化工数据建模与试验优化设计

郑明东　刘炼杰
余　亮　姚伯元　编著

中国科学技术大学出版社

内 容 提 要

本书对概率与数理统计基础、试验设计、试验数据分析以及在 Windows 下直接处理试验数据等内容进行了整合。全书分为三篇，共十四章。分别介绍了数理统计基础，误差理论，参数估计与假设检验，试验数据的表示方法，插值计算，数据平滑，曲线拟合，回归分析与模型预报，以及正交试验设计，调优试验设计和配比试验设计等。每章附有基于 Windows 操作的计算程序和直接使用 Excel 软件的实例。

本书可作为高等工科院校化工、冶金、材料、能源、环境等专业的本科生、研究生教材，也可供科研和生产企业的技术人员参考，尤其适合各类论文阶段的学生使用。

图书在版编目(CIP)数据

化工数据建模与试验优化设计/郑明东，刘练杰，余　亮，姚伯元编著. —合肥：中国科学技术大学出版社，2001. 6
ISBN7-312-01278-7

Ⅰ. 化…　Ⅱ. ①郑…　②刘…　③余…　④姚…　Ⅲ. 化学工业-数学模型-高等学校-教材　Ⅳ. TQ018

中国版本图书馆 CIP 数据核字(2001)第 028172 号

中国科学技术大学出版社出版发行
(安徽省合肥市金寨路 96 号，邮编：230026)
中国科学技术大学印刷厂印刷
全国新华书店经销

开本：787×1092/16　印张：11.5　字数：290 千
2001 年 6 月第 1 版　2001 年 6 月第 1 次印刷
印数：1—4000 册

ISBN 7-312-01278-7/O. 242
定价：16.00 元

前　言

科学研究总离不开科学地进行试验设计和准确快速地进行数据处理，对生产优化和降低成本的作用不可估量。随着计算机的应用，又有新的内容。

该书根据知识、能力、素质培养要求，为体现教学改革思想，整合了《概率论与数理统计》、《计算机在化工中的应用》及《化工试验设计》三门课程的教学内容，形成以数理统计为数学基础，强调近代试验设计方法和试验数据处理技术应用的新体系。并提供了各种镶嵌在 Excel 基础上的计算机算法，可以直接调用，解决了早期软件算法过时和近代流行软件需要编程等问题。

本书原型是作者 1994 年编写的《化工数据模型和试验优化设计》，作为安徽工业大学本科生和部分专业的研究生教材，经过五届的教学实践，尤其是通过国家级教改课题试点班的使用，为该书的修改和完善，提供了丰富的基础素材。

本书第一、二、三章由刘炼杰副教授编写，第四章由姚伯元教授编写，其余各章由郑明东编写，所有计算机程序的算法由余亮同志编写，全书由郑明东统编整理。书中实例大部分是作者多年的科研积累，化工专业的研究生们对例题进行过演算。个别例题为说明问题引用了文献原题。本书在编写过程中还得到学校许多专家、教授的指点和帮助，在此，向他们表示衷心的感谢。

在编写过程中力求区别于纯数学类教材，强调理论联系实际和应用，在不失严谨的前提下，简化数学推导，注重数学基本概念和计算方法，突出对各种方法的理解、掌握和灵活应用。但由于作者的水平有限，书中可能仍有不妥之处，敬请读者批评指正。

编著者

2000 年 10 月

目　次

第一篇　概率与数理统计基础
第一章　概率论基础 …… (1)
1.1　随机事件与概率 …… (1)
1.2　随机变量 …… (3)
1.3　多维随机变量 …… (7)
1.4　随机变量的数字特征 …… (8)
第二章　数理统计基础 …… (12)
2.1　抽样与抽样分布 …… (12)
2.2　参数估计 …… (15)
2.3　假设检验 …… (19)
第三章　误差理论 …… (24)
3.1　误差的基本概念 …… (24)
3.2　误差的表示方法 …… (26)
3.3　有效数字 …… (27)
3.4　间接测量误差的计算 …… (28)
第二篇　实验数据建模
第四章　实验数据整理 …… (32)
4.1　实验结果的表示方法 …… (32)
4.2　计算机绘图 …… (39)
4.3　异常数据的取舍 …… (43)
第五章　数学模型概述 …… (51)
5.1　数学模型的建立 …… (51)
5.2　数学模型的类型 …… (54)
5.3　曲线模型的选择 …… (55)
第六章　插值计算与插值多项式模型 …… (56)
6.1　概述 …… (56)
6.2　线性插值 …… (56)
6.3　拉格朗日插值 …… (57)
6.4　牛顿插值 …… (58)
6.5　样条插值 …… (62)
第七章　实验数据的曲线拟合 …… (63)
7.1　概述 …… (63)
7.2　最小二乘法的数学原理 …… (63)
7.3　最小二乘法的数学描述 …… (64)

7.4 多项式曲线拟合 …… (66)
7.5 最小二乘法的计算机算法 …… (68)
7.6 最小二乘法拟合误差 …… (69)
7.7 非线性化模型参数的求解 …… (70)
7.8 分段曲线拟合 …… (72)
第八章 数据点平滑技术 …… (74)
8.1 图解平滑法 …… (74)
8.2 计算平滑法 …… (74)
第九章 线性代数模型的回归分析 …… (77)
9.1 概述 …… (77)
9.2 一元线性回归分析 …… (78)
9.3 二元线性回归分析 …… (80)
9.4 多元线性回归分析 …… (82)
9.5 非线性相关分析 …… (85)
9.6 回归分析的预报与控制 …… (87)
第三篇 试验优化设计
第十章 试验设计基础 …… (93)
10.1 试验设计的含义 …… (93)
10.2 试验设计的基本概念 …… (94)
10.3 试验设计的基本原则 …… (94)
10.4 试验设计方法 …… (95)
10.5 简单试验设计 …… (96)
第十一章 正交试验优化设计 …… (98)
11.1 概述 …… (98)
11.2 正交试验优化设计的直观分析 …… (100)
第十二章 正交试验的方差检验 …… (110)
12.1 二水平正交试验的方差检验 …… (110)
12.2 三水平正交试验的方差检验 …… (113)
12.3 四水平正交试验的方差检验 …… (114)
12.4 混合型正交表的方差分析 …… (114)
12.5 正交试验中的效应及指标值预估 …… (116)
第十三章 调优运算技术 …… (118)
13.1 概述 …… (118)
13.2 调优运算过程 …… (118)
13.3 二因素调优运算 …… (119)
13.4 三因素调优运算 …… (124)
13.5 多因素调优运算 …… (130)
第十四章 混料配比试验设计 …… (132)
14.1 概述 …… (132)

14.2 单纯形格子设计 …… (132)
14.3 单纯形重心设计 …… (136)
14.4 存在下界约束条件的混料设计 …… (138)
主要参考文献 …… (142)
附录 …… (143)
附录Ⅰ 相关系数 R 表 …… (143)
附录Ⅱ t 分布的双测分位数(t_α)表 …… (145)
附录Ⅲ 随机数表 …… (146)
附录Ⅳ $F(f_1,f_2)$表 …… (147)
附录Ⅴ 常用正交表 …… (149)
附录Ⅵ Excel 及其应用 …… (155)

第一篇　概率与数理统计基础

第一章　概率论基础

概率论是研究随机现象统计规律性的数学分支。它的应用几乎遍及所有科学技术领域和工农业生产之中。本章将介绍概率论的基本概念及一些有关的基本内容。

1.1　随机事件与概率

1.1.1　随机事件

在实践中，有时需要进行这样的试验：它可以在相同的条件下重复进行，每次试验可能的结果有很多个；但试验之前无法确定那一个结果会出现。这样的试验叫做随机试验。如“掷一枚骰子，观察其点数”；“在一大批灯泡中任取一只，测试其寿命”等都是随机试验。

在随机试验中，可能发生、也可能不发生的事件叫做随机事件。常用大写字母 A,B,C 等表示。而把每次试验中必然会发生的事件叫必然事件，必然不会发生的事件叫不可能事件，分别用 S 和 ϕ 来表示。必然事件和不可能事件可看成随机事件的两种特殊情形。如在掷骰子的试验中，“出现偶数点”“出现点数大于 3”都是随机事件，而“出现点数不小于 1”是必然事件，“出现点数大于 6”是不可能事件。

随机试验每一个可能的结果叫做一个基本事件。可以发现，随机事件实际上是某些基本事件的集合。

1.1.2　事件的关系和运算

1. 若事件 A 发生必有事件 B 发生，则称事件 A 是事件 B 的子事件，记 $A\subset B$。若 $A\subset B$ 且 $B\subset A$，称事件 A 与 B 相等，记 $A=B$。

2. 事件 A 与 B 至少有一个发生所构成的事件，称为 A 与 B 的和事件，记 $A\cup B$。

3. 事件 A 与 B 都发生所构成的事件，称为 A 与 B 的积事件，记 $A\cap B$ 或 AB。

4. 事件 A 发生而事件 B 不发生所构成的事件，称为 A 与 B 的差事件，记 $A-B$。

5. 若事件 A 与 B 不可能同时发生，即 $AB=\phi$，则称事件 A 与 B 互不相容（或互斥）。

6. “事件 A 不发生”这一事件，称为事件 A 的对立事件，记为 $\overline{A}$。

1.1.3　概率

1. 古典定义

如果随机试验所有可能的结果是有限个，而每一种试验结果出现的可能性相等，那么称此类问题为古典概型。若可能的结果为 N 种，而事件 A 包含其中的 M 种，则

$$P(A)=\frac{M}{N}$$

如掷骰子出现的点数共有 6 种，而“出现偶数点”包含了{2,4,6}三种，故 $P=\frac{3}{6}=\frac{1}{2}$。

古典概型是概率论早期研究的主要对象，因而将其概率计算公式叫做概率的古典定义。

2. 统计定义

如随机事件 A 在 n 次试验中发生 m 次，称 $f(A)=\frac{m}{n}$ 为事件 A 在 n 次试验中出现的频率。在进行大量重复试验中，事件 A 的频率逐渐稳定在某一常数 P 附近，这个常数 P 就称为事件 A 的概率。

统计定义反映了概率的客观性，但从大量重复试验中找频率的稳定值在实际上难以操作。近代概率论又有概率的公理化定义，因涉及数学知识较多在此不作介绍。

3. 基本性质和公式

(1) $0\leqslant P(A)\leqslant 1$；

(2) $P(S)=1$，$P(\phi)=0$，S 为必然事件，ϕ 为不可能事件；

(3) 若事件 A 与 B 互不相容，则 $P(A\cup B)=P(A)+P(B)$，还可将性质(3)推广：若事件 $A_1,A_2,\cdots,A_n$ 两两互不相容，则 $P(A_1\cup A_2\cup\cdots\cdots\cup A_n)=P(A_1)+P(A_2)+\cdots+P(A_n)$，若事件 $A_1,A_2,\cdots,A_n,\cdots$(可列无穷多个)两两互不相容，则 $P(A_1\cup A_2\cup\cdots\cup A_n\cdots)=P(A_1)+P(A_2)+\cdots+P(A_n)+\cdots$，以上是概率的三条基本性质。由此还可推出

(4) $P(\overline{A})=1-P(A)$

(5) $P(A\cup B)=P(A)+P(B)-P(AB)$　　(加法公式)

1.1.4 条件概率

1. 定义

若 $P(A)>0$，在事件 A 发生了的条件下求事件 B 的概率，称为事件 B 对事件 A 的条件概率，记为 $P(B|A)$。其计算公式为

$$P(B|A)=\frac{P(AB)}{P(A)}$$

2. 乘法公式

当 $P(A)>0$ 时，$P(AB)=P(A)P(B|A)$

当 $P(B)>0$ 时，$P(AB)=P(B)P(A|B)$

例 1.1　一批零件共 100 个，其中有 10 个次品。连续两次从中任取一个零件，作不放回取样。求在第一次取得正品的情况下第二次又取得正品的概率和两次抽取都取得正品的概率。

解：设“第一次取得的零件是正品”为事件 A，“第二次取得的零件是正品”为事件 B。

则
$$P(A)=\frac{90}{100}\qquad P(B|A)=\frac{89}{99}$$

$$P(AB)=P(A)P(B|A)=\frac{90}{100}\cdot\frac{89}{99}=\frac{89}{110}$$

1.1.5 事件的独立性

1. 定义

若 $P(AB)=P(A)P(B)$ 成立，则事件 A 与事件 B 称为相互独立。

当 $P(A)$ 与 $P(B)$ 都不为零时，从独立性定义立即可推出：若事件 A 与事件 B 独立，必有 $P(A|B)=P(A)$，$P(B|A)=P(B)$，反之亦然。

注意“A 与 B 独立”是说事件 A 与 B 的发生互不影响，与“互不相容”是完全不同的概念。“互不相容”是说两事件不能同时发生。

2. 贝努里试验

在一定的条件下进行独立的重复试验，每次试验只有两种可能的结果 A 与 $\overline{A}$，记 $P(A)=p, P(\overline{A})=1-p=q$，这样的实验称为贝努里试验。

贝努里试验在概率论中具有重要地位，从贝努里试验可以得出许多重要的概率分布模型。

1.2 随 机 变 量

1.2.1 随机变量的概念

随机事件千差万别，为了研究的方便，有必要将其数量化，即引进随机变量。所谓随机变量，是指依随机试验的不同结果而取不同数值的变量，常用大写字母 X, Y, Z 等表示。

我们看到有些随机试验结果本来就与数字相联系，如掷骰子出现的点数，测试灯泡寿命所得的小时数等，可将这些数字作为随机变量取值。有的随机试验结果本身不与数字联系，如掷硬币得到“正面”或“反面”，产品检验抽得了“正品”或“次品”等，我们可以给每一结果对应一个数字，如将“出现正面”与“抽得正品”对应数字 1，“出现反面”与“抽得次品”对应数字 0。引进了随机变量之后，一切随机事件都可以用随机变量取某些值来表示。

随机变量主要有离散型随机变量和连续型随机变量两类。离散型随机变量所有可能取的值为有限个或可列无穷多个，而连续型随机变量可以取某一区间内的任一个值。

1.2.2 分布函数

为了对各种随机变量取值的概率分布给出一个统一的表达方法，需要引进随机变量分布函数的概念。

1. 定义

设 X 为随机变量，对任一实数 x，令

$$F(x)=P\{X\leqslant x\}$$

称 $F(x)$ 是 X 的分布函数。

注意到 $F(x)$ 是通常意义下的函数，定义域为 $(-\infty,+\infty)$，值域为区间 $[0,1]$。可以将它与列车时刻表上标明的“列车沿途各站距起点的公里数”相类比，要知道任意两站之间的距离，可以用它们距起点的公里数相减。

3. 性质

(1) $P\{a<X\leqslant b\}=F(b)-F(a)$；

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

(3) $F(x)$ 是单调非减函数，即对任意的 $x_1<x_2$，有 $F(x_1)\leqslant F(x_2)$；

(4) $F(x)$ 是右连续函数。

1.2.3 离散型随机变量及其分布律

1. 分布律

离散型随机变量可将其取各个值的概率用公式给出

$$P\{X=x_K\}=p_K \quad K=1,2,\cdots$$

也可列成表格

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

这里的 p_K 应满足 (1) $p_K\geqslant 0$ (2) $\sum\limits_K p_K=1$

将这种表示概率分布的方式称为 X 的分布律。

2. 常用的离散型概率分布

(1) (0～1)分布

X 只取 0 与 1 两个值,其分布律为

$P(X=1)=p$, $P\{X=0\}=1-p$

(0～1)分布的背景是只有两个结果的随机试验。如“掷硬币”、“记录婴儿性别”、“检验产品是否合格”等等。

(2) 二项分布

X 的分布律为

$$P\{X=K\}=C_n^K p^K(1-p)^{n-K} \quad K=0,1,2,\cdots,n$$

称 X 服从参数为 n 和 p 的二项分布,记为 $b(n,p)$。

二项分布来源于前面介绍过的贝努里试验。若事件 A 发生的概率为 p,在 n 次试验中事件 A 发生 K 次的概率服从二项分布。如产品的次品率一定,任取 n 件中恰有 K 件次品的概率;射击的命中率一定,n 次射击命中 K 次的概率等等。

(3) 泊松分布

X 的分布律为

$$P\{X=k\}=\frac{\lambda^K}{K!}e^{-\lambda} \quad K=0,1,2,\cdots, \quad \lambda>0$$

称 X 服从参数为 λ 的泊松分布,记为 $\pi(\lambda)$。

泊松分布是当二项分布中 n 很大、p 很小、$\lambda=np$ 不很大时的近似。如一页书中的印刷错误个数、在一段时间某种放射性物质发出的到达计数器上的 α 粒子数等都可认为服从泊松分布。

1.2.4 连续型随机变量及其概率密度

1. 概率密度

若对于随机变量 X 的分布函数 $F(x)$,存在非负函数 $f(x)$,使

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t$$

则称 X 为连续型随机变量,并称 $f(x)$ 为 X 的概率密度。

概率密度的性质:

(1) $f(x)\geqslant 0$, $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=1$;

(2) $P\{a \leqslant X \leqslant b\} = \int_a^b f(x)\mathrm{d}x$；此等式对开区间、半开区间都成立。

(3) 若 x 是 $f(x)$的连续点，则有 $F'(x)=f(x)$。

由于当 Δx 较小时，随机变量 X 在$[x,x+\Delta x]$内取值的概率

$$P\{x\leqslant X\leqslant x+\Delta x\}=\int_x^{x+\Delta x} f(t)\mathrm{d}t\approx f(x)\Delta x$$

这表明概率密度的数值反映了随机变量在点 x 附近取值的概率大小。注意到连续型随机变量取个别值的概率 $P\{X=x\}=0$，只能讨论它在某一区间上取值的概率。因而必须牢记：连续型随机变量在任一区间上取值的概率，就等于概率密度在该区间上积分。

2. 常用的连续型概率分布

(1) 均匀分布

X 的概率密度为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b\\ 0, & \text{其它}\end{cases}$$

称 X 在区间(a,b)上服从均匀分布。在数值计算中，四舍五入产生的误差即在$(-0.5,0.5)$上服从均匀分布。

(2) 指数分布

X 的概率密度为 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geqslant 0\\ 0, & x<0\end{cases}$ 其中 $\lambda>0$

称 X 服从指数分布。常用来作为各种“寿命”分布的近似，如电气元件的寿命，人或动物的寿命等等。

(3) 正态分布

X 的概率密度为 $f(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 其中 $\sigma>0$

称 X 服从正态分布，记为 $N(\mu,\sigma^2)$。

正态分布是概率论中最重要、最常用的一种分布。例如测量的误差；炮弹落点的分布；人的身高、体重；化工产品中某种成分的含量；农作物的收获量等等都近似服从正态分布。一般地，如果影响某一数量指标的随机因素很多，而每个因素起的作用都不大，则这个指标可认为服从正态分布，这可用概率论的中心极限定理来证明。

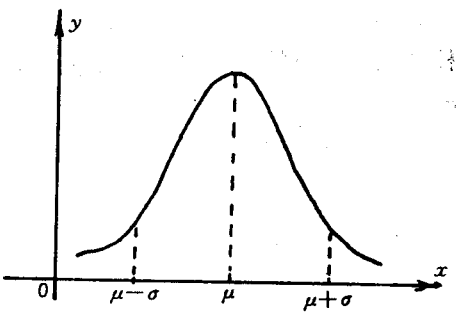

图 1.1

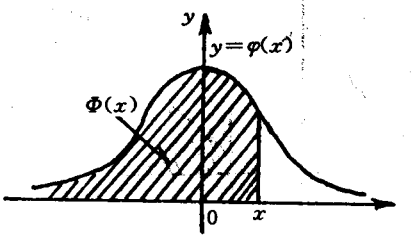

图 1.2

正态分布的概率密度函数 $f(x)$图形如图 1.1。整个图形关于 $x=\mu$ 对称，具有“两头小，中间大”的特征，在 $x=\mu$ 处达到极大。σ 不同时，$f(x)$的形状也不同。σ 越小，图形高而陡峭，分布越集中在 $x=\mu$ 附近。σ 越大，图形低而平坦，分布越分散。

当 $\mu=0,\sigma=1$ 时的正态分布称为标准正态分布，其概率密度与分布函数分别记为 $\varphi(x)$ 和 $\Phi(x)$。即有

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

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

图 1.2 中画出的曲线即 $\varphi(x)$的图像。显然 $\varphi(x)$是偶函数，有 $\varphi(-x)=\varphi(x)$。图中用阴影部分面积表示 $\Phi(x)$，并有

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

为了便于正态分布概率的计算，通常在许多书中都附有 $\Phi(x)$的函数表。对于一般的正态分布 $N(\mu,\sigma^2)$，可通过变换 $Z=\frac{X-\mu}{\sigma}$变为标准正态分布 $N(0,1)$。

由正态分布表容易算出，若 X 服从正态分布 $N(\mu,\sigma^2)$，则

$$P\{|X-\mu|<\sigma\}=68.27\%$$
$$P\{|X-\mu|<2\sigma\}=95.45\%$$
$$P\{|X-\mu|<3\sigma\}=99.73\%$$

即 X 取值落在区间$(\mu-3\sigma,\mu+3\sigma)$之外的概率仅有 0.0027，几乎是不可能的。这称为正态分布的“3σ 规则”。

例 1.2 已知 $X\sim N(1.4,0.0025)$，求 $P\{1.35\leqslant X\leqslant 1.45\}$

解：∵ $\mu=1.4$， $\sigma^2=0.0025$， $\sigma=0.05$

$$Z=\frac{X-1.4}{0.05}\sim N(0,1)$$

$$\therefore\quad P\{1.35\leqslant X\leqslant 1.45\}=P\left\{\frac{1.35-1.4}{0.05}\leqslant\frac{X-1.4}{0.05}\leqslant\frac{1.45-1.4}{0.05}\right\}$$
$$=P\{-1\leqslant Z\leqslant 1\}=\Phi(1)-\Phi(-1)$$
$$=\Phi(1)-[1-\Phi(1)]=2\Phi(1)-1$$
$$=2\times 0.84133-1=0.6827$$

为了应用的需要，对于服从标准正态分布的随机变量 X，称满足条件

$$P\{X>u_\alpha\}=\alpha,\qquad 0<\alpha<1,$$

的点 u_α 为标准正态分布的上 α 分位点(如图 1.3)。如由查表可得

$$u_{0.05}=1.645,$$
$$u_{0.025}=1.96,$$
$$u_{0.005}=2.57。$$

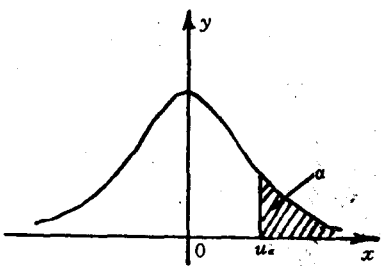

图 1.3

1.3 多维随机变量

在实际当中，随机试验的结果有时需要用两个或两个以上的随机变量来描述，要同时考虑这些随机变量的取值规律和它们之间的联系。称这些在同一随机试验下的多个随机变量为多维随机变量或多维随机向量。以下主要介绍二维情形。

1.3.1 二维随机变量的分布函数

二维随机变量是指有序数组(X,Y)，它的取值是随试验结果而确定的。

1. 联合分布函数

设(X,Y)是二维随机变量，对于任意实数x,y，二元函数

$$F(x,y)=P\{X\leqslant x,Y\leqslant y\}$$

称为二随随机变量(X,Y)的分布函数，或称为X与Y的联合分布函数。

二维分布函数有如下性质：

(1) $P\{x_1<X\leqslant x_2,y_1<Y\leqslant y_2\}=F(x_2,y_2)-F(x_2,y_1)-F(x_1,y_2)+F(x_1,y_1)$

(2) $0\leqslant F(x,y)\leqslant 1$，且

对任意固定的x，$F(x,-\infty)=0$，

对任意固定的y，$F(-\infty,y)=0$，

$F(-\infty,-\infty)=0$，$F(+\infty,+\infty)=1$。

(3) $F(x,y)$对每一个变量都是单调不减、右连续的函数。

2. 边缘分布

二维随机变量(X,Y)作为一个整体，具有分布函数$F(x,y)$，而随机变量X和Y也有各自的分布函数$F_X(x)$和$F_Y(y)$，分别称为二维随机变量(X,Y)关于x和关于y的边缘分布函数。应有

$$F_X(x)=P\{X\leqslant x\}=P\{X\leqslant x,Y<+\infty\}=F(x,+\infty)$$

即
$$F_X(x)=F(x,+\infty)=\lim_{y\to+\infty}F(x,y)$$

同理
$$F_Y(y)=F(+\infty,y)=\lim_{x\to+\infty}F(x,y)$$

1.3.2 二维离散型随机变量的分布律

二维随机变量(X,Y)如所有可能取的值为有限对或可列无穷多对，称为二维离散型随机变量。

设二维随机变量所有可能取的值为(x_i,y_j) $i,j=1,2,\cdots$，若用公式(或表格)给出

$$P\{X=x_i,Y=y_i\}=p_{ij}\quad i,j=1,2,\cdots$$

且具有性质 (1) $p_{ij}\geqslant 0$； (2) $\sum_i\sum_j p_{ij}=1$，

则称它为二维离散型随机变量(X,Y)的分布律(也称为X与Y的联合分布律)。

1.3.3 二维连续型随机变量的概率密度

对二维随机变量(X,Y)的分布函数$F(x,y)$，如存在非负函数$f(x,y)$，使

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\mathrm{d}u\mathrm{d}v$$

则称(X,Y)是二维连续型随机变量。$f(x,y)$称为(X,Y)的概率密度。它应具有性质：

(1) $f(x,y)\geqslant 0$，$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f(u,v)\mathrm{d}u\mathrm{d}v=1$；

(2) $\dfrac{\partial^2 F(x,y)}{\partial x\partial y}=f(x,y)$；

(3) $P\{(x,y)\in G\}=\iint\limits_G f(x,y)\mathrm{d}x\mathrm{d}y$；

G为xoy面上任一区域。

由(X,Y)的联合分布函数与边缘分布函数的关系，有

$$F_X(x)=F(x,+\infty)=\int_{-\infty}^{x}\left[\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y\right]\mathrm{d}x$$

即有
$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y$$

和
$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}x$$

称$f_X(x)$与$f_Y(y)$为(X,Y)关于X与Y的边缘概率密度。

1.3.4 随机变量的独立性

若二维随机变量(X,Y)的分布函数$F(x,y)$与边缘分布函数$F_X(x)$和$F_Y(y)$之间有

$$F(x,y)=F_X(x)F_Y(y)$$

则称随机变量X与Y相互独立。

对连续型随机变量，此条件等价于

$$f(x,y)=f_X(x)f_Y(y)。$$

还可推广到n个随机变量的独立性。设$F(x_1,x_2,\cdots,x_n)$及$F_{X_1}(x_1)$、$F_{X_2}(x_2)$，$\cdots$，$F_{X_n}(x_n)$分别是n维随机变量$(x_1,x_2,\cdots,x_n)$的分布函数及边缘分布函数，如

$$F(x_1,x_2,\cdots,x_n)=F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n)$$

则称随机变量$X_1,X_2,\cdots,X_n$相互独立。

1.4 随机变量的数字特征

随机变量的分布函数能够完整地描述其统计特性。但有些实际问题不需要全面考察随机变量的变化情况，而只需知道能够反映随机变量某些重要特征的数字就够了。例如检验一批化工产品的纯度，不可能将所有的纯度值一一测出，而是关心纯度的平均值，以及纯度对平均值的偏离程度。因偏离程度的大小反映了产品质量的稳定性。这里将介绍几个主要的数字特征。

1.4.1 数学期望

先看一个例子：工厂生产某种零件，每天随机抽取n个检验，若其检查N天，其中出现次品为0个的天数为m_0，次品为1个的天数为m_1，$\cdots$，次品为n个的天数为m_n（显然$m_0+m_1+\cdots+m_n=N$），则平均每天出现的次品件数为

$$\bar{x}=\frac{0\times m_0+1\times m_1+\cdots+n\times m_n}{N}=0\times\frac{m_0}{N}+1\times\frac{m_1}{N}+\cdots+n\times\frac{m_n}{N}=\sum_{K=0}^{n}K\cdot\frac{m_K}{N}$$

其中$\frac{m_K}{N}$是出现K个次品的频率。我们看到次品件数的平均值是件数对于频率的加权平均。对随机变量X来说可以以同样的思路给出X取值的平均值——数学期望。

1. 定义

(1) 设离散型随机变量X的分布律为

$$P\{X=x_K\}=p_K \qquad K=1,2,\cdots$$

若级数$\sum_{K=1}^{\infty}x_Kp_K$绝对收敛，则称级数$\sum_{K=1}^{\infty}x_Kp_K$的和为随机变量$X$的数学期望。记为$E(X)$。

说明：在高等数学中我们知道如级数不绝对收敛，可能因为交换各项的顺序而改变其和。在此若$\sum_{K=1}^{\infty}x_Kp_K$不绝对收敛则造成数学期望无法确定。

(2) 设连续型随机变量X的概率密度为$f(x)$，若积分$\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$绝对收敛，则称积分$\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$的值为随机变量$X$的数学期望，记为$E(X)$。

数学期望也简称为期望或均值，它反映了随机变量取值的平均水平。

例 1.3 计算泊松分布的数学期望。

解：泊松分布的分布律为

$$P\{X=K\}=\frac{\lambda^K\mathrm{e}^{-\lambda}}{K!} \qquad K=0,1,2,\cdots$$

$$E(X)=\sum_{K=0}^{\infty}K\cdot\frac{\lambda^K}{K!}\mathrm{e}^{-\lambda}=\sum_{K=0}^{\infty}K\cdot\frac{\lambda^K}{K!}\mathrm{e}^{-\lambda}$$

$$=\lambda\mathrm{e}^{-\lambda}\sum_{K=0}^{\infty}\frac{\lambda^{K-1}}{(K-1)!}=\lambda\mathrm{e}^{-\lambda}\cdot\mathrm{e}^{\lambda}=\lambda$$

例 1.4 设X服从$[a,b]$上的均匀分布，求$E(X)$。

解：X的概率密度为

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

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x=\int_a^b\frac{1}{b-a}x\mathrm{d}x=\frac{a+b}{2}$$

2. 性质

(1) 若C是常数，则 $E(C)=C$

(2) 若X是随机变量，C是常数，则有

$$E(CX)=CE(X)$$

(3) 若X、Y是两个随机变量，则有

$$E(X+Y)=E(X)+E(Y)$$

(4) 若X、Y是相互独立的随机变量，则有

$$E(XY)=E(X)E(Y)$$

后两条性质都可推广到任意有限个随机变量的情形。

1.4.2 方差

在许多实际问题中，可以看到仅有随机变量的数学期望一个指标是不够的。如有一批电子元件知道平均寿命为 1000 小时。若是所有的元件寿命都在 950～1050 小时之间，我们认为质量是较为可靠的。若是一部分寿命在 1500 小时左右，另一部分寿命在 500 小时左右，我们认为这批元件质量不可靠。这说明我们还需要一个能反映随机变量取值分散程度的量。这就是方差。

1. 定义

设 X 是随机变量，若 $E\{[X-E(X)]^2\}$ 存在，则称其值为随机变量 X 的方差，记为 $D(X)$。

注意到方差的量纲是随机变量 X 量纲的平方。应用中有时用到 $\sqrt{D(X)}$，称为 X 的均方差或标准差，记为 $\sigma(X)$。

计算方差常使用公式

$$D(X)=E(X^2)-[E(X)]^2$$

这是由于

$$\begin{aligned}D(X)&=E\{[X-E(X)]^2\}=E\{X^2-2XE(X)+[E(X)]^2\}\\&=E(X^2)-2E(X)E(X)+[E(X)]^2=E(X^2)-[E(X)]^2\end{aligned}$$

2. 性质

(1) 若 C 是常数，则 $D(C)=0$

(2) 若 X 是随机变量，C 是常数，则有

$$D(CX)=C^2D(X)$$

(3) 若 X、Y 是相互独立的随机变量，则有

$$D(X\pm Y)=D(X)+D(Y)$$

(4) $D(X)=0$ 的充要条件是 X 以概率 1 取常数 C

1.4.3 常用分布的期望与方差

1. (0－1)分布：$E(X)=p$，$D(X)=p(1-p)$

2. 二项分布：$E(X)=np$，$D(X)=np(1-p)$

3. 泊松分布：$E(X)=\lambda$，$D(X)=\lambda$。

4. 均匀分布：$E(X)=\dfrac{a+b}{2}$，$D(X)=\dfrac{1}{12}(b-a)^2$

5. 指数分布：$E(X)=\dfrac{1}{\lambda}$，$D(X)=\dfrac{1}{\lambda^2}$

6. 正态分布：$E(X)=\mu$，$D(X)=\sigma^2$

1.4.4 其它数字特征

1. 矩

设 X 是随机变量，n 是自然数。若 $E(x^n)$ 存在，则称其为 X 的 n 阶原点矩。记为

$$a_n=E(X^n)$$

若 $E\{[X-E(X)]^n\}$ 存在，则称其为 n 阶中心矩，记为

$$\mu_n=E\{[X-E(X)]^n\}$$

显然，X 的数学期望是 X 的一阶原点矩，方差是二阶中心矩。

2. 协方差

(1) 定义：$E\{[X-E(X)][Y-E(Y)]\}$称为随机变量 X 与 Y 的协方差，记为 $COV(X,Y)$。

(2) 性质

① $COV(X,Y)=E(XY)-E(X)E(Y)$

② $D(X\pm Y)=D(X)+D(Y)\pm 2COV(X,Y)$

3. 相关系数

(1) 定义：
$$\rho_{XY}=\frac{COV(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$
称为随机变量 X 与 Y 的相关系数。

(2) 性质：

① $|\rho_{XY}|\leqslant 1$；

② $|\rho_{XY}|=1$ 的充要条件是 X 与 Y 有线性关系的概率为 1。

(即 $P\{Y=aX+b\}=1$)

相关系数 ρ_{XY} 恰是随机变量 X 与 Y 经标准化后的标准随机变量 $X^*=\frac{X-E(X)}{\sqrt{D(X)}}$ 与 $Y^*=\frac{Y-E(Y)}{\sqrt{D(Y)}}$ 的协方差。ρ_{XY} 没有量纲，又不受 X 与 Y 本身取值大小的影响。ρ_{XY} 表示随机变量 X 与 Y 之间线性关系的密切程度。$|\rho_{XY}|$ 较大时，X 与 Y 线性相关的程度较好；$|\rho_{XY}|$ 较小时，X 与 Y 线性相关程度较差。当 $\rho_{XY}=0$ 时，称 X 与 Y 不相关。

当 X 与 Y 相互独立时，$\rho_{XY}=0$，即 X 与 Y 不相关。但当 X 与 Y 不相关时，却不一定相互独立。这是因为“不相关”只表明 X 与 Y 没有线性关系，但还可能有其它关系。