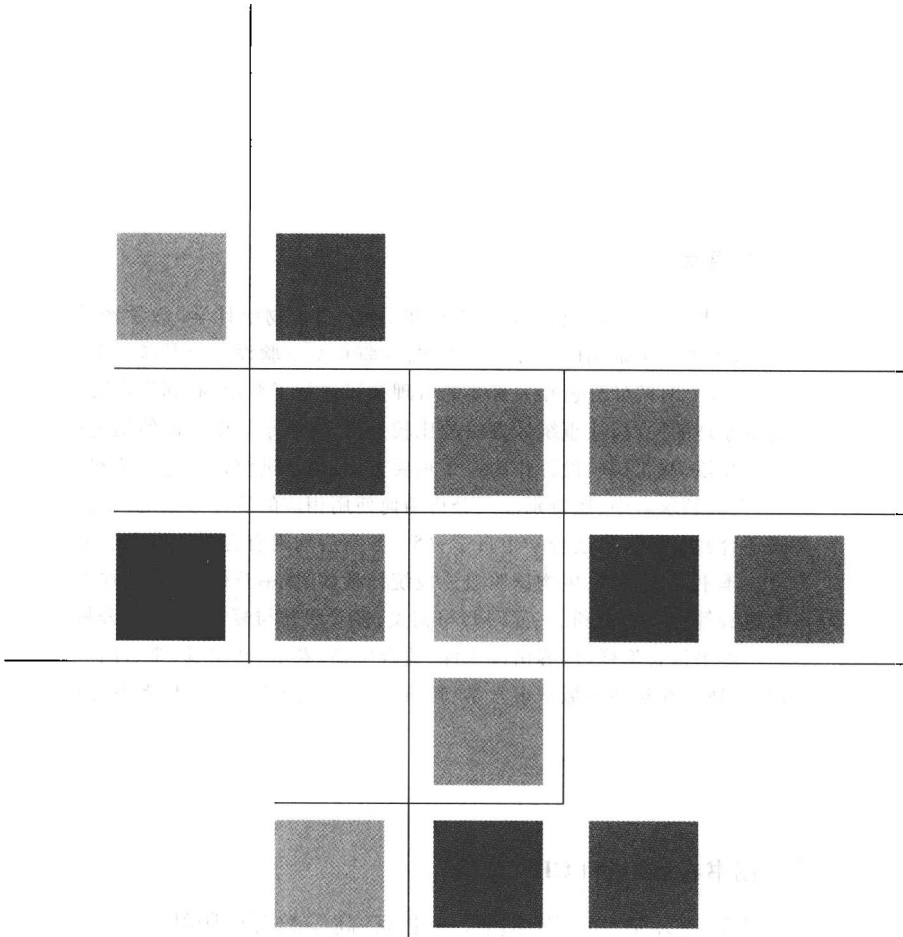
Biostatistics

陈庆富 编著

生物统计学

高等教育出版社
HIGHER EDUCATION PRESS



陈庆富　编著

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

内容提要

本书是编者多年从事大学本科和研究生“生物统计学”教学经验的总结，具有以下特点：(1) 内容全面、完整，基本涵盖了常用的统计分析方法，如将正交试验设计及其统计分析方法、通径分析、聚类分析等内容均纳入教材；(2) 对统计分析相关必要的数理知识进行了介绍，有利于克服高等数学与生物统计学的断层；(3) 在介绍基本原理的同时，注重统计方法的比较分析，有利于学生正确使用统计分析方法；(4) 重视生物科学研究的基本原理介绍，并将试验设计作为一个重要方面，整合进统计原理介绍和例题分析等部分，有利于培养学生的创新能力；在例题求解中，还特别注意分析为何使用相应的统计方法等问题，使读者能正确使用和理解统计分析方法，既适合教师教学也适合学生自学；(5) 各章后均附有提要和丰富的习题，有利于读者抓住重点和理清思路。

本书主要包括：生物试验概述及统计参数、概率及概率分布、样本参数的统计推断(假设检验、区间估计、卡方检验等)、方差分析、一元回归与相关、多元回归与复相关、通径分析和聚类分析等。

本书语言流畅，内容由浅入深，适合生物、农学、林学、医学、环保、园林、食品、教育等专业的本科、研究生使用，也适合在这些领域从事教学、科研和生产实践的科技工作者参考和自学。

图书在版编目(CIP)数据

生物统计学/陈庆富编著. —北京：高等教育出版社，2011.3

ISBN 978-7-04-030993-5

Ⅰ.①生… Ⅱ.①陈… 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(2010)第 257989 号

出版发行	高等教育出版社	购书热线	010-58581118
社　　址	北京市西城区德外大街 4 号	咨询电话	400-810-0598
邮政编码	100120	网　　址	http://www.hep.edu.cn http://www.hep.com.cn
经　　销	蓝色畅想图书发行有限公司	网上订购	http://www.landraco.com http://www.landraco.com.cn
印　　刷	北京人卫印刷厂	畅想教育	http://www.widedu.com
开　　本	787×1092　1/16	版　　次	2011 年 3 月第 1 版
印　　张	17.75	印　　次	2011 年 3 月第 1 次印刷
字　　数	430 000	定　　价	31.00 元

物料号　30993-00

前　言

生物统计学是高等数学在生物学研究中的重要应用，是现代生物学研究不可缺少的工具，是培养学生科学研究和综合分析能力的重要课程，也是生物科学工作者必备的基础；同时，生物统计学又是其他专业课程的重要基础。因此生物统计学在生物、农学、林学、医学、环保、园林、食品、教育等专业的大学本科和研究生教育中占有重要地位。然而，目前“生物统计学”在教学上存在很多问题，其中主要有如下几方面：(1)“生物统计学”与它的主要基础课“高等数学”在教学时间上脱节。“高等数学”常常在大学一年级开设，而“生物统计学”一般在三年级或四年级开设，这种时间断层阻碍了学生对“生物统计学”的学习。(2)“生物统计学”与“高等数学”的教学在知识衔接上脱节。长时间以来，它们一般被作为两门独立课程来开设，前者由生物学教师讲授，后者由数学教师讲授，所讲授知识内容之间常常存在很多不衔接之处，形成了不利于“生物统计学”教学的知识断层。(3) 普通《生物统计学》教材重视数理统计方法的讲授，但在一定程度上忽略了这些数理统计分析方法必须在试验设计正确及其所获资料准确的基础上才能发挥正确的作用。(4) 普通《生物统计学》教材常常忽略一些在生物学研究中使用非常广泛的统计方法，如正交试验设计及其统计分析方法、通径分析、聚类分析。(5)“生物统计学”是大学课程中最难自学的课程之一，不利于统计分析技术的广泛推广。

针对上述问题，编者以近10年来使用的大学本科和研究生《生物统计学》教案为基础修订编写了本教材。本教材主要在以下几个方面进行了尝试：(1) 重新组合了分布于不同教材中的一些常用统计分析方法，使教材包含目前生物学研究中常用的统计分析方法，从而更好地适应现代生物学研究的需要。例如，在介绍基础的生物统计学内容的同时，增加了正交试验设计及其统计分析方法、通径分析、聚类分析等内容。(2) 在介绍统计分析方法之前，融入了一些基本的数理统计知识，以增加读者对统计学分析方法的理解，增加利用这些知识解决生物统计学中参数计算的途径。例如，对抽样分布进行了较详尽的介绍，并以此为基础在统计假设检验原理介绍中提出了累计概率、总体参数、样本参数标准化变量等多种假设检验方法；在多元线性回归中引入了较多的线性代数知识，为求解逆矩阵介绍了逆矩阵定义解法、相似性变换(初等行变换)、列表解法等多种方法。(3) 在第一章增加了生物试验原理和试验设计介绍，在各章统计方法介绍中，也融入了相关试验设计，同时还结合例题尽量详细介绍试验设计中的一些主要环节，以突出试验设计的重要性和形成试验设计样板，以便供读者参考。(4) 文字叙述尽量详尽，逻辑清楚，便于读者自学。(5) 在统计方法介绍中，对一些相关统计方法进行集中对比分析，提出基本的统计分析步骤并分析这些统计方法的差异，然后再用例题加以说明；在例题求解中，还特别注意分析为何使用相应的统计方法等问题，目的是使读者能更好地正确使用统计分析方法，避免乱用或机械地套用统计方法而得出错误结论。此外，本教材还对各章给出了提要，以便读者抓住重点和理清

思路。

在本教材的编写中，很多内容参考了书后所列参考文献，在此对原著者表示衷心的感谢！书中数据只供学习和巩固本学科知识使用，不一定具有真正的实际意义，请广大读者切勿盲目引用。在此书的编写和出版过程中得到贵州师范大学自然地理学博士授予点建设经费的支持，在此表示衷心的感谢。

本书适合生物、农学、林学、医学、环保、园林、食品、教育、自然地理、环境科学、化学、化工等专业的本科和研究生使用，也适合在这些领域从事教学、科研和生产实践中的科技工作者参考和自学。

尽管编者尽心竭力，但是错误及不当之处在所难免，敬请读者不吝指出，编者将不胜感谢（e-mail：cqf1966@163.com）。

编　者

2010年4月26日

目　录

第一章

生物试验的基本方法及其数据资料的统计参数

生物试验要以生物为材料进行研究。通常生物材料的数量很大、甚至无穷大，难以全部参加试验，须通过抽取有代表性的部分个体进行试验、观测和记载，以获得相关的数据资料。这些资料再按照一定的程序，进行科学的整理和统计分析，在犯错误的风险尽可能小的情况下作出结论，从而透过现象看到本质，从感性认识上升到理性认识。这是生物科学研究的一般程序。其中，如何抽取样本、如何进行生物试验、如何对资料进行科学的分析和推断是生物统计学的基本内容，也是深入研究客观事物的重要步骤。显然，生物统计学是生物科学研究方法论的基本内容，是一种重要的分析工具，也是科研能力的重要组成部分。而生物试验又是生物统计分析的基础，因此只有良好的生物试验设计及其所获得的准确试验结果，生物统计方法才能发挥正确的作用。

第一节　生物试验中最基本的统计学术语

1. 总体

总体是指服务于研究目的的、具有共同性质的个体所组成的集团。它常常是设想的或抽象的。总体中的个体数目反映了总体的大小，称为**总体容量**。它可以是无穷多的。比如，在研究水稻品种汕优 63 的产量潜力时，其总体是指此品种在多年（含过去、现在和未来）、多地点（含省内外、国内外）、无数次种植中的所有个体的总和。这种容量（或总量）无法确定的总体，称为**无限总体**。在生物研究中，通常为了得出一般规律，其总体常常是无限总体。当然，总体容量也可以是有限的，如研究某班学生的身高变化规律，研究某奶牛场某时间段奶牛的产奶量，等等。它们所涉及的研究对象数量有限。这种数量有限的能够确定个体数目的总体，就称为**有限总体**。

2. 观察值

观察值是指每个个体的某一性状、特性的特定观察数据。同一总体的各个个体总是有变异的。例如，同一小麦品种，如贵农 10 号，即使在同一条件下种植，由于受许多偶然因素的影响，各个体的植株高度也彼此不完全相同。正如世界上没有完全相同的两个人一样，即使是双胞胎也有一定差异。因此，不同个体的观察值常常是不完全相同的，尤其是量测性状更是如此。

3. 变量（或变数）

变量（或变数）也叫随机变量（或随机变数），是表现出变异的观察值的总称，常常用 X 表示。如研究某奶牛品种产奶量变化规律时，产奶量就是变量，而具体的某个奶牛的产奶量就是观察值，即该变量的具体取值。在研究植株高度变化规律时，株高就是变量，某一个体的具体株高数值就是观察值，依此类推。

4. 总体参数

总体参数是指由总体的全部个体观察值(即总体中变量 X 的所有可能取值)计算出来的能反映总体数据特征特性的参数,也称总体特征数,如总体平均数、总体方差等。

5. 样本

由于研究目的常常是要得出一般性的结论和规律,其总体常常是无限总体,因而研究中不可能对总体中的所有个体都进行观察和研究,所以需要从总体中抽取有代表性的若干个体,进行试验研究,由这些抽取个体所得结果推断出总体特征特性。这些在总体中被抽取的个体所形成的子集,就称为样本。样本中的个体数目即样本大小称为**样本容量**。如小麦品种贵农 10 号株高变异规律研究,我们没有必要对所有植株都进行测量,只需抽取一定数目的植株如 100 株,测量其株高,根据这 100 株所测得的株高数据对总体的株高进行评价。这 100 个个体就是样本。若样本是随机地从总体中抽取的,则该样本称为**随机样本**。不是所有样本都能反映总体和代表总体。一般随机样本能较好地代表和反映总体。一般而言,样本越大,样本就越能代表总体。通常将样本容量大于 30 时的样本称为**大样本**。值得注意的是,不同的研究目的和研究性质,其大样本的数值和标准常常不同。例如,在研究单显性基因遗传规律时,测交 BC_1 遗传分离群体 30 个个体的样本就能被接受,但 F_2 遗传分离群体一般要大于 50 时才被认为是大样本,而在研究多个基因遗传规律时,F_2 遗传分析群体常常要大于 100 个个体才算有代表性。显然,大样本对总体的代表性较好,在研究中应尽量采用大样本进行试验。

6. 样本参数

样本参数是指根据样本中各个个体的观测值数据计算出来的反映样本观测数据特征特性的参数,也称为**样本特征数**或**统计参数**,如样本平均数、样本方差等。它们是总体相应参数的估计值。在实践中,总体参数常常是未知的,我们一般用样本参数估计总体参数,由样本对总体进行评价。

第二节　生物试验的基本方法、类型及基本要求

为了认识生物及其生长发育规律,指导生物工业和农业生产,在有人为控制的条件下所进行的生物观测活动,就称为**生物试验**。生物试验是获得生物统计资料的基本手段。

一、生物试验的基本方法

根据生物试验的目的和特点,生物试验主要有两类最基本的研究方法。

1. 调查和研究总体在自然条件下的现状和变化规律

通过调查和观测自然条件下的特定生态系统、特定生物类型、农业生产或人类其他活动的相关研究总体或其特定样本,根据这些研究总体或样本的调查资料进行统计分析,获取有关总体的特征特性、变化规律等方面的认识,这对于我们了解自然条件下的生物变化规律有重要意义,如人口普查、农业普查、特定生态环境调查、特定资源分布调查、特定生物结构的观察等。此外,日常的教学、科研、试验、生产、管理中所产生的数据资料也可反映相关研究总体的情况,有目的地积累这些数据,结合国内外文献资料,经统计分析和综合分析后,也可获得很有价值的信息或规律。

2. 调查和研究总体在特定条件下的变化规律

调查和研究总体在特定条件下的变化规律也称为专题研究，是指为特定研究目的，人为改变一些因素和条件，观察分析总体或样本特征特性的变化规律，以探明这些因素和条件对总体的效应和影响规律。这是最常见的生物试验，也是科学研究中最常用的方法。

上述两类方法并不矛盾，在实际应用中常常可以并用或交叉使用。因为自然条件不是固定的，常常是变化的，因此在自然条件下也可以研究各因素和条件对总体的效应和影响规律。另外，人为设定的条件也可以是自然条件下可能出现的条件，所得规律也可以用来预测特定自然条件下的总体情况。

二、生物试验的类型

根据试验所考虑因素的多少，生物试验可分为单因子试验和多因子试验。**单因子试验**是只考虑一个因素，而不考虑其他影响因素的试验。如 N 肥增产试验，只考虑 N 肥这个因素，在试验中以不同施 N 量（不同处理水平）进行某种植物的种植试验，观测不同施 N 量下的产量，根据产量高低，找出最佳施 N 量，从而为农业生产提供指导。单因子试验较简单易行且结果明了。**多因子试验**是在试验中同时考虑多个影响因素的试验。如 N、P、K 肥增产试验，是在试验中同时考虑了 N、P、K 三个因素，每个因素都可有不同施用量，一定的 N、P、K 施用量构成一个处理，这样不同处理和试验小区都有不同配比的 N、P、K 施用量。显然，多因子试验比单因子试验更复杂且工作量较大。但是多因子试验不仅可以了解各单因素的单独效应，还可以了解不同因素之间的互作效应。

按试验地点可把试验分为田间试验和室内试验。**田间试验**就是在田间或野外所进行的生物观测活动。其试验条件如温度、光照、湿度、病虫害等常常难以控制，但能反映观测对象在自然生态系统中的变化规律。在田间试验中选择地点的代表性是十分重要的。**室内试验**如花粉培养、组织培养、水培试验、砂培试验、盆栽试验、室内化学试验和物理试验等常常可以控制许多环境条件，较能反映观测对象在特定条件下的变化规律，有利于分析各因素的特殊作用。尽管不同生物分支学科对这两类试验各有侧重，但通常情况下它们相辅相成，都为生物研究所必需。

三、生物试验的基本要求

无论何种类型试验，其试验条件都是不可能完全控制的。因此，为了有效地进行试验，准确获取结果，我们必须注意以下几个方面。

1. 试验项目立项正确

首选生产实践和科学试验中急需解决的重要问题，并从发展的角度出发，适当照顾到长远的或不久的将来可能突出的问题。只有重要的问题才值得我们投入人力、物力、财力去研究。例如当前森林中普遍暴发松毛虫灾害，使松树危在旦夕，必须立即研究和试验出有效的杀虫剂及其最佳使用浓度和使用方法。因为杀虫剂的使用也会同时杀死一些天敌，导致生态平衡的失调，所以此时还要考虑从生态学角度进行综合治理及相关天敌研究等问题，要考虑在消灭松毛虫后可能还会有另一种害虫上升为重要害虫，如何采取相关预防措施，等等。上述这类研究项目都很重要，并且非常有意义。

2. 试验目的要明确、设计要合理

（1）对试验结果和作用要大致了解，这样才能较好地确定试验考察因素和各因素的处理水平。比如N肥增产试验，可设置不同施N量处理水平进行栽培试验，目的是找出一个最佳施N水平，以便经济有效地获得高产。对此，我们需要知道大约要多少用量才有效，各处理至少要相差多少才有益。由于工作量的限制、不能对所有处理水平都进行试验，需要对此作出适当的选择，并使试验充分有意义。比如N肥增产试验中，如果选择施N水平为1，2，3，4（kg/亩），[①]则可能没有意义，因为如此少量的施N量变化可能对产量没有显著影响。根据专业知识和种植经验可以判断，选择施N水平为100，200，300，400（kg/亩），也可能没有意义，因为如此多的施N量可能使植株普遍过度生长或被烧伤、死亡，产量普遍下降、甚至绝收，达不到试验目的。若选择10，20，30，40（kg/亩），则可能合适，因为这些施用量水平中可能有一个最佳的施用量可使产量最高，由此可能找出一个能获得较高产量的施N量，从而达到预期目的。一般而言，当所设置的处理水平范围包含最佳处理水平而且各处理水平之间刚好达到显著差异时，这样的处理水平设计最好。

（2）试验设计要遵循三大基本原则：设置重复、随机排列、局部控制，即重复次数至少3次以上，各处理和重复随机排列，采取一定的控制方法（局部控制）使不同处理之间在试验的外界条件上保持一致。

3. 试验条件和试验材料要有代表性

这对于试验结果的可利用程度有重要意义。若试验结果是为了选出能在某一地区推广的新品种和新技术，则试验条件应尽可能与该地区的常规生产和应用条件相一致。比如，贵州师范大学植物遗传育种研究所要选育出一个新品种在贵州推广，不仅选育过程尽可能在贵州进行，而且所选出品种必须在全省有代表性的多个地点同时进行比较试验（如参加统一组织的贵州省区域试验）。只有在各点或大多数地点都表现好时，才能在全省推广。

4. 试验结果要可靠

这也就是说，试验的准确度和精确度要高。**准确度**是指观察值与相应真值的接近程度。观察值与相应真值越接近，则试验就越准确。但是，一般试验是不知道真值的，故准确度难以确定。**精确度**是指试验中重复观察值彼此的接近程度，即试验误差的大小。它是可以计算的。试验误差越小，则处理间的比较就越精确。当试验没有系统误差（由试验体系如测量工具等引起的误差，常常使观察值或结果普遍增加或减少一定数值）时，精确度与准确度一致。因此，在试验的全过程中，尽量准确地执行各项操作，力求避免人为错误和系统误差。特别要注意试验条件的一致性，以减少误差，提高试验结果的可靠性。

5. 试验结果要能重复

在相同条件下再进行同一或类似试验，要能获得相似的结果。通常情况下，上述各方面都满足要求时，一般可获得相似结果，即试验结果有重复性。

---

① 本书使用亩作为土地面积单位，1亩=1/15公顷=666.666 7平方米。

第三节　试验误差及其控制途径

一、试验误差的概念

影响生物生长发育的因素非常多，在一次试验中难以对它们都进行研究。因此，在生物试验中一般只能选择影响生物的某些因素作为**处理因素**。这些处理所得结果或观察值包含真实效应和非真实效应两方面。**真实效应**是该处理因素本身所能产生的效应。**非真实效应**是许多非处理因素的干扰和影响所产生的效应。例如，N 肥增产试验中，真实效应是 N 肥本身的增产效应；非真实效应是除了 N 肥以外的其他因素如 P、K、微量元素、病虫害、温度、湿度、光照等非处理因素对试验产生的偶然影响，它可使观察值偏离 N 肥效应值（真值）。这种使观察值偏离试验处理真值的偶然影响，就称为**试验误差**（error）。它影响试验的准确度和精确度。试验误差是衡量试验精确度的依据。误差小表示精确度高，误差大则精确度低。显然，只有在误差小时，才能对处理效应作出正确可靠的评价。误差大，可靠性低，甚至有时会导致误差效应超过处理真实效应，而使处理效应不明显。现代科学试验的特点是注意试验设计与统计分析的密切关系，有效地减少和准确计算出试验误差，从而合理地评价试验处理效应，科学地得出结论。

二、试验误差的来源

由于影响试验的因素难以全部控制，因此误差是不可避免的。尤其是田间试验，由于生物受大量的、难以控制的、自然环境条件的影响，其误差常常比物理化学试验误差大得多。为了减少误差，了解误差的来源具有重要意义。

生物试验误差的主要来源如下。

（1）试验设计存在缺陷，设计不科学，使试验容易产生误差。

（2）试验材料固有的差异。试验中各处理的供试材料在遗传和生长发育上或多或少地存在差异，可影响试验结果。

（3）试验操作上的不一致性。如操作人员变动，操作过程不标准、不规范，观测时间、观测方法和所用仪器等方面的不完全一致。

（4）试验外界条件的差异，即试验处理因素以外的各种因素和条件的差异。如农业试验中土壤差异、肥力不均、光照不均、病虫害侵袭、人畜践踏、风雨影响等，它们具有随机性，各处理所受影响不完全相同，导致误差。

上述各项都可在一定程度上影响试验结果，产生误差，甚至导致完全错误的结果。

值得注意的是，试验误差与试验错误是完全不同的。试验错误比试验误差在性质上要严重得多。试验错误或人为失误可导致试验完全失败，必须避免，通常也是可以避免的。

三、试验误差的控制

由于试验中存在许多人为难以控制的因素，试验误差是不可避免的，但是通过一些措施，完全可以减少各种来源的差异，降低试验误差。

控制试验误差的途径主要有如下几个：

（1）试验设计必须合理和完善。一般按生物试验要求进行的生物试验设计，误差较小。

（2）选择同质一致的材料。供试材料的基因型和发育状况要一致。尽量选择纯系、遗传稳定或高度一致的材料进行试验。对于发育上的一致性，可按发育状态分组。若材料很多，可选择同一组的材料用于试验。若材料不足，可以将各组材料按比例混合后用于各处理。这样可使各处理间具有较为一致的材料。例如，对于水稻秧苗，尽管是同一品种同一块地上培育的秧苗，其发育总是存在差异，有的是三叶期，有的是四叶期，有的是五叶期；为了试验的准确性，可按发育期分选成3组，即三叶期组、四叶期组、五叶期组；若材料较多，可选择其中的一组做试验，若材料较少，可按相同比例混合这3组幼苗，形成各处理材料，用于试验。

（3）改进操作、管理技术，使之标准化。总原则是：操作要仔细，尽可能使各处理间和各重复间完全一致。

（4）在试验环境、试验条件上尽可能确保各重复之间和各处理之间一致。可以使用局部控制等方法，首先确保区组内各处理之间的环境条件一致。

第四节　常见的几种生物试验设计

一、完全随机试验设计

若试验外界条件较为一致或对外界条件的特点不了解时，各处理和重复间可以用**完全随机设计**的方式进行安排。具体实施方法就是将各处理和重复编号，采用完全随机的方法（如抽签或使用随机数字表），将这些编号进行随机排列。由于编号过多时难以实现随机化，因此本设计适用于处理和重复数不很多的情况。

二、随机区组试验设计

如试验外界条件较为一致或呈现一定的变化趋势，可把试验环境剖分为几个试验条件基本相同的区域，采用随机区组试验设计方法进行试验。该设计要求设置足够多的重复，重复数通常等于或多于3个。各处理的一个重复构成一个组，称为一个**区组**。各区组内处理间随机排列。各区组可以分别位于环境条件彼此不完全一样的不同试验区域中。但必须使同一区组内的不同处理被安排在同一个环境条件相同的区域中，这种安排称为**局部控制**，即要求同一区组内的处理都处于尽可能相同的外界条件下。例如，N肥增产试验，可设三个处理水平，如10，15，20（kg/亩），分别用1，2，3表示。每个处理水平设三个重复。将每个处理水平的一个重复安排在一个区域中，形成一个区组，即共有三个区组，如区组1为2，1，3，区组2为3，1，2，区组3为3，2，1。各区组中处理水平间是随机排列的。根据实际情况，尽可能使同一区组中的不同处理水平处于同一条件下（即实施局部控制），如同一区组内各处理都在山坡的同一等高线上、土壤条件一致、它们都以相同的情况获取光照等。

三、正交试验设计

当试验因素超过2个时，试验处理数就会大幅度增加，而使试验规模过大，这将导致试验非

常困难、成本也大幅度增加。此时，最好采用正交试验设计。正交试验设计是利用正交表科学地安排多因素不同处理水平试验的方法，能以较少的试验处理数，获得较多的试验信息，较简单、快速地找出多个因素的最佳处理水平或最佳工艺条件等，是非常值得推广和应用的一种试验设计方法，详见第四章。

上述三种试验设计都有一定的随机性，是现代科学试验中最普遍采用的方法，建议广泛推广应用。

但是，有时人们也常采用一些非随机的试验方法，如拉丁方设计、对照和处理相间排列等。这些非随机的试验设计，有时也能获得良好的试验结果。

第五节　生物试验中常用的抽样方法

由于研究目的是为了获得一般规律，所以总体常常是无限总体。对此，我们只能抽取一定样本去推断总体特征。那么，哪种抽样方法才是最好的抽样方法呢？毫无疑问，**只有用与总体相适应的抽样方法抽样，所抽取样本才能最好地代表总体。也只有最能代表总体的样本才是最好的样本。**在对总体特点不了解时，一般用随机抽样方法抽样，所获样本有较好的代表性。当对总体有一定了解时，可对随机抽样方法进行适当改进，甚至采用非随机的抽样方法，也可以改善样本的代表性。常用的抽样方法简介如下。

一、随机抽样法

随机抽样法是在满足随机性的条件下从总体中抽取样本的方法。根据随机性的程度，又可分为完全随机抽样、组内随机抽样和组间随机抽样等方法。

1. 完全随机抽样

这是一种在总体中以完全随机的方法抽取个体、构成样本的抽样方法，其特点是总体中每一个体被抽取的机会均等。实施该抽样方法时，可采取抽签法，即将总体中每个个体逐一编号，并写在纸片上或签条上，将签条或纸片完全混合后，不加选择地从中完全随机抽取一定数目的个体，然后逐个测试、考察、记录，形成样本资料。完全随机抽样法是其他抽样方法的基础，所产生样本有较好的代表性。但是，当总体很大时，对个体编号较为困难。因此实际工作中很少单独采用此法。

2. 组内随机抽样（也称为分层抽样）

它是先将总体中的个体按某种属性特征分成若干组，然后在各组中按比例进行完全随机抽样，抽取一定数目的个体形成样本的抽样方法。其分组的依据可根据研究对象而定，如人的性别（男女）、人的年龄、植株的高矮、麦穗的有芒和无芒、小麦籽粒的角质率高低等。此抽样方法要求：①分组界限要清楚，使组间差异尽可能明显，组内差异尽可能小；②适宜的组内个体数目及要抽取的个体数目；③分组数目不宜太多，否则会降低组间差异。显然，组内随机抽样可增加样本的代表性，较适宜于总体情况复杂、个体数又较多的情况。

3. 组间随机抽样（也称为整群抽样）

该方法先将总体划分为若干组或群，各组或群所含个体数一般是相同或相近的，然后按完全随机的方法抽取若干组或群作为样本。此法的特点是，组间或群间随机，组内或群内不随机。显

然，此法的准确性不如上述两种方法，但是较为省时省力，较适宜于大范围的抽样。

二、非随机抽样

非随机抽样是指采用非随机的方法抽取样本的抽样方法。常用的非随机抽样方法有 2 种。

1. 典型抽样

典型抽样是按研究目的和总体特点，从总体中有意识地、有目的地选取较能代表总体的典型个体或个体群作为样本的抽样方法。比如，小麦田间测产的抽样，目的是了解此小麦品种在此地正常情况下的产量。由于全田面积很大、生长起伏有明显的差异，有的部分甚至遭受意外损害如人畜践踏、植株倒伏、土壤严重肥力不足等，对此可以目测有代表性的田块或地段取点测产。显然，此法较依赖于调查工作者的知识和技能，结果很不稳定，而且不符合随机原理，无法估计抽样误差。但是在特定条件下，用此法抽样比随机抽样更能获得代表总体特点的样本。此法较适宜于从容量巨大、遗传或生长不够一致、部分遭受意外损害的总体中抽取少数个体作为样本的抽样。另外，在品种的提纯复壮中也常常采用这种方法。

2. 顺序抽样（又称为机械抽样）

这是一种按既定顺序抽取一定数目的个体或个体群构成样本的抽样方法。例如，在总体编号中逢单抽样、逢双抽样、逢 1 抽样、逢 5 抽样、逢 10 抽样等。在田间抽样调查中常用的五点式抽样、对角线式抽样、棋盘式抽样、分行抽样、平行线式抽样、“Z”字形抽样等田间抽样方法，均属此类。此法显然是不随机的，但是它抽取的样本能兼顾到总体的各个部分，又简单易行，故而常用。分布和表现较为一致的总体按此法抽样可获得较好的代表性。

上述抽样方法的比较见图 1.5.1。

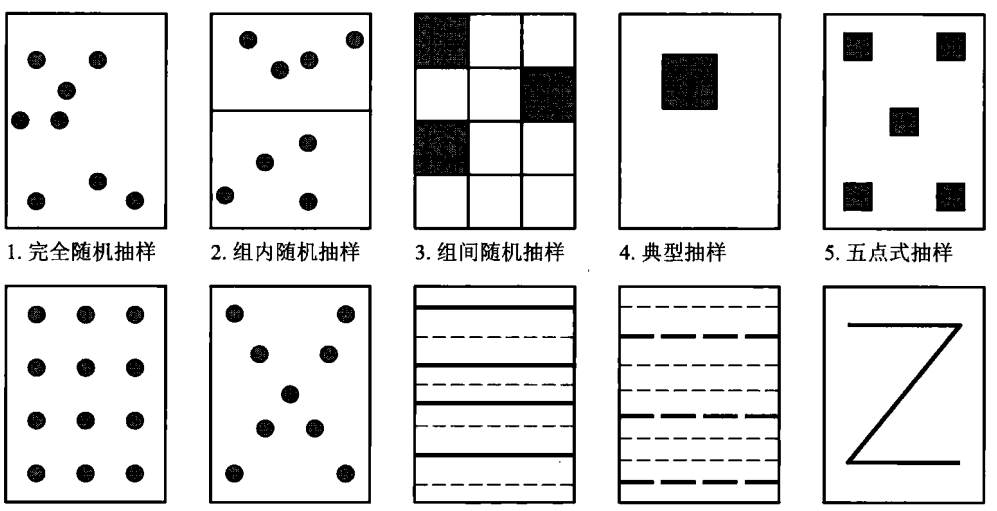


图 1.5.1　几种抽样方法的比较

第六节　生物试验资料的整理

一、资料的类型

按照观测方法的不同，可以将数据资料分为两类。

1. 间断性变数资料

间断性变数资料是用计数方法所获得的数据资料，如基本苗数、穗数、籽粒数，动物头数，细菌的个数，生物细胞数，植物花粉粒数，虫头数，等等。其每个观察值必须用整数来表示，相邻整数间不容许有小数存在，观察值间是不连续的，这种变数就称为间断性变数。由此构成的资料，即间断性变数资料，也称为**不连续性变数资料**。

2. 连续性变数资料

连续性变数资料是用称量、度量和测量等量测方法所获得的数据资料。其各个观察值不限于整数，可以是任何小数，数值间是连续变异的，如小麦穗粒重 2.0 ~ 4.0 g，可以有 2.35 g，2.10 g，2.95 g 等无穷多个数值存在。其小数位数的多少，因计量的精确度不同而异。这种变数称为连续性变数。由此构成的资料，就是连续性变数资料。例如，植株高度、作物产量、蜜蜂采蜜量、籽粒的蛋白质含量及氨基酸含量、奶牛产奶量及体重等资料，都属于此类。

在遗传学上常常把生物性状按其遗传特点分为两类，即数量性状和质量性状。对于数量性状可以用计数和计量两种方法来进行观察。其中用计数观察所得资料，如亩穗数、分蘖数、穗粒数等属于间断性变数资料；而用计量观测所得资料，如高度、产量、体重、粒重、蛋白质含量、氨基酸含量等则属于连续性变数资料。对于质量性状，由于它们是不能量测的，通常表现为某种属性，如颜色、毛等的有无。一般采用计数的方法统计具有某属性的个体数目。例如，在 320 株水稻构成的 F_2 群体中，考察得知有 240 株为紫色柱头，有 80 株为黄色柱头。由于质量性状是属性性状，只有有和无这两种情况；有时可有几种情况如紫色、红色、粉红色、白色等。对此，在统计学上，可以把具有某属性当做“1”，不具有当做“0”，或者把不同的几种属性分别当做“0”，“1”，“2”，“3”…。于是就构成由 0 和 1、或 0，1，2，3…所组成的数据资料。显然，这类资料属于间断性变数资料。

二、资料的整理

无论何种生物资料，都包含有很多观察值，有时甚至是成千上万的。面对未加整理的一大堆混乱的数据，很难从中得出明确的概念和认识。对此，可以将它们按数值大小进行排列分组，制成频数分布表和频数分布图，就可以看出观察值的分布规律。不管是间断性变数资料还是连续性变数资料，都可以整理成频数分布表和频数分布图。

1. 频数分布表

对于取值较少的间断性变数资料，很容易按所取的数值（或为某种属性）的不同直接归类、分组，然后统计各组观察值的频数，就形成由不同取值（或不同属性）与频数构成的频数分布表。

例 1.6.1　某豌豆遗传试验中以圆粒、红花品种作母本，与皱粒、白花品种进行杂交，杂种植株表现为圆粒、红花。杂种自交产生 F_2 代群体。调查该群体各植株的上述性状。所得数据资料