生物统计方法

南京农学院统计遗传教研组编

南京市农业干部学校教务处印

一九八二年三月

前言

一九〇一年出版的生物统计杂志(Biometrika)第一期，最早采用“生物统计学”(Biometry)这一术语。当时著名学者Galton对生物统计学的概念为：“生物统计学是一门科学，应用了统计的现代方法于生物科学上。”八十年来，随着生产实践和科学研究的发展，特别是生物学各领域和统计学的发展，以及计算工具的不断创新，生物统计学这门年轻的学科迅速成长起来，电子计算机的应用，更使这门学科发展到一个新的阶段。现在无论是在农、林、牧、渔、医药卫生、人口理论、环境保护等各方面的科学研究和实际工作中，生物统计学都居有不可替代的重要地位。

生物统计学的主要功用有三方面：（1）试验设计或调查方案的依据和实施指导；（2）试验和调查资料的收集和整理；（3）试验或调查结果的假设测验和统计推断。掌握生物统计学的基本原理和方法，将有助于我们把科研和实际工作做得更加合理、更加准确，并以最少的代价获得最多的试验信息和研究成果。

这本讲义是为已经参加工作的农业科技人员编写的培训教材或自修读物。为适应目前的一般水平，内容力求简明实用，在阐明原理时不作过多的数学论证，也尽量少用线性代数和微积分知识。在读完每一章时，一定要演算完该章的习题，然后再阅读下一章。如作培训教材用，建议各章的讲授学时数如下，可根据学员的程度而增减。

第一章　6—8　学时

第二章　8—10　学时

第三章	8—10 学时
第四章	10—12 学时
第五章	8—10 学时
第六章	8—10 学时
合 计	48—60 学时

学员有了这部分基础后，可以继续学习全国高等农业院校的统编教材《田间试验和统计方法》。

参加本书编写的有陆作楣、翟虎渠、马泽仁等。由于我们水平有限，时间仓促，编写中一定有不少错误之处，请读者批评指正。

一九八一年十二月

目　录

生物统计方法

第一章　次数分布和平均数、变异数

在农业科学实验中，我们可以得到大量的试验资料，这些资料初看起来，往往是杂乱无章的，所以必须按照一定的程序，进行科学的整理和分析，才能透过现象看到本质，从中找出规律性的东西来。

第一节　总体和样本

在科学实验中，我们常用的方法是从样本推论总体，因此先要了解总体和样本的概念。

具有共同性质的个体所组成的集团称为总体。如水稻南京11号的总体，是指南京11号这一品种在多年、多地中的无数次种植中的所有个体，这个总体中的个体数目相当地多，我们无法得到比较准确的数字，象这样的总体称为无限总体。再如一块玉米田中的所有果穗，或一块田里的小地蚕幼虫等这一类的总体，虽然它们的个体数较多，但还是能比较准确地知道其个体数目，这类总体称为有限总体。

同一总体中各个个体的某类性状、特性不可能完全一致。如水稻品种南京11号，即使是栽种在条件相对一致的一块田中，它的株高就彼此不一。每一个体的某一性状或特性的测定值叫观察值，由于偶然因素影响而表现出变异的观察值总称为变数或随机变数，某一特定的观察值称变量。由总体的全部观察值计得的总体特征数，如总体平均数等称为参数，一般用希腊字母表示。如用μ表示总体的平均数。

我们研究的对象是总体，但总体所包含的个体太多，我们无法逐一加以测定，所以，我们一般只是研究总体的一小部分，即从总体中抽取若干个个体来加以研究。这些个体的组成称为样本。由样本而得到的特征数如样本平均数等则称为统计数。统计数是总体相应的参数的估计值，一般用拉丁字母表示。如用$\bar{x}$表示样本的平均数。

既然要从样本估计总体，那就要考虑样本的代表性，样本越能近似地代表总体就越好，这样的样本只能随机地从总体中抽取，即总体内各个体具有同等的被抽取机会，这种样本称为随机样本。

第二节　次数分布

一、试验资料的性质

田间试验中所得到的资料，因研究的性状、特性不同而有不同的性质。一般可分为两大类：

（一）数量性状的资料　获得数量性状资料有计数和量测两种方法，所得变数各不相同。

1.连续性变数：指由称量、度量或测量等量测方法所得的数据；在两个相邻的数值间可以有微量差异的其它值存在。如测定水稻每穗粒重，在2克和3克之间，可以有2.5克或2.532克等等数值存在，其小数位数多少因称量的精确度而不同。这种变数称为连续性变数。象水稻的株高、产量等，均属于此类变数。

2.不连续性变数或间断性变数：指用计数方法获得的资料，其各个观察值必须用整数表示。如基本苗数，棉苗上的棉蚜数，在相邻的整数间不会有带有小数的数值。由于这些变量都是整数，而两个相邻的整数间是不连续的，故称这类变数为不连续性变数或间断性变数。

（二）质量性状的资料　指能观察分类而不能量测的性状，即属性

性状，如花的颜色，芒的有无等。整理这类资料，可采用下列方法：

1.应用统计次数的方法：在一定总体内，统计其具有该种性状的个体数目及不具有该种性状的个体数目，按两类分别计其次数。如在160株水稻植株中有120株有芒，40株无芒。这类资料也称次数资料。

2.数量化方法：给予每个属性以相当的数量，如小麦子粒颜色有白有红，可令呈白色的数量为0，呈红色的数量为1。从这类变数所得的资料，处理方法同间断性变数资料。

二、次数分布表

田间试验和调查研究所得的大量资料，如不加整理，对这些资料很难得出明确的概念。如能按数值大小进行分组，制成次数分布表，就可以看出资料的集中和变异情况，从而对资料有一个初步的概念，并为以后进一步处理资料打下基础。

（一）间断性变数资料的整理　现以棉苗上的单株棉蚜数为例，调查100株棉苗上的单株棉蚜数，得下列数据：

58	57	63	48	58	49	59	31	46	56
38	48	42	90	82	72	42	42	76	33
44	41	57	39	39	36	26	43	48	79
34	68	52	48	33	67	48	27	48	39
45	44	69	43	32	66	43	48	44	29
53	36	28	40	45	74	56	41	60	49
71	53	24	18	17	57	18	14	83	64
88	62	77	62	63	63	33	13	44	6
95	38	53	74	54	24	35	31	53	54
53	48	51	51	4	21	43	59	87	46

因计有100个变量，可分为10组，经整理后将上述100株棉苗上的单株棉蚜数列成表1.1，即为次数分布表。表上左边一列叫分组数列。从表上我们可见：（1）单株棉蚜数在0—99头，大部分分布在30—60头，最多见的在40—49头。（2）次数分布中间多，两头少。（3）各

组组距、组限均为整数，且相邻组限不重叠。

表1.1　　100株棉苗上棉蚜数的次数分布表

组　　距	次　　数(f)
0～9	2
10～19	5
20～29	7
30～39	14
40～49	28
50～59	20
60～69	11
70～79	7
80～89	4
90～99	2
Σ	100

（二）连续性变数资料的整理　连续性资料和间断性资料整理有所不同。为制作次数分布表，先解释如下术语：

高限：各组的最大值，以U表示。

低限：各组的最小值，以L表示。

组距：各组的高、低限之差，用i表示。

组值：又叫组中值，为各组的高、低限的平均数，以X表示。

$$X=\frac{1}{2}(L+U)$$

极差：亦称全距，为整个资料中的最大值和最小值之差，以R表示。

次数：又叫频数，为每一组所包括的测定值个数，以f表示，

$$\Sigma f = n。$$

次数分布表的制作步骤：

1.求极差。

2.确定组数和组距：根据资料的具体情况和参考表1.2决定组数，进而可求得组距 $i=\frac{R}{组数}$，i应取整数或便于计算之数。

表1.2　　样本大小与组数多少的关系

样本内观察值个数	分组时的组数
50～	5～10
100～	8～16
200～	10～20
300～	12～24
500～	15～30
1000～	20～40

3.确定第一组（分组数列中数值最小的那一组）的低限L_1。在确定L_1时应注意：（1）$L_1+\frac{1}{2}i$（即组中值）应是数字位数较少或便于计算的数。（2）第一组的组中值应接近或等于资料中最小的那个测定值。

4.写出分组数列：在写分组数列时要遵守三个原则。

（1）互斥：即各组的数量范围除相邻组有一共同组限外，要互相排斥，一个测定值只能在其中一组，不得重复。故而连续性变数分组的组限小数位数应比观察值多一位。（2）完全：即分组数列要把全部测定值都包括进去。（3）一致：即同一分组数列的组距要一致（开口次数分布表的第一和最后一组除外，见表1.7）

5.统计测定值在分组数列各组中出现的次数。

[例1.1]，表1.3为140行（行长4尺）水稻产量的资料，试制作次数分布表

表1.3　　140行（行长4尺）水稻产量　　单位：（克）

177	215	197	97	123	159	245	119	119	131
149	152	167	104	161	214	125	175	219	118
192	176	175	95	136	199	116	165	214	95
158	83	137	80	138	151	187	126	196	134
206	137	98	97	129	143	179	174	159	165
136	108	101	141	148	168	163	176	102	194
145	173	75	130	149	150	161	155	111	158
131	189	91	142	140	154	152	163	123	205
149	155	131	209	183	97	119	181	149	187
131	215	111	186	118	150	155	197	116	254
239	160	172	179	151	198	124	179	135	184
168	169	173	181	188	211	197	175	122	151
171	166	175	143	190	213	192	231	163	159
158	159	177	147	194	227	141	169	124	159

解：本资料 $n=140$，可以分12—15组，假若分12组，先求出极差，$R=245-75=179$，故 $i=\frac{179}{12}=14.9$，为方便计，取15克作为组距。因最小测定值为75，故第一组的组中值可以定为75，第二组组中值则为90，第三组组中值为105……。各组组中值确定后，减去$\frac{1}{2}$组距得其低限，加上$\frac{1}{2}$组距为其高限。如第一组的低限为 $75-\frac{1}{2}(15)=67.5$，高限为 $75+\frac{1}{2}(15)=82.5$，照此方法写出分组数列于表1.4，然后将各观察值划记号统计于表1.4上，表1.4即为140行水稻产量的次数分布表。实得组数为13组，比预定多一组，这是因为第一组和最后一组各向外延伸了半个组距而造成的。

表1.4　　140行水稻产量次数分布表

分组数列	组中点值（x）	划记号数	数次（f）
67.5～82.5	75	丅	2
82.5～97.5	90	正丅	7
97.5～112.5	105	正丅	7
112.5～127.5	120	正正下	13
127.5～142.5	135	正正正丅	17
142.5～157.5	150	正正正正	20
157.5～172.5	165	正正正正正	25
172.5～187.5	180	正正正正—	21
187.5～202.5	195	正正下	13
202.5～217.5	210	正丅	9
217.5～232.5	225	下	3
232.5～247.5	240	丅	2
247.5～262.5	255	—	1
总次数（n）			140

从表1.4我们可见（1）水稻单行（行长4尺）产量变异范围在67.5—262.5克之间。（2）大部分产量是在112.5—202.5克之间，最常见的是在157.5—172.5克之间。（3）产量的次数分布为中间高两头低的钟形单峰分布。

三、次数分布图

试验资料除用次数分布表来表示外，也可以用图形来表示。次数分布图可以更形象地表明次数分布的情况。常用的图示有方柱形图和多边形图。

不论用那一种形式的图示，都可以用坐标纸作好直角坐标，水平方向为横坐标，用来表示各组的组限或组中值；垂直方向为纵坐标，用来表示各组的次数（频数）。

（一）方柱形图　方柱形图适用于表示连续性变数的次数分布，作图时，首先写出分组数列的低限，并以各组的次数为相应的纵坐标，作

出一个个小矩形。例1.1的资料作方柱形图于图1.1。

（二）多边形图　多边形图也是表示连续性变数资料的一种常用图示法，且可在同一张图纸上表示两组以上的资料。在作图时，首先写出每组的组中值，然后在组中值的上方、该组次数的高度处标记一点，依此方法标出各组的次数点，各点标好后用直线依次连接，所成图形为次数多边形图。多边形图的的折线在左边最小组值和右边最大组值外，应向左右各伸出一个组距，和横轴相交。

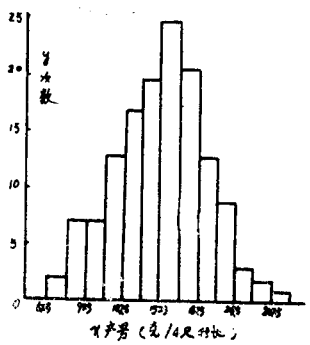

图1.1　140行水稻产量次数分布方柱形图

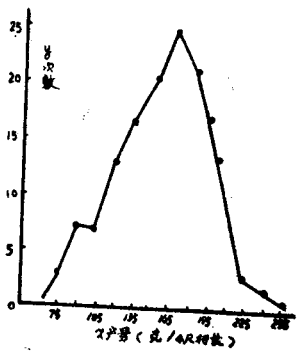

图1.2　140行水稻产量次数分布多边形图

在制作方柱形图和多边形图时应注意图形的大小比例适当，一般横坐标与纵坐标长度比为5：4或6：5。

四、相对次数和相对累加次数

将次数分布表上的各组次数除以总次数，即得各组的相对次数，各组的相对次数依次累加（或各组的次数依次累加，再除以总次数）。即得相对累加次数。

相对次数和相对累加次数作用有二：

1.明瞭测定值在各种不同范围内出现的频率。

2.不同的资料可以相互比较。

[例1.2]、例1.1的资料可以进一步做成相对次数和相对累加次数（表1.5），以及相对累加次数曲线（图1.3）。

表1.5　　例1.1资料的相次数和相对累加次数

分组数列	组中点值（X）	次数（f）	相对次数	相对累加次数
67.5～82.5	75	2	0.0143	0.0143
82.5～97.5	90	7	0.0500	0.0643
97.5～112.5	105	7	0.0500	0.1143
112.5～127.5	120	13	0.0928	0.2071
127.5～142.5	135	17	0.1215	0.3286
142.5～157.5	150	20	0.1428	0.4714
157.5～172.7	165	25	0.1786	0.6500
172.5～187.5	180	21	0.1500	0.8000
187.5～202.5	195	13	0.0929	0.8929
202.5～217.5	210	9	0.0642	0.9571
217.5～232.5	225	3	0.0215	0.9786
232.5～247.5	240	2	0.0143	0.9929
247.5～262.5	255	1	0.0071	1.0000
Σ		140	1.0000	

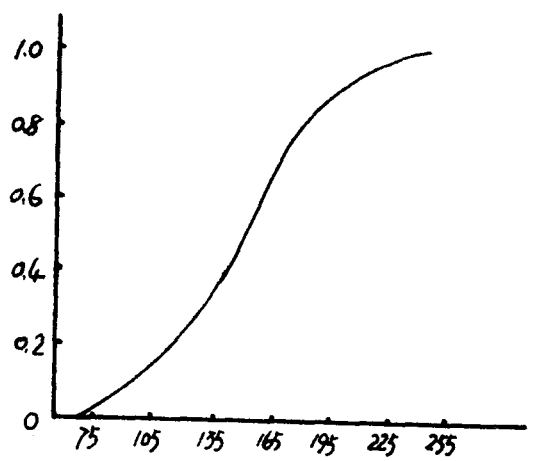

图1.3　例1.1资料的相对累加次数曲线

第三节　平均数和变异数

一、平均数

测定值的代表值有平均数、众数和中位数三种，其中以平均数的应用最为广泛。平均数又分算术平均数、几何平均数和调和平均数等。最常用的是算术平均数。

（一）算术平均数：若有n个测定值为X_1，X_2，X_3，……X_n，则算术平均数为

$$\overline{X}=\frac{x_1+x_2+\cdots\cdots+x_n}{n}=\frac{\Sigma X}{n} \qquad (1.1)$$

n又称为样本容量，也就是测定值的个数。Σ是总和符号，算术平均数的重要特性之一为$\Sigma(X-\overline{X})=0$，证明如下：

$$\Sigma(X-\overline{X})=(X_1-\overline{X})+(X_2-\overline{X})+\cdots\cdots+(X_n-\overline{X})$$

$$=(X_1+X_2+X_3+\cdots\cdots+X_n)-n\overline{X}$$

$$=\Sigma X-n\overline{X}=\Sigma X-n\frac{\Sigma X}{n}=0$$

各观察值和平均数之差称为离均差，上式清楚地表明离均差的总和等于零。

平均数第二个重要特性是离均差平方和$\Sigma(X-\overline{X})^2$为最小。即

$$\Sigma(X-\overline{X})^2<\Sigma(X-a)^2 \qquad a\neq\overline{X}$$

[例1.3]测得8个土壤样品的有机质含量为：0.18、0.19、0.19、0.20、0.20、0.20、0.22、0.23（克）试求$\overline{X}$。

解 $\overline{X}=\frac{0.18+0.19+\cdots\cdots+0.23}{8}=\frac{1.61}{8}=0.20$ （克）

X的有效数字一般只需和测定值相同，至多多一位。若各测定值在总体（或样本中）有不同的比重（统计上称为权），在计算时要把各个X的不同权考虑在内。一般权数用f表示。求得的平均数称加权平均

数。

$$\overline{X}=\frac{f_1x_1+f_2x_2+\cdots\cdots+f_mx_m}{f_1+f_2+\cdots\cdots+f_m}=\frac{\Sigma fx}{\Sigma f}=\frac{\Sigma fx}{n} \qquad (1.2)$$

[例1.4]调查某大队1000亩小麦田的粘虫密度，有100亩每亩500头，700亩每亩900头，200亩每亩1300头，试求该大队每亩麦田的平均粘虫数。

解：$\overline{X}=\frac{(100\times500)+(700\times900)+(200\times1300)}{100+700+200}=\frac{940000}{1000}=940$（头）

（二）几何平均数：用G表示，常用于表示平均增长率

$$G=\sqrt[n]{x_1\cdot x_2\cdots\cdots x_n}=(x_1\cdot x_2\cdots\cdots x_n)^{1/n} \qquad (1.3a)$$

将上式两边取对数得

$$\lg G=\frac{\lg x_1+\lg x_2+\cdots\cdots+\lg x_n}{n}=\frac{\Sigma\lg x}{n} \qquad (1.3b)$$

显然几何平均数是各测定值的对数平均数的反对数。

[例1.5]在诱蛾灯下诱得三化螟各世代的最高蛾数如下表，试求各世代发蛾量的平均增长率。

表1.6　　某诱蛾灯下三化螟各世代的最高蛾数

世　代	蛾　数	相对增长率（X）	1gX
1	50		
2	375	7.5000	0.875061
3	2005	5.3467	0.728086
4	18150	9.0524	0.956764
Σ			2.559911

解：$\lg G=\frac{2.559911}{3}=0.853303$　　$G=\lg^{-1}0.853303=7.1335$

也就是说三化螟蛾数平均下一代较上一代增长7.1335倍。

（三）调和平均数：又称倒数平均数，记作H，用于表示平均速度。其值为各测定值的倒数平均数的倒数。

$$\frac{1}{H}=\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots\cdots+\frac{1}{x_n}\right)=\frac{\Sigma(\frac{1}{x})}{n} \qquad (1.4a)$$

$$H=\frac{n}{\Sigma(\frac{1}{x})} \qquad (1.4b)$$

[例1.6]测得一定条件下10个第一代稻苞虫卵每天的发育进度：4个为0.125，2个为0.143，4个为0.100，试求平均发育进度和平均卵期。

解：$\frac{1}{H}=\frac{1}{10}\left(\frac{4\times1}{0.125}+\frac{2\times1}{0.143}+\frac{4\times1}{0.100}\right)=8.6$（天）

$$H=\frac{1}{8.6}=0.11628$$

这就是说，第一代稻苞虫的平均卵期为8.6天，平均每天发育进度为0.116。

（四）中位数：记作Md，如将测定值按大小顺序排列，当测定值为奇数时，中位数就是最中间的一个测定值；当测定值数目为偶数时，中位数就是最中间二个测定值的算术平均数。中位数是一个地位表征常数，其意义是测定值在其左右各分布50%，植保上常用的LD_{50}（致死中量）、LC_{50}（致死中浓度）、LT_{50}（致死中时间）等亦可认为是一种中位数。如用次数分布表求中位数，其公式如下：

$$Md=L_{md}+\frac{\frac{n}{2}-A}{f_{md}}\times i \qquad (1.5)$$

其中Lmd为中位数所在组的低限，fmd为中位数所在组的次数。n为总次数，A为中位数所在组上面各组的总次数，i为组距。

[例1.7]取三化螟初孵幼虫204头，使其在浸有1:100的敌百虫溶液的滤纸上爬行（在25℃下），得不同时间下死亡头数于表1.7的直行(1)(2)，试求中位数。

表 1.7　　敌百虫的杀螟效果的致死中时间计算

（1）爬行时间（分钟）	（2）致死头数	（3）M_d的位置	（4）累加次数	（5）相对累加次数（%）
<15	22		22	10.76
15—	31		53	25.98
25—	29		82(A)	40.20
35—	36	M_d	118	57.84
45—	25		143	70.10
55—	32		175	85.78
65—	21		196	96.08
≥75	8		204	100.00
Σ	204			

解：由表1.7直行(3)知中位数应在35—44.9分钟这一组，而在35分钟以内已有82头死亡，故在35—44.9分钟这一组内，中位数之前的死亡头数还有$\frac{204}{2}-82=20$头，又知这一组次数为36，i＝10分钟，中位数在这一组所得的值应为$\frac{20}{36}\times10=5.6$，也就是死亡20头约需5.6分钟，再加上中位数所在组的低限35，故有

$$Md=35+5.6=40.6$$

如用公式表示则

$$Md=35+\frac{\frac{204}{2}-82}{36}\times10=40.6$$

由此说明三化螟幼虫在40.6分钟前死掉一半，在其后死掉一半，故此例的致死中时间为40.6分钟。

（五）众　数：记作 Mo。它表示最常出现的测定值，对间断性变数，可直接求得。若为分组资料，即为次数最多的一组的组中值。

[例1.8]测定100头第三代粘虫的一代生活史所需的日数结果于表1.8，试求这代粘虫生活史的众数日期。