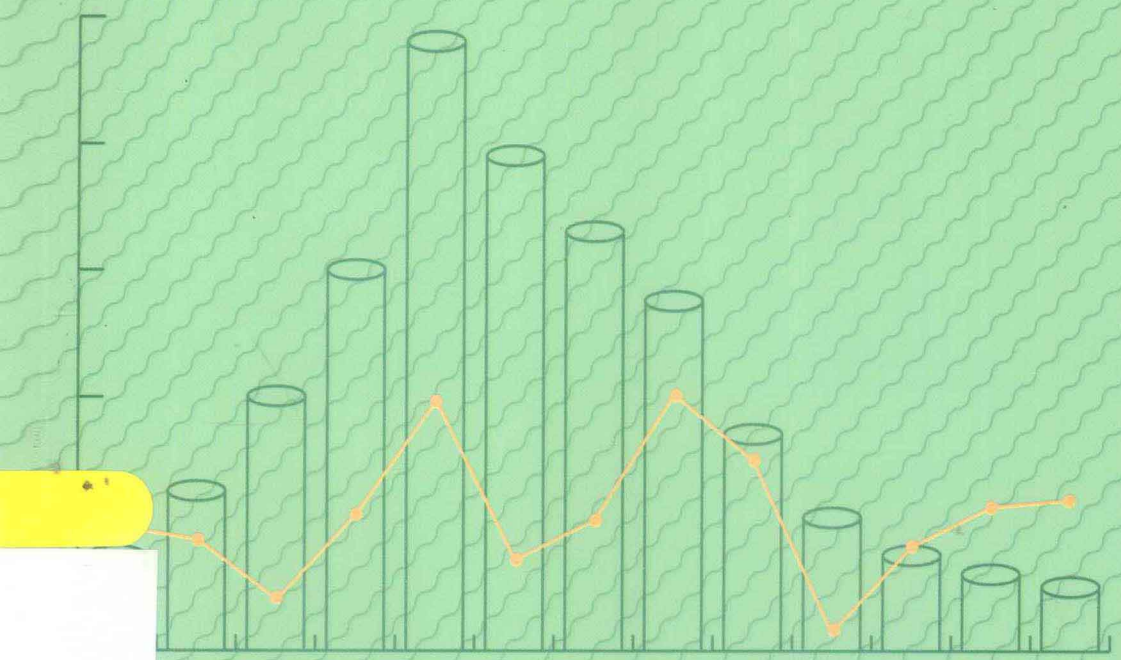
生物统计学

（第4版）

杜荣骞



高等教育出版社
HIGHER EDUCATION PRESS

BIOSTATISTICS

生物统计学

SHENGWU TONGJIXUE

（第4版）

杜荣骞

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目（CIP）数据

生物统计学／杜荣骞编.—— 4 版.——北京：高等教育出版社，2014.1

ISBN 978-7-04-038971-5

Ⅰ.①生… Ⅱ.①杜… 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（2013）第 298087 号

策划编辑 王 莉　　责任编辑 赵晓媛　　封面设计 张 楠
版式设计 于 婕　　责任印制 朱学忠

出版发行 高等教育出版社
社　　址 北京市西城区德外大街 4 号
邮政编码 100120
印　　刷 三河市骏杰印刷厂
开　　本 787mm×1092mm　1/16
印　　张 20
字　　数 510 千字
购书热线 010-58581118
咨询电话 400-810-0598

网　　址 http://www.hep.edu.cn
　　http://www.hep.com.cn
网上订购 http://www.landraco.com
　　http://www.landraco.com.cn

版　　次 1999 年 7 月第 1 版
　　2014 年 1 月第 4 版
印　　次 2014 年 1 月第 1 次印刷
定　　价 38.00 元

物 料 号 38971-00

作者简介

南开大学生命科学学院教授、博士生导师。1941 年生，1964 年毕业于南开大学生物学系，1964 年至 1973 年工作于中国科学院遗传研究所，其后回母校任教，2006 年退休。

先后为本科生、硕士生和博士生主讲过“生物统计学”、“普通遗传学”、“人类遗传学”、“数量及群体遗传学”、“群体遗传学与进化”、“分子遗传学技术”等课程。编写出版了《生物统计学》(1985 年版)、《生物统计学》(第 1 版至第 4 版)、《生物统计学题解及练习》，参编了《遗传学名词》(第 2 版)，遗传学试题库等。

早期以人类细胞遗传学研究为主。20 世纪 90 年代初赴加拿大 McGill 大学访问，回国后从事抗性遗传学研究。先后克隆了植物耐盐相关基因 *KD1*，培育出耐盐牧草“南港 A”等。其后从事昆虫抗菌肽(蛋白)的分离、纯化和抗菌肽基因的克隆等工作。从家蝇中纯化出 4 种新的抗菌肽，克隆了相关的 cDNA 序列。获国务院政府特殊津贴。亲自指导 20 余名博士和硕士研究生，已全部获得学位。除研究工作外，还积极推行产业化转化，已获得 13 项授权的发明专利。积极协助中小企业转型、创新，已获得初步成果。

父母亲对子女的为人要求严格，以“孝悌忠信礼义廉耻”作为处世做人的基本准则。在父母亲的言传身教下，严格要求自己：全心做事，认真做人。母亲生于清末的农村，没有接受过正规教育，自己却勤奋学习，积极进取。自学了西医学、英文、拉丁文，考取了执业执照。经常用韩愈《原毁》的名言：“彼，人也，予，人也；彼能是，而我乃不能是！”激励我们。以双亲为榜样，近 50 年来，工作兢兢业业，一丝不苟。面对学生，以身作则。获德育先进个人和天津市高等学校教学楷模称号等奖励。

数字课程（基础版）

生物统计学

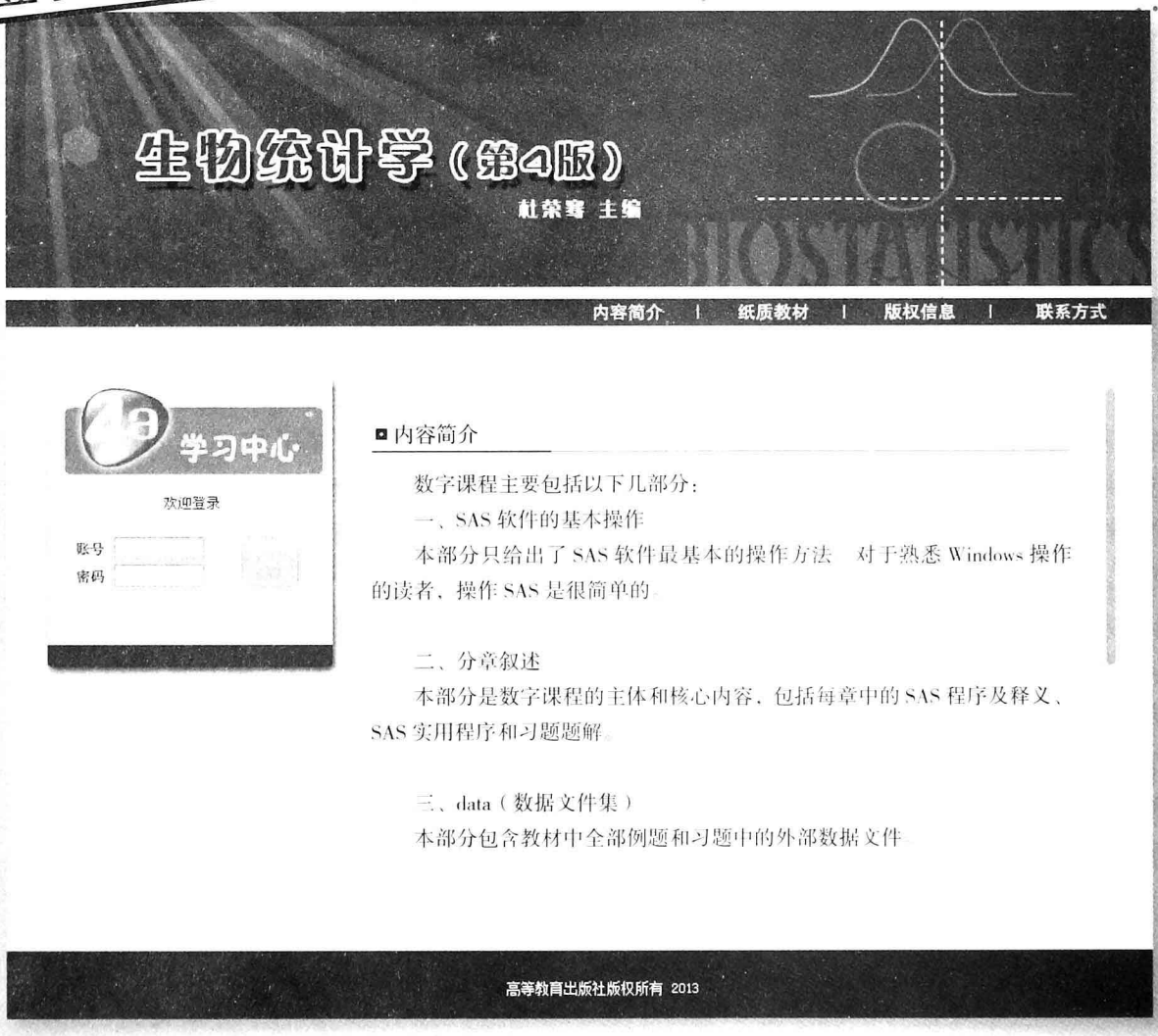
（第4版）

登录以获取更多学习资源！

登录方法：

1. 访问 http://res.hep.com.cn/38971
2. 输入数字课程账号（见封底明码）、密码
3. 点击“LOGIN”
4. 进入学习中心，选择课程

账号自登录之日起一年内有效，过期作废。
使用本账号如有任何问题，
请发邮件至：life@pub.hep.cn



http://res.hep.com.cn/38971

请在具有 IE 内核的浏览器下访问该网站。其他浏览器访问，可能造成课程资源无法正常显示

第 4 版前言

本教材自 1985 年出版以来，深得读者、教师和同学们的抬爱和热心指导，作者非常感谢。随着时代的发展、科技的进步，生物统计学在生命科学中的地位愈发重要，对本门课程的要求也不断提高。作者一直在思考怎样使这门课程更加适应时代的需要，不断提高读者在实验设计和数据处理方面的能力。这是一本面向本科生的教材，教材的知识面和深度在适应本科教学基本要求的基础上，略有提高，以便满足不同读者的需要。特别强调，该书只是一本启迪性的教材，起一个抛砖引玉的作用。在信息传播高度发达的今天，学生们可以通过多种途径获得知识，不仅仅限于教材，也不仅仅限于课堂学习。为适应互联网广泛传播知识的现状，本书在讲解各种统计方法时，不仅要使读者知其然，更重要的是要使读者知其所以然。希望读者在学习过程中多问自己几个为什么，才能够对统计学原理有真正的理解。只有在知其所以然之后，才能正确吸纳多方汲取的知识，并正确应用到科研、生产工作中去。为了能更好地学习这门课程，有以下几点建议：

1. 学习过程一定要循序渐进，要认真理解课程内容，再通过反复练习巩固所学知识。课程内容是连贯的，前面的任何疏忽，都有可能影响到后面的学习，前面的任何欠缺都要及时补上。教材的编写充分考虑到读者自学的需要，很容易读懂，但若不进行大量的练习以巩固所学知识，又很容易忘却或把概念搞混。所以，要读好这门课，不仅要“学”，更要不断地“习”，要“学而时习之”。

2. 一门新知识的学习一定要扎扎实实，不能浮躁，不要急于求成。现在有多种统计软件可以很容易地处理数据，千万不要认为有了统计软件就不必要学习统计学基础知识了。我曾经遇到过这样的同学，一心钻在实验室中做实验，最后得到大量数据，怎样处理这些数据？怎样从这些数据中得到可信的结论？到这时才想到生物统计学。随便找一个软件，把数据输进去，运行后得出一个结果，这就万事大吉，研究工作就算完成了。不去考虑实验是否符合实验设计的要求，也不考虑所采用的数据处理方法是否满足统计学要求。仓促把稿件投了出去，很快就被退了回来。如果单单是计算上的错误，还能挽回，如果是实验设计本身就有问题，几年的辛苦就付之东流了。如果将得到的错误结论应用到医学、农学和其他应用科学领域中，将会酿成大祸。

统计软件只是一个工具，帮助人们从繁重的计算工作中解放出来。它不能代替人们的思维，以及对知识和能力的应用。所以同学们在进入实验室之前，就应当认真设计实验、认真考虑对数据的处理方法。本教材在配套数字课程中提供了常用的 SAS 程序和 SAS Enterprise Guide 使用简介，目的是为了增长读者的知识面，为以后的工作提供方便。在学习阶段，我主张还是要用纸和笔，一丝不苟地去计算，只有这样学习的知识才扎实，才能帮助深入理解统计学的基本思想，提高自己设计实验和处理数据的能力和技巧。现代社会，知识的更新速度非常之快，没有良好的独立思考、独立学习能力，知识很快就会老化。读大学不单单是学点知识，更重要的是利用大学的一切可资利用的资源，通过多种途径充实自己，建立优良的思维方法，提高自己各方面的能力，只有这样才能跟上时代的要求，永远走在时代的前列。

3. 本次修订中，对一些同学们极易出现的错误和疏忽做了特别的强调。例如，在参数统计中，统计假设检验的一个很重要的前提是，未知总体应当服从或近似服从正态分布，所以首先要做正态性检验。作者在阅读文献时，发现不少论文作者在处理数据之前并未做正态性检验。或虽然没有给出检验过程，但强调已经做过正态性检验，抽出样本的总体服从或近似服从正态分布。如果这一条件得不到满足，一切参数统计检验工作都是徒劳的，甚至会产生错误的结果。为此，在本版中特别强调正态性检验，以便引起读者的高度重视。

4. 非参数统计是应用统计学的重要内容。在参数统计中，若正态性不能得到满足时，经常要使用非参数统计的方法进行推断。本书第三版第94页上有这样一段话“如果采用多种变换方法都不能得到较好的正态性，这时就需要用非参数方法进行统计推断。非参数统计已超出本教材的内容，请参考有关书籍。”作者经过再三考虑，决定在第四版中添加四种最常用的非参数统计方法。有了这部分内容，对一些不服从正态分布的问题，给出了具体的解决途径。同时也提示读者，非参数统计是另一类重要的统计分析方法。

5. 前面已经说过：本教材是为本科生编写的，但知识面的宽度和深度略有提高。内容稍微宽泛一些，可以给教师和学生较大的选择自由。任课教师可以根据本专业的特点、教授对象和学时安排选择适合于教学需要的相关内容讲授。学生可以根据自己的能力和爱好，除课堂内容外还可以选读一些课堂外的知识。本教材没有配套的教学课件，作者虽曾考虑过，但最后还是决定不编写，最好由使用本教材的教师自己编写。一方面可以最大程度发挥教师的主观能动性，课程可以讲解得更生动，更符合客观需要，有更多的自主权。我不希望看到各个学校都使用统一的教材、统一的课件，把学生都培养成工业流水线上的产品，这种做法不利于我国高等教育事业的发展。课程的学习应体现出不同学校、不同学生的不同特色，在夯实基础知识的前提下，给教师和学生较大的宽泛度，以充分发挥教和学的主观能动性。

本书在一些细节方面做了进一步修订，叙述更加严谨，更易理解。数字课程部分的安排做了重新调整、补充和修订，方便读者查找和对比阅读。对应内容在书中用ⓔ标出。该部分还添加了SAS Enterprise Guide使用的基础知识，为读者开阔眼界，提供更多解决统计学问题的途径。

作者虽然已经退休，我院刘方院长一如既往地支持本教材的修订工作，为了使本版的修订工作做得更好，特意购买了基于SAS 9.3版本的“SAS学院版基础数据分析工具包”，作者在此深表谢意。特别要感谢我校孙丹老师在修订过程中提供的多方面帮助，以及为本书顺利完成修订工作所做出的贡献。感谢妻子承担起繁琐的家务，为顺利完成本书修订工作所创造的条件。感谢读者多年来对本书的厚爱和指导，读者的关爱是修订好本书的动力。

作　者

2013年7月17日

rongqian@nankai.edu.cn

第 3 版前言

本书是为生命科学学科本科生编写的教材，编写的原则是夯实基础，注重对基础知识和基本概念的理解与应用。书中所选编的内容与深度与本科生的教学是相适应的。在第 3 版中对教学内容没有做更多的补充，没有增加更多的教学工作量，只是根据在教学过程中出现的问题，从以下几方面进行了修订。

1. 本书的第 1 章到第 9 章及第 12 章涉及的都只是因变量（响应变量）的问题，并未讨论自变量与因变量之间的关系。只有第 10 章和第 11 章涉及自变量。但是在因变量的命名上却出现一些混乱，在第 10 章和第 11 章以变量 X 作为自变量，变量 Y 作为因变量，而其余各章均以变量 X 作为因变量，这样命名容易给读者造成混乱。产生这种现象的原因，主要是沿袭了我国应用统计学的命名习惯。这次修订对此做了彻底调整，全书均以变量 Y 作为因变量。因此，有统计学基础的一些读者可能会不太习惯，但是它的科学性更强。

2. 在教学过程中发现，同学们对教材中的一些内容提出的问题比较多。如为什么将原始数据 ±0.5 就是连续性矫正？在用拟合优度检验分布的正态性时，编码变量是怎么确定的？在检验中有什么作用？对于初学者，怎样直观地理解方差分析？这样的一些问题在这次修订中都做了详细的解释，便于读者自学。

3. 本书的第 4 章“抽样分布”是统计假设检验的基础，是很重要的一部分内容。由于罗列了大量的数理统计学的结论，内容比较枯燥，学生在学习和理解上比较困难。这次修订对这些枯燥的内容通过“Monte Carlo”方法将其具体化，应用 SAS 程序进行模拟抽样，将枯燥的公式变成一幅幅动态的直方图。这样做使读者能直观地理解从不同总体中抽取的样本及其样本统计量的分布规律。只有对抽样分布有了深入的理解，才能准确地进行统计假设检验，才不会由于检验的条件不满足而出现统计学错误。

正态性是假设检验的基础，为了正确判断未知总体的正态性，这次修订给出了判断正态性的实用方法及其 SAS 程序。

4. 作者查阅了 1 000 多篇中英文文献，筛选出 100 余篇，编写成练习题，几乎更新了第 2 版中的全部习题。这些习题都是从实践中得到的，同学们通过对这些习题的演算，了解生物统计学是如何解决科研和实践中的具体问题，尽早地接触科研和实践工作。同时也可以帮助读者进一步了解生物统计学可以解决哪些科研和实践中的问题，扩大知识面，提高解决实际问题的能力。

5. 生物统计学是一门实用科学，要求学生除掌握生物统计学的基本原理外，还必须有很强的计算能力。希望学生在学习过程中，对每一条公式，每一种方法，用纸和笔借助计算器进行认真的计算，以便理解公式、熟悉公式、记忆公式。

然而在计算技术发达的今天，必须学会用计算机处理数据。考虑到读者使用的方便，本书附学习卡一张，通过学习卡可访问《生物统计学》（第 3 版）配套网站 http://res.hep.com.cn/bios3 或 http://res.hep.edu.cn/bios3。该网站介绍了国际通用的 SAS 统计软件包的基本操作方法，提

供本书全部例题和习题的外部数据文件,编写了全部例题的SAS程序(共91个基本程序)及部分习题的程序,并结合教材的内容做了深入的解释和说明。同学们可以在自己的计算机上演练,为将来在科研和生产实践中的应用打下初步基础。网站还给出了用SAS程序处理的主教材全部习题的题解,包括SAS程序、输出结果以及根据输出结果得到的结论。同学们在做练习题时,除手算外,还可以尝试用SAS程序进行计算。

在修订本书的过程中,本人得到妻子王方慎的全力支持。她承担了繁琐的家务,创造了良好的工作环境,并对修订的内容提出了很多珍贵建议和意见,在此深表谢忱!

本书的编写得到了“南开大学教材资助立项项目”的资助,在此表示感谢。

最后,衷心感谢参考文献的作者提供了大量的原始实验数据。

作者的E-mail:rongqian@nankai.edu.cn。

作 者

2008年6月23日

第 2 版前言

本书自 1999 年出版以来，深受广大读者欢迎，被多所高校采用作为教材。读者的喜爱便是对编者的激励，为了能更好地为生物统计学教学服务。经多方征求读者意见后，在本书第 2 版的内容编版上做了比较大的调整。

第 2 版删去了第 1 版中“附录：SAS 软件基本操作”部分和各章中给出的 SAS 程序。增加了第 12 章“实验设计”。统计分析与实验设计是密不可分的，只知道统计分析方法，而不知道如何设计符合统计学要求的实验，这样的知识是不全面的。为了提高学生独立分析问题、独立设计实验和独立处理实验结果的能力，我们认为增加实验设计是十分必要的。

对第 1 版中“SAS 软件基本操作”和相关的 SAS 程序做了调整和补充，连同每一章的习题详解及各章的大量复习题另行成册，称为《生物统计学题解及练习》，供读者理解和巩固所学知识以及学习如何用 SAS 软件处理数据。

生物统计学的内容很广泛，根据对本科生的要求和学时数的安排并征求了多方意见，确定了本书所选内容。讲完全书需要 50～60 学时。如果学时安排较少，可以适当减少两因素和多因素方差分析以及多元回归内容。我们建设，尽量保持生物统计学基本原理和统计假设检验内容的完整性，在此基础上，学生通过自学便能很快掌握更多的统计学知识。

希望广大读者在使用本书过程中，对所发现的问题及不足之处能不吝赐教，并希望提出进一步的修改意见，使本书在生物统计学教学中发挥更大作用。对此，编者将致以深切谢意。编者的电子信箱地址为：our gene@ eyou. com。

杜荣骞

2003 年 3 月

第 1 版前言

生物统计学是现代生物学研究不可缺少的工具。不论是传统学科还是现代分子生物学，时时刻刻都在与数字打交道。为了揭示生物体内在规律或生物与环境之间的关系，都离不开因素分析，特别是多元分析。生物统计学不仅在传统生物学、医学和农学中被广泛应用，而且在新兴的分子生物学研究中也发挥着重要作用。例如，绘制连锁图，特别是绘制人类基因连锁图时，制图函数的获得，Lod Score 的计算以及 DNA 序列同源性分析等都是建立在统计学基础上的。没有良好的统计学基础，这些工作只能知其然，而不能知其所以然，对于工作的深入开展是很不利的。因此，生物统计学已经成为每一位生物科学工作者必备的基础。

这本教材是在 1985 年版本的基础上，广泛征求各方面意见重新编写的。为配合生物学的迅速发展，在内容和编排上做了适当调整，删除了一些不常用的内容，增加了一些必要的基础内容，如方差分析中均方期望的推演等。近十几年来电脑在我国的迅速普及，出现了大量的统计软件。许多过去望而却步的繁重计算工作，现在已变得轻而易举。利用统计软件代替繁重的手工计算，是生物统计学发展的必然趋势。SAS 是国际上公认的统计软件，它的包容量大、伸缩性强，在全球范围内被各行各业广泛采用，因此，本书编写了介绍 SAS 软件应用的章节，以满足读者的需要。书内的例题和习题除一部分是编者自己的工作外，很多是从书后所列参考资料中引用的，在这里对原著者深表谢意。为了使例题更具代表性，对其中有些数据做了适当调整，因此，书中例题和习题中的数据只供学习和巩固统计学知识使用，没有真正的科学意义，请广大读者切勿引用。

本书在编写过程中得到了各方面大力支持，四川大学刘天伦先生在内容编排上提出过宝贵建议，本校数学系沈世镒先生，计算机系涂奉生先生曾鼎力相助，生命科学学院王颖老师在资料整理和誊写上做了大量工作，在这里对以上各位先生表示诚挚谢意。

在这里需要特别提出的是，美国 SAS 软件研究所上海办事处为本书的编写提供了 SAS 软件和多方支援，为促成本书起了很大作用。编者在这里对上海办事处的关心和支持表示衷心感谢。

编者在编写时虽已尽心竭力，但错误及不当之处仍在所难免，敬希读者不吝指出，本人将不胜感谢。

编　者

于南开大学生命科学学院

1998 年 12 月

目　录

第 1 章

统计数据的收集与整理

§1.1 总体与样本

1.1.1 统计数据的不齐性

人类在生活、生产和科学研究中经常与数据打交道。在对特定的研究对象进行测量、记录并分析所得数据之后，你会发现，即使从同一类对象中所得到的数据也不完全相同，有大有小、参差不齐。或者说，产生这些数据的个体间存在着广泛变异。

造成生物体变异的原因有很多，概括起来可以分为遗传因素、环境因素及发育噪声(development noise)。遗传因素的影响是显而易见的。就拿身高来说，子女身高直接受父母身高的影响，通常是父母高，子女也高；父母矮，子女也矮。环境因素表现在很多方面。仍以身高为例，包括：食量、蛋白质摄入量、营养成分平衡、维生素和微量元素的获得量、锻炼、劳动强度、睡眠时间、不良嗜好、修养、心理承受力等。我们会发现，即使在遗传与环境因素都得到控制的情况下，个体间仍然存在变异。例如，小麦纯系是经过多代自交得到的，遗传上已经纯合化，个体间遗传成分可以认为是均一的。将自交系的单株后代种植在生长条件都相同的环境中，例如，种植在人工气候室中，使用电脑控制肥力、水分、光照、温度、通风等，即使这样，个体间仍存在变异。它们的株高、穗长、穗重、干物重等还会有一定的波动。这种波动的产生是由发育噪声引起的。或者说是由于在个体发育过程中的某些随机因素造成的。如果把影响生物变异的各种遗传因素、形形色色的环境因素以及种种随机因素自由组合起来，其组合数将是一个天文数字。不同个体的组合方式不同，由此造成了生物个体之间的广泛变异。由此可见，变异性是自然界存在的客观规律。

由于个体间的变异，给我们处理数据带来很多困难。例如，考察我国 18 岁男青年身高，若个体间没有变异，我们随便测量一个人就可以了。然而，由于个体间存在着变异，为了测得 18 岁男青年身高，从理论上讲，应当把全国所有 18 岁男青年身高都测量一遍，用其平均数来代表身高数值，但这是很难做到的。退一步讲，虽然很难做到，但只要投入足够的人力和财力，还是可以测量出这些数据的。如果要测量所有新生儿体重，则无论如何也拿不到全部数据。因为新生儿不断出生，要想收集到所有新生儿体重，就要不断测量，只要有新生儿出生，测量就不能停止。由此可见，测量全部对象既不现实也不可能。我们只能从全部研究对象中抽出一部分个体来，通过对这一部分个体的研究来推断全体的情况。这就出现了我们下面将要提出的两个概念：总体与样本。

1.1.2　总体与样本

统计学研究的核心问题是如何通过样本推断总体。因此，总体与样本是生物统计学中的两个最基本概念。

总体(population)是我们研究的全部对象。总体又分为**无限总体**(infinite population)和**有限总体**(finite population)。例如，我们要研究在某种条件下生长的小麦的株高，因为无法估计出在这种条件下生长的小麦的数量，可以设想这一总体是无限的。或者研究新生儿体重，因为新生儿是不断增加的，所以这一总体也可以设想是无限的。如果我们要调查一所学校今年新生的身高，这一总体则是有限的。生物统计学中所遇到的总体多数都是无限总体。构成总体的每个成员称为**个体**(individual)。

样本(sample)是总体的一部分，样本内包含的个体数目称为**样本含量**(sample size)。

1.1.3　抽样

从总体中获得样本的过程称为**抽样**(sampling)。抽样的目的是希望通过对样本的研究推断其总体。例如，希望由100株“三尺三”高粱的株高，推断在这种条件下生长的该品种的株高。这就要求样本应能在最大程度上代表总体的情况。为此，在从总体中抽取样本时，总体中的每一个个体被抽中的机会必须都一样，不能带有偏见。例如，在小麦育种工作中，我们常常希望得到矮秆品种。为了满足个人愿望，在抽样时便多抽矮秆的，这样得到的样本没有代表性，属于偏性抽样，不能代表总体的情况。我们需要的样本应该是一个总体的缩影。为了达到这个目的，就需要用**随机抽样**(random sampling)的方法获得样本。

随机抽样的方法很多，例如抽签、抓阄等。最好的方法是使用随机数字表(见ⓔ附表1)进行抽样。现举例说明怎样用随机数字表进行抽样。假设需要从包含4 728个个体的总体中，抽出一个含量为20的样本。因为个体总数4 728是一个4位数，所以总体中每一个个体的编号都应是4位数，即从0 001号到4 728号。第Ⅰ步，闭上眼睛用铅笔在随机数字表上任意点上一点，假若点到奇数上，就用第一页表；点到偶数上，就用第二页表。第Ⅱ步，在选定的那一页上，再点一次，决定从哪个字开始。决定了起点以后，开始以四位数字为一节连续读下去，不用考虑数字间的间隙。可以正读、倒读、横向读、纵向读，也可以沿对角线方向读。选出小于等于4 728的数字，大于4 728的则舍弃，直到取满20个数为止。这20个数所对应的个体，即为我们选中的样本。更方便的方法是用随机数函数产生所需要的随机数。

从一有限总体中抽样，可分为**放回式抽样**(sampling with replacement)和**非放回式抽样**(sampling without replacement)。所谓放回式抽样是指：从总体中抽出一个个体，记下它的特征后，放回总体中，再做第二次抽样。这种抽样方式可能会重复抽中某一个体。非放回式抽样是指：从总体中抽出个体后，不再放回。在上述的例子中，若保留重复的随机数字，则为放回式抽样；若舍弃重复的数字，则为非放回式抽样。对于无限总体来说，放回式抽样和非放回式抽样，实际上没有区别。

样本的含量越大越有代表性。但是，太大的样本研究起来是很困难的。因此，样本的含量必须合适。

§1.2　数据类型及频数(率)分布

1.2.1　连续型数据和离散型数据

统计学的最基本工作是收集数据。把原始数据收集上来之后,首先要对数据进行整理并分析这些数据的特性和变化规律。生物统计学中经常遇到的数据有两种类型,一种是连续型数据,另一种是离散型数据。

与某种标准做比较所得到的数据称为**连续型数据**(continuous data),又称为**度量数据**(measurement data)。例如,长度、时间、质量、OD 值、血压值等。这类数据通常是非整数。虽然有时记载的是整数,如身高的厘米数,但是当提高精确度后,总会出现小数。对连续型数据进行分析的方法,通常称为**变量的方法**(method of variable)。

由记录不同类别个体的数目所得到的数据,称为**离散型数据**(discrete data),又称为**计数数据**(count data)。例如,某一类别动物的头数、具有某一特征的种子粒数、血液中不同类型的细胞数目等。所有这些数据全都是整数,而且不能再细分,也不能进一步提高它们的精确度。对离散型数据进行分析的方法,通常称为**属性的方法**(method of attribute)。

在判断数据的类型之后,就要进一步研究数据的变化规律。描述数据变化规律的最简单方法是将这些数据列成**频数表**(frequency table)或绘成**频数图**(frequency graph),根据频数分布进行研究。

1.2.2　频数(率)表和频数(率)图的编绘

离散型数据及连续型数据的频数表和频数图的编绘方法略有不同,下面各举一例说明。先看离散型数据频数(率)表和频数(率)图的编绘方法。

例 1.1　调查每天出生的 10 名新生儿中,体重超过 3 kg 的人数,共调查 120 天。每天的 10 名新生儿中,体重超过 3 kg 的人数,可能有 11 种情况:1 名也没有,有 1 名,有 2 名……10 名都是,如表 1-1 的第一列所示。这一列称为组值(class value)。表 1-1 的第二列所记载的是调查结果。

表 1-1　每 10 名新生儿中体重超过 3 kg 的人数的频数(率)表

组值 (体重超过 3 kg 的人数)	频数计算	频　数	频　率
0		0	0.000
1		0	0.000
2		0	0.000
3	一	1	0.008
4	丅	2	0.017
5	正丅	12	0.100
6	正正丅	19	0.158
7	正正正正正正正丅	39	0.325

续表

组值 （体重超过3 kg的人数）	频数计算	频　数	频　率
8	正正正正正正ㄈ	34	0.283
9	正正	10	0.083
10	下	3	0.025
总计		120	0.999

如第一天调查的结果，有6名超过3 kg的，则在组值为6的一行做个记号，一般使用“正”字或“卌”号表示。全部调查完毕，累加各行结果，填入频数一栏，或者将各行的结果除以总数而得出频率。所谓频率，即将某一类别的数目除以总数所得到的分数。把频数或频率按超过3 kg的人数的顺序排列起来，便得到了**频数分布**（frequency distribution）或**百分率分布**（percentage distribution）。频数表可以比较清楚地描述出数据变化规律。为了更直观地描述数据变化规律，还可以绘成频数图表示（图1－1）。图1－1称为**柱形图**（column diagram），它的横轴表示每10名新生儿中，体重超过3 kg的人数，纵轴表示每一组的频数。若将纵轴改为频率的话，则得到频率图。频率图与频数图的图形完全一样。

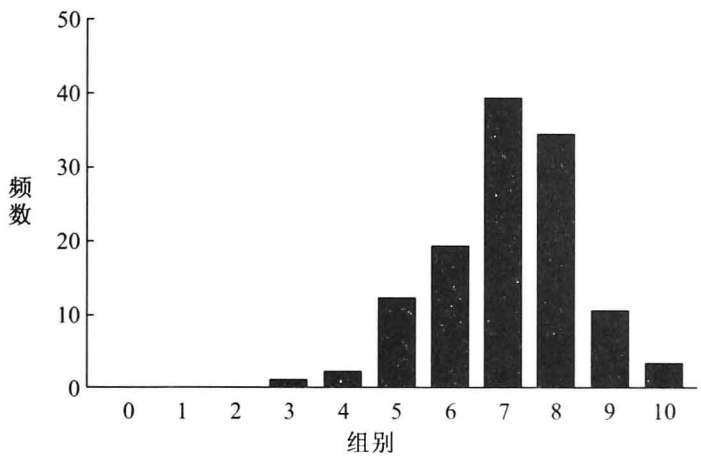


图1－1　频数图

下面这个例子介绍了连续型数据频数（率）表和频数（率）图的编绘方法。

例1.2　表1－2给出的数据是250粒黑莲子籽粒重的测量结果*。

表1－2　黑莲子单粒重/g

1.02	1.03	0.81	1.15	0.71	1.25	1.12	1.15	1.17	1.18	1.08	1.02	1.10
1.01	1.06	0.78	1.25	0.80	1.09	1.03	0.98	1.03	1.01	1.07	1.35	0.85
1.21	1.06	0.80	0.95	0.95	1.45	1.31	1.30	1.01	1.02	1.00	0.75	1.01
1.19	1.13	1.21	1.38	1.20	1.22	0.88	1.33	0.97	1.03	0.94	0.69	0.88
0.88	1.08	1.23	0.93	0.92	0.99	1.21	0.83	1.10	0.95	0.94	1.13	1.18
0.96	1.14	0.87	1.00	0.61	1.28	0.90	0.81	1.26	0.74	0.96	0.94	0.61