生物分类学统计中几个常用特征数字的求取方法（二）

张 伟 权
（中国科学院海洋研究所）

二、标 准 差

1．求取标准差的一般公式

利用算术平均数作为样品的代表，进行统计分析，是目前分类学中探索变异规律时经常采用的最基本的方法之一。优点是算法简便；在相同的环境条件下受取样的影响较小；加上平均数本身是直接由样品内各个变量*求得的，因此具有普遍的代表性。如果分析用的样品内部各变量之间的差异不大，那末用平均数来反映群体（即总体）的特征就比较合适。然而，单靠平均数有时是不够的，这是因为平均数虽然能够反映样品内变量分布的集中性，但是对变量分布的总貌则无法表示，因此单独依靠平均数来刻划和推估总体，有时会导致严重错误。例如有两个样品，各有10个标本组成，其体长的平均数都是50（毫米）。如果只从平均数来看，它们似乎是十分一致的，但是仔细研究一下这两个样品的变量分布，情况就完全不同了（见表1）。

表（1）

样品甲	47	48	49	50	50	50	50	51	52	53	$\overline{X}_1=50$
样品乙	2	7	20	25	36	64	75	80	93	98	$\overline{X}_2=50$

从表中可以看出，样品甲内各变量的波动不大，个体间的差别程度很小，其分布都集中在平均数$\overline{X}_1$的周围（集中程度大），因此平均数的代表意义就大；相反，样品乙的平均数$\overline{X}_2$尽管也是50，但是变量的波动很大，个体间的差异十分明显，各变量与平均数的关系也较疏远（离散程度大），因此用平均数来表示样品特征的可靠性就差。

为了更加全面地考察和描述一个样品，最好是在对变量集中性分析的同时，能够对样品内部的离散程度进行分析。一种最简单的方法是利用极差。极差是指样品中出现的最大数据与最小数据之间的差值。例如某一样品中，最大数据为100（厘米），最小数据为25（厘米），则100－25＝75（厘米）即为该样品的极差。如果极差大，则说明样品内变量的差异程度大，也就是变异的离散程度大；极差越小，则样品内变量之间的差异程度越小，变异比较集中。

利用极差作为衡量差异程度的标准有两个缺点：

①涉及的变量太少，未能充分利用样品给出的全部信息；

②与平均数不发生直接关系。因此反映实际情况的精确程度较差。

目前被普遍接受而用以确定样品变异程度的最好方法是计算标准差（Standard deviation）。样品标准差记为S．D。其定义为离均差（$Xi-\overline{X}$）平方的总和（Σ）被样品内个体数（N）除的平方根。换言之，标准差乃是离均差的量度。上述定义用公式表示为：

$$S.D=\sqrt{\frac{\Sigma(Xi-\overline{X})^2}{N}} \qquad (1)$$

式中：Xi 为样品内的变量
$\overline{X}$ 为样品平均数

* 本文所提变量一词，意思是指样品内各个个体的测量数据。

N 为样品值（即样品所包含的个体数）

$(Xi-\overline{X})$ 称为样品内各变量的离均差

$(Xi-\overline{X})^2$ 为离均差的平方

$\Sigma(Xi-\overline{X})^2$ 是离均差平方总和

而 $\Sigma(Xi-\overline{X})^2/N$ 即所谓方差（variance），记为 S^2。

由于离均差的总和永远为“0”（见前平均数一节）即 $\Sigma(Xi-\overline{X}=0$，不能反映差异的大小，因此式中把它进行平方。这样，既能克服离均差总和无法反映差异的缺点，又能明显地加大差值，增加了度量变异程度的灵敏性。式中分母部分用样品内变量总数 N 相除，目的是要消除不同样品出现的个数差异，便于进行样品间的互相比较。将公式右边加用平方根号的目的是因为离均差经过平方后，原来的度量单位（如厘米、克等）也都相应地变成了平方，因此必须用开平方的方法将其复原。

由于标准差 $S.D$ 的求得，是以样品内各个变量与平均数（$\overline{X}$）之差异程度的比较为标准的，因此代表性强，上述公式显然是比较理想的。但是还必须指出，这个公式在理论上还有一定的缺陷，这是因为在求取标准差的过程中，样品内的每个变量都分别与平均数进行了比较。平均数 $\overline{X}$ 是从每个变量中各取 $1/N$ 而组成的。当各变量分别与平均数比较的同时，无形中也包括了与自己的 $1/N$ 个相比较。也就是说，公式中表面上虽然有 N 个变量的比较，实质上仅相当于 $N-1$ 个独立比较，结果必然导致 $S.D$ 要比实际总体偏低的弊端。

为了尽可能地使样品反映的情况更合理，上述求取 $S.D$ 的公式必须作部分修正，亦即将分母部分的 N 改成 $N-1$*。

$$S.D=\sqrt{\Sigma(Xi-\overline{X})^2/(N-1)} \qquad (2)$$

经过上述修改后的公式（2）与前一个公式（1）相比，更能反映实际，特别是在样品较小的情况下（15个变量以内的小样品中）尤为实用，建议普遍采用。

2，标准差的计算

现在，以表（1）为例，对公式（2）进行实际运算。

设 Xi、Yi 为甲、乙样品内的变量。$\overline{X}$，$\overline{Y}$ 分别代表它们的平均数。

根据公式的要求首先作表。见表（2）。

在完成上述工作后把表中给出的相应数字代入公式（2），得：

样品甲 $S.D_x=\sqrt{\Sigma(Xi-\overline{X})^2/(N-1)}$

$=\sqrt{28/9}=\sqrt{3.11}=1.76$

样品乙 $S.D_y=\sqrt{\Sigma(Yi-\overline{Y})^2/(N-1)}$

$=\sqrt{11748/9}=\sqrt{1305.33}=36.12$

计算表明，这两个样品的标准差分别为1.76和36.12。

在进行上述简单样品计算的同时，有一点要引起注意，即离均差的总和必然为“0”（见平均数的特征）。因此表内第2和第6栏的底数（即 $\Sigma(Xi-\overline{X})$ 和 $\Sigma(Yi-\overline{Y})$）也必须为“0”，平均数的这个特征，可以用来校对计算有无错误**。

以上所提，是求取标准差的简单例子。如果变量已经作了频数统计，那末求出标准差时，公式（2）还需要作相应的修改：

$$S.D=\sqrt{\frac{\Sigma(Xi-\overline{X})^2fi}{N-1}} \qquad (3)$$

式中 fi 代表不同变量出现的频数。

例如海筒螅Tubularia marina Torrey 的一个

---

* 这里的N-1在统计学上称为自由度。所谓自由度，就是样品中可以不受约束而自由变动的变量数目。例如，某一样品包括5个变量，在平均数和总数已经肯定的情况下，这5个变量中有4个可以自由变动，而剩下的一个就不能变动了，否则将影响平均数和总数的改变。因此我们把样品中那些可以随意变动的变量的数目称为该样品的自由度。

** 利用离均差的总和 $\Sigma(-xi-\overline{X})=0$ 校对计算有无错误，适用于变量未经处理的样品（见表2），如果表格内的变量已作频数统计（见表4），或者分组统计（表5）时，则上述校对应分别改为 $\Sigma=(xi-\overline{X})fi=0$ 和 $\Sigma(xi-\overline{X})fi=0$

表（2）

Xi（甲样品的变量分布）	$Xi-\overline{X}$（离均差）	$(Xi-\overline{X})^2$（离均差平方）	备　注	Yi（乙样品的变量分布）	$Yi-\overline{Y}$（离均差）	$(Yi-\overline{Y})^2$（离均差平方）	备　注
47	－3	9		2	－48	2,304	
48	－2	4		7	－43	1,849	
49	－1	1	$\overline{X}=50$	20	－30	900	$\overline{Y}=50$
50	0	0		25	－25	625	
50	0	0	N＝10	36	－14	196	
50	0	0		64	14	196	N＝10
50	0	0		75	25	625	
51	1	1		80	30	900	
52	2	4		83	43	1,849	
53	3	9		98	48	2,304	
Σ（综和）	$\Sigma(Xi-\overline{X})=0$	$\Sigma(Xi-\overline{X})^2=28$		Σ（综和）	$\Sigma(Yi-\overline{Y})=0$	$\Sigma(Yi-\overline{Y})^2=11,748$	

样品（例五），由50个变量组成。其水螅体的茎高（单位毫米）已经整理成频数表（3）的形式。试求其标准差。

表（3）

xi	27	30	33	35	36	40	44	50	$\overline{X}=3468$
fi	3	6	9	14	13	2	2	1	N＝50

根据已经整理的频数统计资料，结合公式（3）的要求做出计算用表

把表中所得相应数值代入公式（3）得：

$$S.D=\sqrt{\Sigma(xi-\overline{X})^2fi/N-1}=$$

$$\sqrt{822.9100/49}=\sqrt{16.795}=4.10\text{（毫米）}$$

如果变量已经采用分组频数统计的形式，计算标准差时，只要把公式（3）中的变量xi改用组中值$\mathring{x}i$代替就行了：

$$S.D=\sqrt{\Sigma(\mathring{x}i-\overline{X})^2fi/N-1}\quad(4)$$

这里再以筒螅的样品为例，说明分组频数

表（4）

xi	fi	$xi-\overline{X}$	$(xi-\overline{X})^2$	$(X-\overline{X})^2fi$	备　注
27	3	—7.68	58.9824	176.9472	N＝50
30	6	—4.68	21.9024	131.4144	$\overline{X}=34.68$
33	9	—1.68	2.8224	25.4016	
35	14	0.32	0.1024	1.4336	
36	13	1.32	1.7424	22.6512	
40	2	5.32	28.3024	56.6048	
44	2	9.32	86.8624	173.7248	
50	1	15.32	234.7024	234.7024	
Σ	Σfi＝50			$\Sigma=(xi-\overline{X})^2fi=822.9100$	

表（5）

xi	$\mathring{x}i$	fi	$\mathring{x}i-\overline{X}$	$(\mathring{x}i-\overline{X})$	$(\mathring{x}i-\overline{X})^2fi$	备　　注
25—28	26.5	3	—8.1	65.61	196.83	
28—31	29.5	6	—5.1	26.01	156.06	C＝3（组距）
31—34	32.5	9	—2.1	4.41	39.69	N＝50
34—37	35.5	27	0.9	0.81	21.87	$\overline{X}$＝34.60
37—40	38.5	0	3.9	15.21	0	
40—43	41.5	2	6.9	47.67	95.34	
43—46	44.5	2	9.9	98.01	196.02	
46—49	47.5	0	12.9	166.41	0	
49—52	50.5	1	15.9	252.81	252.81	
Σ		$\Sigma f_i=50$			$\Sigma(\mathring{x}i-\overline{X})^2fi$ ＝958.62	

统计情况下求取标准差的具体做法。已知变量的分组资料如下：

xi	25—28	28—31	31—34	34—37	37—40	40—43
fi	3	6	9	27	0	2

xi	43—46	46—49	49—52	分9组，组距C＝3
fi	2	0	1	$\overline{X}$＝34.60 N＝50

先求出各组的组中值$\mathring{x}i$，再按公式（4）的要求分别计算和填写其余各栏。

把表（5）中所得的结果代入公式（4）得　$S.D=\sqrt{\Sigma(\mathring{x}i-\overline{X})^2fi/N-1}=$

$\sqrt{958.62/49}=\sqrt{19.56}=4.42$（毫米）

用分组频数公式（4）求得的标准差与公式（2，3）的标准差值略有差异。前者是4.42，后者是4.10，差值0.32（毫米）。这是计算方法不同而引起的。根据Simpson和Roe（1939）提出的测量精度为$\frac{1}{20}$极差的原则，上述这种差别低于精度要求的指值（本例的测量精度只要求到毫米，小数点以下的数字原则上毋需列出），并不影响性状分析质量，因此是允许的。

3．标准差的意义

样品标准差和平均数，都是用以刻划样品内变量分布特征的两项极有价值的参数。然而它们所代表的意义各有不同。样品平均数主要是用来反映变量分布的集中程度，而样品的标准差则主要用来反映变量分布的离散程度（或者叫变异程度）。例如，如果把足够数量的个体，按变量大小依次排列，就可以发现，在绝大多数情况下，最接近平均数的那些变量出现的频率最高，而与平均数差值越大的变量，出现的次数越小。上述现象，如果用座标图表示，则变量频率分布将构成一条两头低中间高的“钟”形曲线，即所谓“正态分布曲线”（见图1）。

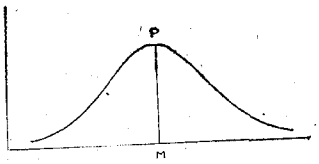

图1　正态分布曲线

图1中“P”为曲线最高点，即变量出现的最大频率，M是变量分布的中点，称为正态分布的平

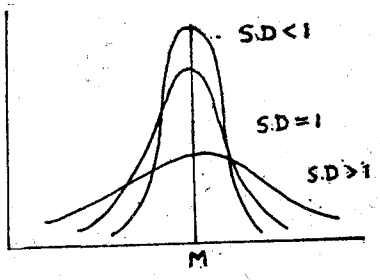

图2　S.D的直观意义

均数（即$\overline{X}$）。正态分布曲线的特点是以最高点（P）为中心，对称地向左右两边下降。但是不同变量分布所构成的正态曲线的肥瘦程度可以不完全相同（见图2），后者是由标准差（S.D）决定的，S.D越大，曲线向两侧下降的速度越慢也就越“胖”，变量就越加分散；S.D越小，曲线下降的速度越快，也就越“瘦”，变量就越加集中。从另一个角度来说，标准差的数值越大，变量在平均数周围散布得越广，曲线就越扁平；如果标准差的数值越小，那末变量在平均数周围的散布就越窄，曲线也就越陡。

正态曲线乃是生物分类研究中总体内变量分布的一般模式，而平均数（它决定曲线中点的位置）、和标准差（决定曲线向中点两侧下降的速率）则是构成正态曲线的两项必不可少的参数。

图3表示为一条标准的正态分布曲线。从图中可以看出，曲线在靠近平均数M处是向外凸的，而在离开平均数不足1/3处则开始内凹。曲线由凸变凹的地方（见图中a、b两个拐点）乃是离开平均数正负各一倍标准差的地方（记为∓S.D）。如果规定正态曲线下整个区间的面积为100%，那末凸起部分的曲线下面积(即图中的a-b-b′-a′区间）占据了整个区间面积的68.27%。从分类学的角度来说，如果把正态曲线下的面积当作变量出现的总频率，那末在离开平均数一倍标准差（∓1S.D）的区间内(即上面提到的a-b-b′-a′）变量的出现频率是68.27%。换言之，整个总体中，将有68.27%

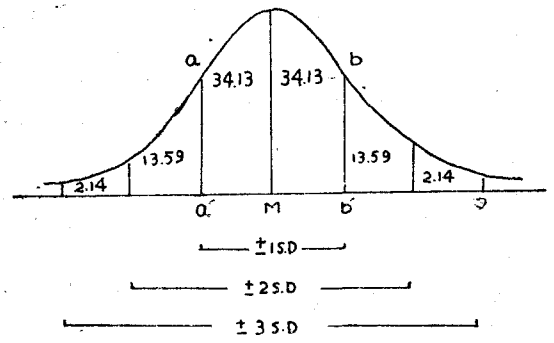

图3　正态曲线下变量所占的面积

的个体(或称变量)在∓S.D的范围内出现。

在取得了上述标准差知识后，就可以用来预测观察范围，因为：

$\overline{X}\mp 1$S.D包括样品内变量的68.27%；

$\overline{X}\mp 2$S.D包括样品内变量的95.45%；

$\overline{X}\mp 3$S.D包括样品内变量的99.73%；

而$\overline{X}\mp 1.96$S.D则包括样品内变量的95%。

例如：假定样品的平均数是50厘米，而它的标准差是3厘米（3厘米=1S.D），则在100个个体中可以期望少于5个个体是在44至56厘米范围（$\overline{X}\mp 2$S.D）之外；

又如：样品平均数是1.405（毫米），标准差是0.047(毫米)。可以认为变量落在1.311—1.499（毫米）区间内的概率（机会）是95.4%

$$(\overline{X}\pm 2\text{S.D}=1.405\pm 2\times 0.04)+\begin{cases}1.311\\ \downarrow\\ 1.499\end{cases}\text{(毫米)}。$$

利用标准差的知识，还可以决定某一特殊观察值将归入某一变异范围的可能机会。例如，前述筒螅的例子〔表(3)〕中，样品的平均数是34.68毫米，标准差是4.10毫米，问水螅体茎高41.5毫米的这个标本应该落在哪一个分布区间（或叫分布范围）？从上述标准差的介绍知道，95.45%的变量照例应在离平均数2倍标准差（$\overline{X}\mp 2$S.D）的范围内。这里

$$\overline{X}\pm 2\text{S.D}=34.68\pm 2\times 4.10=\begin{cases}42.88\\ \downarrow\\ 26.48\end{cases}\text{毫米，}$$

包括了41.5毫米这个长度值，因此这个标本应该在$\overline{X}\mp 2$S.D的区间内出现。

4．联合求取平均数和标准差的几种简易方法

上面我们对平均数$\overline{X}$和标准差S.D的计算方法，单独地作了介绍，但是在实际应用中，这两种特征数通常是必须在一起运算的。例如，只有当知道了平均数后，才能对标准差进行计算等等。这一点可以从

$$\text{S.D}=\sqrt{\frac{\sum(xi-\overline{X})^2 fi}{(N-1)}}$$

的公式中直接看出。此外在求取标准差时，公

式中的离均差平方总和 $\Sigma(x_i-\bar{x})^2$ 计算最为费力，尤其是当变量的数目较多，数值较大时，计算起来就更加繁琐，出现差错的机会也多。为了减轻工作量，并且使计算时尽可能地减少差错，这里特举几种简单有效的计算方法，供在实际应用中参考。

甲．利用简缩变量直接求取 $\bar{x}$ 和S.D

在求取标准差的公式（公式3）

$S.D=\sqrt{\Sigma(x_i-\bar{x})^2 f_i/(N-1)}$ 中把右侧根号内的 $\Sigma(x_i-\bar{x})^2 f_i$ 化开，就可以得到下列结果：

$$\Sigma f_i(x_i-\bar{x})^2=\Sigma f_i x_i^2-2\Sigma f_i x_i\bar{x}+\Sigma f_i\bar{x}^2$$

$$=\Sigma f_i x_i^2-2\Sigma f_i x_i\frac{\Sigma f_i x_i}{N}+\Sigma f_i\left(\frac{\Sigma f_i x_i}{N}\right)^2$$

$$=\Sigma f_i x_i^2-2\frac{(\Sigma f_i x_i)^2}{N}+N\frac{(\Sigma f_i x_i)^2}{N^2}$$

$$=\Sigma f_i x_i^2-\frac{(\Sigma f_i x_i)^2}{N}\qquad(A)$$

把结果（A）代入公式(3)，即成

$$S.D=\sqrt{\frac{\Sigma f_i x_i^2-\frac{(\Sigma f_i x_i)^2}{N}}{N}}\qquad(5)$$

公式（5）的好处是没有平均数（$\bar{x}$），因此可以直接利用样品内变量计算标准(S.D)。但是只做到这一步还不够，因为变量数值太大时，计算起来还有麻烦。因此必须进一步将变量化简。化简的方法与前述平均数的简捷求法相同，即把各个变量(x_i)减去一数a(分组情况下令a=居中组的组中值)，除以另一数值(令c=组距)，把简缩后的变量用U_i表示，则

$U_i=(x_i-a)/c$，于是公式（5）就成了下列形式：

$$S.D=\sqrt{\frac{\Sigma f_i U_i^2-\frac{(\Sigma f_i U_i)^2}{N}}{N-1}}\times C\qquad(6)$$

注意公式右侧根号之外要乘上“C”。这是因为变量在缩简过程中都用C除过，最后必须复原。

值得提出的是公式（6）中并没有补上各变量的共减数“a”。这是因为上式是由离均差平方总和 $\Sigma(x_i-\bar{x})^2$ 演化而来的。前者有一个特点，即各变量减去a值后所得的离均差平方总和，与直接由各变量求得的完全相等。这一点可以通过下述演算得到证明。

设 x_i 为各变量，$\bar{x}$ 为平均数，a 为任意共减数。令x'为简缩变量，$\bar{x}'$为简缩平均值，则$x'=x_i-a$移项得$x_i=x'+a$

在讨论平均数简捷算法时知道，样品平均数 $\bar{x}$ 等于简缩平均数 $\bar{x}'$ 加共减值a。即 $\bar{x}=\bar{x}'+a$，把上面这些有关式代入离均差$(x_i-\bar{x})$的表示式内，则

$$(x_i-\bar{x})=[(x'+a)-(\bar{x}'+a)]$$

$$=(x'-\bar{x}')\qquad(B)$$

把（B）式两端的算式平方，再分别乘上综和符号Σ，则得 $\Sigma(\bar{x}_i-x)^2=\Sigma(x'-\bar{x}')^2$

由此证明，减缩前后的离均差平方总和完全相等。因此在标准差的公式（6）中无需作共减数 a 的还原。

下面是求取$\bar{X}$和$S.D$的用法实例。

表（6）

1.36	1.47	1.45	1.42	1.37	1.39	1.38	1.41	1.31	1.39	1.42	1.37	1.41
1.49	1.39	1.35	1.30	1.37	1.46	1.40	1.48	1.41	1.40	1.40	1.55	1.44
1.43	1.41	1.42	1.34	1.44	1.39	1.36	1.39	1.44	1.38	1.41	1.38	1.48
1.41	1.36	1.39	1.42	1.45	1.53	1.45	1.45	1.44	1.35	1.37	1.42	1.38
1.37	1.40	1.44	1.37	1.32	1.36	1.50	1.37	1.42	1.42	1.46	1.34	
1.40	1.30	1.42	1.36	1.48	1.48	1.43	1.37	1.47	1.43	1.36	1.43	
1.32	1.42	1.39	1.37	1.40	1.40	1.38	1.39	1.35	1.42	1.37	1.42	N=100
1.42	1.42	1.42	1.34	1.45	1.39	1.43	1.45	1.36	1.42	1.27	1.41	单位：厘米

例如，某一样品（例六）有100个变量(见表（5））。

（1）计算前先找出最大变量和最小变量的差值（即找出极差）或称全距。本例中变量的最大值为1.55厘米，最小值为1.27厘米。全距1.55－1.27＝0.28厘米。根据分组原则将其分为10组，确定每组的组距为0.03厘米（如果分为10组，则组距＝全距/组数＝0.28/10＝0.028厘米，小数后的位数太多，为了计算方便，这里采用0.03厘米）；

（2）列表并找出各组的组中值。例如第一组的组中值（1.265＋1.295)/2＝1.28、第二组的组中值是(1.295＋1.325)/2＝1.31……等；

（3）按公式（6）所提的各项要求，列出操作表格，进行计算和填表；

（4）把表（6）所得的相应数字分别代入简捷平均数公式和标准差（6）

$$\bar{X}=\frac{\Sigma u_i f_i}{N}\times c+a=\frac{19}{100}\times 0.03+1.40=1.406\text{厘米}$$

$$S.D=\sqrt{\frac{\Sigma u_i^2 f_i-\frac{(\Sigma u_i f_i)^2}{N}}{N-1}}\times C$$

$$=\sqrt{\frac{261-\frac{19\times 19}{100}}{100-1}}\times 0.03$$

$$=\sqrt{2.5999}\times 0.03=1.61\times 0.03$$

$$=0.048$$

以上即为利用简缩变量直接求取$\bar{X}$和$S.D$的方法。

乙．用简缩变量的累加方法求$\bar{X}$和$S.D$

在变量组组距相等的情况下，还可以采用另外一种简捷的方法求出$\Sigma ui fi$和$\Sigma ui^2 fi$。这种方法所依据的原理是把一系列的乘法改成一系列的加法，以此进一步减少计算误差。为了便于和前面的例子相比较，这里仍用例六加以说明。

① 把例六的组中值和频数列成表格形式，同时在频数栏的右侧补上第一累加栏和第二累加栏（见表7）。

② 计算并填写表内的累加值。

具体的做法是：任意确定一组为中心组（这里确定组中值＝1.40的第五组作为中心组）。规定第三列（即表中第一累加栏）的中心组为“0”值，其它各组的组值分别用第二列（即频数栏）内有关组的累加数填写。例如，第三列第2组的组值是第二列中第1，2两组

表（7）

Xi (变量组)	$\overset{\circ}{Xi}$ (组中值)	f_i (频数)	$u_i=\frac{\overset{\circ}{Xi}-a}{C}$ (缩减后的变量)	$u_i f_i$	$u_i^2 f_i$	备注
1.265—1.295	1.28	1	－4	－4	16	
1.295—1.325	1.31	4	－3	－12	36	
1.325—1.355	1.34	7	－2	－14	28	N＝100
1.355—1.385	1.37	22	－1	－22	22	
1.385—1.415	1.40	24	0	0	0	a＝1.40
1.415—1.445	1.43	24	1	24	24	
1.445—1.475	1.46	10	2	20	40	c＝0.03
1.475—1.505	1.49	6	3	18	54	
1.505—1.535	1.52	1	4	4	16	
1.535—1.565	1.55	1	5	5	25	
Σ				$\Sigma u_i f_i=19$	$\Sigma u_i^2 f_i=261$	

之和（1＋4＝5），第四组的组值是第二列中1，2，3，4这四个组的频数之和（1＋4＋7＋22＝34）……而第三列倒数第三组的组值则是第二列10，9，8，这三组的频数总和（1＋1＋6＝8）等等，依次类推；又规定第四列（即第二累加栏）的中心组及其邻近两组的组值都是“0”，其余各组的组值则用第三列（第一累加栏）中，有关组的累加数填写（方法与上述相同）。填写结束后，把第二累加栏的数值总加在一起，这样便得到了A，B，C，D和E五个数字（见表（7））。

于是：$\Sigma uifi=(A+B)-(C+D)$

$=(42+29)-(34+18)=19$

$\Sigma ui^2fi=2E+A+B+C+D$

$=(2\times 69)+42+29+34+18=261$

把上述相应值分别代入公式（7）* 和(6)得：$\overline{X}=\Sigma uifi/N\times c+a=1.406$（厘米）

$$S\cdot D=\sqrt{\frac{\Sigma ui^2fi-\frac{(\Sigma uifi)^2}{N}}{N-1}}$$

$\times C=0.048$（厘米）

用本法计算的结果与上法完全一致。

丙．利用正态概率纸求取$\overline{X}$和$S\cdot D$

正态概率纸是按正态分布规律而专门设计出来的一种特殊座标纸（见图四）。图内纵座标轴上的刻度表示累计频率，横座标轴的刻度代表变量分布。

现在仍以上例说明正态概率纸的用法。

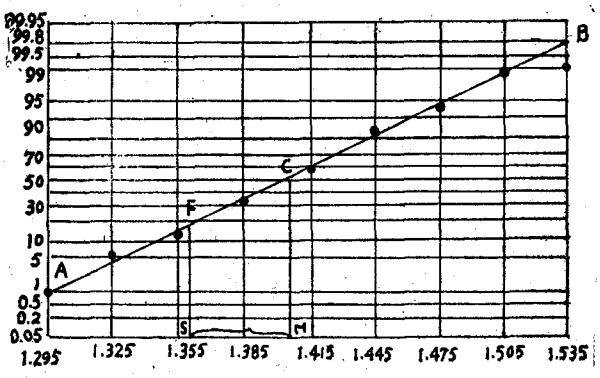

图4　正态概率白纸及其应用

表（8）

	Xi（1）	fi（2）	第一累加（3）	第二累加（4）	备注
（组1）	1.28	1	1	1	
（组2）	1.31	4	5	6	A＝42
（组3）	1.34	7	12	18(D)	
（组4）	1.37	22	34(C)	0	B＝29
（组5）	1.40	24	0	0	C＝34
（组6）	1.43	24	42(A)	0	D＝18
（组7）	1.46	10	18	29(B)	
（组8）	1.49	6	8	11	E＝69
（组9）	1.52	1	2	3	
（组10）	1.55	1	1	1	
	Σ			69(E)	

表（9）

变量组右端点（1）	频数（2）	累计频数（3）	累计频率（4）
1.295	1	1	1%
1.325	4	5	5%
1.355	7	12	12%
1.385	22	34	34%
1.415	24	58	58%
1.445	24	82	82%
1.475	10	92	92%
1.505	16	98	98%
1.535	1	99	99%
1.565	1	100	100%

先在表（6）的基础上列出累计频率（见表（8））。

表中第（3）列各组的数值是第（2）列直到该组为止的频数总和。第（4）列是用百分率表示的各组累记数（即累计频率）。作图时，把表（8）内的第（1）列（变量组的右端点值，或叫上限）当作横座标点，把第（4）列中的累计频率作为纵座标点（见图四）。例如第一点的横座标是1.295，纵座标为1，第二点的横座标是1.325，纵座标是5……依次进行，直到把表中所列的点值全部在图纸上打完

* 此公式见本刊1978年第4期40页

为止。最后把各点相连。如果变量是正态分布，那末各点的连线应该基本上是一条直线**，否则就不是正态分布。从图四的结果看出，本例样品内的变量属于正态分布。

上述作图完成之后，接着就可以估计平均值和标准差 3 。作法是先从纵座标50%这一点引出水平线与直线AB相交于C，再由C点向横座标轴作垂线，交于M点。M点即为平均数。从图上可以估计出，样品平均值$\overline{X}$（即M）大约为1.405厘米。

得出平均数的估计值后，再作标准差的估计。从正态分布曲线的知识知道，离开平均数单侧一倍标准差以外的变量，它们的出现频率是15.87%（见图三）。如果从纵座标15.87%处作一条水平线，则交AB线于F点，由F点向下作垂线交横轴于S点，这个S点就是平均数减去一倍标准差的地方。从图上看这个地方的数值大概是1.358，把平均数（$\overline{X}=1.405$）减去这个数字，剩下的差值（即图四中加括号的一段）就是一倍标准差值。亦即$S\cdot D=1.405-1.358=0.047$厘米。由正态概率纸求出的结果与前述两种简捷算法的结果是比较接近的。

正态概率纸的优点是：1．利用它可以立刻估计出变量是否属于正态分布；2．可以推估样品的平均数和标准差；3．有直观效果，操作也比较方便，只要作图准确，取得的估值也是比较精确的。

** 一般认为，由于样品的随机波动，图内各点不一定完全在一条直线上，这种偏差是允许的，但是中间点离直线的偏差不能太大，否则就可能不是正态分布。

（上接43页）

洋的热带和亚热带的深海海域200—500米的中层水域，或者在陆架边缘的深水中皆有分布。这种分布规律，是受着太平洋南、北赤道流的控制和影响，所分布的海域往往又是高温，高盐的深水海域。也有个别的种类如鲛鰊亚目中的一些种类，则分布在低温、高盐、低压的500—2,500米的深海“微光区”和“无光区”。然而在我国的东海海域，发光鱼类的仔、稚鱼，除七星鱼在东海近岸和外海广泛分布外，其余绝大部份种类仅出现在东海外海的陆缘区，极少在近海出现。这一分布特点，明显地表明它们是在受着北赤道流的影响，随其黑潮暖流在东海外海自西南向东北的流迳过程中，而进入东海的陆架边缘，黑潮暖流所控制的高温（多在26.0℃以上）、高盐（34.00‰左右）的深水（200米水深以上）的黑潮区域。同时，这些发光鱼类在夜间作着垂直移动。当在海上考察，夜间作业时，用浮游生物网垂直拖取和表层挂流网表层拖取时，不仅能采获到比白天采获量大得多的仔鱼和稚鱼样品，而且用手抄网亦能采到它们的幼鱼或成鱼。发光鱼类在东海陆架区生态分布的这一特殊规律性，是东海海域里其他任何鱼类所没有的。（图 8 ）

总之，在我国广阔的东海海域，不仅蕴藏着多种重要的经济鱼类资源，而且也分布着种类繁多的发光鱼类。但是有些大洋性深海底层的发光种类，还有待我们今后进一步的调查、补充和研究。

参 考 文 献

1）孙继仁：1977.
中国近海（南海粤西海区）仔、稚鱼的研究。全国海洋综合调查报告。十册十二章。

2）费鸿年译：1963.
勃朗编著：鱼类生理学 下册，1957年出版。科学出版社。

3）Norman, J.R., 1951.
A. History of Fishes

4）Badcock, J., 1969.
Colour Variation in Two Mesopelagie Fishes and its Carrelation With Ammbient Light Conritions.
Nature, Vol. 221.

5）Karl, F.L., et al., 1977.
Ichthyology. (2nd Edi)
John Wiley & Sons.