**概率论** **Spring 2024**

# 第 3 节: 与期望相关的不等式 矩与偏差 高阶矩

*Lecturer:* 尹一通 *Scribes:* 张枕玮

## §1 与期望相关的不等式

**1.1. Markov 不等式.** 当我们求出一个随机变量期望之后, 自然想知道进行一次实验, 随机变量的取值在均值附近的概率是多少. 为了解决这个问题, 首先介绍 Markov 不等式.

**定理 1.1** (Markov 不等式)**.** 令 $X$ 是一个非负的随机变量, 那么对于任意的 $a > 0$, 有

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

证明. 令指示变量 $I := I(X \geq a)$, 由于 $X \geq 0, a > 0$, 有

$$I = I(X \geq a) \leq \left\lfloor \frac{X}{a} \right\rfloor \leq \frac{X}{a}$$

因此,

$$\Pr(X \geq a) = \mathbb{E}[I] \leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a}$$

□

同样可以使用全期望证明:

证明.

$$\begin{aligned}\mathbb{E}[X] &= \underbrace{\mathbb{E}[X \mid X \geq a]}_{X \geq a\text{是可能的}} \cdot \Pr(X \geq a) + \underbrace{\mathbb{E}[X \mid X < a]}_{X\text{是非负的}} \cdot \Pr(X < a) \\ &\geq a \cdot \Pr(X \geq a) + 0 \cdot \Pr(X < a) \quad = a \cdot \Pr(X \geq a) \\ \Longrightarrow \Pr(X \geq a) &\leq \frac{\mathbb{E}[X]}{a}\end{aligned}$$

□

这样就有了上述的不等式. 一个简单的推论就是对任意的 $c > 1$, $\Pr(X \geq c\mathbb{E}[X]) \leq 1/c$. 这体现了期望的偏差的概率. 其中, 如果 $\forall c > 1, \forall \mu \in \mathbb{R}, \exists$ 非负随机变量 $X$ 满足 $\mathbb{E}[X] = \mu$, 这样才可以取得等号.

更进一步, 这也适用于随机变量的函数.

**推论 1.2.** $X$ 是一个随机变量, $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ 是一个值域非负的函数, 那对任意的 $a > 0$, 都有 $\Pr(f(X) \geq a) \leq \frac{\mathbb{E}[f(X)]}{a}$.

只需要对随机变量 $Y = f(X)$ 证明就好. 这样的好处是如果我们的 $f(X)$ 刻画了 $X$ 的某些有用特征, 这就对后续的分析很有利. 我们在后面会见到.

此外, 我们还希望知道有多少概率偏离了期望的某一个范围. 即如果期望是 $\mu$, $\Pr(|X-\mu| \geq a)$ 的概率是多少? 使用 $Y = |X-\mu|$ 然后使用 Markov 不等式, 会得到 $\Pr(|X-\mu| \geq a) \leq \frac{\mathbb{E}[|X-\mu|]}{a}$. 但是 $\mathbb{E}[|X-\mu|]$ 难以计算.

但是可以对两边取平方, 即定义 $Y = (X-\mu)^2$. 得到

$$\Pr(|X-\mu| \geq a) = \Pr\left((X-\mu)^2 \geq a^2\right) \leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{a^2}.$$

其中 $\mathbb{E}\left[(X-\mu)^2\right]$ 就被称为方差, 或者二阶中心矩. 记作 $\mathrm{Var}[X]$. 后续会继续提到它.

**1.2. Chebyshev 不等式.** 如果我们对方差作为一个随机变量使用 Markov 不等式, 就得到了

**定理 1.3.** 对于随机变量 $X$, 对任意的 $a > 0$, 有

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

这就告诉我们方差和偏离期望的关系, 即对于 $k > 1$, $\Pr(|X - \mathbb{E}[X]| \geq k\sqrt{\mathrm{Var}[X]}) \leq \frac{1}{k^2}$.

此外, 我们还发现一个有趣的关系: 随机变量 $X$ 的均值 $\mu$ 恰好是使得方差最小的 $\mu$. 也就是

**性质 1.4.** 随机变量 $X$ 的均值 $\mu := \mathbb{E}[X]$ 是使得 $\mathbb{E}[(x-\mu)^2]$ 最小的 $\mu$.

证明. 考虑函数 $f(x) = \mathbb{E}\left[(X-x)^2\right] = \mathbb{E}\left[X^2\right] - 2x\mathbb{E}[X] + x^2$, 由于 $f(x)$ 是凸函数, $f'(\mu) = 0$ 就一定是极小值进而是最小值. □

**1.3. 中位数和期望的关系.** 在中学我们学习了中位数的定义.

**定义 1.1** (中位数)**.** 随机变量 $X$ 的**中位数 (median)** 是对于任何变量 $m$, 满足

$$\Pr(X \leq m) \geq 1/2 \text{并且} \Pr(X \geq m) \geq 1/2.$$

类似地指出, 中位数也有类似的计算方法.

**性质 1.5.** 随机变量 $X$ 的中位数 $m$ 是使 $\mathbb{E}[|X-m|]$ 最小的 $m$ 值.

证明. 根据对称性, 假设一个值 $y > m$, 但是 $\Pr(X \geq y) < 1/2$. 那么有

$$\begin{aligned}\mathbb{E}[|X-y| - |X-m|] &= (m-y)\Pr(X \geq y) + \sum_{m<x<y} (m+y-2x)\Pr(X=x) + (y-m)\Pr(X \leq m) \\ &> (m-y)/2 + (y-m)/2 = 0\end{aligned}$$

□

最后, 我们揭示均值, 中位数以及方差之间的关系

**定理 1.6.** 如果 $X$ 是均值为 $\mu$, 方差为 $\sigma^2$, 中位数为 $m$. 有 $|\mu - m| \leq \sigma$.

证明.

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

□

## §2 方差及其计算

**a) 基本定义** 在上一节中简要介绍了方差. 现在给出正式的定义

**定义 2.1** ($k$ 阶 (中心) 矩)**.** 对于整数 $k > 0$, 随机变量 $X$ 的 $k$ **阶矩 ($k$th moment)** 是 $\mathbb{E}\left[X^k\right]$ 的值; $X$ 的 $k$ **阶中心矩 ($k$th central moment)** 是 $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right]$ 的值.

有时候, 如果 $\mathbb{E}[X] = 0$, 会称这个变量已经**中心化 (centralized)** 了. 实际上, 把均值减掉构造的新变量 $Y = X - \mathbb{E}[X]$ 就是中心化的. 特殊地, 随机变量 $X$ 的**方差 (variance)** 是二阶中心矩. 即 $\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$, 如果把它开根号就得到了**标准差 (standard derivation)**$\sigma = \sigma[X] := \sqrt{\mathrm{Var}[X]}$

方差也有简便算法. 由于计算通常很繁琐, 我们考虑对原始公式做简化.

**定理 2.1** (方差的简便算法)**.** 对于随机变量 $X$, 如方差存在, 那么

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

证明.

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

□

另外, 如果 $X$almost surely 是一个常数 (即 $\Pr(X = \mathbb{E}[X]) = 1$), 那么等价于 $\mathbb{E}\left[X^2\right] = \mathbb{E}[X]^2$, 等价于 $\mathrm{Var}[X] = 0$. 这可以根据 Chebyshev 不等式得到.

**2.1. 线性函数的方差.** 对于线性函数的方差, 概括可以总结为

**定理 2.2** (方差的性质)**.** 对于随机变量 $X, Y$, 实数 $a \in \mathbb{R}$, 有

- $\mathrm{Var}[a] = 0$;
- $\mathrm{Var}[X + a] = \mathrm{Var}[X]$(方差是中心矩)
- $\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X]$ (方差是二次的)

- $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$.

这些内容都可以通过上述的简便计算法进行验证. 值得注意的是最后一条, 最后多出了一项 2 倍的 $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. 实际上, 这表示与期望不同, 它还要考虑两个变量之间的某种 "相关性".

**定义 2.2** (协方差)**.** 两个随机变量 $X, Y$ 的**协方差 (covariance)** 定义为

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

实际上协方差像极了对称的双线性函数, 它具有如下的性质.

**定理 2.3** (协方差的性质)**.** 协方差具有如下的性质.

- $\mathrm{Var}[X] = \mathrm{Cov}(X,X)$
- 对称性: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- 双线性: $\mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y)$ 以及 $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.

同时可以发现如果 $X, Y$ 是独立的, 那么 $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$.

$\mathrm{Cov}(X,Y) = 0$ 一定说明随机变量 $X, Y$ 互相独立吗?

推广到多个变量, 就有

**性质 2.4.** 如果 $X_1, X_2, \ldots, X_n$ 是相互独立的随机变量, 那么

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \mathbb{E}\left[\prod_{i=1}^{n-1} X_i\right] \cdot \mathbb{E}\left[X_n\right] = \prod_{i=1}^{n} \mathbb{E}\left[X_i\right].$$

证明. 根据某一函数的期望的规则, 有

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xy \Pr(X = x \cap Y = y) = \sum_{x,y} xy \Pr(X = x)\Pr(Y = y) \\ &= \left(\sum_x x \Pr(X = x)\right)\left(\sum_y y \Pr(Y = y)\right) = \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

□

现在探讨了几个随机变量的乘积的结果. 对于相互独立的情形, 就没有那么复杂. 对于不相互独立的随机变量, 其大致的界限是多少?

**定理 2.5.** 对于随机变量 $X, Y$,

$$\mathbb{E}[XY]^2 \le \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right].$$

更一般地, 对于 $p, q$ 满足 $\frac{1}{p} + \frac{1}{q} = 1$, 就有

$$\mathbb{E}[XY] \le \mathbb{E}\left[|X|^p\right]^{1/p}\mathbb{E}\left[|Y|^q\right]^{1/q}$$

这里可能会联想到这与向量的范数有联系. 这就可以考虑将概率当做向量空间中的向量来考虑, 期望只是类似于某种度量方式.

正是由于上述的界限, 我们才可以定义一个取值在 $[-1,1]$ 的相关系数如下.

**定义 2.3** (相关系数)**.** 随机变量 $X,Y$ 的相关系数为

$$\rho(X,Y)=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}[X]\cdot\operatorname{Var}[Y]}}.$$

两个变量不相关意味着

- $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$
- $\operatorname{Var}[X+Y]=\operatorname{Var}[X]+\operatorname{Var}[Y]$

但不意味着两者独立. 实际上, 独立性是一个强得多的条件.

最后, 我们借助协方差把随机变量的和形式写得好看些. 即可以变为 $\operatorname{Var}[X+Y]=\operatorname{Var}[X]+\operatorname{Var}[Y]+2\operatorname{Cov}(X,Y)$.

对于 $n$ 个随机变量 $X_1,X_2,\ldots,X_n$, 就会变成

$$\operatorname{Var}\left[\sum_{i=1}^{n}X_i\right]=\sum_{i=1}^{n}\operatorname{Var}\left[X_i\right]+\sum_{i\neq j}\operatorname{Cov}\left(X_i,X_j\right)$$

但是对于两两独立的随机变量, 后面的协方差就可以省去了.

**2.2. 常见分布的方差.** 像上一节一样, 同样看一些常见方差的计算.

**a) Bernoulli 随机变量** 回忆 Bernoulli 随机变量满足 $p_X(k)=\Pr(X=k)=\begin{cases}p & \text{if } k=1\\ 1-p & \text{if } k=0\end{cases}$, 且期望为 $p$.

由于 $X^2=X$, 表明 $\mathbb{E}\left[X^2\right]=\mathbb{E}[X]=p$. 所以 $\operatorname{Var}[X]=\mathbb{E}\left[X^2\right]-\mathbb{E}[X]^2=p-p^2=p(1-p)$.

这就使得我们很方便地计算指示器变量的方差. 假设有事件 $A$ 指示变量 $X=I(A)$, $\operatorname{Var}[X]=\Pr(A)(1-\Pr(A))=\Pr(A)\Pr\left(A^c\right)$. 通常这是很方便的.

**b) 离散的均匀分布** 对于 $a\leq b$ 的整数, 令 $X$ 均匀地从 $\{a,a+1,...,b\}$ 中抽取一个, 求随机变量 $X$ 的方差.

首先知道 $\mathbb{E}[X]=\sum_{k=a}^{b}\frac{k}{b-a+1}=\frac{a+b}{2}$, 以及 $\mathbb{E}\left[X^2\right]=\sum_{k=a}^{b}\frac{k^2}{b-a+1}=\frac{2b^2+2ab+2a^2+b-a}{6}$, 根据计算公式得到 $\operatorname{Var}[X]=\mathbb{E}\left[X^2\right]-\mathbb{E}[X]^2=\frac{(b-a)(b-a+2)}{12}$.

**c) Possion 分布** 回忆 Possion 分布满足 $p_X(k)=\Pr(X=k)=\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!},\quad k=0,1,2,\ldots$. 其期望为 $\lambda$.

由于

$$\begin{aligned}\mathbb{E}\left[X^2\right]&=\sum_{k\geq 0}k^2\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}=\sum_{k\geq 1}k\frac{\mathrm{e}^{-\lambda}\lambda^k}{(k-1)!}\\&=\sum_{k\geq 0}(k+1)\frac{\mathrm{e}^{-\lambda}\lambda^{k+1}}{k!}=\lambda\sum_{k\geq 0}(k+1)\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\\&=\lambda\mathbb{E}[X+1]=\lambda(\mathbb{E}[X]+1)=\lambda(\lambda+1)\end{aligned}$$

我们知道 $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda(\lambda+1) - \lambda^2 = \lambda$. Possion 分布的均值和方差一样, 都是 $\lambda$.

**d) 几何分布** 回忆几何分布满足 $p_X(k) = \Pr(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots$, 期望是 $1/p$, 还没有记忆性.

第一种计算方法是直接运算. 由于 $\mathbb{E}[X^2] = \sum_{k\geq 1} k^2(1-p)^{k-1}p = (2-p)p^{-2}$, 自然 $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (2-p)p^{-2} - p^{-2} = (1-p)/p^2$.

第二种方法是使用全期望公式, 以及几何分布无记忆性的特性把条件去掉:

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \mathbb{E}\left[X^2 \mid X > 1\right] \cdot (1-p) + \mathbb{E}\left[X^2 \mid X = 1\right] \cdot p \\ &= \mathbb{E}\left[((X-1)+1)^2 \mid X > 1\right] \cdot (1-p) + p \\ &\xlongequal{\text{无记忆性}} \mathbb{E}\left[(X+1)^2\right] \cdot (1-p) + p \\ &= (1-p)\mathbb{E}\left[X^2\right] + 2(1-p)/p + 1\end{aligned}$$

**e) 二项分布** 回忆二项分布满足 $p_X(k) = \Pr(X = k) = \binom{n}{k}p^k(1-p)^k, \quad k = 0, 1, \ldots, n$, 并且期望为 $np$.

根据定义, 我们要计算

$$\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k(1-p)^{n-k} - (np)^2.$$

但是这个式子相当难算. 实际上, 我们可以把它拆为若干个指示函数. 即看做若干个独立同分布 (iid.) 的 Bernoulli 随机变量的和.

自然, 根据方差的性质, 就有

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = np(1-p).$$

**f) 负二项分布** 回忆参数为 $n, p$ 的负二项分布满足 $p_X(k) = \Pr(X = k) = \binom{k+r-1}{k}(1-p)^k p^r = (-1)^k\binom{-r}{k}(1-p)^k p^r, k = 0, 1, 2, \ldots$, 表示成功概率为 $p$ 的 iid. Bernouli 实验中, 从开始到成功 $r$ 次这一段时间中失败的次数. 自然不会希望按照定义计算.

但是我们发现 $X$ 可以被表示为若干个服从几何分布的随机变量之和. 即 $X_1, X_2, \ldots, X_r$ 是 iid. 服从参数为 $p$ 的几何分布的随机变量, 那么 $X = (X_1 - 1) + \cdots + (X_r - 1)$. 由于它们之间相互独立, 就有

$$\mathrm{Var}[X] = \sum_{i=1}^{r} \mathrm{Var}\left[X_i - 1\right] = \sum_{i=1}^{r} \mathrm{Var}\left[X_i\right] = \frac{r(1-p)}{p^2}.$$

**2.3. 再论 Chebyshev 不等式.** Chebyshev 不等式有很多实际的应用.

**a) 随机算法的去随机化** 下面来看数据结构课上学过的 Hashing 算法. 以字符串 Hashing 为例, 它可以将非常大的整数通过一个函数 $f : S \to [0..n]$ 映射到

不超过 $n$ 的整数. 通常, 这样的函数是由一个质数 $p > 1$, 以及与 $p$ 互质的集合 $[p] := \{0, 1, \ldots, p-1\} =: \mathbb{Z}_p$ 组成.

假定现在有 $i$ 个非常大的整数等待 hashing. 我们从 $[p]$ 中均匀随机选取两个随机变量 $\boldsymbol{a}, \boldsymbol{b}$, 并且让值为 $i$ 的数字 hash 过后的值变为 $r_i := (\boldsymbol{a} \cdot i + \boldsymbol{b}) \bmod p$ , $\forall i = 1, 2, \ldots, p$. 可以发现, $r_1, r_2, ..., r_p \in [p]$ 相互独立. 并且, 每一个 $r_i$ 的值都是均匀分布在 $[p]$ 中的.

证明. 要证明每一个 $r_i$ 等可能地被映成了 $[p]$, 只要看对于任意的 $i \neq j, \forall c, d \in [p]$, 因为

$$\begin{cases} \boldsymbol{a} \cdot i + \boldsymbol{b} \equiv c (\mathrm{mod} p) \\ \boldsymbol{a} \cdot j + \boldsymbol{b} \equiv d (\mathrm{mod} p) \end{cases}$$

都有唯一的 $(a, b) \in [p]^2$ 作为解. 所以说 $\Pr(r_i = c \cap r_j = d) = 1/p^2$. 因此,

$$\Pr(r_i = c) = \Pr(\boldsymbol{a} \cdot i + \boldsymbol{b} \equiv c (\mathrm{mod} p)) = \frac{1}{p} \sum_{a \in [p]} \Pr(\boldsymbol{b} \equiv c - \boldsymbol{a} i (\mathrm{mod} p)) = \frac{1}{p}$$

□

我们讨论了这个映射函数本身的性质. 我们知道 hash 碰撞不是我们希望的, 于是希望减少这样发生碰撞的概率. 其中一个方法是使用两个模数的 hash 算法. 即这个数经过两个 hash 之后的值均是给定的, 我们才认为找到了它.

这就是随机化算法减小错误的一个好办法. 也就是, 我们有一个至少以概率 $1 - \epsilon$ 正确的二分类算法 $\mathcal{A}$, 输入两个参数 $x, r$ 表示输入的要 hash 的数和随机数种子, 其输出结果为 $\{0, 1\}$. 既然一次不可靠, 多跑几次看上去就靠谱一点了.

形式化地说, 我们希望找到一个从多次运行的结果的函数到一个一次最终结果的映射 $f : \{0,1\} \times \{0,1\} \times ... \times \{0,1\} \to \{0,1\}$. 我们规定

- 有错误的时候: $f(x) = 1 \Longrightarrow \Pr(\mathscr{A}(x, r) = 1) \geq \epsilon$
- 没有错误的时候: $f(x) = 0 \Longrightarrow \mathscr{A}(x, r) = 0, \forall r \in [p]$

也就是只要有一个 “1” 我们就认为答案应该返回 1(发生碰撞). 对于更多的情况, 如果跑了很多次都没有出现碰撞的情形, 我们就认为没有发生碰撞.

把这问题形式化一下, 我们的算法就是输入 $x$, 以及运行 $k$ 次. 写做 $\mathscr{A}^k(x, r_1, \ldots, r_k) := \vee_{i=1}^k \mathscr{A}(x, r_i)$. 其中, $k \leq p$, 并且 $\boldsymbol{a}, \boldsymbol{b}$ 是均匀从 $[p]$ 中选取的值, $r_i = (\boldsymbol{a} \cdot i + \boldsymbol{b}) \bmod p$.

这样的算法, 如果 $f(x) = 1$ 就意味着存在一个 $i$ 满足 $\Pr(\mathscr{A}(x, r_i) = 1) \geq \epsilon$. 那么干脆我们把每一次返回值是不是等于 0 作为一个随机变量. 即定义 $X_i = \mathscr{A}(x, r_i)$, 那么最后的随机变量 $X = \sum_{i=1}^k X_i$.

我们现在知道, $X_1, \ldots, X_k$ 是 Bernouli 随机变量, 满足 $\Pr(X_i = 1) \geq \epsilon$. 那么

$$\begin{aligned}\Pr\left(\mathscr{A}^k(x,r_1,\ldots,r_k)=0\right) &= \Pr(X=0)\\ &\le \Pr(|X-\mathbb{E}[X]|\ge \mathbb{E}[X])\\ &\le \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2}\end{aligned}$$

$$\left(\text{期望的线性性 } \mathbb{E}[X]=\sum_{i=1}^{k}\mathbb{E}[X_i]\ge \epsilon k\right.$$

$$\left.\text{两两独立 } \mathrm{Var}[X]=\sum_{i=1}^{k}\mathrm{Var}[X_i]\le\sum_{i=1}^{k}\mathbb{E}\left[X_i^2\right]=\sum_{i=1}^{k}\mathbb{E}[X_i]=\mathbb{E}[X]\right)$$

$$\le \frac{1}{\epsilon k}$$

如果跑两次, 就把正确率从 $1-\epsilon$ 变为了 $1/(\epsilon k)(k\le p)$.

**b) Weierstrass 逼近定理** Weierstrass 认为, 任何一个连续函数, 都可以使用多项式逼近. 这是数学分析课程上面的一个定理.

**定理 2.6** (Weierstrass 逼近定理)**.** 如果 $f:[0,1]\to[0,1]$ 是一个连续函数, 对于任意的 $\epsilon>0$, 都存在一多项式 $p$ 满足

$$\sup_{x\in[0,1]}|p(x)-f(x)|\le\epsilon.$$

证明. 我们选取充分大的 $n$(后面会说明到底要有多大) 并声称一定可以找到这样一个值. 对于 $x\in[0,1]$, 令 $X\sim\frac{1}{n}\mathrm{Bin}(n,x)$ 定义在 $x\in[0,1]$ 上的多项式 $p$ 为

$$p(x)=\mathbb{E}[f(X)]=\sum_{k=0}^{n}f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}$$

那么 $|p(x)-f(x)|=|\mathbb{E}[f(X)-f(x)]|\le\mathbb{E}[|f(X)-f(x)|]$. 回忆函数连续的性质. 如果 $f$ 在 $[0,1]$ 上连续, 那么 $\forall|x-y|<\delta,\exists\delta>0$ s.t. $|f(x)-f(y)|\le\epsilon/2$

那么对 $\mathbb{E}[|f(X)-f(x)|]$ 可以取条件: $|X-x|\le\delta$ 以及 $|X-x|>\delta$, 即

$$\begin{aligned}\mathbb{E}[f(X)] =&\mathbb{E}[|f(X)-f(x)||X-x\,|\le\delta]\cdot\Pr(|X-x|\le\delta)\\ &+\mathbb{E}[|f(X)-f(x)|\,|X-x|>\delta]\cdot\Pr(|X-x|>\delta)\\ \le&\mathbb{E}[\epsilon/2]+|1-0|\cdot\Pr(|X-x|>\delta)\\ \le&\frac{\epsilon}{2}+\frac{x(1-x)}{n\delta^2} && \text{(Chebyshev 不等式)}\\ \le&\frac{\epsilon}{2}+\frac{1}{4n\delta^2}\end{aligned}$$

如果要 $\frac{\epsilon}{2}+\frac{1}{4n\delta^2}\le\epsilon$, 只要选取 $n\ge\frac{1}{2\epsilon\delta^2}$ 即可.

□

## §3 高阶矩

对于更高阶的矩, 同样有对应的名字. 既然矩可以把随机变量用一个实数衡量, 那么如果我们知道很多关于这个随机变量的矩的信息, 是不是可以全面地刻画随机变量呢?

具体地, 如果我们知道 $m_k := \mathbb{E}\left[X^k\right], \forall k \geq 1$, 可否唯一确定 $X$ 的分布? 假设 $X$ 的值域是离散的 $\{x_1, \ldots, x_n\}$, 把概率分布看做向量, 就让我们可以解线性方程组

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix}$$

由于 Vandermonde 矩阵可逆, 自然可以恢复出这变量的 pmf: $p_i = p_X(x_i)$.

更进一步, 我们可以使用多项式簇来 "编码" 每一个矩的值. 也就是把这一列信息包装为一个新函数

$$M_X(t) = \sum_{k \geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!} = \mathbb{E}\left[\mathrm{e}^{tX}\right]$$

我们可以这样做是因为

- 函数簇 $\{t, t^2/2!, t^3/3!, ...\}$ 线性无关. 也就是这簇函数任意两个的线性组合都不能构成这簇函数中的其他函数. 这就可以让我们唯一地编码每一个位置.
- 回忆对于实数的 Taylor 展开. Taylor 展开式告诉我们

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

. 现在, 如果把实数 $x$ 换为随机变量 $X$, 就有

$$e^{sX} = \sum_{k=0}^{\infty} \frac{(sX)^k}{k!} = \sum_{k=0}^{\infty} \frac{X^k s^k}{k!}$$

. 对它取期望, 就得到了最右边的式子. 这解释了 PMF 和 MGF 之间的关系.

这里没有很严格地论述为什么可以把 $x$ 换为 $X$ 的关系, 但是也足够用了.

那么, 我们定义矩的生成函数.

**定义 3.1** (矩的生成函数)**.** 对于随机变量 $X$ 以及参数 $t$, 其**矩的生成函数 (moment generating function)** 是

$$M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] = \sum_{k \geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!}$$

还要不加证明地指出, 如果 $X, Y$ 有相同的矩的生成函数, 那么 $X, Y$ 是同分布的.

实际上, MGF 是一个刻画随机变量的方法, 但是有不足之处. 因为倘若 $\mathbb{E}\left[X^k\right]$ 增长得特别快, 那么 $M_X(t)$ 就不收敛了. 我们在下一节会介绍一个方法.

选取 $e^x$ 这一组基底还有一个重要的原因. 因为见到了这样的多项式, 就可以很方便地得到它的各矩.

**定理 3.1.** 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

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

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

**证明.** 由于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 $M_X^{(n)}(0) = \mathbb{E}\left[X^n\right]$. □

回到离散的变量. 假设我们已经知道了一个离散随机变量的 PMF, 我们如何计算它的 MGF? 关键就是使用 $M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right]$.

PMF 是 Probability Mass Function, 概率质量函数 (告诉了你有多大概率取哪一个值); MGF 是 Moment Generating Function, 上文的矩的生成函数

**例子 3.1.** 设 $X \sim \mathrm{Pos}(\lambda)(P_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, \ldots)$, 下面求 $X$ 的 MGF.

$$\begin{aligned}
M_X(s) &= \mathbb{E}\left[e^{sX}\right] \\
&= \sum_{k=0}^{\infty} e^{sk} e^{-\lambda} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(\lambda e^s\right)^k}{k!} \\
&= e^{-\lambda} e^{\lambda e^s} \quad (\ e^x \text{ 的 Taylor 展开}) \\
&= e^{\lambda(e^s - 1)}, \quad \forall s \in \mathbb{R}.
\end{aligned}$$

**例子 3.2.** 这个例子考察独立的随机变量的和的 MGF. 假设 $X_1, X_2, \ldots, X_n$ 是 $n$ 个独立的随机变量, 随机变量 $Y := X_1 + X_2 + \cdots + X_n$. 假设 $X_1, X_2, ..., X_n$ 的 MGF 为 $M_{X_1}, M_{X_2}, ..., M_{X_n}$, 那么

$$\begin{aligned}
M_Y(s) &= \mathbb{E}\left[e^{sY}\right] \\
&= \mathbb{E}\left[e^{s(X_1 + X_2 + \cdots + X_n)}\right] \\
&= \mathbb{E}\left[e^{sX_1} e^{sX_2} \cdots e^{sX_n}\right] \\
&= \mathbb{E}\left[e^{sX_1}\right] \mathbb{E}\left[e^{sX_2}\right] \cdots \mathbb{E}\left[e^{sX_n}\right] \quad (\text{因为} X_i \text{ 互相独立}) \\
&= M_{X_1}(s) M_{X_2}(s) \cdots M_{X_n}(s).
\end{aligned}$$

这表明独立的随机变量的和反映到 MGF 上是它们的乘积.

**例子 3.3.** 如果 $X \sim \mathrm{Binom}(n, p)$, 求 $X$ 的 MGF. 像往常一样, 将 $X$ 拆成若干次独立的 Bernoili 随机变量的和 $X_i$. 也就是 $X = X_1 + X_2 + \cdots + X_n$, 每个

$X_i \sim \text{Bernoulli}(p)$, 因此

$$\begin{aligned} M_X(s) &= M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s) \\ &= (M_{X_1}(s))^n \quad \text{(因为每个}X_i\text{ 's 都是独立同分布的)} \end{aligned}$$

由于 $M_{X_1}(s) = \mathbb{E}\left[e^{sX_1}\right] = pe^s + 1 - p$, 因此 $M_X(s) = (pe^s + 1 - p)^n$.