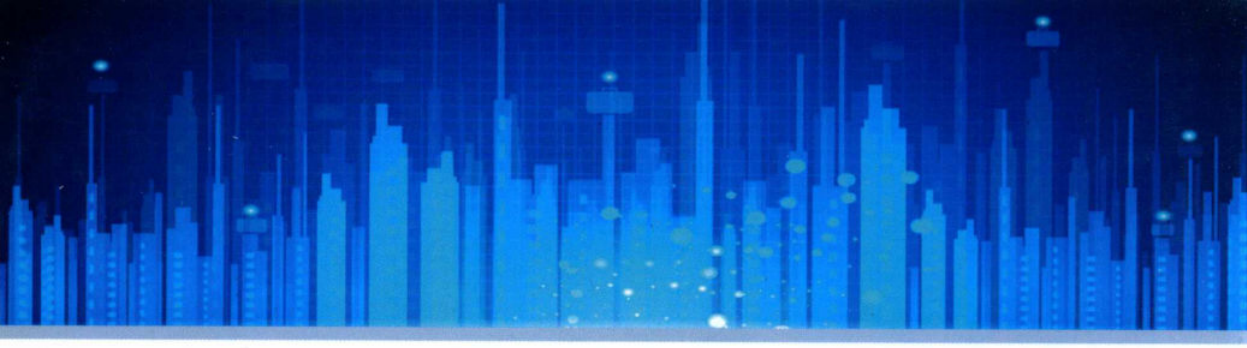
普通高等教育农业部“十二五”规划教材

生物统计学

章元明◎主编

本教材的内容密切联系生物技术
生物科学等专业的专业内容
强调统计推理思想
介绍了正交设计
回归设计
均匀设计三种处理设计
文字简练和通俗易懂

中国农业出版社



封面设计：姜　欣
版式设计：杜　然

欢迎登录：中国农业出版社网站http://www.ccap.com.cn
全国农业教育教材网http://www.qgnyjc.com
欢迎拨打中国农业出版社教材策划部热线：010-59194971，59194972
购书敬请关注中国农业出版社天猫旗舰店：

ISBN 978-7-109-22952-5
9 787109 229525 >
定价：33.50元

普通高等教育农业部“十二五”规划教材

生物统计学

章元明　主编

中国农业出版社

图书在版编目（CIP）数据

生物统计学 / 章元明主编. —北京：中国农业出版社，2017.5
普通高等教育农业部“十二五”规划教材
ISBN 978-7-109-22952-5

Ⅰ.①生… Ⅱ.①章… 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（2017）第 109096 号

中国农业出版社出版
（北京市朝阳区麦子店街 18 号楼）
（邮政编码 100125）
责任编辑 刘 梁
文字编辑 魏明龙

北京万友印刷有 发行
2017 年 5 月 刷

开本：787

编 写 人 员

主　编　章元明

副主编　徐辰武　唐章林

编　委　（按姓氏拼音排序）

崔党群（河南农业大学）

冯建英（南京农业大学）

唐章林（西南大学）

童春发（南京林业大学）

徐辰武（扬州大学）

杨锦忠（青岛农业大学）

张　瑾（南京农业大学）

章元明（华中农业大学）

审　稿　李加纳（西南大学，主审）

邢永忠（华中农业大学）

马朝芝（华中农业大学）

前　言

生物统计学是应用概率论和数理统计学原理及方法来研究怎样以有效的方式搜集、整理、分析带有随机性统计数据的一门学科。该门课程一直是高等院校生物技术、生物科学、植物保护、动植物生产和医学等相关本科专业的基础课程，对于大学生研究课题设计、显著性差异原因分析和或然性归纳推理思想训练发挥着重要作用。随着计算机应用普及、组学数据分析对统计方法提出更高要求，我们针对生物技术和生物科学等专业人才培养的特点与要求，组织了7所高等学校的一线骨干教师，精心编写了本教材。

本教材在简要介绍生物统计学概述的基础上，着重介绍统计数据的整理与描述、理论分布与抽样分布、统计假设检验、方差分析、次数资料的假设检验、回归分析与相关分析和试验设计。本教材密切联系生物技术与生物科学等专业的内容，强调统计推理思想，并且介绍了正交设计、回归设计和均匀设计三种处理设计，力求文字简练和通俗易懂，我们希望对其他专业和科研人员也具有参考价值。

本教材共分为8章，编委分工为：章元明（第一章、第四章第一节）、冯建英（第二章）、童春发（第三章）、崔党群（第四章第二至五节、第六章第四、五节）、唐章林（第五章）、张瑾（第六章第一至三节）、徐辰武（第七章）和杨锦忠（第八章）。全书由章元明统稿，为保持全书统一，避免重复，增加可读性，对全书内容进行了修改。张瑾、冯建英、杨锦忠和崔党群协助整理附表，张瑾和杨锦忠协助整理附录SAS程序，冯建英和童春发重新绘制插图，各位编者及我的研究生温阳俊、布素红和刘庆参与整理全书。西南大学李加纳教授（主审）、华中农业大学邢永忠教授和马朝芝教授审阅了该书。在此，对各位专家的参与和辛勤劳动表示衷心的感谢。

本教材的出版由于种种原因拖延至今。虽然统稿工作经过多次反复讨论与修改，但是鉴于编者水平所限，书中错误在所难免，恳请读者批评指正，以便再版时修改。

章元明

2016年11月

目　录

第一章 绪 论

第一节 生物统计学的基本概念和推理思想

一、生物统计学的基本概念

生物统计学是应用概率论和数理统计学的原理和方法来研究怎样以有效的方式搜集、整理、分析带有随机性的统计数据，并对所研究的问题做出统计推断，对可能做出的对策提供依据或建议的一门学科。

统计数据主要来自统计调查和科学试验。统计调查主要用于社会经济统计学，获得的统计数据表示某一地理区域或者论题的自然经济要素特征、规模、结构和水平等指标；科学试验主要用于生物统计学，其统计数据是在科学试验中获得的试验数据。若比较新培养基A与当前使用培养基B对成苗率的影响，每种培养基设置10个培养皿，获得的20个是否成苗的1和0的数据就是试验数据。

统计数据具有随机性。在统计调查中，为了使调查结果能真实地反映全貌、不偏不倚，调查个体就不能随意选择，就要求每个个体被调查的概率是相等的，即随机调查，这样获得的数据就具有随机性。在科学试验中，每个处理安排在哪个试验单元上也不是随意的，为了消除系统误差，有必要采用随机排列，这样获得的数据也具有随机性。

二、生物统计学的推理思想

生物统计学是通过个别的试验研究得出一般性的结论，因此，其推理应当是归纳推理。归纳推理一般包括简单枚举法和科学归纳法两种。通过简单枚举法，容易得到“所有的天鹅都是白的”、“血是红的”和“乌鸦是黑的”，出现“以偏概全”的错误。科学归纳法虽然比简单枚举法可靠得多，但客观事物的属性往往是多种多样的，不易得出事物与属性间的必然联系。实际上，这两种归纳推理的局限性在于从个别事物推出的一般性结论，已超出前提规定的范围，因而这种或然性推理不宜用必然性结论来表示，应当用或然性或可能性结论来表示，这就是生物统计学的或然性归纳推理或概率归纳推理。

为了理解这种推理思想，用上述培养基例子来说明。显然，培养基A和B平均成苗率的差异有两个原因：一是A和B间的真实差异；二是随机误差。究竟是哪一种原因呢？用肯定性结论似乎缺乏证据。若给出两者出现的概率大小，当随机误差原因出现的概率小到可以忽略的程度时，推断存在真实差异。这是可以接受的。这就是概率归纳推理。在数学中，假定$A=B$和$B=C$，则有$A=C$。其前提条件是$A=B$和$B=C$成立的概率是100%。若它们成立的概率只有20%，且两者相互独立，则$A=C$成立的概率只有4%。这时，经常认为$A\neq C$。

三、生物统计学与大学生素质教育

在高校实施素质教育的过程中，出现了各种人才培养模式和教学内容体系的改革方案。

然而，这些改革方案最终都要体现在课程体系的改革：以构建培养创造力为核心的课程体系改革模式。由于生物统计学的归纳推理思想以及其与科学研究能力培养的关系密切，在课程体系改革模式中备受农业高校农学类各专业与综合性大学生物类各专业的重视。

大学生科学研究素质的培养主要体现在发现问题的能力和研究课题的设计与实施两个层面。发现问题的能力培养是研究生阶段的主要目标之一；研究课题的设计与实施能力的培养则是本科阶段的主要目标之一。生物统计学的教学正可达到这一培养目标。生物统计学包括试验设计和统计方法两个有机联系的组成部分。通过试验设计的教学可提高大学生设计课题试验方案的能力，使之明确课题的研究目的、试验因素与水平以及试验设计方法等方面的内容，这对大学生把握课题是有帮助的。统计方法的教学能让学生找出产生处理间差异的原因，进而获得新的知识。同时，寻找原因是通过观察差异与试验误差的比较来进行的。因此，试验误差及其控制是影响统计结果的重要因素。然而，试验误差的主要来源有很多，就要求我们必须对试验误差进行有效控制。只有有效地控制试验误差才能提高试验精度。因此，有必要教育大学生在科研工作中要做到操作仔细和一丝不苟，从而提高学生的科研素质和水平。此外，在科学研究中，从试验资料中揭示出潜在的规律性极其重要，这需要正确、灵活地运用统计方法。这说明在生物统计学的教学中，除让学生弄清各种统计方法的内涵外，还应使学生能够正确地选择最适的统计方法，以揭示资料潜在的规律，达到研究的最终目的。因此，生物统计学是提高大学生科学研究素质的重要课程。

第二节　生物统计学的功用

一、常用的生物统计学术语

生物统计学是一门应用统计学，涉及较多的统计学概念、计算公式和用表；从推断方式上，要求摆脱传统的必然性推断，接受概率归纳推断。这对初学者来说有一定难度。为了便于初学者学习，本书通过大量实例及与其 SAS 程序相结合的方式，从应用角度来介绍生物统计学的基本概念、基本原理和基本方法，并附有一定数量的习题供初学者练习，以帮助初学者正确理解生物统计学的基本概念和基本原理，掌握并应用所介绍的基本试验设计与统计分析方法，以达到搜集、整理和分析统计数据的目的。

在这一节里，介绍生物统计学中几个最常用的术语。

（一）总体与样本

根据研究目的确定的、符合指定条件的全部研究对象称为总体（population）。构成总体的每一个单元，称为个体（individual）。例如，研究 2015 年上海市 16 岁男中学生的身高。凡是 2015 年上海市 16 岁的男中学生，他们的身高就构成总体，而每一个身高观测值是一个个体。总体可以分为有限总体与无限总体两种。个体数有限的总体称为有限总体（finite population），个体数无限的总体称为无限总体（infinite population）。

从总体中抽取一部分个体的集合就是样本（sample）。样本中所含个体数称为样本容量（sample size），以 n 表示。根据样本容量大小，又分为大样本（$n \geqslant 30$）和小样本。样本是总体的缩影，应该能反映总体的特征特性。但是，它毕竟只是总体的一部分，和总体情况应当有所不同。统计分析的核心在于由样本的信息推断总体的信息，即样本推断总体。为了可靠地推断总体，要求样本具有一定的大小和代表性。只有通过随机抽样（random sampling）

获得的样本才具有代表性。随机抽样就是等概率抽样，即总体中每个个体有同等的机会被独立地抽取到。然而，样本毕竟只是总体的一部分，尽管样本具有一定的大小，也具有代表性，但通过样本来推断总体也不可能是百分之百的正确，即使有很大的可靠性但仍有一定的错误率，这是统计分析的又一特点。一般地，生物统计学中讨论的样本是随机样本。

（二）参数与统计数

总体或样本的数量特征需要用特征数来描述。由总体计算的特征数称为参数（parameter），用希腊字母表示。例如，用 μ 表示总体平均数，用 σ 表示总体标准差。由样本计算的特征数称为统计数（statistic），用拉丁字母表示。例如，用 $\bar{y}$ 表示样本平均数，用 s 表示样本标准差。总体参数由相应的统计数来估计。例如，用样本平均数 $\bar{y}$ 估计总体平均数 μ。

（三）点估计与区间估计

生物统计学的统计推断包括参数估计与假设检验。参数估计可分为点估计（point estimation）和区间估计（interval estimation）。点估计也称定值估计，它是以抽样得到的样本统计数作为总体未知参数的估计值；区间估计是从点估计值和抽样误差出发，按给定的保证概率建立包含待估计参数的区间。其中这个给定的保证概率值称为置信度，这个包含待估参数的区间称为置信区间（confidence interval）。

（四）无偏估计

根据不同样本就会得到不同的总体参数估计值。这样，要确定一个估计值的好坏，就不能仅仅依据某次抽样的结果来衡量，而必须由大量抽样的结果来衡量。因而，自然而基本的衡量标准是要求无系统偏差。换言之，尽管在一次抽样中得到的估计值不一定恰好等于待估参数的真值，但是在大量重复抽样时，所得估计值的平均值应与待估参数真值相同。若估计量的数学期望（平均数）等于未知参数的真值，则该统计量就是总体参数的无偏估计量。例如，样本平均数 $\bar{y}$ 即为总体平均数 μ 的无偏估计量。

（五）随机误差与系统误差

在生物学与农业试验中，试验结果除受试验因素影响外，还受到许多其他非试验因素的干扰，从而产生试验误差。试验中出现的误差主要分为两类：随机误差（random error）与系统误差（systematic error）。由于许多无法控制的内在和外在的偶然因素（如作物种植过程中的栽培管理措施）所造成的误差称为随机误差。尽管在试验中力求一致但做不到绝对一致，因而随机误差具有必然性。虽然随机误差在试验中是不可避免的，但是还是可以减少的。试验误差越小，试验精确性越高。系统误差是一种有原因带有方向性的偏差。在试验过程中，要防止这种偏差，如农事操作和管理的不一致、测量仪器不准等。系统误差影响试验的准确性。

（六）精确度与准确度

科学试验的目的是揭示潜在的规律。因而要求科学试验要准确。用准确度（accuracy）来度量。它是指在试验或调查中，某一试验指标或性状的观测值与其真值接近的程度。设某一试验指标或性状的真值为 μ，观测值为 y。若 y 与 μ 相差的绝对值小，则观测值 y 的准确度高；反之，则低。然而，试验指标或性状真值常常未知。这时，可使用精确度（precision）。它是指在试验或调查中同一试验指标或性状的重复观测值彼此接近的程度。若观测值彼此接近，即观测值 y_i 和 y_j 相差的绝对值 $|y_i - y_j|$ 小，则其精确度高；反之，则低。显然，精确度高不一定准确度高。但是，准确度高则精确度一定高。

生物统计学依赖生物学和统计学两个学科。若将没有生物学联系的统计指标进行统计分析，即使再准确也毫无意义；反之，具有生物学意义的统计分析也会因试验精度低而导致错误的结论。因此，在科学研究做出结论后，还必须再回到实践中加以验证。

二、生物统计学的功用

现代生物统计学已在生命科学研究、生产实践和生物学教育领域得到了极为广泛的应用。其基本功用如下：

（一）提供整理和描述数据的科学方法

在统计调查和科学试验中，可以获得大量的统计数据。例如，2015 年上海市 16 岁男中学生的身高。若将所有观测值都罗列起来，往往庞杂零乱，无法清楚地揭示其潜在的规律性。生物统计学提供了整理统计数据、化繁为简的科学程序，以及由众多观测值归纳出几个能反映其特征的计算方法。通过生物统计学方法对这些统计数据的加工整理，对统计数据进行科学计算，使之条理化，更容易揭示其潜在的特征与规律性。

（二）提供了样本推断总体的科学方法

统计调查与科学试验的目的在于认识总体的特征与规律。然而，总体往往较大，有时还有破坏性，即使有限总体也难以获得其总体全体。因而，在统计调查和科学试验中，通过部分个体（样本）的调查或局部的科学试验来研究总体。这就产生了如何才能由样本科学地推断总体的问题，也就是如何由样本统计数推断总体参数的问题。生物统计学弄清了如何获取随机样本的方法，弄清了样本统计数分布与总体参数间的关系，弄清了由计算的统计假设概率大小推断统计假设是否成立的关系。因而，解决了由样本推断总体的科学问题。

（三）提供了误差分析以鉴定处理效应的科学方法

在研究肥料 A 和 B 的增产效果差异比较的农业试验中，除了肥料因素（处理）不同外，其余因素作为试验条件都要求保持一致，即体现唯一差异原则，以便精确地度量肥料 A 和 B 的产量差异。这里，肥料 A 和 B 的产量平均数差异称为试验的表面效应。实际上，两处理所受的试验条件不可能完全相同，总会有这样或那样的不一致而产生的偶然差异。这种偶然差异称为试验误差（experimental error）。因此，试验的表面效应究竟是处理的不同造成的呢？还是由试验误差造成的呢？换言之，究竟是处理间的本质差异，还是试验误差的偶然差异？这就需要将表面效应与试验误差进行比较。若表面效应显著大于试验误差，说明表面效应是由于处理不同造成的；反之，则由试验误差引起。生物统计学提供了误差分析及其与表面效应比较的科学方法，解决了鉴定处理效应的科学问题。

（四）提供了进行科学试验设计的一些重要原则

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统计数据，尽量降低试验误差，从试验所得的数据中能够无偏地估计处理效应和试验误差，以便从中得出正确的结论；相反，设计不周不仅不能得到正确的试验结果，而且还会带来经济上和其他方面的损失。

总之，生物统计学是一门很有用的工具课程，正确使用这一工具可以使生物科学研究更加有效，使生产效益更高，使教育效果更好。所以，它是每位生物科学研究工作者和生物学教育工作者必须掌握的基本工具。在大数据时代，掌握这种基本工具更加重要。

第三节 生物统计学的发展简史

统计学是随着社会生产水平的发展和适应国家管理的需要而发展起来的一门学科。它可追溯至远古时期。在我国古代，“方以类聚，物以群分”体现了统计分类思想，“称物平施”体现了平均数思想。但是，作为一门独立的学科，统计学至今不过才300多年的历史。从其发展过程来看，可分为古典统计学、近代统计学和现代统计学三个时期，形成国势学派、政治算术学派、社会统计学派和数理统计学派四大主要学派。生物统计学是数理统计学与生物科学、农学和医学形成的交叉学科。

生物统计学的发展是与随机误差的研究和误差控制相联系的。虽然随机误差的正态分布归功于C. F. Gauss（1777—1855），但是J. Kepler（1571—1630）、P. S. Laplace（1749—1827）、G. Galileo（1564—1642）和A. de Moivre（1667—1754）等科学家在1619—1837年间也做了不少工作。K. Pearson（1895）根据二项分布和超几何分布得到了包括正态、χ^2、t和F等分布的Pearson分布族。19世纪误差论的一个重要成果是在误差呈独立同正态分布$N(0, \sigma^2)$的情况下，残差平方和除以σ^2服从自由度（degree of freedom，df）为$n-1$的χ^2分布。Pearson（1899）为检验观测次数与理论次数间的符合程度提出了近似服从χ^2分布的χ^2统计量。分布参数估计方法的提出也是与误差理论的研究相关的：M. Legendre（1805）提出了最小二乘（least squares）法，R. A. Fisher在1912—1922年提出了极大似然（maximum likelihood）法。F. Galton（1886）提出了回归（regression）与相关（correlation），并经K. Pearson等一批学者的完善。这为20世纪上半叶统计方法的发展提供了契机，代表性成果是Fisher（1923）提出的方差分析（analysis of variance，ANOVA），还建立了试验设计的三大基本原则，并提出了随机区组和拉丁方设计。K. Pearson的学生Gosset（1908）用“Student”笔名将t分布发表在K. Pearson创立的《Biometrika》上，并阐明了样本标准差、样本平均数与标准差之比和相关系数的抽样分布，奠定了小样本理论基础。Neyman（1936）和E. S. Pearson（1938）建立了统计推断理论。A. Wald在1939—1950年建立了序贯分析和统计决策理论。W. G. Cochran和G. M. Cox系统地归纳了包括回归设计在内的试验设计和抽样方法的研究进展，田口玄一发展了三次设计，方开泰和王元提出了均匀设计。R. A. Fishes（1890—1962）等利用统计方法研究数量性状的多基因整体效应和作用方式。Sax（1923）、Lander和Botstein（1989）等提出了鉴别单个数量性状基因座的理论框架与统计方法。

在我国，王绶（1935）的《实用生物统计法》和范福仁（1941）的《田间试验技术》推动了我国农业生物统计和田间试验设计的应用。新中国成立后，特别是改革开放后，马育华和吴仲贤在作物科学和动物科学中推广和应用了生物统计和数量遗传学方法，推动了学科的发展。近30年来的代表成果有三倍体胚乳性状遗传分析方法、主基因+多基因混合遗传分离分析方法、数量性状连锁分析与关联分析的统计方法。随着大数据时代的到来，怎样从大数据中发掘潜在的规律性呢？这时，处理海量数据的统计方法就值得大家关注，例如，最小角回归（LARS）、经验Bayes估计和惩罚极大似然方法。未来的生物科学工作者必须掌握最新的统计方法与分析技术，才能更全面地揭示生物学中的科学问题。

第四节 生物统计学的主要内容

生物统计学大致包括如下四个主要内容。

一、描述统计学

描述统计学实际上就是对原始统计数据进行整理，并进行基本分析，其目的是揭示数据潜在规律。它主要包括数据整理和数据基本特征值的计算。数据基本特征值包括数据的集中性、离散性和偏斜性。集中性用平均数表示；离散性用极差、方差和标准差、变异系数和标准误表示；偏斜性用偏度和峰度表示。这部分内容将在第二章介绍。

二、推断统计学

（一）平均数间差异的假设检验

在农业和生物学试验中，经常遇到两个处理的比较。处理间差异显著性检验就是处理平均数间的假设检验。例如，在研究矮壮素使玉米矮化效果的试验中，矮壮素处理数天后，测定处理（甲组）和未处理（乙组，喷水）矮壮素的植株株高。若甲组的株高低于乙组的株高，那么这差异究竟是由矮壮素造成的，还是由其他偶然因素引起的呢？这就需要应用平均数差异显著性检验的统计分析方法，才能做出较可靠的判断，不至于被某些偶然性因素所蒙蔽。相关的理论与方法将在第三、四章介绍。

（二）方差分析

方差分析是多个平均数间差异显著性检验的统计分析方法。将总变异剖分为不同变异原因的变异，以分析不同原因的重要程度，并给出误差的无偏估计。若平均数间差异显著，还需要进行多重比较以获得两两平均数间的差异显著性。相关的理论与方法将在第五章介绍。

（三）次数资料的假设检验

生物学领域中有许多性状不能直接用测量的方法来加以衡量，一般称为属性性状。例如，花的颜色、性别的雌和雄、药物试验的治愈与无效。这些数据可通过对具有相同属性的计数来获得。若需要检验这种属性资料的实际次数与理论次数间的符合程度，可用 χ^2 检验方法。相关的理论与方法将在第六章介绍。

三、线性相关与线性回归

研究变量之间相互关系的密切程度，称为线性相关，以相关系数来表示。例如，人的身高与体重存在着一定程度的相关，一般身高越高，其平均体重越大。相关系数可用来表示两变量间的相关程度与性质。线性回归是指两个或两个以上的变量存在着从属关系，即一个变量（x）变化时，引起另一变量（y）的相应变化的估计。它们的从属关系可以用回归分析方法进行研究，根据实际数据建立的关系式称为回归方程式。本教材主要介绍了直线回归和部分可直线化的曲线回归，多项式回归和多元回归分析可选择讲解。这将在第七章介绍。

四、试验设计

试验设计是指如何选择试验单元，进行合理的分组和安排试验，其目的是为了尽量减少和控制试验误差，并对试验误差做出无偏估计。本教材第八章主要讲述的环境设计有完全随机设计、随机区组设计和拉丁方设计；处理设计有正交设计、回归设计和均匀设计。为了使试验结果成为有用而可靠的科学资料，在开始试验之前，认真地进行试验设计是非常必要的。