数据科学面试

概率论技巧

斯坦福大学统计硕士

FLAG数据科学家

七月在线

Outline

- 数据科学面试
  - 概率论考点
- 古典概型
  - 原理
  - 例题
- 几何概型
  - 原理
  - 例题
- 数学期望
  - 原理
  - 例题
- 贝叶斯
  - 原理
  - 例题
- 练习题

- **Q:** 为什么数据科学面试中会考察概率论问题?

- **Q:** 为什么数据科学面试中会考察概率论问题?
- 工作需要掌握概率论?
- 聪明不聪明?
- 数学基础好不好?
- 概率论题往往较难，可以卡住不少求职者，方便刷人

- **Q:** 数据科学面试中的概率论问题有哪些特点?

- **Q:** 数据科学面试中的概率论问题有哪些特点?
- 题面不复杂
- 题目可以在很短时间给出解答
- 会者不难

Outline

古典概型

原理

- 在一个有限的集合S中随机抽取一个元素，求属于子集T的概率
- 概率为$\frac{|T|}{|S|}$
- 均匀的骰子掷到1的概率:
  $S=\{1,2,3,4,5,6\}, T=\{1\} \Rightarrow \frac{|T|}{|S|}=\frac{1}{6}$
- 均匀的骰子掷到奇数的概率:
  $S=\{1,2,3,4,5,6\}, T=\{1,3,5\} \Rightarrow \frac{|T|}{|S|}=\frac{3}{6}=\frac{1}{2}$

Outline

七月在线 JULYEDU.COM

- **Q:** 54张牌，分成6份，每份9张牌，大小王在一起的概率(美团点评)

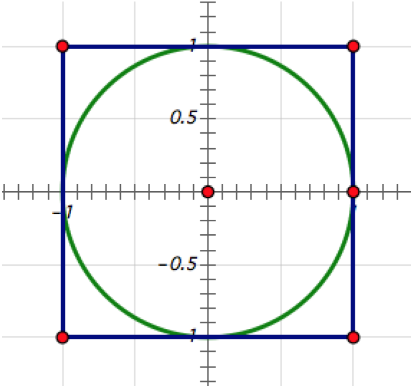
七月在线 JULYEDU.COM

- **Q:** 54张牌，分成6份，每份9张牌，大小王在一起的概率(美团点评)
- $1,2,\cdots,9;10,11,\cdots,18;\cdots;46,47,\cdots,54$
- 将54张牌放入1-54的方法总数为54!
- 其中大小王均属于1-9的方法总数为$9\times8\times52!$
- 大小王均属于10-18的方法总数为$9\times8\times52!,\cdots$
- 大小王在一起的方法总数为$6\times9\times8\times52!$
- 概率为$\frac{6\times9\times8\times52!}{54!}=\frac{8}{53}$
- 巧妙解法?

- **Q:** 54张牌，分成6份，每份9张牌，大小王在一起的概率(美团点评)
- $1, 2, \cdots, 9; 10, 11, \cdots, 18; \cdots; 46, 47, \cdots, 54$
- 将54张牌放入1-54的方法总数为54!
- 其中大小王均属于1-9的方法总数为$9 \times 8 \times 52!$
- 大小王均属于10-18的方法总数为$9 \times 8 \times 52!, \cdots$
- 大小王在一起的方法总数为$6 \times 9 \times 8 \times 52!$
- 概率为$\frac{6 \times 9 \times 8 \times 52!}{54!} = \frac{8}{53}$
- 巧妙解法?
- 固定大王位置，小王和大王在一起的选择有8种，总共有53个位置可以选择。

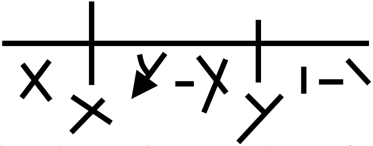
Outline

- 在一个几何形状S中随机抽取一个点，求该点属于子形状T的概率
- 概率为$\frac{|T|}{|S|}$
- 在一个边长为2的正方形内抽取一个点，该点属于其内切单位圆的概率为$\frac{\pi}{4}$

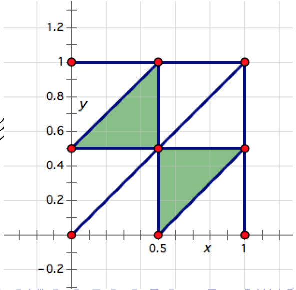


Outline

- **Q:** 一根棍子折三段能组成三角形的概率 (美团点评、百度、Google)



- **Q:** 一根棍子折三段能组成三角形的概率 (美团点评、百度、Google)
- 设棍子长度为1，折在了X和Y，X和Y独立且服从$[0,1]$上的均匀分布, (X,Y)为单位正方形$[0,1]\times[0,1]$上的均匀一点
- 构成三角形的条件是每一段的长度都小于$\frac{1}{2}$
- 考虑$X<Y$的区域，即左上半部分，构成三角形等价于$X<\frac{1}{2}, Y-X<\frac{1}{2}, Y>\frac{1}{2}$
  对应由$(0,\frac{1}{2}),(\frac{1}{2},\frac{1}{2}),(\frac{1}{2},1)$构成的三角形
  面积为$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$
- 同理，$X>Y$的区域对应面积为$\frac{1}{8}$的三角形
- 单位正方形的面积为1
- 概率为$\frac{1}{4}$



Outline

- 直观: 数学期望可以理解为平均值，一个随机变量多次采样后的平均值一般来说会很接近其期望
- 定义: 一个离散随机变量X的数学期望为$\mathbb{E}[X]=\sum_x xp(x)$，其中$p(x)$为$x$处的密度函数
- 以$\frac{1}{2}$的概率取1，以$\frac{1}{2}$的概率取0, $\mathbb{E}[X]=0\times p(0)+1\times p(1)=p(1)=\frac{1}{2}$
- 方法一: 直接按照定义计算
- 方法二: 对于取非负整值的随机变量, $\mathbb{E}[X]=\Pr[X>0]+\Pr[X>1]+\ldots$
- 方法三: 将X拆解为一系列伯努利(取值0/1)变量$X_1,\ldots,X_n$之和，$\mathbb{E}[X]=\mathbb{E}[X_1]+\ldots+\mathbb{E}[X_n]=\Pr[X_1=1]+\ldots+\Pr[X_n=1]$
- 方法四: 第一步分析(first step analysis)，借用马氏链的工具
- 方法五: 矩母函数(MGF)
- 方法六: 对称性

Outline

1 2 3

1 1 1

2 3 2 3 2 1

- **Q:** 从$1, 2, \ldots, n$中有放回地均匀随机采样m次，问出现过的不同数字的个数X的期望是多少？(Google)

X

- **Q:** 从$1, 2, \ldots, n$中有放回地均匀随机采样m次，问出现过的不同数字的个数X的期望是多少？(Google)
- 出现的不同数字的个数X可以看作$X_1 + \ldots + X_n$，其中X_i表示i是否出现，如果出现，为1否则为0
- 例如$n = 5$，只出现了2和4，那么出现的不同数字个数为$X_1(0) + X_2(1) + X_3(0) + X_4(1) + X_5(0) = 2$
- $\mathbb{E}[X_i] = 0 \times \Pr[X_i = 0] + 1 \times \Pr[X_i = 1] = \Pr[X_i = 1] = \Pr[i\text{在}m\text{次采样中出现过}]$
- i没有出现过的概率是每次不出现的概率的乘积$\left(\frac{n-1}{n}\right)^m$
- $\mathbb{E}[X_i] = 1 - \left(\frac{n-1}{n}\right)^m$
- 期望可加: $\mathbb{E}[X] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = n\left[1 - \left(\frac{n-1}{n}\right)^m\right]$

Outline

- 对于两个事件A, B

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B|A]\Pr[A]}{\Pr[B|A]\Pr[A] + \Pr[B|A^c]\Pr[A^c]}$$

- 其中A^c为A的补集

Outline

- **Q:** 某城市发生了一起汽车撞人逃跑事件，该城市只有两种颜色的车，蓝20%绿80%，事发时现场有一个目击者，他指证是蓝车，但是根据专家在现场分析，当时那种条件能看正确的可能性是80%，那么，肇事的车是蓝车的概率是多少？(百度)

- **Q:** 某城市发生了一起汽车撞人逃跑事件，该城市只有两种颜色的车，蓝20%绿80%，事发时现场有一个目击者，他指证是蓝车，但是根据专家在现场分析，当时那种条件能看正确的可能性是80%，那么，肇事的车是蓝车的概率是多少？(百度)
- 由贝叶斯公式

$$
\begin{aligned}
&\Pr[\text{肇事车是蓝车}|\text{目击者指证蓝车}]\\
&=\Pr[\text{肇事车是蓝车}]\Pr[\text{目击者指证蓝车}|\text{肇事车是蓝车}]/\\
&(\Pr[\text{肇事车是蓝车}]\Pr[\text{目击者指证蓝车}|\text{肇事车是蓝车}]\\
&+\Pr[\text{肇事车是绿车}]\Pr[\text{目击者指证蓝车}|\text{肇事车是绿车}])\\
&=\frac{0.2\times 0.8}{0.2\times 0.8+0.8\times 0.2}\\
&=0.5
\end{aligned}
$$

练习题

- **Q:** 某大公司有这么一个规定：只要有一个员工过生日，当天所有员工全部放假一天。但在其余时候，所有员工都没有假期，必须正常上班。这个公司需要雇用多少员工，才能让公司一年内所有员工的总工作时间期望值最大？(美图秀秀)

练习题

- **Q:** 某大公司有这么一个规定：只要有一个员工过生日，当天所有员工全部放假一天。但在其余时候，所有员工都没有假期，必须正常上班。这个公司需要雇用多少员工，才能让公司一年内所有员工的总工作时间期望值最大？(美图秀秀)
- n名员工, X表示不放假的天数，则总工作时间为nX
- 第i天不放假: $X_i = 1$, 第i天放假$X_i = 0$
- $X = X_1 + \ldots + X_{365}$

$$
\begin{aligned}
\mathbb{E}[X_i] &= \Pr[\text{第}i\text{天不放假}] \\
&= \Pr[\text{每个员工都不过生日}] \\
&= (\frac{364}{365})^n
\end{aligned}
$$

$$
\mathbb{E}[nX] = 365n(\frac{364}{365})^n
$$

- $n \leqslant 364$关于n增加，$n > 364$时关于n减少

练习题

- **Q:** 你有三位好友，他们都在西雅图工作，西雅图是出了名的爱下雨，每天下雨的概率高达2/3。假设你的好友视力都不是很好并且办公室的视野也一般，因此以1/3的概率能正确地判断是否在下雨。假如他们中恰好有两位告诉你今天西雅图在下雨，问西雅图实际上在下雨的概率是多少？(LinkedIn)

练习题

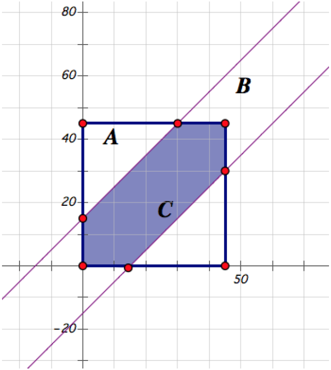
- **Q:** 你有三位好友，他们都在西雅图工作，西雅图是出了名的爱下雨，每天下雨的概率高达2/3。假设你的好友视力都不是很好并且办公室的视野也一般，因此以1/3的概率能正确地判断是否在下雨。假如他们中恰好有两位告诉你今天西雅图在下雨，问西雅图实际上在下雨的概率是多少？(LinkedIn)
- 三位朋友是否说下雨相互独立
- 假设三位朋友为A, B, C: A说今天不下雨，B和C说今天下雨
- $\Pr[\text{A说下雨, B和C说不下雨}|\text{下雨}] = \Pr[\text{A不看错, B和C看错}] = \frac{1}{3}\frac{2}{3}\frac{2}{3}$
- $\Pr[\text{A说下雨, B和C说不下雨}|\text{不下雨}] = \Pr[\text{A看错, B和C不看错}] = \frac{2}{3}\frac{1}{3}\frac{1}{3}$
- $\Pr[\text{下雨}] = \frac{2}{3}, \Pr[\text{不下雨}] = \frac{1}{3}$

练习题

- **Q:** 甲乙约定在某地,忘了约时间,只知道七点到八点见面,他们都打算于七点到七点四十五随机时刻到达某地,然后等十五分钟,如果两人没见面就走人,问他们见面的概率是多少?(瓜子二手车)

练习题

- **Q:** 甲乙约定在某地,忘了约时间,只知道七点到八点见面,他们都打算于七点到七点四十五随机时刻到达某地,然后等十五分钟,如果两人没见面就走人,问他们见面的概率是多少?(瓜子二手车)



谢谢

谢谢!
扫码关注我的知乎专栏