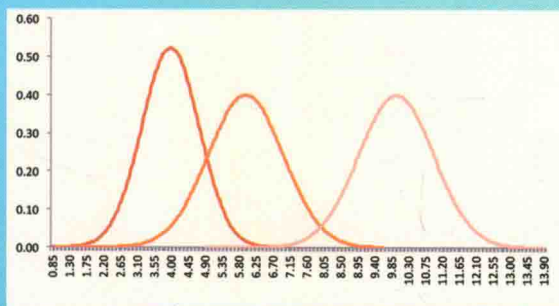
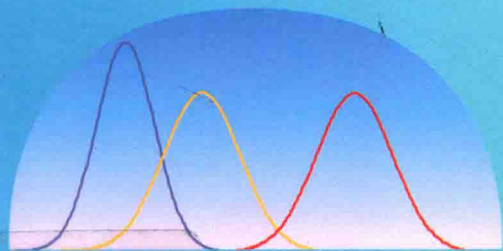
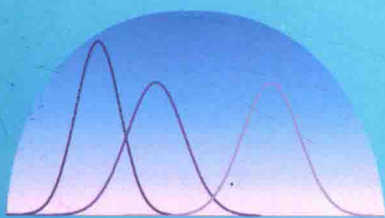




普通高等教育农业部“十二五”规划教材  
全国高等农林院校“十二五”规划教材

# 试验设计与统计分析

方萍 编著



普通高等教育农业部“十二五”规划教材  
全国高等农林院校“十二五”规划教材

# 试验设计与统计分析

方萍 编著



中国农业出版社

## 图书在版编目 (CIP) 数据

试验设计与统计分析/方萍编著. —北京: 中国农业出版社, 2014. 8

普通高等教育农业部“十二五”规划教材 全国高等农林院校“十二五”规划教材

ISBN 978-7-109-19433-5

I. ①试… II. ①方… III. ①试验设计—高等学校—教材②统计分析—高等学校—教材 IV. ①O212

中国版本图书馆 CIP 数据核字 (2014) 第 169760 号

中国农业出版社出版

(北京市朝阳区麦子店街 18 号楼)

(邮政编码 100125)

责任编辑 李国忠

---

中国农业出版社印刷厂印刷 新华书店北京发行所发行

2014 年 8 月第 1 版 2014 年 8 月北京第 1 次印刷

---

开本: 787mm×1092mm 1/16 印张: 19.5

字数: 470 千字

定价: 38.50 元

(凡本版图书出现印刷、装订错误, 请向出版社发行部调换)

# 前 言

## PREFACE

熟练运用相关软件,合理设计试验并对试验资料做出正确的统计分析,是现代高素质科技人才的必备技能。本教材专为农业资源与环境专业编写,也可用作资源与环境、农学类、生命科学类以及相关专业的教材,其目的在于使学员通过本教材的学习,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,并能熟练地运用 Excel、DPS 数据处理系统等统计软件进行试验设计与试验资料的统计分析。

本教材基于笔者执教试验设计与统计分析等相关课程 30 余年所积累的心得而编撰,其架构以笔者 2003 年编著的《试验设计与统计》<sup>[1]</sup>为基础,全书共 10 章,可概括为 6 个部分。第一部分,即第一章,主要介绍试验资料的统计描述、概率分布与抽样分布,是理解与掌握试验设计与统计分析方法所必备的基础知识。第二部分,包括第二章至第四章,在介绍统计假设检验基本原理的基础上,着重介绍基于正态分布、 $t$  分布和  $F$  分布的参数检验和参数区间估计方法,如单个或两个样本平均数及单个或两个样本频率的统计假设检验、单向分组资料的方差分析等;随后介绍包括符号检验、秩和检验及计数资料  $\chi^2$  检验等内容的非参数检验方法与原理;还介绍了数据资料的正态性检验、样本方差的齐性检验以及可疑值的取舍等方法。第三部分,包括第五章与第六章,主要介绍试验设计和抽样调查的基本原理、类型及方法。第四部分,包括第七章与第九章,主要涉及环境、资源和生命科学研究中常用的效应比较性试验结果的统计分析方法,其核心内容是方差分析,并介绍了协方差分析。第五部分,即第八章,介绍变量之间相互关系的统计分析方法,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。第六部分,即第十章,简要介绍逐步回归分析、通径分析、聚类分析、判别分析、主成分分析与典型相关分析等常用多元统计分析方法的基本原理与相应统计软件的应用。为满足知识体系的系统性要求,本教材所涵盖的一些内容已超出本科生试验设计与统计分析等相关课程的教学大纲范围,在教学实践中可选用其中有关章节;也正因如此,本教材还可为环境与资源、农学、生命科学等相关专业研究生高级试验设计与统计分析等相关课所选用,也可供科技工作者参考。

本教材具有如下特色。

①内容侧重于各种统计分析方法的应用，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，而避开复杂烦琐的公式推导。

②所介绍的统计分析方法思路清晰，算法便捷。首先，鉴于各种统计分析方法都会涉及观察值的离均差平方和或乘积和的计算，这是本课程教学过程中令学生倍感烦琐的环节，为此，本书不再采用以往同类教材介绍的适用于笔算的冗长公式和烦琐的算法，而是针对每一种设计的试验结果给出思路清晰的数据整理格式，并利用离均差平方和与方差及离均差乘积和与协方差之间的关系，借助 Excel 插入函数 VARP 及 COVAR 求得相应的离均差平方和或乘积和，以达到既有利于学生掌握各种试验结果的统计分析思路，又简化相应算法之目的。同时，由于 Excel 电子表提供了用于求算基于正态分布、 $t$  分布、 $F$  分布及  $\chi^2$  分布等统计量分布的临界值或概率的插入函数，如 NORMDIST、NORMSDIST、TINV、TDIST、FINV、FDIST、CHINV 和 CHIDIST 等，所以在介绍相关统计分析方法时，不再采用以往同类教材介绍的通过查诸如  $t$  分布临界值表、 $F$  分布临界值表、 $\chi^2$  分布临界值表等附表来获得统计量的显著性水平，而是直接借助这些插入函数来求算。因而本教材中也不再罗列相关附表。

③为便于学生在学习各种试验设计与统计分析方法原理的同时，能熟练掌握 Excel、DPS 等统计软件应用方法，在各章节中穿插介绍了相应统计软件功能的实现过程，并在每章后附有习题，这些正是本教材比 2003 版的重要改进之处。此外，本教材更为强调如何对统计分析结果的展示与表达。

在本教材的编写过程中，参考了国内外的有关书籍和资料，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。浙江大学唐启义教授对本教材的编写提出了许多宝贵意见；浙江大学农业资源与环境专业 2010 级陈妮、李名和周嘉文同学对书中的例题做了全面校验；本书的出版得到了浙江大学“农业资源与环境”特设专业建设项目经费的资助；中国农业出版社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，虽经反复校对修正，但错误和不妥之处在所难免，恳请读者批评指教。

方萍

2014 年 5 月 于杭州紫金港

# 目 录

## CONTENTS

### 前言

第一章 统计学基础知识 .....	1
第一节 几个常用统计术语 .....	1
一、总体与样本 .....	1
二、变数与数据资料 .....	2
三、误差与错误 .....	3
四、准确性与精确性 .....	4
第二节 数据资料的初步整理 .....	4
一、次数分布 .....	5
二、利用 Excel “直方图”工具制作次数分布表与分布图 .....	6
第三节 统计特征数 .....	8
一、表征数据资料集中性趋势的统计特征数 .....	8
二、表征数据资料变异程度的统计特征数 .....	9
三、利用 Excel 插入函数和“描述统计”分析工具求算常用统计特征数 .....	11
第四节 概率论基础 .....	13
一、随机事件与概率 .....	13
二、概率分布 .....	14
三、几种常见的概率分布 .....	15
四、求算相关分布的概率、概率密度值或分布临界值的 Excel 插入函数 .....	17
第五节 抽样分布 .....	18
一、样本平均数分布 .....	18
二、样本平均数差数分布 .....	20
三、 $t$ 分布 .....	21
四、 $\chi^2$ 分布 .....	22
五、 $F$ 分布 .....	23
习题一 .....	24
第二章 统计假设检验 .....	26
第一节 试验结果的直观分析 .....	26
一、试验结果直观分析 .....	26
二、试验数据波动原因 .....	27
第二节 统计假设检验概述 .....	27
一、统计假设检验的意义 .....	27

二、统计假设检验的基本步骤 .....	28
三、统计假设检验的两类错误 .....	29
四、双尾检验与单尾检验 .....	30
五、假设检验应注意的问题 .....	31
第三节 平均数的统计假设检验 .....	32
一、单个总体平均数的假设检验 .....	32
二、两个总体平均数的假设检验 .....	34
三、两样本平均数差异显著性检验的 Excel “数据分析”实现 .....	38
第四节 总体频率的假设检验 .....	41
一、单个样本频率（成数）与总体成数比较的 $Z$ 检验 .....	41
二、两个样本频率（成数）比较的 $Z$ 检验或 $t$ 检验 .....	42
三、“两样本率比较”的 DPS 实现 .....	44
第五节 参数区间估计 .....	46
一、参数区间估计原理 .....	46
二、正态总体平均数的区间估计 .....	47
三、两个总体平均数差数的区间估计 .....	48
四、总体频率与两个总体频率差数区间估计 .....	49
第六节 可疑值的取舍 .....	50
一、 $4\bar{d}$ 法 .....	50
二、 $t$ 检验法 .....	50
习题二 .....	51
<b>第三章 方差分析</b> .....	<b>53</b>
第一节 单向分组资料的方差分析 .....	53
一、试验设计与数据整理 .....	53
二、方差分析步骤 .....	54
三、方差分析应用实例 .....	55
第二节 多重比较 .....	58
一、最小显著差数检验 .....	59
二、最小显著极差检验 .....	63
三、多重比较方法选择 .....	66
第三节 方差分析的基本假定及数据转换 .....	67
一、单向分组资料的线性统计模型 .....	67
二、方差分析应满足的 3 个条件 .....	68
三、数据转换 .....	69
第四节 数据资料的正态性检验 .....	71
一、偏-峰态检验法 .....	71
二、资料分布正态性检验的 DPS 实现 .....	73
第五节 方差同质性检验 .....	73

一、单个样本及两个样本方差的同质性检验 .....	74
二、多个样本方差同质性检验 .....	74
三、多个样本方差同质性检验的 DPS 实现 .....	76
四、单向分组资料统计分析的 DPS 实现 .....	77
习题三 .....	79
<b>第四章 非参数检验 .....</b>	<b>81</b>
<b>第一节 符号检验 .....</b>	<b>81</b>
一、单个样本的符号检验 .....	81
二、配对资料的符号检验 .....	82
<b>第二节 秩和检验 .....</b>	<b>83</b>
一、成组比较的秩和检验 .....	83
二、配对比较的符号秩和检验 .....	86
三、秩和检验的 DPS 实现 .....	88
<b>第三节 计数资料的 <math>\chi^2</math> 检验 .....</b>	<b>89</b>
一、 $\chi^2$ 检验的基本思想 .....	89
二、适合性检验 .....	90
三、独立性检验 .....	91
四、 $\chi^2$ 检验的统计软件实现 .....	92
习题四 .....	94
<b>第五章 试验设计 .....</b>	<b>95</b>
<b>第一节 试验设计概述 .....</b>	<b>95</b>
一、试验设计的意义 .....	95
二、试验研究的基本要求 .....	95
三、与试验有关的术语 .....	96
<b>第二节 试验方案设计 .....</b>	<b>98</b>
一、试验方案设计的基本原则 .....	98
二、效应比较性试验方案设计 .....	99
<b>第三节 试验方法设计 .....</b>	<b>101</b>
一、试验方法设计的目的 .....	101
二、生物试验的误差来源 .....	101
三、试验方法设计的原则 .....	102
四、试验方法设计的类型 .....	102
五、试验方法设计的 DPS 实现 .....	104
<b>第四节 正交设计 .....</b>	<b>107</b>
一、正交设计的基本思想 .....	107
二、正交表 .....	110
三、效应混杂的概念 .....	113



四、正交设计步骤 .....	113
五、正交表的 DPS 获取方法 .....	117
习题五 .....	117
<b>第六章 田间试验与抽样技术 .....</b>	<b>119</b>
<b>第一节 田间试验的特点与误差来源 .....</b>	<b>119</b>
一、田间试验的特点 .....	119
二、田间试验的误差来源 .....	119
<b>第二节 田间试验误差的控制途径 .....</b>	<b>120</b>
一、选择具有代表性的试验地 .....	120
二、试验地的准备 .....	120
三、田间试验小区技术 .....	121
<b>第三节 田间试验实施技术 .....</b>	<b>122</b>
一、编制作业计划书 .....	122
二、田间划区和筑小田埂 .....	123
三、施肥、播种移栽和定植 .....	123
四、田间管理 .....	123
五、观察记载 .....	123
六、收获测产 .....	124
<b>第四节 抽样技术 .....</b>	<b>125</b>
一、抽样研究的基本概念与抽样研究的特点 .....	125
二、抽样调查的步骤与抽样误差的估计 .....	126
三、抽样的基本方法 .....	127
四、抽样方案的制定 .....	130
五、简单随机抽样实例及 Excel “抽样” 工具应用 .....	131
习题六 .....	132
<b>第七章 常用效应比较性试验结果的统计分析 .....</b>	<b>133</b>
<b>第一节 单因素试验结果的统计分析 .....</b>	<b>133</b>
一、随机区组设计的单因素试验结果统计分析 .....	133
二、拉丁方设计单因素试验结果统计分析 .....	137
三、完全随机化设计单因素系统分组试验结果的统计分析 .....	140
<b>第二节 两因素全面实施试验结果的统计分析 .....</b>	<b>142</b>
一、完全随机化设计的两因素试验结果统计分析 .....	143
二、随机区组设计的两因素试验结果统计分析 .....	149
三、裂区设计的两因素试验结果统计分析 .....	156
四、随机区组设计单因素多点试验结果的统计分析 .....	161
<b>第三节 多因素全面实施试验结果统计分析 .....</b>	<b>164</b>
一、完全随机化设计的三因素试验结果统计分析 .....	164

二、随机区组设计的三因素试验结果统计分析 .....	170
三、再裂区设计的三因素试验结果统计分析 .....	173
<b>第四节 正交试验结果的统计分析</b> .....	183
一、正交试验结果的极差分析 .....	184
二、正交试验结果的方差分析 .....	185
三、正交试验结果统计分析需要注意的问题 .....	195
习题七 .....	195
<b>第八章 回归与相关分析</b> .....	199
<b>第一节 回归与相关的基本概念</b> .....	199
一、变量之间相互关系的类型 .....	199
二、回归分析与相关分析的基本概念 .....	200
<b>第二节 一元线性回归分析</b> .....	200
一、散点图 .....	200
二、一元正态线性回归模型 .....	201
三、回归参数 $\beta_0$ 和 $\beta_1$ 的最小二乘法估计 .....	201
四、回归方程的显著性检验 .....	204
五、回归系数的显著性检验和区间估计 .....	206
六、一元线性回归分析的 Excel “回归” 分析工具实现 .....	208
七、依变量重复取值的一元线性回归分析 .....	209
八、回归方程的应用 .....	212
九、线性回归分析注意事项 .....	215
<b>第三节 直线相关分析</b> .....	215
一、相关关系的度量指标——相关系数 .....	216
二、相关系数的显著性检验 .....	217
三、相关分析的 Excel “相关系数” 分析工具实现 .....	218
<b>第四节 多元线性回归与复相关分析</b> .....	219
一、多元线性回归系数的最小二乘估计 .....	220
二、多元线性回归关系的显著性检验 .....	221
三、复相关系数与偏相关系数 .....	222
四、多元线性回归分析的 Excel “回归” 分析工具实现 .....	223
五、多元线性回归方程的应用 .....	226
<b>第五节 一元非线性回归分析</b> .....	226
一、非线性回归分析的内容及方法 .....	226
二、可线性化的一元非线性方程配置方法 .....	228
三、一元非线性模型选优 .....	232
<b>第六节 多项式回归</b> .....	234
一、一元 $k$ 次多项式回归分析 .....	234
二、多元二次多项式模型的建立 .....	238

习题八 .....	244
<b>第九章 协方差分析</b> .....	246
第一节 协方差分析的意义与功用 .....	246
一、协方差分析的意义 .....	246
二、协方差分析的功用 .....	247
第二节 单向分组资料协方差分析 .....	247
第三节 两向分组资料协方差分析 .....	254
一、两向分组资料协方差分析的方法原理 .....	254
二、两向分组资料协方差分析的 DPS 实现 .....	258
习题九 .....	259
<b>第十章 多元统计分析简介及相关软件应用</b> .....	261
第一节 逐步回归分析与通径分析 .....	261
一、逐步回归分析简介 .....	261
二、通径分析简介 .....	263
三、逐步回归分析和通径分析的 Excel 与 DPS 实现 .....	266
第二节 聚类分析 .....	269
一、系统聚类分析简介 .....	269
二、系统聚类分析的 DPS 实现 .....	271
第三节 判别分析 .....	273
一、两类判别 .....	273
二、多类逐步判别 .....	276
第四节 主成分分析 .....	281
一、主成分分析的方法原理 .....	281
二、主成分分析的 DPS 实现 .....	285
第五节 典型相关分析 .....	288
一、典型变量和典型相关系数的概念 .....	288
二、典型相关分析的 DPS 实现 .....	289
习题十 .....	292
<b>附表</b> .....	295
附表 1 Duncan 新复极差检验 $\alpha$ 为 0.05 与 0.01 时的 SSR 值表 .....	295
附表 2 多重比较中的 $q$ 值表 (两尾) .....	296
附表 3 符号检验表 $S_n$ .....	297
附表 4 两样本秩和检验的 $W$ 临界值表 .....	298
附表 5 配对比较符号秩和检验的 $W'$ 临界值表 .....	299
附表 6 $r$ 与 $R$ 值的 0.05 与 0.01 显著性临界值表 .....	299
<b>主要参考文献</b> .....	301

# 第一章 统计学基础知识

试验设计与统计分析运用数理统计原理和误差理论,分析和解释自然与社会科学中的数量关系,帮助研究者正确设计试验和科学分析试验结果,从而揭示自然与社会问题真相,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

#### (一) 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、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3个特征。

**1. 同质性** 同一总体内的各个体,必须在某一方面具有相同的性质,才能把它们集合起来,构成某种属性相同的总体。所谓同质,不是绝对的而是相对的,是随研究目的而变化的。例如土壤碱解氮含量是土壤肥力的重要指标,在研究某地水田土壤供氮能力时,该地区每一块水田的土壤碱解氮含量就是该研究总体的个体。在研究不同施氮量对某地水稻分蘖盛期土壤碱解氮含量影响时,则在某一施氮水平下该地某田块水稻分蘖盛期的土壤碱解氮含量就是该施氮水平下总体的个体。

**2. 变异性** 构成总体的单元在具有同质性前提下,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,若没有变异性则无须统计。对总体的统计研究,实际上就是研究总体内个体间的变异情况。例如要研究某地区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,就是统计分析的基本任务之一。

**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。当总体容量无穷大时,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。

对总体的研究可以做全面普查,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逐一考察,多数情况下只能采用抽样调查的方法来推断总体的特征。

## (二) 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,所以抽样必须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 $n$ 。 $n \geq 30$ 的样本称为大样本; $n < 30$ 的样本称为小样本。有时大样本的统计特征数与小样本的统计特征数具有不同的统计分布形式。

## 二、变数与数据资料

### (一) 变数

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它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、土壤汞含量等都属于随机变量。

### (二) 数据资料

某随机变量的一组具体观测值称为数据资料。例如测定10个土壤样品的全氮含量,将获得10个土样的全氮含量数据资料。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数量属性和质量属性两大类。

**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、地下水中硝酸盐浓度、水体中细菌的数量等都具有数量属性。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资料和间断性变异资料两类。

(1) 连续性变异资料 连续性变异资料是指个体数据间存在很小差别,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。

(2) 间断性变异资料 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** 在环境与资源以及其他科学研究中,有些调查观察对象的一些属性只能观察不能度量,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。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问题。

### 三、误差与错误

#### (一) 误差

在一定条件下某客体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,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可以校正或消除的。例如在环境与资源及其他科学研究的测量中,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称为方法误差。如质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、比色分析中干扰离子的影响等,都会系统地影响测定结果使其偏高或偏低。

(2) 仪器误差 因所用仪器不够精确而引起的误差称为仪器误差,例如滴定管、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称为试剂误差。不纯的试剂或蒸馏水常会引入被测物或干扰物质,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称为操作误差。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都会造成操作误差。另一类是因操作人员的生理条件限制或习惯上的差异所引起的误差,如对滴定的终点颜色变化敏感程度不一,有人偏深有人偏浅。

(5) 地力因素造成的误差 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能通过对照试验、空白试验、仪器校准等办法校正。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** 随机误差又称为偶然误差,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。例如称量坩埚时,即使严格控制冷却时间,但因空气湿度随时都在变化,各次称量也可能有微小差异。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 $\pm 0.1$  mL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,例如供试材料的不均一性,如种子质量、秧苗素质不可能完全一致;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;农事操作的不一致性以及其他不可预测的自然因素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,可能对这份的影响大些,对另一份的影响可能小些。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,所以随机误差也称为不可测误差。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。

## (二) 错误

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，如违反操作规程而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。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六节介绍。

## 四、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。精确性是指对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，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，并减小试验误差提高试验精度就是试验方法设计所要解决的核心问题。

**【例 1.1-1】** 某物体质量为 0.110 8 g，甲、乙、丙 3 人分别对该物体的质量进行 3 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 3 人对某物体质量测定结果的准确性与精确性比较

观测者	观测值			平均数	平均数与真值 之差	最大值与最小值 之差
	1	2	3			
甲	0.110 2	0.110 7	0.110 9	0.110 6	-0.000 2	0.000 7
乙	0.110 1	0.112 7	0.111 5	0.111 4	0.000 6	0.002 6
丙	0.112 5	0.112 6	0.112 4	0.112 5	0.001 7	0.000 2

由 3 人测定结果的平均数与被测物体实际质量偏差可见，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 3 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

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，对其特征难

以一目了然。所以在获取数据资料后，常常需要对其整理和分组。整理时，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。当观测值个数较大时，需按观测值的大小分成若干组，编成次数分布表或制作次数分布图，以直观明了地反映数据资料的分布特征。

## 一、次数分布

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（也称为频数），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。

为更形象直观地表示数据资料的次数分布特征，可根据次数分布表做成次数分布图。常用的分布图包括直方图（也称为方柱形图）和折线图（也称为多边形图）。直方图是以组限为横坐标，以次数或累积频率（%）为纵坐标，由许多直方柱构成的分布图；而多边形图是由折线连接而成的分布图。

下面结合具体例子说明次数分布表及次数分布图的制作过程。

**【例 1.2-1】** 水稻氮素营养丰缺状况与叶绿素含量密切相关，可以采用叶绿素计测定叶片的 SPAD 值来表征叶片的叶绿素含量。用 SPAD-502 型叶绿素计测得某水稻品种分蘖盛期的 200 株水稻主茎顶端第一完全展开叶 SPAD 值如表 1.2-1 所示，试制作次数分布表和分布图。

表 1.2-1 200 株水稻主茎顶端全展叶 SPAD 值

行号 (i)	列号 (j)									
	1	2	3	4	5	6	7	8	9	10
1	<b>33.5</b>	41.4	39.7	37.5	38.4	36.4	38.9	37.9	38.1	37.0
2	38.2	37.8	38.9	36.0	37.6	40.5	37.6	37.5	36.7	36.4
3	36.7	36.2	40.6	38.7	38.9	36.4	37.1	39.5	39.5	33.6
4	39.3	34.1	40.2	35.6	38.4	40.0	38.3	37.3	38.5	35.7
5	38.3	34.4	38.1	34.4	38.3	38.8	38.7	37.8	38.6	39.3
6	37.9	37.5	36.4	40.1	37.5	37.2	39.0	37.1	36.7	39.6
7	37.4	38.1	37.9	37.1	36.5	37.7	38.9	41.2	39.5	36.8
8	39.9	38.3	36.9	37.1	36.9	35.8	37.9	37.2	36.3	36.2
9	39.4	38.4	39.6	38.3	37.7	40.7	36.3	39.0	36.2	38.2
10	39.0	38.4	36.8	38.7	39.3	38.7	37.5	38.3	39.0	34.7
11	36.6	36.5	39.2	39.5	38.1	36.5	39.5	36.7	37.3	36.2
12	39.6	38.6	39.5	38.9	41.5	39.7	35.3	39.1	40.3	37.8
13	38.8	37.8	42.1	37.9	39.0	39.4	38.1	39.5	37.2	39.0
14	38.1	37.4	<b>43.1</b>	36.8	37.4	38.3	38.6	37.6	37.4	38.7
15	37.2	35.6	38.4	39.0	36.6	36.7	37.7	38.0	36.4	38.2
16	34.7	38.4	37.6	37.9	38.6	36.6	36.1	39.8	38.9	35.8
17	36.0	39.2	37.1	37.8	38.0	39.4	40.5	35.1	39.5	37.8
18	37.5	39.3	37.2	40.5	35.7	37.9	40.1	41.8	38.8	36.7
19	36.4	38.1	39.5	39.0	38.8	36.5	38.2	38.3	38.4	38.8
20	36.5	37.2	35.0	36.1	38.0	38.0	37.4	35.5	39.0	37.2



**解：**次数分布表及分布图的制作步骤如下。

1. 求变幅 ( $R$ ) 变幅也称为极差, 记作  $R$ , 是指一群数据资料的最大值 ( $y_{\max}$ ) 与最小值 ( $y_{\min}$ ) 之差, 即  $R=y_{\max}-y_{\min}$ 。表 1.2-1 数据若以代码  $y_{i,j}$  表示, 则最大值  $y_{\max}=y_{14,3}=43.1$ , 最小值  $y_{\min}=y_{1,1}=33.5$ , 所以表中 200 个数据的变幅为:  $R=43.1-33.5=9.6$ 。

2. 预估分组数 ( $k$ ) 次数分布表的分组数 ( $k$ ) 决定于样本量 ( $n$ ) 的大小, 一般可采用 Sturge 公式来粗略估计:  $k=1+3.3\lg(n)$ 。

由于  $n=200$ , 所以,  $k=1+3.3\lg 200=8.6 \approx 9$ , 因此确定本例的预估分组数为 9。

3. 确定组距 ( $C$ ) 组距即同一组的上限与下限之差, 可由变幅  $R$  除以分组数  $k$  来估计, 即  $C=R/k$ 。本例  $R=9.6$ ,  $k=9$ , 则  $C=9.6/9=1.07$ , 为方便起见, 取组距 1.0。

4. 选定各组的上限 ( $L_i$ ) 及总分组数 为避免归组时数据被遗漏, 最小组的上限 ( $L_1$ ) 应大于最小观察值 ( $y_{\min}$ ); 最大组上限 ( $L_k$ ) 大于最大观察值 ( $y_{\max}$ )。本例  $y_{\min}=33.5$ ,  $y_{\max}=43.1$ ,  $C=1$ 。为分组方便可取  $L_1=34$ , 则  $L_2=L_1+C=34+1=35, \dots, L_{11}=L_{10}+1=44$ , 共得到 11 个组的上限, 见表 1.2-2。

5. 观测值归组记数 采用唱票法, 按大于前组上限并小于等于本组上限的数据归于本组的原则, 将各观测值归入相应的组, 并记数每组的观测值个数即次数, 列于表 1.2-2 的第 3 列。即得到次数分布表。

## 二、利用 Excel “直方图” 工具制作次数分布表与分布图

各种版本的 Microsoft Excel 软件, 只要完全安装都带有“数据分析”工具。

1. Microsoft Excel 2003 开启 Excel 电子表格, 依次点击主菜单中的“工具”→下拉菜单中“加载宏”; 勾选弹出窗口中的“分析工具库”, 点击“确定”; 再点击主菜单“工具”, 在下拉菜单中便看到“数据分析”。

2. Microsoft Excel 2007 开启 Excel 电子表格, 依次点击, 左上角“Office 按钮”→选中弹出窗口右下角的“Excel 选项”→在下一弹出窗口点击“加载项”, 选中“分析工具库”, 点击“确定”。再点击主菜单中的“数据”便出现“数据分析”。



点击“数据分析”弹出窗口中有“描述统计”、“直方图”等 19 种分析工具可供选用。利用其中的“直方图”功能可以对一组给定数据制作次数分布表与分布图。

**【例 1.2-2】** 试采用 Excel “直方图” 工具, 对表 1.2-1 数据制作次数分布表及分布图。

**解：**Excel “直方图” 分析工具操作步骤如下。

1. 输入数据 将表 1.2-1 中的 200 个 SPAD 值数据和表 1.2-2 中的组上限数据分别粘贴到 Excel 的空白电子表格。

2. 选择分析工具 点击主菜单栏的“工具”, 在下拉菜单中点击“数据分析”; 从弹出窗口中选择“直方图”, 按“确定”; 弹出“输入及输出选项”对话框。