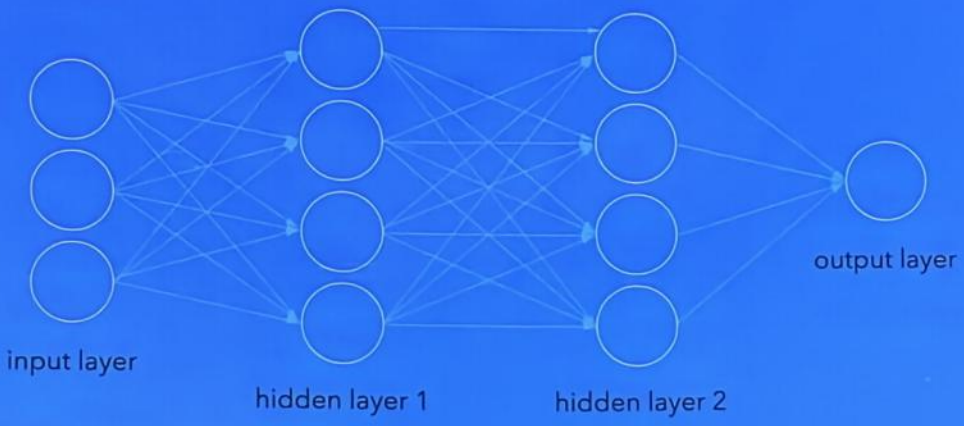
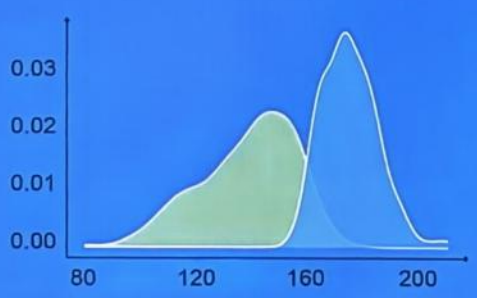
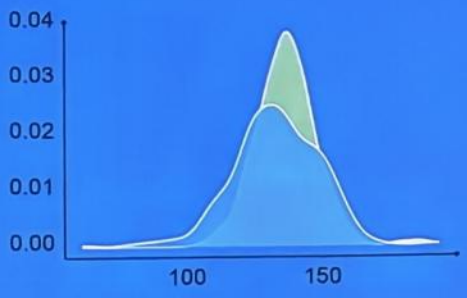
普通高等学校“十四五”规划生命科学类特色教材
普通高等教育新形态一体化教材

生物统计学——生物大数据的概率统计模型与机器学习方法

主编 宁 康







华中科技大学出版社
http://press.hust.edu.cn

普通高等学校“十四五”规划生命科学类特色教材
普通高等教育新形态一体化教材

生物统计学——生物大数据的概率统计模型与机器学习方法

主 编 宁 康
编 委 白 虹 计 磊
钟朝芳 张宇昊

华中科技大学出版社
http://press.hust.edu.cn
中国·武汉

内 容 简 介

本书主要基于作者近年来教授本科生“生物统计学”和研究生“生物信息学”等课程资料，同时参考了国内外众多顶级教学和科研资料编写而成。本书共分为5章：第1章介绍生物统计学的基础概念和基本研究方法；第2章介绍传统生物统计学方法及其应用，包括试验资料的搜集与整理、生物统计量的计算和表征、假设检验及其应用等；第3章介绍生物大数据的特征与挑战，包括生物大数据的特征、生物大数据分析的常规方法、生物大数据经典案例分析等；第4章介绍生物大数据与概率统计模型，包括大数据机器学习基础、聚类降维、概率统计模型方法等；第5章介绍面向生物大数据挖掘的深度学习，包括深度学习的概念及方法、深度学习应用于生物大数据分析的基本流程和经典案例等。每章都附有练习题，供读者参考。

本书具有一定的深度和广度，可以作为生物统计学、生物大数据及机器学习相关课程的教学参考书，也可供生物学、统计学、机器学习、生物大数据等领域的科学工作者阅读。

图书在版编目(CIP)数据

生物统计学：生物大数据的概率统计模型与机器学习方法/宁康主编. —武汉：华中科技大学出版社，2023.7

ISBN 978-7-5680-8635-6

Ⅰ.①生… Ⅱ.①宁… Ⅲ.①生物统计-概率统计-统计模型 ②生物统计-机器学习 Ⅳ.①Q-332

中国国家版本馆CIP数据核字(2023)第126863号

生物统计学：生物大数据的概率统计模型与机器学习方法 宁 康 主编

Shengwu Tongjixue：Shengwu Dashuju de GailüTongji Moxing yu Jiqi Xuexi Fangfa

策划编辑：王汉江

责任编辑：余 涛

封面设计：廖亚萍

责任监印：周治超

出版发行：华中科技大学出版社(中国·武汉) 电话：(027)81321913

武汉市东湖新技术开发区华工科技园 邮编：430223

录 排：武汉市洪山区佳年华文印部

印 刷：武汉市籍缘印刷厂

开 本：787mm×1092mm 1/16

印 张：15.5

字 数：357千字

版 次：2023年7月第1版第1次印刷

定 价：49.80元

线上作业及资源网使用说明

建议学员在PC端完成注册、登录、完善个人信息及验证学习码的操作。

一、PC端学员学习码验证操作步骤

1. 登录

(1) 登录网址 http://bookcenter.hustp.com/，完成注册后单击“登录”按钮。输入账号、密码(学员自设)后，提示登录成功。

(2) 完善个人信息(姓名、学号、班级等信息请如实填写，因线上作业计入平时成绩)，将个人信息补充完整后，单击“保存”按钮即可完成注册登录。

2. 学习码验证

(1) 刮开本书封底所附学习码的防伪涂层，可以看到一串学习码。

(2) 在个人中心页单击“学习码验证”按钮，输入学习码，单击“验证”按钮，即可验证成功。单击“学习码验证”→“已激活学习码”按钮，即可查看刚才激活的图书学习码。

3. 查看课程

在图书搜索框中搜索书名，并单击图书详情页右上角的“加入课程”按钮，返回个人中心，单击“我的课程”按钮，即可看到新激活的课程，单击“课程”按钮，进入课程详情页。

4. 做题测试

在课程详情页可查看相关资源，进入习题页，选择具体章节开始做题。做完之后单击“我要交卷”按钮，随后学员即可看到本次答题的分数统计。

二、手机端学员扫码操作步骤

1. 手机扫描二维码，提示登录；新用户先注册，然后再登录。
2. 登录之后，按页面要求完善个人信息。
3. 按要求输入本书的学习码。
4. 学习码验证成功后，即可扫码看到对应的习题。
5. 习题答题完毕后提交，即可看到本次答题的分数统计。

扫码做题

任课老师可根据学员线上作业情况给出平时成绩。

若在操作上遇到什么问题可咨询陈老师(QQ:514009164)。

线上作业及资源网使用说明

PREFACE

序言

作为研究生命科学最基础的工具性课程之一，“生物统计学”越来越被从事生物学基础教学、生命科学研究的教师和科技工作者高度重视。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，而是需要针对大量调查和测定的数据，应用统计学方法，分析和解释其数量的变化，以制订正确的实验计划，科学地对实验结果进行分析，进而作出符合科学实际的推断。

随着组学技术的快速发展，生命科学所涉及的生物大数据不论从数量上还是类型上，都有了质的飞跃。要从海量异质性的生物大数据中挖掘重要的规律，统计和机器学习方法是非常有效的手段之一。

然而，针对当前“生物统计”领域所要解决的生物大数据挖掘的众多问题，目前的《生物统计学》教材还存在着一系列局限性，例如，部分方法不适应大数据的研究特征，部分内容理论和实践结合不强，部分案例的代表性不强，概率统计模型与机器学习方法的介绍较为粗浅，等等。

本书主要基于作者近年来教授本科“生物统计学”和研究生“生物信息学”等课程资料，同时参考了国内外众多顶级教学和科研资料编写而成。本书共分为5章：第1章和第2章介绍生物统计学的基础及其应用，包括生物统计学的基本概念和基本研究方法、试验资料的搜集与整理、生物统计量的计算和表征等；第3章介绍生物大数据的特征与挑战，包括生物大数据的特征、生物大数据分析的常规方法、生物大数据经典案例分析等；第4章介绍生物大数据与概率统计模型，包括大数据机器学

习基础、聚类降维、概率统计模型方法等;第5章介绍面向生物大数据挖掘的深度学习,包括深度学习的概念及方法、深度学习应用于生物大数据分析的基本流程和经典案例等。每章都附有练习题,供读者参考。

本书通俗易懂,具有一定的深度和广度,可以作为"生物统计学""生物大数据"及"机器学习"相关课程的教学参考书,也可供生物学、统计学、机器学习、生物大数据等领域的科学工作者参考。

限于作者水平和掌握资料的局限性,本书难免存在疏漏和不妥之处,欢迎各位专家和广大读者给予批评指正。

宁 康

2023年6月于华中科技大学

PREFACE

前言

随着组学技术的不断进步，大量不同类别的生物大数据已经产生。合理地分析这些生物大数据，构建可靠的数据模型，将有望发掘重要的生物学规律，指导具体应用。

“大数据开启了一次重大的时代转型。就像望远镜让我们能够感受宇宙，显微镜让我们能够观测微生物一样，大数据正在改变我们的生活以及理解世界的方法，成为新发明和新服务的源泉，而更多的改变正蓄势待发”，互联网专家维克托·迈尔·舍恩伯格在《大数据时代》一书中这样描述大数据。面对海量数据，谁能更好地处理、分析数据，谁就能真正抢得大数据时代的先机。大数据分析对生物医疗行业的发展非常重要。生物医疗行业早就遇到了海量数据和非结构化数据的挑战，大数据分析技术的发展让这些数据的价值得以充分发挥，其中，基因组学是大数据在医疗行业的经典应用。以云计算为基础的大数据分析技术不仅加速了基因序列分析的速度，也让其成本不断降低。

机器学习(machine learning)经常与人工智能（AI）一起讨论，它将是第四次工业革命的主要推动元素。许多专家都提醒我们不能忽视机器学习带来的巨大影响力。在生物医疗领域，机器学习为医疗保健提供者提供了开创性的工具。

本书对学习思维和分析策略的培养，如同传统生物数学和生物信息学的逻辑思维意向，是沿着“实际问题→抽象出的统计问题→统计建模”这条脉络展开。统计分析和深度学习，不是简单地罗列现有的算法技

术,而是试图强调“如何观察问题的结构”、强调“如何基于问题的结构进行统计建模”。求解问题的过程,不应当只是逐个尝试各个模型和技术,也不是纯粹依赖于灵感,而是应该依赖于我们对需要进行统计分析的问题的认识;我们对问题结构认识得越深入,越有助于统计模型的设计和分析。

只讲清楚统计模型和深度学习模型本身不算太困难的任务,但要想讲清楚模型背后的观察、思考和设计过程,却相当困难,当然也是非常重要的一个任务。在本书中,我们常常采用“观察统计问题的结构,先设计一个初步模型,然后观察建模的过程以及问题结构,进而改进模型设计”的方式,试图展现出统计模型背后的思考过程。

基于上述逻辑思维,本书内容组织如下:首先,介绍生物统计学基础概念,传统生物统计学的方法及其应用方法;然后,进入生物大数据与概率统计模型章节,通过贝叶斯推断、隐马尔可夫模型、最大似然推断等方法的层层推进,配合翔实的用例,完整地介绍统计建模方面的知识;最后将统计建模方法延展到面向生物大数据挖掘的深度学习,并介绍相关应用。

编　者

2022 年 9 月

CONTENTS

目录

第1章

生物统计学基础

“对统计学的一知半解常常会造成一些不必要的上当受骗，对统计学的一概排斥往往会造成某些不必要的愚昧无知。”

——C. R. Rao

“在终极的分析中，一切知识都是历史；在抽象的意义下，一切科学都是数学；在理性的基础上，所有的判断都是统计学。”

——C. R. Rao

生物统计学是生物数学中最早形成的一大分支，它是在用统计学的方法和原理研究生物学的客观现象及问题的过程中形成的，生物学中的问题又促使生物统计学中大部分基本方法得到进一步发展。生物统计学是应用统计学的分支，它将统计方法应用到医学及生物学领域。

生物统计学的内容包括试验设计和统计分析。试验设计是指应用数理统计的原理与方法，制定试验方案，选择试验材料，合理分组，降低试验误差，使我们可以利用较少的人力、物力和时间，获得丰富而可靠的数据资料。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，认识客观事物的本质和规律性，使我们对所研究的资料得出合理的结论。由于事物都是相互联系的，统计不能孤立地研究各种现象，而必须通过一定数量的观察，从这些观察结果中研究事物间的相互关系，揭示事物客观存在的规律性。统计分析与试验设计是不可分割的两部分。试验设计必须以统计分析的原理和方法为基础，而正确设计的试验又为统计分析提供了丰富、可靠的信息，两者紧密结合以推断出合理的结论，并不断地推动应用生物科

学研究的发展。

生物统计学已在科学研究和生产实践中得到极为广泛的应用,其基本功能有:

(1) 为科学地整理、分析数据提供方法。

我们做任何工作,都必须掌握基本情况,做到心中有数,才能有的放矢,提高工作质量。在生物学研究中也不例外,必须有计划地搜集资料并进行合理的统计分析,通过调查得到数据,经过加工整理,从中归纳出事物的内在规律性,用于指导相关试验。

(2) 判断试验结果的可靠性。

由于存在试验误差,从试验得到数据资料必须借助统计分析方法才能获得可靠的结论。

(3) 通过统计模型预测发展趋势。

建立统计模型的核心目的,是达成预测效果,尤其是预测事物发展的规律。以疾病的发展趋势为例,针对人群分类、疾病发生发展、术后恢复等生物学领域关键问题,进行统计建模和预测,具有较高的实用价值。

(4) 提供试验设计的原则和方法。

做任何调查或试验工作,事先必须有周密的计划和合理的试验设计,它是决定科研工作成败的一个重要环节。一个好的试验设计可以用较少的人力、物力和时间,最大限度地获得丰富而可靠的资料,尽量降低试验误差。

(5) 为学习其他课程奠定基础。

我们要学好遗传学、育种学等学科,就必须学好生物统计学。例如,数量遗传学就是应用生物统计方法研究数量性状遗传与变异规律的一门学科,如果不懂生物统计学,则很难完全掌握遗传学。此外,阅读中外科技文献也常常会碰到统计分析问题,有生物统计的基础知识才能更容易地理解文献的实验结果。因此,生物科学工作者必须学习和掌握统计方法,才能正确认识客观事物存在的规律性,提高工作质量。

生物统计学通常被应用于样本间的比较和分布分析等,包括统计量的计算和表征、数据的分布分析、数据的分组和比较、假设检验等。

1.1 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,它是应用数理统计的原理和方法来分析、解释试验调查资料,以及生物界各种现象的一门科学。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，以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，应用统计方法对生物学研究进行分析就显得特别重要。生物学研究的实践证明，只有正确地应用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，属于生物数学范畴。生物统计学以数学的概率论为基础，涉及数列、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，主要偏重于统计原理的介绍和具体分析方法的应用。

1.2 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。试验设计主要有对比试验设计、随机区组试验设计以及正交试验设计等。统计分析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量资料所隐藏的信息。

(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。偶然因素的影响是造成试验误差的重要原因。要想正确判断一个试验结果是由试验因素造成的还是试验误差造成的，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,在这种推断中,统计原理和方法提供了理论依据。

(4) 提供试验设计的一些重要原则。为了以较少的人力、物力和财力获取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、时间等条件下,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,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也不可能挽救试验的失败。

1.3 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年)系统地论证了大数定律。后来Bernouli的后代D. BernDouli(1700—1782年)将概率论的理论应用到医学和人类保险。

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年)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获得成功,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年)经过共同努力,于1895年创建了伦敦大学生物统计实验室,于1889年出版了《自然的遗传》一书。在该书中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公式。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(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,遗传研究中的孟德尔豌豆杂交试验,高品质豌豆与低品质豌豆杂交后,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,需通过χ^2进行检验。

Plarson的学生Gosset(1876—1937年)对样本标准差进行了大量研究,于1908年以笔名"Student"在《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，特别是在他出版了《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 Fisher 方差分析问世以来，各种数理统计方法不但在实验室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 年)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，即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。

国内对生物统计学的应用始于 19 世纪 30 年代。新中国成立后，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分析等都取得了较好的成果。

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。由于数学在生物学和农学中的应用，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。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和深入。

1.4 常用统计学术语

1.4.1 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。个体有限的总体称为有限总体，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。个体极多或无限多的总体称为无限总体，例如，某一地区棉田棉铃虫的只数，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