# 谈谈海洋调查研究中的抽样法（Ⅱ）

山东海洋学院　陈敦隆

## 三、分块抽样法

前面，我们已经说过，对于海洋调查研究中的抽样问题来说，总体中所研究指标的方差一般是比较大的。因此，若直接采用上述的随机抽样方法，其所需的样本容量是比较大的。可是过多地增加采样次数，客观上往往又有困难。针对海洋抽样的这些特点，人们自然要问：有沒有别的抽样方法，能使海洋调查收到“事半功倍”的效果呢？我们认为采用大规模抽样理论中的分块抽样法（在抽样方法的书籍中，也有称它为分层抽样法），可能对海洋调查更为有效。所谓分块抽样就是在进行抽样之前，先把抽样总体分成若干块，然后再从每块分别地进行随机抽取。

分块抽样的目的是要提高抽样研究的有效率，因而必须按某种特定的办法进行分块。容易想到，一个好的分块办法应该使各分块内的方差尽可能地小，换言之，应该使块内元素的指标值比较均匀。可以证明：使块内的方差尽可能地小即等价于使块间的方差尽可能地大。为了从量上说明这个问题，我们先来计算总体分块后总体方差的表达式，这个表达式在以后的讨论中还要经常地用到。

设将总体分成$K$块，又对$i=1,2,\cdots\cdots K$，令：$N_i$表示总体内第$i$块中含有的元素个数；

$N=\sum_{i}^{k} N_i$ 表示总体中元素个数；

$ET_i$表示总体内第$i$块中所研究的指标的平均数；

$ET$表示总体中所研究的指标的平均数；

$\sigma^2$表示总体内第$i$块中所研究的指标的方差；

$\sigma^2$表示总体中所研究的指标的方差。这样一来我们可以得到如下的总体方差的表达式：

$$\sigma^2=\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i^2+\sum_{i=1}^{k}\frac{N_i}{N}(ET_i-ET)^2 \qquad (3-1)$$

上式表明：将总体分成若干块后，总体方差可以看作两项之和：第一项是各块内方差的加数平均数，第二项是总体各块间的方差。前者表示块内差异的大小，后者表示块间差异的大小。由于总体方差是固定的，所以分块时分得使块内方差小一些，那么块间的方差就大一些，反之亦然。换言之，就所研究的指标讲，如果分块时使得块内的差异小一点，那么块间的差异就相应地大一点，反之亦然。

上面我们提出了总体分块的准则，但沒有给出具体的分块办法。考虑到系统研究这类问题已有专门的数学分支——聚类分析（它是从数值分类学中分离出来的一个多元统计分析方法的新分支），应用聚类分析的方法不仅能够解决变量（或指标）总体的分类问题，而且更主要的是可以解决多变量总体的分类问题。有关这方面的知识，拟另文介绍，在这里我们也就不作讨论了。

总体分块后，下一步的工作是从每块中分别取样。由于各块内的$N_i$和$\sigma_i^2$不见得都是一样的，所以仍有一些问题需要作进一步的讨论。比如说，对于给定的样本容量$n$，问各块

---

于提高养殖前期紫菜的收获量和质量是有好处的。到养殖后期两种苗源的紫菜产量逐渐接近，有的壳孢子苗绳上紫菜反比单孢子苗有所增高。所以出现这种现象，可能由于经过多次采收的紫菜绳，壳孢子苗长成的紫菜逐渐减少，单孢子苗长成的紫菜逐渐增加，最后全为单孢子苗长成的紫菜。表面看来虽是两种苗源，实际上可能均是单孢子来的，因此紫菜长度的测量结果非常相近，不象养殖前期那样，生长速度明显不同。

内应抽取多大的样品方能提高所得样品的有效性？为了从量上回答这些问题，必须首先算出分块后的总体平均数的表达式。采用上述的记号，容易得到如下的总体平均数的表达式：

$$ET=\sum_{i=1}^{K}\frac{N_i}{N}ET_i \qquad (3-2)$$

上式表明：总体分块后，总体平均数等于各块内平均数的加数平均。

由（3—2）式可以看出，为了用各块中样本的知识来估计总体平均数，我们可取

$$\bar{T}_g=\sum_{i=1}^{K}\frac{N_i}{N}\bar{T}_i \qquad (3-3)$$

作为$ET$的估计，此处，$\bar{T}_i$表示从第$i$块中取出的样本的平均数，下标$g$则专指抽样是按分块方法进行的。

不难证明，$\bar{T}_g$是$ET$的一个无偏估计。事实上，由于$E\bar{T}_i=ET_i$，故由（3—2）和（3—3）式便可得到

$$E\bar{T}_g=ET \qquad (3-4)$$

下面我们来计算$\bar{T}_g$的方差。由（3—2）式和（3—3）式得知$\bar{T}_g$的方差为

$$\sigma^2_{\bar{T}g}=E(\bar{T}_g-ET)^2=E\left\{\sum_{i=1}^{K}\frac{N_i^2}{N^2}(\bar{T}_i-ET_i)^2\right\}$$

因为在每一块中是分别进行取样的，所以利用“相互独立随机变量和的方差等于各随机变量方差之和”这一定理，便有

$$\sigma^2_{\bar{T}g}=\frac{1}{N^2}\sum_{i=L}^{K}N_i^2E(\bar{T}_i-ET_i)^2$$

再利用（1—3）式，即得

$$\sigma^2_{\bar{T}g}=\frac{1}{N^2}\sum_{i=1}^{K}\frac{N_i^2\sigma_i^2}{n_i} \qquad (3-5)$$

此处，$n_i$表示样品中从第$i$块内取出的元素个数。

至此，我们已经导出分块抽样的样本平均数$\bar{T}_g$的方差$\sigma^2_{\bar{T}g}$。同第一节中的$\sigma^2_{\bar{T}}$一样，就估计$ET$而言，$\sigma^2_{\bar{T}g}$越小，由分块抽样所得的代表性（或者说有效性）越大。相反，则越小。

为了区别起见，我们把$\sigma_{\bar{T}g}$称为分块抽样的抽样误差，而把$\sigma_{\bar{T}}$称为不分块抽样的抽样误差。现在人们自然要问：对于给定的样本容量$n$，$\sigma_{\bar{T}}$与$\sigma_{\bar{T}g}$相比哪个大呢？或者说，不分块抽样与分块抽样相比，哪一种抽样方法更有效呢？为了从量上对分块抽样和不分块抽样加以比较，我们先由(1—3)式和(3—1)式写出

$$\sigma^2_{\bar{T}}=\frac{\sigma^2}{n}=\frac{1}{n}\left\{\sum_{i=1}^{K}\frac{N_i}{N}\sigma_i^2+\sum_{i=1}^{K}\frac{N_i}{N}(ET_i-ET)^2\right\} \qquad (3-6)$$

然后再来计算$\sigma^2_{\bar{T}}$与$\sigma^2_{\bar{T}g}$之差。可以证明：

$$\sigma^2_{\bar{T}}-\sigma^2_{\bar{T}g}=\frac{1}{N^2}\sum_{i=1}^{K}N_i\sigma_i^2\left(\frac{N}{n}-\frac{N_i}{n_i}\right)+\frac{1}{n}\sum_{i=1}^{K}\frac{N_i}{N}\left(ET_i-ET\right)^2 \qquad (3-7)$$

上式右端第二项永不为负，且只有在各块内的平均数都相等时才等于零。前面我们已经说过，分块抽样所研究的是各块内平均数彼此不相等的情况。右端第一项可正可负。如果第一项是负的，且它的绝对值大于第二项，那么就有不等式 $\sigma^2_{\bar{T}}-\sigma^2_{\bar{T}g}<0$

在这种情况下，不分块抽样反而比分块抽样好。

当$\frac{N}{n}=\frac{N_i}{n_i}$对任意的$i$都成立时，也就是从任一块中取出的元素个数$n_i$与这块内的元素总个数$N_i$成比例（其比例系数为$\frac{n}{N}$），我们称这种取样的方法为按比例分块抽样法。在这种情况下，（3—7）式右端第一项为零。由此得知，当各块平均数不都相等时，按比例分块抽样比不分块抽样好。

可以证明，若采用按比例分块抽样法，则所得样品平均数的方差为

$$_R\sigma^2_{\bar{T}g}=\frac{1}{n}\sum_{i=1}^{K}\frac{N_i}{N}\sigma_i^2 \qquad (3-8)$$

式中左端方差记号的左下标$R$专指抽样是按比例进行的。事实上，由

$$\frac{N}{n}=\frac{N_i}{n_i}\quad(i=1,\ 2,\ \cdots\cdots,\ K)$$

可得

$$n_i=\frac{n}{N}N_i\quad(i=1,\ 2,\cdots\cdots,K)\qquad(3-9)$$

把(3—9)式代入(3—5)式，便可得(3—8)式。

对比(3—6)与(3—8)两式，容易得到

$$\sigma^2_{\bar{T}}-{}_R\sigma^2_{\bar{T}g}=\frac{1}{n}\sum_{i=1}^{K}\frac{N_i}{N}(ET_i-ET)^2\qquad(3-10)$$

这就是说，同不分块抽样相比，按比例分块抽样的平均数的方差较小，两者的差数恰好是块间平均数的方差的$1/n$（$n$为样本容量）。

按比例分块抽样自然是一种较好的分块抽样法，因为它针对了总体分块后各块内的元素个数$N_i$可能不等的这一事实，提出了一种按(3—9)式对各块中进行抽样的办法。正因为这个缘故，采用按比例分块抽样的结果，提高了抽样研究的有效率〔参见(3—10)〕。但必须指出，按比例取样没有注意到下述另一个事实，即总体分块后各块内的方差$\sigma_i^2$可能是不一样的。前面我们不止一次地说过，为了提高抽样研究的有效率，必须注意到抽样总体的方差，即方差比较大的总体应抽取一个容量比较大的样本；方差比较小的总体抽取一个容量比较小的样本。由于按比例分层抽样在决定各块内的抽样个数时，只是注意到各块内的元素个数$N_i$的差别，而没有注意到各块内方差$\sigma_i^2$的差异，因此可以预料，按比例分块抽样不一定是最好的分块抽样。显然，对于给定的样本容量来说，最有效的分块抽样办法，应该是选择各块内的抽样个数使(3—5)式达到极小。按这个准则而制定的一种分块抽样称为策略分块抽样。

由此可见，寻求策略抽样的办法，在数学上可以归结为求条件极值的问题，即在条件下，

$$\sum_{i=1}^{K}n_i=n\qquad(3-11)$$

寻求使函数

$$\sigma^2_{\bar{T}g}=\frac{1}{N^2}\sum_{i=1}^{K}\frac{N_i^2\sigma_i^2}{n_i}\qquad(3-12)$$

达到最小值的$n_i$（$i=1,2,\cdots\cdots,K$）。

利用拉格朗日乘数法，作函数

$$\Phi=\frac{1}{N^2}\sum_{i=1}^{n}\frac{N_i^2\sigma_i^2}{n_i}+\lambda\left(\sum_{i=1}^{k}n_i-n\right)$$

并让$\Phi$对$n_1,n_2,\cdots\cdots,n_K$的偏导数等于零，得到

$$-\frac{1}{N^2}\frac{N_i^2\sigma_i^2}{n_i^2}+\lambda=0$$

$$i=1,\ 2,\ \cdots\cdots,\ k$$

由此可得

$$n_i=CN_i\sigma_i,\ i=1,\ 2,\cdots\cdots,k\qquad(3-13)$$

此处

$$C=\frac{1}{\sqrt{\lambda}N}$$

把(3—13)式代入(3—11)式便可确定

$$C=\frac{n}{\sum_{i=1}^{k}N_i\sigma_i}\qquad(3-14)$$

最后，把(3—14)式代入(3—13)式中，即得

$$n_i=\frac{n}{\sum_{i=1}^{k}N_i\sigma_i}N_i\sigma_i,$$

$$i=1,\ 2,\ \cdots\cdots,\ k\qquad(3-15)$$

公式(3—14)是非常有趣的。从这个式子中可以清楚地看出，使(3—12)式取得最小值的$n_i$不仅与各块内的元素的总个数有关，而且还与各块内所研究指标的方差有关。这个结论的含义是非常直观的，即那一块内元素的个数比较多，从该块内取出的样品量就应该比较多；与此同时，那一块内所研究指标的标准差比较大，则从该块内取出的样本量也应该比较大。

至此，我们就从量上回答了本节中提出的有关总体分块后如何进行抽样的问题。

在实际工作中，往往是在标准差不知道的情况下进行抽样的。为了确定(3—15)式中$\sigma_i$的值，可以先在各块内抽取一个容量相对地比较小的样本（这是因为总体分块后，$\sigma_i$相对于$\sigma$比较小，因而确定$\sigma_i$所需的样本容量相对于确定$\sigma$而言，是比较小的），并由它算出所研究指标的样品标准差作为$\sigma_i$的近似值。

用（3—15）式的方法决定出的总体中各块抽样个数$n_i$与各块中所研究指标的方差$\sigma_i$有关，因此，对另一种指标来说，抽样研究的有效率便可能减少。通常人们很少只为研究一个指标，而使用一种特定的抽样方法。自然各个不同的指标在各层块内的方差会导致不同的各组$n_i$值。通常我们用下面两种办法之一来解决这个困难。一种是在所研究的指标中取出我们着重研究的一种主要指标，以这种指标为标准按（3—15）式方法进行抽样；而在不易分出主次的情况下可用另一种办法，即在所研究的指标中选出一种与其他指标相关性最好的指标，并以这种指标为标准按（3—15）式抽样。

无疑地，采用策略分块抽样的结果，其所得的抽样误差是最小的，这也就是说，就估计总体平均$ET$而言，它是最有效的。但是如果抽样研究的同时还要估计各块的平均数$ET_i$，那么可能发现在有些块内由（3—15）式所算出的样本个数显得太少了。在容许样本容量扩大时，我们便将样品容量扩大，而各块的$n_i$值也随之扩大，这样一来，也就可以达到既估计$ET$，又估计$ET_i$的目的。但是如果样本容量不可能扩大，而估计$ET_i$却又有重大意义，那么，我们就得放弃策略抽样方法。

为了同按比例分块抽样进行比较，我们来计算一下策略分块抽样的标准误差。显然把（3—15）式代入（3—5），可得

$$ {}_S\sigma^2_{\bar{T}_g} = \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i\right)^2 \qquad (3—16) $$

式中左端方差记号的下标$S$专指抽样是按策略进行的。

对比（3—8）和（3—15）两式，可得

$$ {}_R\sigma^2_{\bar{T}_g} - {}_S\sigma^2_{\bar{T}_g} = \frac{1}{n}\left[\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i{}^2 - \left(\sum_{i=1}^{k}\frac{N_i}{N}\sigma_i\right)\right] $$

在上式中，注意到

$$ \sum_{i=1}^{k}\frac{N_i}{N} = 1 $$

便可得到

$$ {}_R\sigma^2_{\bar{T}_g} - {}_S\sigma^2_{\bar{T}_g} = \frac{1}{n}\left[\sum_{i=1}^{k}\frac{N_i}{N}\left(\sigma_i - \sum_{i=1}^{k}\frac{N_i}{N}\sigma_i\right)^2\right] \qquad (3—17) $$

这就是说，同按比例分块抽样相比较，策略抽样的抽样误差小，两者之差恰好等于块间方差的方差的$1/n$（$n$为样本容量）。

回到第一节中的例子。假定在我们所讨论的海区的表层（这便是我们所讨论的总体），存在着性质彼此不同的$K$个水团（当着所论海区比较大时，这种情况是常见的），这时总体中所研究指标（例如温度、盐度等）的方差是比较大的。因此，在这种情况下，如果我们采用不分块的抽样办法，从整个海区的表层上按简单随机抽样法抽取一个容量不大的样本，那么所得样本的代表性便可能不太好，因而用它的平均数$\bar{T}$来估计整个海区表层的平均温度$ET$可能有较大的误差(参见前两节中的讨论)。但是，如果我们采用分块抽样方法，即在进行抽样之前，首先整个海区表层划分为$K$个水团，每个水团对应于一个分块，然后再从各个水团（或各块）中分别进行抽样，则所得样品的代表性一般要比不分块抽样好。因为总体按水团分块后，各块（或各水团）内所研究指标的方差相对地比较小，或者说各块内元素的指标值比较均匀，因而即使只抽取一个容量不大的样本，其代表性也是比较好的。

划分水团的方法很多，在这里我们暂不作讨论。至于各水团内应抽取多大的样本，本节已经介绍了二种分块抽样概型——按比例分块抽样和策略分块抽样的处理方法。从抽样研究的有效率讲，策略分块抽样比按比例分块抽样好。但是从使用公式上讲，按比例抽样法只要知道各水团内元素的总个数即可，而按策略分块抽样法，除了需要知道各水团内元素的总个数外，还必须知道各水团内所研究指标（如温度、盐度等）的方差。在我们的例子里，决定各水团内元素的总个数是比较容易的，因为尽

（下转第18页）

表 2　三疣梭子蟹的适宜光色

| 照度区 (Lux) | 0.001—0.01 | 0.01—0.1 | 0.1—1.0 | 1.0—10 | 10—100 |
|---|---|---|---|---|---|
| 适宜光色 | 紫（白） | 紫 | 红 | 黄（橙） | 绿（蓝） |

光色时，也应注明相应的光照强度才行。

（三）由上述讨论可知，三疣梭子蟹对颜色光发生最大的趋光反应，必须以亮度和光色的最佳匹配为条件。当获得最佳匹配时，视系统才可处于该环境下的最高兴奋状态。本实验所得的较佳匹配的照度（适宜照度区）如表 3 所示。

表 3　三疣梭子蟹对颜色光反应的适宜照度区

| 光　色 | 红 | 橙 | 黄 | 绿 |
|---|---|---|---|---|
| 适宜照度区 (Lux) | 0.1—1.0 | 1.0—10 | 1.0—10 | 1.0—10 |

| 光　色 | 蓝 | 紫 | 白 |
|---|---|---|---|
| 适宜照度区 (Lux) | 0.01—0.1 | 0.01—0.1 | 0.001—0.01 |

表 3 表明，对橙、黄和绿光1.0—10Lux为适宜照度区。蓝、紫光的适宜照度区为0.01—0.1Lux，这两种光在海水中的透过率较大，这与此蟹所栖息底层的生态环境正相一致。白光是一种混合光，到达底层水时，也只剩下短波光。由此可见，实验结果与自然状态下的情况基本相符。

（四）综上可知，用色灯诱捕三疣梭子蟹是可能的。但在诱与集的作业中，须考虑其亮度与光色的最佳匹配，而宜以光色为主。据此，建议用绿或黄光诱蟹，而用红或紫光集蟹。同样，若以灯诱与笼捕或灯诱与拖网相结合，则亦可考虑如同用弱白光的方法，色灯可置于笼内也可组成“灯链”（集蟹时向网具处逐一熄灭，使蟹向网具或鱼泵处聚集）。灯色以绿或紫为好（因水下灯黄，红光传播不远，诱蟹范围受限制）。

## 参 考 文 献


〔1〕俞大剑、何大仁、郑玉水，1978。厦门大学学报（自然科学版），4：1—13。

〔2〕蔡浩然、马万禄，1978。视觉的分子生理学基础，科学出版社。

〔3〕草下孝也，1959。日本水产学会志，25(1)：17—21。

〔4〕Brown, M. E., 1957. The Physiology of fishes, Vol. Ⅱ. Academic Press inc. (中译本，1963，科学出版社，136—142).

〔5〕Dingle, H., 1962. Am. Naturalist, 96, 151—159.

〔6〕Борисов, П. Г., 1955. Тр. совещания По Вопросам поведения и разведки рыб. Москва, Изд. Анссср, 121—143.

〔7〕Воронцов, Д. С., 1961. Общая Электрофизиология, Государственное издательство Медицискоя литературы, Медгиз, Москва (中译本，1964，人民卫生出版社，118—124).

〔8〕Протасов, В. Р., 1968. Зрение и Ближняя Ориентация рыб. Изд. “Наука”, Москва.


（上接第24页）

管我们无法数出各水团内元素的总个数，但它们的比值却可用各水团的面积比例来代替。因此，对我们的例子来说，为要应用以上已得的结果。只要把本节所有公式的$N_i$和$N$分别换为各水团的面积和整个海区的表层面积即可。至于各水团内所研究指标（如温度）的方差，一般是无法预先知道的。当然，我们可用各水团内的样本方差来近似地代替，但事先总要进行预备抽样才能做到。因此，在实际抽样时，到底是选用按比例分块抽样法，还是采用策略分块抽样法，必须根据海洋调查和研究的目的及具体情况决定。