theorem）**和中心极限定理等深入理论。

Fisher's Theorem（简述） 正态样本具有良好的可分解性，当我们在已知或未知方差的不同情况下，对比均值或方差的假设时，能构造出满足 t、χ^2 或 F 等分布的检验统计量。如：

- $\bar{X} \sim N(\mu, \sigma^2/n)$;
- 若 σ^2 未知，则 $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ 服从自由度为 $n-1$ 的 t 分布;
- 若已知 μ，则 $Q = \frac{(n-1)S^2}{\sigma^2}$ 服从自由度为 $n-1$ 的 χ^2 分布;
- 以及 F 分布的构造常用于比较两个方差是否相等。

1.5.4 分位数 (Quantiles)

为了确定拒绝域的边界，需要用到**分位数**概念。分位数是指分布函数 F 的某个反函数，如 α 分位数满足 $F(q_\alpha) = \alpha$。检验中若需要控制 $P_{H_0}(X \geq c_\alpha) = \alpha$，则 c_α 就是分布相应的 $(1-\alpha)$ 分位数。单边与双边的拒绝域都与分位数直接关联。

1.5.5 常见的四种检验：z 检验, χ^2 检验, t 检验, F 检验

- z **检验**：当方差已知且样本量足够大或总体正态时，对总体均值的假设可使用 z 检验，检验统计量多为 $Z = \dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$。
- χ^2 **检验**：用来对方差进行假设检验，或用于适度性检验（如卡方拟合优度检验）。如 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$。

- t **检验**：在方差未知的正态总体下，对均值进行检验。$T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$。
- F **检验**：常用于比较两个方差是否相等或方差分析中检验不同组均值是否存在差异。若 $U \sim \chi^2_{d_1}$，$V \sim \chi^2_{d_2}$ 且相互独立，则 $\frac{U/d_1}{V/d_2} \sim F_{(d_1,d_2)}$。

对“未拒绝 H_0”结果的进一步思考

最后必须再次强调：**不拒绝 H_0 不等于“接受 H_0 是对的”**。我们往往把第一类错误控制到 α，这意味着在检验设计时，如果 H_0 真实成立，则拒绝它的概率**至多**是 α。但是另一方面，这种保守设计也会导致第二类错误概率（错失拒绝假设的机会）被动地变大。当我们无法拒绝 H_0 时，只能说“以当前数据和所选取的显著性水平，还没有足够证据推翻 H_0”。为此，研究者如果对 H_0 的正确与否依然存在疑问，就需要提高样本量、或降低 α 以减小第一类错误率、或思考如何在实验设计中提升检验的功效（如减少样本变异、增加显著效应等）。

综上所述，假设检验是统计推断的核心方法之一。在了解零假设与备选假设、拒绝域、显著性水平、第一类与第二类错误以及检验功效等关键概念后，就能更深入地理解为何要从概率意义上谨慎地做“拒绝”或“暂不拒绝”的结论，并以合适的方式构造不同分布背景下的检验统计量及其拒绝域。

2 最大似然估计 (Maximum Likelihood Estimation, MLE)

在统计推断中，最基本的问题之一就是：**如何用样本信息给出未知参数的“最好”估计？**最大似然估计（MLE）是这一问题的一种极具影响力的解法。为了更好地理解 MLE 的来龙去脉，我们先简要回顾一下“点估计”这一更基础的概念。

2.1 点估计 (Point Estimation)

- 在频率学派（或称古典学派）的统计框架下，**参数是固定未知值**，而我们的样本是由参数所决定的随机过程产生的。

- **点估计**（Point Estimation）则是指：在获得样本后，通过某种规则或函数，输出一个**单值**去估计这个未知的参数。例如，估计总体均值 μ 时，最常见的点估计量就是样本均值 $\bar{X}$。
- 虽然点估计提供了一个简明的参数数值，但却**无法直接反映估计不确定性的大小**；因此，在本节之后还会介绍另一个可以量化“不确定度”的方法——置信区间。

2.2 似然与似然函数

- 设有样本 $X = (X_1, X_2, \ldots, X_n)$，这些样本来自一个分布函数依赖于未知参数 θ 的总体。直观上说，“数据发生的可能性”取决于 θ 。
- 在频率学派框架中，我们把样本 X 看作**已知**，而参数 θ 视为**自变量**。于是，为了衡量“给定 θ 下样本出现的可能性”，我们定义：

$$L(\theta; X) \;=\; P(X \,|\, \theta).$$

- 若假设 X_i 独立同分布，且其概率密度（或质量）函数为 $f(x|\theta)$，那么

$$L(\theta; X) = \prod_{i=1}^{n} f(X_i|\theta).$$

- **关键点**：似然函数 $L(\theta; X)$ 不是对 θ 的概率分布，而是度量“样本已出现时，参数 θ 的取值有多适宜”。

2.3 最大似然估计：核心思想

- **最大似然估计**（MLE）试图在所有可能的 θ 取值中，找到那个使得 $L(\theta; X)$ “最大化”的 θ 值，即

$$\hat{\theta} \;=\; \arg\max_{\theta} \; L(\theta; X).$$

- 通常，我们更倾向于最大化**对数似然函数** $\ell(\theta; X) = \ln L(\theta; X)$，因为对数运算能把乘积变为加和，简化求导运算：

$$\hat{\theta} \;=\; \arg\max_{\theta} \; \ell(\theta; X).$$

- 在实践中，当我们对 $\ell(\theta; X)$ 求导并令其等于 0 时，往往可以得到封闭形式解（如正态分布、二项分布、泊松分布），或利用数值方法求解（如牛顿迭代、梯度上升等）。

2.4 性质与应用

- **一致性 (Consistency)**：在一定正则条件下，当样本量 $n \to \infty$ 时，最大似然估计 $\hat{\theta}$ 以概率 1 收敛到真实的参数值 θ。
- **渐近正态性 (Asymptotic Normality)**：MLE 在大样本下近似服从以真实参数为中心的正态分布；其方差可用 Fisher 信息量的倒数刻画。
- **有效性 (Efficiency)**：在满足正则条件时，MLE 在渐近意义下能达到 Cramér-Rao 下界，即具有最优的“大样本”估计效率。
- **应用场景**：MLE 在各种常见分布（例如正态、二项、泊松）中有广泛使用；即便在更复杂或高维场景下，MLE 依然是极为重要的估计思想，对机器学习、计量经济学等领域都有重要启示。

对描述的思考

需要注意的是，“似然 (likelihood)”常被混淆为“概率 (probability)”，但两者含义并不完全相同。概率是“给定参数后，样本发生的可能性”，而似然是“给定样本后，哪种参数值最能解释这些数据”。在贝叶斯范式下，似然与先验相结合得到后验，但在频率学派中，MLE 仅寻求能最大化似然的参数点值。

3 无偏估计 (Unbiased Estimation)

当我们用一个估计量 $\hat{\theta}$ 去估计未知参数 θ 时，若该估计量的**数学期望**恰好等于真实参数值，即

$$E[\hat{\theta}] = \theta,$$

则称其为**无偏估计量**。无偏估计强调“平均意义”上的准确性，即它不会在总体上系统性地高估或低估参数。

3.1 核心概念

- 若一个估计量并非无偏，则存在一定的**偏差**：

$$\text{Bias}(\hat{\theta}) \;=\; E[\hat{\theta}] - \theta.$$

- 无偏性是评价估计量品质的一项重要指标，但并不涵盖其方差、稳健性等其他方面的性能。一个估计量可以是无偏的，但方差很大；也可以是带有些许偏差，但整体均方误差（MSE）反而更小。

3.2 经典示例：样本均值与样本方差

- **样本均值**：$\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$。这是总体均值 μ 的无偏估计量，因为 $E[\bar{X}] = \mu$。
- **样本方差**：$S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$。这是总体方差 σ^2 的无偏估计量，即 $E[S^2] = \sigma^2$。
- 若改用 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 而非 $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$，则期望值小于 σ^2，会**系统性地低估**总体方差，从而不再是无偏。

3.3 构造无偏估计量的思路

- 通常可以通过修正系数的方法来消除系统性偏差，从而把估计量转变为无偏的形式。这种修正往往涉及到对样本期望的分析。
- 需要注意的是，在实务中并不总是只追求无偏。有时一个**带有偏差但方差更小**的估计量，在均方误差 (MSE) 角度看，反而更优。这便体现了**方差-偏差权衡**的思想。

3.4 进一步思考

- 无偏只是衡量估计量优劣的一个维度；实际应用中，往往还要结合方差、稳健性以及对先验信息的利用等因素来做整体考量。
- **Gauss-Markov 定理**告诉我们：在线性回归模型中，若满足线性、同方差、无自相关等假设条件，则普通最小二乘法 (OLS) 得到的估计量

是**最佳线性无偏估计** (BLUE)。但这并不意味着 OLS 在所有意义下都是最优，也不一定意味着它能带来最佳预测效果（遇到高维复杂情形时，正则化等技术往往必不可少）。

4 置信区间 (Confidence Interval)

在统计推断中，仅给出一个“点估计”往往不足以回答“这个估计的可靠程度如何”这样的问题。为此，我们需要**区间估计**来对未知参数给出一个可能的范围，并附带一个置信水平说明“在多大程度上我们相信参数会处于这个范围内”。这便是**置信区间**。

4.1 置信空间与置信区间

- **置信空间**是从样本出发，利用参数与数据之间的关系，寻找满足一定“覆盖真值概率不小于 $1-\alpha$”要求的参数集合；在实际应用中，这个集合通常是一个上下界明确的区间，因而称为**置信区间 (Confidence Interval)**。
- **频率学派的解释**：如果我们在同样条件下重复实验很多次，每次都用相同的方法构造相应的区间，那么在这些区间中，大约有 $100(1-\alpha)\%$ 会包含真正的参数。注意，这并不意味着“给定某一个区间，它包含真值的概率是 $100(1-\alpha)\%$”，因为在频率学派下，参数本身并不视为随机；真正随机的是区间在一次次抽样下的变化。

4.2 置信水平与显著性水平

- **置信水平**是区间估计中最常用的指标，常取 95%、99% 等。“高置信水平”意味着我们要求在重复抽样的过程中更高比例的区间能覆盖真值，但这往往会导致区间变宽。
- 显著性水平 α 是与 $1-\alpha$ 相对的概念，表示“不覆盖真值”的概率上限。“5% 显著性水平”对应于 95% 的置信度。

4.3 常见置信区间构造

- **总体均值的置信区间**
  - 当方差 σ^2 已知，且样本量 n 较大或来自正态分布时：

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

    其中 $z_{\alpha/2}$ 是标准正态分布的临界值（如置信水平 95% 时，$z_{0.025} \approx 1.96$）。
  - 当方差未知、且样本量不大时：

$$\bar{X} \pm t_{\alpha/2,\, n-1} \frac{S}{\sqrt{n}},$$

    其中 $t_{\alpha/2,\, n-1}$ 是 t 分布的分位数，S 是样本标准差。
- **总体比例的置信区间** - 若我们观察的是二分类结果，对应的“总体比例”可表示为 p。当样本量足够大时，以正态近似为基础，有

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

  其中 $\hat{p}$ 是样本比例。

4.4 直观理解与实践意义

- 在大部分应用中，如果我们构造了 95% 的置信区间，“这并不代表参数有 95% 的概率落在区间里”，而是说：如果我们不断重复采样和构造区间，那么其中约 95% 的区间能覆盖真实参数。
- 在工程、医学等很多场合，我们往往会希望置信区间**既要准确度高**（即保持合适的置信水平），**又要尽量窄**（便于决策）。因此实际中存在对样本量、分布假设、风险衡量等方面的综合权衡。

5 贝叶斯统计方法：整体性视角与核心内涵

我们前面讲的四种推断方法，更多是“频率学派”视角之下，对于概率统计模型参数的选择；然而，与之相对的，“贝叶斯统计”的方法，同样可

以用来根据样本选择统计模型参数。直觉上，可以把贝叶斯方法理解成一种“稳健的似然估计”，下面会展开细节对此给予解释。

> **概要：** 贝叶斯统计将参数视为随机变量，采样数据作为“证据”来修正我们对参数的不确定性。这种方法与频率学派在思想与操作上均存在差异：一方面，贝叶斯通过先验（prior）与似然（likelihood）的组合，构造后验（posterior）；另一方面，在解释层面，贝叶斯可以直接谈论参数“落在某区间的概率”，而频率学派则更关注“若重复抽样，区间涵盖真实参数的比例”。下文将结合对比与反思，给出一个整体性的贝叶斯方法导论，并探讨其**平滑更新**与**稳健性**背后的逻辑。

5.1 贝叶斯统计的基本概念

1）参数是随机量：从先验到后验

贝叶斯的关键假设在于：**参数**（记为 θ）**本身具有不确定性，可以用概率分布刻画**。我们在分析前就先为 θ 设定一个分布，称之为**先验分布** $p(\theta)$，用来表达在未见到数据之前，我们对参数的可能范围或偏好。

而一旦我们观测到数据 x，就通过贝叶斯公式将先验更新为**后验分布** $p(\theta \mid x)$。具体地说，

$$p(\theta \mid x) \;=\; \frac{p(x \mid \theta)\, p(\theta)}{p(x)},$$

其中：

- $p(\theta)$——**先验**，表示在无数据或少数据时我们对 θ 的“初始信念”；
- $p(x \mid \theta)$——**似然函数**，衡量在给定 θ 下观测到数据 x 的可能性；
- $p(x)$——**证据**或边际似然，是一个归一化常数，保证后验分布积分为 1；它通过积分 $\int p(x \mid \theta)\, p(\theta)\, d\theta$ 来得到。

这就是“**先验** $\times$ **似然** $\Rightarrow$ **后验**”的核心流程。

2) 数据是已知（固定的），不确定性在于参数

与频率学派把“参数固定、数据随机”相反，贝叶斯统计将数据 x 视为已观测到的一次性事实，θ 才是我们不确定的“对象”。因此，在贝叶斯推断中，当数据 x 到手之后，它就不再被重复看待；真正需要考量的是“参数 θ 该如何分布并更新”，从而推导对 θ 的各种概率性描述（如**后验均值**、**后验方差**、**后验区间**等）。

5.2 为何要对 θ 进行积分或求和来获得 $p(x)$？

有时有人会疑惑：为什么我们要对 θ 进行积分，$p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta$？**根本原因**在于，贝叶斯方法强调“$\theta$ 也是随机的”。当我们要获取“观察到 x 的整体概率”时，自然要把**所有可能的 θ 情形**都考虑，并对其先验权重 $p(\theta)$ 进行加权。

$$p(x) \;=\; \int p(x \mid \theta)\, p(\theta)\, d\theta,$$

从而为后验分布的分母提供了**归一化因子**。这保证了在贝叶斯视角里，“θ 的不确定性”被完整地纳入概率运算之中。

5.3 贝叶斯方法的“平滑”更新与稳健性

1) 平滑过程：从先验到后验

与单纯最大似然方法（MLE）“一次性地选择最优点”不同，贝叶斯方法会将先验分布与似然函数**相乘**并用**积分**进行归一化，从而得到一个“更加平滑、连续的修正过程”。

- 当数据量小或噪声大时，**先验**可以阻止估计结果出现极端或不合理的波动；
- 当数据量足够大时，**似然**占主导地位，先验影响相对减弱——这与经典频率学派结果在大样本时往往相吻合。

2) 对异常值更稳健

如果数据中偶然出现离群点或“怪异”样本，纯粹的似然方法（比如 MLE）可能会被极端值严重影响；而贝叶斯方法因为有先验分布做“拉回”

或“约束”，使得估计不至于被少数离群点牵得过远，尤其是当先验蕴含了对参数有限制或正则化的倾向时。

5.4 贝叶斯 vs. 频率学派的对比

1) 参数与数据关注点的区别

- **频率学派**：参数 θ 固定但未知，数据是可以（概念上）反复抽样的随机量。讨论如置信区间之类，强调“若重复抽样，会有多少比例的区间包含真的 θ”。
- **贝叶斯学派**: 数据 x 给定，θ 不确定，是**随机变量**。我们直接谈论“θ 的分布在何区间的概率是多少”。“置信区间”在此就成了**“后验区间”**（或称**可信区间、Bayesian credible interval**），其含义是“在目前信息下，θ 有 95% 的后验概率落于该区间”。

2) 数学形式常有交集，解释视角不同

在大样本极限或在某些先验的选择下（如无信息先验），贝叶斯后验结果往往与频率学派的点估计和区间接近。但**两者的解释框架差别仍旧很大**：

- 频率学派注重“抽样分布”及“重复试验”；
- 贝叶斯注重“后验分布”及“先验如何被数据修正”。

5.5 形式化定义：贝叶斯方法的流程 (简要)

1. **先验建模**：给出 $p(\theta)$，表达对参数 θ 初始的主观或客观判断。
2. **构造似然**：确定 $p(x \mid \theta)$（即数据分布在给定 θ 下的形式）。
3. **更新至后验**：贝叶斯定理

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}, \quad \text{其中} p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta.$$

4. **做推断或预测**：从后验分布出发，可做
   - 后验均值/众数/中位数等点估计，
   - 可信区间 (Credible Interval)，
   - 后验预测分布 $p(x_{\text{new}} \mid x)$ 等。

5.6 思考与总结

对比与深层理解：

- **贝叶斯方法**可被视为“在似然推断的基础上加了一层对参数的‘随机化’处理”，并通过数据**修正**（更新）对参数的信念；在**小样本**或**离群数据**多时展现出更稳健表现。
- 为什么要积分出 $p(x) = \int p(x|\theta)p(\theta)\,d\theta$?

  因为贝叶斯将 θ 当作随机量，需要对所有 θ 可能性加权整合，来保证后验分布是完整且可归一的概率模型。
- **“更关心条件概率 $p(\theta \mid x)$” VS. “分析联合分布 $p(x, \theta)$”**：

  在频率学派视角下，我们更关注“若无穷次重复采样，x 出现的分布如何”；而贝叶斯则更关注“给定实际发生的 x，θ 的概率分布如何变化”。
- 无论如何，贝叶斯并不是简单的“置信区间 + 似然”的混合，而是**将参数本身置于概率空间之中**，强调“先验—后验”的核心思想。它所构造出的“平滑更新”机制，对一些极端或复杂情形尤为有用，也使我们对参数有直观的概率陈述。

总之，贝叶斯统计的精髓就在“把参数当作随机量，对其构建先验，并用数据来修正或更新”的思路上；而那句“观测数据是用来评估 θ 的可能性”也就顺理成章地诠释了“我们为什么要把 $p(\theta \mid x)$ 作为最终目标”，以及为什么要将 $p(x) = \int p(x \mid \theta)p(\theta)\,d\theta$ 视作一个必不可少的归一化过程。