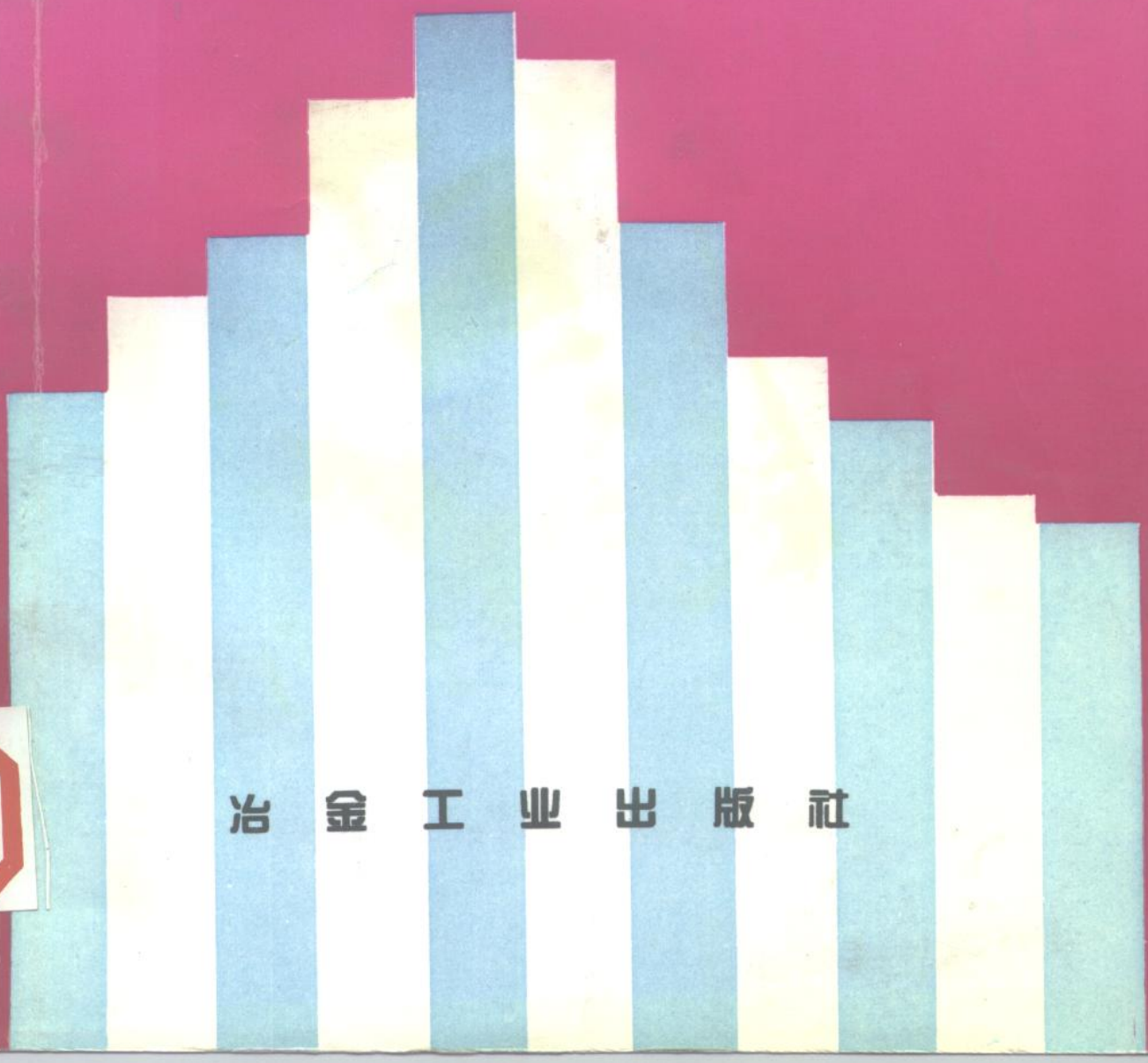
试验设计与数据处理

主　编　吴贵生
副主编　于治福　于淑政　魏效玲



冶　金　工　业　出　版　社

试验设计与数据处理

主　编　吴贵生
副主编　于治福　于淑政　魏效玲

北　京
冶金工业出版社
1997

内 容 简 介

本书是针对高等院校、科研单位和厂矿企业等进行科学研究和生产试验过程中试验方案设计、试验数据处理以及寻求科学结论而编写的。其主要内容为：数理统计的基本理论，正交试验设计及方差分析，SN 比试验设计，产品三次设计，一元线性回归分析及多元线性回归分析等。叙述力求理论联系实际，并着重科研和实践应用。

本书可供科研人员和厂矿企业等单位工程技术人员使用，也可作为高等院校理工科高年级学生及研究生的选修教材。

图书在版编目(CIP)数据

试验设计与数据处理/吴贵生主编. —北京：冶金工业出版社，1997.1

ISBN 7—5024—1982—9

Ⅰ.试… Ⅱ.吴… Ⅲ.①科学—试验—设计②实验数据—数据处理 Ⅳ.N33

中国版本图书馆 CIP 数据核字(96)第 25276 号

出版人 卿启云 （北京沙滩嵩祝院北巷 39 号，邮编 100009）

河北建筑科技学院印刷厂印刷；冶金工业出版社发行；各地新华书店经销

1997 年 1 月第 1 版，1997 年 1 月第 1 次印刷

787mm×1092mm 1/16；15 印张；358 千字；234 页；1—3000 册

定价：24.00 元

前　言

试验设计与数据处理是研究随机现象中变量之间关系的一种数理统计方法。试验设计着重研究如何科学地设计试验方案，正确地分析试验结果，从而寻求最佳生产条件和工艺参数，以开发新产品和改进产品的生产工艺，达到提高产品质量和经济效益的目的。而数据处理主要通过生产和科研中的试验数据，进行点或区间的估计，分析各因素对考察指标影响的显著性，寻求经验公式，制定产品的新标准，探索新工艺和新配方，研究气象与地震预报以及建立生产过程自动控制中数学模型等方面的研究。该门科学在农业、机械、冶金、化工、交通运输，以及经营管理和科学研究等各个领域中得到广泛的应用和发展，并且都取得了显著成绩和非常可观的经济效益。

本书的原型是1989年吴贵生教授编写的《正交试验设计及回归分析》，为河北建筑科技学院校内使用教材，主要用于对本科生开设选修课。经过六年的教学实践，对该教材作了较大改动，增删后于1994年重新编写成《试验设计与数据处理》。又经过两年校内扩展使用，于1996年经过少量改动后，编写成了本书。

参加本书编写的人员有：吴贵生（1、5章），魏效玲（2、3章），于淑政（4、6、7章），于治福（8、9、10章）。

在编写中，力求理论联系实际，结合实例说明基本原理和方法。但由于作者水平有限，书中可能仍有不妥之处，敬请读者批评指正。

编　者

1996年10月

目　录

1 绪 论

试验设计与数据处理是以概率论、数理统计及线性代数为理论基础，经济地、科学地安排试验和分析处理试验结果的一项科学技术。其主要内容是讨论如何合理地安排试验方案和科学地分析处理试验数据和结果，从而达到解决生产中和科学研究中的实际问题。它要求除具备概率论、数理统计和线性代数等基础知识外，还应有较深和较广的专业知识和丰富的实际经验。只有这三者紧密地结合起来，才能取得良好的效果。

1.1 试验设计与数据处理的概念和意义

在科学研究和生产中，经常要做许多试验，并通过试验数据的分析，企图寻求问题的解决办法。如此，就存在着如何安排试验和如何分析试验数据和试验结果的问题，也就是如何进行试验设计和数据处理的问题。

1.1.1 试验设计

试验设计是数理统计学中的一个较大的分支。它主要研究试验数据的合理获得方法，其内容十分丰富。

如果试验安排得合理，试验次数不多，就能得到满意的结果；若试验安排得不合理，试验次数既多，结果还往往又不能令人满意。试验次数过多，既浪费大量的人力和物力，有时还会由于时间拖得很长，使试验条件发生变化而导致试验失败。因此，如何合理地安排试验方案是值得研究的一个重要问题。一项科学合理的试验安排方法应能做到以下三点：(1)试验次数尽可能地少；(2)便于分析和处理试验数据；(3)能得到满意的结果。

对于单因素的试验可以采用 0.618 法、黄金分割法、分数法、平行线法、交替法和调优法等去解决，并且在生产中都取得了显著成效。

而对于多因素的试验安排方法有正交试验设计、SN 比试验设计、产品三次设计、完全随机化试验设计、随机区组试验设计、拉丁方试验设计和正交拉丁方试验设计等。其中目前应用最多的是正交试验设计。该方法是依据数据的正交性(即均匀搭配)来进行试验方案设计的。由于该方法应用广泛，为了方便起见，已经构造出了一套现成规格化的正交表。根据正交表的表头和其中的数字结构就可以科学地挑选试验条件(因素水平)合理地安排试验。它的主要优点是：(1)能在众多的试验条件中选出代表性强的少数试验条件；(2)根据代表性强的少数试验结果数据可推断出最佳的试验条件或生产工艺；(3)通过试验数据的进一步分析处理，可以提供比试验结果本身多得多的对各因子的分析；(4)在正交试验的基础上，不仅可作方差分析，还能使回归分析等数据处理的计算变得十分简单。

1.1.2 数据处理

数据处理也是数理统计学中的一部分重要内容。它主要研究试验测量或观察数据分析计算的处理方法，从而得出可靠和规律性的结果，依据这个规律和结果对工业生产、农业生产、天

气、地震等进行预报和控制，进而掌握和主宰客观事物的发展规律，使之服从和服务于人类。

数据处理的方法很多，如参数估计、假设检验、方差分析和回归分析等。其中参数估计主要对某些重要参数进行点估计和区间估计；方差分析是分析各影响因素对考察指标影响的显著性程度；回归分析是如何获得反映事物客观规律性的数学表达式；假设检验是判断各种数据处理结果的可靠性程度。

1.2 试验设计与数据处理的发展和应用

数理统计是应用概率论的基本理论，而试验设计与数据处理则是数理统计的重要分支和组成部分。因此试验设计与数据处理是在概率论和数理统计的基础上不断完善和发展起来的。

早在17世纪，随机试验是与掷硬币和掷骰子等游戏紧密联系在一起的。硬币和骰子就是最简单的概率模型。数学家赫依琴斯(Huygens)就曾预言过，不要小看这些博弈游戏，它有更重要的应用。

18世纪，法国科学家巴芬(Buffon)对概率论在博弈游戏中的应用深感兴趣，发现了用随机投币试验计算 π 的方法。

1908年，统计学家戈塞特(Gosset)在推导 t 分布的同时，通过抽样的试验方法对总体方差和样本方差的分布进行了研究。

在20世纪初，英国生物统计学家费歇(R. A. Fisher)在统计学的基础上首创了“试验设计”方法。在农业、生物学和遗传学等方面都取得了丰硕成果，使农业大幅度增产。费歇于1935年出版了他的“试验设计”专著。从此开创了试验设计这门新的应用技术科学。

20世纪30年代和40年代，英、美、苏把试验设计推广到采矿、冶金、建筑、纺织、机械和医药等行业，都取得了很好的经济效益。

二次世界大战后，日本从英、美引进了这一技术。于1949年日本的田口玄一博士在试验设计的基础上又创造了“正交试验设计”方法。

1952年田口玄一在日本东海电报公司运用 $L_{27}(3^{13})$ 正交表进行正交试验取得成功。之后，在日本工业生产中得到了迅速推广。仅在1952年至1962年的10年中，试验达到了100万项。其中三分之一的项目都取得了十分明显的效果，并获得了极大的经济效益。其中之一，如他们运用正交试验设计对电讯研究所研制的“线形弹簧继电器”的数十个特性值2000多个变量进行了试验研究，经过7年的努力，制造出了比美国先进的产品。这一产品本身只有几美元，而试验研制花费了几百万美元，但研究成果给该研究所带来几十亿美元的利益。几年后，他们的竞争对手美国西方电器公司不得不停产，转而从日本引进这种先进的继电器。在日本“正交试验设计”技术已成为企业界、工程技术界的研究人员和管理人员必备的技术知识，已成为工程师的共同语言的一部分。

1957年田口玄一博士在正交试验设计的基础上又提出了“信噪比设计”和“产品三次设计”。

信噪比SN(Signal－Noise Ratio)通常被用来表示信号功率与噪音功率的比值，即 $\eta=\frac{S}{N}=\frac{\text{信号功率}}{\text{噪音功率}}$，可以用来评价仪器和设备质量的好坏。

产品三次设计(即系统设计 System design、参数设计 Parameter design、容差设计 Toler-

ance design)是使整机的元器件或零件各参数合理搭配，对于某些地方，采用低级价廉的元器件或零部件仍能保证整机质量稳定和高的可靠性。

在二次世界大战后，日本的工业飞速发展的原因之一，就是在工业领域里普遍推广和应用正交试验设计和产品三次设计的结果。日本的电子产品能够打进美国市场，畅销世界各国的秘诀之一也是运用了正交试验设计和产品三次设计这个得力工具。因此，日本把正交试验设计技术誉为"国宝"是有一定道理的。

数据处理是在大量试验数据基础上，也可在正交试验设计的基础上，通过数学处理和计算，揭示产品质量和性能指标与众多影响因素之间的内在关系，还可以回归出数学表达式，在生产和科研中得到广泛应用，并起到了重要作用和显著效果。

我国从50年代开始研究"试验设计"这门科学，60年代末中国科学院统计数学研究室在"正交试验设计"的观点、理论和方法上都有新的创见，编写了一套较为适用的正交表，创立了简单易懂的"正交试验设计"法。1973年以来，许多科研、生产单位和大专院校应用正交试验设计方法解决了不少科研和生产中的关键问题。例如上海地区，从1978年至1984年有227个单位应用了正交试验设计方法，其中103个单位取得了成效。上海高压油泵厂生产的32MPa高压轴向柱塞泵原来由于摩擦副的结构参数配合不好，经常发生异常发热的质量问题。通过正交试验设计找到了最佳参数组合，不仅降低了止推板和斜盘的精度要求(不平度从0.005放宽到0.01mm)，而且成品合格率由原来的69%提高到了90%以上。

产品三次设计在我国起步较晚，北京761厂在高频负反馈电路中采用了产品三次设计，仅该电路中3DG44GC晶体管正确选择一项，一年可增加经济效益3万余元。杭州电视机厂对西湖牌黑白电视机的OTL电路的中点电压设计中运用了产品三次设计方法。不仅找到了高可靠性、高稳定性等优化方案，而且仅此一项全年收益达13591元。

数据处理在我国各领域也发挥了很大作用，如预报气象和病虫害、制定自动控制中的数学模型、以及参数估计和检验等都要应用到这一数学工具。

70年代以前，我国许多工厂企业为了提高机电产品质量，对元器件或零部件采用层层筛选，专挑质量高、成本高的一级品组装整机，这样使整机昂贵，但质量未必就好。70年代以来，我国很多工厂企业对机电产品积极开展了正交试验设计和产品三次设计，使元器件或零部件的参数合理搭配，从而使我国的很多机械设备和电气产品(如电视机、电冰箱、收录机等)的可靠性和稳定性大幅度提高，许多产品打入了国际市场。

1.3 试验设计与数据处理的基本概念

1.3.1 常用术语

1. 试验考察指标

在试验设计和数据处理中，我们通常根据试验和数据处理的目的而选定用来考察或衡量其效果的特性值称为试验考察指标。试验考察指标可以是产品的质量、成本、效率和经济效益等。

试验考察指标分为定量指标和定性指标两大类。定量化指标(如精度、粗糙度、强度、硬度、合格率、寿命和成本等)可以通过试验直接获得，它方便计算和数据处理。而定性指标(如颜色、

气味、光泽等)不是具体数值,一般要定量化后再进行计算和数据处理。

试验考察指标可以是一个,也可以是几个,前者称为单考察指标试验设计,后者称为多考察指标试验设计。

2. 试验因素

对试验考察指标产生影响的原因或要素称为试验因素。

例如在合金钢 40Cr 的淬火试验中,淬火硬度与淬火温度(如 770、800、850℃)和冷却方式(如水冷、油冷、空冷)有关。其中淬火温度和冷却方式是试验因素,而淬火硬度是试验考察指标。

除上述的试验因素外,在试验过程中由于测量、仪器和环境条件等影响,也会影响到试验考察指标,称这类因素为误差因素。因素一般用大写字母 A、B、C、…来标记。

3. 因素水平

试验因素在试验中所处的状态、条件的变化可能会引起试验指标的变化,我们把因素变化的各种状态和条件称为因素的水平。在试验中需要考虑某因素的几种状态时,则称该因素为几水平因素。如上例 40Cr 的淬火试验中,淬火温度为 770、800、850℃三种状态,则淬火温度这个试验因素为三水平因素。因素的水平应是能够直接被控制的,并且水平的变化能直接影响试验考察指标有不同程度的变化。水平通常用数字 1、2、3…表示。

1.3.2 常用统计量

1. 极差

极差是一组数据中的最大值与最小值之差,其计算公式为:

$$R=x_{\max}-x_{\min} \tag{1-1}$$

极差表示一组数据的最大离散程度,它是统计量中最简单的一个特征参数。在试验设计中会经常用到。

2. 一组数据之和与平均值

在试验设计和数据处理中,设有几个观察值 $x_1,x_2,\cdots x_n$,我们称之为一组数据。这组数据之和与平均值分别为:

$$T=x_1+x_2+\cdots+x_n=\sum_{i-1}^{n}x_i \qquad (i=1,2,\cdots,n) \tag{1-2}$$

$$\overline{A}=\frac{T}{n}=\frac{1}{n}\sum_{i-1}^{n}x_i \qquad (i=1,2,\cdots,n) \tag{1-3}$$

3. 偏差

偏差也称为离差。偏差在数理统计中一般有两种,一种是与期望值 μ 之间的偏差,另一种是与平均值 $\bar{x}$ 之间的偏差。在试验设计和数据处理中往往不知道期望值 μ,而很容易知道平均值 $\bar{x}$,所以常常把与平均值 $\bar{x}$ 之间的偏差作统计量进一步分析研究。

设有 n 个观察值 $x_1,x_2,\cdots,x_n$,则把每个观察值 $x_i(i=1,2,\cdots,n)$ 与平均值 $\bar{x}$ 的差值称为与平均值之间的偏差,简称为偏差。

很显然,与平均值 $\bar{x}$ 之间的偏差的总和为零,即:

$$(x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})=0 \qquad (i=1,2,\cdots,n) \qquad (1-4)$$

4. 偏差平方和与自由度

由式(1－4)可知，一组数据与其平均值的各个偏差值有正、负或零，因此各偏差值的总和为零，所以偏差和不能表明这组数据的任何特征。如果消除掉各个偏差正、负的影响，即以偏差平方和作为这组数据的一个统计量，则偏差平方和能够表征这组数据的分散程度，常以 S 表示。

设有 n 个观察值 $x_1,x_2,\cdots,x_n$，其平均值为 $\bar{x}$，则偏差平方和为：

$$S^2=(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad (i=1,2,\cdots,n) \qquad (1-5)$$

关于自由度的问题可以通过下例来说明。

例如有 4 个数据 3、4、6、7，由于它们之间有一个关系式：

$$\frac{3+4+6+7}{4}=5 \qquad (1-6)$$

数学上称这 4 个数据中只有 4－1(此处"1"指一个关系式)个对其平均值是独立的，也就是说，上述 4 个数据的平均值已知为 5，且其中 3 个数据也已知分别为 3，4，6，那末第四个数据 7 就可由该关系式所确定，这说明第四个数据 7 受其它 3 个独立的数据所束约。自由度是独立数据的个数，所以该例中的自由度 $f=4-1=3$。若有 n 个观察值，与平均值 $\bar{x}$ 的偏差平方和的自由度应为 $n-1$ 个。

5. 方差与均方差

方差也称平均偏差平方和，它表示单位自由度的偏差大小，即偏差平方和 S^2 与自由度 f 的比值 V，V 即是方差。

$$V=\frac{S^2}{f} \qquad (1-7)$$

均方差也称标准偏差。由方差 V 的计算式(1－7)可知，方差 V 的量纲为观察数据 x_i 的量纲的平方，为了与原特性值的量纲相一致，可采用方差 V 的平方根 $\sqrt{V}$ 作为一组数据离散程度的特征参数，即：

$$\sqrt{V}=\sqrt{\frac{S^2}{f}}=\sqrt{\frac{1}{n-1}\sum_{i}^{n}(x_i-\bar{x})} \qquad (i=1,2,\cdots,n) \qquad (1-8)$$

2 样本及其分布

在生产和科学实验中，会碰到大量的数据，如何从这些杂乱无章的数据中，取出有用的情报，帮助解决问题，用于指导生产，为此，需要对数据进行处理。

数据处理，在数理统计中，就是通过随机变量的部分观察值来推断随机变量的特性，例如分布规律和数字特征等。数理统计是具有广泛应用的一个数学分支，它以概率论为理论基础，根据试验或观察得到的数据，对研究对象的客观规律作出合理的估计与判断。

2.1 总体与样本

2.1.1 总体

在数理统计中，人们所研究对象的全体称为总体，而组成总体的每个单元称为个体。任何总体的某项指标，是按一定的规律分布的，因而是一个随机变量，常用大写字母 X、Y、Z 等表示。例如，一批灯泡，以其使用寿命指标来衡量它的质量，若规定寿命低于 1000h 者为次品，要求确定这批灯泡的次品率。显然这个问题可以归结为求灯泡寿命 X 这个随机变量的分布函数 $F(x)$，若已求得 $F(x)$，则 $P\{X<1000\}=F(1000)$就是所求的次品率。如果把每只灯泡的寿命都测出来，问题就得到了圆满的解决。但由于寿命试验是破坏性的，一旦获得全部的试验结果，这批灯泡的灯丝就全部烧断了。因此，是不现实的。

再如有一批晶体管，共 10 万只，若想了解它的某个指标(如直流放大系数)，由于测试不会损坏合格的晶体管，所以最理想的办法是逐一测试。然而，限于人力、物力和时间，也不可能逐一测试。因此只能取总体的一部分来进行试验或测试，然后根据这些试验数据推断总体的指标。

总体的类型随研究的问题而定。它所包含的个体数可以是有限的，也可以是无限的。例如，研究某厂某天生产的某种灯泡的次品率，总体是有限的，其个体数就是该天生产的这种灯泡的总数。但为研究方便，仍以研究灯泡的寿命 X 的分布为例，我们常把相同条件下所生产的这种灯泡的寿命全体，看成一个总体。显然，它是一个无限总体，因而灯泡寿命 X 是一个连续型随机变量。

2.1.2 样本

从总体 X 中随机抽取若干个体观察其某种数量指标的取值过程，称为抽样。从总体中抽取一个个体以作观察或试验，这个抽出的个体在未观察前，它可能取某个值，也可能取另一个值，因此，它也是一个随机变量，常用带下标的大写字母 X_i、Y_i 等表示。

从一个总体中，随机地抽取 n 个个体 $X_1, X_2, \cdots, X_n$，这样取得的$(X_1, X_2, \cdots, X_n)$称为总体 X 的一个样本。样本中个体的数目称为样本容量。对于样本来说，一次抽取、观察的结果是 n 个具体的数据 $x_1, x_2, \cdots, x_n$，称为样本$(X_1, X_2, \cdots, X_n)$的一个观察值，简称样本观察值。而样本观察值的所有可能取值的全体称为样本空间。

为了使抽取的样本能反映总体的性质，要求抽样是完全随机的和独立的，并且每抽取一个个体后总体的成分不变，每次抽样的观察值互不影响，还要求 $X_i(i=1,2,\cdots,n)$ 必须与总体有相同的分布函数 $F(x)$。这样的抽样方法称为简单随机抽样。

如果一个样本中每个个体 X_i 都与总体 X 有相同的分布且相互独立，则称这个样本为简单样本。

综上所述，我们给出如下定义：

设 X 为具有分布函数 $F(x)$ 的随机变量，若 $X_1,X_2,\cdots,X_n$ 为具有相同分布函数 $F(x)$ 的相互独立的随机变量，则称 $(X_1,X_2,\cdots,X_n)$ 为来自总体 X 的容量为 n 的简单随机样本，简称样本。它们的观察值 $x_1,x_2,\cdots,x_n$ 又称为 X 的 n 个独立的观察值。

2.2 样本分布函数与统计量

2.2.1 样本分布函数

实际应用中，总体的分布函数 $F(x)$ 往往是未知的，数理统计的任务之一就是由样本的特性来推断总体的分布。由概率论知，若 $(X_1,X_2;\cdots,X_n)$ 为来自总体 X 的一个样本，则 $X_1,X_2,\cdots,X_n$ 的联合分布函数为

$$F(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i) \tag{2-1}$$

又若 X 具有概率密度 $f(x)$，则 $X_1,X_2,\cdots,X_n$ 具有联合概率密度

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i) \tag{2-2}$$

前面已提到，简单随机样本能很好地反映总体的情况，为了推断总体的分布，这里给出样本分布函数的定义。

设总体 X 的 n 个独立的观察值，按大小次序排列成

$$x_1\leqslant x_2\leqslant\cdots\leqslant x_n$$

若 $x_k\leqslant x<x_{k-1}$，则不大于 x 的观察值的频率为 k/n，因而函数

$$F_n(x)=\begin{cases}0 & x<x_1\\ \dfrac{k}{n} & x_k\leqslant x<x_{k-1}\\ 1 & x_n\leqslant x\end{cases} \tag{2-3}$$

等于在 n 次重复独立试验中，事件 $\{X\leqslant x\}$ 的频率。称之为样本分布函数或经验分布函数。

按经验分布函数的定义，容量为 n 的简单样本 $(X_1,X_2,\cdots,X_n)$ 的经验分布函数 $F_n(x)$ 可能取的值为 $0,\frac{1}{n},\cdots,\frac{k}{n},\cdots,\frac{n-1}{n},1$。“$F_n(x)=\frac{k}{n}$”表示服从总体分布 $F(x)$ 的随机变量 X 取小于 x 值这一事件在 n 次重复独立试验中恰好出现 k 次，也就是说在这 n 次试验中，事件 $\{X\leqslant x\}$ 的频率为 $\frac{k}{n}$。所以按贝努利大数定理，对一个任意的正数 ε，有

$$\lim_{n\to\infty}P\{|F_n(x)-F(x)|\geqslant\varepsilon\}=0$$

在 oxy 面上作出 $y=F(x)$ 及 $y=F_n(x)$ 的图形 C 及 C_n，如图 2－1 所示，该等式表明：对任意给定的正数 ε，在横坐标上任意指定值 x 处，只要 n 足够大，C_n 与 C 上点的纵坐标之差的绝对值

不小于 ε 的概率就能小于任意给定的正数。即，当 n 足够大时，C_n 的图形在不等式

$$F(x)-\varepsilon<y<F(x)+\varepsilon$$

所定的带状区域以外的概率可以小于任意的正数。因此当 n 很大时，样本分布函数 $F_n(x)$ 将近似地等于总体分布函数。

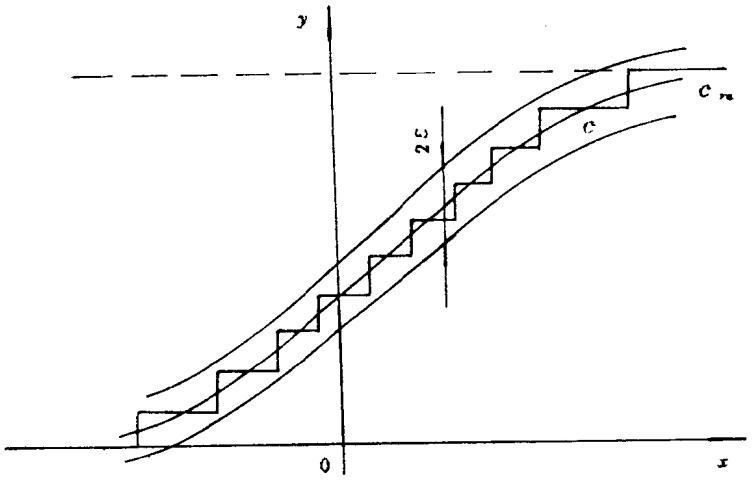


图 2－1　样本分布函数

还可以进一步证明下列格利文科定理

当 $n\to\infty$ 时，$F_n(x)$ 依概率 1 关于 x 均匀地收敛于 $F(x)$，即

$$P\{\lim_{n\to\infty}\max_{-\infty<x<\infty}|F_n(x)-F(x)|=0\}=1$$

这就是我们用样本推断总体的依据。

2.2.2　统计量

对于给定的一个样本的实现 $x_1,x_2,\cdots,x_n$，可以计算它的数字特征，并冠以样本两字，以示与总体数字特征的区别。如，样本 k 阶原点矩为：

$$m_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\quad k=1,2,\cdots \tag{2-4}$$

样本 k 阶中心矩为

$$m'_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k,\quad k=1,2,\cdots \tag{2-5}$$

样本平均值为

$$\bar{x}=m_1=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{2-6}$$

样本方差为

$$s^2=\frac{n}{n-1}m'_2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2-7}$$

s^2 的正平方根 s 称为样本标准离差。

$\bar{x},s^2,m_k,m'_k$ 分别为下列随机变量的观察值：

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{2-8}$$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{2-9}$$

$$M_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\quad k=1,2,\cdots \tag{2-10}$$

$$M'_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,\quad k=1,2,\cdots \tag{2-11}$$

这些随机变量仍分别称为样本平均值、样本方差、样本 k 阶原点矩及样本 k 阶中心矩。可以证明，只要总体的 r 阶矩存在，样本 r 阶矩以概率 1 收敛于总体的 r 阶矩。

在数理统计中，除了用样本矩外，还需要用到另外一些样本数字特征。为此引入如下定义：

设 $(X_1,X_2,\cdots,X_n)$ 为总体 X 的一个样本，$g(X_1,X_2,\cdots,X_n)$ 为一个连续函数。如果 g 中不包含任何未知参数，则称 $g(X_1,X_2,\cdots,X_n)$ 为一个统计量。

如果 $x_1,x_2,\cdots,x_n$ 是样本 $(X_1,X_2,\cdots,X_n)$ 的观察值，则 $g(x_1,x_2,\cdots,x_n)$ 是统计量 $g(X_1,X_2,\cdots,X_n)$ 的一个观察值。

如 $\overline{X}$、S^2、M_k 及 M'_k 都是统计量，其中 $\overline{X}$ 和 S^2 是两个特别重要的统计量。统计量都是随机变量，如果总体的分布函数为已知，则统计量的分布是可以求得的。

2.2.3 顺序统计量

定义　设总体 X 具有连续的分布函数 $F(x)$，$(X_1,X_2,\cdots,X_n)$ 为总体 X 的一个样本，若将样本观察值 $x_1,x_2,\cdots,x_n$ 按从小到大的次序排列：

$$x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(k)}\leqslant\cdots\leqslant x_{(n)}$$

规定统计量 $X_{(k)}$ 为取上述排列的第 k 个值为观察值的随机变量，则称 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 为顺序统计量。其中最小项为 $X_{(1)}=\min(X_1,X_2,\cdots,X_n)$，最大项为 $X_{(n)}=\max(X_1,X_2,\cdots,X_n)$。而统计量 $X_{(n)}-X_{(1)}=\max(X_1,X_2,\cdots,X_n)-\min(X_1,X_2,\cdots,X_n)$ 称为极差。

顺序统计量 $X_{(k)}$ 有自己的分布函数 $F^*_k(x)$。设样本中最大项 $X_{(n)}$ 的分布函数记为 $F^*_n(u)$，则

$$\begin{aligned}F^*_n(u)&=P\{X_{(n)}\leqslant u\}=P\{X_1\leqslant u,X_2\leqslant u,\cdots,X_n\leqslant u\}\\&=[F(u)]^n\end{aligned} \tag{2-12}$$

样本中最小项 $X_{(1)}$ 的分布函数记为 $F^*_1(v)$，则

$$\begin{aligned}F^*_1(v)&=P\{X_{(1)}\leqslant v\}=1-P\{X_{(1)}>v\}\\&=1-P\{X_1>v,X_2>v,\cdots,X_n>v\}\\&=1-[P\{X>v\}]^n=1-[1-F(v)]^n\end{aligned} \tag{2-13}$$

如果总体 X 有概率密度函数 $f(x)$，则 $X_{(n)}$ 和 $X_{(1)}$ 的概率密度函数分别为

$$f^*_n(u)=n[F(u)]^{n-1}\cdot f(u) \tag{2-14}$$

$$f^*_1(v)=n[1-F(v)]^{n-1}\cdot f(v) \tag{2-15}$$

一般地，顺序统计量 $X_{(k)}$ 的分布函数记为 $F^*_k(y)$，$k=1,2,\cdots,n$。我们用 $P\{y\leqslant X_{(k)}<y+\Delta y\}$ 表示事件“在样本 $X_1,X_2,\cdots,X_n$ 中，有一个落在 $[y,y+\Delta y)$，$k-1$ 个落在 $(-\infty,y)$，其余 n

$-k$ 个落在$[y+\Delta y,\infty)$"的概率，于是

$$\Delta F_k^*(y)=P\{y\leqslant X_{(k)}<y+\Delta y\}$$

$$=\binom{n}{1}\Delta F(y)\cdot\binom{n-1}{k-1}[F(y)]^{k-1}\cdot[1-F(y+\Delta y)]^{n-k}$$

因此：

$$f_k^*(y)=\lim_{\Delta y\to 0}\frac{\Delta F_k^*(y)}{\Delta y}$$

$$=n\binom{n-1}{k-1}[F(y)]^{k-1}\cdot[1-F(y)]^{n-k}\cdot f(y) \tag{2-16}$$

$$F_k^*(y)=n\binom{n-1}{k-1}\int_{-\infty}^{y}[F(x)]^{k-1}\cdot[1-F(x)]^{n-k}\cdot f(x)\mathrm{d}x \tag{2-17}$$

2.3 直方图和秩

2.3.1 直方图

上节提到，当样本容量很大时，样本分布函数将近似地等于总体的分布函数。工程上常用直方图来求得随机变量的概率密度函数 $f(x)$，下面介绍直方图的作法。

设 $x_1,x_2,\cdots,x_n$ 是一组数据，为了掌握它变化的规律性，对它加以整理。首先选取 a,b 两数，使得 a 适当小于 $\min(x_1,x_2,\cdots,x_n)$，b 适当大于 $\max(x_1,x_2,\cdots,x_n)$，并用分点 $t_i(a=t_0<t_1<\cdots<t_m=b)$将区间$[a,b]$分成 m 个小区间$[t_{i-1},t_i)$，$i=1,2,\cdots,m$，每个小区间长度为 Δt_i。然后统计 $x_1,x_2,\cdots,x_n$ 落入$[t_{i-1},t_i)$中的个数，设落入$[t_{i-1},t_i)$内有 n_i 个。把每个小区间内的数据称为一组，这样，整批数据就被分成了 m 组。作

$$f_n(x)=\begin{cases}0 & x<a\\ \dfrac{n_i}{n\Delta t_i} & t_{i-1}\leqslant x<t_i,\quad i=1,2,\cdots,m\\ 0 & b\leqslant x\end{cases} \tag{2-18}$$

并绘出 $f_n(x)$的图形。由于 $f_n(x)$的图形呈直方形，因此称为直方图。

在作直方图时，要注意分组问题。组数的多少往往影响着直方图反映数据分布的效应，如果组数过多，每组所占的区间就很狭窄，这不仅造成计算上的麻烦，而且也有可能因随机因素导致某组内数据稀少，甚至没有，这样直方图就不能较好地反映数据所提供的信息；如果组数过少，那么落在每组内的数据就较多，从而掩盖了组内数据变化的情况。在实际应用中，一般当数据多于 100 个时，宜分为 10～20 组，当数据少于 50 个时，分为 5～6 组为宜。

直方图是总体概率密度曲线的一个估计，在这里 n_i 可以看成事件$\{t_{i-1}<x<t_i\}$在 n 次重复，独立试验中出现的频率。令 $\tilde{x}_i=\frac{1}{2}(t_{i-1}+t_i)$。称 $\tilde{x}_i$ 为组中值，则当 n 充分大时，有

$$f(x)\approx f(\tilde{x}_i)\approx\frac{n_i}{n\Delta t_i}=f_n(x)$$

通常称 $f_n(x)$为频率密度函数。

例 2－1 研究某届学生数学成绩的分布，随机抽查了 120 名学生进行测试，得到如下数据：

58　92　69　67　84　94　57　74　74　83

51	62	64	62	72	58	56	76	76	83
83	56	72	98	74	84	68	83	79	85
59	59	73	72	54	69	78	68	82	84
79	78	78	79	77	82	84	82	84	82
81	86	94	79	74	54	72	68	63	45
93	79	42	55	68	70	64	73	73	54
46	64	74	77	76	69	68	66	54	72
50	72	62	63	90	74	54	73	89	68
87	74	86	75	50	82	67	62	88	44
69	88	72	74	55	90	66	76	64	74
65	73	72	69	68	75	60	79	77	80

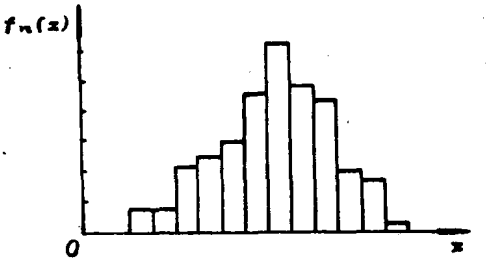


图 2－2　直方图

作出它的直方图。

解　该组数据最小值为 42，最大值为 98，取 $a=40, b=100$，将$[a,b)$分成 12 个小区间，为计算方便，各小区间长度取等值，$\Delta t_i=\dfrac{100-40}{12}=5$，这样将全部数据分成 12 组，计算各组的频数，列于表 2－1 中，由表 2－1 即可画出直方图，如图 2－2 所示。由图可以看出，学生成绩大致是服从正态分布的。

表 2－1　数据统计表

组　号	范　围	频数登记	频数 n_i	频率 $\frac{n_i}{n}$	频率密度
1	[95,100)	一	1	0.008	0.0016
2	[90,95)	正	6	0.050	0.0100
3	[85,90)	正 T	7	0.058	0.0116
4	[80,85)	正 正 正	16	0.133	0.0266
5	[75,80)	正正正 F	18	0.150	0.0300
6	[70,75)	正正正正 F	23	0.192	0.0384
7	[65,70)	正正正 T	17	0.142	0.0284
8	[60,65)	正正	11	0.092	0.0184
9	[55,60)	正	9	0.075	0.0150
10	[50,55)	正 F	8	0.066	0.0132
11	[45,50)	T	2	0.017	0.0034
12	[40,45)	T	2	0.017	0.0034
Σ			120	1	—

2.3.2　秩

在工程试验中，有时会遇到试验样本价格昂贵，试验周期长，限于人力、物力，一般样本容