概率（论）基础

山东省实验中学
宁华

什么是概率？

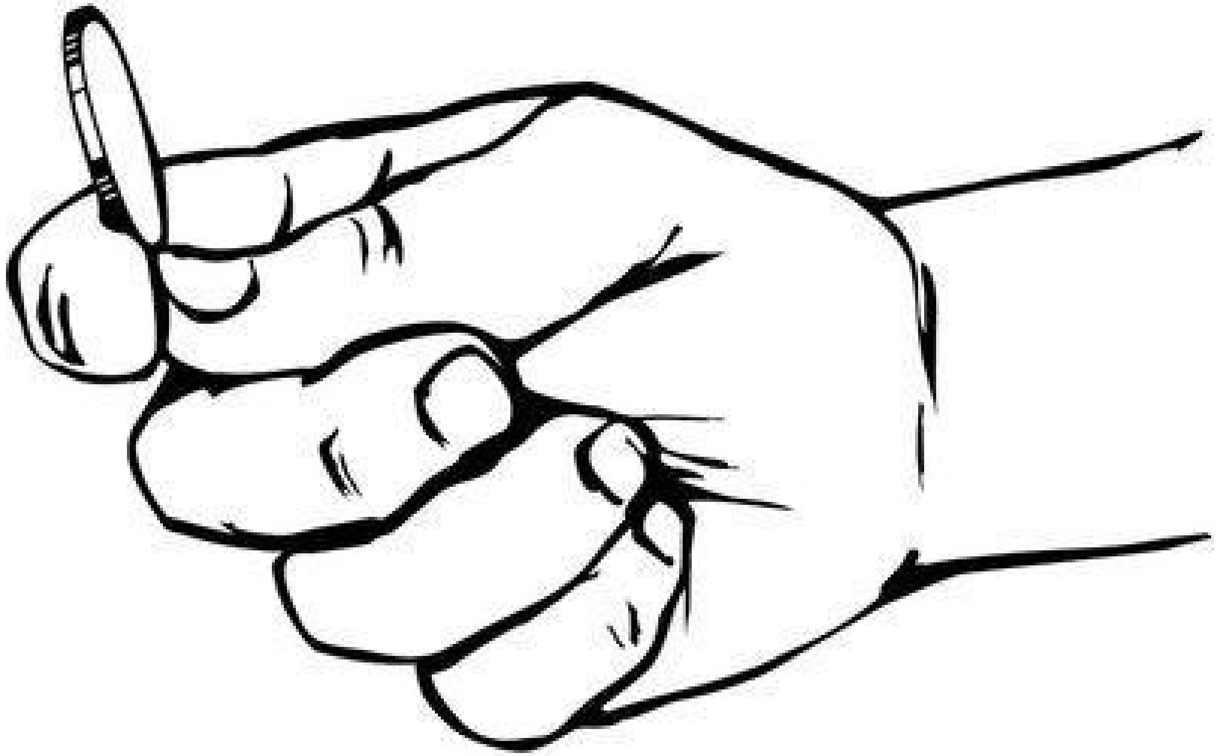
- 对于概率的定义有几个主流的派别：

- 频率派
- 古典派
- 主观派

1、频率派

- 频率派的理论基础是**对过去事实的归纳总结**。

频率



频率

- 如果扔了100次硬币，得到48次正面，52次反面，用正面次数除以总的次数：

$$P_{100}(\text{正面}) = \frac{48}{100} = 0.48$$

- P_{100}(正面) 称为扔100次硬币时，正面出现的频率 。

前人们的试验

	n	n_H	P_n(正面)
德摩根	2048	1061	0.5181
蒲丰	4040	2048	0.5069
皮尔逊	24000	12012	0.5005

C++模拟

```
#include<bits/stdc++.h>
using namespace std;
int main()
{
        int n=10000000;
        srand(time(0));
        int a=0,b=0;
        for(int i=1;i<=n;i++)
        {
                int x=rand()%2;
                if(x)a++;
                else b++;
        }
        cout<<a<<endl<<b<<endl<<double(a)/(a+b)<<endl;
        return 0;
}
```

频率稳定性

- 从试验结果可见，随着 n 的增大，频率越来越趋近于0.5。可见，虽然单次扔硬币的结果是随机的，但多次重复后频率趋于稳定，这种稳定性也称为 **<u>频率稳定性</u>** ，反映了扔硬币存在某种必然性。

频率→概率

- 频率派认为如果频率存在稳定性，即当 n→∞ 时下面极限存在，就得到了概率 （用Probability的首字母P来表示）：

$$P(\text{正面}) = \lim_{n\to\infty} P_n(\text{正面})$$

频率派的缺点

- 通过频率来定义概率的方法比较符合直觉，但缺陷也很明显：

- 首先，需要 n 足够大，但是“足够大”这个词很含糊。
- 其次，需要在相同条件下反复扔硬币，但是“相同条件”这个词也很含糊，也很难保证，比如扔了10000次后，硬币上沾满汗水，那又怎么办？
- 再次，永远也不可能扔无限次硬币，所以得到的概率始终是一个近似值。
- 最后，有些时候根本不具备反复实验的条件，比如火山喷发的概率应该怎么计算？

2、古典派

- 古典派的理论基础是<u>不充分理由原则</u>。

- 在概率论草创阶段，雅各布·伯努利（1654－1705）就提出，如果因为无知，使得我们没有办法判断哪一个结果会比另外一个结果更容易出现，那么应该给予它们相同的概率。比如：

- 硬币：由于不清楚硬币哪一面更容易出现，那么应该给予正面、反面相同的概率，即为 1/2
- 骰子：我们不清楚骰子哪一面更容易出现，那么应该给予每一面相同的概率，即为 1/6
- 此称为不充分理由原则（Insufficient Reason Principle）。

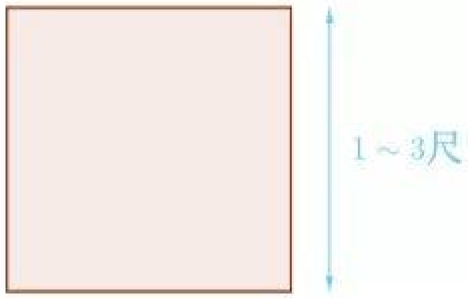
古典概率

- 以不充分理由原则为基础，经由拉普拉斯（皮埃尔-西蒙·拉普拉斯侯爵，1749－1827）之手，确立了 古典概率 的定义，即：**<u>未知的概率都是等概率</u>**。

- 在这之后，古典概率在整个19世纪也被人们广泛接受，我们高中学习的概率，基本都是古典概率。

古典概率

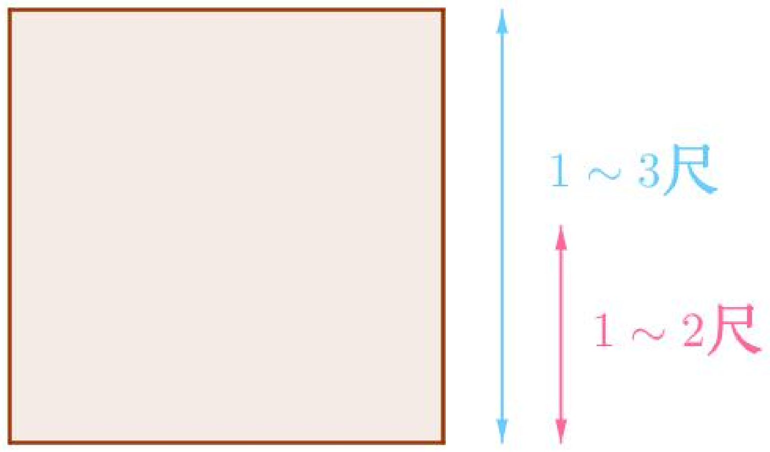
- 比如，有一家原木加工厂，它会把木头切成不同的木方，木方的截面都是正方形，边长会在1 ~ 3尺之间随机浮动：



- 那么根据古典概率，正方形边长在1 ~ 2尺之间的概率为多少？

不充分理由原则

- 根据古典概率的不充分理由原则，我们没有办法判断哪一种边长更容易出现，那么就应该给予它们相同的概率，也就是说1 ~ 3之间每一种长度都是等可能的。1 ~ 2包含了一半的可能长度：



- 所以，正方形边长在1 ~ 2尺之间的概率为 (2-1)/(3-1) = 1/2 。

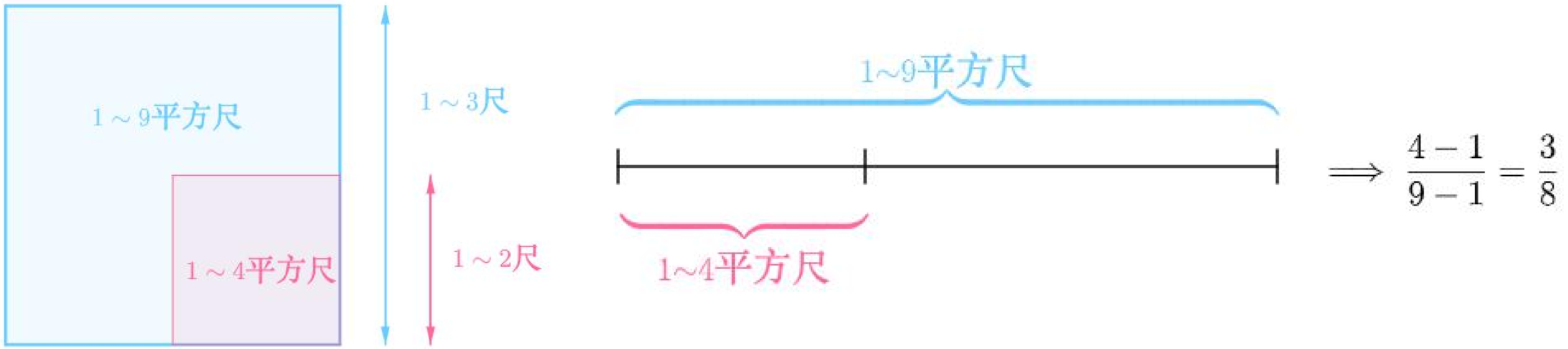
古典派的缺点

- 古典派的缺陷也是非常明显的：

- （1）古典派的概率定义，“未知的概率都是等概率”，有循环定义的嫌疑。
- （2）不充分理由原则没办法处理非等概率的情况，假如被告知硬币两面是非等概率的，但是不知道是哪一面，那么应该怎么办？（拉普拉斯提出还是应该按照等概率来处理）
- （3）还容易产生矛盾——比如：贝特朗悖论

贝特朗悖论

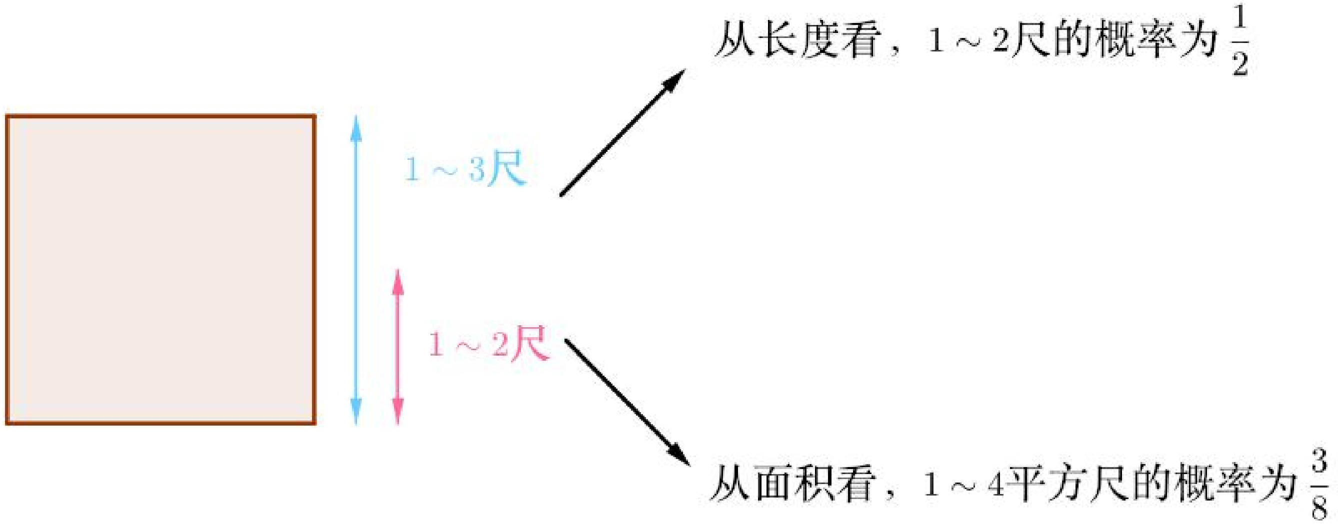
- 贝特朗悖论原型较为复杂，我们仍以原木加工厂的问题为例。该问题还可以转为面积来解答：
- 1 ~ 3尺边长的正方形面积为1 ~ 9平方尺，1 ~ 2尺边长的正方形面积为1 ~ 4平方尺：



- 同样，根据不充分理由原则，1 ~ 9平方尺之间的正方面面积是等可能的，那么正方形面积在1 ~ 4平方尺之间的概率为 3/8 .

哪个对？

- 选择对“长度”还是对“面积”运用不充分理由原则，同一个问题会得到了不同的概率：



反思

- 19世纪不少人相信只要找到适当的等概率，就可以得到问题的唯一解。直到贝特朗悖论出现，人们才开始反思古典概率中的不合理之处：“等概率”的描述实在是太模糊了，存在歧义。
- 在后来数学家的不断努力中，概率论变得越来越严谨，大学中学习的公理化的现代概率论就是集大成者。
- 用现代的概率论重新来审视贝特朗悖论，你会发现其实根本没有矛盾之处。

3、主观派

- 主观派认为概率是信念强度 （degree of belief）。

- 比如说，我个人相信20年后人类从网络时代进入人工智能时代的概率为70%：
- 上面说的概率也就是主观概率，是个人对这个命题的信念强度，换句话说我觉得还是很有可能实现的。
- 虽说是主观概率，其实也有客观的部分，比如刚才对人工智能的判断，就是基于AI的基础设置发展、计算速度的提高等事实。
- 主观概率更贴近人的思考方式，比如我们在作科学研究时，会先给出一个猜想，这就是给出了一个主观概率。
- 所以在人工智能时代，因为要模仿人的行为，主观概率越来越受到重视。

主观派的缺点

- 当然主观派缺陷也很明显，这也是被大家接受困难的原因：
- 说到科学，大家都认为应该是客观的，但是偏偏主观概率不客观，充满了个人偏见
- 因为主观，大家很难对某个主观概率达成共识

小结

- 三个流派大概有以下的区别：

	频率派	古典派	主观派
理论基础 概率定义	过往事实的归纳总结 频率稳定性	不充分理由原则 等概率	知识和直觉 信念强度

- 这三个流派并非泾渭分明、互不相容，反而在发展中犬牙交错。比如要判断火山的喷发概率，就需要总结过往数据（频率派），再加入主观知识（主观派）。
- 为什么概率的定义不明确？可能因为概率本身研究的就是“不明确”。