统 计 学 Statistics

陈 灯 塔

Econ, XJTUCC

CDT.WISE@G

2021 年 4 月 28 日

内容梗概

1. 统计推断：概念
2. 统计推断：估计
3. 统计推断：抽样

抽样推断

利用抽样资料来估计总体数量特征，从而认识总体

1. 日常生活中，我们经常碰到，您想起了哪些例子来了？

1 统计推断：概念
- 抽样调查
- 基本概念
- 抽样分布
- 总结

t-distribution

wiki: Student's t-distribution has the probability density function given by

$$f(x) = \frac{\Gamma((v+1)/2)}{(v\pi)^{1/2}\Gamma(v/2)} \left(1 + \frac{x^2}{v}\right)^{-(v+1)/2}$$

and Γ is the gamma function

$$\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} \, \mathrm{d}x$$

The t-distribution is symmetric and bell-shaped, like the normal distribution. However, the t-distribution has heavier tails, meaning that it is more prone to producing values that fall far from its mean.

Gamma function

$$\Gamma(z+1) = z\Gamma(z) \qquad \Gamma(1) = 1$$

$\Gamma(n) = (n-1)!$

Chi-square distribution

If $Z_1, Z_2, \cdots, Z_k$ are independent, standard normal random variables, then the sum of their squares

$$Q = \sum_{i=1}^{k} Z_i^2$$

is distributed according to the chi-square distribution with k degrees of freedom. This is usually denoted as $Q \sim \chi_k^2$ or $Q \sim \chi^2(k)$

The probability density function (pdf) of the chi-square distribution is

$$f(x; k) = \begin{cases} \frac{x^{k/2-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

where Γ denotes the gamma function, which has closed-form values for integer k

If $X \sim \chi^2(v)$ and $c > 0$, then $cX \sim \Gamma(\alpha = v/2, \beta = \frac{1}{2c})$. (gamma distribution)

抽样调查

抽样推断

抽样估计是按随机原则从全部研究对象中抽取部分个体 (样本) 进行观察，并根据抽样资料对总体的数量特征作出具有一定可靠程度的估计和判断。

- 由部分推断整体的一种认识方法
- 随机原则，也称为等可能原则，每个个体被抽中的可能性相等
- 抽样推断运用概率估计的方法
- 抽样误差不可避免，容许误差可以事先估算或者设定

应用

- 认识无限总体, 全面调查困难
- 破坏性或者消耗性的质量检验　　炮弹, 降落伞, 灯泡
- 节省费用, 不必要全面调查
- 评价和修正全面调查　　人口普查 + 抽样调查
- 生产过程质量控制
- 假设检验，为决策提供依据　　促销是否提高酒店客房入住率

基本概念

总体与样本

总体是根据研究目的确定的所要研究的事物的全体，是由客观存在的、具有同一性质的大量个别事物构成的集合。

- 对于特定的问题来说，总体是唯一的确定的
- 组成总体的个别事物称为个体 (总体单位)，总体所包含个体的数量称为总体容量，通常用大写的字母 N 表示
- 数学上用随机变量记录总体各单位所共同具有的属性或特征。对于特定的问题，该随机变量往往也称为总体 (狭义)

样本是按随机原则从总体中抽取出来的那部分单位组成的集合。

- 样本中所包含的单位个数称为样本容量，一般用小写的字母 n 表示
- 通常将样本容量小于 30 的样本称为小样本，而将样本容量大于 30 的样本称为大样本 [?]
- 与总体是唯一确定的不同，样本不是唯一的，从一个总体中可以抽取很多个样本，全部样本的可能数目与样本容量及随机抽样的方法有关

参数和估计量

总体参数是总体 (狭义) 看成随机变量时，描述其概率分布的参数

- 由于总体是唯一确定的，总体参数也是唯一确定的，只不过通常是未知的
- 一个总体可以有多个参数，从不同方面反映总体的综合数量特征　　总体平均数, 总体比例, 总体方差

估计量：用于估计总体参数的统计量　　如样本均值，样本比例、样本方差等

统计量 vs 估计量: sample statistic is any quantity computed from values in a sample that is used for a statistical purpose; When a statistic is used to estimate a population parameter, the statistic is called an estimator. 没有将 t 统计量称为估计量，因为我们没有使用 t 统计量来估计任何参数

在我们的设定中：参数是数，统计量/估计量是随机变量

参数和估计量

[有限总体] 均值

$$\mu = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

标准差

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{x})^2}{n-1}}$$

以及总体比率 (总体成数)

$$\pi = \frac{N_1}{N} \qquad p = \frac{n_1}{n}$$

对于无限总体、连续总体，总体参数的计算复杂一些。如正态总体，$\mu = \mathrm{E}(X)$ 或者是未知的，通过样本进行估计。

抽样方案

重复抽样 (有放回): n 次相互独立的连续试验　　同一单位有重复抽中的可能。
不重复抽样 (无放回): 一次抽 n 个；或者连续抽 n 次，结果受前面试验的影响

有顺序重复抽样，样本可能数 $M = N^n$. 有顺序不重复抽样 $\binom{N}{n} \cdot n!$

$$M = \frac{N!}{(N-n)!}$$

抽样分布

Central Limit Theorem

Suppose $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, then as n approaches infinity, the asymptotic distribution (converge in distribution) of random variables $\frac{1}{\sigma}\sqrt{n}\cdot\left(\bar{X}_n - \mu\right)$ is standard normal $\mathrm{N}(0,1)$

$$\frac{1}{\sigma}\sqrt{n}\cdot\left(\bar{X}_n - \mu\right) \overset{a}{\sim} \mathrm{N}(0,1)$$

The strong **law of large numbers** (LLN) states that the sample average converges almost surely to the expected value, $\bar{X}_n \to \mu$ a.s. We see that the CLT explores the distribution of the standardized version of $\bar{X}_n$ (demeaned and scaled).

注意：随机变量的极限，有很多种定义 (wiki: Convergence of random variables)

p141 Eq (6-4) $\bar{X}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) \implies \mathrm{N}(\mu, 0)$

抽样分布

样本均值：重复抽样 p136 例子 6-1 (表 6-3, 图 6-2)

$$\mathrm{E}(\bar{x}) = \mu \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

不重复抽样下修正因子 $\frac{N-n}{N-1} \approx 1 - \frac{n}{N} \approx 1$ $(N-1 \approx N)$

$$\mathrm{E}(\bar{x}) = \mu \qquad \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

数值例子验证见 p138 例 6-2 (表 6-4)

样本比例：重复抽样

$$\mathrm{E}(p) = \pi \qquad \sigma_p^2 = \frac{\pi(1-\pi)}{n}$$

不重复抽样下修正因子 $\frac{N-n}{N-1}$

汇总：p140 表 6-5

不重复抽样

此时 X_i 同分布但不独立：X_i 取 x_j 的概率为：$\mathrm{P}(X_1 = x_j) = \frac{1}{N}$. 当 $i \geqslant 2$ 时，前 $i-1$ 次不能取 x_j, 且第 i 次取 x_j

$$\mathrm{P}(X_2 = x_j) = \frac{N-1}{N} \cdot \frac{1}{N-1} = \frac{1}{N}$$
$$\mathrm{P}(X_3 = x_j) = \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{1}{N-2} = \frac{1}{N}$$
$$\vdots$$

通式

$$\mathrm{P}(X_i = x_j) = \frac{N-1}{N}\frac{N-2}{N-1} \cdots \frac{N-(i-2)-1}{N-(i-2)} \cdot \frac{1}{N-(i-1)} = \frac{1}{N} \qquad i \geqslant 2$$

同分布但不独立，因为

$$\mathrm{P}(X_2 = x_3) \neq \mathrm{P}(X_2 = x_3 \,|\, X_1 = x_3) = 0$$

抽样分布

p137 最后两行有问题：如果总体不是正态分布，当 $n < 30$ 时，样本均值的分布不是正态分布，应该用 t 分布来推断 [χ^2 总体的任意大小的样本均值都为 Γ 分布]

- CLT 只是说渐渐分布为正态，没有说趋近的过程一定是 t 分布过程
  - The binomial distribution $\mathrm{B}(n, p)$ is approximately normal with mean np and variance $np(1-p)$ for large n and p not too close to 0 or 1
  - The Poisson distribution with parameter λ is approximately normal with mean λ and variance λ, for large values of λ
  - The chi-squared distribution $\chi^2(k)$ is approximately normal with mean k and variance $2k$, for large k [The sample mean of n i.i.d. $\chi^2(k)$ is distributed according to $\Gamma(\alpha = nk/2, \beta = n/2)$]
  - The Student's t-distribution $t(v)$ is approximately normal with mean 0 and variance 1 when v is large

p139 的倒数第 10 行: 不重复抽样分布近似为正态分布。这里还有一重问题，X_i 是同分布但不独立，适用的不是常用的 iid 版本的 CLT (CLT 有很多个，本科生往往只接触到 iid 的)

30 个的大样本

Why is 30 considered the minimum sample size in some forms of statistical analysis?

- Mostly its what could reasonably fit on a page for a common test, see 'table of t distribution'
- Actually, the "magic number" 30 is a fallacy. See Cohen (1990): *Things I Have Learned (So Far)* some things you learn aren't so
- The choice of $n = 30$ for a boundary between small and large samples is a rule of thumb, only. (似是而非的经验传承)
  1. t-distribution becomes a close fit for the Normal distribution when the number of samples reaches 30. using 30 samples is better than 20 samples but not as good as 40 samples. Clearly, there is no one magic number of samples that you should use based on this argument.
  2. the Law of Large Numbers, which in essence says that the more samples you use the closer your estimates will be to the true population values. Even with more than 70 or 80 samples, the spread of the estimates continues to decrease. So again, there's nothing extraordinary about using 30 samples.

总结

1. 抽样估计的概念 (掌握), 特点和应用 (理解)
2. 参数和估计量 (掌握)
3. 重复抽样与不重复抽样 (掌握)
4. 抽样分布 (了解)

Keyword: 抽样估计, 总体 (广义, 狭义), 样本, 参数, 统计量/估计量, 重复抽样, 不重复抽样, 抽样分布

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- 总体为编号 1–4 的四个球，随机不重复抽取 2 个 (考虑顺序). (1) 计算总体 (球编号) 的期望和方差; (2) 参考 p138 例 6-2 的方法, 将可能样本进行列表，然后用期望和方差的定义公式计算 $\bar{x}$ 的期望和方差, 并对比抽样分布公式计算的结果
- Read
  1. Textbook: S6.5
  2. Reference book: corresponding chapters

2 统计推断：估计

- 估计量
- 区间估计
- 样本容量
- 总结

参数估计

用第一个样本的观测值，或者样本的均值去估计总体的均值，哪个估计量的效果更好？为什么？

估计量

评价估计量

点估计：样本估计量的值作为总体参数的估计值，亦称为定值估计

- 无偏性: 样本估计量的数学期望等于总体参数　Unbiased Estimator. An estimator, W of θ, is an unbiased estimator if $\mathrm{E}(W) = \theta$ for all possible values of θ.
- 有效性：Relative Efficiency. If W_1 and W_2 are two unbiased estimators of θ, W_1 is efficient relative to W_2 when $\mathrm{var}(W_1) \leqslant \mathrm{var}(W_2)$ for all θ, with strict inequality for at least one value of θ.
- 一致性 Consistency. Let W_n be an estimator of θ based on a sample $X_1, X_2, \cdots, X_n$ of size n. Then, W_n is a consistent estimator of θ if for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(|W_n - \theta| > \epsilon) = 0$$

If W_n is not consistent for θ, then we say it is inconsistent. When W_n is consistent, we also say that θ is the probability limit of W_n, written as $\mathrm{plim}(W_n) = \theta$

Convergence of random variables

重复掷一枚均匀硬币，记随机变量 X 表示正面为 1 反面为 0 (理解随机变量实际上是函数)

$$X_n(H) = 1 \qquad X_n(L) = 0$$

定义随机变量 Y 取正面 (Head) 为 0 反面 (taiL) 为 1, 显然 Y 与 X_n 同分布，即

$$\begin{cases} 0 & 1/2 \\ 1 & 1/2 \end{cases}$$

由于 $Y = 1 - X_n$, 有

$$|X_n - Y| = |2Y - 1| = 1$$

显然对于任意的 $0 < \epsilon < 1$

$$\lim_{n\to\infty} \mathrm{P}(|X_n - Y| \geqslant \epsilon) = \lim_{n\to\infty} \mathrm{P}(1 \geqslant \epsilon) = 1 \neq 0$$

X_n 依分布收敛于 Y, 但 X_n 没有依概率收敛于 Y [谈到分布，变量取值的顺序就被抹掉了]

一致性

无偏性 ppt.26; 有效性 ppt.27

一致性 (consistency) ppt.28 感觉图有点不对，一开始的中心位置可以有偏，见 wooldridge 6e p682, 或者 wiki: Consistent estimator 里的图形

可以证明：样本均值和样本比率是总体均值和比率的无偏、有效和一致估计

样本均值是最佳线性无偏估计 best linear unbiased estimator (BLUE). 这里突出的是线性的，即估计量是样本的线性函数。

区间估计

正态分布

如果总体呈正态分布 $\mathrm{N}(\mu, \sigma^2)$

$$z \equiv \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

即

$$\mu = \bar{x} - z\frac{\sigma}{\sqrt{n}}$$

取 $z_{\frac{\alpha}{2}}$ 为 $\frac{\alpha}{2}$ 上分位数, $\mathrm{P}\left(z > z_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$, $\mathrm{P}\left(z < -z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$, 那么

$$\mathrm{P}\left(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

即

$$\mathrm{P}\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

区间估计

点估计 (point estimate) 无法给出估计值接近总体参数程度的信息

区间估计 (interval estimate) 在点估计的基础上，给出总体参数估计的一个区间范围，该区间由样本统计量加减估计误差而得到, 例如

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

根据样本统计量的抽样分布能够对样本统计量与总体参数的接近程度给出一个概率度量

- 将构造置信区间的步骤重复很多次，那么置信区间包含总体参数真值的次数所占的比例称为**置信水平** $1-\alpha$
- α 为是总体参数未在区间内的比例
- 统计学家在某种程度上确信这个区间会包含真正的总体参数，所以给它取名为**置信区间**

区间估计

总体参数以一定的概率落在这一区间　✗ 表述是错误的，为什么？因为总体参数是常数，不是随机变量

用一个具体的样本所构造的区间是一个特定的区间，我们无法知道这个样本所产生的区间是否包含总体参数的真值

区间估计的真正含义：多次抽样，得到多个置信区间，有的区间包含真值，有的没有。有 $(1-\alpha)\cdot M$ 个区间包含真值

做法

区间估计　　E6-9

1. 选定 α 求 Z 值 $z_{\frac{\alpha}{2}}$ `=-NORM.S.INV(alpha/2)`
2. 写出区间 $\left[\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$

α 越小，置信水平为 $1-\alpha$ 越大，可靠性越高；而 α 越小 Z 值 $z_{\frac{\alpha}{2}}$ 越大，区间越宽，精确性越差

- 容许误差 (最大误差) $\Delta_{\bar{x}} = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = z_{\frac{\alpha}{2}}\sigma_{\bar{x}}$
- 如果总体方差 σ^2 未知，用样本方差 s^2 代替 [p143 $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$]
- 不重复抽样下，增加修正因子 $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$

总体成数

简单随机抽样条件下，样本容量很大，或者 $np \geqslant 5$ 且 $(1-p)n \geqslant 5$ 时

$$p \sim \mathrm{N}\left(\pi, \frac{(1-\pi)\pi}{n}\right)$$

标准化后

$$Z = \frac{p-\pi}{\sqrt{\frac{(1-\pi)\pi}{n}}} \sim \mathrm{N}(0,1)$$

置信区间为

$$\left[\bar{x} - z_{\frac{\alpha}{2}}\sqrt{\frac{(1-\pi)\pi}{n}}, \bar{x} + z_{\frac{\alpha}{2}}\sqrt{\frac{(1-\pi)\pi}{n}}\right]$$

π 未知时，用 p 替代。容许误差 $\Delta_p = z_{\frac{\alpha}{2}}\sqrt{\frac{(1-\pi)\pi}{n}}$.

不重复抽样下，增加修正因子 $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$

样本容量

确定样本容量

均值参数估计中的样本容量确定：重复抽样下 $\Delta_{\bar{x}} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$$n = \left(z_{\frac{\alpha}{2}} \frac{\sigma}{\Delta_{\bar{x}}} \right)^2$$

显然容许误差减半，样本容量增加到四倍。不重复抽样下 $\Delta_{\bar{x}} \approx z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$

$$n = \frac{N z_{\frac{\alpha}{2}}^2 \sigma^2}{z_{\frac{\alpha}{2}}^2 \sigma^2 + N\Delta_{\bar{x}}^2} = \frac{z_{\frac{\alpha}{2}}^2 \sigma^2}{\Delta_{\bar{x}}^2 + z_{\frac{\alpha}{2}}^2 \frac{\sigma^2}{N}}$$

总体成数 $\Delta_p = z_{\frac{\alpha}{2}} \sqrt{\frac{(1-\pi)\pi}{n}}$，求出 n 得

$$n = z_{\frac{\alpha}{2}}^2 \frac{(1-\pi)\pi}{\Delta_p^2}$$

不重复抽样下 $\Delta_p \approx z_{\frac{\alpha}{2}} \sqrt{\frac{(1-\pi)\pi}{n}} \sqrt{1 - \frac{n}{N}}$

$$n = \frac{\pi N z_{\frac{\alpha}{2}}^2 - \pi^2 N z_{\frac{\alpha}{2}}^2}{\pi z_{\frac{\alpha}{2}}^2 + N\Delta_p^2 - \pi^2 z_{\frac{\alpha}{2}}^2} = \frac{z_{\frac{\alpha}{2}}^2 (1-\pi)\pi}{\Delta_p^2 + z_{\frac{\alpha}{2}}^2 \frac{(1-\pi)\pi}{N}}$$

总结

1. 点估计与区间估计的区别 (理解)
2. 评价估计量优良性的标准 (了解)
3. 总体参数的区间估计方法 (掌握)
4. 样本容量的确定方法 (掌握)

Keyword: 点估计, 无偏性, 有效性, 一致性, 上分位数, Z 值, 区间估计, 置信水平, 置信区间, 容许误差 (最大误差), 样本容量确定

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- SampleVBA.xlsm 的 Transmission 表单 (个人作业). 其中表单 Sheet3 是品牌知名度例子 (抽样 350 人中 112 人听说过该品牌), 请学习下 (不需要交)
- Read
  1. Textbook: S6.3–4
  2. Reference book: corresponding chapters

3 统计推断：抽样

- 基本原则
- 抽样方法设计
- 总结

抽样

我们将介绍如下概率抽样方法

- 简单随机抽样 (多次抽样，每次 n 个样本, see [SampleVBA.xlsm]Sheet1)
- 类型抽样
- 等距抽样
- 整群抽样
- 阶段抽样

问题：为什么会有这么多抽样方法？

回答：经济现象是复杂的，不是单一的。没有万能的、终极的、随时随地都最优的方法

这也是核心价值观里面为什么要有民主：她是与复杂经济和社会结构相联系的。历史上的专制，都与社会经济的单一为基础，如农业为主。进入工业社会、乃至信息社会，分工更细致，利益多元化共存，东酸西辣南甜北咸，民主共荣是当前人类找到的最好方式。

Random Number Generator

Random number generators can be

- truly random hardware random-number generators (HRNGS), which generate random numbers as a function of current value of some physical environment attribute that is constantly changing in a manner that is practically impossible to model
- or pseudorandom number generators (PRNGS), which generate numbers that look random, but are actually deterministic, and *can be reproduced* if the state of the PRNG is known.
  - For example, a **chaotic logistic map** process is given as follows:

$$X_t = 4X_{t-1}(1 - X_{t-1}), \text{ with } X_0 \in (0, 1)$$

    This is very similar to a so-called **white noise** random process $\{X_t\}$ in certain aspects.
  - Randomness test: in data evaluation, is a test used to analyze the distribution of a set of data to see if it can be described as random (patternless). A well-known and widely-used collection of tests was the Diehard Battery of Tests

基本原则

基本原则

抽样设计的基本原则

1. 保证抽样随机原则的实现
   要保证总体每一单位都有同等的中选机会
2. 抽样效果优化
   在一定的误差要求下选择费用最少的方案；或在一定的费用开支条件下，选择误差最小的方案

抽样误差

见 B-Stat-2021-Ch2 的 “误差”

代表性误差：子样本结构与总体结构不一致，样本不能完全代表总体
代表性误差的影响因素

- 样本量的大小
- 总体的变异大小
- 抽样的方案 (scheme), 放回、顺序
- 抽样方法 (整群抽样, 等距抽样)

抽样方法设计

简单随机抽样

也称为纯随机抽样，是从总体包含的 N 个单位中任意抽取 n 个单位作为样本

- 总体中每个单位可能被抽中的概率相等
- 它是一种最基本的抽样方法
- 它是其他抽样方法的基础

适用与均匀总体: 需要包含所有个体的抽样框，费用可能高 (若空间分散)

- 抽签
- 随机数字
- 伪随机数 (软件)

误差：有放回 ($\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$), 无放回 ($\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$, Ch6 T1 作业 Y104.xlsx)

p146 E6-4 灯泡重复抽样不能直接套用无放回公式，为什么?

类型抽样

亦称为分类抽样或者**分层抽样** (stratified sampling)

1. 将总体按某种特征或原则划分成若干层 (类型、组别)
2. 在每层内独立地、随机地抽取子样本
3. 将子样本合起来构成总体样本

适合于有辅助信息，特别是层之间差异大时

- 层是总体的内部结构信息; 地域或者行政分层
- 等比分层与非等比分层
- 提高抽样精度，对层全面调查，层内非全面调查，尽量缩小层内差异，扩大层间差异

误差：重复抽样 ($\sigma_{\bar{x}} = \sqrt{\frac{\overline{s^2}}{n}}$, $\overline{s^2} = \frac{1}{n}\sum s_i^2 n_i$), 不重复抽样修正因子

等距抽样

课本称为系统抽样

1. 将总体中的所有单位按某一标志排序
2. 在规定的范围内抽取一个单位作为初始单元
3. 按事先定好的间隔 K 确定其他样本单位

误差

- 有关标志排队: 农业产量抽样按亩产 $\rightarrow$ 分层抽样
- 无关标志排队: 收入按门牌号; 成绩按姓氏 $\rightarrow$ 简单随机抽样

整群抽样

总体分为若干群，随机抽取一部分群，对被抽中的群内做全面调查

适用于群内差异大群间差异小 (与分层抽样相反)

优缺点：群内集中，便于调查；代表性可能较差。

多阶段抽样

在抽样时先随机抽总体中某种更大范围的单位，再从中选大单位中随机抽较小范围的单位，逐次类推，最后从更小范围单位中随机抽选样本的基本单位，分阶段来完成抽样的组织工作。

- 当总体很大时，常采用多阶段抽样方法
  - 多级抽样，先抽一级单位，再从一级单位中抽取二级单位，如此继续，直到抽取个体
  - 农业产量调查: 省县乡到农户　　每个阶段可以选择适宜的抽样方法
- 两阶段抽样在组织技术上可以看为是整群抽样第一阶段和类型抽样第二阶段的结合。

非概率抽样

非概率抽样不能用于参数估计, 用于探索性和预备性研究

- 方便抽样：样本的确定主要是基于简便。方便抽样具有相对易于样本选择和搜集数据的优点。
- 判断抽样：由对所研究总体非常了解的人选择总体中他认为最具总体代表性的元素。

总结

1. 抽样组织设计 (理解)

Keyword: 随机原则, 抽样效果, 抽样误差, 简单随机抽样, 类型抽样, 等距抽样, 整群抽样, 多阶段抽样, 方便抽样, 判断抽样

Homework: ($*$ by group)

- Time loss of smart phones, pads and computers in classroom
- 基本概念要牢记，牢记，牢记!
- $*$ Salary-stu.xlsx, 电脑里完成其中 sheet1 里的内容 (纸版或者电子版都不需要提交)
- Read
  1. Textbook: S6.6.1–3
  2. Reference book: corresponding chapters