

# L2 Random Variables

这次讨论的内容稍微有点无聊，然后会涉及到比较多求积分和求级数的内容。但是这部分又是概率论模型的基础——如果要研究实际问题，这些模型都是必不可少的。因此，我们这次先比较全面地给出这些模型，之后的讨论班我们再研究这些模型背后的实际问题（这部分要交给你们来讲，所以大家不妨先自己研究研究一些实际问题）。

## 1 Random Variable Basics

**Definition (Random Variable)** A random variable is a function  $X : \Omega \rightarrow \mathbb{R}$  with the property that  $\{\omega \in \Omega : X(\omega) \in \Sigma\}$  for each  $x \in \mathbb{R}$ . Such a function is said to be  $\Sigma$ -measurable.

**Definition (Distribution Functions)** The distribution function of a random variable  $X$  is the function  $F : \mathbb{R} \rightarrow [0, 1]$  given by  $F(x) = \Pr(X \leq x)$ .

分布函数是表示随机变量最通用也是最根本的方式。因为概率密度函数和概率质量函数只能描述部分的随机变量，两者加起来也无法描述所有的随机变量。

**Lemma (Properties of Distribution Functions)**

1.  $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1$
2. if  $x < y$  then  $F(x) \leq F(y)$
3.  $F$  is right-continuous, that is,  $F(x + h) \rightarrow F(x)$  as  $h \downarrow 0$ .

**Note:** In order to prove the third property, the theorem of **continuity of probability measures** is needed.

我们没有提到概率测度的连续性，因为我想避免掉一些比较高等的数学——比方说，集合序列的极限是怎么定义的？大家学过数列和函数的极限是用 $\epsilon - N/\delta$ 语言定义的，但是集合是不能直接作差的。如果对这部分感兴趣的话，大家可以去看一下实分析的教材。

**Definition (Independence of Random Variable)** Random variables  $X$  and  $Y$  are called **independent** if  $\{X \leq x\}$  and  $\{Y \leq y\}$  are independent events for all  $x, y \in \mathbb{R}$ .

### 1.1 Discrete Random Variable

**Definition (Discrete Random Variable)** The random variable  $X$  is called **discrete** if it takes values in some countable subset  $\{x_1, x_2, \dots\}$ , only, of  $\mathbb{R}$ . The discrete random variable  $X$  has (probability) mass function (pmf)  $f : \mathbb{R} \rightarrow [0, 1]$  given by  $f(x) = \Pr(X = x)$ .

## 1.2 Continuous Random Variable

**Definition (Continuous Random Variable)** The random variable  $X$  is called **continuous** if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^x f(u)du, x \in \mathbb{R}$$

for some integrable function  $f: \mathbb{R} \rightarrow [0, \infty)$  called the **(probability) density function (pdf)** of  $X$ .

Note that the word ‘continuous’ is a misnomer when used in this regard: in describing  $X$  as continuous, we are referring to a property of its distribution function rather than of the random variable (function)  $X$  itself.

所以我们是怎麼区分离散和连续的随机变量的呢？我们看它们的分布函数可以怎麼表示，如果能用pmf表示，那就是离散；如果能用pdf表示，那就是连续。不过，有些随机变量的分布函数可能既不是离散的，也不是连续的。

钟开莱的教材没有使用discrete和continuous这两个词来区分这两类随机变量，而是通过“能用pmf表示”和“能用pdf表示”来区分，因为他认为这两个词是不准确的。也许他的这个做法能让大家更好的理解这两类随机变量的区别。

## 1.3 Moment and Deviation

这部分内容（矩与偏差）目前只需要大概了解即可，之后的讨论班我们会深入讨论。

**Definition (Expectation)** The **mean value**, or **expectation**, or **expected value** of the random variable  $X$  with mass function  $f$  is defined to be

$$\mathbb{E}(X) = \sum_{x: f(x) > 0} xf(x)$$

whenever this sum is absolutely convergent.

**Note:**  $\mathbb{E}(X)$  can be denoted as  $\mathbb{E}X$ .

**Theorem (Linearity of Expectation)** if  $a, b \in \mathbb{R}$  then  $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$

注意：期望的线性不需要随机变量独立的前提

**Definition (Variance)** The **variance** of the random variable  $X$  with mass function  $f$  is defined to be

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

计算方差更多使用  $\mathbb{E}(X^2) - \mathbb{E}X^2$ ，因为通常  $\mathbb{E}(X)$  是已知的。

$\mathbb{E}(X)$  称为一阶原点矩， $\mathbb{E}(X^2)$  称为二阶原点矩， $\text{Var}(X)$  称为二阶中心矩。我们之后的讨论班会（比较）系统地研究随机变量的矩。

**Theorem.**  $X$  and  $Y$  are **independent** and both have finite variances, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

## 2 Elementary Models

这部分会介绍一些常见的基础模型。这些模型是复杂模型的基础，也就是说复杂模型通常都是建立在这些基础模型之上。因此，把基础模型的性质（pmf/pdf/cdf，均值，方差等）掌握好是非常重要的。

其中，不少模型可能会很难理解：为什么这样是对的？为什么这样能去拟合某个现象？如果从概率论的历史看，概率论早年的发展和统计是非常相关的。很多概率模型都脱胎于统计学家统计出来的数据。因此，现实似乎不是人们根据现象的原理去设计模型，而是人们根据现象的数据，找到一个能拟合得不错的模型，然后把模型用在新的现象。其实机器学习也有点这种感觉——大家很少会先全面分析一个数据集的性质，然后决定用什么模型去建模；大家可能通常先把所有模型都用上去，看看验证集哪个表现更好，以及检查一下有无过拟合，再拿那个最好的模型作为预测模型。

那我们会倾向于使用什么模型呢？我觉得大概人们大概会考虑几个方面：

1. 准确性（这是废话，完全不准的模型谁想用？）

2. 简洁性

太过复杂的模型可能导致过拟合，而且也在计算上也不是很方便。

3. 良好的性质

为什么人们喜欢使用连续函数，而不是离散函数呢？很多时候就是因为连续函数有很好的性质，比方说无穷阶可导，比方说积分（积分有时候比求级数好求多了）。

至于选择了一个模型之后，怎么根据实际数据确定模型参数，这部分就是统计的内容了：参数估计（例如矩估计和最大似然估计）。

这些模型的分析需要比较好的微积分的基础（有些离散分布可能需要些组合数学的基础），大家需要复习微积分时，不妨拿这些模型来分析分析 😊

### 2.1 Discrete Models

#### 2.1.1 Bernoulli Trials

**Definition (Bernoulli trials)** A random variable  $X$  takes values 1 and 0 with probabilities  $p$  and  $1 - p$ , respectively.  $\mathbb{E}(X) = p$ ,  $\text{Var}(X) = p(1 - p)$ .

#### 2.1.2 Binomial Distribution

**Definition (Binomial distribution)** We perform  $n$  independent Bernoulli trials  $X_1, X_2, \dots, X_n$  and count the total number of successes  $Y = X_1 + X_2 + \dots + X_n$ . The mass function of  $Y$  is:

$$f(k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, 1, 2, \dots, n.$$

$$\mathbb{E}(Y) = np, \text{Var}(Y) = np(1 - p).$$

### 2.1.3 Geometric Distribution

**Definition (Geometric distribution)** A *geometric* variable  $X$  is a random variable with the geometric mass function

$$f(k) = p(1-p)^{k-1}, k = 1, 2, \dots$$

for some number  $p$  in  $(0, 1)$ .  $\mathbb{E}(X) = p^{-1}$ ,  $\text{Var}(X) = (1-p)p^{-2}$ .

### 2.1.4 Poisson Distribution

**Definition (Poisson distribution)** A *Poisson* variable is a random variable with the Poisson mass function

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

for some  $\lambda > 0$ .  $\mathbb{E}(X) = \lambda$ ,  $\text{Var}(X) = \lambda$ .

## 2.2 Continous Models

### 2.2.1 Uniform Distribution (Continous)

**Definition (Uniform Distribution)** The random variable  $X$  is uniform on  $a, b$  if it has density function  $f$ :

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

$$\mathbb{E}(X) = (a+b)/2, \text{Var}(X) = \frac{(b-a)^2}{12}.$$

### 2.2.2 Exponential Distribution

**Definition (Exponential distribution)** The random variable  $X$  is *exponential* with parameter  $\lambda(> 0)$  if it has density function  $f$ :

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

$$\mathbb{E}(X) = 1/\lambda, \text{Var}(X) = 1/\lambda^2.$$

### 2.2.3 Normal Distribution

正态分布是一个有着比较深刻理论背景模型，等我们学习中心极限定理时，我们会对其进行更加深入的研究。比方说，这个密度函数为什么是这个形式？这个问题并不trivial，我当时学概率论的时候，书里没讲，老师也没讲。

The normal (or Gaussian) distribution with two parameters  $\mu$  and  $\sigma^2$  has density function  $f$ :

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

$$\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2.$$

If  $\mu = 0$  and  $\sigma^2 = 1$  then  $f(x)$ :

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

is the density of the *standard* normal distribution.

## 2.3 Summary

Distribution	Parameter	pmf/pdf	$\mathbb{E}(X)$	$\text{Var}(X)$
Bernoulli	$p$	$f(k) = p^k(1-p)^{(1-k)}, k = 0, 1$	$p$	$p(1-p)$
Binomial	$n, k, p$	$f(k) = \binom{n}{k} p^k(1-p)^{n-k}, k = 0, 1, 2, \dots, n.$	$np$	$np(1-p)$
Geometric	$p$	$f(k) = p(1-p)^{k-1}, k = 1, 2, \dots$	$p^{-1}$	$(1-p)p^{-2}$
Poisson	$\lambda$	$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$	$\lambda$	$\lambda$
Uniform (continuous)	$a, b \in \mathbb{N}$	$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise,} \end{cases}$	$(a+b)/2$	$\frac{(b-a)^2}{12}$
Exponential	$\lambda$	$f(x) = \lambda e^{-\lambda x}, x \geq 0$	$1/\lambda$	$1/\lambda^2$
Normal $N(\mu, \sigma)$	$\mu, \sigma$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty < x < \infty$	$\mu$	$\sigma^2$