术语

- population
  - Population parameter
  - mean μ
  - sd σ
- sample
  - Sample size
  - Sampling Methods
  - sample mean xbar
    - Mean and variance of sample mean
  - Sampling Distributions 抽样分布
- statistic

  Let $X_1, X_2, \ldots, X_n$ be a random sample of size n whose distribution may or may not depend on an unknown parameter θ. Then **the function $T = T(X_1, X_2, \ldots, X_n)$ that does not depend on θ is a statistic.**

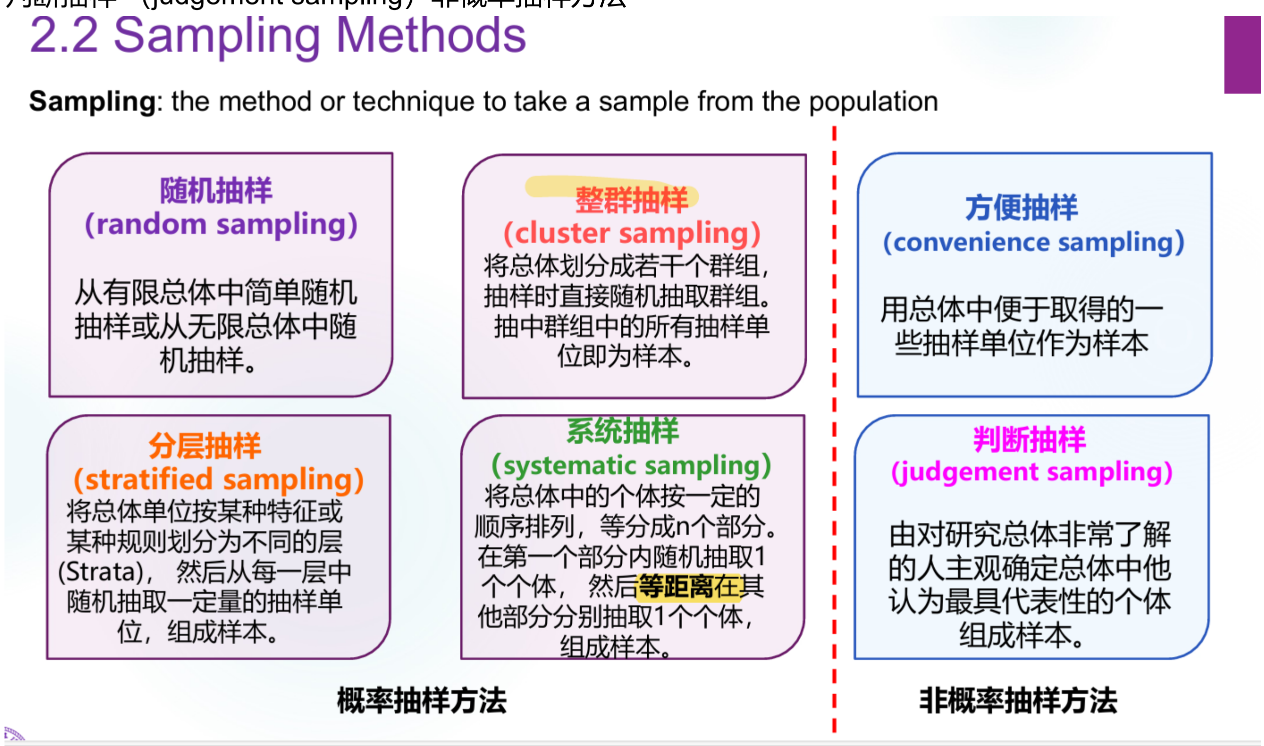
  Examples of statistics $T = T(X_1, X_2, \ldots, X_n)$:

  - $T = \bar{X} = \frac{1}{n}(X_1 + \ldots . + X_n)$ sample mean
  - $T = S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ sample variance
  - $T = S = sqrt(\frac{1}{n-1}\sum(X_i - \bar{X})^2)$ sample standard deviation
  - $T = \mathrm{M_n}$ sample median
  - $T \equiv 7$
  - Order statistics: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$

Sampling Methods

- 随机抽样 （random sampling）
- 分层抽样 （stratified sampling）
- 整群抽样 （cluster sampling）
- 系统抽样 （ systematic sampling
- 方便抽样（convenience sampling 非概率抽样方法

- 判断抽样 （judgement sampling）非概率抽样方法



Random Sample

- independent
- Representative

Mean and variance of sample mean

Let $X_1, X_2, \dots, X_n$ be a random sample of size n from a distribution (population) with mean μ and variance σ^2.

$$E(\bar{X}) = E\left(\frac{X_1 + \ \dots . + X_n}{n}\right) =$$

样本代表整体

$$Var(\bar{X}) = Var\left(\frac{X_1 + \dots + X_n}{n}\right) =$$

Sampling Distributions 抽样分布

For example, $T_n = \bar{X}$. We randomly select a sample of n observations from the population and compute the mean of this sample; call the sample mean $\bar{x}_1$.

If we were to continue this procedure ….

Sampling Distribution of xbar

- When sampling from a normally distributed population

- CLT
- xbar~normal dist

CLT

Given that the distribution of a continuous variable in the underlying population has mean μ and standard deviation σ, the distribution of sample means computed for **samples of size n** has three important properties:

- 1. The mean of the sampling distribution is identical to the population mean μ.
- 2. The standard deviation of the distribution of sample means is equal to $\sigma/\sqrt{n}$.
- 3. Provided that **n is large enough**, the shape of the sampling distribution is approximately normal.

条件

① **Independence**: Sampled observations must be independent.

② **Sample size**:

- the population distribution must be nearly normal

  or

- sample size n > 30 and the population distribution is not normal distributed.

CLT for Bernoulli variable

First, recall that a Binomial variable is just the sum of n Bernoulli variable:

$$S_n = \sum_{i=1}^{n} X_i$$

Notation:

??????????

$$S_n \sim \text{Binomial(n,p)}$$
$$X_i \sim \text{Bernoulli(p)} = \text{Binomial(1, p) for } i = 1, \ldots, n$$

In this case,

$$\hat{p} = \frac{S_n}{n} = \frac{\sum_{i=1} X_i}{n} = \bar{X}$$

$\hat{p}$ is a sample mean!

We can use the CLT when *n* is large.

3.5 Binomial CLT

For a Bernoulli variable,

- $\mu = \text{ mean } = p$
- $\sigma^2 = \text{ variance} = p(1-p)$

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \hat{p} \sim N(p, \frac{p(1-p)}{n})$

检验

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

Currently recommended
np > 15
n(1-p) > 15

校正（? ）

3.6 Improving approximation

Binomial probability:

$$P(7 \le X \le 13) = \sum_{k=7}^{13} \binom{20}{k} 0.5^x (1-0.5)^{20-k}$$

? ? ? ?

Naive approximation:

$$P(7 \le X \le 13) \approx P\left(Z \le \frac{13-10}{\sqrt{5}}\right) - P\left(Z \le \frac{7-10}{\sqrt{5}}\right)$$

Continuity corrected approximation:

$$P(7 \le X \le 13) \approx P\left(Z \le \frac{13+1/2-10}{\sqrt{5}}\right) - P\left(Z \le \frac{7-1/2-10}{\sqrt{5}}\right)$$

连续性校正近似