生物统计学

（第3版）

杜荣骞

高等教育出版社
Higher Education Press

生物统计学

[illegible]

[illegible]

生物统计学

（第3版）

杜荣骞

高等教育出版社
Higher Education Press

图书在版编目(CIP)数据

生物统计学/杜荣骞. —3版. —北京:高等教育出版社,2009.6

ISBN 978-7-04-025745-8

Ⅰ.生… Ⅱ.杜… Ⅲ.生物统计—高等学校—教材 Ⅳ.Q-332

中国版本图书馆CIP数据核字(2009)第013689号

策划编辑 王 莉 **责任编辑** 张晓晶 **特约编辑** 卢 琛 **封面设计** 张 楠

版式设计 范晓红 **责任校对** 金 辉 **责任印制** 韩 刚

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

总 机 010-58581000

经 销 蓝色畅想图书发行有限公司

印 刷 北京民族印务有限责任公司

开 本 787×1092 1/16

印 张 23.5

字 数 590 000

购书热线 010-58581118

免费咨询 800-810-0598

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

畅想教育 http://www.widedu.com

版 次 1999年7月第1版
2009年6月第3版

印 次 2009年6月第1次印刷

定 价 36.80元

物料号 25745-00

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 版的内容编排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。我们建议，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

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，随后回母校任教。

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 版到第 3 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。近年来从事昆虫抗菌肽(蛋白)的分离、纯化和抗菌肽基因的克隆等工作，纯化出 4 种新的抗菌肽，克隆了相关的 DNA 序列。获国务院政府特殊津贴。除研究工作外，积极推行产业化转化，获得十余项授权的发明专利。亲自指导 20 余名博士和硕士研究生，至 2006 年已全部获得学位，大部分在国外从事研究工作。

受父母亲的影响，尊崇儒家思想。常以“古之君子，其责己也重以周，其待人也轻以约”严格要求自己。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“彼，人也，予，人也；彼能是，而我乃不能是！”激励子女。以母亲为榜样，40 多年来，工作兢兢业业，一丝不苟；面对学生，为人师表，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。

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。把所有 18 岁男青年身高都测量一遍是很难做到的。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 千克的人数,共调查 120 天。每天的 10 名新生儿中,体重超过 3 千克的人数,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 千克的人数)	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	正正丅	12	0.100
6	正正正𠂇	19	0.158
7	正正正正正正正𠂇	39	0.325

续表

组值 （体重超过3千克的人数）	频数计算	频　数	频　率
8	正正正正正正下	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3千克的，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3千克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3千克的人数，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。

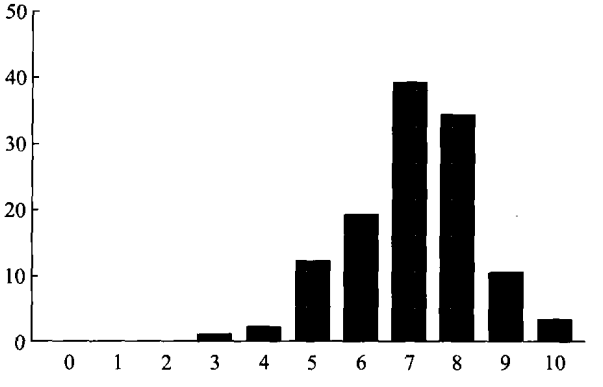

图1－1　频数图

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列出了某农场在做“三尺三”高粱提纯时所调查的100个数据。

表1－2　“三尺三”株高测量结果/cm

155	153	159	155	150	159	157	159	151	152
159	158	153	153	144	156	150	157	160	150
150	150	160	156	160	155	160	151	157	155
159	161	156	141	156	145	156	153	158	161
157	149	153	153	155	162	154	152	162	155
161	159	161	156	162	151	152	154	157	162
158	155	153	151	157	156	153	147	158	155
148	163	156	163	154	158	152	163	158	154
164	155	156	158	164	148	164	154	157	165
158	166	154	154	157	167	157	159	170	158

表中所列出的数据虽然是连续型的,但看上去好像是离散型数据。产生这种误解的原因,是由于株高的单位是"cm",在这种精确度下出现了许多高度相同的植株,当进一步提高精确度后,便很难再找到两个高度相同的植株了。从表 1-2 的原始数据中,除可以找出最大值是 170 cm,最小值是 141 cm 以及估计出它们的平均高度在 150~160 cm 之外,很难再看出什么规律来。但是,当我们将表 1-2 中的数据列成频数表之后,便可以比较清楚地看出这些数据的变化规律。高粱的株高是连续型数据,不是一个个孤立的值。因此,不能像例 1.1 那样制表。连续型数据频数表的制作过程如下:首先将数据分组,一般来说,100 个数据可以分成 8~10 组。根据极差 $R = \max y - \min y = 170 - 140 = 30$,分为 10 组比较合适,每一组的间距刚好是 3 cm。用比较简单的组间距分组,编制频数表比较方便。将分好的组填入表 1-3 的第一列。表 1-3 的第一列称为**组限**(class limit),组限是根据原始记录中的数值确定的。本实验是以 cm 为单位统计数值的,所以第一组的上限"143"cm 的实际值,可能在大于等于 142.5 cm,小于 143.5 cm 范围内。同样,第一组的下限"141"cm 的实际值,可能在大于等于 140.5 cm,小于 141.5 cm 范围内。因此,这一组的全部实际可能值是在大于等于 140.5 cm,小于 143.5 cm 范围内。140.5~称为**组界**(class boundary)。对于其他各组,同样可以定出相应的组界。

表 1-3 "三尺三"株高频数(率)表

组限/cm	组界/cm	中值	频数计算	频数	频率
(141,143)	140.5~	142	一	1	0.01
(144,146)	143.5~	145	丅	2	0.02
(147,149)	146.5~	148	下	4	0.04
(150,152)	149.5~	151	正正下	13	0.13
(153,155)	152.5~	154	正正正正下	23	0.23
(156,158)	155.5~	157	正正正正正下	28	0.28
(159,161)	158.5~	160	正正正	15	0.15
(162,164)	161.5~	163	正正	10	0.10
(165,167)	164.5~	166	下	3	0.03
(168,170)	167.5~	169	一	1	0.01
总计				100	1.00

中值(midvalue)是每一组的两个组限的平均值,但是也有例外。例如,习惯上通常以"岁"为计算年龄的单位。假若有一组的组限是 20~29 岁,上限 29 岁包括这个人可能刚刚 29 岁,也可能即将进入 30 岁。所以这一组既包括 20 岁的,也包括 29 岁的,共有 10 个年龄级。因此,中值应是 $(20+30)\div 2=25$,而不是 $(20+29)\div 2=24.5$。类似这种情况,在计算中值时应特别注意。

频数计算一列,就是将表 1-2 中的数据"对号入座",最常用的方法是采用"唱票"的方式,一人读,一人填。最后将每一组的频数统计出来,记入频数栏,并计算出频率。在制成频数(率)表以后,连续型数据的频数(率)分布规律便清楚多了。

编制连续型数据的频数(率)表,一般需要以下各步:

① 从原始数据表中找出最大值和最小值,并求出极差。

② 决定划分的组数。分组数是由数据的多少决定的,在数据较少时,如 50 ~ 100 个数,可以分为 7 ~ 10 组,数据较多时,可分为 15 ~ 20 组。

③ 根据极差与决定划分的组数,确定组限。

④ 在频数表中列出全部组限、组界及中值。

⑤ 将原始数据表中数据,用唱票的方式填入频数表中,计算出各组的频数和频率。

表 1 - 3 以表格的形式,描述了"三尺三"高粱的株高频数分布。除此之外,还可以用频数图更直观地描述这一分布。下面是三种最常用的频数图。

1. 直方图

在横轴上标明各组的组界,纵轴标明频数。然后以每一组的组界为一个边,相应的频数为另一个边,作矩形,构成**直方图**(histogram)(图 1 - 2)。若纵轴改为频率,则得到频率直方图。直方图又称组织图。频率直方图与频数直方图的图形完全一样。

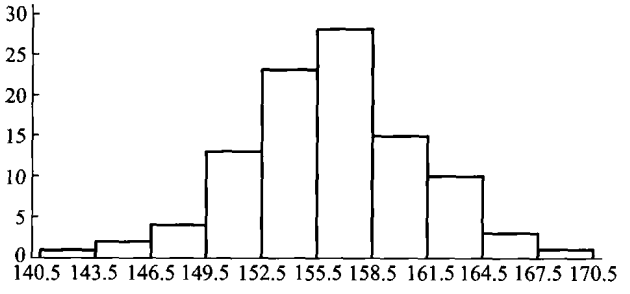

图 1 - 2 "三尺三"株高直方图

2. 多边形图

在横轴上标出各组的中值,纵轴上标出频数(率),在坐标平面内,标出相应的每个点[以中值为横坐标,以该中值对应的频数(率)为纵坐标],用线段连接各点。最低一组非零频数的点,应该直接与相邻的零频数中值相连;最高一组非零频数点,亦应该与相邻的零频数中值点相连。最后得到一个**多边形图**(polygon)(图 1 - 3)。

3. 累积频数图

经常使用的第三种频数图,称为**累积频数图**(cumulative frequency graph)。作图法如下:首先根据表 1 - 3 制成**累积频数表**(表 1 - 4)。在横轴上标出各组的中值,纵轴上标出累积频数(率)。在坐标平面内标出相应的点[以中值为横坐标,以该中值对应的累积频数(率)为纵坐标],连接各点,从而得到累积频数(率)图。

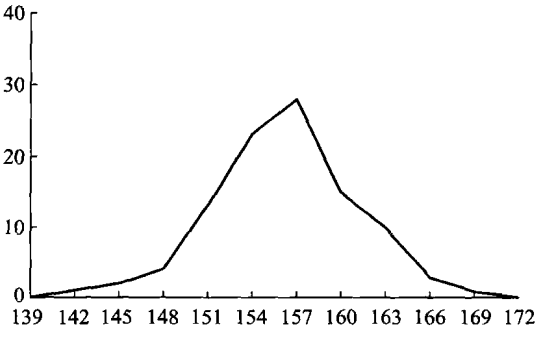

图 1 - 3 "三尺三"株高多边形图

图 1 - 4 就是根据表 1 - 4 所绘制的累积频数图。累积频数图与直方图和多边形图描