全国高等农业院校教材

全国高等农业院校教材指导委员会审定

肥料试验与统计分析

陶勤南 主编

土壤与农业化学专业用

中国农业出版社

全国高等农业院校教材

肥料试验与统计分析

陶勤南　主编

土壤与农业化学专业用

中国农业出版社

全国高等农业院校教材
肥料试验与统计分析
陶勤南　主编

责任编辑　黄慧民
出版　中国农业出版社
（北京市朝阳区农展馆北路 2 号）
发行　新华书店北京发行所
印刷　通县曙光印刷厂

*　　*　　*

开本　787mm×1092mm16 开本
印张　21.75　　字数　500 千字
版、印次　1997 年 5 月第 1 版
1997 年 5 月北京第 1 次印刷
印数　1—2,000 册　　定价 20.60 元

书号　ISBN 7-109-04393-2/S・2719

前　言

西北农业大学主编的《农业化学研究法》教材的内容包括肥料试验与同位素示踪技术两部分，分上下二册出版。由于《核技术在农业上的应用》已定为选修课，原教材的名称已不适合，从本版起改为《肥料试验与统计分析》。

本教材是在以下原则下编写的：

第一，高等数学、概率论、线性代数与微机应用基础均已列为必修课，并在本课程之前已经修毕，因此统计学原理部分是紧密结合试验分析编写的；对多元回归中求解逆矩阵的方法不作详细叙述。为了便于应用，在有关内容之后附以微机计算程序。受课程性质与篇幅限制，本书只列出多项式回归及二元二次多项式回归的计算程序。已学过微机应用基础的同学不难略加修改即可适应各种回归分析应用。

第二，自1977年我国在肥料试验中首次应用回归设计以来，各种类型的回归设计已广泛应用。同时针对土化专业的特点增加了混杂设计。为了便于在生产中快速提出成套技术，对正交设计的内容作了较详细的介绍。对传统的随机区组、拉丁方及裂区试验，也针对土化专业的需要增添了含有假伪处理的裂区设计等内容。为了有利于对学习有兴趣的同学的进一步需要，最后一章介绍了肥料试验技术的进展。

全书分12章。由浙江农业大学陶勤南主编。由以下同志分别担任编写任务：第1—3章方萍执笔；第4—6章吴良欢执笔；第7—10章陶勤南执笔；第11—12章王兴仁执笔。

本教材是在农业部本科教材土化、植保学科组领导下编写的，编写过程中得到浙江农业大学党政领导及农业部教材学科组毛达如教授的支持；本书承杨义群教授、吴平教授审阅并提出许多宝贵意见。

由于编写人员知识有限，内容难免有误，希望提出宝贵意见，以便今后改正。

编　者

1994年6月于杭州

目　录

前言
第一章　生物统计的基础知识 …… 1
第一节　几个常用统计术语 …… 1
第二节　次数分布 …… 2
第三节　统计特征数 …… 5
第四节　概率与概率分布 …… 10
第五节　抽样分布 …… 15
第二章　肥料试验结果的统计假设检验 …… 20
第一节　肥料试验结果的直观分析及存在问题 …… 20
第二节　统计假设检验的原理和步骤 …… 21
第三节　平均数的统计假设检验 …… 24
第四节　计数资料的统计假设检验 …… 29
第五节　方差齐性检验 …… 32
第三章　方差分析 …… 35
第一节　方差分析的基本原理 …… 35
第二节　单向分组资料的方差分析 …… 35
第三节　多重比较 …… 39
第四节　方差分析的数学模型 …… 44
第五节　两向分组资料的方差分析 …… 45
第六节　方差分析的基本假定和数据转换 …… 55
第四章　肥料试验设计的基本原理 …… 57
第一节　肥料试验的种类和要求 …… 57
第二节　肥料试验处理设置的基本原理 …… 58
第三节　肥料试验方法设计的基本原理 …… 61
第五章　田间试验 …… 65
第一节　田间试验的意义与种类 …… 65
第二节　田间试验的误差及其控制途径 …… 66
第三节　田间试验的实施 …… 70
第六章　培养试验 …… 74
第一节　培养试验概况 …… 74
第二节　土培试验 …… 76
第三节　水培试验 …… 77
第四节　砂培试验 …… 81
第五节　特殊培养试验 …… 84

第六节　农化培养室 …… 88
第七章　随机区组、拉丁方与裂区设计 …… 90
第一节　单因素随机区组试验 …… 90
第二节　多因素随机区组试验 …… 92
第三节　多点随机区组试验 …… 95
第四节　拉丁方设计 …… 97
第五节　包含多个拉丁方的试验 …… 99
第六节　裂区试验 …… 101
第七节　再裂区试验 …… 106
第八节　包含假伪处理的裂区试验 …… 109
第九节　分条随机区组试验 …… 112
第十节　裂区拉丁方与分条拉丁方 …… 117
第八章　混杂设计与正交设计 …… 118
第一节　混杂及其在试验设计中的应用 …… 118
第二节　二水平多因素混杂设计 …… 119
第三节　正交表 …… 124
第四节　不设重复的三因素三水平混杂设计 …… 129
第五节　不设重复的正交试验 …… 133
第六节　规模较大的二水平正交表的应用 …… 135
第七节　$L_{27}(3^{13})$ 正交表的应用 …… 140
第八节　并列法 …… 144
第九节　拟因子法 …… 146
第十节　裂区法 …… 152
第九章　线性回归与相关 …… 157
第一节　一元线性回归 …… 157
第二节　有重复的一元线性回归 …… 166
第三节　两个回归方程的比较 …… 169
第四节　相关系数 …… 170
第五节　多元线性回归 …… 175
第六节　部分处理有重复的回归方程 …… 185
第十章　非线性回归分析 …… 191
第一节　可化为线性回归的曲线回归 …… 191
第二节　S形曲线 …… 203
第三节　多项式回归 …… 205
第四节　多元多项式回归 …… 216
第五节　可化为多元线性回归的非线性回归 …… 222
第六节　正交多项式回归 …… 224
第七节　多元正交多项式回归 …… 231
第十一章　回归设计 …… 234
第一节　回归设计的思路 …… 234
第二节　回归设计方案的建立 …… 242

第三节　回归设计的统计分析 …… 251
第四节　回归设计在推荐施肥中的应用 …… 262
第十二章　肥料试验技术的应用与发展 …… 272
第一节　肥料试验技术和应用的进展 …… 272
第二节　肥料试验方法和技术的应用途径 …… 274
第三节　生物试验报告的图示技术 …… 281
附表 …… 288
附表 1　10000 个随机数字 …… 288
附表 2　累积正态分布 Φ（Z）值 …… 292
附表 3　t 分布表 …… 293
附表 4　χ^2 分布表 …… 294
附表 5　F 分布表 …… 295
附表 6　Duncan's 新复极差检验 α 为 0.05 及 0.01 时的 SSR 值表 …… 298
附表 7　多重比较中的 Q 表 …… 300
附表 8　LSD 转化为 Duncan's 新复极差 SSR 值的乘数 R 值表 …… 302
附表 9　检验相关系数 $\rho=0$ 的临界值（r_α）表 …… 303
附表 10　r 与 Z 的换算表 …… 304
附表 11　正态累积概率和概率单位（P）转换表 …… 305
附表 12　正交多项式表 …… 309
附表 13　Yates 氏 N×N 拉丁方变换组 …… 310
附表 14　正交表 …… 314
附表 15　均匀设计表 …… 326
附表 16　较优的回归设计方案 …… 332
参考文献 …… 337

第一章　生物统计的基础知识

生物统计学是数理统计学与生物科学有机结合的一门学科。由于生物体本身具有复杂的生命活动，并在生长发育过程中受到时常变动的各种外界环境条件的影响，因此，生物试验的数据资料普遍表现出变异性，肥料试验数据也不例外。如何从这些变异资料中找出客观规律，正是生物统计学所要解决的问题。生物统计学的功能就在于用适当的方法来确定这种变异中哪一部分是由环境条件引起的，哪一部分是由偶然因素造成的，哪一部分才是生物体本质上具有的数量表现，从而帮助研究者从偶然性所掩盖的试验结果中揭示其在内的必然规律。

第一节　几个常用统计术语

一、总体和样本

研究对象的全体称为总体，组成总体的基本单元称为个体。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。统计学上所说的总体事实上是指所有个体的某种性状的全体观测值。由于受许多偶然因素的影响，同一总体内个体之间往往在一定范围内变化的，但它们仍是同质的。总体容量即总体所包含的个体数目（记作N）若是有限的则称之为有限总体，反之当总体容量为无限时则称之为无限总体。例如，某年某地块上种植西瓜收获时测定所结西瓜的单瓜重，则研究对象是一个有限总体，因为当年该地块上所结的西瓜数是有限的。如果要研究西瓜品种“浙密一号”中心糖度，由于未指明何时何地所产的，因此过去现在将来各地所产的“浙密一号”西瓜均属于该总体，显然这是一个无限总体。

试验研究的目的在于了解总体的特征，但实际工作中由于总体容量往往太大甚至无限，而且许多测定方法具有破坏性，因此只能从总体中抽取若干个体组成样本，并通过对样本的观察研究来推断总体的特征。为了使样本能反映总体的客观规律，抽样不能凭主观意愿，而必须做到随机，也就是说抽样时总体内每个个体有同等的机会被抽取。拈阄就一种最简单的随机抽样方式。样本所含的个体数目称为样本容量，常记作n，肥料试验中n小于30的样本称为小样本，n大于30的样本称为大样本。对于大样本和小样本的统计分析方法是有所区别的。

二、变数与数据

由于受偶然因素的影响，总体内个体间普遍存在着变异性，因而观测值间也表现出波动性。例如，将一块田划分为形状面积相等的小区，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，结果不同小区的产量会因土壤肥力的不均一性及其它偶然因

素的影响而不可能完全相同。又如测定某一土壤样品的全氮含量，由于测定仪器、测定条件及测定者的操作技术等一系列因素的影响，重复测10次可能得到10个不同的观测值。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，常用大写字母X、Y、Z等表示。组成样本或总体的观测值称数据。肥料试验中变数可分为如下两类：

1. 连续性变数　是指用称量、度量、测量或分析化验等量测方法所获得的变数。如作物产量、植株高度、叶面积、酶活性、净光合速率等都是连续性变数，其特点是观测值的取值不限于整数，在任何两个相异的观测值之间可以有更小差异的其它观测值存在，而可观测出的差异大小取决于量测仪器的精度。

2. 不连续性变数　是指通过统计计数方法获得的变数。如水稻的每穗粒数，一批种子的发芽数，单位面积上的害虫数等都是不连续性变数，其特点是观测值的取值只限于非负整数，而没有带任何小数的观测值存在。

三、参数与统计数

用于描述总体特征的一些数值称总体参数如总体平均数、总体标准差，根据样本观测值运算得到的一些数值称为统计数，如样本平均数，样本标准差。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、机误与错误

在一定条件下某一事物所具有的真实数值即为真值。对该事物进行量测所得到的观测值由于受测定过程中的许多偶然因素的影响会与真值产生一定的偏差，这种偏差称为机误或试验误差。由于机误是有偶然因素或未知因素造成的，因而是无法避免的，但是可以通过适当的试验设计使其减小，并对机误的大小作出估计。错误是由于工作疏忽所造成的，如抄错数据、看错仪表读数、计算错误等。只要在工作中认真细心错误是可以避免的。

五、准确性与精确性

准确性是指测定值即观测值与真值的相符性；精确性是指多次重复观测值之间的相符性。所以精确性不等于准确性，但精确性是准确性的前提，如果没有高度的精确性便无良好的准确性可言，但是，若有系统误差存在则即使精确性很高也达不到良好的准确性。

第二节　次数分布

由试验观察或调查所得的数据往往是杂乱无章的，表面上难以看出什么规律性，若通过分类归组，编制成次数分布表或绘成次数分布图就能使其一目了然，并显示出内在规律性。

一、次数分布表

编制次数分布表的方法是将观测值的变异范围划分为等间距的若干个区间，并记下归

入每一区间亦即每一组的观测值次数。由各组的组限、组中值及相应的观测值次数所组成的表就叫做次数分布表。现将表 1-1 所列的 100 株甜菜块根的蔗糖浓度制成的次数分布表列于表 1-2，整理步骤如下：

表 1-1　100 个甜菜块根的蔗糖含量（%鲜重）

株号	蔗糖（%）	株号	蔗糖（%）	株号	蔗糖（%）	株号	蔗糖（%）
1	11.8	26	13.5	51	10.1	76	9.0
2	13.1	27	11.9	52	12.4	77	14.0
3	9.2	28	16.7	53	10.8	78	13.2
4	8.7	29	9.6	54	11.3	79	15.0
5	12.9	30	15.1	55	6.3	80	13.8
6	13.7	31	14.6	56	15.7	81	15.1
7	9.6	32	10.4	57	14.3	82	14.9
8	13.7	33	13.4	58	15.0	83	12.6
9	8.5	34	14.6	59	12.5	84	14.1
10	15.7	35	10.5	60	11.8	85	11.4
11	14.1	36	8.6	61	11.6	86	9.4
12	11.9	37	15.2	62	12.2	87	12.4
13	16.7	38	11.1	63	7.5	88	15.0
14	7.4	39	14.5	64	13.4	89	9.4
15	10.0	40	12.1	65	14.7	90	12.9
16	4.4	41	14.9	66	14.2	91	13.4
17	13.2	42	15.0	67	14.0	92	10.6
18	13.8	43	12.1	68	15.1	93	6.5
19	9.1	44	12.6	69	6.5	94	11.0
20	11.9	45	13.0	70	8.7	95	11.9
21	12.8	46	14.1	71	11.0	96	11.8
22	15.3	47	14.4	72	13.0	97	12.6
23	12.6	48	13.1	73	9.2	98	9.5
24	16.1	49	13.3	74	7.0	99	12.2
25	17.2	50	15.0	75	13.2	100	8.2

1. 计算变幅 R　变幅就是最大观测值 y_{max} 与最小观测值 y_{min} 之差即 $R=y_{max}-y_{min}$。如表 1-1 中 $y_{max}=17.2\%$，$y_{min}=4.4\%$，所以 $R=17.2-4.4=12.8$（%）。

2. 选择组数 K　组数不宜过多或过少，因为组数太少则组距增大，所得次数分布表难以正确反映事物的真实情况，而组数过多使组距太小，不仅计算麻烦，还会使有的组的次数为零或很小，使资料过于分散而达不到分组的目的，同样显示不出资料固有的内在规律。对于大样本，一般可分为 8 至 20 组，也可采用 Sturge 公式：$K=1+3.3\log N$ 来估计并作适当调整，这里的 N 为样本或总体容量。对于表 1-1 来说，$N=100$，$K=1+3.3\log 100=7.6$，因此可分为 8 组。在本例中选用了 9 组。

3. 确定组距 C　组距即每一组所在区间的上、下限之差，可用 R/K 来估计，但组距的最后确定必须根据数据资料的性质及能否使次数分布表显得准确而清晰来决定。例如，表 1-1 的组距可根据 $C=12.8/9=1.4$ 来估计，为使分组方便我们确定组距 C 为 1.5（%）。

4. 决定组限和组中值　每一组所在区间的两个极端值称为组限，大的为上限，小的为下限，一个组的上、下限之和除以 2 得该组的组中值。组限的确定从最小一组的下限开始，以使该组的组中值接近资料的最小值为宜，可用 $y_{min}-\frac{1}{2}C$ 来估计，并为计算便利作适当调整，在每一组的下限基础上加上一个组距便得该的上限及后一组的下限，直到最大一组的上限确定为止。为避免观测值恰好落在组限上，组限的精度可在观测值的最末一位有效数字后增加$\frac{1}{2}$个单位。例如，表 1-1 资料最小值是 4.4（%），组距为 1.5（%）最小一组的下限可取 4.4－1/2×1.5＝3.65。本例中为计算便利取 4.05，它比观测值多一位小数，然而再加 1.5（%）便得到该组的上限和后一组的下限 5.55（%），直到最大一组的上限 17.55（%）确定为止。将各组的组限列于表 1-2 的第一列中，然后计算每一组的组中值列于表 1-2 的第二列中。

5. 统计各组的观测值次数　用唱票的方式将每一观测值归入相应的组，然后统计每组所包含的观测值数目即该组的次数，编制出次数分布表，如表 1-2 所示。

显然次数分布表（表 1-2）比原始数据表（表 1-1）能更清晰地反映甜菜块根蔗糖含量的特征，比如，蔗糖含量在 11.55%至 13.05%这一组次数最大为 24；有 79%的甜菜块根蔗糖含量小于 14.55%；而含量小于 8.55%或大于 16.55%的甜菜很少，分别只占 9%和 4%，而大多数甜菜块根的蔗糖含量在 8.55%至 16.55%之间，占 87%。

表 1-2　由表 1-1 数据编成的次数分布表

组号	组　限	组中值	归组计数	次数（f）	累积次数
1	4.05—5.55	4.8	丨	1	1
2	5.55—7.05	6.3	丨丨丨丨	4	5
3	7.05—8.55	7.8	丨丨丨丨	4	9
4	8.55—10.05	9.3	𝍸 𝍸 丨丨丨	13	22
5	10.05—11.55	10.8	𝍸 𝍸	10	32
6	11.55—13.05	12.3	𝍸 𝍸 𝍸 𝍸 丨丨丨	24	56
7	13.05—14.55	13.8	𝍸 𝍸 𝍸 𝍸 丨丨丨	23	79
8	14.55—16.05	15.3	𝍸 𝍸 𝍸 丨丨	17	96
9	16.05—17.55	16.8	丨丨丨丨	4	100

二、次数分布图

用图示法表示次数分布比次数分布表更加直观，方柱形图和多边形图是常用的次数分布图形式。方柱形图由许多相邻的长方形组成，每个长方形的水平基线以组中值为中点，以上、下限为端点，高度表示次数。多边形图是用折线连成的图形，组中值位于横轴上，次数位于纵轴上，作图时先标出对应于每一组中值和次数的各点，再用折线把相邻点连接起来，这相当于在方柱形图中把各长方形顶边的中点用折线连接起来。现将表 1-2 的次数分布绘成方柱形图和多边形图如图 1-1 所示。用累积次数除以观测值总数 N 就得到累积频率，并绘成图 1-2 的累积频率图。根据累积频率图可以估计观测值低于或等于某一给定值的百分率，或计算落在某一范围内的观测值的百分率。例如，由图 1-2 可见，有 22%的甜菜块根蔗糖含量低于或等于 10.05%，有 79%的甜菜块根蔗糖含量低于或等于 14.55%，有 57%

(79%－22%) 的甜菜块根蔗糖含量在 10.05%到 14.55%之间。

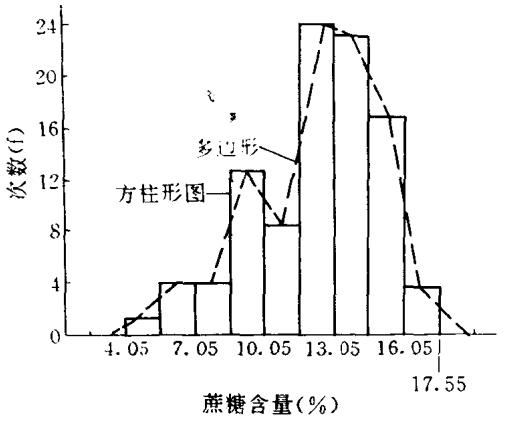

图 1-1 表 1-1 数据的次数分布图

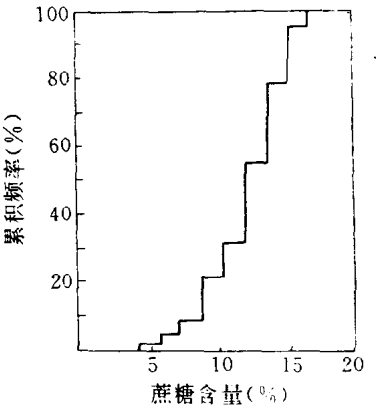

图 1-2 表 1-1 数据的累积频率图

第三节 统计特征数

上节讨论的次数分布反映了数据资料的概貌，可以从中粗略地估计其变异程度和集中趋势，但不能用简明的数字来描述这些特征，因此，需进一步求出能度量资料的集中趋势和分散程度的一些特征数字，称为统计特征数，如平均数、标准差等。

一、趋中性的度量

肥料试验中收集到的许多数据资料作成次数分布表，往往会显示出这样的特征，即次数由少而多，达到最高值后又逐渐减少，大多数观测值集中在中间水平，这种特征就是数据资料的“趋中性”。表示这种趋中特征的统计特征数称为平均数，它可以简明地表示资料的平均水平和中心位置，并可作为资料的代表与另一组同类资料相比较，借以表示二者的差异情况。

(一) 平均数的种类

1. 算术平均数 一组数据的总和除以该组数据的个数所得的商称为算术平均数，简称平均数。对样本而言，所有观测值之和除以样本容量 (n) 即为样本平均数。实际上，它是每个观测值各贡献 $\frac{1}{n}$ 份的总和，因此，算术平均数代表了该组数据的一般水平。

2. 中数 将一组数据由小到大排列，居于中间的那个数据称为中数。若数据个数为偶数，则居中的两个数据的平均值为中数。中数反映了一组数据的居中位置。

3. 众数 在一组数据中，出现次数最多的那个数值称为众数。

4. 几何平均数 如果数据资料是一种比率如增长率，所要求的平均数为平均比率，则应该用几何平均数 G 来表示，G 的计算公式如下：

$$G=\sqrt[n]{y_1 \cdot y_2 \cdots y_n} \qquad (1\text{-}1)$$

式中 G 为几何平均数，n 为样本容量，y_1，y_2，…，y_n 为观测值。

5. 调和平均数　如果数据资料是一种变化率如速率等，求其平均变化率须用调和平均数 H，其计算公式为

$$H=\frac{n}{\frac{1}{y_1}+\frac{1}{y_2}+\cdots+\frac{1}{y_n}} \tag{1-2}$$

例 1　水分在土壤毛管中的上升速度从土表下 30 厘米起，第一个 10 厘米为每分钟 8 厘米，第二个 10 厘米为每分钟 6 厘米，第三个 10 厘米为每分钟 4 厘米，试求平均上升速率。

解　已知 $n=3$，$y_1=8$ 厘米/分，$y_2=6$ 厘米/分，$y_3=4$ 分，根据（1-2）计算平均上升速度

$$H=\frac{n}{\frac{1}{y_1}+\frac{1}{y_2}+\frac{1}{y_3}}=\frac{3}{\frac{1}{8}+\frac{1}{6}+\frac{1}{4}}=5.5\text{（厘米/分）}$$

即土壤水分从土表下 30 厘米到土表的平均上升速率为每分钟 5.5 厘米。

（二）算术平均数的计算

1. 直接法　对于一个容量为 n 的样本，其观测值用 y_1，y_2，$\cdots y_n$ 表示，则平均数 $\bar{y}$（读作 y bar，或 y 杠）由下式求得

$$\bar{y}=\frac{y_1+y_2+\cdots+y_n}{n}=\frac{\sum_{i=1}^{n}y_i}{n} \tag{1-3}$$

表 1-1 的样本平均数

$$\bar{y}=\frac{(11.8+13.1+9.2+\cdots+8.2)}{100}=12.2\text{（%）}$$

对于容量为 N 的有限总体其平均数用 μ 表示，则

$$\mu=\frac{1}{N}\sum_{i=1}^{N}y_i \tag{1-4}$$

式中　y_1，y_2，…，y_N——表示总体的所有个体的取值；

Σ——为加和符号，读作 sigma。

2. 加权法　当一个样本内不同观测值的权重不同时，须用加权法计算平均数，其公式为

$$\bar{y}=\frac{f_1y_1+f_2y_2+\cdots+f_ny_n}{f_1+f_2+\cdots+f_n}$$

或

$$=\frac{\sum_{i=1}^{n}f_iy_i}{\sum_{i=1}^{n}f_i} \tag{1-5}$$

式中　f_1，f_2，…，f_n 分别为 y_1，y_2，…，y_n 的权数，用于权衡不同观测值作用的轻重。当各 f_i 相等时，式（1-3）与（1-5）完全相同。

例 2　为了解某农场小麦平均亩产，共调查了 5 块地，面积各为 10、20、40、15、15 亩，亩产分别为 300、250、250、150、300 公斤/亩，试求平均亩产。

解　$\sum_{i=1}^{5}f_iy_i=10\times300+20\times250+40\times200+15\times150+15\times300=22\,750$（公斤）

$$\sum_{i=1}^{5} f_i = 10+20+40+15+15 = 100 \text{（亩）}$$

$$\therefore \text{小麦平均亩产（}\bar{y}\text{）} = \frac{22\,750}{100} = 227.5 \text{（公斤/亩）}$$

对于次数分布表，也可用加权法计算平均数。用组中值代替 y_i，次数代表 f_i。

例 3　用加权法计算表 1-2 的算术平均数。

解　$\because$
$$\sum_{i=1}^{9} f_i y_i = 1\times 4.8+4\times 6.3+4\times 7.8+13\times 9.3+10\times 10.8$$
$$+24\times 12.3+23\times 13.8+17\times 15.3+4\times 16.8$$
$$=1230$$

$$\sum_{i=1}^{9} f_i = n = 100$$

$\therefore$
$$\bar{y} = \frac{1\,230}{100} = 12.3 \text{（\%）}$$

利用次数分布表计算的平均数限从原始数据计算的平均数稍有出入，这是由于前者以组中值代替每组的观测值来计算平均数，但观测值在组内并不完全是均匀分布的，因而组中值不能完全代表全组观测值，然而这种差异不会太大。

（三）算术平均数的性质

算术平均数是生物统计学中的一个重要统计特征数，它具有如下性质：

1. 离均差即样本中各观测值与样本平均数之差的总和为零，亦即

$$\sum_{i=1}^{n} (y_i-\bar{y}) = 0 \qquad (1\text{-}6)$$

证明如下

$\because$
$$\sum_{i=1}^{n} (y_i-\bar{y}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y}$$
$$= \sum_{i=1}^{n} y_i - n\bar{y}$$
$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i$$

$\therefore$
$$\sum_{i=1}^{n} (y_i-\bar{y}) = 0$$

2. 离均差的平方和最小。这是指样本观测值与平均数之差的平方和跟样本观测值与任一不等于样本平均数的实数之差的平方和相比为最小。

证明：设 $a \neq \bar{y}$，$a = \bar{y}+\Delta$　$\Delta \neq 0$

则
$$\sum_{i=1}^{n} (y_i-a)^2 = \sum_{i=1}^{n} (y_i-\bar{y}+\Delta)^2$$
$$= \sum_{i=1}^{n} [(y_i-\bar{y})+\Delta]^2$$
$$= \sum_{i=1}^{n} (y_i-\bar{y})^2 + 2\Delta\sum_{i=1}^{n} (y_i-\bar{y}) + \sum_{i=1}^{n} \Delta^2$$

$\because$　$\sum (y_i-\bar{y}) = 0$，　$2\Delta\sum_{i=1}^{n} (y_i-\bar{y}) = 0$

$\therefore$　上式 $= \sum_{i=1}^{n} (y_i-\bar{y}) + \Delta^2$

又$\because$　$\Delta \neq 0$　则 $\Delta^2 > 0$

$\therefore$
$$\sum_{i=1}^{n} (y_i-\bar{y})^2 < \sum_{i=1}^{n} (y_i-a)^2 \qquad (1\text{-}7)$$

二、变异程度的度量

总体内部个体间的变异程度是总体的另一个重要特征。如果两个总体具有相同的平均数，但其变异程度不一样，则这两个总体的性质是不一样的。例如，用两台仪器 A 与 B 分别重复测定同一个土样的全氮含量，得结果如下：

仪器 A：0.102%，0.101%，0.100%，$\bar{y}=0.101\%$

仪器 B：0.116%，0.097%，0.090%，$\bar{y}=0.101\%$

虽然两者平均数相等，但很明显仪器 B 所得结果较仪器 A 所得结果更为分散，所以仪器 A 较 B 稳定，结果更可靠。衡量这种分散程度的特征数有如下几种：

1. 极差　即一组数据中最大值与最小值之差。极差反映了一组数据的最大变异幅度，故亦称变幅或全距。由于极差只利用了资料中的两个极端值的信息，不能反映两极端值之间其它数据的变异状况，因而用极差来衡量整个数据资料的变异程度有其缺陷。

2. 方差　即离均差平方之平均值。总体方差用 σ^2 表示，样本方差亦称均方用 s^2 表示。它们的计算公式如下：

$$\sigma^2=\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{N}\text{或简写为}\frac{\sum(y-\mu)^2}{N} \tag{1-8}$$

$$s^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}\text{或简写为}\frac{\sum(y-\bar{y})^2}{n-1} \tag{1-9}$$

式（1-8）是对有限总体而言的，其中 y_i 为总体的任一个体观测值，μ 为总体平均数，N 为总体容量，式（1-9）中 y_i 为样本的任一观测值，$\bar{y}$ 为样本平均数，n 为样本容量。

这里为何要用离均差平方的平均值而不是直接用离均差的平均值来度量变异程度呢？这是因为：第一，虽然离均差反映了观测值偏离平均数的大小，但离均差的代数和为零，因此无论个别离均差有多大，其平均值总等于零，无法用于度量数据的变异程度；第二，离均差通过平方不仅消除了负号，而且加重了较大离均差的分量，可以增加度量变异程度的灵敏度。

既然方差是离均差平方的平均值，为何样本方差公式中分母是 n－1 而非 n 呢？这是由于总体方差往往是未知的，需用样本方差 s^2 来估计，而样本方差的分子部分即离均差的平方和 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 中样本平均数 $\bar{y}$ 代替了总体方差公式中相应的总体平均数 μ，而样本只是总体的一部分，$\bar{y}$ 与 μ 之间会有一定的偏差，即 $\bar{y}\neq\mu$，根据（1-7），$\sum_{i=1}^{n}(y_i-\bar{y})^2<\sum_{i=1}^{n}(y_i-\mu)^2$，如果 s^2 的分母用 n 会使 s^2 比 σ^2 偏小。为此，英国统计学家 W. C. Gosset 建议当 $n<30$ 时即小样本时，改用 n－1 作分母，这样可使小样本的方差成为总体方差的无偏估计。这里分母 n－1 称为自由度，表示独立项的数目，它等于样本容量减去约束条件数。在 s^2 中由于受 $\bar{y}$ 的制约，即受 $\sum(y_i-\bar{y})=0$ 的约束故自由度为 n－1。一般地在估计某个统计数时，如果受到 k 个常数的限制，则自由度（γ）为 n－k。对于 $n>30$ 的大样本，由于用 n 或 n－1 除结果相差不大，故可按 $\frac{\sum(y_i-\bar{y})^2}{n}$ 计算 s^2。

3. 标准差　即方差的平方根。方差是离均差平方的平均，其单位成为原观测值单位的

平方，有时会使其失去意义。标准差又恢复了观测值的度量单位。总体标准差记作 σ，样本标准差记作 s，计算公式分别为

$$\sigma=\sqrt{\frac{\sum (y-\mu)^2}{N}} \tag{1-10}$$

$$s=\sqrt{\frac{\sum (y-\bar{y})^2}{n-1}} \tag{1-11}$$

计算方差或标准差时，分子部分（即离均差的平方和简称平方和记作 SS）的计算较复杂，可以将平方和的公式作如下变换

$$SS=\sum (y-\bar{y})^2=\sum y^2-\frac{(\sum y)^2}{n} \tag{1-12}$$

证明∵ $\sum (y-\bar{y})^2=\sum (y^2-2y\bar{y}+\bar{y}^2)$

$$=\sum y^2-2\bar{y}\sum y+n\bar{y}$$

$$=\sum y^2-2\frac{\sum y}{n}\sum y+n\left(\frac{\sum y}{n}\right)^2$$

$$=\sum y^2-2\frac{(\sum y)^2}{n}+\frac{(\sum y)^2}{n}$$

$$=\sum y^2-\frac{(\sum y)^2}{n}$$

式中 $\frac{(\sum y)^2}{n}$ 称为矫正数，记作 C。

现在电子计算器的使用已相当普遍，一般具有统计功能的计算器，只要按说明书操作，输入观测值后可直接得到 σ、s、$\bar{y}$ 及 $\sum y^2$、$\sum y$ 等值。如果想要计算平方和 SS，可按如下公式计算

$$SS=\sigma^2\times n \tag{1-13}$$

例 4　计算表 1-1 数据的平方和、方差和标准差。

解　∵ $\sum y^2=11.8^2+13.1^2+9.2^2+\cdots+12.2^2+8.2^2=15\ 666.33$

$\sum y=11.8+13.1+9.2+\cdots+12.2+8.2=1\ 224.1$

$n=100$

$C=\frac{(\sum y)^2}{n}=\frac{(1\ 224.1)^2}{100}=14\ 984.2081$

∴ $SS=\sum y^2-C=15\ 666.33-14\ 984.2081=682.1219$

$s^2=\frac{SS}{n-1}=\frac{682.1219}{99}=6.89$

$s=\sqrt{s^2}\doteq 2.6249$

4. 变异系数　标准差是用来度量变异程度的绝对量，可以用来比较性质相同、度量单位一致、平均数较接近的两个样本的变异程度。但是，当平均数相差悬殊、度量单位不同的样本的变异程度用标准差来比较是不适宜的，这时可采用标准差占平均数的百分比这一无量纲的相对量来比较，称之为变异系数，记作 C. V.，其公式为

$$C.V.(\%)=\frac{s}{\bar{y}}\times 100 \tag{1-14}$$

例 5　欲比较两个农场小麦产量的均衡性，测得甲农场小麦平均产量 300 公斤/亩，标

准差30公斤/亩，乙农场小麦平均产量200公斤/亩，标准差25公斤/亩。

解∵ $\bar{y}_{甲}=300 \quad s_{甲}=30$

∴ $C.V._{甲}=\frac{30}{300}\times 100=10\ (\%)$

∵ $\bar{y}_{乙}=200 \quad s_{乙}=25$

∴ $C.V._{乙}=\frac{25}{200}\times 100=12.5\ (\%)$

虽然农场甲的标准差比农场乙大，但其平均产量比后者高，变异系数反而比后者小，因而甲农场小麦生产比乙农场均衡。

第四节　概率与概率分布

肥料试验中由于受许多难以控制的自然环境条件的影响，在试验过程中存在着大量的偶然现象，这就要求我们对这些偶然现象加以研究分析，从中找出规律性的东西。这正是概率论所要研究的课题，所以，概率论是统计分析的基础。例如，在试验中观察或测定得到的是样本，由样本统计数去推断总体的参数时，其可靠程度如何？两个或几个处理之间的差异是纯粹由偶然性变异造成的还是处理效应不同所致？两个变量之间是否存在回归或相关关系？它们的关系密切程度如何？这些问题的回答，必须以概率理论为依据。因此，在讨论各种统计方法之前有必要首先回顾有关概率论的知识。

一、事件与概率

（一）随机事件　自然现象尽管千变万化，若细加考察可以发现它们不外乎如下两类：一类是确定性的，另一类是不确定性的即偶然性的。前者是指在一定条件下必然发生或不发生的现象。例如，在一个大气压下水加热到100℃必然沸腾、同性电荷相互排斥，这种在一定条件下必然发生的事件称为必然事件。与此相反在一定条件下必然不发生的现象称不可能事件，例如，人在无氧条件下长期成活、绿色植物在无光照条件下进行光合作用等。在实际工作中遇到的则是更多的偶然现象，即在一定条件下可能发生也可能不发生的现象，偶然现象也称随机现象或随机事件，数学上常用A、B等字母表示。例如，一粒水稻种子在田间播种后可能发芽也可能不发芽，对某作物施某种肥料后可能有效也可能无效。

个别随机事件的出现带有偶然性，似乎没有规律可循，但对大量同类随机事件进行观察并加以统计分析，可以发现随机事件的发生是具有一定规律性的。例如，掷一枚质地均匀的硬币，币值面朝上或国徽面朝上是不确定的，但如果反复多次投掷这枚硬币，可以发现，出现币值面朝上的次数约为投掷次数的一半，即出现币值面的机会为50%。随机事件的规律性是对一定条件而言的，当条件改变了这个规律也就随之变化了，如果上述硬币质地不均匀，则出现币值面朝上的机会就不再是50%。另一方面，随机事件的规律性是由大量个体组成的一个整体所表现出的某种性质，但这不能简单地归结为个体性质的量的总和。例如，某水稻种子发芽率为90%，这是指大量种子所表现的规律，就其中一粒种子而言要么发芽要么不发芽，而不会发芽90%。这里的90%是针对整个总体而言，并不是针对一个