

全国统计教材编审委员会

“十五”规划教材

DESIGN OF EXPERIMENT
— STUDY GUIDE AND EXERCISE

茆诗松·周纪芴·陈颖·编著

试验设计DOE
— 学习指导与习题

0212.6
20-参

全国统计教材编审委员会

“十五”规划教材

试验设计

—学习指导与习题

Design of Experiment
— Study Guide and Exercise

茆诗松
周纪芈 编 著
陈 颖

中国统计出版社
China Statistics Press



(京)新登字 041 号

图书在版编目(CIP)数据

试验设计——学习指导与习题/茆诗松,周纪芾,陈颖编著.

- 北京:中国统计出版社,2005.6

ISBN 7-5037-4755-2

I. 试…

II. ①茆… ②周… ③陈…

III. 试验设计(数学) - 高等学校 - 教学参考资料

IV. O212.6

中国版本图书馆 CIP 数据核字(2005)第 054105 号

试验设计——学习指导与习题

作 者/茆诗松 周纪芾 陈 颖

责任编辑/乔 杨 吕 军

装帧设计/艺编广告

出版发行/中国统计出版社

通信地址/北京市西城区月坛南街 75 号 邮政编码/100826

办公地址/北京市丰台区西三环南路甲 6 号

电 话/(010)63459084 63266600-22500(发行部)

印 刷/河北天普润印刷厂

经 销/新华书店

开 本/787×1092mm 1/18

字 数/380 千字

印 张/25.125

印 数/1-3000 册

版 别/2005 年 8 月第 1 版

版 次/2005 年 8 月第 1 次印刷

书 号/ISBN 7-5037-4755-2/O·57

定 价/41.00 元

版权所有。未经许可,本书的任何部分不准以任何方式在世界任何地区以任何文字翻印、拷贝、仿制或转载。

中国统计版图书,如有印装错误,本社发行部负责调换。

出版说明

“十五”期间是我国加大教育改革力度，全面推进素质教育，教育体制、教育方法发生重大变革的时期。教材建设必须紧跟教育改革的步伐，建设适应社会主义市场经济和现代化建设需要的高质量教材。为了适应这种新形势的需要，全国统计教材编审委员会制定了《2001—2005年全国统计教材建设规划》（以下简称《规划》），并根据《规划》的要求，主要采取招标的方式组织全国有关院校的专家、学者编写了这批统计学“十五”规划教材。

这批教材力求以第三次全国教育工作会议作出的《中共中央、国务院关于深化改革全面推进素质教育的决定》为指导思想，在充分总结“九五”期间统计教材建设经验的基础上，认真贯彻大胆探索和创新的原则，努力使统计教材具有前瞻性和实用性。选题中不仅包含了一些国内统计研究和教材建设上的空白领域，也包含了统计研究的最新成果。为了配合教师教学、方便地使用这批教材，我们还特地编制了专供教师使用的电子课件，这些课件将在国家统计局统计教育中心网站([hzhttp://edu.stats.gov.cn](http://edu.stats.gov.cn))上挂出，以供需要的教师下载。另外，对于部分教材，我们还编辑出版与之相配套的习题集，以方便教师和学生在学习中使用，也使这批教材在编辑出版上形成一个比较完整的体系。我们相信，这批教材的出版和发行，对于推动我国统计教育改革，加快我国统计教材体系和教材内容更新、改造的步伐，都将起到积极的促进作用，同时，对我国统计教材建设也将起到较好的示范与导向作用。

限于水平和经验，这批教材的编审、出版工作还会有缺点和不足，诚恳欢迎教材的使用单位、广大教师和同学们提出批评和建议。

全国统计教材编审委员会

2004年1月

前 言

这本《试验设计——学习指导与习题》是全国教材编审委员会“十五”规划教材《试验设计》的配套教材。原教材分8章38节编写,本书除第一章(含4节)外,其余7章34节中每节都按内容提要、例题、习题与解答三部分编写。其中内容提要是给出每节重点与要点,并作系统整理而成,是一份很好的复习提纲。若把每节内容提要抽出合并成册,那就是一本试验设计手册。每节的例题是新增的内容,作为教学示例的补充,还对各种试验设计方法的使用和一些理论问题作进一步探讨。每节的习题与解答是本书的主要部分,它把原教材中每道习题都作了较为详尽的解答,有的习题还作了一些深入讨论,以开拓思路,并对原教材中的有误答案作了纠正。另外,原教材后面含有附表15个,篇幅较大。为减少篇幅,本书只附前4个常用分布表,其它附表可参阅原教材。

编写这本《试验设计——学习指导与习题》的目的是想提高学生的分析问题和解决问题的能力,对各种试验设计方法的使用起到示范作用。故对每道习题都作分析,引出试验设计方案,并对数据分析作了较为详尽的叙述。一些中间计算结果也一并列出,便于学生作业查对。我们这样做决不是代替学生去思考,而是想帮助学生思考,使你一进门就能抓住问题的要害。任何知识都要经过自己的思考才能变成自己的知识。假如你自己不去独立思考,而把

这些解答照搬照抄,那对你帮助就太小了,很难成为试验设计的能手。因此独立思考,独立解题才是基本要求。

这本书对实际工作者也是一本很好的参考书。本书是用众多习题与解答叠积而成,每个习题与解答都可看作或大或小的案例。多看此类案例对提高解决试验设计问题是很有好处的。从中你可获得:如何把一个问题转化为试验设计问题?应选用什么试验设计方案去进行试验?数据分析中应注意哪些问题?对数据分析结果如何作出恰当的解释?

本书前三章是茆诗松编写的,中间第四、五章是陈颖编写的,最后三章是周纪芴编写的,最后由茆诗松负责审阅全部书稿,并经几次修正充实,完成此书。在此我们要感谢国家统计局统计教育中心教材处的诸多同志,在他们的多次指导和帮助下,我们顺利地完成了此书编写工作。此外,还要感谢中国统计出版社为此书出版所付出的辛勤劳动。

由于我们水平有限,不当之处在所难免,敬请读者多多批评指正,以便不断修正、充实和提高。

茆诗松 周纪芴 陈 颖

2005 年 3 月 31 日



目 录

第一章 试验设计概要	1
第二章 单因子试验的设计与分析	6
§ 2.1 单因子试验	6
§ 2.2 单因子方差分析	7
§ 2.3 多重比较	25
§ 2.4 效应模型	35
§ 2.5 正态性检验	43
§ 2.6 方差齐性检验	62
第三章 区组设计	71
§ 3.1 随机化完全区组设计	71
§ 3.2 平衡不完全区组设计	99
§ 3.3 链式区组设计	122
第四章 正交设计	140
§ 4.1 多因子试验与正交表	140
§ 4.2 无交互作用情况下的正交设计	143
§ 4.3 有交互作用情况下的正交设计	162
§ 4.4 有重复试验情况下的数据分析	174
§ 4.5 水平数不等情况下的试验设计	182
§ 4.6 裂区法	205
§ 4.7 多指标的数据分析	217
第五章 饱和设计与超饱和设计	225
§ 5.1 饱和设计	225
§ 5.2 图形分析法	227
§ 5.3 数值分析法	234
§ 5.4 超饱和设计	246
第六章 参数设计	251
§ 6.1 参数设计的基本思想	251
§ 6.2 稳健设计	253

目 录

§ 6.3 灵敏度设计	268
§ 6.4 综合噪声因子	280
§ 6.5 动态特性的参数设计	290
第七章 回归设计	316
§ 7.1 回归设计的基本概念	316
§ 7.2 一次回归正交设计	320
§ 7.3 二次回归的中心组合设计	335
§ 7.4 二次回归正交设计	336
§ 7.5 二次回归旋转设计	346
第八章 其它试验设计方法介绍	356
§ 8.1 均匀设计	356
§ 8.2 混料设计	362
§ 8.3 全因子试验的数据分析	370
参考文献	386
附表	387
1. 标准正态分布函数表	387
2. t 分布的 p 分位数表	388
3. χ^2 分布的 p 分位数表	389
4. F 分布的分位数表	390

第一章

试验设计概要

内容提要

1. 试验 目的在于回答一个或几个经过精心构思的问题的实践活动称为试验。

试验广泛应用于工农业生产、国防建设和科学研究中。试验目的上的差异会影响试验的设计与数据的分析。

2. 指标 用于衡量试验结果好坏的特性值称为指标。

常用的指标有两类:定量指标(用测量结果表示的指标)和定性指标(用等级评分等表示的指标)。由于测量数据含有信息丰富,故在试验中要尽量选用定量指标,在不得已场合下才选用定性指标。

3. 因子与水平 影响试验结果的因素称为因子,因子所处的状态称为水平。

因子常可分为两类:可控因子(其水平可控制或可作审慎改变的因子)和不可控因子(在实际操作中不能控制,或难以控制,或要花费昂贵才能控制,或试验人员尚未意识到对试验结果会有影响的因子)。试验应在尽量限制不可控因子的条件下考察可控因子的变化对试验结果的影响,从中寻找可控因子水平的

2 试验设计——学习指导与习题

最佳搭配,使产品的指标接近目标值,有时还要求指标的波动尽量的小。

4. 试验误差 测量值 y 与指标真值 μ 之间的偏差 $\epsilon = y - \mu$ 称为试验误差,简称误差。由于诸多不可控因子的存在,误差是不可避免的,且总是呈现随机性,故 ϵ 是一个随机变量。

据中心极限定理,只要把每个不可控因子都限制在一定范围内,无异常波动出现,随机误差 ϵ 总是呈正态分布 $N(0, \sigma^2)$, 其中标准差 σ 是度量随机误差大小的尺度。 σ 越小试验误差就越小,说明试验的组织实施很好; σ 越大试验误差就越大,说明不可控因子对试验的干扰过大,要努力改进试验的实施。

5. 试验设计 在明确所要考察的可控因子及其水平后对试验进行总体安排,包括试验数据的统计分析方法,称为试验设计。

在试验设计中要注意:①尽量减少试验误差;②尽量减少试验次数;③便于对试验结果进行统计分析。

试验设计的方法有多种,本书将对区组设计、正交设计、饱和设计和超饱和设计、参数设计、回归设计、均匀设计和混料设计等作较为详尽的叙述。

6. 试验设计的三个基本原则:重复、随机化和区组

重复是指一个试验在相同条件下重复进行数次。重复可提供误差的标准差的估计和提高所得结论的可信度。

随机化是指试验材料的分配,各试验点的试验次序都要随机确定。它可使各试验结果相互独立,亦可使不可控因子的影响抵消部分。

区组是指把诸多试验点分为若干小组,使每组内的试验条件相同或近似相同,而组与组之间在试验条件上允许有较大差异,这样的小组被称为区组。使用区组是为了排除或减小试验条件的差异对试验结果的影响,保证统计分析结果的正确性。

7. 试验设计实施的一般指南

- 试验有关人员都要对试验目的清楚、理解和接受。
- 选择好表示试验结果好坏的指标。
- 选定可控因子与水平,尽量限制不可控因子。
- 选择试验设计方案。
- 认真进行试验,记录每次试验结果和试验中出现的异常情况。
- 按预定方法完成数据的统计分析,获得统计推断的结果,做出机理性的解释。
- 对最后结论作出验证试验,确认无误后,写出试验报告并提出今后的行动建议。

例题

例 1.1 试验设计的发展简史。

试验是人们从事的最普遍的活动之一。从家中做菜到技术革新、药品配方、心理测试等都要做试验。一项试验要获得成功常常要做一连串的试验才能实现,这一连串试验要有精心的安排,才能做到撒小网捕大鱼,否则会拖延时间、增加花费,甚至撒大网捕小鱼。愈来愈多的人认识到“试验需要设计”这一道理。近代试验设计的兴起和发展有三个人必需提及,他们是 R. A. Fisher, G. E. P. Box 和田口玄一。

英国统计学家 R. A. Fisher 在二十世纪三十年代在英国一农业试验站对试验设计做了开创性的工作。大家知道,农业试验规模较大、时间长,而且必须妥善处理田间的差异,这些考虑便导致重复、随机化和区组三个基本原则的产生,随之提出一些区组设计方法。为了能有效处理试验数据, Fisher 又提出方差分析方法。数据分析方法从一开始就属于试验设计的一部分。

英国统计学家 G. E. P. Box 在二十世纪五十年代研究化学工业试验的特殊需要——建模与优化中提出回归设计(又称响应曲面设计),其中最著名的是中心组合设计。随着正交性、旋转性和通用性的引入,又提出各种类型的回归设计方案。接着以 D 最优设计为代表的各种最优设计也相继诞生。回归设计的诞生使古老的回归分析只能被动处理数据转化到主动地收集数据和处理数据,从而可以用较少的试验数据获得较为精确的回归模型,这一切都是试验设计的功劳。

日本田口玄一博士早在二十世纪六、七十年代就在日本国内使用正交表安排试验,这不仅逐渐形成独具一格的正交试验设计方法,也对日本产品质量和经济效益提高作出重要贡献。正交试验设计简单易行,不讲究模型,而讲究实效。只要经过半天的培训就能使工程师们学会使用正交试验设计初步。田口另一重要贡献是以降低产品质量特性波动为目的的参数设计(又称稳健设计)。他从工程实际出发把影响质量特性的因子分为可控因子、噪声因子、标示因子、信号因子等,把试验设计不断引向深入,他的动态特性的参数设计把试验设计推向一个高峰。

除此以外,英国的 Scheffe 的混料设计,我国数学家王元和方开泰在 1980 年提出的均匀设计都对试验设计的发展作出了贡献。

例 1.2 试验的指标、因子及其水平的选择。

试验过程相当于一个黑箱操作,黑箱的输入是一些因子,因子就是试验中所研究的变量。若有 20 个因子就有 20 个变量,记为 $x_1, x_2, \dots, x_{20}$ 。黑箱的输出就是指标 y 。试验就是要研究诸变量 x_i 对指标 y 的影响。这种影响可用函数表示

$$y = f(x_1, x_2, \dots, x_{20})$$

这个函数关系在一般场合下人们是未知的,人们想通过试验来研究各个变量的变化对指标 y 的影响。譬如某变量 x_i 的增加可使 y 增加还是减少呢?假如是增加的,那么 x_i 增加一个单位(如 Δx_i) y 增加多少呢?是使得 y 猛增还是微增呢?若是猛增那就是重要因子,若是微增那就是次要因子。在做这些试验前还需对试验的指标、因子及其水平进行选择,使得试验简化而有效。

- 指标的选择。在一个试验中可能有多种指标存在,这时要选择其中最关键的指标,将对其做改进作为试验的目的。这种关键指标最好是定量的,即是可测定的。

- 因子的筛选。把因子分为两类,对不可控因子要尽量设法限制它,譬如灰尘、磁场、湿度等。对可控因子还要再分两类:一类是可确定其水平的因子;另一类是水平要改变的因子。譬如在提高某化学制品的收率试验中有 20 个因子,其中 10 个是不可控因子,设法限制它们,另有 6 个因子根据专业知识和生产经验可以固定在某个水平上,最后剩下的 4 个因子对指标 y 的影响尚不够明确,它们是需要重点考察的对象。但是这 4 个因子变化范围还是要根据专业知识和生产经验确定的。如反应温度在 100°C 到 200°C 之间为宜,催化剂的用量在 2% 到 8% 之间为宜。

- 水平的选择,为了研究可控因子对指标的影响,可以让其取二个或更多个水平,对定量因子若只考察线性效应,取二个水平就足够了,若要考察非线性效应,则需取三个或更多个水平。对定量因子的水平设置要适当分散些,对定性因子的水平的选择就缺乏灵活性。

在上述精心构思的条件下,就可启动试验设计,选用适当的试验设计方案。

例 1.3 不可控因子又称噪声因子,它是引起试验误差的源泉。要罗列一切噪声因子是困难的,但要尽量列举,以便人们研究它、限制它。使它们对试验结果(指标)的影响减少到最小程度。田口玄一建议,可从如下三方面去考察噪声因子,它们是:

- 外部噪声,它是在产品之外引发的一类噪声。如生产过程中的环境温度、湿度、灰尘、磁场、电源电压等,尽管电源电压控制在 220 V,但实际电压还是有波动;

- 内部噪声,它是由产品内部产生的一类噪声。如储期过长、机器老化;
- 产品间的噪声,由产品间的差异引发的一类噪声,如一个电路设计需用 100Ω 的电阻,而从市场上买来的标称值为 100Ω 的电阻未必恰好是 100Ω ,可能是 99Ω ,也可能是 102Ω ,它与标称值的差值对电路输出值会有影响,此种影响是不可控制的。

例 1.4 试验次序的随机化。

试验次序的随机化是为了减少环境条件(不可控因子)的变化所引起的误差而采取的措施。随机化技术多种多样。譬如:

- 为比较某工序中温度(因子 A)的变化对指标 y 的影响,特选二个水平: $A_1 = 80^\circ\text{C}$ 与 $A_2 = 100^\circ\text{C}$ 做试验,各水平都重复 8 次,共做 16 次试验。如何安排试验次序呢?可做 16 张纸签,其中 8 张上写 A_1 ,另 8 张上写 A_2 ,把它放在黑布袋中,随机的一张一张的抽出,可获得如下次序:

$A_1 A_2 A_2 A_1 A_2 A_1 A_1 A_2 A_2 A_2 A_1 A_2 A_1 A_1 A_2 A_1$

试验将按如上次序进行。若允许各水平下重复次数不等,则可用掷硬币的方法,正面做 A_1 ,反面做 A_2 。这样可能导致 A_1 做 9 次重复,而 A_2 做 7 次重复,当没有硬币时,可看手表上的秒针,当秒针在 $1 \sim 30$ 秒之间做 A_1 ,当秒针在 $31 \sim 60$ 秒之间做 A_2 ,这样也可导致重复数不等。重复数相等与不等在统计分析上略有差别,但对分析结果不会产生影响。

- 在多因子试验设计中常要涉及多种水平组合,譬如有 18 个水平组合,那就要做 18 次试验。这时要把试验编号,从 1 到 18,做 18 张纸签,然后按随机抽取的次序进行试验。

- 假如一项试验设计要涉及几十次试验,这时可用随机数表安排试验次序。

第二章

单因子试验的设计与分析

§ 2.1 单因子试验

内容提要

1. 单因子试验的一般概述

设在一个试验中只考察一个因子及其 r 个水平, 因子记为 A , 其 r 个水平记为 $A_1, A_2, \dots, A_r$ 。又设在水平 A_i 下重复进行 m_i 次试验, $i = 1, 2, \dots, r$, 总试验次数为 $n = m_1 + m_2 + \dots + m_r$ 。假如各水平下重复试验次数相等, 即 $m_1 = m_2 = \dots = m_r = m$, 则称此为平衡设计, 否则称为不平衡设计。显然平衡设计是不平衡设计的一个特例。以下叙述对不平衡设计进行。

设 y_{ij} 是在第 i 个水平下的第 j 次重复试验的结果, 这里 i 是水平号, j 是重复号。经过随机化之后, 把所得到的 n 个试验结果按表 2.2.1 顺序排列, 并计算每一水平下数据的和与均值。

2. 单因子试验的三项基本假定

(1) 正态性。在水平 A_i 下的数据 $y_{i1}, y_{i2}, \dots, y_{im_i}$ 是来自正态总体 $N(\mu_i, \sigma_i^2)$ 的一个样本, $i = 1, 2, \dots, r$ 。

(2) 方差齐性。 r 个正态总体的方差相等, 即 $\sigma_1^2 = \sigma_2^2 = \dots$

表 2.2.1 单因子试验的数据

因子 A 的水平	数据	和	均值
A_1	$y_{11}, y_{12}, \dots, y_{1m_1}$	$T_1 = y_{11} + y_{12} + \dots + y_{1m_1}$	$\bar{y}_1 = T_1/m_1$
A_2	$y_{21}, y_{22}, \dots, y_{2m_2}$	$T_2 = y_{21} + y_{22} + \dots + y_{2m_2}$	$\bar{y}_2 = T_2/m_2$
$\vdots$	$\dots\dots\dots$	$\dots$	$\dots$
A_r	$y_{r1}, y_{r2}, \dots, y_{rm_r}$	$T_r = y_{r1} + y_{r2} + \dots + y_{rm_r}$	$\bar{y}_r = T_r/m_r$

$$= \sigma_r^2 = \sigma^2.$$

(3)随机性。所有数据 y_{ij} 都相互独立。

3. 单因子试验中要研究的问题

● 各水平均值 $\mu_1, \mu_2, \dots, \mu_r$ 是否彼此相等? 这要用单因子方差分析去研究。

● 假如这些水平均值不全相等, 哪些均值间的差异是重要的? 如何估计它们? 前者用多重比较去研究, 后者用最小二乘法去估计。

4. 单因子试验模型

单因子试验的三项基本假定用到试验数据上去, 可得到如下统计模型:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2, \dots, r, j = 1, 2, \dots, m_i \quad (2.1.1)$$

其中

- y_{ij} 是因子 A 的第 i 个水平下第 j 次试验结果;
- μ_i 是因子 A 的第 i 个水平的均值, 是待估参数;
- ϵ_{ij} 是因子 A 的第 i 个水平下第 j 次试验误差, 它是随机变量, 诸 ϵ_{ij} 是相互独立同分布的随机变量, 并且共同分布为 $N(0, \sigma^2)$, σ^2 为误差方差, 也是待估参数。

由模型(2.1.1)可知, 诸 $y_{ij} \sim N(\mu_i, \sigma^2)$, 且相互独立。

5. 诸水平均值的估计

用最小二乘法可得诸水平均值 μ_i 的估计, 它们是

$$\hat{\mu}_i = \bar{y}_i, i = 1, 2, \dots, r.$$

§ 2.2 单因子方差分析

内容提要

1. 单因子方差分析是在诸方差相等的条件下对诸正态均值是否彼此相等

的假设检验问题。所涉及的一对假设是：

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$$

$$H_1: \mu_1, \mu_2, \cdots, \mu_r \text{ 不全相等}$$

若在显著性水平 α 上拒绝原假设 H_0 , 则称因子 A 在显著性水平 α 上是显著的, 即诸水平均值 μ_i 不全相等, 亦可简称因子 A 显著; 否则称因子 A 不显著, 即诸水平均值 μ_i 在统计意义下可看作彼此相等, 它们之间的差异完全是由试验误差引起的。

2. k 个数据 $y_1, y_2, \cdots, y_k$ 对其均值 $\bar{y}$ 的偏差的平方和(有时简称平方和)定义为:

$$Q = \sum_{j=1}^k (y_j - \bar{y})^2$$

它是用来度量 k 个数据间的差异(即波动)大小的一个重要的统计量, Q 愈大数据愈分散, Q 愈小数据愈集中。

平方和中独立偏差的个数称为该平方和的自由度, 记为 f , k 个数据的平方和的自由度是 $f = k - 1$ 。自由度是偏差平方和的重要参数。下列三个性质可以用来简化平方和的计算:

- $Q = \sum_{j=1}^k (y_j - \bar{y})^2 = \sum_{j=1}^k y_j^2 - \frac{T^2}{k}$, 其中 $T = y_1 + y_2 + \cdots + y_k$ 。

- $y_1 + c, y_2 + c, \cdots, y_k + c$ 的平方和仍然为 Q 。

- 若 $c \neq 0$, 则 $cy_1, cy_2, \cdots, cy_k$ 的平方和为 $Q' = c^2 Q$, 或 $Q = Q' / c^2$ 。

在正态总体假设下, 平方和有如下基本定理:

定理 2.2.1 假如 $y_1, y_2, \cdots, y_k$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 则有:

- 样本均值 $\bar{y} \sim N(\mu, \sigma^2/k)$;
- 样本的偏差平方和 Q 与方差 σ^2 之比 $Q/\sigma^2 \sim \chi^2(k-1)$;
- $\bar{y}$ 与 Q 相互独立。

3. 总平方和的分解公式

单因子试验共有 $n = m_1 + m_2 + \cdots + m_r$ 个数据, 详见表 2.2.1, 它的总平均值 $\bar{y}$ 与总偏差平方和 S_T 分别为

$$\bar{y} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{m_i} y_{ij} = \frac{1}{n} \sum_{i=1}^r m_i \bar{y}_i$$

$$S_T = \sum_{i=1}^r \sum_{j=1}^{m_i} (y_{ij} - \bar{y})^2, f_T = n - 1$$

其中 $\bar{y}_i$ 为水平 A_i 下 m_i 次重复试验结果的均值。

总平方和有如下分解公式：

$$S_T = S_A + S_e, \quad f_T = f_A + f_e$$

其中 S_A 为因子 A 的平方和，又称组间平方和。 S_e 为误差平方和，又称组内平方和。它们分别为：

$$S_A = \sum_{i=1}^r m_i (\bar{y}_i - \bar{y})^2, f_A = r - 1$$

$$S_e = \sum_{i=1}^r \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2, f_e = \sum_{i=1}^r (m_i - 1) = n - r$$

它们的计算公式分别为：

$$S_T = \sum_{i=1}^r \sum_{j=1}^{m_i} y_{ij}^2 - \frac{T^2}{n}, f_T = n - 1$$

$$S_A = \sum_{i=1}^r \frac{T_i^2}{m_i} - \frac{T^2}{n}, f_A = r - 1$$

$$S_e = S_T - S_A, f_e = n - r$$

4. 统计分析

定理 2.2.2 在单因子方差分析的三个基本假定下，有

$$E(S_e) = (n - r)\sigma^2$$

$$E(S_A) = (r - 1)\sigma^2 + \sum_{i=1}^r m_i (\mu_i - \mu)^2$$

其中 $\mu = \frac{1}{n} \sum_{i=1}^r m_i \mu_i = E(\bar{y})$ 。

由此还可推得如下一些结果：

- 误差平方和 $MS_e = S_e / (n - r)$ 是 σ^2 的无偏估计。
- 在原假设 $H_0: \mu_1 = \mu_2 = \cdots = \mu_r$ 成立下，因子 A 的均方和 $MS_A = S_A / (r - 1)$ 亦是 σ^2 的无偏估计。当原假设 H_0 不成立时，因子 A 的均方和 MS_A 将会偏