西北大学出版社

SHIYAN SHEJI YU
SHUJU CHULI

# 试验设计与数据处理

陈立宇　张秀成◎主编

试验设计与数据处理

西北大学出版社

# 试验设计与数据处理

SHIYAN SHEJI YU SHUJU CHULI

陈立宇　张秀成◎主编



西北大学出版社

**图书在版编目(CIP)数据**

试验设计与数据处理/陈立宇,张秀成主编.—西安:西北大学出版社,2014.11

ISBN 978-7-5604-3389-9

Ⅰ.①试… Ⅱ.①陈…②张… Ⅲ.①化学工程—化学实验—试验设计②化学工程—化学实验—数据处理 Ⅳ.①TQ016

中国版本图书馆 CIP 数据核字(2014)第 276436 号

**试验设计与数据处理**

**主　　编**:陈立宇　张秀成

**出版发行**:西北大学出版社

**地　　址**:西安市太白北路 229 号

**邮　　编**:710069

**电　　话**:029-88305287

**经　　销**:全国新华书店

**印　　装**:西安华新彩印有限责任公司

**开　　本**:787mm×960mm　1/16

**印　　张**:13.75

**字　　数**:240 千字

**版　　次**:2014 年 11 月第 1 版　2014 年 11 月第 1 次印刷

**书　　号**:ISBN 978-7-5604-3389-9

**定　　价**:28.00 元

# 前　言

《试验设计与数据处理》是为高等学校化学工程学科以及有关专业本科生编写的一本教材。

化学工程学科是以实验为基础的学科，大量实验数据的分析处理工作是对事物本质探讨的过程，数据处理方法由简单的数值分析过渡到复杂数学模型，从理论到方法上均有长足的发展。试验设计研究如何科学地设计试验方案、正确地分析试验结果，从而以较少的实验获得较为丰富的信息，达到提高产品质量和经济效益的目的，也成为科技工作者必须掌握的一门技术。

本书主要介绍化学工程学科中常用的试验设计与数据处理方法，内容包括单因素试验设计方法、正交试验设计方法、方差分析、回归分析和误差分析等。

本书讲稿曾作为西北大学化学工程学科《试验设计与数据处理》课程的教材使用，作者在此基础之上，经过充实、取舍而形成了书稿目前的内容和体系，我们编写了此书作为正式教材出版。

全书共分六章，第一章介绍了试验设计原理与数理统计基础，第二章、第三章和第四章介绍了数据处理的回归与相关分析等，第五章、第六章介绍了单因素与多因素试验设计方法，以上内容为化学工程学科有关专业本科生必须掌握的内容，为进行科学研究和撰写毕业论文提供设计和分析方法。

在本书的编写工作中，参考了大量的文献资料，谨向原作者表示最真诚的感谢。

由于编者水平有限，不妥之处在所难免，敬请批评指正。

编　者

2014 年 11 月

# 目　　录

# 绪　论

为了推动化学工程科学的发展，常常要进行大量的科学试验，这些试验都涉及试验方法和试验数据处理的问题。

进行试验之前必须解决的问题是试验方法的问题，即如何获得数据或者如何合理地进行试验设计，以较少的试验代价收集到较多的有代表性的试验数据，从而得出可靠的结论，达到试验的目的。通过试验获得的数据，常常表现出不同程度的变异。产生这种变异的原因有的已被人们所了解，还有许多内在和外在的因素未被人们所认识，由于这些因素的作用，使得试验数据普遍具有变异性。因此，在实际研究过程中必须解决的第二个问题就是试验数据的处理问题，即如何科学地整理、分析所收集的具有变异的试验数据，从而揭示出隐藏在其内部的规律性。

合理地进行试验设计以及科学地处理所收集的数据是本课程的根本任务，这在化工研究中的作用主要体现在以下两个方面。

(1)提供试验设计方法

试验设计主要是指试验因素的选取、试验水平的确定、试验单位的分组和试验顺序的安排。合理的试验设计能减少试验次数，控制和降低试验误差，提高试验的精确性，为试验数据的统计分析提供必要的数据。试验设计中常用的试验设计方法有完全随机设计、随机区组设计、正交设计、均匀设计、回归正交设计和混料设计等。

试验设计的目的是避免系统误差，控制和降低随机误差，考察无偏估计处理效应，从而对样本所在的总体作出可靠、正确的推断。试验设计主要解决如何合理地收集必要且具有代表性数据的问题。

(2)提供整理和分析数据的方法

试验数据可以整理成为多种表达形式，整理试验数据的基本方法是根据试验数据的特性将其进行统计分析，可以表征出试验数据的特性。试验数据处理是数理统计学的原理和方法在化工实验中的应用。

对试验数据进行统计分析其中重要的内容是差异显著性检验。显著性检验的目的是判断数据的变异主要由什么原因产生，是由于试验的数据间(如不同原

料、不同工艺、不同配比）有实质性差异，还是由于无法控制的偶然因素所引起的。显著性检验的方法很多，有显著性检验、方差分析、$\chi^2$检验等。统计分析的另一个重要内容是对变量间的关系进行研究，通过对实验数据进行相关分析、回归分析，可以揭示出变量间的内在联系，为新产品的研发、产品质量的预测和控制提供依据。

# 第一章　概率论及数理统计基础

在对化工过程的分析和认识中，通过实验的方法我们可以得到一系列试验数据，而这些试验数据是带有随机误差的随机数据，用数理统计方法分析处理这些试验数据，就可以得到关于过程本质的一些相关结论。统计学是以概率论为基础的，而概率论是研究随机现象的数学理论。本章主要讲解概率论和数理统计的基础概念，它是本书的基础理论部分，这些基础知识在以后的章节中会得到应用。

本章的主要内容是围绕着随机变量展开的，包括随机变量的描述和随机变量的数理统计方法。

随机变量的描述主要有以下两种方法：

1)随机变量的分布函数描述，包括正态分布、$\chi^2$分布、$t$分布、$F$分布。

2)随机变量的数字特征描述，例如，对于正态分布其数字特征为数学期望$\mu$和方差$\sigma$。

随机变量的数理统计方法主要有以下两种方法：

1)对总体的参数进行估计：点估计，包括数字特征法、顺序统计量法、最大似然法；区间估计，对总体参数给出一定信度下的估值范围。

2)对总体的参数进行假设检验：通过局部观察判断总体的分布是否具有指定的特征。

## 1.1　概率论与数理统计的基本概念

从数学观点看，自然界中的一切现象都存在着规律性。这种规律可分为两大类，即自然现象一类为具有确定性规律的现象，另一类为具有统计规律的现象。

具有确定性规律的现象，即各参数之间呈现出一定的函数关系，或者说在一定条件下可以预言它的结果的现象，如“理想溶液中温度不变时，每一组分在溶液中的饱和蒸汽分压应当与它在溶液中的摩尔分率成正比”。在一定条件下必然会出现的事情，称为必然事件，如“水在标准大气压下被加热至100℃时必定沸腾”；反过来在一定条件下必然不会出现的事情，称为不可能事件，如“水在标准大气压

下被加热至100℃时不沸腾”是一定不会出现的。显然必然事件的反面就是不可能事件。

具有统计规律的现象，各参数初看起来似乎毫无规律，但通过多次的实践分析研究，人们发现它具有统计规律。换句话说，对这类现象不能预言它的结果。例如，掷硬币的试验中，可能出现正面，也可能出现反面，这类不能预言结果的现象称为随机事件或简称事件。就个别事件来看，随机事件的结果带有偶然性，似乎毫无规律，但是在偶然性的背后隐藏着必然性的客观规律。概率论和数理统计就是要从大量的同类随机现象中揭示其规律——统计性规律。

概率论是现代数学的重要分支之一，是研究若干随机事件中数量规律的一门学科，其所具有的统计规律，当测定次数愈多时表现得愈加明显，概率论在实验数据的误差分析以及试验设计中有着广泛的用途。

### 1.1.1 概率

如果事件 $A$ 属于任一随机试验，即这一试验是不能预言结果的，则在 $n$ 次独立试验中 $A$ 出现 $m$ 次时的频率为

$$f(A) = \frac{m}{n} \tag{1-1}$$

例如，掷100次硬币得到53次正面，得正面这一事件的频率为0.53。当试验次数逐渐增大时，频率将围绕在某一常数附近摆动，而逐渐稳定于一个特定的常数。因此，在统计的意义上，随机事件的频率存在着一个极限值，叫做事件 $A$ 的概率，记为 $P\{A\}$，其值就等于这个特定的常数值。将这种描述随机事件发生可能性大小的概念抽象出来，就构成了概率。

$$\lim_{n \to \infty} \frac{m}{n} = P\{A\} \tag{1-2}$$

概率是事件发生频率的抽象，所以说，概率是随机事件的客观存在的一个性质，它的现实意义就是频率。正是在这个基础上，概率论才具备了坚实的客观基础。反过来，也正是基于这一点，才有了利用概率论认识自然现象的可能性。

### 1.1.2 随机变量及分类

事件的概率仅表示试验的某一个结果发生的可能性大小，若要全面了解试验，则必须知道试验的全部可能结果发生的概率，也即必须知道随机试验的概率分布。为了深入研究随机试验，我们引入随机变量的概念。

随机变量是随机试验(如掷硬币等)结果(也称随机事件)的一个数量性的表征,也就是说,由个别试验结果所得到的每个值 $x_i$ 都是一个随机变量,它发生的概率为 $P_i$。随机变量的值与它的概率之间的关系,称为这一随机变量的分布。一个随机变量不仅表明了试验的各种可能的结果,而且还表明了每一种可能结果发生的概率,因此随机变量完整地描述了一个试验。对随机试验概率分布规律的研究就转化成了对随机变量概率分布的研究。

通过实验我们可以得到一系列试验结果,而这些试验结果是带有随机误差的随机数据,在统计学中将试验所得到的随机数据看成是具有一定分布的随机变量。随机试验的可能结果,有的可以用数值表示,但有的与数值并无直接关系。非数值结果的随机试验不便于进行数学处理,为了进行定量的数学处理,必须把随机试验的结果数量化,这就是引进随机变量的原因。

对一次试验来说,其结果有多种可能。每一种可能的结果都可以用一个数值来表示,并把这些数值作为随机变量的取值范围。如果表示试验结果的变量,其可能的取值可以一一列举(即只取有限个取值或者无限个孤立的取值),且以各种确定的概率取这些不同的值,则称该随机变量为离散型随机变量。如果表示试验结果的变量其可能的取值为某范围内的任何数值,且在其取值范围内的任一区间中取值时,其概率是确定的,则称该随机变量为连续型随机变量。

### 1.1.3　总体与个体

根据研究目的确定的符合指定条件的研究对象的全体称为**总体(或母体)**。一般情况下,我们关心的总是对象性质的数量特征,将表示此数量特征的随机变量记作 $X$,因此母体就是全部 $X$ 可能值的集合。$X$ 的分布就是母体的分布。

个体 构成母体的每个基本元素称为**个体**。

样本从母体中抽取部分个体所组成的集合称为**样本**,记为 $x_i(i=1,2,\cdots,n)$。

样本容量样本中个体的数目称为**样本容量**。

统计学中一般是通过样本来了解总体。这是因为总体常常是无限的、假想的,即便是有限的但包含的个体数目也相当多,要获得全部观测值需花费大量人力、物力和时间,或者观测值的获得带有破坏性。我们研究的目的是要了解总体,然而能观测到的却是样本,通过样本来推断总体是数理统计分析的基本特点。为了能可靠地从样本来推断总体,要求样本具有一定的容量和代表性。只有从总体随机抽取的样本才具有代表性。然而,样本毕竟只是总体的一部分,尽管样本

具有一定的容量，也具有代表性，通过样本来推断总体也不可能是百分之百的正确，通常有很大的可靠性，但也具有一定的错误率，这是数理统计分析的又一特点。

参数是由总体计算得到的描述总体数量特征和规律的数值，也称总体特征数；由样本计算得到的特征数称为统计量，也称样本特征数。参数是客观存在的，但常常是未知数，一般用相应的统计量作为其估计值，统计量受抽样变动的影响。

## 1.2 随机变量的描述

在统计学上，习惯用随机变量 $X$、$Y$ 等表示总体，用 $x_1$、$x_2$ 等表示随机变量可能取的值。对于任何一个随机变量，可用随机变量的概率分布函数来描述其特性，这是对随机变量最完善的描述。当我们不需要知道随机变量的全部性质而只要知道其中的一部分性质时，可用随机变量的数字特征来描述其特性。

### 1.2.1 随机变量描述方法一——分布函数法

随机变量可分为连续型随机变量和离散型随机变量，二者均为在自然科学和技术科学领域中经常碰到的两类变量，而前者尤为常见。

1.2.1.1 连续型随机变量的分布函数

(1)分布函数

随机变量 $X$ 的概率分布情况可以用分布函数 $F(x)$ 来表示。

如果考虑连续的随机变量 $X$ 在区间 $(a,b)$ 内可取任何值，并设 $x$ 是此区间内的某个值，如图 1－1 所示，分布函数在 $x$ 处的值等于随机变量取值小于或等于 $X$ 这样一个随机事件的概率：

$$F(x) \equiv P\{X \leqslant x\} \qquad (1-3)$$

$F(x)$ 称为随机变量 $X$ 的概率分布函数，简称分布函数。

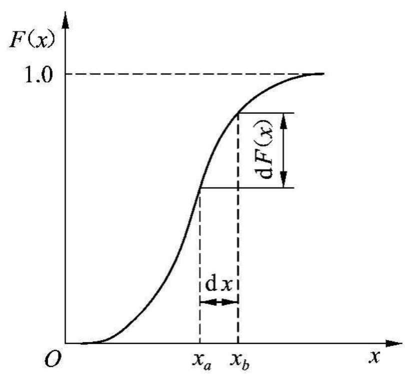

图 1－1 连续型随机变量的分布函数图

从图中可以看出，分布函数具有以下性质：

1) $0 \leqslant F(x) \leqslant 1$，$-\infty < x < +\infty$。

2) $\lim\limits_{x \to -\infty} F(x) = 0$，$\lim\limits_{x \to +\infty} F(x) = 1$。

3)当 $X$ 增大时,$F(x)$非降。

4)对于 $X$ 落在区间$(a,b)$间的概率为

$$P\{x_a \leqslant X \leqslant x_b\} = P\{X \leqslant x_b\} - P\{X \leqslant x_a\} = F(x_b) - F(x_a)$$

若已知随机变量 $X$ 的分布函数 $F(x)$,即可求得 $X$ 落在任意区间上的概率,分布函数 $F(x)$可用来描述随机变量的统计规律性,也可用数学分析的方法研究随机变量。

(2)分布密度函数

分布函数 $F(x)$也可表示为

$$F(x) = P\{X \leqslant x\} = \int_{-\infty}^{x} f(x)\mathrm{d}x \tag{1-4}$$

函数 $f(x)$称为随机变量 $X$ 的概率分布密度函数,同样表达随机变量的概率分布情况,简称分布密度函数。

分布密度函数 $f(x)$的定义为,如果考虑连续的随机变量 $X$ 在区间$(a,b)$内可取任何值,随机变量 $X$ 在区间 $x$ 和 $x+\mathrm{d}x$ 之内取值时的概率为 $\mathrm{d}P$。

显然,此概率 $\mathrm{d}P$ 与 $x$ 有关,并与 $\mathrm{d}x$(当其为无穷小时)成正比。因此得

$$\frac{P\{x \leqslant X \leqslant x + \mathrm{d}x\}}{\mathrm{d}x} = \frac{F(x+\mathrm{d}x) - F(x)}{\mathrm{d}x} = \frac{\mathrm{d}F(x)}{\mathrm{d}x} \equiv f(x) \tag{1-5}$$

也可写为

$$\mathrm{d}P = f(x)\mathrm{d}x$$

曲线 $y = f(x)$ 则称为随机变量 $X$ 的分布密度曲线,如图 1-2 所示。

分布密度函数曲线的重要性质为

1)$f(x) > 0$。

2)对一切实数值,分布密度曲线应该有 $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$。

这是因为在一次实验中随机变量 $X$ 的取值在 $-\infty$ 和 $+\infty$ 范围内,构成一个必然事件,亦即分布密度函数的归一化性质。

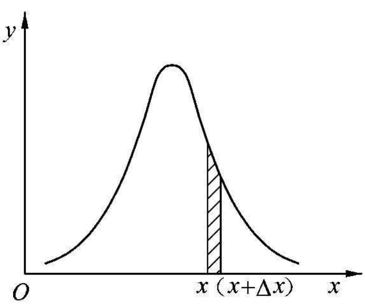


图 1-2 连续型随机变量的分布密度函数图

3)如果知道分布密度函数 $f(x)$,该随机变量在区间 $x_1$ 和 $x_2$ 内的概率可计算如下:

$$F(x)=P\{x_1<X<x_2\}=\int_{x_1}^{x_2}f(x)\mathrm{d}x=F(x_2)-F(x_1)$$

4)在分布密度函数的连续点处，$F'(x)=f(x)$。

事实上，分布密度函数是一个很重要的性质，我们随后对随机变量所做的定义就是由分布密度函数定义的，随机变量的统计规律性也是由分布密度函数来描述的。

1.2.1.2 离散型随机变量的分布函数

若随机变量 $X$ 只取数轴上有限个或无限个孤立值 $x_1,x_2,\cdots$，并且这些值的取得有对应的确定的概率 $p_1,p_2,\cdots$，则称随机变量 $X$ 是离散分布的。即

$$P\{X=x_i\}=p_i \tag{1-6}$$

$$F(X)=P\{X\leqslant x\}=\sum p_i \quad i=1,2,\cdots \tag{1-7}$$

离散型随机变量的分布函数 $F(X)$有如下性质：

$$p_i\geqslant 0 \quad i=1,2,\cdots$$

$$\sum p_i=1$$

离散型随机变量的概率分布具有 $p_i\geqslant 0$ 和 $\sum p_i=1$ 这两个基本性质。

典型的离散型随机变量的分布密度函数曲线如图 1—3 所示。

对于一个随机变量来说，分布函数 $F(x)$是它的最完善的描述。当我们知道一个随机变量的分布函数时，不仅知道该变量取哪些值，还可以知道它是以什么样的概率取这些值的。对于许多科学技术问题，人们的经验证明，并不需要知道它的全部概率性质，往往只需知道这个随机变量的几个数字特征，能反映该变量的变化值的集中位置及离散程度就够了。例如，对于服从正态分布的随机变量最常用的数字特征是数学期望和方差。

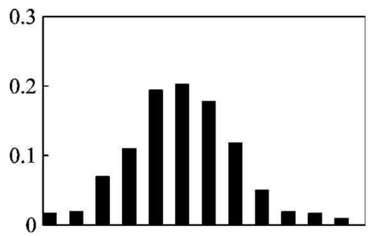

图 1—3 离散型随机变量的分布密度函数图

## 1.2.2 随机变量描述方法二——数字特征法

数字特征及其有关运算在概率论及数理统计中起着重要作用。数字特征按照其作用或性质来分类基本上是两类：一类是位置特征或集中性数字特征，用以表示随机变量可能取值的集中趋势，属于这类的特征数有数学期望、众数及中位数等；另一类是离散性数字特征，用以表示随机变量可能取值的离散程度，属于这

类的数字特征有平均差、方差、极差及变异度等。其中以数学期望和方差最为重要。

1.2.2.1　数学期望

随机变量的数学期望用符号 $\mu$ 表示，是随机变量在数轴上取值的集中位置，它说明随机变量 $X$ 的值大多数集中在哪里。数学期望是一个能够粗略地(未必精确)满足这一要求的平均值，因此有时数学期望也称为均值。但要注意，这里所谓的均值并非指的算术平均值，以后会看到算术平均值只是若干种平均值中的一种特定的形式，与数学期望不同。

数学期望在一定程度上起着“代表”随机变量的作用。例如，我们说某化学产品的纯度是 95%，绝不是说这种产品的纯度绝对准确地为 95%，而只表示该产品的纯度通过大量化验数据的平均值或数学期望，因而起着“代表”的作用。

数学期望的计算可由数学期望算子 $E(x)$ 表示。

数学期望的定义为

$$E(X)=\int_{-\infty}^{+\infty}x\mathrm{d}F(x) \tag{1-8}$$

数学期望的计算公式为

离散型随机变量　$$\mu=E(X)=\sum_{i}x_ip_i \tag{1-9}$$

连续型随机变量　$$\mu=E(X)=\int_{a}^{b}xf(x)\mathrm{d}x \tag{1-10}$$

**【例 1－1】**在射击比赛中，假设以得分多少来评定射击技术的优劣，已知甲、乙二人在射击中得分多少这个随机变量的分布列表如下：

甲运动员

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $P\{X\}$ | 0.4 | 0.1 | 0.5 |

乙运动员

| $y$ | 1 | 2 | 3 |
|---|---|---|---|
| $P\{Y\}$ | 0.1 | 0.6 | 0.3 |

问甲、乙二人谁的射击技术较好，在未来的射击比赛中可预期谁将获胜。

**解**　甲、乙二人射击技术的优劣，可以从在射击中平均每次得分多少的角度来进行比较，也就是说，比较他们得分的数学期望：

$$\mu=E(X)=\sum_{i=1}^{3}x_ip_i=1\times0.4+2\times0.1+3\times0.5=2.1$$

$$\mu=E(Y)=\sum_{i=1}^{3}y_ip_i=1\times0.1+2\times0.6+3\times0.3=2.2$$