关于分合法：大象有几个腰子？

杨正伟
（川北医学院 形态定量研究室，四川 南充市 637007）

假如要抽样估计某大象体内有多少个腰子——肾脏，我们可以这样进行：①把大象任意分成 8 个部分（图 1），从左向右依次用数字 1、2……代表；②任意确定抽样间距，假设确定为 **3**，即等距随机（systematic random）抽取 *1/3* 部分来估计；③在 1~3 之间选择 1 个随机数字，假设选择为 2，然后确定 1~8 之间其余的等距随机数字 5、8；④抽选等距随机数字 2、5 和 8 所代表的大象的头、肢 2 和尾部（图 1），用解剖或扫描等方法确定其中的肾脏数（结果为 **0**）；⑤最后估计该大象体内的肾脏总数：**3** × **0** = 0。即是说，该大象体内肾脏数的无偏估计为 0 个。

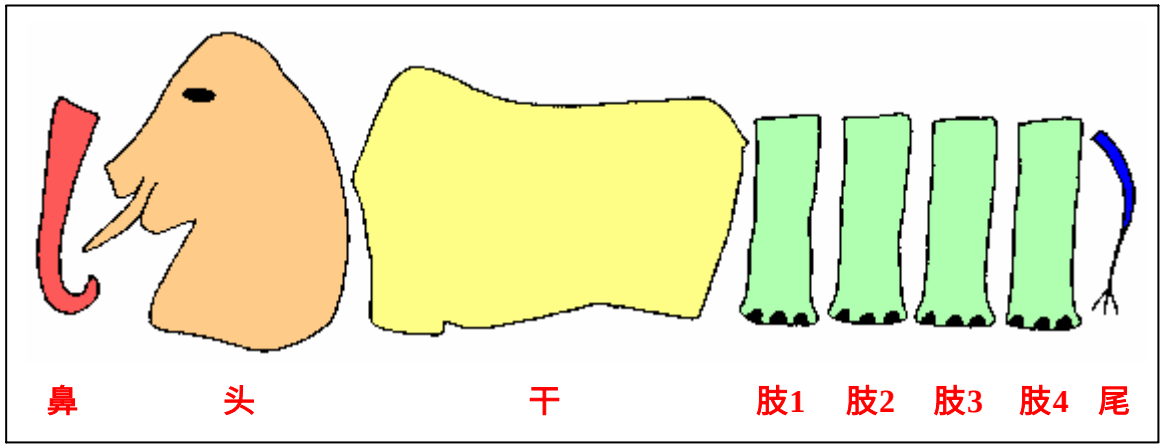

图 1 大象被分成鼻、头等 8 个部分（修饰自 Gundersen 的插图[1]）

以上抽样估计结果是 1 个可能估计结果，用上述方法还同样可能得到另外 2 个估计结果：0、6（当所抽选的等距随机数字组分别为 1、4、7 和 3、6 时；假设该大象只有 2 个肾脏，均在干部内）。因为所有这些可能估计结果的平均值为 2，正好等于该大象体内肾脏数的真值，所以我们说以上任一估计结果都是肾脏数的无偏估计。无偏（unbiased）即无偏差（bias），即所有可能估计结果的平均值或期望值等于真值。无偏性（unbiasedness）是抽样估计的最基本要求；无偏估计可以说就是科学估计，其重要统计特征是：抽样误差（sampling error）总的会随样本含量的增加而减少。换句话讲，“无偏工作”不会“事倍功半”，不会“费力不讨好”。例如，上例中如果把样本含量翻 3 番（即抽取大象的所有 8 个部分）来研究，结果会怎样？

上述利用等距随机抽样的抽样间距来进行总量估计的方法，即 fractionator（分合法）[1]。该法的一般化陈述是：把总体内的所有元素（粒子等）任意分成若干部分，从中等距随机抽取 *1/f*（*f* 为正整数）部分；如果所抽样本内的元素总数为 ***n***，那么总体内所有元素的总数 ***N*** 的无偏估计为：

$$N = f \cdot n$$

分合法于 20 世纪 30 年代提出，但一直被忽视，直到 80 年代被国际著名体视学家 H.J.G. Gundersen（丹麦）再“挖掘”出来[1]后才在体视学领域得到广泛的运用[2]。分合法（原理）是如此的简单、实用，令人不禁感叹“简单就是美”。

例如，为估计小脑内蒲肯野细胞的总数：①先将器官切成平行薄片，然后切成大小相似的组织块，等距随机抽取 $1/f_1$（f_1 为正整数）的组织块；②将所抽选的组织块进行

石蜡包埋，然后完全切成 8 μm 厚的连续切片，等距随机抽取 $1/f_2$（f_2 为正整数）的切片；③计数所抽选切片内的蒲肯野细胞核内的核仁总数 n；④计算小脑内蒲肯野细胞的总数：$f_1 \cdot f_2 \cdot n$。假设每个蒲肯野细胞有且只有 1 个细胞核、1 个核仁，每个核仁能且只能在 1 张切片内被观察到，该项估计就是小脑内蒲肯野细胞总数的无偏估计[3]。

让我们再看秦诗芸（川北医学院外国语言文化系 2004 级本科学生）和笔者的一个研究实例：英语版小说《飘》内的单词总数估计。该书共有 1231 页，每页内的行数不超过 31，每行内的单词数很少超过 15。我们等距随机抽选了 *1/4* 页（第 1、5、9、13、17、21……1225、1229 页），从每页等距随机抽选了 *1/31* 行，从每行等距随机抽选（计数）了 *1/15* 个单词。最后，实际从该书共抽得 308 页，277 行；从这 277 行共抽得（计数）215 个单词。因此，该书中的单词总数的无偏估计为：$(4 \times 31 \times 15 \times 215) = 399900$。（注：如果该研究仅仅是要估计单词总数，可计数所抽取的每行内的所有单词，而不只计数其中 *1/15* 的单词。）

实际运用分合法进行研究时，非常重要的是，要注意通过适当的抽样设计来尽量减少各级抽样中的抽样误差。例如，在切取小脑组织块时切除不包含蒲肯野细胞的髓质[3]；让所切组织块的大小相似以使各组织块内所测细胞的数量相似（假设组织块内所测细胞数量的分布较均匀），或者按组织块的大小依次将组织块排成行后再等距随机抽选[4]。又如，估计单词总数时（上述），因为每页内的行数不超过 31，约 90%的页都有 31 行，所以我们可把页内行的抽样间距确定为 15 或 31，以减少从每页内抽到的行数的变异；如果我们把该间距定为 22 或 50，势必就会加大抽样变异（误差）。

关于分合法估计的抽样误差，一个简单的估计方法是：把所得的 1 个等距随机样本分成 2 个等距子样本（subsample），分别从这 2 个子样本计算估计结果 N_1、N_2，然后用下式估计误差系数（coefficient of error - CE）[5]：

$$\mathrm{CE} = \frac{1}{\sqrt{2}} \cdot \frac{|N_1 - N_2|}{(N_1 + N_2)}$$

例如，估计单词总数时（上述），把所抽选的第 1、9、17……1225 页当作 1 个子样本，把所抽选的第 5、13、21……1229 页当作另 1 个子样本。从这 2 个子样本最后各计数到 108 和 107 个单词，因此，把这 2 个子样本合在一起所得的估计结果（上述）的误差系数为 0.33%（根据上式）。这说明该项估计的抽样误差非常小，该项估计非常准确：根据误差系数（标准误除以均值）的定义，用上述多级抽样方法进行估计可能得到的结果与真值之间的差异，平均只有 0.33%。

参考文献

1. Gundersen HJ. Stereology of arbitrary particles. A review of unbiased number and size estimators and the presentation of some new ones, in memory of William R. Thompson. J Microsc 1986; 143(1): 3-45.
2. Mayhew TM, Gundersen HJ. ‘If you assume, you can make an ass out of u and me’: a decade of the disector for stereological counting of particles in 3D space. J Anat 1996; 188(1): 1-15.
3. 唐勇，杨正伟，崔成虎. 人小脑蒲肯野细胞数目的定量研究. 解剖学杂志 1994; 17(2): 192-3.
4. Gundersen HJ. The smooth fractionator. J Microsc 2002; 207(3): 191-210.
5. Yang Z, Zhang R, Wen X, Huang A. Caveat on the error analysis for stereological estimates. Image Anal Stereol 2000; 19(1): 9-13.