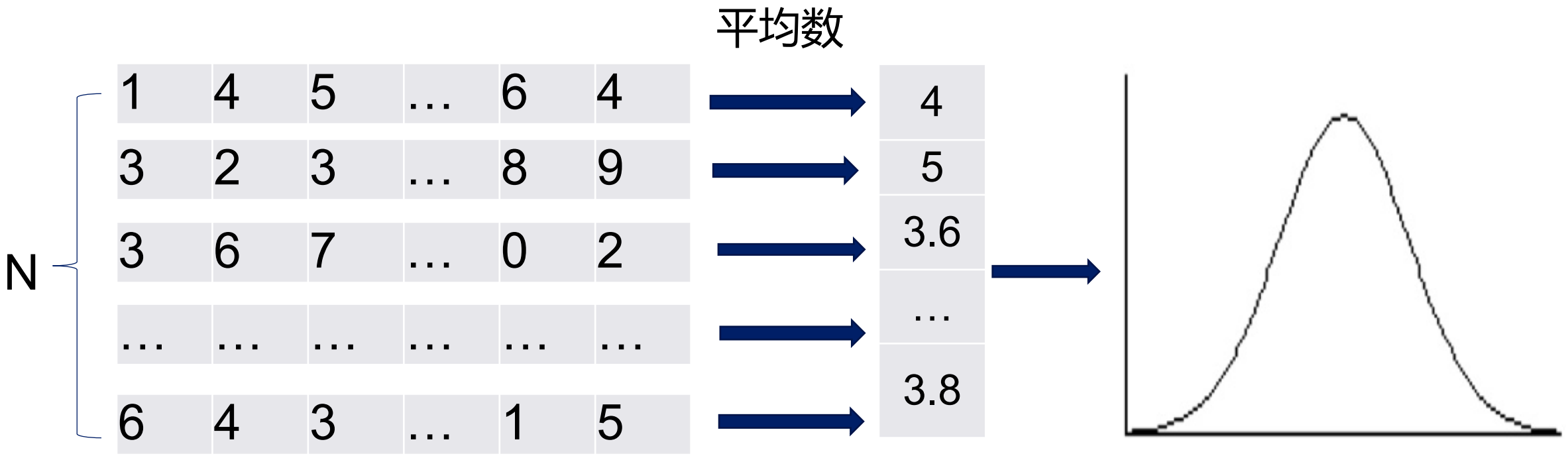
Practice of AI

C2: Machine learning & Data analyze

Jim Xie

2020/10/6

中心极限定理



- N越大，采样越多，曲线越瘦
- N越小，采样越少，曲线越瘦

$$\text{标准差} = \frac{\sigma}{\sqrt{N}}$$

问题

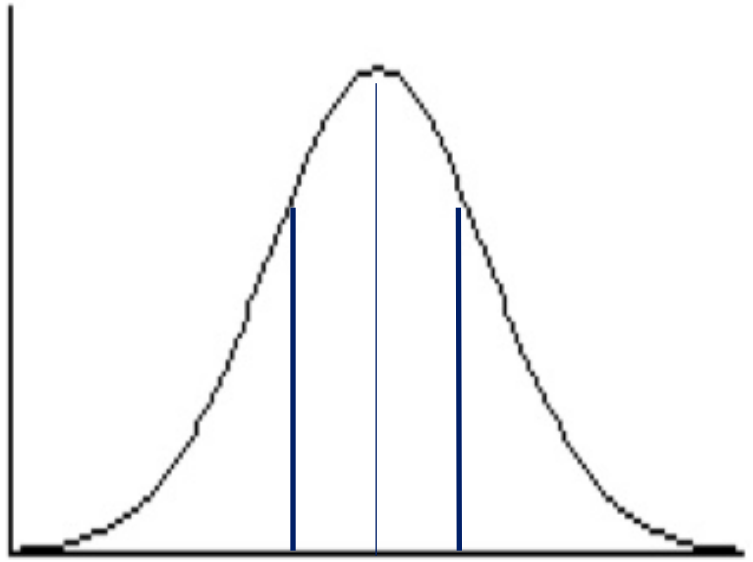
举例：

- 客户平均每天1万封垃圾邮件，标准差是5000；
- 新feature上线后，观察100天，平均每天9000封垃圾邮件；
- 新feature是否有效？

假设

H0: 假设feature是没有作用的，计算出现这种统计结果的概率

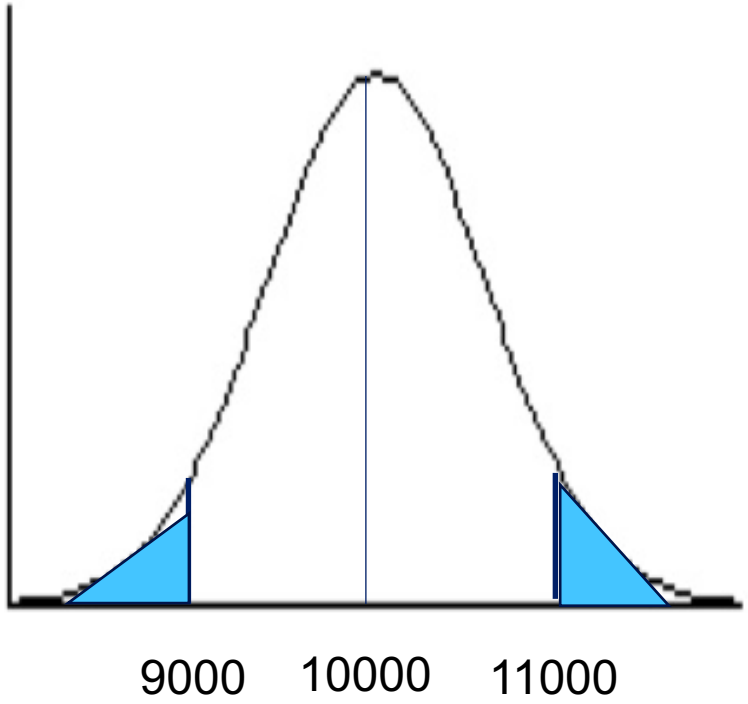
- 应该服从原来的正态分布，由于采样(N)为100.
- 标准差：$\sigma = \frac{5000}{\sqrt{100}} = 500$



10000

计算概率

- 10000-9000=1000正好是2个标准差，概率为95%;
- H0成立情况下，出现这个或更极端的结果，概率为5%;
- 也就是说，有95%的信心拒绝H0;
- 95%称为置信水平;



- 采样100天，有95%信心。
- 采样25天呢， 有68%信心。
- 采样7天呢， 有38%信心。

Thanks

2020-8-15