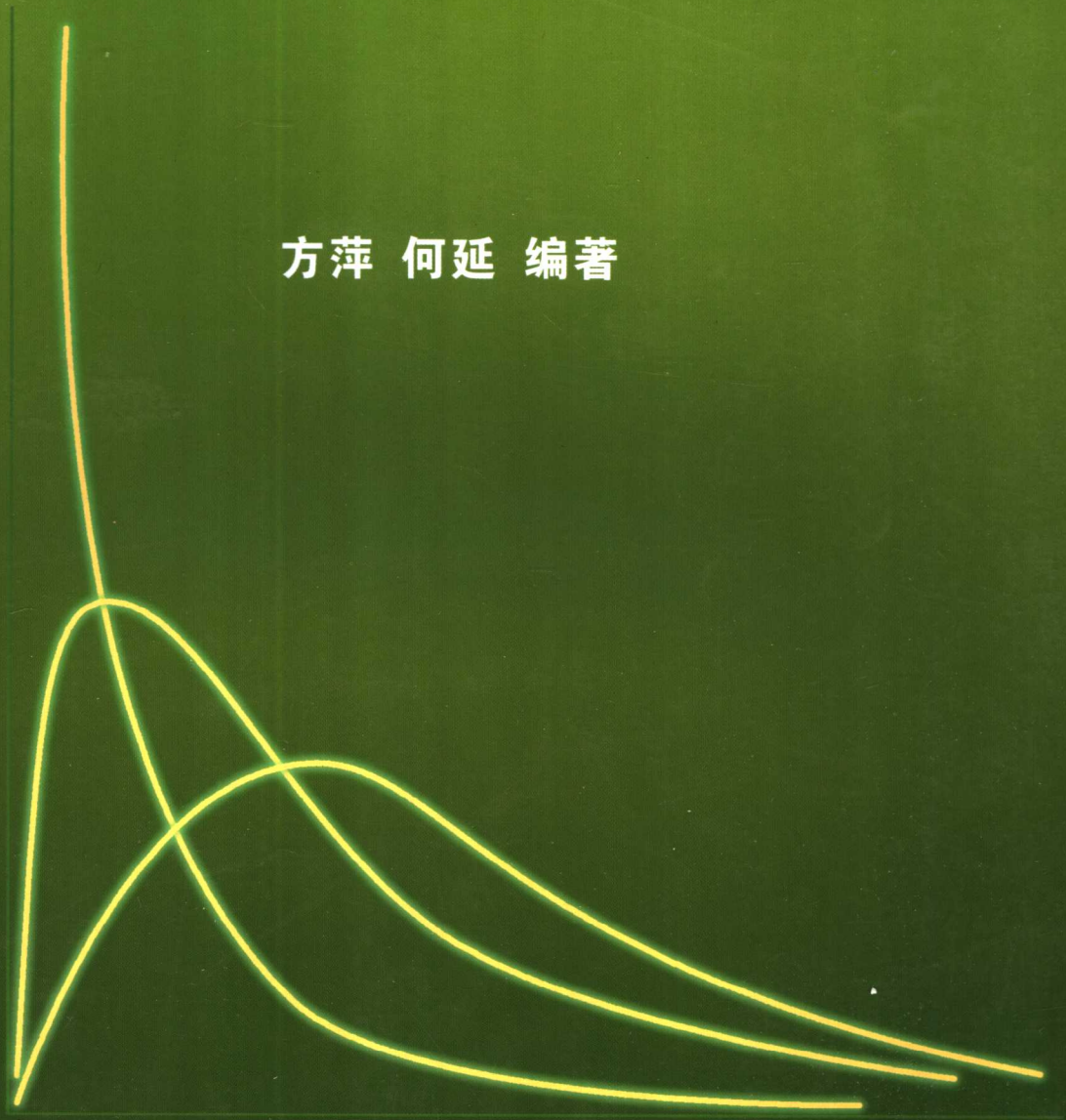
试验设计与统计

方萍 何延 编著



浙江大學出版社

试验设计与统计

方 萍 何 延 编著

浙江大學出版社

图书在版编目（CIP）数据

试验设计与统计 / 方萍，何延编著. —杭州：浙江大学出版社，2003.6
ISBN 7-308-03313-9

Ⅰ. 试... Ⅱ. ①方...②何... Ⅲ. ①统计—试验设计(数学)②统计分析 Ⅳ. C81

中国版本图书馆 CIP 数据核字（2003）第 032573 号

责任编辑 杜玲玲
封面设计 宋纪浔
出版发行 浙江大学出版社
（杭州浙大路 38 号 邮政编码 310027）
（E-mail:zupress@mail.hz.zj.cn）
（网址：http://www.zjupress.com）
排　　版 浙江大学出版社电脑排版中心
印　　刷 浙江大学印刷厂
开　　本 787mm×1092mm 1/16
印　　张 21.5
字　　数 550 千
版 印 次 2003 年 6 月第 1 版 2003 年 6 月第 1 次印刷
印　　数 0001—1000
书　　号 ISBN 7-308-03313-9/C・182
定　　价 30.00 元

内容简介

本书是专门为环境与资源相关专业本科生开设“试验设计与统计”课程而编写的。全书共分九章，在简要叙述统计学的基础知识（如：试验资料整理、统计特征数的意义及计算、概率与概率分布、抽样分布）的基础上，着重介绍了环境与资源研究中常见的试验设计和统计分析的基本原理与方法、常用的多元统计分析方法及相应的统计软件应用。主要内容包括：成组和配对设计的对比试验结果差异显著性 t 测验、参数区间估计、方差同质性检验、数据资料正态性检验、计数资料的 χ^2 检验、非参数检验；完全随机设计、随机区组设计及拉丁方设计、裂区设计、再裂设计及含有假伪处理的裂区设计的多因素和（或）单因素试验结果的方差分析与多重比较，正交设计及其试验结果的统计分析，协方差分析，一元与多元线性回归与相关及曲线回归分析；简要介绍了逐步回归分析、通径分析、聚类分析、判别分析、主成分分析、因子分析、典型相关分析和时间序列分析；特别介绍了 Excel 和 DPS 数据处理系统提供的相应统计分析软件的应用方法。

本书具有以下特点：通俗易懂，避开了繁琐的公式推导而侧重于应用方法介绍；内容丰富，它集基本数理统计方法与多元统计于一体；专业应用性强，它可供环境和资源相关学科及农学类等学科作为教材使用，也适合生命科学、农业科学、医学等领域的科技工作者阅读参考。

前　言

随着科学技术的不断进步，合理地设计试验，正确地分析试验结果，特别是熟练地运用计算机软件进行试验设计与统计分析，已经成为高素质科技人才的必备技能。本书是特地根据环境与资源相关专业对试验设计与统计分析方法的要求而编写的。

本书的写作是在作者多年从事试验设计与统计分析的教学和应用研究的基础上完成的。全书共九章，可概括为六部分：第一部分（第一、二章），为试验设计与统计分析的基本原理，主要介绍试验资料的统计描述、统计假设检验的基本原理与各种统计假设检验方法。第二部分（第三、四章），介绍试验设计和抽样调查的基本原理、类型及方法。第三部分（第五、七章），为各种比较性试验的设计及试验结果统计分析方法，是本书的重点之一，其核心内容是方差分析。第四部分（第六章），为变量之间相互关系的统计分析，主要内容有一元线性回归与相关、多元线性回归、多项式回归、可直线化的曲线回归分析等。第五部分（第八章），简要介绍多元统计分析方法，包括逐步回归分析、通径分析、聚类分析、判别分析、主成分分析、因子分析、典型相关分析和时间序列分析。第六部分（第九章），在介绍运用 *Excel* 电子表格数据处理软件进行统计分析的方法和技巧的基础上，重点叙述应用 *DPS* 数据处理系统进行试验设计与各种统计分析方法。

书中内容侧重于各种统计方法的应用，在统计原理方面，一般只给出概念的介绍和公式的简单推导，而避开复杂繁琐的公式推导。有关统计运算方法，本教材不再采用以往教材介绍的适用于笔算的冗长公式和繁琐的算法，而是介绍思路清晰、方法简单的数据整理和计算器中数据的输入与输出过程。目的在于使读者通过本书的学习，掌握试验设计的基本原理与方法，了解有关统计分析方法的原理与适用条件，能熟练地运用 *Excel*、*DPS* 数据处理系统等统计软件进行试验设计与统计分析，并对统计结果作出正确的专业解释。

在本书的编写过程中参考了国内外的有关书籍和资料（主要书目列于书后），引用了其中的一些内容和实例，在此，对所有作者和译者表示诚挚的感谢。本书的编写得到了浙江大学教材建设委员会和环境与资源学院领导的大力支持；浙江大学理学院陈叔平教授对本书的编写提出了很多建设性的意见；浙江大学出版社领导和工作人员特别是杜玲玲编辑对本书的出版给予了热情的支持和帮助，在此，一并表示衷心的感谢。

由于编者水平所限，虽经反复修改但错误和不妥之处在所难免，恳请读者批评指正。

方　萍

2003 年 1 月于杭州华家池

目　　录

第一章　统计学基础知识

试验设计与统计是运用数理统计原理和误差理论，分析和解释环境与资源科学中的数量关系，帮助研究者正确设计试验和科学分析试验结果，从而揭示环境与资源问题真相的学科。它属于统计学范畴。本章重点介绍总体、样本与误差的基本概念，总体与样本的统计特征数和概率分布与抽样分布的意义及性质，为进一步学习试验设计以及各种统计分析方法奠定基础。

第一节　几个常用统计术语

一、总体与样本

1. 总体

试验研究离不开研究对象。某项试验研究的具体对象的全体称为总体，它是由许多客观存在的具有某种共同性质的总体单元所构成的集合体。构成总体的每个单元称为个体。对个体的某种性状加以考察（如称量、度量、计数或分析化验）所得的数值，称为观测值。总体所包含的个体数目（N）称为总体容量。总体具有以下三个特征：

（1）同质性。同一总体的各个体，必须在某一方面具有相同的性质，才能把它们集合起来，构成某种性质相同的总体。所谓同质，不是绝对的而是相对的，是随研究目的而变化的。例如，西瓜中心糖度是西瓜的重要品质指标，在筛选优质西瓜品种的研究中，每一西瓜品种的中心糖度就是总体的一个个体。在研究施氮量对西瓜品种浙蜜1号品质的影响时，每一施氮水平下浙蜜1号西瓜品种的中心糖度是一个总体单元。

（2）变异性。在构成总体的单元具有同质性的前提下，不同单元之间一般都存在差异，即变异性。这种同质性和变异性是由事物的客观性所决定的。也就是说，凡客观事物都是同质性和变异性的对立统一体。一个统计总体，如果没有同质性便不成其为总体，若没有变异性则无需统计。对总体的统计研究，实际上就是研究组成这些总体的个体之间的变异情况。例如，要研究某地区的土壤重金属镉的污染状况，就需要研究该地区不同地块的土壤镉含量变异情况。在同质性的基础上研究总体的变异程度、集中趋势及规律，就是统计分析的重要任务之一。

（3）大量性。统计研究的目的在于揭示事物的客观规律性，而这种规律性只能在大量事物的普遍联系中表现出来，所以总体是由大量个体构成的。若总体容量无穷大，则称其为无限总体。如大气、水体、土体等连续体是无限可分的，因此总体单元数是无限的。若总体容量是有限的，则称其为有限总体。如某一时刻的全国人口是一个容量很大的有限总体。

对总体的研究可以作全面普查，如全国人口普查、土壤普查等；但全面普查往往耗资巨大，难以实施，特别是对于无限总体或有限大总体，全面普查是不现实的，加之有些观测手段具有破坏性，即使总体容量不太大，也不允许对所有个体加以一一考察，因此，多数情况下只能采用抽样调查的方法。

2. 样本

从总体中抽取一部分个体所组成的集合称为样本。由样本特征来推断总体性质是统计分析的基本手段。为此，样本必须对总体具有代表性，这要求抽样满足随机抽样的要求：(1)等可能性，即每次抽样时各个体具有同等机会被抽取；(2)独立性，即每次抽样不影响下次抽样时各个体被抽取的机会。

样本所包含的个体数目称为样本容量，常记作 n。$n \geqslant 30$ 的样本称大样本；$n < 30$ 时称小样本。有时大样本与小样本具有不同的统计分布特征。

二、变数与数据资料

由于受许多偶然因素影响，总体内部个体之间普遍存在着变异性，因而观测值之间也会表现出波动性。例如，将一块田划分为等面积的不同小区，种植同一品种的水稻并进行相同的田间管理，收获时分区测产，测得各小区的产量会因土壤肥力的不均一性以及其他偶然因素的影响而不尽相同。又如，当重复 10 次测定某一土样的镉含量，其结果也会因测试仪器、测定条件及测试者的操作技能等一系列因素的影响而表现出一定的波动性。这种受许多偶然因素影响而表现出波动性的数量称为随机变量或随机变数，简称变数，常用 X、Y、Z 等大写字母表示。例如作物产量、水体的生物耗氧量、土壤的汞含量等都属于随机变量。

某随机变量的一组具体观测值称为资料，例如测定 10 个土壤样品的全氮含量，将获得这 10 个土壤全氮含量数据的资料。由于测试或调查的手段、工具、方法、对象不同，获取数据资料的性质有很大的差异，有的是数据资料，有的则是文字描述资料。因此，这些资料按照其性质或特性可分为两大类：

1. 数量属性

数量属性是指测试、调查的对象具有可度量或计数的性质。例如，环境污染的面积、污染物的浓度、地下水中硝酸盐的浓度、水体中细菌的数量等。根据个体间数据差异的性质，数量属性资料又可进一步分为连续性变异和间断性变异两类。连续性变异资料是指个体间数据存在很小差别，当总体足够大时，随着度量精度的提高这种差别可达到人为可测的任意精度。一般这类资料是通过称量、度量、测量或分析测试获取的，其取值精度取决于量测工具的精密度。间断性变异资料是指用计数的方法所得到的数据资料，其取值只限于非负整数，例如，环境中活有机体总是以自然数计量的，尽管总体无限大，但两个类群有机体的差数不会小于 1。

2. 质量属性

在环境与资源研究中，有些观察调查对象的一些属性能观察而不能度量，如污染物的不同颜色、不同气味等都是质量属性。可采用赋值法使质量属性数量化，即对某种属性的不同类别赋予不同的数码，例如对污染水体的不同颜色可赋予不同的数值，如取红色为 -1，无色为 0，绿色为 1，等等。通过对质量属性赋予数值后，就可以进一步用统计方法处理所研究的环境或资源问题。

三、参数与统计数

在同质性的前提下总体具有变异性和大量性的特性，因此，统计学上需要用一些参数来反映总体内部个体间的变异程度或集中性趋势等特征，如总体平均数、总体方差、总体标准差等统称为总体参数，简称参数。相应地，利用样本资料计算得到的用于描述样本内部个体间的变异程度或集中性趋势等特征的一些指标，如样本平均数、样本标准差等称为样本统计数，简称

统计数。统计数用于估计相应的参数。参数与统计数有时统称为统计特征数。

四、误差与错误

在一定条件下某客体物所具有的真实数值即为真值。由于受测定过程中许多偶然因素或人为因素的影响，对该客体物进行量测所得到的观测值与其真值之间会有一定的偏差，即误差。任何试验结果都具有误差，在一切科学试验过程中自始至终存在误差，这称为误差公理。根据误差产生原因的可知性，误差可分为系统误差和随机误差两类。

1. 系统误差

系统误差是由某种确定的原因所引起的误差。其特点是在相同条件下重复测定时，以相同的大小和正负性重复出现，因此系统误差是可以测定并校正或消除的。在环境与资源领域的研究中，系统误差的主要来源有以下几方面。

(1) 方法误差：因分析方法本身不够完善所造成的误差。如重量分析中沉淀不完全或有共沉淀现象等，滴定分析反应不够完全，比色分析中存在干扰离子等，都会使测定结果偏高或偏低。

(2) 仪器误差：因所用仪器不够精确而引起的误差。如滴定管、容量瓶的刻度或仪表的刻度不准确等。

(3) 试剂误差：由于试剂不纯造成的误差。不纯的试剂或蒸馏水常会向被测物引入干扰物质，使测定结果偏高或偏低。

(4) 操作误差：由于分析人员主观方面的原因，使实际操作与正确的操作有一定出入而造成的误差。例如滴定时过早读数，坩埚灼烧后没有冷却至室温就称量，称量时未能防止样品吸湿等。因操作人员的生理条件限制或习惯上的差异也会引起操作误差。如对滴定终点的颜色变化敏感程度不一，有人偏深有人偏浅，就会造成结果的差异。

(5) 田间试验中土壤肥力的方向性变化，会造成试验地特定部位的供试作物产量偏高或偏低。

由于系统误差是重复地以固定形式出现的，因此不能通过增加重复测定次数加以消除。对于分析测试中的系统误差可以通过对照实验、空白实验、校准仪器等办法校正。对于田间试验中土壤肥力的方向性变化可通过随机区组设计加以局部控制。

2. 随机误差

随机误差又叫偶然误差，是由很多不可避免且无法控制的偶然因素引起的误差。就分析测试结果而言，这些偶然因素来自分析方法本身、仪器、环境、操作等各个方面。例如坩埚称重时，即使严格控制冷却时间，但因空气湿度随时都在变化，各次称量可能有微小差异。控制滴定终点时，由于视觉判断能力的限制，滴定终点不可能完全一致。读取滴定管体积时虽尽了最大努力，但仍有±0.1ml 的允许误差。对生物试验结果而言，这些偶然因素就更为复杂，例如由于供试材料的不均一性，种子质量、秧苗素质就不可能完全一致；光照、温度、湿度等影响生长的环境因子也可能随时随地发生变化；农时操作的不一致性以及其他不可预测的自然或人为因素的干扰均会使试验结果产生误差。

这些偶然因素的作用，可能现在有，过一会就没有，今天有明天可能又没有；对平行样品进行测定时可能这一份的影响大些，那一份则小些。每个偶然因素引起一部分误差，几种偶然因素随机组合，总的引起一次测定的偶然误差。因此随机误差的特点是：产生原因不确定，其误差大小无规律性，有时大有时小，有时正有时负，不具“单向性”或“重现性”，所以随机误差也称不

可测误差。随机误差虽不可避免，也不能校正，但若在同样条件下对同一试样进行多次测定，就会发现随机误差的出现是服从统计规律的。因此，可以利用数理统计方法对试验数据进行分析处理，通过增加重复试验次数来减小随机误差对试验结果的影响。

在试验过程中由于工作出错造成的观测值与真值的差异，称为疏失误差，即错误。错误与误差在性质上是两个完全不同的概念。由于其产生原因是疏忽大意、操作不当等主观因素，如因违反操作规程而导致测定错误或数据抄错、算错等，观测值将明显歪曲事实，无规律可循，成为坏值或异常值，对于这些数值必须予以剔除。因此，我们在试验研究过程中必须养成认真细致的工作习惯，保持严谨的科学态度，杜绝疏失误差的产生。

一组重复观测值中往往有个别数据与其他数据相差较大，这一数据称为可疑值或极端值，也就是离群值。如果可疑值不是由于明显的过失造成，就要用统计学的方法确定其取舍，具体方法将在第二章第十一节介绍。

五、准确性与精确性

准确性是指观测对象的观测值与其真值的偏离程度，偏离越小则试验越准确。精确性是指同一观测对象的重复观测值之间的彼此相符程度，即试验误差的大小，误差越小则试验越精确。所以，准确性与精确性是不同的概念。

在统计工作中，常用样本统计数来估计总体参数。因此，我们用统计数接近参数的程度来衡量统计数的准确性高低，而用统计数的变异程度来衡量统计数的精确性高低。可见，准确性不等于精确性。准确性表示观测值与真值的相符程度，精确性表示重复观测值间的变异程度。由于一般试验中真值为未知数，所以试验的准确性难以确定。精确性一般是指试验误差，是可以估计的。如何正确估计试验误差，并减小试验误差以提高试验精度是试验方法设计所要解决的核心问题。

例 1.1-1 某物体的质量为 0.1108g，甲、乙、丙 3 人分别对该物体进行三次重复测定，结果如表 1.1-1 所示。试比较三者测定结果的准确性与精确性。

表 1.1-1 甲、乙、丙三人对某物体的质量进行测定之结果的准确性与精确性比较 g

观测者	观测值			平均	与真值偏差	最大值与最小值之差
	1	2	3			
甲	0.1102	0.1107	0.1109	0.1106	0.0002	0.0007
乙	0.1101	0.1127	0.1115	0.1114	0.0006	0.0016
丙	0.1125	0.1126	0.1124	0.1125	0.0017	0.0002

由三人测定结果的平均数与被测物体实际质量的偏差可见，甲的偏差最小，乙次之，丙最大，所以甲测定结果准确性最高，丙最低，乙介于两者之间；从三人测定结果的变异幅度即最大值与最小值之差看，丙最小，甲次之，乙最大，因此，丙的测定结果精确度最高，甲次之，乙最低。

第二节 数据资料的初步整理

环境与资源科学研究中获取的原始资料在未加整理之前，往往是一堆杂乱无章的数据，无法立即从中找出规律性。在对资料进行分析之前，需先对它们进行整理和分组。整理时，首先

应根据资料的类型、性质或测定时间的不同进行分类，将性质相同的资料归纳在一起，使资料系统化，以反映事物的本质。因此，整理资料时只有坚持“同质”的原则，才能显示资料内部的规律性，得出正确的结论。资料按性质分类后，应进一步对同质资料进行分组。一般当资料中观测值在30个以下时，称为小样本，不必再分组，但可以对其排序，使观测值按大小次序排列，以便简洁地反映数据资料的变异情况。当观测值在30个以上时，称为大样本，需按观测值的大小分成若干组，编成次数分布表或绘成次数分布图，以直观明了地反映数据资料的分布特征。

次数分布表的编制方法是，将一群观测值的变异范围划分为互不相容的若干区间，记数归入各区间的观测值次数。由各组的组限（或组中值）及相应次数构成的表格称为次数分布表。下面结合具体例子说明次数分布表的建立过程。

例 1.2-1 试对100个甜菜块根蔗糖含量数据（表1.2-1）制作次数分布表和分布图。

表 1.2-1 100个甜菜块根蔗糖含量 %

编号	1	2	3	4	5	6	7	8	9	0
0	11.8	13.1	9.2	8.7	12.9	13.7	9.6	13.7	8.5	15.7
1	14.1	11.9	16.7	7.4	10.0	4.4	13.2	13.8	9.1	11.9
2	12.8	15.3	12.6	16.1	17.2	13.5	11.9	16.7	9.6	15.1
3	14.6	10.4	13.4	14.6	10.5	8.6	15.2	11.1	14.5	12.1
4	14.9	15.0	12.1	12.6	13.0	14.1	14.4	13.1	13.3	15.0
5	10.1	12.4	10.8	11.3	6.3	15.7	14.3	15.0	12.5	11.8
6	11.6	12.2	7.5	13.4	14.7	14.2	14.0	15.1	6.5	8.7
7	11.0	13.0	9.2	7.0	13.2	9.0	14.0	13.2	15.0	13.8
8	15.1	14.9	12.6	14.1	11.4	9.4	12.4	15.0	9.4	12.9
9	13.4	10.6	6.5	11.0	11.9	11.8	12.6	9.5	12.2	8.2

解

① 求变幅 R：

$$R=y_{max}-y_{min},$$

其中 y_{max} 为最大观测值，y_{min} 为最小观测值。表1.2-1中 $y_{max}=17.2\%$，$y_{min}=4.4\%$，所以，

$$R=17.2-4.4=12.8(\%)。$$

② 选择组数 k。可采用Sturge公式粗略估计：

$k=1+3.3\lg N$，其中，N 为总体容量，表1.2-1中 $N=100$，所以，$k=1+3.3\lg100=7.6$。实际工作中可以确定跟估计值较接近的一个整数，本例定为9组。

③ 确定组距 C，即每组的上限与下限之差，由变幅除以组数来估计：

$$C=R/k,$$

由于 $R=12.8\%$，$k=9$，所以，$C=12.8/9\approx1.4$。为应用方便，本例取1.5。

④ 决定组限和组中值：

组限即一个组所在区间的两个极端值，大的为上限，小的为下限。为观测值归组方便，组限的小数位数可比观测值多一位。最小组的下限记作 L_{11}，一般由最小观测值减二分之一组距来估计，即

$$L_{11}=y_{min}-C/2。$$

已知 $y_{min}=4.4\%$，$C=1.5$，所以 $L_{11}=4.4-1.5/2=3.65$，为分组方便本例取4.05；最小组的上