谈谈海洋调查研究中的抽样法（Ⅰ）

山东海洋学院　陈敦隆

从海洋中抽样，其主要的目的是为海洋科学研究提供各种资料（或称数据）。没有有效的（或者说，没有有代表性的）海洋资料，就谈不到对海洋科研和成果的验证与应用。因此，开展海洋中抽样方法的讨论，自然就成为广大海洋工作者所关心的问题。归纳起来，主要的问题有二：一是根据某项科学研究的目的（比如说为了估计某海区的底栖生物量等等）应进行多少次抽样方能保证所得资料具有足够的代表性？二是根据海洋中抽样的特点，应选用什么样的方法，方能达到“事半功倍”的效果？据笔者所知，为了解决上述问题，在我国的海洋调查工作中已有许多行之有效的方法。然而为了加速我国海洋事业的现代化，有关海洋调查中的抽样问题仍需要从量上加以研究。

本文的主要目的在于探讨在海洋调查中可能有广泛应用的一些新的抽样方法，特别是其中的大规模抽样方法。因限于篇幅，文中仅拟对有关方法作扼要的介绍和讨论，而不作具体的实例计算，希望对此有兴趣的读者来完成这一部份有意义的工作。

一、抽样方法的综述

众所周知，抽样是一种研究手段，其目的是通过适当的方法从研究对象的总体中抽取一定数量的样品来进行分析研究，从而推断研究对象总体的实际情况。为了说明这个问题，让我们来看下面的一个简单例子。

假定我们的研究目的在于估计某海区表层海水的平均温度。显然，在实际海洋调查中，人们不可能对这个海区表层上的所有地点都进行观测（即抽样），而只能在预先选定的测站上逐点测量表层海水的温度，并进行平均运算，，从而把海水的平均温度估计出来。

应该指出，我们在这里仅以估算某海区表层为例子来进行讨论，但从原则上讲，下面讨论的内容也适用于估计任何指标（或称随机变量）的各种特征数（或称参数）。比如说，可以用来估计断面或任何指定时间区间内（如一季或一年）海水的平均温度，也可以用来估计指定测点上平均波高和单位面积水柱内的含沙量或生物量，此外还可以用来估计水温、盐度和氯度的方差以及它们之间的相关系数等等。

在数理统计学中，我们把研究对象的全体统称为总体，更确切地说，我们把所研究的全部元素的指标值所组成的集合称为总体。样本是总体的一部分，或更确切点说，样本是由总体中一定数目的元素的指标值所组成的一个序列。样本中含有的元素的个数称为样品的容量。

必须指出，样本具有两重性，即在取样之前，我们并不知道它是由总体中哪一部分的元素所组成的，但是在取样之后，自然就知道了。为了区别起见，我们把前者（即随抽样机会而定的样本）称为随机样本，而把后者称为样本值。由此可见，样本值只不过是随机样本的一个可能值。

抽样的目的既然是用样本的知识来推断总体在这方面的情况，因此，推断的正确与否便取决于样本在整个总体中代表性的大小。正因为这个缘故，人们自然要问：用什么样的取样办法才能获得有代表性的样本呢？这自然是抽样研究的中心课题之一。

常用的抽样方法是所谓随机抽样。采用随机抽样方法从总体中抽取一个大容量的样本，其所得结果无疑是好的，也就是说，大样本的代表性是好的，相反，如果只抽取一个小容量的样本，其所得的结果可能并不太好，也就是说，小样本不能很好的代表原来的总体。

为了从量上说明上述问题，必须首先搞清

随机抽样的数学含义。所谓随机抽样的方法是指在每一次抽样时，总体中各个元素被取出的概率（也称机率）都相等，但这并不要求在整个抽样过程中各个元素被抽出的概率都是一样的。如果后者成立，则称这种抽样方法为简单的。由简单抽样方法所得的样本称为简单样本。在数理统计学中，几乎所有的推断方法只对简单样本有效。

对于海洋中的抽样来说，几乎所有的总体都是无穷大的，（即包含有无限多个元素）。所以，如果我们采用随机抽样方法，则几乎所有的样品都是简单样本。在本节和下一节里，我们均假定所有的样本都是简单样品。

在数理统计学中已有不少的办法来检验一种抽样方法是否是随机的（但这些检验方法并不都是很简单的），比如说，利用X^2—检验法或游程理论等。因限于篇幅，这里不拟对此作具体讨论，而在实用上，利用现成的随机数表即可方便的从总体中抽取出简单样本来。

下面我们用前述例子来说明简单样本的代表性问题。在这个例子里，总体便是我们所讨论海区表层上所有点的温度值，我们把它记为T。而样本就是由这个总体中抽取出的一定数目的温度值。假定样本的容量为n，我们把它记为（T_1、T_2、……T_n）。现在我们的问题是用这个样本来估计总体T的平均数，并把它记为ET。由数理统计学中的估计理论得知，我们可取样本（T_1、T_2、……T_n）的平均数作为总体平均数ET的估计值。不难证明，样本平均数是总体平均数的一个无偏估计。习惯上，我们把样本（T_1、T_2、……、T_n）的平均数记为$\overline{T}$，它的计算公式如下：

$$\overline{T}=\frac{\sum_{i=1}^{n}T_i}{n} \qquad (1-1)$$

就估计ET而言，如果样本的代表性比较好，则$\overline{T}$与ET的差异应该比较少。现在的问题是什么样的指标来表示这种差异呢？由（1—1）式可以看出：对于给定的一个样本值，$\overline{T}$是一个常数，因而差异也是一个常数。而对随机样本来说，$\overline{T}$是一个随机变量，因而差异也是一个随机变量，或者说是随机差异。表示随机差异的指标常用样本平均数$\overline{T}$的标准差，即平均差异平方的平方根，简称均方根差，用数学的式子来表示，就是

$$\sigma_{\overline{T}}=\sqrt{E\ (\overline{T}-E\ T)^2} \qquad (1-2)$$

在这里，为明确起见，还采用下标$\overline{T}$来表明这是平均温度的均方根差，在抽样方法中，通常把$\sigma_{\overline{T}}$称为抽样误差。

由此可见，就估计ET而言，$\sigma_{\overline{T}}$越小，样品代表性也就越大。反之，则越小。

同样地，在比较不同抽样误差作为依据，我们也应该以随机样本的抽样误差作为依据，即$\sigma_{\overline{T}}$越大，抽样的有效率也就越低，反之，则越高。

对于随机抽样来说，由于随机简单样本中所有元素都是相互独立的随机变量，并且它们具有和总体一样的分布，故由概率论中“相互独立随机变量和的方差等于各随机变量方差之和”这个定理，我们容易由（1—2）式得出：

$$\sigma_{\overline{T}}=\frac{\sigma}{\sqrt{n}} \qquad (1-3)$$

式中，σ为总体中所研究指标的标准差（简称为总体的标准差），而n为随机样本的容量。

由（1—3）式可以清楚看出：简单随机样本的抽样误差$\sigma_{\overline{T}}$与总体的标准差σ成正比，而与样本容量n的平方根值$\sqrt{n}$成反比。因此若采用随机抽样的方法，不管总体的标准差σ如何（但必须有限），只要样本容量n足够大，$\sigma_{\overline{T}}$便可足够的小，特别当$n\to\infty$时，$\sigma_{\overline{T}}\to 0$。这就是说，大样本的代表性无疑是好的。相反，如果样本的容量n比较小，则当总体标准差σ不是太小时，$\sigma_{\overline{T}}$是不会很小的，这就是说，小样本的代表性可能不太好。

对于海洋调查中所遇到的抽样问题，一般具有下面两个特点：（1）由于海洋调查的范围一般是比较大的（尤其是大规模的海洋普查更是如此），再加上海洋中各种要素（也就是

上面所说的指标）易受各种外界因素的影响，所以总体中所研究指标的标准差一般都是比较大的。（2）在一般情况下，由于调查工具的限制，过多的要求增加样本容量，是不可取的。这样做不仅会造成人力和物力的浪费，而且有时甚至是无法实现的。因此，如果我们在海洋调查中，千篇一律地采用随机抽样的方法（目前主要是采用这种方法）那末所得样本的代表性一般是不很理想的。大量的实例也证实了这一点。正因为这个缘故，人们自然要问：在海洋调查中，我们能否采用别的一些更有效的抽样方法来取得更有代表性的样本呢？在本文的最后一节里，我们将专门地探索这个问题。

二、决定样本容量大小的方法

在讨论决定样本大小的方法之前，必须首先明确地指出，不同的抽样方法，其所需的样本容量是不一样的，即使是用同一种抽样方法，但由于样本的使用目的不同，也会有不同的决定样本大小的计算公式，在本节里，为了讨论确定起见，我们只讨论用随机抽样的方法，而样本的使用目的，则是为了估计总体的平均数。

下面，我们将给出几种决定样本大小的方法。

我们已经知道，对于随机抽样来说，样本的容量越大，抽样误差便越小，换言之，大样本的估计精度高。因此，我们可以通过增加样本的容量来提高估计的精度。为了保证抽样误差达到一定的精度要求（比如小于0.01C，这叫绝对精度；或小于0.1ET，这叫相对精度），利用（1—3）式可以算出至少需要多大的样本容量，比如说，为了保证样本估计的精度不低于一个预先规定的小正数η，即

$$\sigma_{\bar{T}}<\eta \qquad (2—1)$$

则由（1—3）式可以算出相应的样本容量至少应为

$$n=\left(\frac{\sigma}{\eta}\right)^2 \qquad (2—2)$$

上式表明：在（2—1）式的精度要求下，至少需要的样本容量与总体方差σ^2成正比，而与η^2成反比。由此可见，精度要求越高，所需的样本容量越大。为了从（2—2）式决定至少需要的样本容量，必须知道σ与η。η的大小自然要由研究工作者依据所研究问题的重要性来定，至于σ，一般是不知道的。为了确定σ的值，可从总体中抽取一个容量比较大的样本，并由它所算出的标准差作为σ的近似值。在数理统计学中，我们常把样品（T_1、T_2、……、T_n）的标准差记为S，它的计算公式如下：

$$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i-\overline{T})^2} \qquad (2—3)$$

上面我们从估计的精度出发（即要求抽样误差小于一个预先指定的正小数η），给出了一种决定样品容量大小的方法。但从实用的观点来说，人们所关心的有时不是抽样误差（也就是平均平方误差的平方根），而是绝对误差。实际工作者的习惯提法是：对于一个具体的研究总体，相应取多大的样本容量，才能保证绝对误差（有时也采用相对误差）不超过一个预先给定的正小数ϵ。由于我们是在抽样之前决定所需的样本容量，所以我们并不能预先知道在每一次抽样中所得的样本是由总体中的哪些元素组成的，简言之，我们只能谈论随机样本，因而由随机样本所得的平均数$\overline{T}$自然是随机变量了。$\overline{T}$既是随机变量，我们就不能要求在每次抽样中，$\overline{T}$与ET的绝对误差都小于ϵ，而只能在概率的意义下加以讨论。由概率论的知识可以证明：对于给定的ϵ，样本的容量n越大，样本平均数$\overline{T}$与总体平均数ET之差的绝对值小于ϵ的概率也就越大。采用概率中的记号来表示就是$P\{|\overline{T}-ET|<\epsilon\}$越大。因此，对于给定的误差范围$\epsilon$和置信概率$P$，我们可以通过下式来算出至少需要的样本容量。

$$P\{|\overline{T}-ET|<\epsilon\}\geqslant P \qquad (2—4)$$

显然，一个好的估计应该是ϵ取值较小而P值较大。在实际工作中，应取ϵ和P等于多少才能保证研究工作的需要，这个问题的讨论一般属于海洋调查的规范问题。ϵ是估计时所容许的最大误差，而P便是置信度的大小。在应用

科学里，常把P称为可靠性概率。一般常取P等于95%。在这种情况下，由（2—4）式算出至少需要的样本容量n能够保证有95%的概率达到容许的误差范围（即绝对误差小于ϵ）。换句话说，如果每次从总体中随机地抽出由(2—4）式算出的容量大小的样本，那么在一百次抽样中，大约有九十五次达到容许的误差范围。利用概率论中实际推断原理（即小概率事件在单次试验中不可能发生的原则），我们便可以说，当P很大时（比如说$P=95\%$），由(2—4)式所决定的样本容量，在单次抽样中，其所得的样本平均数几乎达到容许的误差范围。当然，这里并不排除个别次抽样的结果超出容许的误差范围之外的情况，但出现这种情况的概率是比较小的，它等于$1-P$（比如说，当取$P=95\%$时$1-P=5\%$）在数理统计学中，我们通常把$1-P$称为危险率或风险率（在有的书中则把它称为信度），并记为α。

下面，我们就来说明如何由（2—4）式来决定至少需要的样本容量。为了叙述方便起见，我们将针对几种不同的情况进行讨论。

当我们不知道总体的分布类型时，例如在上一节的例子里，如果我们不知道所讨论海区内表层水温下服从哪一种分布，那么，在这种情形下，我们可以利用概率论中著名的切贝雪夫不等式来决定所需样本容量。事实上，利用这个不等式。我们有

$$P\left\{|\overline{T}-ET|<\epsilon\right\}\geqslant 1-\frac{\sigma^2}{n\epsilon^2} \qquad (2—5)$$

此处：n为样本容量，σ^2为总体的方差。

对比（2—4）和（2—5）两式，我们不难看出：只要样本容量n满足下式

$$1-\frac{\sigma^2}{n\epsilon^2}\geqslant P \qquad (2—6)$$

即可。由此即得至少所需的样品容量为

$$n=\frac{\sigma^2}{\epsilon^2(1-P)} \qquad (2—7)$$

或

$$n=\frac{\sigma^2}{\epsilon^2\alpha} \qquad (2—8)$$

上式表明：在给定的ϵ和P下，其所需的样本容量n与σ^2成正比，而与ϵ^2（$1—P$）成反比。同前面一样，当着我们不知道总体方差σ^2时，一般可用样品的方差S^2来近似地代替。

当着我们知道总体的分布类型时，例如在上一节的例子里，如果我们事先知道所讨论海区的表层水温服从正态分布，那么，在这种情形下，对于给定的置信概率P，利用正态分布的性质，我们有

$$P\left\{|\overline{T}-ET|<u\frac{\sigma}{\sqrt{n}}\right\}=P \qquad (2—9)$$

此处：n为样品容量，σ为总体标准差，u由下式

$$\frac{2}{\sqrt{2\pi}}\int_o^u e^{-\frac{x^2}{2}}dx=P$$

决定。对于给定的置信概率P，u的值可以直接由正态分布表中查出。比如说，当$P=95\%$时，由正态分布表中可查出$u=1.96$。

对比（2—4）和（2—9）两式，我们不难看出：只要样本容量n满足下式

$$u\frac{\sigma}{\sqrt{n}}\leqslant\epsilon \qquad (2—10)$$

即可。由此即得至少所需的样本容量为

$$n=\left(\frac{u\sigma}{\epsilon}\right)^2 \qquad (2—11)$$

上式表明：在给定的ϵ和P下，如果我们知道总体的分布服从正态分布，则至少所需的样本容量n与$(u\sigma)^2$成正比，而与ϵ^2成反比。

对比（2—7）式和（2—11）可知，由它们决定的样本容量往往是不一样的，举个例子来说，若取$P=95\%$，由（2—7）式得所需样品容量为

$$n_1=\frac{20\sigma^2}{\epsilon^2}$$

而由（2—11）式则得所需的样品容量为

$$n_2=\left(\frac{1.96\sigma}{\epsilon}\right)^2$$

两式相除，得

$$\frac{n_1}{n_2}=\frac{20}{(1.96)^2}=5.2$$

这就是说，n_1比n_2大五倍之多。因此，在决定所需样品容量时，我们应当尽可能地利用总体分布的知识。

由于正态分布是自然界中常见的一种分布，所以有必要对（2—10）式作进一步的说明。在总体服从正态分布的情况下，为了决定所需的样本容量，必须首先知道ϵ，u和σ，或者说，必须首先知道ϵ、p和σ。ϵ和p的值自然应该由实际工作者依据所研究问题的重要性来定（比如说，取$\epsilon=0.01$，$p=95\%$，或者说，取$\epsilon=0.01$，$u=1.96$），而σ一般是不知道的，为了确定σ，常用样本的标准差S来代替。但必须指出，如果样品的容量n比较小（比如说，$n<50$），则S与σ可能有相当大的差异。因此，人们要问：在已取得小样本的情况下，怎样来决定所需的样本容量呢？为了回答这一问题，需要利用数理统计学中小样本分布理论。在总体方差不知道的情况下，利用数理统计学中著名的t—分布便可以在一定置信概率下对正态总体的平均数进行区间估计，用数学的式子来表示，就有

$$p\left\{\left|\overline{T}-ET\right|<t\frac{S}{\sqrt{n-1}}\right\}=p \qquad (2—12)$$

这里，p为置信概率，n为样品容量（在统计学中，我们把样品容量减1（即$n-1$）称为t分布的自由度），S由（2—3）式给出，而t由下式①

$$\frac{2}{\sqrt{n-1}\,B\left(\frac{1}{2},\frac{n-1}{2}\right)}\int_0^t\frac{1}{\left(1+\frac{t^2}{n-1}\right)^{\frac{n}{2}}}dt=p$$

①式中，$B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}dx$ $(q>0,\ p>0)$

决定（在统计学中，我们称t为置信限）。对于给定的置信概率和自由度数$n-1$，t的值可以直接由现成的t—分布表中查出。比如说，当$n=10$，$p=99\%$时，则从自由度为9的t—分布表中查出$t=3.25$。

对比（2—4）和（2—12）两式，我们不难看出，只要样本容量n满足下式

$$t\frac{S}{\sqrt{n-1}}<\epsilon \qquad (2—13)$$

即可。

为了从（2—13）式中决定至少需要的样本容量n，必须首先知道t和S，但是t和S又与n有关。因此，在这种情况下，用什么方法来决定（2—13）式中的n呢？作者建议采用如下的逐步逼近法：首先从总体中抽取一个容量比较小的样本，设其样本的大小为n_0，从中计算出S来，再从自由度为(n_0-1)的t分布表查出对应于置信率为p的t值，最后把所得的S和t代入（2—13）式中，定出满足（2—13）式的最小n，我们把它记为n_1，如果$n_1<n_0$，则n_0即可看作是所需的样本容量；如果$n_1>n_0$，则继续从总体中取样直到样品容量是n_1为止，重复以上步骤，定出n_2。如果$n_2<n_1$，则n_1可作为所需的样本容量；如果$n_2>n_1$，则仍须继续重复下去，通常只须重复几次便可求出所需的样本容量来。（待续）

名词解释

近岸波

海浪传至近岸后形成近岸波。由于产生折射、绕射和反射，近岸波十分复杂，至今对其未能深入的了解。海浪传到近岸海区，波速减小，波长变短，波向常发生折射，并趋向于与等深线垂直。在近岸，波高也发生变化。在波向线聚集区，波高增大；在散开区，波高减小。此外，波剖面不断变形，对于斜度较大的海底，波峰的前侧逐渐变陡，后侧逐渐变平缓，直至前侧变成铅直并向前卷倒，发生破碎。如遇障碍物，海浪可发生绕射现象，而传到障碍物的背面，但波高减小。近岸波的研究对于港口的建设有重要作用。（石　谋）