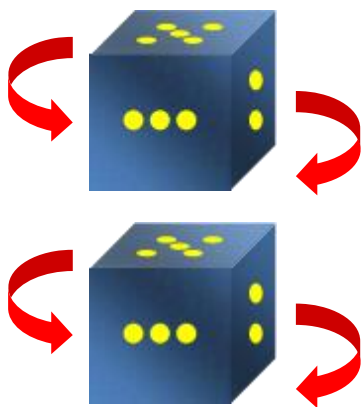




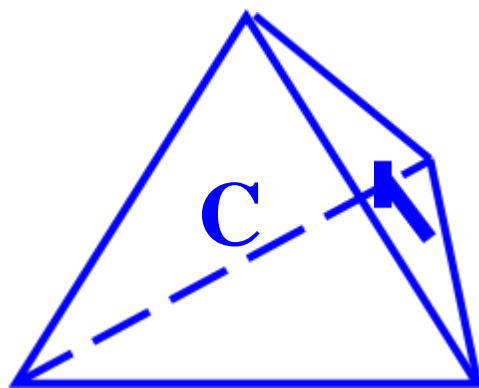
(一) 概率基础

研究随机现象的基本方法是进行随机试验 E , 满足:

- (1) 试验的所有可能结果不止一个, 且可以事先明确;
- (2) 试验结果的不确定性: 不可预言; → 随机性
- (3) 试验的重复性: 试验的条件可重复实现。→ 统计规律



点数1~6中的
任意之一发生



ACGT中的任意之一发生

...GCATGGCTAA...
...ACGCTGCTGA...

AC、AG、... 中的任
意之一发生

基本事件 (generic element) ω : 随机试验E的基本结果。

(基本事件 (elementary event) 的说法是错误的, 应该是“基本结果、基本元”)

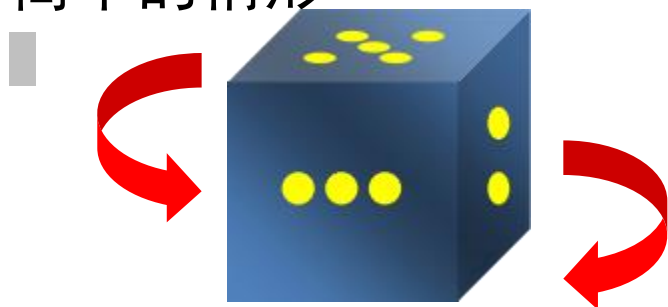
样本空间 (sample space) Ω : 全体基本事件组成的集合, $\Omega = \{\omega\}$ 。

(Ω 是可测空间, 类似于具有质量、密度、体积的物质空间)

(1) 完备性: 每次试验必出现一个基本事件 ω , $\omega \in \Omega$

(2) 互斥性: 每次试验只出现一个基本事件, 任何两个不同的基本事件不同时发生

(3) 最简性: 基本事件是最简单的试验结果, 不能划分为更简单的情形



全部点数: {1; 2; 3; 4; 5; 6}

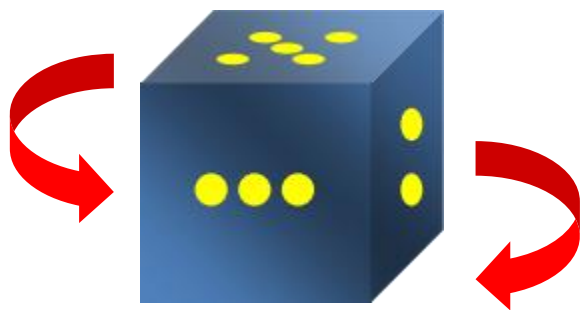
奇数点数: {1; 3; 5}

接连两次奇数: {11; 13; 15; 31; 33;
35; 51; 53; 55}

随机事件(random event) A : 样本空间的子集, $A \subseteq \Omega$ 。

- (1) 基本事件是最简单的随机事件;
- (2) 随机事件是若干基本事件的集合;
- (3) 若 $\omega \in A$, 则 ω 的发生 $\Rightarrow A$ 的发生;

事件域 \mathcal{F} : 某个随机试验相关的所有随机事件的集合, 它们需满足构成 σ 域 (σ -field) 的条件, 是样本空间 的子集族。



投掷3次: $\{1\}; \{1\}; \{5\}$
一次随机事件

- 样本空间(sample space) Ω : $\Omega = \{ \omega \}$ 。
- 事件域 (events field) \mathcal{F} : $\{A\}$ (A 为 Ω 的子集, 满足构成域的条件, 进一步选择满足一定条件的 A 构成 \mathcal{F} , 此处的条件将由Kolmogorov公理化定义阐明)
- 随机事件 (random event) A : $A \subseteq \Omega$, 且 $A \in \mathcal{F}$

如果能定义样本空间上的一种测度 P (起到物理上的质量的测量作用), 使得:

$$P(\Omega)=1$$

$P(A)$ 就是随机事件 A 的概率

设 E 是随机试验， Ω 为其样本空间， \mathcal{F} 为定义在该样本空间的事件域。对于每个随机事件 $A \in \mathcal{F}$ ，定义实值函数 $P(A)$ ($A \in \mathcal{F}$) 满足

- (1) 非负性: $P(A) \geq 0$ ($A \in \mathcal{F}$) ;
- (2) 归一性: $P(\Omega) = 1$;
- (3) 可列可加性: 设 A_i ($i \geq 1$)互不相容, 即 $A_i A_j = \emptyset$ ($i \neq j$), 则有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

称 $P(A)$ ($A \in \mathcal{F}$) 为随机事件 A 的概率。

概率空间: Ω 、 \mathcal{F} 、 $P(A)$ 三要素

频数 (Occurrence): 在 n 次重复试验中, 事件 A 的出现次数 n_A 称为事件 A 的频数

频率 (Frequency): $f_A = n_A / n$

频率的性质:

(1) 非负性: $f_A \geq 0$

(2) 归一性: $f_{\Omega} = 1$

(3) 可加性: 若事件 A 、 B 不同时发生 (不相容), 则 $f_{A \cup B} = f_A + f_B$

频率的稳定性：

在大量次数的试验中，在大多数情况下，随着试验次数的增加，随机事件 A 发生的频率将稳定在某个常数附近。

概率的统计定义：

设 A 是随机事件，由于频率稳定性，在大多数情况下， f_A 将稳定在某个常数附近，称此常数为事件 A 的概率，记为 $P(A)$ 。

概率的描述性定义：

概率是随机事件发生的可能性大小的数量表示，是定义于事件域 \mathcal{F} 上取值于 $[0,1]$ 的函数。

概率与频率的关系：

频率：

- (1) 在一定程度上反映了随机事件发生的可能性；
- (2) 依赖于试验本身（试验者）、试验次数。

概率：

- (1) 反映随机事件发生的可能性，是随机事件本身固有的性质
- (2) 不依赖于具体的试验；
- (3) 以频率稳定性为基础，并通过大量试验中的频率稳定性来表现

历史上的抛掷硬币试验记录

试验者	n	n_A	$f_n(A)$
<i>Buffon</i>	4040	2048	0.5080
<i>Pearson</i>	12000	6019	0.5016
<i>Pearson</i>	24000	12012	0.5005



讨论：

1. 频率定义概率（统计概型）的理论基础是**大数定律**，以描述随机事件A发生概率 $p(A)$ 的**Bernoulli试验**来实现；
2. 统计概率在实践中具有重要意义，它是数理统计的基础；
3. 概率的统计定义需要大量重复试验，在实际应用中具有限制性。统计实践中，在满足一定条件下，**频率值 f_A 是概率值 $P(A)$ 的最大似然估计解。**

概型 (Scheme, probability model)

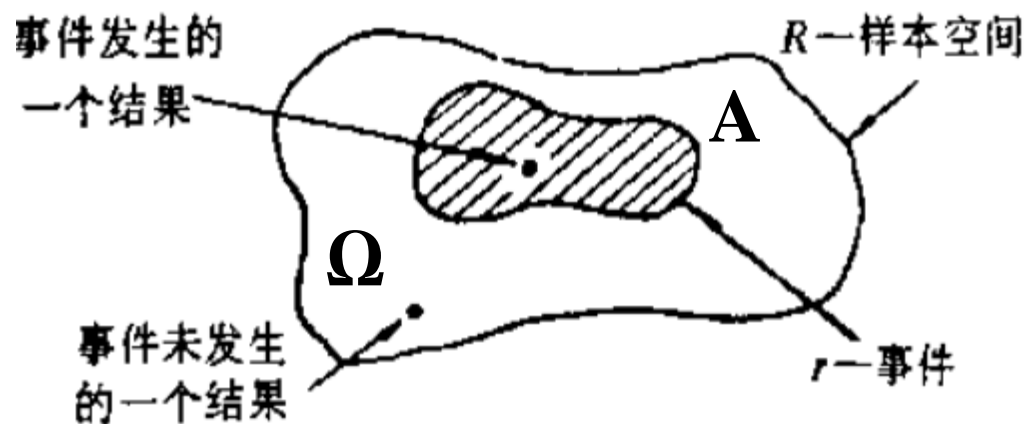
目的：为实现 $P(A)$ 的计算

办法：

$$P(A) = \frac{D(A)}{D(\Omega)}$$

$D(\Omega)$: 样本空间 Ω 的测度

$D(A)$: 随机事件 A 的测度



随机变量 X

在自然界中，有些变量在每次观察前，不可能事先确定其取值；经过大量反复观察，其取值又有一定的规律，这种变量称为随机变量 X 。

离散型随机变量

X 的所有可能取值是有限个或可列个。

连续型随机变量

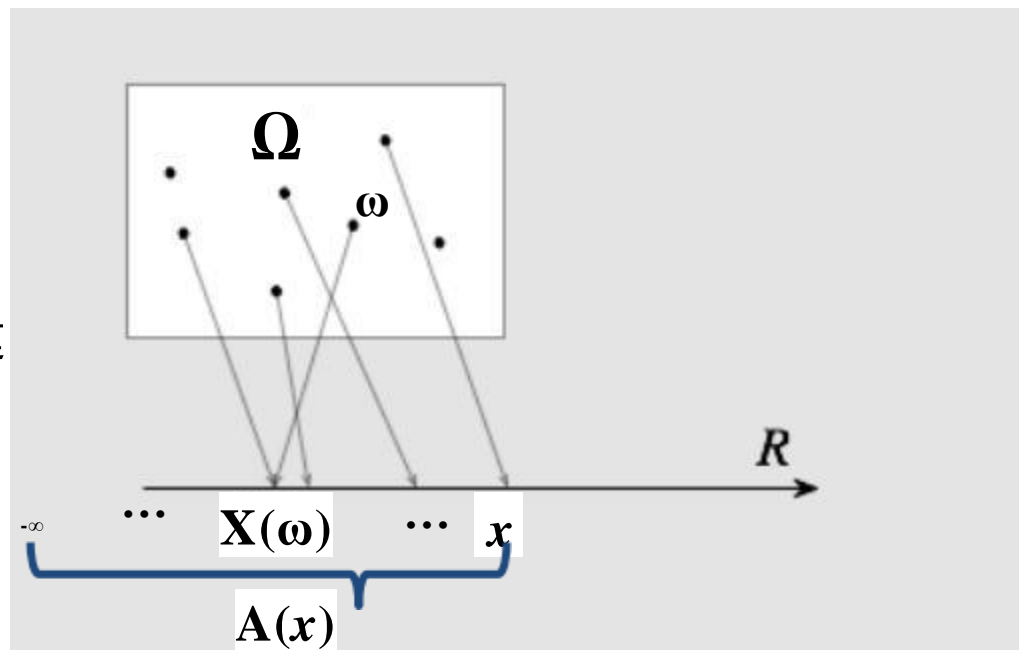
最常见的一类非离散型随机变量。

随机变量 X (random variable):

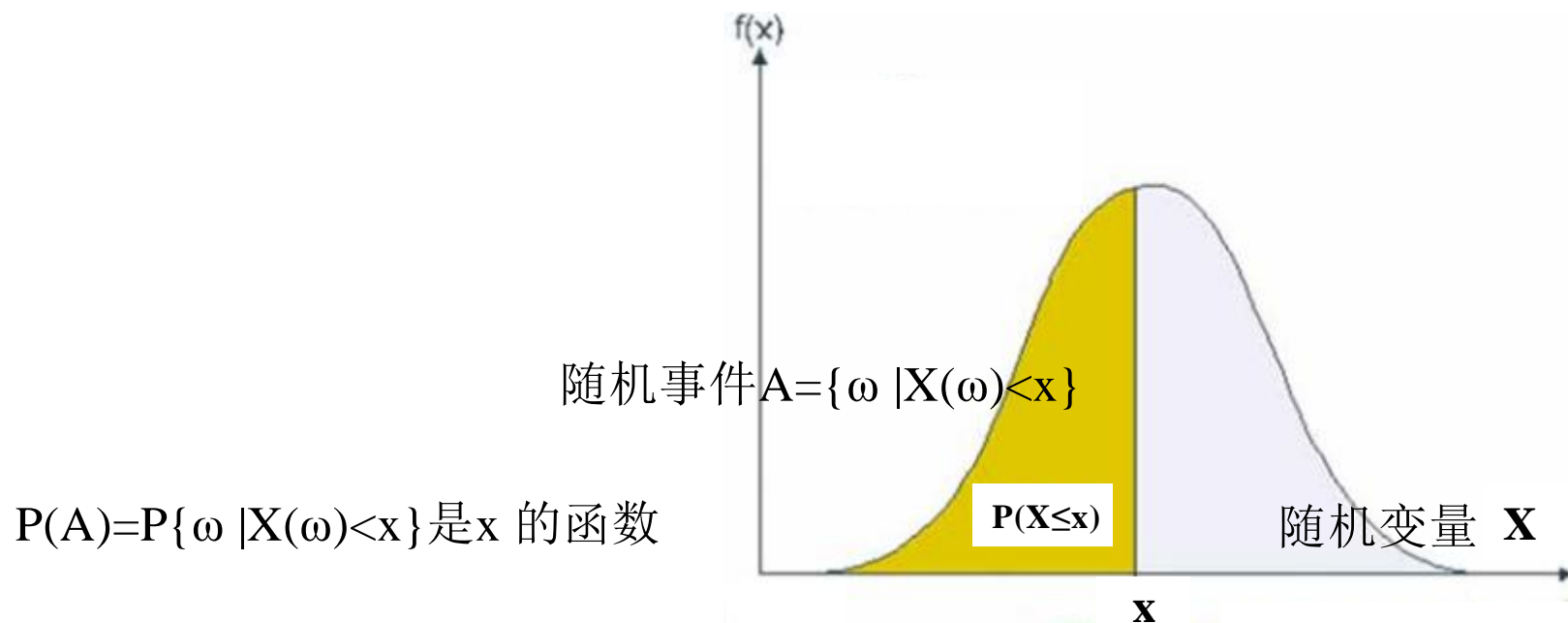
设 (Ω, F_Ω, P) 为概率空间, $X=X(\omega)$ ($\omega \in \Omega$) 是定义在 Ω 上的单值实函数, 若对于任一实数 $x \in \mathbf{R}$, ω 的集合 $\{\omega: X(\omega) \leq x\}$ 是一随机事件, 亦即 $\{\omega \mid X(\omega) \leq x\} \in F_\Omega$, 则称 $X(\omega)$ 为随机变量, 简记为 X 。

不仅定义随机变量的取值
(可测函数), 也定义这些
取值具有一定的规律——即
概率分布特征。

也因此定义了一类事件 A , 建
立了一个良好的数学模型。
(并非所有的事件 A)



- (1) 随机变量的取值具有随机性, 并有一定的概率规律;
- (2) 随机变量实质上是定义在样本空间 Ω 的一个实值函数, 亦即对样本空间 Ω 上的某一样本点 ω 赋值 $X(\omega)$ 来表示该点;
- (3) 注意: 与概率模型中函数 $P(A)$ 的区别。



概率密度函数与概率分布函数

对连续型随机变量，考察事件 $\{a < X < b\}$ 的概率。若存在非负的可积函数 $p(x)$ ，使得：对任意的 $a, b (a < b)$ ，都有

$$P\{a < X < b\} = \int_a^b p(x) dx$$

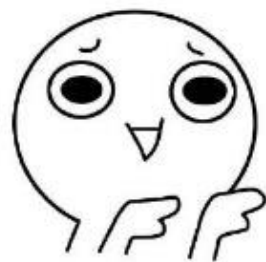
则称 $p(x)$ 为随机变量 X 的概率密度函数。

对所有随机变量 X ，可以定义以下的概率分布函数 $F(x)$ ：

$$F(x) = P\{X \leq x\} = \int_{-\infty}^x p(t) dt \quad \longrightarrow \quad p(x) = F'(x)$$

$p(x)$ 的性质： \longrightarrow

$$\begin{aligned} p(x) &\geq 0 \\ \int_{-\infty}^{+\infty} p(x) dx &= 1 \end{aligned}$$



老师提问时看不到我

Q:请说出你知道的统计分布?

二项分布 (binomial distribution)

Bernoulli试验：连续 n 次独立地重复一个试验，每次试验结果只有两个不同的结果A、非A，它们出现的概率分别是 p 、 q ，且 $p+q=1$ 。

$$P\{X = k\} = C_n^k p^k q^{n-k} \quad k = 0, 1, 2, \dots, n$$

设 n 重Bernoulli试验中事件A出现的次数为 X ，显然 X 为离散型随机变量。

则 X 的概率分布为：

$$\begin{aligned} P\{X = k\} &\geq 0 \quad k = 0, 1, 2, \dots, n \\ \sum_{k=0}^n C_n^k p^k q^{n-k} &= (p + q)^n = 1 \end{aligned}$$

称 X 服从参数为 n, p 的二项分布，记为 $X \sim B(n, p)$ 。

均匀分布 (uniform distribution)

设 X 为连续型随机变量, X 的概率密度为:

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{其它值} \end{cases}$$

称 X 在区间 $[a, b]$ 上服从均匀分布, 记为 $X \sim U(a, b)$ 。

显然有:

$$P\{x_1 \leq X \leq x_2\} = \int_{x_1}^{x_2} p(x) dx = \frac{x_2 - x_1}{b - a}$$

其中 $x_1, x_2 \in [a, b], x_1 < x_2$ 。

Poisson分布 (Poisson distribution)

设 X 为离散型随机变量, X 的概率分布为:

$$P\{X = k\} = \frac{e^{-\lambda} \lambda^k}{k!}, \quad \lambda > 0, \quad k = 0, 1, 2, \dots$$

称 X 服从参数为 λ 的Poisson分布, 记为 $X \sim \Pi(\lambda)$ 。

指数分布 (exponential distribution)

设 X 为连续型随机变量, X 的概率密度为:

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases} \quad \lambda > 0$$

称 X 服从参数为 λ 的指数分布。

正态分布 (Normal distribution)

设随机变量 X 的概率密度为：

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < +\infty$$

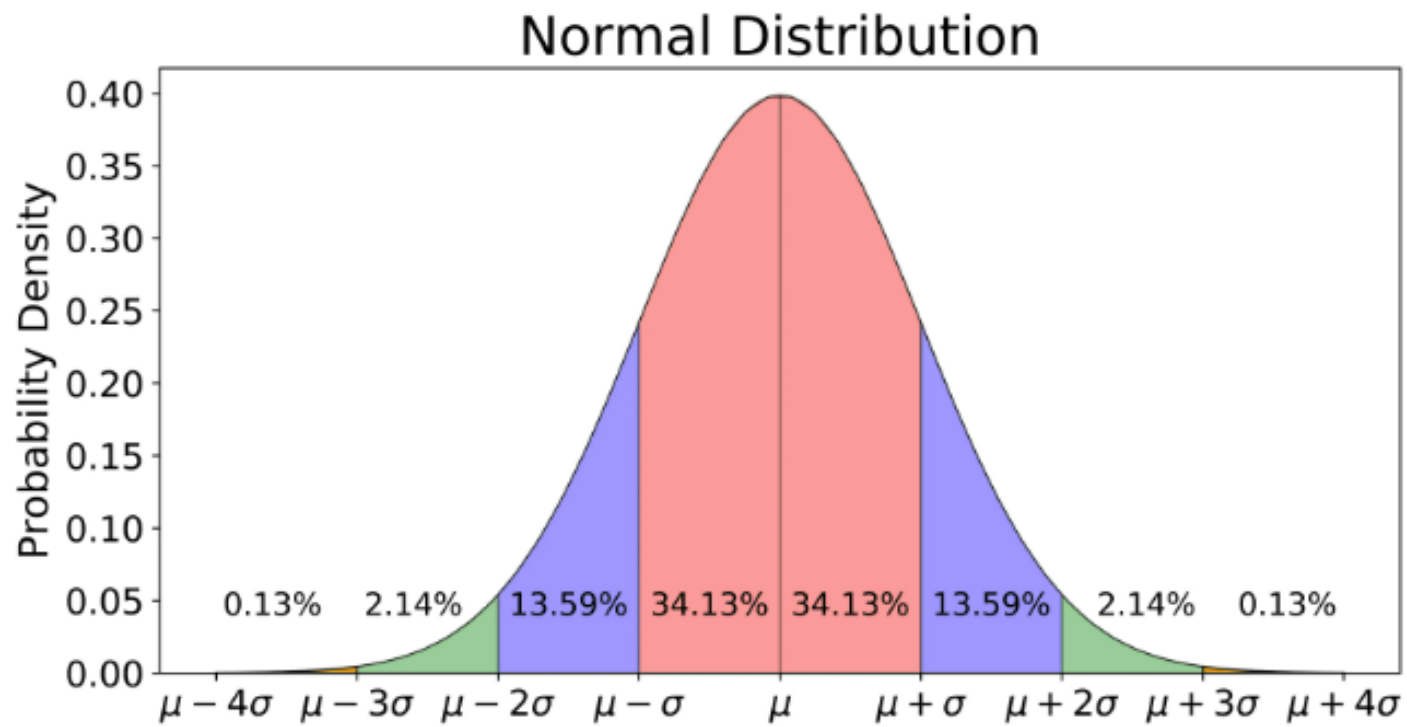
其中 $-\infty < \mu < +\infty$, $\sigma > 0$ 均为常数。称 X 服从参数为 μ, σ 的正态分布，记作 $X \sim N(\mu, \sigma^2)$ 。

μ ：均值； σ ：方差

遵从正态分布的随机变量 X ，其正态分布函数为：

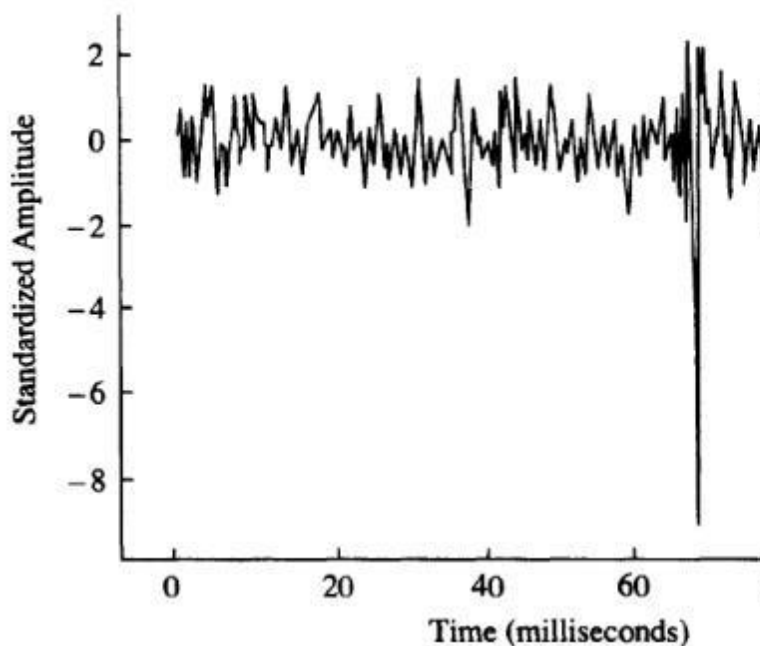
$$P(X < x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \quad -\infty < x < +\infty$$

$\mu=0$ ； $\sigma^2=1$ 时，称为标准正态分布，记为 $X \sim N(0, 1)$ 。



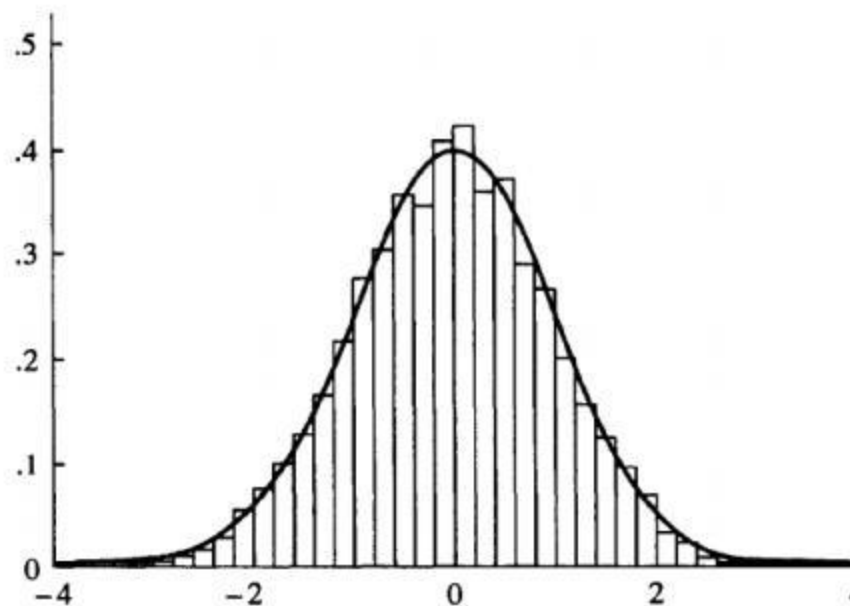
海底声纳 (sonar) 记录的声波由大量的背景噪声组成, 在北冰洋这样的背景噪声部分地由冰块碰撞和拉伸造成。Veith和Wilks (1985年)分析了一组北冰洋海底声波记录数据, 发现背景噪声由两类可区分的信号组成: 一类符合Gauss分布, 一类对应于大尺度的冲击(波)。

A record of undersea noise containing a large burst.



(Veitch J., and Wilks A. (1985). A characterization of Arctic undersea noise. *J. Acoust. Soc. Amer.*, 77: 989-999.)

A histogram from a "quiet" period of undersea noise with a fitted normal density.



——当一个量可看成由许多微小、独立的随机因素作用的总后果时，每种因素在正常状态下都不起压倒性的主导作用（故称**正态分布**），一般都服从或近似服从正态分布

——正态分布是科学和工程领域数据统计分析研究中最重要 的分布，许多统计分析方法的数学前提就是数据遵循某种正态分布

均值(mean) 或数学期望(mathematical expectation)

离散型随机变量的均值

设离散型随机变量 X 的分布律为:

$$P(X = x_i) = p_i, \quad i = 1, 2, 3, \dots$$

若

$$E(X) = \sum_{i=1}^{+\infty} x_i p_i$$

收敛, 则称 $E(X)$ 为随机变量 X 的均值或数学期望。

x_i : 质点 i 的坐标; p_i : 质点 i 的质量

——→ $E(X)$: 质心坐标

连续型随机变量的均值

设 X 为连续型随机变量，它的概率密度函数为 $p(x)$ ，若

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

收敛，则称 $E(X)$ 为随机变量 X 的均值或数学期望。

方差(variance)

设随机变量 X 的均值为 $E(X)$, 则:

$$X \text{ 的方差: } D(X) = E(X - E(X))^2$$

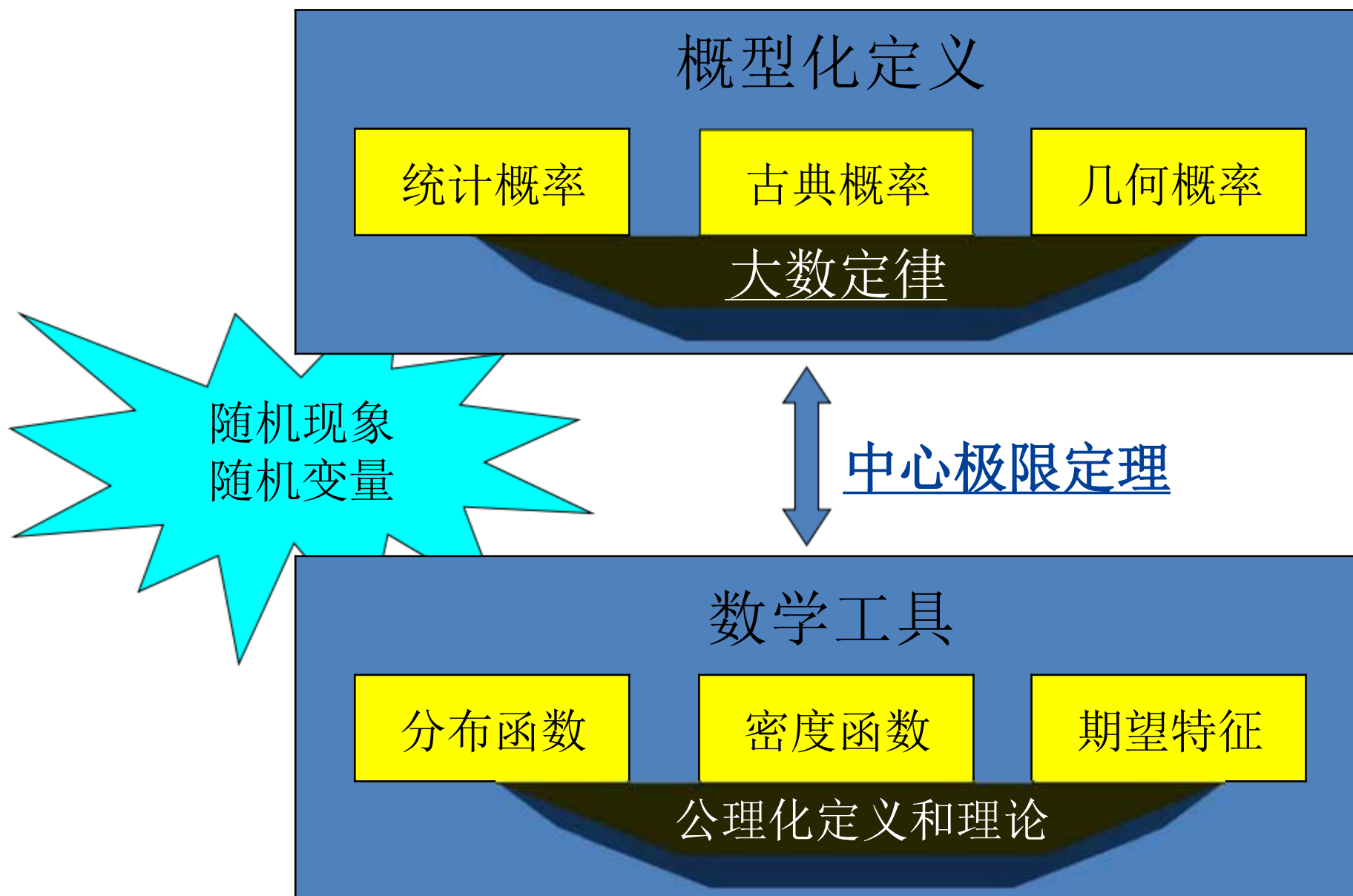
$$X \text{ 的标准差或均方差: } \sqrt{D(X)}$$

对于离散型随机变量 X , 其方差为:

$$D(X) = \sum_{i=1}^{\infty} (x_i - E(X))^2 p_i$$

对于连续型随机变量 X , 其方差为:

$$D(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 p(x) dx$$




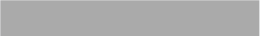
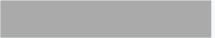








在不在玩个脑筋急转弯啊

Q:首位数字出现的频率是否是均匀分布?

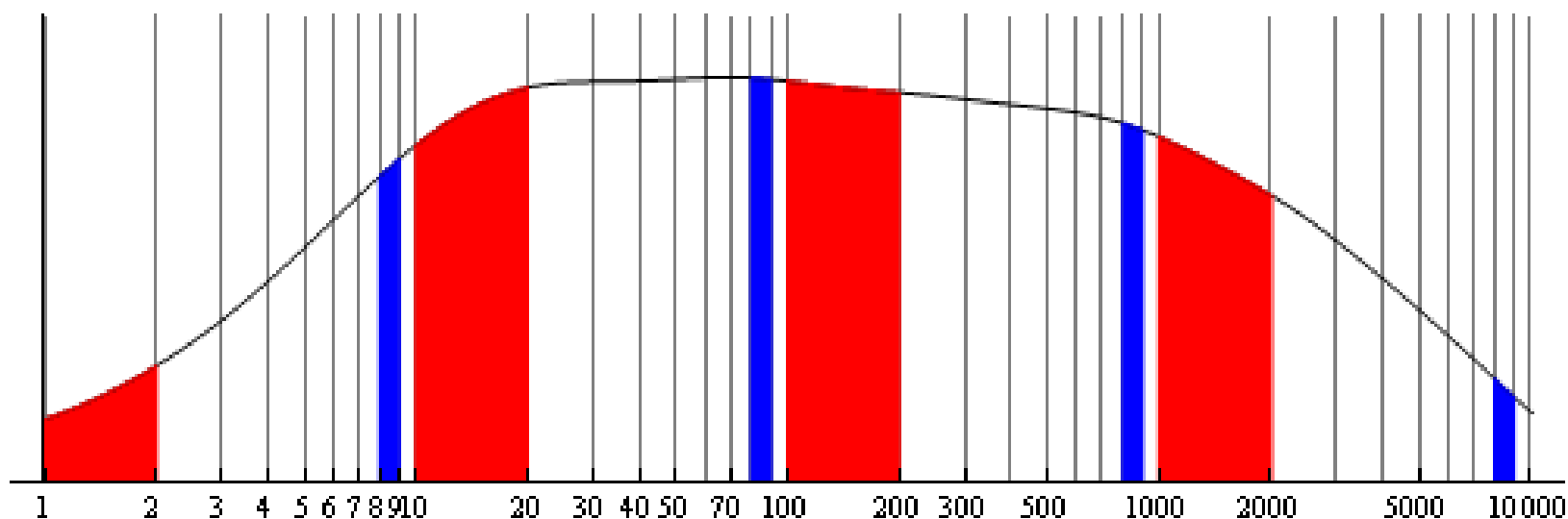
- 一堆从实际生活得出的数据中，以1为起首的数字的出现机率约为总数的三成，接近期望值 $1/9$ 的3倍。
- 推广来说，越大的数，以它起首的数出现的机率就越低。它可用于检查各种数据是否有造假。

d	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

直观理解：

- 从数数来说，从1开始到9，如果9是终点，所有数的起首机会相同
- 从9到10...19，以1为起首的数字大大抛离了其他数字
- 下一次到9起首之前，必然会经过2..8的数。
- 以任何一个数字为终点，以1为起首的出现率一般比9大。

对数坐标系下的1万以内数字的概率密度图



使用条件

- 数据至少有3000笔以上
- 不能人为操控

应用实例

- 检查各种经济数据是否有欺瞒之处，或者检查是否有假帐
- 揭露2009年伊朗总统大选的造假

问题来源

- 1881年天文学家西蒙.纽康发现以1为首的对数表的那几页比较烂
- 1938年物理学家本福特再次发现这个现象，并使用数据做了证实



(二) 统计学基础

揭示复杂现象、复杂行为和复杂过程的不确定性背后隐含的本质规律

- (1) 研究对象的随机性特性
- (2) 随机性的来源：系统的复杂性、方法的局限性
- (3) 统计学方法的有效性：提供有效的信息、确定决定系统发展变化的因素（条件）
- (4) 有助于认识复杂系统的多层次结构

设 $\{\Omega, \mathcal{F}\}$ 为可定义概率函数的可测空间, Φ 为其上的一个概率分布族, 则称三元组 $\{\Omega, \mathcal{F}, \Phi\}$ 为**统计模型** (statistical model) 或**统计结构** (statistical structure)。

设 $\{\Omega, \mathcal{F}, \Phi\}$ 和 $\{\Omega', \mathcal{F}', \Phi'\}$ 为两个统计模型, 则称

$$(\Omega \otimes \Omega', \mathcal{F} \otimes \mathcal{F}', \Phi \otimes \Phi')$$

为它们的**乘积模型**, 记为:

$$(\Omega, \mathcal{F}, \Phi) \otimes (\Omega', \mathcal{F}', \Phi')$$

类似地, 可以给出 n 个统计模型的乘积模型。特别地, n 个相同统计模型 $\{\Omega, \mathcal{F}, \Phi\}$ 的乘积模型称为**重复抽样模型**, 记为 $\{\Omega, \mathcal{F}, \Phi\}^n$ 。

乘积模型在实际中相当于独立观察系统, 重复抽样模型相当于对一个观测对象进行有限次独立抽样结果的描述。

总体 X (population)

研究对象的某种特征值的全体组成的集合。用 X 表示。

样本 X_1, X_2, \dots, X_n (sample)

在总体中选取部分有代表性的子集称为（随机）样本。
一个样本是来自总体 X 的一组相互独立、同 X 分布的随机变量。

样本值 x_1, x_2, \dots, x_n

从总体 X 随机抽取的一组观测值，常用 x_1, x_2, \dots, x_n 来表示样本或样本值。

总体参数

设总体 X 容量为 N , 个体取值为 $\{x_i\}$, $i=1,2,\dots,N$ 。定义:

总体均值 (population mean) : $\mu = \frac{1}{N} \sum_{i=1}^N x_i$

总体方差 (population variance) :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^N (x_i^2 - \mu^2)$$

总体标准差 (population standard deviation) :

$$s = \sqrt{\sigma^2}$$

随机抽样 (random sampling) 简称**抽样**:

——从总体 X 中按照一定的概率抽取若干个体来研究 X 的取值。

——抽样的理论基础: 概率理论、统计学理论

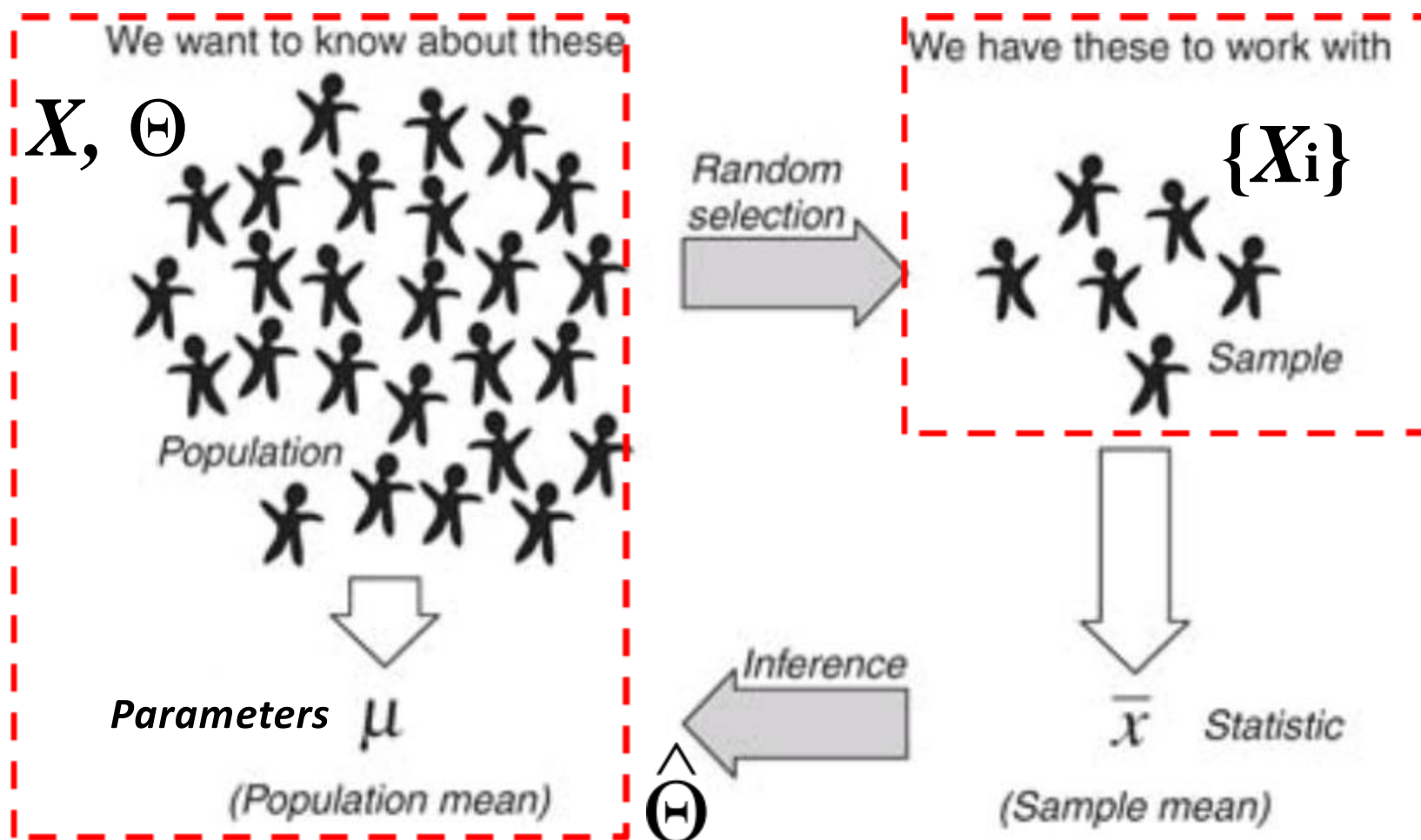
随机样本 (random sample) 简称**样本**:

——按照一定的概率从总体 $X=\{x_i\}$, $i=1, 2, \dots, N$ 中抽取作为总体代表的若干个体的集合 $\{X_1, X_2, \dots, X_n\}$, $n < N$, 称为**容量为 n 的样本**。

X
随机变量

$\{X_i\} \ i=1, 2, \dots, n$
 n 维随机变量

统计学的基本思想：总体 \rightarrow 样本 \rightarrow 总体



简单随机抽样(simple random sampling, SRS)

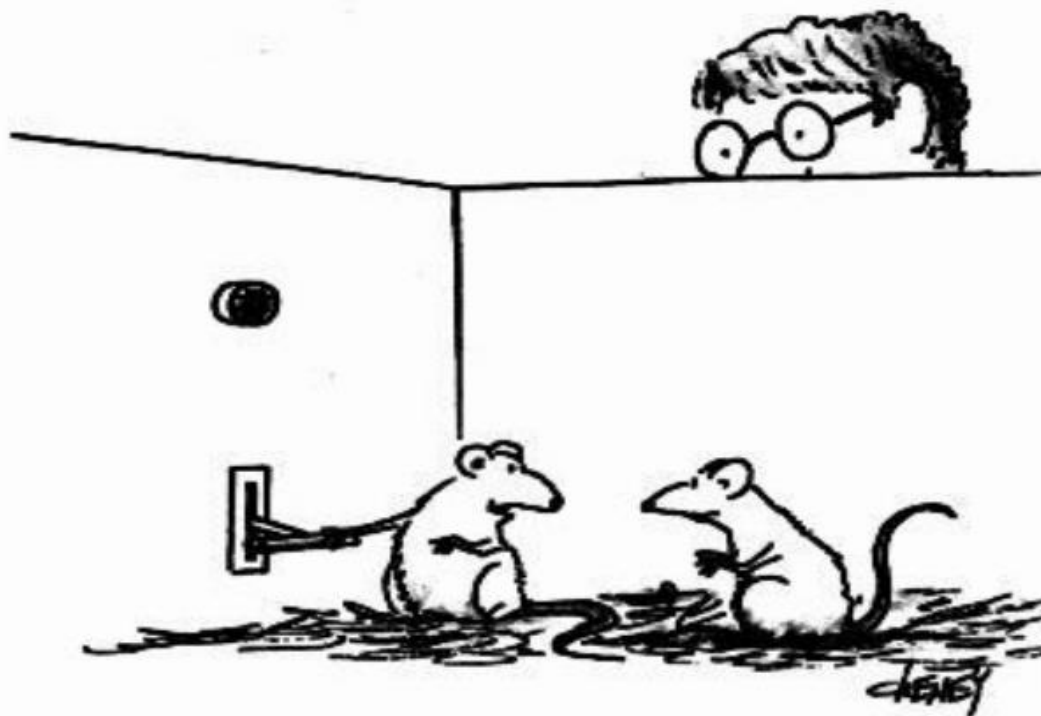
若把取自总体 X 的容量为 n 的样本看成是对同一个统计模型独立进行 n 次试验（观测）的结果，则每次试验的结果是对总体随机变量 X 的一个观测， n 次观测对应 n 个随机变量 (X_1, X_2, \dots, X_n) 的集合。

若这 n 个随机变量 (X_1, X_2, \dots, X_n) 满足两个条件（**独立同分布条件，IID (independent identical distribution) 条件**）：

- (1) n 个随机变量 (X_1, X_2, \dots, X_n) 相互独立；
- (2) n 个随机变量 (X_1, X_2, \dots, X_n) 都与总体随机变量 X 具有相同的概率分布。

则称 n 个随机变量 (X_1, X_2, \dots, X_n) 的集合为一个**简单随机样本**。该随机样本的观测值也记为 (X_1, X_2, \dots, X_n) 。

某博士生想做一个研究某种新型增强记忆的药物对老鼠学走迷宫产生影响的试验，他从实验动物中心购买了同一类型的**20**只老鼠，请问这些老鼠是否属于简单随机样本？



It's a rather interesting phenomenon. Every time I press this lever, that post-graduate student breathes a sigh of relief.

● 统计量(statistical quantity)

设 X_1, X_2, \dots, X_n 为总体 X 的一个样本, $g(x_1, x_2, \dots, x_n)$ 为连续函数, 则称 $g(x_1, x_2, \dots, x_n)$ 为一个统计量。

显然, 统计量 $g(x_1, x_2, \dots, x_n)$ 也是一个随机变量。

● 总体 X 的数字特征——参数

总体均值 μ : 刻划总体的平均取值

总体方差 σ^2 : 刻划总体取值的分散 (涨落) 程度

根据样本值推断总体性质—推断统计学

样本均值 \bar{x} :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

样本方差 s^2 :

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

显然，样本均值、样本方差都属于统计量。

通常用样本均值、样本方差作为总体均值、总体方差的无偏估计量。

无偏估计：当 n 取得充分大，样本均值、样本方差分别逼近总体均值和总体方差。

无偏样本方差：

$$s^2 = \frac{n}{n-1} \cdot \frac{N-1}{N} \cdot \tilde{s}^2 = \frac{1}{n-1} \cdot \frac{N-1}{N} \sum_{i=1}^n (X_i - \bar{X})^2$$

当 $N \gg n$ 时，或 N 未知时，就是常采用的样本方差（近似无偏样本方差）：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

参数估计问题

假定总体 X 的分布函数形式已知，对其中的某些参数进行估计。
估计方法：矩估计法、最小二乘法、最大似然法，

假设检验问题

从样本值出发，判断关于总体分布的某种假设是否成立。

假设检验问题举例

为验证一硬币是否匀称（即正反两面出现的概率是否相等），
做投掷 试验。假定试验结果有以下两个：

（1）正面55次，反面45次；

（2）正面40次，反面60次。

如何判断该硬币是否匀称？

- 1、提出原假设（或称零假设）和备选假设（或称对立假设）
原假设：硬币匀称；备选假设：硬币不匀称
- 2、指定显著性水平 α （一般取 $\alpha = 0.05, 0.01, \dots$ ）
 α 值用以衡量（或拒绝）原假设成立所需证据的指标。
 α 值越小，否定原假设的条件越高，不容易否定原假设；
 α 值越大，否定原假设的条件越低，比较容易否定原假设。
- 3、构造检验统计量 W
 X_i : 第 i 次试验的结果, $X_i = 1$ 表示出现正面, $X_i = 0$ 表示出现反面

$$Y = \sum_{i=1}^{100} X_i \quad (100 \text{ 次试验其出现正面的次为})$$

$$Z = 100 - Y \quad (100 \text{ 次试验其出现反面的次为})$$

$$W = Y - Z \quad (100 \text{ 次试验其出现正反面之差的绝对值})$$

- 4、进行统计试验——收集数据、计算检验统计量及显著性概率值 p 通常已知检验统计量 W 的概率分布性质，如：

$$p = P\{W \geq 10 \text{ 硬币匀称}\} = 0.27$$

$$p = P\{W \geq 20 \text{ 硬币匀称}\} = 0.04$$

- 5、根据显著性水平 α 值进行判断

对于第一个试验结果， $p=0.27 > \alpha (=0.05)$ ，故硬币匀称假设成立；

对于第二个试验结果， $p=0.04 < \alpha (=0.05)$ ，故硬币匀称假设不成立；



(三) 概率模型及重要公式

描述对象的不同结果具有不同的发生概率。

骰子: $p_1, p_2, p_3, p_4, p_5, p_6$

$$p_i \geq 0 \quad \sum_{i=1}^6 p_i = 1$$

连续抛掷 (独立不相关): **1 6 3 3 5**

$$p_1 \cdot p_6 \cdot p_3 \cdot p_3 \cdot p_5$$

DNA序列：4种核苷酸构成的符号序列

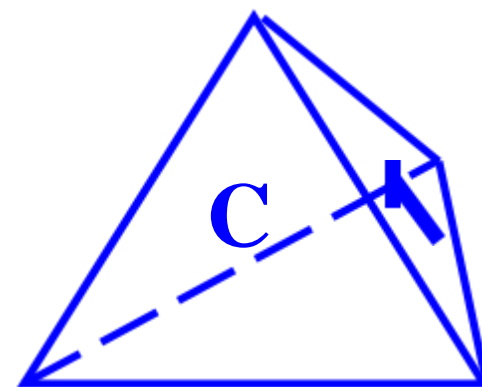
氨基酸序列：20种氨基酸构成的符号序列

序列： $x_1 x_2 \dots x_n$

$$q_{x_1} q_{x_2} \dots q_{x_n} = \prod_{i=1}^n q_{x_i}$$

真实DNA/氨基酸序列?...

随机DNA/氨基酸序列?...



基因组DNA序列上A、C、G、T的出现概率：

$$p_A p_C p_G p_T$$

设A、B为试验E的两个事件，且 $P(B) > 0$ ，称

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

联合概率

为在事件B发生的条件下，事件A发生的**条件概率**。

例子：基因组DNA序列上，CpG岛区域的GC含量显著高于非CpG岛区域

设A、B为试验E的两个事件，满足 $P(A) > 0$ 、 $P(B) > 0$ ，则

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



$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

设 B_1, B_2, \dots, B_n 为试验 E 的一组事件, 满足


$$B_i \cap B_j = \phi \quad (i \neq j)$$

$$\bigcup_{i=1}^n B_i = \Omega$$

$$P(B_i) > 0 \quad i = 1, 2, \dots, n$$

则对任一事件 A , 有

$$P(A) = \sum_{i=1}^n P(B_i) P(A | B_i)$$

$$P(A) = \sum_{i=1}^n P(AB_i) = \sum_{i=1}^n P(B_i) P(A | B_i)$$


设 B_1, B_2, \dots, B_n 为试验 E 的一组事件, 满足

$$B_i \cap B_j = \phi \quad (i \neq j)$$

$$\bigcup_{i=1}^n B_i = \Omega$$

$$P(B_i) > 0 \quad i = 1, 2, \dots, n$$

则对任一事件 A , 满足 $P(A) > 0$, 有

$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^n P(B_j)P(A | B_j)} \quad i = 1, 2, \dots, n$$

Bayes公式：后验概率公式

先验概率
(Prior probability)

似然值
(Likelihood)

$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^n P(B_j)P(A | B_j)}$$

后验概率
(Posterior probability)

讨论： 如果把事件A看作一个试验结果,把构成样本空间划分的事件组 B_1, B_2, \dots, B_n 看作导致A发生的各种原因, 则Bayes公式用于推测实验结果A发生的原因。

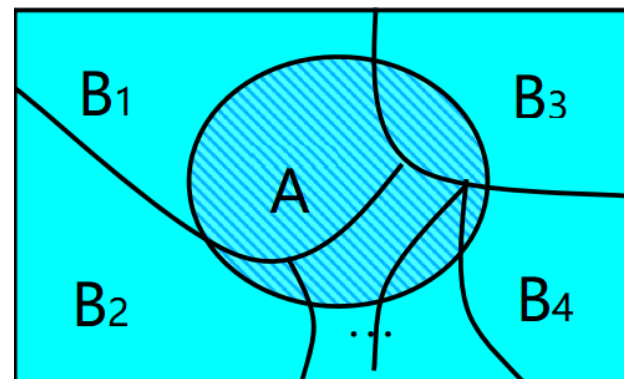
$$P(B_i | A) = \frac{P(B_i)P(A | B_i)}{\sum_{j=1}^n P(B_j)P(A | B_j)}$$

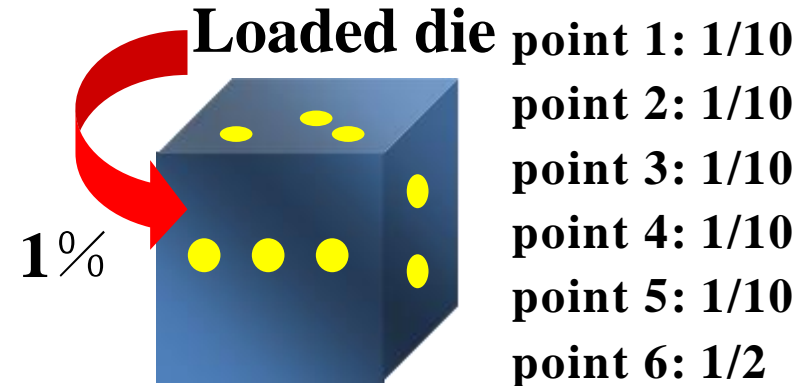
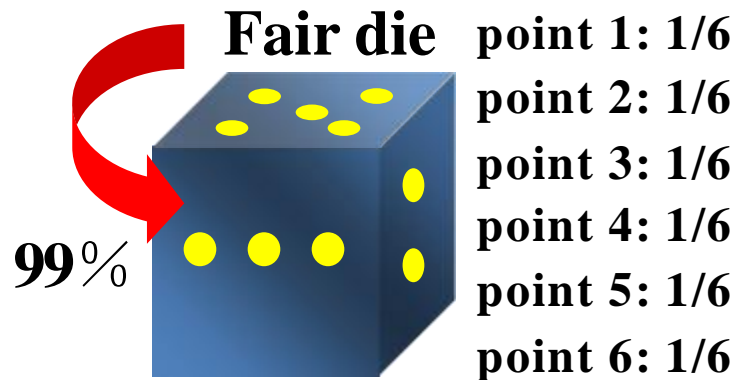
$P(B_i)$: 完备事件组 $\{B_i\}$ 的先验概率 (prior probability)

$P(A)$: 实验结果A的先验概率 (prior probability)

$P(B_i | A)$: 后验概率 (posterior probability)

Bayes概率在机器学习、人工智能、知识发现领域中有极其广泛的应用。是生物医学数据分析、生物信息学等的基本方法之一。





$$P(6 | D_{fair}) = 1 / 6 = 0.1667$$

$$P(6 | D_{loaded}) = 1 / 2 = 0.5$$

$$P(6, D_{fair}) = 0.99 \times 0.1667 = 0.165$$

$$P(6, D_{loaded}) = 0.01 \times 0.5 = 0.005$$

$$P(6) = P(6, D_{fair}) + P(6, D_{loaded}) = 0.165 + 0.005 = 0.170$$

Bayes公式的应用: 根据观察的同一骰子连续掷出的点数, 判断来自哪个骰子?

(1) 观察的点数序列为 “666”时:
已知先验概率为:

$$P(D_{fair}) = 0.99 \quad P(D_{loaded}) = 0.01$$

观察点数序列为 “666”的似然值:

$$P('666' | D_{fair}) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = 0.0046$$

$$P('666' | D_{loaded}) = 0.5 \times 0.5 \times 0.5 = 0.125$$

观察点数序列为 “666” 的后验概率 (Bayes 概率) 为:

$$P(D_{\text{fair}} | '666') = \frac{\left(\frac{1}{6}\right)^3 \times 0.99}{0.5^3 \times 0.01 + \left(\frac{1}{6}\right)^3 \times 0.99} = 0.79$$

$$P(D_{\text{loaded}} | '666') = \frac{0.5^3 \times 0.01}{0.5^3 \times 0.01 + \left(\frac{1}{6}\right)^3 \times 0.99} = 0.21$$

(2) 观察的点数序列为至少多少个连续的 “6” 时, 可以判定来自作弊的骰子?

(直接计算分子作比较即可)

训练数据集 D
(Training set)



频率
 (f_1, f_2, \dots, f_n)



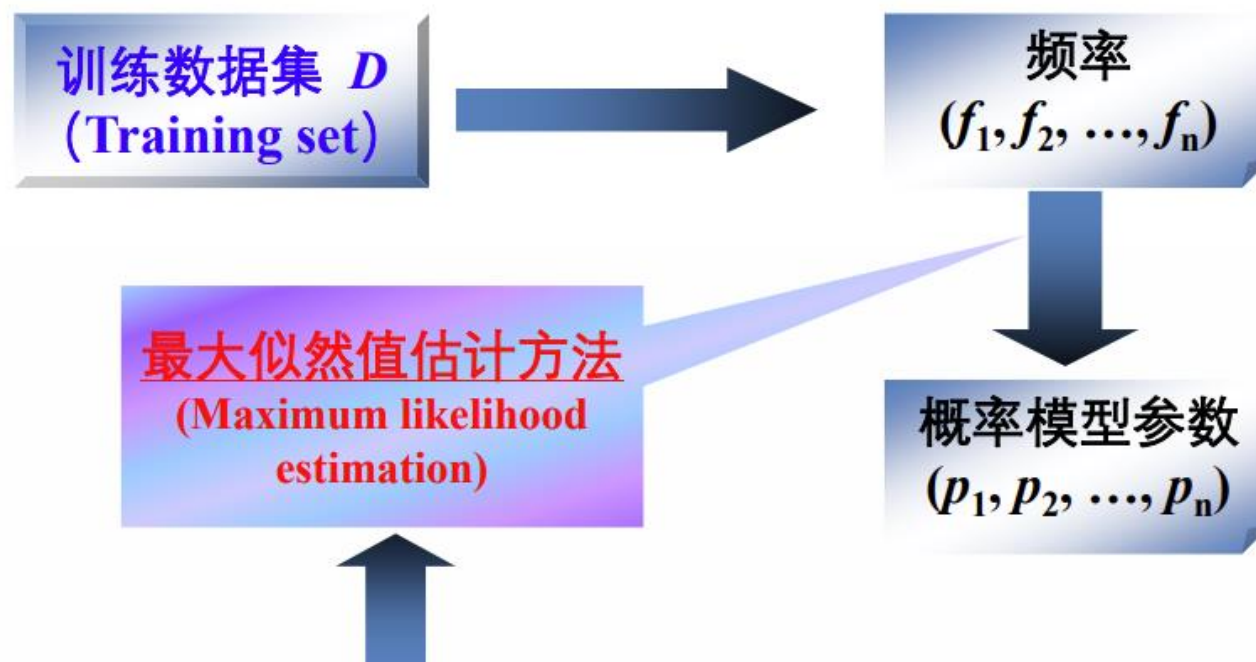
概率模型参数
 (p_1, p_2, \dots, p_n)

对于已知的训练数据集 D , 我们的目标是构造并确定它的概率模型参数:

$$\vec{p} = (p_1, p_2, \dots, p_n)$$

例如对于某一基因组DNA序列:

$$\vec{p} = (p_A, p_C, p_G, p_T)$$



$$\vec{p} = (p_1, p_2, \dots, p_n)$$

$$\vec{p}^{ML} = \operatorname{argmax}_{\vec{p}} P(D | p) = (f_1, f_2, \dots, f_n)$$

讨 论

训练集的推广性：根据训练集**D**得到的概率模型参数，能否同样适用于新的数据集？

特点：处理训练集数据较完整、数据集较大的情况。

训练集的不足 → 参数的不合理

依靠先验知识进行修正 → pseudocount

例如：观察的骰子点数序列： 1 3 4 2 4 6 2 1 2 2

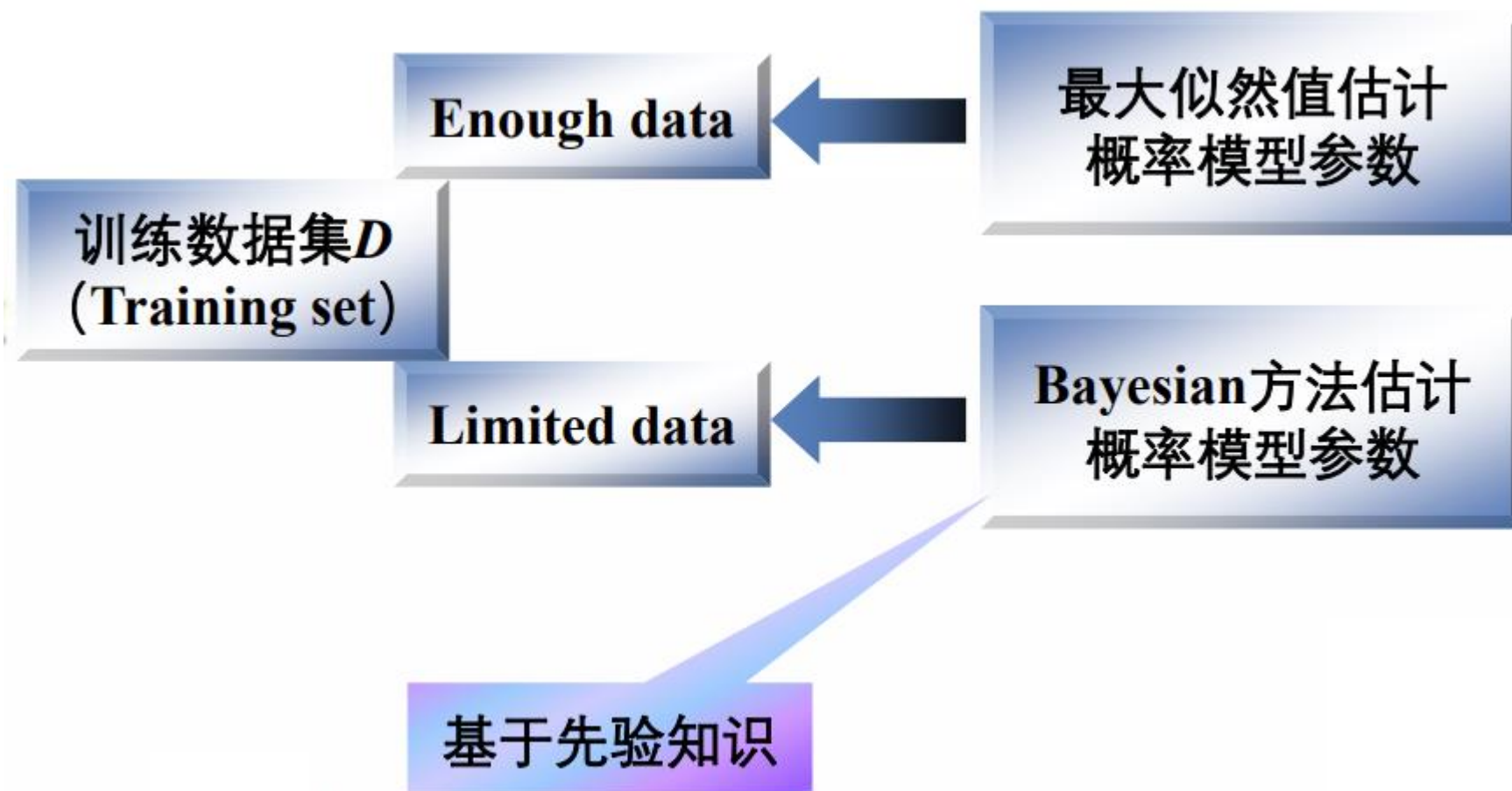
$$p_1 = 2/10 = 0.2 \quad p_2 = 4/10 = 0.4 \quad p_3 = 1/10 = 0.1$$

$$p_4 = 2/10 = 0.2 \quad p_5 = 0 \quad p_6 = 1/10 = 0.1$$



$$p_1 = 3/16 = 0.1875 \quad p_2 = 5/16 = 0.3125 \quad p_3 = 2/16 = 0.125$$

$$p_4 = 3/16 = 0.1875 \quad p_5 = 1/16 = 0.0625 \quad p_6 = 2/16 = 0.125$$



对于已知的训练数据集 \mathbf{D} ，我们的目标是构造它的概率模型参数

$$\vec{p} = (p_1, p_2, \dots, p_n)$$

由Bayes公式，得到计算任一概率模型参数 (p_1, p_2, \dots, p_n) 的后验概率：

$$P(\vec{p} | D) = \frac{P(\vec{p})P(D | \vec{p})}{P(D)}$$

其中，

$$P(D) = \int_{\vec{p}'} P(\vec{p}') P(D | \vec{p}') d\vec{p}'$$

先验估计出合理的
概率分布

$$P(\vec{p} | D) = \frac{P(\vec{p})P(D | \vec{p})}{\int_{\vec{p}'} P(\vec{p}')P(D | \vec{p}')d\vec{p}'}$$

如何选取估计值?

MAP估计 (Maximum a posterior estimation) :

$$\vec{p}^{MAP} = \arg \max_{\vec{p}} P(\vec{p} | D)$$

实际上相当于分子取最大值。

注：当 (p_1, p_2, \dots, p_n) 均匀分布时，等效于ML估计方法。



(四) 向量、矩阵和线性代数初步

向量:

既有大小又有方向, 由模 (长度) 和方向两个变量来确定。
向量是向量空间的成员。

直观理解:

一个向量就是数值的有序序列: $\mathbf{X}=\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$

一个 n 维向量就是 n 维向量空间的一个点。

例: 一个基因组序列的碱基含量 f_A, f_C, f_G, f_T ,
构成向量 $\mathbf{F}=\{f_A, f_C, f_G\}$

向量的基本运算:

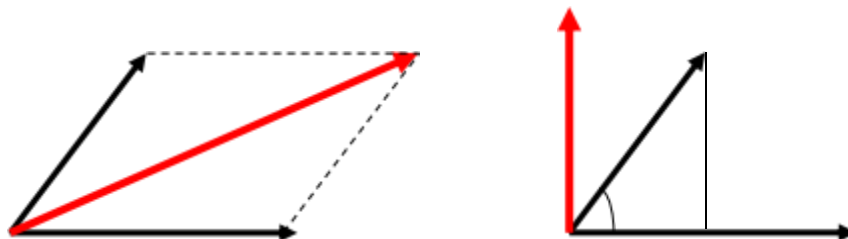
长度: $|X| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$

加法: $X+Y$

数乘: $\lambda \cdot X$

内积 (点积): $X \cdot Y = \sum_i x_i y_i$

叉积 $X \times Y = X Y \sin(X, Y) e$ e : 垂直于 X, Y 的单位向量



矩阵:

矩形阵列，排列成行和列的值

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \dots & & x_{ij} & \\ x_{M1} & x_{M2} & \dots & x_{MN} \end{bmatrix}$$

- (1) 矩阵使得诸多数学、物理问题大为简化
- (2) 矩阵使得线性代数方程组、线性微分方程组、以及偏微分方程的数值求解问题大为简化

矩阵的基本运算：

矩阵相加（减）：仅当二者行数、列数均相等时才有意义

$$X + Y = \begin{bmatrix} x_{ij} \end{bmatrix} + \begin{bmatrix} y_{ij} \end{bmatrix} = \begin{bmatrix} x_{ij} + y_{ij} \end{bmatrix}$$

矩阵的数乘：

$$\lambda \cdot X = \begin{bmatrix} \lambda x_{ij} \end{bmatrix}$$

矩阵的乘法：**X**为**M**行**N**列，**Y**为**N**行**K**列，乘法得到的**Z**为**M**行**K**列

$$Z = XY = \begin{bmatrix} x_{mn} \end{bmatrix} \begin{bmatrix} y_{nk} \end{bmatrix} = \begin{bmatrix} \sum_n x_{mn} y_{nk} \end{bmatrix} = \begin{bmatrix} z_{mk} \end{bmatrix}$$

矩阵的转置：行列进行交换

线性变换和线性方程组的矩阵表示：

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1N}x_N = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2N}x_N = b_2$$

.....

$$a_{N1}x_1 + a_{N2}x_2 + \dots + a_{NN}x_N = b_N$$



$$AX = B$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \dots & & \dots & \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix},$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}$$

矩阵求逆：相当于矩阵的除法， I 为单位矩阵

$$AA^{-1} = A^{-1}A = I$$

矩阵求逆的步骤：初等变换法（略）

求解线性代数方程组：

$$A^{-1}AX = IX = X = A^{-1}B$$

$$X = A^{-1}B$$

逆矩阵法求解线性代数方程组：

高斯消去法（略）