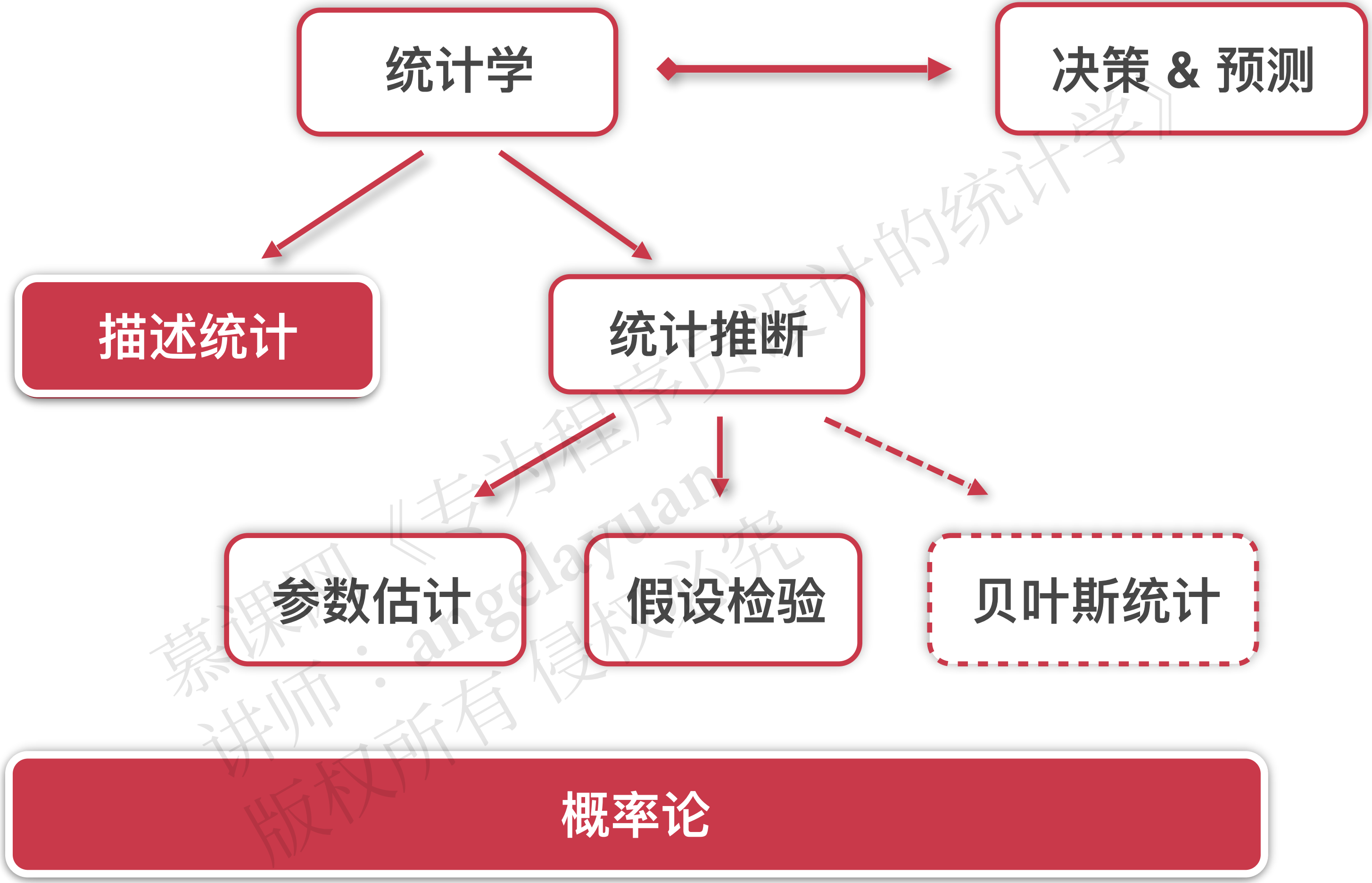
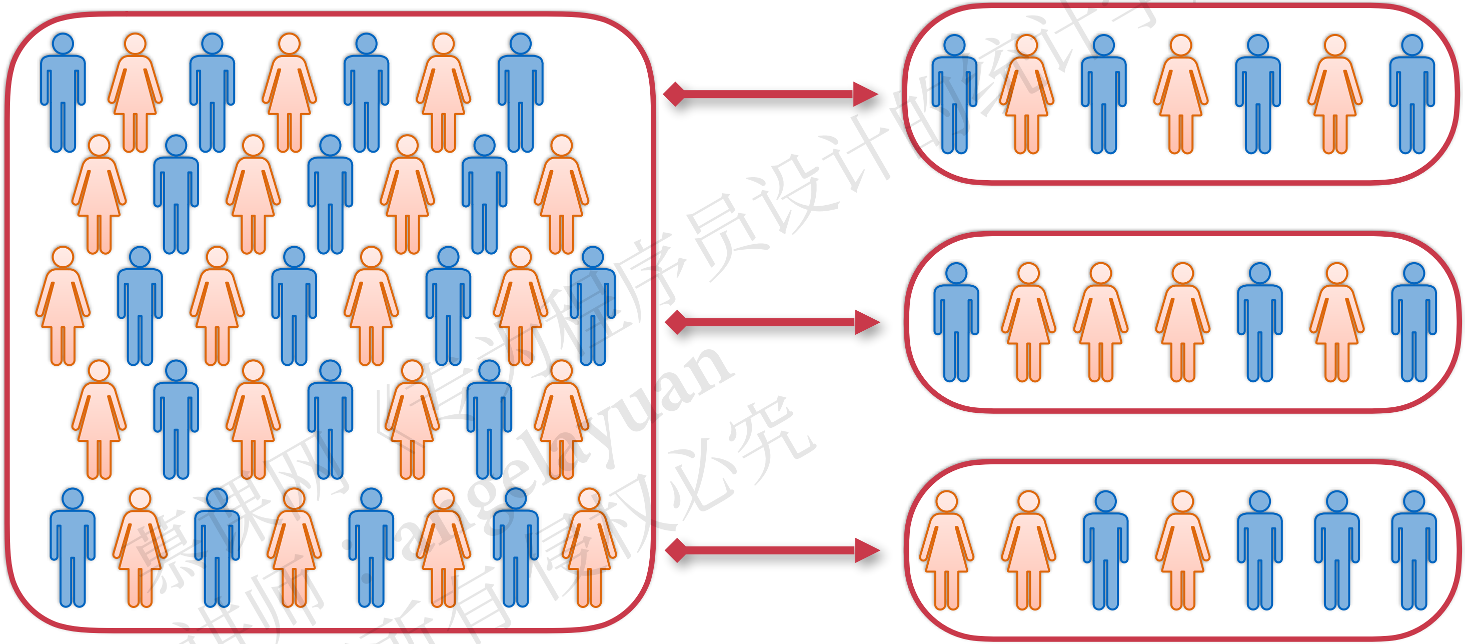

统计学
决策 & 预测
描述统计
统计推断
参数估计
假设检验
贝叶斯统计
概率论

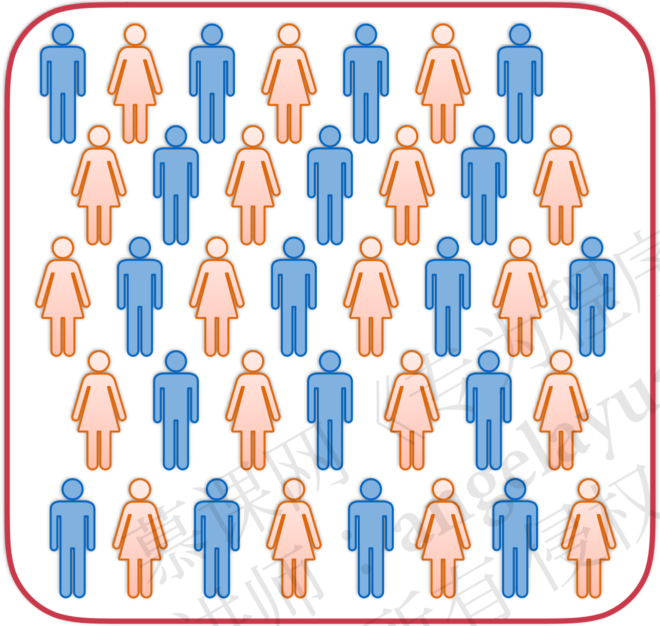
样本及抽样分布

总体与样本

总体(population)　　样本(sample)

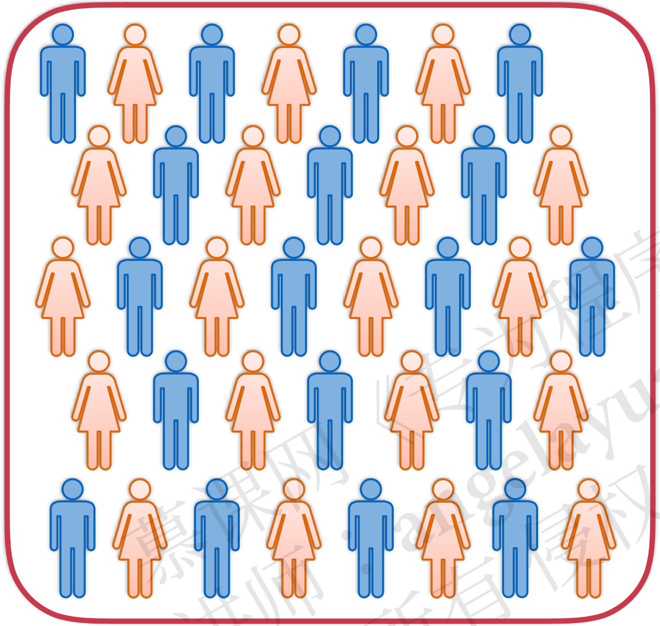


总体(population)



- 总体: 试验的全部可能的观察值
- 个体: 每一个观察值
- 总体的容量: 总体中所包含的个体的个数

总体(population)

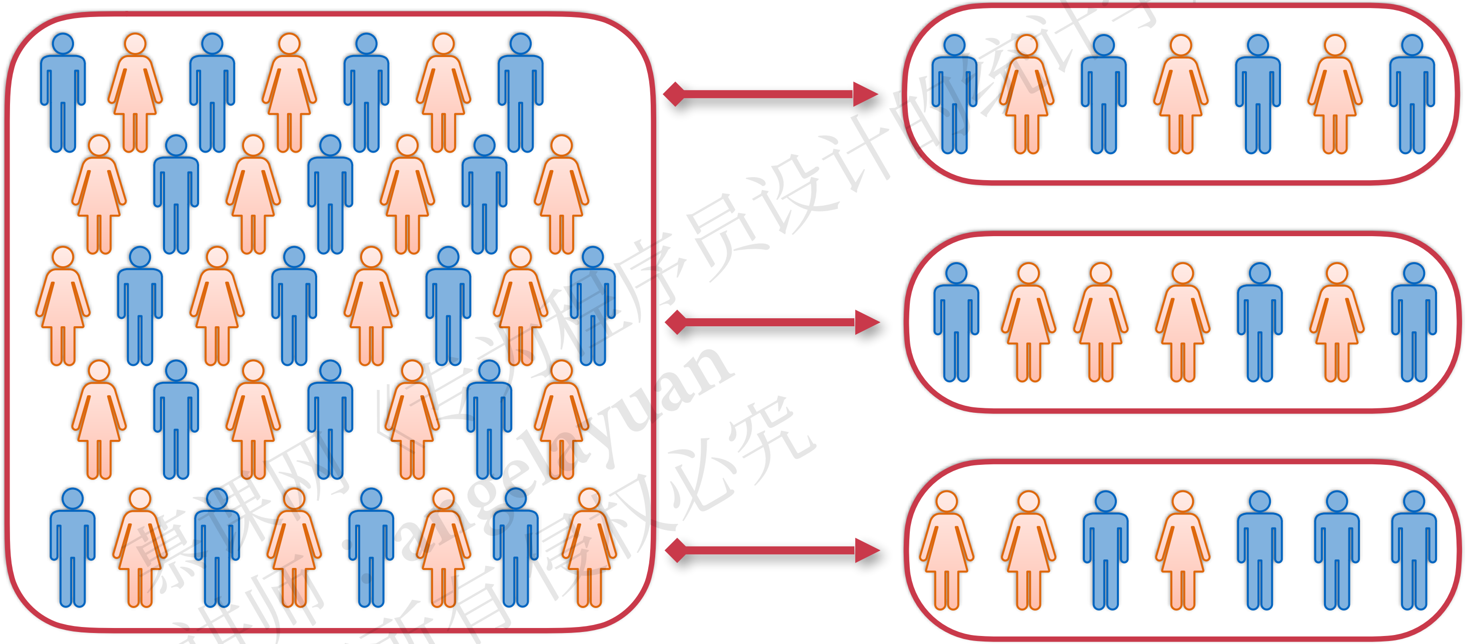


- 一个总体对应于一个随机变量X
- 对总体的研究就是对一个随机变量X的研究，X的分布函数和数字特征就称为总体的分布函数和数字特征

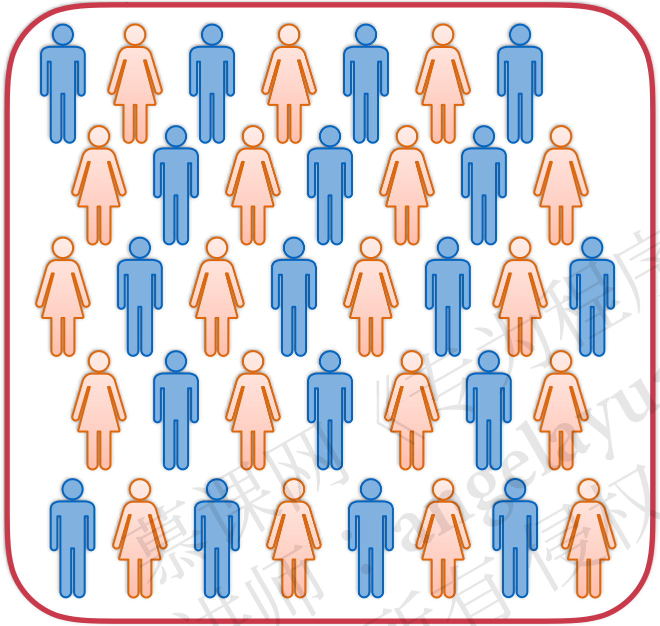
- 在实际中，总体的分布一般是未知的，或只知道它具有某种形式而其中包含着未知参数
- 从总体中抽取一部分个体，根据获得的数据来对总体分布做出推断。被抽出的部分个体叫做总体的一个样本
- 样本是进行统计推断的依据

总体(population)

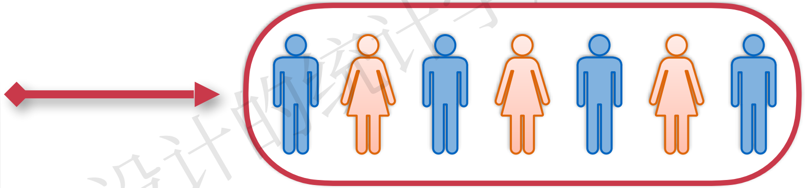
样本(sample)



总体(population)



样本(sample)

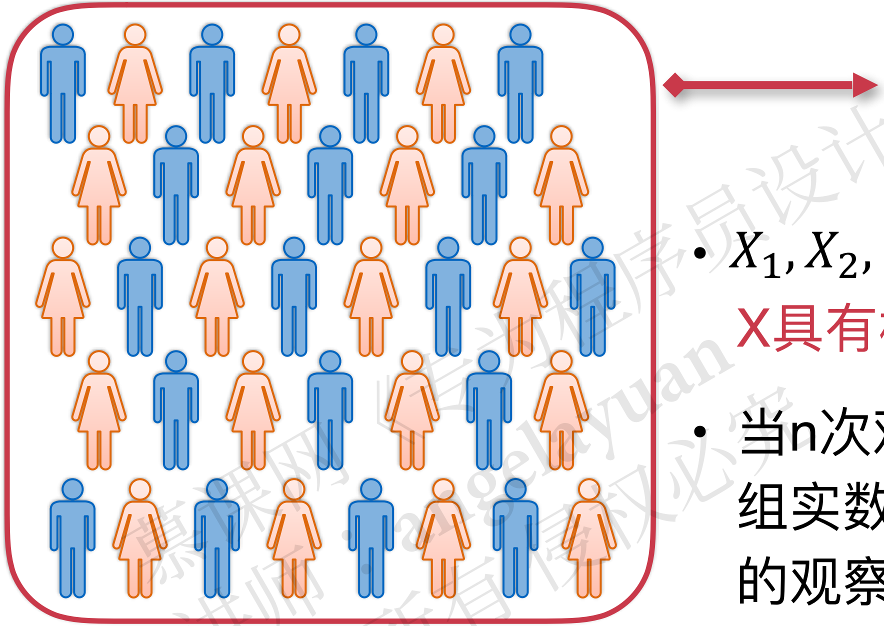


- 从总体抽取一个个体，就是对总体X进行一次观察并记录其结果
- 在相同的条件下对总体X进行n次重复的、独立的观察，将n次观察结果按试验的次序记为 $X_1, X_2, \ldots X_n$

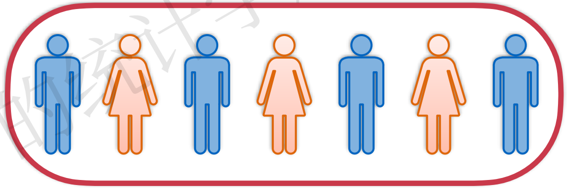
来自总体X的一个简单随机样本

n称为样本容量

总体(population)



样本(sample)



- $X_1, X_2, \ldots X_n$ 是互相独立的，并且都是与X具有相同的分布的随机变量
- 当n次观察一经完成，我们就得到一组实数 $x_1, x_2, \ldots x_n$，它们是$X_1, X_2, \ldots X_n$的观察值，称为样本值

抽样分布

我们往往不是直接使用样本本身，而是针对不同的问题构造样本的适当函数，利用这些样本的函数进行统计推断

统计量

$X_1, X_2, \dots X_n$ 是来自总体X的一个样本

$g(X_1, X_2, \dots X_n)$ 是 $X_1, X_2, \dots X_n$ 的函数

若g中不含未知参数

则称 $g(X_1, X_2, \dots X_n)$ 是一个**统计量**

统计量

$X_1, X_2, \ldots X_n$ (随机变量)

$x_1, x_2, \ldots x_n$ (样本值)

$g(X_1, X_2, \ldots X_n)$ (随机变量)

$g(x_1, x_2, \ldots x_n)$ (观察值)

常用统计量

$X_1, X_2, \dots X_n$ 是来自总体X的一个样本

$x_1, x_2, \dots x_n$ 是这一样本的样本值

- 样本均值 $g(X_1, X_2, \dots X_n) = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

  其观察值为 $g(x_1, x_2, \dots x_n) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

常用统计量

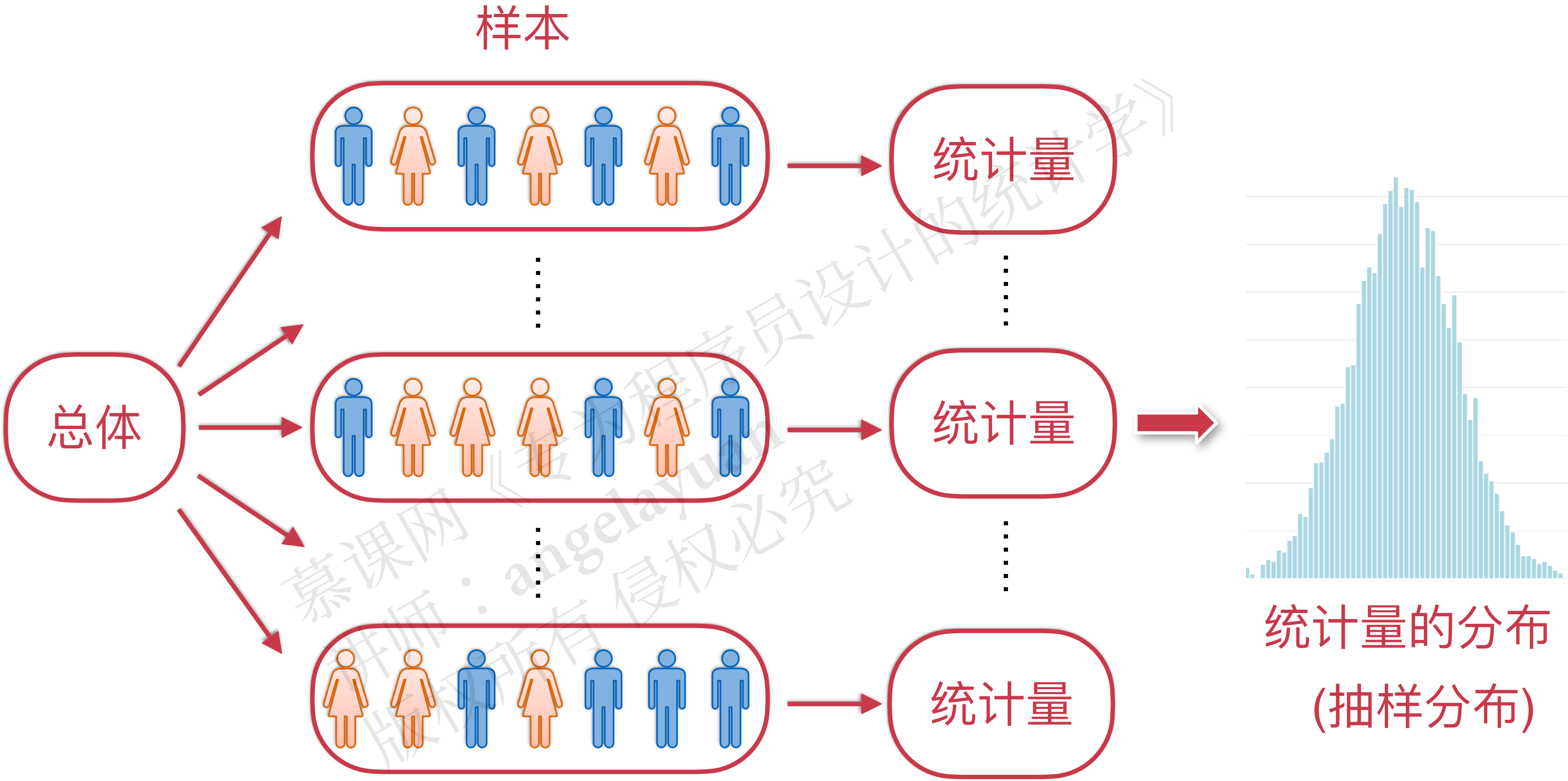
$X_1, X_2, \dots X_n$ 是来自总体X的一个样本

$x_1, x_2, \dots x_n$ 是这一样本的样本值

- 样本方差 $S^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

  其观察值为 $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

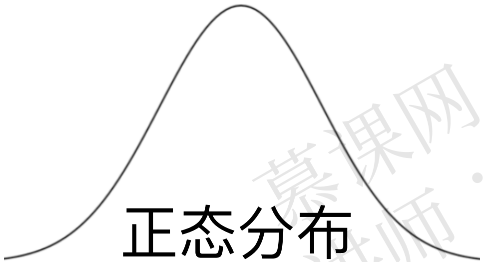
- 样本标准差 $S = \sqrt{S^2}$


样本
总体
统计量
统计量
统计量
统计量的分布
(抽样分布)

来自正态总体的几个常用统计量的分布

总体 → 样本 → 统计量 →

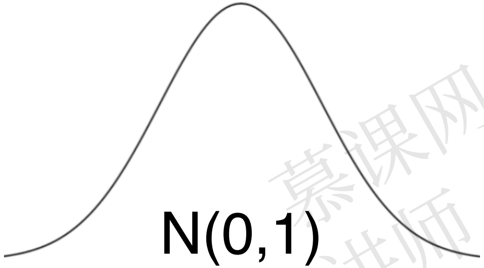
- 卡方分布
- t分布
- F分布



正态分布

卡方分布(chi-square distribution)

总体 → $X_1, X_2, \ldots X_n$ 是来自总体N(0,1)的一个样本 → 统计量



标准正态分布

$$\chi^2 = {X_1}^2 + {X_2}^2 + \cdots + {X_n}^2$$

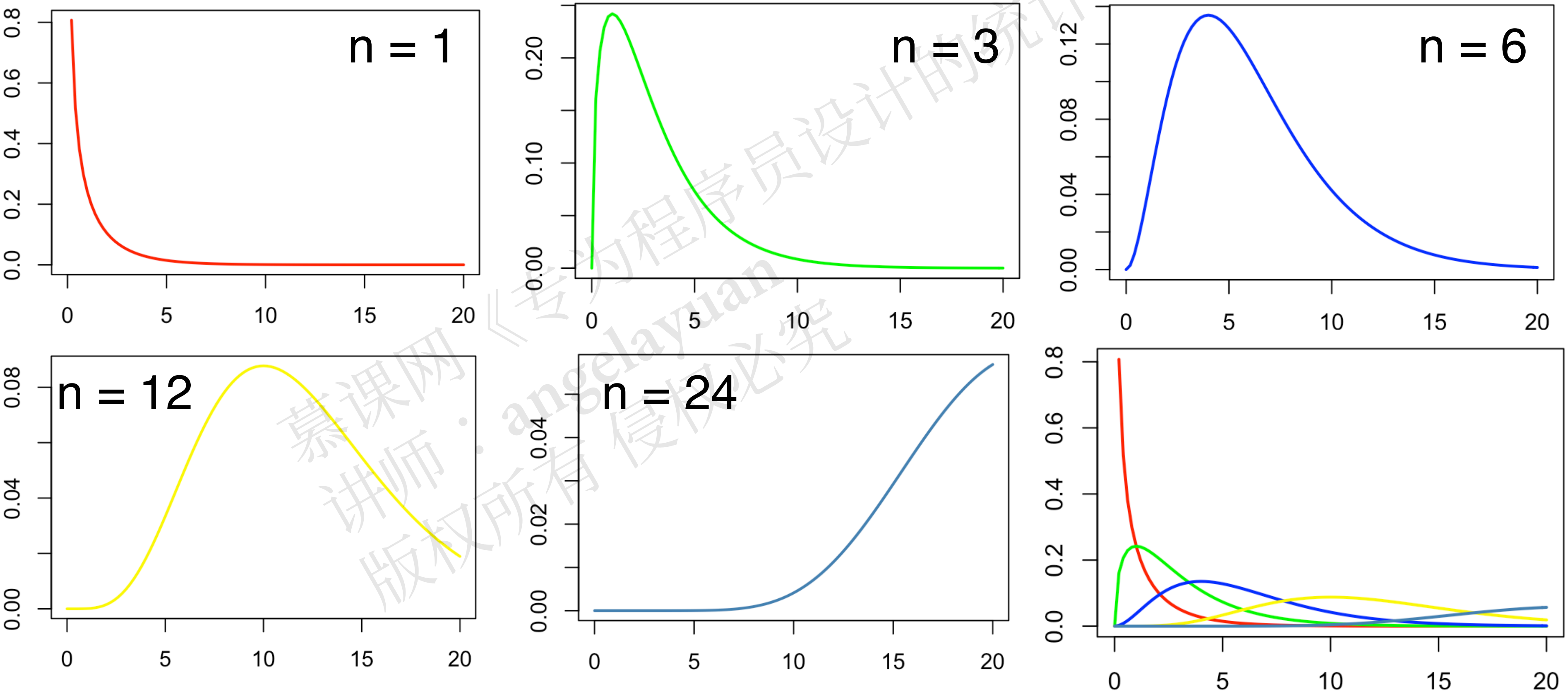
服从自由度为n的卡方分布

记为 $\chi^2 \sim \chi^2(\mathrm{n})$

自由度(df: degree of freedom)

= 独立变量的个数

卡方分布的概率密度函数图



卡方分布(chi-square distribution)

卡方分布的性质

- 可加性 $\chi_1{}^2 \sim \chi^2(n_1), \chi_2{}^2 \sim \chi^2(n_2)$, 并且$\chi_1{}^2, \chi_2{}^2$相互独立

  则有 $\chi_1{}^2 + \chi_2{}^2 \sim \chi^2(n_1 + n_2)$

- 期望(均值)和方差

  $$E(\chi^2) = n, Var(\chi^2) = 2n$$

t分布(t distribution)

总体 → $X \sim N(0,1)$, $Y \sim \chi^2(\mathrm{n})$, X,Y相互独立 → 统计量

N(0,1)

标准正态分布

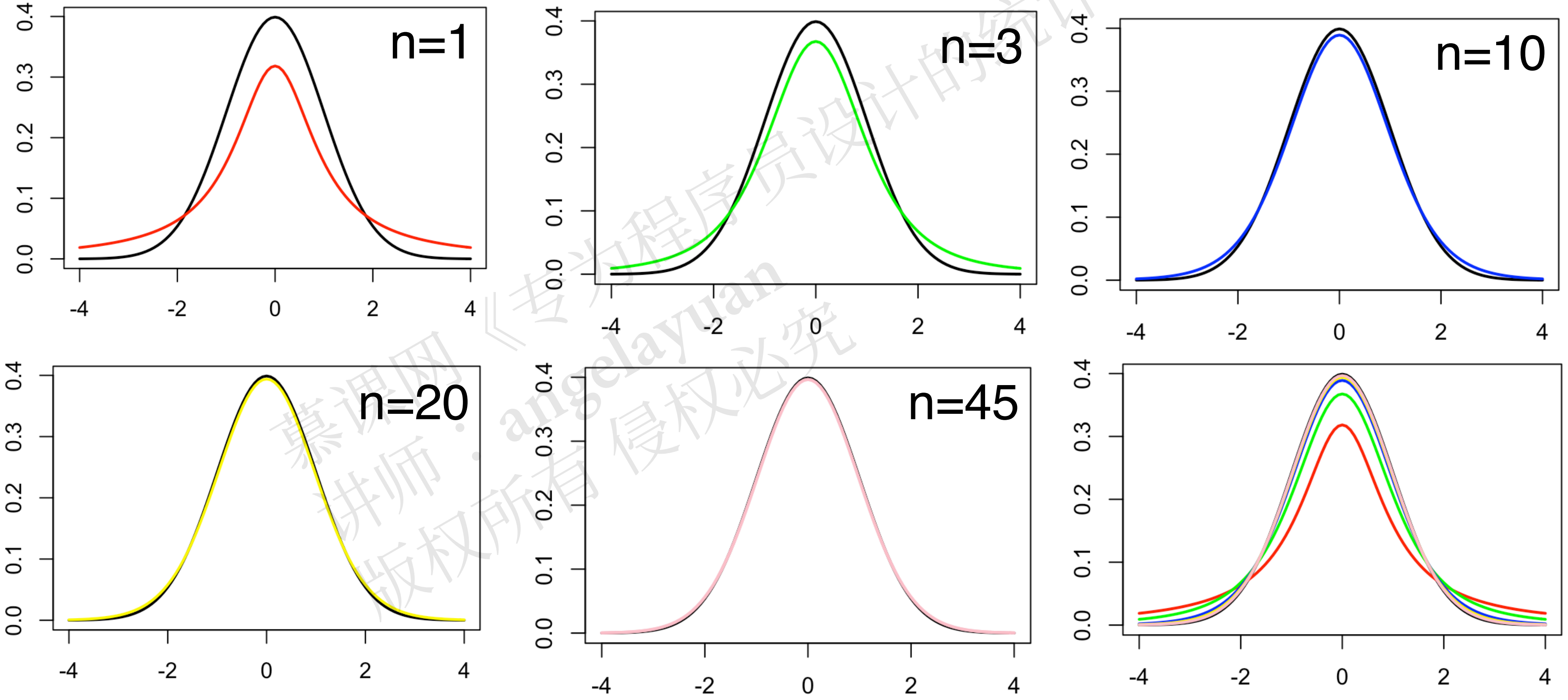
$$t = X/\sqrt{Y/n}$$

服从自由度为n的t分布

记为$t \sim t(n)$

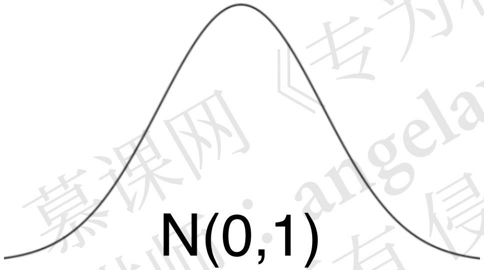
t分布又称Student分布

t分布的概率密度函数图



F分布(F distribution)

总体 → $U \sim \chi^2(n_1)$ $V \sim \chi^2(n_2)$ U,V相互独立 → 统计量



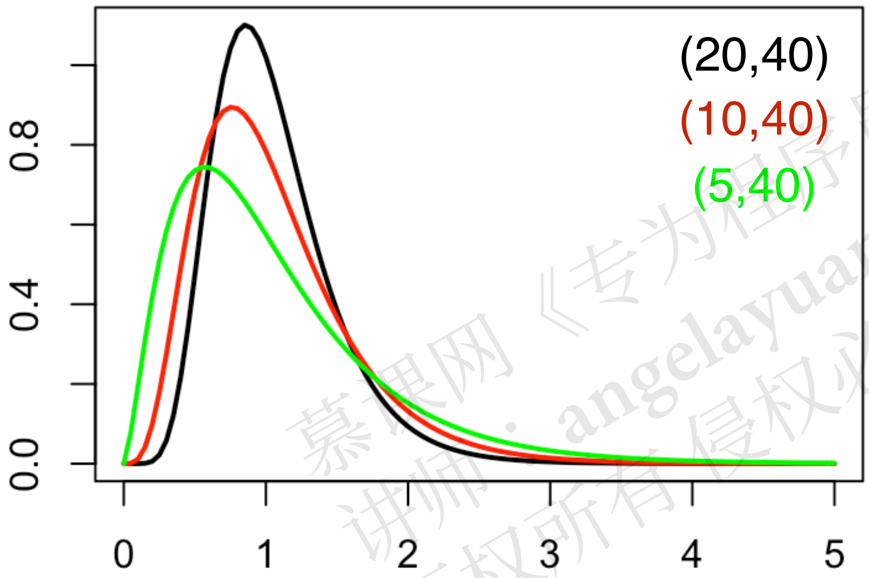
N(0,1)

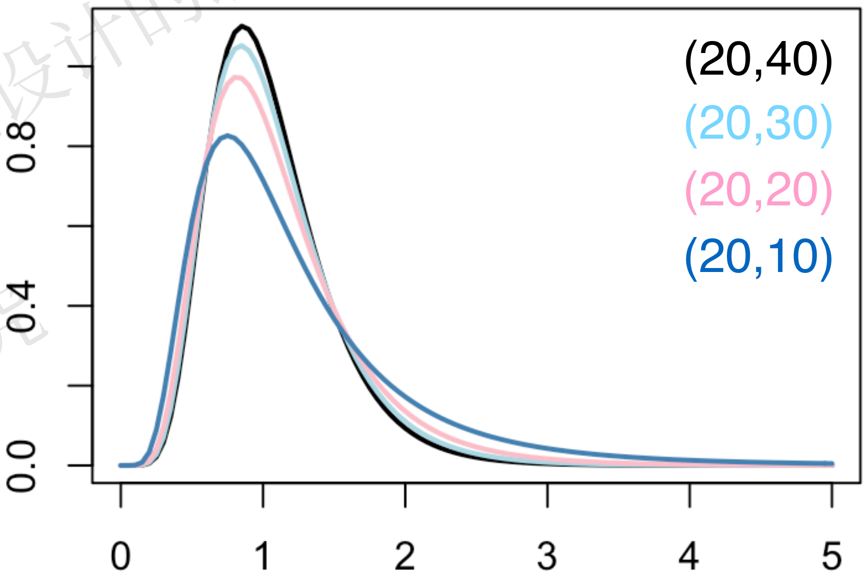
标准正态分布

$$F = \frac{U/n_1}{V/n_2}$$

服从自由度为(n_1, n_2)的F分布，记为$F \sim F(n_1, n_2)$

F分布的概率密度函数图





常用统计量

- 样本均值
- 样本方差

常用统计量的分布

- 卡方～卡方分布
- t～t分布
- F～F分布

正态总体的样本均值和样本方差的分布

设总体X的均值为μ, 方差为 σ^2

$X_1, X_2, \dots X_n$ 是来自总体X的一个样本

$\bar{X}, S^2$ 分别为样本均值和样本方差

则有 $E(\bar{X}) = \mu, \mathrm{Var}(\bar{X}) = \sigma^2/n$

$E(S^2) = \sigma^2$

注: 我们没有限定总体X的分布

证明

$$E(\bar{X}) = E(\frac{X_1 + X_2 + \cdots + X_n}{n})$$

$$= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n)$$

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n}.n\mu = \mu$$

证明

$$Var(\bar{X}) = Var(\frac{X_1 + X_2 + \cdots + X_n}{n})$$

$$= Var(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n)$$

$$= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \cdots + \frac{1}{n^2}Var(X_n)$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \sigma^2/n$$

定理

- **定理一**

  $X_1, X_2, \ldots X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本

  $\bar{X}$ 是样本均值，则有 $\bar{X} \sim N(\mu, \sigma^2/n)$

中心极限定理 (Central Limit Theory)

中心极限定理(CTL)

- 中心极限定理是概率论中的一组定理
- 中心极限定理说明，在适当的条件下，相互独立的随机变量之和经适当标准化后，其分布近似于正态分布；注意，不要求变量本身服从正态分布

样本均值的分布

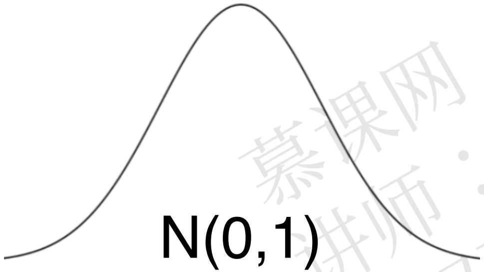
- 样本来自正态总体

总体 → 1000个样本

样本容量= 40

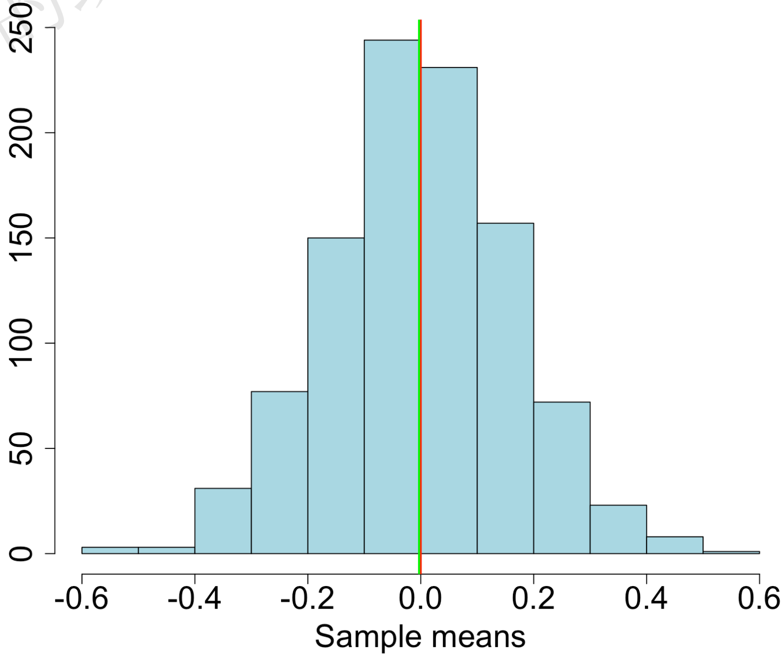
↓

计算1000个样本的样本均值 →



N(0,1)

标准正态分布

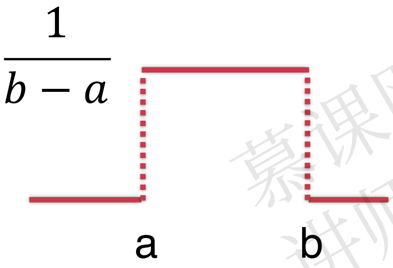


样本均值的分布

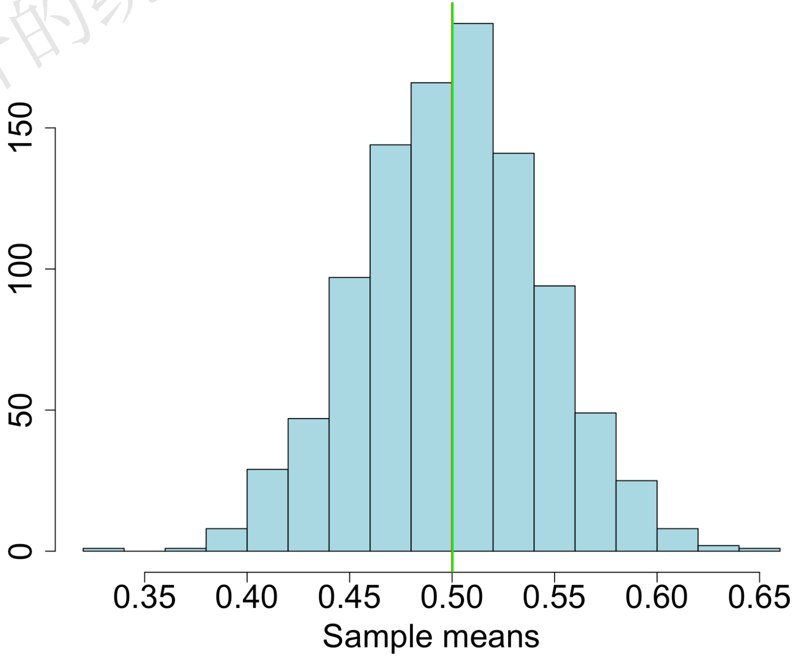
- 样本来自非正态总体

1000个样本
样本容量= 40

计算1000个样本的样本均值



U(0,1) 均匀分布



定理

- **定理二**

  $X_1, X_2, \dots X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本

  $\bar{X}, S^2$ 分别为样本均值和样本方差

  则有 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ $\qquad$ $\bar{X}, S^2$ 相互独立

定理

- **定理三**

  $X_1, X_2, \ldots X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本

  $\bar{X}, S^2$ 分别为样本均值和样本方差

  则有 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

证明

- **定理三**

根据定理一 $\bar{X} \sim N(\mu, \sigma^2/n) \longrightarrow \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

根据定理二 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

$t = X/\sqrt{Y/n}$

$X \sim N(0,1)$

$Y \sim \chi^2(\mathrm{n})$

X,Y相互独立

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}$$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t\,(n-1)$$

定理

- **定理四 (两个正态总体的样本均值和样本方差)**

  $X_1, X_2, \dots X_{n_1}$是来自正态总体$N(\mu_1, {\sigma_1}^2)$的一个样本

  $\bar{X}, {S_1}^2$分别为样本均值和样本方差

  $Y_1, Y_2, \dots Y_{n_2}$是来自正态总体$N(\mu_2, {\sigma_2}^2)$的一个样本

  $\bar{Y}, {S_2}^2$分别为样本均值和样本方差

定理

- **定理四 (两个正态总体的样本均值和样本方差)**

则有 $\dfrac{{S_1}^2/{S_2}^2}{{\sigma_1}^2/{\sigma_2}^2} \sim F\,(n_1-1, n_2-1)$

当${\sigma_1}^2={\sigma_2}^2=\sigma^2$时, $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2)$

$$S_w=\sqrt{\frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}}$$

编程理解中心极限定理

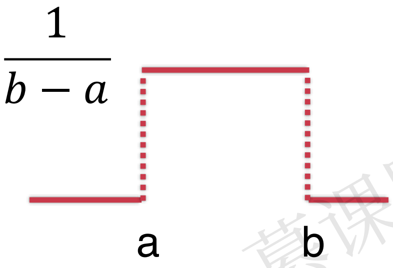
在适当的条件下，相互独立的随机变量之和经适当标准化后，其分布近似于正态分布；不要求变量本身服从正态分布

总体

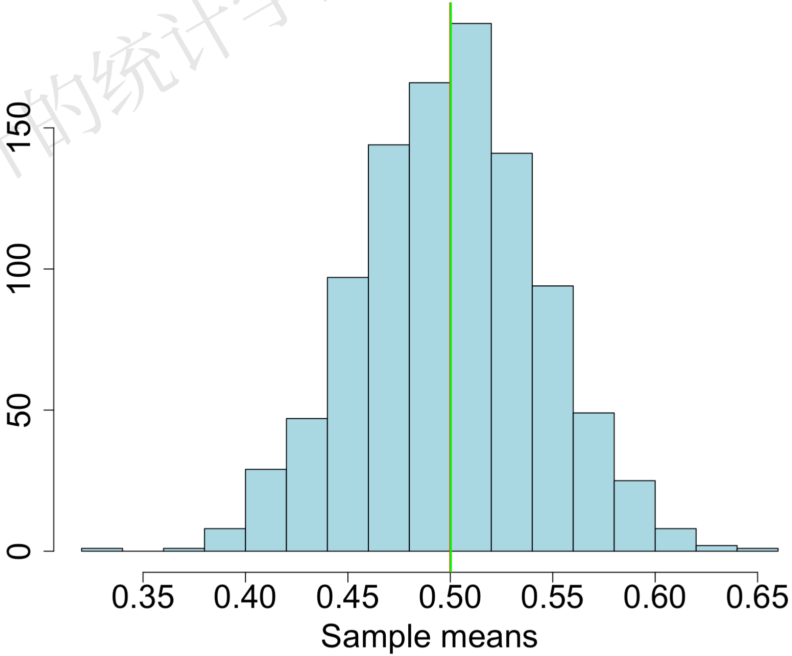
1000个样本

样本容量= 40

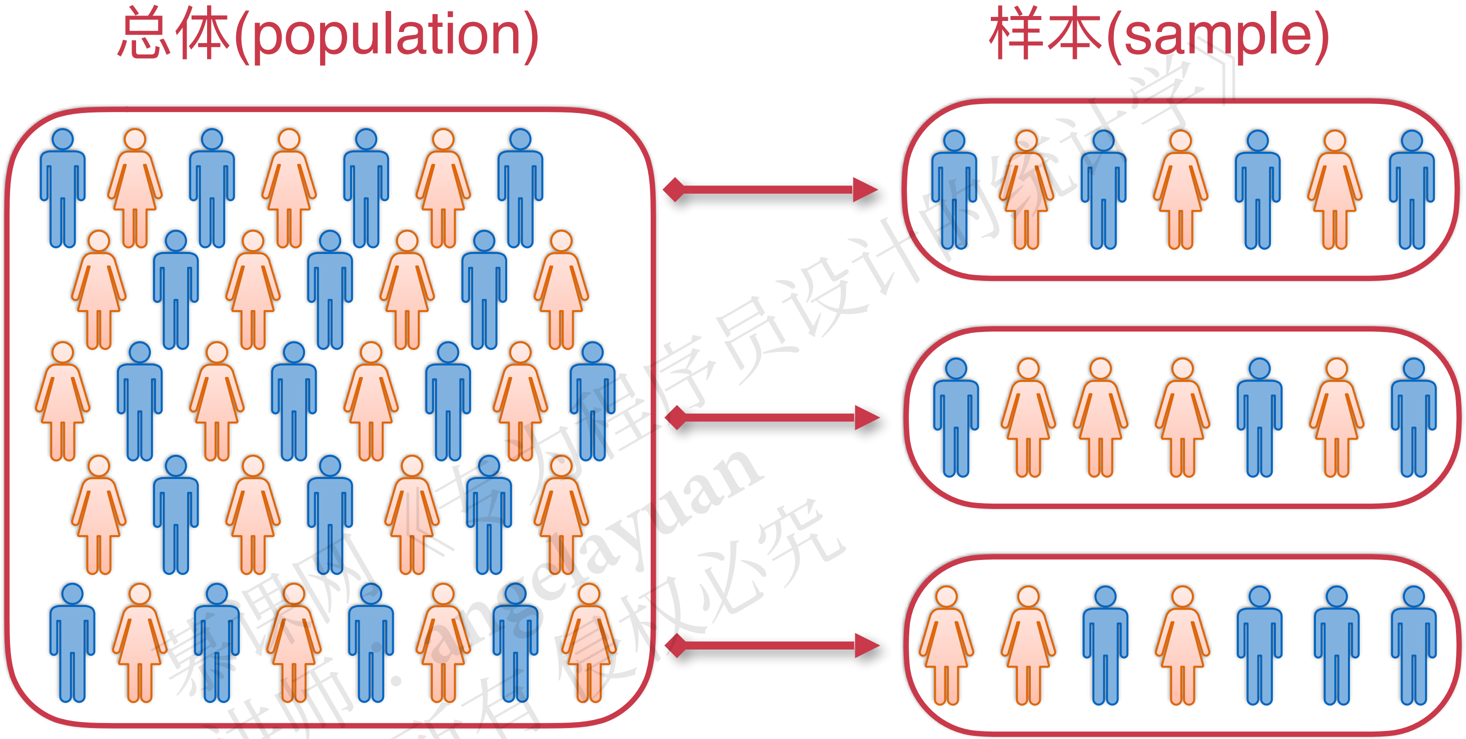
计算1000个样本的样本均值



U(0,1) 均匀分布

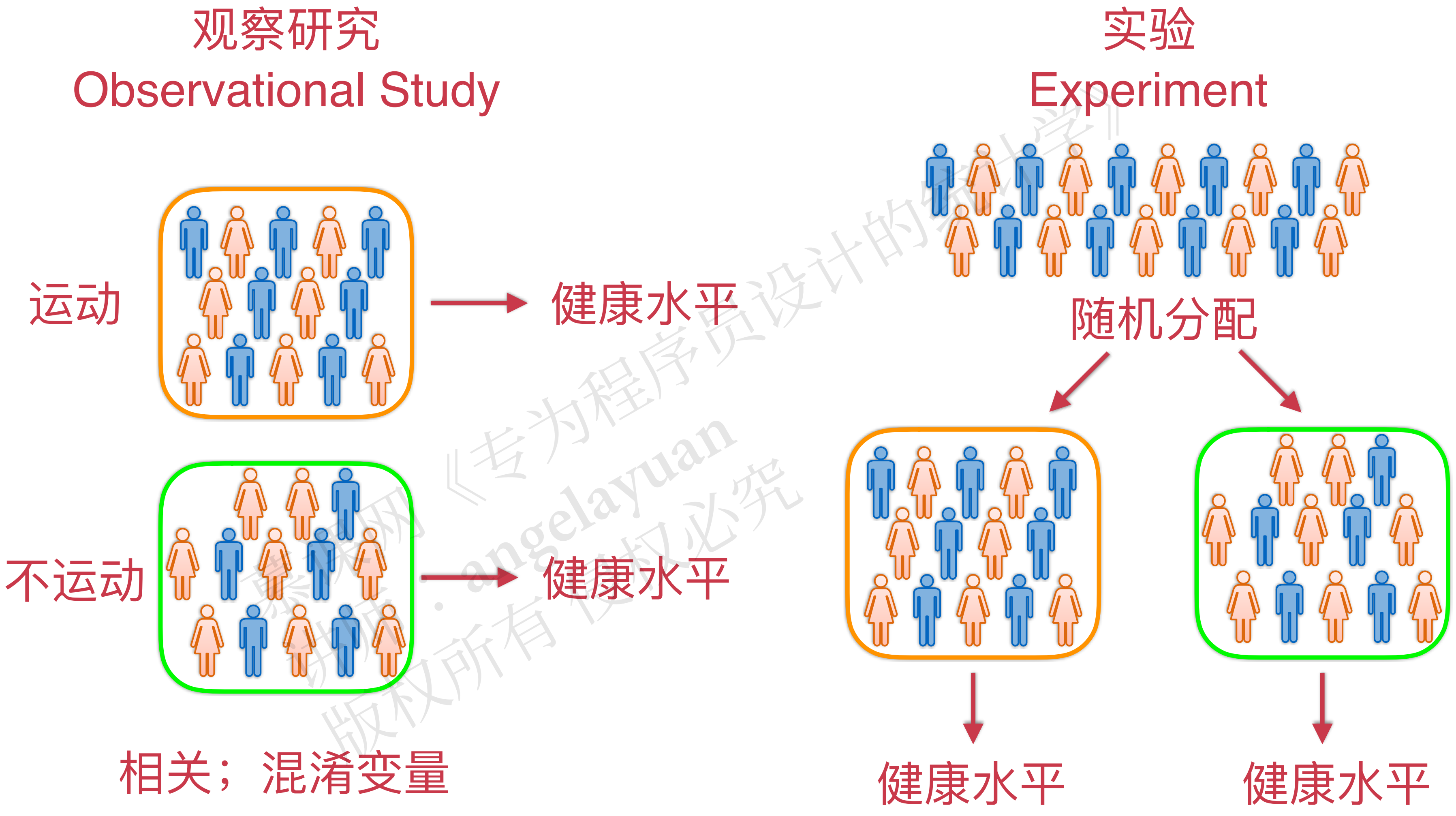


随机抽样，误差源，随机分配


总体(population)
样本(sample)

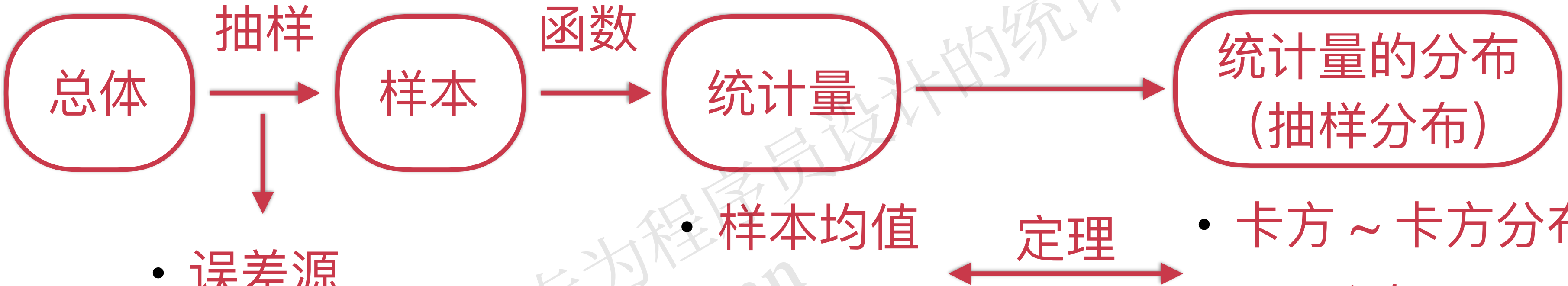
误差源

- 方便性: 容易联系到的人更可能包含进来
- 不回复: 随机抽取的样本中只有一部分人回答了问卷
- 自愿性: 样本由自愿参与的人组成


观察研究
Observational Study
运动
健康水平
不运动
健康水平
相关；混淆变量
实验
Experiment
随机分配
健康水平
健康水平

	随机分配	非随机分配
随机抽样	因果; 可泛化	非因果; 可泛化
非随机抽样	因果; 不可泛化	非因果; 不可泛化

本章小结



- 误差源
- 随机抽样 vs 随机分配
  - 是否随机抽样：是否可泛化
  - 是否随机分配：是否因果

- 样本均值
- 样本方差

- 卡方 ~ 卡方分布
- t ~ t分布
- F ~ F分布
- Z ~ 正态分布