生物统计学

BIOSTATISTICS

谢邦昌 赵雅婷 邬宏潘 耿 直

中国统计出版社
China Statistics Press

生物统计学

BIOSTATISTICS

谢邦昌　赵雅婷　邬宏潘　耿　直

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

生物统计学/谢邦昌等编.
—北京:中国统计出版社,2003.6
ISBN 7-5037-4164-3

Ⅰ.生…
Ⅱ.谢…
Ⅲ.生物统计
Ⅳ.Q—332

中国版本图书馆 CIP 数据核字(2003)第 045384 号

生物统计学

作　　者/谢邦昌等编
责任编辑/吕　军
装帧设计/张建民
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 75 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/(010)63459084、63266600—22500(发行部)
印　　刷/科伦克三莱印务(北京)有限公司
经　　销/新华书店
开　　本/787×1092mm　1/18
字　　数/480 千字
印　　张/26
印　　数/1—3000 册
版　　别/2003 年 7 月第 1 版
版　　次/2003 年 7 月第 1 次印刷
书　　号/ISBN 7-5037-4164-3/Q·1
定　　价/52.00 元

前　言

生物统计学

试验的目的，是想从试验结果所得的数据，下正确的判断或做正确的预测，以解释各种现象。但这种被认为可供给信赖情报的试验结果，也常常因随机性而发生错误。因此如何利用这些具有变动性的数据下正确的判断及推测，就变成了重要的课题。尤其是以生物体为研究对象的各种试验，除因随机性而发生错误外，生物体本身的先天性差异，都易造成试验结果的差异。所以整理试验结果，并将所有可能的信息以最简单的方式表达以利于判断，也即如何应用统计分析方法整理试验结果，成为研究者不可缺少的知识。试验数据或结果的获得可分为二种情况，一是与试验数据取得的“次序”有关的场合，而另一种是与数据取得的次序无关，但与数据本身具有的特色有关的场合。前者与数据的发生时间有关，可用时间序列分析法分析数据；而后者则与试验数据的分布有关，重要的情报可由统计分析结果获得。

本文分四部分，第一部分记述各种统计分析方法，包括：数据的汇整，统计量的计算，各种分布，估计与假设检验的基本观念，平均与变异的假设检验，类别数据的卡方检验，简单线性回归分析，相关与多元回归分析等。第二部分为试验设计，包括：方差分析，实验的基本设计，裂区设计法与折叠分类设计法，直交表的应用与探求最适条件的试验设计法等。第三部分为特论，包括：统计因果推断，时间序列分析与预测，数据的掘取，非参数统计分析方法，抽样方法与抽样分布，官能检查、生物定量法、计量育种的统计分析法、医学上的离散数据分析以及生物信息的数据处理等。

生物统计学

序　言

统计学是以观察数量为主的科学。在数量化过程中使用各种数学方法进行记录、分类及标准化，形成各种方法论，产生了描述统计学与推断统计学。描述统计学是由 K. Pearson 倡导的，推断统计学是由 R. A. Fisher 倡导的，后者已成为今日统计学的主流。

十九世纪末，自然科学史上有三大成就，即演化论、能量不灭理论及细胞学的发展，这三项成就均对统计学的发展产生很大的影响。演化论促成描述统计学的创立，能量不灭理论由热力学的进展而创立，但热力学的各种法则是由气体运动论证明力学而形成的，后来演变成统计力学，导出近代概率论的方法。细胞学的发展由遗传学间接促成，孟德尔的遗传法则是现代统计理论的基础。

Galton 受 Darwin 的“物种起源”(Origin of Species，1859)的影响，对遗传现象发生兴趣，他引用了统计方法来解决演化问题，而当时热衷于弹性论研究的应用数学家 K. Pearson 则对 Galton 的优生学研究发生兴趣，从而创立了生物统计学，即生物统计学是以对 Darwin 的演化论和 Galton 的优生学等理论，提出数学方法为研究目的的学科。Darwin 的理论，为当时英国的作物栽培及畜牧生产提供了理论基础，他的研究方法是通过“观察”得到的；Pearson 从观察、描述等研究方法中找出数学方法，从而形成了描述统计学，其内容涵

盖 Pearson 分布曲线、回归系数、相关系数、重相关系数、偏相关系数与分位数等。描述统计学是以演化和遗传等生物现象为主，由 Galton 的优生学与 K. Pearson 的生物统计学的研究方法抽象化所形成的，描述统计学不但用于观察生物现象，同时也被应用于社会、经济与人口学等方面的探讨。

推断统计学的重要基础有两个，即圃场试验法（田间试验）和大量生产管理。100 年来的田间试验是由“技艺”进入“科学”的过程。1843 年成立于 London 北方的 Rothamsted 农业试验场在统计学的发展史上扮演了重要的角色。这个试验场面临的重要问题是无法把握地力的变化，试验结果常常发生很大的误差，因此他们订出了各种田间试验法，如 Beaven 的正方形试验区的棋盘法、半条播机法等。这些方法是由 W. S. Gusset（Student 氏）所提出，但最后提出革命性变化的是 R. A. Fisher，他将随机化和重复（randomization and replication）的观念引入田间试验，从而形成随机化试验排列，如随机化完全区组设计（RCBD）或拉丁方设计法（Latin Square Method）等，并使用方差分析（ANOVA）建立了新的田间试验法。

二十世纪初，工业界致力于产品规格的统一与产业经营的合理化，从而要求标准化，即将生产方式全面管制，因此产生了品质管理和抽样检查等统计方法。近代生产方式的特征之一是归还作业，即原定规格产品的重复生产。但实际上是无法完全依照原定规格重复生产的，因为在技术上无法达到目标。对这种偶然的变化的管理或变动结果是否达到统计上可被接受的判断变成推断统计学的任务。偶然的变动必须经过技术方面上的探明。大量生产方式依照标准化→生产→检查等方式进行，其理论上的特性即为假设的拟定→实验的执行→假设的检验等对应的科学上的认定。抽样检查，理论上属于估计理论或假设检验论，即品质管理为实现逐次检查或推断统计等提供适用的场合。

形成推测统计学的功劳者为 W. S. Gosset，而后由 R. A. Fisher 在 1920 年左右再发展而成，特别是估计理论、方差分析法、试验设计法等。伦敦大学的 E. S. Pearson 及 J. Neyman 等人用严密的数学公式，建立了假设检验方法。遗传学及优生学，不但与描述统计学的创立和发展有很大的关系，而且和推断统计学的重要试验也有密切关系。抽样法是统计调查中很重要的取样方法，这种方法也受到推断统计学的影响。品质管理也在工业界全面被接受，如 A. Wald 的逐步检验法是为满足这方面的需要而导出的方法。计量经济也被用于美国的财政管理，从而形成计量经济学与推断统计学的结合。其它，如舆论调查与教育统计等也在计量心理学及计量社会学上被重视，目前推断统计学在这些方面的发展也成为重要的课题。推测统计学的应用领域，如试

验设计或生产管理等从行动的观点上强调设计与管理的重要性，在设计上目前管理方法的发展逐渐受到推断统计学的影响。推断统计学在本质上为统计推理(测)，含估计理论(theory of estimation)与假设检验理论 (testing of statistical hypothesis)。

目前人类遗传，特别是基因组分析为最热门的课题，由于计算机使用上的方便，使族群遗传与分子遗传两学科紧密地连结在一起。基因组研究提供大量的 DNA 序列资料及染色体连锁地图(linkage mapping)，这些信息在理论上如何建立分析尚在研究中，要解读这些人类、植物及动物等巨大基因组资料，不但需要借用各种数学及统计方法，对其计算更加需要很多计算机软件的协助。新技术常常被诱导，特别是资料分析方法、计算机绘图、非线性型动态模型等，这些分析方法是针对细胞与分子生物学的需要而导出的，目前则进行类神经科学或类神经网络等的分析，这些生物学上的随机过程可应用于神经系统，其它如随机微分方程及图形判别等。现阶段常被应用到的理论工具有随机过程 ，如马尔可夫过程 (Marcov processes)及马尔可夫链(Marcov chains)等。随机过程也常使用于传染病及生态学上族群模式的建立。早期的族群生长模式常用决定式模型 (deterministic model)，但因为生与死的随机变动，使概率模型也被用于模型的建立，决定式模型可用微分方程得到解答，而概率模型则需要用到非线性型方程式，因此微分方程或偏微分方程有时并不能获得适当解答，尤其是在计算机设备经常无法普遍使用时更是困难。目前由于计算机等计算工具十分方便，这些已不成为问题。最近十年，AIDS 等传染病研究可以借用计算机设备来计算，但仍有待于适当的数学分析模型软件包的建立。尤其是微分方程式及微分差分方程式尚待开发。特别重要的是统计分析方法模型的配合与临界值的推断等技术尚待开发。流行性感冒、麻疹及虐疾等人类流行病的传染模式急待建构。农业上重要病害的传染模式，以往常用决定性模型建构其传染行为，但目前已在这方面改用空间分布的方式加以探讨，特别是在生态学、全球气候变迁、生物多样性、海洋学及气象资料等的探讨中，更加重视空间分布的应用。目前开始发展的地理信息系统及遥测技术等也受到生物统计学者重视。

实验设计的研究以往偏重于农业上的田间试验，但今天已逐渐转移到工业及医药问题的研究上并导出了各种不同的设计方法，如反应曲面的旋转设计(rotatable designs)，育种上常用到的全互交设计(diallel cross)，linear-plateau 及 plateau-linear-plateau 等探讨肥料效应、药理试验及药剂反应实验等。

目前由于分子生物学的研究进展迅速，产生了许多分子生物学的资料。对这些资料的分析可进一步了解遗传的本质，在数量遗传学上重视基因座的定位(QTL)，并由 DNA 核酸资料寻找基因讯息，探讨基因的核酸序列，氨基酸排列与蛋白质结构，并进一步探讨基因结构与基因作用，这些研究都需要借用统计方法，计算机模拟、比对(联配，alignment)计算及逻辑分析等，这些方法形成了生物信息学(Bioinformatics)，并从序列比对产生分子演化(Molecular Evolution)探讨生物的演化方向等。生物统计学的发展正逐渐改变方向与内容，随着生物学研究的开展产生新的方法与技巧，以解决新的生物学问题。

本书分三大部分共二十三章，第一部分为九章生物统计分析内容，介绍生物、农业及医学上常用的统计方法和基本观念，第二部分为实验设计法，介绍实验上常用的五种基本设计及其分析方法，第三部分为特论，介绍九种特殊的统计方法及推断，对分子生物、医学及农业上的特殊问题提供分析方法，协助研究人员能做适当的推论及客观的决论。本书大部分由谢邦昌教授撰写，其它三人负责部分的章节，编写时力求完善，但漏误之处在所难免，希望学者专家能随时修正，提供改进意见。本书编写时蒙辅仁大学统研所杨雅雯小姐大力协助编辑及打字，牺牲整个暑假完成本书，编者们甚为感谢。

编辑委员

2003 年 11 月

生物统计学

目　录

第一部分　生物统计分析

第二部分 实验设计

第三部分　特　　论