

DS概率题目总结 @Cindy @昆茹

写在前面：

因为最近DS的面试题目中，总会在某一轮的面试里面有一个统计题目。最近Facebook包括更多的金融行业的DS会考到很多统计。因此我就把最近的我遇到的题目，我不会的，还有所有的和分布相关的8大分布一起总结了一下。（腾讯文档图片有时候加载有些慢，但是空的地方都有图）

框架：



一. 泊松分布(Possion Distribution)

1.1 常考概率函数 (5★)

泊松分布的概率函数为：

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, \dots$$

泊松分布的参数 λ 是单位时间(或单位面积)内随机事件的平均发生次数。 K 是实际发生的次数，一般是求 概率 $P(k)$ ，泊松分布适合于描述单位时间内随机事件发生的次数。

泊松分布的[期望](#)和[方差](#)均为 λ 。特征函数为

$$\text{特征函数为 } \psi(t) = \exp\{\lambda(e^{it} - 1)\}.$$

1.2.泊松分布与二项分布

当二项分布的 n 很大（实验次数）而 p 很小时（某件事情发生的概率），泊松分布可作为二项分布的近似，其中 λ 为 np 。通常当 $n \geq 20, p \leq 0.05$ 时，就可以用[泊松公式](#)近似得计算。

1.3 经典例题

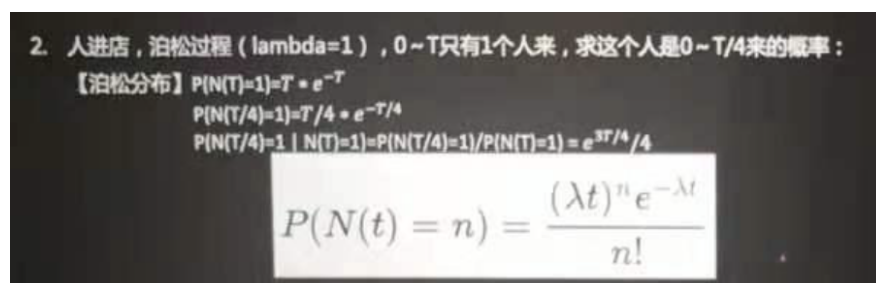
泊松分布适合于描述单位时间（或空间）内随机事件发生的次数。如某一服务设施在一定时间内到达的人数，电话交换机接到呼叫的次数，汽车站台的候客人数，机器出现的故障数，自然灾害发生的次数，一块产品上的缺陷数，显微镜下单位分区内的细菌分布数等等。

观察事物平均发生 m 次的条件下，实际发生 x 次的概率 $P(x)$ 可用下式表示：

$$P(x) = \frac{m^x}{x!} \times e^{-m}$$

$$p(0) = e^{-m}$$

1.4 我遇到过不会的题



【解析】：来的时候就是一个随机过程， k 就是发生一次，这个人来的概率就是一个随机概率，所以相当于一个泊松分布+ 一个条件分布。

图中加上时间的分布叫泊松过程

条件分布的补充

$P(A|B) = \frac{P(AB)}{P(B)}$ 在这个题目: $P(AB) = P(A)$ 因为 t 发生事件的时候, $t/4$ 一定也发生了。

贝叶斯的概率是已知:

$P(B|A)$ 求 $P(A|B)$

二. 指数分布 (5🌟)

2.1 概念:

指数分布 (英语: Exponential distribution) 是一种连续概率分布。指数分布可以用来表示独立随机事件发生的时间间隔, 比如旅客进入机场的时间间隔、电话打进客服中心的时间间隔

2.2 与泊松分布之间的关系

在一个时间段内事件平均发生的次数服从泊松分布, 这个次数在泊松分布中用 λ 表示。这个 λ 在指数分布里面的意义基本是一样的, 也是在一个时间段内事件平均发生的次数。

泊松分布表示的是事件发生的次数, “次数”这个离散变量, 所以泊松分布是离散随机变量的分布。

指数分布是两件事情发生的平均间隔时间, “时间”是连续变量, 所以指数分布是一种连续随机变量的分布。

可以用等公交车作为例子:

某个公交站台一个小时内出现了的公交车的数量 就用泊松分布来表示

某个公交站台任意两辆公交车出现的间隔时间 就用指数分布来表示

计算CLV的时候, 也用到了gamma分布, 即 $\alpha = 1$ 的指数分布

<https://www.jiqizhixin.com/articles/2018-06-21-8>

2.2.1 期望上的区别

每日卖出馒头的数目 X 服从泊松分布，卖出馒头的时间间隔 Y 服从指数分布：

$$X \sim P(\lambda), \quad Y \sim \text{Exp}(\lambda)$$

他们的期望分别为：

$$E(X) = \lambda, \quad E(Y) = \frac{1}{\lambda}$$

根据之前的分析就比较好理解了， $E(X)$ 的含义是平均每日卖出的馒头数，而 $E(Y)$ 是每个馒头之间卖出的平均时间间隔，所以两者是倒数关系：每日卖出的越多自然间隔时间越短，每日卖出的越少自然间隔时间越长。

2.3 具体讲解例子

*（注意是要求导得到概率密度函数）

两次卖出馒头之间的时间间隔大于 t 的概率，根据之前的分析，等同于 t 时间内没有卖出一个馒头的概率，而後者的概率可以由泊松过程给出。至此所需的条件都齐备了，那么开始解题吧，假设随机变量：

$$Y = \text{两次卖出馒头之间的时间间隔}$$

这个随机变量的概率可以如下计算：

$$P(Y > t) = P(X = 0, t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}, \quad t \geq 0$$

进而有：

$$P(Y \leq t) = 1 - P(Y > t) = 1 - e^{-\lambda t}$$

这其实已经得到了 Y 的累积分布函数了：

$$F(y) = P(Y \leq y) = \begin{cases} 1 - e^{-\lambda y}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

对其求导就可以得到概率密度函数：

$$p(y) = \begin{cases} \lambda e^{-\lambda y}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

这就是卖出馒头的的时间间隔 Y 的概率密度函数，也称为 **指数分布**。

2.3 例题：

A government office has two officer's handling people's requests. Suppose that the time between request arrivals at the first officers desk is random and follows an

exponential distribution with $A = u_1$. Similarly, the time between request arrivals at the second officer's desk is also random and follows an exponential distribution with $A = u_2$. The first officer has probability of referring any request he receives to the office supervisor, and the second officer has probability p_2 of doing so. What is the average time between requests referred to the supervisor

3. 2人处理req, req到A手里间隔时间随机指数分布 ($\lambda = u_1$), 到B手里也是 ($\lambda = u_2$), A上报概率 p_1 , B上报概率 p_2 , 求总上报的平均间隔时间:

【指数分布】 $P(t \text{ 内上报}) = 1 - P(t \text{ 内没上报}) = P(A \text{ 收到}) \cdot p_1 + P(B \text{ 收到}) \cdot p_2 - P(A \text{ 收到}) \cdot P(B \text{ 收到}) \cdot p_1 \cdot p_2$

$P(A \text{ 收到}) = 1 - e^{-u_1 t}$, $P(B \text{ 收到}) = 1 - e^{-u_2 t}$, $E(A \text{ 收到}) = 1/u_1$, $E(B \text{ 收到}) = 1/u_2$

求得 $f(t)$, 再0到正 ∞ 积分 $t f(t)$ 得 $E(t)$

$$\frac{p_1}{u_1} \cdot (1 - p_2) + \frac{p_2}{u_2} \cdot (1 - p_1) + \frac{p_1 p_2}{u_1 + u_2}$$

$$P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t}$$

$$E(X) = \int_{-\infty}^{\infty} |x| f(x) dx = \int_0^{\infty} x f(x) dx$$

Or 排除法, 与 $p_1 p_2$ 正相关, 且小于 p_1/u_1 (p_2 非0时) 即单人处理的时间间隔, 且 $p_2=0$ 时只与 p_1, u_1 相关

三.条件概率 (5🌟)

3.1 概念

事件可以是 "独立" 的, 意思是每个事件是 不 受其他事件影响的。

但事件也可以是 "相关" 的意思是 受过去事件影响的,

$P(B|A)$ 的意思是 "在事件 A 发生的条件下, 事件 B 发生的概率"

换句话说, 事件 A 已经发生了, 现在事件 B 发生的可能性是多少?

3.2 计算公式

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

3.3 经典例题

1. 3 coins, 1 fair, 1全正, 1全反, 随机盲取1个抛, 结果是正, 求抛到的那个硬币是fair概率:

【条件概率】 $P(\text{fair, head}) = 1/3 * 1/2 = 1/6$, $P(\text{heads, head}) = 1/3 * 1 = 1/3$, $P(\text{tails, head}) = 0$,
 $P(\text{head}) = 1/6 + 1/3 = 1/2$, $P(\text{fair} | \text{head}) = P(\text{fair, head}) / P(\text{head}) = (1/6) / (1/2) = 1/3$

英文 heads / tails

$P(A \text{ and } B)$ 在两件事情本身是相互独立的情况下, 可以写成 $P(A) * P(B)$

这个题目 $P(A/B)$: B事件是扔到正面的情况

$P(AB)$ 可以想象是先扔出来一个硬币是fair, 然后第二个是他正面。两个步骤, 相互独立(Fair 硬币并不影响他扔到正反面的概率, 所以独立)。所以 $P(AB) = P(A) * P(B) = 1/3 * 1/2 = 1/6$

$P(B)$ 如上面就是是fair 硬币中, 1/2 的情况是正+ 全是tail * 0情况是正 + 全是Head * 1全为正

$P(B) = 1/3 * 1/2 + 1/3 * 0 + 1/3 * 1 = 1/2$

所以: $1/6 | 1/2 = 1/3$

四. 伯努利分布 (3🌟)

4.1 概念

伯努利分布就是说明自变量值为1和0的概率是多少。就是说明概率, 没有其他的什么用, 然后二项分布就是实验n次, 每次的结果都可能正可能负, 然后二项分布的相关计算就是用伯努利的为正为负的概率来计算

4.2 概率公式

0-1分布: $P(X=k) = p^k(1-p)^{1-k}, k=0,1$

五. 二项分布 (相似于伯努利) 4🌟

5.1 概念

n重伯努利试验「成功」次数的离散概率分布

5.2 概率公式

二项分布: $B(n, p)$: $P(X=k) = C_n^k p^k (1-p)^{n-k}, k=0,1,\dots,n$

5.3 例题:

张三参加英语雅思考试, 每次考试通过的概率为 1/3, 不通过的概率为 2/3。如果他连续考试 4 次, 那么恰好通过 2 次的概率为多少?

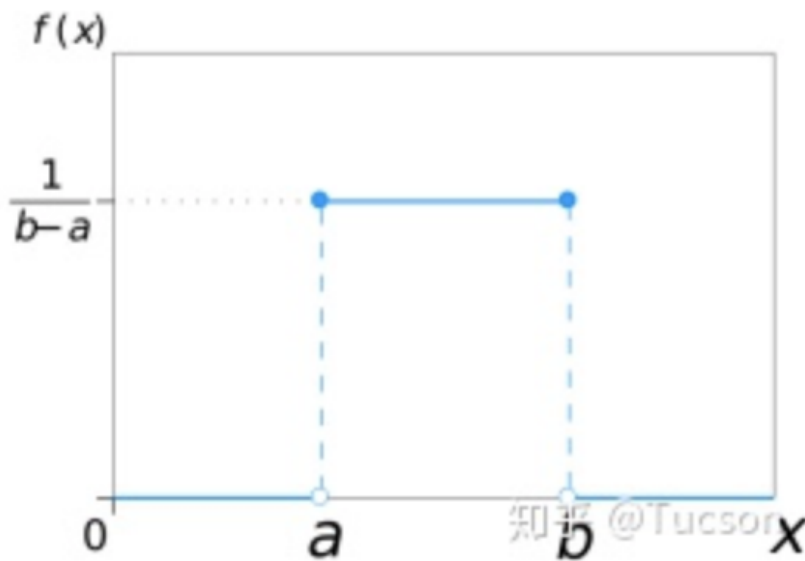
$$C_4^2 \times \left(\frac{1}{3}\right)^2 \left[1 - \left(\frac{1}{3}\right)\right]^{(4-2)} = \frac{4 \times 3}{2 \times 1} \times \frac{1}{9} \times \frac{4}{9} = \frac{8}{27}$$

二项分布的典型例子是扔硬币，硬币正面朝上概率为p, 重复扔n次硬币，k次为正面的概率即为一个二项分布概率。

六. 均匀分布 (Uniform Distribution) 3🌟

6.1 概念

在概率论和统计学中，均匀分布也叫矩形分布，它是对称概率分布，在相同长度间隔的分布概率是等可能的。均匀分布由两个参数a和b定义，它们是数轴上的最小值和最大值，通常缩写为U (a, b)



6.2：性质

均匀分布的概率密度函数为：

$$f(x) = \frac{1}{b-a}, a < x < b$$

$$f(x) = 0, else$$

遵循均匀分布的X的平均值和方差为：

$$\text{平均值} \rightarrow E(X) = (a+b)/2$$

$$\text{方差} \rightarrow V(X) = (b-a)^2/12$$

6.3 例题

(1) 花店每天销售的花束数量是均匀分布的，最多为40，最少为10。我们来计算一下日销售量在15到30之间的概率。

$$\text{日销售量在15到30之间的概率为 } (30-15) * (1/(40-10)) = 0.5$$

同样地，日销售量大于20的概率为 = 0.667

(2) 题目是：在长为L的线段上随机地选取一点，将其分为两段，短的一段与长的一段之比小于1/4的概率是多少？

解答：随机取一点为X

则X服从U(0, L)的均匀分布

$$\text{所以 } P(0 < X < 1/5L) = 1/5$$

$$P(4/5L < X < L) = 1/5$$

短的一段与长的一段之比小于1/4的概率为上面两个之和=2/5

7. 正态分布 (4.5🌟)

7.1 特征

正态分布代表了宇宙中大多数情况的运转状态。大量的随机变量被证明是正态分布的。任何一个分

布只要具有以下特征，则可以称为正态分布：

1. 分布的平均值、中位数和模式一致。
2. 分布曲线是钟形的，关于线 $x = \mu$ 对称。
3. 曲线下的总面积为1。
4. 有一半的值在中心的左边，另一半在右边。

7.2 性质

正态分布与二项分布有着很大的不同。然而，如果试验次数接近于无穷大，则它们的形状会变得十分相似。

遵循正态分布的随机变量X的值由下式给出：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\}} \quad \text{for } -\infty < x < \infty.$$

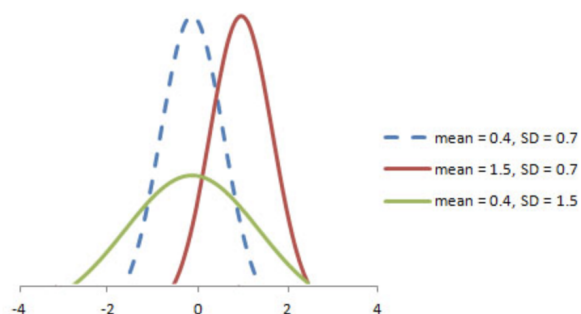
正态分布的随机变量X的均值和方差由下式给出：

均值 $\rightarrow E(X) = \mu$

方差 $\rightarrow \text{Var}(X) = \sigma^2$

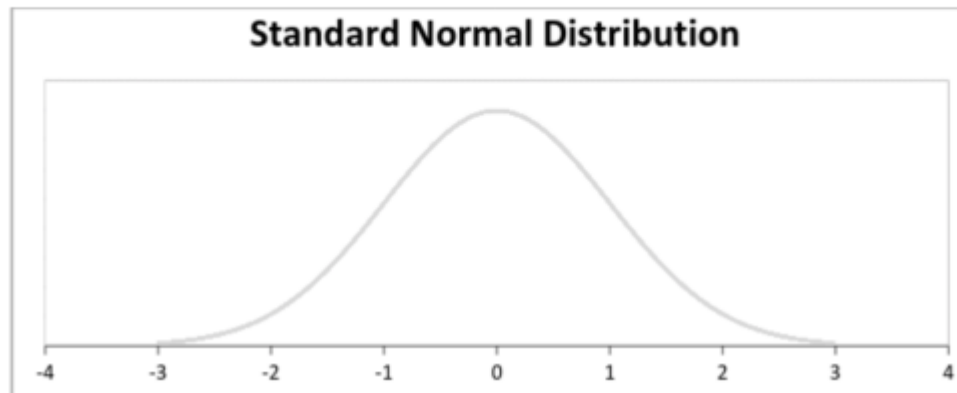
其中， μ （平均）和 σ （标准偏差）是参数。

随机变量 $X \sim N(\mu, \sigma)$ 的图如下所示。



标准正态分布定义为平均值等于0，标准偏差等于1的分布：

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for } -\infty < x < \infty$$



八. Additonal Questiosn

8.1 切金条问题

技巧就是：如果是基数天，基本是2的倍数（1，2，4，8）区分金块，如果是7天就是1，2，4；15天就是1，2，4，8。

You've got someone working for you for seven days and a gold bar to pay him. The gold bar is segmented into seven connected pieces. You must give them a piece of gold at the end of every day. What and where are the fewest number of cuts to the bar of gold that will allow you to pay him 1/7th each day?



解答：

Puzzle Solution:

Lets split the chain as,



Day 1: Give A (+1)

Day 2: Get back A, give B (-1, +2)

Day 3: Give A (+1)

Day 4: Get back A and B, give C (-2, -1, +4)

Day 5: Give A (+1)

Day 6: Get back A, give B (-1, +2)

Day 7: Give A (+1)

Day	我方	人数	
Day 1	12,4,8	1	-1股
Day 2	21,4,8	2	
Day 3	11,4,8	3	
Day 4	41,2,8	4	
Day 5	11,2,8	5	
Day 6	21,1,8	6	
Day 7	11,8	7	
Day 8	81,2,4	8	
Day 9	11,2,4	9	
Day 10	21,4	10	人均有1/5

【类似题目】

You have a worker and a gold bar. The worker will work for you for 15 days. He must have $x/15$ of the gold bar after the x th day of work. What is the minimum number of cuts of the gold bar you need to make to pay the worker? (3刀, 4 bites 1-2-4-8)

8.2 分步概率

【Citadel】A bag contains 8 fair dice as well as two rigged dice with 4 dots on all six sides. You pick a die at random item the bag and roll it for 3 times. What is the probability that all three rolls produce the 4-dot outcome?

A. $5/27$ B. $11/54$ C. $2/9$ D. $13/54$ E. $7/27$

要注意是同一个骰子扔，所以三次中，第一步是选骰子，且不变，只是每次扔骰子的点数发生变化。所以

$$[8/10 * (1/6)^3] + [2/10 * 1^3]$$

三次方应该只针对 扔出去的点数

类似的题目

九. ML 小问题补充

7.1 Which regularizer to use to get a sparse set of regression parameters?

Parse的意思: Lasso: 有很多0, 有weights, 所以是absolute value。

Ridge (weights 很小, 但是well distributed, 每个parameter)

答: The most common sparse regularizer is **sum of absolute values** (so-called Lasso regression). With carefully chosen penalty coefficient, it makes some of less useful parameters exactly zero.

Cardinality penalty exactly imposes sparsity, but it cannot be combined with gradient descent, and usually requires combinatorial optimization. Simply put, it is slow to apply.

Maximum value and **Euclidean norm** do not affect sparsity at all.

7.2 what is the best way to handle missing time indices in csv? (PYTHON)

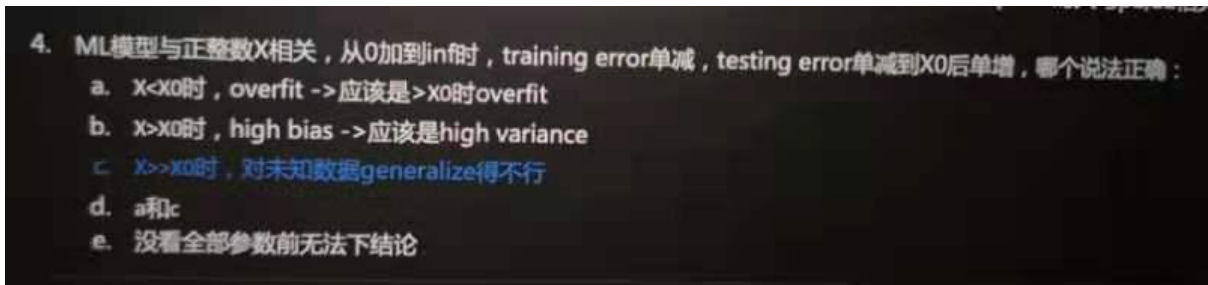
looking for reindex i.e

```
1 df =  
  df.reindex(pd.date_range(df.index.min(),df.index.max(),freq='D'))
```

In case of removing nan rows then

```
1 df = df.dropna()
```

7.3 Testing Error Problem

- 
4. ML模型与正整数X相关, 从0加到inf时, training error单减, testing error单减到X0后单增, 哪个说法正确:
- a. $X < X_0$ 时, overfit -> 应该是 $> X_0$ 时overfit
 - b. $X > X_0$ 时, high bias -> 应该是high variance
 - c. $X > X_0$ 时, 对未知数据generalize得不行
 - d. a和c
 - e. 没看全部参数前无法下结论

十.Reference:

[1] <https://www.zhihu.com/question/24796044>