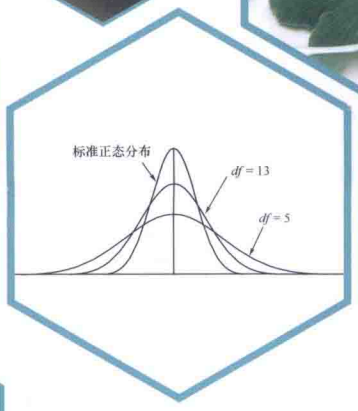

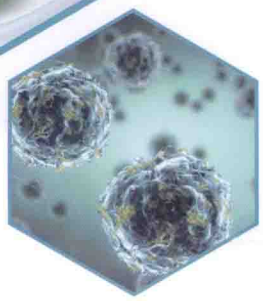

基础生物统计学

吴伟坚　许益镌　何余容　陈科伟◎编

科学出版社

基础生物统计学

吴伟坚　许益镌　何余容　陈科伟　编

科学出版社

北京

内 容 简 介

生物统计学是数理统计的原理和方法在生物科学中的具体应用。在生命科学领域的科研和生产实践中产生大量的数量资料，生物统计学可以帮助研究工作者从大量繁杂的数量资料中整理和分析出准确的信息。本书系统地介绍了数理统计学的基本原理和方法，重点介绍了数量资料的整理，描述性统计量，抽样分布，假设检验的原理，t 检验、方差分析、简单相关与回归，χ^2 适合性和独立性检验，同时对试验设计及其统计分析进行了叙述。

本书附录部分就如何利用 Excel 和 SPSS 解决描述统计、t 检验、单向和双向方差分析及相关和回归分析等相应章节的实例进行了介绍，有利于读者更好地解决科研中遇到的数理统计问题。

本书可供高等院校生命科学类和种植业类专业的本科生作为教材使用，也可供生命科学类专业的科研工作者、教师和研究生参考。

图书在版编目(CIP)数据

基础生物统计学 / 吴伟坚等编. —北京：科学出版社，2015
ISBN 978-7-03-045243-6

Ⅰ. ①基… Ⅱ. ①吴… Ⅲ. ①生物统计 Ⅳ. ①Q-332

中国版本图书馆 CIP 数据核字(2015)第 170144 号

责任编辑：丛 楠 / 责任校对：郑金红
责任印制：赵 博 / 封面设计：铭轩堂

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新科印刷有限公司 印刷
科学出版社发行 各地新华书店经销

*

2015 年 8 月第 一 版 开本：787×1092 1/16
2015 年 8 月第一次印刷 印张：13
字数：308 000

定价：38.00 元

(如有印装质量问题，我社负责调换)

前　言

生物统计学(Biostatistics)是数理统计学的一个分支，是应用数量统计学的原理和方法来分析和解释生物界各种现象和试验调查资料的一门学科。虽然这门学科较年轻，是在1925年由K. Pearson(1857～1936)和R.A. Fisher(1890～1962)所创建，其发祥地是在英国的乐桑试验站(Rothamsted Experimental Station in Harpenden)。但它发展很快，引人入胜，应用甚广，是生物学研究不可缺少的学科课程。

生物统计学发展到今天，已被引入到各个生物学领域。本学科是在正确试验设计的基础上，正确地收集整理试验数据、正确地进行统计推断并作出正确的科学结论的原理和方法。数理统计学研究的对象是随机变量，而随机变量在不同的条件下由于偶然因素影响，具有不确定性。正如生物统计学的缔造者R. A. Fisher所指出的："统计学家不是炼金人，不期望从无价值的材料中产生黄金，她像一个化学家能准确地分析出材料固有的效价并提取其含量，此外并无别的……数理统计学的工作只是为了对事物的固有本质作出正确的判断。"

本书是在植物保护专业"生物统计学"教科书的基础上，参考国内外一些生物统计学教材编写而成的。植物保护专业的生物统计学教材源于华南农业大学尹汝湛教授编写的《昆虫试验统计》，后经多次修订为《生物统计学——应用于植保科学》。尹汝湛教授和古德就教授为植物保护专业生物统计学的教材建设倾注了大量的心血。为了适应现代高等教育的发展和现代生物统计学教学的要求，编者几年前便萌生了写作这样一本教材的念头，当时，在华南农业大学昆虫学科前辈的指导下，开始编写新的教案，并且已在华南农业大学植物保护专业本科生中讲授过本书的大部分内容。

本书共分九章，系统地介绍了生物统计学的基本原理和方法，重点介绍了数量资料的整理、描述性统计量、概率分布、抽样分布、假设检验的原理、t 检验、方差分析、简单相关与回归、χ^2 适合性和独立性检验，同时对试验设计及其统计分析进行了叙述。附录部分就如何利用Excel和SPSS解决描述统计，t 检验、单向和双向方差分析以及相关和回归分析等相应章节的实例进行了介绍。吴伟坚编写了第一至第八章和附录B、C；许益镌编写了第九章和附录A，参与了第二章、第五章的编写；何余容参与了第九章

和附录C的编写；陈科伟参与了第一章、第二章和第三章的编写。每章均附有习题，本书的例题和习题主要源于尹汝湛教授的《昆虫试验统计》和编者近年发表的论文。

本书在出版过程中得到科学出版社和华南农业大学昆虫学科的大力支持，在此一并感谢。

本书可供高等院校生命科学类和植物生产类专业的本科生作为教材使用，也可供生命科学类专业的科研工作者、教师和研究生参考。

由于编者水平有限，本书难免有错漏和不足之处，诚望读者不吝赐教，以便下次修订时完善。

编　者

2015年1月于广州

目　　录

第一章　引　　言

§1.1　数理统计学的定义

在各个领域的研究中，都会碰到数量资料，而且常常会遇到类似下面的一些问题。例如：一种新的农药，如何判断它是否有效？慢性铅中毒患者的血压正常吗？如何抽检几百或几千株植株来估计某种病害的流行程度？温度对某种昆虫产卵量的影响是否存在？昆虫的人工饲料配方有没有明显改进？如何以最少的资源和人力来得到我们所需要的某种信息？等等。这一类问题的共同特点，就是人们只能得到他所关心的事情的不完全信息，或者是单个实验的结果有某种不确定性。

取得数量资料的方法一是全面调查，二是抽样调查。全面调查有时不可能做到，如农药污染了河水，不可能调查全部河水农药的含量。可能做到的往往又代价太大，如2010年开始的中国第六次全国人口普查，前后历时3年，共600多万名普查员参加，花费近80亿元，社会各界投入的人财物力及时间成本巨大。又如，为了知道灯泡合格与否或它的使用寿命，我们常常需要对它做破坏性检验，此时我们显然不能把所有的灯泡都检验一下，而只能满足于对少数几个样品的抽检，这样获得的信息显然是不完全的。再比如，要检验某病原物对植物的致病性，一般来说，接种过病原物的植物不一定全发病，而未接种的也不会全不发病。那么发病与不发病的差别究竟到多大时我们才能认为接种的病原物是有致病性呢？同时，即使我们采用完全一样的实验条件再次进行实验，发病与不发病的植物数量也会有所变化，这说明类似实验的结果具有某种内在的不确定性。要想在这种情况下正确判定病原的致病性，就涉及我们如何评价一些并不确定的实验结果的问题。

要从这样一些问题中得出科学的、可靠的结论，就必须依靠数理统计学。不同的学者曾给数理统计学下过很多定义，如：①数理统计学是一门理论和应用的学科，它用来创造、发展并应用一些技术，使归纳推断所产生的不确定性得到度量；②数理统计学是一门关于数量资料的收集、整理、分析和解释的学科；③数理统计学是一门以概率论为基础，以样本为根据，运用数学模型推断总体的学科。

统计推断是数理统计学的基本任务，为什么要进行统计推断？如果每刻每单位容量的河水的农药含量是完全相等，或者每个人的身高体重完全一致，那么问题就非常简单了，因为可以用一小部分的数据去推断研究对象的总体，也就不需要数理统计这门学科了。可事实并非如此，世界万物的状态总是参差不齐，多姿多彩的。万物状态间的差别是由两种误差造成的：①条件误差：人所能控制或确定的因素的变化而引起的变差；②随机误差：受偶然的无法控制的因素的影响而引起的变差。

在自然界和现实生活中，事物都是相互联系和不断发展的，在它们彼此间的联系和发展中，根据事物间是否存在必然的因果联系，可以分成截然不同的两大类现象，即确定性的现象和不确定性的现象。确定性现象是在一定条件下，必定会导致某种确定的结果。举例来说，在标准大气压下，水加热到100℃，就必然会沸腾，事物间的这种联系是属于必然性

的。通常的自然科学各学科就是专门研究和认识这种必然性的，寻求这类必然现象的因果关系，把握它们之间的数量规律。

不确定性现象是指，在一定条件下，事物的结果是不确定的，可能出现也可能不出现。举例来说，同一个工人在同一台机床上加工同一类型零件若干个，它们的尺寸总会有一些差异。又如，在同样条件下，进行小麦品种的人工催芽试验，各种子的发芽情况也不尽相同，有强弱和早晚的分别。为什么在相同的情况下，会出现这种不确定的结果呢？这是因为，我们说的“相同条件”是针对一些主要条件来说的，除了这些主要条件外，还有许多次要条件和偶然因素是人们无法事先一一掌握的。正因为这样，我们在这一类现象中，就无法用必然性的因果关系对个别现象的结果事先预计出确定的答案。事物间的这种关系是属于偶然性的，这种现象叫做偶然现象，或者叫做随机现象。

在自然界以及人们的生产生活中，随机现象十分普遍，也就是说随机现象是大量存在的。比如：同种昆虫不同个体的体重、同一条生产线上生产的灯泡的寿命等，都是随机现象。因此，我们说：随机现象就是在同样条件下，多次进行同一试验或调查同一现象，所得结果不完全一样，而且无法准确地预测下一次所得结果的现象。随机现象这种结果的不确定性，是由于一些次要的、偶然的因素影响所造成的。

随机现象从表面上看，似乎是杂乱无章的、没有什么规律的现象。但实践证明，如果同类的随机现象大量重复出现，它的总体就呈现出一定的规律性。大量同类随机现象所呈现的这种规律性，随着我们观察次数的增多而愈加明显。比如掷硬币，每一次投掷很难判断是哪一面朝上，但是如果多次重复地掷这枚硬币，就会越来越清楚地发现它们正、反面朝上的次数大体相同。

我们把这种由大量同类随机现象所呈现出来的集体规律性，叫做统计规律性。概率论和数理统计就是研究大量同类随机现象的统计规律性的数学学科。

在一般的科学研究中，随机误差和条件误差往往是混在一起，甚至会把随机误差误认为条件误差。从这个意义上来讲，数理统计学的任务有二：①进行合理的试验设计，减少随机误差；②对随机误差作出适当的估计，从而辨认出是否存在条件误差及条件误差的大小。

由于随机误差的普遍存在，数理统计学渗透到科学技术的每个领域和生活的各个方面。随机误差是数理统计学研究的主要内容，而概率论正是研究这种误差本身的普遍性和规律性的学科，故概率论又是数理统计学的重要依据和基础。

数理统计学在很多领域都被证明了是必不可少的工具，即所谓的工具性学科。工具(tool)泛指生产、生活中使用的器具或用以达到某种目的的东西或手段。天文学家根据统计方法预言天空物体的未来位置；遗传分离定律是由统计方法确定下来的；人寿保险费与赔偿金额是以统计记录为基础的生命表核定的；工程师们发现抽样调查方法在控制产品质量方面的价值是无法估量的；商业领导人和政府的智囊团使用统计方法作出决策。

生物统计学便是数理统计学这种工具在生物学中的应用。生物学是一门实验科学，不管你从事的是生物学的哪一个分支，都不可能完全脱离试验或野外调查。而试验或调查所得到的结果几乎无例外地都带有或多或少的不确定性，即试验误差。在这种情况下不用数理统计学是不可能得到正确的结论的。作为一个实验科学工作者，离开了数理统计学就寸步难行。希望读者通过学习，能够掌握常用的数理统计方法，尤其是它们的条件、适用范围、优缺点等，从而能够应用它们去解决实践中遇到的问题。

§1.2 数理统计学的发展简史

统计是一个古老而时髦的名词。古老:它是作为国家的计算和统计开始的,我们可从亚里士多德的《国家事物》和《圣经》等书籍中找到这些记载。在奴隶社会和封建社会,统计意味着财富统计、人口统计和税收统计等,即国力统计;从数理统计学(Statistics)、统计学家(statist)和国家(state)三个名词中也可看到数理统计的渊源所在。时髦:现国家各级政府均设有统计局,我们常常听到不少的统计数据:人口、粮食产量、物价指数、国民生产总值、失业率等,这些均属社会经济统计范畴。前苏联科学院、苏联中央统计局和苏联高教部于1954年3月召开的联合科学会议上曾把社会经济统计和数理统计严格区别开来,分别列入社会科学和自然科学中,认为社会经济统计的基础是马克思主义哲学和政治经济学。事实上两者均研究数量资料,两者间并无不可逾越的鸿沟。

概率论产生于17世纪,本来是应保险事业的发展而产生的,但是来自于赌博者的需求,却是数学家们思考概率论问题的源泉。早在1654年,有一位法国知识分子赌徒梅累(Mere)向当时的数学家帕斯卡(Blaise Pascal)提出一个使他苦恼了很久的问题:"两个赌徒相约赌若干局,谁先赢 m 局就算赢,全部赌本就归谁。但是当其中一个人赢了 $a(a<m)$ 局,另一个人赢了 $b(b<m)$ 局的时候,赌博中止。问:赌本应该如何分法才合理?"此后,帕斯卡在1642年发明了世界上第一台机械加法计算机。三年后,也就是1657年,荷兰著名的天文、物理兼数学家惠更斯(Christiaan Huygens)企图自己解决这一问题,结果写成了《论机会游戏的计算》一书,这就是最早的概率论著作。概率论是根据大量同类随机现象的统计规律,对随机现象出现某一结果的可能性作出一种客观的科学判断,对这种出现的可能性大小作出数量上的描述;比较这些可能性的大小、研究它们之间的联系,从而形成的一整套数学理论和方法。

16~18世纪,赌博盛行促成了概率论的诞生(以Jakob Bernoulli的《猜测术》为标志);殖民扩张、航海业和保险业的发展使人口统计学(Demography)得到很大的发展;高斯(Gauss)从重复测量一个数量误差的研究中导出了Laplace-Gauss方程;孟德尔的豌豆杂交试验,气象学、社会学、天文学等许多学科大量应用了概率论的原理和方法。

19世纪,Karl Pearson花了大半个世纪研究数理统计。Karl Pearson原为数学物理学家,后来研究遗传学,提出了相关与回归的概念,发展了 χ^2 检验,在文献中引进了"均差"、"标准差"等名词并创办了 *Biometrika* 杂志。William Sealy Gosset(Pearson的学生)以"Student"为笔名在Biometrika上发表了许多关于小样本抽样方面的文章。

20世纪:Ronald Aylmer Fisher及其学生们受Pearson和Gosset的影响,对数理统计学的发展作出了巨大的贡献,如提出零假设的概念,提出 F 检验和方差分析等。

数理统计学作为一门学科的诞生是以Fisher于1925年写的一本著作 *Statistical Methods for Research Workers* 为标志的,故数理统计学是20世纪初的产物,曾被美国一家杂志评为20世纪对人类影响最大的25门学科之一。

根据数量资料提供的信息作出的判断,对日常生活的影响与日俱增。数理统计这一科学序列,已成为处理每个有数量资料出现的领域的必不可少的工具。今天,建立在以概率论为基础的现代统计学,在物理学、生物学、化学、医学与农学等自然科学中,在经济学、教育学和社会学等社会科学中,在政府和企业中,都被证明是不可或缺的助力。

数理统计学的应用范围不尽相同,但所用的基本原理和基本方法则大部分是相同的。

第二章　数据分析导论

§2.1　变量和数据的类型

生物统计学中所需要研究和处理的数据属于变量(variable)。对不同的个体或单位具有的同一性状进行观察的结果,可以获得不一定相同的观察值,则这个性状就称为变量,每一个观察值称为该变量的数据(variate)。生物学上有各种各样的变量,这些变量可以包括形态学上的测量如高度、长度等,生物体内某种化学物质的含量,某种生物过程中不同指标间的比率,某种行为出现的频率和用于生物研究方面的电、光学仪器上的读数,等等。例如,昆虫的体重、虫口密度、昆虫取食量和交配次数、昆虫过冷却点温度、各虫态的历期和单位面积作物产量等,都是变量。变量通常可划分为以下三种类型:定量变量(quantitative variables)、序列变量(ranked variables)和属性变量(categeorical data 或 qualitative variables)。定量变量又分为离散型变量(discrete variables)和连续型变量(continuous variables)。

2.1.1　定量变量

1. 离散型变量

离散型变量中每个数据都是整数,因此数据间的差异也必然是整数,亦称为计数资料(count data)。因为观察时只能一一点数而不能称量。例如,每个调查单位有虫 0 头、1 头、2 头……但是应该指出,经过统计加工的指标,如平均数,则可以是非整数。例如,每个单位平均有虫 1.5 头。

2. 连续型变量

当数据由大到小顺序排列时,每两个数据之间总有可能取多于一个中间数值的变量叫做连续型变量。例如,长度、重量、面积和容量等都属于连续型变量。连续型变量亦称为测量资料(measure data),因为它只能量和度而不能一一点数。它的原始数据是以截取一定小数位数的近似值来表示的。

2.1.2　秩次变量

将已有的计数资料或测量资料或上述等级资料,重新按数值由小到大顺序排列,然后依次给予每值一个秩序值,如 1、2、3……秩次变量可以运用特定的方式进行统计分析(属非参量方法)。

2.1.3　属性变量

1. 二项变量

二项变量也叫名义变量。调查得来的数据只有两种类型,非此即彼,如雌或雄、存活或

死亡、寄生或非寄生、发芽或不发芽等。通常是以其中之一方调查单位数占全部调查单位数的比率(百分比)来表示,称为死亡率、雌(雄)性百分率、化蛹率、寄生率、发芽率等。有些情况可人为地运用 0-1 化处理。

2. 等级变量

按一定的分级标准,把调查对象的表现分为若干等级,每个等级定出级值,如 1、2、3、4、5,或 1、3、5、7、9 等。于是,调查时将每个观测对象评定一个级值,加以记录,以后可将等级资料如同上述的计数资料或测量资料一样进行统计分析。这种做法优点在于简化工作,评级的标准是很灵活多样的,既可据已有的计数资料或测量资料的大小来划分等级范围,也可以据难于以数值表示的特征,如色、香、味、作物长势、虫害程度概况等来分级。昆虫的龄态当然也可以作为分级标准以表示发育进度。

§2.2　总体和样本

一个变量的全部数据构成总体(population)。总体也可以理解为某一性状的全部观测值,如一块稻田上全部的三化螟卵块。总体也可以理解为全部观测单位,总体的含量通常以 N 表示。

一个变量的一部分数据(又称变员数),即总体的一部分,被抽出来代表总体的,叫做样本(samples)。取得样本的方法叫做取样技术。样本通常是由多个取样单位集合而成的,因此,取样技术是指如何决定取样单位的大小形状、个数及位置等。通常要求取样单位要按随机的原则,即让总体内每个调查单位被选取的机会同等。按随机的原则取得的样本称为随机样本(random sample)。我们处理的样本,通常都要求是随机样本。来自随机样本的数据叫做随机变量(random variable)。统计学所研究的对象就是随机变量。

对总体内所有的调查单位都一一加以观察,叫做全面调查。这样得来的信息,当然最能接近总体的真实情况。但是,实施全面调查往往是很困难,甚至是不可能的。因此只能取样,考察样本,通过样本的信息估计总体的实际情况。通常总体属于未知,而样本则来自实测。为此,提高样本的代表性极为重要。但是,从样本得到的信息同总体实况之间总有或大或小的距离,这个差异叫取样误差,取样误差小则表示样本的代表性高。

样本含量(sample size)指的是样本内取样单位的个数,也指样本内数据的个数,一般以 n 表示。通常含量在 30 以上的样本叫做大样本,30 以下的叫小样本。大样本和小样本在分析方法上是不同的。随着样本含量增大,其所含信息对总体的代表性相对提高。但在许多情况下,只能就较小的样本进行研究。从总体的资料计算出来的特征数,如平均数和标准差等,叫做总体特征数。总体特征数又叫做参量或参数(parameter)。参量是定值,是不变的,因此也叫真值,常以希腊字母 μ、σ 等表示。从样本资料计算出来的特征数,叫做统计量(statistic)。统计量是用来估计总体的真值的,因此属于估计值(estimator),常以字母 $\bar{x}$、S 等表示(图 2.1)。同一个总体的不同的样本,它们的统计量可以不相同。即使是同一个总体,也可以有来自不同样本的不一定相同的估计值。至于估计的精确度则同取样误差有密切的关系。从同一个总体所取得的多个样本的统计量,可以被视作一个新的总体的数据,即由统计量构成的总体。它的数据的分布称为抽样分布(sampling distribution)。

图 2.1 统计分析的基本过程

总体与样本的关系是统计学上非常重要的基本知识，在分析过程中，千万不要把总体与样本混为一谈，或只见样本而不知有总体，把样本等同于总体。

§2.3 数据的整理

本节讨论如何整理和表达实际观测得来的样本资料，着重点是显示样本内变员数的分布状况。这里说的是实际观测值的分布，请不要与第三、四章中所介绍的理论分布混淆。

从调查得来的样本原始资料，数值有大有小，调查和记录有先有后，常常看起来是一堆杂乱无章数据，不容易看出观测值分布的趋势，更不便于计算和分析。为此，对原始资料加以整理是很必要的。数据表达的方式通常有频数分布表、统计表与统计图等。

2.3.1 频数分布表

在数据整理工作中最重要的工作是制作频数分布表。第一步是制“依次表”，也就是把各个变员数按由小到大的顺序排列起来，使原始资料系统化。第二步是分组归纳各个变员数至适当的组。落在某一个值或某一组的下限与上限之间的变员数称为频数(frequency)。频数分布表的结构其实很简单，主要的是两列，其一是分组的准则，其二是频数。有时还有第三列，叫比率，是百分比值，它是根据绝对频数计算而得，因此，也叫相对频数(频率)。有了这项相对频数，就便于表内各组之间的相互比较，也便于不同频数分布表之间的相互比较。连续型数据的频数分布表是按组限来分组的，为了进一步计算和图示，往往设立“组中值”这一列，作为各组变员数的代表值。

所谓分组准则，有几种情况，最简单的是直接地依据属性或名义来分组，如作物田类名称或品种，时间上的月份、年份，虫害的轻、重，作物器官或产品色、香、味方面的性状等。表 2.1 是通过一次调查结果制作的以某种昆虫虫(龄)态分布为例的频数分布表，其中每个虫态的个体数就是频数，这种频数分布表广泛应用于虫害预测预报的发育进度调查和生命表技术。

表 2.1 某昆虫各虫态个体数量频数分布表

虫态	频数(f)	比率/%
1 龄幼虫	2	2.86
2 龄幼虫	13	18.57
3 龄幼虫	30	42.86
4 龄幼虫	12	17.14
5 龄幼虫	4	5.71
预蛹	2	2.86
蛹	5	7.14
蛹壳	2	2.86
合计	70	100.00

常见的频数分布表是按数值来分组的，有两种情况：①按单位频数的组值（或级值）分组；②按一定区间，即组限（包括下限和上限）分组。离散型数据大多数以组值分组，个别情况应用组限分组，但连续型数据则一定按组限分组。

1. 离散型数据的频数分布表

【例 2.1】 调查 200 丛水稻遗株，每丛内越冬三化螟（*Tryporyza incertulas*）幼虫数量的原始调查资料如下：

1,1,0,0,2,0,0,1,0,2,1,0,1,1,0,1,0,0,3,0,2,1,0,0,1,0,1,0,0,1,0,1,0,1,0,0,
0,0,5,0,1,0,0,0,0,4,2,0,0,3,0,4,1,3,1,4,0,1,2,6,0,3,2,1,0,2,0,0,1,1,0,0,0,0,
0,0,0,0,2,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
2,0,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,1,1,0,0,0,1,1,1,0,
0,1,1,1,0,1,0,0,0,1,1,0,0,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,0,
0,1,0,1,1,0,0,0,0,0,1,0

可列出频数分布表（表 2.2）。

表 2.2　200 丛水稻遗株内越冬三化螟幼虫数量频数分布表

x（组值）	f（频数）	比率/%
0	134	67.0
1	48	24.0
2	9	4.5
3	4	2.0
4	3	1.5
5	1	0.5
6	1	0.5
合计	200	100.0

这样的表叫做闭式表。有时因为高值组的频数很少，但组数多，就整理成开式表。例如表 2.2 可写成如表 2.3 的开式表。

表 2.3　200 丛水稻遗株内越冬三化螟幼虫数量频数分布表

x（组值）	f（频数）	比率/%
0	134	67.0
1	48	24.0
2	9	4.5
3	4	2.0
4	3	1.5
≥5	2	1.0
合计	200	100.0

【例 2.2】 某些果树的果实受害虫为害后，可按受害程度分级，按级值定义调查后可列成表 2.4 的频数分布表。

表 2.4 某果树果实受害虫为害程度情况

受害等级(级值 x)	各级受害果实数(频数 f)	比率/%
0	20	5.56
1	40	11.11
2	100	27.77
3	120	33.33
4	60	16.67
5	20	5.56
合计	360	100.00

有时资料太繁杂，相同的观测值过少，为了使频数分布表呈紧缩形式，易于看出分布的特点，可以“组限”来分组。

【例 2.3】 1956 年广州市石牌区的第五代三化螟 53 个卵块含卵粒数的资料，以 30 粒卵为一组，可列成表 2.5 的频数分布表。

表 2.5 第 5 代三化螟卵块含卵粒数的频数分布表(广州石牌，1956)

每卵块含卵粒数(组限 x)	卵块数(频数 f)	比率/%
1～30	0	0
31～60	4	7.55
61～90	11	20.75
91～120	14	26.42
121～150	11	20.75
151～180	7	13.21
181～210	4	7.55
211～240	2	3.77
合计	53	100.00

2. 连续型数据的频数分布表

连续型数据的频数分布表总是依据组限来分组的，每组下限至上限的距离叫组距，各组的组距相同。为了便于数据的归类，往往把第一组的下限稍定小一点。第一组的下限确定后，把第一组的下限加上组距，即得第一组上限，以后各组的上、下限都可以连续地推导出来。由于资料是连续的，为了便于变员数的归类，相邻两组的上限和下限的写法有多种形式。如下介绍的是其中一种，即把每组的上限写成比其应有的值(即次组下限的值)略小些，即可写成 0.9 或 0.99 这样带小数的形式。例如，组距为 10，第一组下限为 0，各组可写成如下形式(表 2.6)。

表 2.6　连续型数据频数分布表组限划分格式

组次	下限	上限	频数
1	0	9.9	2
2	10	19.9	4
3	20	29.9	6
⋮	⋮	⋮	⋮

至于组距的大小，要按资料的具体情况和分组数多少而适当决定。可以以 10、20、30 等为组距，适合于一般计数习惯，但也不应受习惯限制。至于分组数的多少，以 6～20 组为适当，同时要考虑样本含量(n)，可按表 2.7 分组。

表 2.7　分组数与样本含量关系

样本含量	宜分组数
40～60	6～8
60～100	7～10
100～200	9～12
200～500	12～17
>500	17～20

以组限分组的频数分布表中的 x 为组中值，组中值＝下限＋组距/2，而组距＝上限－下限。下面介绍一个制作连续型变量的频数分布表的例子。

【例 2.4】 在广州天河区称量 106 头越冬三化螟幼虫体重(单位：mg)，原始资料如下：

13.0　18.4　19.4　23.3　24.3　24.7　25.1　25.2　25.6　26.0　27.6　28.0　28.2
28.2　28.3　28.3　28.5　29.1　29.3　29.8　30.1　30.2　30.3　30.4　30.5　30.7
31.0　31.7　31.8　32.0　32.8　32.8　33.1　34.3　35.2　35.3　35.6　35.8　35.9
36.3　36.3　36.3　36.6　37.0　37.3　37.5　38.0　38.6　38.6　38.6　38.8　39.2
39.3　40.0　40.2　40.3　40.3　40.4　40.6　40.8　41.3　41.6　41.8　41.8　41.8
42.0　42.4　42.5　42.9　42.9　43.1　43.3　43.7　43.8　44.2　44.2　46.1　47.0
47.3　47.9　48.0　48.1　48.3　51.6　52.1　52.9　53.3　53.3　54.5　56.4　58.5
59.1　59.3　59.4　60.0　60.5　61.1　62.5　63.8　69.7　71.8　72.7　76.2　76.7
79.6　86.2

现以 6 mg 为组距，分成 13 组，第一组下限为 10 mg，制作频数分布表如表 2.8 所示。

表 2.8　越冬三化螟幼虫体重(mg)(广州天河，1963)

组限	x(组中值)	f(频数)	比率/%
10～15.9	13	1	0.94
16～21.9	19	2	1.89
22～27.9	25	8	7.55
28～33.9	31	22	20.75

续表

组限	x(组中值)	f(频数)	比率/%
34～39.9	37	20	18.87
40～45.9	43	23	21.70
46～51.9	49	8	7.55
52～57.9	55	6	5.66
58～63.9	61	9	8.49
64～69.9	67	1	0.94
70～75.9	73	2	1.89
76～81.9	79	3	2.83
82～87.9	85	1	0.94
合计		106	100.00

3. 制作频数分布表的意义

由以上几个频数分布表可以看出，原来表面上看来很不规律的变量资料，经过整理、制作成频数分布表后，竟然不是杂乱无章的，而变员数的分布总是有一定的趋势。最常见的情况是频数两头小、中间大，也有头大尾巴长、递增型等状态。由此可见，样本资料的分布在一定程度上体现了变量的规律性。

变量是集中性和分散性的辩证统一。这两个特点是变量规律性中最基本和最本质的东西。分散是主导的一面，没有分散就不称其为变量了。然而各个变员数又有力求集中的一面，虽然集中的程度、位置和形状因具体变量资料而有所不同。我们把样本资料整理成频数分布表，第一个目的就是要使它呈现出离散趋势和集中趋势的状况，给人们以清晰的印象，从而初步掌握这个变量的特点。当然，由于取样误差的干扰，样本数据的分布不可能充分地、完整地代表总体数据的真实分布，但仍可以看出总体分布的趋势。

制作频数分布的第二个目的是便于图示、计算和分析。在试验研究中，为了解变量的离散趋势和集中趋势，不应仅仅应用平均数、最小值和最大值这三者来表示，这不足以充分表示变量分布的特点。譬如，某地在第一代三化螟产卵盛期，用石油乳剂混和乐果(石乐合剂)喷施卵块做试验，试图减少蚁螟孵化从而减少枯心苗数量。结果，经施药的卵块所形成的枯心苗群与对照群相比，枯心苗数大大地减少了。如果只用最小、最多和平均数这三个数值表示，试验(每群枯心苗数)如表 2.9 所示：

表 2.9 石乐合剂喷施三化螟卵块后水稻枯心群苗数调查

处理	枯心群苗数		
	最小	最多	平均
施药	0	34	11.35
对照	15	126	57.91

这样的数据是不便于对试验结果作进一步分析的，用以表示试验结果也是不能令人满