第一章

概　论

本章提要

生物统计学是把数学的语言引入具体的生命科学领域，运用数理统计的原理和方法对生物有机体开展调查和试验，目的是以样本的统计数估计总体的参数，对所研究的总体进行合理的推论。生物统计学主要包括试验设计和统计分析两部分内容，其作用主要有4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。统计学的发展经历了古典记录统计学、近代描述统计学和现代推断统计学3个阶段。本章还介绍了统计学中几组常用的术语。

第一节　生物统计学的概念

统计学(statistics)是把数学的语言引入具体的科学研究领域，将所研究的问题抽象为数学问题的过程，是搜集、分析和解释数据的一门科学。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一门学科，属于应用统计学的一个分支。随着生物学研究的不断发展，生物统计学的应用也越来越广泛。

生物学研究的对象是生物有机体，与非生物相比，它具有特殊的变异性、随机性和复杂性。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学的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。

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而形成的，它是应用数学的一个分支，属于生物数学的范畴。生物统计学是把数学的方法引入具体的生命科学

领域，把生命科学领域中具体的研究问题抽象为数学问题，从大量试验数据中探寻其规律的过程，以数学的概率论和数理统计为基础，涉及数列、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　统计学发展概况

人类的统计实践是随着记数活动而产生的。因此，对统计发展的历史可追溯到远古的原始社会。但是，使人类的统计实践上升到理论予以总结和概括成一门系统的统计学，起源于17世纪英国，其代表人物W. Petty (1623～1687)是政治算术学派的奠基人，代表作是《政治算术》。政治算术学派主张用大量观察和数量分析等方法对社会经济现象进行研究，为统计学的发展开辟了广阔的前景。由于W. Petty对统计学的形成有着巨大的贡献，马克思称他为“统计学的创始人”。统计学的发展经历了古典记录统计学、近代描述统计学和现代推断统计学3个阶段。

一、古典记录统计学

古典记录统计学(record satistics)形成于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。概率论被引进之后，逐渐成为一种较为成熟的方法。

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、统计学家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、物理学的研究中进行了推广应用。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学的理论十分重要，它最早是由法国数学家De Moiver于1733年发现的。德国天文学家和数学家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，并提出了“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。

二、近代描述统计学

近代描述统计学(description statistics)形成于19世纪中叶至20世纪上半叶，这个时期也是统计学应用于生物学研究的开始和发展时期，其“描述”特色是由一批原来研究生物进化的学者们提炼而成的。英国遗传学家F.Galton(1822～1911) 自1882年起开设“人体测量实验室”，分析父母与子女的变异，探寻其遗传规律，应用统计方法研究人种特性和遗传，探索了能把大量数据加以描述与比较的方法和途径，引入了中位数、百分位数、四分位数，以及分布、相关、回归等重要的统计学概念与方法，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。

F.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《生物计量学报》)这一权威杂志。在该杂志的创刊词中,F.Galton和 K.Pearson 首次为他们所运用的统计方法明确提出了"生物统计"(biometry)一词,F.Galton 解释为:所谓生物统计学,就是应用于生物学科中的统计方法。在《自然界的遗传》一文中,K. Pearson 提出了相关与回归分析问题,并给出了简单相关系数和复相关系数的计算公式。1900 年,K. Pearson 在研究样本误差效应时,提出了 χ^2 检验,它在属性资料的统计分析中有着广泛的应用。

三、现代推断统计学

现代推断统计学(inference statistics)形成于 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繁杂的数量关系以及一系列未知的数量变化,单靠记录或描述的统计方法已难以奏效。因此,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* 杂志上发表了论文《平均数的概率误差》,创立了小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 年看成是统计推断理论发展史上的里程碑,也有人推崇 W.S.Gosset 为推断统计学(尤其是小样本研究理论)的先驱者。

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的方法。1915 年,R.A.Fisher 在 *Biometrika* 上发表论文《无限总体样本相关系数值的频率分布》,被称为现代推断统计学的第一篇论文。1925 年,R.A.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、生物学及遗传学的研究与发展。自方差分析问世以来,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。

J. Newman(1894～1981)和 E. 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、K. Mather 对群体遗传学、F. Yates 对田间试验设计等都作出了杰出的贡献。

我国对生物统计学的应用始于 1913 年顾澄教授翻译的英国统计学家 G.U.Yule 在 1911 年出版的关于描述统计学的名著《统计学之理论》,这标志着英国、美国数理统计学传入中国的开始。之后,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及亲代和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，联合国教育、科学及文化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网络技术的发展，SAS(statistical analysis system)、SPSS(statistical package for the social science)等国际通用统计软件的开发和应用，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。

第三节　常用统计学术语

一、总体与样本

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)。

总体按所含个体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。例如，通过临床试验来推断某种药品比另一种药品治愈率高，这里无限总体是指一个理论性总体。个体有限的总体称为有限总体(finite population)。例如，对某一班学生身高进行调查，这时总体是指这一班中每位学生的身高。

要研究总体的性质，一般情况下我们无法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，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。在这种情况下，只能采取抽样的方法，从总体中抽取一部分个体进行研究。

从总体中抽出的若干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)，样本中个体的数目称为样本容量(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，$n < 30$ 的样本称为小样本，$n \geqslant 30$ 的样本称为大样本。在一些计算和分析检验方法上，大、小样本是不同的。

在对事物的研究过程中，人们常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。

二、参数与统计数

参数(parameter)也称为参量，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。统计数(statistic)也称为统计量，是由样本计算所得的数值，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。由于总体一般都很大，有的甚至不可能取得，所以总体参数通常是未知的。正因为如此，我们才进行抽样，由于样本是已经抽出来的，所以统计数是可以计算出来的，我们可以根据样本统计数来估计总体的参数。此外，还有一些统计量是为了进行统计分析而构造出来的，如后续章节中的 u 统计量、t 统计量及 F 统计量等。

三、变量与资料

相同性质的事物间表现差异性的某项特征或性状称为变量或变数(variable)，是研究者在确定了研究目的之后，所观测的试验指标。由于试验目的不同，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、血型等。变量通常记为 x，如10个人的身高为155～180cm，共有158，167，173，155，180，165，175，178，170，162(cm)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。变量的观察结果可以是定量的，也可以是定性的，其结果称为变量值(value of variable)或观测值(observed value)，也称为数据、资料(data)。

根据获取观测值的方式及测量方法所提供的数值信息的差异，变量可以分为定量变量和定性变量。通过测量所获得的、用具体数值与特定计量单位表达的数据称为定量变量(quantitative variable)，也称为数值变量(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，如人的身高(cm)、体重(kg)、脉搏计数(次/min)等。定量变量根据取值的不同，可以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，变量之间是连续的、无限的。例如，小麦的株高为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，并且通常是整数，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。

定性变量(qualitative variable)也称为分类变量(categorical variable)、名义变量(nominative 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。例如，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。例如，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)，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。

对应于变量，常量(constant)是不能给予不同数值的变量，它是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，如总体平均数、标准差、变异系数等。只有在事物的总体发生变动时，常量才随之变化。

四、因素与水平

试验中所研究的影响试验指标的原因或原因组合称为试验因素(experimental factor)或处理因素(treatment factor)，简称为因素或因子(factor)。试验因素常用大写字

母,如 A、B、C 等来表示。

每个试验因素的不同状态(处理的某种特定状态或数量上的差别)称为因素水平(level of factor),简称为水平(level)。例如,研究温度对某种酶活性的影响,所设置的15℃、20℃、25℃、30℃ 分别称为温度因素的一个水平。可见,因素是一个抽象的概念,而水平则是一个较为具体的概念。水平常用代表该因素的字母添加下标(如1、2、3等)来表示,如 A_1、A_2、B_1、B_2 等。

按照性质不同,因素可以分为可控因素和非控因素。在试验中可以人为调控的因素称为可控因素(controllable factor)或固定因素(fixed factor)。该因素的水平可准确控制,且水平固定后,其效应也固定,同时在试验进行重复时可以得到相同的结果。例如,研究3种温度对胰蛋白酶水解产物的影响,因为温度是可以严格控制的,所以在重复该试验时对于相同的温度其水解产物的量也是固定的。温度在此例中即为固定因素。

在试验中不能人为调控的因素称为非控因素(uncontrollable factor)或随机因素(random factor)。该因素的水平不能严格控制,或虽水平能控制,但其效应仍为随机变量,同时在试验进行重复时不易得到相同的结果。例如,研究农家肥不同施用量对作物产量的影响,由于农家肥有效成分较为复杂,不能像控制温度那样,将农家肥有效成分严格地控制在某一固定值上,在重复试验时即使施用相同数量的农家肥,也得不到一个固定的效应值。农家肥在此例中即为随机因素。

五、处理与重复

试验处理(experimental treatment)通常也称为处理(treatment),是指对受试对象给予的某种外部干预(或措施)。其中受试对象(tested subject)又称为试验单位或试验单元(experimental unit),是指在试验中能接受不同试验处理的独立的试验载体。植物个体、动物个体,以及不同的组织、器官等都可以作为试验单位。

处理根据所涉及的因素数可以分为单因素处理和多因素处理。当试验中涉及的因素只有一个时,称为单因素处理(single factor treatment)。在单因素处理中,实施在试验单位上的具体项目就是试验因素的某一水平。例如,饲料的比较试验,实施在试验单位(如某种畜禽)上的具体项目就是饲喂某一种饲料。进行单因素试验时,试验因素的一个水平就是一个处理。

如果试验中涉及两个或两个以上的因素,则称为多因素处理(multiple factors treatment)。可依处理因素数进行具体命名,如二因素试验处理、三因素试验处理等。在多因素试验处理中,实施在试验单位上的具体项目是各因素的某一水平组合。例如,3个播种密度对4个小麦品种的产量影响试验,就是一个二因素试验处理,试验共有 $3\times4=12$ 个水平组合,实施在试验单位(小麦)上的具体项目就是某个种植密度与某个小麦品种的组合。进行多因素试验时,试验因素的一个水平组合就是一个处理。相对于单因素试验,多因素试验不但可以研究因素的主效,同时也可研究因素之间的交互作用。

重复(repetition)是指在试验中,将一个处理实施在两个或两个以上的试验单位上。处理实施的试验单位数即为处理的重复数。例如,研究某种饲料对猪的增重效果,将该种饲料饲喂5头猪,则表明这个处理(饲料)有5次重复。

六、效应与互作

试验因素相对独立的作用称为该因素的主效应(main effect),简称为主效或效应(effection)。例如,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。两个或两个以上处理因素间相互作用所产生的效应,称为互作效应(interaction effect),简称为互作或连应(interaction),如氮、磷肥共施会对作物产量产生互作效应。互作效应有正效应,也有负效应。如果氮、磷肥共施的产量效应大于氮、磷肥单施效应之和,说明氮磷肥互作效应为正效应;如果氮、磷肥共施的产量效应小于氮、磷肥单施效应之和,说明氮磷肥互作效应为负效应。互作效应为零,则称因素间无交互作用,没有交互作用的因素是相互独立的因素。应该注意的是,有时交互作用相当大,甚至可以忽略主效。因素间是否存在交互作用有专门的统计推断方法,有时也可根据专业知识或经验加以判断。

七、准确性与精确性

准确性也称为准确度(accuracy),是指在调查或试验中某一试验指标或性状的观测值与真值接近的程度。精确性也称为精确度(precision),是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。统计学是以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,用样本中各变量间变异程度的大小来衡量该样本统计数精确性的高低。因此,准确性不等于精确性。准确性反映测定值与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允许范围内。

八、误差与错误

误差(error)也称为试验误差(experimental error),是指观测值偏离真值的差异。试验误差可以分为随机误差和系统误差两类。随机误差(random error)也称为抽样误差(sampling error)、偶然误差(accidental error),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值之间的差异,是不可避免的。统计上的试验误差通常就是指随机误差。我们可以通过增加抽样或试验次数降低随机误差,但不能完全消除随机误差。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,如仪器调校的差异、各批次药品间的差异、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以控制的,只要试验工作做得精细,在试验过程中是可以避免的。

错误(mistake)又称为过失性误差(gross error),是指在试验过程中,人为因素所引起的差错。例如,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、数据抄错、计算出现错误等。在科学研究过程中,这类错误是不允许产生的。

第四节　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。试验设计是指应用统计的原理与方法制订试验方案、选择试验材料并进行合理分组，使我们可以用较少的人力、物力和时间获得较多而可靠的数据资料。统计分析是指应用数理统计的原理与方法对数据资料进行分析与推断，主要包括统计描述和统计推断，涉及数据资料的搜集和整理、特征数的计算、假设检验、方差分析、回归和相关分析、协方差分析等。从二者的关系来看，统计分析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。合理地进行调查或试验设计，科学地整理、分析所得到的资料是生物统计学的根本任务。

生物统计学的基本作用可以概括为以下 4 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，这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，就必须运用统计分析的方法。

(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资料和较好的试验结果，在一些生物学研究中，就需要以统计原理为依据，科学地进行试验设计。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。

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　解释以下概念：

总体、个体、样本、样本容量、变量、参数、统计数、因素、水平、处理、重复、效应、互作、试验误差。

习题 1.3　随机误差与系统误差有何区别？

习题 1.4　准确性与精确性有何区别？