Lecture Notes

2023 秋　概率论与数理统计 A

Author
AUGPath
China University of Geosciences (Wuhan)
December 17, 2023

Contents

我们假定读者已经学习或正在学习《高等数学》、《离散数学》、《算法导论》的课程. 其中, 学习并深入理解高等数学中的概念是至关重要的.

Part I

概率论中基本的概念

1 实验. 事件. 概率空间简介

1.1 随机实验

考虑某项试验, 其结果在某一组条件下会有若干种不同的结局 (现象) $\omega_1, \cdots, \omega_N$. 关于这些结局具体是什么并不重要, 我们其实要把他们抽象为一组点. 我们把这些结局 $\omega_1, \cdots, \omega_N$ 称做基本事件, 而把一切结局的全体

$$\Omega = \{\omega_1, \cdots, \omega_N\}$$

称做基本的事件空间, 或样本空间.

例子 1.1. *对于“掷一枚硬币”, 基本事件空间由两个点组成:*

$$\Omega = \{\mathrm{Z}, \mathrm{F}\},$$

其中 Z 表示出现“正面”, 而 F 表示出现“反面”. (这时假设, 诸如“硬币在棱上立着”, “硬币丢失”,$\cdots\cdots$ 的情况不会出现.) 也就是假设不出现“正面”就出现“反面”.

将一枚硬币重复掷 n 次, 基本事件空间为

$$\Omega = \{\omega : \omega = (a_1, \cdots, a_n)\}, a_i = \ Z \text{ 或} \mathrm{F},$$

且基本事件的总数 $N(\Omega) = 2^n$.

除基本事件空间的概念外, 现在引进重要概念: 事件. 事件的概念, 是建立所考察试验的各种概率模型 (“理论”) 的基础. 在试验的结果中, 试验者一般并不关心究竟出现了哪种具体的结局, 而关心出现的结局属于一切结局集合的哪个子集. 满足试验条件的一切子集 $A \subseteq \Omega$, 分为两种类型: “结局 $\omega \in A$” 或 “结局 $\omega \notin A$ ”. 我们称这样的子集 A 为事件.

例子 1.2. *将一枚硬币重复掷三次, 一切可能结局的空间 Ω, 由 8 个点构成:*

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\}$$

其中 0 和 1 分别表示郑出“正面”和“反面”. 如果由“一组条件”可以记录 (确定、

测量等) 所有 3 次郑硬币的结果, 则例如

$$A = \{000, 001, 010, 100\}$$

就是事件: "将一枚硬币重复掷三次" 正面至少出现两次. 假如由 "一组条件" 只能确定第一次郑出的结果, 则 A 已经不能称为事件, 因为关于 "试验的具体结局 ω 是否属于 A", 既不能肯定也不能否定.

在随机试验中, 当事件中的一个样本点出现时, 称该事件发生.

1.2 事件的关系与运算

定义 1.1 (事件的关系)**.** 若事件 A 发生时, 事件 B 一定发生. 则称事件 A 包含于事件 B（或事件 B **包含** A）, 记作

$$A \subset B \text{ (或} B \supset A)$$

对任意事件 A, 有 $\emptyset \subset A \subset \Omega$.
若 $A \subset B$, 且 $B \subset A$, 则称事件 A 与 B **相等**, 记作 $A = B$.

下面来考察事件的运算:

定义 1.2 (事件的并)**.** "事件 A、B 至少有一个发生" 称为事件 A 与 B 的**和**或**并** (union), 记作

$$A \cup B \text{ (或} A + B)$$

也就是

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$$

注: 使用数学归纳法, 事件的并可以推广到多个的情形: 如 n 个事件的并

$$\bigcup_{i=1}^{n} A_i := \text{"事件 } A_1, \cdots, A_n \text{ 至少有一个发生"}$$

可数个事件的并

$$\bigcup_{i=1}^{\infty} A_i := \text{"事件 } A_1, A_2, \cdots \text{ 至少有一个发生"}$$

定义 1.3. "事件 A、B 同时发生" 称为事件 A 与 B 的**积**或**交** (intersection), 记作

$$AB \text{ (或} A \cap B)$$

也就是

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}$$

注: 事件的交可以推广到多个的情形：如 n 个事件的交

$$\bigcap_{i=1}^{n} A_i := \text{“事件 } A_1, \cdots, A_n \text{ 全都发生”}$$

可数个事件的交

$$\bigcap_{i=1}^{\infty} A_i := \text{“事件 } A_1, A_2, \cdots \text{ 全都发生”}$$

定义 1.4. “事件 A 发生, 但 B 不发生” 称为事件 A 与 B 的**差**, 记作

$$A - B$$

也就是

$$A - B := \{\omega : \omega \in A \text{ and } \omega \notin B\}$$

像集合那样, 我们同样可以引入事件的关系:

定义 1.5. 若 $AB = \emptyset$, 则称 A 与 B **互斥**（或称 A 与 B **不相容**）, 即 A 与 B 不可能同时发生.

定义 1.6. 称 $\Omega - A$ 为事件 A 的**对立**事件（或称 A 的**补**）, 记为 $\overline{A}$. 它表示 “事件 A 不发生”.

像集合那样, 事件具有如下的运算规律:

- 交换律
  - $AB = BA$, $A \cup B = B \cup A$
- 结合律
  - $(AB)C = A(BC)$, $(A \cup B) \cup C = A \cup (B \cup C)$
- 分配律
  - $A(B \cup C) = AB \cup AC$, $A(B - C) = AB - AC$
- 对偶律
  - $\overline{AB} = \overline{A} \cup \overline{B}$, $\overline{A \cup B} = \overline{A}\,\overline{B}$

除了在《离散数学》课程中我们学习过较为严格的证明之外, 这些规律可以从集合较为基础的表示法 – Venn 图中看出.

下面来考察一些常见的化简运算关系的等式:

命题 1.7. 对任意两个事件 A 和 B, 总有 $A-B=A-AB$.

命题 1.8. 事件 A、B **对立**当且仅当 A、B **互斥**且 $A\cup B=\Omega$.

例子 1.3. *设 A,B 为两个事件, 则有*

- $A\overline{B}=A-B=A-AB$;
- $A=AB\cup A\overline{B}$.

解答: 用事件运算的分配律:

- $A\overline{B}=A(\Omega-B)=A\Omega-AB=A-AB$;
- $AB\cup A\overline{B}=A(B\cup\overline{B})=A\Omega=A$.

例子 1.4. *A, B, C 表示事件*

- *A 发生: A;*
- *仅 A 发生: $A\cap\bar{B}\cap\bar{C}$;*
- *恰有一个发生: $A\bar{B}\bar{C}\cup\bar{A}B\bar{C}\cup\bar{A}\bar{B}C$;*
- *至少有一个发生: $A\cup B\cup C$;*
- *至多有一个发生: $\bar{A}\bar{B}\bar{C}\cup A\bar{B}\bar{C}\cup\bar{A}B\bar{C}\cup\bar{A}\bar{B}C$;*
- *都不发生: $\bar{A}\bar{B}\bar{C}$;*
- *不全部发生: $\overline{ABC}=\bar{A}\cup\bar{B}\cup\bar{C}$.*

Takeaway Message

可以使用集合描述事件, 离散数学中学过的集合的运算 (或者直观) 将允许我们对于事件进行化简和操作.

2 事件的概率

事件的**概率**: 刻画试验中随机事件发生的**可能性大小**.

2.1 概率的统计定义

定义: 设在 n 次试验中, 事件 A 发生了 m 次, 则称

$$f_n(A):=\frac{m}{n}$$

为事件 A 发生的**频率** (frequency).

定义: 在相同条件下重复进行的试验中, 若随着试验次数 n 的增加, 事件 A 发生的频率稳定在某一常数 p 附近, 则称 p 为事件 A 的**概率**, 记作 $P(A)=p$.

也就是概率是频率的稳定值. 实际应用中常将大量重复试验中事件的频率作为概率的近似估计.

性质: 频率的性质:

- $0 \le f_n(A) \le 1$;
- $f_n(\Omega)=1,\ f_n(\emptyset)=0$;
- 若事件 $A_1, A_2, \cdots, A_k$ 两两互斥, 则

$$f_n\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} f_n(A_i)$$

由于上述的性质, 我们给出概率的数学公理化定义:

定义 2.1 (概率的公理化定义)**.** 设 Ω 是样本空间, 定义概率空间 $(\Omega, \mathcal{F}, P)$. 对每个事件 $A \in \mathcal{F}$ 定义一个实数 $P(A)$ 与之对应. 集合函数 P 满足以下条件:

- 非负性: 对任意事件 A, 均有 $P(A) \ge 0$;
- 规范性: $P(\Omega)=1$;
- 可加性: 若事件序列 $\{A_n\}_{n\ge 1}$ 两两互斥, 则

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

则称 $P(A)$ 为事件 A 的**概率** (probability).

这里的事件用集合的语言描述, 考虑集合 $A \subseteq \Omega$ 的某个集系 $\mathscr{A}_0$, 则利用集合运算 $\cup, \cap$ 与 $\setminus$ 可以由 $\mathscr{A}_0$ 构造新集系, 其中元素也是事件. 给这些事件补充上必然事件 Ω 和不可能事件 $\varnothing$, 得集系 $\mathscr{A}$, 则 $\mathcal{A}$ 是代数. 所谓 "代数" 即 Ω 的这样的集系, 满足

1) $\Omega \in \mathcal{F}$,
2) 若 $A \in \mathcal{F}, B \in \mathcal{F}$, 则集合 $A \cup B, A \cap B, A \backslash B$ 也都属于 $\mathcal{F}$.

例如这些内容

事件可以先简单认为就是 Ω 一堆子集构成的集合, 当然有一些条件需要满足. 对于初学者而言, 下面的就先不用看了. 这些内容只是为了那些学习过离散数学并且知道这种情形的作用的同学准备的.

例子 2.1. *a)* $\mathscr{A} = \{\Omega, \varnothing\}$ *集系由* Ω *和空集* $\varnothing$ *构成, 称做平凡代数;*

b) $\mathscr{A} = \{A, \bar{A}, \Omega, \varnothing\}$ *事件* A *产生的集系;*

c) $\mathscr{A} = \{A : A \subseteq \Omega\}$ Ω *全部子集的集系 (包括空集* $\varnothing$ *).*

这些事件代数可以按分割的方式得到: 我们称集系

$$\mathscr{D} = \{D_1, \cdots, D_n\}$$

构成集合 Ω 的一个分割, 而 $D_1, \cdots, D_n$ 是该分割的原子, 如果 $D_1, \cdots, D_n$ 非空且

两两不相容, 而它们的和等于 Ω :

$$D_1 + \cdots + D_n = \Omega.$$

例如, 假定集合 Ω 由 3 个点构成: $\Omega = \{1, 2, 3\}$, 则存在 5 个不同的分割:

$$\begin{array}{ll} \mathscr{D}_1 = \{D_1\} & D_1 = \{1, 2, 3\} \\ \mathscr{D}_2 = \{D_1, D_2\} & D_1 = \{1, 2\}, D_2 = \{3\} \\ \mathscr{D}_3 = \{D_1, D_2\} & D_1 = \{1, 3\}, D_2 = \{2\} \\ \mathscr{D}_4 = \{D_1, D_2\} & D_1 = \{2, 3\}, D_2 = \{1\} \\ \mathscr{D}_5 = \{D_1, D_2, D_3\} & D_1 = \{1\}, D_2 = \{2\}; D_3 = \{3\}. \end{array}$$

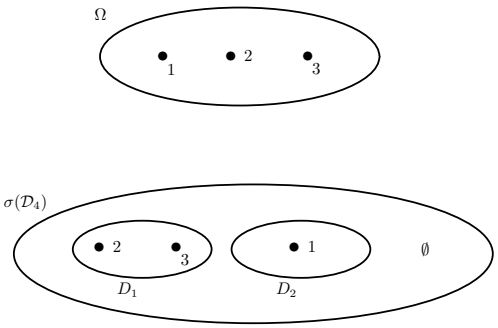

Figure 1: 集合代数

如果考虑 $\mathscr{D}$ 中一切集合的并连同空集 $\varnothing$, 则得到的集系是代数, 称做 $\mathscr{D}$ 产生的代数, 记作 $\sigma(\mathscr{D})$. 于是, 代数 $\sigma(\mathscr{D})$ 的元素由空集 $\varnothing$ 与分割 $\mathscr{D}$ 之原子中集合的和组成. 这样, 如果 $\mathscr{D}$ 是 Ω 的某一分割, 则它与代数 $\mathscr{B} = \sigma(\mathscr{D})$ 一一对应. 如图 1.

逆命题也正确. $\mathscr{B}$ 是有限空间 Ω 的子集的代数, 则存在唯一分割 $\mathscr{D}$, 其原子是代数 $\mathscr{B}$ 的元素, 并且 $\mathscr{B} = \sigma(\mathscr{D})$. 事实上, 假设集合 $\mathscr{D} \in \mathscr{B}$ 并且具有性质: 对于任意 $B \in \mathscr{B}$, 集合 $D \cap B$ 要么与 D 重合, 要么是空集. 那么, 这样集合 D 的全体组成分割 $\mathscr{D}$ 并且具有所要求的性质 $B = \sigma(\mathscr{D})$.

Takeaway Message

概率是在样本空间上面定义的一个函数, 满足:

- (1) 非负性: 每个事件的概率必须大于等于 0;
- (2) 规范性所有的事件概率 “总和” 等于 1;
- (3) 可加性: 互斥事件的概率可以直接相加.

2.2 概率的加法公式

我们已经在互斥的时候, 规定了其概率的计算方法. 也就是公理 (3). 特别地, 当 $n = 2$ 的时候, 就有:

命题 2.2 (加法公式)**.** 若两个事件 A, B 互斥, 则

$$P(A \cup B) = P(A) + P(B).$$

下面我们来看一看不互斥的情形.

注: 由加法公式可得到如下性质:

- 对任意事件 A, 有 $P(A) = 1 - P\left(\overline{A}\right)$.
- 对任意两个事件 A, B, 有

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

Aside

我们实际上可以借此瞥见容斥原理.

注: 若三个事件 A_1, A_2, A_3 两两互斥, 则

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3).$$

对任意三个事件 A_1, A_2, A_3, 有

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = {} & P(A_1) + P(A_2) + P(A_3) \\ & - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) \\ & + P(A_1A_2A_3). \end{aligned}$$

注: 更一般地, 可以使用容斥原理计算: 若 n 个事件 $A_1, A_2, \cdots, A_n$ 两两互斥, 则

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

对任意 n 个事件 $A_1, A_2, \cdots, A_n$, 有

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \left[(-1)^{k+1} \sum_{1 \le i_1 \le \cdots \le i_k \le n} P(A_{i_1} \cdots A_{i_k}) \right].$$

2.3 古典概型模型

定义 2.3. 如果一个随机试验具有以下特点:

- 样本空间只含有限多个样本点;
- 各样本点出现的可能性相等,

则称此随机试验是古典型的. 此时对每个事件 $A \subset \Omega$,

$$P(A) = \frac{\text{事件 } A \text{ 包含的样本点数}}{\text{样本点的总数}} = \frac{n(A)}{n(\Omega)}$$

称为事件 A 的**古典概率**.

根据上述的定义, 我们可以立即得出 $P(\emptyset) = 0$, $P(\Omega) = 1$.

2.4 几何概型

定义 2.4. 设样本空间为有限区域 Ω, 若样本点落入 Ω 内的任何区域 G 中的概率与区域 G 的测度成正比, 则样本点落入 G 内的概率为:

$$p = \frac{|G|}{|\Omega|}$$

我们可以先简单地把 “测度” 理解为面积. 并且我们发现, 如果 $P(A) = 0$, 那么 A 不一定是不可能事件. 比如正方形区域中的一个点, 一个点的面积为 0. 因此正方形中选一个点的概率总为 0.

多知道一点: 一点计数技巧

这里有点困难, 不妨先跳过.

古典概型的计数有时候会变得更有效率, 下面举例几个计数问题.

我们来看著名的 “The Twelvefold Way” 这个问题: 它包括了 12 种从有 n 个球放入有 k 个盒子里的方法. 每种方法具有独特的限制, 包括球和盒子是否是区分的及是否允许空盒子等.

n 个球	k 个盒子	想怎么放怎么放	每个盒子最多 1 个球	不允许有空盒子
不同的球 o0o	不同的盒子 $\boxed{1}\boxed{2}\boxed{3}$	(1)	(2)	(3)
相同的球 ooo	不同的盒子 $\boxed{1}\boxed{2}\boxed{3}$	(4)	(5)	(6)
不同的球 o0o	相同的盒子 $\square\square\square$	(7)	(8)	(9)
相同的球 ooo	相同的盒子 $\square\square\square$	(10)	(11)	(12)

我们下面来看这个问题.

问题 (1): n 个球, k 个盒子, 盒子和球都是不同的, 随便放 我们希望做的事情是 “把 n 个球放入 k 个盒子”. 这时候, 我们对于第一个球的选择就随便选一个就好了. 因此有 k 种方法. 对于第二个球, 因为没有限制, 我们照样可以用 k 种方法... 一直到第 n 个球. 因此总共的方案是 k^n.

问题 (2): n 个球, k 个盒子, 盒子和球都是不同的, 每个盒子最多 1 个球 我们假设盒子的个数多于球, 这样做的事情就会有意义一点. 我们希望做的事情是 “把 n 个球放入 k 个盒子, 每个盒子最多 1 个球”. 这时候, 我们对于第一个球的选择就随便选

一个就好了. 因此有 k 种方法. 对于第二个球, 因为没有限制, 我们可以用 $k-1$ 种方法 (有一个已经占用了)... 一直到第 n 个球, 就有 $k-n+1$ 个. 因此总共的方案是 $k(k-1)(k-2)\cdots(k-n+1)$.

我们一般把这个叫做排列数, 因为它阐述的是从 k 个物品里面选择 n 个数的方法.[1] 同时, 从 k 开始, 往下乘 n 个数也被称作下降幂 (falling power).

定义 2.5 (排列数)**.** 从 n 个物品里面选择 k 个数的方法数记作排列数. 记作 A_n^k. 计算方法为

$$A_k^n = k(k-1)(k-2)\cdots(k-n+1)$$

其中 $k(k-1)(k-2)\cdots(k-n+1)$ 可以被记作下降幂, 写作 $k^{\underline{n}}$.

问题 (3): n 个球, k 个盒子, 盒子和球都是不同的, 不允许有空盒子我们发现当我们的球的数量不少于盒子数量的时候这个内容才有意义.

既然我不允许有空盒子, 我先随便挑出来 k 个球去 "压箱底", 然后剩下的像刚刚一样随便放不就好了? 其实这个方法是不对的. 因为这样会算重复一些方案 – 你默认的要压箱底的和后来放的在这里是考虑次序的, 而原来的问题是不考虑次序的. 那我们该怎么做?

这事实上是集合的一个划分. 每一个划分正好对应一个集合. 我们如果能够把这个集合划分为 k 份, 然后再把每一个划分对应上一个盒子就好了. 第二步很简单, 直接乘上 $k!$ 即可.

关键是如何划分这个集合? 为了方便我们的符号书写, 我们先记 $\left\{ {n \atop k} \right\}$ 为把 n 个集合划分为 k 个部分的个数. 这时候, 我们与其一口吃个胖子, 我们可以一步一步地考虑[2]

要把 $\{1, 2, \ldots, n\}$ 划分为 k 份, 可以借助那些以往的状态可以把我们带到 $\left\{ {n \atop k} \right\}$.

第一种情况是, 我们已经把集合 $\{1, 2, \ldots, n-1\}$ 分为了 k 个部分, 现在的任务是把 n 放入任何这 k 部分的其中之一. 这就给了我们 $k\left\{ {n-1 \atop k} \right\}$ 种方法达到这个目的.

第二种情况是, 我们已经把 $\{1, 2, \ldots, n-1\}$ 分为了 $k-1$ 个部分, 并且让 $\{n\}$ 单独一份. 这样, 我们就有 $\left\{ {n-1 \atop k-1} \right\}$ 种方法.

这两种方法构建的分割是不同的: 因为在第一种方法中, n 始终位于一个大小 > 1 的划分部分中, 而在第二种方法中, $\{n\}$ 始终是一个单独的一部分. 因此这两种情况是不重叠的. 而对于任意一个 n 元素集合分割为 k 份, 必定可以通过这两种方法之一来构建. 因此根据求和法则:

$$\left\{ {n \atop k} \right\} = k \left\{ {n-1 \atop k} \right\} + \left\{ {n-1 \atop k-1} \right\}$$

成立.

要求得这个递归式的表达式是十分难的. 我们一般到此为止了. 事实上, 这个内容叫做**第二类 Stirling 数 (Stirling number of the second kind)**. 要计算第二类 Stirling 数, 我们有如下的公式:

-- 你为什么总是用英文写人名?
-- 省点笔画, 好写.
-- 可是这是打字啊!
-- 确实. 但是

定义 2.6 (第二类 Stirling 数)**.** 将一个大小为 n 的集合划分为 k 个部分的方案数被命名为第二类 Stirling 数. 记作 $\left\{ {n \atop k} \right\}$.

定理 2.7. 第二类 Stirling 数满足关系

$$\left\{ {n \atop k} \right\} = k \left\{ {n-1 \atop k} \right\} + \left\{ {n-1 \atop k-1} \right\}$$

于是, 我们一般使用递推计算的方式计算这个集合. 这个确实需要很多思考, 这就是为什么我们会用一个伟大数学家的名字命名它.

这下子, 我们就得到了第三个问题的答案: $k! \left\{ {n \atop k} \right\}$.

问题 (5): n 个相同的球, k 个不同的盒子, 每个盒子顶多 1 个球 这就要求我们搞清楚到底哪个可以有球, 哪个盒子里面没有球就行了. 所以我们要求从 k 个里面选取 n 个出来. 这个应该如何计算呢? 实际上, 我们可以先从排列数出发, 然后想一想把它们分成若干个组, 也就是从小到大排个序. 这样子就是总共的组合数有 $A_k^n/k!$ 个. 为了方便起见, 我们把这个定义做组合数.

定义 2.8 (组合数)**.** 从 n 个物品里面选取 k 个数的方案数为组合数, 记作 $\binom{n}{k}$, 或者 C_n^k. 定义为

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

问题 (4): n 个相同的球, k 个不同的盒子, 随便放 由于每一个球是相同的, 所以我们需要关注每一个盒子里面被放了多少球. 因此, 我们就相当于要在这几个球的空档里面 “插板”. 由于随意放置, 我们相当于要在 $n+k-1$ 个里面选出 k 个, 于是, 得到了

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!} = \frac{n(n+1)(n+2)\cdots(n+k-1)}{k!}.$$

我们把这个记作多重组合数的系数 (非标准官方译名):

定义 2.9 (多重集合组合数)**.** 多重集合的组合数定义为

$$\left(\binom{n}{k}\right) = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!} = \frac{n(n+1)(n+2)\cdots(n+k-1)}{k!}.$$

其中, $n(n+1)(n+2)\cdots(n+k-1)$ 这样的从 n 开始, 向上乘 k 个数这样的被称为上升幂. 方便起见记作 $n^{\bar{k}}$

于是, 我们得到了这个问题的答案: $\left(\binom{k}{n}\right)$.

问题 (6): n 个相同的球, k 个不同的盒子, 每个盒子不许空 那么我们不妨首先把

前几个球放到前几个球里面, 然后剩下的就得到了不受限制的状况了. 也就是我们这个的答案是 $\binom{n-1}{k-1}$.

问题 (7): n 个不同的球, k 个相同的盒子, 随便放我们可以把 $\{1, 2, \cdots, n\}$ 划分进 i 个非空的盒子, 其中, $i \leq k$. 于是根据加法原理, 这个问题的答案是 $\sum_{i=1}^{k} \left\{ {n \atop i} \right\}$.

问题 (8): n 个不同的球, k 个相同的盒子, 每个盒子顶多一个球事实上, 如果 $n > k$, 那么不可能做到. 根据抽屉原理, 总有一个盒子要装两个球. 反之, 我们就可以做到. 于是这个问题的答案是

$$\begin{cases} 1 & \text{if } n \leq k \\ 0 & \text{if } n > k \end{cases}.$$

问题 (9): n 个不同的球, k 个相同的盒子, 不允许有空的盒子 哈哈! 这不就是我们集合划分的定义吗? 这样, 我们就可以用 $\left\{ {n \atop k} \right\}$ 表示了.

问题 (12): n 个球, k 个盒子, 盒子和球都相同, 不能有空盒子 其实这个是当我们把球放进盒子里面之后, 真正重要的是什么? 事实上, 我们发现我们只要关心每个盒子有几个球就好了, 并且我们不用关心有多少球的顺序. 等价地说, 就是把一个整数分拆. 比如 7 就可以这样分拆成 $1, 2, \cdots, 7$ 部分:

$$\begin{array}{ll} \{7\} & p_1(7) = 1 \\ \{1,6\}, \{2,5\}, \{3,4\} & p_2(7) = 3 \\ \{1,1,5\}, \{1,2,4\}, \{1,3,3\}, \{2,2,3\} & p_3(7) = 4 \\ \{1,1,1,4\}, \{1,1,2,3\}, \{1,2,2,2\} & p_4(7) = 3 \\ \{1,1,1,1,3\}, \{1,1,1,2,2\} & p_5(7) = 2 \\ \{1,1,1,1,1,2\} & p_6(7) = 1 \\ \{1,1,1,1,1,1,1\} & p_7(7) = 1 \end{array}$$

等价地说, 我们的要求是一个数 n 的 k 分拆, 分别记作 $x_1, x_2, \cdots, x_k$, 满足如下的条件 (*):

- $x_1 \geq x_2 \geq \cdots \geq x_k \geq 1$;
- $x_1 + x_2 + \cdots + x_k = n$.

为了方便起见, 我们把整数 n 分拆成 k 部分记作 $p_k(n)$. 读作 "n 的 k-分割" 下面我们同样用类似于递归的方法来求解这个问题:

假设 $(x_1, \ldots, x_k)$ 是 n 的一个 k-分割. 满足刚刚我们提到过的条件 (*).

我们对这个问题分类讨论: 第一种情况是, 如果 $x_k = 1$, 那么 $(x_1, \cdots, x_{k-1})$ 是把 $n-1$ 分割成的一个不同的 $(k-1)$-分割

第二种情况是, 如果 $x_k > 1$, 那么 $(x_1 - 1, \cdots, x_k - 1)$ 是 $n-k$ 的一个不同的 k-分割. 并且每个 $n-k$ 的 k-分割都可以通过这种方式得到. 因此在这种情况下, n 的 k-分割数目为 $p_k(n-k)$.

由于所有的情况都已经讨论完毕, 因此, 我们可以使用加法原理, 把这两个部分加

起来, 得到了 n 的 k-分割数目为 $p_{k-1}(n-1)+p_k(n-k)$, 即

$$p_k(n)=p_{k-1}(n-1)+p_k(n-k)\,.$$

定义 2.10 (分拆数)**.** 定义分拆数 $p_k(n)$ 表示把一个正整数 n 分拆为 k 部分, 分别记作 $x_1,x_2,\cdots,x_k$, 满足如下的条件的个数:

- $x_1\geq x_2\geq\cdots\geq x_k\geq 1$;
- $x_1+x_2+\cdots+x_k=n$.

定理 2.11. 分拆数满足性质

$$p_k(n)=p_{k-1}(n-1)+p_k(n-k)\,.$$

所以我们这个问题的答案就是 $p_n(k)$.

问题 (10): n 个球, k 个盒子, 盒子和球都相同, 随便放 有了分拆数之后, 我们就可以决定到底要分拆多少个了, 于是答案就是 $\sum_{i=1}^{k}p_i(n)$.

问题 (11): n 个球, k 个盒子, 盒子和球都相同, 每个盒子顶多 1 个球 它和第 (8) 问的情况类似. 同样要么能做, 要么不能做. 原理还是依照第八个问题一样.

这样我们就得到了整个表格的全貌:

n 个球	k 个盒子	想怎么放怎么放	每个盒子最多 1 个球	不允许有空盒子
不同的球 o0o	不同的盒子 $\boxed{1}\boxed{2}\boxed{3}$	k^n	$k^{\underline{n}}$	$n!\left\{{n\atop k}\right\}$
相同的球 ooo	不同的盒子 $\boxed{1}\boxed{2}\boxed{3}$	$\binom{k}{n}$	$\binom{k}{n}$	$\binom{\binom{k}{n-k}}{}$
不同的球 o0o	相同的盒子 $\square\square\square$	$\sum_{i=1}^{k}\left\{{n\atop i}\right\}$	$\begin{cases}1 & \text{if } n\leq k\\ 0 & \text{if } n>k\end{cases}$	$\left\{{n\atop k}\right\}$
相同的球 ooo	相同的盒子 $\square\square\square$	$\sum_{i=1}^{k}p_i(n)$	$\begin{cases}1 & \text{if } n\leq k\\ 0 & \text{if } n>k\end{cases}$	$p_k(n)$

不要担心! 这张表格看起来有些复杂. 其实, 这张表格没有记忆的必要. 现在我们只需要学习排列数和组合数就可以建立一个很好的模型了. 这些概念是非常有趣和实用的, 它们能够帮助我们解决很多有趣的问题.

上述材料里面的有时候我们还会遇到更加复杂的问题, 比如对于分拆数, 我们需要将一个数分拆成若干个部分, 并且考虑它们之间的顺序. 都可以通过一些递归的方法来解决. 我们只当做对大家的训练. 一个初学者当然需要看过足够多的例子, 加以大量的思考才能设计出比较好的这方面的内容. 大家完全不必着急.

假设我们有 n 个不同的球, k 个不同的盒子. 我们可以用一个映射的方式来描述不同的放置方法. 具体来说, 我们可以把每个盒子看作一个 “投影”, 而每个球就是我们要放入的 “元素”. 这样, 每一种放置方法就可以看作是一个特定的映射.

那么, 任意的映射就是我们刚刚的 “随便放”; 单射就是我们的 “每个盒子只放一个球”; 满射就是 “每个盒子不能空”. 因此, 这个表格更为一般的情况你就能够看得懂了.

N	M	任何一个 $f: N \to M$	单射 $f: N \xrightarrow[1\text{-}1]{} M$	满射 $f: N \xrightarrow[\text{onto}]{} M$
不同的球 o0o	不同的盒子 $\boxed{1}\boxed{2}\boxed{3}$	k^n	$k^{\underline{n}}$	$n!\left\{ {n \atop k} \right\}$
相同的球 ooo	不同的盒子 $\boxed{1}\boxed{2}\boxed{3}$	$\left(\left({k \atop n} \right)\right)$	$\binom{k}{n}$	$\left(\left({k \atop n-k} \right)\right)$
不同的球 o0o	相同的盒子 $\square\square\square$	$\sum_{i=1}^{k} \left\{ {n \atop i} \right\}$	$\begin{cases} 1 & \text{if } n \le k \\ 0 & \text{if } n > k \end{cases}$	$\left\{ {n \atop k} \right\}$
相同的球 ooo	相同的盒子 $\square\square\square$	$\sum_{i=1}^{k} p_i(n)$	$\begin{cases} 1 & \text{if } n \le k \\ 0 & \text{if } n > k \end{cases}$	$p_k(n)$

3 条件概率

定义 3.1. 设 $P(A) > 0$, 称

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

为在**事件 A 发生条件下, 事件 B 的条件概率**. (记作 $P(B \mid A)$)

比如, 在古典概率模型中,

$$P(B|A) = \frac{\text{事件 } AB \text{ 包含的样本点数}}{\text{事件 } A \text{ 包含的样本点数}} = \frac{n(AB)}{n(A)}.$$

条件概率也是概率, 因此它也具有概率的性质:

$$\begin{aligned} &P(A \mid A) = 1 \\ &P(\varnothing \mid A) = 0 \\ &P(B \mid A) = 1, \quad B \supseteq A \\ &P\left(B_1 + B_2 \mid A\right) = P\left(B_1 \mid A\right) + P\left(B_2 \mid A\right), B_1, B_2 \text{互斥} \end{aligned}$$

更抽象的, 我们有

命题 3.2. 设 $P(A) > 0$, 则

- 对任意事件 B, 均有 $P(B|A) \ge 0$;
- $P(\Omega|A) = 1$;
- 若事件序列 $\{B_n\}_{n\ge1}$ 两两互斥, 则

$$P\left(\bigcup_{n=1}^{\infty} B_n \middle| A\right) = \sum_{n=1}^{\infty} P(B_n|A).$$

由这些性质可见, 对于固定的事件 A, 在概率空间 $(\Omega \cap A, \mathscr{B} \cap A)$ 上的条件概率

$P(\cdot \mid A)$, 以及在空间 $(\Omega, \mathscr{C})$ 上的概率 $P(\cdot)$ 具有同样的性质, 其中

$$\mathscr{A} \cap A = \{B \cap A : B \in \mathscr{B}\}$$

全概率公式

由于条件概率, 我们有概率的乘法公式:

定理 3.3 (乘法公式)**.** 由条件概率的定义, 得到

- 如果 $P(A) > 0$, 则有 $P(AB) = P(A)P(B|A)$.
- 如果 $P(B) > 0$, 则有 $P(AB) = P(B)P(A|B)$.

我们同样可以把这个性质推广到 n 个物品的时候.

推论 3.4. 如果 $P(A_1A_2 \cdots A_{n-1}) > 0$, 则有**乘法公式**

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

另一个简单而重要的公式称做全概率公式, 是利用条件概率计算复合事件的基本工具. 我们首先希望对样本空间进行划分. 然后再进行求解:

定义 3.5. 设 Ω 为某试验的样本空间, $A_1, A_2, \cdots, A_n$ 为一组事件. 如果以下条件成立:

- $A_1, A_2, \cdots, A_n$ 两两互斥,
- $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$,

则称 $A_1, A_2, \cdots, A_n$ 为样本空间 Ω 的一个**划分**.

有了这样的分类之后, 我们就可以给出全概率了.

定理 3.6 (全概率公式)**.** 如果 $A_1, A_2, \cdots, A_n$ 是样本空间的划分, 且都有正概率, 则对任意事件 B 有

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

Proof. 考虑基本事件空间 Ω 的某个分割 $\mathscr{D} = \{A_1, \cdots, A_n\}$, 且 $P(A_i) > 0 (i = 1, 2, \cdots, n)$. (这样的分割又称做不相容事件的完全事件组.) 显然,

$$B = BA_1 + \cdots + BA_n,$$

因此

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

其中

$$P\left(BA_i\right)=P\left(B\mid A_i\right)P\left(A_i\right)$$

□

例子 3.1. 假设要对研究生论文抄袭现象进行社会调查，我们设计两个具有**相同答案**的问题：

- 你的生日是否在 7 月 1 日以前？
- 你做论文时是否有过抄袭行为？

同时提供给受访者一个放有等量红球和白球的袋子，受访者在不被观察的情况下从袋子中随机取一个球观察颜色后放回. 如果是红球回答第一个问题，白球回答第二个问题.

假定受访者有 150 人，统计出共有 60 个回答“是”. 问：有抄袭行为的比率是多少？

解答： 事件 A 表示抽到白球，事件 B 表示回答是，则有

$$P(B)=P(A)P(B|A)+P(\overline{A})P(B|\overline{A}).$$

代入已知的概率，得到

$$\frac{60}{150}=\frac{1}{2}\cdot P(B|A)+\frac{1}{2}\cdot\frac{1}{2}$$

求得

$$P(B|A)=\frac{3}{10}$$

3.1 Bayes 公式

设事件 A 和 B 的概率大于 $0:P(A)>0,P(B)>0$，利用乘法公式，我们可以先看 B 而非 A，得到：

$$P(AB)=P(A\mid B)P(B).$$

和刚刚得到的 $P(AB)=P(B\mid A)P(A)$ 对比，得到了

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

定理 3.7 (Bayes 定理)**.** 设 $0<P(A)<1$，$P(B)>0$，则有

$$P(A|B)=\frac{\mathbb{P}(AB)}{\mathbb{P}(B)}=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\overline{A})P(B|\overline{A})}$$

Web Demonstrate Aside

著名的科普视频频道主 3Blue1Brown 曾经对 Bayes 定律进行了可视化. 可以参考Bilibili: BV1R7411a76r.

假如事件组 $A_1, \cdots, A_n$ 是 Ω 的一个分割, 那么有

推论 3.8. 如果 $A_1, A_2, \cdots A_n$ 是样本空间的一个划分, 且都有正概率, 则对任意正概率的事件 B 有

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}.$$

实际上, 在统计应用中, 事件 $A_1, \cdots, A_n$ 组成事件组 $(A_1 + \cdots + A_n = \Omega)$, 常称做“假设” 或 “假说”, 而 $P(A_i)$ 称做假设 A_i 的先验概率). 条件概率 $P(A_i \mid B)$ 称做假设 A_i 在事件 B 出现后的后验概率.

练习 I.1 假设匣中有两枚硬币: A_1 是一对称的硬币, “正面” Z 出现的概率等于 1/2, 而 A_2 是一枚不对称的硬币, “正面” Z 出现的概率等于 1/3. 随意选出一枚硬币并将其投掷, 结果掷出正面. 问抽到硬币为对称硬币的概率如何? #

解答: 建立相应的概率模型. 这里自然取集合 $\Omega = \{A_1\mathrm{Z}, A_1\ \mathrm{F}, A_2\mathrm{Z}, A_2\ \mathrm{F}\}$, 可以描绘选取和投掷的结局, 其中 A_1Z 表示 “选中硬币” A_1, 结果掷出正面 Z, 等等, 而 F 表示硬币掷出反面. 根据条件, 所考虑结局的概率应该是:

$$P(A_1) = P(A_2) = \frac{1}{2}$$

和

$$P(\mathrm{Z} \mid A_1) = \frac{1}{2}, \quad P(\mathrm{Z} \mid A_2) = \frac{1}{3}.$$

这些条件唯一决定各结局的概率:

$$P(A_1\mathrm{Z}) = \frac{1}{4}, P(A_1\ \mathrm{F}) = \frac{1}{4}, P(A_2\mathrm{Z}) = \frac{1}{6}, P(A_2\ \mathrm{F}) = \frac{1}{3}$$

那么, 根据贝叶斯公式, 所求的概率为

$$P(A_1 \mid \mathrm{Z}) = \frac{P(A_1)\,P(\mathrm{Z} \mid A_1)}{P(A_1)\,P(\mathrm{Z} \mid A_1) + P(A_2)\,P(\mathrm{Z} \mid A_2)} = \frac{3}{5}$$

练习 I.2 袋子中有 10 个白球, 5 个黑球. 现掷一枚均匀的骰子. 掷出几点就从袋中取几个球. 若已知取出的球全为白球, 求掷出 3 点的概率. #

解答: 原问题的意思是在取出的球全为白球的条件下, 掷出三点的概率. 设 $B =$ {取出的球全是白球}, $A =$ {掷出i点}$(i = 1, 2, \cdots, 6)$.

$$\begin{aligned} P(A_3|B) &= \frac{P(A_3)P(B|A_3)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_6)P(B|A_6)} \\ &= \frac{\frac{1}{6} \times \frac{\binom{5}{3}}{\binom{15}{3}}}{\sum_{i=1}^{5} \frac{1}{6} \times \frac{\binom{5}{i}}{\binom{15}{i}} + \frac{1}{6} \times 0} = 0.4835 \end{aligned}$$

4 事件的独立性

定义 4.1. 若两事件 A、B 满足

$$P(AB) = P(A)P(B),$$

则称**事件 A、B 相互独立**.

实际意义: 若 $P(B) > 0$, 则上式等价于

$$P(A|B) = P(A),$$

即**事件 A 的概率不受事件 B 发生与否的影响**. 也就是事件 B 没有给我们任何的信息.

注: "两个事件互斥" 和 "两个事件相互独立" 是不同的概念:

- 互斥 $\Rightarrow P(A \cup B) = P(A) + P(B)$;
- 独立 $\Rightarrow P(AB) = P(A)P(B)$.

但两者也有关系: 如果 $P(A) > 0$ 且 $P(B) > 0$, 则两者不可能既是互斥的又是独立的.

我们接下来看多个事件的独立性:

定义 4.2. 称 $n(n \geq 2)$ 个事件 $A_1, A_2, \cdots, A_n$ 相互独立, 如果对任意一组指标

$$1 \leq i_1 < i_2 < \cdots < i_k \leq n \quad (k \geq 2)$$

都有

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

发现若 A 与 B 相互独立, 且 B 与 C 相互独立, 则 A 与 C **未必**相互独立.

例子 4.1. 从全体有两个孩子的家庭中随机选择一个家庭, 并考虑下面三个事件:

- A 为 "第一个孩子是男孩",

- B 为“两个孩子不同性别”,
- C 为“第一个孩子是女孩”.

容易验证 A 与 B 相互独立, B 与 C 相互独立, 但是 A 与 C 不独立.

同样, 三个两两独立的也不一定都独立.

伯恩斯坦四面体问题: 一个正四面体有三面各涂上红, 白, 黑三种颜色. 第四面同时涂上三种颜色. 这四个面等概率出现在底面. 以 A, B, C 分别表示四面体底面出现红, 白, 黑三种颜色的事件. 问 A, B, C 是否相互独立?

$$P(A) = P(B) = P(C) = 2/4 = 1/2,$$

$$P(AB) = P(BC) = P(AC) = 1/4,$$

$$P(ABC) = 1/4 \neq P(A)P(B)P(C).$$

4.1 多知道一点: 朴素的 Bayes 分类器

我们希望对事情做分类.

我们有 n 个训练示例的集合, 每个对象具有一系列特征: $(X_1, X_2, \cdots, X_m)$, 对于每个 X_i 都在一组特定区域取值. 每个 D_i 由表达其特征的一组值（方便起见写作向量形式）

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$$

对象可能的分类集合 $C = \{C_1, C_2, \cdots, C_t\}$.$C(D_i)$ 是 D_i 的分类. 形式化的说, 我们就有了下面的一组集合供我们训练.

$$\{(D_1, C(D_1)), (D_2, C(D_2)), \cdots, (D_n, C(D_n))\}$$

假设我们有一个足够大的训练集, 然后, 对于每个向量 $y = (y_1, \cdots, y_m)$ 和每一个 c_j, 我们完全可以使用训练集计算一个这样的经验条件概率: 具有特征向量 y 的对象被分类为 C_j.

根据条件概率的定义, 我们知道 $p_{y,j}$ 的计算公式

$$P_{y,j} = \frac{P\{(\forall i, x_i = y_i) \wedge c(D_i) = c_i\}}{P\{\forall i, x_i = y_i\}}$$

当一个具有特征 x^* 的对象过来的时候, 我们计算 $P(c(D^*) = c_j | x^* = (x_1^*, \cdots, x_n^*))$ 的估计. 最后, 我们返回一个向量 $(p_{x^*,1}, p_{x^*,2}, \cdots, p_{x^*,n})$, 表示它被划分到各个类的概率大小.

这种方法的难处在于我们需要大量的准确的样本. 即使我们的特征只能取得 0 和 1 两个值, 如果有 m 个特征, 我们也要获得 2^m 个特征值与之对应.

如果我们的这些特征相互独立, 我们就可以加快这个进程:

$$\begin{aligned} P\left(c\left(D^*\right)=c_j \mid x^*\right) &= \frac{P\left(x^* \mid c\left(D^*\right)=c_j\right) \cdot P\left(c\left(D^*\right)=c_j\right)}{P\left(x^*\right)} \\ &= \frac{\prod_{k=1}^m P\left(x_k^*=x_i \mid c\left(D^*\right)=c_j\right) \cdot P\left(c\left(D^*\right)=c_j\right)}{P\left(x^*\right)}. \end{aligned}$$

其中 x_k^* 是 x^* 的第 k 个分量. 由于每个特征的可能数量一定, 在 m 个特征的时候, 我们只需要学习 $O(m|C|)$ 的概率估计.

训练过程很简单:

- 对于每一类 c_j, 看一看被分类为 c_j 与总共的占比为多少, 并用此计算

$$\hat{P}\left(c\left(D^*\right)=c_j\right)=\frac{\left|\left\{i \mid c\left(D_i\right)=c_j\right\}\right|}{|D|}$$

  这里 $\hat{P}$ 表示我们算的是经验概率.
- 对于每一个特征 X_k 和对应的特征值 x_k, 我们注意这个特征值 x_k 被分到 c_j 那一类的占比. 也就是

$$\hat{P}\left(x_k^*=x_k \mid c\left(D^*\right)=c_j\right)=\frac{\left|\left\{i: x_k^i=x_k, c\left(D_i\right)=c_j\right\}\right|}{\left\{i \mid c\left(D_i\right)=c_j\right\} \mid}.$$

一旦我们训练好了分类器, 当一个具有特征向量 $x^*=\left(x_1^*, \cdots, x_m^*\right)$ 新的对象 D_i^* 来了的时候, 对于每一个 c_j, 我们就可以计算

$$\left(\prod_{k=1}^m \hat{P}\left(x_k^*=x_k \mid c\left(D^*\right)=c_j\right)\right) \hat{P}\left(c\left(D^*\right)=c_j\right)$$

并且取最大值, 得到它最可能在的分类.

总结一下, 我们的算法如算法 1所示.

多知道一点: 事件独立性的表述

在概率论中, 往往不但需要考虑事件 (集合) 的独立性, 而且需要研究事件 (集合) 组的独立性. 下面引进相应的定义.

定义: 称 Ω 子集系的代数 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ (关于概率 P) 为独立的或统计独立的, 如果对于相应地属于 $\mathscr{A}_1$ 和 $\mathscr{A}_2$ 的两个任意子集 A_1 和 A_2 独立.

算法 1: 朴素 Bayes 分类器算法

Data: 所有分类的集合 C, 特征以及对应的特征值 $F_1, F_2, \cdots, F_m$, 用于分类项目的训练集 D

训练阶段

1 对于每一类 $c \in C$, 特征 $k = 1, 2, \cdots, m$, 以及特征值 $x_k \in F_k$, 计算

$$\hat{P}\left(x_k^* = x_k \mid c\left(D^*\right) = c\right) = \frac{\left|\left\{i : x_k^i = x_k, c\left(D_i\right) = c\right\}\right|}{\left\{i \mid c\left(D_i\right) = c\right\} \mid}.$$

2 对于每一类 $c \in C$, 计算

$$\hat{P}\left(c\left(D^*\right) = c\right) = \frac{\left|\left\{i \mid c\left(D_i\right) = c\right\}\right|}{|D|}.$$

给具有特征向量 $x^* = x = (x_1, \ldots, x_m)$ 的新的对象 D^* 归一个类:

1 最有可能的分类:

$$c\left(D^*\right) = \arg\max_{c_j \in C} \left(\prod_{k=1}^{m} \hat{P}\left(x_k^* = x_k \mid c\left(D^*\right) = c_j\right) \right) \hat{P}\left(c\left(D^*\right) = c_j\right).$$

2 计算可能的分类分布:

$$\hat{P}\left(c\left(D^*\right) = c_j\right) = \frac{\left(\prod_{k=1}^{m} \hat{P}\left(x_k^* = x_k \mid c\left(D^*\right) = c_j\right)\right) \hat{P}\left(c\left(D^*\right) = c_j\right)}{\hat{P}\left(x^* = x\right)}$$

Part II

一维随机变量

5 随机变量

下面我们继续就说一说随机变量的初步想法.

我们考虑在有限个试验结局的情形. 设 $(\Omega, \mathscr{A}, P)$ 是某具有有限个结局试验的概率模型, $N(\Omega)$ 是 Ω 中基本事件的个数, 而 $\mathscr{A}$ 是 Ω 中所有子集的代数.

比如, 掷硬币.

例子 5.1. *a) 对于接连两次掷硬币模型, 其基本事件空间为 $\Omega = \{\mathrm{ZZ}, \mathrm{ZF}, \mathrm{FZ}, \mathrm{FF}\}$, 其中 Z 为正面, F 为反面. 我们利用下面的表格定义随机变量 $X = X(\omega)$ 其中 $X(\omega)$ 是对应于 ω 的“正面”出现的次数:*

ω	ZZ	ZF	FZ	FF
$X(\omega)$	*2*	*1*	*1*	*0*

b) 随机变量 X 的另一简单的例子是某集合 $A \in \mathscr{A}$ 的示性函数 (亦称特征函数):

$$X = I_A(\omega)$$

其中,

$$I_A(\omega)=\begin{cases}1, & \omega\in A\\ 0, & \omega\notin A\end{cases}$$

当实验者遇到描绘某些记载或读数的随机变量时, 则他关心的基本问题是, 该随机变量取各个数值的概率如何. 换句话说, 他关心的不是概率 P 在 $(\Omega,\mathscr{A})$ 上的分布, 而是概率在随机变量之可能值的集合上的分布. 对于我们现在研究的情形, Ω 由有限个点构成, 则随机变量 X 的值域 X 也是有限的.

设 $X=\{x_1,\cdots,x_m\}$, 其中 $x_1,\cdots,x_m$ 是 X 的全部可能值.

记 $\mathscr{X}$ 为值域 X 上一切子集的全体, 并设 $B\in\mathscr{X}$. 当 X 是随机变量 X 的值域时, 集合 B 也可以视为某个事件.

在 $(X,\mathscr{B})$ 上考虑由随机变量 X 按照,

$$P_X(B)=P\{\omega:X(\omega)\in B\},\quad B\in\mathscr{B}$$

产生的概率 $P_X(\cdot)$. 显然, 这些概率的值完全决定于:

$$P_X\left(x_i\right)=P\left\{\omega:X(\omega)=x_i\right\},\quad x_i\in X$$

其中组数 $\left\{P_X\left(x_1\right),\cdots,P_X\left(x_m\right)\right\}$ 称做随机变量 X 的概率分布.

随机变量 X 的构造完全由概率分布 $\left\{P_X\left(x_i\right),1,2,\cdots,m\right\}$ 描述. 下面引进的分布函数的概念, 提供随机变量构造的等价描述.

定义: 设 $x\in\mathbb{R}^1$. 函数

$$F_X(x)=P\{\omega:X(\omega)\le x\}$$

称做随机变量 X 的分布函数,

我们发现这和实值函数 (如果对于每一个实数 x 又有唯一的 y 与之对应, 那么称为 y 实变量 x 的函数) 有很多相似之处. 把这个定义的前提推广到 x 不是 "实数" 的情形. 如果我们把这个想法定义在样本空间上, 这样的函数就是随机变量. 意味着

定义: 定义在样本空间上的函数就称为随机变量.

因为他们会为每一个样本点对应一个值, 在我们进行随机试验的时候, 这个值就好像是随机的一样.

定义 5.1 (随机变量)**.** 设 Ω 是某随机试验的样本空间. 如果对于每个 $\omega\in\Omega$, 都有一个实数 $X(\omega)$ 与其对应, 这样就得到一个定义在 Ω 上的函数

$$X=X(\omega),$$

称该函数为随机变量 (random variable).

随机变量一般用大写英文字母 X、Y、Z 或小写希腊字母 ξ、η、γ 来表示.

- 大写字母: 一个实验中的值
- 小写字母: 某个具体实验中的取值

按照研究的顺序, 我们现在先研究离散型 (只能取有限个或者可列个值), 然后使用微积分的技巧研究连续型 (取得某一区间内的任何数值). 在这期间, 我们借助一些手段 (如定义概率分布函数等) 从而可以研究混合型 (一部分连续, 一部分离散) 的随机变量.

仿照事件的独立性的定义, 我们给出随机变量的独立性的定义. 同样, 这表示他们不彼此依赖.

定义 5.2. 设 $X_1, \cdots, X_r$ 在 $\mathbb{R}^1$ 是一组中 (有限) 集合 X 上取值的随机称随机变量 $X_1, \cdots, X_r$ 为 (全体) 独立的, 如果对于任意 $x_1, \cdots, x_r \in X$,

$$P\{X_1 = x_1, \cdots, X_r = x_r\} = P\{X_1 = x_1\} \cdots P\{X_r = x_r\}$$

记 $\mathscr{X}$ 是 X 中所有子集的代数, 上述的内容可以等价的写成: 对于任意 $B_1, \cdots, B_r \in \mathscr{X}$,

$$P\{X_1 \in B_1, \cdots, X_r \in B_r\} = P\{X_1 \in B_1\} \cdots P\{X_r \in B_r\}$$

6 离散型随机变量的概率分布

回顾上一节介绍的基本想法:

定义 6.1 (概率分布)**.** 若离散型随机变量 X 的所有可能值为 $\{x_k\}$, 分别对应概率 $\{p_k\}$, 则称

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

为 X 的概率分布 (分布律).

为了方便起见, 可以把概率列表以得到直观描述:

X	x_1	x_2	$\cdots$	x_k	$\cdots$
P	p_1	p_2	$\cdots$	p_k	$\cdots$

这张表我们称作随机变量 X 的分布列.

由于每一个概率是非负的, 并且我们遍历的整个概率空间, 因此概率分布具有如下的性质: $\begin{cases} p_k \geq 0, \quad k = 1, 2, \cdots \\ \sum\limits_k p_k = 1 \end{cases}$

有时候我们对于小于某一个特定值的变量的取值内容感兴趣. 概率分布函数就是统计小于某一个特定值发生的概率. 同时这种定义也可以让我们从一个具体的点之中脱

身. 我们的视野就可以看得更广了, 这种技巧在连续的时候很有用处.

定义 6.2 (概率分布函数)**.** 设离散型随机变量 X 的概率分布为

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

则定义 X 的分布函数为

$$F_X(x) = P(X \le x) = \sum_{x_k \le x} p_k,$$

这里的和式是对所有满足 $x_k \le x$ 的 p_k 求和.

这里的概率分布函数满足如下的特征:

(1) $F_X(-\infty) = 0, F_X(+\infty) = 1$;

(2) $F_X(x)$ 右连续: $F_X(x^+) = F_X(x)$, 并且是阶梯函数.

Takeaway Message

我们可以考察每个情况出现的概率.
概率分布的前缀和 (对每个 i, 对前 i 项求和构成的数列) 是概率分布函数, 概率分布函数的差分 (每相邻两项后一项减去前一项形成的数列, 且第 0 号数字为 0) 是概率分布.

6.1 经典的离散型随机变量

有了上一章的一番定义之后, 我们来看几个离散型随机变量.

(一) 0-1 分布 (两点分布)

例子 6.1. 100 件产品中, 有 98 件是正品, 2 件是次品, 今从中随机地抽取一件, 若规定

$$X = \begin{cases} 1, & \text{取到正品;} \\ 0, & \text{取到次品;} \end{cases}$$

则随机变量 X 的概率分布表为

X	0	1
P	0.02	0.98

.

定义 6.3 (两点分布 (0-1 分布))**.** 若随机变量 X 的概率分布为

X	0	1
P	$1-p$	p

,

则称 X 服从参数为 p 的两点分布（或 0-1 分布）. 记为

$$X \sim B(1,p)$$

(二-ε) 几何分布

例子 6.2. *抛一枚硬币, 硬币出现正面的概率为 p, 请问前 $k-1$ 次抛出反面, 第 k 次出现正面的概率.*

定义 6.4. 若随机变量 X 的概率分布为

$$P(X=k) = p(1-p)^{k-1}, \quad k = 1,2,3\cdots,$$

则称 X 服从参数为 p 的几何分布, 记为 $X \sim G(p)$.

几何分布是接下来的例子要说的二项分布, 当次数恰好等于 1 时的特例.

(二) 二项分布

例子 6.3. *若某射手每次射击命中的概率均为 p, 现进行 n 次独立射击, 求恰有 k 次命中的概率.*

先研究射击次数 $n=4$ 的特殊情形. 此时有

$k=0$	✗✗✗✗
$k=1$	✓✗✗✗, ✗✓✗✗, ✗✗✓✗, ✗✗✗✓
$k=2$	✓✓✗✗, ✓✗✓✗, ✓✗✗✓, ✗✓✓✗, ✗✓✗✓, ✗✗✓✓
$k=3$	✓✓✓✗, ✓✓✗✓, ✓✗✓✓, ✗✓✓✓
$k=4$	✓✓✓✓

定义 6.5 (n 重 Bernoulli 实验)**.** 只有两种可能结果的试验称为 Bernoulli 试验. 将一 Bernoulli 试验独立重复 n 次称为 n 重 Bernoulli 试验.

我们发现, 上面的问题就是 k 重 Bernoulli 实验恰好成功 k 次的概率. 根据计数的技巧, 我们发上发现:

定理 6.6 (Bernoulli 定理)**.** 设一次试验中事件 A 发生的概率为 $p(0<p<1)$, 则 n 重 Bernoulli 试验中, 事件 A 恰好发生 $k(0 \le k \le n)$ 次的概率为

$$b(k;n,p) := \binom{n}{k} p^k (1-p)^{n-k}.$$

对二项分布 $B(n,p)$, 当 n 充分大、p 很小时, (但是保证 $np_n=\lambda$ 不变) 形成的函数曲线是什么?

$$\begin{aligned}
&\binom{n}{k}p_n^k(1-p_n)^{n-k}\\
=&\frac{n(n-1)\cdots(n-k+1)}{k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}\\
=&\frac{\lambda^k}{k!}\Big(1\cdot(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})\Big)(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k}\\
\to&\frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\cdot 1
\end{aligned}$$

因此我们发现, 当 $n\to\infty$ 的时候, 有

定理 6.7. 设 $\lambda>0$ 为一个常数, n 为任意正整数, 且 $np_n=\lambda$, 则对任意一个固定的非负整数 k,

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

这就是 Poisson 分布.

(三) Poisson 分布 Poisson 分布的定义如下.

定义 6.8 (Poisson 分布)**.** 如果随机变量 X 服从以下分布律

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

其中 $\lambda>0$, 则称 X 服从参数为 λ 的 Poisson 分布, 记为 $X\sim P(\lambda)$.

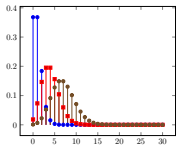

Figure 2: Poisson 分布

图 2展示了不同 λ 对应的分布情况. 蓝色, 红色, 棕色分别对应的是 $\lambda=1,4,7$ 的情形.

Poisson 分布常与单位时间（或单位面积、单位产品等）上的计数过程相联系.

Takeaway Message

三个重要离散型随机变量的分布:

- 几何分布

$$X \sim G(p) : P(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, 3 \cdots$$

- 二项分布

$$X \sim b(k; n) : P(X = p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Poisson 分布:

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

7 连续型随机变量

连续型随机变量 X 的取值范围为一个区间. 此时 X 取某个确定值的概率总是等于零. 我们发现事件 $X = a$ 可能发生, 但 $P(X = a) = 0$.

这就启发我们使用极限的想法来考虑之. X 在某个小区间内取值的概率可以大于零. 将区间分为若干段, 研究 X 在各小区间取值的概率. 分组的频率直方图刻画随机变量的概率分布, 分组越细, 频率直方图就越接近一条连续曲线. 取极限, 就有:

定义 7.1. 概率密度函数 如果存在非负函数 $f(x)$, 满足

$$P(X \leq x) = \int_{-\infty}^{x} f(t)dt,$$

则称 X 为连续型随机变量, 称 $f(x)$ 为 X 的概率密度函数 (pdf, probability density function).

和概率分布函数一样, 具有如下的性质:

$$(1)\, f(x) \geq 0, \quad (2) \int_{-\infty}^{+\infty} f(x)dx = 1.$$

这样确实可以来表述问题. 但是对于一半离散, 一半连续的随机变量, 该怎么统一起来? 实际上, 我们希望 “忽略” 在转折点上面带来的影响. 这样, 我们可以使用上面的分布函数的思想进行.

定义 7.2. 对任意随机变量 X（离散或连续），称函数

$$F_X(x) := P(X \le x), \quad x \in \mathbb{R}$$

为 X 的分布函数.

这是一个对于概率密度函数 (或者概率分布) 的累计, 自然, 它应该满足如下的性质:

- 对任意实数 $a < b$, 总有 $F(a) \le F(b)$;
- $0 \le F(x) \le 1$;
- $F(-\infty) = 0$, $F(+\infty) = 1$.

分别对应着: 概率不可能为负数; 所有的概率必须在 0 和 1 之间; 并且所有的可能性加起来必须等于 1(整个 Ω 的概率).

Takeaway Message

我们可以考察每个小区间出现的概率.
概率密度函数的积分是概率分布函数, 概率分布函数求导得到概率密度函数.
不论是离散型的或非离散型的随机变量 X, 都可以借助分布函数 $F(x) = P(X \le x), \quad -\infty < x < \infty$ 来描述.

7.1 常见的连续性随机变量

(一) 均匀分布 这可能是最基本的一个分布, 这个分布在一个区间内每个点的取值相同:

定义 7.3. 若随机变量 X 有概率密度

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

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

(二) 指数分布 假设某医院平均每小时出生 λ 个婴儿, 也就是说

- 1 小时内出生婴儿个数 $N(1)$ 服从 Poisson 分布 $P(\lambda)$.
- t 小时内出生婴儿个数 $N(t)$ 服从 Poisson 分布 $P(\lambda t)$.

我们希望研究婴儿出生的时间间隔 X 的概率分布.

当 $t<0$ 时, 有 $F(t)=P(X\leq t)=0$. 当 $t\geq 0$ 时,

$$\begin{aligned} F(t)&=P(X\leq t\}=1-P\{X>t)\\ &=1-P(N(t)=0)=1-e^{-\lambda t}. \end{aligned}$$

因此, 当 $t<0$ 时, $f(t)=0$; 当 $t\geq 0$ 时, $f(t)=\lambda e^{-\lambda t}$.
基于此, 我们给出指数分布的描述:

定义 7.4. 如果随机变量 X 有以下概率密度

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

其中 $\lambda>0$, 则称 X 服从参数为 λ 的指数分布, 记为

$$\boldsymbol{X}\sim \mathrm{Exp}(\lambda).$$

其分布函数为

$$F(x)=\begin{cases}1-e^{-\lambda x} & x\geq 0\\ 0 & x<0\end{cases}.$$

指数分布经常作为时间间隔或等待时间的分布. 其有一个非常重要的特性: 无记忆性.

命题 7.5 (指数分布的无记忆性)**.** 随机变量 X 服从参数为 λ 的指数分布. 设 $s,t>0$, 即有

$$P(X>s+t|X>s\}=P\{X>t).$$

Proof. 证明无记忆性, 只要证明 $\forall s,t\geq 0$, 有 $P(X\leq s\}=P(X\leq s+t\mid X\geq t\}$ 容易计算, $P(X\leq s+t\mid X\geq t\}=\frac{P\{t<X\leq s+t)}{P\{X\geq t)}=\frac{\int_1^{s+t}\lambda e^{-\lambda x}dx}{\int_t^{+\infty}\lambda e^{-\lambda x}dx}=\frac{e^{-\lambda t}-e^{-\lambda(s+t)}}{e^{-\lambda t}}=1-e^{-\lambda s}$
$P\{X\leq s)=\int_0^s\lambda e^{-\lambda x}dx=1-e^{-\lambda s}$ □

(三) 正态分布 在《高等数学 II》中我们使用极坐标的换元法了解到了如下的事实:

$$\int_{-\infty}^{+\infty}e^{-x^2}dx=\sqrt{\pi}$$

实际上, 这个性质比较好的函数稍加改造便可得到一个非常重要的分布函数 – 正态分布函数. 我们在后面会发现, 当样本总体是大量的随机变量求和的时候, 分布将不可避免地趋向于正态分布.

定义 7.6. 如果随机变量 X 有以下概率密度

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

其中 μ,σ 为常数且 $\sigma>0$, 则称 X 服从正态分布, 简记为

$$X\sim N(\mu,\sigma^2).$$

称 $N(0,1)$ 为标准正态分布.

带入数据, 我们发现标准正态分布的概率密度函数 (PDF) 为

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},$$

前面奇怪的系数是保证它在做积分的时候满足规范化的条件.

在这个问题中, 其分布函数

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

是没有显式的积分结果的. 通常这个结果需要查表得到.

如果发现一个分布不是标准正态分布, 可以用线性变换把它变成标准正态分布.

命题 7.7. 随机变量 $X\sim N(\mu,\sigma^2)$, 则 $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$.

Proof. $Z=\frac{X-\mu}{\sigma}$ 的分布函数为

$$\begin{aligned}P(Z\le x\}&=P\left\{(X-\mu)/\sigma\le x\right\}=P\{X\le\mu+\sigma x)\\&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\mu+\alpha t}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}\,\mathrm{d}t\end{aligned}$$

做变量代换, 令 $\frac{t-\mu}{\sigma}:=u$, 得

$$P(Z\le x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-u^2/2}\,\mathrm{d}u=\Phi(x),$$

由此知 $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$. □

Takeaway Message

常见的连续性概率分布:

- 均匀分布: $X \sim U(a,b)$: 的 PDF 是

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

- 指数分布 $X \sim \text{Exp}(\lambda)$ 的 PDF 是

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- 正态分布 $X \sim N(\mu, \sigma)$ 的 PDF 是

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

7.2 条件概率密度函数

一个概率密度函数的条件概率我们应当如何定义? 使用微积分的方法, 我们不研究单个点上面的情况, 我们转而研究一个 “小区间”.

研究在给定 $\xi < X_1 \leq \xi + h$ 下事件 $X_2 \leq \eta$ 的条件概率. 也就是

$$\mathrm{P}\left\{X_2 \leq \eta \mid \xi < X_1 \leq \xi + h\right\} = \frac{\int_{\xi}^{\xi+h} \int_{-\infty}^{\eta} f(x,y)\mathrm{d}y \ \mathrm{d}x}{\int_{\xi}^{\xi+h} f_X(x)\mathrm{d}x}$$

, 把分子和分母同除以 h, 当 $h \to 0$ 时, 右边趋于

$$U_{\xi}(\eta) = \frac{1}{f_X(\xi)} \int_{-\infty}^{\eta} f(\xi, y)\mathrm{d}y$$

所以, 对于固定的 ξ, 这是 η 的一个分布函数, 它的密度为

$$u_{\xi}(\eta) = \frac{1}{f_X(\xi)} f(\xi, \eta).$$

我们把这个定义做这种情形下的条件概率.

8 随机变量的函数的分布

在物理实验中, 我们希望测量直径 d, 从而得到圆的面积 $S = \frac{1}{4}\pi d^2$. 但是测量是有误差的. 这时候我们发现, 随机变量 S 是随机变量 d 的函数. 记得随机变量是定义在样本空间上的函数; 而这样的一个样本空间上面的函数又可以定义另一个函数. 这就是我们做抽象的好处.

更一般地, 我们希望找到如下问题的一类解法:

- 已知的随机变量 X 的概率分布
- 去求得它的函数 $Y = g(X)$ (g 连续, 且已知)

首先来看一个离散情形下的例子. 这里我们主要演示这件事情主要在做什么事情.

例子 8.1. *设随机变量 X 具有以下的分布律, 试求 $Y = (X-1)^2$ 的分布律.*

X	-1	0	1	2
p_k	0.2	0.3	0.1	0.4

解答: Y 所有可能取的值为 $0, 1, 4$. 由

$$\begin{aligned} P\{Y=0\} &= P\left\{(X-1)^2=0\right\} = P\{X=1\} = 0.1, \\ P\{Y=1\} &= P\{X=0\} + P\{X=2\} = 0.7, \\ P\{Y=4\} &= P\{X=-1\} = 0.2, \end{aligned}$$

所以 Y 的分布律为:

Y	0	1	4
p_k	0.1	0.7	0.2

再一个简单的例子, 有别于离散的情况可以一个点一个点的考虑, 我们可以使用分布函数这样的累计量来求积分得到.

例子 8.2. *设随机变量为 $X \sim N(0,1)$, 求 $Y = e^X$ 的概率密度函数. ($N(0,1)$ 意味着 $X \sim \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$)*

解答: 当 $y \leq 0$ 时,

$$F_Y(y) = P\{Y \leq y\} = 0$$

故,

$$f_Y(y) = 0$$

当 $y > 0$ 时,

$$F_Y(y) = P\{Y \leq y\} = P\{e^X \leq y\} = P\{X \leq \ln^y\} = F_X(\ln^y)$$

对两端对 x 求导, 得到

$$f_Y(y) = f_X(\ln^y)(\ln^y)' = \frac{1}{y}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\ln^y)^2}$$

我们发现在这种情形下, 有我们可以用如下的步骤清楚地表述我们所做的事情:

定理 8.1 (两个随机变量的联系 (关系是严格单调函数))**.** 设随机变量 X 具有概率密度 $f_X(x), -\infty < x < \infty$, 又设函数 g(x) 处处可导且恒有 $g'(x) > 0$(或恒有 $g'(x) < 0$), 则 $Y = g(X)$ 是连续型随机变量, 其概率密度为

$$f_Y(y) = \begin{cases} f_X[h(x)] \mid h'(x) \mid, & \alpha < y < \beta, \\ 0, & \text{其他}, \end{cases}$$

其中 $\alpha = \min\{g(-\infty), g(+\infty)\}, \beta = \max\{g(-\infty), g(+\infty)\}$,$h(y)$ 是 $g(x)$ 的反函数.

Proof. 我们只证 $g'(x) > 0$ 的情况, 此时 $g(x)$ 在 $(-\infty, \infty)$ 内严格单调增加, 它的反函数 $h(y)$ 存在, 且在 (α, β) 内严格单调增加、可导.

分别记 X,Y 的分布函数为 $F_X(x), F_Y(y)$. 现在先来求 Y 的分布函数 $F_Y(y)$.

因为 $Y = g(X)$ 在 (α, β) 内取值, 故当 $y \le \alpha$ 时, $F_Y(y) = P\{Y \le y\} = 0$; 当 $y \ge \beta$ 时, $F_Y(y) = P\{Y \le y\} = 1$.

当 $\alpha < y < \beta$ 时,

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} = P\{g(X) \le y\} \\ &= P\{X \le h(y)\} = F_X[h(y)] \end{aligned}$$

将 $F_Y(y)$ 关于 y 求导数, 即得 Y 的概率密度.

$$f_Y(y) = \begin{cases} f_X[h(y)]h'(y), & \alpha < y < \beta, \\ 0, & \text{其他}. \end{cases} \tag{1}$$

对于 $g'(x) < 0$ 的情况可以同样地证明, 此时有

$$\begin{cases} f_Y(y) = f_X[h(y)][-h'(y)], & \alpha < y < \beta, \\ 0, & \text{其他}. \end{cases} \tag{2}$$

合并 (1)(2) 式可证定理成立. □

那么, 在不单调的情况下, 我们应该怎么办呢? 实际上, 我们应该分类讨论, 一个一个的来看. 如下所示.

例子 8.3. 设随机变量为 $X \sim N(0,1)$,$Y = 2X^2 + 1$ 的概率密度函数.

解答: 当 $y < 1$ 时,

$$F_Y(y) = P\{Y \le y\} = 0, f_Y(y) = 0$$

当 $y \geq 1$ 时,

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\left\{2X^2 + 1 \leq y\right\} \\ &= P\left\{-\sqrt{\frac{y-1}{2}} \leq X \leq \sqrt{\frac{y-1}{2}}\right\} \\ &= \Phi\left(\sqrt{\frac{y-1}{2}}\right) - \Phi\left(-\sqrt{\frac{y-1}{2}}\right) \\ &= 2\Phi\left(\sqrt{\frac{y-1}{2}}\right) - 1 \\ f_Y(y) &= \frac{\mathrm{d}[2\Phi(\sqrt{\frac{y-1}{2}}) - 1]}{\mathrm{d}y} = \frac{1}{2\sqrt{\pi(y-1)}} e^{\frac{-(y-1)}{4}} \end{aligned}$$

例子 8.4. 设随机变量 X 在区间 $(0,1)$ 上服从均匀分布, 求 $Y = \frac{1}{1+X}$ 的概率密度.

解答:

$$\begin{aligned} F_Y(y) &= P\{Y \leq y\} = P\left\{\frac{1}{X+1} \leq y\right\} \\ &= P\left\{X \geq \frac{1-y}{y}\right\} = 1 - P\left\{X \leq \frac{1-y}{y}\right\} \\ &= 1 - F_X\left(\frac{1-y}{y}\right) \end{aligned}$$

对上式两边求导得:

$$f_Y(y) = -f_X\left(\frac{1-y}{y}\right)\left(\frac{1-y}{y}\right)' = \begin{cases} \frac{1}{y^2}, & \frac{1}{2} < y < 1, \\ 0, & \text{其他} \end{cases}$$

Part III

多维随机变量及其分布

9 二维随机变量及其分布函数

有时候样本空间不一定受一维量的影响. 例如研究学龄前的儿童发育情况, 对这一地区的儿童进行抽查. 比如观察到: 身高 H, 体重 W. 样本空间: $S = \{e\} = \{$ 某地区的全部学龄前儿童 $\}$, 那么 $H(e)$ 和 $W(e)$ 是定义在 S 上的两个随机变量. 又如观察炮弹着陆点, 是由横坐标 x, 纵坐标 y 确定的. 那么, 对于二维的随机变量, 我们应该如何分析?

定义 9.1 (二维随机变量)**.** 一般地, 设 E 是一个随机试验, 它的样本空间是 $S=\{e\}$, 设 $X=X(e)$ 和 $Y=Y(e)$ 是定义在 S 上的随机变量, 由它们构成的一个向量 (X,Y), 叫做**二维随机向量**或**二维随机变量**. 第二章讨论的随机变量也叫一维随机变量.

最终的二维随机变量 (X,Y):

- 与 X 有关, 与 Y 有关
- 与 X,Y **之间的关系**有关

因此通常将 (X,Y) 作为一个整体来研究.

注: 这个实际上可以推广. 对于 n 维的随机变量也是有类似的记号.

像上面的例子一样, 我们同样可以定义随机变量的分布函数, 表示累积量的关系, 便于后面的问题分析.

定义 9.2 (多维变量的分布函数)**.** 设 (X,Y) 是二维随机变量, 对于任意实数 x,y, 二元函数:

$$F(x,y)=P((X\leq x)\cap(Y\leq y)\}:=P\{X\leq x,Y\leq y)$$

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

二维的情形下, 如何求出落入小区间的概率? 即: 随机点 (X,Y) 落在矩形域 $\{(x,y)\mid x_1<x\leq x_2,y_1<y\leq y_2\}$ 的概率是什么? 根据图像:

$$\begin{aligned}&P\left\{x_1<X\leq x_2,y_1<Y\leq y_2\right\}\\&\quad=F\left(x_2,y_2\right)-F\left(x_2,y_1\right)+F\left(x_1,y_1\right)-F\left(x_1,y_2\right).\end{aligned}$$

9.1 分布函数的性质

像一维的时候那样, 同样发现分布函数要满足这些基本的性质:

命题 9.3 (分布函数的性质)**.** 分布函数满足如下的性质:

- **单调**.$F(x,y)$ 是变量 x 和 y 的不减函数
  - 对于任意固定的 y, 当 $x_2>x_1$ 时 $F\left(x_2,y\right)\geqslant F\left(x_1,y\right)$;
  - 对于任意固定的 x, 当 $y_2>y_1$ 时 $F\left(x,y_2\right)\geqslant F\left(x,y_1\right)$.
- $0\leq F(x,y)\leq 1$
  - $\forall$ 固定的 $y,F(-\infty,y)=0$,
  - $\forall$ 固定的 $x,F(x,-\infty)=0$,
  - $F(-\infty,-\infty)=0,F(\infty,\infty)=1$.

- 右连续.$F(x+0,y)=F(x,y), F(x,y+0)=F(x,y)$, 即 $F(x,y)$ 关于 x 右连续, 关于 y 也右连续.
- 非负. 对于任意 $(x_1,y_1),(x_2,y_2),x_1<x_2,y_1<y_2$,$F(x_2,y_2)-F(x_2,y_1)+F(x_1,y_1)-F(x_1,y_2)\geqslant 0$.

9.2 离散型随机变量

我们先从离散的情形看到一些概念:

定义 9.4 (离散型随机变量)**.** 如果二维随机变量 (X,Y) 全部可能取到的值是有限对或可列无限多对, 则称 (X,Y) 是离散型的随机变量.

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

我们同样希望使用表格来 “枚举” 每一种情况的概率. 这种表格我们称为 “联合分布律”.

定义 9.5. 设二维离散型随机变量 (X,Y) 所有可能取的值为 $(x_i,y_j),i,j=1,2,\cdots$, 记 $P\{X=x_i,Y=y_j\}=p_{ij},i,j=1,2,\cdots$, 由概率的定义有

$$p_{ij}\geqslant 0,\quad \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}p_{ij}=1.$$

称 $P\{X=x_i,Y=y_j\}=p_{ij},i,j=1,2,\cdots$ 为二维离散型随机变量 (X,Y) 的分布律, 或随机变量 X 和 Y 的联合分布律.

实际上, 这种问题我们可以在未来处理连续性问题的时候带来一个平滑的转化.

例子 9.1. 设随机变量 X 在 $1,2,3,4$ 四个整数中等可能地取一个值, 另一个随机变量 Y 在 $1\sim X$ 中等可能地取一整数值. 试求 (X,Y) 的分布律.

解答: 可由乘法公式求得 (X,Y) 的分布律.

$$P(X=i,Y=j\}=P(Y=j\mid X=i)P\{X=i\}=\frac{1}{i}\cdot\frac{1}{4},$$
$$i=1,2,3,4,j\leq i.$$

于是 (X,Y) 的分布律为

$Y\backslash X$	1	2	3	4
1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{16}$
2	0	$\frac{1}{8}$	$\frac{1}{12}$	$\frac{1}{16}$
3	0	0	$\frac{1}{12}$	$\frac{1}{16}$
4	0	0	0	$\frac{1}{16}$

有了联合分布律, 自然可以得到联合分布函数. 按照定义 9.2写一下就会发现,

$$F(x,y)=\sum_{x_i\leq x}\sum_{y_j\leq y}p_{ij},$$

下面我们把上面的定义 9.2推广到连续情形.

9.3 二维连续型随机变量

定义 9.6. 对于二维随机变量 (X,Y) 的分布函数 $F(x,y)$, 如果存在非负的函数 $f(x,y)$ 使对于任意 x,y 有

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

则称 (X,Y) 是连续型的二维随机变量, 函数 $f(x,y)$ 称为二维随机变量 (X,Y) 的*概率密度*, 或称为随机变量 X 和 Y 的*联合概率密度*. F 称为随机变量 X,Y 的联合分布函数.

观察到, 只是上述的 $\sum$ 换成了 $\int$. 而且它同样遵循一些换元规则. 我们可以使用矩阵的记号描述之.

和离散的情形一样, 联合分布函数也满足一些基本性质 – 和命题 9.3所指示的基本类似.

- *非负*. $f(x,y)\geqslant 0$.
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(u,v)\mathrm{d}u\ \mathrm{d}v=F(\infty,\infty)=1$.
- 设 G 是 xOy 平面上的区域, 点 (X,Y) 落在 G 内的概率为

$$P((X,Y)\in G)=\iint_{G}f(u,v)\mathrm{d}u\ \mathrm{d}v.$$

- 若 $f(x,y)$ 在点 (x,y) 连续, 则有

$$\frac{\partial^2 F(x,y)}{\partial x\partial y}=f(x,y).$$

实际上, 上述的第 (4) 条性质有一个直观的解释: 若 $f(x,y)$ 在点 (x,y) 处连续, 则当 $\Delta x,\Delta y$ 很小时 $P(x<X\leq x+\Delta x,y<Y\leq y+\Delta y\}\approx f(x,y)\Delta x\Delta y$. 对于一个联合密度函数, 如果其表达式为 $z=f(x,y)$, 直观来看就是表示空间中的一个曲面. 这个

曲面有一些特别之处: 介于它和 xOy 平面的空间区域的体积为 1. 并且概率分布的值就是 $P\{(X,Y)\in G)$ 的值等于以 G 为底, 以曲面 $z=f(x,y)$ 为顶面的柱体体积.

使用极限的语言也容易证明上述的第四条性质:

$$\begin{aligned}
&\lim_{\substack{\Delta x\to 0^+\\ \Delta y\to 0^+}} \frac{P(x<X\le x+\Delta x, y<Y\le y+\Delta y)}{\Delta x\Delta y}\\
&=\lim_{\substack{\Delta x\to 0^+\\ \Delta y\to 0^+}} \frac{1}{\Delta x\Delta y}[F(x+\Delta x,y+\Delta y)-F(x+\Delta x,y)\\
&\quad -F(x,y+\Delta y)+F(x,y)]\\
&=\frac{\partial^2 F(x,y)}{\partial x\partial y}=f(x,y)
\end{aligned}$$

注: * 在微积分中, 我们了解了换元法. 并且很多时候它可以带给我们方便. 这里我们同样介绍类似的方法. 例如 $P\left\{a_1<X_1\le b_1, a_2<X_2\le b_2\right\}=\int_{a_1}^{b_1}\int_{a_2}^{b_2} f\left(x_1,x_2\right)\mathrm{d}x_1\ \mathrm{d}x_2$. 有时候为了方便起见, 我们会进行变量替换: $X_1=a_{11}Y_1+a_{12}Y_2,\quad X_2=a_{21}Y_1+a_{22}Y_2$. 这个变量替换是可逆的, 因为其行列式 $\Delta=a_{11}a_{22}-a_{12}a_{21}\neq 0$.

带入这样的变量代换, 就将原来的区域变换为了 $\mathrm{P}\{\Omega\}=\iint_{\Omega_*} f\left(a_{11}y_1+a_{12}y_2, a_{21}y_1+a_{22}y_2\right)\cdot \Delta\mathrm{d}y_1\ \mathrm{d}y_2$

区域 Ω_* 由所有这样的点 (y_1,y_2) 组成: 它的像 (x_1,x_2) 在 Ω 中. 因为事件 $(X_1,X_2)\in\Omega$ 和 $(Y_1,Y_2)\in\Omega_*$ 是相等的, 由此 (Y_1,Y_2) 的联合密度由下式给出:

$$g\left(y_1,y_2\right)=f\left(a_{11}y_1+a_{12}y_2, a_{21}y_1+a_{22}y_2\right)\cdot\Delta$$

实际上, 像上面的内容进行的变量代换可以写作线性方程组的形式. 行向量的利用需要把由 $\mathbf{R}^r$ 到 $\mathbf{R}^m$ 的线性变换写成形式

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{A}$$

即

$$y_k=\sum_{j=1}^{r} a_{j_k}x_j\quad k=1,\cdots,m$$

因此, 引入矩阵记号对我们有很大的帮助.

Aside

回顾矩阵的一些基本内容:

- 访问: 一个矩阵在访问元素的时候, 首先是行, 其次是列. 因此, $\alpha\times\beta$ 矩阵 $\boldsymbol{A}$ 含有 α 行和 β 列, 它的元素记为 a_{j_k}, 第 1 个下标表示行, 第 2 个下标表示列. 可以认为表现为一个线性方程组.
- 矩阵乘法: 类似于线性方程组的代换. 一般的规则是: 如果有 $\boldsymbol{A}$ 含有 α 行和 β 列, 其元素为 a_{j_k}; 如果 $\boldsymbol{B}$ 是含有元素 b_{j_k} 的 $\beta\times\gamma$ 矩阵乘积 AB 是

含有元素 $a_{j_1}b_{1k}+a_{j_2}b_{2k}+\cdots+a_{j_\beta}b_{\beta k}$ 的 $\alpha\times\gamma$ 矩阵.

– 不一定满足交换律, 但是满足结合律.

- 只含有一行的 $1\times\alpha$ 矩阵称为行向量, 只含有一列的矩阵称为列向量.

矩阵的内积: 两个行向量 $\boldsymbol{x}=(x_1,\cdots,x_\alpha)$ 和 $\boldsymbol{y}=(y_1,\cdots,y_\alpha)$ 的内积定义为

$$\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}=\boldsymbol{y}\boldsymbol{x}^{\mathrm{T}}=\sum_{j=1}^{\alpha}x_jy_j$$

二次型: 如果 $a_{jk}=a_{kj}$, 即 $\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}$, 则方阵 $\boldsymbol{A}$ 是对称的, 与对称的 $r\times r$ 矩阵 $\boldsymbol{A}$ 有关的二次型定义为

$$\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^{\mathrm{T}}=\sum_{j,k=1}^{r}a_{jk}x_jx_k,$$

其中 $x_1,\cdots,x_r$ 是未定的. 如果对于所有的非零向量 $\boldsymbol{x}$ 有 $\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^{\mathrm{T}}>0$, 是称矩阵 $\boldsymbol{A}$ 是正定的. 由上述准则推出, 正定矩阵是可逆的.

例子 9.2. 设二维随机变量 (X,Y) 具有概率密度

$$f(x,y)=\begin{cases}2\mathrm{e}^{-(2x+y)}, & x>0,y>0,\\ 0, & \text{其他}.\end{cases}$$

(1) 求分布函数 $F(x,y)$*; (2)* 求概率 $P(Y\leq X)$.

解答: (1)
$$\begin{aligned}F(x,y)&=\int_{-\infty}^{y}\int_{-\infty}^{x}f(u,v)\mathrm{d}u\ \mathrm{d}v\\&=\begin{cases}\int_0^y\int_0^x 2\mathrm{e}^{-(2u+v)}\mathrm{d}u\ \mathrm{d}v, & x>0,y>0,\\ 0, & \text{其他}.\end{cases}\end{aligned}$$

即有 $F(x,y)=\begin{cases}\left(1-\mathrm{e}^{-2x}\right)\left(1-\mathrm{e}^{-y}\right), & x>0,y>0,\\ 0, & \text{其他}.\end{cases}$

(2) 将 (X,Y) 看作是平面上随机点的坐标. 即有 $\{Y\leq X\}=\{(X,Y)\in G\}$, 其中 G 为 xOy 平面上直线 $y=x$ 及其下方的部分. 于是

$$\begin{aligned}P(Y\leq X\}=P\{(X,Y)\in G)&=\iint_G f(u,v)\mathrm{d}u\ \mathrm{d}v\\&=\int_0^\infty\int_y^\infty 2\mathrm{e}^{-(2u+v)}\mathrm{d}u\ \mathrm{d}v=\frac{1}{3}.\end{aligned}$$

9.4 n 维随机变量及其分布函数

可以推广到 $n>2$ 的情形.

定义 9.7. 设 E 是一个随机试验, 它的样本空间是 $S=\{e\}$, 设 $X_1=X_1(e), X_2=X_2(e), \cdots, X_n=X_n(e)$ 是定义在 S 上的随机变量, 由它们构成的一个 n 维向量 $(X_1, X_2, \cdots, X_n)$ 叫做 n 维随机向量或 n 维随机变量.

定义 9.8. 对于任意 n 个实数 $x_1, x_2, \cdots, x_n$, n 元函数

$$F(x_1, x_2, \cdots, x_n)=P\{X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n\}$$

称为 n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 的分布函数或随机变量 $X_1, X_2, \cdots, X_n$ 的联合分布函数.

其具有类似于二维随机变量的分布函数的性质.

把分布函数的定义推广到 n 维空间, 就给出了如下的定义.

定义 9.9. n 维随机变量的 $(X_1, X_2, \ldots, X_n)$ 的分布函数定义为

$$F(x_1, x_2, \ldots, x_n)=P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n),$$

其中 $x_1, x_2, \ldots, x_n$ 为任意实数.

若存在非负可积函数 $f(x_1, x_2, \ldots, x_n)$, 使对于任意实数 $x_1, x_2, \ldots, x_n$ 有

$$F(x_1, x_2, \ldots, x_n)=\int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}}\cdots\int_{-\infty}^{x_1} f(x_1, x_2, \ldots, x_n)dx_1dx_2\ldots dx_n,$$

则称 $f(x_1, x_2, \ldots, x_n)$ 为 $(X_1, X_2, \ldots, X_n)$ 的概率密度函数.

10 边缘分布

我们可以用多种角度来看二维随机变量及其分布函数:

- 作为整体: $F(x, y)$, 其中 X, Y 都是随机变量.
- 关注局部: X 和 Y 也有他们自己的分布函数. 记为 $F_X(x), F_Y(y)$.

我们 "选择性地" 忽略一个变量, 就得到了如下的定义:

定义 10.1 (边缘分布)**.** 二维随机变量 (X, Y) 作为一个整体, 具有分布函数 $F(x, y)$. 而 X 和 Y 都是随机变量, 各自也有分布函数, 将它们分别记为 $F_X(x), F_Y(y)$, 依次称为二维随机变量 (X, Y) 关于 X 和关于 Y 的**边缘分布函数**. 边缘分布函数可以由 (X, Y) 的分布函数 $F(x, y)$ 所确定, 并且 $F_X(x)=P(X \leq x\}=P\{X \leq x, Y<\infty)=F(x, \infty)$.

10.1 离散情形的边缘分布

根据定义 10.1和离散型变量的特点, 我们可以考虑离散状态下二维随机变量的边缘分布:

假设有二维离散型随机变量 (X,Y) , 其所有可能取的值为 $(x_i,y_j),i,j=1,2,\cdots$, 记 $P\{X=x_i,Y=y_j\}=p_{ij}$.

定义: 考察 X 的边缘分布:

$$F_X(x)=F(x,\infty)=\sum_{x_i\le x}\sum_{j=1}^{\infty}p_{ij}$$

- X 的分布律: $P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij},\quad i=1,2,\cdots$.
- Y 的分布律: $P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij},\quad j=1,2,\cdots$.

  为了方便, 可以引入如下的记号

$$p_{i\bullet}:=\sum_{j=1}^{\infty}p_{ij}=P\{X=x_i\},\quad i=1,2,\cdots,$$

$$p_{\bullet j}:=\sum_{i=1}^{\infty}p_{ij}=P\{Y=y_j\},\quad j=1,2,\cdots,$$

分别称 $p_{i\bullet}(i=1,2,\cdots)$ 和 $p_{\bullet j}(j=1,2,\cdots)$ 为 (X,Y) 关于 X 和关于 Y 的边缘分布律.

其中, 下标中的 $\bullet$ 类似通配符: $p_{i\bullet}$ 是由 p_{ij} 关于 j 求和后得到的, 反之亦然. 这样的记号将有助于理解连续的情形.

在一门学习了正则表达式的课程里面我们知道了 *. 我们可以确信概率论和正则表达式是很相关的（确信）

10.2 连续情形的边缘分布

定义: 对于连续型随机变量 (X,Y), 设它的概率密度为 $f(x,y)$, 由于

$$F_X(x)=F(x,\infty)=\int_{-\infty}^{x}\left[\int_{-\infty}^{\infty}f(u,v)\mathrm{d}v\right]\mathrm{d}u$$

- X 是一个连续型的随机变量, 其概率密度为 $f_X(x)=\int_{-\infty}^{\infty}f(x,v)\mathrm{d}v$.
- Y 是一个连续型的随机变量, 其概率密度为 $f_Y(y)=\int_{-\infty}^{\infty}f(u,y)\mathrm{d}u$.

  分别称 $f_X(x),f_Y(y)$ 为 (X,Y) 关于 X 和关于 Y 的边缘概率密度.

例子 10.1. 设随机变量 X 和 Y 具有联合概率密度

$$f(x,y)=\begin{cases}6, & x^2\le y\le x,\\ 0, & \text{其他}.\end{cases}$$

求边缘概率密度 $f_X(x),f_Y(y)$.

解答: 根据定义:

$$f_X(x)=\int_{-\infty}^{\infty} f(u,v)\mathrm{d}v=\begin{cases}\int_{x^2}^{x} 6\ \mathrm{d}v=6\left(x-x^2\right), 0\leq & x\leq 1,\\ 0, & \text{其他.}\end{cases}$$

$$f_Y(y)=\int_{-\infty}^{\infty} f(u,v)\mathrm{d}u=\begin{cases}\int_{y}^{\sqrt{y}} 6\ \mathrm{d}u=6(\sqrt{y}-y), & 0\leq y\leq 1,\\ 0, & \text{其他.}\end{cases}$$

最后一个问题是有边缘分布, 能推出联合分布吗? 事实上是不行的. 我们用正态分布做一个例子:

例子 10.2. 设二维随机变量 (X,Y) 的概率密度为

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\},$$

其中 $\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$ 都是常数, 且 $\sigma_1>0,\sigma_2>0,-1<\rho<1$. 我们称 (X,Y) 为服从参数为 $\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$ 的二维正态分布 (这五个参数的意义将在下一章说明), 记为 $(X,Y)\sim N\left(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho\right)$. 试求二维正态随机变量的边缘概率密度.

解答: 要计算 $f_X(x)=\int_{-\infty}^{\infty} f(x,y)\mathrm{d}y$, 由于

$$\frac{(y-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}=\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)^2-\rho^2\frac{(x-\mu_1)^2}{\sigma_1^2},$$

于是

$$f_X(x)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)^2}\ \mathrm{d}y$$

令

$$t:=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right)$$

则有

$$f_X(x)=\frac{1}{2\pi\sigma_1}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\int_{-\infty}^{\infty}\mathrm{e}^{-t^2/2}\ \mathrm{d}t$$

即

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},\quad -\infty<x<\infty,$$

同理有

$$f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_2}\mathrm{e}^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},\quad -\infty<y<\infty.$$

我们发现二维正态分布的两个边缘分布都是一维正态分布, 而且并不依赖 ρ. 因此, 边缘分布一般不能决定联合分布.

10.3 推广到 n 维

当然, 高维空间里面的边缘分布也是通过 “选择性忽略” 有些值定义的.

定义 10.2. 设 $(X_1, X_2, \ldots, X_n)$ 的分布函数 $F(x_1, x_2, \ldots, x_n)$ 为已知, 则 $(X_1, X_2, \ldots, X_n)$ 的 $k(1 \leq k < n)$ 维边缘分布函数就随之确定, 例如 $(X_1, X_2, \ldots, X_n)$ 关于 X_1、关于 (X_1, X_2) 的边缘分布函数分别为

$$F_{X_1}(x_1) = F(x_1, \infty, \infty, \ldots, \infty),$$

$$F_{X_1,X_2}(x_1, x_2) = F(x_1, x_2, \infty, \infty, \ldots, \infty).$$

又若 $f(x_1, x_2, \ldots, x_n)$ 是 $(X_1, X_2, \ldots, X_n)$ 的概率密度, 则 $(X_1, X_2, \ldots, X_n)$ 关于 X_1、关于 (X_1, X_2) 的边缘概率密度分别为

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_2 dx_3 \ldots dx_n.$$

11 相互独立的随机变量

在离散的情形, 我们定义两个事件是独立的当且仅当 $P(AB) = P(A)P(B)$. 那么对于随机变量, 我们也给出同样的定义:

定义 11.1 (相互独立的随机变量)**.** 设 $F_{X,Y}(x, y)$ 及 $F_X(x), F_Y(y)$ 分别是二维随机变量 (X, Y) 的分布函数及边缘分布函数. 若对于所有 x, y 有

$$P(X \leq x, Y \leq y\} = P(X \leq x)P\{Y \leq y),$$
$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

则称随机变量 X 和 Y 是相互独立的.

实际上, 刚刚的定义给了我们一点提示: 如果 (X, Y) 是连续型随机变量,$f_{X,Y}(x, y), f_X(x), f_Y(y)$ 是概率密度以及边缘概率密度, 那么:

- X 和 Y 相互独立 $\iff$ $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ 几乎处处成立.
- (除了在平面上面积为 0 的集合)

  也就是说, 如果 (X, Y) 是离散型随机变量:
- $P\{X = x_i, Y = y_j\} = P\{X = x_i\} P\{Y = y_j\}$.

11.1 推广到 n 维

对于多个事件的独立性, 我们知道任意的一个事件的子集都必须满足 $P(A_1 \cdots A_k) = P(A_1) \cdots P(A_n)$. 高维的情形也是类似的:

定义:

设 $(X_1, X_2, \ldots, X_n)$ 的分布函数 $F(x_1, x_2, \ldots, x_n)$ 为已知, 则 $(X_1, X_2, \ldots, X_n)$ 的 $k(1 \leq k < n)$ 维边缘分布函数就随之确定, 例如 $(X_1, X_2, \ldots, X_n)$ 关于 X_1、关于 (X_1, X_2) 的边缘分布函数分别为

$$F_{X_1}(x_1) = F(x_1, \infty, \infty, \ldots, \infty),$$

$$F_{X_1, X_2}(x_1, x_2) = F(x_1, x_2, \infty, \infty, \ldots, \infty)$$

. 又若 $f(x_1, x_2, \ldots, x_n)$ 是 $(X_1, X_2, \ldots, X_n)$ 的概率密度, 则 $(X_1, X_2, \ldots, X_n)$ 关于 X_1、关于 (X_1, X_2) 的边缘概率密度分别为

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_2 dx_3 \ldots dx_n,$$

$$f_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n) dx_3 dx_4 \ldots x_n,$$

若对于所有的 $x_1, x_2, \ldots, x_n$ 有

$$F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_n}(x_n),$$

则称 $X_1, X_2, \ldots, X_n$ 是相互独立的.
若对于所有的 $x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_n$ 有

$$F(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n) = F_1(x_1, x_2, \ldots, x_m) F_2(y_1, y_2, \ldots, y_n),$$

其中 F_1,F_2,F 依次为随机变量 $(X_1, X_2, \ldots, X_m)$,$(Y_1, Y_2, \ldots, Y_n)$ 和 $(X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n)$ 的分布函数, 则称随机变量 $(X_1, X_2, \ldots, X_m)$ 和 $(Y_1, Y_2, \ldots, Y_n)$ 是相互独立的.

定理 11.2. 设 $(X_1, X_2, \ldots, X_m)$ 和 $(Y_1, Y_2, \ldots, Y_n)$ 相互独立, 则 $X_i (i = 1, 2, \ldots, m)$ 和 $Y_j (j = 1, 2, \ldots, n)$ 相互独立, 又若 h, g 是连续函数, 则 $h(X_1, X_2, \ldots, X_m)$ 和 $g(Y_1, Y_2, \ldots, Y_n)$ 相互独立.

12 两个随机变量的函数的分布

我们先来看一个问题: X, Y 是相互独立的离散型随机变量, 等概率地取 $[0, 3]$ 区间的整数. 问 $X + Y = 3$ 的概率是多少?

解答也不难. 考虑一共有 $(0,3),(1,2),(2,1),(3,0)$ 四种情况, 相加即可. 那么我们把这个 $X+Y$ 看做新的随机变量, 应该如何求? 实际上我们只要枚举就好了.

$$\begin{aligned} P_W(w) &= P(X+Y=w) \\ &= \sum_x P(X=x)P(Y=w-x) \\ &= \sum_x P_X(x)P_Y(w-x). \end{aligned}$$

我们现在来系统的考察对于任意的分布函数, 几个常见的两个随机变量参与运算之后的概率密度.

(一) $Z=X+Y$ 的分布 设 (X,Y) 是二维连续型随机变量, 它具有概率密度 $f_{X,Y}(x,y)$. 那么连续型随机变量 $Z=X+Y$ 的概率密度是多少?

大致思路:

- 先来求 $Z=X+Y$ 的分布函数 $F_Z(z)$
- $F_Z(z)=P(Z\le z)=\iint_{x+y\le z} f_{X,Y}(x,y)\mathrm{d}x\ \mathrm{d}y$

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z-y} f(x,y)\mathrm{d}x\right]\mathrm{d}y \\ &\xlongequal{x:=u-y} \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z} f(u-y,y)\mathrm{d}u\right]\mathrm{d}y \\ &= \int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f(u-y,y)\mathrm{d}y\right]\mathrm{d}u \end{aligned}$$

这里的换元法实际上是使用给定的不等关系 $x+y\le z$ 为了消去 x, 减少变量个数.

根据定义, 求导得到: $f_{X+Y}(z)=\int_{-\infty}^{\infty} f(u-y,y)\mathrm{d}y=\int_{-\infty}^{\infty} f(z-y,y)\mathrm{d}y$.

我们用定理的形式总结这一事实:

定理 12.1. 设 (X,Y) 是二维连续型随机变量, 它具有概率密度 $f_{X,Y}(x,y)$. 则 $Z=X+Y$ 仍为连续型随机变量, 其概率密度为

$$f_{X+Y}(z)=\int_{-\infty}^{\infty} f(z-y,y)\mathrm{d}y$$

或

$$f_{X+Y}(z)=\int_{-\infty}^{\infty} f(x,z-x)\mathrm{d}x$$

如果 X,Y 相互独立的话, 上述公式可以进一步地写作

- $f_{X+Y}(z)=\int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\mathrm{d}y$;
- $f_{X+Y}(z)=\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\mathrm{d}x$

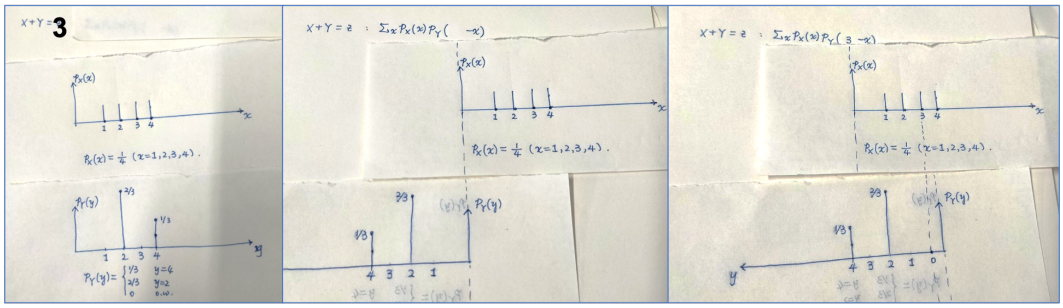

Figure 3: 直观理解卷积

Aside

卷积: 刚刚常见的操作其实有一个名字: 卷积. 比如我们在求两个多项式的乘积的时候, $(\sum_{i=0}^{\infty} a_i x^i)(\sum_{j=0}^{\infty} b_j x^j)$.
我们可以用这样的公式计算:

$$\sum_{k=0}^{\infty}\left(\sum_{i+j=k} a_i b_j\right) x^k = \sum_{k=0}^{\infty}\left(\sum_{i=0}^{k} a_i b_{k-i}\right) x^k$$

刚刚的那个问题同样和这个问题有类似的性质, 只是求和号变为了积分号.

为了方便起见, 这两个公式称为 f_X 和 f_Y 的卷积公式, 记为 $f_X * f_Y$, 即

$$f_X * f_Y = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\mathrm{d}y = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\mathrm{d}x$$

Aside

直观理解卷积: 先把纸片翻一下, 然后平移, 最后对应相乘相加. 如图 3.

(二)$Z = Y/X, Z = XY$ 的分布 设 (X,Y) 是二维连续型随机变量, 它具有概率密度 $f_{X,Y}(x,y)$. 连续型随机变量 $Z = Y/X$ 的概率密度是多少?

我们还是还是首先设出 $F_{Y/X}(z) = P(Y/X \le z)$. 但是这里由于函数的不连续性, 需要分两种情况: $x < 0, x > 0$ 分别考虑:

$$
\begin{aligned}
F_{Y/X}(z) = P(Y/X \le z) &= \iint_{\substack{y/x \le z \\ x<0}} f(x,y)\mathrm{d}y\ \mathrm{d}x + \iint_{\substack{y/x \le z \\ x>0}} f(x,y)\mathrm{d}y\ \mathrm{d}x \\
&= \int_{-\infty}^{0} \left[\int_{zx}^{\infty} f(x,y)\mathrm{d}y\right]\mathrm{d}x + \int_{0}^{\infty} \left[\int_{-\infty}^{zx} f(x,y)\mathrm{d}y\right]\mathrm{d}x \\
&\xlongequal{y:=xu} \int_{-\infty}^{0} \left[\int_{z}^{-\infty} xf(x,xu)\mathrm{d}u\right]\mathrm{d}x + \int_{0}^{\infty} \left[\int_{-\infty}^{z} xf(x,xu)\mathrm{d}u\right]\mathrm{d}x \\
&= \int_{-\infty}^{0} \left[\int_{-\infty}^{z} (-x)f(x,xu)\mathrm{d}u\right]\mathrm{d}x + \int_{0}^{\infty} \left[\int_{-\infty}^{z} xf(x,xu)\mathrm{d}u\right]\mathrm{d}x \\
&= \int_{-\infty}^{z} \left[\int_{-\infty}^{+} \infty|x|f(x,xu)\mathrm{d}u\right]\mathrm{d}x
\end{aligned}
$$

遵循同样的模式, 同样可以求出: $Z = XY$ 的概率分布.

$$
\begin{aligned}
&F_{XY}(z) = P(XY \le z) \\
&= \iint_{\substack{xy \le z \\ x<0}} f(x,y)\mathrm{d}y\mathrm{d}x + \iint_{\substack{xy \le z \\ x>0}} f(x,y)\mathrm{d}y\mathrm{d}x \\
&= \int_{-\infty}^{0} \left(\int_{z/x}^{+\infty} f(x,y)\mathrm{d}y\right)\mathrm{d}x + \int_{0}^{+\infty} \left(\int_{-\infty}^{z/x} f(x,y)\mathrm{d}y\right)\mathrm{d}x \\
&\xlongequal{y:=u/x} \int_{-\infty}^{0} \left(\int_{z/x}^{+\infty} f\left(x,\frac{u}{x}\right) d\left(\frac{u}{x}\right)\right)\mathrm{d}x + \int_{0}^{+\infty} \left(\int_{-\infty}^{z/x} f\left(x,\left(\frac{u}{x}\right)\mathrm{d}x\right)\right) \\
&= \int_{-\infty}^{0} \left(\left(\frac{1}{x}\right)\int_{z}^{-\infty} f\left(x,\frac{u}{x}\right)\mathrm{d}u\right)\mathrm{d}x + \int_{0}^{+\infty} \left(\frac{1}{x}\int_{-\infty}^{z} f\left(x,\frac{u}{x}\right)\mathrm{d}u\right)\mathrm{d}x \\
&= \int_{0}^{z} \left(\int_{-\infty}^{\infty} \frac{1}{|x|} f\left(x,\frac{u}{x}\right)\mathrm{d}u\right)\mathrm{d}x
\end{aligned}
$$

(三) $M = \min\{X,Y\}$ 的分布 X,Y 是两个相互独立的随机变量, 它们的分布函数分别为 $F_X(x)$ 和 $F_Y(y)$. 求 $M = \min\{X,Y\}$ 的分布函数.

$$
\begin{aligned}
F_{\min}(z) = P(N \le z\} &= 1 - P\{N > z) \\
&= 1 - P(X > z, Y > z\} = 1 - P(X > z) \cdot P\{Y > z)
\end{aligned}
$$

也就是

$$
F_{\min}(z) = 1 - [1 - F_X(z)]\,[1 - F_Y(z)]\,.
$$

上述的三个情况都可以推广到 n 维的情形.

Part IV

随机变量的数字特征

13 数学期望

有这样的一个游戏: 花 2 元并投掷一颗均匀的骰子. 如果事件 A $=\{1,2,3\}$ 发生, 收到 1 元. 如果事件 B $=\{4,5\}$ 发生, 收到 2 元. 如果事件 C $=\{6\}$ 发生, 收到 6 元. 你会参加这个游戏吗?

Web Demonstrate Aside

实际上, 真理元素的频道主实际上真的在路边做了这个实验. 可以参看他们的视频: Bilibili: BV1Xx411b7rM

可能我们的第一考虑是看看 "平均" 能得多少. 这样的随机变量 X,

$$X=\begin{cases}1 & \text{如果事件 A 发生}\\ 2 & \text{如果事件 B 发生}\\ 6 & \text{如果事件 C 发生}\end{cases}$$

事件 A、B、C 的概率分别是:

$$P(A)=\frac{3}{6},\quad P(B)=\frac{2}{6},\quad P(C)=\frac{1}{6}$$

那么, 求结果的平均值 $1\cdot P(A)+2\cdot P(B)+6\cdot P(C)=\frac{13}{6}$.

刚刚我们做的事情, 用更正式的语言, 实际上是就是有了一个 $(\Omega,\mathscr{A},P)$ 这样的有限概率空间, 而 $X=X(\omega)$ 是某一随机变量, 其值域为 $\{x_1,\cdots,x_k\}$. 如果设 $A_i=\{\omega:X(\omega)=x_i\}$, 则显然 X 可以表示为

$$X(\omega)=\sum_{i=1}^{k}x_iI\left(A_i\right)$$

记 $p_i=P\{X=x_i\}$. 直观上显然: 如果在 n 次独立重复试验中观测随机变量 X 的取值, 则取 x_i 的值大致应该出现 $np_i(i=1,\cdots,k)$ 次. 因此, 根据 n 次试验的结果, 计算的该随机变量的 "平均值" 大致为

$$\frac{1}{n}\left[np_1x_1+\cdots+np_kx_k\right]=\sum_{i=1}^{k}p_ix_i$$

下面的歌词阐述了我们在做选择的时候内心的权衡: ``可万一对了呢　会不会我会不会更快乐 可万一错了呢 这一切　我是否可以承受''
--《万一对了呢》ChiliChill

定义 13.1 (离散型随机变量的期望)**.** 设离散型随机变量 X 的分布律为

$$P\{X = x_k\} = p_k, \quad k = 1, 2, \cdots.$$

若级数

$$\sum_{k=1}^{\infty} x_k p_k$$

绝对收敛, 则称级数 $\sum_{k=1}^{\infty} x_k p_k$ 的和为随机变量 X 的数学期望, 记为 $\mathbb{E}(X)$. 即

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} x_k p_k$$

仿照离散型随机变量的期望的定义, 我们同样给出连续形随机变量的定义:

定义 13.2. 连续型随机变量的期望 设连续型随机变量 X 的概率密度为 $f(x)$, 若积分

$$\int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

绝对收敛, 则称积分 $\int_{-\infty}^{\infty} x f(x) \mathrm{d}x$ 的值为随机变量 X 的数学期望, 记为 $\mathbb{E}(X)$. 即

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

实际上, 连续型随机变量的数学期望有另一种理解. 看下面的例子:

例子 13.1. 连续性随机变量 $X \geq 0$, 其概率密度函数为 $f_X(x)$. 证明 $\mathbb{E}(X) = \int_0^{\infty} P(X > t)dt$.

证明直接展开交换积分次序即可.

下面我们看若干常见分布的数学期望.

13.1 常见分布的数学期望

Bernouli 分布 它描述的是只先进行一次事件试验, 该事件发生的概率为 p, 不发生的概率为 $1 - p$. 也就是

- $P(X = 1) = p$
- $P(X = 0) = 1 - p$

因此, $\mathbb{E}(X) = 0 \cdot (1 - p) + 1 \cdot p = p$.

Poisson 分布 它的分布是: 对于大于 0 的参数 λ,

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

期望的话, 就是 $\sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!}$. 我们可以使用《高等数学 II》中的公式计算它:

$$\begin{aligned} E(X) &= \lambda e^{-\lambda} \sum_{k\geq 1} \frac{1}{(k-1)!}\lambda^{k-1} && \text{去掉} k=0 \text{ 的那一项} \\ &= \lambda e^{-\lambda} \sum_{j\geq 0} \frac{\lambda^j}{j!} && \text{变量代换} j := k-1 \\ &= \lambda e^{-\lambda} e^{\lambda} && \text{指数函数的 Taylor 级数展开} \\ &= \lambda \end{aligned}$$

几何分布 回忆: 对与 $G(X=k)$ 的几何分布, 表达的意思是前 $k-1$ 次皆失败, 第 k 次成功的概率. 因此, 其概率是

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

所以它的期望表达式是:

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

由于 p 是常数, 我们可以把它从求和号中拿出来, 得到 $p\sum_{k=1}^{\infty} k(1-p)^{k-1}$

- 使用: $\frac{1}{1-x} = 1 + x + x^2 + \cdots = \sum_{k=0}^{\infty} x^k$, 当 $|x| < 1$ 时:
- 两端求导: $\frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} kx^{k-1}$.
- 令 $x = 1-p$, 得到 $\frac{1}{p^2} = \sum_{k=1}^{\infty} k(1-p)^{k-1}$
- 因此期望为 $1/p$.

实际上, 这个操作正是对应着它的无记忆性. 在介绍了期望的性质之后, 我们发现使用无记忆性来说明这个性质就不用如此大费周章了.

均匀分布 回顾: 均匀分布的概率密度函数

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

带入计算之后, 得到了一个很平凡的结论:

$$\int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}.$$

这也符合直觉 – 如果你往这里加一个支点的话, 它就会支撑起整个概率密度函数.

正态分布 回忆: 正态分布的概率密度函数

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

这个问题, 可以根据对称性进行求解. 注意到关于 μ 对称.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} xe^{-(x-\mu)^2/2\sigma^2}dx = \mu$$

*** 二项分布** 它描述的是在 n 次独立重复的 Bernoili 试验中, 设每次试验中事件 A 发生的概率为 p. 用 X 表示 n 重伯努利试验中事件 A 发生的次数. 其概率概率分布为

$$P(X=k) = \binom{n}{k}p^k q^{n-k}.$$

像这样对二项式的操作在计算机科学中是很有必要了解的. 因为我们总是在分析算法的时候用到他们.

因此

$$\begin{aligned}\mathbb{E}(x) =& 0\cdot\binom{n}{0}p^0(1-p)^n + 1\cdot\binom{n}{1}p^1(1-p)^{n-1} + \cdots \\ &+ (n-1)\binom{n}{n-1}p^{n-1}(1-p)^1 + n\binom{n}{n}p^n(1-p)^0 \\ =& \sum_{i=0}^{n} i\binom{n}{i}p^i(1-p)^{n-i}\end{aligned}$$

我们发现它难以计算. 一个原因是里面那个青色的 i 太难受了. 如果我们能够把它去掉就好了. 幸运的是, 组合恒等式里面有这样的一条性质, 可以帮助我们完成这件事.

回忆: 二项式系数的定义是 $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots(1)} = \frac{n!}{k!(n-k)!}$. 其有一个有趣的等式, 可以帮助我们把东西吸收/提取出来. 也就是:

$$\binom{r}{k} = \frac{r}{k}\binom{r-1}{k-1}, \text{整数} k \neq 0$$

这个一般形象地称为 “吸收-提取恒等式”. 这个证明可以使用其定义展开即可.

$$\begin{aligned}\binom{r}{k} &= \frac{r!}{k!(r-k)!} \\ &= \frac{r}{k}\cdot\frac{(r-1)!}{(k-1)!(r-k)!} \\ &= \frac{r}{k}\cdot\frac{(r-1)!}{(k-1)!((r-1)-(k-1))!} \\ &= \frac{r}{k}\cdot\binom{r-1}{k-1}\end{aligned}$$

好. 有了这样的一个想法, 根据吸收-提取恒等式:

$$\begin{aligned}&\sum_{i=1}^{n} i\binom{n}{i}p^i(1-p)^{n-i} \\ =&\sum_{i=1}^{n} n\binom{n-1}{i-1}p^i(1-p)^{n-i} \\ =&np\sum_{i=1}^{n}\binom{n-1}{i-1}p^{i-1}(1-p)^{n-i} \\ =&np(p+(1-p))^{n-1} = np\end{aligned}$$

这个为什么结果如此简单? 实际上因为期望有一些性质. 我们下一节会提到原因.

有些分布没有数学期望. 比如 Cauchy 分布: $f_X(x) = \frac{4}{\pi}\frac{4}{1+x^2}, -\infty < x < +\infty$. 我们在求它的积分的时候, 就会发现

$$\begin{aligned}&\frac{4}{\pi}\int_{-\infty}^{\infty}|x|\frac{4}{1+x^2}\\=&\int_0^{+\infty}\frac{2x}{1+x^2}dx\\=&\int_0^{+\infty}\frac{1}{1+y}dy=\ln(1+y)<+\infty.\end{aligned}$$

发散. 因此 Cauchy 分布没有数学期望.

13.2 数学期望的重要性质

1.(常数的数学期望) 设 C 是常数, 则有 $\mathbb{E}(C) = C$.

事实上, 直观来看, 每一次我们得到的都一定是 C. 因此有其合理性.

2. (常数进出数学期望) 设 C 是常数, X 是一个随机变量, 则有

$$\mathbb{E}(CX) = C\mathbb{E}(X).$$

这是因为常数可以自由进出求和号/积分号.

3. (两个随机变量之和) 设 X, Y 是两个随机变量, 则有

$$E(X+Y) = E(X) + E(Y).$$

事实上, 这是因为分配率得到的结果. 设二维随机变量 (X, Y) 的概率密度为 $f_{X,Y}(x, y)$, 其边缘概率密度为 $f_X(x), f_Y(y)$

$$\begin{aligned}E(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)\,f_{X,Y}(x,y)\,dxdy\\&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy\\&=E(X)+E(Y).\end{aligned}$$

上面的 2° 和 3° 合在一起被称作期望的线性性质. 这和我们在线性代数中的线性空间的理论中的一部分很相似.

4. (两个独立的随机变量之积) 设 X, Y 是相互独立的随机变量, 则有

$$E(XY) = E(X)E(Y).$$

这是因为只有独立的情况下, 联合随机变量的内容才可以得到拆开.
若 X 和 Y 相互独立,

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,dxdy \\ &= \left[\int_{-\infty}^{\infty} xf_X(x)\,dx\right]\left[\int_{-\infty}^{\infty} yf_Y(y)\,dy\right] = E(X)E(Y). \end{aligned}$$

对于上面的二项分布, 我们可以注意到每一次实验, 我们都可以看做一个随机变量. 每一次的期望都是 p, 自然, 所有的实验, 期望就是 np.

练习 IV.1 A, B, C, D 四人竞拍, 价高者得. 假设你是 A, 已知 B, C, D 三人竞拍价互相独立, 切服从 $U(7, 11)$. 如 A 中标可以以 10 转让, 如何报价使得 A 获得的期望收益最大? #

解答: 以 $G(A)$ 表达 A 的收益. $P(G(X) = 10 - X) = ((x-7)/4)^3$, $P(G(X) = 0) = 1 - ((x-7)/4)^3$. 那么 $\mathbb{E}(G(X)) = (10-x)(\frac{x-7}{4})^3$. 得到 x 的极值为 $37/4$.

13.3 随机变量函数的数学期望

下面我们来研究随机变量函数的数学期望.

1. 由随机变量 X 求 $\mathbb{E}(f(X))$ 设 $g(x)$ 是连续函数, X 是随机变量.

- 如果 X 是离散型随机变量, 且分布律为 $P(X = x_k) = p_k (k = 1, 2, \cdots)$, 若 $\sum_k g(x_k)p_k$ 绝对收敛, 则

$$E[g(X)] = \sum_k g(x_k)p_k.$$

- 如果 X 是连续型随机变量, 且概率密度函数为 $f(x)$, 若 $\int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x$ 绝对收敛, 则

$$\int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x.$$

也就是只要把 "系数" 变为了对应的函数值. 事实上, 在对应的离散的状态下, 我们

有如下的说明:

$$
\begin{aligned}
p_Y(y) &= \sum_{\{x|f(x)=y\}} p_X(x) \\
E[f(X)] &= E[Y] \\
&= \sum_y y p_Y(y) \\
&= \sum_y y \sum_{\{x|f(x)=y\}} p_X(x) \\
&= \sum_y \sum_{\{x|f(x)=y\}} y p_X(x) \\
&= \sum_y \sum_{\{x|f(x)=y\}} f(x) p_X(x) \\
&= \sum_x f(x) p_X(x)
\end{aligned}
$$

2. 用 (X,Y) 得到 $\mathbb{E}(g(X,Y))$ 设 $g(x,y)$ 为连续函数,

离散情形:

- 若 (X,Y) 为离散型随机变量, 分布律为 $p_{ij} = P(X = x_i, Y = y_j)(i,j = 1,2,\cdots)$
- 且级数 $\sum_i \sum_j g(x_i,y_j)p_{ij}$ 绝对收敛
- 则

$$\mathbb{E}(g(X,Y)) = \sum_i \sum_j g(x_i,y_j)p_{ij}.$$

连续情形:

- 若 (X,Y) 为连续型随机变量, 联合概率密度函数为 $f(x,y)$
- 且积分 $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)\mathrm{d}x\mathrm{d}y$ 绝对收敛
- 则

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)\mathrm{d}x\mathrm{d}y.$$

多知道一点: 快速排序的期望运行时间

我们在《算法导论》的课程上了解了如下的随机快速排序算法. 使用自然语言大概可以看做算法 2.

算法 2: 随机快速排序算法

Input: 待排序的数组 $S = [x_1, x_2, \cdots, x_n]$
Output: 排序后的数组 S.

- 如果 S 只有一个或者零个元素, 返回 S. 否则继续.
- 随机选择 S 中的元素 s 作为基准元素.
- 把 S 分为两个小的列表 S_1, S_2. 其中, 任何一个 S_1 中的元素比 s 要小, 任何一个 S_2 中的元素比 s 要大.
- 对 S_1, S_2 进行快速排序.
- 返回列表 $[S_1, x, S_2]$.

我们声称: 每一次从元素中独立地随机选取基准, 那么对于任意的输入, 快速排序比较的期望次数为 $2n\lg n + O(n)$.

Proof. 设 $y_1, y_2, \cdots, y_n$ 是输入值 $x_1, x_2, \cdots, x_n$ 按照升序排列的结果. 我们定义 $X_{ij}(i < j)$ 是一个随机变量. 如果在算法执行的某一时刻 y_i 和 y_j 发生了比较, X_{ij} 取值为 1, 否则为 0. 那么比较的总次数满足

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}$$

根据期望的线性性, $\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{ij}\right) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}(X_{ij})$.

由于 X_{ij} 只能取 0 和 1, 是指示变量, $\mathbb{E}(X_{ij})$ 是 X_{ij} 等于 1 的概率.

什么时候 y_i 和 y_j 会发生比较呢? 我们发现 y_i 和 y_j 发生比较, 当且仅当 y_i 或 y_j 是从集合 $Y_{ij} = \{y_i, y_{i+1}, \cdots, y_{j-1}, y_j\}$ 中选取的一个基准元素. 否则, 他们会被分在不同的子列表中, 因而不会比较.

由于我们的基准元素是独立选取的, 因此 y_i 和 y_j 是从 Y_{ij} 中选取的一个基准元素的概率, 也就是 X_{ij} 取 1 的概率, 是 $2/(j-i+1)$. 也就是

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{2}{j-i+1} \\
&\xlongequal{k:=j-i+1} \sum_{i=1}^{n-1}\sum_{k=2}^{n-i+1}\frac{2}{k} = \sum_{k=2}^{n}\sum_{i=1}^{n+1-k}\frac{2}{k} \\
&= \sum_{k=2}^{n}(n+1-k)\frac{2}{k} = \left((n+1)\sum_{k=2}^{n}\frac{2}{k}\right) - 2(n-1) \\
&= (2n+2)\sum_{k=1}^{n}\frac{1}{k} - 4n.
\end{aligned}$$

□

14 条件数学期望

定义 14.1. 设随机变量 X 在 $y = y_i$ 的条件下的条件分布列为

$$p_{i|j} = P(X = x_i | Y = y_j),$$

若级数 $\sum_i x_i p_{i|j}$ 绝对收敛, 那么称此级数 $Y = Y_i$ 条件下 X 的条件数学期望, 记为 $\mathbb{E}(X|Y = y_i)$.

正如条件概率也是概率, 条件数学期望也是数学期望. 自然, 它就满足数学期望对应的性质.

例子 14.1. 有一根长度为 l 的棍子, 在 $(0,l)$ 内均匀地取一点 X 折断. 再次在 $(0,a)$ 选一点 Y 折断. 求 Y 的期望.

解答: 假设 Y 是一个具体值的情形. $\mathbb{E}(X \mid Y=y) = \frac{y}{2}$ 是一个数字. 但是实验之前, 这个值是随机的. 我们干脆把它写作 $\mathbb{E}(X|Y)$ 作为一个随机变量. 其中, 记作 $\mathbb{E}(X|Y) = Y/2$.

下面来看有时候如何方便地计算期望. 我们可以把样本空间构成一组不相交的集合 $A_1, A_2, \cdots, A_n$, 因此可以在每个小块上面计算期望值. 也就是:

$$\mathbb{E}(X) = P(A_1)\mathbb{E}(X \mid A_1) + \cdots + P(A_n)\mathbb{E}(X \mid A_n).$$

这个可以使用全概率公式, 两边乘上其对应的系数得到. 这就是**全期望公式**.

有了这样的记号, 我们给出指数分布的期望的又一说明:

例子 14.2. 我们把原来的事件划分为不相交的两类.

$$A_1 : \{x = 1\}, \quad A_2 : \{x > 1\}.$$
$$\mathbb{E}(X) = P(X=1)\mathbb{E}(X \mid x=1) + P(X>1)\mathbb{E}(X \mid x>1).$$

那么计算它就得到了:

$$\begin{aligned}\mathbb{E}(X) &= P(X=1)E[X \mid x=1] + P(X>1)E[X \mid x>1].\\ &= p \cdot 1 + (1-p)\boxed{\ ?\ }\end{aligned}$$

我们关注 $E[X \mid x>1]$ 的情形,

$$\begin{aligned}E[X \mid x-1>0] &= E[X-1 \mid x-1>0] + 1\\ &= \mathbb{E}(X) + 1\end{aligned}$$

由于指数分布的无记忆性, 上面的式子就是 $\mathbb{E}(X)+1$.

上面的式子经过整理得到 $\mathbb{E}(X) = p \cdot 1 + (1-p)(\mathbb{E}(X)+1)$. 我们由此解答出来 $\mathbb{E}(X)$, 因此就得到期望值为 $1/p$.

14.1 迭代的期望

我们接下来考虑期望的迭代公式, 也就是 $\mathbb{E}[\mathbb{E}[X \mid Y]]$. 这个符号看上去很难懂. 我们先从简单来看这个公式:

- $\mathbb{E}[X \mid Y=y]$ 表示当随机变量 Y 取值 y 的时候, 随机变量 X 的期望.
- 考虑所有可能的 Y, 其构成的一组期望就是 $\mathbb{E}_y[X \mid Y=y], \forall y$.
- 这一组期望, 是一个关于 Y 的随机变量. 为了简便, 我们把这样的一组简写为 $\mathbb{E}[X \mid Y]$.
- 对于这样的一个随机变量, 自然可以求它的期望.

实际上, 这个公式有一个更简洁的表示. 我们按照定义展开:

$$\begin{aligned}
\text{左手边} &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \frac{f_{x,Y}(x,y)}{f_Y(y)} dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X,Y}(x,y) dx dy. \\
&= \int_{-\infty}^{+\infty} x f_x(x) dx = \mathbb{E}[x] = \text{右手边}
\end{aligned}$$

这个公式有一个名字, 叫做 "迭代期望定律 (Law of iterated expectations)".

定理 14.2. 迭代的期望公式:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

15 期望相关的不等式

1. Jensen 不等式 先从一个简单的例子开始. 假设我们从 $[1, 99]$ 的范围内选取一个正方形的边长 X, 问面积的期望是多少? 根据上面的公式, 我们知道 $\mathbb{E}(X^2) = \int_0^{99} x^2 dx = 9950/3$.

在式子变得复杂之后, 这样的运算就稍显麻烦. 能不能偷懒计算 $\mathbb{E}(X)^2$ 代替上述的计算? 并不可以, 因为 $\mathbb{E}(X)^2 = 50^2 = 2500$. 但是我们可以看到大小关系: $\mathbb{E}(X^2) \geq \mathbb{E}(X)^2$. 并且可以证明它是对的:

Proof. 考虑 $Y = (X - \mathbb{E}(X))^2$, 随机变量 Y 非负, 意味着其期望也是非负的. 所以有

$$\begin{aligned}
0 \leq \mathbb{E}[Y] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X\mathbb{E}[X]] + (\mathbb{E}[X])^2 \\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2
\end{aligned}$$

□

实际上, 这只是 Jensen 不等式的一个例子. Jensen 不等式会告诉我们, 对于任意的一个凸函数 f, 都会有 $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$.

-- 高数老师告诉过我, 高等数学里面和数学分析里面函数的凹凸性定义是相反的. 求偏导的时候求的顺序也是相反的. 那我应该相信谁? -- 相信定义.

Aside

回顾凸函数: 一个函数 $f : \mathbb{R} \to \mathbb{R}$ 是凸函数, 当且仅当 $\forall x_1, x_2, 0 \leq \lambda \leq 1$, 有

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

并且我们知道, 如果 f 是二次可微函数, 则 f 是凸函数当且仅当 $f''(x) \geq 0$.

定理 15.1. 如果 f 是凸函数, 那么 $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.

Proof. 假设 f 有 Taylor 展开式. 令 $\mu = \mathbb{E}[X]$. 根据 Taylor 中值定理, 存在一个值 c 使得

$$\begin{aligned} f(x) &= f(\mu) + f'(\mu)(x-\mu) + \frac{f''(c)(x-\mu)^2}{2} \\ &\geq f(\mu) + f'(\mu)(x-\mu) \end{aligned}$$

由于 f 是凸函数, 得到 $f''(c) > 0$. 两边同时取期望, 得到:

$$\begin{aligned} \mathbb{E}[f(X)] &\geq \mathbb{E}\left[f(\mu) + f'(\mu)(X-\mu)\right] \\ &= \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[X] - \mu) \\ &= f(\mu) = f(\mathbb{E}[X]). \end{aligned}$$

□

2. Cauchy-Schwarz 不等式 这个不等式说的是下面的一个事情:

对于随机变量 X, Y, 我们有 $\mathbb{E}[|XY|]^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$

这使我想到了线性代数的时候学习内积空间的时候的痛苦...

首先考虑 $\mathbb{E}\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f_X(x), \mathbb{E}\left(X^2\right) = \int_{-\infty}^{\infty} y^2 f_Y(y)$. 为了方便起见, 不妨设 $\mathbb{E}\left(X\right), \mathbb{E}\left(Y\right)$ 都大于 0. 假若我们引入新变量, 使得我们的 $\tilde{X}, \tilde{Y}$ 的值都在 0 到 1 之间:

$$\tilde{X} := \frac{X}{\sqrt{\mathbb{E}\left(X^2\right)}}, \tilde{Y} := \frac{Y}{\sqrt{\mathbb{E}\left(Y^2\right)}}$$

这个时候由于任意的值有基本不等式, 故随机变量也要满足基本不等式描述的: $2|\tilde{X}\tilde{Y}| \leq \tilde{X}^2\tilde{Y}^2$. 对两边取期望, 有 $2\mathbb{E}\left(\tilde{X}\tilde{Y}\right) \leq \mathbb{E}\left(\tilde{X}^2\right)\mathbb{E}\left(\tilde{Y}^2\right) = 2$. 因此有 $\mathbb{E}\left(\tilde{X}\tilde{Y}\right) \leq 1$, 自然

$$\mathbb{E}\left(|XY|\right)^2 = \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$$

Aside

这个不等式我们在线性代数课程上面讲解向量的模的时候也有一个类似的不等式. 也就是如果一个线性空间 V 中 $u, v \in V$, 那么有

$$|\langle u, v\rangle| \leq \|u\|\|v\|.$$

其中, $\|\cdot\|$ 是 $\cdot$ 的范数.
向量的范数一般定义为 $\sqrt{\langle v, v\rangle}$.

3. Markov 不等式 概率和期望之间有什么联系? 在例 13.1中, 我们了解到了期望计算和概率相联系的一种形式: $\mathbb{E}\left(X\right) = \int_0^{\infty} P(X > t)dt$. 自然的, $P(X > \epsilon)$ 和 $\mathbb{E}\left(X\right)$ 有联系.

定理 15.2. Markov 不等式:

$$\forall \varepsilon > 0, P(X > \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}$$

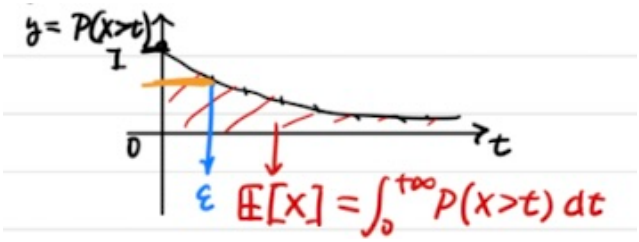

Figure 4: Markov 不等式限制的方式

我们采用最粗暴的限制限制手段. 如图 4, 仅仅考虑 $0 \leq t \leq \epsilon, 0 < y < P(x > \epsilon)$ 的部分. 因此, 我们可以得到证明:

$$\begin{aligned}\mathbb{E}[x] &\geqslant \int_0^{\varepsilon} P(x > t)dt \\ &\geqslant \int_0^{\varepsilon} P(x > \varepsilon)dt \\ &= \varepsilon P(x > \varepsilon)\end{aligned}$$

于是我们可以得到想要的不等式.

-- 怎么说把这个东西切了就切了? 有点太粗暴了吧!
-- 实际上期望是它的一阶矩, 如果我们知道了更高阶的矩, 我们得到的信息就多了.

练习题: 数学期望

练习 IV.2 某个大楼有 10 层, 某次有 25 人在一楼搭乘电梯上楼. 假设每人都等可能地在 2 到 10 层中的任何一层出电梯, 并且出电梯与否互相独立. 同时在 2 到 10 层中没有人上电梯. 并且电梯只有在有人要出电梯的时候才停止. 求电梯停下的总次数的数学期望. #

提示或解答: 每个人在第 i 层下的概率为 $P_i = 1/9, i = 1, 2, \cdots, 10$. 记第 k 个人在第 i 层下电梯记作 $A_{k,i}$. 那么对于任意的 i, $P(A_k) = 1/9, P(\overline{A_k}) = 8/9.(k = 1, 2, \cdots, 25)$. 又因为 $A_{1,i}, \cdots, A_{25,i}$ 相互独立, 那么第 i 层无人下电梯的概率为

$$P\left(\prod_{i=1}^{25} \overline{A_k}\right) = \prod_{i=1}^{25} P\left(\overline{A_k}\right) = \left(\frac{8}{9}\right), (k = 1, 2, \cdots, 25).$$

设 X_i 是指示第 i 层有没有人下的示性函数, 那么, $P(X = 0) = (8/9)^{25}$, $P(X = 1) = 1-(8/9)^{25}$. 因此电梯停的总次数为 $X = \sum_{i=2}^{10} X_i = 9 \times \left(1 - (8/9)^{25}\right)$

练习 IV.3 设随机变量 X, Y 相互独立, 且都服从 $N(\mu, \sigma^2)$. 设 $Z = \max\{X, Y\}$. 求 $\mathbb{E}(Z)$. #

提示或解答: 正则化变量. 令 $U := \frac{X-\mu}{\sigma}, V := \frac{Y-\mu}{\sigma}$. 那么 $X = \sigma U + \mu, Y = \sigma V + \mu$. 由于 X, Y 独立, 因此 U, V 独立. 且 $U \sim N(0, 1), V \sim N(0, 1)$. 发现 $Z = \max\{X, Y\} = \max\{\sigma U + \mu, \sigma V + \mu\} = \sigma \max(U, V) + \mu$.

要求 $\max(U,V)=\frac{1}{2}(U+V+|U-V|)$, 我们要知道 $|U-V|$ 的分布. 令 $T:=U-V\sim N(0,2)$. 考虑

$$E(|T|)=\int_{-\infty}^{+\infty}|t|\frac{1}{\sqrt{2\pi}\cdot\sqrt{2}}e^{-\frac{t^2}{2\times 2}}dt=\frac{2}{\sqrt{\pi}}.$$

因此得到 $E[\max(U,V)]=\frac{1}{2}(EU+EV+E|U-V|)=\frac{1}{\sqrt{\pi}}$. 从而 $E[\max\{X,Y\}]=\sigma E[\max(U,V)]+\mu=\frac{\sigma}{\sqrt{\pi}}+\mu$.

练习 IV.4 对于任意的整数 $k>1$, 证明 $\mathbb{E}\left(X^k\right)\geq(\mathbb{E}(X))^k$. #

提示或解答: 使用 Jensen 不等式. 令 $f(x):=x^n$.

练习 IV.5 (负二项分布) 现在考虑投掷一枚硬币, 直到第 k 次出现正面的投掷次数的 X 的分布. 其中每次投掷硬币出现是独立的, 概率为 p. 证明: 对于 $n>k$, 这个分布是

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

#

提示或解答: 可以考虑在出现最后一次正面之前, 我们需要在 $n-1$ 个空格里面插入 $k-1$ 个正面. 其余的是反面. 由此可以得到这个等式的意义.

练习 IV.6 对于一枚每次投出正面的概率是 p 的硬币, 每次投掷之间互相独立, 直到第 k 次正面出现的时候, 期望的投掷次数是多少? #

提示或解答:

方法 1. 令 $\mathbb{E}(N_k):=$ 有 n 个正面投掷的期望数. 令 X_1 是第一次投掷的结果. 如果第一次的结果是反面, 那么还会回到 N_k, 但是期望会加 1. 反之, 就会到达 N_{k-1}, 同样期望会加一. 也就是

$$\mathbb{E}(N_k\mid X_1=T)=1+\mathbb{E}(N_k)$$

$$\mathbb{E}(N_k\mid X_1=H)=1+E[N_{k-1}]$$

因此有表达式 $\mathbb{E}(N_k)=\mathbb{E}(N_k\mid X_1=H)\cdot P(X_1=H)+\mathbb{E}(N_k\mid X_1=T)\cdot P(X_1=T)$. 解之, 得到 $\mathbb{E}(N_k)=\mathbb{E}(N_{k-1})+\frac{1}{p}$. 因此是 $\mathbb{E}(N_k)=n/p$.

高中生: 考虑 f_i 代表现在已经扔出了 i 个正面, 还需要扔的期望数. 根据这个定义 $f_k=0$, $f_{k-1}=(1-p)(f_{k-1}+1)+p(f_k+1)$.

练习 IV.7 我们在一个 n 张不同的卡牌中进行带放回的抽卡, 抽出每一张卡的概率相等. 请问我们期望抽取多少次, 才能把所有的卡牌见到一遍? #

提示或解答: 可以考虑当前已经见到了 k 个卡牌, 那么还有 k/n 的概率回到当前的状态, $(n-k)/n$ 的概率到达见到的少一个的状态.

练习 IV.8 我们一遍一遍地投掷一个均匀的筛子. 在连续出现两个 6 之前, 期望的投掷次数是多少? (答案不是 36). #

提示或解答: 想办法搞明白各个你的划分之间的转移关系. 或者看 Markov 链的相关内容.

练习 IV.9 数字 $1, 2, \cdots, n$ 可以用如下的排列函数 $\pi : [1..n] \to [1..n]$ 表示. 其中, $\pi(i)$ 是 i 在这个排列中的序号. 排列 π 的不动点是满足 $\pi(x) = x$ 的 x 的集合. 求从所有排列中选取一个排列的不动点的期望. #

提示或解答: 一个简单的方法: 定义 $f(\sigma)$ 为随机一个排列 σ 的不动点的数量. 那么

$$\mathbb{E}[f(\sigma)] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_{\sigma(i)=i}\right] = \sum_{i=1}^{n} \mathbb{E}\left[1_{\sigma(i)=i}\right]$$

对于每个 i，i 成为 σ 的一个不动点的概率等于 $\sigma(i)$ 等于 i 的概率，因此等于 $\frac{1}{n}$。利用这个结果，我们得到 $\mathbb{E}[f(\sigma)] = 1$。

练习 IV.10 假设 $a_1, a_2, \ldots, a_n$ 是 1, 2, ..., n 的一个随机排列，等可能地是 $n!$ 种可能的排列之一。当对列表 $a_1, a_2, \ldots, a_n$ 进行排序时，元素 a_i 必须移动 $|a_i - i|$ 个位置才能到达其在排序后的位置。求出

$$\mathbb{E}\left[\sum_{i=1}^{n} |a_i - i|\right]$$

也就是元素移动的期望总距离. #

提示或解答: 和上面一样, 使用期望的线性性.

练习 IV.11 排列 $\pi : [1, n] \to [1, n]$ 可以表示为一张图. 我们可以这样做: 为每个数字 $i, i = 1, \ldots, n$ 设立一个顶点。如果排列将数字 i 映射到数字 $\pi(i)$，则从顶点 i 到顶点 $\pi(i)$ 绘制一个有向边。这导致一个由不相交环构成的图。注意，其中一些环可能是自环。在 n 个数字的随机排列中，期望的环的数量是多少? #

提示或解答: 设 $h(n)$ 为 [n] 的随机排列中环的平均数;我声称 h 满足以下递推关系:$h(n) = \frac{n-1}{n} h(n-1) + \frac{1}{n}(h(n-1)+1)$。因为: 设 π 是 2,3,⋯,n 的任意排列，用环形式表示。假设忽略括号，从左到右排列的 π 的条目是 $\pi_1, \ldots, \pi_{n-1}$。现在，在 π 中以以下两种方式之一插入 1。

- 在 π_1 的左边作为 (1)，形成一个独立的环。
- 立即在 π_k 的某个 $k = 1, \ldots, n-1$ 的位置之后插入，与 π_k 同一个环中。

每个 $[n]$ 的排列都可以唯一地通过这种方式从 2,⋯,n 的一个唯一排列 π 获得。

一个 2,⋯,n 的随机排列的平均环的数量当然是 $h(n-1)$。递推关系 (1) 立即由以下事实推出：上述操作 (1) 增加了 1 个环，并占据了所有情况中的 $\frac{1}{n}$，而操作 (2) 不改变环的数量，并占据了剩余的 $\frac{n-1}{n}$ 情况。

计算可知,

$$h(n) = H_n = \sum_{k=1}^{n} \frac{1}{k}.$$

16 方差

随机变量 X 的方差, 表征 X 取值的散布程度. 最简单的想法是: 可以使用 $\mathbb{E}(|X - \mathbb{E}(X)|)$ 来表征. 但是, 由于带着绝对值在求导的时候会带来不少麻烦, 我们考虑使用平方来操作, 也就是 $\mathbb{E}((X - \mathbb{E}(X))^2)$.

定义 16.1. 定义设 X 是一个随机变量，若 $E\{[X - E(X)]^2\}$ 存在，则称 $\mathbb{E}((X - \mathbb{E}(x))^2)$ 为 X 的方差, 记为 $D(X)$ 或 $\mathrm{Var}(X)$, 即

$$D(X) = \mathrm{Var}(X) := \mathbb{E}\left((X - \mathbb{E}(X))^2\right).$$

在应用上还引入量 $\sqrt{D(X)}$, 记为 $\sigma(X)$, 称为标准差或均方差.

其实际意义是明显的: 我们查看方差的大小

- $D(X)$ 较小: X 的取值比较集中在 $\mathbb{E}(X)$ 的附近
- $D(X)$ 较大: 表示 X 的取值较分散.

其实, 方差说特别也不特别, 它实际上就是随机变量 X 的函数 $g(X) = (X - E(X))^2$ 的期望.

对这个式子做一些简单的化简, 我们发现, 要计算一个数的方差, 我们只需要计算 $\mathbb{E}(X)$ 以及 $\mathbb{E}(X^2)$ 即可.

命题 16.2. 随机变量 X 的方差可按下列公式计算.

$$D(X) = E\left(X^2\right) - [E(X)]^2.$$

其证明如下:

Proof. 由数学期望的性质 $1°, 2°, 3°$ 得

$$\begin{aligned} D(X) &= E\left\{[X - E(X)]^2\right\} = E\left\{X^2 - 2XE(X) + [E(X)]^2\right\} \\ &= E\left(X^2\right) - 2E(X)E(X) + [E(X)]^2 \\ &= E\left(X^2\right) - [E(X)]^2. \end{aligned}$$

□

16.1 方差的性质

我们首先从抽象的层面说明方差具有哪些性质. 实际上, 这只是个关于期望的性质的练习 – 方差只是求解一堆数学期望而已. 我们先假设下面所遇到的随机变量其方差存在:

1. 常数的方差是 0 设 C 是常数, 则 $D(C)=0$.

Proof. 代入定义, 有 $D(C)=E\left\{[C-E(C)]^2\right\}=0$. □

2. 一个变量的线性变换后的方差 设 X 是随机变量, C 是常数, 则有

$$D(CX)=C^2D(X),\quad D(X+C)=D(X).$$

Proof.

$$\begin{aligned}D(CX)&=E\left\{[CX-E(CX)]^2\right\}=C^2E\left\{[X-E(X)]^2\right\}=C^2D(X)\\D(X+C)&=E\left\{[X+C-E(X+C)]^2\right\}=E\left\{[X-E(X)]^2\right\}=D(X)\end{aligned}$$

□

3. 两个随机变量和的方差 设 X,Y 是两个随机变量, 则有

可不要忘记了这个 2. !

$$D(X+Y)=D(X)+D(Y)+2E\{(X-E(X))(Y-E(Y))\}.$$

特别, 若 X,Y 相互独立, 则有

$$D(X+Y)=D(X)+D(Y).$$

这一性质可以推广到任意有限多个相互独立的随机变量之和的情况.

Proof.

$$\begin{aligned}D(X+Y)=&E\left\{[(X+Y)-E(X+Y)]^2\right\}\\=&E\left\{[(X-E(X))+(Y-E(Y))]^2\right\}\\=&E\left\{(X-E(X))^2\right\}+E\left\{(Y-E(Y))^2\right\}\\&+2E\{[(X-E(X)][Y-E(Y)]\}\\=&D(X)+D(Y)+2E\{[X-E(X)][Y-E(Y)]\}.\end{aligned}$$

对于其中的 $E\{[(X-E(X)][Y-E(Y)]\}$ 一项:

$$\begin{aligned}
&2E\{[(X-E(X)][Y-E(Y)]\} \\
&=2E\{XY-XE(Y)-YE(X)+E(X)E(Y)\} \\
&=2\{E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\} \\
&=2\{E(XY)-E(X)E(Y)\}.
\end{aligned}$$

□

16.2 常见分布的方差

1. 均匀分布 由于期望 $\mathbb{E}(X)$ 我们已经求过了, 我们来看平方的期望 $\mathbb{E}(X^2)$. 经过简单的计算, 我们发现

$$E\left(X^2\right)=\int_{-\infty}^{\infty}x^2f(x)dx=\frac{b^2+ab+a^2}{3}$$

因此

$$D(X)=E\left(X^2\right)-E(X)^2=\frac{(b-a)^2}{12}.$$

2. Poisson 分布 我们回忆参数为 λ 的 Poisson 分布的概率分布是 $f(k)=\frac{e^{-\lambda}\lambda^k}{k!}$. 同样的,

$$\begin{aligned}
E\left(X^2\right)&=\sum_{k=0}^{\infty}k^2\frac{\lambda^ke^{-\lambda}}{k!}=\sum_{k=1}^{\infty}ke^{-\lambda}\lambda\frac{\lambda^{k-1}}{(k-1)!} \\
&=\sum_{k=1}^{\infty}(k-1+1)\lambda e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \\
&=\sum_{k=1}^{\infty}(k-1)\lambda e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!}+\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\
&=\lambda^2e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda \\
&=\lambda^2+\lambda \\
D(X)&=\lambda^2+\lambda-\lambda^2=\lambda
\end{aligned}$$

3. 指数分布 回忆它的概率分布函数为

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

并且 $\mathbb{E}(X) = 1/\lambda$. 我们还是关注 $\mathbb{E}(X^2)$.

$$\begin{aligned} EX^2 &= \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx \\ &= -\int_0^{+\infty} x^2 de^{-\lambda x} = -\left. x^2 e^{-\lambda x} \right|_0^{+\infty} + \int_0^{+\infty} 2x e^{-\lambda x} dx \end{aligned}$$

因此

$$\mathrm{Var(X)} = EX^2 - (EX)^2 = \frac{1}{\lambda^2}$$

4. 正态分布 我们发现正态分布的通用表达式比较麻烦. 我们首先计算标准正态分布的方差, 然后试图通过线性变换的方法得到原来的正态分布表达式. 回忆 $N(0,1) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

计算的大概过程大概如下:

- 写出表达式: $E(X^2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx$
- 由于是偶函数, $E(X^2) = 2\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx$
- 希望计算出 $\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx$.
- 极坐标换元, 得到值为 $1/2$.
- 因此 $\mathbb{E}(X^2) = 1$.
- $D(X) = E(X^2) - (E(X))^2 = 1 - 0^2 = 1$
- 通过标准化得到方差

接着我们做线性变换代换回去.

$$\begin{aligned} &X \sim N\left(\mu, \sigma^2\right), Z \sim N(0,1) \\ &Z = \frac{X-\mu}{\sigma} \sim N(0,1) \\ &D(Z) = D\left(\frac{X-\mu}{\sigma}\right) = D\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right) = D\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma^2} D(X) = 1 \end{aligned}$$

得到 $D(x) = \sigma^2$.

5. 几何分布 我们回忆其分布函数为 $P(X = k) = (1-p)^{k-1} p$ $\frac{1}{p}$, 以及 $\mathbb{E}(X) = 1/p$. 方便起见, 令 $q := 1 - p$.

那么

$$\begin{aligned}
D(X) &= E\left(X^2\right) - (E(X))^2 \\
&= \sum_{k=1}^{\infty} k^2 q^{k-1} p - \frac{1}{p^2} \\
&= p\left[\sum_{k=1}^{\infty}(k+1)kq^{k-1} - \sum_{k=1}^{\infty} kq^{k-1}\right] - \frac{1}{p^2} \\
&= p\left(\frac{\mathrm{d}^2}{\mathrm{d}q^2}\sum_{k=1}^{\infty} q^{k+1} - \frac{\mathrm{d}}{\mathrm{d}q}\sum_{k=1}^{\infty} q^k\right) - \frac{1}{p^2} \\
&= p\left[\frac{2}{(1-q)^3} - \frac{1}{(1-q)^2}\right] - \frac{1}{p^2} = \frac{q}{p^2}.
\end{aligned}$$

6. 二项分布 我们在上一节里面用非常痛苦的方法求解了它的期望. 这里, 我们还是先不用任何的性质, 再来痛苦地求解一下它的方差.

$$\begin{aligned}
E\left(k^2\right) &= \sum_{k=0}^{n} k^2 p(k) \\
&= \sum_{k=1}^{n} k^2 \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k} \\
&= \sum_{k=1}^{n} [k(k-1)+k] \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k} \\
&= \sum_{k=1}^{n} k(k-1) \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k} + \sum_{k=1}^{n} k \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k} \\
&= \sum_{k=1}^{n} k(k-1)\frac{n!}{k!(n-k)!} p^k q^{n-k} + np \\
&= n(n-1)p^2 \sum_{k=1}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} q^{(n-2)-(k-2)} + np \\
&= n(n-1)p^2 + np
\end{aligned}$$

因此 $D(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}\left(X\right))^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)$.

这和我们期望的性质也是吻合的. 毕竟, 这就是 n 次独立的 0-1 分布. 我们的结果也是支持这一点的.

总结 总的来说, 我们有表 1的表格总结我们的结果.

16.3 Chebyshev 不等式

利用随机变量的期望和方差, 我们可以推出 Chebushev 不等式, 来得到更好的界限.

分布名称	PDF	均值	方差
二项分布	$P(X=k)=C_n^k p^k(1-p)^{n-k}$	np	$np(1-p)$
几何分布	$P(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{(1-p)}{p^2}$
正态分布	$f\left(x \mid \mu, \sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
均匀分布	$f(x \mid a,b)=\begin{cases}\frac{1}{b-a} & \text{for } a\le x\le b\\ 0 & \text{for } x<a \text{ or } x>b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布	$f(x)=\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Poisson 分布	$f(k)=\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ

Table 1: 常见分布的均值和方差

定理 16.3. 对于任意的 $a>0$, 有

$$P(|X-\mathbb{E}(X)|\geq a)\leq\frac{D(X)}{a^2}$$

Proof. 由于有绝对值, 我们首先将绝对值去掉, 得到

$$P(|X-\mathbb{E}[X]|\geq a)=P\left((X-\mathbb{E}[X])^2\geq a^2\right)$$

由于 $(X-\mathbb{E}(X))^2$ 是一个非负的变量, 使用 Markov 不等式得到:

$$P\left((X-\mathbb{E}[X])^2\geq a^2\right)\leq\frac{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]}{a^2}=\frac{\mathrm{Var}[X]}{a^2}$$

□

这个定理在说什么? 这意味着, 它与期望相差的程度的概率可以使用方差的相关性质来描述.

有了这个定理, 我们可以得到关于方差的另一个重要的性质:

4. 方差为 0 的充要条件 $D(X)=0$ 的充要条件是 X 以概率 1 取常数 $E(X)$, 即

$$P(X=E(X))=1$$

Proof. 充分性. 设 $P(X=E(X)\}=1$, 则有 $P\{X^2=[E(X)]^2)=1$, 于是

$$D(X)=E\left(X^2\right)-[E(X)]^2=0.$$

必要性: 用反证法假设 $P(X=E(X)\}<1$, 则对于某一个数 $\varepsilon>0$, 有 $P(|X-E(X)|\geqslant\varepsilon)>0$, 但由切比雪夫不等式, 对于任意 $\varepsilon>0$, 由刚才的不等关系 $P\{|X-\mu|\geqslant\varepsilon)\leq\frac{\sigma^2}{\varepsilon^2}$, $\sigma^2=0$, 由于有

$$P(|X-E(X)|\geqslant\varepsilon)=0,$$

矛盾, 于是 $P(X=E(X))=1$. □

17 协方差. 矩

1. 随机变量的协方差 在上一节中, 在看两个随机变量的和的方差的时候, 我们注意到 $\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]+\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$. 为了方便起见, 我们使用记号 $\mathrm{Cov}(X,Y)$ 来代表 $\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$. 我们很快会看到这样做的好处.

定义 17.1. 两个随机变量 X 和 Y 的*协方差*定义做

$$\mathrm{Cov}(X,Y):=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$$

自然地, 我们首先来看看矩的性质. 实际上, 两个随机变量的矩类似于双线性函数.

- 交换性: $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$.
- 双线性性: $\mathrm{Cov}(aX,bY)=ab\,\mathrm{Cov}(X,Y).a,b$ 是常数.
- 可加性: $\mathrm{Cov}\left(X_1+X_2,Y\right)=\mathrm{Cov}\left(X_1,Y\right)+\mathrm{Cov}\left(X_2,Y\right)$.

这些性质可以直接带入定义证明.

另外, 如果把它按照期望的性质乘开, 自然有 $\mathrm{Cov}(x,y)=\mathbb{E}\left(XY\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$. 这是一个在实际运算中比较方便的性质.

2. 随机变量的相关系数 我们希望对于两个随机变量考察是不是有线性关系. 也就是, 假设我们有两个随机变量 X,Y. 我们希望以 X 的线性函数 $a+bX$ 来近似表示 Y. 自然, 我们可以用 $E\left[(Y-(a+bX))^2\right]$ 表示拟合的好坏程度. 这个期望值越小越好. 我们的目标是求最佳近似式 $a+bX$ 中的 a,b . 也就是最小化 e.

$$\begin{aligned}e&:=E\left[(Y-(a+bX))^2\right]\\&=E\left(Y^2\right)+b^2E\left(X^2\right)+a^2-2bE(XY)+2abE(X)-2aE(Y)\end{aligned}$$

这个式子中, a,b 是变量. 对 e 求关于 a,b 的偏导数, 看一看可能的极值点.

$$\begin{cases}\frac{\partial e}{\partial a}=2a+2bE(X)-2E(Y)=0,\\ \frac{\partial e}{\partial b}=2bE\left(X^2\right)-2E(XY)+2aE(X)=0.\end{cases}\Longrightarrow\begin{aligned}b_0&=\tfrac{\mathrm{Cov}(X,Y)}{D(X)}\\ a_0&=E(Y)-E(X)\tfrac{\mathrm{Cov}(X,Y)}{D(X)}.\end{aligned}$$

回带, 得到

$$\begin{aligned}\min_{a,b}E\left\{[Y-(a+bX)]^2\right\}&=E\left\{\left[Y-(a_0+b_0X)\right]^2\right\}\\&=\left(1-\frac{\mathrm{Cov}^2(X,Y)}{D(X)D(Y)}\right)D(Y).\end{aligned}$$

注意上述表达式中的青色的内容. 我们发现这样一个事情可以用来衡量对应变量之间的相关性. 于是我们给这个一个新定义:

定义 17.2. 我们定义

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

为随机变量 X 与 Y 的相关系数.

根据上述的推演, 我们猜测: 相关系数总是小于 1 的. 以及如果相关系数等于 1 的话, 那么一定会有相关系数的绝对值等于 1. 下面我们描述这两个性质并给出证明:

- $|\rho_{XY}| \leq 1$.
- $|\rho_{XY}| = 1$ 的充要条件是, 存在常数 a, b 使 $\{Y = a + bX\} = 1$.

Proof. 1° 因为 $E\left\{[Y - (a_0 + b_0X)]^2\right\}$ 及 $D(Y)$ 的非负性, 知 $1 - \rho_{XY}^2 \geq 0$, 亦即 $|\rho_{XY}| \leq 1$.

2° 若 $|\rho_{XY}| = 1$, 有:

$$E\left\{\left[Y - (a_0 + b_0X)^2\right]\right\} = 0.$$

从而:

$$0 = E\left\{\left[Y - (a_0 + b_0X)^2\right]\right\}$$

$$= D\left[Y - (a_0 + b_0X)\right] + \left\{E\left[Y - (a_0 + b_0X)\right]\right\}^2,$$

故有:

$$D\left[Y - (a_0 + b_0X)\right] = 0$$

$$E\left[Y - (a_0 + b_0X)\right] = 0$$

$$P(Y - (a_0 + b_0X) = 0\} = 1, P\{Y = a_0 + b_0X) = 1.$$

反之, 若存在常数 a^*, b^* 使得:

$$P(Y = a^* + b^*X\} = 1, P\{Y - (a^* + b^*X) = 0) = 1,$$

于是 $P([Y - (a^* + b^*X)]^2 = 0) = 1.$

即得 $E\{[Y - (a^* + b^*X)]^2\} = 0.$

故有 $0 = E\left\{[Y - (a^* + b^*X)^2]\right\}$

$$= E\left\{\left[Y - (a_0 + b_0X)^2\right]\right\} = \left(1 - \rho_{XY}^2\right) D(Y).$$

即得 $|\rho_{XY}| = 1.$ □

由于 e 是 $|\rho_{XY}|$ 的严格单调减少函数. 当 $|\rho_{XY}|$ 较大时 e 较小, 表面 X, Y(就线性关系而言) 联系较紧密. 但是, 如果我们发现 $|\rho_{XY}| = 0$, 那么说不说明两个随机变量独立呢? 实际上并不一定. 因为不相关只是就线性关系而言, 而相互独立是就一般关系而言. 没有线性关系, 倒是有可能有平方关系, 以及各种各样奇怪的关系. 只是用线性的方式拟合并不好.

所以我们要特别强调:

相互独立的变量相关系数一定等于 0, 相关系数等于 0 的变量不一定独立.

例子 17.1. 设 (X,Y) 的分布律为

$Y \ddots X$	-2	-1	1	2	$P(Y=j)$
1	0	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{1}{2}$
4	$\frac{1}{4}$	0	0	$\frac{1}{4}$	$\frac{1}{2}$
$P(X=i)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1

解答: 易知 $E(X)=0, E(Y)=\frac{5}{2}, E(XY)=0$, 于是 $\rho_{XY}=0$,X,Y 不相关, 这表示不存在线性关系, 但, $P\{X=-2,Y=1\}=0\neq P\{X=-2\}P\{Y=1\}$, 知 X,Y 不是相互独立的.

例子 17.2. 对于两个正态分布而言, 如果相关系数等于 0, 那么他们一定独立吗? 我们回顾这样的性质: 对于二维正态随机变量 $(X,Y),X$ 和 Y 相互独立的充要条件是参数 $\rho=0$.

$$\mathrm{Cov}(X,Y)=0=\rho\sigma_1\sigma_2$$

$$\text{于是}\quad \rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\rho$$

这就是说, 二维正态随机变量 (X,Y) 的概率密度函数中的参数 ρ 就是 X 和 Y 的相关系数, 因而二维正态随机变量的分布可完全由 X,Y 各自的数学期望、方差以及它们的相关系数所确定. 所以, 对于二维正态随机变量 (X,Y) 而言, X 和 Y 不相关与 X 和 Y 相互独立是等价的.

3. 矩的相关概念 方差中, 我们用到了 $\mathbb{E}(X^2)$ 和 $\mathbb{E}(X)$ 进而求出 $D(X)$. 实际上, 在有些时候, 我们可能会求解 $\mathbb{E}(X^3)$, $\mathbb{E}(X^4),\cdots$, 进而运用期望的线性性求解有关多项式的期望. 在这里, 为了统一语言, 我们给出一连串的如下的定义, 以保证在未来说到的时候会清楚的理解.

定义 17.3. (1) 设 X 和 Y 是随机变量, 若 $E\left(X^k\right),\quad k=1,2,\cdots$ 存在, 称它为 X 的 k 阶原点矩, 简称 k 阶矩.
(2) 若 $\quad E\left(X^kY^l\right),\quad k,l=1,2,\cdots$ 存在, 称它为 X 和 Y 的 $k+l$ 阶混合矩.
(3) 设 X 和 Y 是随机变量, 若

$$E\left\{[X-E(X)]^k\right\},k=2,3,\cdots$$

存在, 称它为 X 的 k 阶中心矩.
(4) 若 $E\left\{[X-E(X)]^k[Y-E(Y)]^l\right\},\quad k,l=1,2,\cdots$ 存在, 称它为 X 和 Y 的 $k+l$ 阶混合中心矩.

18 中位数和平均值

我们在初中的时候学过了一列数的中位数. 下面我们试着把这个定义拓展到随机变量中.

一个朴素的想法是, 我们定义 $P(X \le n) \ge 1/2$, 以及 $P(X \ge m) \ge 1/2$ 为随机变量 X 的中位数. 因为有离散的奇数个随机变量的取值的时候, 把它们的值按照某一次序排列 $x_1, x_2, \cdots, x_{2k+1}$, 那么中位数就是 x_{k+1}. 如果有偶数个值 $x_1, x_2, \cdots, x_{2k}$, 那么 (x_k, x_{k+1}) 上面的任何一个值都是中位数.

随机变量 X 的期望 $\mathbb{E}(X)$ 通常和中位数是不一样的. 那么他们之间有什么关系呢?

定理 18.1. 对于期望存在, 中位数存在的随机变量 X, 设其期望为 $\mathbb{E}(X)$, 中位数为 m, 那么:

- 期望 $\mathbb{E}(X)$ 是使得 $\mathbb{E}((X-c)^2)$ 最小的 c 的值;
- 中位数 m 是使得 $\mathbb{E}(|X-c|)$ 最小的 c 的值;

Proof. 第一个性质可以使用期望的线性性说明. $\mathbb{E}[(X-c)^2] = \mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2$. 对于 c 求导数, 即可得到我们要的性质.

第二个性质我们可以考虑: 对于任何一个不是中位数的值 c 和任何中位数 m, 有 $\mathbb{E}[|X-c|] > \mathbb{E}[|X-m|]$. 等价地, 也就是说 $\mathbb{E}[|X-c| - |X-m|] > 0$.

不是一般性, 假设我们能够让上述式子最小的 c 的值大于中位数 m 的值, 也就是 $c > m$.(我们会看到如果 $c < m$ 也可以有类似的方法证明得到) 由于 c 不是中位数, $P(X > c) < 1/2$. 我们接下来考虑等式里面的任何一个量 x 和 c 的关系. 对于 x, c 的关系, 有

- 如果 $x \ge c, |x-c| - |x-m| = m - c$.
- 如果 $m < x < c, |x-c| - |x-m| = c + m - 2x > m - c$.
- 如果 $x \le m, |x-c| - |x-m| = c - m$.

综上考虑, 我们有

$$\begin{aligned}&\mathbb{E}[|X-c| - |X-m|]\\ &\quad = P(X \ge c)(m-c) + \sum_{x:m<x<c} P(X=x)(c+m-2x) + P(X \le m)(c-m).\end{aligned}$$

我们发现关键就是探讨中间的那一项求和会造成什么影响. 这时候, 我们需要考虑两种情况:

1° 如果 $P(m < X < c) = 0$, 那么

$$\begin{aligned}\mathbb{E}[|X-c| - |X-m|] &= P(X \ge c)(m-c) + P(X \le m)(c-m)\\ &> \frac{1}{2}(m-c) + \frac{1}{2}(c-m) \qquad (\text{因为} P(X \ge c) < 1/2 \text{ 且} m < c)\\ &= 0,\end{aligned}$$

2° 如果 $P(m < X < c) \neq 0$, 那么

$$\begin{aligned}
&\mathbb{E}[|X-c| - |X-m|] \\
&\quad = P(X \geq c)(m-c) + \sum_{x:m<x<c} P(X=x)(c+m-2x) \\
&\qquad + P(X \leq m)(c-m) \\
&\quad > P(X > m)(m-c) + P(X \leq m)(c-m) \qquad (\text{因为}\forall x. c+m-2x > m-c) \\
&\quad > \frac{1}{2}(m-c) + \frac{1}{2}(c-m) \\
&\quad = 0,
\end{aligned}$$

综上所述, 这个总是大于 0. 因此我们的定理成立. □

并且, 我们会发现标准差, 中位数, 期望之间有一个很有趣的关系:

定理 18.2. 如果 X 是随机变量, 且具有有限的标准差 σ, 期望 μ 和中位数 m, 那么

$$|\mu - m| \leq \sigma.$$

意味着中位数和期望差的不远, 只要我们的方差足够小.

Proof.

$$\begin{aligned}
|\mu - m| &= |\mathbb{E}[X] - m| \\
&= |\mathbb{E}[X-m]| \\
&\leq \mathbb{E}[|X-m|] && (\text{Jensen 不等式}) \\
&\leq \mathbb{E}[|X-\mu|] && (\text{中位数会最小化}\mathbb{E}(|X-c|)) \\
&\leq \sqrt{\mathbb{E}[(X-\mu)^2]} && (\text{Jensen 不等式}) \\
&= \sigma.
\end{aligned}$$

□

练习题: 方差. 矩. 协方差. 中位数.

练习 IV.12 假设我们扔一个均匀的骰子 100 次. 令 X 为这 100 次中出现的次数和. 使用 Chebyshev 不等式得到 $P(|X-350| \geq 50)$. #

提示或解答: Chebyshev 不等式: $P(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$

练习 IV.13 每次抛硬币，正面朝上的概率为 p，并且相互独立。求直到第 k 次出现正面的抛硬币次数的方差。 #

提示或解答： 计算得到 $\mathbb{E}(X)=\frac{i}{1-p}=\sum_{k=0}^{\infty} ip(1-p)^{k-1}$. 并且 $\mathbb{E}\left(X^2\right)=\sum_{k=0}^{\infty} i^2p(1-p)^{k-1}=\frac{i^2}{1-p}$. 那么方差就是 $\mathbb{E}\left(X^2\right)-\mathbb{E}(X)^2=\frac{i^2}{1-p}-(\frac{i}{1-p})^2$.

练习 IV.14 一个简单的股票市场模型表明，每天，价格为 q 的股票以概率 p 增长到 qr，以概率 $1-p$ 下跌到 $\frac{q}{r}$。假设我们从价格为 1 的股票开始，找到 d 天后股票价格的期望值和方差的公式。 #

提示或解答： 类似于二项分布.

练习 IV.15 当一个排列 $\pi:[1,n]\to[1,n]$ 满足 $\pi(x)=x$ 时，x 是该排列的一个不动点。求从所有排列中随机选择的排列中不动点数量的方差。 #

提示或解答： 可以使用 $D(X)=\mathbb{E}\left(X^2\right)-\mathbb{E}(X)^2$.

19 Chernoff 界和 Hoeffding 界

我们导出两个比较有用的界限. 这两个界限可以告诉我们关于分布的尾部其实是呈指数式衰减的. 这些界限是通过将 Markov 不等式应用于随机变量的矩生成函数而得出的. 在讨论之前, 我们先介绍重要的一个重要的概念, 也就是矩的生成函数.

19.1 矩的生成函数

定义 19.1. 对于一个随机变量 X, 它的矩生成函数为

$$M_X(t)=\mathbb{E}\left[\mathrm{e}^{tX}\right].$$

我们主要考察它在 0 附近的存在性和性质. 这样的一个函数 $M_X(t)$ 实际上捕捉了 X 的所有矩. 我们马上会看到这一点.

定理 19.2. 假设随机变量 X 的矩生成函数为 $M_X(t)$. 在可以交换期望值和微分操作的前提下, 对于所有 $n>1$, 我们可以得到

$$E[X^n]=M_X^{(n)}(0),$$

这里的 $M_X^{(n)}(0)$ 表示 $M_X(t)$ 在 $t=0$ 处的第 n 阶导数的值.

Proof. 假设我们可以交换积分和期望的次序, 那么有 $M_X^{(n)}(t)=\mathbb{E}\left[X^n\mathrm{e}^{tX}\right]$. 带入 $t=0$, 得到 $M_X^{(n)}(0)=\mathbb{E}\left[X^n\right]$. □

例子 19.1. 下面我们来看一个例子, 矩的生成函数是如何编码一个随机变量的所有的矩的. 对于一个参数为 p 的几何分布 ($P(X=n)=(1-p)^{n-1}p$), 那么对于 $t<-\ln(1-p)$, 有

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[\mathrm{e}^{tX}\right] \\ &= \sum_{k=1}^{\infty}(1-p)^{k-1}p\mathrm{e}^{tk} \\ &= \frac{p}{1-p}\sum_{k=1}^{\infty}(1-p)^{k}\mathrm{e}^{tk} \\ &= \frac{p}{1-p}\left(\left(1-(1-p)\mathrm{e}^{t}\right)^{-1}-1\right) \end{aligned}$$

也就是

$$\begin{aligned} &M_X^{(1)}(t) = p\left(1-(1-p)\mathrm{e}^{t}\right)^{-2}\mathrm{e}^{t} \quad \text{并且} \\ &M_X^{(2)}(t) = 2p(1-p)\left(1-(1-p)\mathrm{e}^{t}\right)^{-3}\mathrm{e}^{2t} + p\left(1-(1-p)\mathrm{e}^{t}\right)^{-2}\mathrm{e}^{t} \end{aligned}$$

带入得到 $\mathbb{E}[X]=1/p$, 以及 $\mathbb{E}\left[X^2\right]=(2-p)/p^2$. 与我们先前计算的结果相符.

另一个有用的特性是, 随机变量的矩生成函数（或者说变量的所有矩）可以独特地描述其分布. 不过, 证明太过复杂, 我们略去.

定理 19.3. 假设 X 和 Y 是两个随机变量. 如果存在一个 $\delta>0$, 使得在区间 $(-\delta,\delta)$ 内对于所有 t 都满足 $M_X(t)=M_Y(t)$, 那么 X 和 Y 将拥有相同的分布.

上面的定理在确定独立随机变量之和的分布的时候非常有用.
下面的一个性质是关于两个随机变量的矩的和的.

定理 19.4. 如果 X 和 Y 是独立随机变量, 那么,

$$M_{X+Y}(t)=M_X(t)M_Y(t).$$

Proof. 使用变量的独立性, 再使用幂的性质, 就得到了:

$$M_{X+Y}(t)=\mathbb{E}\left[\mathrm{e}^{t(X+Y)}\right]=\mathbb{E}\left[\mathrm{e}^{tX}\mathrm{e}^{tY}\right]=\mathbb{E}\left[\mathrm{e}^{tX}\right]\mathbb{E}\left[\mathrm{e}^{tY}\right]=M_X(t)M_Y(t).$$

□

19.2 推导得到 Chernoff 界

对于一个随机变量, 我们要想得到 Chernoff 界, 我们需要对于矩的生成函数 e^{tx} 选择一个比较好的 t, 然后使用 Markov 不等式. 也就是, 对于任意的 $t>0$, 有

$$P(X\geq a)=P\left(\mathrm{e}^{tX}\geq\mathrm{e}^{ta}\right)\leq\frac{\mathbb{E}\left[\mathrm{e}^{tX}\right]}{\mathrm{e}^{ta}}$$

我们取出所有 $t>0$ 中使它最小的一个, 有

$$P(X \geq a) \leq \min_{t>0} \frac{\mathbb{E}\left[\mathrm{e}^{tX}\right]}{\mathrm{e}^{ta}}.$$

同样的, 对于 $t<0$, 有

$$P(X \leq a) = P\left(\mathrm{e}^{tX} \geq \mathrm{e}^{ta}\right) \leq \frac{\mathbb{E}\left[\mathrm{e}^{tX}\right]}{\mathrm{e}^{ta}}$$

因此

$$P(X \leq a) \leq \min_{t<0} \frac{\mathbb{E}\left[\mathrm{e}^{tX}\right]}{\mathrm{e}^{ta}}$$

从这种方法导出的界叫做 Chernoff 界. 由于实际选取的 t 不一样, 我们得到的界的形式可能并不唯一.

现在, 我们推导 Chernoff 界最常用的一种形式. 也就是若干个 Poisson 试验的和. 这些实验不一定同分布, 也就是令 $X_1, X_2, \cdots, X_n$ 是一系列独立的 Poisson 实验, 也就是 $P\left(X_i=1\right)=p_i$. 令 $X=\sum_{i=1}^{n} X_i$, 定义均值 μ 为

$$\mu = \mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \sum_{i=1}^{n} p_i$$

现在我们想关心 X 偏离其期望的概率, 比如偏离 $\delta\mu(\delta>0)$ 的概率是多少. 也就是看看 $P(X \geq (1+\delta)\mu)$ 以及 $P(X \leq (1-\delta)\mu)$ 的大小是多少.

要得到 Chernoff 界, 对于每一个 X_i 首先我们使用矩生成函数:

$$\begin{aligned} M_{X_i}(t) &= \mathbb{E}\left[\mathrm{e}^{tX_i}\right] \\ &= p_i\mathrm{e}^t + (1-p_i) \\ &= 1 + p_i\left(\mathrm{e}^t - 1\right) \\ &\leq \mathrm{e}^{p_i\left(\mathrm{e}^t-1\right)}, \qquad \text{(使用不等式} 1+y \leq e^y\text{)} \end{aligned}$$

由于 n 个随机变量等于矩生成函数相乘, 以及每一个矩生成函数唯一决定了一个分布, 我们知道这些随机变量的和的矩的生成函数为:

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &\leq \prod_{i=1}^{n} \mathrm{e}^{p_i\left(\mathrm{e}^t-1\right)} \\ &= \exp\left\{\sum_{i=1}^{n} p_i\left(\mathrm{e}^t-1\right)\right\} \\ &= \mathrm{e}^{\left(\mathrm{e}^t-1\right)\mu}. \end{aligned}$$

有了这样的矩的生成函数, 我们就可以得到对应的 Chernoff 界.

定理 19.5. 假设 $X_1, X_2, \cdots, X_n$ 是互相独立的 Poisson 实验, 满足 $P(X_i = 1) = p_i$. 令 $X = \sum_{i=1}^{n} X_i$, 均值 $\mu = \mathbb{E}[X]$, 有如下的 Chernoff 界:

- 对于任意的 $\delta > 0$,

$$P(X \geq (1+\delta)\mu) \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

- 对于任意的 $0 < \delta \leq 1$,

$$P(X \geq (1+\delta)\mu) \leq \mathrm{e}^{-\mu\delta^2/3}$$

- 对于 $R \geq 6\mu$,

$$P(X \geq R) \leq 2^{-R}$$

该定理的第一个界是最强的, 我们正是从这个界限推导出其他两个界限. 这些更加简单的界很多情况下容易叙述和计算.

Proof. 使用 Markov 不等式, 对于任意的 $t > 0$, 我们有:

$$\begin{aligned} P(X \geq (1+\delta)\mu) &= P\left(\mathrm{e}^{tX} \geq \mathrm{e}^{t(1+\delta)\mu}\right) \\ &\leq \frac{\mathbb{E}\left[\mathrm{e}^{tX}\right]}{\mathrm{e}^{t(1+\delta)\mu}} \\ &\leq \frac{\mathrm{e}^{\left(\mathrm{e}^{t}-1\right)\mu}}{\mathrm{e}^{t(1+\delta)\mu}}. \end{aligned}$$

我们置 $t = \ln(1+\delta) > 0$ 得到

$$P(X \geq (1+\delta)\mu) \leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}.$$

也就是第一个等式. 要得到第二个等式, 我们需要说明, 对于 $0 < \delta \leq 1$, 有

$$\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}} \leq \mathrm{e}^{-\delta^2/3}$$

两边同时取对数, 就转化为普通高中生也会做的问题了. 具体地, 我们取 $f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3} \leq 0$, 计算 $f(\delta)$ 的导数,

$$\begin{aligned} f'(\delta) &= 1 - \frac{1+\delta}{1+\delta} - \ln(1+\delta) + \frac{2}{3}\delta \\ &= -\ln(1+\delta) + \frac{2}{3}\delta \\ f''(\delta) &= -\frac{1}{1+\delta} + \frac{2}{3}. \end{aligned}$$

我们看到对于 $0 \leq \delta < \frac{1}{2}$, 有 $f''(\delta) < 0$, 而对于 $\delta > \frac{1}{2}$, 有 $f''(\delta) > 0$. 因此在区间 $[0, 1]$

上, $f'(\delta)$ 先减小后增加. 考虑到 $f'(0)=0$ 和 $f'(1)<0$, 我们可以推断在区间 $[0,1]$ 中 $f'(\delta)\leq 0$. 由于 $f(0)=0$, 所以在该区间内 $f(\delta)\leq 0$. 于是得到了第二个式子.

对于第三个式子, 令 $R=(1+\delta)\mu$, 那么, 对于 $R\geq 6\mu$, $\delta=R/\mu-1\geq 5$. 因此, 使用第一个式子, 有

$$\begin{aligned} P(X\geq(1+\delta)\mu) &\leq \left(\frac{\mathrm{e}^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \\ &\leq \left(\frac{\mathrm{e}}{1+\delta}\right)^{(1+\delta)\mu} \\ &\leq \left(\frac{\mathrm{e}}{6}\right)^{R} \\ &\leq 2^{-R}. \end{aligned}$$

□

上面探讨了高于平均值的概率. 其实对于低于平均值偏差为 δ 的概率, 我们也有类似的结果:

定理 19.6. 假设 $X_1, X_2, \cdots, X_n$ 是互相独立的 Poisson 实验, 满足 $P(X_i=1)=p_i$. 令 $X=\sum_{i=1}^{n}X_i$, 均值 $\mu=\mathbb{E}[X]$, 对于 $0<\delta<1$, 有

- $$P(X\leq(1-\delta)\mu)\leq\left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$
- $$P(X\leq(1-\delta)\mu)\leq\mathrm{e}^{-\mu\delta^2/2}$$

Proof. 对于 $t<0$, 使用 Markov 不等式, 有

$$\begin{aligned} P(X\leq(1-\delta)\mu) &= P\left(\mathrm{e}^{tX}\geq\mathrm{e}^{t(1-\delta)\mu}\right) \\ &\leq \frac{\mathbb{E}\left[\mathrm{e}^{tX}\right]}{\mathrm{e}^{t(1-\delta)\mu}} \\ &\leq \frac{\mathrm{e}^{\left(\mathrm{e}^{t}-1\right)\mu}}{\mathrm{e}^{t(1-\delta)\mu}}. \end{aligned}$$

对于 $0<\delta<1$, 仿照上述, 设 $t=\ln(1-\delta)$, 得到第一个式子

$$P(X\leq(1-\delta)\mu)\leq\left(\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$

仿照上述的证明该方法, 同样可以使用 $\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{(1-\delta)}}\leq\mathrm{e}^{-\delta^2/2}$ 成立, 证明第二个式子.

□

所以对于泊松分布而言, 距离均值的偏差到底是多少? 我们根据上面两个定理的第二个进行推论:

推论 19.7. 假设 $X_1, X_2, \cdots, X_n$ 是互相独立的 Poisson 实验, 满足 $P(X_i = 1) = p_i$. 令 $X = \sum_{i=1}^{n} X_i$, 均值 $\mu = \mathbb{E}[X]$, 对于 $0 < \delta < 1$, 有

$$P(|X - \mu| \geq \delta\mu) \leq 2\mathrm{e}^{-\mu\delta^2/3}.$$

例子 19.2 (投硬币)**.** 令随机变量 X 表示 n 次独立的投掷中, 正面向上的次数. 使用 *Chernoff* 界, 我们得到

$$\begin{aligned} P\left(\left|X - \frac{n}{2}\right| \geq \frac{1}{2}\sqrt{6n\ln n}\right) &\leq 2\exp\left\{-\frac{1}{3}\frac{n}{2}\frac{6\ln n}{n}\right\} \\ &= \frac{2}{n} \end{aligned}$$

这表示均值聚集在 $n/2$ 的过程是十分快速的. 大多数时候, 与平均值的偏差约为 $O(\sqrt{n\ln n})$.

考虑我们想要一个序列中不多于 $n/4$ 个正面, 不少于 $3n/4$ 个反面, 使用 *Chebyshev* 不等式, 得到 $P\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{4}{n}$. 而使用刚刚的 *Chernoff* 界我们就发现随着 n 增大, 我们实际上在指数衰减. 也就是

$$\begin{aligned} P\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) &\leq 2\exp\left\{-\frac{1}{3}\frac{n}{2}\frac{1}{4}\right\} \\ &\leq 2\mathrm{e}^{-n/24}. \end{aligned}$$

因此, *Chernoff* 界给出的结果比使用 *Chebyshev* 不等式获得的界要小得多.

19.3 Hoeffding 界

如果我们的随机变量是有界的, 那么可以使用 Hoeffding 界来获得更加紧的界. 我们先来陈述这一事实.

定理 19.8 (Hoeffding 界)**.** 令 $X_1, X_2, \cdots, X_n$ 是互相独立的随机变量, 并且对于任意的 $1 \leq i \leq n, \mathbb{E}(X_i) = \mu, P(a \leq X_i \leq b) = 1$, 那么

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2\mathrm{e}^{-2n\epsilon^2/(b-a)^2}$$

要证明这个定理, 首先需要证明一个引理.

引理: 设 X 是一个随机变量, $P(X \in [a, b]) = 1$ 并且 $\mathbb{E}[X] = 0$. 那么对于任意的 $\lambda > 0$, 有

$$\mathbb{E}\left[\mathrm{e}^{\lambda X}\right] \leq \mathrm{e}^{\lambda^2(b-a)^2/8}.$$

Proof. 由于 $f(x)=\mathrm{e}^{\lambda x}$ 是一个凸函数, 对任意 $\alpha\in(0,1)$ 有

$$f(\alpha a+(1-\alpha)b)\leq \alpha \mathrm{e}^{\lambda a}+(1-\alpha)\mathrm{e}^{\lambda b}$$

对于 $x\in[a,b]$, 令 $\alpha=\frac{b-x}{b-a}$, 这样一来, $x=\alpha a+(1-\alpha)b$. 我们就有

$$\mathrm{e}^{\lambda x}\leq \frac{b-x}{b-a}\mathrm{e}^{\lambda a}+\frac{x-a}{b-a}\mathrm{e}^{\lambda b}$$

考虑取 $\mathrm{e}^{\lambda X}$ 的期望, 由于 $\mathbb{E}(X)=0$, 我们有

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{\lambda X}\right]&\leq \mathbb{E}\left[\frac{b-X}{b-a}\mathrm{e}^{\lambda a}\right]+\mathbb{E}\left[\frac{X-a}{b-a}\mathrm{e}^{\lambda b}\right]\\&=\frac{b}{b-a}\mathrm{e}^{\lambda a}-\frac{\mathbb{E}[X]}{b-a}\mathrm{e}^{\lambda a}-\frac{a}{b-a}\mathrm{e}^{\lambda b}+\frac{\mathbb{E}[X]}{b-a}\mathrm{e}^{\lambda b}\\&=\frac{b}{b-a}\mathrm{e}^{\lambda a}-\frac{a}{b-a}\mathrm{e}^{\lambda b}\end{aligned}$$

下面对于最终的表达式做一点操作. 令 $\phi(t)=-\theta t+\ln(1-\theta+\theta\mathrm{e}^{t})$, 对于 $\theta=\frac{-a}{b-a}>0$. 有

$$\begin{aligned}\mathrm{e}^{\phi(\lambda(b-a))}&=\mathrm{e}^{-\theta\lambda(b-a)}\left(1-\theta+\theta\mathrm{e}^{\lambda(b-a)}\right)\\&=\mathrm{e}^{\lambda a}\left(1-\theta+\theta\mathrm{e}^{\lambda(b-a)}\right)\\&=\mathrm{e}^{\lambda a}\left(\frac{b}{b-a}-\frac{a}{b-a}\mathrm{e}^{\lambda(b-a)}\right)\\&=\frac{b}{b-a}\mathrm{e}^{\lambda a}-\frac{a}{b-a}\mathrm{e}^{\lambda b},\end{aligned}$$

这和我们推出的上界 $\mathbb{E}(\exp(\lambda X))$ 相等. 我们发现 $\varphi(0)=\varphi'(0)=0$, 对于任意的 $t,\varphi''(t)\leq 1/4$. 根据 Taylor 展开, 对于任意的 $t>0$, 都有 t' 使得 $t'\in[0,t]$,

$$\phi(t)=\phi(0)+t\phi'(0)+\frac{1}{2}t^2\phi''(t')\leq\frac{1}{8}t^2$$

因此, 对于 $t=\lambda(b-a)$, 有

$$\phi(\lambda(b-a))\leq\frac{\lambda^2(b-a)^2}{8}$$

于是

$$\mathbb{E}\left[\mathrm{e}^{\lambda X}\right]\leq\mathrm{e}^{\phi(\lambda(b-a))}\leq\mathrm{e}^{\lambda^2(b-a)^2/8}$$

□

有了这个引理, 我们下面证明 Hoeffding 界.

Proof. 令 $Z_i=X_i-\mathbb{E}[X_i]$, $Z=\frac{1}{n}\sum_{i=1}^{n}Z_i$.

对于任意的 $\lambda > 0$, 根据 Markov 不等式,

$$P(Z \geq \epsilon) = P\left(\mathrm{e}^{\lambda Z} \geq \mathrm{e}^{\lambda\epsilon}\right) \leq \mathrm{e}^{-\lambda\epsilon}\mathbb{E}\left[\mathrm{e}^{\lambda Z}\right] \leq \mathrm{e}^{-\lambda\epsilon}\prod_{i=1}^{n}\mathbb{E}\left[\mathrm{e}^{\lambda Z_i/n}\right]$$
$$\leq \mathrm{e}^{-\lambda\epsilon}\prod_{i=1}^{n}\mathrm{e}^{\lambda^2(b-a)^2/n^2} \leq \mathrm{e}^{-\lambda\epsilon+\lambda^2(b-a)^2/8n}$$

在倒数第二个步骤中, 我们运用了上述引理, 并利用了 Z_i/n 被限制在 $\frac{a-\mu}{n}$ 和 $\frac{b-\mu}{n}$ 之间这一事实。设置 $\lambda = \frac{4n\epsilon}{(b-a)^2}$ 得到:

$$P\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mu \geq \epsilon\right) = P(Z \geq \epsilon) \leq \mathrm{e}^{-2n\epsilon^2/(b-a)^2}$$

对于 $Z \leq -\epsilon$ 的情况, 有同样的情况. 因此定理得证.

□

Part V

概率的极限定理

20 随机序列的收敛性

1. 引例 设 $X_1, X_2, \cdots$ 是一列随机变量, 我们希望知道 n 很大时 X_n 近似地是什么样的随机变量. 这也称为随机变量的收敛性.

我们在起初的时候说明了公理的方法描述了概率. 但是, 这些公理化的方法是不是真的描述了我们想要的呢? 也就是: 具有直观上的 "频率的稳定值" 的含义呢?(见第 2.1 节.) 更加正式地说, 若 A 在 n 次独立试验中发生了 μ_n 次, 问: 当 n 很大时, $\frac{\mu_n}{n}$ 是否与 $P(A)$ 很接近?

例子 20.1. 先考虑使用示性函数做的一个简单的例子. 令

$$X_i = \begin{cases} 1, & \text{当第} i \text{ 次试验时} A \text{ 发生}, \\ 0, & \text{当第} i \text{ 次试验时} A \text{ 不发生} \end{cases}$$

$(i = 1, 2, \cdots)$, 则 $X_1, X_2, \cdots$ 是随机变量序列. 所谓 n 次独立试验, 就是指随机变量 $X_1, \cdots, X_n$ 相互独立. 显然 A 发生的次数 $\mu_n = \sum_{i=1}^{n} X_i$. 记 $X_n = \frac{1}{n}\sum_{i=1}^{n} X_i (n \geqslant 1)$. 问题是: 当 n 很大时, 随机变量 X_n 是否与常数 $P(A)$ 很接近?

我们可以对于任何随机变量 X, 我们可以使用比较简单的离散型随机变量 X^* 来

近似 $X, 0 \leq X - X^* < \varepsilon$, 这里我们选取比较简单的分段函数 X^* , 它与 ε 有关.

$$X^* = \begin{cases} 0, & 0 \leq X < \varepsilon, \\ -\varepsilon, & -\varepsilon \leq X < 0, \\ \varepsilon, & \varepsilon \leq X < 2\varepsilon, \\ \cdots\cdots & \cdots\cdots \\ k\varepsilon, & k\varepsilon \leq X < (k+1)\varepsilon \quad (k \text{ 是任何整数}), \\ \cdots\cdots & \cdots\cdots \end{cases}$$

实际上, $X^* = \left\lfloor \frac{1}{\varepsilon} X \right\rfloor \varepsilon$. 令 $\varepsilon = \frac{1}{n}, X_n = X^*$, 则 n 很大时 X_n 与 X 任意接近.

例子 20.2. 一个射手向一目标连续射击 *6000* 次, 每次射中的概率是 $\frac{1}{6}$, 问: 射中次数在 *900* 至 *1100* 之间的概率是多少? 这个问题从理论上不难回答, 从第一章知这个概盗等于 $\sum_{k=900}^{1100} \binom{6000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{5000-k}$, 但具体数值如何算出, 这就不容易了. 用 μ_{6000} 表示 *6000* 次射击中射中的次数, 能否找到比较简单的随机变量 η, 其分布函数比较好算 (或其数值可从造好的表中查出), 使得 μ_{6000} 与 η 很接近或者说

$$P\left(900 \leq \mu_{6000} \leq 1100\right) \approx P(900 \leq \eta \leq 1100).$$

2. 三个收敛性质 我们下面考察三种重要的定义: 我们假设 η 和 $X_1, X_2, \cdots$ 都是随机变量. 我们再次强调: 这些随机变量 $\eta = \eta(\omega), X_1 = X_1(\omega), \cdots, X_n = X_n(\omega), \cdots$, 实质上都是概率空间 $(\Omega, \mathscr{F}, P)$ 上的实值函数!

但是为了简便起见, 我们在事件的表述上常常省去了 ω, 实际上:

$$\begin{gathered} \{|X_n - \eta| \geqslant \varepsilon\} := \{\omega : |X_n(\omega) - \eta(\omega)| \geqslant \varepsilon\}, \\ \left\{\lim_{n\to\infty} X_n = \eta\right\} := \left\{\omega : \lim_{n\to\infty} X_n(\omega) := \eta(\omega)\right\}, \\ \{X_n \leq x\} := \{\omega : X_n(\omega) \leq x\}, \quad \{\eta \leq x\} = \{\omega : \eta(\omega) \leq x\}. \end{gathered}$$

我们在下面的论述中, 常常省略 ω.

我们假设 η 和 $X_1, X_2, \cdots$ 都是随机变量.

定义 20.1. 称 $X_1, X_2, \cdots$ 依概率收敛于 η, 若对任何 $\varepsilon > 0$, 成立

$$\lim_{n\to\infty} P\left(|X_n - \eta| \geqslant \varepsilon\right) = 0.$$

此时记做 $X_n \xrightarrow{P} \eta$.

定义 20.2. 称 $X_1, X_2, \cdots$ 概率为 1 (或几乎必然) 地收敛于 η, 若

$$P\left(\lim_{n\to\infty} X_n = \eta\right) = 1.$$

此时记做 $X_n \xrightarrow{\text{a.s.}} \eta$ 或 $X_n \to \eta$ (a.s.). a. s. 是英文 almost surely 的缩写.

定义 20.3. 称 $X_1, X_2, \cdots$ 若收敛 (依分布收敛) 于 η, 若对 η 的分布函数 $F(x)$ 的任何连续点 x, 皆成立

$$\lim_{n\to\infty} P\left(X_n \le x\right) = P(\eta \le x).$$

此时记做 $X_n \xrightarrow{w} \eta$ (或 $X_n \xrightarrow{d} \eta$).

其中依分布收敛和随机变量的矩母函数很有关联. 如果我们设 $Y_1, Y_2, \ldots$ 是一系列随机变量, Y_i 的分布函数为 F_i. 自然, 它的矩母函数为 $M_X(t) = \mathbb{E}\left(e^{tX}\right)$, 是一个随机变量. 我们 (不加证明地) 给出下面的定理. 这在后面证明大数定律的时候很有用处.

定理 20.4. 如果我们设 $Y_1, Y_2, \ldots$ 是一系列随机变量. Y_i 的分布函数为 F_i, 矩母函数为 M_i. 令 Y 是一个随机变量, 且 Y 的分布函数为 F, 矩母函数为 M. 如果 $\forall t$, 都有 $\lim_{n\to\infty} M_n(t) = M(t)$, 那么对于所有使得 $F(t)$ 连续的 t, 都有 $F_n \xrightarrow{w} F$.

3. 三个收敛性的联系

定理 20.5. 设 $X_n \xrightarrow{\text{a. s.}} \eta$, 则 $X_n \xrightarrow{P} \eta$.

Proof. 假设有一个集合 $A = \{\omega : X_1(\omega), X_2(\omega), \cdots \text{ 不收敛于 } \eta(\omega)\}$. 从假设知 $P(A) = 0$. 对任何 $\varepsilon > 0$, 令 $B = \{\omega : \text{有无穷多个 } n \text{ 使得 } |X_n(\omega) - \eta(\omega)| \geqslant \varepsilon\}$, $B_m = \{\omega : \text{有 } n \geqslant m \text{ 使得 } |X_n(\omega) - \eta(\omega)| \geqslant \varepsilon\}$, 则 $B_m \supset B_{m+1}, B = \bigcap_{m=1}^{\infty} B_m$.

于是 $\lim_{m\to\infty} P\left(B_m\right) = P(B) \le P(A) = 0$. 因为 $P\left(|X_m - \eta| \geqslant \varepsilon\right) \le P\left(B_m\right)$. 所以 $\lim P\left(|X_m - \eta| \geqslant \varepsilon\right) = 0$. 这就证明了 $X_n \xrightarrow{P} \eta$. □

应注意的是, 上述定理的逆不成立. 下面的例子给出一个反例:

例子 20.3. *设 $\Omega = (0,1), \mathscr{F}$ 由 $(0,1)$ 中所有 Borel 子集组成, P 是这样的概率测度: 对任何区间 $(a,b)(0 \le a < b \le 1), P((a,b)) = b - a$. 在概率空间 $(\Omega, \mathscr{F}, P)$ 上考虑下列随机变量序列.*

对任何正整数 k 及 $j=1,2,\cdots,2^k$, 令

$$X_{k1}=\begin{cases}1, & 0<\omega<\frac{1}{2^k},\\ 0, & \text{其他;}\end{cases}$$

$$X_{kj}=\begin{cases}1, & \frac{j-1}{2^k}\leqslant\omega<\frac{j}{2^k}, \quad (j>1).\\ 0, & \text{其他}\end{cases}$$

这些 $\left\{X_{kj}:k\geqslant 1,j=1,2,\cdots,2^k\right\}$ 可排成一个序列: $X_{11},X_{12},X_{21},X_{22},X_{23},X_{24},X_{31},X_{32},\cdots,X_{38},X_{41},$ (用 "字典排列法", 按第 1 个下标 k 从小到大排, 第 1 个下标相同者则按第 2 个下标从小到大排). 把这个序列依次记为 $X_1,X_2,\cdots$. 易知, 对每个 $n\geqslant 1$ 有 k_n 和 j_n 使得 $X_n=X_{k_nj_n}$ 对任何 $\varepsilon\in(0,1),P\left(|X_n|\geqslant\varepsilon\right)=P\left(X_n=1\right)=\frac{1}{2^{k_n}}$. 由于 $n\to\infty$ 时 $k_n\to\infty$, 故有 $\lim_{n\to\infty}P\left(|X_n|\geqslant\varepsilon\right)=0$. 这表明 $X_n\xrightarrow{P}0$.

但是, 对任何 $\omega\in(0,1),\lim_{n\to\infty}X_n(\omega)$ 不存在. 实际上, 对任何 ω 和 k 有唯一的 j_k 使得 $X_{kj_k}(\omega)=1$, 从而 $j\neq j_k$ 时 $X_{kj}(\omega)=0$. 由此可见, 数列 $X_1(\omega),X_2(\omega),\cdots$ 中有无穷多个是 1 , 又有无穷多个是 0 , 因而 $\lim_{n\to\infty}X_n(\omega)$ 不存在.

定理 20.6. 设 $X_n\xrightarrow{P}\eta$, 则 $X_n\xrightarrow{w}\eta$.

Proof. 设 x_0 是 η 的分布函数 $F(x)$ 的连续点. 记 $F_n(x)=P\left(X_n\leq x\right)(n\geqslant 1)$. 易知, 对任何 $\varepsilon>0$, 有

$$\begin{aligned}\left\{X_n\leq x_0\right\}&=\left\{X_n-\eta+\eta\leq x_0\right\}\\&\subset\left\{X_n-\eta\leq-\varepsilon\right\}\cup\left\{\eta\leq x_0+\varepsilon\right\},\end{aligned}$$

于是

$$P\left(X_n\leq x_0\right)\leq P\left(X_n-\eta\leq-\varepsilon\right)+P\left(\eta\leq x_0+\varepsilon\right).$$

故

$$F_n\left(x_0\right)-F\left(x_0\right)\leq P\left(|X_n-\eta|\geqslant\varepsilon\right)+F\left(x_0+\varepsilon\right)-F\left(x_0\right).$$

类似地, 有

$$\left\{X_n\leq x_0\right\}\supset\left\{X_n-\eta\leq-\varepsilon,\eta\leq x_0-\varepsilon\right\}.$$

于是

$$\begin{aligned}P\left(X_n\leq x_0\right)&\geqslant P\left(X_n-\eta\leq\varepsilon\text{ 且}\eta\leq x_0-\varepsilon\right)\\&\geqslant P\left(\eta\leq x_0-\varepsilon\right)-P\left(X_n-\eta>\varepsilon\right)\end{aligned}$$

(因为 $P(A\cap B)\geqslant P(B)-P(\bar{A})$). 故

$$F_n\left(x_0\right)\geqslant F\left(x_0-\varepsilon\right)-P\left(|X_n-\eta|\geqslant\varepsilon\right),$$

$$F_n\left(x_0\right)-F\left(x_0\right)\geqslant F\left(x_0-\varepsilon\right)-F\left(x_0\right)-P\left(|X_n-\eta|\geqslant\varepsilon\right)$$

由于上述两个红色的式子可见,

$$|F_n(x_0) - F(x_0)| \leq F(x_0+\varepsilon) - F(x_0-\varepsilon) + P(|X_n - \eta| \geqslant \varepsilon).$$

由于 x_0 是 $F(x)$ 的连续点, 因此对任何 $\delta > 0$, 有 $\varepsilon > 0$ 满足 $F(x_0+\varepsilon)\ F(x_0-\varepsilon) < \frac{\delta}{2}$. 再取 n_0, 当 $n \geqslant n_0$ 时, $P(|X_n - \eta| \geqslant \varepsilon) < \frac{\delta}{2}$. 于是对一切 $n \geqslant n_0$ 有 $|F_n(x_0) - F(x_0)| < \delta$. 这就证明了 $F_n(x_0) \to F(x_0)\,(n \to \infty)$. 故 $X_n \xrightarrow{w} \eta$. □

同样, 这个定理的逆命题不成立. 我们举出一个反例:

例子 20.4. 设 $X \sim N(0,1)$, 令

$$X_{2n-1} = X, \quad X_{2n} = -X \quad (n \geqslant 1).$$

易知所有的 X_n 有相同的分布函数 $\Phi(x)$. 这个 $\Phi(x)$ 是标准正态分布函数. 当然 $X_n \xrightarrow{w} X$. 但是, 对 $\varepsilon > 0$ 有

$$P(|X_n - X| \geqslant \varepsilon) = \begin{cases} 0, & n \text{ 是奇数}, \\ P\left(|X| \geqslant \frac{\varepsilon}{2}\right), & n \text{ 是偶数}. \end{cases}$$

可见 X_n 并不依概率收敛于 X.

定理 20.7. 设 $X_n \xrightarrow{w} X, \eta_n \xrightarrow{P} 0$, 则 $X_n + \eta_n \xrightarrow{w} X$.

Proof. 设 x_0 是 X 的分布函数 $F(x)$ 的连续点. 对于 $\varepsilon > 0$, 易知

$$\begin{aligned} P(X_n + \eta_n \leq x_0) &\leq P(\eta_n \leq -\varepsilon) + P(X_n \leq x_0 + \varepsilon) \\ &\leq P(|\eta_n| \geqslant \varepsilon) + P(X_n \leq x_0 + \varepsilon), \end{aligned}$$

于是

$$\begin{aligned} &P(X_n + \eta_n \leq x_0) - F(x_0) \\ \leq &P(|\eta_n| \geqslant \varepsilon) + P(X_n \leq x_0 + \varepsilon) - F(x_1 + \varepsilon) \\ &+ F(x_0 + \varepsilon) - F(x_0). \qquad (*) \end{aligned}$$

另一方面,

$$\begin{aligned} P(X_n + \eta_n \leq x_0) &\geqslant P(X_n \leq x_0 - \varepsilon, \eta_n \leq \varepsilon) \\ &\geqslant P(X_n \leq x_0 - \varepsilon) - P(\eta_n > \varepsilon) \\ &\geqslant P(X_n \leq x_0 - \varepsilon) - P(|\eta_n| \geqslant \varepsilon) \qquad (**) \end{aligned}$$

于是

$$\begin{aligned} &P(X_n + \eta_n \leq x_0) - F(x_0) \\ \geqslant &P(X_n \leq x_0 - \varepsilon) - F(x_0 - \varepsilon) + F(x_0 - \varepsilon) \\ &- F(x_0) - P(|\eta_n| \geqslant \varepsilon). \end{aligned}$$

任给定 $\delta > 0$, 取 $\varepsilon_1 > 0$ 足够小使得 $F(x_0 + \varepsilon_1) - F(x_0) < \frac{\delta}{3}$ 且 $x_0 + \varepsilon_1$ 是 $F(x)$ 的连续点 (单

调函数在任何小区间内均有连续点.). 由于 $X_n \xrightarrow{w} X, \eta_n \xrightarrow{P} 0$, 有 n_1 使得对一切 $n \geqslant n_1$,

$$P\left(X_n \leq x_0+\varepsilon_1\right)-F\left(x_0+\varepsilon_1\right)<\frac{\delta}{3}, \quad P\left(\left|\eta_n\right| \geqslant \varepsilon_1\right)<\frac{\delta}{3}.$$

于是, 从 (*) 式知, 当 $n \geqslant n_1$ 时,

$$P\left(X_n+\eta_n \leq x_0\right)-F\left(x_0\right)<\delta. \qquad (\circ)$$

再取 $\varepsilon_2>0$ 使得 $F\left(x_0\right)-F\left(x_0-\varepsilon_2\right)<\frac{0}{3}$ 且 $x_0-\varepsilon_2$ 是 $F(x)$ 的连续点, 于是有 n_2 使得当 $n \geqslant n_2$ 时, 恒有

$$P\left(X_n \leq x_0-\varepsilon_2\right)-F\left(x_0-\varepsilon_2\right)>-\frac{\delta}{3},$$

$$P\left(\left|\eta_n\right| \geqslant \varepsilon_2\right)<\frac{\delta}{3},$$

于是从 (**) 式知, 当 $n \geqslant n_2$ 时, 有

$$P\left(X_n+\eta_n \leq x_0\right)-F\left(x_0\right)>-\delta. \qquad (\circ\circ)$$

从 $(\circ)$ 和 $(\circ\circ)$ 式知 $n \geqslant \max\left(n_1, n_2\right)$ 时,

$$\left|P\left(X_n+\eta_n \leq x_0\right)-F\left(x_0\right)\right|<\delta.$$

这就证明了 $\lim_{n\to\infty} P\left(X_n+\eta_n \leq x_0\right)=F\left(x_0\right)$. 故 $X_n+\eta_n \xrightarrow{w} X$. □

定理 20.8. 设 $X_n \xrightarrow{w} X, \eta_n \xrightarrow{P} 1$, 则

$$X_n \eta_n \xrightarrow{w} X.$$

有了这些定义, 我们可以探究一些当 n 特别大的时候, 会发生什么.

21 大数定律

大数定律及其重要推论 我们先来看构建概率极限定理的一些工具.

定理 21.1 (Chebyshev 大数律)**.** 设 $X_1, X_2, \cdots$ 是相互独立的随机变量序列, $\mathbb{E}\left(X_i\right)=\mu_i, \operatorname{var}\left(X_i\right)=\sigma_i^2 (i \geqslant 1)$ 且 $\left\{\sigma_i^2, i \geqslant 1\right\}$ 有界, 设 $S_n=\sum_{i=1}^n X_i$ $(n \geqslant 1)$, 则

$$\frac{S_n-\mathbb{E}\left(S_n\right)}{n} \xrightarrow{P} 0 \quad (n \to \infty).$$

Proof. 设 $\sigma_i^2 \leq M$(一切 $i \geqslant 1$). 利用 Chebyshev 不等式知

$$\begin{aligned} P\left(\left|\frac{S_n-\mathbb{E}\left(S_n\right)}{n}\right| \geqslant \varepsilon\right) &= P\left(\left|S_n-\mathbb{E}\left(S_n\right)\right| \geqslant n\varepsilon\right) \\ &\leq \frac{1}{n^2\varepsilon^2} \operatorname{Var}\left(S_n\right). \end{aligned}$$

由于 $X_1, \cdots, X_n$ 两两不相关, $\operatorname{Var}(S_n) = \sum_{i=1}^n \operatorname{Var}(X_i) \leq nM$. 于是

$$P\left(\left|\frac{S_n - \mathbb{E}(S_n)}{n}\right| \geqslant \varepsilon\right) \leq \frac{M}{n\varepsilon^2} \quad (\text{一切}\varepsilon > 0).$$

由此得证.

□

我们就立刻可以得到这样的一个有用的推论:

推论 21.2. 设 $X_1, X_2, \cdots$ 是相互独立同分布的随机变量序列, $\mu = \mathbb{E}(X_1)$ 和 $\sigma^2 = \operatorname{var}(X_1)$ 都存在, $S_n = \sum_{i=1}^n X_i (n \geqslant 1)$, 则

$$\frac{S_n}{n} \xrightarrow{P} \mu \quad (n \to \infty).$$

推论 21.3. 设单次试验中事件 A 发生的概率是 p, 在 n 次独立试验 $(n \geqslant 2)$ 中 A 发生了 ν_n 次, 则

$$\frac{\nu_n}{n} \xrightarrow{P} p (n \to \infty).$$

Proof. 令

$$X_i = \begin{cases} 1, & \text{第} i \text{ 次试验中} A \text{ 发生}, \\ 0, & \text{第} i \text{ 次试验中} A \text{ 不发生} \end{cases}$$

$(i = 1, 2, \cdots)$, 则 $\frac{\nu_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i$. 由于 $X_1, X_2, \cdots$ 是相互独立同分布的随机变量序列, $\mathbb{E}(X_i) = p, \operatorname{var}(X_i) = p(1-p) (i \geqslant 1)$. 因此成立. □

上面的两个推论, 尤其是第一个推论是我们最常用到的大数律, 其中方差存在的条件可以去掉, 而且可以证明更强的结论, 但证明较复杂, 我们在这里就不证明了. 那么如果对于数学期望不存在的时候, 是否存在常数 a 使得 $\sum_{i=1}^n X_i \xrightarrow{P} a \quad (n \to \infty)$? 我们看一个例子:

例子 21.1. 设 $X_1, X_2, \cdots$ 是相互独立同分布的随机变量序列, 共同分布是柯西分布, 即密度函数是

$$p(x) = \frac{1}{\pi(1 + x^2)}.$$

记 $S_n = \sum_{i=1}^n X_i (n \geqslant 1)$. 可以证明对任何 $n \geqslant 1, \frac{1}{n}S_n$ 与 X_1 有相同的分布函数. 因此对任何实数 a 和 $\varepsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - a\right| \geqslant \varepsilon\right) \equiv P(|X_1 - a| \geqslant \varepsilon) > 0.$$

故 $\frac{S_n}{n}$ 不能以概率收敛于 a.

那么什么时候类似的定律成立? 我们给出如下事实:

定理 21.4 (Cantelli 强大数律)**.** 设 $X_1, X_2, \cdots$ 是相互独立的随机变量序列, $\mathbb{E}(X_i) = \mu_i, \mathbb{E}(X_i - \mu_i)^4 \leq M$ (一切 $i \geqslant 1; M$ 是一个常数), $S_n = \sum_{i=1}^n X_i (n \geqslant 1)$, 则当 $n \to \infty$ 时,

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{\text{a.s.}} 0.$$

其证明较为复杂.

先证明一个引理: $X_1, \cdots, X_n$ 相互独立 ($n \geqslant 2$), 且 $\mathbb{E}(X_i) = 0, \mathbb{E}\left(X_i^4\right) \leq M (i = 1, \cdots, n)$, 则

$$\mathbb{E}\left(\sum_{i=1}^n X_i\right)^4 \leq 3n^2 M$$

Proof. 用数学归纳法. 显然上述式对 $n = 1$ 成立. 设 $n = k$ 时 (2.5) 式成立, 则

$$\begin{aligned}\left(\sum_{i=1}^{k+1} X_i\right)^4 &= \left(\sum_{i=1}^k X_i + X_{k+1}\right)^4 \\ &= \left(\sum_{i=1}^k X_i\right)^4 + 4\left(\sum_{i=1}^k X_i\right)^3 X_{k+1} + 6\left(\sum_{i=1}^k X_i\right)^2 X_{k+1}^2 \\ &\quad + 4\left(\sum_{i=1}^k X_i\right) X_{k+1}^3 + X_{k+1}^4\end{aligned}$$

于是

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{k+1} X_i\right)^4 &= \mathbb{E}\left(\sum_{i=1}^k X_i\right)^4 + 4\mathbb{E}\left(\sum_{i=1}^k X_i\right)^3 \cdot \mathbb{E}(X_{k+1}) \\ &\quad + 6\mathbb{E}\left(\sum_{i=1}^k X_i\right)^2 \cdot \mathbb{E}\left(X_{k+1}^2\right) + 4\mathbb{E}\left(\sum_{i=1}^k X_i\right) \cdot \mathbb{E}\left(X_{k+1}^3\right) + \mathbb{E}\left(X_{k+1}^4\right).\end{aligned}$$

由于$\mathbb{E}\left(X_i^2\right) \leq \left(\mathbb{E}\left(X_i^4\right)\right)^{\frac{1}{2}}, \mathbb{E}(X_i) = 0$ 且$\mathbb{E}\left(\sum_{i=1}^k X_i\right)^2 = \mathbb{E}\left(X_1^2\right) + \cdots +$

$\mathbb{E}\left(X_k^2\right) \leq k\sqrt{M}$, 故

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{k+1} X_i\right)^4 &\leq 3k^2 M + 0 + 6kM + 0 + M \\ &\leq 3(k+1)^2 M\end{aligned}$$

□

下面开始证明定理:

Proof. 注意定理中的条件 $X_i = X_i(\omega)(i \geqslant 1), S_n$

$$S_n(\omega) = \sum_{i=1}^n X_i(\omega). \text{ 令}$$

$$D = \left\{\omega : \sum_{n=1}^\infty \left(\frac{S_n(\omega) - \mathbb{E}(S_n)}{n}\right)^4 \text{ 发散}\right\}.$$

我们来证明 $P(D)=0$. 任给定 $A>0$, 令

$$D_N=\left\{\omega:\sum_{n=1}^{N}\left(\frac{S_n(\omega)-\mathbb{E}\left(S_n\right)}{n}\right)^4>A\right\}\quad(N\geqslant 1),$$

则 $D\subset\bigcup_{N=1}^{\infty}D_N$, 由此有

$$P(D)\leq P\left(\bigcup_{N=1}^{\infty}D_N\right)=\lim_{N\to\infty}P\left(D_N\right).$$

另一方面,

$$AI_{D_N}(\omega)\leq\sum_{n=1}^{N}\left(\frac{S_n(\omega)-\mathbb{E}\left(S_n\right)}{n}\right)^4$$

于是

$$\begin{aligned}\mathbb{E}\left(AI_{D_N}(\omega)\right)&\leq\mathbb{E}\left(\sum_{n=1}^{N}\left(\frac{S_n(\omega)-\mathbb{E}\left(S_n\right)}{n}\right)^4\right)\\&=\sum_{n=1}^{N}\mathbb{E}\left(\frac{S_n(\omega)-\mathbb{E}\left(S_n\right)}{n}\right)^4\\&\leq\sum_{n=1}^{N}\frac{1}{n^4}3n^2M\quad\text{(上述引理)}\\&\leq 3M\sum_{n=1}^{\infty}\frac{1}{n^2}.\end{aligned}$$

由于 $\mathbb{E}\left(AI_{D_N}(\omega)\right)=A\mathbb{E}\left(I_{D_N}(\omega)\right)=AP\left(D_N\right)$, 因此

$$P\left(D_N\right)\leq\frac{3M}{A}\sum_{n=1}^{\infty}\frac{1}{n^2}.$$

令 $N\to\infty$, 从 (2.6) 式得 $P(D)\leq\frac{3M}{A}\sum_{n=1}^{\infty}\frac{1}{n^2}$. 令 $A\to\infty$, 知 $P(D)=0$, 故 $P\left(D^c\right)=1$. 当 $\omega\in D^c$ 时级数 $\sum_{n=1}^{\infty}\left(\frac{S_n(\omega)-\mathbb{E}(S_n)}{n}\right)^4$ 收敛, 从而

$$\lim_{n\to\infty}\frac{S_n(\omega)-\mathbb{E}\left(S_n\right)}{n}=0.$$

这表明定理式成立. □

这就给了我们一个很直接的推论:

推论 21.5. 设 $X_1,X_2,\cdots$ 是相互独立同分布的随机变量序列, $\mu=\mathbb{E}\left(X_1\right)$ 和 $\mathbb{E}\left(X_1^4\right)$ 存在, $S_n=\sum_{i=1}^{n}X_i(n\geqslant 1)$, 则当 $n\to\infty$ 时,

$$\frac{S_n}{n}\xrightarrow{\text{a.s.}}\mu.$$

这个推论是上面的定理的直接结果. 它就回答了我们什么情况下我们可以使用大数定律.

大数定律的应用 大数定律是很多统计方法的理论依据. 例如, 为了估计随机变量 X 的期望, 若 $X_1,X_2,\cdots,X_n$ 是 X 的 n 次观测值, 人们常用平均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 作

为 $\mathbb{E}(X)$ 的估计量 (近似值). 由于强大数律: 当 $n\to\infty$ 时, $\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{\text{a.s.}} \mathbb{E}(X)$, 故 n 较大时用 X 估计 $\mathbb{E}(X)$ 是合理的. 对于 X 的方差, 人们常用 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 作为 $\operatorname{var}(X)$ 的估计量. 利用强大数律知

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 &= \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\frac{\sum_{i=1}^{n}X_i}{n}\right)^2\right\}\\ &= \mathbb{E}\left(X_1^2\right)-\left(\mathbb{E}\left(X_1\right)\right)^2=\operatorname{var}(X)\quad\left(\text{a.s.}\ \right).\end{aligned}$$

这表明, 当 n 较大时用 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 估计 X 的方差是合理的.

除此之外, 大数定律同样告诉了我们使用计算机模拟的可行性. 为了计算随机变量 X 的期望 $\mathbb{E}(X)$, 若能够产生与 X 有相同概率分布的相互独立的随机变量序列 $X_1,X_2,\cdots$, 则据强大数律, $X=\frac{1}{n}\sum_{i=1}^{n}X_i$ 就是 $\mathbb{E}(X)$ 的近似值 (当 n 很大时).

怎样得到与 X 有相同分布的相互独立的随机变量序列 $X_1,X_2,\cdots$ 呢? 设 X 的分布函数是 $F(x),U_1,U_2,\cdots$ 是服从 $(0,1)$ 上均匀分布的相互独立的随机变量序列. 令 $X_i=F^{-1}\left(U_i\right)(i\geqslant 1)$, 这里

$$F^{-1}(u)=\min\{x,F(x)\geqslant u\}\quad(0<u<1)$$

是随机变量 X 的 u 中位数. 并且 $X_1,X_2,\cdots$ 是相互独立同分布的随机变量序列, 联合分布恰好是 $F(x)$.

例子 21.2. 我们在计算积分 $I=\int_a^b f(x)\mathrm{d}x$ 的时候, 可以使用随机模拟法进行计算. 不失一般性, 假定被积函数是非负的. 对于一般情形, 设 $f(x)$ 有下界 A. 令 $f^*(x)=f(x)-A$, 则 $f^*(x)\geqslant 0$ 且

$$\int_a^b f(x)\mathrm{d}x=\int_a^b f^*(x)\mathrm{d}x+A(b-a),$$

故只需考虑非负函数的积分. 设 $u_1,u_2,\cdots$ 是服从 $(0,1)$ 上均匀分布的相互独立的随机变量序列, 令 $\xi_i=a+(b-a)u_i$, 则 ξ_i 服从 (a,b) 上的均匀分布. 依强大数律, 当 $n\to\infty$ 时,

$$\frac{1}{n}\sum_{i=1}^{n}f\left(\xi_i\right)\xrightarrow{a.s.}\mathbb{E}f\left(\xi_1\right).$$

由于 $\mathbb{E}f\left(\xi_1\right)=\int_a^b f(x)\frac{1}{b-a}\,\mathrm{d}x$, 故 n 很大时,

$$\int_a^b f(x)\mathrm{d}x\approx(b-a)\frac{1}{n}\sum_{i=1}^{n}f\left(\xi_i\right).$$

由此可见, 只要得到服从 $(0,1)$ 上均匀分布的随机数 $u_1,\cdots,u_n$, 就可得到 $\int_a^b f(x)\mathrm{d}x$ 的近似值. 并且, 这个方法可推广用于计算高维的数值积分

$$\int\cdots\int_D f\left(x_1,\cdots,x_m\right)\mathrm{d}x_1\cdots\mathrm{d}x_m,$$

具体叙述从略.

22 中心极限定理

许多随机变量是大量的相互独立的随机因素的综合影响之和所形成的. 我们说明, 这种随机变量往往近似地服从正态分布.

我们首先查看独立同分布的情况.

定理 22.1 (中心极限定理)**.** 设 $X_1, \ldots, X_n$ 是 n 个独立同分布的随机变量, 其均值为 μ, 方差为 σ^2。令 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$。那么对于任意的 a 和 b,

$$\lim_{n\to\infty} P\left(a \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le b\right) \xrightarrow{D} \Phi(b) - \Phi(a)$$

这里 Φ 是标准正态分布函数。

Proof. 首先标准化变量: 令 $Z_i = (X_i - \mu)/\sigma$, 那么期望为 0, 方差为 1. 并且得到

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}.$$

要用定理 20.4的依收敛分布性质, 需要说明

$$\lim_{n\to\infty} \mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} Z_i/\sqrt{n}}\right] = \mathrm{e}^{t^2/2}$$

设 $M(t) = \mathbb{E}\left[\mathrm{e}^{tZ_i}\right]$ 是 Z_i 的矩母生成函数, 那么 $Z_i/\sqrt{n}$ 的矩母生成函数为 $\mathbb{E}\left[\mathrm{e}^{tZ_i/\sqrt{n}}\right] = M\left(\frac{t}{\sqrt{n}}\right)$. 由于独立同分布, 有 $\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} Z_i/\sqrt{n}}\right] = \left(M\left(\frac{t}{\sqrt{n}}\right)\right)^n$. 令 $L(t) = \ln M(t)$, 由于 $M(0) = 1$, 得到 $L(0) = 0$.

观察边界点的一阶, 二阶导数:$L'(0) = \frac{M'(0)}{M(0)} = \mathbb{E}(Z_i) = 0$, $L''(0) = \frac{M(0)M''(0) - \left(M'(0)\right)^2}{(M(0))^2} = \mathbb{E}[Z_i^2] = 1$.

现在, 我们只要说明 $(M(t/\sqrt{n}))^n \to \mathrm{e}^{t^2/2}$, 即 $nL(t/\sqrt{n}) \to t^2/2$ 即可. 使用洛必达法则两次就可以成功说明之.

$$\begin{aligned}
\lim_{n\to\infty}\frac{L(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty}\frac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \\
&= \lim_{n\to\infty}\frac{L'(t/\sqrt{n})t}{2n^{-1/2}} \\
&= \lim_{n\to\infty}\frac{-L''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \\
&= \lim_{n\to\infty} L''(t/\sqrt{n})\frac{t^2}{2} \\
&= \frac{t^2}{2}.
\end{aligned}$$

□

实际上, 在数学期望, 方差都存在, 但是不同的时候, 同样也有大数定律成立:

定理 22.2. 设 $X_1, X_2, \cdots$ 是相互独立的随机变量序列, $\mu_i = \mathbb{E}(X_i), \sigma_i^2 = \operatorname{var}(X_i)\ (i \geqslant 1)$ 都存在, 且存在 $r > 2$ 使得下式成立:

$$\lim_{n\to\infty} \frac{1}{B_n^r} \sum_{i=1}^{n} \mathbb{E}\left|X_i - \mu_i\right|^r = 0$$

这里 $B_n = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}(n \geqslant 1)$. 设 $S_n = \sum_{i=1}^{n} X_i (n \geqslant 1)$, 则对一切 x 成立:

$$P\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\operatorname{var}(S_n)}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{u^2}{2}}\, \mathrm{d}u$$

从这两个定理知 (我们的条件实际应用的过程中一般都是满足的), 当 n 较大时,

$$P\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\operatorname{var}(S_n)}} \leq x\right) \approx \Phi(x)$$

我们时常使用这个关系求解问题.

例子 22.1. *一加法器同时收到 20 个噪声电压 $V_k (k = 1, \cdots, 20)$, 兼设它们是相互独立的, 且都服从 $(0, 10)$ 上的均匀分布. 设 $V = \sum_{k=1}^{20} V_k$, 求 $P(V > 105)$*

解答: 由假设知 $\mathbb{E}(V_1) = 5, \operatorname{var}(X_1) = \frac{100}{12}$. 从定理 22.1知 $\frac{V - 20\times 5}{\sqrt{20\times 100/12}}$ 近似服从 $N(0, 1)$.

于是

$$\begin{aligned} P(V > 105) &= P\left(\frac{V - 20\times 5}{\sqrt{20\times 100/12}} > \frac{105 - 20\times 5}{\sqrt{20\times 100/12}}\right) \\ &= P\left(\frac{V - 20\times 5}{\sqrt{20\times 100/12}} > 0.387\right) \\ &\approx 1 - \Phi(0.387) = 0.348. \end{aligned}$$

例子 22.2. *一份考卷由 99 道题组成, 按从易到难的次序排列. 某学生答对第 1 题的概率是 0.99 , 答对第 2 题的概率是 $0.98, \cdots$, 答对第 i 题的概率是 $1 - i/100 (i = 1, 2, \cdots, 99)$. 若规定正确回答 60 道题以上 (含 60 道题) 才算通过考试, 试问: 该学生通过考试的可能性有多大?*

解答: 对 $i = 1, 2, \cdots, 99$, 令

$$X_i = \begin{cases} 1, & \text{若该学生答对第} i \text{ 题}, \\ 0, & \text{若该学生未答对第} i \text{ 题}, \end{cases}$$

则 $P(X_i=1)=p_i=1-i/100, P(X_i=0)=1-p_i$. 显然, 该学生通过考试的可能性由概率 $P\left(\sum_{i=1}^{99} X_i \geqslant 60\right)$ 来刻画. 为了计算这个概率, 我们可以设想还有 $X_{100}, X_{101}, \cdots$ 使得 $\{X_n, n \geqslant 1\}$ 是相互独立的随机变量序列且 X_{99+i} 与 X_{99} 有相同的分布 (一切 $i \geqslant 1$). 易知

$$\mathbb{E}(X_i)=\begin{cases} p_i, & 1 \leq i \leq 99, \\ p_{99}, & i>99 \end{cases}$$

$$\operatorname{var}(X_i)=\begin{cases} p_i(1-p_i), & 1 \leq i \leq 99, \\ p_{99}(1-p_{99}), & i>99. \end{cases}$$

于是当 $n \geqslant 99$ 时,

$$B_n^2 \triangleq \sum_{i=1}^{n} \operatorname{var}(X_i)=\sum_{i=1}^{99} p_i(1-p_i)+(n-99)p_{99}(1-p_{99}).$$

由于 $|X_i-\mathbb{E}(X_i)|^3 \leq (X_i-\mathbb{E}(X_i))^2$, 知

$$\sum_{i=1}^{n} \mathbb{E}|X_i-\mathbb{E}(X_i)|^3 \leq \sum_{i=1}^{n} \mathbb{E}(X_i-\mathbb{E}(X_i))^2=B_n^2.$$

于是

$$\lim_{n\to\infty} \frac{1}{B_n^3} \sum_{i=1}^{n} \mathbb{E}|X_i-\mathbb{E}(X_i)|^3=0.$$

这表明 定理 22.2 要求的条件满足 (取 $r=3$). 易知

$$\mathbb{E}\left(\sum_{i=1}^{99} X_i\right)=\sum_{i=1}^{99} \mathbb{E}(X_i)=\sum_{i=1}^{99} p_i=49.5,$$

$$B_{99}^2=\sum_{i=1}^{99} \operatorname{var}(X_i)=\sum_{i=1}^{99} p_i(1-p_i)=16.665.$$

利用定理 22.2知

$$\begin{aligned} P\left(\sum_{i=1}^{99} X_i \geqslant 60\right) &= P\left(\frac{\sum_{i=1}^{99} X_i-49.5}{\sqrt{16.665}} \geqslant \frac{60-49.5}{\sqrt{16.665}}\right) \\ &= P\left(\frac{\sum_{i=1}^{99} X_i-49.5}{\sqrt{16.665}} \geqslant 2.5735\right) \approx 1-\Phi(2.5735) \\ &= 0.005. \end{aligned}$$

这表明, 该学生通过考试的可能性很小, 大约只有千分之五.

Part VI

随机过程

23 随机过程简介

通常来讲, 随机过程研究的是无穷多个随机变量, 并将他们当做一个整体来探讨. 更正式地说, 随机过程就是

定义 23.1. 给定无穷集 $T \subset (-\infty, +\infty)$. 如果对于每一个 t, 对应一个随机变量 X_t, 则称随机变量簇 $\{X_t, t \in T\}$ 是随机过程. 有时候简称为过程. (或写作 $X(t)$)

例子 23.1. *用 X_t 表示某邮箱每天从 0 时刻到 t 时刻所接到的信件的个数. 那么 $\{X(t), t \in [0, +\infty)\}$ 是随机过程.*

例子 23.2. *水中的花粉在做布朗运动. 这时候, 如果记 X_t 为花粉在时刻 t 所在的位置的坐标, 那么 $\{X_t, t \in [0, +\infty)\}$ 便是一个随机过程.*

一般我们用 E 表示这些 X_t 所取的值组成的集合, E 叫做状态空间. 如果 $X_t = x$, 我们称随机过程 $\{X_t, t \in T\}$ 在时刻 t 处于状态 X. 当 t 是可列无穷集的时候, $\{X_t, t \in T\}$ 称为离散时间的随机过程. 比如上面的例子 $T = \{0, 1, 2, \}$ 当 T 是一个区间 (包括无穷区间) 的时候, $\{X_t, t \in T\}$ 叫做连续时间的随机过程. 比较常见的是 $T = [0, +\infty)$.

给定 T 中的 n 个数 $t_1, t_2, \cdots, t_n$, 记 $(X_{t_1}, X_{t_2}, \cdots, X_{t_n})$ 的分布函数为 $F_{t_1,t_2,\cdots,t_n}(x_1, x_2, \cdots, x_n)$. 这种分布函数的全体 $\{F_{t_1,t_2,\cdots,t_n}(x_1, x_2, \cdots, x_n), n \geq 1, t_1, t_2, \cdots, t_n \in T\}$ 称为有限维分布簇. 这个分布簇刻画了随机过程的概率的特性.

随机过程其实可以看做时间 t 和可能的结果 ω 的二元函数. 如果我们固定 t, 那么我们就得到了在某一个特定的时间下可能的分布; 相反, 如果我们固定 ω, 那么 $X_t(\omega)$ 就会是 t 的函数. 有时候简写为 $X(t)$. 这个函数就是随机过程的一个 "实现". 我们在一个时间段上面对于随机过程进行观察, 得到的记录就是这样的一个随机过程的实现的一段.

接下来我们来定义两个随机过程的等价性.

定义 23.2. 设两个过程 $\{X_t, t \in T\}, \{Y_t, t \in T\}$ 是随机等价的, 若 $P(X_t = Y_t) = 1$.

24 Markov 链

我们现在看一个重要的问题, 即 Markov 链. 如果我们把现实生活看做状态随时间的依次推演, 那么我们自然希望通过当前的状态推知未来的状态. 这里, 我们做简化: 也就是让下一个时刻处于什么状态仅仅由当前在什么状态决定.

定义 24.1. 如果 E 是一个有限集或可数的无限集, 一个取值为 E 的随机变量序列 $\{X_n, n \geq 0\}$ 是 Markov 链, 如果对于任何非负整数列 $t_1 < t_2 < \cdots < t_n < t_{n+1}$, 以及 E 中的元素 $i_1, i_2, \cdots, t_{n+1}$ 均成立如下的等式:

$$P(X_{t_{n+1}} = i_{n+1} | X_{t_1} = i_1, \cdots, X_{t_n} = i_n) = P(X_{t_{n+1}} = i_{n+1} | X_{t_n} = i_n).$$

当 $P(X_{t_1} = i_1, \cdots, X_{t_n} = i_n) \geq 0$.

这样的一个定义就表明, 状态 X_t 仅仅由 X_{t-1} 决定, 并且并不由 X_{t-2} 等等的历史因素决定. 这个性质叫做 Markov 性质或者无记忆性. 换句话说, X_{t-1} 已经足够告诉我们 X_t 有多少概率会走向哪里了.

不失一般性, 我们可以令集合 E 为 $\{0, 1, 2, \cdots, n\}$. 下面, 我们自然希望得到从一个状态转移到另一个状态的概率是多少. 比如我现在在 i 状态, 下一个时刻可能在 j 的概率

$$P_{i,j} := P(X_t = j | X_{t-1} = i).$$

如果我们把所有的可能的转移都画出来, 也就是 $P_{0,0}, P_{0,1}, \cdots, P_{n,n}$, 就构成了一个矩阵. 我们把它叫做转移矩阵. 用 $\mathbf{P}$ 表示.

$$\mathbf{P} = \begin{pmatrix} P_{0,0} & P_{0,1} & \cdots & P_{0,j} & \cdots \\ P_{1,0} & P_{1,1} & \cdots & P_{1,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ P_{i,0} & P_{i,1} & \cdots & P_{i,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix}$$

根据上面的描述, 我们会发现这个矩阵满足对于任意的 i, 有 $\sum_{j \geq 0} P_{i,j} = 1$. 我们把这个写作矩阵形式, 最方便的一点就是: 在我们经过了很多次迭代之后, 问一问到达每一个状态的概率会稳定吗? 如果稳定, 那么大概是多少? 如果可以稳定的话, 它就称作稳定状态. 下面我们来看一看如何得到稳定状态.

令 $p_i(t)$ 表示在时刻 t 在状态 i 的概率. 把时刻 t 的每一个状态收集到一个向量里面, 就有 $\vec{p}(t) := (p_0(t), p_1(t), p_2(t), \ldots)$. 那么, 要从 $t-1$ 时刻的分布向量 $\vec{p}(t-1)$ 得到 t 时刻的分布向量, 根据这次只与上次相关的性质, 就有对于每一个可能转移过来的状态, 乘上从上个时刻转移到这个时刻的概率, 也就是

$$p_i(t) = \sum_{j \geq 0} p_j(t-1) P_{j,i},$$

为了方便起见, 使用矩阵的形式写出来:

$$\vec{p}(t) = \vec{p}(t-1)\mathbf{P}.$$

所以我们现在知道了从上一时刻到这一时刻的时候概率的变化. 下面, 我们来定义

从 i 到 j, 经历了 m 步, 停在某一个状态的概率:

$$P_{i,j}^{m} := P\left(X_{t+m} = j \mid X_t = i\right)$$

把上面的行向量拼起来, 对于每一行施以同样的操作, 我们就有

$$P_{i,j}^{m} = \sum_{k \geq 0} P_{i,k} P_{k,j}^{m-1}.$$

这时候, 我们就需要依赖经历了 $m-1$ 步的基础上面再走一步.

如果我们使用矩阵的记号进行表述的话, 就会有更神奇的. 令 $\mathbf{P}^{(m)}$ 表示经过 m 步转移之后的概率. 这个矩阵中的每一项就表示 $P_{i,j}^{\mathbf{m}}$, 这时候把上面的求和号的式子改写为矩阵的形式, 就有:

$$\mathbf{P^{(m)}} = \mathbf{P} \cdot \mathbf{P^{(m-1)}};$$

如果我们对它施以归纳法, 就得到了

$$\mathbf{P^{(m)}} = \mathbf{P}^{m}.$$

这就表明对于任意的 $t \geq 0, m \geq 1$, 有

$$\vec{p}(t+m) = \vec{p}(t)\mathbf{P}^{m}$$

我们可以使用有向图来表示 Markov 链的转移过程.

Part VII

数理统计简介

25 数理统计的基本概念

1. 引入: 什么是数理统计 (摘自 [?]) Freedman 在他的《统计学》中生动的描述了统计学研究什么:

> (数理) 统计学是什么? 统计学[3]是对令人图惑费解的问题作出数学设想的艺术. 应该怎样设计实验来演定新药的疗效? 什么东西引起父母与孩子之间的相像, 并且那种力量有多强? 通货澎胀率如何测定? 失业率呢? 它们怎样联系起来? 赌场为什么在轮盘赌上得益? 盖洛普民意测验怎么能够使用仅仅几千人的样本预测选举结果?

可以看出, 统计学研究的对象十分广泛. 与先前学过的概率论相比, 统计学更加关心由我们的 "数据" 是如何得到 "模型" 的. 而非概率论关心的由 "模型" 得到 "数据".

[3]"统计学" 与 "数理统计学" 实质是同一个学科. 通常, 在统计研究者强调数学方法时, 将 "统计学" 前面加上一个 "数理" 的形容词.

首先我们的研究对象是数据. 什么是数据呢? 广义地讲, 数据培贸我们在实际工作中的记录. 例如某工厂为考查灯泡的使用寿命, 随机地抽取了 18 台产品做试验, 测得寿命数据 (单位: h) 如下:

$$17, 29, 50, 68, 100, 130, 140, 270, 280, 340,$$
$$410, 450, 520, 620, 190, 210, 800, 1100.$$

这 18 个寿命数据就是我们的研究对象. 若不对这些数据进行合理的抽象, 就不可能对这些数据进行深层次的分析, 进而从中获得更多的信息. 在数据处理时我们通常用 x 表示数据, 这里的 x 既可以是一个数, 也可以是一个向量或其他的量. 当明确表示向量时, 我们用 $\vec{x}$ 表示之. 这时数据的主要形式是 $\vec{x} = (x_1, \cdots, x_n)$.

在实际问题中, 有时候单一个字母是不够用于表达数据的. 例如, 在连续 10 天的气象记录中, 得到 $m_1, \cdots, m_{10}, M_1, \cdots, M_{10}$, 其中 $m_i (i = 1, \cdots, 10)$ 是每天的最低气温, $M_i (i = 1, \cdots, 10)$ 是每天的最高气温. 此时的数据为 $\vec{x} = \{(m_i, M_i), i = 1, \cdots, 10\}$. 在学习统计学的时候, 用 $\vec{x} = (x_1, \cdots, x_n)$ 表示数据是最方便并且能够抓住数据本质的一种方法. 本节开头引入的寿命数据可表达成 $\vec{x} = (x_1, \cdots, x_{18})$ 或 $\vec{x} = (x_1, \cdots, x_n), n = 18$.

引入数据的概念以后, 我们要记住统计工作的核心任务是对数据进行分析, 进而对所考查的问题作出推断. 在寿命数据的问题中我们的任务是考察该厂生产的电子产品的使用寿命. 我们收集到的 18 台电子产品的寿命数据是该厂生产的一部分产品的数据. 此处特别强调, 我们的目的是要了解该厂生产的电子产品的使用寿命, 而不是这 18 台产品的使用寿命. 这 18 台产品的使用寿命是已经明摆着的数据, 不必再进行细究. 为了研究产品的使用寿命, 我们必须弄明白, 什么是工厂生产的电子产品的使用寿命, 而且还要弄清楚这 18 台产品的寿命与该厂生产的电子产品的使用寿命之间的联系. 最后我们要以这 18 台产品的位用寿命为依据, 对该厂生产的电子产品的使用寿命作某些推断.

根据经验知, 一个工厂生产的产品的使用寿命是带随机性的. 因此, 我们把一个工厂所生产的电子产品的使用寿命 X 看成一个随机变量. 作这样的拙象以后, 可以把人们思想中直观的概念精确化成为一个数学概念. 若没有这种抽象, 就不可能对工厂生产的电子产品的使用寿命进行精确地研究. 什么是我们所需要的信息? 随机变量的某些持征是我们最关心的. 例如, X 的期望 $\mathbb{E}(X), \mathbb{E}(X)$ 越大, 说明产品的使用寿命越长. $\mathbb{E}(X)$ 的大小说明该厂生产的产品质量. 除了 $\mathbb{E}(X), X$ 的标准差 $\sigma(X) = \sqrt{\mathrm{var}(X)}$ 也是一个很重要的指标. 当然 X 的分布体现了工厂所生产产品的使用寿命的全部信息, 因此若我们要了解工厂生产的电子产品的使用寿命, 只需了解随机变量 X 的分布. 而数据又是什么? 它与随机变量 X 的关系是什么? 从数据形成的过程可知, 电子产品的寿命 x_1 是工厂生产的某台产品的寿命, 它是 X 的一个观察值, 也可以看做是与 X 同分布的随机变量 X_1 的观察值. 同样 $x_n (n = 2, \cdots, 18)$ 是与 X 同分布的随机变量 $X_n (n = 2, \cdots, 18)$ 的观察值. 这样, $x = (x_1, \cdots, x_{18})$ 是 $\boldsymbol{X} = (X_1, \cdots, X_{18})$ 的观察值. 有经验的实际工作者一定会明白, 我们收集 18 台数据的目的是为了了解工厂所生产产品的质量, 所以在采样时一定不会为某种利益去故意选择好的产品或坏的产品进行检查. 因此所选的产品一定是代表工厂产品质量的随机变量 X 的观察值, 而且这 18 台产品也是相互独立地

采样而得到的. 用数学的语言来描述, $X_1, \cdots, X_{18}$ 为相互独立且同分布的随机变量序列, 而其共同分布与 X 的分布相同. 由于数据 $\boldsymbol{x}$ 与 X 有这样一层关系, 我们就指望从 $\boldsymbol{x}$ 得到 X 的分布信息.

2. 基本的概念和定义

由上面的例子, 我们可以看出:

定义 25.1. 在数理统计中, 把研究对象的全体称为总体, 组成总体的每个元素称为个体. 总体中随机抽取若干个个体构成的集合称为总体的样本. 样本所含个体的个数称为样本容量.

其中一类很重要的问题是, 根据以往的经验, 我们知道它大概是什么分布, 但是没有办法确定它的参数. 比如, 在考察某电子设备的使用寿命的时候. 我们通过观察猜测 X 服从指数分布, 即

$$X \sim \frac{1}{\theta} \exp\left\{-\frac{x}{\theta}\right\}, \quad x > 0, \theta > 0.$$

但现在我们并不知道 θ 是多少. 有没有办法让我们得到这个 θ 的值呢? 这就是参数估计的概念.

26 极大似然估计 (最大似然估计)

把上一节的内容总结一下. 我们通常用

$$X = (X_1, \cdots, X_n) \sim P_0, \quad \theta \in \Theta$$

的形式来表示一个统计模型. 这样的一个统计模型给出了若干个备选的样本分布 Θ, 但是并没有告诉我们应该选择哪一个 θ 作为我们的估计. 因此, 如何选择 θ 就成了我们要解决的问题.

根据离散和连续之别, 我们给出离散和连续的统计模型的定义:

定义 26.1. 设 $(X_1, \cdots, X_n)$ 为互相独立选取的样本, 其中 $X_i (i = 1, \cdots, n)$ 为离散型随机变量, 样本分布列具有下列一般形式:

$$P_\theta\left((X_1, \cdots, X_n) = (x_1, \cdots, x_n)\right) = \prod_{i=1}^{n} P_\theta\left(X_i = x_i\right), \quad \theta \in \Theta,$$

此处 θ 为参数.

由于 $(X_1, \cdots, X_n)$ 的分布与 θ 有关, 常把其有关事件的概率 $P(\cdot)$ 记为 $P_\theta(\cdot)$.

定义 26.2. 对于连续型随机变量而言, 此时 $X_i(i=1,\cdots,n)$ 为连续型随机变量, 样本 $(X_1,\cdots,X_n)$ 具有联合密度

$$\prod_{i=1}^{n} p(x_i,\theta), \quad \theta\in\Theta$$

当观察值 $(x_1,\cdots,x_n)$ 得到以后, 在许多待选的总体参数 θ 中, 哪个与此数据最匹配呢? 比如在离散的时候, 事件 $\{(X_1,\cdots,X_n)=(x_1,\cdots,x_n)\}$ 的概率. 我们希望挑选使 $P_\theta((X_1,\cdots,X_n)=(x_1,\cdots,x_n))$ 达到最大的 θ 值作为真值的估计. 也就是调整 θ, 使得我们上述定义中的两个式子的值达到最大值.

为了方便起见, 我们把上述的两个式子中表示的称作似然函数 $L(\theta)$: 即离散情况下 $L(\theta):=P_\theta((X_1,\cdots,X_n)=(x_1,\cdots,x_n))=\prod_{i=1}^{n}P_\theta(X_i=x_i), \quad \theta\in\Theta$; 连续情况下 $L(\theta):=\prod_{i=1}^{n}p(x_i,\theta), \quad \theta\in\Theta$.

这就引出我们的第一种估计参数的方法: 极大似然估计法 (Maximal Likelihood Estimation). 为了简单起见, 我们称他为 ML 估计.

定义 26.3. 设 $\theta\in\Theta$ 为统计模型 $(X_1,\cdots,X_n)\sim P_\theta$ 的参数. 统计模型可为连续型, 也可为离散型. 设 $x_1,\cdots,x_n$ 为总体的样本值. 若存在 $\hat{\theta}(x_1,\cdots,x_n)$ 使得

$$L\left(\hat{\theta}(x_1,\cdots,x_n)\right)=\max_{\theta\in\boldsymbol{\theta}}L(\theta),$$

其中 $L(\cdot)$ 为 定义 26.1 或 定义 26.2 所给出的似然函数, 则称 $\hat{\theta}(x_1,\cdots,x_n)$ 为 θ 的最大似然估计 (简称 ML 估计).

需要说明的是, 这种估计方法并不能称为绝对的准则. 但是它是对于大多数统计模型都可以工作的一个方法. 我们用如下的例子具体体会一下其感觉.

例子 26.1. 假设一个坛子内有 3 个球, 其中有黑球, 也有白球. 此时只有两种情况发生, 坛子中有一个黑球和两个白球或两个黑球和一个白球, 但不知道那一种情况是真实的. 我们用 θ 表示坛子中黑球的个数, 则 θ 只可能有两种情况, $\theta=1$ 或 $\theta=2$. 现在从中随机地抽取一个球, 用 X 表示摸到球的状况: $X=1$ 表示摸到的是黑球, $X=0$ 表示摸到的是白球. X 是一个随机变量, 它的分布就刻画了坛子之中黑、白球的分布状况. 设 $\theta=1$, 即坛子内有一个黑球和两个白球, 此时, X 的分布为

$$P_1(X=1)=\frac{1}{3}, \quad P_1(X=0)=\frac{2}{3}$$

当 $\theta=2$ 时, 即坛子中有两个黑球和一个白球, X 的分布为

$$P_2(X=1)=\frac{2}{3}, \quad P_2(X=0)=\frac{1}{3}.$$

由此看出, 坛子内球的状况不同, X 的分布也不同. 由 X 的分布可以定坛子内球

的状态. 这样, 我们建立了一个统计模型

$$P_\theta(X=1)=\frac{\theta}{3},\quad P_\theta(X=0)=1-\frac{\theta}{3},\quad \theta=1,2.$$

这个问题原本是一个猜测问题, 坛子中有几个黑球是不知道的, 即 $\theta=1$ 或 $\theta=2$ 是未知的. 现在假定我们摸到的是一个黑球, 即事件 $\{X=1\}$ 发生. 参数 θ 的取值有两种可能, 当参数 $\theta=2$ 时, $P_2(X=1)=\frac{2}{3}$; 当参数 $\theta=1$ 时, $P_1(X=1)=\frac{1}{3}$. 究竟猜 $\theta=2$, 还是猜 $\theta=1$? 根据最大似然的思想, 毫无疑问, 选择 $\hat{\theta}=2$, 即认定坛子内有两个黑球. 事实上, 从直观的角度看来, 取 $\hat{\theta}=2$ 的理由也是十分充分的. 设想当 $\theta=2$, 即坛子中有两个黑球时, 摸到黑球的可能性比当 $\theta=1$, 即坛子中有一个黑球时, 摸到黑球的可能性大. 要我们猜 $\theta=1$ 或 $\theta=2$ 时, 当然应该猜成 $\hat{\theta}=2$. 类似地, 当我们摸到的是一个白球时, $P_2(X=0)=\frac{1}{3}, P_1(X=0)=\frac{2}{3}$, 此时应选择 $\hat{\theta}=1$, 即认定坛子中有两个白球和一个黑球. 注意, $\hat{\theta}$ 不过是一个估计, 我们不能确知坛子内黑球个数的真实情况, 除非一次摸出两个球来观察它们的颜色. 这就是最大似然基本思想的来历.

理解了这样的思想, 我们来进行简单的应用.

例子 26.2. 设 $X\sim b(1,p).X_1,X_2,\cdots,X_n$ 是来自 X 的一个样本, 试求参数 p 的最大似然估计量.

解答: 设 $x_1,x_2,\cdots,x_n$ 是相应于样本 $X_1,X_2,\cdots,X_n$ 的一个样本值. X 的分布律为

$$P(X=x)=p^x(1-p)^{1-x},\quad x=0,1.$$

故似然函数为

$$L(p)=\prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i}=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

取对数, 求最大值, 有 $\ln L(p)=\left(\sum_{i=1}^{n}x_i\right)\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p),$

$$令\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p)=\frac{\sum_{i=1}^n x_i}{p}-\frac{n-\sum_{i=1}^n x_i}{1-p}=0,$$

解得 p 的最大似然估计值

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}.$$

p 的最大似然估计量为

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}.$$

例子 26.3. 设 $X\sim N(\mu,\sigma^2),\mu,\sigma^2$ 为未知参数, $x_1,x_2,\cdots,x_n$ 是来自 X 的一个样本值. 求 μ,σ^2 的最大似然估计量.

解答: X 的概率密度为

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right],$$

似然函数为

$$\begin{aligned}L\left(\mu,\sigma^2\right)&=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right]\\&=(2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right].\end{aligned}$$

而

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2.$$

$$\begin{cases}\frac{\partial}{\partial\mu}\ln L=\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i-n\mu\right)=0,\\ \frac{\partial}{\partial\sigma^2}\ln L=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2=0.\end{cases}$$

由前一式解得 $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$, 代入后一式得 $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$. 因此得 μ 和 σ^2 的最大似然估计量分别为

$$\hat{\mu}=\bar{X},\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$$

上面的问题我们都没有或者很少用到约束条件. 下面我们来看一看用到约束条件的情形. 毕竟根据定义, 我们的目标是最大化, 而不是取对数之后求导.

例子 26.4. *设总体 X 在 $[a,b]$ 上服从均匀分布, a,b 未知, $x_1,x_2,\cdots,x_n$ 是一个样本值. 试求 a,b 的最大似然估计量.*

解答: 记 $x_{(1)}=\min\left\{x_1,x_2,\cdots,x_n\right\},x_{(n)}=\max\left\{x_1,x_2,\cdots,x_n\right\}.X$ 的概率密度是

$$f(x;a,b)=\begin{cases}\frac{1}{b-a}, & a\le x\le b,\\ 0, & \text{其他.}\end{cases}$$

似然函数为

$$L(a,b)=\begin{cases}\frac{1}{(b-a)^n}, & a\le x_1,x_2,\cdots,x_n\le b,\\ 0, & \text{其他.}\end{cases}$$

把 $x_1,\cdots,x_n$ 按照大小从小到大重排为 $x_{(1)},\cdots,x_{(n)}$. 由于 $a\le x_1,x_2,\cdots,x_n\le b$, 等价于 $a\le x_{(1)},\cdots,x_{(n)}\le b$. 似然函数可写成

$$L(a,b)=\begin{cases}\frac{1}{(b-a)^n}, a\le x_{(1)}, & b\geqslant x_{(n)},\\ 0, & \text{其他.}\end{cases}$$

于是对于满足条件 $a \leq x_{(1)}, b \geqslant x_{(n)}$ 的任意 a, b 有

$$L(a,b) = \frac{1}{(b-a)^n} \leq \frac{1}{\left(x_{(n)} - x_{(1)}\right)^n}.$$

即 $L(a,b)$ 在 $a = x_{(1)}, b = x_{(n)}$ 时取到最大值 $\left(x_{(n)} - x_{(1)}\right)^{-n}$. 故 a, b 的最大似然估计值为

$$\hat{a} = x_{(1)} = \min_{1 \leq i \leq n} x_i, \quad \hat{b} = x_{(n)} = \max_{1 \leq i \leq n} x_i.$$

a, b 的最大似然估计量为 $\hat{a} = \min_{1 \leq i \leq n} X_i, \quad \hat{b} = \max_{1 \leq i \leq n} X_i$.

27 矩估计

设 $X_1, \cdots, X_n$ 为来自总体 $X \sim F_\theta(\theta \in \theta)$ 的一个样本. 根据大数定律有

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}(X) := \mathbb{E}_\theta(X)$$

由于上式中 X 的分布与参数 θ 有关, 常把 X 的期望 $\mathbb{E}(X)$ 记为 $\mathbb{E}_\theta(X)$.

由于当 n 充分大时 $\bar{X}$ 与 $\mathbb{E}_\theta(X)$ 非常靠近, 我们干脆就利用 $a_1 \triangleq \bar{X}$ 作为 $\alpha_1 \triangleq \mathbb{E}_\theta(X)$ 的估计, 这个估计称为 $\mathbb{E}_g(X)$ 的矩估计. 通裳 $\alpha_t \triangleq \mathbb{E}_\theta\left(X^l\right)$ 称为 l 阶总体矩, 而 $a_l \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i^l$ 称为 l 阶样本矩. 由大数定律知, 可用各阶样本矩去估计相应的总体矩. 由此而得矩估计这个名称. 依这个思想, 可利用

$$a_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

作为 $\alpha_2 = \mathrm{E}_\theta(X^2)$ 的估计. 只要这些矩存在. 下面我们给这样的想法一个定义.

定义 27.1. 设 $X_1, \cdots, X_n$ 为来自总体 $X \sim F_\theta(\theta \in \Theta)$ 的一个样本. 若所涉及的矩存在, 则

(1) l 阶样本矩 $a_l = \frac{1}{n} \sum_{i=1}^{n} X_i^l$ 为相应的总体矩 $\alpha_l = \mathbb{E}_\theta\left(X^l\right)$ 的矩估计, $l = 1, 2, \cdots$;

(2) 若存在连续函数 ϕ 使 $g(\theta) = \phi\left(\alpha_1, \cdots, \alpha_k\right)$ 成立, 则 $g(\theta)$ 的矩估计定义为

$$\widehat{g(\theta)} = \phi\left(a_1, \cdots, a_k\right),$$

其中 a_l 为相应于总体矩 $\alpha_l (l = 1, \cdots, k)$ 的样本矩.

同样的, 我们给出一个简单的例子说明它的原理.

例子 27.1. 在某炸药制造厂, 一天中发生着火现象的次数 X 是一个随机变量, 假设它服从以 $\lambda > 0$ 为参数的泊松分布, 参数 λ 为未知. 现有以下的样本值, 试估计参数 λ.

着火次数 k	0	1	2	3	4	5	6	
发生 k 次着火的天数 n_k	75	90	54	22	6	2	1	$\sum = 250$

解答: 用样本均值来估计总体均值.

$$\begin{aligned}\bar{x} &= \frac{\sum_{k=0}^{6} k n_k}{\sum_{k=0}^{6} n_k} \\ &= \frac{1}{250}[0\times 75 + 1\times 90 + 2\times 54 + 3\times 22 + 4\times 6 + 5\times 2 + 6\times 1] \\ 0 &= 1.22, \text{ 即} E(X) = \lambda \text{ 的估计为} 1.22.\end{aligned}$$

让我们回顾一下上一节中做过的一些例子, 看看矩估计和 ML 估计得到的结果是否一致.

例子 27.2. *设总体 X 的均值 μ 及方差 σ^2 都存在, 且有 $\sigma^2 > 0$. 但 μ, σ^2 均为未知. 又设 $X_1, X_2, \cdots, X_n$ 是来自 X 的样本. 试求 μ, σ^2 的矩估计量.*

解答: 由于 $\begin{cases} \mu_1 = E(X) = \mu, \\ \mu_2 = E(X^2) = D(X) + [E(X)]^2 = \sigma^2 + \mu^2. \end{cases}$ 解得

$$\begin{cases} \mu = \mu_1, \\ \sigma^2 = \mu_2 - \mu_1^2. \end{cases}$$

分别以 A_1, A_2 代替 μ_1, μ_2, 得 μ 和 σ^2 的矩估计量分别为

$$\begin{aligned}&\hat{\mu} = A_1 = \bar{X}, \\ &\hat{\sigma^2} = A_2 - A_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.\end{aligned}$$

例子 27.3. *设总体 X 在 $[a, b]$ 上服从均匀分布, a, b 未知. $X_1, X_2, \cdots, X_n$ 是来自 X 的样本, 试求 a, b 的矩估计量.*

解答: 即

$$\begin{aligned}&\mu_1 = E(X) = (a+b)/2, \\ &\mu_2 = E\left(X^2\right) = D(X) + [E(X)]^2 \\ &= (b-a)^2/12 + (a+b)^2/4. \\ &\begin{cases} a + b = 2\mu_1, \\ b - a = \sqrt{12\left(\mu_2 - \mu_1^2\right)}. \end{cases}\end{aligned}$$

解这一方程组得

$$a = \mu_1 - \sqrt{3\left(\mu_2 - \mu_1^2\right)}, \quad b = \mu_1 + \sqrt{3\left(\mu_2 - \mu_1^2\right)}.$$

分别以 A_1, A_2 代替 μ_1, μ_2, 得到 a, b 的矩估计量分别为 (注意到 $\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2\Big)$

$$\hat{a} = A_1 - \sqrt{3\left(A_2 - A_1^2\right)} = \bar{X} - \sqrt{\frac{3}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2},$$

$$\hat{b} = A_1 + \sqrt{3\left(A_2 - A_1^2\right)} = \bar{X} + \sqrt{\frac{3}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}.$$

可以看到, 对于有些问题, 矩估计和 ML 估计得到的结果是不一样的. 就这个例子而言, 在上一节中我们得到的结果 $\max_{1\le i\le n}\{X_i\}$ 不相符合. 矩估计 $2\bar{X}$ 还有一个很不好的性质, 当出现情况 $2\bar{X} < \max_{1\le i\le n}\{X_i\}$ 时, 就会变得很不合理. 在本例的模型中, 数据不会超过 $[0,\theta]$ 这个范围, 因此 θ 的估计值 $\hat{\theta}$ 也应该满足要求 $x_i \le \hat{\theta}, i = 1, \cdots, n$. 但条件 $2\bar{X} < \max_{1\le i\le n}\{X_i\}$ 说明数据已经超出了 $[0,\hat{\theta}] = [0, 2\bar{X}]$ 这个范围, 出现了矛盾. 这是矩估计的不足之处.

此外, 矩估计不是唯一的. 如果这个例子中我们使用二阶矩估计的话, 就有

$$\alpha_2 = \mathbb{E}_\theta\left(X^2\right) = \frac{1}{\theta}\int_0^\theta x^2 \,\mathrm{d}x = \frac{1}{\theta}\frac{x^3}{3}\Big|_0^\theta = \frac{\theta^2}{3},$$

故 $a_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ 是 $\frac{\theta^2}{3}$ 的矩估计. 由 $\alpha_2 = \frac{\theta^2}{3}$ 解出 $\theta = \sqrt{3\alpha_2}$, 也就是 θ 的矩估计为 $\sqrt{3a_2} = \sqrt{\frac{3}{n}\sum_{i=1}^n X_i^2}$.

既然矩估计不唯一, 我们选择哪个好? 这就是我们接下来要解答的问题.

28 估计的无偏性. 样本数字特征

28.1 估计的无偏性

在前面两节的内容中, 我们介绍了两种估计的方法: 最大似然法和矩估计法. 并且看到使用不同的方法, 结果往往是不同的. 那么, 哪一种方式更好? 为了解决这个问题, 人们提出了估计量的评价标准. 其中之一就是无偏性.

定义 28.1. 设 $X_1, \cdots, X_n \sim \text{iid}\, F(x,\theta), \theta \in \Theta$ 为一个统计模型, $g(\theta)$ 为待估量. 统计量 $T\left(X_1, \cdots, X_n\right)$ 称为 $g(\theta)$ 的无偏估计, 如果 T 满足

$$\mathbb{E}_\theta\left(T\left(X_1, \cdots, X_n\right)\right) = g(\theta), \quad \forall \theta \in \Theta.$$

直观上来看, 无偏估计在平均意义下是准确的.

例子 28.1. 设某工厂产品的不合格品率为固定的常数 $p, p \in (0,1)$, 即在正常的情况下工厂产品的不合格品率保持不变. 产品检验员每天抽取 5 件产品进行检验, 并采用 $X/5$ 作为不合格品率的估计, 此处 X 为 5 件产品中的不合格品件数. 这样不合格品率的估

计值只可能是下列 6 个数之一: $0, 0.2, 0.4, 0.6, 0.8, 1$. 显然估计的精度不会高, 这是由样本量所决定的. 经计算知, $\hat{p} = X/5$ 是 p 的一个无偏估计. 利用大数定律可知, 将每天的估计值进行平均, 长时期的积累, 这个平均值会与真值接近. 无偏估计的优点也就在于此. 对于产品检验员来说, 由一次检验的结果得到的估计不会很精确, 但重复使用无偏估计, 长期积累的平均数会趋于不合格品率的真值.

例子 28.2. 在例 *26.3*中, 我们知道了正态分布的 *ML* 估计为:

$$\hat{\mu} = \bar{X} \quad \text{和} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

现在考虑他们是不是无偏的.

对于 $\hat{\mu} = \bar{X}$, 利用期望的性质和正态分布的定义可得

$$\mathbb{E}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right) = \mathbb{E}\left(X_1\right) = \mu,$$

即 $\hat{\mu}$ 是 μ 的无偏估计.

现在来看 $\hat{\sigma}^2$ 的无偏性. 对其做变形:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[X_i - \mu - (\bar{X} - \mu)\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 + (\bar{X} - \mu)^2 - \frac{2}{n}(\bar{X} - \mu)\sum_{i=1}^{n}\left(X_i - \mu\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - (\bar{X} - \mu)^2.
\end{aligned}$$

因此

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \mu\right)^2\right] - \mathbb{E}\left[(\bar{X} - \mu)^2\right] \qquad (*)$$

由于 $X_1, \cdots, X_n \sim \text{iid}\, N\left(\mu, \sigma^2\right)$, 其中 σ^2 就是总体 $X \sim N\left(\mu, \sigma^2\right)$ 的方差, 故

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \mu\right)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

对于 $(*)$ 式右边第二项, 利用独立随机变量和的方差等于各随机变量方差的和的性质, 可知

$$\mathbb{E}\left[(\bar{X} - \mu)^2\right] = \text{var}(\bar{X}) = \frac{1}{n}\,\text{var}\left(X_1\right) = \frac{1}{n}\sigma^2$$

因此带入 $(*)$ 得到

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

因此, 上述的估计不是 σ^2 的无偏估计. 不过, 我们可以对于其稍作修改, 估计 σ^2

为

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

上例中, 我们将 σ^2 的 ML 估计 $\hat{\sigma}^2$ 作适当的修正, 得到 σ^2 的无偏估计 $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$. 其实, 总体方差 σ^2 的无偏估计 S_n^2 具有普遍适用性, 即使我们所研究的总体不是正态总体, S_n^2 仍然是 $\mathrm{var}(X)$ 的无偏估计. 这是我们对于样本方差的定义.

定理 28.2. 设总体 X 的方差 $\mathrm{var}(X)$ 存在且为有限, $X_1, \cdots, X_n$ 为 X 的一个样本, 则

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

为 $\mathrm{var}(X)$ 的无偏估计.

Proof. 首先将 $\sum_{i=1}\left(X_i - \bar{X}\right)^2$ 进行化简:

$$\begin{aligned}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 &= \sum_{i=1}^{n}\left[X_i - \mu - (\bar{X} - \mu)\right]^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2,\end{aligned}$$

其中 $\mu = \mathbb{E}(X)$. 然后对两边求期望, 得

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n\mathbb{E}\left[(\bar{X} - \mu)^2\right] \\ &= n\,\mathrm{var}\left(X_1\right) - n\,\mathrm{var}(\bar{X}) \\ &= n\,\mathrm{var}\left(X_1\right) - \mathrm{var}\left(X_1\right) \\ &= (n-1)\,\mathrm{var}(X)\end{aligned}$$

□

在上面的推导中, 我们没有用到任何关于正态分布的性质. 因此我们更倾向于选择这个为样本的方差.

28.2 样本的数字特征

于是, 对于样本而言, 我们做如下的定义:

定义 28.3. 若 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 令 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 则称 $\bar{X}$ 为样本均值或样本平均数.

样本均值也是一个统计量. 它是一个随机变量, 表征样本观测值的 "中心". 因此, 当给出一组数据 $x_1, x_2, \cdots, x_n$, 样本均值为 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

既然 $\bar{X}$ 是一个随机变量, 我们可以求出它的均值和方差.

定理 28.4. 若 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 若 $\mathbb{E}(X) = \mu$, $D(X) = \sigma^2$, 那么

- $\mathbb{E}\left(\bar{X}\right) = \mu, D(X) = \frac{1}{n}\sigma^2$.
- 当 n 充分大的时候, $\bar{X}$ 近似服从正态分布 $N(\mu, \sigma^2/n)$.

Proof. 我们只说明 (1). (2) 可以使用独立同分布的中心极限定理说明. 下面说明 (1). 因为 $X_1, X_2, \cdots, X_n$ 是 iid. rv. 所以

$$\mathbb{E}\left(\bar{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} n\mu = \mu.$$

对于方差,

$$D\bar{X} = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\sigma = \frac{\sigma^2}{n}.$$

□

接下来看样本方差. 正如我们之前定义的, 我们选用哪个无偏的进行估计.

定义 28.5. 若 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, 令

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

为样本方差. 且称 $\sqrt{S^2}$ 为样本标准差.

样本方差的期望就是我们的总体方差, 就像我们刚刚证明的一样.

29 统计量与抽样分布

很多统计推断都是基于正态分布的总体假定的. 我们看几个重要的统计量.

1. U 统计量及分布 U 统计量是标准化的正态分布.

定理 29.1. 若总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \cdots, X_n$ 是它的一个样本. 那么统计量

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Proof. 因为 $\mathbb{E}\left(\bar{X}\right) = \mu, D(\bar{X}) = \sigma^2/n$. 由于相互独立且服从正态分布的随机变量的线性组合依然服从正态分布. 所以 $\bar{X} \sim N(\mu, \sigma^2/n)$. 因此 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. □

由于它的分布就是标准正态函数. 因此省略.

2. χ^2 统计量及分布 对于 n 正态总体抽取的样本, 有时候需要两个随机变量之间的"距离", 因此需要考虑平方操作. 他们的平方相加服从什么分布呢? 推导显示, 他们服从 χ^2 分布. 也就是

定义 29.2. 设总体 $X \sim N(0,1)$, $X_1, X_2, \cdots, X_n$ 是它的一个样本. 那么

$$\sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

其中

$$\chi_n^2(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

这里面 n 是一个参数, 表示这个分布的自由度.

对于自由度为 n 的 χ^2 分布, 我们不加证明地给出如下的常用的性质:

定理 29.3.

- 期望和方差:$\mathbb{E}(\chi_n^2) = n, \mathbb{D}(\chi_n^2) = 2n$.
- 当 n 充分大的时候, χ_n^2 分布近似服从正态分布 $N(n, 2n)$.
- 可加性: 设 ξ_1, ξ_2 是相互独立的随机变量, 且 $\xi_1 \sim \chi_{n_1}^2, \xi_2 \sim \chi_{n_2}^2$, 那么

$$\xi_1 + \xi_2 = \chi_{n_1+n_2}^2.$$

并且将我们的样本方差经过类似标准化的操作就能推出这样的结果:

定理 29.4. 设总体 $X \sim N(0,1)$, $X_1, X_2, \cdots, X_n$ 是它的一个样本. 则统计量

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{(n-1)}^2.$$

我们称 $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ 为 χ^2 统计量.

例子 29.1. 设 $X \sim N(0,1)$, $X_1, X_2, \cdots, X_5$ 是来自正态总体 $N(0,9)$ 的样本. 已知统计量

$$Y = a(X_1 + X_2)^2 + b(X_3 + X_4 + X_5)^2$$

服从 χ^2 分布, 求出待定的系数 a, b, 以及统计量的自由度 λ.

解答: 由条件, $X_1 + X_2 \sim N(0,18), X_3 + X_4 + X_5 \sim N(0,27)$. 于是 $(X_1 + X_2)/\sqrt{18}$ 与 $(X_1 + X_2 + X_3)/\sqrt{27}$ 互相独立且服从正态分布 $N(0,1)$. 由于可加性:

$$\left(\frac{X_1 + X_2}{\sqrt{18}}\right)^2 + \left(\frac{X_3 + X_4 + X_5}{\sqrt{27}}\right)^2 = \frac{1}{18}(X_1 + X_2)^2 + \frac{1}{27}(X_3 + X_4 + X_5)^2 \sim \chi_2^2.$$

3. T 统计量及分布 很多时候, 我们并不知道方差. 把 (1.) 中的真实方差 σ 替换为样本方差, 就得到了这样的分布:

定义 29.5. 设总体 $X \sim N(\mu, \sigma^2)$, $X_1, X_2, \cdots, X_n$ 是它的一个样本. 那么统计量

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}.$$

其中,

$$t_{n-1}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}$$

是一个自由度为 n 的 t 分布.

除了这个定义中所述的方法. 实际上还有另一个方法得到这样的分布. 我们这里同样省略它的证明.

定理 29.6. 设 $X \sim N(0,1), Y \sim \chi_n^2$, 并且 X, Y 相互独立. 则

$$t = \frac{X}{\sqrt{Y/n}} \sim t(n).$$

30 区间估计

设 $X_1, \cdots, X_n \sim \text{iid} F_\theta(\theta \in \Theta)$ 为某统计模型, 其中 θ 可为向量参数. 现设我们的目的是估计 $g(\theta)$. 本节以前所讨论的估计是用一个 "点" $T(X_1, \cdots, X_n)$ 去估计 $g(\theta)$ 的值, 称为点估计. 但实际工作者对点估计并不满意, 其主要原因是我们不好把握点估计 $T(X_1, \cdots, X_n)$ 与 $g(\theta)$ 之间的差距. 为解决这个问题, 统计学家提出用置信区间来估计 $g(\theta)$.

例子 30.1. 在测量问题中, 设 $X_1, \cdots, X_n \sim \text{iid}\, N(a, \sigma^2)$, 其中参数 $a \in (-\infty, +\infty), \sigma^2 > 0$. 我们的目的是要知道被测对象 a 的真值. 由于误差的存在, 由统计学的知识知, 要精确地知道 a 的值是不可能的, 因此只能得到 a 的近似值. 若用点估计 $T(X_1, \cdots, X_n)$ 去估计 a 的值, 我们不知道 T 是比 a 大还是比 a 小, 也不知道 T 离 a 有多远. 虽然我们能知道误差的大致分布, 但实际工作者对点估计还是不放心. 不过如果给出两个统计量 $\underline{T}$ 和 $\bar{T}$, 满足 $\underline{T} \leq a \leq \bar{T}$, 这样实际工作者就放心了. 但是 $\underline{T}$ 和 $\bar{T}$ 都是随机变量, 我们不能保证 $P(\underline{T} \leq a \leq \bar{T}) = 1$. 从而, 对给定很小的正数 α, 若能保证 $P(\underline{T} \leq a \leq \bar{T}) \geq 1 - \alpha$, 我们就满意了. 现在, 我们称 $[\underline{T}, \bar{T}]$ 为置信度是 $1 - \alpha$ 的置信区间.

定义 30.1. 设 $X_1, \cdots, X_n \sim \text{iid}\, F_\theta(\theta \in \Theta)$ 为某统计模型, 其中 θ 可为向量参数. 又设 $g(\theta)$ 为 θ 的实值函数 (在统计中 $g(\theta)$ 也称为参数或一维参数).

(1) 设 $\underline{T}$ 和 $\bar{T}$ 为满足条件 $\underline{T} < \bar{T}$ 的两个统计量, $\alpha \in (0,1)$ 为某常数. 若对任意 $\theta \in \Theta$, 有

$$P_\theta(\underline{T} \leq g(\theta) \leq \bar{T}) \geqslant 1 - \alpha,$$

则称 $[\underline{T}, \bar{T}]$ 为 $g(\theta)$ 的置信度是 $1 - \alpha$ 的置信区间.

(2) 设 $\underline{T}$ 为某统计量, $\alpha \in (0,1)$ 为某常数. 若对任意 $\theta \in \Theta$, 有

$$P_\theta(g(\theta) \geqslant \underline{T}) \geqslant 1 - \alpha,$$

则称 $\underline{T}$ 为 $g(\theta)$ 的置信度是 $1 - \alpha$ 的置信下限.

(3) 设 $\bar{T}$ 为某统计量, $\alpha \in (0,1)$ 为某常数. 若对任意 $\theta \in \theta$, 有

$$P_\theta(g(\theta) \leq \bar{T}) \geqslant 1 - \alpha,$$

则称 $\bar{T}$ 为 $g(\theta)$ 的置信度是 $1 - \alpha$ 的置信上限.

由于置信下限和置信上限只是置信区间的一种特殊情况, 因此在以后的讨论中会讨论得少一些.

我们下面只看一种特殊的构造置信区间的方法. 也就是估计单个正态分布总体的置信区间.

例子 30.2. 设 $(X_1, X_2, \cdots, X_n)$ 是取自正态总体 $X \sim N(\mu, \sigma_0^2)$ 的样本. σ_0^2 已知. 求参数 μ 的置信度为 $1 - \alpha$ 的置信区间.

解答: 首先寻找未知参数的一个良好的估计. μ 的最大似然估计是样本均值 $\bar{X}$. 现在要求一个常数 δ 使得

$$P(|X - \mu| < \delta) = 1 - \alpha.$$

我们发现 $\bar{X} \sim \left(\mu, \frac{\sigma_0^2}{n}\right)$. 将其标准化为 Z, 就有

$$Z := \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0,1)$$

于是上式变为

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}\right| < \frac{\delta}{\sigma_0/\sqrt{n}}\right) = P\left(|Z| < \frac{\delta}{\sigma_0/\sqrt{n}} = 1 - \alpha.\right)$$

由于标准正态分布上面的分位数, 知道 $\frac{\delta}{\sigma_0/\sqrt{n}} = z_{\alpha/2}$. 所以当 σ_0 已知的时候, 参数 μ 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\bar{X} - \frac{\sigma_0}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma_0}{\sqrt{n}} z_{\alpha/2}\right)$$

接下来我们考察四个情况, 分别对应方差已知/未知, 均值已知/未知的情形. 其大致的求解步骤正如上例子. 只不过把标蓝的分布更改一下就行了.

我们总结如下表.

待估参数	其它参数	统计量	置信区间
μ	σ^2 已知	$u = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$	$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} u_{\alpha/2}\right)$
μ	σ^2 未知	$t = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$	$\left(\bar{X} \pm \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right)$
σ^2	μ 未知	$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} \cdot \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$
σ^2	μ 已知	$\chi^2 = \frac{(n)S^2}{\sigma^2} \sim \chi^2(n)$	$\left(\frac{(n)S^2}{\chi^2_{\alpha/2}(n)} \cdot \frac{(n)S^2}{\chi^2_{1-\alpha/2}(n)}\right)$