1. 概率的数学定义：

   三元组的表示法：(Ω, F, P)。其中Ω 表示所有平行世界的集合，总的面积为1，P表示部分的Ω （也就是它的子集）的面积，F表示某一平行世界的随机变量的值。随机变量是一个很重要的概念。

2. 联合概率和边缘概率

   一个平行世界的面积需要用到多个随机变量来描述的时候就是联合概率分布。

在多个随机变量中只考虑部分随机变量的情况的就是边缘概率

如果考虑在一个时间已经发生的情况下的一个时间的概率就是条件概率，这个时候相当于是把整个世界的面积切割成一个的更小的范围。在这个小的宇宙中找寻真理

1. 条件概率、联合概率、边缘概率是可以互相转化的。多远的随机变量的时候尤为重要，能够很好的推导出公式是很重要的一步。

   从联合概率到条件概率

   $$P(A,B,C) = P(A|B,C) * P(B|C) * P(C) = P(B|A,C) * P(A|C) * P(C) = P(C|A,B) * P(A|B) * P(B)$$

   从条件概率到联合概率

   $$P(A|B,C,D) = P(A,B|C,D) * P(C|D) * P(D)$$

2. 贝叶斯公式

   贝叶斯公式的精髓是在已经知道结果的基础上，推测原因出现的概率。其中在不管结果如何的时候，知道的原因概率分布，这个就会是先验概率分布。需要求解的知道结果后的原因的概率分布是后验概率分布。基本的贝叶斯应用的场景是下面的模型

   - 已知所有的P（原因）和P（结果|原因）一览
   - 求结果P（原因|结果）

   比如邮件过滤系统中，这封邮件是不是垃圾邮件，这是一个因，它有自己的概率分布P（垃圾邮件），和系统是否判断这是不是一个垃圾邮件没有关系。系统判断的结果正确与否也有一个概率分布，这个和它本身是不是垃圾邮件也是有关的，也就是概率分布P（判断结果|是否是垃圾邮件）。现在要求在系统判断这是一封垃圾邮件的情况下，这封邮件确实垃圾邮件的概率是多少P（是否是垃圾邮件|系统判断为垃圾邮件）？

   公式内容：

   $$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

通常来说既是两个步骤：（B是原因，A是结果）

- 计算B在所有情况下的P（A,B）的联合概率
- 计算所需要的特定原因值B0和A的联合分布。就可以倒推出在结果已经出现的情况下原因出现的概率分布。

3. 随机变量的独立理解

   1. 独立和均匀分布并不一样
   2. 独立和独立同分布不一样

   事件独立的定义

> 如果事件A和事件B满足P（A|B）=P（A|非B），称为A和B独立。

事件独立性的几种不同的表述

- 条件概率和条件无关
- 添加条件或者删除条件不影响概率
- 联合概率之比相同
- **联合概率** 是边缘概率的乘积

  P(AB) = P(A)P(B)

两个随机变量独立的几种描述

- 联合概率能够分解为仅含A和B的式子。$P（A,B）=h（A）*g（B）$
- 其他的表述和事件独立的表述一致

多个随机变量的独立性（有些地方是有些迷惑的）

- 部分随机变量的互相独立不能代表所有的随机变量是独立的
- 多个随机变量独立的定义：比如4个随机变量A,B,C,D是独立的，那么这个事实说明了下面两个点：
  - 其中任意的三个事件都是独立的
  - $P(A,B,C,D)=P(A)_P(B)_P(C)*P(D)$

  只有第二个条件是没法说明四个随机变量独立的，很关键。

4. 随机变量的期望和方差
   - 一个随机变量的期望表示这个随机变量的平均水平，方差表示分布的分散程度。有一个很有意思的点就是

比如随机变量X 的期望是μ 这意味着$E(X-\mu)=0$,但是$E[(X-,u)^2]\neq 0$,后面的这个式子也就是指方差了。这么说方差也只是一种特殊的期望而已。

- 不同的随机变量的期望以及方差之和。

  这是一个很有意思的问题，其实很多时候可以根据期望和方差的性质进行数学推到得到这个结论，但是这里还是说一下：相互独立的随机变量和的方差等于方差的和，
  $V(Y_1+Y_2+...+Y_n)=V(Y_1)+V(Y_2)+...+V(Y_n)$

  二项分布的方差推到就可以根据这个性质得到。

- 方差和期望的关系

$$V(X)=E(X^2)-E(X)^2$$

这个公式要记住方差本质就是一种期望，更可以说是一种二阶期望，所以$E(X^2)$ 某种意义是一个方差，利用变量替换的方法$Z=X-\mu$ 就可以推导出这个公式。

5. 大数定律

一个具有稳定的概率的事件，对于一次事件发生结果的预测我们是无能为力的，我们无法准确预测这个事件的结果，但是在很多这样的事件发生的情况下，我们就可以对于所有这些事件发生的规律做出准确地结论。这就是中心极限定律的要说明的事情。

6. 独立同分布

是指那些概率分布模型一样，但是又是相互独立的随机变量组成的一个模型成为独立同分布。

比如掷20次骰子，就是20次的一个独立同分布。

这里与有一个重要的概念上的分别：

比如掷骰子n次，这是一个正常的骰子，对于一次投掷来说，每一个点数的概率都是1/6，点数的期望是3.5，记住这点很重要。这n次组成一个独立同分布，现在这20个点数的平均值构成一个新的随机变量 $Y = \frac{X_1+X_2+...+X_n}{n}$ ，很容易的，我们可以通过期望的性质得到$E(Y) = \mu, V(Y) = \frac{\sigma^2}{n}$ 。这里我们可以得到一个的结论：**我们只要不断重复的做一个事件，而且它们独立，也就是**$n \to \infty$ **的过程，就能使得Y的方差接近于0。记得方差表示的是随机变量的随机程度，这就意味着Y这个随机变量变得不随机了，是一个基本上可以确定的量。** 这正是著名的大数定律

- 对于大数定理，可以这么理解它的重大意义。本来对于一个事件来说，我们是无法预测它的结果的，它的期望是一种上帝视角，对多个平行世界结果的总结，我们身在其中一个世界，按道理来说只能知道其中的一种的结果。但是中心极限定理给了我们一种工具，只要不断地重复做这件事，我们就能得到上帝视角，知晓这件事件的期望或者说是概率分布。

7. 条件期望

   条件期望$E(Y|X = a)$ 这个期望是指$\sum_b b * P(Y = b|X = a)$

连续分布的概率

1. 在连续分布的模型中，如果像离散分布的那样通过 值 — 概 率 值 的描述方法进行描述，将得不到有用的信息，因为任意一点的概率对于连续分布的模型来说都是零 $P(X = c) = 0$ (a<X<b)$。所以我们需要一种叫欧诺个全新的描述方式，这个就是概率密度函数和累计分布函数
2. 累计分布函数$F_X(a) = P(X \leq a)$

   概率密度函数$f_X(a) = F'_X(a)$ 。f_X 的值本身并不表示概率的值，所以可以大于1，它表示的是概率的密度。

   f_X 本身需要满足一定的要求，那就是概率的归一化和非负性的特点，所以并不是所有的函数都是和做概率密度函数
3. 概率密度的变量替换

   e.g 知道X的概率密度函数f_X ，现在有$Y = 3X + 3$ ，如何求得f_Y ？

   这时候很容易混淆一个概念就是直接带入$X = \frac{Y-3}{3}$ 到f_X 中。但是变量替换的过程不仅只是变量值的变换。这其中还包括了变量的拉伸，可以通过可拉伸的胶带上的墨水密度（深浅）来理解这个。

   比如如果只是均匀的伸展，每个地方的密度都会变稀，而且每个对应点的位置都会变化。

   如果做不均匀的拉伸，情况会更加复杂，拉伸更剧烈的地方会密度变化更大。这种变化程度可以通过微分$g'(x)$来描述，数学上可以做出严格的证明，这里不展开。最后可以得到公式

$$f_Y(y) = |\frac{f_X(x)}{g'(x)}|, y = g(x)$$

4. 联合分布、边缘分布、条件分布

   既然概率分布的表示方式有所不同，那么联合分布、边缘分布、条件分布这些公式和离散值的分布特征也会有一定的却别。

   - 联合分布

     $f_{X,Y}(x, y) = f(x, y)$ 体积表示一块区域的概率。

   - 边缘分布

     对于$f_X(x)$ 就是一个边缘分布，$f_X(x = a)$表示的是 $x = a$ 对于联合分布函数（以XOY坐标面为底）一个截面的面积。

- 条件分布

  按照离散值的条件概率定义

  $$P(X=a|Y=b)=\frac{P(X=a,Y=b)}{P(Y=b)}$$

  但是在连续值中，这个会产生0/0的错误，所以需要新的表示方法。于是在连续值中产生了一种表示条件分布的概率密度函数，它使用$x=a$对于联合分布函数（以XOY坐标面为底）截面的面积来描述。

  $g(y)=f_{X,Y}(a,y)$，但是这样的函数不一定会符合概率密度函数的性质，所以需要添加一个常数c对$g(y)$进行拉伸

  $h(y)=\frac{g(y)}{c}$，使得$\int_{-\infty}^{+\infty}h(y)dy=1/c\int_{-\infty}^{+\infty}g(y)dy=1$可以的到c的取值

  $c=\int_{-\infty}^{+\infty}g(y)dy$很容易看到c的表达式就是一个边缘概率的特殊值$f_X(a)$

  由此我们得到一个完整的函数可以用来表示条件概率密度了，那就是$h(y)$,这样连续值的条件概率就有了定义

  $$f_{Y|X}(b|a)=\frac{fX,Y(a,b)}{f_X(a)}$$

- 贝叶斯公式

  和离散值的贝叶斯公式相似，仅仅是把级数的操作替换成积分而已。

- 独立性

  两个随机变量独立可以看作是对于联合概率密度函数的图像

  在x轴任意一个值取得截面形状都是相同的，这里的形状相同不是能够等比例放大或缩小得到的图像，而是拉伸能够得到的图像，用数学式子表达就是

  $f_{X,Y}(a,y)=cf_{X,Y}(a',y)$恒成立，其中c是一个由a和a'决定的常量

  最后得到的公式和离散值是一样的，因为没有涉及级数，所以不用改成积分

  满足$f_{X,Y}(x,y)=f_X(x)f_Y(y)$的变量属于互相独立的

- 变量变换

  对于联合分布$f_{X,Y}(x,y)$来说，如果采取

  - $Z=aX+bY,W=cX+dY$的变换，则根据单变量的变换理论来理解，就相当于在一块油画布上进行斜向拉伸，每个点的位置会发生变化，而且点聚集的密度也会变化，可以根据线性代数的理论来求解

    $$\begin{pmatrix}Z\\W\end{pmatrix}=A\begin{pmatrix}X\\Y\end{pmatrix},A=\begin{pmatrix}a&b\\c&d\end{pmatrix}$$

    $$f_{Z,W}(z,w)=\frac{1}{|detA|}f_{X,Y}(x,y)$$

  - 采用有曲率的一般性变换

    $$\begin{cases}Z=g(x,y)\\W=h(x,y)\end{cases}$$

变换之后，点(x,y)移动到了点(z,w)，且一一对应。

点（x，y）附近的面积扩大倍率为$|\partial(z,w)/\partial(x,y)|$（数学上可以解释，这里不做说明），$\partial(z,w)/\partial(x,y)$ 称作雅可比式

$$\frac{\partial(z,w)}{\partial(x,y)} = det\begin{pmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \end{pmatrix}$$

这样变换的点位置变换和面积倍率变化都出来了，总结为下面的公式：

$$f_{Z,W}(z,w) = \frac{1}{|\partial(z,w)/\partial(x,y)|} f_{X,Y}(x,y)$$

- 期望值和方差
  - 连续值的期望定义是和离散值一致的，都是相应的随机变量值和对应概率的乘积在求和。在连续值中不能直接得到一个值的概率，但是可以使用微分的概念求一小段区间的概率，这一段的随机变量取值使用区间起始值代替，最后就是得到一个积分了

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

至于期望的那些基本性质和离散值的期望一致。

推广到二维随机变量的期望

$$E(Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y f_{X,Y}(x,y) dx dy$$
$$E(X) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X,Y}(x,y) dx dy$$
$$E(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{X,Y}(x,y) dx dy$$
$$E(h(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x,y) f_{X,Y}(x,y) dx dy$$

- 方差定义的时候并没有使用某个具体值的概率，所以离散值的推导过程对于连续值还是适用。

$$V(X) = E(X^2) - E(X)^2$$

5. 正态分布

   标准正态分布的公式：

$$f(x) = \square e^{-\frac{x^2}{\triangle}}$$

上面的两个系数先不给出，我们需要通过一番计算求解。

标准正态分布是一种概率分布，既然如此就要满足概率分布密度函数的性质——和坐标轴围成的面积为1，由此可以确定□ 的值。

这里需要一个著名的公式进行辅助：高斯积分公式

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

$$\xrightarrow{\text{推广}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{c}}\,dx = \sqrt{c\pi}$$

由此可以知道$\square = \frac{1}{\sqrt{\triangle\pi}}$，现在还不能确定两个系数，其实标准正态分布还有一个很特殊的性质，方差为1

$$V[X] = E[X^2] = \square \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{\triangle}}\,dx = 1$$

最终可以得到标准正态分布的公式

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

通过平移和伸缩变化可以得到一般的正太分布

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

- 正态分布有几个性质
  1. 只要随机变量X的概率密度满足下面的要求：

     $$f_X(x) = const * e^{x\text{的二次式}},\ -\infty < x < +\infty$$

     就一定是正态分布。（形如$ax^2 + bx + c$都是二次式）
  2. 只要独立的随机变量X，Y中一个是正态分布，$W = X + Y$就满足正态分布，也就是“独立的正态分布经过加法运算仍然是正太分布”。

为什么正态分布这么重要呢?

还是因为现实世界中的很多事物都有这种“以某个值μ为中心，在这个中心点出现的概率比较高，离这个中心点越远的地方出现的概率越低”的特征，这就说名很多事物的都可以通过正态分布进行建模，进行数学处理了。至于为什么世界呈现这种形态，有一种说法是误差的扰动造成的。比如一个运动员投篮的过程，即使每次投篮的位置一样，但是每次的投篮的命中点却不可能一致。这是因为有很多微小的扰动，比如投篮的力度、姿势、出手的时机，这些都有不可控的扰动存在，使得命中点没有一个确定的位置，但是这些扰动是随机的，对于位置的偏移的影响也是随机的，这就是造成了投篮的位置在某一个位置附近按照正态分布的特点分布。

从投篮这件事情，我们可以引申到现实世界的很多事物，也可以帮助我们理解中心极限定理！！！！

假设这些影响结果的扰动设为一些符合$i.i.d$（也就是相互独立的，现实世界中往往是成立的）的随机变量$X_1, X_2, X_3 \ldots X_{n-1}, X_n$。这些误差是随机的，可能往左也可能往右偏，得到下面的结论

$$E[X_1] =\ldots= E[X_n] = 0$$
$$V[X_1] =\ldots= V[X_n] = \sigma^2 > 0$$

现在我们需要考虑的是这些随机变量$X_1, X_2, X_3 \ldots X_{n-1}, X_n$ 叠加起来会产生怎样的影响，尤其是n 足够大的时候。

单纯的叠加的只能得到

$[X_1 + X_2 + X_3 \ldots + X_n] = n\sigma^2 \to \infty$

没有讨论的意义。所以我们需要对分布进行标准化处理。

$$W_n = \frac{X_1 + X_2 + X_3 \ldots + X_n}{\sqrt{n}\sigma}$$

通过复杂的数学证明（可以自行查阅）可以知道W_n 服从正态分布。

最后我们能得到一个结论就是

任何微小的误差在大量的叠加之后会符合正态分布，无论原本的分布是怎么样的

协方差矩阵、多元正态分布与椭圆

1. 协方差

   先介绍两个随机变量X和Y 的协方差$Cov[X, Y]$ 的概念：

   $$Cov[X, Y] = E[(X - \mu)(Y - v)]$$

   协方差能够表达的是

   > - 协方差为正时：当一个变量增大时，另一个变量相应增大
   > - 协方差为负时：当一个变量增大时，另一个变量相应减小
   > - 协方差为0时：当一个变量增大时，另一个变量不会相应变化。

   如果协方差为正，表示两个随机变量正相关；为负，则表示负相关；为零，表示不相关。

2. 协方差的性质

   $$Cov[X, Y] = Cov[X + a, Y + b]$$

   - 根据协方差的定义很容易看出来，期望在协方差的定义中是减法，会消掉常数。

   $$abCov[X, Y] = Cov[aX, bY]$$

   - 按照定义也不难推导，X和它的期望值μ都是按同样的比例增大，得到的当然是可以提出公因数。特别的，$V[X] = Cov[X, X]$ ，这个性质其实和方差的一个性质有关就是$V[aX] = a^2V[X]$ 。
   - 如果两个随机变量独立，那么$Cov[X, Y] = 0$ ，也就是说独立性是不相关性的充分条件（独立一定不相关，不相关不一定独立）
   - $Cov[X, Y] = E[XY] - E[X]E[Y]$

3. 之前有提到过协方差能够表示随机变量的相关性，但是协方差的值能不能表明相关的程度呢？

这是一个值得考虑的问题，一开始看上去好像是可以有这种效果的，协方差的绝对值越大，相关程度越低。但是仔细一想，其实只要的给出一个例子就能知道$Cov[X,Y]=3.7, Cov[100X,100Y]=37000$ 显而易见我们只是同比例的扩大了随机变量，按理来说变量的相关程度并不会发生变化，但是协方差却不一致。

其实我们也很容易看到问题的所在，协方差距离表征相关程度只差了一个标准化的过程。

令

$$\hat{X}=\frac{X}{\sigma_X}, \hat{Y}=\frac{Y}{\sigma_Y}$$

$$\text{其中}\sigma_X=\sqrt{V[X]}, \sigma_Y=\sqrt{V[Y]}$$

$$\text{经过变换，变量的方差标准化为一}Cov[\hat{X},\hat{Y}]=\frac{Cov[X,Y]}{\sigma_X\sigma_Y}$$

$$\text{定义这个标准化的量作为相关系数}$$

$$\rho_{XY}=\frac{Cov[X,Y]}{\sigma_X\sigma_Y}=\frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}}$$

4. 相关系数的性质
    - 随机变量的相关性不受变量缩放系数的影响
    - 取值范围[-1,+1]
    - 相关系数越接近+1，（X,Y）就越接近一条左下右上方向的直线
    - 相关系数越接近-1，（X,Y）就越接近一条左上右下方向的直线
    - 如果X和Y独立，相关系数为0
5. 相关系数能够反映的东西也是有限的，不能盲目相信相关系数。

    比如即使两个变量的相关系数为0，两个变量还是存在关系，并不是独立的

    比如大学饭堂销量和总务处的失物招领成正相关，但这并不表示的这两个事物具有因果联系。很有可能只是因为学校放假了，两个变量都和放假相关而已。
6. 协方差矩阵

    前面我们知识界研究了两个变量的相关性，但是对于多变量，我们的可以采取列表的方式，展现两两之间的协方差值

	X_1	X_2	X_3
X_1	$Cov[X_1,X_1]$	$Cov[X_2,X_1]$	$Cov[X_3,X_2]$
X_2	$Cov[X_1,X_2]$	$Cov[X_2,X_2]$	$Cov[X_3,X_2]$
X_3	$Cov[X_1,X_3]$	$Cov[X_2,X_3]$	$Cov[X_3,X_3]$

    但是这样只能展现局部的特征，我们需要需要寻找一种能够展现整体联系的表现方式——那就是矩阵。

    n个变量间需要$n\times n$ 的矩阵表征相关性，比如X_1，X_2，X_3 需要下面的矩阵来表示

$$\begin{pmatrix} V[X_1] & Cov[X_2,X_1] & Cov[X_3,X_1] \\ Cov[X_1,X_2] & V[X_2] & Cov[X_3,X_2] \\ Cov[X_1,X_3] & Cov[X_2,X_3] & V[X_3] \end{pmatrix}$$

    协方差矩阵是一个对称矩阵。

    有了矩阵，我们就不需要每次都列写整个矩阵的内容个，而是可以矩阵进行计算了。

令

$$\text{随机变量的列向量}\mathbf{X}=\begin{pmatrix}X_1\\X_2\\X_3\\\vdots\\X_n\end{pmatrix}$$

$$E[\mathbf{X}]=\begin{pmatrix}E[X_1]\\E[X_2]\\E[X_3]\\\vdots\\E[X_n]\end{pmatrix}$$

$$\mu=E[\mathbf{X}]$$

现在我们可以通过矩阵运算表示协方差矩阵了

$$V[\mathbf{X}]=E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T]$$

其实可以说随机变量的向量的方差就是协方差矩阵

读者可以通过矩阵运算的定义计算验证这个表达式和协方差矩阵的定义是一致的。

通过矩阵运算实现协方差并不是只是节省纸墨，还有更深远的意义

7. 向量和矩阵运算的性质

$$E[ARB]=AE[R]B$$

其中A和B都是确定的矩阵值，R是随机变量的矩阵。

其实在实数中的很多结论都可以在矩阵中成立。对矩阵的操作都是对矩阵中每个元素的进行同样的操作。比如概率密度函数

$$f_{\mathbf{X}}(x)=f_{X_1,X_2,X_3\ldots X_n}(x_1,x_2,x_3\ldots x_n)$$

$$x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}$$

8. 协方差的变量替换

$$V[aX]=a^2V[X],a\text{是常数}$$

$$V[a^TX]=a^TV[X],\ a\text{是确定的列向量}$$

$$V[AX]=ATV[X]A^T,\ a\text{是确定的矩阵}$$

当然矩阵和向量的乘法都要是满足定义的

我们早就知道方差能够表征随机变量在期望附近的发散程度，在多变元的系统中，单个随机变量的大小$V[X_1]$还是可以表示在n维空间中沿着X_1 这个轴的发散程度。自然我们想到协方差可以表示的各个方向的发散程度。现在想像一朵概率密度之云飘散在n维的向量空间，这朵云有特异的形状，现在我们从任意一个方向飞来，我们知道这朵云的质心所在，想要测量这朵云在我么观测的方向分散程度，这就需要协方差矩阵进行计算了。