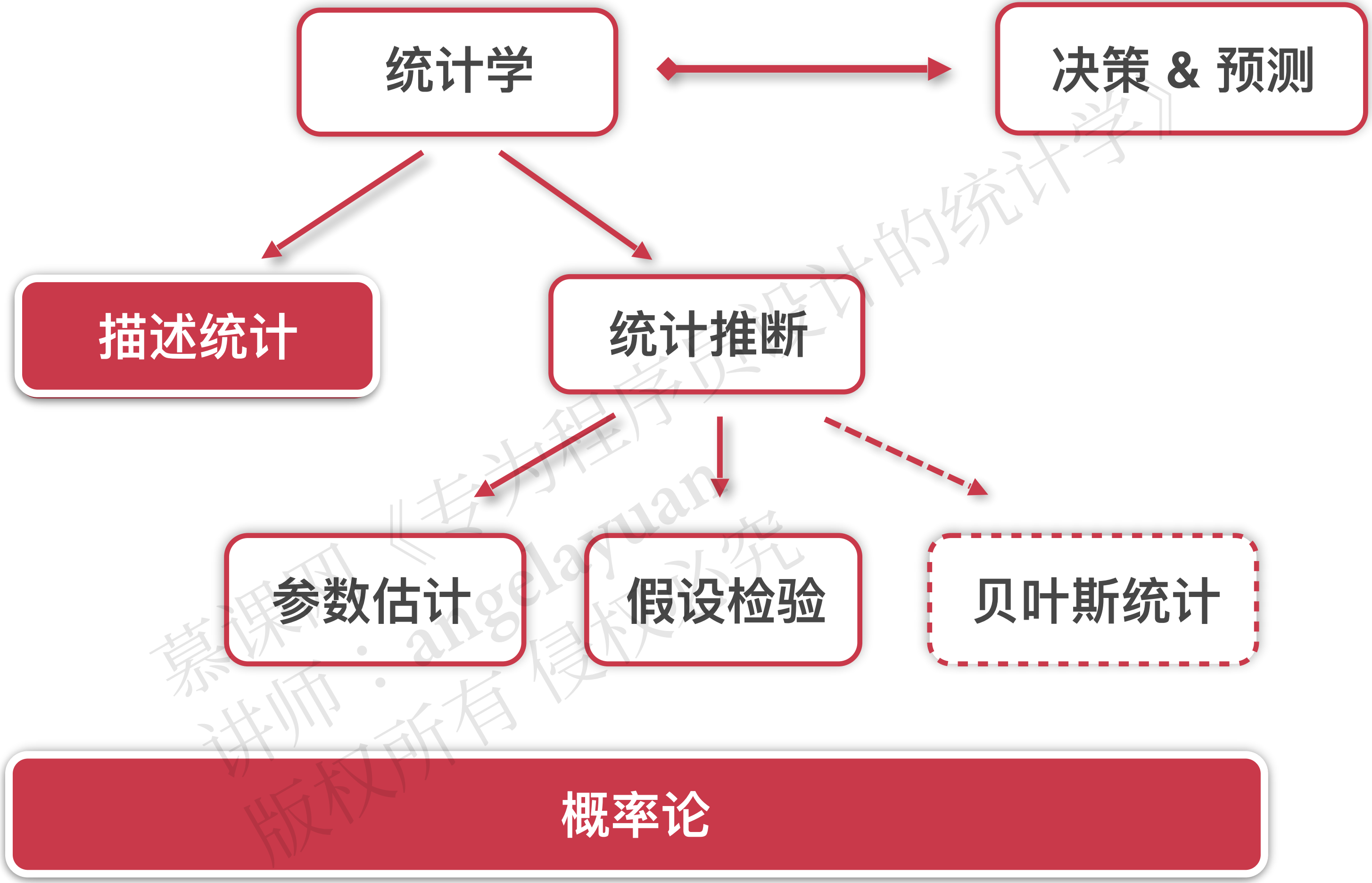
# 必须了解的概率论知识


统计学
决策 & 预测
描述统计
统计推断
参数估计
假设检验
贝叶斯统计
概率论

# 什么是概率论

# 概率论

- **确定性现象**

  抛起的硬币必然下落

  在标准大气压下，水加热到100度必然沸腾

  定义：在一定条件下必然会发生的现象

# 概率论

- **随机现象**

抛一次硬币的结果{正面, 反面}　　不确定,但肯定是正面或反面

抛一次的结果{1, 2, 3, 4, 5, 6}　　不确定,但肯定是六个点数之一

抛10000次硬币:{正, 反, 反, 正, …..}　　正面:反面 ≈ 1:1

抛10000次:{1,5,3,3,1,2,2,5,5,6,4, …..}　　1:2:3:4:5:6 ≈ 1:1:1:1:1:1

# 概率论

- **随机现象**

不确定,但肯定是正面或反面

不确定,但肯定是六个点数之一

在个别试验中

结果具有不确定性

正面:反面 ≈ 1:1

1:2:3:4:5:6 ≈ 1:1:1:1:1:1

在大量重复试验中

结果呈现出固有规律性

# 概率论

在个别试验中
结果具有不确定性

随机试验(Experiment)

- 可以在相同条件下重复进行
- 可能的结果不止一个并且能够事先明确所有可能结果
- 进行试验前不能确定哪个结果会出现

在大量重复试验中
结果呈现出固有规律性

统计规律性

# 概率论

- **样本空间(Space)**

定义: 随机试验E的所有可能结果组成的集合

试验: 扔一次硬币进行观察 ➡ S: {Head, Tail}

试验: 扔二次硬币进行观察 ➡ S: {HH, HT, TH, TT}

试验: 扔三次硬币进行观察 ➡ S: {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}

# 概率论

## ◆ 随机事件

定义: 样本空间S的子集称为随机试验E的随机事件, 简称事件

S: {Head, Tail} ➡ 子集: {H}, {T}　　基本事件

S: {HH, HT, TH, TT} ➡ 子集: {HH}, {HT}, {TH}, {TT},

{HH, HT}, {HT, TH}, …..

{HH, HT, TH}, {HT, TH, TT}, …..

空集ø也是S的子集, 它在每次试验中都不发生，称为不可能事件

# 概率论

- **随机事件**

定义: 样本空间S的子集称为随机试验E的随机事件, 简称事件

S: {HH, HT, TH, TT} → {HH, HT, TH} = {至少一次正面}

事件发生: 在每次试验中, 当且仅当这一子集的一个样本点出现

样本空间是必然事件

随机现象
在个别试验中结果具有不确定性
在大量重复试验中结果呈现出固有规律性
（统计规律性）
• 可在相同条件下重复进行
• 能够事先明确所有可能结果
• 试验前不确定哪个结果会出现
样本空间
子集
(随机)事件

# 概率

# 个人经验/知识

**明天的天气**

S: {晴，雨}

事件: {雨}

P(雨) = 0.2

P(雨) = 0.7

……

**美国总统选举**

S: {希拉里赢，希拉里输}

事件: {希拉里赢}

P(希拉里赢) = 0.7

P(希拉里赢) = 0.4

……

- 猜测
- 个人经验/知识
- 主观
- 可信度低

# 频率

在相同条件下，进行了n次试验; 在这n次试验中，

事件A发生的次数 $n_A$ 称为事件A发生的频数

比值 $n_A/n$ 称为事件A发生的频率，记为 $f_n(A)$

# 频率

进行n = 100次抛硬币

A = {H}　　$n_A = 40$

$f_n(A) = n_A/n$ = 40/100 = 0.4

**性质**　　$0 \le f_n(A) \le 1$

$f_n(S) = 1$

若 $A_1, A_2, \ldots, A_k$ 是两两互不相容事件

则 $f_n(A_1 \cup A_2 \cup \ldots \cup A_k) = f_n(A_1) + f_n(A_2) + \ldots + f_n(A_k)$

# 概率

对于随机试验E的每一个事件A赋予一个实数，记为P(A)，称为事件A的概率，如果集合函数P(.)满足下列条件

- 非负性：对于每一个事件A，有 $P(A) \geq 0$
- 规范性：对于必然事件S，有 $P(S) = 1$
- 可列可加性：设 $A_1, A_2, \ldots$ 是两两互不相容事件，有

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots$$

# 频率 vs 概率

**为什么可以用频率近似概率?**

A = {H}

抛n = 10次硬币，计算频率

抛n = 60次硬币，计算频率

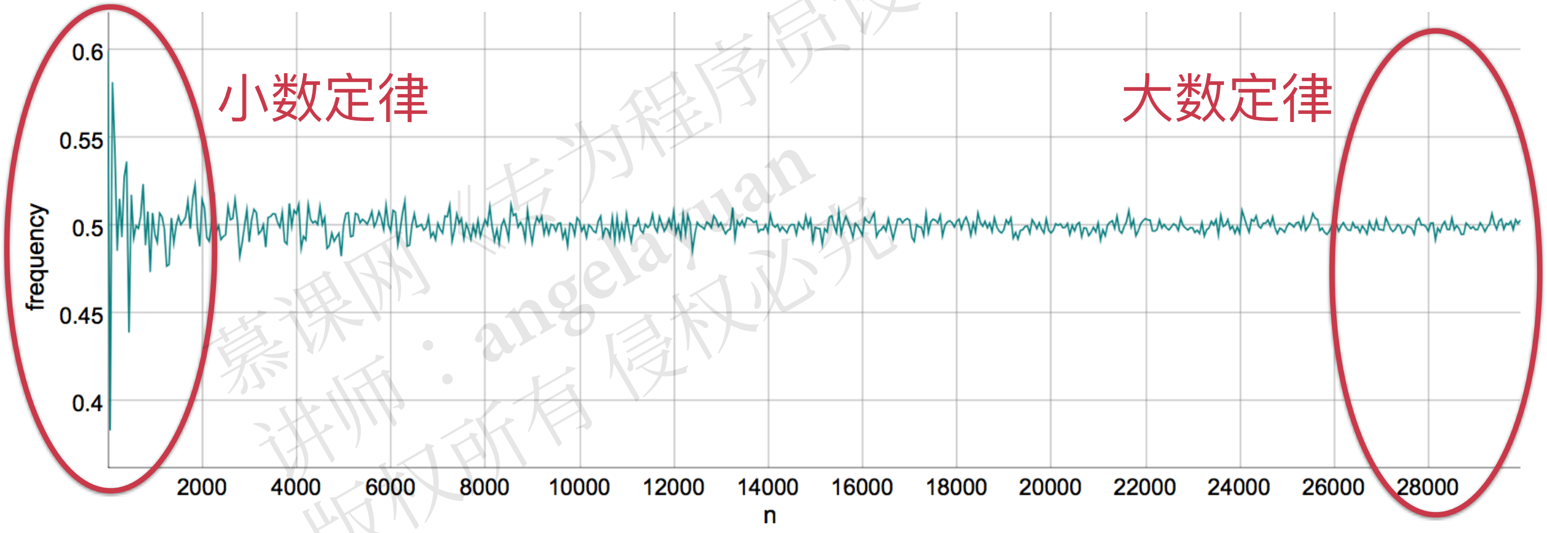
抛n = 110次硬币，计算频率

·······

抛n = 10000次硬币，计算频率

# 频率 vs 概率

为什么可以用频率近似概率?




$$n \to \infty, f_n(A) \to P(A)$$

# 编程理解小数和大数定律

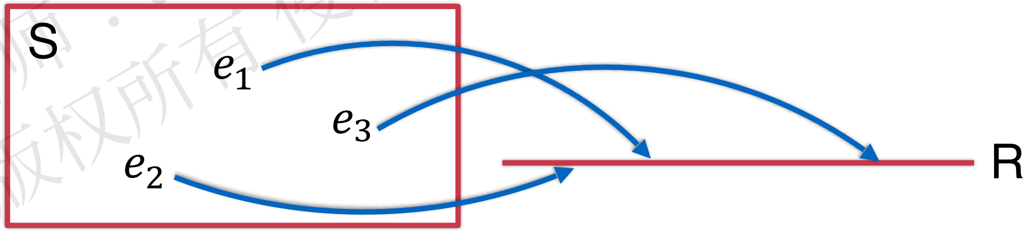
# 再谈变量

# 变量

数据是由变量组成的; 一个变量至少包含2个不同取值

如何把数据/变量与概率论中的概念联系起来?

➡ 引入一个法则, 将样本空间S的每一个元素(即随机试验E的每一个结果)与实数对应起来



# 随机变量

设随机试验E的样本空间为S = {e}, X = X(e)是定义在样本空间S上的实值单值函数, 称X = X(e)为随机变量

试验E: 抛硬币

样本空间S: {H,T}

变量X: X(H) = 1

X(T) = 0

名目尺度, 定性变量

试验E: 观测教育程度

样本空间S: {小学,初中,高中,大学}

变量X: X(小学) = 1

X(初中) = 2

X(高中) = 3

X(大学) = 4

次序尺度

定性变量

# 随机变量

如果随机试验的结果本身就是一个数，即e本身是一个数，令X = X(e) = e，X就是一个随机变量

试验E: 某大学学生的出生年份

样本空间S: {2000, 2001, 2002, 2003}

变量X: X(2000) = 2000

X(2001) = 2001

X(2002) = 2002

X(2003) = 2003

等距尺度

定量变量

# 随机变量

如果随机试验的结果本身就是一个数，即e本身是一个数，
令X = X(e) = e, X就是一个随机变量

试验E: 某大学学生的身高

样本空间S: {1.55, 1.56, ....., 1.90}

变量X: X(1.55) = 1.55
X(1.56) = 1.56
……
X(1.90) = 1.90

等比尺度
定量变量

以大写字母如X, Y, Z…表示变量
以小写字母如x, y, z…表示实数

# 随机变量

- **性质**

随机变量的取值随试验的结果而定

试验的各个结果的出现有一定的概率
因而随机变量的取值有一定的概率

在试验之前不能预知它取什么值

# 随机变量的分类

## 离散型

取值有限个或可列无限多个

## 连续型

在一定区间内可以任意取值

# 概率分布

# 离散型随机变量及其分布

# 分布律

**离散型随机变量**: 可能取到的值是有限个或可列无限多个

**随机变量X的分布律 (Probability Mass Function)**

设X所有可能取的值为 $x_k\ (k=1,2,\dots)$

X取各个可能值的概率，即事件 $\{X=x_k\}$ 的概率为

$$P\{X=x_k\}=p_k\quad(k=1,2,\dots)$$

# 分布律

$$P\{X = x_k\} = p_k \quad (k = 1,2,\dots)$$

**抛硬币** X所有可能的取值为 $x_1 = 0, x_2 = 1$

X取各个可能值的概率为 $P\{X = 0\} = 0.5; P\{X = 1\} = 0.5$

**掷** X所有可能的取值为

$$x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5, x_6 = 6$$

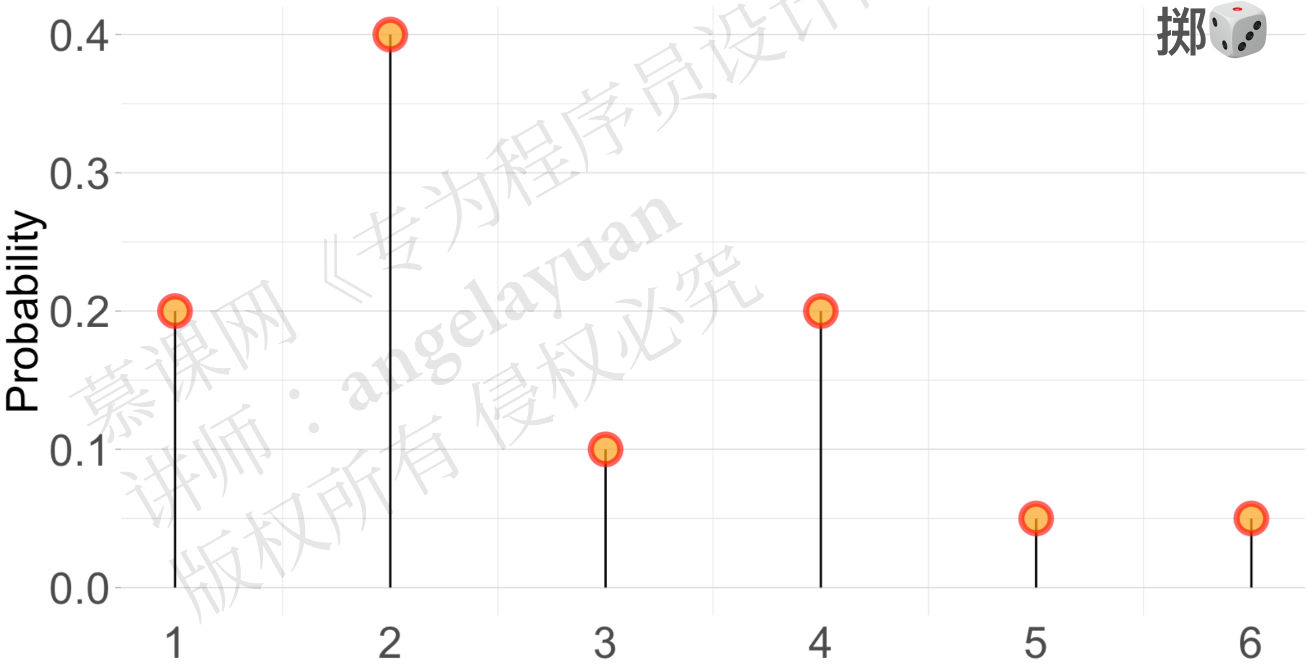
X取各个可能值的概率为

$$P\{X = 1\} = 0.2; P\{X = 2\} = 0.4; P\{X = 3\} = 0.1$$

$$P\{X = 4\} = 0.2; P\{X = 5\} = 0.05; P\{X = 6\} = 0.05$$

# 分布律

$$P\{X = x_k\} = p_k \quad (k = 1,2, \dots)$$



# 两种重要的分布

- (0-1)分布/两点分布 (Bernoulli distribution)
- 伯努利试验，二项分布 (Binomial distribution)

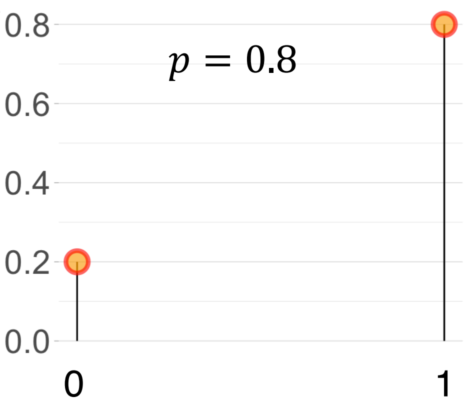
# (0-1)分布/两点分布 (Bernoulli distribution)

设随机变量X只可能取0与1两个值，它的分布律是

$$P\{X=k\}=p^k(1-p)^{1-k},(k=0,1;0<p<1)$$

则称X服从以 $p$ 为参数的(0-1)分布或两点分布

| $X$ | 0 | 1 |
|---|---|---|
| $P\{X=k\}$ | $1-p$ | $p$ |



# (0-1)分布/两点分布 (Bernoulli distribution)

| $X$ | 0 | 1 |
|---|---|---|
| $P\{X=k\}$ | $1-p$ | $p$ |

对于随机试验E, 如果其样本空间S只包含两个元素, 总能够在S上定义一个服从(0-1)分布的随机变量来描述这个随机试验的结果

性别: X=0 当性别为女; $P\{女\}=1-p$

X=1当性别为男; $P\{男\}=p$

# 伯努利试验，二项分布 (Binomial distribution)

设试验E只有两个可能的结果: $A, \bar{A}$ , 则称E为伯努利试验

抛硬币　性别

设 $P(A) = p \ \ (0 < p < 1)$, 此时 $P(\bar{A}) = 1 - p$

将E独立重复地进行n次, 则称这一串重复的独立试验为n重伯努利试验

抛n次硬币　记录n个人的性别

# 伯努利试验，二项分布 (Binomial distribution)

以X表示n重伯努利试验中事件A发生的次数, X是一个随机变量, X的所有可能的取值为0,1,2,…,n, 其分布律为

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\ldots,n;0<p<1)$$

H H ….. H T T ….. T

前k次正面朝上 后(n-k)次反面朝上

H T T H ….. H T H

T H T H ….. H H T

# 伯努利试验，二项分布 (Binomial distribution)

以X表示n重伯努利试验中事件A发生的次数, X是一个随机变量, X的所有可能的取值为0,1,2,…,n, 其分布律为

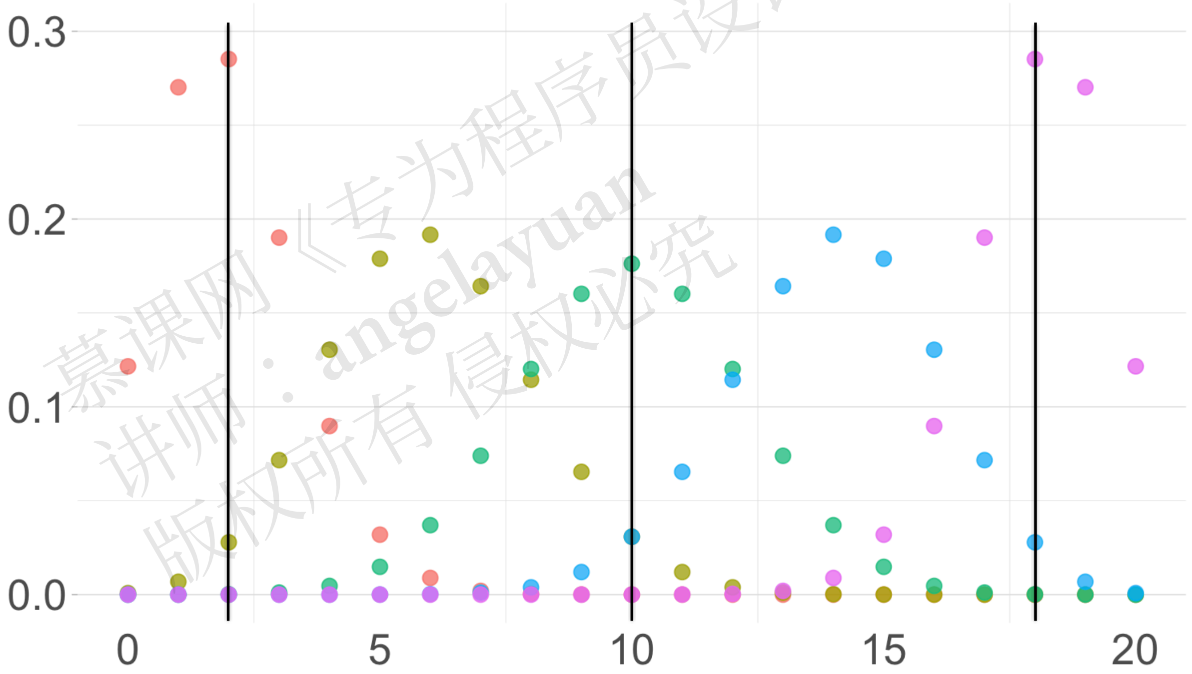
$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\dots,n;0<p<1)$$

其中 $\binom{n}{k}=\dfrac{n!}{k!\,(n-k)!}$ $\quad n!=n\times(n-1)\times\cdots\times 2\times 1$

则称X服从参数为 $\mathrm{n},p$ 的二项分布, 记为 $X\sim b(\mathrm{n},p)$

# 伯努利试验，二项分布 (Binomial distribution)

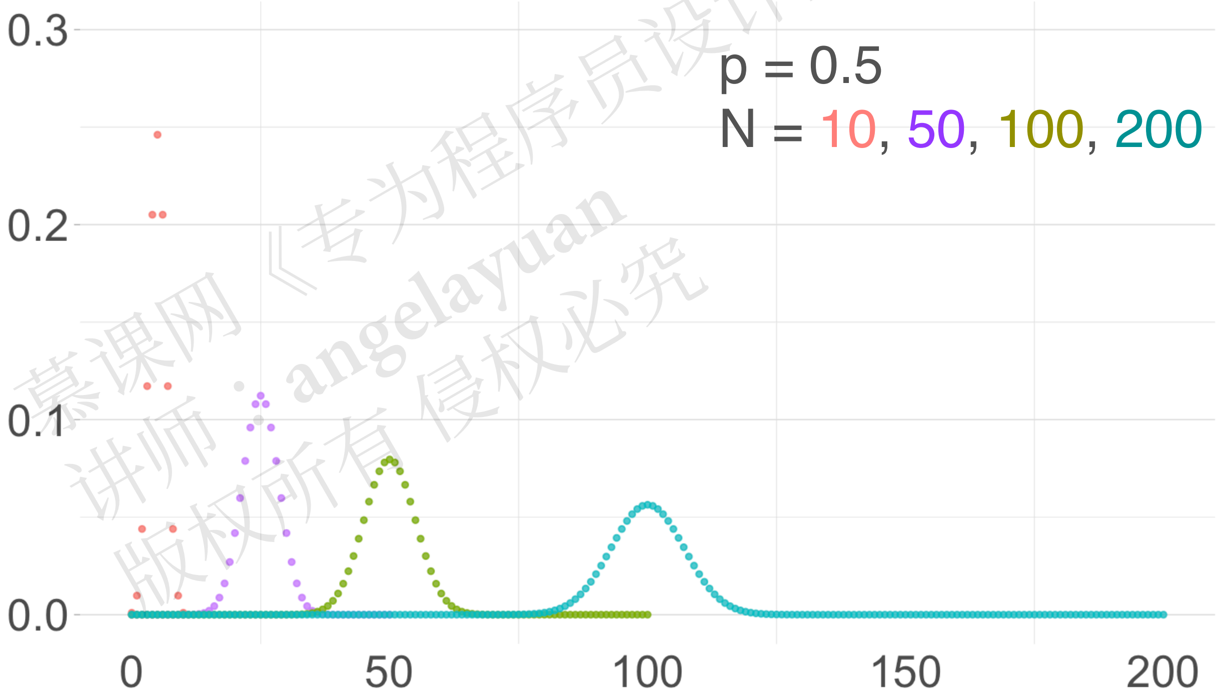
$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\ldots,n;0<p<1)$$




n = 20
p = 0.1, 0.3,
0.5, 0.7, 0.9

# 伯努利试验，二项分布 (Binomial distribution)

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k}, (k=0,1,2,\dots,n; 0<p<1)$$



# 伯努利试验，二项分布 (Binomial distribution)

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\ldots,n;0<p<1)$$

每次抛硬币, 正面朝上的概率是p

抛n次硬币, 其中有k次正面朝上的概率是多少?

某地区, 一个新生儿是女孩的概率是p

现有n个新生儿, 其中有k个是女孩的概率是多少?

# 连续型随机变量及其分布

# 分布函数

设X是一个随机变量，x是任意实数，函数

$$F(x) = P\{X \leq x\}, -\infty < x < \infty$$

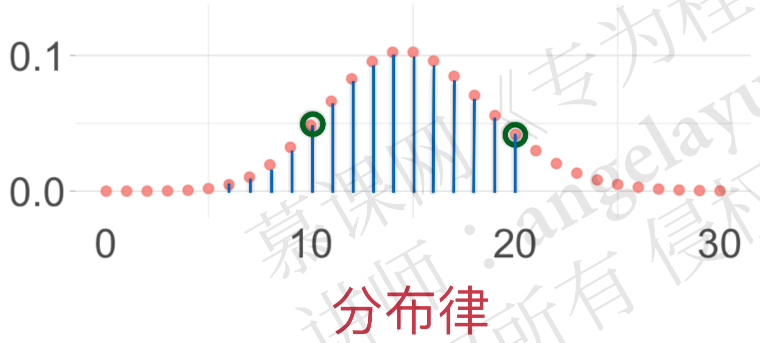
称为X的分布函数(Cumulative Distribution Function)

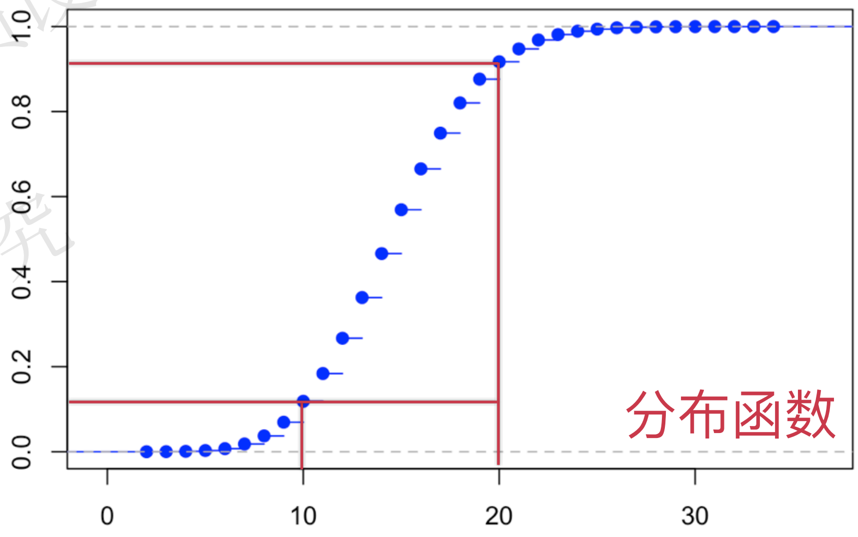
分布函数适用于离散型随机变量和连续型随机变量

# 分布函数

$$P\{X = x_k\} = p_k \ \ (k = 1,2,\ldots)$$

$$F(x) = P\{X \le x\}, -\infty < x < \infty$$



# 分布函数

若已知X的分布函数，就可知X落在任一区间$(x_1, x_2]$的概率

分布函数完整地描述了随机变量的统计规律性

# 概率密度

- **连续型随机变量**

如果对于随机变量X的分布函数F(x),存在非负函数f(x)

使对于任意实数x有 $F(x) = \int_{-\infty}^{x} f(t)d(t)$

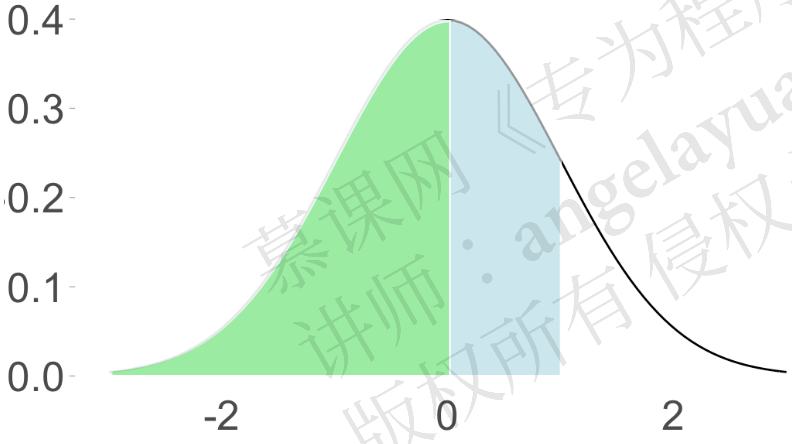
则称X为连续型随机变量

函数f(x)称为X的概率密度函数(Probability Density Function)

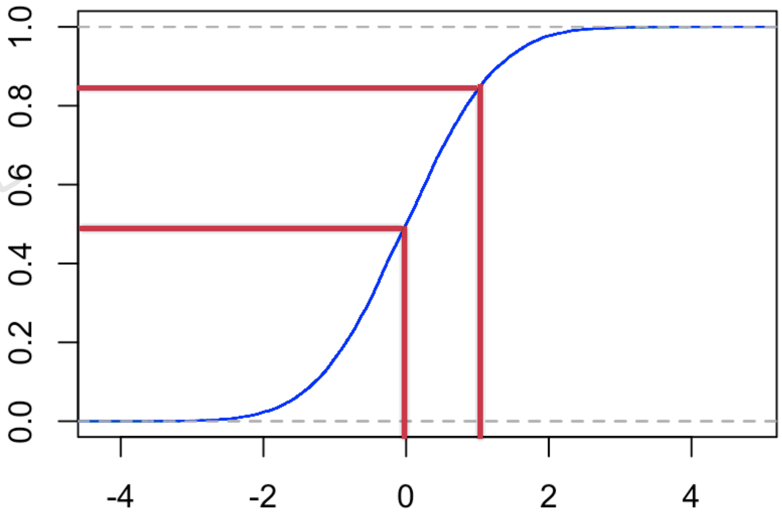
# 概率密度

$$F(x) = \int_{-\infty}^{x} f(t)d(t)$$

概率密度 $f(x)$




分布函数 $F(x)$



# 概率密度

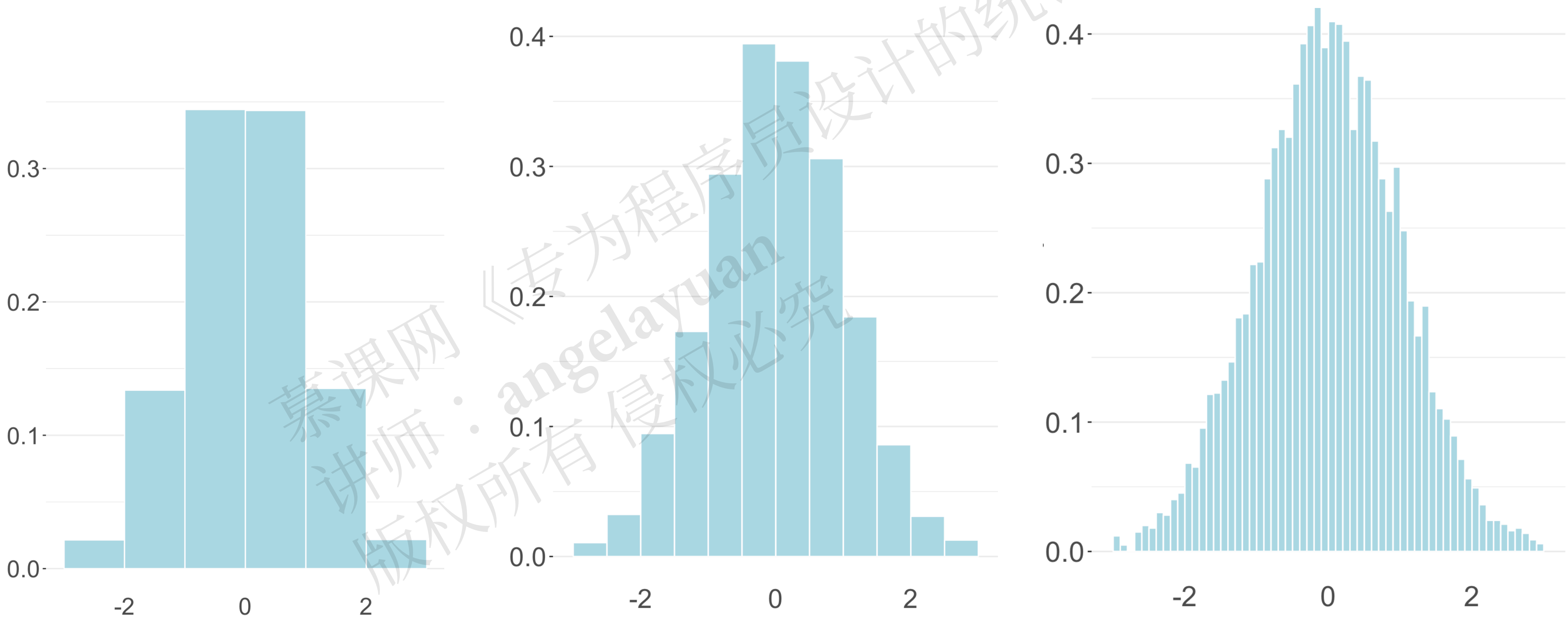
对于连续型随机变量X而言，它取任一指定实数值a的概率均为0

$$P\{X = a\} = 0$$

若A是不可能事件，则有P(A) = 0

若P(A) = 0，并不一定意味着A是不可能事件

# 概率密度



# 正态分布

# 正态分布 (Normal distribution)

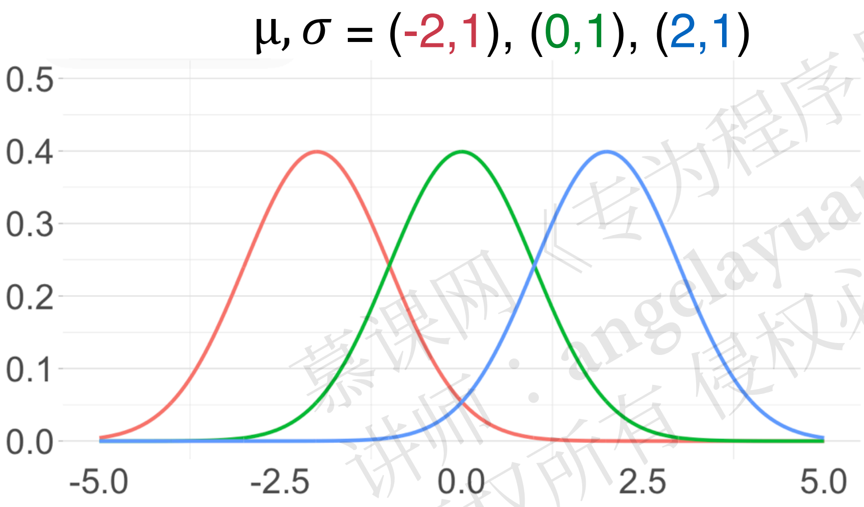
若连续型随机变量X的概率密度为

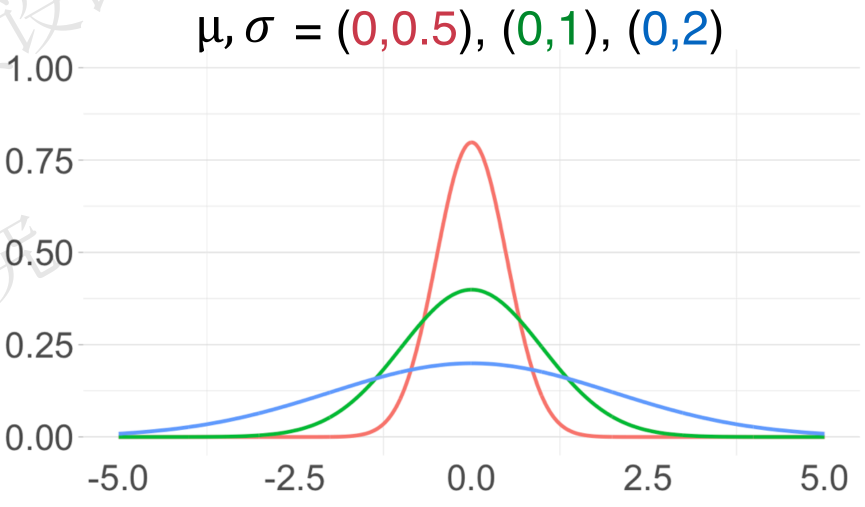
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},(-\infty<x<\infty)$$

其中 $\mu,\sigma(\sigma>0)$ 为常数, 则称X服从参数为 $\mu,\sigma$

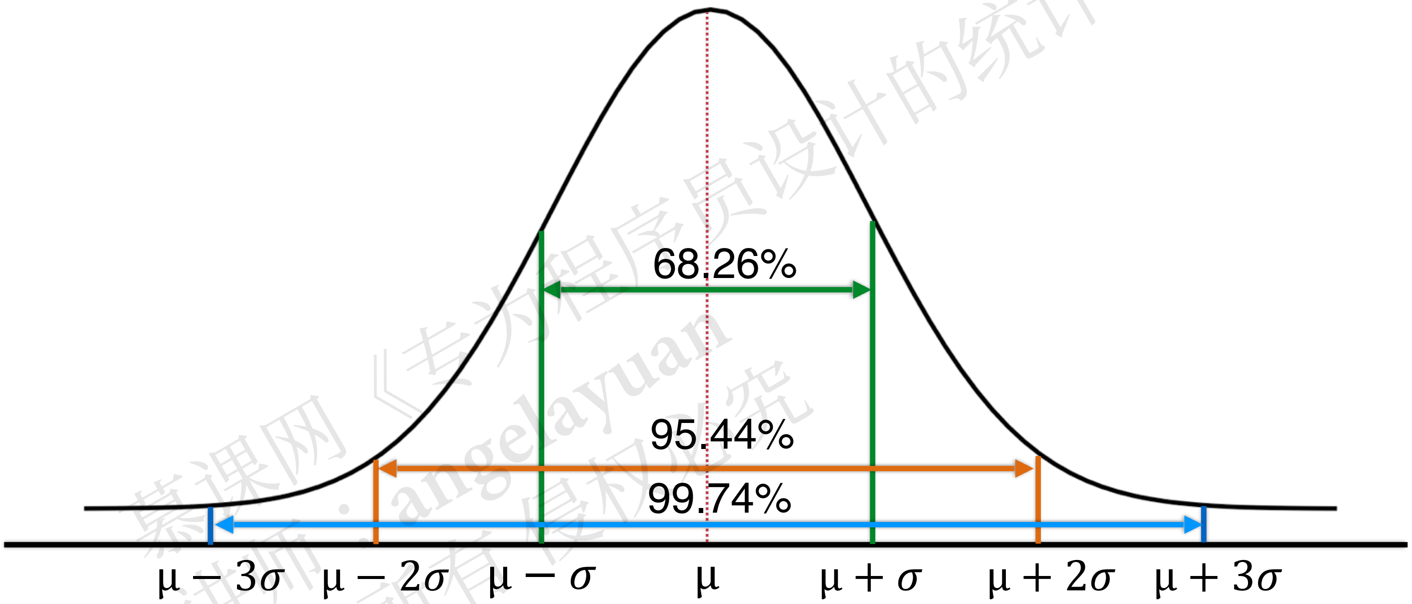
的正态分布或高斯(Gauss)分布, 记为 $X\sim N(\mu,\sigma^2)$

# 正态分布 (Normal distribution)



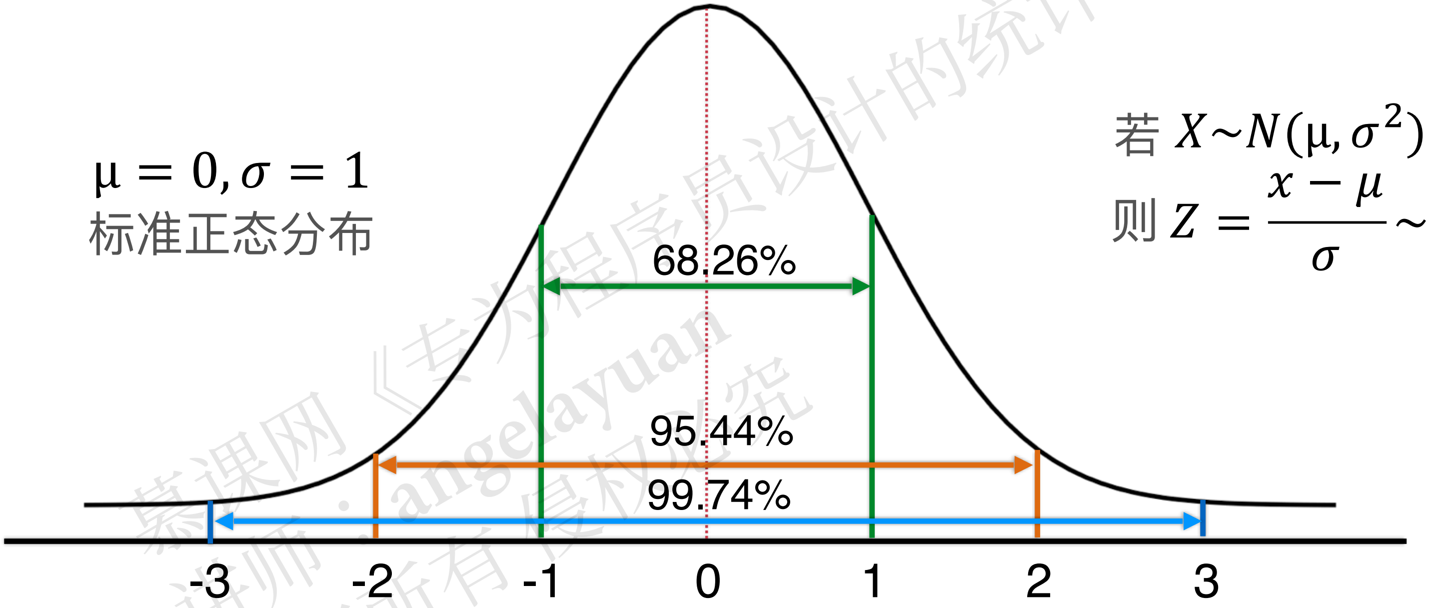
# 正态分布 (Normal distribution)




3σ法则: 正态变量的取值范围是正负无穷, 但它的值几乎肯定落在$(\mu-3\sigma, \mu+3\sigma)$

# 正态分布 (Normal distribution)

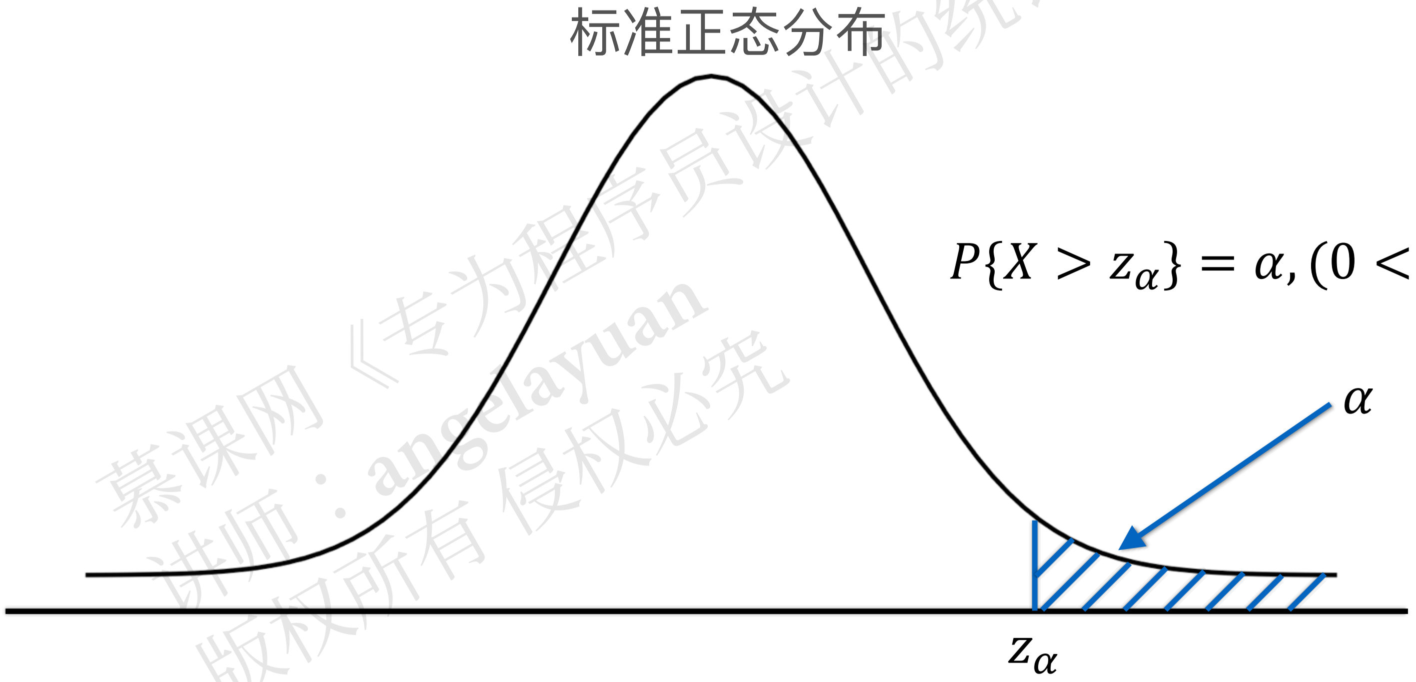
$\mu = 0, \sigma = 1$
标准正态分布




若 $X \sim N(\mu, \sigma^2)$

则 $Z = \dfrac{x-\mu}{\sigma} \sim N(0,1)$

# 正态分布 (Normal distribution)




$$P\{X > z_\alpha\} = \alpha, (0 < \alpha < 1)$$

标准正态分布的上$\alpha$分位点

# 本章小结

## 必须了解的概率论知识

- 随机现象
- 样本空间
- 随机事件

- 概率分布
  - 两点分布、二项分布
  - 正态分布

- 把数据/变量和概率论中的概念建立联系

- 概率
- 频率vs概率
- 小数/大数定律