

概率论从挂科到重修

Velen Kong

2019 年 8 月 11 日

摘要

这个系列主要是出于随机过程期末考的爆炸遭遇，果然跳过一年的数学课还是有点勉强，拿到了大一最差数学成绩。于是打算重新复习（预习？）一遍概率论，主要基于

(美) Ross S.M. 著. A First Course in Probability 概率论基础教程(原书第9版) [M] 童行伟 & 梁宝生. 译. 北京:机械工业出版社 2014

也可能涉及

(美) Bertsekas D.P. & Tsitsiklis J.N. 著. Introduction to Probability 概率导论(第二版) [M] 郑中国 & 童行伟. 译. 北京:人民邮电出版社 2009

(俄) Shiryaev A.N. 著. Probability-1 概率:第一卷(修订和补充第三版) [M] 周概容. 译. 北京:高等教育出版社 2007。

（如果有时间也许会顺手往下写随机过程的续集，嗯，一定不会咕咕咕）

1 绪论

绪论就不写具体内容了，简单写一下概率的现代测度定义吧，看不懂也不影响学基础概率论。定义设 Ω 是一个非空集合， \mathcal{F} 是 Ω 中的子集类，如果满足：

- (1) $\emptyset, \Omega \in \mathcal{F}$;
- (2) 若 $A \in \mathcal{F}, A^c \in \mathcal{F}$
- (3) 若 $A_1, A_2, \dots \in \mathcal{F}, \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

那么称 \mathcal{F} 是 Ω 上的 σ -代数，称 (Ω, \mathcal{F}) 为可测空间。

再设映射 $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ ，如果满足：

- (1) $\mathbb{P}(\Omega) = 1$
- (2) 若 $A_1, A_2, \dots \in \mathcal{F}$ 两两互斥，则

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

则称 \mathbb{P} 是 (Ω, \mathcal{F}) 上的概率测度，称 $(\Omega, \mathcal{F}, \mathbb{P})$ 是概率空间。

其实理解起来也不是很难。首先是可测空间，空集和全集肯定需要可测，然后如果某个事件的概率可测，那相反的事件概率肯定也可测，并且一些可测子集的并同样可测。例如最简单的非平凡概率空间，其可测空间只有一个事件：

$$\mathcal{F} = \{\emptyset, \Omega, A, \overline{A}\}$$

然后再让概率满足全集映射到1，互斥事件概率具有可列可加性即可。在上述非平凡概率空间中，测度 \mathbb{P} 为

$$\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1$$

$$\mathbb{P}(A) = p, \mathbb{P}(\overline{A}) = 1 - p$$

几何概率空间则是一个相对而言更加抽象的例子，设 $\Omega = (a, b), \mathcal{F}$ 是由 Ω 中所有开集生成的 σ -代数。令 m 是 (Ω, \mathcal{F}) 的Lebesgue测度，对 $A \in \mathcal{F}$

$$\mathbb{P}(A) = \frac{m(A)}{b - a}$$

那么称 $(\Omega, \mathcal{F}, \mathbb{P})$ 是一个几何概率空间。其中Lebesgue测度可以简单理解为

$$m(c, d) = d - c, \quad (c, d) \subseteq (a, b)$$

接下来我们从测度出发来定义随机变量，逐渐开始不讲人话。

定义设 (Ω, \mathcal{F}) 和 (Ω', \mathcal{F}') 是给定的两个可测空间， ξ 是 Ω 到 Ω' 上的映射。如果对任何 $A \in \mathcal{F}'$ 有 $\xi^{-1}(A) \in \mathcal{F}$ ，则称 ξ 为可测映射。

特别地，当 $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B})$ 时，称 ξ 为可测函数，也叫做随机变量。

其中 \mathcal{B} 是 \mathbb{R} 上的Borel代数。Borel代数属于实变函数内容了，大概就是包含所有开集的最小 σ -代数。

定义设 $\xi : \Omega \rightarrow \mathbb{R}$, $\sigma(\xi) = \xi^{-1}(\mathcal{B})$ 称为由 ξ 生成的 σ -代数。它是令 ξ 成为随机变量的最小 σ -代数。

更一般地，设 T 是一个指标集，对 $i \in T$ ，有 $\xi_i : \Omega \rightarrow \mathbb{R}$ 。

记 $\sigma(\xi_i : i \in T) = \sigma\{\xi_i^{-1}(\mathcal{B}) : i \in T\}$

这即是由 $\{\xi_i^{-1} : i \in T\}$ 生成的 σ -代数，它是使得所有 $\xi_i (i \in T)$ 成为随机变量的最小 σ -代数。

可以粗略理解为将每一个事件映射到实数集上从而形成一个随机变量，如果被映射到的是整数集，那么就相当于离散型随机变量。

搞这么麻烦就是为了给概率一个抽象而精确的数学定义，从测度论开始概率算是脱离了玄学进入了现代数学，这点上概率与统计不同。然而这对于概率论基础学习而言其实并没有什么卵用，我自己到现在也没有完全明白，单纯用来水一下绪论。

2 概率公理和条件概率

2.1 组合分析入门

组合分析(combinatorial analysis)是关于计数的数学理论,也是高中接触到的概率论内容。入门比较简单,这里列一下知识点,不知道的请自行百度。

计数的加法原理和乘法原理、排列计算、组合计算、二项式定理、简单不定方程非负整数解。

其中提一下组合计算,如果将 n 个元素分成大小为 n_1, n_2, \dots 的组,不同组合数为

$$\binom{n}{n_1, n_2, \dots} = \frac{n!}{n_1! n_2! \dots}$$

从而有二项式定理的推广,多项式定理

$$(x_1 + x_2 + \dots)^n = \sum_{\sum n_i = n} \binom{n}{n_1, n_2, \dots} x_1^{n_1} x_2^{n_2} \dots$$

最后列几个组合恒等式,证明读者可留作习题

$$k \binom{n}{k} = (n - k + 1) \binom{n}{k-1} = n \binom{n-1}{k-1}$$

$$\binom{2n}{n} = \sum_{k=0}^n \binom{n}{k}^2$$

$$\binom{n}{k} = \sum_{i=k}^n \binom{i-1}{k-1} \quad n \geq k \quad (\text{费马组合恒等式})$$

$$\sum_{k=1}^n \binom{n}{k} k = 2^{n-1} n$$

$$\sum_{k=1}^n \binom{n}{k} k^2 = 2^{n-2} n(n+1)$$

$$\sum_{k=1}^n \binom{n}{k} k^3 = 2^{n-3} n^2(n+3)$$

$$\binom{n}{n_1, n_2, \dots} = \binom{n-1}{n_1-1, n_2, \dots} + \binom{n-1}{n_1, n_2-1, \dots} + \dots$$

$$\sum_{x_1+\dots+x_r=n} \frac{n!}{x_1!x_2!\dots x_r!} = r^n \quad (\text{考虑}r\text{个字母组成长度为}n\text{的序列种数})$$

2.2 概率论公理

如果看懂了绪论里的测度论基础可以跳过这一部分。如果没有看懂，这一节会用比较简单的方式定义概率。

我们把一个随机事件所有可能的结果构成的集合称为样本空间(sample space)，记为 S 。把样本空间的任一子集 E 称为事件(event)。

从而定义交(intersection)、并(union)、补(complement)、互斥(mutually exclusive)。同时这些事件集合自然满足容斥原理以及交并补运算的德摩根律(De Morgan Law)。

然后我们假定概率论的三大公理

- (1) $0 \leq P(E) \leq 1$
- (2) $P(S) = 1$
- (3) 对于互不相容的事件 E_1, E_2, \dots 有

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

实际上概率论的公理是以

$$P(E) = \lim_{n \rightarrow \infty} \frac{n(E)}{n}$$

作为理论的基础前提的，其中 $n(E)$ 是 n 次实验中，事件 E 发生的次数。因而在测度论出现之前，概率和统计半斤八两，都是玄学，不过如今的概率已经有了严格的抽象理论，统计依然处于鄙视链底端(统计学子暴哭)。

接下来介绍一个有趣的悖论。假设有个无限大的箱子，以及无限个按照正整数编号的球，先考虑第一个实验。

实验一：在离午夜十二时差1秒时放入1-10号球，然后把10号球拿出来；在差 $\frac{1}{2}$ 秒时放入11-20号球，然后把20号球拿出来...

那么显然，最后箱子子里有无限个球。

实验二：在离午夜十二时差1秒时放入1-10号球，然后把1号球拿出来；在差 $\frac{1}{2}$ 秒时放入11-20号球，然后把2号球拿出来...

最后结果是只剩下了一个空箱子，因为任何某个球都会被拿出。详细的严谨证明大概需要用到一定的分析学知识，取出的球和放入的球之间能形成一一对应。

到现在为止似乎都很简单，那么我们来看最后一个实验。

实验三：在离午夜十二时差1秒时放入1-10号球，然后随机把箱子里的一个球拿出来；在差 $\frac{1}{2}$ 秒时放入11-20号球，然后再随机把箱子里的一个球拿出来...

要知道实验三的最后剩下了几个球，我们需要先证明概率论的一条基本定理。

定义事件序列如果满足

$$E_1 \subset E_2 \subset \dots$$

那么称其为递增事件序列。

同样我们可以定义递减事件序列为

$$E_1 \supset E_2 \supset \dots$$

对于递增序列，定义

$$\lim_{n \rightarrow \infty} P(E_n) = \bigcup_{i=1}^{\infty} E_i$$

同样地，对于递减序列定义

$$\lim_{n \rightarrow \infty} P(E_n) = \bigcap_{i=1}^{\infty} E_i$$

那么我们有以下定理成立

$$\lim_{n \rightarrow \infty} P(E_n) = P(\lim_{n \rightarrow \infty} E_n)$$

证明对于递增序列 $\{E_n\}$ ，我们定义事件序列 $\{F_n\}$ 为

$$F_1 = E_1$$

$$F_n = E_n \left(\bigcup_{i=1}^{n-1} E_i \right)^c = E_n E_{n-1}^c, \quad n > 1$$

显然， $\{F_n\}$ 是一系列互斥事件，同时满足

$$\bigcup_{i=1}^n F_i = \bigcup_{i=1}^n E_i, \quad n \geq 1$$

因此

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n P(F_i) = \lim_{n \rightarrow \infty} P\left(\bigcup_{i=1}^n F_i\right) \\ &= \lim_{n \rightarrow \infty} P\left(\bigcup_{i=1}^n E_i\right) = \lim_{n \rightarrow \infty} P(E_n) \end{aligned}$$

对于递减序列同理可证。

回到实验三，我们考虑1号球第 n 次操作后还在箱子里的概率

$$P(E_n) = \frac{9 \times 18 \times 27 \dots (9n)}{10 \times 19 \times 28 \dots (9n+1)}$$

由于之前所证的定理，1号球午夜还在箱子里的概率为

$$P\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n \rightarrow \infty} P(E_n) = \prod_{n=1}^{\infty} \left(\frac{9n}{9n+1}\right)$$

又因为

$$\begin{aligned} \prod_{n=1}^{\infty} \left(\frac{9n+1}{9n}\right) &\geq \prod_{n=1}^m \left(1 + \frac{1}{9n}\right) \\ &> \frac{1}{9} + \frac{1}{18} + \dots + \frac{1}{9m} \\ &= \frac{1}{9} \sum_{i=1}^m \frac{1}{i} \end{aligned}$$

所以极限概率 $P(F_i) = 0$ ，而整个箱子非空的概率为 $P(\bigcup_{i=1}^{\infty} F_i)$ ，利用布尔不等式可得

$$P\left(\bigcup_{i=1}^{\infty} F_i\right) \leq \sum_{i=1}^{\infty} P(F_i) = 0$$

由此我们证明了实验三最终箱子将为空。

2.3 条件概率和独立性

假定 F 发生的条件下 E 发生的条件概率为 $P(E|F)$ ，那么可以有定义对 $P(F) > 0$,

$$P(E|F) = \frac{P(EF)}{P(F)}$$

接下来对条件概率的定义做推广，很自然地得到乘法法则

$$P(E_1 E_2 E_3 \dots) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2) \dots$$

然后我们介绍条件概率的一大应用——贝叶斯公式。

首先我们设 E 和 F 两个事件，我们有

$$E = EF \cup EF^c$$

所以可以得到

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)(1 - P(F)) \end{aligned}$$

推广后可得全概率公式，令 $\{F_n\}$ 表示一系列互斥且穷举的事件

$$P(E) = \sum_{i=1}^n P(EF_i) = \sum_{i=1}^n P(E|F_i)P(F_i)$$

然后我们有贝叶斯公式，令 $\{F_n\}$ 表示一系列互斥且穷举的事件

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(EF_j)}{\sum_{i=1}^n P(E|F_i)P(F_i)}$$

接下来举一个条件概率的应用例子。

某次桥牌比赛上，A和B被指控作弊，控方认为其打的牌与作弊通牌的情况下打的牌是一样的，就像开了上帝视角。辩方认为他们的打法没有作弊的情况下也是最优解。假设控辩双方都是正确的，那么A和B完美的出牌这一项证据能否增大其作弊的概率呢？

答案显然是不能，因为在不作弊的情况下这也是完美打法。我们令 H 是某个假设(如A和B作弊)，令 E 表示某项证据，那么

$$\begin{aligned} P(H|E) &= \frac{P(HE)}{P(E)} \\ &= \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)} \\ &= \frac{P(H)}{P(H) + (1 - P(H)) \frac{P(E|H^c)}{P(E|H)}} \end{aligned}$$

由此，为了研究条件更新对事件概率的影响，我们定义优势比

定义事件 A 的优势比为

$$\frac{P(A)}{P(A)^c} = \frac{P(A)}{1 - P(A)}$$

产生新的证据 E 后, 假设 H 的新优势比为

$$\frac{P(H|E)}{P(H|E^c)} = \frac{P(H)P(E|H^c)}{P(H^c)P(E|H)}$$

然后我们研究独立事件, 如果

$$P(EF) = P(E)P(F)$$

那么称它们是独立的(independent), 否则称为相依的(dependent)。而要注意的, 三个事件之间的独立需要事件之间的任意组合都独立。

如果某个试验的子试验都相互独立同分布, 那么称其为重复试验。

下面我们看一个重要的条件概率试验。

假设盒子中有 $k+1$ 枚硬币, 同时第 i 枚硬币正面朝上的概率为 $\frac{i}{k}$, 那么从盒子中取出一枚硬币, 连续 n 次正面朝上, 求下一次正面朝上的概率。

我们令 C_i 表示取出的是第 i 枚硬币, F_n 表示前 n 次结果都为正面朝上, 则所求的概率为

$$P(H|F_n) = \sum_{i=0}^k P(H|F_n C_i) P(C_i|F_n)$$

那么对于第 i 枚硬币被取出的结果, 我们有

$$P(H|F_n C_i) = \frac{i}{k}$$

$$\begin{aligned} P(C_i|F_n) &= \frac{P(C_i F_n)}{P(F_n)} \\ &= \frac{P(F_n C_i)}{\sum_{j=0}^k P(F_n C_j) P(C_j)} \\ &= \frac{\left(\frac{i}{k}\right)^n \frac{1}{k+1}}{\sum_{j=0}^k \left(\frac{j}{k}\right)^n \frac{1}{k+1}} \end{aligned}$$

因此

$$P(H|F_n) = \frac{\sum_{i=0}^k \left(\frac{i}{k}\right)^{n+1}}{\sum_{j=0}^k \left(\frac{j}{k}\right)^n}$$

当 $k \rightarrow \infty$ 时, 利用积分近似得

$$\frac{1}{k} \sum_{i=0}^k \left(\frac{i}{k}\right)^{n+1} \approx \int_0^1 x^{n+1} dx = \frac{1}{n+2}$$

$$\frac{1}{k} \sum_{j=0}^k \left(\frac{j}{k}\right)^n \approx \int_0^1 x^n dx = \frac{1}{n+1}$$

所以对于 $k \rightarrow \infty$ 有

$$P(H|F_n) = \frac{n+1}{n+2}$$

这个试验被称为拉普拉斯继承准则, 将其抽象为对于 n 个互斥且穷举的假设, 其先验(prior)概率(初始概率)为 $P(H_i)$, 然后由贝叶斯公式得知事件 E 发生后的后验(posterior)概率.

另外我们可以得知事件 E_1 和 E_2 发生后的后验(posterior)概率为

$$P(H_i|E_1E_2) = \frac{P(E_1E_2|H_i)P(H_i)}{\sum_j P(E_1E_2|H_j)P(H_j)}$$

前提是 E_1 和 E_2 对于 H_j 条件独立。

3 离散型随机变量

3.1 基本概念

如果不考虑测度论，那么我们可以简单地把定义在样本空间上的实值函数称为随机变量(random variable)。然后得到分布函数(cumulative distribution function)

$$F(x) = P(X \leq x), \quad -\infty < x < \infty$$

如果一个随机变量的可能取值是至多可数的，那么称为离散型随机变量。对于一个离散型随机变量 X 我们定义概率分布列(probability mass function)

$$p(a) = P(X = a)$$

$$\sum_i p(x_i) = 1$$

定义期望为

$$E(X) = \sum_{x:p(x)>0} xp(x)$$

同时容易证明

$$\begin{aligned} E(g(X)) &= \sum_i g(x_i)p(x_i) = \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i) \\ &= \sum_j y_j P(g(X) = y_j) = E(g(X)) \end{aligned}$$

同时期望满足线性关系

$$E(aX + b) = aE(X) + b$$

随机变量期望的和与它们和的期望相等

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$$

期望也被称为一阶矩(first moment)。接下来介绍二阶矩，也就是方差(variance)。

定义随机变量的方差为

$$\text{Var}(X) = E((X - \mu)^2) = E(X^2) - (E(X))^2$$

方差满足

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

定义随机变量的标准差(standard deviation)为

$$SD(X) = \sqrt{\text{Var}(X)}$$

3.2 二项随机变量

假设 n 次试验, 每次成功的概率为 p , 用随机变量 X 表示成功次数, 那么称 X 是参数为 (n, p) 的二项(binomial)随机变量。其分布列为

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \dots$$

下面研究其一些性质, 先计算期望和方差。

$$E(X^k) = \sum_{i=1}^n i^k \binom{n}{i} p^i (1-p)^{n-i}$$

利用 $i \binom{n}{i} = n \binom{n-1}{i-1}$ 可知

$$\begin{aligned} E(X^k) &= np \sum_{i=1}^n i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} \\ &= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\ &= E((Y+1)^{k-1}) \end{aligned}$$

其中 Y 是参数为 $(n-1, p)$ 的二项随机变量。分别令 $k=1, k=2$ 可得

$$E(X) = np$$

$$E(X^2) = np((n-1)p + 1)$$

$$\text{Var}(X) = np(1-p)$$

同时我们可以递推计算二项分布函数

$$P(X = k+1) = \frac{p}{1-p} \frac{n-k}{k+1} P(X = k)$$

3.3 Poisson随机变量

独立重复进行 n 次试验，每次成功概率为 p ，当 n 充分大而 p 非常小时，成功次数的分布接近参数为 $\lambda = np$ 的Poisson分布。其中参数为 λ 的Poisson分布的分布列为

$$p(i) = e^{-\lambda} \frac{\lambda^i}{i!}$$

下面计算Poisson分布的期望和方差

$$E(X) = \sum_{i=0}^{\infty} \frac{i e^{-\lambda} \lambda^i}{i!} = \lambda \sum_{j=0}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} = \lambda$$

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

$$\text{Var}(X) = \lambda$$

接下来考虑 n 个事件，发生概率相互独立为 p_i 且都很小，则总发生次数近似于参数为 $\sum_{i=1}^n p_i$ 的Poisson分布。这被称为Poisson范例，具体的Poisson过程分解与合成定理将会在随机过程中使用。

Poisson随机变量有着严格且抽象的数学定义。对于一系列事件，如果存在正常数 λ 满足下列条件

- (1) 在任意长度为 h 的时间区间内恰好发生一个事件的概率为 $\lambda h + o(h)$
- (2) 在任意长度为 h 的时间区间内发生两个及以上事件的概率为 $o(h)$
- (3) 任意不相交的时间区间内发生的事件个数相互独立

则称其为强度为 λ 的Poisson过程，在任何长度 t 的时间区间内，发生事件的个数是以 λt 为参数的Poisson随机变量。

证明过程思路比较清晰，我们考虑区间 $(0, t)$ ，将其分成 n 份，我们有

$$\begin{aligned} P(N(t) = k) &= P(\text{其中正好有 } n \text{ 个子区间各发生一次事件}) \\ &\quad + P(N(t) = k, \text{至少有一个区间多于一个事件}) \end{aligned}$$

将右边设为事件A和B

$$\begin{aligned}
 P(B) &\leq P(\text{至少有一个区间多于一个事件}) \\
 &\leq \sum_{i=1}^n P(\text{第}i\text{个区间多于一个事件}) \\
 &= \sum_{i=1}^n o\left(\frac{t}{n}\right) \\
 &= t \frac{o\left(\frac{t}{n}\right)}{\frac{t}{n}}
 \end{aligned}$$

所以对于 $n \rightarrow \infty$ 我们有

$$P(B) \rightarrow 0$$

同时又有

$$P(A) = \binom{n}{k} \left(\frac{\lambda t}{n} + o\left(\frac{t}{n}\right)\right)^k \left(1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right)\right)^{n-k}$$

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

另外Poisson分布还有方便的递推计算公式

$$\frac{P(X = i + 1)}{P(X = i)} = \frac{\lambda}{i + 1}$$

3.4 其他重要分布

在独立重复试验中，假设每次成功概率 p ，令 X 表示首次成功需要的试验次数，则

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

称 X 为参数为 p 的几何(geometric)随机变量，且

$$E(X) = \frac{1}{p}$$

$$Var(X) = \frac{1 - p}{p^2}$$

负二项(negative binomial)随机变量是几何随机变量的一种推广, X 表示直到试验成功 r 次需要的次数, 则

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

$$\begin{aligned} E(X^k) &= \sum_{n=r}^{\infty} n^k \binom{n-1}{r-1} p^r (1-p)^{n-r} \\ &= \frac{r}{p} \sum_{n=r}^{\infty} n^{k-1} \binom{n}{r} p^{r+1} (1-p)^{n-1-(r-1)} \\ &= \frac{r}{p} E((Y-1)^{k-1}) \end{aligned}$$

从而有

$$E(X) = \frac{r}{p}$$

$$Var(X) = \frac{r(1-p)}{p^2}$$

接下来介绍超几何(hyper geometric)随机变量。箱子里有 m 个白球, $N-m$ 个黑球, 从中随机无放回取出 n 个球, X 表示取出的白球个数, 那么有

$$P(X = i) = \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}} \equiv P_i(N)$$

这时候令 $P_i(N)$ 最大, 就得到一个合理估计, 称为极大似然估计(maximum likelihood)。具体求解也很简单, 对于超几何随机变量有

$$\frac{P_i(N)}{P_{i-1}(N)} = \frac{(N-m)(N-n)}{N(N-m-n+i)}$$

显然在 $N \leq \frac{mn}{i}$ 时 $P_i(N)$ 单调递增, 之后单调递减, 所以满足此式的最大整数就是 N 的极大似然估计。

对于超几何随机变量, 我们有

$$\begin{aligned} E(X^k) &= \sum_{i=0}^n i^k P(X = i) = \sum_{i=1}^n i^k \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}} \\ &= \frac{nm}{N} \sum_{i=1}^n i^{k-1} \frac{\binom{m-1}{i-1} \binom{N-m}{n-i}}{\binom{N-1}{n-1}} \\ &= \frac{nm}{N} E((Y+1)^{k-1}) \end{aligned}$$

所以

$$E(X) = \frac{nm}{N}$$

$$Var(X) = np(1-p)(1 - \frac{n-1}{N-1})$$

当 N 对于 n 很大时，可以近似为二项随机变量。

然后介绍一个重要的分布—— ζ 分布。

先介绍Riemann ζ 函数

$$\zeta(s) = \sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^s$$

接下来我们令

$$P(X = k) = C \frac{1}{k^{\alpha+1}}, \quad k = 1, 2, \dots$$

其中

$$C = (\zeta(s))^{-1} = \left(\sum_{k=1}^{\infty} \left(\frac{1}{k}\right)^s\right)^{-1}$$

4 连续型随机变量

4.1 基本概念

如果存在一个非负函数 f 满足对于任何实数集 B ，都有

$$P(X \in B) = \int_B f(x)dx$$

那么称 X 为连续型(continuous)随机变量， f 为随机变量 X 的概率密度函数(probability density function)。

由定义知连续型随机变量取任何固定值的概率都为0。

类似于离散型随机变量，我们定义连续型随机变量的期望和方差为

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

同理有

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

为了证明上述公式，我们需要先证明对于非负随机变量 Y

$$E(Y) = \int_0^{\infty} P(Y > y)dy$$

$$\begin{aligned} \int_0^{\infty} (Y > y)dy &= \int_0^{\infty} \int_y^{\infty} f_Y(x)dx dy \\ &= \int_0^{\infty} \int_0^x dy f_Y(x)dx \\ &= E(Y) \end{aligned}$$

由此我们可以简单证明对于某个非负函数 $g(x)$

$$\begin{aligned} E(g(X)) &= \int_0^{\infty} P(g(x) > y)dy \\ &= \int_0^{\infty} \int_{x:g(x)>y} f(x)dx dy \\ &= \int_{x:g(x)>y} \int_0^{\infty} dy f(x)dx \\ &= \int_{-\infty}^{+\infty} g(x)f(x)dx \end{aligned}$$

同样，离散型随机变量的期望和方差定义及运算法则也适用于连续型随机变量。

4.2 均匀随机变量

如果随机变量 X 的密度函数为

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{others} \end{cases}$$

那么称其为 $(0, 1)$ 上的均匀分布(uniformly distribution)。同理可以推广到区间 (α, β) 上的均匀分布。

对于 (α, β) 上的均匀分布, 我们有

$$E(X) = \int_{\alpha}^{\beta} x f(x) dx = \frac{\alpha + \beta}{2}$$

同时又有

$$E(X^2) = \int_{\alpha}^{\beta} x^2 f(x) dx = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}$$

所以

$$\text{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

4.3 正态随机变量

正态分布(normal distribution), 又称高斯分布(Gaussian distribution)。

我们定义正态随机变量 X 的密度函数为

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

称其为参数为 (μ, σ^2) 的正态随机变量。

我们对于 $y = (x - \mu)/\sigma$ 可知

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy$$

下证

$$\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi}$$

$$\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{y^2+x^2}{2}} dy dx$$

接下来利用极坐标变换可得

$$\int_0^{\infty} \int_0^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr = 2\pi$$

正态分布的线性组合也是正态分布。定义标准正态分布的参数为(0, 1)。下面证明对于任意正态随机变量 X , $Y = aX + b$ 也是正态随机变量, 且参数为 $a\mu + b$ 和 $a^2\sigma^2$ 。

$$F_Y(x) = P(X \leq \frac{x-b}{a}) = F_X(\frac{x-b}{a})$$

求导可知

$$f_Y(x) = \frac{1}{a} f_X(\frac{x-b}{a}) = \frac{1}{\sqrt{2\pi}a\sigma} \exp\{-\frac{(x-b-a\mu)^2}{2a^2\sigma^2}\}$$

标准正态分布的期望为0, 方差则为

$$\begin{aligned} Var(Z) &= E(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \left(-xe^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right) \\ &= 1 \end{aligned}$$

所以正态分布的均值为 μ , 方差为 σ^2 。

一般将标准正态随机变量的分布函数记为

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy$$

$$F_X(a) = \Phi(\frac{a-\mu}{\sigma})$$

当 $np(1-p)$ 较大时, 正态分布可以很好地近似二项分布, 反之可以利用Poisson分布近似。

一个重要的结论就是De Moivre-Laplace定理, 在 n 次独立重复实验中, 设成功概率为 p , 记成功次数为 S_n , 那么当 $n \rightarrow \infty$ 时有

$$P(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b) \rightarrow \Phi(b) - \Phi(a)$$

具体证明可参考中心极限定理。

4.4 指数随机变量

如果随机变量 X 的密度函数为

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

则称为参数为 λ 的指数随机变量。

$$F(a) = \int_0^a \lambda e^{-\lambda x} dx = 1 - e^{-\lambda a}$$

$$\begin{aligned} E(X^n) &= \int_0^\infty x^n \lambda e^{-\lambda x} dx \\ &= -x^n e^{-\lambda x} \Big|_0^\infty + \int_0^\infty n x^{n-1} e^{-\lambda x} dx \\ &= \frac{n}{\lambda} E(X^{n-1}) \end{aligned}$$

所以有

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ Var(X) &= \frac{1}{\lambda^2} \end{aligned}$$

指数分布是仅有的具有无记忆性的分布，我们定义一个随机变量 X 是无记忆的(memoryless)，如果满足

$$P(X > s + t | X > t) = P(X > s)$$

指数型随机变量可变形为双指数型随机变量，又称拉普拉斯随机变量。其密度函数为

$$f(x) = \frac{1}{2} \lambda e^{-\lambda|x|} \quad -\infty < x < +\infty$$

分布函数为

$$F(x) = \begin{cases} \frac{1}{2} e^{\lambda x} & x \leq 0 \\ 1 - \frac{1}{2} e^{-\lambda x} & x > 0 \end{cases}$$

4.5 危险率函数

考虑一个连续型非负随机变量 X ，将其解释为某个零件的寿命，其危险率函数(hazard rate, failure rate)定义如下

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)}, \quad \bar{F} = 1 - F$$

表示使用时间为 t 的零件不能继续使用的概率。对于指数分布

$$\lambda(t) = \lambda$$

同时分布函数可以由危险率函数唯一确定

$$F(x) = 1 - \exp\left\{-\int_0^x \lambda(t) dt\right\}$$

而如果

$$\lambda(t) = a + bt$$

那么有

$$F(t) = 1 - e^{-at - \frac{1}{2}bt^2}$$

$$f(t) = (a + bt)e^{-at - \frac{1}{2}bt^2}$$

$a = 0$ 时, 其被称为瑞利分布(Rayleigh distribution)。

4.6 其他重要分布

设随机变量 X 具有密度函数

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

其中 $\Gamma(\alpha)$ 被称为 Γ 函数, 则称 X 是参数为 (α, λ) 的 Γ 分布。

Γ 函数是对阶乘的一种扩展, 定义为

$$\Gamma(n) = \int_0^{\infty} e^{-y} y^{n-1} dy$$

易知

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)!$$

实践中经常作为Poisson过程中发生 n 个事件所需要的时间分布。所以当 $\alpha = 1$ 时其退化为指数分布。

$$\begin{aligned} E(X) &= \frac{1}{\lambda \Gamma(\alpha)} \int_0^{\infty} \lambda e^{-\lambda x} (\lambda x)^{\alpha} dx \\ &= \frac{\Gamma(\alpha+1)}{\lambda \Gamma(\alpha)} = \frac{\alpha}{\lambda} \end{aligned}$$

而参数为 $(\frac{n}{2}, \frac{1}{2})$ 的 Γ 分布则被称为自由度为 n 的卡方分布, 即 χ^2 分布。

韦布尔分布在生命现象中有着广泛应用，特别当某个对象符合最弱链模型时，其寿命服从韦布尔分布。其分布函数具有如下形式

$$F(x) = \begin{cases} 0 & x \leq \nu \\ 1 - \exp\left\{-\left(\frac{x-\nu}{\alpha}\right)^\beta\right\} & x > \nu \end{cases}$$

称为参数 (ν, α, β) 的韦布尔随机变量。求导得密度函数为

$$f(x) = \begin{cases} 0 & x \leq \nu \\ \frac{\beta}{\alpha} \left(\frac{x-\nu}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x-\nu}{\alpha}\right)^\beta\right\} & x > \nu \end{cases}$$

我们在 $(0, 1)$ 点放置一个激光源，令其与 x 轴夹角为 θ 在 x 轴上得到一点 X ，其中 θ 满足 $(-\frac{\pi}{2}, \frac{\pi}{2})$ 上的均匀分布。那么 X 的坐标分布为

$$F(x) = P(\tan\theta \leq x) = \frac{1}{2} + \frac{\arctan x}{\pi}$$

$$f(x) = \frac{1}{\pi(1+x^2)} \quad -\infty < x < +\infty$$

其满足柯西分布。柯西分布的分布函数为

$$f(x) = \frac{1}{\pi(1+(x-\theta)^2)} \quad -\infty < x < +\infty$$

接下来介绍一个常用于为取值于某个有限区间的随机现象建立模型的 β 分布。其分布函数为

$$f(x) = \begin{cases} \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)} & 0 < x < 1 \\ 0 & \text{others} \end{cases}$$

其中

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

$\beta > \alpha$ 则密度函数向左倾斜，即取小值的概率增大。 $a = b = 1$ 时变为均匀分布，随着 a 和 b 的增大，取值于 $\frac{1}{2}$ 附近的概率也会增大。

可以证明

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

对于参数为 (a, b) 的 β 随机变量有

$$E(X) = \frac{a}{a+b}$$

$$Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

4.7 随机变量函数的分布

定理设 X 是一个连续型随机变量，密度函数为 f_x 。设 $g(x)$ 是单调可微函数，那么随机变量 $Y = g(x)$ 的密度函数为

$$f_x(y) = \begin{cases} f_x(g^{-1}(y)) \left| \frac{d}{dx} g^{-1}(y) \right| & \exists x, y = g(x) \\ 0 & \forall x, y \neq g(x) \end{cases}$$

对于 $g(x)$ 递增的情况证明

$$F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

求导得

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)$$

所以假设 X 是一个非负随机变量，则对于 $g(x) = x^n$ 有

$$f_Y(y) = \frac{1}{n} y^{\frac{1}{n}-1} f(y^{\frac{1}{n}})$$

对于参数为 (μ, σ^2) 的正态随机变量 X ，设

$$Y = e^X$$

称其为参数为 (μ, σ^2) 的对数正态随机变量，即 $\ln(y)$ 满足正态分布。常用于描述证券价格相对于昨天的比率分布。

同时可得

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left\{-\frac{(\ln(y) - \mu)^2}{2\sigma^2}\right\}$$

5 随机变量的联合分布

5.1 联合分布函数

定义两个随机变量 X, Y 的联合概率分布函数(joint cumulative probability distribution function)为

$$F(a, b) = P(X \leq a, Y \leq b) \quad -\infty < a, b < +\infty$$

$$F_X(a) = P(X \leq a, Y < \infty) = \lim_{b \rightarrow \infty} F(a, b)$$

其中 F_X 被称为 X 的边缘分布(marginal distribution)。

理论上所有关于 X, Y 的概率问题都可以由联合分布函数来计算。例如

$$P(X > a, Y > b) = 1 - F_X(a) - F_Y(b) + F(a, b)$$

更进一步有

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F(a_2, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, b_1)$$

当 X, Y 都是离散型随机变量时, 可以很容易定义其联合概率分布列(joint probability mass function)

$$p(x, y) = P(X = x, Y = y)$$

$$p_X(x) = \sum_{y: p(x, y) > 0} p(x, y)$$

另一方面, 我们称 X, Y 是联合连续的(jointly continuous), 如果存在一个定义于 \mathbb{R}^2 上的函数 $f(x, y)$ 满足对任意集合 C 有

$$P((X, Y) \in C) = \iint_{(x, y) \in C} f(x, y) dx dy$$

则称其为 X, Y 的联合密度函数(joint probability density function)。

5.2 独立随机变量

5.3 独立随机变量的和

5.4 离散情形下的条件分布

5.5 连续情形下的条件分布

5.6 次序统计量

5.7 随机变量函数的联合分布

← To be continue