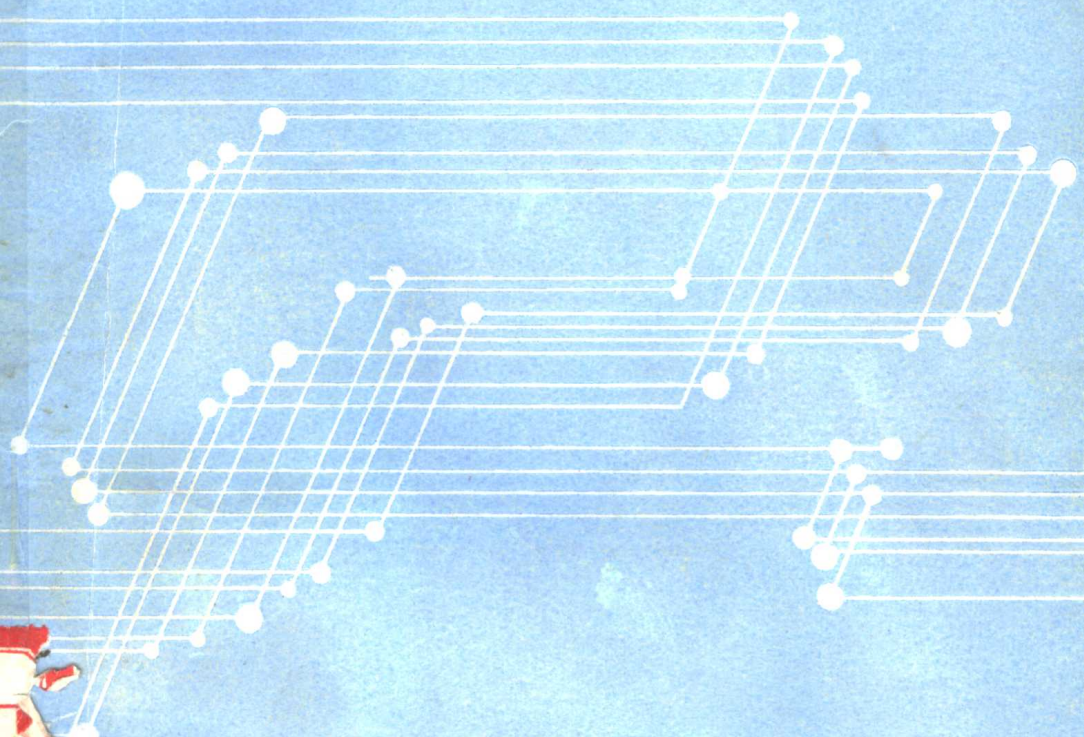
朱中南
戴迎春 编著

化工数据处理与实验设计



烃加工出版社

化工数据处理与实验设计

朱中南　戴迎春　编著

烃 加 工 出 版 社

内 容 提 要

本书从概率论和数理统计的基本知识入手，介绍了建立化工过程数学模型的参数估计、模型检验和实验设计等方面的内容，并通过对例题的分析介绍了计算机在化工过程模型化方面的应用。

本书可供化工、石油化工等有关专业的科技人员阅读，也可供高等院校有关专业的研究生作教材和参考书

化工数据处理与实验设计

宋中南 魏迎春 编著

烃加工出版社出版

同兴印刷厂排版

仰山印刷厂印刷

新华书店北京发行所发行

850×1168毫米 32开本 12 5/8印张 327千字 印1—3500

1989年7月北京第1版 1989年9月北京第1次印刷

ISBN 7-80043-035-9/TQ·026 定价：5.35元

前　言

过程模型化目前已成为化工过程开发研究工作中的一个主要手段。为了建立定量的化工过程数学模型，除了需要有关学科的专业基础理论外，主要还是依赖实验方法去探求和掌握被研究对象的规律，研究各变量之间的相互关系。

解决这个任务的方法可概括为两类：分析法和经验法。分析法在对研究对象的物理和化学过程深入理解的基础上，经简化、推导求得一个数学模型，再通过实验，求得模型中的待定参数，并对模型进行检验，求得一个能够反映过程机理的，又在实验范围内与实验数据相一致的数学模型。经验法则并不要求对过程本质有深切的理解，它完全依赖于实验，根据数据输入、输出的关系，用统计回归分析方法建立一个能够描述变量外部联系的数学模型。

可见，无论分析法还是经验法，最终都必须依赖于实验才能使化工过程模型化得以实现。这主要是由于化学工程研究对象的复杂性造成的，如对数量众多的物理参数以及复杂的边界条件难以作定量地描述。此外，研究对象的复杂性还表现在多变量上，各个变量之间交互作用，并大都呈非线性关系，加之实验测定中总存在一定的误差，因而使得从实验的综合结果中分析出各个变量之间的相互关系更加困难。传统的单变量的实验方法存在着工作量过大的弊病，有时甚至缺乏实际可能性。因此，极需研究过程模型化方法中有关数据处理和实验设计等一些共性问题，如线性、非线性代数模型和常微分模型的参数估值方法，模型的统计性检验和筛选以及有效的实验安排方法。上述问题不仅在化工领域而且在其它工程领域都是十分重要和非常活跃的课题。据统计

70年代国外就有数本专著出版，国内在80年代初也出版了一些论文和译著。但是，这些书籍往往过于数学化，系统性也欠理想，因此，作为我国目前工程技术人员继续教育的教材是不适宜的。我们在科研工作中深深感到，如果能从应用的角度对此作较系统的介绍，将有助于读者较快地掌握这一领域的一些基本方法和基本原理，并使数据处理和实验设计与电子计算技术有机结合，成为过程开发研究工作的一个重要手段。

本书的内容可概括为预备知识、参数估计、模型检验和实验设计四大部分。第一章介绍了有关概率论和数理统计方面的基本知识，专为缺少这方面知识的读者参考。已有这方面知识的读者可以直接从第二章开始阅读。第二章介绍有关误差理论和误差分析。这两章构成了第一部分的内容。后三部分，则是对线性或非线性模型进行叙述，其中第三章和第六章分别介绍了线性代数模型的参数估计、模型检验和实验设计。第四、五、七章分别介绍了非线性代数模型和常微分模型的参数估计、模型检验和实验设计。最后，在附录中还对各章主要的计算方法，配上以ALGOL系统719语言编写的电子计算机原程序，用例子说明调用方法，供研究人员参考。

本书是我们以1979年和1983年为华东化工学院研究生编写的“过程模型化方法”以及“数据处理和实验设计”讲义为基础，作了补充改写而成。由于作者水平有限，实践经验不足，一定有不少错误和缺点，请读者批评指正，不胜感激。

朱中南、戴迎春

1985年12月

目　录

前言

第一章　概率和数理统计基础知识 …… 1

第一节　概述 …… 1

第二节　随机事件的概率 …… 3

第三节　随机变量的分布 …… 7

第四节　联合概率分布 …… 11

第五节　随机变量的特征数字——数学期望，方差 …… 12

第六节　正态分布 …… 17

第七节　样本的平均值和方差 …… 22

第八节　χ^2分布 …… 25

第九节　t分布和F分布 …… 28

第十节　点估计和区间估计 …… 32

第十一节　假设检验 …… 36

本章主要符号表 …… 48

第二章　实验测定量的误差估计 …… 49

第一节　概述 …… 49

第二节　随机误差的特性 …… 51

第三节　测量值的表示方法（已知标准差） …… 55

第四节　实验测定误差标准差的估计方法 …… 57

第五节　测量值的表示方法（已知标准差的估计值） …… 59

第六节　随机误差方差的传递 …… 62

本章主要符号表 …… 71

第三章　线性代数模型的回归分析方法 …… 72

第一节　概　述 …… 72

第二节　线性代数模型参数的最小二乘估计法 ……75
第三节　参数最小二乘估计值的数学期望和方差 ……79
第四节　回归方程的显著性检验 ……86
第五节　回归系数的显著性检验 ……94
第六节　逐步回归分析法 ……101
第七节　预测和控制 ……114
本章主要符号表 ……121
第四章　非线性模型参数估计方法 ……122
第一节　概述 ……122
第二节　模型的通式 ……124
第三节　参数估计的目标函数 ……128
第四节　非线性代数模型的最小二乘估计方法 ……136
第五节　常微分模型参数估计方法 ……142
第六节　用最优化方法解决参数估计问题 ……155
第七节　参数估计的几个具体问题 ……171
第八节　参数估计值的置信域 ……175
第九节　参数估计应用举例 ……184
本章主要符号表 ……201
第五章　模型检验 ……202
第一节　概述 ……202
第二节　方差分析 ……204
第三节　相关系数及其显著性检验 ……207
第四节　残差分析 ……214
第五节　非本征参数法 ……220
第六节　根据模型的固有特征来鉴别模型 ……230
第七节　模型化过程的实例 ……239
本章主要符号表 ……246
第六章　回归正交实验设计 ……248
第一节　概述 ……248

第二节　一次回归正交实验设计所处理的数学模型……284
第三节　一次回归正交实验设计的基本思想……251
第四节　一次回归正交实验设计步骤……256
第五节　一次回归正交设计的应用……269
第六节　二次回归正交实验设计所处理的数学模型
和组合实验设计……274
第七节　二次回归的组合实验设计的正交性……277
第八节　二次回归正交实验设计步骤……286
第九节　二次回归正交实验设计举例……290
第十节　利用回归正交实验设计寻找过程最佳条件……297
本章主要符号表……308
第七章　序贯实验设计……309
第一节　概述……309
第二节　参数估计的序贯实验设计……315
第三节　参数估计序贯实验设计举例……323
第四节　模型筛选的序贯实验设计……331
第五节　模型筛选序贯实验设计举例……341
本章主要符号表……349
主要参考书目和文献……350
附录一　计算机程序……351
一、逐步回归计算程序……351
二、二次回归正交设计的算程序……357
三、阻尼最小二乘法计算程序……360
四、单纯形计算程序……367
附录二　向量和矩阵的概念及运算法则……372
附表Ⅰ　正态分布表……376
附表Ⅱ　正态分布的双侧分位数（u_α）表……378
附表Ⅲ　χ^2分布表……379
附表Ⅳ　t分布表……382
附表Ⅴ　t分布的双侧分位数（t_α）表……384
附表Ⅵ　F检验的临界值（F_α）表……386

第一章　概率和数理统计基础知识

第一节　概　　述

在自然界中存在两类不同的现象。一类是在一定条件下必然会发生的。例如水在标准大气压下，加热到100℃时必然会沸腾。这类在一定条件下必然会发生的事情就称为必然事件。反之，那类在一定条件下必然不会发生的事情就称为不可能事件。尽管必然事件和不可能事件表现的形式相反，但它们的实质却是相同的。所有这类现象我们称之为必然现象。它们具有确定性的规律。与必然现象存在着本质区别的另一类现象是在一定的条件下，或者发生，或者不发生的现象。例如同一个样品，在相同条件下进行分析，可发现所得结果并不一致，而存在微小差异。这种在一定的相同条件下，可能发生，可能不发生的现象称为随机现象。对于随机现象关心的是其结果是否出现，这些结果称为随机事件（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，在偶然性的背后总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、同类的随机现象中揭示其规律——统计性规律。当用数学模型来描述上述两类现象时，可把数学模型分为确定性模型与随机性模型。确定性模型的数学描述中没有随机性的因素（也就是说，变量的值与参数都是确定的数），而且得到的模型解是确定的值。随机性模型允许在数学描述中存在随机因素，因而随机性模型的解并不是一个确定的值，仅仅知道它取某个值的概率，更确切地说是它在某个区间内的概率。在化学工程领域的数据处理中遇到的现象，大多数就其本身来说具有确定的规律，只是由于在测量中引入了随机性的因素，

使结果具有随机性。为此，常采用大量的实验，用统计方法来研究其规律性。例如数学模型参数估值是通过大量的实验测定结果间接求得模型参数值。尽管反映事物客观规律的模型参数（例如反应动力学中的活化能）本身是具有客观性的真值，但由于测定误差的随机性，必然造成参数估计值的随机性，往往这些参数估计值的随机离散程度较大地超过了个别测量值误差的离散程度。这就需要我们从概率角度去认识这些数学模型的结果。为了运用模型进行计算，就要考虑参数的不确定因素，即不仅要估计参数值，而且要估计参数的置信区间。

概率论是从大量的随机实验中研究随机的规律性，但实际上所允许的实验次数永远只能是有限的，有时甚至是少量的。因此，必须有效地利用有限的信息，排除由于信息不足所引起的干扰来研究现象的内在规律性，并作出一定精确程度的判断和预测，将这些研究结果加以归纳整理。在统计学中，我们把所研究的全部元素组成的集合称为总体，而把组成总体的每个元素称为个体。若总体中的元素有限，则称总体的容量是有限的。若总体的元素是无穷的，则称总体的容量是无穷的。为了分析总体，需要从一个总体中抽取一些元素代表这个总体，这些元素的集合称为总体的样本(子样)。样本所包括个体的数目叫做样本容量。数理统计就是根据样本探求有关总体的种种知识，以及利用样本的随机性来检验总体的种种假设。

例如，戊烷异构化反应动力学机理研究中，有单位模型和双位模型二种机理可供选择，要通过实验测定结果判断哪一个模型是正确的，并确定其模型参数。

$$\text{单位模型：}\quad r=\frac{kK_2\,(p_2-p_3/K)}{1+K_1p_1+K_2p_2+K_3p_3}$$

$$\text{双位模型：}\quad r=\frac{kK_2\,(p_2-p_3/K)}{(1+K_1p_1+K_2p_2+K_3p_3)^2}$$

对这类化学工程中常常遇到的研究课题，就可以用概率和数理统

计知识给予新的认识。

第二节　随机事件的概率

在一定实验条件下，现象A可能发生，也可能不发生，我们把发生现象A的事件叫做随机事件A。如果在既定的条件下进行一组实验，总共进行n次，其中事件A发生了n_A次，则该组实验中事件A的频率是比值n_A/n。重复进行很多组这样的实验，我们就会发现随机事件的频率具有某种规律性：频率总是在某个值的上下摆动，并且随着每组实验次数n的增多，频率上下摆动的平均幅度趋于减小，出现显著大或显著小的频率的可能性减少。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，记作$P(A)$。

$$\lim_{n\to\infty}\frac{n_A}{n}=P(A) \qquad (1\text{-}1)$$

随机事件的频率在其概率值的上下波动，并且随实验次数的增多趋近于概率值，是随机现象存在着内在规律性的表现。

若事件A与事件B是两个不同的随机事件，定义事件A和事件B的"和事件"$A+B$是指A与B至少有一个发生的事件。如果用两个圆分别表示事件A和事件B的集合，如图1-1所示，则两个圆的总和（图1-1中阴影部分）就代表事件$A+B$的集合。显然，只有A发生或只有B发生，或者A、B同时发生的事件都是事件$A+B$。

定义事件A与事件B的"积事件"AB为A和B同时发生的事件。图1-2中两个事件重叠的区域就是积事件的区域。

同样，可类似地定义多个事件的和与积。若事件$A_1, A_2, \cdots$，A_n是n个不同事件，则和事件$A_1+A_2+A_3+\cdots+A_n=\sum_i^n A_i$是$A_1$，$A_2$，…，$A_n$中至少有一个发生的事件，积事件$A_1A_2\cdots A_n=$

$\prod_{i}^{n} A_i$是A_1，A_2，…，A_n同时发生的事件。

关于概率事件的概率公式如下：

一、和事件的概率公式

一般情况下，和事件$A+B$的概率不等于事件A与事件B的概率之和$P(A)+P(B)$。因为对两个圆的重叠部分的概率，即A和B同时发生的那些事件的概率已计算了两次，所以，和事件概率公式应当是

$$P(A+B)=P(A)+P(B)-P(AB) \qquad (1-2)$$

如果A和B不可能在一次实验中同时发生，即$P(AB)=0$，则称事件A与事件B是互斥事件。

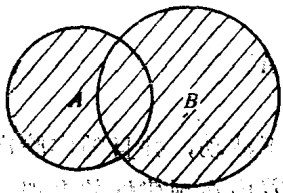


图 1-1 事件的和

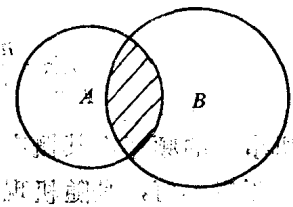


图 1-2 事件的积

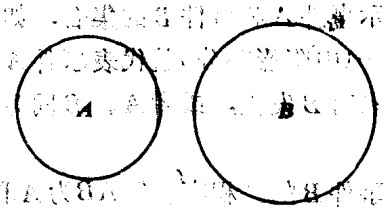


图 1-3 互斥事件

互斥事件A与B的和事件概率公式为

$$P(A+B)=P(A)+P(B) \qquad (1-3)$$

推广到n个两两互斥的随机事件A_1，A_2，…A_n，其和事件的概率为

$$P\left(\sum_{i}^{n} A_i\right)=\sum_{i}^{n} P(A_i) \tag{1-4}$$

二、积事件概率公式

在B发生的条件下A发生的概率，叫做A对于B的条件概率，记作$P(A/B)$或$P_B(A)$。

如果在n次实验中，B发生N_B次，A和B同时发生N_{AB}次，按条件概率的定义有

$$\lim_{n\to\infty}\frac{N_{AB}}{N_B}=P(A/B) \tag{1-5}$$

同时，又有

$$\lim_{n\to\infty}\frac{N_{AB}}{N_B}=\lim_{n\to\infty}\frac{N_{AB}/n}{N_B/n}=\frac{P(AB)}{P(B)}$$

所以

$$P(A/B)=P(AB)/P(B)$$

由式(1-5)可得积事件概率公式

$$P(AB)=P(B)P(A/B) \tag{1-6}$$

同理，可推得

$$P(AB)=P(A)P(B/A) \tag{1-6'}$$

同样，可导得多个事件之积的概率

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2/A_1)\cdot P(A_3/A_1A_2)\cdots P(A_n/A_1A_2\cdots A_{n-1}) \tag{1-7}$$

对于事件A和事件B，如果A事件的概率不受B是否发生的影响，即

$$P(A/B)=P(A) \tag{1-8}$$

则称事件A独立于事件B。

如果事件A独立于事件B，显然事件B也独立于A，即

$$P(B/A)=P(B) \tag{1-9}$$

对于互相独立的事件A和B，积事件概率公式为

$$P(AB)=P(A)P(B) \tag{1-10}$$

反之，如果事件A和事件B的概率满足式(1-8)、(1-9)和(1-10)

中任意一个，则事件A和B就是互相独立的事件。

三、全概率公式

如果n个互斥的事件B_1，B_2，…，B_n中只有一个事件出现时，事件A才可能出现，则事件A的概率

$$P(A)=\sum_{i}^{n}P(B_i)P(A/B_i) \qquad (1\text{-}11)$$

称之为全概率公式。全概率是指出现A的全部概率。

显然有

$$\sum_{i}^{n}P(B_i)=1$$

由于事件B_1，B_2，…，B_n两两互斥，而事件A只能伴随B_1，B_2，…，B_n中的一个同时发生。所以，事件A可以表示成下面互斥的积事件之和

$$A=AB_1+AB_2+\cdots AB_n=\sum_{i}^{n}AB \qquad (1\text{-}12)$$

利用式(1-4)和(1-6)可得

$$P(A)=\sum_{i}^{n}P(AB_i)=\sum_{i}^{n}P(A/B_i)P(B_i) \qquad (1\text{-}13)$$

四、贝叶斯（*Bayes*）公式

若事件B能且只能与两两互不相容事件A_1，A_2，…，A_n中之一同时发生，即

$$B=\sum_{i}^{n}BA_i$$

由于 $$P(A_iB)=P(B)P(A_i/B)=P(A_i)P(B/A_i)$$

故 $$P(A_i/B)=\frac{P(A_i)P(B/A_i)}{P(B)} \qquad (1\text{-}14)$$

又由全概率公式

$$P(B)=\sum_{i}^{n}P(A_i)P(B/A_i)$$

即得

$$P(A_i/B)=\frac{P(A_i)P(B/A_i)}{\sum_{i}^{n}P(A_i)P(B/A_i)} \qquad (1\text{-}14)$$

(1-14) 式称贝叶斯公式。

贝叶斯公式在概率论和数理统计中有广泛应用。假定 A_1, A_2, …, A_n 是导致实验结果的各种可能的"原因"，$P(A_i)$ 称为先验概率。它反映了各种原因发生的可能性大小，一般是以往经验的总结，在实验前已经知道。现在若实验结果产生了事件 B，这个信息将有助于探讨事件发生的"原因"。条件概率 $P(A_i/B)$ 称为后验概率，它反映了实验之后对各种"原因"发生可能性大小的新知识。贝叶斯定理在模型检验中有较多的应用。例如对某一实际过程可能的数学模型有 A、B、C 三种，在实验前三种模型可能性的概率均相等，即都等于1/3。现在进行了某次实验，我们根据贝叶斯定理可以计算各模型在已知概率的条件下，出现特定实验结果的概率，从而推断实验以后第 i 个模型是正确的概率。

第三节　随机变量的分布

只用一个单独的数值显然不能代表一个随机变量，即使列举出随机变量的全部可能值，也仍然不能算是完全描述了一个随机变量。完整地掌握一个随机变量，必须了解它取各种可能值的概率，即须了解随机变量的概率分布。

随机变量 X 的概率分布可以用分布函数 $F(x)$ 来表示。分布函数在 x 处的值等于随机变量 X 取值小于或等于 x 这样一个随机事件的概率

$$F(x)=P(X\leqslant x) \qquad (1\text{-}15)$$

显然，任何一个分布函数都必须满足

$$F(x=-\infty)=0, \qquad F(x=\infty)=1. \tag{1-16}$$

对离散型随机变量X只能取可数的数值x（$x=x_1, x_2\cdots$）。除了分布函数外还可以用概率函数$p(x)$来描述它的概率分布。概率函数在某一点x处的值等于随机变量X取x值的概率，即

$$p(x)=P(X=x) \tag{1-17}$$

由分布函数和概率函数的定义，它们之间有下列关系

$$F(x)=\sum_{x_i<x} p(x_i) \tag{1-18}$$

式中，$\sum_{x_i<x}$表示对所有满足$x_i\leqslant x$的x_i求和，

$$p(x_i)=F(x_i)-F(x_{i-1}) \tag{1-19}$$

并且

$$\sum_{x} p(x)=F(x=\infty)=1 \tag{1-20}$$

式中，$\sum_{x}$表示对所有可以取的x值求和。

离散型随机变量概率函数和分布函数的形状见图1-4。

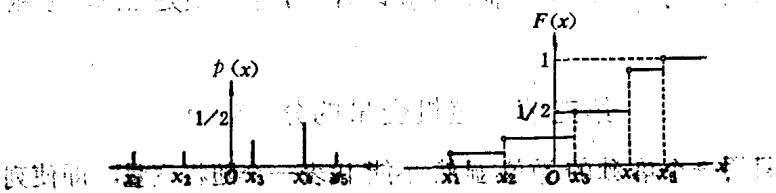


图1-4 离散型随机变量概率函数和分布函数

对于连续型随机变量X，由于随机量取值连续，可以定义概率密度函数

$$p(x)\equiv\frac{dF(x)}{dx} \tag{1-21}$$

下面公式可以说明概率密度函数的意义

$$P(x<X\leqslant x+dx)=dF(x)=p(x)dx \tag{1-22}$$

上面等式应该从极限意义来理解，即

$$p(x)=\lim_{\Delta x\to 0}\frac{P(x<X\leqslant x+\Delta x)}{\Delta x} \tag{1-23}$$

概率密度函数在某一点的值是随机变量在该点的概率密度，随机变量的值落入该点附近一个无限小区间内的概率，等于该点的概率密度和区间长度的乘积（由于概率不可能取负值，所以(1-22)式右侧的dx应该理解为dx的绝对值）。

概率密度函数和分布函数的关系还可以写成

$$F(x)=\int_{-\infty}^{x}p(x)\mathrm{d}x \tag{1-24}$$

且有

$$\int_{-\infty}^{\infty}p(x)\mathrm{d}x=F(x=\infty)=1 \tag{1-25}$$

式(1-20)和(1-25)叫做归一化条件。任何概率函数或概率密度函数都必须满足归一化条件。

概率密度函数曲线是一条连续的曲线。概率密度曲线图在横轴上任一点左侧曲线下的面积就是分布函数在该点的值。由于归一化条件，密度曲线下的总面积为1。分布函数曲线是一条单调上升到1的曲线。分布密度函数曲线和分布函数曲线见图1-5。

概率密度曲线是频率分布直方图在大样本情况下的极限。即

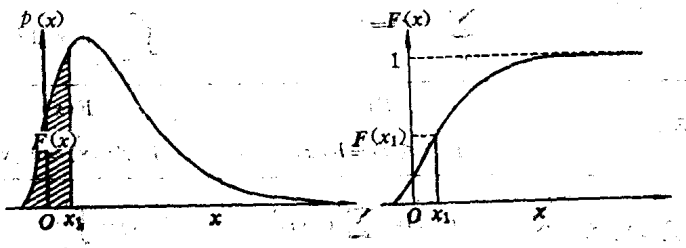


图1-5　连续型随机变量的概率密度函数和分布函数