1 频率与概率

频率：

在相同情况下，进行了n次试验，在这n次试验中，事件A出现了m次，

则事件A在n次试验中出现的概率为：$f_n(A) = \frac{m}{n} = \frac{\text{事件}A\text{出现的次数}}{\text{试验的总次数}}$

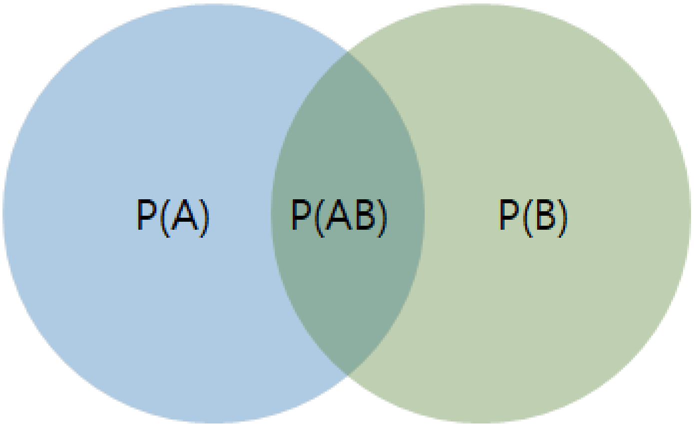
概率：

在大量重复试验中，若事件A发生的概率稳定地在某一常数p附近摆动，则称常数p为事件A发生的概率，即P(A)=p

2 条件概率

事件A发生的条件下，事件B发生的概率：$P(A|B)=\frac{P(AB)}{P(B)}$，同理：$P(B|A)=\frac{P(AB)}{P(A)}$

乘法公式： $P(AB)=P(A|B)P(B)=P(B|A)P(A)$

> 理解，画两个圆的相交
>
> 
>

3 事件的独立性：

一般情况：$P(AB)=P(A)P(B|A)$

独立事件：$P(AB)=P(A)P(B)$　　（事件A的发送与否 与 事件B的发生没有关系）

即：$P(A|B)=P(A)$

- 互斥事件：两事件的交集为空，A∩B=∅，P(AB)=0

4 全概率公式

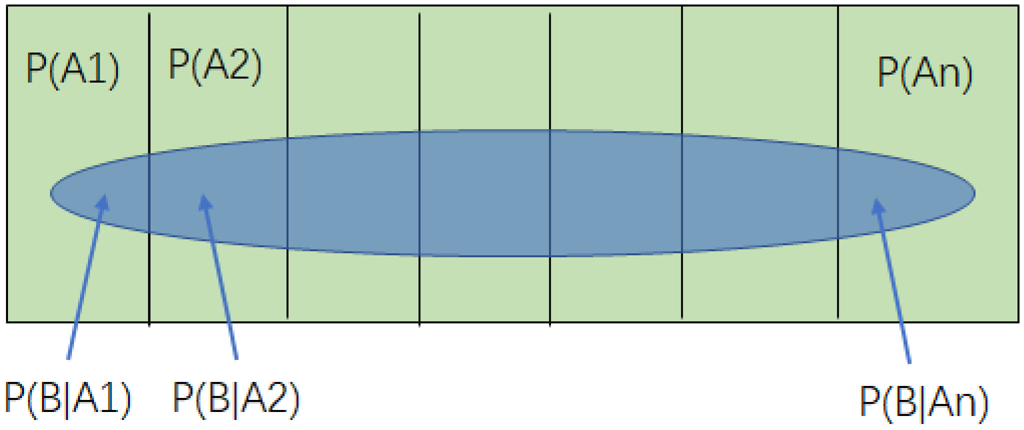
两两互斥的事件$A_1,A_2,...A_n$，$P(A_i)>0$，且$\sum_i^n A_i=\Omega$，则对任意事件B有：

$P(B)=\sum_i^n P(A_i)P(B|A_i)$

满足全概率公式中的事件组$A_1,A_2,...A_n$叫做 **完备事件组。**

```
全概率公式是“由因推果”的思想，当知道某件事的原因后，推断由某个原因导致这件事发生的概率为多少。
```

> 理解：



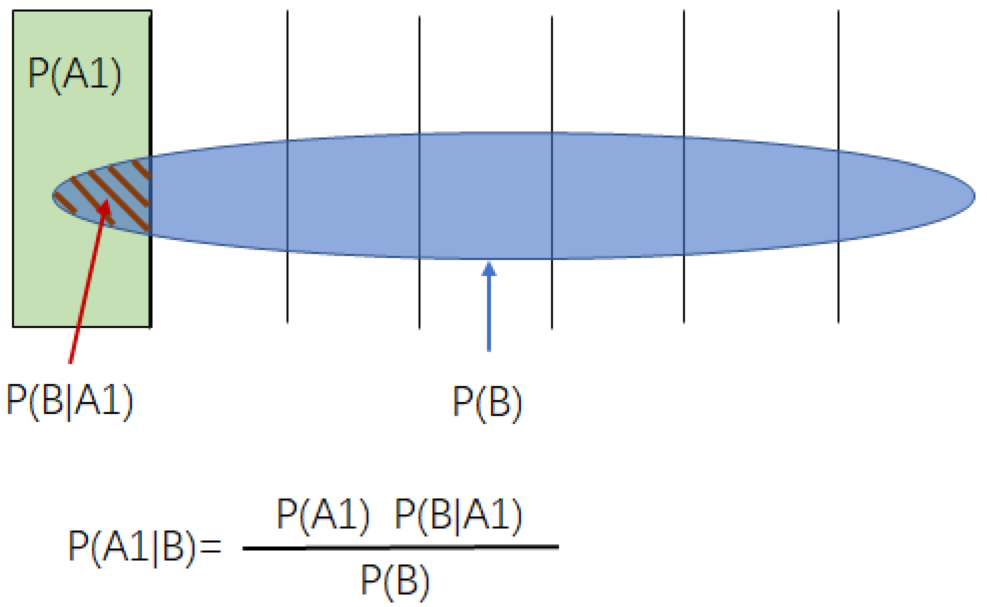
5 贝叶斯公式：

设$A_1,A_2,...A_n$构成完备事件组，则对任一事件B，有：

$$P(A_i|B)=\frac{P(A_iB)}{P(B)}=\frac{P(A_i)P(B|A_i)}{\sum_i^n P(A_i)P(B|A_i)}$$

> 贝叶斯公式是“由果溯因”的思想，当知道某件事的结果后，由结果推断这件事是由各个原因导致的概率为多少。

理解：



6 先验、后验概率

先验概率：

事情**未发生**，根据以往数据统计，分析事情发生的可能性，
如**全概率公式**，它往往作为"**由因求果**"问题中"因"出现的概率，
即：当知道某件事的原因后，推断由某个原因导致这件事发生的概率为多少。

后验概率：

　某件事**已经发生**，想要计算这件事发生的原因是由某个因素引起的概率，
　如**贝叶斯公式**，它往往作为“**由果溯因**”问题中“果”出现的概率，
即：当知道某件事的结果后，由结果推断这件事是由各个原因导致的概率为多少。

关系：

　已知先验概率，通过全概率公式可以求出 后验概率
　已知后验概率，通过贝叶斯公式可以求出 先验概率

7 离散随机变量分布

分布	描述	表达式	期望	方差
两点分布 X~B(1,p)	一次伯努利试验，只有两个结果0,1	$P(X=k)=p^kq^{(1-k)}$	p	$p(1-p)$
二项分布 X~B(n,p)	n重伯努利试验中，事件A出现的次数 (n较大、p较小，可近似为泊松分布,$\lambda=np$)	$P(X=k)=p^kq^{n-k}$	np	$np(1-p)$
泊松分布 X~P(λ)		$P(X=k)=\frac{\lambda^ke^{-k}}{k!}$	λ	λ
几何分布	有放回地抽取，首次试验成功所需做的试验次数 X	$P(X=k)=p(1-p)^k$	$\frac{1}{p}$	
超几何分布	抽出 n 个对象，成功抽出 k 次指定种类的对象的概率	$P(X=k)=\frac{C_M^kC_{N-M}^{n-k}}{C_N^n}$	$\frac{1-p}{p^2}$	

8 连续型随机变量分布

分布	概率密度函数	期望	方差
均匀分布 X~U(a,b)	$f(x)=\begin{cases}\frac{1}{b-a} & \text{a<x<b}\\0 & \text{else}\end{cases}$	$\frac{a+b}{2}$	$\frac{a-b}{12}$
指数分布 X~E(λ)	$f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{x>=0}\\0 & \text{else}\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
正态分布 X~N(μ,σ^2)	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

μ决定左右位置，σ决定分布的胖瘦，方差越小，PDF越瘦高（越集中）

9 概率密度函数（probability density function，PDF）

> 连续型随机变量，问结果出现在**某个点的概率**，等于0。
>
> 比如，问投硬币正面概率为0.7的概率，P(P(正面)=0.7)，结果是一个无穷小量，再比如，求投中靶上某一点的概率，结果也是一个无穷小量。
>
> 但是，我们可以把这个问题化为，求落在**某个范围内的概率**

PDF意义：

密度函数$f(x)$反映概率在某点x附近的"密集程度"。

一个随机变量出现在2个值区间的概率，等于 2个值之间曲线下方的面积，因此有：

$$P(a<X<b)=\int_a^b f(x)dx$$

(因为任一个点处的概率为0，这里符号加不加等号都可以)

注意：

- 一个事件概率为0，不一定是不可能事件
- 一个事件概率为1，不一定是必然事件

> 概率密度函数，可以看做是频数直方图的一种极限。
>
> - 每个小长方形的 面积=频率，高度=$\frac{Prob.}{\Delta x}$ (即概率密度)
> - 所有小长方形的面积之和=1
> - 组距$\Delta x \rightarrow 0$
>
> 而概率密度函数也满足：
>
> - $f(x)>=0$
> - $f(x)$ 下的面积恒为1

10 期望、方差

随机变量：

随机变量(Random Variable) X 是一个映射，把随机试验的结果与实数建立起了一一对应的关系。而期望与方差是随机变量的两个重要的数字特征。

数学期望/均值：

随机变量的期望：是试验中每次可能结果的概率乘以其结果的总和，它反映随机变量**平均取值的大小**。
样本均值（样本数目无穷多时，样本均值会无穷接近于数学期望，这是大数定律之一

方差：

随机变量的方差：是用来度量**随机变量偏离均值的程度。**
样本方差：是每个样本值与全体样本值的平均数之差的平方值的平均数。

11 协方差、相关系数

（1）协方差：

用于**衡量两个变量的总体误差**。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。

从数值来看，协方差的数值越大，两个变量同向程度也就越大。反之亦然。

$$Cov(X,Y)=E[(X-EX)(Y-EY)]=E(XY)-EXEY$$

如果两个变量的**变化趋势一致**，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。 如果两个变量的**变化趋势相反**，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。

（2）相关系数：

随机变量X和Y的相关系数： $\rho=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$

用 X、Y 的协方差除以 X 的标准差和 Y 的标准差。

相关系数也可以看成协方差：一种剔除了两个变量量纲影响、标准化后的特殊协方差。

它消除了两个变量变化幅度的影响，而只是单纯反应两个变量每单位变化时的相似程度。

> **两个因素会影响协方差的值：**
>
> 1、两个变量各自的方差不变的情况下，两个变量的正相关性越强烈，协方差越大，负相关性越强烈，协方差越小；
>
> 2、两个变量的相关性不变的情况下，x或y变量的方差越大，协方差的绝对值越大。(“或”的意思是，x的方差大，或者y的大，或者它俩的都大);
>
> 因素1对协方差的影响是“绝对”大小（带符号），因素2影响的是“绝对值”的大小

> **反过来的推论: 如果协方差的值是个很大的正数，我们可以得到两个结论：**
>
> （1） 两者有很大概率是正相关的；
>
> （2） 这个值很大到底是因为①：正相关很强烈造成的呢？还是②：x或y的方差很大造成的呢，这个①和②我们是区分不出来的

> 那么如何衡量正负相关性呢，显然要**把x或y的方差，从对协方差的影响中剔除掉**，这样协方差剩余的部分就能看出相关性的强烈程度了。剔除的方法也很简单，协方差除以xy的标准差就行了。得出的结果就被成为相关系数：$\rho=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$

> 上面讲的是两个变量之间的协方差，如果有n个变量X1、X2、…Xn，两两之间的协方差，就可以组成**协方差矩阵**，我们定义：
>
> $$\vec{X}=\begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_3 \end{bmatrix}$$
>
>
>
> 那么上述n个变量的协方差矩阵就是：

> $$P_X = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \cdots & \cdots & \sigma_{nn} \end{bmatrix}$$
>
>
>
> 是**半正定矩阵**

（3）不相关与独立

不相关：指没有线性关系。$Cov(X,Y)=0$

独立：指没有任何关系，包括线性、非线性关系等各种关系。$f(x,y)=f(x)f(y)$

- 独立 则一定 不相关，不相关 不一定 独立。

> **相关系数或协方差为 0 的时候能否说明两个分布无关？为什么？**
>
> 只能说明不线性相关，不能说明无关。因为在数学期望存在的情况下，独立必不相关，不相关未必独立。

（4）其他

- $D(X \pm Y)=D(X)+D(Y) \pm 2Cov(X,Y)$ 。当X与Y相互独立时，一定有Cov(X,Y)=0
- $|\rho_{XY}|=1$的充要条件：P(Y=aX+b)=1

12 若干正态分布相加、相乘后得到的分布分别是什么？

相加：（独立的前提下）都服从正态分布。

相乘：

结论：两个分别服从 $N(\mu_1,\sigma_1^2)$ 和 $N(\mu_2,\sigma_2^2)$ 的正态分布的概率密度函数相乘后，新函数等价于正态分布 $N(\mu_0,\sigma_0^2)$ 的概率密度函数乘以缩放因子 A 。其中，缩放因子

$A=\dfrac{e^{-\frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}}}{\sqrt{2\pi\left(\sigma_1^2+\sigma_2^2\right)}}$ ，正态分布的均值 $\mu_0=\dfrac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}$ ，正态分布的方差 $\sigma_0^2=\dfrac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}$ 。

13 独立与互斥的关系

不是一个层面上的问题：

事件 A 与事件 B 独立的定义是： $P(AB)=P(A)P(B)$ 。

事件 A 与事件 B 互斥的定义是：集合 A 与集合 B 没有相同的样本点，即 $A\cap B=\phi$ 。（ ϕ 代表空集， $A\cap B$ 也可以简记为 AB ）

如果事件 A 或事件 B 发生的概率**都不为0**，那么独立和互斥有这样一层关系：**互斥不独立，独立不互斥。**

14 切比雪夫不等式

设随机变量X的数学期望EX与方差DX存在，则对任意的$\epsilon>0$，有：$P(|X-EX|\geq\epsilon)\leq\frac{DX}{\epsilon^2}$

当给定误差ϵ 时，可以估计X偏离期望的概率。

15 大数定律——大量随机试验的样本均值

（阐明大量随机现象平均结果的稳定性）

则对任意正数 $\boldsymbol{\epsilon}$ >0，伯努利大数定律表明：

$$\lim_{n\to\infty}P\left\{\left|\frac{n_x}{n}-p\right|<\varepsilon\right\}=1$$

样本数量很大的时候，**样本均值和数学期望充分接近**。也就是说当我们大量重复某一相同的实验的时候，其最后的实验结果可能会稳定在某一数值附近。

16 中心极限定理——大量随机试验的样本均值的分布

- 大量（n→∞）、独立、同分布的随机变量之和，近似服从于一维正态分布。

定义：设X1、X2…是具有相同分布、相互独立的一系列随机变量，$E(X_i)=\mu,D(X_i)=\sigma^2$

则近似有：$\sum_i^n X_i\sim N(n\mu,n\sigma^2)$

17. 大数定律和中心极限定理的区别

前者更关注的是**样本均值**，后者关注的是**样本均值的分布**，比如说掷色子吧，假设一轮掷色子 n 次，重复了 m 轮，当 n 足够大，大数定律指出这 n 次的均值等于随机变量的数学期望，而中心极限定理指出这 m 轮的均值分布符合围绕数学期望的正态分布。

19. 最大似然估计（极大似然估计）是什么?

极大似然估计就是一种参数估计方法。

最大似然估计的目的是：**利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值。**

原理：极大似然估计是建立在极大似然原理的基础上的一个统计方法，是概率论在统计学中的应用。极大似然估计提供了一种给定观察数据来评估模型参数的方法，即：“**模型已定，参数未知**”。通过若干次试验，观察其结果，利用试验结果得到某个参数值能够使样本出现的概率为最大，则称为极大似然估计。

方程的解只是一个估计值，只有在样本数趋于无限多的时候，它才会接近于真实值。

- 求最大似然估计量 $\hat{\theta}$ 的一般步骤：

```
[1]    写出似然函数；
[2]    对似然函数取对数，并整理；
[3]    求导数；
[4]    解似然方程。
```

- 最大似然估计的特点：

```
[1]    比其他估计方法更加简单；
[2]    收敛性：无偏或者渐近无偏，当样本数目增加时，收敛性质会更好；
[3]    如果假设的类条件概率模型正确，则通常能获得较好的结果。但如果假设模型出现偏差，将导致非常差的
```

离散型随机变量的最大似然估计

离散型随机变量 X 的分布律为 $P\{X=x\}=p(x;\theta)$，设 $X_1,\cdots,X_n$ 为来自 X 的样本，$x_1,\cdots,x_n$ 为相应的观察值，θ 为待估参数。

在参数 θ 下，分布函数随机取到$x_1,\cdots,x_n$ 的概率为

$$p(x|\theta)=\prod_{i=1}^{n}p(x_i;\theta)$$

构造似然函数：

$$L(\theta|x)=p(x|\theta)=\prod_{i=1}^{n}p(x_i;\theta)$$

可知似然函数是一个关于 θ 的函数，要找到最大概率生成 x 的参数，即找到当 $L(\theta|x)$ 取最大值时的 θ。

求解出最大值，通常的方法就是求导=0：

$$\frac{d}{d\theta}L(\theta|x)=0$$

由于式子通常是累乘的形式，我们借助对数函数来简化问题：

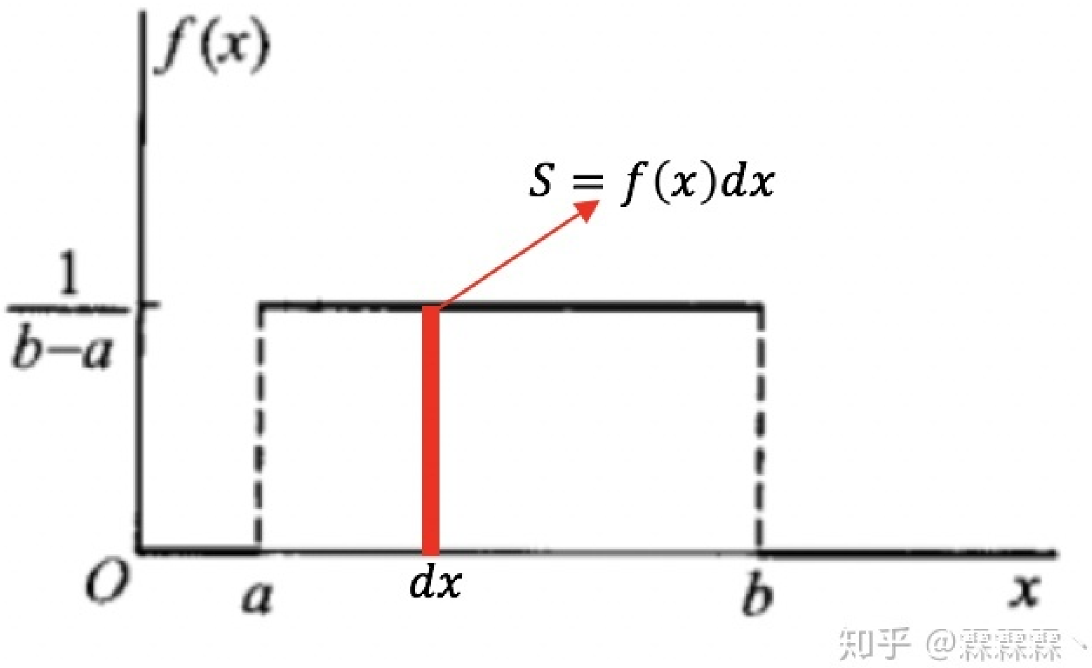
$$\frac{d}{d\theta}lnL(\theta|x)=0$$

上式也通常被称作**对数似然方程**。如果 θ 包含多个参数 $\theta_1,\cdots,\theta_k$，可对多个参数分别求偏导来连立方程组。

连续型随机变量的最大似然估计

连续型随机变量 X 的概率密度为 $f(x;\theta)$，设 $X_1,\cdots,X_n$ 为来自 X 的样本，$x_1,\cdots,x_n$ 为相应的观察值，同样地，θ 为待估参数。

概率密度的图像与横轴所围成的面积大小代表了概率的大小，当随机变量 X 取到了某一个值 x_1，可看做是选取到了 $f(x_1;\theta)$ 与 dx 所围成的小矩形。如图所示：



接着与离散型随机变量类似，随机取到观察值 x 的概率为：

$$p(x;\theta)=\prod_{i=1}^{n} f(x_i;\theta)dx$$

构造似然函数：

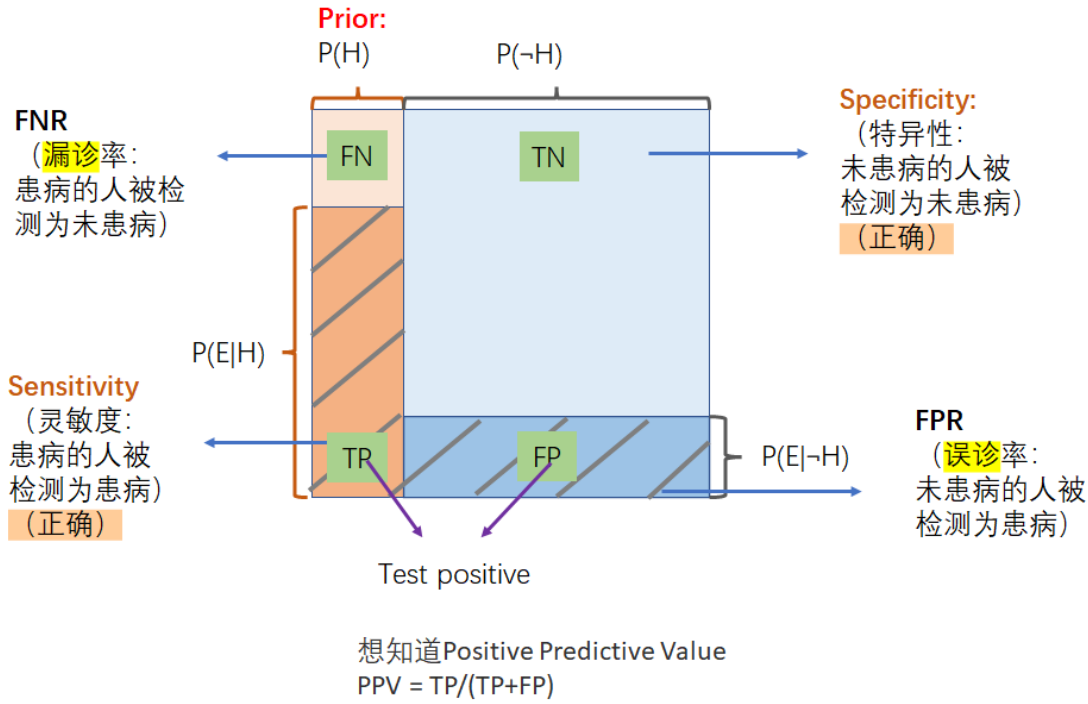
$$L(\theta|x)=\prod_{i=1}^{n} f(x_i;\theta)dx$$

由于 $\prod_{i=1}^{n} dx$ 不随参数变化，故我们选择忽略，似然函数变为：

$$L(\theta|x)=\prod_{i=1}^{n} f(x_i;\theta)$$

接着计算步骤和离散型类似，取对数求导等于0。

补充：

> **用 先验 去修正 后验**
>
> 这里仅以0、1表示结果的两种情况，分别对应Negative、Positive，以患病with/without cancer举例：
>
> 在试验前，知道患病率P(1)=95%，知道灵敏度Sensitivity=90%，特异性Specificity=91%
>
> 经过一次试验后，假如测得有cancer，想知道真正有cancer的概率。
>
> 
>
>
> $PPV = \frac{TP}{TP+FP} = \frac{prior*Sens.}{prior*Sens.+(1-prior)*FPR}$
>
> 参考：3Blue1Brown 概率系列视频