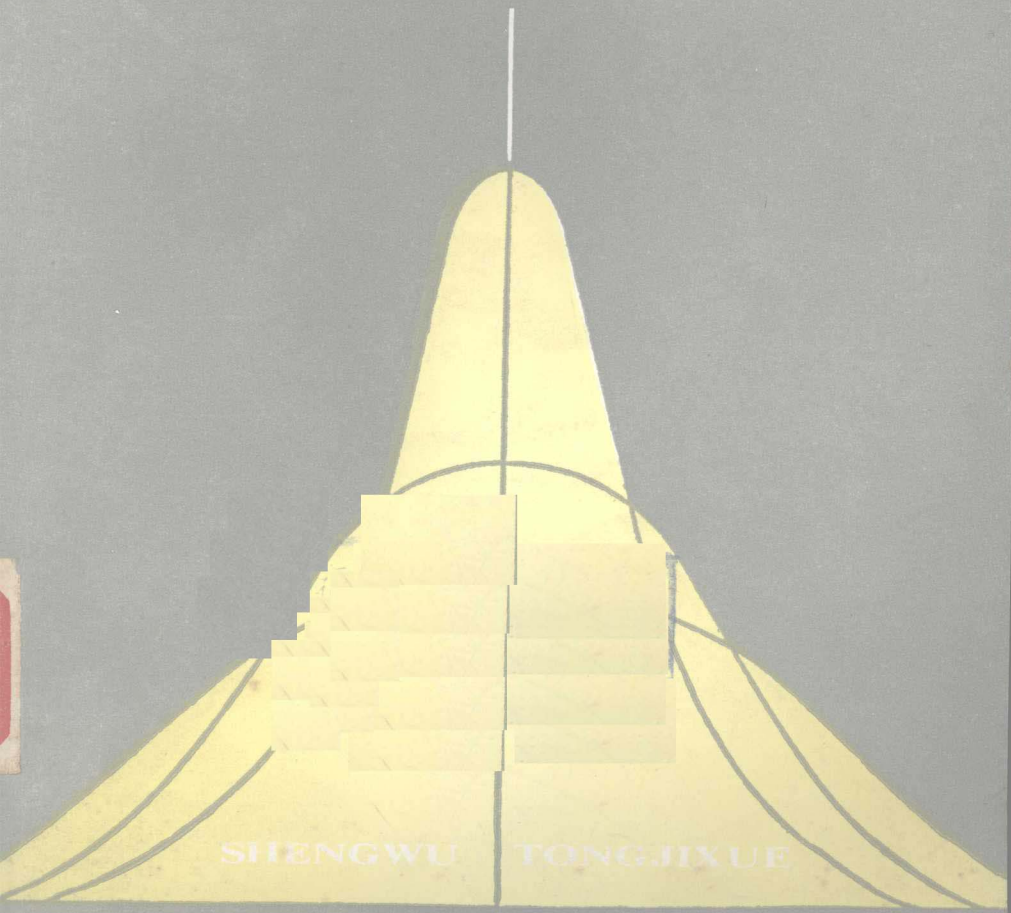
杜荣骞 编

高等学校试用教材

生物统计学



高等教育出版社

高等学校试用教材

生物统计学

杜荣骞 编

高等教育出版社

内 容 提 要

本书是作者在多年教学实践中修改而成。内容包括数据整理、统计推断、方差分析、回归分析和实验设计等。本书可供综合大学和师范院校生物系以及农林医院校有关专业师生使用，也可供有关科研工作者参考。

责任编辑：张念辉

高等学校试用教材

生 物 统 计 学

朴荣森 编

高等教育出版社出版
新华书店北京发行所发行
河北省香河县印刷厂印装

开本850×1168 1/32 印张16 字数380,000
1985年10月第1版 1987年8月第3次印刷
印数 10,321—15,330
书号 13010·01146 定价 2.85元

前　言

今天，生物学的发展已进入了定量地研究生命现象的阶段。需要运用数理统计学原理，分析和解释生物学上的数量变化，以正确设计试验及正确处理试验结果，从而推导出较为客观的结论。因此，广大生物学工作者和高等院校师生迫切希望有一本适合生物系学生使用的生物统计学教科书。

本书是作者在1974年以来历年教学讲义的基础上修改而成。内容侧重于各种统计方法在生物学中的应用，而不强调各种公式的严格推导。但是尽可能地将各种公式的来龙去脉交待清楚，使读者能够对统计学的基本原理有一全面的了解。全书共分九章，可在54学时内讲完(每周3学时)。在此期间，可组织同学用5—6学时做一次抽样试验(习题3.15)和一次模拟两因素实验方差分析(习题5.7)练习。书中列举了生物学中经常遇到的大量实例，每章之后附有一定数量习题，可供练习。由于篇幅所限，所举的例子不可能面面俱到。实际上，在生物学的各个领域中，甚至包括医学和农学在内，统计学的原理及方法并无很大差异，只要掌握统计学的基本原理，对于各种类型的问题都可得到解决。

书中的例子，多数是作者近年来所接触到的实际问题，也有部分例子，选自其他作者编写的书刊，在此深表谢意。

本书在编写过程中，得到了周概容同志的热情指导和王方慎同志的大力协助，在此谨致谢忱！

由于作者的学识所限，错误在所难免，希望使用本教材的教师、学生和生物学工作者们提出宝贵意见。以便再版时订正。

编　者

1981年2月

目　录

第一章 数据整理

§1.1 频数分布 (freguency distribution)

1.1.1 连续型数据 (continuous data) 和离散型数据 (discrete data)

统计学的最基本工作是收集数据。把原始数据收集上来之后，首先要整理和分析这些数据的特性和变化规律。生物统计学中经常遇到的数据有两种类型，一种是连续型数据，一种是离散型数据。

与某种标准做比较所得到的数据称为**连续型数据**。又称**度量数据** (measurement data)。例如，长度、时间和重量。这类数据通常是非整数，虽然有时记载的是整数，如身高的厘米数，但是当提高精确度后，总会出现小数。对连续型数据的分析方法，通常称为**变量的方法**(method of variable)。

由记录不同类别个体的数目所得到的数据，称为**离散型数据**。又称**计数数据**(count data)。例如，某一类别动物的头数，具有某一特征的种子粒数，血液中不同类型的细胞数目等。所有这些数据全都是整数，而且不能再细分，也不能进一步提高它们的精确度。离散型数据的分析方法，通常称为**属性的方法** (method of attribute)。

在判断数据的类型之后，就要进一步研究数据的变化规律。描述数据变化规律的最简单方法是将这些数据列成**频数表** (frequency table)或绘成**频数图**(frequency graph)，根据频数分布进行研究。

1.1.2 频数表和频数图的编绘

离散型数据及连续型数据的频数表和频数图的编绘方法略有不同，下面各举一例说明。

例 1.1 调查每天出生的 10 个新生婴儿中，体重超过 3 公斤的人数，共调查 120 天。每天的 10 个新生婴儿中，体重超过 3 公斤的人数，可能有 11 种情况：1 个也没有，有 1 个，有 2 个，…，10 个都是，如表 1-1 的第一列所示。表 1-1 的第 2 列是记载的调查

表 1-1 每 10 个新生儿中体重超过 3 公斤的人数的频数(率)表

组值(class value) (体重超过 3 公斤的人数)	频数计算	频 数	频 率
0		0	0.000
1		0	0.000
2		0	0.000
3	一	1	0.008
4	丅	2	0.017
5	正丅	12	0.100
6	正正正𠂉	19	0.158
7	正正正正正正正𠂉	39	0.325
8	正正正正正正𠂉	34	0.283
9	正正	10	0.083
10	丅	3	0.025
总 计		120	0.999

结果。如第 1 天调查有 6 个超过 3 公斤的，则在组值为 6 的一行做个记号，一般使用“正”字或“卌”号表示。全部调查完毕，累加各行结果，填入频数一栏。或者，将各行的结果除以总数而得出频率。把频数或频率按超过 3 公斤的人数的顺序排列起来，就得到了**频数分布**(frequency distribution)或**频率分布**(relative frequency distribution)。频数表可以比较清楚地描述数据变化规律。有时为了更直观地描述数据变化规律，还可以绘成频数图表

示(图 1-1)。图 1-1 的横轴表示每 10 个孩子中，体重超过 3 公斤的人数，纵轴表示每一组的频数。若将纵轴改为频率的话，则得到频率图。频率图与频数图的图形完全一样。

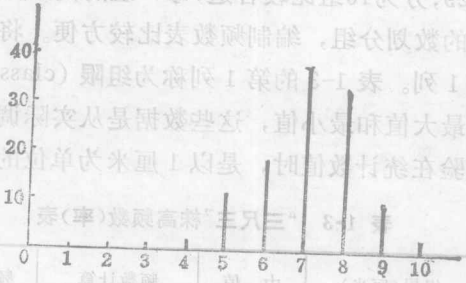


图 1-1 频数图

例 1.2 表 1-2 列出了某农场在做高粱“三尺三”提纯时，所调查的 100 个数据。从表 1-2 的原始数据中，除可以找出最大值是 170 厘米，最小值是 141 厘米以及估计出它们的平均高度大约在 150—160 厘米之外，很难再看出什么规律来。但是，当我们将表 1-2 中的数据列成频数表之后，情况就会有明显的不同。从频数表中，可以比较清楚地看出这些数据的变化规律。高粱的株高

表 1-2 “三尺三”株高测量结果

155	158	159	155	150	159	157	159	151	152
159	158	153	153	144	156	150	157	160	150
150	150	160	156	160	155	160	151	157	155
159	161	156	141	156	145	156	153	158	161
157	149	153	153	155	162	154	152	162	155
161	159	161	156	162	151	152	154	157	162
158	155	153	151	157	156	153	147	158	155
148	163	156	163	154	158	152	163	158	154
164	155	156	158	164	148	164	154	157	165
158	166	154	154	157	167	157	159	170	158

是连续型数据，不是一个个孤立的值。因此，不能象例 1.1 那样制表。连续型数据频数表的制作过程如下：首先将数据分级，一般来说，100 个数可以分 8—10 组。根据极差 $R=\max x-\min x=170-141=29$，分为10组比较合适，每一组的间距，刚好是 3 厘米。用比较简单的数划分组，编制频数表比较方便。将分好的组填入表 1-3 的第 1 列。表 1-3 的第 1 列称为组限 (class limit)，组限是这一组的最大值和最小值，这些数据是从实际调查的数据中得到的。本试验在统计数值时，是以 1 厘米为单位的。所以第一组

表 1-3 “三尺三”株高频数(率)表

组限(厘米)	组界(厘米)	中 值	频数计算	频 数	频 率
141—143	140.5—143.5	142	一	1	0.01
144—146	143.5—146.5	145	丅	2	0.02
147—149	146.5—149.5	148	正	4	0.04
150—152	149.5—152.5	151	正正丅	13	0.13
153—155	152.5—155.5	154	正正正正丅	23	0.23
156—158	155.5—158.5	157	正正正正正丅	28	0.28
159—161	158.5—161.5	160	正正正	15	0.15
162—164	161.5—164.5	163	正正	10	0.10
165—167	164.5—167.5	166	丅	3	0.03
168—170	167.5—170.5	169	一	1	0.01
总 和				100	1.00

的上限“143”厘米的实际值，可能在 142.5—143.5 厘米范围内。同样，第一组的下限“141”厘米的实际值，可能在 140.5—141.5 厘米范围内。因此，这一组的全部实际可能值是在 140.5—143.5厘米范围内。140.5—143.5 这个范围称为组界(class boundary)。对于其它各组，同样可以定出相应的组界。

中值(midvalue)是每一组的两个组限的平均值。但是也有例外。例如，习惯上通常以“岁”为计算年龄的单位。假若有一组的组限是 20—29 岁，上限 29 岁包括这个人可能刚刚 29 岁，也可能

即将进入30岁。所以这一组既包括20岁的，也包括29岁的，共有10个年龄级。因此，中值是(20+30)÷2=25，而不是(20+29)÷2=24.5。类似这种情况，在计算中值时应特别注意。

频数计算一列，就是将表1-2中的数据"对号入座"，最常用的方法是采用"唱票"的方式，一人读，一人填。最后将每一组的频数统计出来，记入频数栏，并计算出频率。在制成频数(率)表以后，同样可以看出连续型数据的频数(率)分布规律。

编制连续型数据的频数(率)表，一般需要以下各步：

1 从原始数据表中，找出最大值和最小值，并求出极差。

2 决定划分的组数，分组数是由数据的多少决定的，在数据较少时，如50—100个数，可以分为7—10组。数据较多时，可分为15—20组。

3 根据极差与决定划分的组数，确定组限。

4 在频数表中列出全部组限、组界及中值。

5 将原始数据表中的数据，用唱票的方式填入频数表中，计算出各组的频数和频率。

表1-3以表格的形式，描述了高粱品种"三尺三"的株高频数分布，还可以用频数图更直观地描述这一分布。下面讲三种最常用的频数图。

1. 直方图(histogram)

在横轴上标明各组的组界，纵轴标明频数。然后以每一组的

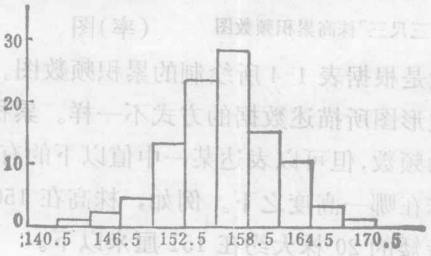


图 1-2 "三尺三"株高直方图

组界为一个边，相应的频数为另一个边，作矩形，构成直方图。见图 1-2。若纵坐标改为频率，则得到频率直方图。频率直方图与频数直方图的图形完全一样。

2. 多边形图(polygon)

在横轴上标出各组的中值，纵轴上标出频数(率)，在坐标平面内，标出相应的每个点。然后，连接各点。并且最低一组非零频数的点，应该直接与相邻的零频数中值点相连；最高一组非零频数点，亦应该与相邻的零频数中值点相连。最后得到一个多边形图。

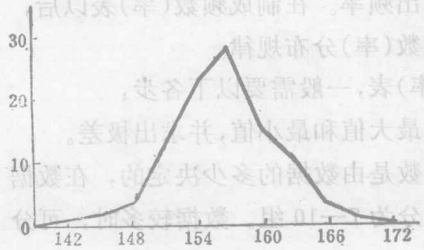


图 1-3 "三尺三"株高多边形图

3. 累积频数图(cumulative frequency graph)

经常使用的第三种频数图，称为累积频数图。作图法如下。首先根据表 1-3 制成累积频数表(表 1-4)。在横轴上标上各组的中值，纵轴上标上累积频数(率)。在坐标平面内标上对应的点。连接各点，从而得到累积频数(率)图。

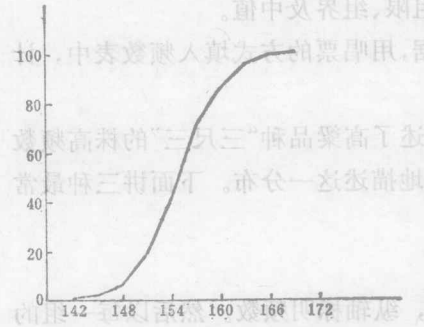


图 1-4 "三尺三"株高累积频数图

图 1-4 就是根据表 1-4 所绘制的累积频数图。累积频数图与直方图和多边形图所描述数据的方式不一样。累积频数图不能表达任何中值的频数，但可以表达某一中值以下的有多少株，以及一定数量的植株在哪一高度之下。例如，株高在 150 厘米以下的大约有 15 株，最矮的 20 株大约在 151 厘米以下。

表 1-4 “三尺三”株高的累积频数表

中　值	累积频数	中　值	累积频数
142	1	157	71
145	3	160	86
148	7	163	96
151	20	166	99
154	43	169	100

1.1.3 研究频数分布的意义

根据编绘的频数表或频数图，可以明显地看出数据的三个重要特征。

首先，根据频数(率)分布，可以看出数据的集中情况。一般来说，不论是离散型数据，还是连续型数据都有聚集于某一范围内的趋势，常常用**平均值**(average value)表示全部数据的集中点。使用最广泛的**平均**(average)是**算术平均数**(arithmetic mean)，其次是**中位数**(median)和**众数**(mode)。算术平均数是一群数据的重心所在。第二种平均是中位数，它是在累积频数图中$\frac{1}{2}$总频数位置上的数值。例 1.2 中，总频数的中点在 50—51 之间。从累积分布图中，可以找到，株高大约为 155 厘米。第三种平均是众数。离散型数据的众数是频数图中频数最高的数。连续型数据的众数是频数图中频数最高的中值。

其次，从频数表或频数图中，可以直观地看出数据的变异情况：这群数据是集中在平均数附近，还是分散在平均数的两侧。如果数据大部分集中在平均数附近，远离平均数的两侧数据比较少，则这样的数据是比较整齐的；若分布在平均数附近的数据与分布在远离平均数的两侧数据相差无几，则这样的数据是比较分散

的。

第三，从频数分布图中，还可以看出曲线的形状。例如，有些分布从零频数开始平稳地上升，直到最高频数，然后平稳地下降直到零频数。结果得到一个对称的直方图或多边形图。而另一些分布，在上升阶段可能要经过很多步，到达最高频数后，突然下降；或者相反，上升很快，下降很慢。

此外，频数表或频数图还可以显示一些不规则的情况。例如，在一个分布中，出现一个或几个频数突然高出正常频数的情况，这是一种异常分布，可能是由于条件不一致，或由于度量时的失误造成的。当出现这样一些不规则情况时，需要认真研究，尽可能找出原因。

1.1.4 总体(population)和样本(sample)

统计学的核心问题，是研究总体与样本之间的关系。因此，总体与样本，是生物统计学中的两个最基本概念。

总体是我们研究的全部对象。总体又分为**无限总体** (infinite population) 和**有限总体** (finite population)。例如，我们要研究在某种条件下生长的小麦的株高，因为无法估计出在这种条件下，生长的小麦的数量，可以设想这一总体是无限的。或者研究新生婴儿体重，因为新生婴儿是无止境的，所以这一总体也可以设想是无限的。如果我们要调查一所学校，今年新生的身高。这一总体则是有限的。生物统计学中所遇到的总体多数都是无限总体。构成总体的每个成员称为**个体**(individual)。

样本是总体的一部分，例 1.1 和例 1.2 即各为一个样本。样本内个体的数目称为**样本含量**(sample size)。抽样的目的，是希望通过对样本的研究，推断其总体。例如希望由 100 株“三尺三”高粱的株高，推断在这种条件下生长的该品种的株高。这就要求样本应能在最大程度上，代表总体的情况。为此，在从总体中抽取样本时，就应做到总体中的每一个个体被抽中的机会，都必须一

样，不能带有偏见。例如，在小麦育种工作中，我们常常希望得到矮秆品种，为了满足个人愿望，在抽样时如果多抽矮秆的，这样得到的样本就没有代表性。属于偏性抽样，不能代表总体的情况。我们需要的样本应该是一个总体的缩影。为了达到这个目的，就需要用**随机抽样**(random sampling)的方法获得样本。

随机抽样的方法很多，例如抽签、拈阄等。最好使用随机数字表(见附表1)进行抽样。现举例子说明，如何用随机数字表抽样。假设需要从包含4728个个体的总体中，抽出一个含量为20的样本。先将总体中的每一个个体从0000号编到4728号。第一步，闭上眼睛用铅笔在随机数字表上任意划点，假若点到奇数上，就用第一页表；点到偶数上，就用第二页表。第二步，在选定的那一页上，再点一次。根据点中的数字决定从哪一行读起。最后再点一次，决定从哪个字开始。决定了起点以后，开始以四位数字为一节读下去，小于等于4728的数字就被选中，大于4728的则舍弃，直到取满20个数为止。这20个数所对应的个体，即为我们选中的样本。

从一有限总体中抽样，可分为**放回式抽样**(sampling with replacement)和**非放回式抽样**(sampling without replacement)。所谓放回式抽样是指：从总体中抽出一个个体，记下它的特征后，放回总体中。再做第二次抽样。这种抽样方式可能会重复抽中某一个体。非放回式抽样是指：从总体中抽出个体后，不再放回。在上述的例子中，若保留重复的随机数字，则为放回式抽样；若舍弃重复的数字，则为非放回式抽样。对于无限总体来说，放回式抽样和非放回式抽样，实际上没有区别。

样本的含量越大越有代表性。但是，太大的样本研究起来是很困难的。因此，样本的含量必须合适。

总之，在用统计学方法进行研究时，都要问以下几个问题：

1 样本是否是随机选取的？

2 选取的样本含量是否合理？

3 所研究的对象是一个总体，还是几个总体？

§ 1.2 平均数(mean)和标准差 (standard deviation)

频数表和频数图，只能定性地描述一组数据。对于生物统计学工作来说，这种描述远远不够。为了更客观地描述这些数据，需要借助于以下三种分析工具的帮助。它们是：数据集中点的度量——**平均数**，数据变异程度的度量——**标准差**和曲线形状的度量——**偏斜度**和**峭度**。这些数字是描述样本频率分布特征的，称为**样本数字特征**或简称为**样本特征数**(sample characteristic)。

1.2.1 几种平均(average)

求平均的目的，是为了给出一个数。用这个数来描述由许多数组成的样本。如果样本中的所有的数都是一样的，那么平均值就是这个数。若样本中的数不一样，则针对不同的目的可使用不同的平均。在生物统计学中，运用最多的是算术平均数 (arithmetic mean)。样本算术平均数的符号是$\bar{x}$，读做"x杠"或"杠x"。若用x_1，x_2，…，x_n表示组成样本的所有的数，则它们的算术平均数为：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}, \tag{1.1}$$

或简写为：

$$\bar{x}=\frac{\sum x}{n}.$$

其中$\sum$是求和的符号，n为样本含量。$i=1$是相加数的下限表示从x_1开始相加，$\sum$符号上的n是相加数的上限，表示一直加到x_n。