概率论与随机过程

一、概率与分布

1.1 条件概率与独立事件

1. 条件概率：已知 A 事件发生的条件下 B 发生的概率，记作 $P(B \mid A)$，它等于事件 AB 的概率相对于事件 A 的概率，即：$P(B \mid A)=\frac{P(AB)}{P(A)}$ 。其中必须有 $P(A)>0$ 。

2. 条件概率分布的链式法则：对于 n 个随机变量 $X_1, X_2, \cdots, X_n$，有：

$$P(X_1, X_2, \cdots, X_n)=P(X_1) \prod_{i=2}^{n} P(X_i \mid X_1, \cdots, X_{i-1})$$

3. 两个随机变量 X, Y 相互独立的数学描述：$P(X, Y)=P(X) P(Y)$ 。记作：$X \perp Y$ 。
4. 两个随机变量 X, Y 关于随机变量 Z 条件独立的数学描述：$P(X, Y \mid Z)=P(X \mid Z) P(Y \mid Z)$ 。
   记作：$X \perp Y \mid Z$ 。

1.2 联合概率分布

1. 定义X 和 Y 的联合分布为：$P(a, b)=P\{X \leq a, Y \leq b\}, \quad -\infty<a, b<+\infty$ 。
   - X 的分布可以从联合分布中得到：
   $$P_X(a)=P\{X \leq a\}=P\{X \leq a, Y \leq \infty\}=P(a, \infty), \quad -\infty<a<+\infty$$
   - Y 的分布可以从联合分布中得到：
   $$P_Y(b)=P\{Y \leq b\}=P\{X \leq \infty, Y \leq b\}=P(\infty, b), \quad -\infty<b<+\infty$$

2. 当 X 和 Y 都是离散随机变量时， 定义 X 和 Y 的联合概率质量函数为：$p(x, y)=P\{X=x, Y=y\}$
   则 X 和 Y 的概率质量函数分布为：
$$p_X(x)=\sum_y p(x, y) \quad p_Y(y)=\sum_x p(x, y)$$

3. 当 X 和 Y 联合地连续时，即存在函数 $p(x, y)$ ，使得对于所有的实数集合 $\mathbb{A}$ 和 $\mathbb{B}$ 满足：
$$P\{X \in \mathbb{A}, Y \in \mathbb{B}\}=\int_{\mathbb{B}} \int_{\mathbb{A}} p(x, y) dx dy$$
   则函数 $p(x, y)$ 称为 X 和 Y 的概率密度函数。
   - 联合分布为：$P(a, b)=P\{X \leq a, Y \leq b\}=\int_{-\infty}^{a} \int_{-\infty}^{b} p(x, y) dx dy$ 。
   - X 和 Y 的分布函数以及概率密度函数分别为：

$$P_X(a)=\int_{-\infty}^{a}\int_{-\infty}^{\infty}p(x,y)dxdy=\int_{-\infty}^{a}p_X(x)dx$$
$$P_Y(b)=\int_{-\infty}^{\infty}\int_{-\infty}^{b}p(x,y)dxdy=\int_{-\infty}^{b}p_Y(y)dy$$
$$p_X(x)=\int_{-\infty}^{\infty}p(x,y)dy$$
$$p_Y(y)=\int_{-\infty}^{\infty}p(x,y)dx$$

.

二、期望和方差

2.1 期望

1. 期望描述了随机变量的平均情况，衡量了随机变量 X 的均值。它是概率分布的泛函（函数的函数）。
   - 离散型随机变量 X 的期望：$\mathbb{E}[X]=\sum_{i=1}^{\infty}x_ip_i$ 。

     若右侧级数不收敛，则期望不存在。
   - 连续性随机变量 X 的期望：$\mathbb{E}[X]=\int_{-\infty}^{\infty}xp(x)dx$ 。

     若右侧极限不收敛，则期望不存在。
2. 定理：对于随机变量 X ，设 $Y=g(X)$ 也为随机变量，$g(\cdot)$ 是连续函数。
   - 若 X 为离散型随机变量，若 Y 的期望存在，则：$\mathbb{E}[Y]=\mathbb{E}[g(X)]=\sum_{i=1}^{\infty}g(x_i)p_i$ 。

     也记做：$\mathbb{E}_{X\sim P(X)}[g(X)]=\sum_x g(x)p(x)$ 。
   - 若 X 为连续型随机变量，若 Y 的期望存在，则：$\mathbb{E}[Y]=\mathbb{E}[g(X)]=\int_{-\infty}^{\infty}g(x)p(x)dx$ 。

     也记做：$\mathbb{E}_{X\sim P(X)}[g(X)]=\int g(x)p(x)dx$ 。

   该定理的意义在于：当求 $\mathbb{E}(Y)$ 时，不必计算出 Y 的分布，只需要利用 X 的分布即可。

   该定理可以推广至两个或两个以上随机变量的情况。对于随机变量 X,Y，假设 $Z=g(X,Y)$ 也是随机变量，$g(\cdot)$ 为连续函数，则有：$\mathbb{E}[Z]=\mathbb{E}[g(X,Y)]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)p(x,y)dxdy$ 。也记做：$\mathbb{E}_{X,Y\sim P(X,Y)}[g(X,Y)\int g(x,y)p(x,y)dxdy$ 。
3. 期望性质：
   - 常数的期望就是常数本身。
   - 对常数 C 有：$\mathbb{E}[CX]=C\mathbb{E}[X]$ 。
   - 对两个随机变量 X,Y，有：$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]$ 。

     该结论可以推广到任意有限个随机变量之和的情况。
   - 对两个相互独立的随机变量，有：$\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$ 。

     该结论可以推广到任意有限个相互独立的随机变量之积的情况。

2.2 方差

1. 对随机变量 X ，若 $\mathbb{E}[(X-\mathbb{E}[X])^2]$ 存在，则称它为 X 的方差，记作 $Var[X]$ 。

   X 的标准差为方差的开平方。即：

$$Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$
$$\sigma = \sqrt{Var[X]}$$

- 方差度量了随机变量 X 与期望值偏离的程度，衡量了 X 取值分散程度的一个尺度。
- 由于绝对值 $|X - \mathbb{E}[X]|$ 带有绝对值，不方便运算，因此采用平方来计算。

  又因为 $|X - \mathbb{E}[X]|^2$ 是一个随机变量，因此对它取期望，即得 X 与期望值偏离的均值。

2. 根据定义可知：

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

3. 对于一个期望为 μ， 方差为 $\sigma^2, \sigma \neq 0$ 的随机变量 X，随机变量 $X^* = \frac{X-\mu}{\sigma}$ 的数学期望为0，方差为1。 称 X^* 为 X 的标准化变量。
4. 方差的性质：
   - 常数的方差恒为 0 。
   - 对常数 C ，有 $Var[CX] = C^2 Var[X]$ 。
   - 对两个随机变量 X, Y，有： $Var[X + Y] = Var[X] + Var[Y] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

     当 X 和 Y 相互独立时，有 $Var[X + Y] = Var[X] + Var[Y]$ 。这可以推广至任意有限多个相互独立的随机变量之和的情况。
   - $Var[X] = 0$ 的充要条件是 X 以概率1取常数。

2.3 协方差与相关系数

1. 对于二维随机变量 (X, Y) ，可以讨论描述 X 与 Y 之间相互关系的数字特征。
   - 定义 $\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ 为随机变量 X 与 Y 的协方差，记作 $Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$。
   - 定义 $\rho_{XY} = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$ 为随机变量 X 与Y的相关系数，它是协方差的归一化。
2. 由定义可知：

$$Cov[X, Y] = Cov[Y, X]$$
$$Cov[X, X] = Var[X]$$
$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

3. 协方差的性质：
   - $Cov[aX, bY] = abCov[X, Y]$， a, b 为常数。
   - $Cov[X_1 + X_2, Y] = Cov[X_1, Y] + Cov[X_2, Y]$
   - $Cov[f(X), g(Y)] = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])]$
   - $\rho[f(X), g(Y)] = \frac{Cov[f(X),g(Y)]}{\sqrt{Var[f(X)]}\sqrt{Var[g(Y)]}}$
4. 协方差的物理意义：
   - 协方差的绝对值越大，说明两个随机变量都远离它们的均值。
   - 协方差如果为正，则说明两个随机变量同时趋向于取较大的值或者同时趋向于取较小的值；如果为负，则说明一个随变量趋向于取较大的值，另一个随机变量趋向于取较小的值。
   - 两个随机变量的独立性可以导出协方差为零。但是两个随机变量的协方差为零无法导出独立性。

因为独立性也包括：没有非线性关系。有可能两个随机变量是非独立的，但是协方差为零。如：假设随机变量 $X \sim U[-1,1]$ 。定义随机变量 S 的概率分布函数为：

$$P(S=1)=\frac{1}{2}P(S=-1)=\frac{1}{2}$$

定义随机变量 $Y=SX$，则随机变量 X,Y 是非独立的，但是有： $Cov[X,Y]=0$ 。

5. 相关系数的物理意义：考虑以随机变量 X 的线性函数 $a+bX$ 来近似表示 Y。以均方误差

$$e=\mathbb{E}[(Y-(a+bX))^2]=\mathbb{E}[Y^2]+b^2\mathbb{E}[X^2]+a^2-2b\mathbb{E}[XY]+2ab\mathbb{E}[X]-2a\mathbb{E}[Y]$$

来衡量以 $a+bX$ 近似表达 Y 的好坏程度。 e 越小表示近似程度越高。

为求得最好的近似，则对 a,b 分别取偏导数，得到：

$$a_0=\mathbb{E}[Y]-b_0\mathbb{E}[X]=\mathbb{E}[Y]-\mathbb{E}[X]\frac{Cov[X,Y]}{Var[X]}$$

$$b_0=\frac{Cov[X,Y]}{Var[X]}$$

$$\min(e)=\mathbb{E}[(Y-(a_0+b_0X))^2]=(1-\rho_{XY}^2)Var[Y]$$

因此有以下定理：

- $|\rho_{XY}|\leq 1$ （$|\cdot|$ 是绝对值）。
- $|\rho_{XY}|=1$ 的充要条件是：存在常数 a,b 使得 $P\{Y=a+bX\}=1$。

6. 当 $|\rho_{XY}|$ 较大时， e 较小，意味着随机变量 X 和 Y 联系较紧密。于是 ρ_{XY} 是一个表征 X、Y 之间线性关系紧密程度的量。

7. 当 $\rho_{XY}=0$ 时，称 X 和 Y 不相关。

- 不相关是就线性关系来讲的，而相互独立是一般关系而言的。
- 相互独立一定不相关；不相关则未必独立。

2.4 协方差矩阵

1. 设 X 和 Y 是随机变量。

- 若 $\mathbb{E}[X^k],k=1,2,\cdots$ 存在，则称它为 X 的 k 阶原点矩，简称 k 阶矩。
- 若 $\mathbb{E}[(X-\mathbb{E}[X])^k],k=2,3,\cdots$ 存在，则称它为 X 的 k 阶中心矩。
- 若 $\mathbb{E}[X^kY^l],k,l=1,2,\cdots$ 存在，则称它为 X 和 Y 的 $k+l$ 阶混合矩。
- 若 $\mathbb{E}[(X-\mathbb{E}[X])^k(Y-\mathbb{E}[Y])^l],k,l=1,2,\cdots$ 存在，则称它为 X 和 Y 的 $k+l$ 阶混合中心矩。

因此：期望是一阶原点矩，方差是二阶中心矩，协方差是二阶混合中心矩。

2. 协方差矩阵：

- 二维随机变量 (X_1,X_2) 有四个二阶中心矩（假设他们都存在），记作：

$$\begin{aligned}c_{11}&=\mathbb{E}[(X_1-\mathbb{E}[X_1])^2]\\c_{12}&=\mathbb{E}[(X_1-\mathbb{E}[X_1])(X_2-\mathbb{E}[X_2])]\\c_{21}&=\mathbb{E}[(X_2-\mathbb{E}[X_2])(X_1-\mathbb{E}[X_1])]\\c_{22}&=\mathbb{E}[(X_2-\mathbb{E}[X_2])^2]\end{aligned}$$

称矩阵

$$\mathbf{C}=\begin{bmatrix}c_{11}&c_{12}\\c_{21}&c_{22}\end{bmatrix}$$

为随机变量 (X_1, X_2) 的协方差矩阵。

- 设 n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 的二阶混合中心矩
  $c_{ij} = Cov[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$ 都存在，则称矩阵

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

  为 n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 的协方差矩阵。

  由于 $c_{ij} = c_{ji}, i \neq j, i, j = 1, 2, \cdots, n$ 因此协方差矩阵是个对称阵。

3. 通常 n 维随机变量的分布是不知道的，或者太复杂以致数学上不容易处理。因此实际中协方差矩阵非常重要。

三、大数定律及中心极限定理

3.1 切比雪夫不等式

1. 切比雪夫不等式：假设随机变量 X 具有期望 $\mathbb{E}[X] = \mu$， 方差 $Var(X) = \sigma^2$ ，则对于任意正数 ε ，下面的不等式成立：

$$P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

   - 其意义是：对于距离 $\mathbb{E}[X]$ 足够远的地方 （距离大于等于 ε ），事件出现的概率是小于等于 $\frac{\sigma^2}{\varepsilon^2}$。即事件出现在区间 $[\mu - \varepsilon, \mu + \varepsilon]$ 的概率大于 $1 - \frac{\sigma^2}{\varepsilon^2}$ 。

     该不等式给出了随机变量 X 在分布未知的情况下， 事件 $\{|X - \mu| \leq \varepsilon\}$ 的下限估计。如：
     $P\{|X - \mu| < 3\sigma\} \geq 0.8889$ 。

   - 证明：

$$\begin{aligned} P\{|X - \mu| \geq \varepsilon\} &= \int_{|x-\mu| \geq \varepsilon} p(x)dx \leq \int_{|x-\mu| \geq \varepsilon} \frac{|x - \mu|^2}{\varepsilon^2} p(x)dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx = \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

2. 切比雪夫不等式的特殊情况：设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，且具有相同的数学期望和方差：$\mathbb{E}[X_k] = \mu, Var[X_k] = \sigma^2$。 作前 n 个随机变量的算术平均： $\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$， 则对于任意正数 ε 有：

$$\lim_{n \to \infty} P\{|\overline{X} - \mu| < \varepsilon\} = \lim_{n \to \infty} P\{|\frac{1}{n} \sum_{k=1}^{n} X_k - \mu| < \varepsilon\} = 1$$

   证明：根据期望和方差的性质有：$\mathbb{E}[\overline{X}] = \mu$， $Var[\overline{X}] = \frac{\sigma^2}{n}$ 。根据切比雪夫不等式有：

$$P\{|\overline{X} - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{n\varepsilon^2}$$

   则有 $\lim_{n \to \infty} P\{|\overline{X} - \mu| \geq \varepsilon\} = 0$ ，因此有：$\lim_{n \to \infty} P\{|\overline{X} - \mu| < \varepsilon\} = 1$ 。

3.2 大数定理

1. 依概率收敛：设 $Y_1,Y_2,\cdots,Y_n,\cdots$ 是一个随机变量序列， a 是一个常数。

   若对于任意正数 ε 有： $\lim_{n\to\infty}P\{|Y_n-a|\le\varepsilon\}=1$ ，则称序列 $Y_1,Y_2,\cdots,Y_n,\cdots$ 依概率收敛于 a 。

   记作： $Y_n\xrightarrow{P}a$

2. 依概率收敛的两个含义：

   - 收敛：表明这是一个随机变量序列，而不是某个随机变量；且序列是无限长，而不是有限长。
   - 依概率：表明序列无穷远处的随机变量 Y_∞ 的分布规律为：绝大部分分布于点 a ，极少数位于 a 之外。且分布于 a 之外的事件发生的概率之和为0。

3. 大数定理一： 设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，且具有相同的数学期望和方差：$\mathbb{E}[X_k]=\mu,Var[X_k]=\sigma^2$。 则序列： $\overline{X}=\frac{1}{n}\sum_{k=1}^nX_k$ 依概率收敛于 μ ， 即 $\overline{X}\xrightarrow{P}\mu$ 。

   注意：这里并没有要求随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 同分布。

4. 伯努利大数定理： 设 n_A 为 n 次独立重复实验中事件 A 发生的次数， p 是事件 A 在每次试验中发生的概率。则对于任意正数 ε 有：

$$\lim_{n\to\infty}P\{|\frac{n_A}{n}-p|<\varepsilon\}=1$$
$$or:\quad\lim_{n\to\infty}P\{|\frac{n_A}{n}-p|\ge\varepsilon\}=0$$

   即：当独立重复实验执行非常大的次数时，事件 A 发生的频率逼近于它的概率。

5. 辛钦定理：设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，服从同一分布，且具有相同的数学期望：$\mathbb{E}[X_k]=\mu$。 则对于任意正数 ε 有：

$$\lim_{n\to\infty}P\{|\frac{1}{n}\sum_{k=1}^nX_k-\mu|<\varepsilon\}=1$$

   - 注意：这里并没有要求随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 的方差存在。
   - 伯努利大数定理是亲钦定理的特殊情况。

3.3 中心极限定理

1. 独立同分布的中心极限定理：设随机变量 $X_1,X_2,\cdots,X_n$ 独立同分布，且具有数学期望和方差：$\mathbb{E}[X_k]=\mu,Var[X_k]=\sigma^2$， 则随机变量之和 $\overline{SX_n}=\sum_{k=1}^nX_k$ 的标准变化量：

$$Y_n=\frac{\overline{SX_n}-\mathbb{E}[\overline{SX_n}]}{\sqrt{Var[\overline{SX_n}]}}=\frac{\overline{SX_n}-n\mu}{\sqrt{n}\sigma}$$

   的概率分布函数 $F_n(x)$ 对于任意 x 满足：

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\{Y_n\le x\}=\lim_{n\to\infty}P\{\frac{\sum_{k=1}^nX_k-n\mu}{\sqrt{n}\sigma}\le x\}$$
$$=\int_{-\infty}^x\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x)$$

   - 其物理意义为：均值方差为 μ,σ^2 的独立同分布的随机变量 $X_1,X_2,\cdots,X_n$ 之和 $\overline{SX_n}=\sum_{k=1}^nX_k$ 的标准变化量 Y_n，当 n 充分大时，其分布近似于标准正态分布。

     即： $\overline{SX_n}=\sum_{k=1}^nX_k$ 在 n 充分大时，其分布近似于 $N(n\mu,n\sigma^2)$ 。
   - 一般情况下，很难求出 n 个随机变量之和的分布函数。因此当 n 充分大时，可以通过正态分布来做理论上的分析或者计算。

2. `Liapunov` 定理：设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，具有数学期望和方差：$\mathbb{E}[X_k] = \mu_k, Var[X_k] = \sigma_k^2$ 。记：$B_n^2 = \sum_{k=1}^n \sigma_k^2$。若存在正数 δ，使得当 $n \to \infty$ 时，$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}[|X_k - \mu_k|^{2+\delta}] \to 0$ 。则随机变量之和 $\overline{SX_n} = \sum_{k=1}^n X_k$ 的标准变化量：

$$Z_n = \frac{\overline{SX_n} - \mathbb{E}[\overline{SX_n}]}{\sqrt{Var[\overline{SX_n}]}} = \frac{\overline{SX_n} - \sum_{k=1}^n \mu_k}{B_n}$$

的概率分布函数 $F_n(x)$ 对于任意 x 满足：

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\{Z_n \le x\} = \lim_{n\to\infty} P\{\frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n} \le x\}$$
$$= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x)$$

   - 其物理意义为：相互独立的随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 之和 $\overline{SX_n} = \sum_{k=1}^n X_k$的衍生随机变量序列 $Z_n = \frac{\overline{SX_n} - \sum_{k=1}^n \mu_k}{B_n}$，当 n 充分大时，其分布近似与标准正态分布。
   - 这里并不要求 $X_1, X_2, \cdots, X_n, \cdots$ 同分布。

9. `Demoiver-Laplace`定理：设随机变量序列 $\eta_n, n = 1, 2, \ldots$ 服从参数为 (n, p) 的二项分布，其中 $0 < p < 1$ 。则对于任意 x , 有：

$$\lim_{n\to\infty} P\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2|2} dt = \Phi(x)$$

   该定理表明，正态分布是二项分布的极限分布。当 n 充分大时，可以利用正态分布来计算二项分布的概率。

五、常见概率分布

5.1 均匀分布

1. 离散随机变量的均匀分布：假设 X 有 k 个取值： $x_1, x_2, \cdots, x_k$，则均匀分布的概率密度函数(`probability mass function:PMF`)为：

$$p(X = x_i) = \frac{1}{k}, \quad i = 1, 2, \cdots, k$$

2. 连续随机变量的均匀分布： 假设 X 在 `[a,b]` 上均匀分布，则其概率密度函数(`probability density function：PDF`)为：

$$p(X = x) = \begin{cases} 0, & x \notin [a, b] \\ \frac{1}{b-a}, & x \in [a, b] \end{cases}$$

.

5.2 伯努利分布

1. 伯努利分布：参数为 $\phi \in [0, 1]$。随机变量 $X \in \{0, 1\}$ 。
   - 概率分布函数为：$p(X = x) = \phi^x (1 - \phi)^{1-x}$, $x \in \{0, 1\}$ 。
   - 期望： $\mathbb{E}[X] = \phi$ 。方差： $Var[X] = \phi(1 - \phi)$ 。
2. `categorical` 分布：它是二项分布的推广，也称作 `multinoulli` 分布。假设随机变量 $X \in \{1, 2, \cdots, K\}$ ，其概率分布函数为：

$$
\begin{aligned}
p(X=1)&=\theta_1\\
p(X=2)&=\theta_2\\
&\vdots\\
p(X=K-1)&=\theta_{K-1}\\
p(X=K)&=1-\sum_{i=1}^{K-1}\theta_i
\end{aligned}
$$

其中 θ_i 为参数，它满足 $\theta_i \in [0,1]$，且 $\sum_{i=1}^{K-1}\theta_i \in [0,1]$ 。

5.3 二项分布

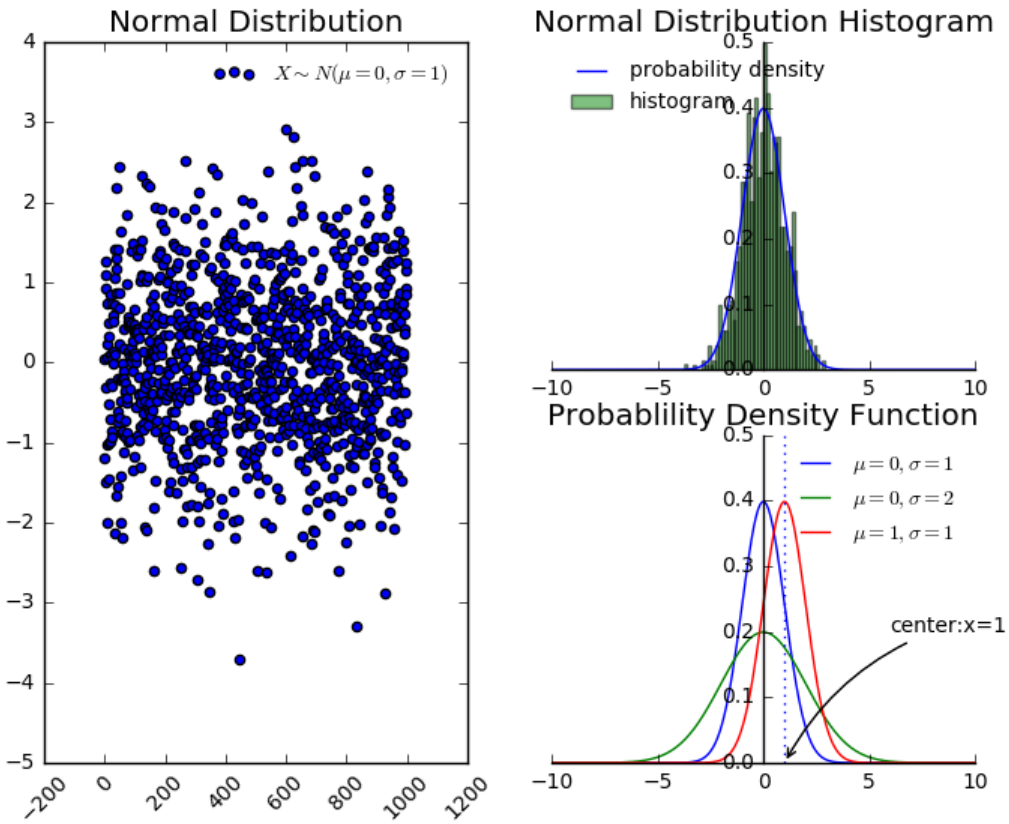
1. 假设试验只有两种结果：成功的概率为 ϕ，失败的概率为 $1-\phi$。 则二项分布描述了：独立重复地进行 n 次试验中，成功 x 次的概率。
   - 概率质量函数：
   $$p(X=x)=\frac{n!}{x!(n-x)!}\phi^x(1-\phi)^{n-x}, x\in\{0,1,\cdots,n\}$$
   - 期望： $\mathbb{E}[X]=n\phi$。 方差： $Var[X]=n\phi(1-\phi)$ 。

5.4 高斯分布

1. 正态分布是很多应用中的合理选择。如果某个随机变量取值范围是实数，且对它的概率分布一无所知，通常会假设它服从正态分布。有两个原因支持这一选择：
   - 建模的任务的真实分布通常都确实接近正态分布。中心极限定理表明，多个独立随机变量的和近似正态分布。
   - 在具有相同方差的所有可能的概率分布中，正态分布的熵最大（即不确定性最大）。

5.4.1 一维正态分布

1. 正态分布的概率密度函数为 :
   $$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, -\infty<x<\infty$$
   其中 $\mu,\sigma(\sigma>0)$ 为常数。
   - 若随机变量 X 的概率密度函数如上所述，则称 X 服从参数为 μ,σ 的正态分布或者高斯分布，记作 $X\sim N(\mu,\sigma^2)$。
   - 特别的，当 $\mu=0,\sigma=1$ 时，称为标准正态分布，其概率密度函数记作 $\varphi(x)$，分布函数记作 $\Phi(x)$ 。
   - 为了计算方便，有时也记作：$\mathcal{N}(x;\mu,\beta^{-1})=\sqrt{\frac{\beta}{2\pi}}\exp\left(-\frac{1}{2}\beta(x-\mu)^2\right)$ ，其中 $\beta\in(0,\infty)$ 。
2. 正态分布的概率密度函数性质：
   - 曲线关于 $x=\mu$ 对称。
   - 曲线在 $x=\mu$ 时取最大值。
   - 曲线在 $x=\mu\pm\sigma$ 处有拐点。
   - 参数 μ 决定曲线的位置； σ 决定图形的胖瘦。



3. 若 $X \sim N(\mu, \sigma^2)$ 则：
   - $\frac{X-\mu}{\sigma} \sim N(0,1)$
   - 期望：$\mathbb{E}[X]=\mu$。方差：$Var[X]=\sigma^2$。
4. 有限个相互独立的正态随机变量的线性组合仍然服从正态分布：若随机变量 $X_i \sim N(\mu_i, \sigma_i^2), i=1,2,\cdots,n$ 且它们相互独立，则它们的线性组合：$C_1X_1+C_2X_2+\cdots+C_nX_n$ 仍然服从正态分布（其中$C_1, C_2, \cdots, C_n$ 不全是为 0 的常数），且：$C_1X_1+C_2X_2+\cdots+C_nX_n \sim N(\sum_{i=1}^n C_i\mu_i, \sum_{i=1}^n C_i^2\sigma_i^2)$。

5.4.2 多维正态分布

1. 二维正态随机变量 (X,Y) 的概率密度为：

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

   根据定义，可以计算出:

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x-\mu_1)^2/(2\sigma_1^2)}, -\infty < x < \infty$$
$$p_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(y-\mu_2)^2/(2\sigma_2^2)}, -\infty < y < \infty$$
$$\mathbb{E}[X] = \mu_1$$
$$\mathbb{E}[Y] = \mu_2$$
$$Var[X] = \sigma_1^2$$
$$Var[Y] = \sigma_2^2$$
$$Cov[X,Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_1)(y-\mu_2)p(x,y)dxdy = \rho\sigma_1\sigma_2$$
$$\rho_{XY} = \rho$$

2. 引入矩阵：

$$\vec{\mathbf{x}} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$\mathbf{\Sigma}$ 为 (X,Y) 的协方差矩阵。其行列式为 $\det\mathbf{\Sigma} = \sigma_1^2\sigma_2^2(1-\rho^2)$，其逆矩阵为：

$$\mathbf{\Sigma}^{-1} = \frac{1}{\det\mathbf{\Sigma}} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

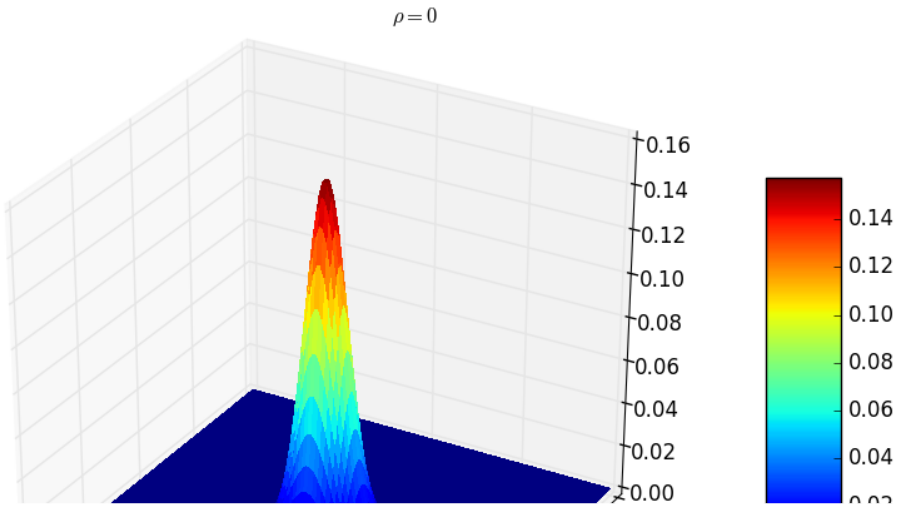
于是 (X,Y) 的概率密度函数可以写作 $(\vec{\mathbf{x}}-\vec{\mu})^T$ 表示矩阵的转置：

$$p(x,x) = \frac{1}{(2\pi)(\det\mathbf{\Sigma})^{1/2}} \exp\{-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\vec{\mathbf{x}}-\vec{\mu})\}$$

其中：

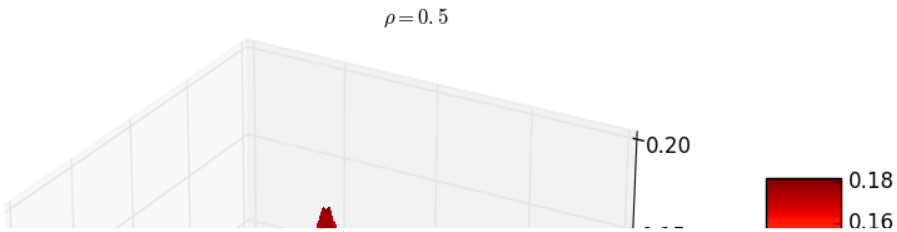
- 均值 μ_1, μ_2 决定了曲面的位置（本例中均值都为0）。
- 标准差 σ_1, σ_2 决定了曲面的陡峭程度（本例中方差都为1）。
- ρ 决定了协方差矩阵的形状，从而决定了曲面的形状。
  - $\rho = 0$ 时，协方差矩阵对角线非零，其他位置均为零。此时表示随机变量之间不相关。

    此时的联合分布概率函数形状如下图所示，曲面在 $z = 0$ 平面的截面是个圆形：



- $\rho = 0.5$ 时，协方差矩阵对角线非零，其他位置非零。此时表示随机变量之间相关。

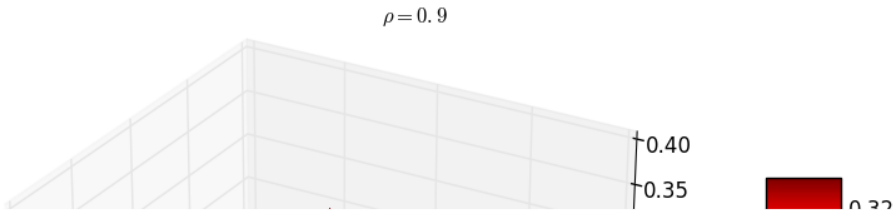
  此时的联合分布概率函数形状如下图所示，曲面在 $z = 0$ 平面的截面是个椭圆，相当于圆形沿着直线 $y = x$ 方向压缩：



  - $\rho = 1$ 时，协方差矩阵对角线非零，其他位置非零。

    此时表示随机变量之间完全相关。此时的联合分布概率函数形状为：曲面在 $z = 0$ 平面的截面是直线 $y = x$，相当于圆形沿着直线 $y = x$ 方向压缩成一条直线 。

    由于 $\rho = 1$会导致除数为 0，因此这里给出 $\rho = 0.9$：



3. 多维正态随机变量 $(X_1, X_2, \cdots, X_n)$，引入列矩阵：

$$\vec{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

$\mathbf{\Sigma}$ 为 $(X_1, X_2, \cdots, X_n)$ 的协方差矩阵。则：

$$p(x_1, x_2, x_3, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{\Sigma})^{1/2}} \exp\{-\frac{1}{2}(\vec{\mathbf{x}} - \vec{\mu})^T \mathbf{\Sigma}^{-1}(\vec{\mathbf{x}} - \vec{\mu})\}$$

记做：$\mathcal{N}(\vec{\mathbf{x}}; \vec{\mu}, \mathbf{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n det(\mathbf{\Sigma})}} \exp\big(-\frac{1}{2}(\vec{\mathbf{x}} - \vec{\mu})^T \mathbf{\Sigma}^{-1}(\vec{\mathbf{x}} - \vec{\mu})\big)$ 。

4. n 维正态变量具有下列四条性质：

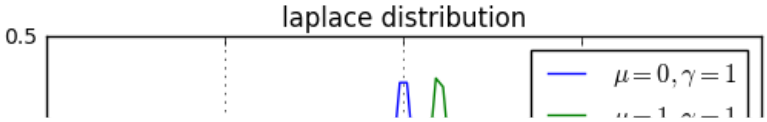
   - n 维正态变量的每一个分量都是正态变量；反之，若 $X_1, X_2, \cdots, X_n$ 都是正态变量，且相互独立，则 $(X_1, X_2, \cdots, X_n)$ 是 n 维正态变量。
   - n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布的充要条件是：$X_1, X_2, \cdots, X_n$ 的任意线性组合：$l_1 X_1 + l_2 X_2 + \cdots + l_n X_n$ 服从一维正态分布，其中 $l_1, l_2, \cdots, l_n$ 不全为 0 。
   - 若 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布，设 $Y_1, Y_2, \cdots, Y_k$ 是 $X_j, j = 1, 2, \cdots, n$ 的线性函数，则 $(Y_1, Y_2, \cdots, Y_k)$ 也服从多维正态分布。

     这一性质称为正态变量的线性变换不变性。
   - 设 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布，则 $X_1, X_2, \cdots, X_n$ 相互独立 $\Longleftrightarrow$ $X_1, X_2, \cdots, X_n$ 两两不相关。

5.5 拉普拉斯分布

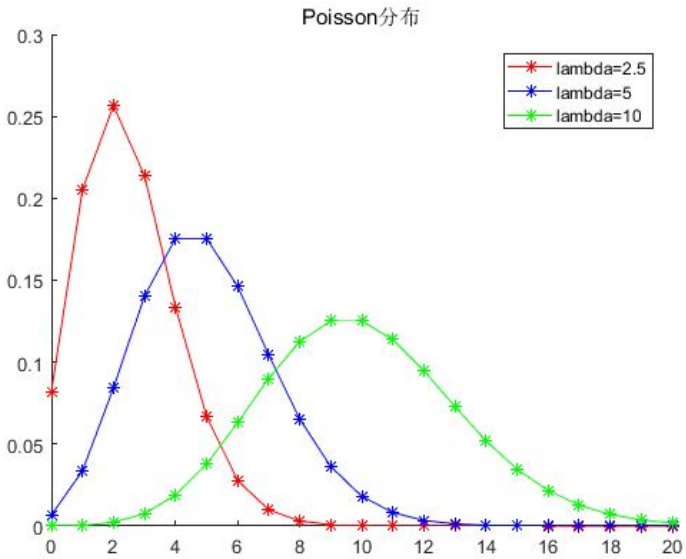
1. 拉普拉斯分布：

- 概率密度函数：$p(x;\mu,\gamma)=\frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)$ 。
- 期望： $\mathbb{E}[X]=\mu$ 。方差： $Var[X]=2\gamma^2$ 。



5.6 泊松分布

1. 假设已知事件在单位时间（或者单位面积）内发生的**平均**次数为 λ ，则泊松分布描述了：事件在单位时间（或者单位面积）内发生的具体次数为 k 的概率。
   - 概率质量函数：$p(X=k;\lambda)=\frac{e^{-\lambda}\lambda^k}{k!}$ 。
   - 期望： $\mathbb{E}[X]=\lambda$。 方差： $Var[X]=\lambda$ 。



2. 用均匀分布模拟泊松分布：

```
def make_poisson(lmd,tm):
    '''
    用均匀分布模拟泊松分布。 lmd为 lambda 参数； tm 为时间
    '''
    t=np.random.uniform(0,tm,size=lmd*tm) # 获取 lmd*tm 个事件发生的时刻
    count,tm_edges=np.histogram(t,bins=tm,range=(0,tm))#获取每个单位时间内，事件发生的次数
    max_k= lmd *2 # 要统计的最大次数
    dist,count_edges=np.histogram(count,bins=max_k,range=(0,max_k),density=True)
    x=count_edges[:-1]
    return x,dist,stats.poisson.pmf(x,lmd)
```

该函数：

- 首先随机性给出了 `lmd*tm` 个事件发生的时间（时间位于区间 `[0,tm]` ）内。
- 然后统计每个单位时间区间内，事件发生的次数。
- 然后统计这些次数出现的频率。
- 最后将这个频率与理论上的泊松分布的概率质量函数比较。

5.7 指数分布

1. 若事件服从泊松分布，则该事件前后两次发生的时间间隔服从指数分布。由于时间间隔是个浮点数，因此指数分布是连续分布。
   - 概率密度函数： （ t 为时间间隔）

$$p(t;\lambda)=\begin{cases}0, & t<0\\ \frac{\lambda}{\exp(\lambda t)}, & t\geq 0\end{cases}$$

   - 期望： $\mathbb{E}[t]=\frac{1}{\lambda}$ 。方差： $Var[t]=\frac{1}{\lambda^2}$ 。

2. 用均匀分布模拟指数分布：

```
def make_expon(lmd,tm):
    '''
    用均匀分布模拟指数分布。 lmd为 lambda 参数； tm 为时间
    '''
    t=np.random.uniform(0,tm,size=lmd*tm) # 获取 lmd*tm 个事件发生的时刻
    sorted_t=np.sort(t) #时刻升序排列
    delt_t=sorted_t[1:]-sorted_t[:-1] #间隔序列
    dist,edges=np.histogram(delt_t,bins="auto",density=True)
    x=edges[:-1]
    return x,dist,stats.expon.pdf(x,loc=0,scale=1/lmd) #scale 为 1/lambda
```

5.8 伽马分布

1. 若事件服从泊松分布，则事件第 i 次发生和第 $i+k$ 次发生的时间间隔为伽玛分布。由于时间间隔是个浮点数，因此指数分布是连续分布。
   - 概率密度函数： $p(t;\lambda,k)=\frac{t^{(k-1)}\lambda^k e^{(-\lambda t)}}{\Gamma(k)}$ ， t 为时间间隔 。
   - 期望： $\mathbb{E}[t]=\frac{k}{\lambda}$ 。方差： $Var[t]=\frac{k}{\lambda^2}$ 。

2. 上面的定义中 k 必须是整数。事实上，若随机变量 X 服从伽马分布，则其概率密度函数为：

$$p(X;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}X^{\alpha-1}e^{-\beta X},\quad X>0$$

记做 $\Gamma(\alpha,\beta)$ 。其中 α 称作形状参数，β 称作尺度参数。

- 期望 $\mathbb{E}[X]=\frac{\alpha}{\beta}$，方差$Var[X]=\frac{\alpha}{\beta^2}$ 。
- 当 $\alpha\leq 1$ 时，$p(X;\alpha,\beta)$ 为递减函数。
- 当 $\alpha>1$ 时，$p(X;\alpha,\beta)$ 为单峰函数。

3. 性质：
   - 当 $\beta=n$ 时， 为 `Erlang` 分布。
   - 当 $\alpha=1,\beta=\lambda$ 时，就是参数为 λ 的指数分布。
   - 当 $\alpha=\frac{n}{2},\beta=\frac{1}{2}$ 时，就是常用的卡方分布。
4. 伽马分布的可加性：设随机变量 $X_1,X_2,\cdots,X_n$ 相互独立并且都服从伽马分布： $X_i\sim\Gamma(\alpha_i,\beta)$，则：

$$X_1+X_2+\cdots+X_n\sim\Gamma(\alpha_1+\alpha_2+\cdots+\alpha_n,\beta)$$

5. 用均匀分布模拟伽玛分布：

```
def make_gamma(lmd,tm,k):
    '''
    用均匀分布模拟伽玛分布。 lmd为 lambda 参数； tm 为时间；k 为 k 参数
    '''
    t=np.random.uniform(0,tm,size=lmd*tm) # 获取 lmd*tm 个事件发生的时刻
    sorted_t=np.sort(t) #时刻升序排列
    delt_t=sorted_t[k:]-sorted_t[:-k] #间隔序列
    dist,edges=np.histogram(delt_t,bins="auto",density=True)
    x=edges[:-1]
    return x,dist,stats.gamma.pdf(x,loc=0,scale=1/lmd,a=k) #scale 为 1/lambda,a 为 k
```

5.9 贝塔分布

1. 贝塔分布是定义在 $(0,1)$ 之间的连续概率分布。

   如果随机变量 X 服从贝塔分布，则其概率密度函数为：