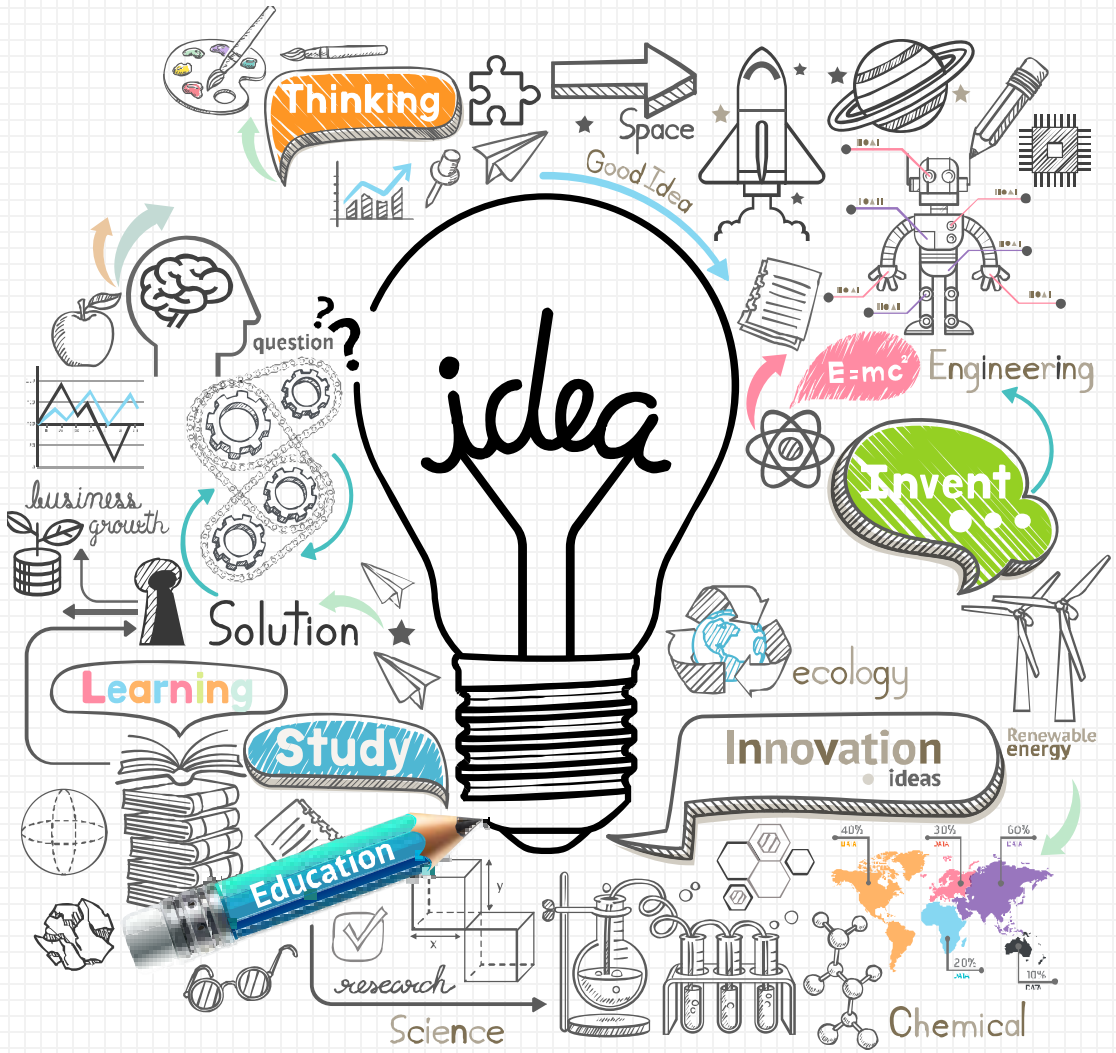
“正大杯”第十二届市调大赛公益培训课

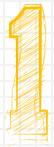
概率调查样本的统计推断

李莉莉
青岛大学经济学院



CONTENTS

一 概述



基于设计统计推断

基于模型统计推断

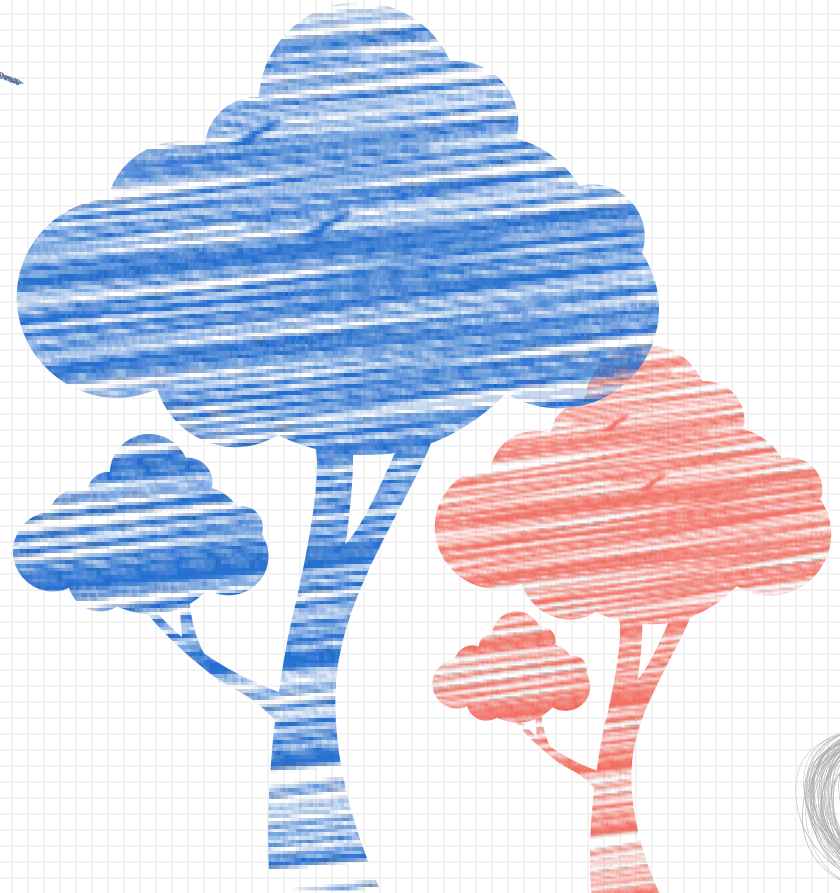
模型辅助推断

总体推断目标

(1) 均值： $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ ；

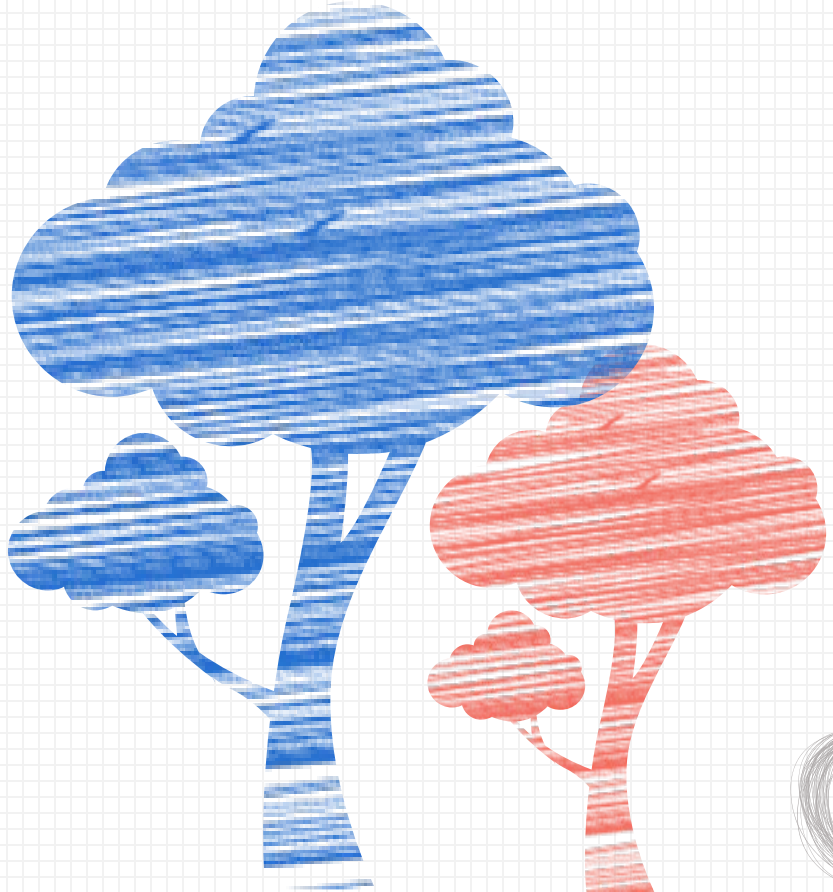
(2) 总和： $Y = N\bar{Y}$ ；

(3) 比例： $P = \frac{N_A}{N}$



1 基于设计的统计推断

- 传统的抽样推断方法
- 通常假定总体是固定和有限的，样本是根据某种抽样设计从总体中随机抽取，结合入样概率对总体进行推断。



2 基于模型的统计推断

基本思想

假设所研究的总体是从无限超总体中抽取的一个样本量为N的样本。

3 模型辅助推断

基本思想

基于设计的统计推断和基于模型的统计推断相结合

二、基于设计统计推断

1 概率抽样

2 非概率抽样

1 非概率抽样
(non-probability sampling)

用一种主观的方法从总体中抽选单元。

- 随意抽样
- 志愿者抽样
- 判断抽样
- 配额抽样

非概率抽样的缺点

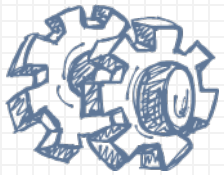
为了对总体进行推断，需要对样本的代表性做很强的假定。

怎样利用非概率抽样的样本得到可靠的估计值以及抽样误差估计值。

2 概率抽样
（probability sampling）

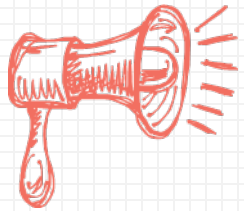
两条基本准则：

单元是随机抽取的；



调查总体中的每个单元都有一个非零的入样概率，并且能计算出这些概率。

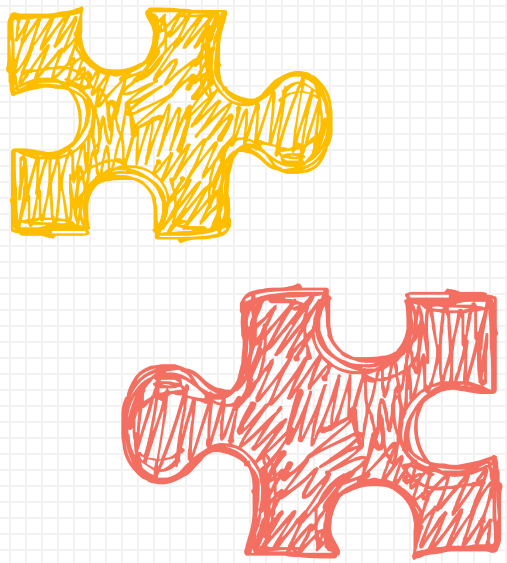
概率抽样的优点



能对总体进行推断

- 能得到总体的可靠估计值
- 能计算估计值的抽样误差

如何抽到一个“好”样本

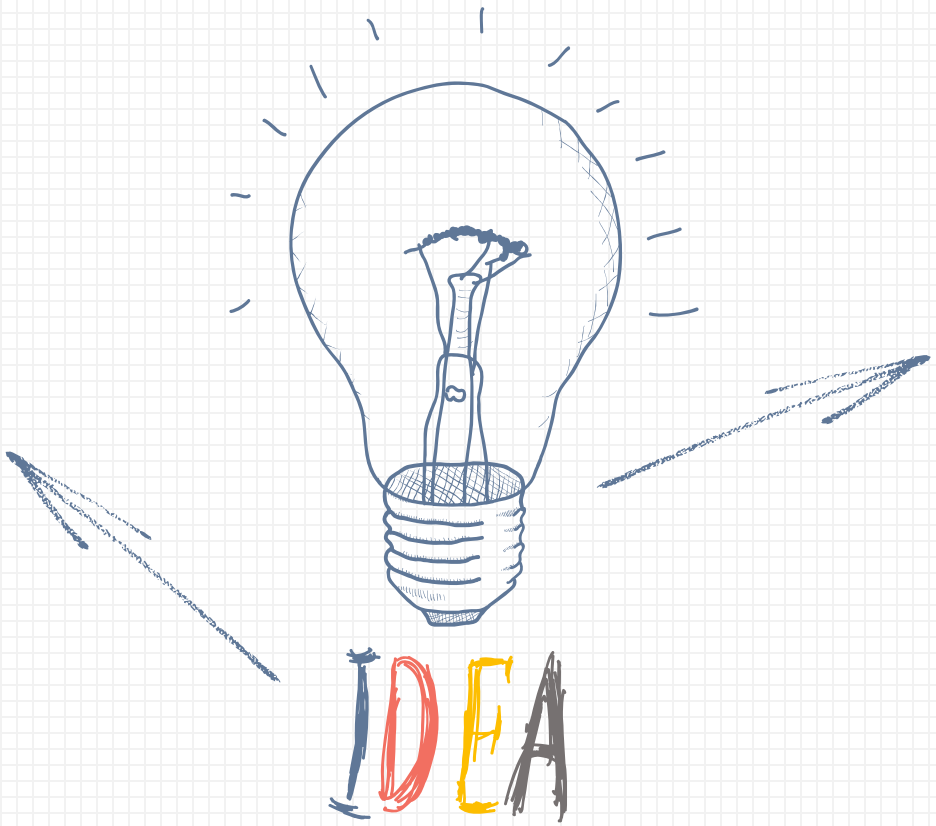


1. 用好的抽样方法
2. 用好的估计方法

好的抽样策略
辅助信息（变量）的充分利用

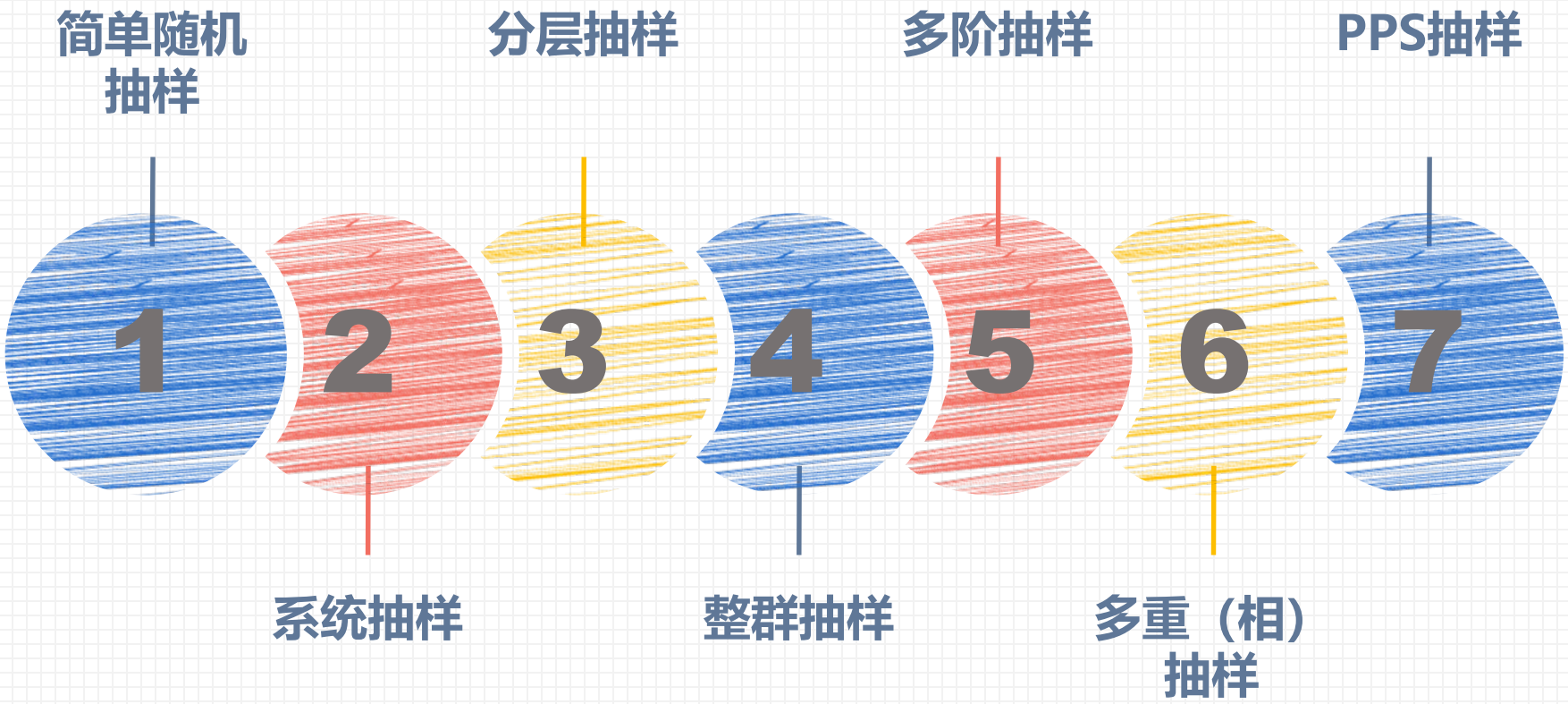
3 概率抽样方法的类型

放回抽样与
不放回抽样

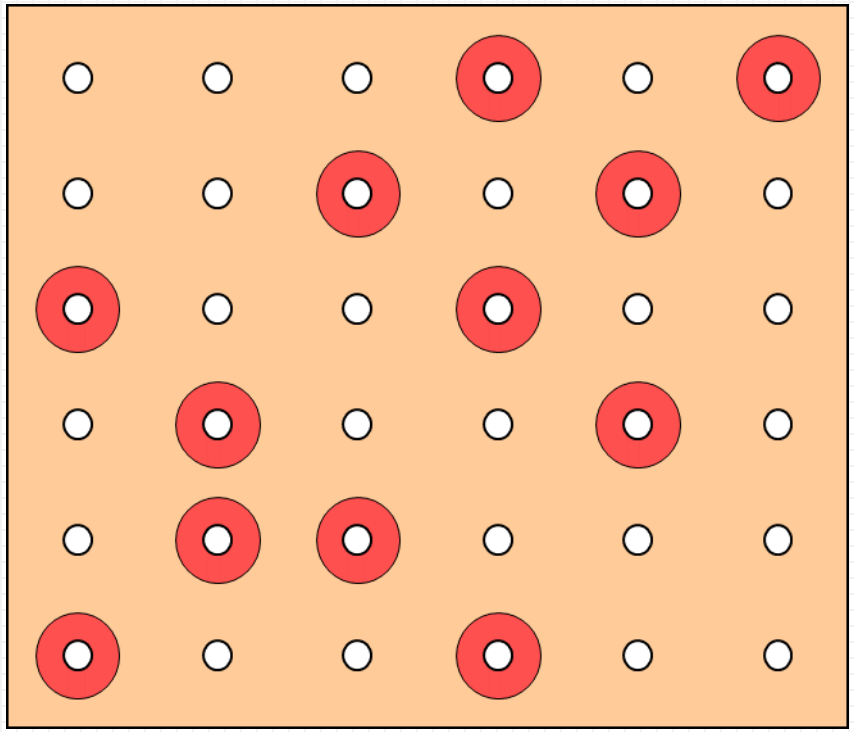


等概率抽样与
不等概率抽样

4 概率抽样方法

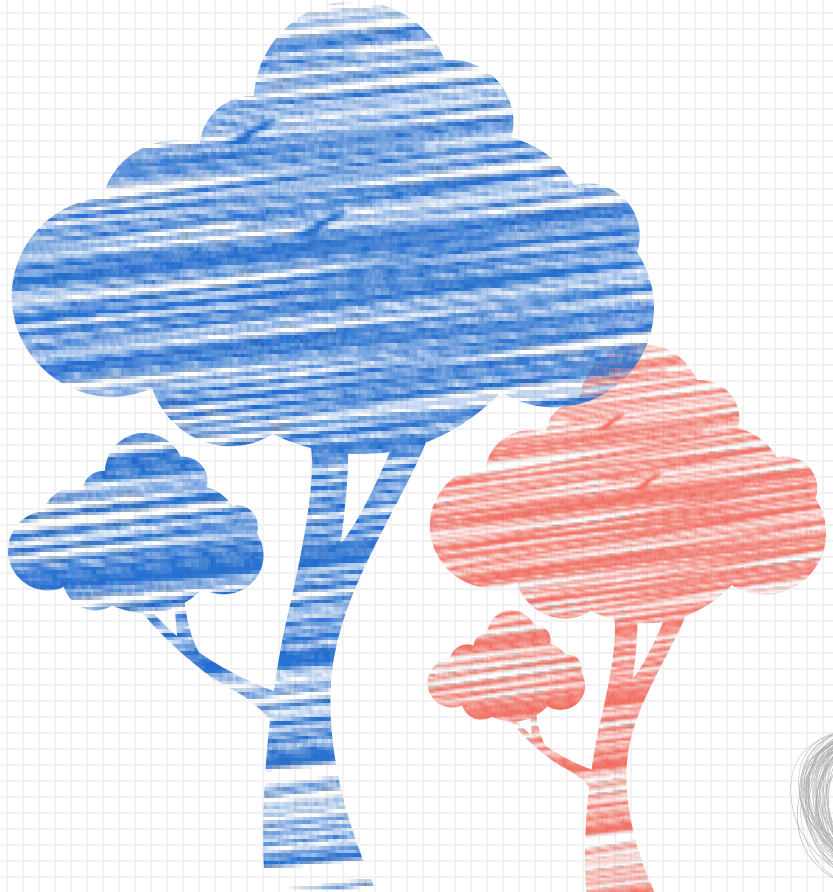


4.1 简单随机抽样（SRS）
(simple random sampling)



（1）简单随机抽样实施方法

- 抽签法
- 随机数法：
  - 随机数表法
  - 随机数骰子产生随机数
  - 计算机产生：随机数发生器



(1) 简单随机抽样实施方法

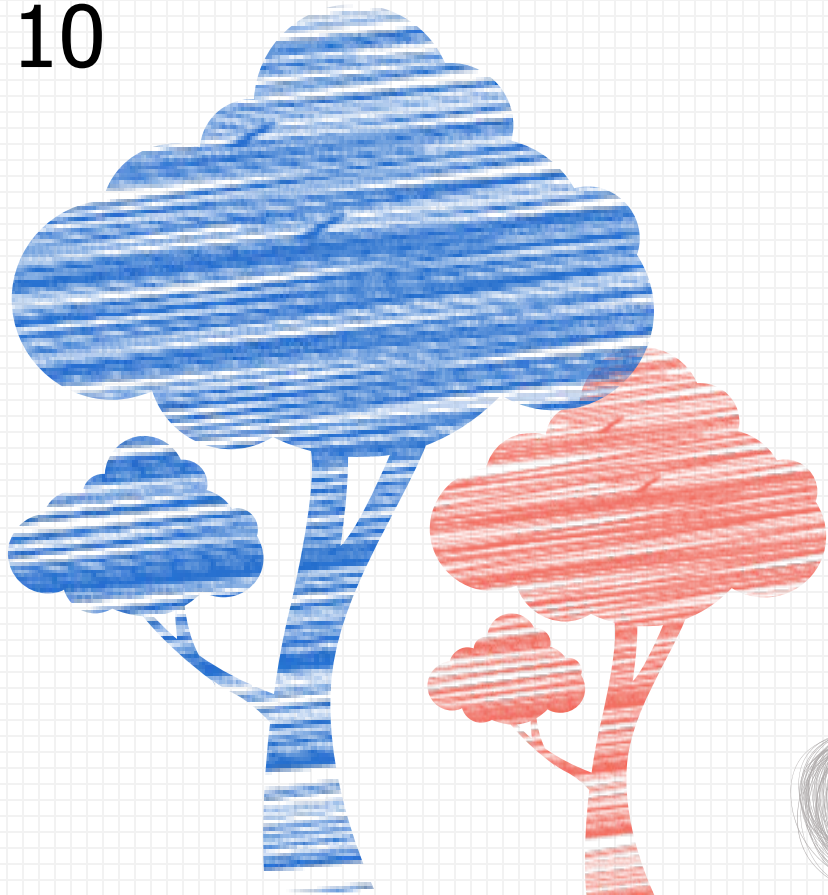
40 50 60 100 120 140 340 260 180 210

1 2 3 4 5 6 7 8 9 10

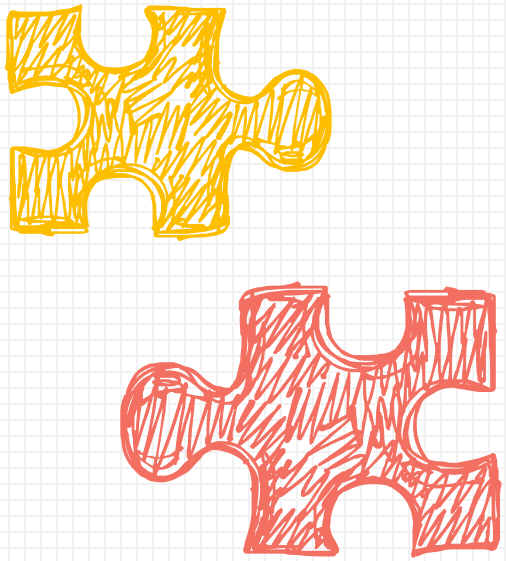
1 2 3 50

5 6 7 200

5 6 10 156.67



（2）简单随机抽样的优点



1. 是最简单的抽样技术；

2. 抽样框不需要其他（辅助）信息，唯一需要的只是一个关于调查总体所有单元的一个完全的清单和与其如何联系的信息；

3. 关于样本量的确定、总体估计与方差估计都有现成的标准公式可以利用，因此技术发展已经成熟。

（3）简单随机抽样的缺点

1. 抽样框中即使有现成的辅助信息也不加利用，使得估计的统计效率较其他利用辅助信息的样本设计低；
2. 由于样本在总体中的地理分布范围比较广，如果采用面访，费用较高；
3. 有可能抽到一个“差的”样本；
4. 如果不用计算机，而用随机数表抽一个大样本将十分单调劳神。

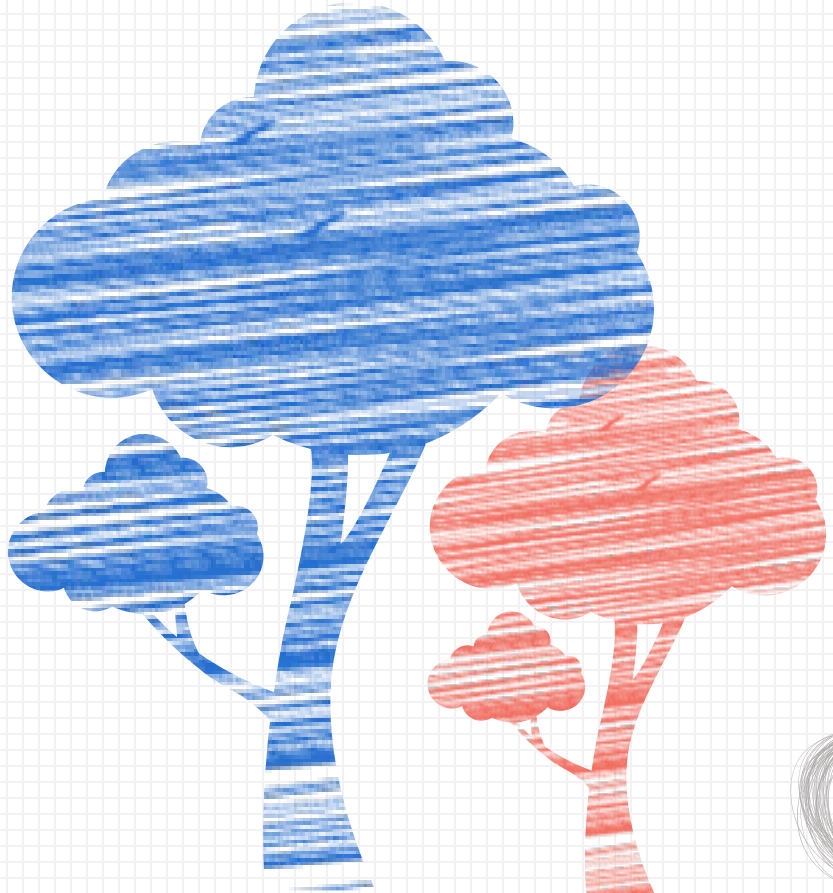
（4）简单随机抽样对总体的推断

对于简单随机抽样, 有 $E(\bar{y})=\bar{Y},\ E(\hat{Y})=Y.$

$$\hat{\bar{Y}}=\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$\hat{Y}=N\bar{y}=\frac{N}{n}\sum_{i=1}^{n}y_i$$

$$E(\bar{y})=\bar{Y}$$



（4）简单随机抽样对总体的推断

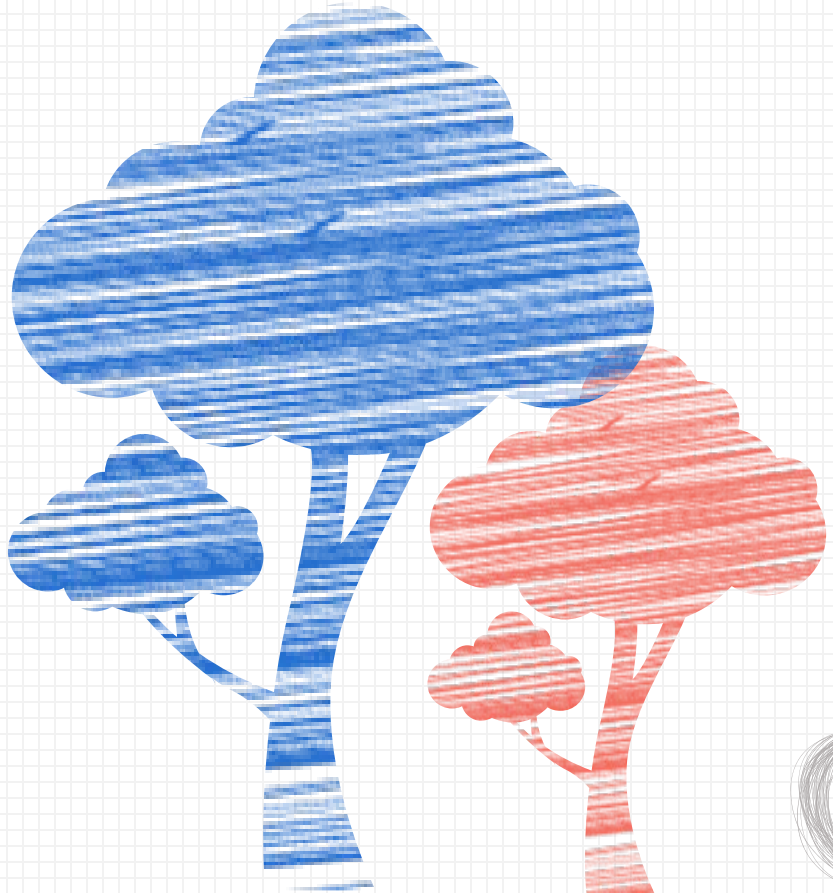
对于简单随机抽样，有

$$V(\bar{y}) = \frac{1-f}{n} S^2 = \frac{N-n}{nN} S^2,$$

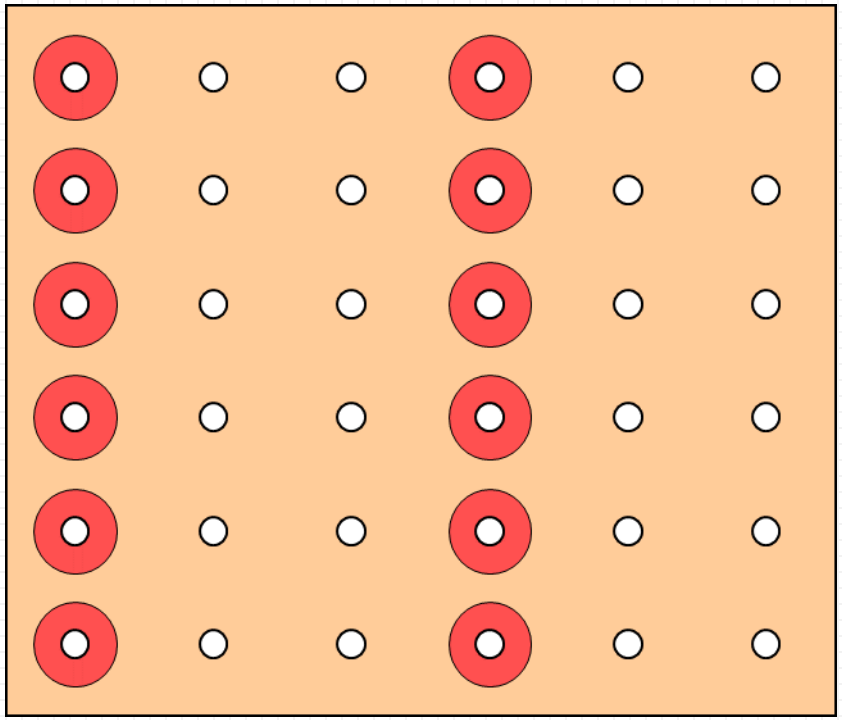
$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2。$$

$$v(\bar{y}) = \frac{1-f}{n} s^2 \text{ 是 } V(\bar{y}) \text{ 的U.E.,}$$

$$\text{其中} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

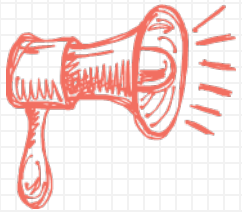


4.2系统抽样（SYS）
(systematic sampling)



（1）系统抽样定义

将总体单元按一定的顺序排列，先随机抽取起始单元，然后按照既定的某种规则抽取其他样本单元。



等距抽样是系统抽样的一种。

（2）等距抽样实施方法

等距抽样的实施方法

设总体含N个单元，从中抽取一个容量为n的样本。将N个总体单元直线排列，编上1至N 的号码。取一个与N/n 最接近的整数 k。在1~k中随机选取一个整数r，然后以第r个单元为起始单元，每隔k个单元抽取一个样本单元。

例，N=10, n=3, 取 k=3,

若 r=1, 则抽中{1,4,7,10};

r=2, 则抽中{2,5,8};

r=3, 则抽中{3,6,9}.

缺点：当N/n非整数时，样本单元数可能为n，也可能为n+1 或 n-1。

（2）等距抽样实施方法

2) Lahiri 的圆周抽样法

将1~N个单元排成一个圆周。从1~N中随机选取一个整数r，以第r个单元为起点，每隔k个单元抽取一个样本单元，直到抽足n个为止。

特点：每个单元等概率入样，样本单元数严格等于n。

（3）系统抽样优缺点

系统抽样的优点

（1）没有抽样框时，可代替简单随机抽样方法简单

（2）不需要辅助的抽样框信息；

系统抽样的缺点

（1）若抽样间隔与总体的某种周期性变化一致，会得一个差的样本；

（2）不使用辅助信息使抽样效率不高；

（3）使用概念框时，不能预先知道样本量；

（4）没有一个无偏的方差估计量；

（5）当N不能被n整除时会得到样本量不同的样本。

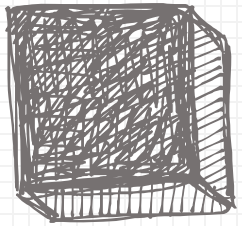
系统抽样的特点及其应用

优点1

方便易行，适宜在现场抽样。
例如，在生产线上抽产品检验；
对在校学生的调查；

优点2

抽样框编制简单。



缺点

精度与单元排列顺序密切相关。

5. N=nk 时的等距样本

设总体含N个单元，从中抽取一个容量为n的样本

$$
\begin{array}{cccc}
y_1 & y_2 & \cdots & y_k \\
y_{k+1} & y_{k+2} & \cdots & y_{2k} \\
\cdots & \cdots & \cdots & \cdots \\
y_{(n-1)k+1} & y_{(n-1)k+2} & \cdots & y_N
\end{array}
$$

$$
\begin{array}{cccc}
y_{1,1} & y_{2,1} & \cdots & y_{k,1} \\
y_{1,2} & y_{2,2} & \cdots & y_{k,2} \\
\cdots & \cdots & \cdots & \cdots \\
y_{1,n} & y_{2,n} & \cdots & y_{k,n} \\
--- & --- & --- & --- \\
\bar{y}_1 & \bar{y}_2 & \cdots & \bar{y}_k
\end{array}
$$

- 把每一列看作一个群，则等距抽样相当于只抽一个群的整群抽样；

- 把每一行看作一个层，那等距抽样相当于每层只抽一个单元的分层抽样。

一、估计量

假如用 SRS 抽取了第 r 个等距样本，

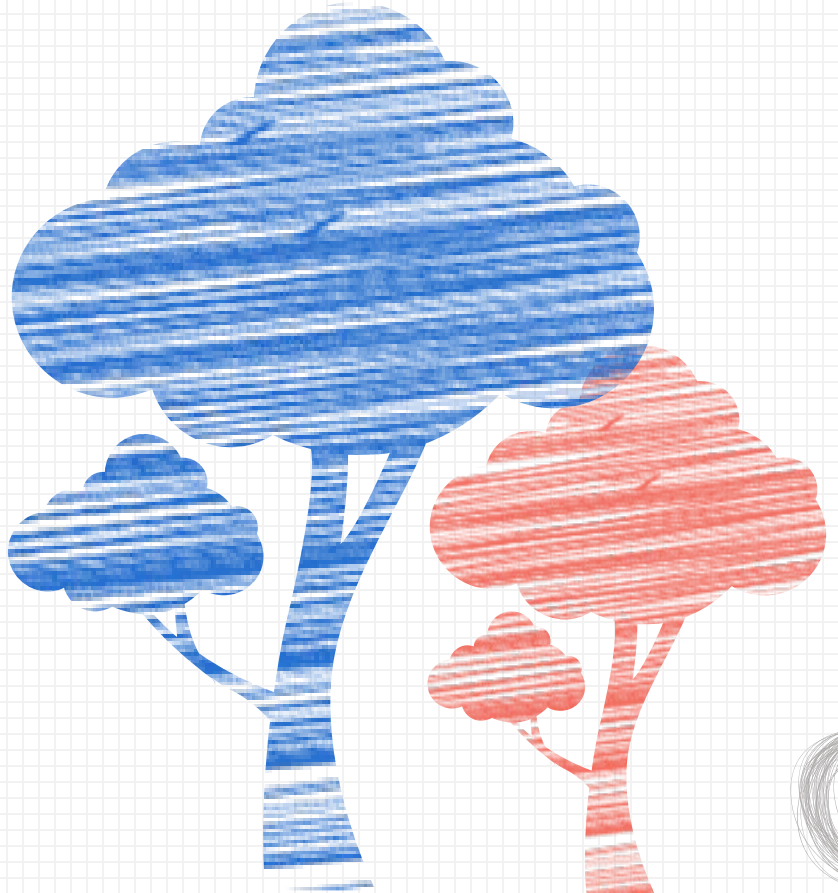
则总体均值 $\bar{Y}$ 常用样本均值来估计，即

$$\bar{y}_{sy} = \bar{y}_r = \frac{1}{n}\sum_{j=1}^{n} y_{rj}$$

性质： （$N = kn$ 时）

1. $\bar{y}_{sy}$ 是 $\bar{Y}$ 的无偏估计；

2. $V(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2$



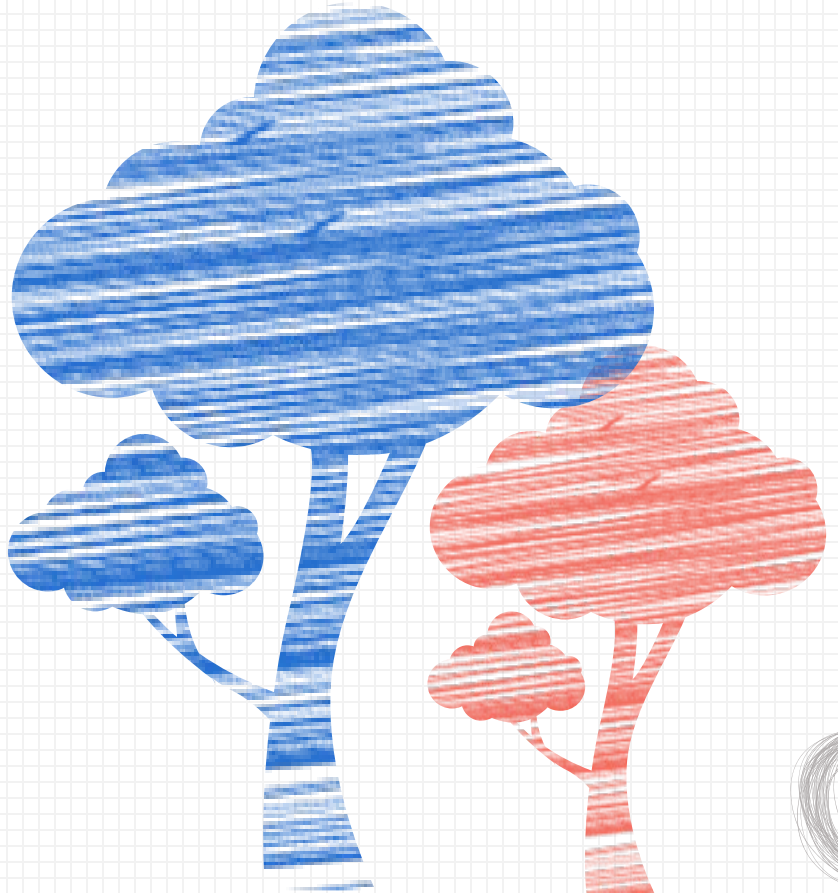
二、方差的三种表示法

用等距样本内方差表示估计量的方差

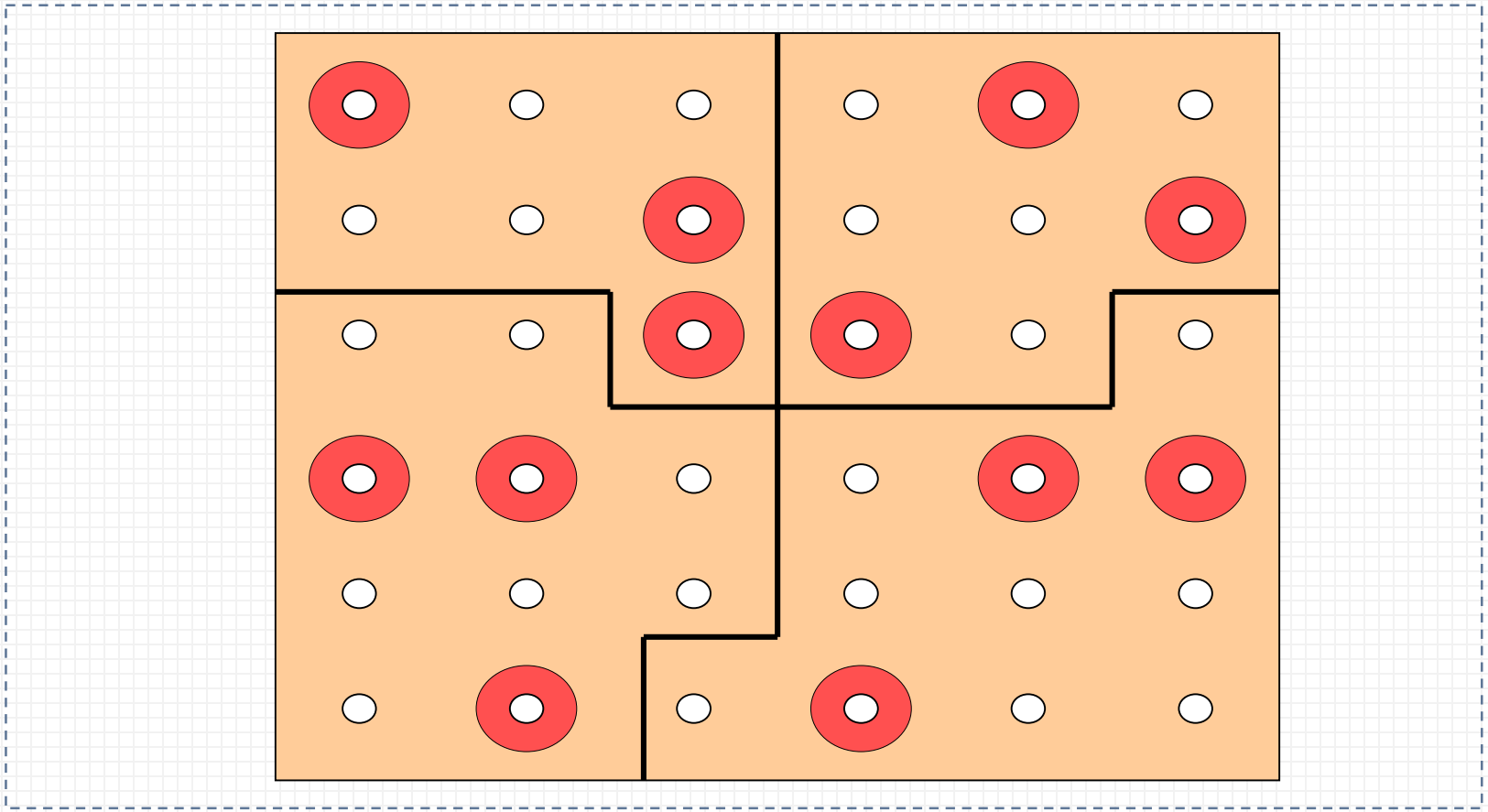
$$V(\bar{y}_{sy}) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_{wsy}^2 ,$$

$$S_{wsy}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$$，称为**样本内方差**。

- S_{wsy}^2越大，则 $V(\bar{y}_{sy})$ 越小。
- 当 $S_{wsy}^2 > S^2$ 时，等距抽样的精度优于SRS。
- S_{wsy}^2的大小与总体单元的排列次序有关，故等距抽样的精度与总体单元的排列次序有关。



4.3 分层抽样
(stratified sampling)



分层抽样优点和缺点

1）在调查中不仅需要对总体进行参数估计，也需要对层的参数进行估计。
2）使样本更具代表性。
3）便于组织管理和数据汇总
4）对不同层可以按照不同情况和条件，具体采用不同的抽样方法。
5）分层抽样可以提高估计量的精度
分层抽样的缺点：
（1）对抽样框的要求比较高，必须有分层的辅助信息；
（2）收集或编制抽样框的费用比较高；
（3）若调查变量与分层的变量不相关，效率可能降低；
（4）估计值的计算比简单随机抽样复杂。

1. $\bar{Y}$ 的估计

若 $\hat{\bar{Y}}_h$ 是 $\bar{Y}_h$ 的估计，则常用其按 W_h 的加权平均估计 $\bar{Y}$,

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \hat{\bar{Y}}_h$$

$$\hat{Y}_h = N_h \hat{\bar{Y}}_h$$

$$\hat{Y}_{st} = N \hat{\bar{Y}}_{st} = \sum_{h=1}^{L} N_h \hat{\bar{Y}}_h$$

2. 性质

对一般的分层抽样，若 $\hat{\bar{Y}}_h$ 是 $\bar{Y}_h$ 的U.E.，$h = 1, \cdots, L$，则

(1) $\hat{\bar{Y}}_{st}$ 是 $\bar{Y}$ 的U.E.；

(2) $V(\hat{\bar{Y}}_{st}) = \sum_{h=1}^{L} W_h^2 V(\hat{\bar{Y}}_h)$。

二、分层随机抽样

对每一层都进行SRS的分层抽样称为分层随机抽样，是最常用、最简单的一种分层抽样方法。

估计量 $\bar{y}_{st}=\sum_{h=1}^{L}W_h\bar{y}_h=\frac{1}{N}\sum_{h=1}^{L}N_h\bar{y}_h$，是 $\bar{Y}$ 的U.E.；

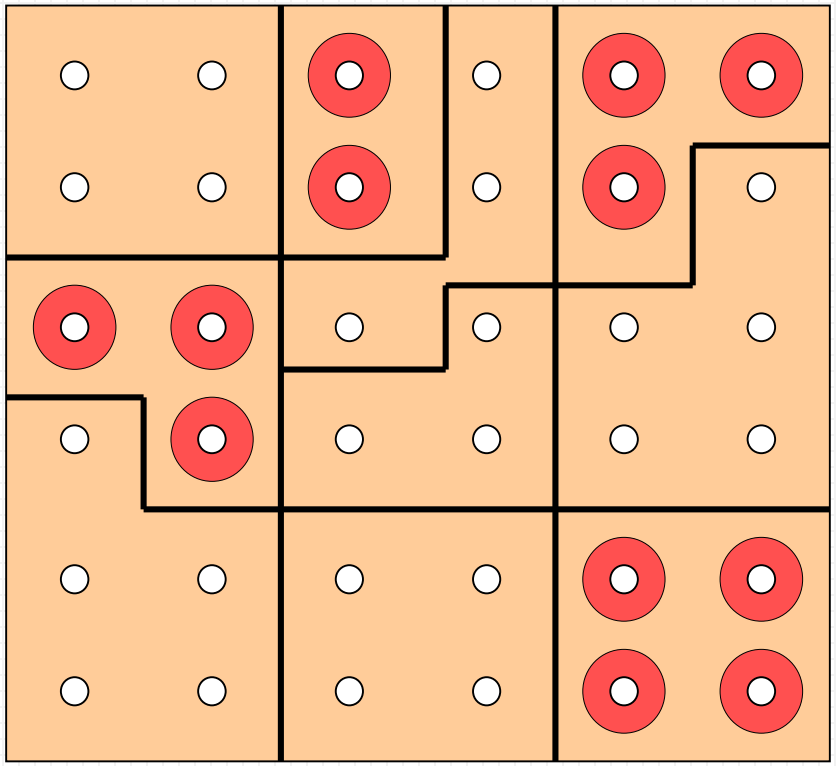
方差 $V(\bar{y}_{st})=\sum_{h=1}^{L}W_h^2V(\bar{y}_h)$

$$=\sum_{h=1}^{L}W_h^2\frac{1-f_h}{n_h}S_h^2$$

方差的一个U.E. $v(\bar{y}_{st})=\sum_{h=1}^{L}W_h^2\frac{1-f_h}{n_h}s_h^2$

$$=\sum_{h=1}^{L}\frac{W_h^2s_h^2}{n_h}-\frac{1}{N}\sum_{h=1}^{L}W_hs_h^2。$$

4.4 整群抽样
(cluster sampling)



什么是整群抽样

例. 实施下面几项调查，分别可能采用两种抽样方案，比较这两种方案。

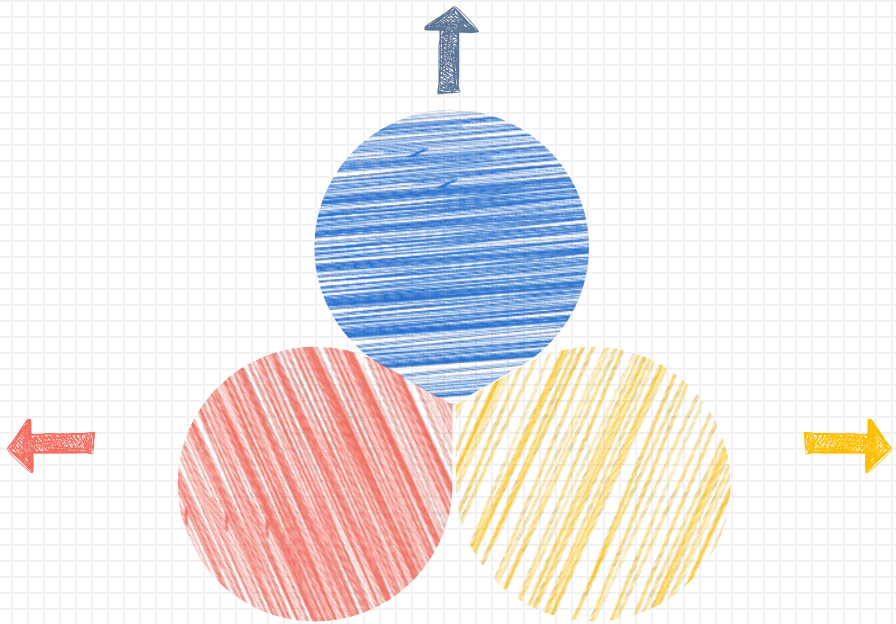
1. **学校需了解住宿学生对宿舍管理的意见。**
   a) 抽个人，在所有住宿生中进行SRS;
   b) 抽寝室，在所有寝室中进行SRS，对抽中寝室中的每个人作调查。
2. **一批产品的合格检验。产品每20个装一箱。**
   a) 从所有产品中抽一个简单随机样本;
   b) 抽取若干箱，整箱检验。

一、定义

将总体分割成若干个较大的单元，然后对这些大单元进行抽样，对抽中的大单元中的所有个体实施全面调查。这种抽样方法称为整群抽样。

这些大单元称为初级单元(即群)，
个体称为次级单元。

单阶整群抽样。



若每个初级单元还可以分割为若干个较小的次级单元，每个次级单元包含若干个个体(三级单元)。先抽初级单元，再在抽中初级单元中抽取若干个次级单元，对抽中次级单元作全面调查。这种抽样方法称为二阶整群抽样。

二、实施抽样的优缺点

优点：

1. **抽样框编制简便**

只须以群为基本单位来编制抽样框。

2. **样本单元相对集中**

便于实施、节约时间、费用，也便于人员培训、保证调查数据的质量。

缺点：

与容量同样的SRS相比，精度一般较差。

三、如何分群?

1. **从方便事实调查的角度**

以自然单位划分群，比如，一个寝室、一箱产品、一个班级、一栋居民楼、一个企业等等。

2. **从精度角度看，要求群的代表性好。**

原则：群内差异大、群间差异小。

与分层原则刚好相反。

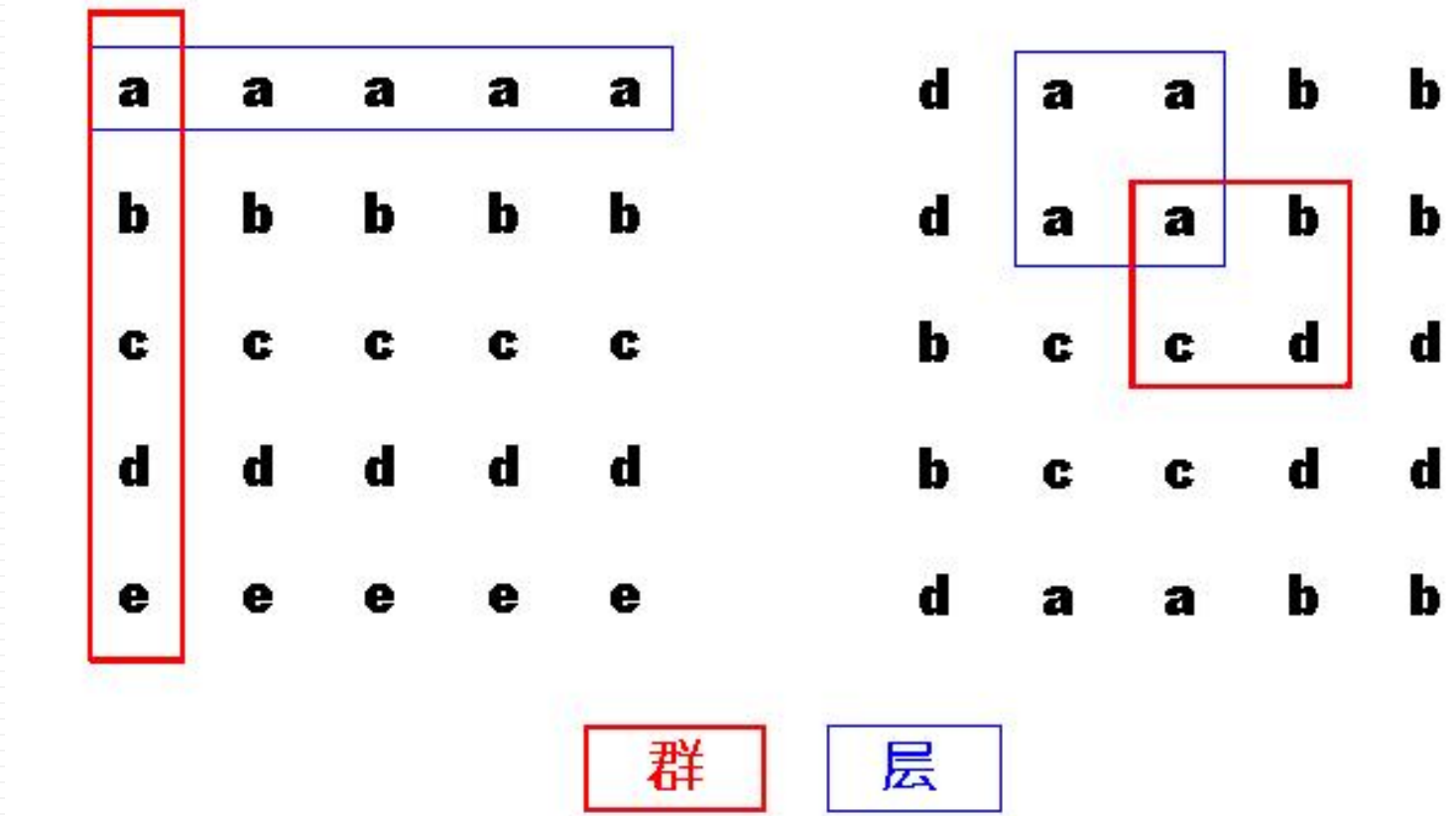
3. **群的大小**

最好相等，便于统计分析。

群的规模要适中，过大，则费用省但精度差；

过小，则精度高但费用也高。

划分群、层的示意图


a a a a a
b b b b b
c c c c c
d d d d d
e e e e e
d a a b b
d a a b b
b c c d d
b c c d d
d a a b b
群
层

整群抽样的优点

（1）大大减低收集数据的费用；

（2）当总体单元自然形成的群时，容易取得抽样框，抽样也更容易；

（3）当群内单元差异大，而不同群之间的差异小时，

可以提高效率。

缺点：

（1）若群内个单元有趋同性，效率将会降低；

（2）通常无法预先知道总样本量，因为不知道群内有多少单元；

（3）方差估计比简单随机抽样更为复杂

可以综合利用分层和整群抽样技术，采取分层整群抽样，比如人体尺寸调查，采用分层提高样本代表性，采用整群抽样，便于数据的收集。

估计量及其性质

1. Y 的估计（群大小相等）

估计量：$\hat{\bar{Y}} = \bar{y}$，$\hat{Y} = N \cdot \bar{y}$

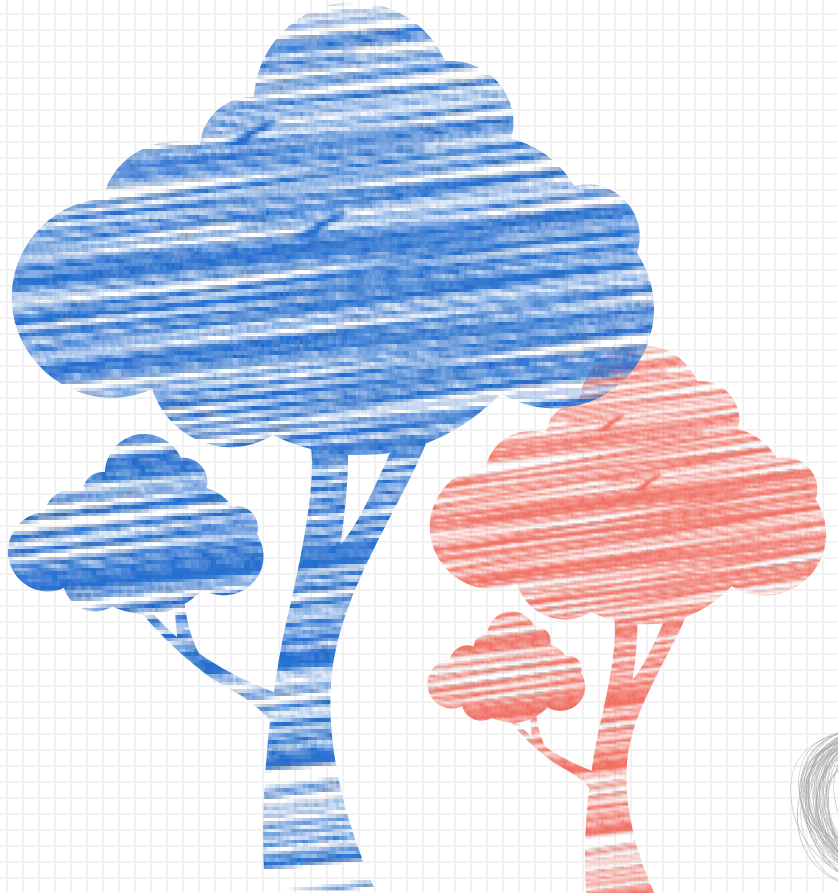
性质：

(1) $\hat{Y}$ 是 Y 的 U.E.；

(2) $V(\hat{Y}) = N^2 V(\bar{y}) = N^2 \dfrac{1-f}{n} \dfrac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$ 。

方差估计量：

$v(\hat{Y}) = N^2 \dfrac{1-f}{n} \dfrac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{y})^2$ 是 $V(\hat{Y})$ 的U.E.。



二、简单估计量

2. $\bar{\bar{Y}}$ 或 P 的估计 （前提：M_0 已知）

估计量： $\bar{\bar{Y}} = \dfrac{\bar{Y}}{M}$ $\hat{\bar{\bar{Y}}} = \dfrac{\bar{y}}{M} = \bar{\bar{y}}$

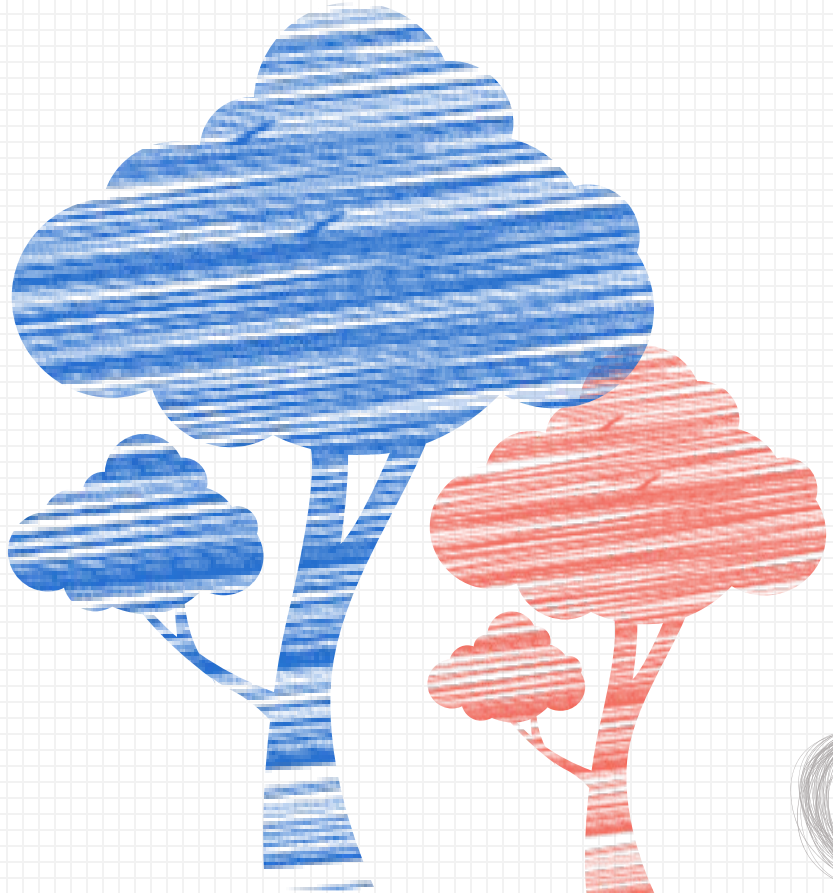
性质：

(1) $\hat{\bar{\bar{Y}}}$ 是 $\bar{\bar{Y}}$ 的 U.E.;

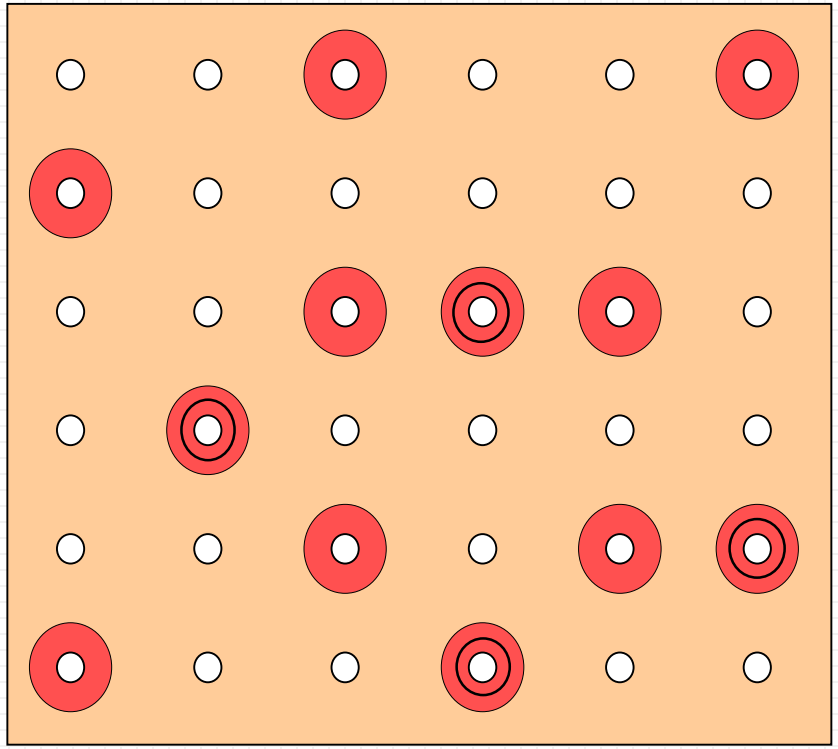
(2) $V(\hat{\bar{\bar{Y}}}) = \dfrac{1-f}{n}\dfrac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 = \dfrac{1-f}{nM}S_b^2$

方差估计量：

$$v(\hat{\bar{\bar{Y}}}) = \frac{1-f}{nM}s_b^2$$



4.5 多重（相）抽样
(multi-phase sampling)



二重抽样定义

抽样分两步进行：

1）第一重抽样：从总体中抽取比较大的样本，获取有关总体的某些辅助信息；

2）第二重抽样：所抽取的样本比较少，此次调查为主调查，第二重样本是第一重样本的子样本，也可以从总体中独立抽取。

注：本书考虑的第二重样本是第一重样本的子样本，第一重抽样限于简单随机抽样。

二重抽样的应用

1. 从总体单元中筛选主要调查对象；
2. 在多指标调查中总体指标差异较大或对目标量的估计的精度要求不同，并不需要相同的样本量；
3. 对于那些为提高抽样效率在抽样或构造估计量时需要总体某些辅助信息；
4. 在连续性调查中，同一单元不同时间的指标值相关，利用相关性，采用回归估计可提高精度。

二重分层抽样的样本抽选方法

总体　　$1,\cdots,\ N$

第一重抽样　$1,\cdots,\ n'$（SRS抽样）

第一重样本分层$n_1^{'}\cdots,\ n_L^{'}$

第二重抽样在第一重样本中进行分层随机抽样，抽取样本量为n，每层样本量为n_h,每层抽样比（指定）为γ_h。

层权估计$w_h^{'}=\dfrac{n_h^{'}}{n'},\ E(w_h^{'})=\dfrac{N_h}{N},$

$$\gamma_h=\frac{n_h}{n_h^{'}}$$

一、二重分层抽样的估计方法

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi},$$

$$\bar{y}_{stD} = \sum_{h=1}^{L} w_h' \bar{y}_h$$

$$V(\bar{y}_{stD}) = \left(\frac{1}{n'} - \frac{1}{N}\right)S^2 + \frac{1}{n'}\sum_{h=1}^{L}\left(\frac{1}{\gamma_h} - 1\right)W_h S_h^2$$

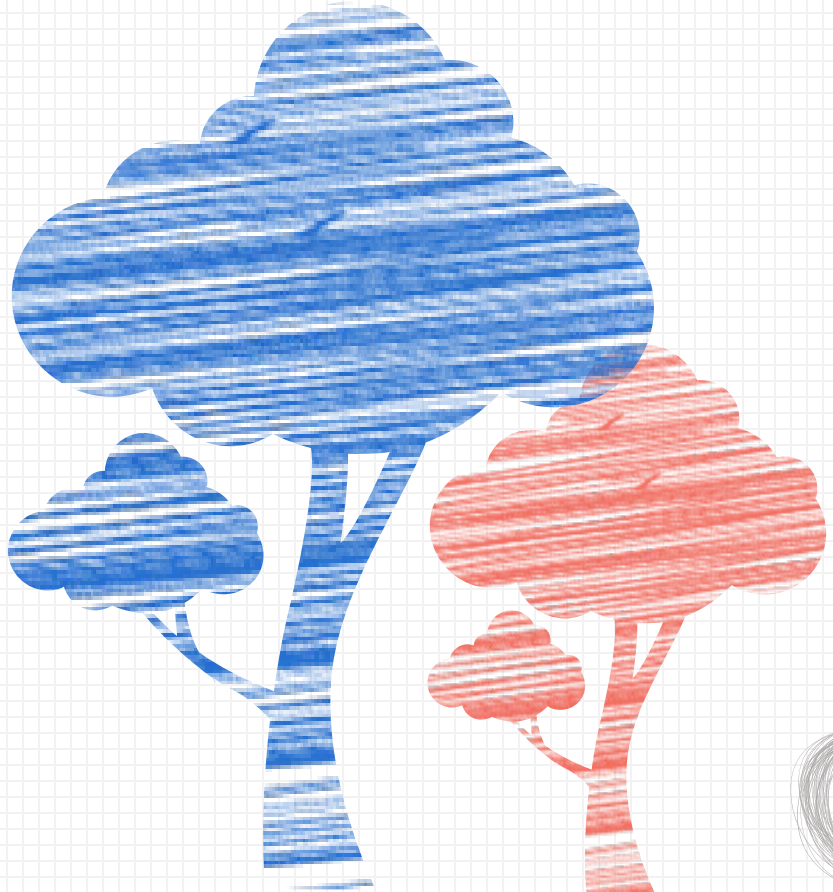
$$v(\bar{y}_{stD}) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{n_h'}\right)w_h'^2 s_h'^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)\sum_{h=1}^{L} w_h'(\bar{y}_h - \bar{y}_{stD})^2$$

$$v(\bar{y}_{stD}) \approx \sum_{h=1}^{L}\frac{w_h'^2 s_h'^2}{n_h} + \frac{1}{n'}\sum_{h=1}^{L} w_h'(\bar{y}_h - \bar{y}_{stD})^2$$

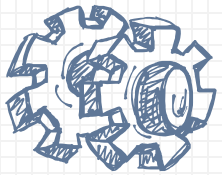
4.6 不等概率抽样

为什么采用不等概率抽样？

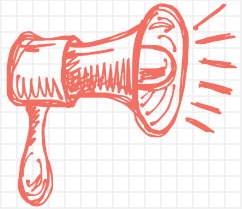
等概率抽样每个单元被抽中的概率相同，总体单元相差很大，等概率抽样的效果不好。



不等概率抽样优点



提高估计精度



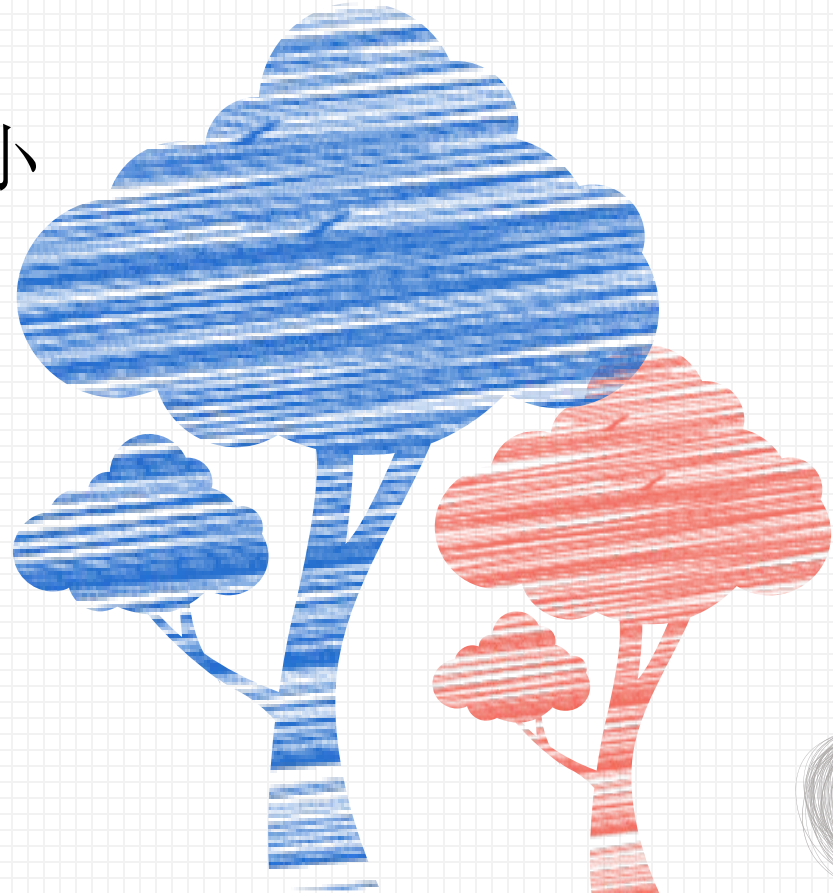
减少抽样误差

不等概率抽样分类

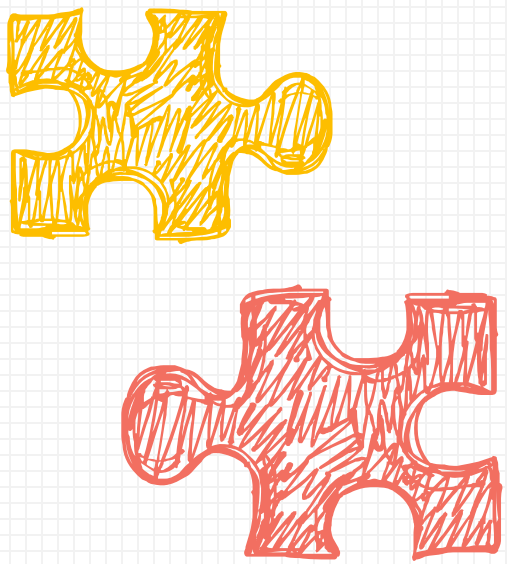
1、是否放回

- 放回不等概率抽样
- 不放回不等概率抽样

2、单元被抽中的概率是否严格与单元大小成比例



放回不等概率抽样



1. 多项抽样与PPS抽样
2. 多项抽样的实施方法
3. 汉森-赫维茨估计量及其性质

多项抽样与PPS抽样

设总体包含N个单元，进行放回抽样，抽中第i个单元的概率为$Z_i, i=1,2,\cdots,N, \sum_{i=1}^{N} Z_i = 1$，独立抽样$n$次，称这种不等概率抽样为多项抽样。

记t_i为总体第i个单元被抽中的次数，则$1 \le t_i \le n, \sum_{i=1}^{N} t_i = n$

则$(t_1, t_2 \cdots t_N)$的联合分布是多项分布

$$\frac{n!}{t_1!t_2!\cdots t_N!} Z_1^{t_1} Z_2^{t_2} \cdots Z_N^{t_N}$$

多项抽样与PPS抽样

若每个单元说明其大小或规模的度量为M_i时，

$$Z_i = \frac{M_i}{M_0}$$

$M_0 = \sum_{i=1}^{N} M_i$，则每个单元被抽中的概率与单元大小成比例，称这种多项抽样为（放回的）与大小成比例的概率抽样(*sampling with probablity proportional to size*)，简称*PPS*抽样。

多项抽样的实施方法

代码(Hansen-Hurwitz)法

拉希里(Lahiri) 法

代码(Hansen-Hurwitz)法

- 第一步：找一个整数M_0，使每个$M_i = M_0 Z_i$都为一个整数$(i = 1, \ldots, N)$
- 第二步：将$[1，M_0]$分割为N个区域，使这N个区域与总体中N个单元格建立起一一对应关系，具体对应方式为：

  单元1——$[1，M_1]$

  单元2——$[M_1 + 1，M_1 + M_2]$

  …

  单元N——$[M_1 + \cdots + M_{N-1} + 1, M_0]$
- 第三步：设抽取的样本数为n，则在$[1，M_0]$中产生n个随机数，与这些随机数所在的n个区域相对的单元就是样本单元。

【例题1】

设某个总体有N=10个单元，相应的单元大小M_i及其代码数如右表所示，要在其中产生一个n=3的样本，试用代码法抽取样本。

i	M_i	$M_i * 10$	累计$M_i * 10$	代码
1	0.6	6	6	1~6
2	14.5	145	151	7~151
3	1.5	15	166	152~166
4	13.7	137	303	167~303
5	7.8	78	381	304~381
6	15	150	531	382~531
7	10	100	631	532~631
8	3.6	36	667	632~667
9	6	60	727	668~727
10	1.1	11	738	728~738
总计	$M_0 = 73.8$	738	——	——

若在[1,738]中产生三个随机数为354，553，493，则该抽出哪几个总体单元？

汉森-赫维茨估计量及其性质

对于多项抽样，设$y_1, y_2, \ldots, y_n$是按照Z_i入样概率进行多项抽样得到的观测值，相应的Z_i的值为$z_1, z_2, \ldots, z_n$，则总体总和的估计为

$$\hat{Y}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i}$$

汉森-赫维茨估计量性质

(1) $E(\hat{Y}_{HH}) = Y;$

(2) $V(\hat{Y}_{HH}) = \frac{1}{n}\sum_{i=1}^{N} Z_i \left(\frac{Y_i}{Z_i} - Y\right)^2$

(3)若$n > 1$，则

$$v(\hat{Y}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{z_i} - \hat{Y}_{HH}\right)^2$$

是$V(\hat{Y}_{HH})$的无偏估计。

不放回不等概率抽样

一、**入样（包含）概率的定义**

每个单元被包含到样本中的概率。

二阶入样（包含）概率：两个单元同时被包含到样本中的概率。

入样（包含）概率的性质

对于固定的 n，包含概率 π_i 与 π_{ij} 满足以下性质：

1） $\sum_{i=1}^{N}\pi_i = n$；

$$\sum_{i\neq j}^{N}\pi_{ij} = \sum_{i\neq j}^{N}\Pr(i)\Pr(j|i) = \pi_i\sum_{i\neq j}^{N}\Pr(j|i) = (n-1)\pi_i$$

2） $\sum_{i\neq j}^{N}\pi_{ij} = (n-1)\pi_i$ ；

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_{ij} = \frac{1}{2}(n-1)\sum_{i=1}^{N}\pi_i = \frac{1}{2}n(n-1)$$

3） $\sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij} = \frac{1}{2}n(n-1)$ 。

霍维茨-汤普森估计量

总体总量Y的估计可采用霍维茨-汤普森（Horvitz-Thompson）估计量：

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

霍维茨-汤普森估计量性质

1）若 $\pi_i > 0, \quad i = 1, 2, \cdots, N$，则 $\hat{Y}_{HT}$ 是 Y 的无偏估计，且它的方差为：

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j$$

当 n 固定时，又有

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

霍维茨-汤普森估计量性质

2）若 $\pi_i > 0, \quad \pi_{ij} > 0, \quad i, j = 1, 2, \cdots, N, \quad i \neq j$，则

$$v(\hat{Y}_{HT}) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j$$

是的$V(\hat{Y}_{HT})$无偏估计。当n固定时，下面耶茨(Yates)-格伦迪(Grundy)-(sen)估计量$v(\hat{Y}_{YGS})$也是$V(\hat{Y}_{HT})$无偏估计。

n=2的严格πPS抽样

1）布鲁(Brewer)方法

要求每个$Z_i < \frac{1}{2}$,

第一个样本按照与$\frac{Z_i(1-Z_i)}{1-2Z_i}$成比例的概率抽取；

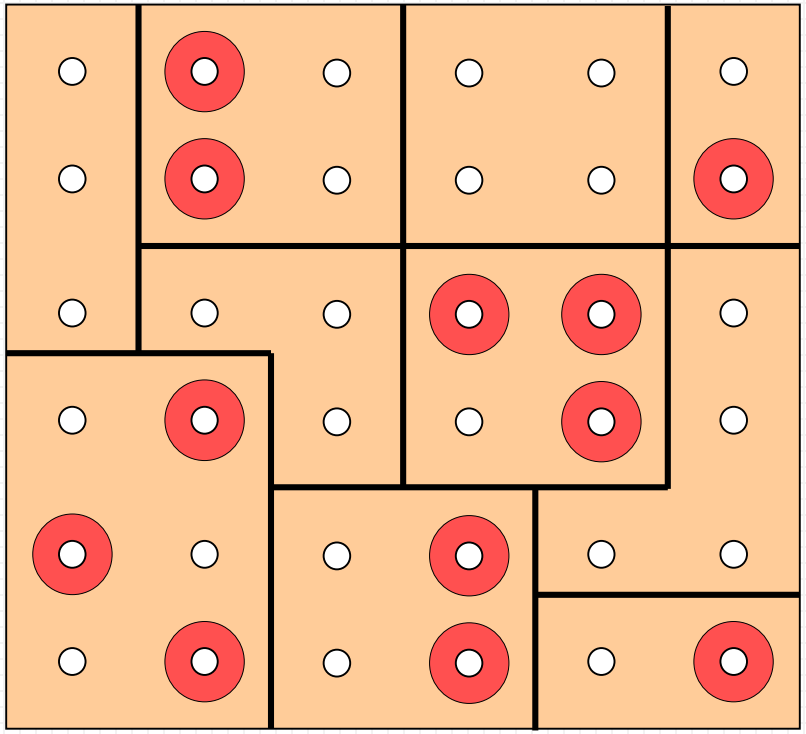
第二个样本在其余单元中按照与Z_j成比例的概率抽取。

i	M_i	$Z_i = M_i / M_0$	$\frac{Z_i(1-Z_i)}{1-2Z_i}$	累计 $\frac{Z_i(1-Z_i)}{1-2Z_i}$	累计 $Z_j (j \neq 5)$
1	23	0.092	0.1024	0.1024	0.092
2	45	0.180	0.2306	0.3330	0.272
3	11	0.044	0.0461	0.3791	0.316
④	68	0.272	0.4342	0.8133	0.588
⑤	37	0.148	0.1791	0.9924	—
6	20	0.080	0.0876	1.0800	0.668
7	27	0.108	0.1229	1.2029	0.776
8	19	0.076	0.0828	1.2857	0.852
$\sum$		1.0000	D =1.2857		

$$\hat{Y}_B = \frac{y_1}{\pi_1} + \frac{y_2}{\pi_2} = \frac{1}{2}\left(\frac{y_1}{z_1} + \frac{y_2}{z_2}\right)$$

$$v_{YGS}(\hat{Y}_B) = \frac{\pi_1\pi_2 - \pi_{12}}{\pi_{12}}\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2$$

4.7 多阶抽样
(multi-stage sampling)



次级单元均值的估计量及其性质

考虑两阶段都采用SRS的情形。且第二阶段的抽样比均相同。

若两阶段都是简单随机抽样，则

1) $\bar{\bar{y}}$ 是 $\bar{\bar{Y}}$ 的无偏估计；

2) $V(\bar{\bar{y}}) = \dfrac{1-f_1}{n}S_1^2 + \dfrac{1-f_2}{nm}S_2^2$；

3) $v(\bar{\bar{y}}) = \dfrac{1-f_1}{n}s_1^2 + \dfrac{f_1(1-f_2)}{nm}s_2^2$ 是 $V(\bar{\bar{y}})$ 的无偏估计。

其中，$f_1 = n/N$ 为第一阶段的抽样比，$f_2 = m/M$ 为第二阶段的抽样比。

4.8 比估计与回归估计

（1）如何估计两个指标的总和（或均值）之比 R=Y/X；

（2）如何利用变量 X与 Y之间的关系，以及 X的辅助信息，对Y 之总和或均值作更好的估计。

例： 家庭开支和家庭收入的比值

抽样调查中经常采用辅助信息来提高估计的精度。

辅助信息往往是先于（或独立于）本次抽样调查得到的，常体现在若干个辅助变量中。

有些抽样框一开始就配有一个或若干个辅助变量，例如：

（1）各街道最近依次人口普查时的居民人数

（2）学校在校学生的年龄、性别、平均绩点

辅助变量的利用（假定与待研究的变量有关系）：

（1）用于抽样设计

（2）用于构造估计

R的估计(Ratio Estimator)

对于SRS，$\hat{R}=\bar{y}/\bar{x}$

1) $\hat{R}$ 有偏

2) $|B(\hat{R})|/\sqrt{MSE(\hat{R})}=O(1/\sqrt{n})$

若已知辅助变量X的总量$X=\sum_{i=1}^{N}X_i$

或均值$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$，则$Y$的比估计量为

$$\bar{y}_R=\frac{\bar{y}}{\bar{x}}\bar{X}=\hat{R}\bar{X}$$

$$\hat{Y}_R=\frac{\bar{y}}{\bar{x}}X=\hat{R}\bar{X}=N\bar{y}_R=N\hat{R}\bar{X}$$

1) $\hat{R}$有偏，即 $E(\hat{R}) \neq R$。

考虑一个N=6的总体

i	x_i	y_i
1	2	5
2	3	7
3	5	10
4	5	11
5	7	15
6	8	18
总和	30	66
R	2.2	

从中抽取n=4的SRS

样本号	样本单元	x bar	y bar	R hat
1	(1,2,3,4)	3.75	8.25	2.2000
2	(1,2,3,5)	4.25	9.25	2.1765
3	(1,2,3,6)	4.50	10.00	2.2222
…	…	…	…	…
13	(2,3,5,6)	5.75	12.50	2.1739
14	(2,4,5,6)	5.75	12.75	2.2174
15	(3,4,5,6)	6.25	13.50	2.1600
aver.		5	11	2.2014

对于SRS，当n较大时，有

$$E(\hat{R}) \approx R$$

$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2}\frac{\sum_{i=1}^{N}(y_i - Rx_i)^2}{N-1}$$

$$= \frac{1-f}{n\bar{X}^2}(S_y^2 + R^2 S_x^2 - 2RS_{yx})$$

$$= \frac{1-f}{n\bar{X}^2}(S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x)$$

其中 $S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}),\ \rho = \frac{S_{yx}}{S_y S_x}$

方差的估计量

$$v_1(\hat{R}) = \frac{1-f}{n\bar{X}^2}\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1}$$

$$v_2(\hat{R}) = \frac{1-f}{n\bar{x}^2}\frac{\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2}{n-1}$$

回归估计(Regression Estimator)

如果Y与辅助变量X大致上呈正比例关系时，那么可以利用X的总体信息提高对Y的总体总和或均值的估计精度。

实际问题中，Y与X之间还可能呈现除正比例关系外别的关系，比如，较一般的线性关系。这时，又该如何利用这种关系来提高估计精度呢?

估计量

设想每个总体单元的两个指标值 (Y_i, X_i)之间有近似的线性关系：

$$Y_i = a + bX_i + \varepsilon_i, \quad i = 1, \cdots, N,$$

其中ε_i表示偏离这种线性关系的误差，一种自然的要求是 $\sum_{i=1}^{N} \varepsilon_i = 0$。

因此有：

$$\overline{Y} = a + b\overline{X}。$$

可以设想,$(\overline{y}, \overline{x})$偏离该直线不会太远，因此近似地有：

$$\overline{y} \doteq a + b\overline{x}。$$

结合上述二式得：

$$\overline{Y} \doteq \overline{y} + b(\overline{X} - \overline{x})。$$

1. 估计量

从而，我们得到如下形式的回归估计量：

总体均值的估计：

$$\hat{\bar{Y}}_{lr} = \bar{y} + b(\bar{X} - \bar{x}),$$

总体总和的估计：

$$\hat{Y}_{lr} = N \cdot \hat{\bar{Y}}_{lr},$$

其中 $\bar{X}$ 已知。

直观解释：是用 $b(\bar{X} - \bar{x})$ 去调整 $\bar{y}$ 对 $\bar{Y}$ 的估计。

特例：

1) $b = 0$ 时，$\hat{\bar{Y}}_{lr} = \bar{y}$，即简单估计量；

2) $b = \bar{y}/\bar{x} = \hat{R}$ 时，$\hat{\bar{Y}}_{lr} = \hat{R}\bar{X} = \hat{\bar{Y}}_{R}$，即比估计量。

B的含义

用直线 $y=a+bx$去拟合$(Y_i,X_i),i=1,\cdots,N.$

自然地，可要求取 a,b 最小化

$$Q(a,b)=\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{N}(Y_i-a-bX_i)^2$$

Q是 a,b的二次函数，求导求极值可得极值点为：

$$\begin{cases}\hat{a}=\overline{Y}-\hat{b}\overline{X},\\ \hat{b}=B=\dfrac{\sum_{i=1}^{N}(Y_i-\overline{Y})(X_i-\overline{X})}{\sum_{i=1}^{N}(X_i-\overline{X})^2}.\end{cases}$$

B称为总体回归系数。

用样本估计b

b未知时，自然地希望用样本估计 B作为b，以期获得方差最小的回归估计量。

B的估计常取为：

$$\hat{B}=\frac{s_{yx}}{s_x^2}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

但取 $b=\hat{B}$时，$\hat{\bar{Y}}_{lr}$不再是样本观测$(y_i,x_i),i=1,\cdots,n$的线性函数，一般也不再无偏。性质较复杂，只能讨论其大样本性质：

当n较大时，有

(1) $E(\hat{\bar{Y}}_{lr})\approx\bar{Y}$,

(2) $MSE(\hat{\bar{Y}}_{lr})\approx V(\hat{\bar{Y}}_{lr})\approx\frac{1-f}{n}S_y^2(1-\rho^2)$.

4. 用样本估计b

方差估计量可取：

$$v(\hat{\bar{Y}}_{lr}) = \frac{1-f}{n(n-2)} \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{B}(x_i - \bar{x}) \right]^2$$

$$= \frac{1-f}{n(n-2)} \left\{ \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{B}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}$$

5. 回归估计、比估计、简单估计量的比较(大样本情况下)

$$V(\hat{\bar{Y}}_{lr}) \doteq \frac{1-f}{n} S_y^2 (1-\rho^2)$$

$$V(\hat{\bar{Y}}_{R}) \doteq \frac{1-f}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x)$$

$$V(\bar{y}) = \frac{1-f}{n} S_y^2$$

(1) $V(\hat{\bar{Y}}_{lr}) \le V(\bar{y})$, “=” 仅当$\rho = 0$时成立。

(2) $V(\hat{\bar{Y}}_{R}) - V(\hat{\bar{Y}}_{lr}) \approx \frac{1-f}{n}(RS_x - \rho S_y)^2 = \frac{1-f}{n} S_x^2 (R-B)^2 \ge 0,$

“=” 仅当 $RS_x = \rho S_y$ 或 $R = B$时成立。

三 基于模型的统计推断

$$Y=\sum_{k\in s}Y_k+\sum_{k\notin s}Y_k$$

$$\hat{T}(s)=\sum_{k\in s}y_k+U(s)$$

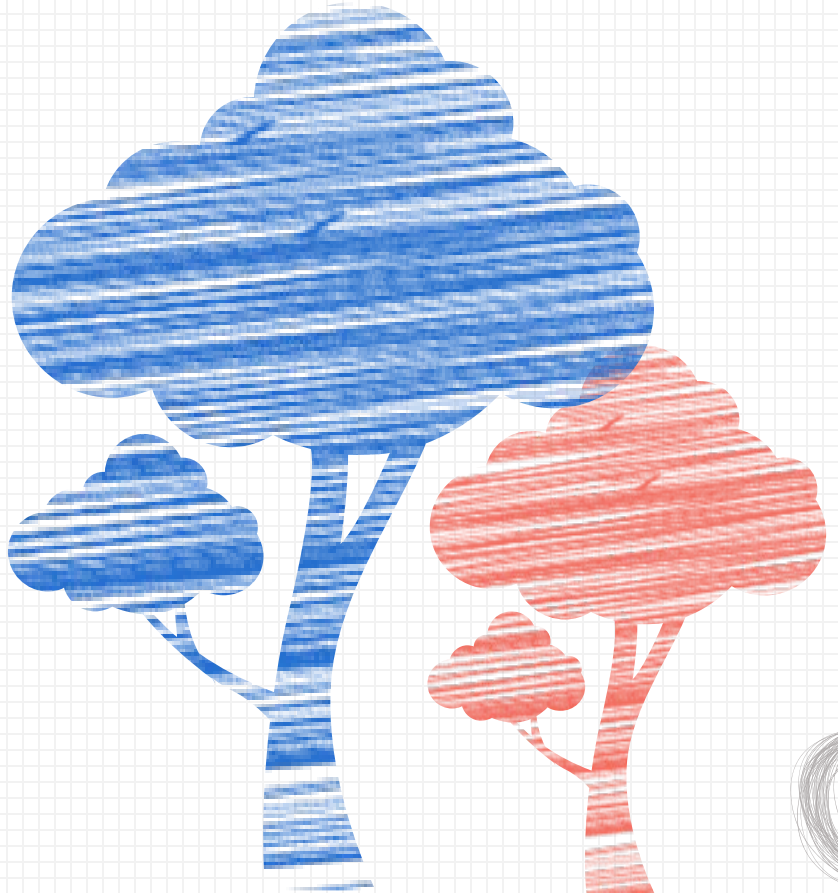
$$\begin{cases} y=X\beta+e \\ \varepsilon(e)=0, Var(e)=\sigma^2 V \end{cases}$$

$\beta=(\beta_0,\beta_1,\cdots,\beta_r)'$， $\sigma^2>0$为未知参数。

$y=(y_1,\cdots,y_N)'$， $X=(x_{ij})_{N\times(r+1)}=(X_1,\cdots,X_N)'$

$y=(y_s^{'},y_{\bar{s}}^{'})', X=(X_s^{'},X_{\bar{s}}^{'})'$

$$V=\begin{pmatrix} V_s & V_{s\bar{s}} \\ V_{\bar{s}s} & V_{\bar{s}} \end{pmatrix}$$

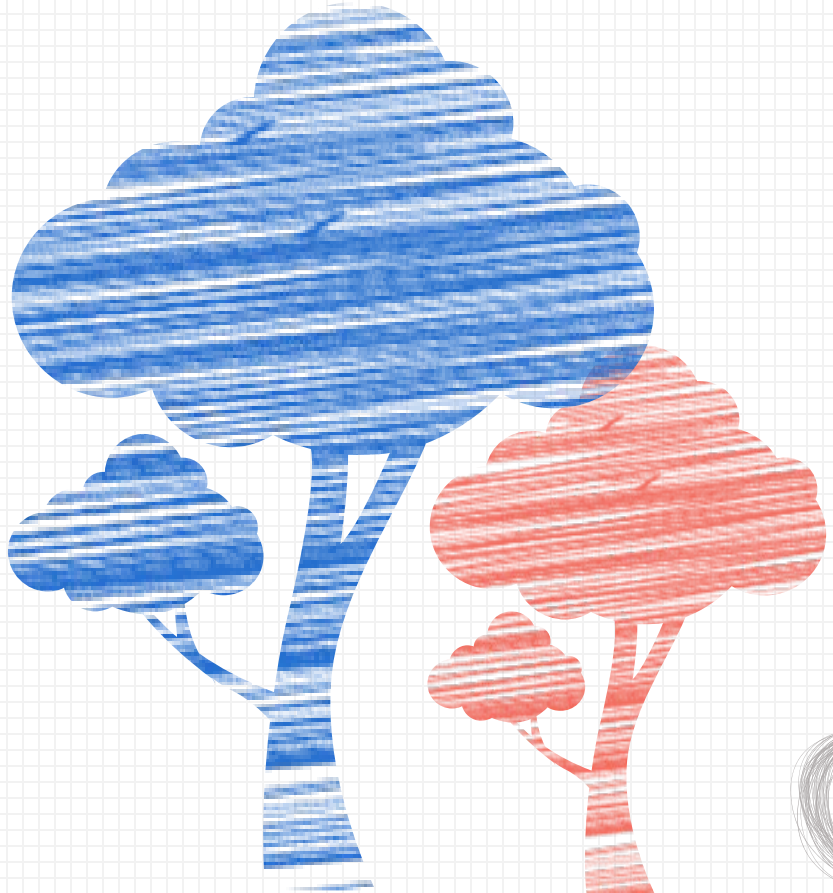


三 基于模型的统计推断

$$Y=\sum_{k\in s}Y_k+\sum_{k\notin s}Y_k$$

$$\hat{Y}_m=\sum_{k\in S}y_k+\hat{\beta}\sum_{k\in U\setminus S}x_k$$

$$\hat{\beta}=\frac{\sum_{k\in S}y_k x_k/v(x_k)}{\sum_{k\in S}x_k^2/v(x_k)}$$



三 基于模型的统计推断

$$\begin{cases} y = X\beta + e \\ \varepsilon(e) = 0, Var(e) = \sigma^2 V \end{cases}$$

$w'y_s$（其中$w = (w_{1s}, \cdots, w_{ns})'$）为线性估计类中的最优线性估计

$$\Leftrightarrow w'X_s = 1_N' X_s, \text{且} V_s w - K1_N \in C(X_s)$$

$$\hat{T}_{MR}(s) = 1_n' y_s + 1_{N-n}' [X_{\bar{s}}\hat{\beta}^* + V_{s\bar{s}} V_s^{-1}(y_s - X_s\hat{\beta}^*)]$$

其中$\hat{\beta}^* = (X_s' V_s^{-1} X_s)^{-1} X_s' V_s^{-1} y_s$是线性估计类中的最优估计。

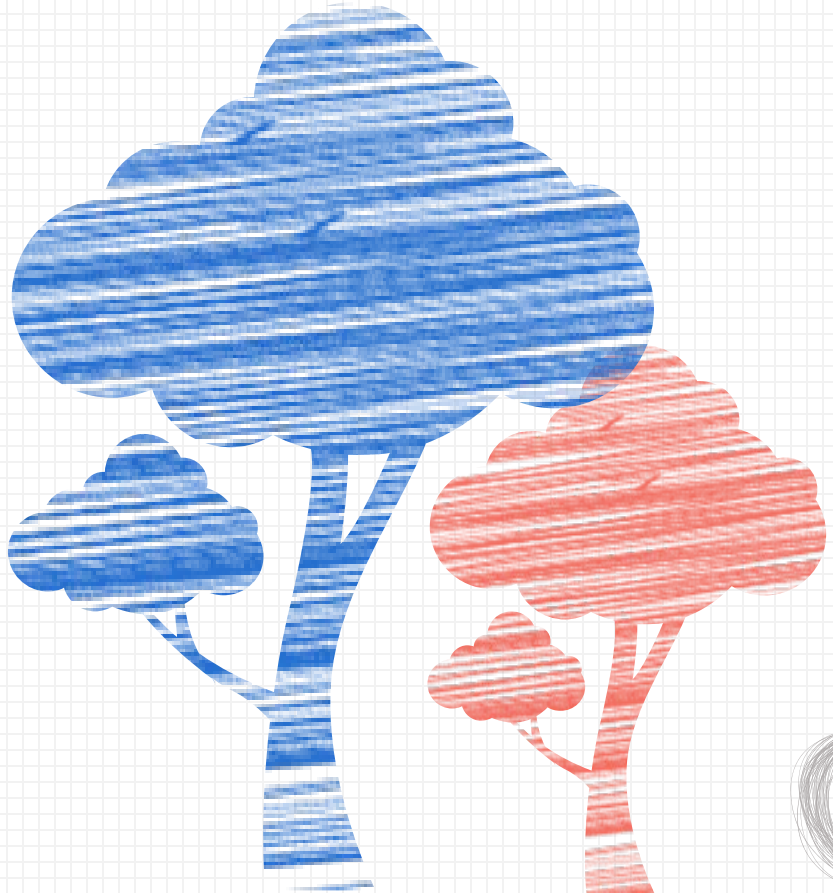
四 模型辅助推断

$$\hat{Y}_{GR} = \hat{B}_0 \sum_{k\in U} x_k + \sum_{k\in U} \hat{\varepsilon}_k$$

$$= \hat{B}_0 \sum_{k\in U} x_k + \sum_{k\in S} \frac{\hat{\varepsilon}_k \Big| B_0 = \hat{B}_0}{\pi_k}$$

$$= \sum_{k\in S} \frac{y_k}{\pi_k} + \hat{B}_0 \left(\sum_{k\in U} x_k - \sum_{k\in S} \frac{x_k}{\pi_k} \right)$$

$$\hat{B}_0 = \frac{\sum_{k\in S} y_k x_k / v(x_k)\pi_k}{\sum_{k\in S} x_k^2 / v(x_k)\pi_k}$$



THANKS