湘潭大学 数学与计算科学学院

分层随机抽样

Stratified Random Sampling

—— 隆浙帆

个人主页：www.longzf.com

引言(*Introduction*)

抽样技术的一个核心目标就是：在一定人力、物力、财力的限制下，去找到一个精度较高的总体特征估计量。

精度通常由估计量的均方误差衡量，均方误差是偏差平方与方差的和，偏差是期望与总体特征之差，所以推导总体特征不同估计量的期望与方差，占据了我们教材的大部分版面。

引言（*Introduction*）

总体特征包括均值、总值、比例、比率。通常，教材只会给出总体均值各种估计量（简单估计、比估计、回归估计）期望与方差的详细推导过程。这是因为：

1. 对特定总体，总值是均值的常数倍，比例是特殊的均值，一旦知道了总体均值估计量的期望与方差，我们就可以很容易地得到总体总值、总体比例估计量的期望与方差；

引言(*Introduction*)

2. 虽然无法通过总体均值估计量的期望与方差直接推得总体比率估计量的期望与方差，但估计总体比率的需求并不常见，因此教材只给出了简单随机抽样下的总体比率估计量的期望与方差。事实上，总体比率估计量的期望与方差结果，更多地承担着中介的角色，因为总体特征的比估计与比率估计量息息相关。

引言（*Introduction*）

下面，我们回顾一下分层随机抽样中，总体均值的简单估计、比估计、回归估计的期望与方差。我们不从简答估计开始看，而从回归估计开始看。这是因为：

与简单随机抽样一致，对分层随机抽样而言，其总体均值的简单估计、比估计也是总体均值回归估计的特例。

回归估计(*Regression Estimator*)

分别回归估计 (separate regression estimator)

$$\bar{y}_{lrs} = \sum_{i=1}^{L} W_h \bar{y}_{lrh} = \sum_{i=1}^{L} W_h \left[\bar{y}_h + \beta_h(\bar{X}_h - \bar{x}_h)\right]$$

联合回归估计 (combined regression estimator)

$$\bar{y}_{lrc} = \bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st}) = \sum_{i=1}^{L} W_h \bar{y}_h + \beta\left(\bar{X} - \sum_{i=1}^{L} W_h \bar{x}_h\right)$$

回归估计（*Regression Estimator*）

β_h (或 β) 事先给定：

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

$$V(\bar{y}_{lrs}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 - 2\beta_h S_{xyh} + \beta_h^2 S_{xh}^2\right)$$

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

$$V(\bar{y}_{lrc}) = \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 - 2\beta S_{xyh} + \beta^2 S_{xh}^2\right)$$

回归估计（*Regression Estimator*）

β_h (或 β) 事先未定，各层样本量 n_h (或样本量 n)较大:

$$E(\bar{y}_{lrs}) \approx \bar{Y}, B_h = \frac{S_{xyh}}{S_{xh}^2}$$

$$V(\bar{y}_{lrs}) \approx \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 - 2B_h S_{xyh} + B_h^2 S_{xh}^2\right)$$

$$E(\bar{y}_{lrc}) \approx \bar{Y}, B_c = \frac{\sum_{h=1}^{L} a_h B_h}{\sum_{h=1}^{L} a_h}, a_h = \frac{W_h^2(1-f_h)S_{xh}^2}{n_h}$$

$$V(\bar{y}_{lrc}) \approx \sum_{h=1}^{L} W_h^2 \frac{1-f_h}{n_h}\left(S_{yh}^2 - 2B_c S_{xyh} + B_c^2 S_{xh}^2\right)$$

三种估计的关系

令 $\beta_h = \bar{y}_h/\bar{x}_h$ (或 $\beta = \bar{y}/\bar{x}$)，根据 β_h (或 β) 事先给定条件下回归估计期望与方差的公式，以及 n_h (或 n) 较大时 $\bar{y}_h/\bar{x}_h \approx \bar{Y}_h/\bar{X}_h = R_h$ (或 $\bar{y}/\bar{x} \approx \bar{Y}/\bar{X} = R$) 的性质，可立即得到分别 (或联合) 比估计的期望与方差公式。

令 $\beta_h = 0$ (或 $\beta = 0$)，根据 β_h (或 β) 事先给定条件下回归估计期望与方差的公式，可立即得到简单估计的期望与方差公式。

三种估计的比较

1. 当各层样本量 n_h 较大、总体 Y 与总体 X 各层变异系数一致 ($S_{yh}/\bar{Y}_h = S_{xh}/\bar{X}_h$) 且各层相关系数大于 1/2 ($\frac{S_{xyh}}{S_{yh}S_{xh}} > 1/2$) 时，分别比估计精度高于简单估计;

2. 当样本量 n 较大、各层相对于总体的变异程度一致 ($S_{yh}/\bar{Y} = S_{xh}/\bar{X}$) 且各层相关系数大于 1/2 ($\frac{S_{xyh}}{S_{yh}S_{xh}} > 1/2$) 时，联合比估计精度高于简单估计;

三种估计的比较

3. 当各层样本量 n_h (或样本量 n) 较大时，分别 (或联合) 回归估计精度不低于简单估计;

4. 当存在某些层样本量 n_h 不大 (或样本量 n 不大) 时，比估计、回归估计既存在偏差，也没有方差的准确或近似公式，此时无法从理论上判断三者精度的关系。随机模拟显示：此时回归估计与比估计方差相差不大，但其均方误差显著大于比估计。

三种估计的比较

5.当各层样本量 n_h 较大、各层比率 $R_h = \frac{s_{xyh}}{s_{xh}^2}$ 时，分别比估计精度不低于联合比估计；

6.当各层样本量 n_h 较大时，分别回归估计精度不低于联合回归估计；

7.当样本量 n 较大，但存在某些层样本量 n_h 不大时，由于此时没有分别比(回归)估计方差的准确或近似公式，以及联合估计无需辅助变量各层特征的便利性，如果要选择比(回归)估计，应优先选择联合比(回归)估计；