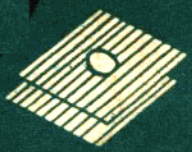
水产生物统计

（2．统计分析方法）

孙尽善　编著



农业部中等水产教育研究会

编者的话

近年来全国各地的中等水产专业学校，纷纷开设了《水产生物统计》这门专业基础课。但至今该门课的统编教材尚未正式出版。为了满足当前各校的教学急需，我受农业部中等水产教育研究会和《中等水产教育》编辑部的委托，编写了“水产生物统计分析方法”这本小册子。本书主要介绍了试验数据的整理和分析方法、统计推断方法、方差分析方法以及相关与回归分析方法等。本书对于《水产生物统计》中所常用的几个统计分析方法都作了简明介绍。因而，本书与我编的《水产科学试验设计方法》一书配套，可作为中等水产学校水产养殖专业《水产生物统计》课的代用教材。在选作代用教材时，最好先讲《水产科学试验设计方法》的前六章，再讲本书的内容。当本书的内容讲完后，再讲《水产科学试验设计方法》的第七和第八章。这样处理的好处是：能使各种统计分析方法有个再复习和巩固的过程，有利于提高《水产生物统计》这门课的教学质量。水产科技人员学习《水产生物统计》时，也应依此顺序学习为宜。

本书在编写上着重统计分析方法的介绍和应用，对于统计理论未克兼顾。文字上力求通俗易懂，便于读者自学和应用。由于编者水平有限，编写时间仓促，加上印刷条件限制，书中难免存在缺点错误，恳切希望读者批评指正，以便今后有机会时再修改提高。

孙尽善

1989年10月于山东省水产学校

目　录

第一章　数据的整理与分析

在水产生产和水产生物科学研究中，我们所遇到的现象一般都是随机现象。水产生物统计方法，就是研究随机现象的统计规律性的一门学科。研究随机现象的统计规律性，需要进行观测或试验，取得原始资料，利用这些资料进行整理和分析，对所研究的问题通过观察对比作出估计和推断，以找出事物的一般规律性，并用它来改造客观世界的实践。本章主要结合水产养殖实例介绍描述统计学的一般内容。

§1　总体、个体、样本

一、几个基本概念

例如，当我们研究某养虾池中对虾的体长时，对于每一尾对虾都有一个确定的体长，它们都是研究的对象。统计学中把研究的对象的全体叫做**总体（或母体）**，总体中的每一个对象叫做**个体**，总体中的一部分叫做**样本（或子样）**，习惯上还把总体中的个体的个数叫做**总体的容量**，记为N。把样本中所包含个体的的个数叫做**样本的容量**，记为n。通常当$n<50$时，认为属于小样本，$n\geq 50$时，认为属于大样本。

总体中包含个体的数目可以是有限的，也可以是无限的。包含有限个体的总体称为**有限总体**。包含无限个个体的总体称为**无限总体**。

二、标志和标志值

在生物统计中，经常是研究总体中个体的某个特征，如研究某池鲢鱼的体重、体长。其中体长、体重都是标志，当标志可以用数量表示时，称为数量标志，个体在某项数量标志上的具体值称为标志值。例如，把某池的鲢鱼体长作为标志，随机抽测10尾鲢鱼的体长(单位：cm)为：14.8，14.9，14.5，14.4，14.1，14.6，15.3，15.1，13.9，14.7。这10个体长数据就是体长的标志值，这10个标志值就构成了鲢鱼体长的样本数据资料。

如果总体中个体的某种标志不能用数量表示，则称这种标志为非数量标志或称质量（属性）标志。如鱼虾贝藻的品种、对虾的雄雌、鱼病的好转与痊愈等都是属性标志。对于属性标志在实际研究中，都是首先利用统计次数法或评分法把它们化为数量标志后再进行研究。

水产生物统计方法，就是从样本数据资料出发，应用概率和数理统计数学原理，对总体中的个体在某项标志上的标志值进行整理分析，以作出对总体的估计或推断，掌握总体的一般规律。

三、常用抽样方法

1．简单随机抽样法

怎样从总体中抽取个体组成样本呢？一般都是采用简单随机抽样的办法。这种抽样方法的主要特点就是随机性，即总体中各个体被抽中的可能性完全相等。为了保证这一点，人们的主观意志完全排除，因此称简单随机抽样为纯随机抽样。具体的方法有抽签法、查随机表法和经验数据法，这些方法已在《水产科学试验设计方法》（简称“设计方法”）第二章介绍过，此地从略。

2．机械抽样法（又叫等距抽样）

这种抽样方式是先将总体的各单位，按某一有关标志排队，然后取相等间距抽取调查单位。例如，某总体含有500个个体，想抽50个作为样本，则可先将总体的个体编号（可明编也可暗编）为1，2，……500，抽取号码为10，20，30，……500的个体组织成样本，这就是等距的机械抽样法。又如，对于某地养殖滩涂的底质调查，可沿潮位线方向，每隔20米定一个调查点，这也属于等距机械抽样。

机械抽样法的缺点是，当总体存在周期性变化时，则可能得到标志值偏小或偏大的样本。

3．整群抽样法

整群抽样法，是先把要研究的总体分为若干群（组），以群为单位进行抽样调查，对抽到的群再作全面（或抽样）调查。例如，全国淡水鱼亩产量调查，可以分级进行，可先抽查有代表性的若干个省，在抽到的省中再抽若干个县，在抽到的县中再抽若干个区，进行亩产实测，这就是三级整群抽样。

在水产资源调查中，对大的水面或滩涂进行整群抽样时，应当注意群（组）分布的均匀性，为此，多采用五点式、对角线式和平行线式等。

4．分等按比例抽样法（分层按比例抽样）

这种方法是先按一定标志把总体分为若干层，再按随机原则抽取一定比例的个体组成样本。在对总体构成有一定了解时采用此法较好。例如，对于一个大型混养成鱼池进行估产，已知该池混养的主要品种和放养尾数比例是：白鲢70%，花鲢10%，团头鲂15%，鲤鱼5%。计划抽80尾检查

估计产量，用分等按比例抽样法，即白鲢要抽56尾，花鲢要抽8尾，团头鲂要抽12尾，鲤鱼要抽4尾，各类鱼分别求出平均体重，再依各类鱼放养的总量进行估产就比较准确。

§2 样本特征数

在《试验设计方法》一书的第六章中，我们已经介绍了间断性和连继性样本数据资料的整理方法，并列出了频率分布表和频率分布图。从表和图中我们不仅知道了在这个样本中随机变数取了哪些数值，而且可以看到取得这些数值的机会多寡的情况。这就是说样本的频率分布可以粗略的反映了总体分布的情况。但是，样本的频率分布并不能完全地代替总体分布规律。

在水产生产或水产科学研究中，有时只需用个别数字刻划总体的某些特征就可以了。刻划总体数量特征和规律性的数字称为总体特征数。相应地刻划样本数量特征或规律的数字称为样本特征数。本节阐述样本特征数，下节介绍总体特征数。

一、样本平均数

1．算术平均数

设样本中各个体的标志值分别为X_1，X_2……X_n，则

$$\bar{X}=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i\cdots\cdots(1-1)$$

称为样本算术平均数。

例1 抽查某场的十个成鱼养鱼池折合亩产量(公斤/亩)分别是：482，493，457，471，510，446，435，418，

394，469。试求该场成鱼亩产的算术平均数。

解：由（1－1）式知该场成鱼亩产的算术平均数为：

$$\overline{X}=\frac{1}{10}\sum_{i=1}^{10}X_i=\frac{1}{10}〔482+493+\cdots+469〕$$

$$=457.5（公斤／亩）$$

很显然，$\overline{X}$表示X_1，X_2，$\cdots X_n$的平均水平。例1中某场十个成鱼池的平均亩产量是457.5公斤。

2．加权平均数

若样本的标志值X_1，X_2，$\cdots X_m$出现的频数分别为f_1，f_2，$\cdots f_m$，$(\sum_{i=1}^{m}f_i=n)$，

则

$$\overline{X}=\frac{X_1f_1+X_2f_2+\cdots+X_mf_m}{f_1+f_2+\cdots+f_m}=\frac{1}{n}\sum_{i=1}^{m}X_if_i\cdots(1-2)$$

称为样本的加权平均数。

当样本的标志值.整理出分组材料时，如果用组中值代表各组的值，在计算样本平均数时，就要用到加权平均数的计算公式（1－2）。

例2　从某苗种场待售的一批鲢鱼种中，随机抽取50尾鱼种组成样本，测得各尾鱼的体重如下：（单位：g）

22.3	21.3	19.2	16.6	23.1	23.9	24.8	26.4	26.6	21.3
24.8	23.9	23.2	23.3	21.4	19.8	18.3	20.0	21.5	21.0
18.7	22.4	26.6	23.9	24.8	18.8	27.1	20.6	25.0	23.6
22.5	23.5	23.9	25.3	23.5	22.6	21.5	20.6	25.8	24.5
24.0	23.5	22.6	21.8	20.8	19.5	20.9	22.1	22.7	23.6

整理成分布表如下表1—1。

表1—1　50尾鲢鱼种体重频率分布表

组　限	组中值X_i	频数f_i	频率（%）
15—17	16	1	2
17—19	18	3	6
19—21	20	8	16
21—23	22	14	28
23—25	24	17	34
25—27	26	6	12
27—29	28	1	2
总　和		50	100

试计算样本的平均数。

解：①资料未分组时，由公式（1—1）得

$$\bar{X}=\frac{1}{50}(22.3+21.3+\cdots+23.6)=22.59$$

②材料分组后，用组中值代表各组内体重数值，用公式（1—2）得

$$\bar{X}=\frac{1}{50}(16\times 1+18\times 3+20\times 8+\cdots+28\times 1)$$

$$=22.6$$

由①、②看，两种计算结果相差很小，可以忽略不计。因此，当样本进行分组整理后，平均数的计算可用计算加权平均

数的公式（1—2）进行。

显然，加权平均数只是算术平均数的一种缩写形式。为什么叫加权平均数呢？这是因为在公式（1—2）中，f_i是发生的频数，也叫x_i的“权”数，这表示x_i在计算中所占的比重不同，以频数为权数的平均数就是加权平均数。算术平均数与加权平均数在记法上都采用$\bar{X}$。当数据分组整理时，$\bar{X}$指加权平均数，如果未作分组整理，$\bar{X}$就指算术平均数。

平均数$\bar{X}$刻划了X_1，X_2，…X_n的平均水平，也就是样本频率分布的中心值。例1中鲢鱼的平均亩产量$\bar{X}=457.5$公斤。例2中鲢鱼种的平均体重$\bar{X}=22.6$（g），都是样本频率分布的中心值。因此，样本特征数$\bar{X}$是样本集中性测定的一个标志。

3．算术平均数的两个基本性质

① $\sum_{i=1}^{n}(X_i-\bar{X})=0$，即离均差的总和为0。

证：$\because \sum_{i=1}^{n}(X_i-\bar{X})=\sum_{i=1}^{n}X_i-n\bar{X}$

$=\sum_{i=1}^{n}X_i-n\cdot\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=0$ （证完）

② 离均差的平方和$\sum_{i=1}^{n}(X_i-\bar{X})^2$最小。

即，对于任何一个异于$\bar{X}$的数值α，都有：

$$\sum_{i=1}^{n}(X_i-\bar{X})^2 < \sum_{i=1}^{n}(X_i-\alpha)^2$$

证：$\because \sum_{i=1}^{n}(X_i-\alpha)^2 = \sum_{i=1}^{n}\left[(X_i-\bar{X})+(\bar{X}-\alpha)\right]^2$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2+2(\bar{X}-\alpha)\sum_{i=1}^{n}(X_i-\bar{X})+n(\bar{X}-\alpha)^2$$

而由①知 $\sum_{i=1}^{n}(X_i-\bar{X})=0$

$$\therefore \sum_{i=1}^{n}(X_i-\alpha)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\alpha)^2$$

而 $\bar{X}\neq\alpha$，故 $n(\bar{X}-\alpha)^2>0$，

因而有 $\sum_{i=1}^{n}(X_i-\bar{X})^2 < \sum_{i=1}^{n}(X_i-\alpha)^2$（证完）

二、中位数和众数

刻划样本的特征，有时也用中位数和众数。所谓中位数就是样本标志值排成依次表后，中间的那一个数。若样本的容量是偶数时，则中位数就是中间两个数的平均值。当样本的频率分布左右对称时，中位数就是算术平均数。当样本的标志值有极端数据时，中位数比算术平均数有较好的代表性。

所谓众数，是指在样本数据中出现次数最多的数值，在分组材料中常用频数最多的那一组的组中值作为众数。如表

1—1中，第五组的频数最多，它的组中值24g就是50尾鲢鱼种体重的众数。当样本的频数分布左右完全对称时，算术平均数、中位数和众数三者就一致了，否则三者是有差异的。

三、样本的方差和标准差

我们先看一个例子。分别在甲、乙两池各抽鲢鱼苗7尾，甲池的体重（单位：g）为：8，8，9，10，11，12，12；乙池的体重（单位：g）为：5，6，8，10，12，14，15。显然，这两池抽样所得的样本平均数是相等的，即$\bar{X}_{甲}=\bar{X}_{乙}=10g$。但甲池的鱼苗体重比较均匀，乙池的鱼苗体重大小不一参差不齐。因此，只用一个平均数来刻划样本的特征是不够的，还需要有刻划样本变异程度（离中性）的特征数。这里介绍的样本方差和标准差就是这样的特征数。

要刻划X_1，X_2，…X_n的变异程度，就要寻找一个比较标准，这就是频率分布的中心值$\bar{X}$。我们把离均差的平方和除以样本容量n作为刻划样本数值X_1，X_2，……X_n变异程度的特征数，并把它称为样本方差，其平方根称为样本标准差。即

设样本各个体的标志值为X_1，X_2，…X_n，则

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \cdots\cdots\cdots (1-3)$$

称为样本方差。

$$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} \cdots\cdots (1-4)$$

称为样本标准差。样本标准差S与X_i具有相同的单位。

在实用中一般由公式（1—3）和（1—4）算得的样本方差和标准差往往偏小，在水产生物统计中常用修正的样本方差和标准差，即

$$S^2_{\cdot}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \cdots\cdots\cdots\cdots (1-5)$$

$$S_{\cdot}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad \cdots\cdots\cdots\cdots (1-6)$$

在统计学上$n-1$称为离均差平方和$\sum_{i=1}^{n}(X_i-\bar{X})^2$的自由度，记作$df=n-1$。本书今后使用中所提到的样本方差和标准差就是指公式（1—5）和（1—6）式，为书写方便而略“※”号，即仍用S^2作样本方差，它的平方根S称为样本标准差。

四、样本标准差计算方法

1．未分组时样本标准差的计算方法

在利用公式（1—6）计算样本标准差时，离均差的平方和$\sum_{i=1}^{n}(X_i-\bar{X})^2$直接运算不太方便，而且利用中间运算结果$\bar{X}$往往会增大运算误差。下面给出直接用原始数据$X_1$，$X_2$，……$X_n$计算离均差平方和及样本标准差的公式。

$$\because \quad \sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}(X_i^2-2\bar{X}X_i+\bar{X}^2)$$

$$=\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2$$

$$=\sum_{i=1}^{n}X_i^2-2\left(\frac{\Sigma X}{n}\right)\cdot\sum_{i=1}^{n}X_i+n\left(\frac{\Sigma X}{n}\right)^2$$

$$=\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\Big/n$$

将它代入公式（1—6）得：

$$S=\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\Big/n\right]} \quad \cdots\cdots(1-7)$$

例3　从同一个网箱养的罗非鱼中抽50尾测其体长资料如下：（单位cm）

11.7	11.2	10.8	7.2	11.1	9.4	9.3	8.0	12.5	12.2
13.6	10.2	7.2	4.2	7.6	13.0	10.5	11.1	8.5	10.6
12.0	10.0	5.1	10.3	11.3	9.5	9.5	7.1	10.2	8.2
7.0	14.2	11.6	8.7	6.6	10.9	8.4	5.9	10.9	10.3
12.6	14.5	8.5	10.5	9.0	9.7	11.5	8.7	9.1	10.2

试计算体长平均数和标准差。

解：此地$n=50$

$$\because \quad \sum_{i=1}^{50} X_i^2 = 11.7^2+13.6^2+\cdots\cdots+10.2^2 = 5063.36$$

$$\sum_{i=1}^{50} X_i = 11.7+13.6+\cdots\cdots+10.2 = 491$$

$$\therefore \quad \bar{X} = \frac{1}{50}\times 491 = 9.82\ (\text{cm})$$

$$S = \sqrt{\frac{1}{49}\left[5063.36-(491^2/50)\right]} = 2.221(\text{cm})$$

2．分组材料计算样本标准差的方法

当样本数据作分组整理时，计算样本标准差的公式为：

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{m} f_i(X_i-\bar{X})^2}$$

$$= \sqrt{\frac{1}{n-1}\left\{\sum_{i=1}^{m} f_i X_i^2 - \left(\sum_{i=1}^{m} f_i X_i\right)^2/n\right\}} \quad \cdots\cdots(1-8)$$

其中：m为分组数，f_i为频数，X_i为组中值。n为样本容量。

例4　对例3中，50尾罗非鱼的体长作分组整理，绘出频率分布图，并计算样本平均数和标准差。

解：①作数据分组整理如表1—2

表 1—2　50尾罗非鱼体长频率分布表

组　限	组中值X_i	频率f_i	频　率(%)
3 — 5	4	1	2
5 — 7	6	3	6
7 — 9	8	12	24
9 — 11	10	21	42
11 — 13	12	9	18
13 — 15	14	4	8
Σ		50	100

②绘频率分布直方图

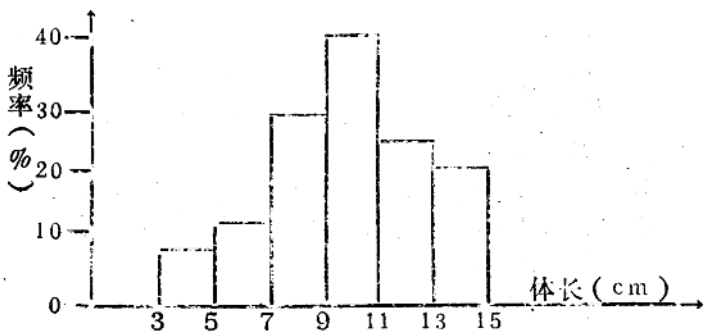


图 1—2　50尾罗非鱼体长频率分布图

③计算样本平均数与标准差