生物统计

常振江 编著

内容提要

全书共设九章，对统计理论、统计技术方法及统计方法的应用等做了较详细的阐述。同时结合实际选编了大量的例题和习题。

本书可作为高等师范院校生物专业的教材，也可供工、农、医、卫等有关专业人员参考使用。

生 物 统 计

常振江 编著

特约责任编辑 徐 新

*

学 术 期 刊 出 版 社 出 版

北京海淀区学院南路86号

辽宁师范大学印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

*

1988年12月第1版 开本：787×1092 1/32

1988年12月第1次印刷 印张：9 1/5

印数：0001—3000 字数：210 000

ISBN 7-80045-291-3/Q·9

定价：4.00元

前　言

生物统计方法已经成为现代生物科学试验的重要手段之一。随着生产和科学试验的发展，人们迫切需要学习一些生物统计技术知识，为满足广大读者的需要，我们编写了《生物统计》这本书。本书是在辽宁师范大学生物系本科生使用的《生物统计》讲义的基础上，结合教学实践，经修改补充而成的。

全书分两大部分，共九章三十九节。第一部分主要介绍了生物统计基础理论，设五章，计21节；第二部分主要讨论了常用统计技术方法，设四章，计18节。各章均附有习题，希望有助于提高教学效果。

本书在编写过程中，曾广泛参照了各兄弟院校及有关专家学者的有关教材和论著，并得到辽宁师范大学校、生物系、数学系领导和生物系鄂永昌教授的热情指导和大力帮助。鄂永昌教授阅读过全部书稿，提出了宝贵的修改意见；数学系不少老师对本书的编写工作也提出了不少有益的建议。书中的图形由刘玉珍同志编制。对于他们的热情支持和帮助，谨此一并表示衷心的感谢。

本书可作为高等师范院校生物专业的教材，也可作为工、农、医、卫等专业人员的学习用书。由于编者的水平有限，错误和不当之处在所难免，尚祈广大读者批评指正。

编　著　者

1988年7月

目　录

第一章　生物统计数据的简单分析

要对事物在数量上有客观的认识，研究其量的关系，唯一可靠的方法就是以有效的方式收集、整理试验数据，然后进行分析，对所观察的问题作出推断、预测。本章叙述的计算平均数、标准差、中位数、组织图等，是生物统计中数据分析的简单方法，为下述几章的概念提供直观的认识。

§1　数据集中性的度量——均值

设给一组数据 $x_1, x_2, \cdots, x_n$，我们把

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

叫做这组数据的算术平均数，简称为平均数或均值。为什么要求均值呢？因为它能集中反映这组数据的基本情况，或者说它具有代表性。

现在，我们来讨论均值的代表性问题。对给定的 n 个数 $x_1, x_2, \cdots, x_n$，$x_i - C$ (C为给定的常数)反映了 C 偏离 x_i 的程度。$x_i - C$ 可能为正或负，为了消除符号的影响，用 $(x_i - C)^2$ 来衡量 C 与 x_i 的偏离程度。显然，C 与“$x_1, x_2, \cdots, x_n$”偏离程度最小应使 $\sum_{i=1}^{n} (x_i - C)^2$ 达到最小。而均值 $\bar{x}$ 恰是符合这一要求的唯一的数。换句话说，$\bar{x}$ 是“$x_1, x_2, \cdots, x_n$”最具有

代表性的值。为了说明这一点，我们需要先证

公式 1
$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \qquad (1\cdot1\cdot1)$$

证： 由 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

即有

$$n\bar{x}=\sum_{i=1}^{n}x_i=(x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})+n\bar{x}$$

移项就有

$$(x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})=0$$

(1·1·1)式得证·

公式 2　任给一常数 C，成立等式

$$\sum_{i=1}^{n}(x_i-C)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-C)^2 \qquad (1\cdot1\cdot2)$$

证：
$$\begin{aligned}\sum_{i=1}^{n}(x_i-C)^2&=\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-C)^2\\&=\sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-C)\sum_{i=1}^{n}(x_i-\bar{x})\\&\quad+n(\bar{x}-C)^2\\&=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-C)^2\end{aligned}$$

从(1·1·2)式可以看出

$$\sum_{i=1}^{n}(x_i-C)^2\geqslant\sum_{i=1}^{n}(x_i-\bar{x})^2$$

且只有在 $C=\bar{x}$ 时才达到最小值 $\sum_{i=1}^{n}(x_i-\bar{x})^2$·　这就说明了

均值 $\bar{x}$ 是 "$x_1, x_2, \cdots, x_n$" 最具有代表性的值，它是表示数据集中性的一个量。

计算均值 $\bar{x}$ 有一些常用的方法，我们给出以下公式计算 $\bar{x}$. 这 n 个公式在统计学中处理数据时，有简化计算的作用。

若 $y_i = x_i - a$, $i = 1,2,\cdots,n$, 则

$$\bar{y} = \bar{x} - a \tag{1·1·3}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1·1·4}$$

证： $$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} x_i - na$$

两边同除 n，即得

$$\bar{y} = \bar{x} - a$$

又

$$y_i - \bar{y} = x_i - a - (\bar{x} - a) = x_i - \bar{x}$$

两边平方求和就得(1·1·4)式。

若 $y_i = bx$, $i = 1,2,\cdots,n$, 则

$$\bar{y} = b\bar{x} \tag{1·1·5}$$

$$\sum_{i=1}^{n}(x - \bar{x})^2 = b^{-2}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1·1·6}$$

证： $$\sum_{i=1}^{n} y_i = b\sum_{i=1}^{n} x_i$$

两边同除 n，即得

$$\bar{y} = b\bar{x}$$

又

$$y_i - \bar{y} = b(x_i - \bar{x})$$

两边平方求和，即

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

所以

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=b^{-2}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

若 $y_i=bx_i-a$，$i=1,2,\cdots,n$，则

$$\bar{y}=b\bar{x}-a,\ \sum_{i=1}^{n}(x_i-\bar{x})^2=b^{-2}\sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad (1\cdot1\cdot7)$$

注意到上面的证明，即可得(1·1·7)的二式。

在求 $x_1,x_2,\cdots,x_n$ 的均值时，如果数据 $x_i(i=1,2,\cdots,n)$ 中有相同的，自然可以合并。不妨设不同的只有 k 个值 a_1，$a_2,\cdots,a_k$，并且 $a_i(i=1,2,\cdots,k)$ 出现了 n_i 次($n_1+n_2+\cdots+n_k=n$)。这时

$$x_1+x_2+\cdots+x_n=n_1a_1+n_2a_2+\cdots+n_ka_k$$

因此

$$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$$

$$=\frac{1}{n}\sum_{i=1}^{k}x_ia_i=\sum_{i=1}^{k}a_i\frac{n_i}{n} \qquad (1\cdot1\cdot8)$$

例 有一个班级(31人)，在一次数学考试中成绩如下表(表1·1)

表1·1

成绩(a_i)	5	4	3	2
人数(n_i)	4	13	11	3

求出这次考试的平均成绩。

我们可以把每个学生的成绩加起来，再被班级总人数

除，即使用 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，求出 $\bar{x}=3.58$。现在，也可以这样来求平均成绩：

$$\bar{x}=\sum_{i=1}^{4}a_i\frac{n_i}{n}=5\times\frac{4}{31}+4\times\frac{13}{31}+3\times\frac{11}{31}+2\times\frac{3}{31}=3.58$$

从(1·1·8)式可以看到，均值 $\bar{x}$ 与 a_i 的关系是由比值 n_i/n 决定的，就是说 $\bar{x}$ 与 a_i 的关系与 a_i 在全部数据中所占的比例有关，比值 n_i/n 越大，$\bar{x}$ 受 a_i 的影响就大，n_i/n 越小，$\bar{x}$ 受的影响就小。

把(1·1·8)式的意义加以推广，就得到加权平均数的概念：

给定一组数据 $x_1, x_2, \cdots, x_n$，又给了一组正数 $p_1, p_2, \cdots, p_n$，且 $\sum_{i=1}^{n}p_i=1$，则

$$p_1x_1+p_2x_2+\cdots+p_nx_n$$

就叫做 $x_1,x_2,\cdots,x_n$ 的加权平均数，$p_1,p_2,\cdots,p_n$ 称为 $x_1,x_2,\cdots,x_n$ 相应的权。当权 $p_1=p_2=\cdots=p_n=1/n$ 时，加权平均数就是算术平均数。“权”p_i 就是衡量 x_i 在数据 $x_1,x_2,\cdots,x_n$ 平均时的重视程度，因此，在实际问题中如何合理地决定权是有重要意义的问题。

§2 离中性的度量——标准差

在上一节中我们讨论了对给定的一组数据，其均值 $\bar{x}$ 的代表性，即在 $\sum_{i=1}^{n}(x_i-C)^2$ 中当 $C=\frac{1}{n}\sum_{i=1}^{n}x_i$ 时达到最小值。本节进一步讨论数据 $x_1, x_2, \cdots, x_n$ 与 $\bar{x}$ 的偏差问题，

给出度量方法。

对给定的一组数据 $x_1, x_2, \cdots, x_n$ 来说

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

就叫做数据 $x_1, x_2, \cdots, x_n$ 的方差，它的算术平方根叫做标准差，分别记作 S^2 和 S。即

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \qquad (1\cdot2\cdot1)$$

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} \qquad (1\cdot2\cdot2)$$

一组数据的标准差 S 越大，这组数据就越“分散”，或者说这组数据的变异性（即相互不同的程度）就越大；S 越小，这组数据的变异性就小，也就更“集中”。当 $x_1=x_2=\cdots=x_n$ 时，$S=0$。标准差 S 是表征数据离中性（偏离数据均值 $\overline{x}$）的一个量。

对一组数 $x_1, x_2, \cdots, x_n$ 进行分析时，$\overline{x}$ 与 S（或 S^2）是最常用的两个量，一个是代表性的值，一个是描述数据变异性的值。在实际应用中，我们可以利用本章§1中介绍的几个计算 $\overline{x}$ 与 S^2 的简化公式进行计算。

例 1 从调查资料得到健康人血清粘蛋白含量 (mg/100 ml) 与矽肺病人血清粘蛋白含量数据如下表(表1·2)

表1·2

健康人	42.84	48.19	48.19	52.48	58.90	64.26	69.61	80.22
二期矽肺病人	65.45	69.63	69.73	74.97	80.44	80.44	95.20	96.39

试求出健康人和二期矽肺人这两组数据的均值和方差。

解：我们用 $x_1, x_2, \cdots, x_8$ 表示健康人的八个数据，将 x_i 均减去50后乘以100，即得 $y_i=100(x_i-50)$，就不需要用小数来表示了。于是得表1·3。

表1·3

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
−716	−181	−181	248	890	1426	1961	3022

$$\sum_{i=1}^{8} y_i=6469 \qquad \bar{y}=\frac{1}{8}\times 6469=808.63$$

于是

$$\bar{x}=\frac{\bar{y}}{100}+50=58.0863$$

将 x_i 逐个减去 $\bar{x}$，然后平方求和，可算得 $S^2=160.17$，标准差 $S=12.66$.

在(1·1·2)式中，取 $C=0$，得

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2 \qquad (1\cdot 2\cdot 3)$$

这是在计算数据的方差时常常用到的一个公式。我们用(1·2·3)式再计算例1的方差：

$$\sum_{i=1}^{8} x_i^2=28113.326$$

$$n\bar{x}^2=8\times 3374.0182=26992.146$$

$$\sum_{i=1}^{8}x_i^2-n\bar{x}^2=1121.180$$

$$S^2=1121.180/7=160.17$$

$$S=12.66$$

与刚才那种直接计算的结果相同。(1·2·3)式是一个在统

计中很有用的公式。

对矽肺病人的数据进行类似的计算，就得到均值是79.03，方差是134.88，标准差是11.61。

从数据上看矽肺人的均值高，方差小，健康人的均值小，方差大，数据更分散些。

现在，我们可以理解到(1·1·4)～(1·1·7)式的含意是：

(1) 若将各数据加同一常数 a，则均值也加同一常数 a，而方差不变。

(2) 若将各数据同乘一常数 b，则均值也乘同一常数 b，而方差则乘以 b^2。

描述数据之间变异性，也可以用别的量，如下面的一些量：

1. 极差 $R=\max\{x_1,x_2,\cdots,x_n\}-\min\{x_1,x_2,\cdots,x_n\}$。式中 $\max\{x_1,x_2,\cdots,x_n\}$ 和 $\min\{x_1,x_2,\cdots,x_n\}$ 分别表示 x_1，x_2，…，x_n 中最大值和最小值。它反映了数据之间最大的距离是多少。

2. 变异系数 x_1，x_2，…，x_n 的方差 S^2 或标准差 S 都是有单位的量。单位不同时不好比较，为了消除单位不同的影响，考察相对的变异性，这时采用变异系数

$$CV=S/|\bar{x}|$$

来表达相对波动的大小。式中对均值取绝对值是为了使变异系数只取正值。

例 2 用例 1 的数据算出相应的极差与变异系数。

解：对于正常人

极差 $R=80.22-42.84=37.38$

变异系数 $CV=12.66/58.0863=0.2180$

对于矽肺病人

$R=96.39-65.45=30.94$

$CV=11.61/79.03=0.1469$

从极差和变异系数来看，也是健康人的变化幅度大，矽肺病的数据比较集中。

3. 中位数 将数据 $x_1, x_2, \cdots, x_n$ 按大小次序排列，当 n 是奇数时，居中的一个就是；当 n 是偶数时，居中的两个取均值就是中位数。

§3 组织图、累积次数图

在生物统计中常常要对实际观测所得到的数据进行分析、研究，从而推断研究对象的某些性质，首先我们要对这些数据加以整理。这一节介绍的组织图(直方图)、累积次数图是整理、分析数据的常用方法。

在数据很多(一般数据不少于30)时，为了简便起见，可以把数据分组整理。标明落入第 i 组数据的个数，记为 n_i，叫做第 i 组相应的频数。全部数据 n 叫做总频数。根据分组的间隔，在第 i 组相应的间隔上画一矩形，其面积就是该组的相对频数(频数/总频数)。这样的图形就叫做组织图。

我们结合下面的实例，讨论如何绘制组织图。

例 1 从某块地上随机抽取50株树本，观测这50株树的高，得到下列数据：

22.3	21.2	19.2	16.2	23.1	23.9	24.8	26.4	26.6
24.8	23.9	23.2	23.3	21.4	19.8	18.3	20.0	21.5
18.7	22.4	26.6	23.9	24.8	18.8	27.1	20.6	25.0
23.5	23.9	25.3	23.5	22.6	21.5	20.6	25.8	24.0
22.3	25.6	21.8	20.8	19.5	20.9	22.1	22.7	23.6
24.5	23.6	21.0	21.3	22.5				

本例数据最大值为27.1，最小值为16.2.

决定组距和组数　在数据比较多时，通常分成 10～20 组，数据少于50时分成5～7组。本例确定分 7 组。

每一组的区间长度叫做*组矩*。分组时一般要求各组的组距相等。连续型变量所求得的组距不一定是整数，为了便于计算可以采用整数作为组距。组距的大小是根据全距与组数的关系决定的：组距＝全距/组数。

$$\text{本例组距}=\frac{27.1-16.2}{7}=1.56\doteq 2\ .$$

求组中值　组中值是每组的上限与下限的中间值。即

$$\text{组中值}=\frac{\text{组上限}+\text{组下限}}{2}$$

确定组限。在分组时为了避免第一组中变数过多，一般第一组的组中值最好取接近或等于资料中的最小值。这样，在本例中第一组的下限定为15，第一组的下限确定后，其余各组的组限可由以下若干个不重叠的子区间

$$[15,17)、[17,19)、\cdots、[25,27)、[27,29)$$

确定。此处用半开区间是使得各个子区间没有公共点。这种整理方法叫*上限排外法*。也可用下限排外法，不管是上限排外或下限排外，都要按统一办法来处理。

在分组后所得的实际组数，有时和最初确定的组数不同。如第一组下限和数据中最小值相差较大或实际组距比计算的组距为小时，那么实际分的组数将比原定组数为多，反之则少。出现这种情况稍加调整并不影响分组或计算。

整理时，可用划计法在划记表(表1·4)上进行。

表1·4

树高分组（单位：米）	划记
15—17	一
17—19	下
19—21	正 下
21—23	正 正 丅
23—25	正 正 正 一
25—27	正 丅
27—29	一

划记表

例如第1株树高是22.3，就在21—23的一组中划上一笔，如此类推。恰好是25.0的数据应归入25—27这一组，这是由上限排外法决定的。

我们用每一组的组中值来代表属于这一组的各个观测值。这自然要有些误差，组数越多（即组距越小）误差也越小，但组数很多就达不到简化计算的目的。下面用表1·5表示各组的组频数(表1·4中的划记数)和相对频数，各子区间分别用其组中值来代表。

现在作组织图。在横轴上标出所有的子区间，以每一子区间为底，f/nh为高作矩形，就得到组织图。其中f是组频数，h是组距，n是总频数，矩形的面积等于