

# **大数据基础： 概率论与数理统计**

宁波海大物联科技有限公司

Frank Wen

2020.06.20

# 《大数据基础》系列

1. 概率论与数理统计
2. 数据可视化
3. 机器学习与数据挖掘

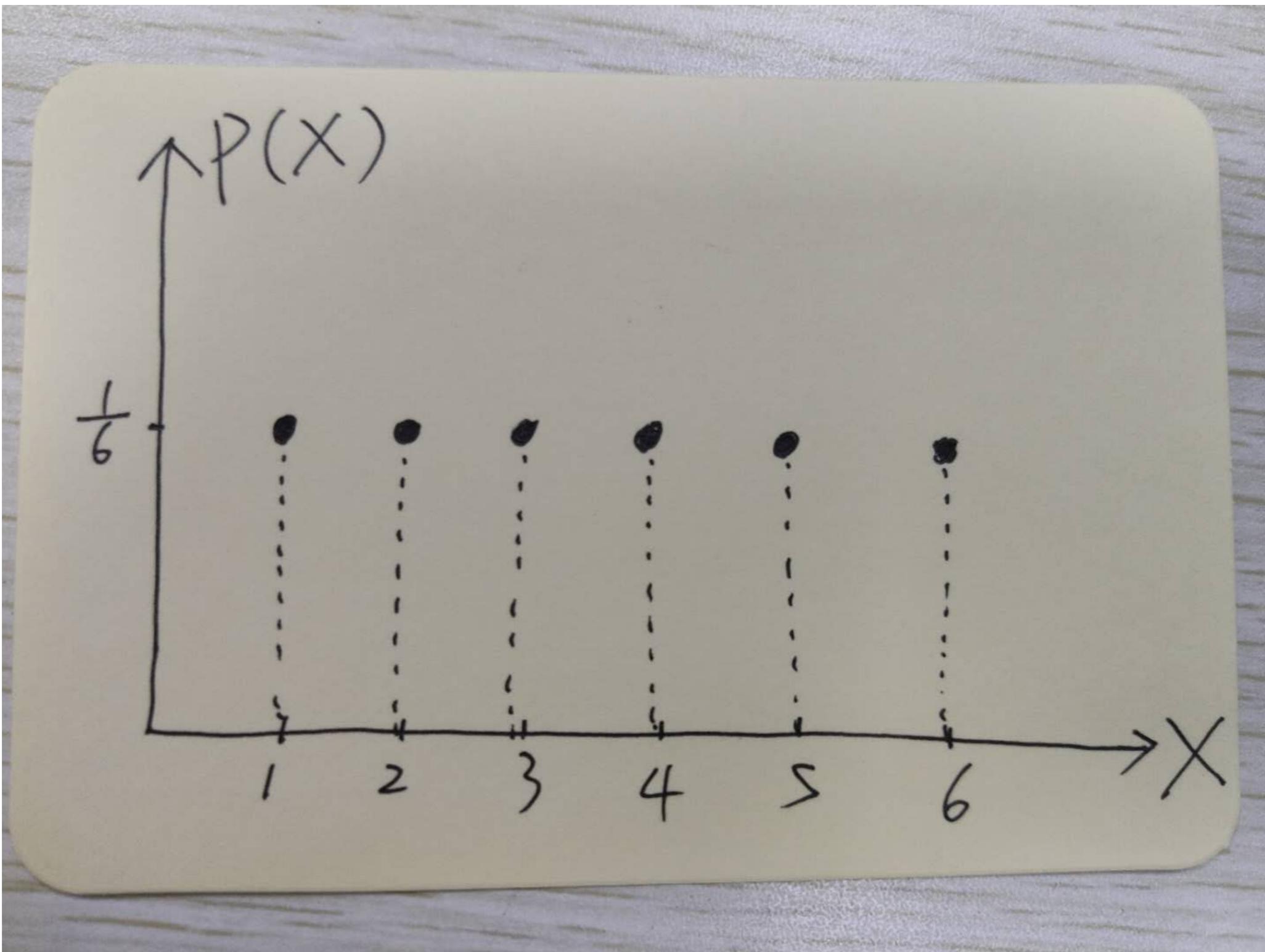
# Part 1. 概率

- 随机变量
- 随机事件
- 概率

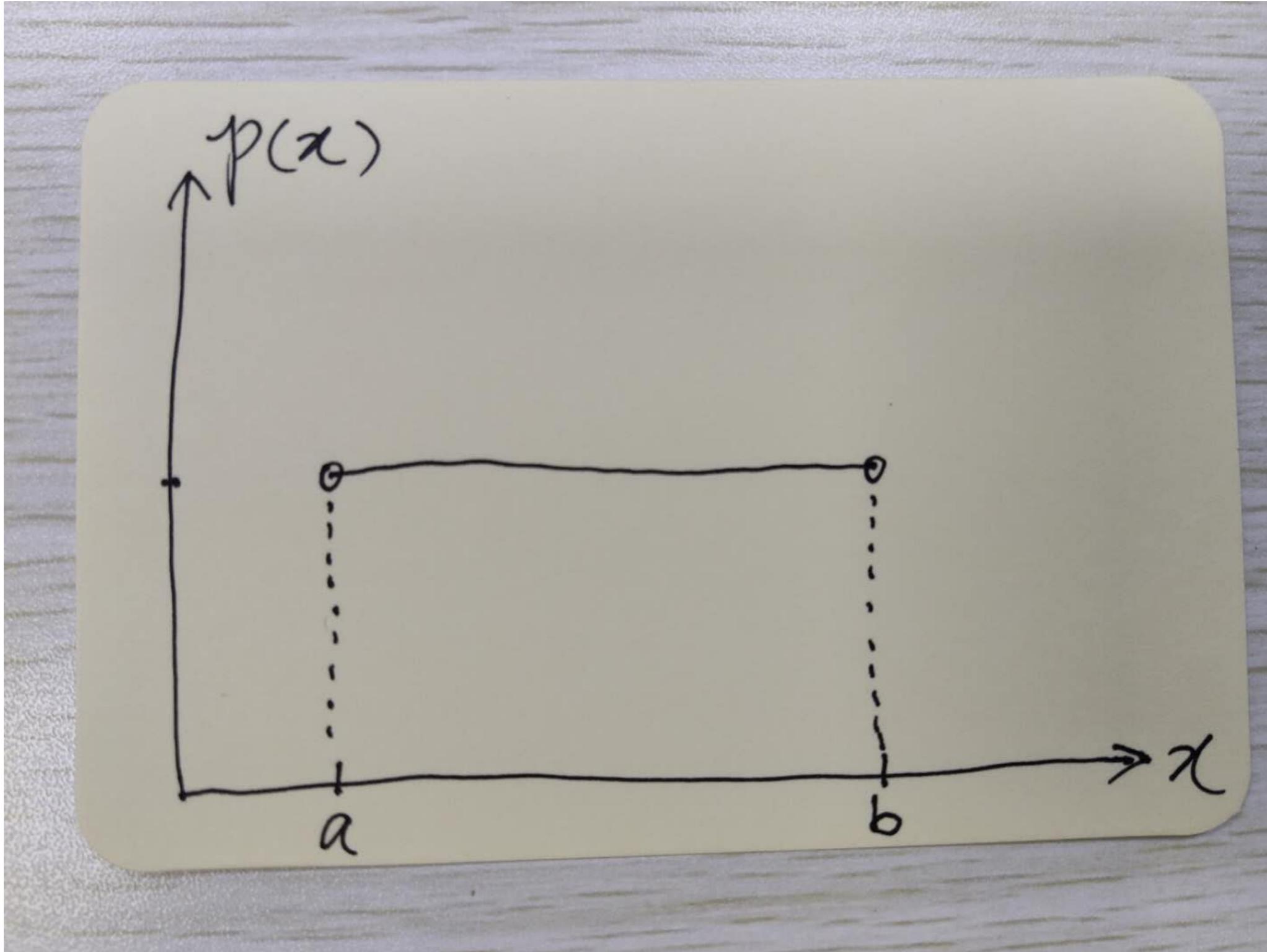
- 随机变量—— $X$ : 扔一个色子出现的点数
- 随机事件—— $A$ : 结果为6点
- 概率—— $P(X=A) = P(A) = 1/6$



# 离散型



# 连续型



# 连续型

$p(x)$  —— 概率密度函数

$$\int_{-\infty}^{+\infty} p(x) dx = 1$$

## 常见的概率分布

均匀分布

正态分布

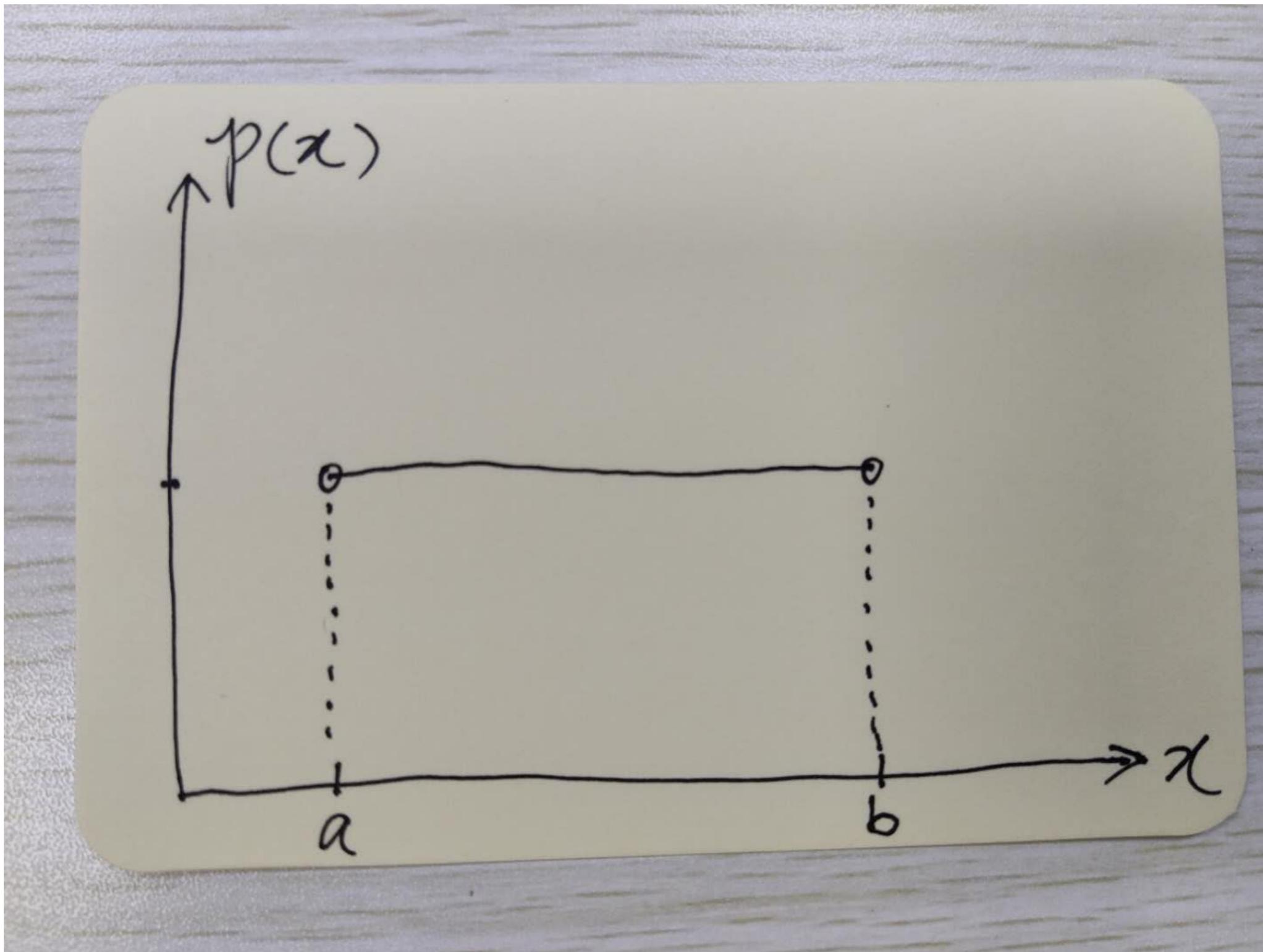
指数分布

.....

# 均匀分布

$$p(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{其他} \end{cases}$$

# 均匀分布

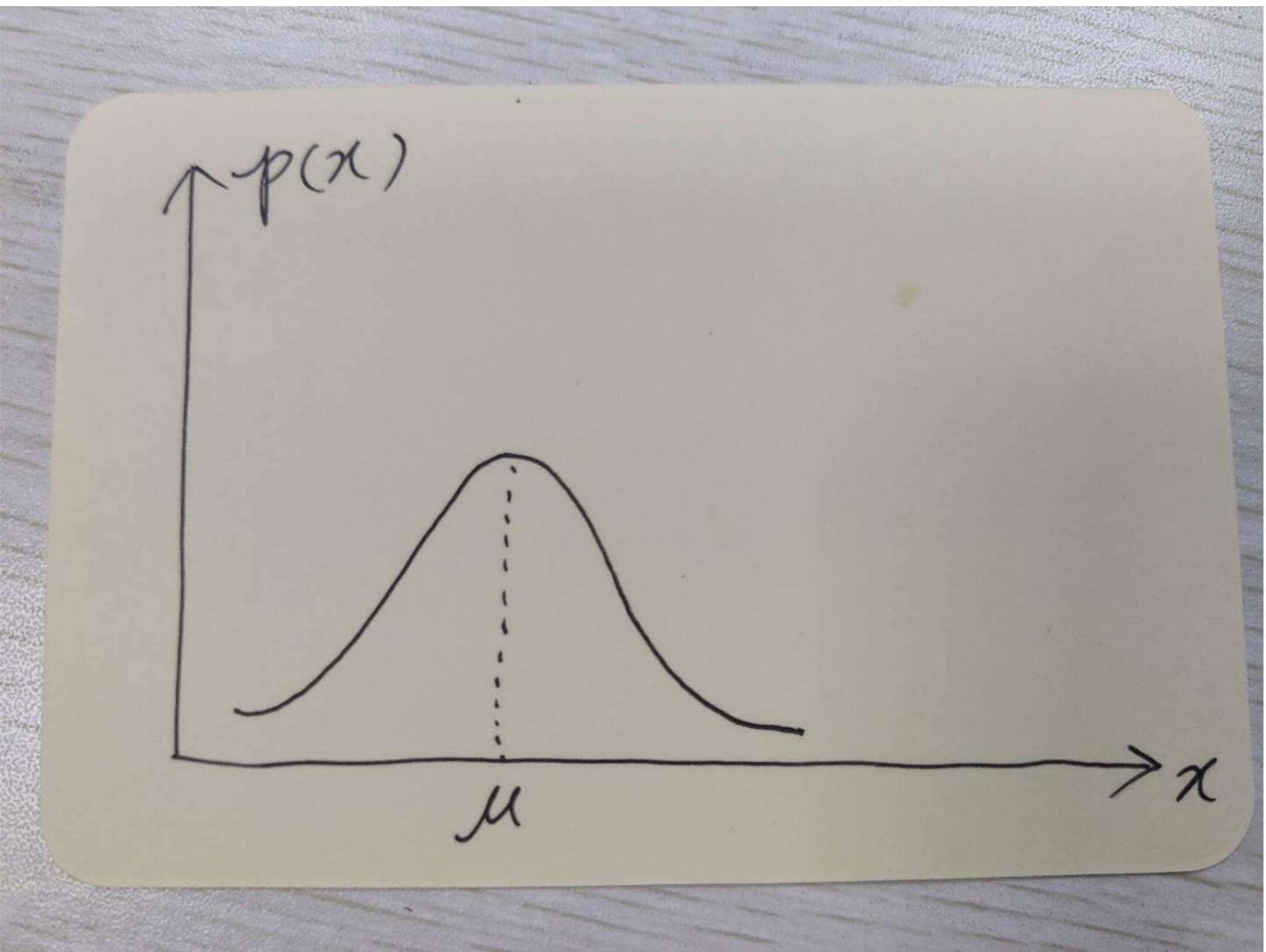


# 正态分布

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

其中  $-\infty < x < \infty$

# 正态分布



- 记作  $X \sim N(\mu, \sigma^2)$

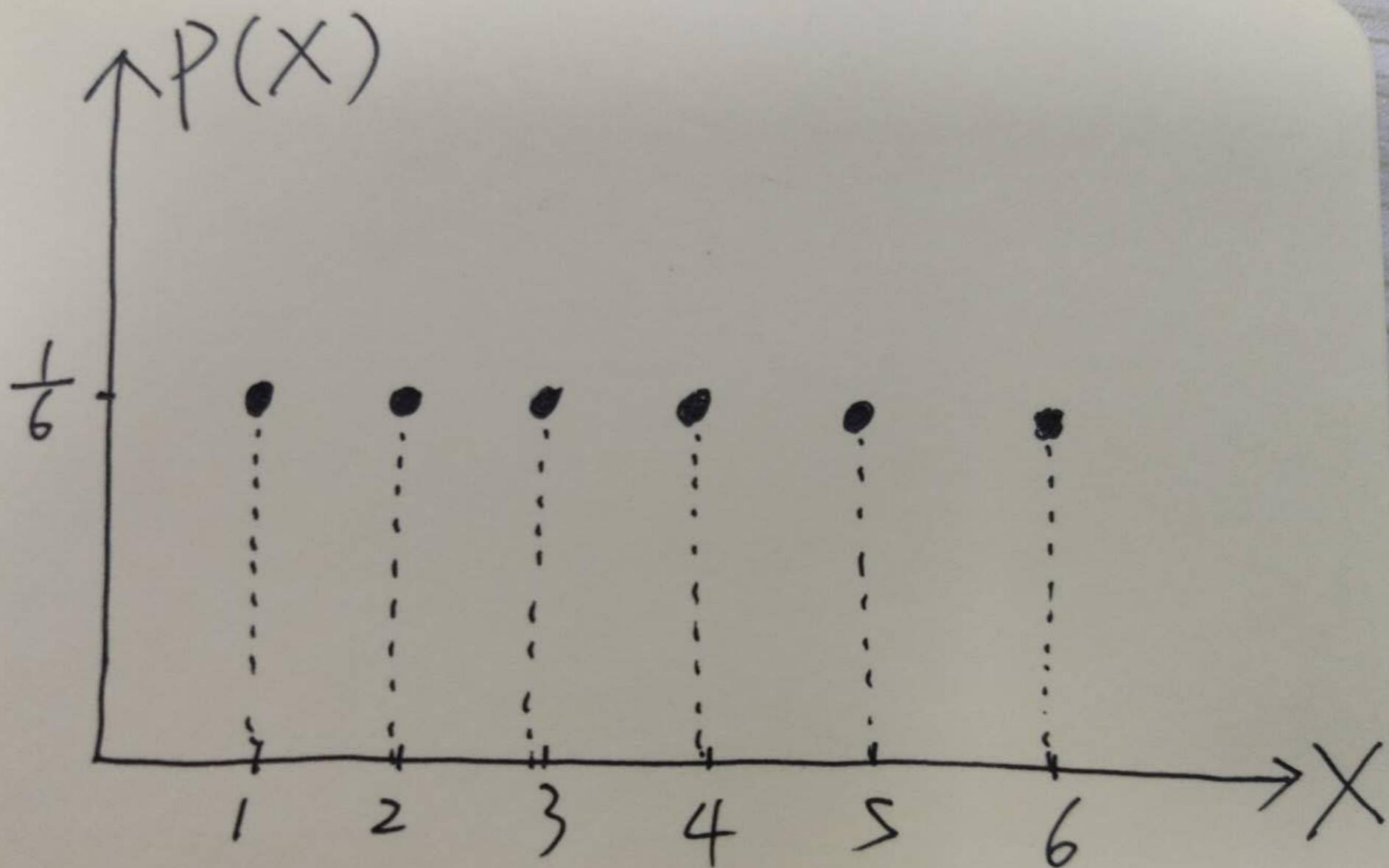
# Part 2. 统计

# 数学期望：Expectation

- 又叫均值：mean
- 与平均数average不一定相等

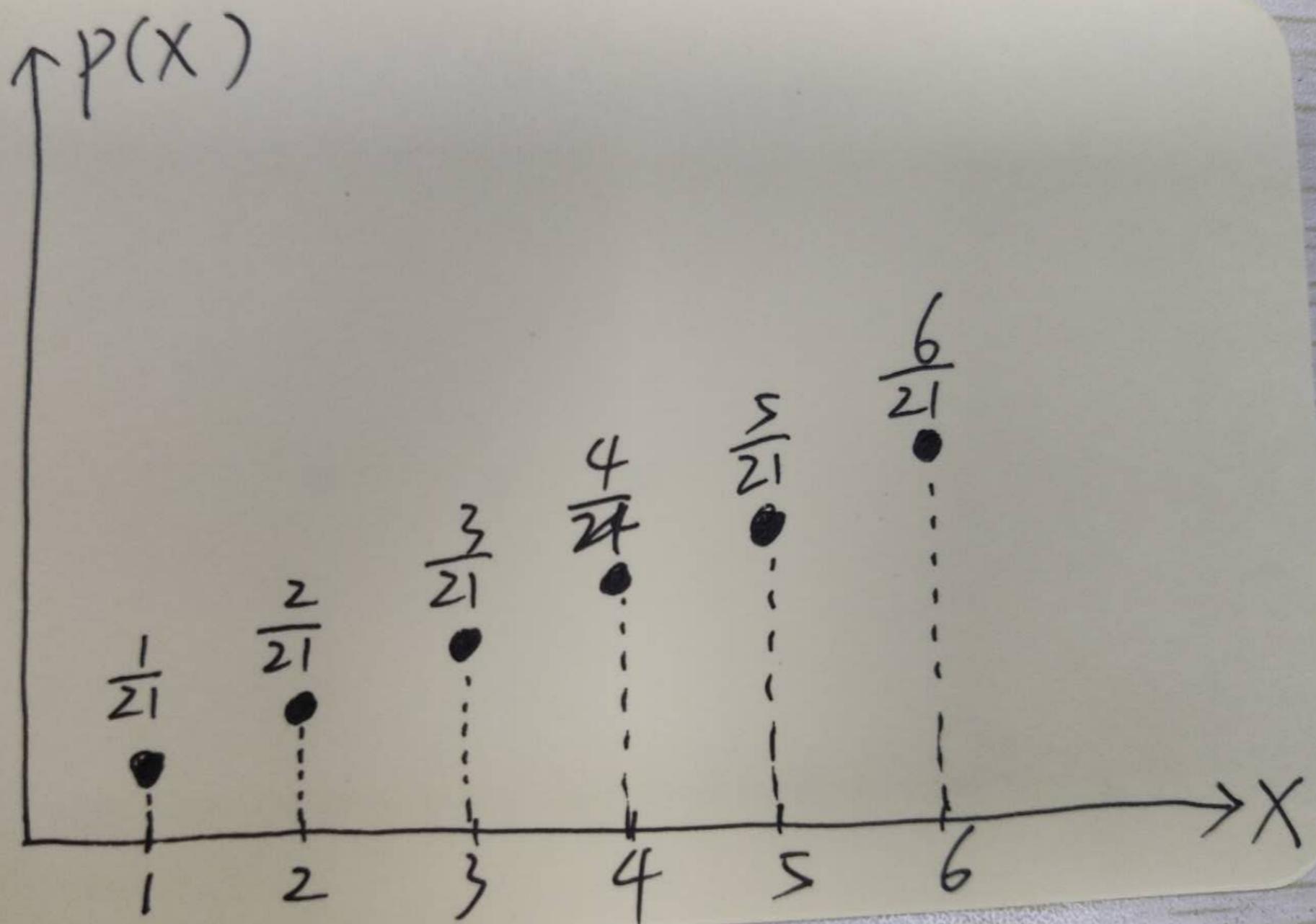
$$E(X) = \sum_{i=1}^{\infty} x_i p(x_i)$$

$$E(X) = \int_{-\infty}^{+\infty} x p(x) dx$$



$$\begin{aligned}E(X) &= \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 \\&\quad + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 \\&= \frac{1}{6} \times (1+2+3+4+5+6) \\&= 3.5\end{aligned}$$

- 均匀分布情况下，期望==平均数



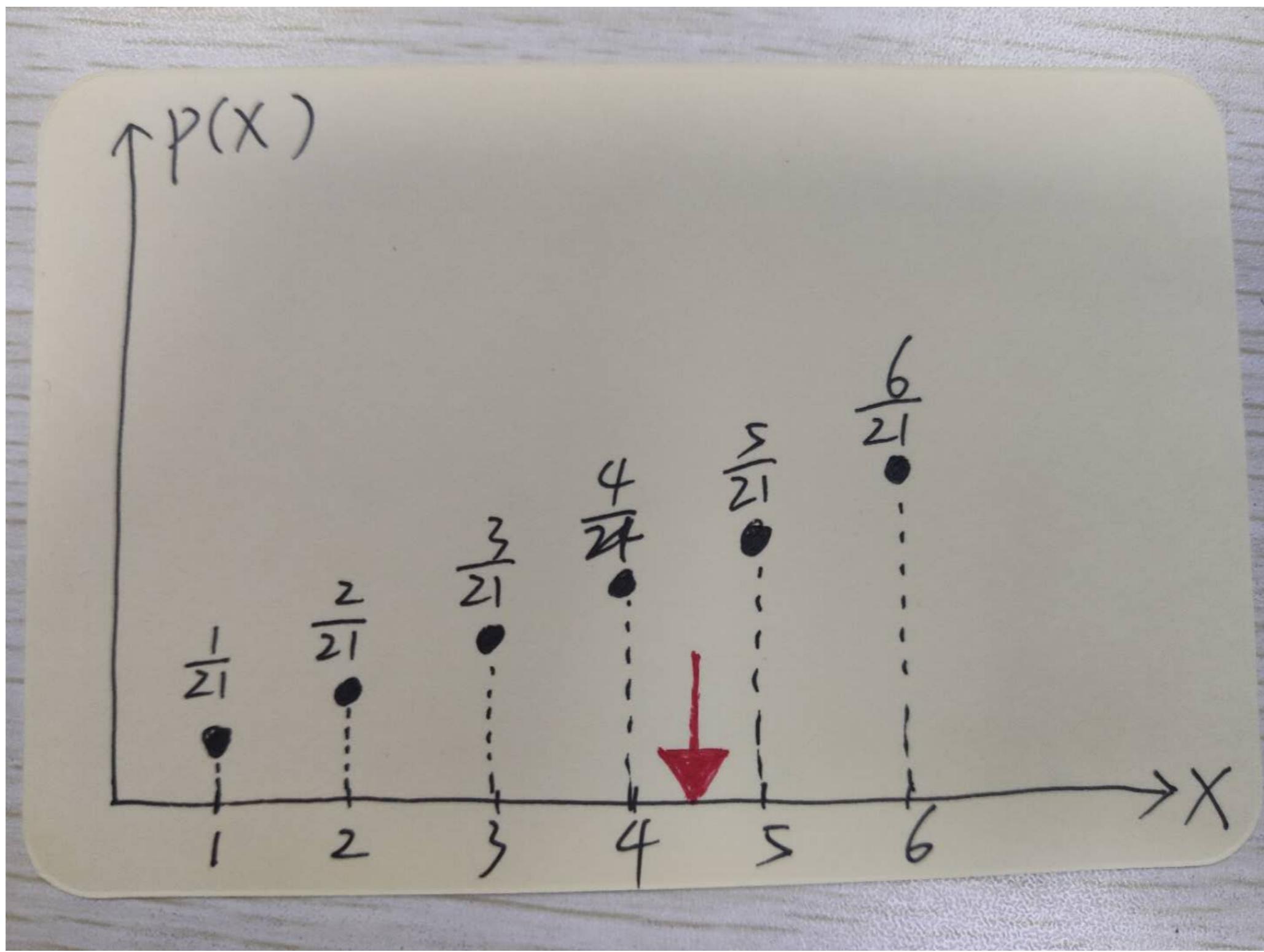
$$E(X) = 1 \times \frac{1}{21} + 2 \times \frac{2}{21} + 3 \times \frac{3}{21}$$

$$+ 4 \times \frac{4}{21} + 5 \times \frac{5}{21} + 6 \times \frac{6}{21}$$

$$= \frac{1}{21} \times (1 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2)$$

$$= 4.333$$

- 数学期望体现了概率分布的重心所在位置



# 方差：Variance

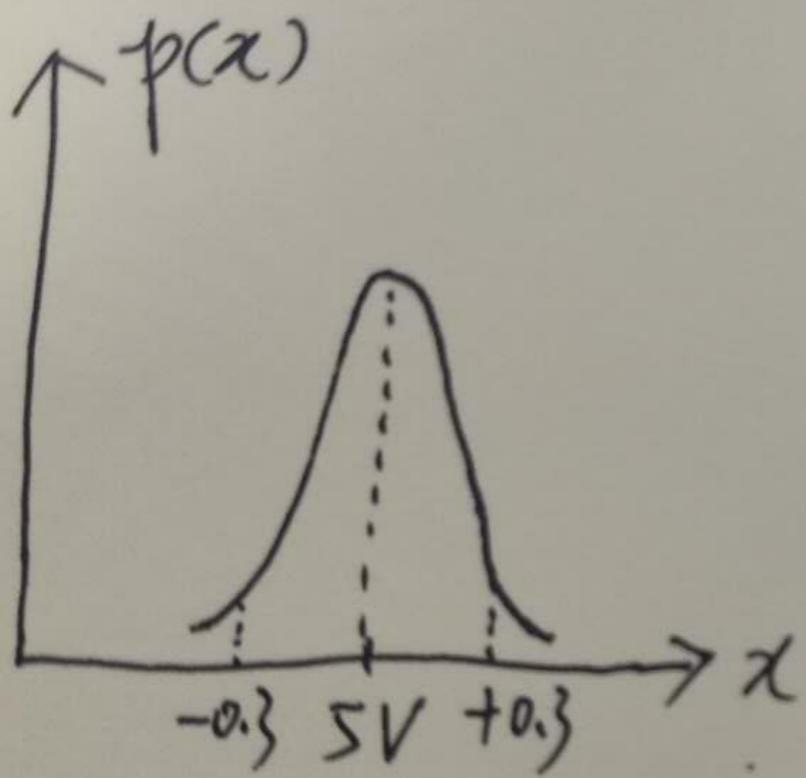
$$\text{Var}(X) = E(X - E(X))^2$$

# 标准差： standard deviation

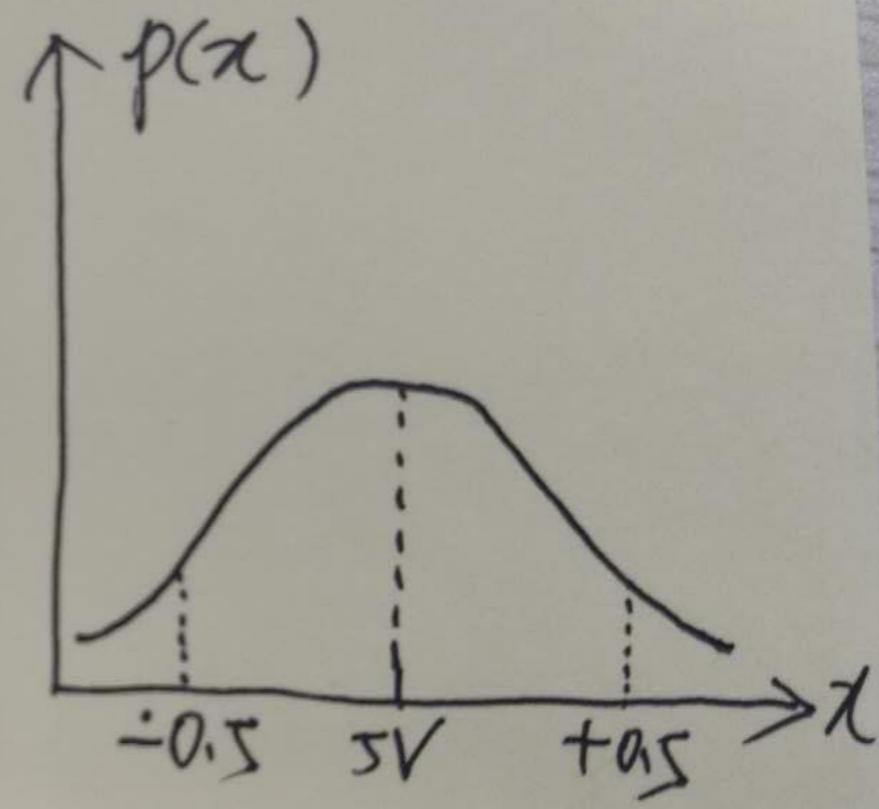
$$\sigma(X) = \sqrt{\text{Var}(X)}$$

- 方差和标准差，体现了随机变量的集中与分散程度。
- 方差和标准差越小，随机变量的取值越集中。
- 方差和标准差越大，随机变量的取值越分散。

电压： $5V \pm 0.5V$



方差小



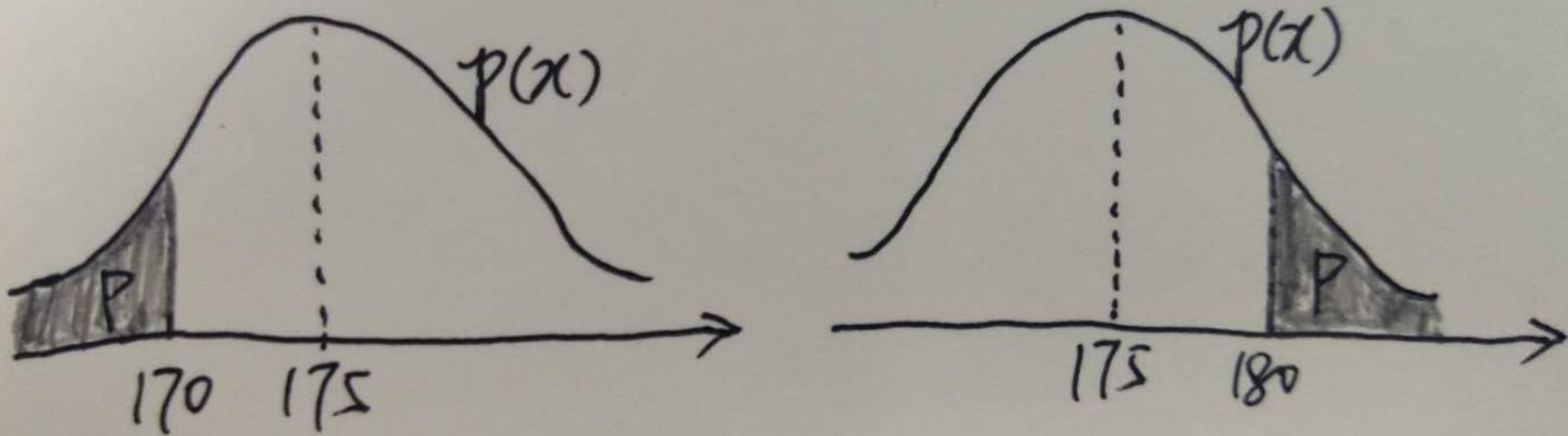
方差大

# 分位数：quantile

$$P = \int_{-\infty}^{x_p} p(x) dx \quad (\text{下侧分位数})$$

$$P' = \int_{x'_p}^{\infty} p(x) dx \quad (\text{上侧分位数})$$

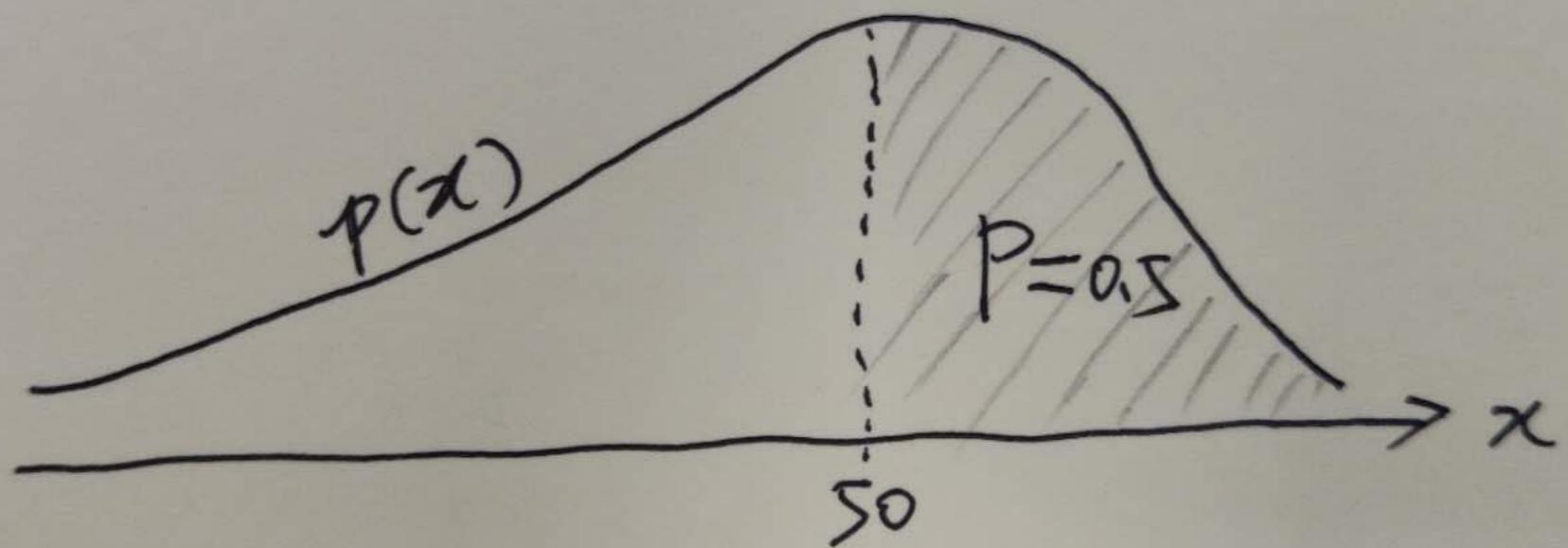
例：身高分布



# 中位数： median

- 中位数是一个特殊的分位数
- $p=0.5$
- 常记做  $m_{0.5}$

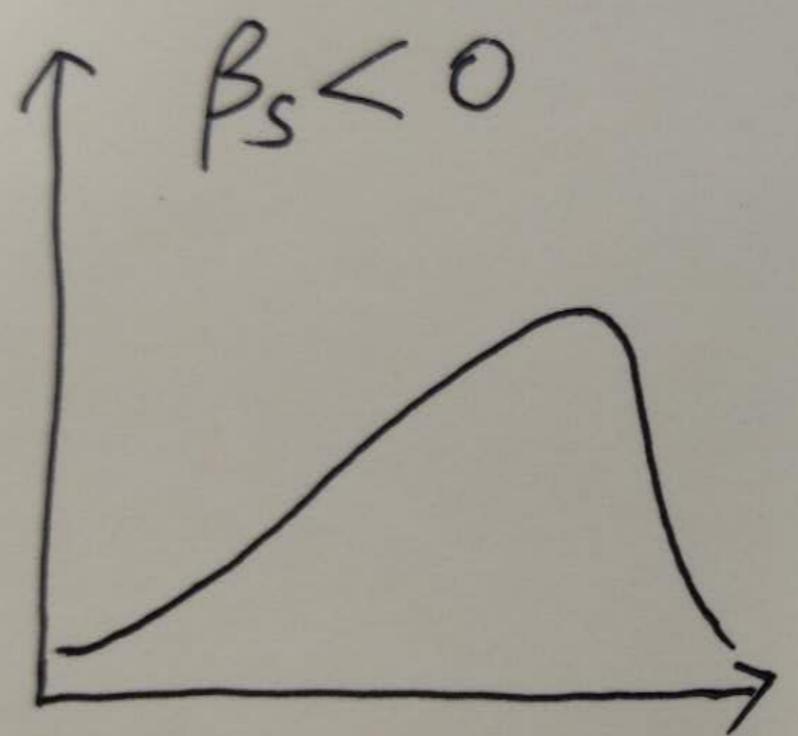
例：年齡分布



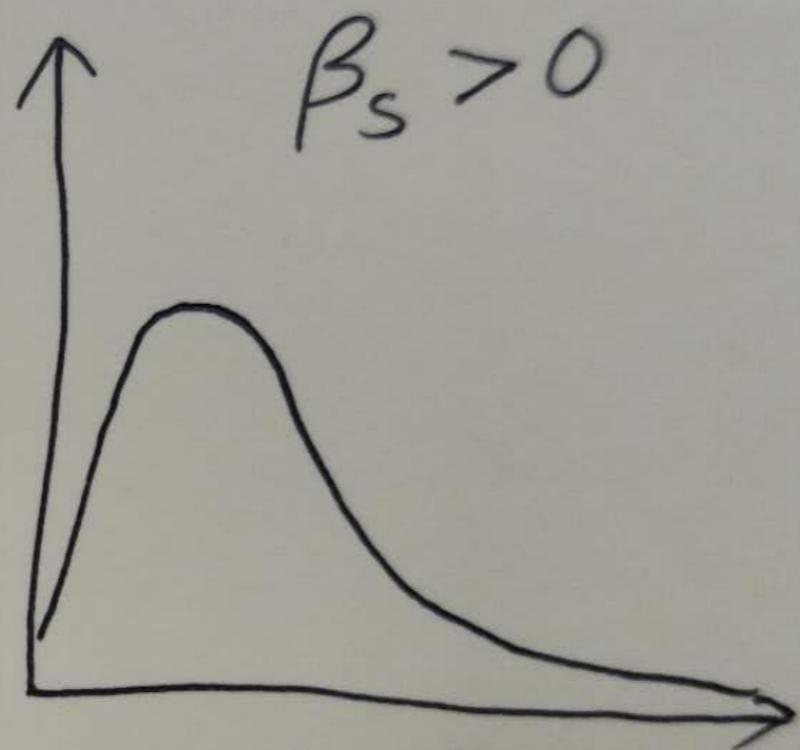
# 偏度系数：skewness

$$\beta_s = \frac{E(X - E(X))^3}{[Var(X)]^{3/2}}$$

- 偏度系数体现了概率分布偏离对称性的程度



左偏

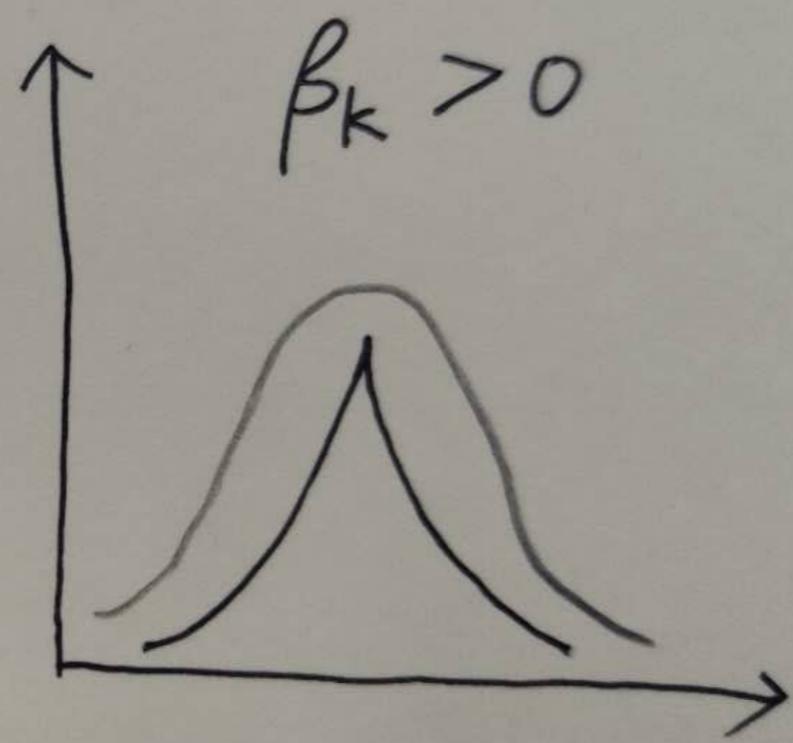
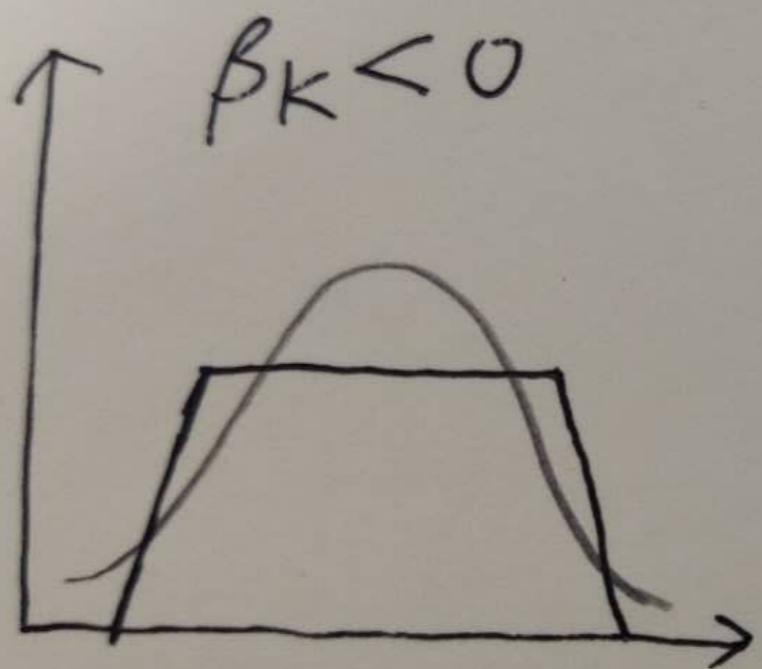


右偏

# 峰度系数：kurtosis

$$\beta_k = \frac{E(X - E(X))^4}{[Var(X)]^2} - 3$$

- 峰度系数体现了概率分布的尖峭程度



# Part 3. 实践

- 概率论：概率分布已知
- 实际情况：概率分布未知

- 总体: population
- 样本: sample
- 抽样: 从总体中随机地抽取n个个体



# 矩(法)估计

- 用样本均值估计总体均值
- 用样本方差估计总体方差
- 用样本中位数估计总体中位数
- 用样本偏度估计总体偏度
- .....

# 样本均值

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$= \frac{1}{n} \sum_{i=1}^n x_i$$

- 进行n次抽样， 得到的样本均值不一定相等。

# 样本方差

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- 注意，分母是 $n-1$
- 无偏估计

## \*对比：总体方差

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- 有了均值、方差，就可以计算标准差、中位数、偏度、峰度等等，不再赘述。

下一讲

《大数据基础：数据可视化》

谢谢！