概率

量化随机

小胖

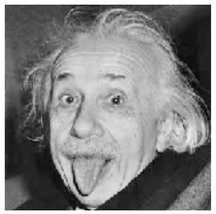
目录

概率的定义

什么是概率

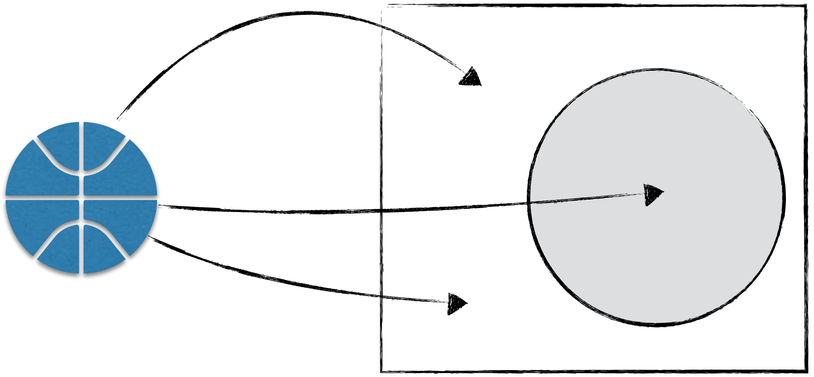
在现实世界里，充满了各种随机事件

- 彩票的中奖、掷骰子的点数
- 微观世界里，粒子出现的位置



概率是用来刻画随机的一种数学工具

球落入圆圈的概率 = 圆圈面积 / 方框面积



随机是一个很模糊的哲学概念

概率又是数学领域人们理解最差的一个分支

数学上的概率 ≠ 人类语言中的概率

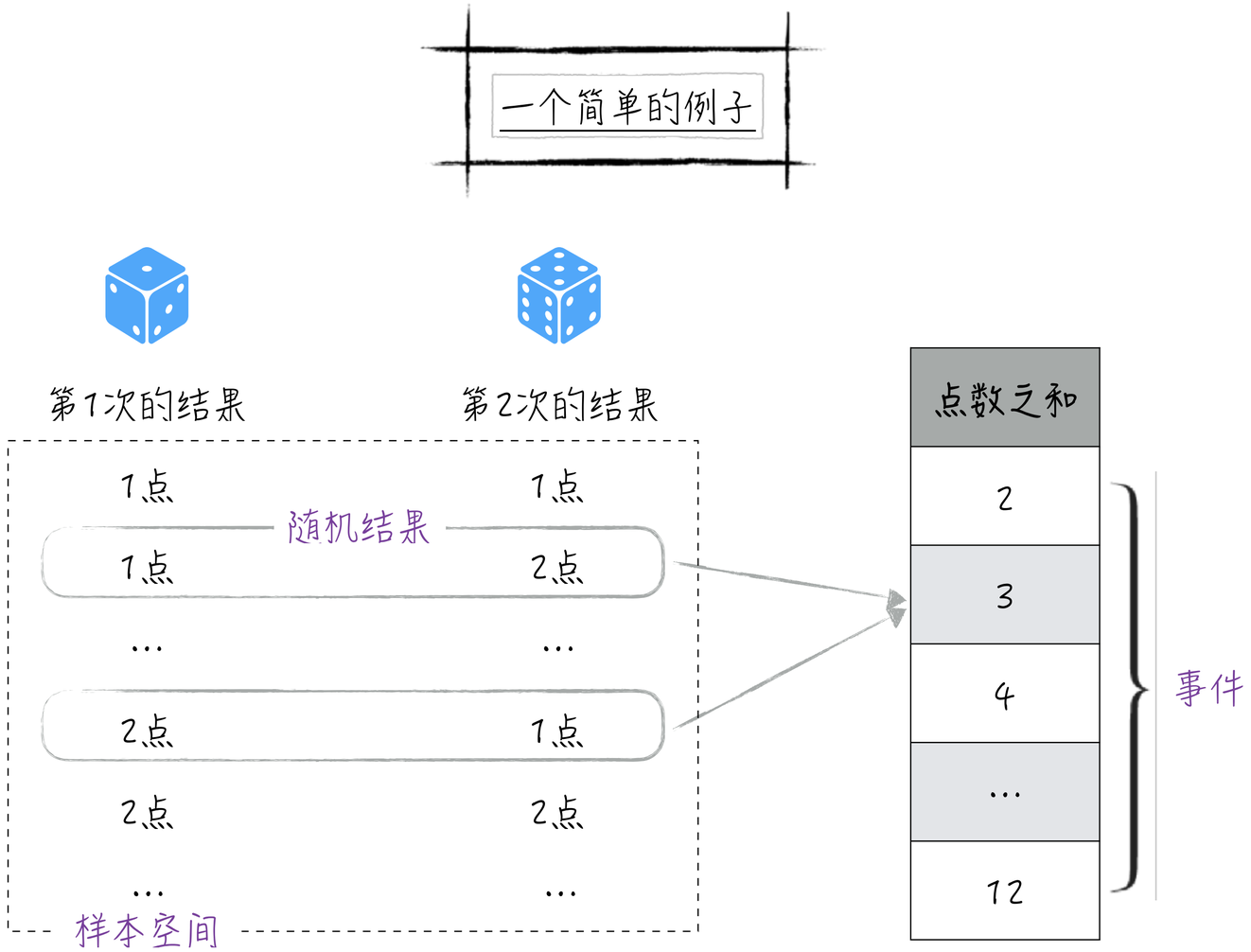
假设我们将从自然数里面随机选出一个数字

选择1的“概率”是多少？

这个“概率”，数学上无法定义

概率的定义

事件与概率空间



在随机结果有限的情况下：

- 定义样本空间S：所有随机结果ω组成的集合
- 定义概率：满足如下三个条件的，从样本空间到实数的函数

$$P(\omega) \geq 0$$

$$\sum_{\omega \in S} P(\omega) = 1$$

$$P(E) = \sum_{\omega \in E} P(\omega)$$，其中E为任意一个事件

概率的定义

事件与概率空间

在随机结果有限的情况下：

- 定义样本空间S：所有随机结果ω组成的集合
- 定义概率：满足如下三个条件的，从样本空间到实数的函数
  - $P(\omega) \geq 0$
  - $\sum_{\omega \in S} P(\omega) = 1$
  - $P(E) = \sum_{\omega \in E} P(\omega)$，其中E为任意一个事件

以上的定义可以推广到随机结果无限的情况

数学上的概率就是满足柯尔莫果洛夫公理（Kolmogorov axioms）的函数

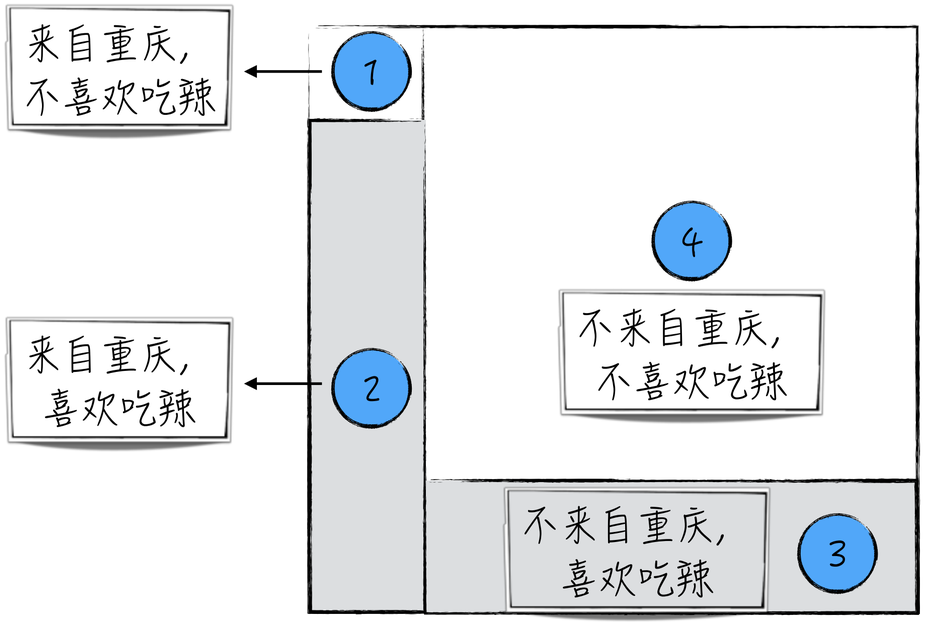
- 对于任意一个事件E，则$P(E) \geq 0$；
- 对于所有可能事件的集合Ω，则$P(\Omega) = 1$；
- 任意两两互不相交的事件可数序列$E_1, E_2, \ldots$，则$P(E_1 \cup E_2 \cup \ldots) = \sum P(E_i)$。

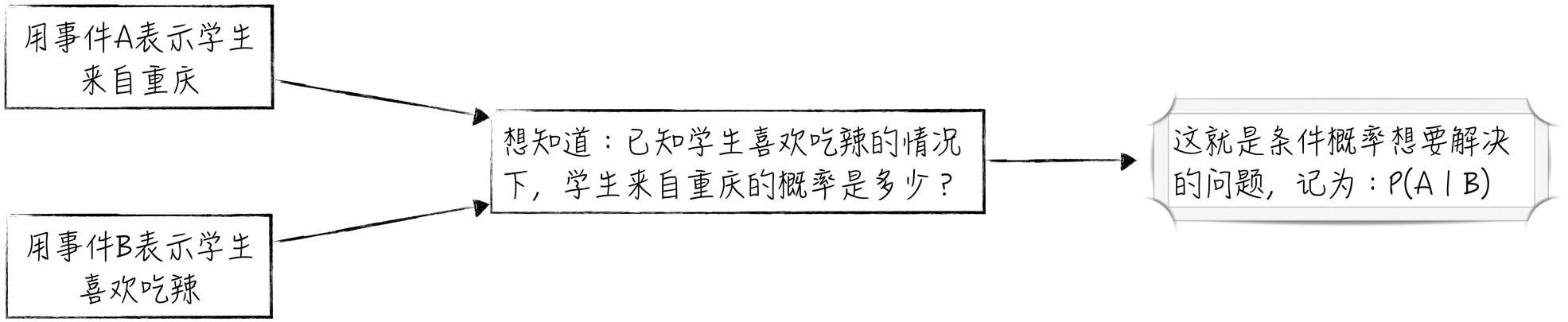
概率的定义

条件概率

假设在一个大学班级里：

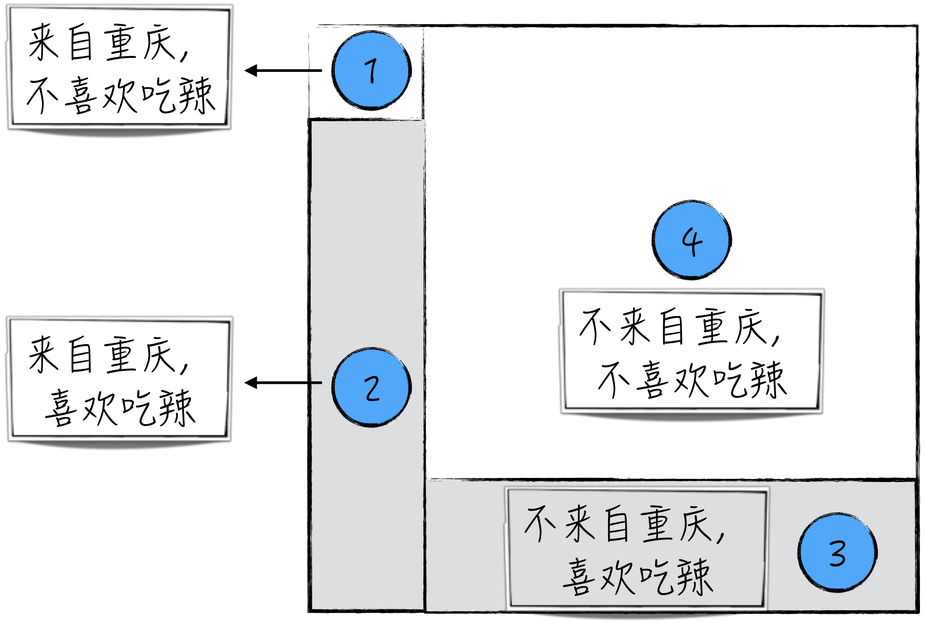
- 来自重庆的学生比例为10%，而这批学生中喜欢吃辣的比例为90%
- 来自其他省份的学生比例为90%，而这些学生中喜欢吃辣的比例为10%





概率的定义

条件概率



来自重庆的概率：$P(A)=(1+2)\div(3+4+1+2)$

喜欢吃辣的条件下，来自重庆的概率：$P(A|B)=2\div(2+3)$

对于两个事件A、B

- 两个事件同时发生记为$A \cap B$
- 定义已知B发生的情况下，A发生的条件概率为

$$P(A \mid B)=\frac{P(A \cap B)}{P(B)}$$

同理可以得到：

$$P(B \mid A)=\frac{P(A \cap B)}{P(A)}$$

结合上面两个条件概率的定义，可以得到如下的贝叶斯公式

$$P(B \mid A)=\frac{P(A \mid B)P(B)}{P(A)}$$

概率的定义

条件概率

再次回顾问题的描述

假设在一个大学班级里：

- 来自重庆的学生比例为10%，而这批学生中喜欢吃辣的比例为90%
- 来自其他省份的学生比例为90%，而这些学生中喜欢吃辣的比例为10%

记号描述

用事件A表示学生来自重庆

用事件B表示学生喜欢吃辣

翻译成数学语言

已知：

$$P(A)=0.1, P(B|A)=0.9$$

$$P(A^C)=0.9, P(B|A^c)=0.1$$

求解

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

$P(A)=0.1$

已知一个学生喜欢吃辣这一条信息

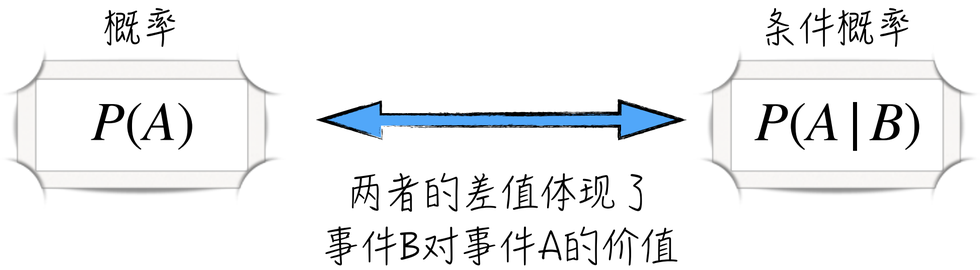
他是重庆的人概率从10%上升到25%

量化信息的价值

$P(A|B)=0.5$

概率的定义

独立事件



若 $P(A) = P(A \mid B)$ 则称A、B为相互独立事件

$$P(A \cap B) = P(A)P(B)$$

同理可以定义任意多个相互独立的事件

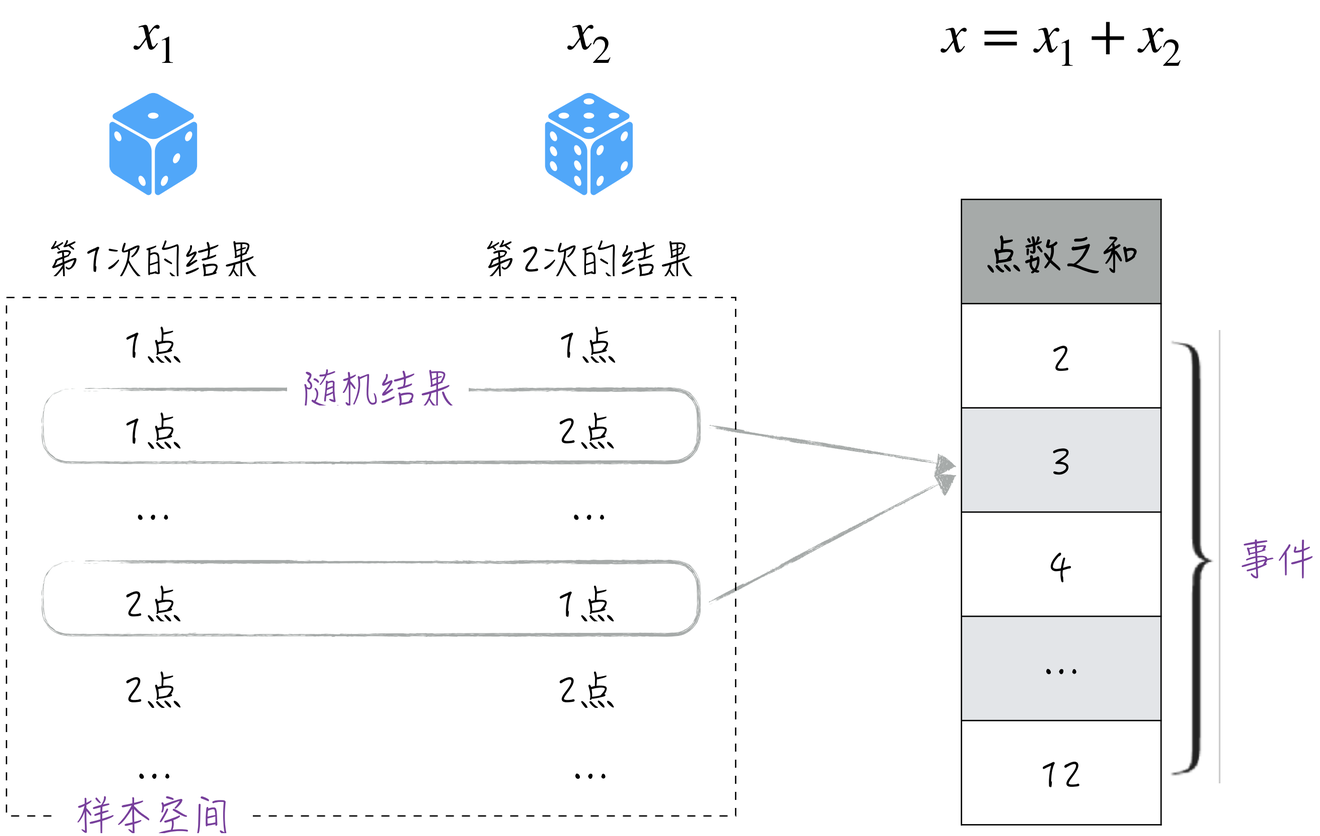
事件A1,…, An相互独立，当且仅当，对于任意子集，Ai1, …, Aik，都成立

$$P(A_{i1} \cap \ldots A_{ik}) = P(A_{i1}) \ldots P(A_{ik})$$

目录

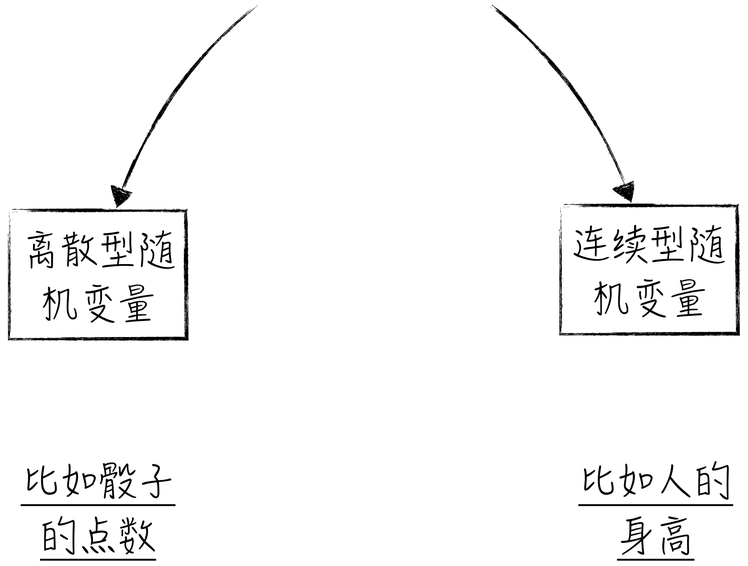
随机变量

随机变量种类



直观上，随机变量可以理解为取值为随机数的变量

严谨的定义：随机变量是将随机事件映射成实数的可测函数



随机变量

刻画随机变量的方法

离散型随机变量

x_1

对于离散型的随机变量，使用概率分布函数来刻画它

$$P(x_1 = i) = 1/6; i = 1,...,6$$

连续型随机变量

对于连续型的随机变量，使用概率密度函数来刻画它

$$P(a \leq x \leq b) = \int_a^b f_x(t)dt$$

累积分布函数

$$C_x(a) = P(x <= a)$$

期望

$$E[x] = \sum_i P(x = x_i)x_i$$

$$E[x] = \int xf(x)dx$$

方差

$$Var(x) = E[(x - E[x])^2]$$
$$= E[x^2] - (E[x])^2$$

协方差

$$Cov(x, y) = E[(x - E[x])(y - E[y])]$$
$$= E[xy] - E[x]E[y]$$

随机变量

刻画随机变量的方法

对于期望、方差、协方差有如下的性质：

$$E[x+y] = E[x] + E[y]$$

$$Cov(x,x) = Var(x)$$

$$Var(ax+by) = a^2Var(x) + b^2Var(y) + 2abCov(x,y)$$

如果x，y是两个相互独立的随机变量

$$P(x=a, y=b) = P(x=a)P(y=b) \quad \text{或者} \quad f_{xy} = f_x f_y$$

这个时候，有如下的性质

$$E[xy] = E[x]E[y] \qquad Cov(x,y) = 0$$

$$Var(ax+by) = a^2Var(x) + b^2Var(y)$$

目录

正态分布与P-value

正态分布、中心极限定理

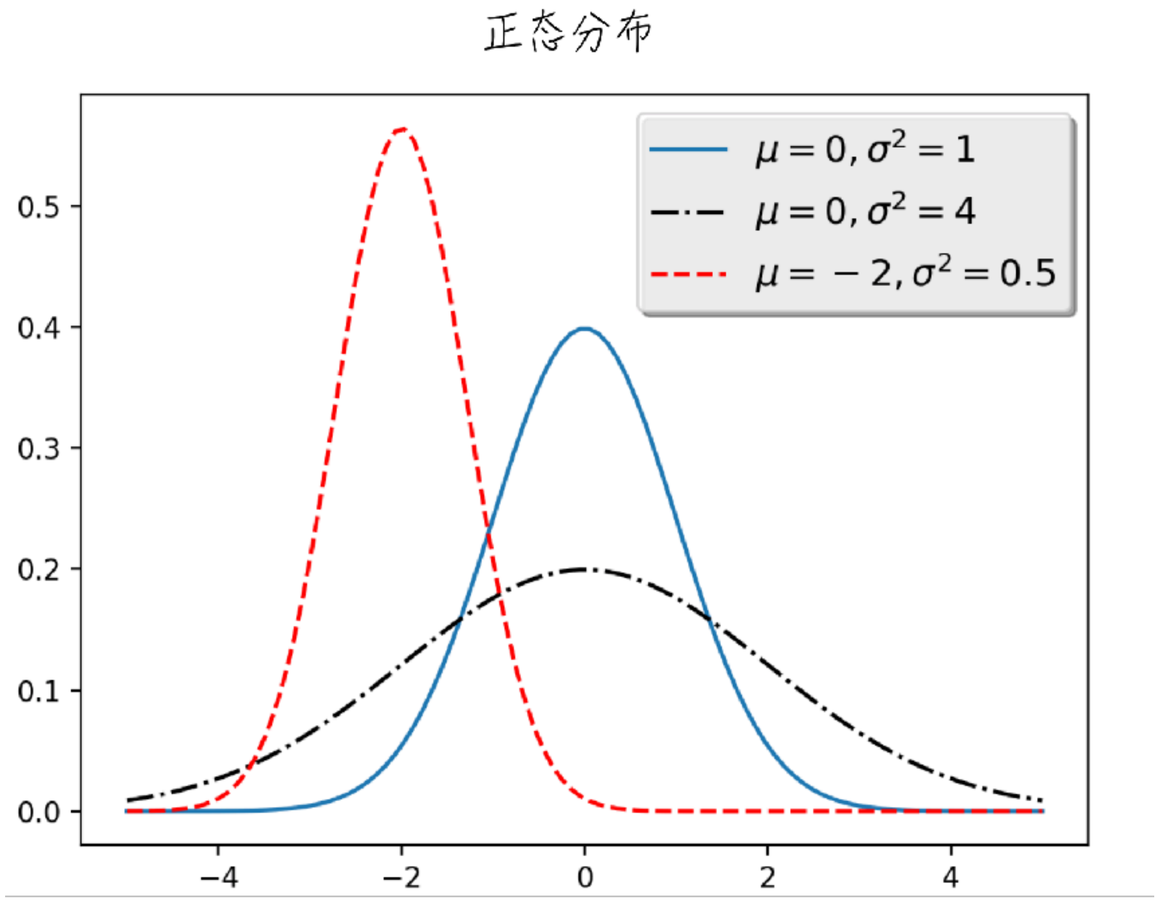
正态分布也称高斯分布，是最为重要的一种概率分布

- 如果一个随机变量服从正态分布，则它可能的取值是任意实数
- 相应的概率密度函数如下

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

容易证明：

$$E[x] = \mu, Var(x) = \sigma^2$$

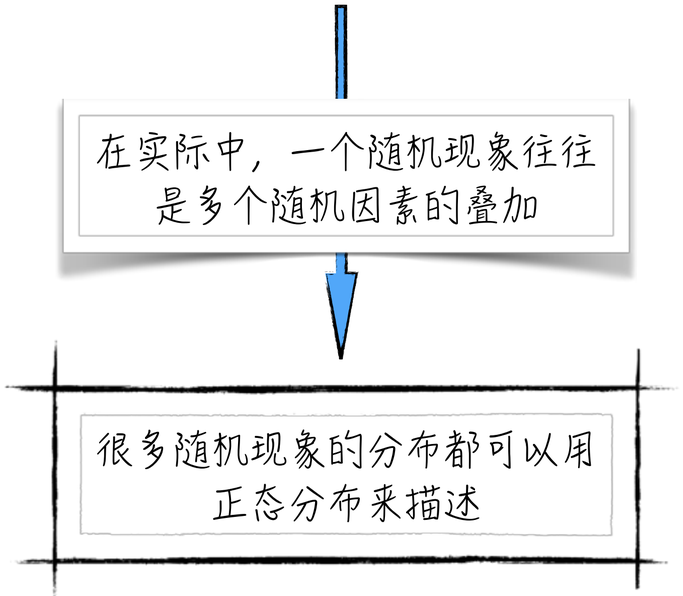


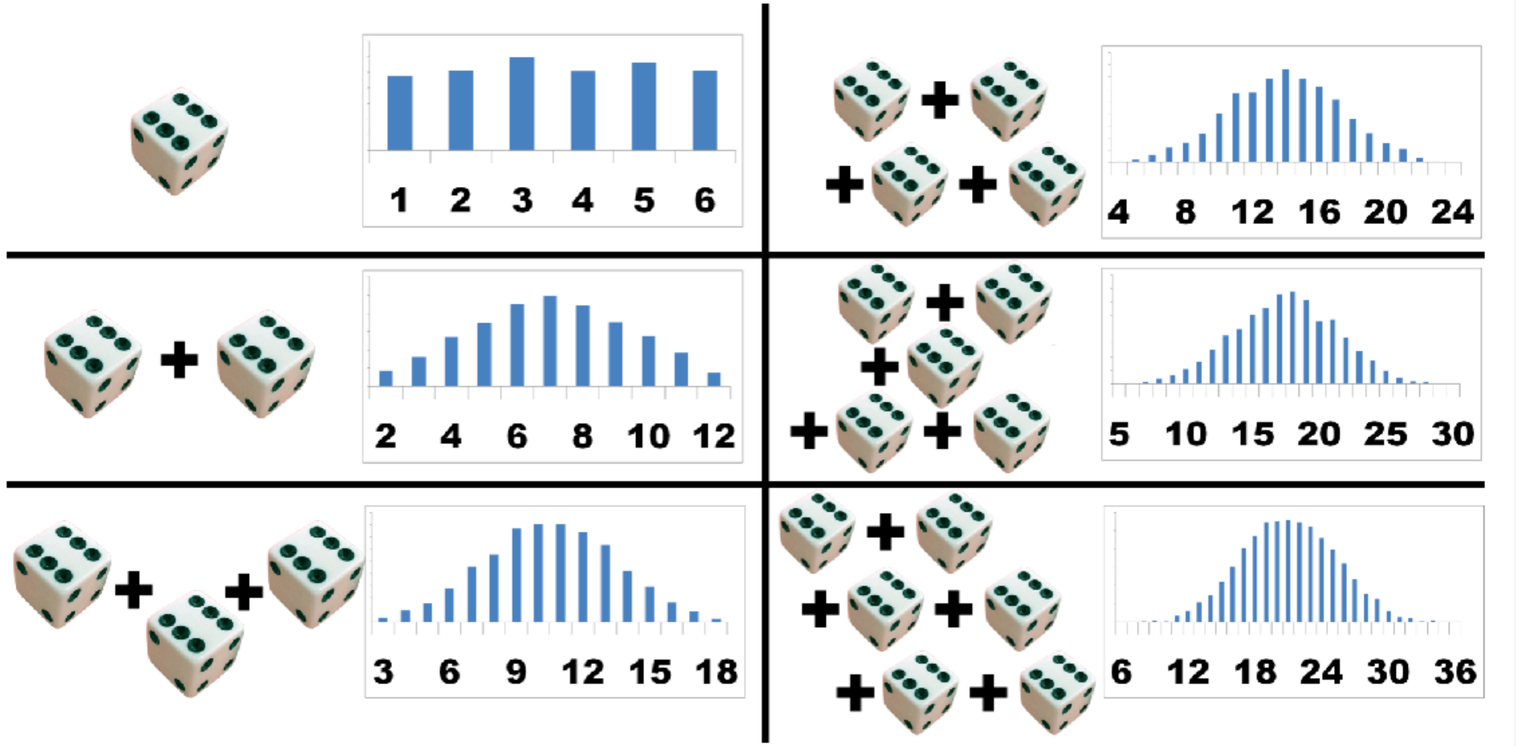
正态分布与P-value

正态分布、中心极限定理

中心极限定理

n个独立同分布的随机变量相叠加，得到的和将越来越近似于一个**正态分布**



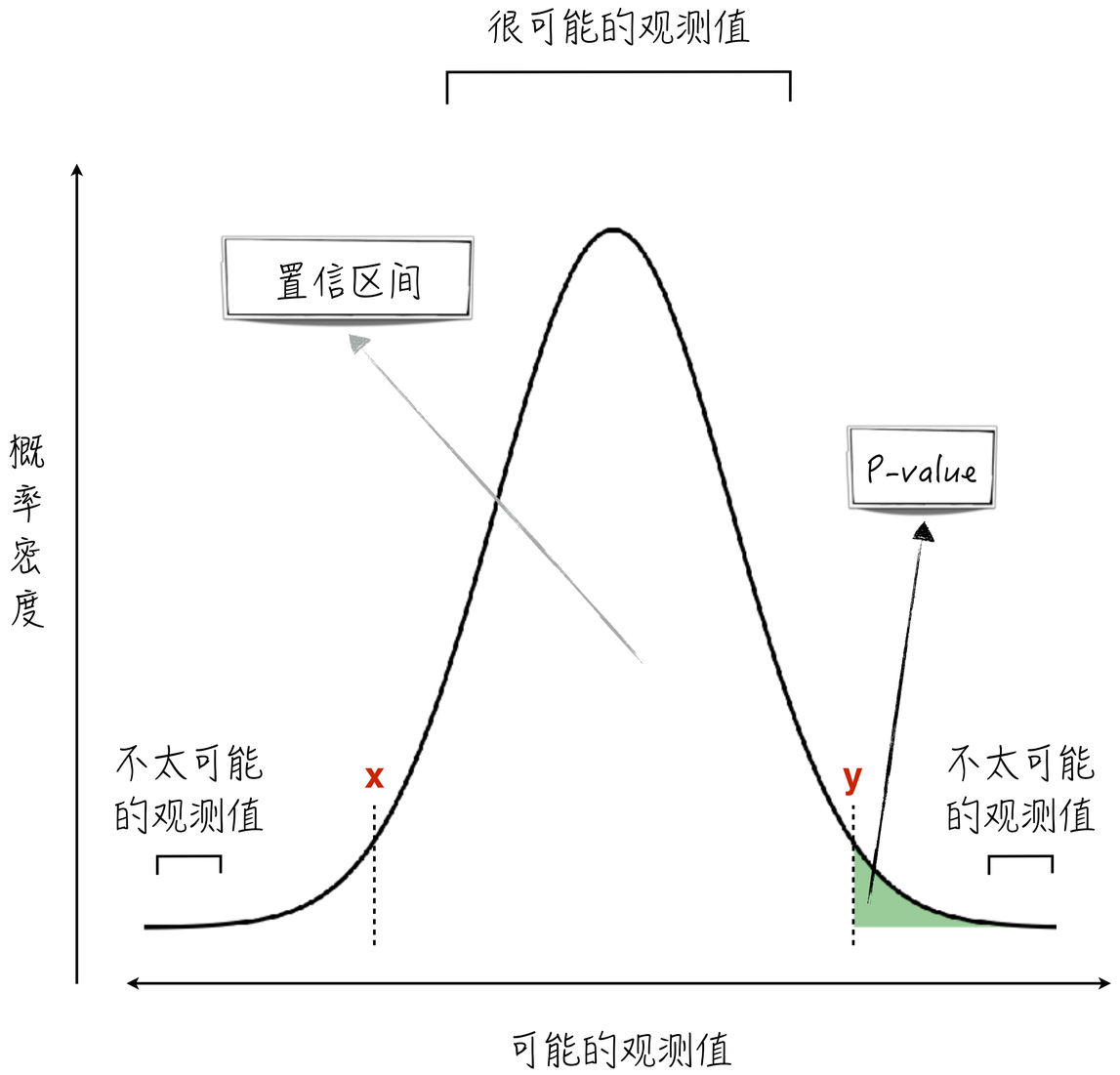


正态分布与P-value

置信区间与P-value

置信区间：概率值等于a，且以期望为中心的对称区域（在实际中a常常等于0.95或者0.99）

对于置信区间的两个边界值x，y，它们的P-value为(1 - a)/2



THANK YOU