21 世纪高等院校教材——生物科学系列

生 物 统 计 学

（第三版）

李春喜　姜丽娜　邵　云　王文林　编著

科 学 出 版 社

北　京

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、生物学研究中试验资料的整理、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、χ^2 检验、方差分析、直线回归与相关分析、可直线化的非线性回归分析、协方差分析、多元回归与多元相关分析、逐步回归与通径分析和多项式回归分析，同时对抽样原理和方法、试验设计原理及对比设计、随机区组设计、平衡不完全区组设计、裂区设计、拉丁方设计、正交设计等常用试验设计及其统计分析也进行了详细叙述。在上述内容的基础上，对聚类分析、判断分析、主成分分析、因子分析、典型相关、时间序列分析等多元分析也作了简要介绍。

本书可供综合性大学、师范院校生物类及其相关专业的本科生作为教材使用，也可作为从事生命科学、农业科学、林业科学、医学、畜牧兽医、水产科学等专业的科研工作者、教师和研究生的参考书。

图书在版编目(CIP)数据

生物统计学/李春喜等编著.—3 版.—北京:科学出版社,2005.7
(21 世纪高等院校教材——生物科学系列)
ISBN 7-03-015155-0

Ⅰ.生… Ⅱ.①李…②姜…③邵…④王… Ⅲ.生物统计-高等学校-教材 Ⅳ.Q-332

中国版本图书馆 CIP 数据核字(2005)第 035171 号

责任编辑:周 辉 张晓春/责任校对:鲁 素
责任印制:安春生/封面设计:陈 敬

科学出版社 出版
北京东黄城根北街 16 号
邮政编码:100717
http://www.sciencep.com

印刷

科学出版社发行 各地新华书店经销
*
1997 年 8 月第 一 版 开本:850×1168 1/16
2000 年 11 月第 二 版 印张:23 1/2
2005 年 7 月第 三 版 字数:488 000
2005 年 7 月第一次印刷 印数:1—5 000

定价:28.00 元

(如有印装质量问题,我社负责调换〈环伟〉)

第三版前言

进入21世纪，生命科学已成为现代科学技术发展中最活跃、最具有活力和最富有挑战性的学科领域之一。生命科学从古到今，从对生物体整体的观察到种群、群落、生态圈、生物多样性的研究和利用显微技术对细胞结构的认识、利用分子生物学技术对生命活动机理的探索无不都伴随着数学方法的运用。随着生命科学研究的深入发展和相关学科的融合与渗透，数学方法运用于生命科学的范围将会越来越广泛，越来越深入。数学在生命科学上的应用，总的说来主要包括三个方面：一是应用数学知识来解决生命科学中的实际问题，如应用数学模型研究生物生长发育的过程，利用计算机模拟对其生育过程进行预测与控制等；二是数学方法与生命科学融合交叉产生新的学科，如数量生态学、数量遗传学、生物信息学等；三是从生命科学中提炼出数学问题进行研究，进而发展新的数学理论，如基因如何从时间、空间上来精确发育过程，多个核苷酸是如何构成功能基因的等。可见，生物数学对生命科学的发展是必不可少的。生物统计学作为生物数学的重要分支，在生命科学研究与探索的过程中发挥了巨大的推动作用。因此，各高等学校生物类专业都将生物统计学作为一门重要的基础课程纳入到教学体系中，促进了新型综合性生物学高级人才的培养。

为适应21世纪生命科学发展和生物学人才培养的要求，在本书1997年第一版、2000年第二版的基础上，对全书内容重新进行了编排和审核，增加了部分内容，修订和改正了原书中存在的一些错误。与前两版相比，本书突出了以下几个特点：①内容更加丰富，增加了平衡不完全区组设计、倒数函数曲线、通径分析等内容；②编排更科学，将全书分解为16章，各章节的安排更加注重了内容的循序渐进；③针对性更加明确，内容突出了本书主要作为生物类及相关专业教材这个重点，更换了第二版中部分针对性不强的例子，对各章思考练习题及其答案重新进行了核对，对相关名词与术语增加了英文标注，并重新编排了中英文对照索引，以便于学习和检索。

在本书第三版的修订和出版过程中，得到了科学出版社和河南师范大学的大力支持，周辉先生在书稿的编审方面做了大量工作，周其源、鲁旭阳、苗永平、李丹丹、侯小丽、张蓓蓓、肖长新、冯淑利、刘玲9位研究生承担了书稿的部分核对工作和习题答案的核对工作，在此一并感谢。

本书能够第三次出版，是广大读者支持和厚爱的结果。希望各位读者在阅读和使用过程中对本书的谬误和不妥之处给予批评指正。

李春喜
于河南师范大学
2005年1月

第二版前言

近代生物科学的发展有两个显著的特点:一个是向微观方向的发展,通过显微技术对生物的细胞和细胞结构进行深入研究;另一个是向宏观方向的发展,从生物体的器官、整体到种群、群落、生态圈进行研究。这两个发展方向的共同趋势都是需要运用数学方法对生物体、生物器官、细胞及分子结构所观察和实验的结果进行综合分析,研究各种因素间的相互作用,通过建立数学模型,并对模型进行数学推理,来发现和解释新的生命现象。随着科学的发展,数学方法在生物学研究中的应用会越来越广泛,其作用也将会越来越重要。因此,一门新兴的边缘性学科——生物数学也就应运而生了。在生物数学领域中,生物统计学是应用最早也最广泛的一门学科,起先是应用生物学科,后来是纯生物学科,它们都对生物统计学的应用有一定的深度和广度,特别是信息科学的迅猛发展和计算机的迅速普及,为在生物学研究中运用生物统计学原理和方法提供了更为广阔的空间。生物统计学作为基础性工具课程,越来越为高校生物类各专业所重视。

本书第一版出版以来,在部分高校生物类专业作为教材使用,一些科学研究单位也作为工具书应用,总的说来反映是良好的。有不少读者对本书进行了仔细研究,提出了不少修改意见,对本书第一版中出现的错误诚恳地提出了批评。根据读者的意见和生物统计学应用的需要,这次修订,对第一版各章节作了较大幅度的调整,将全书分为十四章,补充了拉丁方设计和裂区设计两种试验设计方法,将抽样原理和方法、常用试验设计及其统计分析放在了可直线化的非线性回归分析之后进行介绍,使章节编排体系更符合读者学习的要求。书中还增加了对全文关键词汇和术语的索引,并在书后附上了各章部分思考练习题答案。同时,对本书第一版中的不妥和错误之处进行了订正,更换了部分引用例题,以使这些例题更能反映本章内容和便于读者的学习和理解。

在本书第二版的修订和出版过程中,得到了河南省科委和科学出版社的大力支持,尚玉磊、邱宗波、董媛、史小琴同志承担了部分书稿的校对工作,在此一并表示感谢。

由于作者水平有限,谬误和不当之处,敬请读者批评指正!

李春喜
于河南师范大学
2000 年 4 月

第一版前言

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，以正确制订试验计划，科学地对试验结果进行分析，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，对高校生物类专业，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、样本频率的检验、方差同质性检验、非参数检验和 χ^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做了大量工作，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

目　　录

第一章

概　　论

本章提要　生物统计学是数理统计的原理和方法在生命科学领域的具体应用，它是运用数理统计原理和方法对生物有机体开展调查和试验，目的是以样本的特征来估计总体的特征，对所研究总体进行合理的推论，得到对客观事物本质和规律性的认识。生物统计学的主要研究内容包括试验设计和统计分析两大部分，其作用主要有四个方面：提供整理、描述数据资料的科学方法并确定其数量特征，判断试验结果的可靠性，提供由样本推断总体的方法，提供试验设计的原则。本章还介绍了生物统计学的发展概况和六组统计学常用术语。

第一节　生物统计学的概念

生物统计学（biostatistics）是数理统计（mathematical statistics）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和解释生物界各种现象和试验调查资料的科学，属于应用统计学的一个分支。随着生物学研究的不断发展，运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在自己的研究领域越来越普遍地应用生物统计分析方法，并把它作为学科自身发展的需要。

生物学研究的对象是生物有机体，与非生物相比，它具有更加特殊的复杂性。生物有机体的生长发育、生理活动、生化变化及有机体受外界各种随机因素的影响等，都使生物学研究的试验结果有较大的差异性，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迷惑，从而被人们忽视。因此，在生物学研究中，应用生物统计学就显得特别重要。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，目的是为了以样本的特征对未知总体进行推断，从特殊推导一般，对所研

究的总体作出合乎逻辑的推论，得到对客观事物的本质和规律性的认识。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。生物统计学的过程就是把数学的语言引入具体的生命科学领域，把具体生命科学领域中亟待研究的问题抽象为数学问题的过程。它是以数学的概率论和数理统计为基础，这其中要涉及到数列、排列、组合、矩阵、微积分等知识。作为一门工具课，生物统计学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 生物统计学的内容与作用

生物统计学的基本内容，概括起来主要包括试验设计(experimental design)和统计分析(statistical analysis)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，其中主要的有对比设计、随机区组设计、平衡不完全区组设计、拉丁方设计、裂区设计以及正交设计等。在统计分析中，主要包括数据资料的搜集和整理、数据特征数的计算、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制出一定形式的图表，使研究者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，其他条件都应控制一致，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他偶然因素的影响。偶然因素的影响就是造成试验误差的重要原因。要正确判断一个试验结果是由试验因素造成的还是由试验误差造成的，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。这是生物统计学的精髓所在，也是其成为一门科学的缘由。在这种推断中，统计学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，可以在较少人力、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，究其原因多是由于缺乏科学的统计学知识，从而使试验的效率大大降低。尽管统计学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，这种试验绝对不会成功，统计学原理和分析方法也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，一是政治科学的需要，二是当时贵族阶层对机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。统计学的发展过程大体上可以分为古典记录统计学、近代描述统计学和现代推断统计学三个阶段。

一、古典记录统计学

古典记录统计学(record satistics)的形成是在17世纪中叶至19世纪中叶。在最初兴起时，通过用文字或数字如实记录与分析国家社会经济状况，初步建立了统计研究的方法和规则。到概率论被引进之后，才逐渐成为一项成熟的方法。

瑞士数学家J.Bernoulli(1654～1705)系统论证了大数定律。后来，J.Bernoulli的后代D.Bernoulli(1700～1782)将概率论的理论应用到医学和人类保险。

法国天文学家、数学家P.S.Laplace(1749～1827)发展了概率论的研究，建立了严密的概率数学理论，并在天文学、物理学的研究中推广应用了概率论。他研究了最小二乘法，提出了“拉普拉斯定理”(中心极限定理的一部分)，初步建立了大样本推断的理论基础，为后人开创了抽样调查的方法。

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。德国天文学和数学家G.F.Gauss(1777～1855)在研究观察误差理论时，也独立推导出测量误差的概率分布方程，提出了“误差分布曲线”。这条分布曲线称为Gauss分布曲线，也就是正态分布曲线。Gauss对统计学的另一重要贡献是首先提出了统计上非常重要的最小二乘法。

二、近代描述统计学

近代描述统计学(description statistics)的形成是在19世纪中叶至20世纪上半叶，这个时期也是统计学用于生物学研究的开始和发展时期。

1870年，英国遗传学家F.Galton(1822～1911)在19世纪末应用统计方法研究人种特性和遗传，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，但由于他开创性地将统计方法应用于生物学研究，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.Pearson(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，1889 年发表了《自然界的遗传》一文，并于 1901 年创办了 Biometrika(生物统计学报或称为生物计量学报)权威杂志。在该杂志的创刊词中，Galton 和 Pearson 首次为他们所运用的统计方法论明确提出了“生物统计”(biometry)一词，Galton 解释为：所谓生物统计学，就是应用于生物学科中的现代统计方法。在《自然界的遗传》一文中，K.Pearson 首先提出了相关与回归分析问题，并给出了计算简单相关系数和复相关系数的计算公式。K.Pearson 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标卡方(χ^2)的检验问题，它在属性统计分析中有着广泛的应用。例如，遗传学孟德尔豌豆杂交试验中，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 χ^2 进行检验。

三、现代推断统计学

现代推断统计学(inference statistics)的形成是在 20 世纪初至 20 世纪中叶。随着社会科学和自然科学领域研究的不断深入，各种事物与现象之间的表面关系及未知的一些数量变化，要求采用推断的方法来掌握事物之间的真正联系并对事物进行预测。从描述统计学到推断统计学，这是统计学发展过程中的一个巨大飞跃。

K.Pearson 的学生 W.S.Gosset(1876～1937)对样本标准差进行了大量研究，于 1908 年以笔名“Student”在 Biometrika 上发表论文《平均数的概率误差》，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。因此，许多统计学家把 1908 年看作是统计推断理论发展史上的里程碑。

英国统计学家 R.A.Fisher(1890～1962)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，创立了方差和方差分析。在从事农业试验及数据分析研究时，他提出了随机区组法、拉丁方法和正交试验设计。1925 年，Fisher 发表了《试验研究工作中的统计方法》，对方差分析及协方差分析进一步作了完整的解释，从而推动和促进了农业科学、生物学和遗传学的研究与发展。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。

Newman(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设检验学说。假设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。

我国对生物统计学的应用始于 1913 年顾澄教授翻译的统计名著《统计学之理论》。这是英国统计学家在 1911 年出版的关于描述统计学的著作，也是英美数理统计学传入中国的开始。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、医学和农学的应用，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、网络技术的发展，SAS（statistical analysis system）、SPSS（statistical package for the social science）等国际通用软件的开发和应用以及生命科学研究领域的不断深入，生物统计学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、个体与样本

具有相同性质的个体所组成的集合称为总体（population），它是指研究对象的全体，而组成总体的基本单元称为个体（individual）。

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（infinite population）。例如，某一地区棉田棉铃虫的头数，可以认为是无限总体。另外，也可从抽象意义上来理解无限总体，比如通过临床试验来推断某一种药品比另一种药品的治愈率高，这里无限总体指的是一个理论性总体。个体有限的总体称为有限总体（finite population），如对某一班学生身高进行调查，这时总体是指这一班中每一名学生的身高。

要研究总体的性质，一般情况下我们无法一一对总体中的个体全部取出进行调查或研究。因为在实际研究过程中，我们常常会遇到两种难以克服的困难：一是总体的个体数目较多，甚至无限多；二是有时总体的数目虽然不多，但试验具有破坏性，或者试验费用很高，不允许做更多的试验，因而只能采取抽样的方法，从总体中抽取一部分个体进行研究，作为统计的依据。

从总体中抽出的若干个个体所构成的集合称为样本（sample），构成样本的每个个体称为样本单位（sample unit），样本个体数目的大小称为样本容量（sample size），记为 n。样本的作用在于估计总体。例如可以调查某一地区棉田 100 株棉花上的棉铃虫头数，来推断该地区棉铃虫的发生状况，以采取相应的对策。一般在生物学研究中，样本容量 $n < 30$ 的称为小样本，样本容量 $n \geqslant 30$ 的称为大样本。在一些计算和分析检验方法上，大、小样本是不同的。

二、变量与常数

相同性质的事物间表现差异性或差异特征的数据称为变量或变数（variable）。由于试验目的的不同，所选择的变量也不相同。如植物叶片叶绿素的含量，人体身高、体重、血糖含量，同窝动物的身长及生理指标等。变量通常记为 x，如 10 个人的身高在 155～

180cm 之间，共有 158，167，173，155，180，165，175，178，170，162cm 10 个变量值，记作 x_i（$i=1,2,\cdots,10$），表示 x_1 到 x_{10} 之间任一数值。变量的测得值称为变量值（value of variable）或观测值（observed value），亦称为资料（data）。

变量按其性质可分为连续变量和非连续变量。连续变量（continuous variable）表示在变量范围内可抽出某一范围的所有值，这种变量之间是连续的、无限的。如小麦的株高在 80～90cm，在此范围内可以取得无数个变量。非连续变量（discontinuous variable），也称为离散型变量（discrete variable），表示在变量数列中，仅能取得固定数值。如菌落中的菌数、单位面积水稻的茎数、小白鼠每胎产仔数等。

变量可以是定量的，也可以是定性的。定量变量（quantitative variable）亦称为数值变量（numerical variable），其变量值是定量的，表现为数值大小，一般有度量衡单位。如每个人的身高（cm）、体重（kg），出栏时猪的重量（kg）等。定性变量（qualitative variable）亦称为分类变量（categorical variable），其变量值是定性的，表示某个体属于几种互不相容的类型中的一种，如果蝇的翅有长翅与残翅，人的血型有 A、B、AB 和 O 型，豌豆花的颜色有白色、红色和紫色，等等。变量的类型是根据研究目的而确定的。根据需要，各类变量可以互相转化。如以人作为研究对象，观察某人群成年男子的血红蛋白含量（$mg.L^{-1}$），属于定量变量；若按血红蛋白含量正常与偏低分为两类，则属于定性变量。

常数（constant）是不能给予不同数值的变量，它代表事物特征和性质的数值，通常由变量计算而来，在一定过程中是不变的。如某样本平均数、标准差、变异系数等。

三、参数与统计数

参数（parameter）也称参量，是对一个总体特征的度量，常用希腊字母表示。如总体平均数 μ、总体标准差 σ 等均为参数。

总体一般都很大，有的甚至不可能取得，所以总体参数一般不可能计算出来。而可以通过对总体抽取样本，计算样本的统计数，来估计总体参数。从样本中计算所得的数值称为统计数（statistic），它是描述样本特征的数量，常用英文字母表示，如样本平均数 $\bar{x}$、样本标准差 s 等。样本统计数是总体参数的估计值。

四、效应与互作

引起试验差异的作用称为效应（effection），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（interaction），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、误差与错误

误差（error）也叫试验误差（experimental error），是指试验中不可控因素所引起的观测

值偏离真值的差异。试验中出现的误差可以分为两类:随机误差和系统误差。随机误差(random error)也称为抽样误差(sampling error)、偶然误差(accidental error),它是由于试验中许多无法控制的偶然因素所造成的试验结果与真实结果之间产生的误差,是不可避免的。我们可以通过试验设计和精心管理设法减小随机误差,而不能完全消除。随机误差影响试验的精确性。统计上的试验误差就是指随机误差。增加抽样或试验次数,可以降低随机误差。系统误差(systematic error)也称为片面误差(lopsided error),是由于试验处理以外的其他条件明显不一致所产生的带有倾向性的或定向性的偏差。系统误差主要由一些相对固定的因素引起,例如仪器调校的差异,各批药品间的差异,不同操作者操作习惯的差异等。系统误差在某种程度上是可控制的,只要试验工作做得精细,在试验过程中是可以克服的。

错误(mistake)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。这类错误原则上是不允许产生的。

六、准确性与精确性

准确性也称为准确度(accuracy),指在调查或试验中某一试验指标或性状的观测值与其真值接近的程度。精确性也称精确度(precision),指调查或试验中同一试验指标或性状的重复观测值彼此接近程度的大小。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,而精确性则是反映多次测定值的变异程度。

不同研究对精确度的要求是不一样的,一般来说,化学测量应当有较高的精确性,动物实验或医学临床试验由于试验对象个体差异及测定条件的影响,较难控制精确性,但应尽量将其控制在专业规定的容许范围内。

思考练习题

习题 1.1 什么是生物统计学?生物统计学的主要内容和作用是什么?

习题 1.2 解释并举例说明以下概念:总体、个体、样本、样本容量、变量、参数、统计数、效应、互作、随机误差、系统误差、准确性、精确性。

习题 1.3 误差与错误有何区别?

第二章

试验资料的整理与特征数的计算

本章提要 试验资料的搜集与整理是数据资料处理的首要环节。资料的搜集常用的方法有调查和试验，资料的整理一般需要通过对原始资料进行检查、核对，制作次数分布表和次数分布图来完成。试验资料均具有集中性和离散性两种基本特征，平均数是反映集中性的特征数，主要包括算术平均数、中位数、众数、几何平均数等，而反映离散性的特征数是变异数，主要包括极差、方差、标准差和变异系数等。

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(data)。这些资料在未整理之前，一般是分散的、零星的和孤立的，是一堆无序的数字。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，然后列成统计表，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(quantitative character)和质量性状(qualitative character)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(data of quantitative character)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(enumenation data)，也称为非连续变量资料(data of dis-

continous variable)，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，…，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(measurement data)，也称为连续变量资料(data of continuous variable)，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(data of qualitative character)，也称属性资料(attribute data)，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，小麦芒的有无，茸毛的有无；果蝇的长翅与残翅；人血型的 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(frequency counting)　于一定总体内，根据某一质量性状的类别统计其次数或频数(frequency)，以次数或频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(point system)　这种方法是用数字级别表示某现象在表现程度上的差别。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)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进行数量化。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进行。

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，归根结底是用样本来估计总体的问题。因此，样本资料的搜集(collection)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。无论是调查还是试验，统计学对原始资料都要求完整和准确。

(一) 调查

资料的调查(survey)有两种方法，一种是普查，另一种是抽样调查。普查(census)是指对研究对象的每一个个体都进行测量或度量的一种全面调查(complete survey)，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内进行，主要目的是摸清研究对

象的基本情况。在生物学研究中,普查仅仅是在极少数情况下才能进行的调查,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(sampling survey)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(random sample),也就是用随机抽样(random sampling)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他个体的影响。第二条适合于无限总体,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,将在第七章叙述。

(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(experiment)来获取样本资料,安排这些试验时,要设置试验处理,遵循随机、重复和局部控制三项基本原则进行设计。常见的试验设计方法主要有:对比设计、随机区组设计、平衡不完全区组设计、裂区设计、拉丁方设计、正交设计等,具体内容将在第八章和第九章中介绍。

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资料(row data)后,要对全部数据进行检查与核对,才能进行数据的整理(collation)。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。主要核对原始资料的测量和记载有无差错,检查原始资料有无遗失、重复不合理的归并和特大、特小异常值的出现。对个别缺失的数据,可以进行缺失数据估计,对重复、错误和异常值应予以删除或订正,但不能随意改动,必要时要进行复查或重新试验。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(频数)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，就需将数据分成若干组，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（频数）分布表（frequency table），作出次数（频数）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（grouping method of monomial）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 7 组，将各组所属数据进行统计，得出各组次数，计算出各组的频率（frequency）和累积频率（cumulative frequency），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	频　率	累积频率
11	2	0.02	0.02
12	7	0.07	0.09
13	19	0.19	0.28
14	35	0.35	0.63
15	21	0.21	0.84
16	11	0.11	0.95
17	5	0.05	1.00

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。如研究不同小麦品种的 300 个麦穗，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 9 个组，将 300 个麦穗的资料进行归组，计算出各组的次数、频率和累积频率，结果见表 2-3，就可明显表示出其分布情况，即大部

分麦穗的粒数在 28～52 之间。

表 2-3 不同小麦品种 300 个麦穗穗粒数的次数分布表

穗粒数	次 数	频 率	累积频率
18～22	3	0.0100	0.0100
23～27	18	0.0600	0.0700
28～32	38	0.1267	0.1967
33～37	51	0.1700	0.3667
38～42	68	0.2267	0.5934
43～47	53	0.1766	0.7700
48～52	41	0.1367	0.9067
53～57	22	0.0733	0.9800
58～62	6	0.0200	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(grouping method of class interval)。分组时需先确定全距、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(range)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(cm)。

(2)确定组数和组距。组数(number of classes)是根据样本观测数的多少及组距的大小来确定的,同时也考虑到对资料要求的精确度以及进一步计算是否方便。组数与组距有密切的关系。组数多些,组距相应就变小,组数越多所求得的统计数就越精确,但不便于计算;组数太少,组距就相应增大,虽然计算方便,但所计算的统计数的精确度较差。为了使两方面都能够协调,组数不宜太多或太少。在确定组数和组距时,应考虑样本容量的

大小、全距的大小、便于计算、能反映出资料的真实面貌等因素。通常划分组数可参照表2-5 样本容量与分组数的关系来确定。

表 2-5　样本容量与分组数的关系

样本容量	分组数
30～60	5～8
60～100	7～10
100～200	9～12
200～500	10～18
500 以上	15～30

组数确定好后，还须确定组距(class interval)。组距是指每组内的上下限范围。分组时要求各组的距离相同。组距的大小是由全距和组数所确定的：

$$\text{组距} = \frac{\text{全距}}{\text{组数}}$$

表 2-4 鲢鱼体长的样本容量为 150，查表 2-5，组数为 9～12 组，这里取 10 组，则组距应为：

$$\frac{48}{10} = 4.8(\text{cm})$$

为分组方便，以 5cm 作为组距。

(3)确定组限和组中值。组限(class limit)是指每个组变量值的起止界限。每个组有两个组限，一个下限和一个上限。在确定最小一组的下限时，必须把资料中最小的数值包括在内，因此，下限要比最小值小些。为了计算方便，组限可取到 10 分位或 5 分位数上，如表 2-4 中最小值为 37cm，第一组的下限可定为 35cm，上限定为 40cm，即 35～40cm 为第一组，凡大于 35cm 小于 40cm 的变量均归于这一组，等于或大于 40cm 的变量列入下一组。确定最大一组的上限时，必须大于资料中的最大值。为了使各组界限明确，避免重叠，目前在写法上，每组只写下限，不写上限，如表 2-4 资料分组写成 35～，40～，…，85～。

组中值(class midvalue)是两个组限下限和上限的中间值。在资料分组时，为了避免第一组中观测数过多，一般第一组的组中值最好接近或等于资料中的最小值。其计算公式为：

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

或

$$\text{组中值} = \text{下限} + \frac{1}{2}\text{组距} = \text{上限} - \frac{1}{2}\text{组距}$$

(4)分组，编制次数分布表。确定好组数和各组上下限后，可按原始资料中各观测数的次序，把各个数值归于各组，即进行分组(class fication)。一般用“正”字划计法或卡片法来计算各组的观测数次数。全部观测数归组后，即可求出各组的次数、频率和累积频率，制成一个次数分布表(表 2-6)。这种次数分布表不仅便于观察，而且可根据它绘制成次数分布图，计算平均数和标准差等特征数。

表 2-6 150 尾鲢鱼体长(cm)的次数分布表

组限(cm)	组中值(cm)	次 数	频 率	累积频率
35～	37.5	3	0.0200	0.0200
40～	42.5	4	0.0267	0.0467
45～	47.5	17	0.1133	0.1600
50～	52.5	28	0.1867	0.3467
55～	57.5	40	0.2666	0.6133
60～	62.5	25	0.1667	0.7800
65～	67.5	17	0.1133	0.8933
70～	72.5	6	0.0400	0.9333
75～	77.5	7	0.0467	0.9800
80～	82.5	2	0.0133	0.9933
85～	87.5	1	0.0067	1.0000

(三) 次数(频数)分布图

次数(频数)分布图(frequency chart)就是把次数分布资料画成统计图形。次数分布图可以更直观地观察各组变量次数分布的情况,形象地把资料特征表达出来。常用的次数分布图有条形图、饼图、直方图、多边形图和散点图等。

1. 条形图 条形图又称柱形图(bar chart),适合于表示计数资料和属性资料的次数分布。作图时,用横坐标表示变量的自然值,纵坐标表示次数,每一个次数数据于相应自然值的位置分别截取一定的宽度和相应次数高度的长方形。每个长方形之间要隔出一定距离,以区别于下面要介绍的直方图。以 100 只来亨鸡每月产蛋数的次数分布为例作出条形图(图 2-1)。

2. 饼图 饼图(pie chart)适合于表示计数资料和属性资料的次数分布。作图时,把饼图的全面积看成 1,求出各观测值次数占观测值总数的百分比,即构成比(或频率),按构成比将圆饼分成若干份,以扇形面积大小分别表示各个变量的比例。以 100 只来亨鸡每月产蛋数的次数分布为例作出饼图(图 2-2)。

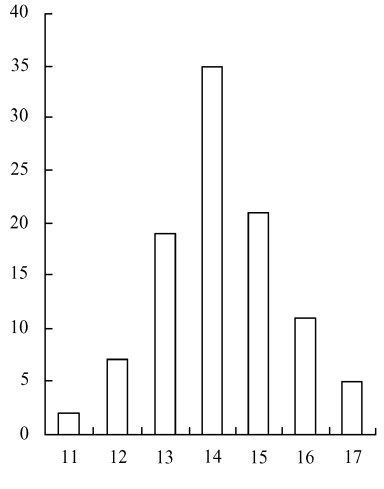

图 2-1 来亨鸡月产蛋数次数分布条形图

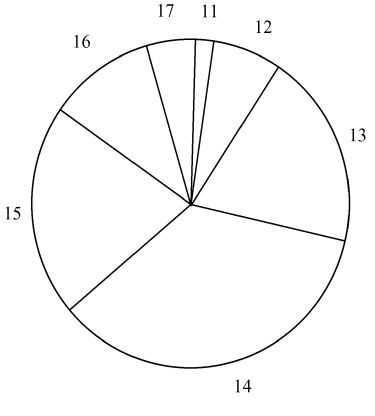

图 2-2 来亨鸡月产蛋数次数分布饼图

3. 直方图　直方图(histogram)又称矩形图,适合于表示计量资料的次数分布。其作图方法与条形图相似,以横坐标表示各组组限,纵坐标表示次数,截取一定距离代表组限大小和次数多少,用直线连接起来,构成一个个长方形。各组间一般没有距离,前一组上限与后一组下限可合并公用。以 150 尾鲢鱼体长的次数分布为例作出直方图(图 2-3)。

4. 多边形图　多边形图(polygon chart)也称折线图(broken-line chart),也是表示计量资料次数分布的一种方法。作折线图时,以横坐标表示各组组中值,纵坐标表示次数,在各组组中值的垂线上,按该组次数应占高度标记一个点,把相邻的点用直线段顺次连接起来,即成多边形图。以 150 尾鲢鱼体长的次数分布为例作出多边形图(图 2-4)。

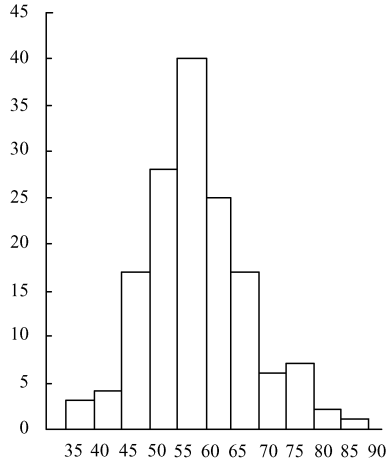

图 2-3　鲢鱼体长次数分布直方图

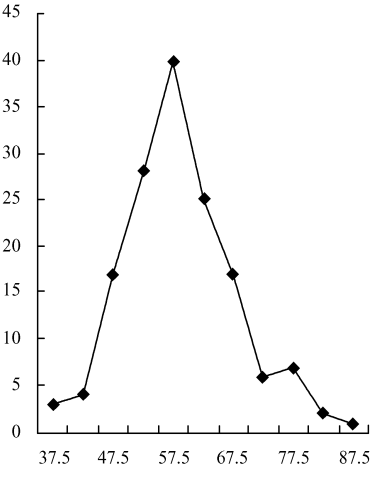

图 2-4　鲢鱼体长次数分布多边形图

5. 散点图　散点图又称散布图(scatter chart),适合于表示计数资料和计量资料的次数分布。图中横坐标表示 x 变量,纵坐标表示 y 变量。它是以点的分布反映变量之间相关情况,根据图中的各点分布走向和密集程度来判断变量之间关系的。以 100 只来亨鸡

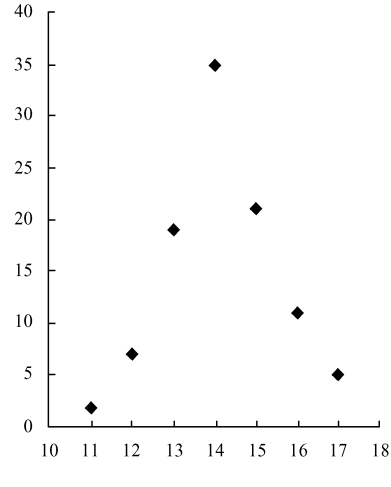

图 2-5　来亨鸡月产蛋数次数分布散点图

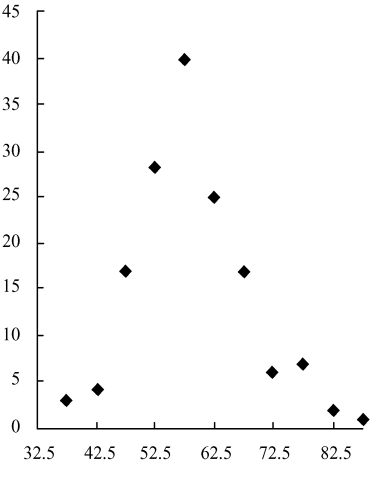

图 2-6　鲢鱼体长次数分布散点图

每月产蛋数的次数分布和 150 尾鲢鱼体长的次数分布为例作出散点图(图 2-5、图 2-6)。

通过以上次数分布图,我们可以比次数分布表更直观地看出各观测资料的变化趋势,各资料的分布中心及其变异趋势均可以很直观地得到描述。同样,也可以按照资料分组的频率值绘制成频率分布图。

第二节 试验资料特征数的计算

由上节所述的次数分布,我们可以看出变量的分布具有两种明显的基本特征,即集中性和离散性。集中性(centrality)是变量在趋势上有着向某一中心聚集,或者说以某一数值为中心而分布的性质。离散性(discreteness)是变量有着离中分散变异的性质。为了反映变量分布的这两个基本性质,必须计算它们的特征数(eigenvalue)。反映集中性的特征数是平均数,其中应用最普遍的是算术平均数。此外还有几何平均数、中位数和众数等。反映离散性的特征数为变异数,常用的指标是极差、方差、标准差和变异系数等,其中最为常用的是标准差,它是变量的平均变异程度的度量。

一、平 均 数

平均数(mean)是计量资料的代表值,表示资料中观测值的中心位置,并且可作为资料的代表与另一组资料相比较,以确定二者相差的情况。

(一) 平均数的种类

平均数的种类较多,主要有以下四种:

1. 算术平均数 总体或样本资料中各个观测值的总和除以观测值的个数所得的商,称为算术平均数(arithmetic mean)。对于一具有 N 个观测值的有限总体,其观测数为 $x_1, x_2, \cdots, x_N$,则该总体算术平均数(arithmetic mean of the population)为:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

对于具有 n 个观测数 $x_1, x_2, \cdots, x_n$ 的样本,其样本算术平均数(arithmetic mean of the sample)为:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

式(2.1)和式(2.2)中,$\sum$ 为求和符号,$\sum_{i=1}^{n} x_i$ 表示从 x_i 的 $i=1$ 一直加到 $i=n$,也可简写为 $\sum_i x_i$ 或 $\sum x$,$\bar{x}$ 是 μ 的估计值。因 $\bar{x}$ 应用广泛,常简称平均数或均数 。

2. 中位数 将试验或调查的资料中所有观测值依大小顺序排列,居于中间位置的观测值称为中位数(median),以 M_d 表示。当观测值个数 n 为奇数时,中位数是第 $(n+1)/2$ 位置的观测值;当观测值个数 n 为偶数时,中位数是第 $n/2$ 和 $n/2+1$ 位置的两个观测值之和的 1/2。

3. 众数　资料中出现次数最多的那个观测值或次数最多一组的中点值(组中值),称为众数(mode),以 M_o 表示。

4. 几何平均数　资料中有 n 个观测数,其乘积开 n 次方所得的数值,称为几何平均数(geometric mean)。几何平均数适用于变量 x 为对数正态分布,经对数转换后呈正态分布的资料。其计算公式为:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{2.3}$$

上述四种平均数中,算术平均数是最常用的平均数,中位数、众数和几何平均数使用较少。

(二) 算术平均数的计算方法

1. 直接计算法　当样本较小时可根据算术平均数的定义直接进行。

【例 2.1】 随机抽取 20 株小麦,其株高(cm)分别为 82,79,85,84,86,84,83,82,83,83,84,81,80,81,82,81,82,82,82,80,求小麦的平均株高。

根据平均数的定义,由式(2.2),得:

$$\bar{x} = \frac{1}{n}\sum x = \frac{1}{20} \times (82 + 79 + \cdots + 80) = 82.3(\text{cm})$$

2. 减去(或加上)常数法　若变量 x_i 的值都较大(或较小),且接近某一常数 a 时,可将它们的值都减去(或加上)常数 a,得到一组新的数据,然后再计算平均数,最后重新加上(或减去)常数 a 即得到 $\bar{x}$。以减去常数 a 为例,设 $y_1 = x_1 - a, y_2 = x_2 - a, \cdots, y_n = x_n - a$,则有 $x_1 = y_1 + a, x_2 = y_2 + a, \cdots, x_n = y_n + a$,于是有:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}(y_i + a) = \frac{1}{n}\sum_{i=1}^{n} y_i + \frac{1}{n}(na) = \frac{1}{n}\sum y + a \tag{2.4}$$

【例 2.2】 利用减去常数法,计算例 2.1 的平均数 $\bar{x}$。

设 $a = 80$,则有 $y_1 = 82 - 80 = 2, y_2 = 79 - 80 = -1, \cdots, y_{20} = 80 - 80 = 0$,代入式(2.4),得:

$$\bar{x} = \frac{1}{20} \times [2 + (-1) + \cdots + 0] + 80 = 82.3(\text{cm})$$

3. 加权平均法　在具有 n 个观测数的样本中,如果观测数 x_1 出现 f_1 次,观测数 x_2 出现 f_2 次,……,观测数 x_m 出现 f_m 次,且 $f_1 + f_2 + \cdots + f_m = n$,这时则有:

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_m x_m}{f_1 + f_2 + \cdots + f_m} = \frac{1}{n}\sum_{i=1}^{m} f_i x_i \tag{2.5}$$

这里,f_i 可理解为 x_i 在平均数中的"权数(weight)",即数值相同的观测数出现的次数,因而上式所求得的 $\bar{x}$ 称为加权平均数(weighted mean)。

【例 2.3】 利用加权平均法,计算例 2.1 的加权平均数。

先整理 20 个小麦株高数据如表 2-7。

由式(2.5),得:

$$\bar{x} = \frac{1}{20} \times (79 \times 1 + 80 \times 2 + \cdots + 86 \times 1) = 82.3(\text{cm})$$

表 2-7 小麦 20 个株高(cm)数据的次数分布

株高 x	次数 f	fx	fx^2
79	1	79	6241
80	2	160	12800
81	3	243	19683
82	6	492	40344
83	3	249	20667
84	3	252	21168
85	1	85	7225
86	1	86	7396
总和	$\sum f = 20$	$\sum fx = 1646$	$\sum fx^2 = 135524$

(三) 算术平均数的重要特性

(1)样本中各观测值与其平均数之差——离均差(deviation from mean)——的总和等于零。证明如下:

$$\begin{aligned}\sum(x-\bar{x}) &= (x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})\\ &= (x_1+x_2+\cdots+x_n)-n\bar{x}\\ &= \sum x-n\bar{x}\end{aligned}$$

因为 $\bar{x}=\dfrac{\sum x}{n}$, 所以 $\sum x = n\bar{x}$, 故:

$$\sum(x-\bar{x}) = \sum x - n\bar{x} = 0 \tag{2.6}$$

(2) 样本中各观测值与其平均数之差平方的总和,较各观测值与任一数值离差的平方和为小,即离均差平方和(mean deviation sum of squares)为最小。设 a 为 $\bar{x}$ 以外的任何数值,则 $\sum(x-\bar{x})^2 < \sum(x-a)^2$。证明如下:

$$\begin{aligned}\sum(x-a)^2 &= \sum[(x-\bar{x})+(\bar{x}-a)]^2\\ &= \sum[(x-\bar{x})^2+2(x-\bar{x})(\bar{x}-a)+(\bar{x}-a)^2]\\ &= \sum(x-\bar{x})^2+2\sum(x-\bar{x})(\bar{x}-a)+\sum(\bar{x}-a)^2\\ &= \sum(x-\bar{x})^2+n(\bar{x}-a)^2\end{aligned}$$

已知 $\sum(x-\bar{x})=0$, 因此 $2\sum(x-\bar{x})(\bar{x}-a)=0$。

因为 $n(\bar{x}-a)^2$ 必大于 0,所以,有:

$$\sum(x-\bar{x})^2 < \sum(x-a)^2 \tag{2.7}$$

(四) 算术平均数的作用

算术平均数是描述观测资料的重要特征数,它的作用主要有以下两点:

(1)指出一数据资料内变量的中心位置,标志着资料所代表性状的数量水平和质量水平;

(2)作为样本或资料的代表数与其他资料进行比较。