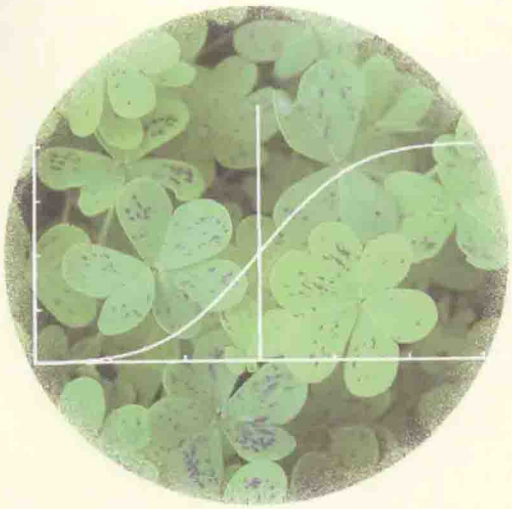
全国普通高等院校
生命科学类“十二五”规划教材

十二五

生物统计学

彭明春　马纪　主编

Biostatistics



华中科技大学出版社
http://www.hustp.com

全国普通高等院校生命科学类“十二五”规划

生物统计学

主　编　彭明春　马　纪

副主编　陈其新　王有武　施文正　聂呈荣
　　王　玉　万海清　陈　国

编　委　（按姓氏拼音排序）
　　陈　国　陈其新　耿丽晶　侯沁文
　　胡　颖　李转见　刘小宁　马　纪
　　聂呈荣　彭明春　施文正　石培春
　　万海清　王　玉　王文龙　王晓俊
　　王有武

华中科技大学出版社
中国·武汉

内 容 简 介

本书分为基础理论和软件应用两个部分，较系统地介绍了生物统计学的基本原理和方法，以及统计分析工作的软件实现。基础理论部分，在简要介绍生物统计学的内容、发展的基础上，介绍了数据的描述性分析、概率与概率分布，重点介绍平均数与方差的统计推断、方差分析、一元回归与相关分析、协方差分析、非参数检验，抽样的原理与方法、常用试验设计与统计分析，多元数据的线性回归与复相关分析、聚类分析、主成分分析和典型相关分析，涵盖了各类数据的常用统计分析方法。软件应用部分，重点介绍用 SPSS 软件进行统计分析工作的基本操作、过程和参数的选择与设置，并简要介绍了 Excel 的生物统计应用。

本书充分考虑了生物科学、生物技术、生态环境、农业科学、林业科学、医学卫生、食品科学等专业的统计分析需要，可作为综合性大学、师范院校、农业、林业、医学等院校相关专业本科生和研究生的教材，也可作为相关专业科研工作者、教师的参考书。

图书在版编目(CIP)数据

生物统计学/彭明春，马纪主编. —武汉：华中科技大学出版社，2014.10
ISBN 978-7-5609-9716-2

Ⅰ.①生… Ⅱ.①彭… ②马… Ⅲ.①生物统计-高等学校-教材 Ⅳ.①Q-332

中国版本图书馆 CIP 数据核字(2014)第 250940 号

生物统计学 彭明春 马 纪 主编

策划编辑：王新华
责任编辑：王新华
封面设计：刘 卉
责任校对：张会军
责任监印：周治超
出版发行：华中科技大学出版社(中国·武汉)
武昌喻家山 邮编：430074 电话：(027)81321915
录 排：华中科技大学惠友文印中心
印 刷：武汉鑫昶文化有限公司
开 本：787mm×1092mm 1/16
印 张：18.75
字 数：489 千字
版 次：2015 年 3 月第 1 版第 1 次印刷
定 价：42.00 元

全国普通高等院校生命科学类“十二五”规划教材
编 委 会

主任委员

余龙江　华中科技大学教授，生命科学与技术学院副院长，2006—2012 教育部高等学校生物科学与工程教学指导委员会生物工程与生物技术专业教学指导分委员会委员，2013—2017 教育部高等学校生物技术、生物工程类专业教学指导委员会委员

副主任委员（排名不分先后）

胡永红　南京工业大学教授，南京工业大学研究生院副院长

李　钰　哈尔滨工业大学教授，生命科学与技术学院院长

任国栋　河北大学教授，2006—2012 教育部高等学校生物科学与工程教学指导委员会生物学基础课程教学指导分委员会委员，河北大学学术委员会副主任

王宜磊　菏泽学院教授，2013—2017 教育部高等学校大学生物学课程教学指导委员会委员

杨艳燕　湖北大学教授，2006—2012 教育部高等学校生物科学与工程教学指导委员会生物科学专业教学指导分委员会委员

曾小龙　广东第二师范学院教授，副校长，学校教学指导委员会主任

张士璀　中国海洋大学教授，2006—2012 教育部高等学校生物科学与工程教学指导委员会生物科学专业教学指导分委员会委员

委员（排名不分先后）

陈爱葵	胡仁火	李学如	刘宗柱	施文正	王元秀	张　峰
程水明	胡位荣	李云玲	陆　胤	石海英	王　云	张　恒
仇雪梅	贾建波	李忠芳	罗　充	舒坤贤	韦鹏霄	张建新
崔韶晖	金松恒	梁士楚	马　宏	宋运贤	卫亚红	张丽霞
段永红	李　峰	刘长海	马金友	孙志宏	吴春红	张　龙
范永山	李朝霞	刘德立	马三梅	涂俊铭	肖厚荣	张美玲
方　俊	李充璧	刘凤珠	马　尧	王端好	徐敬明	张彦文
方尚玲	李　华	刘　虹	马正海	王金亭	薛胜平	郑永良
耿丽晶	李景蕻	刘建福	毛露甜	王伟东	闫春财	周　浓
郭晓农	李　梅	刘　杰	聂呈荣	王秀利	杨广笑	朱宝长
韩曜平	李　宁	刘静雯	彭明春	王永飞	于丽杰	朱长俊
侯典云	李先文	刘仁荣	屈长青	王有武	余晓丽	朱德艳
侯义龙	李晓莉	刘忠虎	邵　晨	王玉江	昝丽霞	宗宪春

全国普通高等院校生命科学类“十二五”规划教材
组编院校

（排名不分先后）

北京理工大学
广西大学
广州大学
哈尔滨工业大学
华东师范大学
重庆邮电大学
滨州学院
河南师范大学
嘉兴学院
武汉轻工大学
长春工业大学
长治学院
常熟理工学院
大连大学
大连工业大学
大连海洋大学
大连民族学院
大庆师范学院
佛山科学技术学院
阜阳师范学院
广东第二师范学院
广东石油化工学院
广西师范大学
贵州师范大学
哈尔滨师范大学
合肥学院
河北大学
河北经贸大学
河北科技大学
河南科技大学
河南科技学院
河南农业大学
菏泽学院
贺州学院
黑龙江八一农垦大学
华中科技大学
华中师范大学
暨南大学
首都师范大学
南京工业大学
湖北大学
湖北第二师范学院
湖北工程学院
湖北工业大学
湖北科技学院
湖北师范学院
湖南农业大学
湖南文理学院
华侨大学
华中科技大学武昌分校
淮北师范大学
淮阴工学院
黄冈师范学院
惠州学院
吉林农业科技学院
集美大学
济南大学
佳木斯大学
江汉大学文理学院
江苏大学
江西科技师范大学
荆楚理工学院
军事经济学院
辽东学院
辽宁医学院
聊城大学
聊城大学东昌学院
牡丹江师范学院
内蒙古民族大学
仲恺农业工程学院
云南大学
西北农林科技大学
中央民族大学
郑州大学
新疆大学
青岛科技大学
青岛农业大学
青岛农业大学海都学院
山西农业大学
陕西科技大学
陕西理工学院
上海海洋大学
塔里木大学
唐山师范学院
天津师范大学
天津医科大学
西北民族大学
西南交通大学
新乡医学院
信阳师范学院
延安大学
盐城工学院
云南农业大学
肇庆学院
浙江农林大学
浙江师范大学
浙江树人大学
浙江中医药大学
郑州轻工业学院
中国海洋大学
中南民族大学
重庆工商大学
重庆三峡学院
重庆文理学院

前　言

生物统计学是应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数据的科学，是生命科学及其相关领域的研究者必备的一门工具。如同计算机对于所有自然科学和社会科学的重要作用一样，生物统计学的作用已经体现在生命科学研究的方方面面，而生物统计学本身的发展不仅促进了生命科学的发展，也为数理统计学的建立和发展以及新课题的提出作出了重要贡献。

在生物统计学已经在各级各类高校普遍开设的情况下，编写一本浅显易懂、将理论性和应用性紧密结合的生物统计学教材，是许多高校的迫切需求。本书由来自全国十余所高校从事生物统计学教学多年、有丰富教学和科研经验的教师，集各家之所长，参考国内外同类教材编写而成。面向生命科学及其相关的农学、医学等专业背景的学习者，针对相关专业生物统计学或类似课程的教学需要，强调基础理论，突出实践应用。本书在编写方式上遵循由浅入深的认知规律，结合统计软件在生物统计学中的广泛应用，提供了针对常用的 SPSS 软件和 Excel 软件进行基本统计学运算的指导，具有很强的实用性。

本书分为基础理论和软件应用两部分。基础理论以讲解生物统计的原理、方法和过程为目标，系统介绍生物统计学的数理统计基础、数据统计分析的原理和方法，内容涵盖了各类数据的常用统计分析方法。软件应用以讲解软件操作的数据格式、分析步骤和结果解释为目标，重点介绍统计分析工作在 SPSS 软件中的实现，并简要介绍了 Excel 软件的生物统计应用。

全书共分 14 章，第 1 章由马纪（新疆大学）编写，第 2 章由刘小宁（新疆大学）编写，第 3 章由侯沁文（长治学院）编写，第 4 章由王玉（浙江中医药大学）、王晓俊（长春工业大学）编写，第 5 章由胡颖（哈尔滨工业大学）编写，第 6 章和第 14 章由万海清（湖南文理学院）、王文龙（湖南文理学院）编写，第 7 章由耿丽晶（辽宁医学院）编写，第 8 章和第 9 章由聂呈荣（佛山科学技术学院）、施文正（上海海洋大学）编写，第 10 章由陈国（华侨大学）编写，第 11 章和第 13 章由陈其新（河南农业大学）、李转见（河南农业大学）编写，第 12 章由彭明春（云南大学）、王有武（塔里木大学）、石培春（石河子大学）编写。彭明春、马纪对各章节内容进行了修改补充。云南大学党承林教授通审了全书，并提出了许多有益的建议，在此表示衷心感谢。

由于编者水平有限，书中不足之处在所难免，恳请读者批评指正，以便再版时修改完善。

编　者

2014 年 11 月

目　录

第1章 绪 论

1.1 生物统计学简介

统计学(statistics)是研究如何收集、整理、分析和解释数据的科学。它把数学语言引入具体的科学领域,并把具体科学领域中要解决的问题抽象为数学问题进行分析处理。统计学分为描述统计学(descriptive statistics)和推断统计学(inference statistics)两部分。描述统计学研究对客观现象进行数量计量、数据收集、加工、概括和表示的方法,不同的领域对数据的统计描述略有差异;推断统计学研究根据样本数据去推断总体的方法,是统计学的核心和主要内容。统计学也可分为理论统计学和应用统计学,理论统计学研究统计学的数学原理和方法,应用统计学将理论统计学的研究成果作为工具应用于各个科学领域。

统计学是数学的分支学科,但统计学并非简单地等同于数学。数学研究的是没有量纲或单位的抽象的数及其数量关系,统计学则是研究具体的、实际调查或试验获得的数据的数量规律;统计学与数学所用的逻辑方法不同,数学主要使用演绎法,而统计学是演绎法与归纳法相结合,以归纳法为主。

生物统计学(biostatistics)是研究收集、整理、分析和解释生物科学试验数据的科学,是统计学原理在生物学研究领域的应用。在生物科学学科体系中,生物统计学属于生物数学的范畴,在统计学学科体系中,生物统计学是应用统计学的分支。

生物科学是一门实验科学,在研究过程中会产生大量试验数据,这些数据由于受到随机因素的干扰,带有随机误差而具有不确定性,即同一个试验的多次重复,试验结果并不完全一样,需要对各种试验结果出现的概率大小作出判断,从偶然的不确定性中找出内在的规律性,剔除试验误差的干扰。生物统计学为我们提供解决这类问题的方法。

1.2 生物统计学的主要内容

生物统计学包括试验数据的获取、整理和分析等相关内容,具体来说,包括试验或调查设计、数据的整理(描述统计学)、概率论基础(统计理论基础)、统计推断方法(推断统计学)等内容。

1.试验或调查设计方法

在着手开展一项科学研究之前,需要根据提出的问题设计试验方案或调查方案,以便控制

试验误差。由于不同类型的数据或者不同设计方法所获得的数据在误差估计方面有较大的不同，生物统计学提供了一系列不同的数据分析处理方法。因此，如何科学地获得试验或调查数据，以便采用适当的统计方法进行数据分析，是获得合理、可信、可以解释的试验结果的一个重要前提。对此，生物统计学提供了若干试验或调查设计的方法。

在生态学、农学、流行病学、食品卫生检查等领域，现状研究的主要方法是调查，开展大范围抽样是常用的研究方法之一。调查设计是指整个调查计划的制订，包括调查研究的目的、对象与范围，调查项目及调查表内容，抽样方法的选取，抽样单位和抽样数量的确定，数据处理方法，调查组织工作，调查报告撰写等内容。合理的调查设计能够控制与降低抽样误差，提高调查的精确性，为获得总体参数的可靠估计提供必要的数据。

在研究某个因素的效应大小，或两个以上因素的主效应或相互作用时，开展有控制的比较试验设计是常用的研究方法。试验设计是指试验单位的选取、生物学重复数的确定及试验单位的分组等。合理的试验设计能够有效地控制和降低试验误差，提高试验的精确性，为统计分析获得试验处理效应和试验误差的无偏估计提供有效的数据。

简而言之，试验或调查设计主要解决合理地收集必要而有代表性资料的问题。统计学在不同学科应用时，针对不同的调查对象，试验和调查设计的具体方法会有所不同。

2.数据整理方法

通过试验，尤其是调查，获得大量数据后，由于数据的随机性，同样的试验或调查所得数据往往有一定范围的变异，需要对数据进行整理，以便发现数据内部所包含的规律，比如集中性、变异性、数据的分布形态等。数据整理的基本方法是根据数据出现的频率，编制频数统计表或绘制频数统计图。通过统计表图可以直观地看出所得数据的集中或离散情况，并计算代表样本资料数量特征的统计数（如平均数、标准差等），以此估计相应的总体参数（总体的平均数、标准差等）。

3.统计推断的基础理论

生物统计学的重要任务是建立由样本统计结果推断总体参数的方法，而这些方法都是以数据（随机变量）的概率以及概率分布为基础建立起来的，因此需要对概率的基础知识有所了解。生物统计学中所涉及的概率知识非常基础，主要包括随机事件和概率的定义、概率的分布、抽样分布等基础概率知识，非数学专业的学生在理解上基本没有障碍。

4.数据统计分析方法

数据的统计分析，是指通过样本数据推断总体参数（平均数、方差等）的过程，在统计学上称为统计推断。通过试验或调查获得了具有变异性的资料后，要了解资料之间的试验指标产生差异的原因，即变异是由处理效应所引起还是由随机误差所导致的。例如不同盐浓度对某种植物的生长有无影响，荒漠耐盐植物与非耐盐植物在耐盐性方面有无差异，农业害虫棉铃虫的解毒酶是否强于其他昆虫等。要回答上述问题，可利用显著性检验排除那些无法控制的偶然因素的干扰，将处理间是否存在本质差异这一问题以概率的形式揭示出来。假设检验的方法很多，常用的有 t 检验、方差分析、χ^2 检验等。

统计分析的另一个重要内容是研究成对或成组变量（试验指标或性状）间的关系，即相关分析与回归分析。通过对资料进行相关与回归分析，可以揭示变量间的内在联系。利用回归分析可以找出影响某个变量的因素，并据此开展预测预报研究。多元统计分析近 30 年来得到了迅速发展，并在自然科学和社会科学的许多领域得到广泛应用。

5.统计分析软件的应用

近些年来，随着计算机科学的发展和普及，各类统计软件迅速发展，统计学分析方法得到了极大的应用和推广，常用的统计软件有 SPSS、Excel、SAS、S-plus、Minitab、Statistica、Eviews 等。本书针对本课程相关专业学生的知识结构和软件的易用性，结合相关统计分析内容，介绍 SPSS 和 Excel 软件的应用。

通过对生物统计学内容的简要介绍，不难看出生物统计学在生物科学研究中的重要性，它是每一个从事生物科学研究的人必须掌握的基本工具。随着生物统计学方法的普及、统计学软件的不断发展，已有越来越多的科技工作者掌握并在实际研究工作中应用了生物统计学方法，并取得了显著的成效。

1.3 生物统计学发展简史

在介绍一门学科时，通常需要了解这门学科的发展简史，从而对该学科的内涵有深刻的理解。统计学的建立和发展与人类的统计实践活动密不可分。如果说统计学是随着人类计数活动而产生的，那么统计发展史可以追溯到距今有 5000 多年的远古社会。但是，能使人类的统计实践上升到理论，并概括总结成为一门系统的学科是近代的事情，距今只有 300 多年的历史。统计学发展的概貌大致可划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时期。

1.古典记录统计学

17 世纪中叶至 19 世纪中叶，在利用文字或数字记录与分析国家社会经济状况的过程中，初步建立了统计研究的方法和规则。统计学在这一阶段的意义和范围还不太明确，概率论被引入之后，这些方法逐渐成熟。

统计学是从拉普拉斯(P.S.Laplace，1749—1827)开始的，他是法国天文学家、数学家、统计学家，他的主要贡献包括建立了概率论，代表作是《概率分析理论》，该书把数学分析方法运用于概率论研究，建立了严密的概率数学理论。拉普拉斯还推广了概率论的应用，解决了一系列实际问题，例如在人口统计、误差理论中的应用。拉普拉斯提出了大数定律并尝试了大样本推断(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。他根据法国 30 个县市的人口出生率推算了全国的人口，这种利用样本来推断总体的思想方法为后人开创了抽样调查的方法。

德国著名数学家、物理学家、天文学家、大地测量学家高斯(C.F.Gauss，1777—1855)对统计学的误差理论作出了重要贡献。调查测量中的误差不仅不可避免，而且无法把握。高斯以他丰富的天文观察和土地测量经验，总结发现误差变异大多服从正态分布，运用极大似然法及其他数学知识，推导出测量误差的概率分布公式，并提出了“误差分布曲线”，即高斯分布曲线，也就是今天所说的正态分布曲线。1809 年高斯发表了统计学中最常用的最小二乘法。

2.近代描述统计学

近代描述统计学的形成在 19 世纪中叶至 20 世纪上半叶。这种“描述”特色是由一批研究生物进化的学者提炼而成的，代表人物是英国的高尔顿和他的学生皮尔逊。

高尔顿(F.Galton，1822—1911)是英国生物学家、统计学家。他于 1882 年建立了人体测量试验室，测量了 9337 人的身高、体重、呼吸力、拉力和压力、手击的速率、听力、视力、色觉等

人体资料，得出了“祖先遗传法则”，引入了中位数、百分位数、四分位数、四分位差以及分布、相关、回归等重要的统计学概念与方法。1901 年，高尔顿创办了《Biometrika》(生物计量学)杂志，首次提出了“Biometry”(生物统计学)一词，认为“生物统计学是应用于生物学科中的现代统计方法”。高尔顿及其学生虽然开展的是生物统计学研究，但在这一过程中，他们更重要的贡献是发展了统计学方法本身。

皮尔逊(K.Pearson，1857—1936)是英国数学家、哲学家、统计学家。他将生物统计学提升到了通用方法论的高度，首创了频数分布表与频数分布图，提出了分布曲线的概念。1900 年皮尔逊发现了 χ^2 分布，并提出了有名的 χ^2 检验法，后经费歇尔(R.A.Fisher，1890—1962)补充，成为小样本推断统计的早期方法之一。皮尔逊还发展了回归与相关的概念，提出复相关、总相关、相关比等概念，不仅发展了高尔顿的相关理论，还为之建立了数学基础。

3.现代推断统计学

现代推断统计学形成时间大致是 20 世纪初叶至 20 世纪中叶，此时无论是社会领域还是自然领域都向统计学提出了更多的要求。人们开始深入研究事物与现象间的关系，对其中繁杂的数量关系以及一系列未知的数量变化，单靠描述的统计方法已难以奏效，因而产生了“推断”的方法来掌握事物总体的真正联系以及预测未来的发展。

从描述统计学到推断统计学是统计学发展过程中的一大飞跃，这场深刻变革是在农业田间试验领域完成的，英国统计学家戈塞特和费歇尔对现代推断统计学的建立作出了卓越贡献。

1908 年，戈塞特(W. S. Gosset，1876—1937)首次以“学生”(Student)为笔名，在《Biometrika》杂志上发表了“平均数的概率误差”，为“学生氏 t 检验”提供了理论基础，成为统计推断理论发展史上的里程碑。后来，戈塞特又连续发表了相关系数的概率误差、非随机抽样的样本平均数分布、从无限总体随机抽样平均数的概率估算表等重要论文，为小样本理论奠定了基础。由于戈塞特的理论使统计学开始由大样本向小样本、由描述向推断发展，因此，可以认为是戈塞特开创了推断统计学。

费歇尔对统计学作出了很多重要贡献，他强调统计学是一门通用方法论。1924 年，费歇尔综合研究了 t 分布、χ^2 分布和 u 分布，使 t 检验也能适用于大样本，χ^2 检验也能适用于小样本。1925 年在《供研究人员使用的统计方法》中对方差分析和协方差分析进行了完整表述：方差分析法是一种在若干组能相互比较的资料中，把产生变异的原因加以区分的方法与技术，方差分析简单实用，大大提高了试验分析的效率，对大样本和小样本都可使用。1925 年提出了随机区组设计和拉丁方设计；1926 年发表了试验设计方法梗概；1935 年这些方法得到进一步完善，并首先在卢桑姆斯坦德农业试验站得到检验与应用，后来又被他的学生推广到许多其他科学领域。1938 年费歇尔与耶特斯合编了《F 分布显著性水平表》，为方差分析的研究与应用提供了方便。

费歇尔在统计学发展史上的地位是显赫的，他的研究成果特别适用于农业与生物学领域，但已经渗透到一切应用统计学中，由此所形成的推断统计学已经被广泛地应用。美国统计学家约翰逊(P.O.Johnson)在 1959 年出版的《现代统计方法：描述和推断》一书指出：“从 1920 至今的这段时期，称之为统计学的费歇尔时代是恰当的。”

1.4 统计学的常用术语

在正式进入统计学学习之前，需要先了解一些统计学的常用术语，这些术语定义了统计学

的基本元素，理解这些术语有助于进一步的学习。

1.总体、个体与样本

总体(population)是研究对象的全体，总体中的一个研究单位称为个体(individual)。例如，要了解某大学大一新生的身高，那么该校全体大一新生就是研究总体，每一个学生就是组成这个总体的个体。

样本(sample)是从总体中抽取的用于代表总体的一部分个体。通常情况下，了解大一新生的身高，可以选择几个专业的学生作为代表，被选中的学生就是代表这个总体的样本。可以用样本的平均数来估计总体的平均数，样本中个体数量越多，对总体的代表性越好。

包含有限多个个体的总体称为有限总体(finite population)，包含无限多个个体的总体称为无限总体(infinite population)。例如研究某一地区新生儿的体重，因为新生儿的出生是无止境的，所以这一总体是一个无限总体；要调查某大学大一学生的身高，这一总体则是有限的，其中每个学生身高的测定值为这一总体的一个个体。在实际研究中还有一类假想总体，例如进行几种饲料的饲养试验，实际上并不存在用这几种饲料进行饲养的总体，只是假设有这样的总体存在，把所进行的每一次试验看成假想总体的一个个体。在生物科学研究中，广泛采用假想总体及其样本开展研究。

前面提到，统计学的逻辑关系是以归纳法为主，其含义就是通过样本来推断总体。为什么不直接研究总体呢？因为对于无限总体和假想总体，无法对其进行完全调查或观测；对于个体数量很多的有限总体，要获得全部观测值须花费大量人力、物力和时间或者观测值的获得带有破坏性，也不适用于直接研究总体。因此，通过样本来推断总体是统计分析的基本特点。

样本中所包含的个体数量称为样本容量或样本大小(sample size)。样本容量记为 n，通常把 $n\leqslant 30$ 的样本称为小样本，$n>30$ 的样本称为大样本，它们在统计推断方法上有若干区别。为了能可靠地从样本来推断总体，要求样本具有一定的个体数量和代表性。

只有从总体随机抽取的样本才具有代表性。所谓随机抽样(random sampling)，是指总体中的每一个个体都有同等的被抽取的机会组成样本。

2.参数与统计数

统计学上，总体或样本的特征用数值来描述，称为特征数(eigenvalue)；这种特征包括集中性和离散性两个方面，通常用平均数描述总体或样本的集中性，用标准差描述总体或样本的离散性。

由总体计算的特征数称为参数(parameter)，由样本计算的特征数称为统计数(statistic)。通常用希腊字母表示参数，例如用 μ 表示总体平均数，用 σ 表示总体标准差；用拉丁字母表示统计数，例如用 $\bar{x}$ 表示样本平均数，用 s 表示样本标准差。

对总体和样本的特征数加以区别是很有必要的，它们之间有一种逻辑关系，在统计学上由样本特征数可以推断或估计总体的特征数。总体参数由相应的样本统计数来估计，例如用 $\bar{x}$ 估计 μ，用 s 估计 σ。平均数与标准差是一对非常重要的特征数。

3.准确性与精确性

准确性与精确性是对在试验中所获得样本数据的质量的一种度量。准确性(accuracy)也叫准确度，是指在试验中某一试验指标的观测值与其真值接近的程度。直观上理解，观测值与真值接近，则其准确性高，反之则低。精确性(precision)也叫精确度，是指同一试验指标的重复观测值彼此接近的程度。若观测值彼此接近，则观测值精确性高，反之则低。由于真值常常不知道，所以准确性只是一个概念，不易度量，而精确性在统计学中可以通过随机误差的大小

加以度量。

4.随机误差与系统误差

试验中的误差问题是统计学的核心问题。观测数据之所以表现出随机性波动，主要是由随机误差引起的，正确估计出试验中的误差，对于统计推断的效率至关重要。前面提到的试验设计问题，实际上也是如何控制试验误差的问题。

试验中出现的误差分为两类：随机误差(random error)与系统误差(systematic error)。

随机误差是由无法控制的内在和外在的偶然因素所造成的，是客观存在的，在试验中，即使十分小心也难以消除。如试验材料的初始条件、培养条件、管理措施等尽管在试验中力求一致，但不可能绝对一致。随机误差影响试验的精确性，随机误差愈小，试验的精确性愈高。因为各个样本平均数之间的差异实际上是由抽样造成的，所以随机误差也叫抽样误差(sampling error)。

系统误差也叫片面误差(lopsided error)，是由试验材料的初始条件不同或测量仪器不准等引起的倾向性或定向性偏差。如供试对象年龄、初始重、性别、健康状况等存在差异，或饲料种类、品质、数量、饲养条件不完全相同，或测量的仪器调试存在差异等情况会导致系统误差。系统误差影响试验的准确性，应当通过采用适当的试验设计、精心完成试验操作来加以控制。

习题

习题 1.1　什么是生物统计学？生物统计学有哪些主要内容？

习题 1.2　解释以下概念：总体、个体、样本、样本容量、参数、统计数。

习题 1.3　准确性与精确性有何不同？

习题 1.4　随机误差与系统误差有何不同？

第2章 数据的描述性分析

在生物学研究中，通过在一定条件下对某种事物或现象进行调查或试验，可获得大量的数据(data)，或称为资料。这些数据在未整理之前，是一堆无序的数字。描述性分析就是要通过对这些数据的整理归类，制作统计表、绘制统计图，计算平均数、标准差等特征数来反映数据的特征，揭示数据的内在规律。

2.1 数据的类型

对调查或试验获得的数据进行分类是统计归纳的基础，如果不进行分类，大量的原始数据就不能系统化、规范化，不能反映数据本身的特征和规律。在调查或试验中，由于使用的方法和研究的性状特征不同，数据的性质也就不同，生物的性状可以大致分为数量性状和质量性状两大类，取得的数据也可以是定量的或定性的，分别称为数量性状数据和质量性状数据。

1.数量性状数据

数量性状数据(data of quantitative character)是指通过测量、度量或计数取得的数据。根据数据的特征又分为连续型数据和离散型数据。

1) 连续型数据

连续型数据(continuous data)或称为计量数据(measurement data)，是指用测量或度量方式得到的数量性状数据，即用度、量、衡等计量工具直接测定获得的数据。如身高、作物产量、蛋白质含量等。这类数据的观测值可以是整数，也可以是带小数的数值，其小数位数由测量工具或统计要求的精度而定，数据之间的变异是连续的，因此也称为连续性变量数据。

2) 离散型数据

离散型数据(discrete data)或称为计数数据(enumeration data)，是指用计数方式得到的数量性状数据。如不同血型的人数、鱼的数量、白细胞数等。这类数据的观察值只能以整数表示，不会出现带小数的数值。观察值是不连续的，因此也称为非连续性变量数据。

2.质量性状数据

质量性状数据(data of qualitative character)或称为属性数据(attribute data)，是指对某种现象进行观察而不能测量的数据。如土壤的颜色、植物叶的形状等。在统计分析中，质量性状数据需要进行数量化以后才能参与统计分析。

质量性状数据数量化的方法主要有二值化和等级化两种方法。二值化是用1和0分别表示某一特征的有和无。等级化是将数据用若干等级表示，如植物的抗病能力可划分为3(免疫)、2(高度抵抗)、1(中度抵抗)、0(易感染)4个等级。数量化后的质量性状数据参照离散型

数据的处理方法进行处理。

2.2 数据的整理

数据的整理是指根据数据的数量和数值范围，对数据进行分组和各组的频数统计，然后编制次数(频数)分布表或绘制次数(频数)分布图。

对原始数据进行检查核对后，根据数据中观测值的数量确定是否分组。当观测值不多($n \leqslant 30$)时，一般不分组直接进行数据整理。当观测值较多($n > 30$)时，需将观测值分成若干组，制成次数(频数)分布表，观察数据的集中性和变异性情况。不同类型的数据，其整理的方法略有不同。

1.离散型数据的整理

离散型数据基本上采用单项式分组法整理，其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，可将数据中每个变量分别归入相应的组内，然后制成次数(频数)分布表。下面以100只芦花鸡每月产蛋数(表2-1)为例，说明离散型数据的整理。

表2-1 100只芦花鸡每月产蛋数

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

当所调查数据的变量值较少时，以每个变量值为一组；当数据较多、变量值范围较大时，以几个相邻观察值为一组，适当减少组数，这样资料的规律性更明显，对资料进一步计算分析也比较方便。

表2-1的数据，变量值为11～17，可分成7组。然后以唱票方式记录每个变量值(产蛋数)出现的次数，便可得到次数(频数)分布表(表2-2)。

原来无序的原始数据经整理后，从中可以发现有35%的芦花鸡每月产蛋数为14枚，有72%的芦花鸡月产蛋数为13～15枚；产蛋12枚及以下的有12%的个体，产蛋16枚及以上的有16%的个体。

表2-2 产蛋数的次数(频数)分布表

产蛋数	次数	频率	累计频率
11	4	0.04	0.04
12	8	0.08	0.12
13	18	0.18	0.30
14	35	0.35	0.65
15	19	0.19	0.84
16	11	0.11	0.95
17	5	0.05	1.00

2.连续型数据的整理

连续型数据不能按离散型数据的分组方法进行整理，一般采用组距式分组法，即在分组前确定全距、组数、组距、组中值及各组上下限，然后将全部观测值按照大小归入相应的组。下面以100例30～40岁健康男子血清总胆固醇含量(mmol/L)测定结果(表2-3)为例，说明其整理的方法及步骤。