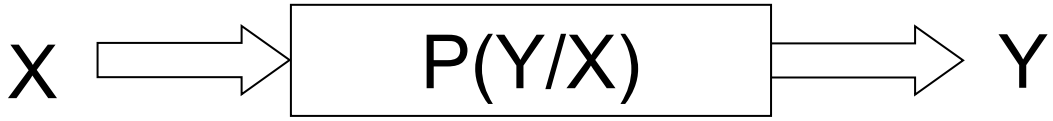
脑科学和类脑

7 信息熵和贝叶斯定理

于玉国
yuyuguo@fudan. edu. cn
计算神经生物学实验室
复旦大学智能复杂体系基础理论与关键技术实验室

P(推测？ |观察）



对于大脑，信息是什么？

大脑对世界的认知是基于精确求解还是概率推断？

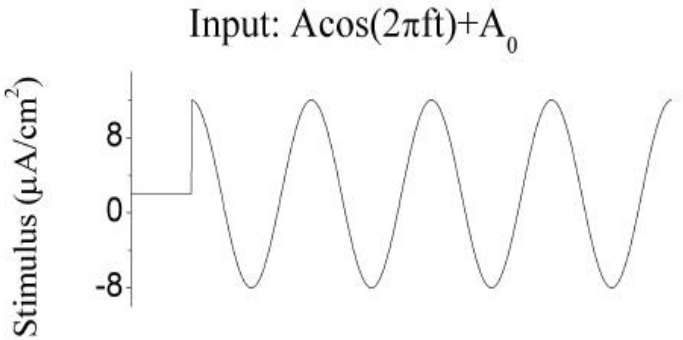
信息（information）是什么?

如何计算？


普通高等教育“十一五”国家级规划教材
概率论与数理统计
第四版
浙江大学 盛 骤 谢式千 潘承毅 编

信号

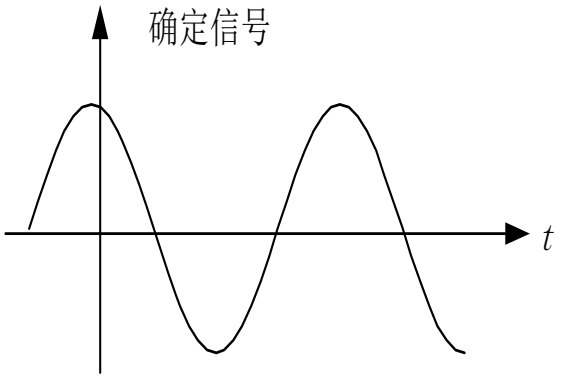
- 物理： 信号是携带信息的一种物理变化
- 信息论：可用数学函数表示的一种信息流。
- 数学： 信号是一个或多个变量的函数或序列
- 自变量：时间、位移、周期、频率、幅度、相位

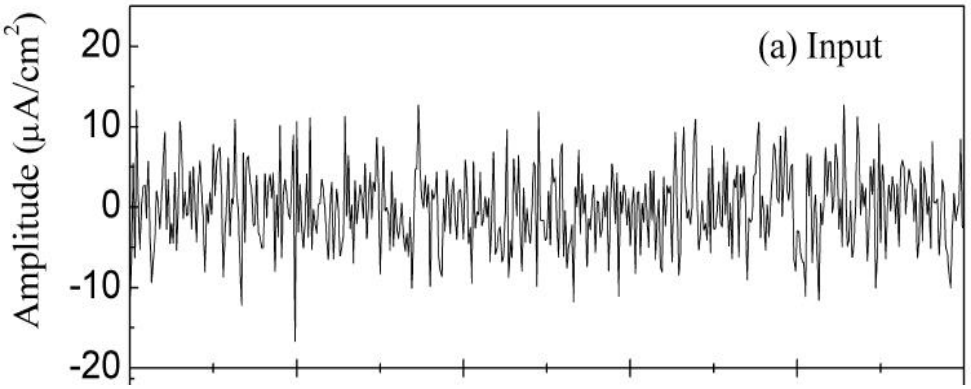


信号的分类

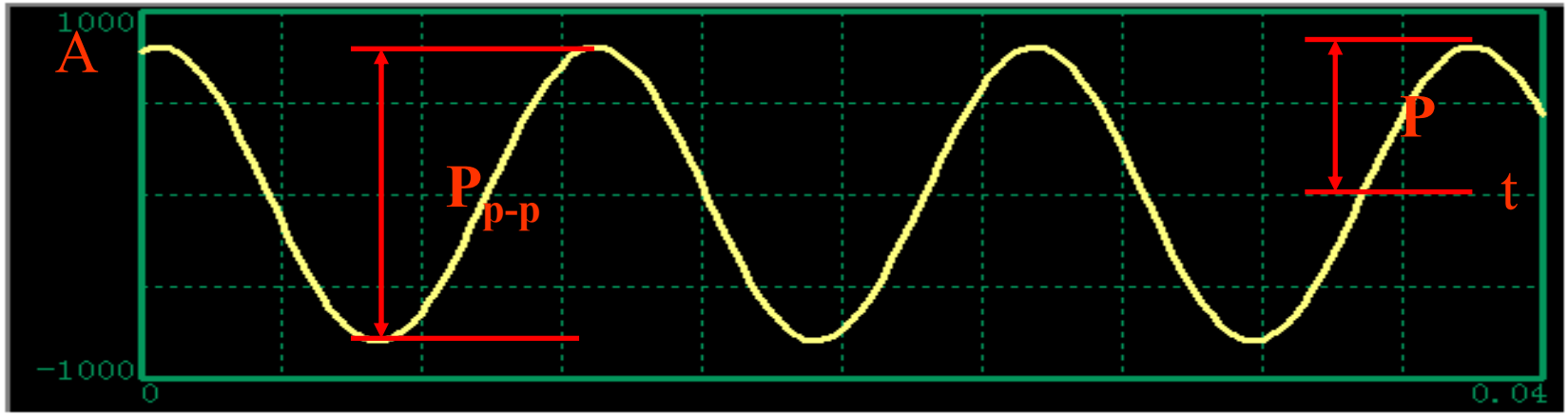
- 确定信号：信号可以用一个确定的时间函数（或序列）表示。**Y=Sin(2pi*f*t)**

- 随机信号：信号在任意时刻由于某些“不确定性”或“不可预知性”的因素而造成信号无法用一个确定的时间函数（或序列）表示。





1 峰值



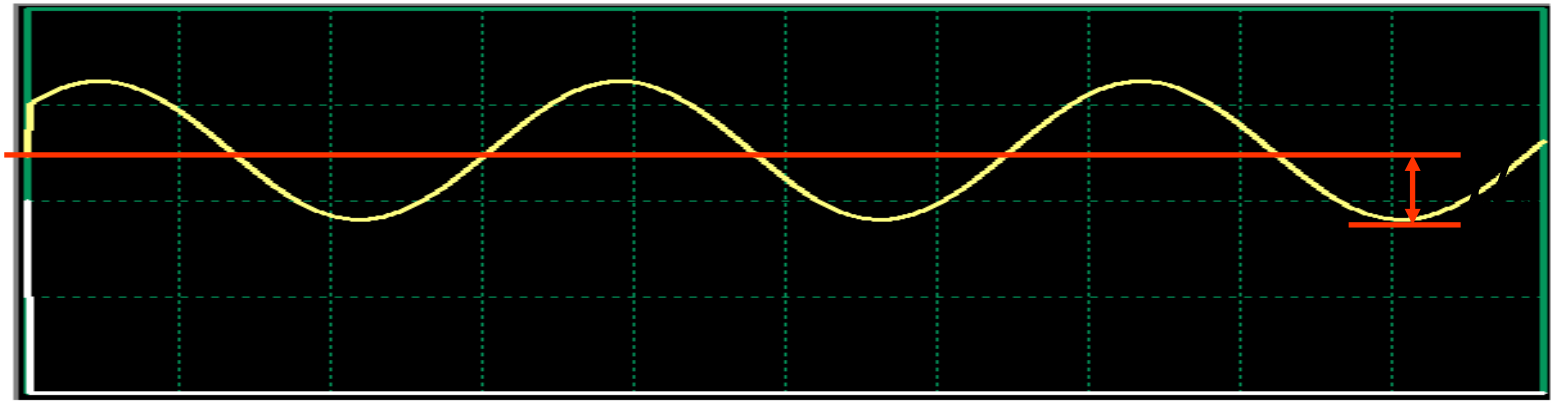
- 峰值(peak value)P----信号的最大瞬时强度
- 双峰值(double peak value)P_{p-p}

2 均值(mean value)

- 均值$E[x(t)]$表示集合平均值或数学期望值。

$$\mu_x = E[x(t)] = \lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt$$

$$\mu_x = E[x(t)] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i$$



3 均方值(mean square value)

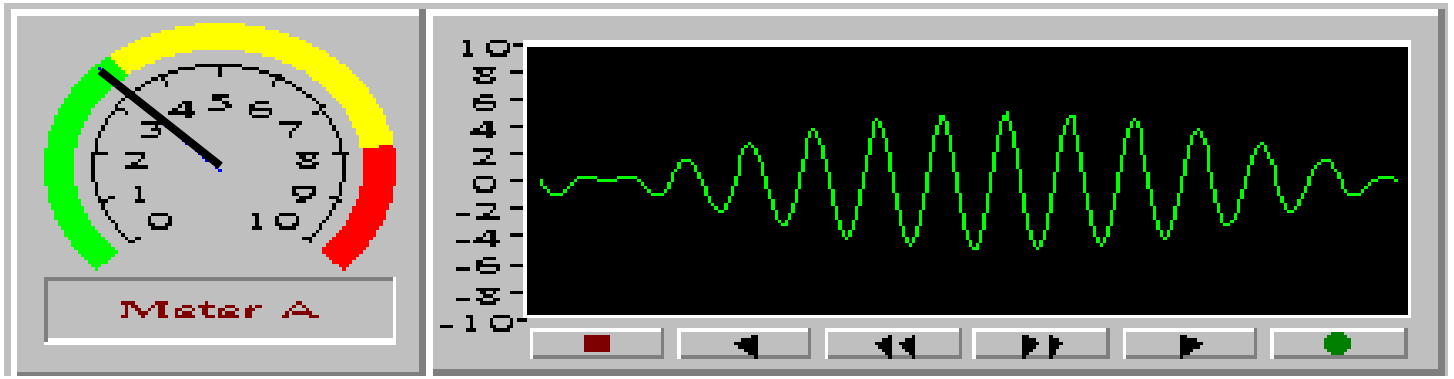
- 信号的均方值$E[x^2(t)]$，表达了信号的强度

$$\psi_x^2 = E[x^2(t)] = \lim_{T\to\infty}\frac{1}{T}\int_0^T x^2(t)dt$$

$$\psi_x^2 = E[x^2(t)] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i^2$$

有效值(RMS)：均方值的正平方根值，相当于信号平均功耗

$$X_{r.m.s.} = \sqrt{\psi_x^2} = \psi_x$$



信号有效值　　信号波形

3 均方值(mean square value)

- 物理意义：信号的平均功率
- 借鉴电学功率：单位时间的能耗

$$P = \frac{V^2}{R} = I^2 R$$

$$p = \frac{1}{T}\int_0^T v^2(t)/R dt = \frac{1}{T}\int_0^T i^2(t) R dt$$

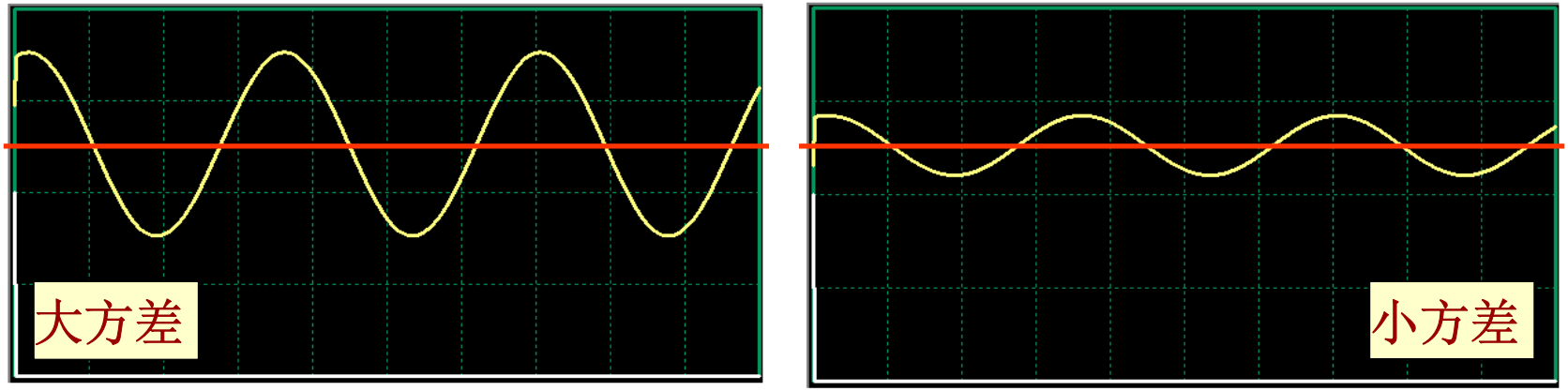
$$= \frac{1}{T}\int_0^T v^2(t) dt = \frac{1}{T}\int_0^T i^2(t) dt \quad (R=1)$$

4 方差(variance): 量化信号围绕均值的离散程度

- 方差概念的提出
  - 欲了解信号相对于均值的离散程度 $x(t)-\mu_x$
- 对任意t，信号偏离的量为 $[x(t)-\mu_x]^2$
- 定义数据的总体方差为：

$$\sigma_x^{\,2} = E[(x(t)-E[x(t)])^2] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}[x_i-\mu_x]^2$$

$$\sigma_x = \sqrt{\sigma_x^{\,2}}$$ （标准方差 standard deviation)



4 方差(variance)

- 方差的物理意义
  - 信号的纯波动分量
  - 去除直流分量后，信号的平均功率

- 信号总功率＝交流量功率＋直流量功率

$$\psi_x^2 = \sigma_x^2 + \mu_x^2$$

样本方差

样本方差：$s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

样本方差可以理解成是对所给总体方差的一个估计。

母体方差：$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

也叫总体方差

样本标准差：$s = \sqrt{s^2} = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

母体标准差：$\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

为什么样本方差的分母是$n-1$，而总体方差的分母却是n？

为什么样本方差的分母是$n-1$，而总体方差的分母却是n？这里贡献一个简单的启发式想法。

容量为n的样本方差公式：

$$s^2=\frac{\sum(x-\bar{x})}{n-1}$$

个数为N的总体方差的公式：

$$\delta^2=\frac{\sum(X-\bar{X})}{N}$$

对于样本方差来说，假如从总体中只取一个样本，即$n=1$，那么样本方差公式的分子分母都为0——方差完全不确定。这很好理解，因为样本方差是用来估计总体中个体之间的变化大小，只拿到一个个体，当然完全看不出变化大小。反之，如果公式的分母不是$n-1$而是n，计算出的方差就是0——这是不合理的，因为不能只看到一个个体就断定总体的个体之间变化大小为0。

对于总体方差来说，假如总体中只有一个个体，即N=1，那么方差，即个体的变化，当然是0。如果分母是N-1，总体方差为0/0，即不确定，却是不合理的——总体方差不存在不确定的情况。

由此，样本方差分母为$n-1$，总体方差分母为n，才是合理的。

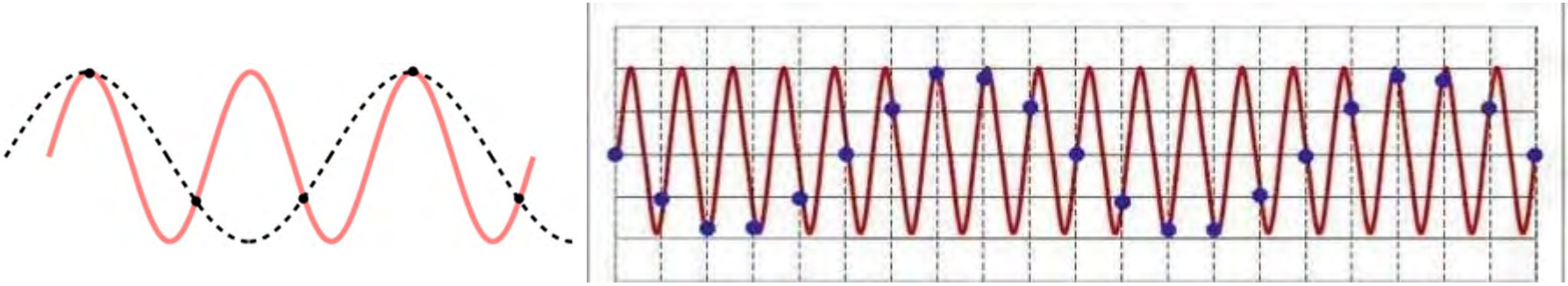
采样频率以及采样定理

1.采样频率：定义了每秒从实际给定的连续信号中提取并组成离散信号的采样点数，它用赫兹（Hz）来表示。采样频率的倒数是采样时间，是采样点之间的时间间隔。通俗的讲采样频率是指计算机每秒钟采集多少个信号点。

2.采样定理：又称香农采样定理(或奈奎斯特采样定理），如果信号含有的最高频率为f_{max}，那么采样频率f_s应至少高于信号中最高频率f_{max}的2倍，这样，原来的连续信号可以从采样数据中基本重建出来,即 $f_s > 2\ f_{max}$

实际应用中采样频率往往应大于信号最高频率的**2.56～4倍**，防止数字信号失真。

3. 采样点越密集，那越接近信号波原始的样子，损失信息越少，越方便还原信号。



不能上述采样条件，采样后信号的频率就会重叠，即高于采样频率一半的频率成分将被重建成低于采样频率一半的信号。这种频谱的重叠导致的失真称为混叠

标准正交基

欧氏空间中，单位向量 $\vec{e}_x$，$\vec{e}_y$ 满足如下内积关系式

$$\vec{e}_x \bullet \vec{e}_x = 1$$
$$\vec{e}_y \bullet \vec{e}_y = 1$$
$$\vec{e}_x \bullet \vec{e}_y = e_x * e_y * \cos(90^\circ) = 0$$

则称这2个单位向量构成标准正交基.
X-Y平面空间的任何一个向量都可化解为一个x方向和一个y方向的向量的矢量加合。

傅立叶变换

“周期信号都可表示为一系列满足正交关系的正弦和余弦信号的加权和”；“非周期信号也都可用正弦信号的加权积分表示”

$$f_T(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos n\omega_1 t + b_n \sin n\omega_1 t)$$

直流分量　　基波分量 n =1　　谐波分量 n>1

$$\omega_1 = \frac{2\pi}{T_1} \qquad \omega_n = n\omega_1$$

Jean Baptiste Joseph Fourier
1768-1830

定理 1. 组成三角级数的函数系

$$1, \cos x, \sin x, \cos 2x, \sin 2x, \cdots, \cos nx, \sin nx, \cdots$$

在 $[-\pi, \pi]$ 上正交，即其中任意两个不同的函数之积在 $[-\pi, \pi]$ 上的积分等于 0.

证： $\int_{-\pi}^{\pi} 1 \cdot \cos nx \, \mathrm{d}x = \int_{-\pi}^{\pi} 1 \cdot \sin nx \, \mathrm{d}x = 0 \quad (n = 1, 2, \cdots)$

$$\int_{-\pi}^{\pi} \cos kx \cos nx \, \mathrm{d}x$$

$$= \frac{1}{2} \int_{-\pi}^{\pi} [\cos(k+n)x + \cos(k-n)x] \, \mathrm{d}x = 0 \ (k \neq n)$$

同理可证：$\int_{-\pi}^{\pi} \sin kx \sin nx \, \mathrm{d}x = 0 \quad (k \neq n)$

$$\int_{-\pi}^{\pi} \cos kx \sin nx \, \mathrm{d}x = 0$$

直流系数 $$a_0 = \frac{1}{T_1}\int_{t_0}^{t_0+T_1} f(t).dt$$

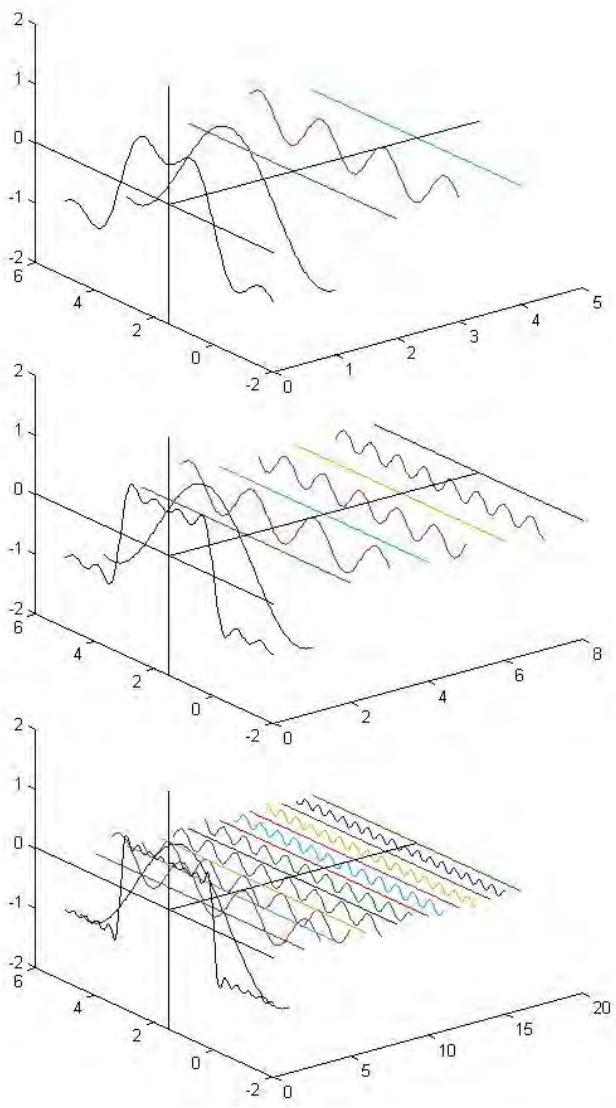
余弦分量系数 $$a_n = \frac{2}{T_1}\int_{t_0}^{t_0+T_1} f(t).\cos n\omega_1 t.dt$$

正弦分量系数 $$b_n = \frac{2}{T_1}\int_{t_0}^{t_0+T_1} f(t).\sin n\omega_1 t.dt$$

离散振幅频谱： $$\left|F_n(\omega)\right| = \frac{1}{2}\sqrt{a_n^2 + b_n^2}$$

$F_n(\omega)$称为信号x(t)的频谱它反映了信号f(t)中各种频率成分的分布状况，f(t)可以表示成谐分量的无限叠加。可证明：对一般实信号x(t)，其频谱是ω的复函数

在欧几里得空间当中，可以通过选取一组正交基，使得空间内的所有向量都可以由这组正交基线性表出。



$f_T(t)$是以T为周期的函数

$$f_T(t)=\frac{a_0}{2}+\sum_{n=1}^{\infty}(a_n\cos n\omega t+b_n\sin n\omega t)$$

$$\omega=2\pi/T,$$

$$a_n=\frac{2}{T}\int_{-T/2}^{T/2}f_T(t)\cos n\omega t dt \qquad (n=0,1,2,\cdots)$$

$$b_n=\frac{2}{T}\int_{-T/2}^{T/2}f_T(t)\sin n\omega t dt \qquad (n=1,2,\cdots)$$

引入欧拉公式

$$\mathbf{e^{ix}=\cos x+i\sin x,}\qquad -\infty<x<+\infty.$$

$$f_T(t)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(a_n\frac{e^{in\omega t}+e^{-in\omega t}}{2}+b_n\frac{e^{in\omega t}-e^{-in\omega t}}{2i}\right)$$

$$=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(\frac{a_n-ib_n}{2}e^{in\omega t}+\frac{a_n+ib_n}{2}e^{-in\omega t}\right)$$

$$f(t)=\sum_{n=-\infty}^{\infty}F(n\omega_1)e^{jn\omega_1 t}$$

周期函数的复指数级数

- 由前知

$$f_1(t) = a_0 + \sum_{n=1}^{\infty}(a_n \cos n\omega_1 t + b_n \sin n\omega_1 t)$$

- 由欧拉公式

$$f(t) = \sum_{n=-\infty}^{\infty} F(n\omega_1) e^{jn\omega_1 t}$$

- 其中

$$F(0) = a_0 \qquad F(n\omega_1) = \frac{1}{2}(a_n - jb_n)$$

引入了负频率

$$F(-n\omega_1) = \frac{1}{2}(a_n + jb_n)$$

欧拉公式 $\mathbf{e^{ix} = \cos x + i \sin x,} \quad -\infty < x < +\infty.$

指数形式的傅里叶级数的系数

$$F(n\omega_1)=F_n$$

$$F_n=\frac{1}{T_1}\int_{t_0}^{t_0+T_1} f(t)e^{-jn\omega_1 t}dt$$

两种傅氏级数的系数间的关系

$$F_0=c_0=d_0=a_0$$

$$F_n=\left|F_n\right|e^{j\varphi_n}=\frac{1}{2}(a_n-jb_n)$$

$$F_{-n}=\left|F_{-n}\right|e^{-j\varphi_n}=\frac{1}{2}(a_n+jb_n)$$

对任何一个非周期函数$f(t)$都可以看成是由某个周期函数$f_T(t)$当$T\to\infty$时转化而来的.

作周期为T的函数$f_T(t)$，使其在$[-T/2,T/2]$之内等于$f(t)$，在$[-T/2,T/2]$之外按周期T延拓到整个数轴上，则T越大，$f_T(t)$与$f(t)$相等的范围也越大，这就说明当$T\to\infty$时，周期函数$f_T(t)$便可转化为$f(t)$，即有

$$\lim_{T\to+\infty} f_T(t)=f(t)$$

两种傅氏级数的系数间的关系

$$|F_n| = |F_{-n}| = \frac{1}{2}c_n = \frac{1}{2}d_n = \frac{1}{2}\sqrt{a_n^2 + b_n^2}$$

$$|F_n| + |F_{-n}| = c_n$$

$$F_n + F_{-n} = a_n$$

$$j(F_n - F_{-n}) = b_n$$

$$c_n^2 = d_n^2 = a_n^2 + b_n^2 = 4F_nF_{-n}$$

能量谱密度和功率谱密度

$$\int_{-\infty}^{\infty}[x(t)]^2dt=\int_{-\infty}^{\infty}x(t)\cdot\frac{1}{2\pi}\int_{-\infty}^{\infty}F_x(\omega)e^{j\omega t}d\omega dt$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}F_x(\omega)\cdot\int_{-\infty}^{\infty}x(t)e^{j\omega t}dtd\omega$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}F_x(\omega)F_x(-\omega)d\omega dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}F_x(\omega)F_x^*(\omega)d\omega dt$$

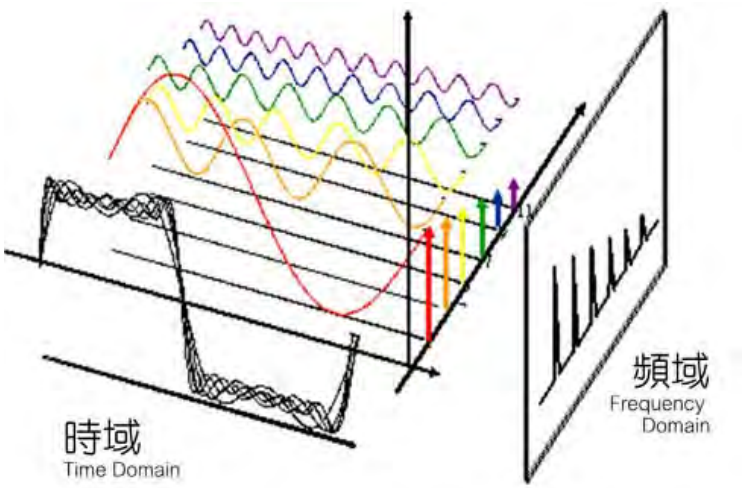
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}|F_x(\omega)|^2\,d\omega$$ ←帕赛瓦尔定理

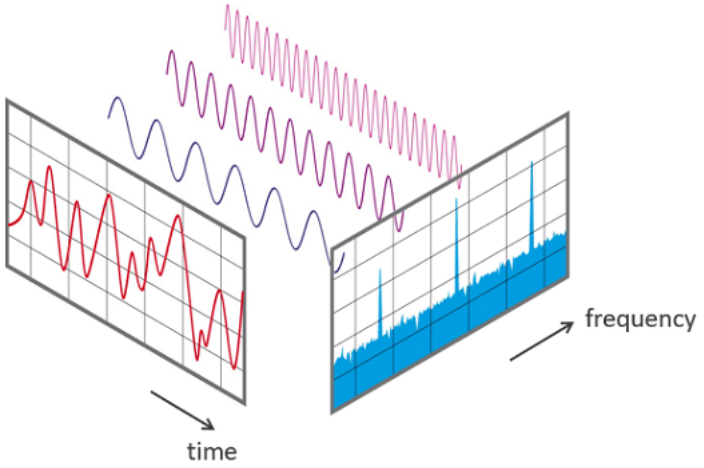
由于左边是：x(t)在时间(-∞,∞)上的总能量＝$|F_x(w)|^2$ 在整个频域上的积分。
因此 $|F_x(w)|^2$ 表示 x(t) 在不同频率上总能量的分布密度。

能量谱密度：$\varepsilon(\omega)=|F(\omega)|^2$

功率谱密度：$S(\omega)=\lim_{T\to\infty}\frac{|F_T(\omega)|^2}{T}$

频谱分析





绝大部分时域或空间上的信号都可以分解成不同频率的许多正弦波成分，
时域到频域就是将一个随时间变化的动态信号分解成各个静态的不同频率正弦信号

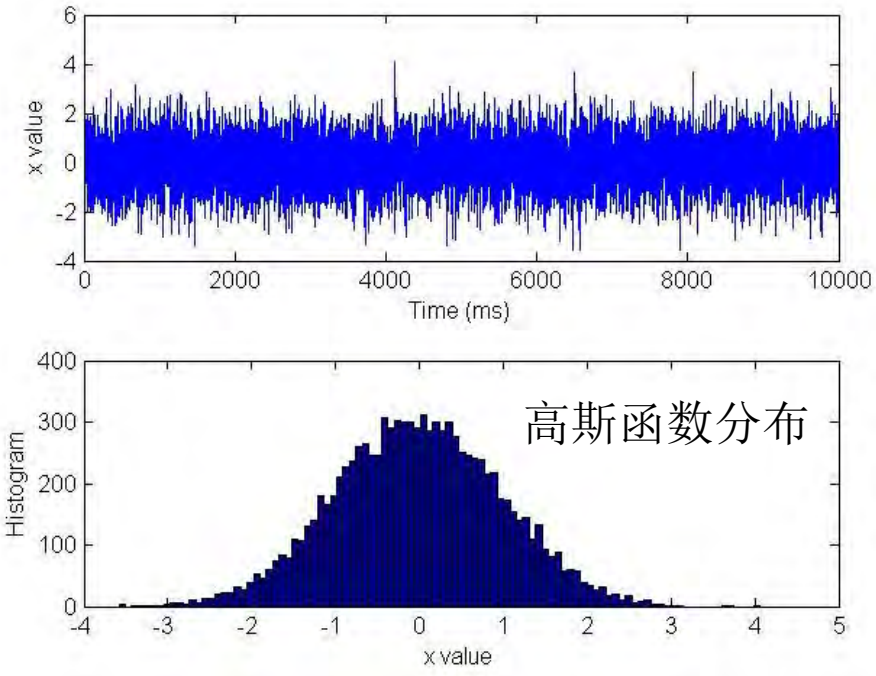
高斯白噪声 Gaussian White Noise

高斯**白噪声**：一个噪声序列，每个值彼此相互独立，**任意时刻出现的噪声幅值都是随机的**，但每个值大小的概率呈现高斯分布，且所有频率成分的出现是均匀分布的，

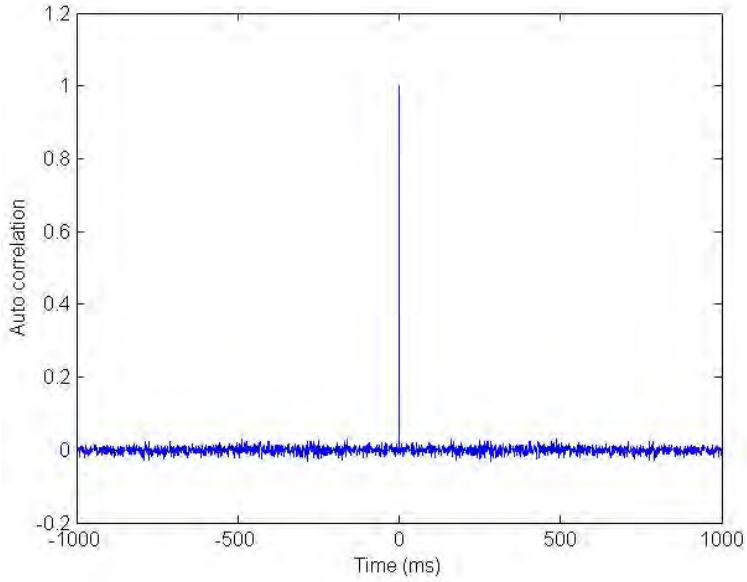
则称它为高斯白噪声。均值m,标准差为 σ

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-m)^2/2\sigma^2}$$

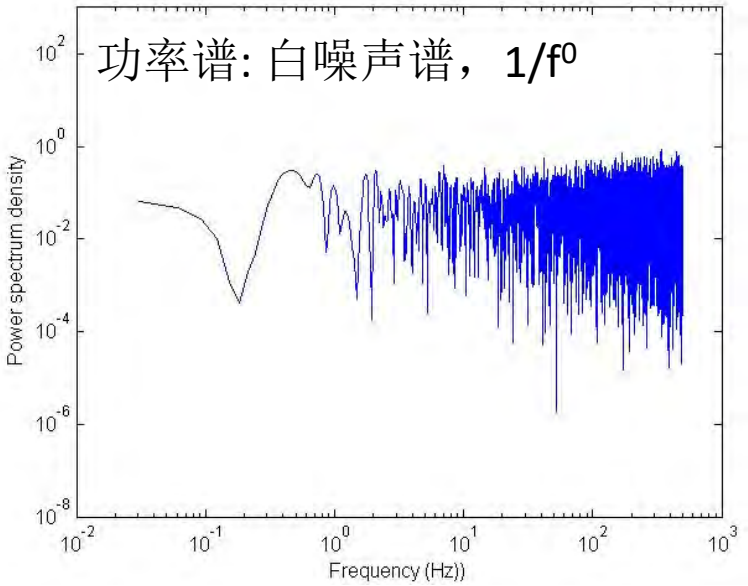
概率分布函数



时间空间，没有相关性

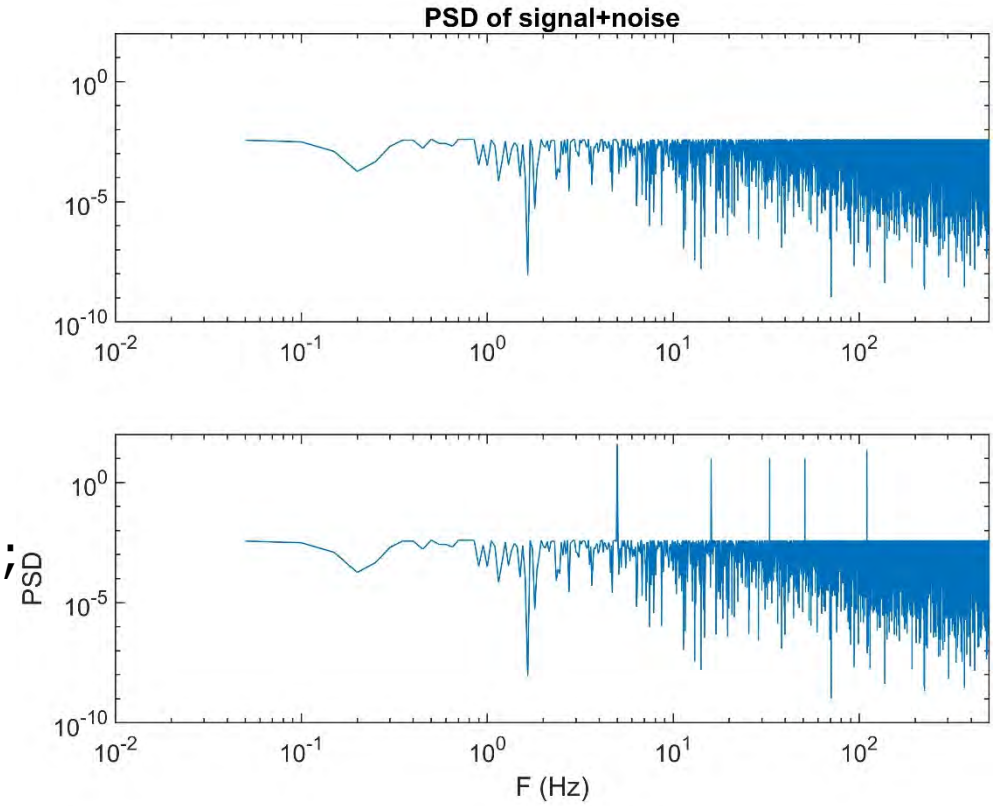


频率空间：平谱



Matlab 求周期信号的功率谱程序

```
clear;
dt=1; t=1:dt:20000;
F0=1000/dt; nx=128; %32768;
f=33;%signal frequency
s1=sin(2.*pi.*t.*0.001*f);
%0.001 scale factor to change
the default 1 sec to be 1 ms
s2=sin(2.*pi.*t.*0.001*16);
s3=sin(2.*pi.*t.*0.001*5)*2;
s4=sin(2.*pi.*t.*0.001*51);
s5=sin(2.*pi.*t.*0.001*110)*1.5;
s=s1+s2+s3+s4+s5;
n1=noise(0,length(s));n1=n1';
s=s+n1;
```




```
[psn,fn] = periodogram(n1,rectwin(length(s)),length(s),F0);
[psx,fs] = periodogram(s,rectwin(length(s)),length(s),F0);
figure;subplot(2,1,1),loglog(fn,psn)
 xlim([1e-2 5e2]);ylim([1e-10 1e2])
subplot(2,1,2);loglog(fs,psx)
 xlim([1e-2 5e2]);ylim([1e-10 1e2])
 xlabel('F (Hz)'); ylabel('PSD')
```

例 1 求矩形脉冲函数 $f(t)=\begin{cases}1, & |t|\le 1\\ 0, & |t|>1\end{cases}$ 的付氏变换及其

积分表达式。

$$F(\omega)=\int_{-\infty}^{+\infty} f(t)e^{-i\omega t}dt=\int_{-1}^{1} e^{-i\omega t}dt=\frac{e^{-i\omega t}}{-i\omega}\Big|_{-1}^{1}$$

$$=-\frac{1}{i\omega}(e^{-i\omega}-e^{i\omega})=\frac{2\sin\omega}{\omega}$$

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)e^{i\omega t}d\omega=\frac{1}{\pi}\int_{0}^{+\infty} F(\omega)\cos\omega t d\omega$$

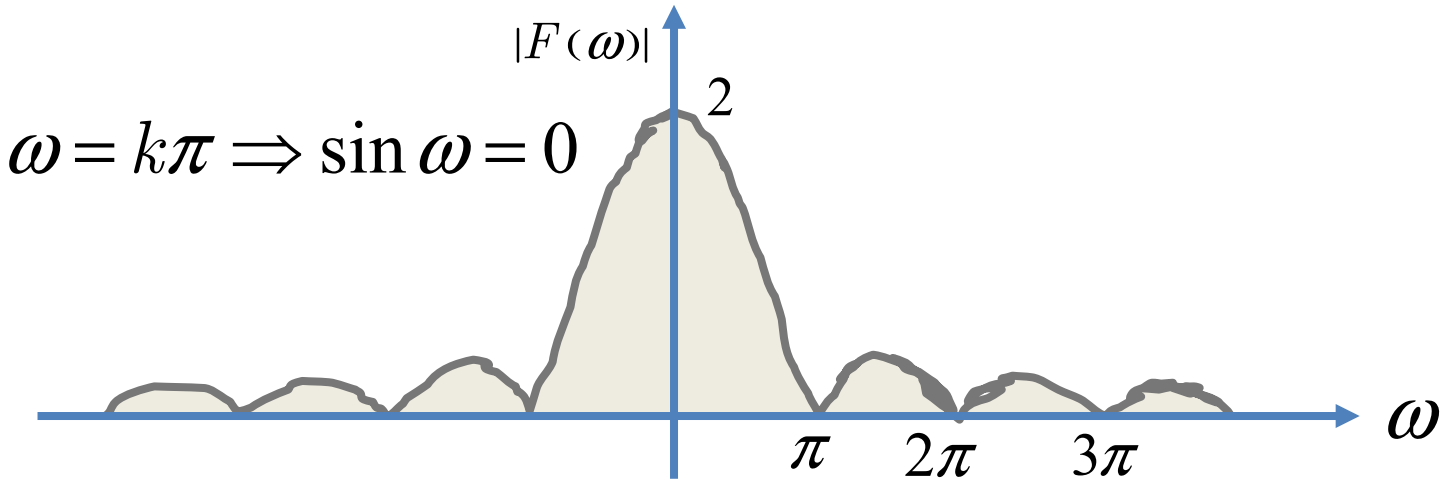
$$=\frac{1}{\pi}\int_{0}^{+\infty}\frac{2\sin\omega}{\omega}\cos\omega t d\omega=\frac{2}{\pi}\int_{0}^{+\infty}\frac{\sin\omega\cos\omega t}{\omega}d\omega$$

$$\int_0^{+\infty} \frac{\sin\omega\cos\omega t}{\omega}\mathrm{d}\omega = \begin{cases} \frac{\pi}{2} & |t|<1 \\ \frac{\pi}{4} & |t|=1 \\ 0 & |t|>1 \end{cases}$$

因此可知当$t=0$时, 有

$$\int_0^{+\infty} \frac{\sin x}{x}\mathrm{d}x = \int_0^{+\infty} \mathrm{sinc}(x)\mathrm{d}x = \frac{\pi}{2}$$

另外，由$|F(\omega)|=2\left|\frac{\sin\omega}{\omega}\right|$可作出频谱图:



定义事件发生频率和概率

在一个时间序列信号上任意一个数值A出现的次数(频数)为： n_A

信号所有数值出现的总数： N

则f(A)= n_A /N 定义为 A在这N个信号数值中出现的**频率**，

当信号时间序列数值N足够多时，某数值A出现的频率f(A)趋于稳定值p，

可定义f(A)=p为A出现的概率，记为P(A)=p

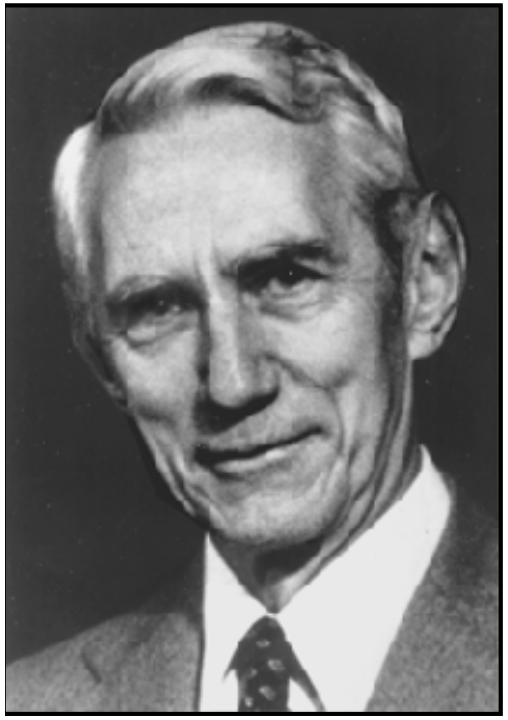
满足如下性质：

1） $0 \le P(A) \le 1$

2）若$A_1, A_2, \cdots, A_k$两两互不相容，则

$$P(\bigcup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i) = 1$$

信息论：如何量化信息？



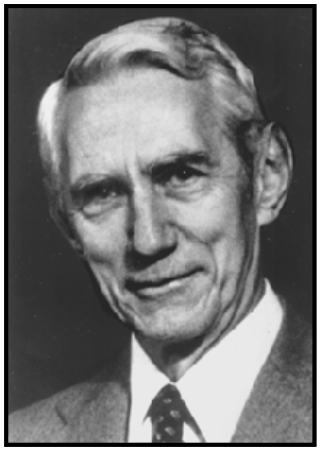
Claude Elwood Shannon (April 30, 1916 - February 24, 2001) has been called “the father of information theory”.

1948: « A Mathematical Theory of Communication »

克劳德·艾尔伍德·香农：美国数学家、信息论的创始人
Shannon在题为“通讯的数学理论”的论文中指出：

“信息是用来消除随机不定性的东西”

概率=》不确定性和信息熵



克劳德·香农

对于A事件发生的概率P（A），发生的概率范围（0～1）值越大，其发生的概率越大，不确定性就越小，我们可定义不确定性为：

$h = -\log_2 P$

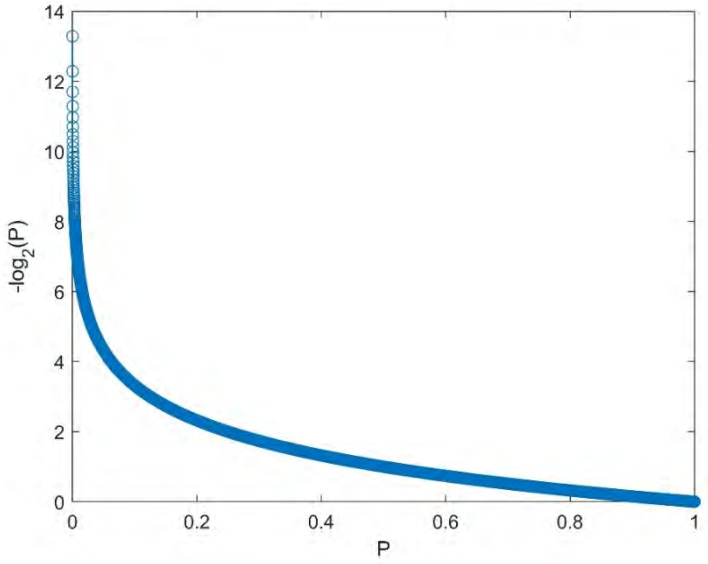
概率越小，不确定性越大！

假定一事件发生有x_1、x_2…x_N种结果

每一种结果出现的概率为$P(x_i)$，

简写为P_i，为了衡量这样一个概率分布的不确定性，香农引入信息熵定义：

$$H = -\sum_{i=1}^{N} P_i \log_2 P_i$$



1948: « A Mathematical Theory of Communication »

信息量

假定有一事件可能有x_1、$x_2\cdots x_N$种结果，每一种结果出现的概率为$P(x_i)$，或简写为P_i，香农把这类事件信息不确定性定义为信息熵：

$$H = -\sum_{i=1}^{N} P_i \log_2 P_i$$

$$\text{if } P_i = \frac{1}{N},\ H_0 = -\left[\frac{1}{N}\log_2\frac{1}{N} + \frac{1}{N}\log_2\frac{1}{N}\cdots + \frac{1}{N}\log_2\frac{1}{N}\right]$$

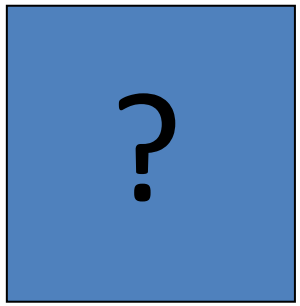
$$H_0 = -\log_2\frac{1}{N} = \log_2 N$$

观察过后，概率确定，此时信息熵H_{end}=0

这个过程信息量的获得：

$\Delta I = H_0 - H_{end} = \log_2 N$　等概率事件观察获得信息量

信息 = 不确定性减少
(Shannon,1948)

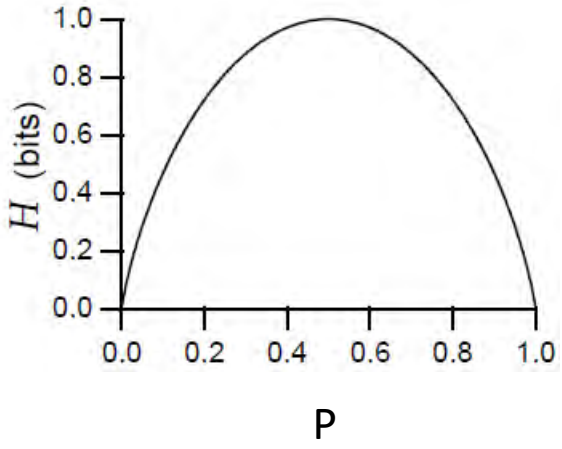


正面或反面？

观察后

正面概率为P(A)=p
反面概率为P(B)=1-p
则未知正反的时候其信息熵就是：$H(p) = -p*\log_2 p - (1-p)*\log_2(1-p)$
当p值从0.01逐渐增加到1时，看信息熵变化曲线：

发现p=0.5时对应信息熵最大，
系统那时的不确定性最大



一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以，信息熵也可以说是系统通讯过程有序化程度的一个度量。

信息熵和信息量量单位

- 在信息论中常用的对数底是2，信息量的单位为比特(bit)；
- 若取自然对数，则信息量的单位为奈特(nat);
- 若以10为对数底,则信息量的单位为笛特(det)

$$1\ \text{nat} = \log_2 e \approx 1.433\ \text{bit}$$

$$1\ \text{det} = \log_2 10 \approx 3.322\ \text{bit}$$

在缺少足够的输入情况下，大脑对世界的观察信息熵多大？
观察后获得多少信息量？

对确定性信息的提取和判断

男人还是女人？大约多少岁了？那是什么季节？他想干什么？灯光的颜色是什么？……从不同的角度来看，这一张普通的照片里包含了很多信息，确定性的，不确定性的。。。

观察前后信息量的获得

如何计算信息量

男人还是女人？

- 开始：男 （p1=50%）女 (p2=50%)

$$I_1 = -\sum_{i=1}^{N} P_i \log_2 P_i = -（0.5 \times \log_2(0.5) + 0.5 \times \log_2(0.5))$$

$$= -(\log_2(\frac{1}{2}) = \log_2(2) = 1 \text{ (bits 比特)}$$

观察后：男（p1=100%) 女 (p2=0%)

$$I_2 = -\sum_{i=1}^{N} P_i \log_2 P_i = -（1 \times \log_2(1) = 0 \text{ (bits 比特)}$$

$$\Delta I = I_1 - I_2 = 1 \text{ (bits)}$$

→ 观察过程掌握的信息量之一

如何计算信息量

什么季节？

春夏秋冬 （可能性 pi=1/4）

$$I_1 = -\sum_{i=1}^{N} P_i \log_2 P_i = -1 \times \log_2(\frac{1}{4}) = 1 \times \log_2(4) = 2 \text{ (bits)}$$

$$I_2 = -1 \times \log_2(1) = 0 \text{ (bits)}$$

$$\Delta I = I_1 - I_2 = 2 \text{ (bits)}$$

→ 观察过程掌握的信息量之二

春夏秋冬 （可能性 pi=1/4）

如何计算信息量

多少岁？

年龄

观察前：1—100岁 （可能性 pi=1/100）

观察后： 可能50—59岁左右 （p50,p51….p59=1/10)

$$I_1 = -\sum_{i=1}^{N} P_i \log_2 P_i = -1 \times \log_2(\frac{1}{100}) = 1 \times \log_2(100) = 6.6439 \text{ (bits)}$$

$$I_2 = -1 \times \log_2(\frac{1}{10}) = 3.322 \text{ (bits)}$$

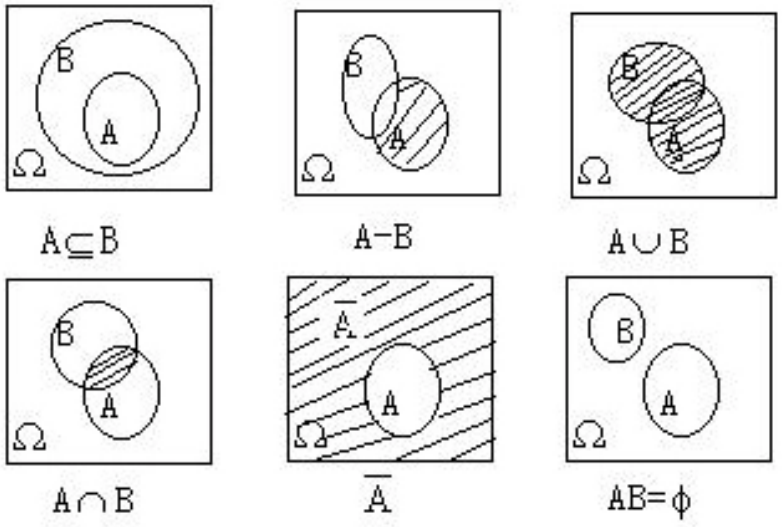
$$\Delta I = I_1 - I_2 = 3.322 \text{ (bits)}$$

→ 观察过程掌握的信息量之三

事件发生的概率

联合概率 P(AB)表示两个事件共同发生的概率，A与B的联合概率也可表示为 P(A ∩ B)

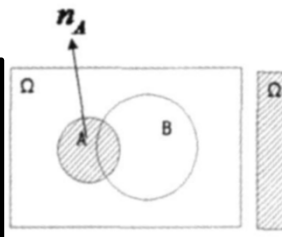
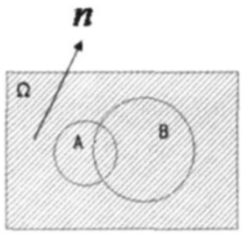
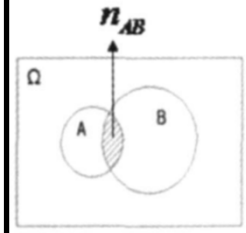
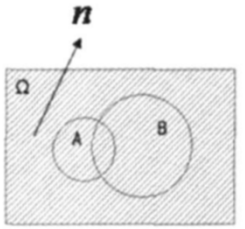
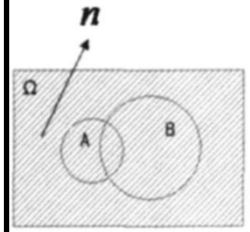
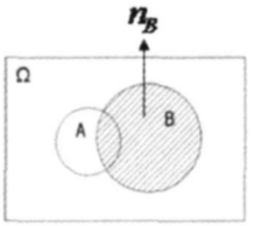
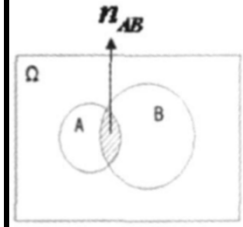
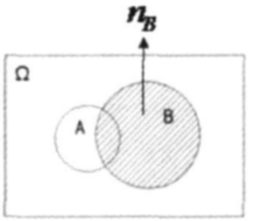
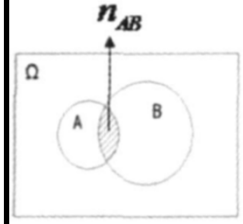
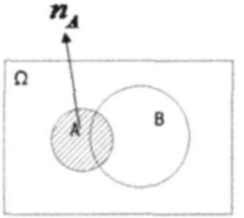
P(A∩B)是样本空间S中A事件和B事件同时发生的概率，也就是A和B相交的区域



P(AB)=P(A ∩ B) 是A、B同时发生的概率

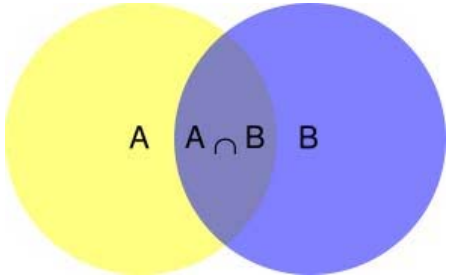
P(AUB)是A或者B发生的概率，A与B的积事件

表 1　概率的计算与韦恩图

序号	所求概率	古典定义	韦恩图 分子	韦恩图 分母
		$P(A)=n_A/n$		
		$P(AB)=n_{AB}/n$		
		$P(B)=n_B/n$		
		$P(A\|B)=n_{AB}/n_B$		
		$P(B\|A)=n_{BA}/n_A$		

条件概率和联合概率

1. 联合概率：P(AB)表示两个事件共同发生的概率
A与B的联合概率也可表示为 P(A ∩ B)



联合概率公式：P(AB)= P(B) P(A|B)= P(A) P(B|A)

如A和B是相互独立，没有关联，则 P(AB)= P(A)P(B)

如A和B是互不相容，具有互斥性，则 P(AB)= 0
(A和B不能同时发生)

条件概率P(A|B)大小等于AB相交面积/B事件面积

条件概率P(B|A)大小等于AB相交面积/A事件面积

2. 条件概率：设有两个事件A, B，我们把在事件B出现的条件下会关联到事件A发生的概率定义为条件概率P(A|B)，

1. 可以通过如下公式计算得到：**P(A|B)= P(AB)/P(B)**

右图：B事件发生后A事件也发生的概率**P(A|B)= P(AB)/P(B)**

A事件发生后B事件也发生的概率**P(B|A)= P(AB)/P(A)**

条件熵

给定一个事件，一个随机变量的条件熵

- 在给定一个y_j条件下，X集合的条件熵

$$H(X/y_j)=-\sum_i p(x_i/y_j)\log_2 p(x_i/y_j)$$

已知一个随机变量，另一个随机变量的条件熵

- 则在给定Y（即所有y_j）条件下，X集合的条件熵

$$H(X/Y)=\sum_j p(y_j)H(X/y_j)=-\sum_j p(y_j)\sum_i p(x_i/y_j)\log_2 p(x_i/y_j)$$

$$=-\sum_{i,j} p(x_i y_j)\log_2 p(x_i/y_j)$$

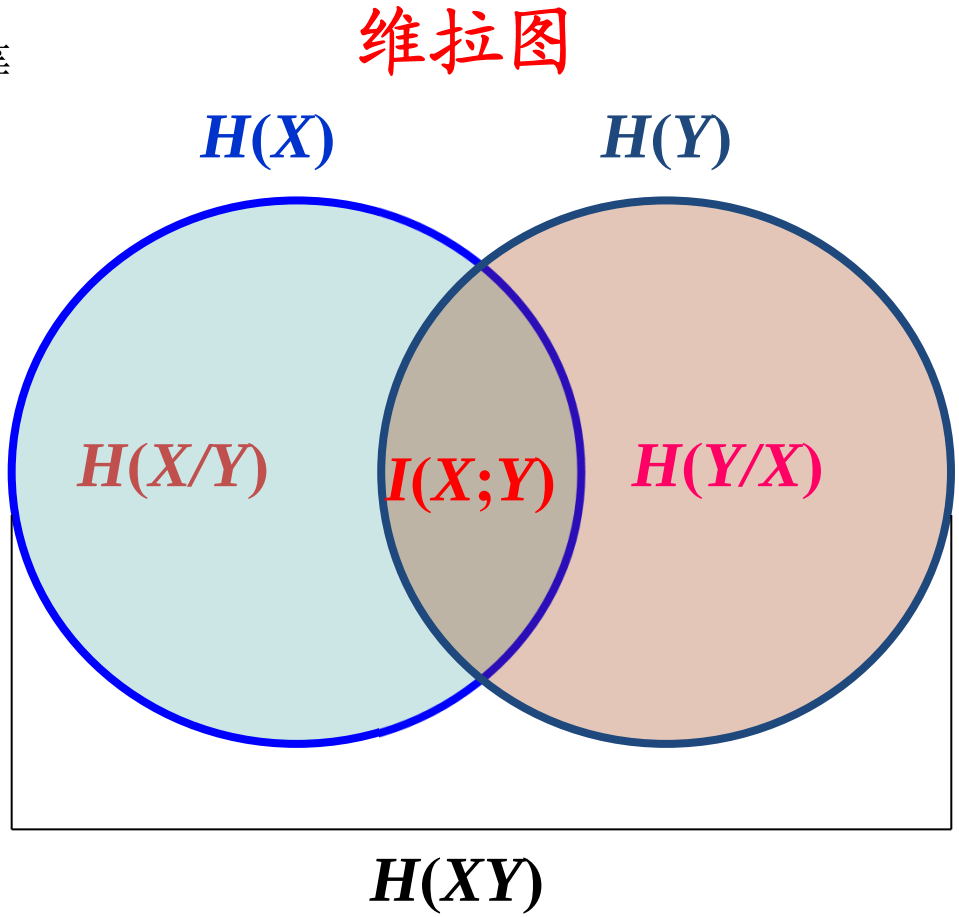
条件熵H(X/Y)也称为在观察到Y事件情况下条件X出现的熵

互信息量与熵

- 熵是平均不确定性的描述
- 互信息：接收端所获得的信息量等于不确定性的消除(两熵之差)

$$I(X;Y) = H(X) - H(X/Y)$$
$$= H(Y) - H(Y/X)$$
$$= H(X) + H(Y) - H(XY)$$

左侧圈表示 H(X)，
右侧圈表示 H(Y)
两个圈覆盖的所有区域表示 H(XY)
左侧（不包含重合部分）区域表示 H(X|Y)
右侧（不包含重合部分）区域表示 H(Y|X)



联合熵

- 联合熵是联合事件集合 *XY*上的信息熵

$$H(XY) = -\sum_{i,j} p(x_i y_j) \log_2 p(x_i y_j)$$

- 联合熵*H*(*XY*)表示*X*和 *Y*同时发生的不确定度;
  或两个随机变量联合概率的自信息的数学期望。

定理证明

证明：联合熵、信源熵和条件熵之间的关系

$$H(XY)=H(X)+H(Y/X)=H(Y)+H(X/Y)$$

$$H(XY)=-\sum_{i,j}p(x_iy_j)\log p(x_iy_j)$$

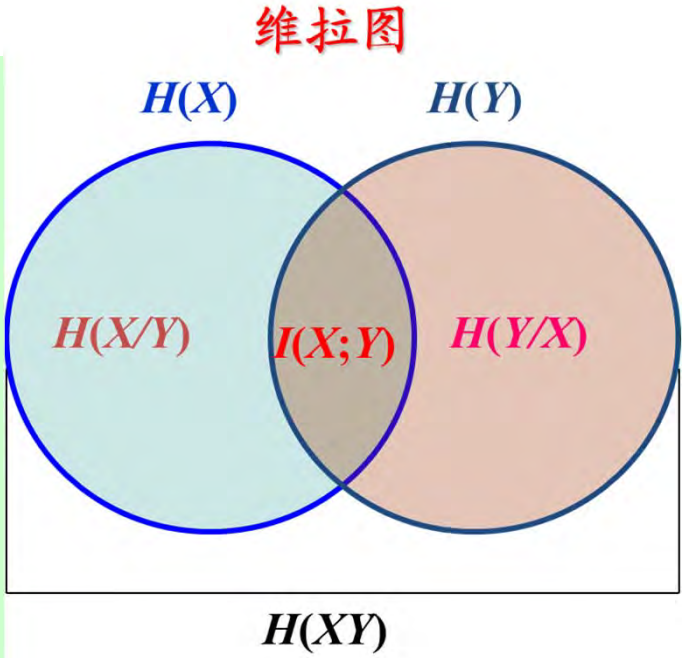
$$=-\sum_{i,j}p(x_iy_j)\log[p(x_i)p(y_j/x_i)]$$

$$=-\sum_{i,j}p(x_iy_j)\log p(x_i)-\sum_{i,j}p(x_iy_j)\log p(y_j/x_i)$$

$$=-\sum_{i}[\sum_{j}p(x_iy_j)]\log p(x_i)-\sum_{i,j}p(x_iy_j)\log p(y_j/x_i)$$

$$=-\sum_{i}p(x_i)\log p(x_i)+H(Y/X)$$

$$=H(X)+H(Y/X)$$



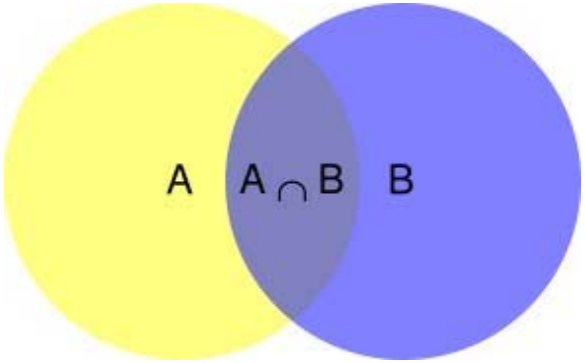
条件概率和联合概率

1. 联合概率：P(AB)表示两个事件共同发生的概率
A与B的联合概率也可表示为 P(A ∩ B)

联合概率公式：P(AB)= P(B) P(A|B)= P(A) P(B|A)

如A和B是相互独立，没有关联，则 P(AB)= P(A)P(B)

如A和B是互不相容，具有互斥性，则 P(AB)= 0
(A和B不能同时发生)



条件概率P(A|B)大小等于AB相交面积/B事件面积

条件概率P(B|A)大小等于AB相交面积/A事件面积

2. 条件概率：设有两个事件A, B，我们把在事件B出现的条件下会关联到事件A发生的概率定义为条件概率P(A|B)，

1. 可以通过如下公式计算得到：**P(A|B)= P(AB)/P(B)**

右图：B事件发生后A事件也发生的概率**P(A|B)= P(AB)/P(B)**

A事件发生后B事件也发生的概率**P(B|A)= P(AB)/P(A)**

右于P(AB)= P(A)P(B|A)= P(B)P(A|B)

所以： P(A|B)=P(A)P(B|A)/ P(B)

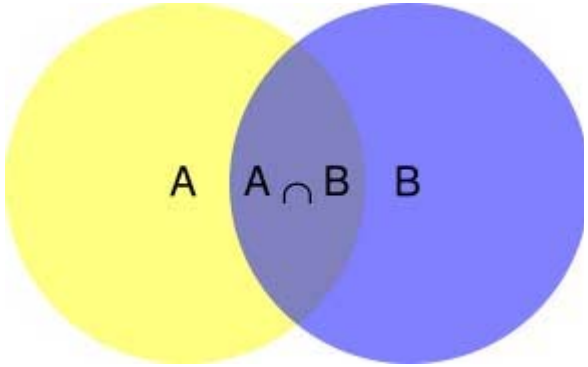
贝叶斯公式

由 P(AB)= P(A)P(B|A)= P(B)P(A|B)

==》

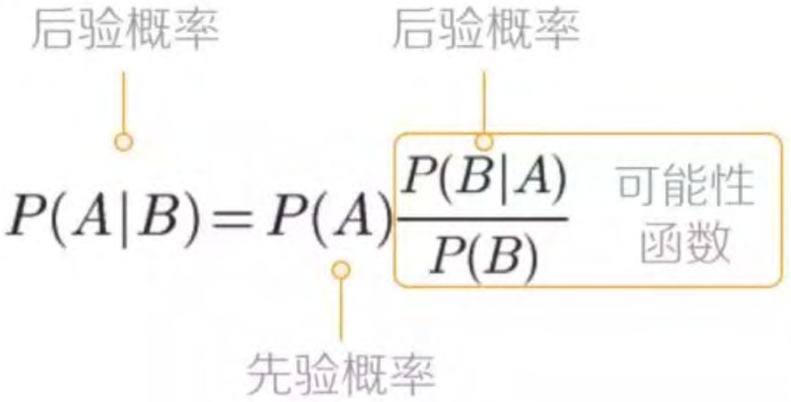
P(A|B) = P(A)P(B|A)/P(B)

贝叶斯公式



P(B|A)也称为“似然函数”（Likelyhood）等于给定参数 A 后变量 B发生的概率，当我们最大化这个函数时，可以最佳估算P(A|B)

后验概率（posterior probability）在给出相关证据或现象后得到某事件发生的条件概率

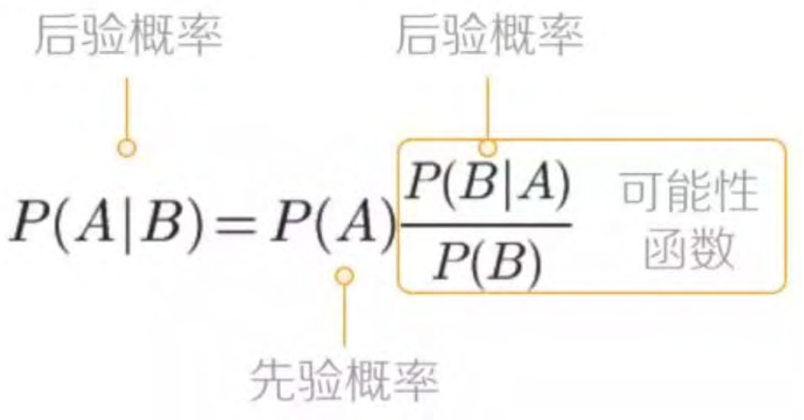


先验概率（prior probability）根据以往经验和分析得到的概率

贝叶斯，英国数学家, 创立贝叶斯统计理论，

https://seeing-theory.brown.edu/bayesian-inference/cn.html

Thomas Bayes
(1702-1761)



1、P(A)是A的先验概率或边缘概率，称作"先验"是因为它不考虑B因素。
2、P(A|B)是已知B发生后A的条件概率，也称作A的后验概率。
3、P(B|A)是已知A发生后B的条件概率，也称作B的后验概率，这里称作似然度。
4、P(B)是B的先验概率或边缘概率，这里称作标准化常量。
P(B)称为证据(evidence)，即无论事件如何，事件B(或evidence)的可能性有多大
5、P(B|A)/P(B)称作标准似然度。
贝叶斯法则又可表述为：A后验概率=(似然度*先验概率)/标准化常量
=标准似然度*A先验概率

P(A|B)随着P(A)和P(B|A)的增长而增长，随着P(B)的增长而减少，即如果B独立于A时被观察到的可能性越大，那么B对A的支持度越小。

贝叶斯公式举例

比如：5月1日去黄山旅游，早上发现多云(B事件），问下雨（A事件）发生的概率？

已知历史数据经验：

1）5月份的早上多云（B事件）的概率是40%

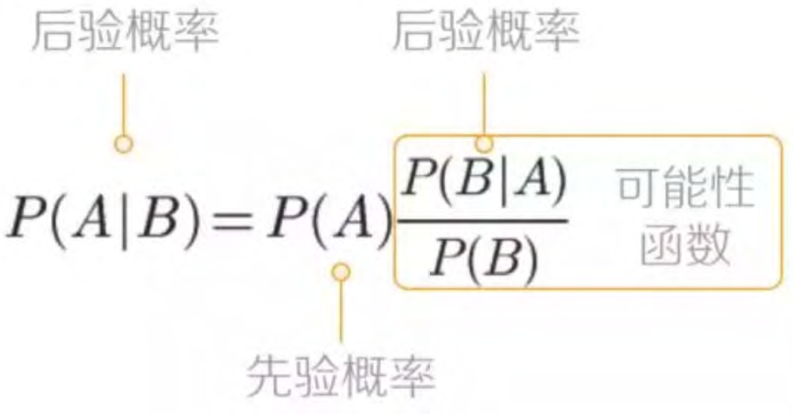
2）5月份下雨（A事件）的概率是10%

3）5月份，早上下雨时多云的概率是50%：P(B|A)

P(云)=40%

P(雨)=10%

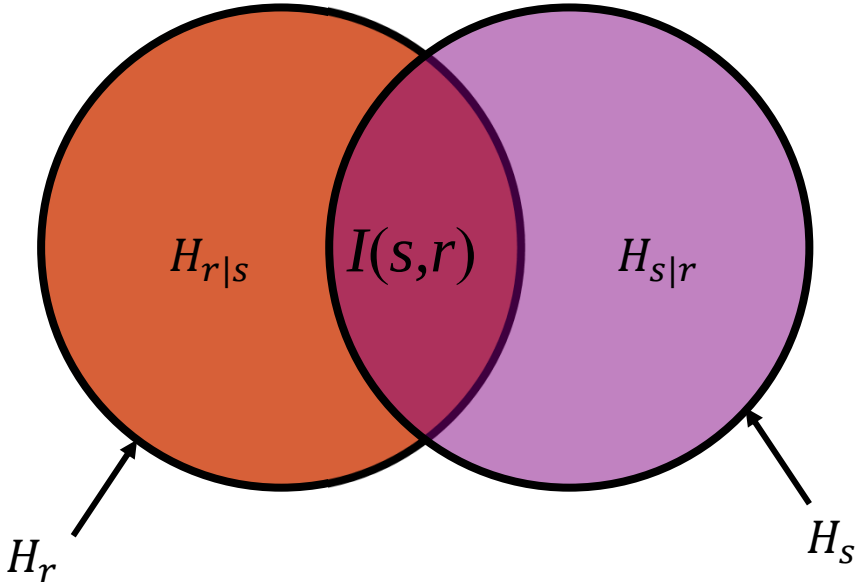
P(云|雨)=50%



P(雨|云)=P(雨)*P(云|雨)/P(云)

=0.1*0.5/0.4=12.5%

神经元对刺激s的响应r：互信息



$H_{r|s}$ 给定刺激s(t)观察到神经信号的不确定反应r(t)的条件信息熵

$H_{s|r}$ 记录到神经信号反应r(t)时，观察到不确定刺激信号s(t) 条件信息熵

互信息(Mutual information)

$$I(r,s) = H_r - H_{r|s}$$
$$= I(s,r) = H_s - H_{s|r}$$

互信息衡量了神经反应中有多少信息是用于编码刺激的

$H_{r|s}$ 也可写成H_{noise}，即神经响应中因背景噪声存在的不确定性

$$I(S;R) = \sum_{r,s} P(s)P(r|s)\log_2 \frac{P(r|s)}{P(r)}$$

互信息Mutual Information

神经元对一个给定刺激s的反应熵：

$$H_s = -\sum_r P[r|s] \log_2 P[r|s]$$

噪声熵 noise entropy: $H_{\text{noise}} = \sum_s P[s] H_s = -\sum_{s,r} P[s] P[r|s] \log_2 P[r|s]$

This is the entropy associated with that part of the response variability that is not due to changes in the stimulus, but arises from other sources.

$$I_{\text{m}} = H - H_{\text{noise}} = -\sum_r P[r] \log_2 P[r] + \sum_{s,r} P[s] P[r|s] \log_2 P[r|s]$$

$$P[r] = \sum_s P[s] P[r|s]$$

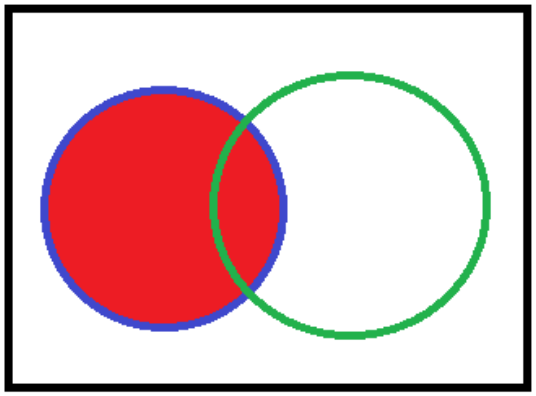
Mutual Information

$$I(S;R) = \sum_{r,s} P(s) P(r|s) \log_2 \frac{P(r|s)}{P(r)}$$

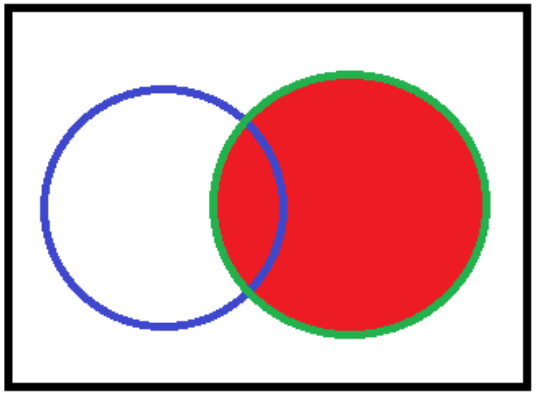
课后作业：

1 把10枚相同的硬币抛在地面上，每枚硬币有正反两种可能，问观察前后信息熵的变化，或观察过程获取的信息量时多少？

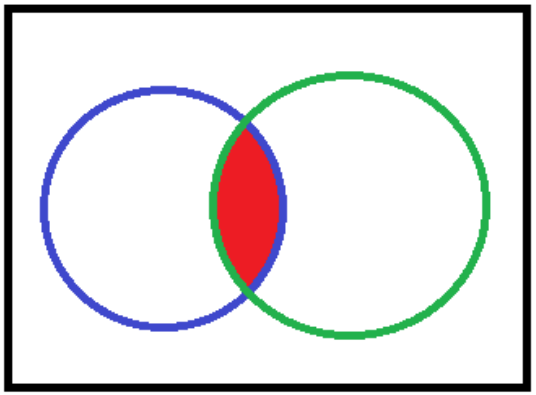
2 有一同学，考试成绩数学不及格的概率是0.15，语文不及格的概率是0.05，两者都不及格的概率为0.03，在一次考试中，已知他数学不及格，那么他语文不及格的概率是多少？



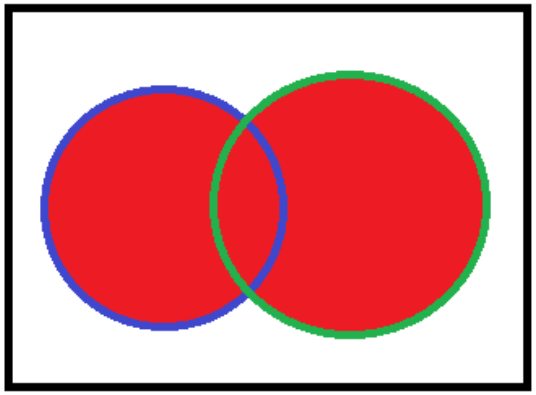
P(A)



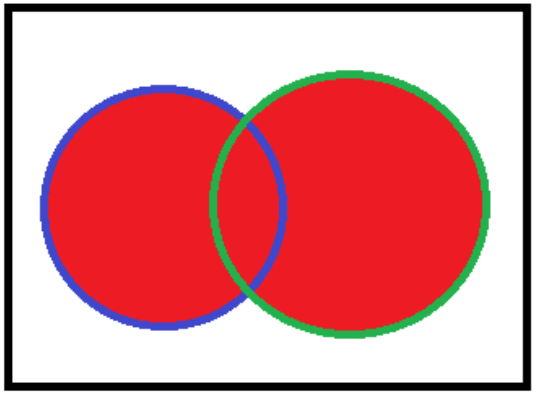
P(B)



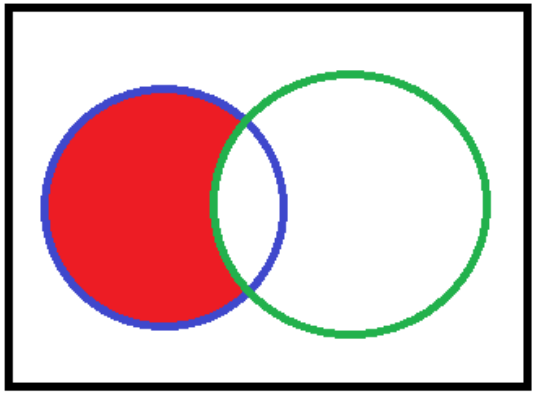
P(AB)



P(A+B)



P(A∪B)



P(A−B)