



概率论笔记

作者：肖程哲

时间：July 15, 2022



苟日新，日日新，又日新

目录

第 1 章 概率基础	1
1.1 概率空间	1
1.2 古典概型与几何概型	3
1.2.1 古典概型	3
1.2.2 几何概型	3
1.3 条件概率	3
第 2 章 随机变量	5
2.1 随机变量的分布	5
2.2 多元随机变量	7
2.2.1 边际分布	8
2.2.2 独立	8
2.2.3 条件分布	8
2.3 随机变量的函数	9
2.3.1 分布函数法	10
2.3.2 Copula	11
2.3.3 概率密度函数法	12
2.3.4 矩母函数法	13
2.3.5 次序统计量	13
第 3 章 随机变量的数值特征	15
3.1 期望	15
3.1.1 均值	15
3.1.2 方差	16
3.1.3 协方差	17
3.2 矩母函数与特征函数	18
3.2.1 矩	18
3.2.2 矩母函数	18
3.2.3 联合特征函数	19
3.2.4 特征函数	19
3.3 条件期望	20
3.4 熵与信息	20
第 4 章 常见分布	17
4.1 离散分布	17
4.2 连续分布	17
4.3 正态分布及其导出分布	17
第 5 章 概率极限	18
5.1 收敛	18
5.2 大数定理	18
5.3 中央极限定理	18

第 A 章 基本数学工具	19
A.1 排列与组合	19

第3章 随机变量的数值特征

3.1 期望

定义 3.1

对于实值随机向量 $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ 和 (可测) 函数 $g : \mathbb{R}^n \rightarrow \mathbb{R}$, 称

$$\mathbb{E}[g(X)] = \int_{\Omega} g(X(\omega)) d\mathbb{P}(\omega) = \int_{\mathbb{R}} g(x) dF_X(x)$$

为 $g(X)$ 的期望 (expectation).



注 当 $F_X(x)$ 在 x_0 处连续可导时, $dF_X(x_0) = f_X(x_0)dx$; 当 x_0 为区间断点时时, $dF_X(x_0) = p_X(x_0)\delta(x_0)dx$. 期望算子 \mathbb{E} 是一个线性泛函, 仅适用于可积的随机变量.

定义 3.2

- 当 $g(x) = x$ 时, $\mathbb{E}[g(x)] = \mathbb{E}[X]$ 称作 X 的均值 (mean), 记为 μ_X
- 当 $g(x) = (x - \mu_X)^2$ 时, $\mathbb{E}[g(x)] = \mathbb{E}[(X - \mathbb{E}[X])^2]$ 称作 X 的方差 (variance), 记为 σ_X^2 . 其平方根称作 X 的标准差 (standard deviation), 记为 σ_X
- 当 $g(x, y) = (x - \mu_X)(y - \mu_Y)$ 时, $\mathbb{E}[g(X, Y)] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ 称作 X 与 Y 的协方差 (covariance), 记为 $\text{Cov}(X, Y)$ 或 σ_{XY} .
- 定义 X 与 Y 的相关系数 (correlation coefficient) 为: $\sigma_{XY}/(\sigma_X \sigma_Y)$, 记为 $\text{Cor}(X, Y)$ 或 ρ_{XY} . 若 $\rho_{XY} = 0$, 则称 X 与 Y 不相关



3.1.1 均值

注

- 随机变量的均值可看作其加权平均, 权重为其 pdf 或 pmf, 也即其质心. 从大数定律 (5) 的角度看, 也可解释为其长期均值.
- 方差为随机变量距其均值的均方偏差, 刻画了 X 的变动程度
- 随机变量的均值与标准差的单位和其本身相同, 方差的为其平方

定理 3.1

均值为随机变量的线性映射, 即:

$$\mathbb{E}(a + \sum_{i=1}^n b_i X_i) = a + \sum_{i=1}^n b_i \mathbb{E}(X_i)$$



定理 3.2 (0)

若 X, Y 独立, 则

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

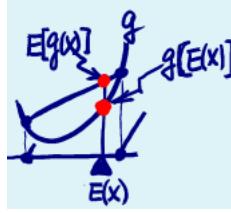


注 由于 $\mathbb{E}(X/Y) = \mathbb{E}(X)\mathbb{E}(\frac{1}{Y})$, 而 $\mathbb{E}(\frac{1}{Y}) \neq \frac{1}{\mathbb{E}(Y)}$, 所以 $\mathbb{E}(X/Y) \neq \mathbb{E}(X)/\mathbb{E}(Y)$

定理 3.3

若 g 为下凸 (convex) 函数, 则 $\mathbb{E}[g(X)] \geq g[\mathbb{E}(X)]$; 若 g 为上凸 (concave) 函数, 则 $\mathbb{E}[g(X)] \leq g[\mathbb{E}(X)]$;





一个重要结果是, 若 $g(X) \geq 0$, 则 $\mathbb{E}[g(X)] = 0 \implies g(X) \stackrel{\text{a.s.}}{=} 0$, 即 $\mathbb{P}\{g(X) = 0\} = 1$. 其证明可通过 **Markov 不等式**

$$\mathbb{P}\{g(X) \geq \varepsilon\} \leq \mathbb{E}[g(X)]/\varepsilon, \quad \forall \varepsilon > 0$$

完成, 其中需要用到概率的连续性, 即 $\lim_{n \rightarrow \infty} A_n = A \implies \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

预处理随机变量有两个常用变换:

- 中心化 (centralization) $X \mapsto X - \mathbb{E}X$;
- 标准化 (standardization) $X \mapsto \frac{X - \mathbb{E}X}{\sqrt{\text{Var}(X)}}$.

3.1.2 方差

定理 3.4

$$\sigma_X^2 = \text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = E(X^2) - \mu_X$$



定理 3.5

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



定理 3.6

$$\text{Var}(a + \sum_{i=1}^n b_i X_i) = \sum_{i=1}^n b_i^2 \text{Var}(X_i) + \mathbf{b}^T \Sigma \mathbf{b}$$

其中 Σ 为协方差矩阵, $\Sigma_{i,j} = \text{Cov}(X_i, X_j)$



定理 3.7

若 X_1, \dots, X_n 相互独立, 则:

$$\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$$



考虑均方误差 (mean squared error)

$$\text{MSE}(X; \theta) = \mathbb{E}[|X - \theta|^2], \quad \theta \in \mathbb{R},$$

通过方差偏差分解 (variance-bias decomposition)

$$\text{MSE}(X; \theta) = \text{Var}(X) + |\mathbb{E}X - \theta|^2$$

可以说明 $\theta \mapsto \text{MSE}(X; \theta)$ 在 $\mathbb{E}X$ 处取到最小值 $\text{Var}(X)$.

投影 (projection) 和正交分解的思想在各种内积空间中应用广泛, 这里是 $\mathbb{E} = \text{proj}_{\mathbb{R}}$, 概率论中关于子事件域 \mathcal{G} (随机元 X , resp.) 的条件期望几何直观是 $\text{proj}_{\mathcal{G}}(\text{proj}_{\sigma(X)}, \text{resp.})$, 线性模型 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ 中拟合值为 $\hat{\mathbf{y}} = \text{proj}_{\text{Col}(\mathbf{X})} \mathbf{y}$.

定理 3.8 (Chebyshev 不等式)

设随机变量 X 的均值与方差分别为: μ, σ^2 , 则:

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$$



证明 设 $f(x)$ 为 X 的概率密度函数, 令 $R = \{x : |x - \mu| > t\}$

$$\begin{aligned}\mathbb{P}(|x - \mu| > t) &= \int_R 1 \cdot f(x) dx \leq \int_R \frac{(x - \mu)^2}{t^2} f(x) dx \\ &\leq \int_{-\infty}^{\infty} \frac{(x - \mu)^2}{t^2} f(x) dx \\ &= \frac{\sigma^2}{t^2}\end{aligned}$$

注 若令 $t = k\sigma$, 则 $\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}$, 即标准差可代表随机变量偏离均值的概率单位距离.

推论 3.1

$$\text{Var}(X) = 0 \implies P(X = \mu) = 1$$



3.1.3 协方差

协方差代表了 X 与 Y 之间的联合变化倾向, 或者说他们间的相关程度, 但其间未必有因果关系.

定理 3.9

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mu_X \mu_Y$$

**定理 3.10**

$$\text{Cov}\left(a + \sum_{i=1}^n b_i X_i, c + \sum_{j=1}^m d_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m b_i d_j \text{Cov}(X_i, Y_j) = \mathbf{b}^T \Sigma \mathbf{d}$$

**定理 3.11**

独立是不相关的充分条件, 但不是必要条件

**定理 3.12**

$-1 \leq \rho_{XY} \leq 1$, 当且仅当 X 与 Y 间为线性关系时取等号



证明

定理 3.13

平移与缩放随机变量都不影响其协方差, 即:

$$|\text{Cov}(a + bX, c + dY)| = |\text{Cor}(X, Y)|$$



3.2 矩母函数与特征函数

3.2.1 矩

定义 3.3

对于随机变量 X , 定义其 k 阶矩 (moment) 为 $E(X^k)$, 记为 μ_k ; 定义其 k 阶中心矩 (central moment) 为 $E((X - \mu_X)^k)$, 记为 v_k ;



易知矩与中心矩间存在以下关系:

$$\begin{aligned} v_k &= \sum_{i=0}^k \binom{k}{i} \mu_i (-\mu_X)^{n-i} \\ \mu_k &= \sum_{i=0}^k \binom{k}{i} v_i (\mu_X)^{n-i} \end{aligned}$$

特别的有:

$$\begin{aligned} E(X) &= \mu_1 \\ \text{Var}(X) &= v_2 = \mu_2^2 - 2\mu_1 \cdot \mu_1 + \mu_2 = \mu_2 - \mu_1^2 \end{aligned}$$

3.2.2 矩母函数

定义 3.4

对于随机变量 X , 若下式期望存在:

$$M_X(t) = \mathbb{E}(e^{tX})$$

则称其为矩母函数 (moment generating function, mgf).



注 此表达式等价于对概率质量函数或密度函数作 Laplace 变换, 当 t 取某些特定值时, 可能不存在. (若 $t = 0$ 则永远存在)

定理 3.14

若当 t 属于一个包含零点的开区间时, 矩母函数一直存在, 则其唯一对应一个概率分布.



定理 3.15

若当 t 属于一个包含零点的开区间时, 矩母函数一直存在, 则:

$$M_X^{(k)}(0) = E(X^k)$$



注 由此可方便地计算各阶矩, 故称其为矩母函数. 反过来, 若已知各阶矩, 通过 Tayler 展开式 $M_X(t) = \sum_{k=0}^{\infty} \frac{M_X^{(k)}(0)}{k!} t^k$ 还原矩母函数, 进而得出概率分布.

命题 3.1

若 a, b 为常数, 则

$$M_{a+bX}(t) = e^{at} M_X(bt)$$



定理 3.16

若 X, Y 独立, 则

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

泛化情况: 若 X_1, \dots, X_n 相互独立, 则

$$M_{X_1+\dots+X_n} = \prod_{i=1}^n M_{X_i}(t)$$



3.2.3 联合特征函数

定义 3.5

对于随机变量 X_1, \dots, X_n , 若下式期望存在:

$$M_{X_1\dots X_n}(t_1, \dots, t_n) = \mathbb{E}(e^{t_1 X_1 + \dots + t_n X_n})$$

则称其为**联合矩母函数** (joint moment generating function, joint mgf)。



注 此处为多元 函数

命题 3.2

$$M_{X_i}(t_i) = M_{X_1\dots X_n}(0, \dots, t_i, \dots, 0)$$



定理 3.17

当且仅当:

$$M_{X_1\dots X_n}(t_1, \dots, t_n) = \prod_{i=1}^n M_{X_i}(t_i)$$

时, X_1, \dots, X_n 相互独立



注 与累计函数、密度函数、质量函数 的情况类似, 变量相互独立等价于联合函数可拆分为边缘函数的乘积

定理 3.18

$$\frac{\partial^{r_1+\dots+r_n}}{\partial t_1^{r_1}\dots\partial t_n^{r_n}} M_{X_1\dots X_n}(t_1, \dots, t_n) = E(X_1^{r_1}\dots X_n^{r_n})$$



3.2.4 特征函数

由于有时矩母函数可能不存在, 为避免此缺陷, 构造出与之特性类似的特征函数。

定义 3.6

对于随机变量 X , 定义其**特征函数** (characat function, chf) 为:

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX))$$

对于随机变量 X_1, \dots, X_n , 定义其**联合特征函数** (joint characat function, joint chf) 为:

$$\phi_{X_1\dots X_n}(t_1, \dots, t_n) = \mathbb{E}(e^{i(t_1 X_1 + \dots + t_n X_n)})$$



注 此表达式等价于对概率质量函数或密度函数作 Fourier 变换

命题 3.3

随机变量的特征函数总是存在



命题 3.4

若矩母函数存在，则其与特征函数之间满足关系：

$$\phi_X(t) = M_X(it)$$

定理 3.19

特征函数可通过以下逆变换得到分布：

$$\text{离散 } p_X(x) = \lim_{T \rightarrow \infty} \int_{-T}^T e^{-itx} \phi_X(t) dt$$

$$\text{连续 } f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t) dt$$



3.3 条件期望

定义 3.7

条件分布(定义??)的数学期望(若存在)称为**条件期望**,其定义如下:

$$E(Y|X=x) = \begin{cases} \sum_j y_j P(Y=y_j|X=x), & (X,Y) \text{ 为二维离散随机变量;} \\ \int_{-\infty}^{+\infty} y p(y|x) dy, & (X,Y) \text{ 为二维连续随机变量.} \end{cases}$$



3.4 熵与信息