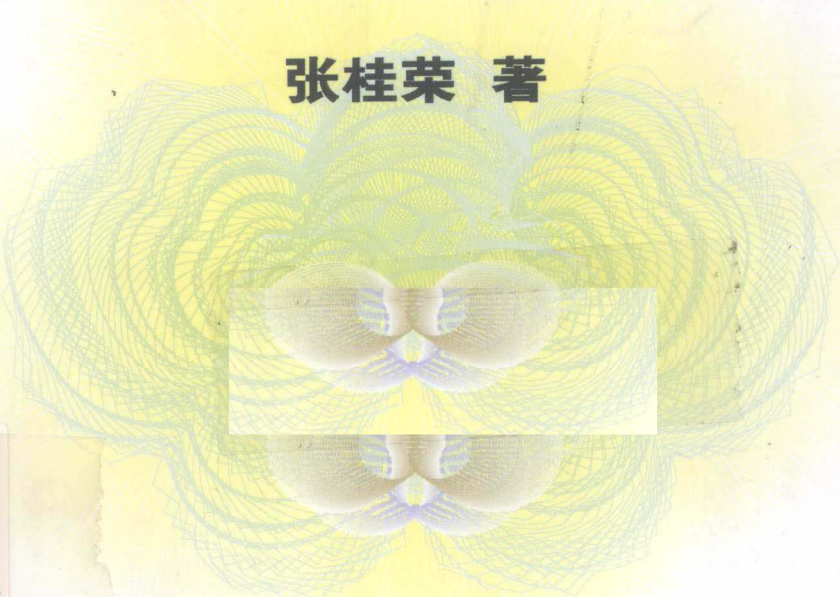
生物统计发展与应用

张桂荣 著



中国农业科学技术出版社

生物统计发展与应用

[redacted] 著

[redacted]

生物统计发展与应用

张桂荣 著

中国农业科学技术出版社

图书在版编目（CIP）数据

生物统计发展与应用/张桂荣著. —北京：中国农业科学技术出版社，2009.7
ISBN 978-7-80233-957-6

Ⅰ.生… Ⅱ.张… Ⅲ.生物统计 Ⅳ.Q-332

中国版本图书馆 CIP 数据核字（2009）第 115215 号

责任编辑 刘 建
责任校对 贾晓红

出 版 者 中国农业科学技术出版社
北京市中关村南大街 12 号 邮编：100081
电 话 (010) 82106638（编辑室）(010) 82109704（发行部）
(010) 82109703（读者服务部）
传 真 (010) 82109709
网 址 http://www.castp.cn
经 销 者 新华书店北京发行所
印 刷 者 北京富泰印刷有限责任公司
开 本 889 mm×1 194 mm 1/32
印 张 9.25
字 数 200 千字
版 次 2009 年 7 月第 1 版 2009 年 7 月第 1 次印刷
定 价 28.00 元

前　言

生物统计是运用数理统计学的原理与方法收集、整理、分析生物信息数据，解释生物现象，寻求其内在规律的科学。与一般数理统计学相比，它不仅要通过事物外在数量表现，去揭示事物可能存在的规律性，而且要根据专业知识去探讨、解释为什么会存在这种规律性；它不仅以抽象出来的随机变量为研究对象，而且必须紧密联系实际，使其具有更好的实践意义。生物统计方法对于认识世界、推动生物科学研究和农业生产不断发展具有重要作用，是进行生物类科学试验与研究的有力武器。生物统计的研究对象是带有随机性的生物有机体及其生态条件，其主要研究内容，一是如何运用科学的观察或试验的方式收集资料，目的在于取得样本；二是运用生物统计方法，通过对局部（样本）的认识去推断总体规律性，即统计推断问题，目的在于辨识信息，是人们认识客观事物的深化过程。

人类迈入21世纪，计算机技术突飞猛进，计算机的强大计算能力及其应用已渗透到社会的各个领域，各种统计分析方法层出不穷，并得到空前的发展，计算机技术与生物统计的结合势在必行。本书将传统生物统计与计算机统计软件相结合，介绍了利用计算机技术对常用生物统计方法进行运算分析，以代替繁重的人工计算分析，降低了工作难度，提高了工作效率，实现了现代计算技术与生物统计的联姻。

本书共有10个部分，第一部分主要介绍生物统计的发展情况和现代试验方法；第二部分主要介绍田间试验的意义与基本要求；第三部分主要介绍试验设计的原理、方法和试验的管理、总

结；第四部分主要介绍试验资料的整理与特征数及 SAS 在生物统计上的应用；第五部分主要介绍理论分布与抽样分布及利用 SAS 进行正态分布的拟合；第六部分主要介绍统计假设测验的基本原理、方法及 SAS 在非参数统计的应用；第七部分主要介绍方差分析的基本原理及 SAS 在方差分析中的应用；第八部分主要介绍卡平方（x^2）测验方法及 SAS 在适合性、独立性检验的应用；第九部分主要介绍直线回归和相关及 SAS 在直线回归和相关的应用；第十部分主要介绍试验结果的统计分析及 SAS 在随机区组试验设计结果分析中的应用。本书一方面介绍了国际上著名的数据分析软件 SAS 在生物统计上的应用，一方面注重严格的理论推导和方法步骤的详细介绍以及对各方法统计思想的阐述和对分析结果的解释，使读者既知其然，也知其所以然。

本书在编写过程中得到山东农业大学、菏泽学院等院校的有关专家教授给予的大力支持、帮助和指导，并提出了宝贵意见，在此一并表示感谢！限于作者的水平和对生物统计的理解，书中难免会有不妥和错误之处，恳请读者给予批评指正。

作 者

2008 年 12 月

目　录

1　生物统计的发展 …… (1)
1.1　试验设计与统计学发展 …… (1)
1.1.1　古典记录统计学阶段 …… (2)
1.1.2　近代描述统计学阶段 …… (3)
1.1.3　现代推断统计学阶段 …… (6)
1.2　农业和生物学领域的科学研究 …… (8)
1.3　科学研究的基本过程和方法 …… (9)
1.3.1　科学研究的基本过程 …… (9)
1.3.2　科学研究的基本方法 …… (10)
2　田间试验 …… (13)
2.1　田间试验的意义、任务及基本要求 …… (13)
2.1.1　田间试验的意义与任务 …… (13)
2.1.2　试验的要求 …… (14)
2.2　试验的种类和试验方案 …… (17)
2.2.1　基本概念 …… (17)
2.2.2　试验种类 …… (18)
2.3　试验误差及其控制 …… (21)
2.3.1　试验误差的概念及来源 …… (21)
2.3.2　试验误差的控制途径 …… (22)
3　试验设计与实施 …… (25)
3.1　试验设计的原则 …… (25)
3.1.1　重复（replication） …… (25)
3.1.2　随机排列（random assortment） …… (25)

3.1.3 局部控制（local control） …… (26)
3.2 试验的小区技术 …… (27)
3.2.1 试验小区的面积、形状和方向 …… (27)
3.2.2 重复设置及排列 …… (30)
3.2.3 对照区和保护区的设置 …… (31)
3.3 常用的田间试验设计 …… (32)
3.3.1 顺序排列的试验设计 …… (32)
3.3.2 随机排列的试验设计 …… (34)
3.4 温室与实验室的试验 …… (38)
3.5 试验的布置与管理 …… (38)
3.5.1 试验计划的制定 …… (39)
3.5.2 试验地的准备和区划 …… (40)
3.5.3 种子准备和播种或移栽 …… (41)
3.5.4 试验管理 …… (42)
3.5.5 收获与室内考种 …… (43)
3.6 试验的观察记载和测定 …… (44)
3.6.1 试验的观察记载 …… (44)
3.6.2 试验的抽样 …… (45)
3.7 试验总结 …… (47)
3.7.1 试验总结的主要内容 …… (47)
3.7.2 试验总结写作的特点和要求 …… (49)
4 试验资料的整理与特征数 …… (51)
4.1 常用统计术语 …… (51)
4.2 资料的整理 …… (52)
4.2.1 试验资料的性质与分类 …… (52)
4.2.2 资料的整理 …… (54)
4.3 资料的特征数 …… (63)
4.3.1 平均数 …… (63)

4.3.2　变异数 …… (66)
4.4　SAS 在生物统计上的应用 …… (73)
4.5　利用 SAS 描述样本数据 …… (74)
5　理论分布与抽样分布 …… (81)
5.1　事件、概率和随机变量 …… (81)
5.1.1　事件 …… (81)
5.1.2　事件的概率 …… (82)
5.1.3　事件间的关系 …… (83)
5.1.4　计算事件概率的法则 …… (85)
5.1.5　随机变量 …… (86)
5.2　二项分布 …… (88)
5.2.1　贝努利试验及其概率公式 …… (88)
5.2.2　二项分布的定义及其特点 …… (89)
5.2.3　二项分布的概率函数及计算 …… (91)
5.2.4　二项分布应用条件 …… (92)
5.2.5　二项分布的平均数与标准差 …… (92)
5.3　泊松分布 …… (94)
5.3.1　泊松分布的定义及特点 …… (94)
5.3.2　泊松分布的概率计算及应用条件 …… (95)
5.4　正态分布 …… (96)
5.4.1　正态分布的定义及其特征 …… (96)
5.4.2　正态分布的标准化 …… (99)
5.4.3　正态分布的概率计算 …… (99)
5.5　抽样分布 …… (103)
5.5.1　样本平均数的抽样分布 …… (103)
5.5.2　标准误 …… (106)
5.5.3　样本平均数差数分布 …… (108)
5.5.4　样本平均数差数标准误 …… (109)

5.5.5 t 分布 …… (109)
5.6 利用 SAS 进行正态分布的拟合与检验 …… (111)
6 统计假设测验 …… (113)
6.1 统计假设测验的基本原理 …… (113)
6.1.1 统计假设测验的概述 …… (113)
6.1.2 统计假设测验的基本方法 …… (114)
6.1.3 两尾测验与一尾测验 …… (119)
6.1.4 统计假设测验的两类错误 …… (120)
6.2 平均数的假设测验 …… (123)
6.2.1 单个样本平均数的假设测验 …… (123)
6.2.2 两个样本平均数相比较的假设测验 …… (125)
6.2.3 二项资料的百分数假设测验 …… (132)
6.3 参数的区间估计 …… (135)
6.3.1 总体平均数 μ 的区间估计 …… (135)
6.3.2 两总体平均数差数（$\mu_1 - \mu_2$）的区间估计 …… (138)
6.4 SAS 在非参数统计的应用 …… (140)
6.4.1 单个样本平均数假设测验（t 检验） …… (140)
6.4.2 成对资料假设测验（t 检验） …… (142)
7 方差分析 …… (145)
7.1 方差分析的基本原理 …… (145)
7.1.1 方差分析的意义 …… (145)
7.1.2 方差分析的步骤 …… (146)
7.2 单向分组资料的方差分析 …… (159)
7.2.1 组内观测值数目相等的单向分组资料的方差分析 …… (159)
7.2.2 组内观测值数目不等的单向分组资料的方差分析 …… (162)

7.3　两向分组资料的方差分析 …… (165)
7.3.1　两向分组资料无重复测值试验资料的方差分析 …… (165)
7.3.2　两向分组资料有重复观测值试验的方差分析 …… (170)
7.4　SAS 在方差分析中的应用 …… (185)
7.4.1　SAS 在单因素完全随机化设计方差分析的应用 …… (185)
7.4.2　SAS 在随机区组设计（两向分组）方差分析应用 …… (186)
8　卡平方（x^2）测验 …… (189)
8.1　卡平方（x^2）分布 …… (189)
8.1.1　卡平方（x^2）定义与分布 …… (189)
8.1.2　卡平方（x^2）分布的特点 …… (190)
8.1.3　卡平方（x^2）的连续性矫正 …… (191)
8.2　适合性测验 …… (191)
8.2.1　适合性测验的意义 …… (191)
8.2.2　适合性测验的方法 …… (192)
8.3　独立性测验 …… (195)
8.3.1　独立性测验的意义 …… (195)
8.3.2　独立性测验的方法 …… (196)
8.4　SAS 在独立性检验的应用 …… (203)
9　直线回归和相关 …… (205)
9.1　回归和相关的概念 …… (205)
9.1.1　直线回归和相关的概念 …… (205)
9.1.2　应用直线回归与相关分析时的注意事项 …… (208)
9.2　直线回归 …… (209)
9.2.1　直线回归方程 …… (209)

9.2.2 直线回归假设测验 …… (214)
9.3 直线相关 …… (216)
9.3.1 相关系数和决定系数 …… (216)
9.3.2 相关系数和决定系数计算 …… (219)
9.3.3 相关系数假设测验 …… (220)
9.4 应用 SAS 作相关分析 …… (222)
10 试验结果的统计分析 …… (225)
10.1 顺序排列设计试验结果统计分析 …… (225)
10.1.1 对比法试验结果统计方法 …… (225)
10.1.2 间比法试验结果统计方法 …… (228)
10.2 随机排列设计试验结果统计分析 …… (230)
10.2.1 单因素随机区组试验结果统计方法 …… (230)
10.2.2 复因素随机区组试验和统计方法 …… (236)
10.2.3 裂区试验结果统计方法 …… (243)
10.3 SAS 在随机区组试验设计结果分析中的应用 … (250)
10.3.1 单因素完全随机化设计应用 …… (250)
10.3.2 单因素随机区组设计应用 …… (251)
10.3.3 复因素随机区组设计应用 …… (252)
参考文献 …… (285)

1 生物统计的发展

生物统计是运用数理统计学的原理与方法收集、整理、分析数据，解释生物现象，寻求其内在规律的科学。与一般数理统计学相比，它不仅要通过事物外在数量表现，去揭示事物可能存在的规律性，而且要根据专业知识去探讨、解释为什么会存在这种规律性；它不仅以抽象出来的随机变量为研究对象，而且必须紧密联系实际，使其具有更好的实践意义。生物统计方法对于认识世界，发展科学和推动农业生产不断发展具有重要作用，是进行生物类科学试验的有力武器。生物学的研究对象是带有随机性的生物有机体及其生态条件，其主要研究内容，一是如何运用科学的观察或试验的方式收集资料，目的在于取得样本；二是运用生物统计方法，通过对局部（样本）的认识去推断总体规律性，即统计推断问题，目的在于辨识信息，是人们认识客观事物的深化过程。因此，我们可以这样说，生物统计的基本任务是如何运用观察或试验的方式取得样本，即试验设计与抽样技术；进而运用统计推断的方法由样本去推断总体，即统计分析方法。

1.1 试验设计与统计学发展

20 世纪以来，由于生物统计学的进展，使生物科学和农业科学逐渐成为可以用数学方法来处理和研究的科学。试验统计学作为一门系统的学科开始于 1925 年英国统计学家 R. A. Fisher 的著作 Statistical Methods for Research Workers，该书形成了试验统计学较为完整的体系。以后随着农业和生物学研究的发展，生物

统计、试验设计和抽样理论得到了快速的同步发展，并随着工业研究和数理科学研究的发展而进一步推动了应用数理统计学的发展，反过来又推动了试验统计学的不断发展。试验统计学的发展，大致可划分为3个阶段。

1.1.1 古典记录统计学阶段

古典记录统计学形成期间大致在17世纪中叶至19世纪中叶。统计学在这个兴起阶段，还是一门意义和范围不太明确的学问，在它用文字或数字如实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。17世纪Pascal和Fermat的概率论；18世纪De Moive、P. S. Laplace和Gauss的正态分布理论；最卓有成效地把古典概率论引进统计学的是法国天文学家、数学家、统计学家拉普拉斯（P. S. LapLace，1749～1827）。因此，后来比利时大统计学家凯特勒指出，统计学应从LapLace开始。

（1）拉普拉斯的主要贡献。

拉普拉斯的主要贡献：①发展了概率论的研究，LapLace第一篇关于概率论的表述发表于1774年。1812年发表的《概率分析理论》（先后出过4版）是他的代表作。LapLace最早系统地把数学分析方法运用到概率论研究中，建立了严密的概率数学理论；②推广了概率论在统计中的应用，主要表现在人口统计、观察误差理论和概率论对于天文问题的应用。1809～1813年，LapLace结合概率分布模型和中心极限思想来研究最小二乘法，首次为统计学中这项后来最常用的手段奠定了理论基础；③明确了统计学的大数法则，LapLace发现在观察天体运动现象时，当次数足够多时，能使个体的特征趋于消失，而呈现出某种同一现象。LapLace认为这其中一定存在着某些原因，而决非出于偶然；④进行了大样本推断的尝试，在统计发展史上，人口的推算

问题，多少年来成为统计学家耿耿于怀的难题。直到19世纪初，LapLace才用概率论的原理迈出了关键的一步。1781～1786年提出“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。在统计发展史上，他利用样本来推断总体的思想方法，开创了一条抽样调查的新路子。

（2）高斯的主要贡献。

另一位在概率论与统计学结合的研究上做出贡献的是德国大数学家高斯（C. F. Gauss，1777～1855）。他的主要贡献有：①建立最小二乘法。1795年，Gauss设想以残差平方和$\sum (Y_i - a - bx_i)^2$为最小的情况下，求得的a与b来估计α与β。1798年完成最小二乘法的整个构思与结构，正式发表于1809年。②发现高斯分布。Gauss以他丰富的数学实践经验，发现观察值x与真正值μ的误差变异大量服从现代人们最熟悉的正态分布。他运用极大似然法及其他数学方法，推导出测量误差的概率分布公式。“误差分布曲线”，这个术语就是Gauss提出来的，后人为了纪念他，称该分布曲线为高斯曲线，也就是今天的正态分布曲线。Gauss所发现的一般误差概率分布曲线以及据此来测定误差的方法，不仅在理论上，而且在应用上都有极为重要的意义。

1.1.2 近代描述统计学阶段

近代描述统计学形成期间大致在19世纪中叶至20世纪上半叶。由于生物学家们为了解决达尔文进化论中的复杂问题，经常需要借助统计学手段，而在这个过程中，原有的统计学方法的不足与局限性逐步地暴露出来。因此，许多学者在改善手段方面做了许多工作。19世纪达尔文应用统计方法研究生物界的连续性变异；孟德尔应用统计方法发现显性、分离、独立分配等遗传定律。由于这种“描述”特色由一批研究生物进化的学者们提炼而成，因此历史上称他们为生物统计学派。生物统计学派的创始

人是英国的高登（F·Galton，1822～1911），主要发展是由 Galton 的得意门生 K·泊松（K·Poisson，1857～1936）完成的。

（1）高登的主要贡献。

高登的主要贡献：①初创生物统计学。Galton 自 1882 年起开设“人体测量实验室”，在连续 6 年中，共测量了 9 337 人的“身高、体质量、阈度、呼吸力、效力和压力、手击的速率、听力、视力、色觉及个人的其他资料”，他深入钻研那些资料中隐藏着的内在联系，最终得出“祖先遗传法则”。在极其广泛地收集资料的同时，为了能使他的遗传理论建立在比较精确的基础上，出色地引入了中位数（Median）、百分位数（Persentile）、四分位数（Quartile）、四分位差（Quviation）以及分布、相关、回归等重要的统计学概念和方法。他在著作《Natural Inheritan》中首先提出了生物统计学（Biometry）一词，指出“所谓生物统计学，是应用于生物学中的现代统计方法”；②对统计学的贡献。变异是进化论中的重要概念，高登首先以统计方法加以处理，最终导致英国生物统计学派的创立。1889 年，Galton 把总体的定量测定法引入遗传研究中。通过总体测量发现，对动物或植物的每一个种别都可以决定一个平均类型。在一个种别中，所有个体都围绕着这个平均类型，并把它当做轴心向多方面变异。这就是他提出的“平均离差法则”。关于“相关”，统计相关法是由高登创造的。关于相关研究的起因，最早是他因度量甜豌豆的大小，觉察到子代在遗传后有“返于中亲”的现象。1877 年，他搜集大量人体身高数据后，计算分析高个子父母、矮个子父母以及一高一矮父母的后代各有多少个高个子和矮个子子女，从而把“父母高的后代高个子比较多，父母矮的后代高个子比较少”这一定性认识具体化为父母与子女之间在身高方面的定量关系，并提出了相关函数（即现在常用的相关系数）的计算公式；关于“回归”，1870

年，高登在研究人类身长的遗传时发现，高个子父母的子女，其身高有低于他们父母身高的趋势；相反，矮个子父母的子女，其身高却往往有高于他们父母身高的趋势。这就是统计学上“回归”的最初含义。1886 年，高登在论文“在遗传的身高中向中等身高的回归”中，正式提出了“回归”概念。

（2）K·泊松的主要贡献。

对生物统计学倾注心血，并把它上升到通用方法论高度的是 K·Poisson，他对统计学的主要贡献有：①变异数据的处理。生物统计中所取得的数据常常是零乱的，很难看出其规律。K·Poisson 首创的频数分布表与频数分布图成为统计方法中最基本的手段之一；②分布曲线的选配。19 世纪以前，人们认为以频数分布描述变异值，最终都表现为正态分布曲线。但是，K·Poisson 从生物统计资料的经验分布中，注意到许多生物上的度量不具有正态分布，而常常呈偏态分布，甚至倾斜度很大，而且也不一定都是单峰，也有非单峰的。1894 年，他在“关于不对称频率曲线的分解”一文中首先把非对称的观察曲线分解为几个正态曲线。利用所谓“相对斜率”的方法得到 12 种分布函数型，其中包括正态分布、矩形分布、J 型分布、U 型分布或菱型分布等；③卡方检验的提出。1900 年 K·Poisson 独立地又重新发现了 x^2 分布，并提出了有名的“卡方检验法”（test of x^2）。在自然现象的范围内，x^2 检验法运用得很广泛。以后经 R. A. Fisher 补充，成了小样本推断统计的早期方法之一；④回归与相关的发展。回归与相关，经 K·Poisson 进一步发展后，这两个出自于生物统计学领域的概念，便被推广为一般统计方法论的重要概念。此外，K·Poisson 还提出复相关、总相关、相关比等概念，不仅发展了 Galton 的相关理论，还为之建立了数学基础。