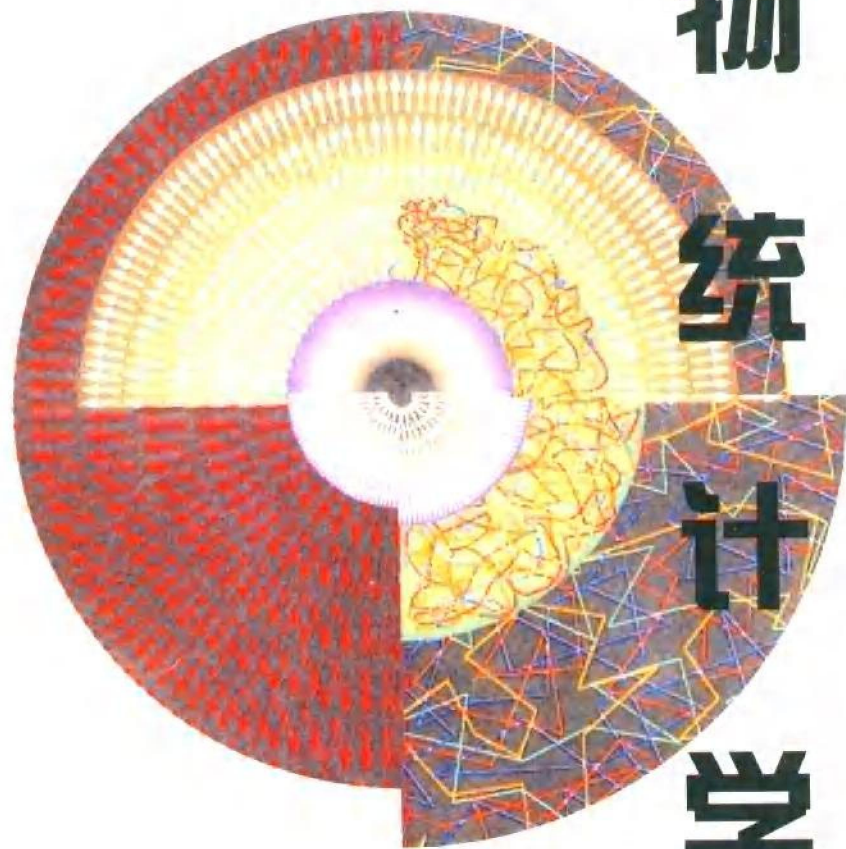


# 生物统计学



李春喜 王文林 等 编著

科学出版社

Q-332  
LCX

128/07  
生物统计学

李春喜 王文林 等编著

科学出版社

## 内 容 简 介

本书较为系统地介绍了生物统计学的基本原理和方法,在简要叙述了生物统计学的产生、发展及其研究对象与作用、生命科学中试验资料的整理、特征数的计算、概率和概率分布、抽样分布基础上,着重介绍了平均数的统计推断、 $\chi^2$  检验、方差分析、直线回归与相关分析、可直线化的曲线回归分析、多项式回归分析、多元回归与相关分析和协方差分析,同时对常用试验设计、抽样原理和方法也进行了详细叙述。在上述内容的基础上,对聚类分析、典型相关、因子分析等多元分析也作了简要介绍。本书通俗易懂,具有一定的深度和广度,适合生命科学、农业科学、医学工作者阅读,也可供本、专科院校生物类专业作为教材使用。

### 图书在版编目(CIP)数据

生物统计学/李春喜等编著. —北京:科学出版社,1997. 8

ISBN 7-03-006069-5

I. 生… II. 李… III. 生物统计 IV. Q-724

中国版本图书馆 CIP 数据核字(97)第 07412 号

科学出版社出版

北京东黄城根北街16号

邮政编码:100717

北京双青印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

\*

1997年8月第一版 开本:787×1092 1/16

1997年8月第一次印刷 印张:17 1/4

印数:1-3 550 字数:391 000

定价:25.00元

## 《生物统计学》编委会名单及作者分工

主 编：李春喜 王文林  
副主编：陈士林 李建军 郑爱珍  
编 委：詹克慧 成庆利 陈晓萍 孙希增  
审 稿：崔党群

第 一 章：郑爱珍  
第 二 章：陈晓萍  
第 三 章：郑爱珍  
第 四 章：李建军  
第 五 章：王文林  
第 六 章：王文林  
第 七 章：詹克慧  
第 八 章：陈士林  
第 九 章：李春喜  
第 十 章：孙希增  
第 十 一 章：成庆利  
第 十 二 章：李春喜  
附 表：成庆利(附表 1—3)  
          陈晓萍(附表 4—10)  
          孙希增(附表 11—12)

# 前 言

生物统计学是运用数理统计的原理和方法来分析和解释生物界各种现象和试验调查资料的一门科学。随着生物学的不断发展,对生物体的研究和观察已不再局限于定性的描述,而是需要从大量调查和测定数据中,应用统计学方法,分析和解释其数量上的变化,以正确制定试验计划,科学进行试验结果的分析,从而作出符合科学实际的推断。目前,生物统计学在农学、林学、畜牧、医药、卫生、生态、环保等领域已有广泛应用,但在纯生物学研究方面的应用,不管是在深度上还是广度上都不及上述领域。有鉴于此,在生物学研究中,迫切需要加强生物统计学的应用,对大学生物类专业,它也是一门应被十分重视的工具课程。本书正是为了满足这些需要而编写的。

本书的写作是在作者多年从事生物统计学教学和应用研究的基础上完成的。书中所介绍的内容主要侧重于各种统计方法的应用,在统计原理方面,一般只作概念上的介绍和公式的简单推导,对有些较复杂的统计公式则只给出公式,其目的主要是为了让读者不但对统计学原理有较全面的了解,更重要的是结合实例了解和掌握各种常用统计方法。在本书的安排上,全书共分十二章,概括起来主要有五个方面:第一章至第三章介绍统计和概率的基础知识,包括生物统计学的概念和内容、数据的搜集与整理、平均数和变异数的计算、概率和概率分布等;第四章、第五章介绍统计推断,包括样本平均数的检验、样本频数的检验、方差同质性检验、非参数检验和 $\chi^2$ 检验;第六章至第九章介绍统计分析方法,主要内容有方差分析、直线回归与相关分析、可直线化的曲线回归分析、多元回归与相关分析、逐步回归分析、多项式回归、协方差分析;第十章、第十一章介绍抽样与试验设计,主要包括抽样误差估计、抽样方法、抽样方案制订及常见的试验设计如对比设计、随机区组设计、正交设计及其相应的统计分析方法;第十二章对近年来应用越来越多的多元统计分析进行了简单介绍。每章都附有一定数量的思考练习题,供读者参考。

本书中的例子主要有两个来源,一个是近年来有关生物学、农学、林学、医学、畜牧、水产、环保等领域或学科的实际研究资料,另一个是有关著作中的一些例题。崔党群教授在百忙中通审了全书,并提出了富有建设性的建议。贾玉书同志承担了本书的大部分绘图工作。姜丽娜同志在本书的录排中做了大量工作。在本书的出版过程中,得到了科学出版社的大力支持,特别是张晓春同志在书稿的编审和发行方面作了大量工作,在此一并表示谢意。

本书通俗易懂,具有一定的深度和广度,适合生物学、农学、医学、畜牧、水产、环保等领域或学科的科学工作者阅读,也可供本、专科院校生物类专业作为教材使用。

由于作者水平的限制和资料占有的局限性,本书难免会有错误和不妥之处,敬请读者批评指正,以便日后修订完善。

李春喜 王文林

1997年3月

· v ·

# 目 录

前言	v
<b>第一章 概论</b>	<b>1</b>
第一节 生物统计学的概念	1
第二节 生物统计学的主要内容	1
第三节 生物统计学发展概况	2
第四节 常用统计学术语	3
思考练习题	5
<b>第二章 试验资料的整理和特征数的计算</b>	<b>6</b>
第一节 试验资料的搜集与整理	6
第二节 试验资料特征数的计算	12
思考练习题	18
<b>第三章 概率与概率分布</b>	<b>19</b>
第一节 概率基础知识	19
第二节 二项分布	24
第三节 泊松分布	27
第四节 正态分布	29
第五节 统计数的分布	33
思考练习题	40
<b>第四章 统计推断</b>	<b>41</b>
第一节 假设检验的原理与方法	41
第二节 样本平均数的假设检验	45
第三节 样本频率的假设检验	53
第四节 参数的区间估计与点估计	56
第五节 方差的同质性检验	61
第六节 非参数检验	63
思考练习题	68
<b>第五章 <math>\chi^2</math> 检验</b>	<b>70</b>
第一节 $\chi^2$ 假设检验的原理与方法	70
第二节 适合性检验	71
第三节 独立性检验	74
思考练习题	78
<b>第六章 方差分析</b>	<b>79</b>
第一节 方差分析的基本原理	79
第二节 单因素方差分析	88
第三节 二因素方差分析	91
第四节 多因素方差分析	100
第五节 方差分析缺失数据的估计	104

第六节	方差分析的基本假定和数据转换 .....	106
思考练习题	.....	109
<b>第七章</b>	<b>直线回归与相关分析</b> .....	<b>112</b>
第一节	回归和相关的概念 .....	112
第二节	直线回归 .....	113
第三节	直线相关 .....	121
第四节	可直线化的非线性回归分析 .....	124
思考练习题	.....	134
<b>第八章</b>	<b>多元回归与多项式回归分析</b> .....	<b>135</b>
第一节	多元回归分析 .....	135
第二节	逐步回归 .....	144
第三节	多元相关分析 .....	153
第四节	多项式回归分析 .....	157
思考练习题	.....	166
<b>第九章</b>	<b>协方差分析</b> .....	<b>167</b>
第一节	协方差分析的意义和作用 .....	167
第二节	单向分组资料的协方差分析 .....	168
第三节	两向分组资料的协方差分析 .....	172
第四节	协方差分析的数学模型和基本假定 .....	175
思考练习题	.....	176
<b>第十章</b>	<b>抽样原理与方法</b> .....	<b>177</b>
第一节	抽样误差的估计 .....	177
第二节	样本容量的确定 .....	179
第三节	抽样的基本方法 .....	181
第四节	抽样方案的制订 .....	184
思考练习题	.....	185
<b>第十一章</b>	<b>常用试验设计及其统计分析</b> .....	<b>186</b>
第一节	试验设计的基本原理 .....	186
第二节	对比试验设计及其统计分析 .....	189
第三节	随机区组试验设计及其统计分析 .....	191
第四节	正交试验设计及其统计分析 .....	198
思考练习题	.....	204
<b>第十二章</b>	<b>多元统计分析简介</b> .....	<b>206</b>
第一节	数据矩阵与相似系数 .....	206
第二节	聚类分析 .....	211
第三节	判别分析 .....	214
第四节	主成分分析 .....	217
第五节	因子分析 .....	222
第六节	典型相关分析 .....	227
第七节	时间序列分析 .....	231
思考练习题	.....	235
<b>附表</b>	.....	<b>238</b>

附表 1	正态分布的累积函数 $F(u)$ 值表 .....	238
附表 2	正态离差 ( $u$ ) 值表(双尾) .....	240
附表 3	$t$ 值表(双尾) .....	240
附表 4	$\chi^2$ 值表(右尾) .....	241
附表 5	$F$ 值表(右尾) .....	242
附表 6	符号检验表 .....	246
附表 7	秩和检验表 .....	246
附表 8	新复极差检验 $SSR$ 表 .....	247
附表 9	$q$ 值表(双尾) .....	248
附表 10	$r$ 与 $R$ 的临界值表 .....	249
附表 11	常用正交表 .....	250
附表 12	正交多项式系数表 .....	259
<b>参考文献</b>	.....	<b>263</b>



# 第一章 概 论

## 第一节 生物统计学的概念

生物统计学是数理统计在生物学研究中的应用,它是用数理统计的原理和方法来分析 and 解释生物界各种现象和试验调查资料的科学。随着生物学研究的不断发展,应用统计学方法,来认识、推断和解释生命过程中的各种现象,也越来越广泛。尽管生物统计学在应用过程中曾经受到过一些批评,但绝大多数生物学家、农学家、园艺学家、育种学家、畜牧学家、医学工作者以及人口学家还是越来越普遍地在自己的研究领域里应用生物统计学方法。

生物学研究的对象是复杂的有机体,与非生物相比,它具有更加特殊的复杂性。有机体本身的生理活动和生化变化及有机体受外界环境因素的影响等,都使生物学研究的试验结果产生许多较大的差异性,这种差异性往往会掩盖生物体本身的特殊规律。在生物学研究中,大量试验资料内在的规律性,也容易被杂乱无章的数据所掩盖,容易被人们所忽视。因而,应用生物统计对生物学研究就显得特别重要。生物学研究的实践证明,只有正确地应用生物统计原理和分析方法对生物学试验进行合理设计,对数据进行客观分析,才能得出科学的结论。

在对事物的研究过程中,人们往往是通过某事物的一部分(样本),来估计事物全部(总体)的特征的,目的是为了以样本推断总体,从特殊推导一般,对所研究的总体作出合乎逻辑的推论,得到对客观事物本质的和规律性的认识。在生物学研究中,我们所期望的是总体,而不是样本。但是在具体的试验过程中,我们所得到的却是样本而不是总体。因此,从某种意义上讲,生物统计学是研究生命过程中以样本来推断总体的一门学科。

生物统计学是在生物学研究过程中,逐渐与数学的发展相结合所形成的,它是应用数学的一个分支,属于生物数学的范畴。生物统计学以数学的概率论为基础,也涉及到数列、排列、组合、矩阵等知识。生物统计学作为一门重要的工具课,一般不过多讨论数学原理,而主要偏重于统计原理的介绍和具体分析方法的应用。

## 第二节 生物统计学的主要内容

生物统计学的基本内容,概括起来主要包括试验设计和统计分析两大部分。在试验设计中,主要介绍试验设计的有关概念、试验设计的基本原则、试验设计方案的制定、常用试验设计方法,其中主要的有对比试验设计、随机区组试验设计以及正交试验设计等。在统计分析中,主要包括数据资料的搜集和整理、数据特征的度量、统计推断、方差分析、回归

和相关分析、协方差分析、主成分分析、聚类分析等。

从生物统计学的基本作用上来讲,其任务可以概括为以下几个方面:

(1) 提供整理和描述数据资料的科学方法,确定某些性状和特性的数量特征。一批试验或数据资料,若不整理则杂乱无章,不能说明任何问题。统计方法提供了整理资料、化繁为简的科学程序,它可以从众多的数据资料中,归纳出几个特征数或绘出一定形式的图表,使试验研究者能从少数的特征数或一些简单的图表中了解大批资料所蕴藏的信息。

(2) 判断试验结果的可靠性。一般在试验中要求除试验因素以外,其它条件都应控制一致,但在实践中无论试验条件控制得如何严格,其试验结果总是受试验因素和其它偶然因素的影响。偶然因素的影响就是造成试验误差的重要原因。一个试验结果,是由试验因素造成的还是试验误差造成的,要正确判断就必须应用统计分析方法。

(3) 提供由样本推断总体的方法。试验的目的在于认识总体规律,但由于总体庞大,一般无法实施,在研究过程中都是抽取总体中的部分作为样本,用统计方法以样本来推断总体的规律性,在这种推断中,统计原理和方法起到了理论上的保证作用。

(4) 提供试验设计的一些重要原则。为了以较少的人力、物力和财力取得较多的试验信息和较好的试验结果,在一些生物学研究中,就需要科学地进行试验设计,如对样本容量的确定、抽样方法、处理设置、重复次数的确定以及试验的安排等,都必须以统计学原理为依据。从统计分析和试验设计的关系来看,统计学原理可以为试验设计提供合理的依据,而试验设计又是统计分析方法的进一步运用。以统计学原理为指导,进行科学合理的试验设计时,可以使在较少人力、物力、时间和条件下,得出可靠而准确的数据和信息。以往有一些试验资料,由于设计不当而丧失了大量的试验信息,究其原因多半是由于缺乏一定的统计知识,使试验的效率大大降低。当然,统计原理和分析方法对试验设计有着积极的指导意义,但它绝对不可能代替试验设计。如果试验目的、要求不明确,设计不合理,试验条件不合适,统计数据不准确,这种试验也绝对不会成功,统计原理和分析方法都不可能挽救试验的这种失败。

### 第三节 生物统计学发展概况

现代统计学起源于17世纪,它主要有两个来源,一是政治的需要,二是当时贵族阶层对概率数学理论很感兴趣而发展起来的。另外,研究天文学的需要也促进了统计学的发展。瑞士数学家 J. Bernouli(1654—1705)系统论证了大数定律。后来 Bernouli 的后代 D. Bernouli(1700—1782)将概率论的理论应用到医学和人类保险。

正态分布理论对研究生物统计的理论是十分重要的,它最早是由 De Moiver 于 1733 年发现的,后来德国天文学和数学家 Gauss(1777—1855)在研究观察误差理论时,也独立发现了正态分布的理论方程,因此,常有人将正态分布称为 Gauss 分布。

统计学用于生物学的研究,开始于 19 世纪末。1870 年,英国遗传学家 Galton(1822—1911)在 19 世纪末叶应用统计方法研究人种特性,分析父母与子女的变异,探索其遗传规律,提出了相关与回归的概念,开辟了生物学研究的新领域。尽管他的研究当时并未成功,但由于他开创性应用统计方法来进行生物学研究,后人推崇他为生物统计学的创始人。

在此之后, Galton 和他的继承人 K. Plarson(1857—1936)经过共同努力于 1895 年成

立了伦敦大学生物统计实验室,于1889年发表了《自然的遗传》一书。在该书中,K. Plarson 首先提出了回归分析问题,并给出了计算简单相关系数和复相关系数的计算公式。K. Plarson 在研究样本误差效应时,提出了测量实际值与理论值之间偏离度的指数卡方( $\chi^2$ )的检验问题,它在属性统计分析中有着广泛的应用。例如,在遗传上孟德尔豌豆杂交试验,高豌豆品种与低豌豆品种杂交后,它的后代理论比率应该是高3:低1,但实际后代数是否符合3:1,需用 $\chi^2$ 进行检验。

K. Plarson 的学生 Gosset(1876—1937)对样本标准差进行了大量研究,于1908年以笔名“Student”在该年的《生物统计学报》(Biometrika)上发表论文,创立了小样本检验代替大样本检验的理论和方法,即  $t$  分布和  $t$  检验法。 $t$  检验已成为当代生物统计工作的基本工具之一,它也为多元分析的理论形成和应用奠定了基础。

英国统计学家 Fisher 于1923年发展了显著性检验及估计理论,提出了  $F$  分布和  $F$  检验。他在从事农业试验及数据分析研究时,创立了正交试验设计和方差分析。在生物统计中,方差分析有着广泛的应用,特别是在他发表了《试验研究工作中的统计方法》专著后,对推动和促进农业科学、生物学和遗传学的研究与发展,起到了奠基作用。自20年代 Fisher 的方差分析问世以来,各种数理统计方法不但在实验室中成为研究人员的析因工具,而且在田间试验、饲养试验、临床试验等农学、医学和生物学领域也得到了广泛应用。

Neyman(1894—1981)和 S. Pearson 进行了统计理论的研究工作,分别于1936年和1938年提出了一种统计假说检验学说,假说检验和区间估计作为数学上的最优化问题,对促进统计理论研究和对试验作出正确结论具有非常实用的价值。

另外,P. C. Mabeilinrobis 对作物抽样调查、A. Waecl 对序贯抽样、Finney 对毒理统计、K. Mather 对生统遗传学、F. Yates 对田间试验设计等都做出了杰出的贡献。

国内对生物统计学的应用始于30年代。中华人民共和国成立以后,许多生物学研究工作者积极从事统计学理论和实践的应用研究,使生物统计学在农业科学、医学科学、生物学、遗传学、生态学等学科领域发挥了重要作用。应用试验设计方法和统计分析理论,进行农作物品种产量比较试验、病虫害的预测预报、动物饲养试验、饲料配方、毒理试验、动植物资源的调查与分析、动植物育种中遗传资源和亲子代遗传的分析等都取得了较好成果。

近年来,生物统计学发展迅速,从中又分支出生统遗传学(群体遗传学)、生态统计学、生物分类统计学、毒理统计学等。由于数学与生物学和农学的应用,使生物数学成为一门新的学科,生物统计学只是它的一个分支学科。1974年,联合国教科文组织在编制学科分类目录时,第一次把生物数学作为一门独立的学科列入生命科学类中。随着计算机的普及和生物学研究的不断深入,生物统计的研究和应用必将越来越广泛,越来越深入。

## 第四节 常用统计学术语

### 一、总体与样本

总体是指研究对象的全体,而组成总体的基本单元称为个体。总体按总体单位的数目可分为有限总体和无限总体。个体极多或无限多的总体称为无限总体。例如,某一地区棉



#### 四、效应与互作

引起试验差异的作用称为效应,如不同饲料使动物的体重增加表现出差异,不同品种的玉米产量不同等。互作,也称连应,是指两个或两个以上处理因素间的相互作用产生的效应。如氮、磷肥共施会对作物产量产生互作效应。互作有正效应,也有负效应,如果氮、磷共施的产量效应大于氮、磷单施效应之和,说明氮磷互作为正效应,如果氮、磷共施的产量效应小于氮、磷单施效应之和,说明氮磷互作为负效应。

#### 五、机误与错误

机误,也叫试验误差,是指试验中由于无法控制的随机因素所引起的差异。如在抽样中,会出现较大或较小的数据,这是由于总体中的个体间存在一定的差异,它是不可避免的,试验中只能设法减小,而不能完全消灭。增加抽样或试验次数,可以降低机误的数值。

错误是指在试验过程中,人为的作用所引起的差错。如试验人员粗心大意,使仪器校正不准、药品配制比例不当、称量不准确、将数据抄错、计算出现错误等都是由于人为因素造成的,在试验中是完全可以避免的。

#### 六、准确性与精确性

统计工作是用样本的统计数来推断总体参数的。我们用统计数接近参数真值的程度,来衡量统计数准确性的高低,用样本中的各个变量间变异程度的大小,来衡量该样本精确性的高低。因此,准确性不等于精确性。准确性是说明测定值对真值符合程度的大小,而精确性则是多次测定值的变异程度。

#### 思考练习题

1. 生物统计学的主要内容和作用是什么?
2. 举例说明什么叫总体。什么叫样本。
3. 对下列问题进行判断,哪些属于有限总体?哪些属于无限总体?
  - (1) 一批鱼苗的体长;
  - (2) 某地区的地下水位;
  - (3) 某年某地区棉田棉铃虫的数量;
  - (4) 某林区的林木蓄积量;
  - (5) 全国所有自然保护区物种的数量。
4. 什么是参数?什么是统计数?二者有何联系和区别?
5. 什么是机误?什么是错误?如何避免试验错误的发生?

## 第二章 试验资料的整理和特征数的计算

在生物学试验及调查中,能够获得大量的原始数据,这是在一定条件下,对某种具体事物或现象观察的结果,我们称之为资料。这些资料在未整理之前,一般是分散的、零星的和孤立的,是一堆杂乱无章的数字。统计分析就是要依靠这些资料,通过整理分析进行归类,使其系统化,列成统计表,绘出统计图,计算出平均数、变异数等特征数。

### 第一节 试验资料的搜集与整理

#### 一、试验资料的类型

对试验资料进行分类是统计归纳的基础,若不进行分类,大量的原始资料就不能系统化、规范化。对试验资料进行分类整理时,必须坚持“同质”的原则。只有“同质”的试验数据,才能根据科学原理来分类,使试验资料正确反映事物的本质和规律。

对于生物学试验及调查所得的资料,由于使用方法和研究的性状特性不同,其资料性质也不相同。根据生物的性状特性,大致可分为数量性状和质量性状两大类,因而,我们所得到的资料有时是定量的,有时则是定性的,所以这些资料可以分为数量性状资料和质量性状资料。

##### (一) 数量性状资料

数量性状资料一般是由计数和测量或度量得到的。由计数法得到的数据称为计数资料,也称作非连续变量资料,如鱼的尾数、玉米果穗上籽粒行数、种群内的个体数、人的白细胞计数等。计数资料的变量值以正整数出现,不可能带有小数。如鱼的尾数只可能是1, 2, 3, ...,  $n$ , 绝对不会出现2.5、4.8等这样的数据。

由测量或度量所得的数据称为计量资料,也称为连续变量资料,数据通常用长度、重量、体积等单位表示,如人的身高、玉米的果穗重量、仔猪的体重、奶牛的产奶量等。计量资料不一定是整数,在相邻值之间有微小差异的数值存在。如小麦的株高为80—95cm,可以是85cm,也可以是86cm,甚至可以是86.5cm或86.54cm等变量值,随小数位数的增加,可以出现无限个变量值。至于小数位数的多少,要依试验的要求和测量仪器或工具的精度而定。

##### (二) 质量性状资料

质量性状资料,也称属性性状资料,是指对某种现象只能观察而不能测量的资料。如水稻花药、籽粒、颖壳的颜色,芒的有无,茸毛的有无;果蝇长翅与残翅;人血型有A、B、AB、O型;动物的雌、雄;疾病治疗的疗效有痊愈、好转、无效等。为了统计分析,一般需先把质量性状资料数量化,可以采取下面两种方法:

1. 统计次数法 于一定总体内,根据某一质量性状的类别统计其次数,以次数来作为质量性状的数据。在分组统计时可按质量性状的类别进行分组,然后统计各组出现的次数。因此,这类资料也称次数资料。例如,红花豌豆与白花豌豆杂交,统计 $F_2$ 代不同花色的植株时,在1000株植株中,有红花266株、紫花494株、白花240株,可以计算出三种颜色花出现的次数百分率分别为:26.6%、49.4%和24.0%。

2. 评分法 这种方法是用数字级别表示某现象在程度上的差别。如小麦感染锈病的严重程度可划分为0(免疫)、1(高度抵抗)、2(中度抵抗)、3(感染)四级;家畜精液品质可以评为三级,好的评为10分,较好的评为8分,差的评为5分。这样,就可以将质量性状资料量化了。经过数量化的质量性状资料的处理方法可以参照计数资料的处理方法。

## 二、试验资料的搜集

从统计学意义上讲,生物学所研究的一切问题,归根结蒂是用样本来估计总体的问题。因此,搜集样本资料是统计分析的第一步,也是全部统计工作的基础。资料的来源一般有两个,一是调查,二是试验。无论调查还是试验,统计学对原始资料的要求都是要完整、准确。

### (一) 调查

资料的调查方法有两种,一种是普查,另一种是抽样调查。普查是指对研究对象的每个个体都进行测量或度量的一种全面调查,比如人口普查、土壤普查等。普查一般要求在一定的时间或范围进行,主要目的是摸清研究对象的家底情况。在生物学研究中,普查仅仅是在极少数情况下才能进行,多数情况还是抽样调查,比如某一地区的生物资源调查、棉田某一病害发病率调查等,都需要抽样调查。

抽样调查是一种非全面调查,它是根据一定的原则对研究对象抽取一部分个体进行测量或度量,把得到的数据资料作为样本进行统计处理,然后利用样本特征数对总体进行推断。要使样本无偏差地估计总体,除了样本容量要大之外,重要的是采用科学的抽样方法,抽取有代表性的样本,取得完整而准确的数据资料。实践证明,正确的抽样方法不仅能节约人力、物力和财力,而且与相应的统计分析方法相结合,可以做出比较准确的估计和推断。

生物学研究中,由于研究的目的和性质不同,所采取的抽样方法也各不相同。以概率论和数理统计的原理为依据,用来推断总体的样本必须是随机样本,就是用随机抽样方法所得到的样本,只有这种样本才能正确估计出抽样误差,才能用来准确地推断总体。随机抽样必须满足两个条件:①总体中每个个体被抽中的机会是均等的。②总体中任意一个个体是否被抽中是相互独立的,即个体是否被抽中不受其它个体的影响。②适合于无限总体,但对生物学研究来说,部分研究的抽样对象属于有限总体,要完全符合随机样本的理论要求就非常困难。关于抽样的原理与方法,将在第十章叙述。

### (二) 试验

在生物学研究中,对于一些理论性的无限总体,一般需要通过设置各种类型的试验来获取样本资料,设置这些试验时,要遵循随机、重复和局部控制三项基本原则。常见的试验设计方法主要有:单因子随机区组试验、复因子随机区组试验、裂区设计试验、正交设计试验、拉丁方设计试验等。

### 三、试验资料的整理

#### (一) 原始资料的检查与核对

通过调查或试验取得原始数据资料后,要对全部数据进行检查与核对后,才能进行数据的整理。对原始资料进行检查与核对应从数据本身是否有错误、取样是否有差错和不合理数据的订正三方面进行。数据的检查和核对,在统计处理工作中是一项非常重要的工作。只有经过检查和核对的数据资料,保证数据资料的完整、真实和可靠,才能通过统计分析,来真实地反映出调查或试验的客观情况。

#### (二) 次数分布表

调查或试验所得的数据资料,经过检查与核对后,根据样本资料的多少确定是否分组。一般样本容量在 30 以下的小样本不必分组,可直接进行统计分析。如果样本容量在 30 以上时,就须将数据分成若干组,以便进行统计分析。数据经过分组归类后,可以制成有规则的次数分布表,作出次数分布图。

1. 计数资料的整理 计数资料基本上采用单项式分组法进行整理,它的特点是用样本变量自然值进行分组,每组均用一个或几个变量值来表示。分组时,可将数据资料中每个变量分别归入相应的组内,然后制成次数分布表。例如,从某鸡场调查 100 只来亨鸡每个月的产蛋数,原始数据结果见表 2.1。

表 2.1 100 只来亨鸡每月的产蛋数

15	17	12	14	13	14	12	11	14	13
16	14	14	13	17	15	14	14	16	14
14	15	15	14	14	14	11	13	12	14
13	14	13	15	14	13	15	14	13	14
15	16	16	14	13	14	15	13	15	13
15	15	15	14	14	16	14	15	17	13
16	14	16	15	13	14	14	14	14	16
12	13	12	14	12	15	16	15	16	14
13	14	16	15	15	15	13	13	14	14
13	15	17	14	13	14	12	17	14	15

每月产蛋数变动在 11—17 范围内,把 100 个观测值按照每月产蛋数加以归类,共分 6 组,这样经整理后可得出每月产蛋数的次数分布表,见表 2.2。

表 2.2 100 只来亨鸡每月产蛋数的次数分布表

每月产蛋数	次数( $f$ )	频率
11	2	0.02
12	7	0.07
13	19	0.19
14	35	0.35
15	21	0.21
16	11	0.11
17	5	0.05

从表 2.2 可以知道,一堆杂乱无章的原始数据资料,经初步整理后,就可了解这些资



料的大概情况,其中以每月产蛋数为14的最多。这样,经过整理的资料也就便于进一步的分析。对于变量较多而变异范围较大的计数资料,若以每一变量值划分一组,则显得组数太多而每组变量数目较少,看不出数据分布的规律性。如研究不同小麦农家品种的每穗粒数为18—62粒,如果按一个变量分为一组,需要分45组,显得十分分散。为了使次数分布表表现出规律性,可以按5个变量分为一组,共分18—22、23—27、28—32、33—37、38—42、43—47、48—52、53—57、58—62九个组,取300个麦穗的资料,进行整理,结果见表2.3,就可明显表示出其分布情况,大部分麦穗的粒数在28—52之间。

表 2.3 小麦农家品种 300 个麦穗每穗粒数的次数分布表

每穗粒数	次 数( $f$ )	频 率
18—22	3	0.0100
23—27	18	0.0600
28—32	38	0.1267
33—37	51	0.1700
38—42	68	0.2267
43—47	53	0.1766
48—52	41	0.1367
53—57	22	0.0733
58—62	6	0.0200
合 计	300	1.0000

2. 计量资料的整理 计量资料的整理不可能按计数资料的归组方法进行,一般采用组距式分组法。分组时须先确定全距组数、组距、各组上下限,然后按观测值的大小来归组。下面以150尾鲢鱼的体长资料(表2.4)为例,来说明计量资料的整理方法和具体步骤。

表 2.4 150 尾鲢鱼的体长(cm)

56	49	62	78	41	47	65	45	58	55
52	52	60	51	62	78	66	45	58	58
56	46	58	70	72	76	77	56	66	58
63	57	65	85	59	58	54	62	48	63
58	52	54	55	66	52	48	56	75	55
63	75	65	48	52	55	54	62	61	62
54	53	65	42	83	66	48	53	58	57
60	54	58	49	52	56	82	63	61	48
70	69	40	56	58	61	54	53	52	43
58	52	56	61	59	54	59	64	68	51
55	47	56	58	64	67	72	58	54	52
46	57	38	39	64	62	63	67	65	52
59	60	58	46	53	57	37	62	52	59
65	62	57	51	50	48	46	58	64	68
69	73	52	48	65	72	76	56	58	63

(1) 求全距。全距是样本数据资料中最大观测数与最小观测数的差值。它是整个样本的变异幅度。由表2.4,可以看出,鲢鱼体长最大值为85cm,最小值为37cm,因此,全距为 $85 - 37 = 48(\text{cm})$ 。

(2) 确定组数和组距。组数是根据样本观测数的多少及组距的大小来确定的,同时也