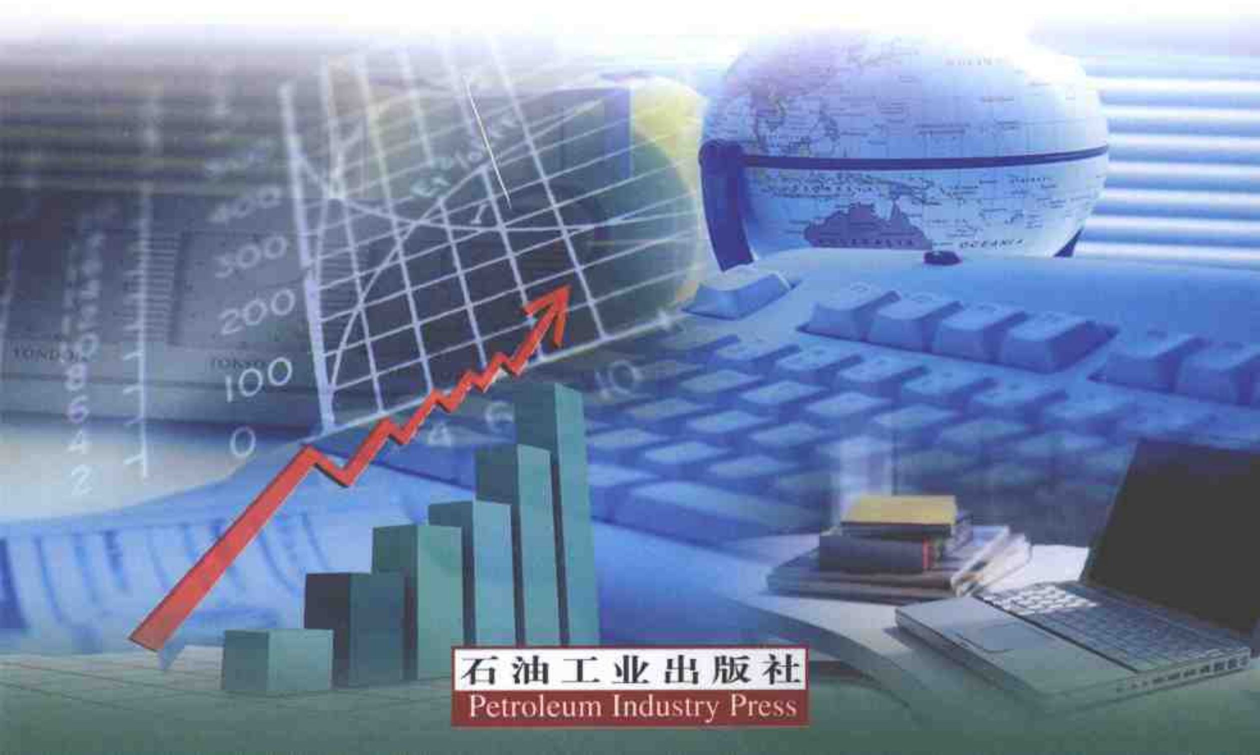
高等学校教材

应用统计方法

常兆光　王清河　杜彩凤　编著



石油工业出版社
Petroleum Industry Press

高等学校教材

应用统计方法

常兆光　王清河　杜彩凤　编著

石油工业出版社

内容提要

本书系统介绍了处理随机数据的常用统计方法及其应用，内容包括预备知识、数理统计的基本概念、参数估计、假设检验、回归分析、方差分析、试验设计、判别分析、聚类分析、主成分分析、因子分析等. 书中收录了许多实例，各章有例题及练习题，书末有附录.

本书可作为高等理工院校非数学专业研究生教材，也可作为具有相当数学准备(初等微积分、概率论初步及少量矩阵知识)的应用统计工作者和工程技术人员的参考书.

图书在版编目(CIP)数据

应用统计方法/常兆光，王清河，杜彩凤编著.
北京：石油工业出版社，2009.11
高等学校教材
ISBN 978-7-5021-7420-0

Ⅰ.应…
Ⅱ.①常… ②王…③杜…
Ⅲ.应用统计学
Ⅳ.C8

中国版本图书馆 CIP 数据核字(2009)第 183163 号

出版发行：石油工业出版社
(北京安定门外安华里 2 区 1 号 100011)
网　址：www.petropub.com.cn
编辑室：(010)64523574　发行部：(010)64523620
经　销：全国新华书店
印　刷：石油工业出版社印刷厂

2009 年 11 月第 1 版　2009 年 11 月第 1 次印刷
787×960 毫米　开本：1/16　印张：20
字数：392 千字

定价：34.00 元
(如出现印装质量问题，我社发行部负责调换)

前　言

“应用统计方法”是研究和揭示随机现象统计规律性的数学学科，自20世纪以来发展较快，已成为当代应用数学重要组成部分. 近代科学发展最显著的特点之一是学科交叉、相互渗透，新兴学科不断涌现. 尤其是计算机的日益普及，进一步推动了该学科的迅猛发展，并已渗透于工业、农业、水文、医学、力学、保险数学、物理学、计量经济学、化学、教育学、体育学、心理学及气象预报、产量预报、地震预报、石油勘探开发、可靠性工程等许多领域.

目前，“应用统计方法”已成为高等学校文、理、工、医、农、经济等学科硕士研究生重要的必修课程. 现在出版的这方面的教材，包括国人自著和译作，为数不算少，也各有其优缺点和使用范围，但大都较偏重理论而轻应用，对应用广泛的各种方法的使用背景和具体实现介绍不够详尽，缺乏既简单易理解而又能对阐明基本概念有帮助的例题. 这在某种程度上给教材的选择带来较大的困难. 因此，编写一部体系基本完整、内容适合实际需要的教材是十分必要的.

本书在体系上参考了大量的国内外优秀教材，是笔者多年教学经验的总结. 本书力求重点突出，具有较强的实用性. 本书在重视基本概念、基本理论、基本运算的同时，加强实用统计方法的介绍，较充分地体现了应用数学课程以实用为主要目的的特点，对实用统计方法的介绍比一般同类教材详细，尤其对目前使用广泛的处理多维数据的统计方法均进行了详略得当、难易适中的介绍；精心选取了既有实用背景又能对阐明基本概念有帮助、提高读者兴趣的例题及习题；介绍实用统计方法时，重视计算机上的实现步骤，结合统计软件(SAS，SPSS，EXCEL)的使用配备上机练习.

本书的特点：一是应用性强，基于实际问题提出统计方法的使用，便于使用者真正掌握统计方法及其应用；二是深入浅出，通俗易懂，全书在不失严谨的前提下尽量避免繁琐的公式推导而侧重应用统计方法的介绍，使学生掌握统计方法应用的背景要求和前提条件；三是在统计计算上强调运用统计软件(尤其是EXCEL的使用)来完成，为统计方法在实际中能广泛应用奠定基础.

本书在编写过程中，得到了中国石油大学(华东)研究生院、数学学院的大力支持和帮助，参考了王才经教授多年讲授应用统计的手稿。王才经教授审阅了全书的初稿，给出了许多指导性意见；常秦应用 EXCEL 或其他统计软件对书中的例题及课后大部分习题进行了较详细的解答，在此表示衷心的感谢.

由于作者水平有限，书中难免仍有错误或不恰当之处，竭诚地希望读者批评指正.

编　者

2009 年 7 月

目　　录

第一章　预 备 知 识

应用统计是一个应用非常广泛的数学分支．它以概率论与数理统计作为理论基础．它的任务是研究如何用有效的方法去搜集、整理和分析带有随机性影响的数据，并对所关心的问题作出推断和预测，直接为决策行动提供依据和建议．凡是有大量数据出现的地方，一般都要用到应用统计的方法．也就是说，应用统计方法是直接从随机现象的观察值去研究它的客观规律．在本章及第二章中，我们把概率论与数理统计初步的有关知识介绍给读者，熟悉这些内容的读者可从第三章看起，想看这两章的读者最好把正文中未给出证明的结果作为练习，这将为以后各章的学习带来很大的方便．

第一节　随机事件与概率

一、随机事件与样本空间

在工农业生产、科学实验和现实生活中，我们遇到过各种各样的随机试验（简称为试验，记作 E）．我们将试验 E 的所有可能出现的结果组成的集合称为 E 的**样本空间**，记作 Ω．Ω 中 E 的每个最基本（最简单）的试验结果称为**样本点**．

样本空间包含了试验 E 的所有可能结果，我们将每一个可能的结果称为随机事件（简称为事件），通常用大写字母 A,B,C 等表示．只包含一个样本点的事件称为**基本事件**；而有两个或两个以上的基本事件（样本点）组成的事件称为**复合事件**．在每次试验中，一定发生的事件称为**必然事件**，也记作 Ω；而一定不发生的事件称为**不可能事件**，记作 $\varnothing$．

需要指出的是：无论是必然事件、随机事件，还是不可能事件，都是相对"一定条件"而言的．条件发生变化，事件的性质也发生变化．例如，抛掷两颗骰子，"出现的点数之和为 5 点"及"出现的点数之和大于 5 点"都是随机事件；若同时抛掷 6 颗骰子，"出现的点数之和为 5 点"则是不可能事件了，而"出现的点数之和大于 5 点"则是必然事件了．为了以后讨论问题方便，通常将必然事件和不可能事件看成是特殊的随机事件．

若记 Ω 为样本空间（必然事件），$\varnothing$ 为不可能事件，e 为基本事件，$A,B,A_1,A_2,\cdots,A_n$ 为随机事件，则有事件之间的关系与运算如下所述．

(1) **包含关系**．若 $A\subset B$，则称事件 B 包含事件 A．它表示事件 A 发生必导致事件 B 发生．对任一事件 A，都有 $\varnothing\subset A\subset\Omega$．

(2)**相等关系**. 若 $A \subset B$ 且 $B \subset A$,则称事件 A 等于事件 B,记作 $A=B$.

(3)**和事件**. 事件 $A \cup B$ 称为事件 A 与事件 B 的和事件,它表示事件 A 与事件 B 至少有一个发生. 和事件的运算可推广至任意有限个事件及无限多个事件.

(4)**积事件**. 事件 $A \cap B$ 称为事件 A 与事件 B 的积事件,简记为 AB,它表示事件 A 与事件 B 同时发生. 积事件的运算可推广至任意有限个事件及无限多个事件.

(5)**差事件**. 事件 $A-B$ 称为事件 A 与事件 B 的差事件,表示事件 A 发生而事件 B 不发生.

(6)**互不相容(互斥)事件**. 若 $AB=\varnothing$,则称事件 A 与事件 B 互不相容,它表示事件 A 与事件 B 不同时发生. 基本事件是互不相容的.

类似地,若 $\forall i \neq j$, $A_iA_j=\varnothing$, $i,j=1,2,\cdots$,则称 $A_1,A_2,\cdots,A_n$ **两两互不相容**.

若 $A_1,A_2,\cdots,A_n$ 两两互不相容,且在每次试验中事件 $A_1,A_2,\cdots,A_n$ 必发生其中之一,即 $\bigcup\limits_{i=1}^{n} A_i=\Omega$,且 $A_iA_j=\varnothing$, $i \neq j$,则称 $A_1,A_2,\cdots,A_n$ 为**互不相容事件完备组**.

(7)**对立事件(逆事件)**. 若 $AB=\varnothing$ 且 $A \cup B=\Omega$,则称事件 A 与事件 B 互逆,又称事件 A 为事件 B 的对立事件(或事件 B 为事件 A 的对立事件),记作 $A=\overline{B}$($\overline{B}$ 表示 B 不发生),或 $B=\overline{A}$.

需要指出的是:由(6),(7)可知,若事件 A 与事件 B 互逆,则事件 A 与事件 B 一定互不相容,但反之则不一定.

(8)**德·摩根定律(对偶原理)**.

$$\overline{A \cup B}=\overline{A} \cap \overline{B},\ \overline{A \cap B}=\overline{A} \cup \overline{B}$$

$$\overline{\bigcup_{i=1}^{n} A_i}=\bigcap_{i=1}^{n} \overline{A_i},\ \overline{\bigcap_{i=1}^{n} A_i}=\bigcup_{i=1}^{n} \overline{A_i}$$

(9)**运算规律**.

设 A,B,C 为三个事件,则有

交换律:$A \cup B=B \cup A$, $A \cap B=B \cap A$.

结合律:$A \cup (B \cup C)=(A \cup B) \cup C$,

$A \cap (B \cap C)=(A \cap B) \cap C$.

分配律:$A \cup (B \cap C)=(A \cup B) \cap (A \cup C)$,

$A \cap (B \cup C)=(A \cap B) \cup (A \cap C)$.

二、频率与概率的几种定义

定义 1-1 设 E 为随机试验,A 为其中任一个事件,$n(A)$为事件 A 在 n 次重

复试验中发生的次数，则称比值$\frac{n(A)}{n}$为n次试验中A发生的频率，记为

$$f_n(A) \triangleq \frac{n(A)}{n} \tag{1-1}$$

式中　$n(A)$——事件A在n次重复试验中发生的频数．

当n增大时，$f_n(A)$逐渐稳定于某一个确定值$P(A)$，称$P(A)$为频率的稳定值．

定义 1-2(概率的统计性定义)　在不变条件下做大量重复试验，称在重复试验中事件A发生的频率的稳定值p为事件A的概率，记为$P(A)$.

定义 1-3(概率的公理化定义)　设E为随机试验，Ω为它的样本空间，对E中的每一个事件A都赋予一个实数，记为$P(A)$，且满足

(1)非负性：$0 \leqslant P(A) \leqslant 1$.

(2)规范性：$P(\Omega)=1$.

(3)可加性：若$A_1, A_2, \cdots, A_n$两两互不相容，即

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{1-2}$$

则称$P(A)$为事件A的概率．

由定义可得概率$P(A)$的性质如下：

(1)$P(\varnothing)=0$.

(2)若$A_1, A_2, \cdots, A_n$两两互不相容，则

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \tag{1-3}$$

式(1-3)称为概率的有限可加性．

(3)若A的对立事件记为$\overline{A}$，则$P(A)=1-P(\overline{A})$.

(4)若$A \subset B$，则$P(B-A)=P(B)-P(A)$，且$P(A) \leqslant P(B)$.

(5)$P(A \cup B)=P(A)+P(B)-P(AB)$.　　(1-4)

性质(5)可推广至任意有限个事件$A_1, A_2, \cdots, A_n$的情形

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \leqslant i < j \leqslant n} P(A_i A_j) + \sum_{1 \leqslant i < j < k \leqslant n}^{n} P(A_i A_j A_k) + \cdots + (-1)^{n-1} P(A_1 A_2 \cdots A_n) \tag{1-5}$$

特别的，对于事件A_1, A_2, A_3，有

$$P(A_1 \cup A_2 \cup A_3)=P(A_1)+P(A_2)+P(A_3)-P(A_1A_2)-P(A_1A_3)-P(A_2A_3)+P(A_1A_2A_3)$$

定义 1-4(概率的古典定义)　设 E 为随机试验,Ω 为它的样本空间,如果 Ω 中的全体基本事件只有有限个,且每个基本事件出现的机会都是均等的,则对 Ω 中的任意事件 A,有

$$P(A)=\frac{A\text{ 中包含基本事件的个数}}{\Omega\text{ 中基本事件的总数}}$$

简记为

$$P(A)=\frac{n(A)}{n(\Omega)} \tag{1-6}$$

并称 $P(A)$ 为事件 A 的**古典概率**.

定义 1-5　设 Ω 是一个几何区域,且其大小可以度量,记为 $\mu(\Omega)$. 向 Ω 中任掷一点,落在该区域任一点处都是等可能的,$A=\{$掷点落在 A 内$\}$为 Ω 中任一事件,且其度量为 $\mu(A)$,则

$$P(A)=\frac{\mu(A)}{\mu(\Omega)} \tag{1-7}$$

并称由此求得的概率为**几何概率**.

三、条件概率

定义 1-6　设 A,B 为样本空间 Ω 中的两个事件,$P(B)>0$,则称

$$P(A\mid B)=\frac{P(AB)}{P(B)} \tag{1-8}$$

为事件 B 发生的条件下事件 A 发生的条件概率.

由条件概率定义可得

$$P(AB)=P(B)P(A\mid B),P(B)>0 \tag{1-9}$$

$$P(AB)=P(A)P(B\mid A),P(A)>0 \tag{1-10}$$

通常称式(1-9)、式(1-10)为概率的乘法公式.

上述公式可推广到任意有限个事件的情形,即若 $A_1,A_2,\cdots,A_n$ 为 n 个事件,且 $P(A_1A_2\cdots A_{i-1})>0,i=2,3,\cdots,n$,则

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2\mid A_1)P(A_3\mid A_1A_2)\cdots P(A_n\mid A_1A_2\cdots A_{n-1}) \tag{1-11}$$

定理 1-1(全概率公式)　设 $A_1,A_2,\cdots,A_n$ 为两两互不相容事件,且 $\bigcup\limits_{i=1}^{n}A_i=\Omega$,

$P(A_i)>0, i=1,2,\cdots,n$,则$\forall A\subset\Omega$,有

$$P(A)=\sum_{i=1}^{n}P(A_i)P(A\mid A_i) \tag{1-12}$$

需要指出的是:在定理1-1中用条件"$A\subset A_1\cup A_2\cup\cdots\cup A_n\subset\Omega$"代替条件"$\bigcup\limits_{i=1}^{n}A_i=\Omega$",定理结论仍成立.

定理1-2(Bayes公式) 设$A_1,A_2,\cdots,A_n$为两两互不相容事件,且$\bigcup\limits_{i=1}^{n}A_i=\Omega$,$P(A_i)>0, i=1,2,\cdots,n$,则$\forall A\subset\Omega, P(A)>0$,有

$$P(A_i\mid A)=\frac{P(A_i)P(A\mid A_i)}{\sum\limits_{j=1}^{n}P(A_j)P(A\mid A_j)}, i=1,2,\cdots,n \tag{1-13}$$

四、事件的独立性

定义1-7 设A,B为两个事件,如果

$$P(AB)=P(A)P(B) \tag{1-14}$$

则称A,B为相互独立事件.

由定义1-7可得,若事件A与事件B相互独立,则

(1)A与$\overline{B}$,$\overline{A}$与B,$\overline{A}$与$\overline{B}$也相互独立;

(2)$P(A)=P(A|B), P(B)=P(B|A)$.

定义1-8 设$A_1,A_2,\cdots,A_n$为n个随机事件,若对任意的k,$1<k\leqslant n$及任意的$1\leqslant i_1<i_2<\cdots<i_k\leqslant n$,有

$$P(A_{i_1}A_{i_2}\cdots A_{i_k})=P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k})$$

则称$A_1,A_2,\cdots,A_n$相互独立.

定义1-9 设$A_1,A_2,\cdots,A_n$为n个随机事件,若其中任意两个事件A_i,A_j,$i\neq j$相互独立,则称$A_1,A_2,\cdots,A_n$两两独立.

由定义1-8、定义1-9可知,n个事件相互独立一定两两独立,反之不真.

第二节 随机变量及其分布

为了描述随机现象,我们已引入了样本空间等概念.我们知道,样本空间里的样本点可以是数量性质的,也可以是非数量性质的.例如:掷骰子试验中,观察朝上一面的点数,样本空间$\Omega=\{1,2,3,4,5,6\}$;掷硬币试验中,观察朝上一面出现的

情况，样本空间 $\Omega=\{H,T\}$. 前者样本空间是数量化的，后者样本空间与数量无关. 我们也知道，以前所学过的数学大都是从数量上来反映客观世界的. 为了利用已有的数学工具来研究随机现象，我们应设法对样本空间进行数量化，即需要把样本空间里的样本点与数联系起来，建立样本空间 Ω 与实数(或复数)域或其一部分的对应关系. 为此引入随机变量的概念.

一、随机变量的分布函数

定义 1-10 设 E 是随机试验，它的样本空间为 $\Omega=\{e\}$，如果对于每一个 $e\in\Omega$，都有唯一的实数 $X(e)$ 和它对应，且对于任何实数 x，$\{X(e)\leqslant x\}$ 具有确定的概率，则称 $X(e)$（简记为 X）为**随机变量**.

由定义 1-10 可知，若 X 为一随机变量，对于任何实数 x，$\{X(e)\leqslant x\}$ 具有确定的概率. 为了刻画随机变量取值的概率规律，我们引入分布函数概念.

定义 1-11 设 X 为一随机变量，$\forall x\in\mathbf{R}$，令

$$F(x)=P\{X\in(-\infty,x]\}=P\{X\leqslant x\},-\infty<x<+\infty \tag{1-15}$$

则称 $F(x)$ 为随机变量 X 的分布函数.

由定义 1-11 可知，当 $a<b$ 时，

$$P\{a<X\leqslant b\}=F(b)-F(a) \tag{1-16}$$

分布函数 $F(x)$ 具有如下性质：

(1) $0\leqslant F(x)\leqslant 1$.

(2) $F(x)$ 是一个非降函数.

(3) $F(x)$ 右连续，即 $\lim\limits_{x\to a^+}F(x)=F(a)$.

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

对于一个已知函数 $F(x)$，如何知道它是否可以作为某个随机变量的分布函数呢？这只要验证它是否满足性质(2)，(3)，(4)就可以了.

二、离散型随机变量及其概率分布(分布律)

定义 1-12 设 X 为离散型随机变量，其可能取值为 x_k，$k=1,2,\cdots$，且

$$P\{X=x_k\}\triangleq p_k,k=1,2,\cdots \tag{1-17}$$

则称 p_k，$k=1,2,\cdots$ 为 X 的概率分布(或分布律)，简记为 $\{p_k\}_1^\infty$.

显然，概率分布 $\{p_k\}_1^\infty$ 满足：

(1) $p_k\geqslant 0$, $k=1,2,\cdots$（非负性）.

(2) $\sum\limits_{k=1}^{\infty}p_k=1$（归一性）.

概率分布$\{p_k\}_1^{\infty}$也可用表格形式表示，即

X	x_1	x_2	$\cdots$	x_n	$\cdots$
p_k	p_1	p_2	$\cdots$	p_n	$\cdots$

此表称为 X 的**分布列(或概率分布表)**.

由概率分布可得 X 的分布函数 $F(x)$ 为

$$F(x)=P\{X\leqslant x\}=\sum_{x_k\leqslant x}p_k \tag{1-18}$$

且可能取值 x_k 是 $F(x)$ 的跳跃间断点，其跃度为

$$F(x_k)-F(x_k-0)=p_k,k=1,2\cdots$$

三、连续型随机变量的概率分布

在非离散型随机变量中，仅介绍在应用中经常遇到的连续型随机变量．连续型随机变量取值的特点是充满一个区间或若干个区间．

定义 1-13 若随机变量 X 的分布函数 $F(x)$ 可表示成

$$F(x)=\int_{-\infty}^{x}f(u)\mathrm{d}u \tag{1-19}$$

其中 $f(x)$ 为一非负可积函数，则称 X 为**连续型随机变量**，$f(x)$ 称为 X 的**概率密度(或概率分布、分布密度)**.

由式(1-19)可知，连续型随机变量 X 的分布函数 $F(x)$ 是连续函数．

由定义 1-13 可得，概率密度 $f(x)$ 具有如下性质：

(1) $f(x)\geqslant 0,-\infty<x<+\infty$.

(2) $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$.

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

(4)若 $f(x)$ 在点 x 处连续，则 $F'(x)=f(x)$.

(5)连续型随机变量 X 取值为任一实数 a 的概率均等于零，即

$$P\{X=a\}=0$$

连续型随机变量 X 取值为任一实数 a 的概率均等于零，这一性质与离散型随机变量的情形是截然不同的．由此可见，概率为零的事件并不一定是不可能事件．也就是说，若 A 是不可能事件，则 $P(A)=0$，反之则不一定成立．

由性质(3)、性质(5)可知，在计算连续型随机变量 X 落在某一区间上的概率时，可以不必区分该区间是开区间、闭区间还是半开半闭区间，即对于任意的实数

a,b,有

$$P\{a \leqslant X < b\} = P\{a < X < b\} = P\{a < X \leqslant b\} = P\{a \leqslant X \leqslant b\}$$
$$= \int_a^b f(x)\mathrm{d}x = F(b) - F(a)$$

四、几种常见随机变量的概率分布

1. (0—1)分布

若 X 的分布列为

X	0	1
p_k	$1-p$	p

即

$$P\{X=k\} = p^k(1-p)^{1-k} \quad ,k=0,1;0<p<1$$

则称 X 服从参数为 p 的(0—1)分布,记作 $X \sim B(1,p)$.

2. 二项分布

如果随机变量 X 的所有可能取值为 $0,1,2,\cdots,n$,其分布律为

$$P_n(k) = P\{X=k\} = C_n^k p^k (1-p)^{n-k}, k=0,1,2,\cdots,n \qquad (1-20)$$

称式(1-20)为二项概率公式,并称随机变量 X 服从参数为 n,p 的二项分布,记为 $X \sim B(n,p)$. 当 $n=1$ 时,即为(0—1)分布.

3. 泊松分布

如果随机变量 X 的所有可能取值为 $0,1,2,\cdots$,其分布律为

$$P\{X=k\} = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!}, k=0,1,2,\cdots;\lambda>0 \qquad (1-21)$$

则称 X 服从参数为 λ 的泊松(Poisson)分布,记为 $X \sim P(\lambda)$.

4. 超几何分布

若随机变量 X 的所有可能取值为 $0,1,2,\cdots,L$, $L=\min\{M,n\}$,其概率分布为

$$P\{X=k\} = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, k=0,1,2,\cdots,L \qquad (1-22)$$

其中整数 $N>0,M>0$,且 $n \leqslant N-M$,则称 X 服从参数为 n,M,N 的超几何分布,简记为 $X \sim H(n,M,N)$.

超几何分布主要用于产品的不放回抽样,如用它确定次品数或正品数的概率分布.

事实上，当 n 固定，$\frac{M}{N}=p$ 固定，$N\to\infty$时，有

$$\lim_{N\to\infty}\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}=C_n^k p^k q^{n-k} \tag{1-23}$$

式(1-23)表明了超几何分布与二项分布之间的联系，即超几何分布的极限分布为二项分布．

5. 均匀分布

设随机变量 X 在有限区间$[a,b]$内取值，且其分布密度为

$$f(x)=\begin{cases}\frac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & 其他\end{cases}$$

则称 X 在有限区间$[a,b]$上服从均匀分布，记为 $X\sim U(a,b)$.

6. 正态分布

若随机变量 X 的分布密度为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2},-\infty<x<+\infty \tag{1-24}$$

其中 $\mu,\sigma(\sigma>0)$为常数，则称 X 服从参数为 μ,σ 的正态分布或高斯(Gauss)分布，记为 $X\sim N(\mu,\sigma^2)$.

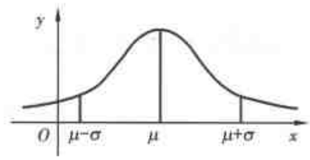


图 1-1

$f(x)$的图形如图 1-1 所示．由图 1-1 可以看出，X 落在 μ 附近的概率较大，且满足：

(1) $f(x)\geqslant 0$　　(非负性).

(2) $\int_{-\infty}^{+\infty}f(x)\mathrm{d}x=1$　　(归一性).

$f(x)$及其图形有如下性质：

(1) $f(x)$的图形关于 $x=\mu$ 对称．

(2) 当 $x=\mu$ 时，$f(\mu)=\frac{1}{\sqrt{2\pi}\sigma}$为最大值．

(3) $f(x)$以 x 轴为渐近线，$x=\mu\pm\sigma$ 处存在拐点，其分布函数为

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(t-\mu)^2}\mathrm{d}t,-\infty<x<+\infty$$

特别的，当 $\mu=0,\sigma=1$ 时，图形关于 $x=0$ 对称，则密度函数为

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},-\infty<x<+\infty \tag{1-25}$$

其分布函数为

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}dt,-\infty<x<+\infty$$

此时称 X 服从**标准正态分布**，记为 $X\sim N(0,1)$.

$\varphi(x)$的图形如图 1-2 所示，由对称性得

$$\Phi(-x)=1-\Phi(x)$$

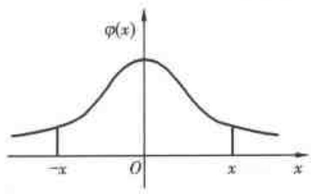


图 1-2

为了解决正态分布的计算问题，人们编制了标准正态分布函数 $\Phi(x)$的部分函数值表以供查阅(附表 1). 对于一般正态分布函数 $F(x)$的计算可以通过下述式(1-26)化成标准正态分布函数 $\Phi(x)$来计算．正态分布的分布函数与标准正态分布的分布函数有如下关系：

$$F(x)=\Phi(\frac{x-\mu}{\sigma}) \tag{1-26}$$

由式(1-26)可知，若 $X\sim N(\mu,\sigma^2)$，得

$$P\{|X-\mu|<3\sigma\}\approx 0.9974$$

即当 $X\sim N(\mu,\sigma^2)$时，“$|X-\mu|\geqslant 3\sigma$”是一小概率事件．如果 x 是 X 的一个观察值，理所当然地认为 x 应满足$|X-\mu|<3\sigma$，否则我们认为它是一个异常数据，通常称为**离群值**，这时可将其剔除．这就是人们进行数据处理时所常用的方法，称为 3σ 准则．但应该指出的是，在有些情况下，我们不能把离群值简单地剔除，应对其进行详细研究，因为它很可能蕴含着某种重要信息．

7. 指数分布

若随机变量 X 具有概率密度

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中 $\lambda>0$ 为常数，则称 X 服从参数为 λ 的指数分布，记为 $X\sim e(\lambda)$.

8. 柯西(Cauchy)分布

若随机变量 X 具有概率密度