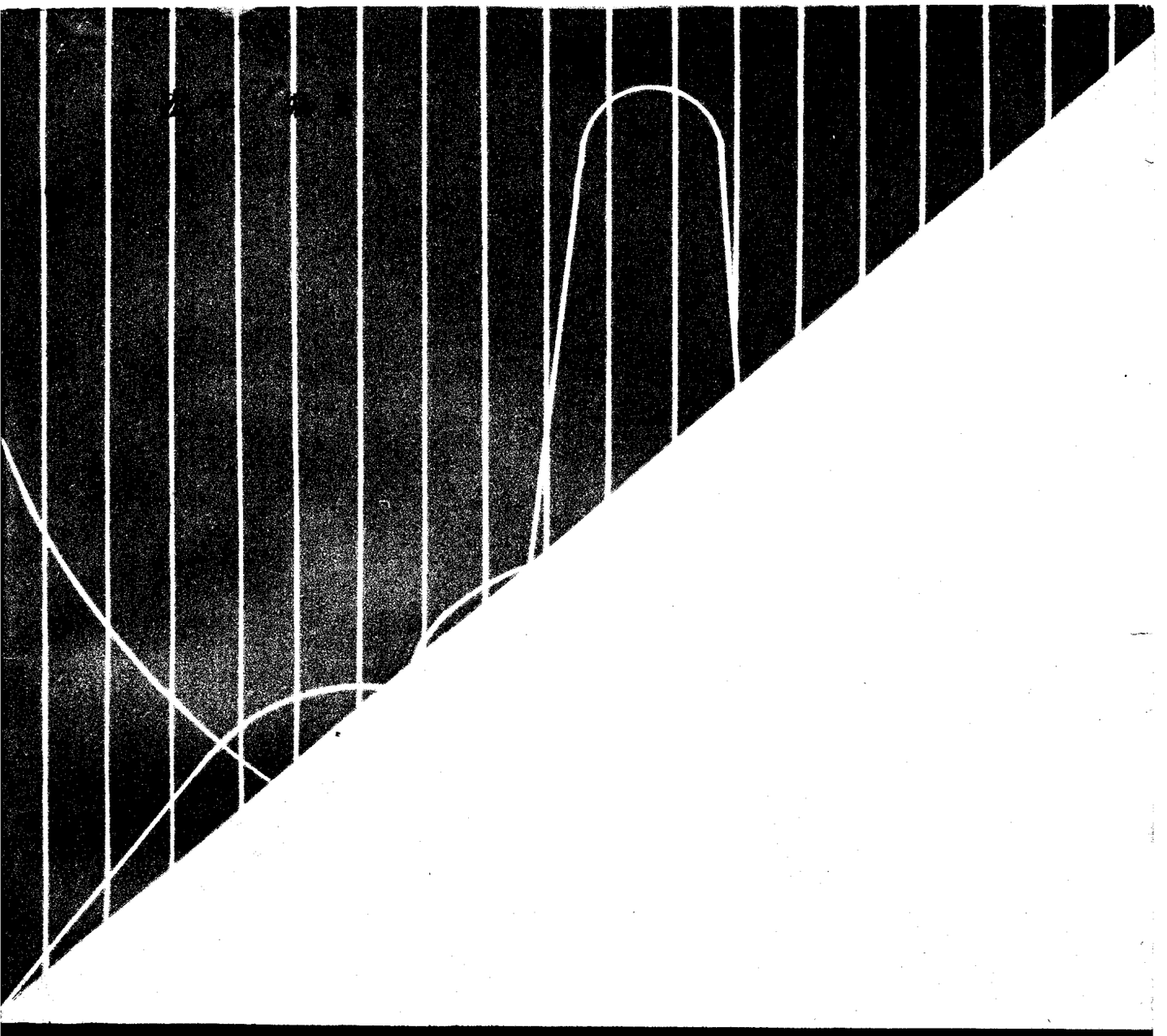

生物统计学

内 容 提 要

生物统计学是综合性大学生物系和有关专业的一门重要工具课。因此本书涉及的内容除考虑到综合性大学生物系各专业对生物统计学的要求外，还考虑了有关林、农、牧、医专业及从事生物科学研究的实际需要。全书共分十二章，比较全面地介绍了生物统计学的内容。并于每章后附有习题，以作巩固学习之用。同时，为了教学及自学方便，书后还附有五个附录及参考文献。

本书可作为综合性大学生物系及有关农、林、牧专业的教材，并可供自学者参考。

生物统计学

王宏年 编著

兰州大学出版社出版

（兰州大学校内）

兰州八一印刷厂印刷 甘肃省新华书店发行

开本：787×1092毫米 1/16 印张：22

1988年5月第1版 1988年5月第1次印刷

字数：506千字 印数：1—2000册

ISBN 7—311—00069—6/Q·6

书号：13402·26 定价：4.37元

前　言

生物统计学是应用数学的一个分支。近年来，随着广泛的应用而得到迅速发展，新的理论和新的实验也不断出现。对综合性大学生物系及有关专业来说，它是一门重要的工具课。考虑到这一点，本书不去追求数学理论的介绍和数学上的严密性，而是恰当地对统计学原理给予详细介绍，重点放在讲解如何应用、运算方法，并附以例题加以深化，以适应教学上的需要。

本书是在多年来为学生编写的讲义的基础上，吸收有关文献的精华及有关教材的优点编著而成。涉及的内容既考虑到综合性大学生物系各专业对生物统计学的要求，也考虑到有关农、林、牧、医专业及从事生物科学研究的实际需要，全书共分十二章，比较全面的介绍了生物统计学的基本内容。在每章后附以习题，以作巩固学习之用。为了教学及自学上的方便，书后附有五个附录。

由于编著者水平有限，在编写过程中，尽管做了很大努力，疏漏及错误之处在所难免。恳请使用本书的同行、读者或学生给予热情指正。

本书编写过程中得到中国农业科学院兰州畜牧研究所吴仁润教授的指导，兰州大学出版社张芸湘副编审协助，景文野同志清绘了插图，编者在此表示衷心的感谢。

编者

1987年10月9日

序

目前，在我国已出版的生物统计学专著并不多，而可以做为综合性大学生物系和农林畜牧院校相关专业的生物统计学教材也不能满足需要。编著者积多年教学之经验，几易其稿撰写的《生物统计学》，对于弥补这一不足，无疑具有重要意义。

生物科学使人类认识了生命过程中的许多规律，并在实践中解决了农业、林业、畜牧业和医学中的许多课题。然而，在认识生命现象规律的过程中，除了物理学、化学起了很大作用以外，数学的作用也绝不可低估。数学进入生物学，使生物学从定性描述走向定量分析，对于加深人们对生命现象本质的理解意义颇为深远。由于生物统计学的开创性贡献，从而使生物统计学成为综合大学生物系和农、林、牧、医等高等院校相关专业的必修课。生物科学研究的对象是复杂的有机体，与非生物相比，生物具有更加特殊的复杂性。由于有机体要常常受到内外环境因素的影响，从而使生物科学试验结果必须产生许多较大的差异性，这种较大的差异性常常会掩盖了有机体内在的特殊规律性；在生物科学试验中，如果对杂乱无章的数据不加适当的处理和分析，也就很难发现实验资料所体现的有机体的某种规律性。正由于如此，生物统计学所阐明的生物统计、分析方法对生物科学研究就显得十分重要。事实上、只有合理地应用生物统计原理和方法对生物实验进行设计，对数据进行处理和分析，才能得出合乎科学的结论。所以，无论是初学生物学专业或相关专业的学生，还是生物学的教学、科研人员，熟悉和掌握生物统计的原理和方法就显得十分重要。生物统计的原理和方法在生物学的各个领域，在农学、育种学、畜牧学和医学等方面的科学研究中的广泛应用，也充分地说明了这一点。

可以说，生物统计学是应用数学的一个分支。它是把数理统计的原理和方法用于生物科学研究中的数量资料的搜集、整理和分析的科学；也就是说使用数理统计的基本原理和方法进行生物科研的实验设计，资料的整理和归纳，以分析、推断生命的规律性。作为应用数学分支的生物统计学，并不去过多地讨论数学原理，而是把其原理作为解释生命现象科学方法的依据，内容偏重于统计原理和方法的介绍，而不去追求数学理论的介绍和数学上的严紧性。因此可以说，生物统计学是一门工具课。

王宏年先生有生物统计学的丰富教学经验。他编著的这本书，在国内已出版的同类著作中，有其独到之处。这就是他不拘泥于数学理论上的推导，又紧紧围绕基本的统计原理的讨论，着力向读者详细地介绍统计原理在生物学中的应用。该书对其所述的原理和方法讲得都比较全面而详细，并附以例题；特别是对例题解答得十分详细。因此，该书无论是当作教材，还是自学者自己去读，都是很适宜的。

吴仁润

1987年10月10日

目　录

第一章　绪　论

一、生物统计分析的意义

在认识自然界规律过程中，固然每门学科都以研究某一方面的规律为自身的重要任务，然而，各学科之间相互配合、应用合理的工具，对于人们认识自然、掌握规律，却有其重要的作用。随着生物科研工作的不断发展，应用统计分析，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（biometry）在应用过程中曾经受到过挫折，但绝大多数生物学工作者，农学家，育种学家、畜牧学家和医学工作者还是越来越普遍地把生物统计方法用在自己的科学研究中，并且变为各学科自身发展的需要，受到人们的重视。

生物学研究的对象是复杂的有机体，有植物、动物和微生物，与非生物相比，它具有更加特殊的复杂性。有机体内在的生理活动，生化变化；有机体受外界环境作用的影响等，都使生物科学试验结果产生了许多较大的差异性，这种较大的差异性往往会掩盖了生物体内在的特殊规律。在生物科学试验中，大量实验资料内部潜在的规律性，也容易被杂乱无章的数据所迷惑，容易被人们所忽视。因而，生物统计方法对生物科学研究就显得更为重要。实践证明，也只有正确地应用生物统计原理和方法对实验进行合理设计，对数据进行客观地分析，才能科学地得出合理地结论。

众所周知，应用数学逻辑，分析解释数量资料，称为数理统计或统计分析。数理统计方法可以用于处理各类数量资料，它是从数量方面综合认识事物规律性的一种重要工具，是通过对事物的量的研究达到认识其质的一种科学方法。把数理统计的科学原理和方法用于生物科学研究中的数量资料的搜集整理和分析的科学称为生物统计学，即用数理统计的原理和方法进行生物科研实验设计，资料的整理、归纳，以分析、推断生物界各种生命现象的本质，寻找出生命现象各方面的规律。

生物统计学是应用数学的一个分支，它以数学的概率论为基础。除概率论外，生物统计也涉及到数学中的其它知识，如数列、矩阵、排列、组合等。但生物统计学并不去过多地讨论数学原理，而是把其原理作为解释生命现象科学方法的一种依据。另外，生物统计学作为一门重要工具课，其内容偏重于统计原理和方法的介绍，而不大注意追究其数学推导和数学理论的研究。

二、生物统计的基本内容及特点

生物科学研究中所遇见的变异量，都是服从于一定的理论分布的。简单地说所谓理论分布就是指在某一性状或特性的总体中，各个个体部分出现的次数占总体的百分数。由于自然界生命现象的复杂性，生物体的性状或特性的变异量的分布表现出多样性，有的表现出连续性变异的特性，如小麦株高，动物体重的变异，其变异量是服从常态分布的；另一些性状则表现为非连续性变异，如动物的性比例，小麦芒的有无等是服从于二项分布的；还有的现象

是服从普阿松分布的。因此，研究这些现象特征，需要有概率和理论分布的基础知识。

从生物统计的基本作用上来讲，其内容可概括为以下三个方面：

1. 提供整理资料的方法、确定某些性状和特性的数量特征，是生物统计的基础部分，即描述统计（descriptive statistics）部分，如资料的整理，把观察、记载、调查的个体变数条理化；然后计算一群变数的平均数，掌握个体间的共同特性，即代表性；再对变量变异程度大小进行计算，估计、了解变量离散程度等。这是统计分析的最基础部分。例如，测得500株小麦植株高度的资料，统计分析的第一步就是把杂乱无章的一堆数据条理化，了解最高、最低植株的高度，计算500个数据的平均数和标准差，了解该小麦品种株高的集中情况和离散情况并以图表示之。

2. 判断试验结果的可靠性。进行科学试验，往往会得到不同的结果、即使同样试验重复一次，结果也不尽相同。不同品种，不同处理，不同药物作用等，均可得到一些有差异的资料；要了解品种、药物的作用，就需要用统计分析方法进行科学分析，使我们对资料间差异情况的可靠性有一个正确地认识，了解其差异是否真实。在判断试验结果的可靠程度时，主要包括两个方面，一种是判断试验结果与理论数值之间的符合程度，如用χ^2测验处理分析的资料；另一种是判断试验处理间（或品种间、药物间）差异的可靠程度或差异的显著程度，如用t测验、F测验所分析的资料。

3. 测定两个或两个以上变数群之间的关系，了解两种事物或现象之间的内在联系，了解它们之间联系的形式，方向和程度，进而了解一事物变化对另一事物变化的影响程度，从而对现象进行预测和判断，也称作统计推断方法。（method of estimative statistics）。如研究肥料与作物产量之间的关系时，不仅要研究二者间联系的方向、程度，而且，进而了解肥料增减对产量有多大的影响。研究亲子代之间的关系，同胞间的关系、玉米茎的粗细与果穗长短之间的关系等，均要用统计推断方法进行分析，即用相关分析法和回归分析法进行分析。

生物统计为试验设计提供了科学依据，因而研究试验设计的原理和方法的也是统计学内容之一。

生物统计分析的特点是根据有限数量的样本观察数据来推断和估计事物的总体。在一般情况下，总体的平均数和标准差是无法用实验求得的。要了解这个事物的特性，我们只能从总体中抽取一部分个体进行测量，求得样本的平均数和标准差，用它们来估计总体的性质。因此，生物统计分析总是带有其不确定性，需要使用一个概率量度来给以补充。因为是用样本来估计总体，给我们提出了另一个问题，样本对总体的代表性有多大、即用样本估计总体的偏差有多大？这个问题，除了与抽样方法有关系外，更重要地是样本包含的个体数量有多大，即与样本含量大小有关。在一定可能的条件下，样本包含个体数量越大，对总体的代表性越大，估计的偏差越小，这是根据大数定理（law of large number），即资料愈多愈能反映出事物的客观规律，资料太少，就易于导致错误的结论。但是，统计分析也不认为观测的个体数量越多越好，因为数量太多，进行实验的次数也就加多，记载、调查统计工作，不仅需要更多的人力，物力和时间，而且由于年度、季节的影响，药物来源及批号的变动，工作人员的调动更换等，反而会使实验结果参杂着许多不易控制的因素，降低了数据的质量，增加了误差，这样的结果，即使有大量资料得，出的结论也不一定可靠。所以，个体数量的多少，要视研究的对象和研究问题的性质来决定。例如在确定进行实验的动物数

量时，要看动物个体间差异大小，实验的目的、要求，动物的来源以及实验的其它条件来确定。另外，还要注意观测资料的同质性，即材料的来源，特点、特性要相同，否则就不具备研究对象的共性，当然结果也就失去了意义。

统计分析过程是依实验结果所得数据为依据对原设立的“假说”进行客观评定。应该了解的是，在评估“假说”过程中，我们只能否定那些不准确的“假说”，而不能去证明那些准确的“假说”。

三、生物统计与实验设计

生物科研工作都是要通过试验进行的，要使试验工作准确而有成效，一方面要提高试验技术，使试验达到更高的准确度（accuracy）；另一方面，也要对试验进行精心设计，要尽量利用现有条件取得最充分的资料。

生物统计与科学试验有密切关系，它可以为试验设计提供合理地依据，而试验设计又是生物统计方法的进一步运用。

生物统计可以帮助我们充分的认识试验误差和掌握误差，这就使我们在进行试验设计时有足够的理论依据。要比较全面地考虑到各种影响试验效果的可能条件。比如对试验重复数的设置，对随机排列和随机取样的运用，要考虑引起误差的因素。在进行动物饲料营养价值的试验中，选择试验用的动物，要注意其一致性（consistency），不能只注意动物年龄大小一致、体重相似，而忽视动物的健康状况。不能只注意在一定试验时间内动物体重增加的情况，而忽视上述各因素的影响。统计理论要求进行试验设计必须考虑影响试验结果出现差异的各种因素，这样才能对试验因素的作用作出比较合理的结论。

以统计理论为指导，进行实验设计时，要考虑到在较少人力、物力和时间的条件下，得出可靠的信息，取得准确的数据。在简单易行，节省人力、物力和时间的情况下，还要取得好的结果。例如有人研究两种抗组织胺在人体内是否有相加作用，如果按一般设计方法，实验要作640次反应测定，需耗费8500个工时，而采用拉丁方设计，只作 $9 \times 9 = 81$ 次测定、花费1100个工时就够了，效率提高了近8倍，并且节省了财力和物力。

科学试验的目的是要了解一个或几个实验因子的作用，因而对试验中遇到的其它因子则要进行严格地控制。实验者的任务就是要严格控制其它因子的作用，有意识的了解实验因子预期应达到的作用。根据统计理论，对试验设计的基本要求是对实验中无法消除的误差能够提出足够的统计量，把所有的比较和计算量都建立在可靠的基础上，并在同一试验中取得更多的资料。要达到这一点，统计理论对试验设计提出的基本要求是：必须遵循进行重复、设立对照和随机性三个原则，这一点将在第十二章内详细叙述。

当然，生物统计对试验设计有着积极的指导意义，但它绝不能去代替科学试验。如果试验目的、要求不明确、试验设计不合理，试验条件不合适，统计数据不准确，这种试验也绝不会成功，统计理论和方法都不能挽救试验的这种失败。因此，正确地使用生物统计理论是非常重要的。

四、正确地使用统计分析方法

生物统计分析方法在生物科研试验及其它有关试验的各个方面都得到应用，并且越来越广泛。为试验结果分析提供数学依据、以提高分析的可靠程度。但是，统计分析必竟不是科学试验，只是一种为科学试验服务的辅助工具，绝不能利用统计分析方法来改变试验研究的结果。因此，使用统计分析方法时，我们要始终明确这一点，既要反对那种认为统计无用的说法，也要反对认为统计分析什么都能解决的思想。

应用统计分析方法分析、推断和解释试验结果时，我们必须以唯物辨证法为指导思想，以生物科学知识为基础，要全面地认识问题，既要看到生物统计分析的重要意义，也要在对生物界定性认识的基础上进行统计分析。如研究二个性状之间的关系时，既要明确统计分析的科学性，也要用生物学知识了解二性状间是否确实存在着直接地内在联系，而不能被两组变数数据的巧合所迷惑；还应注意统计分析所起作用的范围，不能到处乱用还要看到影响生物科学试验的内、外因素是非常复杂的，是不断变化的。性状和特性的表现是受生物学规律所支配，而不会被统计规律所制约，因此，要得出比较合理的结论、真正地认识事物的本质，首先必须以生物科学知识为基础、结合试验具体情况，然后正确应用统计方法进行分析、推论。

应该着重指出，科学试验数据是统计分析的依据，它要求试验数据必须是准确的。绝不能不加分析、随便收集一些杂乱不全的不准确资料进行统计分析。在科学试验中，没有全面进行试验设计、不精心操作、粗心大意，企图藉用统计分析方法想找出一个本来不存在的规律、那只能适得其反。统计分析只能阐明、揭示规律，而不能“创造”规律。

五、统计学常用术语

1 **总体与样本**（population and sample）

在自然界中许多事物和现象客观上都构成一个总体，所谓总体就是同质的个体所组成的集合、即事物的全体。如人的身高、体重；小麦籽粒含蛋白质百分数，同一批号药物的效价值，患某病的患者等，都各自构成一个总体。

总体包含的个体可以为无穷大，且往往只是设想的或是抽象的，即使个数有限，在实际工作中也是无法知道其总体的全貌，如某小麦品种的株高，50万瓶青霉素效价值等，了解它们，我们只能对一部分个体进行观察、测量，这种从总体中抽取部分个体的过程叫做抽样（*sampling*）。

从总体的性质来看，可将总体分为有限总体和无限总体。有限总体（finite population）指总体内元素是可数的，无限总体（infinite population）指总体包含的个体是不可数的。

从研究工作实际来看，总体是相对的，有大有小。如人的身高，可指全世界所有人的身高，也可以把某地区包含成员身高当作一个总体。

样本是指从总体中抽取的一部分个体，是总体的一部分，它是总体的代表。在一个样本内可以包含有不同的个体数、样本内包含的个体数目称为样本含量（*size* of sample）。根

据样本含量多少，可将样本分为大样本和小样本两种(large sample and small sample)，大样本指含量在30个个体以上，30个个体以下的称小样本。

在生物科学研究中，人们期望知道的是总体的参数（parameter of population），而不是样本的统计量（*sample* statistics），而我们在实际工作中，在具体试验中，所能得到的只是样本统计量、而不是总体参数。生物统计学就是解决这个矛盾，用样本正确地推断、估计总体。

2 参数与统计量（parameter and statistics）

参数也称参量，是指由总体中计算的指标数值。如总体平均数，总体标准差，总体的成数等均是参数。统计量是指由样本计算的标志数值，即由观察值所定出的量，如样本平均数、样本标准差，通常把统计量作为参数的估计值。

3 变数与常数（veriate aud constants）

相同性质的事物间表现差异性或变异特征的数据，即在一个界限内变动着的表示性状的数值称作变数或变量。自然界同质事物间，都存在着不同程度的变异，事物之间其所以可以区分，就是由于它们之间存在着差异性。绝对相同的两个事物是没有的。如人的身高、体重各不相同，小麦的穗长、株高、各有差异，同种细菌的的菌落各有大小，同窝动物个体之间表现出的生理、生化指标各不一样。所有这些差异均可用量来表示，通常记作x，如10株小麦植株，株高在84到91厘米之间，共有84，85、86、87、88、89、90、91八个变数值，记作x_i（i＝1，2，…，8），表示x_1到x_n之间任一个数值，亦称x_i为随机变数（random *variable*）。

变数按其性质可分为连续变数和非连续变数。

连续变数（continuous *variable*）表示在变数行列中可抽出某一范围内的所有值。变数之间是连续的、无限的。如正常男子的体重在55至65公斤之间，在这个数值之间可以取得无穷个变数。表示长度、重量的数值均属于这类变数。

非连续变数（discontinuous variable），也称离散变数。在变量行列中，仅能取得固定数值，通常是整数。如小麦的小穗数、每穗籽粒数，母猪每窝产仔猪数等。

变数可以是定性的，也可以是定量的。定性的、如当某个体属于几种互不相容的类别中的一种时，它的特性是“有”或“无”，是这一种或那一种，二者或多者居其一。如小麦的芒有长芒、短芒、无芒三种，香豌豆的花色有白色、红色和紫色三种，疗效有治愈、好转、死亡等等。而定量的则是可测可量的，如人的白血球计数，视野内细菌数目，小麦的千粒重，仔猪断奶重等。

常数表示能代表事物特征和性质的数值，通常是由变数计算出来的，在一定过程中是不变的。如总体的平均数，标准差，变异系数等。只有在事物总体发生变动时，它才随之变动。

常数就来源而言，可将其分为：

常规数：指从总体中测定所得的数值，是该总体的真正数值，反映了总体的性质和特性。生物科学中无法测定常规数。

估计常规数：由样本计算出的用以估计或代表总体常规数的数值。由于样本间有差异，估计常规数常常是变化的，它只能与常规数接近、在常规数两边摆动、很难完全相同，统计学中所用的常数一般都是指估计常规数。

4 效应与连应 (effect and interaction)

引起试验对象出现真正差异的作用称为效应。如肥料使作物产量提高，不同饲料引起动物体重表现出真正差异等。根据引起效应的作用的事物不同，可将效应分为几种情况：如用肥料、药物或栽培方法使某小麦显然生产力提高、增产效果显著、称为处理效应；不同作物品种进行产量比较试验，由品种不同引起产量差异显著、称作品种效应。

连应是指两个或两个以上因素间的相互作用所产生的效应，也称交互作用。如肥料试验中、磷肥与氮肥的作用不同，氮肥的效应大小，常常随磷肥的的存在与否而变化；磷肥效应的大小，也随着氮肥的存在与否而不同，共同施用，其效应超过了各自的效应之和。如单独施氮肥，单位面积内可增产20斤，单独施磷肥，可增产10斤，如果共同施用，可增产35斤，其中多余的5斤是由氮肥和磷肥共同作用所产生的。连应有产生正作用的，也有产生负作用的，如肥料试验有增产的，也有减产的，氮、磷肥共施可以增产、氮、钾肥共施有时使产量降低，这就是负连应作用。

5 零值假设 (null hypothesis)

零值假设也称无效假设或解消假设。在对两个样本统计量进行差异显著性测定时，统计学都要提出一个假设，总是把两个样本间表现出的差异认为是来自同一个总体，样本间没有根本性差别。假设差异是由抽样误差所致、然后进行统计分析、依据实验所得数据，按照一定概率标准来推断这种假设是否成立，从而判断两个样本统计量差异的真假，得出统计量显著性测定的结论。

根据研究问题的对象不同，假设可以用不同形式提出。如测定某种药物的效价值，可先假设药剂无效，如果比较两种药物的效应大小，则假设两种药物效果相同，然后按实际数据计算这种假设出现的概率。依概率的大小来判断假设的真假，确定二事物间差异的真假。

应该指出统计学上的证据不是证明，即使样本很大，P值很小，假设不能成立，也不能绝对否定假设，必竟还会有那一部分很小的可能性在某种情况下是会出现的。

6 概率 (probability)

概率是可能性的定量计量，即把反映某一事件发生的可能性大小用数值表示出来。如在一对等位基因的杂交试验中，隐性性状在杂种第二代出现的可能性大小，我们可以用数值表示出来。在人群中，色盲在男性和女性中的比率是不同的，男、女性出现色盲的可能性可以用数值表示出来。这个数值就是指某一事件出现的次数比较稳定的占全事件出现的总次数的百分率。某些事件在一定条件下必然出现，称必然事件，它的概率是1。某些事件是不可能事件，它的数字概率为0。另一些事件，在一定条件下可能出现、也可能不出现，其出现的概率在0到1之间。

概率在统计学上用P表示，如$P \leqslant 0.05$表示某事件出现的概率小于或等于百分之五，$P \leqslant 0.01$，表示某事件发生的可能性小于或等于百分之一。

7 置信限度 (confidence limit)

置信限度也称可信限，是指在一定概率下，用样本均数估计总体均数可能的范围，即总体均数的可能区间范围。由于样本均数是一个变量，不能确切的确定总体均数。因此，比较合理的是给总体参数一个范围，即在一定概率下，总体均数可能在多么大一个范围内，这个范围也称区间估计。

8 自由度 (degree of freedom)，

是指能够自由活动的变量的个数，也指可以自由活动的范围。从几何学上的点来看自由度，在平面上的一点，通常用（x、y）表示，如果没有条件限制，这个点可以沿着纵横二轴任意移动，其自由活动的范围有两个。如果有一个条件限制它，如$x+y=5$时，x和y就只有一个能任意选择，即当$x=2$时，y只能等于3，且（x,y）点只能在$x+y=5$这条直线上活动。故自由度只有一个。在空间的一个点是由三个数字表示的，即（x，y，z），如不加以限制，它有三个自由度，可以向三个方向变动，如果使$x+y+z=10$，这个空间点就只有两个数可以自由活动，自由度就只有两个。因此，每加上一个条件，自由度就减少一个。由此推广，数学上把x_1、x_2、x_3……xn作为几度空间里的点，如无条件限制，这个点就有几个自由度，如加上下面一个条件，即

$$x_1+x_2+x_3+\cdots\cdots+x\text{n}=\Sigma x;$$

有了这个条件的限制，自由度就减少一个，成为n－1。如计算样本标准差时[2]因受样本估计常规数均数的限制，其自由度就为变数的个数减去1，即n－1。如果同时存在两个或三个条件的限制，则自由度n相应的减少2或3、即为n－2，n－3。如计算相关，回归有关统计量时，自由度受到两个估计常规数的限制，其自由度为n－2。

生物统计学其所以应用自由度，是为了在小样本情况下，不降低其精确程度，因为通常计算采用的小样本只是总体中极小的一部分，其全距远比总体的全距为小。这样用样本估计总体，对总变量平均时，不用样本包含个数n而用自由度n－1去除。

9 **错误与机误**（error）

错误指工作人员在试验中疏忽大意、粗枝大叶或有意识的加大减少数据等造成的差错，使试验结果不真实，如在试验中称量药物时未校正天平，培养基中药物比例数不当，少放或忘记加入某种药物，这是人为的差错；可以避免。

机误是指试验中出现的差异是由机会造成的。如在随机抽样中，出现较大或较小的数据较多，这是由于总体中个体间存在着差异，由于机会作用，出现差异，它是不可避免的，只能设法减小、而不能完全消灭，但增加次数或加大含量可以减小，如抛镍弊计算正面向上的次数：

抛5次，出现的比率有0.0，0.2，0.4，0.6，0.8，1.0，差距为0.0—1.0之间。相差为±0.5；

抛50次，其差距在0.36—0.64之间，相差为±0.14；

抛500次，差距为0.488—0.512之间，相差为±0.012；

抛5040次，差距在0.4931—0.5069之间，相差为±0.0069；

抛24000次，差距在0.4995—0.5005之间，相差为±0.0005

可见，次数越多，误差越小。

10 **百分比与百分率**（percentage）

百分比与百分率为两种不同统计指标。百分比是结构指标，表示事件占的比例，即部分对全体之比，如技术人员占全厂人员的比重；百分率则是频率指标，表示事件出现的频率、如某地区某病发病率。

第二章　资料的收集与整理

在生物科学试验及调查中，能够取得大量的原始数据，这是在一定的具体条件下，对某种事物或现象观察的结果，我们称之为资料（data）。统计分析就是要依靠这些准确而完整的资料所提供的信息，去了解事物或现象内部潜在的问题，认识事物内在的规律。科学试验或调查所取得的原始资料一般是分散的，零星的和孤立的，所提供的情报并非一目了然，要从这些分散的资料中揭露出事物的内在联系，了解其本质，就必须先对资料进行整理，对原始数据进行去粗取精，去伪存真。用统计特有的方法把零乱的数据理顺，归纳、使其系统化，并且列成统计表，绘出统计图。

一、数量资料表示的方式

对资料进行分类是统计归纳的基础，不进行分类，大量的原始数据就不能系统化，规律化。对资料进行分类整理时，首先，必须坚持“同质”（homogeneity）的原则。只有“同质”的数据，才能使资料反映出事物的本质和规律。生物的性状特性大致可分为两类情况：一类是质量性状；另一类是数量性状。因而，我们取得的数据可能是定性的，也可能是定量的，这些资料可以分为质量性状资料和数量性状资料。

1 质量性状资料

所谓质量性状（qualitative character），一般表明一种现象的性质，用计量工具无法度量。统计学上只能记录具有某一属性的个体数，如香碗豆花的颜色，红花、白花、紫花各占多少株；果蝇长翅与残翅的个体数；长芒、短芒、无芒的小麦植株数；人的血型，A、B、AB及O型各多少个；雌、雄个体数目等。还有一些属性的表示方法是用数字级别表示某现象在程度上的差别，即用评分法表示类别、程度，如小麦感染诱病的严重程度；作物倒伏的严重程度；家畜精液品质；绵羊油汗色泽，都是用评分法表示。对于上述质量性状取得的资料，一般均采用记数方式来表示。每个变数均以整数出现，两变数之间是不连续的，如红花香碗豆与白花香豌豆杂交，统计杂种第二代不同花色的植株时，1000株植株中，红花266株、紫花494株、白花240株。某种公绵羊85头后代中油汗色泽评分统计，1级5头，2级8头、3级16头，4级21头、5级35头等。

有些质量性状在遗传上是受多基因控制的，如绵羊油汗色泽，小麦籽粒颜色，均受多基因控制，但前者可以区分由白到黄五个级别，而后者一般只按红、白二色计数，实际上在红粒小麦籽粒中颜色的深浅程度仍有差别，表现出连续性的变异。因而也可以把它们看作数量性状，由于测量困难，常采用计数方法来表示。

在观察总体中各个单位时，我们有时只关心每个个体“具有某种特征”，还是“不具有某种特征”，如质量检查中计“合格”和“不合格”，血型调查只统计“O”型人数等，总

体单位的有些特性，只有两种对立表现，或者是根据研究的目的，人为的把各种表现归并为对立的两种表现——“是A”和“非A”，把一个是非型标志改变为统计变量，即成数计算中的“0—1”变量，也可当作质量性状看待。

2 数量性状资料

生物有机体某些性状和特性的差异是用数量表示，如人的身高、体重、小麦的千粒重、分蘖数、每穗小穗数，土壤成分中的含氮量，单位面积的害虫数、药物处理后植株成活数，害虫死亡数等。统计这类现象的资料，根据变数的性质又可将其分为计数资料（不连续变异）和计量资料（连续变异）。

A 计数资料（enumeration data）：计数资料是指用计数方式取得的资料、表示事物的个数，变数值是整数、整数间的数值是不连续的，如玉米果穗上的籽粒行数，小麦每穗粒数，猪每窝产仔数，均可用统计个数来表示，这类资料称为计数资料。变数值以整数出现不可能带有小数，不能划分成最小单位的连续性变数。如小鼠每窝产仔数为1—10个，则其变数为1、2、3……10，绝不会出现3.5个，2.8个等。

B 计量资料（measurent data）：计量资料是通过直接计量而得来的，变数值是通过称、量得到的，数据是用长度、重量、容积等单位表示。如玉米籽粒百粒重，小麦株高、单株粒重，仔猪出生重、断乳重等。这类现象所测得的变数不一定是整数，而是在两个相接的整数间可以出现任何数值，表现出连续性变数。如玉米植株高度1.89米，2.04米，1.99米等，在1.99至2.00米之间还会出现1.991，1.9991……等变数值，可以出现无限个变数值。至于小数后位数的多少，那是根据实验要求和度量工具的精确度而定。

应该指出，生物的某些性状是可以用计数，计量两种方式表示的。比如，同为植株的高度，一般用计量方式表示，取得的资料是计量资料。如果植株高度相差悬殊，中间连续不起来，如孟德尔试验中所使用的高茎、矮茎豌豆植株、在株高上明显地可分为高、矮两种，植株高度属于质量性状、统计高、矮植株各有多少株，这样得来的资料就是计数资料。

二、数量资料的收集

收集资料（collective data）是统计分析的第一步，也是全部统计工作的基础。生物科学数量资料的来源主要是通过试验，实验和调查。统计学对原始资料的要求是数据要完整、准确，就是说，要注意数据本身有无差错，如记录不全、测量不准，取样方法正确与否、非同质数据有否进行合并等，必须经过反复检查和核对。

统计分析解决问题的主要的方法是用样本正确地估计总体，抽样的目的主要是对总体某种特征的推断，要使样本无偏差的估计总体，除了样本包含的个体数量要足够大以外，重要的是要采用科学取样方法，抽取具有代表性的样本，采用计数或计量的方法取得一定的可靠试验数据。

在科学研究试验中，由于研究的目的与性质不同，可以采取各种各样行之有效的取样方法。从理论上讲，由于是以概率论和数理统计的原理为依据，用样本推断总体，因而，这种样本必须是随机样本，就是用随机取样方法得到的样本。对于随机取样，从数学上已经严格地论证了：随机抽样结果，靠近总体要推断的指标的概率比远离总体要推断的指标的概率要大得多。只有这种样本才能正确地估计出抽样误差，才能用来合理的估计总体。

随机抽样（random sampling）必须满足以下两个条件：

①总体中每个个体，都具有同等被抽选的机会；

②总体中任一个体被选中的可能性，不受其它个体被选中的影响，各个个体被选中是相互独立的。后一条对无限总体是合适的，如果总体个体数有限，进行无退还抽样，则每选出一个个体后，其余个体被选中的可能性就不同于取样前，因此，应该把这一条理解为“总体中任一个体被选中的可能性受任何其它个体被选中的影响是相等的。”

完全满足随机取样在理论上的要求是很困难的，统计学家为了使取样尽可能地满足这些要求，根据研究对象和研究目的的不同，制订出许多随机取样的方法。

1 **单纯随机取样**（simple random samplin）：

取样时，对总体不作任何分类排队的处理，混匀总体全部个体，使各个体在抽样中有同等的相互独立的被选机会，一个一个地随机抽取。抽取方式可利用随机数字表或抽签方式。随机数字表是一种由许多随机数字排列起来的表格（附表13），是经过随机性检验而确定的，它的使用方式很多，通常使用时，首先将抽样对象按顺序编号，然后在随机数字表中任取一个数字作为起点，可向任何方向摘录数字，根据抽样个体数多少和抽样对象的多少、选择数字位数、选好后把大于抽样对象数目的数字均减去抽样对象数如果还大，再减去一倍，使所抽数字均小于抽样对象数，这些数字就是我们随机抽样的对象。例如在800人中抽取50人作调查，先将800人编号（实际工作中并不一定要逐人编号），然后在随机数字表中任取一数，假如是从第五行第五列的64开始，向右摘取三位数字即为：643，854，824，622,316,243，099，006，184，432，……，将大于800的数字分别减去800，则分别得：643，054，026,622，316，243，099，006，184，432，……，被编为这些号码的人，就组成我们所需的单纯随机抽样的样本。

正确运用随机数字表，可保证随机抽样的随机性。简单随机取样法适用于个体间差异较小，个体分布比较集中的研究对象，一般抽取的个体数也较少。对个体数量大的总体，此法费时费力，有局限性、不宜使用。

2 **分区随机抽样**（stratified random sampling）：

抽样时，把总体分成若干部分，然后从每个部分随机抽取若干个体组成样本。这种方法的优点是中选的个体在总体中的分布比单纯随机抽样更均匀，样本代表性较好。

3 **系统抽样法**（systematic sampling）

系统抽样也称机械抽样。是把总体各单位按一定标志排队，然后按相等距离抽取个体，或机械地每隔若干单位抽取一个个体。如排队编号为3000个单位的总体，抽120个个体为样本，可每隔25个抽取一个，抽取时可从25号以前任何一个编号开始。如从13号开始，则第二个被抽对象是38号、第三个是63号……。这种抽样方法的优点是方法简便，抽样误差比随机抽样为小。不足之处是由于是机械地每隔一定距离抽取一个，各个号码不是相互独立，故在某些情况下，这种机械性抽样对总体的代表性是不如随机抽样，可能出现某些偏倚。

4 **整群抽样**（cluster sampling）：

以整群或集团为单位进行抽样，不是一个个体一个个体的抽取，如调查小学生某种情况时，在每个学校内可抽取一到几个班，对抽取的每个班的成员进行调查。这种方法适于大规模的调查，易于组织且节省人力、物力。但群体间差异要小，抽样时抽取群体的数目要多，否则抽样误差较大。