生物分类学统计中几个常用特征数字的求取方法(一)

张 伟 权

(中国科学院海洋研究所)

前 言

"分类"研究是一切学科研究的基础。海洋生物分类的目的，就是要将有关生物之间、错综复杂的差异分析整理成易于认识的类群，找出它们的重要性状以及不同类群之间的恒定区别，并且在此基础上分门别类，列出系统，使物种形成和生物进化的事实，最终能够在各级分类阶段中得到客观、合理的反映。

据估计，地球上现存的生物约有150万种(其中动物110多万种，植物为30余万种)。要是把亚种的数目也估计在内，则上述数字可能超出300万，而且新类群的描述还在不断增加。可以想象，如果不把这些种类进行科学的分类，那末一切更加深入的研究，就会变得毫无价值。但是，想要达到上述分类的目的，采用传统的方法是不够的，这是因为随着现代研究工作的深入，新的试验和观察内容不断增多，千差万别的记录数字必将大量涌现，对于这样的资料，采用一般手段是很难弄清其内在规律的。长期以来，人们在探索自然的过程中，逐渐认识到，利用数理统计的方法可以比较容易地透过现象，把握事物的本质。由于通常的统计方法具有简单易行的特点，因此把它用来作为学科研究的一种手段，是值得大力提倡的。目前统计方法已经成为许多专业进行研究的主要技术标准之一，并且越来越多地被广泛采用。

本文试图从海洋生物分类的角度出发，对生物统计技术的基本方法进行一般性的探讨，目的是引起初学分类同志的注意，以便在日后的科学工作中推广选用。下面就分类工作中常用的几个特征数字(包括算术平均数、标准差、变异系数、比率、标准误差、相关系数和差异系数)以及差异显著性测定等，依次进行常识性质的介绍，介绍尽量以举例为主，各种方法原则上只作直观说明，不作数理推导。由于作者水平有限，差错肯定难免，殷切希望广大读者批评指正。

* * *

分类学工作者在进行生物分类研究的同时，有二个问题必须向自己提问：1．你所选择的性状是否适用于分类识别？如果回答是肯定的，那末，2．用什么方法才能比较全面地表达这种性状？

对于上述二个问题，一般的回答是：要尽可能地利用统计学手段。

尽管，有些分类学者习惯于经验而忽视了统计学方法，但是事实上他们不仅在列举平均数和体形大小范围时使用了统计数字，而且在表示身体比例和进行相似种类的比较时也不知不觉地使用了统计学。这是由于统计数字无可否认地可以增加性状描记的正确性的缘故。例如，在描述中国对虾 *Penaeus orientalis* Kishinouye 的体形大小时，说它的体长"较大"，总不如说"体长平均为20厘米"更为确切；同样，在描记毛蚶 *Anadara* (*A.*) *sub-*

crenata Lischke时，说“壳表的放射肋有32条”要比笼统地说“壳表有肋”更接近实际。

过去，一个“模式”观点的分类学家，相信种是绝对的，无空间向度的，很少涉及到地理变异。因此他们只要凭借一、两个生物标本就可以用来进行描记或者确定它们的分类位置。但是随着近代分类学研究的逐渐深入，种群概念代替了种的模式概念。物种不再被认为是固定不变的了。大量事实证明，物种是由种群组成的，而种群与种群、甚至同一种群内部各个体之间，彼此都不可能完全相同。因此，如果不采用统计的方法，就不可能恰当地进行物种描述。也就更加无法确切地进行种群间差异程度的比较。分类统计的功用，主要在于能够将数量资料做系统的分析和估计，以此推估自然种群的基本特征。

可以这样说，任何分类学描记，如果不包括起码的数量资料，其结果就不够完整。

一、算术平均数（Mean）

1．求取平均数的一般公式

分类学工作者的基本任务之一，是要把自然界中，个体间错综复杂的性状差异加以分析，从中找出规律性的东西，使之成为易于认识的类群，并且在此基础上进行分类和探索物种起源。要想达到这个目的，单凭一、两个标本当然是不够的。但是一大堆杂乱无章的样品，也同样无实用价值可言。为了便于找出个体间性状差异的实质性区别，首先应该对样品进行初步整理。例如把业已确定用于统计分析的某一分类性状，按标本进行逐个测量，并把测量的结果根据变量大小进行分组和作出频数统计表格等（见后例子），接下去便可进行分析和统计了。

生物分类研究中，最经常遇到的一个统计项目是算术平均数（arithmetic mean）。用符号$\overline{X}$表示（读作X-bar），或者用M表示。平均数的定义是：样品内的变量之和$\sum_{i=1}^{n} Xi$除以样品内的标本总数N。例如，在一个由N个变量（即个体的代表值）组成的样品内，其各变量分别用X_1，X_2，$X_3 \cdots X_n$表示，则平均数$\overline{X}$可由下式求得：

$$\overline{X}=\frac{X_1+X_2+X_3+\cdots+X_N}{N} \qquad (1)$$

上式也可以写成：

$$\overline{X}=\frac{X_1}{N}+\frac{X_2}{N}+\frac{X_3}{N}+\cdots+\frac{X_N}{N} \qquad (2)$$

为了简化起见，将上述（2）和（1）改成如下形式：

$$\overline{X}=\sum_{i=1}^{n} X_i/N \qquad (3)$$

式中Σ称为综和符号（读作Sijma，是大写的希腊字母），X_i代表变量，$\sum_{i=1}^{n}$（以后简号为Σ）的意思是把所有的变量（X_1，X_2，$X_3 \cdots X_N$）相加到一起，即由X_1一直加到X_N为止。

例如复瓦哈鳞虫 *Harmothoë imbricata* (*Linné*) 的一个样品由8个标本组成。它们的体节数目分别为35，35，36，37，35，36，37，36（节），试求其平均数？

已知该样品共有8个标本，即$N=8$

把上列有关的体节数目直接代入简单式（3）

则 $$\overline{X}=\frac{35+35+36+37+35+36+37+36}{8}$$

$$=\frac{288}{8}=36$$

上述样品的体节平均数是36节。

上面（3）这个公式在样品较小的情况下比较适用，但是如果样品内包含的个数很多，为了计算方便，需要将变量进行频数统计。样品经过频数统计后，求取平均数的公式变化如下：

$$\overline{X}=\Sigma f_i x_i/N \qquad (4)$$

公式（4）中，f_i代表样品内不同变量出现的次数（即频数）、x_i代表不同的变量。

用频数统计法求取平均数的例子如下。设有30尾中国毛虾*Acetes chinensis* Hansen组成的一个样品，其尾肢内枝基部排列的红色斑

点数目各为3,4,4,6,3,3,4,5;4,3,4,4,5,4,3,
,5,3,4,4,6,3,4,4,5,3,4,4,4,4，求其平均数。

先将变量按大小次序作出频数分布表

表（1）

X_i 变量分布)	f_i (各变量出现的频数)	f_iX_i (变量×频数)	备注
3	8	24	
4	16	64	N=
5	4	20	30
6	2	12	
Σ(综和)	$\Sigma f_i=60$	$\Sigma f_iX_i=120$	

将表内求得的相应数字代入公式（4）

则 $\overline{X}=\Sigma f_iX_i/N=120/30=4$

即这一毛虾样品的斑点平均数为4（个）

当样品所含的变量数目很大，特别是当样品内变量分布呈连续性质时（即变量不是正数1,2,3,4…N而是1.2，1.34，1.39，2.00，2.10…N，两变量之间不具明显的间断），就需要将变量进行分组频数统计，否则计算起来会变得十分困难。分组情况下求取样品平均数的公式是：

$$\overline{X}=\Sigma f_iX_i^0/N \qquad (5)$$

式中f_i代表不同变量组内，变量出现的次数；X_i^0代表各变量组的组中值（组中值是由变量组的最大变量maxX（或称上限）加最小变量minX（或称下限）除2而得。用算式表示为：$(\max X+\min X)/2$。

为了使同志们熟悉分组情况下求取平均数的具体方法，下面用一个简单的、变量分布为非连续性质的例子加以说明（连续性分布的例子可参阅平均数的简捷算法一节）。

例如，中华蜾蠃蜂 *Corophium sinense* Zhang 100个雌体抱卵量的实际观察值如下表。试用分组频数法求其平均数。

表（2）

84	36	51	31	41	28	19	19	29	16
23	37	17	17	40	28	20	41	50	14
32	30	24	53	28	27	38	18	48	74
41	28	30	38	36	42	34	43	21	27
15	29	73	20	29	18	40	52	31	30
27	27	62	31	43	32	26	27	17	30
39	41	16	29	45	30	16	42	36	37
35	52	66	36	21	37	34	68	32	31
28	40	35	32	20	33	41	20	41	35
37	43	42	17	42	36	35	70	50	82

第一步：确定分组数目和组距。上述观察记录中，每尾中华蜾蠃蜂的最小抱卵数目为14粒，最大抱卵数为84粒，其变量的差值（即极差**）为70，根据通常的原则，当变量的总数至100个以上时，一般将样品分成8—20组（注：组数越少，则组距越大，计算起来比较方便，但所得的精确度差；组数越多，则组距越小，由计算得出的数字其精确度越高，但计算麻烦。为了尽可能地做到既使计算方便又保证适当的精确程度，习惯上当样品所含的个数在100以上时将其分成8—20组，个数在100以下时分成5—8组）。这里将样品分为8组，每组的组距为10。

第二步：按公式（5）要求，列出频数统计表如下：

** 极差或称全距（记为R），为样品内最大变量与最小变量的差值。极差、组距和组数的关系为：组数＝极差/组距。变量分组时通常用这一公式来进行划分。

表（3）

X_i（变量组）	X^0_i（变量组的组中值）	f_i（各变量组内变量出现的次数）		f_iX_i（频数×变量）	备　注
10—20*	15	13	正正下	195	
20—30*	25	23	正正正下	575	N=100
30—40*	35	33	⋮	1155	（样品内个体
40—50*	45	18	⋮	810	的总数）
50—60*	55	5	正	275	C=10（组距）
60—70*	65	3	下	195	
70—80*	75	3	下	225	
80—90*	85	2	下	170	
Σ（综和）		$\Sigma f_i=100$		$\Sigma f_iX_i=3600$	

在制作上述分组频数表时有几点应该注意：①最小变量（例如表中的15）应尽量放在第一变量组的靠近组中值处，最大变量（例如85）应尽量排在最末一组的近组中值处，也就是说，最小变量和最大变量不要分别用作第一组的始值和最末组的终值（这样出现的统计误差较小）；②为了使所有的变量都能明确无误地记入相应的变量组内，可以采用划记“正”号的方法（象选举时的唱票那样）将变量分划入组，划记结束后，再用阿拉伯字母代替，将“正”号用橡皮擦去（见表1,3的第3列）。③为了避免混乱，凡变量与各组中带有*号的数值相同时，该变量应一律归入下一组。例如观察表中“20”这个变量共出现3次，因为它与第一组的上限（即带*号者）的数值相同，因此在统计频数时应该将其放在第二组（表中20—30*的那一组）而不是放在第一组的频数栏内。

第三步，将分组统计表中所得的相应数字代入公式（5）得：$\overline{X}=\Sigma f_iX_i^0/N=3,600/100=36$（粒）

2．样品平均数的意义和特点

样品平均数是一项极有价值的参数，可以用它推估自然总体的特征。因为1．平均数是由样品内各变量求得的，也就是说是从每个变量中分别提取1/N组合而成的（见公式（2）），计算平均数时参予统计的每一个变量都同等重要，只要增减全体变量中的任何一个，结果就会受到影响。因此平均数有照顾全体的特点，代表性强（平均数的代表性）；2．在同一环境下采集同种样品时，由各样品分别算出的平均数彼此比较接近，换言之平均数受取样的影响要较其它集中性指标的特征数字（例如中位数*—median）为小（即平均数的稳定性）。3．任何变量都是围绕平均数分布的。越靠近平均数的变量、出现的机会越多（见正态分布一节），离开平均数越远的变量出现的机会越小（平均数的集中性）。

除了上述特点外，平均数还有以下两个特征：

1．任何样品内，各变量与平均数的差值之总和为零（即离均差的总和$\Sigma(X_i-\overline{X}=0)$）。

---

* 中位数是指变量分布的居中数。当变量为奇数个时，居中数只有一个。例如1,2,3,4,5中的“3就是中位数；当变量为偶数个时，中位数就等于二个居中数的算术平均数。例如1,2,3,4,5,6的中位数是(3+4)/2=3.5。中位数有时也可以用来代表样品的平均水平，但其代表性和稳定性都不如平均数，只要在变量分布的任何一端增加或者减少一个数字，中位数的位置就会改变，因为在生物统计中极少采用。

例如：7,6,5,4,3这5个变量(属奇数个)，其平均数$\overline{X}=5$。设X_i为变量，$(X_i-\overline{X})$为各变量与平均数之差（离均差），则：

$$\Sigma(X_i-\overline{X})=(7-5)+(6-5)+(5-5)+(4-5)+(3-5)=2+1+0-1-2=0$$

又如，7,6,5,4,3,2，这6个变量（属偶数个）的$\overline{X}=4.5$

则：$\Sigma(X_i-\overline{X})=(7-4.5)+(6-4.5)+(5-4.5)+(4-4.5)+(3-4.5)+(2-4.5)=2.5+1.5+0.5-0.5-1.5-2.5=0$

2．样品中，各个变量与任何一个数值之差的平方总和，永远要比同一样品内各变量与平均数之差的平方总和为大（换言之，即离均差的平方之和为最小）。

例如，一样品的变量分别为5,4，4，3，其

表（4）

X_i	$X_i-\overline{X}$	$(X_i-\overline{X})^2$	X_i-a_1	$(X_i-a_1)^2$	X_i-a_2	$(X_i-a_2)^2$	备　注
5	1	1	0	0	+3	9	
4	0	0	-1	1	+2	4	$N=4$
4	0	0	-1	1	+2	4	$\overline{X}=4$
3	-1	1	-2	4	+1	1	$a_1=5$
Σ		$\Sigma(X_i-\overline{X})^2=2$		$\Sigma(X_i-a_1)^2=6$		$\Sigma(X_i-a_2)^2=18$	$a_2=2$

表（5）　长蛇鳚鱼一个样品的体长观察记录（厘米）

192.0	205.2	209.3	213.6	216.7	219.1	221.8	225.1	228.0	232.8	239.9
193.0	205.4	210.0	213.9	216.0	219.3	221.0	225.0	228.1	233.4	242.0
194.0	205.8	210.1	214.1	216.0	219.7	222.0	225.9	228.1	233.7	242.4
195.0	205.7	210.4	214.0	217.0	219.3	222.3	225.9	229.5	234.2	243.5
196.5	206.0	210.5	214.3	217.0	219.4	222.4	225.3	229.4	234.1	243.7
197.0	206.1	210.3	214.5	217.4	220.0	222.0	225.0	229.8	234.9	245.1
198.0	206.3	210.1	214.0	217.7	220.0	222.7	226.1	230.0	234.6	247.5
198.0	206.5	210.2	214.7	217.8	220.1	222.6	226.3	230.1	235.3	249.1
199.5	206.4	211.3	214.9	218.0	220.4	222.4	226.3	230.3	235.4	252.0
200.3	206.7	211.0	215.0	218.0	220.6	222.1	226.0	230.4	235.0	
201.4	206.9	211.2	215.2	218.8	220.3	223.0	226.1	230.2	236.1	
201.5	207.0	211.3	215.3	218.4	220.7	223.4	226.4	230.1	236.7	
201.8	207.0	211.3	215.4	218.3	221.1	223.4	226.5	231.1	237.1	
202.0	207.2	212.0	215.4	218.9	221.0	223.1	226.0	231.4	237.0	
202.0	207.4	212.4	215.0	218.4	221.0	223.2	227.0	231.2	237.4	
202.4	207.5	212.5	215.7	218.2	221.3	223.0	227.0	231.6	237.0	
202.7	208.3	212.3	215.9	218.1	221.4	223.2	227.4	231.7	238.4	
202.8	208.5	212.6	215.0	218.0	221.0	224.1	227.3	232.3	239.4	
203.1	208.0	213.4	216.3	219.0	221.2	224.2	227.6	232.3	239.4	
203.3	208.7	213.4	216.4	219.0	221.4	224.3	227.1	232.0	239.0	
204.0	209.0	213.5	216.1	219.0	221.7	225.0	228.3	232.7	239.3	

表（6）利用简捷算方求取平均数的分组统计表

Xi （变量组）	X_i^0 （组中值）	$Ui=(X_i^0-a)/c$ （变量缩简值）	fi （变量在各组出现的频数）	fiui 变量频数X变量缩简值）	备注
190—194	192	(192—222)/4=－7.5	2	－15.0	
194—198	196	(196—222)/4=－6.5	4	－26.0	
198—202	200	(200—222)/4=－5.5	7	－38.5	
202－206	204	(204—222)/4=－4.5	12	－54.0	
206—210	208	(208—222)/4=－3.5	19	－66.5	N=222
210—214	212	(212—222)/4=－2.5	24	－60.0	
214—218	216	(216—222)/4=－1.5	27	－40.5	a=222
218—222	220	(220—222)/4=－0.5	35	－17.5	
222—226	224	(224—222)/4=0.5	26	13.0	C=4
226—230	228	(228—222)/4=1.5	21	31.5	
230—234	232	(232—222)/4=2.5	18	45.0	
234—238	236	(236—222)/4=3.5	13	45.5	
238—242	240	(240—222)/4=4.5	6	27.0	
242—246	244	(244—222)/4=5.5	5	27.5	
246—250	248	(248—222)/4=6.5	2	13.0	
250—254	252	(252—222)/4=7.5	1	7.5	
Σ				$\Sigma fiui=-108$	

$\overline{X}=4$，设a_1和a_2各为一任意数，并令$a_1>\overline{X}>a_2$（即a_1大于平均数，a_2小于平均数），例中设$a_1=5$，$a_2=2$，则上述结果可以由下表看出：

由表（4）可知，离均差平方之总和$\Sigma(X_i-\overline{X})^2$为“2”，而其它任何一数（例如本例中的$a_1$和$a_2$）与各个变量的差值平方和都较离均差平方之和要大（表中分别为6和18）。

记住平均数的上述两项特征，对于以后的计算和进一步理解问题将会带来好处。

3．计算平均数的简捷方法

如果一个样品所含变量的数目太多，而且数值较大，采用一般的平均数公式计算有时仍嫌麻烦，理想的方法是使用简缩运算法。先将样品内的各个变量简化，求出简化平均数，然后再行还原。这种方法在生物统计学中比较常用，现举例如下。

假设长蛇鮫鱼*Saurida elongata* T. e+S. 的一个样品，共有222个标本组成，其体长的变异范围在192—252厘米之间，具体观察值见表，(5)试用简捷算法求其平均体长。

已知上述变量的分布范围为192—252厘米，则全距(252—192)=60厘米。由于本例样品所含的个数较多（N=222），根据前述变量分组的原则，将全距分成16组，每组的组距为4厘米（注意最小变量和最大变量应分别置于第一组和最末组的近中部处）。

计算时先将各变量的组中值x^0_i分别减去一数a和除上一数c（a和c的选取无严格规定，但一般以分组表中，居中组（或者接近居中组的某一组）的组中值作为a，以组距作为c）。这里令a=222（即居中组的组中值），c=4（组距）。Ui为缩简后的变量。则$Ui=(x^0_i-a)/c$。这样，求取缩简平均数（用$\overline{x}_c$表示缩简平均数）的公式就成了如下形式：

$$\overline{x}_c=\Sigma f_iU_i/N \qquad (6)$$

接下去，根据公式（6）的要求，列出样品分组频数统计表，并按照各栏所示进行计算和填写（见表6）

（下转第46页）

会议消息

中国海洋湖沼学会恢复大会暨学术讨论会在青岛召开

中国海洋湖沼学会恢复大会暨学术讨论会于11月20日至28日在青岛召开。该学会成立于一九五〇年，文化大革命以来，由于林彪、“四人帮”的破坏干扰，学会一直停止了活动。这次会议的召开，是该学会十五年来的第一次盛会。参加会议的有学会第二届理事会理事，《海洋与湖沼》学报编委会委员，还有中国科学院五局、国家海洋局、国家地质总局、科学出版社、中国大百科全书出版社和来自十五个省市有关科研、教学单位的代表和列席代表138人。全国科协、山东省科协、中共青岛市委的负责同志也出席了会议，并讲了话。

这次标志着中国海洋湖沼学会获得新生的盛会，检阅了十几年来我国海洋湖沼科学的发展成就。会议收到论文、报告和摘要共235篇，在会上宣读了85篇。这些论文不论在数量和质量方面，或是在研究内容的广度深度方面，都是学会成立以来所未有的。有些论文有一定的理论水平和创见，有些对生产实践有一定的指导意义，有的还填补了我国部分空白学科。老科学家饶钦止、赫崇本、曾呈奎、毛汉礼、施成熙、刘健康等都在大会上作了学术报告。饶钦止、赫崇本教授在报告中还详细阐明了海洋科学与湖沼科学是姐妹学科，它们之间是有机联系和互相促进的，受到与会同志的赞同。

会议代表回顾我国海洋湖沼科学的发展情况，愤怒揭批林彪、“四人帮”，扫除了思想障碍。大家在讨论中都能各抒已见，畅所欲言，充分表达了自己的意见，并对今后的科研工作和学会的活动等问题提了许多好的设想和建议。会议代表认为，当前我国海洋科学和湖沼科学力量都很薄弱，在海洋湖沼学会恢复以后，当务之急是把这两方面的力量团结起来，集中力量办好学会，加快海洋湖沼科学研究的步伐，为实现我国的四个现代化，赶超世界先进水平作出贡献。会议经过充分讨论，修改了海洋湖沼学会会章，增补了学会副理事长、常务理事及《海洋与湖沼》学报编委，推选了童第周、孙云铸、伍献文、朱元鼎等教授为中国海洋湖沼学会名誉理事长，补选了曾呈奎教授为学会理事长、刘瑞玉副教授为秘书长。会议圆满地完成了任务，达到了预期目的。

（蔡浩然）

（上接第45页）

上述填表结束后将表中给出的相应数字代入公式（6）即为缩简后的平均数。$\overline{X_0}=\Sigma f_iU_i/N=-108/222=-0.486$

最后，在缩简平均数（$\overline{X_0}$）的基础上补回原来减去和除去的a值和c值，所得的结果就是我们要求的样品平均值$\overline{X}$。

即 $$\overline{X}=\overline{X_0}\cdot c+a \qquad (7)$$

上例中$\overline{X}=\overline{X_0}\cdot c+a=-0.486\times 4+222=22,005$（有效值220）

亦即蛇长鳚鱼样品的平均体长为220厘米。

上面举的是变量同时减a和除c的例子，但是也可以只采用其中的一项进行缩简。这要看具体情况而灵活掌握，如果变量不是除数的倍数，那么为了避免计算过程中出现新的麻烦（有小数位出现），可以只用减a法。后一种場合下，统计表中第三例的$U_i=\frac{x_i-a}{c}$的形式要相应地改成$U_i=x_i-a$，以后的计算步骤则完全不变。